Cuckoo Birds

- Cuckoo birds have a behavior in which they lay their eggs in other birds' nests.
- The other birds then raise and care for the newly hatched cuckoos.
- Cuckoos return year after year to the same territory and lay their eggs in the nests of a particular host species.
- Furthermore, cuckoos appear to mate only within their territory.
- Therefore, geographical sub-species are developed, each with a dominant foster-parent species.
- A general question is, are the eggs of the different sub-species distinct so that they are adapted to a particular foster-parent species?
- Specifically, we can ask, are the mean lengths of the cuckoo eggs the same in the different sub-species?

Display of Cuckoo Bird Egg Lengths

Here is a plot of egg lengths (mm) of cuckoo bird eggs categorized by the species of the host bird.

Analysis of Variance

- Analysis of variance (ANOVA) is a statistical procedure for analyzing data that may be treated as multiple independent samples with a single quantitative measurement for each sampled individual.
- ANOVA is a generalization of the methods we saw earlier in the course for two-independent samples.
- The bucket of balls model is that we have $I$ different buckets of balls, each of which contains numbered balls.
- The populations means and standard deviations of the numbers in each bucket are $\mu_i$ and $\sigma_i$ respectively for $i = 1, \ldots, I$.
- In ANOVA, we often assume that all of the population standard deviations are equal.
Notation

This notation is used to describe calculations of variability within samples and variability among samples, although for historical reasons of poor grammar, the term between samples is more commonly used.

\[ y_{ij} \] = the \( j \)th observation in the \( i \)th group
\[ I \] = the number of groups
\[ n_i \] = the \( i \)th sample size
\[ \bar{y}_i \] = the mean of the \( i \)th sample
\[ n^* = \sum_{i=1}^{I} n_i \] = the total number of observations
\[ \bar{y} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} y_{ij}}{n^*} \] = the grand mean

Sums of Squares within Groups

We measure variability by sums of squared deviations. The sums of squares within groups, or SS(within), is a combined measure of the variability within all groups.

\[
SS(\text{within}) = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \\
= \sum_{i=1}^{I} (n_i - 1)s_i^2
\]

Notice that this measure of variability is a weighted sum of the sample variances where the weights are the degrees of freedom for each respective sample.

The Big Picture

- ANOVA is a statistical procedure where we test the null hypothesis that all population mean are equal versus the alternative hypothesis that they are not all equal.
- The test statistic is a ratio of the variability among sample means over the variability within sample means.
- When this ratio is large, this indicates evidence against the null hypothesis.
- The test statistic will have a different form than what we have previously seen. The null distribution is an \( F \) distribution, named after Ronald Fisher.
- An ANOVA table is an accounting method for computing the test statistic.
- We introduce a lot of new notation on the way...
Sums of Squares Between (Among) Means

- We measure variability by sums of squared deviations. The sums of squares between groups, or \( SS(\text{between}) \), is a measure of the variability among sample means.
  \[
  SS(\text{between}) = \sum_{i=1}^{I} n_i (\bar{y}_i - \bar{y})^2
  \]
- Notice that this measure of variability is a weighted sum of the deviations of the sample means from the grand mean, weighted by sample size.

Degrees of Freedom

- The degrees of freedom within samples is simply the sum of degrees of freedom for each sample.
- This is equal to the total number of observations minus the number of groups.
  \[
  df(\text{within}) = \sum_{i=1}^{I} (n_i - 1) \quad = \quad n^* - I
  \]

Mean Square Within

- In ANOVA, a mean square will be the ratio of a sum of squares over the corresponding degrees of freedom.
  \[
  MS(\text{within}) = \frac{SS(\text{within})}{df(\text{within})} = \frac{(n_1 - 1)s_1^2 + \cdots + (n_I - 1)s_I^2}{n^* - I}
  \]
- In other words, the mean square within is a weighted average of the sample variances where the weights are the degrees of freedom within each sample.
- The square root of the mean square within is the estimate of the common variance for all the \( I \) populations.
  \[
  s_{\text{pooled}} = \sqrt{MS(\text{within})}
  \]
The F Statistic

- The F statistic is the ratio of the mean square between over the mean square within.
  \[ F = \frac{\text{MS(between)}}{\text{MS(within)}} \]
- If the populations are normal, the population means are all equal, the standard deviations are all equal, and all observations are independent, then the F statistic has an F distribution with I − 1 and n* − I degrees of freedom.
- An F distribution is positive and skewed right like the chi-square distribution, but it has two separate degrees of freedom, the numerator degrees of freedom and the denominator degrees of freedom.
- If X_1 and X_2 are independent χ² random variables with k_1 and k_2 degrees of freedom respectively, then
  \[ F = \frac{X_1}{k_1} \frac{X_2}{k_2} \]
  has an F distribution with k_1 and k_2 degrees of freedom.

Mean Square Between

In ANOVA, a mean square will be the ratio of a sum of squares over the corresponding degrees of freedom.

\[ \text{MS(between)} = \frac{\text{SS(between)}}{\text{df(between)}} = \frac{(n_1 - 1)s_1^2 + \cdots + (n_I - 1)s_I^2}{n^* - I} \]

ANOVA Table for the Cuckoo Example

```
> fit = aov(eggLength ~ birdSpecies)
> anova(fit)
Analysis of Variance Table

  Df  Sum Sq Mean Sq F value  Pr(>F)
birdSpecies  5   42.94  8.588  10.388 3.152e-08 ***
Residuals 114  94.24  0.827
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- In R, the columns are in an unconventional order and there is no row for totals.
- R names the row corresponding to between by the corresponding categorical variable.
- R names the row corresponding to within Residuals.
- There are six groups and 120 total observations, which explains the degrees of freedom column.
- Each mean square is the ratio of the corresponding sum of squares and degrees of freedom.

Total Sum of Squares

- If we treated all observations as coming from a single population (which would be the case if all population means were equal and all population standard deviations were equal as well), then it would make sense to measure deviations from the grand mean.
- This is the total sum of squares.
  \[ \text{SS(total)} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}.)^2 \]
- It turns out that the total sum of squares can be decomposed into the sum of squares within and the sum of squares between.
  \[ \text{SS(total)} = \text{SS(within)} + \text{SS(between)} \]
- Similarly, the total degrees of freedom would be n* − 1.
- There is a similar decomposition.
  \[ \text{df(total)} = \text{df(within)} + \text{df(between)} \]
  \[ n^* - 1 = (n^* - I) + (I - 1) \]
Multiple Comparisons

- In general, we may wish to follow up on a general F test.
- If the population means are not all the same, can we further identify the differences?
- In a sense, we would like to simultaneously test the equality of each possible pair of population means, correcting for the fact that when we make multiple comparisons, it is more likely that at least one p-value will be small (or significant), even when all of the population means are equal.
- There are several methods for correcting for multiple comparisons.
- Your textbook describes two of these, the Newman-Keuls procedure which is valid only for balanced one-way ANOVA analyses for hypotheses involving differences in population means, and the Bonferroni procedure which is much more general, but highly conservative.

Assumptions

For ANOVA to be applicable, the following conditions must hold.

- Each group must be thought of as a random sample from a population.
- Observations within each group should be independent of each other.
- The $I$ samples must be independent of each other.
- The $I$ populations must be approximately normal with a common standard deviation.

In practice, the central limit theorem steps in if the populations differ from normality — the methods are valid for nonnormal populations if the sample sizes are large enough.

The method is also fairly robust to the assumption of equal variances. If sample variances differ by a fairly large factor, then you might want to consider a modified method.

Newman-Keuls

- The Newman-Keuls procedure is based on the distribution of the differences of order statistics of sample means when all population means are equal and sample sizes are equal.
- The order statistics are the sample means sorted from smallest to largest.
- The procedure is as follows:
  - Sort the sample means in increasing order.
  - Compute critical values $R_i = q_i \cdot \text{pooled}/\sqrt{n}$.
  - Make pairwise comparisons.
- It is easiest to learn from a numerical example.
Bonferroni

- The Bonferroni method is a general method for multiple comparisons. Unlike Newman-Keuls which is especially tailored to balanced one-way ANOVA designs considering all pair-wise comparisons,
- Bonferroni is completely general and is applicable to any prespecified set of k hypotheses, independent or not, in any setting.
- Because of this generality, the method is conservative.
- The basic idea if there are k tests is to test each of them at the $\alpha/k$ level. Then, the probability of at least one significant p-value is no more than $\alpha$ (but often is considerably less).
- In the setting of one-way ANOVA with I groups, there are $k = I(I - 1)/2$ pairwise comparisons.
- For a balanced design, we can construct confidence intervals of the same width for each difference in population means.

Newman-Keuls (cont.)

- Consider Exercise 11.24. The sorted sample means are:

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>A</th>
<th>B</th>
<th>E</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3.70</td>
<td>4.37</td>
<td>4.76</td>
<td>5.38</td>
<td>5.41</td>
</tr>
</tbody>
</table>

- Each sample has size 9.
- The scale factor is $s_{pooled}/\sqrt{n} = \sqrt{0.2246/9} = 0.158$.
- For significance level $\alpha = 0.05$, we have this table.

<table>
<thead>
<tr>
<th>i</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q_i$</td>
<td>2.86</td>
<td>3.44</td>
<td>3.79</td>
<td>4.04</td>
</tr>
<tr>
<td>$R_i$</td>
<td>0.45</td>
<td>0.54</td>
<td>0.60</td>
<td>0.64</td>
</tr>
</tbody>
</table>

- The values for $q_i$ come from Table 11 on page 697 of the textbook. There are total of 40 degrees of freedom (45-5).
- The value of 4.04 implies that the statistic $\frac{\bar{y}_{max} - \bar{y}_{min}}{s_{pooled}/\sqrt{9}}$ has a 5% probability of exceeding 4.04 when there are five independent random samples of size nine each taken from normal populations with the same mean and standard deviation.

Bonferroni Example

- For the data in the previous example, there are $I = 5$ groups and hence $k = 5 \times 4/2 = 10$ comparisons.
- A 95% confidence interval corresponds to a hypothesis test with $\alpha = 0.05$.
- We can construct ten 99.5% confidence intervals for each difference in population means and be at least 95% confident that all ten contain the true differences.
- Each confidence interval has a margin of error of the form

$$t \times s_{pooled} \times \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}.$$

- The pooled estimate is $s_{pooled} = \sqrt{0.2246} = 0.474$.
- There are 40 degrees of freedom — the 0.9975 quantile is 2.971.
- The margin of error is 0.664. Notice that this is larger than the Newman-Keuls method.
- In this case, the pairs AB, BE, and AD are not significantly different.

Newman-Keuls (cont.)

- To complete the problem, we make the pairwise comparisons from large groups to small.
- The maximum minus the minimum for the only group of five consecutive means is 5.41 - 3.70 = 1.71, much larger than the critical value of 0.64.
- There are two sets of four means, $E - C = 1.68$ and $D - A = 1.04$, both of which are larger than 0.60.
- There are three sets of three consecutive means, and the differences (1.06, 1.01, 0.65) are all larger than 0.54.
- There are four sets of two consecutive means, two of which have differences smaller than 0.45. So we conclude that A has the smallest population mean, A and B are next, but cannot be distinguished, and that E and D are the largest, but cannot be distinguished at the 5% significance level.
Example — Soybean Experiment

- A researcher investigates the growth of soybeans under four experimental conditions in a $2 \times 2$ factorial design.
- One factor is shaking. Soybean seedlings are either shaken for 20 minutes two times daily or not shaken at all. The levels for this factor are stress and control.
- Another factor is light. Soybeans are grown with light levels of low or moderate.

There are four possible treatment combinations.

- 1. Control, low light
- 2. Stress, low light
- 3. Control, moderate light
- 4. Stress, moderate light

There are 13 observations for each treatment combination.

Plots of Data

```r
> soy = read.table("soybean.txt", header = T)
> attach(soy)
> group = as.factor(paste(shaking, light))
> plot(area ~ group)
```

A model for one-way ANOVA

- It can be useful to model one-way ANOVA in terms of a treatment effect instead of population means.
  
  \[ y_{ij} = \mu + \tau_i + \epsilon_{ij} \]

- In this notation, $\mu$ is a common constant, $\tau_i$ is a treatment effect, and $\epsilon_{ij}$ is the error term.
- The expected value of the $j$th observation in treatment group $i$ is $\mu + \tau_i$.
- Think of the treatment effects as being adjustments relative to some base level.
- The null hypothesis of equal means is then equivalent to the null hypothesis of no treatment effects.

\[ H_0: \tau_1 = \cdots = \tau_I = 0 \]

Two-way ANOVA

- Two-way ANOVA is similar to one-way ANOVA except that there are two categorical explanatory variables and a single quantitative response variable.
- A two-way setting can arise with a blocked design where treatments are allocated at random within two or more blocks.
- There would then be a possible block effect, a treatment effect, and maybe even an interaction.
- We also may have a factorial experiment where there are two treatments of interest and we observe individual observations for each possible treatment combination.
- A model for this is
  
  \[ y_{ijk} = \mu + \tau_i + \beta_j + \epsilon_{ijk} \]

  where $\tau_i$ is one treatment effect, $\beta_j$ is another treatment effect, and $\epsilon_{ijk}$ is the random error term.
A model without interaction

The following table then shows the expected growth for each experimental condition in the parameters of the problem.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Light</th>
<th>Shaking</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Control</td>
<td>μ</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>Stress</td>
<td>μ + β₂</td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>Control</td>
<td>μ + τ₂</td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>Stress</td>
<td>μ + τ₂ + β₂</td>
<td></td>
</tr>
</tbody>
</table>

Observations

- Each sample looks to be fairly symmetric.
- The amount of variability within each sample looks to be about the same.
- There are no extreme outliers.
- A model that assumes normality within groups will be reasonable.
- The group means look different.
- Light level seems to have a greater effect than shaking — there is a bigger difference between the moderate groups and the low groups than there is between the control groups and the stress groups.
- There does not appear to be much of an interaction effect.

Using R to fit the model

If we are only interested in the global $F$ test, we can use the `aov` function to fit the model and produce the ANOVA table.

\[
\begin{align*}
\text{Df} & \quad \text{Sum Sq} & \quad \text{Mean Sq} & \quad \text{F value} & \quad \text{Pr(}F\text{)} \\
\text{light} & \quad 1 & \quad 42752 & \quad 42752 & \quad 4.8714 & \quad 7.136e-09 & \quad *** \\
\text{shaking} & \quad 1 & \quad 14858 & \quad 14858 & \quad 1.6931 & \quad 0.0001479 & \quad *** \\
\text{Residuals} & \quad 49 & \quad 43003 & \quad 878 & \quad 878 & \quad & \quad \\
\end{align*}
\]

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Sometimes we may be interested in estimation in addition to testing. In this case, the function `lm` (for linear model) is a better choice.

A model without interaction

The mathematical notation for this model is

\[
y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ijk}.
\]

The parameter $\tau$ is the effect due to light, the parameter $\beta$ is the effect due to shaking, and $\epsilon$ represents the independent normal error.

When a categorical variable is included in a linear model, a constraint must be imposed for the parameters to be identifiable.

There are several ways to do this. In R, the levels are ordered alphabetically and the effect due to the first level for each factor is set to 0.

We have the factor `light` with levels `low` and `moderate`, and the factor `shaking` with levels `control` and `stress`. 
Residual Plots in R

We use the functions `resid` and `fitted` to extract the residuals and fitted values from each model and plot them using the function `plot`. The `abline` function adds a line to an existing plot. I added some options to make the plot prettier.

Residual Plots in R

```r
> plot(fitted(fit1), resid(fit1), xlab = "Fitted Values", ylab = "Residuals", + pch = 16) > abline(h = 0)
```

Using R to estimate the parameters

```r
> fit1b = lm(area ~ light + shaking) > summary(fit1b)
Call:
  lm(formula = area ~ light + shaking)
Residuals:
   Min      1Q  Median      3Q     Max
-68.5577 -19.0288  -0.3846  18.8462  59.7885
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)        246.019     7.116   34.575  < 2e-16 ***
lightmoderate      57.346     8.216    6.980  7.14e-09 ***
shakingstress      -33.808     8.216   -4.115  0.000148 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.62 on 49 degrees of freedom
Multiple R-Squared: 0.5726, Adjusted R-squared: 0.5551
F-statistic: 32.82 on 2 and 49 DF, p-value: 9.029e-10

The summary function shows the parameter estimates along with estimates of standard errors. This allows us to make confidence intervals of parameters in the model.

Residual Plots

- The model had four total free parameters, $\mu$, $\tau_2$, and $\beta_2$, the parameters for the mean and $\sigma$, the common standard deviation for all error distributions.
- The estimates of these parameters are denoted $\hat{\mu}$, $\hat{\tau}_2$, $\hat{\beta}_2$, and $\hat{\sigma}$. You should know how to identify these estimates from R output.
- These estimates provide is estimates for the means of each treatment combination, also called fitted values.
- The difference between each individual observation and its fitted value is called a residual.
- It is always a good idea to plot residuals versus fitted values to examine if the model assumptions are reasonable.
A model with interaction

Here is the mathematical display of a model with an interaction term.

\[ y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk} \]

While there are nominally four interaction variables, by convention we set any with \( i = 1 \) or \( j = 1 \) to be 0. So in a \( 2 \times 2 \) factorial experiment, there is only a single free parameter for the interaction effect.

With the previous three free parameters, this makes four, equal to the number of treatment combinations.

Interaction

- The previous model had three free parameters for the mean, but there were four treatment combinations.
- The model implied no interaction.
- In other words, the effect of stress on expected growth was the same for plants in low light and moderate light conditions.
- We can test this assumption informally with an interaction plot.
- In an interaction plot, the \( x \)-axis is one of the categorical variables, the \( y \)-axis is the response variable, the means of each treatment combination are drawn, and lines connect points with the same level of a second explanatory factor.

Fitting the model in R

In R, the interaction term is `light:shaking`. The symbol \( * \) in the formula is a shorthand to include each factor and their interaction in the model.

\[ \text{> fit2 = aov(area ~ light * shaking)} \]
\[ \text{> anova(fit2)} \]

Interaction Plots in R

The first two arguments are explanatory factors, the third is the continuous response variable. Nearly parallel lines are consistent with negligible interaction.

\[ \text{> interaction.plot(shaking, light, area)} \]
Estimating the parameters

> fit2b = lm(area ~ light * shaking)
> summary(fit2b)

Call:
  lm(formula = area ~ light * shaking)

Residuals:
   Min     1Q   Median     3Q    Max
-67.8462 -19.3846  -0.3846  19.5385  59.0769

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2 45.308    8.2 99.559 < 2e-16 ***
lightmoderate 58.769    11.736 5.007 7.86e-06 ***
shakingstress -32.385    11.736 -2.759 0.00817 **
lightmoderate:shakingstress -2.846    16.598 -0.171 0.86457

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 . ‘ ’ 1

Residual standard error: 29.92 on 48 degrees of freedom
Multiple R-Squared: 0.5729,  Adjusted R-squared: 0.5462
F-statistic: 21.46 on 3 and 48 DF,  p-value: 5.867e-09