ReCap Part I (Chapters 1,2,3,4) Quantitative reasoning is based on models, including statistical analysis based on models.

ReCap Part II (Chapters 5,6,7)
Hypothesis testing uses the logic of the null hypothesis to declare a decision.
Estimation is concerned with the specific value of an unknown population parameter.

ReCap (Ch 9,10,11) The General Linear Model with a single explanatory variable.
ReCap (Ch 12) GLM with more than one regression variable (multiple regression)
ReCap (Ch 13) GLM with more than one categorical variable (ANOVA).

Today: Special case of two factor ANOVA: Hierarchical ANOVA
Both factors random.
The logical relation of one factor to another is hierarchical:
   one factor is nested within another.
This is in contrast to crossed designs, such as two-way ANOVA and randomized blocks.

Wrap-up. Comparison of hierarchical with two-way ANOVA.

Two-way ANOVA has an interaction term. Testing starts with this term.

In randomized blocks, the interaction term is present logically, but assumed to be zero if treatments assigned randomly to blocks.

Hierarchical ANOVA differs from crossed design. The interaction term is known to be zero, because units of analysis cannot be matched across treatments.
Introduction.

We used the GLM to analyze a response variable in relation to two categorical variables (two-way ANOVA). In the examples so far the levels within the one explanatory variable (e.g. food type) could be matched with levels in a second explanatory variable (e.g. sex). Consequently, we can display the data in a two-way table. This is called a crossed design: levels within one factor can be matched with levels in another factor.

<table>
<thead>
<tr>
<th>Response variable</th>
<th>Explanatory variable</th>
<th>Explanatory variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Limpet respiration</td>
<td>Species (fixed)</td>
<td>Salinity (fixed by experiment)</td>
</tr>
<tr>
<td>Rat weight gain</td>
<td>Protein source (fixed)</td>
<td>Protein Level (fixed)</td>
</tr>
<tr>
<td>Hours of extra sleep</td>
<td>Subject (random)</td>
<td>Drug (fixed)</td>
</tr>
<tr>
<td>Tribolium dry weight</td>
<td>Block (random)</td>
<td>Genotype (fixed)</td>
</tr>
</tbody>
</table>

Sokal and Rohlf (1995 Box 10.1) report fly wing lengths of 4 flies in each of 3 cages.

<table>
<thead>
<tr>
<th>Response variable</th>
<th>Explanatory variable</th>
<th>Explanatory variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fly wing length</td>
<td>Cage (random)</td>
<td>Fly (random)</td>
</tr>
</tbody>
</table>

We again have two explanatory variables. But now we cannot match the levels of one variable – fly -- across the other (cage). We have no information to match a fly in Cage I (Fly A, B, C, D) to a fly in Cage II (Fly E, F, G, H). Fly is nested within cage. Here is a diagram.

Nor do we have information to match wing measurements across flies, except perhaps as first or second. Measurements are nested within fly.

Tree. A graphical expression for the idea of hierarchical ANOVA is a tree. One can draw a few branches representing genera, then for each branch draw twigs representing species. The twigs cannot be aligned across branches, so the design is hierarchical.

Mobile. Another visualization is a hanging mobile. Near the top of the mobile are branches that rotate around a balance point. Beneath each branch there are sub-branches, all of which rotate around a balance point below the branch. The sub-branches beneath one branch rotate independently of those beneath another branch. They cannot be aligned. This represents a hierarchical design.
Crossed versus nested factors.
Two crossed factors can also be shown as a branching diagram. To avoid this we inspect a two way table.

<table>
<thead>
<tr>
<th>Fly within Cage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cage</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

The table has more empty than occupied cells because of the nesting. As a result we cannot estimate the interaction term. Instead, the interaction term is joined with the lower level term as follows:

\[
\text{Fly} + \text{Cage} \times \text{Fly} \rightarrow \text{Fly(Cage)}
\]

More generally:

\[
B = + A \times B \rightarrow B(A)
\]

The symbol \( B(A) \) is read from left to right as “B within A.” At first glance the nested factor seems to belong inside the brackets. However, \( B(A) \) is easier to use in reading, speaking, and thinking because “B within A” reads naturally from left to right. The logically correct notation also reads left to right: \( B \subset A \).

Notation such as \( A/B \) for nesting of \( B \) within \( A \) does occur, but almost all statistical packages use the conventional notation found in textbooks. Here is an accounting of degrees of freedom showing how joining an interaction term with the lower level term produces the df for the nested term.

<table>
<thead>
<tr>
<th>Crossed</th>
<th>Nested</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>df</td>
</tr>
<tr>
<td>Factor A</td>
<td>3-1 = 2</td>
</tr>
<tr>
<td>Factor B</td>
<td>4-1 = 3</td>
</tr>
<tr>
<td>A \times B</td>
<td>2 \times 3 = 6</td>
</tr>
</tbody>
</table>

\[
B + A \times B \rightarrow B(A)
\]

Factor B df + Factor A \times B df = Factor B(A) df

\[
3 + 6 = 9
\]

A nested factor ANOVA can be fully random, or mixed.

<table>
<thead>
<tr>
<th>Both Random</th>
<th>Mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>B crossed with A</td>
<td>B nested in A</td>
</tr>
<tr>
<td>A Random</td>
<td>A Random</td>
</tr>
<tr>
<td>B Random</td>
<td>B(A) Random</td>
</tr>
<tr>
<td>A \times B</td>
<td>Random</td>
</tr>
</tbody>
</table>

Not all random factors are nested. Crossed random factors occur in experiments where two sources of variation are controlled statistically, such as a latin square.
Example Data from Sokal and Rohlf (1995) Box 10.1
Two measurements made on the left wings of 4 mosquitos in each of 3 cages.
This is a nested design. We cannot match flies across cages. We cannot match wing
measurements across flies.

1. Construct model
Verbal model.
Does mosquito wing length vary among cages as well as among mosquitos?

Graphical model. Plot of the means of the measurements for each fly in each cage.

Response variable is winglength
L = micrometer units (ratio type of scale)

Explanatory variable is cage
X_{cage} = (categorical variable)

Second explanatory variable is fly within cage
or equivalently
X_{fly \subset cage} (categorical variable) 
X_{fly(cage)}

Each fly gets a different label. We avoid using the label “Fly A” in cage II or III
because each fly is another sample from a population.

* X_{fly(cage)} = Fly A  Fly B  Fly C  Fly D  in cage I
* X_{fly(cage)} = Fly E  Fly F  Fly G  Fly H  in cage II
* X_{fly(cage)} = Fly J  Fly K  Fly L  Fly M  in cage III

Write GLM: Y = \beta_0 + \beta_{cage}X_{cage} + \beta_{fly(cage)}X_{fly(cage)} + res
Component Y_{ijk} = \mu + A_i + B_{ij} + \varepsilon_{ijk}

GLM notation, showing parameters, is shown along with component notation, showing
subscripts and two random factors, labelled with roman letters.
2. Execute analysis.
Place data in model format:
- Column labelled \( L \), with response variable fly wing length
- Column labelled \( X_{\text{cage}} \), with explanatory variable \( X_{\text{cage}} = \text{I, II, or III} \)
- Column labelled \( X_{\text{fly(cage)}} \) with label (number) for each fly

Code model statement in statistical package according to the GLM

\[
Len = \beta_0 + \beta_{\text{cage}} \cdot X_{\text{cage}} + \beta_{\text{fly(cage)}} \cdot X_{\text{fly(cage)}} + \varepsilon
\]

<table>
<thead>
<tr>
<th>Minitab</th>
<th>R-code</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTB &gt; anova 'Len' = 'cage' 'fly'('cage'); SUBC&gt; random 'cage' 'fly'('cage').</td>
<td>FlyMod8&lt;-aov(Len ~ Cage + Error(Cage/Fly), SRBx10_1)</td>
</tr>
</tbody>
</table>

This produces estimates of the overall mean (\( \hat{\beta}_0 \)) and the mean for each cage (\( \hat{\beta}_0 + \hat{\beta}_{\text{cage}} \cdot X_{\text{cage}} \)).

The fitted values are the means for each fly (\( \hat{\beta}_0 + \hat{\beta}_{\text{cage}} \cdot X_{\text{cage}} + \hat{\beta}_{\text{fly(cage)}} \cdot X_{\text{fly(cage)}} \)).

\[
\hat{\beta}_0 = \text{mean}(L) = 24.11 \#599.2 = 66.633 : m
\]

\[
\hat{\beta}_0 + \hat{\beta}_{\text{cage}} = \text{mean}(L_{\text{cage=I}}) = 8.11 \#82.7 = 72.84 : m \\
\hat{\beta}_0 + \hat{\beta}_{\text{cage}} = \text{mean}(L_{\text{cage=II}}) = 8.11 \#79.7 = 59.96 : m \\
\hat{\beta}_0 + \hat{\beta}_{\text{cage}} = \text{mean}(L_{\text{cage=III}}) = 8.11 \#36.8 = 67.10 : m
\]

3. Evaluate the model  Plot residuals versus fitted values.
3. Evaluate the model

A. No line fitted in model, so skip evaluation of straight line assumption.


No systematic change in residuals with increase in fitted values (i.e. no cones). Normal? Yes.

Independent? Lag plot cannot be used when replication is less than 4 at the lowest level in the analysis (at the level of fly in this analysis). The correlation of adjacent residuals is by definition \( r = -0.5 \) (Anscombe, F.J. Tukey J.W. 1963. The examination and analysis of residuals. *Technometrics* 5: 141-160) where there are two observations at the lowest level, as in the fly winglength data. Induced correlation is evident in the downward trend in plot of adjacent values.

The errors are acceptably homogeneous and normal.
4. **State population and whether sample is representative.**

This is a laboratory study with a well-established measurement protocol that generates data that meets Hacking’s (1965) requirement for likelihood inference (Hacking. I. 1965 *The Logic of Statistical Inference*). Inference requires procedural statement that generates comparable data, which becomes the population. A better and better estimate of the value of a parameter emerges from the law of large numbers as Hacking’s data producing machine runs longer and longer.

Inference is to a population of wing measurements according to a measurement protocol. Inference is to cages similar to those in this study.

4. **Is hypothesis testing appropriate?**

The goal of the analysis is an estimate of variance at different levels. Estimates such as these are central to statistical design. We will calculate and report likelihood ratios as a measure of evidence.

5. **Partition df and SS according to model.**

Compute total df. \( n-1 = 24-1 = 23 \)

Partition according to model

- cage term. 3 cages hence \( 3-1 = 2 \) df
- fly term. 4 flies per cage hence \( 4-1 = 3 \) df for each cage.
- \( 3^2 \) df for the fly within cage term.

Calculate \( SS_{\text{total}} = 23 \times \text{Var}(Len) = 23 \times 104.43 = 2401.98 \)

By hand: \( SS_{\text{total}} = \Sigma Y^2 - n^{-1}(\Sigma Y)^2 = 108962 - 2^{-1} \cdot 1599.22^2 = 2401.98 \)

In Minitab:

```
MTB > let kl = ssq('Len')
MTB> print kl
kl  2401.98
```

In a spreadsheet

```
=DevSQ(A1:A24)
```

In R

```
> sum( (x - mean(x) )^2 )
[1] 2401.98
```

Use statistical software to partition the \( SS_{\text{total}} \) and produce the ANOVA table.

\[
\text{GLM: } Y = \beta_0 + \beta_{\text{cage}} X_{\text{cage}} + \beta_{\text{fly(cage)}} X_{\text{fly(cage)}} + \text{res}
\]

The subcommand `SUBC> random` forces Minitab to compute the correct F-ratio. If not present, both F-ratios are relative to \( MS_{\text{residual}} \).

Here is R-code to obtain correct df, SS, and MS:

```
FlyMod8 <- aov(WLEN ~ Cage + Error(Cage/Fly), SRBx10_1)
```
5. Calculate SS, partition according to model.

Here is the partitioning of the SS\text{total}.

<table>
<thead>
<tr>
<th>GLM:</th>
<th>Y = $o + $\text{cage} X \text{cage} + $\text{fly \ diam} X \text{fly \ cage} + \text{res}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source:</td>
<td>\text{total} = \text{cage} \text{fly(cage)} \text{res}</td>
</tr>
<tr>
<td>24–1 =</td>
<td>3–1 + 3*(4–1) + 12</td>
</tr>
<tr>
<td>2402 =</td>
<td>665.68 + 1720.68 + 15.62</td>
</tr>
</tbody>
</table>

$1 - R^2 = (15.62/2401.97) = 0.65 \%$

$LR = (1-R^2)^{-24/2} = 1.7 \times 10^{26}$

The evidence for the model is strong so we proceed to a calculation of the likelihood for each component of variation. In a model with random factors the likelihood ratio and the F-ratio are formed relative to the correctly nested ratio of expected mean squares, as described in texts in experimental design (e.g. Cochran and Cox 1957, Quinn and Keough). The likelihood ratio for the cage term is not formed relative to the residual SS. It is formed relative to the random term below it in the ANOVA table.

<table>
<thead>
<tr>
<th>df</th>
<th>SS</th>
<th>1+ SS\text{ratio}</th>
<th>LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>665.68</td>
<td>1.39</td>
<td>51</td>
</tr>
<tr>
<td>9</td>
<td>1720.68</td>
<td>111.16</td>
<td>3.6 \times 10^{24}</td>
</tr>
<tr>
<td>12</td>
<td>15.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>2401.98</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The likelihood ratio for Fly(Cage) is formed relative to the random term below it.

The evidence for the fly within cage term is far stronger than the evidence for the cage term.


The parameters in this analysis were not of interest. The variance due to each factor is of interest. The interest is in the amount of variation, not the contrasts in means. The amount of variability due to each factor is used to design efficient experiments.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cage</td>
<td>2</td>
<td>665.68</td>
<td>27.7</td>
</tr>
<tr>
<td>FlydCage</td>
<td>9</td>
<td>1720.68</td>
<td>71.6</td>
</tr>
<tr>
<td>Error</td>
<td>12</td>
<td>15.62</td>
<td>0.7</td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>2401.97</td>
<td></td>
</tr>
</tbody>
</table>

Most of the variability (72\%) is among flies. As a result, in designing an experiment we would want to use many flies to reduce the MS at this level. The percent variability is less at the level of cages, so we would not need to use many cages to reduce the MS at the cage level. The variability due to measurement error is negligible, so we do not need to devote effort to repeating measurements on a single fly.