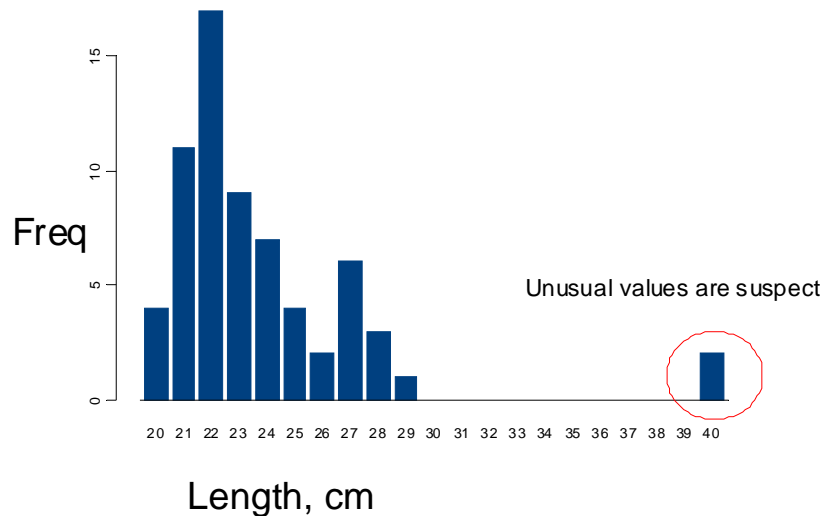# Data Quality Control Procedures

John M. Hoenig

Prepared for the 2nd CRFM

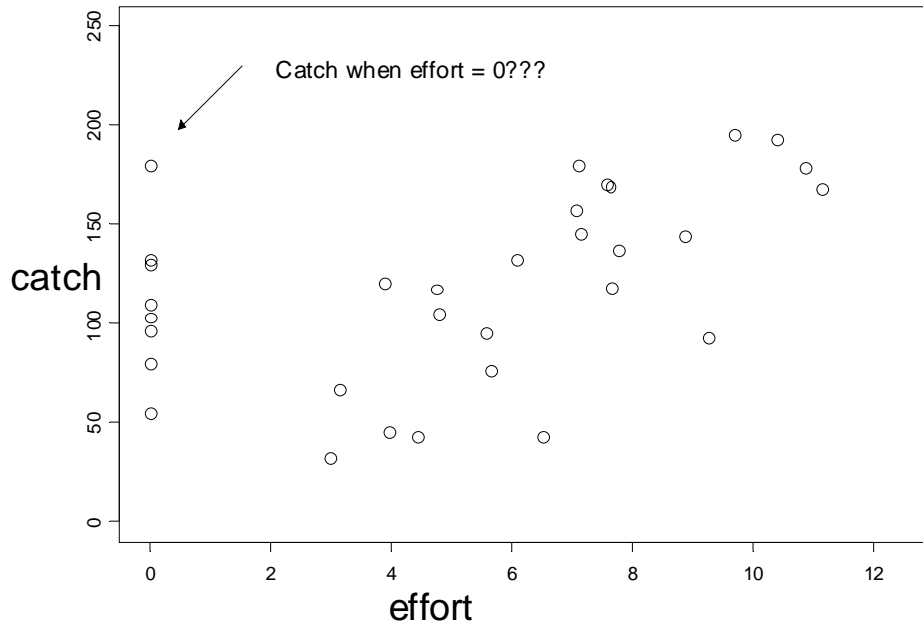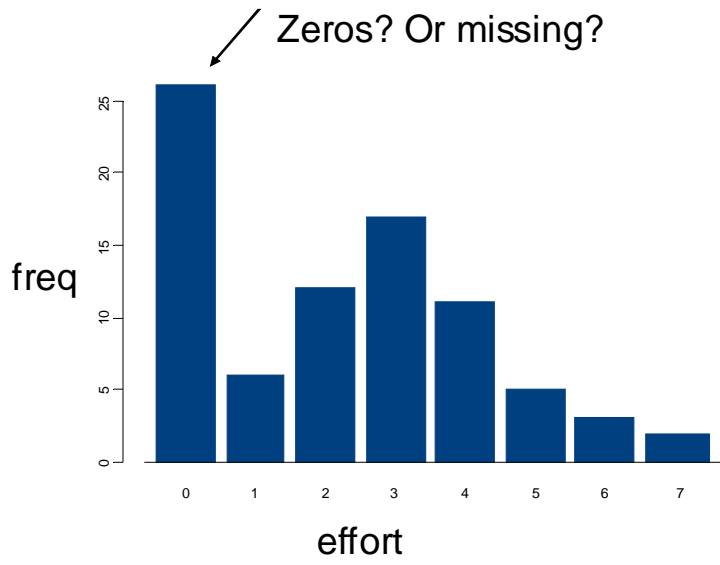Methods Working Group Meeting

St. George's, Grenada

June 2006

1) Every time you make a change to a database, enter your name, the date, and a description & justification of the change

2) Data must be proofread at least once. Enter name of person doing checking and date the data were checked

3) Every database should have embedded in it an explanation of every column heading, every symbol used, etc.

4) A person examining a database should be able to understand it without having to look elsewhere for explanations.
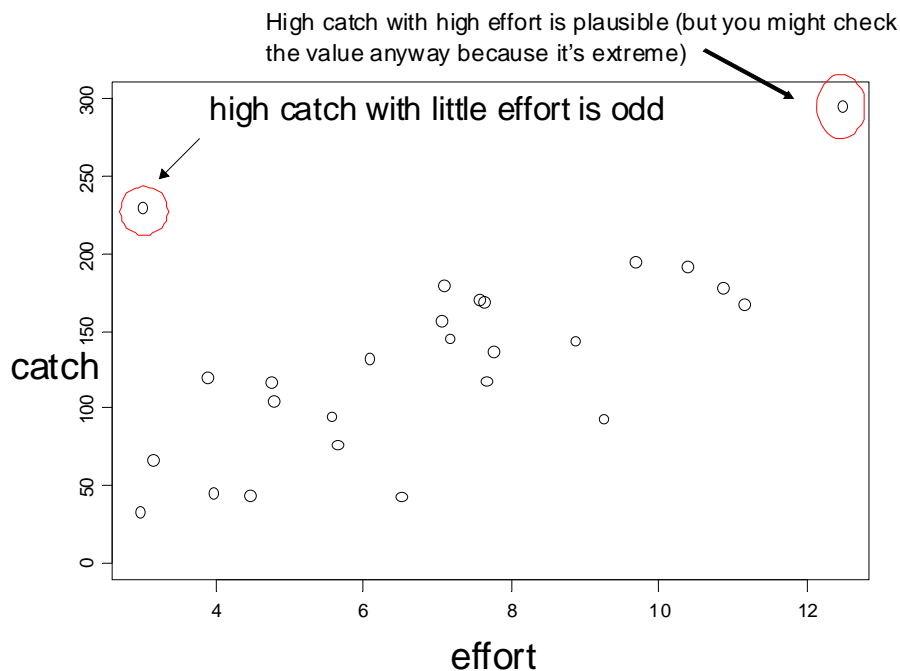
LOOK at the data.

• are there impossible values, e.g., length = 47g.2 cm

• have missing values accidentally been converted to
  zeros? (see end of this presentation for an
  example of how this can occur.)

Histograms

Freq

Unusual values are suspect

Length, cm

Zeros? Or missing?

freq

effort

Catch when effort = 0???

catch

effort

High catch with high effort is plausible (but you might check
the value anyway because it's extreme)

high catch with little effort is odd

catch

effort

Example: how missing values can be converted into zeros

In Excel, enter a column of numbers but leave a couple of cells blank. Use the average function to compute the sample average. (for example, =average(b3:b10)

Now highlight this column of numbers and copy it to a new page. Notice that the blank cells now have zeros. Use the average command to find the average. Note that it's different from what you got in the previous sheet. The reason is that Excel now thinks the missing cells are real data that should be included in the average.

The point is that errors can creep into databases in a variety of ways and sometimes it's not even a case of typing errors. So, you need to check your data. Also, the wise data analyst does not trust Excel completely.