Statistics 3515 **Lecture 1**

Single Factor Completely Randomized Experiments

In an experiment to compare different treatments, each treatment must be applied to several different experimental units. This is because the response from different units varies, even if the units are treated identically. The simplest experimental design in which, in order to compare a treatments, treatment i is applied to n_i units, i=1,....,a. In many experiments n_1=....=n_a, but this is not necessary or even not desirable. This design, in which the only recognizable difference between units is the treatments which are applied to those units, is called a ***single factor completely randomized design*.** With this design, the assignments are made completely at random. This complete randomization provides that every experimental unit has an equal chance to receive any one of the treatments or, equivalently that all combinations of experimental units are assigned to the different treatments are equally likely.

A completely randomized design is particularly useful when the experimental units are quite homogeneous. This design is very flexible; it accommodates any number of treatments and permits different sample sizes for each of the treatments. Its chief disadvantage is that, when the experimental units are heterogeneous, this design is not as efficient as other statistical designs.

Terminology:

A *factor* is an independent variable to be studied in an investigation. For example in an investigation of how cotton content affects the tensile strength of a new synthetic fiber, the factor studied is cotton content. similarly, to study the relationship between sales of cereal and four different package designs, the factor is package design.

A *level* of a factor is a particular form of that factor. In the synthetic fiber study, the product development engineer has selected fibers with 15%, 20%, 25%, 30% and 35% cotton. Those are the five levels of the factor in that study. In the cereal study, there are four levels for the factor of package design. In the first example the factor is a *quantitative* one, while in the second example it is a *qualitative* one.

In a single factor experiment, a *treatment* corresponds to a factor level. In multi factor studies, a treatment will correspond to a combination of factor levels. In a single factor experiment if the levels of the factors are chosen at random, we will say it is a *random (effects) model*, otherwise it will be called a *fixed (effects) model*.

Type of Data:

Observational Data - data that is obtained without controlling the independent variable(s) of interest.

Experimental Data - data that is obtained by the experimenter by controlling the independent variable(s).

Example: New synthetic fiber study

Data (lb/inch square): tensile strength of the new synthetic fiber.

	Observation						
Cotton Percentage	1	2	3	4	5	Total	Mean
15	7	7	15	11	9	49	9.8
20	12	17	12	18	18	77	15.4
25	14	18	18	19	19	88	17.6
30	19	25	22	19	23	108	21.6
35	7	10	11	15	11	54	10.8
						376	15.04

It is always a good idea to examine experimental data graphically. For example one can present a boxplot and/or a scatter plot of tensile strength vs cotton percentage. In the SAS output the letters are the individual observations and the rectangles in the boxplot are the sample means. Both graphs indicate that tensile strength increases as cotton content increases, up to 30%. Beyond 30% cotton there is a sizable decrease in tensile strength. The scatter diagram supports the fact that the variability does not depend on cotton content. From the graphical display one would suspect that cotton content affects tensile strength and that around 30% cotton one would get maximum strength.

Analysis of the Fixed Effects Model

Model:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}\ ,\quad j=1,....,n_i \ \text{ and } i=1,....,a.$$

Y_{ij} is the jth observation for the ith treatment, μ is the overall mean representing the common effect for the entire experiment, τ_i is the effect of the ith treatment, and ε_{ij} is the random error present in the jth observation for the ith treatment.

Assumptions: 1. ε_{ij} are iid $N(0,\sigma_e^2)$

2. $\sum_{i=1}^{a} \tau_i = 0.$

From the expression for the model it follows that for $1 \le j \le n_i$ and $1 \le i \le a$

$$E(Y_{ij}) = \mu + \tau_i = \mu_i$$

is the mean of the observations in the ith group. The analysis of this experiment consists of testing

$H_0 :\ \tau_i = 0$ for all i vs $H_a :$ not all τ_i are 0,

which equivalent to testing

$H_0 :\ \mu_i = \mu$ for all i vs $H_a :$ not all $\mu_i = \mu$.

To test the above hypothesis the F test in a one way analysis of variance is used. The anova approach has two purposes. First, it provides a subdivision of the total variability between the experimental units into separate components, each component representing a different source of variability, so that the relative importance of the different sources can be assessed. Second, and more important, it gives an estimate of the underlying variability between units which provides a basis for inferences about the effects of the applied treatments. We now proceed to develop these for our model.

Notation: $Y_{i.} = \sum_{j=1}^{n_i} Y_{ij}$, $Y_{..} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} Y_{ij}$, $N = \sum_{i=1}^{a} n_i$

$$\bar{Y}_{i.} = Y_{i.} / n_i, \quad \bar{Y}_{..} = Y_{..} / N = \sum_{i=1}^{a} n_i \, Y_{i.} / N.$$

From the model we get the following identity

$$Y_{ij} - \mu = (Y_{ij} - \mu_i) + (\mu_i - \mu),$$

which remains valid when we replace the parameters by their estimates:

$$Y_{ij} - \bar{Y}_{..} = (Y_{ij} - \bar{Y}_{i.}) + (\bar{Y}_{i.} - \bar{Y}_{..}).$$

The above equation states that the deviation of each observation from the overall mean can be decomposed into two parts: the deviation of the observation from its treatment mean plus the deviation of the treatment mean form the overall mean. If we square both sides of the above equation and sum over i and j we get the following *fundamental equation of the analysis of variance:*

$$\sum_{i=1}^{a} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 + \sum_{i=1}^{a} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2,$$

or

$$SS_{total} = SS_{error} + SS_{treatment}.$$

The term on the left hand side represents the total variability in the data. The first term on the right hand side of the identity represents the total variability within each of the a treatments.
Since we have assumed that the variances within the a treatments are equal if we divide that term by

$$\sum_{i=1}^{a} (n_i - 1) = N - a$$

we get an unbiased estimator of the variance σ^2, which valid regardless of the null hypothesis being true or not true.

Now, if H_0 is true, then the second term divided by a - 1, is also an unbiased estimator of σ^2. Moreover the two estimators are independent of each other and their quotient denoted by

$$F = \frac{SS_{treatment}/(a-1)}{SS_{error}/(N-a)}$$

has an F distribution with a-1 and N-a degrees of freedom. Since the numerator gets large when H_0 is not true, while the denominator remains stable, we reject H_0 for large values of F. Table IV on pages A-6 to A-10 gives the critical values for the upper tail area = α for the F distribution, for selected values of α.

For this example one can show that:

$SS_{total} = 636.96$

$SS_{treatment} = 475.76$

$SS_{error} = 161.20$

F = 14.76

From Table IV, page A-10 in the Appendix, we get that the critical value for our data set for α=.01 is 4.43 ($\upsilon_1=4, \upsilon_2=20$). Therefore, we can reject the null hypothesis at the .01 level. A more accurate result can be obtained from SAS output.