# Introduction to Analysis of Variance (AOV, ANOVA)

R. A. Fisher, 1920s

## 1-way AOV

Idea: examine differences in means among several populations

$H_0$: $\mu_1 = \mu_2 = \cdots = \mu_4$ (common mean)

$H_a$: $\mu_1 \neq \mu_2 \neq \cdots \neq \mu_4$ (means are different somewhere)

Important assumption: all populations have a common variance ($\sigma^2$)

**Examples**

1. Experimental setting

Potted plants in a greenhouse; 4 fertilizer mixes; randomly select 5 plants for each treatment

$Y_{ij}$ = growth yield of the $j$th plant ($j = 1, 2, 3, 4, 5$) treated with mix $i$ ($i = 1, 2, 3, 4$)

$$Y_{1j} \sim \text{normal}(\mu_1, \sigma^2)$$

$$Y_{2j} \sim \text{normal}(\mu_2, \sigma^2)$$

$$Y_{3j} \sim \text{normal}(\mu_3, \sigma^2)$$

$$Y_{4j} \sim \text{normal}(\mu_4, \sigma^2)$$

2. Observational setting

Interested in quantifying differences in household incomes among 3 ethnic groups; randomly select $n_i$ households from the $i$th ethnic group

$Y_{ij}$ = annual income (or log(annual income)) of $j$th household selected ($j = 1, 2, ..., n_i$) in the $i$th ethnic group ($i = 1, 2, 3$)

$$Y_{1j} \sim \text{normal}(\mu_1, \sigma^2)$$

$$Y_{2j} \sim \text{normal}(\mu_2, \sigma^2)$$

$$Y_{3j} \sim \text{normal}(\mu_3, \sigma^2)$$

(Think about differences in *interpreting* the results between experimental and observational settings)

**Details of AOV**

Data:

| population | data | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | $y_{11}$ | $y_{12}$ | $\cdots$ | $y_{1n_1}$ | $n_1$ | $\overline{y}_{1\cdot}$ | ${s_1}^2$ |
| 2 | $y_{21}$ | $y_{22}$ | $\cdots$ | $y_{2n_2}$ | $n_2$ | $\overline{y}_{2\cdot}$ | ${s_2}^2$ |
| $\vdots$ | | | $\vdots$ | | | | |
| $t$ | $y_{t1}$ | $y_{t2}$ | $\cdots$ | $y_{tn_t}$ | $n_t$ | $\overline{y}_{t\cdot}$ | ${s_t}^2$ |

Also, $n_T = n_1 + n_2 + \cdots + n_t$ (total number of observations), and $\overline{y}_{\cdot\cdot} = \frac{1}{n_T}\sum_{i=1}^{t}\sum_{j=1}^{n_i} y_{ij}$ (mean of all the observations)

Under $H_0$, $\mu_1 = \mu_2 = \cdots = \mu_t = \mu$. Estimate $\mu$ under $H_0$ with $\overline{y}_{\cdot\cdot}$ . Estimate $\sigma^2$ under $H_0$ with

$$s_T^2 = \frac{\sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{\cdot\cdot})^2}{n_T - 1}$$

Numerator: $\Sigma\Sigma(y_{ij} - \overline{y}_{\cdot\cdot})^2 = \text{TSS}$ "total sum of squares" (total variability in the data)

Under $H_a$, $\mu_1 \neq \mu_2 \neq \cdots \neq \mu_t$ . Separate means are needed to describe the data. Estimate them with $\overline{y}_{1\cdot}$, $\overline{y}_{2\cdot}$, ..., $\overline{y}_{t\cdot}$ . Estimate $\sigma^2$ with the pooled estimator (using separate means for each group:

$$s_W^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_t - 1)s_t^2}{n_T - t}$$

Numerator: $\sum_i \sum_j (y_{ij} - \overline{y}_{i\cdot})^2 = \text{SSW}$ "within-sample sum of squares" (sum of squared departures from their group means).

Test statistic for hypothesis test based on following concept: does the model $H_a$ meaningfully reduce the amount of variability (noise) left over in the data? (i.e. are the extra mean parameters worthwhile?)

Test statistic is essentially a comparison of $s_W^2$ with $s_T^2$:

$\frac{s_W^2}{s_T^2}$ *small*: $H_a$ favored

$\frac{s_W^2}{s_T^2}$ *large*: $H_0$ favored

Note: can show that

$$\text{TSS} = (n_T - 1)s_T^2 = \sum_i \sum_j (y_{ij} - \overline{y}_{..})^2$$

$$= \underbrace{\sum_i \sum_j (y_{ij} - \overline{y}_{i.})^2}_{\text{SSW}} + \underbrace{\sum_i n_i (\overline{y}_{i.} - \overline{y}_{..})^2}_{\text{SSB}}$$

(SSB: "between-sample sum of squares")

Also can show that

$$\frac{s_W^2}{s_T^2} = \frac{\text{SSW}/(n_T - t)}{\text{TSS}/(n_T - 1)} = \frac{(n_T - 1)\text{SSW}}{(n_T - t)\text{TSS}}$$

$$= \frac{(n_T - 1)\text{SSW}}{(n_T - t)[\text{SSW} + \text{SSB}]}$$

$$= \frac{(n_T - 1)}{(n_T - t)[1 + (\text{SSB}/\text{SSW})]}$$

$$= \frac{(n_T - 1)}{(t - 1)\left[\frac{(n_T - t)}{(t-1)} + \frac{\text{SSB}/(t-1)}{\text{SSW}/(n_T - t)}\right]}$$

$$\frac{(n_T - 1)}{(t - 1)\left[\frac{(n_T - t)}{(t-1)} + f\right]}$$

Here

$$f = \frac{\text{SSB}/(t-1)}{\text{SSW}/(n_T - t)} = \frac{s_B^2}{s_W^2} \qquad \frac{\text{(between-sample variance)}}{\text{(within-sample variance)}}$$

$f$ *large*: $\text{H}_\text{a}$ favored

$f$ *small*: $\text{H}_0$ favored

If $\text{H}_0$ is true, then

$$F = \frac{S_B^2}{S_W^2} \sim \text{F}(t-1, n_T - t)$$

**Hypothesis test**

Hypotheses:

$\text{H}_0$: $\mu_1 = \mu_2 = \cdots = \mu_t = \mu$

$\text{H}_\text{a}$: $\mu_1 \neq \mu_2 \neq \cdots \neq \mu_t$

Test statistic:

$$f = \frac{\text{SSB}/(t-1)}{\text{SSW}/(n_T - t)} = \frac{s_B^2}{s_W^2}$$

Rejection region:

reject $\text{H}_0$ if $f \geq f_a$, the $100(1-\alpha)$th percentile of an $\text{F}(t-1, n_T - t)$ distribution