

A statistical model of an experiment

<http://marekrychlik.com/node/29>

The response variable captures the measurements of the characteristic of the experimental units or subjects observed after the treatment is applied. We will denote the response variable by y .

The statistical model for comparative studies is based on an assumption that there is a reference population of subjects or experimental units. The population is often conceptual. Thus, we may consider a sample of 8 car engines in reference to the population of all car engines with similar characteristics (make, model, etc.). The experimental units are assumed to be selected at random from their reference population.

The population variance σ^2 is assumed to be the same for each of the populations in a comparative study and unaffected by the treatment. This is called homoskedasticity (or homoscedasticity) or homogeneity of variances. It is important to know to what degree this assumption holds. There are ways to test populations for homoscedasticity. An example violation would be if the treatment resulted in drastically increased variance.

Let us assume that we have t groups of observational units, with group i consisting of r_i units.

The cell mean model can be described by a formula:

$$y_{ij} = \mu_i + \epsilon_{ij}$$

where $i = 1, 2, \dots, t$ and $j = 1, 2, \dots, r_i$. If it can be assumed that all groups are of the same size then $r_i = r$. Under this assumption, some parts of the mathematical analysis can be simplified.

The meaning of the symbols is as follows:

- μ_i represents the mean of the i -th treatment population;
- ϵ_{ij} is the experimental error.

This is a linear statistical model. The experimental error variance σ^2 is assumed to be the same for all treatment populations, and it coincides with the variance of the reference population (by homoscedasticity).

The reduced model is represented by the formula:

$$y_{ij} = \mu + \epsilon_{ij}$$

The difference is that the treatment group mean is assumed to be independent of the treatment. The full model represents the hypothesis that the means of the treatment populations differ (the alternative hypothesis: H_a). The reduced model represents the hypothesis that the means are unaffected by the treatment (the null hypothesis: H_0).

Thus:

- H_0 says that $\mu_i = \mu$ are the same for all treatment populations;
- H_a says that for at least one pair (i, j) we have $\mu_i \neq \mu_j$.

The statistical package R has special notation and an underlying data object to represent linear models. See help for the symbol ' '.

The method of least squares is used to fit the model to the experimental data.