Statistical Science

Chapter 1.2 Model Based Statistical Inference and Analysis

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
- Perplexing questions
- Uncertainty
- Verbal, Graphical, and Formal Models
- Role of models in statistics
- Course Structure
- Opinions
- Looking Ahead

on chalk board

```
First Day
 Introductions
 Syllabus
 Questionnaire
Yellow chalk
Lab 1
```

Wrap-up. Model Based Statistical Inference and Analysis

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 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 NOT a course in mathematics.

The emphasis will be on applying mathematics, not on the mathematical apparatus.

The focus will be on data, and good practice summarization (tables, graphs,), inference with a statistical model, and interpreting analytic results.

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:

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 audieance 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, 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 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 Statistical Inference and Analysis (today)

Challenging questions

Here is 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 intrinsic 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 ethical 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 practical 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 33×10^6 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 ?

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 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 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 when the number of fish was overestimated, from 1985 through 1987. The disaster played out from 1989 until the fishery was closed in 1992.

A sixth challenging question: Is the plant 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. Note the appearance of the word “likely” in 1995 and its subsequent use. This refers to likelihood, a 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 extremely likely 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.”

Uncertainty.

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

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 ? (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.

Fish 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

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 increase 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.

Factors that cannot be controlled by manipulation:

-Individual variation in epidemiological or drug effectiveness studies

-Effects of weather, change in climate

-List another example _____

Causation at multiple spatial scales.

-Weather more uncertain at weeks than tomorrow, even more uncertain next month

-Species numbers increase with increase in habitat diversity, which is greater at regional than local scales.

-List another example _____

Multiple causation:

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 draw conclusions an act in the face of uncertainty. The result is that people make different estimates, weight things differently, and arrive at different conclusions and different decisions.

Conclusions and 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 will have an effect, sometimes large.
2. Cigarette smoking. Our own behavior will affect our health, and those breathing the same air. But smoking by many people will place a substantial burden on the health care system, in the form of poor health outcomes.
3. Fish numbers. Conclusions and actions by any one person will have little effect on fish numbers. But fishing by many people, and by governments (through funding for economic development) will eventually drive any species to commercial extinction.

How do we draw science based conclusions and estimates of uncertainty when these affect many people, including people we do not know or will never meet? Statistical inference addresses this question.

To draw conclusion that affect many people, or that involve public funds or public policy, we need to make estimates that are rational (based on logic) and that are based on evidence. Statistical science is the use of the logic of math to measure the strength of evidence and to arrive at estimates of uncertainty based on clearly stated assumptions.

What is statistical science?

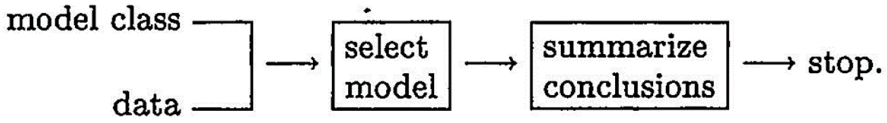
At the dawn of this century Nelder (1999) defined statistical science as follows.

Statistical science is not just a branch of mathematics;
it is not a purely deductive system,
it is concerned with quantitative inferences from data obtained from the real world.

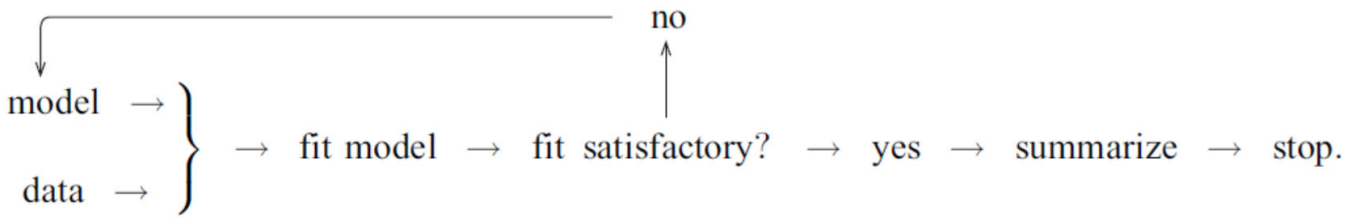
If we statisticians are to become statistical scientists we must become thoroughly familiar with the processes of science.

The subject should be renamed statistical science and be focused on the experimental cycle design-execute-analyse-predict.

The process of statistical analysis as presented in many textbooks appears to take the form:



I call this the read—calculate—print—stop mode of working. It assumes the model class is adequate. The new style incorporates a model checking loop:



In this course you will learn statistical science using this model checking loop. It was introduced in this course in 1992, with a generic recipe that included a model checking step for the normal error assumption and the straight line assumption of a linear regression. The model checking step was included in the course as a consequence of learning the statistical package GLIM. This package was developed by a working group of the Royal Statistical Society under the chairmanship of John Nelder in the 1970s.

Statistical science, as defined by Nelder, has two parts, data and models, which are used to draw inferences that extend beyond statistical summaries of the data. Thus, statistical science is not applied mathematics; it is a science that draws inferences from explanatory models based on measurement.

Nelder makes it clear that there are three modes of inference, not two.

there is a flourishing school of likelihood inference, to which I belong. It is relevant to the model selection stage, and it differs from Bayesian analysis in not incorporating prior distributions into the model.

What is statistical science?

Nelder lists non-scientific practices in need of removal.

> *this will mean replacing the calculation of P-values by measures of effects and their uncertainties, discarding multiple-comparison tests, playing down distribution-free methods and replacing them by the modelling of errors.*

This list can be expanded:

> Replacing data transformation with the modelling of errors and link functions
> Reporting results with units
> Using the residuals, not the response variable, to check the error model
> Reporting the likelihood ratio, not the p-value, as the measure of evidence
> Treating measured quantities as stochastic variables, not random variables

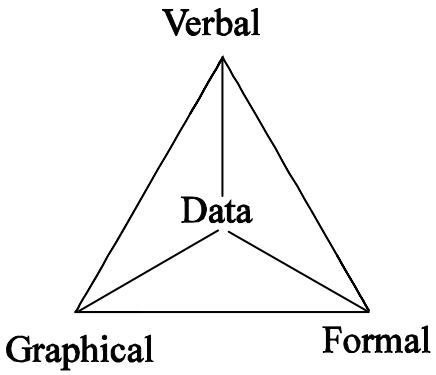
Statistical Science is Founded on Measurements

It is a truism to state that scientific data is based on sound measurement practice and clear understanding of how data are generated. Nevertheless, in talking with students, it is becoming increasingly common to encounter derived data, where what is measured is no longer visible. For example DNA loss is reported as a percentage. Percentage of what? Measurement turns out to be that of area covered by DNA in a gel, either scattered (lost) or intact in a tightly coiled structure. Percent loss is then the area of the scattered DNA, divided by the area covered by both forms. This matters because some knowledge of the measurements is needed in making an informed first choice for the error structure used in statistical analysis. Should we begin by using a normal error? In this case, with data as a percentage bounded at zero and one, a normal error might be acceptable if values are close to 50%. Unfortunately a normal error will produce misleading estimates if the data extends outward toward the boundaries of the data bounded at zero and one. A normal error with values near the boundaries easily produces 95% confidence limits that extend beyond zero, into the realm of a DNA deficit—more DNA lost than total DNA in a cell.

A clear understanding of the data results from asking: What are the units of measurement? For a per capita rate of change in population size, with units of percent per unit time, what are the units of time? Percentage of what – numbers? biomass? The practice of reporting data without units is amplified by statistical practice, which reduces measured variables to unitless ratios. Commonly used statistics—*t, F*, Chisquare, standard deviations—are unitless in the interest of generality. Parameter estimates—means and slopes—retain the units of the underlying variables. Unfortunately, means and slopes are too often reported without their units. In disciplines outside of physics and chemistry units are more often missing than present (Schneider 1994, 2009)

Statistical Science - Verbal, graphical, and formal models

Once we have a clear grasp of how are data were generated, the next step is a verbal model. Verbal simplifications is the beginning point for developing an understanding the relation of one variable to another. Graphical presentation is the next step. This facilitates the third step, writing a formal model that expresses the relation of one variable (a response variable) to another {explanatory variable).



The diagram shows a typical route:

1st, measured quantities (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 these allow us to make calculations, such as number of species expected in 1 hectare of tropical forest.

Quantitative reasoning entails repeated cycling from formal, to data, to picture, to formal model.It is not an arcane practice restricted to "modelers." It is a way of reasoning about physical, chemical, and biological phenomena. Statistical methods are based on models, they are not just recipes.

Formal models have many 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 science are not hard to find:
confusing correlation with causation
poor reporting of methods --> irreproducible results.
Poor practice affects us all--health and medicine.
--environment and ecology.

Statistical analysis is based on an explanatory model and an error model.

Statistics are a formal (*i.e.,* mathematical) way of
-discovering generalizations

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

-evaluating generalizations expressed as models

Statistical methods are used:
to interpret observations,
to evaluate complex evidence
to disentangle multiple causation
to increase efficiency in carrying out experiments
to measure the strength of evidence.
to measure uncertainty
to infer from data to model and from sample to poplation.

Data	=	Model	+	Residual
Y	=	$\Sigma \beta X$	+	ε
Observed	=	Expected	+	Residual
Response variable		Explanatory variables		Residual
Observed		Explanatory Model		Error Model

Looking ahead.

Part I of this course covers sound practice in
- defining measured variables
- rescaling measured quantities
- use of symbolic notation to express, evaluate, and test scientific concepts

Part II covers
- data equations
- frequency distributions
- modes of inference
- measuring evidence (likelihood ratios)
- measuring uncertainty (p-values and confidence limits)
- goodness of fit and likelihood ratio tests

Part III introduces a generic recipe for statistical analysis for one explanatory variable

Part IV extends the generic recipe to any number of explanatory variables

Part V further extends the general linear model (normal errors) to the generalized linear model (non-normal error models)

Part VI provides brief overview of exploratory data analysis.
It extends the general linear model to
- multiple response variables
- multivariate analysis
- correlation analysis
- generalized likelihood (AIC)

Course Structure

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

(graph here).

Some beliefs in need of updating

"Lies, damn lies, and statistics."

Variously attributed to Benjamin Disraeli and Mark Twain.
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" 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.

"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 have become commonplace in statistics. In the past mathematically trained statisticians sometimes looked down on this.

We will use numerical methods in this course, to calculate p-values when a normal distribution or other error model is not applicable, based on model-checking.

We will also use numerical methods to check our judgements of graphical displays used in model checking.