BITS Pilani

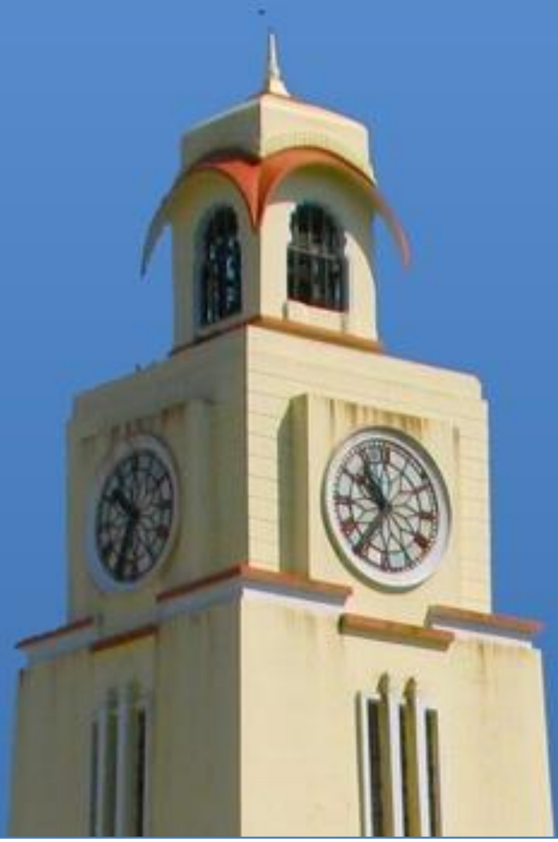
Pilani | Dubai | Goa | Hyderabad

# Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani



# Session 10

# (Maximum Likelihood Estimation & ANOVA)

## $12^{th}$ and $13^{th}$ Aug 2023

# Maximum Likelihood Estimation (MLE)

- Method of Maximum Likelihood Estimation is the best and most popular one among all methods to obtain an almost good or best estimator for a population parameter.
- It is a method of obtaining an estimator which most (maximum) likely estimates the true value of the parameter i.e., finding an estimator that can give most likely nearer value for the unknown true value of parameter.
- The corresponding estimator is called maximum likelihood estimator (MLE).

# Maximum Likelihood Estimation (MLE)

Suppose we have a random sample $x_1, x_2, .., x_n$ whose assumed probability distribution depends on some unknown parameter $\theta$.

Ex:

1) For Binomial unknown parameters are n, p.

2) For Poisson unknown parameter is $\lambda$.

Our goal is to find good estimator of $\theta$(population parameter) using sample and which can be done with the help of MLE.

# Maximum Likelihood function

- Let $x_1, x_2, .., x_n$ be i.i.d. random variables drawn from some probability distribution that depends on some unknown parameter θ.
- The goal of MLE to maximize likelihood function

$$L(\theta) = f(x_1, x_2, \ldots, x_n \mid \theta)$$
$$= f(x_1|\theta) f(x_2|\theta) \ldots f(x_n|\theta)$$

$$L(\theta) = \prod_{i=1}^{n} f(x_i / \theta)$$

# Maximum Likelihood Estimation (MLE)

- The maximum likelihood estimate (MLE) of θ is that value of θ that maximizes likelihood(θ).

It is defined as

$$L(\theta) = \prod_{i=1}^{n} f(x_i / \theta)$$

$$\log L(\theta) = \sum_{i=1}^{n} \log f(x_i / \theta)$$

For maximization, we have

$$\frac{dL}{d\theta} = 0 \quad ; \quad \frac{d^2L}{d\theta^2} < 0$$

Example: An unfair coin is flipped 100 times, and 61 heads are observed. The coin either has probability $\frac{1}{3}, \frac{1}{2}$, or $\frac{2}{3}$ of flipping a head each time it is flipped. Which of the three is the MLE?

Solution: Here the distribution is the binomial distribution with $n = 100$.

$$P\left(H = 61 \mid p = \frac{1}{3}\right) = \binom{100}{61}\left(\frac{1}{3}\right)^{61}\left(\frac{2}{3}\right)^{39} \approx 9.6 \times 10^{-9}$$

$$P\left(H = 61 \mid p = \frac{1}{2}\right) = \binom{100}{61}\left(\frac{1}{2}\right)^{61}\left(\frac{1}{2}\right)^{39} = 0.007$$

$$P\left(H = 61 \mid p = \frac{2}{3}\right) = \binom{100}{61}\left(\frac{2}{3}\right)^{61}\left(\frac{1}{3}\right)^{39} = 0.040$$

p.m.f.

$$\boxed{\begin{gathered} P(x) = \binom{n}{x} p^x (1-p)^{n-x} \\ 0 \le p \le 1 \\ x = 0,1,2,\dots,n; \end{gathered}}$$

Since $P\left(H = 61 \mid p = \frac{2}{3}\right)$ is maximum and hence MLE is $p = \frac{2}{3}$

**Example:** An unfair coin is flipped 100 times, and 61 heads are observed. What is the MLE when nothing is previously known about the coin?

**Solution:** Since the distribution follow is Binomial distribution, with parameter $p$.

Here $n = 100$. The likelihood function (MLE) is

$$P(H = 61|p) = \binom{100}{61} p^{61}(1-p)^{39}$$

**For maximization**

$$\frac{d}{dp} P(H = 61|p) = 0$$

$$\Rightarrow \binom{100}{61}\left[61p^{60}(1-p)^{39} - 39p^{61}(1-p)^{38}\right] = 0$$

$$\Rightarrow p^{60}(1-p)^{38}(61 - 100p) = 0$$

$$\Rightarrow p = 0, \frac{61}{100}, 1$$

Thus, the likelihoods are

$$P(H = 61|p = 0) = 0$$

$$P\left(H = 61\middle|p = \frac{61}{100}\right) = \binom{100}{61}\left(\frac{61}{100}\right)^{61}\left(\frac{39}{100}\right)^{39}$$

$$P(H = 61|p = 1) = 0$$

Since $P\left(H = 61\middle|p = \frac{61}{100}\right)$ is maximum and hence $p = \frac{61}{100}$ is the MLE.

# Maximum Likelihood for a Binomial distribution

- Suppose we wish to find the maximum likelihood estimate (MLE) of θ for a Binomial distribution,

$$p_k(k,\theta) = nC_k \theta^k (1-\theta)^{n-k}$$

$$\log p_k(k,\theta) = \log(nC_k) + k\log(\theta) + (n-k)\log((1-\theta)$$

$$\frac{\partial \log p_k(k,\theta)}{\partial \theta} = 0 \Rightarrow 0 + \frac{k}{\theta} - \frac{n-k}{1-\theta} = 0$$

$$k - k\theta = n\theta - k\theta \Rightarrow \theta = \frac{k}{n}$$

## Example

If 48 of 60 transceivers passed inspection. (a) Obtain the maximum likelihood estimate of the probability that a transceiver will pass inspection. (b) Obtain the maximum likelihood estimate that the next two transceivers tested will pass inspection.

Solution:

a) n=60, k=48, the MLE P(to pass test) = $\frac{48}{60}$=0.8

b) For next 2 consecutive test, the MLE to pass test = (0.8)(0.8)

## Example:

Consider a sample 0,1,0,0,1,0 from a binomial distribution, with the form P[X=0]=(1-p), P[X=1]=p. Find the maximum likelihood estimate of p.

Soln :

# Maximum Likelihood Estimation

## Example:

**Consider a sample 0,1,0,0,1,0 from a binomial distribution, with the form P[X=0]=(1-p), P[X=1]=p. Find the maximum likelihood estimate of p.**

Soln :

L(p)=P[X=0] P[X=1] P[X=0] P[X=0] P[X=1] P[X=0]

=(1-p) p (1-p) (1-p) p (1-p)

$=(1-p)^3p^2$.

Log L(p)=$\log[(1-p)^3p^2.]=\log[(1-p)^3]+\log[(p^{2)}.]=3\log(1-p)+2\log p$

$$\frac{\partial Log L(p)}{\partial p}=0 \qquad \frac{-3}{1-p}+\frac{2}{p}=0 \Rightarrow \frac{-3p+2-2p}{p(1-p)}=0 \Rightarrow p=2/5$$

That is , there is 1/3 chance to observe this sample if we believe the population to be Binomial distributed .

# MLE for Poisson Distribution Parameter

Let $X_1, X_2, \ldots, X_n \in R$ be a random sample from a Poisson distribution

The p.d.f. of a Poisson Distribution is :

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$ ; where x =0,1,2,...

The likelihood function is:

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = e^{-\lambda n} \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i}$$

The log-likelihood is:

$$lnL(\lambda) = -\lambda n + \sum_{i=1}^{n} xi.\, ln(\lambda) - ln(\prod_{i=1}^{n} xi)$$

Setting its derivative with respect to $\lambda$ to zero, we have:

$$\frac{d}{d\lambda} lnL(\lambda) = -n + \sum_{i=1}^{n} xi.\frac{1}{\lambda} = 0$$

giving,

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{X}$$

which is the maximum likelihood estimate

## Example:

In one area along the interstate, the number of dropped wireless phone connections per call follows a Poisson distribution. From four calls, the number of dropped connections is 2, 0, 3, 1

(a) Find the maximum likelihood estimate of λ.

(b) Obtain the maximum likelihood estimate that the next two calls will be completed without any accidental drops.

Solution: a) $\bar{x} = \lambda = \frac{2+0+3+1}{4} = 1.5$

b) For next one call, P(x=0) = $\frac{e^{-1.5}(1.5)^0}{0!} = e^{-1.5}$

For next two calls = P(X=0)P(X=0)= $(e^{-1.5})(e^{-1.5})$= $e^{-3}$

**EXAMPLE** **Maximum likelihood estimator of function of $\lambda$**

The number of defective hard drives produced daily by a production line can be modeled as a Poisson distribution. The counts for ten days are

$$7 \quad 3 \quad 1 \quad 2 \quad 4 \quad 1 \quad 2 \quad 3 \quad 1 \quad 2$$

Obtain the maximum likelihood estimate of the probability of 0 or 1 defectives on one day.

**Solution** From the maximum likelihood estimate of $\lambda$ is $\widehat{\lambda} = \overline{x} = 26/10 = 2.6$. Consequently, by the invariance property, the maximum likelihood estimate of

$$P(X = 0 \quad \text{or} \quad 1) = e^{-\lambda} + \frac{\lambda e^{-\lambda}}{1!}$$

is

$$e^{-\widehat{\lambda}} + \frac{\widehat{\lambda} e^{-\widehat{\lambda}}}{1!} = e^{-2.6} + \frac{2.6 \cdot e^{-2.6}}{1!} = 0.267$$

There will 1 or fewer defectives on just over one-quarter of the days. ■

The method of maximum likelihood applies to more than one parameter.

# Testing of Hypothesis ⇨ Analysis of Variance (ANOVA)

## Types of ANOVA

- One-way
- Two-way
- Three-way (LSD)
- Repeated measures of ANOVA (Rm ANOVA)

**Note:** In this unit we discuss only One-way ANOVA and Two ANOVA in detail

# Testing of Hypothesis ⇨ Why Analysis of Variance (ANOVA)

Student's t-test cannot be applied

⇩

No. of groups are more than two (say k) and are independent

⇩

If t-test is applied, the type-I error will increase

ANOVA Used to test equality of more than two population means against not equal

# Testing of Hypothesis ➡ One-way Analysis of Variance

**ANOVA** ➡ Testing equality of k group means against not equal

**ANOVA**

- Samples are drawn from normal population
- The population variances should be equal
- The sample size should be less than 30 (i.e., n < 30)
- Groups should be independent
- Subjects should be allocated randomly to both groups
- However even if sample size more than 30 (i.e., n > 30) ANOVA should be continue to apply, because of central limit theorem it approaches normal.

# One-way ANOVA

ANOVA is a statistical technique used to determine whether differences exist among three or more population means.

In one-way ANOVA the effect of one factor on the mean is tested. It is based on independent random samples drawn from $k$ – different levels of a factor, also called treatments.

The following notations are used in one-way ANOVA. The data can be represented in the following tabular structure.

## Data representation for one-way ANOVA

| Treatments | | | | | Total |
|---|---|---|---|---|---|
| Treatment 1 | $x_{11}$ | $x_{12}$ | $\cdots$ | $x_{1n_1}$ | $x_{1.}$ |
| Treatment 2 | $x_{21}$ | $x_{22}$ | $\cdots$ | $x_{2n_2}$ | $x_{2.}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\cdots$ | $\vdots$ | $\vdots$ |
| Treatment $k$ | $x_{k1}$ | $x_{k2}$ | $\cdots$ | $x_{kn_k}$ | $x_{k.}$ |

$x_{ij}$ - the $j$ th sample value from the ith treatment, $j = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, k$

$k$ - number of treatments compared.

$x_{i.}$ - the sample total of $i^{th}$ treatment.

$n_i$ - the number of observations in the $i^{th}$ treatment.

$$\sum_{i=1}^{k} n_i = n$$

## Test Procedure

Let the observations $x_{ij}$, $j = 1, 2, \ldots, n_i$ for treatment $i$, be assumed to come from $N(\mu_i, \sigma^2)$ population, $i = 1, 2, \ldots, k$ where $\sigma^2$ is unknown.

**Step 1 : Framing Hypotheses**

**Null Hypothesis** $H_0 : \mu 1 = \mu 2 = \ldots = \mu_k$

That is, there is no significant difference among the population means of $k$ treatments.

**Alternative Hypothesis**

$H_1$: $\mu_i \neq \mu_j$ for atleast one pair $(i,j)$; $i, j = 1,2,\ldots,k$; $i \neq j$

That is, at least one pair of means differ significantly.

**Step 2 : Data**

Data is presented in the tabular form as described in the previous section

**Step 3 : Level of significance : $\alpha$**

**Step 4 : Test Statistic**

$F$ = $MST/MSE$ which follows $F_{(k-1,\, n-k)}$, under $H_0$

To evaluate the test statistic we compute the following:

(i) Correction factor: $C.F = \frac{G^2}{n}$ where $G = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}$

(ii) Total Sum of Squares: $TSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - C.F$

(iii) Sum of Squares between Treatments: $SST = \sum_{i=1}^{k}\frac{x_{i.}^2}{n_i} - C.F$,

where $x_{i.} = \sum_{j=1}^{n_i} x_{ij}$, $i = 1, 2, \ldots, k$

(iv) Sum of Squares due to Error: $SSE = TSS - SST$

## Mean Sum of Squares

Mean Sum of Squares due to treatment: $MST = SST / k - 1$

Mean Sum of Squares due to Error: $MSE = SSE / n - k$

**Step 5 : Calculation of Test statistic**

## ANOVA Table (one-way)

| Source of variation | Sum of squares | Degrees of freedom | Mean sum of squares | F-ratio |
|---|---|---|---|---|
| Treatments | $SST$ | $k$-1 | $MST = \frac{SST}{k-1}$ | $F_0 = \frac{MST}{MSE}$ |
| Error | $SSE$ | $n$-$k$ | $MSE = \frac{SSE}{n-k}$ | |
| Total | $TSS$ | $n$-1 | | |

**Step 6 : Critical value**

$f_e = f_{(k\text{-}1,\, n\text{-}k),\alpha}$.

**Step 7 : Decision**

If $F_0 < f_{(k\text{-}1,\, n\text{-}k),\alpha}$ then reject $H_0$.

innovate achieve lead

# Example:1

Three different techniques namely medication, exercises and special diet are randomly assigned to (individuals diagnosed with high blood pressure) lower the blood pressure. After four weeks the reduction in each person's blood pressure is recorded. Test at 5% level, whether there is significant difference in mean reduction of blood pressure among the three techniques.

| | | | | | |
|---|---|---|---|---|---|
| Medication | 10 | 12 | 9 | 15 | 13 |
| Exercise | 6 | 8 | 3 | 0 | 2 |
| Diet | 5 | 9 | 12 | 8 | 4 |

*Solution:*

**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 atleast 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 atleast 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 statistic**

| | | | | | | 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 |

$$\sum\sum x_{ij}^2 = 1162$$

1. Correction Factor: $CF = \frac{G^2}{n} = \frac{(116)^2}{15} = \frac{13456}{15} = 897.06$

2. Total Sum of Squares: $TSS = = \sum\sum x_{ij}^2 - C.F$

$$= 1162 - 897.06 = 264.94$$

3. Sum of Squares between Treatments: $SST = \frac{\sum x_i^2}{n_i} - C.F$

$$= \frac{5286}{5} - 897.06$$

$$= 1057.2 - 897.06$$

$$= 160.14$$

4. Sum of Squares due to Error: $SSE = TSS - SST$

$$= 264.94 - 160.14 = 104.8$$

## ANOVA Table (one-way)

| Source of variation | Sum of squares | Degrees of freedom | Mean sum of squares | $F$-ratio |
|---|---|---|---|---|
| Between treatments | 160.14 | $3 - 1 = 2$ | 80.07 | $F_0 = \frac{80.07}{8.73} = 9.17$ |
| Error | 104.8 | 12 | 8.73 | |
| Total | 264.94 | $n - 1 = 15 - 1$ $= 14$ | | |

**Step 6 : Critical value**

$f_{(2,\ 12),0.05} = 3.8853.$

**Step 7 : Decision**

As $F_0 = 9.17 > f_{(2,\ 12),0.05} = 3.8853$, the null hypothesis is rejected. Hence, we conclude that there exists significant difference in the reduction of the average blood pressure in atleast one pair of techniques.

# Example :2

Three composition instructors recorded the number of spelling errors which their students made on a research paper. At 1% level of significance test whether there is significant difference in the average number of errors in the three classes of students.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Instructor 1 | 2 | 3 | 5 | 0 | 8 | | |
| Instructor 2 | 4 | 6 | 8 | 4 | 9 | 0 | 2 |
| Instructor 3 | 5 | 2 | 3 | 2 | 3 | 3 | |

*Solution:*

**Step 1 : Hypotheses**

Null Hypothesis: $H_0 : \mu_1 = \mu_2 = \mu_3$

That is there is no significant difference among the mean number of errors in the three classes of students.

**Alternative Hypothesis**

$H_1 : \mu_i \neq \mu_j$ for at one pair $(i, j)$; $i,j = 1,2,3$; $i \neq j$.

That is, atleast one pair of groups differ significantly on the mean number of errors.

**Step 2 : Data**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Instructor 1 | 2 | 3 | 5 | 0 | 8 | | |
| Instructor 2 | 4 | 6 | 8 | 4 | 9 | 0 | 2 |
| Instructor 3 | 5 | 2 | 3 | 2 | 3 | 3 | |

innovate achieve lead

Step 3 : Level of significance $\alpha = 5\%$

Step 4 : Test Statistic

$F_0 = MST / MSE$

Step 5 : Calculation of Test statistic

| | | | | | | | | Total | Square |
|---|---|---|---|---|---|---|---|---|---|
| Instructor 1 | 2 | 3 | 5 | 0 | 8 | | | 18 | 324 |
| Instructor 2 | 4 | 6 | 8 | 4 | 9 | 0 | 2 | 33 | 1089 |
| Instructor 3 | 5 | 2 | 3 | 2 | 3 | 3 | | 18 | 324 |
| | | | | | | | | 69 | |

innovate achieve lead

## Individual squares

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Instructor 1 | 4 | 9 | 25 | 0 | 64 | | |
| Instructor 2 | 16 | 36 | 64 | 16 | 81 | 0 | 4 |
| Instructor 3 | 25 | 4 | 9 | 4 | 9 | 9 | |

$$\sum\sum x_{ij}^2 = 379$$

Correction Factor:

$$CF = \frac{G^2}{n} = \frac{(69)^2}{18} = \frac{4761}{18} = 264.5$$

Total Sum of Squares:

$$TSS = \sum\sum x_{ij}^2 - C.F$$
$$= 379 - 264.5 = 114.5$$

Sum of Squares between Treatments: $SST = \dfrac{\sum x_i^2}{n_i} - C.F$

$$= \left(\frac{324}{5} + \frac{1089}{7} + \frac{324}{6}\right) - 264.5$$

$$= (64.8 + 155.6 + 54) - 264.5$$

$$= (274.4) - 264.5$$

$$= 9.9$$

Sum of Squares due to Error: $SSE = TSS - SST$

$$= 114.5 - 9.9$$

$$= 104.6$$

## ANOVA Table

| Source of variation | Sum of squares | Degrees of freedom | Mean sum of squares | $F$-ratio |
|---|---|---|---|---|
| Between treatments | 9.9 | $3 - 1 = 2$ | $\frac{9.9}{2} = 4.95$ | $F_0 = \frac{4.95}{6.97} = 0.710$ |
| Error | 104.6 | 15 | $\frac{104.6}{15} = 6.97$ | |
| Total | | $n - 1 = 18 - 1 = 17$ | | |

## Step 6 : Critical value

The critical value $= f_{(15,\ 2),0.05} = 3.6823$.

## Step 7 : Decision

As $F_0 = 0.710 < f_{(15,\ 2),0.05} = 3.6823$, null hypothesis is not rejected. There is no enough evidence to reject the null hypothesis and hence we conclude that the mean number of errors made by these three classes of students are not equal. p-value equals **0.508553**

# TWO-WAY CLASSIFICATION

In two-way ANOVA a study variable is compared over three or more groups, controlling for another variable. The grouping is taken as one factor and the control is taken as another factor. The grouping factor is usually known as Treatment. The control factor is usually called Block. The accuracy of the test in two -way ANOVA is considerably higher than that of the one-way ANOVA, as the additional factor, block is used to reduce the error variance.

In two-way ANOVA, the data can be represented in the following tabular form.

**Blocks**

**Blocks**

| Groups or Treatments | 1 | 2 | 3 | … | $m$ | $x_{i.}$ |
|---|---|---|---|---|---|---|
| 1 | $x_{11}$ | $x_{12}$ | $x_{13.}$ | … | $x_{1m}$ | $x_{1.}$ |
| 2 | $x_{21}$ | $x_{22}$ | $x_{2.}$ | … | $x_{2m}$ | $x_{2.}$ |
| 3 | $x_{31}$ | $x_{32}$ | $x_{3.}$ | … | $x_{3m}$ | $x_{3.}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $k$ | $x_{k1}$ | $x_{k2}$ | $x_{k3}$ | … | $x_{km}$ | $x_{k.}$ |
| $x_{.j}$ | $x_{.1}$ | $x_{.2}$ | $x_{.3}$ | … | $x_{.m}$ | $G$ |

## 1. Test Procedure

Steps involved in two-way ANOVA are:

**Step 1 :** In two-way ANOVA we have two pairs of hypotheses, one for treatments and one for the blocks.

### Framing Hypotheses

### Null Hypotheses

$H_{01}$: There is no significant difference among the population means of different groups (Treatments)

$H_{02}$: There is no significant difference among the population means of different Blocks

### Alternative Hypotheses

$H_{11}$: Atleast one pair of treatment means differs significantly

$H_{12}$: Atleast one pair of block means differs significantly

**Step 2 :** **Data** is presented in a rectangular table form as described in the previous section.

**Step 3 :** **Level of significance** $\alpha$.

**Step 4 :** **Test Statistic**

$F_{0t}$ (treatments) = $MST / MSE$

$F_{0b}$ (block) = $MSB / MSE$

To find the test statistic we have to find the following intermediate values.

i) Correction Factor: $C.F = \frac{G^2}{n}$ where $G = \sum_{j=1}^{m} \sum_{i=1}^{k} x_{ij}$

ii) Total Sum of Squares: $TSS = \sum_{i=1}^{k} \sum_{j=1}^{m} x_{ij}^2 - C.F$

iii) Sum of Squares between Treatments: $SST = \sum_{i=1}^{k} \frac{x_{i.}^2}{m} - C.F$

iv) Sum of squares between blocks: $SSB = \sum_{j=1}^{m} \frac{x_{.j}^2}{k} - C.F$

v) Sum of Squares due to Error: $SSE = TSS\text{-}SST\text{-}SSB$

vi) Degrees of freedom

| Degrees of freedom (d.f.) | d.f. |
|---|---|
| Total Sum of Squares | $n-1$ |
| Treatment Sum of Squares | $k-1$ |
| Block Sum of Squares | $m-1$ |
| Error of Sum Squares | $(m-1)(k-1)$ |

vii) Mean Sum of Squares

Mean sum of Squares due to Treatments: $MST = \frac{SST}{k-1}$

Mean sum of Squares due to Blocks: $MSB = \frac{SSB}{m-1}$

Mean sum of Squares due to Error: $MSE = \frac{SSE}{(k-1)(m-1)}$

**Step 5 : Calculation of the Test Statistic**

## ANOVA Table (two-way)

| Source of variation | Sum of squares | Degrees of freedom | Mean sum of squares | F-ratio |
|---|---|---|---|---|
| Treatments | $SST$ | $k$-1 | $MST$ | $F_{0t} = \frac{MST}{MSE}$ |
| Blocks | $SSB$ | $m$-1 | $MSB$ | $F_{0b} = \frac{MSB}{MSE}$ |
| Error | $SSE$ | $(k-1)(m-1)$ | $MSE$ | |
| Total | $TSS$ | $n$-1 | | |

**Step 6 : Critical values**

Critical value for treatments = $f_{(k\text{-}1,(m\text{-}1)(k\text{-}1)),\alpha}$

Critical value for blocks = $f_{(m\text{-}1,\,(m\text{-}1)(k\text{-}1)),\alpha}$

**Step 7 : Decision**

For Treatments: If the calculated $F_{0t}$ value is greater than the corresponding critical value, then we reject the null hypothesis and conclude that there is significant difference among the treatment means, in atleast one pair.

For Blocks: If the calculated $F_{0b}$ value is greater than the corresponding critical value, then we reject the null hypothesis and conclude that there is significant difference among the block means, in at least one pair.

## Example:1

A reputed marketing agency in India has three different training programs for its salesmen. The three programs are Method – A, B, C. To assess the success of the programs, 4 salesmen from each of the programs were sent to the field. Their performances in terms of sales are given in the following table.

| 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.

*Solution:*

**Step 1 : Hypotheses**

**Null Hypotheses:** $H_{01} : \mu_{M1} = \mu_{M2} = \mu_{M3}$ (for treatments)

That is, there is no significant difference among the three programs 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 programs.

$H_{12}$ : 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

$$F_{0t}(\text{treatment}) = \frac{MST}{MSE}$$

$$F_{0b}(\text{block}) = \frac{MSB}{MSE}$$

## Step-5 : Calculation of the Test Statistic

| | 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 | 256 |
| 4 | 7 | 5 | 4 | 16 | 256 |
| $x_i$ | 22 | 28 | 16 | 66 | 1140 |
| $x_{i.}^2$ | 484 | 784 | 256 | 1524 | |

**Squares**

| | | |
|---|---|---|
| 16 | 36 | 4 |
| 36 | 100 | 36 |
| 25 | 49 | 16 |
| 49 | 25 | 16 |
| | | $\sum\sum x_{ij}^2 = 408$ |

Correction Factor: $$CF = \frac{G^2}{n} = \frac{(66)^2}{12} = \frac{4356}{12} = 363$$

Total Sum of Squares: $$TSS = \sum\sum x_{ij}^2 - C.F$$
$$= 408 - 363 = 45$$

Sum of Squares due to Treatments: $$SST = \frac{\sum_{i=1}^{k} x_{.j}^2}{k} - C.F$$
$$= \frac{1140}{3} - 363$$
$$= 380 - 363 = 17$$

Sum of Squares due to Blocks:

$$SSB = \frac{\sum_{i=1}^{k} x_{.j}^{2}}{k} - C.F$$

$$= \frac{1524}{4} - 363$$

$$= 381 - 363$$

$$= 18$$

Sum of Squares due to Error:

$$SSE = TSS - SST - SSB$$

$$= 45 - 17 - 18 = 10$$

Mean sum of squares:

$$MST = \frac{SST}{k-1} = \frac{17}{3} = 5.67$$

$$MSB = \frac{SSB}{m-1} = \frac{18}{2} = 9$$

$$MSE = \frac{SSE}{(k-1)(m-1)} = \frac{10}{6} = 1.67$$

## ANOVA Table (two-way)

| Sources of variation | Sum of squares | Degrees of freedom | Mean sum of squares | $F$-ratio |
|---|---|---|---|---|
| Between treatments (Programs) | 17 | 3 | 5.67 | $F_{ot} = \frac{5.67}{1.67} = 3.40$ |
| Between blocks (Salesmen) | 18 | 2 | 9 | $F_{ob} = \frac{9}{1.67} = 5.39$ |
| Error | 10 | 6 | 1.67 | |
| Total | | 11 | | |

**Step 6 : Critical values**

$f_{(3,\ 6),0.05} = 4.7571$ (for treatments)

$f_{(2,\ 6),0.05} = 5.1456$ (for blocks)

**Step 7 : Decision**

(i) Calculated $F_{0t} = 3.40 < f_{(3,\ 6),0.05} = 4.7571$, the null hypothesis is not rejected and we conclude that there is significant difference in the mean sales among the three programs.

(ii) Calculate $F_{0b} = 5.39 > f_{(2,\ 6),0.05} = 5.1456$, the null hypothesis is rejected and conclude that there does not exist significant difference in the mean sales among the four salesmen.

# Problems

1)

Four brands of flashlight batteries are to be compared by testing each brand in five flashlights. Twenty flashlights are randomly selected and divided randomly into four groups of five flashlights each. Then each group of flashlights uses a different brand of battery. The lifetimes of the batteries, to the nearest hour, are as follows.

| Brand *A* | Brand *B* | Brand *C* | Brand *D* |
|---|---|---|---|
| 42 | 28 | 24 | 20 |
| 30 | 36 | 36 | 32 |
| 39 | 31 | 28 | 38 |
| 28 | 32 | 28 | 28 |
| 29 | 27 | 33 | 25 |

Preliminary data analyses indicate that the independent samples come from normal populations with equal standard deviations. At the 5% significance level, does there appear to be a difference in mean lifetime among the four brands of batteries?

2. The times required by three workers to perform an assembly-line task were recorded on five randomly selected occasions. Here are the times, to the nearest minute

| Hank | Joseph | Susan |
|---|---|---|
| 8 | 8 | 10 |
| 10 | 9 | 9 |
| 9 | 9 | 10 |
| 11 | 8 | 11 |
| 10 | 10 | 9 |

Construct the one way ANOVA table for the data at 5% level of significance.

3. Three drying formulas for curing glue are studied

| | | | | | | |
|---|---|---|---|---|---|---|
| Formula A | 13 | 10 | 8 | 11 | 8 | |
| Formula B | 13 | 11 | 14 | 14 | | |
| Formula C | 17 | 14 | 13 | 10 | 11 | 12 |

Test at 5% level of significance whether is any difference in the mean curing time of glue?

4.

The illness caused by a virus in a city concerning some restaurant inspectors is not consistent with their evaluations of cleanliness of restaurants. In order to investigate this possibility, the director has five restaurant inspectors to grade the cleanliness of three restaurants. The results are shown below.

| Inspectors | Restaurants | | |
|---|---|---|---|
| | I | II | III |
| 1 | 71 | 55 | 84 |
| 2 | 65 | 57 | 86 |
| 3 | 70 | 65 | 77 |
| 4 | 72 | 69 | 70 |
| 5 | 76 | 64 | 85 |

Carry out two-way ANOVA at 5% level of significance.

5. Future group wishes to enter the frozen shrimp market. They contract a researcher to investigate various methods of groups of shrimp in large tanks. The researcher suspects that temperature and salinity are important factor influencing shrimp yield and conducts a two-way analysis of variance with their levels of temperature and salinity. That is each combination of yield for each (for identical gallon tanks) is measured. The recorded yields are given in the following chart:

**Categorical variable Salinity (in pp)**

| Temperature | 700 | 1400 | 2100 | Total | Mean |
|---|---|---|---|---|---|
| $60^0$ F | 3 | 5 | 4 | 12 | 4 |
| $70^0$ F | 11 | 10 | 12 | 33 | 11 |
| $80^0$ F | 16 | 21 | 17 | 54 | 18 |
| Total | 30 | 36 | 33 | 99 | 11 |

Compute the ANOVA table for the model.

innovate achieve lead

# Thank you