Bhoris Dhanjal

Analysis of Variance & Design of Experiments

Lecture Notes
for SSTA402

Contents

Chapter 1

Analysis of Variance (Fixed effect models)

1.1 Chi-squared distribution

Definition 1.1. *Consider the standard normal distribution*

$$Z = \frac{x - \mu}{\sigma} \sim SN(0, 1)$$

The square of the standard normal distribution gives us chi-squared with degree of freedom 1.

In general for iid SN variates Z_i we can say

$$\sum_{i=1}^{n} Z_i^2 \sim \chi^2_{(n)}$$

1.1.1 Tests to use depending on type of data

- For one categorical feature (is there a difference in the proportion) we will use **one sample proportion test.**
- When two categorical features we will use **chi-squared test.**
- When one numeric data (is there is difference in the mean) we will use **t-test.**
- when two numeric data we will use the **t-test for two samples.**

1.2 Analysis of variance

Analysis of variance is a powerful statistical tool for test of significance.

Definition 1.2 (Statistical significance)**.** *We have evidence that the result v_c in the sample also exists in population*

Very large sample: Most results will be statistically significant.
Very small sample: Most results will not be statistically significant.

Definition 1.3 (p-value)**.** *Probability value, it indicates, how likely it is that the result occurred by chance alone.*

The test of significance based on t-distribution is adequate only for testing the significance h difference between two sample means.

When we have 3 or more samples (and we need to test if they all come from the same population) we need an alternative procedure.

Example 1.4. *5 fertilizers are applies to four plots of wheat and yield of wheat on each of the plot is given. We may be interested is to find out whether the effect of these fertilizers on the yields is significantly different or not.*

The basic purpose of ANOVA is to test the homogeneity of several means.

Variation is inherent in nature, the total variation in any set of numerical data is due to a number of causes which may be classified as

1. Assignable causes
2. Chance causes

The variation due ot assignable causes can be detected and measured, whereas the variation due to chance causes is beyond the control of human hand and cannot be traced separately.

Definition 1.5 (Analysis of variance)**.** *According to R.A Fischer, ANOVA is the "separation of variance ascribable to one group of causes from the variance ascribable to other group".*

Assumptions for ANOVA: ANOVA test is based on the test statistics F (or variance ratio). For the validity of the F-test in ANOVA, the following assumptions are made

1. The observations are independent.
2. Parent population from which observations are taken is normal

3. Various treatment and environment effects are additive in nature.;

Theorem 1.6 (Cochran's theorem)**.** *Let $x_1, x_2, \ldots, x_n$ denote a random sample from normal population $N(0, \sigma^2)$. Let the sum of the squares of theese values be written in in the form*

$$\sum_{i=1}^{n} x_i^2 = Q_1 + Q_2 + \cdots + Q_k$$

Where Q_j is a quadratic form in $x_1, x_2, \ldots, x_n$ with rank (degrees of freedom) $r_j = 1, 2, \ldots, k$ then the random variables $Q_1, Q_2, \ldots, Q_k$ are mutually independent and $\frac{Q_i}{\sigma^2}$ is χ_k^2 variable with r_j degrees of freedom iff $\sum_{j=1}^{k} r_j = n$

1.3 One-way classification

Let us suppose that N observations $y_{ij}, (i = 1, 2, \ldots, k; j = 1, 2, \ldots, n)$ of a random variable Y are grouped on some basis, into k classes of sizes $n_1, n_2, \ldots, n_k$ respectively, $\left(N = \sum_{i=1}^{k} n_i\right)$ as exhibited in below table.

$$\begin{bmatrix} \text{Class} & \text{Sample observations} & & & & \text{Total} & \text{Mean} \\ 1 & y_{11} & y_{12} & \cdots & y_{1n_1} & T_1 & \overline{y_{1.}} \\ 2 & y_{21} & y_{22} & \cdots & y_{2n_1} & T_2 & \overline{y_{2.}} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ k & y_{k1} & y_{k2} & \cdots & y_{kn_k} & T_k & \overline{y_{k.}} \end{bmatrix}$$

The total variation in the observation y_{ij} can be split into the following two components:

1. The variation between the classes or the variation due to different bases of the classification, commonly known as treatments.

2. The variation within the classes, i.e. the inherent variation of the random variable within the observation of a class.

The first type of variation is due to assingable cause which can be detected and controlled by human endeavour and second type of varation is due to chance causes, which are beyond control of human hand.

The main object of analysis of variance technique is to examine if there is significant difference ebtween the claass means in view of the inherent vairaiblity within the seperate classes.

Example 1.7. *In particular, let us consider the effect of k different rations on the yield in milk of N cows (of the same breed and stock) divided into k classes of sizes $n_1, n_2, \ldots, n_k$ respectively, $N = \sum_{i=1}^{k} n_i$.*

The sources of variation is,

1. Effect of treatments (i.e. classes).
2. Error (ε) produced by numerical causes

1.3.1 Mathematical model

$$\begin{aligned} y_{ij} &= \mu_i + \varepsilon_{ij} \\ &= \mu + \alpha_i + \varepsilon_{ij} \\ &= \mu + (\mu_u - \mu) + \varepsilon_{ij} \qquad\qquad i = 1, \ldots, k; j = 1, \ldots, n_i \end{aligned}$$

Where, $\alpha_i = \mu_i - \mu$, effect due to i^{th} treatment.

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

y_{ij} j^{th} observation receiving i^{th} treatment, μ is general mean effect, $\mu = \sum_{i=1}^{k} \frac{\mu_i n_i}{N}$, α_i = effect due to i^{th} level of factor $i = 1, \ldots, k$,

$$\sum_{i=1}^{k} n_i \alpha_i = \sum_{i} n_i(\mu_i - \mu) = \sum n_i \mu_i - \sum n_i \mu = 0$$

ε_{ij} is random error or eror component (assume $\varepsilon_{ij} \sim N(0, \sigma^2)$)

1.3.2 Layout of one way ANOVA

$$\begin{bmatrix} \text{Factor level} & \text{Observations} & \text{Total} & \text{Mean} \\ 1 & y_{11} y_{12} \ldots y_{1n_1} & \sum_{j=1}^{n_1} y_{1j} = y_{1\cdot} & \frac{y_{1\cdot}}{n_1} = \overline{y_{1\cdot}} \\ 2 & y_{21} y_{22} \ldots y_{2n_1} & \sum_{j=1}^{n_2} y_{2j} = y_{2\cdot} & \frac{y_{2\cdot}}{n_2} = \overline{y_{2\cdot}} \\ \vdots & \vdots \quad \vdots \quad \ddots \quad \vdots & \vdots & \vdots \\ k & y_{k1} y_{k2} \ldots y_{kn_k} & \sum_{j} y_{kj} = y_{k\cdot} & \frac{y_{k\cdot}}{n_k} = \overline{y_{k\cdot}} \end{bmatrix}$$

y_{ij} is j^{th} observation due to i^{th} level of factor, $i = 1, 2, \ldots, k; j = 1, 2, \ldots, n_i$

$$\sum_{i=1}^{k} n_i = n_1 + n_2 + \cdots n_k = N$$

$$y_i. = \sum_{j=1}^{n_i} y_{ij}, \text{ when, } i = 1, 2, \ldots, k \text{ total of } i^{th} \text{ row}$$

$$y.. = \sum_{i=1}^{k} y_i. = G = \text{Grand total} = \sum_{i=1}^{k} \left[\sum_{j=1}^{n_i} y_{ij} \right]$$

$$x\overline{y_i.} = \frac{y_i.}{n_i} = \sum_{j=1}^{n_i} \frac{y_{ij}}{n_i} = \text{Mean of obs of } i^{th} \text{ level of factor}, i = 1, 2, \ldots, k$$

$$\overline{y..} = \frac{1}{N} \sum_{i} \sum_{j} y_{ij} = \frac{1}{N} \sum_{i} n_i \overline{y_i.} = \text{Overall mean}$$

1.3.3 Assumptions in the model

1. All the observations are independent and $y_{ij} \sim N(\mu_i, \sigma_e^2)$
2. Different effects are additive in nature.
3. ε_{ij} are i.i.d $N(0, \sigma_o^2)$ i.e. $E[\varepsilon_{ij}] = 0$ and $V[\varepsilon_{ij}] = 0 \ \forall i, j$

1.3.4 Hypothesis to be tested

Null hypothesis

We want to test the equality of the population means, i.e. the homogeneity of different treatment. Hence, the null hypothesis is given by

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

which reduces to

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$$

Alternate hypothesis

At least two of the means $(\mu_1, \mu_2, \ldots, \mu_k)$ are different.

1.3.5 Estimating the parameters of the model

The model is

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, i = 1, \ldots, k; j = 1, \ldots, n$$

To estimate the parameters we use least squares method

Theorem 1.8. *Estimating μ*

Proof.

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \varepsilon_{ij}^2 = \sum_i \sum_j (y_{ij} - \mu - \alpha_i)^2$$

Differentiate E w.r.t μ and α_i

$$\begin{aligned} \frac{\partial E}{\partial \mu} &= \frac{\partial}{\partial \mu} \left[\sum_i \sum_j (y_{ij} - \mu - \alpha_i)^2 \right] \\ &= 2 \sum_i \sum_j (y_{ij} - \mu - \alpha_i)(-1) \end{aligned}$$

Equating with 0

$$\begin{aligned} \frac{\partial E}{\partial \mu} = 0 &\implies \sum_{i=1}^{k} \sum_{j=1}^{n_1} = 0 \\ &\implies \sum_i \sum_j y_{ij} - \sum_i \sum_j \mu - \sum_i \sum_j \alpha_i = 0 \\ &\implies \sum_i \sum_j y_{ij} - \mu \sum_i \sum_j 1 - \sum_i \sum_j \alpha_i = 0 \\ &\implies \sum_i \sum_j y_{ij} - \mu N - 0 = 0 \\ &\implies \sum_i \sum_j y_{ij} - N\hat{\mu} = 0 \\ &\implies \hat{\mu} = \frac{\sum_i \sum_j y_{ij}}{N} = \overline{y_{..}} = G \end{aligned}$$

□

Theorem 1.9. *Estimating α_i*

Proof. Begin similarly but now different w.r.t α_i

$$\begin{aligned}
\frac{\partial E}{\partial \alpha_i} &= \frac{\partial}{\partial \alpha_i}\left[\sum_i \sum_j (y_{ij} - \mu - \alpha_i)^2\right] \\
&= 2\sum_j (y_{ij} - \mu - \alpha_i)(-1) \\
&\implies \sum_j (y_{ij} - \mu - \alpha_i) = 0 \\
&\implies \sum_j y_{ij} - \sum_j \mu - \sum_j \alpha_i = 0 \\
&\implies y_{i\cdot} - n_i\mu - n_i\hat{\alpha}_i = 0 \\
&\implies \hat{\alpha}_i = \frac{y_{i\cdot}}{n_i} - \hat{\mu} = \overline{y_{i\cdot}} - \overline{y_{\cdot\cdot}} \\
&\implies \hat{\alpha_i}_{\cdot} = \overline{y_{i\cdot}} - \overline{y_{\cdot\cdot}}
\end{aligned}$$

Substituting these in the equation

$$\begin{aligned}
y_{ij} &= \hat{\mu} + \hat{\alpha}_i + \varepsilon_{ij} \\
y_{ij} &= \overline{y_{\cdot\cdot}} + (\overline{y_{i\cdot}} + \overline{y_{\cdot\cdot}}) + (y_{ij} - \overline{y_{i\cdot}})
\end{aligned}$$

Observation = Grand mean + deviation due to i^{th} treatment + residual or error.
Subtracting $x\overline{y_{\cdot\cdot}}$ from both sides, squaring and summing over i, j.

$$\begin{aligned}
\sum_i \sum_j (y_{ij} - \overline{y_{\cdot\cdot}})^2 &= \sum_i \sum_j ((\overline{y_{i\cdot}} - \overline{y_{\cdot\cdot}}) + (y_{ij} - \overline{y_{i\cdot}}))^2 \\
&= \sum_i \sum_j (\overline{y_{i\cdot}} - \overline{y_{\cdot\cdot}})^2 + \sum_i \sum_j (y_{ij} - \overline{y_{i\cdot}})^2 \\
&\quad + 2\sum_i \sum_j (\overline{y_{1\cdot}} + \overline{y_{\cdot\cdot}})(y_{ij} - \overline{y_{i\cdot}})
\end{aligned}$$

But second term is zero as it is sum of deviations from their mean

$$\sum_i \sum_j (y_{ij} - \overline{y_{\cdot\cdot}})^2 = \sum_i n_i (\overline{y_{i\cdot}} - \overline{y_{\cdot\cdot}})^2 + \sum_i \sum_j (y_{ij} - \overline{y_{i\cdot}})^2$$

Sum of squares of total variation = variation due to level of factors + variation due to error

Total sum of squares = sum of squares due to factor + sum of squares due to error

Total s.s = between s.s. + within s.s.

□

1.3.6 Degrees of freedom

The total no. of observations is equal to N. The total no of s.s. is computed from N observations which are subjected to one restriction.
Therefore, total degrees of freedom (d.f.)=$N-1$.
Degree of freedom due to factors is $k-1$. So now we can say degree of freedom for ss due to error is $n-k=x$

1.3.7 Expectation of various sum of squares

The model is,

$$\begin{aligned}
y_{ij} &= \mu + \alpha_i + \varepsilon_{ij} \\
\overline{y_{i\cdot}} &= \sum_{j=1}^{n_i} \frac{y_{ij}}{n_i} \\
&= \sum_{j} \frac{\mu + \alpha_i + \varepsilon_{ij}}{n_i} \\
&= \frac{n_i\mu + n_i\alpha_i + \varepsilon_{i\cdot}}{n_i} = \mu + \alpha_i + \frac{\varepsilon_{i\cdot}}{n_i} \\
\overline{y_{\cdot\cdot}} &= \sum_i \sum_j \frac{y_{ij}}{N} \\
&= \frac{N\mu}{N} + \frac{\sum_i n_i\alpha_i}{N} + \frac{\varepsilon_{\cdot\cdot}}{N} \\
\overline{y_{\cdot\cdot}} &= \mu + \overline{\varepsilon_{\cdot\cdot}}
\end{aligned}$$

Assume $\varepsilon_{ij} \sim N(0, \sigma^2)$

$$\begin{aligned}
\overline{\varepsilon_{i\cdot}} &= \sum_j \frac{\varepsilon_{ij}}{n_i} \\
E[\overline{\varepsilon_{i\cdot}}] &= 0 \\
V[\overline{\varepsilon_{i\cdot}}] &= \frac{\sigma^2}{n_i}
\end{aligned}$$

Expectation of s.s. due to factor

$$E[s.s.duetofacor] = E\left[\sum_i n_i(\overline{y_i} - \overline{y..})^2\right]$$
$$= E\left[\sum_i n_i\left((\mu + \alpha_i + \overline{\varepsilon_i.}) - (\mu + \overline{\varepsilon..})\right)^2\right]$$

Expand this out...

$$= \sum_i n_i\alpha_i^2 + E\left[\sum_i n_i(\overline{\varepsilon_i.} - \overline{\varepsilon..})^2\right] + 0$$
$$E\left[\sum_i n_i(\overline{e_i.} - \overline{e..})^2\right] = \sum_i n_i[E[\overline{e_i.}^2] + E[\overline{e..}^2] - 2E[\overline{e_i.} - \overline{e..}]]$$
$$= \sum_i n_i\left[\frac{\sigma^2}{n_i} + \frac{\sigma^2}{N} - 2E[\overline{e_i.} - \overline{e..}]\right]$$
$$= k\sigma^2 + \sigma^2 - 2\sum_i n_i(\overline{e_i.} - \overline{e..})$$
$$E[s.s.duetofactors] = \sum_i n_i\alpha_i^2 + k\sigma^2 + \sigma^2 - E\left[2\sum_i n_i(\overline{e_i.} - \overline{e..})\right]$$
$$E\left[\sum_i n_i(\overline{e_i.} - \overline{e..})\right] = E\left[\overline{e_i.}\sum_i n_i(e../n_i)\right]$$
$$= E[\overline{e..}N\overline{e..}]$$
$$= NE[\overline{e..}^2] = \frac{N\sigma^2}{N} = \sigma^2$$
$$E[s.s.duetofactors] = \sum_i n_i\alpha_i^2 + k\sigma^2 - \sigma^2$$
$$= \sum_i n_i\alpha_i^2 + (k-1)\sigma^2$$

We can now also say,

$$\text{Mean s.s} = \frac{\text{s.s. due to factor}}{\text{degrees of freedom}}$$

$$E\left[\frac{\text{s.s. due to factors}}{k-1}\right] = \frac{1}{k-1}\sum_i n_i\alpha_i^2 + \sigma^2$$

If H_0 is true $\forall i$ then, $E[\text{m.s.s. due to factor}] = \sigma^2$

Expectation of error s.s.

$$E[\text{Error s.s.}] = E\left[\sum_i \sum_j (y_{ij} - \overline{y_{i\cdot}})^2\right]$$
$$= E\left[\sum_i \sum_j (y_{ij} - (\mu + \alpha_i + \overline{e_{i\cdot}}))^2\right]$$

magic

$$= E[\sum_i \sum_j (e_{ij} - \overline{e_{i\cdot}})^2]$$
$$= E\left[\sum_i \sum_j (e_{ij}^2) + \sum_i \sum_j (\overline{e_{i\cdot}}^2) - 2\sum_i \sum_j (e_{ij} - \overline{e_{i\cdot}})\right]$$
$$= N\sigma^2 + k\sigma^2 - 2\sum_i \sum_j E[e_{ij} - \overline{e_{i\cdot}}]$$

Consider now,

$$\sum\sum E[e_{ij} - \overline{e_{i\cdot}}] = \sum_i n_i E[\overline{e_{i\cdot}}^2] = \sum_i \frac{n_i\sigma^2}{n_i} = k\sigma$$
$$E[\text{error s.s.}] = N\sigma^2 + k\sigma^2 - 2k\sigma^2$$
$$= (N-k)\sigma^2$$

Now we can say,

$$E[\text{m.s.s. for error}] = \sigma^2$$

From these two subsections we can say $E[\text{m.s.s for factor}] \geq E[\text{m.s.s for error}]$. Equality holds when H_0 is true.

1.3.8 Test statistics

Using Cochran's theorem.
We have $\frac{s.sduetofactor}{\sigma^2}$, $\frac{ssduetoerror}{sigma^2}$ follows chi squared with $(k-1)$ and $(N-k)$ degrees of freedom respectively.
Therefore, the test statistics are defined as follows

$$f = \frac{\chi_1^2/d.f.}{\chi_2^2/d.f.} = \frac{\text{s.s. due to factors}/\sigma^2/(k-1)}{\text{s.s. due ot error}/\sigma^2/(N-k)}$$
$$\frac{\text{M.s.s for factor}}{\text{m.s.s. for error}}$$

We will reject $H_0 : \alpha_i = 0 \forall i$ if $F_{cal} > F_{tab}$ as two independent χ^2 variates with $(k-1), (N-k)$ d.f. respectively.

1.3.9 Computations forms of s.s.

Computational form of total s.s.

$$\begin{aligned}
\text{Total s.s.} &= \sum_i \sum_j (y_{ij} - \overline{y..})^2 \\
&= \sum \sum (y_{ij}^2 + \overline{y..}^2 - 2y_{ij}\overline{y..}) \\
&= \sum \sum y_{ij}^2 + N\overline{y..}^2 - 2\overline{y..} \sum \sum y_{ij} \\
&= \sum \sum y_{ij}^2 - \frac{Ny^2..}{N^2} \\
\text{Total s.s.} &= \sum_i \sum_j y_{ij}^2 - \frac{y^2..}{N} = \sum_i \sum_j y_{ij}^2 - \frac{G^2}{N} \\
\text{Total s.s.} &= \text{Raw s.s.} - G.F
\end{aligned}$$

Computational form of s.s. due to factors

$$\text{s.s. due to factor} = \sum_i n_i (\overline{y_i} - \overline{y..})^2$$

Magic...

$$\text{s.s. due to factor} = \sum_i \frac{T_i}{n_i} - \frac{G^2}{N}$$

1.3.10 One way ANOVA Table

Replace R with T below here,

$$\begin{bmatrix}
\text{Source of variance} & d.f. & s.s. & m.s.s & F \\
\text{Between factor (treatment)} & k-1 & \sum_i \frac{R_i^2}{n_i} - \frac{G^2}{N} = S_R^2 & s_R^2 = \frac{S_R^2}{k-1} & F = \frac{s_R^2}{s_E^2} \\
\text{Within factor (eror)} & N-k & + = S_E^2 & s_E^2 = \frac{s_E^2}{N-k} & - \\
\text{Total} & N-1 & \sum_i \sum_j y_{ij}^2 - \frac{G^2}{N} & - & -
\end{bmatrix}$$

1.3.11 Example of One-way classification ANOVA

Type this out later

1.4 Two-way classification

Suppose n observations are classification into k categories (or classes), say $A_1, A_2, \ldots, A_k$ according to some criterion A; and into h categories, say $B_1, B_2, \ldots, B_h$ according to some criterion B having kh combinations (A_i, B_j) $i = 1, 2, \ldots k; j = 1, 2, \ldots, h$; often called cells.

This scheme of classification according to two factors or criteria is called two-way classification and its analysis is called two way ANOVA. The number of observations in each cell may be equal or different, but we shall consider the case of one observation per cell so that $n = kh$, i.e. the total no of cell is $n = kh$.

Example 1.10. *If we have* 4 *different fertilizers say* A, B, C, D *and* 5 *different types of seeds then we will have* 20 *plots with each of oe having one of the four fertilizers and one of the five seeds. The yields from these plots will be analysed to check whether there is significant difference between fertilizers or between seeds there are two was in which we can compare the data the analysis is called two-way analysis.*

1.4.1 Layout of two-way classification

Levels	1	2	...	j	...	h	Total	Mean
1	y_{11}	y_{12}	...	y_{1j}	...	y_{1h}	$y_{1\cdot}$	$\overline{y_{1\cdot}}$
2	y_{21}	y_{22}	...	y_{2j}	...	y_{2h}	$y_{2\cdot}$	$\overline{y_{2\cdot}}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
i	y_{i1}	y_{i2}	...	y_{ij}	...	y_{ih}	$y_{i\cdot}$	$\overline{y_{i\cdot}}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
k	y_{k1}	y_{k2}	...	y_{kj}	...	y_{kh}	$y_{k\cdot}$	$\overline{y_{k\cdot}}$
Total	$y_{\cdot 1}$	$y_{\cdot 2}$	...	$y_{\cdot j}$	...	$y_{\cdot h}$		
Mean	$\overline{y_{\cdot 1}}$	$\overline{y_{\cdot 2}}$	...	$\overline{y_{\cdot j}}$	...	$\overline{y_{\cdot h}}$		

1.4.2 Notation

$$N = kh$$
$$y.. = \sum_i \sum_j y_{ij} = G$$
$$\overline{y..} = \sum_i \sum_j \frac{y_{ij}}{N} = \frac{y..}{kh}$$

y_{ij} = The observation receiving i^{th} level of factor A and j^{th} level of factor B

$$R_i = y_i. = \sum_{j=1}^{h} y_{ij}$$
$$\overline{y_i.} = \frac{y_i.}{h}$$
$$C_j = y_{\cdot j} = \sum_{i=1}^{k} y_{ij}$$
$$\overline{y_{\cdot j}} = \frac{y_{\cdot j}}{k}$$

1.4.3 Assumptions of two-way classification

1. Observations are independent.
2. Different effects are additive in nature
3. e_{ij} are i.i.d $N(0, \sigma^2)$

1.4.4 Model of two-way classification

$$y_{ij} = \mu_{ij} + e_{ij}, i = 1, \ldots, k; j = 1, \ldots, h$$

Here, e_{ij} is effect due to chance causes. $E[y_{ij}] = \mu_{ij}$ =fixed effect due to assingable causes where y_{ij} are indepdent $N(\mu_{ij}, \sigma^2)$.
μ_{ij} is further split into the followring parts

1. The general mean effect is given by

$$\mu = \sum_i \sum_j \frac{\mu_{ij}}{N}$$

2. α_i effect due to i^{th} level of factor A, $\alpha_i = \mu_i - \mu, i = 1, 2, \ldots, k$ and $\mu_i. = \sum_j \frac{\mu_{ij}}{h}, i = 1, 2, \ldots, k$. Sum of α_i is zero.

   Proof.

$$\begin{aligned}\sum_{i=1}^{k} \alpha_i &= \sum_i (\mu_i. - \mu) \\ &= \sum_i \mu_i. - \mu k \\ &= k \sum_i \sum_j \frac{\mu_{ij}}{hk} - \mu k = 0\end{aligned}$$

□

3. β_j effect due to j^{th} level of factor B. $\beta_j = \mu_{.j} - \mu; j = 1, \ldots, h$ and same thing as above write it out.

4. The interaction effect γ_{ij} when the i^{th} level of first factor and j^{th} level of factor B occurs simultaneously and is given by

$$\gamma_{ij} = \mu_{ij} - \mu_i. - \mu_{.j} - \mu$$

   And its summation across j and i is zero.

Thus, we have $\mu_{ij} = \mu + (\mu_i. - \mu) + (\mu_{.j} - \mu) + (\mu_{ij} - \mu_i. - \mu_{.j} - \mu)$

As there is only one observation per cell we cannot estimate interactive effect. Hence interaction effect is zero and the model reduces to.

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

for $i = 1, \ldots, k; j = 1, \ldots, h$

1.4.5 Hypothesis

Null hypothesis

H_{0_α} : There is no significant difference of factor A,

$$\begin{aligned}\mu_1. = \mu_2. = \cdots = \mu_k. = \mu \\ \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0\end{aligned}$$

H_{0_β} : There is no significant difference of factor B,

$$\begin{aligned}\mu_{.1} = \mu_{.2} = \cdots = \mu_{.k} = \mu \\ \beta_1 = \beta_2 = \cdots = \beta_h = 0\end{aligned}$$

Alternative hypothesis

$H_{1\alpha}$: At least two of the $\mu_i.$'s are different.
i.e. At least one of the α_i's is not zero.
$H_{1\beta}$: At least two of the $\mu.{}_j$'s are different.
i.e. At least one of the β_j's is not zero.

1.4.6 Least square estimators of parameters

To obtain $\hat{\mu}, \hat{\alpha_i}, \hat{\beta_j}$

$$E = \sum_i \sum_j e_{ij}^2 = \sum_i \sum_j (y_{ij} - \mu - \alpha_i - \beta_j)^2$$

First derivate it with respect to μ and equate it to zero to get $\hat{\mu}$

$$\begin{aligned}
\frac{\partial E}{\partial \mu} = 0 &\implies 2\sum_i \sum_j (y_{ij} - \mu - \alpha_i - \beta_j)(-1) = 0 \\
&\implies \sum_i \sum_j y_{ij} - \sum_i \sum_j \mu - \sum_i \sum_j \alpha_i - \sum_i \sum_j \beta_j = 0 \\
&\implies y.. - \mu N - 0 - 0 \\
&\implies \hat{\mu} = \frac{y..}{N} = \overline{y..}
\end{aligned}$$

Derivative it with respect to α_i

$$\begin{aligned}
\frac{\partial E}{\partial \alpha_i} = 0 &\implies 2\sum_j (y_{ij} - \mu - \alpha_i - \beta_j)(-1) = 0 \\
&\implies \sum_j y_{ij} - \sum_j \mu - \sum_j \alpha_i - \sum_j \beta_j = 0 \\
&\implies \sum_j y_{ij} - h\mu - \sum_j \alpha_i - \sum_j \beta_j = 0
\end{aligned}$$

$\sum_j \beta_j = 0$ so,

$$\begin{aligned}
&\implies \sum_j y_{ij} - h\widehat{\mu} = h\widehat{\alpha_i} \\
&\implies \widehat{\alpha_i} = \frac{1}{h}\sum_j y_{ij} - \widehat{\mu} \\
&\implies \widehat{\alpha_i} = \overline{y_i.} - \overline{y..}
\end{aligned}$$

Differentiate it with respect to β_j

$$\begin{aligned}\frac{\partial E}{\partial \beta_j} = 0 &\implies 2\sum_i (y_{ij} - \mu - \alpha_i - \beta_j)(-1) = 0 \\ &\implies \sum_i y_{ij} - \sum_i \mu - \sum_i \alpha_i - \sum_i \beta_j = 0\end{aligned}$$

Similar steps as above

$$\implies \hat{\beta}_j = \overline{y_{\cdot j}} - \overline{y_{\cdot\cdot}}$$

Now we can find the value of e_{ij},

$$\begin{aligned}y_{ij} &= \mu + \alpha_i + \beta_j + e_{ij} \\ &= \overline{y_{\cdot\cdot}} + \overline{y_{i\cdot}} - \overline{y_{\cdot\cdot}} + \overline{y_{\cdot j}} - \overline{y_{\cdot\cdot}} + e_{ij} \\ &= \overline{y_{i\cdot}} + \overline{y_{\cdot j}} - \overline{y_{\cdot\cdot}} + e_{ij} \\ \implies e_{ij} &= y_{ij} - \overline{y_{i\cdot}} - \overline{y_{\cdot j}} + \overline{y_{\cdot\cdot}}\end{aligned}$$

The model is then,

$$y_{ij} = \overline{y_{\cdot\cdot}} + (\overline{y_{i\cdot}} - \overline{y_{\cdot\cdot}}) + (\overline{y_{\cdot j}} - \overline{y_{\cdot\cdot}}) + (y_{ij} - \overline{y_{i\cdot}} - \overline{y_{\cdot j}} - \overline{y_{\cdot\cdot}})$$

Subtracting $\overline{y_{\cdot\cdot}}$ from both the sides squaring both sides and summing over i and j.

$$\begin{aligned}\sum_i \sum_j (y_{ij} - \overline{y_{\cdot\cdot}})^2 &= \sum_i \sum_j [(\overline{y_{i\cdot}} - \overline{y_{\cdot\cdot}}) + (\overline{y_{\cdot j}} - \overline{y_{\cdot\cdot}}) + (y_{ij} - \overline{y_{i\cdot}} - \overline{y_{\cdot j}} - \overline{y_{\cdot\cdot}})]^2 \\ &= \sum_i \sum_j (\overline{y_{i}} - \overline{y_{\cdot\cdot}})^2 + \sum_i \sum_j (\overline{y_{\cdot j}} - \overline{y_{\cdot\cdot}})^2 + \sum_i \sum_j (y_{ij} - \overline{y_{i\cdot}} - \overline{y_{\cdot j}} - \overline{y_{\cdot\cdot}})^2\end{aligned}$$

All cross product terms vanish

Total s.s. = Factor a.s.s+Factor b.s.s.+Error s.s.

1.4.7 Degrees of freedom

- Total d.f. $N - 1$.
- Factor A d.f. $k - 1$.
- Factor B d.f. $h - 1$
- Error d.f. $N - k - h + 1 = (k - 1)(h - 1)$

1.4.8 Mean sum square

- $MSS_A = \frac{SS_A}{k-1}$
- $MSS_B = \frac{SS_B}{h-1}$
- $MSS_{err} = \frac{ErrSS}{(k-1)(h-1)}$

1.4.9 Expectation of various SS

The model is

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

for $i = 1, \ldots, k; j = 1, \ldots h$

$$\overline{y_{i\cdot}} = \mu + \alpha_i + \overline{e_{i\cdot}}$$

Similarly

$$\overline{y_{\cdot j}} = \mu + \beta_j + \overline{e_{\cdot j}}$$

$$\overline{y_{\cdot\cdot}} = \mu + \overline{e_{\cdot\cdot}}$$

We assume the following,

$$\begin{aligned} e_{ij} &\sim N(0, \sigma^2) \\ \overline{e_{i\cdot}} &\sim N\left(0, \frac{\sigma^2}{h}\right) \\ \overline{e_{\cdot j}} &\sim N\left(0, \frac{\sigma^2}{k}\right) \\ \overline{e_{\cdot\cdot}} &\sim N\left(0, \frac{\sigma^2}{N}\right) \end{aligned}$$

We now find expectation of error ss.

$$\begin{aligned} E[SS_{err}] &= E\left[\sum_i \sum_j (y_{ij} + \overline{y_{i\cdot}} + \overline{y_{\cdot j}} + \overline{y_{\cdot\cdot}})^2\right] \\ &= E\left[\sum_i \sum_j \left(y_{ij} - (\mu + \alpha_i + \overline{e_{i\cdot}}) - (\mu + \beta_j - \overline{e_{\cdot j}}) + (\mu + \overline{e_{\cdot\cdot}})\right)^2\right] \\ &= E\left[\sum \sum (e_{ij} + \overline{e_{i\cdot}} + \overline{e_{\cdot j}} + \overline{e_{\cdot\cdot}})^2\right] \end{aligned}$$

All the product terms are zero,

$$\begin{aligned}
&= E[\sum_i \sum_j eij^2 - h\sum_i \overline{e_i.}^2 - k\sum_j \overline{e.j}^2 + N\overline{e..}^2] \\
&= N\sigma^2 - hk\frac{\sigma^2}{k} - kh\frac{\sigma^2}{h} + N\frac{\sigma^2}{N} \\
&= (N - h - k + 1)\sigma^2 \\
&= (kh - h - k + 1)\sigma^2 \\
E[SS_{err}] &= (h-1)(k-1)\sigma^2
\end{aligned}$$

So we also have then

$$\begin{aligned}
E\left[\frac{SS_{err}}{(k-1)(h-1)}\right] &= \sigma^2 \\
E[MSSE] &= \sigma^2
\end{aligned}$$

Therefore MSSE is unbiased estimator of σ^2.

Now we will find expectation of ss due to factor A,

$$\begin{aligned}
E[SS_A] &= E\left[\sum_i h\,(\overline{y_{i\cdot}} - \overline{y_{\cdot\cdot}})^2\right] \\
&= E\left[\sum_i h\,(\mu + \alpha_i + \overline{e_{i\cdot}} - \mu + \overline{e_{\cdot\cdot}})^2\right] \\
&= E\left[\sum_i h\,(\alpha_i + \overline{e_{i\cdot}} + \overline{e_{\cdot\cdot}})^2\right] \\
&= E\left[\sum_i h\left(\alpha_i^2 + (\overline{e_{i\cdot}} + \overline{e_{\cdot\cdot}})^2 + 2\alpha_i(\overline{e_{i\cdot}} + \overline{e_{\cdot\cdot}})\right)\right] \\
&= E\left[\sum_i h\alpha_i^2 + h\sum_i(\overline{e_{i\cdot}} - \overline{e_{\cdot\cdot}})^2 + 2\sum_i \alpha_i(\overline{e_{i\cdot}} - \overline{e_{\cdot\cdot}})\right] \\
&= \sum_i h\alpha_i^2 + E\left[h\sum_i(\overline{e_{i\cdot}} - \overline{e_{\cdot\cdot}})^2\right] + 2\sum_i \alpha_i E[\overline{e_{i\cdot}} - \overline{e_{\cdot\cdot}}] \\
&= h\sum_i \alpha_i^2 + E\left[h\left(\sum_i \overline{e_i}^2 + \sum_i \overline{e_{\cdot\cdot}}^2 - 2\sum_i \overline{e_{i\cdot}\overline{e_{\cdot\cdot}}}\right)\right] \\
&= h\sum_i \alpha_i^2 + h\sum_i E[\overline{e_{i\cdot}}^2] + hkE[\overline{e_{\cdot\cdot}}^2] - 2khE[\overline{e_{\cdot\cdot}}^2] \\
&= h\sum_i \alpha_i^2 + kh\frac{\sigma^2}{h} + hk\frac{\sigma^2}{N} - 2N\frac{\sigma^2}{N} \\
&= h\sum_i \alpha_i^2 + (k-1)\sigma^2
\end{aligned}$$

Therefore we have $E[MSS_A] = \frac{h}{k-1}\sum_i \alpha_i^2 + \sigma^2$.
So this will be a unbiased estimator of σ^2 when H_{0_α} is true, i.e. $\sum \alpha_i^2 = 0$.

Now we will find expectation of ss due to factor B in the exact same

manner,

$$\begin{aligned}
E[SS_B] &= E\left[\sum_j k\left(\overline{y_{i\cdot}} - \overline{y_{\cdot\cdot}}\right)^2\right] \\
&= E\left[\sum_j k\left(\mu + \beta_j + \overline{e_{\cdot j}} - \mu + \overline{e_{\cdot\cdot}}\right)^2\right] \\
&= E\left[\sum_j k\left(\beta_j + \overline{e_{\cdot j}} + \overline{e_{\cdot\cdot}}\right)^2\right] \\
&= E\left[\sum_j k\left(\beta_j^2 + (\overline{e_{\cdot j}} + \overline{e_{\cdot\cdot}})^2 + 2\beta_j(\overline{e_{\cdot j}} + \overline{e_{\cdot\cdot}})\right)\right] \\
&= E\left[\sum_j k\beta_j^2 + k\sum_j(\overline{e_{\cdot j}} - \overline{e_{\cdot\cdot}})^2 + 2\sum_j \beta_j(\overline{e_{\cdot j}} - \overline{e_{\cdot\cdot}})\right]
\end{aligned}$$

I can't be bothered to type this out again its the exact same f thing just do it.

$$= k\sum_j \beta_j^2 + (k-1)\sigma^2$$

So now we have $E[MSS_B] = \frac{k}{h-1}\sum_j \beta_j^2 + \sigma^2$. So we have that MSS_B is an unbiased estimator of σ^2 when H_{0_β} is true i.e. $\sum_j \beta_j^2 = 0$.

1.4.10 Test statistics

Total s.s. = s.s. due to factor A + s.s. due to factor B + SSE $N - 1 = (k-1) + (h-1) + (k-1)(h-1)$

Therefore by Cochran's theorem,

$$\begin{aligned}
\chi_1^2 &= \frac{SS_A}{\sigma^2} \sim \chi^2_{k-1} \\
\chi_2^2 &= \frac{SS_B}{\sigma^2} \sim \chi^2_{h-1} \\
\chi_3^2 &= \frac{SS_{err}}{\sigma^2} \sim \chi^2_{(k-1)(h-1)}
\end{aligned}$$

And we have the test statistics as follows,

$$F_1 = \frac{\chi_1^2/(k-1)}{\chi_3^2/(h-1)(k-1)}$$
$$F_2 = \frac{\chi_2^2/(h-1)}{\chi_3^2/(h-1)(k-1)}$$

Also $E[MSS_A] > E[MSSE]$ then H_{0_1} is rejected.
$E[MSS_B] > E[MSSE]$ then H_{0_2} is rejected.

1.4.11 Two way ANOVA Table

Replace R with T below here,

Source of variance	$d.f.$	$S.S.$	$S.S.S$	F_{cal}
Between factor A	$k-1$	$RSS = \sum_i^k \frac{R_i^2}{n_i} - \frac{G^2}{N}$	$MRSS = \frac{RSS}{k-1}$	$F_1 = \frac{MRSS}{MESS}$
Between factor B	$h-1$	$CSS = \sum_j^h \frac{R_j^2}{n_j} - \frac{G^2}{N}$	$MCSS = \frac{CSS}{h-1}$	$F_2 = \frac{MLSS}{MESS}$
Error	$(k-1)(h-1)$	$ESS = +$	$MESS = \frac{CSS}{(h-1)(k-1)}$	$-$
Total	$N-1$	$T_{ss} = \sum_i \sum_j y_{ij}^2 - \frac{G^2}{N}$	$-$	$-$

1.5 Critical difference (CD)

- When we reject H_0 we can conclude that the treatment mean are not all equal, but we cannot be more sure than this.
- We do not know whether all the means are different from one another or only some of them are different.
- Therefore, once we reject the null hypothesis, we conduct further tests to find out which of the means are actually different.
- This is achieved by comparison of means with the critical difference value.

1.5.1 Aim of Critical difference

- To determine the difference between a pair of means that will be significant.
- To compare that value with teh calculated differences between all pairs of group means.

If the difference between two means is greater than the critical difference it can be concluded that the difference between this pair of means is significant.

- In ANOVA, the null and alternative hypothesis are ....
- When H_0 is false then we have to find out those differences between the means which are significant.
- For this we calculate standard error of difference between means

1.5.2 Standard error of difference

Definition 1.11.

$$SE_d = \sqrt{EMS\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

- n_1, n_2 are the number of observations based on the number of replications of different means under comparisons, if $n_1 = n_2$ then

$$SE_d = \sqrt{\frac{2EMS}{r}}$$

- r =number of replications which is equal for all treatments under comparison
- The critical difference at desired level of significant is calculated and the observed difference are very easily compared with CD-value.

The CD value is calculated as follows,

$$CD = SE_d \cdot T_{tab}$$

Use t-table. **Least significant difference (LSD):** CD value at 5% LOS.

Example 1.12. *Five variates of feeds were equally and randomly allotted to 20 pigs of same sex and approximately of the same body weight. The gain body weight recorded in the feeding experiment were as given in the Table. Analyse data for comparison of feeds.*

Source of Variation	d.f.	s.s.	m.s.s.	F_{cal}	F_{tab}	
					5%	1%
Between feeds	4	155.2	38.8	11.18	3.06	4.89
Within feeds	15	52.0	3.47			
Total	19	207.2				

Proof.

- However, the F-test is significant even if any of two means differ.
- Thus, when a significant effect is found using ANOVA, we don't know which means actually differ significantly from each other.
- Compute $SE_d = \frac{2EMS}{r} = \sqrt{\frac{2(3.47)}{4}} = 1.317$. And we get $CD = 2.80$
- For comparison of means under different variatees of feeds, we arrange means in descending order of magnitude as follows.
  - $F_3 = 16$
  - $F_4 = 12$
  - $F_2 = 11$
  - $F_5 = 9$
  - $F_1 = 8$
- Compute the differences now and we get that $(F_4, F_3), (F_3, F_2), (F_3, F_5), (F_3, F_1), \cdots$ have significant differences, if difference is greater than CD.
- To make this easily interpretable we add same superscript over all those means which do not differ from each other in the following way

In our example we could see the following conclusions

- F_3 gave the maximum weight gain and was significantly different from all other means.
- F_4, F_2 did not differ from each other.
- Similarly, there were no significant difference between feed variates F_2, F_5 are also between F_5, F_1.
- Therefore, F_3 is a recommended feed.

□

Chapter 2

Design of Experiments

Ratio of design D_1 with D_2 are the ratio of their error variance $E = \frac{\sigma_2^2}{\sigma_1^2}\frac{n_1}{n_2}$

2.1 Completely randomized design

2.1.1 Statistical analysis

y_{ij} denotes response measurement of j^{th} experiment unit receiving i^{th} treatment for $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n$.

Mathematical model is the same as that of ANOVA one way classification.

$y_{ij} \sim N(\mu + \alpha_i, \sigma^2)$ $\mu = \frac{\sum_i^k n_i \mu_i}{\sum n_i}$ TSS=SST+SSE $TSS = \sum(y_{ij} - \overline{y_{\cdot\cdot}})^2$ with $N-1$d.f. $SST = \sum n_i(y)$ **complete this up later its same as anova**

2.1.2 Estimation of treatment constants

Let $c_1, c_2, \ldots, c_k$ be constants such that their sum is zero then the linear combination $\sum_i c_i\alpha_i$ is called treatment constant.

An estimate of $\sum_i c_i\alpha_i$ is $\sum_i c_i\overline{y_{i\cdot}}$,

So we have

$$V(\sum_i c_i\hat{\alpha}_i) = \sum c_i^2 \frac{\sigma^2}{n_i}$$

So estimate of standard error of $\sum c_i\hat{\alpha}_i$ is

$$\sqrt{MSE \sum_i \frac{c_i^2}{n_i}}$$

So the confidence intervals is

$$\left[\sum_i c_i\overline{y_{i\cdot}} - t_{(N-k,\alpha/2)}\sqrt{MSSE\sum\frac{c_i^2}{n_i}}, \overline{y_{i\cdot}} + t_{(N-k,\alpha/2)}\sqrt{MSSE\sum\frac{c_i^2}{n_i}}\right]$$

2.1.3 Advantages, disavantages and applications

Advantages

1. It is easy to layout the design
2. All experimental units can be used, i.e. it results in maximum use of experimental material.
3. Complete flexibility is allowed, any number of treatments and replicates may be used.
4. Relatively easy statistical analysis, even with variable replicates and variable experimental errors for different treatments.
5. Analysis remains simple when data are missing.
6. Provides the maximum number of degrees of freedom for error for a given number of experimental units and treatments.

Disadvantages

1. Relatively low accuracy due to lack of restrictions which allows environmental variation to enter experimental error.
2. Not suited for large number of treatments because a relatively large amount of experimental material is needed which increases the variation.

Applications

1. Under conditions where the experimental material is homogenous, i.e. laboratory, or green house experiments or growth chamber experiments.
2. Where a fraction of the experimental units is likely to be destroyed or fail to respond.
3. In small experiments where there is a small number of degrees of freedom.

2.2 Randomized block designs

Chapter 3

Latin square design (LSD)