STA 580 — Fall 2008 — Dr. Charnigo

Lecture 10

Multi-sample problems and the one-way analysis of variance

Introduction. In Lecture 7, I described how to use two independent samples to test whether two population means are equal. However, sometimes we may want to test whether three or more population means are equal.

A recurring example will come from "Obstetrics" on page 620 and pages 1 through 4 of {ANOVAExamples.pdf}. Here we are interested in testing $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ against the complementary alternative,[1] where μ_1 is the mean birthweight of infants born to mothers who never smoked, μ_2 is the mean birthweight of infants born to mothers who gave up smoking during pregnancy, μ_3 is the mean birthweight of infants born to mothers who smoked less than a pack a day while pregnant, and μ_4 is the mean birthweight of infants born to mothers who smoked at least a pack a day while pregnant.

The statistical model. Suppose that we intend to take independent samples of sizes $n_1, \ldots, n_k$ from k populations. Let Y_{ij} denote the random conceptualization of the measurement for individual j in sample i. Suppose that (Equation 12.1)

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

where (Equation 12.2) $\mu + \alpha_i =: \mu_i$ is the mean for population i and the ϵ_{ij} are independent normal random variables with mean 0 and unknown variance σ^2. This prescription for ϵ_{ij} implies a belief that all populations are

[1] The alternative is not $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$ since, for instance, μ_1 and μ_2 could be equal but different from μ_3 and μ_4. The most concise way to express the alternative mathematically is $\sum_{i=1}^{k} \sum_{j=1}^{k} (\mu_i - \mu_j)^2 > 0$. However, using a verbal descriptor such as "complementary" seems easier.

normal and share a common variance σ^2.

Based on what I have said so far, μ and the α_i are not uniquely determined.[2] For instance, suppose that $k = 2$, $\mu + \alpha_1 = 3$, and $\mu + \alpha_2 = 4$. Then we could have either $\mu = 2$, $\alpha_1 = 1$, and $\alpha_2 = 2$ or $\mu = 3.5$, $\alpha_1 = -0.5$, and $\alpha_2 = 0.5$. Therefore, a common convention is to impose a restriction such as

$$\sum_{i=1}^{k} n_i \alpha_i = 0.$$

Then the parameters are uniquely determined and are interpretable: μ is an "overall" mean (more precisely, a weighted average of the population means in which the weights are proportional to the sample sizes), and α_i is the difference between the mean for population i and the overall mean. We view α_i as fixed (i.e., not random) but unknown.

Sums of squares. Once the data have been collected, we may compute the mean for each sample and the mean across all samples. Let these quantities be denoted $\bar{y}_1, \ldots, \bar{y}_k$ and $\bar{\bar{y}}$. Define a "Total Sum of Squares" (Definition 12.2)

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{\bar{y}})^2,$$

which may be interpreted as a gross measure of variability. Indeed, if we merged all of the samples into a single "meta-sample", then the Total Sum of Squares would be the numerator of the meta-sample variance.

The Total Sum of Squares can be decomposed (Equation 12.4) as

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{\bar{y}})^2.$$

[2] A statistician would say that these parameters are not "identifiable".

The first piece is called (Definition 12.3) the "Within Sum of Squares" (also "Error Sum of Squares") and measures dissimilarity within groups. The second piece is called (Definition 12.4) the "Between Sum of Squares" (also "Treatment Sum of Squares" or "Model Sum of Squares") and measures dissimilarity between groups.

The top panel in Figure 12.1 shows a data set for which the Within Sum of Squares is small relative to the Between Sum of Squares (and for which rejecting H_0 should be easy), while the bottom panel shows a data set for which the Within Sum of Squares is large relative to the Between Sum of Squares (and for which rejecting H_0 should be difficult).

Computational formulas for desk calculators. Put $n := n_1 + \cdots + n_k$, and let s_i^2 denote the i^{th} sample variance. The Between SS may be computed as $\sum_{i=1}^k n_i \bar{y}_i^2 - n\bar{\bar{y}}^2$, while the Within SS may be computed as $\sum_{i=1}^k (n_i - 1)s_i^2$ (Equation 12.5).

Mean squares and degrees of freedom. Let the "Between Mean Square" be defined (Definition 12.5) as Between SS$/(k-1)$, and let the "Within Mean Square" be defined (Definition 12.6) as Within SS$/(n-k)$.

The "degrees of freedom" $(k-1)$ and $(n-k)$ may be intuitively understood as follows. For Between SS, we are interested in the fluctuations of the k sample means around μ. However, since we don't know μ, we replace it by the estimate $\bar{\bar{y}}$ and "lose" one degree of freedom. For Within SS, we are interested in the fluctuations of the n measurements around their corresponding population means. However, since we don't know μ_1 through μ_k, we replace them by the estimates $\bar{y}_1$ through $\bar{y}_k$ and "lose" k degrees of freedom.

If $k = 2$, then $n - k = n_1 + n_2 - 2$ and the Within MS is precisely the pooled variance estimate in Equation 8.10. Thus, the Within MS is a multi-sample generalization of the pooled variance estimate. In particular, the Within MS is an estimate of σ^2.

The one-way analysis of variance. Once the Between MS and the Within MS have been obtained, we calculate (Equation 12.6)

$$f := \text{Between MS} \;/\; \text{Within MS}.$$

We reject H_0 if $f > f_{k-1,n-k,1-\alpha}$ (or, equivalently, if the computer-supplied p-value is less than α). The rejection criterion is based on the principle that a large Between SS is indicative of differing population means and the fact that the random conceptualization of f follows the F distribution on $(k-1), (n-k)$ degrees of freedom when H_0 is true.

This procedure is called the one-way analysis of variance because we study populations characterized by *one* factor and *analyze* how a gross measure of *variability* decomposes into pieces reflecting dissimilarities between and within groups.

Example (the one-way analysis of variance). Refer to page 2 of {ANOVAExamples.pdf}, which provides results for the "Obstetrics" example mentioned previously. The first box shows that Total SS = 31.9763 and Between SS = 11.6727. Noting that $n = 27$ and $k = 4$, you could now fill in the rest of the table yourself (except for the p-value) using a desk calculator. We have Within SS = 31.9763 − 11.6727 = 20.3036, Between MS = 11.6727/3 = 3.8909, Within MS = 20.3036/23 = 0.8828, and $f = 3.8909/0.8828 = 4.41$. The p-value is 0.0137, so we can reject $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ at a significance level of $\alpha = 0.05$. The critical value for a level $\alpha = 0.05$ test is $f_{3,23,0.95} = 3.028$. This critical value is unavailable from Table 9 but can be approximated by interpolating between $f_{3,20,0.95} = 3.10$ and $f_{3,30,0.95} = 2.92$.

The second box on page 2 of {ANOVAExamples.pdf} provides the "Root MSE" and the "BWT Mean". The former is the square root of the Within MS (hence, estimates σ), while the latter is $\bar{\bar{y}}$.

Follow-up tests and multiple comparisons

Introduction. If we reject the "omnibus null hypothesis" $H_0 : \mu_1 = \cdots = \mu_k$, we may want to assess which population means differ. To this end, we can perform follow-up tests (also "post-hoc tests") concerning pairs of means. That is, we can test $H_0 : \mu_i = \mu_j$ against $H_1 : \mu_i \neq \mu_j$ for some or all pairs (i, j) with $i \neq j$.

At first glance, how to do this seems obvious. We can employ Equation 12.12, which is identical to Equation 8.11 except that Within MS is used instead of s^2 and the T distribution on $(n - k)$ degrees of freedom is used instead of the T distribution on $(n_1 + n_2 - 2)$ degrees of freedom. However, Equation 12.12 ignores the inflation of overall Type I error probability.

Inflation of overall Type I error probability. Suppose that $k = 3$ and that we want to perform all three follow-up tests concerning pairs of means. If each follow-up test has a Type I error probability of α, then the overall Type I error probability may be considerably larger than α. Indeed, we can get a simple upper bound for the overall Type I error probability as follows.

Let A_1, A_2, A_3 denote the events that the first, second, and third follow-up null hypotheses are rejected. If all three null hypotheses are true, then the probability that we incorrectly reject at least one of them is

$$P(A_1 \cup A_2 \cup A_3) \leq P(A_1 \cup A_2) + P(A_3) \leq P(A_1) + P(A_2) + P(A_3) = 3\alpha.$$

Above I have twice used the fact from Lecture 2 that

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B)$$

for any two events A and B. Thus, if the Type I error probability for each follow-up test is 0.05, all I can promise you is that the overall Type I error probability is no more than 0.15.

More generally, if there are m follow-up tests and each one has a Type I error probability of 0.05, all I can promise you is that the overall Type I error probability is no more than the smaller of $0.05m$ and 1. Hence, if you perform a lot of follow-up tests, each at level 0.05, then you are almost guaranteed to obtain a statistically significant result, even if all of the follow-up null hypotheses are true!

Bonferroni adjustment. If we want the overall Type I error probability across m follow-up tests to be no more than α, we can set the Type I error probability for each follow-up test to α/m. This is referred to as a Bonferroni adjustment for multiple comparisons. In particular, if we wish to test $H_0 : \mu_i = \mu_j$ against $H_1 : \mu_i \neq \mu_j$ for all pairs (i, j), then we have $m = k(k-1)/2$ and can set the Type I error probability for each follow-up test to $\alpha/(k(k-1)/2)$. Rosner describes a Bonferroni adjustment in Equation 12.14 with $m = k(k-1)/2$.

The decision whether to make a Bonferroni adjustment is controversial. First, such an adjustment is conservative in that the overall Type I error probability with a Bonferroni adjustment typically turns out to be much less than α, implying that the power to reject false follow-up null hypotheses is unnecessarily low.[3] Second, if we promise not to perform the follow-up tests unless the omnibus null hypothesis is rejected, then the overall Type I error probability is automatically reduced to a number no greater than the Type I error probability of the omnibus test. In particular, if the probability that we perform the follow-up tests is only 0.05, then the probability of incorrectly rejecting a follow-up null hypothesis cannot be any greater than 0.05. Hence, the practical relevance of a Bonferroni adjustment in this setting is questionable.[4]

Your textbook author gives sound advice: If k is not too large and the follow-up tests are specifically planned before the data are collected (i.e., if each follow-up test is of scientific interest in its own right), then a Bonferroni adjustment is unnecessary. If k is rather large or the follow-up tests are not planned before the data are in hand, then using a Bonferroni adjustment is safer (i.e., renders your work less susceptible to criticism by those who like the idea of making adjustments for multiple comparisons).

[3]Let $A_1, \ldots, A_m$ denote the events that the 1^{st} through m^{th} follow-up null hypotheses are rejected at the common individual significance level of α/m. The only way that the overall Type I error probability can equal α exactly is if $A_1, \ldots, A_m$ are mutually exclusive. But such mutual exclusivity is not typical. If $A_1, \ldots, A_m$ are independent and m is large, then the relation $(1 - \alpha/m)^m \approx \exp[-\alpha]$ implies that the overall Type I error probability is approximately $1 - \exp[-\alpha]$, which is only slightly less than α. But even independence is not typical. If $A_1, \ldots, A_m$ are positively associated, in that they tend to occur together, then the overall Type I error probability is much less than α. Unfortunately for us, this positive association is typical.

[4]A colleague of mine believes that there is a flaw in this reasoning. He believes that the omnibus null hypothesis is never true in practice, so that the probability of performing the follow-up tests is really much greater than 0.05 and often close to 1. In particular, while a false omnibus null hypothesis suggests that some follow-up null hypotheses must be false, we can have a false omnibus null hypothesis even if the vast majority of follow-up null hypotheses are true. For instance, suppose that $\mu_1 = \mu_2 = \cdots = \mu_{k-1} \neq \mu_k$. Then the omnibus null hypothesis is false, but there are $(k-1)(k-2)/2$ true follow-up null hypotheses if all pairs of means are to be compared. My colleague believes that a Bonferroni adjustment is relevant for preventing excessively many Type I errors among these $(k-1)(k-2)/2$ follow-up tests.

Example (Bonferroni adjustment). Refer to pages 3 and 4 of {ANOVAExamples.pdf}. We have $\bar{y}_1 = 7.5857$, $\bar{y}_2 = 7.2400$, $\bar{y}_3 = 6.3286$, and $\bar{y}_4 = 6.0125$. For the test of $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ we have

$$t = \frac{7.5857 - 7.2400}{\sqrt{0.8828(1/7 + 1/5)}} = 0.628.$$

If no adjustment is made, we compare $|t|$ to $t_{23,0.975} = 2.069$ and accept H_0. The p-value is 0.5359. If we make the Bonferroni adjustment with $m = k(k-1)/2 = 6$, then we compare $|t|$ to $t_{23,0.99583} = 2.886$ and accept H_0. Note that the 0.99583 is obtained as $1 - \alpha/(2m) = 1 - 0.05/12$ and that we can (crudely) approximate $t_{23,0.99583}$ from Table 5 as $t_{23,0.995} = 2.807$. A p-value reflecting the Bonferroni adjustment is 1.0000. In general, the formula

$$p_{adj} = \min\{p_{unadj} m, 1\}$$

can be used to convert unadjusted p-values to adjusted p-values.

Linear contrasts. Notice that $\mu_1 - \mu_2$ can be written in the form $\sum_{i=1}^{k} c_i \mu_i$ with $c_1 = 1, c_2 = -1, c_3 = 0, \ldots, c_k = 0$. In general, an expression of the form $\sum_{i=1}^{k} c_i \mu_i$ for which $\sum_{i=1}^{k} c_i = 0$ is referred to as a linear contrast.

Example (linear contrasts). Suppose that, among mothers who did not smoke during pregnancy, 80% never smoked (i.e., 20% gave up smoking during pregnancy). Suppose also that, among mothers who did smoke during pregnancy, 40% smoked less than a pack a day (i.e., 60% smoked at least a pack a day).

If we had to guess, we might say that the mean birthweight of infants born to mothers who did not smoke during pregnancy should be $0.80\mu_1 + 0.20\mu_2$. In fact, this is correct. Let X denote the birthweight of a randomly selected infant born to a mother who did not smoke during pregnancy, and let A be the event that the mother never smoked. In analogy to the rule of total probability

$$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$$

from Lecture 2, there is a rule of total expectation

$$E[X] = E[X|A]P(A) + E[X|\bar{A}]P(\bar{A}).$$

Thus, we obtain $E[X] = \mu_1\, 0.80 + \mu_2\, 0.20$. Similarly, the mean birthweight of infants born to mothers who did smoke during pregnancy is $0.40\mu_3 + 0.60\mu_4$.

The linear contrast $c_1\mu_1 + c_2\mu_2 + c_3\mu_3 + c_4\mu_4$ with $c_1 = 0.80$, $c_2 = 0.20$, $c_3 = -0.40$, and $c_4 = -0.60$ represents the difference between the mean birthweight of infants born to mothers who did not smoke during pregnancy and the mean birthweight of infants born to mothers who did smoke during pregnancy.

Follow-up test involving a linear contrast. To test $H_0 : \sum_{i=1}^k c_i\mu_i = 0$ against $H_1 : \sum_{i=1}^k c_i\mu_i \neq 0$, where $\sum_{i=1}^k c_i = 0$, we compute

$$t := \frac{\sum_{i=1}^k c_i\bar{y}_i}{\sqrt{\text{Within MS } \sum_{i=1}^k (c_i^2/n_i)}}.$$

If we do not wish to adjust for multiple comparisons, we simply compare $|t|$ to $t_{n-k,1-\alpha/2}$ (Equation 12.13). If we want to adjust for a specific number of multiple comparisons, we can use the Bonferroni approach. However, since there are infinitely many linear contrasts for which we can perform follow-up tests, a Scheffé adjustment is more commonly employed. A Scheffé adjustment justifies an unlimited number of follow-up tests and entails (Equation 12.16) comparing $|t|$ to $\sqrt{(k-1)f_{k-1,n-k,1-\alpha}}$.

Example (follow-up test involving a linear contrast). Refer to page 2 of {ANOVAExamples.pdf}. Let us test

$$H_0 : 0.80\mu_1 + 0.20\mu_2 - 0.40\mu_3 - 0.60\mu_4 = 0$$

against

$$H_1 : 0.80\mu_1 + 0.20\mu_2 - 0.40\mu_3 - 0.60\mu_4 \neq 0.$$

We see that $\sum_{i=1}^{k} c_i \bar{y}_i = 1.378$ and $\sqrt{\text{Within MS } \sum_{i=1}^{k}(c_i^2/n_i)} = 0.384$, whence $t = 3.58$. If we make no adjustment for multiple comparisons, then $|t|$ is compared to $t_{23,0.975} = 2.069$. The p-value turns out to be 0.0016.

If we make the Scheffé adjustment, then $|t|$ is compared to $\sqrt{3f_{3,23,0.95}} = \sqrt{9.084} = 3.014$. The p-value turns out to be 0.0153. Unfortunately, in this setting there is no simple formula for converting an unadjusted p-value to an adjusted p-value.

In any case, we reject H_0.