Statistical Science

Chapter 1.2 Model Based Analysis and Statistical Inference

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 analysis and statistical inference

Slide presentation Ch1_1.ppt

Today

Chapter 1.2: Model-based Analysis and Statistical Inference
- Perplexing questions
- Uncertainty
- Verbal, Graphical, and Formal Models
- Role of models in statistics
- Course Structure
- Opinions
- Looking Ahead

on chalk board

Wrap-up. Analysis and Inference in Statistical Science

One goal of this course is to introduce you to effective ways of thinking quantitatively in the biological and environmental sciences.

A second goal is to show you how to write a model to set up your own analysis of scientific data.

A third goal is to give you the practice you need to increase your skill and confidence in setting up and interpreting the results of an analysis of data.

A fourth goal is to develop your critical capacity, both for your own work and that of others.

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

It is NOT a course in statistics.
- The emphasis will be on applying mathematics, not on the mathematical apparatus.
- The focus will be on data, on good practice summarization (tables, graphs,), on inference with a statistical model, and on interpreting analytic results.
- The emphasis will be on the practical application of quantitative methods to interesting questions and perplexing problems in biology.

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:

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".

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

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

It enforces on an audience a linear progression through that hierarchy. In contrast, with handouts, readers could browse and relate items at their leisure).

It results in poor typography and chart layout from poorly designed templates and default settings. Scientific notation is a particular problem—symbolic notation is divorced from meaning and reference to concrete examples.

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, as detailed lecture notes. At each meeting of the course, I will provide a narrative that covers key 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 Khan Academy 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 Analysis in Statistical Inference (today)

Challenging questions

Here is a pair of challenging questions and important questions:

How many species are there? 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.
I distinctly remember the unexpected variety of specimens from a single deep sea dredge haul, in a lab at Woods Hole. During a post-doc in Panama, I felt the same astonishment at the diversity of intertidal benthic organisms in Panama, compared to similar habitats in Massachusetts.
The flip side of the question is, how come there are so many fewer species (and much larger populations) at higher latitudes on the land?
We would like to understand what processes lead to proliferation of species at low latitudes on land, or at greater depths in the ocean.

It is also a question with an <u>ethical</u> side to it.
Does our species have a responsibility to conserve rare species?
Are we in fact causing extinctions ? If so, how fast ?
Which policies lead to species 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.

The question of number of species has intrinsic interest to a biologist.
The question of rate of extinction is of practical importance.
Drugs, alternative varieties for crops, most food comes from a few varieties.
The question of rate of extinction is a matter of ethics.
What responsibility do we have to reverse anthopogenic loss of species diversity ?

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 100,000 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 ?

This is no longer a challenging question but it once was.

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, until 1970 we had no direct evidence that smoking caused the increase. To address this question we can undertake an 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 signal from the noise, using simple rules and much computation.

We use a combination of simplification (known rules) and substantial computation, to pick out structure from noise.

A fourth challenging and interesting question is the secondary structure of the protein, which determines function of the protein and insight in to how to disarm it,.

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. Once a set of these are established, predictions can be made against known structure. From these we identify those with the best evidence (most likely) and perhaps add some new rules. We keep cycling in this fashion to distinguish coding sequences from noise.

Another solution is to calculate associations between sequences and structure of the protein. Then take the sequence that is most often associated with a structure as a hypothesis to be challenged with new data.

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 Newfoundland 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 from 1985 through 1987 when increasing uncertainty in estimate of stock size led to overestimation, visible only in retrospect. The disaster played out from 1989 until the fishery was closed in 1992.

A sixth challenging question: Is the planet warming due to human activities?
This was a challenging question in the 20th century because of the variability in the data and because of economically motivated misinformation. The evidence grew stronger each year. The word "likely" appeared in 1995. This refers to likelihood, a statistical measure of the strength of evidence.
Conclusions of the IPCC (Intergovernmental Panel on Climate Change)
1990: "…emissions resulting from human activities are substantially increasing atmospheric concentrations of greenhouse gases… These increases will enhance the greenhouse effect, resulting on average in an additional warming of the Earth's surface"
1995: "Most of these studies have detected a significant change and show that the observed warming trend is <u>unlikely</u> to be entirely natural in origin"
2001: "There is new and stronger evidence that most of the warming observed over the last 50 years is attributable to human activities."
2007: "Warming of the climate system is unequivocal. Most of the observed increase in global average temperatures since the mid-20th century is <u>very likely</u> due to the observed increase in anthropogenic greenhouse gas concentrations."
2013: "Warming of the climate system is unequivocal. It is <u>extremely likely</u> that human influence has been the dominant cause of the observed warming since the mid-20th century.
2022: "the chance of exceeding tipping points, such as sea level rise due to collapsing ice sheets or ocean circulation changes, cannot be excluded from future planning. Their likelihood increases with greater warming."

In this course you will learn more about likelihood as a measure of the strength of evidence. It is not a part of standard courses in statistics, but it is readily understood. If I say the odds are 50:50 on the flip of a coin, that is the same as saying heads are just as likely as tails.

Uncertainty.

All of these questions posed above arise from uncertainty on estimates and on cause.

How many species and what is rate of extinction?	Estimation uncertainty
Does cigarette smoking cause cancer?	Estimation and causal uncertainty
How many fish in the sea?	Estimation uncertainty
Is climate change due to human activities?	Estimation and causal uncertainty

In every case there is estimation uncertainty. The true value is not known. The true value lies either above the estimate (half the time we hope) or below this estimate (half the time we hope).

Species extinction example.
An environmental group comes to you and asks:
How far from the true value is this estimate of extinction that we have?
In other words: How much weight should be put on this estimate when advocating policy?

Poll for a list of sources of uncertainty, post to chalkboard)
True value of number of species = 20 million ? 50 ?
Poll for a list of extinction rates (0.5 %/yr 1%/yr ?)
How do we quantify the question?
species per unit area ?
total area of each type of habitat? In an ecosystem?
something else?
Poll for 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 ? (*e.g*. agriculture)

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 a clear effect on risk of lung tumors.
We changed our conclusion when we took into account the degree of uncertainty in each experiment.

Fish example. Uncertainty in estimate of fish stock size. Problems arise 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

Measurement error is always present. In some cases it can be reduce. In others, it cannot be reduced mechanically. It can only be reduced by increased evidence.

Sampling is usually necessary.

A full census, to get the true value of a rate, is usually impossible to make.

It is impossible to count all species completely in a single watershed, let alone an entire river system.

Sources of variation cannot always be controlled by manipulation:

-Individual variation in epidemiological or drug effectiveness studies

-Effects of weather, change in climate

-List another example ___________________________

Causation operates at multiple spatial scales.

-Weather is more uncertain at weeks than day by day, even more uncertain next month

-Species numbers do not increase in a one to one fashion with area surveyed.

-List another example __________________________

-Multiple causes may be present:

Simultaneous causes.

Several processes acting at once will each have different effects on rate of species loss. Separating these effects is often difficult.

Example: effect of farm roads versus highways on tree diversity species loss depends on access to tropical forest

List another example _____________________

Interacting causes: effects of several practices may interact.

Mortality rates change in response to global warming.

List another example _______________________

Shifting causation.

Changes in land use practices alter species diversity

List another example ________________________

Unknown and unidentified causes.

For example the spread of a disease

List another example _______________________

Uncertainty.

Individual versus public evaluation of uncertainty.

In science, 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. One such rule is to "update our priors." Here is an example.

What is your prior probability that you will do well in this course?
-It might be based on utility: how useful is this course to you?
-It might also be based on a prior statistics course. (I hope not!)
-It might be based on testimony from students who have taken the course.
-There might be other factors, known best by you.

During this course you will see evidence -weekly lab, several assignments, weekly quiz.

And so there will be frequent opportunity to update your priors. Here is the formula:

Updated (posterior) probability = Evidence x Prior Probability

For marked pieces of work (Scored as %) the Evidence is a likelihood ratio LR.

LR = % / (1−%)

In principle we should multiply these likelihood ratios to obtain the final score. However, for consistency with the way things are done, we use the sum the percentages according to weights on the syllabus.

From this example we can see that people will weight things differently, and they can arrive at different conclusions and decisions.

Public evaluation of uncertainty -- Confirmation bias

People tend to give more weight to evidence that confirms their beliefs. They tend to give less weight to evidence contrary to their beliefs. This is called confirmation bias.

What happens if we update with confirmation bias? We don't expect this to be an issue for the example above. But in everyday life there is considerable room for confirmation bias. An obvious example is climate change denial. As scientists, we have the responsibility to lean in with evidence, as best we can.

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

1. Rates of extinction. Our own decisions are unlikely to have much effect on extinction, but decisions made by many people do have an effect, sometimes large.
2. Cigarette smoking. Our own decision affects our health, and those breathing the same air. Decisions made by many people will determine the 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) can eventually drive any species to commercial extinction.

Public evaluation of uncertainty – The role of statistical science

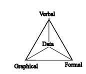
How do we make collective decisions, which affect many people, including people we do not know or will never meet? Statistics were developed to address this question.

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 evidence with the logic of math to arrive at estimates based on clearly stated assumptions.

Verbal, Graphical, and Formal models

The need to make assumptions cannot be escaped.
This includes <u>statistical model</u>s, such as the normal distribution.
 We need a statistical model in order to make statements of evidence and uncertainty.
 Statistical science is based on models, it is not a collection of recipes for statistical tests.



This includes a <u>structural model</u> that relates one (or more) response variables to one (or more) response explanatory variables.
Forming the structural model begins with a <u>verbal model</u>.
 This is a statement in words, of the relation of response to explanatory variables. It can be formed from prior data or research. It can be formed from theory.
The verbal model is converted to a <u>graphical model</u> that displays its relation to data.
Variables are assigned symobols in order to write a <u>formal model.</u>

<u>Typical route</u>: 1st, data
 2nd, verbal (simplification right away)
 3rd, picture (is worth 10^3 words)
 4th, symbolic form (to make calculations)

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.

<u>Advantages of formal models come with a price.</u>
Formal expression (equations) are important because they allow us to make calculations about biological phenomena, such as rates of species loss per hectare of tropical forest.
The price is that 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.

Verbal, Graphical, and Formal models

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 are not hard to find
confusing correlation with causation
poor reporting of methods --> irreproducible results
using a measure of uncertainty (the p-value) as a measure of evidence.
Bad practice affects us all: in medical and health science research
in biological and environmental science

Role of Models in Statistical Science

Statistical inference and analysis require models. We encounter them daily. For example, a mean value is a "model." It is not a measurement, it instead represents a collection of measurements, with equal weight to each measurement.

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

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

-evaluating generalizations via 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 infer beyond the data at hand
to evaluate generality of results.

Course Structure

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

Some opinions about statistics

"Lies, damn lies, and statistics."

Variously attributed to Benjamin Disraeli, Mark Twain, etc.

In fact it is 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" This book displays common forms of deceptive statistical presentation. It is useful in recognizing these, either intentional or inadvertent.

An example is presenting mean values for a highly skewed set of numbers – this will produce a misleadingly high value. It is better to present also the median (half of observations above and half below).

Another example is average number of offspring. In many economically important species (*e.g.* lobsters, codfish) number of offspring depends on body size. Thus the small number of large animals make a disproportionate contribution to offspring numbers. Intuitively, from the mean/ adult, we expect that removing a few large lobsters will have little effect on total number of offspring from a population, when in fact it will have a very large effect.

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

- Attributed to 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, Canada).

The implication is obvious: statistics are just decoration, to make research look "scientific."

Counter example 1: 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

Counter example 2: Funding in many areas of the biology and environmental science requires good statistical design. This is particularly true in medical research. It is often true in environmental impact assessment.

Counter example 3: In the first decade of this century the US National Park Service required all of its monitoring programs to justify funding by showing the program was not “doomed to failure.” In other words, that monitoring program through high variability was unable to produce statistically significant (5% false positive) results at a set rate of 20% false negatives.