



**BITS Pilani**  
Pilani|Dubai|Goa|Hyderabad

# BSDCHZC355

## Statistical Inference & Applications

Shaibal Kr. Sen  
Session 01



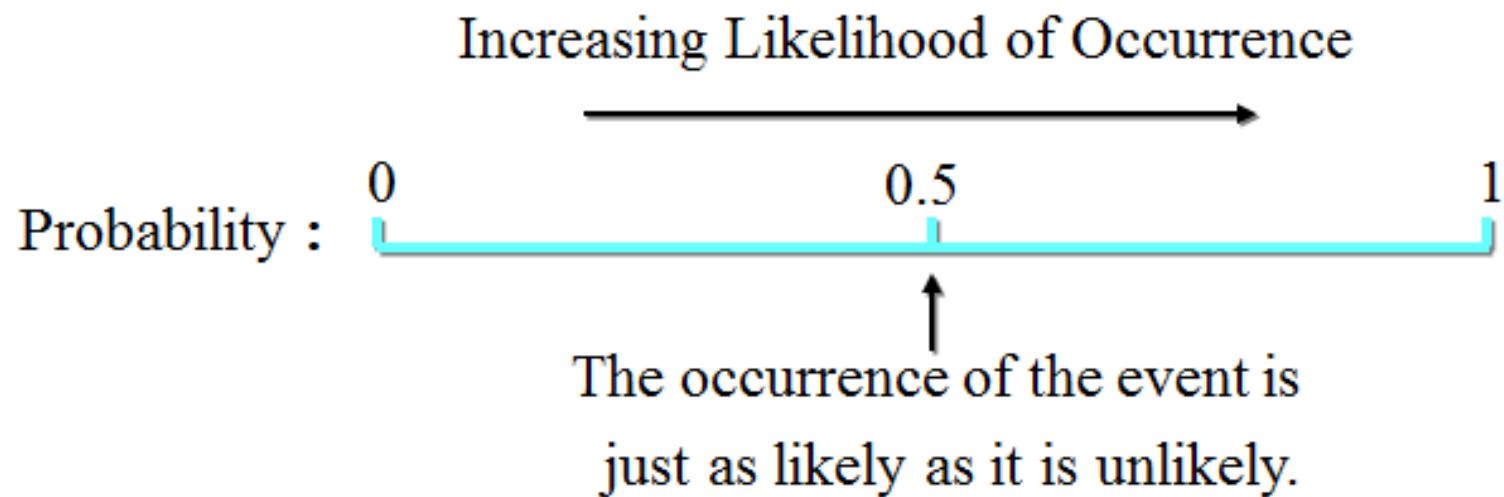
# **STATISTICAL INFERENCES & APPLICATIONS**

## Scheduled Topics to be covered

**Review of Probability & Statistics**

**Probability Distribution - Z, t , F distributions**

## Probability as a Numerical Measure of the Likelihood of Occurrence (Uncertainty).



Classical Probability : Probability is a numerical measure of the likelihood that an event will occur.

Probability of an Event = No. of Favourable outcomes / Total No. of Possible outcomes

---

Empirical Probability : Empirical or Relative Frequency is the 2<sup>nd</sup> type of objective probability. It is based on the number of times an event occurs as a proportion of a known no. of trials.

Mathematically,

$$\text{Empirical Probability} = \frac{\text{No. of times the Event occurs}}{\text{Total no. of observations}}$$

The empirical approach to probability is based on what is called the law of large numbers. The key to establishing probabilities empirically is that more observations will provide a more accurate estimate of the probability.

Law of large nos. : Over a large no. of trials, the empirical probability of an event will approach its true probability.

---

## Axioms of Probability

Probability is a number that is assigned to each member of a collection of events from a random experiment that satisfies the following properties :

If  $S$  is the sample space and  $E$  is any event in a random experiment,

$$(1) P(S) = 1$$

$$(2) 0 \leq P(E) \leq 1$$

(3) For two mutually exclusive events events  $E_1$  and  $E_2$ ;  $E_1 \cap E_2 = \emptyset$  (Null set)

$$\text{and } P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) = P(E_1) + P(E_2)$$

---

Ex 1 : In a certain residential hub, 60% of all households get internet service from the local cable company, 80% get the television service from that company, and 50% get both services from that company.

If a household is randomly selected, what is the probability that it gets at least one of these two services from the company, and what is the probability that it gets exactly one of these services from the company?

Soln : Let household getting Internet services be A, getting TV services be B and getting both TV & Internet services will be  $A \cap B$ . Then as per given data –

$$P(A) = 0.6, \quad P(B) = 0.8 \quad \text{and} \quad P(A \cap B) = 0.5.$$

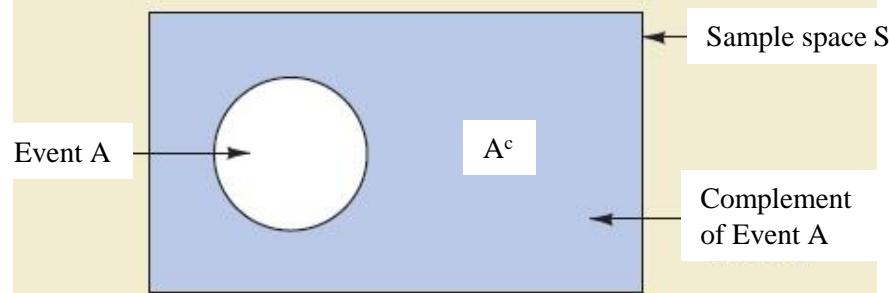
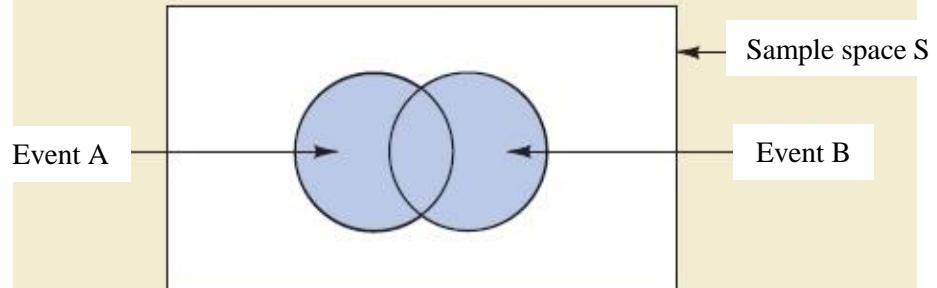
Required to find  $P(\text{exactly one of the services})$

$$\text{We know, } P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.6 + 0.8 - 0.5 = 0.9$$

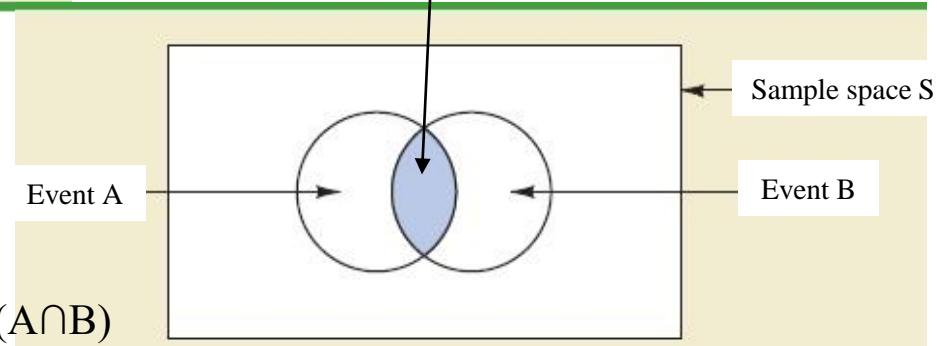
$$\text{so, } P(\text{exactly one of the services}) = P(A \cup B) - P(A \cap B) = 0.9 - 0.5 = 0.4$$

Refer the Venn diagram in next slide

$P(A) = 1 - P(A^c)$ . It leads to  $P(A^c) = 1 - P(A)$



$$P(A \cap B)$$



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

$$P(\text{exactly one of the services}) = P(A \cup B) - P(A \cap B)$$

## Conditional Probability

Definition : For any two events A and B with  $P(B) > 0$ , the conditional probability of A given that B has occurred is defined by –

$$P(A|B) = P(A \cap B) / P(B)$$

The definition of conditional probability yields the following result, obtained by multiplying both sides of equation by  $P(B)$ .

Multiplication Rule :  $P(A \cap B) = P(A|B) \cdot P(B)$

This rule is important because it is often the case that  $P(A \cap B)$  is desired, whereas both  $P(B)$  and  $P(A|B)$  can be specified from the problem description.

## Independent Events

---

If A and B are independent events then

$$P(A|B) = P(A \cap B) / P(B) = P(A) \cdot P(B) / P(B) = P(A)$$

and

$$P(B|A) = P(A \cap B) / P(A) = P(A) \cdot P(B) / P(A) = P(B)$$

---

---

Ex 2 : Suppose a card is drawn from a pack of 52 cards. The probability of it being a card of diamond (D) is  $13/52 = 1/4$ . Now, suppose that it is told that the card drawn is a red card (R), would the probability of it being a D card be  $1/4$  as before or will it be different given the additional information that the card drawn is R.

Soln :  $P(D|R) = P(D \cap R) / P(R) = (13/52) / (26/52) = (1/4) / (1/2) = 1/2$

Here, Diamond (D) is a subset of Red (R)  
ie.  $D \subset R$

$$n(D) = 13, n(R) = 26, n(D \cap R) = 13$$

---

## Bayes' Theorem

Let  $A_1, A_2, \dots, A_k$  be a collection of  $k$  mutually exclusive and exhaustive events with prior probabilities  $P(A_i)$  where  $i = 1, 2, \dots, k$ . Then for any other event  $B$  for which  $P(B) > 0$ , the posterior probability of  $A_j$  given that  $B$  has occurred is –

$$P(A_j|B) = P(A_j \cap B) / P(B) = P(B|A_j) \cdot P(A_j) / \sum_{i=1}^k P(B|A_i) \cdot P(A_i)$$

$$j = 1, 2, \dots, k$$

---

Ex 3 : Three persons A,B and C are competing for the post of CEO of a company. The chances of they becoming CEO are 0.2, 0.3 and 0.5 respectively. The chances of they taking employees beneficial decisions are 0.50, 0.45 and 0.6 respectively.

What are the chances of having employees beneficial decisions after having new CEO.

If employees beneficial decision has been taken, what is the probability that it has been taken by A as new CEO. Similarly, B and C as new CEOs.

Soln :

$$\begin{aligned} P(\text{Ben}) &= 0.2 \times 0.5 + 0.3 \times 0.45 + 0.5 \times 0.6 \\ &= 0.10 + 0.135 + 0.30 = 0.535 \end{aligned}$$

---

$$P(A|Ben) = P(A).P(Ben|A) / P(Ben) = 0.2 \times 0.5 / 0.535 = 0.1869$$

$$P(B|Ben) = P(B).P(Ben|B) / P(Ben) = 0.3 \times 0.45 / 0.535 = 0.2523$$

$$P(C|Ben) = P(C).P(Ben|C) / P(Ben) = 0.5 \times 0.6 / 0.535 = 0.5607$$

## Statistics

---

Statistics may be defined as science that is employed to –

- Collect the data
- Organize and Present the data in a systematic manner
- Analyse the data
- Infer about the data
- Take decision from the data.

Statistics may be defined as numerical data with a view to analyse it.

## Main Branches of Statistics

Descriptive Statistics : Used to study the patterns of past and present data. Classification, tabulation, visualization, central tendency and dispersion measures etc. are some of the important tools of Descriptive Statistics.

Inferential Statistics : Used to estimate and predict the behavior of Big Data (Population) with the help of Small Data (Sample). Sampling, Estimation, Testing of Hypotheses are some of the important tools of Inferential Statistics.

## Mean (Important Measure of Central Tendency)

Also referred as the “**Arithmetic Average**”.

The most commonly used measure of the Center of Data.

Numbers that describe what is average or typical of the distribution.

Computation of Sample Mean :

$\bar{Y} = \Sigma Y / N = (Y_1 + Y_2 + Y_3 + \dots + Y_n) / N$  where “Y bar” equals the sum of all the scores, Y, divided by the number of scores, N.

Computation of the Mean for grouped Data

$\bar{Y} = \Sigma f_i y_i / N$ , where  $f_i y_i$  = a score multiplied by its frequency,  
 $N = \Sigma f_i$ .

## The Standard Deviation (Important measure of Dispersion)

Most common and most important measure of variability is the standard deviation

- A measure of the standard, or average, distance from the mean
- Describes whether the scores are clustered closely around the mean or are widely scattered.

Calculation differs for population and samples.

Variance is a necessary *companion concept* to standard deviation but *not the same* concept.

For Population Variance is  $\sigma^2 = \sqrt{\{(x_i - \bar{x})^2\}} / N$ , Standard Deviation  $\sigma = \sqrt{(\text{Variance})} = \sqrt{[\{\sum(x_i - \bar{x})^2\} / N]}$ .

For Sample Variance is  $S^2 = \{\sum(x_i - \bar{x})^2\} / (n - 1)$  and Standard Deviation  $S = \sqrt{[\{\sum(x_i - \bar{x})^2\} / (n - 1)]}$ .

---

## Ex 4 : Computation of Deviations & Squared Deviations about Mean for the class size Data

No. of students in a Class ( $x_i$ )	Deviation about Mean ( $x_i - \bar{x}$ )	Squared deviation about Mean ( $x_i - \bar{x}$ ) $^2$
46	2	4
54	10	100
42	-2	4
46	2	4
32	-12	144
Total	220	256

$$\text{Mean } \bar{x} = 220 / 5 = 44$$

$$S^2 = \sum(x_i - \bar{x})^2 / (n - 1) = 256 / (5 - 1) = 64$$

$$S = \sqrt{[\sum(x_i - \bar{x})^2 / (n - 1)]} = \sqrt{64} = 8 \text{ (Positive square root)}$$

## Random Variable

---

Random Variable (RV) : A random variable is a real valued function which is a mapping from the sample space  $\Omega$  to the set of real numbers, ie.,  $X : \Omega \rightarrow \mathbb{R}$ . There are two types of RV viz.  
– (i) Discrete RV and (ii) Continuous RV

- (i) Discrete RV take on countable numbers (may be finite or countable infinite values) i.e., without decimal like Natural numbers, Whole numbers, Integers etc.
- (ii) Continuous RV take on any values in an interval i.e., in the set of real numbers which includes, negative, positive, rational, irrational, decimal etc.

## Probability Distributions based on RVs

Two types of Probability distributions are –

(1) Discrete Probability Distribution : A probability distribution based on discrete RV is called discrete probability distribution.  $p(X)$  satisfies -

(i)  $0 \leq p(x) \leq 1$ , for all  $x$

(ii)  $\sum_{\text{All } x_i} p(x) = 1$

(2) Continuous Probability Distribution : A probability distribution based on continuous RV is called continuous probability distribution. An  $f(x)$  is called a probability density function of a continuous random variable if it satisfy the following properties –

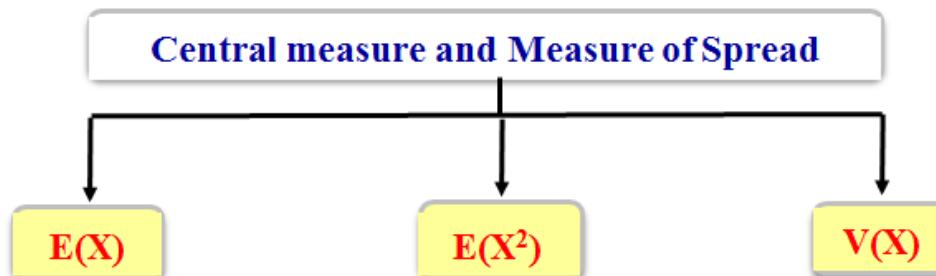
(i)  $f(x) \geq 0$ ,  $x \in (a, b)$

(ii)  $\int_a^b f(x) dx = 1$

(iii)  $P(a \leq X \leq b) = \int_a^b f(x) dx$

## Expected value and Variance (Probability Distribution)

Like mean and standard deviation are computed to describe data measured by quantitative variable, a similar measures viz., expected value (mean) and variance for random variable  $X$  are computed for describing the probability distribution using the formula



For a discrete random variable  $X$  with probability mass function  $p(x)$ ,  
 Mean or Expected value  $= \mu = E(X) = \sum_{\text{over all } x} xp(x)$

$$E(X^2) = \sum_{\text{over all } x} x^2 p(x)$$

$$\text{Variance } = V(x) = \sigma^2 = E[(x - \mu)^2] = \sum (x - \mu)^2 p(x) = E(x^2) - (E(x))^2$$

$$\text{Standard deviation } \sigma = \sqrt{V(x)}$$

Ex 5 : For the given values of  $p(x)$  for various values of  $x$  find the Expected value, Variance and  $p(x)$  for  $x < 2$  and  $x > 4$ .

$$p(x) = \begin{cases} 0, & x < 0 \\ 0.06, & x = 0 \\ 0.13, & x = 1 \\ 0.20, & x = 2 \\ 0.28, & x = 3 \\ 0.25, & x = 4 \\ 0.05, & x = 5 \\ 0.03, & x = 6 \end{cases}$$

Soln : Mean  $\mu = E(X) = 0.06 \times 0 + 0.13 \times 1 + 0.2 \times 2 + 0.28 \times 3 + 0.25 \times 4 + 0.05 \times 5 + 0.03 \times 6$   
 $= 0.13 + 0.4 + 0.84 + 1 + 0.25 + 0.18 = 2.8$

$E(X^2) = 0.06 \times 0^2 + 0.13 \times 1^2 + 0.2 \times 2^2 + 0.28 \times 3^2 + 0.25 \times 4^2 + 0.05 \times 5^2 + 0.03 \times 6^2$   
 $= 0.13 + 0.8 + 2.52 + 4 + 1.25 + 1.08 = 9.78$

$$V(X) = E(X^2) - (E(X))^2 = 9.78 - 2.8^2 = 1.94$$

$$SD(X) = \sqrt{1.94} = 1.392$$

$$\text{Now, } P(X < 2) = P(x = 1) + P(x = 0) = 0.13 + 0.06 = 0.19$$

$$P(X > 4) = P(x = 6) + P(x = 5) = 0.03 + 0.05 = 0.08$$

As in case of discrete probability distribution, the expected value  $E(X)$  and variance  $V(X)$  can be computed for the continuous probability distribution –

$$E(X) = \int_{-\infty}^{\infty} xf(x) dx$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx$$

$$V(X) = E(X^2) - (E(X))^2$$

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

Ex 6 : Let  $x$  be a random variable with pdf given by –

$$f(x) = \begin{cases} cx^2 & \text{for } -1 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

- (i) Find constant ‘c’
- (ii) Find  $E(x)$
- (iii) Find  $V(x)$
- (iv) Find  $P(x \geq 1/2)$

Soln : (i) We know Total Probability  $= \int_{-\infty}^{\infty} f(x) dx = 1$  and as per given data

$$c \int_{-1}^1 cx^2 dx = 1. \quad \text{Hence, } c (x^3 / 3) \Big|_{-1}^1 = 1 \text{ or, } c \{1 / 3 - (-1)^3 / 3\} = 1$$

$$\text{or, } c (1 / 3 + 1 / 3) = 1 \quad \text{ie. } c = 3 / 2 = 1.5$$

$$\text{Therefore, } f(x) = cx^2 = 1.5x^2 \quad \text{for } -1 \leq x \leq 1$$

$$\begin{aligned} \text{(ii) } \mu = E(X) &= \int_{-1}^1 xf(x) dx = \int_{-1}^1 x(1.5x^2) dx = 1.5 \int_{-1}^1 x^3 dx = 1.5[x^4 / 4] \Big|_{-1}^1 \\ &= 1.5(1 / 4 - 1 / 4) = 0 \end{aligned}$$

$$\text{(iii) } V(X) = E(X^2) - (E(X))^2$$

$$E(X^2) = \int_{-1}^1 x^2 f(x) dx = \int_{-1}^1 x^2 (1.5x^2) dx = 1.5(x^5 / 5) \Big|_{-1}^1 = 1.5(1 / 5 + 1 / 5) = 0.6$$

$$\text{So, } V(X) = 0.6 - 0^2 = 0.6$$

$$\text{(iv) } P(x \geq 1 / 2) = \int_{1/2}^1 f(x) dx = \int_{1/2}^1 (1.5x^2) dx = 1.5[x^3 / 3] \Big|_{1/2}^1 = 1.5(0.875) / 3 = 0.4375$$

## Binomial Distribution

A random variable ‘X’ is said to have Binomial distribution if its probability mass function is given by

$$P(x) = \begin{cases} {}^n C_x p^x q^{n-x}, & x = 0, 1, 2, 3, \dots, n \\ 0, & \text{elsewhere} \end{cases}$$

Binomial distribution is a discrete probability distribution.

Binomial distribution will be applied under the following experimental conditions.

- 1) The number of trials (n) is finite.
- 2) The trials are independent of each other.
- 3) The probability of success p is constant for each trial.
- 4) Each trial results in two mutually exclusive events known as success and failure.

Ex 6 : The probability that a man aged 60 years will remain alive till 70 is 0.65. Find – (i) the probability that out of 10 such men at least 7 would be alive at 70 and (ii) Exactly 4 out of 10 would be alive till 70

Soln : We know,  $P(x) = {}^nC_x p^x q^{n-x}$ ,  $x = 0, 1, 2, 3, \dots, n$

where,  $p = 0.65$ ; so,  $q = 1 - 0.65 = 0.35$ ;  $n = 10$

(i) Required to find  $P(x \geq 7) = P(x = 7) + P(x = 8) + P(x = 9) + P(x = 10)$

$$= {}^{10}C_7 (0.65)^7 (0.35)^{10-7} + {}^{10}C_8 (0.65)^8 (0.35)^{10-8} + {}^{10}C_9 (0.65)^9 (0.35)^{10-9} + {}^{10}C_{10} (0.65)^{10} (0.35)^{10-10}$$

$$= {}^{10}C_7 (0.65)^7 (0.35)^3 + {}^{10}C_8 (0.65)^8 (0.35)^2 + {}^{10}C_9 (0.65)^9 (0.35)^1 + {}^{10}C_{10} (0.65)^{10} (0.35)^0$$

$$= 10! / (7! \times 3!) \{(0.65)^7 (0.35)^3\} + 10! / (8! \times 2!) \{(0.65)^8 (0.35)^2\} + 10! / (9! \times 1!) \{(0.65)^9 (0.35)^1\} + 10! / (10! \times 0!) \{(0.65)^{10} (0.35)^0\} = 0.515$$

$$(ii) P(x = 4) = {}^{10}C_4 p^4 (1-p)^{10-4} = {}^{10}C_4 (0.65)^4 (0.35)^6$$

$$= \{10! / 4!(10-4)!\} (0.65)^4 (0.35)^6 =$$

$$= 10 \times 9 \times 8 \times 7 \times 6! / 4 \times 3 \times 2 \times 6! (0.65)^4 (0.35)^6$$

$$= 210 \{(0.65)^4 (0.35)^6\} = 210 \times 0.1785 \times 0.001838 = 0.06889$$

## Poisson Distribution

A random variable ‘X’ is said to have Poisson distribution if its probability mass function is given by –

$$P(x) = \begin{cases} e^{-\lambda}(\lambda^x / x!), & x = 0, 1, 2, \dots \dots \\ 0, & \text{elsewhere} \end{cases}$$

Poisson distribution is the discrete probability distribution of a discrete random variable X, which has no upper bound. It is defined for non negative values of x.

Poisson distribution is suitable for rare events for which the probability of occurrence ‘p’ is very small and the number of trials ‘n’ is very large.

Also, Binomial distribution can be approximated by Poisson distribution when  $n \rightarrow \infty$  and  $p \rightarrow 0$  such that  $\lambda = np = \text{constant}$ .

## Mean & variance of Poisson Distribution

### For Poisson Distribution

$$\text{Mean} = E(X) = \lambda$$

$$\text{Variance} = V(X) = \lambda$$

$$\text{Mean} = \text{Variance} = \lambda$$

$$\text{SD} = \text{sqrt}(\text{variance}) = \text{sqrt}(\lambda)$$

### For Binomial Distribution

$$\text{Mean} = np$$

$$\text{Variance} = V(X) = npq$$

$$\text{Mean} > \text{Variance}$$

$$\text{SD} = \text{sqrt}(\text{variance}) = \text{sqrt}(npq)$$

---

Ex 7 : If the probability of a bad reaction from a certain injection is 0.001. Determine the chance that out of 2000 individuals more than two will get a bad reaction.

Soln : Given  $n = 2000$ ;  $p = 0.001$

$$\text{Mean } \lambda = np = 2000 \times 0.001 = 2$$

We know,  $e^{-\lambda}(\lambda^x / x!)$ ,  $x = 0, 1, 2, \dots$

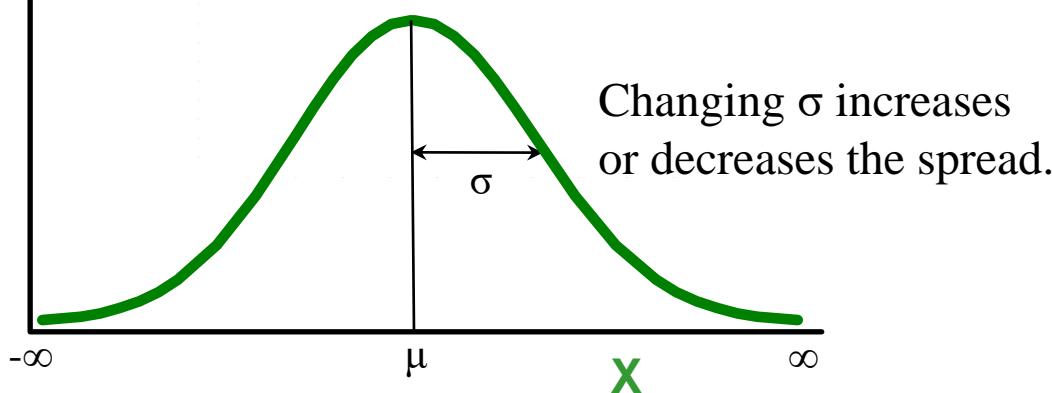
$$\begin{aligned} \text{Now, find } P(x > 2) &= 1 - P(x \leq 2) = 1 - \{P(0) + P(1) + P(2)\} \\ &= 1 - \{e^{-\lambda}(\lambda^0 / 0!) + e^{-\lambda}(\lambda^1 / 1!) + e^{-\lambda}(\lambda^2 / 2!)\} \\ &= 1 - e^{-\lambda} \{1 + \lambda^1 + \lambda^2 / 2\} = 1 - e^{-2} \{1 + 2 + 2^2 / 2\} \\ &= 1 - e^{-2}(5) = 1 - 0.1353 \times 5 = 1 - 0.6765 = 0.3235 \end{aligned}$$

---

**The Normal Distribution** : The Normal Probability Density is one of the special probability densities, usually referred to simply as the Normal Distribution (the words density & distribution are often used interchangeably in the literature of applied statistics). The equation of the normal probability density, whose graph (shaped like the cross section of a bell with different centers & spreads depending on  $\mu$  and  $\sigma$ ).

$f(X)$

Changing  $\mu$  shifts the distribution left or right.



Changing  $\sigma$  increases or decreases the spread.

## Properties:

1. Normal curve is bell shaped and symmetric about the mean.
2. Mean = Mode = Median.
3. Total area under normal curve is equal to 1.
4. Normal curve approaches but never touches the x axis as it extends farther and farther away from the mean.

---

Ex 8 : The average score of students in a statistics test was 78 with a standard deviation of 8. If the test scores are normally distributed, find the probability that a student receives a test score – (i) less than 90, (ii) More than 90 and (iii) between 80 and 96.

Soln : Given  $\mu = 78$  and  $\sigma = 8$

(i) Required to find  $P(x < 90)$ .

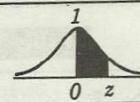
Converting it to Std Normal  $P(Z < 90)$  where,

$$Z = (x - \mu) / \sigma = (90 - 78) / 8 = 1.5 \text{ ie. } P(Z < 1.5) = 0.9332 \text{ (From Z-table)}$$

$$(ii) P(x > 90) = P(Z > 1.5) = 1 - 0.9332 = 0.0668$$

$$(iii) P(80 < x < 96) = P[(80 - 78) / 8 < Z < (96 - 78) / 8] = P(1.5 < Z < 2.25) \\ = 0.9878 - 0.9332 = 0.0546$$

Table V : Area under the Normal curve



<i>z</i>	0	1	2	3	4	5	6	7	8	9
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0754
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2258	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2518	0.2549
0.7	0.2580	0.2612	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2996	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993

Table 3 (continued from page 514)

$z$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5973	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998									
4.0	0.99997									
5.0	0.9999997									

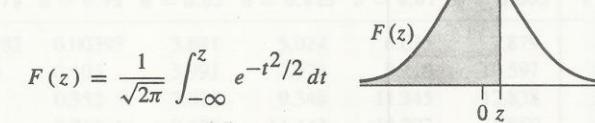
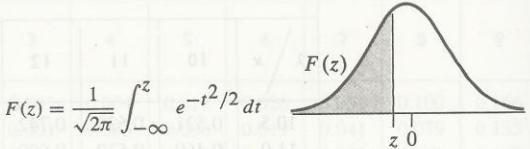


Table 3 Standard Normal Distribution Function

<i>z</i>	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-5.0	0.0000003									
-4.0	0.00003									
-3.5	0.0002									
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0006	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

(continued on following page)



## Sampling

---

Sampling is a method or process of selecting samples from populations.

Data are gathered from samples and conclusions are drawn about the population as a part of the inferential statistics process.

A sample provides a reasonable means for gathering useful decision-making information that might be otherwise unattainable and unaffordable.

## Sampling Distribution-Z distribution

Sampling distribution is a set of collection of all calculated values of a statistic where each value is obtained from each sample of a total possible random samples each of size  $n$  that can be taken from a population. The sampling distribution follows probability distribution.

**Central Limit Theorem (CLT)** : If  $\bar{x}$  is the mean of a sample size  $n$  taken from a population having the mean  $\mu$  and the finite variance  $\sigma^2$ , then  $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$  is a random variable whose distribution approaches to standard normal distribution as  $n \rightarrow \infty$ .

**Note :** Central limit theorem is better applicable if the sample size  $n \geq 25$  but in general for  $n \geq 30$ .

Ex 9 : The time at the counter for a customer to be served at a post office can be modelled as a random variable having mean 176 seconds and variance  $256 \text{ (seconds)}^2$ . The sample mean  $\bar{x}$  will be obtained from the times for a random sample of 100 customers. What is the probability that  $\bar{x}$  will be between 175 and 178 seconds?

Soln : Given  $\mu = 176$ ,  $\sigma^2 = 256$  implies  $\sigma = 16$ ,  $n = 100 (> 30)$   
 Then by using CLT,  $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - 176}{16/\sqrt{100}} = \frac{\bar{x} - 176}{1.6}$  is a standard normal variable.

The probability that  $\bar{x}$  will be between 175 and 178 seconds is

$$P(175 < \bar{x} < 178) = P\left(\frac{175 - 176}{1.6} < Z < \frac{178 - 176}{1.6}\right) =$$

$$P(-0.625 < Z < 1.25) = F(1.25) - F(-0.625) = 0.8944 - 0.2659 \text{ (Taken from the Z - table values of standard normal distribution)} = 0.628$$

## Exact Sampling Distributions (Small Sampling Distributions)

t- Distribution : Let  $(x_1, x_2, \dots, x_n)$  be a small sample drawn at random from a normal population with mean  $\mu$  and un-known variance  $\sigma^2$ .

Let  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  be the sample mean which is the estimator of population mean  $\mu$  and  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$  be the sample variance which is the estimator of population variance  $\sigma^2$ .

Then the distribution of  $\bar{x}$  has the mean  $E(\bar{x}) = \mu$  and variance  $V(\bar{x}) = \frac{\sigma^2}{n} = \frac{S^2}{n}$  (since,  $\sigma^2$  is not known then it is replaced by its estimator  $S^2$ ) and  $SE(\bar{x}) = \sqrt{V(\bar{x})} = S/\sqrt{n}$ .

Then the statistic  $t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$  is a random variable follows Student's t distribution with  $\vartheta(n-1)$  degree of freedom (parameter) where  $-\infty < t < \infty$ .

t distribution is symmetric distribution. It is symmetric at mean 0 and variance greater than unity ( $= \vartheta / (\vartheta - 2)$ ). t distribution behaves similar to standard normal distribution.

(Degrees of freedom  $\vartheta = n - k$  is the difference between 'n' the sample size and k is the number of population parameters which are calculated using the sample data).

The t-distribution curve is symmetric about the mean zero, unimodal, bell shaped and asymptotic on both sides of t-axis.

Ex 10 : A random sample of size 25 from a normal population has the mean

$\bar{x} = 47.5$  and the standard deviation  $S = 8.4$ . Does this information tend to support or refute the claim that the mean of the population  $\mu = 42.1$ ?

Soln : Given  $n = 25$ ,  $\bar{x} = 47.5$ ,  $S = 8.4$ ,  $\mu = 42.1$

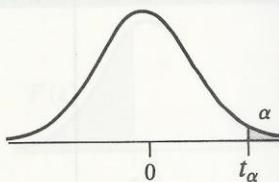
$$\text{Then } t_{(n-1)df} = \frac{\bar{x} - \mu}{S/\sqrt{n}} \Rightarrow t_{24df} = \frac{47.5 - 42.1}{8.4/\sqrt{25}} = 5.4/1.68 = 3.21$$

From Table of t distribution we observe,

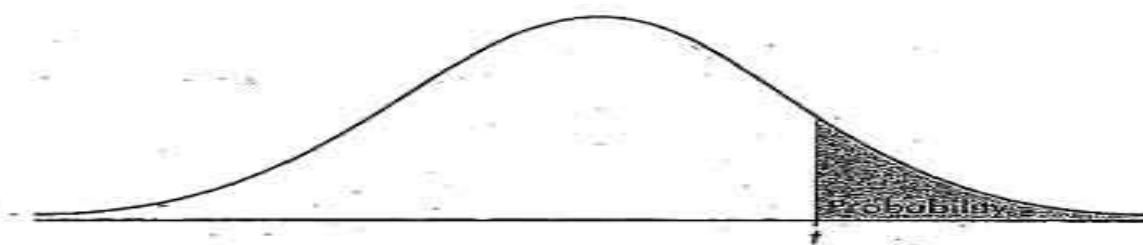
$P(t_{24df} > 2.797) = 0.005$  which is the area of  $t_{24df}$  between 2.797 and  $\infty$ .

Then  $P(t_{24df} > 3.21) < 0.005$  since the area of  $t_{24df}$  between 3.21 and  $\infty$  is less than that of the area between 2.797 and  $\infty$  as it is known that  $2.797 < 3.21 < \infty$ .

Therefore,  $P(t_{24df} > 3.21) < 0.005$  tends to refute (reject) the claim that the mean of the population  $\mu = 42.1$

Table 4 Values of  $t_\alpha$ 

$v$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.00833$	$\alpha = 0.00625$	$\alpha = 0.005$	$v$
1	3.078	6.314	12.706	31.821	38.204	50.923	63.657	1
2	1.886	2.920	4.303	6.965	7.650	8.860	9.925	2
3	1.638	2.353	3.182	4.541	4.857	5.392	5.841	3
4	1.533	2.132	2.776	3.747	3.961	4.315	4.604	4
5	1.476	2.015	2.571	3.365	3.534	3.810	4.032	5
6	1.440	1.943	2.447	3.143	3.288	3.521	3.707	6
7	1.415	1.895	2.365	2.998	3.128	3.335	3.499	7
8	1.397	1.860	2.306	2.896	3.016	3.206	3.355	8
9	1.383	1.833	2.262	2.821	2.934	3.111	3.250	9
10	1.372	1.812	2.228	2.764	2.870	3.038	3.169	10
11	1.363	1.796	2.201	2.718	2.820	2.891	3.106	11
12	1.356	1.782	2.179	2.681	2.780	2.934	3.055	12
13	1.350	1.771	2.160	2.650	2.746	2.896	3.012	13
14	1.345	1.761	2.145	2.624	2.718	2.864	2.977	14
15	1.341	1.753	2.131	2.602	2.694	2.837	2.947	15
16	1.337	1.746	2.120	2.583	2.673	2.813	2.921	16
17	1.333	1.740	2.110	2.567	2.655	2.793	2.898	17
18	1.330	1.734	2.101	2.552	2.639	2.775	2.878	18
19	1.328	1.729	2.093	2.539	2.625	2.759	2.861	19
20	1.325	1.725	2.086	2.528	2.613	2.744	2.845	20
21	1.323	1.721	2.080	2.518	2.602	2.732	2.831	21
22	1.321	1.717	2.074	2.508	2.591	2.720	2.819	22
23	1.319	1.714	2.069	2.500	2.582	2.710	2.807	23
24	1.318	1.711	2.064	2.492	2.574	2.700	2.797	24
25	1.316	1.708	2.060	2.485	2.566	2.692	2.787	25
26	1.315	1.706	2.056	2.479	2.559	2.684	2.779	26
27	1.314	1.703	2.052	2.473	2.553	2.676	2.771	27
28	1.313	1.701	2.048	2.467	2.547	2.669	2.763	28
29	1.311	1.699	2.045	2.462	2.541	2.663	2.756	29
inf.	1.282	1.645	1.960	2.326	2.394	2.498	2.576	inf.



**TABLE B:  $t$ -DISTRIBUTION CRITICAL VALUES**

df	Tail probability $p$											
	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
$\infty$	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291

## F distribution

Let  $(x_1, x_2, \dots, x_{n_1})$  be a small sample drawn at random from a normal population with mean  $\mu_1$  (un-known) and variance  $\sigma^2_1$ .

Let  $\bar{x} = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i$  be the sample mean which is the estimator of population mean  $\mu_1$  and  $S^2_1 = \frac{1}{n_1-1} \sum_{i=1}^{n_1} (x_i - \bar{x})^2$  be the sample variance which is the estimator of population variance  $\sigma^2_1$ .

Let  $(y_1, y_2, \dots, y_{n_2})$  be a small sample drawn at random from another independent normal population with mean  $\mu_2$  (un-known) and variance  $\sigma^2_2$ .

Let  $\bar{y} = \frac{1}{n_2} \sum_{i=1}^{n_2} y_i$  be the sample mean which is the estimator of population mean  $\mu_2$  and  $S^2_2 = \frac{1}{n_2-1} \sum_{i=1}^{n_2} (y_i - \bar{y})^2$  be the sample variance which is the estimator of population variance  $\sigma^2_2$ .

Also assume  $\sigma^2_1 = \sigma^2_2 = \sigma^2$  (unknown).

---

A Random variable having F-Distribution – Theorem : If  $S^2_1$  and  $S^2_2$  are the variances of independent random samples of size  $n_1$  and  $n_2$  respectively, taken from two normal populations having the same variance, then -

$$F = S^2_1 / S^2_2$$

is a random variable having F-Distribution with parameters  $v_1 = n_1 - 1$  and  $v_2 = n_2 - 1$ .

F Values for  $\alpha = 0.05$  $d_1$  $d_2$ 

	1	2	3	4	5	6	7	8	9
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5
2	18.51	19.00	19.16	19.25	19.3	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39

9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96
inf	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

---

For F-Distribution  $F = S^2_1 / S^2_2$  with  $[(n_1 - 1), (n_2 - 1)]$  degrees of freedom.

Ex : If two independent random samples of size  $n_1 = 7$  and  $n_2 = 13$  are taken from a normal population. What is the probability that the variance of the first sample will be at least 3 times as large as that of the second sample?

Soln : To find  $P(S^2_1 \geq 3 S^2_2) = P(S^2_1 / S^2_2 \geq 3)$

$= P(F \geq 3) = P(F([(n_1 - 1), (n_2 - 1)]) \geq 3) =$

$P(F(6,12) \geq 3) = 0.05$  observed from F tables.

Values of  $F_{0.05}$ 

F-table

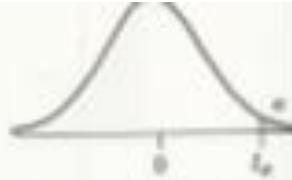
$v_2$ = Degrees of Freedom for Denominator	$v_1$ = Degrees of Freedom for Numerator																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	60	120	$\infty$
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.46	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.63	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.52	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.83	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.40	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.11	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.89	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.73	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.60	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.50	2.47	2.38	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.41	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.34	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.28	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.23	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.18	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.14	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.07	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.02	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.00	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.97	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.88	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.78	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.69	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.60	1.55	1.50	1.43	1.35	1.25
$\infty$	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.51	1.46	1.39	1.32	1.22	1.00

Values of  $F_{0.01}$ 

## F-table

$v_2 = \text{Degrees of Freedom for Denominator}$	$v_1 = \text{Degrees of Freedom for Numerator}$																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	60	120	$\infty$
1	4,052	5,000	5,403	5,625	5,764	5,859	5,928	5,982	6,023	6,056	6,106	6,157	6,209	6,240	6,261	6,287	6,313	6,339	6,366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.57	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.58	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.91	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.45	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.30	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.06	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.26	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.71	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.31	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.01	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.76	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.57	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.41	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.28	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.16	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.07	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	2.98	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.91	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.84	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.79	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.73	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.69	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.64	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.60	2.54	2.45	2.36	2.27	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.45	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.27	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.10	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.93	1.86	1.76	1.66	1.53	1.38
$\infty$	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.77	1.70	1.59	1.47	1.32	1.00

t-table



$n$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.00033$	$\alpha = 0.00025$	$\alpha = 0.0001$	$\alpha$
1	3.078	6.314	12.706	31.821	38.204	50.923	63.657	1
2	1.886	2.903	4.203	6.965	7.650	8.390	9.925	2
3	1.638	2.353	3.182	4.541	4.857	5.392	5.841	3
4	1.533	2.132	2.776	3.747	3.961	4.315	4.804	4
5	1.476	2.015	2.571	3.345	3.534	3.810	4.032	5
6	1.440	1.943	2.447	3.143	3.288	3.523	3.707	6
7	1.415	1.895	2.363	2.998	3.128	3.335	3.499	7
8	1.397	1.860	2.306	2.895	3.016	3.206	3.355	8
9	1.383	1.833	2.262	2.821	2.934	3.111	3.250	9
10	1.372	1.812	2.228	2.764	2.870	3.038	3.169	10
11	1.363	1.796	2.201	2.718	2.820	2.991	3.106	11
12	1.356	1.782	2.179	2.681	2.780	2.954	3.055	12
13	1.350	1.771	2.160	2.650	2.746	2.936	3.032	13
14	1.345	1.761	2.145	2.624	2.718	2.924	2.977	14
15	1.341	1.753	2.131	2.602	2.694	2.897	2.947	15
16	1.337	1.746	2.120	2.583	2.673	2.813	2.921	16
17	1.333	1.740	2.110	2.567	2.655	2.793	2.898	17
18	1.330	1.734	2.101	2.553	2.639	2.773	2.878	18
19	1.328	1.729	2.093	2.539	2.625	2.759	2.861	19
20	1.325	1.725	2.086	2.528	2.613	2.744	2.845	20
21	1.323	1.721	2.080	2.518	2.602	2.732	2.831	21
22	1.321	1.717	2.074	2.508	2.591	2.720	2.819	22
23	1.319	1.714	2.069	2.500	2.582	2.710	2.807	23
24	1.318	1.711	2.064	2.492	2.574	2.700	2.797	24
25	1.316	1.708	2.060	2.485	2.566	2.692	2.787	25
26	1.315	1.706	2.056	2.479	2.559	2.684	2.779	26
27	1.314	1.703	2.053	2.473	2.553	2.676	2.771	27
28	1.313	1.701	2.048	2.467	2.547	2.669	2.763	28
29	1.311	1.699	2.045	2.462	2.541	2.663	2.756	29
30	1.302	1.645	1.980	2.326	2.394	2.498	2.576	30

## STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361

## Practice Problems

Q1. If two fair dice are rolled, what is the probability that sum of the numbers appearing is 9, given that at least one 4 appears?

Q2. Suppose  $X$  has the following probability mass function  $p(0) = 0.2$ ,  $p(1) = 0.5$ ,  $p(2) = 0.3$ . Calculate  $E[X^2]$ .

Q3. A fair die is rolled. If even number appears then a coin is flipped. Draw the probability tree and find the probability of getting -

- (i) An even number & Tail.
- (ii) 6 and Head
- (iii) 3 and Head.

Q4. If the probability is 0.05 that a certain wide-flange column will fail under a given axial load, what are the probabilities that among 16 such columns

- (a) At most two will fail
- (b) At least four will fail

---

Q5. Compute Mean and Variance of  $X$ , when  $X$  represents the outcome when we roll a fair die.

Q6. Average income of a group of employees in a company is Rs. 12,000 with standard deviation of Rs. 1,100. Find the percentage of employees earning –

- i) Between Rs. 9,000 and 11,000
- ii) More than Rs. 13,000
- iii) Less than Rs. 10,000

Q7. Average IQ of a class is 120 with a variance 64. Find -

- i) Minimum score of the 3% students who have a higher IQ.
- ii) Maximum score of the 5% students who have a lower IQ.

---

Q8. In a town on an average 10 accidents occur within a span of 50 days. Assuming the no. of accidents per day follow Poisson distribution, find the probability that there will be 3 or more accidents in a day. (Note :  $e = 2.7183$ )

Q9. Using normal probabilities to determine the mean fill of jars -  
The actual amount of instant coffee that a filling machine puts into “4-ounce” jars may be looked upon as a random variable having a normal distribution with  $\sigma = 0.04$  ounce. If only 2 % of the jars are to contain less than 4 ounces, what should be the mean fill of these jars ?

Q10. Suppose the probability that an item produced by a certain machine will be defective is 0.1. Find the probability that a sample of 10 items will contain at most one defective item. Assume that the quality of successive items is independent.



**BITS** Pilani  
Pilani | Dubai | Goa | Hyderabad



# BSDCHZC355 Statistical Inference & Applications

Shaibal Kr. Sen  
Session 02



# STATISTICAL INFERENCES & APPLICATIONS

## Scheduled Topics to be covered

Review of Testing of Hypothesis

Parametric Tests –

- (i) Test of mean for single population using Z-test and t-test
- (ii) Testing the difference between two means using Z-test
- (iii) Testing the difference between two means using t-test
- (iv) Testing of hypothesis for proportion of two populations

Concept of Non-Parametric Tests

Supposed or Proposed explanation made on the basis of limited evidence is a starting point for further investigation. In the context of Probability, Hypothesis is a statement or claim about the population parameter. Correctness of which is subject to test prior to accepting / rejecting – Accept if the test report of the claim is favourable and reject if the test report of the claim is unfavourable. Explained with relevant examples during the session.

### **Null and Alternate Hypothesis**

If a certain claim is made about a population parameter then it is compared against another hypothesis, which is accepted when original hypothesis is rejected.

Generally the hypothesis which we wish to establish is taken as alternate hypothesis and denoted by  $H_1$ . The hypothesis to be contested is called null hypothesis, denoted by  $H_0$ .

---

Type I error or  $\alpha$  error : There is a chance that null hypothesis is rejected when it is true, that is, we have committed type I error.

Probability of Type I error is  $\alpha = P(H_0 \text{ is rejected} \mid H_0 \text{ is true})$ .  
This is also called level of significance.

Type II error or  $\beta$  error : There is a chance that null hypothesis is accepted when it is false, that is, we have committed type II error.

Probability of Type II error is  $\beta = P(H_0 \text{ is accepted} \mid H_0 \text{ is false})$ .

## Types of Tests

---

A statistical rule which decides whether to accept or reject the null hypothesis on the basis of data. Following Tests are commonly employed –

Parametric Tests : Based on the assumption of some probability distribution.

Non-Parametric Tests : Not based on any assumption of probability distribution. It is assumed that the data do not follow any probability distribution which is not characterized by any parameters - Distribution - free tests – Chi-Square Test.

## Parametric tests

It is assumed that the data do follow some probability distribution which is characterized by any parameters.

**Large Sample Test**

$n \geq 30$

**Standard Normal Test**

Z-Test

**Small Sample Test**

$n < 30$

**Student's t-test**

Unpaired t-Test

Paired t-Test

## Parametric tests

Z-test



This is a test based on Standard Normal Distribution

Used for testing the

Mean of a single population ( $\mu$ )

Difference between means of two populations ( $\mu_1 - \mu_2$ )

Proportion of a single population (P)

Difference between proportions of two populations ( $P_1 - P_2$ )

## One sided and both sided intervals

### One Sided

1a) If  $H_0 : \mu \leq \mu_0$

$H_1 : \mu > \mu_0$  (ie. it is greater than  $\mu_0$ )

Lower tail is Acceptance region & Upper tail is Rejection region

1b) If  $H_0 : \mu \geq \mu_0$

$H_1 : \mu < \mu_0$  (ie. it is lesser than  $\mu_0$ )

Upper tail is Acceptance region & Lower tail is Rejection region

### Both Sided

2) If  $H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$  (ie. it is either greater or lesser than  $\mu_0$ )

Central Region (CR) is Acceptance Region and both ends beyond CR is Rejection Region

## Test of Mean for single population

Ex 1 : It is claimed that sports-car owners drive on the average 18,580 kms per year. A consumer firm believes that the average mileage is not same as the claim. To check, the consumer firm obtained information from randomly selected 10 sports-car owners that resulted in a sample mean of 17,352 kms with a sample standard deviation of 2,012 kms. What can be concluded about this claim at 5% level of significance (LOS)?

Soln : Given  $\mu_0 = 18,580$  kms,  $s = 2,012$  kms, and  $n = 10$ ,  $\bar{x} = 17,352$  kms

Here,  $H_0 : \mu = \mu_0 = 18,580$  and  $H_1 : \mu \neq \mu_0 \neq 18,580$

$$\begin{aligned} \text{It is a parametric test, } t &= \sqrt{n} (\bar{x} - \mu_0) / (s) = (17,352 - 18,580) / (2012/\sqrt{10}) \\ &= (-1228) / 636.25 = -1.93 \end{aligned}$$

Critical value for  $\alpha = 0.05$  and for two tailed test for  $\alpha/2 = 0.025$  is 2.262 for  $n - 1 = 9$  degree of freedom (obtained from t-table)

Since  $|t| = 1.929 < 2.262$ , Hypothesis is accepted ie. sports car owner's claim is accepted.

Ex 2 : Average Height of 10 males of a given locality is found to be 66 inches & variance is 10 in<sup>2</sup>. Is it reasonable to believe that the average height is greater than 64 inches? Test at 5% level of significance. (Obtained from t table, t at 9 d.f. at 5% level of significance is 1.833).

Soln : Given;  $n = 10$ ,  $\mu_0 = 64$ ,  $\alpha = 0.05$ ,  $\bar{x} = 66$ ,  $s = \sqrt{10}$ ,  $t_{0.05, 9} = 1.833$

Null hypothesis should be  $H_0 : \mu \geq 64$  and alternative hypothesis  $H_1 : \mu < 64$

$$\text{Test statistic } t = \frac{(\bar{x} - \mu_0)\sqrt{n}}{s} = \frac{(66 - 64)\sqrt{10}}{\sqrt{10}} = 2$$

Now, as the critical value of  $|t_{0.05, 9}|$  is 1.833, the acceptance region is  $t = -1.833$  to  $\infty$  and rejection region is  $-\infty$  to  $-1.833$  as shown in the diagram (in next slide). As the test statistic (2) is within the acceptance region, the hypothesis that average height is greater than 64 inches is accepted.

**Ex 3 :** A sample of 40 sales receipts from a grocery store has average \$137 and  $\sigma = \$30.2$ . Use these values to test whether or not the mean of sales at the grocery store are different from \$150.

**Soln :** Step 1 : Set the null and alternative hypothesis

$$H_0 : \mu = 150$$

$$H_1 : \mu \neq 150$$

Step 2 : Calculate the test statistic

$$Z = \sqrt{n} ( \bar{x} - \mu_0 ) / \sigma = (137 - 150) / 30.2 / \sqrt{40} = -2.722$$

Step 3 : Set Rejection Region

Looking at the the picture below, we need to put half of  $\alpha$  in the left tail and the other half of  $\alpha$  in the right tail . Thus  $R : |Z| > 2.575$  (assumed  $\alpha = 1\%$ )

*Make the normal graph showing both regions beyond 2.575 (0.005) – as shown above*

Step 4 : Conclude

We can see that  $|-2.722| = 2.722 > 2.575$ , thus our test statistic is in the rejection region. Therefore we reject the null hypothesis in favor of the alternative. We can conclude that the mean is significantly different from \$150, thus it is proven that the mean sales at the grocery store is not \$150.

**Ex 4 :** A sample of 35 sales receipts of a grocery store reveals average sale of Rs. 10000 at  $\sigma$  = Rs. 300. Use these values to test whether or not the mean sale at the grocery store is different from Rs. 11000. Take level of significance (LOS) as 1%. [Given  $Z_{\text{critical}} = 2.575$ ]

**Soln :** Step 1 : Set the null and alternative hypothesis

$$H_0 : \mu = 11000$$

$$H_1 : \mu \neq 11000$$

Step 2 : Calculate the test statistic

$$Z = \sqrt{n} (\underline{x} - \mu_0) / \sigma = (10000 - 11000) / 300 / \sqrt{35} = -1000 / 50.71 = 19.73$$

Step 3 : Set Rejection Region

Looking at the the picture below, we need to put half of  $\alpha$  in the left tail and the other half of  $\alpha$  in the right tail . Thus  $R : |Z| > 2.575$  (assumed  $\alpha = 1\%$ )

*Make the normal graph showing both regions beyond 2.575 (0.005) – as shown above*

Step 4 : Conclude

We can see that  $|-2.722| = 2.722 > 2.575$ , thus our test statistic is in the rejection region. Therefore we reject the null hypothesis in favor of the alternative. We can conclude that the mean is significantly different from \$150, thus it is proven that the mean sales at the grocery store is not \$150.

**Ex 5 :** A company mfrs steel bars. If the production process is working properly, it turns out steel bars with an average length of at least 2.8 mts with a std deviation of 0.2 mts (as determined from Engg specs of the prodn eqpt involved). Longer steel bars can be used or altered, but the shorter bars must be scrapped. A sample of 25 bars is selected from the prodn line. The sample indicates an average length of 2.73 mts. The company wishes to test the hypothesis at the 0.05 level of significance, what decision would it make using the critical value approach to hypothesis testing ?

**Soln :** Step 1 : Set the null and alternative hypothesis

$$H_0 : \mu \geq 2.8$$

$$H_1 : \mu < 2.8$$

Step 2 : Calculate the test statistic

$$\begin{aligned} t &= \sqrt{n} (x - \mu) / \sigma = \sqrt{25} (2.73 - 2.8) / 0.2 \\ &= -1.75 \end{aligned}$$

Step 3 : Set Rejection Region - At  $\alpha = .05$  reject  $H_0$  if  $t_{0.05, 24} (-1.711) > -1.75$

Step 4 : Conclude - We can see that  $-1.75 < -1.711$ , thus our test statistic is in the rejection region (at  $\alpha = 0.05$ ). Therefore, we reject the null hypothesis. We can conclude that average length of steel bars are less than 2.8 mts.

---

**Ex 6 :** A car manufacturing company claims that their cars do not require servicing before running for on an average 10,000 miles. To verify the claim, the manufacturer tests 20 cars under simulated conditions and gets a mean breakdown of 9500 miles with a standard deviation of 1150 miles. What conclusion can be drawn at a significance level of 5%.

**Soln :** Given :  $\bar{x} = 9500$ ,  $\mu = 10000$ ,  $n = 20$ ,  $s = 1150$

$$H_0 : \mu \geq \mu \text{ and } H_1 : \mu < \mu$$

Test statistic  $t = \sqrt{n} (\bar{x} - \mu) / s$   
 $= (9500 - 10000) / (1150/\sqrt{20}) = -500 / 257.15 = -1.94$

From t-table we find,  $t_{0.05, df = 19} = -1.729$  .

Since the test statistic value is beyond the acceptance region, the Hypothesis can not be accepted. In other words the manufacturer's claim is rejected.

## Testing the difference between two means using Z-test

For testing the significant difference between two population means using large samples ( $n \geq 30$ ), the test statistic is -

$$Z = (\bar{x}_1 - \bar{x}_2) / [\sqrt{(\sigma_1)^2/n_1 + (\sigma_2)^2/n_2}] = (\bar{x}_1 - \bar{x}_2) / [\sqrt{(s_1)^2/n_1 + (s_2)^2/n_2}]$$

and the significant value is  $Z_\alpha$  for one tailed test and  $Z_{\alpha/2}$  for two tailed test.

$\mu_1$  and  $\mu_2$  are two population means respectively.

$(\sigma_1)^2$  and  $(\sigma_2)^2$  are two population variances respectively and if they are not known they are replaced by their respective sample variances  $(s_1)^2$  and  $(s_2)^2$ .

$n_1$  and  $n_2$  are two sample sizes respectively (each  $\geq 30$ )

Here,  $H_0 : \mu_1 = \mu_2$  and  $H_1 : \mu_1 \neq \mu_2$

If the calculated  $|Z| \leq$  significant value (obtained from Z-table), then  $H_0$  is accepted; otherwise  $H_0$  is rejected.

Ex 7 : In a survey of buying habits. 400 women shopper's are chosen at random in super Market A located in a certain city. Their average weekly food expenditure is Rs.250 with a SD of Rs.40. For 400 women shopper's are chosen at random in super Market B in another city. The average weekly food expenditure is Rs.220 with a SD of Rs.55.

Test at 1%  $\alpha$  ie. Level of Significance (LOS) whether the average weekly food expenditure of the two populations of shoppers are equal.

Soln : Given  $n_1 = 400$ ,  $n_2 = 400$ ,  $\bar{x}_1 = 250$ ,  $\bar{x}_2 = 220$ ,  $s_1 = 40$ ,  $s_2 = 55$

LOS  $\alpha = 0.01$ , corresponding value from Z-table = 2.58

Now, Null Hypo  $H_0 : \mu_1 = \mu_2$  and Alternate Hypo  $H_1 : \mu_1 \neq \mu_2$  (2-tailed test)

$$\begin{aligned} \text{Test stat } Z &= (\bar{x}_1 - \bar{x}_2) / \{ \sqrt{(s_1^2 / n_1 + s_2^2 / n_2)} \} = (250 - 220) / \{ \sqrt{(40^2 / 400 + 55^2 / 400)} \} \\ &= 30 / \sqrt{11.5625} = 30 / 3.4 = 8.82 \end{aligned}$$

As,  $Z_{\text{stat}} (8.82) > Z_{\text{tab}} (2.58)$ , Null Hypothesis is rejected (ie. Food expenditure of 2 population of shoppers are not equal).

**Ex 8 :** There is an assumption that there is no significant difference between boys and girls wrt intelligence. Tests are conducted on two groups and the following are the observations.

	Size	Mean	SD
Girls	60	75	8
Boys	100	73	10

Validate this at 5% Level of Significance.

**Soln :**  $H_0 : \mu_1 = \mu_2$  and  $\mu_1 \neq \mu_2$  (Two tailed test)

$$\text{Test statistic } Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= (75 - 73) / \sqrt{(8^2/60 + 10^2/100)} = 1.3912$$

Since test statistic value 1.3912 falls within acceptance region the hypothesis is accepted.

Ex 9 : The researchers reported the following sample statistics.

In a sample of 45 women dining with other women, the average number of calories ordered was 850, and the standard deviation was 252.

In a sample of 32 women dining with men, the average number of calories ordered was 719, and the standard deviation was 322.

- (i) At 5% LOS can we conclude that average calories ordered by both the group of women are same?
- (ii) Also evaluate the P-value.

Soln : (i) Using the steps applied in previous numerical we proceed to find

$$\text{Test statistic, } Z_{\text{stat}} = \{(x_1 - x_2)\} / [\sqrt{(s_1^2/n_1 + s_2^2/n_2)}]$$

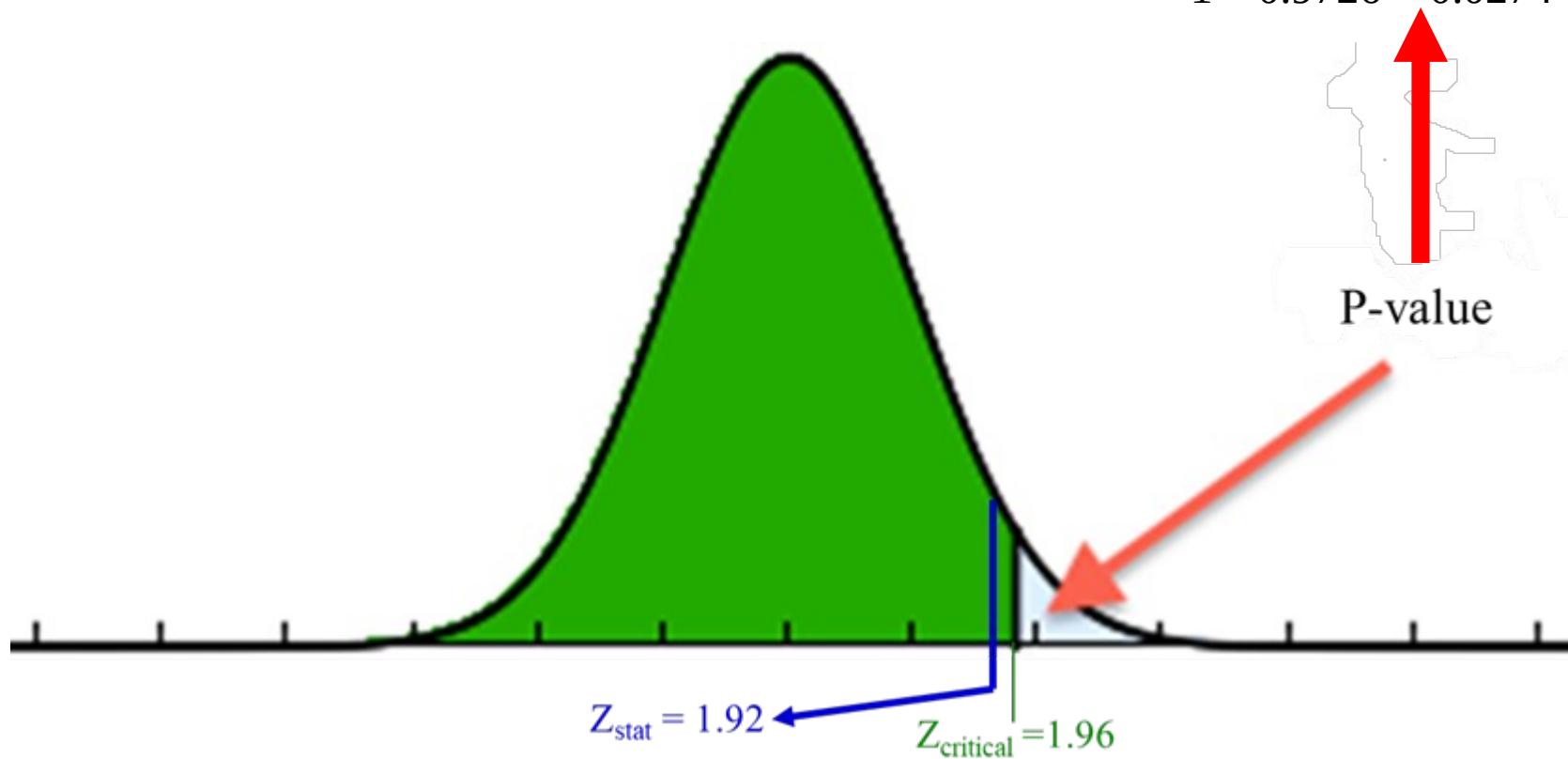
$$= (850 - 719) / [\sqrt{(252^2 / 45 + 322^2 / 32)}] \approx 131 / 72.47 \approx 1.92$$

$$Z_{\text{critical}} \text{ (obtained from Z-table)} = 1.96$$

Since,  $Z_{\text{stat}} (1.92) < Z_{\text{critical}} (1.96)$ , null hypothesis is accepted.

(ii) Area to the left of  $Z_{\text{stat}}$  (1.92) = 0.9726. Hence area to the right of 1.92 is -

$$1 - 0.9726 = 0.0274$$



$P\text{-value} = 0.0274$  which is more than the significance level 0.025. Hence, we accept null hypothesis and reject alternative hypothesis.

## Testing the difference between two means using t-test

Assumptions : Two independent populations follow (i) Normality with (ii) Generally with Equal unknown Variances ( $s^2$ ). (iii) The two random Sample Sizes ( $n_1$  and  $n_2$ ) are less than 30. The averages of sample 1 & sample 2 are respectively  $\bar{x}_1$  &  $\bar{x}_2$ .

### Algorithm :

Step 1 - State Null and Alternative Hypothesis -

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_1 : \mu_1 < \mu_2 \text{ or, } \mu_1 > \mu_2 \text{ or, } \mu_1 \neq \mu_2$$

Step 2 - Specify the level of significance 'α'

Step 3 - Assuming Standard Normal Distribution compute test statistic 't'.

$$\text{Test Statistic } t = (\bar{x}_1 - \bar{x}_2) / [s\sqrt{\{1/n_1 + 1/n_2\}}]$$

$$\text{where, } s^2 = [(n_1 - 1)(s_1)^2 + (n_2 - 1)(s_2)^2] / (n_1 + n_2 - 2) \text{ (if } s_1 \neq s_2\text{)}$$

Step 4 – Find the t-value from t-table for  $t_{\alpha, df}$  & Define the critical region ie. Acceptance & Rejection criteria. Here, df ie. degree of freedom =  $n_1 + n_2 - 2$

Step 5 – Draw inference whether to Accept or Reject the Hypothesis based on critical value or P-value.

**Ex 10 :** The manager of a courier service believes that packets delivered at the beginning of the month are heavier than those delivered at the end of month. As an experiment, he weighed a random sample of 15 packets at the beginning of the month and found that the mean weight was 5.25 kg. A randomly selected 10 packets at the end of the month had a mean weight of 4.26 kg. It was observed from the past experience that the sample variances are 1.20 kg<sup>2</sup> and 1.15 kg<sup>2</sup>. Determine - At 5% level of significance, can it be concluded that the packets delivered at the beginning of the month weigh more?

**Soln :** Given :  $\bar{x}_1 = 5.25$ ,  $\bar{x}_2 = 4.26$ ,  $n_1 = 15$ ,  $n_2 = 10$ ,  $(s_1)^2 = 1.2$ ,  $(s_2)^2 = 1.15$

Degree of freedom =  $n_1 + n_2 - 2 = 15 + 10 - 2 = 23$ ;

from t-table  $t_{0.05/2, 23} = 2.069$  (because it is both sided test)

$$s = \sqrt{[(n_1 - 1)(s_1)^2 + (n_2 - 1)(s_2)^2 / (n_1 + n_2 - 2)]} = \sqrt{[(14 \times 1.2 + 9 \times 1.15) / 23]}$$

Now, Null Hypothesis  $H_0 : \mu_1 = \mu_2$  and Alt Hypo  $H_1 : \mu_1 \neq \mu_2$

$$\begin{aligned} \text{Test statistic } t &= (x_1 - x_2) / \{s\sqrt{(1/n_1 + 1/n_2)}\} \\ &= (5.25 - 4.26) / \{1.086\sqrt{(1/15 + 1/10)}\} = 2.233 \end{aligned}$$

Since,  $t_{\text{stat}} (2.233) > t_{\text{critical}} (2.069)$   $H_0$  is rejected.

**Ex 11 :** As a manager of finance you are assigned the task of choosing the best product in terms of life of the product. Use 5% level of significance.

Product A - Mean 1456 hours with SD 423, size 10

Product B - Mean 1280 hours with SD 398, size 17

Soln :  $H_0: \mu_1 = \mu_2$  and  $H_1: \mu_1 \neq \mu_2$

$$s^2 = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 / (n_1 + n_2 - 2) = 9(423)^2 + 16(398)^2 / (10 + 17 - 2) \\ = 1,65,793; \text{ Hence, } s = 407.18$$

$$\text{Test stat } t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{1/n_1 + 1/n_2}} \\ = \frac{(1456 - 1280)}{407.18(1/10 + 1/17)} = 1.085$$

$$\text{Degree of Freedom} = n_1 + n_2 - 2 = 10 + 17 - 2 = 25$$

We obtain from t-table,  $t_{0.05/2, 25} = 2.06$

Since test statistic  $t$  falls within acceptance region, we accept the null hypothesis.

**Ex 12 :** Random samples of 15 and 10 were selected from two thermocouples. The sample means were 315, 303 and sample standard deviations were 3.8, 4.9 respectively. Test whether there is any significant difference in the means of two thermocouples at 5% level of significance

**Soln :** Given  $\bar{x}_1 = 315$ ,  $\bar{x}_2 = 303$ ,  $s_1 = 3.8$ ,  $s_2 = 4.9$ ,  $n_1 = 15$ ,  $n_2 = 10$

LOS  $\alpha = 0.05$ , Degree of freedom =  $n_1 + n_2 - 2 = 15 + 10 - 2 = 23$

$$\begin{aligned}s^2 &= (n_1 - 1)(s_1)^2 + (n_2 - 1)(s_2)^2 / (n_1 + n_2 - 2) \\ &= 14(3.8)^2 + 9(4.9)^2 / (15 + 10 - 2) = 18.185; \text{ Hence, } s = 4.26\end{aligned}$$

$$t_{0.05/2, 23} = 2.069 \text{ (From t-table)}$$

$$H_0 : \mu_1 = \mu_2 \text{ and } H_1 : \mu_1 \neq \mu_2$$

$$\begin{aligned}\text{Test statistic } t &= (x_1 - x_2) / \{s\sqrt{(1/n_1 + 1/n_2)}\} \\ &= (315 - 303) / \{4.26\sqrt{1/15 + 1/10}\} = 12 / 1.739 = 6.9\end{aligned}$$

Test statistics value is beyond acceptance region, so we can not accept the null hypothesis.

## Testing of Hypothesis for proportion of two populations

Ex 13 : Break up of no. of disconnections of BSNL telephones in the 2 cities – Bangalore & Delhi are respectively, 387 disconnected out of 1500 connections & 310 disconnected out of 1200 connections. Ascertain if there is significant difference between the two proportions of telephone disconnection by BSNL at a confidence level of 99% (LOS = 1%).

Soln : Given;  $x_1 = 387$ ,  $n_1 = 1500$ ,  $x_2 = 310$ ,  $n_2 = 1200$

So,  $p_1 = x_1 / n_1 = 387 / 1500 = 0.258$  and  $p_2 = x_2 / n_2 = 310 / 1200 = 0.2583$

$P = (x_1 + x_2) / (n_1 + n_2) = (387 + 310) / (1500 + 1200) = 697 / 2700 = 0.2581$

$Q = 1 - P = 1 - 0.2581 = 0.7419$

Now,  $H_0 : p_1 = p_2$  and  $H_1 : p_1 \neq p_2$  (two tailed test)

$$\begin{aligned}
 Z_{\text{stat}} &= (p_1 - p_2) / [\sqrt{PQ(1/n_1 + 1/n_2)}] \\
 &= (0.258 - 0.2583) / [\sqrt{(0.2581 \times 0.7419)(1/1500 + 1/1200)}] = -0.0003 / 0.0169 \\
 &= -0.017
 \end{aligned}$$

$Z_{\text{tab}}$  at 1% LOS = 2.58. Since  $Z_{\text{test}} < Z_{\text{tab}}$   $H_0$  is accepted ie. no significant difference in the proportion of telephone disconnection between Bangalore & Delhi.

Ex 14 : It is found that 290 errors in the randomly selected 400 lines of code from Team A and 160 errors in 300 lines of code from Team B. At 5% LOS can we say that Team B's performance is superior to that of Team A?

$$\text{Soln : } p_1 = x_1 / n_1 = 290 / 400 = 0.725, \quad p_2 = x_2 / n_2 = 160 / 300 = 0.533$$

$$P = (x_1 + x_2) / (n_1 + n_2) = (290 + 160) / (400 + 300) = 0.6428; \text{ so } 1 - P = 0.3572$$

Null Hypo  $H_0 : p_2 \geq p_1$  and Alternate Hypo  $H_1 : p_2 < p_1$

$$\begin{aligned} Z_{\text{test}} &= (p_1 - p_2) / [\sqrt{P(1 - P)(1/n_1 + 1/n_2)}] \\ &= (0.725 - 0.533) / [\sqrt{(0.6428 \times 0.3572)((1/400 + 1/300))}] = 5.33 \end{aligned}$$

At 5% LOS  $Z_{\text{tab}} = -1.645$ .

Since test stat is in the Acceptance region, we accept the hypothesis & infer that Team B's performance is superior to that of Team A.

## Non - Parametric tests

It is assumed that the data do not follow any probability distribution which is not characterized by any parameters.

Chi - Square Test is Distribution - free tests

## Testing of Hypothesis : Chi square Test Non-Parametric Test

Independence of  
Two categorical Variables

Goodness-of-fit  
Used for Discrete Distribution

Should be applied only for frequencies. Not for Percentages, Ratios, Mean etc.

The Hypothesis to be tested for independence will be –

$H_0$  : The two categorical variables may be Independent (may not be associated).

$H_1$  : The two categorical variables may not be Independent (may be associated).

Statistical Test is  $\chi^2 = \sum_i^k (O_i - E_i)^2 / E_i$  ,  $k = r \times c$  no. of cells (total no. of cells in  $r \times c$  contingency table)

The test-statistic follows Chi-square distribution with  $(r - 1)(c - 1)$  degrees of freedom where,  $r$  = no. of rows &  $c$  = no. of columns

Chi-square is calculated by –

$$\chi^2 = \sum_{i=1}^k (O_i - E_i)^2 / E_i \approx \chi^2[(r - 1)(c - 1)]$$

where,  $k = r \times c$  is total no. of cells in the  $r \times c$  contingency table,  $r$  is total no. of rows and  $c$  is total no. of columns.

Expected frequencies  $E_{ij} = r_i c_j / n$ , for  $i = 1, 2, 3, \dots, m$  &  $j = 1, 2, 3, \dots, n$

$r_i$  =  $i^{\text{th}}$  row total,  $c_j$  =  $j^{\text{th}}$  column total and  $n$  = overall total

### Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of $\chi^2$								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21

## Testing of Hypothesis - Chi-square test : Independence

Interpretation -

$H_0$  : Smoking habit and Cancer of lung may be independent (may not be associated).

$H_1$  : Smoking habit and Cancer of lung may not be independent (may be associated).

$$\chi^2 = 5.555$$

$df = 1$ , Critical value at  $\alpha = 0.05$  is 3.841,  $P = 0.035$ .

Inference : There may be association between smoking and Cancer of lung.

Ex 15 : Three pension plans independent of job classification. Use  $\alpha = 0.05$ . The opinion of a random sample of 500 employees are shown below -

Job Classification	Pension Plan			Total
	1	2	3	
Salaried Workers	166	86	68	320
Hourly Workers	84	64	32	180
Total	250	150	100	500

$$E_1 = \{(r_1 c_1) / n\} = \{(320 \times 250) / 500\} = 106.24$$

$$E_2 = \{(r_1 c_2) / n\} = \{(320 \times 150) / 500\} = 96.00$$

$$E_3 = \{(r_1 c_3) / n\} = \{(320 \times 100) / 500\} = 64.00$$

$$E_4 = \{(r_2 c_1) / n\} = \{(180 \times 250) / 500\} = 90.00$$

$$E_5 = \{(r_2 c_2) / n\} = \{(180 \times 150) / 500\} = 54.00$$

$$E_6 = \{(r_2 c_3) / n\} = \{(180 \times 100) / 500\} = 36.00$$

Sl No.	$O_i$	$E_i$	$O_i - E_i$	$(O_i - E_i)^2$	$(O_i - E_i)^2 / E_i$
1	166	106.24	59.76	3571.26	33.62
2	86	96	-10	100	1.04
3	68	64	4	16	0.25
4	84	90	-6	36	0.40
5	64	54	10	100	1.85
6	32	36	-4	16	0.44
Chi-square value				37.6	

$$\chi^2 = \sum_{i=1}^k (O_i - E_i)^2 / E_i, \quad k = r \times c \text{ no. of cells (total no. of cells in } r \times c \text{ contingency table)}$$

---

$H_0$  : Job satisfaction and pension plan may be independently distributed (not associated).

$H_1$  : Job satisfaction and pension plan may not be independently distributed (Associated).

$$\chi^2 = 37.6$$

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

$$\chi^2_{0.05, df2} \text{ (from table) for LOS 0.05} = 5.99$$

Inference : Reject  $H_0$  and accept  $H_1$  ie. Job satisfaction and pension plan are Associated.

---

## $\chi^2$ – Test of Goodness of fit

It is a test of fitness between observed and expected frequencies. Here –

$H_0$  : The fit is good between observed (O) and expected frequencies (E)  
ie.  $O \approx E$

$H_1$  : The fit is not good.

The test statistic is  $\chi^2 = \sum (O - E)^2 / E \sim \chi^2_{(k-1) \text{ df}}$ . Here E are found as stated in the problem. If Calculated  $\chi^2 \leq$  Critical  $\chi^2_{(k-1) \text{ df}}$  at  $\alpha$  LOS then  $H_0$  is accepted otherwise  $H_0$  is rejected.

Ex 16 : Following is the record of number of accidents which took place during the various days of the week in India.

Days	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
No. of Accidents	184	148	145	153	150	154	116

Test whether accidents are uniformly distributed over the entire week at 5% LOS.

Soln :  $H_0$  : Accidents are uniformly distributed.

$H_1$  : Accidents are not uniformly distributed.

Then we take  $E$  = mean of accidents for each day.

$$= (184 + 148 + 145 + 153 + 150 + 154 + 116) / 7 = 150$$

O	E	$(O - E)^2$	$(O - E)^2/E$
184	150	1156	7.71
148	150	4	0.03
145	150	25	0.17
153	150	9	0.06
150	150	0	0
154	150	16	0.11
116	150	1156	7.71
Chi-square			15.79

## Practice Problems

Q1. A sample of 25 sales receipts from a grocery store has average Rs. 950 and variance = 400. Use these values to test whether or not the mean sales at the grocery store are different from Rs. 960.

Q2. An insurance company is reviewing its current policy rates. When originally setting the rates they believed that the average claim amount was \$1,800. They are concerned that the true mean is actually higher than this, because they could potentially lose a lot of money. They randomly select 40 claims, and calculate a sample mean of \$1,950. Assuming that the standard deviation of claims is \$500, and set  $\alpha = 0.05$  test to see if the insurance company should be concerned.

Q3. From long experience with a process for manufacturing an alcoholic beverage it is known that the yield is normally distributed with a mean of 500 and a standard deviation of 96 units. For a modified process the yield is 535 units for a sample of size 50. At  $\alpha = .05$  does the modified process increase the yield?

## Practice Problems

Q4. Trying to encourage people to stop driving to campus, the university claims that on average it takes people 30 minutes to find a parking space on campus. I don't think it takes so long to find a spot. In fact I have a sample of the last five times I drove to campus, and I calculated  $\bar{x} = 20$ . Assuming that the time it takes to find a parking spot is normal, and that  $\sigma = 6$  minutes, then perform a hypothesis test with level  $\alpha = 0.10$  (ie. 10%) to see if my claim is correct.

Q5. Average IQ of the students of an Institute was stated to be utmost 130 with a variance of 144. A sample of 25 students were tested and their IQ was found to be 125.

Test the significance of the statement at a confidence level of 95%.

Q6. A sample of size 9 from a normal population is given below. Find the 90% confidence interval (CI) for the mean  $\mu$  of the population. Also find the 90% confidence interval for the variance  $\sigma^2$  of the population. 0, 1, -1, 1, 1, 0, -1, -2, 3.

## Practice Problems

---

Q7. A car manufacturing company claims that their cars do not require servicing before running for an average 10,000 miles. To verify the claim, the manufacturer tests 30 cars under simulated conditions and gets a mean breakdown of 9500 miles with a standard deviation of 1150 miles. What conclusion can be drawn at a significance level of 5%.

Q8. Find the probabilities that a random variable having standard normal distribution will take on a value

- (a) Between 0.87 and 1.28
- (b) Greater than 0.85
- (c) Between -0.34 and 0.62;



# BSDCHZC355

## Statistical Inference & Applications

**BITS** Pilani  
Pilani | Dubai | Goa | Hyderabad

Shaibal Kr. Sen  
Session 03



# STATISTICAL INFERENCES

&  
APPLICATIONS

## Scheduled Topics to be covered

**Introduction to Design of Experiments (DOE)**

**Basic Principles & Guidelines of DOE**

**Inferences about Differences in Means &  
Randomised Designs**

## What is Design of Experiments (DOE)?

**Design** means **developing** a method or procedure or process or way.

**Experiment** means a Way or Method or Procedure or Process of obtaining an answer to a real life problem.

**Design of Experiments** means developing a method to obtain valid inference (conclusion or answer) to a problem.

Statistical design of experiments refers to the process of planning the experiment so that appropriate data will be collected and analyzed by statistical methods, resulting in valid and objective conclusions.

The statistical approach to experimental design is necessary if we wish to draw meaningful conclusions from the data.

## What is Design of Experiments (DOE)?

When the problem involves data that are subject to experimental errors, statistical methods are the only objective approach to analyse.

Thus, there are two aspects to any experimental problem –

Design of the Experiment and

Statistical Analysis of the data.

These two subjects are closely related because the method of analysis depends directly on the design employed.

## Terminology

Before discussing the principles of designs, it is proper to explain the terminology used in this context. The terms commonly used are – **Experiment, Treatment, Experimental Unit, Experimental Error and Precision.**

**Experiment** : It is a means of getting an answer to the question that the experimenter has in mind. This may be to decide which of several pain-relieving drugs that are available in the market is the most effective or whether they are equally effective. An experiment may be planned to compare the Chinese method of cultivation with the standard method used in India. In planning an experiment, we clearly state our objectives & formulate the hypotheses we want to test.

**Treatment** : The different procedures under comparison in an experiment are the different treatments eg. – in an agricultural experiment, the different varieties of a crop or the different manures

---

will be the treatments. In a dietary or medical experiment, the different diets or medicines, etc. are the treatments.

**Experimental Units** : In carrying out an experiment, we should be clear as to what constitutes the experimental unit. An experimental unit is the material to which is applied the treatment and on which the variable under study is measured. In an agricultural field experiment, the plot of land and not the individual plant, will be the experimental unit; in a feeding experiment of cows, the whole cow is the experimental unit; in human experiments in which the treatment affects the individual, the individual will be the experimental unit.

**Experimental Error** : A fundamental phenomenon in replicated experiments is the variation in the measurements made on different experimental units even when they get the same treatment. A part of this variation is systematic and can be explained, whereas the

---

remainder is to be taken to be of the random type. The unexplained random part of the variation is termed the experimental error. This is a technical term and does not mean a mistake, but includes all types of extraneous (irrelevant or unrelated to the subject) variation due to –

- (i) Inherent variability in the experimental units.
- (ii) Errors associated with the measurements made.
- (iii) Lack of representativeness of the sample to the population under study.

The experimental error provides a basis for the confidence to be placed in the inference about the population. So it is important to estimate & control the experimental error. An estimate of the experimental error can only be obtained by replication, and it is controlled by the principle of local control, to be explained shortly.

---

The precision of an experiment is measured by the reciprocal of the variance of a mean -

$$1 / (\sigma)^2 / n = n / \sigma^2$$

As  $n$  (replication number) increases, precision also increases. Another means of increasing precision is to control  $\sigma^2$ . The smaller the value of  $\sigma^2$ , the greater the precision.

---

Whether in Engineering, R&D, or Science Lab, understanding the basics of experimental design can help to achieve more statistically optimal results applying experimental techniques and improve output quality.

Using Design of Experiments (DOE) techniques, one can determine the individual and interactive effects of various factors that can influence the output results of measurements. It can also be used to gain knowledge and estimate the best operating conditions of a System, Process or Product.

DOE applies to many different investigation objectives, but can be especially important early on in a screening investigation to help in determining the most important factors which can influence the responses or critical quality attributes.

---

## Basic Principles of DOE

Basic Principles of DOE are – Randomization, Replication, Blocking, Multifactor Designs and Confounding

The first three here are perhaps the most important –

**Randomization** : By randomization we mean that both the allocation of the experimental material and the order in which the individual runs of the experiment are to be performed are randomly determined.

Statistical methods require that the observations (or errors) be independently distributed random variables. Randomization usually makes this assumption valid.

By properly randomizing the experiment, we also assist in “averaging out” the effects of extraneous (unnecessary-unwanted-untargeted for study) factors that may be present. It will help to eliminate potential biases from the conclusions.

---

**Replication** : By replication we mean an independent repeat run of each factor combination.

Replication has two important properties –

First, it allows the experimenter to obtain an estimate of the experimental error.

Second, it permits the experimenter to obtain a more precise estimate of the parameter.

---

**Blocking (Local Control)** : Blocking is a design technique used to improve the precision with which comparisons among the factors of primary interest are made.

Often blocking is used to reduce or eliminate the variability transmitted from nuisance (extraneous) factors - that is, factors that may influence the experimental response but in which we are not directly interested.

Generally, a block is a set of relatively homogeneous experimental conditions and the variability within a block would be expected to be smaller than the variability between blocks.

The experimenter divides the observations from the statistical design into groups that are run in each block.

---

**Multi-Factor Designs** : These designs are meant to evaluate multiple factors set at multiple levels. One approach is called a Full Factorial experiment, in which each factor is tested at each level in every possible combination with the other factors and their levels. Full factorial experiments that study all paired interactions can be economic and practical if there are few factors and only 2 or 3 levels per factor. The advantage is that all paired interactions can be studied. However, the number of runs goes up exponentially as additional factors are added. Experiments with many factors can quickly become unwieldy and costly to execute, as shown by the chart in next slide. To avoid complexity, factors which are having major impact on the response variable should be taken and factors of minor significance should be avoided.

Experimental work when many factors are likely to be investigated. It

## Relationship of Experimental Run with Factor-Level-Replicate

provides the smallest number of runs with which  $k$  factors can be studied in a complete factorial design. Consequently, these designs are widely used in *factor screening experiment*.

### Relationship between Experimental Run and Factor-Level-Replicate

No. of Factors (F)	No. of Levels per Factor (L)	No. of Replicates (R)	No. of Runs Full Factorial $(L^F) \times R$
2	2	1	$(2^2) \times 1 = 4$
2	3	2	$(3^2) \times 2 = 18$
3	2	2	$(2^3) \times 2 = 16$
3	3	3	$(3^3) \times 3 = 81$
4	2	2	$(2^4) \times 2 = 32$
4	3	4	$(3^4) \times 4 = 324$
5	2	3	$(2^5) \times 3 = 96$
5	3	4	$(3^5) \times 4 = 972$
6	2	3	$(2^6) \times 3 = 192$
6	3	5	$(3^6) \times 5 = 3645$

---

Confounding - is something that is usually considered bad! Here is an example. Let's say we are doing a medical study with drugs X and Y. We put 10 subjects on drug X and 10 on drug Y. If we categorize our subjects by gender, how should we allocate our drugs to our subjects? Let's make it easy and say that there are 10 male and 10 female subjects. A balanced way of doing this study would be to put five males on drug X and five males on drug Y, five females on drug X and five females on drug Y. This is a perfectly balanced experiment such that if there is a difference between male and female at least it will equally influence the results from drug X and the results from drug Y.

An alternative scenario might occur if patients were randomly assigned treatments as they came in the door. At the end of the study they might realize that drug X had only been given to the male subjects and drug Y was only given to the female subjects. We would call this design

---

---

totally confounded. This refers to the fact that if you analyze the difference between the average response of the subjects on X and the average response of the subjects on Y, this is exactly the same as the average response on males and the average response on females. You would not have any reliable conclusion from this study at all. The difference between the two drugs X and Y, might just as well be due to the gender of the subjects, since the two factors are totally confounded.

Confounding is something we typically want to avoid but when we are building complex experiments we sometimes can use confounding to our advantage. We will confound things we are not interested in order to have more efficient experiments for the things we are interested in. This will come up in multiple factor Experiments. We may be interested in main effects but not interactions so we will confound the interactions in this way in order to reduce the sample size, and thus the cost of the experiment, but still have good information on the main effects.

---

## Key Benefits of DOE

1. Optimization of Processes : Helps in identifying the most influential factors in a process and determining the optimal settings for those factors. This leads to improved efficiency and performance, as processes can be fine-tuned based on empirical data.
2. Efficient Utilization of Resource : By designing experiments to collect the most relevant data with fewer trials, DOE minimizes the number of experiments needed. This saves time, reduces costs, and uses fewer resources.
3. Improved Product Quality : With DOE, it's possible to identify the factors that most significantly impact product quality. By controlling and optimizing these factors, organizations can produce higher-quality products more consistently.
4. Data-Driven Decisions : DOE uses statistical methods to analyze data

---

and draw conclusions, reducing the reliance on guesswork or intuition. This ensures that decisions are based on solid data and not on assumptions or biases.

5. Identifying Interactions Between Variables : One of the key strengths of DOE is its ability to detect interactions between different factors. Understanding these interactions helps in making more informed decisions about how multiple variables affect outcomes.

6. Reduction in Variability : By understanding the relationship between factors and their influence on variability, DOE helps reduce unwanted variation, leading to more stable and predictable processes.

7. Better Risk Management : DOE allows you to predict and control the impact of various factors on the final outcome, helping to identify and mitigate risks early in the process or product development.

---

- 
8. Improved Decision-Making : The insights gained from DOE experiments provide a clearer understanding of cause-and-effect relationships, allowing for better decisions in product development, manufacturing, and other areas.
  9. Flexibility : DOE can be applied across various industries and disciplines, from manufacturing and engineering to healthcare, marketing and agriculture, making it a versatile tool for experimentation and improvement.
  10. Scalability : DOE is not only useful in small-scale experiments but can also be scaled up to large, complex systems. As businesses grow, DOE helps scale their processes while maintaining quality and efficiency.

In summary, the benefits of Design of Experiments lie in its ability to optimize processes, improve decision-making, enhance product quality, and reduce variability - all while using resources efficiently.

Already covered in CS2

## Z test of difference of two means

For testing the significant difference between two population means using large samples ( $n \geq 30$ ), the test statistic is -

$$Z = (\bar{x}_1 - \bar{x}_2) / [\sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)}] = (\bar{x}_1 - \bar{x}_2) / [\sqrt{(s_1^2/n_1 + s_2^2/n_2)}]$$

and the significant value is  $Z_\alpha$  for one tailed test and  $Z_{\alpha/2}$  for two tailed test.

$\mu_1$  and  $\mu_2$  are two population means respectively.

$(\sigma_1)^2$  and  $(\sigma_2)^2$  are two population variances respectively and if they are not known they are replaced by their respective sample variances  $(s_1)^2$  and  $(s_2)^2$ .

$n_1$  and  $n_2$  are two sample sizes respectively (each  $\geq 30$ )

Here,  $H_0 : \mu_1 = \mu_2$  and  $H_1 : \mu_1 \neq \mu_2$

If the calculated  $|Z| \leq$  significant value (obtained from Z-table), then  $H_0$  is accepted; otherwise  $H_0$  is rejected.

## Testing the difference between means using t-test

Assumptions : Two independent populations follow (i) Normality with (ii) Generally with Equal unknown Variances ( $s^2$ ). (iii) The two random Sample Sizes ( $n_1$  and  $n_2$ ) are less than 30. The averages of sample 1 & sample 2 are respectively  $\bar{x}_1$  &  $\bar{x}_2$ .

### Algorithm :

Step 1 - State Null and Alternative Hypothesis -

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_1 : \mu_1 < \mu_2 \text{ or, } \mu_1 > \mu_2 \text{ or, } \mu_1 \neq \mu_2$$

Step 2 - Specify the level of significance 'α'

Step 3 - Assuming Standard Normal Distribution compute test statistic 't'.

$$\text{Test Statistic } t = \frac{(\bar{x}_1 - \bar{x}_2)}{[s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}]}$$

$$\text{where, } s^2 = \frac{[(n_1 - 1)(s_1)^2 + (n_2 - 1)(s_2)^2]}{(n_1 + n_2 - 2)} \text{ (if } s_1 \neq s_2\text{)}$$

Step 4 – Find the t-value from t-table for  $t_{\alpha, df}$  & Define the critical region ie. Acceptance & Rejection criteria. Here, df ie. degree of freedom =  $n_1 + n_2 - 2$

Step 5 – Draw inference whether to Accept or Reject the Hypothesis based on critical value or P-value.

**Ex 1 :** The manager of a courier service believes that packets delivered at the beginning of the month are heavier than those delivered at the end of month. As an experiment, he weighed a random sample of 15 packets at the beginning of the month and found that the mean weight was 5.25 kg. A randomly selected 10 packets at the end of the month had a mean weight of 4.26 kg. It was observed from the past experience that the sample variances are 1.20 kg<sup>2</sup> and 1.15 kg<sup>2</sup>. Determine -

- (i) At 5% level of significance, can it be concluded that the packets delivered at the beginning of the month weigh more?
- (ii) Also find P-value.

Soln : Given  $\bar{x}_1 = 5.25$ ,  $\bar{x}_2 = 4.56$ ,  $n_1 = 15$ ,  $n_2 = 10$ ,  $(s_1)^2 = 1.2$ ,  $(s_2)^2 = 1.15$

Degree of freedom =  $n_1 + n_2 - 2 = 15 + 10 - 2 = 23$ ;  $t_{0.05, 23} = 1.714$

(from t-table).

$$s = \sqrt{[(n_1 - 1)(s_1)^2 + (n_2 - 1)(s_2)^2 / (n_1 + n_2 - 2)]} = \sqrt{[(14 \times 1.2 + 9 \times 1.15) / 23]}$$

(i)  $H_0 : \mu_1 - \mu_2 = 0$  and  $H_1 : \mu_1 - \mu_2 \neq 0$

Test statistic  $t = (\bar{x}_1 - \bar{x}_2) / [s\sqrt{1/n_1 + 1/n_2}]$

$$= (5.25 - 4.26) / [1.086\sqrt{1/15 + 1/10}] = 2.234$$

Since,  $t_{\text{stat}} (2.234) > t_{\text{critical}} (1.714)$  reject  $H_0$ .

(ii) P- value corresponding to  $t = 2.234$  is  $0.01 < P < 0.025$  ie. less than 0.05.

**Ex 2 :** As a manager of finance you are assigned the task of choosing the best product in terms of life of the product. Use 5% level of significance.

Product A - Mean 1456 hours with SD 423, size 10

Product B - Mean 1280 hours with SD 398, size 17

Soln :  $H_0: \mu_1 = \mu_2$  and  $H_1: \mu_1 \neq \mu_2$

$$s^2 = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 / (n_1 + n_2 - 2) = 9(423)^2 + 16(398)^2 / (10 + 17 - 2)$$

$$= 1,65,793 \quad s = 407.18$$

$$\text{Test stat } t = \{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)\} / \{s\sqrt{(1/n_1 + 1/n_2)}\}$$

$$= \{(1456 - 1280) - 0\} / \{407.18(1/10 + 1/17)\} = 1.085$$

$$\text{Degree of Freedom} = n_1 + n_2 - 2 = 10 + 17 - 2 = 25$$

We obtain from t-table,  $t_{0.05, 25} = 2.06$

Since test statistic  $t$  falls within acceptance region, we accept the null hypothesis.

**Ex 3 :** Random samples of 15 and 10 were selected from two thermocouples. The sample means were 315, 303 and sample standard deviations were 3.8, 4.9 respectively. Test whether there is any significant difference in the means of two thermocouples at 5% level of significance

**Soln :** Given  $\bar{x}_1 = 315$ ,  $\bar{x}_2 = 303$ ,  $s_1 = 3.8$ ,  $s_2 = 4.9$ ,  $n_1 = 15$ ,  $n_2 = 10$

LOS  $\alpha = 0.05$ , Degree of freedom =  $n_1 + n_2 - 2 = 15 + 10 - 2 = 23$

$t_{0.05, 23} = 1.714$  (From t-table)

$H_0 : \mu_1 = \mu_2$  and  $H_1 : \mu_1 \neq \mu_2$

Test statistic  $t = [(x_1 - x_2) - (\mu_1 - \mu_2)] / [\sqrt{(s_1)^2/n_1 + (s_2)^2/n_2}]$

$$= (315 - 303) / [\sqrt{(3.8)^2/15 + (4.9)^2/10}] = 12 / 1.834 = 6.54$$

Test statistics value is beyond acceptance region, so we can not accept the null hypothesis.

Ex 4 : The researchers reported the following sample statistics.

In a sample of 45 women dining with other women, the average number of calories ordered was 850, and the standard deviation was 252.

In a sample of 27 women dining with men, the average number of calories ordered was 719, and the standard deviation was 322.

- (i) At 5% LOS can we conclude that average calories ordered by both the group of women are same?
- (ii) Also evaluate the P-value.

Soln : Using the steps applied in previous numerical we proceed to find

$$(i) t_{\text{stat}} = (850 - 719) / [\sqrt{(252)^2 / 45 + (322)^2 / 27}] \approx 131 / 72.47 \approx 1.81$$

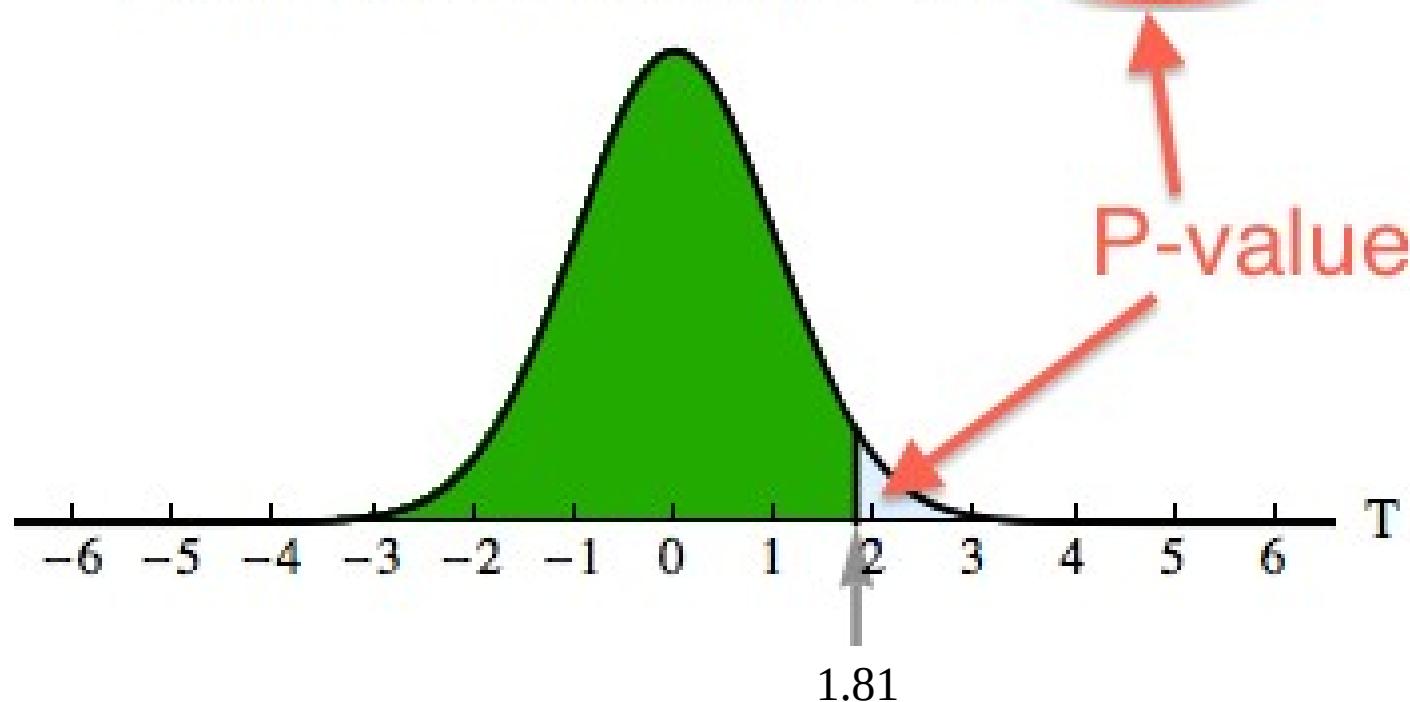
$$t_{\text{critical}} = 1.645$$

Since,  $t_{\text{stat}} > t_{\text{critical}}$  , null hypothesis is rejected.

(ii)

The green area to the left of the  $T$  value = 0.9615

The blue area to the right of the  $T$  value = 0.0385



$P$ -value = 0.0385 which is less than the significance level 0.05. Hence, we reject null hypothesis and accept alternative hypothesis.

## Practice Problems

---

Q1. Explain usefulness of Design of Experiment (DOE)?

Q2. Random samples of 25 and 20 were selected from two thermocouples. The sample means were 321, 308 and sample standard deviations were 4.1, 4.7 respectively. Test whether there is any significant difference in the means of two thermocouples at 1% level of significance

Q3. Is it true that experiments enhance performance of Production Process? Justify your answer.

Q4. State the salient features of DOE and explain two important ones.

Q5. Explain Assignable causes of Quality Variation.

Q7. Narrate the Experimental Design Process.

## Practice Problems

Q8. There is an assumption that there is no significant difference in the productivity of men and women workers in a production center. Tests were conducted on two groups and the following are the observations.

	Sample Size	Average Productivity	SD
Men	50	100	9
Women	60	95	8

Validate this at 5% Level of Significance.

Q9. Fill-up the blank column –

No. of Factors	No. of levels	No. of runs for full Factorial Design
2	2	
2	5	
3	3	
4	3	
5	4	



# BSDCHZC355

## Statistical Inference & Applications

**BITS** Pilani  
Pilani | Dubai | Goa | Hyderabad

Shaibal Kr. Sen  
Session 04



# STATISTICAL INFERENCES & APPLICATIONS

## Scheduled Topics to be covered

---

**Experiment with a Single Factor : Analysis of Variance  
(ANOVA)**

**ANOVA Introduction**

**One Way ANOVA Classification**

## Analysis of Variance (ANOVA)

When do you use an ANOVA?

ANOVA is used to determine whether there are any significant differences between the means of three or more independent (unrelated) groups. Here we will provide a brief introduction to the one-way ANOVA, including the assumptions of the test and when you should use this test.

Analysis of variance (ANOVA) is a collection of statistical models used in order to analyze the differences among group means and their associated procedures (such as “variation” among and between groups), developed by statistician and evolutionary biologist Ronald Fisher. In the ANOVA setting, the observed variance in a particular variable is partitioned into

---

components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether or not the means of several groups are equal, and therefore generalizes the *t*-test to more than two groups. As doing multiple two-sample *t*-tests would result in an increased chance of committing a statistical type I error, ANOVAs are useful in comparing (testing) three or more means (groups or variables) for statistical significance.

Analysis of variance seeks to identify sources of variation in numerical dependent variable Y (the response variable). Variation in the response variable about its mean either is explained by one or more categorically independent variable/s (the factors) or is unexplained (random error).

ANOVA is a comparison of means. Each possible value of a factor or combination of factors is a treatment. Sample observations within each treatment are viewed as coming from populations with possibly different means. We test whether each factor has a significant effect on  $Y$ , and sometimes we test for interaction between factors. The test uses the F distribution.

F-Distribution : Is variance ratio distribution. If we choose two independent random samples from two normal population having equal variances ( $\sigma^2$ ).

Let data values of 1<sup>st</sup> sample be  $x_1, x_2, x_3, \dots, x_{n_1}$  sample size =  $n_1$  and Mean =  $\bar{x}$   
 data values of 2<sup>nd</sup> sample be  $y_1, y_2, y_3, \dots, y_{n_2}$  sample size =  $n_2$  and Mean =  $\bar{y}$

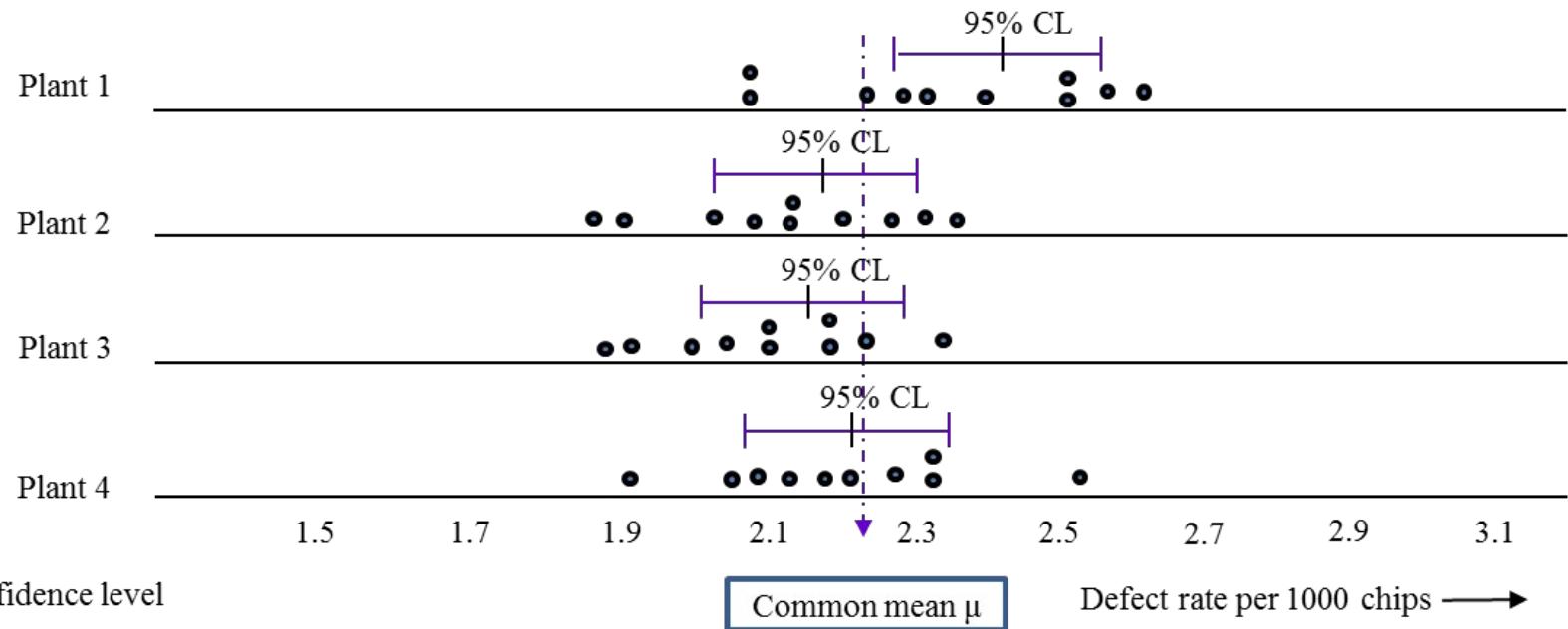
$$\text{Then, } S_1^2 = \left\{ \sum_{i=1}^{n_1} (x_i - \bar{x})^2 \right\} / (n_1 - 1) \quad \text{and} \quad S_2^2 = \left\{ \sum_{i=1}^{n_2} (y_i - \bar{y})^2 \right\} / (n_2 - 1)$$

‘F’ is defined as,  $F = S_1^2 / S_2^2$  (Note : Larger variance is placed in the numerator)

---

ANOVA can handle any number of factors, but the researcher often is interested only in a few. Also, data collection costs may impose practical limits on the number of factors or treatments we can choose. We will focus on ANOVA with one or two factors.

Manufacturing defect rates : Figure below shows a dot plot of daily defect rates for automotive computer chips mfgd at four plant locations. Samples of 10 days production were taken at each plant. Are the observed differences in the plants' sample mean defect rates merely due to random variation? Or are the observed differences between the plants' defect rates too great to be attributed to chance? This is the kind of question that ANOVA is designed to answer.

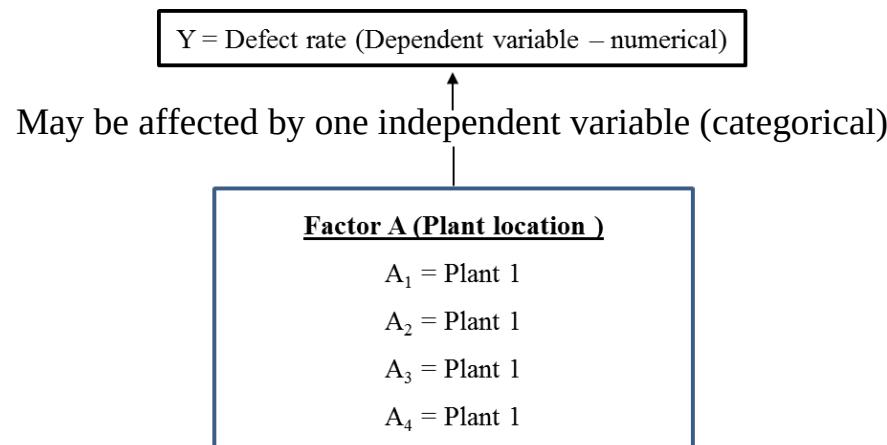


A simple way to state the ANOVA hypothesis is

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$  (mean defect rates are same at all four plants)

$H_1 : \text{Not all the means are equal (at least one mean differs from the others)}$

If we can not reject  $H_0$ , then we conclude that the observations within each treatment or group actually have a common mean  $\mu$  (represented by dotted line) in the previous slide. This one factor ANOVA model may be visualized as below –



## Hospital length of stay

To allocate resources and fix costs correctly, hospital management needs to test whether a patient's length of stay (LOS) depends on the diagnostic related grp (DRG) code and the patient's age grp. Consider the case of a bone fracture. LOS is a numerical response variable (measured in hours). The hospital organizes the data by using five diagnostic codes for age grp (under 18, 18 to 64 & 65 and above). Although patient age is a numerical variable, it is coded into 3 categories based on stages of bone growth. Figure below illustrates two possible ANOVA models (one factor and two factor). We could also test for interaction between factors, as you will see later on.

### One Factor ANOVA

Y = Hospital LOS in Hrs  
(Dependent variable numerical)

May be affected by one independent variable (categorical)

#### Factor A (Type of fracture)

$A_1$  = Facial  
 $A_2$  = Radius or Ulna  
 $A_3$  = Hip or Femur  
 $A_4$  = Other Lower Extremity  
 $A_5$  = All Other

### Two Factor ANOVA

Y = Hospital LOS in Hrs  
(Dependent variable numerical)

May be affected by two independent variable (categorical)

#### Factor A (Type of fracture)

$A_1$  = Facial  
 $A_2$  = Radius or Ulna  
 $A_3$  = Hip or Femur  
 $A_4$  = Other Lower Extremity  
 $A_5$  = All Other

#### Factor B (Age group)

$B_1$  = under 18  
 $B_2$  = 18 to 64  
 $B_3$  = 65 and above

---

Similarly, we can design various ANOVA examples for more such cases viz. – Automobile Paintings (Different paint suppliers eg. – Asian Paints, Berger, Jenson & Nicholson etc.), Car Manufacturing (Maruti, Honda, Hyundai etc. etc.), FMCG Regional Distributor (Eastern, Western, Northern and Southern workstations) services etc. etc.

ANOVA calculations are usually too tedious to do by calculator, so after we choose an ANOVA model and collect the data, we rely on software [eg. Excel, Megastat, Minitab, SPSS (Statistical Package for the Social Sciences)] to do the calculations. In some applications viz. – Accounting, Finance, HR, Marketing large samples can easily be taken from existing records, while in others viz. – Engg, Mfg, Computer Systems experimental data collection is

---

so expensive that small samples are used. Large samples increase the power of the test, but power also depend on degree of variation in  $Y$ . Lowest power would be in small samples with higher variation in  $Y$ , and conversely. Specialised software is needed to calculate power for ANOVA experiments.

## **ANOVA Assumptions**

Observations in  $Y$  are independent.

Populations being sampled are normal.

Populations being sampled have equal variances.

Fortunately, ANOVA is somewhat robust to departures from the normality and equal variance assumptions.

---

One factor ANOVA (Completely randomized model) : Data format - In the following illustration, if we are only interested in comparing of C-grps (treatments or factor levels), we have a one factor ANOVA. This is by far the most common ANOVA model that covers many business problems. The one factor ANOVA is usually viewed as a comparison between several columns of data, although the data could also be presented in rows.

Table below illustrates the data format for a one-factor ANOVA with C treatments, denoted  $A_1, A_2, A_3, \dots, A_c$ . The group means are  $y_1, y_2, y_3, \dots, y_c$ .

### One-factor ANOVA data in column

A <sub>1</sub>	A <sub>2</sub>	....	A <sub>c</sub>
y <sub>11</sub>	y <sub>12</sub>	....	y <sub>1c</sub>
y <sub>21</sub>	y <sub>22</sub>	....	y <sub>2c</sub>
y <sub>31</sub>	y <sub>32</sub>	....	y <sub>3c</sub>
....	....	....	....
....	....	....	....
n <sub>1</sub> obsns	n <sub>2</sub> obsns	....	n <sub>c</sub> obsns
y <sub>1</sub>	y <sub>2</sub>	....	y <sub>c</sub>

### One-factor ANOVA data in rows

A <sub>1</sub>	y <sub>11</sub>	y <sub>21</sub>	y <sub>31</sub>	....	....	n <sub>1</sub> obsns	y <sub>1</sub>
A <sub>2</sub>	y <sub>12</sub>	y <sub>22</sub>	y <sub>32</sub>	....	....	n <sub>2</sub> obsns	y <sub>2</sub>
....	....	....	....	....	....	....	....
A <sub>c</sub>	y <sub>1c</sub>	y <sub>2c</sub>	y <sub>3c</sub>	....	....	n <sub>c</sub> obsns	y <sub>c</sub>

---

Within treatment  $j$  we have  $n_j$  observations on  $y$ . Sample sizes within each treatment do not need to be equal, although there are advantages to having balanced sample sizes. The total number of observations is the sum of the sample sizes for each treatment

$$n = n_1 + n_2 + n_3 + \dots + n_c$$

Hypothesis to be tested : The question of interest is whether the mean of  $Y$  varies from treatment to treatment. The hypothesis to be tested are –

Null Hypothesis  $H_0 : \mu_1 = \mu_2 = \mu_3 = \dots = \mu_c$  (ie. all the means are equal)

Alternative Hypo  $H_1$  : Not all the means are equal (at least one mean is different)

Since one-factor ANOVA is a generalization of the test for equality of two means, why not just compare all possible pairs

---

of means by using repeated two sample t-test. Consider our experiment comparing the four mfg plant average defect rates. To compare pairs of plant averages we would have to perform six ( ${}^4C_2$ ) different t-tests. If each t-test has a Type I error probability equal to 0.05, then the probability that at least one of these test results in a Type I error is  $1 - (0.95)^6 = 0.2649$ . ANOVA tests all the means simultaneously and therefore does not inflate our Type I error.

One factor ANOVA as a linear model : An equivalent way to express the one factor model is to say that observations in treatment  $j$  came from a population with a common mean ( $\mu$ ) plus a treatment effect ( $A_j$ ) plus random error ( $e_{ij}$ ) :

$$y_{ij} = \mu + A_j + e_{ij} \text{ where, } j = 1, 2, \dots, c \text{ & } i = 1, 2, \dots, n_j$$

---

The random error is assumed to be normally distributed with zero mean and same variance for all treatments. If we are interested only in what happens to the response for the particular levels of the factor that were selected (a fixed-effect model), then the hypothesis to be tested are –

$H_0 : A_1 = A_2 = \dots = A_c = 0$  (all treatment effects are zero)

$H_1 : \text{Not all } A_j \text{ are zero}$  (some treatment effects are non-zero)

If the null hypothesis is true ( $A_j = 0$  for all  $j$ ), then knowing that an observation  $x$  came from treatment  $j$  does not help explain the variation in  $Y$  and the ANOVA model collapses to

$$y_{ij} = \mu + e_{ij}$$

---

Group means : The mean of each group is calculated in the usual way by summing the observations in the treatment and dividing by the sample size :

$$\bar{y}_j = (1 / n_j) \sum_{i=1}^{n_j} y_{ij}$$

The overall sample mean or grand mean  $\bar{y}$  can be calculated either by summing all the observations and dividing by  $n$  or by taking a weighted average of the  $c$  sample means :

$$\bar{y} = (1 / n) \sum_{j=1}^c \sum_{i=1}^{n_j} y_{ij} = (1 / n) \sum_{j=1}^c n_j \bar{y}_j$$

Partitioned sum of squares : To understand the logic of ANOVA, consider that for a given observation  $y_{ij}$  the following relationship must hold (on the RHS we just add & subtract  $\bar{y}_j$ ) –

$$(y_{ij} - \bar{y}) = (\bar{y}_j - \bar{y}) + (y_{ij} - \bar{y}_j)$$

This says that any deviation of an observation from the grand mean  $\bar{y}$  may be expressed in two parts - deviation of column mean ( $y_j$ ) — from the grand mean ( $\bar{y}$ ), or between treatments, and the deviation of the observation ( $y_{ij}$ ) from its own column mean ( $\bar{y}_j$ ), — or within treatments. We can show that this relationship also holds for sums of squared deviations, yielding the partitioned sum of squares.

$$\begin{aligned}
 \sum_{j=1}^c \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2 &= \sum_{j=1}^c \sum_{i=1}^{n_j} \{(\bar{y}_j - \bar{y}) + (y_{ij} - \bar{y}_j)\}^2 \\
 &= \sum_{j=1}^c \sum_{i=1}^{n_j} \{(\bar{y}_j - \bar{y})^2 + 2(\bar{y}_j - \bar{y})(y_{ij} - \bar{y}_j) + (y_{ij} - \bar{y}_j)^2\} \\
 &= \sum_{j=1}^c n_j (\bar{y}_j - \bar{y})^2 + \sum_{j=1}^c \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2
 \end{aligned}$$

$$\sum_{j=1}^c \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2 = \sum_{j=1}^c n_j (\bar{y}_j - \bar{y})^2 + \sum_{j=1}^c \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

This important relationship may be expressed simply as

$$SST = SSA + SSE \text{ (Partitioned Sum of Squares)}$$

ie. Sum of squares total = sum of squares between treatments + sum of squares within treatments  
(Explained by factor A) (Unexplained random error)

If the treatment means do not differ greatly from grand mean, SSA will be small and SSE will be large (and conversely). The sums SSA and SSE may be used to test the hypothesis that the treatment means differ from the grand mean. However, we first divide each sum of squares by its degree of freedom (to adjust for group sizes). The test statistic is the ratio of the resulting mean squares. These calculations can be arranged in the tabular format as shown in the next slide.

Source of variation	Sum of squares	Degree of freedom	Mean of squares	F-statistic
Treatment (between groups)	$SSA = \sum_{j=1}^c n_j (\bar{y}_j - \bar{y})^2$	Total no. of Grps - 1 $c - 1$	$MSA = SSA / (c - 1)$	
Error (within groups)	$SSE = \sum_{j=1}^c \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$	Total no. Obs - no. of grps $n - c$	$MSE = SSE / (n - c)$	$F = MSA / MSE$
Total	$SST = \sum_{j=1}^c \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2$	$(c - 1) + (n - c) = n - 1$		

---

The ANOVA calculations are mathematically simple but involve tedious sums. These calculations are almost always done on computer.

Test statistic : At the beginning of this chapter we described the variation in  $Y$  as consisting of explained variation and unexplained variation. To test whether the independent variable explains a significant proportion of the variation in  $Y$ , we need to compare the explained (due to treatments) and unexplained (due to error) variation. Recall that the F-distribution describes the ratio of two variances. Therefore, it makes sense that the ANOVA test statistic is the F-test statistic. The F statistic is the ratio of the variance due to treatments to the variance due to error. MSA is the mean square due to treatment and MSE is the

---

mean square within treatments. Following equation shows the F-statistic and its degree of freedom.

$$F = MSA / MSE = \{(SSA / (c - 1)) / SSE / (n - c)\}$$

where,  $c$  = number of groups &  $n$  = total number of observations

If there is little difference among treatments, we would expect MSA to be near zero because the treatment means  $y_j$  would be near the overall mean  $y$ . Thus, when  $F$  is near zero we would not expect to reject the hypothesis of equal group means. But how large must  $F$  be to convince the mean differ? Just as with Z-test or a t-test, we need a decision rule.

Decision Rule : The F-distribution is right-skewed distribution that starts at zero (F can not be negative since variances are sums

---

of squares) and has no upper limit (since the variances can be of any magnitude). For ANOVA the F-test is a right tailed test. For a given level of significance  $\alpha$ , we can use F-table to obtain the right-tail critical value of F.

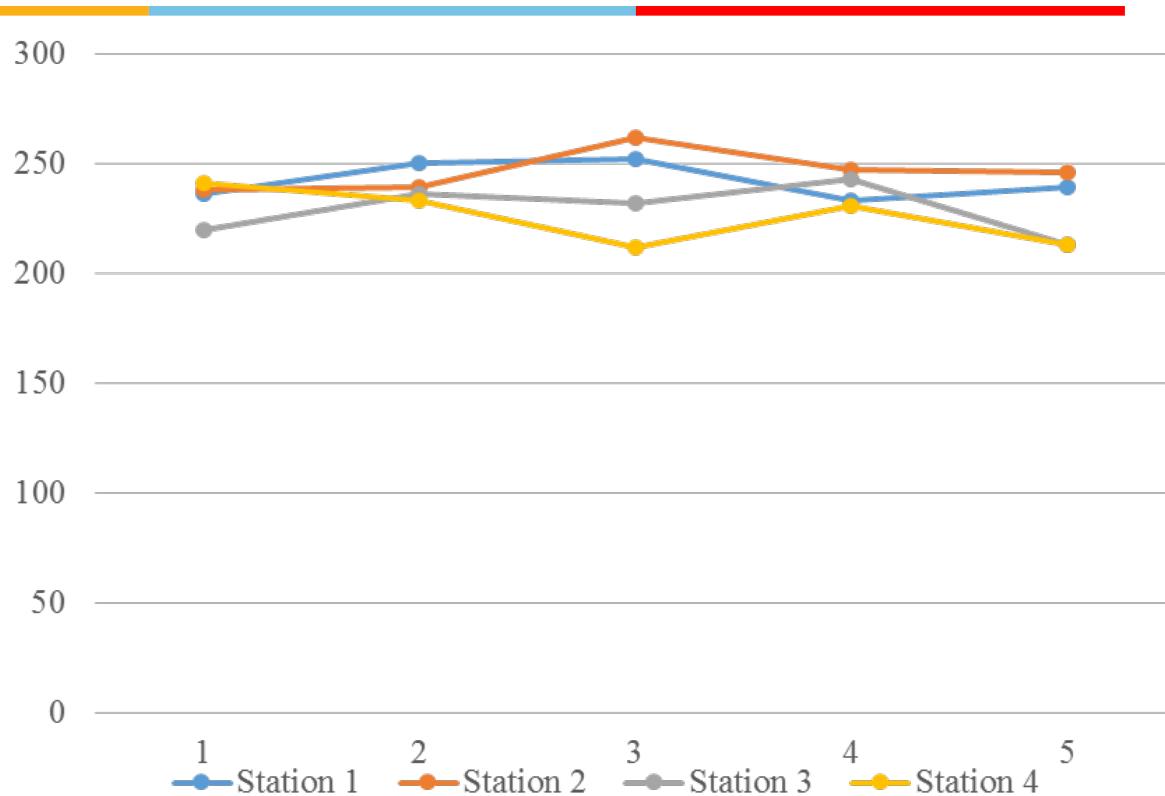
If  $F_{\text{calculated}} > F_{\text{critical}}$  value then the hypothesis is rejected and if  $F_{\text{calculated}} < F_{\text{critical}}$  the hypothesis can not be rejected.

---

**Ex1 :** A cosmetic manufacturer's regional distribution center has four workstations that are responsible for packing cartons for shipment to small retailers. Each workstation is staffed by 2 persons. The task involves assembling each order, placing it in shipping carton, inserting packing material, taping the carton and placing a computer generated shipping label on each carton. Generally, each station can pack 200 cartons a day and often more. However, there is variability, due to differences in orders, labels and cartons. Table below shows the no. of cartons packed per day during a recent week. Is the variation among stations within the range attributable to chance, or do these samples indicate actual differences in the means ?

	Station 1	Station 2	Station 3	Station 4
236	238	220	241	
250	239	236	233	
252	262	232	212	
233	247	243	231	
239	246	213	213	
Sum	1210	1232	1144	1130
Mean	242	246.4	228.8	226
Std deviation	8.515	9.607	12.153	12.884
n	5	5	5	5

Soln : As a preliminary step, we plot the data (as shown below) to check for any time pattern and just to visualize the data. We see some potential differences in means, but no obvious time Pattern (otherwise we would have to consider obsn order as a second factor). We proceed with the hypothesis test.



Step 1 : State the Hypothesis -

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$  (the means are same)

$H_1 : \text{Not all means are equal (at least one mean is different)}$

## Step 2 : State the decision rule -

There are  $c = 4$  groups and  $n = 20$  obsns, so degree of freedom for the F-test are –

Numerator :  $df_1 = c - 1 = 4 - 1 = 3$  (between treatments, factors)

Denominator :  $df_2 = n - c = 20 - 4 = 16$  (within treatments, error)

We will use  $\alpha = 0.05$  for the test. The 5% right tail critical value from the F-table is  $F_{0.05,3,16} = 3.24$ .

## Step 3 : Perform the calculations – As per table at significance level 0.05.

(Refer next slide)

Groups	Count	Sum	Average	Variance ( $\sigma^2$ )
Station 1	5	1210	242	72.5
Station 2	5	1232	246.4	92.3
Station 3	5	1144	228.8	147.7
Station 4	5	1130	226	166

Groups	Count	Sum	Average	Variance ( $\sigma^2$ )
Station 1	5	1210	242	72.5
Station 2	5	1232	246.4	92.3
Station 3	5	1144	228.8	147.7
Station 4	5	1130	226	166

Degree of freedom –  
 Between groups =  $c - 1 = 4 - 1 = 3$   
 Within groups =  $n - c = 20 - 4 = 16$   
 Total =  $n - 1 = 20 - 1 = 19$

### ANOVA Table

Source of variation	Sum of squares	Degree of freedom	Mean square	$F_{cal}$	P-value	F-critical
Between groups (SSA)	1479.2	3	493.0667	4.121769	0.024124	3.24
Within groups (SSE)	1914	16	119.625			
Total (SST)	3393.2	19				

---

## Step 4 : Make the decision –

Since the test statistic  $F = 4.12$  exceeds the critical value  $F_{.05} = 3.24$ , we can reject the hypothesis of equal means. Since excel gives the p-value, you don't actually need excel's critical value. The p-value (0.024124) is less than the the level of significance ( $\alpha = 0.05$ ) which confirms that we should reject the hypo of equal means. The same decision is arrived by Megastat's ANOVA table.

Note : Refer F-table and t-table

## Sample calculation

$\bar{y}_j = 242, 246.6, 228.8$  and  $226$

$$\sum_{j=1}^c \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2 = \sum_{j=1}^c n_j (\bar{y}_j - \bar{y})^2 + \sum_{j=1}^c \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

$$Y = (242 + 246.6 + 228.8 + 226) / 4 = 235.8$$

$$SSA = \sum_{j=1}^4 n_j (\bar{y}_j - \bar{y})^2, \quad df_1 = 4 - 1 = 3$$

$$SSE = \sum_{j=1}^4 \sum_{i=1}^5 (y_{ij} - \bar{y}_j)^2, \quad df_2 = 20 - 4 = 16$$

$$SST = \sum_{j=1}^4 \sum_{i=1}^5 (y_{ij} - \bar{y})^2$$

$$SSA = 5 \{(242 - 235.8)^2 + (246.6 - 235.8)^2 + (228.8 - 235.8)^2 + (226 - 235.8)^2\} = 1479.2$$

Ex 2: ANOVA table for 7 days's mean productivity of a manufacturing house at 4 different production centers is given in the following table –

Source of variation	Sum of squares
Between Production Centers (SSA)	6400
Error (SSE)	8600
Total (SST)	15000

Performing test of hypothesis at a significance level of 5% can we infer that the mean productivity at different production centers are same.

Soln :

Source of variation	Sum of squares	Degree of freedom	Mean square	$F_{\text{stat}}$	$F_{0.05, 3, 24}$
Between Prod Centers (SSA)	6400	$4 - 1 = 3$	2133.33	5.95	3.01
Error (SSE)	8600	$27 - 3 = 24$	358.33		
Total (SST)	15000	$4 \times 7 - 1 = 27$			

$F_{\text{critical}} = 3.01$  obtained from F-table. It is evident from the F-distribution (shown below) that  $F_{\text{stat}} > F_{\text{critical}}$ . Hence the inference drawn is - the mean productivity at different production centers are not the same.

Experimenter has to initiate corrective action to bring in uniformity in production at all the production centers.

**Ex 3:** In many integrated circuit manufacturing steps, wafers are completely coated with a layer of material such as silicon dioxide or metal. The unwanted material is then selectively removed by etching through a mask at different levels of watts of radio frequency power. It is interested to test whether the average etching rates of unwanted material at all 4 radio frequency (RF) levels are equal?

RF Power (W)	Observed Etch Rate (A/Min)				
	1	2	3	4	5
160	575	542	530	539	570
180	565	593	590	579	610
200	600	651	610	637	629
220	725	700	715	685	710

**Soln :**  $H_0$  : The average etching rates of unwanted material at all 4 radio frequency(RF) levels are equal  
 (vs)

$H_1$  : The average etching rates at all 4 radio frequency(RF) levels are not equal

RF Power (W)	Observed Etch Rate (A/Min)					Totals $y_i$	Averages $\bar{y}_i$
	1	2	3	4	5		
160	575	542	530	539	570	2756	551.2
180	565	593	590	579	610	2937	587.4
200	600	651	610	637	629	3127	625.4
220	725	700	715	685	710	3535	707.0
						$y_i = 12,355$	$\bar{y}_i = 617.75$

$$SST : \sum_{i=1}^4 \sum_{j=1}^5 (y_{ij})^2 - (y_{..})^2 / N$$

$$= (575)^2 + (542)^2 + \dots + (710)^2 - (12,355)^2 / 20 = 72,209.75$$

$$SSA = (1/n) \sum_{i=1}^4 y_{i..}^2 - y_{..}^2 / N = (1/5)[(2756)^2 + \dots + (3535)^2] - (12,355)^2 / 20 = 66,870.55$$

We will use the analysis of variance to test

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$  against the alternative  $H_1 : \text{Some means are different.}$

The sums of squares required are computed using Equations 3.8, 3.9 and 3.10 as follows –

$$SSE = SST - SSA = 72,209.75 - 66,870.55 = 5339.20$$

### ANOVA for the Plasma Etching Experiment

Source of variation	Sum of Square	Degree of Freedom	MS	$F_{cal} = (MSA / MSE)$
RF Power	66,870.55	3	$MSA = 22,290.18$	66.80
Error	5339.20	16	$MSE = 333.70$	
Total	72,209.75	19		

RF Power or between-treatment mean square (22,290.18) is many times larger than the within-treatment or error mean square (333.70). This indicates that it is unlikely that the treatment means are equal. More formally, we can compute the F ratio  $F_{cal} = 22,290.18/333.7 = 66.80$  and compare this to an appropriate upper-tail percentage point of the  $F_{3,16}$  distribution. To use a fixed significance level approach, suppose that the experimenter has selected  $\alpha = 0.05$ . From F-table we find that  $F_{0.05,3,16} = 3.24$ . Because  $F_{cal} (66.80) > F_{critical} (3.24)$ , we reject  $H_0$  and conclude the treatment means differ; that is, the RF Power setting significantly affects the mean etch rate.

**Ex 4:** Suppose 3 drying formulas for curing a glue are studied and the following times (in minutes) are observed –

Formula A	Formula B	Formula C
13	13	4
10	11	1
8	14	3
11	14	4
8		2
		4

Test whether mean curing times due to 3 different formulae are equal at 0.01 level of significance.

Soln :

$H_0$  : Mean curing times due to 3 different formulae are equal i.e.,  $\mu_1 = \mu_2 = \mu_3$

(vs)

$H_1$  : Mean curing times due to 3 different formulae are not equal

Curing Times of		Curing Times of		Curing Times of		$y_1^2$	$y_2^2$	$y_3^2$
A	B	C	$y_1$	$y_2$	$y_3$			
13	13	4	169	169	16			
10	11	1	100	121	1			
8	14	3	64	196	9			
11	14	4	121	196	16			
8		2	64		4			
		4						16
Total	50	52	18	518	682	62		

$$K = 3, n_1 = 5, n_2 = 4, n_3 = 6, n = 15, T = 50 + 52 + 18 = 120,$$

$$T(y^2) = 518 + 682 + 62 = 1262, SST = 1262 - (120^2 / 15) = 302,$$

$$SSA = (50^2/5) + (52^2/4) + (18^2/6) - (120^2/15) = 270, SSE = 302 - 270 = 32$$

### Anova Table

Source of variation	Sum of squares	Degree of freedom	Mean squares	$F_{\text{cal}}$ (MSA/MSE)	$F_{\text{critical}}$ $[F_{\alpha,(k-1),(n-k)}]$
Treatments (Formulae)	$\text{SSA} = 270$	$k - 1 = 3 - 1 = 2$	$\text{MSA} = \text{SSA}/(k - 1) = 270/2 = 135$	$135/2.667 = 50.6$	$F_{0.01,(2, 12)} = 6.93$
Error	$\text{SSE} = 32$	$n - k = 15 - 3 = 12$	$\text{MSE} = \text{SSE}/(n - k) = 32/12 = 2.667$		
Total	$\text{SST} = 302$	$n - 1 = 15 - 1 = 14$			

Conclusion : Since  $F_{\text{cal}} 50.6 > F_{\text{critical}} (6.93)$ ,  $H_0$  is rejected ie. Mean curing times due to 3 different formulae are not equal.

**Ex 5:** I belong to a golf club in my neighbourhood. I divide the year into three seasons – Summer (June – September), Winter (November – March) and Shoulder (October, April & May). I believe that I play my best golf during Summer (because I have more time and the course isn't crowded) and Shoulder (because the course isn't crowded) season and my worst golf is during the Winter (because when all of the part-year residents show up, the course is crowded, play is slow and I get frustrated). Data from the last year are shown in the following table.

Season	Observations									
Summer	83	85	85	87	90	88	88	84	91	90
Shoulder	91	87	84	87	85	84	83			
Winter	94	91	87	85	87	91	92	86		

Do the data indicate that my opinion is correct? Use  $\alpha = 0.05$ .

**Soln :**  $H_0$  : The average golf scores are equal between among three seasons  
 (vs)

$H_1$  : The average scores are not equal

From each observation 85 is subtracted then we obtain

**Summer :** -2, 0, 0, 2, 5, 3, 3, -1, 6, 5;  $T_1 = 21$ ,  $T_1^2 = 441$ ,  $T_1^2/10 = 44.1$

**Shoulder :** 6, 2, -1, 2, 0, -1, -2;  $T_2 = 6$ ,  $T_2^2 / 7 = 36/7 = 5.14$

**Winter :** 9, 6, 2, 0, 2, 6, 7, 1;  $T_3 = 33$ ,  $T_3^2 / 8 = 136.125$

$k = 3$ ,  $n = 10 + 7 + 8 = 25$ ,  $T = 21 + 6 + 33 = 60$ ,  $T^2/n = 60^2/25 = 144$

$$T(y^2) = (-2)^2 + (0)^2 + \dots + (7)^2 + (1)^2 = 374$$

$$SST = T(y^2) - T^2/n = 374 - 144 = 230$$

$$SSA = T(T_i^2/n_i) - T^2/n = 44.1 + 5.14 + 136.125 - 144 = 41.365$$

$$SSE = SST - SSA = 230 - 41.365 = 188.635$$

$$MSA = SSA/(k-1) = 41.365/2 = 20.6825,$$

$$MSE = SSE/(n-k) = 188.635/22 = 8.5743$$

$$F_{cal} = MSA / MSE = 20.365/8.5743 = 2.375$$

$$F_{((k-1) = 2, (n-k) = 22)} \text{ at 0.05 level of significance is 3.44}$$

ie.  $F_{0.05, 2, 22} = 3.44$

Since  $F_{cal} < F_{critical}$  then we accept  $H_0$ .

Source of variation	SS	DF	MS	$F_{cal}$	$F_{critical}$
Treatment (Between Seasons)	$SSA = 41.365$	$k - 1 = 2$	$MSA = 20.6825$	$MSA/MSE = 2.412$	3.44
Error	$SSE = 188.635$	$n - k = 22$	$MSE = 8.5743$		
Total	$SST = 230$	$n - 1 = 24$			

Conclusion : Since  $F_{cal}$  (2.412)  $<$   $F_{critical}$  (3.44), accept the Null Hypothesis ie. golf scores are equal in all the three seasons.

## Practice Problems

- Q1. Highlight the difference between one-way and two-way Anova.
- Q2. Conduct test of hypothesis at a significance level of 5% to infer whether mean productivity at different work stations are same. Analysis of variance for the Average productivity for 7 days at 5 different workstations is tabulated below.

Source of variation	Sum of squares
Between Production Centers (SSA)	4650
Error (SSE)	13730

- Q3. Analysis of Variance for the mean defects of six days' production at four different production centers is given in the following table.

Source of variation	Sum of squares
Between Production Centers (SSA)	2800
Error (SSE)	3900

## Practice Problems

Q4. Three different kinds of food are tested on three groups of rats for 5 weeks. The objective is to check the difference in mean weight (in grams) of the rats per week. Apply one-way ANOVA using a 0.05 significance level to the following data –

Food I	Food II	Food III
8	4	11
12	5	8
19	4	7
8	6	13
6	9	7
11	7	9



# BSDCHZC355

## Statistical Inference & Applications

**BITS** Pilani  
Pilani | Dubai | Goa | Hyderabad

Shaibal Kr. Sen  
Session 05



**STATISTICAL INFERENCES**

**&**

**APPLICATIONS**

## Scheduled Topics to be covered

Two –way ANOVA Classification

---

A two-way ANOVA tests the effect of two independent variables on a dependent variable on the expected outcome along with their relationship to the outcome itself. Random factors would be considered to have no statistical influence on a data set, while systematic factors would be considered to have statistical significance. Whereas a one-way ANOVA is a type of statistical test that compares the variance in the group means within a sample whilst considering only one independent variable or factor. It is a hypothesis-based test, meaning that it aims to evaluate multiple mutually exclusive theories about our data.

By using ANOVA, a researcher is able to determine whether the variability of the outcomes is due to chance or to the factors in the analysis. ANOVA has many applications in Finance, Economics, Science, Engineering, DOE, Medicine, and Social Science.

---

---

An ANOVA test is the first step in identifying factors that influence a given outcome. Once an ANOVA test is performed, a tester may be able to perform further analysis on the systematic factors that are statistically contributing to the data set's variability.

Analysis of variances is helpful for testing the effects of variables on one another. It is similar to multiple two-sample t-tests. However, it results in fewer type 1 errors and is appropriate for a range of issues. An ANOVA test groups differences by comparing the means of each group and includes spreading out the variance across diverse sources. It is employed with subjects, test groups, between groups and within groups.

One-Way ANOVA -vs- Two-Way ANOVA : There are two main types of analysis of variance - one-way (or unidirectional) and two-way (bidirectional). One-way or two-way refers to the number of

---

independent variables in your analysis of variance test. A one-way ANOVA evaluates the impact of a sole factor on a sole response variable. It determines whether the observed differences between the means of independent (unrelated) groups are explainable by chance alone, or whether there are any statistically significant differences between groups.

A two-way ANOVA is an extension of the one-way ANOVA. With a one-way, you have one independent variable affecting a dependent variable. With a two-way ANOVA, there are two independents. For example, a two-way ANOVA allows a company to compare worker productivity based on two independent variables, such as department and gender. It is utilized to observe the interaction between the two factors. It tests the effect of two factors at the same time.

---

---

A three-way ANOVA, also known as three-factor ANOVA, is a statistical means of determining the effect of three factors on an outcome.

The only difference between one-way and two-way ANOVA is the number of independent variables. A one-way ANOVA has one independent variable, while a two-way ANOVA has two. Examples of -

**One-way ANOVA** - Testing the relationship between shoe brand (Nike, Adidas, Saucony, Hoka) and race finish times in a marathon.

**Two-way ANOVA** - Testing the relationship between shoe brand (Nike, Adidas, Saucony, Hoka), runner age group (junior, senior, master's), and race finishing times in a marathon.

All ANOVAs are designed to test for differences among three or more groups. If you are only testing for a difference between two groups, use a t-test instead.

---

As we have seen earlier Two-way is an extension of the one way ANOVA in terms of taking the second factor into the analysis.

One-way or two-way refers to the number of independent variables in the experiment.

It is important to note that two-way ANOVA test is used when the number of observations in the subclasses are equal.

One-way ANOVA

A	23	21	45	42	
B	25	24	36		
C	28	27	33	40	36

Two-way ANOVA

Factor 2 ↓ Factor 1	P	Q	R	S
A	23	21	45	42
B	25	24	36	23
C	28	27	33	40

Sum of the Squares (SS) -

Correction Factor (CF) =  $G^2/N$ ; where, G is grand total, N is total number of elements.

(1) SS between Rows =  $\sum\{(T_i)^2/n_i\} - CF$ ; where,  $T_i$  is total of  $i^{\text{th}}$  row.

(2) SS between columns =  $\sum\{(T_j)^2/n_j\} - CF$ ; where,  $T_j$  is total of  $j^{\text{th}}$  column.

(3) SS Total =  $\sum\sum(x_{ij})^2 - CF$

(4) Within Error = (3) - (1) - (2)

**Ex 1 :** Three varieties of coal were analysed by four chemists and the ash contents in the varieties were found as tabulated below -

Varieties	Chemists			
	I	II	III	IV
A	8	5	5	7
B	7	6	4	4
C	3	6	5	4

Discuss the significance of the difference - (i) Chemists and (ii) Varieties of coal in respect of Ash Content.

Test at 5% levcl of significance

**Soln :** For Varieties -  $H_{0V} : \mu_V$  ie.  $\mu_A = \mu_B = \mu_C = \mu$

$H_{1V} : \mu_V \neq \mu$

For Chemists -  $H_{0C} : \mu_C$  ie.  $\mu_I = \mu_{II} = \mu_{III} = \mu_{IV} = \mu$

$H_{1C} : \mu_C \neq \mu$

Soln :

Source of Variation	DF	SS	MSS = SS/DF	F-ratio
Between Varieties	$3 - 1 = 2$			
Between Chemists	$4 - 1 = 3$			
Within Error	$(3-1)(4 - 1) = 6$			
Total	$3 \times 4 - 1 = 11$			

Varieties	Chemists				Total
	I	II	III	IV	
A	8	5	5	7	25
B	7	6	4	4	21
C	3	6	5	4	18
Total	18	17	14	15	$G = 64$

Correction Factor (CF) =  $G^2/N = 64^2/3 \times 4 = 4096/12 = 341.33$

$$SS \text{ Total} = \sum \sum (x_{ij})^2 - CF = 8^2 + 5^2 + \dots + 5^2 + 4^2 - 341.33 \approx 24.667$$

$$SS \text{ between Varieties} = \sum \{(T_i)^2/n_i\} - CF = (25^2 + 21^2 + 18^2)/4 - 341.33$$

$$\approx 6.167$$

$$SS \text{ between Chemists} = \sum \{(T_j)^2/n_j\} - CF$$

$$= (18^2 + 17^2 + 14^2 + 15^2)/3 - 341.33 \approx 3.333$$

$$SSError = SS \text{ Total} - SS \text{ between Varieties} - SS \text{ between Chemists}$$

$$= 24.667 - 6.167 - 3.333 = 15.167$$

### ANOVA Table

Source of Variation	DF	SS	MSS = SS/DF	$F_{\text{calculated}}$	$F_{\text{critical}}$ (From F-table)
Between Varieties	$3 - 1 = 2$	6.167	$6.167/2 = 3.083$	$\text{MSV/MSE} = 1.219$	$F_{0.05,2,6} = 5.14$
Between Chemists	$4 - 1 = 3$	3.333	$3.333/3 = 1.111$	$\text{MSC/MSE} = 0.439$	$F_{0.05,3,6} = 4.76$
Within Error	$(3-1)(4 - 1) = 6$	15.167	$15.167/6 = 2.528$		
Total	$3 \times 4 - 1 = 11$	24.667			

---

Since  $F_{cal}$  (1.219) <  $F_{critical}$  (5.14) for Varieties we can not reject the null hypothesis ie. there is no significant difference between the Varieties.

Since  $F_{cal}$  (0.439) <  $F_{critical}$  (4.76) for Chemists we can not reject the null hypothesis ie. there is no significant difference between the Chemists.

**Ex 2 :** Complete the following Anova Table assuming  $\alpha = 0.05$  -

Source of Variation	DF	SS	MSS = SS/DF	F <sub>calculated</sub>	F <sub>critical</sub> (From F-table)
Between Blocks	4	26.8			
Between Treatments	3		2.5		
Within Error					
Total		85.3			

**Soln :** From the information given in the table, we find there are two variables viz. – Blocks & Treatments. Hence, it is a two-way ANOVA with Number of Blocks = DF + 1 = 4 + 1 = 5 and Number of Treatments = DF + 1 = 3 + 1 = 4.

We know in a two-way ANOVA number of elements in the blocks (rows) & treatments (columns) are same ie. 5 & 4 respectively. Hence, total DF =  $5 \times 4 - 1 = 19$ .

Source of Variation	DF	SS	MSS = SS/DF	F <sub>calculated</sub>	F <sub>critical</sub> (From F-table)
Between Blocks	4	26.8	26.8/4 = 6.7	6.7/2.5 = 2.68	F <sub>0.05, 4, 12</sub> = 3.26
Between Treatments	3	28.5	28.5/3 = 9.5	9.5/2.5 = 3.8	F <sub>0.05, 3, 12</sub> = 3.49
Within Error	$19 - (4 + 3)$ = 12	x = 30	2.5		
Total	$5 \times 4 - 1 =$ 19	85.3			

Soln : From the information given in the table, we find there are two variables viz. – Blocks & Treatments. Hence, it is a two-way ANOVA with Number of Blocks = DF + 1 = 4 + 1 = 5 and Number of Treatments = DF + 1 = 3 + 1 = 4.

We know in a two-way ANOVA number of elements in the blocks (rows) & treatments (columns) are same ie. 5 & 4 respectively. Hence, total DF = 5x4 – 1 =19.

So, DF of Treatments =  $19 - (4 + 3) = 12$ .

Source of Variation	DF	SS	MSS = SS/DF	F <sub>calculated</sub>	
Between Blocks	5	32.7	$32.7/5 = 6.54$		
Between Treatments	4	$85.3 - (32.7 + 20) = 32.6$	$32.6/4 = 8.15$		
Within Error	$29 - (5 + 4) = 20$	$20 \times 1 = 20$	1		
Total	$6 \times 5 - 1 = 29$	85.3			

Let SS Treatments be x. As per given data  $x/12 = 2.5$  ie.  $x = 30$

Since, SS Total = SS Blocks + SS Treatments + SS Error

or,  $85.3 = 26.8 + \text{SS Treatments} + 30$  Hence, SS Treatments =  $85.3 - 56.8 = 28.5$

Source of Variation	DF	SS	MSS = SS/DF	F <sub>calculated</sub>	F <sub>critical</sub> (From F-table)
Between Blocks	4	26.8	$26.8/4 = 6.7$	$6.7/2.5 = 2.68$	$F_{0.05, 4, 12} = 3.26$
Between Treatments	3	28.5	$28.5/3 = 9.5$	$9.5/2.5 = 3.8$	$F_{0.05, 3, 12} = 3.49$
Within Error	$19 - (4 + 3) = 12$	$x = 30$	2.5		
Total	$5 \times 4 - 1 = 19$	85.3			

Since, for blocks  $F_{cal} (2.68) < F_{critical} (3.26)$  null hypothesis can not be rejected but for treatments  $F_{cal} (3.8) > F_{critical} (3.49)$  null hypothesis can not be accepted.

---

**Ex 3 :** A medical device manufacturer produces vascular grafts (artificial veins). These grafts are produced by extruding billets of polytetrafluoroethylene (PTFE) resins combined with a lubricant into tubes. Frequently, some of the tubes in a production run contain small, hard protrusions on the external surface. These defects are known as “flicks”. The defect is cause for rejection of the unit.

The product developer responsible for the vascular grafts suspects that the extrusion pressure affects the occurrence of flicks and therefore intends to conduct an experiment to investigate this hypothesis. However, the resin is manufactured by an external supplier and is delivered to the medical device manufacturer in batches. The engineer also suspects that there may be significant batch-to-batch variation.

Please analyse and infer your finding at 5% LOS.

		$F_2$					
		Batch of Resin (Block)					
		1	2	3	4	5	6
$F_1$	8500	90.3	89.2	98.2	93.9	87.4	97.9
	8700	92.5	89.5	90.6	94.7	87.0	95.8
	8900	85.5	90.8	89.6	86.2	88.0	93.4
	9100	82.5	89.5	85.6	87.4	78.9	90.7

Soln :

		$F_2$						Treatment Total	
		Batch of Resin (Block) [q or b = 6]							
		1	2	3	4	5	6		
$F_1$ [p or a = 4]	8500	90.3	89.2	98.2	93.9	87.4	97.9	556.9	
	8700	92.5	89.5	90.6	94.7	87	95.8	550.1	
	8900	85.5	90.8	89.6	86.2	88	93.4	533.5	
	9100	82.5	89.5	85.6	87.4	78.9	90.7	514.6	
Block Totals		350.8	359	364	362.2	341.3	377.8	y.. = 2155.1	

Null Hypothesis  $H_0 F_1$  : All extrusion pressures equally affect the mean yield.

$H_0 F_2$  : All batches of resins are equally effective.

To perform analysis of variance, we need the following sums of squares –

$$SSTotal = \sum_{i=1}^4 \sum_{j=1}^6 (y_{ij})^2 - (y_{..})^2/N = 193,999.31 - (2155.1)^2/24 = 480.31$$

$$SSTreatment = (1/b) \left\{ \sum_{i=1}^4 (y_{i.})^2 \right\} - (y_{..})^2/N$$

$$= (1/6) \left\{ (556.9)^2 + (550.1)^2 + (533.5)^2 + (514.6)^2 \right\} - (2155.1)^2/24 = 178.17$$

$$SSBlocks = (1/a) \left\{ \sum_{j=1}^6 (y_{.j})^2 \right\} - (y_{..})^2/N$$

$$= (1/4) \left\{ (350.8)^2 + (359.0)^2 + \dots + (377.8)^2 \right\} - (2155.1)^2/24 = 192.25$$

$$SSE = SSTotal - SSTreatment - SSBlocks = 480.31 - 178.17 - 192.25 = 109.89$$

The ANOVA table and inferences are shown in the next slide.

<u>ANOVA Table for Vascular Graft Experiment</u>					
Sources of Variation	Sum of Squares (SS)	Degrees of Freedom (DF)	Mean Square (SS/DF)	$F_{cal}$	P-value
Treatments (Extrusion Pressure)	178.17	$4 - 1 = 3$	$178.17/3 = 59.39$	$59.39/7.33 = 8.10$	0.0019 $< 0.05$
Blocks (Batches)	192.25	$6 - 1 = 5$	$192.25/5 = 38.45$	$38.45/7.33 = 5.24$	0.005 $< 0.05$
Error	109.89	15	$109.89/15 = 7.33$		
Total	480.31	$4 \times 6 - 1 = 23$			

Conclusion : Using  $\alpha = 0.05$ , the critical value of F ie.  $F_{0.05,3,15} = 3.29$  &  $F_{0.05,5,15} = 2.9$  . Since the calculated values of F for both extrusion pressure (8.10) and resin batch effect (5.24) are beyond the acceptance region  $F_{critical}$  values (3.29 & 2.9), we conclude that null hypothesis can not be accepted ie. extrusion pressure affects the mean yield and resin batches seem to differ significantly. It is also evident from the small P-value ( $< \alpha = 0.05$ ).

Ex 4 :

Solve using Two-way ANOVA method

Observation	A	B	C
1	1,4,0,7	13,5,7,15	9,16,18,13
2	15,6,10,13	6,18,9,15	14,7,6,13

Soln :

Row and column sums

	A	B	C	Row total ( $x_a$ )
1	12	40	56	108
2	44	48	40	132
Col total ( $x_b$ )	56	88	96	240

$$\sum x^2 = 1^2 + 4^2 + 0^2 + \dots + 7^2 + 6^2 + 13^2 = 3010 \dots (A)$$

$$\sum (x_b)^2 / ra = (56^2 + 88^2 + 96^2) / (4 \times 2) = (3136 + 7744 + 9216) / 8 = 20096 / 8 = 2512 \dots (B)$$

$$\sum(x_a)^2/rb = (108^2 + 132^2)/4 \times 3 = (11664 + 17424)/12 = 29088/12 = 2424 \dots (C)$$

$$\sum\sum(x_{ab})^2/r = (12^2 + 40^2 + 56^2 + 44^2 + 48^2 + 40^2)/4$$

$$= (144 + 1600 + 3136 + 1936 + 2304 + 1600)/4 = 10720/4 = 2680$$

$$(\sum x)^2/rab = (240)^2/(4 \times 2 \times 3) = 57600/24 = 2400 \dots (D)$$

Sum of squares total

$$SST = \sum x^2 - (\sum x)^2/n = (A) - (D) = 3010 - 2400 = 610$$

Sum of squares between rows

$$SSA = \sum(x_a)^2/rb - (\sum x)^2/rab = (C) - (D) = 2424 - 2400 = 24$$

Sum of squares between columns

$$SSB = \sum(x_b)^2/ra - (\sum x)^2/n = (B) - (D) = 2512 - 2400 = 112$$

Sum of squares between columns

$$SSAB = \sum\sum(x_{ab})^2/r - (\sum x)^2/rab - SSA - SSB = 2680 - 2400 - 24 - 112 = 144$$

$$= 2680 - 2400 - 24 - 112 = 144$$

Sum of squares Error (residual)

$$SSE = SST - SSA - SSB - SSAB = 610 - 24 - 112 - 144 = 330$$

ANOVA Table

Source of Variation	Sum of Squares (SS)	Degree of freedom (DF)	Mean Square (MS)	$F_{cal}$	P-value
A	$SSA = 24$	$a - 1 = 1$	$MSR = 24/1 = 24$	$24/18.333 = 1.3091$	0.2675
B	$SSB = 112$	$b - 1 = 2$	$MSC = 112/2 = 56$	$56/18.333 = 3.0545$	0.0721
AB	$SSAB = 144$	$(a - 1)(b - 1) = 2$	$MSAB = 144/2 = 72$	$72/18.333 = 3.9273$	0.0384
Error (Residual)	$SSE = 330$	$rab - ab = 18$	$MSE = 330/18 = 18.333$		
Total	$SST = 610$	$rab - 1 = 23$			

---

Conclusion :

1. F for between columns for df 1,2 and LOS  $\alpha = 0.05$   
ie.  $F_{0.05,1,2} = 4.4139$

As calculated  $F_{\text{cal}}$  (between rows) = 1.3091 < 4.4139

So,  $H_0$  is accepted, Hence there is no significant difference between rows.

2. F for between columns for df 2, 2 and LOS  $\alpha = 0.05$   
 $F_{0.05,2,2} = 3.5546$

As calculated  $F_{\text{Cal}}$  (between columns) = 3.0545 < 3.5546

So,  $H_0$  is accepted, Hence there is no significant difference between columns.

---

## Ex 5 : Solve using Two-way ANOVA method

Observation	A	B	C	D	E	F
1	1200	1000	980	900	750	800
2	1000	1100	700	800	500	700
3	890	650	1100	900	400	350

Soln :

	A	B	C	D	E	F	Row total ( $x_r$ )
1	1200	1000	980	900	750	800	5630
2	1000	1100	700	800	500	700	4800
3	890	650	1100	900	400	350	4290
Col total ( $x_c$ )	3090	2750	2780	2600	1650	1850	14720

$$\sum x^2 = 13010000 \dots (A)$$

$$\begin{aligned}\sum (x_c)^2 / r &= (3090^2 + 2750^2 + 2780^2 + 2600^2 + 1650^2 + 1850^2) / 3 \\ &= (9548100 + 7562500 + 7728400 + 6760000 + 2722500 + 3422500) / 3 = 937744000 / 3 \\ &= 12581333.3333 \dots (B)\end{aligned}$$

$$\begin{aligned}\sum (x_r)^2 / c &= (5630^2 + 4800^2 + 4290^2) / 6 = (31696900 + 23040000 + 18404100) / 6 \\ &= 73141000 / 6 = 12190166.67 \dots (C) \\ &= (\sum x)^2 / n = (14720)^2 / 18 = 216678400 / 18 = 12037688.8889 \dots (D)\end{aligned}$$

Sum of squares total

$$SST = \sum x^2 - (\sum x)^2 / n = (A) - (D) = 13010000 - 12037688.89 = 972311.11$$

Sum of squares between rows

$$SSR = \sum (x_r)^2 / c - (\sum x)^2 / n = (C) - (D) = 12190166.67 - 12037688.89 = 152477.78$$

## Sum of squares between columns

$$SSC = \sum(x_c)^2/r - (\sum x)^2/n = (B) - (D) = 12581333.33 - 12037688.89 = 543644.44$$

## Sum of squares Error (residual)

$$SSE = SST - SSR - SSC = 972311.11 - 152477.78 - 543644.44 = 276188.89$$

ANOVA table

Source of Variation	Sums of Squares (SS)	Degrees of freedom (DF)	Mean Squares (MS)	$F_{cal}$
Between rows	$SSR = 152477.78$	$r - 1 = 2$	$MSR = 152477.78/2 = 76238.89$	$76238.89/27618.89 = 2.7604$
Between columns	$SSC = 543644.44$	$c - 1 = 5$	$MSC = 543644.44/5 = 108728.88$	$108728.89/27618.89 = 3.9368$
Error (residual)	$SSE = 276188.89$	$(r - 1)(c - 1) = 10$	$MSE = 276188.89/10 = 27618.89$	
Total	$SST = 972311.11$	$rc - 1 = 17$		

## Example of One way ANOVA

Ex 6 : ANOVA table for 7 days's mean productivity of a manufacturing house at 4 different production centers (pc) is given in the following table –

Source of variation	Sum of squares
Between Production Centers (SSA)	6400
Error (SSE)	8600
Total (SST)	15000

Performing test of hypothesis at a significance level of 5% can we infer that the mean productivity at different production centers are same.

Soln :  $H_0$  : Mean Productivity of all the four Centers are the same.

$H_1$  : Mean Productivity of all the four Centers are not the same.

Source of variation	Sum of squares	Degree of freedom	Mean square	$F_{\text{stat}}$	$F_{0.05, 3, 24}$
Between Prod Centers (SSA)	6400	$c - 1 = 4 - 1 = 3$	$6400/3 = 2133.33$	MSA/MSE = 5.95	3.01
Error (SSE)	8600	$n - c = 4 \times 7 - 4 = 24$	$8600/24 = 358.33$		
Total (SST)	15000	$4 \times 7 - 1 = 27$			

$F_{\text{critical}} = 3.01$  obtained from F-table. It is evident from the F-distribution (shown below) that  $F_{\text{stat}} > F_{\text{critical}}$ . Hence the inference drawn is - the mean productivity at different production centers are not the same.

Experimenter has to initiate corrective action to bring in uniformity in production at all the production centers.

## Practice Problems

Q1. A study examining differences in life satisfaction between young adult, middle adult, and older adult men and women was conducted. Each individual who participated in the study completed a life satisfaction questionnaire. A high score on the test indicates a higher level of life satisfaction. Test scores are recorded below.

Group	Young Adult	Middle Adult	Older Adult
Male	4	7	10
	2	5	7
	3	7	9
	4	5	8
	2	6	11
Female	7	8	10
	4	10	9
	3	7	12
	6	7	11
	5	8	13

contd.

---

(i) Complete the following ANOVA table.

<u>Source</u>	<u>SS</u>	<u>df</u>	<u>MS</u>	<u>F</u>
Age	180	2		
Gender	30	1		
Age x Gender	0	2		
<u>Within</u>	<u>44</u>	<u>24</u>		
Total				

- (ii) Are there any significant main effects or an interaction effect.  
(iii) Interpret your answer.
-

Q2. Perform a 2-way ANOVA on the data given below -

Teachers	Students				
	I	II	III	IV	V
A	30	24	33	36	27
B	26	29	24	31	35
C	38	28	35	30	35

- (i) Shift the origin to 30. Perform the ANOVA for the transformed data.
- (ii) How do the results compare with those obtained for the original data?

Q3. A research study was conducted to examine the impact of eating a high protein breakfast on adolescents' performance during a physical education physical fitness test. Half of the subjects received a high protein breakfast and half were given a low protein breakfast. All of the adolescents, both male and female, were given a fitness test with high scores representing better performance. Test scores are recorded below.

Group	High Protein	Low Protein
Males	10	5
	7	4
	9	7
	6	4
	8	5
Females	5	3
	4	4
	6	5
	3	1
	2	2

contd.

(i) Complete the following table -

Source	SS	df	MS	F
Protein Level	20	1		
Gender	45	1		
Protein Level x Gender	5	1		
Within	36	16		
Total				

- (ii) Are there any significant main effects or an interaction effect.  
 (iii) Interpret your answer.

Q4. Three different kinds of food are tested on three groups of rats for 5 weeks. The objective is to check the difference in mean weight (in grams) of the rats per week. Apply one-way ANOVA using a 0.05 significance level to the following data -

Food I	Food II	Food III
8	4	11
12	5	8
19	4	7
8	6	13
6	9	7
11	7	9

Q5. On the basis of following information, complete the ANOVA Table and solve using Two-way ANOVA method

Observation	A	B	C
1	10, 8, 7, 9, 6	7, 4, 3, 2	11, 9, 10, 9, 11
2	1, 2, 1, 4, 2	6, 7, 6, 5	4, 3, 6, 4, 3
3	3, 2, 3, 3, 4	2, 1, 2, 3	5, 6, 4, 5, 5

ANOVA Table

Source of Variation	Sums of Squares (SS)	Degrees of freedom (DF)	Mean Squares (MS)	F
A	$SSA = 145.33$			
B	$SSB = 45.24$			
AB	$SSAB = 93.33$			
Error (residual)	$SSE = 48$			
Total				



# BSDCHZC355

## Statistical Inference & Applications

**BITS** Pilani  
Pilani | Dubai | Goa | Hyderabad

Shaibal Kr. Sen  
Session 06



**STATISTICAL INFERENCES**

**&**

**APPLICATIONS**

## Scheduled Topics to be covered

---

Design of Experiments : Complete Block Designs

Completely Randomized Design (CRD)

Randomized Block Design (RBD)

## Completely Randomized Design (CRD)

The simplest design using the two essential principles of replication and randomization is the CRD. Suppose that we have  $t$  treatments (or  $t$  levels of a factor) under comparison and the  $i^{\text{th}}$  treatment is to be replicated  $r_i$  times, for  $i = 1, 2, \dots, t$ . Then the total number of experimental units necessary for this experiment is  $n = \sum_{i=1}^t r_i$ . In the CRD, we allocate the  $t$  treatments completely at random to the  $n$  units subject to the condition that the  $i^{\text{th}}$  treatment appears in  $r_i$  units, for  $i = 1, 2, \dots, t$ . A Particular case of this equal replication for different treatments, where  $r_1 = r_2 = \dots = r_t = r$ , so that  $n = rt$ .

---

## Description of Design -

Simplest design to use.

Design can be used when experimental units are essentially homogeneous.

Because of the homogeneity requirement, it may be difficult to use this design for field experiments.

The CRD is best suited for experiments with a small number of treatments.

## Randomization Procedure -

Treatments are assigned to experimental units completely at random. Every experimental unit has the same probability of receiving any treatment.

Randomization is performed using a random number table, computer programme etc.

Example of Randomization : Given that there are 4 treatments (A, B, C and D) and there are 5 replicates, how many experimental units would you have?

1	2	3	4	5	6	7	8	9	10
D	D	B	C	D	C	A	A	B	D
11	12	13	14	15	16	17	18	19	20
C	B	A	B	C	B	C	D	A	A

Note that there is no “blocking” of experimental units into replicates.

Every experimental unit has the same probability of receiving any treatment.

---

## Advantages of a CRD –

1. Very flexible design (ie. number of treatments & replications is only limited by the available number of experimental units).
2. Statistical analysis is simple compared to other designs.
3. Loss of information due to missing data is small compared to other designs due to the larger number of degrees of freedom for the error source of variation.

## Disadvantages –

1. If experimental units are not homogeneous and you fail to minimize this variation using blocking, there may be a loss of precision.
2. Usually the least efficient design unless experimental units are homogeneous.
3. Not suited for large number of treatments.

The analysis of CRD is exactly same as that of One-Way ANOVA

### Linear Additive Model for CRD -

$$Y_{ij} = \mu + \tau_i + e_{ij} \text{ where,}$$

$Y_{ij}$  is the  $j^{\text{th}}$  observation of  $i^{\text{th}}$  treatment,

$\mu$  is the population mean,

$\tau_i$  is the treatment effect of  $i^{\text{th}}$  treatment and

$e_{ij}$  is random error with mean 0 and variance  $\sigma^2$ .

After analysis it becomes –

$$\text{SS Total} = \text{SS due to treatment} + \text{SS due to error}$$

$$(S_{\text{total}})^2 = (S_{\text{Treatment}})^2 + (S_E)^2$$

We are interested in testing  $H_0 : \tau_1 = \tau_2 = \dots = \tau_t$  against the alternatives that  $\tau$ 's are not all equal. The analysis in the present case is the same as that of one-way classified data learnt earlier (in Session 4 & 5).

ANOVA Table for CRD

Sources of Variation	Sum of Square (SS)	Degree of Freedom (d.f.)	Mean Sum of Square (MSS)	Variation Ratio ( $F_{cal}$ )
Between Treatments	$(S_{Treatment})^2$	$t - 1$	$MS_{Treatment} = MS_T = (S_{Treatment})^2 / (t - 1)$	$MS_T / MS_E$
Error	$(S_{Error})^2$	$n - t$	$MS_{Error} = MS_E = (S_{Error})^2 / (n - 1)$	
Total	$(S_{Total})^2$	$n - 1$		

We reject  $H_0$  at the LOS  $\alpha$  if  $F_{cal} [MS_T / MS_E] > F_{critical} [F_{\alpha, (t-1), (n-t)}]$ ; otherwise  $H_0$  is accepted. When  $H_0$  is rejected, we may decide further course of action.

Ex 1 : Given the following data, test whether mean effects due to the treatments are equal.

<u>Treatment</u>			
A	B	C	D
2.0	1.7	2.0	2.1
2.2	1.9	2.4	2.2
1.8	1.5	2.7	2.2
2.3		2.5	1.9
1.7		2.4	

Soln : Step 1

Hypothesis to be tested  $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$  - vs -

Alternative Hypothesis  $H_1$  : At least one of the means is different

Replicate	<u>Treatment</u>			
	A	B	C	D
1	2.0	1.7	2.0	2.1
2	2.2	1.9	2.4	2.2
3	1.8	1.5	2.7	2.2
4	2.3		2.5	1.9
5	1.7		2.4	
Total	10.0	5.1	12.0	8.4
$\Sigma(Y_{ij})^2$	20.26	8.75	29.06	17.7

$G = Y_{..} = 35.5$

## Step 2

Calculate the Correction Factor (CF) =  $G^2/N = (Y_{..})^2/\Sigma r_i = 35.5^2/17 = 74.132$

## Step 3

Calculate the Total SS

$$\begin{aligned}
 SS_{\text{Total}} &= \Sigma(Y_{ij})^2 - CF = (2.0^2 + 2.2^2 + 1.8^2 + \dots + 1.9^2) - CF = 75.77 - 74.132 \\
 &= 1.638
 \end{aligned}$$

## Step 4

$$\text{Calculate } SS_{\text{Treatment}} = \sum (Y_{i.})^2 / r_i - CF$$

$$= \{(10)^2/5 + (5.1)^2/3 + (12)^2/5 + (8.4)^2/4\} - 74.132 = 0.978$$

## Step 5

$$\text{Calculate } SS_{\text{Error}} = SS_{\text{Total}} - SS_{\text{Treatment}} = 1.638 - 0.978 = 0.66$$

## Step 6

ANOVA Table for CRD

Sources of Variation	Sum of Squares (SS)	Degree of Freedom (d.f.)	Mean Sum of Square (MSS)	Variation Ratio ( $F_{\text{cal}}$ )
Between Treatments	0.978	$t - 1 = 4 - 1 = 3$	$MS_{\text{Treatment}} = 0.978/3 = 0.326$	$MS_T/MS_E = 6.42$
Error	0.660	$n - t = 17 - 4 = 13$	$MS_{\text{Error}} = 0.660/13 = 0.05076$	
Total	1.638	$n - 1 = 17 - 1 = 16$		

## Step 7

We find the following  $F_{critical}$  values from F-table

$$F_{0.05,3,13} = 3.41 \quad \text{and} \quad F_{0.01,3,13} = 5.74$$

## Step 8

Compare  $F_{cal}$  with  $F_{critical}$  to draw inference –

- (i) Since  $F_{cal} (6.42) > F_{critical} (3.41)$  at 95% level of confidence we reject the null hypothesis.
- (ii) Since  $F_{cal} (6.42) > F_{critical} (5.74)$  at 99% level of confidence we reject the null hypothesis.

## Randomized Block Design (RBD)

Suppose we want to compare the effects of  $t$  treatments, each treatment being replicated an equal number of times, say  $r$  times. Then we need  $n = rt$  experimental units, and these units are not perhaps homogeneous. The RBD consists of two steps. The first step is to divide the units into  $r$  more or less homogeneous groups. In each group or block we take as many units as there are treatments. Thus the number of blocks are the same as the common replication number ( $r$ ). The same technique should be applied to the units of the block. Variation in technique, if any, should be made between the blocks. In agricultural field experiments sometimes a fertility

The second step is to assign the treatments at random to the units of a block. This is the difference of RBD from CBD. In an RBD randomization is restricted within a homogeneous block.

With this design each treatment will have the same number of

## Randomized Block Design (RBD)

replication. If we want additional replications for some treatments, each of these may be applied to more than one unit in a block.

Let us consider five treatments – A, B, C, D and E replicated 4 times. We divide the whole experimental area into four relatively homogeneous strata or blocks and each block into 5 units or plots, treatments are then allocated at random to the plots of a block, fresh randomization done for each block. A particular layout may be as follows –

Block I	A	E	B	D	C
Block II	E	D	C	B	A
Block III	C	B	A	D	E
Block IV	A	D	E	C	B

---

Advantages of RBD : Chief advantages of RBD are as follows –

- (i) Accuracy : This design has been shown to be more efficient or accurate than CRD for most types of experimental work. The elimination of between sum of squares from residual sum of squares usually results in a decrease of error mean sum of squares.
- (ii) Flexibility : In RBD no restrictions are placed on the number of treatments or the number of replicates.
- (iii) Ease of analysis : Statistical analysis is simple, rapid and straight forward.

Disadvantages of RBD : (i) RBD may give misleading results if blocks are not homogeneous.

- (ii) RBD is not suitable for large number of treatments because in that case the block size will increase and it may not be possible to keep large blocks homogeneous.
- (iii) If the data on more than two plots is missing the statistical analysis becomes tedious and complicated.

## The analysis of RBD is exactly same as that of ANOVA-Two Way

If in an RBD, a single observation is made on each of the experimental units, then its analysis is same as ANOVA for fixed effect model for a two-way classified data with one observation per cell.

### Null Hypothesis

$H_{01}$  : Treatments are homogeneous ie.  $\tau_1 = \tau_2 = \tau_3 = \dots = \tau_t$

$H_{02}$  : the blocks are homogeneous ie.  $b_1 = b_2 = b_3 = \dots = b_r$

### Alternative Hypothesis

$H_{11}$  : At least two  $\tau_i$ 's are different

$H_{12}$  : At least two  $b_j$ 's are different

The linear model is given by

$$y_{ij} = \mu + \tau_i + b_j + e_{ij}; \text{ where, } i = 1, 2, \dots, t \text{ and } j = 1, 2, \dots, r$$

## Partitioning of the sum of squares

$$(S_T)^2 = (S_t)^2 + (S_b)^2 + (S_e)^2$$

$= \sum \sum (y_{ij} - \bar{y}_{..})^2$  is the total sum of squares  $(S_T)^2$

$(S_t)^2 = \sum (\bar{y}_i - \bar{y}_{..})^2$  is sum of squares due to treatments  $(S_t)^2$

$(S_b)^2 = \sum (\bar{y}_i - \bar{y}_{..})^2$  is sum of squares due to blocks  $(S_b)^2$

$(S_e)^2 = \sum \sum (y_{ij} - \bar{y}_i - \bar{y}_j + \bar{y}_{..})^2$  is the sum of squares due to error  $(S_e)^2$

$$RSS = \sum Y^2, \quad CF = G^2/n$$

$$TSS = RSS - CF$$

$$SST = \sum (T_i)^2 / r - CF$$

$$SSRe(bl) = \sum (R_i)^2 / t - CF$$

$$SSE = TSS - SST - SSR$$

## Partitioning of the sum of squares

$$(S_T)^2 = (S_t)^2 + (S_b)^2 + (S_e)^2$$

$= \sum \sum (y_{ij} - \bar{y}_{..})^2$  is the total sum of squares  $(S_T)^2$

$(S_t)^2 = \sum (\bar{y}_{i.} - \bar{y}_{..})^2$  is sum of squares due to treatments  $(S_t)^2$

$(S_b)^2 = \sum (\bar{y}_{i.} - \bar{y}_{..})^2$  is sum of squares due to blocks  $(S_b)^2$

$(S_e)^2 = \sum \sum (y_{ij} - \bar{y}_{i.} - \bar{y}_{j.} + \bar{y}_{..})^2$  is the sum of squares due to error  $(S_e)^2$

$$RSS = \sum Y^2, \quad CF = G^2/n$$

$$TSS = RSS - CF$$

$$SST = \sum (T_i)^2/r - CF$$

$$SSRe(bl) = \sum (R_i)^2/t - CF$$

$$SSE = TSS - SST - SSR$$

### ANOVA Table for RBD

Sources of Variation	Sum of Squares	Degree of Freedom (d.f.)	Mean Sum of Square	Variation Ratio ( $F_{cal}$ )
Between Treatments	$(S_t)^2$	$(t - 1)$	$MS_t = (S_t)^2/(t - 1)$	$F_t = MS_t/MS_e$
Blocks or replicates	$(S_b)^2$	$(r - 1)$	$MS_b = (S_b)^2/(r - 1)$	$F_b = MS_b/MS_e$
Error or residual	$(S_e)^2$	$(t - 1)(r - 1)$	$MS_e = (S_e)^2/\{(t - 1)(r - 1)\}$	

**Ex 2 :** A randomized block design experiment is run with three treatment. The three treatmeans are – 6, 7 and 11. Total sum of squares is 220. ANOVA table takes the following form.

Source of Variations	DF	SS	MSS	F-ratio
Treatments				
Blocks		132		
Error				
Total	11	220		

Find all the missing entries in the ANOVA table and draw inference at LOS %.

**Soln :** Since, it is given that RBD is run with 3 treatment, so DF for treatments  $= 3 - 1 = 2$ . As the total df is 11, total number of observation will be  $11 + 1 = 12$  ie. number of blocks  $= 12/3 = 4$ . Hence DF for blocks  $= 4 - 1 = 3$  and DF for error  $= 11 - 2 - 3 = 6$ .

Source of Variations	DF	SS	MSS = SS/DF	Treatment Means	Treatment total (Mean x No. of treatments)
Treatments	2	56	56/2 = 28	6	6x4 = 24
Blocks	3	132	132/3 = 44	7	7x4 = 28
Error	6	220 - (132 + 56) = 32	5.333	11	11x4 = 44
Total	11	220			96

$$\begin{aligned}
 \text{We know SSTreatment} &= \sum(T_i)^2/n_i - CF = (24^2 + 28^2 + 44^2)/4 - 768 \\
 &= 824 - 768 = 56
 \end{aligned}$$

$$CF = G^2/N = 96^2/3 \times 4 = 9216/12 = 768$$

Source of Variations	DF	SS	MSS = SS/DF	$F_{ratio} = \frac{MSS}{MSE}$	$F_{critical}$ (From F-table)
Treatments	2	56	28	5.25	$F_{0.05, 2, 6} = 5.14$
Blocks	3	132	44	8.25	$F_{0.05, 3, 6} = 4.76$
Error	6	32	5.333		
Total	11	220			

Conclusion : Now since,  $F_{ratio}$  for Treatments and Blocks are  $> F_{critical}$  we cannot accept the null hypothesis.

**Ex 3 :** Test whether all 4 replicates (blocks) are equally effective and all 6 seeding rates (treatments) are equally effective in giving the yield from the following data:

Grain yield of rice at six seeding rates (Mg/Ha)

Replicate	Seeding rate (kg/ha)					
	25	50	75	100	125	150
1	5.1	5.3	5.3	5.2	4.8	5.3
2	5.4	6.0	5.7	4.8	4.8	4.5
3	5.3	4.7	5.5	5.0	4.4	4.9
4	4.7	4.3	4.7	4.4	4.7	4.1

Note : There are 4 rows & 6 columns

**Soln :**

$H_{0t}$  : All six seeding rates (treatments) are equally effective.

$H_{0b}$  : All four replicates (blocks) are equally effective.

Replicate	Seeding rate (kg/ha)						$Y_{.j}$
	25	50	75	100	125	150	
1	5.1	5.3	5.3	5.2	4.8	5.3	31
2	5.4	6.0	5.7	4.8	4.8	4.5	31.2
3	5.3	4.7	5.5	5.0	4.4	4.9	29.8
4	4.7	4.3	4.7	4.4	4.7	4.1	26.9
$Y_{i.}$	20.5	20.3	21.2	19.4	18.7	18.8	118.9
$\Sigma(Y_{ij})^2$	105.35	104.67	112.92	94.44	87.53	89.16	594.07

$$= G = Y_{..}$$

Step 1 Calculate Correction Factor (CF) =  $(Y_{..})^2/(txr) = 118.9^2/(6 \times 4) = 589.05$

Step 2 Calculate Total SS =  $\Sigma(Y_{ij})^2 - CF$

$$= (5.1^2 + 5.4^2 + 5.3^2 + \dots + 4.1^2) - 589.05 = 5.02$$

Step 3 Calculate the Replicate SS (Rep SS) =  $\Sigma(Y_{.j})^2/t - CF$

$$= (31.0^2 + 31.2^2 + 29.8^2 + 26.9^2)/6 - 589.05 = 591.015 - 589.05 = 1.965$$

Step 4 Calculate the Treatment SS (Trt SS) =  $\sum(Y_{.j})^2/r - CF$

$$= (20.5^2 + 20.3^2 + 21.2^2 + 19.4^2 + 18.7^2 + 18.8^2)/4 - 589.05 = 1.2675$$

Step 5 Calculate the error SS = Total SS – Rep SS – Trt SS

$$= 5.02 - 1.965 - 1.2675 = 1.7875$$

Step 6

ANOVA Table

Sources of Variation	Sum of Squares	Degree of Freedom (d.f.)	Mean Sum of Square	Variation Ratio (F <sub>cal</sub> )
Between Replicates	1.965	r - 1 = 4 - 1 = 3	MS <sub>r</sub> = 1.965/3 = 0.655	F <sub>r</sub> = MS <sub>r</sub> /MS <sub>e</sub> = 5.495
Between Treatments	1.2675	t - 1 = 6 - 1 = 5	MS <sub>t</sub> = 1.2675/5 = 0.2535	F <sub>t</sub> = MS <sub>t</sub> /MS <sub>e</sub> = 2.127
Error or residual	1.7875	(r - 1)(t - 1) = 15	MS <sub>e</sub> = 1.7875/15 = 0.1192	
Total	5.02	tr - 1 = 23		

## Step 7 Obtain the $F_{critical}$ values from F-table for LOS ( $\alpha$ ) 0.05 and 0.01

For Replicate		For Treatment	
$F_{0.05, 3, 15}$	3.29	$F_{0.05, 5, 15}$	2.90
$F_{0.01, 3, 15}$	5.42	$F_{0.01, 5, 15}$	4.56

## Step 8 Draw the inference -

Replicates : Since  $F_{calc}$  (5.495)  $>$   $F_{tab}$  (3.29 & 5.42) at 95% & 99% levels of confidence, we reject the null hypothesis (ie. all replicate means are equal).

Treatment : Since  $F_{calc}$  (2.127)  $<$   $F_{tab}$  (2.90 & 4.56) at 95% & 99% levels of confidence, we fail to reject the null hypothesis (ie. all treatment means are equal). In other words, we accept the null hypothesis.

---

### Conclusion --

Replicates : Since  $F_{\text{calc}}$  (5.495)  $>$   $F_{\text{tab}}$  (3.29 & 5.42) at 95% & 99% levels of confidence, we reject the null hypothesis (ie. all replicate means are equal).

Treatment : Since  $F_{\text{calc}}$  (2.127)  $<$   $F_{\text{tab}}$  (2.90 & 4.56) at 95% & 99% levels of confidence, we fail to reject the null hypothesis (ie. all treatment means are equal). In other words, we accept the null hypothesis.

## Efficiency of RBD over CRD

If  $r_i = r$  for all  $i = 1, 2, \dots, t$  treatments, then for such data, we can apply both CRD and RBD for testing the equality of effects of treatments. In such case, RBD is more efficient than CRD and the relative efficiency of RBD over CRD is calculated as –

$$E = [r(t - 1)MSE_r + (r - 1)MSRe(B)] / [(rt - 1)MSE_r]$$

For the previous example 2,

$$\begin{aligned} E &= [4(6 - 1)0.1192 + (4 - 1)0.655] / [(4 \times 6 - 1)0.1192] \\ &= (2.384 + 1.965) / 2.7416 = 4.349 / 2.7416 = 1.586 \approx 1.59 \end{aligned}$$

Therefore, RBD is approximately 59% more efficient than CRD.

Note :  $E > 1$  means more efficient,  $E < 1$  means less efficient and  $E = 1$  means equally efficient.

## RBD with one missing observation

In RBD, if any one observation is missed then it is estimated as

$X = (rR' + tT' - G')/[(r - 1)(t - 1)]$  and correction factor is calculated as

$B (\text{Bias}) = [R' - (t - 1)X]^2/[t(t - 1)]$ , where, r and t are number of replicates and treatments respectively and  $R'$ ,  $T'$ ,  $G'$  are the corresponding replicate, treatment and grand totals.

After inserting  $X$  in the original data, all sum of squares are calculated as usual similar to normal RBD without the calculation of error sum of squares.

Corrected total sum of squares, corrected treatment sum of squares and corrected error sum of squares are calculated as

$$\text{CTSS} = \text{TSS} - 2B, \text{CSST} = \text{SST} - B, \text{CSSE} = \text{CTSS} - \text{CSST} - \text{SSB}$$

In the ANOVA table, both  $\text{SST}$ ,  $\text{SSE}$  and  $\text{TSS}$  are replaced by  $\text{CSST}$ ,  $\text{CSSE}$  and  $\text{CTSS}$  respectively and also error df is taken as  $[(r - 1)(t - 1) - 1]$  and total df is taken as  $(rt - 2)$ .

Remaining procedure is as usual similar to normal RBD.

Ex 4 : Suppose that the value for treatment 2 is missing in replication III. The data will then be as presented in the table below.

Treatment (t = 5)	Replication (r = 4)				Total Treatment
	I	II	III	IV	
1	22.9	25.9	39.1	33.9	121.8
2	29.5	30.4	X missing = 33.1	29.6	89.5 (T')
3	28.8	24.4	32.1	28.6	113.9
4	47	40.9	42.8	32.1	162.8
5	28.9	20.4	21.1	31.8	102.2
Total	157.1	142	135.1 (R')	156	590.2 (G')

$$T' = 89.5, R' = 135.1, G' = 590.2$$

$H_{0t}$  : All treatment mean effects are equal and

$H_{1t}$  : The treatment mean effects are not equal.

$H_{0R}$  : All replicate means are homogeneous and

$H_{1R}$  : All replicate means effects are not homogeneous

$$\begin{aligned}
 X &= (rR' + tT' - G') / \{(r-1)(t-1)\} = (4 \times 135.1 + 5 \times 89.5 - 590.2) / (3 \times 4) \\
 &= (540.4 + 447.5 - 590.2) / 12 = 33.1
 \end{aligned}$$

$$\begin{aligned}
 \text{The upward Bias, } B &= \{R' - (t-1)(X)\}^2 / \{t(t-1)\} \\
 &= \{135.1 - (4 \times 33.1)\}^2 / (5 \times 4) = 7.29 / 20 = 0.3645
 \end{aligned}$$

After substituting the estimated missing value, we get –

$$\text{Treatment 2 total } (T' + X) = 89.5 + 33.1 = 122.6$$

$$\text{Replication 3 total } (R' + X) = 135.1 + 33.1 = 168.2 \text{ and}$$

$$\text{The Grand total } (G' + X) = 590.2 + 33.1 = 623.3$$

$$\text{Treatment SS} = \sum (T_i)^2 / r - CF$$

$T' = 89.5, R' = 135.1, G' = 590.2$   
 $t = 5, r = 4, CF = G^2 / rt = 623.3^2 / 4 \times 5$

$$\begin{aligned}
 &= \{121.8^2 + 122.6^2 + 113.9^2 + 162.8^2 + 102.2^2\} / 4 - 623.3^2 / 20 \\
 &= 19946.9725 - 19425.1445 = 521.828
 \end{aligned}$$

$$\text{Corrected Treatment SS (Treatment SS} - B) = 521.828 - 0.3645 = 521.4635$$

TSS (including X)

$$\begin{aligned} &= \Sigma Y^2 - CF = (22.9^2 + 25.9^2 + \dots + 33.1^2 + \dots + 31.8^2) - 623.3^2/20 \\ &= 20364.47 - 19425.1445 = 939.3255 \end{aligned}$$

Corrected TSS (TSS - 2B) = 939.3255 - 2x0.3645 = 938.5965

$$\begin{aligned} SSB(Re) &= \Sigma (R_i)^2/t - CF = (157.1^2 + 142^2 + 168.2^2 + 156^2)/5 - 623.3^2/20 \\ &= 19494.33 - 19425.1445 = 69.1855 \end{aligned}$$

Corrected SSE = Cor TSS - Cor SST - SSB (Re)

$$= 938.5965 - 521.4635 - 69.1855 = 347.9475$$

ANOVA Table ( $r = 4, t = 5$ )

Sources of Variation	Sum of Squares	Degree of Freedom (d.f.)	Mean Sum of Square	Variation Ratio ( $F_{cal}$ )
Replication (Block)	69.1855	$r - 1 = 3$	$MS_r = 69.1855/3 = 23.0618$	$F_r = MS_r/MSe = 23.0618/31.6316 = 0.729$
Treatment (Corrected)	521.4635	$t - 1 = 4$	$MS_t = 521.4635/4 = 130.3659$	$F_t = MS_t/Mse = 130.3659/31.6316 = 4.121$
Error (Corrected)	347.9475	$(r - 1)(t - 1) - 1 = 11$	$MSe = 347.9475/11 = 31.6316$	
Total (Corrected)	938.5965	$tr - 2 = 18$		

Conclusion :  $F_{critical}$  at LOS  $\alpha = 5\%$  for DF 3,11 ie.  $F_{0.05,3,11} = 3.59$  and  $F_{cal} = 0.729$  for replication (block). Since  $F_{cal} < F_{crit}$   $H_0$  for replicates is accepted ie. replicate effects are homogeneous.

$F_{critical}$  at LOS  $\alpha = 5\%$  for DF 4,11 ie.  $F_{0.05,4,11} = 3.36$  and  $F_{cal} = 4.121$  for Treatment. Since  $F_{cal} > F_{crit}$   $H_0$  for treatment effect can not be accepted ie. treatment effects are not homogeneous.

---

Conclusion :  $F_{critical}$  at LOS  $\alpha = 5\%$  for DF 3,11 ie.  $F_{0.05,3,11} = 3.59$  and  $F_{cal} = 0.729$  for replication (block). Since  $F_{cal} < F_{crit}$   $H_0$  for replicates is accepted ie. replicate effects are homogeneous.

$F_{critical}$  at LOS  $\alpha = 5\%$  for DF 4,11 ie.  $F_{0.05,4,11} = 3.36$  and  $F_{cal} = 4.121$  for Treatment. Since  $F_{cal} > F_{crit}$   $H_0$  for treatment effect can not be accepted ie. treatment effects are not homogeneous.

## Practice Problems

Q1. A researcher reports the Relative Efficiency of a Randomized Block Design, relative to a Completely Randomized Design of 5. The Randomized Block Design had 5 treatments and 8 blocks. How many observations would be needed to have as precise of estimates of treatment means, if the experiment was conducted as to a Completely Randomized Design?

Q2. An experiment is conducted as a Randomized Block Design, with 10 blocks and 5 treatments, reports a relative efficiency (relative to a Completely Randomized Design) of  $RE = 5$ . How many TOTAL subjects would be needed if we ran a CRD, to obtain the same standard error of a treatment mean?

i) 50 ii) 10 iii) 25 iv) 250 v) 500

- (i) Ascertain if it is One-way ANOVA or Two-way ANOVA?  
(ii) Construct ANOVA Table.

## Practice Problems

---

Q3. Given the following date, test the hypothesis.

$H_0$  : All the means are equal;  $H_1$  : At least 2 means are different

$n_1 = 4, \bar{x}_1 = 27, s_1 = 4; n_2 = 7, \bar{x}_2 = 25, s_2 = 9; n_3 = 5, \bar{x}_3 = 28, s_3 = 5$

(i) Is it One-Way or Two-Way ANOVA?

(ii) Construct ANOVA Table.



# BSDCHZC355

## Statistical Inference & Applications

**BITS** Pilani  
Pilani | Dubai | Goa | Hyderabad

Shaibal Kr. Sen  
Session 07



# STATISTICAL INFERENCES & APPLICATIONS

## Scheduled Topics to be covered

**Design of Experiments : Complete Block Designs contd.**

**Latin Square Design (LSD)**

**Analysis of RBD and LSD with one missing observation**

**Comparison of efficiencies of CRD, RBD and LSD**

## Latin Square Design (LSD)

In LSD, blocking (local control) principle is applied in two perpendicular directions. The output times of a program for a common problem written by 3 different programmers - Krishna, Rani, Sathya (Rows) using 3 different software Python, R, SAS (Columns) executed in three different brands of laptops - Apple, Lenovo, Acer are given in the following design.

Time (in seconds)	Python	R	SAS
Krishna	Lenovo	Apple	Acer
	20	24	22
Rani	Apple	Acer	Lenovo
	23	26	28
Sathya	Acer	Lenovo	Apple
	20	27	29

## Latin Square Design (LSD)

The interest is to test whether average output times due to three programs are equal, three software are equal and due to three different brands of Laptops are equal. Then LSD is the suitable design in this case.

We set up - Null Hypothesis -

$H_{0R}$  : The average output times due to three programs are equal.

$H_{0C}$  : The average output times due to three software are equal.

$H_{0T}$  : The average output times due to three branded laptops are equal.

And Alternative Hypothesis -

$H_{1R}$  : The average output times due to three programs are not equal.

$H_{1C}$  : The average output times due to three software are not equal.

$H_{1T}$  : The average output times due to three branded laptops are not equal.

LSD is defined for eliminating the variation of two factors called row and column in this design. The number of treatments is equal to number of replications. Thus in case of  $m$  treatments there have to be  $\max^m m^2$  experimental units. The whole of experimental area is divided into  $m^2$  experimental units (plots) arranged in a square so that each row as well as each column contain  $m$  units.

The  $m$  treatments are allocated at random to these rows and columns in such a way that every treatment occurs only once in each row and in each column. Such a layout is LSD.

3 x 3 Layout		
A	B	C
B	C	A
C	A	B

Ex : An animal feeding experiment where the column groups may correspond with initial weight and the row group with age.

Let us set up null hypothesis –

For row effects  $H_{0r} : r_1 = r_2 = \dots = r_m = 0$

For column effects  $H_{0c} : c_1 = c_2 = \dots = c_m = 0$

For treatment effects  $H_{0t} : t_1 = t_2 = \dots = t_m = 0$

Alternative hypothesis –

For row effects  $H_{1r} : \text{At least 2 } r_i \text{ s are different}$

For column effects  $H_{1c} : \text{At least 2 } c_i \text{ s are different}$

For treatment effects  $H_{1t} : \text{At least 2 } t_i \text{ s are different}$

Let  $y_{ijk}$  ( $i, j, k = 1, 2, 3, \dots, m$ ) denote the response from the unit in the  $i^{\text{th}}$  row,  $j^{\text{th}}$  column and receiving the  $k^{\text{th}}$  treatment.

The model is  $y_{ijk} = \mu + r_i + c_j + t_k + e_{ijk}$ ; where  $i, j, k = 1, 2, 3, \dots, m$

where  $\mu$  is the constant mean effect;  $r_i$ ,  $c_j$  and  $t_k$  are due to  $i^{\text{th}}$  row,  $j^{\text{th}}$  column and  $k^{\text{th}}$  treatment respectively and  $e_{ijk}$  is the error effect due to random component assumed to be normally distributed with mean zero and variance  $\sigma_e^2$  ie.  $e_{ijk} \sim N(0, \sigma_e^2)$ .

If we write,

$G$  = Total of all the  $m^2$  observations.

Note : Total of all observations,  $N = m^2$

$R_i$  = Total of the  $m$  observations in the  $i^{\text{th}}$  row.

$C_j$  = Total of the  $m$  observations in the  $j^{\text{th}}$  column.

$T_k$  = Total of the  $m$  observations from the  $k^{\text{th}}$  treatment.

## ANOVA Table

Source of Variation	DF	Sum of Squares	Mean Sum of Sq	Variance Ratio (Fcal)
Rows	$m - 1$	$SSR = (S_R)^2$	$MSR = (S_R)^2/(m - 1)$	$FR = MSR/MSE$
Columns	$m - 1$	$SSC = (S_C)^2$	$MSC = (S_C)^2/(m - 1)$	$FC = MSC/MSE$
Treatments	$m - 1$	$SST = (S_T)^2$	$MST = (S_T)^2/(m - 1)$	$FT = MST/MSE$
Error	$(m - 1)(m - 2)$	$SSE = (S_E)^2$	$MSE = (S_E)^2/(m - 1)(m - 2)$	
Total	$m^2 - 1$	TSS		

---



---



---

Total sum of square (TSS) =  $\sum_{ijk} (y_{ijk})^2 - G^2/m^2$

Row sum of square (SSR) =  $(S_R)^2 = \sum (R_i)^2/m - G^2/m^2$

Column sum of square (SSC) =  $(S_C)^2 = \sum (C_j)^2/m - G^2/m^2$

Treatment sum of square (SST) =  $(S_T)^2 = \sum_k (T_k)^2/m - G^2/m^2$

SSE = TSS – SSR – SSC – SST

Obtain the  $F_\alpha$  value for  $DF = \{(m-1), (m-1)(m-2)\}$  at the level of significance  $\alpha$ . If  $F_R > F_\alpha$  reject  $H_{0R}$  and if  $F_R < F_\alpha$  can not reject  $H_{0R}$ . Similarly, it can be tested for  $H_{0C}$  and  $H_{0T}$ .

Also, if  $p_R < \alpha$ , then reject  $H_0$  and if  $p_R > \alpha$  accept  $H_0$ . Similarly, it can be inferred for  $p_C$  and  $p_T$ .

---

## Advantages of LSD –

With two-way grouping LSD controls more of the variation than CRD or RBD.

The two-way elimination of variation as a result of cross grouping often results in small error mean sum of squares.

LSD is an incomplete 3-way layout. Its advantage over the complete 3-way layout is that instead of  $m^3$  experimental units only  $m^2$  units are needed. Thus a 4x4 LSD results in saving  $m^3 = 4^3 - 4^2 = 48$  observations over a complete 3-way layout.

The statistical analysis is simple though slightly complicated than for RBD. Even 1 or 2 missing observations the analysis remains relatively simple.

More than one factor can be investigated simultaneously.

## Disadvantages of LSD –

LSD is suitable for the number of treatments between 5 and 10 and for more than 10 to 12 treatments the design is seldom used. Since in that case, the square becomes too large and does not remain homogeneous. In case of missing plots the statistical analysis becomes quite complex. If one or two blocks in a field are affected by some disease or pest, we can't omit because the number of rows columns and treatments have to be equal.

### Ex 1 : Perform ANOVA of the following LSD.

An experiment was carried out to determine the effect of claying the ground on the field of barley grains; amount of clay used were as follows –

A : No clay.

B : Clay at 100 per acre.

C : Clay at 200 per acre.

D : Clay at 300 per acre.

The yields were in plots of 8 meters by 8 meters and are given in the table.

D	B	C	A
29.1	18.9	29.4	5.7
C	A	D	B
16.4	10.2	21.2	19.1
A	D	B	C
5.4	38.8	24.0	37.0
B	C	A	D
24.9	41.7	9.5	28.9

## Soln :

$H_{0r}$  : Row means are all equal - vs-  $H_{1r}$  : Row means are not all equal.

$H_{0c}$  : Column means are all equal - vs-  $H_{1c}$  : Column means are not all equal.

$H_{0t}$  : Treatment means are all equal - vs-  $H_{1t}$  : Treatment means are not all equal.

	I	II	III	IV	Row Totals (R <sub>i</sub> )
I	D 29.1	B 18.9	C 29.4	A 5.7	83.1
II	C 16.4	A 10.2	D 21.2	B 19.1	66.9
III	A 5.4	D 38.8	B 24	C 37	105.2
IV	B 24.9	C 41.7	A 9.5	D 28.9	105
Column Totals (C <sub>j</sub> )	75.8	109.6	84.1	90.7	360.2 = G

The four treatment totals are –

$$A : 30.8$$

$$(5.7 + 10.2 + 5.4 + 9.5 = 30.8)$$

$$B : 86.9$$

$$(18.9 + 19.1 + 24 + 24.9 = 86.9)$$

$$C : 124.5$$

$$(29.4 + 16.4 + 37 + 41.7 = 124.5)$$

$$D : 118.0$$

$$(29.1 + 21.2 + 38.8 + 28.9 = 118)$$

$$\text{Grand Total } G = 360.2, \quad N = m^2 = 16; \quad CF = G^2/N = 360.2^2/16 = 8109.0025$$

$$\text{Raw SS (RSS)} = \sum (y_{ijk})^2 = 29.1^2 + 18.9^2 + \dots + 9.5^2 + 28.9^2 = 10,052.08$$

$$\text{Total SS (TSS)} = RSS - CF = 10,052.08 - 8109.0025 = 1943.0775$$

$$\text{SSR} = \sum (R_i)^2/m - CF = (83.1^2 + 66.9^2 + 105.2^2 + 105^2)/4 - 8109.0025 = 259.3125$$

$$\text{SSC} = \sum (C_j)^2/m - CF = (75.8^2 + 109.6^2 + 84.1^2 + 90.7^2)/4 - 8109.0025 = 155.2725$$

$$\text{SST} = \sum (T_k)^2/m - CF = (30.8^2 + 86.9^2 + 124.5^2 + 118^2)/4 - 8109.0025 = 1372.1225$$

$$\begin{aligned} \text{Error SS} &= TSS - SSR - SSC - SST \\ &= 1943.0775 - 259.3125 - 155.2725 - 1372.1225 = 156.37 \end{aligned}$$

### ANOVA Table for LSD

Source of Variation	DF	SS	MSS = SS/DF	$F_{cal}$ = MSS/MSE
Rows	$m - 1 = 3$	259.3125	86.4375	$FR = 3.316$
Columns	$m - 1 = 3$	155.2725	51.7575	$FC = 1.986$
Treatments	$m - 1 = 3$	1372.123	457.3742	$FT = 17.549$
Error	$(m - 1)(m - 2) = 6$	156.37	26.0616	
Total	$m^2 - 1 = 15$	1943.0775		

$F_{critical}$  value obtained from F-table for  $\alpha = 0.05$  and df 3, 6 ie.  $F_{0.05, 3, 6} = 4.76$ . Now, comparing the  $F_{cal}$  with  $F_{critical}$  we conclude that the null hypothesis for rows and columns can not be rejected whereas for treatments null hypothesis is rejected ie. different levels of clay have significant effect on the yield.

Rows

$$F_{\text{cal}} (3.316) < F_{\text{critical}} (4.76)$$

So, Null hypothesis can not be rejected.

Columns

$$F_{\text{cal}} (1.986) < F_{\text{critical}} (4.76)$$

So, Null hypothesis can not be rejected.

Treatments

$$F_{\text{cal}} (17.549) > F_{\text{critical}} (4.76)$$

So, Null hypothesis can not be accepted.

**Table 6(a) Values of  $F_{0.05}$** 

$v_2$ = Degrees of Freedom for Denominator	$v_1$ = Degrees of Freedom for Numerator																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	60	120	$\infty$
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.46	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.63	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.52	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.83	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.40	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.11	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.89	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.73	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.60	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.50	2.47	2.38	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.41	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.34	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.28	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.23	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.18	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.14	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.07	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.02	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.00	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.97	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.88	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.78	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.69	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.60	1.55	1.50	1.43	1.35	1.25
$\infty$	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.51	1.46	1.39	1.32	1.22	1.00

## One missing observation in LSD

Let us suppose that in  $[mxm]$  LSD, the observation occurring in the  $i^{\text{th}}$  row,  $j^{\text{th}}$  column and receiving the  $k^{\text{th}}$  treatment is missing. Let's assume that its value is  $x$ , ie.  $y_{ijk} = x$ . Now let -

$R_i'$  = Total of the known observations in the  $i^{\text{th}}$  row.

$C_j'$  = Total of the known observations in the  $j^{\text{th}}$  column.

$T_k'$  = Total of the known observations receiving  $k^{\text{th}}$  treatment.

$G'$  = Grand Total

$$X = \{m(R_i' + C_j' + T_k') - 2G'\} / \{(m - 1)(m - 2)\}$$

Ex 2 : In the following data, one value is missing. Estimate this value and analyse the given data.

A	C	B	D
12	19	10	8
C	B	D	A
18	12	6	7
B	D	A	C
22	<b>X</b>	5	21
D	A	C	B
12	7	27	17

Soln :

Col Row	I	II	III	IV	Row Totals (R <sub>i</sub> )
I	A 12	C 19	B 10	D 8	49
II	C 18	B 12	D 6	A 7	43
III	B 22	D <span style="color: red;">X</span>	A 5	C 21	48 + X <span style="color: red;">R<sub>3</sub>' = 48</span>
IV	D 12	A 7	C 27	B 17	63
Column Totals (C <sub>i</sub> )	64	38 + X	48	53	203 + X <span style="color: red;">G' = 203</span>
		<span style="color: red;">C<sub>2</sub>' = 38</span>			

$$D = 8 + 6 + X + 12 = 26 + X \quad T_4' = 26$$

$$\text{Here, } m = 4, \quad R_3' = 48, \quad C_2' = 38, \quad T_4' = 26, \quad G' = 203$$

Applying the missing estimation formula :

$$X = \{(m(R_3' + C_2' + T_4') - 2G')\}/\{(m - 1)(m - 2)\} = \{4(48 + 38 + 26) - 2 \times 203\}/3 \times 2 = 7$$

Inserting the estimated value of X, we get the following observations -

Row \ Col	I	II	III	IV	Row Totals (R <sub>i</sub> )
I	A 12	C 19	B 10	D 8	49
II	C 18	B 12	D 6	A 7	43
III	B 22	D 7	A 5	C 21	55
IV	D 12	A 7	C 27	B 17	63
Column Totals (C <sub>j</sub> )	64	45	48	53	G = 210

The four treatment totals are –

$$A (T_1) : 31$$

$$12 + 7 + 5 + 7 = 31$$

$$B (T_2) : 61$$

$$10 + 12 + 22 + 17 = 61$$

$$C (T_3) : 85$$

$$19 + 18 + 21 + 27 = 85$$

$$D (T_4) : 33$$

$$8 + 6 + 7 + 12 = 33$$

$$CF = G^2/m^2 = 210^2/4^2 = 44100/16 = 2756.25$$

$$\text{Raw SS(RSS)} = \sum (y_{ijk})^2 = 12^2 + 18^2 + \dots + 21^2 + 17^2 = 3432$$

$$\text{Total SS (TSS)} = \text{RSS} - \text{CF} = 3432 - 2756.25 = 675.75$$

$$\begin{aligned} \text{Row Sum of Squares (SSR)} &= \sum (R_i)^2/m - CF \\ &= (49^2 + 43^2 + 55^2 + 63^2)/4 - 2756.25 = 54.75 \end{aligned}$$

$$\begin{aligned} \text{Column Sum of Squares (SSC)} &= \sum (C_j)^2/m - CF \\ &= (64^2 + 45^2 + 48^2 + 53^2)/4 - 2756.25 = 52.25 \end{aligned}$$

$$\begin{aligned} \text{Sum of Squares of Treatment (SST)} &= \sum (T_k)^2/m - CF \\ &= (31^2 + 61^2 + 85^2 + 33^2)/4 - 2756.25 = 417.75 \end{aligned}$$

$$\begin{aligned} \text{Error SS (SSE)} &= \text{TSS} - \text{SSR} - \text{SSC} - \text{SST} \\ &= 675.75 - 54.75 - 52.25 - 417.75 = 151 \end{aligned}$$

ANOVA Table

Source of Variation	DF	SS	MSS = SS/DF	Variance Ratio		Conclusion
				Calculated	Tabulated $F_{0.05,3,5}$	
Rows	$4 - 1 = 3$	54.75	18.25	0.60	5.41	Insignificant
Columns	$4 - 1 = 3$	52.25	17.42	0.58	5.41	Insignificant
Treatments	$4 - 1 = 3$	417.75	139.25	4.61	5.41	Insignificant
Error	$(m - 1)(m - 2) - 1$ $= 5$	151	30.25			
Total	14	675.75				

### Remark 1 : Efficiency of LSD over RBD

There may be two cases to judge the relative efficiency of LSD over RBD –

1. Relative efficiency of LSD over RBD, when rows are taken as blocks is

$$= \{MSSC + (m - 1)MSSE\}/(mxMSSE)$$

where, MSSC = Mean Sum of Squares due to columns &

MSSE = Mean Sum of Squares due to error

2. Relative efficiency of LSD over RBD, when columns are taken as blocks is

$$= \{MSSR + (m - 1)MSSE\}/(mxMSSE)$$

where, MSSR = Mean Sum of Squares due to rows, m = no. of rows or no. of treatments.

### Remark 2 : Efficiency of LSD over CRD

Relative efficiency of LSD over CRD is given by

$$= \{MSSR + MSSC + (m - 1)MSSE\}/\{(m + 1)MSSE\}$$

## To calculate efficiency of LSD over RBD and CRD with respect to

In example 1,  $m = 4$ ,  $MSSR = 86.44$ ,  $MSSC = 51.76$ ,  $MSSE = 26.06$

Relative efficiency of LSD over RBD

If rows are assumed as blocks then  $e = \{MSSC + (m - 1)MSSE\}/(mxMSSE)$   
 $= (51.76 + (4 - 1) \times 26.06)/(4 \times 26.06) = 1.25$  then 25% more efficient than RBD.

If columns are assumed as blocks then

$e = \{MSSR + (m - 1)MSSE\}/(mxMSSE)$   
 $= (86.44 + (4 - 1) \times 26.06)/(4 \times 26.06) = 1.58$  then 58% more efficient than RBD.

Relative efficiency of LSD over CRD

$e = \{MSSR + MSSC + (m - 1)MSSE\}/\{(m + 1)MSSE\}$   
 $= \{51.76 + 86.44 + (4 - 1) \times 26.06\}/\{(4 + 1) \times 26.06\} = 1.66$  then 66% more efficient than CRD.

---

**Ex :** Perform ANOVA for the following LSD and obtain the relative efficiencies of the design over RBD and CRD. Use 0.01 level of significance.

Time(in seconds)	Python	R	SAS
Krishna	Lenovo 20	Apple 24	Acer 22
Rani	Apple 23	Acer 26	Lenovo 28
Sathya	Acer 20	Lenovo 27	Apple 29

$H_{0R}$  : The three programme instructions are equally effective.

$H_{1R}$  : The three programme instructions are not equally effective.

$H_{0C}$  : The three softwares are equally effective.

$H_{1C}$  : The three softwares are not equally effective.

$H_{0T}$  : The three brands of computers are equally effective.

$H_{1T}$  : The three brands of computers are not equally effective.

				Total (T <sub>k</sub> )
T <sub>1</sub>	0	8	7	15
T <sub>2</sub>	4	3	9	16
T <sub>3</sub>	2	6	0	8

	C <sub>1</sub>	C <sub>2</sub>	C <sub>3</sub>
R <sub>1</sub>	20	24	22
R <sub>2</sub>	23	26	28
R <sub>3</sub>	20	27	29

$$m = 3, N = m^2 = 3^2 = 9$$

$$Y = 20$$

	C <sub>1</sub>	C <sub>2</sub>	C <sub>3</sub>	Total (R <sub>i</sub> )
R <sub>1</sub>	0 T <sub>1</sub>	4 T <sub>2</sub>	2 T <sub>3</sub>	6
R <sub>2</sub>	3 T <sub>2</sub>	6 T <sub>3</sub>	8 T <sub>1</sub>	17
R <sub>3</sub>	0 T <sub>3</sub>	7 T <sub>1</sub>	9 T <sub>2</sub>	16
Total (C <sub>j</sub> )	3	17	19	39 (G)

$$RSS = \sum Y^2 = 0^2 + 4^2 + 2^2 + 3^2 + 6^2 + 8^2 + 0^2 + 7^2 + 9^2 = 259$$

$$CF = G^2/N = 39^2/3^2 = (39/3)^2 = 13^2 = 169$$

$$TSS = RSS - CF = 259 - 169 = 90$$

$$SSR = \sum (R_i)^2/m - CF = (6^2 + 17^2 + 16^2)/3 - 169 = 193.66 - 169 = 24.66$$

$$SSC = \sum (C_j)^2/m - CF = (3^2 + 17^2 + 19^2)/3 - 169 = 184.66 - 169 = 15.66$$

$$SST = \sum (T_k)^2/m - CF = (15^2 + 16^2 + 8^2)/3 - 169 = 181.66 - 169 = 12.66$$

$$SSE = TSS - SSR - SSC - SST = 90 - 24.66 - 15.66 - 12.66 = 37.02$$

### ANOVA Table for LSD

SV	SS	DF	MS = SS/DF	Variance Ratio (Fcal)	$F_{critical}$ ( $F_{0.01, 2, 2}$ )
Rows	24.66	$m - 1 = 2$	$24.66/2 = 12.33$	$MSR/MSE = 12.33/18.51 = 0.666$	99
Columns	15.66	$m - 1 = 2$	$15.66/2 = 7.83$	$MSC/MSE = 7.83/18.51 = 0.423$	99
Treatments	12.66	$m - 1 = 2$	$12.66/2 = 6.33$	$MST/MSE = 6.33/18.51 = 0.342$	99
Error	37.02	$(m - 1)(m - 2) = 2$	$37.02/2 = 18.51$		
Total	90	$m^2 - 1 = 8$			

Since the calculated values of F for Rows, Columns and Treatments are  $< F_{critical}$  values as depicted in the above ANOVA table, all the null hypothesis are accepted.

## Efficiency of LSD

(i) Efficiency of LSD over RBD if rows are blocks –

$$e = \{\text{MSC} + (m - 1)\text{MSE}\}/m\text{xMSE} = (7.83 + 2 \times 18.51)/3 \times 18.51 = 0.8 < 1$$

ie. LSD is less efficient than RBD by 20%.

(ii) Efficiency of LSD over RBD if columns are blocks –

$$e = \{\text{MSR} + (m - 1)\text{MSE}\}/m\text{xMSE} = (12.33 + 2 \times 18.51)/3 \times 18.51 = 0.88 < 1$$

ie. LSD is less efficient than RBD by 12%

(iii) Efficiency of LSD over CRD if columns are blocks –

$$e = \{\text{MSSR} + \text{MSSC} + (m - 1)\text{MSSE}\}/\{(m + 1)\text{MSSE}\} = (12.33 + 7.83 \times 18.51)/4 \times 18.51 = 0.77 < 1$$

ie. LSD is less efficient than CRD by 23%

## Miscellaneous

---

1. An experimenter randomly allocated 125 male turkeys to five treatment groups: control and treatments A, B, C, and D. There were 25 birds in each group, and the mean results were 2.16, 2.45, 2.91, 3.00, and 2.71, respectively. The sum of squares for experimental error was 153.4. Test the null hypothesis that the five group means are the same against the alternative that one or more of the treatments differs from the control.

Soln :  $H_0 : \mu_A = \mu_B = \mu_C = \mu_D = \mu_{\text{control mean}}$

$H_1$  : At least one or more treatment means are not equal to control mean.

Given  $n_1 = n_2 = n_3 = n_4 = n_5 = 25$ ;  $N = 5 \times 25 = 125$ ,  $k = 5$

$$\bar{x}_1 = 2.16, \bar{x}_2 = 2.45, \bar{x}_3 = 2.91, \bar{x}_4 = 3, \bar{x}_5 = 2.71$$

$$T_1 = n_1(\bar{x}_1) = 54, T_2 = n_2(\bar{x}_2) = 61.25, T_3 = n_3(\bar{x}_3) = 72.75,$$

$$T_4 = n_4(\bar{x}_4) = 75, T_5 = n_5(\bar{x}_5) = 67.75$$

Given, SSE = 153.4. By CRD (ANOVA One-way),

$$CF = G^2/N, \text{ where } G = 25(2.16 + 2.45 + 2.91 + 3 + 2.71) = 330.75;$$

$$\text{and } N = 5 \times 25 = 125$$

$$CF = 330.75^2/125 = 875.1645$$

$$SST = \sum (T_i)^2/n_i - CF = \{54^2/25 + 61.25^2/25 + 72.75^2/25 + 75^2/25 + 67.75^2/25\}$$

$$-870.1645 = 887.0075 - 875.1645 = 11.843$$

Soln :  $H_0: \mu_A = \mu_B = \mu_C = \mu_D = \mu_{\text{control mean}}$

$H_1$  : At least one or more treatment means are not equal to control mean.

Given  $n_1 = n_2 = n_3 = n_4 = n_5 = 25$ ;  $N = 5 \times 25 = 125$ ,  $k = 5$

$$\bar{x}_1 = 2.16, \bar{x}_2 = 2.45, \bar{x}_3 = 2.91, \bar{x}_4 = 3, \bar{x}_5 = 2.71$$

$$T_1 = n_1(\bar{x}_1) = 54, T_2 = n_2(\bar{x}_2) = 61.25, T_3 = n_3(\bar{x}_3) = 72.75,$$

$$T_4 = \bar{n}_4(\bar{x}_4) = 75, T_5 = n_5(\bar{x}_5) = 67.75$$

Given,  $SSE = 153.4$ . By CRD (ANOVA One-way),

$$CF = G^2/N, \text{ where } G = 25(2.16 + 2.45 + 2.91 + 3 + 2.71) = 330.75;$$

$$\text{and } N = 5 \times 25 = 125$$

$$CF = 330.75^2/125 = 875.1645$$

$$SST = \sum(T_i)^2/n_i - CF = \{54^2/25 + 61.25^2/25 + 72.75^2/25 + 75^2/25 + 67.75^2/25\}$$

$$-870.1645 = 887.0075 - 875.1645 = 11.843$$

ANOVA One-Way Table for CRD

SV	SS	DF	MS = SS/DF	$F_{\text{cal}}$	$F_{\text{critical}}$ ( $F_{0.02,4,120}$ )
Treatment	11.843	$k - 1 = 4$	$11.843 / 4 = 2.96$	2.316	2.45
Error	153.4	$n - k = 125 - 5 = 120$	$153.4 / 120 = 1.278$		
Total	165.243	124			

Since  $F_{\text{cal}} < F_{\text{critical}}$ , we accept the null hypothesis.

2. The leaves of certain plants in the genus *Albizzia* will fold and unfold in various light conditions. We have taken fifteen different leaves and subjected them to red light for 3 minutes. The leaves were divided into three groups of five at random. The leaflet angles were then measured 30, 45, and 60 minutes after light exposure in the three groups. Data from W. Hughes. Analyze these data to test the null hypothesis that delay after exposure does not affect leaflet angle.

Delay (mins)	Angles (degrees)				
30	140	138	140	138	142
45	140	150	120	128	130
60	118	130	128	118	118

$$\begin{array}{lll}
 T_i & T_i^2 & \\
 T_1 & T_1^2 & n_1 = 5 \\
 T_2 & T_2^2 & n_2 = 5 \\
 T_3 & T_3^2 & n_3 = 5 \\
 G & & N = 15
 \end{array}$$

## CRD ANOVA One-Way

$H_0$  : The three types delay does not effect the leaflet angles  $\mu_{30} = \mu_{45} = \mu_{60}$

$$TSS = RSS - CG = \sum Y^2 - G^2/N$$

$$SST = \sum (T_i)^2/n_i - CF = (T_1)^2/n_1 + (T_2)^2/n_2 + (T_3)^2/n_3 - G^2/N$$

$$SSE = TSS - SST$$

SV	SS	DF	MS = SS/DF	F <sub>cal</sub>	F <sub>0.05,2,12</sub>
Treatments		k - 1 = 2	MSTr	MSTr/MSE	3.89
Error		n - k = 12	MSE		
Total		14			

3. Three different methods of analysis M1, M2, M3 are used to determine of a certain constituent in the sample. Each method is used by five analysis in the results, and the results are given in the results are given in table. Do these results indicate a significant variation either between the methods or between the analysts?

Analysts	Methods		
	M1	M2	M3
1	7.5	7.0	7.1
2	7.4	7.2	6.7
3	7.3	7.0	6.9
4	7.6	7.2	6.8
5	7.4	7.1	6.9

## ANOVA-I way

Three different techniques namely – Medication, Exercises and Special Diet are randomly assigned to (individuals diagnosed with high blood pressure) to lower the blood pressure. After four weeks the reduction in each person is recorded. Test at 5% level, whether there is significant difference in mean reduction of blood pressure among three techniques.

Medication	10	12	9	15	13
Exercise	6	8	3	0	2
Diet	5	9	12	8	4

Soln :

Step 1 : Hypotheses

Null Hypothesis :  $H_0 : \mu_1 = \mu_2 = \mu_3$

That is, there is no significant difference among the three groups on the average reduction in blood pressure.

Alternative hypothesis :  $H_1 : \mu_i \neq \mu_j$  for at least one pair  $(i, j)$ ;  $i, j = 1, 2, 3$ ;  $i \neq j$

That is, there is significant difference in the average reduction in blood pressure in at least one pair of treatments.

Step 2 : Data

Medication	10	12	9	15	13
Exercise	6	8	3	0	2
Diet	5	9	12	8	4

Step 3 : Level of significance  $\alpha = 0.05$

Step 4 : Test statistic,  $F_0 = MST/MSE$

Step 5 : Calculation of Test Statistics

						Total	Square
Medication	10	12	9	15	13	59	3481
Exercise	6	8	3	0	2	19	361
Diet	5	9	12	8	4	38	1444
						$G = 116$	5286

Individual Squares

Medication	100	144	81	225	169
Exercise	36	64	9	0	4
Diet	25	81	144	64	16

$$\Sigma \Sigma (x_{ij})^2 = 10^2 + 12^2 + 9^2 + 15^2 + 13^2 + 6^2 + 8^2 + 3^2 + 2^2 + 5^2 + 9^2 + 12^2 + 8^2 + 4^2 = 1162;$$

$$1. \text{ Correction Factor (CF)} = G^2/N = 116^2/15 = 13456/15 = 897.06$$

2. Total Sum of Squares (TSS) :  $\Sigma \Sigma (x_{ij})^2 - CF = 1162 - 897.06 = 264.94$

3. Sum of Squares between Treatments (SST) :  $\Sigma \Sigma (x_i)^2 / n_i - CF$   
 $= 5286/5 - 897.06 = 1057.2 - 897.06 = 160.14$

4. Sum of Squares due to Error (SSE) :  $TSS - SST = 264.94 - 160.14 = 104.8$

ANOVA Table (One-way)

SV	SS	DF	MS = SS/DF	F <sub>ratio</sub>	F <sub>crit</sub> = F <sub>0.05, 2,12</sub>
Between Treatments	160.14	3 - 1 = 2	160.14/2 = 80.07	MSTr/MSE = 9.17	3.89
Error	104.8	12	104.8/12 = 8.73		
Total	264.94	n - 1 = 15 - 1 = 14			

Step 6 : Critical value F<sub>0.05, 2,12</sub> = 3.89

Step 7 : Decision – As F<sub>ratio</sub> (9.17) > F<sub>critical</sub> (3.89), the null hypothesis is rejected. Hence, we conclude that there exists significant difference in the reduction of the average blood pressure in at least one pair of techniques.

## ANOVA II Way

A reputed marketing agency in India has three different training programmes for its salesmen. The three programmes are Method – A, B, C. To assess the success of the programmes, 4 salesmen from each programmes were sent to the field. Their performance in terms of sales are given in the table below -

Salesmen	Methods		
	A	B	C
1	4	6	2
2	6	10	6
3	5	7	4
4	7	5	4

Test whether there is significant difference among methods and among salesmen.

---

**Soln :**

Step 1 : Hypotheses

Null Hypotheses :  $H_{01} : \mu_{M1} = \mu_{M2} = \mu_{M3}$  (for methods / treatments)

That is, there is no significant difference among the three programmes in their mean sales.

$H_{02} : \mu_{s1} = \mu_{s2} = \mu_{s3} = \mu_{s4}$  (for blocks)

Alternative Hypotheses :

$H_{11} :$  At least one average is different from the other, among the three programmes.

$H_{12} :$  At least one average is different from the other, among the four salesmen.

## Step 2 : Data

Salesmen	Methods		
	A	B	C
1	4	6	2
2	6	10	6
3	5	7	4
4	7	5	4

Step 3 : Level of significance  $\alpha = 5\%$

Step 4 : Test statistic :  $F_{ot}(\text{treatment}) = \text{MST}/\text{MSE}$  &  $F_{ob}(\text{blocks}) = \text{MSB}/\text{MSE}$

## Step 5 : Calculation of Test statistic

Salesmen	Methods			Total $x_i$	$x_i^2$
	A	B	C		
1	4	6	2	12	144
2	6	10	6	22	484
3	5	7	4	16	196
4	7	5	4	16	196
$\bar{x}_i$	22	28	16	66	1140
$(\bar{x}_i)^2$	484	784	256	1524	

Squares		
16	36	4
36	100	36
25	49	16
49	25	16
		$\Sigma\Sigma(x_{ij})^2 = 408$

Salesmen	Methods			Total $x_i$	$x_i^2$
	A	B	C		
1	4	6	2	12	144
2	6	10	6	22	484
3	5	7	4	16	196
4	7	5	4	16	196
$\bar{x}_i$	22	28	16	66	1140
$(\bar{x}_i)^2$	484	784	256	1524	

Correction Factor (CF) =  $G^2/N = 66^2/(3 \times 4) = 4356/12 = 363$

Total Sum of Squares (TSS) =  $\sum \sum (x_{ij})^2 - CF = 408 - 363 = 45$

Sum of squares due to treatment (SST) =  $\sum_{i=1}^k (x_{i.})^2/k - CF = 1140/3 - 363 = 17$

$$\begin{aligned}\text{Sum of Squares due to Blocks (SSB)} &= \sum (x_{.j})^2/k - CF = 1524/4 - 363 \\ &= 381 - 363 = 18\end{aligned}$$

$$\text{Sum of Squares due to Error (SSE)} = TSS - SST - SSB = 45 - 17 - 18 = 10$$

$$\text{Mean Sum of Squares : MST} = SST/(k - 1) = 17/3 = 5.67$$

$$\text{MSB} = SSB/(m - 1) = 18/2 = 9$$

$$\text{MSE} = SSE/(k - 1)(m - 1) = 10/6 = 1.67$$

ANOVA Table (Two-way)

SV	SS	DF	MS = SS/DF	Fratio	Fcritical
Between Treatments (Programmes)	17	$4 - 1 = 3$	$17/3 = 5.67$	$MSTr/MSE = 3.4$	$F_{0.05,3,6} = 4.76$
Between Blocks (Salesmen)	18	$3 - 1 = 2$	$18/2 = 9$	$MSB/MSE = 5.39$	$F_{0.05,2,6} = 5.14$
Error	10	6	$10/6 = 1.67$		
Total	45	11			

## Step 6 : Critical values

$F_{0.05,3,6} = 4.76$  (for treatments)

$F_{0.05,2,6} = 5.14$  (for blocks)

## Step 7 : Decision

- (i) Calculated  $F_{0t}(3.4) < F_{0.05,3,6} (4.76)$ , the null hypothesis is not rejected and we conclude that there is significant difference in the mean sales among the three programmes.
- (ii) Calculated  $F_{0b}(5.39) > F_{0.05,2,6} (5.14)$ , the null hypothesis is rejected and we conclude that there does not exist significant difference in the mean sales among the four salesmen.

---

## Conclusion --

For Treatments : Since  $F_{0t}(3.4) < F_{0.05,3,6} (4.76)$ , the null hypothesis is not rejected and we conclude that there is significant difference in the mean sales among the three programmes.

For Blocks : Since Calculated  $F_{0b}(5.39) > F_{0.05,2,6} (5.14)$ , the null hypothesis is rejected and we conclude that there does not exist significant difference in the mean sales among the four salesmen.



# BSDCHZC355

## Statistical Inference & Applications

**BITS** Pilani  
Pilani | Dubai | Goa | Hyderabad

Shaibal Kr. Sen  
Session 09



**STATISTICAL INFERENCES**

**&**

**APPLICATIONS**

## Scheduled Topics to be covered

---

Design of Experiments : Incomplete Block Designs

Balanced Incomplete Block Design (BIBD)

Partially Balanced Incomplete Block Design (PBIBD)

## Incomplete Block Designs

Randomised Block Design (RBD) is a complete block design in which each block size (number of experimental units or plots in a block) is same as the number of treatments and the number of blocks or replicates is less than or more than or equal to number of treatments.

On the other hand an Incomplete Block Design is such that each block size (number of experimental units or plots) is less than the number of treatments and the number of blocks is more than or equal to number of treatments.

---

These designs were introduced by Yates in order to eliminate heterogeneity to a greater extent than is possible with Randomized Blocks and Latin Squares when the number of treatments is large. The precision of the estimate of a treatment effect depends on the number of replications of the treatment – if larger is the number of replications, the more is the precision. Similar is the case for the precision of estimate of the difference between two treatment effects. If a pair of treatment occurs together more number of times in the design, the difference between these two treatment effects can be estimated with more precision. To ensure equal or nearly equal precision of comparisons of different pairs of treatment effects, the treatments are so allocated to the experimental units in different

---

blocks of equal sizes such that each treatment occurs at most once in a block and it has an equal number of replications and each pair of treatments has the same or nearly the same number of replications. When the number of replications of all pairs of treatments in a design is the same, then we have an important class of designs called Balanced Incomplete Block Designs (BIBD) and when there are unequal number of replications for different pairs of treatments, then the designs are called as Partially Balanced Incomplete Block Design (PBIBD).

---

Balanced Incomplete Block Designs (BIBDs) are formed mainly when number of plots in each block is less than the number of treatments.

A BIBD is an arrangement of  $v$  treatments in  $b$  blocks each of size  $k$  ( $< v$ ) such that –

- (i) Each treatment occurs at most once in a block.
- (ii) Each treatment occurs in exactly  $r$  blocks.
- (iii) Each pair of treatments occurs together in exactly  $\lambda$  block.

Hence, the parameters of the design are  $v$  = treatments,  $b$  = blocks, block size =  $k < v$ , Each treatment occurs in exactly  $r$  blocks, each pair of treatments occur in exactly  $\lambda$  block.

**Ex :** A BIBD for  $v = b = 5$ ,  $r = k = 4$  and  $\lambda = 3$  in the following -

	(I)	1	2	3	4
Blocks	(II)	1	2	3	5
	(III)	1	2	4	5
	(IV)	1	3	4	5
	(V)	2	3	4	5

The symbols  $v$ ,  $b$ ,  $r$ ,  $k$ ,  $\lambda$  are called the parameters of the design. These parameters satisfy the relations :  $vr = bk$ .....(i) and  $\lambda(v - 1) = r(k - 1)$ .....(ii) where,  $b \geq v$ .

Each side of equation (i) represents the total number of experimental units or plots in the design. Equation (ii) can be established by noting that a given treatment occurs with  $(k - 1)$  other treatments in each of  $r$  blocks and also occurs with each of the other  $(v - 1)$  treatments in  $\lambda$  blocks.

## Statistical Analysis

Consider the model :  $y_{ij} = \mu + \tau_i + \beta_j + e_{ij}$

Observation = General Mean + Treatment Effect + Block Effect + Random Error

Random errors are assumed to be independently and identically distributed normally with mean zero and constant variance  $\sigma^2$ . On minimizing the error sum of squares with respect to the parameters, we get a set of normal equations which can be solved to get the estimates of different contrasts of various treatment & block effects.

Now we compute -

G = Grand Total of observations.

$\bar{y}$  = Grand Mean =  $G/n$ , where  $n = vr = bk$  = Total No. of observations.

$T_i$  = Sum of observations for treatment  $i$ , ( $i = 1, 2, \dots, v$ )

$B_j$  = Sum of observations in block  $j$ , ( $j = 1, 2, \dots, b$ )

CF =  $G^2/n$

$Q_i$  = Adjusted  $i^{\text{th}}$  treatment total

=  $T_i - (\text{Sum of block totals in which treatment } i \text{ occurs})/\text{Block size (k)}$

A solution for the  $i^{\text{th}}$  treatment effect is,  $\hat{\tau}_i = (kQ_i)/(\lambda v)$       ( $i = 1, 2, \dots, v$ )

Adjusted treatment mean for treatment  $i$  =  $i^{\text{th}}$  treatment effects ( $\hat{\tau}_i$ ) + grand mean ( $\bar{y}$ )

---

Various sums of squares can be obtained as follows –

- (i) Total Sum of Squares (TSS) =  $\Sigma(\text{observation})^2 - \text{CF}$
- (ii) Treatment Sum of Squares unadjusted ( $\text{SST}_U$ ) =  $[\Sigma(T_i)^2]/r - \text{CF}$
- (iii) Block Sum of Squares unadjusted ( $\text{SSB}_U$ ) =  $[\Sigma(B_j)^2]/k - \text{CF}$
- (iv) Treatment Sum of Squares adjusted ( $\text{SST}_A$ ) =  $\Sigma \hat{\tau}_i Q_i$
- (v) Error Sum of Squares (SSE) = TSS –  $\text{SSB}_U - \text{SST}_A$
- (vi) Blocks Sum of Squares adjusted ( $\text{SSB}_A$ ) =  $\text{SST}_A + \text{SSB}_U - \text{SST}_U$

Analysis of Variance for a BIB is given in next slide -

### ANOVA for BIB (v, b, r, k, $\lambda$ ) Design

Source of Variation	DF	SS	MS	$F_{\text{calculated}}$	$F_{\text{critical}}$
SSTreatment (unadjusted)	$v - 1$	$SST_U$			
SSBlocks (unadjusted)	$b - 1$	$SSB_U$			
SSTreatment (adjusted)	$v - 1$	$SST_A$	$MST$	$MST/MSE$	$F_{(v-1),(n-v-b+1)}$
SSBlocks (adjusted)	$b - 1$	$SSB_A$	$MSB$	$MSB/MSE$	$F_{(b-1),(n-v-b+1)}$
SSError	$n - v - b + 1$	$SSE$	$MSE$		
SSTotal	$n - 1$	$TSS$			

Ex : The following data relate to an experiment conducted using a BIB design with parameters  $v = 4$ ,  $b = 4$ ,  $r = 3$ ,  $k = 3$ ,  $\lambda = 2$ . The layout plan and yield figures (in coded units) are tabulated below -

Block Number	Treatments & Yield figures			
1	(1)	77	(2)	85
2	(1)	70	(2)	67
3	(1)	69	(3)	62
4	(2)	72	(3)	63
			(4)	55

Carry out the analysis.

Soln : Grand Total,  $G = 77 + 85 + 60 + \dots + 55 = 774$

No. of observations,  $n = 12$ , Grand mean,  $\bar{y} = G/n = 774/12 = 64.5$

No. of replication,  $r = 3$ , Block size,  $k = 3$ , CF =  $G^2/n = 599076/12$

$$= 49923$$

1	2	3	4	5	6	7	8	9
Treat. / Block No.	( $T_i$ )	( $B_j$ )	Block Nos. in which Treat. $i$ occurs	$\sum_{j(i)} B_j$	$\sum_{j(i)} B_j/k$	$Q_i$	$\hat{t}_i = kQ_i/\lambda v$ ith treat. Effect	Adj. treat. mean
1	216	222	1, 2, 3	584	194.67	21.33	8.00	72.50
2	224	191	1, 2, 4	603	201	23.00	8.63	73.13
3	185	171	1, 3, 4	583	194.33	-9.33	-3.50	61.00
4	149	190	2, 3, 4	552	184	-35.00	-13.13	51.37

$$T_1 = 77 + 70 + 69 = 216, \dots, T_4 = 54 + 40 + 55 = 149,$$

$$B_1 = 77 + 85 + 60 = 222, \dots, B_4 = 72 + 63 + 55 = 190,$$

$$Q_i = T_i - \sum_{j(i)} B_j/k$$

Treat. Total  $T_1 = 77 + 70 + 69 = 216$ , etc., Block Total  $B_1 = 77 + 85 + 60 = 222$

Total of blocks in which treat.  $i$  occurs  $\sum_{j(i)} B_j = 222 + 191 + 171 = 584$ , etc.

Adj. treat. Total ( $Q_i$ ) =  $T_i - \sum_{j(i)} B_j/k = 216 - 584/3 = 21.33$ , etc.

Total SS (TSS) =  $\sum (\text{observation})^2 - CF = 77^2 + 85^2 + \dots + 55^2 - CF$   
 $= 51442 - 49923 = 1519$

Treatment SS unadj. ( $SST_U$ ) =  $[\sum (T_i)^2]/r - CF = 153258 - 49923 = 1163$

Block SS unadj. ( $SSB_U$ ) =  $[\sum (B_j)^2]/k - CF = 151106/3 - 49923 = 445.67$

Treatment SS adj. ( $SST_A$ ) =  $\sum \hat{\tau}_i Q_i = 861.34$

Error SS (SSE) =  $TSS - SSB_U - SST_A = 1519 - 445.67 - 861.34 = 211.99$

Block SS adj. ( $SSB_A$ ) =  $SST_A + SSB_U - SST_U = 861.34 + 445.67 - 1163 = 144.01$

Refer ANOVA table in next slide -

1	2	3	4	5	6	7	8	9
Treat. / Block No.	( $T_i$ )	( $B_j$ )	Block Nos. in which Treat. $i$ occurs	$\sum_{j(i)} B_j$	$\sum_{j(i)} B_j/k$	$Q_i$	$\hat{\tau}_i = kQ_i/\lambda v$ ith treat. Effect	Adj. treat. mean
1	216	222	1, 2, 3	584	194.67	21.33	8.00	72.50
2	224	191	1, 2, 4	603	201	23.00	8.63	73.13
3	185	171	1, 3, 4	583	194.33	-9.33	-3.50	61.00
4	149	190	2, 3, 4	552	184	-35.00	-13.13	51.37

Treat. Total  $T_1 = 77 + 70 + 69 = 216$ , etc., Block Total  $B_1 = 77 + 85 + 60 = 222$

Total of blocks in which treat.  $i$  occurs  $\sum_{j(i)} B_j = 222 + 191 + 171 = 584$ , etc.

Adj. treat. Total ( $Q_i$ ) =  $T_i - \sum_{j(i)} B_j/k = 216 - 584/3 = 21.33$ , etc.

$$\begin{aligned}\text{Total SS (TSS)} &= \sum (\text{observation})^2 - \text{CF} = 77^2 + 85^2 + \dots + 55^2 - \text{CF} \\ &= 51442 - 49923 = 1519\end{aligned}$$

$$\text{Treatment SS unadj. (SST}_U) = [\sum (T_i)^2]/r - \text{CF} = 153258 - 49923 = 1163$$

$$\text{Block SS unadj. (SSB}_U) = [\sum (B_j)^2]/k - \text{CF} = 151106/3 - 49923 = 445.67$$

$$\text{Treatment SS adj. (SST}_A) = \sum \hat{\tau}_i Q_i = 861.34$$

$$\text{Error SS (SSE)} = \text{TSS} - \text{SSB}_U - \text{SST}_A = 1519 - 445.67 - 861.34 = 211.99$$

$$\text{Block SS adj. (SSB}_A) = \text{SST}_A + \text{SSB}_U - \text{SST}_U = 861.34 + 445.67 - 1163 = 144.01$$

Refer ANOVA table in next slide -

<u>ANOVA Table</u>					
Source of Variation	DF	SS	MS	$F_{\text{calculated}}$	$F_{\text{critical}}$
SSBlocks (unadj.)	3	445.67			
SSTreatments (adj.)	3	861.34	287.11	6.77	5.41
SSBlocks (adj.)	3	144.01	48.00	1.13	5.41
SSTreatments (unadj.)	3	1163.00			
SSError	$11 - 6 = 5$	211.99	42.40		
SSTotal	$12 - 1 = 11$	1519.00			

Since for Treatments  $F_{\text{calculated}} (6.77) > F_{\text{critical}} (5.41)$ , Treatment effects are significant.

## Ex : Analyze the following BIBD

Balanced Incomplete Block Design for Catalyst Experiment

Treatment (Catalyst)	Block (Batch of Raw Material)			
	1	2	3	4
1	73	74	-	71
2	-	75	67	72
3	73	75	68	-
4	75	-	72	75

B1	B2	B3	B4
1	1	2	1
3	2	3	2
4	3	4	4

$H_{0t}$  : All treatment effects are equal –vs-  $H_{at}$  : All treatment effects are not equal

$H_{0b}$  : All block effects are equal –vs-  $H_{ab}$  : All block effects are not equal

### Balanced Incomplete Block Design for Catalyst Experiment

Treatment (Catalyst)	Block (Batch of Raw Material)				
	1	2	3	4	$y_{i.}$
1	73	74	-	71	218
2	-	75	67	72	214
3	73	75	68	-	216
4	75	-	72	75	222
$y_{.j}$	221	224	207	218	$870 = y_{..}$

This is a BIBD with  $v = a = 4$ ,  $b = 4$ ,  $k = 3$ ,  $r = 3$ ,  $\lambda = 2$  and  $N = 12$ . The analysis of this data is as follows. The Total Sum of Squares is –

$$\begin{aligned} SST &= \sum (obsn)^2 - G^2/N = \sum_i \sum_j (y_{ij})^2 - (y_{..})^2/12 \\ &= (73^2 + 74^2 + \dots + 75^2) - 870^2/12 = 63156 - 63075 = 81 \end{aligned}$$

$$\begin{aligned} SSBlocks &= 1/3 \left[ \sum_{j=1}^4 (y_{.j})^2 \right] - G^2/N = 1/3 [221^2 + 224^2 + 207^2 + 218^2] - 870^2/12 \\ &= 63125 - 63070 = 55 \end{aligned}$$

$$Q_1 = T_1 - 1/3(B_1 + B_2 + B_4) = 218 - 1/3(221 + 224 + 218) = -9/3 = -3$$

$$Q_2 = T_2 - 1/3(B_2 + B_3 + B_4) = 214 - 1/3(224 + 207 + 218) = -7/3$$

$$Q_3 = T_3 - 1/3(B_1 + B_2 + B_3) = 216 - 1/3(221 + 224 + 207) = -4/3$$

$$Q_4 = T_4 - 1/3(B_1 + B_3 + B_4) = 222 - 1/3(221 + 207 + 218) = 20/3$$

$$\begin{aligned}
 \text{SSTreatment (adjusted)} &= \text{SSTreat.}_{(\text{adj})} = [k \sum_{i=1}^4 (Q_i)^2] / \lambda a = \sum \hat{\tau}_i Q_i \\
 &= 3[9 - 9/3)2 + (-7/3)2 + (-4/3)2 + (20/3)2] / (2 \times 4) = 22.75
 \end{aligned}$$

The error sum of squares is obtained by subtraction as -

$$\text{SSE} = \text{SST} - \text{SSTreat.}_{(\text{adj})} - \text{SSBlocks} = 81 - 22.75 - 55 = 3.25$$

$$Q'_1 = B_1 - 1/3(T_1 + T_3 + T_4) = 221 - 1/3(218 + 216 + 222) = 7/3$$

$$Q'_2 = B_2 - 1/3(T_1 + T_2 + T_3) = 224 - 1/3(218 + 214 + 216) = 24/3$$

$$Q'_3 = B_3 - 1/3(T_2 + T_3 + T_4) = 207 - 1/3(214 + 216 + 222) = -31/3$$

$$Q'_4 = B_4 - 1/3(T_1 + T_2 + T_4) = 218 - 1/3(218 + 214 + 222) = 0$$

$$SS_{\text{Blocks}_{(\text{adj.})}} = [r \sum (Q'_i)^2] / \lambda b = 3[(7/3)^2 + (24/3)^2 + (-31/3)^2 + (0)^2] / (2 \times 4) = 66.08$$

$$SST_{\text{Treat.}} = \sum (T_i)^2 / r - G^2 / N = [(218)^2 + (214)^2 + (216)^2 + (222)^2] / 3 - (870)^2 / 12 \\ = 11.67$$

#### ANOVA including both Treatments & Blocks

Source of Variation	DF	SS	MS	F <sub>calculated</sub>	F <sub>critical</sub>
SSTreatments (adj.)	3	22.75	22.75/3 = 7.58	7.58/0.65 = 11.66	5.41
SSTreatments (unadj.)	3	11.67			
SSBlocks (unadj.)	3	55.00			
SSBlocks (adj.)	3	66.08	66.08/3 = 22.03	22.03/0.65 = 33.89	5.41
SSError	11 - 6 = 5	3.25	3.25/5 = 0.65		
SSTotal	11	81.00			

Value of F<sub>3,5</sub> for 5% LOS, obtained from F-table, is 5.41 ie. F<sub>critical</sub> = 5.41

Since both Treatments & Blocks F<sub>calculated</sub> (11.66) & (33.89) > F<sub>critical</sub> (5.41),  
 Reject both H<sub>0t</sub> and H<sub>0b</sub>

## Advantages and Disadvantages of BIBDs

Designs are complex and are looked up in tables.

### Advantages

Small blocks are more homogeneous than large blocks, so experimental error is lower.

Use when there is variability within larger blocks.

Use to increase precision.

### Disadvantages

Design can require a fixed number of treatments, a fixed number of reps or both.

More reps are needed.

The complexity of the analysis is increased.

There is unequal precision for certain comparisons of treatment means.

### Uses

Use to reduce block size in single factor experiments when the number of treatments is large.



# BSDCHZC355

## Statistical Inference & Applications

**BITS** Pilani  
Pilani | Dubai | Goa | Hyderabad

Shaibal Kr. Sen  
Session 10



# STATISTICAL INFERENCES & APPLICATIONS

## Scheduled Topics to be covered

---

Design of Experiments : Incomplete Block Designs contd.

Simple Lattice Design

Youden Square Design

## Simple Lattice Design (Quasi Factorial Design)

Lattice designs form an important class of useful Incomplete Block Designs.

Balanced Incomplete Block Designs (BIBD) are the most efficient designs in the class of Binary Incomplete Block Designs but these designs require usually a large number of replications and are not available for all combinations of parametric values. To overcome these difficulties Yates (1936) introduced a class of designs known as Quasi Factorial or Lattice Designs. The characteristic features of these designs are that the number of treatments is a perfect square and the block size is square root of this number. Moreover, incomplete blocks are combined in groups to form separate replications. The

## Simple Lattice Design

---

number of replications of the treatments are flexible in these designs and are useful for situations in which a large number of treatments are to be tested. If the design has two replications in the treatments, it is called a simple lattice; if it has 3 replications it is called a triple-lattice and so on. In general, if the number of replications is  $m$ , it is called an  $m$ -ple lattice. Square Lattice designs can be constructed as depicted in next slide.

## Statistical Analysis

Similar to BIBD, consider the model :  $y_{ij} = \mu + \tau_i + \beta_j + e_{ij}$

Observation = General Mean + Treatment Effect + Block Effect + Random Error.

Random errors are assumed to be independently and identically distributed normally with mean zero and constant variance  $\sigma^2$ .

Now we compute -

$G$  = Grand Total of observations.

$\bar{y}$  = Grand Mean =  $G/n$ , where  $n = ms^2 = 2s^2$ , where,  $s$  = block size

$T_i$  = Sum of observations for treatment  $i$ , ( $i = 1, 2, \dots, s^2$ )

$B_j$  = Sum of observations in block  $j$ , ( $j = 1, 2, \dots, ms$ ), where  $m = 2$

$CF = G^2/n$ , where  $n = ms^2 = 2s^2$

$Q_i$  = Adjusted  $i^{\text{th}}$  treatment total

=  $T_i - (\text{Sum of block totals in which treatment } i \text{ occurs})/\text{Block size (s)}$

## Statistical Analysis

$S_R(Q_i)$  = Sum of Q's for treatments including i which are in the same row of treatment i in the standard array.

$S_C(Q_i)$  = Sum of Q's for treatments including i which are in the same column of treatment i in the standard array.

Adjusted treatment mean for treatment i =  $i^{\text{th}}$  treatment effect ( $\hat{\tau}_i$ ) + Grand Mean ( $\bar{y}$ )

Various Sums of Squares can be obtained as follows –

- (i) Total Sum of Squares (TSS) =  $\Sigma(\text{observations})^2 - \text{CF}$
- (ii) Treatments Sum of Squares unadjusted ( $SST_U$ )  
=  $\Sigma(T_i)^2/m - \text{CF}$
- (iii) Blocks Sum of Squares unadjusted ( $SSB_U$ ) =  $\Sigma(B_j)^2/s - \text{CF}$
- (iv) Treatments Sum of Squares adjusted ( $SST_A$ ) =  $\Sigma t_i Q_i$
- (v) Error Sum of Squares (SSE) = TSS –  $SSB_U - SST_A$
- (vi) Blocks Sum of Squares adjusted ( $SSB_A$ )  
=  $SST_A + SSB_U - SST_U$

The Analysis of variance for an  $m$ -ple lattice with  $v = s^2$  treatments in blocks of size 's' is given below -

ANOVA including both Treatments & Blocks

Source of Variation	DF	SS	MS	$F_{\text{calculated}}$
SSBlocks (unadj.)	$ms - 1$	$SSB_U$		
SSTreatments (adj.)	$s^2 - 1$	$SST_A$	$MST$	$MST/MSE$
SSBlocks (adj.)	$ms - 1$	$SSB_A$	$MSB$	$MSB/MSE$
SSError	$(s - 1)(ms - s - 1)$	$SSE$	$MSE$	
SSTotal	$ms^2 - 1$	$TSS$		

Ex : The following table gives the synthetic yields per plot of an experiment conducted with  $3^2 = 9$  treatments using a simple lattice design.

Replication 1				Replication 2			
Blocks	Treatments (yield per plot)			Blocks	Treatments (yield per plot)		
1	1 (8)	7 (5)	4 (3)	4	8 (2)	7 (2)	9 (7)
2	3 (3)	6 (2)	9 (6)	5	4 (3)	5 (3)	6 (3)
3	8 (3)	5 (7)	2 (3)	6	2 (2)	3 (4)	1 (6)

Analyze the data.

Analysis :

$$\text{Grand Total, } G = 8 + 5 + \dots + 6 = 72$$

$$\text{No. of observations, } n = 18$$

$$\text{Grand Mean, } \bar{y} = G/n = 72/18 = 4, \text{ No. of replications} = 2$$

$$\text{Block size, } k = 3, \text{ Correction Factor (CF)} = G^2/n = 72^2/18 = 288$$

---

Note :  $T_1 = 8 + 6 = 14$ , etc.

$$B_1 = 8 + 5 + 3 = 16$$

Total of Block, in which treatment 1 occurs,

$$\sum_{j(1)} B_j = (8 + 5 + 3) + (2 + 4 + 6) = 16 + 12 = 28, \text{ etc.}$$

$$\text{Adjusted Treatment Total, } Q_1 = T_1 - \sum_{j(1)} B_j / k = 14 - 28/3 = 4.67$$

$$\hat{\tau}_1 = Q_1/2 + [S_R(Q_1) + S_C(Q_1)]/(2 \times 3) = 4.67/2 + 1.001/6 = 2.50$$

Specimen examples :  $T_1 = 8 + 6 = 14, \dots, T_5 = 7 + 3 = 10, \dots, T_9 = 6 + 7 = 13$

$B_1 = 8 + 5 + 3 = 16, \dots, B_4 = 2 + 2 + 7 = 11, \dots, B_6 = 2 + 4 + 6 = 12$

$\sum_{j(1)} B_j = (8 + 5 + 3) + (2 + 4 + 6) = 16 + 12 = 28$ , (Total of Block, in which treatment 1 occurs)....,

$\sum_{j(5)} B_j = (8 + 5 + 3) + (2 + 4 + 6) = 16 + 12 = 28$  etc.

$$\hat{\tau}_1 = Q_1/2 + [S_R(Q_1) + S_C(Q_1)]/(2 \times 3) = 4.67/2 + 1.001/6 = 2.50$$

Adjusted treatment mean for treatment  $i = i^{\text{th}}$  treatment effect ( $\tau_i$ ) + Grand Mean ( $\bar{y}$ )

Grand Mean,  $\bar{y} = G/n = 72/18 = 4$

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treat/ Block No.	$T_i$	$B_j$	Block Nos. in which treat $i$ occurs	$\sum_{j(i)} B_j$	$\sum_{j(i)} B_j/k$	$Q_i$ (2) - (6)	$\hat{\tau}_i$	Adj. treat. Mean
1	14	16	1, 6	28	9.33	4.66	2.50	6.50
2	5	11	3, 6	25	8.33	-3.33	-2.16	1.84
3	7	13	2, 6	23	7.66	-0.66	0.33	4.33
4	6	11	1, 5	25	8.33	-2.66	-1.33	2.66
5	10	9	3, 5	22	7.33	2.66	0.50	4.50
6	5	12	2, 5	20	6.66	-1.66	-0.50	3.50
7	7		1, 4	27	9.00	-2.00	-0.83	3.16
8	5		3, 4	24	8.00	-3.00	-2.00	2.00
9	13		2, 4	22	7.33	5.66	3.50	7.50

Total Sum of Squares (TSS)	$= \sum (\text{observation})^2 - \text{CF}$ $= 8^2 + 5^2 + \dots + 6^2 - 72^2/18 = 66$
Treatment Sum of Squares unadjusted ( $\text{SST}_U$ )	$= [\sum (T_i)^2]/m - \text{CF}$ $= (14^2 + \dots + 13^2)/2 - 72^2/18 = 49$
Block Sum of Squares unadjusted ( $\text{SSB}_U$ )	$= [\sum (B_j)^2]/s - \text{CF}$ $= (16^2 + \dots + 12^2)/3 - 72^2/18 = 9.33$
Treatment Sum of Squares adjusted ( $\text{SST}_A$ )	$= \sum \hat{\tau}_i Q_i = 51.44$
Block Sum of Squares adjusted ( $\text{SSB}_A$ )	$= \text{SST}_A + \text{SSB}_U - \text{SST}_U$ $= 51.44 + 9.33 - 49 = 11.77$
Error Sum of Squares (SSE)	$= \text{TSS} - \text{SSB}_U - \text{SST}_A$ $= 66 - 51.44 - 9.33 = 5.23$

ANOVA Table is shown in the next slide.

**ANOVA Table**

Source of Variation	DF	SS	MS	Fcalculated	Fcritical
SSBlocks (unadj.)	5	9.33			
SSTreatments (adj.)	8	51.44	6.43	4.91	6.04
SSBlocks (adj.)	5	11.77	2.35	1.79	6.26
SSTreatments (unadj.)	8	49.00			
SSError	17 - (5 + 8) = 4	5.23	1.31		
SSTotal	17				

$F_{0.05,8,4}$  and  $F_{0.05,5,4}$  values obtained from F-table are 6.04 & 6.26 respectively.

As  $F_{\text{calculated}}$  values  $< F_{\text{critical}}$  value we conclude that Treatment effects and block effects are not significantly different.

## Advantages and Disadvantages of Lattice Designs

The main advantage of Balanced Lattices is that a large number of treatments may be compared within relatively small blocks. Another advantage of balanced lattice designs is that each pair of treatments is compared with the same degree of precision because each treatment occurs together in the same block with every other treatment an equal number of times (usually once). Hence, to obtain a balanced lattice, some restrictions on the number of treatments and the number of blocks in the design are required. Consequently, balanced lattices are not available for 36, 100 and 144 treatments. The disadvantages of the design are the limitations for the number of allowable treatments, block sizes and replication. The analysis also becomes more complex and the designs are more difficult to construct as the number of treatments increase.

## Youden Square Design

A Youden square design (YSD) is a design with incomplete columns by means of which two sources of variation can be eliminated. The rows of YSD form a Randomised Block Design (RBD) and the columns form a Balanced Incomplete Block Design (BIBD). These are basically symmetrical BIBD by which the block to block variation can be eliminated. The  $k$  units in each block be thought of occupying  $k$  different positions. With the help of YSD, the effects of such positions can be eliminated.

Youden square is a reduction from Randomized Complete Blocks to a Balanced Incomplete Design in which one row or column has been removed from a Latin square (so it is no longer square). It is a useful design for balancing out the effects of treatment order in a repeated-measures sequence.

## Youden Square Design

Definition : A YSD is an arrangement of  $v$  treatments in a  $k \times v$  rectangular array such that every symbol occurs exactly once in each row and the columns form a symmetrical Balanced Incomplete Block Design (BIBD) with parameters  $v = b$ ,  $r = k$ ,  $\lambda$ . Here,  $k$  = no. of rows,  $v$  = no. of columns,  $b$  = no. of treatment,  $r$  = no. of replicate and  $\lambda$  = no. of pair replicate.

YSD with 7 treatments arranged in 3 rows and 7 columns

1	2	3	4	5	6	7
2	3	4	5	6	7	1
4	5	6	7	1	2	3

## Model of YSD

$$y_{ij(m)} = \mu + \rho_i + c_j + \tau_m + e_{ij(m)} \quad \dots \dots \dots \quad (a)$$

Where  $\mu$  is the general mean,  $\rho_i$  is the  $i^{\text{th}}$  row effect,  $c_j$  is the  $j^{\text{th}}$  column effect,  $\tau_m$  is the  $m^{\text{th}}$  treatment effect and  $e_{ij(m)}$  are random errors assumed to be independently normally distributed. The model of response for a YSD is the same as in (a) with  $i = 1, 2, \dots, k$ ;  $j = 1, 2, \dots, v$  and  $m = 1, 2, \dots, v$ .

Let  $R_1, R_2, \dots, R_k$  be the  $k$  row totals;  $C_1, C_2, \dots, C_v$  be the  $v$  column totals,  $T_1, T_2, \dots, T_v$  be the  $v$  treatment totals and  $G$  be the grand total.

Further let  $Q_m$  be adjusted  $m^{\text{th}}$  treatment total obtained by subtracting the sum of column means in which  $m^{\text{th}}$  treatment occurs from  $T_m$  ( $m = 1, 2, \dots, v$ ). Further let  $Q = (Q_1, Q_2, \dots, Q_m)'$ .

$$\hat{\tau}_m = (k/\lambda v)Q_m, \quad m = 1, 2, \dots, v \quad \text{where, } Q_m = T_m - \sum_{j(m)} C_j / k$$

$H_{0R}$  : All Row effects are equal or All Row Means are equal

$H_{0C}$  : All Column effects are equal or All Column Means are equal

$H_{0T}$  : All Treatment effects are equal or All Treatment Means are equal

Analyze the following YSD. The yields are given below with 4 treatments.

	$C_1$	$C_2$	$C_3$	$C_4$
$R_1$	2 (1)	3 (2)	1 (3)	4 (2)
$R_2$	3 (2)	1 (3)	4 (5)	2 (4)
$R_3$	4 (4)	2 (1)	3 (2)	1 (5)

Here,  $k$  = no. of rows = 3,  $v$  =  $b$  = no. of columns = 4 = no. of treatments, no. of replication = 4,  $n$  = 12.

$$R_1 = 1 + 2 + 3 + 2 = 8; R_2 = 2 + 3 + 5 + 4 = 14; R_3 = 4 + 1 + 2 + 5 = 12$$

$$C_1 = 1 + 2 + 4 = 7; C_2 = 2 + 3 + 1 = 6; C_3 = 3 + 5 + 2 = 10; C_4 = 2 + 4 + 5 = 11$$

$$T_1 = 3 + 3 + 5 = 11; T_2 = 1 + 1 + 4 = 6; T_3 = 2 + 2 + 2 = 6; T_4 = 4 + 5 + 2 = 11$$

$$G = 8 + 14 + 12 = 34; \text{ CF} = G^2/n = 34^2/12 = 96.33$$

$$\begin{aligned} \text{RSS} = \sum y^2 &= 1^2 + 2^2 + 3^2 + 2^2 + 2^2 + 3^2 + 5^2 + 4^2 + 4^2 + 1^2 + 2^2 + 5^2 \\ &= 18 + 54 + 46 = 118 \end{aligned}$$

$$Q_1 = T_1 - (C_2 + C_3 + C_4)/3 = 11 - (6 + 10 + 11)/3 = 2$$

$$Q_2 = T_2 - (C_1 + C_2 + C_4)/3 = 6 - (7 + 6 + 11)/3 = -2$$

$$Q_3 = T_3 - (C_2 + C_1 + C_3)/3 = 6 - (6 + 7 + 10)/3 = -1.66$$

$$Q_4 = T_4 - (C_4 + C_3 + C_1)/3 = 11 - (11 + 10 + 7)/3 = 1.66$$

$$\text{TSS} = \sum y^2 - \text{CF} = 118 - 96.33 = 21.67$$

$$\text{SSR} = [\sum (R_i)^2]/v - \text{CF} = (8^2 + 14^2 + 12^2)/4 - 96.33 = 101 - 96.33 = 4.67$$

$$\text{SSC} = [\sum (C_i)^2]/k - \text{CF} = (7^2 + 6^2 + 10^2 + 11^2)/3 - 96.33 = 102 - 96.33 = 5.67$$

$$\text{SST} = [k \sum (Q_m)^2]/\lambda v = [3\{(2)^2 + (-2)^2 + (-1.66)^2 + (1.66)^2\}]/3 \times 4 = 3.36$$

$$\text{SSE} = \text{TSS} - \text{SSR} - \text{SSC} - \text{SST} = 21.67 - 4.67 - 5.67 - 3.36 = 7.97$$

ANOVA Table

Source of Variation	DF	SS	MS	$F_{\text{calculated}}$	$F_{\text{critical}}$
SSRows	$k - 1 = 2$	4.67	$4.67/2 = 2.335$	$2.335/2.66 = 0.88$	9.55
SSColumns	$v - 1 = 3$	5.67	$5.67/3 = 1.89$	$1.89/2.66 = 0.71$	9.28
SSTreatments (eliminating rows & columns)	$v - 1 = 3$	3.36	$3.36/3 = 1.12$	$1.12/2.66 = 0.42$	9.28
SSError	$vk - 2v - k + 2$ $= 12 - 8 - 3 + 2 = 3$	7.97	$7.97/3 = 2.66$		
SSTotal	$vk - 1 = 11$	21.67			

$k$  = no. of rows = 3,  $v$  = b = no. of columns = 4 = no. of treatments, no. of replication = 4,  $n$  = 12.

The significance of treatment effects can be tested by comparing the F-value of table for  $F_{0.05,(v-1),(vk-2v-k+2)}$ . The calculation of treatment SS is identical to the one for a BIB design and the other calculations like Rows SS, Columns SS are the routine ones.

Since the calculated F for all parameters less than Fcritical values, null hypothesis for  $H_{0R}$ ,  $H_{0C}$  and  $H_{0T}$  are all acceptable.

Note : Error SS = Total SS – Rows SS – Columns SS -Treatments (eliminating rows & columns)  
 $SS = 21.67 - 4.67 - 5.67 - 3.36 = 7.97$

Illustration : In one of the experiments, the experimenter is interested in making comparison among 7 treatments and there are 28 experimental units available. These 28 experimental units are arranged in a Youden Square Design with 4 rows and 7 columns with one observation per cell. The parameters of the design are  $v$  (number of treatments) = 7,  $p$  (number of rows) = 4,  $q$  (number of columns) = 7,  $r$  (replication of treatments) = 4. The layout of the design along with the observations is given below -

2(4.00)	3(5.30)	4(1.10)	5(16.90)	6(16.90)	7(10.30)	1(294.00)
7(17.50)	1(220.00)	2(12.20)	3(15.50)	4(11.00)	5(26.50)	6(27.20)
6(37.00)	7(26.00)	1(310.00)	2(22.70)	3(24.20)	4(21.40)	5(31.30)
5(46.80)	6(44.20)	7(34.30)	1(282.00)	2(33.70)	3(33.70)	4(30.50)

The analysis of variance for the above design for testing the equality of treatment effects is obtained in the next slide.

---

ANOVA Table

Source of Variation	DF	SS	MS	Fcalculated	Fcritical
Due to Rows	3	3319.0068	1106.336	4.89	3.49
Due to Columns (unadjusted)	6	27595.4086	4599.2348	20.31	3.00
Treatments (adjusted)	6	195027.72	32504.6207	143.56	3.00
Error	12	2717.0814	226.4235		
Total	27	228659.2211			

no. of rows = 4, no. of columns = 7, no. of treatments = 7, no. of replication = 4.

$$F_{0.05,3,12} = 3.49; \quad F_{0.05,6,12} = 3$$



# BSDCHZC355

## Statistical Inference & Applications

**BITS** Pilani  
Pilani | Dubai | Goa | Hyderabad

Shaibal Kr. Sen  
Session 11



**STATISTICAL INFERENCES**

**&**

**APPLICATIONS**

## Scheduled Topics to be covered

---

Factorial Experiments -

Basic Definitions and principles

Interactions and Analysis of -

$2^2$  Factorial experiments

$2^3$  Factorial experiments

## Factorial Designs

Main Effects and Interaction Effects Simultaneously:

To study the significant difference between treatment (or levels of factor) effects (main effects).

To study the significant difference between interactive treatment effects.

---

Experiments where the effects of more than one factor, say variety, manure etc. are considered together are called factorial experiments, while experiments with one factor, say only variety or manure, may be called simple experiments. Consider a simple case If there are  $p$  different varieties, then we shall say that there are  $p$  levels of the factor 'variety'. Similarly, the second factor 'manure' may have  $q$  levels, ie. there may be  $q$  different manures or different doses of the same manure. Then this factorial design will be called a  $p \times q$  experiment. As a different example, the two factors may be two different manures, say nitrogen and phosphate and at different doses, and  $p$  and  $q$  different doses, respectively. Then this will also give a  $p \times q$  experiment. We shall consider only the simplest cases, viz. cases of a  $n$  factors each at 2 levels or what are known as  $2^n$  experiments, where  $n$  is any positive integer greater than or equal to 2.

---

---

Let us examine a two-factor factorial so that we can set up some basic notation as well as introduce some fundamental ideas. Notationally the designs are  $2^K$ , where  $K$  = number of factors, and  $2^K$  = number of treatment combination runs (usually just called “treatments”).

For example if  $K = 2$ , the design requires  $2^2 = 4$  runs.

---

A  $2^2$  experiment : Let us consider two factors – A and B, each at 2 levels. Following Yates, we denote by a or b one of two levels at which the corresponding factor (denoted by capital letter) occurs and for definiteness we shall call this the second level. The first level of A or B will be signified by the absence of the corresponding letter in the treatment combination. Now with 2 factors, each at 2 levels, there will be  $2 \times 2 = 4$  treatment combinations. They are enumerated below –

- (1) : A and B both at first levels or at lower level,  
a : A at second level and B at first level or A at high level & B at low level,  
b : A at first level and B at second level or A at low level & B at high level,  
ab : A and B both at second levels or A at high level & B at high level,

---

These four treatment combinations may be compared using a CRD (Completely Randomized Design) Or RBD (Randomized Block Design) or an LSD (Latin Square Design). For a  $2^2$  experiment in  $r$  blocks, the treatment combinations  $t = 4$ . And the analysis of a  $2^2$  experiment in a Latin square design  $m = 4$ . In a factorial experiment one is more interested in the separate tests about Main Effects and Interaction Effects, which are performed by splitting the treatment SS carrying 3 d.f. into 3 orthogonal components, each carrying a single d.f. and each associated either with a main effect or an interaction.

Main effect and Interaction effect : The symbols  $[a]$  and  $(a)$  will be used to denote the total and mean (respectively) of all observations receiving the treatment combination  $a$ . The letters  $A$ ,  $B$  and  $AB$ , when they refer to numbers, will be used to stand for main effects due to factors  $A$  and  $B$  and  $AB$  due to the interaction of the two factors.

---

It is convenient to obtain the factorial effects and their SSs from the treatment totals rather than from the treatment means. We define the factorial effect totals as follows –

$$\left. \begin{array}{l} [A] = [ab] - [b] + [a] - [1] \\ [B] = [ab] + [b] - [a] - [1] \\ [AB] = [ab] - [b] - [a] + [1] \end{array} \right\} \dots \dots \dots \quad (i)$$

Then the SS due to any main effect or the interaction effect is obtained by multiplying the square of the effect total with the reciprocal of  $4r$ , where  $r$  is the common replication number. Thus

$$\begin{aligned}
 \text{SS due to main effect of A} &= [A]^2/4r \text{ with 1 d.f.} \\
 \text{SS due to main effect of B} &= [B]^2/4r \text{ with 1 d.f.} \\
 \text{SS due to interaction effect of AB} &= [AB]^2/4r \text{ with 1 d.f.}
 \end{aligned} \quad \dots \dots \dots \text{(ii)}$$

### Yates' Method for a $2^2$ Experiment

Treatment combination (1)	Total (2)	(3)	(4)	
(1)	[1]	[1] + [a]	$[1] + [a] + [b] + [ab] = \text{Grand total}$	
Fixed order	a	[a]	$[a] - [1] + [ab] - [b] = [A]$	
	b	[b]	$[b] + [ab] - [1] - [a] = [B]$	Fixed order
	ab	[ab]	$[ab] - [b] - [a] + [1] = [AB]$	

It is then a simple matter to express the factorial effect totals or the SSs in terms of the factorial effects, main or interaction, remembering that a factorial effect total is  $2r$  times the corresponding factorial effect. Thus the factorial effects are as follows –

$$\left. \begin{array}{l} \text{Main effect of A} = [A]/2r \\ \text{Main effect of B} = [B]/2r \\ \text{Interaction AB} = [AB]/2r \end{array} \right\} \dots \dots \dots \text{(iii)}$$

and the SS due to a factorial effect is  $rx(\text{factorial effect})^2$ .

The test for the significance of any factorial effect, main effect or interaction, may now be obtained by computing –

$$F = (\text{MS due to factorial effect})/\text{MSE}$$

### ANOVA Table for $2^2$ experiment in r Randomised Blocks

Source of variation	d.f.	SS	MS	$F_{\text{calculated}}$
Blocks	$r - 1$	SS (Blocks)	MS (Blocks)	
Main effect A	1	$[A]^2/4r$	MSA	MSA/MSE
Main effect B	1	$[B]^2/4r$	MSB	MSB/MSE
Interaction AB	1	$[AB]^2/4r$	MS(AB)	MS(AB)/MSE
Error	$3(r - 1)$	By subtraction	MSE	
Total	$4r - 1$	$\sum (y_{ij} - y_{\text{oo}})^2$		

---

Ex : A  $2^2$  experiment : in six ( $Y = 6$ ) randomized blocks was conducted in order to obtain an idea of the interaction : spacing (s) x number of seedlings per hole (N) along with the effects of different types of spacing and different numbers of seedlings per hole, while adopting the Japanese method of cultivation.

The levels of two factors are –

$S : \begin{cases} 8'' \text{ spacings in between} \\ 10'' \text{ spacings in between} \end{cases}$

and

$N : \begin{cases} 3 \text{ seedlings per hole} \\ 4 \text{ seedlings per hole} \end{cases}$

The field plan and yield of dry Aman paddy (in kg.) are given in next slide.

Block 1				Ttl	Block 2				Ttl	Block 3				Ttl
(1)	s	ns	n		ns	(1)	s	n		(1)	n	s	ns	
117	106	109	114	446	114	120	117	114	465	111	117	114	106	448
Block 4					Block 5					Block 6				
ns	n	s	(1)	439	ns	s	(1)	n	283	n	(1)	ns	s	361
98	121	112	108		75	97	73	38		58	81	105	117	

Analyse the data to find out if there are any significant treatment effects – main or interaction.

We apply Yates' method to find the total effects.

### Yates' Method for a $2^2$ Experiment

Treatment combination (1)	Total yield from all blocks (2)	(3)	(4)	Average effect
(1)	$[1] = 610$	$[1] + [n] = 1172$	$[1] + [n] + [s] + [ns] = 2442 = \text{Grand total (G)}$	
n	$[n] = 562$	$[s] + [ns] = 1270$	$[n] - [1] + [ns] - s = -104 = [N]$	$[N]/2 \times 6$
s	$[s] = 663$	$[n] - [1] = -48$	$[s] + [ns] - [1] - [n] = 98 = [S]$	$[S]/2 \times 6$
ns	$[ns] = 607$	$[ns] - [s] = -56$	$[ns] - [s] - [n] + [1] = -8 = [NS]$	$[NS]/2 \times 6$

Now perform the randomized block analysis –

Treatment totals are :  $[1] = 117 + 120 + 111 + 108 + 73 + 81 = 610$

$[n] = 114 + 114 + 117 + 121 + 38 + 58 = 562$

$[s] = 106 + 117 + 114 + 112 + 97 + 117 = 563$  and

$[ns] = 109 + 114 + 106 + 98 + 75 + 105 = 607.$

Grand Total (G) = [1] + [n] + [s] + [ns] = 610 + 663 + 562 + 607 = 2442

Six block totals are : 446, 465, 448, 439, 283 and 361.

Raw total SS =  $\Sigma y^2 = 117^2 + 106^2 + \dots + 117^2 = 2,59,024$

Correction Factor (CF) =  $G^2/N = (2442)^2/24 = 2,48,473.5$

Total SS (TSS) = RSS – CF = 2,59,024 – 2,48,473.5 = 10,550.5

$$\begin{aligned} \text{Block SS (BSS)} &= (446^2 + 465^2 + \dots + 361^2)/4 - CF \\ &= 2,54,744 - 2,48,473.5 = 6270.5 \end{aligned}$$

$$\begin{aligned} \text{Treatment SS (TrSS)} &= \{(610^2 + \dots + 607^2)/6\} - 2,48,473.5 \\ &= 1495962/6 - 2,48,473.5 = 853.5 \end{aligned}$$

Error SS = TSS – BSS – TrSS = 10,550.5 – 6,270.5 - 853.5 = 3,426.5

---

Also, SS due to N = [N]<sup>2</sup>/4r = (-104)<sup>2</sup>/24 = 450.667

SS due to S = [S]<sup>2</sup>/4r = (98)<sup>2</sup>/24 = 400.167

SS due to NS = [NS]<sup>2</sup>/4r = (-8)<sup>2</sup>/24 = 2.67

ANOVA Table is presented in next slide

ANOVA Table for  $2^2$  experiment

Source of variation	d.f.	SS	MS	$F_{\text{calculated}}$	$F_{\text{critical}}$
Blocks	5	6,270.50	1254.1		$F_{0.05,1,15} = 4.54$
N	1	450.667	450.667	1.973	
S	1	400.167	400.167	1.752	
NS	1	2.667	2.667	<1	
Error	$23 - 8 = 15$	3,426.50	228.433		
Total	23	10,550.50			

There is no significant main or interaction effects present in the above experiment, as in each of the cases the computed value of F is less than the corresponding theoretical value (critical value) at the 5% level.

---

A  $2^3$  experiment : We now consider the case of three factors – A, B and C, each at 2 levels, where a, b and c will denote the second levels of the factors, respectively,  $2 \times 2 \times 2 = 8$  treatment combinations written in the systematic order are : (1), a, b, ab, c, ac, bc, abc.

The 8 treatment combinations may be compared in any of the designs viz. – CRD, RBD or LSD. The analysis will be same as in the corresponding design, the number of treatments being  $t = 8$  in CRD and RBD and  $m = 8$  in LSD. The treatment SS has 7 d.f. We next divide it into 7 orthogonal contrasts of the 8 treatment means (or totals) with the help of the main effects and interactions. In a three-factor experiment there are three main effects – A, B, C; and three first order interactions – AB, AC, BC; and one second-order (or three-factor) interaction – ABC.

#### Interaction effects

$$AB = \{(1) - a - b + ab + c - ac - bc + abc\} / 4$$

$$AC = \{(1) - a + b - ab - c + ac - bc + abc\} / 4$$

$$BC = \{(1) + a - b - ab - c - ac + bc + abc\} / 4$$

$$ABC = \{-(1) + a + b - ab + c - ac - bc + abc\} / 4$$

SS due to factorial effects and tests of significance of factorial effects :

We define factorial effect totals by combining the 8 treatment totals with signs given in the above table. Thus

$$[A] = [abc] - [bc] + [ac] - [c] + [ab] - [b] + [a] - [1]$$

and similarly the other effect totals are obtained.

The SS due to factorial effect is obtained by multiplying the square of

---

the corresponding effect total by the reciprocal of  $8r$ , where  $r$  is common replication number. Thus, the test for the significance of any factorial effect, main effect or interaction may now be obtained by computing

$$F = \text{MS due to factorial effect} / \text{MSE}$$

where, MSE is the error MS of the analysis of variance table of the corresponding design. This  $F$  follows the  $F$  distribution with  $1, 7(r - 1)$  d.f. Hence, the hypothesis of the absence of the factorial effect is rejected at the level  $\alpha$  if for our data

$$F > F_{\alpha, 1, 7(r-1)};$$

otherwise, the hypothesis is accepted.  $7(r - 1)$  is the error d.f. for a  $2^3$  experiment conducted in  $r$  randomized blocks.

Run	Labels	A	B	C	AB	AC	BC	ABC
1	(1)	-	-	-	+	+	+	-
2	a	+	-	-	-	-	+	+
3	b	-	+	-	-	+	-	+
4	ab	+	+	-	+	-	-	-
5	c	-	-	+	+	-	-	+
6	ac	+	-	+	-	+	-	-
7	bc	-	+	+	-	-	+	-
8	abc	+	+	+	+	+	+	+

### Interaction effects

$$AB = \{(1) - a - b + ab + c - ac - bc + abc\} / 4$$

$$AC = \{(1) - a + b - ab - c + ac - bc + abc\} / 4$$

$$BC = \{(1) + a - b - ab - c - ac + bc + abc\} / 4$$

$$ABC = \{-(1) + a + b - ab + c - ac - bc + abc\} / 4$$

Yates' method of computing factorial effect totals for a $2^3$ experiment				
Treatment combination (1)	Total (2)	(3)	(4)	(5)
1)	[1]	[1] + [a]	[b] + [ab] + [1] + [a]	$[bc] + [abc] + [c] + [ac] + [b] + [ab] + [1] + [a] = \text{Grand total (G)}$
a	[a]	[b] + [ab]	[bc] + [abc] + [c] + [ac]	$[abc] - [bc] + [ac] - [c] + [ab] - [b] + [a] - [1] = [A]$
b	[b]	[c] + [ac]	[ab] - [b] + [a] - [1]	$[bc] + [abc] - [c] - [ac] + [b] + [ab] - [1] - [a] = [B]$
ab	[ab]	[bc] + [abc]	$[abc] - [bc] + [ac] - [c]$	$[abc] - [bc] - [ac] + [c] + [ab] - [b] - [a] + [1] = [AB]$
Half	c	[c]	[a] - [1]	$[bc] + [abc] + [c] + [ac] - [b] - [ab] - [1] - [a] = [C]$
	ac	[ac]	[ab] - [b]	$[abc] - [bc] + [ac] - [c] - [ab] + [b] - [a] + [1] = [AC]$
	bc	[bc]	[ac] - [c]	$[bc] + [abc] - [c] - [ac] - [b] - [ab] + [1] + [a] = [BC]$
	abc	[abc]	[abc] - [bc]	$[abc] - [bc] - [ac] + [c] - [ab] + [b] + [a] - [1] = [ABC]$

ANOVA Table for a  $2^3$  experiment in  $r$  Randomised Blocks

Source of variation	d.f.	SS	MS	F
Blocks	$r - 1$	SS (Blocks)	MS (Blocks)	
Main effect A	1	$[A]^2/8r$	MSA	MSA/MSE
Main effect B	1	$[B]^2/8r$	MSB	MSB/MSE
Main effect C	1	$[C]^2/8r$	MSC	MSC/MSE
Two factor Interaction AB	1	$[AB]^2/8r$	MS(AB)	MS(AB)/MSE
Two factor Interaction AC	1	$[AC]^2/8r$	MS(AC)	MS(AC)/MSE
Two factor Interaction BC	1	$[BC]^2/8r$	MS(BC)	MS(BC)/MSE
Three factor Interaction ABC	1	$[ABC]^2/8r$	MS(ABC)	MS(ABC)/MSE
Error	$7(r - 1)$	SSE (by subtraction)	MSE	
Total	$8r - 1$	$\Sigma(y_{ij} - y_{oo})^2$		

## **Yates' method of computing factorial effect totals**

Yates gives a systematic method of obtaining the various effect totals for any  $2^n$  experiment without writing down the algebraic expressions. We shall describe it for the  $2^2$  experiment, but it can be easily extended to the case of any  $2^n$  experiment.

The steps are as follows –

- (i) First write down the 4 treatment combinations systematically in the 1<sup>st</sup> column, starting with the treatment combination (1) and then introducing the letters a, b in turn. After introducing a letter, write down its combination with all the previous treatment combinations and then introduce a new letter. Repeat it until all the letters (n letters in case of a  $2^n$  experiment) have exhausted.
- (ii) Next, write down the treatment total from all the replicates in the 2<sup>nd</sup> column against the appropriate treatment combination.

## **Yates' method of computing factorial effect totals**

- (iii) The first two columns we get from the observed data. For obtaining column 3, we break the even number of values in the 2<sup>nd</sup> column into consecutive pairs (1, 2; 3, 4; etc.) Then in the 1<sup>st</sup> half of the 3<sup>rd</sup> column we write down the sums of the values in these pairs in order and in the 2<sup>nd</sup> half of 3<sup>rd</sup> column we write down in order the differences of the values in the pairs in the 2<sup>nd</sup> column (the 1<sup>st</sup> member subtracted from 2<sup>nd</sup> member of a pair).
- (iv) We next break the values in the 3<sup>rd</sup> column into consecutive pairs and put the sums and differences of the members of these pairs in order in the 4<sup>th</sup> column.

For a 2<sup>2</sup> experiment the 4<sup>th</sup> column values give the factorial effect totals corresponding to the treatment combinations occurring in the corresponding position of the 1<sup>st</sup> column.

## Yates' method of computing factorial effect totals

For a  $2^n$  experiment we are to repeat  $n$  times the operations of columns 3 and 4 and then the values in the  $(n + 2)^{\text{nd}}$  column will be the factorial effect totals, the 1<sup>st</sup> entry in the last column being always the grand total.



# BSDCHZC355

## Statistical Inference & Applications

**BITS** Pilani  
Pilani | Dubai | Goa | Hyderabad

Shaibal Kr. Sen  
Session 12



**STATISTICAL INFERENCES**

**&**

**APPLICATIONS**

## Scheduled Topics to be covered

---

3<sup>2</sup> Factorial experiments

Split Plot Design

## 3<sup>2</sup> Design

When factors are taken at three levels instead of two, the scope of an experiment increases. It becomes more informative. A study to investigate if the change is linear or quadratic is possible when the factors are at three levels. The more the number of levels the better, yet the number of the levels of the factors cannot be increased too much as the size of the experiment as the size of the experiment increases too rapidly with them. Let us begin with two factors A and B, each at three levels, say 0, 1 and 2 (3<sup>2</sup> factorial experiment). The treatment combinations are –

00	$a_0b_0$	(1)	Factors A and B both at 1 <sup>st</sup> levels
10	$a_1b_0$	a	Factors A is at 2 <sup>nd</sup> level and B is at 1 <sup>st</sup> level
20	$a_2b_0$	$a^2$	Factors A is at 3 <sup>rd</sup> level and B is at 1 <sup>st</sup> level
01	$a_0b_1$	b	Factors A is at 1 <sup>st</sup> level and B is at 2 <sup>nd</sup> level
11	$a_1b_1$	ab	Factors A and B both at 2 <sup>nd</sup> level
21	$a_2b_1$	$a^2b$	Factors A is at 3 <sup>rd</sup> level and B is at 2 <sup>nd</sup> level
02	$a_0b_2$	$b^2$	Factors A is at 1 <sup>st</sup> level and B is at 3 <sup>rd</sup> level
12	$a_1b_2$	$ab^2$	Factors A is at 2 <sup>nd</sup> level and B is at 3 <sup>rd</sup> level
22	$a_2b_2$	$a^2b^2$	Factors A and B both at 3 <sup>rd</sup> levels

## 3<sup>2</sup> Design

As learnt during previous session there will be 7 factorial effects – 3 main factor effects (A, B & C), 3 two factor Interaction effects (AB, AC & BC) and 1 three factor Interaction effect (ABC).

Any standard design can be adopted for the experiment. The main effects A, B can respectively be divided into linear and quadratic components each with 1 df as  $A_L$ ,  $A_Q$ ,  $B_L$  and  $B_Q$ . Accordingly AB can be partitioned into four components as  $A_L B_L$ ,  $A_L B_Q$ ,  $A_Q B_L$ ,  $A_Q B_Q$ , each 1 df. The coefficients of the treatment combinations to obtain the above effects are given as –

Nine (3<sup>2</sup>) Treatment Totals : [1] [a] [a<sup>2</sup>] [b] [ab] [a<sup>2</sup>b] [b<sup>2</sup>] [ab<sup>2</sup>] [a<sup>2</sup>b<sup>2</sup>]

## Split-Plot Design.

In some multifactor factorial experiments, we may not be able to completely randomize the order of the runs. This often results in a generalization of the factorial design called a split-plot design.

As an example, consider a paper manufacturer who is interested in three different pulp preparation methods. The methods differ –

- (i) in the amount of hardwood in the three different pulp mixtures and
- (ii) four different cooking temperatures for the pulp.

The experimenter wishes to study the effect of these two factors on the tensile strength of the paper. Each replicate of a factorial experiment requires  $3 \times 4 = 12$  observations and the experimenter has decided to run three replicates. This will require total 36 runs. The experimenter decides to conduct the experiment as follows. A batch of pulp is produced by one of the three methods under study. Then this batch is

---

divided into four samples and each sample is cooked at one of the four temperatures. Then a second batch of pulp is made up using another of the three methods. This second batch is also divided into four samples that are tested at the four temperatures. The process is then repeated until all three replicates (36 runs) are obtained. The data are shown in table above.

Initially, we might consider this to be factorial experiment with three levels of preparation method (factor A) and four levels of temperature (factor B). If this is the case, then the order of experimentation within

---

each replicate should be completely randomized. That is, we should randomly select a treatment combination (a preparation method and a temperature) and obtain an observation, then we should randomly select another treatment combination and obtain a second observation, and so on, until all 36 observations have been taken. However, the experimenter did not collect the data this way. He made up a batch of pulp and obtained observations for all four temperatures from that batch. Because of the economics of preparing the batches and the size of the batches, this is the only feasible way to run this experiment. A completely randomized factorial experiment would require 36 batches of pulp, which is completely unrealistic. The split-plot design requires only 9 batches total. Obviously, the split-plot design has resulted in considerable experimental efficiency.

---

## Experiment on Tensile Strength of Paper

### Pulp Preparation

Method (Factor A) 

Temperature (°F)	Replicate 1			Replicate 2			Replicate 3		
	1	2	3	1	2	3	1	2	3
(Factor B)	200	30	34	29	28	31	31	31	35
	225	35	41	26	32	36	30	37	40
	250	37	38	33	40	42	32	41	39
	275	36	42	36	41	40	40	44	45

---

The design used in our example is a split-plot design. In this split-plot design we have 9 whole plots, and the preparation methods are called the whole plot or main treatments. Each whole plot is divided into four parts called subplots (or split-plots), and one temperature is assigned to each. Temperature is called the subplot treatment. Note that if other uncontrolled or undesigned factors are present and if these uncontrolled factors vary as the pulp preparation methods are changed, then any effect of the undesigned factors on the response will be completely compounded with the effect of the pulp preparation methods. Because the whole-plot treatments in a split-plot design are confounded with the whole-plots and the subplot treatments are not confounded, it is best to assign the factor we are most interested in to the subplots, if possible.

---

This example is fairly typical of how the split-plot design is used in an industrial setting. Notice that the two factors were essentially “applied” at different times. Consequently, a split-plot design can be viewed as two experiments “combined” or superimposed on each other. One “experiment” has the whole-plot factor applied to the large experimental units (or it is a factor whose levels are hard to change) and the other “experiment” has sub-plot factor applied to the smaller experimental units (or it is a factor whose levels are easy to change).

A good example of split-plot design is about a paper manufacturer who wants to analyze the effect of three pulp preparation methods and four cooking temperatures on the tensile strength of the paper. The experimenter wants to perform three

---

---

replicates of this experiment on three different days each consisting of 12 runs ( $3 \times 4$ ). The important issue here is the fact that making the pulp by any of the methods is cumbersome. Thus method is a “hard to change” factor. It would be economical to randomly select any of the preparation methods, make the blend and divide it into four samples and cook each of them with one of the four cooking temperatures. Then the second method is used to prepare the pulp and so on. As we can see, in order to achieve this economy in the process, there is a restriction on the randomization of the experimental run.

---

In field experiments, sometimes a factor has to be applied to a large experimental unit. This is true when the different methods of ploughing or irrigation are to be compared. And in such cases it is possible to introduce a second factor, which does not require large plots, with a small number of levels into the same experiment, at a little extra cost. This is done by splitting the plots (called whole plots) of the first factor into as many sub-plots as there are levels of the second factor.

A split-plot design with an RBD for the first set of treatments (called “the whole-plot treatments”) is obtained by allotting the whole-plot treatments at random to the whole plots of a block and then randomizing the second set of treatments (“called the sub-plot treatments”) to the sub-plots within each whole plot.

---

**Ex :** A variety-manurial experiment was conducted by allotting the three varieties  $V_1$ ,  $V_2$  and  $V_3$  at random to the plots of four randomized blocks and then, splitting each plot into four sub-plots, the four manures  $M_1$ ,  $M_2$ ,  $M_3$  and  $M_4$  were applied at random within each plot. The plan and yield are shown below. Analyse the data to find out if there are any effects due to manure or variety of interaction between variety and manure.

Variety	BLOCK			
	I	II	III	IV
$V_1$	609	450	488	545
$V_2$	920	870	833	1118
$V_3$	1067	1072	1093	905

$p = 3 = \text{number of varieties (A)}$

$q = 4 = \text{number of manures (B)}$

$r = 4 = \text{number of replicates or blocks}$

$N = pqr = 3 \times 4 \times 4 = 48$

		V <sub>1</sub>	V <sub>3</sub>	V <sub>2</sub>			V <sub>2</sub>	V <sub>1</sub>	V <sub>3</sub>
Block I	M <sub>1</sub>	94	M <sub>4</sub>	440	M <sub>2</sub>	250			
	M <sub>3</sub>	220	M <sub>2</sub>	297	M <sub>1</sub>	147			
	M <sub>2</sub>	185	M <sub>3</sub>	218	M <sub>3</sub>	248			
	M <sub>4</sub>	110	M <sub>1</sub>	112	M <sub>4</sub>	275			
Total		609	1067	920			870	450	1072
Block III	V <sub>1</sub>	V <sub>2</sub>	V <sub>3</sub>		V <sub>1</sub>	V <sub>3</sub>	V <sub>2</sub>		
	M <sub>1</sub>	78	M <sub>3</sub>	196	M <sub>2</sub>	235			
	M <sub>3</sub>	135	M <sub>4</sub>	262	M <sub>3</sub>	260			
	M <sub>4</sub>	130	M <sub>1</sub>	155	M <sub>1</sub>	115			
	M <sub>2</sub>	145	M <sub>2</sub>	220	M <sub>4</sub>	483			
Total		488	833	1093			545	905	1118
M <sub>1</sub> V <sub>1</sub> = 94 + 71 + 78 + 81 = 324; M <sub>1</sub> V <sub>3</sub> = 112 + 140 + 115 + 145 = 512					M <sub>2</sub> V <sub>3</sub> = 297 + 222 + 235 + 246 = 1000; M <sub>3</sub> V <sub>3</sub> = 218 + 340 + 260 + 191 = 1009				

Soln : We draw up the block-variety table for obtaining the whole-plot analysis

Variety	BLOCK				Total
	I	II	III	IV	
V <sub>1</sub>	609	450	488	545	2092
V <sub>2</sub>	920	870	833	1118	3741
V <sub>3</sub>	1067	1072	1093	905	4137
Total	2596	2392	2414	2568	9970

$$CF = G^2/N = 9970^2/12 = 9970^2/3 \times 4 \times 4 = 2,070,852.08333$$

$$\begin{aligned}
 \text{Block SS} &= (2596^2 + 2392^2 + 2414^2 + 2568^2)/12 - CF \\
 &= 24882900/12 - 2,070,852.0833 = 2,073,575 - 2,070,852.0833 \\
 &= 2,722.91667
 \end{aligned}$$

$$\begin{aligned}
 \text{Variety SS} &= (2092^2 + 3741^2 + 4137^2)/16 - CF \\
 &= 35486314/16 - 2,070,852.0833 = 2,217,894.625 - 2,070,852.0833 \\
 &= 147,042.5417
 \end{aligned}$$

Error (I) SS =  $(609^2 + 920^2 + \dots + 1118^2 + 905^2)/4 - 9970^2/48$  – Variety SS  
 - Block SS

$$8957010/4 - 2,070,852.08333 - 147,042.54167 - 2,722.91667 = 18,634.9583$$

Now draw up the Variety-Manure table to obtain the Manure SS and Interaction SS

Variety	V <sub>1</sub>	V <sub>2</sub>	V <sub>3</sub>	Total
Manure				
M <sub>1</sub>	324	559	512	1395
M <sub>2</sub>	665	900	1000	2565
M <sub>3</sub>	593	1005	1009	2607
M <sub>4</sub>	510	1277	1616	3403
Total	2092	3741	4137	9970

$$\begin{aligned} \text{Manure SS} &= (1395^2 + 2565^2 + 2607^2 + 3403^2)/12 - 9970^2/48 \\ &= 26902108/12 - 2,070,852.08333 = 170,990.25 \end{aligned}$$

---

$$\begin{aligned}\text{Variety x Manure SS} &= (324^2 + 665^2 + \dots + 1009^2 + 1616^2)/r - 9970^2/48 \\ &\quad - \text{Manure SS} - \text{Variety SS} \\ &= 9813866/4 - 2,388,884.875 = 64,581.625\end{aligned}$$

$$\begin{aligned}\text{Raw Total SS} &= \sum y^2 = \text{Sum of squares of 48 values from Variety-Manure} \\ &\quad \text{square} = 94^2 + 440^2 + \dots + 323^2 + 450^2 = 2,500,068\end{aligned}$$

$$\text{Total SS} = \text{RSS} - \text{CF} = 2,500,068 - 2,070,852.0833 = 429,215.91667$$

$$\begin{aligned}\text{Error (II) SS} &= \text{Total SS} - \text{Variety SS} - \text{Var x Man SS} - \text{Man SS} - \text{Block SS} \\ &= 25,243.62493 \text{ (By subtraction)}\end{aligned}$$

### ANOVA of the Split-Plot Design

Source of variation	df	SS	MS = SS/df	Fcalculated
Blocks	$r - 1 = 3$	2,722.92	907.6388 (1)	(1)/(3)
Varieties	$p - 1 = 2$	147,042.54	73521.27 (2)	(2)/(3)
Error (I)	$(r - 1)(p - 1) = 6$	18,634.96	3105.83 (3)	
Manures	$q - 1 = 3$	170,990.25	56996.75 (4)	(4)/(6)
Variety x Manure	$(p - 1)(q - 1) = 6$	64,581.63	10763.6 (5)	(5)/(6)
Error (II)	27	25,243.62	934.95 (6)	
Total	$N - 1 = 47$	429,215.92		

Since  $F$  for interaction  $10763.6/934.95 = 11.512 > F_{critical}$  (appx 3.56 at LOS 1%), hypothesis of no interaction effects is rejected. As such, we do not perform the test for main effects of the Factors and hence corresponding  $F$ 's are not shown in the table.

$$CF = G^2/N = 9970^2/pqr = 9970^2/3 \times 4 \times 4 = 2,070,852.08333$$

$$\begin{aligned} \text{Block SS} &= (2596^2 + 2392^2 + 2414^2 + 2568^2)/pq - CF \\ &= 24882900/12 - 2,070,852.0833 = 2,073,575 - 2,070,852.0833 \\ &= 2,722.91667 \end{aligned}$$

$$\begin{aligned} \text{Variety SS} &= (2092^2 + 3741^2 + 4137^2)/rq - CF \\ &= 35486314/16 - 2,070,852.0833 = 2,217,894.625 - 2,070,852.0833 \\ &= 147,042.5417 \end{aligned}$$

$$\begin{aligned} \text{Manure SS} &= (1395^2 + 2565^2 + 2607^2 + 3403^2)/rp - 9970^2/48 \\ &= 26902108/12 - 2,070,852.08333 = 170,990.25 \end{aligned}$$

$$\begin{aligned} \text{Variety x Manure SS} &= (324^2 + 665^2 + \dots + 1009^2 + 1616^2)/r - 9970^2/48 \\ &\quad - \text{Manure SS} - \text{Variety SS} \\ &= 9813866/4 - 2,388,884.875 = 64,581.625 \end{aligned}$$

$$\begin{aligned} \text{Raw Total SS} &= \sum y^2 = \text{Sum of squares of 48 values from Variety-Manure square} \\ &= 94^2 + 440^2 + \dots + 323^2 + 450^2 = 2,500,068 \end{aligned}$$

$$\text{Total SS} = \text{RSS} - CF = 2,500,068 - 2,070,852.0833 = 429,215.91667$$

$$\begin{aligned} \text{Error (II) SS} &= \text{Total SS} - \text{Variety SS} - \text{Var x Man SS} - \text{Man SS} - \text{Block SS} \\ &= 25,243.62493 \text{ (By subtraction)} \end{aligned}$$

## Advantages and Disadvantages

The split-plot design has two errors, of which  $E_{II}$  is smaller than  $E_I$ . Hence usually, the B and AB effects will be estimated and tested more precisely than the A effects. The main advantage of the design is that often it is possible to introduce the second factor B, requiring small experimental material, along with A in a split-plot arrangement at little extra cost. If we have a choice for the allocation of factor A and factor B to the whole plots and split-plots, we shall apply the factor which is more important to the split-plot.

The disadvantages of this design are that the presence of two errors makes the analysis difficult and sometimes the error  $E_I$  may be too large.

Although the experimental error for sub-plot treatments and interaction is smaller than that for whole-plot treatment, it can be shown that the

---

average experimental error over all treatment comparisons is the same for a split-plot design and the corresponding factorial experiment in an RBD.

In summary, when one of the treatment factors needs more replication or experimental units (material) than another or when it is hard to change the level of one of the factors, these design become important. The primary disadvantage of these designs is the loss in precision in the whole plot treatment comparison and the statistical complexity.



# BSDCHZC355

## Statistical Inference & Applications

**BITS** Pilani  
Pilani | Dubai | Goa | Hyderabad

Shaibal Kr. Sen  
Session 13, 14 &15



**STATISTICAL INFERENCES**

**&**

**APPLICATIONS**

## Scheduled Topics to be covered during Session 13, 14 & 15

Statistical Quality Control -

Importance of SQC in industry.

Statistical basis of Shewhart control charts.

Construction of control charts for variables (mean, range and standard deviation) and attributes ( p , np, & c- charts with fixed and varying sample sizes).

Interpretation of control charts.

Natural tolerance limits and specification limits,

Process Capability Index.

Concept of Six Sigma and its importance.

---

What are the determinants for purchasing a product (or availing a service) by a customer?

Necessary Requirement of the product or service for the customer –

Acceptable Quality of the product or service.

Affordable Price of product or service.

---

**Importance of Quality** : Controlling and improving quality has become an important business strategy for many organizations; Manufacturers, Distributors, Transportation Companies, Financial Services organizations; Health care providers and Government agencies. Quality is a competitive advantage. A business that can delight customers by improving and controlling quality can dominate its competitors.

---

Quality : We may define quality in many ways. Most people have a conceptual understanding of quality as relating to one or more desirable characteristics that a product or service should possess. Although this conceptual understanding is certainly a useful starting point, we will give a more precise and useful definition.

Definition : Quality means fitness for use.

---

Quality characteristics : Every product possesses a number of elements that jointly describe what the user or consumer thinks about quality. These parameters are often called quality characteristics. Sometimes these are called Critical-to-Quality (CTQ) characteristics. Quality characteristics may be of several types –

1. Physical : Length, Weight, Voltage, Viscosity.
2. Sensory : Taste, Appearance, Colour
3. Time Orientation : Reliability, Durability, Serviceability

---

## Dimensions of Quality :

- (1) Performance – Will the product do the intended job?
  - (2) Reliability – How often does the product fail?
  - (3) Durability – How long does the product last?
  - (4) Serviceability – How easy is it to repair the product?
  - (5) Aesthetics – What does the product look like?
  - (6) Features – What does the product do?
  - (7) Perceived Quality – What is the reputation of the company or its product?
  - (8) Conformance to Standards – Is the product made exactly as the designer intended?
-

---

Definition : Quality is inversely proportional to variability. This definition implies that if the variability in the important characteristics of product decreases, the quality of product increases. (Note : Here variability is referred as unwanted or harmful variability).

Quality Engineering : is the set of Operational, Managerial & Engineering activities that a company uses to ensure that the quality characteristics of the product are at nominal or required levels and that the variability around these desired levels is minimum.

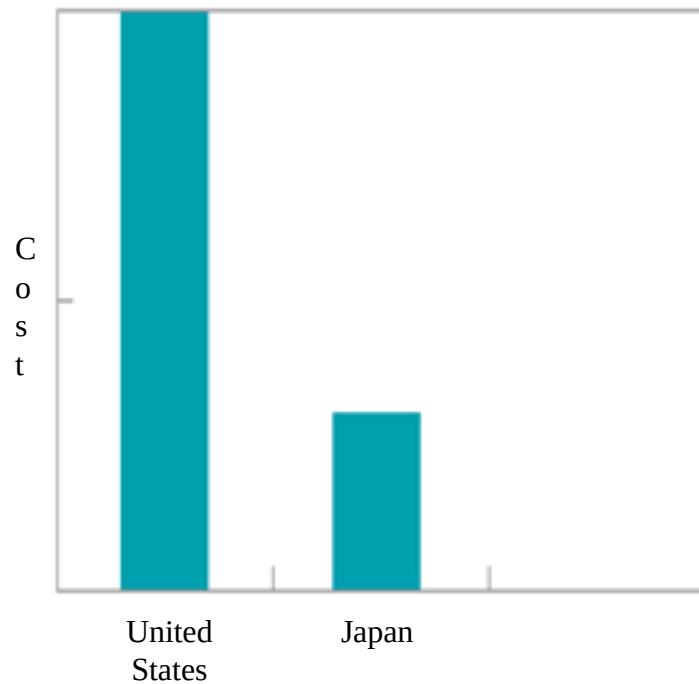
Quality improvement is the reduction of variability in processes and products.

---

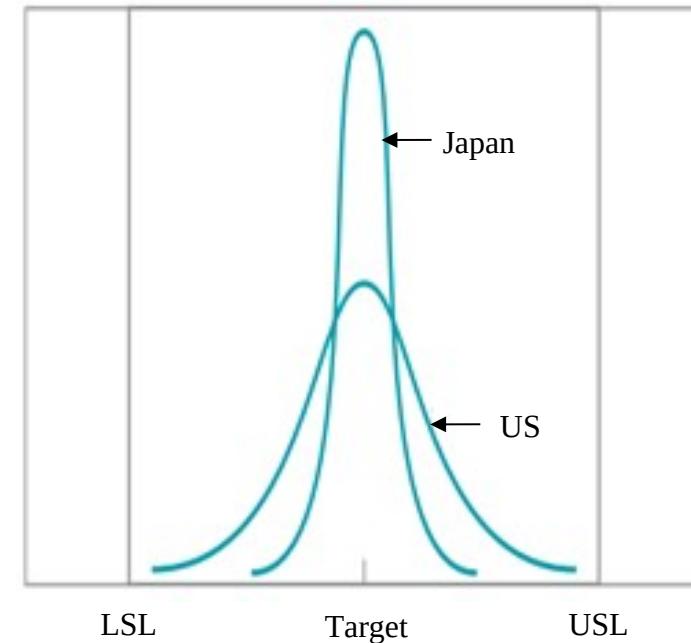
---

Most organizations find it difficult and expensive to provide the customer with products that have quality characteristics that are always identical from unit to unit or are at levels that match customer expectations. A major reason for this is variability. There is certain amount of variability in every product; consequently, no two products are ever identical. For example, the thickness of the blades on a jet turbine engine impeller is not identical even on the same impeller. Blade thickness will also differ between impellers. If this variation in blade thickness is small, then it may have no impact on the customer. However, if the variation is large, then the customer may perceive the unit to be undesirable and unacceptable. Sources of this variability include differences in materials, differences in the performance and operation of the manufacturing equipment and differences in the way the operators perform their tasks. This line of thinking led to the previous definition of quality improvement.

---



Warranty Costs for  
Transmissions



Distributions of Critical  
Dimensions for Transmissions

---

Since Variability can only be described in statistical terms, Statistical Methods play a central role in Quality improvement efforts. In the application of statistical methods to quality engg., it is fairly typical to classify data on quality characteristics as either attributes or variables data. Variables data are usually Continuous Measurements, such as length, voltage or viscosity. Attributes data, on the other hand, are usually Discrete Data, often taking the form of counts. Such as the loan applications that could not be properly processed because of missing required information, or the number of emergency room arrivals that have to wait more than 30 minutes to receive medical attention. We will describe statistical-based quality engg. tools for dealing with both types of data.

---

Statistical methods and their application in quality improvement have had a long history. In 1924, Walter A. Shewhart of the Bell Telephone Laboratories developed the Statistical Control Chart concept, which is often considered the formal beginning of Statistical Quality Control (SQC). Towards the end of 1920s, Harold F. Dodge and Harry G. Romig, both of Bell Telephone Laboratories, developed statistically based acceptance sampling as an alternative to 100% inspection. By the middle of the 1930s, statistical quality control methods were in wide use at Western Electric, the manufacturing arm of the Bell System.

A value of measurement that corresponds to the desired value for the quality characteristic is called the ‘Nominal’ or ‘Target Value’ for that characteristic. These target values are usually

---

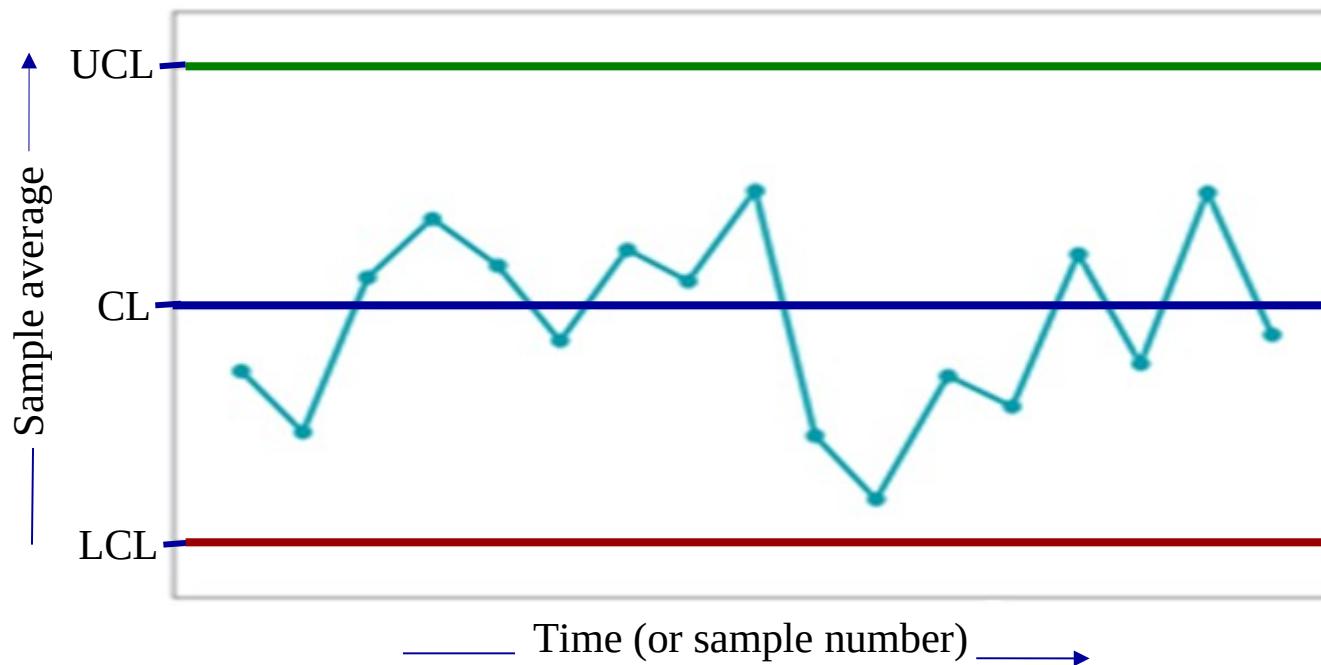
bounded by a range of values that, most typically, we believe will be sufficiently close to the target so as to not impact the function or performance of the product if the quality characteristic is in that range. The largest allowable value for a quality characteristic is called the Upper Specification Limit (USL) and the smallest allowable value for a quality characteristic is called the Lower Specification Limit (LSL). Some quality characteristics have specification limits on only one side of the target. For example, the compressive strength of a component used in an automobile bumper likely has a target value and a lower specification limit, but not an upper specification limit.

---

A Control Chart is one of the primary techniques of Statistical Process Control (SPC). A typical control chart is shown here. This chart plots the averages of measurements of quality characteristic in samples taken from process versus time (or the sample number). The chart has a central line (CL) and upper & lower control limits (UCL & LCL) as shown in the figure. The central line shows where this process characteristic should fall if there are no unusual sources of variability present. The control limits are determined from some simple charts are applied to the output variable (s) in a system as depicted in the figure. However, in some cases they can be usefully applied to the inputs as well.

The control chart is a very useful process monitoring technique; when unusual sources of variability are present, sample averages will plot

outside the control limits. This is a signal that some investigation of the process should be made and corrective action to remove these unusual sources of variability taken. Systematic use of a control chart is an excellent way to reduce variability.

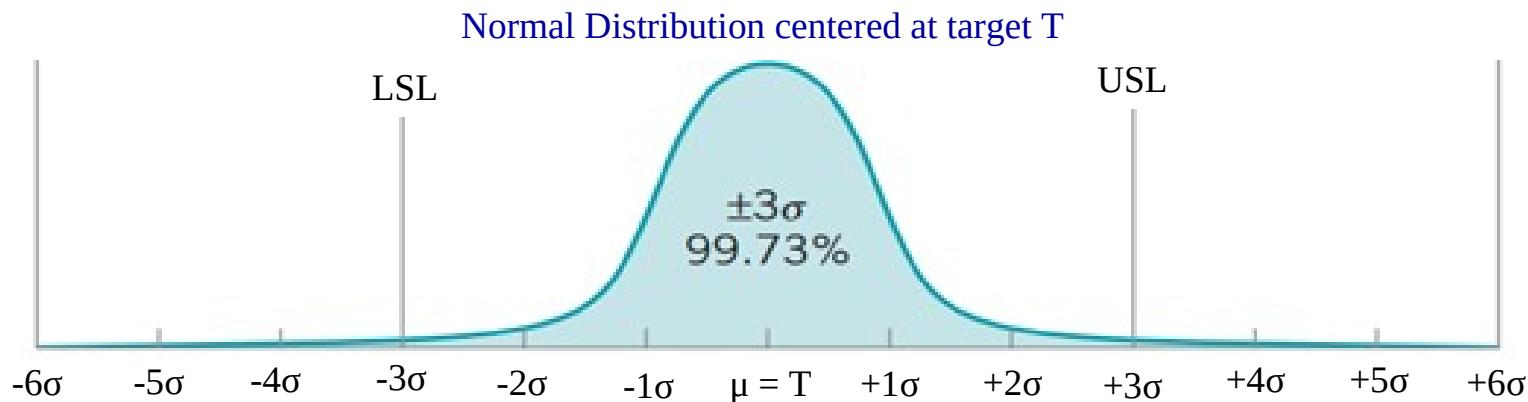


## Management Aspects of Quality Improvement

Statistical techniques, including Statistical Process Control (SPC) and Designed Experiments, along with other problem-solving tools are the technical basis for Quality Control and Improvement. However, to be used most effectively, these techniques must be implemented within and be part of a management system that is focused on quality improvement. The management system of an organization must be organized to properly direct the overall quality improvement philosophy and ensure its deployment in all aspects of the business. The effective management of quality involves successful execution of three activities : Quality Planning, Quality Assurance and Quality Control & Improvement.

# Six-Sigma

Products with many components typically have many possibilities for failure or defects to occur. Motorola developed the Six-Sigma Programme in the year 1980 as a response to the demand for their products. The focus of Six-Sigma is reducing variability in key product quality characteristics to the level at which failure or defects are extremely unlikely.



Spec. Limit	% inside Spec	% Defective	PPM Defective
$\pm 1\sigma$	68.27	31.73	317300
$\pm 2\sigma$	95.45	4.55	45500
$\pm 3\sigma$	99.73	0.27	2700
$\pm 4\sigma$	99.9937	0.0063	63
$\pm 5\sigma$	99.999943	0.000057	0.57
$\pm 6\sigma$	99.999998	0.000002	0.002

Figure in slide shows a normal probability distribution as a model for a quality characteristic with the specification limits at three standard deviations on either side of the mean. Now it turns out that in this situation the probability of producing a product within the specification is 0.9973 (or 99.73%), which corresponds to 2700 parts per million (PPM) defective. This is referred to as three-sigma quality performance, and it actually sounds pretty good. However, suppose we have a product that consists of an assembly of 100 independent components or parts and all 100 of these parts must be non-defective for the product to function satisfactorily. The probability that any specific unit of product is non-defective is -

$$0.9973 \times 0.9973 \times \dots \times 0.9973 = (0.9973)^{100} = 0.7631$$

That is about 23.69% (ie. 100 – 76.31) of the products produced under

---

this is not acceptable situation, because many products used by to-day's society are made up of many components. Even a relatively simple service activity, such as a visit by a family of four to a fast-food restaurant, can involve the assembly of several dozen components. A typical automobile has about 100,000 components and an airplane has between one and two million!

The Motorola six-sigma concept is to reduce the variability in the process so that the specification limits are at least six standard deviation from the mean. Then, as shown in the figure, there will only be about 2 parts per billion defective. Under Six-Sigma Quality, the probability that any specific unit of the hypothetical product above is non-defective is 0.9999998 or 0.002 PPM defective, a much better situation.

## Shewhart's Control Charts

Chance & Assignable causes of Quality Variation : In any production process, regardless of how well designed or carefully maintained it is, a certain amount of inherent or natural variability will always exist. This natural variability or “background noise” is the cumulative effect of many small, essentially unavoidable causes. In the framework of SQC, this natural variability is often called a “stable system of chance causes”. A process that is operating with only Chance Causes of Variation present is said to be in Statistical Control. In other words, the chance causes are inherent part of the process.

Other kinds of variability may occasionally be present in the output of process. This variability in key quality characteristics usually arises from three sources – improperly adjusted or uncontrolled machines, operator errors or defective raw material. Such variability is generally large when compared to the background noise and it usually represents

## Shewhart's Control Charts

an unacceptable level of process performance. We refer to these sources of variability that are not part of the chance cause pattern as Assignable Causes of Variation. A process that is operating in the presence of assignable causes is said to be an Out-of-Control process.

Popularly there are seven SPC tools – (i) Histogram, (ii) Check sheet, (iii) Pareto chart, (iv) Cause-and-effect diagram, (v) Stratification (alternately, flow chart), (vi) Scatter diagram and (vii) Control charts.

Shewhart control chart is probably the most technically sophisticated. It was developed in the 1920s by Walter A. Shewhart of the Bell Telephone Laboratories. To understand the statistical concepts that form the basis of SPC, we must first describe Shewhart's theory of variability – (i) Chance and (ii) Assignable Causes.

---

Chance and Assignable Causes of Quality Variation : In any production process, regardless of how well designed or carefully maintained it is, a certain amount of inherent or natural variability will always exist. This natural variability or “background noise” will always exist. This natural variability or “background noise” is the cumulative effect of many small, essentially unavoidable causes. In the framework of statistical quality control, this natural variability is often called a “stable system of chance causes.” A process that is operating with only chance causes of variation present is said to be in statistical control. In other words, the chance causes are an inherent part of the process.

Other kinds of variability, Assignable causes of variation, may occasionally be present in the output of a process. This variability in key quality characteristics usually arises from three sources :

(i) Improperly adjusted or uncontrolled machines, (ii) operator errors,

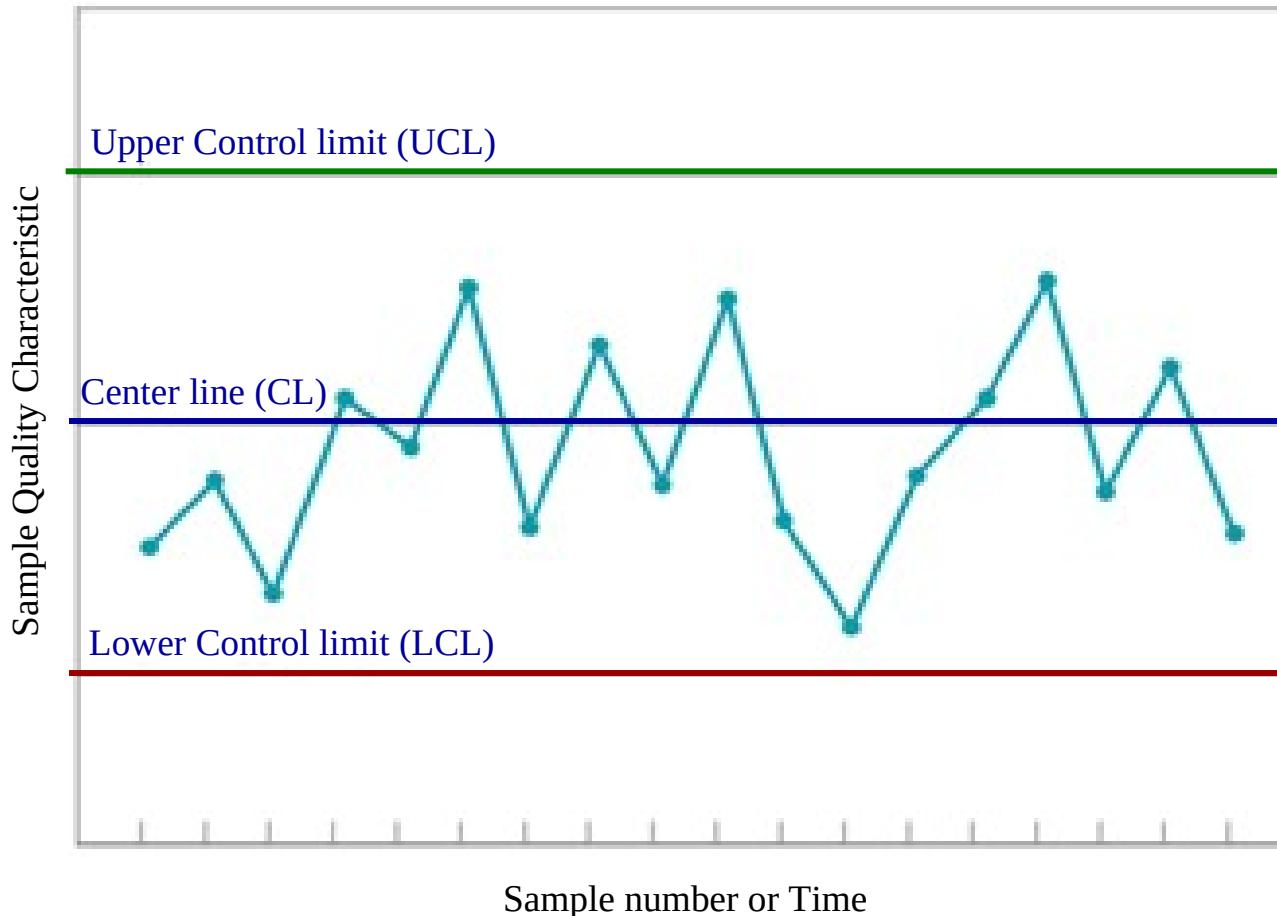
---

or (iii) defective raw material. Such variability is generally large when compared to the background noise and it usually represents an unacceptable level of process performance. A process that is operating in the presence of assignable causes is said to be an out-of-control process.

## Statistical Basis of the Control Chart

A typical control chart is shown here. The control chart is a graphical display of quality characteristic that has been measured or computed from a sample versus the sample number or time. The chart contains a Center Line (CL) that represents the average value of the quality characteristic corresponding to the in-control state (that is, only chance causes are present). Two other horizontal lines – called the Upper Control Limit (UCL) and the Lower Control Limit (LCL), are also shown on the chart. These control limits are chosen so that if the process is in control, nearly all the sample points will fall between them. As long as the points plot within the control limits, the process is assumed to be in control and no action is necessary. However, a point that plots outside of the control limits is interpreted as evidence that the process is out of control and investigation and corrective action are required to find and eliminate the assignable cause or causes responsible for this behaviour. It is customary to connect the sample points on the control chart.

### A typical Control Chart



We may give a general model for a control chart. Let  $w$  be a sample statistic that measures quality characteristic of interest and suppose that the mean of  $w$  is  $\mu_w$  and the standard deviation is  $\sigma_w$ . Then the center line, Upper Control Limit (UCL) and Lower Control Limit (LCL) become –

$$UCL = \mu_w + L\sigma_w = \mu_w + 3\sigma_w$$

$$\text{Center line} = \mu_w$$

$$LCL = \mu_w - L\sigma_w = \mu_w - 3\sigma_w$$

where  $L$  is the distance of the control limits from the center line, expressed in standard deviation units. This general theory of control charts was first proposed by Walter A. Shewhart and control charts developed according to these principles are often called Shewhart control charts.

$$\begin{aligned} \text{UCL/LCL} &= \mu \pm 3\sigma \\ \text{UWL/LWL} &= \mu \pm 2\sigma \end{aligned}$$

UCL - Upper Control Limit

LCL - Lower Control Limit

UWL - Upper Warning Limit

LWL - Lower Warning Limit

■ APPENDIX VI

Factors for Constructing Variables Control Charts

Observations in	Chart for Averages					Chart for Standard Deviations				Chart for Ranges						
	Factors for Control Limits			Factors for Center Line		Factors for Control Limits				Factors for Center Line		Factors for Control Limits				
	Sample, $n$	$A$	$A_2$	$A_3$	$c_4$	$1/c_4$	$B_3$	$B_4$	$B_5$	$B_6$	$d_2$	$1/d_2$	$d_3$	$D_1$	$D_2$	$D_3$
2	2.121	1.880	2.659	0.7979	1.2533	0	3.267	0	2.606	1.128	0.8865	0.853	0	3.686	0	3.267
3	1.732	1.023	1.954	0.8862	1.1284	0	2.568	0	2.276	1.693	0.5907	0.888	0	4.358	0	2.574
4	1.500	0.729	1.628	0.9213	1.0854	0	2.266	0	2.088	2.059	0.4857	0.880	0	4.698	0	2.282
5	1.342	0.577	1.427	0.9400	1.0638	0	2.089	0	1.964	2.326	0.4299	0.864	0	4.918	0	2.114
6	1.225	0.483	1.287	0.9515	1.0510	0.030	1.970	0.029	1.874	2.534	0.3946	0.848	0	5.078	0	2.004
7	1.134	0.419	1.182	0.9594	1.0423	0.118	1.882	0.113	1.806	2.704	0.3698	0.833	0.204	5.204	0.076	1.924
8	1.061	0.373	1.099	0.9650	1.0363	0.185	1.815	0.179	1.751	2.847	0.3512	0.820	0.388	5.306	0.136	1.864
9	1.000	0.337	1.032	0.9693	1.0317	0.239	1.761	0.232	1.707	2.970	0.3367	0.808	0.547	5.393	0.184	1.816
10	0.949	0.308	0.975	0.9727	1.0281	0.284	1.716	0.276	1.669	3.078	0.3249	0.797	0.687	5.469	0.223	1.777
11	0.905	0.285	0.927	0.9754	1.0252	0.321	1.679	0.313	1.637	3.173	0.3152	0.787	0.811	5.535	0.256	1.744
12	0.866	0.266	0.886	0.9776	1.0229	0.354	1.646	0.346	1.610	3.258	0.3069	0.778	0.922	5.594	0.283	1.717
13	0.832	0.249	0.850	0.9794	1.0210	0.382	1.618	0.374	1.585	3.336	0.2998	0.770	1.025	5.647	0.307	1.693
14	0.802	0.235	0.817	0.9810	1.0194	0.406	1.594	0.399	1.563	3.407	0.2935	0.763	1.118	5.696	0.328	1.672
15	0.775	0.223	0.789	0.9823	1.0180	0.428	1.572	0.421	1.544	3.472	0.2880	0.756	1.203	5.741	0.347	1.653
16	0.750	0.212	0.763	0.9835	1.0168	0.448	1.552	0.440	1.526	3.532	0.2831	0.750	1.282	5.782	0.363	1.637
17	0.728	0.203	0.739	0.9845	1.0157	0.466	1.534	0.458	1.511	3.588	0.2787	0.744	1.356	5.820	0.378	1.622
18	0.707	0.194	0.718	0.9854	1.0148	0.482	1.518	0.475	1.496	3.640	0.2747	0.739	1.424	5.856	0.391	1.608
19	0.688	0.187	0.698	0.9862	1.0140	0.497	1.503	0.490	1.483	3.689	0.2711	0.734	1.487	5.891	0.403	1.597
20	0.671	0.180	0.680	0.9869	1.0133	0.510	1.490	0.504	1.470	3.735	0.2677	0.729	1.549	5.921	0.415	1.585
21	0.655	0.173	0.663	0.9876	1.0126	0.523	1.477	0.516	1.459	3.778	0.2647	0.724	1.605	5.951	0.425	1.575
22	0.640	0.167	0.647	0.9882	1.0119	0.534	1.466	0.528	1.448	3.819	0.2618	0.720	1.659	5.979	0.434	1.566
23	0.626	0.162	0.633	0.9887	1.0114	0.545	1.455	0.539	1.438	3.858	0.2592	0.716	1.710	6.006	0.443	1.557
24	0.612	0.157	0.619	0.9892	1.0109	0.555	1.445	0.549	1.429	3.895	0.2567	0.712	1.759	6.031	0.451	1.548
25	0.600	0.153	0.606	0.9896	1.0105	0.565	1.435	0.559	1.420	3.931	0.2544	0.708	1.806	6.056	0.459	1.541

---

## Types of Control Charts –

- I. Charts for Variables – a) Average chart (x-bar chart)  
b) Range chart (R chart)
- II. Chart for Attributes – a) Charts for defective items –
  - i) Fraction defective (Varied sample size) p chart
  - ii) No. of defective (Constant sample size) np chart  
b) Charts for defects per unit –
  - i) Varied sample size u chart
  - ii) Constant sample size c chart

---

There are two types of control charts that we deal with.

Variables Control Charts : These charts are applied to data that follow a continuous distribution. Variables include length, thickness, diameter, breaking strength, temperature, viscosity etc. It is used to monitor characteristics that can be measured and have a continuum of values, such as height, weight, or volume.

Attributes Control Charts : These charts are applied to data that follow a discrete distribution. Attributes deal with qualitative information, eg.  $40 \pm 0.6$ . Another example of attributes data is the count of defects. A control chart for attributes is used to monitor characteristics that have discrete values and can be counted.

Classifications such as conforming and nonconforming are commonly used in quality control.

---

Control Chart of Variables

Mean Chart

Range Chart

Standard deviation Chart

## Control Charts for $\bar{x}$ and R

Statistical Basis of the Charts : Suppose that a quality characteristic is normally distributed with mean  $\mu$  and standard deviation  $\sigma$ , where both  $\mu$  and  $\sigma$  are known. If  $x_1, x_2, \dots, x_n$  is a sample of size  $n$ , then the average of this sample is

$$\bar{x} = (x_1 + x_2 + \dots, x_n)/n = \sum x_i/n \quad (\text{limit of } i \text{ is from 1 to } n)$$

and we know that  $\bar{x}$  is normally distributed with mean  $\mu$  and standard deviation  $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ . Furthermore, the probability is  $1 - \alpha$  that any mean will fall between

—

$$\mu + Z_{\alpha/2}\sigma_{\bar{x}} = \mu + Z_{\alpha/2}(\sigma/\sqrt{n}) \quad \text{and} \quad \mu - Z_{\alpha/2}\sigma_{\bar{x}} = \mu - Z_{\alpha/2}(\sigma/\sqrt{n}) \quad \dots \dots \text{(i)}$$

Therefore, if  $\mu$  and  $\sigma$  are known, equation (i) could be used as Upper & Lower Control Limits on a control chart for a sample means. It is customary to replace  $Z_{\alpha/2}$  by 3, so that three sigma limits are employed. If a sample mean falls outside of these limits, it is an indication that the process mean is no longer equal to  $\mu$ .

In practice, we usually will not know  $\mu$  and  $\sigma$ . Therefore, they must be estimated from preliminary samples or subgroups taken when the process is thought to be in control. These estimates should usually be based on at least 20 to 25 samples. Suppose that  $m$  samples are available, each containing  $n$  observations on the quality characteristic. Typically  $n$  will be small, often either 4, 5 or 6. These small sample sizes usually result from the construction of rational subgroups and from the fact that the sampling and inspection costs associated with variables measurements are usually relatively large. Let  $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_m$  be the average of each sample. Then the best estimator of  $\mu$ , the process average, is the grand average - say,

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \dots + \bar{x}_m}{m}$$

=

Thus,  $\bar{x}$  would be used as the center line on the  $\bar{x}$  chart.

To construct the control limits, we need an estimate of the standard deviation  $\sigma$ .

We may estimate  $\sigma$  from either the standard deviations or the ranges of  $m$  samples. For the present, we will use the range method. If  $x_1, x_2, \dots, x_n$  is a sample of size  $n$ , then the range of the sample is the difference between the largest and smallest observation; that is,

$$R = x_{\max} - x_{\min}$$

Let  $R_1, R_2, \dots, R_m$  be the ranges of  $m$  samples. The average range is

$$R = (R_1 + R_2 + \dots + R_m)/m$$

We may now give the formulas for constructing the control limits on the  $\bar{x}$  chart. They are as follows –

**Control Limits for the  $\bar{x}$  Chart**

$$\text{UCL} = \bar{x} + A_2 \bar{R}$$

$$\text{Center line} = \bar{x}$$

$$\text{LCL} = \bar{x} - A_2 \bar{R}$$

The constant  $A_2$  is tabulated for various sample sizes in Appendix Table VI.

Process variability may be monitored by plotting values of the sample range  $R$  on a control chart. The center line and control limits of the  $R$  chart are as follows -

**Control Limits for the  $R$  Chart**

$$\text{UCL} = D_4 \bar{R}$$

$$\text{Center line} = \bar{R}$$

$$\text{LCL} = D_3 \bar{R}$$

The constant  $D_3$  &  $D_4$  are tabulated for various values of  $n$  in Appendix Table VI.

## APPENDIX VI

### Factors for Constructing Variables Control Charts

Observations in	Chart for Averages					Chart for Standard Deviations				Chart for Ranges						
	Factors for Control Limits			Factors for Center Line		Factors for Control Limits				Factors for Center Line		Factors for Control Limits				
	Sample, $n$	$A$	$A_2$	$A_3$	$c_4$	$1/c_4$	$B_3$	$B_4$	$B_5$	$B_6$	$d_2$	$1/d_2$	$d_3$	$D_1$	$D_2$	$D_3$
2	2.121	1.880	2.659	0.7979	1.2533	0	3.267	0	2.606	1.128	0.8865	0.853	0	3.686	0	3.267
3	1.732	1.023	1.954	0.8862	1.1284	0	2.568	0	2.276	1.693	0.5907	0.888	0	4.358	0	2.574
4	1.500	0.729	1.628	0.9213	1.0854	0	2.266	0	2.088	2.059	0.4857	0.880	0	4.698	0	2.282
5	1.342	0.577	1.427	0.9400	1.0638	0	2.089	0	1.964	2.326	0.4299	0.864	0	4.918	0	2.114
6	1.225	0.483	1.287	0.9515	1.0510	0.030	1.970	0.029	1.874	2.534	0.3946	0.848	0	5.078	0	2.004
7	1.134	0.419	1.182	0.9594	1.0423	0.118	1.882	0.113	1.806	2.704	0.3698	0.833	0.204	5.204	0.076	1.924
8	1.061	0.373	1.099	0.9650	1.0363	0.185	1.815	0.179	1.751	2.847	0.3512	0.820	0.388	5.306	0.136	1.864
9	1.000	0.337	1.032	0.9693	1.0317	0.239	1.761	0.232	1.707	2.970	0.3367	0.808	0.547	5.393	0.184	1.816
10	0.949	0.308	0.975	0.9727	1.0281	0.284	1.716	0.276	1.669	3.078	0.3249	0.797	0.687	5.469	0.223	1.777
11	0.905	0.285	0.927	0.9754	1.0252	0.321	1.679	0.313	1.637	3.173	0.3152	0.787	0.811	5.535	0.256	1.744
12	0.866	0.266	0.886	0.9776	1.0229	0.354	1.646	0.346	1.610	3.258	0.3069	0.778	0.922	5.594	0.283	1.717
13	0.832	0.249	0.850	0.9794	1.0210	0.382	1.618	0.374	1.585	3.336	0.2998	0.770	1.025	5.647	0.307	1.693
14	0.802	0.235	0.817	0.9810	1.0194	0.406	1.594	0.399	1.563	3.407	0.2935	0.763	1.118	5.696	0.328	1.672
15	0.775	0.223	0.789	0.9823	1.0180	0.428	1.572	0.421	1.544	3.472	0.2880	0.756	1.203	5.741	0.347	1.653
16	0.750	0.212	0.763	0.9835	1.0168	0.448	1.552	0.440	1.526	3.532	0.2831	0.750	1.282	5.782	0.363	1.637
17	0.728	0.203	0.739	0.9845	1.0157	0.466	1.534	0.458	1.511	3.588	0.2787	0.744	1.356	5.820	0.378	1.622
18	0.707	0.194	0.718	0.9854	1.0148	0.482	1.518	0.475	1.496	3.640	0.2747	0.739	1.424	5.856	0.391	1.608
19	0.688	0.187	0.698	0.9862	1.0140	0.497	1.503	0.490	1.483	3.689	0.2711	0.734	1.487	5.891	0.403	1.597
20	0.671	0.180	0.680	0.9869	1.0133	0.510	1.490	0.504	1.470	3.735	0.2677	0.729	1.549	5.921	0.415	1.585
21	0.655	0.173	0.663	0.9876	1.0126	0.523	1.477	0.516	1.459	3.778	0.2647	0.724	1.605	5.951	0.425	1.575
22	0.640	0.167	0.647	0.9882	1.0119	0.534	1.466	0.528	1.448	3.819	0.2618	0.720	1.659	5.979	0.434	1.566
23	0.626	0.162	0.633	0.9887	1.0114	0.545	1.455	0.539	1.438	3.858	0.2592	0.716	1.710	6.006	0.443	1.557
24	0.612	0.157	0.619	0.9892	1.0109	0.555	1.445	0.549	1.429	3.895	0.2567	0.712	1.759	6.031	0.451	1.548
25	0.600	0.153	0.606	0.9896	1.0105	0.565	1.435	0.559	1.420	3.931	0.2544	0.708	1.806	6.056	0.459	1.541

## Ex : $\bar{x}$ and R chart for a manufacturing process

A hard-bake process is used in conjunction with photolithography in semiconductor manufacturing. We wish to establish statistical control of the flow width of the resist in this process using  $\bar{x}$  and R charts. Twenty five samples, each of size five wafers, have been taken when we think the process is in control. The interval of time between samples or subgroups is one hour. The flow width measurement data (in microns) from these samples is shown in the table.

Flow width measurements (microns) for Hard-Bake Process

Sample Number	Wafers				
	1	2	3	4	5
1	1.3235	1.4128	1.6744	1.4573	1.6914
2	1.4314	1.3592	1.6075	1.4666	1.6109
3	1.4284	1.4871	1.4932	1.4324	1.5674
4	1.5028	1.6352	1.3841	1.2831	1.5507
5	1.5604	1.2735	1.5265	1.4363	1.6441
6	1.5955	1.5451	1.3574	1.3281	1.4198
7	1.6274	1.5064	1.8366	1.4177	1.5144
8	1.4190	1.4303	1.6637	1.6067	1.5519
9	1.3884	1.7277	1.5355	1.5176	1.3688
10	1.4039	1.6697	1.5089	1.4627	1.5220
11	1.4158	1.7667	1.4278	1.5928	1.4181
12	1.5821	1.3355	1.5777	1.3908	1.7559
13	1.2856	1.4106	1.4447	1.6398	1.1928
14	1.4951	1.4036	1.5893	1.6458	1.4969
15	1.3589	1.2863	1.5996	1.2497	1.5471
16	1.5747	1.5301	1.5171	1.1839	1.8662
17	1.3680	1.7269	1.3957	1.5014	1.4449
18	1.4163	1.3864	1.3057	1.6210	1.5573
19	1.5796	1.4185	1.6541	1.5116	1.7247
20	1.7106	1.4412	1.2361	1.3820	1.7601
21	1.4371	1.5051	1.3485	1.5670	1.4880
22	1.4738	1.5936	1.6583	1.4973	1.4720
23	1.5917	1.4333	1.5551	1.5295	1.6866
24	1.6399	1.5243	1.5705	1.5563	1.5530
25	1.5797	1.3663	1.6240	1.3732	1.6887



Sample Number	Wafers					$\bar{x}_i$	$R_i$
	1	2	3	4	5		
1	1.3235	1.4128	1.6744	1.4573	1.6914	1.5119	0.3679
2	1.4314	1.3592	1.6075	1.4666	1.6109	1.4951	0.2517
3	1.4284	1.4871	1.4932	1.4324	1.5674	1.4817	0.1390
4	1.5028	1.6352	1.3841	1.2831	1.5507	1.4712	0.3521
5	1.5604	1.2735	1.5265	1.4363	1.6441	1.4882	0.3706
6	1.5955	1.5451	1.3574	1.3281	1.4198	1.4492	0.2674
7	1.6274	1.5064	1.8366	1.4177	1.5144	1.5805	0.4189
8	1.4190	1.4303	1.6637	1.6067	1.5519	1.5343	0.2447
9	1.3884	1.7277	1.5355	1.5176	1.3688	1.5076	0.3589
10	1.4039	1.6697	1.5089	1.4627	1.5220	1.5134	0.2658
11	1.4158	1.7667	1.4278	1.5928	1.4181	1.5242	0.3509
12	1.5821	1.3355	1.5777	1.3908	1.7559	1.5284	0.4204
13	1.2856	1.4106	1.4447	1.6398	1.1928	1.3947	0.4470
14	1.4951	1.4036	1.5893	1.6458	1.4969	1.5261	0.2422
15	1.3589	1.2863	1.5996	1.2497	1.5471	1.4083	0.3499
16	1.5747	1.5301	1.5171	1.1839	1.8662	1.5344	0.6823
17	1.3680	1.7269	1.3957	1.5014	1.4449	1.4874	0.3589
18	1.4163	1.3864	1.3057	1.6210	1.5573	1.4573	0.3153
19	1.5796	1.4185	1.6541	1.5116	1.7247	1.5777	0.3062
20	1.7106	1.4412	1.2361	1.3820	1.7601	1.5060	0.5240
21	1.4371	1.5051	1.3485	1.5670	1.4880	1.4691	0.2185
22	1.4738	1.5936	1.6583	1.4973	1.4720	1.5390	0.1863
23	1.5917	1.4333	1.5551	1.5295	1.6866	1.5592	0.2533
24	1.6399	1.5243	1.5705	1.5563	1.5530	1.5688	0.1156
25	1.5797	1.3663	1.6240	1.3732	1.6887	1.5264	0.3224

$$\Sigma \bar{x}_i = 37.6400 \quad \Sigma R_i = 8.1302$$

$$\bar{\bar{x}} = 1.5056 \quad \bar{R} = 0.32521$$

When setting up  $\bar{x}$  and R charts, it is best to begin with the R chart because the control limits on the  $\bar{x}$  chart depend on the process variability, unless process variability is in control, these limits will not have much meaning. Using the data in the table of the exercise, we find the center line for R chart is

$$\bar{R} = \left\{ \sum_{i=1}^{25} (R_i) / 25 \right\} = 8.1302 / 25 = 0.32521$$

For samples of  $n = 5$ , we find from Appendix Table VI that  $D_3 = 0$  and  $D_4 = 2.114$ . Therefore, the control limits for the R chart are

$$\underline{LCL} = \underline{R} D_3 = 0.32521(0) = 0$$

$$UCL = RD_4 = 0.32521(2.114) = 0.68749$$

The R chart is shown in subsequent slide. Both control charts were constructed by Minitab which uses more decimal places in the calculation than we did. When the 25 sample ranges are plotted on the R chart there is no indication of an out-of-control condition.

Since R chart indicates that process variability is in control, we may now construct  $\bar{x}$  chart. The center line is

$$\bar{\bar{x}} = (\sum \bar{x}_i)/25 = 37.64/25 = 1.5056 \text{ (Summation limit } i = 1 \text{ to } 25\text{)}$$

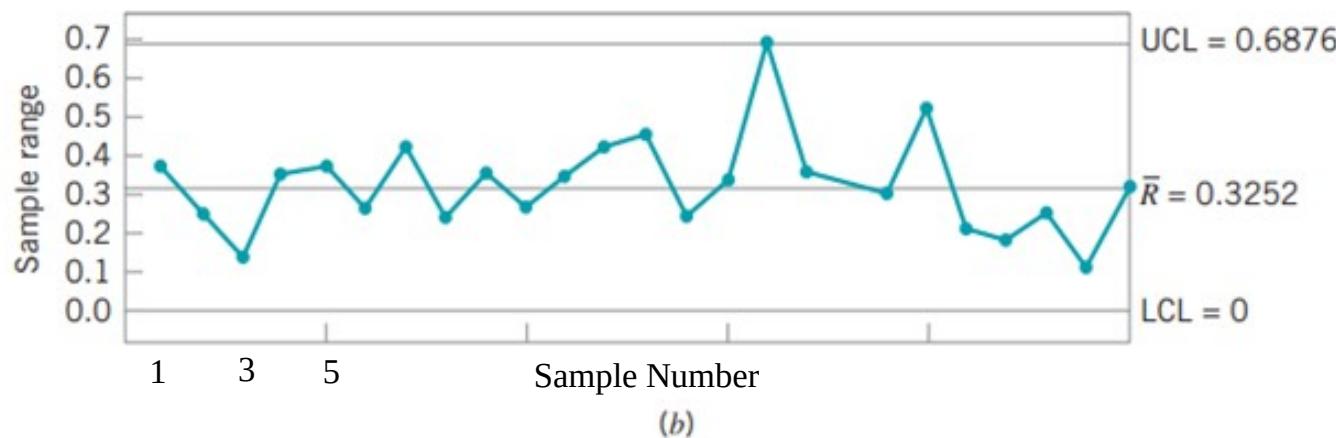
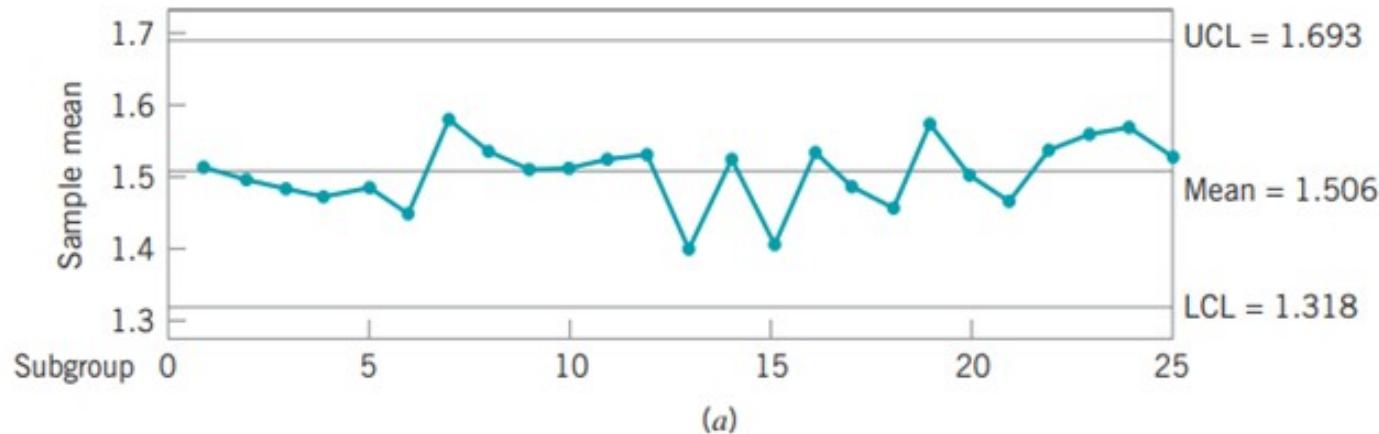
To find the control limits on the  $\bar{x}$  chart, we use  $A_2 = 0.577$  from Appendix Table VI for sample of size  $n = 5$  and the equation of control limit for x chart to find

$$\begin{aligned} \text{UCL} &= \bar{x} + A_2 R = 1.5056 + (0.577)(0.32521) = 1.69325 \text{ and} \\ &= \bar{x} - A_2 R \end{aligned}$$

$$\text{UCL} = \bar{x} - A_2 R = 1.5056 - (0.577)(0.32521) = 1.31795$$

The x chart is also shown. When the preliminary sample averages are plotted on this chart, no indication of an out-of-control condition is observed.

Therefore, since x and R charts exhibit control, we would conclude that the process is in control at the stated levels and adopt the trial control limits for use in phase II, where monitoring of future production is of interest.



$\bar{x}$  (a) and R charts (b) (from Minitab) for flow width in hard-bake process

## Interpretation of $\bar{x}$ and R charts

If all points fall in between UCL and LCL in both mean and range charts without showing in specific pattern then the process is said to be in the state of under statistical control otherwise the process is lack or out of statistical control.

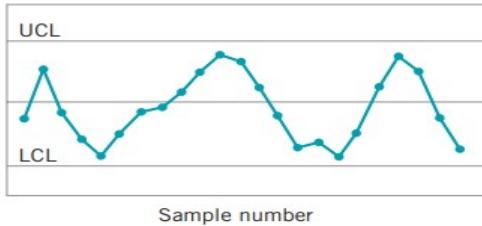
Lack of control in Mean chart indicates machine problems.

Lack of control in Range chart indicates workers problems.

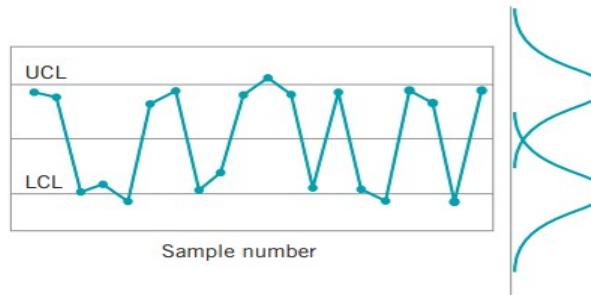
### **Cases of Lack of control :**

If at least one point falls outside UCL and / or LCL.

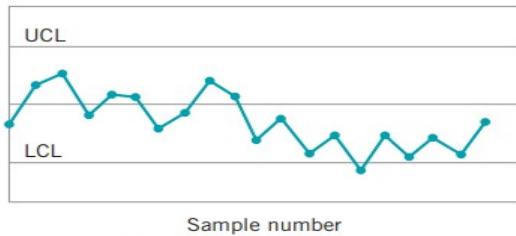
If all points fall in between the limits but showing specific patterns which are shown in the next slide.



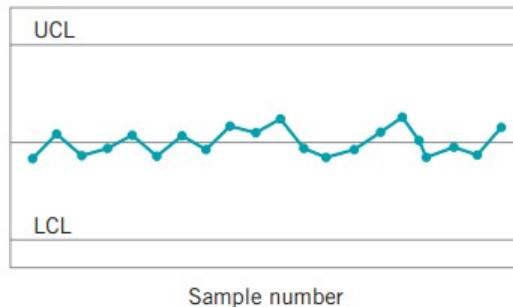
■ **FIGURE 6.8** Cycles on a control chart.



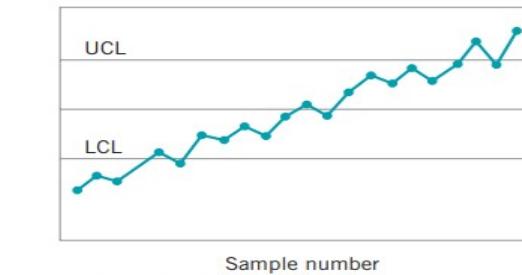
■ **FIGURE 6.9** A mixture pattern.



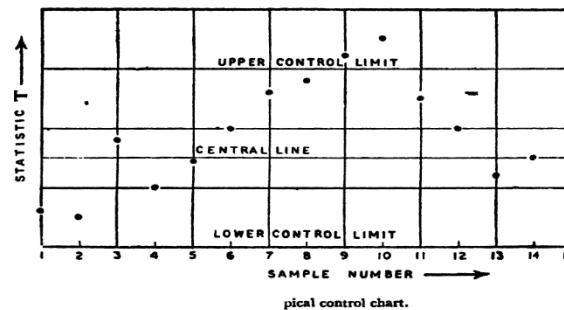
■ **FIGURE 6.10** A shift in process level.



■ **FIGURE 6.12** A stratification pattern.

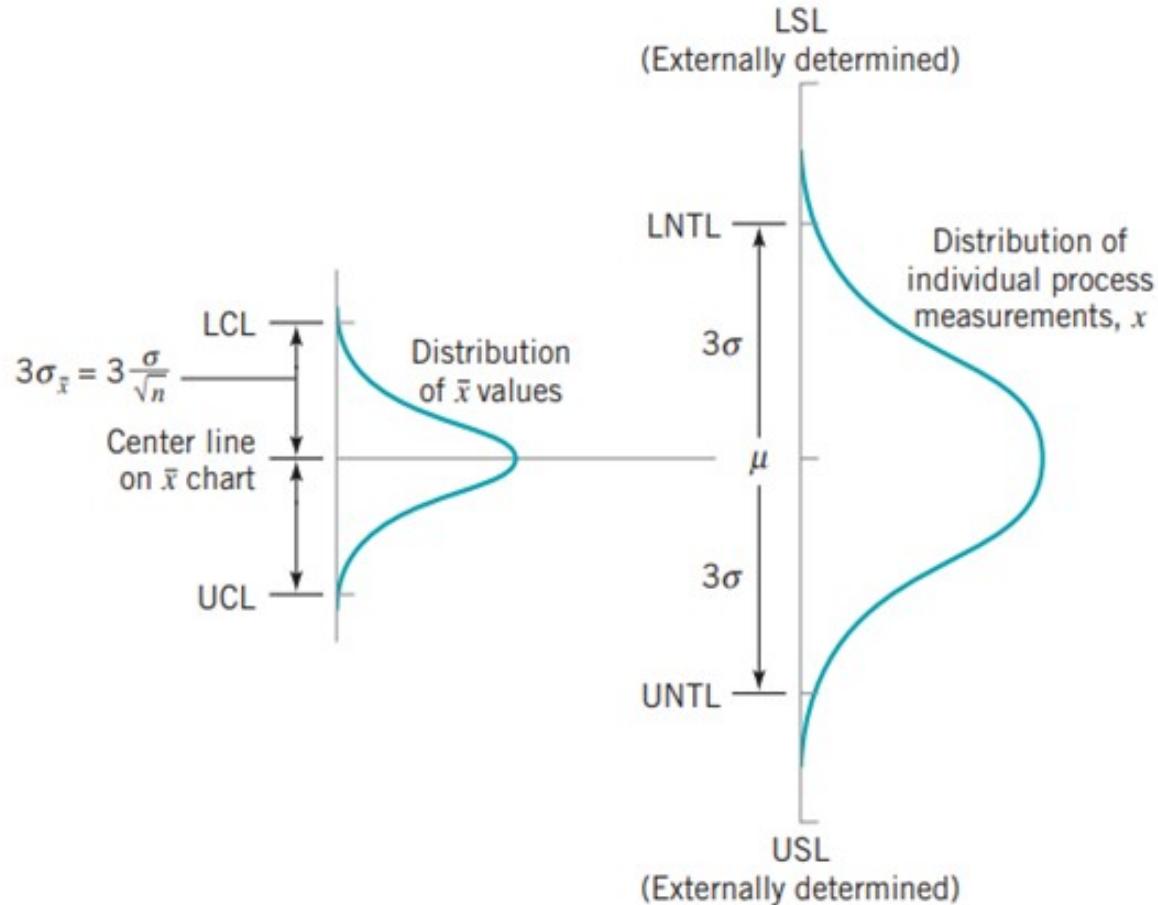


■ **FIGURE 6.11** A trend in process level.



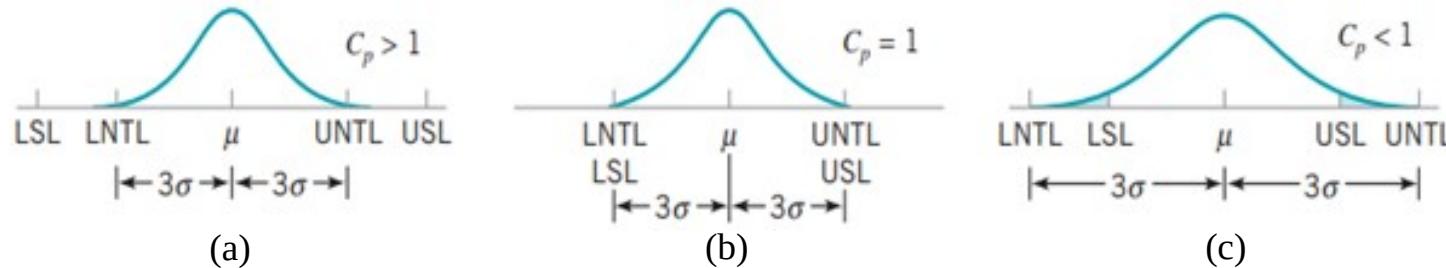
From the above chart, e.g., it appears that the process has been out of control in the 9th and 10th samples.

Control Limits, Specification Limits and Natural Tolerance Limis. A point that should be emphasized is that there is no connection or relationship between the control limits on the  $x$  and  $R$  charts and the specification limits on the process. The control limits are driven by the natural variability of the process (measured by the process standard deviation  $\sigma$ ), that is, by the natural tolerance limits of the process. It is customary to define the upper and lower natural tolerance limits, say UNTL and LNTL, as  $3\sigma$  above and below the process mean. The specification limits, on the other hand, are determined externally. They may be set by management, the manufacturing engineers, the customer or the product developers / designers. One should have knowledge of inherent process variability when setting specifications, but remember that there is no mathematical or statistical relationship between the control limits and specification limits. The situation is summarized in the figure in next slide. We have encountered practitioners who have plotted specification limits on the  $x$  control chart. This practice is completely incorrect and should not be done.



Relationship of natural tolerance limits, control limits and specification limits.

Following figure illustrates three cases of interest relative to the PCR  $C_p$  and process specification. In figure (a) PCR  $C_p$  is greater than unity. This means that the process uses up much less than 100% of the tolerance band. Consequently, relatively few non-conforming units will be produced by this process.



Process fallout and the Process Capability Ratio  $C_p$

Figure (b) shows a process for which the PCR  $C_p = 1$ ; ie. the process uses up all the tolerance band. For a normal distribution this would imply about 0.27% (or 2700 ppm) nonconforming units. Finally figure (c) presents the process for which the PCR  $C_p < 1$ ; ie. the process uses up more than 100% of the tolerance band. In this case the process is very yield-sensitive and a large number of nonconforming units will be produced. Note that all the above cases assume that the process is centered at the midpoint of the specification. In many situations this will not be the case.

Estimating Process Capability : The  $\bar{x}$  and R charts provide information about the performance or process capability of the process. From the  $\bar{x}$  chart, we may estimate the mean flow width of the resist in the hard-bake process as  $\bar{x} = 1.5056$  microns. The process standard deviation may be estimated as

$$\hat{\sigma} = \bar{R}/d_2 = 0.32521/2.326 = 0.1398 \text{ microns}$$

where the value of  $d_2$  for samples of size 5 is found in Appendix Table VI. The specification limits on flow width are  $1.50 \pm 0.50$  microns. The control chart data may be used to describe the capability of the process to produce wafers relative to these specifications. Assuming that flow width is a normally distributed random variable, with mean 1.5056 and standard deviation 0.1398, we may estimate the fraction of nonconforming wafers produced as

$$p = P\{x < 1.00\} + P\{x > 2.99\}$$

$$= \Phi\{(1.00 - 1.5056)/0.1398\} + 1 - \Phi\{(2.99 - 1.5056)/0.1398\}$$

$$= \Phi(-3.61660) + 1 - \Phi(3.53648) \approx 0.00015 + 1 - 0.99980 \approx 0.00035$$

This is about 0.035% [350 ppm] of the wafers produced will be outside of the specifications.

Another way to express process capability is in terms of the process capability ratio (PCR)  $C_p$ , which for a quality characteristic with both upper & lower specification limits (USL & LSL respectively) is

$$C_p = (USL - LSL)/6\sigma$$

Note that the  $6\sigma$  spread of the process is the basic definition of process capability. Since  $\sigma$  is usually unknown, we must replace it with an estimate. We frequently use

$\hat{\sigma} = \bar{R}/d_2$  as an estimate of  $\sigma$ , resulting in an estimate  $C_{p\hat{\sigma}}$  of  $C_p$ . For the hard-bake process, since  $\bar{R}/d_2 = \hat{\sigma} = 0.1398$ , we find that

$$C_{p\hat{\sigma}} = (2 - 1)/6(0.1398) = 1/0.8388 = 1.192$$

This implies that the “natural” tolerance limits in the process ( $3\sigma$  above & below the mean) are inside the lower & upper specification limits. Consequently, a moderately small number of nonconforming wafers will be produced. The PCR  $C_p$  may be interpreted another way. The quantity

$$P = (1/C_p)100\%$$

---

is simply the percentage of the specification band that the process uses up. For the hard-bake process an estimate of  $P$  is

$$\hat{P} = (1/C_{\hat{p}})100\% = (1/1.192)100\% = 83.89$$

That is, the process uses up about 84% of the specification band.

Charts based on Standard Values : When it is possible to specify standard values for the process mean & standard deviation, we may use these standards to establish the control charts for  $\bar{x}$  and R without analysis of past data. Suppose that the standards given are  $\mu$  and  $\sigma$ . Then the parameters of  $\bar{x}$  chart are

$$UCL = \mu + 3\sigma/\sqrt{n}$$

$$\text{Center line} = \mu$$

$$LCL = \mu - 3\sigma/\sqrt{n}$$

Thus the parameters of the R chart with standard  $\sigma$  given are

$$UCL = D_2\sigma$$

$$\text{Center line} = d_2\sigma$$

$$LCL = D_1\sigma$$

Control Charts for  $\bar{x}$  and  $s$  : Although  $\bar{x}$  and R charts are widely used, it is occasionally desirable to estimate the process standard deviation directly instead of indirectly through the use of the Range R. This leads to control charts for  $\bar{x}$  and  $s$ , where  $s$  is the sample standard deviation. Generally,  $\bar{x}$  and  $s$  charts are preferable to their more familiar counterparts,  $\bar{x}$  and R charts when either

1. The sample size  $n$  is moderately large say,  $n > 10$  or 12. (Recall that the range method for estimating  $\sigma$  loses statistical efficiency for moderate to large samples).
2. The sample size  $n$  is variable.

Consequently, the parameters of the  $s$  chart with a standard value for  $\sigma$  given become

$$UCL = B_6 \sigma$$

$$\text{Center line} = C_4 \sigma$$

$$LCL = B_5 \sigma$$

Values of  $B_5$  &  $B_6$  are tabulated for various sample sizes in Appendix Table VI. The parameters of the corresponding  $x$  chart have been given earlier.

If no standard is given for  $\sigma$ , then it must be estimated by analyzing past data. Suppose that  $m$  preliminary samples are available, each of size  $n$ , and let  $s_i$  be the standard deviation of the  $i^{\text{th}}$  sample. The average of the  $m$  standard deviation is

$$\bar{s} = (1/m) \left( \sum_{i=1}^m s_i \right)$$

$$s_i = \sqrt{\left\{ \sum_{i=1}^n (x_i - \bar{x})^2 / (n - 1) \right\}}$$

Consequently, we may write the parameters of the  $s$  chart as

$$\text{UCL} = B_4 \bar{s}$$

$$\text{Center line} = \bar{s}$$

$$\text{LCL} = B_3 \bar{s}$$

Let the ~~constant~~  $A_3 = 3/(C_4 \sqrt{n})$ . Then the  $x$  chart parameters become

$$\text{UCL} = \bar{x} + A_3 \bar{s}$$

$$\text{Center line} = \bar{x}$$

$$\text{LCL} = \bar{x} - A_3 \bar{s}$$

Ex :  $\bar{x}$  and s charts for the Piston Ring Data – Construct & interpret x & s charts using the piston ring inside diameter measurements in the following table.

Inside Diameter Measurements (mm) for Automobile Engine Piston Ring

Sample Number	Observations				
1	74.030	74.002	74.019	73.992	74.008
2	73.995	73.992	74.001	74.011	74.004
3	73.988	74.024	74.021	74.005	74.002
4	74.002	73.996	73.993	74.015	74.009
5	73.992	74.007	74.015	73.989	74.014
6	74.009	73.994	73.997	73.985	73.993
7	73.995	74.006	73.994	74.000	74.005
8	73.985	74.003	73.993	74.015	73.988
9	74.008	73.995	74.009	74.005	74.004
10	73.998	74.000	73.990	74.007	73.995
11	73.994	73.998	73.994	73.995	73.990
12	74.004	74.000	74.007	74.000	73.996
13	73.983	74.002	73.998	73.997	74.012
14	74.006	73.967	73.994	74.000	73.984
15	74.012	74.014	73.998	73.999	74.007
16	74.000	73.984	74.005	73.998	73.996
17	73.994	74.012	73.986	74.005	74.007
18	74.006	74.010	74.018	74.003	74.000
19	73.984	74.002	74.003	74.005	73.997
20	74.000	74.010	74.013	74.020	74.003
21	73.982	74.001	74.015	74.005	73.996
22	74.004	73.999	73.990	74.006	74.009
23	74.010	73.989	73.990	74.009	74.014
24	74.015	74.008	73.993	74.000	74.010
25	73.982	73.984	73.995	74.017	74.013

Soln :

Sample Number	Observations					$\bar{X}_i$	$S_i$
1	74.030	74.002	74.019	73.992	74.008	74.010	0.0148
2	73.995	73.992	74.001	74.011	74.004	74.001	0.0075
3	73.988	74.024	74.021	74.005	74.002	74.008	0.0147
4	74.002	73.996	73.993	74.015	74.009	74.003	0.0091
5	73.992	74.007	74.015	73.989	74.014	74.003	0.0122
6	74.009	73.994	73.997	73.985	73.993	73.996	0.0087
7	73.995	74.006	73.994	74.000	74.005	74.000	0.0055
8	73.985	74.003	73.993	74.015	73.988	73.997	0.0123
9	74.008	73.995	74.009	74.005	74.004	74.004	0.0055
10	73.998	74.000	73.990	74.007	73.995	73.998	0.0063
11	73.994	73.998	73.994	73.995	73.990	73.994	0.0029
12	74.004	74.000	74.007	74.000	73.996	74.001	0.0042
13	73.983	74.002	73.998	73.997	74.012	73.998	0.0105
14	74.006	73.967	73.994	74.000	73.984	73.990	0.0153
15	74.012	74.014	73.998	73.999	74.007	74.006	0.0073
16	74.000	73.984	74.005	73.998	73.996	73.997	0.0078
17	73.994	74.012	73.986	74.005	74.007	74.001	0.0106
18	74.006	74.010	74.018	74.003	74.000	74.007	0.0070
19	73.984	74.002	74.003	74.005	73.997	73.998	0.0085
20	74.000	74.010	74.013	74.020	74.003	74.009	0.0080
21	73.982	74.001	74.015	74.005	73.996	74.000	0.0122
22	74.004	73.999	73.990	74.006	74.009	74.002	0.0074
23	74.010	73.989	73.990	74.009	74.014	74.002	0.0119
24	74.015	74.008	73.993	74.000	74.010	74.005	0.0087
25	73.982	73.984	73.995	74.017	74.013	73.998	0.0162
					$\Sigma = 1850.028$	0.2351	
					$\bar{x} = 74.001$	$\bar{s} = 0.0094$	

The grand average and the average standard deviation are

$$\bar{\bar{x}} = (1/25) \sum_{i=1}^{25} \bar{x}_i = (1/25)(1850.028) = 74.001 \text{ and}$$

$$\bar{s} = (1/25) \sum_{i=1}^{25} s_i = (1/25)(0.2351) = 0.0094$$

respectively. Consequently, the parameters for the  $\bar{x}$  chart are

$$\text{UCL} = \bar{\bar{x}} + A_3 \bar{s} = 74.001 + (1.427)(0.0094) = 74.014$$

$$\text{CL} = \bar{x} = 74.001$$

$$\text{LCL} = \bar{\bar{x}} - A_3 \bar{s} = 74.001 - (1.427)(0.0094) = 73.988$$

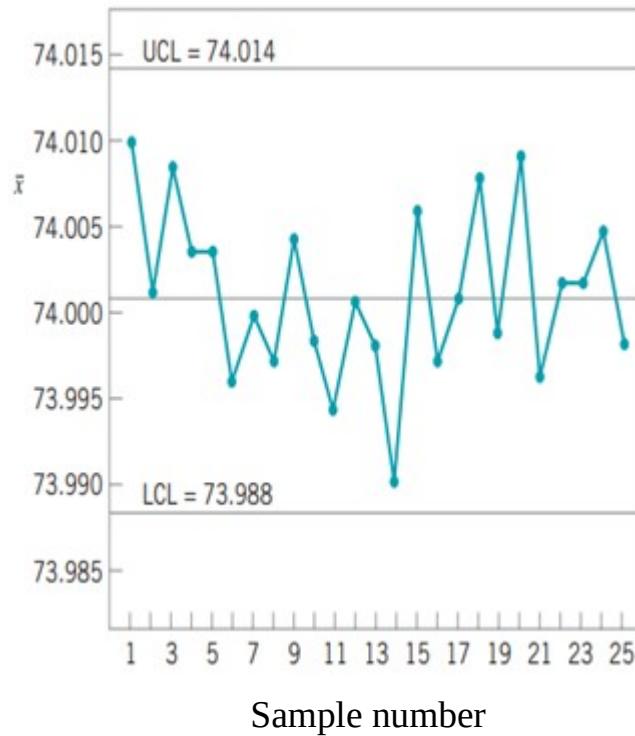
and the  $s$  chart

$$\text{UCL} = B_4 s = (2.089)(0.0094) = 0.0196$$

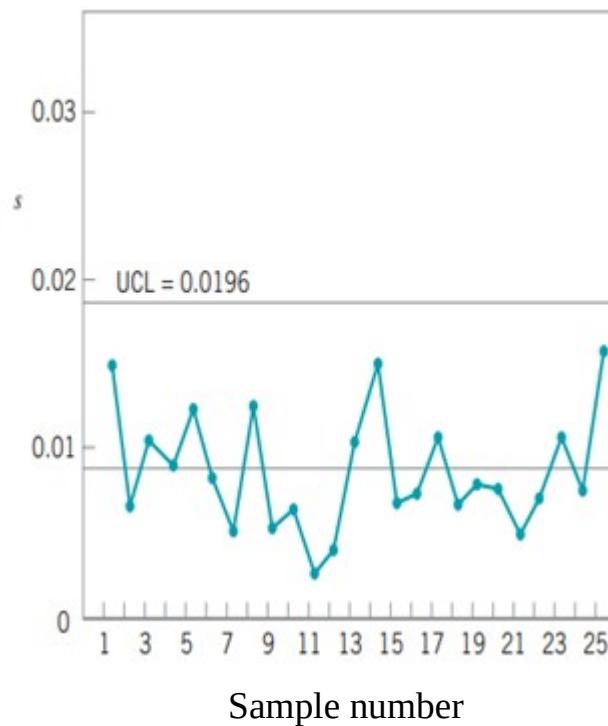
$$\text{CL} = s = 0.0094$$

$$\text{LCL} = B_3 s = (0)(0.0094) = 0$$

The control charts are shown here. There is no indication that the process is out of control, so those limits could be adopted for phase II monitoring of the process.



(a)



(b)

(a)  $\bar{x}$  chart and (b) s chart for the given example

## Control Chart of Attribute

p-Chart

np-Chart

C-Chart

## Control Chart for Fraction Nonconforming or Defective (p)

The fraction nonconforming or defective (p) is defined as the ratio of the number of nonconforming items in a population to the total number of items in that population. The items may have several quality characteristics that are examined simultaneously by the inspector. If the item does not conform to standard on one or more of these characteristics, it is classified as nonconforming.

The statistical principles underlying the control chart for fraction nonconforming are based on the binomial distribution. Suppose the production process is operating in a stable manner, such that probability that any unit will not conform to specifications is  $p$  and successive units produced are independent. Then each unit produced is a realization of Bernoulli random variable with parameter  $p$ . If the random sample of  $n$  units of product is selected and if  $D$  is the number of units of product that are nonconforming then

## Control Chart for Fraction Nonconforming or Defective (p)

$$P(D = x) = {}^nC_x p^x (1 - p)^{n-x}, \text{ for } x = 0, 1, \dots, n$$

We know for binomial random variable D mean & variance are  $np$  &  $np(1 - p)$  respectively.

The sample fraction nonconforming is defined as the ratio of the number of nonconforming units in the sample D to the sample size n; that is

$$\hat{p} = D/n$$

---

Suppose that the true fraction nonconforming  $p$  in the production process is known or is specified standard value. Then of the center line & control limits of the fraction nonconforming control chart would be as follows –

Fraction Nonconforming Control Chart : Standard Given

$$UCL = p + 3\sqrt{p(1-p)/n}$$

$$\text{Center line} = p$$

$$LCL = p - 3\sqrt{p(1-p)/n}$$

Depending on the values of  $p$  and  $n$ , sometimes the lower control limit  $LCL < 0$ . In these cases, we customarily set  $LCL = 0$  and assume that the control chart only has an upper control limit.

When the process fraction nonconforming  $p$  is not known, then it must be estimated from observed data. The usual procedure is to select  $m$  preliminary samples, each of size  $n$ . As the general rule,  $m$  should be at least 20 or 25. Then if there are  $D_i$  nonconforming units in sample  $i$ , we compute the fraction nonconforming in the  $i^{\text{th}}$  sample as

$$\hat{p}_i = D_i/n, \quad i = 1, 2, \dots, m$$

and the average of these individual samples fractions nonconforming is

$$\bar{p} = \left\{ \left( \frac{\sum_1^m D_i}{mn} \right) \right\} = \left\{ \left( \frac{\sum_1^m \hat{p}_i}{m} \right) \right\}$$

The statistic  $\bar{p}$  estimates the unknown fraction nonconforming  $p$ . The center line and control limits of the control chart for fraction nonconforming are computed as given in next slide.

### Fraction Nonconforming Control Chart: No Standard Given

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Center line =  $\bar{p}$  (7.8)

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

This control chart is also often called the p-chart.

## Ex : Construction & operation of a Fraction Nonconforming Control Chart

Frozen orange juice concentrate is packed in 6-oz cardboard cans. These cans are formed on a machine by spinning them from cardboard stock and attaching a metal bottom panel. By inspection of a can, we may determine whether, when filled, it could possibly leak either on the side seam or around the bottom joint. Such a nonconforming can has an improper seal on either the side seam or the bottom panel. Set up a control chart to improve the fraction of nonconforming cans produced by this machine.

Data for Trial Control  
Limits, Sample size = 50

Sample Number	Number of Nonconforming Cans, $D_i$	Sample Number	Number of Nonconforming Cans, $D_i$
1	12	17	10
2	15	18	5
3	8	19	13
4	10	20	11
5	4	21	20
6	7	22	18
7	16	23	24
8	9	24	15
9	14	25	9
10	10	26	12
11	5	27	7
12	6	28	13
13	17	29	9
14	12	30	6
15	22		
16	8		

Sample Number	Number of Nonconforming Cans, $D_i$	Sample Fraction Nonconforming, $\hat{p}_i$	Sample Number	Number of Nonconforming Cans, $D_i$	Sample Fraction Nonconforming, $\hat{p}_i$
1	12	0.24	17	10	0.20
2	15	0.30	18	5	0.10
3	8	0.16	19	13	0.26
4	10	0.20	20	11	0.22
5	4	0.08	21	20	0.40
6	7	0.14	22	18	0.36
7	16	0.32	23	24	0.48
8	9	0.18	24	15	0.30
9	14	0.28	25	9	0.18
10	10	0.20	26	12	0.24
11	5	0.10	27	7	0.14
12	6	0.12	28	13	0.26
13	17	0.34	29	9	0.18
14	12	0.24	30	6	0.12
15	22	0.44		347	$\bar{p} = 0.2313$
16	8	0.16			

To establish the control chart, 30 samples of sample size,  $n = 50$  cans each were selected at half-hour intervals over a three-shift period in which the machine was in continuous operation. The data are shown in the table that follows.

We construct a phase I control chart using the preliminary data to determine if the process was in control when these data were collected. Since the 30

samples contain  $\sum_{i=1}^{30} D_i = 347$  nonconforming cans, find

$$\bar{p} = \frac{\sum_{i=1}^{30} D_i}{mn} = 347/30 \times 50 = 0.2313$$

Using  $\bar{p}$  as an estimate of the true process fraction nonconforming, we can now calculate the upper & lower control limits as

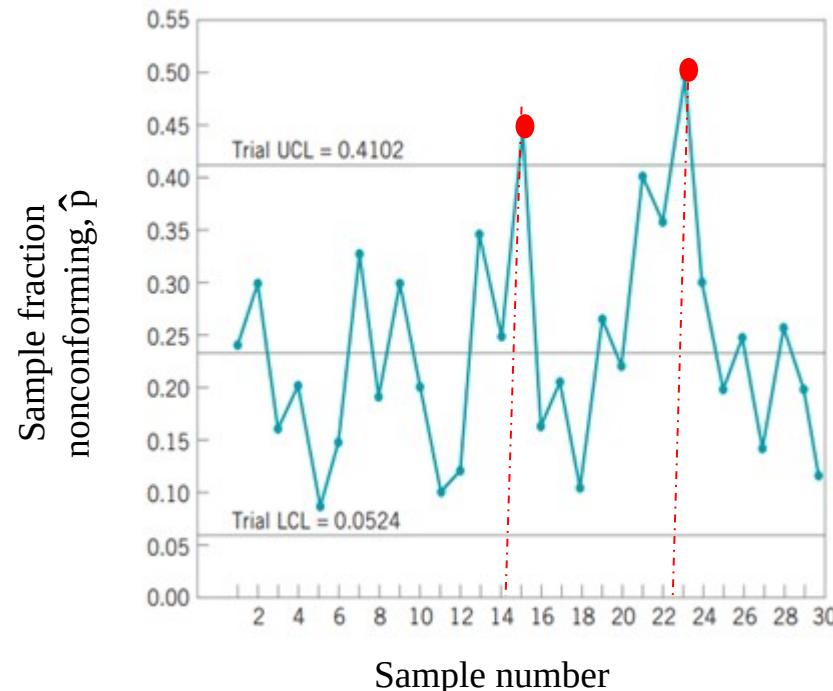
$$\bar{p} \pm 3\sqrt{\bar{p}(1 - \bar{p})/n} = 0.2313 \pm 3\sqrt{0.2313(0.7687)/50} = 0.2313 \pm 0.1789$$

Therefore,

$$UCL = \bar{p} + 3\sqrt{\bar{p}(1 - \bar{p})/n} = 0.2313 + 0.1789 = 0.4102$$

$$LCL = \bar{p} - 3\sqrt{\bar{p}(1 - \bar{p})/n} = 0.2313 - 0.1789 = 0.0524$$

The control chart with center line at  $\bar{p} = 0.2313$  and the above UCL & LCL are shown in the figure. The sample fraction nonconforming from each preliminary sample is plotted on this chart. We need that two points, those from sample 15 & 23, plot above the UCL, so the process is not in control. These points must be investigated to see whether an assignable cause can be determined.



Initial phase I fraction nonconforming control chart for given data

The np Control Chart : It is possible to base a control chart on the number non-conforming rather than the fraction nonconforming. This is often called an number nonconforming (np) control chart. The parameters of this chart are as follows.

### The *np* Control Chart

$$\text{UCL} = np + 3\sqrt{np(1-p)}$$
$$\text{Center line} = np$$
$$\text{LCL} = np - 3\sqrt{np(1-p)}$$

If a standard value for  $p$  is unavailable, then  $\bar{p}$  can be used to estimate  $p$ . Many nonstatistically trained personnel find the np chart easier to interpret than the usual fraction nonconforming control chart.

---

Ex : An np Control Chart - Set up an np control chart for the orange juice concentrate can process in earlier example.

Using the data in the table, we found that  $\bar{p} = 0.2313$ ,  $n = 50$

Soln : The parameters of the np control chart would be

$$UCL = n\bar{p} + 3\sqrt{n\bar{p}(1 - \bar{p})} = 50(0.2313) + 3\sqrt{50(0.2313)(0.7687)} = 20.510$$

$$CL = n\bar{p} = 50(0.2310) = 11.565$$

$$LCL = n\bar{p} - 3\sqrt{n\bar{p}(1 - \bar{p})} = 50(0.2313) - 3\sqrt{50(0.2313)(0.7687)} = 2.620$$

Now in practice, the number of nonconforming units in each sample is plotted on the np control chart, and the number of nonconforming units is an integer. Thus, if 20 units are nonconforming the process is in control, but if 21 occur the process is out of control. Similarly, there are three nonconforming units in the sample and the process is in control, but two nonconforming units would imply an out-of-control process. Some practitioners prefer to use integer values for the

---

control limits on the np chart instead of their decimal fraction counterparts. In this example we could choose 2 and 21 as the LCL and UCL respectively and the process would be considered out-of-control if a sample value of np plotted at or beyond the control limits.

---

**Variable Sample Size :** In some applications of the control chart for fraction nonconforming, the sample is a 100% inspection of process output over some period of time. Since different numbers of units could be produced in each period, the control chart would then have a variable sample size. There

**Variable-Width Control Limits :** The first and perhaps the most simple approach is to determine control limits for each individual sample that are based on the specific sample size. That is, if the  $i^{\text{th}}$  sample is of size  $n_i$ , then the upper and lower control limits are  $\bar{p} \pm 3\sqrt{\{p(1 - p)/n_i\}}$ . Note that the width of the control limits is inversely proportional to the square root of the sample size.

To illustrate this approach, consider the data in Table 7.4. These data came from the purchasing group of a large aerospace company. This group issues purchase orders to the company's suppliers. The sample size in Table 7.4 are the total number of purchase orders issued each week. Obviously, this is not constant. A nonconforming unit is a purchase order with an error. Among the most common errors are specifying incorrect part numbers, wrong delivery dates and wrong

---

---

supplier information. Any of these mistakes can result in a purchase order change, which takes time and resources and may result in delayed delivery of material.

**TABLE 7.4**

Purchase Order Data for a Control Chart for Fraction Nonconforming with Variable Sample Size

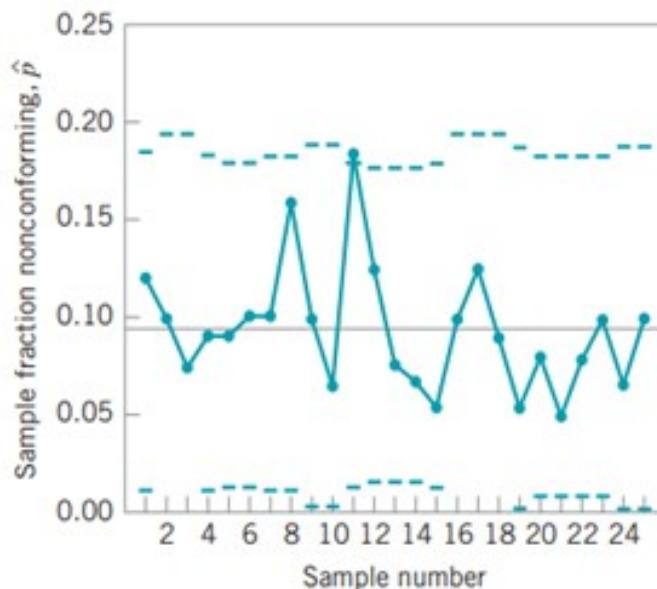
Sample Number, $i$	Sample Size, $n_i$	Number of Nonconforming Units, $D_i$	Sample Fraction Nonconforming, $\hat{p}_i = D_i/n_i$	Standard Deviation		Control Limits	
				$\hat{\sigma}_{\hat{p}} = \sqrt{\frac{(0.096)(0.904)}{n_i}}$		LCL	UCL
1	100	12	0.120	0.029		0.009	0.183
2	80	8	0.100	0.033		0	0.195
3	80	6	0.075	0.033		0	0.195
4	100	9	0.090	0.029		0.009	0.183
5	110	10	0.091	0.028		0.012	0.180
6	110	12	0.109	0.028		0.012	0.180
7	100	11	0.110	0.029		0.009	0.183
8	100	16	0.160	0.029		0.009	0.183
9	90	10	0.110	0.031		0.003	0.189
10	90	6	0.067	0.031		0.003	0.189
11	110	20	0.182	0.028		0.012	0.180
12	120	15	0.125	0.027		0.015	0.177
13	120	9	0.075	0.027		0.015	0.177
14	120	8	0.067	0.027		0.015	0.177
15	110	6	0.055	0.028		0.012	0.180
16	80	8	0.100	0.033		0	0.195
17	80	10	0.125	0.033		0	0.195
18	80	7	0.088	0.033		0	0.195
19	90	5	0.056	0.031		0.003	0.189
20	100	8	0.080	0.029		0.009	0.183
21	100	5	0.050	0.029		0.009	0.183
22	100	8	0.080	0.029		0.009	0.183
23	100	10	0.100	0.029		0.009	0.183
24	90	6	0.067	0.031		0.003	0.189
25	90	9	0.100	0.031		0.003	0.189
	2450	234	2.383				

For 25 samples, we calculate  $\bar{p} = \sum_{i=1}^{25} D_i / \sum n_i = 234/2450 = 0.096$

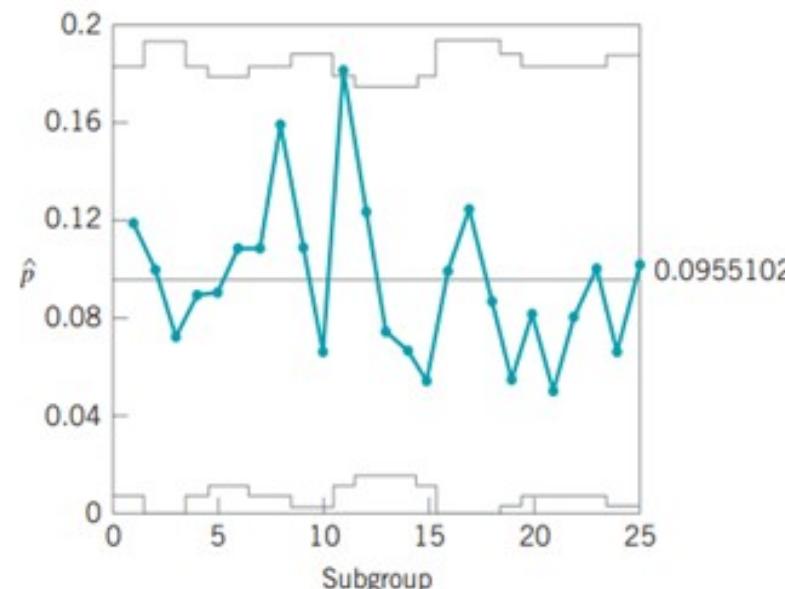
Consequently, the center line is at 0.096 and control limits are

$$UCL = \bar{p} + 3\hat{\sigma}_p = 0.096 + 3\sqrt{\{(0.096)(0.904)/n_i\}}$$

$$LCL = \bar{p} - 3\hat{\sigma}_p = 0.096 - 3\sqrt{\{(0.096)(0.904)/n_i\}}$$



Control chart for fraction nonconforming with variable sample size



Control chart for fraction nonconforming with variable sample size using Minitab

## **C-Control Chart for Nonconformities (Number of Defects per item:C)**

A nonconforming item is a unit of product that does not satisfy one or more of the specifications for that product. Each specific point at which a specification is not satisfied results in a defect or nonconformity. Consequently, a nonconforming item will contain at least one nonconformity. However, depending on their nature and severity, it is quite possible for a unit to contain several nonconformities and not be classified as nonconforming. As an example, suppose we are manufacturing personal computers. Each unit could have one or more very minor flaws in the cabinet finish, and since these flaws do not seriously affect the unit's functional operation, it could be classified as conforming. However, if there are too many of these flaws, the personal computer should be classified as nonconforming, since the flaws will be very noticeable to the customer and might affect the sale of the unit. There are many practical situations in which we prefer to work directly with the number of defects or nonconformities rather than the fraction nonconforming. These include the number of defective welds in 100 m of oil pipeline, the number of broken rivets in an aircraft wing, the number of functional defects in an electronic logic device, the number of errors on a document and so forth.

Consider the occurrence of nonconformities in an inspection unit of product. In most cases, the inspection unit will be a single unit of product, although this is not necessarily always so. The inspection unit is simply an entity for which it is convenient to keep records. It could be a group of 5 units of product, 10 units of product and so on. Suppose that defects or nonconformities occur in this inspection unit according to the Poisson distribution; that is

$$p(x) = e^{-c} c^x / x! \quad x = 0, 1, 2, \dots$$

where  $x$  is the number of nonconformities and  $c > 0$  is the parameter of the Poisson distribution. We know that the mean and variance of the Poisson distribution are the parameter  $c$ . Therefore, a control chart for nonconformities, or  $c$  chart with three sigma limits would be defined as follows.

Control Chart for Nonconformities : Standard Given

$$\begin{aligned} UCL &= c + 3\sqrt{c} \\ \text{Center line} &= c \\ LCL &= c - 3\sqrt{c} \end{aligned}$$

Assuming that a standard value for  $c$  is available. Should these calculations yield a negative value for the LCL, set  $LCL = 0$

## Control chart for Nonconformities : No Standard Given

$$UCL = \bar{c} + 3\sqrt{\bar{c}}$$

$$\text{Center line} = \bar{c}$$

$$LCL = \bar{c} - 3\sqrt{\bar{c}}$$

Ex : Nonconformities in Printed Circuit Boards - Table 7.7 below presents the number of nonconformities observed in 26 successive samples of 100 PCBs. Note that, for reasons of convenience, the inspection unit is defined as 100 boards. Set up a c chart for these data.

[Table 7.7 Data on the Number of Nonconformities in Samples of 100 PCBs](#)

Sample Number	Number of Nonconformities	Sample Number	Number of Nonconformities
1	21	14	19
2	24	15	10
3	16	16	17
4	12	17	13
5	15	18	22
6	5	19	18
7	28	20	39
8	20	21	30
9	31	22	24
10	25	23	16
11	20	24	19
12	24	25	17
13	16	26	15

Soln : Since 26 samples contain 516 total nonconformities, we estimate  $c$  by

$$\bar{c} = 516/26 = 19.85$$

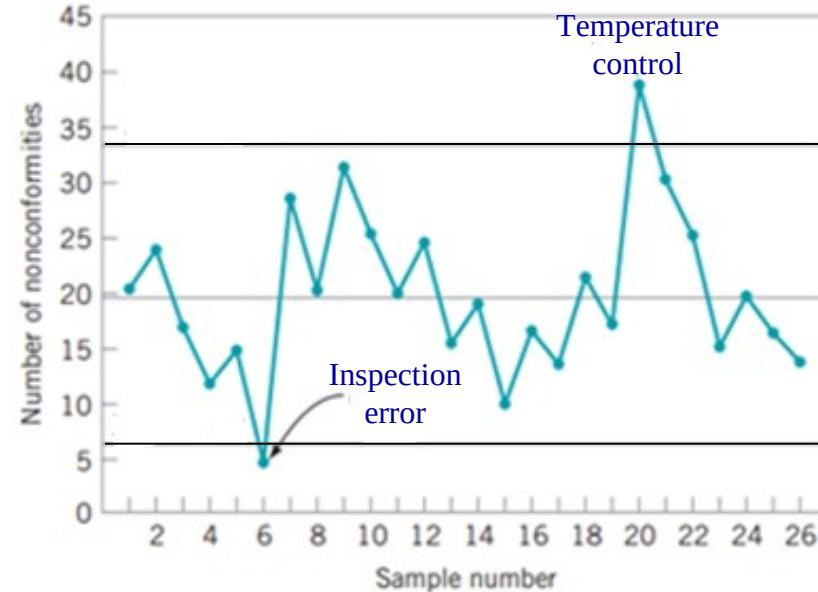
Therefore, the trial control limits are given by

$$UCL = \bar{c} + 3\sqrt{c} = 19.85 + 3\sqrt{19.85} = 33.22$$

$$CL = \bar{c} = 19.85$$

$$LCL = \bar{c} - 3\sqrt{c} = 19.85 - 3\sqrt{19.85} = 6.48$$

The control chart for the nonconformities is shown here. The number of observed nonconformities from the preliminary samples is plotted on this chart. Two points (samples 6 & 20) plot outside the control limits. Investigation of sample 6 revealed that a new inspector had examined the PCBs in this sample & that he did not recognize several of the types of nonconformities that could have been present. Furthermore, the unusually large number of nonconformities in sample 20 resulted from a temperature control problem in the wave soldering machine, which was subsequently repaired.



Ex : The data shown in the table are  $\bar{x}$  and  $R$  values for 24 samples of size  $n = 5$  taken from a process producing bearings. The measurements are made on the inside diameter of the bearing, with only last three decimals recorded (ie., 34.5 should be recorded 0.50345).

- (a) Set up  $\bar{x}$  and  $R$  charts on this process. Does the process seem to be in statistical control?
- (b) If specifications on this diameter are  $0.5030 \pm 0.0010$ , find the percentage of nonconforming bearings produced by this process. Assume that diameter is normally distributed.

Bearing Diameter Data

Sample Number	$\bar{x}$	$R$	Sample Number	$\bar{x}$	$R$
1	34.5	3	13	35.4	8
2	34.2	4	14	34.0	6
3	31.6	4	15	37.1	5
4	31.5	4	16	34.9	7
5	35.0	5	17	33.5	4
6	34.1	6	18	31.7	3
7	32.6	4	19	34.0	8
8	33.8	3	20	35.1	4
9	34.8	7	21	33.7	2
10	33.6	8	22	32.8	1
11	31.9	3	23	33.5	3
12	38.6	9	24	34.2	2

Samples of  $n = 6$  items each are taken from a process at regular intervals. A quality characteristic is measured, and  $\bar{x}$  and  $R$  values are calculated for each sample. After 50 samples, we have

$$\sum_{i=1}^{50} \bar{x}_i = 2000 \quad \text{and} \quad \sum_{i=1}^{50} R_i = 200$$

Assume that the quality characteristic is normally distributed.

- (a) Compute control limits for the  $\bar{x}$  and  $R$  control charts.
- (b) All points on both control charts fall between the control limits computed in part (a). What are the natural tolerance limits of the process?
- (c) If the specification limits are  $41 \pm 5.0$ , what are your conclusions regarding the ability of the process to produce to produce items within these specifications?
- (d) Assuming that if an item exceeds the upper specification limit it can be reworked and if it is below the lower specification limit it must be scrapped, what percent scrap and rework is the process producing?

The number of nonconforming switches in samples of size 150 are shown in the following table. Construct a fraction nonconforming control chart for these data. Does the process appear to be in control?

**Number of Nonconforming Switches**

Sample Number	Number of Nonconforming Switches	Sample Number	Number of Nonconforming Switches
1	8	11	6
2	1	12	0
3	3	13	4
4	0	14	0
5	2	15	3
6	4	16	1
7	0	17	15
8	1	18	2
9	10	19	3
10	6	20	0

The Table below represents the results of inspecting all units of a personal computer produced for the past ten days. Does the process appear to be in control?

Personal Computer Inspecting Results

Day	Units Inspected	Nonconforming Units	Fraction Nonconforming
1	80	4	0.050
2	110	7	0.064
3	90	5	0.056
4	75	8	0.107
5	130	6	0.046
6	120	6	0.050
7	70	4	0.057
8	125	5	0.040
9	105	8	0.076
10	95	7	0.074

Based on the data in the given Table if an np chart is to be established, what would you recommend as the center line and control limits? Assume that n = 500.

#### Data for the Exercise

<b>Day</b>	<b>Number of Nonconforming Units</b>
1	3
2	4
3	3
4	2
5	6
6	12
7	5
8	1
9	2
10	2

The tabulated data represent the number of nonconformities per 1000 meters in telephone cable. From analysis of these data would you conclude that the process is in statistical control? What control procedure would you recommend for future production?

**Telephone Cable Data for the exercise**

<b>Sample Number</b>	<b>Number of Nonconformities</b>	<b>Sample Number</b>	<b>Number of Nonconformities</b>
1	1	12	6
2	1	13	9
3	3	14	11
4	7	15	15
5	8	16	8
6	10	17	3
7	5	18	6
8	13	19	7
9	0	20	4
10	19	21	9
11	24	22	20



# BSDCHZC355

## Statistical Inference & Applications

**BITS** Pilani  
Pilani | Dubai | Goa | Hyderabad

Shaibal Kr. Sen  
Session 16 (Contact Hr 1& 2)



## STATISTICAL INFERENCESES

&

## APPLICATIONS

(Session 16)

---

# ➤ REVISION SESSION

---

**Ex 1:** A leading Marketing Consultancy Group conducted a research to investigate how product type and advertisement platform affect the success of their campaigns. An online advertisement campaign was undertaken for two types of products viz. - Consumer Durables and Office Automation, on two different platforms - Google Ads and LinkedIn Ads and recorded the average number of viewers for each combination as tabulated below.

Product Type	Google Ads	LinkedIn
Consumer Durables	4500	3300
Office Automation	3900	2400

The Consultancy Group want to analyze - whether the platform (Google Ads vs. LinkedIn Ads) matters, whether product type matters and whether there is an interaction (i.e., if effect of platform depends on the product type).

- Based on the given averages, is there any visible interaction between product type and platform? Explain briefly.
- Which factor appears to have a stronger influence on the number of views: the platform used or the product type? Support your answer with reasoning

Product Type	Google Ads	LinkedIn
Consumer Durables	4500	3300
Office Automation	3900	2400

Soln : (a) It is evident that the performance drops from Google Ads to LinkedIn for each product as below -

Consumer Durables :  $4500 - 3300 = 1200$

Office Automation :  $3900 - 2400 = 1500$

Since the drop in views is different for the two products, this suggests that the platform's effect depends on the product type. Therefore, it can be inferred that there appears to be an interaction between platform and product type.

(b) We can compare magnitude of changes to verify which factor has stronger influence -

Difference between products on Google Ads =  $4500 - 3900 = 600$

Difference between products on LinkedIn =  $3300 - 2400 = 900$

But across platforms, the drop is larger -

Consumer Durables : 1200

Office Automation : 1500

This shows that the platform, Google Ads and LinkedIn causes a bigger change in reach than product type which means platform has a stronger effect on viewership.

**Ex 2 :** In a Completely Randomized Design (CRD), a laboratory tested the effectiveness of three cleaning solutions -  $CS_1$ ,  $CS_2$  and  $CS_3$  on stain removal. Each solution was randomly applied to four glass slides and the stain removal effectiveness was rated on a scale from 1 to 10. The results were as follows -

Solution  $CS_1$  : 8, 9, 8, 10 = 35

Solution  $CS_2$  : 4, 5, 6, 7 = 22

Solution  $CS_3$  : 6, 7, 8, 8 = 29

(a) Compute the Correction Factor

(b) Compute the Total Sum of Squares (TSS)

**Soln :** (a) Grand Total (G) = 8 + 9 + 8 + 10 + 4 + 5 + 6 + 7 + 6 + 7 + 8 + 8 = 86

$$CF = G^2 / N = 86^2 / 12 \approx 616.33 \text{ (Here, } G = 86 \text{ and } N = 12\text{)}$$

$$(b) TSS = \sum(Y_{ij})^2 - CF$$

$$= (8^2 + 9^2 + 8^2 + 10^2 + 4^2 + 5^2 + 6^2 + 7^2 + 6^2 + 7^2 + 8^2 + 8^2) - 616.33$$

$$= (64 + 81 + 64 + 100 + 16 + 25 + 36 + 49 + 36 + 49 + 64 + 64) - 616.33$$

$$= 648 - 616.33 = 31.67$$

---

**Ex 3 :** An experimenter performs a Randomized Block Design where six different seeding rates (Treatments  $T_1$  to  $T_6$ ) were tested over four field blocks. The following summary values were provided from the analysis:

Grand Total (G) = 285.36

Correction Factor (CF) = 1413.72

Total Sum of Squares (TSS) = 12.048

Sum of Squares for Blocks (SSB) = 4.716

Sum of Squares for Treatments (SST) = 3.042

(a) Calculate the Error Sum of Squares (SSE).

(b) Based on the given sums of squares, which factor - blocks or treatments has a larger impact on variation? Justify your answer.

**Soln :** (a) 
$$\begin{aligned} \text{SSE} &= \text{TSS} - \text{SSB} - \text{SST} \\ &= 12.048 - 4.716 - 3.042 = 4.29 \end{aligned}$$

(b) Since  $\text{SSB} = 4.716$  is larger than  $\text{SST} = 3.042$ , we conclude that blocks have a stronger effect on the response variable than the treatments

---

**Ex 4 :** Monthly sales (in lakhs of Rs.) of a distributor in the four Regional Centers during consecutive four months are given below –

Northern Center - 100, 105, 110, 108

Southern Center - 82, 86, 79, 84

Eastern Center - 87, 91, 93, 95

Western Region - 111, 118, 121, 116

Using one-way ANOVA, analyse if there is any noticeable difference in sales among the Regions.

- Compute the Grand Mean of Sales based on the given sales figures for the month.
- Based only on the group averages, is there likelihood of significant difference between the sales of the regions? Justify your answer.

Northern Center - 100, 105, 110, 108  
 Southern Center - 82, 86, 79, 84  
 Eastern Center - 87, 91, 93, 95  
 Western Region - 111, 118, 121, 116

Soln : Grand Total of sales of all the four regions

$$\begin{aligned}
 &= 100 + 105 + 110 + 108 + 82 + 86 + 79 + 84 + 87 + 91 + 93 + 95 + 111 \\
 &\quad + 118 + 121 + 116 = 1586
 \end{aligned}$$

Number of observations = 16

$$\begin{aligned}
 \text{Hence, Grand Mean} &= \text{Total Sales} / \text{Number of observations} \\
 &= 1586/16 \approx 99.13
 \end{aligned}$$

(b) Group mean sales of the different Regional Centers is :

$$\text{Northern Center} - (100 + 105 + 110 + 108)/4 = 423/4 = 105.75$$

$$\text{Southern Center} - (82 + 86 + 79 + 84)/4 = 331/4 = 82.75$$

$$\text{Eastern Center} - (87 + 91 + 93 + 95)/4 = 366/4 = 91.25$$

$$\text{Western Region} - (111 + 118 + 121 + 116)/4 = 466/4 = 116.5$$

The difference between group means is quite large. West average (Rs. 116.5 Lakhs) is much higher & South (Rs. 82.75 L) is much lower than the grand mean (Rs. 99.13 L). Therefore, it is likely that ANOVA would detect a statistically significant difference between these groups.

---

**Ex 5 :** An experiment with Completely Randomized Design was conducted to study the effect of three different fertilizers –  $M_1$ ,  $M_2$  and  $M_3$  on crop yield in 3 homogeneous plots. Each fertilizer was randomly applied to three homogeneous plots. The yields (in kg) from each treatment are given below –

$M_1$  : 7.75, 8.25, 8

$M_2$  : 9.5, 9.25, 9.75

$M_3$  : 7.25, 7, 7.5

(a) Calculate the grand total of all the observations.

(b) Using treatment totals, compute the sum of squares due to treatment (SST).

$M_1 : 7.75, 8.25, 8$

$M_2 : 9.5, 9.25, 9.75$

$M_3 : 7.25, 7, 7.5$



Soln : (a) Total yield due to  $M_1 = 7.75 + 8.25 + 8 = 24$

Total yield due to  $M_2 = 9.5 + 9.25 + 9.75 = 28.5$

Total yield due to  $M_3 = 7.25 + 7 + 7.5 = 21.75$

Grand Total (G) =  $24 + 28.5 + 21.75 = 74.25$

(b)  $SST = (24^2 + 28.5^2 + 21.75^2)/3 - G^2/3^2$

$$\approx 620.44 - 74.25^2/9 = 620.44 - 612.56 = 7.88$$

**Ex 6 :** In LSD, blocking (local control principle) is applied in two perpendicular directions. The output times of a program for a common problem written by 3 different Programmers - Krishna, Rani, Sathya (Rows) using 3 different Softwares - Python, R, SAS (Columns) executed in three different Brands of Laptops - Apple, Lenovo, Acer are given in the following design.

		Softwares →	Python	R	SAS
		Programmers ↓	Lenovo	Apple	Acer
Krishna		20	24	22	
Rani	Apple		Acer	Lenovo	
		23	26	28	
Sathya	Acer		Lenovo	Apple	
		20	27	29	

**Note :**  
 Softwares in violet  
 Programmers in Blue  
 Computer Brands in Black  
 Output time in Red

- Compute the grand total and verify the total number of observations (N).
- Based on the software-wise totals, which software appears to have the lowest average execution time? Show your reasoning.

Programmers ↓	Softwares →	Python	R	SAS
Krishna	Lenovo	Apple	Acer	
	20	24	22	
Rani	Apple	Acer	Lenovo	
	23	26	28	
Sathya	Acer	Lenovo	Apple	
	20	27	29	



**Soln :** (a) Total =  $20 + 24 + 22 + 23 + 26 + 28 + 20 + 27 + 29 = 219$

Total Number of observations (N) = 3×3 = 9

$$\text{Python} = 20 + 23 + 20 = 63 \quad \longrightarrow \quad 63/3 = 21.0$$

$$R = 24 + 26 + 27 = 77 \quad \longrightarrow \quad 77/3 \approx 25.67$$

$$\text{SAS} = 22 + 28 + 29 = 79 \quad \longrightarrow \quad 79/3 \approx 26.33$$

Hence, Python takes the lowest average execution time ie. it is the fastest.

**Ex 7 :** Perform ANOVA of the following LSD.

An experiment was carried out to determine the effect of claying the ground on the field of barley grains; amount of clay used were as follows –

A : No clay.

B : Clay at 100 per acre.

C : Clay at 200 per acre.

D : Clay at 300 per acre.

The yields were in plots of 8 meters by 8 meters and are given in the table.

D	B	C	A
<b>29.1</b>	<b>18.9</b>	<b>29.4</b>	<b>5.7</b>
C	A	D	B
<b>16.4</b>	<b>10.2</b>	<b>21.2</b>	<b>19.1</b>
A	D	B	C
<b>5.4</b>	<b>38.8</b>	<b>24.0</b>	<b>37.0</b>
B	C	A	D
<b>24.9</b>	<b>41.7</b>	<b>9.5</b>	<b>28.9</b>

**Note :** Figures in Bold Blue shows the yield of the plots.

A, B, C & D are plots of grounds.

Soln :

$H_{0r}$  : Row means are all equal - vs-  $H_{1r}$  : Row means are not all equal.

$H_{0c}$  : Column means are all equal - vs-  $H_{1c}$  : Column means are not all equal.

$H_{0t}$  : Treatment means are all equal - vs-  $H_{1t}$  : Treatment means are not all equal.

	I	II	III	IV	Row Totals ( $R_i$ )
I	D 29.1	B 18.9	C 29.4	A 5.7	83.1
II	C 16.4	A 10.2	D 21.2	B 19.1	66.9
III	A 5.4	D 38.8	B 24	C 37	105.2
IV	B 24.9	C 41.7	A 9.5	D 28.9	105
Column Totals ( $C_j$ )	75.8	109.6	84.1	90.7	360.2 = G

	I	II	III	IV	Row Totals (R <sub>i</sub> )
I	D 29.1	B 18.9	C 29.4	A 5.7	83.1
II	C 16.4	A 10.2	D 21.2	B 19.1	66.9
III	A 5.4	D 38.8	B 24	C 37	105.2
IV	B 24.9	C 41.7	A 9.5	D 28.9	105
Column Totals (C <sub>j</sub> )	75.8	109.6	84.1	90.7	360.2 = G

The four treatment totals are –

$$A : 30.8$$

$$(5.7 + 10.2 + 5.4 + 9.5 = 30.8)$$

$$B : 86.9$$

$$(18.9 + 19.1 + 24 + 24.9 = 86.9)$$

$$C : 124.5$$

$$(29.4 + 16.4 + 37 + 41.7 = 124.5)$$

$$D : 118.0$$

$$(29.1 + 21.2 + 38.8 + 28.9 = 118)$$

$$\text{Grand Total } G = 360.2, \quad N = m^2 = 16; \quad CF = G^2/N = 360.2^2/16 = 8109.0025$$

$$\text{Raw SS (RSS)} = \sum (y_{ijk})^2 = 29.1^2 + 18.9^2 + \dots + 9.5^2 + 28.9^2 = 10,052.08$$

$$\text{Total SS (TSS)} = RSS - CF = 10,052.08 - 8109.0025 = 1943.0775$$

$$\text{SSR} = \sum (R_i)^2/m - CF = (83.1^2 + 66.9^2 + 105.2^2 + 105^2)/4 - 8109.0025 = 259.3125$$

$$\text{SSC} = \sum (C_j)^2/m - CF = (75.8^2 + 109.6^2 + 84.1^2 + 90.7^2)/4 - 8109.0025 = 155.2725$$

$$\text{SST} = \sum (T_k)^2/m - CF = (30.8^2 + 86.9^2 + 124.5^2 + 118^2)/4 - 8109.0025 = 1372.1225$$

$$\text{Error SS} = TSS - SSR - SSC - SST$$

$$= 1943.0775 - 259.3125 - 155.2725 - 1372.1225 = 156.37 \quad 16$$

SSR =  $\sum(R_i)^2/m - CF = (83.1^2 + 66.9^2 + 105.2^2 + 105^2)/4 - 8109.0025 = 259.3125$

SSC =  $\sum(C_j)^2/m - CF = (75.8^2 + 109.6^2 + 84.1^2 + 90.7^2)/4 - 8109.0025 = 155.2725$

SST =  $\sum(T_k)^2/m - CF = (30.8^2 + 86.9^2 + 124.5^2 + 118^2)/4 - 8109.0025 = 1372.1225$

Error SS = TSS - SSR - SSC - SST  
= 1943.0775 - 259.3125 - 155.2725 - 1372.1225 = 156.37

ANOVA Table for LSD

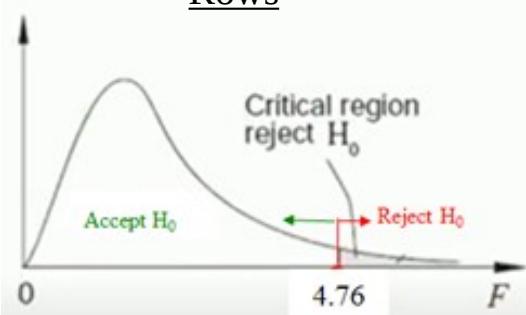
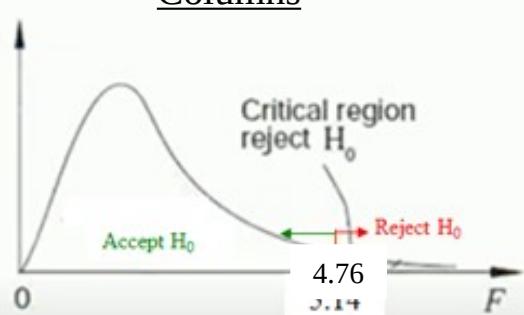
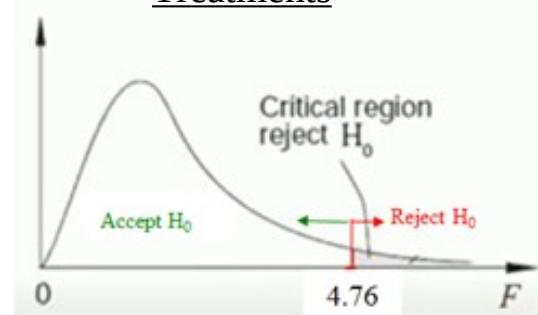
Source of Variation	DF	SS	MSS = SS/DF	F <sub>cal</sub> = MSS/MSE
Rows	m - 1 = 3	259.3125	86.4375	FR = 3.316
Columns	m - 1 = 3	155.2725	51.7575	FC = 1.986
Treatments	m - 1 = 3	1372.123	457.3742	FT = 17.549
Error	(m - 1)(m - 2) = 6 = 15 - 9	156.37	26.0616	
Total	m <sup>2</sup> - 1 = 15	1943.0775		Error DF = m <sup>2</sup> - 1 - {(m - 1) + (m - 1) + (m - 1)} = (m + 1)(m - 1) - 3(m - 1) = (m - 1)(m - 2)

F<sub>critical</sub> value obtained from F-table for  $\alpha = 0.05$  and df 3, 6 ie. F<sub>0.05, 3, 6</sub> = 4.76.

Now, comparing the F<sub>cal</sub> with F<sub>critical</sub> we conclude that the null hypothesis for rows and columns can not be rejected whereas for treatments null hypothesis is rejected ie. different levels of clay have significant effect on the yield.

ANOVA Table for LSD

Source of Variation	DF	SS	MSS = SS/DF	$F_{cal}$ = MSS/MSE
Rows	$m - 1 = 3$	259.3125	86.4375	$FR = 3.316$
Columns	$m - 1 = 3$	155.2725	51.7575	$FC = 1.986$
Treatments	$m - 1 = 3$	1372.123	457.3742	$FT = 17.549$
Error	$(m - 1)(m - 2) = 6$	156.37	26.0616	
Total	$m^2 - 1 = 15$	1943.0775		

Rows

Columns

Treatments


$$F_{cal} (3.316) < F_{critical} (4.76)$$

So, Null hypothesis can not be rejected.

$$F_{cal} (1.986) < F_{critical} (4.76)$$

So, Null hypothesis can not be rejected.

$$F_{cal} (17.549) > F_{critical} (4.76)$$

So, Null hypothesis can not be accepted.

## Ex 8 : Analyze the following BIBD

Balanced Incomplete Block Design for Catalyst Experiment

Treatment (Catalyst)	Block (Batch of Raw Material)			
	1	2	3	4
1	73	74	-	71
2	-	75	67	72
3	73	75	68	-
4	75	-	72	75

B1	B2	B3	B4
1	1	2	1
3	2	3	2
4	3	4	4

$H_{0t}$  : All treatment effects are equal -vs-  $H_{at}$  : All treatment effects are not equal

$H_{0b}$  : All block effects are equal -vs-  $H_{ab}$  : All block effects are not equal

Soln :

Treatment (v) = 4, Block (b) = 4, Each treatment occurs in blocks (r) = 3, Block size k < v = 3, Each pair of treatment occurring ( $\lambda$ ) = 2

Balanced Incomplete Block Design for Catalyst Experiment

Treatment (Catalyst)	Block (Batch of Raw Material)				
	1	2	3	4	$y_{i.}$
1	73	74	-	71	218
2	-	75	67	72	214
3	73	75	68	-	216
4	75	-	72	75	222
$y_{.j}$	221	224	207	218	870 = $y_{..}$

Treatment (Catalyst)	Block (Batch of Raw Material)				
	1	2	3	4	y <sub>i..</sub>
1	73	74	-	71	218
2	-	75	67	72	214
3	73	75	68	-	216
4	75	-	72	75	222
y <sub>.j</sub>	221	224	207	218	870 = y <sub>..</sub>

Treatment (v) = 4, Block (b) = 4, Each treatment occurs in blocks (r) = 3, Block size k < v = 3, Each pair of treatment occurring ( $\lambda$ ) = 2

This is a BIBD with  $v = a = 4$ ,  $b = 4$ ,  $k = 3$ ,  $r = 3$ ,  $\lambda = 2$  and  $N = 12$ . The analysis of this data is as follows. The Total Sum of Squares is –

$$\begin{aligned} SST &= \sum (obsn)^2 - G^2/N = \sum_i \sum_j (y_{ij})^2 - (y_{..})^2/12 \\ &= (73^2 + 74^2 + \dots + 75^2) - 870^2/12 = 63156 - 63075 = 81 \end{aligned}$$

$$\begin{aligned} SSBlocks &= 1/3 \left[ \sum_{j=1}^4 (y_{.j})^2 \right] - G^2/N = 1/3 [221^2 + 224^2 + 207^2 + 218^2] - 870^2/12 \\ &= 63125 - 63070 = 55 \end{aligned}$$

$$Q_1 = T_1 - 1/3(B_1 + B_2 + B_4) = 218 - 1/3(221 + 224 + 218) = -9/3 = -3$$

$$Q_2 = T_2 - 1/3(B_2 + B_3 + B_4) = 214 - 1/3(224 + 207 + 218) = -7/3$$

$$Q_3 = T_3 - 1/3(B_1 + B_2 + B_3) = 216 - 1/3(221 + 224 + 207) = -4/3$$

$$Q_4 = T_4 - 1/3(B_1 + B_3 + B_4) = 222 - 1/3(221 + 207 + 218) = 20/3$$

This is a BIBD with  $v = a = 4$ ,  $b = 4$ ,  $k = 3$ ,  $r = 3$ ,  $\lambda = 2$  and  $N = 12$ . The analysis of this data is as follows. The Total Sum of Squares is –

$$\begin{aligned} SST &= \sum (obsn)^2 - G^2/N = \sum_i \sum_j (y_{ij})^2 - (y_{..})^2/12 \\ &= (73^2 + 74^2 + \dots + 75^2) - 870^2/12 = 63156 - 63075 = 81 \end{aligned}$$

$$\begin{aligned} SSBLOCKS &= 1/3 \left[ \sum_{j=1}^4 (y_{.j})^2 \right] - G^2/N = 1/3 [221^2 + 224^2 + 207^2 + 218^2] - 870^2/12 \\ &= 63125 - 63070 = 55 \end{aligned}$$

$$Q_1 = T_1 - 1/3(B_1 + B_2 + B_4) = 218 - 1/3(221 + 224 + 218) = -9/3 = -3$$

$$Q_2 = T_2 - 1/3(B_2 + B_3 + B_4) = 214 - 1/3(224 + 207 + 218) = -7/3$$

$$Q_3 = T_3 - 1/3(B_1 + B_2 + B_3) = 216 - 1/3(221 + 224 + 207) = -4/3$$

$$Q_4 = T_4 - 1/3(B_1 + B_3 + B_4) = 222 - 1/3(221 + 207 + 218) = 20/3$$

Treatment (Catalyst)	Block (Batch of Raw Material)				
	1	2	3	4	$y_{..}$
1	73	74	-	71	218
2	-	75	67	72	214
3	73	75	68	-	216
4	75	-	72	75	222
$y_{..}$	221	224	207	218	870 = $y_{..}$

$$\begin{aligned} SSTreatment\ (adjusted) &= SSTreat_{(adj)} = [k \sum_{i=1}^4 (Q_i)^2] / \lambda a = \sum \hat{\tau}_i Q_i \\ &= 3[9-9/3)2 + (-7/3)2 + (-4/3)2 + (20/3)2] / (2 \times 4) = 22.75 \end{aligned}$$

The error sum of squares is obtained by subtraction as -

$$SSE = SST - SSTreatment_{(adj)} - SSBLOCKS = 81 - 22.75 - 55 = 3.25$$

$$Q'_1 = B_1 - 1/3(T_1 + T_3 + T_4) = 221 - 1/3(218 + 216 + 222) = 7/3$$

$$Q'_2 = B_2 - 1/3(T_1 + T_2 + T_3) = 224 - 1/3(218 + 214 + 216) = 24/3$$

$$Q'_3 = B_3 - 1/3(T_2 + T_3 + T_4) = 207 - 1/3(214 + 216 + 222) = -31/3$$

$$Q'_4 = B_4 - 1/3(T_1 + T_2 + T_4) = 218 - 1/3(218 + 214 + 222) = 0$$

$$SS_{\text{Blocks}_{(\text{adj.})}} = [r \sum (Q'_i)^2] / \lambda b = 3[(7/3)^2 + (24/3)^2 + (-31/3)^2 + (0)^2] / (2 \times 4) = 66.08$$

$$SST_{\text{Treat.}} = \sum (T_i)^2 / r - G^2 / N = [(218)^2 + (214)^2 + (216)^2 + (222)^2] / 3 - (870)^2 / 12 \\ = 11.67$$

#### ANOVA including both Treatments & Blocks

Source of Variation	DF	SS	MS	$F_{\text{calculated}}$	$F_{\text{critical}}$
SSTreatments (adj.)	3	22.75	22.75/3 = 7.58	7.58/0.65 = 11.66	5.41
SSTreatments (unadj.)	3	11.67			
SSBlocks (unadj.)	3	55.00			
SSBlocks (adj.)	3	66.08	66.08/3 = 22.03	22.03/0.65 = 33.89	5.41
SSError	11 - 6 = 5	3.25	3.25/5 = 0.65		
SSTotal	11	81.00			

Value of  $F_{3,5}$  for 5% LOS, obtained from F-table, is 5.41 ie.  $F_{\text{critical}} = 5.41$

Since both Treatments & Blocks  $F_{\text{calculated}} (11.66) \& (33.89) > F_{\text{critical}} (5.41)$ ,  
 Reject both  $H_{0t}$  and  $H_{0b}$

**Ex 9 :** The following table gives the synthetic yields per plot of an experiment conducted with  $3^2 = 9$  treatments using a simple lattice (2 replication) design.

Replication 1				Replication 2			
Blocks	Treatments (yield per plot)			Blocks	Treatments (yield per plot)		
1	1 (8)	7 (5)	4 (3)	4	8 (2)	7 (2)	9 (7)
2	3 (3)	6 (2)	9 (6)	5	4 (3)	5 (3)	6 (3)
3	8 (3)	5 (7)	2 (3)	6	2 (2)	3 (4)	1 (6)

Analyze the data. LOS 5%

Soln :

$$\text{Grand Total, } G = 8 + 5 + \dots + 6 = 72$$

$$\text{No. of observations, } n = 2s^2 = 2 \times 3^2 = 18$$

Grand Mean,  $\bar{y} = G/n = 72/18 = 4$ , No. of replications = 2 (as it is a simple lattice)

Block size,  $k = 3$ , Correction Factor (CF) =  $G^2/n = 72^2/18 = 288$

Replication 1				Replication 2			
Blocks	Treatments (yield per plot)			Blocks	Treatments (yield per plot)		
1	1 (8)	7 (5)	4 (3)	4	8 (2)	7 (2)	9 (7)
2	3 (3)	6 (2)	9 (6)	5	4 (3)	5 (3)	6 (3)
3	8 (3)	5 (7)	2 (3)	6	2 (2)	3 (4)	1 (6)

Note :  $T_1 = 8 + 6 = 14$ , etc.

$$B_1 = 8 + 5 + 3 = 16$$

Total of Block, in which treatment 1 occurs,

$$\sum_{j(1)} B_j = (8 + 5 + 3) + (2 + 4 + 6) = 16 + 12 = 28, \text{ etc.}$$

$$\text{Adjusted Treatment Total, } Q_1 = T_1 - \sum_{j(1)} B_j / k = 14 - 28/3 = 4.67$$

$$\hat{\tau}_1 = Q_1/2 + [S_R(Q_1) + S_C(Q_1)]/(2 \times 3) = 4.67/2 + 1.001/6 = 2.50$$

Replication 1				Replication 2			
Blocks	Treatments (yield per plot)			Blocks	Treatments (yield per plot)		
1	1 (8)	7 (5)	4 (3)	4	8 (2)	7 (2)	9 (7)
2	3 (3)	6 (2)	9 (6)	5	4 (3)	5 (3)	6 (3)
3	8 (3)	5 (7)	2 (3)	6	2 (2)	3 (4)	1 (6)

$$T_2 = 3 + 2 = 5, T_6 = 2 + 3 = 5, T_9 = 6 + 7 = 13$$

$$B_2 = 3 + 2 + 6 = 11, B_4 = 2 + 2 + 7 = 11, B_6 = 2 + 4 + 6 = 12$$

$$\sum_{j=3} B_j = 3 + 2 + 6 + 2 + 4 + 6 = 23; \sum_{j=6} B_j = 3 + 2 + 6 + 3 + 3 + 3 = 20$$

Blocks	Treatments (yield per plot)			Blocks	Treatments (yield per plot)		
1	1 (8)	7 (5)	4 (3)	4	8 (2)	7 (2)	9 (7) 
2	3 (3)	6 (2)	9 (6)	5	4 (3)	5 (3)	6 (3)
3	8 (3)	5 (7)	2 (3)	6	2 (2)	3 (4)	1 (6)

Specimen examples :  $T_1 = 8 + 6 = 14, \dots, T_5 = 7 + 3 = 10, \dots, T_9 = 6 + 7 = 13$

$B_1 = 8 + 5 + 3 = 16, \dots, B_4 = 2 + 2 + 7 = 11, \dots, B_6 = 2 + 4 + 6 = 12$

$\sum_{j(1)} B_j = (8 + 5 + 3) + (2 + 4 + 6) = 16 + 12 = 28$ , (Total of Block, in which treatment 1 occurs) ...,

$\sum_{j(5)} B_j = (8 + 5 + 3) + (2 + 4 + 6) = 16 + 12 = 28$  etc.

$$\hat{\tau}_1 = Q_1/2 + [S_R(Q_1) + S_C(Q_1)]/(2 \times 3) = 4.67/2 + 1.001/6 = 2.50$$

Adjusted treatment mean for treatment  $i = i^{\text{th}}$  treatment effect ( $\hat{\tau}_i$ ) + Grand Mean 

Grand Mean,  $\bar{y} = G/n = 72/18 = 4$

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treat/ Block No.	$T_i$	$B_j$	Block Nos. in which treat $i$ occurs	$\sum_{j(i)} B_j$	$\sum_{j(i)} B_j/k$	$Q_i$ (2) - (6)	$\hat{\tau}_i$	Adj. treat. Mean
1	14	16	1, 6	28	9.33	4.66	2.50	6.50
2	5	11	3, 6	25	8.33	-3.33	-2.16	1.84
3	7	13	2, 6	23	7.66	-0.66	0.33	4.33
4	6	11	1, 5	25	8.33	-2.66	-1.33	2.66
5	10	9	3, 5	22	7.33	2.66	0.50	4.50
6	5	12	2, 5	20	6.66	-1.66	-0.50	3.50
7	7		1, 4	27	9.00	-2.00	-0.83	3.16
8	5		3, 4	24	8.00	-3.00	-2.00	2.00
9	13		2, 4	22	7.33	5.66	3.50	7.50

Replication 1			Replication 2		
Blocks	Treatments (yield per plot)		Blocks	Treatments (yield per plot)	
1	1 (8)	7 (5)	4	8 (2)	7 (2)
2	3 (3)	6 (2)	5	4 (3)	5 (3)
3	8 (3)	5 (7)	6	2 (2)	3 (4)
					1 (6)

Total Sum of Squares (TSS)  $= \sum (\text{observation})^2 - \text{CF}$

$$= 8^2 + 5^2 + \dots + 6^2 - 72^2/18 = 66$$

Treatment Sum of Squares unadjusted ( $SST_U$ )  $= [\sum (T_i)^2]/m - \text{CF}$

$$= (14^2 + \dots + 13^2)/2 - 72^2/18 = 49$$

Block Sum of Squares unadjusted ( $SSB_U$ )  $= [\sum (B_j)^2]/s - \text{CF}$

$$= (16^2 + \dots + 12^2)/3 - 72^2/18 = 9.33$$

Treatment Sum of Squares adjusted ( $SST_A$ )  $= \sum \hat{\tau}_i Q_i = 51.44$

Block Sum of Squares adjusted ( $SSB_A$ )  $= SST_A + SSB_U - SST_U$

$$= 51.44 + 9.33 - 49 = 11.77$$

Error Sum of Squares (SSE)  $= TSS - SSB_U - SST_A$

$$= 66 - 51.44 - 9.33 = 5.23$$

ANOVA Table is shown in the next slide.

$= 8^2 + 5^2 + \dots + 6^2 - 72^2/18 = 66$

$= [\Sigma (T_i)^2]/m - \text{CF}$

$= (14^2 + \dots + 13^2)/2 - 72^2/18 = 49$

$= [\Sigma (B_j)^2]/s - \text{CF}$

$= (16^2 + \dots + 12^2)/3 - 72^2/18 = 9.33$

$= \Sigma \hat{\tau}_i Q_i = 51.44$

$= SST_A + SSB_U - SST_U$

$= 51.44 + 9.33 - 49 = 11.77$

$= TSS - SSB_U - SST_A$

$= 66 - 51.44 - 9.33 = 5.23$

ANOVA Table is shown in the next slide.

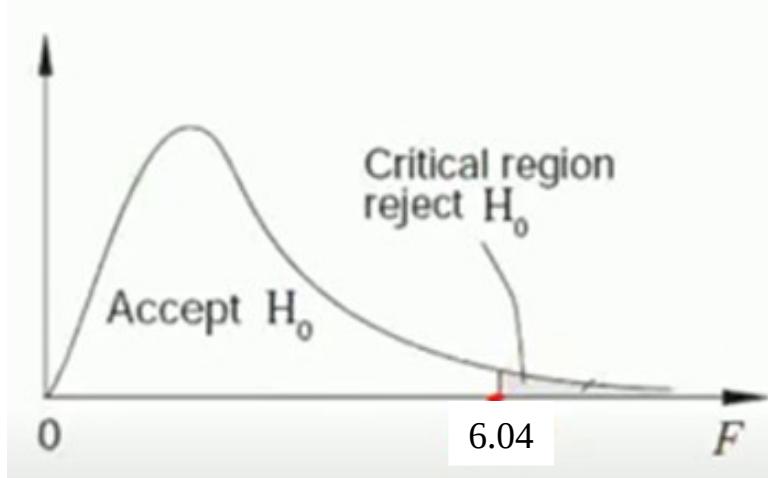
### ANOVA Table

Source of Variation	DF	SS	MS	F <sub>calculated</sub>	F <sub>critical</sub>
SSBlocks (unadj.)	$2 \times 3 - 1 = 5$	9.33			
SSTreatments (adj.)	$3^2 - 1 = 8$	51.44	6.43	4.91	6.04
SSBlocks (adj.)	$2 \times 3 - 1 = 5$	11.77	2.35	1.79	6.26
SSTreatments (unadj.)	$3^2 - 1 = 8$	49.00			
SSError	$17 - (5 + 8) = 4$	5.23	1.31		
SSTotal	$2 \times 3^2 - 1 = 17$				

F<sub>0.05,8,4</sub> and F<sub>0.05,5,4</sub> values obtained from F-table are 6.04 & 6.26 respectively.

As F<sub>calculated</sub> values < F<sub>critical</sub> value we conclude that Treatment effects and block effects are not significantly different.

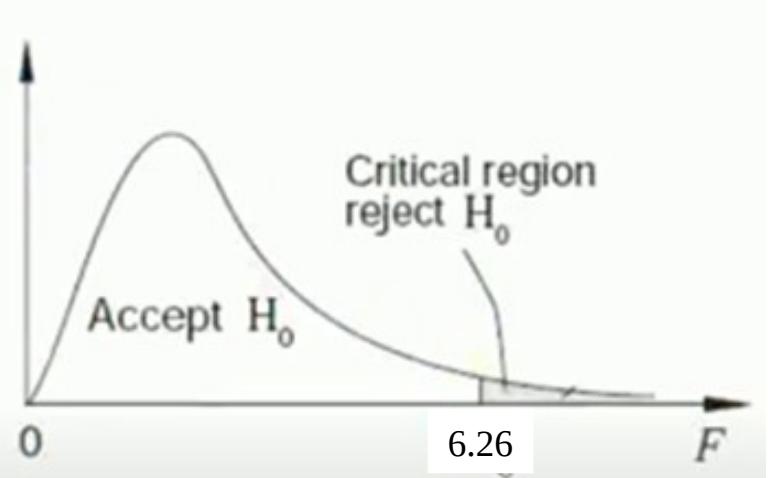
### For Treatment



$$F_{\text{cal}} (4.91) < F_{\text{cri}} (6.04)$$

Accept the Null Hypo

### For Blocks



$$F_{\text{cal}} (1.79) < F_{\text{cri}} (6.26)$$

Accept the Null Hypo

---

**Ex 10 :** A  $2^2$  experiment in six ( $Y = 6$ ) randomized blocks was conducted in order to obtain an idea of the interaction : spacing (s) x number of seedlings per hole (N) along with the effects of different types of spacing and different numbers of seedlings per hole, while adopting the Japanese method of cultivation.

The levels of two factors are –

$S : \begin{cases} 8'' \text{ spacings in between} \\ 10'' \text{ spacings in between} \end{cases}$

and

$N : \begin{cases} 3 \text{ seedlings per hole} \\ 4 \text{ seedlings per hole} \end{cases}$

The field plan and yield of dry Aman paddy (in kg.) are given in next slide.

Block 1				Ttl	Block 2				Ttl	Block 3				Ttl
(1)	s	ns	n		ns	(1)	s	n		(1)	n	s	ns	
117	106	109	114	446	114	120	117	114	465	111	117	114	106	448
Block 4					Block 5					Block 6				
ns	n	s	(1)	439	ns	s	(1)	n	283	n	(1)	ns	s	361
98	121	112	108		75	97	73	38		58	81	105	117	

Analyse the data to find out if there are any significant treatment effects – main or interaction.

We apply Yates' method to find the total effects.

(1)	(1)
a	s
b	n
ab	ns

Block 1				Ttl	Block 2				Ttl	Block 3				Ttl	innovate achieve lead		
(1)	s	ns	n		ns	(1)	s	n		(1)	n	s	ns				
117	106	109	114	446	114	120	117	114	465	111	117	114	106	448			

Block 4				Block 5				Block 6						
ns	n	s	(1)	ns	s	(1)	n	n	(1)	ns	s			
98	121	112	108	439	75	97	73	38	283	58	81	105	117	361

From slide 19

$$\begin{aligned}
 [S] &= -[1] + [s] - [n] + [sn] = [sn] - [n] + [s] - [1] \\
 [N] &= -[1] - [s] + [n] + [sn] = [sn] - [s] + [n] - [1] \\
 [NS] &= +[1] - [s] - [n] + [sn] = [sn] - [n] - [s] + [1]
 \end{aligned}$$

### Yates' Method for a 2<sup>2</sup> Experiment

Treatment combination (1)	Total yield from all blocks (2)	(3)	(4)	Average effect
(1)	[1] = 610	[1] + [n] = 1172	[1] + [n] + [s] + [ns] = 2442 = Grand total (G)	
n	[n] = 562	[s] + [ns] = 1270	[n] - [1] + [ns] - [s] = -104 = [N]	[N]/2x6
s	[s] = 663	[n] - [1] = -48	[s] + [ns] - [1] - [n] = 98 = [S]	[S]/2x6
ns	[ns] = 607	[ns] - [s] = -56	[ns] - [s] - [n] + [1] = -8 = [NS]	[NS]/2x6

Now perform the randomized block analysis –

Treatment totals are : [1] = 117 + 120 + 111 + 108 + 73 + 81 = 610

[s] = 106 + 117 + 114 + 112 + 97 + 117 = 663; [n] = 114 + 114 + 117 + 121 + 38 + 58 = 562

and [ns] = 109 + 114 + 106 + 98 + 75 + 105 = 607; G = 610 + 663 + 562 + 607 = 2442

[S] = 607 - 562 + 663 - 610 = 98; [N] = 607 - 663 + 562 - 610 = -104

[NS] = 607 - 562 - 663 + 610 = -8

(1) (1)  
a s  
b n  
ab ns



Block 1				Ttl	Block 2				Ttl	Block 3				Ttl
(1)	s	ns	n		ns	(1)	s	n		(1)	n	s	ns	
117	106	109	114	446	114	120	117	114	465	111	117	114	106	448
Block 4					Block 5					Block 6				
ns	n	s	(1)		ns	s	(1)	n		n	(1)	ns	s	
98	121	112	108	439	75	97	73	38	283	58	81	105	117	361

$$\text{Grand Total (G)} = [1] + [n] + [s] + [\text{ns}] = 610 + 663 + 562 + 607 = 2442$$

Six block totals are : 446, 465, 448, 439, 283 and 361.

$$\text{Raw total SS (RSS)} = \sum y^2 = 117^2 + 106^2 + \dots + 117^2 = 2,59,024$$

$$\text{Correction Factor (CF)} = G^2/N = (2442)^2/24 = 2,48,473.5$$

$$\text{Total SS (TSS)} = \text{RSS} - \text{CF} = 2,59,024 - 2,48,473.5 = 10,550.5$$

$$\begin{aligned} \text{Block SS (BSS)} &= (446^2 + 465^2 + \dots + 361^2)/4 - \text{CF} \\ &= 2,54,744 - 2,48,473.5 = 6270.5 \end{aligned}$$

$$\begin{aligned} \text{Treatment SS (TrSS)} &= \{(610^2 + \dots + 607^2)/6\} - 2,48,473.5 \\ &= 1495962/6 - 2,48,473.5 = 853.5 \end{aligned}$$

$$\text{Error SS} = \text{TSS} - \text{BSS} - \text{TrSS} = 10,550.5 - 6,270.5 - 853.5 = 3,426.5$$

Here, number of Blocks or replicates is 6 and number of treatment combinations is 4. Since, each block contains 4 treatment combinations then each block total square should be divided by 4 while finding block sum of squares. Similarly, each treatment combination replicates as many times as the number of Blocks i.e, 6 then each treatment combination total square should be divided by 6 while finding treatment sum of squares.

From slide 25

$$[S] = 607 - 562 + 663 - 610 = 98; [N] = 607 - 663 + 562 - 610 = -104$$

$$[NS] = 607 - 562 - 663 + 610 = -8$$

From slide 19

$$\text{SS due to main effect of A} = [A]^2/4r \text{ with } 1 \text{ d.f.}$$

$$\text{SS due to main effect of B} = [B]^2/4r \text{ with } 1 \text{ d.f.}$$

$$\text{SS due to interaction effect of AB} = [AB]^2/4r \text{ with } 1 \text{ d.f.}$$

Also, SS due to N =  $[N]^2/4r = (-104)^2/24 = 450.667$

$$\text{SS due to S} = [S]^2/4r = (98)^2/24 = 400.167 \quad (1) \quad (1)$$

$$\text{SS due to NS} = [NS]^2/4r = (-8)^2/24 = 2.67 \quad \begin{matrix} a & s \\ b & n \\ ab & ns \end{matrix}$$

ANOVA Table is presented in next slide

$$\text{Total SS (TSS)} = \text{RSS} - \text{CF} = 2,59,024 - 2,48,473.5 = 10,550.5$$

$$\begin{aligned}\text{Block SS (BSS)} &= (446^2 + 465^2 + \dots + 361^2)/4 - \text{CF} \\ &= 2,54,744 - 2,48,473.5 = 6270.5\end{aligned}$$

$$\text{SS due to N} = [N]^2/4r = (-104)^2/24 = 450.667$$

$$\text{SS due to S} = [S]^2/4r = (98)^2/24 = 400.167$$

$$\text{SS due to NS} = [NS]^2/4r = (-8)^2/24 = 2.67$$

ANOVA Table for  $2^2$  experiment

Source of variation	d.f.	SS	MS	$F_{\text{calculated}}$	$F_{\text{critical}}$
Blocks	5	6,270.50	1254.1		$F_{0.05,1,15} = 4.54$
N	1	450.667	450.667	1.973	
S	1	400.167	400.167	1.752	
NS	1	2.667	2.667	<1	
Error	$23 - 8 = 15$	3,426.50	228.433		
Total	$6 \times 4 - 1 = 23$	10,550.50			

There is no significant main or interaction effects present in the above experiment, as in each of the cases the computed value of F is less than the corresponding theoretical value (critical value) at the 5% level of significance.

**Ex 11 :** A  $3^2$  experiment was conducted to study the effects of the two factors – Nitrogen (N) & Phosphorus (P) each at 3 levels – 0, 1, 2. Two replications of nine plots each were used. The table shows the Response of the experiment.

Plan and Response of the $3^2$ experiment							
Replication	Treatment		Response	Replication	Treatment		Response
	n	p			n	p	
I	0	1	14	II	1	2	20
	2	0	15		1	0	19
	0	0	16		1	1	17
	2	1	15		0	0	18
	0	2	16		2	1	19
	1	2	18		0	1	16
	1	1	17		0	2	16
	1	0	19		2	2	19
	2	2	17		2	0	16

- Present the Treatment Combination Table
- Treatment Sum of Squares
- Replication Sum of Squares

(i)

Treatment Combinations Table

Replicat e	(1)	n	$n^2$	p	np	$n^2p$	$p^2$	$np^2$	$n^2p^2$	Total
	00	10	20	01	11	21	02	12	22	
I	16	19	15	14	17	15	16	18	17	147 (R <sub>1</sub> )
II	18	19	16	16	17	19	16	20	19	160 (R <sub>2</sub> )
Total	34	38	31	30	34	34	32	38	36	307 (G)
	(T <sub>1</sub> )	(T <sub>2</sub> )	(T <sub>3</sub> )	(T <sub>4</sub> )	(T <sub>5</sub> )	(T <sub>6</sub> )	(T <sub>7</sub> )	(T <sub>8</sub> )	(T <sub>9</sub> )	
	[1]	[n]	[n <sup>2</sup> ]	[p]	[np]	[n <sup>2</sup> p]	[p <sup>2</sup> ]	[np <sup>2</sup> ]	[n <sup>2</sup> p <sup>2</sup> ]	

 (ii) Treatment Sum of Squares (SST) = { $\sum$ (Treatment Total)<sup>2</sup>}/r - CF

 CF = G<sup>2</sup>/n; Here r = 2, G = 307 and n = r(3<sup>2</sup>) = 18

 Hence, CF = G<sup>2</sup>/n = 307<sup>2</sup>/18 = 5236

 Treatment SS (SST) = {Sum (Treatment Total)<sup>2</sup>}/r - CF

$$= (34^2 + 38^2 + \dots + 38^2 + 36^2)/2 - 5236.0556 = 5268.5 - 5236.0556 = 32.4444$$

 (iii) Replication SS (SSR) = {(R<sub>1</sub>)<sup>2</sup> + (R<sub>2</sub>)<sup>2</sup>}/9 - CF = (147<sup>2</sup> + 160<sup>2</sup>)/9 - 5236.0556

$$= 5245.4444 - 5236.0556 = 9.3888$$

**Ex 12 :** A variety-manurial experiment was conducted by allotting the three varieties  $V_1$ ,  $V_2$  and  $V_3$  at random to the plots of four randomized blocks and then, splitting each plot into four sub-plots, the four manures  $M_1$ ,  $M_2$ ,  $M_3$  and  $M_4$  were applied at random within each plot. The plan and yield are shown below. Analyse the data to find out if there are any effects due to manure or variety of interaction between variety and manure.

Variety	BLOCK			
	I	II	III	IV
$V_1$	609	450	488	545
$V_2$	920	870	833	1118
$V_3$	1067	1072	1093	905

Number of varieties (V) = 3

Number of manures (M) = 4

Number of replicates or (Blocks) = 4

$N = V \times M \times R = 3 \times 4 \times 4 = 48$

	V1	V3	V2	
	M1	M4	M2	
Block I	94	440	250	
	220	297	147	
	185	218	248	
	---	112	275	
Block II	V2	V1	V3	
	M1	135	M2	160
	M4	290	M4	---
	180	M3	124	M3
Block III	265	M1	71	M2
	V1	V2	V3	
	M1	78	M3	196
	M3	135	M4	262
Block IV	130	M1	155	M1
	145	M2	220	115
	---	M4	---	
	114	M4	323	M4

- Specify the missing data (indicated by red dash ---).
- Draw up the block-variety table
- Find the correction factor.

Variety	BLOCK			
	I	II	III	IV
V <sub>1</sub>	609	450	488	545
V <sub>2</sub>	920	870	833	1118
V <sub>3</sub>	1067	1072	1093	905

Soln :

	V1	V3	V2		V2	V1	V3					
	M1	94	M4	440	M2	250	M1	135	M2	160	M4	370
Block I	M3	220	M2	297	M1	147	M4	290	M4	95	M1	140
	M2	185	M3	218	M3	248	M3	180	M3	124	M3	340
	M4	110	M1	112	M4	275	M2	265	M1	71	M2	222

	V1	V2	V3		V1	V3	V2					
	M1	78	M3	196	M2	235	M1	81	M2	246	M3	296
Block III	M3	135	M4	262	M3	260	M2	175	M3	191	M2	260
	M4	130	M1	155	M1	115	M4	175	M1	145	M1	112
	M2	145	M2	220	M4	483	M3	114	M4	323	M4	450

As indicated above :

Missing numbers in Block I - V1 :  $609 - (94 + 220 + 185) = 110$ ;

Missing numbers in Block II – V1 :  $450 - (160 + 124 + 71) = 95$ ;

Missing numbers in Block III – V3 :  $1093 - (235 + 260 + 115) = 483$ ;

Missing numbers in Block IV - V1 :  $= 545 - (175 + 175 + 114) = 81$

### (ii) Block-variety table

Variety	BLOCK				Total
	I	II	III	IV	
V <sub>1</sub>	609	450	488	545	2092
V <sub>2</sub>	920	870	833	1118	3741
V <sub>3</sub>	1067	1072	1093	905	4137
<b>Total</b>	<b>2596</b>	<b>2392</b>	<b>2414</b>	<b>2568</b>	<b>9970</b>

---

$$(iii) CF = G^2/N = 9970^2/VMR = 9970^2/3 \times 4 \times 4 = 2,070,852.08333$$

Ex 13 : A machine drills hole in a pipe with a mean diameter of 0.532 cm and a standard deviation of 0.002 cm. Calculate the control limits for mean of samples 5.

Soln : Given  $\bar{x} = 0.532$ ,  $\sigma = 0.002$ ,  $n = 5$

Hence, control limit for  $\bar{x}$  chart is

$$UCL = \bar{x} + 3(\sigma/\sqrt{n}) = 0.532 + 3(0.002/\sqrt{5}) = 0.5347$$

$$CL = \bar{x} = 0.532$$

$$LCL = \bar{x} - 3(\sigma/\sqrt{n}) = 0.532 - 3(0.002/\sqrt{5}) = 0.5293$$

Ex 14 : The following data gives the readings for 6 samples of size 6 each in the production of a certain product. Find the control limits using mean chart. [Given for sample size  $n = 6$ ,  $A_2 = 0.483$ ],

Sample	1	2	3	4	5	6
Mean	300	342	351	319	326	333
Range	25	37	20	28	30	22

Soln :

Sample	1	2	3	4	5	6	Total
Mean	300	342	351	319	326	333	1971
Range	25	37	20	28	30	22	162

Hence, control limit for Mean chart is

$$\bar{\bar{x}} = (\sum \bar{x}) / \text{number of samples} = 1971/6 = 328.5;$$

$$\bar{R} = (\sum R) / \text{number of samples} = 162/6 = 27$$

$$UCL = \bar{\bar{x}} + A_2 \bar{R} = 328.5 + 0.483(27) = 341.54$$

$$CL = \bar{\bar{x}} = 328.5$$

$$LCL = \bar{\bar{x}} - A_2 \bar{R} = 328.5 - 0.483(27) = 315.45$$

**Ex 15 :** The data shows the sample mean and range for 10 samples for size 5 each. Find the control limits for mean chart and range chart and comment on the state of control of the process.

Sample	1	2	3	4	5	6	7	8	9	10
Mean	21	26	23	18	19	15	14	20	16	10
Range	5	6	9	7	4	6	8	9	4	7

Soln :

Sample	1	2	3	4	5	6	7	8	9	10	Total
Mean	21	26	23	18	19	15	14	20	16	10	182
Range	5	6	9	7	4	6	8	9	4	7	65

The control limits for Mean chart ( $\bar{x}$ -chart) is –

$$\bar{\bar{x}} = (\sum \bar{x}) / \text{number of samples} = 182/10 = 18.2$$

$$\bar{R} = (\sum R) / \text{number of samples} = 65/10 = 6.5$$

$$UCL = \bar{\bar{x}} + A_2 \bar{R} = 18.2 + 0.577(6.5) = 21.95$$

$$CL = \bar{\bar{x}} = 18.2$$

$$LCL = \bar{\bar{x}} - A_2 \bar{R} = 18.2 - 0.577(6.5) = 14.45$$

The control limits for Range chart is –

$$UCL = D_4 \bar{R} = 2.114(6.5) = 13.741$$

$$CL = \bar{R} = 6.5$$

$$LCL = D_3 \bar{R} = 0(6.5) = 0$$

### Control Limits for Mean chart

$$UCL = \bar{x} + A_2 \bar{R} = 18.2 + 0.577(6.5) = 21.95$$

$$CL = \bar{x} = 18.2$$

$$LCL = \bar{x} - A_2 \bar{R} = 18.2 - 0.577(6.5) = 14.45$$

### Control Limits for R chart

$$UCL = D_4 \bar{R} = 2.114(6.5) = 13.741$$

$$CL = \bar{R} = 6.5$$

$$LCL = D_3 \bar{R} = 0(6.5) = 0$$

lead

### Comparing the Given Sample Data with Control Limits of Mean Chart

Sample	1	2	3	4	5	6	7	8	9	10	Total
Mean	21	26	23	18	19	15	14	20	16	10	182
Range	5	6	9	7	4	6	8	9	4	7	65

It is observed that the Sample nos. 2 & 3 are above the UCL and Sample nos. 7 & 10 are below the LCL of Mean Chart. We know that Lack of control in Mean Chart indicates Machine Problems. Hence remedial action should be initiated to rectify the machine problem.

### Now, Comparing the Given Sample Data with Control Limits of R Chart

Sample	1	2	3	4	5	6	7	8	9	10	Total
Mean	21	26	23	18	19	15	14	20	16	10	182
Range	5	6	9	7	4	6	8	9	4	7	65

All the sample points are within the control limits of R Chart. We know that Lack of control in Range Chart indicates Workers Problems. Since there is no lack of control in Range Chart, we can infer that there is no problem due to workers.

**Ex 16 :** The Values of sample mean ( $\bar{x}$ ) and the range (R) for ten samples of size 5 each is tabulated below. Draw mean chart and comment on the state of control of the process.

Sample	1	2	3	4	5	6	7	8	9	10
Mean	43	49	37	44	45	37	51	46	43	47
Range	5	6	5	7	7	4	8	6	4	6

Given the following control chart constraint for :  $n = 5$ ,  $A_2 = 0.577$ ,  $D_3 = 0$  and  $D_4 = 2.114$ .

**Soln :**

Sample	1	2	3	4	5	6	7	8	9	10	Total
Mean	43	49	37	44	45	37	51	46	43	47	442
Range	5	6	5	7	7	4	8	6	4	6	58

$$\bar{\bar{x}} = (\sum \bar{x})/\text{number of samples} = 442/10 = 44.2$$

$$\bar{R} = (\sum R)/\text{number of samples} = 58/10 = 5.8$$

The control limits for mean chart in next slide –

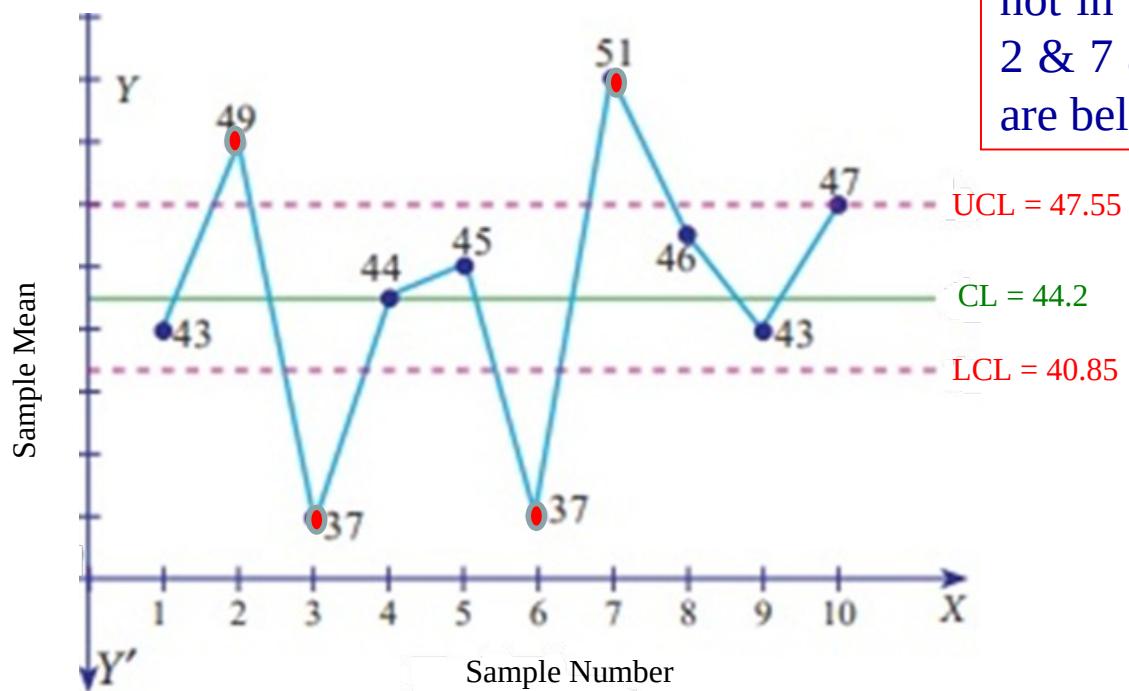
Sample	1	2	3	4	5	6	7	8	9	10	Total
Mean	43	49	37	44	45	37	51	46	43	47	442
Range	5	6	5	7	7	4	8	6	4	6	58

$$UCL = \bar{\bar{x}} + A_2 \bar{R} = 44.2 + 0.577(5.8) = 47.55$$

$$CL = \bar{\bar{x}} = 44.2$$

$$LCL = \bar{\bar{x}} - A_2 \bar{R} = 44.2 - 0.577(5.8) = 40.85$$

Control Chart



It is evident that the process is not in control - Sample numbers 2 & 7 are above UCL and 3 & 6 are below LCL.

## APPENDIX VI

### Factors for Constructing Variables Control Charts

Observations in Sample, $n$	Chart for Averages					Chart for Standard Deviations				Chart for Ranges						
	Factors for Control Limits			Factors for Center Line		Factors for Control Limits				Factors for Center Line		Factors for Control Limits				
	$A$	$A_2$	$A_3$	$c_4$	$1/c_4$	$B_3$	$B_4$	$B_5$	$B_6$	$d_2$	$1/d_2$	$d_3$	$D_1$	$D_2$	$D_3$	$D_4$
2	2.121	1.880	2.659	0.7979	1.2533	0	3.267	0	2.606	1.128	0.8865	0.853	0	3.686	0	3.267
3	1.732	1.023	1.954	0.8862	1.1284	0	2.568	0	2.276	1.693	0.5907	0.888	0	4.358	0	2.574
4	1.500	0.729	1.628	0.9213	1.0854	0	2.266	0	2.088	2.059	0.4857	0.880	0	4.698	0	2.282
5	1.342	0.577	1.427	0.9400	1.0638	0	2.089	0	1.964	2.326	0.4299	0.864	0	4.918	0	2.114
6	1.225	0.483	1.287	0.9515	1.0510	0.030	1.970	0.029	1.874	2.534	0.3946	0.848	0	5.078	0	2.004
7	1.134	0.419	1.182	0.9594	1.0423	0.118	1.882	0.113	1.806	2.704	0.3698	0.833	0.204	5.204	0.076	1.924
8	1.061	0.373	1.099	0.9650	1.0363	0.185	1.815	0.179	1.751	2.847	0.3512	0.820	0.388	5.306	0.136	1.864
9	1.000	0.337	1.032	0.9693	1.0317	0.239	1.761	0.232	1.707	2.970	0.3367	0.808	0.547	5.393	0.184	1.816
10	0.949	0.308	0.975	0.9727	1.0281	0.284	1.716	0.276	1.669	3.078	0.3249	0.797	0.687	5.469	0.223	1.777
11	0.905	0.285	0.927	0.9754	1.0252	0.321	1.679	0.313	1.637	3.173	0.3152	0.787	0.811	5.535	0.256	1.744
12	0.866	0.266	0.886	0.9776	1.0229	0.354	1.646	0.346	1.610	3.258	0.3069	0.778	0.922	5.594	0.283	1.717
13	0.832	0.249	0.850	0.9794	1.0210	0.382	1.618	0.374	1.585	3.336	0.2998	0.770	1.025	5.647	0.307	1.693
14	0.802	0.235	0.817	0.9810	1.0194	0.406	1.594	0.399	1.563	3.407	0.2935	0.763	1.118	5.696	0.328	1.672
15	0.775	0.223	0.789	0.9823	1.0180	0.428	1.572	0.421	1.544	3.472	0.2880	0.756	1.203	5.741	0.347	1.653
16	0.750	0.212	0.763	0.9835	1.0168	0.448	1.552	0.440	1.526	3.532	0.2831	0.750	1.282	5.782	0.363	1.637
17	0.728	0.203	0.739	0.9845	1.0157	0.466	1.534	0.458	1.511	3.588	0.2787	0.744	1.356	5.820	0.378	1.622
18	0.707	0.194	0.718	0.9854	1.0148	0.482	1.518	0.475	1.496	3.640	0.2747	0.739	1.424	5.856	0.391	1.608
19	0.688	0.187	0.698	0.9862	1.0140	0.497	1.503	0.490	1.483	3.689	0.2711	0.734	1.487	5.891	0.403	1.597
20	0.671	0.180	0.680	0.9869	1.0133	0.510	1.490	0.504	1.470	3.735	0.2677	0.729	1.549	5.921	0.415	1.585
21	0.655	0.173	0.663	0.9876	1.0126	0.523	1.477	0.516	1.459	3.778	0.2647	0.724	1.605	5.951	0.425	1.575
22	0.640	0.167	0.647	0.9882	1.0119	0.534	1.466	0.528	1.448	3.819	0.2618	0.720	1.659	5.979	0.434	1.566
23	0.626	0.162	0.633	0.9887	1.0114	0.545	1.455	0.539	1.438	3.858	0.2592	0.716	1.710	6.006	0.443	1.557
24	0.612	0.157	0.619	0.9892	1.0109	0.555	1.445	0.549	1.429	3.895	0.2567	0.712	1.759	6.031	0.451	1.548
25	0.600	0.153	0.606	0.9896	1.0105	0.565	1.435	0.559	1.420	3.931	0.2544	0.708	1.806	6.056	0.459	1.541

---

**ALL THE BEST**  
**for**  
**ENSUING COMPREHENSIVE**  
**EXAMS**

---

# Thank You