MA5701: Statistical Methods

Chapter 6 : Inferences for Two or More Means

Kui Zhang, Mathematical Sciences

Example 6.1 – Soil Silt Content

- **Example 6.1** (Soil Silt Content) - A study was done to compare soil mapping units. Study area consists of eight sites and samples were obtained in each site at 11 random points. The soil property considered was the silt content, expressed as percentages of the total silt, clay, and sand content.
- Questions asked:
  - Is there a difference in silt content among the soils from different sites?
  - If there is a difference, can we identify the sites having the largest and smallest silt content?
- Possible solutions
  - ANOVA - analysis of variance.
  - Why not pair-wise comparison? – not as powerful as ANOVA.

Example 6.1 - Data

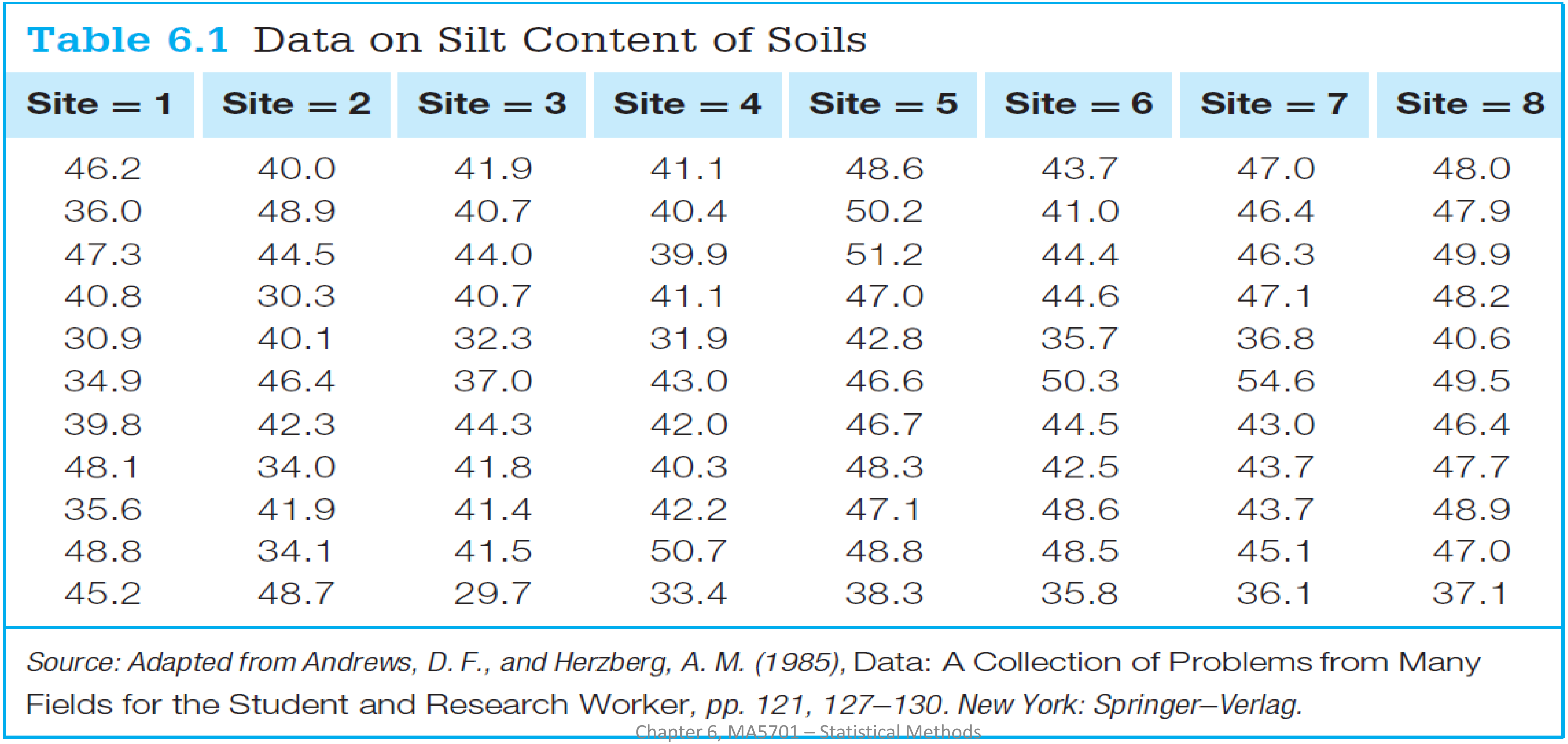
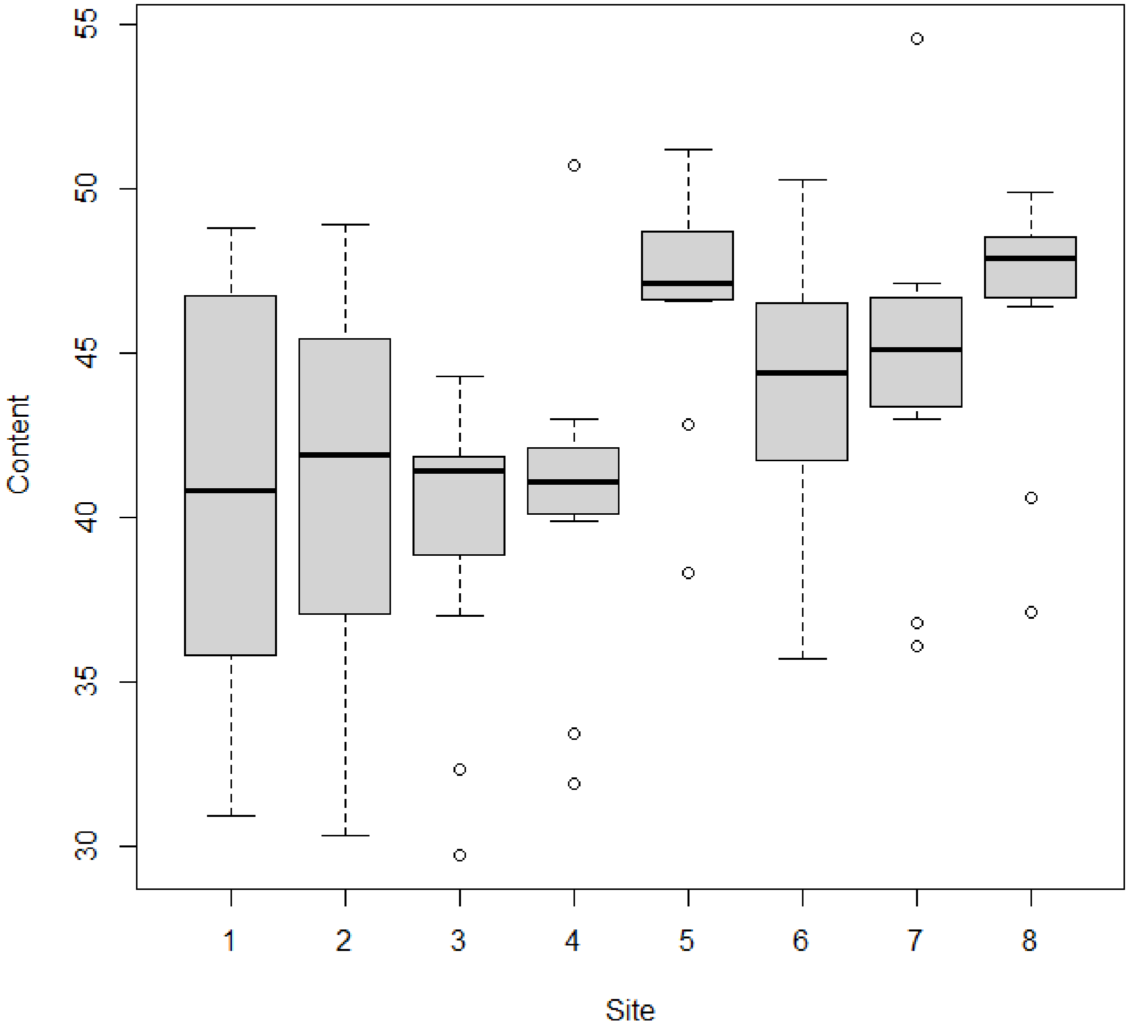


Table 6.1 Data on Silt Content of Soils

Site = 1	Site = 2	Site = 3	Site = 4	Site = 5	Site = 6	Site = 7	Site = 8
46.2	40.0	41.9	41.1	48.6	43.7	47.0	48.0
36.0	48.9	40.7	40.4	50.2	41.0	46.4	47.9
47.3	44.5	44.0	39.9	51.2	44.4	46.3	49.9
40.8	30.3	40.7	41.1	47.0	44.6	47.1	48.2
30.9	40.1	32.3	31.9	42.8	35.7	36.8	40.6
34.9	46.4	37.0	43.0	46.6	50.3	54.6	49.5
39.8	42.3	44.3	42.0	46.7	44.5	43.0	46.4
48.1	34.0	41.8	40.3	48.3	42.5	43.7	47.7
35.6	41.9	41.4	42.2	47.1	48.6	43.7	48.9
48.8	34.1	41.5	50.7	48.8	48.5	45.1	47.0
45.2	48.7	29.7	33.4	38.3	35.8	36.1	37.1

Source: Adapted from Andrews, D. F., and Herzberg, A. M. (1985), Data: A Collection of Problems from Many Fields for the Student and Research Worker, *pp. 121, 127–130. New York: Springer–Verlag.*

Example 6.1 – Box Plot



Introduction

- It is common to compare more than two means.
- The pooled t-test for comparing two means cannot be generalized to the comparison of more than two means.
- We will cover the **ANOVA** (analysis of variance) method.
- Discuss the assumptions necessary for the validity of the results from such an analysis and discuss alternative methods if these assumptions are not satisfied.
- Introduce procedures for specific comparisons among selected means.

Data Table 6.2

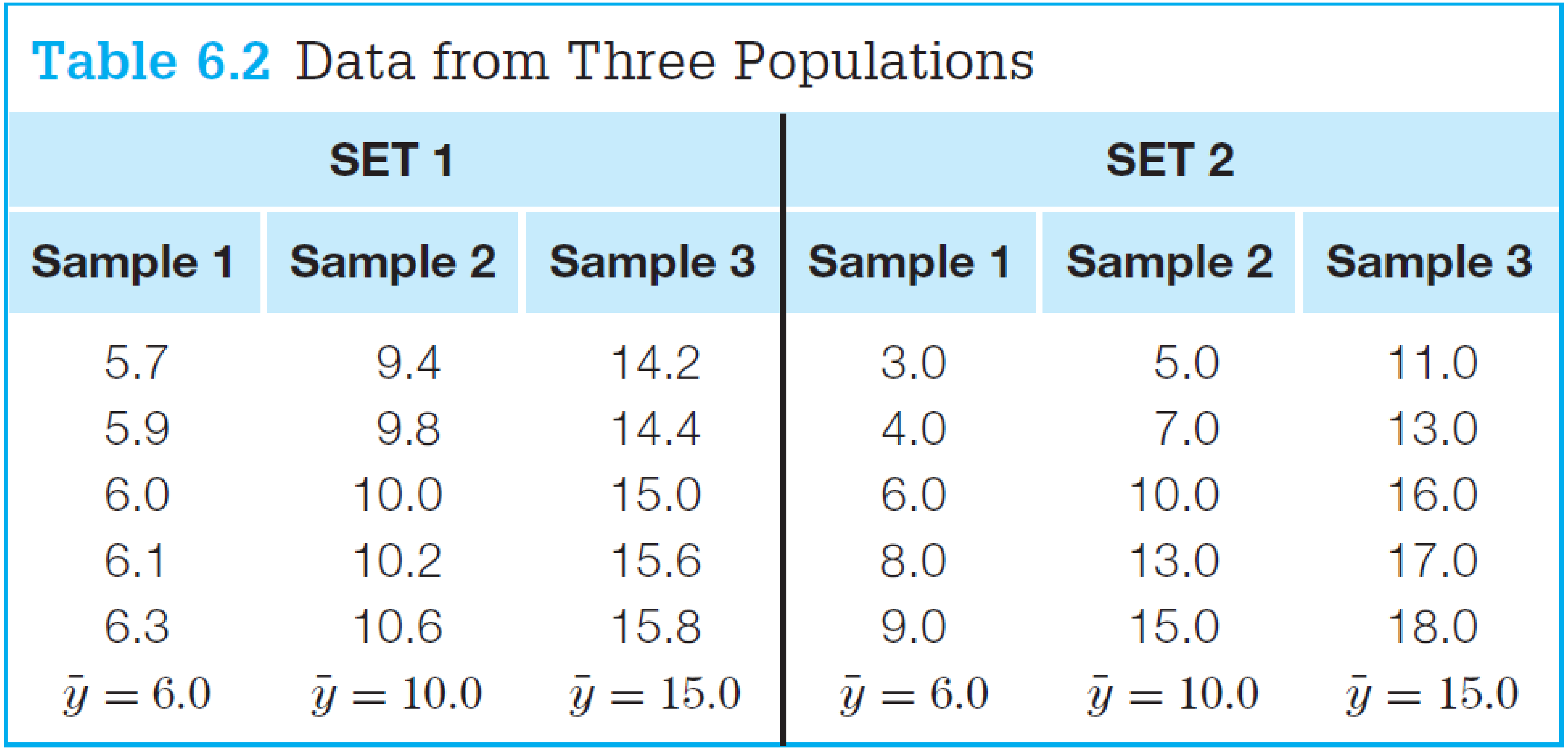
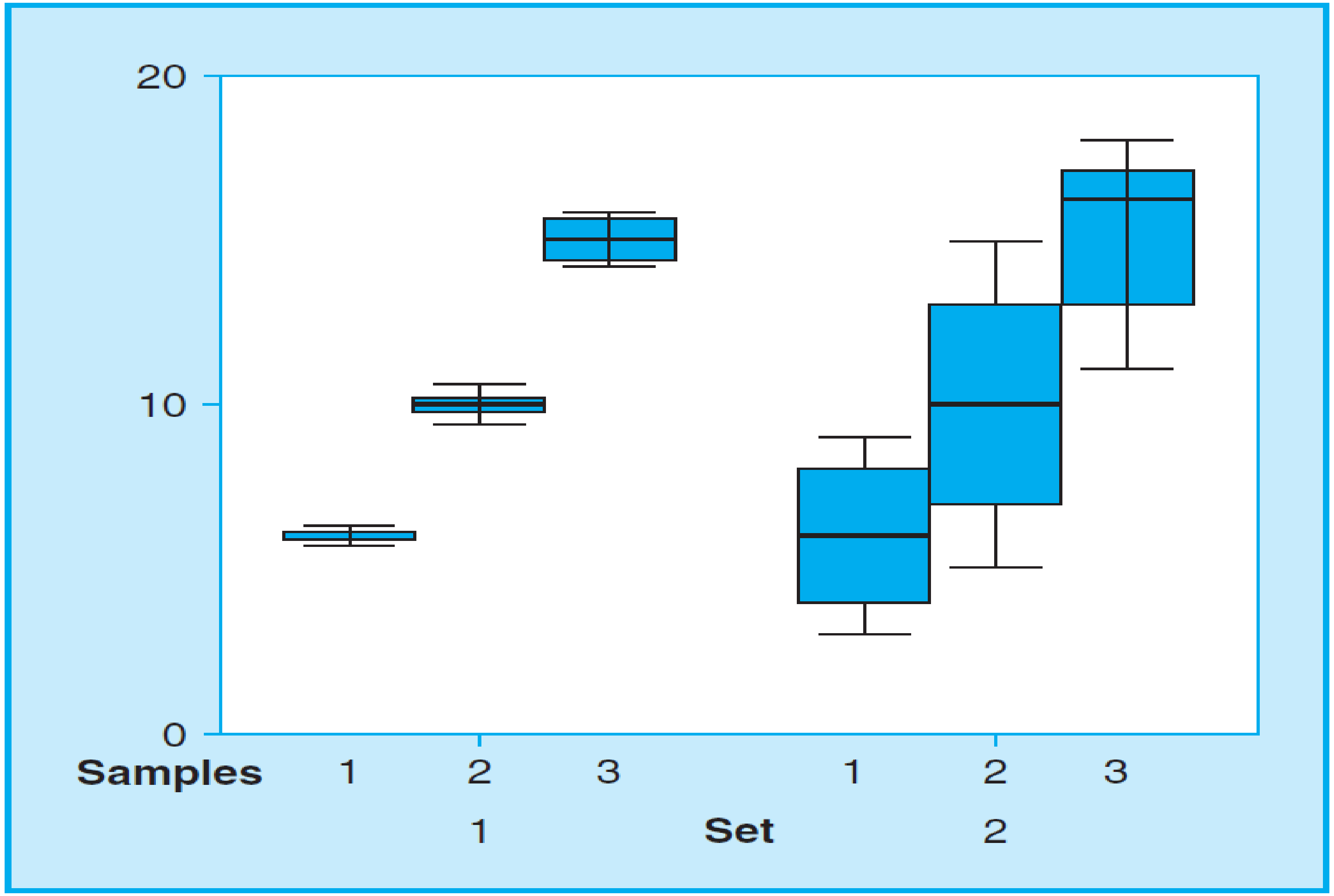


Table 6.2 Data from Three Populations

SET 1			SET 2		
Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
5.7	9.4	14.2	3.0	5.0	11.0
5.9	9.8	14.4	4.0	7.0	13.0
6.0	10.0	15.0	6.0	10.0	16.0
6.1	10.2	15.6	8.0	13.0	17.0
6.3	10.6	15.8	9.0	15.0	18.0
$\bar{y} = 6.0$	$\bar{y} = 10.0$	$\bar{y} = 15.0$	$\bar{y} = 6.0$	$\bar{y} = 10.0$	$\bar{y} = 15.0$

Box Plot for Data from Table 6.2



Analysis of Variance

- There is stronger evidence of differences among means in Set 1 than among means in Set 2.
- First, the variances ***among*** (***between***) the means for the two sets are identical. Here, the variance is the sample variance of 6, 10, and 15.
- Second, the observations ***within*** the samples are more closely bunched in Set 1 than they are in Set 2, and we know that sample means from populations with smaller variances will also be less variable. Thus the variance among the observations ***within*** the individual samples is smaller for Set 1.

Analysis of Variance

- Such observation is the basis for using the analysis of variance for making inferences about differences among means.
- The analysis of variance is based on the comparison of the **variance *among* (*between*) the means** of the populations to the variance among sample observations ***within*** the individual populations.

Notation for One-Way ANOVA

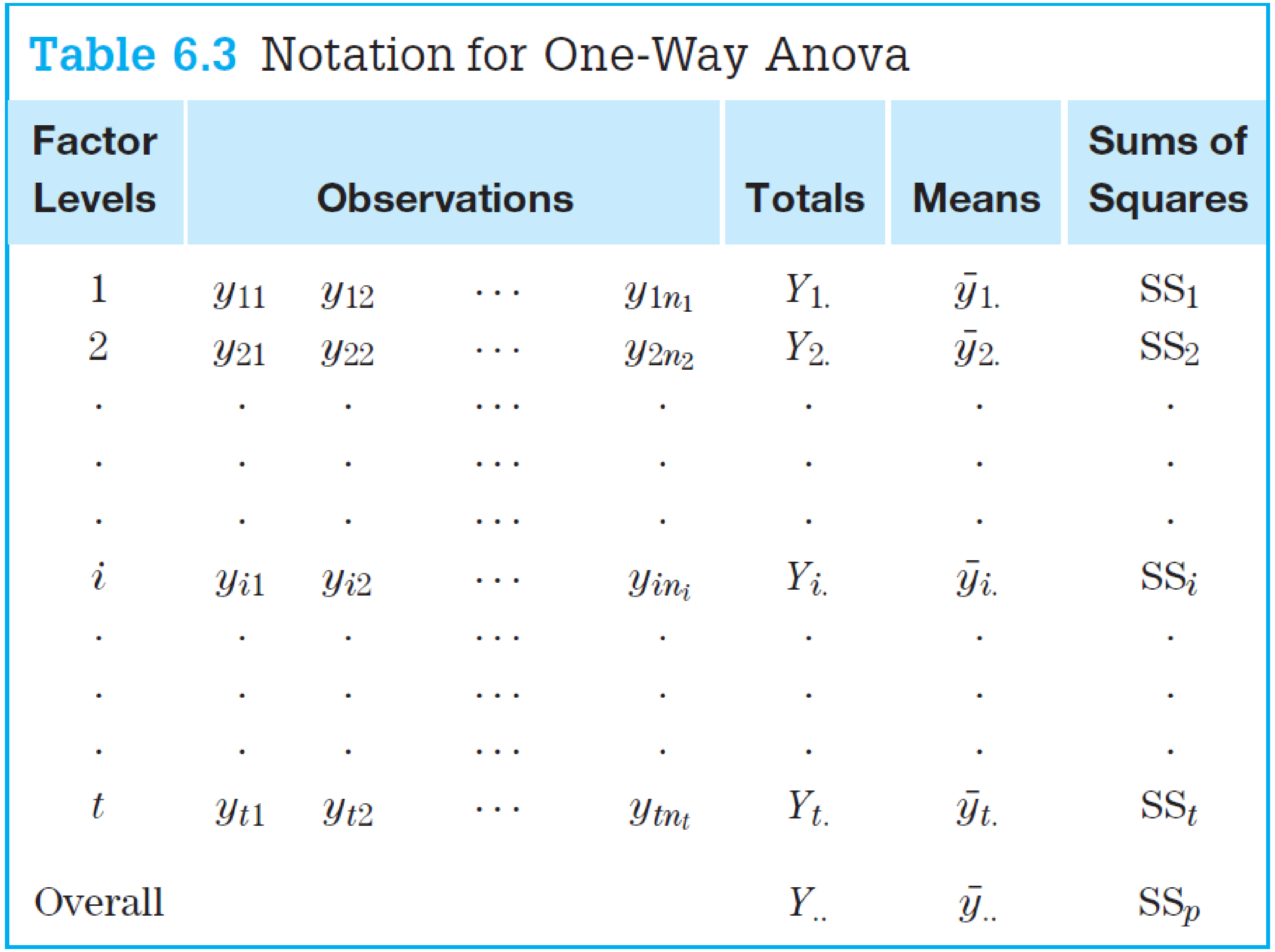


Table 6.3 Notation for One-Way Anova

Factor Levels	Observations				Totals	Means	Sums of Squares
1	y_{11}	y_{12}	$\cdots$	y_{1n_1}	$Y_{1.}$	$\bar{y}_{1.}$	SS_1
2	y_{21}	y_{22}	$\cdots$	y_{2n_2}	$Y_{2.}$	$\bar{y}_{2.}$	SS_2
.	.	.	. . .	.	.	.	.
.	.	.	. . .	.	.	.	.
.	.	.	. . .	.	.	.	.
i	y_{i1}	y_{i2}	$\cdots$	y_{in_i}	$Y_{i.}$	$\bar{y}_{i.}$	SS_i
.	.	.	. . .	.	.	.	.
.	.	.	. . .	.	.	.	.
.	.	.	. . .	.	.	.	.
t	y_{t1}	y_{t2}	$\cdots$	y_{tn_t}	$Y_{t.}$	$\bar{y}_{t.}$	SS_t
Overall					$Y_{..}$	$\bar{y}_{..}$	SS_p

Notation for One-Way ANOVA

- Let t be the number of groups (or sets) from t populations ($t \geq 2$).
- For the ith group of samples ($i = 1, \cdots, t$):
  - The sample size is n_i.
  - The samples are $y_{i1}, y_{i2}, \cdots, y_{in_i}$.
  - The sample mean, sample variance, and sample standard deviation are $\bar{y}_{i\cdot}$, s_i^2, and s_i, respectively.

Notation for One-Way ANOVA

- Note that for the ith set of samples ($i = 1, \cdots, t$):
  - The sample mean: $\bar{y}_{i\cdot} = \frac{1}{n_i}(y_{i1} + y_{i2} + \cdots + y_{in_i})$.
  - The sample variance: $s_i^2 = \frac{1}{n_i - 1}((y_{i1} - \bar{y}_{i\cdot})^2 + \cdots + \left(y_{in_i} - \bar{y}_{i\cdot}\right)^2)$
  - The sample standard deviation: $s_i = \sqrt{s_i^2}$.
  - The (corrected) sum of squares (within group) is:

$$SS_i = (y_{i1} - \bar{y}_{i\cdot})^2 + \cdots + \left(y_{in_i} - \bar{y}_{i\cdot}\right)^2 = (n_i - 1) * s_i^2$$

Notation for One-Way ANOVA

- For all the data, we can calculate:
  - The sample mean:

$$\bar{y}_{..} = \frac{1}{n_1 + \cdots + n_t}(y_{11} + \cdots + y_{1n_1} + \cdots + y_{t1} + \cdots + y_{tn_i})$$

  - The sample variance:

$$s^2 = \frac{1}{n_1 + \cdots + n_t - 1}((y_{i1} - \bar{y}_{..})^2 + \cdots + \left(y_{1n_1} - \bar{y}_{..}\right)^2 + \cdots + (y_{t1} - \bar{y}_{..})^2 + \cdots + \left(y_{tn_t} - \bar{y}_{..}\right)^2)$$

  - The (corrected) sum of squares (Total):

$$TSS = (n_1 + \cdots + n_t - 1) * s^2 =$$

$$(y_{i1} - \bar{y}_{..})^2 + \cdots + \left(y_{1n_1} - \bar{y}_{..}\right)^2 + \cdots + (y_{t1} - \bar{y}_{..})^2 + \cdots + \left(y_{tn_t} - \bar{y}_{..}\right)^2$$

Sample Mean of ANOVA

- Since $\bar{y}_{..} = \frac{1}{n_1+\cdots+n_t}(y_{11}+\cdots+y_{1n_1}+\cdots+y_{t1}+\cdots+y_{tn_i})$, we have
$$\bar{y}_{..} = \frac{1}{n_1+\cdots+n_t}(n_1\bar{y}_{1.}+n_2\bar{y}_{2.}+\cdots+n_t\bar{y}_{t.})$$
- If the sample sizes are same, i.e., $n_1 = n_2 = \cdots = n_t = n$, then we have
$$\bar{y}_{..} = \frac{1}{t}(\bar{y}_{1.}+\bar{y}_{2.}+\cdots+\bar{y}_{t.})$$
- If the sample sizes are not same, in general,
$$\bar{y}_{..} \neq \frac{1}{t}(\bar{y}_{1.}+\bar{y}_{2.}+\cdots+\bar{y}_{t.})$$

Sums of Squares/Sample Mean/Sample Variance

Let n, SS, $\bar{y}$, s^2, and $y_1, \cdots, y_n$ be the sample size, (corrected) sum of squares, sample mean, sample variance, and data from a data set, respectively:

$$\bar{y} = (y_1 + \cdots + y_n)/n$$

$$SS = \left((y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2\right) = (n-1) * s^2$$

- Different calculations result in same values but have difference efficiency.

Notation for One-Way ANOVA

- Note that for such data, the data file only contains two useful columns: one is for the group variable and the other is for the observation.
- So in R, you can use
  - R function such as mean(), var() to calculate the overall sample mean, sample variance (sample standard deviation), and sum of squares.
  - You can use R function tapply to calculate the sample mean, sample variance (sample standard deviation), and sum of squares for each group of samples.

Exercises – Sick Leave by Branch

- **Exercise (Sick Leave by Branch)** - A local bank has three branch offices. The bank has a liberal sick leave policy, and a vice-president was concerned about employees taking advantage of this policy. She thought that the tendency to take advantage depended on the branch at which the employee worked. To see whether there were differences in the time employees took for sick leave, she asked each branch manager to sample employees randomly and record the number of days of sick leave taken during 2008. Ten employees were chosen, and the data are listed here.
  - **Branch 1:** 15, 20, 19, 14
  - **Branch 2:** 11, 15, 11
  - **Branch 3:** 18, 19, 23

Exercises – Sick Leave by Branch

- What is the number of groups (t)?
- What is the sample size for each group?
- What are y_{21}, y_{22}, y_{23}?
- How will you calculate $\bar{y}_{1.}, \bar{y}_{2.}, \bar{y}_{3.}$, and $\bar{y}_{..}$?
- How will you calculate s_1^2, s_2^2, s_3^2,and s^2?
- How will you calculate SS_1, SS_2, SS_3 and TSS?
- How will you use R to create the data and do all the corresponding calculations?

Sums of Squares among Means

We have three types of sum of squares:

- **Total sum of squares (TSS) with $n_1 + \cdots + n_t - 1 = \sum_{i=1}^{t} n_i - 1$ degrees of freedom.**
- **Sum of squares within groups (SSW) with $n_1 + \cdots + n_t - t = \sum_{i=1}^{t} n_i - t$ degrees of freedom.**
- **Sum of squares between groups (SSB) with $t - 1$ degrees of freedom.**

Total Sum of Squares

- **The total sum of squares (TSS)** is:

$$TSS = (y_{i1} - \bar{y}_{..})^2 + \cdots + \left(y_{1n_1} - \bar{y}_{..}\right)^2 + \cdots + (y_{t1} - \bar{y}_{..})^2 + \cdots + \left(y_{tn_t} - \bar{y}_{..}\right)^2$$

$$TSS = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$$

$$\boldsymbol{TSS = (n_1 + \cdots + n_t - 1) * s^2}$$

$$TSS = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij})^2 - (n_1 + \cdots + n_t)(\bar{y}_{..})^2$$

- **The degrees of freedom of the total sum of squares is:**

$$n_1 + \cdots + n_t - 1 = \sum_{i=1}^{t} n_i - 1.$$

Sum of Squares within Groups

- **Sum of squares within groups (SSW) is:**

$$SSW = (SS_1 + \cdots + SS_t)$$

$$\boldsymbol{SSW = (n_1 - 1) * S_1^2 + \cdots + (n_t - 1) * S_t^2}$$

$$SSW = \left(y_{11}^2 + \cdots + y_{1n_1}^2 - n_1 * (\bar{y}_{1.})^2\right) + \cdots + \left(y_{t1}^2 + \cdots + y_{tn_t}^2 - n_t * (\bar{y}_{t.})^2\right)$$

$$SSW = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij})^2 - \sum_{i=1}^{t} n_i * (\bar{y}_{i.})^2$$

- **The degrees of freedom of Sum of squares within groups (SSW) is:**

$$n_1 + \cdots + n_t - t = \sum_{i=1}^{t} n_i - t$$

Sum of Squares within Groups

- **The Sum of squares between groups (SSB) is:**

$$SSB = n_1(\bar{y}_{1.} - \bar{y}_{..})^2 + \cdots + n_t\,(\bar{y}_{t.} - \bar{y}_{..})^2$$

$$SSB = \sum_{i=1}^{t} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$$

- **The degrees of freedom The Sum of squares between groups (SSB) is:**

$$t - 1$$

Overall Test in ANOVA

- For sums of squares, we have:

$$TSS = SSW + SSB;$$

- For the degrees of freedom (df) of sums of squares, we have

$$df\ of\ TSS = df\ of\ SSB + df\ of\ SSW$$

- SSB and SSW are independent.

Partition of Sums of Squares

- In ANOVA, the null hypothesis is:

$$H_0: \mu_1 = \cdots = \mu_t$$

- The alterative hypothesis is:

$$H_a\text{: at least one equality is not satisfied}$$

- The test statistic is:

$$F = \frac{SSB/(t-1)}{SSW/(\sum n_i - t)} = \frac{MSB}{MSW}$$

- It has a *F*-distribution with $t - 1$ and $\sum_{i=1}^{t} n_i - t$ degrees of freedom.
- We reject the null hypothesis when $F > F_{t-1, \sum_{i=1}^{t} n_i - t, \alpha}$
- The p-value is: $\Pr(F_{t-1, \sum_{i=1}^{t} n_i - t} > F)$.

Partition of Sums of Squares

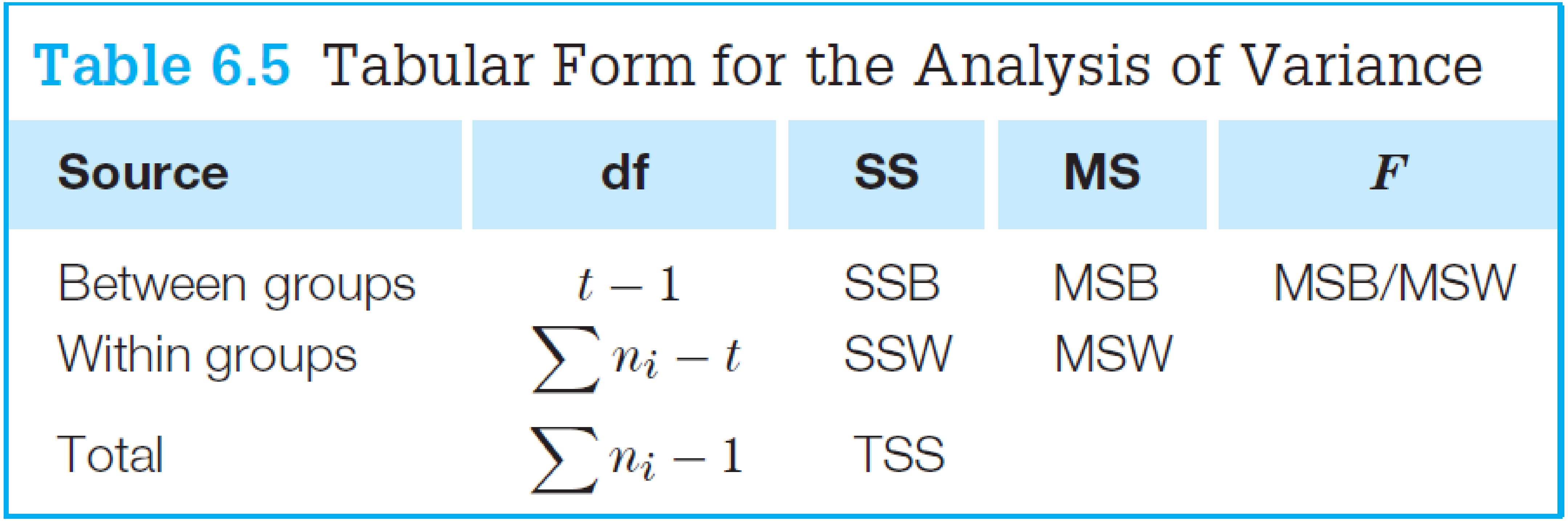


Table 6.5 Tabular Form for the Analysis of Variance

Source	df	SS	MS	F
Between groups	$t - 1$	SSB	MSB	MSB/MSW
Within groups	$\sum n_i - t$	SSW	MSW	
Total	$\sum n_i - 1$	TSS		

Rational of Overall Test

- $F = \frac{MSB}{MSW} = \frac{SSB/(t-1)}{SSW/(\sum n_i - t)} = \frac{\sum_{i=1}^{t} n_i(\bar{y}_{i.} - \bar{y}_{..})^2/(t-1)}{((n_1-1)*S_1^2 + \cdots + (n_t-1)*S_t^2)/(\sum n_i - t)}$
- The numerator, is the estimated (squared) difference of group means and the overall mean. It can be considered as the sample variance among group means. It is expected to be small if all group means are same.
- The denominator, is the estimate of the common variance, σ^2.
- The statistic can be considered as the square of the ratio of the test statistic and its estimated standard deviation.

$\boldsymbol{F}$-test in ANOVA and Two Sample $\boldsymbol{t}$-test

- When $t = 2$ (two groups), then the F-test statistic is equivalent to the two sample t-test statistic for $H_0: \mu_1 = \mu_2$ versus $H_a: \mu_1 \neq \mu_2$.
- See the deviation from another file.

Example 6.2 – Yield of Rice

Example 6.2 (Yield of Rice) - An experiment to compare the yield of four varieties of rice was conducted.

- Each of 16 plots on a test farm where soil fertility was fairly homogeneous was treated alike relative to water and fertilizer.
- Four plots were randomly assigned each of the four varieties of rice. Note that this is a designed experiment, specifically a completely randomized design.
- The yield in pounds per acre was recorded for each plot.
- Do the data presented in Table 6.4 indicate a difference in the mean yield between the four varieties?

Example 6.2 – Data

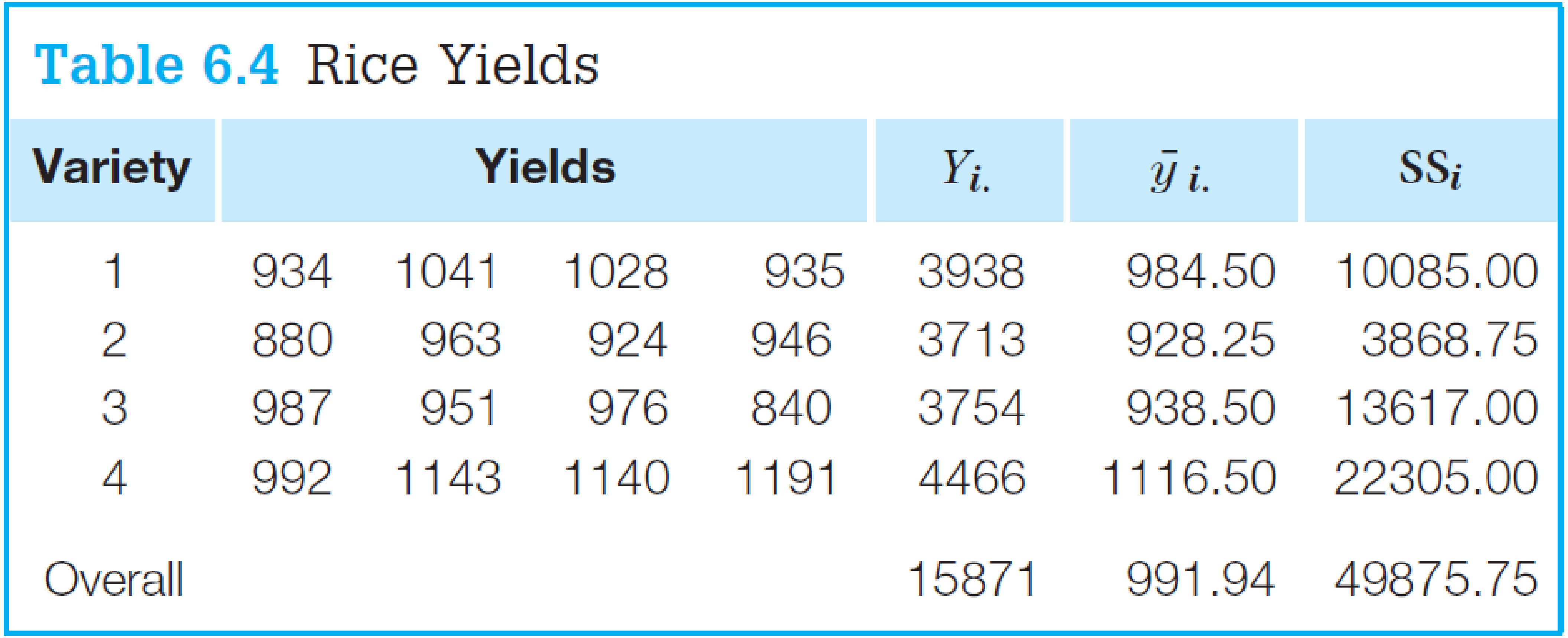
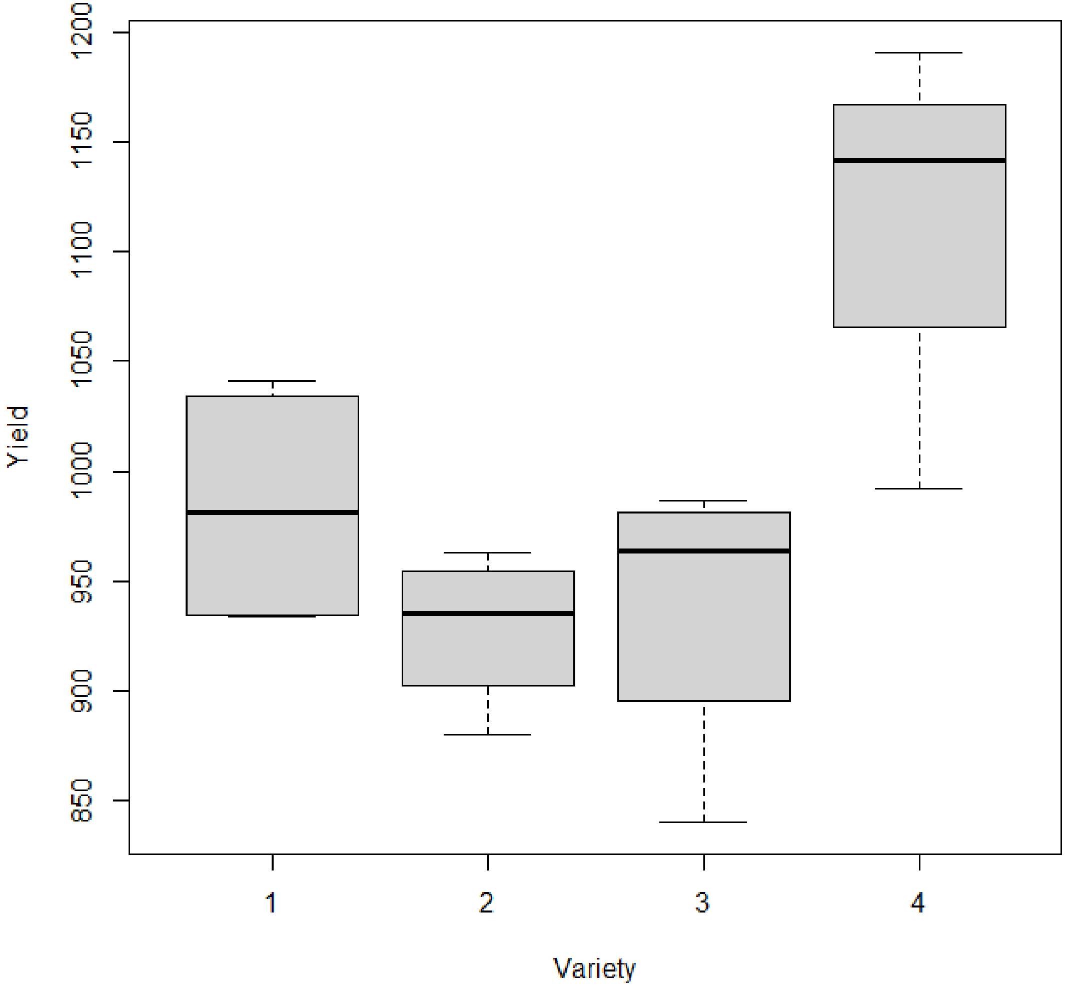


Table 6.4 Rice Yields

Variety	Yields				$Y_{i.}$	$\bar{y}_{i.}$	SS_i
1	934	1041	1028	935	3938	984.50	10085.00
2	880	963	924	946	3713	928.25	3868.75
3	987	951	976	840	3754	938.50	13617.00
4	992	1143	1140	1191	4466	1116.50	22305.00
Overall					15871	991.94	49875.75

Example 6.2 - Plot



Example 6.2 – ANOVA Table

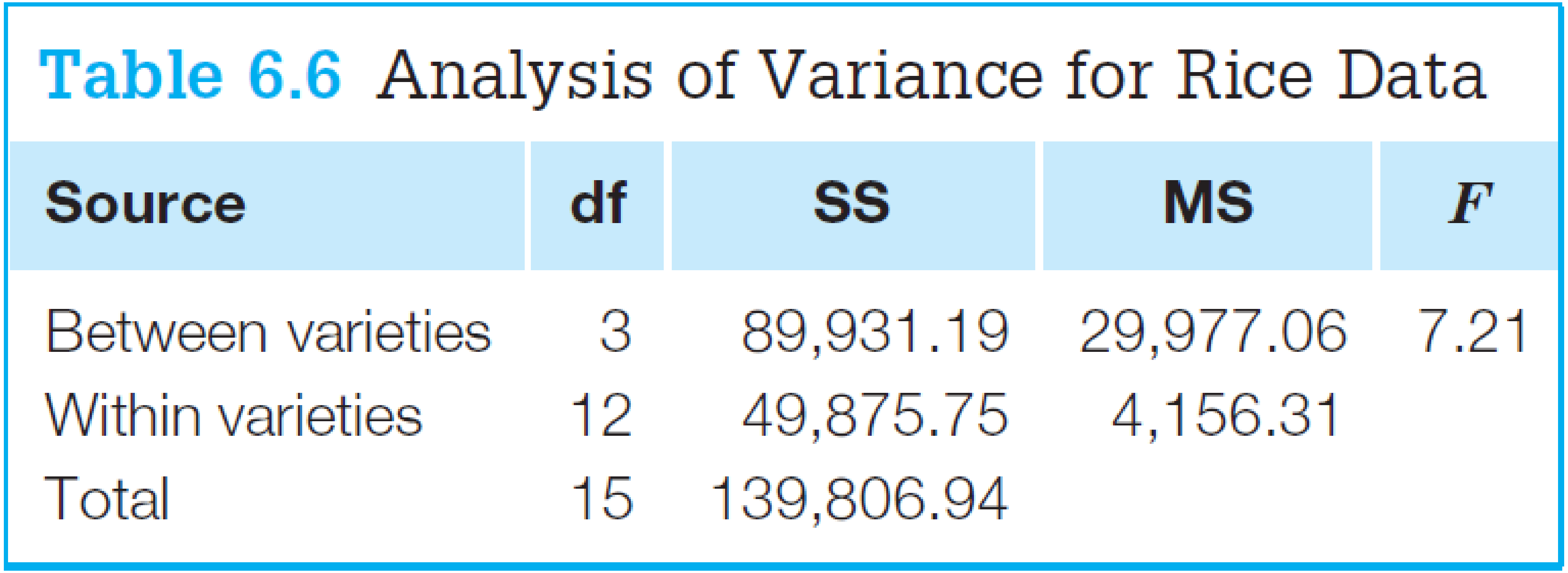


Table 6.6 Analysis of Variance for Rice Data

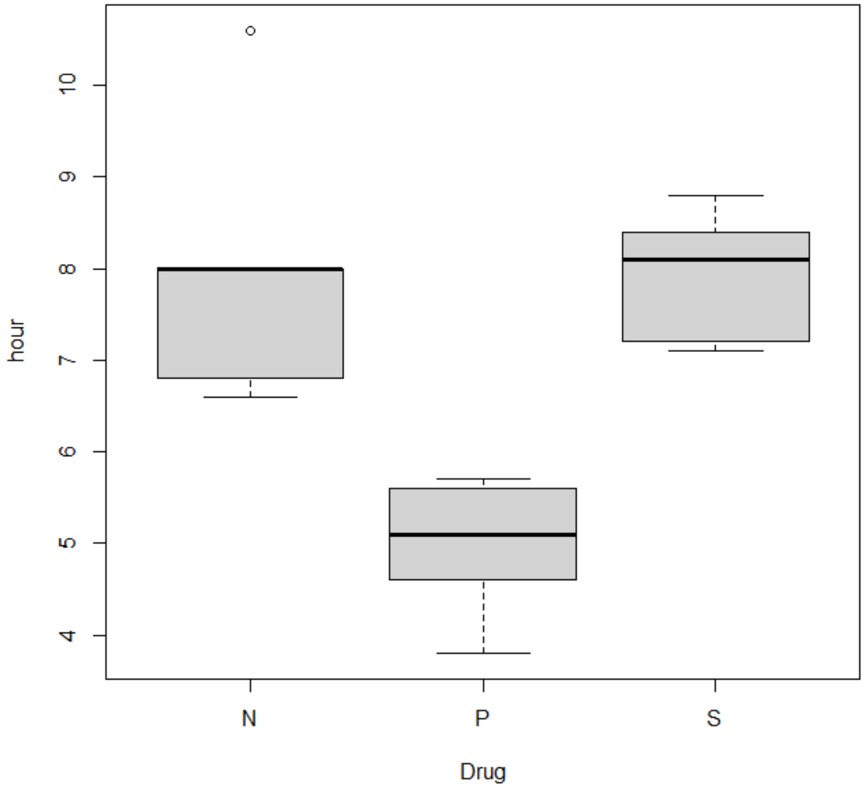
Source	df	SS	MS	*F*
Between varieties	3	89,931.19	29,977.06	7.21
Within varieties	12	49,875.75	4,156.31	
Total	15	139,806.94		

Exercise – Sleeping Drug

Exercise – Sleeping drug - in an experiment to determine the effectiveness of sleep-inducing drugs, 18 insomniacs were randomly assigned to three treatments: (1) placebo (no drug); (2) standard drug; (3) new experimental drug.

- The response is average hours of sleep per night for a week.
- Due to some reasons, some data are missing.
- Data from placebo: 5.6, 5.7, 5.1, 3.8, 4.6
- Data from standard drug: 8.4, 8.2, 8.8, 7.1, 7.2, 8.0
- Data from new drug: 10.6, 6.6, 8.0, 8.0, 6.8

Exercise – Sleeping Drug



Assumptions of ANOVA Under Linear Model

- **Linear Model**: $y_{ij} = \mu_i + \varepsilon_{ij}, i = 1, \cdots, t; j = 1, \cdots, n_i$; where
  - y_{ij} is the observation
  - μ_i is the mean of that group
  - ε_{ij} is the error – difference between the observation and group mean
- **Assumptions of ANOVA**:
  - The specified model and its parameters adequately represent the behavior of the data – self-evident for ANOVA so we do not need to check it.
  - The error (ε_{ij}) are normally distributed random variables with mean zero and variance σ^2.
  - The error are independent – ok with independent samples.

Normality and Homogeneous Variance

- **Normality assumption of error** – Required for the validity *F* distribution of MSB/MSW. Fortunately, its validity is not severely affected by relatively minor violations of the normality assumption. Therefore, the ANOVA test is known as a relatively robust test. However, extreme non-normality, especially extremely skewed distributions, or the existence of outliers may result in biased tests and other methods should be used.
- **Equal variance assumption of error** - Required for the validity *F* distribution of MSB/MSW. Again, minor violations of it do not have a significant effect on the analysis, while major violations may cast doubt on the usefulness of inferences on means.

Detection of Violated Assumptions

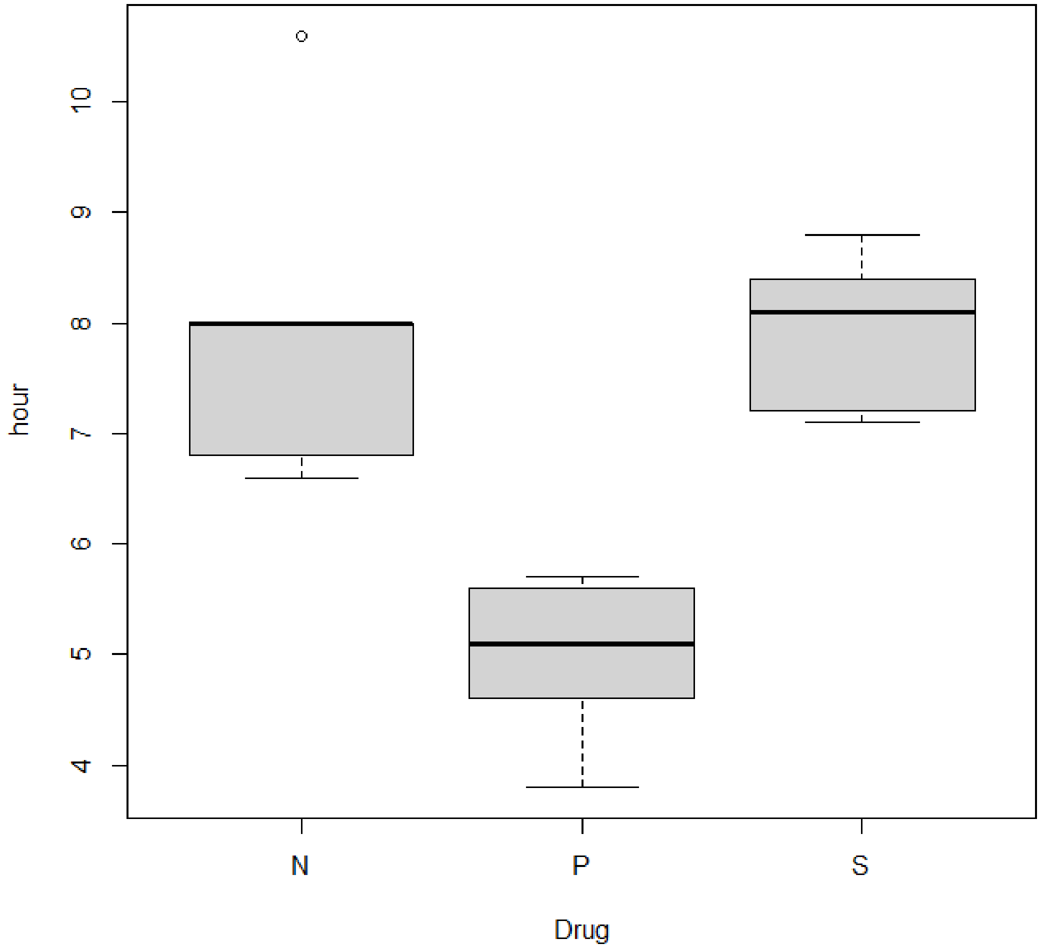
- **Normality assumption** – Look at the error, $y_{ij} - \bar{y}_{i.}$ and draw Q-Q plot.
- **Equal variance assumption** – There are many methods that are implemented in SAS and can be used.
- Here we introduce one test – **Levene Test**:
  - It is robust against serious departures from normality, and does not require equal sample sizes.
  - It computes the absolute difference between the value of each observation and its group mean and performs a one-way analysis of variance on these differences. *F* statistic from this analysis of variance is used as a test for homogeneity of variances.

Exercise – Sleeping Drug

Exercise – Sleeping drug - in an experiment to determine the effectiveness of sleep-inducing drugs, 18 insomniacs were randomly assigned to three treatments: (1) placebo (no drug); (2) standard drug; (3) new experimental drug.

- The response is average hours of sleep per night for a week.
- Due to some reasons, some data are missing.
- Data from placebo: 5.6, 5.7, 5.1, 3.8, 4.6
- Data from standard drug: 8.4, 8.2, 8.8, 7.1, 7.2, 8.0
- Data from new drug: 10.6, 6.6, 8.0, 8.0, 6.8

Exercise – Sleeping Drug



Levene Test from Sleeping Drug Data

- library(car)
- leveneTest(hour ~ factor(drug), data = drug, center = mean)
- R Output:

```
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  2  0.8596 0.4461
      13
```

If Assumptions Are Violated

- If assumptions are violated:
  - Always, to reexamine closely the data and data collection procedures to determine that the data have been correctly measured and recorded.
  - It is also important to verify the model specification, since defects in the model often show up as violations of assumptions.
  - A transformation may be used – logarithm or square root transformation.
  - Alternative analyses may be necessary.

Variance Stabilizing Transformation

- Sometimes, there is a relationship between its variance and its characteristic.
  - Large plants or large animals vary more in size than do small ones.
- Some commonly used transformations:
  - If standard deviation is proportional to mean, use logarithm.
  - If variance is proportional to mean, take the positive square root.
  - If the data are proportions or percentages, use $\arcsin(\sqrt{y_{ij}})$, where y_{ij} are proportions.

Specific Comparisons and Estimations

- A statistically significant *F* test of ANOVA simply indicates that some differences exist among the means but does not indicate what specific differences.
- Often we are also interested in some specific hypotheses , such as (1) Is the mean response for a specific level superior to that of the others? (2) Is there some natural grouping of factor level responses?
- We only consider specific comparisons that are based on certain types of linear combinations of means, which is called **contrasts** .

Contrasts

- **Definition 6.1 -** A **contrast** is a linear function of means whose coefficients add to 0. That is, a linear function of population means, $L = \sum_{i=1}^{t} a_i u_i$ is a contrast if $\sum_{i=1}^{t} a_i = 0$
- There are four groups in Example 6.2 – Yield of Rice. Which of the following is a contrast:
  - $L = \mu_1$
  - $L = \mu_1 - \mu_2$
  - $L = \mu_1 + \mu_2 - 2\mu_3$
  - $L = \mu_1 + \mu_3 - \mu_2 - \mu_4$
- Use ESTIMATE statement of PROC GLM to estimate L and test if $L = 0$