ONE-WAY ANALYSIS OF VARIANCE MODEL

The simplest type of analysis of variance considers a situation that generalizes the two-sample t-test situation by allowing more than two groups.

The situation:

1. We have a *response variable* Y.

2. We have v populations $G_1, G_2, \ldots G_v$ on which the response variable is defined. In an experiment, these will be defined by the levels of a treatment factor ("treatment groups"): G_i is the population that has received the ith treatment (that is, the ith level of the treatment factor)

3. Let Y_i denote the response for the population G_i. (i.e, $Y_i = Y|G_i$, Y restricted to G_i.)

4. Let μ_i be the mean of Y on the ith population G_i. (i.e., $\mu_i = E(Y_i) = E(Y|G_i)$ (μ_i is sometimes called the *true mean* for the ith treatment or population.)

5. Let $\varepsilon_i = Y_i - \mu_i$. ($\varepsilon_i$ is a new random variable, called the ith *error*)

Example: Continuing with the example of testing computer packages to teach a programming language, suppose we were comparing 3 such packages rather than 2. Then
        Y = score on the test given to the subjects
        v = 3
        G_i = all conceivable engineers who have used the ith package.
        μ_i = mean score of all conceivable engineers who have used the ith package

Note that (5) can be re-expressed as

    $Y_i = \mu_i + \varepsilon_i$

(These *model equation* is sometimes called a *linear* or *additive* model. This form of the model equations is called the *means model*.)

*Model assumptions***:**

    1. For each i, a simple random sample of size r_i from population G_i is gathered.

2. The samples are independent.

3. Each error variable ε_i is normally distributed.

4. All error variables ε_i have the same variance σ^2.

Comments:
1. For an experiment, assumptions (1) and (2) can be combined to say that *experimental units are randomly assigned to treatments*, subject only to the constraint that the sample size for the ith treatment is r_i. An experimental design in which experimental units are randomly assigned to treatments (subject only to the constraint that the sample size for the ith treatment is r_i) is said to be *completely randomized*.
2. If all the r_i's are equal, then the design is said to be *balanced*.
3. Assumptions (3) and (4) can be combined to say $\varepsilon_i \sim N(\mu_i, \sigma^2)$
4. Students who have had regression should recognize the similarities to a linear regression model with indicator variables representing a categorical variable.

Alternate formulations of the model equation.

1. Let $\mu = E(Y)$ (the *overall mean*) and let $\tau_i = \mu_i - \mu$. Then the model equation can be rephrased as

$$Y_i = \mu + \tau_i + \varepsilon_i .$$

(τ_I is called the *effect* of the ith treatment on the response. This version of the model is sometimes called the *effects model*.)

2. Many people prefer to state the model equation in terms of the sample random variables as follows:

Let Y_{it} be the random variable that represents the response from the tth observation from G_i (in terms of an experiment: the response from the tth observation of the ith treatment). Let $\varepsilon_{it} = Y_{it} - \mu_i$. Then the model equation is stated as:

$$Y_{it} = \mu_i + \varepsilon_{it}$$

or $\quad Y_{it} = \mu + \tau_i + \varepsilon_{it} .$

In this formulation, the model assumptions become:

    a) The ε_{it} are independent random variables.

    b) For each i and t, $\varepsilon_{it} \sim N(\mu_i, \sigma^2)$

Note: The alternate model formulation as presented here is called a *fixed effects model*, since we are assuming that we have specified treatments fixed by the experimenter. Note that in the fixed effects model, the τ_i's are parameters. This model could be generalized to a situation where the treatments are a random sample from a larger population of treatments. In this case, the τ_i's are random variables. This generalization is called a *random effects model*. Random effects models are discussed in Chapter 17.

Dot notation

The following notational conventions are convenient when we are dealing with lots of subscripts:

- A dot in a subscript position means "add over all values of the subscript in that position."
- If there is a bar over the variable as well as a dot in the subscript position, then divide by the number of possibilities for the subscript as well as add over all values of the subscript. (In other words, take the average over all values of the subscript.)

Examples:

$$Y_{i\cdot} = \sum_{t=1}^{v} Y_{it} \qquad \bar{Y}_{i\cdot} = \frac{1}{v}\sum_{t=1}^{v} Y_{it} \qquad Y_{\cdot t} = \sum_{i=1}^{r_i} Y_{it} \qquad Y_{\cdot\cdot} = \sum_{t=1}^{v}\sum_{i=1}^{r_i} Y_{it}$$