

**Assignment #14 — Due Wednesday, May 6, 2009, by 5:00 P.M.**

Turn in homework in lecture, discussion, or your TA's mailbox. Indicate the discussion section in which you expect to attend to pick up this assignment on the assignment.

**311:** Monday 1:20–2:10

**312:** Monday 12:05–12:55

This assignment is about ANOVA.

1. (*ANOVA Theory*) The model that underlies ANOVA is the following:  $Y_{ij} \sim N(\beta_i, \sigma^2)$  where all of the random variables are mutually independent,  $i \in \{1, \dots, a\}$  is the index of the group,  $j \in \{1, \dots, n_i\}$  is the index of the observation within group  $i$ ,  $n_i$  is the sample size of group  $i$ ,  $n = n_1 + \dots + n_a$  is the total number of observations,  $\beta_i$  is the population mean of group  $i$ , and  $\sigma^2$  is the common variance for each individual random observation.

The data  $\{y_{ij}\}$  is summarized within each sample and overall. The individual sample means are  $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ . The grand mean of all observations is  $\bar{y} = \sum_{i=1}^a \sum_{j=1}^{n_i} y_{ij}/n = \sum_{i=1}^a (n_i/n)\bar{y}_i$  so the grand mean is the weighted average of the sample means, weighted by their sample sizes. The individual sample variances are  $s_i^2 = \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2/(n_i - 1)$ . The grand sum of squared differences between observations and sample means divided by  $n - a$  is

$$\frac{\sum_{i=1}^a \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{n - a} = \sum_{i=1}^a \left( \frac{n_i - 1}{n - a} \right) s_i^2$$

which is a weighted average of the sample standard deviations, weighted by the degrees of freedom  $(n_i - 1)$  within each sample.

Note that with this setup,  $\bar{y}_i \sim N(\beta_i, \sigma^2/n_i)$  for each  $i$ .

ANOVA depends on examining the distribution of two independent estimates of  $\sigma^2$ , one based on variation among the sample means, the other based on variation within the samples. This homework problem asks you verify that each of these two independent estimates is, in fact, an unbiased estimator of  $\sigma^2$ . The algebra gets a little hairy, so I provide some steps for each derivation. At some point in the derivation, you need to expand a product of the form

$$\left( \sum_{i=1}^n a_i \right)^2 = \left( \sum_{i=1}^n a_i \right) \left( \sum_{j=1}^n a_j \right) = \sum_{i=1}^n \sum_{j=1}^n a_i a_j = \sum_{i=1}^n a_i^2 + 2 \sum_{i < j} a_i a_j$$

This is straightforward algebra, but may not be familiar to you.

- (a) Show that  $E(\sum_{i=1}^a \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2) = (n - a)\sigma^2$ .
  - i. Show that  $y_{ij} - \bar{y}_i = (y_{ij} - \beta_i) - \frac{1}{n} \sum_{k=1}^{n_i} (y_{ik} - \beta_i)$ .
  - ii. Show that  $E((y_{ij} - \beta_i)(y_{ik} - \beta_i))$  is 0 when  $j \neq k$  and  $\sigma^2$  when  $j = k$ .
  - iii. Evaluate  $E(\sum_{i=1}^a \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2)$  by making the substitution to represent the data in terms of  $y_{ij} - \beta_i$ , square the expression (keeping  $y_{ij} - \beta_i$  together!), exchange the order of summation and expectation, and simplify.
- (b) Show that  $E(\sum_{i=1}^a n_i (\bar{y}_i - \bar{y})^2) = (a - 1)\sigma^2 + \sum_{i=1}^a n_i (\beta_i - \beta)^2$  where  $\beta = \sum_{i=1}^a (n_i/n)\beta_i$  is the weighted mean of the means from each normal distribution.
  - i. Show that  $\bar{y}_i - \bar{y} = (\bar{y}_i - \beta_i) + (\beta_i - \beta) - \frac{1}{n} \sum_{j=1}^a n_j (\bar{y}_j - \beta_j)$ .

- ii. Make the above replacement and square out the trinomial expression  $(A + B - C)^2 = A^2 + B^2 + C^2 + 2AB - 2AC - 2BC$ .
- iii. Multiply by  $n_i$  and take expectations, recalling that  $E(\bar{y}_i - \beta_i) = 0$ ,  $E(\bar{y}_i - \beta_i)^2 = \sigma^2/n_i$ , and  $E(y_i - \beta_i)(y_j - \beta_j) = 0$  when  $i \neq j$  since  $\bar{y}_i$  and  $\bar{y}_j$  are then independent. This shows that  $E(n_i(\bar{y}_i - \bar{y})^2) = n_i(\beta_i - \beta)^2 + \sigma^2(1 - n_i/n)$ .
- iv. Take the final sum, and simplify to complete the problem.
- (c) Conclude  $(\sum_{i=1}^a \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2)/(n-a)$  is an unbiased estimator of  $\sigma^2$  for any  $\{\beta_i\}$ , but that  $\sum_{i=1}^a n_i(\bar{y}_i - \bar{y})^2$  is an unbiased estimator of  $\sigma^2$  only if the  $\{\beta_i\}$  are all equal.
2. (*ANOVA Practice*) Enter the data from Exercise 10.4.4 into R.
- (a) Use `xyplot()` to make side-by-side dot plots of the data.
- (b) Compute means and standard deviations of each group.
- (c) Use `lm()` to fit the linear model and summarize with `summary()`. Make a connection between the estimated parameter values and the sample statistics.
- (d) Use `anova()` to show the corresponding ANOVA table. Summarize the results of this hypothesis test.

**Work to do, but not turn in.**

- Study for the final exams!
-