ANOVA Notes

Math 481

April 27, 2006

Given k normal populations with a common variance σ^2, we wish to perform the following hypothesis test:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1 &: \mu_i \neq \mu_j \text{ for at least one } i,\, j, \end{aligned}$$

where μ_i is the mean of the i-th population; $1 \leqq i \leqq k$.

We will construct a test statistic a test statistic which compares the variability *between* samples to the variability *within* samples.

The variability within samples is measured by

$$s_W^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \cdots + (n_k-1)s_k^2}{(n_1-1)+(n_2-1)+\cdots+(n_k-1)}.$$

Notice that in the $k = 2$ case the above formula reduces to the familiar "pooled variance" s_p^2 that we encountered when testing hypotheses involving the difference of two means with small samples.

The variability between samples is measured by constructing a data set whose points are the means of the samples drawn from the k populations, and then calculating the sample variance of this data set. In other words, draw a sample from population 1, calculate its mean $\bar{x}_1$ and write it down. Then draw a sample from population 2, calculate its mean $\bar{x}_2$ and write it down. Continue until you have repeated this process for all k populations. We now have written down a data set

$$S = \{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k\}.$$

Now find the sample varaiance of S and call this s_B^2. The test statistic we define is the ratio s_B^2/s_W^2, which turns out to have an F distribution with $k-1$ and $(n_1-1)+(n_2-1)+\cdots+(n_k-1)$ degrees of freedom.

In practice, since the quantities s_B^2 and s_W^2 are rather tedious to calculate directly from the definitions, we simplify the necessary calculations using some algebraic shortcuts. The price that we pay for this is that additional notation must be introduced, and the motivation behind the definition of the test statistic is obscured.

- One sample is drawn from each of the k populations; the size of the ith sample is denoted n_i.
- The data from the sample drawn from the ith population is

$$x_{i1}, x_{i2}, \ldots, x_{in_i}.$$

- The total sample size is $n_T = n_1 + n_2 + \cdots + n_k$.
- The sum of the sample data drawn from the ith population is denoted

$$x_{i\cdot} = x_{i1} + x_{i2} + \cdots + x_{in_i}.$$

- The mean of the sample data drawn from the ith population is denoted

$$\bar{x}_{i\cdot} = \frac{x_{i\cdot}}{n_i}.$$

- The *grand total* of all observations is denoted

$$x_{\cdot\cdot} = x_{1\cdot} + x_{2\cdot} + \cdots + x_{k\cdot} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}.$$

- The *grand mean* is denoted

$$\bar{x}_{\cdot\cdot} = \frac{x_{\cdot\cdot}}{n_T} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}}{\sum_{i=1}^{k} n_i}.$$

- The *total sum of squares* is

$$SS_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}^2 - \frac{x_{\cdot\cdot}^2}{n_T}.$$

- The *sum of squares between samples* is

$$SS_B = \sum_{i=1}^{k} \frac{x_{i\cdot}^2}{n_i} - \frac{x_{\cdot\cdot}^2}{n_T}.$$

- The *sum of squares within samples* is

$$SS_W = SS_T - SS_B.$$

- The *mean square between samples* is $s_B^2 = \frac{SSB}{k-1}$.
- The *mean square within samples* is $s_W^2 = \frac{SSW}{n_T - k}$.

Often the most important quantities from the above are summarized in tabular form called an *analysis of variance* table.

Source	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*	*Test statistic*
Between samples	SS_B	$k-1$	s_B^2	s_B^2/s_W^2
Within samples	SS_W	n_T-k	s_W^2	
Totals	SS_T	n_T-1		

We reject the null hypothesis at the significance level α if the test statistic is greater than $F_{\alpha,k-1,n_T-k}$.

Example:

Given three normal populations with a common variance σ^2, we wish to perform the following hypothesis test:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \mu_3 \\ H_1 &: \mu_1 \neq \mu_2 \text{ or } \mu_1 \neq \mu_3 \text{ or } \mu_2 \neq \mu_3, \end{aligned}$$

where μ_i is the mean of the i-th population; $1 \leqq i \leqq 3$.

The sample from the first population is 1, 4, 7.

The sample from the second population is 2, 2, 3, 4.

The sample from the third population is 2, 2, 5, 6, 7, 1.

We summarize the information in the following ANOVA table:

Source	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*	*Test statistic*
Between samples	3.65	2	1.82	0.35
Within samples	51.58	10	5.16	
Totals	55.23	12		

Since $0.35 < 4.10 = F_{0.05,2,10}$, we fail to reject the null hypothesis.

Suggested exercises: Use the data provided from textbook exercises 15.16, 15.17, 15.18, and 15.19 to perform analyses of variance to test appropriate hypotheses as to whether the population means are all equal or not.

Homework to turn in on April 27: In addition to the three problems described at `http://www.math.rutgers.edu/~asills/teach/spr06/481hw.txt`, use the data from textbook exercise 15.20 to test the hypothesis

$$\begin{aligned} H_0 &: \mu_X = \mu_Y = \mu_Z \\ H_1 &: \mu_X \neq \mu_Y \text{ or } \mu_X \neq \mu_Z \text{ or } \mu_Y \neq \mu_Z, \end{aligned}$$

at the level $\alpha = 0.05$ level using the analysis of variance technique.