Statistical Science

Chapter 1.3 Quantitative Reasoning

Chapter 1.1 The Role of Statistics in Science
Slide presentation on website
Chapter 1.2 Model Based Analysis and Inference

Inferential Cards Lab 1

Chapter 1.3 Quantitative Reasoning
Two Examples
Advantages of the Model-Based Approach

Not here last time?
Syllabus (course website}

Questionnaire results
Yellow chalk
Lab 1
Bring Cards
Location: see syllabus

on chalk board

ReCap. Model Based Analysis and Statistical Inference

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 science.

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

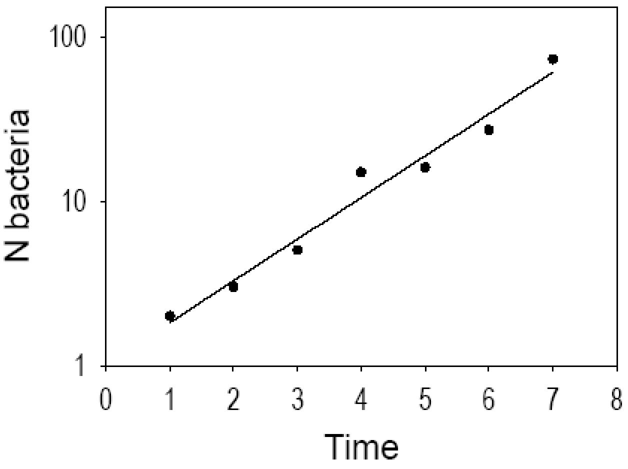
Example of model: $N = N_o\, e^{r\,t}$ N = bacterial numbers, t = time.

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.

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



Then draw the graph: a line (yellow) through the data (black dots on white or white dots on a chalk board)

This line is a graphical model.

We define the rate of growth in symbolic form as: $1/N\ dN/dt = r$

Here is the same idea in a different form $N = N_o e^{r\,t}$

This is a formal model without a specific value of *r*.

Taking the logarithm of both sides of the equation we have: $ln(N) = ln(N_o) + r\,t$

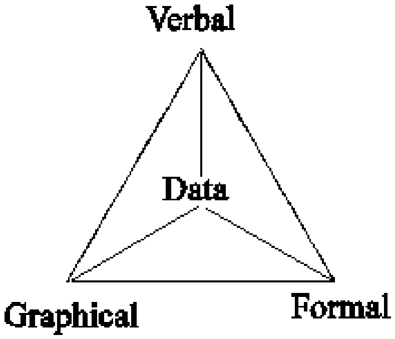
We use a statistical technique (regression) to estimate *r*, the slope of the line.

The regression estimate is 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 formal model with a specific value of *r*.

A formal model states an idea about the relation of measured quantities (*N* and *t*), expressed in symbolic form. Here is a diagram of what we just did.



```
        Construct triangle
Write DATA
Write VERBAL MODEL above DATA
   connect to DATA with line
Write GRAPHICAL MODEL
   connect to VERBAL MODEL by line
Write FORMAL MODEL
   connect to GRAPHICAL with a line
   to complete triangle
```

Another example of quantitative reasoning

Draw data in white chalk.

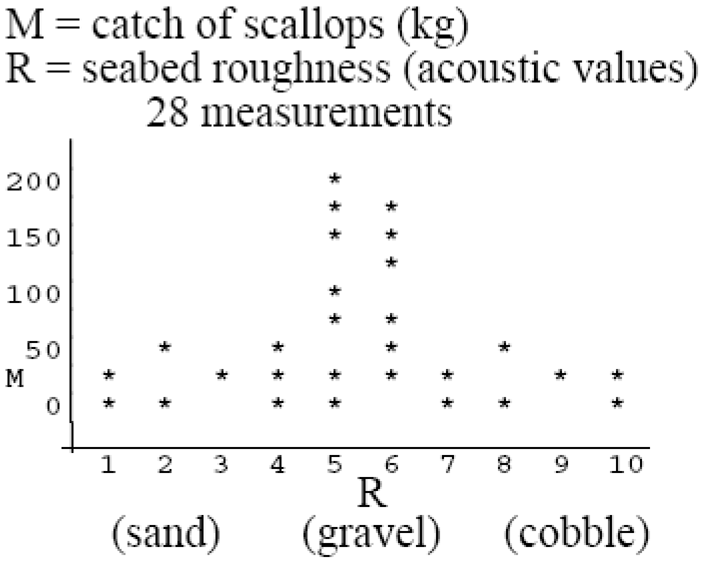
State verbal model.
Catch is higher in gravel than in finer (sand) or coarser (cobble) substrates

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) integrates "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 correlations.

We are going to use models to quantify uncertainty and draw conclusions from the data. This is called inferential statistics. It takes several forms.

1. Evidentialist. We infer from the data to a model of the data. We report likelihood ratios as a measure of the strength evidence calculated from our data.
2. Frequentist. We use likelihood ratios to infer from the data to a population.
3. Priorist. We use a likelihood ratio to infer from a prior to a posterior probability.

What then is a likelihood ratio? It is a measure of evidence for one model (say a regression line) relative to another (regression line with zero slope). In this course you will learn about and use both a measure of evidence (the likelihood ratio) and a measure of uncertainty (the p-value).

Advantages of model based approach to statistics in science.

Statistics are traditionally taught separately from 'models.'

Example of statistics: regression.
Example of model: $1/N \, dN/dt = r$

This course does not present statistics as a collection of tests. The course teaches you to write the statistical model based on a science question. Why this approach?

REASON 1 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 2 Model approach leads to learning of concepts & principles, rather than memorizing a collection of techniques.

REASON 3 Statistics and modelling are closely related. To illustrate:
Models underlie the most widely used statistical methods.
Statistical analysis is commonly used to develop and defend a model.

REASON 4 A model-based approach has the advantage of carryover.
We use what we know about biological models to improve statistical analysis
We use what we know about statistics to evaluate models.

Statistics are traditionally taught as a series of techniques "101 Statistical Tests"
This is the way we identify birds from a field guide. Or how we identify plants from a key. But it is not a description of how we do science.

Instead, we learn concepts that connect terms.

```
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 2. Learning concepts and principles. More about this.

Learning to write the model has several advantages
Work from general principles of survey and experimental design.
Reduction in arbitrary material to learn.
Extends analysis beyond the list of named tests.
Extends analysis beyond text-book cases, which don't always fit our data.

But there are disadvantages
Principles are abstract, and so are harder to learn.
We need specific cases, often several, to grasp a concept.

Statistics are often taught as series of prescriptions, because this 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 are more limited than writing the model
(*e.g.* Chisquare tests vs logistic regression model).
Some widely held beliefs are wrong (*e.g.* 'data must be normal').
Key assumptions are 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 – Course goals and course structure

Learning principles and how to implement them on a computer will serve us better than learning prescriptions. Learning prescriptions will not serve us well, beyond the classroom.

Methods and principles will be taught together in same course
-methods in the lab sections (including use of computer),
-principles in lectures.

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

The course begins with the basics of data collection and explanatory models (Part I):
the concept of a well defined, measurable Quantity
working with models in symbolic form (practice in Lab 2)
Part II covers quantitative measures of evidence and uncertainty:
Frequency distributions
Likelihood ratios
Type I and II error
Bayes' rule and confidence intervals
Part III covers statistical analysis for a single explanatory variable.
Linear regression
t-tests
One-way ANOVA
Part IV covers statistical analysis for multiple explanatory variables.
Multiple regression
Multiway ANOVA
ANCOVA
Designs that do not fit into these categories.