ONE-WAY ANALYSIS OF VARIANCE MODEL

- The simplest type of analysis of variance
- Generalizes the two-sample equal variance t-test situation to more than two groups.

The situation:

1. *Response variable* Y (e.g., score on exam)

2. v populations $G_1, G_2, \ldots G_v$ on which the response variable is defined.
    - e.g., "treatment groups": G_i is the population that has received the i^{th} treatment (or: the i^{th} level of the treatment factor)

3. Y_i: response for the population G_i.
    - i.e, $Y_i = Y|G_i$, Y restricted to G_i.

4. μ_i = the mean of Y on the i^{th} population G_i.
    - i.e., $\mu_i = E(Y_i) = E(Y|G_i)$
    - μ_i is sometimes called the *true mean* for the i^{th} treatment or population.

5. $\varepsilon_i = Y_i - \mu_i$.
    - ε_i is a new random variable
    - ε_i = *i^{th} error*

Example: Testing computer packages to teach a programming language, but comparing 3 such packages rather than 2.

- Y =
- v =
- G_i =
- μ_i =

Note:
- (5) can be re-expressed as $Y_i = \mu_i + \varepsilon_i$
- *model equations*
- a *linear* or *additive* model.
- *means model*.

***Model assumptions*:**

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

2. The samples are independent.

3. Each ε_i is normally distributed.

4. All ε_i's 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 i^{th} treatment is r_i. i.e.,the experiment is *completely randomized*.

2. *Balanced design*: When all r_i's are equal.

3. Assumptions (3) and (4) combined: $\varepsilon_i \sim N(0, \sigma^2)$

4. Note similarities to a linear regression model with indicator variables representing a categorical variable.

Alternate formulations of the model equations.

1. Letting $\mu = E(Y)$ (the *overall population mean*) and $\tau_i = \mu_i - \mu$, the model equation becomes:

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

- τ_i : the *effect* of the i^{th} treatment on the response.
- " *effects model*"

2. In terms of the sample random variables:

Y_{it} = the random variable giving the response from the t^{th} observation from G_i (e.g., the response from the t^{th} observation of the i^{th} treatment).

$$\varepsilon_{it} = Y_{it} - \mu_i.$$

The model equation becomes:

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

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

Model assumptions become:

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

b) For each i and t, $\varepsilon_{it} \sim N(0, \sigma^2)$

Note: This is a *fixed effects model*: We are assuming that we have specified treatments fixed by the experimenter. So the τ_i's are parameters.

A generalization: The treatments are a random sample from a larger population of treatments. So the τ_i's are random variables. (*random effects model* -- discussed in Chapter 17.)

Dot notation

Convenient notational conventions:

- A dot in a subscript position means "add over all values of the subscript in that position."

Examples:

$$Y_{i\bullet} = \sum_{t=1}^{r_i} Y_{it} \qquad Y_{\bullet t} = \sum_{i=1}^{v} Y_{it} \qquad Y_{\bullet\bullet} = \sum_{i=1}^{v}\sum_{t=1}^{r_i} Y_{it}$$

- A bar over the variable as well as a dot in the subscript position means: divide by the number of possibilities for the subscript as well as add over all values of the subscript. (i.e., take the average over all values of the subscript.)

Example:

$$\overline{Y}_{i\bullet} = \frac{1}{r_i}\sum_{t=1}^{r_i} Y_{it}$$

More examples: