

Some helpful notes for COMPONENT 2, SECTION 4.

Before doing a nested ANOVA, you will screen your data based on the following criteria:

1. *Select only records for which tarsus is non missing.*
2. *Keep only individuals for which you have 2 or more measurements.*
3. *Keep only 2 measures per bird.*
4. *Keep only years for which you have at least 10 males AND 10 females with 2 records each.*

Selecting only records for which tarsus is non-missing is straightforward and requires a simple conditional statement. The second step is a bit more complicated...

In order to keep only individuals which have 2 or more measurements, you will need to create a new dataset (ALL) with a conditional statement that eliminates individuals with only 1 measurement. First, the data need to be sorted by CWS. Then a new dataset is created that deletes all observations for which the first and last CWS numbers are the same:

```
DATA ALL;
  SET ALL;
  BY CWS;
  IF FIRST.CWS AND LAST.CWS THEN DELETE;
RUN;
```

If there is only one CWS number (one measurement for that individual) then both conditions, FIRST.CWS and LAST.CWS, are true, that is, they each equal 1 and would be deleted.

Next you want to only keep 2 measures per bird. This can be done by keeping only the first and last measurements for each bird. From dataset ALL, two new datasets are created (i.e., FIRST, LAST) that each contain records for only the first or last observations. For example:

```
DATA FIRST;
  SET ALL;
  BY CWS;
  IF FIRST.CWS;
RUN;
```

The same would be repeated for dataset LAST. The two datasets are then combined not merged:

```
DATA COMBINED;
  SET FIRST LAST;
RUN;
```

When datasets are combined they are simply added together, one on top of the other, so that no observations from the source files are deleted. Try merging the two datasets (i.e., MERGE FIRST LAST; BY CWS; RUN;) and notice how the new dataset is different.

Before you go to step 4, you will need to screen the data for a known error, where certain individuals were recorded as both a female and male in different observations. These records need to be identified and placed in a new file. This file can then be merged against the

COMBINED dataset as an exclusion (translation: it is subtracted from COMBINED). To do this, recall the dataset with only the first observations (FIRST) and only keep variables CWS and SEX:

```
DATA FIRST2;
  SET FIRST;
  KEEP CWS SEX;
RUN;
```

Then do the same for the LAST dataset, except rename SEX to SEX2:

```
DATA LAST2;
  SET LAST;
  SEX2=SEX;
  KEEP CWS SEX2;
RUN;
```

Then create a new dataset (SEXCHNG) by merging the two datasets by CWS. Next, you will screen SEXCHNG to only contain records where the sexes are different (i.e., SEX and SEX2 are not equal) for each CWS number:

```
DATA SEXCHNG2;
  SET SEXCHNG;
  IF SEX=SEX2 THEN DELETE;
  KEEP CWS;
RUN;
```

Use the following code to create a new dataset (FINAL) that excluded these sex change individuals from COMBINED:

```
DATA FINAL;
  MERGE COMBINED (IN=A) SEXCHNG2 (IN=B);
  BY SEX;
  IF A AND NOT B;
RUN;
```

The same result could be achieved using other approaches, however, creating a separate file with the 'bad' information can be a handy tool in other data screening situations. For example, imagine that you supervised the collection of this goose data up in the arctic. During data collection, it became obvious that one technician consistently had trouble sexing the birds. Rather than do repeated screening operations for each analysis to remove questionable data recorded by this person, it would be easier to just create a separate file with all observations taken by this 'challenged' technician and exclude it from all analyses where sex is used.

Once the data are cleaned up, you can proceed with step 4 (hint: use PROC FREQ!) and the nested analysis...good luck!