CPH 931 — Fall 2008 — Dr. Charnigo

Lecture 3

Repeated measures analysis of variance

Introduction. In your first statistics course you learned about the independent samples T test for equality of two means, which entailed computing (with obvious notation)

$$t_{independent} := \frac{\bar{x} - \bar{y}}{\sqrt{s^2_{pooled}/n_x + s^2_{pooled}/n_y}}$$

and comparing its absolute value to $t_{n_x+n_y-2,1-\alpha/2}$.

You also learned about the paired samples T test for equality of two means, which entailed computing

$$t_{paired} := \frac{\bar{x} - \bar{y}}{\sqrt{s^2_{x-y}/n_{x-y}}}$$

and comparing its absolute value to $t_{n_{x-y}-1,1-\alpha/2}$.

As the name suggests, the paired samples T test is potentially appropriate — we must be mindful of the normality assumption — when each observation in the first sample corresponds to one and only one observation in the second sample, either because we have "matched" subjects in the two samples or (more commonly) because the subjects in the two samples are actually the same people.

On the other hand, the independent samples T test is potentially appropriate when there are no such correspondences between observations in the first sample and observations in the second sample.

What if more than two means are to be compared? If we have independent samples, then a potentially appropriate method is the one-way analysis of variance (ANOVA) that you also learned about in your first statistics course. This entailed computing

$$f_{independent} := \frac{\text{Between Mean Square}}{\text{Within Mean Square}}$$

and comparing it to $f_{df_{between}, df_{within}, 1-\alpha}$.

Yet, a typical first statistics course does not say what to do if we wish to compare more than two means and the samples are not independent, as would be case if three or more observations were made on each subject. This problem is not covered in CPH 930 either. Therefore, we will address it here.

What does this problem have to do with ordinary least squares and linear regression? Plenty! However, we will wait until Lecture 4 to make the connection explicit.

Scenario and Terminology. There are k experimental conditions or "treatments". Each of n subjects is observed once on each treatment. The observation for subject i on treatment j is denoted Y_{ij}. We assume that

$$Y_{ij} = \mu_j + \alpha_i + \epsilon_{ij}. \tag{1}$$

The quantities $\mu_1, \ldots, \mu_k$ are called "fixed effects" and represent the unknown treatment means. The quantities $\alpha_1, \ldots, \alpha_n$ are called "random effects" and represent individual subjects' tendencies to score higher or lower than average. The quantities $\epsilon_{11}, \epsilon_{12}, \ldots, \epsilon_{nk}$ are called "error terms" and represent differences between the actual observations and the treatment means adjusted for individual subjects' tendencies to score higher or lower than average.

The random effects are assumed to be independent normal random variables with mean 0 and common but unknown variance σ_α^2. The error terms are assumed to be independent normal random variables with mean 0 and common but unknown variance σ_ϵ^2. The random effects and the error terms are assumed to be independent.

If $\sigma_\alpha^2 = 0$, then $\alpha_1 = \cdots = \alpha_n = 0$ and all of the observations are independent. In this case the data can be analyzed using the one-way ANOVA from your first statistics course. However, there is usually little reason to believe that $\sigma_\alpha^2 = 0$.

If $\sigma_\alpha^2 > 0$, then repeated observations on the same subject are correlated. This "intra-class correlation" is given by $\sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_\epsilon^2)$ and, unlike an ordinary Pearson correlation, is necessarily a positive number. The one-way ANOVA from your first statistics course is inappropriate in this case because it does not take into account that part of the within-treatment variability is due to the random effects.

Motivating Example and Further Terminology. The digit symbol substitution test (DSST) is used by psychologists to assess attention and concentration in relation to, for instance, substance abuse. The data in {DSST.xls} pertain to 16 subjects who were asked to perform the DSST under four experimental conditions: on placebo following a low-sensation activity, on placebo following a high-sensation activity, on d-amphetamine (10 mg/70 kg) following a low-sensation activity, and on d-amphetamine (10 mg/70 kg) following a high-sensation activity. Actually, the subjects completed the DSST four times under each experimental condition, and the whole experiment was replicated. For now, however, we confine attention to the third time the DSST was completed ("TIME" $=3$) in the first run of the experiment ("REPLICAT" $=0$).

The response variable is "CORRECTTRIALS" and represents the number of times that the DSST was performed correctly in 90 seconds. We refer to "CONDITION", the variable identifying the experimental conditions, as the explanatory variable. We refer to "SUBJECT", the variable identifying the subjects, as the blocking variable.

The term "blocking variable" arises from agricultural experiments on crop yield that used model (1), with the "treatments" being different fertilizers and the "subjects" being different farms or blocks of land. Each fertilizer was applied to a small portion of the land within each block. This approach was considered preferable to applying a single fertilizer within a block because, in that case, a block that happened to have better soil would create a misleadingly favorable impression about the fertilizer applied to it.

Notation and Sums of Squares. Consider model (1). For $j = 1, \ldots, k$ let

$$\bar{Y}_{.j} := n^{-1} \sum_{i=1}^{n} Y_{ij}$$

be the average observation on treatment j over all subjects. For $i = 1, \ldots, n$ let

$$\bar{Y}_{i.} := k^{-1} \sum_{j=1}^{k} Y_{ij}$$

be the average observation on subject i over all treatments. Let

$$\bar{Y}_{..} := (nk)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k} Y_{ij}$$

be the average observation over all subjects and treatments.

We refer to

$$SST := \sum_{i=1}^{n} \sum_{j=1}^{k} (Y_{ij} - \bar{Y}_{..})^2$$

as the Total Sum of Squares, to

$$SSA := \sum_{i=1}^{n} \sum_{j=1}^{k} (\bar{Y}_{.j} - \bar{Y}_{..})^2 = n \sum_{j=1}^{k} (\bar{Y}_{.j} - \bar{Y}_{..})^2$$

as the Treatment Sum of Squares, to

$$SSB := \sum_{i=1}^{n}\sum_{j=1}^{k}(\bar{Y}_{i.} - \bar{Y}_{..})^2 = k\sum_{i=1}^{n}(\bar{Y}_{i.} - \bar{Y}_{..})^2$$

as the Block Sum of Squares, and to

$$SSE := \sum_{i=1}^{n}\sum_{j=1}^{k}(Y_{ij} - \bar{Y}_{.j} - \bar{Y}_{i.} + \bar{Y}_{..})^2$$

as the Error Sum of Squares. One can show that

$$SST = SSA + SSB + SSE.$$

Sums of Squares in the Motivating Example. Let Y_{ij} denote the number of times that subject i performed the DSST correctly in 90 seconds under experimental condition j. Again, we are confining attention to the third time the DSST was completed ("TIME" = 3) in the first run of the experiment ("REPLICAT" = 0). Straightforward if tedious computations show that

$$\bar{Y}_{.1} = 66.5,\ \bar{Y}_{.2} = 66.4375,\ \bar{Y}_{.3} = 69.75,\ \bar{Y}_{.4} = 71.1875,$$

$$\bar{Y}_{1.} = 74.5,\ \bar{Y}_{2.} = 66.25, \ldots, \bar{Y}_{15.} = 78.25,\ \bar{Y}_{16.} = 60.5,$$

$$\bar{Y}_{..} = 68.46875,$$

$$\begin{aligned} SST &= (76 - 68.46875)^2 + (70 - 68.46875)^2 + \cdots \\ &\quad + (60 - 68.46875)^2 + (64 - 68.46875)^2 \\ &= 4585.9375, \end{aligned}$$

$$\begin{aligned} SSA &= 16[(66.5 - 68.46875)^2 + (66.4375 - 68.46875)^2 \\ &\quad + (69.75 - 68.46875)^2 + (71.1875 - 68.46875)^2] \\ &= 272.5625, \end{aligned}$$

$$\begin{aligned} SSB &= 4[(74.5 - 68.46875)^2 + (66.25 - 68.46875)^2 + \cdots \\ &\quad +(78.25 - 68.46875)^2 + (60.5 - 68.46875)^2] \\ &= 3404.4375, \end{aligned}$$

and

$$SSE = SST - SSA - SSB = 4585.9375 - 272.5625 - 3404.4375 = 908.9375.$$

The Logic of Repeated Measures ANOVA. Consider the null hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$. If the null hypothesis is true, then there is a common treatment mean that we may call μ. In this case

$$\bar{Y}_{.j} = n^{-1} \sum_{i=1}^{n} (\mu + \alpha_i + \epsilon_{ij}) = \mu + n^{-1} \sum_{i=1}^{n} \alpha_i + n^{-1} \sum_{i=1}^{n} \epsilon_{ij}$$

and

$$\bar{Y}_{..} = (nk)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k} (\mu + \alpha_i + \epsilon_{ij}) = \mu + n^{-1} \sum_{i=1}^{n} \alpha_i + (nk)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k} \epsilon_{ij},$$

so that

$$\bar{Y}_{.j} - \bar{Y}_{..} = n^{-1} \sum_{i=1}^{n} \epsilon_{ij} - (nk)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k} \epsilon_{ij}$$

has mean zero. Hence,

$$SSA = n \sum_{j=1}^{k} \left(n^{-1} \sum_{i=1}^{n} \epsilon_{ij} - (nk)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k} \epsilon_{ij} \right)^2$$

is not anticipated to be large. In fact, probability theory shows that SSA/σ_ϵ^2 has a chi-square distribution on $(k-1)$ degrees of freedom, regardless of n.

On the other hand, if $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ is false, then

$$\bar{Y}_{.j} - \bar{Y}_{..} = \mu_j - \left(\frac{\mu_1 + \cdots + \mu_k}{k} \right) + n^{-1} \sum_{i=1}^{n} \epsilon_{ij} - (nk)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k} \epsilon_{ij},$$

which has mean $\mu_j - \left(\frac{\mu_1+\cdots+\mu_k}{k}\right)$. Consequently,

$$SSA = n\sum_{j=1}^{k}\left(\mu_j - \left(\frac{\mu_1+\cdots+\mu_k}{k}\right) + n^{-1}\sum_{i=1}^{n}\epsilon_{ij} - (nk)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{k}\epsilon_{ij}\right)^2$$

is anticipated to be large, and to become immense as n increases. Note that the statement about SSA/σ_ϵ^2 having a chi-square distribution on $(k-1)$ degrees of freedom does not apply when H_0 is false.

Thus, a test of $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ can be based on the size of SSA. If σ_ϵ^2 were known, we could simply calculate SSA/σ_ϵ^2 and reject H_0 if SSA/σ_ϵ^2 exceeded $\chi^2_{k-1,1-\alpha}$. However, since σ_ϵ^2 is not known, we must express SSA relative to another quantity. A natural candidate is SSE.

Probability theory shows that SSE/σ_ϵ^2 has a chi-square distribution on $(n-1)(k-1)$ degrees of freedom, regardless of whether H_0 is true. Moreover, SSE and SSA are independent. Thus, if H_0 is true,

$$\frac{SSA/[(k-1)\sigma_\epsilon^2]}{SSE/[(n-1)(k-1)\sigma_\epsilon^2]} = \frac{SSA/[(k-1)]}{SSE/[(n-1)(k-1)]}$$

has an F distribution on $(k-1)$ numerator and $(n-1)(k-1)$ denominator degrees of freedom. Therefore, we may define the mean squares

$$MSA := SSA/[(k-1)] \quad \text{and} \quad MSE := SSE/[(n-1)(k-1)]$$

and then reject H_0 if the ratio MSA/MSE exceeds $f_{k-1,(n-1)(k-1),1-\alpha}$.

Testing H_0 in this manner is referred to as performing a repeated measures ANOVA.

Repeated Measures ANOVA in the Motivating Example. With $k = 4$ and $n = 16$, we find that

$$MSA = 272.5625/3 = 90.854 \quad \text{and} \quad MSE = 908.9375/45 = 20.199.$$

The ratio of mean squares is 4.498, which exceeds $2.812 = f_{3,45,0.95}$. So, at significance level 0.05 we reject the null hypothesis that the mean DSST score is the same for all four experimental conditions.

Reading the SAS Output. All of the above computations — and some computations not described above — can be carried out in SAS using PROC MIXED. The SAS output for this example is shown in {DSST.rtf}.

At the bottom of page 1 we have SSA, SSB, and SSE in the "Sum of Squares" column, preceded by the corresponding degrees of freedom in the "DF" column and followed by the corresponding mean squares in the "Mean Square" column. The ratio of MSA to MSE is the first entry in the "F Value" column, and the first entry in the "Pr > F" column states that the corresponding p-value is 0.0076.

Page 2 of {DSST.rtf} lists $\bar{Y}_{.1}$, $\bar{Y}_{.2}$, $\bar{Y}_{.3}$, and $\bar{Y}_{.4}$ in the "Solution for Fixed Effects" box. These are estimates of μ_1, μ_2, μ_3, and μ_4 respectively. Page 2 also displays estimates for σ^2_α and σ^2_ϵ in the "Covariance Parameter Estimates" box.

Moreover, page 2 shows the results of post-hoc tests for equality of μ_1 to μ_2, μ_1 to μ_3, μ_1 to μ_4, μ_2 to μ_3, μ_2 to μ_4, and μ_3 to μ_4. The p-values are not adjusted for multiple comparisons.

If you wish to adjust the p-values for multiple comparisons, one approach (Bonferroni) is to simply multiply the p-values by the number of comparisons and then truncate the results at 1. For instance, the adjusted p-value for testing equality of μ_1 to μ_2 would be 1, while the adjusted p-value for testing equality of μ_1 to μ_3 would be 0.2802. Whether to adjust p-values for multiple comparisons in this setting is controversial; personally I do not consider it necessary, but the ill-tempered reviewer at the journal to which you submit your manuscript may disagree!

A Comparison to One-Way ANOVA. For further insight into the merit of using repeated measures ANOVA with model (1), let us analyze the DSST data using the one-way ANOVA from your first statistics course.

The Between Sum of Squares is the same as SSA, 272.5625, and has the same degrees of freedom, 3.

The Within Sum of Squares equals $SSB + SSE = 3404.4375 + 908.9375 = 4313.375$ and has 60 degrees of freedom.

Thus, we find that

$$f_{independent} = \frac{\text{Between Mean Square}}{\text{Within Mean Square}} = \frac{90.854}{71.890} = 1.264$$

and fail to reject H_0 because 1.264 is less than $2.758 = f_{3,60,0.95}$.

What happened? Because SSB is enormous compared to SSE, the Within Mean Square is much larger than MSE, even though the Within Sum of Squares has more degrees of freedom than SSE. Consequently, the Between Mean Square $= MSA$ is compared to an unnecessarily large quantity and does not exceed that quantity by a large enough margin to permit rejection of H_0.

Discussion questions

1. Given data, how might you assess whether the random effects in model (1) were normally distributed? What about the error terms?

2. Apart from uncertainty about whether the random effects and error terms in model (1) were normally distributed, what other reservations might you have about applying model (1) to analyze the DSST data in the manner that we did?