Basics of Design of Experiments and ANOVA

So far we concentrated on analysis of a given experiment or data. Structure of the experiment is now explored. Design of experiments is a study of construction and analysis of experiments where purposeful changes are made to input variables of a process or system so as to observe and identify the reasons for changes in the output response. A cause-effect mechanism is of interest here.

Example. Different or different amounts of fertilizers versus yield of a crop

Experimental designs are used mostly for comparative experiments:
Comparing treatments in a clinical trial
Comparing factors (fertilizers, crop patterns etc.) in agricultural experiments

Randomization, replication, blocking and confounding of effects are some important concepts in this context. Randomization means that each subject has the same chance of being placed in any given experimental group. Then factors which cannot be controlled need not be considered since their effects are averaged out.
Replication means having multiple subjects in all experimental groups, ensuring that 'within group' variation can be estimated.
Blocking and confounding of effects will be considered later.

Consider the following example of a completely randomized design.

Example. Monosodium glutamate (MSG), a common ingredient of preserved food is known to cause brain damage in various mammals. In a study of the other effects, weight of ovaries (mg), both for a sample of rats treated with MSG and for an independent control sample of similar but untreated rats were obtained:

	sample size (n_i)	sample mean	sample s.d.
MSG	10	29.35	4.55
Control	12	21.86	10.09

Consider the linear model,

$$y_i = \begin{cases} \mu_m + \epsilon_i & i = 1, 2, \ldots, n_1; \\ \mu_c + \epsilon_i & i = n_1 + 1, n_1 + 2, \ldots, n_1 + n_2, \end{cases}$$

ϵ_i i.i.d $N(0, \sigma^2)$. Write it in the vector/matrix form, $Y = X\beta + \epsilon$:

$$
\begin{pmatrix} y_1 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \end{pmatrix}
=
\begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix}
\begin{pmatrix} \mu_m \\ \mu_c \end{pmatrix} + \epsilon,
$$

$\epsilon \sim N_{n_1+n_2}(0, \sigma^2 I)$. Then

$$
X'X = \left\{ \begin{array}{cc} n_1 & 0 \\ 0 & n_2 \end{array} \right), \quad \text{so } (X'X)^{-1} = \begin{pmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{pmatrix}.
$$

Therefore,

$$
\begin{pmatrix} \hat{\mu}_m \\ \hat{\mu}_c \end{pmatrix} = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_m \\ \mu_c \end{pmatrix}, \sigma^2 \begin{pmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{pmatrix} \right),
$$

independent of

$$
\text{RSS} = \sum_{i=1}^{n_1} (y_i - \bar{y}_1)^2 + \sum_{i=n_1+1}^{n_1+n_2} (y_i - \bar{y}_2)^2 \sim \sigma^2 \chi^2_{n_1+n_2-2}.
$$

We want to compare μ_m with μ_c. $H_0 : \mu_m = \mu_c = 0$ is meaningless; $H_0 : \mu_m = \mu_c$ is of interest. i.e., $H_0 : (1 \quad -1) \begin{pmatrix} \mu_m \\ \mu_c \end{pmatrix} = 0$. Note,

$$
\frac{(\bar{y}_1 - \bar{y}_2 - (\mu_m - \mu_c)) / \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}{\sqrt{\text{RSS}/(n_1 + n_2 - 2)}} \sim t_{n_1+n_2-2}.
$$

If H_0 is true, then

$$
\frac{(\bar{y}_1 - \bar{y}_2) / \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}{\sqrt{\text{RSS}/(n_1 + n_2 - 2)}} \sim t_{n_1+n_2-2},
$$

or equivalently,

$$
\frac{(\bar{y}_1 - \bar{y}_2)^2 / \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}{\text{RSS}/(n_1 + n_2 - 2)} \sim F_{1, n_1+n_2-2}.
$$

The design in this experiment has complete randomization. The observations inside the groups are independent, and also the two samples are independent.

2

For this reason, the design is called a completely randomized design. We can generalize this procedure if we want to compare k means, as will be done later.

Paired differences - example of a block design

Sometimes independent samples, such as the ones in a completely randomized design, from two (or $k > 2$) populations is not an efficient way for comparisons. Consider the following example.

Example. It is of interest to compare an enriched formula with a standard formula for baby food. Weights of infants vary significantly and this influences weight gain more than the difference in food quality. Therefore, independent samples (with infants having very different weights) for the two formulas will not be very efficient in detecting the difference. Instead, pair babies of similar weight and feed one of them the standard formula, and the other the enriched formula. Then observe the gain in weight:

pair	1	2	3	...	n
enriched	e_1	e_2	e_3	...	e_n
standard	s_1	s_2	s_3	...	s_n

However, the samples may not be treated as independent but correlated. The n pairs of observations, $(e_1, s_1), \ldots, (e_n, s_n)$ may still be treated to be uncorrelated (or even independent). These n pairs are like n independent blocks, inside each of which we can compare enriched with standard. This is the idea of blocking and block designs. Blocks are supposed to be homogeneous inside, so comparison of treatments within blocks becomes efficient.