Math 373  Lecture 26

ANOVA: Analysis of Variation

We want an optimal treatment for growing tomatoes. Tomato plants are the experimental units. The health of a tomato plant depends on the amounts of phosphate fertilizer, of nitrogen fertilizer, of water and of shade. Factors are the variables (amounts of fertilizer and water) which we control. The value of a factor is its level (if ½ lb. of nitrogen fertilizer is used, the nitrogen factor has level ½). A treatment is a particular combination of factor levels. The measurement we want to maximize is the response. Suppose we want to maximize the total weight of tomatoes produced by each plant. Then weight is the response. To determine how the various factors affect the response (total weight), we plant tomato plants in some number \(k\) of plots. All plants in the same plot get the same treatment; different plots get different treatments. If there are \(k\) plots, then \(k\) treatments will be tested. Plants in the first plot might have a treatment with low levels of fertilizer but a high level of water, plants in the second plot might have both a high level of fertilizer and water, plants in the third might have a high level of phosphate, a low level of nitrogen and a medium level of water, ... We want a treatment (the combination of levels of fertilizer, water and shade) which produces the optimal response (the most pounds of tomatoes).

The responses of the plants in the \(k\) plots (one plot for each treatment) are samples from the \(k\) treatment populations. Some of the variation between the responses (total weight of tomatoes) of two plants is due to random chance or natural variation or statistical error. The variation between plants in the same plot is due entirely to this “error” since they get the same treatment. The variation between plants in different plots will be partly due to sampling error and partly due to the difference in treatment. Our job is to separate the variation due to the treatment (treatment variation) from differences due to random variation (error variation).

In the following two figures, \(\{a,b,c\}\) are responses to one treatment and \(\{x,y,z\}\) are the set of responses to a second treatment.

\[
\begin{array}{cccc}
  a & b & c \\
  x & y & z \\
  a & x & b & y & z & c
\end{array}
\]

In the left experiment, there is very little variation within \(\{a,b,c\}\) or \(\{x,y,z\}\) but a substantial difference between the two population means. In the right experiment, there is only a small difference between the two population means but there is a substantial variation within each population. Hence the left experiment has a small error variation and a large treatment variation; the right experiment has a large error variation and a small treatment variation.

We wish to determine if any two of the \(k\) populations differ significantly from each other. We could test each pair of populations for a significant difference between their means. But if there are 11 populations, the number of possible pairs is \(k(k-1)/2 = 11 \times 10/2 = 55\). The following method is more efficient and has less error.

Given: \(k\) independent “treatment” populations, i.e., \(k\) blocks, with means \(\mu_1, \mu_2, ..., \mu_k\). We assume the populations are normal and have the same variance \(\sigma^2\). Suppose we take random samples of sizes \(n_1, n_2, ..., n_k\) from these respective populations. Let \(\bar{x}, \bar{x}_2, ..., \bar{x}_k\) and \(s_1, s_2, ..., s_k\) be their means and std. devs.

Let \(n = n_1 + n_2 + ... + n_k\) be total number of observations and \(\bar{x}\) be the mean of all the observations. Since the populations have the same std. dev. \(\sigma\), \(s_1, s_2, ..., s_k\) are all estimates of \(\sigma\).

The degrees of freedom, \(df\), is given for each sum of squares below. To get the total variation (the mean value of the sum of squares) we divide the sum of squares by the number of degrees of freedom.

\[
s_{total}^2 = \frac{\sum(x_i - \bar{x})^2}{(n-1)}
\]

**DEFINITION.** The total sum of squares is

- Total SS = \(\sum_{i=1}^{n} (x_i - \bar{x})^2 = (n-1)s_{total}^2\), \(df = n-1\),
- \(MS, mean\ squares\ (total\ variation)\), \(= total\ SS/df\),
- \(SST, sum\ of\ squares\ for\ treatments\), \(= \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2\), \(df_t = k-1\),
- \(MST, mean\ squares\ for\ treatments\), \(= SST/df_t\)

Special case. If \(n_1 = n_2 = ... = n_k = b\), then \(SST = bSS_t = b(k-1)s_t^2\) where \(SS_t\) is the treatment sum of squares and \(s_t\) is the std. dev. of the treatment squares.

- \(SSE, sum\ of\ squares\ for\ error\), \(= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{i,j} - \bar{x}_i)^2 = \sum_{i=1}^{k} SS_i + SS_2 + ... + SS_k\)
- \(MSE, mean\ squares\ for\ error\), \(= SSE/df_e\)

\(s\), the pooled estimate of the common \(\sigma\), \(= \sqrt{MSE}\).

\(F = MST/MSE\). Note the differences between \(s, s_{total}\) and \(s_t\).

**THEOREM.** Total SS = SST + SSE. Likewise, \(df = df_t + df_e\). The last is easy: \(df = n-1 = (k-1) + (n-k) = df_t + df_e\).

Two of the \(k\) populations differ significantly from each other \(\iff\) the treatment variation MST significantly exceeds the error variation MSE. As before, we compare variances by taking ratios, \(F = MST/MSE\).

\(MST > MSE\ \iff\ \text{F} > 1\) if \(F > 1\). Since \(F\) is a ratio of squares, it has the \(F\) distribution.

**F test for determining if two or more population means differ significantly.**

- \(H_0: \mu_1 = \mu_2 = ... = \mu_k\) \(H_0: \text{At least two means differ.}\)
- \(H_0: MST \leq MSE\ \iff\ H_0: \text{MST > MSE}\)
- \(H_0: F \leq 1\ \iff\ H_1: \text{F} > 1\) null region: \(F \in [0,1]\)

Test statistic: \(F = MST/MSE\) with \(df_t = df_f, df_e = \infty\). Acceptance region for \(F\): \([0, F_{0.05}]\).

**Suppose we have 3 populations each with samples of 11 observations. Suppose the sample means are 40, 45 and 50 and the sample std. devs. are 4, 3, and 5. Do any two of the means differ significantly?**

- \(SST = bSS_t = b(k-1)s_t^2 = 11(3-1)s_t^2 = 11(50) = 550\)
- \(SSE = (11-1)(4^2) + (11-1)(3^2) + (11-1)(5^2) = 500\)
- \(SST = 550\ \iff\ SSE = 500\ \iff\ \alpha = 0.05\)
- \(df_t = df_f = (3-1) = 2\ \iff\ df_e = 30\ \iff\ F_{0.05} = 3.32\)
- \(MST = SST/df_t = 275, MSE = SSE/df_e = 55.56, F = MST/MSE = 4.95\)
- Acceptance region for \(F\): \([0, 3.32]\)

Do two means differ significantly? Yes, \(F = 4.95 \notin [0, 3.32]\) = acceptance region for no significant difference.