Model Based Statistics in Biology

Chapter 1.2 Model Based Statistics in Biology

Chapter 1.1 The Role of Statistics in Science
- The Dark Side of Statistics
- Statistics are Balderdash - Get rid of them!
- Statistics are like taxes - inevitable.
- Hypothesis testing is statistical flotsam
- Model Based Statistics
- Discarding the flotsam and jetsam
- Learning model-based statistics

Slide Show Ch1_1.ppt

Today

Chapter 1.2: Model Based Statistics in Biology
- Perplexing questions
- Uncertainty
- Verbal, Graphical, and Formal Models
- Role of models in statistics
- Course Structure
- Opinions

on chalk board

```
First Day
 Roster: Names+Email
 Handout Syllabus
 Questionnaire
Yellow chalk
Lab 1
 Bring Cards
 Location: cf syllabus
```

Wrap-up. Model Based Statistics in Biology

One goal of this course is to introduce you to <u>effective ways of thinking</u> quantitatively about biological phenomena.

A <u>second goal</u> is to give you practice you need to <u>increase your skill and confidence</u> in the application of quantitative methods.

A <u>third</u> goal is to <u>develop your critical capacity</u>, both for your own work and that of others.

This is NOT a course in mathematics, *per se*.
- Emphasis will be on applying mathematics, not on the mathematical 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.

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

Chapter 1.1 The Role of Statistics in Science (on the course website).

The *only* power point presentation in this course. Why?

In his essay "The Cognitive Style of PowerPoint", Edward Tufte criticizes many properties and uses of powerpoint software:

It is used to guide and to reassure a presenter, rather than to enlighten the audience;

It has unhelpfully simplistic tables and charts, resulting from the low resolution of early computer displays;

The outliner causes ideas to be arranged in an unnecessarily deep hierarchy, itself subverted by the need to restate the hierarchy on each slide;

Enforcement of the audience's linear progression through that hierarchy (whereas with handouts, readers could browse and relate items at their leisure);

Poor typography and chart layout, from presenters who are poor designers and who use poorly designed templates and default settings (in particular, difficulty in using scientific notation);

Simplistic thinking, from ideas being squashed into bulleted lists, and stories with beginning, middle, and end being turned into a collection of disparate, loosely disguised points. This may present an image of objectivity and neutrality that people associate with science, technology, and "bullet points".

Tufte argues that the most effective way of presenting information in a technical setting, such as an academic seminar or a meeting of industry experts, is by distributing a brief written report that can be read by all participants in the first 5 to 10 minutes of the meeting.

In this course you have the handouts, they are on the web, for each lecture. At each meeting of the course, I will provide a narrative that covers the points. When it comes to learning statistical procedures, I'll be providing step by step narratives of those procedures, something I have been doing in this course for decades. For a recent example of the same narrative approach in science see http://www.khanacademy.org/

Instead of the khanacademy moving hand with narrative voice, you're going to see me develop concepts, show their relation to each other, and narrate the computational steps on a large chalkboard.

Chapter 1.2: Model Based Statistics in Biology (today)

A pair of challenging questions: How many species are there and how fast are species going extinct?

I'll tell you why I think these are important questions, especially to a biologist.
How many species are there ? thousands? hundreds of thousands? millions? tens of millions? What is the order of magnitude of the number of species?
This is a question of <u>intrinsic</u> interest to biologists.
One of the most remarkable things about living organisms is that they come in such a diversity of species. This is one of the attractions of biology.
Why should there be so many species? It is especially striking if we go to the tropics, or or take a dredge haul from the deep sea.
Narrative of seeing variety of specimens from deep sea collections at Woods Hole. Narrative of intertidal research in Panama.
The flip side of the qustion is, how come there are so many fewer species (and much larger populations) at higher latitudes ?
We would like to understand what processes lead to proliferation of species seen in some regions, and not in others.

It is also a question with an <u>ethical</u> side to it.
Do humans have the right to displace other species from this planet ?
Are we in fact causing extinctions ? If so, how fast ?
Which policies lead to extinction ?

It is also a question of <u>practical</u> value. As many people know, plant diversity is a storehouse of secondary compounds, used for spices and medicinals. Loss of diversity diminishes the number of compounds available.

Question of number of species has intrinsic interest to a biologist.
Question of rate of extinction is of practical importance.
    Drugs, alternative varieties for crops, most food comes from a few varieties.
Question of rate of extinction is a matter of ethics.
    What right to squeeze out the other species on this planet ?

How many species are there ? Any thoughts at this point on how we could obtain an estimate ? (responses usually involve some form of survey, which requires a model).

According to World Conservation Union, forests cover 33 x 10^6 km^2

Of the 80,000 to 10,000 species of trees once on the planet, nearly 100 have gone extinct, and more than 8,600 are headed that way (8 Sept 1998 Globe and Mail)

A second challenging and important question:

Does the risk of cancer depend on the number of cigarettes smoked ?

We can do a survey on cigarette smoking, and measure whether the percent of people with tumors increases with number of cigarettes smoked. While there is an association, we have no evidence that smoking caused the increase. To address this question we can undertake and experiment where we force rats in a lab to breath cigarette smoke. If the experiment is well controlled, we have better evidence that cigarette smoking increases the risk of cancer. Such experiments were common in the 1950s and 1960s, as the evidence mounted that cigarette smoking was responsible for an epidemic in lung cancer. Some experiments showed clear effects (especially those with large numbers of rats with similar genetic make-up). Other experiments failed to show clear effects (especially those with few rats from different sources). The differences in outcome were the source of considerable controversy, because the collision of health concerns with the fact that many people depended on the cigarette industry for their livelihood. We can’t throw large numbers of people out of work unless we are certain that cigarette smoking causes cancer. But the results seemed very uncertain because some experiments showed an increase in risk, while others did not. If we quantify the uncertainty associated with each experiment, we find that our measure of uncertainty is far higher for the experiments with few rats. When we quantify uncertainty, we conclude that the risk of cancer increases for rats exposed to cigarette smoke under experimentally controlled conditions. We change our conclusion when we take into account the uncertainty of each experiment.

A third perplexing and important question:

What is the structure of a protein that plays a role in the spread of certain cancers?
Science 261:844 (13 August 1993)

Structural information should help guide efforts to disarm the protein and reduce the cancer's threat. Let's imagine you have a fully automated lab that can generate 400,000 bases worth of DNA sequence data per week. This of course consists of strings of A,s G's, T's and C's. And we know that only about 2% of this consists of protein coding information. The rest is "junk" much of it silent. Or it consists of regulatory regions, which are not of much use in identifying protein structure.

One solution is to write some rules (TAA = stop codon) then use computer to pick out protein coding segments. Pick out the signl from the noise, using simple rules and much computation.

A combination of simplification (known rules) and much computation, to pick out structure from noise.

A fourth challenging and interesting question, related to function of the protein and how to disarm it, is the secondary structure of the protein. How do the strands twist and fold ? When do you get a helix ? When do you get a loopy strand ? Can THIS be predicted from the one dimensional sequence of A G T C ?

Any ideas on solving this problem ??

One solution is to program simple predictive rules into computer, then have computer check accuracy of predictions. The computer "learns" by comparing predictions of structure (from sequences) to known structure. This requires simple rules or "models" of how sequence translates into structure. Establish a set of these. Then evaluate several pedictions against known structure. Then choose "best" set of rules, perhaps with some new rules. Keep cycling in this fashion to distinguish coding sequences from noise.

Another solution is to calculate associations between sequences and structure of the protein. Take the sequence that is most often associated with a structure. Use this as hypothesis. Then test by other means.

A fifth challenging question: How many fish in the sea?
The question is of considerable importance to people whose livelihoods depend on the sea. By extension it is important in areas where the fishery is an important part of the local economy. The largest lay-off in Canadian history occurred in July 1992, when the fishery was closed and 30,000 people were out of work. The stage for economic disaster was set in 1985, when warnings from inshore fishermen were ignored. The potential for disaster grew when the number of fish was overestimated, from 1985 through 1987. The disaster was then played out from 1989 until the fishery was closed in 1992.

Uncertainty. How do we evaluate uncertainty?

How many species and what is rate of extinction?
Does cigarette smoking cause cancer?
How many fish in the sea?

The number of species currently extent, and the rate of extinction, are both <u>estimates</u>. The true value is not known. The true value lies either above this estimate (half the time) or below this estimate (half the time). So an environmental group comes to you and asks:

How far is this estimate from the true value ?

In other words: How much weight should be put on this estimate when making plans?

list sources of uncertainty (on board)
    true value of number of species (20 million ? 50 ?)
true value of extinction rate (0.5 %/yr   1%/yr ?)
true value of measured quantities used in making estimate ?
    (e.g. species per unit area)
    (e.g. total area of each type of habitat)
other sources of uncertainty besides biological uncertainty
  what are effects of a particular policy on land use ?
  will this policy continue, or is it likely to change ?
  what are the competing pressures ? (agriculture, etc)

Cigarette example. We saw that uncertainty varied among experiments, depending on whether we had many rats (more certain result) or few rats (less certain result). We started with one conclusion about the experiment: exposure to cigarette smoke under experimental conditions does not have clear effect on risk of lung tumors.
We changed out conclusion when we took into account the degree of uncertainty in each experiment.

Another example. Uncertainty in estimate of fish stock size. Problems created when an estimate (with high uncertainty) is presented as a single number, implying a high degree of certainty. If the true value could be half or twice the estimate, then this should be communicated along with the estimate.

Uncertainty has many sources

Factors that cannot be controlled by manipulation:
-Effects of weather, change in climate,
-Individual variation in epidemiological or drug effectiveness studies
Causation at multiple spatial scales :
Weather more uncertain at weeks than tomorrow,
even more uncertain next year
Local agricultural practices less reliable regionally than locally

Uncertainty often high in biological systems, especially compared to inanimate systems, as in chemistry and physics.

Multiple causation:
simultaneous causes: two or three policies acting at once
each will have different effects on rate of species loss.
how to separate these effects ?
Example: effect of farm roads versus highways on tree diversity
species loss depends on access to tropical forest
species loss depends on kind of access
interacting causes: effects of several practices may interact.
farm roads + highways lead to rapidly clearing.

Shifting causation: effect of changes in land use practices on species diversity

Unknown and unidentified causes (many diseases)
undiagnosed disease could change local estimates of
species per unit area

Uncertainty arises when making an estimate.
It is impossible to make complete survey all of the globe, and count all species
It is impossible even to count completely in a single watershed of say
20,000 km^2.

Uncertainty. Individual versus public evaluation of uncertainty. In biological systems, we are dealing with complexity. We are dealing with uncertainty. How do we handle this?

In our everyday lives we use heuristic rules to act in the face of uncertainty. The result is that people make different estimates, weight things differently, and arrive at different decisions.

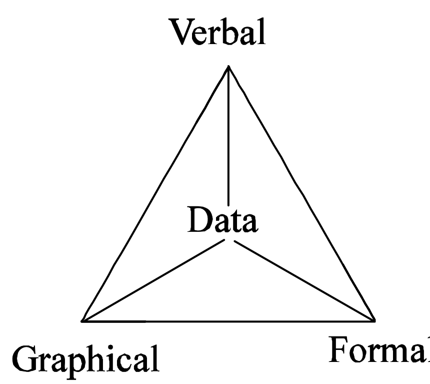
Decisions that affect many people differ from those that we make for ourselves.

1. Rates of extinction. Our own decisions unlikely to have much effect on extinction, but decisions made by many people will have an effect, sometimes large.
2. Cigarette smoking. Our own decision affects our health, and those breathing the same air. But decisions made by many people will place a substantial burden on the health care system.
3. Fish numbers. Decision by any one person will have little effect on fish numbers. But decisions by many people, and by governments (through funding for economic development) will eventually drive any species to commercial extinction.

How do we make decisions when those decisions affect many people, including people we do not know or will never meet? Statistics were developed to address this question. To make decisions that affect many people, or that involve public funds or public policy, we need to make decisions that are rational (based on logic) and that are based on evidence. Statistics combine the logic of math with evidence to arrive at estimates based on clearly stated assumptions.

Verbal, graphical, and formal models

Inescapable need to make assumptions, devise models.
Verbal simplifications in order to understand a situation.
Model of uncertainty in order to make statements in face of uncertainty.
Statistical methods are based on models, they are not just recipes.



Typical route: 1st, data
2nd, verbal (simplification right away)
3rd, picture (is worth 10^3 words)
4th, make calculations (requires formal expression)

Formal expression (equations) are important because they are the ones that allow us to make calculations, such as 1000 species per hectare of tropical forest

Quantitative reasoning entails repeated cycling from formal, to data, to picture, to formal

It is not arcane and inaccessible. It is a way of reasoning about biological phenomena.

Verbal, graphical, and formal models

Formal models have a number of advantages, but they have a price.
They are not as familiar as verbal or graphical models
Happily, skill and confidence in using model based statistics can be gained by practice and frequent contact with real data.

Some characteristics of quantitative methods. (D.S. Riggs, Chapter 1)
Brevity of expression.
Right or wrong, given the assumptions.
Good and bad practices, as in any human activity
Good practices lead to effective action,
bad practices lead to confusion and waste of time.
Examples of bad practice in biology not hard to find
confusing correlation with causation
poor reporting of methods --> irreproducible results
Affects us all--health and medicine.
--environment and ecology.

Role of Models in Statistics

Statistics, believe it or not, are based on models. Statistics are simplifications of data, based on models. You have all encountered mean values. A mean value is a "model" because it is not a measurement, it instead represents a large collection of values.

Statistics are a formal (ie mathematical) way of
-discovering generalizations (models)

Extinction = function (land use, response, ___, ___)
$Y = f(X)$
Effect = function(Cause)
Expected = function(Observed)

-evaluating generalizations (models)

Data	=	Model	+	Residual
Y	=	βX	+	ε
Observed	=	Expected	+	Residual

Statistical methods are used:
to interpret observations,
to evaluate complex evidence
to disentangle multiple causation
to increase efficiency in carrying out experiments
to evaluate reliability. How certain can we be ?
to evaluate generality of results.

Course Structure – Model based statistics

Learning occurs best with frequent contact with material.
Learn most by active contact –weekly quizzes, + assignment + lab
These activities weighted by course evaluation of students.
Curve of work in this course, for undergrads, for grad students.

Some opinions about statistics

"Lies, damn lies, and statistics."

Variously attributed to Benjamin Disraeli, Mark Twain, etc.
In fact due to Leonard Courtney, from a speech published in the Journal of the Royal Statistics Society.

There is no question that it is possible to lie with statistics. There is even a book entitled: "How to Lie with Statistics" Useful to learn some tricks of statistical presentation, in order to recognize the commoner forms of deception. For example, presenting mean values for a highly skewed set of numbers will produce a misleadingly high value. Better to present also the median (half above and half below). Example of average number of offspring. In many invertebrate species a small number of individuals have no offspring, a very small number have many. Such as oysters. Hence average is not representative. Removal of a few large individuals may appear to have little effect on average, when in fact it will have a very large effect.

"If your experiment needs statistics, you ought to have done a better experiment"
- Ernest Rutherford

1. Not all systems can be manipulated.
   Ethical considerations (medical research)
   Large scale environmental effects (weather)
2. Statistics are an important way of doing a better experiment, as we will see.

"Statistics are just frosting on the cake" (ethologist at Queens University).

meaning that statistics are just decoration, to make it look "scientific")
But: Fisher's Fundamental Theorem. This is one of the most important ideas in evolutionary biology. It is basically a statistical concept: the rate of change in gene frequency depends on the amount of additive genetic variance
But: No funding in many areas of biology without good statistical design.
But: No funding in many areas of health sciences without good statistical design.

More opinions about statistics

"Not an elegant solution"

Mathematicians prefer elegant proofs, for good reason: generality

But elegant math is not always capable of solving problems.

Often what is called a numerical solution has to be used.

For example, the weather can be predicted from fluid dynamics equations, but these equations cannot be solved analytically

They can be solved by massive computation.

This is exactly what is done.

Numerical methods, which are not "elegant" are increasingly used in statistics, even though mathematically trained statisticians sometimes look down on this.

We will use numerical methods in this course, to calculate p-values when the underlying theoretical distribution is unknown