Design and Analysis of Experiments

03 - Single factor Analysis of Variance

Ahmed A.(Msc)

DIRE DAWA UNIVERSITY
COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES
DEPARTMENT OF SATATISTICS

October 9, 2024

Outline

Review of Lecture 2: Comparing Two Conditions or Treatments

In **Lecture 2**, we discussed methods for comparing two conditions or treatments.

- For example, the Portland cement tension bond experiment involved two different mortar formulations.
- Another way to describe this experiment is as a single-factor experiment with two levels of the factor:
  - The factor is mortar formulation.
  - The two levels are the two different formulation methods.
- Many experiments of this type involve more than two levels of the factor.
- This chapter focuses on methods for the design and analysis of single-factor experiments with an arbitrary number *a* levels of the factor (or treatments).
- We assume that the experiment has been completely randomized.

Single-Factor Experiment: Two Levels (Mortar Formulation)

Treatment (Level)	Observations	Totals y	Averages $\bar{y}$
1 (M Mortar)	$y_{11} = 16.85, \ldots, y_{1n}$	$y_{1\cdot} = \sum_{j=1}^{10} y_{1j}$	$\bar{y}_1 = \frac{y_{1\cdot}}{10}$
2 (Unm Mortar)	$y_{21} = 16.62, \ldots, y_{2n}$	$y_{2\cdot} = \sum_{j=1}^{10} y_{2j}$	$\bar{y}_2 = \frac{y_{2\cdot}}{10}$

Table: Typical Data for a Single-Factor Experiment (Mortar Formulation)

Single factor Analysis of Variance

Example: RF Power vs. Etch Rate

Objective: To Investigate the relationship between RF power and etch rate using a single-factor experiment. The goal is to find the power setting that achieves the target etch rate.

Experimental Design:

- **Factor:** RF Power
- **Levels:** 4 (160, 180, 200, 220 W)
- **Replicates:** 5 wafers per power level
- **Total Runs:** 20 (randomized)

Example...

Randomization:

- Use a spreadsheet (e.g., Excel) to list the 20 runs.
- Generate a column of random numbers using the `RAND()` function.
- Sort the run order by this column to ensure randomization.

Example Run Order Generation:

1. Enter RF power levels and replicates in rows.
2. Generate random numbers with `=RAND()` in an adjacent column.
3. Sort all data by the random number column.

Example...

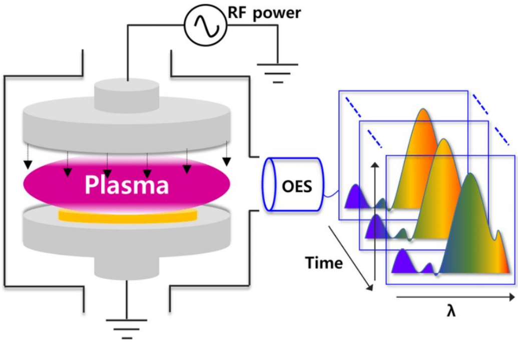

Figure: 3.1 A single-wafer plasma etching tool

Example...

Test Sequence	Excel Random Number (Sorted)	Power
1	12417	200
2	18369	220
3	21238	220
4	24621	160
5	29337	160
6	32318	180
7	36481	200
8	40062	160
9	43289	180
10	49271	200
11	49813	220
12	52286	220
13	57102	160
14	63548	160
15	67710	220
16	71834	180
17	77216	180
18	84675	180
19	89323	200
20	94037	200

Importance of Randomization in Experimental Design

Problem with Non-Randomized Runs:

- Running tests in a nonrandomized order (e.g., all 160 W tests first, then 180 W, etc.) can introduce **nuisance variables**.
- One such nuisance variable could be a **warm-up effect** in the etching tool, where prolonged tool usage lowers the etch rate.
- If the warm-up effect occurs, it could systematically distort the results, leading to invalid conclusions.

Solution: Randomization

- By running the 20 wafers in a random order, the effects of the warm-up are spread out across all RF power levels.
- This randomization prevents the warm-up effect from contaminating the results for any one power setting.

Example...

Graphical Examination of Data:

- **Box Plots:** A box plot for etch rates at each RF power level can help visualize variation and central tendency.
- **Scatter Plot:** A scatter plot of etch rate vs. RF power shows the trend and possible linear relationships.
- Both graphs suggest that etch rate increases as RF power increases.

Conclusion:

- Randomization is critical in experimental design to avoid biases introduced by uncontrolled variables, such as equipment effects or environmental changes.
- Visualizing data with box plots and scatter plots helps identify trends and relationships in the experiment.

Example...

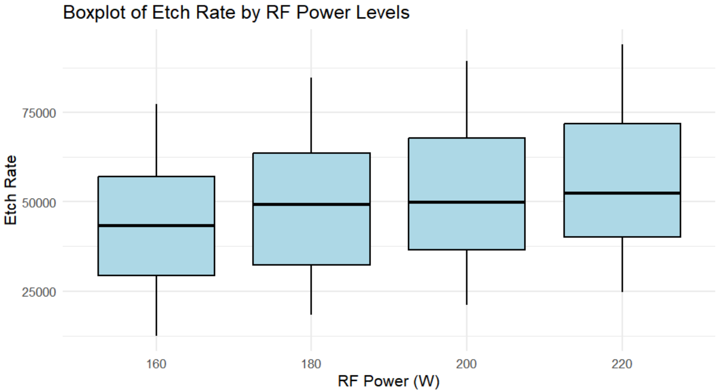

Figure: 3.2 Box plots of the etch rate data

Example...

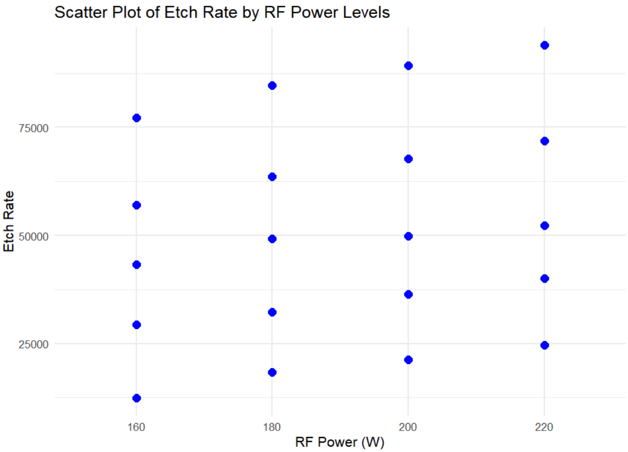

Figure: 3.3 scatter diagram of the etch rate data

Example...

■ **TABLE 3.1**

Etch Rate Data (in Å/min) from the Plasma Etching Experiment

Power (W)	Observations 1	2	3	4	5	Totals	Averages
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

Figure: 3.3 scatter diagram of the etch rate data

- Suppose we have a treatments or different levels of a single factor that we wish to compare.
- The observed response from each of the a treatments is a random variable. The data would appear as in Table 3.2.
- An entry in Table 3.2 (e.g.,$y_i j$) represents the jth observation taken under factor level or treatment i. There will be, in general, n observations under the ith treatment.

The analysis of variance

■ **TABLE 3.2**

Typical Data for a Single-Factor Experiment

Treatment (Level)	Observations				Totals	Averages
1	y_{11}	y_{12}	$\cdots$	y_{1n}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	$\cdots$	y_{2n}	$y_{2.}$	$\bar{y}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	y_{a1}	y_{a2}	$\cdots$	y_{an}	$y_{a.}$	$\bar{y}_{a.}$
					$y_{..}$	$\bar{y}_{..}$

The analysis of variance

- We will describe observations from an experiment with a model
- Two main models: means model and effects model
- Both are linear statistical models
- We'll focus on one-way or single-factor ANOVA

Means Model

The means model is given by:

$$y_{ij} = \mu_i + \epsilon_{ij} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n \end{cases} \tag{3.1}$$

Where:

- y_{ij} is the ijth observation
- μ_i is the mean of the ith factor level or treatment
- ϵ_{ij} is a random error component

Effects Model

The effects model is derived by defining $\mu_i = \mu + \tau_i$:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n \end{cases} \tag{3.2}$$

Where:

- μ is the overall mean
- τ_i is the ith treatment effect

This Model is called the **one-way or single-factor analysis of variance**(**ANOVA**) model because only one factor is investigated.

Model Assumptions

For hypothesis testing:

- Errors are normally and independently distributed
- Errors have mean zero and variance σ^2
- Variance σ^2 is constant for all factor levels
- Observations are mutually independent

This implies:

$$y_{ij} \sim N(\mu + \tau_i, \sigma^2)$$

Fixed vs Random Factor

Two scenarios:

1. Fixed Effects Model
    - Treatments specifically chosen by experimenter
    - Conclusions apply only to factor levels considered
    - Estimate model parameters (μ, τ_i, σ^2)
2. Random Effects Model
    - Treatments are a random sample from a larger population
    - Conclusions can be extended to all treatments in the population

Analysis of the Fixed Effects Model

In this section, we develop the single-factor analysis of variance for the fixed effects model.

Notation

- $y_{i.}$: total of observations under the ith treatment
- $\bar{y}_{i.}$: average of observations under the ith treatment
- $y_{..}$: grand total of all observations
- $\bar{y}_{..}$: grand average of all observations

Mathematical Expressions

$$
\begin{aligned}
y_{i.} &= \sum_{j=1}^{n} y_{ij} \\
\bar{y}_{i.} &= \frac{y_{i.}}{n} \quad i = 1, 2, \ldots, a \\
y_{..} &= \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij} \\
\bar{y}_{..} &= \frac{y_{..}}{N}
\end{aligned} \tag{3.3}
$$

where $N = an$ is the total number of observations.
The "dot" subscript notation implies summation over the subscript that it replaces.

Hypotheses for Treatment Means

We are testing the equality of a treatment means: $E(y_{ij}) = \mu + \tau_i = \mu_i$, $i = 1, 2, \ldots, a$

$$\begin{aligned} &H_0 : \mu_1 = \mu_2 = \cdots = \mu_a \\ &H_1 : \mu_i \neq \mu_j \text{ for at least one pair } (i, j) \end{aligned} \tag{3.4}$$

Effects Model Breakdown

In the effects model:

- $\mu_i = \mu + \tau_i$
- μ is the overall mean
- $\sum_{i=1}^{a} \mu_i = a\mu$
- $\sum_{i=1}^{a} \tau_i = 0$

Treatment effects τ_i can be thought of as deviations from the overall mean.

Equivalent Hypotheses

An equivalent way to write the hypotheses in terms of treatment effects τ_i:

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0$$
$$H_1 : \tau_i \neq 0 \text{ for at least one } i$$

Equivalent Hypotheses

- We test the equality of treatment means
- Alternatively, we test if treatment effects are zero
- The appropriate procedure for testing is the analysis of variance

Decomposition of the Total Sum of Squares

- Analysis of Variance (ANOVA) partitions total variability into component parts
- We use the total corrected sum of squares as a measure of overall variability

Total Corrected Sum of Squares

The total corrected sum of squares (SST) is defined as:

$$SST = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 \tag{3.5}$$

where:

- a is the number of treatments
- n is the number of replicates per treatment
- y_{ij} is the jth observation in the ith treatment
- $\bar{y}_{..}$ is the grand mean

Decomposition of SST

SST can be decomposed as:

$$\begin{aligned} SST &= \sum_{i=1}^{a}\sum_{j=1}^{n}[(y_{i.} - \bar{y}_{..}) + (y_{ij} - y_{i.})]^2 \\ &= n\sum_{i=1}^{a}(y_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - y_{i.})^2 \end{aligned} \tag{3.6}$$

Fundamental ANOVA Identity

Equation 3.6 is the fundamental ANOVA identity:

$$SST = SS_{Treatments} + SSE$$

where:

- $SS_{Treatments}$ is the sum of squares due to treatments (between treatments)
- SSE is the sum of squares due to error (within treatments)

Degrees of Freedom

- Total: $N - 1$ (where $N = an$)
- Treatments: $a - 1$
- Error: $a(n-1) = an - a = N - a$

Error Sum of Squares

The error sum of squares is:

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - y_{i.})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2$$

Interpretation

- $SS_{Treatments}$: Measures differences between treatment means
- SSE: Measures differences due to random error within treatments
- The decomposition allows us to compare these sources of variation

Sample Variance Within Treatmentsn

We will examine:

- Sample variance within treatments
- Pooled estimate of common variance
- Mean squares and their expected values

Sample Variance Within Treatments

The sample variance in the *i*th treatment is:

$$S_i^2 = \frac{\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2}{n-1}, \quad i = 1, 2, \ldots, a$$

Pooled Estimate of Common Variance

Combining a sample variances:

$$\frac{\sum_{i=1}^{a}(n-1)S_i^2}{(n-1)+(n-1)+\cdots+(n-1)} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2}{a(n-1)} = \frac{SSE}{N-a}$$

$SSE/(N-a)$ is a pooled estimate of the common variance within each of the a treatments.

Estimating σ^2 from Treatment Means

If there were no differences between the a treatment means:

$$\frac{\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2}{a-1}$$

estimates σ^2 if the treatment means are equal.

Intuitive Explanation

- $\frac{\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2}{a-1}$ estimates σ^2/n
- So $n\frac{\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2}{a-1}$ estimates σ^2
- This holds if there are no differences in treatment means

Mean Squares

We define two mean squares:

$$MS_{Treatments} = \frac{SS_{Treatments}}{a-1} = \frac{n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2}{a-1}$$
$$MSE = \frac{SSE}{N-a}$$

Expected Value of MSE

Consider:

$$E(MSE) = E\left[\frac{1}{N-a}\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2\right]$$
$$= \frac{1}{N-a}E\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}^2 - 2y_{ij}\bar{y}_{i.} + \bar{y}_{i.}^2)\right]$$

$$= \frac{1}{N-a}E\left[\sum_{i=1}^{a}\sum_{j=1}^{n}y_{ij}^2 - 2n\sum_{i=1}^{a}\bar{y}_{i.}^2 + n\sum_{i=1}^{a}\bar{y}_{i.}^2\right]$$
$$= \frac{1}{N-a}E\left[\sum_{i=1}^{a}\sum_{j=1}^{n}y_{ij}^2 - n\sum_{i=1}^{a}\bar{y}_{i.}^2\right]$$

Expected Value of MSE...

- Substituting the model (Equation 3.1) into this equation, we obtain:

$$E(MSE) = \frac{1}{N-a} E\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(\mu + \tau_i + \epsilon_{ij})^2 - \frac{1}{n}\sum_{i=1}^{a}\left(\sum_{j=1}^{n}\mu + \tau_i + \epsilon_{ij}\right)^2\right]$$

- When squaring and taking expectation, terms involving ϵ and τ are replaced by σ^2 and $n\tau^2$, respectively, because $E(\epsilon_{ij}) = 0$.
- All cross products involving ϵ_{ij} have zero expectation.

Expectation of Mean Squares

- After squaring and taking expectation, we get:

$$E(MSE) = \frac{1}{N-a}\left(N\sigma^2 + n\sum_{i=1}^{a}\tau_i^2 + N\sigma^2 - N\mu^2 - n\sum_{i=1}^{a}\tau_i^2 - a\sigma^2\right)$$

$$E(MSE) = \sigma^2$$

- Similarly, we can show that:

$$E(MS_{\text{Treatments}}) = \sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a-1}$$

- If there are no differences in treatment means ($\tau_i = 0$), $MS_{\text{Treatments}}$ also estimates σ^2.
- If treatment means differ, the expected value of the treatment mean square is greater than σ^2.

Statistical Analysis

- We can test the hypothesis of no difference in treatment means by comparing $MS_{\text{Treatments}}$ and MSE.
- Null hypothesis:
  $H_0 : \mu_1 = \mu_2 = \ldots = \mu_a$, or equivalently, $H_0 : \tau_1 = \tau_2 = \ldots = \tau_a = 0$
- Assumptions:
  - Errors ϵ_{ij} are normally and independently distributed
  - Mean zero and variance σ^2

Distribution of Sum of Squares

- SST/σ^2 is distributed as chi-square with $N-1$ degrees of freedom
- SSE/σ^2 is chi-square with $N-a$ degrees of freedom
- If $H_0 : \tau_i = 0$ is true, $SS_{\text{Treatments}}/\sigma^2$ is chi-square with $a-1$ degrees of freedom
- Note: $SS_{\text{Treatments}}$ and SSE add up to SST

Cochran's Theorem

- Cochran's theorem is useful in establishing the independence of SSE and $SS_{\text{Treatments}}$
- This theorem helps in formulating the F-test for ANOVA
- The F-test compares the ratio of $MS_{\text{Treatments}}$ to MSE with the F-distribution

Cochran's Theorem

THEOREM 3-1
Cochran's Theorem

Let Z_i be NID(0, 1) for $i = 1, 2, \ldots, \nu$ and

$$\sum_{i=1}^{\nu} Z_i^2 = Q_1 + Q_2 + \cdots + Q_s$$

where $s \leq \nu$, and Q_i has ν_i degrees of freedom ($i = 1, 2, \ldots, s$). Then $Q_1, Q_2, \ldots, Q_s$ are independent chi-square random variables with $\nu_1, \nu_2, \ldots, \nu_s$ degrees of freedom, respectively, if and only if

$$\nu = \nu_1 + \nu_2 + \cdots + \nu_s$$

Cochran's Theorem and Degrees of Freedom

- The degrees of freedom for $SS_{\text{Treatments}}$ and SSE add to $N - 1$, the total number of degrees of freedom.
- Cochran's theorem implies that $SS_{\text{Treatments}}/\sigma^2$ and SSE/σ^2 are independently distributed chi-square random variables.

F-Distribution

- If the null hypothesis of no difference in treatment means is true, the ratio

$$F_0 = \frac{MS_{\text{Treatments}}}{MSE} \tag{3.7}$$

is distributed as F with $a-1$ and $N-a$ degrees of freedom.

- Equation (3.7) is the test statistic for the hypothesis of no differences in treatment means.

Expected Mean Squares

- MSE is an unbiased estimator of σ^2.
- Under the null hypothesis, $MS_{\text{Treatments}}$ is also an unbiased estimator of σ^2.
- If the null hypothesis is false, the expected value of $MS_{\text{Treatments}}$ is greater than σ^2.

Rejecting the Null Hypothesis

- Under the alternative hypothesis, the numerator of the test statistic is greater than the denominator.
- Therefore, we reject H_0 for large values of the test statistic.
- This implies an upper-tail, one-tail critical region:

$$F_0 > F(\alpha, a-1, N-a)$$

Computing Sums of Squares

- The sums of squares can be computed in several ways.
- One direct approach is:

$$SS_{\text{Treatments}} = \frac{1}{n}\sum_{i=1}^{a} y_{i}^2 - \frac{y_{\dots}^2}{N} \tag{3.8}$$

$$SST = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{\dots}^2}{N} \tag{3.9}$$

$$SSE = SST - SS_{\text{Treatments}} \tag{3.10}$$

Using a Spreadsheet

- A spreadsheet can be used to compute these terms for each observation.
- Sum the squares to obtain SST, $SS_{\text{Treatments}}$, and SSE.

ANOVA Table for the Single-Factor, Fixed Effects Model

TABLE 3.3

The Analysis of Variance Table for the Single-Factor, Fixed Effects Model

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Between treatments	$SS_{\text{Treatments}} = n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2$	$a-1$	$MS_{\text{Treatments}}$	$F_0 = \frac{MS_{\text{Treatments}}}{MS_E}$
Error (within treatments)	$SS_E = SS_T - SS_{\text{Treatments}}$	$N-a$	MS_E	
Total	$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2$	$N-1$		

ANOVA Example

- The engineer is interested in determining if the RF power setting affects the etch rate.
- A completely randomized experiment with four levels of RF power and five replicates was conducted.
- Data from the experiment is shown below (from Table 3.1). We will use the analysis of variance to test
- H0:$\mu_1 = \mu_2 = \mu_3 = \mu_4$ against the alternative
- H1: some means are different.
- The sums of squares required are computed as follows:

ANOVA Example

RF Power (W)	Observed Etch Rate (Å/min) 1	2	3	4	5	Totals $y_{i.}$	Averages $\bar{y}_{i.}$
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
						$y_{..} = 12{,}355$	$\bar{y}_{..} = 617.75$

Sums of Squares Calculation

$$SS_T = \sum_{i=1}^{4}\sum_{j=1}^{5} y_{ij}^2 - \frac{y_{..}^2}{N} = \left[(575)^2 + (542)^2 \cdots + (710)^2\right] - \frac{(12,355)^2}{20}$$

$$= 72,209.75$$

$$SS_{\text{Treatments}} = \frac{1}{n}\sum_{i=1}^{4} y_{i.}^2 - \frac{y_{..}^2}{N} \quad = \frac{1}{5}\left[(2756)^2 + \cdots + (3535)^2\right] - \frac{(12,355)^2}{20}$$

$$= 66,870.55$$

$$SS_E = SS_T - SS_{\text{Treatments}} \qquad = 72,209.75 - 66,870.55 = 5339.20$$

ANOVA Example

■ **TABLE 3.4**

ANOVA for the Plasma Etching Experiment

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-Value
RF Power	66,870.55	3	22,290.18	$F_0 = 66.80$	<0.01
Error	5339.20	16	333.70		
Total	72,209.75	19			

ANOVA Example

- We compute the F-ratio as:

$$F_0 = \frac{22,290.18}{333.70} = 66.80$$

- We compare this to the critical value from the $F_{3,16}$ distribution at $\alpha = 0.05$:

$$F_{0.05,3,16} = 3.24$$

- Since $F_0 = 66.80 > 3.24$, we reject the null hypothesis H_0 and conclude that the treatment means differ significantly.

Conclusion

- Since $F_0 = 66.80$ is much larger than the critical value 3.24, we reject H_0 and conclude that the RF power setting significantly affects the mean etch rate.
- The P-value for this test statistic is very small, indicating strong evidence against H_0.

Estimators for the Single-Factor Model

We now present estimators for the parameters in the single-factor model and confidence intervals on the treatment means.

- Reasonable estimates of the overall mean and the treatment effects are given by (Equation 3.11).
- The overall mean is estimated by the grand average of the observations:

$$\hat{\mu} = \bar{y}_{..}$$

- The treatment effect for the ith treatment is estimated as the difference between the treatment average and the grand average:

$$\hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..}, \quad i = 1, 2, \ldots, a \tag{3.11}$$

Confidence Interval on the ith Treatment Mean

- A point estimator of μ_i is $\hat{\mu}_i = \bar{y}_{i.}$.
- Assuming normally distributed errors, each treatment average is distributed as:

$$\bar{y}_{i.} \sim NID(\mu_i, \sigma^2/n)$$

- If σ^2 were known, we could use the normal distribution to define the confidence interval.
- Using MSE as an estimator of σ^2, we base the confidence interval on the t-distribution.
- A $100(1-\alpha)\%$ confidence interval on the ith treatment mean μ_i is given by:

$$\bar{y}_{i.} - t_{\alpha/2,N-a}\sqrt{\frac{MSE}{n}} \leq \mu_i \leq \bar{y}_{i.} + t_{\alpha/2,N-a}\sqrt{\frac{MSE}{n}} \tag{3.12}$$

Confidence Interval for the Difference Between Two Treatment Means

Differences in treatments are often of great practical interest.

- A $100(1-\alpha)\%$ confidence interval on the difference between two treatment means, $\mu_i - \mu_j$, is given by:

$$\bar{y}_{i.} - \bar{y}_{j.} - t_{\alpha/2,N-a}\sqrt{\frac{2MSE}{n}} \leq \mu_i - \mu_j \leq \bar{y}_{i.} - \bar{y}_{j.} + t_{\alpha/2,N-a}\sqrt{\frac{2MSE}{n}} \quad (3.13)$$

Unbalanced Single-Factor Experiments

- In some single-factor experiments, the number of observations taken within each treatment may differ.
- We refer to this design as **unbalanced**.
- The analysis of variance (ANOVA) described earlier can still be used, but slight modifications must be made in the sum of squares formulas.
- Let n_i represent the number of observations under treatment i $(i = 1, 2, \ldots, a)$, and $N = \sum n_i$ be the total number of observations.

Modified Sum of Squares Formulas

The sum of squares total (SS_T) and sum of squares for treatments ($SS_{\text{Treatments}}$) are adjusted as follows:

- The manual computational formula for the total sum of squares (SS_T) is:

$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{\cdot\cdot}^2}{N} \tag{3.14}$$

- The sum of squares for treatments ($SS_{\text{Treatments}}$) is:

$$SS_{\text{Treatments}} = \sum_{i=1}^{a} \frac{y_{i\cdot}^2}{n_i} - \frac{y_{\cdot\cdot}^2}{N} \tag{3.15}$$

unbalanced designs

- Other than the modifications in the sum of squares formulas, no further changes are required in the analysis of variance for unbalanced designs.
- The resulting analysis follows the same framework as balanced designs but accommodates unequal sample sizes.

Advantages of a Balanced Design

There are two main advantages of using a balanced design:

1. **Insensitivity to Variance Assumptions**:
   - When the sample sizes are equal across treatments, the test statistic is relatively insensitive to small departures from the assumption of equal variances across the a treatments.
   - This is not the case for **unequal sample sizes**, where unequal variances can have a larger impact on the test results.
2. **Maximized Power of the Test**:
   - The power of the test is **maximized** when the sample sizes are equal across treatments.
   - Equal sample sizes allow the test to more effectively detect differences in treatment means.

Model Adequacy Checking

The decomposition of the variability in the observations through an analysis of variance identity (Equation 3.6) is a purely algebraic relationship. However, the use of partitioning to test for no differences in treatment means requires that certain assumptions be satisfied.

- Observations are adequately described by the model.
- Errors are normally and independently distributed with mean zero and constant but unknown variance σ^2.

If these assumptions are valid, the ANOVA procedure is an exact test of the hypothesis of no difference in treatment means. However, in practice, these assumptions may not hold exactly.

Importance of Checking Assumptions

It is usually unwise to rely on ANOVA until the validity of these assumptions has been checked. Violations of the basic assumptions and model adequacy can be investigated by examining residuals.

$$e_{ij} = y_{ij} - \hat{y}_{ij} \tag{3.16}$$

where $\hat{y}_{ij}$ is the estimated observation obtained as:

$$\begin{aligned} \hat{y}_{ij} &= \hat{\mu} + \bar{\tau}_i \\ &= \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) \\ &= \bar{y}_{i.} \end{aligned} \tag{3.17}$$

Examination of the residuals should be an automatic part of any ANOVA. If the model is adequate, the residuals should be structureless and contain no obvious patterns.

Residuals and Model Inadequacies

Through the analysis of residuals, many types of model inadequacies and violations of the underlying assumptions can be discovered.

Graphical Analysis of Residuals:

- Residuals should be plotted to check for patterns.
- A structured pattern indicates model inadequacy.

Normality Assumption

A check of the normality assumption can be made by plotting a histogram of the residuals. If the $NID(0, \sigma^2)$ assumption on the errors is satisfied, the histogram should resemble a normal distribution centered at zero.

- With small samples, significant fluctuations may occur.
- Moderate departures from normality do not necessarily indicate serious violations.
- Gross deviations from normality are serious and require further analysis.

Normal Probability Plot

An extremely useful procedure is to construct a normal probability plot of the residuals.

- If the underlying error distribution is normal, this plot will resemble a straight line.
- Emphasize the central values of the plot more than the extremes.