Unit : 1

Probability and Random Variables

Axioms of Probability

1. $0 \leq P(E) \leq 1$
2. $P(S) = 1$
3. For any sequence of mutually exclusive events $E_1, E_2, \ldots,$

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

where $P(E)$ is prob/: of the event E.

Theorems on Probability

* Theorem : 1

The prob/: of an impossible event is zero (or) the null event of prob/: is zero.

(ie) $P(\phi) = 0$.

* Theorem : 2

If A^c is the complementary event of A, then $P(A^c) = 1 - P(A) \leq 1$.

* Theorem : 3

If $B \subset A$, $P(B) \leq P(A)$

* Theorem : 4 (Additive law of probability)

If A and B are any two events, and are not disjoint, then

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

(or) $$P(A \cup B) = P(A) + P(B) - P(AB)$$

* Theorem : 5

If A, B and C are any three events, then

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

Theorem: 6

If $A_1, A_2, \ldots, A_n$ are n mutually exclusive events then the prob/. of happening of one of them,

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$$

Note:

Multiplication Theorem

If two events A and B are independent and can happen simultaneously, the prob of their joint occurrence

$$P(A \cap B) = P(A) \cdot P(B).$$

Theorem: 7

If the events A and B are independent, then

i) $\bar{A}$ and $\bar{B}$ are independent

ii) $\bar{A}$ and B are independent

iii) A and $\bar{B}$ are independent.

Problems:

Find the prob/. that exactly one head appears in a single throw of a fair coin.

Sol

$$P(A) = \frac{n(A)}{n(S)}.$$

$$S = \{H, T\} \Rightarrow n(S) = 2.$$

$$n(A) = 1$$

$$P(A) = 1/2.$$

2) Four persons are chosen at random from a group containing 3 men, 2 women and 4 children. S.T the chance that exactly two of them will be children is $10/21$

Sol:

Total no/. of persons $= 9$

4 persons can be chosen out of 9 person $= {}^9C_4$ ways.

$$= \frac{9 \cdot 8 \cdot 7 \cdot 6}{1 \cdot 2 \cdot 3 \cdot 4} = 126 \text{ way}.$$

The no/: of ways of choosing 2 children out of 4 children $= 4C_2$ ways.

$$= \frac{4 \cdot 3}{1 \cdot 2} = 6 \text{ ways}.$$

The remaing two persons can be chooren from 5 persons (3 m + 2 w) $= 5C_2$ ways

$= 10$ ways.

$2C_2$
$6 \times 3 + 6 \times 1 + 6 \times 3 \times 2$
$\frac{18 + 6 + 36}{126} = \frac{60}{126} = \frac{20}{42} = \frac{10}{21}$

$\therefore$ The no/: of favourable case $= 4C_2 \times 5C_2$

$= 60$ ways

Required probability $= \frac{10}{21}$.

3 Two dice are thrown together. Find prob that

a) the total of the nos/. on the top face is 9

b) the top faces are same.

Sol

a) ~~Let~~ Let A be the event. which gives the sum of the top no: as 9.

$\therefore A = \{(4,5)\ (5,4)\ (3,6)\ (6,3)\}$

$n(A) = 4$

$n(S) = 36$.

$\therefore P(A) = 4/36$.

$2C, 2M = 4C_2$
$2C, 2W$
$2C, 1M, 1W$

b) Let B be the event which gives the top faces are same.

$B = \{(1,1), (2,2), (3,3), (4,4), (5,5)\ (6,6)\}$

$n(B) = 6$.

$$P(B) = \frac{n(B)}{n(S)} = \frac{6}{36} = \frac{1}{6}$$

4) One card is drawn from a deck of 52 cards. What is the prob. of the card being either red or a king.

Sol Let $A = \{$an event that the card drawn is red$\}$

$B = \{$an event that the card drawn is king$\}$

$\therefore P(A) = \frac{n(A)}{n(S)} = \frac{26}{52} = \frac{1}{2}$

$P(B) = \frac{n(B)}{n(S)} = \frac{4}{52} = \frac{1}{13}$.

$n(A \cap B) = n$ (there are two red coloured king cards)

$= 2$.

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

$= \frac{26}{52} + \frac{4}{52} - \frac{2}{52}$

$= \frac{28}{52} = \frac{7}{13}$.

5) If A and B are independent events with $P(A) = 0.4$ and $P(B) = 0.5$, find $P(A \cup B)$.

Sol

$\because$ A and B are independent.

$P(A \cap B) = P(A)\,P(B) = 0.4 \times 0.5 = 0.20$

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

$= 0.4 + 0.5 - 0.20$

$= 0.9 - 0.2 = 0.7$.

Conditional Probability

* **Marginal Prob**

A prob/. of only one event that takes place is called a marginal prob/.

* **Joint Prob/:**

The prob/. of occurence of both events A and B together, denoted by $P(A \cap B)$, is known as joint prob/. of A and B.

* **Conditional Prob**

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \quad \text{if } P(B) \neq 0.$$

Problems

1. A box contains 4 bad and 6 good tubes. Two are drawn out from the box at a time. One of them is tested and found to be good. What is the prob/. that other one is also good?

Sol:

A = One of the tubes drawn is good.

B = Other tube is good.

$P(A \cap B) = P[\text{both tubes are good}]$

$$= \frac{6C_2}{10C_2} = \frac{6 \cdot 5/1 \cdot 2}{10 \cdot 9/1 \cdot 2} = \frac{1}{3}.$$

- Knowing that one tube is good, the conditional prob/. that the other tube is also good is required.

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

$$= \frac{1/3}{6/10} = \frac{1}{3} \times \frac{10}{6} = 5/9$$

2) Given a binary communication channel, where A is the input and B is the output. Let $P(A) = 0.4$, $P(B/A) = 0.9$, $P(\bar{B}/\bar{A}) = 0.6$. Find

i) $P(A/B)$, ② $P(A/\bar{B})$.

Sol

Given $P(A) = 0.4$

$P(B|A) = 0.9$

$\Rightarrow \frac{P(A \cap B)}{P(A)} = 0.9$

$\Rightarrow P(A \cap B) = 0.9 \times P(A) = 0.9 \times 0.4$

$= 0.36.$

Also given: $P(\bar{B}/\bar{A}) = \frac{P(\bar{A} \cap \bar{B})}{P(\bar{A})} = 0.6$

$P(\bar{A} \cap \bar{B}) = 0.6 \times P(\bar{A})$

$1 - P(A \cup B) = 0.6 \, [1 - P(A)]$

$1 - [P(A) + P(B) - P(A \cap B)] = 0.6 \, (1 - 0.4)$

$1 - [0.4 + P(B) - 0.36] = (0.6)(0.6) = 0.36$

$1 - 0.4 - P(B) + \cancel{0.36} = \cancel{0.36}$

$0.6 - P(B) = 0$

$\boxed{P(B) = 0.6}$

$\therefore$ i) $P(A/B) = \frac{P(A \cap B)}{P(B)} = 0.6$

2) $P(A/\bar{B}) = \frac{P(A \cap \bar{B})}{P(\bar{B})} = \frac{P(A) - P(A \cap B)}{1 - P(B)}$

$= \frac{0.4 - 0.36}{1 - 0.6} = 0.1.$

Baye's Theorem

Let $B_1, B_2, \ldots, B_n$ be an exhaustive and mutually exclusive random experiments and A be an event related to that B_i, then

$$P(B_i/A) = \frac{P(B_i)\, P(A/B_i)}{\sum_{i=1}^{n} P(B_i) P(A/B_i)} \; ; i = 1, 2, 3, \ldots, n.$$

1. A bag contains 3 black and 4 white balls. Two balls are drawn at the time without replacement.
   1) What is the prob/: that the 2nd ball drawn is white?
   2) What is the conditional prob/: that the first ball drawn is white if the 2nd ball is known to be white?.

Sol:

Given: 3 black balls, 4 White balls.

Total no/: of balls = 7.

Let A → first ball drawn is white

B → Second ball is white, it happen is two mutually exclusive ways:

1. First ball is white and 2nd ball is white
2. 1st ball is black and 2nd ball is white

i) $P(B) = P(1) + P(2)$

$= \frac{4}{7} \times \frac{3}{6} + \frac{3}{7} \times \frac{4}{6} = \frac{12}{42} + \frac{12}{42} = \frac{24}{42}$

$= 4/7$

2) $P(A/B) = \frac{P(A \cap B)}{P(B)}$

$P(A \cap B) = P(\text{Both Balls are white})$.

$= \frac{4}{7} \times \frac{3}{6} = 2/7$

$\therefore P(A/B) = \frac{2/7}{4/7} = 1/2$.

2. A bag contains 5 balls and it is not known how many of them are white. Two balls are drawn at random from the bag and they are noted to be white. What is the prob/. that all the balls in the bag are white?

Sol:

A bag contains 5 balls. Two balls are drawn at random and found to be white.

$B_1 \rightarrow$ the bag contains 2 W and 3 diff/. colour balls.

$B_2 \rightarrow$ Bag contains 3 W and 2 diff colour balls.

$B_3 \rightarrow$ Bag contains 4 W and 1 diff colour balls

$B_4 \rightarrow$ Bag contains 5 W balls.

Let A be the event of drawing 2 white balls.

Let $P(B_1) = P(B_2) = P(B_3) = P(B_4) = 1/4$.

$$P(A/B_1) = \frac{2C_2}{5C_2} = \frac{1}{10}$$

$$P(A/B_2) = \frac{3C_2}{5C_2} = \frac{3}{10}$$

$$P(A/B_3) = \frac{4C_2}{5C_2} = \frac{6}{10}$$

$$P(A/B_4) = \frac{5C_2}{5C_2} = 1.$$

$P(A) = P($drawing 2 W Balls from all bags$)$

$$= P(B_1) \cdot P(A/B_1) + P(B_2)\,P(A/B_2) + P(B_3)\,P(A/B_3) + P(B_4)\,P(A/B_4)$$

$$= \frac{1}{4} \times \frac{1}{10} + \frac{1}{4} \times \frac{3}{10} + \frac{1}{4} \times \frac{6}{10} + \frac{1}{4} \times 1$$

$$= \frac{20}{40} = \frac{1}{2}.$$

To find $P(B_4/A)$.

(ii) P(2 white balls drawn from the Bag contains 5 w, B)

$= P(B_4/A)$.

By Baye's thm,

$$P(B_4/A) = \frac{P(B_4) \cdot P(A/B_4)}{\left(P(B_1) \cdot P(A/B_1) + P(B_2) \cdot P(A/B_2)\right) + P(B_3) \cdot P(A/B_3) + P(B_4) \cdot P(A/B_4)}$$

$$= \frac{\frac{1}{4} \times 1}{\frac{1}{2}} = \frac{1}{4} \times 2 = \frac{1}{2} = 0.5.$$

3. In a bolt factory, Machines A, B and C manufacture 25%, 35% and 40% of the total output respectively. of the total their output 5%, 4% and 2% are defective bolts. A bolt is drawn at random and is found to be defective. What is the prob% that it was manufactured by Machine B?.

Sol:

Let B_1 be Bolt manufactured from machine A

B_2 be bolt manufactured from machine B

B_3 be bolt manufactured from machine C.

and A be defective bolts.

$\therefore P(B_1) = 25\% = 0.25$

$P(B_2) = 35\% = 0.35$

$P(B_3) = 40\% = 0.40$

$P(A/B_1) = 5\% = 0.05$

$P(A|B_2) = 4\% = 0.04$

$P(A|B_3) = 2\% = 0.02$.

$\therefore$ $P(A) = P($ bolt is drawn and found to be defective$)$

$= P(B_1) P(A|B_1) + P(B_2) \cdot P(A|B_2) + P(B_3) \cdot P(A|B_3)$

$= 0.25 \times 0.05 + 0.35 \times 0.04 + 0.40 \times 0.02$

$= 0.0345$.

To find $P(B_2|A)$

By Baye's thm,

$$P(B_2|A) = \frac{P(B_2) \cdot P(A|B_2)}{P(B_1) \cdot P(A|B_1) + P(B_2) \cdot P(A|B_2) + P(B_3) \cdot P(A|B_3)}$$

$$= \frac{0.35 \times 0.04}{0.0345} = \frac{28}{69}.$$

Binoma

Q1. Two food delivery services "S" and "Z" have 1000 orders and 500 orders respectively per day in a city. Both have some miss-delivery of 1% and 2% per day. In a typical day if a food is ordered by a customer, then

i) What is the probability that the delivery is missing?

ii) If the delivery was missing, what is the probability that the customer has ordered through "S"?

Sol Let S be B_1 and Z be B_2

Total Orders $= 1000 + 500$
$= 1500$

Order in S $= 1000$

$P(B_1) = \frac{1000}{1500}$

$= \frac{2}{3}$

Orders in Z $= 500$

$P(B_2) = \frac{500}{1500}$

$= \frac{1}{3}$

Let x be the missing orders of

each food service

$P(A/B_1) = 1\%$

$= \frac{1}{100}$

$= 0.01.$

$P(A/B_2) = 2\%$

$= 0.02$

i) P (delivery is missing) = P(A) (Total prob. Theorem)

$= P(B_1) \cdot P(A/B_1) + P(B_2) \cdot P(A$

$= \frac{2}{3} \times 0.01 + \frac{1}{3} \times 0.02$

$= 0.0133$

ii) $P(B_1/A) = \frac{P(B_1) \cdot P(A/B_1)}{P(B_1) \cdot P(A/B_1) + P(B_2) \cdot P(A/B_2)}$

$= \frac{2/3 \times 0.01}{0.0133}$

$= 0.5013$

Q3. There is a garage with 5 green 6 blue and 10 black cars. One car comes out, what is the probability that it is green? What are the chances it is black? what are the chances that it is not blue? what are the chances that it is either green or black? What are the chances that it is white? If the car drives away and happens to be blues, and then a second car comes out what are the chances that it is green?

Sl.

No/: of Cars in a garage = 5G + 6B + 10 Black
= 21

i) P (green colour car coming out) = $\frac{5}{21}$

ii) P (Black colour car) = $\frac{10}{21}$

iii) P (not Blue colour car) = P(Green and Black colour car)

$= \frac{15}{21}$

iv) P (white colour car) = 0.

v) P (Green or Black car) = P (Green car) + P (Black car)

$= \frac{5}{21} + \frac{10}{21} = \frac{15}{21}$

vi) if Blue colour car drives away, then

No. of Cars = 5 G + 5 Blue + 10 Black

= 20 cars.

$P(\text{2nd car } \cancel{\text{comes}} \text{ is green colour}) = \frac{5}{20}$

$= \frac{1}{4}$.

Q4. There are two garages A and B. Garage A has 10 green and 5 blue cares. Garage B has 5 green and 10 blue cars. A car coming out of one of the garages happens to be blue. What are the chances it came out of Garage A? What are the chances it came from Garage B? Assume there is no bias in letting cars out of the two garages.

Sol:

Let B_1 be garage A
B_2 be garage B.

Let $P(B_1) = \frac{1}{2}$ and $P(B_2) = 1/2$

Let 'A' be Blue car coming out from Garages.

$\therefore P(A/B_1) = \frac{5}{15} = \frac{1}{3}$

$P(A/B_2) = \frac{10}{15} = \frac{2}{3}$.

Now;

$P(B_1) . P(A/B_1) + P(B_2) P(A/B_2)$

$= \frac{1}{2} \times \frac{1}{3} + \frac{1}{2} \times \frac{2}{3}$

$= \frac{1}{6} + \frac{2}{6} = \frac{3}{6} = \frac{1}{2}$.

i) P (Blue car coming from Garage A)

$= P(B_1/A)$

By Baye's thm;

$$P(B_1/A) = \frac{P(B_1).P(A/B_1)}{P(B_1).P(A/B_1)+P(B_2).P(A/B_2)}$$

$$= \frac{\frac{1}{2}\times\frac{1}{3}}{\frac{1}{2}} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}.$$

(ii) P(Car coming from Garage B).

$$= P(B_2/A)$$

$$= \frac{P(B_2).P(A/B_2)}{P(B_1).P(A/B_1)+P(B_2).P(A/B_2)}$$

$$= \frac{\frac{1}{\cancel{2}}\times\frac{\cancel{2}}{3}}{\frac{1}{2}} = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}.$$

Q6. Box 1 contains 1 white and 999 red balls. Box 2 contains 1 red and 999 white balls. A ball is picked from a randomly selected box. If the ball is red, what is the probability that it came from box 1?

Sol

A = ball from 1^{st} box
B = ball from 2^{nd} box
R = red ball

To find $P(A/R) = ?$

$$P(R) = P(A)\,P(R/A) + P(B)\,P(R/B)$$

$$P(R/A) = \frac{999}{1+999} = \frac{999}{1000}$$

$$P(R/B) = \frac{1}{1+999} = \frac{1}{1000}$$

$$P(A) = P(B) = \frac{1}{2}$$

$$P(R) = P(A)\,P(R/A) + P(B)\,P(R/B)$$

$$P(R) = \frac{1}{2} \times \frac{999}{1000} + \frac{1}{2} \times \frac{1}{1000}$$

$$P(A/R) = \frac{P(A) \cdot P(R/A)}{P(A) \cdot P(R/A) + P(B) \cdot P(R/B)}$$

$$= \frac{\frac{1}{2} \times \frac{999}{1000}}{\frac{1}{2} \times \frac{999}{1000} + \frac{1}{2} \times \frac{1}{1000}} = \frac{\frac{1}{2} \times \frac{999}{1000}}{\frac{1}{2000}\left[999+1\right]}$$

$$= \frac{\frac{1}{\cancel{2}} \times \frac{999}{1000}}{\frac{1}{\cancel{2000}}(\cancel{1000})} = \frac{999}{1000}$$

Random Variable

A real valued fun/. defined on the outcome of a prob/. experiment is called a R.V.

Discrete Random Variable

A random variable whose set of possible value is either finite or countably infinite is called discrete Random variable.

Continuous Random Variable

A random variable X is said to be continuous random variable if it takes all possible values b/w certain limits or in an interval which may be finite or infinite.

Probability fun/. of R.V.

Discrete

1. $P(x_i) \geq 0 \quad \forall \, i = 1, 2, \ldots$

2. $\Sigma P(x_i) = 1$

3. If $x = 0, 1, 2, \ldots, n$

$$P(X < x_i) = P(X = 0) + P(X = 1) + \cdots + P(X = x_1)$$

4) $P(X > x_i) = 1 - P(X \leq x_i)$

$P(X \geq x_i) = 1 - P(X < x_i)$

5) Cumulative distribution:

$F(x) = P(X \leq x)$.

Continuous.

1. $f(x) \geq 0, \quad x \in (-\infty, \infty)$

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

3) $P(X < a) = \int_{-\infty}^{a} f(x)\,dx$

$P(X > a) = \int_{a}^{\infty} f(x)\,dx.$

$P(a \leq x \leq b) = \int_{a}^{b} f(x)\,dx.$

4) $P(X \leq a) = P(X < a)$

$P(X > a) = P(X \geq a)$

$P(a \leq X \leq b) = P(a < X < b)$

$= P(a \leq X < b) = P(a < X \leq b)$

5) $F(x) = \int_{-\infty}^{x} f(x)\,dx$

1) $\frac{d}{dx} F(x) = f(x)$.

$\Rightarrow F'(x) = f(x)$.

2) $\lim_{x \to -\infty} F(x) = 0$; $\lim_{x \to \infty} F(x) = 1$.

3) $P(X = x_i) = F(x_i) - F(x_{i-1})$

4) $P(a < X \le b) = F(b) - F(a)$.

Problems:

1. If RV X takes the values 1, 2, 3 and 4 such that $2P(X=1) = 3P(X=2) = P(X=3) = 5P(X=4)$. Find the p.d.f and cumulative distribution of X.

Sol:

Let $2P(X=1) = k$.

$P(X=1) = k/2$

$3P(X=2) = k$

$P(X=2) = k/3$

$P(X=3) = k$

$5P(X=4) = k$

$P(X=4) = k/5$.

W.K.T Total probability = 1.

$k/2 + k/3 + k + k/5 = 1$.

$k\left(\frac{15+10+30+6}{30}\right) = 1$

$k\left(\frac{61}{30}\right) = 1$

$$\boxed{k = \frac{30}{61}}$$

$\therefore$ Pdf

$X = x$	1	2	3	4	Total
$P(X = x)$	$\frac{15}{61}$	$\frac{10}{61}$	$\frac{30}{61}$	$\frac{6}{61}$	$\frac{61}{61} = 1$.

To find cdf

When $x < 1$.

$$F(x) = 0.$$

When $x = 1$

$$F(1) = P(X \le 1) = 0 + P(X=1) = \frac{15}{61}$$

$$F(2) = P(X \le 2) = 0 + \frac{15}{61} + \frac{10}{61} = \frac{25}{61}$$

$$F(3) = P(X \le 3) = 0 + \frac{15}{61} + \frac{25}{61} + \frac{30}{61} = \frac{55}{61}$$

$$F(4) = P(X \le 4) = \frac{55}{61} + \frac{6}{61} = \frac{61}{61} = 1$$

3 A R.V X has the foll: probability fun,

x:	0	1	2	3	4	5	6	7
$P(X=x)$	0	k	$2k$	$2k$	$3k$	k^2	$2k^2$	$7k^2+k$.

i) Find k

ii) Find $P(X<6)$, $P(X\geq 6)$ and $P(0<X<5)$

iii) Find minimum value of 'a' $\ni$: $P(X\leq a) > 1/2$

iv) Find the distribution fun: of X.

Sol:

W.K.T Total prob: = 1.

$0+k+2k+2k+3k+k^2+2k^2+7k^2+k=1$

$10k^2+9k-1=0$.

$\boxed{k=\frac{1}{10}}$ or $k=-1$ (not possible).

X:	0	1	2	3	4	5	6	7
$P(x)$:	0	$\frac{1}{10}$	$2/10$	$2/10$	$3/10$	$1/100$	$2/100$	$17/100$.

ii) $P(X<6) = P(X=0)+P(X=1)+P(X=2)+P(X=3)$
$+P(X=4)+P(X=5)$.

$= \frac{81}{100}$

ii) $P(X\geq 6) = 1-P(X<6) = 1-\frac{81}{100} = \frac{19}{100}$.

(ii) $P(0<X<5) = P(X=1)+P(X=2)+P(X=3)+P(X=4)$

$= \frac{4}{5}$

iii) $P(X\leq 3) = P(X=0,1,2,3) = \frac{1}{2}$

$P(X\leq 4) = P(X=0,1,2,3,4) = 4/5 > 1/2$

$\therefore\ a = 4$.

iv) The distribution fun/: of x is $F(x)$.

x	0	1	2	3	4	5	6	7
$F(x)$	0	$1/10$	$3/10$	$5/10$	$8/10$	$81/100$	$\frac{83}{100}$	1

4) If c.d.f of R.V X is given by

$$F(x) = \begin{cases} 1 - 4/x^2, & \text{if } x > 2 \\ 0, & \text{if } x \le 2 \end{cases}$$

find (i) $P(X < 3)$ ii) $P(4 < X < 5)$ iii) $P(X \geq 3)$

Sol:

Given: $F(x) = 1 - \frac{4}{x^2}$, $x > 2$.

$$f(x) = \frac{d}{dx} F(x) = 0 - 4(-2)x^{-2-1}$$

$$= 8x^{-3} = 8/x^3.$$

i) $$P(X<3) = \int_{-\infty}^{3} f(x)dx = \int_{2}^{3} 8/x^3\, dx = \cancel{8}^{4}\left[\frac{x^{-2}}{\cancel{-2}}\right]_2^3$$

$$= -4\left[\frac{1}{x^2}\right]_2^3 = -4\left[\frac{1}{9} - \frac{1}{4}\right]$$

$$= 5/9$$

ii) $$P(4 < X < 5) = \int_4^5 f(x)dx = 8\int_4^5 x^{-3}dx \ne \frac{9}{100}$$

$$= 8\left[\frac{x^{-2}}{-2}\right]_4^5 = -4\left[\frac{1}{25} - \frac{1}{16}\right]$$

$$= \frac{9}{100}$$

iii) $$P(X \geq 3) = 1 - P(X<3) = 1 - \frac{5}{9} = \frac{4}{9}.$$

Binomial distribution

$$P(X=x) = nC_x \, p^x q^{n-x};$$

The Binomial distribution is a discrete distribution with parameters n and p. If p and q are equal it is symmetrical, otherwise it is non-symmetrical.

here $n \to$ no/: of trials.
$x \to$ no/: of success.
$p \to$ probability of success.
$q \to$ probability of failure.

and $p+q = 1$.

$nC_r \to$ no/: of combinations of n things taken r at time.

$$= \frac{n!}{(n-r)!\, r!}$$

~~The pars~~

Assumptions of Binomial distribution

* The no/: of observations is fixed.
* Each trial has two mutually exclusive possible outcomes (success or failure)
* Each trial is independent of other trials.

Characteristics of Binomial distribution

1. The discrete probability distribution which is based on Binomial theorem is called Binomial distribution.
2. Parameters are (n and p)

3) $\boxed{\begin{array}{l} \text{mean} = np \\ \text{Variance} = npq \end{array}}$

4) if $p = q = \frac{1}{2}$ $\Rightarrow$ symmetric distribution.

5) ~~if, $p > 1/2$ and if $p < 1/2$~~.

5) if $p > 1/2$ $\Rightarrow$ negatively skewed distribution.

if $p < 1/2$ $\Rightarrow$ positively skewed distribution.

6) Binomial distribution tends to Poisson distribution if n is very large. and p is very small.

7) Binomial distribution tents to Normal distribution if n is very large ; p and q are not small.

Applications of B.D

1. It is used to find prob/: of getting x succes in 'n' independent Bernoulli trials.

2. It is also used in sampling, inspection plans, genetic experiments.

Poisson distribution

$$P(X=x) = \frac{e^{-m} \cdot m^x}{x!} = \frac{e^{-\lambda} \cdot \lambda^x}{x!}$$

where $x \rightarrow$ no/. of success.

$\lambda \rightarrow$ mean $= n \cdot p$.

Assumptions

1. The no/: of events is discrete in a given interval.

2. The prob/: is approximately proportional to the length of the interval for an event which may occur in a short interval $[t, t+\Delta t]$

3. The occurrences of events are independent in non-overlapping intervals.

4. The prob/: of two events is negligible in a short interval $[t, t+\Delta t]$.

Characteristic of Poisson Distribution

1. The outcomes that occur in the result of the experiment can be classified as success or failures.

2. $\boxed{\begin{array}{l} \text{Mean} = \lambda. \\ \text{Variance} = \lambda \end{array}}$

3. Sum of two Poisson variables X and Y is also Poisson variate with parameter $\lambda_1 + \lambda_2$

4. $\text{Skewness} = \frac{1}{\sqrt{\lambda}}$

$\text{Kurtosis} = \frac{1}{\lambda}$.

Applications of Poisson Distribution.

1. To count the no/: of defects of an item used in quality control statistics.

2. To used to find the no/: of typing errors per page in a typed material

Problems

1. Fit a binomial distribution to foll/: distribution of 156 samples:

Sol:

$$P(X=x) = nC_x\, p^x q^{n-x} = 7C_x\, p^x q^{7-x}.$$

No/: of Defective items	7	6	5	4	3	2	1	0
No/: of samples.	1	6	32	36	48	24	7	2

here $n = 7$; $N = 156$.

To find p.

W.k.T mean $= np$.

$$\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i} = np = 7p.$$

$$\Rightarrow 7 \cdot p = \frac{(7\times1)+(6\times6)+(5\times32)+(4\times36)+(3\times48)+(2\times24)+(1\times7)+(0\times2)}{1+6+32+36+48+24+7+2}.$$

$$= \frac{546}{156} = 3.5.$$

$$\Rightarrow p = \frac{3.5}{7} = 0.5$$

$$q = 1-p = 1-0.5 = 0.5.$$

$$\therefore P(X=x) = 7C_x\,(0.5)^x (0.5)^{7-x}.$$

$$= 7C_x\,(0.5)^{x+7-x}$$

$$= 7C_x\,(0.5)^7.$$

x	$P(x) = 7C_x\,(0.5)^7$	$E(x) = N\,P(x) = 156 \times P(x)$
0	$7C_0\,(0.5)^7 = 0.0078125$	$156\times0.0078125 = 1.22$

1	$7C_1 (0.5)^7 = 0.0547$	8.53
2	$7C_2 (0.5)^7 = 0.1641$	25.59
3	$7C_3 (0.5)^7 = 0.2734$	42.65.
4	$7C_4 (0.5)^7 = 0.2734$	42.65
5	$7C_5 (0.5)^7 = 0.1641$	25.59
6	$7C_6 (0.5)^7 = 0.0547$	8.53
7	$7C_7 (0.5)^7 = 0.0078$	1.22

2. The mean and variance of a Binomial variable X are 4 and 2 respectively. Find the prob. that X takes values greater than 3.

Sol:

Given: X follow Binomial distribution.

mean $= 4$ Variance $= 2$

$\Rightarrow np = 4 \rightarrow (1)$ $npq = 2 \rightarrow (2)$

$$\frac{(1)}{(2)} \Rightarrow \frac{np}{npq} = \frac{4}{2}$$

$$\frac{1}{q} = 2 \Rightarrow \boxed{q = \frac{1}{2}}$$

$$\therefore \boxed{p = 1-q = 1-\frac{1}{2} = 1/2}$$

$$P(X=x) = nC_x\, p^x\, q^{n-x}.$$

$$= nC_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{n-x}$$

Sub/: $p = 1/2$ in (i).

$$n \times \frac{1}{2} = 4$$

$$\Rightarrow \boxed{n = 8}$$

$$\therefore P(X = x) = 8C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{8-x} = 8C_x \left(\frac{1}{2}\right)^8$$

$$P(X > 3) = P(X = 4, 5, 6, 7, 8)$$

$$= P(X=4) + P(X=5) + P(X=6) + P(X=7) + P(X=8)$$

$$= 8C_4 \left(\frac{1}{2}\right)^8 + 8C_5 \left(\frac{1}{2}\right)^8 + 8C_6 \left(\frac{1}{2}\right)^8 + 8C_7 \left(\frac{1}{2}\right)^8 + 8C_8 \left(\frac{1}{2}\right)^8$$

$$= 0.636.$$

3. The incidence of occupational disease in an industry is such that the workers have a 20% chance of suffering from it. What is the prob/. that out of six workers 3 or more will contract the disease?

Sol:

Given: $p = 20\% = 0.20$; $q = 0.80$

$n = 6$.

$$P(X = x) = nC_x \, p^x \, q^{n-x}$$

$$= 6C_x \, (0.20)^x (0.80)^{6-x}.$$

$$P(3 \text{ or more}) = P(X \geq 3)$$

$$= P(X = 3, 4, 5, 6)$$

$$= P(X=3) + P(X=4) + P(X=5) + P(X=6).$$

$= 6C_3 (0.20)^3 (0.80)^3 + 6C_4 (0.20)^4 (0.80)^2$

$+ 6C_5 (0.20)^5 (0.80) + 6C_6 (0.20)^6$

$= 0.098$.

Differnce b/w Binomial and Poisson distribution

Binomial	Poisson
1. $P(x) = nC_x\, p^x q^{n-x}$	1. $P(x) = \dfrac{e^{-\lambda} \cdot \lambda^x}{x!}$
2. n and p are parameters	2. λ is parameter
3. mean $= np$ variance $= npq$	3. mean $= \lambda$ variance $= \lambda$.
4. No/. of trials is fixed.	4. No/. of trial is infinite

3. A Machine manufacturing screws is known to produce 5% defective. In a R.S of 15 screws, what is the prob that there are i) exactly 3 defectives ii) not more than 3 defectives.

Sl:

$$p = \frac{5}{100} = 0.05 \qquad P(X=x) = nC_x\, p^x q^{n-x}$$

$$= 15C_x (0.05)^x (0.95)^{x}.$$

$$q = \frac{95}{100} = 1-P \;; \quad n = 15.$$

i) $P(\text{exactly 3 defectives}) = P(X=3)$

$$= 15C_3 (0.05)^3 (0.95)^{12}$$

$$= 0.0307.$$

ii) $P(\text{not more than 3 defectives}) = P(X \le 3).$

$$= P(X = 0,1,2,3)$$

$$= 15C_0 (0.05)^0 (0.95)^{15} + 15C_1 (0.05)^1 (0.95)^{14}.$$

$$+ 15C_2 (0.05)^2 (0.95)^{13} + 15C_3 (0.05)^3 (0.95)^{12}$$

$$= 0.9944$$

3. Out of 800 families with 4 children each, how many families would be expected to have i) 2 boys and 2 girls (ii) atleast 1 boy iii) almost 2 girls and (iv) Children of both genders. Assume equal prob/: for boys and girls

Sl.

Let $p = q = \frac{1}{2}$; $N = 800$; $n = 4$.

$$P(X=x) = nC_x\, p^x q^{n-x} = 4C_x (0.5)^x (0.5)^{4-x}.$$

$$= 4C_x (0.5)^4$$

$p \rightarrow$ prob/: of success which i.e being boy.

i) $P(\text{2 boys and 2 girls}) = P(\text{2 boys}) = P(X=2)$

$$= 4C_2 (0.5)^4 = 3/8.$$

No/. of faimilies $= N\,P(X=x) = 800 \times 3/8 = 300$

ii) P(at least one boy) = $P(X \geq 1)$

$= 1 - P(X < 1)$

$= 1 - P(X = 0)$

$= 1 - \left(\frac{1}{2}\right)^4 = 15/16$.

No/: of families $= 800 \times \frac{15}{16} = 750$.

iii) P(atmost 2 girls) = P[atleast 2 boys]

$= P[X \geq 2] = 1 - P[X < 2]$.

$= 1 - \{P(X=0) + P(X=1)\}$

$= 1 - \left\{\left(\frac{1}{2}\right)^4 + 4\left(\frac{1}{2}\right)^4\right\}$

$= 11/16$.

No/: of families $= 800 \times \frac{11}{16} = 550$.

iv) P(Children of both genders) = 1 − P(Children of same gender)

= 1 − {P(all are boys) + P(all are girls)}

$= 1 - \{P(X=4) + P(X=0)\}$

$= 1 - \left[\left(\frac{1}{2}\right)^4 + \left(\frac{1}{2}\right)^4\right]$

$= \frac{7}{8}$.

No/: of families $= \frac{7}{8} \times 800 = 700$.

CamScanner

1. If X is a Poisson Variate such that $2P(X=0)+P(X=2) = 2P(X=1)$, find $E(X)$.

Sol:

$$P(X=x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!}$$

Given $2P(X=0)+P(X=2) = 2P(X=1)$

$$2\frac{e^{-\lambda} \cdot \lambda^0}{0!} + \frac{e^{-\lambda} \cdot \lambda^2}{2!} = 2\frac{e^{-\lambda}\lambda}{1!}$$

$$2e^{-\lambda} + \frac{e^{-\lambda} \cdot \lambda^2}{2} = 2e^{-\lambda} \cdot \lambda.$$

(X) by $2e^{\lambda}$.

$$4 + \lambda^2 = 4\lambda.$$

$$\lambda^2 - 4\lambda + 4 = 0.$$

$$(\lambda-2)^2 = 0$$

$$\boxed{\lambda = 2}$$

$$E(X) = \text{mean} = \lambda = 2.$$

2. The no/. of monthly breakdown of a computer is a RV having a Poisson dist/. with mean equal to 1.8. Find the prob/. that this computer will fun/. for a month. i) without a breakdown ii) with only one breakdown iii) with atleast one breakdown.

Sol:

Given: mean $= \lambda = 1.8$

$$P(X=x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!} = \frac{e^{-1.8}(1.8)^x}{x!}$$

i) $P(X=0) = \frac{e^{-1.8}(1.8)^0}{0!} = e^{-1.8} = 0.1653$

ii) P(with only one breakdown) $= P(X=1)$

$$= \frac{e^{-1.8}(1.8)^1}{1!} = 0.2975$$

iii) P(with atleast 1 breakdown) $= P(X \geq 1) = 1 - P(X < 1)$

$$= 1 - P(X = 0)$$

$$= 1 - 0.1653 = 0.8347.$$

Q7. Suppose we investigate the safety of a dangerous intersection. Past police records indicate a mean of five accidents per month at this intersection. The number of accidents is distributed according to poisson distribution and the highway safety Division wants us to calculate the probability in any month of exactly 0, 1, 2, 3 or 4 accidents. Calculate the probability of more than 3 accidents.

Sol

$\lambda = 5$

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x} \quad x = 0, 1, 2$$

$$= \frac{e^{-5}5^{x}}{x} \quad x = 0, 1, 2$$

$$P(X = 0) = \frac{e^{-5}5^{0}}{0} = 0.006738$$

$$P(X = 1) = \frac{e^{-5}5^{1}}{1} = 0.03369$$

$$P(X = 2) = \frac{e^{-5}5^{2}}{2} = 0.08422$$

$$P(X = 3) = \frac{e^{-5}5^{3}}{3} = 0.14037$$

$$P(X=4) = \frac{e^{-5}5^4}{4}$$

$$= 0.17547$$

$$P(x > 3) = 1 - P(0, 1, 2 \text{ or } 3)$$

$$= 1-(0.006738 + 0.03$$
$$+ 0.08422 + 0.14$$

$$= 0.73498$$

Q5. A blade manufacturer manufac-tures and supplies blades in packets of 10. There is a 0.2% probability for any blade to be defective. Find approximately the number of packets containing two defective blades in a consignment of 20,000 packets.

Sol:

$n = 10$; $N = 20000$

$p = 0.2\% = \frac{0.2}{100} = 0.002$.

Binomial.

$P(X=x) = nC_x\, p^x q^{n-x}$.

Poisson distribution

$$P(X=x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!}$$

where $\lambda = np = 10 \times 0.002$
$= 0.02$

$$P(X=2) = \frac{e^{-0.02} \cdot (0.02)^2}{2!}$$
$= 0.000196$

No. of Packets containing two defective Blades $= N \times P(X=2)$
$= 20,000 \times 0.000196$.
$= 3.9207$
≈ 4 packets

9a. The following table gives the number of days in a 60-days period during which road accidents occurred in a own. find poisson distribution to the data

No. of accidents	0	1	2	3	4
No. of days	31	18	7	3	1

Sol

$N = 60$

W.K.T, The poisson distribution

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$\lambda = \text{mean} = \frac{\Sigma f_i x_i}{\Sigma f_i}$$

$$= \frac{0+18+14+9+4}{31+18+7+3+1}$$

$$= \frac{45}{60} = 0.75$$

$$\lambda = 0.75$$

$$P(X = x) = \frac{e^{-0.75}(0.75)^{x}}{x!}$$

Expected frequency $= NP(X=x)$

$$E(x) = 60\ \frac{e^{-0.75} \cdot 0.75^{x}}{x!}$$

$$E(0) = \frac{60 \times e^{-0.75} \cdot 0.75}{0!} = \frac{60 \times e^{\wedge}(0.75) \times 1}{1}$$

$$= 60 \times 0.4724 = 28.34$$

$$E(1) = \frac{60 \times e^{-0.75} \times 0.75}{1!} = 60 \times e^{0.75}$$

$$= 60 \times 0.3543 = 21.26$$

$$E(2) = \frac{60 \times e^{-0.75} \, 0.75}{2!} = 60 \times 0.1329$$

$$= 7.97$$

$$E(3) = \frac{60 \times e^{-0.75} \times 0.75^{3}}{3!} = 60 \times 0.332$$

$$= 1.99$$

$$E(4) = \frac{60 \times e^{-0.75} \times 0.75^{4}}{4!} = 60 \times 0.0062$$

$$= 0.37$$

Q8. i) A Sales representative can convert a customer as potential buyer with the probability of 70%. If he is able to meet the 10 customers in day, find the probability of converting.

A. atleast one customer
B. Not even a single customer
C. exactly one customer.

ii) $N = 1000$; $n = 5$; $p = 50\%$. then find $P(X = 2)$ by binomial distribution.

Sol i) Let x be the no. of customers

$$n = 10$$
$$p = \frac{70}{100} = 0.7$$

$$P(X = x) = nC_x \, p^x q^{n-x}$$
$$= 10C_x (0.7)^x (0.3)^{10-x},$$
$$x = 0, \cancel{x}1, 2 \ldots : 10$$

1) Probability of converting at least one customer

$= \cancel{10 C_x}$

TNPL

$= P(x \geq 1) = 1 - P(X = 0)$
$= 1 - 10C_0 (0.7)^0 (0.3)^{10}$
$= 0.9999$

(2) Probability not even a single custo

$= P(X = 0) = 10C_0 (0.7)^0 (0.3)^{10}$
$= 0.0000059$

(3) Exactly one customer

$= P(X = 1)$
$= 10C_1 (0.7)^1 (0.3)^9$
$= 0.0001378$

ii) $N = 1000$
$n = 5$
$p = 50\%$

$P(X = x) = nC_x p^x q^{n-x}$

$= 5C_x (0.5)^x (0.5)^{5-x}$;
$x = 0, 1, 2, \ldots, 5$

$P(X = 2) = 5C_2 (0.5)^2 (0.5)^3$

$= 0.3125$

Normal distribution

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

If x is normally distributed Random Variable and μ and σ are mean and S.D, then, normal curve is defined as

$$z = \frac{x-\mu}{\sigma}.$$

parameters are ~~N~(μ,~~ (μ and σ)

1. The daily income of 2000 employees is normally distributed around a mean of Rs. 80 and with S.D of ₹5. Estimate the no/: of employees whose daily income will be.

   i) B/w ₹80 and ₹82.

   ii) More than ₹85.

   iii) Less than ₹73.

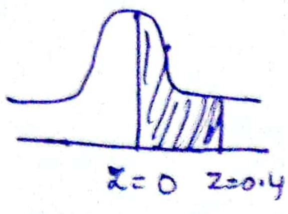
Sol:

$$z = \frac{\bar{x}-\mu}{\sigma} = \frac{\bar{x}-80}{5}.$$

$N = 2000$.

i) $P(80 < x < 82) = P(0 < x < 0.4) = 0.1554$

$x = 80 \Rightarrow z = \frac{80-80}{5} = 0$

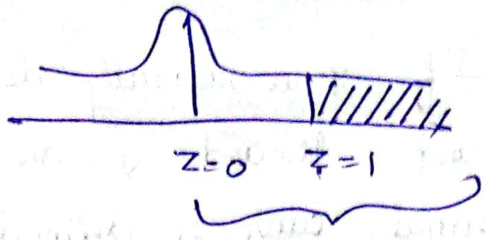


$x = 82 \Rightarrow z = \frac{82-80}{5} = 0.4$

No/: of employees $= 2000 \times 0.1554$
$= 311$

ii) $P(X>85) = P(z>1) = 0.5 - P(0<z<1)$

$X = 85 \Rightarrow z = \frac{85-80}{5} = 1$



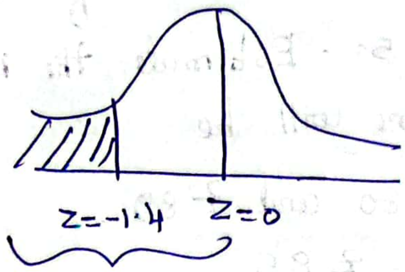
0.5

$= 0.5 - 0.\cancel{4192}\,3413 = \cancel{0.0808}\ 0.1587$

No/: of employees $= 2000 \times 0.1587 = 317$

iii) $X = 73 \Rightarrow z = \frac{73-80}{5} = -1.4$

$P(X<73) = P(z<-1.4)$



0.5

$= 0.5 - P(-1.4<z<0)$

$= 0.5 - P(0<z<1.4)$

$= 0.5 - 0.4192 = 0.0808$.

No/: of employees $= 2000 \times 0.0808 = 162$

Q2. A cable TV operator collects monthly charge Rs. 300 as an average and standard deviation of Rs. 100 Monthly charge will vary by normal distribution since it depends on the choices of package by customer. If a household is choosen at random what is the probability that the customer pays

i) Below Rs 200
ii) Rs 200 - 400
iii) Above Rs 400

Sol. S.D = 100
M = 300

Normal Distribution

$$Z = \frac{x - M}{\sigma}$$

i) Below Rs 200

$x = 200$

$$Z = \frac{200 - 300}{100}$$

$= 1,$

$P(x < 200) = P(z < -1)$

$P(z < -1) = 0.5 - P(-1 < z < 0)$

$= 0.5 - P(0 < z < 1)$

$= 0.5 - 0.3413$

$= 0.1587$

ii) Rs 200 - 400

$x = 200 \quad Z = \frac{200 - 300}{100} = 1$

$x = 400 \quad z = \frac{400 - 300}{100} = 1$

$P(200 < x < 400) = P(-1 < x < 1)$

$P(-1 < z < 0) + P(0 < z < 1)$

$= 0.3413 + 0.3413$

$= 0.6826$

iii) Above Rs 400

$(x > 400)$

$x = 400 \quad z = \frac{400 - 300}{100} = 1$

$P(x > 400) = P(z > 400)$

$= 0.5 - P(0 < z < 1)$

$= 0.5 - 0.3413$

$= 0.1587.$

1 An electrical firm manufactures light bulbs that have a life, before burn out, that is normally distributed with mean equal to 800 hrs and a S.D of 40 hrs. Find

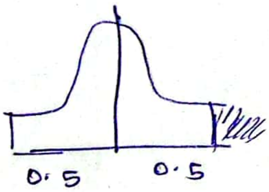
i) prob that bulb burns more than 834 hrs.

ii) prob that bulb burns b/w 778 and 834 hrs.

Sol:

Given X follow N.D with $\mu = 800$ and $\sigma = 40$.

Now $Z = \frac{X-\mu}{\sigma} = \frac{X-800}{40}$.



i) $P(X > 834) = P(Z > 0.85)$.

$= 0.5 - P(z < 0.85)$

$= 0.5 - P(0 < z < 0.85)$.

$= 0.5 - 0.3023 = 0.1977$.

ii) $P(778 < X < 834)$

$= P[-0.55 < z < 0.85]$

$= P[-0.55 < z < 0] + P[0 < z < 0.85]$

$= P[0 < z < 0.55] + P[0 < z < 0.85]$

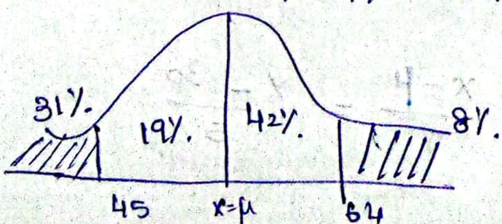
$= 0.2088 + 0.3023 = 0.5111$.

2. In a normal dist^{n}: 31% of the items under 45 and 8% are over 64. Find μ and σ.

Sol:

Given: $P(X < 45) = 31\% = 0.31$

$P(X > 64) = 8\% = 0.08$.



W.K.T $Z = \frac{X-\mu}{\sigma}$.

When $X = 45$, take $Z = Z_1$

$X = 64$, take $Z = Z_2$.

$P(X < 45) = P(Z < Z_1) = 0.31$.

$\Rightarrow P(Z_1 < Z < 0) = 0.19$.

$\Rightarrow$ from table $Z_1 = -0.49$.

(ie) $\frac{45-\mu}{\sigma} = -0.49$.

$45 - \mu = -0.49\sigma$

$\mu + 0.49\sigma - 45 = 0. \rightarrow (1)$

|||ly $P(X > 64) = P(Z > Z_2) = 0.08$.

$\Rightarrow P(0 < Z < Z_2) = 0.42$

from table, $Z_2 = 1.40$.

$\frac{64-\mu}{\sigma} = 1.40$.

$64 - \mu = 1.40\sigma$.

$\mu + 1.40\sigma - 64 = 0. \rightarrow (2)$

By solving, these two

we get $\sigma = 10$; $\mu = 50$.

and $\sigma^2 = 100$

3. X is a normal variate with mean 30 and S.D 5. Find prob that

i) $26 \le X \le 40$ ii) $X \ge 45$ iii) $|X - 30| > 5$

Sol:

$Z = \frac{X-\mu}{\sigma} = \frac{X-30}{5}$.

i) $P(26 \le X \le 40) = P(-0.8 \le Z \le 2)$

$= P(-0.8 \le Z \le 0) + P(0 \le Z \le 2)$

$= P(0 \le Z \le 0.8) + P(0 \le Z \le 2)$

$= 0.2881 + 0.4772$

$= 0.7653$.

ii) $P(X \geq 45) = P(Z \geq 3)$

$= 0.5 - P(Z \le 3)$

$= 0.5 - P(0 \le Z \le 3)$

$= 0.5 - 0.4987 = 0.0013$.

iii) $P(|X - 30| \le 5) = P(25 \le X \le 35)$

$= P(-1 \le Z \le 1) = 2P(0 \le Z \le 1)$

$= 2 \times 0.3413 = 0.6826$.

iv) $P(|X - 30| \geq 5) = 1 - P(|X - 30| \le 5)$

$= 1 - 0.6826 = 0.3174$ //.

Uniform distribution:

$$f(x) = \frac{1}{b-a} \; ; \quad a \le x \le b.$$

$$\text{Mean} = \frac{a+b}{2}.$$

$$S.D = \frac{b-a}{\sqrt{12}} \quad ; \quad \text{Variance} = \frac{(b-a)^2}{12}.$$

Note

$$P(X_1 \le x \le X_2) = \frac{X_2 - X_1}{b-a}.$$

1\. A random variable X has a uniform distribution over $(-3,3)$ compute.

i) $P(X<2)$ ii) $P[|X|<2]$ iii) $P(|X-2|<2)$.

iv) find k for which $P(X>k) = 1/3$.

Sol:

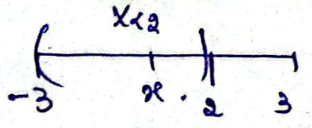
Given: X is a Uniform random variable.

$$f(x) = \frac{1}{b-a} = \frac{1}{3-(-3)} = \frac{1}{6}.$$

W.K.T.

$$P(x_1 < X < x_2) = \frac{X_2 - X_1}{b-a} = \frac{X_2 - X_1}{6}.$$

i) $P(X<2) = P(-3 < X < 2) = \frac{2-(-3)}{6}$



$$= \frac{5}{6}$$

ii) $P(|X|<2) = P(-2 < X < 2) = \frac{2-(-2)}{6}$

$$= \frac{4}{6} = \frac{2}{3}$$

(iii) $P(|x-2| < 2)$

$$= P(-2 < x-2 < 2)$$

$$= P(-2+2 < x-2+2 < 2+2)$$

$$= P(0 < x < 4)$$

$$= \cancel{\frac{4-0}{6}} = \cancel{k}$$

$$= P(0 < x < 3)$$

$[\because x \text{ lies b/w } (-3,3)]$

$$= \frac{3-0}{6} = \frac{3}{6} = \frac{1}{2}.$$

(iv) Given: $P(x > k) = \frac{1}{3}$

$$\Rightarrow P(k < x < 3) = \frac{1}{3}.$$

$$\Rightarrow \frac{3-k}{6} = \frac{1}{3}$$

$$\Rightarrow 3-k = \frac{6}{3} = 2.$$

$$\Rightarrow 3-k = 2$$

$$\Rightarrow -k = 2-3 = -1$$

$$\Rightarrow \boxed{k=1}$$

10. The national Association of Insurance Commissioners, India conducted a survey in which it found that on an average the automobiles are insured for the amount of ₹ 691 yearly. Let the insurance cost be uniformly distributed in the country with in a range of ₹ 200 to ₹ 1.182, then calculate

1) Standard deviation
2) Height of distribution
3) Probability of a person spending ₹ 410 to ₹ 825 for the automobile insurance.

Sol 1) Standard deviation

$\overline{x} = 691$

$a = 200$

$b = 1,182$

$$\sigma = \frac{b-a}{\sqrt{12}} = \frac{1,182-200}{\sqrt{12}}$$

$$= \frac{982}{\sqrt{12}}$$

$$= 283.5$$

ii) The height of distribution

$$= \frac{1}{b-a} = \frac{1}{1,182-200}$$

$$= \frac{1}{982} = 0.001$$

iii) Probability of a person spending ₹ 410 to ₹ 825 for the automobile

$x_1 = 410$

$x_2 = 825$

$$P(410 \le x \le 825) = \frac{825-410}{1,182-200}$$

$$= \frac{415}{982}$$

$$= 0.4226$$

PART-A (Additional Questions)

1. Let X be the lifetime in years of a mechanical part. Assume that X has the cdf $F(x) = 1 - e^{-x}$. Find $P[1 < X \leq 3]$.

Sol:

Given $F(x) = 1 - e^{-x}$.

$$f(x) = \frac{d}{dx} F(x) = 0 - e^{-x}(-1)$$
$$= e^{-x}$$

$$P(1 < X \leq 3) = F_X(3) - F_X(1)$$
$$= (1 - e^{-3}) - (1 - e^{-1})$$
$$= 0.318.$$

2. Suppose that X has a poisson distribution with parameter $\lambda = 2$. Compute $P(X \geq 1)$.

Sol:

X is a poisson variate.

$$\therefore P(X = x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!} = \frac{e^{-2} \cdot 2^x}{x!}$$

$$P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0)$$
$$= 1 - \left[\frac{e^{-2}\, 2^0}{0!}\right] = 1 - 0.1353$$
$$= 0.\cancel{0}8647$$

3. The average no/: of defective chips manufactured daily at a plant is 5. Assume the no/: of defects is a poisson random variable X.

compute mean and variance of x if $P(X=0) = 0.0497$.

Sol:

Given: $P(x=0) = 0.0497$;

$$P(X=x) = \frac{e^{-\lambda} \cdot \lambda^{x}}{x!}$$

$$P(X=0) = \frac{e^{-\lambda} \cdot \lambda^{0}}{0!} = e^{-\lambda} = 0.0497.$$

$$-\lambda = \log(0.0497)$$

$$-\lambda = -3.0018$$

$$\lambda = 3.0018.$$

$\Rightarrow$ mean $= \lambda = 3.0018$

$\Rightarrow$ variance $= \lambda = 3.0018$.

4. Give an example for discrete and Continuous Random variable.

Discrete RV [Binomial distribution, Poisson distribution]

* The no/: of tables in Restaurant or no/: of rooms in a lodge.

continuous RV [Normal and Uniform distribution]

* The age of students in a school, heights and weight etc.

CamScanner

Chapter: 2 Sampling distribution and Estimation.

The sampling distribution of a statistic is the probability distribution of all possible values the statistic may take, when computed from random sample of same size, drawn from a specified population.

Standard Error (S.E)

The S.D of the sampling distribution of a statistic is of particular importance in test of hypothesis and is called the standard error of the statistic

Properties of Sampling distribution

1. The mean of the population and the sampling distributions mean are equal.

2. According to the normal distribution, the categorisation of the population mean in terms of standard deviation

   i) Approximately 68% of all sample means into 1 standard deviation.

   ii) 95% into 1.96 standard deviations

   and iii) 99% into the 3 standard deviation.

3. The S.E of the mean is the standard deviation of the sampling distribution.

Sampling Techniques

Types of Sampling Design.

- Probability sampling
  - Simple Random sample.
  - Systematic Sample
  - Stratified Sample.
  - Cluster Sample.
- Non-Probability Sampling:
  - Convenience Sampling
  - Panel sampling
  - Purposive sampling
  - Snowball sampling

Simple Random Sampling:

This is the most famous and simple method of sampling where each unit of the population is equally probable of getting included in the sample.

Simple random sampling says that:

1. There is an equal chance for each elt of the population to be included in the sample and the choices are independent to each other.

2. Each possible sample combination has an equal chance of being chosen.

Methods of Simple Random Sampling

1. Lottery method
2. By using random numbers

Advantages of simple Random sampling

* Freedom from Basis
* Representativeness
* Ease of sampling and Analysis.

Disadvantages of Simple Random Sampling

* Simple random sampling usage is limited by cost factor.
* Availability of a current Listing of universe elements.
* Statistical Efficiency.
* Administrative Difficulties.

Systematic Sampling

This method is applicable when the size of the population is finite and on the basis of any system the units of the universe are arranged such as ~~ap~~ alphabetic arrangement, numerical arrangement or geographical arrangements.

Advantages:

1. Simple and convenient
2. Gives similar Results.
3. Independent.
4. Little chance of Biasness.
5. Helps in Random selection

Disadvantages:

1. High sampling error.
2. Possibility of selecting impracticable units.
3. Biased.
4. Not suitable for large population.

Stratified Random Sampling

In the stratified random sampling, the sample is selected from different homogeneous strata or parts of a universe instead of heterogeneous universe as a whole. The summary of this sampling procedure are as follows:

1. The sampled universe is divided (or stratified) into groups that are mutually exclusive and include all items in the universe.

2. A simple random sample is then chosen independently from each group or stratum.

Advantages:

1. More representatives.
2. Certainty
3. Greater Precision
4. Administrative convenience.

Disadvantages:

1. Needs more attention
2. Time consuming
3. Complicated
4. Expensive.

Cluster sampling

According to this method there is further noticeable sub-division of the universe into clusters. Simple random sampling is performed and clusters are drawn accordingly constituting a sample of all the units belonging to the selected clusters.

Advantages:

1. Cheap, quick and easy.
2. Larger sample size
3. Convenient to obtain
4. Cost Effective.

Disadvantages

1. Least Representative
2. High Sampling Error
3. Less Efficient
4. Sometimes not appropriate.

Difference b/w stratified and cluster sampling

Stratified Sampling	Cluster sampling
* One divide the population into a few sub-group. i) There are many elements in each ~~group~~ sub-group. ii) Selection of the sub-gp depending upon the criterion related to the variables under study.	* One divide the population into many sub groups. i) There are few elts in each sub group. ii) According to some criterion of ~~ease~~ each subgroups are selected or availability in data collection.
* Homogeneity is secured withing sub-groups	* Heterogeneity is secured within sub groups.
* Securing heterogeneity b/w sub-groups	* Securing homogeneity b/w sub-groups.

Advantages of Probability Sampling

1. Unbiased Estimate
2. Relative Efficiency
3. Less universe knowledge required
4. Every item in the population has an equal chance of being selected and analysed.
5. Easy data analysis and error calculation is allowed by this method of sampling.

Disadvantages of probability Sampling

1. Less efficient.
2. Non-utilisation of additional knowledge
3. Complex and time consuming
4. High Level Skills.
5. More time required.
6. High Costs.

Difference b/w Probability and Non-Probability Sampling

Probability Sampling	Non-Probability Sampling
* Sampling error can be controlled.	* Sampling error can not be controlled.
* The selection process is not influenced by the expertise of the researcher because it depends on the specific technique.	* There may be existence of higher level selection biasness.
* The involvement of time and costs may be high.	* The involvement of cost is very low and quicker alternative.
* Possibility of testing the hypothesis through formal, rigorous tests of significance in obtaining more reliable results.	* The reliability of results is not very high because parametric tests of significance are not applicable.
* If the population is heterogeneous then it is more reliable and representative.	* For homogeneous population it may be more useful.

* If the population is high then the accuracy may be poor.	* Accuracy in such situations may be scattered.
* In probability sampling, formal sampling frames are required.	* It is effective even in the absence of an elaborate sampling frame.

Central Limit Theorem (CLT) [Lindberg - Levy's form]

* If $X_1, X_2, \ldots, X_n$ be a sequence of independent itentically distributted RV's with: $E(X_i) = \mu$; var $(X_i) = \sigma^2$, $i = 1, 2, \ldots n$ and if $S_n = X_1 + X_2 + \cdots + X_n$, then under certain general conditions, S_n follow a normal distribution with mean $n\mu$ and variance $n\sigma^2$ as $n \to \infty$.

* If the average of RV's follows normal distribution, then $\bar{X}$ follow $N(\mu, \sigma/\sqrt{n})$

By CLT,

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

* If the discrete RV's follow normal distribution, then $\bar{X}$ follow $N(\mu, \sigma)$ by CLT.

$$Z = \frac{\bar{X} - \mu}{\sigma}.$$

Applications of CLT.

1. Various assumptions for the relationship b/w the sample statistics and the population parameters can be made using CLT.
2. The probability of getting different sample means can be calculated by this theorem.

3. As the sample means of sample size $n \geq 30$ are normally distributed. Hence the normal dist/: for sample data taken from any population can be calculated. This is the main use of CLT.

Uses of CLT

1. It states that almost all theoretical distributions converge to normal distribution as $n \to \infty$.

2. It helps to find out the distribution of the sum of a large no/: of independent R.V's.

1. In a sample of 25 observations from a normal distribution with mean 98.6 and S.D 17.2.

1. What is $P(92 < \bar{x} < 102)$?

2. Find the corresponding prob/: given a sample of 36.

Sol

i) Given: $\mu = 98.6$; $\sigma = 17.2$; $n = 25$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - 98.6}{17.2/\sqrt{25}} = \frac{\bar{x} - 98.6}{3.44}$$

When $\bar{x} = 92 \Rightarrow z = \frac{92 - 98.6}{3.44}$

$$= -1.92.$$

$\bar{x} = 102 \Rightarrow z = \frac{102 - 98.6}{3.44} = 0.99.$

$$\therefore P(92 < \bar{x} < 102) = P(-1.92 < z < 0.99)$$

$$= P(-1.92 < z < 0) + P(0 < z < 0.99).$$

$= P(0 < z < 1.92) + P(0 < z < 0.99)$

$= 0.4726 + 0.\cancel{14}3389 = 0.8115$

(ii) when $n = 36$.

$\therefore z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \dfrac{\bar{x} - 98.6}{17.2/\sqrt{36}} = \cancel{2.87} \dfrac{\bar{x} - 98.6}{2.87}$

when $\bar{x} = 92 \Rightarrow z = \dfrac{92 - 98.6}{2.87} = -2.30$.

$\bar{x} = 102 \Rightarrow z = \dfrac{102 - 98.6}{2.87} = 1.18$.

$P(92 < \bar{x} < 102) = P(-2.30 < z < 1.18)$

$= P(-2.30 < z < 0) + P(0 < z < 1.18)$

$= P(0 < z < 2.30) + P(0 < z < 1.18)$

$= 0.4861 + 0.3599$

$= 0.846$.

2. The mean value of a random sample of 60 items was found to be 145, with S.D of 40. Find the 95% confidence limits for the population mean. What size of the sample required to estimate the population mean within 5 of its actual value with 95% or more confidence using the sample mean.

Sol: Given $n = 60$; ~~$\mu = 140$~~ $\bar{x} = 145$.

$\sigma = 40$.

$z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \cancel{\dfrac{\bar{x} - 145}{40/\sqrt{60}}}$

95% confidence limit is

$\mu = \bar{x} \pm z \cdot \sigma/\sqrt{n}$.

$$\mu = 145 \pm 1.96 \times \frac{40}{\sqrt{60}}$$

$$\mu = \left(145 - 1.96 \times \frac{40}{\sqrt{60}},\ 145 + 1.96 \times \frac{40}{\sqrt{60}}\right)$$

$$\mu = (134.9,\ 155.1)$$

$$\Rightarrow \boxed{134.9 \le \mu \le 155.1}$$

To find n: here $E = \pm 5$ (given),

Given: $P(\bar{x} - \mu \le 5) \geq 0.95$.

$$\boxed{\begin{aligned} n &= \left(\frac{z_\alpha \times \sigma}{E}\right)^2 \\ &= \left(\frac{1.96 \times 40}{5}\right)^2 \\ &= 245.86 \\ &\approx 246. \end{aligned}}$$

$$\cancel{P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \le \frac{5}{\sigma/\sqrt{n}}\right) \geq 0.95}$$

$$\cancel{P\left(z \le \frac{5\sqrt{n}}{\sigma}\right) \geq 0.95}$$

$$\cancel{\mu = \bar{x} \pm z_\alpha \cdot \sigma/\sqrt{n}}$$

$$\cancel{|\mu - \bar{x}| = z_\alpha \cdot \sigma/\sqrt{n}}$$

$$\cancel{5 = 1.96 \times \frac{40}{\sqrt{n}}}$$ (95% confidence limit z_α is 1.96)

$$\cancel{5\sqrt{n} = \frac{1.96 \times 40}{5}}$$

$$\cancel{n = \left(\frac{1.96 \times 40}{5}\right)^2 = 245.86}$$

$$\therefore n \approx 246.$$

Sampling distribution of the mean (μ) and difference of mean ($\mu - \bar{x}$) [large sample]

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

here $\mu_{\bar{x}} = \mu$. [(or) Sample mean and population mean are same).

$$S.E = \sigma/\sqrt{n} = \sigma_{\bar{x}}.$$

Note

In sampling from a finite population, the sampling distribution of a sample mean will have mean $\mu_{\bar{x}} = \mu$ and

$$S.E = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}.$$

* $$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

* $$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2.$$

$$\boxed{S.E = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.}$$

1. ~~In a certain group of people, 150 samples are taken in the study of 3 variables height, weight and age: The result are as follows.~~

~~Height : Mean - 165 cm, S.D - 20 cm~~

Confidence interval for the population mean for large sample.

$$\mu = \left(\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right)$$

* 95% confidence interval ⇒ $Z_{\alpha/2} = 1.96$
* 99% confidence interval ⇒ $Z = 2.58$
* 98% confidence interval ⇒ $Z = 2.33$.
* 90% confidence interval ⇒ $Z = 1.645$.

Sampling distribution of Proportions:

$$\text{mean} = \mu_P = P$$

$$S \cdot E = \sigma_P = \sqrt{\frac{PQ}{n}}$$

where $P+Q=1$

Note: For large values of n $(n \geq 30)$, the sampling distribution of proportion is very closely normally distributed.

Sampling distribution of the difference of two proportions

$$\text{mean} = \mu_{P_1-P_2} = P_1 - P_2$$

$$S.E = \sigma_{P_1-P_2} = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$$

where $P_1 + Q_1 = 1$

$P_2 + Q_2 = 1$

Confidence limits of population

* Single proportions

$$P = p \pm z \cdot \sqrt{\frac{PQ}{n}}$$

* Difference proportion

$$P_1 - P_2 = (p_1 - p_2) \pm z \cdot \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$$

~~Sampling~~ Confidence limit of population mean (small sample)

$$\mu = \bar{x} \pm t_{\alpha/2}\, s/\sqrt{n}$$

where $S^2 = \frac{1}{n-1} \Sigma (x-\bar{x})^2$ with d.f $= n-1$

Confidence interval for the difference b/w two population means for small samples

$$\mu_1 - \mu_2 = \left((\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \cdot S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

where $S^2 = \frac{1}{n_1+n_2-2} \left[\Sigma (x_1 - \bar{x}_1)^2 + \Sigma (x_2 - \bar{x}_2)^2 \right]$

α = Level of significance . at $d.f = n_1+n_2-2$.

Determining the sample size.

* sample size for estimating population mean :

$$n \neq \left(\frac{Z_\alpha \cdot n}{S} \right)$$

$$S.E = Z_\alpha \cdot \frac{\sigma}{\sqrt{n}} = E$$

$$\Rightarrow \sqrt{n} = \frac{Z_\alpha \cdot \sigma}{E}$$

$$\Rightarrow n = \left(\frac{Z_\alpha \cdot \sigma}{E} \right)^2$$

* Sample size for estimating a population proportion.

$$S.E = Z_\alpha \cdot \sqrt{\frac{PQ}{n}} = E$$

$$\Rightarrow n = \left(\frac{Z_\alpha \cdot \sqrt{PQ}}{E} \right)^2$$

Note : S.E will be given in our question itself.

1\. In a certain group of people, 150 samples are taken in the study of three variables, height, weight and age. The result are as follows:

Height : Mean $= 165$ cm, S.D $= 20$ cm

Weight : Mean $= 70$ kg, S.D $= 10$ kg.

Age : Mean $= 45$ years, S.D $= 5$ years.

Compute standard error in each case. What is the range within you will have 95% confidence level for the estimate?

Sol:

Given: $n = 150$ (large sample)

$$S.E = \frac{\sigma}{\sqrt{n}} = \frac{\sigma}{\sqrt{150}}$$

i) Height: $\bar{x} = 165$ $\sigma = 20$.

$$S.E = \frac{20}{\sqrt{150}} = 1.633$$

95% confidence limit:

$$\mu = \bar{x} \pm z \cdot \frac{\sigma}{\sqrt{n}}.$$

$$= 165 \pm 1.96\,(1.633)$$

$$= 165 \pm 3.201$$

$$= (161.799, 168.20)$$

ii) Weight: $\bar{x} = 70$; S.D $= \sigma = 10$.

$$S.E = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{150}} = 0.816.$$

95% confidence limit

$$\mu = \bar{x} \pm z \cdot \left(\frac{\sigma}{\sqrt{n}}\right) = 70 \pm 1.96\,(0.816)$$

$= 70 \pm 1.6$

$\mu = (68.4, 71.6)$

(iii) Age: $\bar{x} = 45$, $\sigma = 5$

$S.E = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{150}} = 0.408$.

95% confidence limit:

$\mu = \bar{x} \pm z \cdot \frac{\sigma}{\sqrt{n}}$

$= 45 \pm 1.96 (0.408)$

$= 45 \pm 0.8$

$= (44.2, 45.8)$.

2. A random sample of 100 observations yields sample mean $\bar{x} = 150$ and sample variance $s^2 = 400$. Compute 95% and 99% confidence interval for the population mean.

Sol: Given: $n = 100$ (large sample)

$\bar{x} = 150$

$s^2 = 400 \Rightarrow s = \sqrt{400} = 20$.

$S.E = \frac{s}{\sqrt{n}} = \frac{20}{\sqrt{100}} = 2$.

95% confidence Limit

$\mu = \bar{x} \pm z \cdot \frac{s}{\sqrt{n}}$.

$= 150 \pm 1.96 (2)$.

$= 150 \pm 3.92 =$, ~~(153.92, 14~~

$= (146.08, 153.92)$.

99% confidence Limit

$\mu = \bar{x} \pm z (s/\sqrt{n})$

$\mu = 150 \pm 2.58 (2) = 150 \pm 5.16.$

$= (144.84, 155.16).$

3. A random sample of 10 employees reveals the foll. family dental expenses (in thousan ₹) in the previous year: 11, 37, 25, 62, 51, 21, 18, 43, 32, 20. Set up 99% confidence interval of the average family dental expenses for the employees of this Organisation.

Sol:

$n = 10$ (small sample).

$\therefore df = 10 - 1 = 9.$

Tab. t value at 1% = ~~1.833~~ 3.250.

99% confidence limit

$$\mu = \bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}.$$

$$= \bar{x} \pm 3.250 \frac{s}{\sqrt{n}}$$

$$s^2 = \frac{1}{n-1} \Sigma (x - \bar{x})^2.$$

x	$x - \bar{x}$	$(x - \bar{x})^2$
11	-21	441
37	5	25
25	-7	49
62	30	900
51	19	361
21	-11	121
18	-14	196
43	11	121
32	0	0
20	-12	144.
320	0	2358.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{320}{10} = 32.$$

$$S^2 = \frac{1}{10-1}(2358) = \frac{1}{9}(2358) = 262$$

$$S = \sqrt{262} = 16.\,9$$

$$\therefore \mu = 32 \pm 3.25\left[(16.\,9)/\sqrt{10}\right]$$

$$= 32 \pm 52.62/\sqrt{10}$$

$$= 32 \pm 16.64$$

$$= (15.36,\ 48.64)$$

4. In a random sample of 100 men taken from village A, 60 were found to be consuming alcohol. In another sample of 200 men taken from village B, 100 were found to be consuming alcohol. Construct 95% confidence interval in respect of difference in the proportions of men who consume alcohol.

Sol: Given: $n_1 = 100$; $n_2 = 200$

$$P_1 = \frac{60}{100} = 0.6$$

$$Q_1 = 1 - P_1 = 0.4.$$

$$P_2 = \frac{100}{200} = 0.5.$$

$$Q_2 = 1 - P_2 = 0.5.$$

95% confidence Limit

$$p_1 - p_2 = (P_1 - P_2) \pm Z\left(\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}\right)$$

$$S.E = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$$

$$= \sqrt{\frac{0.6 \times 0.4}{100} + \frac{0.5 \times 0.5}{200}}$$

$= 0.06$

$$\therefore\ p_1 - p_2 = (0.6 - 0.5) \pm 1.96\,(0.06)$$

$$= 0.1 \pm 1.96 \times 0.06$$

$$= 0.1 \pm 0.1176$$

$$= (0.018, 0.218)$$

5. A survey of 748 randomly selected employees of dot com companies showed that 35% feel secure about their jobs. Give a 90% confidence interval for the proportion of dot.com compan who feel secure about their jobs.

Sol:

Given: $n = 748$

$$P = \frac{35}{100} = 0.35$$

$$Q = 1 - P = 0.65$$

$$S.E = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{0.35 \times 0.65}{748}}$$

$$= 0.0174$$

90% confidence interval

$$p = P \pm z\sqrt{\frac{PQ}{n}}$$

$$= 0.35 \pm 1.645\,(0.174)$$

$$p = (0.3214, 0.3786)$$

6. A firm wishes to estimate with a maximum allowable error of 0.05 and 95% level of confidence, the proportion of customers who prefer its product. How large a sample will be required in order to make such as estimate

if the preliminary sales report indicate that 25% a of all consumers prefer the firms product.

Sol.

Given $P = 25\% = 0.25$

$Q = 1 - P = 0.75$

$E = 0.05$

$Z_\alpha = 1.96$ (at 95% level of confidence).

To find sample size

$$n = \left(\frac{Z_\alpha \cdot \sqrt{PQ}}{E}\right)^2$$

$$= \left(\frac{1.96 \times \sqrt{0.25 \times 0.75}}{0.05}\right)^2$$

$$\boxed{n = 288}$$

Estimation:

1. Point estimation
2. Interval estimation.

Point Estimation

When a single value is used as an estimate, the estimate is called a point estimate of the population parameter.

Ex: Sample mean ($\bar{x}$) is the sample statistic used as an estimate of population mean μ.

Interval Estimation

An estimate of a population parameter given by two numbers b/w which the parameter may be considered to lie is called an interval estimate of the parameter.

Characteristic of a good estimator (point estimator)

i) Unbiasedness

An estimator is said to be unbiased if its expected value is equal to the population parameter it estimates.

Ex: $E[\bar{x}] = \mu$.

(ie) sample mean is a unbiased estimator of a population mean μ.

ii) Efficiency

An estimator is efficient if it has a relatively smaller variance.

iii) Consistency

An estimator is said to be consistent if probability of being close to the parameter it estimates increases as the sample size increases.

The sample mean $\bar{x}$ is said to be a consistent estimator of θ.

iv) Sufficiency

An estimator is said to be sufficient if it contains all the information in the data about the parameter it estimates.

1. S.T the sample mean $\bar{x}$ is an unbiased estimator of population mean μ.

Sl:

$$\bar{x} = \frac{\sum_i x_i}{n}$$

$$E[\bar{x}] = E\left[\frac{\sum x_i}{n}\right] = \frac{1}{n} E\left[\sum x_i\right]$$

$$= \frac{1}{n}\left\{E(x_1) + E(x_2) + \cdots + E(x_n)\right\}$$

$$= \frac{1}{n}\left\{\mu + \mu + \mu + \cdots + \mu\right\} \quad (n \text{ times})$$

$$= \frac{1}{\cancel{n}}\left[\cancel{n}\mu\right]$$

$E(\bar{x}) = \mu.$

$\Rightarrow$ $\bar{x}$ is unbiased estimator of μ.

2 Below you are given the values obtained from an infinite population 38, 34, 35, 39.

1. Find a point estimate for μ. Is this an unbiased estimate of μ? Explain.

2. Find a point estimate for σ^2 (variance).

3) Find a point estimate for σ.

4) What can be said about the sampling distribution of $\bar{x}$?

Sol:

1) . Point Estimation of $\mu = E(\bar{x})$

$$= \frac{38 + 34 + 35 + 39}{4} = \frac{140}{4}$$

$$= 35$$

2) Point Estimation of $\sigma^2 = \frac{1}{n-1}\Sigma(x - \bar{x})^2$

$$= \frac{1}{3}\left[(38-35)^2 + (34-35)^2 + (35-35)^2 + (39-35)^2\right]$$

$$= \frac{1}{3}(9 + 1 + 0 + 16) = 8.667$$

3) Point Estimation of $\sigma = \sqrt{\sigma^2}$

$$= \sqrt{8.667} = 2.944.$$

4)

* Sampling distribution of $\overline{X}$ is defined as the probability of all the possible means of the samples.

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

The shape of sampling distribution is normal curve.

* The S.D of the distribution of sampling statistics is the S.E of the statistics.

Difference B/w Point estimate and Interval Estimate of Populatiation Parameter

Point Estimate	Interval Estimate
* A point estimate of a population parameter is a single value of statistic.	* An interval estimate is defined by two nos/. b/w which a population parameter is said to lie.
* Ex/. sample mean $\overline{X}$.	Ex: $a < \bar{x} < b$.

Additional Problems

1. ~~In a shipment of manufacture~~

1. The age of employees in a company follows normal distribution with its mean and variance as 40 years and 121 years respectively. If a random sample of 36 employees is taken from a finite normal population 1000, what is the prob/. that the sample mean is

i) less than 40 (ii) greater than 42

(iii) b/w 40 and 42.

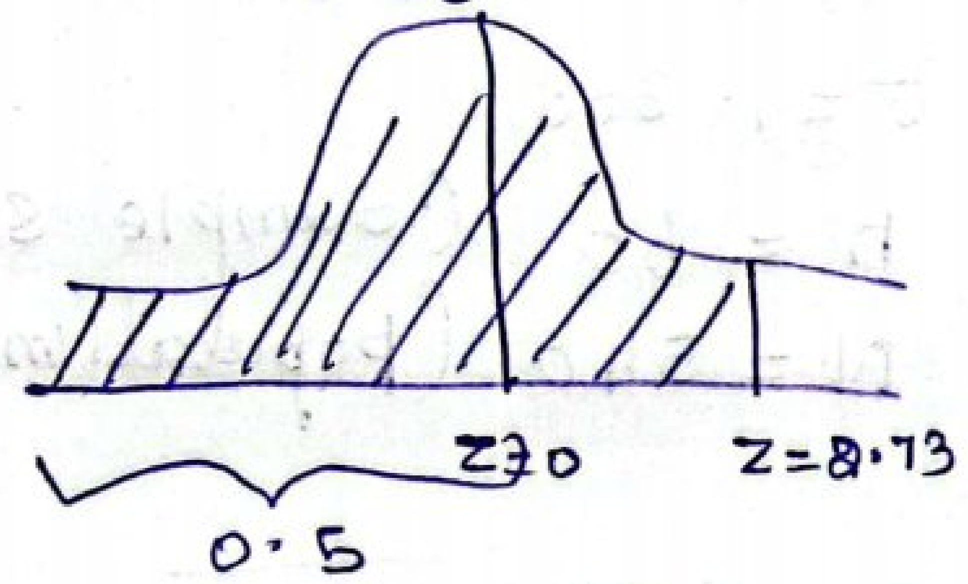
8.

Given: $n=36$; $\mu=40$; $\sigma^2=121 \Rightarrow \sigma=11$

$$Z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}=\frac{\bar{x}-40}{11/\sqrt{36}}=\frac{\bar{x}-40}{1.83}$$

i) $P(\bar{x}<45)=P(z<2.73)$

$$Z=\frac{45-40}{1.83}=2.73$$



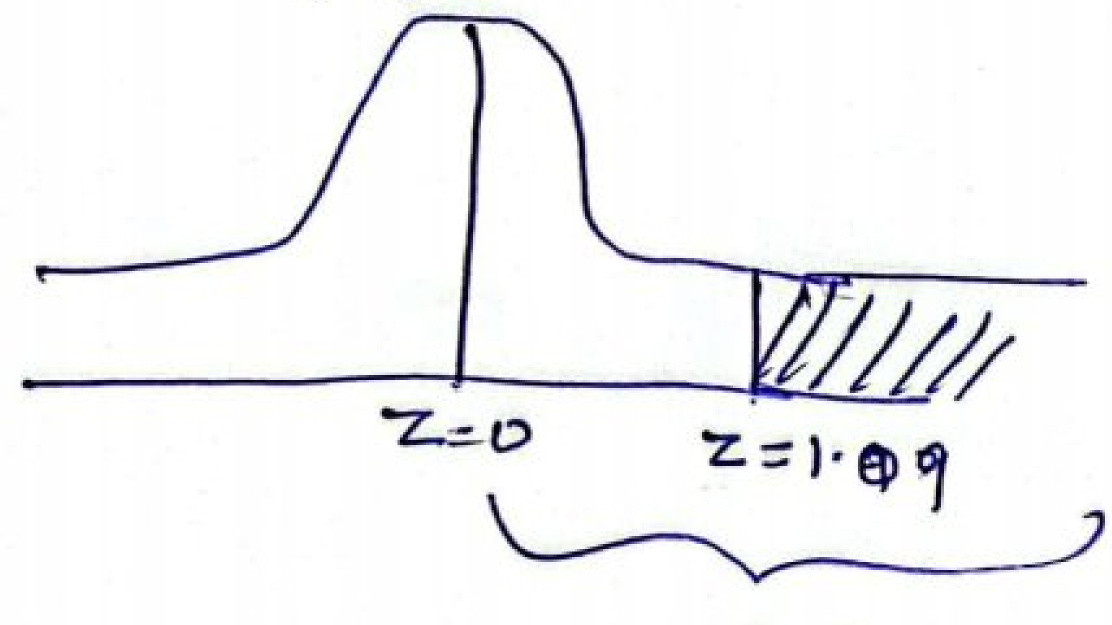
$= 0.5+P(0<z<2.73)$

$= 0.5+0.4968$

$= 0.9968$

ii) $P(\bar{x}>42)=P(z>1.09)$

$$Z=\frac{42-40}{1.83}=1.09$$



$=0.5-P(0<z<1.09)$

$=0.5-0.3621=0.1379.$

iii) $P(40<\bar{x}<42)=P(0<z<1.09)$

$=0.3621.$

CS CamScanner

8) From a population of 540, a sample of 60 individual is taken. From this sample, the mean is found to be 6.2 and S.D 1.368. Find estimated S.E of the mean.

Sol.

$$S.E = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}.$$

here. $\sigma = 1.368$.

$n = 60$ (sample size).

$N = 540$ (population size given)

$$\therefore S.E = \frac{1.368}{\sqrt{60}} \sqrt{\frac{540-60}{540-1}} = 0.167.$$

— x —

Chapter : 3 Testing of Hypothesis - Parametric Tests

On the basis of sample information, we make certain decisions about the population. In taking such decisions, we make certain assumptions. These assumptions are known as Statistical hypotheses. These hypotheses are tested.

Null hypothesis

Null hypothesis is based on analysing the problem, Null hypothesis is the hypothesis of no difference.

It is denoted by H_0.

Alternative hypothesis:

Any hypothesis which is complementary to the null hypothesis (H_0) is called an alternative hypothesis, denoted by H_1.

Rule:

* If we want to test the significance of the difference b/w statistic and the parameter, then $H_0 : \mu_0 = \bar{x}$.

* If we want to test any statement about the population, then $H_0 : \mu = \mu_0$.

* If $H_0 : \mu = \mu_0$, then alternative hypothesis will be

$H_1 : \mu \neq \mu_0$ (two tailed test)

$H_1 : \mu < \mu_0$ (Left tailed test)

$H_1 : \mu > \mu_0$ (Right tailed test).

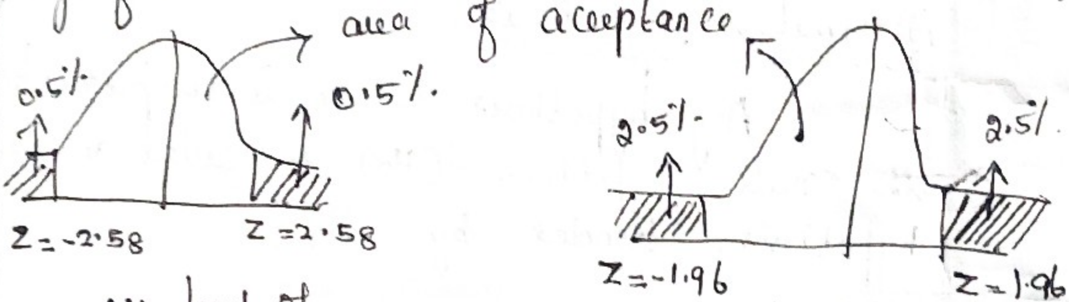
Critical Region

A region, corresponding to a statistic t, in the sample space S which amounts to rejection of the null hypothesis H_0 is called as critical region or region of rejection.

The region of the sample space S which amounts to the acceptance of H_0 is called acceptance region.

Level of Significance.

The prob. that the value of the statistic lies in the critical region is called the level of significance.



1% level of significance

5% level of significance.

Errors

Type I Error

If H_0 is rejected, while it should have been accepted. Rejection of H_0 when it is actually true

Type II Error

If H_0 is accepted while it should have been rejected. Acception of H_0 when it is actually false.

Degree of freedom.

The no. of elts in the calculation which are able to vary.

Large Sample test ($n \geq 30$)

1. Test for a specified mean
2. Test for the equality of two means
3. Test for specified proportion
4. Test for equality of two proportions

Table for critical values on using normal probability

Critical values	Level of significance (α) 1%	5%	10%
Two tailed test	$\|Z\| = 2.58$	1.96	1.645
Left-tailed test	$Z = 2.33$	1.645	1.28
Right-tailed test	$Z = -2.33$	-1.645	-1.28

Confidence limits

* 95% confidence limits will be lies in the interval $(\mu - 1.96\sigma, \mu + 1.96\sigma)$
* 99% confidence limit will be lie in the interval $(\mu - 2.58\sigma, \mu + 2.58\sigma)$

$\therefore$ The numbers 1.96 and 2.58 are called confidence co-eff:

~~Test for a specified mean and equality of two means~~

Types of Hypothesis Test

* Test of significance for small sample
* Test of significance for Large sample
* Parametric test
* Non-Parametric test

Difference b/w Large and Small sample Tests

	Large Sample	Small Sample.
Sample size	>30	<30.
Assumption for test of Significance.	The random sampling distribution of a statistic is approximately normal	Assumption of normality is made unless state
Comparison from population	Different values of samples are close to population values	Sample values and population values differ significantly.
S.E	This concept is used while testing for significance	Generally, this concept is not used
Types of hypothesis	Z-test χ^2-test	t-test (mean) F-test (variance) Z-test

Difference b/w Parametric and Non-Parametric Test

Parametric Test	Non-Parametric Test
* Information about population is completely known.	* No information about the population is available
* Specific assumptions are made regarding the population.	* No assumptions are made regarding the population.
* Null hypothesis is made on parameters of the population distribution.	* Ho is free from parameters.

* Test statistic is based on the distribution	* Test statistic is arbitrary
* Parametric tests are applicable only for variables.	* Applicable for both variables and attributes.
* This test is powerful if it exists	It is not as powerful as parametric test

Formulation of Hypothesis

* Set up a hypothesis
* Set up suitable significance level
* Test statistics.
* Doing Computation
* Making Decision.

Z-test (large sample $n>30$).

* For single mean

$$z = \frac{\bar{x} - \mu}{S.E}, \quad S.E = \sigma/\sqrt{n} = \sqrt{\sigma^2/n}$$

$\mu \rightarrow$ population mean

$\bar{x} \rightarrow$ Sample mean.

$\sigma \rightarrow$ population standard deviation

$s \rightarrow$ sample standard deviation.

$n \rightarrow$ sample size.

* For two difference mean

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

The mean lifetime of a sample of 100 tube lights produced by a company is found to be 1600 hrs with S.D of 95 hrs. Test the hypothesis that the mean lifetime of the bulbs produced by company is 1620 hrs.

Sol:

Given $\bar{x} = 1600$, $n = 100$

$s = 95$; $\mu = 1620$

Null hypothesis: H_0

$$\mu = 1620 = \mu_0$$

Alternative hypothesis: (H_1)

$$\mu \neq 1620 \text{ (two tailed test)}.$$

here $\alpha = 5\%$, d.r.f

tab $z = 1.96$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{1600 - 1620}{95/\sqrt{100}}$$

$$= \frac{-20}{9.5} = -2.1053$$

$$|z| = 2.1053.$$

Conclusion

Cal $|z| >$ tab z.

$\therefore$ H_0 is rejected

H_1 is accepted.

Scanned with ACE Scanner

using newtons forward.

find $x = 5$, satisfying the following data.

2) Intelligence test given to two groups of boys and girls gave the foll% information.

	Mean score	S.D	Number
Girls	75	10	50
Boys	70	12	100

Is the difference in the mean scores of boys and girls statistically significant?

Sol.

Given $n_1 = 50$ $\quad \bar{x}_1 = 75$ $\quad s_1 = 10$

$n_2 = 100$ $\quad \bar{x}_2 = 70$ $\quad s_2 = 12$

Null hypothesis: H_0

$$\mu_1 = \mu_2$$

Alternative hypothesis: H_1

$$\mu_1 \neq \mu_2$$

$\alpha = 5\%$

Tab $z = 1.96$.

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \cancel{1500} - \cancel{18} \frac{75-70}{\sqrt{\frac{100}{50} + \frac{144}{100}}}$$

$$= \frac{5}{\sqrt{3.44}} = 2.6958$$

Conclusion

Cal $z = 2.6958 > 1.96 =$ Tab z.

H_0 is rejected.

Scanned with ACE Scanner

3) Two cities are studied for weight. Results are as follows.

City A: sample 200, mean 75kg, S.D 10kg.

City B: Sample 250, Mean 85kg, S.D 5kg

1) Test whether city A population could have an average weight of 80.

2) Also test whether city A is heavier than city B.

Sol:

Given: $n_1 = 200$ $\quad \overline{x}_1 = 75$ $\quad s_1 = 10$

$n_2 = 250$ $\quad \overline{x}_2 = 85$ $\quad s_2 = 5$

1) Null hypothesis (H_0)

$$\mu_1 = 80$$

Alternative hypothesis (H_1)

$$\mu_1 \neq 80 \text{ (two tailed, at } 5\% \text{)}$$

$\therefore$ tab $z = 1.96$

$$z = \frac{\bar{x}_1 - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x}_1 - \mu_1}{s_1/\sqrt{n_1}} = \frac{75-80}{10/\sqrt{200}}$$

$$= -5 \times \frac{\sqrt{200}}{10}$$

$$= -7.07106$$

$$|z| = 7.07106$$

Conclusion

cal z > tab z.

H_0 is rejected.

2) Null hypothesis (H_0) There is no significant difference b/w means.

$$\mu_1 = \mu_2$$

Alternative hypothesis (H_1)

$$\mu_1 > \mu_2 \text{ (Right tailed test, at } 5\% \text{)}$$

tab/. $z = -1.645$.

Now, $$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{75-85}{\sqrt{\frac{100}{200} + \frac{25}{250}}}$$

$$z = -12.9099$$

Conclusion:

here cal z < tab z.

$\Rightarrow$ H_0 is accepted.

Scanned with ACE Scanner

H.W

Two independent samples of observations were collected. For the first sample of 60 elts, the mean was 86 and S.D 6. The second sample of 75 elts has a mean of 82 and a S.D of 9. Using $\alpha = 0.01$, test whether the two samples can reasonably be considered to have come from population with the same mean.

Sol: Given: $n_1 = 60$ $\bar{x}_1 = 86$ $s_1 = 6$

$n_2 = 75$ $\bar{x}_2 = 82$ $s_2 = 9$.

Null hypothesis Ho

$$\mu_1 = \mu_2$$

Alternative hypothesis $\mu_1 \neq \mu_2$ (two tailed, 1%)

$$tab\ |z| = 2.58$$

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \frac{86-82}{\sqrt{\frac{36}{60} + \frac{81}{75}}} = \frac{4}{1.2961}$$

$$= 3.0861$$

$$|z| = 3.0861.$$

Conclusion

$\cancel{\text{tab}}$ cal $|z| >$ tab $|z|$

Ho is rejected.

Testing of hypothesis about proportion

* For single proportion

$$z = \frac{p - P}{\sqrt{PQ/n}}$$

* For two different proportion

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{PQ}{n_1} + \frac{PQ}{n_2}}}$$

where $P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

$Q = 1 - P$.

1. A social experiment shows that in a group 20% people are ready to sell their votes for money when they are offered a small amount. In another group, 40% people are ready to sell their vote when they are offered huge sum money. In both the cases, 1000 members each were participated. Test at 5% level of significance (two sided) that there is a difference two proportions.

Sol:

Given: $p_1 = 20\% = \frac{20}{100} = 0.2$, $n_1 = 1000$

$p_2 = 40\% = \frac{40}{100} = 0.4$, $n_2 = 1000$

Null hypothesis (H_0)

There is no significant difference b/w proportions.

$\therefore P_1 = P_2$

Alternative hypothesis (H_1)

$P_1 \neq P_2$ (two tailed, 5%).

tab $|z| = 1.96$

$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

$= 0.3$

$Q = 1 - P = 0.7$

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{PQ}{n_1} + \frac{PQ}{n_2}}}$$

$PQ = 0.21$

$$Z = \frac{0.2 - 0.4}{\sqrt{\frac{0.21}{1000} + \frac{0.21}{1000}}} = -9.7590$$

$|z| = 9.7590$.

Conclusion

Cal z > tab z.

H_0 is rejected.

2. A coin is tossed 256 times and 132 heads are obtained. Would you conclude that the coin is biased one?

Sol:

$n = 256$.

No. of heads appeared $= 132$.

$\therefore$ prop. of event $= \frac{132}{256} = 0.516 = p$.

$P = \frac{1}{2}$.

$Q = \frac{1}{2}$.

Null hypothesis: H_0.

Coin is unbiased. $(P = \frac{1}{2})$.

Alternative hypothesis: H_1

Coin is biased $(P \neq \frac{1}{2})$ (two tailed)

$|z| = 1.96$

$$z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} = \frac{0.516 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{256}}} = 0.512$$

$|z| = 0.512$

Conclusion:

$$\text{Cal } |z| < \text{tab } |z|$$

H_0 is accepted.

$\therefore$ Coin is unbiased.

H.W

In a random sample of 1000 people from city A, 400 are found to be consumers of wheat. In a sample of 800 from city B, 400 are found to be consumers of wheat. Does this data give a significant difference b/w the two cities as far as the proportion of wheat consumers is concerned?

t-Test for testing sample mean (n<30)

* For single sample

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{\bar{x} - \mu}{\sqrt{S^2/n}}$$

where $S^2 = \frac{1}{n-1} \Sigma (x_i - \bar{x})^2$ (if S.D not given)

$S^2 = \frac{n s^2}{n-1}$ (if sample S.D given).

degrees of freedom = n-1

here $\bar{x}$ = sample mean = $\frac{\Sigma x}{n}$.

s = sample. S.D.

S^2 = unbiased estimate of population Variance.

* For different mean

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $S^2 = \frac{1}{n_1+n_2-2} \left[\Sigma (x_1 - \bar{x}_1)^2 + \Sigma (x_2 - \bar{x}_2)^2 \right]$

(if S.D not given)

$S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1+n_2-2}$ (if S.D given)

degrees of freedom = $n_1 + n_2 - 2$

* Paired t-test for difference Mean

$$t = \frac{\bar{d}}{S/\sqrt{n}} = \frac{\bar{d}}{S/\sqrt{n}}$$

where $\bar{d} = \frac{\sum_i d_i}{n}$ = Mean of difference.

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

degrees of freedom $= n-1$

n = number of paired observations in the sample.

Problems

1. Two independent samples of size 8 and 7 contained the following values:

Sample I	19	17	15	21	16	18	16	14
Sample II	15	14	15	19	15	18	16	

Q: Is the difference b/w the sample mean significant.?

Sol: Given: $n_1 = 8$ $n_2 = 7$.

Null hypothesis (H_0)

There is no significant diff/: b/w means. $\mu_1 = \mu_2$

Alternative hypothesis

$\mu_1 \neq \mu_2$ (two tailed, 5% level).

degrees of freedom $= n_1 + n_2 - 2$
$= 15 - 2 = 13$.

$\therefore \boxed{\text{tab/: } t = 2.16.}$

Now, $t = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $\left[\because \mu_1 = \mu_2 \Rightarrow \mu_1 - \mu_2 = 0\right]$

Scanned with ACE Scanner

$\overline{x_1} = \Sigma x_1$

x_1	$x_1 - \overline{x_1}$ $x_1 - 17$	$(x_1-\overline{x_1})^2$	x_2	$x_2 - \overline{x_2}$ $x_2 - 16$	$(x_2-\overline{x_2})^2$
19	2	4	15	-1	1
17	0	0	14	-2	4
15	-2	4	15	-1	1
21	4	16	19	3	9
16	-1	1	15	-1	1
18	1	1	18	2	4
16	-1	1	16	0	0
14	-3	9			
136.		36.	112		20

$$\overline{x_1} = \frac{\Sigma x_1}{n_1} = \frac{136}{8} = 17$$

$$\overline{x_2} = \frac{\Sigma x_2}{n_2} = \frac{112}{7} = 16$$

$$S^2 = \frac{1}{n_1+n_2-2}\left[\Sigma(x_1-\overline{x_1})^2 + \Sigma(x_2-\overline{x_2})^2\right]$$

$$= \frac{1}{13}\left[36+20\right] = 4.3077$$

$$\boxed{S = 2.0755}$$

$$t = \frac{17-16}{2.0755\sqrt{\frac{1}{8}+\frac{1}{7}}} = 0.9309.$$

Conclusion:

Cal. $t = 0.9309 <$ tab. $t = 2.16$

$\therefore$ Ho is accepted.

Scanned with ACE Scanner

2) Given a sample mean of 83, a sample s.d of 12.5 and sample size of 22, test the hypothesis that the value of population mean is 70 against the alternative that it is more than 70. Use the 0.025 significance level.

Sol:

Given $\bar{x} = 83$

$s = 12.5$

$n = 22 \quad (< 30)$

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} \quad \text{where} \quad S^2 = \frac{ns^2}{n-1}.$$

Null hypothesis (H_0)

$\mu = 70$.

Alternative hypothesis (H_1)

$\mu > 70$ (Right tailed - One tailed).

$\alpha = 0.025$

$= 2.5\%$ level (one tailed)

$= 5\%$ level (two tailed).

$d.f = n - 1 = 22 - 1 = 21.$

$\boxed{\text{tab}/.\ t = 2.080}$

Now, $t = \dfrac{\bar{x} - \mu}{S/\sqrt{n}} = \dfrac{83 - 70}{S/\sqrt{22}}$

$S^2 = \dfrac{ns^2}{n-1} = \dfrac{22\,(12.5)^2}{21} = 163.6905$

$\boxed{S = 12.7942}$

$\therefore t = \dfrac{83 - 70}{12.7942/\sqrt{22}} = 4.7659$

Scanned with ACE Scanner

Conclusion:

Cal $t = 4.7659 > \text{tab}\% \ t = 2.080$.

$\Rightarrow$ Ho is rejected.

3. An agency conducting weight reduction program claims that participants in their program achieve weight reduction of at least 6 kg after two weeks of the program. In evidence, they have given the foll% data of 10 participants who have undergone this program. On the basis of this sample evidence, can the claim of the agency on weight reduction be said to valid?

Before (kg)	85	91	99	92	87	79	87	92	95	106
After (kg)	76	83	81	86	79	73	79	85	95	96

Sol: Given: $n = 10$.

Null hypothesis (H_0)

There is no changes in weight of the participants

$\mu_1 = \mu_2$.

Alternative hypothesis (H_1)

$\mu_1 \neq \mu_2$ (two tailed, 5%)

Before	After	d	$d-\bar{d}$	$(d-\bar{d})^2$
85	76	9	1	1
91	83	8	0	0
99	81	18	10	100
92	86	6	-2	4
87	79	8	0	0
79	73	6	-2	4
87	79	8	0	0
92	85	7	-1	1
95	95	0	-8	64
106	96	10	2	4
		80		178

$$\bar{d} = \frac{\Sigma d}{n} = \frac{80}{10} = 8$$

$$S^2 = \frac{1}{n-1} \Sigma (d_i - \bar{d})^2 = \frac{1}{9}(178)$$

$$= 19.7778$$

$$S = \sqrt{19.7778} = 4.4472.$$

$$t = \frac{\bar{d}}{S/\sqrt{n}} = \frac{8}{4.4472/\sqrt{10}} = 5.6886.$$

conclusion:

cal degrees of freedom $= n-1 = 9$.

tab/: t at 5% level $= 2.262$.

hence cal/: t > tab/: t

Ho is rejected.

(ie) there is changes in weight.

4 The heights A R.S of 10 boys had the foll/: I.Q's 70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Do these data support the assumption of a population mean I.Q of 100? Find a reasonable range in which most of the mean I.Q values of samples of 10 boys lie.

Sol:

Given: $n = 10$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{70+120+110+101+88+83+95+98+107+100}{10}$$

$$= \frac{972}{10} = 97.2.$$

Null Hypothesis (Ho) $\mu = 100$

Alternative Hypotheses (H1) $\mu \neq 100$ (two tailed, 5%)

$d.f = n-1 = 10-1 = 9$

$tab\% \; t = 2.262$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{97.2 - 100}{s/\sqrt{10}}$$

$$s^2 = \frac{1}{n-1} \Sigma (x - \bar{x})^2$$

$$= \frac{1}{9} (1833.60) = 203.73$$

$$S = \sqrt{203.73} = 14.2734$$

$$\therefore t = \frac{97.2 - 100}{14.2734/\sqrt{10}} = -0.6203$$

$$|t| = 0.6203$$

Conclusion:

$cal\; t\% < tab\% \; t$

Ho is accepted.

Range:

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

$$t(s/\sqrt{n}) = \bar{x} - \mu.$$

$$\mu = \bar{x} \pm t\,(s/\sqrt{n})$$

$$= 97.2 \pm (2.262)(4.514)$$

$$= 97.2 \pm 10.21067$$

$$= (86.9893, 107.4107).$$

Scanned with ACE Scanner

F-test (Variance)

$$F = \frac{S_1^2}{S_2^2} \text{ (or) } \frac{S_2^2}{S_1^2}$$

where $S_1^2 = \frac{1}{n_1 - 1} \Sigma (x_1 - \bar{x}_1)^2$

$$S_2^2 = \frac{1}{n_2 - 1} \Sigma (x_2 - \bar{x}_2)^2$$

degrees of freedom $= (v_1, v_2) / (v_2, v_1)$

$= (n_1 - 1, n_2 - 1) / (n_2 - 1, n_1 - 1)$

1. Two samples of 6 and 7 items have the foll/: values for a variable.

Sample 1	39	41	42	42	44	40	
Sample 2	40	42	39	45	38	39	40

Do the sample variances differ significantly?

Sol: Given: $n_1 = 6$ $\quad n_2 = 7$.

$\bar{x}_1 = 41.3333$ $\quad \bar{x}_2 = 40.~~4333~~$ 4286

x_1	$x_1 - \bar{x}_1$	$(x_1 - \bar{x}_1)^2$	x_2	$x_2 - \bar{x}_2$	$(x_2 - \bar{x}_2)^2$
			40	−0.4286 −0.4333	0.1837 0.1877
39	−2.3333	5.4443	42	1.5714 1.5667	2.4693 2.4545
41	−0.3333	0.1111	39	−1.4286 −1.4333	2.0409 2.0543
42	0.6667	0.4445	45	4.5714 4.5667	20.8977 20.8547
42	0.6667	0.4445	38	−2.4286 −2.4333	5.8981 5.9209
44	2.6667	7.1113	39	−1.4286 −1.4333	2.0409 2.0543
40	−1.3333	1.7777	40	−0.4286 −0.4333	0.1837 0.1877
248		15.334	283		33.7141 33.7143

$$S_1^2 = \frac{1}{n_1 - 1} \Sigma (x_1 - \bar{x}_1)^2 = \frac{1}{5}(15.3334)$$

$$= 3.0667$$

Scanned with ACE Scanner

$$S_2^2 = \frac{1}{n_2 - 1} \Sigma (x_2 - \bar{x}_2)^2$$

$$= \frac{1}{6} (33.4141) = 5.6190$$

$$= 5.6191$$

Null hypothesis: H_0

There is no significant difference b/w variance.

Alternative Hypothesis: H_1

There is significant diff/. b/w variance

$$F = \frac{S_2^2}{S_1^2} \quad \left[\because S_2^2 > S_1^2 \text{ \& } F > 1 \right]$$

$$= \frac{5.6190}{3.0668} = 1.8322$$

$$d.f = (V_2, V_1) = (n_2 - 1, n_1 - 1)$$

$$= (6, 5).$$

tab/. F = 4.95.

Conclusion:

Cal/. F < tab/. F

H_0 is accepted.

2. Two random samples given the foll data.

Sample	size	mean	Variance
1	8	9.6	1.2
2	11	16.5	2.5

Can we conclude that the two samples have been drawn from the same normal population

Sl:

$n_1 = 8 \quad \bar{x}_1 = 9.6 \quad s_1^2 = 1.2$

$n_2 = 11 \quad \bar{x}_2 = 16.5 \quad s_2^2 = 2.5$

$$S^2 = \frac{1}{n_2 - 1} \Sigma (x_2 - x_2$$

F-test

Null hypothesis (H_0)

There is no significant diff/: b/w population variances.

$$\sigma_1^2 = \sigma_2^2$$

Alternative hypothesis (H_1)

There is significant diff/: b/w variances.

$$\sigma_1^2 \neq \sigma_2^2 \quad (\text{two tailed, } 5\%)$$

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{8 \times 1.2}{7} = 1.3714$$

$$S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{11}{10} (2.5) = 2.75$$

$$F = \frac{S_2^2}{S_1^2} = 2.0053$$

$$d.f = (n_2 - 1, n_1 - 1) = (10, 7) =$$

$$\text{tab } F = 3.63$$

Conclusion

Cal F < tab F

H_0 is accepted.

t-test

Null hypothesis: (H_0)

There is no significant diff/: b/w means.

$$\mu_1 = \mu_2$$

Alternative hypothesis (H_1)

There is significant diff/: b/w means.

$$\mu_1 \neq \mu_2 \quad (\text{two tailed, } 5\%)$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $S^2 = \dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = 2.1824$

$$S = 1.4773$$

$$t = \frac{9.6 - 16.5}{1.4773\left(\sqrt{\frac{1}{8}+\frac{1}{11}}\right)} = -10.0513$$

$$|t| = 10.0513$$

6

$$d.f = n_1 + n_2 - 2 = 17.$$

$$tab\ t = 2.110\ .$$

Conclusion

Cal t > tab t

H_0 is rejected.

Final conclusion:

Two samples have not come from same population.

H.W

Two random samples given the foll result

Sample	Size	mean	Sum of squares of deviations from the mean
1	10	15	90
2	12	14	108.

Test whether the samples come from the same normal population at 5% level of significance.

8

One way classification

Complete Randomized design (CRD)

Procedure:

Null hypothesis (H_0)

There is no significant diff/. b/w columns.

Alternative hypothesis (H_1)

There is significant diff/. b/w columns.

step: 1

Find N = No/. of observations

step: 2

Find T = Total of all observation values

step: 3

Find $CF = \frac{T^2}{N}$

step: 4

$$TSS = \Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2 + \cdots - \frac{T^2}{N}$$

step: 5

$$SSC = \frac{(\Sigma x_1)^2}{n_1} + \frac{(\Sigma x_2)^2}{n_2} + \frac{(\Sigma x_3)^2}{n_3} + \cdots - \frac{T^2}{N}$$

where $n_1, n_2, n_3, \ldots$ = no/. of observations in each columns

step: 6

$$SSE = TSS - SSC.$$

step: 7

Prepare ANOVA Table.

Source of Variations	d.f	Sum of Squares	Mean Square	Cal F	Tab/. F
B/w columns	$C-1$	SSC	$MSC = \frac{SSC}{d.f}$	$F_c = \frac{MSC}{MSE}$ (or) $= \frac{MSE}{MSC}$	
Error	$N-C$	SSE	$MSE = \frac{SSE}{d.f}$		

Scanned with ACE Scanner

Conclusion

* If Cal F_c < tab. F_c

$\Rightarrow$ H_0 is accepted

* If Cal F_c > tab. F_c

$\Rightarrow$ H_0 is rejected

Two Way Classification

Randomized Block Design

Procedure

Null hypothesis (H_0)

i) There is no significant diff% b/w columns

ii) There is no significant diff% b/w rows

Alternative hypothesis (H_1)

i) There is significant diff% b/w columns

ii) There is significant diff% b/w rows

Step: 1

Find N

Step: 2

Find T

Step: 3

Find $CF = \frac{T^2}{N}$

Step: 4

$TSS = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \dots - \frac{T^2}{N}$

Step: 5

$SSC = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} + \dots - \frac{T^2}{N}$

Step: 6

$SSR = \frac{(\Sigma Y_1)^2}{m_1} + \frac{(\Sigma Y_2)^2}{m_2} + \frac{(\Sigma Y_3)^2}{m_3} + \dots - \frac{T^2}{N}$

Step: 7

$SSE = TSS - SSC - SSR$

Scanned with ACE Scanner

Step:8 Prepare ANOVA Table

Source of variations	d.f	Sum of Squares	Mean square	Cal/F	tab/F
B/w columns	$C-1$	SSC	$MSC = \frac{SSC}{d.f}$	$F_C = \frac{MSC}{MSE}$ (or) $\frac{MSE}{MSC}$	
B/w rows	$r-1$	SSR	$MSR = \frac{SSR}{d.f}$	$F_R = \frac{MSR}{MSE}$ (or) $\frac{MSE}{MSR}$	
Error	$(c-1)(r-1)$	SSE	$MSE = \frac{SSE}{d.f}$		

Conclusion:

* If cal F_C < tab F_C

$\Rightarrow$ $H_0(i)$ is accepted

If cal F_C > tab F_C

$\Rightarrow$ $H_0(i)$ is rejected

* If cal F_R < tab F_R

$\Rightarrow$ $H_0(ii)$ is accepted

If cal F_R > tab F_R

$\Rightarrow$ $H_0(ii)$ is rejected.

Pbm 1. Four doctors each test four treatments for a certain disease and observe the no/. of days each patient takes to recover. The results are as follows.

Doctor	Treatment 1	2	3	4
A	10	14	19	20
B	11	15	17	21
C	9	12	16	19
D	8	13	17	20

Analyse significant difference of Doctor and treatment

Q:

Null hypothesis (H_0)

i) There is no significant diff/. b/w columns (Treatment)

ii) There is no significant diff/. b/w rows (Doctor)

Alternative Hypothesis (H_1)

i) There is significant diff/. b/w columns

ii) There is significant diff/. b/w rows

	x_1	x_2	x_3	x_4	Total	x_1^2	x_2^2	x_3^2	x_4^2
y_1	10	14	19	20	63	100	196	361	400
y_2	11	15	17	21	64	121	225	289	441
y_3	9	12	16	19	56	81	144	256	361
y_4	8	13	17	20	58	64	169	289	400
Total	38	54	69	80	241	366	734	1195	1602

Step: 1 $N = 16$

Step: 2 $T = 241$

Step: 3
$$\frac{T^2}{N} = \frac{(241)^2}{16} = 3630.0625$$

Step: 4
$$TSS = \Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2 + \Sigma x_4^2 - \frac{T^2}{N}$$
$$= 366 + 734 + 1195 + 1602 - 3630.0625$$
$$= 266.9375$$

Step: 5
$$SSC = \frac{(\Sigma x_1)^2}{n_1} + \frac{(\Sigma x_2)^2}{n_2} + \frac{(\Sigma x_3)^2}{n_3} + \frac{(\Sigma x_4)^2}{n_4} - \frac{T^2}{N}$$

$$= \frac{38^2}{4} + \frac{54^2}{4} + \frac{69^2}{4} + \frac{80^2}{4} - 3630.0625$$

$$= 250.1875$$

Step: 6

$$SSR = \frac{(\Sigma y_1)^2}{m_1} + \frac{(\Sigma y_2)^2}{m_2} + \frac{(\Sigma y_3)^2}{m_3} + \frac{(\Sigma y_4)^2}{m_4} - \frac{T^2}{N}$$

$$= \frac{63^2}{4} + \frac{64^2}{4} + \frac{56^2}{4} + \frac{58^2}{4} - 3630.0625$$

$$= 11.1875$$

Step: 7

$$SSE = TSS - SSC - SSR$$

$$= 5.5625$$

Step: 8 Prepare ANOVA table

Source of Variations	d.f	Sum of Squares	Mean square	Cal. F	tab. F at 5%
B/w columns	$C-1=3$	250.1875	MSC = 83.3958	F_c = 134.93	(3,9) = 3.86
B/w rows	$r-1=3$	11.1875	MSR = 3.7292	F_R = 6.033	(3,9) = 3.8
Errors	$(c-1)(r-1)$ = 9	5.5625	MSE = 0.6181		

Conclusion.

* Cal F_c > tab F_c

  H_0 (i) is rejected (treatment)

* Cal F_R > tab F_R

  H_0 (ii) is rejected (Doctor)

2. The foll/. are the nos/ of mistakes made in 5 successive days of 4 technicians working for a photographic laboratory

Tech I	Tech (II)	Tech - III	Tech IV
6	14	10	9
14	9	12	12
10	12	7	8
8	10	15	10
11	14	11	11

Test at the level of significance $\alpha = 0.01$ whether the differences among the 4 samples can be attributed to chance?

Sl:

Null hypotheses (H_0)

There is no significant diff/ b/w columns (treatment)

Alternative hypotheses (H_1)

There is significant diff/ b/w columns

X_1	X_2	X_3	X_4	Total	X_1^2	X_2^2	X_3^2	X_4^2
6	14	10	9	39	36	196	100	81
14	9	12	12	47	196	81	144	144
10	12	7	8	37	100	144	49	64
8	10	15	10	43	64	100	225	100
11	14	11	11	47	121	196	121	121
49	59	55	50	213	517	717	639	510

Step: 1

$N = \cancel{16}\ 20$

Step: 2

$T = 213$

Step : 3

$$CF = \frac{T^2}{N} = \frac{213^2}{20} = 2268.45$$

Step : 4

$$TSS = \Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2 + \Sigma x_4^2 - \frac{T^2}{N}$$

$$= 517 + 717 + 639 + 510 - 2268.45$$

$$= 114.55$$

Step : 5

$$SSC = \frac{(\Sigma x_1)^2}{n_1} + \frac{(\Sigma x_2)^2}{n_2} + \frac{(\Sigma x_3)^2}{n_3} + \frac{(\Sigma x_4)^2}{n_4} - \frac{T^2}{N}$$

$$= \frac{49^2}{5} + \frac{59^2}{5} + \frac{55^2}{5} + \frac{50^2}{5} - 2268.45$$

$$= 12.95$$

Step : 6

$$SSE = TSS - SSC$$

$$= 114.55 - 12.95 = 101.6$$

Step : 7

Prepare ANOVA Table.

Source of Variations	d.f	Sum of squares	Mean Square	Cal/. F	Tab/. F at 1%
B/w columns	$c-1 = 3$	$SSC = 12.95$	$MSC = 4.3167$	$F_c = \frac{MSE}{MSC} = 1.4710$	$(16,3) = 26.9$
Error	$N-c = 16$	$SSE = 101.6$	$MSE = 6.35$		

Conclusion :

* Cal F_c < tab F_c

⇒ H_0 is accepted (treatment)

~ x ~

Chapter: 4 Testing of hypothesis (Non-Parametric term)

χ^2 - test for goodness of fit

Tests of goodness of fit are used when we want to determine whether an actual sample distribution matches a known theoretical distribution.

$$\chi^2 = \sum (O-E)^2/E$$

$$d.f = n-1.$$

1. The theory predicts that the proportion of beans in the 4 gps A, B, C and D should be 9:3:3:1. In an experiment among 1600 Beans, the nos/. In 4 gps were 882, 313, 287, 118. Do the experimental results support the theory?

Sl:

Null hypothesis (H0):

The experimental results support the theory.

H1 The experimental results do not support the theory.

$$\chi^2 = \sum (O-E)^2/E^2.$$

$$d.f = n-1 = 4-1 = 3.$$

Level of significance = 5%.

Tab/. χ^2 at 5% = 7.815.

Given: Expected ratios : 9 : 3 : 3 : 1

Let Expected values of A, B, C and D be $9x$, $3x$, $3x$ and x.

$$9x + 3x + 3x + x = 1600$$

$$16x = 1600$$

$$\boxed{x = 100}$$

$\therefore$ $E(A) = 9x = 900$

$E(B) = 3x = 300$

$E(C) = 3x = 300$

$E(D) = x = 100$.

O	E	$(O-E)^2/E$
882	900	0.36
313	300	0.5633
287	300	0.5633
118	100	3.24
1600	1600	4.7266

$\therefore$ Cal $\chi^2 = 4.7266$.

Conclusion

Cal χ^2 > tab χ^2.

Ho is rejected.

CamScanner

2. Verify whether Poisson dist/. can be assumed from the data given below:

No/. of defects	0	1	2	3	4	5
Observed frequency	6	13	13	8	4	3

Sol:

Ho: Poisson dist/. fit to the given data

H1: Poisson dist/. do not fit to the given data.

W.K.T The poisson dist/. of X is defined by

$$P(X=x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!} = \frac{e^{-2} \cdot 2^x}{x!}$$

here λ = mean = $\frac{\Sigma f_i x_i}{\Sigma f_i}$

$$= \frac{0\times6 + 1\times13 + 2\times13 + 3\times8 + 4\times4 + 5\times3}{6+13+13+8+4+3}$$

$$= \frac{94}{47} = 2.$$

E(0) = Expected frequency in Poisson distribution is $= N \cdot P(X=x)$

$= 47\ P(X=x)$.

$$E(0) = 47\ P(X=0) = 47 \times \frac{e^{-2} \times 2^0}{0!}$$

$$= 47 \times e^{-2} = 6.3608$$

$$E(1) = 47\ P(X=1) = \frac{47 \times e^{-2} \times 2^1}{1!}$$

$$= 12.7215$$

$$E(2) = 47\ P(X=2) = \frac{47 \times e^{-2} \times 2^2}{2!}$$

$$= 12.7215$$

$$E(3) = 47\, P(X=3) = 47 \frac{e^{-2}\, 2^3}{3!}$$
$$= 8.4810$$

$$E(4) = 47\, P(X=4) = \frac{47 \times e^{-2} \times 2^4}{4!} = 4.2405$$

$$E(5) = 47\, P(X=5) = \frac{47 \times e^{-2} \times 2^5}{5!} = 1.6962$$

O	E	$(O-E)^2/E$
6	6.3608	0.0204
13	12.7215	0.0061
13	12.7215	0.0061
8	8.4810	0.0273
4	4.2405	0.0136
3	1.6962	1.0022
		1.0696

Cal $\chi^2 = 1.0696$.

d.f for poisson dist/: $= n - 2$
$= 6 - 2 = 4$.

Lab χ^2 at 5% $= 9.49$.

Conclusion

Cal χ^2 > Lab χ^2

Ho is accepted.

Rank sum test

Rank sum test is a non parametric test for identifying differences b/w two or more population based on the analysis of two or more independent samples one from each population are used.

1. Mann-Whitney 'U' test (two population)
2. Kruskal-Wallis test (or) H-test. (More than two population).

Mann-Whitney U-Test

H_0: $\mu_1 = \mu_2$ (ie) two populations are identical.

H_1: $\mu_1 \neq \mu_2$ (ie) two populations are non-identical.

Procedure

1. Assign ranks to all samples (from smallest to the largest)
2. Assign the average of the rank if the sample values are same (ie) there are tie scores.
3. Find the sum of the ranks for each of the sample. Let us denote these sums by R_1 and R_2. Also n_1 and n_2 are their respective sample sizes.

   For our convenience choose $n_1 \leq n_2$ (if they are unequal).
4. Calculate U-statistic

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \text{ (for sample I)}$$

(or)

$$U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \text{ (for sample 2)}$$

mean of $U = \frac{n_1 n_2}{2} = E(U)$

Variance of $U = Var(U) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$.

$$\therefore Z = \frac{U - E(U)}{\sqrt{Var(U)}}$$

5) If $|Z| \leq Z_\alpha$,

$\Rightarrow$ we accept H_0

where α is level of significance.

1. The foll are the no/: of mistakes counted on pages randomly selected from reports typed by a company's two secretaries.

Male Secretary : 15 10 5 6 8 10 12

Female secretary : 12 8 7 9 10 5 4

Use 'U' test at 2%. level of significance to test the null hypothesis that the 2 secretaries average equal mistakes per page.

Sol:

H_0 : Two populations are identical.

$\mu_1 = \mu_2$

H_1 : Two populations are non-identical.

X_1	R_1	X_2	R_2
15	4	12	12.5
10	10	8	6.5
5	2.5	7	5
6	4	9	8
8	6.5	10	10
10	10	5	2.5
12	12.5	4	1
	59.5		45.5

Here $n_1 = 7$, $n_2 = 7$, $R_1 = 59.5$, $R_2 = 45.5$

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$= 7 \times 7 + \frac{7(8)}{2} - 59.5$$

$$= 17.5$$

$$E(U) = \frac{n_1 n_2}{2} = \frac{49}{2} = 24.5.$$

$$Var(U) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} = \frac{49(15)}{12}$$

$$= 61.25$$

Now, $z = \frac{U - E(U)}{\sqrt{Var(U)}}$

$$= \frac{17.5 - 24.5}{\sqrt{61.25}}$$

$$= -0.8944$$

$$|z| = 0.8944.$$

Tab/. z at 2% $= 2.326$

Conclusion:

Cal/. z < tab/. z.

$\Rightarrow H_0$ is accepted.

2. The nicotine content of two brands of cigarettes measured in milligrams, was found to be as follows

Brand A : 2.1 4.0 6.3 5.4 4.8 3.7 6.1 3.3

Brand B : 4.1 0.6 3.1 2.5 4.0 6.2 1.6 2.2 1.9 5.4

Use rank sum test, test the hypothesis, at 0.05 level of significance, that the average nicotine contents of the two brands are equal against the alternative that they are unequal.

Sol:

H_0 : Two populations are identical

H_1 : Two populations are non-identical

X_1	R_1	X_2	R_2
2.1	4	4.1	12
4.0	10.5	0.6	1
6.3	18	3.1	7
5.4	14.5	2.5	6
4.8	13	4.6	10.5
3.7	9	6.2	17
6.1	16	1.6	2
3.3	8	1.9	3
		5.4	14.5
	93		78

here $n_1 = 8$; $n_2 = 10$; $R_1 = 93$; $R_2 = 78$.

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$= 8 \times 10 + \frac{8(8+1)}{2} - 93$$

$$= 80 + 36 - 93 = 23$$

$$E(U) = \frac{n_1 n_2}{2} = \frac{8 \times 10}{2} = 40$$

$$Var(U) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} = \frac{80 \times 19}{12}$$

$$= 126.6667$$

Now, $$Z = \frac{U - E(U)}{\sqrt{Var(U)}} = \frac{23 - 40}{\sqrt{126.667}}$$

$$= -1.5105$$

$$|Z| = 1.5105$$

At 5% level of significance for two tailed test $z = 1.96$.

Here cal z < tab z.

H_0 is accepted.

Kruskal - Wallis Test or H - Test

The Mann - Whitney U test can be used to test whether two populations are identical. It has been extended to the case of 3 or more populations by Kruskal and Wallis. The hypothesis for K-W test with $k \geq 3$ can be written as follows.

$H_0: \mu_1 = \mu_2 = \mu_3$.

(ie) All populations are identical.

H_1: All populations are non-identical.

* K-W test can be computed as follows

$$H \text{ (or) } W = \frac{12}{n(n+1)} \left[\sum_{i=1}^{k} \frac{R_i^2}{n_i} \right] - 3(n+1)$$

where n_i = the no/. of items in sample i

k = no/. of populations (or samples).

$n = n_1 + n_2 + \dots + n_k$

R_i = Sum of the Ranks of all items in sample i.

Note * The sampling distribution of 'W' can be approximated by a χ^2-distribution with $(k-1)$ d.f.

1. Use Kruskal - Wallis test to test for differences in mean among 3 samples. If $\alpha = 0.01$, what are your conclusion.

Sample I : 95 97 99 98 99 99 99 94 95 98

Sample II : 104 102 102 105 99 102 111 103 100 103

Sample III : 119 130 132 136 141 172 145 150 144 135.

Sol:

$H_0 : \mu_1 = \mu_2 = \mu_3$.

All population means are identical.

$H_1 : \mu_1 \neq \mu_2 \neq \mu_3$.

Sample I	R_1	Sample II	R_2	Sample III	R_3
95	2.5	104	18	119	21
97	4	102	14	130	22
99	9	102	14	132	23
98	5.5	105	19	136	25
99	9	99	9	141	26
99	9	102	14	172	30
99	9	111	20	145	28
94	1	103	16.5	150	29
95	2.5	100	12	144	27
98	5.5	103	16.5	135	24
	57		153		255

Here $n = n_1 + n_2 + n_3 = 10 + 10 + 10 = 30$.

$$W = \frac{12}{n(n+1)} \left[\sum_{i=1}^{3} \frac{R_i^2}{n_i} \right] - 3(n+1)$$

$$= \frac{12}{30(31)} \left[\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \frac{R_3^2}{n_3} \right] - 3(31)$$

$$= \frac{12}{30(31)} \left[\frac{57^2}{10} + \frac{153^2}{10} + \frac{255^2}{10} \right] - 93$$

$$= 25.30.$$

The χ^2 value at 1% leve ~~wrt~~ ~~cat~~ with

d. f $= K-1 = 3-1 = 2$ is

$$\boxed{\chi^2_\alpha = 9.21.}$$

Here. Cal W > tab/: χ^2_α.

$\Rightarrow$ Ho is rejected.

2 A company's trainees are randomly assigned to groups which are taught a certain industrial inspection procedure by 3-different methods. At the end of the instruction period they are tested for inspection performance quality. The foll/: their scores.

Method A: 80 83 79 85 90 68

Method B: 82 84 60 72 86 67 91

Method C: 93 65 77 78 88

Use H-test to determine at 5% level of significance whether 3-methods are equally effective.

Sol:

$H_0: \mu_1 = \mu_2 = \mu_3$ (All populations are identical)

$H_1: \mu_1 \neq \mu_2 \neq \mu_3$.

Method A	R_1	Method B	R_2	Method C	R_3
80	9	82	10	93	18
83	11	84	12	65	2
79	8	60	1	77	6
85	13	72	5	78	7
90	16	86	14	88	15
68	4	67	3		
		91	17		
	61		62		48

Here $n_1 = 6$; $n_2 = 7$; $n_3 = 5$

$n = n_1 + n_2 + n_3 = 18$.

$K = 3$ (number of method).

$$W = \frac{12}{n(n+1)} \left[\sum_{i=1}^{K} \frac{R_i^2}{n_i} \right] - 3(n+1)$$

$$= \frac{12}{18 \times 19} \left[\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \frac{R_3^2}{n_3} \right] - 3(18+1)$$

$$= \frac{12}{18 \times 19} \left[\frac{61^2}{6} + \frac{62^2}{7} + \frac{48^2}{5} \right] - 3(19)$$

$$= 57.8169 - 57$$

$$W = 0.8169.$$

The χ^2 value at 5% level with $d.f = K - 1 = 2$.

$\therefore \chi^2_{\alpha} = 5.991$.

Conclusion:

Cal: $W > \chi^2$

$\Rightarrow$ H_0 is accepted.

Kolmogorov - Smirnov Test (K-S Test)

* It is a simple non-parametric test for testing whether there is a significance between an observed frequency distribution and a theoritical frequency distribution.

* The K-S test is another measure of the goodness of fit of a frequency distribution as was the χ^2 - Test.

* $D_n = \max |F_e - F_o|$.

$\Rightarrow$ maximum absolute deviation of expected relative frequency F_e and observed relative frequency F_o.

Advantages :

* It is more powerful test.
* It is easier to use, since it does not require that the data be grouped in any way.

1 Below is the table of observed frequencies along with the frequency to the observed under a normal distribution.

a) Calculate the K-S statistic.

b) Can we conclude that this distribution does infact follow a normal distribution?

Use 0.10 level of significance.

Test score :	51-60	61-70	71-80	81-90	91-100
Observed freq/. :	30	100	440	500	130
Expected freq/. :	40	170	500	390	100.

St:

Ho : This distribution follows a normal distribution.

H_1 : This distribution does not follow a normal distribution.

Observed frequency	Observed Cumulative frequency	Fo	Expected frequency	Expected Cumulative frequency
30	30	$\frac{30}{1200} = 0.025$	40	40
100	130	$\frac{130}{1200} = 0.108$	170	210
440	570	$\frac{570}{1200} = 0.475$	500	710
500	1070	$\frac{1070}{1200} = 0.891$	390	1100
130	1200	$\frac{1200}{1200} = 1$	100	1200

CS CamScanner

Fe	$\lvert Fe - Fo \rvert$
$\frac{40}{1800} = 0.033$	0.008
$\frac{210}{1800} = 0.175$	0.067
$\frac{710}{1800} = 0.592$	0.117
$\frac{1100}{1800} = 0.920$	0.029
$\frac{1800}{1800} = 1$	0

$\therefore D_n = \max \lvert Fe - Fo \rvert = 0.117$.

Level of significance $\alpha = 0.10 = 10\%$.

~~Tab/: of χ^2 at 10% of d.f = 5-1 = 4~~

Tab/: of D_n at 10% of $n=5$ is 0.510.

Conclusion

* Cal/: D_n < tab/: D_n.

$\Rightarrow$ Ho is accepted

Sign Test

Sign test is conducted under the following circumstances

1. When there are pair of observations on two things being compared.

2. For any given pair, each of the two observation is made under similar conditions.

3. No assumptions are made regarding the Parent population.

CS CamScanner

Working Rule

1. Omitting zero differences, find the no/: of +ve deviations in $d_i = x_i - y_i$. Let it be k.

When $n \leq 30$

2. Find $p' = P(u \leq k)$

$$= \left(\frac{1}{2}\right)^n \sum_{x=0}^{k} \binom{n}{x}$$

$$= \left(\frac{1}{2}\right)^n \sum_{x=0}^{k} nC_x \quad [\text{if } K \text{ is no/: of +ve deviation}]$$

Find $p' = P(u \geq k)$

$$= \left(\frac{1}{2}\right)^n \sum_{x=k}^{n} nC_x \quad [\text{if } K \text{ is no/: of } -\text{ve deviation}]$$

3. * If $p' \leq 0.05 \Rightarrow H_0$ is rejected

* If $p' > 0.05 \Rightarrow H_0$ is accepted.

When $n > 30$,

1. Find $Z = \dfrac{u - \frac{n}{2}}{\sqrt{n/4}} \sim N(0,1)$

where.

2. * If $|z| \leq 1.96$ (two failed)

$\Rightarrow H_0$ is accepted at 5% Level of Significance.

* If $|z| \leq 2.58$ - (two tailed)

$\Rightarrow H_0$ is accepted at 1% LOS.

1. The following data show the employee's rates of defective work before and after a change in the wage incentive plan. Compare the foll/. two sets of

data to see whether the charge lowered the defective units produced. Using the sign test with $\alpha = 0.01$.

Before:	8	7	6	9	7	10	8	6	5	8	10	8
After:	6	5	8	6	9	8	10	7	5	6	9	8

Sol:

Null hypothesis: $H_0 : p = 0.5$.

(ie) There is no significant change in the defective units produced.

$H_1 : p \neq 0.5$.

There is significant change in the defective units produced.

x	y	$d = x_i - y_i$	sign
8	6	2	+
7	5	1	+
6	8	−2	−
9	6	3	+
7	9	−2	±
10	8	2	+
8	10	−2	±
6	7	−1	∓
5	5	0	0
8	6	2	+
10	9	1	+
8	8	0	0

No. of +ve deviations = 6

No. of −ve deviations = 4

$\therefore n = 6 + 4 = 10 < 30$.

Take $k = 4$ (as +ve deviations).

$$\therefore p' = P(u \leq k) = \left(\frac{1}{2}\right)^n \sum_{x=0}^{k} nC_x.$$

$$p' = \left(\frac{1}{2}\right)^{10} \sum_{x=0}^{6} 6C_x$$

$$= \left(\frac{1}{2}\right)^{10} \left[6C_0 + 6C_1 + 6C_2 + 6C_3 + 6C_4 + 6C_5 + 6C_6\right]$$

$$= \left(\frac{1}{2}\right)^{10} \left[1 + 6 + 15 + 20 + 15 + 6 + 1\right]$$

$$p' = 0.0624$$

Conclusion:

$p' = 0.0624 > 0.05$

$\Rightarrow H_0$ is accepted.

2. The following data in tons, are the amounts of sulphur oxides emitted by a large industrial plant in 40 days.

24	15	20	29	19	18	22	25	27	9
17	20	17	6	24	14	15	23	24	26
19	23	28	19	16	22	24	17	20	13
19	10	23	18	31	13	20	17	24	14

use the sign test to test the null hypothesis $\mu = 21.5$ against alternative hypothesis $\mu > 21.5$ at the 0.01 LOS.

Sol:

$H_0 : \mu = 21.5$ $(\cancel{p \neq 1/2})$

$H_1 : \mu > 21.5$. $(\cancel{p \neq 1/2})$ (one tailed test).

Given $\cancel{n = 40} > \cancel{30}$ ∴ $n = 40 > 30$

$\cancel{\therefore z =}$ ∴ $z = \dfrac{u - \frac{n}{2}}{\sqrt{\frac{n}{4}}}$

where u is no/. of +ve deviation.

Now, Replacing each value exceeding 21.5 with plus sign, each value less than 21.5 with minus sign, we get

+ − − + − − + − − + + + − − −
− − + − − + + + − + + − − + +
− − − − − + − + − − − + −

here u = no/. of + signs = 16.

$n = 16 + 24 = 40$.

$$z = \frac{16 - \frac{40}{2}}{\sqrt{40/4}} = \frac{-4}{3.16} = -1.26.$$

$|z| = 1.26$.

tab/. z at 1% Los for one tailed test is 2.33.

Conclusion

~~tab~~ cal z/. < tab z.

$\Rightarrow$ H_0 is accepted with $\mu = 21.5$.

One Sample Run Test

A run is a subsequence of one or more identical symbols representing a common property of the data.

Working Rule

H_0 : Observations are generated Randomly.

H_1 : Observations are not randomly generated (two tailed test).

$$z = \frac{R - E(R)}{\sqrt{V(R)}} = \frac{R - \mu}{\sigma}$$

where $\mu = \frac{2n_1 n_2}{n_1+n_2} + 1$

$$\sigma = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1+n_2)^2 (n_1+n_2-1)}}$$

* If $|Z| \leq 1.96 \Rightarrow H_0$ accepted at $\alpha = 5\%$.
* if $|Z| \leq 2.58 \Rightarrow H_0$ accepted at $\alpha = 1\%$.

1. The following is an arrangement of 25 men (M) and 15 women (W) lined up to purchase tickets for a premier picture show

M	WW	MMM	W	MM	W	M	W	M	WWW	MMM
1	2	3	4	5	6	7	8	9	10	11

W	MM	WWW	MMMMMM	WWW	MMMMMM
12	13	14	15	16	17.

Test for randomness at 5% LOS.

Sol:

H_0: Arrangement of M and W are random

H_1: Arrangement of M and W are not random.

By Run test,

$$Z = \frac{R - \mu}{\sigma}$$

where R = No/: of runs = 17.

n_1 = no/: of Mens = 25

n_2 = no/: of Women = 15

$$\mu = \frac{2n_1 n_2}{n_1+n_2} + 1 = \frac{2(25)(15)}{25+15} + 1$$

$= 19.75$.

$$\sigma = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$$

$$= \sqrt{\frac{2\times25\times15\,(2\times25\times15 - 25 - 15)}{(25+15)^2\,(25+15-1)}}$$

$$= \sqrt{\frac{750\,(710)}{1600\times39}} = 2.92$$

$$\therefore z = \frac{17 - 19.75}{2.92} = -0.94$$

$$|z| = 0.94$$

Conclusion

Cal z < tab/: $z = 1.96$ at 5% LOS.

$\Rightarrow$ Ho is accepted.

2. The foll/: are the prices in Rs./kg of a commodity from 2 random samples of shops from 2 cities A and B.

City A : 2.73 3.82 4.35 3.23 4.74 3.65 3.8 4.15 2.76 2.85 3.25 3.45 3.85 4.45 4.95 3.95 4.78.

City B : 3.75 5.37 4.78 3.69 4.75 4.85 6.0 4.8 4.9 3.84 3.9 4.8 5.23 6.1 3.6 3.83

Apply the run test to examine whether the distribution of prices of commodity in the two cities is the same.

Sl:

Ho: The distribution of prices of commodity in the 2 cities is same.

H1: Distribution of 2 cities is not same.

$\boxed{\text{LOS} = 5\% = \alpha.}$

Let us combine all Observations from both cities and arrange them in ascending Order.

Assign the letter A to city A.

Assign the letter B to city B.

A	A	A	A	A	A
2.73	2.76	2.85	3.23	3.25	3.45

1

B	A	B	B	A	A	B	B
3.6	3.65	3.69	3.75	3.8	3.82	3.83	3.84
2	3	4		5		6	

A	B	A	A	A	A	A	A
3.85	3.9	3.95	4.15	4.35	4.45	4.72	4.74
7	8	9					

B	B	B	B	B	B	A
4.75	4.78	4.8	4.8	4.85	4.9	4.95
			10			11

B	B	B	B
5.23	5.27	6.0	6.1

12.

R = total run = 12.

n_1 = no. of Letter A = 17.

n_2 = no. of Letter B = 16.

$$\mu = \frac{2n_1n_2}{n_1+n_2} + 1 = 17.485$$

$$\sigma = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2\ (n_1+n_2-1)}} = \sqrt{7.977} = 2.824$$

$$Z = \frac{R-\mu}{\sigma} = \frac{12 - 17.485}{2.824} = 1.94$$

$|Z| = 1.94$.

Tab. value of Z at 5% LOS = 1.96.

Conclusion.

Cal. $|Z|$ < tab. $|Z|$

$\Rightarrow$ H_0 is accepted.

χ^2 - distribution of different attributes

χ^2-test is used to find out whether two or more attributes are associated or not. This test helps in finding the association or independence of two or more attributes.

1. A certain college is studied for gender and UG background. The results are as follows.

	Engineering	Non-Engineering.
Boys	70	30
Girls	25	95

Q: Frame the hypothesis and does χ^2-test for this data to identify association b/w ~~gen~~ gender and UG ~~gen~~ qualification.

Sol:

H_0: Genders and UG qualifications are independent.

H_1: Genders and UG qualifications are not independent.

LOS ⇒ $\alpha = 5\%$

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}.$$

			Total
	70	30	100
	25	95	120
Total	95	125	220

$$E(70) = \frac{100 \times 95}{220} = 43.182.$$

$$E(30) = \frac{125 \times 100}{220} = 56.818$$

$$E(25) = \frac{95 \times 120}{220} = 51.818$$

$$E(95) = \frac{125 \times 120}{220} = 68.182$$

O	E	$(O-E)^2$	$(O-E)^2/E$
70	43.184	719.205	16.655
30	56.818	719.205	12.658
25	51.818	719.205	13.879
95	68.182	719.205	10.548
			53.74

$\chi^2 = 53.74$.

$d.f = (c-1) \times (r-1) = (2-1) \times (2-1) = 1$.

Tab χ^2 at $5\% = 3.841$.

Conclusion

Cal χ^2 > tab χ^2.

Ho is rejected.

2. Various countries are compared using two variables composition of economy and growth band as shown in the foll/. cross table.

	High growth	Medium Growth	Low Growth
Predominant Agriculture	20	25	5
Predominant Manufacturing	40	5	6
Predominant Services	5	55	20

Test whether the predominant fun/. in an economy has an impact on the growth of the economy using χ^2-test.

H_0 : Predominant fun/. and growth of economy are independence.

H_1 : Both are not independent.

				Total
	20	25	5	50
	40	5	6	51
	5	55	20	80
Total	65	85	31	181

$$E(20) = \frac{50 \times 65}{181} = 17.95.$$

$$E(25) = \frac{50 \times 85}{181} = 23.5.$$

$$E(5) = \frac{50 \times 31}{181} = 8.55.$$

$$E(40) = \frac{51 \times 65}{181} = 18.32.$$

$$E(5) = \frac{51 \times 85}{181} = 23.95.$$

$$E(6) = \frac{51 \times 31}{181} = 8.73$$

$$E(5) = \frac{80 \times 65}{181} = 28.73.$$

$$E(55) = \frac{80 \times 85}{181} = 37.57$$

$$E(20) = \frac{80 \times 31}{181} = 13.7.$$

O	E	$(O-E)^2$	$(O-E)^2/E$
20	17.95	4.2025	0.234123
25	23.5	2.25	0.095745
5	8.55	12.6025	1.473977
40	18.32	470.0224	25.65604

5	~~23.73~~ 23.95	~~563.xxx~~ 359.1025	~~19.60017~~ 14.99384
6	8.73	7.4529	0.853711
5	28.73	563.1129	19.60017
55	37.57	303.8049	8.086369
20	13.7	39.69	2.89708.
			73.89186

$$\chi^2 = 73.8913.$$

$$d.f = (c-1)\times(r-1) = (3-1)(3-1)$$

$$= 2\times 2 = 4.$$

tab χ^2 at 5% LOS = 9.488.

Conclusion:

Cal. χ^2 > tab. χ^2

$\Rightarrow$ Ho is rejected.

Unit-5

Correlation and Regression

Correlation

Correlation is a statistical tool used to measure the relationship b/w two sets of variables and express each in a precise manner.

Types of correlation.

* Positive and Negative correlation.
* ~~Ra~~ Simple, Multiple and Partial Correlation.
* Linear and Non-linear correlation.

Positive Correlation

Correlation is positive when two variables vary in the same direction.

Ex: Correlation b/w sales and expenses.

Negative Correlation

Correlation is negative when both the variables vary in the opposite direction.

Ex: Correlation b/w production and price of crop.

Linear Correlation

Changes in the value of one variable has a fixed ratio to the variation in the values of other variable. When the variables are plotted on a graph, it will fall on straight line.

Non-Linear Correlation (Curvilinear)

Changes in values of one variable does not have a fixed ratio to the variation in the value of other variable. When we plotted these points on a graph, the plotted points would fall on a curve.

Simple correlation

When we measure the linear relationship b/w two variables then this interpretation is known as simple correlation.

Ex: relationship b/w sales and expenses and income and consumption etc.

Partial correlation.

If we have various related variables and try to find out the relationship b/w two variables then it is known as partial correlation.

Multiple correlation:

It is defined as, the measurement of the effect of multiple variables on one variable.

Degree of correlation

Degree	+ve correlation co-eff	-ve correlation co-eff.
1. Perfect	1	-1
2. Limited		
i) High	B/w 0.75 to 1	B/w -0.75 to -1
ii) Moderate	B/w 0.25 to 0.75	B/w -0.25 to -0.75
iii) Low	B/w 0 to 0.25	B/w 0 to -0.25
3. Absence	—	—

Difference B/w co-eff of Determination and co-eff of correlation

Correlation (r)	Determination (r^2)
1. It is used to measure a linear relationship b/w two variables.	1. The squared correlation gives proportion of common variance b/w two variables.
2. The co-eff of correlation indicates the amount of information common to two variables.	2. The co-eff of determination is used to analyze how differences in one variable can be explained by a difference in a second variable.
3. $-1 \leq r \leq 1$	3. $0 \leq r^2 \leq 1$.

Advantage of Correlation Analysis

* Observe Relationships
* A good starting point for Research
* Uses for further studies.
* Simple metrices.

Applications of Correlation.

1. The nature, direction and degree of relationship B/w two or more variables are determined by the use of correlation analysis.
2. It is used for estimating the change in value of one variable occurs due to the changes in the value of other variable.
3. Correlation analysis is helpful in making predictions.

Methods of computing correlation

1. Scatter Diagram.
2. Karl Pearson's co-eff of correlation.
3. Spearman's Rank correlation.

Scatter diagram:

* Graphical representation of the relationship b/w two variables calculated on the same set of individual is known as scatter diagram.
* Scatter diagram is a dot chart specially used to show the correlation.

Advantages.

1. It is very simple and non-mathematical technique.
2. It is not influenced by the size of extreme item.
3. It is the very basic step to find out the relationship b/w two variables.

Disadvantages

* The main disadvantage of this technique is that it cannot find out an exact degree of ~~p~~ correlation b/w two variables.
* We can only view the visual form of correlation and direction on the chart.

Karl Pearson's co-eff of correlation.

To calculate the magnitude of linear relationship b/w two variables Karl Pearson gave a quantitative technique.

This technique is known as Pearsonian co-eff of correlation (r) and is extensively used in practice.

Properties:

* Ideal measure of correlation and is independent of the units of the variables.
* Free from change of origin and scale.
* Based on all observations.
* Lies b/w -1 and $+1$.

if $r = -1 \Rightarrow$ perfect -ve correlation

if $r = 1 \Rightarrow$ Perfect +ve correlation.

if $r = 0 \Rightarrow$ no correlation b/w two variables.

Advantages:

The main advantage is that it gives the result in one value and also summarize the degree of correlation and direction.

Disadvantages

1. Every time assumes only a linear relationship b/w variables.
2. It is a time consuming method.
3. Does not convey the cause and effect relationship.
4. Significance of correlation co-eff is affected by the extreme values.
5. Interpreting the value of correlation co-eff is difficult.

Formula:

$$r = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y} = \frac{E(XY) - E(X) \cdot E(Y)}{\sqrt{E(X^2) - E(X)^2}\ \sqrt{E(Y^2) - E(Y)^2}}.$$

$$= \frac{\Sigma (x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma (x-\bar{x})^2}\ \sqrt{\Sigma (y-\bar{y})^2}}$$

1) Find the co-efficient of correlation between x and y using the following data:-

x	65	67	66	71	67	70	68	69
y	67	68	68	70	64	67	72	70

Sollution:

$$r_{xy} = \frac{cov(x,y)}{\sigma_x \sigma_y}.$$

x	y	x^2	y^2	xy
65	67	4225	4489	4355
67	68	4489	4624	4556
66	68	4356	4624	4488
71	70	5041	4900	4970
67	64	4489	4096	4288
70	67	4900	4489	4690
68	72	4624	5184	4896
69	70	4761	4900	4830
543	546	36885	37306	37073

$$r_{xy} = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

$$E(X) = \frac{\Sigma x}{n} = \frac{543}{8} = 67.88$$

$$E(Y) = \frac{\Sigma y}{n} = \frac{546}{8} = 68.25$$

$$E(x^2) = \frac{\Sigma x^2}{n} = \frac{36885}{8} = 4610.66$$

$$E(y^2) = \frac{\Sigma y^2}{n} = \frac{37306}{8} = 4663.25$$

$$E(xy) = \frac{\Sigma xy}{n} = \frac{37073}{8} = 4634.13$$

$$\text{cov}(x,y) = E(xy) - E(x)\,E(y)$$

$$= 4634.13 - 67.88 \times 68.25$$

$$= 4634.13 - 4632.81$$

$$= 1.32$$

$$\sigma_x = \sqrt{\text{var}\,x} = \sqrt{\Sigma(x^2) - \Sigma(x)^2}$$

$$= \sqrt{4610.66 - (67.88)^2}$$

Fd

$$= \sqrt{4610.66 - 4607.69}$$

$$= \sqrt{2.97}$$

$$= 1.72$$

$$\sigma_y = \sqrt{\text{var}\,y} = \sqrt{\Sigma(y^2) - (\Sigma y)^2}$$

$$= \sqrt{4663.25 - (68.25)^2}$$

$$= \sqrt{4663.25 - 4658.06}$$

$$= \sqrt{5.19}$$

$$= 2.27$$

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y}$$

$$= \frac{1.32}{1.72 \times 2.27}$$

$$= \frac{1.32}{3.90}$$

$$= 0.33$$

5. Two cities A and B are compared for temperatures on certain days. Ten samples are taken. Find the coefficient of correlation between the temperatures of the two cities.

city A	25	29	33	45	41	28	21	20	27
city B	22	32	31	46	44	26	25	25	23

Sollution:-

$$r = \frac{N\Sigma dx dy - \Sigma dx \, \Sigma dy}{\sqrt{N\Sigma dx^2 - (\Sigma dx)^2}\ \sqrt{N\Sigma dy^2 - (\Sigma dy)^2}}$$

$\therefore$ Assumed mean for x is 25

Assumed mean for y is 31.

X	Y	dx = X − A A = 25	dx^2	dy = Y − A A = 31	dy^2	dxdy
25	22	0	0	−9	81	0
29	32	4	16	1	1	4
33	31	8	64	0	0	0
45	46	20	400	15	225	300
41	44	16	256	13	169	208
28	26	3	9	−5	25	−15
21	25	−4	16	−6	36	24
20	25	−5	25	−6	36	30
27	23	2	4	−8	64	−16
		40 184	790	−5	637	535

$$r = \frac{N\Sigma dxdy - \Sigma dx \, \Sigma dy}{\sqrt{N\Sigma dx^2 - (\Sigma dx)^2}\,\sqrt{N\Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{9 \times 535 - (44 \times -5)}{\sqrt{9 \times 790 - (44)^2}\,\sqrt{9 \times 637 - (-5)^2}}$$

$$= \cancel{\frac{5035}{\sqrt{5510}\,\sqrt{5708}}} \qquad = \frac{5035}{\sqrt{5174}\,\sqrt{5708}}$$

$$= \cancel{\frac{5015}{74.23 \times 75.84}} \qquad = \frac{5035}{71.93 \times 75.55}$$

$$= \cancel{\frac{5015}{5632.55}} \qquad \boxed{r = 0.9265}$$

$$= \cancel{0.8903}$$

Spearman's Rank Correlation

This technique is used when value of variable cannot be calculated quantitatively.

To apply the spearman's rank correlation technique we need to first arrange the value of variable is serial order.

$$R = 1 - \frac{6 \Sigma d^2}{n(n^2-1)}$$

where R = Rank co-eff of correlation.

d = difference b/w two ranks $(R_1 - R_2)$

Σd^2 = sum of squares of difference of ranks.

n = number of pair of observations.

Note

1. Rank correlation always lies b/w -1 and $+1$
2. If no rank is given, then we first calculate the rank of the given data. after arr,
3. Repeated Ranks

When values are repeated, then we calculate the average of rank and assign it to the repeated values.

$$R = 1 - \frac{6\cdot\left(\Sigma d^2 + m_1(m_1^2-1) + m_2(m_2^2-1) + \ldots\right)}{N(N^2-1)}$$

where m_i = no. of items having common rank

Merits:

* Easy to understand and simple to calculate.
* When data are qualitative.
* It also applies when actual data are giv

7. Calculate coefficient of Rank Correlation:

capital in lakh	66	55	46	33	22	18	11	8	7	11
Profit in lakh	58	43	36	27	15	9	12	15	6	14

Sollution:

$$R = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

X	R_1	Y	R_2	$d = R_1 - R_2$	d^2
66	1	58	1	0	0
55	2	43	2	0	0
46	3	36	3	0	0
33	4	27	4	0	0
22	5	15	5.5	-0.5	0.25
18	6	9	9	-3	9
11	7.5	12	8	-0.5	0.25
8	9	15	5.5	3.5	12.25
7	10	6	10	0	0
11	7.5	14	7	0.5	0.25
					22

also 11 is repeated 2 times.

$$CF_1 = \frac{m(m^2-1)}{12}$$

$$= \frac{2(2^2-1)}{12}$$

$$= 0.5$$

15 is repeated 2 times

$$CF_2 = \frac{m(m^2-1)}{12}$$

$$= \frac{2(2^2-1)}{12}$$

$$= 0.5.$$

$$R = 1 - \frac{6\sum d_i^2 + CF_1 + CF_2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 22 + 0.5 + 0.5}{10(10^2-1)}$$

$$= 1 - \frac{133}{990}$$

$$= 1 - 0.1343$$

$$= 0.8657.$$

8. Ten competitors in a beauty contest are ranked by three judges in the following order.

I Judge	1	6	5	10	3	2	4	9	7	8
II Judge	3	5	8	4	7	10	2	1	6	9
III Judge	6	4	9	8	1	2	3	10	5	7

(i) r_{12} between the ranks of judges I and II.

(ii) r_{23} between the ranks of judges II and III

(iii) r_{13} between the ranks of judges I and III

Sollution:

Judge I	Judge II	Judge III	D_{12}	D_{12}^2	D_{23}	D_{23}^2	D_{13}	D_{13}^2
R_1	R_2	R_3	R_1-R_2		R_2-R_3		R_1-R_3	
1	3	6	-2	4	-3	9	-5	25
6	5	4	1	1	1	1	2	4
5	8	9	-3	9	-1	1	-4	16
10	4	8	6	36	-4	16	2	4
3	7	1	-4	16	6	36	2	4
2	10	2	-8	64	8	64	0	0
4	2	3	2	4	-1	1	1	1
9	1	10	8	64	-9	81	-1	1
7	6	5	1	1	1	1	2	4
8	9	7	-1	1	2	4	1	1
			0	200	0	214	0	60

Formulas:

$$r_{12} = 1 - \frac{6 \sum D_{12}^2}{n(n^2-1)} = 1 - \frac{6 \times 200}{10 \times 99} = 1 - 1.212 = -0.212.$$

$$r_{23} = 1 - \frac{6 \sum D_{23}^2}{n(n^2-1)} = 1 - \frac{6 \times 214}{10 \times 99} = 1 - 1.297 = -0.297.$$

$$r_{13} = 1 - \frac{6 \sum D_{13}^2}{n(n^2-1)} = 1 - \frac{6 \times 60}{10 \times 99} = 1 - 0.364 = +0.636$$

Conclusion:

The Judges I and III have nearest approach to common tastes.

Regression Analysis

Regression Analysis is a statistical tool used to calculate a continuous dependent variable from various independent variables and is commonly used for prediction and forecasting.

Application

1. Used to know the relationship b/w different (one or more) variables.
2. Used to find out the co-eff. of correlation (r) and co-eff of determination (r^2).
3. In corporate sector it is useful to check the quality.
4. Also very useful to determine the statistical curve (demand, supply etc).

Difference b/w correlation and Regression

Correlation	Regression
1. Study of the linear relationship b/w two variables is known as correlation	1. It is statistical tool used to calculate a continuous dependent variable from various independent variables.
2. $-1 \leq r \leq 1$	2. It is possible that co-eff is >1.
3. Correlation co-eff is symmetric $r_{xy} = r_{yx}$.	3. regression co-eff is not symmetric $b_{xy} \neq b_{yx}$.
4. Used to test the relationship b/w two variables.	4. Used for prediction and forecasting.

Determination of Linear Regression Equation

Method of Least squares Method

* Regression eqn/: of x on y is defined by

$$x = a + by \rightarrow (1)$$

and
$$\Sigma x = Na + b\Sigma y \rightarrow (2)$$
$$\Sigma xy = a \cdot \Sigma y + b\Sigma y^2 \rightarrow (3)$$

We get the values of a and b from eqn (2) and (3).

* Regression eqn/: of y on x is defined by

$$y = a + b x \rightarrow (1)$$

and
$$\Sigma y = a \cdot N + b \cdot \Sigma x \rightarrow (2)$$
$$\Sigma xy = a \cdot \Sigma x + b \cdot \Sigma x^2 \rightarrow (3)$$

We get the values of a and b from eqn/: (2) and (3).

Problems:

Regression Equations when Deviation taken from Actual mean

* Regression eqn/: of x on y is defined by

$$(x - \bar{x}) = b_{xy}(y - \bar{y}), \text{ where}$$

$$b_{xy} = r \cdot \frac{\sigma_x}{\sigma_y} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2}.$$

* Regression eqn/: of y on x is defined ~~Actual Mean~~. by

$$(y - \bar{y}) = b_{yx}(x - \bar{x}), \text{ where}$$

$$b_{yx} = r \cdot \sigma_y / \sigma_x = \Sigma(x-\bar{x})(y-\bar{y}) / \Sigma(x-\bar{x})^2$$

Properties of Regression co-eff.

1. bxy and byx are called as regression co-eff.

2. Correlation co-eff $(r) = \pm\sqrt{b_{xy} \cdot b_{yx}}$.

3. * If both bxy and byx is positive, then r is positive.
   * If both bxy and byx is Negative, then r is negative.

4. * If one regression co-eff is greater than unity, then remaining must be smaller than unity.

Co-eff. of Determination (r^2).

$$\text{co-eff of Determination} = \frac{\text{Explained Variation}}{\text{Total Variance}}.$$

Co-eff of Non-Determination (k^2).

$$\text{co-eff of Non-Determination} = \frac{\text{Unexplained Variation}}{\text{Total variance}}$$

$$= 1 - r^2.$$

Probable and Standard error of Estimate

$$S.E = \frac{1 - r^2}{\sqrt{N}}$$

$$P.E = 0.6745 \times S.E$$

$$= 0.6745 \times \frac{1 - r^2}{\sqrt{N}}.$$

CamScanner

Utility of P.E

1. If $|r| > 6 \cdot P.E$

$\Rightarrow$ co-eff of correlation (r) is significant.

2. If $|r| < 6 \cdot P.E$

$\Rightarrow$ co-eff of correlation (r) is insignificant.

1. Obtain the equations of the regression lines from the foll. data. Using method of least square. Hence find the co-eff correlation b/w x and y. Also estimate the value of

1. y when $x = 38$
2. x when $y = 18$

x:	20	26	29	30	31	31	34	35
y:	20	20	21	29	27	24	27	31

Sol:

X	Y	X^2	Y^2	XY
20	20	400	400	400
26	20	676	400	520
29	21	841	441	609
30	29	900	841	870
31	27	~~729~~ 961	729	837
31	24	961	576	744
34	27	~~729~~ 1156	729	918
35	31	1225	961	1085
236	199	7120	5077	5983

Regression eqn/: of X on Y

$$X = a \cdot + bY \rightarrow (1)$$

and

$$\Sigma X = a \cdot N + b \cdot \Sigma Y$$

$$\Rightarrow 236 = 8a + 199b \rightarrow (2)$$

$$\Sigma XY = a\Sigma Y + b\,\Sigma Y^2$$

$$5983 = 199a + 5077b \rightarrow (3)$$

$(2) \times 199 \Rightarrow$ $1592\,a + 39601\,b = 46964$

$(3) \times 8 \Rightarrow$ $1592\,a + 40616\,b = 47864$

$(-) \quad (-) \quad (-)$

$$-1015\,b = -900$$

$$b = \frac{-900}{-1015} = 0.89$$

Sub/. $b = 0.89$ in (2)

$$8a + 199(0.89) = 236.$$

$$8a = 236 - 177.11 = 58.89$$

$$a = 7.36$$

$\therefore (1) \Rightarrow$ $\boxed{X = 7.36 + 0.89\,Y}$

Regression eqn/: Y on X.

$$Y = a + bX. \rightarrow (1)$$

and

$\Sigma Y = a \cdot N + b \cdot \Sigma X \Rightarrow 199 = 8 \cdot a + 236 \cdot b \rightarrow (2)$

$\Sigma XY = a \cdot \Sigma X + b\,\Sigma X^2 \Rightarrow 5983 = 236 \cdot a + 7120\,b \rightarrow (3)$

$(2) \times 236 \Rightarrow$ $1888\,a + 55696\,b = 46964$

$(3) \times 8 \Rightarrow$ $1888\,a + 56960\,b = 47,864$

$(-) \quad (-) \quad (-)$

$$-1264\,b = -900$$

$b = \cancel{3.14090}\ 0.7120$ $\quad 0.7120$

Sub/: in (2) $\Rightarrow 8a = 199 - 236\,(\cancel{3.14090})$

$8a = \cancel{605.524}\ 30.968$

$a = \cancel{-75.6905}\ 3.871$

$Y = \cancel{-75.6905}\ 3.871 + \cancel{3.14090}\ 0.7120\,X.$

$b_{yx} = \cancel{3.14090}\ 0.7120$

$b_{xy} = 0.89$

$r = \cancel{\sqrt{3.14090 \times 0.89}}\ \sqrt{0.7120 \times 0.89}$

$= 0.7960.$

2. Find the standard error of estimate of Y on X and X on Y from the foll/: data

X:	1	2	3	4	5
Y:	2	5	9	13	14.

Sol:

$S_y = \sqrt{1-r^2}\,.\,\sigma_y.$

$S_x = \sqrt{1-r^2}\,.\,\sigma_x.$

X	Y	X^2	Y^2	XY
1	2	1	4	2
2	5	4	25	10
3	9	9	81	27
4	13	16	169	52
5	14	25	196	70
15	43	55	475	161

$$r = \frac{E(xy) - E(x)\cdot E(y)}{\sqrt{E(x^2) - E(x)^2}\ \sqrt{E(y^2) - E(y)^2}}$$

$$E(xy) = \frac{1}{n}\Sigma xy = \frac{161}{5} = 32.2$$

$$E(x^2) = \frac{1}{n}\Sigma x^2 = \frac{55}{5} = 11$$

$$E(y^2) = \frac{1}{n}\Sigma y^2 = \frac{475}{5} = 95$$

$$E(x) = \frac{1}{n}\Sigma x = \frac{15}{5} = 3$$

$$E(y) = \frac{1}{n}\Sigma y = \frac{43}{5} = 8.6$$

$$r = \frac{32.2 - 3 \times 8.6}{\sqrt{11 - 3^2}\ \sqrt{95 - 8.6^2}} = 0.9866$$

$$\sigma_x = \sqrt{11 - 9} = 1.4142$$

$$\sigma_y = \sqrt{95 - 8.6^2} = 4.5869$$

$$\therefore\ S_y = \sqrt{1 - 0.9866^2} \times 4.5869 = 0.7484$$

$$S_x = \sqrt{1 - 0.9866^2} \times 1.4142 = 0.2307$$

2) Obtain the equations of the lines of regression from the following data:

X	1	2	3	4	5	6	7
Y	9	8	10	12	11	13	14

Solution:-

Regression equations of X

$$(x-\bar{x}) = b_{xy}(y-\bar{y})$$

$$\therefore b_{xy} = r\frac{\sigma x}{\sigma y} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2}$$

Regression equation of y

$$\therefore b_{yx} = r\frac{\sigma y}{\sigma x} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$

$$r = \pm\sqrt{b_{xy} \cdot b_{yx}}$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{28}{7} = 4$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{77}{7} = 11$$

x	y	$x-\bar{x}$ $(x-4)$	$(x-\bar{x})^2$	$(y-\bar{y})$ $(y-11)$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
1	9	-3	9	-2	4	6
2	8	-2	4	-3	9	6
3	10	-1	1	-1	1	1
4	12	0	0	1	1	0
5	11	1	1	0	0	0
6	13	2	4	2	4	4
7	14	3	9	3	9	9
28	77	0	28	0	28	26

$$bxy = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2} = \frac{26}{28} = 0.92$$

$$byx = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \frac{26}{28} = 0.92$$

Regression equation of x

$$(x-\bar{x}) = bxy\,(y-\bar{y})$$

$$(x-28) = 0.92\,(y-11)$$

$$x = 0.92y - 10.12 + 28$$

$$\boxed{x = 0.92y + 17.88}$$

Regression equation of y

$$(y - \bar{y}) = b_{yx}(x - \bar{x})$$

$$(y - 11) = 0.92(x - 4)$$

$$y = 0.92x - 3.68 + 11$$

$$\boxed{y = 0.92x + 7.32}$$

4. A Certain Product's demand over 10 months is studied and recorded estimate the equation for this demand as a function of time and compute the co-efficient of determination. (r^2)

t	1	2	3	4	5	6	7	8	9	10
Demand	20	20	30	27	35	36	40	38	45	49

Sollution:

$$r = \frac{N\Sigma dxdy - \Sigma dx \, \Sigma dy}{\sqrt{N\Sigma dx^2 - (\Sigma dx)^2}\,\sqrt{N\Sigma dy^2 - (\Sigma dy)^2}}$$

x	y	$dx = x - A$, $A = 5$	dx^2	$dy = y - A$, $A = 27$	dy^2	$dxdy$
1	20	-4	16	-7	49	28
2	20	-3	9	-7	49	21
3	30	-2	$\cancel{8}$4	3	9	-6
4	27	-1	1	0	0	0
5	35	0	0	8	64	0
6	36	1	1	9	81	9
7	40	2	4	13	169	26
8	38	3	9	11	121	33
9	45	4	16	18	324	72
10	49	5	25	22	484	110
		5	8$\cancel{4}$5	70	1350	$\cancel{305}$ 293

$$r = \frac{N \Sigma dxdy - \Sigma dx \Sigma dy}{\sqrt{N \Sigma dx^2 - (\Sigma dx)^2} \sqrt{N \Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{10 \times 293 - 5 \times 70}{\sqrt{10 \times 85 - (5)^2} \sqrt{10 \times 1350 - (70)^2}}$$

$$= \frac{2580}{\sqrt{825} \sqrt{8600}} = \frac{2500}{28.72 \times 92.73}$$

$$\cancel{= \frac{2700}{29.06 \times 92.73}}$$

$$\cancel{= 1.001}$$

$$\boxed{r = 0.9386}$$

$$r^2 = 0.9385$$

1. If $r = -0.8$ and $N = 36$ calculate

a) Standard Error

b) Probable Error

c) Limit of population correlation. Also state whether the value of r is significant.

Sol:

Given: $r = -0.8$; $N = 36$.

a) $S.E = \frac{1-r^2}{\sqrt{N}} = \frac{1-0.64}{6} = 0.06$.

b) $P.E = 0.6745 \times S.E$

$= 0.6745 \times 0.06 = 0.04$

c) Limits of population correlation

$= r \pm P.E$

$= -0.8 \pm 0.04$

$= (-0.84, -0.76)$

d) Ration of r to $P.E = \frac{|r|}{P.E}$

$= \frac{0.8}{0.04} = 20$ times.

Since r is more than 6 times of P.E

(ie) $|r| > 6 \times P.E$

$\Rightarrow r$ is significant

CamScanner