ljgormezano 2005

DATA CLEANING GUIDE

In general, if it appears that a value was entered incorrectly, then it should be converted to a 'missing' value. The entire observation should not be deleted because there still may be usable data in other variables. For continuous variables, you must decide whether especially large or small values are outliers (and should be converted to 'missing') or whether these values are plausible measurements that reflect the 'normal' variation in the data. This is a judgment call that you as the biologist must make. There is no correct answer and it is expected that final results will differ between projects. However, each conditional statement used to clean the data should have a comment with it explaining why you decided to make the changes that you did.

In the gosling dataset there are a few observations that were recorded for ross goslings (as opposed to snow goslings) – these should be permanently deleted from the dataset. The following includes some facts about the data that you will help you decide what to delete and what to convert to 'missing':

1. For gosling color at the time of banding (COLOR), 'B' and 'W' mean blue and white plumage, respectively. Values of 'R' and '7' are recorded for ross geese and a value of '1' means the bird was a white ('W') snow gosling.

2. For SEX, 'F', 'M' and 'U' denote female, male and unknown, respectively.

3. Gosling colors at hatching (GCOLOR) are recorded as either 'B', 'W' and '0'. A value of '0' means the color was not determined.

4. The conditions of goslings in each nest were checked simultaneously whether or not all birds were hatched. Six different conditions (GOS_COND) were recorded:
   a. 'E' – egg is not yet hatched
   b. 'P' – egg is pipped (chipped – when hatchling is starting break through)
   c. '5' – half of egg shell is broken away as hatchling is emerging
   d. 'W' – hatchling is wet
   e. 'D' – hatchling is damp
   f. 'F' – hatchling is fluffy

5. Development of feathers can be highly dependent on environmental factors, therefore, they can vary significantly from year to year. In the gosling dataset, measurements of both the primary and midtail feathers reflect this variation and both exhibit bimodal distributions with a very small second peak. Keep this in mind when screening for outliers.

6. Banding of goslings began about 3-4 weeks after hatching. Therefore, banding age (BNDAGE), which is recorded in estimated days after hatching, should not be much less than 21 days. Smaller values are either errors or a coding for 'missing' values.