ISyE 6739–The Analysis of Variance (ANOVA) Single Factor (Chapter 13)

Instructor: Kamran Paynabar
H. Milton Stewart School of
Industrial and Systems Engineering
Georgia Tech

Kamran.paynabar@isye.gatech.edu
Office: Groseclose 436

Hypothesis Test on Means of Multiple Normal Distributions

Example:

A manufacturer of paper used for making grocery bags is interested in improving the tensile strength of the product. Product engineering thinks that tensile strength is a function of the hardwood concentration in the pulp and that the range of hardwood concentrations of practical interest is between 5 and 20%. A team of engineers responsible for the study decides to investigate four levels of hardwood concentration: 5%, 10%, 15%, and 20%.

Question of interest: Is hardwood concentration an important factor in improving tensile strength?

$$\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H_a: \text{at least one mean differs from others} \end{cases}$$

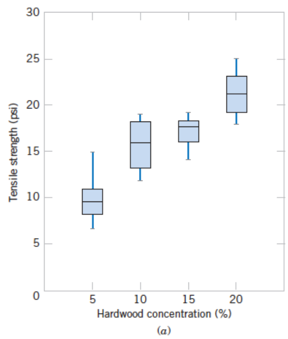
Hypothesis Test on Means of Multiple Normal Distributions

Table 13-1 Tensile Strength of Paper (psi)

Hardwood Concentration (%)	Observations 1	2	3	4	5	6	Totals	Averages
5	7	8	15	11	9	10	60	10.00
10	12	17	13	18	19	15	94	15.67
15	14	18	19	17	16	18	102	17.00
20	19	25	22	23	18	20	127	21.17
							383	15.96

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H_a : \text{at least one mean differs from others} \end{cases}$$

- The levels of the factor are sometimes called treatments.
- Each treatment has six observations or replicates.
- The runs are run in random order.
- This setting is known as Completely Randomized Single-Factor Experiment.



Hypothesis Test on Means of Multiple Normal Distributions

Table 13-2 Typical Data for a Single-Factor Experiment

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

$i = 1, 2, \cdots, a \longrightarrow$ Number of populations

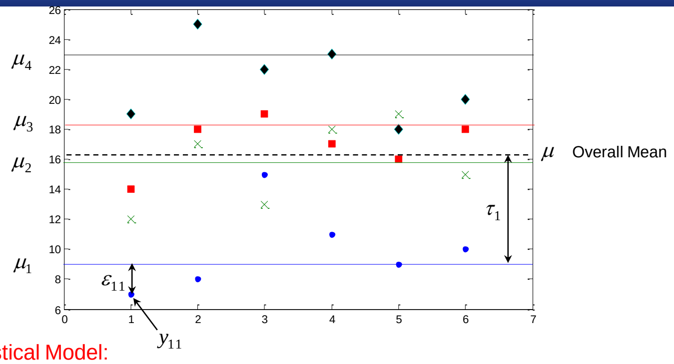
$j = 1, 2, \cdots, n \longrightarrow$ Sample size

y_{ij} Observation j from population i

Population i Total $\quad y_{i\cdot} = \sum_{j=1}^{n} y_{ij}$ $\quad$ Population i Average $\quad \bar{y}_{i\cdot} = y_{i\cdot}/n \quad i = 1, 2, \ldots, a$

Grand Total $\quad y_{\cdot\cdot} = \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}$ $\quad \bar{y}_{\cdot\cdot} = y_{\cdot\cdot}/N$ Grand Average

Hypothesis Test on Means of Multiple Normal Distributions



Linear Statistical Model:

$$Y_{ij} = \mu_i + \epsilon_{ij} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n \end{cases} \qquad Y_{ij} = \mu + \tau_i + \epsilon_{ij} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n \end{cases}$$

Assumptions:

1. $\varepsilon_{ij} \sim NID(0, \sigma^2)$
2. $\sum_{i=1}^{a} \tau_i = 0$
3. Populations (factors) have equal variances

Hypothesis Test on Means of Multiple Normal Distributions

We wish to test the hypotheses:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H_a : \text{at least one mean differs from others} \end{cases}$$

We know that $\mu_i = \mu + \tau_i$

Therefore, the hypothesis test can be written as

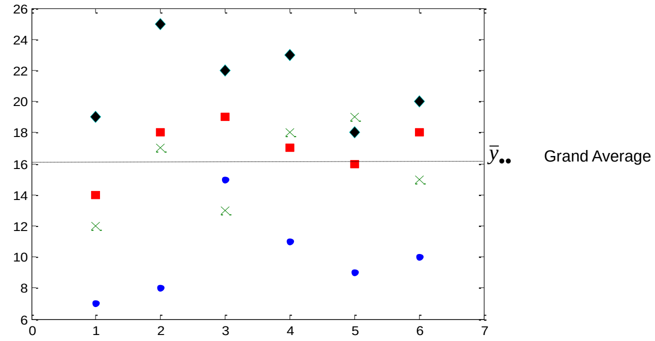
$$H_0: \tau_1 = \tau_2 = \cdots = \tau_a = 0$$
$$H_1: \tau_i \neq 0 \quad \text{for at least one } i$$

Analysis of Variance (ANOVA)

$$\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H_a: \text{at least one mean differs from others} \end{cases}$$

ANOVA partitions the total variability into two parts

Total Variations = Between-group Variations + Within-group Variations



$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2 = n\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2 \qquad (13\text{-}5)$$

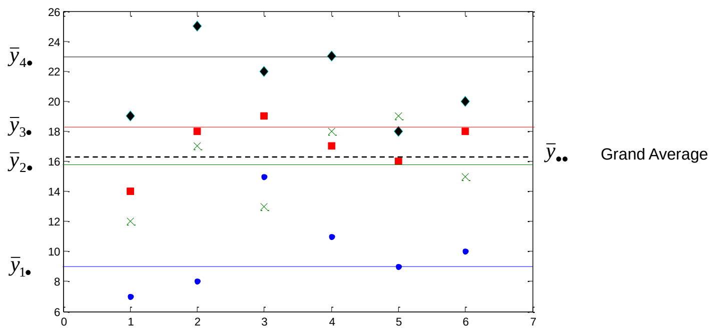
Analysis of Variance (ANOVA)

$$\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H_a: \text{at least one mean differs from others} \end{cases}$$

ANOVA partitions the total variability into two parts

Total Variations = Between-group Variations + Within-group Variations

SST = SSB + SSW or (SST = $SS_{treatments}$ + SS_{Error})



$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2 = n\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2 \qquad (13\text{-}5)$$

Analysis of Variance (ANOVA)

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \cdots = \mu_a \\ H_a : \text{at least one mean differs from others} \end{cases}$$

ANOVA partitions the total variability into two parts

$$\text{SST} = \text{SS}_{\text{treatments}} + \text{SS}_{\text{Error}}$$

The appropriate test statistic is

$$F_0 = \frac{SS_{\text{Treatments}}/(a-1)}{SS_E/[a(n-1)]} = \frac{MS_{\text{Treatments}}}{MS_E} \qquad (13\text{-}7)$$

We would reject H_0 if

$$F_0 > F_{\alpha, a-1, a(n-1)} \quad \text{or} \quad F_0 > F_{\alpha, a-1, N-a}$$

Analysis of Variance (ANOVA)

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \cdots = \mu_a \\ H_a : \text{at least one mean differs from others} \end{cases}$$

$$\text{SST} = \text{SS}_{\text{treatments}} + \text{SS}_{\text{Error}}$$

The sums of squares computing formulas for the ANOVA with equal sample sizes in each treatment are

$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}^2 - \frac{y_{..}^2}{N} \qquad (13\text{-}8)$$

and

$$SS_{\text{Treatments}} = \sum_{i=1}^{a} \frac{y_{i.}^2}{n} - \frac{y_{..}^2}{N} \qquad (13\text{-}9)$$

The error sum of squares is obtained by subtraction as

$$SS_E = SS_T - SS_{\text{Treatments}} \qquad (13\text{-}10)$$

ANOVA Table

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \cdots = \mu_a \\ H_a : \text{at least one mean differs from others} \end{cases}$$

Table 13-3 The Analysis of Variance for a Single-Factor Experiment, Fixed-Effects Model

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Treatments	$SS_{\text{Treatments}}$	$a - 1$	$MS_{\text{Treatments}}$	$\dfrac{MS_{\text{Treatments}}}{MS_E}$
Error	SS_E	$a(n - 1)$	MS_E	
Total	SS_T	$an - 1$		

We would reject H_0 if

$$F_0 > F_{\alpha, a-1, a(n-1)} \quad \text{or} \quad F_0 > F_{\alpha, a-1, N-a}$$

ANOVA

Example:

A manufacturer of paper used for making grocery bags is interested in improving the tensile strength of the product. Product engineering thinks that tensile strength is a function of the hardwood concentration in the pulp and that the range of hardwood concentrations of practical interest is between 5 and 20%. A team of engineers responsible for the study decides to investigate four levels of hardwood concentration: 5%, 10%, 15%, and 20%.

Hardwood Concentration (%)	Observations 1	2	3	4	5	6	Totals	Averages
5	7	8	15	11	9	10	60	10.00
10	12	17	13	18	19	15	94	15.67
15	14	18	19	17	16	18	102	17.00
20	19	25	22	23	18	20	127	21.17
							383	15.96

$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}^2 - \frac{y_{..}^2}{N}$$

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

$$SS_E = SS_T - SS_{\text{Treatments}}$$

ANOVA for Unbalanced Experiments

The sums of squares computing formulas for the ANOVA with unequal sample sizes n_i in each treatment are

$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}^2}{N} \qquad (13\text{-}13)$$

$$SS_{\text{Treatments}} = \sum_{i=1}^{a} \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{N} \qquad (13\text{-}14)$$

and

$$SS_E = SS_T - SS_{\text{Treatments}} \qquad (13\text{-}15)$$

Confidence Intervals on Means in ANOVA

A $100(1 - \alpha)$ percent confidence interval on the mean of the ith treatment μ_i is

$$\bar{y}_{i.} - t_{\alpha/2,a(n-1)} \sqrt{\frac{MS_E}{n}} \le \mu_i \le \bar{y}_{i.} + t_{\alpha/2,a(n-1)} \sqrt{\frac{MS_E}{n}} \qquad (13\text{-}11)$$

Example: For 20% hardwood, the resulting confidence interval on the mean is?

Confidence Intervals on Mean Differences in ANOVA

A $100(1-\alpha)$ percent confidence interval on the difference in two treatment means $\mu_i - \mu_j$ is

$$\bar{y}_{i\cdot} - \bar{y}_{j\cdot} - t_{\alpha/2,a(n-1)}\sqrt{\frac{2MS_E}{n}} \le \mu_i - \mu_j \le \bar{y}_{i\cdot} - \bar{y}_{j\cdot} + t_{\alpha/2,a(n-1)}\sqrt{\frac{2MS_E}{n}} \qquad (13\text{-}12)$$

Example: For 15% and 10% hardwood, the resulting confidence interval on the mean difference is?

Multiple Comparisons Following ANOVA

The least significant difference (LSD) is

$$\text{LSD} = t_{\alpha/2,a(n-1)}\sqrt{\frac{2MS_E}{n}} \qquad (13\text{-}16)$$

If the sample sizes are different in each treatment: $\text{LSD} = t_{\alpha/2,N-a}\sqrt{MS_E\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$

If $\bar{y}_{i\bullet} - \bar{y}_{j\bullet} > LSD$, then, mean treatment *i* differs from mean treatment *j*

Example

EXAMPLE 13-2
We will apply the Fisher LSD method to the hardwood concentration experiment. There are $a = 4$ means, $n = 6$, $MS_E = 6.51$, and $t_{0.025,20} = 2.086$. The treatment means are

$$\bar{y}_{1\cdot} = 10.00 \text{ psi}$$
$$\bar{y}_{2\cdot} = 15.67 \text{ psi}$$
$$\bar{y}_{3\cdot} = 17.00 \text{ psi}$$
$$\bar{y}_{4\cdot} = 21.17 \text{ psi}$$

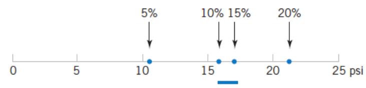


Figure 13-2 Results of Fisher's LSD method in Example 13-2.

Use LSD to find which pairs of treatments have different means.

Model Adequacy Checking

Validity of assumptions is checked by residual analysis

$$Y_{ij} = \mu_i + \epsilon_{ij} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n \end{cases} \qquad \hat{y}_{ij} = \bar{y}_{ij} \implies e_{ij} = y_{ij} - \hat{y}_{ij}$$ residuals

Table 13-6 Residuals for the Tensile Strength Experiment

Hardwood Concentration (%)	Residuals					
5	−3.00	−2.00	5.00	1.00	−1.00	0.00
10	−3.67	1.33	−2.67	2.33	3.33	−0.67
15	−3.00	1.00	2.00	0.00	−1.00	1.00
20	−2.17	3.83	0.83	1.83	−3.17	−1.17

Assumptions:

1. $\varepsilon_{ij} \sim NID(0, \sigma^2)$
2. $\sum_{i=1}^{a} \tau_i = 0$
3. Populations (factors) have equal variances

Normality Check

- Normal Plot
- Histogram
- Goodness of fit tests

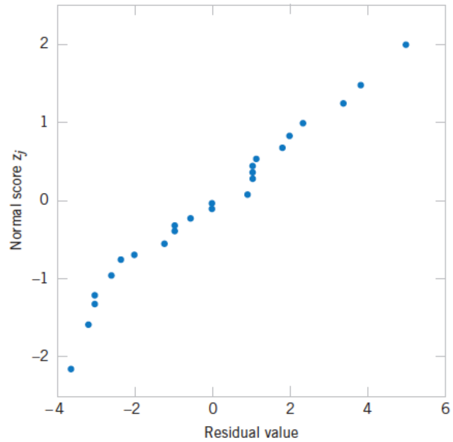


Figure 13-4 Normal probability plot of residuals from the hardwood concentration experiment.

Variance Consistency

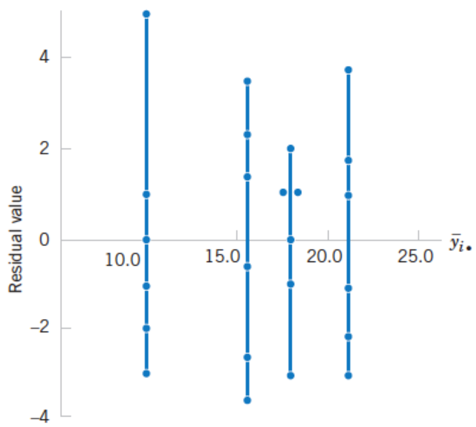


Figure 13-6 Plot of residuals versus $\overline{y}_i$

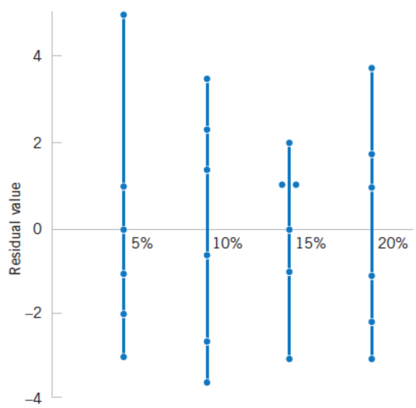


Figure 13-5 Plot of residuals versus factor levels (hardwood concentration).