ANOVA (analysis of variance) is useful in comparisons involving several population means

One-way ANOVA deals with the effect of a single nominal factor on a single continuous response variable. When that one factor is a fixed factor, one-way ANOVA (often referred to as fixed-effects one-way ANOVA) involves a comparison of several (two or more) population means.

The Assumptions

1. **Independent random samples** (individuals, animals, etc.) have been selected from each of *k* populations or groups.
2. A value of a specified dependent variable has been recorded for each experimental unit (individual, animal, etc.) sampled.
3. The dependent variable is **normally distributed** in each population.
4. The **variance** of the dependent variable is the **same in each population** (this common variance is denoted as σ^2).

ANOVA Forms

One-way ANOVA model:

$$Y_{ij} = \mu_i + \epsilon_{ij}; j = 1, \ldots, n_i, i = 1, \ldots, k$$

where μ_i's are deterministic and $\epsilon_{ij} \sim_{iid} N(0, \sigma^2)$

Or

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}; j = 1, \ldots, n_i, i = 1, \ldots, k$$

where μ, α_i's are deterministic and $\epsilon_{ij} \sim_{iid} N(0, \sigma^2)$.

Some constraints on α_i's for avoiding over-parametrization, like

(1) $\sum_{i=1}^{k} n_i \alpha_i = 0$

Estimations: $\hat{\mu}_i = \bar{Y}_{i.} = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}, \hat{\mu} = \bar{Y}_{..} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}}{N}, \hat{\alpha}_i = \hat{\mu}_i - \hat{\mu} = \bar{Y}_{i.} - \bar{Y}_{..},$

$$\widehat{\sigma^2} = MSE = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{i.}\right)^2}{N-k}$$

(2) $\sum_{i=1}^{k} \alpha_i = 0$

Estimations: $\hat{\mu}_i = \bar{Y}_{i.} = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}, \hat{\mu} = \frac{\sum_{i=1}^{k} \bar{Y}_{i.}}{k}, \hat{\alpha}_i = \hat{\mu}_i - \hat{\mu}$

Balanced design (equal sample sizes): estimations are the same with the estimations under $\sum_{i=1}^{k} n_i \alpha_i = 0$ constraints.

(3) $\alpha_1 = 0$

Estimations: $\hat{\mu}_i = \bar{Y}_{i.} = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}, \hat{\mu} = \bar{Y}_{1.}, \hat{\alpha}_i = \hat{\mu}_i - \hat{\mu}$

Why constraints?

In order to get LSE of the parameters, we need to solve the following equations:

for any i = 1,2,…,k

$$\sum_{j=1}^{n_i} Y_{ij} = n_i\hat{\mu} + n_i\hat{\alpha}_i$$

Then we have k independent equations and k+1 parameters $(\hat{\mu}, \hat{\alpha}_1, \ldots, \hat{\alpha}_k)$ to estimate, so there are infinitely many solutions. To solve this problem, we need to impose the constraints.

Contrast

Functions of the parameters that have the form $\sum c_i \alpha_i$ where $\sum c_i = 0$ are called *contrasts*. For example, each difference of effects $\alpha_i - \alpha_j$ is a contrast. This is a quantity that is often of interest in an experiment. Other contrasts, such as differences of averages, may be of interest as well in certain experiments.

Example: An experimenter is trying to determine which type of non-rechargeable battery is most economical. He tests five types and measures the lifetime per unit cost for a sample of each. He also is interested in whether basic or heavy-duty batteries are most economical as a group. He has selected two types of heavy duty (groups 1 and 2) and three types of basic batteries (groups 3, 4, and 5). So, to study his second question, he tests the difference in averages:

$$H_0: \frac{\alpha_1 + \alpha_2}{2} = \frac{\alpha_3 + \alpha_4 + \alpha_5}{3}; H_A: \frac{\alpha_1 + \alpha_2}{2} \neq \frac{\alpha_3 + \alpha_4 + \alpha_5}{3}$$

Every difference of averages is a contrast.

Exercise: Every contrast $\sum c_i \alpha_i$ is a linear combination of the effect differences $\alpha_i - \alpha_j$ and is estimable, with least squares estimate $\sum c_i \hat{\alpha}_i = \sum c_i (\bar{Y}_{i.} - \bar{Y}_{..}) = \sum c_i \bar{Y}_{i.}$(under $\sum_{i=1}^{k} n_i \alpha_i = 0$ constraints)

Hypothesis Test

$$H_0: \mu_1 = \mu_2 = .. = \mu_k; H_A: k \ population \ means \ are \ not \ all \ equal$$

Reject→multiple-comparison procedures

F Test

$$F = \frac{between - group}{within - group} = \frac{\frac{\sum n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2}{k-1}}{\frac{\sum\sum (Y_{ij} - \bar{Y}_{i.})^2}{N-k}} \sim F(k-1, N-k) \ under \ H_0$$

Thus, for a given α, we would reject H_0 and conclude that some (i.e., at least two) of the population means differ from one another if

$F \geq F_{k-1, n-k, 1-\alpha}$

where $F_{k-1, n-k, 1-\alpha}$ is the $100(1-\alpha)\%$ point of the F distribution with $(k-1)$ and $(n-k)$

Reference:

Kleinbaum, David G., et al. *Applied regression analysis and other multivariable methods*. Cengage Learning, 2013.