Model Based Statistics in Biology

Chapter 1.3 Quantitative Reasoning

Chapter 1.1 The Role of Statistics in Science
Chapter 1.2 Model Based Statistics in Biology

Lab 1 Verbal Models (Guess the Process)

Chapter 1.3 Quantitative Reasoning
Two Examples
Reasons for Model-Based Approach

on chalk board

```
Not here last time?
 Roster: Names+Email
 Handout Syllabus

Questionnaire results
Yellow chalk
Lab 1
  Bring Cards
 Location: cf syllabus
```

ReCap. Model Based Statistics in Biology

One Goal of this course is to introduce you to effective ways of thinking quantitatively about biological phenomena.
A second goal is to give you practice you need to increase your skill and confidence in the application of quantitative methods.
A third goal is to develop your critical capacity, both for your own work and that of others.

It is NOT a course in mathematics. It *is* a course in applied mathematics.

Limited treatment of mathematical apparatus.
Emphasis will be on applying this apparatus.
Will work with data, summarizations of data (tables, graphs, statistics, models).
The emphasis will be on the practical application of quantitative methods to interesting questions and perplexing problems in biology.

It IS a course in how to think with biologically interesting quantities.

Today Examples of Quantitative Reasoning

Wrap-Up

In this course we will adopt a model based approach to statistics.
There are several advantages.

1 Statistics and modelling are closely related – stats are based on models.
2 Advantage of integration is carryover
3 We have a broader capacity to evaluate uncertainty in the analysis of biological data, than if we learn a series of tests.
4 Model approach leads to learning of concepts & principles, rather than collection of techniques

An example of quantitative reasoning

Statistics are traditionally taught separately from 'models'

Example of statistics: regression

Example of model: $N = N_o\, e^{r\, t}$ N = bacterial numbers

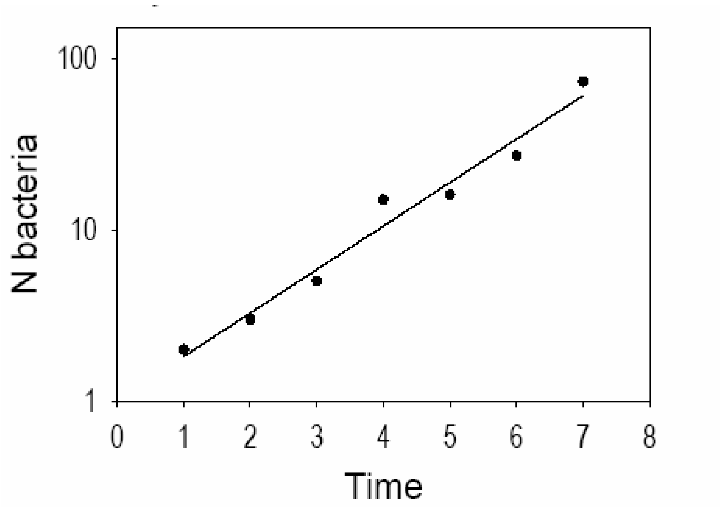
This is an equation for exponential growth in bacterial numbers N

This course will integrate both equations and statistics in reasoning about biological problems. Here is an example (move around triangle).

Start with data on bacterial numbers *N* at hourly intervals *t*.

t=hr	*N*
1	2
2	3
3	5
4	15
5	16
6	27
7	73

State <u>verbal</u> <u>model</u>



Then draw the graph.
The line thru the data (in yellow)
is the graphical model.

We define the rate of growth as:

$1/N\ dN/dt = r$

Here is the same idea in a different form

$N = N_o e^{r\, t}$

Taking the logarithm of both sides
of the equation we have:

$ln(N) = ln(N_o) + r\, t$

```
Construct triangle at each step
Write DATA
Write VERBAL MODEL,
     connect with line
Write GRAPHICAL MODEL
     connect with line
Write FORMAL MODEL
     connect with line
     to complete triangle
```

We use a statistical technique (regression) to estimate *r*, the slope of the line in the graph shown above.

The estimate of r = 0.006/hr or 0.6%/hour

We rewrite the equation, this time with the estimated value of *r*.

$N = N_o e^{0.6t}$

This is a <u>formal model</u>
It is an idea about the relation of scaled quantities, expressed in symbolic form.

Another example of quantitative reasoning

Draw data in white chalk.

State verbal model.

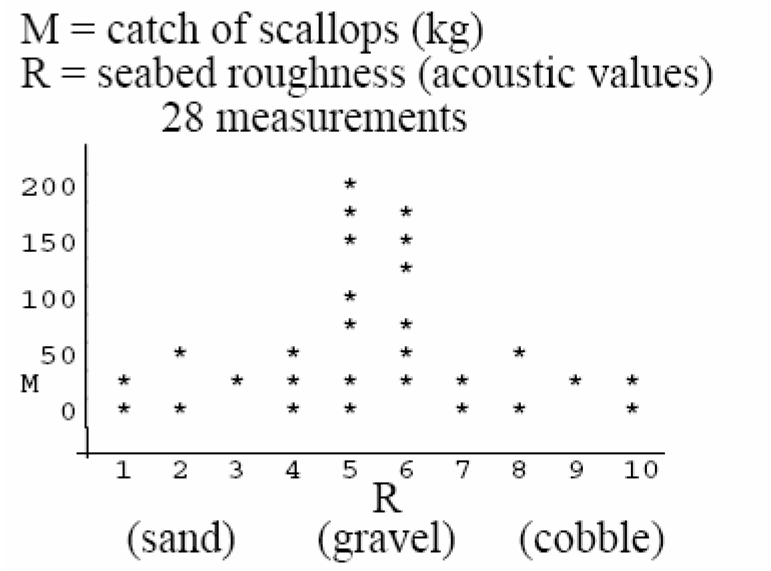
Catch is higher in gravel than in finer (sand) or coarser (cobble) substrates

Ask for graphical model.

Draw this model in yellow.

Then draw other models

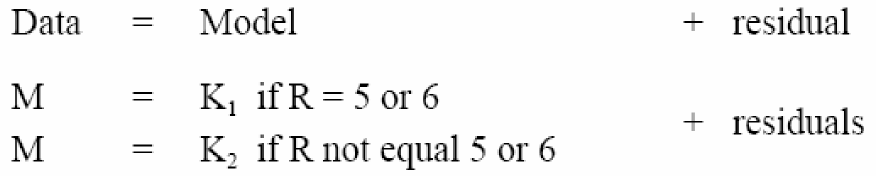
- housetop,
- two means, gravel or not
- normal curve



Write equation for two-mean model $M = K_1$ if R = 5 or 6 (gravel)
$M = K_2$ if R not equal 5 or 6

Let's review what we did. We began with verbal model, then moved to graphical model, and finally to a formal model. This illustrates quantitative reasoning.

Both models can be compared to data.



Data = Model + residual

$M = K_1$ if R = 5 or 6 + residuals
$M = K_2$ if R not equal 5 or 6

The two-mean model happens to be the statistical model used in a t-test, which is a test of whether two means differ by more than just chance levels.

This format (Data = Model + Residual) shows how we will integrate "modeling" with "statistics"

We are going to use models of data to summarize data, as in descriptive statistics. Common examples are means, standard deviations, and slopes in regression.

We are going to use models of data to make decisions in the face of uncertainty. This is called inferential statistics. Common examples are t-tests, ANOVA, regression.

Reasons for model based approach to statistics.

Statistics are traditionally taught separately from 'models'
Example of statistics: t-test
Example of model: $1/N\ dN/dt = r$

This course presents statistics from a modelling point of view. Why ?

REASON 1 Statistics and modelling are closely related. To illustrate:
Models underlie many statistical methods.
Statistics commonly used to develop and defend a model.

REASON 2 Advantage of model-based approach is carryover
Use what we know about biological models to improve statistical analysis
Use what we know about statistics to evaluate models

REASON 3 With the modelling approach we are no longer dependant on the machinery of hypothesis testing. We have a broader range of ways to evaluate uncertainty in the analysis of biological data.

REASON 4 Model approach leads to learning of concepts & principles, rather than memorizing a collection of techniques.

Statistics are traditionally taught as a series of techniques "101 Statistical Tests"
Mathematical biology is also traditionally taught as a series of case studies.
This is not the way the rest of biology is taught (if taught well).

Instead, learn concepts along with definitions.

```
To illustrate, scatter these 8 terms on board
     sporophyte,  metaphase,  blastula,  morula,
     gametophyte,  prophase,  telophase,  gastrula

Probably no one can still define these (I can't either)
But I bet you can still match them by concept
                                                  cell cycle
                                      alternation of generations
                               early development (embryogenesis)

Which terms pertain to the concept of "Cell cycle" ?
     (they'll get it right.  draw circle around these terms.)
Which terms pertain to  "Alternation of generations" ?
     (students call out terms, draw circle around these)
Which terms pertain to "embryogenesis" ?  etc.
```

Reasons for modeling approach.
REASON 4. Learning of concepts and principles. More about this.

This has several advantages

greater generality--can work from general principles
less amount of arbitrary material to learn
wider repertoire of methods because we don't have to 'name the test'
can get beyond text-book cases, which don't always fit our data.

But there are disadvantages

Principles are abstract, and so are harder to learn.
Need specific cases, often several, to grasp the concept.

Statistics are often taught as series of prescriptions, because it is less abstract and all we have to do is follow the recipe

A few prescriptions are highly useful, serving well in many cases (for example, t-test)

But while prescriptions are readily learned in a classroom setting (follow the recipe), they are not going to serve us well outside the classroom.

There may not be a textbook case that fits our data.
The search for a better recipe can be laborious and confusing.
We end up having to learn corrections (e.g. arcsin transform for % data).
The corrections may be a waste of time (e.g. arcsin transform).
Several prescriptions fit our data, but give different results
(e.g. ANOVA versus Kruskal Wallis test).
Standard prescriptions more limited than need be
(e.g. Chisquare tests vs logistic regression).
Prescriptions as they are taught are wrong (e.g. 'data must be normal.')
Key assumptions sometimes missing from the prescription
(e.g. homogeneity of slopes when using ANCOVA for statistical control).
When we focus on learning a series of tests we don't learn general principles.

Principles and general techniques (e.g. constructing the model, evaluating the residuals) will serve us best when it comes to

designing an experiment,
designing a survey,
evaluating statistical conclusions in the published literature

Model based statistics.

Four reasons

Learning prescriptions will not serve us well, beyond the classroom.

Learning principles and how to implement them on a computer will serve us better.

Methods and principles will be taught together in same course

- methods in the lab sections (including use of computer)
- principles in lectures

Goal is to learn to think with quantities, as well as to develop skill in applying specific methods.

Begin with a few basics (Part I):

    - the concept of a well defined, measurable Quantity
    - then equations (practice in Lab 2)
- then quantifying uncertainty (Part II):
    - frequency distributions
    - the peculiar logic of the null hypothesis
    - hypothesis testing
    - confidence intervals
- then statistical evaluation of quantitative relations (Part III).