

Rejoinder: Approximating data[☆]

P.L. Davies^{*}

*Department of Mathematics, University Duisburg-Essen, Germany
Department of Mathematics, Technical University Eindhoven, Holland, The Netherlands*

Received 15 April 2008
Available online 11 June 2008

1. Beran

In his discussion Professor Beran has taken the opportunity to take a much broader and wide ranging look at the state, and future, of statistics as an empirical science. He mentions amongst other topics algorithms, randomness, truth, and complexity on which I give some brief comments. However there are two important points on which, I think, we agree. Firstly, the data-generating mechanism is very often unknown. Secondly, the measure of success is the empirical one.

When attempting to explain the source of the ‘randomness’ we see in the world, we are perhaps led to quantum mechanics. The conventional Copenhagen interpretation is that the random collapse of the wave-packet is caused by the act of measurement. The first formal expression with ‘measurement’ replaced by ‘interaction’ was given in [Ghirardi, Rimini, and Weber \(1986\)](#). On the other hand there is the deterministic ‘hidden variable’ version of quantum mechanics due to Bohm (see for example [Bell \(1987\)](#), [Goldstein \(1998a,b\)](#) and [Dürr \(2001\)](#)) which leaves us with no definite answer. Even within mathematics the difference between ‘deterministic’ and ‘stochastic’ systems is not straightforward: deterministic systems such as the prime numbers can exhibit behaviour which can be described in stochastic terms (see [Kac \(1959\)](#) and [Tennenbaum \(1995\)](#)). Finally, although the Kolmogorov concept of complexity can be used to show that the binary expansion of π is non-random, it is not applicable to finite data. The definition is asymptotic and as there is no objective Turing machine, the asymptotics are non-uniform. Attempts to make it applicable to real data such as minimum description length are forced to fudge the issue (see Section 3.6 of [Rissanen \(1998\)](#)).

I come to the more specific points made by Beran.

- His comments on likelihood are certainly true. It is well known that likelihood can behave pathologically and requires regularity conditions to prevent this. Nevertheless, even with all these in place, it remains blind. Whether it is in the coffin is debatable: I have the impression that it is still very much alive. As evidence I cite Nils Hjort’s contribution to the discussion and the books [Pawitan \(2001\)](#), [Burnham and Anderson \(2002\)](#) and [Held \(2008\)](#), the latter arriving on my desk whilst writing the rejoinder.
- The use of the Kolmogorov metric is indeed restricted to the real line; in practice I prefer Kuiper metrics. The half-space metric mentioned by Beran is the obvious higher-dimensional generalization, others are mentioned in

[☆] This research was supported in part by Sonderforschungsbereich 475, University of Dortmund.

^{*} Corresponding address: Department of Mathematics, University Duisburg-Essen, Universitaetsstrasse 3, 45117 Essen, Germany.
E-mail address: laurie.davies@uni-duisburg-essen.de.

Davies (1993). In general any Vapnik–Chervonenkis class of Borel sets will provide a weak metric. For a justification see Chapter 2 of Vapnik (1998).

- I am not sure what Beran means when asking of the relevance of the distribution function, density function distinction for non-random data. Many very different densities can produce the same distribution function to any given degree of accuracy. If the analysis is to be density based, then some form of choice is necessary and this will in general entail a form of regularization. One possibility is a regularization by shape which can be done at the level of distribution functions as in Davies and Kovac (2004).
- There are indeed many bootstrap methods but even when using bootstrap I find the idea of approximation useful. If the bootstrap samples look very different from the original data in a manner which is relevant for the analysis, then I would be unwilling to place much confidence in the result. Moreover to the extent that bootstrap techniques are used for confidence intervals, they must, in addition to being close to the real data, assume a strong form of smoothness, namely differentiability.
- Finally my concept of truth in this paper is at the basic level of whether it rained today or not: these are the ‘everyday truths’ of Williams (2002). As no statistical model has this level of immediacy, I prefer to do without the term. A deeper analysis of truth may have to take account Gödel’s theorem, but I have not given the matter sufficient thought to make any comment on this. At the level of rain today or not, it is possible to do without it.

Beran gives statisticians the Kepler problem as a challenge for a discipline lacking recognized open problems such as the Riemann hypothesis. His description is very brief so that different interpretations are possible. In the Tychonic system the earth was at the centre, the sun and moon revolved around the earth but all other planets revolved around the sun. Within this system Kepler’s three laws are not compatible with the data. Kepler himself placed the sun at the centre with all the planets, including the earth, moving round the sun, which is the Copernican system. If the challenge is to move from the Tychonic system to the Copernican system on the basis of Tycho Brahe’s data alone, it is a formidable one. If however we start from the Copernican system and are then required to deduce Kepler’s three laws of motion, then it may at least be possible to deduce the first two, the movement in ellipses and the speeds of planets. The third one however may require the correct question being posed.

Finally Beran expresses the hope that statistics may become like physics, ‘which made the transition from philosophy to science centuries ago’. If one reads the discussions on interpretations of quantum mechanics and the relevance of string theory (see for example Hooft (1997)), then one sees that philosophy is still alive and well in physics. The intensity of these discussions in the physics community makes statisticians appear well-mannered.

2. Hampel

Hampel states that the concept of approximation is non-parametric. This is true of the examples in the paper but it can be applied to parametric models. Davies (1995) already gives examples of approximation regions for parametric models. Fig. 1 shows two different approximation regions for the parameter (μ, σ) of the normal distribution for the following data:

6.125	−8.375	1.000	2.000	0.750	2.875	3.500	5.125
1.750	3.625	7.000	3.000	9.375	7.500	−6.000	

They are taken from Andrews and Herzberg (1985) and were collected by Darwin and give the differences in inches in growth of cross- and self-fertilized plants. The upper panel of Fig. 1 gives the 95% ‘confidence’ region for (μ, σ) based on the usual t -statistic for μ and on the variance for σ ; 0.025 of probability was ‘spent’ on each statistic. The lower panel shows a 95% approximation region based on the four statistics

- (1) $\sqrt{n}(\text{median}(X_n) - \mu)/\text{mad}(X_n)$,
- (2) $\text{mad}(X_n)/\sigma$,
- (3) $\max_i |X_i - \mu|/\sigma$,
- (4) $\max_{a < b} |F_n(b) - F_n(a) - (\Phi(\sigma b + \mu) - \Phi(\sigma a + \mu))|$,

where F_n denotes the empirical distribution function of the data; an amount 0.0125 of probability was spent on each statistic.

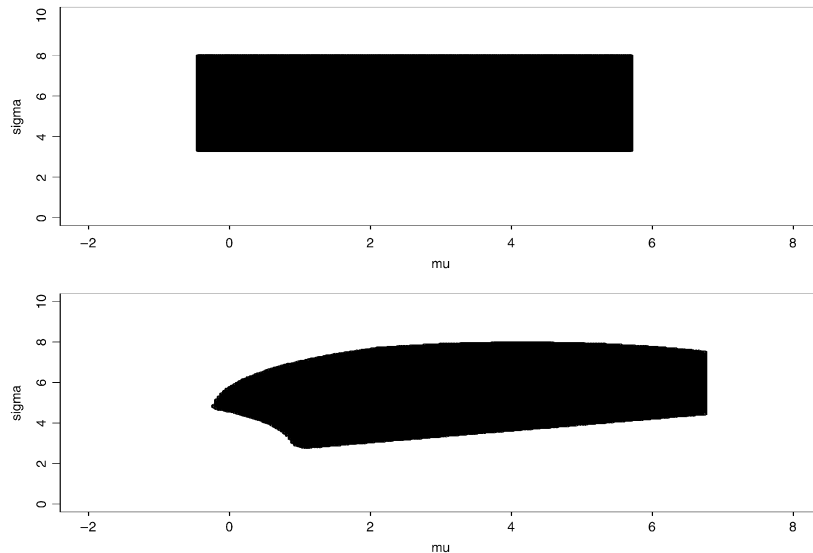


Fig. 1. The upper panel shows a standard 95% confidence region for (μ, σ) for the Darwin data given in the text on self- and cross-fertilized plants. The lower panel gives a confidence region for the same data based on the median, the MAD, the largest deviation and the distance from the empirical distribution function as described in the text.

Hampel mentions the ‘arbitrariness’ of the choice of features in the definition of ‘typical’. Although arbitrary it is a form of constrained arbitrariness. The choices will have to be relevant ones for answering the scientific problem at hand. I can do no better than refer to Tukey’s comment in Section 7 on this very problem.

Professor Hampel indicates his preference for the Prohorov metric and his example seems to support his case. I disagree however. The definition of the Prohorov metric is

$$d_{pv}(P, Q) = \inf\{\varepsilon > 0 : P(B^\varepsilon) \leq Q(B) + \varepsilon, B \in \mathfrak{B}\} \quad (1)$$

where

$$B^\varepsilon = \{x : \text{there exists a } y \in B \text{ with } |x - y| < \varepsilon\}. \quad (2)$$

The ε in (1) plays two roles. In (2) it is a measure of error and will therefore in general have units such as grams or centimeters. The second use in (1) is as a probability which is a dimensionless quantity. The definition of the Prohorov metric conflates the two: in my opinion this disqualifies it. If $\mathbb{P}_n$ is an empirical distribution and P a probability model then $d_{pv}(\mathbb{P}_n, P)$ will depend on the units of measurement making it completely uninterpretable. This can be avoided by considering affinely invariant metrics, that is metrics d for which $d(P^A, Q^A) = d(P, Q)$ for all affine transformations A . The Kolmogorov metric is such a metric. It is not difficult to see that no affinely invariant metric can generate the weak topology. We have a conflict and for me the choice is clear. Although the Kolmogorov metric does not topologize weak convergence it can be used to prove and formulate weak convergence. The prime example is the central limit theorem which is nothing more than the convergence in the Kolmogorov metric. Davies (1993) contains a ‘version’ of the Prohorov metric which uses a proportional instead of an absolute error. Both uses of ε are now dimensionless but the metric is still affine invariant which means that it does not metricize weak convergence. In his discussion Hampel states that we can ‘statistically prove that the “true” model is in a certain (weak-topology) neighbourhood...’. One of my aims in this paper is to produce an account of statistics where we do not need to take recourse to either true or “true”. Even in his example of the binomial distribution, the independence of the trials is beyond any form of statistical proof: the binary expansion of π indicates the problems involved.

Hampel, Ronchetti, Rousseeuw, and Stahel (1986) is indeed one of the few books where approximation extends to models themselves and is not just restricted to some form of numerical approximation as in the central limit theorem. However not all books on robustness are as explicit. The word is not contained in the index of one of the most recent ones written by leading experts in the area, Maronna, Martin, and Yohai (2006). Although on the formal level it is possible to give precise definitions of approximation as is done in the paper, once we move to the practical level of

procedures there is an inevitable vagueness. If we wish to approximate data by a normal distribution as say a summary statistic, then whether this works well or not depends on the problem at hand and the data. There is no possibility of describing such problems or data precisely. At best one can give guidelines, examples and counterexamples. I agree with Hampel on the problem of symmetry but have nothing of my own to add to the comments he made.

Hampel states that robustness is the theory of approximate parametric models and this is indeed the usual definition. I regard any stochastic model as an approximation without any reference to robustness. If however we want the chosen definition of approximation to be stable in that small changes in the data or the parametric model or both do not lead large changes in the analysis, then we are naturally lead to problems of robustness in the sense of Hampel. I point out however that some of the limitations of likelihood given in Section 2 of “Approximating Data” still hold: even robustified likelihood is still blind.

I agree with Hampel about the asymptotics. I had hoped to use the ‘small sample asymptotics’ of Hampel (1973), Field and Hampel (1982) and Field and Ronchetti (1990) for the location-scale functional of Section 11. However a conversation with Elvezio Ronchetti indicated that the results were not applicable. As it is the small sample behaviour for $n \leq 500$ was obtained by means of simulations, which I do not find aesthetically satisfactory even if it works in practice.

My students learn the ‘mean after elimination of the outliers’ as the ‘5.2 Hampel rule’ (Hampel, 1985) and are advised to use it before calculating the standard t -statistic. As Hampel states, smoothly redescending M-estimators behave better, and I indeed have a version of the joint location-scale functional based on such a redescending function. It was not used simply because it takes somewhat more space to define it. I was responsible for the inclusion of Hampel’s three-part redescending M-estimator (Hampel et al., 1986, page 150) in the official German standard, DIN 38402-45, for the conduct of interlaboratory tests for the examination of water, waste water and sludge. I do not think I ever told him this, but he knows now.

3. Pötscher

Professor Pötscher points out that, without further restrictions, the parameters μ and σ in the model $F((\cdot - \mu)/\sigma)$ are not identifiable. One restriction which goes some way to solve the problem is to require that the functional T_L for specifying μ be affinely equivariant as in (27). From this it follows that $T_L(F) = 0$ for any F which is symmetric about 0 and in consequence $T_L(F((\cdot - \mu)/\sigma)) = \mu$ for any such F . In particular $T_L(N(\mu, \sigma^2)) = \mu$. This is as far as one can go. Although returning the location measure of 842 as a summary of 8.2, 7.8, 8.0 may seem odd, the oddness disappears when it is pointed out that the original data are height measurements from street level whilst the summary is measured from sea level. Once we move away from symmetric models then there is a degree of arbitrariness. Tukey writes

For symmetric models we can be concerned with both bias and variance, but all definitions of bias for unsymmetric models still seem essentially arbitrary. We probably do better by allowing finite-sample robust procedures to give us a sort of analytic continuation from the symmetric to the unsymmetric.

I make the following comments:

- (1) If we are interested in differences, as in the one-way table, then *all* affinely equivariant location measures give the same answer. No symmetry is required.
- (2) Suppose we interpret Tukey’s *analytical continuation* as local Fréchet differentiability. If now the data (or model) is nearly symmetric, that is, close to some symmetric distribution in the Kolmogorov metric, then the analytical continuations will also be close.
- (3) The only way of choosing between different analytical continuations is the empirical one of their performance in practice. In the case of interlaboratory tests, where prepared samples can be used to check performance, it seems, as already mentioned in the introduction, that the elimination of outliers followed by the mean works better than the median. This indicates that M-functionals with a redescending ψ -function would, in this situation, be better than the ψ -function of (28).

There is nothing corresponding to symmetry for the scale measure so that the requirement of affine equivariance for T_S as in (27) will never specify $T_S(F((\cdot - \mu)/\sigma))$ uniquely. What does follow is $T_S(F((\cdot - \mu)/\sigma)) = c(F)\sigma$ for all σ and for some $c(F)$. If scale is only a nuisance parameter then the lack of uniqueness does not matter; scale is

required to calculate an approximation interval for T_L but this does not depend on $c(F)$ as it cancels out. If we are interested in a value for the scale then the only possibility I see is to calibrate T_S at some distribution F_0 , that is to make it Fisher consistent at F_0 . For the MAD and taking F_0 to be the Gaussian distribution, this means multiplying by 1.482602 which is the default version of the MAD in R.

4. Hjort

Professor Hjort has given a spirited defence of likelihood to counter the criticisms of Section 2. An adequate response to his arguments would require much more space than is available, but it is at least possible to make the disagreements clearer. Such disputes are often better conducted over a glass of wine and there is no doubt that I will have the opportunity of doing this with Nils Hjort, but for the moment I have only his written position with the possibility of misunderstandings.

Suppose we have data such as the copper data and a parametric model, P_θ and we wish to estimate θ on the basis of the data $\mathbf{x}_n$. If $\mathbf{x}_n$ was generated under P_{θ_0} then the problem is well posed. If however $\mathbf{x}_n$ was not generated under P_θ for any θ we are trying to estimate something which is not there. The position taken in my paper is to distinguish between the parameter θ and the state of the world. In the copper example we are trying to estimate the quantity of copper, not some parameter in a statistical model. As far as the parameter is concerned we try to specify those parameter values, if any, which approximate the data. In a speculative act and assuming that some parameter values are consistent with the data, we associate one with the other. The success or failure is an empirical question. The answer given by Hjort, as I understand it, is that we are indeed trying to estimate a parameter, namely that one which is closest to the truth. This at least makes one difference clear; I am trying to find which θ if any are consistent with the data, he is trying to get a value θ for which $f(\cdot|\theta)$ is close to the truth g behind the data. The trouble with the truth behind the data is that it is very complicated. In the copper example the truth will consist of a series of complicated physical and chemical interactions at quantum level, and the final result will be affected amongst other things by the cleanliness of the container for the sample, the matrix of other chemicals in the sample and the skill of the operator. There is no possibility of this corresponding to a ‘true data-generating density g' ’. The conclusion is that we cannot make any sense of estimating that parameter ‘closest to the truth’ for the truth is complicated and unknown. What Hjort proposes is something else. He has two models, both densities, g and $f(\cdot|\theta)$, one for the truth and the other for the data. However let us accept this in the test-bed sense and examine it in this context. We take the parametric model to be the Gaussian model and consider two mechanisms for the truth, both of which generate i.i.d. data with mean zero and variance 1. For each data set we are allowed only to calculate the likelihood function. As this depends only on the mean and variance of the samples the likelihood gives us no way of deciding which mechanism was used to generate the data. I interpret Hjort’s position as indicating that, although this is true, in both cases we are getting as close as we can to the underlying truth in the sense of Kullback–Leibler. We now reveal that the generating mechanism 1 is the standard $N(0, 1)$ distribution, and that mechanism 2 generates random variables taking the values 1 and -1 with equal probabilities. In case 1 the divergence is finite and even better than that: in Kullback–Leibler sense the truth is on your doorstep. In case 2 the divergence is infinite, further than the furthest galaxy. Not only that, you do not know which is which. Getting as close as possible to something which is infinitely far away is not an argument I find convincing. At first glance this may seem to be an extreme example but it is not: it applies to any situation where the model is continuous and the generating mechanism is discrete. Examples are counts such as the thin-film data or in finance or internet-traffic data, or indeed the copper data. Copper comes in discrete units, atoms, and water comes in discrete units, molecules. The truth behind the copper data is discrete, the number of copper atoms per 10^{23} water molecules and for any continuous model such as the Gaussian, the Kullback–Leibler discrepancy is infinite. A final example: if the truth is $\text{binom}(10000, 0.5)$ and the model Gaussian, the Kullback–Leibler distance is infinite, the Kolmogorov distance to the $N(5000, 2500)$ is 0.00398. Which has more statistical importance, the former or the latter? To make sense of Kullback–Leibler one requires ‘weak conditions’ or ‘natural and weak conditions’ where the truth is not only required have a density g , which in turn is sometimes required to have a continuous third derivative $g^{(3)}$. Such conditions are neither weak nor natural. Davies (1987) requires the existence of a strictly monotone differentiable density to prove uniqueness of the S-estimator of covariance. It was the implausibility and non-testability of this assumption (and similar ones in Davies (1990, 1992)) for real data which caused me to question their innocuousness. Even now some members of the robustness community have no qualms about assuming densities, the assumptions

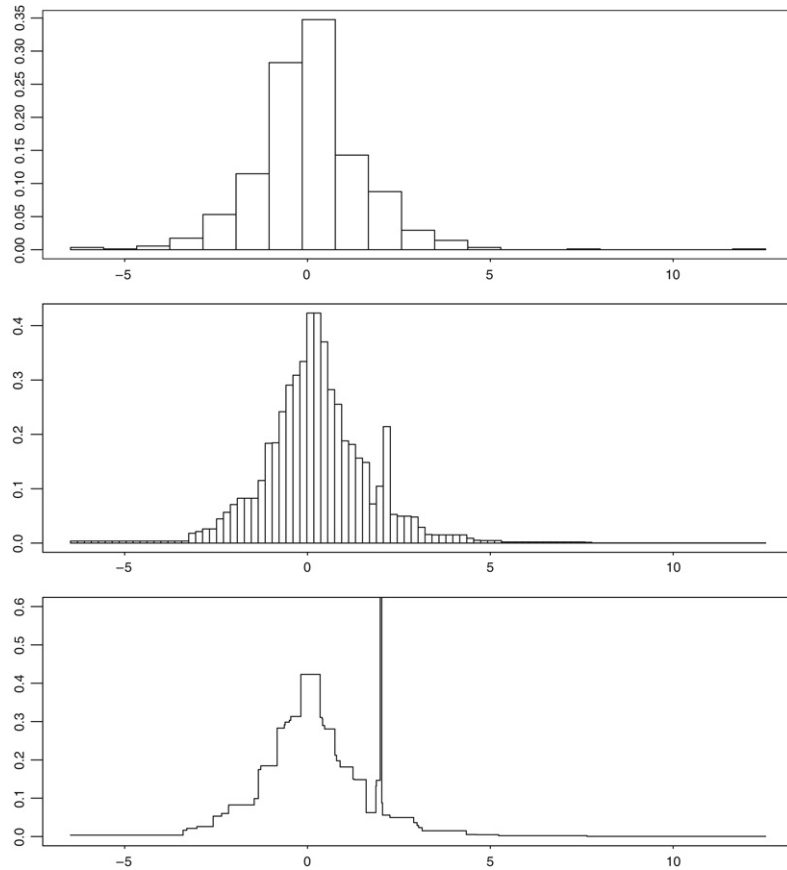


Fig. 2. Three histograms for a sample of 1000 Laplace and 20 $N(2, 0.01^2)$ random variables. The top histogram is chosen by AIC, the bottom is the default taut string histogram. The centre 'histogram' derives from the bottom but with 100 equally spaced bins.

also being described as 'weak'. They are not. They are an admission of failure and a warning that something has gone wrong. All this is, of course, just a variation of (i).

Hjort mentions densities at several points in his response, including the use of histograms. He writes that 'a statistician looking at a histogram and deciding that the data are not normal (e.g. by visible skewness, or by having two clear modes) has used a density-based goodness-of-fit test'. This may or may not be true, it depends on the histogram. Suppose we have data $\mathbf{x}_n$ and wish to specify a density $f(\cdot|\boldsymbol{\pi})$, $\boldsymbol{\pi} = (\pi_1, \dots, \pi_k)$ which is a mixture of k uniform distributions on the intervals $[x_{(1)} + i(x_{(n)} - x_{(1)})/k, x_{(1)} + (i+1)(x_{(n)} - x_{(1)})/k]$ for $i = 0, \dots, k-1$. One method of choosing k is to use AIC. The upper panel of Fig. 2 shows the result for a data set of size $n = 1020$ consisting of 1000 Laplace and 20 $N(2, 0.01^2)$ random variables. The selected value of k was 21. The density has 4 peaks but that caused by the 20 $N(2, 0.01^2)$ values is not visible. The centre panel shows a density with $k = 100$. In terms of the AIC criterion it is much worse than the density of the top panel. On the other hand there is a good reason for preferring it: it has the correct number of peaks, namely 2, and at the correct places. The centre panel is not a histogram as the heights of the bins do not correspond to the bin frequencies. The bottom panel shows a density which is also a histogram, but with unequal and yet automatically chosen bin lengths. This was obtained using the default procedure of the taut string method of Davies and Kovac (2004). This produces a sequence of models of increasing complexity but with complexity measured not by the dimension $k-1$ of the parameter mixing parameter $\boldsymbol{\pi}$ but by the number of local extreme values. The fidelity of a distribution function F_n to the data is measured by the closeness of the empirical distribution function transformed data $\tilde{\mathbf{x}}_n = F_n(\mathbf{x}_n)$ to a uniform distribution on $[0, 1]$. This in turn is measured by increments of Kuiper metrics of up to order 19 Davies and Kovac (2004) and ratios of order statistics $(\tilde{x}_{(j)} - \tilde{x}_{(i)})/(\tilde{x}_{(k)} - \tilde{x}_{(i)})$ (Dümbgen & Walther, in press). The fidelity is fixed and we take the simplest model which satisfies the conditions. In contrast to AIC there is no trade-off between complexity and fidelity. Furthermore the definition of fidelity is at the level of distribution functions (and not at the level of densities). It is perfectly possible

Table 1

The table gives the results for the taut string method which corresponds to those of Table 1 of Hjort and Claeskens (2004)

n	(a)	(b)	(c)
25	55.8	32.7	35.4
50	91.8	51.1	64.6
100	99.9	84.0	94.3

to ‘estimate’ densities in the weak d_{ko} -topology but only if these are subject to some form of shape or smoothness regularization. A statistician basing a decision on the top panel of Fig. 2 is indeed using a density-based goodness-of-fit. A decision based on the bottom histogram is not a density-based goodness-of-fit test, it is a combination of a Kuiper metric test and a multiresolution goodness-of-fit test, both in the weak d_{ko} -topology. In this particular case the method outperforms the AIC method in finding the peaks or modes mentioned by Hjort. It does this not only for this data set but consistently over a whole range of test-beds (see Davies, Gather, and Weinert (2008)). This is not surprising. The AIC measures fits of complicated models to complicated data sets by a single, one-dimensional number, the log-likelihood. The complexity is also a single, one-dimensional number and not subject to choice, namely the dimension. It is simply impossible to read off modality or smoothness from these two numbers. The whole concept is flawed. Finally we point out that all three models are infinitely distant from the truth in the Kullback–Leibler sense, but I do not suppose that anybody is worried by this. If however the distribution functions of the densities are far away from the empirical distribution function of the data, we should be worried.

Hjort argues, and gives examples to support his claim, that not all ‘diagnostics and model validation are F questions’, and I am willing to concede this as a description of actual practice. What I argue is that they should be, and that questions of f should be questions of F plus regularization. Hjort mentions the two-sample hypothesis $f = g$. Although I have worked on the corresponding problem in non-parametric regression I have no method to offer for the density problem. However the plan of attack is clear. The question would be reformulated as ‘Does there exist a density f_0 which adequately approximates both samples?’ where ‘adequately approximates’ is at the level of distribution functions and, at least as a first attempt, can be taken to be that of Davies and Kovac (2004).

Hjort also mentions the hypothesis $f = f_0$ with f_0 specified and starts with the simple testing problem $f = f_0$ against $f = f_1$. As he points out, the optimal test is density based. Optimality is one of the siren songs of likelihood, the other being the likelihood principle. Tukey’s comments on the milk bottle are relevant here. Optimality results can be interesting and useful but only for regularized problems with uninformative f_0 and f_1 . Otherwise the optimal procedure can be useless, as we have already seen, by taking f to be continuous under H_0 and any discrete distribution under H_1 . The more general problem of testing $f = f_0$ against $f \neq f_0$ can be tackled using the taut string method: one simply checks whether the associated distribution function F_0 approximates the data in the specified sense. The results corresponding to Table 1 of Hjort and Claeskens (2004) are given in Table 1. The results are not bad. Under (a) it is only slightly worse than the best two cos-methods. Here the polynomial methods dominate which is not surprising, given the generating mechanism. The worst performance is under (b) but here it beats the two cos-methods to which it lost out under (a). Under (c) it beats all the polynomial methods and again the two cos-methods to which it lost out under (a). But of course it can do much more than this, for example the bottom two histograms of Fig. 2.

Hjort’s interpretation of blandness is not what is meant. Blandness is a property of the distribution, irrespective of whether this is named by the distribution function or by the density. In the location model blandness is defined in terms of the Fisher information. For an arbitrary distribution function F we follow Huber (1981) and define it by

$$I(F) = \sup_{u \in \mathcal{C}_K^1} \frac{\left(\int u^{(1)}(x) dF(x) \right)^2}{\int u(x)^2 dF(x)}$$

where $\mathcal{C}_K^1$ denotes the set of all continuously differentiable functions u with compact support and $\int u^2 dF > 0$. It is finite if and only if F has an absolutely continuous density f with derivative $f^{(1)}$ and in this case

$$I(F) = \int \frac{f^{(1)}(x)^2}{f(x)} dx < \infty.$$

The $N(0, 1)$ distribution minimizes $I(F)$ under all F with variance 1. Optimality results are interesting if they hold for a worst-case distribution, such as the normal distribution, or, if the distribution has some ‘horn’ allowing an increase in precision, then this should have some exterior justification. As Tukey writes, we are ‘lucky’ that the Gaussian distribution is bland, although its very blandness is probably implied by the central limit theorem.

Fortunately I did not write that d_{ko} is the natural metric (although, in a very weak moment, I could well have done that), but that d_{ko} is the natural topology, something different. The Kuiper metrics of Davies and Kovac (2004) are much more efficacious for detecting peaks than is the Kolmogorov metric, yet they generate the same open sets. The topologies of $\int |F - G|$ and d_{ko} are not comparable. I like metrics which are well-defined and finite on the space of all probability measures: $\int |F - G|$ does not satisfy this. It is also not affine invariant which disqualifies it for some purposes. I can see no arguments at all in favour of the L_1 metric $\int |f - g|$. Its connection with total variation makes this clear

$$\int |f - g| = 2 \sup_{B \in \mathfrak{B}} |F(B) - G(B)|$$

with a slight abuse of notation. This implies that the distance between any discrete distribution and any continuous one is the maximum value of 2. This is the same problem we had with the Kullback–Leibler discrepancy. If the chemists start telling us the concentration of copper in the number of copper atom per 10^{23} molecules of water, which corresponds better to physical reality, then we have a ‘discrete truth’ and the L_1 distance to any continuous model is 2. I see nothing natural or authoritative in this, on the contrary. Suppose we replace the class of Borel sets $\mathfrak{B}$ by a smaller class $\mathcal{C}$ and define a metric by

$$d_{\mathcal{C}}(F, G) = \sup_{B \in \mathcal{C}} |F(B) - G(B)|.$$

Now if $\mathcal{C}$ is a Vapnik–Chervonenkis class then we get something sensible and we do not have to alter our concept of distance if the chemists suddenly start talking in terms of numbers of molecules. For work in higher dimensions on densities which are based on V–C-classes I refer to Polonik (1995).

The Kolmogorov metric depends on F which in turn depends on the density f , so, yes, the Kolmogorov distance depends on the densities. However, and this is the whole problem, the differential operator D connecting F and f via $D(F) = f$ is pathologically discontinuous. More concretely, for any given $\varepsilon > 0$ there are densities f and g with $\int |f - g| > 2 - \varepsilon$ and $d_{ko}(F, G) < \varepsilon$. In statistics the latter is the relevant quantity. Here is another way of looking at the problem. Given two densities f and g we have

$$\int |f - g| = \int_{f > g} f - g + \int_{g > f} g - f.$$

How complicated are the sets $f > g$ and $g > f$? In one dimension the only sort of sets we can deal with are intervals and unions of intervals. Suppose therefore that $f > g$ is a union of κ_1 disjoint intervals and $g > f$ is a union of κ_2 disjoint intervals. Then it is a simple exercise to show that $\int |f - g| = d_{ku}^{\kappa}(F, G)$ with $\kappa = \kappa_1 + \kappa_2$ where d_{ku}^{κ} denotes the Kuiper metric of order κ . Thus $\int |f - g|$ is only going to be of statistical relevance if it can be closely approximated by a Kuiper metric $d_{ku}^k(F, G)$ with k relatively small. The default value in Davies and Kovac (2004) for samples of size larger than 100 is $k = 19$. I suspect that something very similar happens in higher dimensions. If $\int |f - g|$ cannot be closely approximated by $d_{\mathcal{C}}(F, G)$ where the V–C-dimension of $\mathcal{C}$ is small, then it is not of statistical relevance. In other words, Kuiper and associated metrics have authority, and in cases of disagreement with the L_1 metric, they are to be preferred.

Hjort refers to the copper data and noticed two errors for which I apologize and for which I thank him. Firstly, he is correct in that the variance of the mixture model is not 1 as claimed. A square-root is missing and (8) should read

$$x_i = \frac{x'_i}{\sqrt{\sigma^2 + \frac{1}{k} \sum_{j=1}^k x_j'^2}} \quad \text{where } x'_i = \Phi^{-1}(i/(k+1)), i = 1, \dots, k. \quad (3)$$

This also makes the model closer to the standard normal model with a Kolmogorov distance of 0.0122. I am rather embarrassed by this, I should have seen immediately that the distance 0.029 given in the paper was too large. Secondly

the value of the log-likelihood for the normal model is not 19.805 as claimed but 20.315 as given by Hjort. I often forget that R uses $n - 1$ in the definition of the variance and I failed to take this into account. Another problem is that I failed to make the idea of approximation of a model to data sufficiently precise. It is implicit in the reference to ‘finite precision’ and Tukey’s remark about the irrationality of Gaussian random variables but it was not explicit. When comparing real data with simulated samples, the latter should be given to the same precision as the former. Otherwise the task is trivial. Without this it may indeed be possible to reject the mixture model but if the truncation is performed, I doubt whether this can be done. Hjort states that the mixture model is a ‘highly wriggly affair’ with 50 sharp peaks and would therefore be ruled out on ‘a priori’ grounds. Actually the proposal of 50 peaks is extremely modest compared with the truth of 10^{23} peaks. The mixture model was nevertheless tongue-in-cheek: what was intended is the fact that there are many models consistent with the data. Besides the mixture model we could take the Laplace distribution and operate optimally for this distribution. The maximum likelihood estimates are the median and the mean absolute deviation from the median. The resulting 95% confidence interval is [1.996, 2.064] which is shorter than that based on the Gaussian distribution. The value of the log-likelihood 20.089 is somewhat lower but not drastically so. We could also use various t -distributions or Huber distributions. Many of these would pass even Hjort’s tests and the question is, which one do we use and why? One answer is to take the most uninformative one, that is bland or hornless, and to base the analysis on this. In this spirit one could minimize the Fisher information within a prescribed Kuiper ball of the empirical distribution and use this as the model. This leads to a nonlinear optimization problem but could possibly give good results for large sample sizes. Hampel proposes using simple models, where ‘simple’ can also be interpreted as blandness, and estimating the parameters robustly based on this model. He even proposes this when the model is clearly wrong, and there is much to be said in his favour. Nevertheless, even in this case, the model cannot be too wrong. What I propose in this situation is to change tack completely and think in terms, not of models, but of procedures, as already indicated in Section 5.2. What do we want our procedure to accomplish? It should provide a point value for location, a range of reasonable values, and a reasonable value for the scale. It should be affinely equivariant and allow a stable analysis in that small alterations in the data should cause only small alterations in the results. Here small changes in the data are measured in the Kolmogorov metric and allow for large alterations in a few data points. Moreover we wish the procedure to behave stably on different test-beds. On symmetric test-beds the point value should be close to the point of symmetry and the range of reasonable values means that it should include the point of symmetry with a specified probability, say 0.95. Finally we wish its performance in a chosen worst-case scenario, say that of a Gaussian test-bed, to be reasonable when compared with the optimal strategy. It is possible to include other properties such as bias behaviour when outliers are one-sided but this is sufficient to explain the approach. The problem is now to construct such a procedure. The demand that both the coverage probabilities and the lengths of the intervals should be stable over different test-beds leads to the requirement of Fréchet differentiability with respect to the weak d_{ko} -topology. Without this we could not replace the P' in (34) by the empirical distribution $\mathbb{P}_n$. Fréchet differentiability with respect to the L_1 metric is not sufficient as $\int |p_n - f| = 2$ whereas $d_{ko}(P_n, P) = O(1/\sqrt{n})$. This is yet another version of the same argument that the L_1 metric is irrelevant for statistics. Constructing a functional which is Fréchet differentiability with respect to the weak d_{ko} -topology is more difficult than the construction one which is Fréchet differentiability with respect to the strong L_1 -topology; there are fewer open sets. As pointed out by Hampel in his contribution, some statisticians thought it was not possible until Clarke (1983). The result is the procedure of Section 11. When applied to the copper data it gives a point value of 2.021 and an interval [1.974, 2.068]. If we replace the observations 1.7 and 1.86 by 0.17 and 0.186 respectively the point value is 2.019 with interval [1.970, 2.069], which demonstrates the stability of the procedure. Finally I point out that the procedure cannot be derived by a likelihood approach based on some location/scale model.

Hjort is correct in that the approximation regions can be seen as inverted tests. This was also noted by Pötscher who pointed out that simply to treat the procedure as one of inverting tests would be missing the point. Expressed as a test I would be interested in testing some null hypothesis about the mode of generation of the data, but I see no point in testing a hypothesis which I hold, with good reason, to be false a priori. By regarding such tests as expressing some form of typical (controlled by the size of the test) behaviour under the model, and then asking whether the data also exhibit this typical behaviour, the emphasis is moved from some hidden ‘truth’ to the data. The choice of ‘typical’ cannot be prescribed by an appeal to some form of optimality, even if optimal tests under *regularized, non-informative, bland* models may give important indications of what is attainable and what is not attainable. The inverted test leading to the approximation region (18) are those of testing that the mean is zero on the intervals in question. The model is the worst-case uninformative Gaussian distribution. Dümbgen and Spokoiny (2001) also invert optimal tests but for

testing for monotonicity on the intervals, and lead to weighted sums of residuals. This is (j) of Section 2. Inverting tests has one advantage over the usual confidence intervals in that the answer can be ‘no’. Nevertheless, in spite of a formal correspondence in this aspect I find it much more satisfying to ask the question ‘Do the data look like P -random variables?’ than to ask the test question ‘Are they P -random variables?’. The former is a question of approximation, the latter a question of truth.

In conclusion I claim that Hjort has failed to make his case. Likelihood remains blind and leaves us with no idea whether we are close to the truth in terms of Kullback–Leibler or infinitely far away. I suspect always the latter.

References

- Andrews, D. F., & Herzberg, A. M. (1985). *A collection of problems from many fields for the student and research worker*. New York: Springer.
- Bell, J. S. (1987). *Speakable and unspeakable in quantum mechanics*. Cambridge: Cambridge University press.
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference* (2nd ed.). New York: Springer.
- Clarke, B. R. (1983). Uniqueness and Fréchet differentiability of functional solutions to maximum likelihood type equations. *Annals of Statistics*, 11, 1196–1205.
- Davies, P. L. (1987). Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *Annals of Statistics*, 15, 1269–1292.
- Davies, P. L. (1990). The asymptotics of S-estimators in the linear regression model. *Annals of Statistics*, 18, 1651–1675.
- Davies, P. L. (1992). The asymptotics of Rousseeuw’s minimum volume ellipsoid. *Annals of Statistics*, 20, 1828–1843.
- Davies, P. L. (1993). Aspects of robust linear regression. *Annals of Statistics*, 21, 1843–1899.
- Davies, P. L. (1995). Data features. *Statistica Neerlandica*, 49, 185–245.
- Davies, P. L., Gather, U., & Weinert, H. (2008). Nonparametric regression as an example of model choice. *Communications in Statistics - Simulation and Computation*, 37, 274–289.
- Davies, P. L., & Kovac, A. (2004). Densities, spectral densities and modality. *Annals of Statistics*, 32(3), 1093–1136.
- Dümbgen, L., & Spokoiny, V. G. (2001). Multiscale testing of qualitative hypotheses. *Annals of Statistics*, 29(1), 124–152.
- Dümbgen, L., & Walther, G. (2008). Multiscale inference about a density. *Annals of Statistics* (in press).
- Dürr, D. (2001). *Bohmsche mechanik als grundlage der quantenmechanik*. Berlin, Heidelberg: Springer.
- Field, C., & Hampel, F. (1982). Small-sample asymptotic distributions of M-estimators of location. *Biometrika*, 69, 29–46.
- Field, C., & Ronchetti, E. (1990). *Monograph series, Small sample asymptotics*. Hayward, CA: Institute of Mathematical Statistics.
- Ghirardi, G. C., Rimini, A., & Weber, T. (1986). Unified dynamics for microscopic and macroscopic systems. *Physical Review D*, 34(2), 470–491.
- Goldstein, S. (1998a). Quantum theory without observers- part one. *Physics Today*, 42–46.
- Goldstein, S. (1998b). Quantum theory without observers- part two. *Physics Today*, 38–42.
- Hampel, F. (1973). Some small sample asymptotics. In *Proceedings of the prague symposium on asymptotic statistics*, (pp. 109–126).
- Hampel, F. R. (1985). The breakdown points of the mean combined with some rejection rules. *Technometrics*, 27, 95–107.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. New York: Wiley.
- Held, L. (2008). *Methoden der statistischen Inferenz: Likelihood und Bayes*. Heidelberg: Spektrum Akademischer Verlag.
- Hjort, N. L., & Claeskens, G. (2004). Goodness of fit via nonparametric likelihood ratios. *Scandinavian Journal of Statistics*, 31, 487–513.
- Hooft, G. t. (1997). *In search of the ultimate building blocks*. Cambridge, UK: Cambridge University Press.
- Huber, P. J. (1981). *Robust statistics*. New York: Wiley.
- Kac, M. (1959). *The Carus Mathematical Monographs: Vol. 12. Statistical independence in probability, analysis and number theory*. Wiley.
- Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Wiley series in probability and statistics. Robust statistics: Theory and methods*. Wiley.
- Pawitan, Y. (2001). *In all likelihood: Statistical modelling and inference using likelihood*. Oxford university Press.
- Polonik, W. (1995). Density estimation under qualitative assumptions in higher dimensions. *Journal of Multivariate Analysis*, 55, 61–81.
- Rissanen, J. (1998). *Series in computer science: Vol. 15. Stochastic complexity in statistical inquiry* (2nd ed.). Singapore: World Scientific.
- Tennenbaum, G. (1995). *Cambridge Studies in Advanced Mathematics: Vol. 46. Introduction to analytic and probabilistic number theory*. Cambridge University Press.
- Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley.
- Williams, B. (2002). *Truth and truthfulness*. Princeton University Press.