ANOVA: Distribution of SSB and SSW under H_0

Setup and Definitions

Let $X = (X_1, \dots, X_n) \in \{1, \dots, I\}^n$ be the group labels and $Y = (Y_1, \dots, Y_n) \in \mathbb{R}^n$ the responses. Define:

$$N_i = \sum_{k=1}^{n} \mathbf{1}\{X_k = i\}, \qquad \overline{Y}_i = \frac{1}{N_i} \sum_{k=1}^{n} Y_k \mathbf{1}\{X_k = i\}, \qquad \overline{Y} = \frac{1}{n} \sum_{k=1}^{n} Y_k$$

The **variance decomposition** reads:

$$\underbrace{\sum_{k=1}^{n} (Y_k - \overline{Y})^2}_{SST} = \underbrace{\sum_{i=1}^{I} N_i (\overline{Y}_i - \overline{Y})^2}_{SSB} + \underbrace{\sum_{i=1}^{I} \sum_{k=1}^{n} \mathbf{1}\{X_k = i\} (Y_k - \overline{Y}_i)^2}_{SSW}$$

Distribution under H_0

Model: $Y_k = \mu + \varepsilon_k$, $\varepsilon_k \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, i.e. $\mu_1 = \cdots = \mu_I = \mu$ (unknown).

> **Proposition**
>
> Under H_0, assuming $\varepsilon_k \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$:
>
> $$\frac{SSB}{\sigma^2} \sim \chi^2(I-1), \qquad \frac{SSW}{\sigma^2} \sim \chi^2(n-I)$$
>
> and $SSB \perp SSW$. Consequently,
>
> $$F = \frac{SSB/(I-1)}{SSW/(n-I)} \sim \mathcal{F}(I-1,\, n-I)$$

Proof

Step 1: Reduction to ε.

Under H_0, $\overline{Y} = \mu + \overline{\varepsilon}$ and $\overline{Y}_i = \mu + \overline{\varepsilon}_i$, so the μ's cancel:

$$SSW = \sum_{i=1}^{I} \sum_{k=1}^{n} \mathbf{1}\{X_k = i\}(\varepsilon_k - \overline{\varepsilon}_i)^2, \qquad SSB = \sum_{i=1}^{I} N_i(\overline{\varepsilon}_i - \overline{\varepsilon})^2$$

Step 2: Matrix formulation.

Let $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top \sim \mathcal{N}(0, \sigma^2 I_n)$. Define the following orthogonal projection matrices:

- $P_1 = \frac{1}{n}\mathbf{1}\mathbf{1}^\top$, projecting onto span($\mathbf{1}_n$) (global mean), with rank(P_1) = 1.
- $P_2 = \text{diag}(P_{G_1}, \dots, P_{G_I})$ where $P_{G_i} = \frac{1}{N_i}\mathbf{1}_{N_i}\mathbf{1}_{N_i}^\top$, so $(P_2\varepsilon)_k = \overline{\varepsilon}_{X_k}$ (within-group means), with rank(P_2) = I.

One checks that $P_1P_2 = P_2P_1 = P_1$ (since $\mathbf{1}_n \in \text{Im}(P_2)$). Then:

$$SSB = \varepsilon^\top(P_2 - P_1)\varepsilon, \qquad SSW = \varepsilon^\top(I_n - P_2)\varepsilon$$

Step 3: The two matrices are orthogonal projections with orthogonal ranges.

- $P_2 - P_1$ is an orthogonal projection: it is symmetric and $(P_2 - P_1)^2 = P_2^2 - P_2P_1 - P_1P_2 + P_1^2 = P_2 - P_1 - P_1 + P_1 = P_2 - P_1$.
- $I_n - P_2$ is an orthogonal projection (standard complement of P_2).
- Their ranges are orthogonal since $(I_n - P_2)(P_2 - P_1) = P_2 - P_1 - P_2^2 + P_2P_1 = P_2 - P_1 - P_2 + P_1 = 0$.

Step 4: Degrees of freedom.

$$\text{rank}(P_2 - P_1) = \text{tr}(P_2 - P_1) = I - 1$$
$$\text{rank}(I_n - P_2) = \text{tr}(I_n - P_2) = n - I$$

Step 5: Chi-squared distributions and independence (Cochran's theorem).

For $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and an orthogonal projection P of rank r:

$$\frac{\varepsilon^\top P\, \varepsilon}{\sigma^2} \sim \chi^2(r)$$

Two such quadratic forms $\varepsilon^\top P\varepsilon$ and $\varepsilon^\top Q\varepsilon$ are independent if and only if $PQ = 0$, i.e. their ranges are orthogonal. Applying this:

$$\frac{SSB}{\sigma^2} = \frac{\varepsilon^\top (P_2 - P_1)\varepsilon}{\sigma^2} \sim \chi^2(I-1)$$

$$\frac{SSW}{\sigma^2} = \frac{\varepsilon^\top (I_n - P_2)\varepsilon}{\sigma^2} \sim \chi^2(n-I)$$

and $SSB \perp SSW$ since $(P_2 - P_1)(I_n - P_2) = 0$.

Step 6: Fisher distribution.

By definition of the Fisher distribution, the ratio of two independent chi-squared variables divided by their respective degrees of freedom follows an $\mathcal{F}$ distribution:

$$F = \frac{SSB/(I-1)}{SSW/(n-I)} = \frac{\chi^2(I-1)/(I-1)}{\chi^2(n-I)/(n-I)} \sim \mathcal{F}(I-1,\, n-I) \qquad \blacksquare$$