BCCS 2008/09: Graphical models and complex stochastic systems: Lecture 2: Statistical inference

A module of 10 lectures, 2 to 13 February 2009.
Given by Peter Green, of the Statistics group in the School of Mathematics.
Office: 4.11, School of Mathematics; tel: 87967; email: P.J.Green@bristol.ac.uk
Web page: http://www.stats.bris.ac.uk/∼peter/Teach/css.html

The module covers some modern aspects of statistical science. These make use of ideas in probability and graph theory (theory module 1), but are aimed at modelling and analysis of data from real complex systems, systems that exhibit uncertainty/randomness/variation/stochasticity. Methods using these ideas are used successfully in interdisciplinary work in many fields, from archaeology to zoology (my own recent work has touched on astronomy, agricultural science, genomics, neuroscience, protein structure, cryptanalysis, biometrics, medical image analysis, geographical epidemiology,. . .). There are also connections at a more theoretical level with models in statistical physics and theoretical chemistry.

The word **statistics** is used in many senses; a statistic can be a number or a measurable function of a random variable; statistics can be collections of either, or the discipline itself, and even that is a bit ambiguous – does it include probability? – does it include artificial intelligence or machine learning? – does it include all data analysis? We will usually use it as an abbreviation for **statistical inference**.

Basic ideas of statistical inference

Inference is about generalising from observed data (numbers, images, signals, text, . . .) to say something about underlying mechanisms. So it is not restricted to summarising or displaying the data you have. In a sense it is more about the data you might have had, or the data you will have 'next time', so it involves thinking about the 'mechanism' generating the data, and not only the data themselves. This is a philosophical challenge! What does it mean to say something about underlying mechanisms when data are uncertain? Just as a given mechanism might have generated data different from the data you have, *the data you have might have been generated from more than one mechanism*! So we are not talking about *deduction*.

To be a little pretentious about it, what we are trying to answer is the fundamental question in the philosophy of science. The discipline of statistical inference arose as an attempt to provide a model for scientific inference (in the presence of uncertainty, i.e. always!) that is logically sound, mathematically rigorous, and scientifically grounded. Statistics is the only part of the mathematical sciences where the interface to the 'real world' is part of the subject itself.

2.1 What's the big idea?

Statistical inference is the 'inverse of probability theory'.

Suppose we give a particular drug treatment to n patients with a certain medical condition. Let the random variable X be the number who respond positively. Let θ be the probability that a randomly chosen patient (one of the n, or one in the future?) will respond positively.

Probability theory tells us how to get X from θ – a reasonable *model* might be to assume that X is binomially distributed with parameters (n, θ), that is

$$P\{X = x\} = \binom{n}{x} \theta^x (1-\theta)^{(n-x)}$$

Statistical inference is about getting θ from X. Just as θ does not *determine* X, neither does X determine θ, so we might draw conclusions about θ of different kinds:

- we can give a 'best guess' or *estimate* of θ – for example, the number X/n might be a good guess

- we can give a range of values for θ, for example a *confidence interval* – the interval $[X/n - 1.96\surd(X(1-X)/n), X/n + 1.96\surd(X(1-X)/n)]$ is a common choice

- we can inquire whether there is any evidence against a pre-assumed value for θ – for example the value that characterises some other drug therapy – that is, *test a hypothesis* about θ.

Let us step back a little from this example. What are the ingredients of this set-up?

1. A model for the mechanism generating the data. It is a probability statement 'predicting' data X. It is an assumption, not a fact – its origin will be a question of scientific judgement, based on understanding of the real mechanism, experience, precedent, physical laws, convenience, ...

2. The model usually involves a parameter θ, whose value is unknown.

3. Both X and θ may be (very-) high dimensional. Either (but especially X) may be discrete or continuous.

4. A statistical model, just like any other kind of model, can be useful without being 'true'.

5. Statistical inference will provide us with

   (a) principles for turning the '$\theta \rightarrow X$' relationship into '$X \rightarrow \theta$' (we will meet maximum likelihood and Bayesian analysis)

   (b) methods, which are determined by model+principle

   (c) mathematical and computational techniques for deriving and implementing methods

   (d) tools for evaluating the quality of methods (so we can, e.g., say one estimate is better than another)

   (e) tools for criticising models (so the underlying assumptions can be challenged)

2.2 Maximum likelihood

The assumed model gives us a function connecting θ and X – as in the example above $P\{X = x\} = \binom{n}{x}\theta^x(1-\theta)^{(n-x)}$ – we will write this as $p(x|\theta)$. Notice two things about this – we suppress mentioning irrelevant constants like n, and more importantly, we simplify by omitting the symbol X for the random variable we are talking about. Take care if you do this!

In classical (frequentist) statistics, when the observed values of the data are substituted for x, and this function is regarded as a function of θ alone, it is called the *likelihood function*. Its value at a particular θ is the probability of getting the data you did get, assuming the model, for this value of θ.

The principle of maximum likelihood says that we should estimate θ by the value that make the data most probable – i.e. the value that maximises $p(x|\theta)$. It is not correct to call it the 'most probable' value of θ.

Example. If you differentiate $p(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{(n-x)}$ with respect to θ (NB not x!) – or, more simply differentiate $\log p(x|\theta)$ – and set to zero, you find that $p(x|\theta)$ (or its log) is maximised at $\theta = x/n$. (We should check that it is actually a maximum, either by calculus or algebraically.) So x/n is the *maximum likelihood estimate* of θ based on these data.

2.3 Repeated sampling

The conceptual basis for evaluating the performance of any statistical method is *repeated sampling*. Imagine the experiment/observation that produces X can be repeated indefinitely, under the same conditions. So we get a sequence of data sets: $x_1, x_2, \ldots$. Apply our statistical method (e.g. maximum likelihood estimation) to each data set, and look at the collection of results. If we are doing estimation, we expect the collection of estimates (e.g. $x_1/n, x_2/n, \ldots$) to cluster around θ (we talk about the bias and variance of an estimator to quantify this). If we are calculating confidence intervals, we expect the intervals to be as short as possible, but to cover the true θ with the stated confidence. If we are testing hypotheses, we expect both error probabilities to be small: that of declaring the hypothesis to be true when it is false, and that of declaring it to be false when it is true.

All these calculations can be done using probability theory (and calculus and algebra, typically).

Statistical theory yields various results to guide our choice of principle – e.g. that for a particular model, an *optimal* estimator can be found, or that maximum likelihood estimation performs very well in many circumstances (according to some criterion), but not universally, etc.

2.4 Bayesian analysis

The key idea in Bayesian analysis is to treat both parameters and data as the same kinds of object – random variables – and to use the same quantity – probability – to quantify both kinds of uncertainty, that is the uncertainty you have about X when you know θ, and the uncertainty in θ having observed X. This approach completely avoids all the logical brain-twisting needed to talk about tests of hypotheses and confidence intervals, and allows inferences to be stated in a pleasingly direct way, by computing the probability distribution of θ give the observed value of X – called the *posterior distribution*, $p(\theta \mid X)$.

There have long been philosophical controversies about whether this view of inference is acceptable. One issue is whether the two kinds of uncertainty mentioned above really can be treated 'the same'. A second issue is that if θ is (treated as) a random variable, then it must have a probability distribution even before you observe X – the *prior* distribution, $p(\theta)$. That expresses what you believe about θ before you have any data. On the positive side, this allows you to bring in any scientific information you have before conducting your experiment, in a principled way, and combine it with the information in the data, using

Bayes' theorem:

$$p(\theta \mid X) = \frac{p(\theta)p(X \mid \theta)}{\int p(\theta')p(X \mid \theta')d\theta'}$$

or simply $p(\theta \mid X) \propto p(\theta)p(X \mid \theta)$. On the negative side, some say that it is wrong to allow the result to be affected by prior information, which may be subjective and differ between different analysts.

Bayesian analysis has become enormously more accepted in recent years, because of a combination of (a) realisation that often, although the prior affects results, it may not affect them very much; (b) computational methods now exist to do Bayesian analysis on a serious scientific scale; and (c) thanks to these, for complicated models, Bayesian methods are actually more likely to be implementable than non-Bayesian ones. The rest of this module has a lot more to say on this point.

2.5 Reading

The basic ideas of statistical inference are found ubiquitously in every intermediate text book – anything that is both not so basic as to deal only with descriptive statistics and not so advanced that it sets out the ideas in an unnecessarily abstract way: just explore the library shelves (especially around QA276) or online catalogue. Covering the Bayesian approach is not quite so common – for these ideas, best to consult either a book that is explicitly an introduction to Bayesian statistics (e.g. *Bayesian inference in statistical analysis* by George E.P. Box and George C. Tiao or *Bayes and empirical Bayes methods for data analysis* by Bradley P. Carlin and Thomas A. Louis), or one that sets out to develop both the classical and the Bayesian ideas between the same covers (e.g. *Statistical inference: an integrated approach* by H.S. Migon and D. Gamerman).