M4056 Analysis of Variance, I **November 22-24, 2010**

1. Introduction and Goal

Let X be a normal random variable with mean μ_X and variance σ^2. Let Y be another normal random variable with mean μ_Y and the same variance σ^2 as X. In the lectures of November 17 and 19, we examined how to test the hypothesis $H_0 : \mu_X = \mu_Y$ using the evidence obtained from a sample $(X_1, \ldots, X_n)$ from the X distribution and a sample $(Y_1, \ldots, Y_m)$ form the Y distribution. The key technical result that makes this possible is the fact that

$$t = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{S_p\sqrt{1/n + 1/m}}$$

is a t distribution with $m + n - 2$ degrees of freedom.

Now suppose that we have several normal random variables $Y_1, \ldots, Y_m$. We shall assume they all have the same variance σ^2. The means may be different. Let μ_i be the mean of Y_i, $i \in \{1, \ldots, m\}$. Let μ be the average of the μ_i. In summary,

$$Y_i \sim \text{normal}(\mu_i, \sigma^2) \text{ for } i = 1, \ldots m$$

$$\mu = \frac{1}{m}\sum_{i=1}^{m} \mu_i.$$

From each distribution, we take a sample $(Y_{i1}, \ldots, Y_{in})$. Thus, we have an $m \times n$ matrix of independent random variables:

$$\begin{matrix} Y_{11} & Y_{12} & \cdots & Y_{1n} \\ Y_{11} & Y_{12} & \cdots & Y_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ Y_{m1} & Y_{m2} & \cdots & Y_{mn} \end{matrix}$$

Here, the i^{th} row is i.i.d. Y_i. Our goal is to devise a test for the null hypothesis:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_m.$$

For example, suppose m different treatments were applied to m different groups. To determine if there is any evidence that any of the treatments are effective, this is the hypothesis we would test. A significant violation of the null hypothesis would count as evidence of that at least one group was affected differently than the others. Note that in the scenario we are imagining, we have m samples, each of size n. A more general situation arises if the samples have different sizes, but we will delay consideration of this till later.

2. Some Notation

We introduce the following abbreviations:

$$\overline{Y}_{i\cdot} := \frac{1}{n}\sum_{j=1}^{n} Y_{ij} \quad \text{(an estimate of } \mu_i\text{)},$$

$$\overline{Y}_{\cdot\cdot} := \frac{1}{m\,n}\sum_{i=1}^{m}\sum_{j=1}^{n} Y_{ij} \quad \text{(an estimate of } \mu\text{)}.$$

The statistic we will use for testing H_0 is based on the following:

Fact.

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_{i\cdot})^2 + n\sum_{i=1}^{m}(\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot})^2.$$

Comment. This is saying that the sum of the squares of the deviations of all the observations from the grand mean (SS_{TOT}) is equal to the sum of the squares of the deviations of the observations *within* each group from the group mean (SS_W) *plus* the sum of the squares of the deviations of the group means from the grand mean (SS_B—“B” stands for *between* groups):

$$SS_{TOT} = SS_W + SS_B.$$

Proof. First, observe that:

$$\begin{aligned}\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_{\cdot\cdot})^2 &= \sum_{i=1}^{m}\sum_{j=1}^{n}[(Y_{ij}-\overline{Y}_{i\cdot})+(\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot})]^2 \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_{i\cdot})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n}(\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot})^2 \qquad (*) \\ &\quad + 2\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_{i\cdot})(Y_{i\cdot}-\overline{Y}_{\cdot\cdot}).\end{aligned}$$

Now

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_{i\cdot})(\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot}) = \sum_{i=1}^{m}(\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot})\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_{i\cdot}),$$

but for each i,

$$\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_{i\cdot}) = 0,$$

since the sum of the deviations from the mean is zero. Thus, the term with coefficient 2 in $(*)$ is zero. /////

Lemma. *Let X_j, $j = 1,\ldots,n$, be independent random variables with* $\mathrm{E}(X_j) = \mu_j$ *and* $\mathrm{Var}(X_j) = \sigma^2$. *Let $\mu = \frac{1}{n}\sum_{j=1}^{n}\mu_j$. Then:*

$$\mathrm{E}(X_j-\overline{X})^2 = (\mu_j-\mu)^2 + \frac{n-1}{n}\sigma^2.$$

Proof. First, observe that $\mathrm{E}(X_j-\overline{X})^2 = \mathrm{E}(X_j^2) - 2\mathrm{E}(X_j\overline{X}) + \mathrm{E}(\overline{X}^2)$. Calculate each term on the right:

a) $\mathrm{E}(X_j^2) = (\mathrm{E}X_j)^2 + \mathrm{Var}X_j = \mu_j^2 + \sigma^2$.

b) $\mathrm{E}X_j\overline{X} = \frac{1}{n}\sum_{k=1}^n \mathrm{E}(X_jX_k) = \frac{1}{n}\left(\sum_{k=1}^n \mu_j\mu_k + \sigma^2\right) = \mu_j\mu + \frac{1}{n}\sigma^2$. The second equality here comes about because $\mathrm{E}(X_jX_k) = \mu_j\mu_k$ if $j \neq k$ (because X_j and X_k are independent), while $\mathrm{E}(X_jX_j) = \mu_j^2 + \sigma^2$ as in a).

c) $\mathrm{E}(\overline{X}^2) = (\mathrm{E}\overline{X})^2 + \mathrm{Var}\overline{X} = \mu^2 + \frac{1}{n}\sigma^2$.

Now add them up: $\mu_j^2 + \sigma^2 - 2(\mu_j\mu + \frac{\sigma^2}{n}) + \mu^2 + \frac{\sigma^2}{n} = (\mu_j - \mu)^2 + \sigma^2 - \frac{\sigma^2}{n}$. /////

3. The Expected Values of SS_W and SS_B

Let us apply the lemma with $X_j = Y_{ij}$, the variables defined in the introduction. (We treat i as fixed throughout the discussion, but what we say applies to any i.) Since all the variables have the same expected value, $\mathrm{E}Y_{ij} = \mu_i$, $j = 1, \ldots, n$, we get:

$$\mathrm{E}(SS_W) = \sum_{i=1}^m\sum_{j=1}^n \mathrm{E}(Y_{ij} - \overline{Y}_{i\cdot})^2 = \sum_{i=1}^m\sum_{j=1}^n \frac{n-1}{n}\sigma^2 = m(n-1)\sigma^2. \qquad (\mathcal{W})$$

(The second equality uses the fact that $\mu_{ij} = \mu_i$ for all j, so the difference $\mu_{ij} - \overline{\mu_i}$ vanishes.) This shows among other things that $SS_W/m(n-1)$ is an unbiased estimator for σ^2.

Let us apply the lemma, with i in place of j, and $X_i = \overline{Y}_{i\cdot}$. (The μ_i are the numbers in the introduction: $\mu_i = \mathrm{E}(\overline{Y}_{i\cdot})$. Recall that we are using the symbol μ to stand for the average of the μ_i.) We get

$$\begin{aligned}\mathrm{E}(SS_B) &= n\sum_{i=1}^m \mathrm{E}(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 \\ &= n\sum_{i=1}^m\left[(\mu_i - \mu)^2 + \frac{m-1}{m}\frac{\sigma^2}{n}\right] \qquad (\mathcal{B}) \\ &= (m-1)\sigma^2 + n\sum_{i=1}^m(\mu_i - \mu)^2.\end{aligned}$$

This shows that SS_B is sensitive to differences in the group means, since if they are not all the same, then $\sum_{i=1}^m(\mu_i - \mu)^2 > 0$.

4. The Distribitions of SS_W and SS_B

Now recall that if $X_1, \ldots, X_n$ are i.i.d. normal(μ, σ^2), then

$$\frac{1}{\sigma^2}\sum_{i=1}^n(X_i - \overline{X})^2 = (n-1)S^2/\sigma^2 \sim \chi^2_{n-1}. \qquad (**)$$

For each $i = 1, \ldots, m$, $(**)$ applies to $Y_{i1}, \ldots, Y_{in}$, showing that

$$\frac{1}{\sigma^2} \sum_{j=1}^{n} (Y_{ij} - \overline{Y}_{i\cdot})^2 \sim \chi^2_{n-1}.$$

Since these sums are independent for different i,

$$SS_W/\sigma^2 = \frac{1}{\sigma^2} \sum_{i=1}^{m} \sum_{j=1}^{n} (Y_{ij} - \overline{Y}_{i\cdot})^2 \sim \chi^2_{m(n-1)}.$$

Note that this is true regardless of whether or not the null hypothesis $H_0 : \mu_1 = \cdots = \mu_m$ is true.

Equation $(**)$ also applies to $\sum_{i=1}^{m} (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2$, but only in case the null hypothesis is true. In this case, the $\overline{Y}_{i\cdot}$ are independent and identically distributed normal$(\mu, \sigma^2/n)$ variables. Thus,

$$SS_B/\sigma^2 = \frac{n}{\sigma^2} \sum_{i=1}^{m} (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 \sim \chi^2_{m-1}.$$

Finally, we consider the independence of SS_W and SS_B. We showed earlier in the course that if $X_1, \ldots, X_n$ are i.i.d. normal random variables, then $\overline{X}$ and the vector $(X_1 - \overline{X}, \ldots, X_n - \overline{X})$ are independent of one another. Thus, for each $i = 1, \ldots, m$, $\overline{Y}_{i\cdot}$ and the vector $(Y_{i1} - \overline{Y}_{i\cdot}, \ldots, Y_{in} - \overline{Y}_{i\cdot})$ are independent of one another. But SS_B is a function of the $\overline{Y}_{i\cdot}$, $i = 1, \ldots, m$, while SS_W is a function of the $Y_{ij} - \overline{Y}_{i\cdot}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$. (This independence result does not depend on the null hypothesis; it is true regardless of whether the null hypothesis holds.)

5. The Test Statistic

Theorem. *Under the null hypothesis, the statistic*

$$F := \frac{SS_B/(m-1)}{SS_W/m(n-1)}$$

has an F distribution with $m-1$ and $m(n-1)$ degrees of freedom.

Proof. This follows from the definition of the F distribution. /////

Under the null hypothesis, the expected value of this statistic is 1. If the null hypothesis is false, the expected value is larger than 1. A test of level α has rejection region $F \in (r, \infty)$, where r is chosen so that $P(F > r) = \alpha$. The required value of r can be determined from a table of the F distribution. (Or, see the Mathematica Notebook, FRatioDistribution.)