There are two main parts in the analysis:

1. An overall test to see if there is statistical evidence that there exist *any* differences.
2. A more detailed *follow-up* analysis to decide which of the populations differ, and to estimate how large the differences are.

Let $r =$ number of levels of the explanatory variable (number of treatment groups for example). Let $n_i =$ number of cases (experimental units) in i^{th} group, and let $n_T = \Sigma_{i=1}^{r} n_i$ be the total number of observations.

Let the population mean parameters be μ_i, $i = 1, \ldots, r$.

Hypotheses to be tested:

H_0: $\mu_1 = \mu_2 = \cdots = \mu_r$ vs.

H_a: at least two of the means are not equal.

Assumptions:

- Each of the r population distributions is normal.
- The r standard deviations are all equal.
- All n_T observations are taken independently.

Let $\mu_i =$ the mean sales volume (number of cases) that would be seen in the whole population of stores, similar to those in the study, if Package Design i was used.

Independence assumption really has two parts. We assume the r groups are independent, and that observations within each group are independent.

Alternative hypothesis is the opposite of the null hypothesis.
H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
H_a: Not H_0
Better: H_a: For at least one pair of means μ_i, μ_j, $\mu_i \neq \mu_j$

Common error: H_a: At least one mean is not equal to the others ✗

H_a: $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$ ✗

Cell-means model

Let Y_{ij} denote the response of the j^{th} unit in the i^{th} group (i^{th} level of the factor).

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

where

- the μ_i are parameters,
- the errors, ϵ_{ij}, are independent $\mathcal{N}(0, \sigma^2)$

Another way to state the model:

Y_{ij} are independent $\mathcal{N}(\mu_i, \sigma^2)$, $\ \text{i} = 1, \ldots, r; \ \text{j} = 1, \ldots, n_i$

1. Y is sum of fixed and random components
2. $E(Y_{ij}) = \mu_i$
3. The variance of Y_{ij} is constant, equal to σ^2
4. Y_{ij} is normally distributed
5. The Y_{ij} are all independent

Example Kenton Food Company

Suppose we know $\mu_1 = 15$, $\mu_2 = 16$, $\mu_3 = 20$, $\mu_4 = 28$, $\sigma = 1.5$

Two of the observations are:

Y	package	id
11	1	1
19	2	4

In this hypothetical situation, find the error terms for these two observations.

Answer:

$\epsilon_{11} = Y_{11} - \mu_1 = 11 - 15 = -4$

$\epsilon_{24} = Y_{24} - \mu_2 = 19 - 16 = 3$

Cell means model is a linear model

Illustrate why, using a case involving $r = 3$ treatments and two replicates per treatment.

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon},$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{pmatrix}.$$

Notation for cell and grand means

Let $Y_{i.} = \sum_{j=1}^{n_i} Y_{ij}$, be the sum of the observations in Group i, for $i = 1, \ldots, r$.

Then the mean of the i^{th} group is:

$$\overline{Y}_{i.} = \frac{1}{n_i} Y_{i.}$$

The group means are also called the *cell means*.

Let $Y_{..} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} Y_{ij}$ be the total of all the observations.

Then the *grand mean* is:

$$\overline{Y}_{..} = \frac{1}{n_T} Y_{..}$$