1 Motivating example

Example 1. We would like to estimate the likability index of four different instructors, Instructor 1, Instructor 2, Instructor 3, Instructor 4, by checking at their reviews from three students. Let $X_i$ (for $i \in \{1, 2, 3, 4\}$) by the (random) likability index of Instructor $i$, which is a normal random variable with mean $\mu_i$ and variance $\sigma^2$ (all instructors share the same variance). The review scores from the three students are given by

<table>
<thead>
<tr>
<th></th>
<th>Student A</th>
<th>Student B</th>
<th>Student C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instructor 1</td>
<td>13</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Instructor 2</td>
<td>15</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>Instructor 3</td>
<td>8</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>Instructor 4</td>
<td>11</td>
<td>15</td>
<td>10</td>
</tr>
</tbody>
</table>

The hypothesis are:

\[ H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \]
\[ H_1: \text{at least one } \mu_i \text{ is different} \]

\[ F = \frac{MSB}{MSE} \]
\[ dfB = k - 1 \]
\[ dfE = (n - 1)(k - 1) \]
\[ p\text{-value} = \Pr(F > F_{dfB, dfE}) \]

\[ F_{dfB, dfE} \text{ is the } \alpha\text{ critical value of the } F \text{-distribution} \]

\[ \text{Critical value: } F_{0.05, 3, 12} = 3.89 \]

The hypothesis are:

\[ H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \]
\[ H_1: \text{at least one } \mu_i \text{ is different} \]

\[ F = \frac{MSB}{MSE} \]
\[ dfB = k - 1 \]
\[ dfE = (n - 1)(k - 1) \]
\[ p\text{-value} = \Pr(F > F_{dfB, dfE}) \]

\[ F_{dfB, dfE} \text{ is the } \alpha\text{ critical value of the } F \text{-distribution} \]

\[ \text{Critical value: } F_{0.05, 3, 12} = 3.89 \]

The hypothesis are:

\[ H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \]
\[ H_1: \text{at least one } \mu_i \text{ is different} \]

\[ F = \frac{MSB}{MSE} \]
\[ dfB = k - 1 \]
\[ dfE = (n - 1)(k - 1) \]
\[ p\text{-value} = \Pr(F > F_{dfB, dfE}) \]

\[ F_{dfB, dfE} \text{ is the } \alpha\text{ critical value of the } F \text{-distribution} \]

\[ \text{Critical value: } F_{0.05, 3, 12} = 3.89 \]

The hypothesis are:

\[ H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \]
\[ H_1: \text{at least one } \mu_i \text{ is different} \]

\[ F = \frac{MSB}{MSE} \]
\[ dfB = k - 1 \]
\[ dfE = (n - 1)(k - 1) \]
\[ p\text{-value} = \Pr(F > F_{dfB, dfE}) \]

\[ F_{dfB, dfE} \text{ is the } \alpha\text{ critical value of the } F \text{-distribution} \]

\[ \text{Critical value: } F_{0.05, 3, 12} = 3.89 \]

The hypothesis are:

\[ H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \]
\[ H_1: \text{at least one } \mu_i \text{ is different} \]

\[ F = \frac{MSB}{MSE} \]
\[ dfB = k - 1 \]
\[ dfE = (n - 1)(k - 1) \]
\[ p\text{-value} = \Pr(F > F_{dfB, dfE}) \]

\[ F_{dfB, dfE} \text{ is the } \alpha\text{ critical value of the } F \text{-distribution} \]

\[ \text{Critical value: } F_{0.05, 3, 12} = 3.89 \]
• The null hypotheses $H_0$: $\mu_1 = \mu_2 = \ldots = \mu_4$;

• The alternative hypotheses $H_1$: $\mu_1 \neq \mu_2$, or $\mu_1 \neq \mu_3$, or $\mu_1 \neq \mu_4$.

Can we reject $H_0$ with significance level $\alpha = 0.05$? \hfill \triangle

## 2 One-factor analysis of variance

Our problem is of the following form.

• **Object:**
  - $X_1, X_2, \ldots, X_m$ are **independent** random variables with **unknown** mean $\mu_1, \mu_2, \ldots, \mu_m$ and **unknown** variance $\sigma^2$.

• **Hypotheses:**
  - **Null Hypothesis** $H_0$: $\mu_1 = \mu_2 = \ldots = \mu_m$.
  - **Alternative Hypothesis** $H_1$: $\mu_1 \neq \mu_2$, or $\mu_1 \neq \mu_3$, \ldots, or $\mu_1 \neq \mu_m$.

• **Input:** Significance level $\alpha$, and $n_1$ many random samples for $X_1$, $n_2$ many random samples for $X_2$, \ldots, $n_m$ many random samples for $X_m$, i.e.,

```latex
\begin{array}{cccc}
\text{Samples for } X_1 & X_{11} & X_{12} & \ldots & X_{1n_1} \\
\text{Samples for } X_2 & X_{21} & X_{22} & \ldots & X_{2n_2} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\text{Samples for } X_m & X_{m1} & X_{m2} & \ldots & X_{mn_m} \\
\end{array}
```

• **Methodology:**
  - Compute $n = n_1 + n_2 + \ldots + n_m$;
  - For each $i \in \{1, 2, \ldots, m\}$, compute
    \[ \overline{X}_i := \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}, \]
    and
    \[ \overline{X} = \frac{1}{n} \sum_{i=1}^{m} n_i \overline{X}_i. \]
  - Compute SS(TO), SS(T), SS(E).
Reject $H_0$ if 

$$\frac{\text{SS}(T)}{(m - 1)} \geq F_\alpha(m - 1, n - m),$$

where $F_\alpha(m - 1, n - m)$ can be computed from Table VII in the textbook.

**Definition 2.** The total sum of squares is

$$\text{SS( TO)} := \sum_{i=1}^{m} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2.$$ 

The error sum of squares is

$$\text{SS(E)} := \sum_{i=1}^{m} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2.$$ 

The between-treatment sum of squares is

$$\text{SS(T)} := \sum_{i=1}^{m} n_i (\bar{X}_i - \bar{X})^2.$$ 

△

Note that there are formulas equivalent to Definition 2 that sometimes are simpler to compute:

$$\text{SS( TO)} = \left( \sum_{i=1}^{m} \sum_{j=1}^{n_i} X_{ij}^2 \right) - n(\bar{X})^2;$$

$$\text{SS(T)} = \left( \sum_{i=1}^{m} n_i (\bar{X}_i)^2 \right) - n(\bar{X})^2;$$

$$\text{SS(E)} = \text{SS( TO)} - \text{SS(T)}.$$ 

**Answer to Example 1.** From the sample data, we have

$$n = n_1 + n_2 + n_3 + n_4 = 12;$$

$$\bar{X}_1 = \frac{13 + 8 + 9}{3} = 10;$$

$$\bar{X}_2 = \frac{15 + 11 + 13}{3} = 13;$$

$$\bar{X}_3 = \frac{8 + 12 + 7}{3} = 9;$$

$$\bar{X}_4 = \frac{11 + 15 + 10}{3} = 12;$$

$$\bar{X} = \frac{(3)(10) + (3)(13) + (3)(9) + (3)(12)}{12} = 11.$$
which gives us

$$SS(TO) = (13 - 11)^2 + (8 - 11)^2 + (9 - 11)^2 + (15 - 11)^2 + (11 - 11)^2 + (13 - 11)^2$$
$$+ (8 - 11)^2 + (12 - 11)^2 + (7 - 11)^2 + (11 - 11)^2 + (15 - 11)^2 + (10 - 11)^2$$
$$= 80,$$

and

$$SS(T) = (3)(10 - 11)^2 + (3)(13 - 11)^2 + (3)(9 - 11)^2 + (3)(12 - 11)^2 = 30,$$

and

$$SS(E) = SS(TO) - SS(T) = 80 - 30 = 50.$$

This gives us

$$\frac{SS(T)}{(m - 1)} \cdot \frac{SS(E)}{(n - m)} = \frac{30}{3} \cdot \frac{50}{8} = 1.6.$$ 

On the other hand, the value for $F_\alpha(m - 1, n - m)$ is

$$F_\alpha(m - 1, n - m) = F_{0.05}(3, 8) = 4.07.$$

Since the former is smaller than the latter, we conclude that the test is inconclusive.

\[ \square \]

**Remark 3.** The notation $X_i$ here is written as $\mathbf{X}_i$ in the textbook. The notation $\mathbf{X}$ here is written as $\mathbf{X}_s$ in the textbook.

**Remark 4.** We can drop the assumption that $X_1, \ldots, X_m$ are normal random variables, assuming that $n_1, n_2, \ldots, n_m$ are large enough.

\[ \triangle \]

### 3 ANOVA table

**Definition 5.** The *analysis-of-variance table* (ANOVA table) is
<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares (SS)</th>
<th>Degrees of freedom</th>
<th>Mean square (MS)</th>
<th>F ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>SS(T)</td>
<td>m − 1</td>
<td>MS(T) = \frac{SS(T)}{m-1}</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MS(T) \cdot MS(E)</td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>SS(E)</td>
<td>n − m</td>
<td>MS(E) = \frac{SS(E)}{n-m}</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>SS(TO)</td>
<td>n − 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For example, the ANOVA table for Example 1 is

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares (SS)</th>
<th>Degrees of freedom</th>
<th>Mean square (MS)</th>
<th>F ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>30</td>
<td>3</td>
<td>10</td>
<td>1.6</td>
</tr>
<tr>
<td>Error</td>
<td>50</td>
<td>8</td>
<td>6.25</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>80</td>
<td>11</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>