On the construction and comparison of statistical models for scientific discovery

J.K. Lindsey
Biostatistics, Limburgs Universitair Centrum, Diepenbeek
Email: jlindsey@luc.ac.be

Abstract

The scientific process can be thought of as having two distinct stages, *discovery* and *confirmation* by replication. The second corresponds to many standard statistical procedures but the first is much more difficult to formalize.

This survey looks at the various steps in the model building process in the discovery stage: conception and study design, construction, selection, diagnostics, uncertainty, and interpretation. Some limitations of present practices are suggested and a number of outstanding problems described.

Particularly important problems arise in the comparison and selection of models involving different functions. These include how to allow for the uncertainty arising from fitting several distinct models to a data set and how to measure the relative complexity of models other than simply by counting the number of unknown parameters. Both are closely related to the discovery process in science.

KEYWORDS: Diagnostics, embedding, mechanistic model, model selection, model uncertainty, profile likelihood.

1 Introduction

Since its first development, the field of statistics has played important roles in many areas of society. Its beginnings can be found in astronomical prediction, social statistics, and epidemiology. Solid foundations were established through work in agronomy and genetics. Recently, a major impetus has come from clinical medicine. None of this should be surprising. We live in a society of uncertainty and statistics specializes in the study of uncertainty.

Modern statistics is primarily an invention of the twentieth century. Classical statistics developed between the two world wars, the greatest name of that period being Fisher. The 1950s and 1960s were a period of consolidation. In contrast, the last thirty years have seen a continuous revolution of statistical practice. Before that time, such important fields as survival and discrete data analysis, to name but two, did not even exist. These recent developments have arisen from at least two major impetuses: the computer revolution and the requirements of medicine and the pharmaceutical industry.

Statisticians like to believe that they are *the* experts in the study of uncertainty. Some feel threatened by recent developments, such as chaos theory, neural networks, data mining, and so on, that treat uncertainty in different ways than does the classical statistical approach. Statistics has powerful means for handling small data sets and for drawing general conclusions about the larger populations from which these are supposed to come. Statisticians are more at ease analyzing a sample survey than a complete census. However, they do not have a monopoly over the ways in which uncertainty can be handled.

Many statisticians have the unfortunate weakness of tending to believe that their favourite procedures are applicable in almost any circumstances. The frequentist, Bayesian, and likelihood schools all make claims to superiority in drawing inferences. Each has its strengths in specific contexts; none can provide the final solution. The frequentist approach was designed for decision-making in a repetitive situation, such as industrial quality control. The Bayesian approach emphasizes incorporation of individual subjective beliefs, appropriate for example in financial decision-making. The likelihood (Fisherian) approach concentrates on obtaining the maximum information from the presently available data without taking into account how it will be used, a primary goal of the empirical stage of scientific research.

In a similar way, many statisticians have their favourite statistical technique or model, often because they sacrificed an enormous amount of time and effort on it for their doctoral dissertation. Then, throughout their career, they attempt to apply it in all possible circumstances. My particular weaknesses are likelihood inference and principles of modelling.

Here I shall look at some aspects of inference specifically related to modelling that I did not cover in Lindsey (1999). I can only discuss one small area, delimited by the likelihood approach and scientific applications, making no claim that these ideas are more generally applicable. For example, I ignore completely decision-making problems, descriptive statistics, industrial applications, and so on.

Although science is only concerned with *repeatable* phenomena, I shall not concentrate on this aspect of the statistical endeavour, the verification of models. Rather, I shall look at the discovery and development of appropriate models up to the stage when it becomes feasible to entertain the possibility of repeatability. Then, the scientific community takes over.

Scientific discovery can arise in at least two distinct ways: new theoretical developments may point to something that then has to be empirically checked or new empirical data may contradict existing theory pointing to a modification or a new theory. The first poses a relatively simple statistical problem. The second is much more difficult: how can we assess that the given empirical data support a new theory derived from them better than the old, thus indicating that scientific replication will be necessary for confirmation?

2 Model conception

2.1 Understanding the scientific question

In spite of the pretensions of some statisticians (and names of journals), statistics is not a science; it has no subject matter in nature or society that it specializes in studying. It is rather a collection of *methods* for treating empirical data involving uncertainty. It forms an important part of the epistemology of some areas of science, particularly those involving living beings, where variability can be large. Thus, it is close to mathematics (which is even further from science), not only in using the latter discipline but also in that it thrives on abstraction from specific problems. But it is also far from mathematics, and closer to science, in that it proceeds from the specific to the general and that it necessarily involves empirical applications.

Consider, for example, a scientist who comes to a statistician wishing to fit the Michaëlis-Menten equation to some assay data. This model gives the initial velocity of an enzyme-catalysed reaction as a function of substrate concentration, x:

$$\mu(x) = \frac{V_{max} x}{K_m + x} \tag{1}$$

where $\mu(x)$ is the mean initial velocity, V_{max} is the maximum velocity (in practice, divided by a calibration constant), and K_m is the Michaëlis constant. Because this equation has the logistic form, many statisticians will immediately suggest the more general and 'much better' model for such assays,

$$\mu(x) = \alpha_0 + \frac{\alpha_1}{1 + \mathrm{e}^{\beta_0 + \beta_1 \log(x)}} \tag{2}$$

(I resist discussing the nonparametricians who point out that their splines, local polynomial smoothing, ... are superior.) Indeed, this equation can be rewritten

$$\mu(x) = \frac{V_{max} x^{\beta_1} + K_m V_0}{K_m + x^{\beta_1}}$$

with $\alpha_0 = V_{max}$, $\alpha_1 = V_0 - V_{max}$, $\beta_0 = -\beta_1 \log(K_m)$, and the concentration power-transformed by β_1. Admittedly, this latter equation can yield a measure of goodness of fit of Equation (1), in one particular direction, but at the loss of the mechanistic model that interests the scientist. How many statisticians will then check if $\alpha_1 = -\alpha_0$ and $\beta_1 = 1$ in Equation (2) are reasonable so that the original model is recovered?

Most statisticians have no scientific training, their background being primarily mathematical. (When they get together with medical doctors, most of whom also have no scientific training, the results can be close to tragic.) In the statistical literature, we often find statements such as 'we shall analyze a *real* data set' (as if most statistical data are not real) chosen to show that a favourite model is useful, without out any scientific context being provided about the data or any scientific theory behind the model, or 'scientific interest centres on ...' to defend that favourite model in some abstract context where no specific scientific problem has even been stated.

When faced with a scientific problem, statisticians cannot construct suitable models in isolation, without detailed interaction with the scientists. On the other hand, many scientists have insufficient mathematical knowledge to translate their theories into equations susceptible to confrontation with empirical data, and to combat the statistician's unscientific mathematical distortions of their theories. Thus, the first element of any statistical process within science must be the close cooperation and interaction among the actors involved.

2.2 Systematic and random aspects

The responsibility of the scientists is to provide the theory; that of the statistician is to translate it into a mathematical/statistical form, most often being a prime contributor of probabilistic/stochastic elements to handle the variability.

By 'model', I mean some function that allows one to calculate the probability of any possible relevant data set, perhaps after fixing the values of some unknown parameters. For example, the partial likelihood for Cox proportional hazards does not correspond to a model in this sense. Statistical models can generally be decomposed into two distinct parts. Some probability distribution is used to describe the *random* variability. Then, parameters within that distribution function are allowed to vary in *systematic* ways with covariates relating to subgroups of the population, time, space, and so on. The systematic part tells how the random part changes shape when these covariates change.

The most familiar parameter that is allowed to vary systematically is the mean or other location parameter. Usually, the scientist/statistician interaction process is not too difficult for conceptualizing changes in this parameter. It may simply involve solving some set of differential equations, for example.

Scientists (and most statisticians!) have much more difficulty in conceptualizing the form of the variability about this mean equation, and even more problem in allowing that variability to change with covariates (abandoning the constant variance hypothesis). When only measurement error is involved, as often is the case in chemistry for example, then a classical normal model is usually reasonable. But this is rarely appropriate when living beings are studied. Then, how can the scientist, or the statistician, judge *a priori* whether, say, a gamma, a Weibull, or a skewed stable distribution is most suitable? So much emphasis has been placed on the linear normal model and nonparametric procedures that we have accumulated little experience as to which distributions are really scientifically most appropriate in different circumstances.

2.3 Study design

For many people not using statistics frequently, this discipline simply involves supplying a cookbook of equations so that one knows when to use a Student-t test instead of a Chi-squared test, perhaps logistic regression instead of ordinary linear regression. To them, study design is not part of

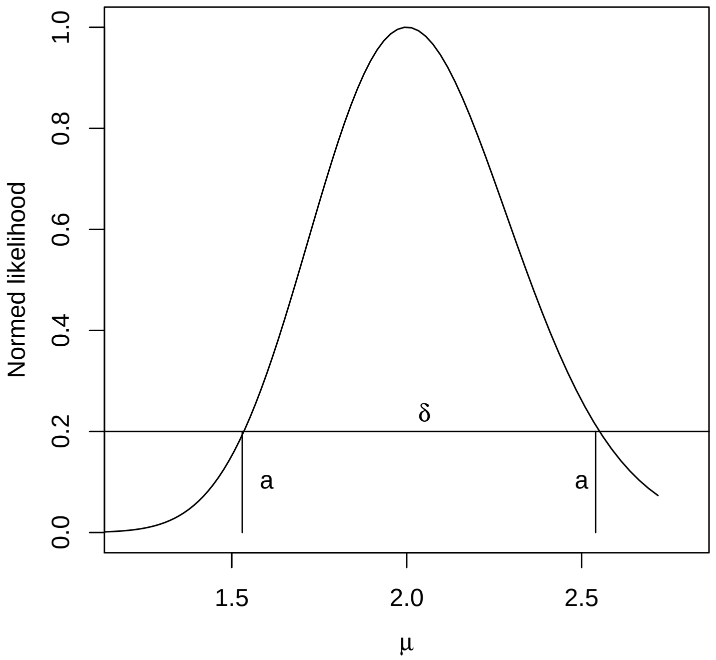

Figure 1: Sample size calculations for a Poisson distribution with mean, $\mu = 2$. Here, $\delta = 1$ is the scientifically-interesting difference and $a = 0.2$ is the plausibility required.

statistics; it is 'preparing a research project'! Nevertheless, in industry and science, statistics in the twentieth century has earned its position as a valued partner primarily for its contributions of randomization and blinding to avoid biases and of optimal treatment allocation and sample size calculations to reduce costs. The contribution of statistical analysis, including modelling, has been rather minor. This must change if statistics is to survive as a distinct and viable discipline in the next century.

Let us consider briefly the procedure for sample size calculation. To carry out any such calculation, we must, paradoxically, actually 'know' the parameter values that we wish to estimate. Thus, it is only feasible in very simple cases. Consider a study to estimate the mean, μ, of a Poisson distribution. After the study is performed, the maximum likelihood estimate will make the data most probable under the assumed model function. Suppose that we are only interested in models that make the data at least a proportion, $0 < a < 1$, of this maximum probability, where we can arbitrarily choose this value. For this given level of plausibility, the *precision level* of the parameter, we wish to obtain an interval of size δ around the estimated mean. This is illustrated in Figure 1 when we believe that $\mu = 2$ and choose $a = 0.2$ and $\delta = 1$.

The width of the likelihood curve in this Figure can only vary as a function of the estimated mean and of the sample size. However, we have fixed the mean so that we can calculate the required sample size. In this example, $N = 25$. Of course, if our guess at the mean is too small, our sample will be too large because the likelihood curve will be narrower, and inversely if the guess is too large.

This simple example illustrates the main principles involved in any exact sample size calculation. The three values, N, δ, and a, are intimately linked; we only have two 'degrees of freedom' to choose them. However, in more complex cases, approximations often need to be used.

3 Model building

3.1 Generality

Perhaps because of its close relationship with mathematics, statistics strives to produce *general* procedures that are applicable in a wide variety of situations. This has certain advantages but it can also have the unfortunate consequence that the techniques may not be very good in any specific circumstances. Thus, in a certain sense, modern statistics often shows fundamental ignorance of scientific principles, attempting to impose its 'generally applicable' methods in all situations instead of trying to understand each specific scientific problem and to develop specific procedures for it.

The classical linear model is the archetypical case of generality: it is widely believed that most problems can be transformed in some way so that least-squares multiple regression can provide a solution (at least if one's favourite technique is not applicable). If the relationship is nonlinear, a polynomial can be used. After all, it can be interpreted as a Taylor series expansion of some nonlinear function. But what how does that bring us closer to understanding what that unknown function actually might be?

3.2 Description versus explanation

Much of both classical and modern statistics is purely *descriptive*. It tries to describe empirical observations in some appropriate way without out any attempt at *understanding* the underlying phenomenon, the data generating mechanism.

In some areas, such as spatial statistics, little more seems possible at present. There, no alternative to descriptive techniques, such as nonparametrics, appears reasonable. In simple factorial experiments, with two or three levels of each factor, classical linear models may be suitable, although one may often question whether the normal distribution adequately approximates the variability. They provide a technological answer to what happens to the response when those factors are modified, without indicating why.

In contrast, the goal of science is to understand a phenomenon as completely and generally as possible. This can only be accomplished by developing a mechanistic model, such as the Michaëlis-Menten equation referred to above, to approximate the data generating process sufficiently well. However, by definition, as scientists always emphasize, any such model is never true or correct; no matter how appropriate, it is still an approximation to reality.

Nevertheless, much of modern statistics prefers empirical models to mechanistic ones, the extreme example being nonparametric statistics. Those areas of statistics that have escaped from this rule (for example, communications theory, statistical mechanics, population and molecular genetics, pharmacokinetics, computer science, pricing methods in financial mathematics) have almost exclusively been developed by non-statisticians.

3.3 Minimal assumptions

Much of modern statistics seems obsessed with avoiding making unnecessary or unfounded assumptions. This is an appropriate position when one is only interested in reaching a decision in difficult circumstances without really attempting to understand the phenomenon involved. Any assumptions that might lead to a wrong decision must be avoided. The procedure adopted must be robust to any remaining false assumptions.

In scientific research, this is essentially a dead-end approach. The only way that knowledge can be advanced is by making assumptions and seeing how they correspond empirically to reality. A nonparametric test of a difference between two treatments can reliably tell us whether or not such a difference exists under the conditions in which the experiment was carried out—so that an appropriate decision can be made. But, it can never tell us anything about *why* there was such a difference.

One especially pernicious effect of this reluctance to make assumptions is that little knowledge has accumulated about what distributions are suitable in various circumstances. For example, for survival data in medicine, little is known about the applicability of the many available distributions in various contexts because of the wide use of the semi-parametric proportional hazards model. (In additional, research workers are now having to face the fact that the strong assumption of proportionality is itself rarely supported by the data so that conclusions about treatment differences may often have been wrong.) Although masses of survival data have been collected over the last decades, we have learned virtually nothing about the mechanisms of survival of people with various diseases. This contrasts with the advances made in engineering applications using more mechanistic models based, among others, on the Weibull and inverse Gaussian distributions.

On the other hand, in disciplines where complex models are widely used, such as in pharmacokinetics, strong, unverified, assumptions about distributions are often made. Almost everyone in this field firmly believes that drug concentrations in the body have a log normal distribution. Unfortunately, this has only recently been checked empirically, still in only a few cases, and it very often proves to be wrong.

3.4 Key concepts

The fundamental concepts of model-building, such as rates (of flow) with their differential equations, intensities (of events), latent variables, state spaces, transition probabilities, and so on, are not encountered in basic statistics courses. In fact, model-building itself is only rarely studied in these courses.

Understanding the asymptotic properties of some t-test or developing a new score test is considered far more important than studying the wide range of ways in which observations may vary. Until very recently, most of the multivariate distributions available related to tests for multivariate data; they were not models to be fitted to data.

If the strength of statistics is in its handling of variability, then statisticians should do more to promote their prime tool for describing variability: the variety of probability distributions and stochastic processes available for fitting to data (not those used for the distributional theory of frequentist statistics or the prior distributions of Bayesians).

4 Model selection

4.1 Known model function

Classical statistics, whether frequentist or Bayesian, is almost exclusively concerned with situations in which the model function is known and the only uncertainty is about the values of the parameters in that function. Thus, one simplifies a model by testing if a parameter might be zero and examines the uncertainty about a (non-zero) parameter by finding confidence or credibility intervals for it. In fact, if the parameter is zero, the model function has changed. For the frequentist school at least, the important thing is that the (conditional) distribution has not changed its functional form. That is what its tests and intervals are based on.

In certain special situations, parameter estimation can be separated from examining goodness of fit of the model function. In the linear exponential family, minimal sufficient statistics, say $\mathbf{t}$, exist for the parameters and $f(\mathbf{y}|\mathbf{t})$ can be used to examine goodness of fit. Unfortunately, most mechanistic models do not fit into this framework so that such a separation is not possible. The distribution is not in the exponential family and/or the model is nonlinear so that the minimal sufficient statistic is usually $\mathbf{y}$.

4.2 Selection criteria

Model selection criteria, such as the AIC (Akaike, 1973) and the BIC (Schwarz, 1978), have been developed, but these have a fundamentally different basis than the classical Bayesian and frequentist procedures (Burnham and Anderson, 1998; Lindsey, 1999). These criteria can provide results that directly contradict the classical Bayesian and frequentist ones in many situations.

Model selection criteria are fundamentally likelihood based. They do not require a probabilistic interpretation of the conclusions being drawn. The likelihood function provides a measure of how close a given model is to the data. However, a more complex model has more chance of being close to the data so that this must be taken into account. Then, the $(-\log)$ likelihood is penalized by some function of the number of estimated parameters.

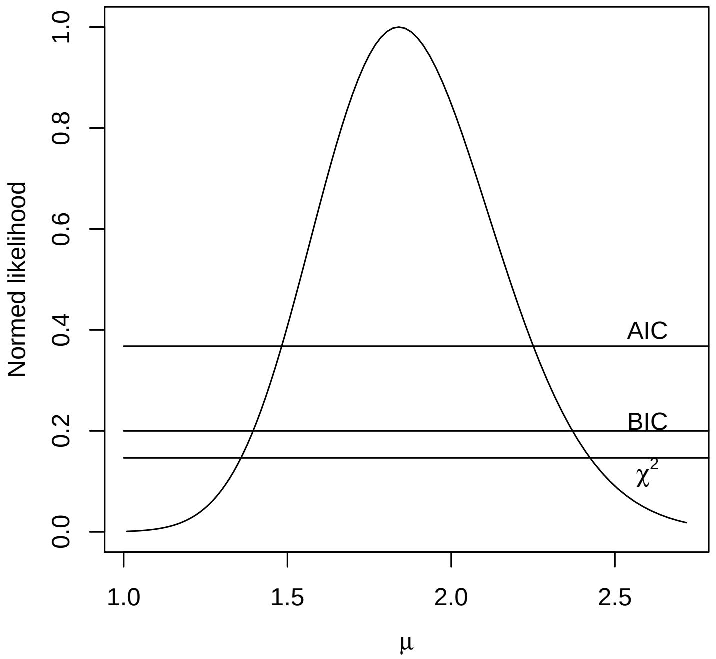

Figure 2: Three criteria for a plausibility interval about the mean of a Poisson distribution.

Any model selection criterion, even classical step-wise regression using Student t or Chi-squared tests, can be interpreted as the construction of plausibility intervals about parameters. A parameter is eliminated if the point at which it disappears from the model, often zero or one, is included in the interval. For three such standard criteria, this is illustrated in Figure 2 for data relating to the same Poisson mean problem as in Figure 1, where we found $N = 25$.

In the notation used above (Figure 1), the classical Chi-squared criterion sets $a = \exp(-\chi^2_1/2)$, the standard AIC has $a = 1/\mathrm{e}$, and the BIC has $a = 1/\sqrt{N}$. The difference between the classical procedures, whether Bayesian or frequentist, and proper model selection criteria lies in how the level changes for regions involving more than one parameter. For classical procedures, the change arises from the difference in distribution as the degrees of freedom (say p, the number of estimated parameters) change: for example, $a = \exp(-\chi^2_p/2)$. In contrast, for proper model selection criteria, the level is given by a^p. This ensures that we avoid the contradictions in Bayesian and frequentist inference that can arise when different numbers of parameters are estimated. Model selection criteria yield inferences that remain *compatible* when the numbers of parameters differ (Lindsey, 1999).

Although the level for the standard AIC with one estimated parameter seems very high in Figure 2, this quickly changes as the number of parameters increases. For more than seven parameters, the level, $\exp(-p)$, given by the standard AIC is lower than that, $\exp(-\chi^2_p/2)$, for a Chi-squared region at the 95% level. This is scientifically reasonable as more complex models are more highly penalized. Note, however, that the plausibility level for the AIC need not be fixed at $a = 1/\mathrm{e}$ for one parameter but can be chosen so as to obtain any desired precision level, as suggested above for sample size calculations.

On the other hand, notice that the plausibility level of the BIC involves N so that the sample size is fixed with the plausibility level, in contrast to the other two. We have lost a 'degree of freedom' in calculating sample size. This is generally an undesirable characteristic.

It is also important to emphasize that regions defined by one fixed value of a are only a crude summary of the complete likelihood surface. A set of them for various values of a is more informative in summarizing the shape of the likelihood function. But, those for a fixed a do provide us with a means of comparing regions arising from likelihood surfaces of different dimension, something that is impossible without such a criterion of calibration

4.3 Can complexity be measured?

Once a model has more than one parameter, things rapidly become more complex. Let me continue with my Poisson example. Above, I had a sample ($N = 25$) whose mean I suspect may be about $\mu_1 = 2$. I now take a second sample of 25 under conditions where I think the mean is about $\mu_2 = 3$, for a total sample size of $N = 50$. I am interested in the ratio of means, say $\lambda = \mu_1/\mu_2$. As a complementary parameter, I shall simply take μ_1. Recall that we have chosen to make inferences using $a = 0.2$. The likelihood surface for these two parameters is plotted as contours in Figure 3. The outer contour, $a^2 = 0.04$, is the appropriate one for a joint likelihood region at this plausibility level. (The second one from the outside is $a = 0.2$.) But then how do we proceed to produce informative graphics when we have more than two parameters?

In constructing a theory, scientists are interested in obtaining the simplest possible explanation for the phenomenon under study. They start with the simplest reasonable model and introduce no additional parameters or variables unless they are absolutely necessary. This contrasts with the approach of many statisticians. For example, in multiple regression or generalized linear model problems, one often starts with the most complex (saturated) model and tries to simplify it. As is well known, starting from the simple and from the complex will often not produce the same final result (unless all subsets regression is used). However, here the basic difference in philosophy of model building is more important than the difference in results in specific cases.

Classically, statistics measures the complexity of models, in relation to the available information, in terms of the degrees of freedom. This is closely related to the model selection penalties. In both cases, one may question the adequacy of such measures of complexity simply in terms of the numbers of estimated parameters. For example, is a linear model less complex than a nonlinear one with the same number of parameters? Are the gamma and Weibull distributions twice as complex as the exponential distribution because they have twice as many parameters?

This question is of direct relevance to scientific discovery. Suppose that the data indicate that some new model function is appropriate and that this new model has as many unknown parameters as the old one. Is the better fit, both to the current data and in future replication of the study,

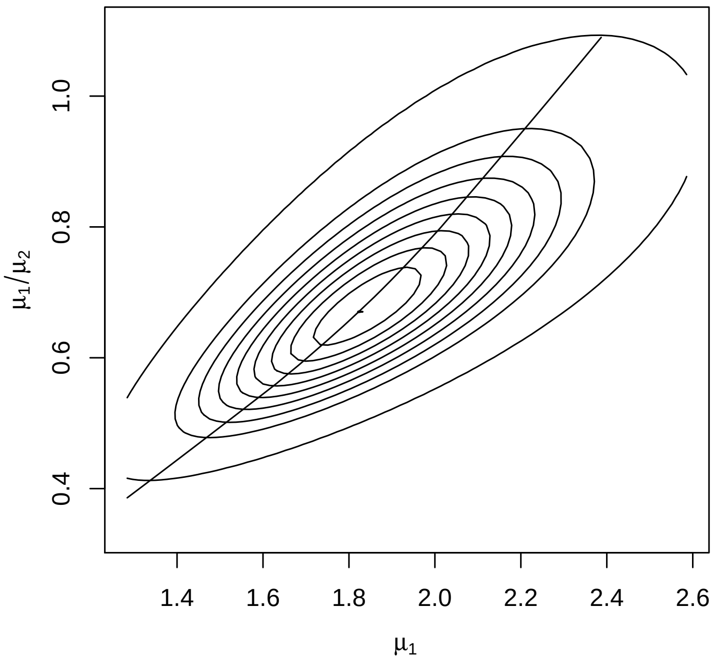

Figure 3: Contours (0.04, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9) of normed likelihood for the mean and the ratio of two Poisson means. The diagonal line shows the profile likelihood for the ratio of means.

simply due to the greater complexity of the new function?

Information theory has been much concerned with measures of complexity. Unfortunately, exact values, such as Kolmogorov complexity, are not computable. Many approximations, such as those developed by minimum description length (Rissanen, 1983, 1987), use approximations, most of which result in some familiar model selection criterion or a modification of one. In that context, my a is the precision of the parameter space, for example $1/\sqrt{N}$ yielding the BIC. The latter represents the magnitude of the estimation error in a parameter.

However, we have seen that a has a likelihood interpretation in terms of precision as well. It is the proportion of the maximum probability of the observed data (their likelihood) that we are prepared to accept for a model in the retained set. In this sense, it is not something inherently fixed. However, it evidently should not be less than $1/\sqrt{N}$. For the standard AIC, this only means that 1/e is too small for sample sizes of seven or less!

Unfortunately, these measures from information theory do not solve our problem. They calculate the minimum number of bits required to transmit the (discrete) data *given* the statistical model function and its parameter values. These are measured respectively by the negative log likelihood function and the number of bits required to transmit the parameter values themselves (at a given level of precision), the penalty. They do not take into account the cost of transmitting the definition of the statistical model function itself. This must vary with the complexity of that function. The person who develops a more appropriate (likelihood-based!) measure of complexity than the number of estimated parameters will become famous.

4.4 Comparing functions

Much of statistics can be seen as a model selection problem: Should my regression model be modified to include this covariate in it? Is a model with a mean of 3.4 more appropriate than one with a mean of 4.3? What set of parameter values (that is, subset of models) should I select as appropriate for these data?

For a given model function, each different parameter value defines a distinct model. Thus, constructing a confidence or credibility interval for a parameter can be interpreted as selecting the set of models having parameter values in that interval. Classical statistics, whether Bayesian or frequentist, is good at studying parameter values for a given fixed model function.

Any model function with given, fixed parameter values allows one to calculate the probability of the observed data: the likelihood function. Classical statistics can easily compare such models when the parameter values are varied. The problem is considered to be much more difficult when one wants to compare, say, gamma, log normal, and Weibull distributions. And yet, for fixed parameter values in each, the probabilities of the data can still be calculated, and compared.

One classical solution is to embed the models of interest in a more global one. For example, the above three distributions can be embedded in the generalized gamma distribution. The problem then reduces to one of studying a new parameter, with specific discrete values corresponding to each of the model functions of interest. In contrast, likelihood-based model selection criteria allow direct comparison of different model functions without the need for such embedding. (This, of course, is not meant to imply that embedding is not useful.)

5 Model diagnostics

5.1 Questioning the model and the data

Scientists are wary of models that describe their data too well. They know that some part of the data will certainly be found to be wrong. The scientists that I have met argue vehemently against letting the data speak for themselves (just as they spontaneously argue against allowing personal opinions to enter into account, without knowing that Bayesian statistics even exists). A major discrepancy between model and data may indicate a scientific breakthrough so that great care must be taken.

On the other hand, models for which it is worth collecting empirical data, and the theories behind them, are generally supported by a wide variety of sources. Unless the experiment is a *crucial* test of the theory, the data set arising from it will generally not be sufficient cause for a model to be completely rejected. No scientific model, or theory, will be abandoned unless a better one is available to replace it: model comparison, not testing. Models and their theories must be testable, not in terms of null hypotheses, but as compared to competing models and their theories. This is exactly what likelihoods are about.

Much of modern statistics has attacked this problem of confronting data and models from the other end. Instead of using rigorous models with strong assumptions to determine which observations may be wrong and which theory is supported, it has concentrated on developing general procedures with supposedly weak assumptions that are 'robust' to data errors and general methods for detecting 'outliers'.

Most model diagnostics, particularly those based on residuals, were developed specifically for linear normal models. Often, they are based on the mean of the observations, not taking into account the changing form of the distribution around the mean, for example as covariates change. Their adaptation to other contexts, even to generalized linear models, is rather *ad hoc* and often not very informative. In many realistic models, the information for checking the model cannot be separated from that for estimating the parameters, as mentioned in Section 4.1. It has been my experience that standard diagnostics can often indicate no problem with a given model and yet a rigorous model selection procedure would reject it in favour of some other much better fitting model.

5.2 Amending the model

If the data have been properly cleaned and checked and if careful model selection has been carried out, inspection of model diagnostics should almost never reveal anything unexpected. Outliers that are erroneous values should have been detected by the cleaning process, although scientists know that this is never infallible. All reasonable alternative models should have been considered in the selection process and the ones best fitting the data retained. The remaining possibility, if diagnostics detect an anomaly, is that the data are indicating some unforeseen modification to the model, or some completely new model that was not previously under consideration. This is the substance of scientific discovery; it does not happen often in one's lifetime!

Non-erroneous outliers can only be defined in terms of a given model. If they prove important with respect to that model, it must be modified to accommodate them. This may involve introducing missing covariates, developing a more appropriate nonlinear model, using a more 'robust' distribution with heavier tails, and so on.

6 Model uncertainty

6.1 The role of prior knowledge

If a model selection procedure has been used, this implies that several, even a large number of, models have been fitted to the data. Some have argued that this model uncertainty should be taken into account in drawing conclusions from a study. Should these conclusions be penalized by the number of models tried, in a similar way to the model selection penalty for the number of parameters estimated? On the other hand, one might argue that model selection has reduced uncertainty by eliminating clearly unacceptable models.

The answer to these questions will depend, among other things, on how the various models came to be tried for the given data set. If an exhaustive list of possibilities (known competing theories) was compiled before data collection and only those tried, the situation will not be the same as if the best model found was suggested by the data set itself (a possible scientific discovery). If an exhaustive list of possible models could be pre-established, then we are in a case of at least partial confirmation of previous results, the repeatability of science, not discovery. If the chosen model was suggested by the data, then only new data from future independent studies by the scientific community can confirm the choice.

6.2 Inferences about individual parameters

Once a reasonable model function has been selected, one often wishes to make inference individually about one or more of the parameters. This is still a model selection problem: selecting a subset of models specified by a given range of parameter values.

A first criterion for proceeding is that we do not find any contradictions with respect to our model selection process. For example, a plausibility interval of reasonable values for a parameter that has remained in the model should not contain the value indicating that it should be removed from the model.

Except in very special cases of orthogonality of parameters, inferences about any one parameter must depend on the values of the others. If we look at different fixed values of μ_1 in Figure 3, it is clear that, for each, our conclusions about λ will change. How can this uncertainty be taken into account? Statisticians have spent a lot of time working on this problem. The Bayesian solution, a marginal posterior distribution which is an *average* over models with different values of μ_1, is unintelligible in terms of likelihood. I believe that it is scientifically meaningless: reasonable values of λ are changing depending on the value of μ_1 so that no average is interpretable. Frequentist solutions, such as conditional and modified profile likelihood, are equally suspect as these corrections can *narrow* the plausibility region in the face of this uncertainty!

Let us instead consider ways of summarizing this likelihood surface in one dimension for the parameter that interests us. Because the plausibility of our parameter of interest varies with μ_1, let us first take a series of cuts through the surface for various values of this latter parameter, as superimposed in the left graph of Figure 4. The outline of this graph is the well-known normed profile likelihood, but this way of producing it is more informative (at least when there are only two parameters). It shows how the values of the parameter of interest become less plausible as the second parameter moves away from its maximum likelihood estimate.

Of course, the normed profile likelihood can also be obtained directly. It is the line of highest

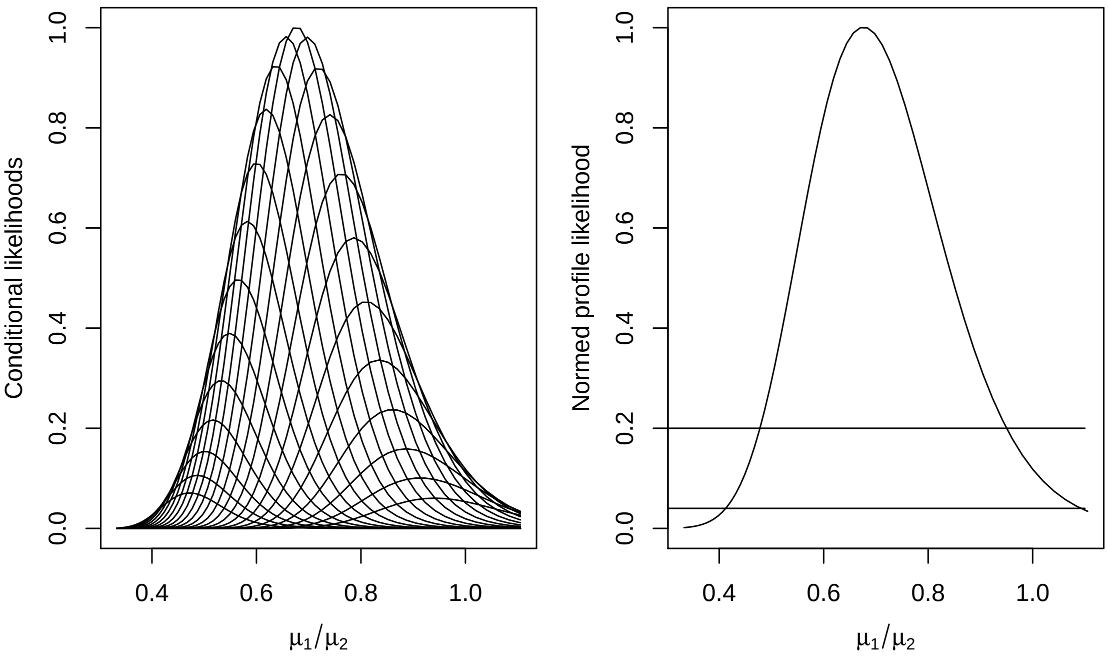

Figure 4: Inferences about the ratios of two Poisson means. Left graph: superimposed cuts through the likelihood surface of Figure 3 for a series of fixed values of the first mean. Right graph: normed profile likelihood with plausibility levels of 0.2 and 0.2^2 indicated.

likelihood when viewed from the axis of the parameter of interest, as shown by the diagonal line in Figure 3. This is plotted in the right graph of Figure 4. With some misgivings, the frequentist school uses this as if it were an ordinary one-parameter likelihood instead of a summary of a multidimensional surface.

Our problem here is to decide what plausibility level we should use with such a summary likelihood curve. The values of 0.2 and 0.2^2 are shown, corresponding to the points where the diagonal line in Figure 3 cuts the second and outermost contours. The former would be the frequentist choice: treat the curve, at least approximately, as an ordinary one-parameter likelihood. The problem is that this assumes that, at each point on the graph, μ_1 takes exactly its maximum likelihood estimate for the corresponding fixed value of λ. However, our model clearly has two estimated parameters so that the model selection criteria must be based on the latter, a^p (here with $p = 2$); otherwise, we risk drawing incompatible inferences. The wider interval allows for the unknownness in μ_1 at the proper level of uncertainty; it is a *maximum* rather than an average. However, it is not clear if this need be the narrowest interval possible for the given level of plausibility.

6.3 A global model

In looking more closely at model uncertainty, let us consider first the simplest case where all of the models examined are based on the same distributional assumption. The order in which they were fitted should be unimportant as only the total set of models examined should play a role in any measure of uncertainty. For example, we might be in a standard linear multiple regression situation where the distributional assumptions are not in question. A global model will exist that contains all covariates tried (including transformations, interactions, and so on). Notice that the number of models examined may be much larger than the number of parameters in this global model, as for example with all subsets regression. The chosen model contains a subset of these covariates.

Now suppose that a number of different distributions were also considered, in a simple case, say the generalized linear models based on the log normal, gamma, and inverse Gaussian distributions. As suggested above, in some cases, such comparisons can be conducted by embedding all possibilities within a more complex distribution with extra parameters. Suppose however that we are not interested in intermediate distributions, but only exactly those specified, because they correspond to distinct scientific theories. (Recall my example of the treatment of the Michaëlis-Menten equation in Section 2.1). Then, we can set up a global likelihood function containing indicator functions as to which distribution is actually used:

$$\mathrm{L}(\boldsymbol{\theta}, \phi) = I(\phi = 1) f_{LN}(\boldsymbol{\theta}_{LN}) + I(\phi = 2) f_G(\boldsymbol{\theta}_G) + I(\phi = 3) f_{IG}(\boldsymbol{\theta}_{IG}) \tag{3}$$

The indicator function, taking values zero or one, contains an unknown parameter, ϕ, with discrete values 1 corresponding to the log normal, 2 to the gamma, and 3 to the inverse Gaussian distribution. Except for the discreteness of ϕ, this likelihood function differs little from those arising from embedding. But how many parameters does it contain? $\boldsymbol{\theta}_{LN}$, $\boldsymbol{\theta}_G$, and $\boldsymbol{\theta}_{IG}$ all have the same dimension, but only one of the three actually appears in the likelihood function, depending on the value of ϕ. Thus, we are again in a situation where all models can be nested in a global model. We have a legitimate likelihood function that allows us to calculate the probability of the observed data for all possible parameter values.

6.4 Defining the problem (if there is one)

Plotting profile likelihoods for the parameters of most interest in a model is one simple way of providing us with indications of the uncertainty about the coefficients in this model function, *given that it is the only model function under consideration.* The height of the curve defining a plausibility (confidence or credibility) region will depend on the number of estimated parameters in that model (determined by the AIC, χ^2, or other criterion). The more estimated parameters, the lower this height and the larger the region. In what way should this height be lowered even further to account for the uncertainty arising from the number of other models tried?

Care must be taken here. If we lower the height defining the region of acceptable parameter values at this stage, after model selection has been completed, the enlarged region for a parameter may include zero values so that the previous conclusions from model selection are altered and a simpler model function indicated. Hence, for such a correction to work without contradictions, that is, provide compatible inferences, the complete set of models to be tried must be known in advance.

Numerical procedures, whether Newton-Raphson, simulated annealing, or other, to obtain optimal parameter values are a model selection process: they automatically examine many models to find an optimal one. We may ask if, from a likelihood point of view, such maximization procedures differ fundamentally from all subsets regression or the procedures necessary to find the optimal model in a global function such Equation (3). Certainly, from a frequentist viewpoint, they do.

Nevertheless, whether to penalize for the number of models examined, and if so how, still remains as a fundamental statistical problem.

7 Model interpretation

7.1 Parameters

In a certain sense, parameters are arbitrary, only serving to specify some given model function. They can generally be transformed without fundamentally changing the meaning of the model. This is reflected in the invariance of inferences from the likelihood function to reparametrization of a model. For example, in a regression model, the essential thing is how the probability of the various possible responses changes with the covariates: the changing shape of the (conditional) probability distribution about the regression curve. I like to remind my students that, for continuous response variables, the probability of an observation lying exactly on the regression line is theoretically zero, in spite of the fact that it is confusingly called the 'expected value'!

However, in a mechanistic model, each parameter often has a specific meaning. For example, it should make sense that any parameter, and not just the mean (or those referring to it), can vary with covariates in an interpretable way. Of course, for many, this will be found empirically not to be the case.

7.2 Extrapolation

To a very large extent, advancement of science is based on the construction of new theories, supported by models, that produce verifiable predictions outside the range of those produced by existing theories. Their success often hinges on being able to predict what will be observed in cases outside the data available to construct the theory. In other words, science depends, fundamentally, on the production of theories that are successful at *extrapolation*.

This contrasts with the way in which statisticians usually proceed. A regression model is fitted to data, but only considered useful for predictions *within* the range of those observed data. Extrapolation is considered to be dangerous and foolish. An important exception is, of course, the work in time series prediction, but unfortunately much of this is not based on mechanistic scientific models.

8 Conclusions

Science involves

- developing theories and accumulating knowledge to understand, not just to describe, nature and society;
- doing this without any view as to how they will be used;
- setting up simple models based on some specific theory;
- clearly stating assumptions;
- confronting the models with empirical data, with an outlook to discovering new models;
- but being wary of those data;
- extrapolating outside the observed data;
- the community of scientists checking repeatability of the results;
- only abandoning a model if a better one is available.

Many of these principles are in direct contradiction with much of current statistical teaching and practice. Thus, it is an unfortunate fact of life that much of modern statistics is anti-scientific. Most statisticians have been trained in mathematics departments out of contact with science. In the meantime, top-level scientists have had to work out their own new statistical techniques for their specific problems, occasionally adapting what they can from the statistical literature.

Nevertheless, statistics has come to play an important role in certain areas of research and development. Often, as in clinical trials, this is primarily due to its promotion of basic *design* principles such as randomization, blinding, and so on. Much remains to be done.

A few of the unsolved statistical problems raised above include

- What probability distribution is most appropriate to describe each specific scientific phenomena?
- What is the best way to represent likelihood regions in more than two dimensions?
- How can the plausibility level of a normed profile likelihood for one parameter, in the presence of several others, best be calibrated?
- How can the complexity of a model function better be measured other than simply by the number of unknown parameters?
- What diagnostics should be used outside the linear normal model, especially when the minimal sufficient statistic for the parameters involves the complete data?

- Should we account for model uncertainty arising from examining several models and, if so, how?

After thirty years of constant revolution in statistics, we may well ask if we are headed in the right direction.

Acknowledgments Göran Arnoldsson, Philippe Lambert, Patrick Lindsey, David Sprott, and Pablo Verde provided valuable comments on an earlier draft.

References

[1] Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In Petrov, B.N. and Csàki, F., *Second International Symposium on Inference Theory*, Budapest: Akademiai Kiàdo, pp. 267–281.

[2] Burnham, K.P. and Anderson, D.R. (1998) Model Selection and Inference: A Practical Information-Theoretic Approach. Berlin: Springer-Verlag.

[3] Lindsey, J.K. (1999) Some statistical heresies (with discussion). *Journal of the Royal Statistical Society* **D48**, 1–40.

[4] Rissanen, J. (1983) A universal prior for integers and estimation by minimum description length. *Annals of Statistics* **11**, 416–431.

[5] Rissanen, J. (1987) Stochastic complexity. *Journal of the Royal Statistical Society B* **49**. 223–265.

[6] Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.