Chapter 3. Analysis of Variance

Reading Assignment

Chapter 13 (skip Sections 8, 9, 10).

Analysis of Variance

Comparison of means of samples with different independent variables.

1. n_i: sample size from treatment i.
2. $\{Y_{ij} : 1 \leq j \leq n_i\}$: responses (samples) from treatment i.
3. $\bar{Y}_i$: sample mean from treatment i.
4. $\bar{Y}$: sample mean from all the responses.
5. k: number of treatments.
6. $n = n_1 + n_2 + \cdots + n_k$: total sample size.

The Statistical Model

Let Y_{ij} be the j-th response from treatment i.

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad 1 \le i \le k, 1 \le j \le n_i.$$

- μ_i: the population mean for treatment i.
- ε_{ij}: random error, with distribution $N(0, \sigma^2)$.

$$H_0\text{: } \mu_1 = \mu_2 = \cdots = \mu_k$$

Procedure of ANOVA

Total variation:

$$\text{Total SS} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2.$$

1. Variation between treatments:

$$\text{SST} = \sum_{i=1}^{k} n_i (\bar{Y}_i - \bar{Y})^2$$

2. Variation within treatment i:

$$\text{SSE} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2.$$

$$\text{Total SS} = \text{SST} + \text{SSE}$$

Source of Variation	d.f.	SS	MS	F
between treatments	$k-1$	SST	$\text{MST} = \dfrac{\text{SST}}{k-1}$	$F = \dfrac{\text{MST}}{\text{MSE}}$
Within treatment	$n-k$	SSE	$\text{MSE} = \dfrac{\text{SSE}}{n-k}$	
Total	$n-1$	Total SS		

Under H_0, $F \sim F(k-1, n-k)$. Therefore

$$\text{P-value} = P(F(k-1, n-k) \geq F)$$

$$\text{Rejection Region} = \{F > F_\alpha(k-1, n-k)\}$$

Justification of ANOVA

1. An illustration of the effect of the difference between μ_i

   Let $Y_{ij} \sim N(\mu_i, 1)$ [or $\sigma = 1$].

(a) Case I: $\mu_1 = \mu_2 = \mu_3 = 0$.

Data:

	Treatment 1	Treatment 2	Treatment 3
	0.17	1.62	0.33
	0.20	-0.80	0.00
mean	0.185	0.41	0.175

Source of Variation	d.f.	SS	MS	F
between treatments	2	0.074	0.037	$F = 0.037$
Within treatment	3	2.98	0.99	
Total	5	3.05		

P-value = 0.96

(b) Case II: $\mu_1 = -1, \mu_2 = 0, \mu_3 = 2.$

Data:

	Treatment 1	Treatment 2	Treatment 3
	-0.83	1.62	2.33
	-0.80	-0.80	2.00
mean	-0.815	0.41	2.175

Source of Variation	d.f.	SS	MS	F
between treatments	2	8.97	4.485	$F = 4.53$
Within treatment	3	2.98	0.99	
Total	5	11.95		

P-value = 0.12

(c) Case III: $\mu_1 = -2, \mu_2 = 0, \mu_3 = 4.$

Data:

	Treatment 1	Treatment 2	Treatment 3
	-1.83	1.62	4.33
	-1.80	-0.80	4.00
mean	-1.815	0.41	4.175

Source of Variation	d.f.	SS	MS	F
between treatments	2	36.54	18.27	$F = 18.45$
Within treatment	3	2.98	0.99	
Total	5	39.52		

P-value = 0.02

2. The special case of $k = 2$

Equivalent to the t-test.

$$F = T^2 = \left(\frac{\bar{Y}_1 - \bar{Y}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right)^2$$

where

$$S_p \doteq \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

$$S_i \doteq \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2.$$

3. The expected values of MST and MSE

$$E[\text{MST}] = \sigma^2 + \frac{1}{k-1}\sum_{i=1}^{k} n_i(\mu_i - \bar{\mu})^2, \qquad \bar{\mu} = \frac{1}{n}\sum_{i=1}^{k} n_i\mu_i.$$

$$E[\text{MSE}] = \sigma^2$$

Proofs

Want to show that, under H_0, SSE $\sim \chi^2(n-k)$, SST $\sim \chi^2(k-1)$ and they are independent.

Step 0: Assume $\mu_1 = \mu_2 = \ldots = \mu_i = 0$.

Step 1: SSE + SST = Total SS: straightforward.

Step 2: SSE and SST are independent: $\bar{\varepsilon}$ only depends on $\bar{\varepsilon}_i$.

Step 3: SSE$\sim \sigma^2\chi^2(n-k)$ is simple.

Step 4: SST$\sim \sigma^2\chi^2(k-1)$.

Estimation of the differences of μ_i

$$S^2 \doteq \text{MSE} = \frac{\text{SSE}}{n-k}.$$

1. The $(1-\alpha)$ confidence interval for μ_i is

$$\bar{Y}_i \pm t_{\alpha/2}(n-k)\frac{S}{\sqrt{n_i}}$$

2. The $(1-\alpha)$ confidence interval for $\mu_i - \mu_j$ is

$$(\bar{Y}_i - \bar{Y}_j) \pm t_{\alpha/2}(n-k)S\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}$$

3. In general, the $(1-\alpha)$ confidence interval for

$$\sum_{i=1}^{k} \lambda_i \mu_i$$

is

$$\sum_{i=1}^{k} \lambda_i \bar{Y}_i \pm t_{\alpha/2}(n-k) S \sqrt{\sum_{i=1}^{k} \frac{\lambda_i^2}{n_i}}$$

Example

The grades from three sections.

Section 1	Section 2	Section 3
95	85	79
32	90	92
47	79	63
75	50	68
83	32	76
84	84	20
73	78	37
68	95	74
	65	86
	80	

$$k = 3, \quad n_1 = 8, \quad n_2 = 10, \quad n_3 = 9, \quad n = n_1 + n_2 + n_3 = 27$$

1. Test $H_0 : \mu_1 = \mu_2 = \mu_3$ with $\alpha = 0.05$.

$$\bar{Y}_1 = 69.6, \quad \bar{Y}_2 = 73.8, \quad \bar{Y}_3 = 66.1$$

$$\bar{Y} = \frac{\qquad\qquad\qquad\qquad\qquad}{} = 70$$

$$\text{SST} = \sum_{i=1}^{3} n_i(\bar{Y}_i - \bar{Y})^2 = 282.57.$$

$$S_1^2 = 428.55, \quad S_2^2 = 379.51, \quad S_3^2 = 547.36.$$

$$\text{SSE} = \frac{\qquad\qquad\qquad\qquad\qquad\qquad}{} = 10794.32$$

$$S^2 = \frac{\text{SSE}}{n-k} = 449.76$$

$$F = \frac{\text{MST}}{\text{MSE}} = \frac{\text{SST}/(k-1)}{\text{SSE}/(n-k)} = \frac{141.29}{449.76} = 0.31$$

$$\text{P-value} = P(F(2,24) > 0.31) = 0.74$$

We cannot reject H_0.

2. Find a 95% confidence interval for μ_2.

$$\begin{aligned} \bar{Y}_2 \pm t_{0.025}(24)\frac{S}{\sqrt{n_2}} &= 73.8 \pm 2.064 \cdot \sqrt{449.76} \cdot \sqrt{\frac{1}{10}} \\ &= [60.0, 87.6]. \end{aligned}$$

3. Find a 95% confidence interval for $(\mu_1 + \mu_2)/2 - \mu_3$.

$$
\begin{aligned}
&[(\bar{Y}_1 + \bar{Y}_2)/2 - \bar{Y}_3] \pm t_{0.025}(24) S \sqrt{\frac{(1/2)^2}{n_1} + \frac{(1/2)^2}{n_2} + \frac{(-1)^2}{n_3}} \\
&= 5.6 \pm 2.064 \cdot \sqrt{449.76} \cdot \sqrt{\frac{(1/2)^2}{8} + \frac{(1/2)^2}{10} + \frac{(-1)^2}{9}} \\
&= [-12.3, 23.5].
\end{aligned}
$$

Relations with multiple linear regression

ANOVA can be solved using linear regression models.

- Transform ANOVA into a linear model.

  Define the *explanatory* variables $(x_1, x_2, \ldots, x_{k-1})$ such that

$$x_i = \begin{cases} 1, & \text{if the response is from Treatment } i \\ 0, & \text{otherwise.} \end{cases}$$

  [Why we do not need x_k?] Then

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-1} x_{k-1} + \varepsilon$$

  with

$$\beta_0 = \mu_k, \quad \beta_1 = \mu_1 - \mu_k, \quad \cdots, \quad \beta_{k-1} = \mu_{k-1} - \mu_k,$$

$$\varepsilon \sim N(0, \sigma^2).$$

- Properties of the linear model.

1. The least square estimates are

$$\hat{\beta}_0 = \bar{Y}_k, \quad \hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_k, \quad \cdots, \quad \hat{\beta}_{k-1} = \bar{Y}_{k-1} - \bar{Y}_k$$

2. SSE for regression = SSE in ANOVA:

$$\text{SSE} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2.$$

We denote this SSE by SSE_C [C = complete model].

- Relations with ANOVA.

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

is equivalent to

$$H_0 : \beta_1 = \cdots = \beta_{k-1} = 0.$$

The F-test statistics

$$F = \frac{(\text{SSE}_R - \text{SSE}_C)/(k-1)}{\text{SSE}_C/(n-k)}.$$

SSE_R = Total SS in ANOVA[Why?], SSE_C = SSE in ANOVA

$$\text{SSE}_R - \text{SSE}_C = \text{SST in ANOVA}.$$

Same F-statistics.

- Linear regression model can do more.

Example

1. Use linear regression to solve ANOVA problem

Treatment	Response
A [1]	47, 42, 43, 46, 44, 42
B [2]	51, 58, 62, 49, 53, 51, 50, 59
C [3]	37, 39, 41, 38, 39, 37, 42, 36, 40
D [4]	42, 43, 42, 45, 47, 50, 48

$$k = 4, \quad n_1 = 6, \quad n_2 = 8, \quad n_3 = 9, \quad n_4 = 7, n = 30$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon.$$

- For response from Treatment A, $x = (x_1, x_2, x_3) = (1, 0, 0)$.
- For response from Treatment B, $x = (x_1, x_2, x_3) = (0, 1, 0)$.
- For response from Treatment C, $x = (x_1, x_2, x_3) = (0, 0, 1)$.
- For response from Treatment D, $x = (x_1, x_2, x_3) = (0, 0, 0)$.

$$\hat{\beta}_0 = 45.3, \quad \hat{\beta}_1 = -1.3, \quad \hat{\beta}_2 = 8.8, \quad \hat{\beta}_3 = -6.5$$

$$[\hat{\beta}_0 = \bar{Y}_4,\ \hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_4,\ \hat{\beta}_2 = \bar{Y}_2 - \bar{Y}_4,\ \hat{\beta}_3 = \bar{Y}_3 - \bar{Y}_4.]$$

$$\mathrm{SSE} = \mathrm{SSE}_C = 277.86.$$

$$S^2 = \frac{\mathrm{SSE}}{n-k} = 10.687, \quad S = \sqrt{S^2} = 3.269.$$

$$(X'X)^{-1} = \begin{bmatrix} 0.143 & -0.143 & -0.143 & -0.143 \\ -0.143 & 0.310 & 0.143 & 0.143 \\ -0.143 & 0.143 & 0.268 & 0.143 \\ -0.143 & 0.143 & 0.143 & 0.2534 \end{bmatrix}$$

(a) Test $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ or $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$.

$$\text{SSE}_R = \text{SSE for regression model "}Y = \beta_0 + \varepsilon\text{"} = 1293.67$$

$$F = \frac{(1293.67 - 277.86)/(4-1)}{277.86/(30-4)} = 31.7$$

$$\text{P-value} = P(F(3, 26) \geq 31.7) \approx 0.$$

(b) Give 95% confidence interval for $\mu_1 - \mu_4$.

Note that $\mu_1 - \mu_4 = \beta_1 = a'\beta$ with

$$a = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

And 95% confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{0.025}(26) S \sqrt{a'(X'X)^{-1}a}$$

or

$$-1.3 \pm 2.056 \cdot 3.269 \cdot \sqrt{0.310} = [-5.04, 2.44]$$

(c) Give 95% confidence interval for $\mu_1 - \mu_2$.

Use

$$\mu_1 - \mu_2 = (\beta_0 + \beta_1) - (\beta_0 + \beta_2) = \beta_1 - \beta_2 = a'\beta$$

with

$$a = \begin{bmatrix} 0 \\ 1 \\ -1 \\ 0 \end{bmatrix}$$

(d) Give 95% confidence interval for $\mu_1 + \mu_3 + \mu_4$.

Use

$$\mu_1 + \mu_3 = (\beta_0 + \beta_1) + (\beta_0 + \beta_3) + \beta_0 = a'\beta$$

with

$$a = \begin{bmatrix} 3 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$