

Bhoris Dhanjal

Analysis of Variance & Design of Experiments

Lecture Notes
for SSTA402

Contents

1	Analysis of Variance	1
	(Fixed effect models)	
1.1	Chi-squared distribution	1
1.1.1	Tests to use depending on type of data	1
1.2	Analysis of variance	2
1.3	One-way classification	3
1.3.1	Mathematical model	4
1.3.2	Layout of one way ANOVA	4
1.3.3	Assumptions in the model	5
1.3.4	Hypothesis to be tested	5
1.3.5	Estimating the parameters of the model	6
1.3.6	Degrees of freedom	8
1.3.7	Expectation of various sum of squares	8
1.3.8	Test statistics	10
1.3.9	Computations forms of s.s.	11
1.3.10	One way ANOVA Table	11
1.3.11	Example of One-way classification ANOVA	12
1.4	Two-way classification	12
1.4.1	Layout of two-way classification	12
1.4.2	Notation	13
1.4.3	Assumptions of two-way classification	13
1.4.4	Model of two-way classification	13
1.4.5	Hypothesis	14
1.4.6	Least square estimators of parameters	15
1.4.7	Degrees of freedom	16
2	Design of Experiments	17
3	Latin square design (LSD)	18

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

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

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 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 applied 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 to 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, \dots, x_n denote a random sample from normal population $N(0, \sigma^2)$. Let the sum of the squares of these values be written in the form*

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

Where Q_j is a quadratic form in x_1, x_2, \dots, x_n with rank (degrees of freedom) $r_j = 1, 2, \dots, k$ then the random variables Q_1, Q_2, \dots, 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 $y_{ij}, (i = 1, 2, \dots, k; j = 1, 2, \dots, n)$ of a random variable Y are grouped on some basis, into k classes of sizes n_1, n_2, \dots, n_k respectively, $(N = \sum_{i=1}^k n_i)$ as exhibited in below table.

Class	Sample observations				Total	Mean
1	y_{11}	y_{12}	\dots	y_{1n_1}	T_1	$\overline{y_1}$
2	y_{21}	y_{22}	\dots	y_{2n_1}	T_2	$\overline{y_2}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots	\vdots
k	y_{k1}	y_{k2}	\dots	y_{kn_k}	T_k	$\overline{y_k}$

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 assignable cause which can be detected and controlled by human endeavour and second type of variation 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 between the class means in view of the inherent variability within the separate 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, \dots, 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

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

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 $i = 1, \dots, k$,

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

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

1.3.2 Layout of one way ANOVA

Factor level	Observations	Total	Mean
1	$y_{11} y_{12} \dots y_{1n_1}$	$\sum_{j=1}^{n_1} y_{1j} = y_{1\cdot}$	$\frac{y_{1\cdot}}{n_1} = \bar{y}_{1\cdot}$
2	$y_{21} y_{22} \dots y_{2n_2}$	$\sum_{j=1}^{n_2} y_{2j} = y_{2\cdot}$	$\frac{y_{2\cdot}}{n_2} = \bar{y}_{2\cdot}$
\vdots	$\vdots \quad \vdots \quad \ddots \quad \vdots$	\vdots	\vdots
k	$y_{k1} y_{k2} \dots y_{kn_k}$	$\sum_j y_{kj} = y_{k\cdot}$	$\frac{y_{k\cdot}}{n_k} = \bar{y}_{k\cdot}$

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

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

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

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

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

$$\overline{y_{\cdot\cdot}} = \frac{1}{N} \sum_i \sum_j y_{ij} = \frac{1}{N} \sum_i n_i \overline{y_{i\cdot}} = \text{Overall mean}$$

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 $(\mu_1, \mu_2, \dots, \mu_k)$ are different.

1.3.5 Estimating the parameters of the model

The model is

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

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

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

□

Theorem 1.9. *Estimating α_i*

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

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

Substituting these in the equation

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

Observation = Grand mean + deviation due to i^{th} treatment + residual or error.

Subtracting $x\bar{y}_{\cdot\cdot}$ from both sides, squaring and summing over i, j .

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

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

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

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,

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

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

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

Expectation of s.s. due to factor

$$\begin{aligned}
E[s.s. \text{ due to factor}] &= E \left[\sum_i n_i (\bar{y}_i - \bar{y}_{..})^2 \right] \\
&= E \left[\sum_i n_i ((\mu + \alpha_i + \bar{\varepsilon}_{i.}) - (\mu + \bar{\varepsilon}_{..}))^2 \right]
\end{aligned}$$

Expand this out...

$$\begin{aligned}
&= \sum_i n_i \alpha_i^2 + E \left[\sum_i n_i (\bar{\varepsilon}_{i.} - \bar{\varepsilon}_{..})^2 \right] + 0 \\
E \left[\sum_i n_i (\bar{\varepsilon}_{i.} - \bar{\varepsilon}_{..})^2 \right] &= \sum_i n_i [E[\bar{\varepsilon}_{i.}^2] + E[\bar{\varepsilon}_{..}^2] - 2E[\bar{\varepsilon}_{i.} - \bar{\varepsilon}_{..}]] \\
&= \sum_i n_i \left[\frac{\sigma^2}{n_i} + \frac{\sigma^2}{N} - 2E[\bar{\varepsilon}_{i.} - \bar{\varepsilon}_{..}] \right] \\
&= k\sigma^2 + \sigma^2 - 2 \sum_i n_i (\bar{\varepsilon}_{i.} - \bar{\varepsilon}_{..}) \\
E[s.s. \text{ due to factors}] &= \sum_i n_i \alpha_i^2 + k\sigma^2 + \sigma^2 - 2 \sum_i n_i (\bar{\varepsilon}_{i.} - \bar{\varepsilon}_{..}) \\
\sum_i n_i (\bar{\varepsilon}_{i.} - \bar{\varepsilon}_{..}) &= E \left[\bar{\varepsilon}_{i.} \sum_i n_i (\bar{\varepsilon}_{i.} / n_i) \right] \\
&= E[\bar{\varepsilon}_{..} N \bar{\varepsilon}_{..}] \\
&= NE[\bar{\varepsilon}_{..}^2] = \frac{N\sigma^2}{N} = \sigma^2 \\
E[s.s. \text{ due to factors}] &= \sum_i n_i \alpha_i^2 + k\sigma^2 - \sigma^2 \\
&= \sum_i n_i \alpha_i^2 + (k-1)\sigma^2
\end{aligned}$$

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.

$$\begin{aligned}
E[\text{Error s.s.}] &= E \left[\sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2 \right] \\
&= E \left[\sum_i \sum_j (y_{ij} - (\mu + \alpha_i + \bar{e}_{i\cdot}))^2 \right]
\end{aligned}$$

magic

$$\begin{aligned}
&= E \left[\sum_i \sum_j (e_{ij} - \bar{e}_{i\cdot})^2 \right] \\
&= E \left[\sum_i \sum_j (e_{ij}^2) + \sum_i \sum_j (\bar{e}_{i\cdot}^2) - 2 \sum_i \sum_j (e_{ij} - \bar{e}_{i\cdot}) \right] \\
&= N\sigma^2 + k\sigma^2 - 2 \sum_i \sum_j E[e_{ij} - \bar{e}_{i\cdot}]
\end{aligned}$$

Consider now,

$$\begin{aligned}
\sum_i \sum_j E[e_{ij} - \bar{e}_{i\cdot}] &= \sum_i n_i E[\bar{e}_{i\cdot}^2] = \sum_i \frac{n_i \sigma^2}{n_i} = k\sigma^2 \\
E[\text{error s.s.}] &= N\sigma^2 + k\sigma^2 - 2k\sigma^2 \\
&= (N - k)\sigma^2
\end{aligned}$$

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 $\frac{s.s \text{ due to factor}}{\sigma^2}$, $\frac{s.s \text{ due to error}}{\sigma^2}$ follows chi squared with $(k-1)$ and $(N-k)$ degrees of freedom respectively.

Therefore, the test statistics are defined as follows

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

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.

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

Computational form of s.s. due to factors

$$\text{s.s. due to factor} = \sum_i n_i (\bar{y}_i - \bar{y}_{..})^2$$

Magic...

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

1.3.10 One way ANOVA Table

Replace R with T below here,

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

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, \dots, A_k according to some criterion A ; and into h categories, say B_1, B_2, \dots, B_h according to some criterion B having kh combinations (A_i, B_j) $i = 1, 2, \dots, k; j = 1, 2, \dots, h$; 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

$$\begin{aligned}
 N &= kh \\
 y_{..} &= \sum_i \sum_j y_{ij} = G \\
 \overline{y_{..}} &= \sum_i \sum_j \frac{y_{ij}}{N} = \frac{y_{..}}{kh}
 \end{aligned}$$

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

$$\begin{aligned}
 R_i &= y_{i.} = \sum_{j=1}^h y_{ij} \\
 \overline{y_{i.}} &= \frac{y_{i.}}{h} \\
 C_j &= y_{.j} = \sum_{i=1}^k y_{ij} \\
 \overline{y_{.j}} &= \frac{y_{.j}}{k}
 \end{aligned}$$

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

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

Here, e_{ij} is effect due to chance causes. $E[y_{ij}] = \mu_{ij}$ = fixed effect due to assignable causes where y_{ij} are independent $N(\mu_{ij}, \sigma^2)$.

μ_{ij} is further split into the following 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 , $\alpha_i = \mu_i - \mu, i = 1, 2, \dots, k$ and $\mu_{i\cdot} = \sum_j \frac{\mu_{ij}}{h}, i = 1, 2, \dots, k$. Sum of α_i is zero.

Proof.

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

□

3. β_j effect due to j^{th} level of factor B . $\beta_j = \mu_{\cdot j} - \mu; j = 1, \dots, h$ 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

$$\gamma_{ij} = \mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} - \mu$$

And its summation across j and i is zero.

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

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 $i = 1, \dots, k; j = 1, \dots, h$

1.4.5 Hypothesis

Null hypothesis

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

$$\begin{aligned} \mu_{1\cdot} &= \mu_{2\cdot} = \dots = \mu_{k\cdot} = \mu \\ \alpha_1 &= \alpha_2 = \dots = \alpha_k = 0 \end{aligned}$$

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

$$\begin{aligned} \mu_{\cdot 1} &= \mu_{\cdot 2} = \dots = \mu_{\cdot k} = \mu \\ \beta_1 &= \beta_2 = \dots = \beta_h = 0 \end{aligned}$$

Alternative hypothesis**Complete this****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

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

Differentiate it with respect to β_j

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

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

$$\begin{aligned}
 y_{ij} &= \mu + \alpha_i + \beta_j + e_{ij} \\
 &= \bar{y}_{..} + \bar{y}_{i.} - \bar{y}_{..} + \bar{y}_{.j} - \bar{y}_{..} + e_{ij} \\
 &= \bar{y}_{i.} + \bar{y}_{.j} - \bar{y}_{..} + e_{ij} \\
 \implies e_{ij} &= y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}
 \end{aligned}$$

The model is then,

$$y_{ij} = \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})$$

Subtracting $\bar{y}_{..}$ from both the sides squaring both sides and summing over i and j .

$$\begin{aligned}
 \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 &= \sum_i \sum_j [(\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})]^2 \\
 &= \sum_i \sum_j (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_i \sum_j (\bar{y}_{.j} - \bar{y}_{..})^2 + \sum_i \sum_j (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2
 \end{aligned}$$

All cross product terms vanish

$$\text{Total s.s.} = \text{Factor a.s.s.} + \text{Factor b.s.s.} + \text{Error s.s.}$$

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. $N - k - h + 1 = (k - 1)(h - 1)$

Chapter 2

Design of Experiments

Chapter 3

Latin square design (LSD)