Model Based Statistics in Biology

Chapter 2.3 Data collection, recording, and error checking

ReCap (Ch 1)
Quantities (Ch2)
2.1 Five part definition
2.2 Types of measurement scale
2.3 Data collection, recording, and error checking
2.4 Graphical and tabular display of data
Critique of graphs and tables
2.5 Ratio Scale Units
Base
Standard Multiples
Commonly used units in biology
2.6 Dimensions

on chalk board

Not here last time?
Course Outline
Name on roster
Questionnaire results

Discussion of Cards Lab:
Anybody come up with "wrong" rule that works?
In 1997 mutually exclusive pairs introduced ("test" cards), before going to multiple working hypotheses ("crucial" cards).
Ask for discussion of this, comparison of "test" and "crucial."
In 1998 crucial cards only.

Recap Chapter 1
Quantitative reasoning: Example of scallops, which combined stats and models
Biological reasoning will take the lead.
Stats will be a general analytic tool, not a list of tests.
Model based analysis will by the tool we use in statistical analysis.
Models (equations) are ideas about the relation of quantities.

Recap Chapter 2
Quantities: 5 components
Name
Symbol
Procedural Statement (write out and review: Could I use this recipe?)
Set of numbers collected into a vector
Units on a defined measurement scale
Type of scale: Nominal
Ordinal
Interval (or cardinal)
Ratio (or cardinal)

Today: Now that we have general definition of quantity, we will look at recording and displaying of quantities.

Wrap-up
Documentation: Notebooks. Label files, put information in files.
Public vs private files.
Tables for documentation, figures for discovery or communication of pattern.
Graphs and tables are an important part of quantitative work.

```
This material is presecriptive and not related to examples.
Add examples, retain in written notes
Discuss card lab data file (why it is so structured).
Discuss spreadsheet structure and labelling
Bring Tufte, 10 minute tour.
Could develop and revise a data sheet as exercise, or in lab.
```

Collecting data

The most important facet of good data collection is a clear statement of what is being measured, and why. If this is not clear, then the resulting data will be difficult, if not impossible, to analyze.

It may not always be possible to have complete procedural statement at the beginning of a project. However, it helps to have a preliminary protocol written out. This can be modified as needed during the early stages of the project.

Notebooks are useful devices for both lab and field work.

- Keep a record of what is being measured and why.
- This becomes a permanent record of what was done and why.
- Keep a record of how measurements were made (protocol)
- Notebooks should be complete and detailed, with no need to rely on memory later.
- Could someone else use the written protocol in the notebook?
- Keep record in notebook of any peculiarities of a particular measurement
- Notebooks have become a standard of valid science. Questions of authenticity of results are addressed by producing original notebooks.

Data forms are another useful device.

- These aid in organizing data.
- They also act as prompts, so that routine data that is part of the project (*e.g.* weather) is not overlooked or omitted from a record.
- Leave room on a form for notes (unusual conditions, possible errors in data or instruments, *etc*).
- Organize form for easy usage.
    - Separate rapidly changing from slowly changing variables.
    - For example,
        - slowly changing variables at top of sheet, in header.
        - rapidly changing variables close together, to reduce jumping across sheet
- Paper forms versus digital forms
    - Becoming easier to take digital forms, displayed on screen, into field.
    - Limits non=digital data capture (underwater, battery life) are decreasing.

Collecting data (continued)

Computer-assisted data capture is becoming increasingly important in both laboratory and field research.

Why: It reduces errors
It produces data with better resolution (especially, time).
How: Specialized equipment with standard output. *E.g.* scale with digital output.
Specialized equipment is often cheap, but not always flexible.
Programmable data loggers are more flexible.

- Portable devices (phone, laptops) are often used in this manner
- A wide variety of programmable data loggers have come on the market
- Mobile devices allow direct recording of data to digital format.
- Interface usually a forms program.
  - Error checking routines have obvious advantages
  - Error checking routines costly to produce, sometimes too rigid.
- Electronic pen interface
  - Advantage of flexibility and record like handwritten record
  - Disadvantage of errors in conversion to numeric data.

Instrumental data

A continuing trend is the increased use of electronic devices to obtain data, often with proxies to the quantity of interest. A simple example is measuring conductivity of water to obtain its salinity, defined as mass of dissolved matter to the mass of water in which it was dissolved. Here is a checklist applicable to instrumental data.

Correct for measurement error.
Choose between available estimators (Wald, 2SLS, GMM, LIML).
Understand the exclusion assumption.
Test for exclusion violations.
Detect weak instruments to avoid weak-instrument bias.
Deal with heterogeneous effects.
Work with compliance classes.
Estimate local/complier average treatment effects (LATE).
Detect violations of the monotonicity assumption.

Recording data

1. Separate original from derived data.
   - Notebooks
     - Original is a matter of record
     - Never erase the original. If an error is made, cross off the original entry with a light pencil mark, and then add the correction.
     - Derived data set has corrections. These are used in analysis
   - Digital files from handwritten sheets: Same principle applies.
     - Keep first digital copy, after removing any transcription errors.
     - Updates or errors identified during analysis are recorded in an updated file, not the first digital file. Why? Because corrections are sometimes made in error, the original was correct. Example of extreme values.
   - Digital files produced by automatic instruments: Same principle applies.
     - Keep first digital copy.
     - Updates and error corrections go in another copy, which is updated.

2. Keep a record of your data in a public (non-proprietary) format.

To an ever increasing degree, data are being stored in proprietary files. These are files produced by a particular program, and readily used by that program. Spreadsheets are the commonest example. These have embedded formatting and special characters that are not universally readable. In contrast, ‘public’ files (with the extension *.txt) can be read by any word processor, by any spreadsheet, and by any statistical package. They have only a few hidden characters such as tabs and line feeds. The key is that they can be opened by any program. This is important because cut and paste (as from a spreadsheet) only works if you can open the proprietary file (such as a spreadsheet file).

Proprietary data files are the default option for many programs. For example, the excel spreadsheet program produces *.xls and *.xlsx files, which are proprietary. These propriety files are quickly and accurately read by the program that created them because the formatting information (names of variables, location, *etc*) are stored along with the data. Some proprietary files (such as *.xls files) can be read by other programs. But in many cases proprietary files (such as *.mtw, *.sav) can be read by only a single program or software package. They are not universally readable. In the end, the only program that can be guaranteed to read a proprietary file is the program that generated the file.

Public data files are produced in a format that is not proprietary to any one program. Files exported as *.txt files are public – readable by any other program. These are also called ascii files because they consist of numbers and letters in ascii format (ascii translates strings of eight zeros and ones into numbers and letters). There is no formatting information in the file, other than tabs and line feeds. Commas are sometimes used instead of spaces or tabs to separate items, resulting in *.csv files (which can be read as *.txt files).

Recording data (continued)

So what's wrong with proprietary files?

More than one researcher has lost data because it was stored in a digital format that turned out to be unreadable by all but one program. An example is Ian Stenhouse, a Memorial graduate student who collected data on geese for the Scottish Research Council. He stored and worked with the data in package called "Super-Insight." Several years after completing the study he went back to re-analyze this data. No-one in North America had ever heard of "Super-Insight." Worse, the Scottish Research Council had discontinued the package, the package was no longer available, and so most of his data was lost. Luckily, he happened to have saved some of the data as ASCII files (*.txt files), which was able to use.

Lesson: Always keep a digital copy in publicly readable format. This means ascii files (*.txt or *.csv files). If you care about your data being accessible in the future, then save it in well documented files in public format.

The same applies to written documents. Wordperfect was once ubiquitous, until bought by a competitor and left to wither away. Wordperfect files are no longer usable, despite the promises of the competitor. Word is now ubiquitous, except in areas of science where mathematical notation is prevalent. Will *.doc and *.docx files be readable in 20 years? Perhaps. In the meantime, save your documents with markup language (*.rtf).

3. Use readable codes, not unreadable codes.

In the past, computer storage space was at a minimum. This forced the use of space efficient codes such as 1 = feeding, 2 = not feeding for the variable FS = foraging status. The phenomenal rise in cheap computer storage capacity has made space efficient coding unnecessary. Readability is now far more important than space efficiency. Readable codes eliminate coding errors that arise when codes are opaque and arbitrary, such as 1 = feeding. The limitation is that codes or variable names longer than 8 characters are cumbersome in many statistical packages. A reasonable solution is 'FORAGE' for the variable 'foraging status' and 'Yes' 'No' for each data entry. For the variable windspeed the natural name is 'WINDSP' natural codes are N NNE, *etc*, not 1, 2, 3, *etc*.

4. Distinguish true zero from missing data

How:

never code 'missing' as a zero. Save the numeral 0 for true zero.
code 'missing' as blank (space bar) if using fixed column format
code 'missing' as special value, if using free form or space delimited format.

5. Assign a symbol for wild or out of range values.

wild value = negative count if using data that can only be positive (counts)
wild value = letter if data can only be number.
wild value = number outside of permissible range.

Data structure

1. Use simple structure whenever possible: case (row) by variable (column).
   These are called "flat" files.
   More complicated formats arise when there are several types of records.
   Avoid the temptation of putting these all into the same file.
   Create one file for each type of record, so that all the records (rows) have exactly the same format.
   In a spreadsheet program, data with different structures are stored in different sheets (tabs at the bottom).
   To tie different types of records together, create a new file that contains the relevant data, and the variable used to tie the data.
   Data columns in public format files are separated by delimiters (spaces, tabs, or commas).

2. Document the structure. List all variables, the units of each variable, and all possible codes for each variable. These should appear in the public file. The data files that will be sent to you in this course will be documented at the end. Documentation at the top or beginning of the file used to interfere with reading the data, but with modern software, this is becoming less of a problem.

If you use a complex data structure, prepare an external catalogue (either in notebook, or in a separate digital file). This should include protocols, list of variables, all possible codes. For more about this topic search ‘metadata.’

Error checking

This is an important step in research.

Why: Errors are inevitable. People are not machines.
Errors are usually due to transcription error, sometimes due to machine error.
Computation errors, which were a problem before computers, have largely been eliminated.

How:
When transcribing to digital form, two people are 10 times better than one.
first person looks at original, other looks at digital version.
one person looking back and forth is far less effective.
Keep it simple. Use any device that reduces need for concentration on several things, or that requires memory. This reduces errors in checking.