Modelling: the classical approach

E. F. HARDING

Department of Pure Mathematics and Mathematical Statistics, Statistical Laboratory, Cambridge University, 16 Mill Lane, Cambridge, CB2 1SB, U.K.

Abstract. The kernel of the 'classical approach' to statistical modelling is the use of procedures inspired by the 'classical' objectives of Hypothesis Testing, Confidence Intervals, and Parameter Estimation, in the context of statistical models formulated (usually) in a highly specific way, and evaluated by the techniques of Sampling Theory.

Used sensitively, intelligently and flexibly, these procedures offer a powerful and adaptable approach to statistical problems arising in scientific contexts (though as usually taught, these methods appear limited and rigid). Their power lies in that they permit sharp focus on specific aspects of a theory, their flexibility in that they offer a wide choice of aspects to examine; but these are only realised when the corresponding statistical models are continually re-evaluated in the wider logic of the scientific context. These remarks are illustrated with a variety of examples.

Introduction

Given the diverse and sometimes fragmentary character of 'modelling' as practised, the opening contributors have been asked to attempt unifying presentations from three main theoretical standpoints: 'Classical', 'Likelihood' and 'Bayesian'.

Modelling is setting up a relationship between theory, data and reality. I shall interpret *classical* as denoting mainly tests of significance, hypothesis-testing and confidence intervals, and estimation—*classical* in almost the musical sense of a relatively traditional stylistic formalism, well understood, familiar and easily assimilated.

I must try not to pre-empt what Professor Aitkin† will say about likelihood, or Professor Smith about Bayesian methods. Classical modelling is often (though by no means always) non-Bayesian, but use of the likelihood function is unavoidable; I shall use it, however, in a quite classical way.

The approach adopted here will use examples to demonstrate the power of the classical approach, point out some limitations and dangers, and exemplify its methods. A theme of the Conference is that users seek *models developed from and supported by data, leading to an increasingly crucial rôle for Statistics in the modelling process*. I do not take this to mean that a model is good merely because it gives a good fit to the data: a satisfactory relationship with ideas and theories specific to the investigation is also essential. On the other hand, *the use of Statistics within modelling* will be a constant feature.

Foundations

The logical kernel of the classical approach is the significance test of a null hypothesis, and its extension to hypothesis-testing within a family of alternative hypotheses.

Let H_0 be a hypothesis to explain data D_0, and $\Delta(D_0; H_0)$ a *measure of discrepancy* between D_0 and H_0: the larger Δ, the more remote H_0 as explanation of D_0. When H_0 holds, let $\Delta(D; H_0)$ have a definite distribution when D varies randomly under H_0, and let $\delta_0 = \Delta(D_0; H_0)$.

† Was Aitken in the original

The specification of Δ induces a nested structure on the sample space, for given H_0, in terms of subsets such that each subset contains all observations D for which $\Delta(D; H_0) \geq \delta$ for some value of δ. Conversely, given a family $\{H\}$ of hypotheses, and observed data D_0, a nested structure is induced on $\{H\}$ according to $\Delta(D_0; H) \geq \delta$.

To each subset in the sample-space nesting can be attached its probability

$$\alpha = P_{H_0}[\,\Delta(D; H_0) \geq \delta\,]$$

under H_0. Therefore there is an inverse monotonic relationship between the defining discrepancy-level δ and the probability α: the smaller α, the greater δ. This is the basis for a test of a given hypothesis H_0, since when data D_0 are observed and we also have, for sufficiently small given α,

$$P_{H_0}[\,\Delta(D; H_0) \geq \delta_0\,] = P_{H_0}[\,\Delta(D; H_0) \geq \Delta(D_0; H_0)\,] \leq \alpha$$

then we can assert that the observed datum belongs to an extreme (discrepant) class whose total probability is implausibly small, such hypotheses being 'rejected at significance level α'.

Conversely, given a family $\{H\}$ of hypotheses, each possible datum D_0 maps into the subset of $\{H\}$ not rejected at level α by the test based on $\Delta(D_0; H)$ when H is taken as null hypothesis. This set of hypotheses is a confidence set at level $p = 1 - \alpha$, since if any $H_0 \in \{H\}$ is true, then the probability is at least p that the set so constructed contains H_0.

The above formulation is clearly very general and flexible, in that the choice of discrepancy function $\Delta(D; H)$ is open, and no fixed level of significance (α) or of confidence (p) is set. When 'powerful' test procedures are available, these are embraced; but ad-hoc or expedient procedures, non-parametric or distribution-free methods, and approaches to the testing of one dimension of a multiple parameter are also covered.

I regard the above dualism between hypothesis tests and confidence intervals as primary in the classical approach. Estimation (while it may be the main objective on a given occasion) is a derivative procedure. By *estimation* I mean the production of 'point estimates' as such: if they are accompanied by appraisals of precision (e.g. standard deviations), then (implicitly) the formalism of (possibly approximate) confidence intervals is being used. From this standpoint, an appropriate point estimate is a parameter value common to confidence intervals at all confidence levels.

Mode of Application

In the above formulation, all discrepancy measures which are one-to-one monotonic functions of a given measure Δ are equivalent in that they will, for a given hypothesis H_0, give rise to the same nested structure of subsets of the sample space, and to the same assignment of probabilities α to the subsets; and likewise on a family $\{H\}$ of hypotheses is induced, for given data D_0, always the same nested family of confidence sets with the same confidence levels p.

In short,

$$\alpha(D_0; H_0) = P_{H_0}[\Delta(D; H_0) \geq \Delta(D_0; H_0)]$$

is itself a measure of discrepancy, and can be taken as canonical representative of the set of all equivalent measures Δ. This amounts to transforming a given measure Δ to a 'universal' scale, that of the probability α, and this scale also has a universal operational interpretation, namely the chance (when H_0 is true) that so large a

discrepancy as $\Delta = \delta_0$ should occur. From this point of view there is no essential difference between equivalent measures Δ.

In applications, however, a given measure of discrepancy Δ will not arise arbitrarily: it will represent or summarise what the investigator perceives as being relevant and important features of the relationship between data and reality. In the course of any extended or complex investigation, the relevant features considered will vary kaleidoscopically as the problem is viewed under different aspects. The investigator will choose, among equivalent measures, one that immediately reflects his intuitive or reasoned perception of the current aspect.

The rôle of the hypothesis H may be multiple. On the one hand, H may be formulated in terms from the domain of reality, on the other hand it may be an abstract label specifying a particular probability distribution. Often it has both rôles. By definition, a *statistical hypothesis* is a sentence specifying a unique probability distribution (as a probability model). Two such sentences have the same meaning, as probability models, if they specify the same distribution even when, expressed in different real terms, they have different real meanings.

For instance, the two sentences:

(A) Events occur in a unit interval as a homogenous Poisson process whose rate μ has a Gamma distribution with index α and scale parameter β;

(B) events occur in a unit interval as a Poisson process whose rate μ at time t, given that r events have already occurred, is given by $\mu = \lambda + \gamma r$;

both imply

(C) the probability of n events in the interval is

$$P(n) = \binom{\rho + n - 1}{\rho - 1} p^{\rho} (1-p)^n \quad (n = 0, 1, 2, \ldots),$$

i.e. a negative binomial distribution, where we have:

(A) $\rho = \alpha \qquad p = (1+\beta)^{-1}$

(B) $\rho = \dfrac{\lambda}{\gamma} \qquad p = e^{-\gamma}$.

Thus (A) and (B) are the same statistical hypothesis, but have quite different real meanings, and cannot be distinguished by data on n alone.

It is a common fallacy to start with one hypothesis stated in real terms (such as (A)) and deduce a statistical hypothesis (here (C)), verify the latter's goodness of fit, and complacently infer that (A) is the truth. It is also common to choose a distribution (C) because its fit is good, and infer that a mechanism (such as (A)) that the investigator knows about is the case, not suspecting that there is a statistically equivalent, but really different mechanism (B). There is no theoretical way out of such dilemmas, which can only be resolved by understanding of the real terms of the investigation, or by taking account of further data of a different kind.

The hypothesis H, then, on the one hand expresses the way the real features of the problem influence the data, and on the other hand specifies the probability model for the influence of sampling on the variability of the data.

The pure significance test

A pure significance test primarily addresses the question: *Is there anything there*? Hackneyed textbook examples are archaic and non-distinctive. Applications which flatter the classical approach include goodness-of-fit and simulation. We shall consider an example of each.

The pure significance test, based on $\Delta(D; H_0)$, refers explicitly to only one hypothesis H_0. What constitutes departure from H_0 is subsumed in the form of Δ. It follows that significantly large values of $\Delta(D; H_0)$ evoke implicit *alternative* hypotheses H_1 for which such data D† should be more probable under H_1 than under H_0. For given Δ, only such alternatives are potentially 'visible'.

This ability to invoke alternatives is a valuable aid to modelling, and is fundamental in the context of an actual investigation. In this respect the classical approach shares the spirit of modern 'Data Analysis'—indeed was its precursor—and is chiefly marked by explicit dependence on a calculated significance level α (largely eschewed in the procedures of 'Exploratory Data Analysis').

Example 1. Goodness of fit and the χ^2 test

Data $D_0 = (n_1, \ldots, n_k)$ are the numbers out of n falling into each of k categories C_i $(i = 1, \ldots, k)$. The hypothesis H_θ asserts that

A_1 the n items assort independently of each other

A_2 the probability that an item falls in C_i is $p_i(\theta)$

A_3 this probability is the same for all items.

The goodness of fit question is whether the data-frequencies $\{n_i\}$ are compatible with H_θ for some θ, i.e. to test the composite hypothesis $H = \{H_\theta : \theta \in \Omega\}$. For H_θ define the discrepancy

$$\Delta(\theta) = \chi^2(\theta) = \sum_{i=1}^{k} \frac{(n_i - np_i(\theta))^2}{np_i(\theta)}$$

and for H let the discrepancy be

$$\Delta(D; H) = \min[\, \Delta(\theta) : \theta \in \Omega \,]$$

Then it is a classical result that if for some θ all of A_1, A_2 and A_3 hold, then $\Delta(D; H)$ has, for large n, asymptotically the mathematically defined χ^2 distribution on $\nu = k - 1 - r$ degrees of freedom, where r is the dimension of Ω. Symbolically:

$$A_1 \wedge A_2 \wedge A_3 \;\Rightarrow\; \Delta \sim \chi^2_\nu .$$

Note that $\Delta(D; H)$ is invariant under permutation of the category labels $\{i\}$. Clearly Δ is formulated to directly express differences between $\{n_i\}$ and $\{np_i(\theta)\}$. If A_1 and A_3 hold, but the probabilities are $\{\pi_i\}$, then (asymptotically for large n) the distribution of $\Delta(D; H)$ is the non-central $\chi^2_\nu(\delta^2)$, where the non-centrality is

$$\delta^2 = n \sum_{i=1}^{n} \frac{(\pi_i - \hat{p}_i)^2}{\hat{p}_i}$$

and the $\{\hat{p}_i\}$ may be taken as the $\{p_i(\theta)\}$ that minimise δ^2.

Significantly large Δ are evidence against H. If A_1 and A_3 are maintained, variation of H corresponds to variation of A_2, hence of δ^2. H corresponds to $\delta^2 = 0$, and the usual usage of the χ^2 test of goodness of fit amounts to a significance test of the null hypothesis $\delta^2 = 0$. Dually, the values of δ^2 which, as null hypotheses, are not rejected at‡ significance level α for given data D_0 form a confidence set for δ^2 at level $p = 1 - \alpha$, and can be re-expressed in terms of a confidence region for $\{\pi_i - p_i\}$, i.e. for the degree of departure of $\{\pi_i\}$ from H. Likewise, a confidence set for $\{\pi_i\}$ is the set not rejected when adopted as values of $\{p_i(\theta)\}$.

†[not in the original: '(i.e. data D such that Δ is large)'] ‡[was as in the original]

Less familiarly, rejection of H because Δ seems not to have the χ^2 distribution negates the above implication and implies the negation of it first term, viz.

$$\text{not}\,(\Delta \sim \chi^2_\nu) \Rightarrow (\text{not } A_1) \vee (\text{not } A_2) \vee (\text{not } A_3),$$

expressing the fact that when a hypothesis is under test *every* element is vulnerable, not just the one of prime interest. The formal symbolic negation is a list of all the ways (not mutually exclusive) for the hypothesis H to fail and, again, the classical hypothesis test's capacity to explicitly generate alternatives is a useful and powerful aid to modelling.

A classic instance of the latter reasoning is Fisher's (1936) re-examination of Mendel's data. Mendel's inheritance hypothesis implied whole-number ratios for expected numbers of phenotypes, such as 3:1, 9:3:3:1 and 27:9:9:9:3:3:3:1 for uni-, bi- and tri-factorial heterozygous crossings respectively, on the assumption that the parental genotypes are known for certain. Using χ^2, Fisher evaluated the goodness of fit of Mendel's observations, and obtained small χ^2 values such as would be *exceeded* on typically more than 95%, and, overall, on more than 99.9% of occasions. This fit is 'too good to be true', and H should be rejected. But A_2 is certainly not contradicted; therefore A_1 or A_3 must go (or both). When allowance is made for uncertainty of genotype due to finite numbers of test progeny, different frequencies are to be expected from which the observed frequencies now *differ* significantly. Fisher concludes that Mendel's results were, in one way or another, falsified so as to agree closely with his expectations. Such a process is a failure of A_1.

Example 2. Quantograms and simulation

The significance-test formalism can be used for data where there is no clearly appropriate objective sampling model, if applied to results of suitable computer simulations. The illustrative example will be Kendall's 'cosine quantogram' originally applied to the search for a possible quantum of length (the 'megalithic yard') in the diameters of megalithic stone circles (Kendall, 1974), as proposed by Thom (see, e.g., Thom, 1955, 1967). In this particular application the 'experiment' is *intrinsically* unrepeatable (as historical events generally are). The data are diameters $(X_1, \ldots, X_N)$, and the simplest quantal hypothesis is of the form

$$X = Mq + \varepsilon$$

where M is an integer, q the 'quantum' of length, and ε a perturbation. An alternative non-quantal hypothesis is that X is distributed somehow 'smoothly' over the whole range. Testing one against the other by conventional means requires precise specification of the distributions, which could be discussed at length and inconclusively (this question is thoroughly treated by Kendall).

Alternatively, consider working with a measure of discrepancy between the data and a 'non-quantal' hypothesis H, where Δ is chosen to immediately reflect a potential quantal structure. The implied near-periodicity of the distribution of the X-values suggests the 'cosine quantogram'

$$\phi(\tau) = \sqrt{(2/N)} \sum_{j=1}^{N} \cos(2\pi X_j \tau)$$

where $\tau = 1/q$. Thus $\phi(\tau)$ is effectively the Fourier cosine transform of the sample, and is the real part of the empirical characteristic function. The cosine quantogram of a lattice distribution (i.e. one with an exact quantum) will give a δ-function peak at the corresponding frequency τ. If there is no quantum effect, $\phi(\tau)$ will be distributed (for large N) like $N(0,1)$ for any fixed τ. The range (τ_0, τ_1) of τ-values should correspond to

the range of *a priori* reasonable q-values (in this application, say $q = 1$ ft to $q = 10$ ft) but should be no wider than necessary. The computed course of the empirical cosine quantogram can then be taken as *derived primary data* (D_0) for further analysis (Fig. 1).

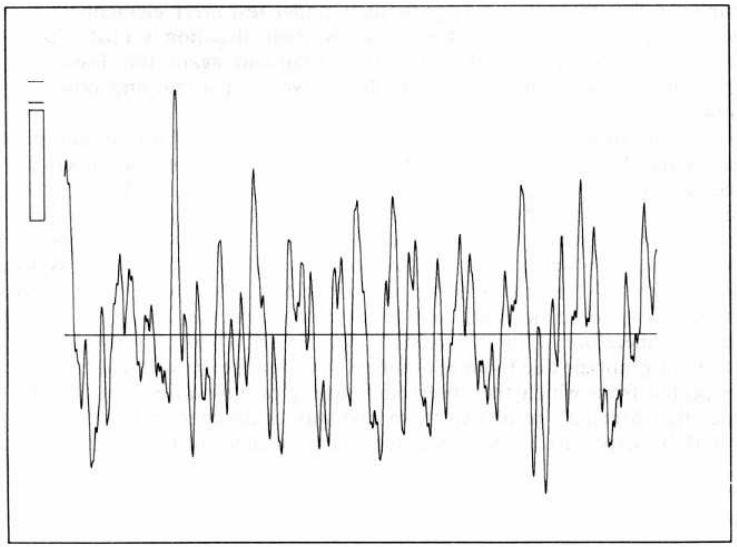

Fig. 1. The empirical cosine quantogram $\phi(\tau)$ computed from diameters of 169 stone circles from Scotland. England and Wales (Tables 5.1 and 5.2 of Thom (1967)—raw, not 'unrounded' data) for τ ranging from 0.09 to 0.59. The highest peak occurs at $\hat{q} = 1/\hat{\tau} = 5.44$ ft. On the left of the figure the two horizontal bars show the two greatest values of maximum peak height in 200 simulations (From Fig. 9 of Kendall, 1974), and the rectangle shows the range of values of maximum peak height attained in the remaining 198. Thus the maximum peak height for the real data was exceeded only once in 200 simulation trials of the 'non-quantal' hypothesis (note that Kendall used 'unrounded' data—*loc.cit.*).

It is now natural to take as measure of discrepancy $\Delta(D_0; H)$ between the empirical quantogram D and a non-quantal hypothesis H the quantity

$$\Delta(D_0; H) = \max[\phi(\tau): \tau_0 \le \tau \le \tau_1].$$

Calculation of D does not depend on specifying H. However, in order to evaluate its significance a precise specification must be made. The approximate $N(0,1)$ distribution for *fixed* q cannot be used for a value $\hat{q}$ maximising $\phi(1/q)$, since $\hat{q}$ has been selected by the data. Kendall's solution was to take "a 'random' set of data, similar in all respects to the actual data save only in definitely *not* having any underlying quantal effect", and subject it repeatedly to simulation, where the stated qualifications amount to

(i) using the same N (though this is not critical),

(ii) ensuring a similar coarse-grained structure, as exhibited by a spline-transform smoothing of the raw data.

Then random data $D = (\xi_1, \ldots, \xi_N)$ are repeatedly generated according to a suitable distribution, and $\Delta(D; H)$ is calculated each time. The situation of the observed $\Delta(D_0; H)$ within the distribution of the simulated values then provides an estimate of

$$\alpha = P_H[\Delta(D; H) \ge \Delta(D_0; H)].$$

Kendall (with many refinements of the above argument) then obtains α = ca. 0.01 to α = ca. 0.07 for various sets of stone circles in England and Wales, and in Scotland, the observed peak quantogram generally occurring at $\hat{q}$ = ca. 5.4 ft. He can thus conclude that although the evidence that there *is* a quantum is less conclusive than one would wish, it is nonetheless strong enough to justify the expense of improved and more accurate survey of the sites. He also repeats the simulation with artificial data simulated according to a *quantal* hypothesis, and obtains an estimated standard deviation for $\hat{q}$ of 0.0181 ft., all 25 simulated estimates being in the range (5.41, 5.50), i.e. within 1 inch.

The question received a Bayesian analysis by Freeman (1976). He states that "a Bayesian approach is unable to encompass a clear test of whether or not a quantum exists, owing to the lack of an alternative model for the data in the absence of a quantum". The conclusion may be right, but surely such an alternative model is just what Kendall used (granted an element of judicious choice in the precise form of the "null distribution").

The fundamental difficulty is absence of a prior probability (π_0) that the quantum exists. A Bayesian analysis could at best offer π_1 (the posterior probability) as a function of π_0, in the typical form

$$\pi_1 = \frac{\pi_0 B}{A + \pi_0(B - A)}$$

† where

$$A = P(\{X_j\} \mid \bar{Q})$$

is the probability of the data assuming no quantum ($\bar{Q}$) and

$$B = \int P(\{X_j\} \mid Q; q)\, g(q)\, dq$$

is the probability of the data assuming a quantum, in which $g(q)$ is the prior distribution of the quantum q and $P(\{X_j\} \mid Q; q)$ is the distribution of the data given a quantum equal to q. If there is strong evidence of any kind in the data for the existence of a quantum, then π_1 will be near 1 except for small values of π_0. This explicit trade-off between π_0 and π_1 is a potentially useful feature of the Bayesian approach. On the other hand, the significance level of a 'classical' test of the no-quantum hypothesis is an absolute figure, appropriate for assessing what is, *a priori*, simply an open question.

As for estimation of q, assuming a quantum to exist, the results of Freeman and of Kendall agree closely, both in estimated values and in estimated precision. Reasons why this should be so are given by Silverman (1976), who shows that Freeman's posterior density is closely related to Kendall's quantogram.

Example 3. Transient effects in experiments yielding time-series

The classical approach is especially—perhaps uniquely—appropriate when deliberate randomisation has been done. The use of the 'randomisation distribution' in analysing data from standard randomised experimental designs is well known; this example will exhibit its use in a less tractable application, the typical context being a biological or medical experiment to study the effect of a treatment on a time-dependent quantity.

Consider a medical experiment in which say blood pressure, breathing rate or urine production is being monitored, or a biological or psychological experiment where the variable being observed is say rate of occurrence of neural action potential, movement of an animal, rate of performance of a task, or some other measure of activity. If the

† Here, and in the following lines, X_i (original) has been changed to X_j for consistency.

variable is simply passively recorded, the result will be a fluctuating time series which may have a complex structure for which there is no obvious model. At some moment, the experimenter intervenes—to inject a drug, apply a stimulus, etc. Does the treatment have an effect? If so, it may be transient, and difficult to distinguish from a fluctuation that might have occurred anyway at that moment.

If there is a model for the expected effect, and also a model for the serial stochastic structure of the series, then standard approaches to signal detection in the presence of known noise could be used. Suppose such models are not available.

The investigator, examining the data series, may be able to identify certain features of the record that look like possible effects; should these occur within a reasonable time lag after the treatment, there is a suggestion—but no more—that the treatment has had an effect. How can a significance level (α) be associated with such an observation? The problem is that there is no known distribution to use for calculating α.

But in many experiments one can be imposed. Let $(0, T)$ be the interval of observation of one individual subject. Let (T_1, T_2) be within $(0, T)$. Apply the treatment at a random moment X chosen according to a distribution $f_0(x)$ on (T_1, T_2).

Next, let the entire record be analysed by any means likely to respond to the presence of an effect in the record. This may be as simple as a running mean, or some more complex filter. In some cases, the investigator himself must, by inspection, judge whether a given stretch of the record is potentially an effect (but if it is done this way, the moment of treatment, X, must be concealed from him). In any case, the result will be a series of measurements $(t_1, \ldots, t_n)$ marking the onsets of events in the record which are *candidate effects*. The $\{t_i\}$ are the statistics, determined from the record and from background knowledge, and from nothing else.

The analysis now turns on the intervals $\{(X, t_i)\}$. Normally (except in situations where the subject might anticipate the treatment-event) moments t_i with $t_i < X$ can be excluded as not corresponding to possible effects. Consider the remainder, and define the lags

$$\tau_i = t_i - X.$$

Suppose there is a (possibly parametrised) distribution $\psi(t; \theta)$ for the time lag τ between treatment and effect (if present). In many applications, a reasonable distribution for the lag may be more readily available than explicit models for the effect or for the time series. On the null hypothesis (H_0) that the treatment has no effect, X has the null distribution $f_0(x)$ on (T_1, T_2). On an alternative hypothesis (H_1) that with parameter value θ one of the t_i marks the true effect, and *conditional on the entire record*, the distribution of X is given by

$$f_1(X) = \frac{\sum f_0(X)\, \psi(t_i - X; \theta)}{\int \sum f_0(X)\, \psi(t_i - X; \theta)\, dX}$$

$$= f_0(X) \cdot G(X; \theta), \text{ say,}$$

[where $\psi(t)$ is the distribution of the lag, *(Added 8 November 2009)*]
summing over the remaining t_i and integrating over (T_1, T_2).

Using the Neyman-Pearson Lemma, the most powerful test of H_0 versus H_1 is to reject H_0 if the likelihood ratio $f_1(X; \theta) / f_0(X)$ is sufficiently large, i.e. if $G(X; \theta)$ exceeds a critical level λ. Choose λ so that the integral of $f_0(X)$ over the x-values for which $G(X; \theta) \geq \lambda$ is equal to the desired size of test, α.

When θ is unknown and has to be estimated from the data, maximum-likelihood estimation may be used, choosing θ to maximise $G(X; \theta)$; it may be necessary to use records from several subjects. A desired size α may sometimes be approximately achieved as above, integrating $f_0(X)$ over $\{X : G(X; \theta) \geq \lambda\}$ for $\theta = \hat{\theta}$. In general, allowance must be made for the dependence of $\hat{\theta}$ on X. *Conditionally on the record*, i.e. on the $\{t_i\}$,

this can be straightforwardly achieved by simulation since, with fixed $\{t_i\}$, there is no need to simulate different possible records.

Further complications include the possibility that an effect might be masked by an intrinsic (treatment-independent) event in the record. This could be approached by incorporating a probability p of such masking as an additional parameter to be estimated, but then many records may be required.

In the simplest possible case, there is only one candidate effect time t, θ is known, and X has been chosen according to a uniform distribution on (T_1, T_2). The likelihood ratio criterion is then $\psi(t-X) \geq \lambda$. Let L_λ be the total length of the set on which this is satisfied; the size of the test is $\alpha = L_\lambda / L$, where $L = T_2 - T_1$. In this case the dominant design consideration becomes clear: L should be as large as possible.

Confidence sets for parameters

Turning to confidence sets, the next example will show how the duality between hypothesis tests and confidence sets can be applied to produce a solution in classical terms to a problem which has, in the past, generated considerable controversy among 'classical' statisticians. It will also exhibit a typical likelihood-based analysis.

Example 4. Calibration

Data $(x_1, y_1), \ldots, (x_n, y_n)$ are obtained. The x-values are, typically, controlled (i.e. there is no stochastic model for their values). The corresponding y-values are random, according to

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad (i = 1, \ldots, n)$$

where the ε_i are independent 'random errors', each with $N(0, \sigma^2)$ distribution.

A further independent observation Y is made of y, corresponding to an unobserved value X of x, for which

$$Y = \alpha + \beta X + \varepsilon$$

It is required to infer plausible values for X, given the data $(x_1, y_1), \ldots, (x_n, y_n)$ and Y.

An obvious estimate of X is obtained by 'reading the regression in reverse' as $\hat{X} = (Y - \hat{\alpha}) / \hat{\beta}$, but its sampling distribution has neither mean nor variance. An early discussion of the problem is in Berkson (1969). Krutchkoff (1967) suggested the apparently improper procedure of regressing X on Y, and began a controversy which still rumbles. It has been treated by a Bayesian approach in Hoadley (1970) and in Hunter & Lamboy (1981), an empirical Bayes approach in Lwin & Maritz (1982), by likelihood in Minder & Whitney (1975) and by structural inference in Kalotoy (1971). Brown (1982) reviews the controversy and the various approaches and considers the multivariate case; his result (1.2) is equivalent to a result derived below and underlies the solution proposed here which is a directly derived likelihood-based confidence interval.

The unobserved X will be treated as if it were the value of a parameter in the problem, so that there are four parameters, viz. $(\alpha, \beta, \sigma^2, X)$. A set of plausible values for parameters will be interpreted as a set of values, all with sufficiently high values of a likelihood function for those parameters, given the data.

The likelihood function for X will be obtained here as

$$L(X) = \{\sup_{(\alpha,\beta,\sigma^2)} L(\alpha, \beta, \sigma^2, X)\} / \{\sup_{(\alpha,\beta,\sigma^2,X)} L(\alpha, \beta, \sigma^2, X)\}$$

(Maximum Relative Likelihood or Profile Likelihood). On the hypothesis that X has

any specific value X_0, the likelihood $L(X_0)$ has a fixed distribution. $L(X_0)$ is in fact a function of a ratio of quadratic forms in the 'errors' $\varepsilon_1, \ldots, \varepsilon_n$ and ε, and does not depend at all on the true values of α, β and σ^2. Therefore, on the hypothesis that $X = X_0$, and for any given value L_0 between 0 and 1, the probability

$$P(L(X_0) > L_0) = p(L_0)$$

depends only on L_0. If, therefore, L_0 is chosen so that $p(L_0)$ is any desired confidence level p_0, say, we can calculate from the data the limits on the set of X-values such that $L(X) > L_0$. Then this set of X-values contains the true X-value, X_0 (whatever it may be), if and only if $L(X_0) > L_0$; an event which has known probability p_0. Therefore the set of X-values such that $L(X) > L_0$ is a confidence set for the true value X_0, at confidence level p_0, and contains all X-values which, given the data, have sufficiently high likelihood. It can appropriately be called a *likelihood-based confidence set*.

This confidence set will usually be a simple (and preferably short) interval, but sometimes it may consist of two disjoint semi-infinite intervals, or even the whole real line. This arises when the original data are insufficient to establish a well-defined calibration, or when the Y-value is well outside the range for which the calibration was established. In any case, it is a clear indication that the data are inadequate to give a satisfactory estimate of the unknown value of X corresponding to the observed Y-value.

For the given data, assuming any given parameter-values $(\alpha, \beta, \sigma^2, X)$, the likelihood is

$$\sigma^{-(n+1)} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 - (Y - \alpha - \beta X)^2 / 2\sigma^2 \right\}$$

For simplicity, set the origins of x and y at the means, $\bar{x}$ and $\bar{y}$, of $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ so that the x_i and y_i sum to zero. The Maximum-Likelihood Estimates (MLEs) of α, β, σ^2 and X are then given by

$$\hat{\alpha} = 0$$

$$\hat{\beta} = \sum_{i=1}^{n} x_i y_i \Big/ \sum_{i=1}^{n} x_i^2$$

$$\hat{X} = Y / \hat{\beta}$$

$$\hat{\sigma}^2 = \frac{1}{n+1} \sum_{i=1}^{n} (y_i - \hat{\beta} x_i)^2$$

On the other hand, for an arbitrary fixed value of X, the MLEs are given by

$$\tilde{\alpha} = \frac{1}{D} \left\{ Y \sum_{i=1}^{n} x_i^2 - X \sum_{i=1}^{n} x_i y_i \right\} \qquad [\dagger]$$

$$\tilde{\beta} = \frac{1}{D} \left\{ (n+1) \sum_{i=1}^{n} x_i y_i + nXY \right\}$$

$$\tilde{\sigma}^2 = \frac{1}{n+1} \left\{ \sum_{i=1}^{n} (y_i - \tilde{\alpha} - \tilde{\beta} x_i)^2 + (Y - \tilde{\alpha} - \tilde{\beta} X)^2 \right\}$$

where

$$D = (n+1) \sum_{i=1}^{n} x_i^2 + nX^2$$

[†] The original had $\tilde{\alpha} = \frac{1}{D} \left\{ \sum_{i=1}^{n} x_i^2 - X \sum_{i=1}^{n} x_i y_i \right\}$ which wrongly omitted the 'Y'.

The likelihood ratio is therefore now given by

$$L(X) = (\hat{\sigma}/\tilde{\sigma})^{n+1}$$

and so the construction of the confidence region as above depends on the properties of the ratio

$$R(X) = \hat{\sigma}^2/\tilde{\sigma}^2.$$

A confidence region for X_0 will be the set of X-values such that $R(X)$ exceeds a given value R_0, and the confidence level p_0 associated with it will be the probability

$$P(R(X_0) > R_0) = p_0,$$

calculated on the hypothesis that the true value of X is X_0.

It can be established that

$$(n-2)\frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2} = (n-2)\frac{1-R(X)}{R(X)}$$

has the F-distribution $F_{1:(n-2)}$. Therefore the hypothesis that the unknown x corresponding to the observed Y has a particular value X can be tested using this distribution: a significantly large value of this ratio is evidence against the hypothetical value X. The set of all X-values not rejected at a given significance level α constitutes a confidence set of values of X, acceptable at a confidence level $p = 1-\alpha$.

Table 1. Fisheries research data used for Example 4 (Calibration) and Fig. 2. Estimated spawning biomass (x) and larval abundance index (y) in successive years. Given for illustrative purposes only.

Spawning biomass (000 tonnes) (x)	Larval abundance index (y)
2130	484
2210	372
1940	204
1230	112
150	88
670	52

As an example we shall analyse the hypothetical case in which the data of Table 1 are given, together with an 'observation' $Y = 500$, and we seek a likelihood function and confidence intervals for the corresponding X.

Fig. 2 is a composite diagram in which the data and the results are displayed. The data are plotted relative to the (X, Y) axes, and the likelihood function $L(X)$ given by (3.5) is plotted with $Y = 500$ as a baseline, relative to a scale for $L(X)$ on the right of the figure. The 50%, 75%, 90%, 95% and 99% confidence intervals were calculated. Their boundary points correspond to levels of $L(X)$ which are shown on the figure, labelled with the corresponding confidence levels and projected down onto the base line to show the X-values.

Example 5. Estimating stock-recruitment relationships in fisheries research

In modelling the population dynamics of fish stocks, an important factor is the dependence of future recruitment (R) on current stock level (N). A convenient general form is

$$R = \alpha(N) \cdot N.$$

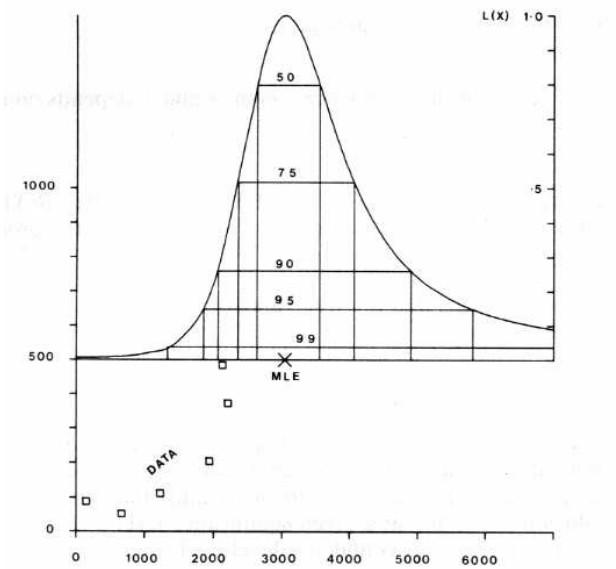

Fig. 2. Likelihood function $L(X)$ for the calibration problem. The data (from Table 1) are plotted in the lower part of the figure. The value $Y = 500$ is given and the corresponding value of X is to be inferred. Supported on the line $Y = 500$ is the graph of $L(X)$ (scale of L on the right) and the MLE of X is marked with '×'. The 50%, 75%, 90%, 95% and 99% confidence intervals for X are shown, as derived from the levels of $L(X)$.

(It may be appropriate to disaggregate N according to age-group, but in many cases this is not necessary.) The multiplier $\alpha(N)$ is a 'density-dependent recruitment rate', and cannot be constant if the stock is to be stable. Furthermore, the degree of non-linearity, as expressed by $\alpha(N)$, is an important index of the capacity of the stock to sustain intensive exploitation. Introductory discussions of this question can be found in Pielou (1977), Hoppensteadt (1982) and Weatherley (1972). Cushing (1975, Chapter 7) discusses it at length.

A variety of parametrised functional forms has been proposed of which many have been summarised by May *et al.* (1978). Their range of behaviours can be emulated by the single form

$$\alpha(N) = \frac{\alpha}{1 + (N/M)^{\beta}}$$

depending on the value of β, which has been discussed by Shepherd (1982). The value of α gives the exponential rate of growth of a small stock, and the parameter M is an index of the 'carrying capacity' of the environment. As stock size $N \to \infty$, recruitment grows indefinitely, or tends upwards to an asymptote (αM), or rises to a maximum and then tends downward to 0, according as $\beta < 1$, $\beta = 1$ or $\beta > 1$ (Fig.3), all being biologically possible. The greater the value of β, the greater the resilience of the stock to† exploitation.

†*was* 'of'

As an element in fishery management, therefore, the estimation of parameters, especially β, must be assessed. The two most important parameters are M and β and, not unexpectedly, they are somewhat linked.

Fig. 4 shows data for North Sea herring for the years 1952–1974 (N in megatonnes, R in billions). The two highest R-values are exceptional and will not be included in the

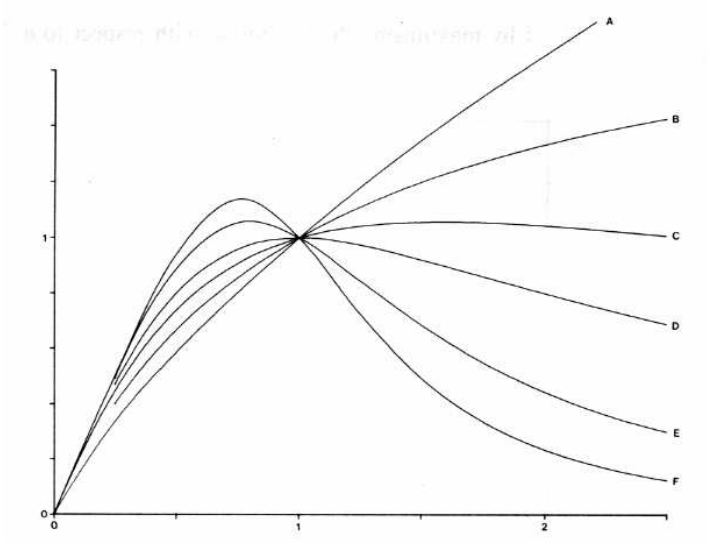

Fig. 3. Curves illustrating the Stock–Recruitment relationship in dimension-free form as $y = x/(1+x^{\beta})$, where $y = (R/\alpha M)$ and $x = (N/M)$.† The curves A–F correspond to values of $\beta = 0.5, 1.0, 1.5, 2.0, 3.0, 4.0$.

† [The original erroneously had $x = (P/M)$. Also, the scale-mark '1' on the Y-axis should be '0.5']

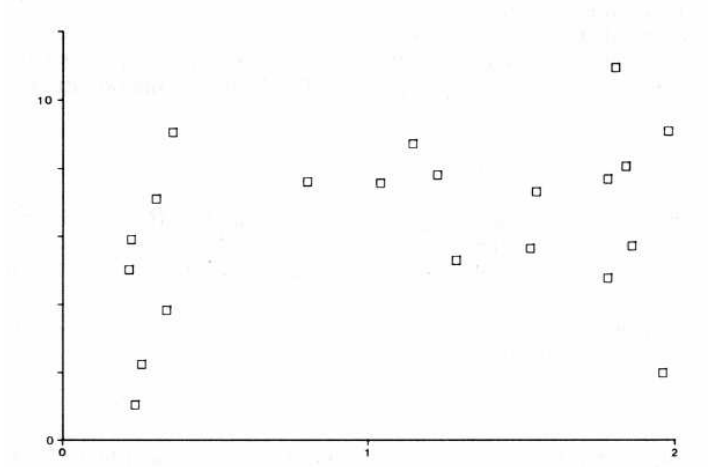

Fig. 4. Stock-Recruitment data for North Sea Herring (1952–1974, except 1956 and 1960), shown for illustrative purposes only. Abscissa: spawning population biomass (megatonnes); ordinate: numbers of recruits ($\times 10^9$).

analysis. The above law was incorporated into a stochastic model assuming multiplicative error, so that the model can be written

$$\log R = \log\alpha + \log N - \log(1 + (N/M)^{\beta}) + \varepsilon \qquad \text{[Notation trivially changed]}$$

Assuming for working purposes that ε has an $N(0, \sigma^2)$ distribution, likelihood contours

for M and β were obtained by maximising the likelihood with respect to α. They are shown in Fig. 5.

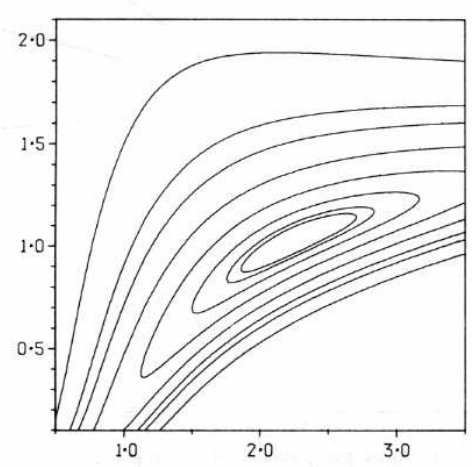

Fig. 5. Likelihood contours for the parameters β and M in the Stock–Recruitment model. Contour levels are chosen from the χ^2_2 distribution corresponding to confidence levels $P = 0.05$ (innermost), 0.1, 0.25, 0.5, 0.75, 0.9 and 0.99. Abscissa: values of β; Ordinate: values of M.

These contours are approximate confidence sets for (M, β) jointly, according to the following argument. Let any given point (M, β) be taken as a null hypothesis H_0; the parameter space then has one dimension (for α). As alternative hypothesis H_1 let (M, β) be unrestricted, giving three dimensions in all. The Wilks likelihood ratio test statistic is

$$\lambda(M, \beta) = 2\{\log L(\hat{\alpha}, \hat{M}, \hat{\beta}) - \log L(\tilde{\alpha}, M, \beta)\}$$

where $\tilde{\alpha}$ is the maximising value of α for fixed (M, β) and $(\hat{\alpha}, \hat{M}, \hat{\beta})$ are the maximising values for (α, M, β) unrestricted. For large samples, $\lambda(M, \beta)$ has asymptotically a χ^2_{ν} distribution, where ν is the difference in dimensions of H_1 and H_0, in this case 2 (Kendall & Stuart II, 1979, section 24.7).

Let $U_{\nu,p}$ be such that $P[\chi^2_{\nu} < U_{\nu,p}] = p$. Then to within the approximation involved in using the asymptotic distribution, if

$$\lambda(M, \beta) < U_{\nu,p}$$

then H_0 is not rejected at significance level $1 - p$, and therefore belongs to a confidence set, confidence level p, consisting of all such values (M, β). Therefore the contours of (M, β) on the figure are the boundaries of approximate confidence sets; on the figure these have been labelled with the values of p derived from $U_{\nu,p}$.

It is clear that M and β are near-aliases, and that acceptable values of β (with corresponding values of M) range from $\beta < 1.0$ to $\beta > 3.0$, since the 90% confidence set extends beyond this range. Considering that β is an important parameter, the stock–recruitment relationship cannot be satisfactorily estimated from this series of herring data. The (M, β) contour diagram permits rather a wide spectrum of curves to be fitted to the data, with strong implications for using the model to assess sustainable yields, and stability of the population.

Classical methods in complex investigations

Example 6. Danish elvers

Boëtius (1976) reports the results of observations on several thousand elvers taken during the 1972 elver run at Højer and Esrom in Denmark. These have been analysed (Harding, 1985) as part of an extended investigation of the spawning of the European eel (*Anguilla anguilla*) (Boëtius & Harding, 1985). Historically, an important issue was whether the entire European eel population was, morphologically, completely homogenous. The example provides a complex and interesting illustration of the deployment of classical methods in probing for possible explanations of a phenomenon.

The Højer samples were taken at three stages of the run (I, II, III) and subdivided according to five developmental stages as determined by pigmentation (A–E).

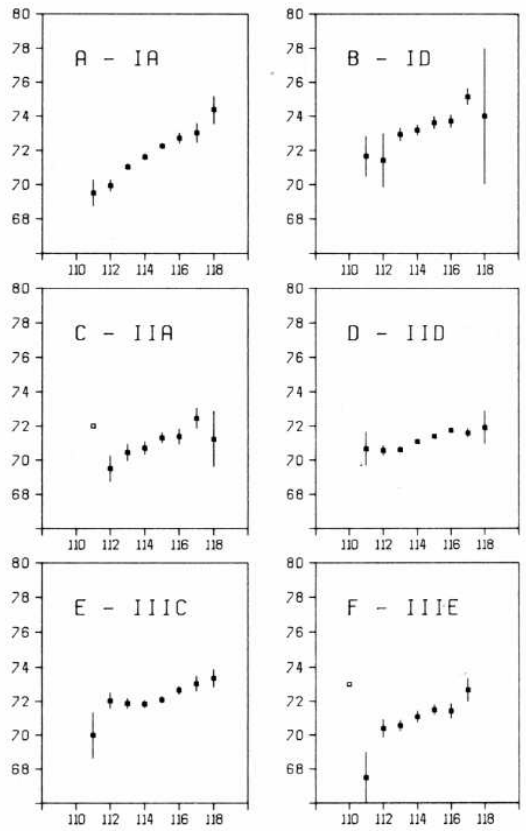

Fig. 6. Data for Højer samples: mean length and SD of the mean for each value of Total Number of Vertebrae (TNV) (filled squares denote mean length; vertical bars extend to ±1 standard deviation of the mean; open squares denote single specimens). Abscissa for each graph: TNV. Ordinate for each graph: length (mm).

Graphs correspond to samples and development stages as follows:

A: Sample I, Stage A C: Sample II, Stage A E: Sample III, Stage C
B: Sample I, Stage D D: Sample II, Stage D F: Sample III, Stage E

Fig. 6 shows mean length (L) vs. total number of vertebrae (TNV) for Højer subsamples IA, ID, IIA, IID, IIIC and IIIE (4289 specimens in all). Fig. 7 shows the same for the Esrom sample (2150 specimens). Several explanations could be suggested for the observed dependence:

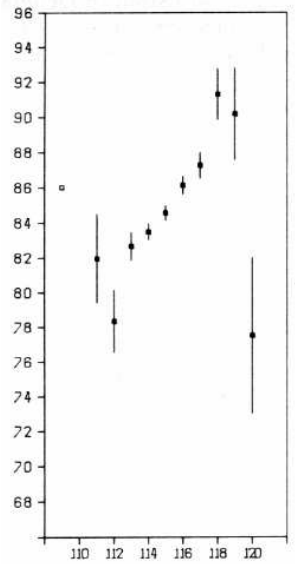

Fig. 7. Data for Esrom sample. Interpretation of the graph as for Fig. 6.

(A) Simple proportionality: the more segments to an individual, the longer it is (other things equal).
(B) Individuals with higher genetic potential for growth have higher numbers of vertebrae.
(C) Numbers of vertebrae increase as growth proceeds.
(D) Application of criteria for determining the tailmost vertebra may sometimes give different results according to length, as a consequence of morphological changes occurring during growth (structures in the tail tip are complex and variable).

The hypothesis that the slopes of the 6 Højer regressions of L on TNV are equal is rejected at $\alpha = 0.006$, and the Esrom slope is significantly very different from the Højer slopes. These results weigh heavily against explanation (A), which is however roughly compatible with the Højer samples.†

† [Also, in general the growth of fish is more a function of their size, activity, food consumption and surface areas of internal organs than of precisely how many vertebrae they have. *Note added 21 October 2009*]

Explanation (C) seems unlikely for so definite a structure as vertebrae, and there is no relationship between mean length or mean TNV and sub-sample, which would be expected if (C) held. Explanations (B) and (D) cannot be directly tested. There is in any case interest in finding an explanation of the data which permits L to be independent of TNV in elvers of common origin. Such a one is:

(E) The Højer specimens are a mixture of distinct groups, where in each group L is independent of TNV, but the groups differ in distribution of length and in distribution of TNV. In such a mixture, L can have a very closely linear relationship with TNV (Fig. 8).

A hypothesis such as (E) is in fact directly suggested by inspection of histograms of length for each TNV (Fig. 9). There is a central peak in a constant position for TNV = 113–116 (mean = ca. 72.5 mm), the length distributions for TNV = 112–114 are

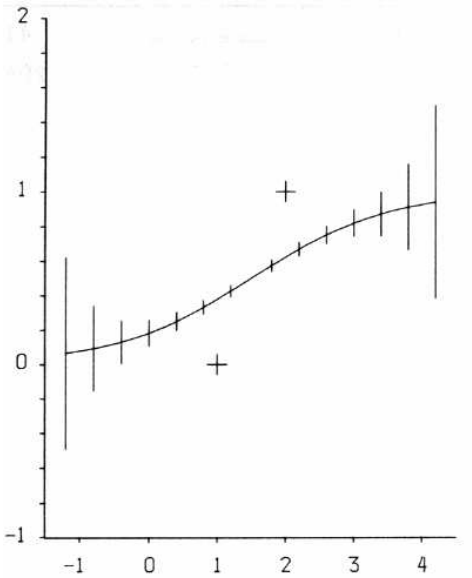

Fig. 8. Graph to illustrate the nearly linear dependence of mean Y on X when (X, Y) is sampled from a mixture of two normal distributions, in each of which X and Y are independent.
The two crosses mark the means of the two distributions in which:
(a) $X \sim N(1, 1)$, $Y \sim N(0, 1)$;
(b) $X \sim N(2, 1)$, $Y \sim N(1, 1)$;
The curve is the expected value of Y given X in a random sample from an equal mixture of (a) and (b). The vertical bars are proportional to the standard deviations of the samples of Y-values when the X-values are equally grouped, taking into account the frequencies of occurrence of X.

negatively skew, and positively skew for TNV = 116–118. The mixture model adopted attributes the negative skewness to a component with relatively low mean TNV and L, and the positive skewness to a component with relatively high mean TNV and L, in each case in the presence of a component with intermediate TNV and L. Thus the probability that a specimen will have length = L and TNV = V can be written as

$$P(V, L) = \pi_1 P_1(V) \cdot Q_1(L) + \pi_2 P_2(V) \cdot Q_2(L) + \pi_3 P_3(V) \cdot Q_3(L)$$

where the π_i are the proportions of the three components, and the $P_i(V)$, $Q_i(L)$ are grouped normal distributions. There are 14 parameters in this model.

Fitting the parameters is complicated by the fact, evident from Fig. 9, that certain lengths are favoured over their neighbours, an artefact probably due to observer bias. Fitting such a model depends delicately on fine detail of the shapes of distributions. Initial approximations were determined by a somewhat elaborate procedure involving probability plotting, components of variance, and fitting by moments (see Harding, 1985), and then an iterative procedure was entered in which the parameters were interactively varied so as to increase the likelihood on each iteration. Care was taken to arrest iteration when the resulting fitted distributions began to be unduly influenced by the artefactual irregularities noted in the length distributions.

The resulting fit is shown superimposed on the original histograms in Fig.10. The

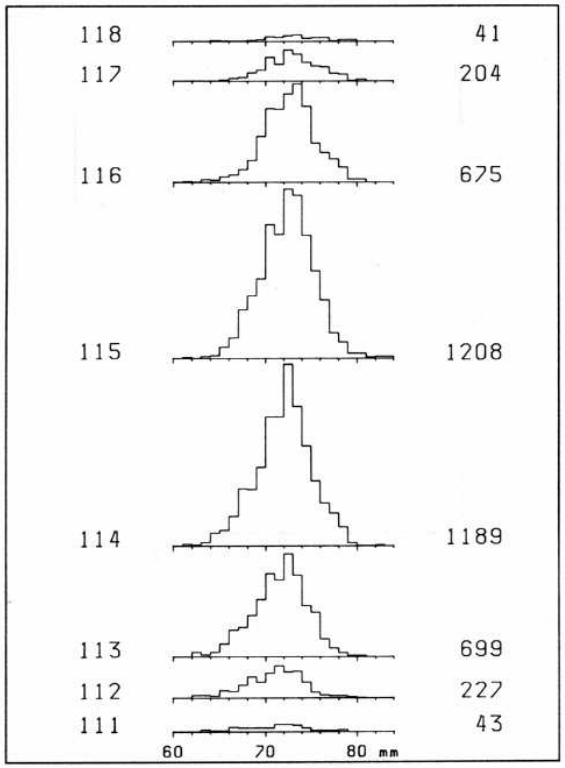

Fig. 9. Histograms of the length distributions for each TNV, for the Højer material (all three samples pooled). On the left: TNV. On the right: observed numbers of specimens for each TNV.† Heights of histogram bars are proportional to absolute numbers. Histogram for TNV=119 (three specimens only) not shown.

shapes are well matched. On the right are expected numbers for each TNV calculated from

$$E_V = 4289(\pi_1 P_1(V) + \pi_2 P_2(V) + \pi_3 P_3(V))$$

The agreement is close—'χ^2' = 4.98 for 8 classes (but on how many degrees of freedom…?).

In conclusion, it is *possible*, at least, to represent the Højer sample as a mixture of three distinct though overlapping components, in each of which length is statistically independent of TNV, and thus account for the dependence of L on TNV. If this tripartite representation *is* the explanation, then all three groups are present in all sub-samples (though, as further analysis can show, in slightly different proportions), since the length–TNV relationship is equally manifest in all.

Therefore, on this hypothesis, the tripartite structure is *not* a consequence of the fact that the Højer specimens were taken on three separate occasions, and would imply that there are present three groups that, at least, have experienced different environments in early development‡ and may possibly have been spawned in different places. Such a hypothesis is of great interest for the major investigation, since Johannes Schmidt's Sargasso Sea theory requires the utmost morphological homogeneity in the entire Euopean population (Schmidt, 1922).

‡ [It is known, experimentally, that the number of vertebrae in a fish of given species is influenced, for instance, by temperature during embryonic development in the egg. *Note added 21 October 2009*]

† [A duplicate of this sentence has been removed]

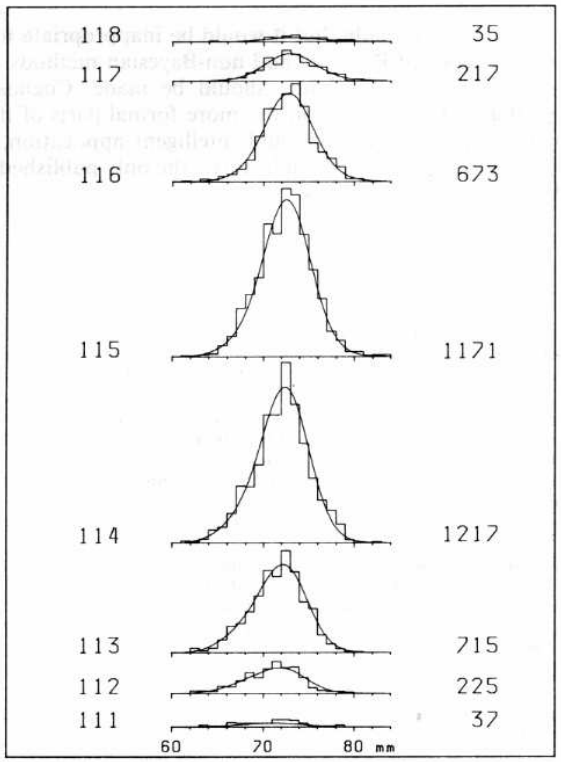

Fig. 10. The fitted 'mixture' model described in the text, superinposed on the histograms of Fig.9. On the right: expected numbers (E_V) calculated as explained in the text.

Discussion

The 'classical' statistical procedures of hypothesis testing, confidence intervals and estimation, based on sampling properties of statistics, often appear as rigidly formulated ends in themselves, with limited scope, and yielding conclusions of limited interest, as encountered in standard courses and textbooks.

The general formulation presented at the start of this paper allows freedom, flexibility and power in applications of classical statistical methods. Their power lies in that they permit sharp focus on specific aspects of an investigation, their flexibility in that they offer a wide range of aspects to examine.

These properties of classical methods just mentioned are valuable tools for the discovery, formulation and verification of feasible models, but their full power and flexibility are only realised when the problem is repeatedly re-examined under different aspects, and the corresponding statistical models are continually re-stated in the real terms of the investigation.

These remarks have been illustrated by a variety of examples. Full discussion of any one of these would exceed the space of the present paper, but I hope that, taken together, they give a good impression of the scope, limitations and, on occasion, dangers of the classical approach.

Some examples were chosen to exhibit the central rôle played by the likelihood function in many applications, conjoined with appropriate sampling theory (as opposed to direct interpretation of the likelihood function itself). Some reference to

Bayesian approaches has been made, but it would be inappropriate to go further into the relative merits and scope of Bayesian and non-Bayesian methods.

Finally, an acknowledgment, long due, should be made. Cognoscenti will have recognised George Barnard's influence on the more formal parts of this presentation. For this, and for his example in flexible and intelligent application, my thanks and appreciation. (Apart from presentations in lectures, the only published account I know is in Barnard (1962), especially pp. 42–49.)

References

BARNARD, G.A. (1962) Prepared contribution to discussion of L.J. Savage's paper, in: L.J. Savage *et al.*, *Foundations of Statistical Inference*, edited by G.A. Barnard & D.R. Cox (London, Methuen).

BERKSON, J. (1969) Estimation of a linear function for a calibration line. *Technometrics*, 11, pp. 649–660.

BOËTIUS, J. (1976) Elvers, *Anguilla anguilla* and *Anguilla rostrata* from two Danish locations. Size, body weight, developmental stage and number of vertebrae related to time of ascent. *Meddelelser fra Danmarks Fiskeri- og Havundersøgelser, New Series*, 7, pp. 199–220.

BOËTIUS, J. & HARDING, E.F. (1985) A re-examination of Johannes Schmidt's Atlantic eel investigations. *Dana*, 4, pp. 129–162.

BROWN, P.J. (1982) Multivariate calibration. *Journal of the Royal Statistical Society Series B*, 44, pp. 287–321.

CUSHING, D.H. (1975) *Marine Ecology and Fisheries* (Cambridge University Press).

FISHER, R.A. (1936) Has Mendel's work been rediscovered? *Annals of Science*, 1, pp. 115–137. [Reprinted in *Collected Papers of R. A. Fisher*, Volume III, pp. 514–536 (University of Adelaide) 1973.]

FREEMAN, P.R. (1976) A Bayesian analysis of the megalithic yard. *Journal of the Royal Statistical Society Series A*, 138, pp. 20–55.

HARDING, E.F. (1985) On the homogeneity of the European Eel population (*Anguilla anguilla*), *Dana*, 4, pp. 49–66.

HOADLEY, B. (1970) A Bayesian look at inverse linear regression. *Journal of the Americam Statistical Association*, 65, pp. 356–369.

HOPPENSTEADT, F.C. (1982) *Mathematical Methods of Population Biology* (Cambridge Studies in Mathematical Biology 4) (Cambridge University Press).

HUNTER, W.G. & LAMBOY, W.F. (1981) A Bayesian analysis of the linear calibration problem, *Technometrics*, 23, pp. 323–328.

KALOTAY, A.J. (1971) Structural solution to the linear calibration problem, *Technometrics*, 13, pp. 761–769.

KENDALL, D.G. (1974) Hunting quanta, *Philosophical Transactions of the Royal Society of London*, 276, pp. 231–266.

KENDALL, M.G. & STUART, A. (1979) *The Advanced Theiry of Statistics*, 4th edn. vol. II (London, Charles Griffin).

KRUTCHKOFF, R.G. (1967) Classical and inverse regression methods of calibration, *Technometrics*, 9, pp. 425–439.

KRUTCHKOFF, R.G. (1969) Classical and inverse regression methods of calibration in extrapolation, *Technometrics*, 11, pp. 605–608.

LWIN T. & MARITZ, J.S. (1982) An analysis of the linear-calibration controversy from the perspective of compound estimation, *Technometrics*, 24, pp. 235–242.

MAY, R.M., BEDDINGTON, J.R., HORWOOD, J.W. & SHEPHERD, J.G. (1978) Exploiting natural populations in an uncertain world. *Mathematical Biosciences*, 42, pp. 219

MINDER, CH.E. & WHITNEY, J.B. (1975) A likelihood analysis of the linear calibration problem, *Technometrics*, 17, pp. 463–471.

Pielou, E.C. (1977) *Mathematical Ecology* (New York, John Wiley).

SCHMIDT, JOHS. (1922) The breeding places of the eel, *Philosophical Transactions of the Royal Society of London Series B*, 211, pp. 179–208.

SHEPHERD, J.G. (1976) A versatile new stock–recruitment relationship for fisheries and the construction of sustainable yield curves, *Journal du Conseil permanent international pour l'Exploration de la Mer*, 40, pp. 67–75.

SILVERMAN, B.W. (1982) Discussion of Freeman (1982). *Journal of the Royal Statistical Society Series A*, 139, pp. 44–45.

THOM, A. (1955) A statistical examination of the megalithic sites in Britain, *Journal of the Royal Statistical Society Series A*, 118, pp. 275–295.

THOM, A. (1967) *Megalithic Sites in Britain* (Oxford, Clarendon Press).

WEATHERLEY, A.H., (1972) *Growth and Ecology of Fish Populations* (New York, Academic Press).