The Annals of Statistics
2002, Vol. 30, No. 5, 1225–1310

WHAT IS A STATISTICAL MODEL?[1]

BY PETER MCCULLAGH

University of Chicago

This paper addresses two closely related questions, "What is a statistical model?" and "What is a parameter?" The notions that a model must "make sense," and that a parameter must "have a well-defined meaning" are deeply ingrained in applied statistical work, reasonably well understood at an instinctive level, but absent from most formal theories of modelling and inference. In this paper, these concepts are defined in algebraic terms, using morphisms, functors and natural transformations. It is argued that inference on the basis of a model is not possible unless the model admits a natural extension that includes the domain for which inference is required. For example, prediction requires that the domain include all future units, subjects or time points. Although it is usually not made explicit, every sensible statistical model admits such an extension. Examples are given to show why such an extension is necessary and why a formal theory is required. In the definition of a subparameter, it is shown that certain parameter functions are natural and others are not. Inference is meaningful only for natural parameters. This distinction has important consequences for the construction of prior distributions and also helps to resolve a controversy concerning the Box–Cox model.

1. Introduction. According to currently accepted theories [Cox and Hinkley (1974), Chapter 1; Lehmann (1983), Chapter 1; Barndorff-Nielsen and Cox (1994), Section 1.1; Bernardo and Smith (1994), Chapter 4] a statistical model is a set of probability distributions on the sample space $\mathcal{S}$. A parameterized statistical model is a parameter set Θ together with a function $P : \Theta \to \mathcal{P}(\mathcal{S})$, which assigns to each parameter point $\theta \in \Theta$ a probability distribution P_θ on $\mathcal{S}$. Here $\mathcal{P}(\mathcal{S})$ is the set of *all* probability distributions on $\mathcal{S}$. In much of the following, it is important to distinguish between the model as a function $P : \Theta \to \mathcal{P}(\mathcal{S})$, and the associated set of distributions $P\Theta \subset \mathcal{P}(\mathcal{S})$.

In the literature on applied statistics [McCullagh and Nelder (1989); Gelman, Carlin, Stern and Rubin (1995); Cox and Wermuth (1996)], sound practical advice is understandably considered to be more important than precise mathematical

Received January 2000; revised July 2001.

[1]Supported in part by NSF Grants DMS-97-05347 and DMS-00-71726.

AMS 2000 subject classifications. Primary 62A05; secondary 62F99.

Key words and phrases. Aggregation, agricultural field experiment, Bayes inference, Box–Cox model, category, causal inference, commutative diagram, conformal model, contingency table, embedding, exchangeability, extendability, extensive variable, fertility effect, functor, Gibbs model, harmonic model, intensive variable, interference, Kolmogorov consistency, lattice process, measure process, morphism, natural parameterization, natural subparameter, opposite category, quadratic exponential model, representation, spatial process, spline model, type III model.

definitions. Thus, most authors do not offer a precise mathematical definition of a statistical model. Typically, the preceding definition is taken for granted, and applied to a range of sensibly constructed models.

At a minimum, a Bayesian model requires an additional component in the form of a prior distribution on Θ. A Bayesian model in the sense of Berger (1985), Smith (1984) or Bernardo and Smith (1994) requires an extra component in the form of a judgment of infinite exchangeability or partial exchangeability in which parameters are *defined* by limits of certain statistics. Although Bayesian formulations are not the primary focus of this paper, the notion that the model is extendable to a sequence, usually infinite, is a key concept.

The parameterization is said to be identifiable if distinct parameter values give rise to distinct distributions; that is, $P_\theta = P_{\theta'}$ implies $\theta = \theta'$. Thus, the parameter is identifiable if and only if $P : \Theta \to \mathcal{P}(\mathcal{Y})$ is injective. Apart from these conditions, the standard definition permits arbitrary families of distributions to serve as statistical models, and arbitrary sets Θ to serve as parameter spaces.

For applied work, the inadequacy of the standard definition is matched only by the eccentricity of the formulations that are permitted. These examples make it abundantly clear that, unless the model is embedded in a suitable structure that permits extrapolation, no useful inference is possible, either Bayesian or non-Bayesian.

To be fair, most authors sound a note of warning in their discussion of statistical models. Thus, for example, Cox and Hinkley [(1974), Chapter 1], while admitting that "it is hard to lay down precise rules for the choice of the family of models," go on to offer a range of recommendations concerning the model and the parameterization. In particular, "the model should be consistent with known limiting behavior" and the parameterization should be such that "different parameter [components] have individually clear-cut interpretations." The intention of this article is to define some of these concepts in purely algebraic terms.

2. Examples.

2.1. *Twelve statistical exercises.* The following list begins with four exercises in which the models are plainly absurd. The point of the exercises, however, is not so much the detection of absurd models as understanding the sources of absurdity. From a more practical viewpoint, the more interesting exercises are those in which the absurdity is not obvious at first sight.

EXERCISE 1 (A binary regression model). Consider a model for independent binary responses in which certain covariates are prespecified. One of these covariates is designated the treatment indicator. The model specifies a logit link if the number of subjects is even, and a probit link otherwise. Inference is required for the treatment effect. Predictions are required for the response of a future subject with specified covariate values.

EXERCISE 2 (A randomized blocks model). Consider a randomized blocks design with b blocks and k varieties. In the statistical model, all observations are independent and normally distributed with unit variance. The expected yield μ_{ij} of variety i in block j is assumed to be expressible in the following manner for some real-valued function α on varieties and β on blocks:

$$\mu_{ij} = \begin{cases} \alpha_i + \beta_j, & \text{if } k+b \text{ is even,} \\ \exp(\alpha_i + \beta_j), & \text{otherwise.} \end{cases}$$

On the basis of this model, inference is required in the form of a confidence interval or posterior distribution for a particular variety contrast $\alpha_1 - \alpha_2$.

EXERCISE 3 (A linear regression model). In the standard linear regression model $Y \sim N(X\beta, \sigma^2 I_n)$ on $\mathcal{R}^n$, the parameter (β, σ^2), is a point in $\mathcal{R}^p \times [0, \infty)$. In our modified eccentric version, the parameter space is $\Theta = \mathcal{R}^p \times [n, \infty)$, so that $\sigma^2 \geq n$. A prediction interval is required for the value of the response on a new subject whose covariate value is $x \in \mathcal{R}^p$.

EXERCISE 4 [An i.i.d. normal model (Section 6.6)]. In this model, the observations are independent, identically distributed and normal. The parameter space is $\Theta = \mathcal{R}^2$. If n is even, the mean is θ_1 and the variance is θ_2^2: otherwise, the mean is θ_2 and the variance is θ_1^2. On the basis of observed values $(y_1, \ldots, y_n)$, inference is required in the form of confidence limits or a posterior distribution on Θ.

EXERCISE 5 [The type III model (Section 6.6)]. Consider a randomized blocks design as in Exercise 2 above, all observations being independent, normal with unit variance. According to the type III model as described in the SAS manual [Littell, Freund and Spector (1991), pages 156–160], the vector μ lies in the linear subspace III_{kb} of dimension $bk - k + 1$ such that the k variety means are equal,

$$III_{kb} = \{\mu \in \mathcal{R}^{kb} : \bar{\mu}_{1\cdot} = \cdots = \bar{\mu}_{k\cdot}\}.$$

On the assumption that the fit is adequate, what conclusions can be drawn about variety differences on either a subset of the blocks or on other blocks similar to some of those used in the experiment?

EXERCISE 6 [The Box–Cox model (Section 7)]. In the Box–Cox model [Box and Cox (1964)], it is assumed that after some componentwise power transformation $Y_i \mapsto Y_i^\lambda$, the transformed response variable satisfies the standard normal-theory linear model with mean $E(Y^\lambda) = X\beta$ and constant variance σ^2. In the problem posed by Bickel and Doksum (1981), inference is required in the form of confidence intervals or a posterior distribution for the parameter β or a component thereof.

EXERCISE 7 [A model for clustered data (Section 6.6)]. In a cluster of size k, the response Y has joint density with respect to Lebesgue measure on $\mathcal{R}^k$ proportional to

$$\exp\left(-\frac{1}{2}\theta_1 \sum y_i^2 + \frac{1}{2}\theta_2 \sum_{i\neq j} \frac{y_i y_j}{k-1}\right)$$

for some $\theta_1 > 0$ and $0 \le \theta_2 < \theta_1$. Thus, the vector Y is normally distributed with zero mean and exchangeable components. Observations on distinct clusters are independent. On the basis of the observed data $(k_1, Y_1), \ldots, (k_n, Y_n)$, in which k_r is the size of cluster r and Y_r normally distributed on $\mathcal{R}^{k_r}$, a confidence set is required for θ. In particular, if the observed clusters are all of size 2, inference is required for the distribution of the cluster mean in a cluster of size 8.

EXERCISE 8 [An i.i.d. Cauchy model (Section 6.6)]. By the two-parameter Cauchy family is meant the set of distributions on $\mathcal{R}$ with densities

$$\left\{\frac{|\theta_2|\,dy}{\pi(\theta_2^2 + (y-\theta_1)^2)} : \theta_2 \neq 0,\ \theta_1 \in \mathcal{R}\right\}.$$

Let $Y_1, \ldots, Y_n$ be n independent and identically distributed random variables with distribution in the two-parameter Cauchy family. A confidence interval or posterior distribution is required for the parameter θ_1. The catch here is that the Cauchy family is closed under the real fractional linear group, and the confidence interval is required to have the corresponding property. In other words, for any real numbers a, b, c, d such that $ad - bc \neq 0$, the random variables $Z_i = (aY_i + b)/(cY_i + d)$ are i.i.d. Cauchy. If we write $\theta = \theta_1 \pm i\theta_2$ as a conjugate pair of complex numbers, the transformed parameter is $\psi = (a\theta + b)/(c\theta + d)$ [McCullagh (1992, 1996)]. The procedure used must be such that if the values $Z_1, \ldots, Z_n$ are reported as i.i.d. Cauchy(ψ), and a confidence interval is requested for $\theta_1 = \Re((d\psi - b)/(a - c\psi))$, the same answer must be obtained regardless of the values a, b, c, d.

EXERCISE 9 (A model for a spatial process). The temperature in a room is modelled as a stationary isotropic Gaussian process in which the mean temperature is constant $E(Y_x) = \mu$, and the covariance function is $\operatorname{cov}(Y_x, Y_{x'}) = \sigma^2 \exp(-\lambda|x - x'|)$. The parameter space is

$$(\mu, \sigma^2, \lambda) \in \mathcal{R} \times (0, \infty)^2.$$

A confidence interval is required for $\theta = \mu/\sigma$.

EXERCISE 10 [Regression and correlation (Section 6.2)]. Consider the standard normal-theory linear regression model with one covariate in which

the observations are independent, normally distributed with conditional mean $E(Y|x) = \alpha + \beta x$ and constant variance σ^2. The parameter space is

$$(\alpha, \beta, \sigma^2) \in \mathcal{R}^2 \times [0, \infty).$$

A confidence interval or posterior distribution is required for the correlation coefficient.

EXERCISE 11 [Spatial simultaneous equation model (Section 6.5)]. Let Y be a spatial process observed on a rectangular lattice of sites. The joint distribution is defined by a set of simultaneous equations

$$Y_i = \sum_{j \sim i} \beta Y_j + \varepsilon_i,$$

in which $j \sim i$ means that site $j \neq i$ is a neighbour of site i. This expression is interpreted to mean that $(I - B)Y = \varepsilon$ is standard normal. The components of B are zero except for neighboring sites for which $b_{ij} = \beta$. The system is such that $I - B$ is invertible, so that Y is normal with zero mean and inverse covariance matrix $(I - B)^T (I - B)$. A confidence interval is required for β.

EXERCISE 12 [Contingency table model (Section 6.4)]. All three factors in a contingency table are responses. In principle, each factor has an unbounded number of levels, but some aggregation has occurred, and factor B is in fact recorded in binary form. The log-linear model $AB + BC$ is found to fit well, but no log-linear submodel fits. What conclusions can be drawn?

2.2. *Remarks.* These exercises are not intended to be comparable in their degree of absurdity. They are intended to illustrate a range of model formulations and inferential questions, some clearly artificial, some a little fishy, and others bordering on acceptability. In the artificial class, I include Exercises 1, 2, 4, and possibly 3. Exercise 9 ought also to be regarded as artificial or meaningless on the grounds of common sense. The type III model has been criticized from a scientific angle by Nelder (1977) in that it represents a hypothesis of no scientific interest. Despite this, the type III model continues to be promoted in text books [Yandell (1997), page 172] and similar models obeying the so-called weak heredity principle [Hamada and Wu (1992)] are used in industrial experiments. The absurdity in the other exercises is perhaps less obvious. Although they appear radically different, from an algebraic perspective Exercises 2 and 5 are absurd in rather similar ways.

Each of the formulations satisfies the standard definition of a statistical model. With respect to a narrowly defined inferential universe, each formulation is also a statistical model in the sense of the definition in Sections 1 and 4. Inference, however, is concerned with natural extension, and the absurdity of

each formulation lies in the extent of the inferential universe or the scope of the model. Although a likelihood function is available in each case, and a posterior distribution can be computed, no inferential statement can breach the bounds of the inferential universe. To the extent that the scope of the model is excessively narrow, none of the formulations permits inference in the sense that one might reasonably expect. We conclude that, in the absence of a suitable extension, a likelihood and a prior are not sufficient to permit inference in any commonly understood sense.

3. Statistical models.

3.1. *Experiments.* Each statistical experiment or observational study is built from the following objects:

1. A set $\mathcal{U}$ of statistical units, also called experimental units, plots, or subjects;
2. A covariate space Ω;
3. A response scale $\mathcal{V}$.

The design is a function $x : \mathcal{U} \to \Omega$ that associates with each statistical unit $i \in \mathcal{U}$ a point x_i in the covariate space Ω. The set $\mathcal{D} = \Omega^{\mathcal{U}}$ of all such functions from the units into the covariate space is called the design space.

The response is a function $y : \mathcal{U} \to \mathcal{V}$ that associates with each statistical unit i, a response value y_i in $\mathcal{V}$. The set $\mathcal{S} = \mathcal{V}^{\mathcal{U}}$ of all such functions is called the sample space for the experiment. In the definition given in Section 1, a statistical model consists of a design $x : \mathcal{U} \to \Omega$, a sample space $\mathcal{S} = \mathcal{V}^{\mathcal{U}}$ and a family of distributions on $\mathcal{S}$.

A statistical model $P : \Theta \to \mathcal{P}(\mathcal{S})$ associates with each parameter value θ a distribution $P\theta$ on $\mathcal{S}$. Of necessity, this map depends on the design, for example, on the association of units with treatments, and the number of treatment levels that occur. Thus, to each design $x : \mathcal{U} \to \Omega$, there corresponds a map $P_x : \Theta \to \mathcal{P}(\mathcal{S})$ such that $P_x\theta$ is a probability distribution on $\mathcal{S}$. Exercise 2 suggests strongly that the dependence of P_x on $x \in \mathcal{D}$ cannot be arbitrary or capricious.

3.2. *The inferential universe.* Consider an agricultural variety trial in which the experimental region is a regular 6×10 grid of 60 rectangular plots, each seven meters by five meters, in which the long side has an east–west orientation. It is invariably understood, though seldom stated explicitly, that the purpose of such a trial is to draw conclusions concerning variety differences, not just for plots of this particular shape, size and orientation, but for comparable plots of various shapes, sizes and orientations. Likewise, if the trial includes seven potato varieties, it should ordinarily be possible to draw conclusions about a subset of three varieties from the subexperiment in which the remaining four varieties are ignored. These introductory remarks may seem obvious and unnecessary, but they have far-reaching implications for the construction of statistical models.

The first point is that before any model can be discussed it is necessary to establish an inferential universe. The mathematical universe of experimental units might be defined as a regular 6×10 grid of 7×5 plots with an east–west orientation. Alternatively, it could be defined as the set of all regular grids of rectangular plots, with an east–west orientation. Finally, and more usefully, the universe could be defined as a suitably large class of subsets of the plane, regardless of size, shape and orientation. From a purely mathematical perspective, each of these choices is internally perfectly consistent. From the viewpoint of inference, statistical or otherwise, the first and second choices determine a limited universe in which all inferential statements concerning the likely yields on odd-shaped plots of arbitrary size are out of reach.

In addition to the inferential universe of statistical units, there is a universe of response scales. In an agricultural variety trial, the available interchangeable response scales might be bushels per acre, tones per hectare, kg/m^2, and so on. In a physics or chemistry experiment, the response scales might be °C, °K, °F, or other suitable temperature scale. In a food-tasting experiment, a seven-point ordered scale might be used in which the levels are labelled as

$$\mathcal{V} = \{\text{unacceptable, mediocre, } \ldots, \text{ excellent}\},$$

or a three-point scale with levels

$$\mathcal{V}' = \{\text{unacceptable, satisfactory, very good}\}.$$

It is then necessary to consider transformations $\mathcal{V} \to \mathcal{V}'$ in which certain levels of $\mathcal{V}$ are, in effect, aggregated to form a three-level scale. Finally, if the response scale is bivariate, including both yield and quality, this should not usually preclude inferential statements concerning quality or quantity in isolation.

Similar comments are in order regarding the covariate space. If, in the experiment actually conducted, fertilizer was applied at the rates, 0, 100, 200 and 300 kg/ha, the inferential universe should usually include all nonnegative doses. Likewise, if seven varieties of potato were tested, the inferential universe should include all sets of varieties because these particular varieties are a subset of other sets of varieties. This does not mean that informative statements can be made about the likely yield for an unobserved variety, but it does mean that the experiment performed may be regarded as a subset of a larger notional experiment in which further varieties were tested but not reported.

3.3. *Categories.* Every logically defensible statistical model in the classical sense has a natural extension from the set of observed units, or observed varieties, or observed dose levels, to other unobserved units, unobserved blocks, unobserved treatments, unobserved covariate values and so on. Thus, in constructing a statistical model, it is essential to consider not only the observed units, the observed blocks, the observed treatments and so on, but the inferential universe of all

relevant sample spaces, all relevant sets of blocks, all relevant covariate values and so on. Every sensible statistical model does so implicitly. The thesis of this paper is that the logic of every statistical model is founded, implicitly or explicitly, on categories of morphisms of the relevant spaces. The purpose of a category is to ensure that the families of distributions on different sample spaces are logically related to one another and to ensure that the meaning of a parameter is retained from one family to another.

A category $\mathcal{C}$ is a set of objects, together with a collection of arrows representing maps or morphisms between pairs of objects. In the simplest categories, each object is a set, and each morphism is a map. However, in the category of statistical designs, each design is a map and each morphism is a pair of maps, so a morphism need not be a map between sets. To each ordered pair (S, S') of objects in $\mathcal{C}$, there corresponds a set $\hom_{\mathcal{C}}(S, S')$, also denoted by $\mathcal{C}(S, S')$, of morphisms $\varphi : S \to S'$, with domain S and codomain S'. For certain ordered pairs, this set may be empty. Two conditions are required in order that such a collection of objects and arrows should constitute a category. First, for each object S in $\mathcal{C}$, the identity morphism $1 : S \to S$ is included in $\hom(S, S)$. Second, for each pair of morphisms $\varphi : S \to S'$ and $\psi : S' \to S''$, such that $\operatorname{dom}\psi = \operatorname{cod}\varphi$, the composition $\psi\varphi : S \to S''$ is a morphism in $\hom(S, S'')$ in $\mathcal{C}$. In particular, the set $\hom(S, S)$ of morphisms of a given object is a monoid, a semigroup containing the identity and closed under composition. A glossary of certain category terminology is provided in the Appendix. For all further details concerning categories, see Mac Lane (1998).

In the discussion that follows, the symbol $\mathrm{cat}_{\mathcal{U}}$ represents the category of morphisms of units. The objects in this category are all possible sets $\mathcal{U}, \mathcal{U}', \ldots$ of statistical units. These sets may have temporal, spatial or other structure. The morphisms $\varphi : \mathcal{U} \to \mathcal{U}'$ in $\mathrm{cat}_{\mathcal{U}}$ are maps, certainly including all insertion maps $\varphi : \mathcal{U} \to \mathcal{U}'$ (such that $\varphi u = u$) whenever $\mathcal{U} \subset \mathcal{U}'$. In general, $\mathrm{cat}_{\mathcal{U}}$ is the category of morphisms that preserves the structure of the units, such as equivalence relationships in a block design [McCullagh (2000), Section 9.5] or temporal structure in time series. In typical regression problems therefore, $\mathrm{cat}_{\mathcal{U}}$ is identified with the the generic category $\mathit{1}$ of injective, or 1–1, maps on finite sets. These are the maps that preserve distinctness of units: $u \neq u'$ in $\mathcal{U}$ implies $\varphi(u) \neq \varphi(u')$ in $\mathcal{U}'$.

A response is a value taken from a certain set, or response scale. If the response is a temperature, each object in $\mathrm{cat}_{\mathcal{V}}$ is a temperature scale, including one object for each of the conventional scales °C, °F and °K. To each ordered pair of temperature scales $(\mathcal{V}, \mathcal{V}')$ there corresponds a single invertible map, which is an affine transformation $\mathcal{V} \to \mathcal{V}'$. Likewise, yield in a variety trial may be recorded on a number of scales such as bushels/acre, tones/ha or $\mathrm{kg/m^2}$. To each response scale there corresponds an object $\mathcal{V}$ in $\mathrm{cat}_{\mathcal{V}}$, and to each pair of response scales there corresponds a single invertible map, which is a positive scalar multiple. For a typical qualitative response factor with ordered levels, the morphisms are order-preserving surjections, which need not be invertible. Special applications call for

other sorts of morphisms, such as censoring in survival data. The essential point here is that each response scale determines a distinct object in $\mathrm{cat}_{\mathcal{V}}$, and the maps $\mathcal{V} \to \mathcal{V}'$ are surjective (onto).

Since statistical inferences are invariably specific to the scale of measurement, it is absolutely essential that the various quantitative response scales not be fused into one anonymous scale labelled $\mathcal{R}$ or $\mathcal{R}^+$. An affine transformation of temperature scales is a transformation $\varphi : \mathcal{V} \to \mathcal{V}'$ from one scale into another. Unless $\mathcal{V} = \mathcal{V}'$, such a transformation is not composable with itself, a critical distinction that is lost if all scales are fused into $\mathcal{R}$. Thus, although $\mathrm{cat}_{\mathcal{V}}$ may be a category of invertible maps, it is ordinarily not a group.

Likewise, the symbol cat_{Ω} represents the category whose objects are the covariate spaces, and whose morphisms are maps, ordinarily injective, between covariate spaces. Because of the great variety of covariate spaces, these maps are more difficult to describe in general. Nonetheless, some typical examples can be described.

A quantitative covariate such as weight is recorded on a definite scale, such as pounds, stones or kg. To each scale Ω there corresponds a set of real numbers, and to each pair (Ω, Ω') of quantitative scales there corresponds a 1–1 map $\Omega \to \Omega'$, usually linear or affine. The objects in cat_{Ω} are some or all subsets of each measurement scale. For some purposes, it is sufficient to consider only bounded intervals of each scale: for other purposes, all finite subsets may be sufficient. Generally speaking, unless there is good reason to restrict the class of sets, the objects in cat_{Ω} are *all* subsets of *all* measurement scales. The morphisms are those generated by composition of subset insertion maps and measurement-scale transformation. Thus, cat_{Ω} may be such that there is no map from the set {ten stones, eleven stones} into {140 lbs, 150 lbs, 160 lbs} but there is one map into $\{x \text{ lbs} : x \geq 0\}$ whose image is the subset {140 lbs, 154 lbs}.

For a typical qualitative covariate such as variety with nominal unordered levels, the objects are all finite sets, and cat_{Ω} may be identified with the generic category $\mathcal{I}$, of injective maps on finite sets. If the levels are ordinal, cat_{Ω} may be identified with the subcategory of order-preserving injective maps. Factorial designs have several factors, in which case cat_{Ω} is a category in which each object is a product set $\Omega = \Omega_1 \times \cdots \times \Omega_k$. The relevant category of morphisms is usually the product category, one component category for each factor. For details, see McCullagh (2000).

4. Functors and statistical models.

4.1. *Definitions.* It is assumed in this section that the response is an *intensive* variable, a $\mathcal{V}$-valued function on the units. Extensive response variables are discussed in Section 8. In mathematical terms, an intensive variable is a function on the units: an extensive variable such as yield is an additive set function. The importance of this distinction in applied work is emphasized by Cox and Snell

[(1981), Section 2.1]. The implications for the theory of random processes are discussed briefly by Kingman [(1984), page 235].

In terms of its logical structure, each statistical model is constructed from the following three components:

1. A category $\mathrm{cat}_{\mathcal{U}}$ in which each object $\mathcal{U}$ is a set of statistical units. The morphisms $\mathcal{U} \to \mathcal{U}'$ in $\mathrm{cat}_{\mathcal{U}}$ are all injective maps preserving the structure of the units. In typical regression problems, $\mathrm{cat}_{\mathcal{U}}$ may be identified with the category $\mathcal{I}$ of all injective maps on finite sets.
2. A category cat_{Ω} in which each object Ω is a covariate space. The morphisms $\Omega \to \Omega'$ are all injective maps preserving the structure of the covariate spaces.
3. A category $\mathrm{cat}_{\mathcal{V}}$ in which each object $\mathcal{V}$ is a response scale. The morphisms $\mathcal{V} \to \mathcal{V}'$ are all maps preserving the structure of the response scale. These are typically surjective.

These three categories are the building blocks from which the design category, the sample space category and all statistical models are constructed.

Given a set of units $\mathcal{U}$ and a covariate space Ω, the design is a map $x : \mathcal{U} \to \Omega$ associating with each unit $u \in \mathcal{U}$ a point x_u in the covariate space Ω. In practice, this information is usually coded numerically for the observed units in the form of a matrix X whose uth row is the coded version of x_u. The set of all such designs is a category $\mathrm{cat}_{\mathcal{D}}$ in which each object x is a pair $(\mathcal{U}, \Omega)$ together with a map $x : \mathcal{U} \to \Omega$. Since the domain and codomain are understood to be part of the definition of x, the set of designs is the set of all such maps with $\mathcal{U}$ in $\mathrm{cat}_{\mathcal{U}}$ and Ω in cat_{Ω}, in effect, the set of model matrices X with labelled rows and columns. Each morphism $\varphi : x \to x'$ in which $x' : \mathcal{U}' \to \Omega'$, may be associated with a pair of injective maps $\varphi_d : \mathcal{U} \to \mathcal{U}'$ in $\mathrm{cat}_{\mathcal{U}}$, and $\varphi_c : \Omega \to \Omega'$ in cat_{Ω} such that the diagram

$$\begin{array}{ccc} \mathcal{U} & \xrightarrow{x} & \Omega \\ {\scriptstyle\varphi_d}\downarrow & & \downarrow{\scriptstyle\varphi_c} \\ \mathcal{U}' & \xrightarrow{x'} & \Omega' \end{array}$$

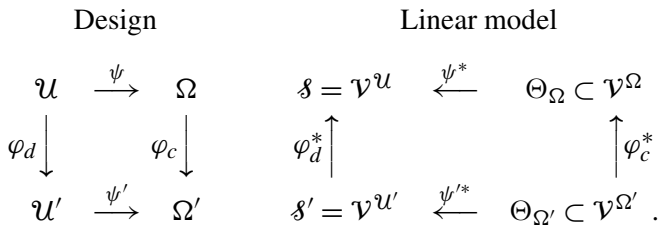
commutes [Tjur (2000)]. In other words $x'\varphi_d$ and $\varphi_c x$ represent the same design $\mathcal{U} \to \Omega'$. In matrix notation, $X' = UXW$ in which U is a row selection matrix, and W is a code-transformation matrix.

The general idea behind this construction can be understood by asking what it means for one design to be embedded in another. Here we consider simple embeddings obtained by selection of units or selection of covariate values. First, consider the effect of selecting a subset $\mathcal{U} \subset \mathcal{U}'$ of the units and discarding the remainder. Let $\varphi_d : \mathcal{U} \to \mathcal{U}'$ be the insertion map that carries each $u \in \mathcal{U}$ to itself as an element of $\mathcal{U}'$. The design that remains when the units not in $\mathcal{U}$ are discarded is the composition $x'\varphi_d : \mathcal{U} \to \Omega'$, which is the restriction of x' to $\mathcal{U}$. The diagram

may thus be completed by taking $\Omega = \Omega'$, φ_c the identity, and $x = x'\varphi_d$. Next, consider the effect of selection based on covariate values, that is, selecting only those units $\mathcal{U}$ whose covariate values lie in the subset $\Omega \subset \Omega'$. Let $\varphi_c : \Omega \to \Omega'$ and $\varphi_d : \mathcal{U} \to \mathcal{U}'$ be the associated insertion maps. The design map $x : \mathcal{U} \to \Omega$ is given by x' restricted to the domain $\mathcal{U}$ and codomain $\Omega \subset \Omega'$. Finally, if $\mathcal{U} = \mathcal{U}'$, $\Omega = \Omega'$ and φ_c is a permutation of Ω, the design $\varphi_c x$ is simply the design x with a rearrangement of labels. The same design could be obtained by suitably permuting the units before using the design map x, that is, the design $x\varphi_d$.

4.2. *Model.* Let $\mathcal{V}$ be a fixed response scale. A response y on $\mathcal{U}$ is a function $y : \mathcal{U} \to \mathcal{V}$, a point in the sample space $\mathcal{V}^{\mathcal{U}}$. To each set $\mathcal{U}$ there corresponds a sample space $\mathcal{V}^{\mathcal{U}}$ of $\mathcal{V}$-valued functions on $\mathcal{U}$. Likewise, to each injective morphism $\varphi : \mathcal{U} \to \mathcal{U}'$ in $\mathrm{cat}_{\mathcal{U}}$ there corresponds a coordinate-projection map $\varphi^* : \mathcal{V}^{\mathcal{U}'} \to \mathcal{V}^{\mathcal{U}}$. For $f \in \mathcal{V}^{\mathcal{U}'}$, the pullback map defined by functional composition $\varphi^* f = f \circ \varphi$ is a $\mathcal{V}$-valued function on $\mathcal{U}$. Thus $(\mathcal{V}, *)$ is a functor on $\mathrm{cat}_{\mathcal{U}}$, associating with each set $\mathcal{U}$ the sample space $\mathcal{V}^{\mathcal{U}}$, and with each morphism $\varphi : \mathcal{U} \to \mathcal{U}'$ the map $\varphi^* : \mathcal{V}^{\mathcal{U}'} \to \mathcal{V}^{\mathcal{U}}$ by functional composition. The identity map $\mathcal{U} \to \mathcal{U}$ is carried to the identity $\mathcal{V}^{\mathcal{U}} \to \mathcal{V}^{\mathcal{U}}$ and the composite map $\psi\varphi : \mathcal{U} \to \mathcal{U}''$ to the composite $\varphi^*\psi^* : \mathcal{V}^{\mathcal{U}''} \to \mathcal{V}^{\mathcal{U}}$ in reverse order.

Before presenting a general definition of a statistical model, it may be helpful to give a definition of a linear model. Let $\mathcal{V}$ be a vector space, so that the sample space $\mathcal{V}^{\mathcal{U}}$ is also a vector space. A linear model is a subspace of $\mathcal{V}^{\mathcal{U}}$, suitably related to the design. In the functor diagram below, each map in the right square is a linear transformation determined by functional composition. Thus, for $f \in \mathcal{V}^{\Omega'}$, the pullback by φ_c is $\varphi_c^* f = f \circ \varphi_c$, which is a vector in $\mathcal{V}^{\Omega}$. Likewise, $\psi'^* f = f \circ \psi'$ is a vector in $\mathcal{V}^{\mathcal{U}'}$,



As defined in McCullagh (2000), a linear model is determined by a subrepresentation $\Theta_\Omega \subset \mathcal{V}^{\Omega}$, together with the design pullback map ψ^*. A subrepresentation Θ in the standard representation $\mathcal{V}^{\Omega}$ is a sequence of subspaces $\{\Theta_\Omega \subset \mathcal{V}^{\Omega}\}$, indexed by the objects in cat_Ω such that, for each map $\varphi_c : \Omega \to \Omega'$ in cat_Ω, the linear transformation $\varphi_c^* : \mathcal{V}^{\Omega'} \to \mathcal{V}^{\Omega}$ also satisfies $\varphi_c^* \Theta_{\Omega'} = \Theta_\Omega$.

This skeleton description of a linear model focuses only on the linear subspace, and says nothing about probability distributions on the sample space. In the usual complete formulation, the parameter space is extended to $\Theta \times \mathcal{R}^+$ by the inclusion of a dispersion parameter σ^2. For the design $\psi : \mathcal{U} \to \Omega$, the sample space is

$\mathcal{S} = \mathcal{V}^{\mathcal{U}}$, and the full normal-theory linear model is the set of normal distributions with mean vector in the subspace $\psi^*\Theta_\Omega$, and variance matrix proportional to the identity.

The skeletal description is shared by nonnormal linear models, certain nonlinear models and all generalized linear models [McCullagh and Nelder (1989)]. The only modification required is the step by which the representation Θ determines a family of distributions on $\mathcal{S}$. In generalized linear models, this step is determined by the linear predictor $\eta = \psi^*\theta$ from Θ into $\mathcal{V}^{\mathcal{U}}$, an invertible link function $\eta = g(\mu)$ that is also a natural componentwise transformation $\mathcal{V}^{\mathcal{U}} \to \mathcal{V}^{\mathcal{U}}$, together with a suitable error distribution with independent components.

The general construction follows the same lines. For each fixed $\mathcal{V}$, the parameter $(\Theta, *)$ is a contravariant functor $\mathrm{cat}_\Omega \to \mathcal{K}$ associating with each covariate set Ω a parameter set Θ_Ω, and with each (injective) covariate morphism $\varphi_c : \Omega \to \Omega'$ a surjective parameter map $\varphi_c^* : \Theta_{\Omega'} \to \Theta_\Omega$. Frequently, $\mathcal{K}$ is the category of surjective linear transformations on vector spaces, in which case Θ is called a representation of cat_Ω. A statistical model is a functor on $\mathrm{cat}_\mathcal{D}$ associating with each design object $\psi : \mathcal{U} \to \Omega$ a model object, which is a map $P_\psi : \Theta_\Omega \to \mathcal{P}(\mathcal{S})$ such that $P_\psi\theta$ is a probability distribution on $\mathcal{S}$. The set of distributions thus generated is $\mathcal{F}_\psi = P_\psi\Theta_\Omega$. To each morphism $(\varphi_d, \varphi_c) : \psi \to \psi'$ in $\mathrm{cat}_\mathcal{D}$, the functor associates a map $(\varphi_d^\dagger, \varphi_c^*) : P_{\psi'} \to P_\psi$ as illustrated below:

$$
\begin{array}{cccccccc}
 & \text{Design} & & & \text{Sample space} & & \text{Model} & \\
(1) & \mathcal{U} & \xrightarrow{\psi} & \Omega & \mathcal{S} = \mathcal{V}^{\mathcal{U}} & \mathcal{P}(\mathcal{S}) & \xleftarrow{P_\psi} & \Theta_\Omega \\
 & \varphi_d \downarrow & & \varphi_c \downarrow & \varphi_d^* \uparrow & \varphi_d^\dagger \uparrow & & \varphi_c^* \uparrow \\
 & \mathcal{U}' & \xrightarrow{\psi'} & \Omega' & \mathcal{S}' = \mathcal{V}^{\mathcal{U}'} & \mathcal{P}(\mathcal{S}') & \xleftarrow{P_{\psi'}} & \Theta_{\Omega'} \ .
\end{array}
$$

As usual in such diagrams, the maps in both squares are assumed to commute. Some consequences of this definition are as follows:

1. The sample-space transformation $\varphi_d^* : \mathcal{S}' \to \mathcal{S}$ also carries each distribution F on $\mathcal{S}'$ to the transformed distribution $\varphi_d^\dagger F = F \circ \varphi_d^{*-1}$ on $\mathcal{S}$.
2. Commutativity: $P_\psi\varphi_c^* = \varphi_d^\dagger P_{\psi'} : \Theta_{\Omega'} \to \mathcal{P}(\mathcal{S})$. In other symbols, for each $\theta \in \Theta_{\Omega'}$, $P_\psi\varphi_c^*\theta = \varphi_d^\dagger P_{\psi'}\theta$.

Condition (1) ensures that the family of distributions on $\mathcal{S}$ is suitably embedded in the family of distributions on $\mathcal{S}'$. The maps $\varphi_d^* : \mathcal{S}' \to \mathcal{S}$ that define this embedding depend on the category of morphisms on units. For φ_c equal to the identity on Ω, and thus φ_c^* the identity on Θ_Ω, the consistency condition (2) asserts that the distribution $P_\psi\theta$ on $\mathcal{S}$ is the same as the marginal distribution of $P_{\psi'}\theta$ on $\mathcal{S}'$ by the embedding map $\varphi_d^* : \mathcal{S}' \to \mathcal{S}$.

The association $\mathcal{U} \mapsto \mathcal{V}^{\mathcal{U}}$, $\varphi_d \mapsto \varphi_d^*$ determines the structure of the sample spaces as functor on $\mathrm{cat}_{\mathcal{U}}$, and thus as a functor on $\mathrm{cat}_{\mathcal{D}}$. The set of morphisms of sample spaces is in fact a category isomorphic with $\mathrm{cat}_{\mathcal{U}}^{\mathrm{op}}$, the category obtained by reversing all arrows in $\mathrm{cat}_{\mathcal{U}}$. Likewise, the association $\mathcal{U} \mapsto \mathcal{P}(\mathcal{V}^{\mathcal{U}})$, $\varphi_d \mapsto \varphi_d^\dagger$ is also a functor on $\mathrm{cat}_{\mathcal{U}}$ whose image is isomorphic with $\mathrm{cat}_{\mathcal{U}}^{\mathrm{op}}$. The mapping $\mathcal{S} \mapsto \mathcal{P}(\mathcal{S})$, $\varphi_d^* \mapsto \varphi_d^\dagger$ is an isomorphism of categories. The parameter Θ is a functor on cat_{Ω} and thus also a functor on $\mathrm{cat}_{\mathcal{D}}$. The model $P : \Theta \to \mathcal{P}(\mathcal{S})$ is thus a natural transformation of functors on $\mathrm{cat}_{\mathcal{D}}$.

Let $P, Q : \Theta \to \mathcal{P}(\mathcal{S})$ be two models having the same domain Θ. For each $0 \le \alpha \le 1$, the convex combination $\alpha P + (1-\alpha)Q$ is also a statistical model, so the set of models $\Theta \to \mathcal{P}(\mathcal{S})$ is convex. The majority of non-Bayesian models that occur in practice are extreme points of this set [Lauritzen (1988)].

A glance at almost any book on statistical modelling shows that the concepts of a model "making sense," and a parameter "having a meaning" are deeply ingrained in statistical thinking and well understood at an instinctive level. The purpose of the preceding definitions is to state in purely mathematical terms what is meant by a model "making sense" logically (as opposed to making sense physically or biologically). The sense of a model and the meaning of a parameter, whatever they may be, must not be affected by accidental or capricious choices such as sample size or experimental design. The definition of the map $\varphi_d^\dagger$, together with the commutativity condition (2), ensures that the meaning of the parameter is retained for families on different sample spaces and different designs. The parameter sets Θ_Ω, and the maps P and φ_c^* are not otherwise prescribed, so the framework provides ample opportunity for imaginative construction of specific models. Note, for example, that if φ_c is the identity on Ω, $\varphi_c^\dagger$ is the identity on Θ_Ω, so the marginal distribution of $P_{\psi'}\theta$ by the map $\varphi_d^\dagger$ is equal to $P_\psi \theta$.

4.3. *Composition, mixtures and hierarchical models.* To each set S we associate the vector space $\mathrm{vect}(S)$ of formal linear combinations $\alpha_1 s_1 + \alpha_2 s_2 + \cdots$ of the elements of S. By restricting these formal linear combinations to formal convex combinations in which $\alpha_i \ge 0$ and $\sum \alpha_i = 1$, we obtain the set $\mathcal{P}(S) \subset \mathrm{vect}(S)$ of probability distributions on S. If S is countable, each formal convex combination $\alpha_1 s_1 + \alpha_2 s_2 + \cdots$ may be interpreted as an actual operation in which an element of $\{s_1, s_2, \ldots\}$ is chosen according to the probability distribution $\{\alpha_1, \alpha_2, \ldots\}$. The point of this abstract construction is that each function $h : S \to T$ determines a linear transformation $h^* : \mathrm{vect}(S) \to \mathrm{vect}(T)$ by

$$h^*(\alpha_1 s_1 + \alpha_2 s_2 + \cdots) = \alpha_1 h(s_1) + \alpha_2 h(s_2) + \cdots .$$

Evidently h^* also carries the subset $\mathcal{P}(S)$ into $\mathcal{P}(T)$. The extreme points of $\mathcal{P}(S)$ may thus be identified with the elements of S. This correspondence between $h : S \to T$ and $h^* : \mathrm{vect}(S) \to \mathrm{vect}(T)$ is so natural that, wherever possible, the

notational distinction is avoided. Given $h : S \to T$, formal linearity determines $h : \mathcal{P}(S) \to \mathcal{P}(T)$.

Consider a design $\psi : \mathcal{U} \to \Omega$ and a distribution F on Θ_Ω. The model $P : \Theta \to \mathcal{P}(\mathcal{Y})$ associates with each point $\theta \in \Theta_\Omega$ a distribution $P_\psi\theta$ on $\mathcal{Y}_\mathcal{U}$. Since F is a formal convex combination of points in Θ_Ω, $P_\psi F$ is a similar formal combination of points in $\mathcal{P}(\mathcal{Y})$, that is, a joint distribution on $\Theta_\Omega \times \mathcal{Y}_\mathcal{U}$. In other words, to each model component $P_\psi : \Theta_\Omega \to \mathcal{P}(\mathcal{Y}_\mathcal{U})$ there corresponds a map $P_\psi : \mathcal{P}(\Theta_\Omega) \to \mathcal{P}(\Theta_\Omega \times \mathcal{Y}_\mathcal{U})$, associating with each prior distribution F on Θ_Ω a joint distribution $P_\psi F$ on $\Theta_\Omega \times \mathcal{Y}_\mathcal{U}$ by formal convex combinations. The marginal distribution on $\mathcal{Y}_\mathcal{U}$ is obtained by integration over Θ_Ω, that is, by treating the formal linear combination F as an actual linear combination.

A hierarchical model is a chain $\Theta^1, \Theta^2, \ldots, \Theta^k$ of parameter functors, together with a chain of natural transformations,

$$P_1 : \Theta^1 \to \mathcal{P}(\Theta^2), \quad P_2 : \Theta^2 \to \mathcal{P}(\Theta^3), \ldots, \quad P_k : \Theta^k \to \mathcal{P}(\mathcal{Y}).$$

By the compositional operation described above, we have

$$(P_k P_{k-1} \cdots P_1) : \Theta^1 \to \mathcal{P}(\Theta^2 \times \cdots \times \Theta^k \times \mathcal{Y}).$$

This composite function is not a statistical model according to the definition, but it can be converted into one by the natural transformation of integration in which the formal linear combination $P_{k-1} \cdots P_1$ is interpreted as an actual linear combination,

$$\Theta^1 \xrightarrow{P_k \cdots P_1} \mathcal{P}(\Theta^2 \times \cdots \times \Theta^k \times \mathcal{Y}) \to \mathcal{P}(\mathcal{Y}).$$

A Bayesian model is a hierarchical model in which $k > 1$ and $\Theta^1 = \{0\}$ is the trivial functor, a one-point set with identity map.

4.4. *Submodel.* Let $\Theta, \Xi : \mathrm{cat}_\Omega \to \mathcal{K}$ be two contravariant functors from cat_Ω into the category of surjective maps. To each model $P : \Theta \to \mathcal{P}(\mathcal{Y})$ and natural transformation $h : \Xi \to \Theta$ there corresponds a submodel $Ph : \Xi \to \mathcal{P}(\mathcal{Y})$ by restriction of P to the image of h. These relationships are illustrated by the commutative diagram

$$\begin{array}{ccccccc}
\psi & & \mathcal{P}(\mathcal{Y}) & \xleftarrow{P_\psi} & \Theta_\Omega & \xleftarrow{h_\Omega} & \Xi_\Omega \\
{\scriptstyle(\varphi_d, \varphi_c)}\downarrow & & \uparrow{\scriptstyle \varphi_d^\dagger} & & \uparrow{\scriptstyle \varphi_c^*} & & \uparrow{\scriptstyle \varphi_c'} \\
\psi' & & \mathcal{P}(\mathcal{Y}') & \xleftarrow{P_{\psi'}} & \Theta_{\Omega'} & \xleftarrow{h_{\Omega'}} & \Xi_{\Omega'} \ .
\end{array}$$

4.5. *Subparameter.* The notion that only certain functions of a parameter "make sense," or "have a meaning" is well understood and formalized to some extent by dimensional analysis in physics. Thus, in an i.i.d. normal model for the weight distribution of apples at harvest, it might make sense to compare the

subparameter σ/μ for different varieties, but the "subparameters" $\mu + \sigma^2$ and $\mu + \sigma/\mu$ are clearly meaningless. We are thus led to the concept of a natural subparameter.

A *natural subparameter* is a natural transformation $g : \Theta \to \Xi$ of functors on $\mathrm{cat}_{\mathcal{D}}$. To each design $\psi : \mathcal{U} \to \Omega$ the natural transformation g associates a map $g_\psi : \Theta_\Omega \to \Xi_\Omega$ in such a way that the diagram shown below commutes

$$\begin{array}{ccccc} \psi & & \Theta_\Omega & \xrightarrow{g_\psi} & \Xi_\Omega \\ {\scriptstyle (\varphi_d, \varphi_c)}\downarrow & & \varphi_c^*\uparrow & & \varphi_c'\uparrow \\ \psi' & & \Theta_{\Omega'} & \xrightarrow{g_{\psi'}} & \Xi_{\Omega'} \ . \end{array}$$

In other words $g_\psi \varphi_c^* = \varphi_c' g_{\psi'}$.

Let $g : \Theta \to \Xi$ and $h : \Theta \to \Psi$ be two natural subparameters such that $(g, h) : \Theta \to \Xi \times \Psi$ is a natural isomorphism of functors. When such a pair exists, we say that g, h are complementary subparameters in Θ. In general, for a given subparameter g there need not exist a complementary subparameter. When a complementary parameter exists, it need not be unique. The notion of orthogonal parameters [Cox and Reid (1987)] does not arise here unless the parameter sets Θ_Ω are inner product spaces, and the maps φ_c^* preserve inner products. Such functors do arise in rather specialized applications.

4.6. *Identifiability.* The parameter Θ is said to be identifiable at the design $\psi : \mathcal{U} \to \Omega$ if distinct parameter values give rise to distinct distributions on the sample space. In other words, Θ is identifiable at ψ if the associated map $P_\psi : \Theta_\Omega \to \mathcal{P}(\mathcal{S})$ is injective.

Likewise, a natural subparameter Ξ is identifiable at ψ if distinct ξ-values give rise to nonoverlapping sets of distributions. To each point $\xi \in \Xi_\Omega$ there corresponds a set of points $g_\psi^{-1}\xi$ in Θ_Ω, and a set of distributions $P_\psi g_\psi^{-1}\xi$ in $\mathcal{F}_\psi = P_\psi \Theta_\Omega$. The subparameter $g\Theta \to \Xi$ is said to be identifiable at ψ if, for each $\xi \neq \xi'$ in Ξ_Ω, the sets

$$P_\psi g_\psi^{-1}\xi \quad \text{and} \quad P_\psi g_\psi^{-1}\xi'$$

are nonempty and nonoverlapping. An identifiable subparameter is thus a surjective natural transformation $g : \Theta \to \Xi$ that generates a partition of $\mathcal{F}_\psi$ by nonoverlapping sets.

4.7. *Response-scale morphisms.* In the discussion thus far, the response scale has been kept fixed. That is to say, the analysis has not considered the effect on the model of changing the units of measurement from, say, bushels per acre to tonnes per hectare, or of aggregating selected levels of an ordinal response variable. To complete the story, it is necessary to consider the various morphisms $\gamma : \mathcal{V} \to \mathcal{V}'$

of the response scale and their effect on the sample space and model. For each response scale $\mathcal{V}$ in $\mathrm{cat}_{\mathcal{V}}$, the sample space is a contravariant functor on $\mathrm{cat}_{\mathcal{U}}$, and the model is a contravariant functor on the design. To each morphism $\gamma : \mathcal{V} \to \mathcal{V}'$ in $\mathrm{cat}\,\mathcal{V}$ there corresponds a natural transformation of sample spaces as illustrated in the commutative diagram below:

$$\begin{array}{ccccc} & & \mathcal{V} & \xrightarrow{\gamma} & \mathcal{V}' \\ \\ \mathcal{U} & & \mathcal{V}^{\mathcal{U}} & \xrightarrow{\gamma_{\mathcal{U}}} & \mathcal{V}'^{\mathcal{U}} \\ \varphi \downarrow & & \varphi^*_{\mathcal{V}} \uparrow & & \uparrow \varphi^*_{\mathcal{V}'} \\ \mathcal{U}' & & \mathcal{V}^{\mathcal{U}'} & \xrightarrow{\gamma_{\mathcal{U}'}} & \mathcal{V}'^{\mathcal{U}'} \end{array} .$$

The category of morphisms of sample spaces, $\mathrm{cat}_{\mathcal{S}}$, is the category whose objects are all sample spaces $\mathcal{S} = \mathcal{V}^{\mathcal{U}}$ with $\mathcal{V} \in \mathrm{cat}_{\mathcal{V}}$ and $\mathcal{U} \in \mathrm{cat}_{\mathcal{U}}$. The morphisms $\mathcal{V}^{\mathcal{U}'} \to \mathcal{V}'^{\mathcal{U}}$ are all compositions $\gamma_{\mathcal{U}}\varphi^*_{\mathcal{V}} = \varphi^*_{\mathcal{V}'}\gamma_{\mathcal{U}'}$ in which $\gamma : \mathcal{V} \to \mathcal{V}'$ is a morphism of the codomain, and φ^* is a functional composition with the domain morphism $\varphi : \mathcal{U} \to \mathcal{U}'$. In this way, the category of morphisms on sample spaces is built up from a category of injective morphisms on units, and a category of morphisms on response scales.

In fact, $\mathrm{cat}_{\mathcal{S}}$ is isomorphic with the product category $\mathrm{cat}_{\mathcal{V}} \times \mathrm{cat}^{\mathrm{op}}_{\mathcal{U}}$, where $\mathrm{cat}^{\mathrm{op}}_{\mathcal{U}}$ is the opposite category derived from $\mathrm{cat}_{\mathcal{U}}$ by the process of reversing all arrows. The categories $\mathrm{cat}^{\mathrm{op}}_{\mathcal{D}}$ and $\mathrm{cat}^{\mathrm{op}}_{\Omega}$ are defined in the same way. The logical structure of a statistical model is then illustrated by the following functor sequences, which show that $\mathrm{cat}_{\mathcal{S}}$ and the parameter Θ are covariant functors on $\mathrm{cat}_{\mathcal{V}} \times \mathrm{cat}^{\mathrm{op}}_{\mathcal{D}}$.

$$\begin{aligned} &\mathrm{cat}_{\mathcal{V}} \times \mathrm{cat}^{\mathrm{op}}_{\mathcal{D}} \to \mathrm{cat}_{\mathcal{V}} \times \mathrm{cat}^{\mathrm{op}}_{\mathcal{U}} \cong \mathrm{cat}_{\mathcal{S}} \cong \mathcal{P}(\mathcal{S}), \\ &\mathrm{cat}_{\mathcal{V}} \times \mathrm{cat}^{\mathrm{op}}_{\mathcal{D}} \to \mathrm{cat}_{\mathcal{V}} \times \mathrm{cat}^{\mathrm{op}}_{\Omega} \xrightarrow{\Theta} \mathcal{K}. \\ &\text{Model: } P : \Theta \to \mathcal{P}(\mathcal{S}). \end{aligned} \tag{2}$$

The functor $\mathrm{cat}_{\mathcal{S}} \to \mathcal{P}(\mathcal{S})$ is an isomorphism, associating with each sample space $\mathcal{S} = \mathcal{V}^{\mathcal{U}}$ the set of all probability distributions on $\mathcal{S}$, and with each morphism $f : \mathcal{S} \to \mathcal{S}'$ of sample spaces a morphism on distributions defined by composition with the inverse image of f. A model is a natural transformation $P : \Theta \to \mathcal{P}(\mathcal{S})$ associating with each pair $(\mathcal{V}, \psi : \mathcal{U} \to \Omega)$ a map $P_\psi : \Theta \to \mathcal{P}(\mathcal{V}^{\mathcal{U}})$ satisfying the properties in (1).

In the absence of covariate effects, the commutative diagram of a statistical model takes the following simplified form in which the model $P : \Theta \to \mathcal{P}(\mathcal{S})$ is indexed by sample spaces $\mathcal{S}, \mathcal{S}', \ldots$. First, the parameter space $(\Theta, \dagger)$ is a covariant functor on the category $\mathrm{cat}_{\mathcal{V}}$, associating with each response scale $\mathcal{V}$, a parameter set $\Theta_{\mathcal{V}}$, and with each morphism $\gamma : \mathcal{V} \to \mathcal{V}'$ a parameter morphism

$\gamma^\dagger : \Theta_{\mathcal{V}} \to \Theta_{\mathcal{V}'}$. Second, to each morphism of sample spaces $\gamma\varphi^* : \mathcal{V}^{\mathcal{U}'} \to \mathcal{V}^{\mathcal{U}}$, the transformation on the model is described by the commutative diagram,

$$\begin{array}{ccccc} \mathcal{S}' = \mathcal{V}'^{\mathcal{U}} & & \mathcal{P}(\mathcal{S}') & \xleftarrow{P_{\mathcal{S}'}} & \Theta_{\mathcal{V}'} \\ {\scriptstyle \gamma\varphi^*}\uparrow & & \uparrow{\scriptstyle \circ(\gamma\varphi^*)^{-1}} & & \uparrow{\scriptstyle \gamma^\dagger} \\ \mathcal{S} = \mathcal{V}^{\mathcal{U}'} & & \mathcal{P}(\mathcal{S}) & \xleftarrow{P_{\mathcal{S}}} & \Theta_{\mathcal{V}} \quad . \end{array}$$

To see what this means, consider a random variable Y taking values in $\mathcal{S} = \mathcal{V}^{\mathcal{U}'}$ with distribution $P_{\mathcal{S}}\theta$ for some $\theta \in \Theta_{\mathcal{V}}$. Each morphism $\mathcal{S} \to \mathcal{S}'$ is the composition of a coordinate projection map φ^* and a response-scale morphism γ. The marginal distribution of Y is $(P_{\mathcal{S}}\theta) \circ (\gamma\varphi^*)^{-1}$ on $\mathcal{S}'$. Commutativity ensures that this distribution is also equal to the image under $P_{\mathcal{S}'}$ of the induced parameter $\gamma^\dagger\theta$. In other words, $P : \Theta \to \mathcal{P}(\mathcal{S})$ is a natural transformation of functors.

5. Special cases. The statement in equations (1) and (2), that a statistical model is a natural transformation $P : \Theta \to \mathcal{P}(\mathcal{S})$ of functors, is a consistency condition ensuring that different ways of computing the probability of equivalent events give the same answer. As a consequence, equivalent events determine the same likelihood. Some familiar instances are as follows.

Equivariance and transformation models. Let $\mathcal{C} = \mathcal{G}$ be a group acting on the response scale $\mathcal{V}$. The group induces a homomorphic action $\mathcal{S} \to \mathcal{S}$ on the sample space, usually by componentwise transformation. Likewise, the same group induces a homomorphic action $\Theta \to \Theta$ on the parameter space. The group actions on $\mathcal{S}$ and on Θ are conventionally denoted by the same symbol $y \mapsto gy$ and $\theta \mapsto g\theta$, respectively. Ordinarily, this abuse of notation leads to no confusion because the functors $\mathcal{G} \to \mathcal{S}$ and $\mathcal{G} \to \Theta$ are usually group isomorphisms. The group transformation formulation is an assertion that the event-parameter pairs $(E; \theta)$ and $(gE; g\theta)$ are equivalent. The commutativity condition may then be written in the form $P(E; \theta) = P(gE; g\theta)$. Examples include the Cauchy model, location-scale models, and the Box–Cox model (Section 7).

A natural, or equivariant, subparameter [Lehmann and Casella (1998), page 160], or a permissible subparameter [Helland (1999a, b)] is a function $h : \Theta \to \Xi$ such that $h(\theta_1) = h(\theta_2)$ implies $h(g\theta_1) = h(g\theta_2)$ for each $g \in \mathcal{G}$. Provided that h is surjective, this condition ensures that $hgh^{-1} : \Xi \to \Xi$ is well defined, and is a group homomorphism. The term "invariantly estimable parameter" has been used by Hora and Buehler (1966), but this terminology is doubly misleading. First, a natural parameter is equivariant and not ordinarily invariant under the group, that is, hgh^{-1} need not be the identity on Ξ. Second, estimability depends on the design, and a natural parameter need not be estimable or even identifiable at a given design.

Insertion maps and Kolmogorov consistency. Let $\mathrm{cat}_{\mathcal{U}}$ be the category of insertion maps on finite sets, $\mathcal{V} = \mathcal{R}$ a fixed set, $\mathrm{cat}_{\Omega} = \{0\}$ a one-point set, and let $\mathcal{S} = \mathcal{R}^{\mathcal{U}}$ be the standard contravariant representation. Then Θ is necessarily a set with one object and identity map. For each finite set $\mathcal{U} = \{i_1, \ldots, i_m\} \subset \mathcal{U}' = \{1, \ldots, n\}$ the insertion map $\varphi : \mathcal{U} \to \mathcal{U}'$ carries each $i \in \mathcal{U}$ to itself ($\varphi i = i$) as an element in $\mathcal{U}'$. The pullback map $\varphi^* : \mathcal{R}^{\mathcal{U}'} \to \mathcal{R}^{\mathcal{U}}$ is a linear coordinate-projection map that deletes all coordinates except those in $\mathcal{U}$. For simplicity of notation, assume that $i_r = r$, so that $\mathcal{U}$ consists of the first m elements of $\mathcal{U}'$. The implication of using insertion maps is that the event $E = A_1 \times \cdots \times A_m \subset \mathcal{R}^{\mathcal{U}}$ and the event

$$E' = \varphi^{*-1} E = A_1 \times \cdots \times A_m \times \mathcal{R} \times \cdots \times \mathcal{R} \subset \mathcal{R}^{\mathcal{U}'}$$

are equivalent. The natural transformation $P : \Theta \to \mathcal{P}(\mathcal{R}^{\Omega})$ in equation (1) associates with each parameter point $\theta \in \Theta$ a probability distribution $P_{\mathcal{U}}(\cdot; \theta)$ on $\mathcal{R}^{\mathcal{U}}$ in such a way that these equivalent events have the same probability $P_{\mathcal{U}}(E; \theta) = P_{\mathcal{U}'}(E'; \theta)$. Commutativity is equivalent to the statement that the Kolmogorov existence condition for a stochastic process [Billingsley (1986), equation (36.3)], is satisfied at each parameter point.

Interference and insertion maps. Let $\mathrm{cat}_{\mathcal{U}}$ be the category of insertion maps on finite sets, $\mathcal{V} = \mathcal{R}$ a fixed set, and let $\mathcal{S} = \mathcal{R}^{\mathcal{U}}$ be the standard contravariant representation. Let $\mathcal{U} = \{1, \ldots, m\}$, $\mathcal{U}' = \{1, \ldots, n\}$ and let $\varphi_d : \mathcal{U} \to \mathcal{U}'$ be the insertion map, implying $m \le n$. Consider a morphism of designs $\psi \mapsto \psi'$ in which $\varphi_c : \Omega \to \Omega$ is the identity. In other words, ψ is obtained from ψ' by ignoring, or forgetting, the units $\{m+1, \ldots, n\}$. The commutativity condition in equation (1) is simply the statement that, for each $\theta \in \Theta_{\Omega}$, the events E and E' as defined above have the same probability $P_{\psi}(E; \theta) = P_{\psi'}(E'; \theta)$. This is a stronger condition than Kolmogorov consistency because the design ψ' carries information on the association of covariates, or assignment of treatments, to all units including those unobserved units $\{m+1, \ldots, n\}$ for which the event E' does not determine a value. In the literature on experimental design, this assumption is called lack of interference [Cox (1958), Section 2.4; Rubin (1986)].

If cat_{Ω} includes insertion maps, $\varphi_c : \Omega \to \Omega'$ need not be the identity map. The commutativity condition (1) $P_{\psi}(E; \varphi_c^* \theta) = P_{\psi'}(E'; \theta)$ implies not only that lack of interference holds for each family $P_{\psi} : \Theta_{\Omega} \to \mathcal{P}(\mathcal{R}^{\mathcal{U}})$, but also that, for $\Omega \ne \Omega'$ the distributions in each family are related in a natural way. For each θ' such that $\varphi_c^* \theta' = \theta$, $P_{\psi}(E; \theta) = P_{\psi'}(E'; \theta')$.

Exchangeability and injection maps. Let $\mathrm{cat}_{\mathcal{U}} = \mathcal{I}$ be the category of injective maps on finite sets, $\mathcal{V} = \mathcal{R}$ a fixed set, $\mathrm{cat}_{\Omega} = \{0\}$ a one-point set, and let $\mathcal{S} = \mathcal{R}^{\mathcal{U}}$ be the standard contravariant representation. Then Θ is necessarily a set with one object and identity map. To each injective map $\varphi : \mathcal{U} \to \mathcal{U}'$ in $\mathcal{I}$, the standard

(contravariant) representation associates the surjective linear map $\varphi^*: \mathcal{R}^{\mathcal{U}'} \to \mathcal{R}^{\mathcal{U}}$ by functional composition: $(\varphi^* f)(i) = f(\varphi(i))$ for $i \in \mathcal{U}$. For $\mathcal{U} = \mathcal{U}'$, the linear map φ^* permutes coordinates: more generally, φ^* is a coordinate-projection map.

The commutativity condition in (1) is the statement that, to each $\theta \in \Theta$ there corresponds a distribution $P_{\mathcal{U}}\theta$ on $\mathcal{R}^{\mathcal{U}}$ that satisfies the following conditions.

1. For each permutation $\varphi: \mathcal{U} \to \mathcal{U}$, the coordinate-permutation map $\varphi^*: \mathcal{R}^{\mathcal{U}} \to \mathcal{R}^{\mathcal{U}}$ carries $P_{\mathcal{U}}\theta$ to itself. In other words, $P_{\mathcal{U}}\theta$ is invariant under coordinate permutation.
2. For each injection $\varphi: \mathcal{U} \to \mathcal{U}'$, the coordinate-projection map $\varphi^*: \mathcal{R}^{\mathcal{U}'} \to \mathcal{R}^{\mathcal{U}}$ carries the distribution $P_{\mathcal{U}'}\theta$ on $\mathcal{R}^{\mathcal{U}'}$ to $P_{\mathcal{U}}\theta$ on $\mathcal{R}^{\mathcal{U}}$.

In other words, to each point $\theta \in \Theta$ there corresponds a sequence $\{P_{\mathcal{U}}\theta\}$ that is the distribution of an infinitely exchangeable sequence of random variables in the sense of de Finetti [(1975), Section 11.4]. The set $\mathrm{Inv}_{\mathbb{1}}(\mathcal{R}^{\mathcal{U}})$ of all natural transformations $\{0\} \to \mathcal{P}(\mathcal{R}^{\mathcal{U}})$ is the set of all infinitely exchangeable probability models on the standard representation.

Partial exchangeability and $\mathbb{1}^2$. The preceding example may be modified by restricting the objects and maps in $\mathrm{cat}_{\mathcal{U}}$. Let $\mathrm{cat}_{\mathcal{U}}$ be the product category $\mathbb{1}^2$ in which the objects are all Cartesian product sets $\mathcal{U} \times \mathcal{W}$ and the morphisms $\mathcal{U} \times \mathcal{W} \to \mathcal{U}' \times \mathcal{W}'$ are all injective maps preserving the product structure. Each morphism is an ordered pair (φ_1, φ_2) in which $\varphi_1: \mathcal{U} \to \mathcal{U}'$ and $\varphi_2: \mathcal{W} \to \mathcal{W}'$ are injective maps acting componentwise. The standard representation of $\mathbb{1}^2$ associates with each product set $\mathcal{U} \times \mathcal{W}$ the vector space $\mathcal{S}_{\mathcal{U}\mathcal{W}} = \mathcal{R}^{\mathcal{U}\times\mathcal{W}}$, and with each injective morphism (φ_1, φ_2), a surjective linear map $(\varphi_1, \varphi_2)^*: \mathcal{R}^{\mathcal{U}'\times\mathcal{W}'} \to \mathcal{R}^{\mathcal{U}\times\mathcal{W}}$ by functional composition,

$$(\varphi_1, \varphi_2)^* f(i, j) = f\big(\varphi_1(i), \varphi_2(j)\big)$$

for $f \in \mathcal{R}^{\mathcal{U}'\times\mathcal{W}'}$. For simplicity, assume that Θ is a set containing one element. The commutativity condition in (1) is the statement that, to each rectangular index array $\{1, \ldots, m\} \times \{1, \ldots, n\}$ there corresponds a distribution P_{mn} on $\mathcal{R}^{mn}$ such that, for $m' \geq m$ and $n' \geq n$, P_{mn} is the marginal distribution of $P_{m'n'}$ under all coordinate-projection maps $(\varphi_1, \varphi_2)^*$ in the standard representation. In particular, P_{mn} is invariant under row permutation and column permutation. In other words, P is the distribution of an infinite partially exchangeable array in the sense of Aldous (1981). Conversely, each infinite partially exchangeable array of random variables has a distribution that corresponds to an $\mathbb{1}^2$-invariant distribution. A model is thus a set of such distributions indexed by Θ. The set $\mathrm{Inv}_{\mathbb{1}^2}(\mathcal{S})$ of all $\mathbb{1}^2$-natural transformations $\{0\} \to \mathcal{P}(\mathcal{S})$ is the set of all invariant probability models on the standard $\mathbb{1}^2$-representation.

A completely randomized design. Consider a design in which $\mathrm{cat}_{\mathcal{U}} = \mathbb{1}$ implying that the units are infinitely exchangeable. We assume also that cat_{Ω} includes all insertion maps. In a linear or generalized linear model, Θ is a representation of cat_{Ω} by surjective linear transformations. For generalized linear models with unknown dispersion, the most common choice is the representation $\Theta_{\Omega} = \mathcal{R} \oplus \mathcal{R}^{\Omega}$, or the subrepresentation $\mathcal{R} \oplus 1$ implying equal treatment effects. Depending on the choice of $P : \Theta \to \mathcal{P}(\mathcal{S})$, the representation Θ may serve as an index set for various families, typically with independent components, such as $Y_i \sim N(\theta_{\psi(i)}, \theta_0^2)$, $Y_i \sim \mathrm{Cauchy}(\theta_{\psi(i)}, |\theta_0|)$, $Y_i \sim Po(\exp(\theta_{\psi(i)}))$, $Y_i \sim Bi(1, 1/(1 + \exp(-\theta_{\psi(i)})))$ in which $\psi(i)$ is the treatment associated with unit i. In addition to the two component parameters $\theta_0 \in \mathcal{R}$ and $\theta \in \mathcal{R}^{\Omega}$, the only natural linear subparameter $g : \Theta \to \Xi$ is the quotient projection $\mathcal{R}^{\Omega} \to \mathcal{R}^{\Omega}/1$, corresponding to the space of treatment contrasts [McCullagh (1999)].

Block designs. Let $\mathrm{cat}_{\mathcal{U}} = \mathcal{ND}$ be the category in which each object $\mathcal{U}$ is a set together with the equivalence relation, $u \sim u'$ if units u and u' belong to the same block. The morphisms $\varphi : \mathcal{U} \to \mathcal{U}'$ are all injective maps that preserve the equivalence relationship: $u \sim u'$ in $\mathcal{U}$ if and only if $\varphi u \sim \varphi u'$ in $\mathcal{U}'$. To say the same thing in a slightly different way, each object in $\mathcal{ND}$ is a set of plots together with a map $b : \text{plots} \to \text{blocks}$. If b is surjective, we may write $\mathcal{U} = \{(i, b(i)) : i \in \text{plots}\}$. The morphisms $\varphi : b \to b'$ are all ordered pairs (φ_d, φ_b) of injective maps such that the following diagram commutes:

$$\begin{array}{ccc} \text{plots} & \xrightarrow{\ b\ } & \text{blocks} \\ {\scriptstyle \varphi_d}\downarrow & & {\scriptstyle \varphi_b}\downarrow \\ \text{plots}' & \xrightarrow{\ b'\ } & \text{blocks}' \ . \end{array}$$

Let $\mathrm{cat}_{\Omega} = \{0\}$ be a one-point set, so that Θ is also a set. According to (1), a model for a block design with no specific factors is a natural transformation $P : \Theta \to \mathcal{P}(\mathcal{S})$ in which $\mathcal{S}$ is the representation $\mathcal{R}^{\mathcal{U}}$, meaning real-valued functions on plots. In other words, each $P\theta$ is a distribution in $\mathrm{Inv}_{\mathcal{ND}}(\mathcal{R}^{\mathcal{U}})$.

Let $\varepsilon_0, \{\varepsilon_i\}, \{\eta_b\}$ be independent standard normal random variables. Let Y be an infinite sequence of random variables indexed by the objects $\mathcal{U}$ such that, for some measurable function $g : \mathcal{R}^3 \to \mathcal{R}$, $Y_{\mathcal{U}}$ has the same distribution as $X_{\mathcal{U}}$ on $\mathcal{R}^{\mathcal{U}}$, where

$$X_{\mathcal{U}}(i) = g(\varepsilon_0, \varepsilon_i, \eta_{b(i)})$$

for $i \in \mathcal{U}$. Then Y has a distribution in $\mathrm{Inv}_{\mathcal{ND}}(\mathcal{S})$. The usual normal-theory variance-components model is obtained if g is a linear combination of ε_i and $\eta_{b(i)}$. Thus, we may have $\Theta = \mathcal{R}^3$, and $P\theta$ equal to the distribution of Y, where $Y_i = \theta_0 + \theta_1 \varepsilon_i + \theta_2 \eta_{b(i)}$.

Causality and nonspecific effects. Although the same category of morphisms arises in these last two examples, the standard models are quite different. If the units are regarded as having the infinitely exchangeable block structure associated with $\mathcal{ND}$, we arrive at a random-effects model. If the units are regarded as infinitely exchangeable modulo specific covariate effects, the model contains specific parameters for those effects. Nonspecific factors such as blocks, temporal structure and spatial structure, for which no level-specific inference is required, are best regarded as a property defining the structure of the units. Specific factors such as treatment, variety, age, race, sex or religion, for which level-specific inferences may be required, must be regarded as objects in cat_Ω. Level-specific effects may be ascribed to specific factors, whether intrinsic or not, but causal effects are not ordinarily ascribed to intrinsic factors [Holland (1986); Cox (1986)]. Thus, longevity may vary with race, sex and religion, in a specific manner, but it would be at best an abuse of language to assert that race, sex or religion is a cause of longevity.

The definitions given in Section 4 deliberately avoid all reference to causal mechanisms, which are necessarily context dependent in ways that categories and morphisms are not. A given model such as quasi-symmetry, arising in very different fields of application from linkage disequilibrium in genetics to citation studies in science, may be capable of numerous mechanistic interpretations. Also, models exist for which no mechanististic interpretation is readily available, or which are in conflict with accepted scientific laws. Such alternative models are necessary in order that scientific laws may be tested.

6. Examples.

6.1. *Location-scale models.* For simplicity of exposition in this section, we consider only models for independent and identically distributed scalar responses. Such models are determined by their one-dimensional marginal distributions on the response scale, so we may ignore $\mathrm{cat}_\mathcal{U}$ and cat_Ω. The motivation for location-scale models comes from transformations of the response scale, that is, the morphisms in $\mathrm{cat}_\mathcal{V}$. For definiteness, take $\mathrm{cat}_\mathcal{V}$ to be the category of morphisms of temperature scales. The objects in this category are all temperature scales, °F, °C, °K, and possibly others not yet invented. Associated with each scale $\mathcal{V}$ is a certain set of real numbers, and to each pair $(\mathcal{V}, \mathcal{V}')$ of scales there corresponds a single invertible affine map $(a, b) : \mathcal{V} \to \mathcal{V}'$, which transforms y on the $\mathcal{V}$-scale into $a + by$ on the $\mathcal{V}'$-scale. If $\mathcal{V} =$°C and $\mathcal{V}' =$°F, the map $(32, 9/5)$ carries y in °C to $32 + 9y/5$ in °F.

One example of a functor is the map $T : \mathrm{cat}_\mathcal{V} \to GA(\mathcal{R})$ in which each object in $\mathrm{cat}_\mathcal{V}$ is associated with $\mathcal{R}$, and each map $(a, b) : \mathcal{V} \to \mathcal{V}'$ is associated with an affine transformation $(a, b)^\dagger : \mathcal{R} \to \mathcal{R}$ defined by $x \mapsto a + bx$. If $\mathrm{cat}_\mathcal{V}$ contains the three objects °C, °F and °K, and nine maps, the image maps $\mathcal{R} \to \mathcal{R}$ in $T\,\mathrm{cat}_\mathcal{V}$ are the identity, the affine maps $(32, 9/5)$, $(273, 1)$, $(523.4, 9/5)$ and their inverses,

making a total of seven maps. These are not closed under composition, so the image $T\,\mathrm{cat}_{\mathcal{V}}$ is not a category. However, if all conceivable scales are included in $\mathrm{cat}_{\mathcal{V}}$, the image of T is equal to the affine subgroup with $b > 0$ acting on $\mathcal{R}$.

Another example of a functor is the map $\Theta : \mathrm{cat}_{\mathcal{V}} \to GA(\mathcal{R}^2)$ in which each $(a, b) : \mathcal{V} \to \mathcal{V}'$ is associated with the affine map $\mathcal{R}^2 \to \mathcal{R}^2$ defined by

$$(a, b)^\dagger : (\theta_1, \theta_2) \mapsto (a + b\theta_1, b\theta_2),$$

and we may take Θ as the parameter space for a location-scale model. Two examples of models $\Theta \to \mathcal{P}(\mathcal{V}^{\mathcal{U}})$ with independent components are

$$(\theta_1, \theta_2) \mapsto N(\theta_1, \theta_2^2) \quad \text{and} \quad (\theta_1, \theta_2) \mapsto \mathrm{Cauchy}(\theta_1 + i\theta_2).$$

Since the points $(\theta_1, \pm\theta_2)$ give rise to the same distribution, the parameter is not identifiable. Provided that $b > 0$, the subobject $\mathcal{R} \times \mathcal{R}^+$ determines a subfunctor, and the submodel with $\theta_2 \geq 0$ generates the same family. These models are equivalent to the usual group formulation by affine transformations acting componentwise on $\mathcal{R}^{\mathcal{U}}$ only if $\mathrm{cat}_{\mathcal{V}}$ contains all scales, that is, all pairs (a, b) with $b > 0$. Otherwise, the image of Θ is not a group.

To each affine transformation (a, b) of response scales, there corresponds a transformation $(a, b)^\dagger : \mathcal{R}^2 \to \mathcal{R}^2$ on the object in Θ. The first component of this transformation $\theta_1 \mapsto a + b\theta_1$ does not depend on θ_2, so the coordinate projection $(\theta_1, \theta_2) \mapsto \theta_1$ is a natural transformation. The mean or median is a natural subparameter. The same is true for the second component, so $\theta \mapsto \theta_2$ is also a natural subparameter. However the coefficient of variation $\tau = \theta_1/\theta_2$ is not a natural subparameter because the transformation $(\theta_1, \theta_2) \mapsto \theta_2/\theta_1$ is not natural for the category of location-scale transformations. The location-scale morphism (a, b) carries $\tau = \theta_1/\theta_2$ to $|b|\theta_2/(a + b\theta_1)$, which cannot be expressed as a function of (τ, a, b) alone. On the other hand, for the subcategory $(0, b)$ of componentwise nonzero scalar multiplication, the coefficient of variation is a natural subparameter. The induced transformation is $\tau \mapsto \mathrm{sign}(b)\tau$.

The analysis for other location-scale families follows the same lines. Consider, for example, the location-scale t family,

$$\mathcal{F} = \{t_\nu(\mu, \sigma^2) : \mu \in \mathcal{R},\ \sigma^2 \geq 0,\ \nu = 1, 2, \ldots\}$$

with integer degrees of freedom. Then ν is invariant under all maps in the category, and the induced transformations for (μ, σ^2) are those described above. Relative to the category of location-scale and coordinate-projection maps, μ and σ^2 are natural subparameters, but combinations such as $\tau = \sigma/\mu$ or $\mu + \sigma$ are not. However $\mu + \sigma$ and each percentile is a natural subparameter relative to the subcategory in which $b \geq 0$.

6.2. *Regression and correlation.* The issue in Exercise 10 concerns the effect on the parameter $\theta = (\alpha, \beta, \sigma)$, and on the correlation coefficient of selection of experimental units. The correlation coefficient

$$\rho = \frac{\beta\sigma_\psi}{\sqrt{(\sigma^2 + \beta^2\sigma_\psi^2)}}$$

depends on the design through the design variance σ_ψ^2. To show that the map $(\alpha, \beta, \sigma) \mapsto \rho$ is not natural, observe that if $\varphi_c = 1$ in the diagram in Section 4.5, the induced maps φ_c^*, φ_c' are both identities. In other words, for any design morphism $(\varphi_d, 1)$, the induced parameter map is the identity and $g\theta$ is invariant under these maps. But σ_ψ^2 is not invariant under selection of units, so ρ is not a natural transformation.

Any inferential statement concerning ρ must necessarily be interpreted relative to a suitable model and category for which ρ is a natural transformation of functors. One way to achieve this is to regard the pair (y, x) as a bivariate response on the units. Then $\Omega = \{0\}$ and there is only one design. Let $\mathrm{cat}_\mathcal{V}$ be the group of componentwise affine transformations, $\mathcal{R}^2 \to \mathcal{R}^2$. We write $y' = a + by$, $x' = c + dx$ in which $bd \neq 0$. The induced morphism on the parameter space is

$$\alpha' = a + b\alpha - bc\beta/d; \qquad \beta' = b\beta/d; \qquad \sigma' = |b|\sigma;$$

so Θ is a functor or group homomorphism $\mathcal{R}^3 \to \mathcal{R}^3$. In addition, the scalar multiple $\mathrm{sign}(bd)$, acting as a map $[-1, 1] \to [-1, 1]$, is also a group homomorphism. Since $\rho' = \mathrm{sign}(bd)\rho$, the correlation coefficient is a natural subparameter. Within the limitations of this subcategory, inference for ρ may be sensible.

6.3. *Splines and curve estimation.* Let the objects in cat_Ω be all bounded intervals of $\mathcal{R}$, and let the morphisms $\Omega \to \Omega'$ be all invertible affine maps. This is a subcategory of the category of injective maps on subsets of $\mathcal{R}$ because each object in cat_Ω is a bounded interval, and each map is invertible (injective and surjective). Let $\mathcal{H}_\Omega^{(k)}$ be the vector space, of dimension $2k + 4$ of cubic splines on $\Omega = (a, b)$ with k knots and $k + 1$ intervals of equal length. For each invertible affine map $\varphi : \Omega \to \Omega'$ and f a cubic spline in $\mathcal{H}_{\Omega'}^{(k)}$, the composition $\varphi^* f = f \circ \varphi$ is a vector in $\mathcal{H}_\Omega^{(k)}$. Thus $\mathcal{H}^{(k)} = \{\mathcal{H}_\Omega^{(k)}\}$ is a contravariant functor on cat_Ω.

Since the objects $\mathcal{H}_\Omega^{(k)}$ are vector spaces and the maps $\varphi^* : \mathcal{H}_{\Omega'}^{(k)} \to \mathcal{H}_\Omega^{(k)}$ are linear, $\mathcal{H}^{(k)}$ is a representation of cat_Ω, a finite-dimensional subrepresentation in the standard representation $\{\mathcal{R}^\Omega\}$. Note that the subrepresentations of this category in $\{\mathcal{R}^\Omega\}$ are very different from the subrepresentations of the affine group $GA(\mathcal{R})$ in the standard representation by real-valued functions on $\mathcal{R}$. Each finite-dimensional subrepresentation of the affine group in $\mathcal{R}^\mathcal{R}$ is a vector space of nonhomogeneous polynomials on $\mathcal{R}$.

The linear model in which the response Y satisfies

$$E(Y|x) = \mu(x); \qquad \operatorname{cov}(Y|x) = \sigma^2 I_n$$

for some $\mu \in \mathcal{H}_\Omega^{(k)}$ satisfies the conditions for a statistical model. The parameter functor is $\Theta = \mathcal{H}^{(k)} \times \mathcal{R}^+$.

The preceding category includes only invertible affine maps, and the models are not closed under restriction to subintervals of the x-scale. If it is required that the model also be closed under restriction, we may construct a statistical model as follows. Let cat_Ω be the category in which each object Ω is an interval of the real line, including infinite intervals if so desired. The morphisms are all affine maps, not necessarily invertible. Let $\mathcal{H}_\Omega^{(k)}$ be the set of functions on Ω such that each $f \in \mathcal{H}_\Omega^{(k)}$ is a cubic spline with k or fewer knots, not necessarily equally spaced. This set is not a vector space because it is not closed under addition of functions. But, for each $f \in \mathcal{H}_{\Omega'}^{(k)}$, and $\Omega \subset \Omega'$, the restriction of f to Ω is a function in the set $\mathcal{H}_\Omega^{(k)}$. Moreover, for each injective map $\varphi : \Omega \to \Omega'$, the pullback map $\varphi^* f = f \circ \varphi$ satisfies $\varphi^* \mathcal{H}_{\Omega'}^{(k)} = \mathcal{H}_\Omega^{(k)}$, so φ^* is surjective. With this modification, the conditions for a statistical model relative to the extended category are satisfied.

6.4. *Contingency-table models.* In Exercise 12, the model $AB + BC$ invites the conclusion that variables A and C are conditionally independent given B. This unqualified conclusion does not stand up to scrutiny because it is not invariant under aggregation of levels of B as the context appears to demand. The statement cannot be true at every level of aggregation unless either A is independent of (B, C) or C is independent of (A, B), and the data indicate that neither of these models fits. The stated conclusion, that A and C are conditionally independent given that B is recorded at a particular level of aggregation, evidently accords preferential status to a particular response scale. In practice, although considerable aggregation may have occurred, the nature of such aggregation is seldom arbitrary. For convenience of tabulation, aggregation tends to occur over levels that are sparse or levels that are considered to be homogeneous. Thus, despite the preferential status accorded to the aggregated levels of B, the conditional independence conclusion may be defensible. The point of this discussion is that the bald statement of conditional independence is not an automatic consequence of the fitted model.

From the viewpoint of categories and functors, to each response scale with unstructured levels it is natural to associate the category Surj of surjective maps on finite nonempty sets, or at least a suitable subcategory of certain surjective maps. Since there are three response factors, $\mathrm{cat}_\mathcal{Y} = \mathrm{Surj}^3$, the parameter Θ is a covariant functor on Surj^3, and $P\Theta$ is a class of distributions that is closed under marginal transformations. Among log-linear models, the only families having this property are those associated with the model formulas

$$A + B + C, \quad AB + C, \quad AC + B, \quad A + BC, \quad ABC,$$

in which each letter occurs exactly once. No formulation such as $AB + BC$ whose interpretation necessarily accords preferential status to the recorded levels can be a statistical model in the sense of being a covariant functor on Surj^3.

If in fact, some or all factors have ordered levels, we may restrict Surj to the subcategory $\mathrm{Surj}_{\mathrm{ord}}$ of weakly monotone surjective maps. The set of models is thereby greatly increased. In the bivariate case, the set of distributions having constant global cross-ratio [Pearson (1913); Plackett (1965); Dale (1984)] is closed under monotone marginal transformation. The sets of distributions having additive or log-additive global cross-ratios are likewise closed.

The importance of the distinction between response and explanatory factor is evident in the functor sequence (2). Suppose that A is the response and B, C are explanatory, all having unordered levels. It may then be appropriate to choose $\mathrm{cat}_{\mathcal{V}} = \mathrm{Surj}$ for surjective morphisms on the levels of A, and $\mathrm{cat}_{\Omega} = \mathcal{I}^2$ for selection of the levels of B and C. Then Θ is a covariant functor on $\mathrm{Surj} \times \mathcal{I}^{\mathrm{op}} \times \mathcal{I}^{\mathrm{op}}$. Among log-linear formulations, each factorial expression that includes $A + BC$ represents a model in the sense of (2). For two-way tables, the set of nonnegative matrices of rank $\leq k$ is a functor corresponding to a model not of the log-linear type. If the levels of A are ordered and attention is restricted to order-preserving surjective maps, further models are available in the form of cumulative logit and related formulations [McCullagh (1980)]. Apart from the rank 2 canonical correlation models, none of the association models described by Goodman (1979, 1981) for two-way tables with ordered levels is a covariant functor on either $\mathrm{Surj}^2_{\mathrm{ord}}$ or $\mathrm{Surj}_{\mathrm{ord}} \times \mathcal{I}^{\mathrm{op}}_{\mathrm{ord}}$. Goodman's association structure is not preserved under aggregation of adjacent response levels. In the absence of an extension to other levels of aggregation, all conclusions derived from fitting such models are necessarily specific to the particular level of aggregation observed, so the scope for inference is rather narrow.

6.5. *Embeddability and stochastic processes.* The point of Exercise 11 is that such a specification need not define a process that extends beyond the observed lattice. Let the observed 3×3 lattice be embedded as the central square of a larger $n \times n$ lattice. The distribution of the response on the sublattice may be computed from the n^2-dimensional normal distribution with zero mean and inverse variance matrix $(I_n - B_n)^T(I_n - B_n)$ by integrating out the $n^2 - 9$ unwanted variables. The matrices I_n and B_n are of order $n^2 \times n^2$. But the result of this computation is different for each n, and is not equal to the normal distribution with inverse covariance matrix $(I_3 - B_3)^T(I_3 - B_3)$. In other words, Exercise 11 does not identify an embeddable process. Similar issues arise in connection with edge effects in Gibbs models for finite lattice systems: for details, see Besag and Higdon [(1999), Section 2.3.1].

6.6. *Comments on selected exercises.* With the usual understanding of what is meant by natural embedding, the formulations in Exercises 1–5 do not determine a unique probability distribution or likelihood function. In Exercise 1, which asks for predictions for a future subject, the observed event is either $E \subset \mathcal{R}^n$ or the equivalent set $E' = E \times \mathcal{R} \subset \mathcal{R}^{n+1}$. The likelihood function may thus legitimately be calculated either by $P_n(E;\theta)$ or by $P_{n+1}(E';\theta)$. The absurdity of Exercise 1 lies in the fact that these expressions are not equal. To the extent that this embedding is natural, the likelihood function is not well defined. The nature of the embeddings may be different, but the same objection applies to Exercises 1–5 and 11. Although the set $P\Theta$ in Exercise 4 is the set of all i.i.d. normal models, the parameterization is unnatural and the likelihood function is not well defined.

The logical contradiction in the type III model in Exercise 5 is exactly the same as in Exercises 1 and 2, but the formulation is less flagrant and consequently more persistent. Let $\mathcal{F}_{kb}$ be the family of normal distributions on $\mathcal{R}^{kb}$ with constant variance $\sigma^2 > 0$, such that μ lies in the linear subspace III_{kb}. It is understood that the sample space morphisms include the coordinate projection map $\mathcal{R}^{kb+k} \to \mathcal{R}^{kb}$ in which one entire block is deleted or ignored. It is easy to see that the fixed-sum constraint is not satisfied by the blocks that remain. In other words, the sequence of vector spaces $\{III_{kb}\}$ is not closed under such maps, and consequently the sequence $\mathcal{F} = \{\mathcal{F}_{kb}\}$ of families is not closed either. Put more bluntly, the observation $y \subset \mathcal{R}^{kb}$ is equivalent to the incomplete observation $y' = y \times \mathcal{R}^k \subset \mathcal{R}^{k(b+1)}$ taken from a larger design in which one entire block is unobserved. But the likelihood function based on $(y, \mathcal{F}_{kb})$ is not equivalent to the likelihood function based on $(y', \mathcal{F}_{k(b+1)})$. Given this embedding, no unique likelihood can be associated with the type III model.

Gibbsian models have the property that the conditional distribution of certain subsets of the variables given the values of all other variables is also of the same type. This property makes it straightforward to construct families that are closed under conditioning. Exercise 7 illustrates a familiar problem in the use of quadratic exponential families and Gibbs models. Although the normal family is closed under coordinate projection, the implied parameterization, in which (θ_1, θ_2) does not depend on cluster size, does not yield a commutative model diagram. In other words, if $(Y_1, \ldots, Y_k)$ have the exchangeable normal distribution with parameter (θ_1, θ_2), the marginal distribution of $(Y_1, \ldots, Y_{k-1})$ is exchangeable and normal, but the parameter is not θ. By contrast, the conventional parameterization by variance and nonnegative covariance independent of cluster size, does yield a commutative model diagram. Even though the two formulations may coincide when applied to an example in which the cluster size is constant, their extensions to variable-sized clusters are different, and this difference is critical for prediction and inference. The version in which the variance and covariance are nonnegative and independent of cluster size is a statistical model. The version described in Exercise 7 is not.

The problem with the Cauchy model is that, although the family is closed under the group of real fractional linear transformations, the projection $(\theta_1, \theta_2) \mapsto \theta_1$ is not a natural transformation. Since the components are independent, it is sufficient to take $\mathcal{S} = \mathcal{R}$ and $\varphi y = (ay + b)/(cy + d)$ in the diagram below:

$$\begin{array}{ccccccc} y & & C(\theta) & \longleftarrow & \theta & \longrightarrow & g(\theta) = \theta_1 \\ \varphi\downarrow & & \varphi^{\dagger}\downarrow & & \varphi^{*}\downarrow & & \downarrow\varphi' = ? \\ \frac{ay+b}{cy+d} & & C(\psi) & \longleftarrow & \psi = \frac{a\theta+b}{c\theta+d} & \longrightarrow & g(\psi) = \psi_1 \end{array}$$

In fact, $\psi_1 = \Re((a\theta + b)/(c\theta + d))$ depends on θ_2 and is not expressible as a function of (θ_1, φ) alone. The real part is not a natural parameter relative to the group of fractional linear transformations. It appears that the only natural transformations are the identity on Θ and the constant function $\Theta \to \{0\}$.

In each particular application, it is essential to establish the category of relevant morphisms of sample spaces at the outset. Even though the observations are modelled by Cauchy distributions, it is perfectly acceptable, and frequently entirely reasonable, to choose the family of location-scale and coordinate projection maps. Unless the response is an overt ratio of outcomes, fractional linear transformations are seldom appealing in applied work, and they should be excluded on that basis. The category of morphisms on sample spaces must not be determined by the chosen family of distributions, but by the context. The mere fact that the Cauchy family happens to be closed under reciprocals does not mean that the category must include reciprocals.

The implicit category of morphisms in Exercise 9 is generated by affine transformations of temperature, together with scalar and rotational transformations of physical space. When the temperature scale is morphed from Fahrenheit to Celsius, there is a corresponding morph,

$$(\mu, \sigma^2, \lambda) \mapsto \left(5(\mu - 32)/9, (5\sigma/9)^2, \lambda\right)$$

of the parameter space. But there exists no corresponding morph for $\theta_1 = \mu/\sigma$ or $\theta_2 = \mu/\lambda$, or for any other nonsensical combination such as $\sigma + \lambda$.

7. The Box–Cox model.

7.1. *Natural subparameter.* It is sufficient in what follows to consider the simplified Box–Cox model in which the observations $y_1, \ldots, y_n$ are independent and identically distributed. The sample spaces are all finite-dimensional real vector spaces $\{\mathcal{R}^n : n \geq 0\}$. For $m \geq n$, the morphisms $\mathcal{R}^m \to \mathcal{R}^n$ include all coordinate permutation and coordinate projection maps. In addition, all scalar multiples $y \mapsto \gamma y$ for $\gamma > 0$ and all componentwise power transformations $y \mapsto \pm|y|^{\lambda}$ for $\lambda \neq 0$, are also included in $\mathcal{C}$. The inferential difficulties that arise in the Box–Cox model are due principally to the interaction between the power transformation

and the multiplicative scalar maps rather than coordinate projection maps. In the discussion of sample-space morphisms, therefore, we restrict our attention to the power transformation and scalar multiplication maps.

Since the observations are independent and identically distributed by assumption, the family of distributions on $\mathcal{R}^n$ is determined by the one-dimensional marginal family. Let $N(\beta, \sigma^2, \lambda)$ be the distribution on $\mathcal{R}$ such that when $Y \sim N(\beta, \sigma^2, \lambda)$, then $Y^\lambda \sim N(\beta, \sigma^2)$. Minor complications associated with negative values may be handled by interpreting Y^λ as $|Y|^\lambda \operatorname{sign}(Y)$. It follows that $Y^\alpha \sim N(\beta, \sigma^2, \lambda/\alpha)$ for $\alpha \neq 0$. In addition, for each scalar $\gamma > 0$, $\gamma Y \sim N(\gamma^\lambda \beta, \gamma^{2\lambda}\sigma^2, \lambda)$.

The parameter is $(\beta, \sigma^2, \lambda)$, and the parameter set is $\Theta = \mathcal{R} \times \mathcal{R}^+ \times \{\mathcal{R} \setminus 0\}$. To each scalar multiple $\gamma > 0$ the associated parameter map $\gamma^* : \Theta \to \Theta$ is the group homomorphism,

$$\gamma^* : (\beta, \sigma^2, \lambda) \mapsto (\gamma^\lambda \beta, \gamma^{2\lambda}\sigma^2, \lambda).$$

For each power transformation map $y \mapsto y^\alpha$ on the sample space, the parameter map is $\alpha^* : \lambda \mapsto \lambda/\alpha$, with β and σ^2 unaffected. The Box–Cox model is thus a parameterized statistical model according to our definition.

The fundamental difficulty, clearly evident in the discussion [Box and Cox (1982)] is that the coordinate projection $(\beta, \sigma^2, \lambda) \mapsto \beta$ is not a natural transformation of functors. The problem is evident from the diagram in which $\gamma : \mathcal{R} \to \mathcal{R}$ is a positive scalar multiple:

$$\begin{array}{ccccc} \mathcal{R} & & (\beta, \sigma^2, \lambda) & \longrightarrow & \beta \\ {\scriptstyle\gamma}\downarrow & & {\scriptstyle\gamma^*}\downarrow & & \downarrow{\scriptstyle\gamma' = ?} \\ \mathcal{R} & & (\gamma^\lambda\beta, \gamma^{2\lambda}\sigma^2, \lambda) & \longrightarrow & \gamma^\lambda\beta \end{array}$$

Commutativity requires the map γ' acting on $\beta \in \mathcal{R}$, that is, the pair (γ, β), to deliver the value $\gamma^\lambda\beta$, an impossible task. Consequently, neither β nor σ^2 nor (β, σ^2) is a natural parameter. Some examples of natural parameters include

$$\lambda, \quad \beta/\sigma, \quad \beta^{1/\lambda}, \quad (\beta, \lambda) \quad \text{and} \quad (\sigma^2, \lambda).$$

The fact that β is not a natural parameter according to our definition is an implicit statement that inference for β is meaningless in this system, that is, relative to the given category of morphisms, power transformations and positive scalar multiples.

One objection to the preceding analysis runs as follows: "The data Y were generated according to a particular distribution $N(\beta_0, \sigma_0^2, \lambda_0)$, and I want to know the true value β_0." The difficulty here is that the equivalent data $2Y$ were generated according to the distribution $N(2^{\lambda_0}\beta_0, 2^{2\lambda_0}\sigma_0^2, \lambda_0)$. But there is no way to transform the true value of β into the true value of $2^\lambda\beta$. Evidently, the phrase "the true value of β" is meaningless within the context of the group of scalar multiples acting on the sample space.

7.2. *The modified power transformation.* The problem of parameter interpretability is greatly reduced, though not entirely eliminated, by considering the modified power transformation $\mathcal{R}^n \to \mathcal{R}^n$ in which $y^{(\alpha)}$ is a vector with components

$$y_i^{(\alpha)} = y_i^{\alpha} / \dot{y}^{\alpha-1},$$

where $\dot{y}$ is the geometric mean of $\{y_1, \ldots, y_n\}$ [Hinkley and Runger (1984)]. The first algebraic obstacle is that, for $\mathrm{cat}_{\mathcal{U}} = \mathcal{U}$, the transformation $y \mapsto y^{(\alpha)}$ is not a natural transformation $\mathcal{R}^n \to \mathcal{R}^n$ on the standard representation. Although the geometric mean is invariant under permutation of units, it is not invariant under coordinate projection maps. The geometric mean of a proper subset of $\{y_1, \ldots, y_n\}$ is not the same as the geometric mean of all n values. In other words, modified Box–Cox transformation followed by coordinate projection is not the same as coordinate projection followed by modified Box–Cox transformation. Such differences may, however, be sufficiently small to be overlooked in practice.

In the following analysis, the category is reduced by restriction to a single object, a particular set of n units and the associated sample space, $\mathcal{S} = \mathcal{R}^n$. The morphisms are coordinate permutation, modified power transformation, and positive scalar multiplication, each of which commutes with the modified power transformation. The modified transformation is natural relative to the restricted category. Let $N_n(\beta, \sigma^2, \lambda)$ be the distribution on $\mathcal{R}^n$ such that if $Y \sim N_n(\beta, \sigma^2, \lambda)$ then $Y^{(\lambda)} \sim N_n(\beta, \sigma^2, 1)$, with independent normal components. This family is a functor on the restricted category: for multiplicative scalar morphisms $Y \mapsto \gamma Y$, the associated parameter morphism is

$$\gamma^* : (\beta, \sigma^2, \lambda) \mapsto (\gamma\beta, \gamma^2\sigma^2, \lambda).$$

For modified power morphisms $Y \mapsto Y^{(\alpha)}$ on the sample space, the parameter morphism is $\lambda \mapsto \lambda/\alpha$ with (β, σ^2) fixed. In both cases, the transformation acts componentwise.

Thus, relative to this restricted and artificial category, the modified Box–Cox family is a parameterized statistical model in which each component is a natural subparameter. It should come as no surprise that β/λ is not a natural subparameter.

8. Extensive response variable.

8.1. *Spatial aggregation and measure processes.* Let $\mathcal{D}$ be a fixed domain in the plane. An algebra $\mathcal{A} = \{A_1, \ldots\}$ is a collection of subsets of $\mathcal{D}$, containing the empty set and closed under finite set unions and set differences. A measure Y on $\mathcal{A}$ is an additive set function taking the value zero on the empty set and additive for disjoint sets. If A, B are elements of $\mathcal{A}$, then $A \cup B$, $A \setminus B$ and $A \cap B$ are also in $\mathcal{A}$, and

$$Y(A \cup B) + Y(A \cap B) = Y(A) + Y(B).$$

The yield of crop in a field trial is an additive set function. Typically, the process is defined on Borel sets, and observed on the algebra generated by the plots.

In the discussion that follows, it is assumed that the process Y takes values in a set $\mathcal{V}$ that includes zero and is closed under addition (an additive semigroup). In practice, this usually means that $\mathcal{V}$ is the set of nonnegative integers, the set of nonnegative reals, the set of real or complex numbers, or the Cartesian product of such sets.

Let $\mathcal{C}$ be the category in which each object $\mathcal{A}$ is a finite algebra of Borel-measurable subsets of $\mathcal{D}$, and each morphism $\varphi : \mathcal{A} \to \mathcal{A}'$ is the insertion map in which $\mathcal{A} \subset \mathcal{A}'$. The sample space is a contravariant functor on $\mathcal{C}$ that associates with each algebra $\mathcal{A}$ the set $\text{meas}_{\mathcal{V}}(\mathcal{A})$ of $\mathcal{V}$-valued additive set functions on $\mathcal{A}$, and with each insertion map $\varphi : \mathcal{A} \to \mathcal{A}'$ the map $\varphi^* : \text{meas}_{\mathcal{V}}(\mathcal{A}') \to \text{meas}_{\mathcal{V}}(\mathcal{A})$ by restriction to $\mathcal{A} \subset \mathcal{A}'$. A probability model P for the process is a contravariant functor on $\mathcal{C}$ that associates with each $\mathcal{A}$ a probability distribution $P_{\mathcal{A}}$ on $\text{meas}_{\mathcal{V}}(\mathcal{A})$ in such a way that, when $S \subset \text{meas}_{\mathcal{V}}(\mathcal{A})$ is $P_{\mathcal{A}}$-measurable, the inverse image $\varphi^{*-1}S \subset \text{meas}_{\mathcal{V}}(\mathcal{A}')$ is $P_{\mathcal{A}'}$-measurable, and $P_{\mathcal{A}}(S) = P_{\mathcal{A}'}(\varphi^{*-1}S)$. In other words, $P_{\mathcal{A}}$ is the marginal distribution of $P_{\mathcal{A}'}$,

$$\begin{array}{ccc} \mathcal{A} & \text{meas}_{\mathcal{V}}(\mathcal{A}) & P_{\mathcal{A}} \\ \varphi \downarrow & \varphi^* \uparrow & \uparrow \circ \varphi^{*-1} \\ \mathcal{A}' & \text{meas}_{\mathcal{V}}(\mathcal{A}') & P_{\mathcal{A}'} \end{array} .$$

This is a category-style statement of the Kolmogorov consistency condition that must be satisfied by the finite-dimensional distributions of a measure process on $\mathcal{D}$ [Kingman (1984)]. The term "process" refers both to the random variable Y and to the probability model P on $\mathcal{D}$ satisfying the preceding conditions.

A point process is a process for which each outcome is a countable set of points in $\mathcal{D}$. In other words, a point process is a nonnegative integer-valued random measure such that, with probability 1, Y has countable support and no multiple points. Then $Y(A)$ is the number of points in A. A Poisson process [Kingman (1993)] is a point process in which, for some weakly finite nonatomic measure μ on $\mathcal{D}$, (a) $Y(A)$ has the Poisson distribution with mean $\mu(A)$; (b) for nonoverlapping sets $A_1, A_2, \ldots$, the random variables $Y(A_1), Y(A_2), \ldots$ are independent. A process satisfying condition (2) is said to be completely random.

A measure process is called Gaussian if, for each finite collection of subsets $\{A_1, \ldots, A_n\}$, the random variable $(Y(A_1), \ldots, Y(A_n))$ is normally distributed.

8.2. *Domain morphisms.* We consider first some of the issues connected with statistical models for processes in which covariate effects do not arise. Ordinarily we require a family that is extendable to a large class of subsets of the plane. Accordingly, we begin with a category $\text{cat}_{\mathcal{D}}$ in which the objects $\mathcal{D}, \mathcal{D}', \ldots$ are some or all such subsets of the plane. The class of morphisms $\varphi : \mathcal{D} \to \mathcal{D}'$ depends on the context, but is restricted to injective maps that preserve Borel sets. In

practice, the morphisms are usually translations, rotations and scalar multiples that preserve some aspect of Euclidean structure. Every injective map $\varphi : \mathcal{D} \to \mathcal{D}'$ preserves Boolean structure in the sense that, for $A, B \subset \mathcal{D}$, $\varphi(A \cup B) = \varphi(A) \cup \varphi(B)$, and $\varphi(A \cap B) = \varphi(A) \cap \varphi(B)$. Thus, φ carries each algebra of Borel subsets of $\mathcal{D}$ into an algebra of Borel subsets of $\mathcal{D}'$.

In the diagram below, to each domain $\mathcal{D}$ there corresponds a sample space, $\mathcal{S}_{\mathcal{D}}$, of $\mathcal{V}$-valued measures on (the Borel subsets of) $\mathcal{D}$. For each point $Y \in \mathcal{S}_{\mathcal{D}'}$, a measure on $\mathcal{D}'$, $(\varphi^* Y)(A) = Y(\varphi A)$ is the composition measure on $\mathcal{D}$.

$$
\begin{array}{ccccccc}
\mathcal{D} & Y \circ \varphi & \mathcal{S}_{\mathcal{D}} & \mathcal{P}(\mathcal{S}_{\mathcal{D}}) & \xleftarrow{P_{\mathcal{D}}} & \Xi_{\mathcal{D}} \\
\downarrow \varphi & \uparrow & \uparrow \varphi^* & \uparrow \varphi^\dagger & & \uparrow \varphi' \\
\mathcal{D}' & Y & \mathcal{S}_{\mathcal{D}'} & \mathcal{P}(\mathcal{S}_{\mathcal{D}'}) & \xleftarrow{P_{\mathcal{D}'}} & \Xi_{\mathcal{D}'} \ .
\end{array} \tag{3}
$$

The map $\varphi^\dagger$ on probability models is defined by composition with the inverse image of φ^*, that is, $(\varphi^\dagger F)(S) = F(\varphi^{*-1} S)$ for $S \subset \mathcal{S}_{\mathcal{D}}$ and $F \in \mathcal{P}(\mathcal{S}_{\mathcal{D}'})$.

Typically, the process Y to be studied is modelled as a measure on $\mathcal{D}$, assigning values to each of the Borel sets. Although we refer to the sample space as the set of measures on $\mathcal{D}$ it is important to understand that Y is ordinarily not observed on the Borel sets, but on the finite subalgebra generated by the plots.

An invariant probability model is a natural transformation $\{0\} \to \mathcal{P}(\mathcal{S})$ that associates with each object $\mathcal{D}$ a process, $P_{\mathcal{D}}$ on meas($\mathcal{D}$), in such a way that, for each domain morphism $\varphi : \mathcal{D} \to \mathcal{D}'$, the sample-space map $\varphi^* : \mathcal{S}_{\mathcal{D}'} \to \mathcal{S}_{\mathcal{D}}$ satisfies $P_{\mathcal{D}} = P_{\mathcal{D}'} \circ \varphi^{*-1}$. For example, the Poisson process with constant unit intensity function is invariant under translation and rotation of domains. For many categories of domain morphisms, however, invariant processes do not exist. For example, there exists no nontrivial process that is invariant under translation and scalar multiplication. However, the *family* of Poisson processes with constant intensity function is closed under similarity transformations. Likewise, the family of stationary Gaussian processes with constant intensity and isotropic covariance density

$$\sigma^2 \exp(-\gamma |x - x'|)\, dx\, dx', \qquad \sigma^2, \gamma \geq 0,$$

is also closed under similarity transformations.

A statistical model $P : \Xi \to \mathcal{P}(\mathcal{S})$ is a natural transformation of functors on $\mathrm{cat}_{\mathcal{D}}$. For example, if $\mathrm{cat}_{\mathcal{D}}$ is the group of similarity transformations acting on $\mathcal{R}^2$, we may take $\Xi_{\mathcal{D}} = \mathcal{R}^+$ and φ' equal to the Jacobian of φ. If $P_{\mathcal{D}}\xi$ is the homogeneous Poisson process on $\mathcal{D}$ with intensity ξ, the diagram commutes.

More generally, let $\Xi_{\mathcal{D}}$ be the set of nonnegative measures on $\mathcal{D}$. For each $\varphi : \mathcal{D} \to \mathcal{D}'$ the induced map φ' on measures is given by $(\varphi' \mu)(A) = \mu(\varphi A)$ for $\mu \in \Xi_{\mathcal{D}'}$ and $A \subset \mathcal{D}$. If μ is nonatomic, so also is $\varphi' \mu$. The set of nonatomic measures is thus a subfunctor of Ξ. If $P_{\mathcal{D}'} \mu$ is the Poisson process on $\mathcal{D}'$ with nonatomic mean measure μ, the induced process $\varphi^\dagger P_{\mathcal{D}'} \mu$ on $\mathcal{D}$ is Poisson with

nonatomic mean measure $\varphi'\mu$, and the diagram commutes. An arbitrary measure μ with atoms gives rise to a process in which multiple events occur with nonzero probability at the same point. Although $P_{\mathcal{D}}\mu$ is a Poisson process, it is not a point process according to the usual definition [Kingman (1993)]. Further submodels exist if $\mathrm{cat}_{\mathcal{D}}$ is restricted to continuous maps.

8.3. *Covariate effects in spatial processes.* In a field trial over a domain $\mathcal{D}$, the design is a map $\psi : \mathcal{D} \to \Omega$, associating with each point u in the domain a point ψu in Ω. In practice, since observations are made on plots, not points, ψ is constant on each plot. Each morphism $\psi \to \psi'$ of designs is a pair of injective maps (φ_d, φ_c) such that $\psi'\varphi_d = \varphi_c\psi$.

To construct a model for the effect of covariates on a measure-valued process, it is essential to begin with a suitable baseline family, or uniformity model, for the process when covariate effects are absent. Such a family is necessarily closed under the category of morphisms of the domains. In the diagram shown below, the parameter space is a contravariant functor associating with each design $\psi : \mathcal{D} \to \Omega$ a parameter set $\Xi_{\mathcal{D}} \times \Theta_{\Omega}$, and with each morphism $(\varphi_d, \varphi_c) : \psi \to \psi'$ a map $(\varphi'_d, \varphi'_c) : \Xi_{\mathcal{D}'} \times \Theta_{\Omega'} \to \Xi_{\mathcal{D}} \times \Theta_{\Omega}$. A statistical model is a natural transformation of functors on the design. The baseline uniformity model is obtained by choosing $\Theta = \{0\}$.

(4)
$$\begin{array}{ccccccc}
 & \text{Design} & & & \text{Sample space} & \text{Model} & \\
\mathcal{D} & \xrightarrow{\psi} & \Omega & & \mathcal{S}_{\mathcal{D}} = \mathrm{meas}(\mathcal{D}) & \mathcal{P}(\mathcal{S}_{\mathcal{D}}) \xleftarrow{P_{\psi}} & \Xi_{\mathcal{D}} \times \Theta_{\Omega} \\
\varphi_d \downarrow & & \varphi_c \downarrow & & \varphi_d^* \uparrow & \varphi_d^{\dagger} \uparrow & \varphi'_d \uparrow \quad \uparrow \varphi'_c \\
\mathcal{D}' & \xrightarrow{\psi'} & \Omega' & & \mathcal{S}_{\mathcal{D}'} = \mathrm{meas}(\mathcal{D}') & \mathcal{P}(\mathcal{S}_{\mathcal{D}'}) \xleftarrow{P_{\psi'}} & \Xi_{\mathcal{D}'} \times \Theta_{\Omega'}
\end{array}$$

In order to see what this diagram means in practice, let $\varphi : \mathcal{D} \to \mathcal{D}'$ be a similarity transformation with scalar multiple $\sqrt{2}$. Then the area of the image $\varphi_d\mathcal{D} \subset \mathcal{D}'$ is twice the area of $\mathcal{D}$, and the mean intensity of $Y\varphi_d = \varphi_d^* Y$ on $\mathcal{D}$ is twice the intensity of Y on $\mathcal{D}'$. Thus, in the baseline model, we may choose $\Xi_{\mathcal{D}} = \mathcal{R}$ and $\varphi'_d = |\partial\varphi_d/\partial u| = 2$ as a scalar multiple.

Let φ_c be the identity on Ω, so that φ'_c is the identity on Θ_{Ω}. Suppose that the design ψ' on $\mathcal{D}'$ uses two varieties and that the yield for variety 2 exceeds the yield for variety 1 by one ton per acre. As observed on $\mathcal{D}$, the yield per unit area for each variety is doubled, and the difference is also doubled. For the two varieties, the processes on $\mathcal{D}'$ are $P_{\psi'}(\xi, \theta_1)$ and $P_{\psi'}(\xi, \theta_2)$ with intensity functions $\lambda(\xi, \theta_1), \lambda(\xi, \theta_2)$. On $\mathcal{D}$ we obtain $P_{\psi}(2\xi, \theta_1)$ and $P_{\psi}(2\xi, \theta_2)$ with intensity functions $\lambda(2\xi, \theta_1), \lambda(2\xi, \theta_2)$. Commutativity requires a model parameterized in such a way that the difference between the intensity functions on $\mathcal{D}$ is twice the difference between the intensity functions on $\mathcal{D}'$. So far as

expected yields are concerned, the variety effects must be multiplicative, not additive.

Definition (4) is sufficiently flexible to accommodate interference in the form of carry-over effects such as fertilizer seepage, or prolonged residual effects in cross-over trials. To do so, however, the covariate space must be extended so that ψ associates with each unit the treatment applied to that unit and the treatments applied to the neighbouring units. The morphisms $\psi \to \psi'$ preserve designs in this extended sense, and P_ψ may incorporate specific carry-over effects.

A model for yield in agricultural field trials that is closed under spatial aggregation has obvious conceptual advantages. However, there are competing issues to be considered, including model flexibility and ease of fitting. First, uniformity trials seem to indicate that most agricultural processes are not spatially stationary, so a family of nonstationary baseline processes may be required. It is not at all clear what sorts of nonstationarity should be accommodated in such a baseline model. Partly for these reasons, the model presented by Besag and Higdon (1999) uses a baseline family that is not closed under spatial aggregation. For further commentary on this point, see the discussion by Best, Ickstadt and Wolpert (1999), Harville and Zimmerman (1999) and the reply by the authors (page 740). The Besag–Higdon formulation is thus not a statistical model for an extensive variable in the sense of the present paper. Algebra does not easily deal with approximations or inequalities, so it is hard to say in what sense the Besag–Higdon formulation might be approximately closed under spatial aggregation, rotation or other domain morphisms.

8.4. *Conformal models.* A conformal model is a natural transformation of functors in which $\mathrm{cat}_{\mathcal{D}}$ is the category of conformal maps acting on domains in the plane. The category of conformal maps is a natural choice for real spatial processes because it preserves local Euclidean structure.

A completely random process on $\mathcal{D}$ is characterized by its one-dimensional distributions. Because it represents an additive set function, such a process on Borel sets is necessarily infinitely divisible, so that the log characteristic function of the one-dimensional distributions has the form

$$\log E(e^{itY(A)}) = m(A) \times \log \phi(t)$$

for some nonnegative measure m on $\mathcal{D}$. A family of completely random processes is determined by a particular choice of characteristic function ϕ together with a suitable family of intensity measures closed under the required category of domain morphisms. For example, if the category includes scalar multiples, the family of measures must be closed under scalar multiplication. Likewise, under conformal mapping, $\varphi : \mathcal{D} \to \mathcal{D}'$, the process $\varphi^* Y = Y\varphi$ on $\mathcal{D}$ has log characteristic function

$$\log E(e^{it(\varphi^* Y)(A)}) = \log E(e^{itY(\varphi A)}) = m\big(\varphi A\big) \times \log \phi(t)$$

for $A \subset \mathcal{D}$. The measure m on $\mathcal{D}'$ is thus transformed to the measure $m\varphi$ on $\mathcal{D}$ by functional composition. If m has a density dm with respect to Lebesgue measure on $\mathcal{D}'$, the induced measure has a density $\varphi' dm$ with respect to Lebesgue measure on $\mathcal{D}$ given by

$$\varphi' dm(x) = dm(\varphi x) \times |\partial\varphi(x)/\partial x|.$$

As will be shown below, the family of intensity measures whose log densities are harmonic functions is closed under conformal mapping. Thus, an infinitely divisible distribution together with such a family of mean measures determines a conformal family of completely random processes.

Some further examples of conformal models can be described in terms of cumulant measures. The first cumulant measure of the process Y on $\mathcal{D}$ is assumed to have a positive density such that

$$E\big(Y(dx)\big) = \exp\big(\lambda_1(x)\big)\,dx.$$

Thus

$$E\big(Y(A)\big) = \int_A \exp\big(\lambda_1(x)\big)\,dx$$

for $A \subset \mathcal{D}$. The variance measure on $\mathcal{D}$ and covariance measure on $\mathcal{D} \times \mathcal{D}$ are also assumed to have densities expressible as follows:

$$\operatorname{var}\big(Y(dx)\big) = \exp\big(\lambda_2(x)\big)\,dx,$$
$$\operatorname{cov}\big(Y(dx), Y(dx')\big) = \exp\big(\lambda_{11}(x, x')\big)\,dx\,dx',$$

for $x \neq x'$. Thus,

$$\operatorname{cov}\big(Y(A), Y(B)\big) = \int_{A\cap B} \exp\big(\lambda_2(x)\big)\,dx + \int_{A\times B} \exp\big(\lambda_{11}(x, x')\big)\,dx\,dx'.$$

Let Y be a process on $\mathcal{D}'$ with logarithmic intensity functions $\lambda_1', \lambda_2', \lambda_{11}'$. Let $\varphi : \mathcal{D} \to \mathcal{D}'$ be a conformal transformation, and let $Y\varphi$ be the composition process on $\mathcal{D}$. Then the first and second cumulant measures of the transformed process at x in $\mathcal{D}$ have logarithmic intensities such that

$$\lambda_1(x) = \lambda_1'(\varphi x) + \log J(x),$$
$$\lambda_2(x) = \lambda_2'(\varphi x) + \log J(x),$$
$$\lambda_{11}(x, x') = \lambda_{11}'(\varphi x, \varphi x') + \log J(x) + \log J(x'),$$

where $J(x) = |\partial\varphi(x)/\partial x|$ is the Jacobian of the transformation. If $\varphi : \mathcal{D} \to \mathcal{D}'$ is conformal, $\log J : \mathcal{D} \to \mathcal{R}$ is harmonic. Note also that $\lambda_2 - \lambda_1$ is an absolute invariant.

Let $\mathcal{H}_{\mathcal{D}}$ be the vector space of harmonic functions on $\mathcal{D}$. In particular, if $\varphi : \mathcal{D} \to \mathcal{D}'$ is conformal, $\log J$ is harmonic and thus a vector in $\mathcal{H}_{\mathcal{D}}$. If the intensity functions λ_1', λ_2' are also in $\mathcal{H}_{\mathcal{D}}$, it is evident that $\lambda_1 = \lambda_1' \circ \varphi$ and

$\lambda_2 = \lambda_2' \circ \varphi$ are in $\mathcal{H}_{\mathcal{D}'}$. Further, if $\lambda_{11} \in \mathcal{H}_{\mathcal{D}}^{\otimes 2}$, the corresponding transformed parameter λ_{11}' is a vector in $\mathcal{H}_{\mathcal{D}'}^{\otimes 2}$. For example, $\log|x - x'|$ is a vector in $\mathcal{H}_{\mathcal{D}'}^{\otimes 2}$ (excluding the diagonal set $x = x'$ in $\mathcal{D} \times \mathcal{D}$), and $\log|\varphi(x) - \varphi(x')|$ is a vector in $\mathcal{H}_{\mathcal{D}}^{\otimes 2}$, also excluding the diagonal.

The baseline processes considered here are those conformal processes for which the log-transformed cumulant densities are harmonic functions. The subfamily of completely random processes for which $\lambda_{11} = -\infty$, and thus $\lambda_2(x) = \lambda_1(x) + \text{const}$, is of particular interest in the analysis of field experiments.

8.5. *A conformal model with covariates.* We consider a model in which, conditionally on the real-valued function λ, the mean and variance of the process Y at $x \in \mathcal{D}$ are modelled as follows:

$$\begin{aligned} \mu(dx) &= E\big(Y(dx)\big) = \exp\big(\lambda(x) + (\psi^*\theta)(x)\big)\,dx, \\ \operatorname{var}\big(Y(dx)\big) &= \sigma^2 \mu(dx), \\ \operatorname{cov}\big(Y(dx), Y(dx')\big) &= 0, \end{aligned} \tag{5}$$

for some volatility parameter $\sigma^2 > 0$. Following Besag and Higdon (1999), we may interpret the term $\lambda(x)$ as the fertility intensity at x. Although the notation $(\psi^*\theta)(x)$ may appear unfamiliar, the effects of variety, treatment and other covariates are modelled in the usual manner, though their effects on the mean are multiplicative rather than additive. The multiplicative form in (5) is necessary to ensure that the diagram (4) is commutative for the category of conformal transformations. It is possible but not necessary that Y should be a Gaussian process.

The key issue concerns the assumptions to be made about fertility effects. It is generally understood that, for field experiments, the assumption of uniform fertility is unduly optimistic. Neighboring plots tend to have similar fertilities, but there can be substantial fertility gradients over the domain of interest. At the other extreme, if no assumptions are made about fertility patterns, that is, if λ is regarded as an arbitrary continuous function in $\mathcal{R}^{\mathcal{D}}$, the variety effects are not identifiable. One intermediate option is to assume that fertility variation can be modelled as a quadratic function or a rational function of suitable low order. In this section, we explore a third option, namely assuming that the fertility intensity λ on $\mathcal{D}$ lies in the vector space of harmonic functions $\mathcal{H}_{\mathcal{D}}$. This assumption means that the fertility at $x \in \mathcal{D}$ is equal to the average fertility on each disk $D(x, r) \subset \mathcal{D}$ with center x and radius $r > 0$. The model allows considerable flexibility for fertility gradients, but it does imply that the fertility function cannot have a local maximum or minimum in the interior of $\mathcal{D}$. As with all such assumptions born of algebraic convenience, empirical work is needed to see if the assumption is reasonable. The available numerical evidence is limited, but very positive.

A harmonic function has the property that the Laplacian at $x = (x_1, x_2)$ vanishes:

$$\nabla\lambda(x) = \frac{\partial^2\lambda}{\partial x_1^2} + \frac{\partial^2\lambda}{\partial x_2^2} = 0$$

for each $x \in \mathcal{D}$. Formal application of the Laplacian operator to our model gives

$$\nabla \log \frac{d\mu}{dx} = \nabla(\psi^*\theta)(x)$$

since the Laplacian is a linear operator. Provided that the the fertility function is sufficiently smooth to be approximated by a vector in $\mathcal{H}_\mathcal{D}$, no prior distribution is required. Even though the vector space $\mathcal{H}_\mathcal{D}$ of fertility intensities is infinite dimensional, certain treatment and variety contrasts are identifiable. Identifiability requires that $\nabla(\psi^*\theta)$ not be identically zero. Numerical calculation shows that, in many cases, all variety contrasts remain identifiable.

8.6. *Lattice approximation.* We assume here that the response measure Y is observed on a regular $m \times n$ grid of rectangular plots, $\Delta_x \times \Delta_y$, sufficiently small that the Laplacian can be approximated at each internal plot. Let the rows be indexed by $1 \le i \le m$, and the columns by $1 \le j \le m$. On the assumption that the plot sizes are sufficiently small that λ is effectively constant on plots, the multiplicative model in the preceding section gives

$$\mu(A) = E\big(Y(A)\big) = \exp\big(\lambda_A + (\psi^*\theta)_A\big) \times |A|,$$
$$\operatorname{var}\big(Y(A)\big) = \sigma^2\mu(A)$$

in which $(\psi^*\theta)_A$ is the effect of the treatment or variety associated with plot A. Reverting to more classical notation in which the plots are indexed by $(i, j) \in m \times n$, we have

$$\log \mu_{ij} = \lambda_{ij} + (X\theta)_{ij} + \log|A_{ij}|.$$

In the terminology of generalized linear models [McCullagh and Nelder (1989)], the offset is $\log|A|$, the link function is log, and the variance function is the mean. Provided that the plot areas are constant, the term $\log|A|$ can be absorbed into the harmonic λ.

The simplest lattice version of the Laplacian is proportional to

$$(\nabla Y)_{ij} = (Y_{i-1,j} - 2Y_{ij} + Y_{i+1,j})/\Delta_y^2 + (Y_{i,j-1} - 2Y_{ij} + Y_{i,j+1})/\Delta_x^2$$

for $2 \le i \le m-1$ and $2 \le j \le n-1$. It is absolutely essential to take account of the geometry of the plots by including the factors Δ_x and Δ_y in the definition of the Laplacian. If the plots are square, the Laplacian at (i, j) may be approximated by the sum of the four neighbors minus four times the value at (i, j).

The kernel of the linear transformation $\nabla : \mathcal{R}^{mn} \to \mathcal{R}^{(m-2)(n-2)}$ is the subspace $\mathcal{H}$ of lattice harmonic functions, of dimension equal to the number of boundary plots. The linear projection $R^{mn} \to \mathcal{R}^{mn}$,

$$H = I - \nabla'(\nabla\nabla')^{-1}\nabla$$

has the property that the image of H is the kernel of ∇, which is the subspace $\mathcal{H}$. The model formula is thus $\ker\nabla + X$, or $H + X$, with log link and variance proportional to the mean. An approximate estimate of the variety contrasts may be obtained by solving

$$(X'WX)\tilde{\beta} = X'W(\log Y),$$

where $W = \nabla'(\nabla\nabla')^{-1}\nabla$. Since ∇ annihilates constants and linear functions, the matrix $X'WX$ is not invertible. At best, only contrasts are estimable.

In principle, the nonlinear model can be fitted by standard software such as GLIM or the `glm()` function in Splus. In practice, since H is $mn \times mn$ of rank $h = 2(m + n - 2)$, it is necessary to eliminate the redundant columns. This can be done by using an equivalent full-rank matrix $H' = HJ$ in which J is $mn \times h$ of full rank. A matrix of independent uniform random numbers is adequate for this purpose.

The data from the wheat uniformity trial reported by Mercer and Hall (1911), and reproduced in Andrews and Herzberg (1985), were reanalysed in the light of the preceding discussion. The design is a grid of 20×25 nearly square plots, each $1/500$ of an acre in area. Each plot contains 11 drill rows 10.82 feet long. From the given total area, we infer that the drill rows are approximately nine inches apart and that the plot size is 10.82×8.05 in feet. After allowing for multiplicative row and column effects, it is found that the mean square due to harmonic variations is 0.304 on 82 degrees of freedom, whereas the residual mean square is 0.107 on 374 degrees of freedom. At the expense of the boundary plots, a 25% reduction in the variance of treatment contrasts is achieved by the elimination of harmonic trends. The observed variance ratio of 2.85 provides strong evidence for the presence of at least two variance components in these data, and the ratio might well have been considerably larger had the region not been selected on the basis of its uniform appearance. Numerically, it matters little whether the model is additive or multiplicative.

The real and imaginary parts of $(x_1 + ix_2)^k$ are homogeneous polynomials of degree k, both of which are harmonic functions. The set $\mathcal{H}_k \subset \mathcal{H}$ of harmonics of degree k or less on $\mathcal{D} = \mathcal{R}^2$ is a subspace of dimension $2k + 1$ closed under similarity transformations, that is, planar rotation, translation and scalar multiplication. That is to say, $\mathcal{H}_k$ and $\mathcal{H}/\mathcal{H}_k$ are representations of this group, and there is a corresponding invariant decomposition of the harmonic sum of squares. In the Mercer–Hall data, low-order harmonics tend to dominate, but this dominance is much less pronounced than I had anticipated. After allowing for

row and column effects, the seven unaliased harmonics of degree five or less account for only 17% of the total harmonic sum of squares. The F-ratio of 2.19 on 7, 75 degrees of freedom gives a p-value of 4.5%. It should be borne in mind, however, that this one-acre region of the field was selected by visual inspection on account of its evident uniformity. Had the harvested region been predetermined, or otherwise selected at random, large-amplitude low-order harmonic trends might well have been more pronounced.

Curiously, Besag [(1974), Table 16] obtains exactly the same residual mean square using a second-order Gibbs model with a linear trend adjustment. This coincidence can be explained in part by noting that the average of the four first-order coefficients in the fitted Gibbs model is almost exactly 0.25, while the second-order coefficients are very small. These values suggest nonstationarity [Besag and Kooperberg (1995)], and are consistent with the hypothesis that the process is white noise plus a (random) harmonic function. An intrinsic Gaussian process with a harmonic generalized covariance density of the form $-\log|x - x'|$ for $x \neq x'$ is a possibility. It should be noted that the orientation of the plots is not entirely clear from the description given. If the plots were in fact 8.05×10.82, so that the total area is 201.25×216.4 rather than 161.0×270.5, the preceding analysis would give a residual mean square of 0.114. The Gibbs model, which ignores such physical spatial information, is unaffected.

8.7. *A parting proposal.* Many traditional field experiments use block designs, in which the model matrix X includes block effects in addition to treatment and variety effects. Certain directional block effects and discontinuities may be attributable to drainage, ploughing, harvesting or other field management practices, and specific allowance for these may be required. Other artificial block effects may also be included to compensate for natural spatial variation. The fertility patterns modelled by block effects are constant on blocks but discontinuous at block boundaries. The fertility patterns modelled by harmonic functions are smooth trends having the characteristic mean-value property. In the absence of a compelling argument for fertility discontinuities beyond those described above, it should not be necessary to include artificial block effects in the model in addition to harmonic variations. This line of argument leads to the modest, but challenging, proposal to augment the theory of block designs with a theory of efficient harmonic designs for field experiments. It is inefficient, for example, to assign the same treatment to neighboring plots.

APPENDIX

Glossary on categories. *Category theory asks of every type of mathematical object*: *"What are the morphisms?"*; *it suggests that these morphisms should be described at the same time as the objects* [Mac Lane (1998)]. The morphisms

determine the structure of the objects, and, in a sense, they are more important than the objects.

A category $\mathcal{C}$ is determined by the objects $\Omega, \Omega', \ldots$ it contains, and the morphisms, maps or arrows, $\varphi : \Omega \to \Omega'$ between pairs of objects. The set of morphisms $\Omega \to \Omega'$ in $\mathcal{C}$ is denoted by $\mathcal{C}(\Omega, \Omega')$ or $\hom_{\mathcal{C}}(\Omega, \Omega')$. Two conditions are required in order that a given collection of objects and arrows should constitute a category. First, for each object Ω, the identity arrow $1 : \Omega \to \Omega$ is a morphism in $\mathcal{C}$. Second, to each pair of arrows $\varphi : \Omega \to \Omega'$, $\psi : \Omega' \to \Omega''$ in $\mathcal{C}$ such that the domain of ψ is equal to the codomain of φ, the composition arrow $\psi\varphi : \Omega \to \Omega''$ is a morphism in $\mathcal{C}(\Omega, \Omega'')$. A category is thus a collection of morphisms that contains each identity and is closed under composition of composable arrows.

Each set $\mathcal{C}(\Omega, \Omega)$ contains the identity and is closed under composition, that is, $\mathcal{C}(\Omega, \Omega)$ is a monoid (semigroup with identity). A category with all arrows invertible is a groupoid; a category with all arrows identities is a set, the set of objects; a category with exactly one object is a monoid; a category with one object and invertible arrows is a group.

Product category. Let $\mathcal{C}, \mathcal{K}$ be two categories. The objects in the product category $\mathcal{C} \times \mathcal{K}$ are the Cartesian products $\Omega \times \Gamma$ in which $\Omega \in \mathcal{C}$ and $\Gamma \in \mathcal{K}$. The morphisms $\Omega \times \Gamma \to \Omega' \times \Gamma'$ are all ordered pairs $(\varphi : \Omega \to \Omega', \psi : \Gamma \to \Gamma')$ of morphisms acting componentwise: $(\varphi, \psi)(i, j) = (\varphi i, \psi j)$.

Functor. A functor $T : \mathcal{C} \to \mathcal{K}$ is a morphism of categories, preserving category structure (identity and composition). Since the functor acts on objects as well as arrows, we sometimes write $T = (T, *)$ in which T is the object map and $*$ is the arrow map. The object map carries Ω to the object $T\Omega = T_\Omega$ in $\mathcal{K}$. In a covariant functor, the arrow map carries the arrow $\varphi : \Omega \to \Omega'$ to the arrow $\varphi^* : T_\Omega \to T_{\Omega'}$ in $\mathcal{K}$. Two conditions are required. First, the identity $1 : \Omega \to \Omega$ in $\mathcal{C}$ is carried to the identity $1 : T_\Omega \to T_\Omega$ in $\mathcal{K}$. Second, for each $\psi : \Omega' \to \Omega''$, the composition arrow $\psi\varphi : \Omega \to \Omega''$ is carried to the composition $\psi^*\varphi^* : T_\Omega \to T_{\Omega''}$ in $\mathcal{K}$.

In a contravariant functor, the arrow map carries $\varphi : \Omega \to \Omega'$ to $\varphi^* : T_{\Omega'} \to T_\Omega$, reversing the sense of the arrow. Apart from this reversal, the functor preserves category structure. First, each identity $1 : \Omega \to \Omega$ in $\mathcal{C}$ is carried to an identity $1 : T_\Omega \to T_\Omega$ in $\mathcal{K}$. Second, the composition arrow $\psi\varphi : \Omega \to \Omega''$ is carried to the composition $\varphi^*\psi^* : T_{\Omega''} \to T_\Omega$, reversing the order of composition.

The example in Section 6.1 shows that the set of arrows in the image of T need not be closed under composition; that is, $T\mathcal{C}$ need not be a subcategory of $\mathcal{K}$.

A group homomorphism is a functor in which $\mathcal{C}$ and $\mathcal{K}$ are groups. On account of isomorphism, the distinction between covariant and contravariant functors does not arise. The image $T\mathcal{C} \subset \mathcal{K}$ is a subgroup in $\mathcal{K}$.

Isomorphism. An isomorphism of categories is a covariant functor $T:\mathcal{C}\to\mathcal{K}$ that is a bijection on both objects and arrows. A contravariant functor $T:\mathcal{C}\to\mathcal{K}$ that is a bijection on both objects and arrows is an opposite-isomorphism $\mathcal{C}\to\mathcal{C}^{\mathrm{op}}\cong\mathcal{K}$.

Representation. Let $\mathcal{K}$ be the category of linear transformations on vector spaces. A functor $T:\mathcal{C}\to\mathcal{K}$ is said to be a representation of $\mathcal{C}$ by linear transformations, or simply a representation of $\mathcal{C}$.

The standard contravariant representation of $\mathcal{C}$ associates with each object Ω the vector space $\mathcal{R}^{\Omega}$, and with each morphism $\varphi:\Omega\to\Omega'$ the linear transformation $\varphi^*:\mathcal{R}^{\Omega'}\to\mathcal{R}^{\Omega}$ by functional composition: $\varphi^* f = f\circ\varphi$.

Dual representation. Let $\mathcal{K}$ be the category of linear transformations on vector spaces. The dual functor associates with each vector space $\mathcal{V}$ in $\mathcal{K}$ the dual vector space $\mathcal{V}'$ of linear functionals on $\mathcal{V}$. To each morphism, or linear transformation, $A:\mathcal{V}\to\mathcal{W}$ in $\mathcal{K}$ there corresponds a dual linear transformation $A':\mathcal{W}'\to\mathcal{V}'$, also called the vector-space adjoint transformation. This operation reverses the direction of each arrow, so the dual is a contravariant functor $\mathcal{K}\to\mathcal{K}$. In matrix notation, A' is the transpose of A.

The dual of a covariant representation $T:\mathcal{C}\to\mathcal{K}$ is the composition $\mathcal{C}\xrightarrow{T}\mathcal{K}\xrightarrow{'}\mathcal{K}$ associating with each object Ω the dual vector space, T'_{Ω}, and with each morphism $\varphi:\Omega\to\Omega'$ the dual linear transformation $(T\varphi)':T'_{\Omega'}\to T'_{\Omega}$.

Opposites. To each category $\mathcal{C}$ we make correspond an abstract category $\mathcal{C}^{\mathrm{op}}$ by retaining the objects, and reversing all arrows. The functor $\mathcal{C}\to\mathcal{C}^{\mathrm{op}}$ is contravariant, and the order of composition is reversed. Each category of invertible maps (groupoid) is isomorphic with its opposite. The category obtained by replacing each object Ω of $\mathcal{C}$ by the power set 2^{Ω}, and each map $\varphi:\Omega\to\Omega'$ by its inverse image $\varphi^{-1}:2^{\Omega'}\to 2^{\Omega}$, is isomorphic with $\mathcal{C}^{\mathrm{op}}$. Other concrete instances of opposites include the standard representation and the dual category of linear transformations.

Natural transformation. Let $\mathcal{C},\mathcal{K}$ be two categories, and let S,T be two covariant functors $\mathcal{C}\to\mathcal{K}$. A natural transformation $g:S\to T$ associates with each object Ω in $\mathcal{C}$ an arrow $g_{\Omega}:S_{\Omega}\to T_{\Omega}$ in such a way that the arrow diagram commutes.

$$\begin{array}{ccc} \Omega & S_{\Omega} & \xrightarrow{g_{\Omega}} \quad T_{\Omega} \\ \varphi\big\downarrow & \varphi_s^*\big\downarrow & \qquad\quad \varphi_t^*\big\downarrow \\ \Omega' & S_{\Omega'} & \xrightarrow{g_{\Omega'}} \quad T_{\Omega'} \end{array}$$

In other words, for each $\varphi : \Omega \to \Omega'$ in $\mathcal{C}$ the image arrows φ_s^* and φ_t^* satisfy $\varphi_t^* g_\Omega = g_{\Omega'} \varphi_s^* : S_\Omega \to T_{\Omega'}$.

If S and T are contravariant functors, the direction of the arrows φ_s^* and φ_t^* is reversed. The commutativity condition then becomes $\varphi_t^* g_{\Omega'} = g_\Omega \varphi_s^* : S_{\Omega'} \to T_\Omega$.

Acknowledgments. I am grateful to Yali Amit, Steen Andersson, Jim Berger, Peter Bickel, David Cox, Art Dempster, Gary Glonek, Inge Helland, Saunders Mac Lane, Nick Polson, Nancy Reid, Michael Stein, Stephen Stigler, Burt Totaro, Ernst Wit and the referees for helpful discussion and comments on various points.

REFERENCES

ALDOUS, D. (1981). Representations for partially exchangeable arrays of random variables. *J. Multivariate Analysis* **11** 581–598.

ANDREWS, D. F. and HERZBERG, A. (1985). *Data*. Springer, New York.

BARNDORFF-NIELSEN, O. E. and COX, D. R. (1994). *Inference and Asymptotics*. Chapman and Hall, London.

BARTLETT, M. S. (1978). Nearest neighbour models in the analysis of field experiments (with discussion). *J. Roy. Statist. Soc. Ser. B* **40** 147–174.

BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, New York.

BERNARDO, J. M. and SMITH, A. F. M. (1994). *Bayesian Theory*. Wiley, New York.

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **36** 192–236.

BESAG, J. and HIGDON, D. (1999). Bayesian analysis of agricultural field experiments (with discussion). *J. Roy. Statist. Soc. Ser. B* **61** 691–746.

BESAG, J. and KOOPERBERG, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82** 733–746.

BEST, N. G., ICKSTADT, K. and WOLPERT, R. L. (1999). Contribution to the discussion of Besag (1999). *J. Roy. Statist. Soc. Ser. B* **61** 728–729.

BICKEL, P. and DOKSUM, K. A. (1981). An analysis of transformations revisited. *J. Amer. Statist. Assoc.* **76** 296–311.

BILLINGSLEY, P. (1986). *Probability and Measure*, 2nd ed. Wiley, New York.

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations (with discussion). *J. Roy. Statist. Soc. Ser. B* **26** 211–252.

BOX, G. E. P. and COX, D. R. (1982). An analysis of transformations revisited, rebutted. *J. Amer. Statist. Assoc.* **77** 209–210.

COX, D. R. (1958). *Planning of Experiments*. Wiley, New York.

COX, D. R. (1986). Comment on Holland (1986). *J. Amer. Statist. Assoc.* **81** 963–964.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

COX, D. R. and REID, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **49** 1–39.

COX, D. R. and SNELL, E. J. (1981). *Applied Statistics*. Chapman and Hall, London.

COX, D. R. and WERMUTH, N. (1996). *Multivariate Dependencies*. Chapman and Hall, London.

DALE, J. R. (1984). Local versus global association for bivariate ordered responses. *Biometrika* **71** 507–514.

DE FINETTI, B. (1975). *Theory of Probability* **2**. Wiley, New York.

GELMAN, A., CARLIN, J. B., STERN, H. and RUBIN, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall, London.

GOODMAN, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *J. Amer. Statist. Assoc.* **74** 537–552.

GOODMAN, L. A. (1981). Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *J. Amer. Statist. Assoc.* **76** 320–334.

HAMADA, M. and WU, C. F. J. (1992). Analysis of designed experiments with complex aliasing. *J. Qual. Technology* **24** 130–137.

HARVILLE, D. A. and ZIMMERMANN, D. L. (1999). Contribution to the discussion of Besag (1999). *J. Roy. Statist. Soc. Ser. B* **61** 733–734.

HELLAND, I. S. (1999a). Quantum mechanics from symmetry and statistical modelling. *Internat. J. Theoret. Phys.* **38** 1851–1881.

HELLAND, I. S. (1999b). Quantum theory from symmetries in a general statistical parameter space. Technical report, Dept. Mathematics, Univ. Oslo.

HINKLEY, D. V. and RUNGER, G. (1984). The analysis of transformed data (with discussion). *J. Amer. Statist. Assoc.* **79** 302–320.

HOLLAND, P. (1986). Statistics and causal inference (with discussion). *J. Amer. Statist. Assoc.* **81** 945–970.

HORA, R. B. and BUEHLER, R. J. (1966). Fiducial theory and invariant estimation. *Ann. Math. Statist.* **37** 643–656.

KINGMAN, J. F. C. (1984). Present position and potential developments: Some personal views. Probability and random processes. *J. Roy. Statist. Soc. Ser. A* **147** 233–244.

KINGMAN, J. F. C. (1993). *Poisson Processes*. Oxford Univ. Press.

LAURITZEN, S. (1988). *Extremal Families and Systems of Sufficient Statistics*. *Lecture Notes in Statist.* **49**. Springer, New York.

LEHMANN, E. L. (1983). *Theory of Point Estimation*. Wiley, New York.

LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. Springer, New York.

LITTELL, R., FREUND, R. J. and SPECTOR, P. C. (1991). *SAS System for Linear Models*, 3rd ed. SAS Institute, Cary, NC.

MAC LANE, S. (1998). *Categories for the Working Mathematician*, 2nd ed. Springer, New York.

MCCULLAGH, P. (1980). Regression models for ordinal data (with discussion). *J. Roy. Statist. Soc. Ser. B* **42** 109–142.

MCCULLAGH, P. (1992). Conditional inference and Cauchy models. *Biometrika* **79** 247–259.

MCCULLAGH, P. (1996). Möbius transformation and Cauchy parameter estimation. *Ann. Statist.* **24** 787–808.

MCCULLAGH, P. (1999). Quotient spaces and statistical models. *Canad. J. Statist.* **27** 447–456.

MCCULLAGH, P. (2000). Invariance and factorial models (with discussion). *J. Roy. Statist. Soc. Ser. B* **62** 209–256.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

MCCULLAGH. P. and WIT, E. (2000). Natural transformation and the Bayes map. Technical report.

MERCER, W. B. and HALL, A. D. (1911). The experimental error of field trials. *J. Agric. Research* **50** 331–357.

NELDER, J. A. (1977). A re-formulation of linear models (with discussion). *J. Roy. Statist. Soc. Ser. A* **140** 48–77.

PEARSON, K. (1913). Note on the surface of constant association. *Biometrika* **9** 534–537.

PLACKETT, R. L. (1965). A class of bivariate distributions. *J. Amer. Statist. Assoc.* **60** 516–522.

RUBIN, D. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58.

RUBIN, D. (1986). Comment on Holland (1986). *J. Amer. Statist. Assoc.* **81** 961–962.

SMITH, A. F. M. (1984). Present position and potential developments: some personal views. Bayesian statistics. *J. Roy. Statist. Soc. Ser. A* **147** 245–259.

TJUR, T. (2000). Contribution to the discussion of McCullagh (2000). *J. Roy. Statist. Soc. Ser. B* **62** 238–239.

WHITTLE, P. (1974). Contribution to the discussion of Besag (1974). *J. Roy. Statist. Soc. Ser. B* **36** 228.

YANDELL, B. S. (1997). *Practical Data Analysis for Designed Experiments*. Chapman and Hall, London.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
5734 UNIVERSITY AVENUE
CHICAGO, ILLINOIS 60637–1514
E-MAIL: pmcc@galton.uchicago.edu

DISCUSSION

BY JULIAN BESAG[1]

University of Washington

Introduction. I am glad of the opportunity to discuss some aspects of Peter McCullagh's paper. Parametric statistical formulations have recently come under intense attack [e.g., Breiman (2001)] but I strongly disagree with the notion that they are no longer relevant in contemporary data analysis. On the contrary, they are essential in a wealth of applications where one needs to compensate for the paucity of the data. Personally, I see the various approaches to data analysis (frequentist, Bayesian, machine learning, exploratory or whatever) as complementary to one another rather than as competitors for outright domination. Unfortunately, parametric formulations become easy targets for criticism when, as occurs rather often, they are constructed with too little thought. The lack of demands on the user made by most statistical packages does not help matters and, despite my enthusiasm for Markov chain Monte Carlo (MCMC) methods, their ability to fit very complicated parametric formulations can be a mixed blessing. So, in that sense, McCullagh's paper is timely and of course it makes many valid points but I also think it is misconceived, both in its general aims and in the agricultural application discussed in Section 8.

General comments. My overall impression of McCullagh's framework is that it really concerns mathematical models, whereas statistical models are more subtle, which makes the subject in some ways more difficult and more rewarding. That

[1]Supported in part by the Center for Statistics and the Social Sciences with funds from the University Initiatives Fund at the University of Washington.

said, I have always disliked the term "statistical model" because I think it accords a status to formulations that is seldom earned in practice. In spatial statistics, I failed to persuade anyone to adopt the term "scheme" [Besag (1974)] or "prescription" [Besag (1975)] and these days often fall back on "representation" (with or without its Bayesian connotations). I prefer to reserve the term "model" for something that has a physical justification, such as the Poisson process in certain applications. I doubt whether generalized linear models, for example, usually fall into this category, any more than do spatial autoregressions. McCullagh seems to demand the universality of a physical model, but without doing any of the physics, and surely this is too much to ask? A statistical formulation can still be very useful without universality, so long as one restricts inferences to what might loosely be described as "interpolation" rather than "extrapolation." Is this not what we teach our students? Yet several of McCullagh's criticisms and examples involve "extrapolation" to parts of a space in which there is no claim that the original formulation holds. This is different from the valid points that are made about ensuring consistency between different measurement scales, say.

Moreover, if I were forced by law to call something a statistical model, then I would insist that it must be useable for and assessable by data analysis, whereas a McCullagh statistical model is a mathematical object involving data but not necessarily data analysis. As a concrete example, I think McCullagh views the formulation in Section 8.6 as a lattice-based approximation to a previous statistical model, whereas I would interpret it as a statistical scheme motivated by a previous (I believe untenable) mathematical model.

Maybe the above point is a minor one, but it is related to another that to me is crucial and yet seems absent from McCullagh's framework, namely the role of approximation. Indeed, he writes, "Algebra does not easily deal with approximations or inequalities." It seems to me that, except for some rather limited randomization analyses, the initial choice of an appropriate level of approximation is a vital ingredient in almost any statistical investigation (though maybe requiring review in the light of the available data) and is perhaps the most creative aspect of our subject. McCullagh seeks to produce rules within which creativity must operate, but it is not obvious that his formalism takes us constructively beyond sound common sense. Furthermore, it can impose quite unreasonable constraints. Obviously, I agree that one should always recognize and seek to rectify mathematical inconsistencies but not slavishly so if this means abandoning a scientifically useful approximation in favor of a formulation that is universally self-consistent but unsustainable in the region of interest.

Although, in principle, the purpose of an investigation should govern the formulation and the type and the amount of data collected, practice is often harsher and the specification one adopts depends qualitatively on the available sample size. Fisher (1922) wrote, "More or less elaborate forms will be suitable according to the volume of the data." Indeed, this interdependence provides the salvation of frequentist p-values against the argument that failure to "reject" a

model occurs merely because the sample size is too small. That is, the argument becomes irrelevant if one recognizes that the generality of statistical formulations should normally increase with sample size, until ultimately one may indeed be allowed the luxury of machine learning. I may need to add that my interpretation of a p-value is merely as an exploratory device that quantifies inconsistency between the observed data and a particular formulation. McCullagh seems not to accept that sample size should be an important ingredient in statistical modeling; see Section 4.2. Of course, I agree that generally there should be coherence between formulations in going from sample size n to $n + 1$ but this process also takes us from n to n^2 and then there will often be overriding concerns and a need for greater flexibility.

Markov random fields. Markov random fields (MRFs) are distributions specified via their full conditionals (originally called local characteristics). The identification between any particular MRF and a corresponding Gibbs distribution and vice versa follows from the Hammersley–Clifford theorem [e.g., Besag (1974)]. The only restriction on either class is a minor positivity condition, so Gibbs (say) distributions are not required to have the property McCullagh ascribes to them in Section 6.6.

Exercise 7 is supposedly an example of the inconsistencies that plague MRFs but seems badly chosen. I agree of course that the formulation is bizarre but, without a context, there is no reason why the distribution of Y in a cluster of size k, marginalized over its kth component, should be required to coincide with that of Y in a cluster of size $k - 1$. For example, suppose that k refers to the litter sizes in a study of piglets and that a measurement is made on each piglet. Then exchangeability within litters might be a reasonable assumption but marginalizing over a piglet in a litter of size k does not produce a litter of size $k-1$. And whatever the formulation, it seems reckless even to contemplate drawing inferences about litters of size eight from data merely on litters of size two!

A better example of the contradictions that arise in MRF formulations is mentioned in McCullagh's discussion of Exercise 11. That is, a parametric MRF on a grid (say) is generally inconsistent with the corresponding MRF on a subset of the grid. In principle, consistency can be restored by conditioning on an appropriate boundary but this is usually too wasteful in practice. Partial fixes may be available by using marginalizations of MRFs on $\mathcal{Z}^2$ or otherwise; see Besag and Kooperberg (1995), Besag and Higdon (1999) and Rue and Tjelmeland (2002). However, spatial effects are often of secondary importance, as in variety trials, and the main intention is to absorb an appropriate level of spatial variation in the formulation, rather than produce a spatial model with scientifically interpretable parameters. Nevertheless, McCullagh's basic point is well taken. For example, I view the use of MRFs in geographical epidemiology [e.g., Besag, York and Mollié (1991)] as mainly of exploratory value, in suggesting additional spatially related covariates whose inclusion would ideally dispense with the need for a spatial formulation;

see Byers and Besag (2000) for an example on prostate cancer and ethnicity. A particularly blatant abuse of MRFs occurs in the analysis of social networks, where the parameters in Markov random graphs are often ascribed substantive interpretations that are meaningless, if only because they depend on the size of the system. I anticipate that MRFs will play a diminishing role in statistical analysis but currently they still have useful computational advantages when MCMC is used.

Agricultural field experiments. Although field experiments no longer feature in most graduate programs, their design and analysis comprise an important area of application for statistics. Variety trials usually involve say 25 to 75 varieties of a crop, with very limited replication, perhaps three or, even worse, only two plots being assigned to each variety. Here I exclude early generation trials, often having very large numbers of varieties and no replication but with check plots of a standard variety used as controls.

It has always been recognized that generally a crop will perform much more similarly on two plots close together than on plots further apart. Thus, Fisher [(1928), page 229] wrote, "... the peculiarity of agricultural field experiments lies in the fact, verified in all careful uniformity trials, that the area of ground chosen may be assumed to be markedly heterogeneous, in that its fertility varies in a systematic, and often a complicated manner from point to point." Soil measurements, such as pH, are not generally available to make adjustments for fertility. Fisher's solution to the problem resulted in the design and analysis of experiments, an approach that provides rigorous inference via randomization analysis but, for modern-day experiments, can be very inefficient when compared to model-based analysis.

McCullagh refers frequently to agricultural experiments and in Section 8 proposes a spatial formulation based on harmonic functions. Despite initial resistance, spatial methods have become increasingly popular: for example, frequentist spatial analysis is now used in some 5000 experiments annually in Australia alone [Gilmour, Cullis, Smith and Verbyla (1999)]. In several places, McCullagh mentions Besag and Higdon (1999), henceforth BH, though with no obvious enthusiasm. As background, BH describes a Bayesian approach to statistical analysis, with some emphasis on variety trials, and examines several complicated data sets; easier examples are analyzed in Besag and Higdon (1993), in Besag, Green, Higdon and Mengersen [(1995), Section 5] and in the rejoinder to discussion in BH. A first-order Gaussian intrinsic autoregression is used as a simple but flexible baseline representation of the spatial variation in fertility. BH does not pretend to provide a complete solution and indeed discusses some current deficiencies. Below I will confine my comments to points raised by McCullagh.

Response scale (Section 3.2). McCullagh makes the basic point that statistical analysis should not depend on response scale. BH achieves this by standardizing the raw data, which is effective but rather untidy in a Bayesian framework.

Covariate space (Section 3.2). McCullagh states that, for a trial in which fertilizer is applied at rates in the range 0–300 kg/ha, he requires the inferential universe to extend to all nonnegative rates. Yet this seems pointless without a corresponding extension of the model itself, even though any such extension cannot be assessed. Indeed, I presume that McCullagh himself would be unwilling to draw inferences at a rate of, say, 400 kg/ha, without additional information. Similarly, in his example on potatoes, I would not address varieties that are not in the experiment or remain yet to be invented. Of course, this is not the case if the tested varieties form a random sample from a definite population.

Experimental units (Section 3.2). In discussing variety trials, McCullagh claims, "It is invariably understood, though seldom stated explicitly, that the purpose of such a trial is to draw conclusions concerning variety differences, not just for plots of this particular shape, size and orientation, but for comparable plots of various shapes, sizes and orientations." Here I would replace "invariably" by "almost never." McCullagh seems to confuse variety trials with uniformity trials in which a single variety is grown. Uniformity trials (e.g., the Mercer and Hall data in Section 8.6) were used long ago to investigate optimal plot size for genuine experiments, but they are performed very rarely now and plot dimensions are often determined by management practice rather than by statistical criteria. However, in passing I note that the asymptotic logarithmic behavior of the variogram for the BH intrinsic autoregression is in good agreement with the empirical findings from uniformity trials in Fairfield Smith (1938) and Pearce (1976).

Of course, in a genuine variety trial, one might want to predict what the aggregate yield over the entire field would have been for a few individual varieties but this does not require any extension of the formulation to McCullagh's conceptual plots. Indeed, such calculations are especially well suited to the Bayesian paradigm, both theoretically, because one is supposed to deal with potentially observable quantities rather than merely with parameters, and in practice, via MCMC, because the posterior predictive distributions are available rigorously. That is, for the aggregate yield of variety A, one uses the observed yields on plots that were sown with A and generates a set of observations from the likelihood for those that were not for each MCMC sample of parameter values, hence building a corresponding distribution of total yield. One may also construct credible intervals for the difference in total yields between varieties A and B and easily address all manner of questions in ranking and selection that simply cannot be considered in a frequentist framework; for example, the posterior probability that the total yield obtained by sowing any particular variety (perhaps chosen in the light of the experiment) would have been at least 10% greater than that of growing any other test variety in the field.

I am aware that the previous paragraph may be misconstrued. David Higdon and I are primarily "spatialists" rather than card-carrying Bayesians and BH merely explores some consequences of a Bayesian approach. There are other caveats. The above arguments assume that the formulation is correct, though one can and should carry out sensitivity analysis; also any model-based formulation should leave room for outliers and other aberrations, which BH discusses via hierarchical t's and other robust procedures that relax the baseline Gaussian distributions.

More to the point, neither the agronomists who make recommendations nor the farmers themselves care too much about the difference in yield between varieties A and B grown on one particular experimental field. I presume McCullagh would include other "similar" fields in his inferential universe, something I feel is useful only if there is additional external information. Ideally, and often in practice, several trials are carried out in a range of environments, in which case progress can be made, perhaps invoking a hierarchical formulation; see BH for an example on corn (maize) grown at six different sites in North Carolina. This also downgrades the importance of the specific assumptions that are made at plot level (and at any finer scale). If trials are not conducted at several sites, then recommendations need to be moderated appropriately. Incidentally, the use of posterior predictive distributions, with their additional variability, might be helpful in curbing the frustration of plant breeders when the "best" variety experimentally does not subsequently perform as well.

Fertility (Section 8). Some care is required in what is meant by "fertility": both McCullagh and BH are rather vague. Fertility does not exist in abstraction, nor even in a particular field, because it means different things for different crops. To me, "fertility" represents the plot-to-plot variation due to the environment, if a single variety in the trial is grown throughout. This is well defined only if one can make the usual assumption that variety effects are additive, perhaps after a transformation of yield. My version of fertility usually includes standard fixed effects caused by human intervention, such as those due to blocks (if these have a physical presence) or to boustrophedon harvesting, but here I shall assume such fixed effects are taken into account separately. However, fertility must also cater for all other environmental (usually thought of as random) effects and it is generally recognized that these include indeterminate plot-aligned effects due to management practice. McCullagh writes, "It is absolutely essential to take account of the geometry of the plots" but he ignores the influence of other effects that destroy any assumed isotropy in the natural variation of the soil; and underlying isotropy is itself a somewhat dubious assumption in the first place. I agree that McCullagh's rule would often be adequate in practice (e.g., for the Mercer and Hall data); indeed, many trials these days use plots that are sufficiently long and narrow for one-dimensional representations to suffice. However, I would still prefer to include a parameter for directionality, as in BH, rather than rely on the rule. This may involve a substantial computational overhead, which McCullagh

conveniently avoids, but it sacrifices negligible information even when the rule holds. For an example where one-dimensional adjustment suffices but in the *opposite* direction to that expected, see Besag and Higdon [(1993), Section 1]. McCullagh's "absolutely essential" seems to me to indulge an inappropriate mathematical assumption.

Variety effects (Section 8.3). McCullagh claims, "variety effects must be multiplicative," as a result of certain mathematical assumptions. This would lead me to reject the assumptions! Variety effects cannot be perfectly additive on yield itself, because yields cannot be negative, but this is a different point and is usually irrelevant in practice. Of course, I agree that one often needs to transform the data to achieve approximate additivity.

Incidentally, the need to attach a prior to variety effects in a Bayesian analysis will be seen by many as a handicap but I think this is mistaken. Experience suggests close *numerical* agreement between frequentist and Bayesian results for corresponding spatial formulations when an approximately uniform variety prior is adopted. However, in practice, varieties within known groups may be genetically very similar, in which case it is natural to adopt appropriate Gaussian (or perhaps t) priors, which result in appropriate shrinkage of the estimates, particularly when there is little replication. This would again help prevent the frustration of plant breeders mentioned above.

Stationarity (Section 8.3). McCullagh agrees that stationarity of the environmental (random) effects is dubious (at the scale of the experiment). My experience in fitting stationary lattice schemes is that typically one obtains parameter values on the nonstationary edge of the parameter space, a finding that is not restricted to agricultural data. Thus, following Künsch (1987), I have preferred instead to use limiting intrinsic versions. For example, in one spatial dimension, a first-order stationary autoregression becomes a random walk, with an arbitrary level rather than being tied down somewhere. The increments need not be Gaussian. In two dimensions, a first-order Gaussian intrinsic autoregression can be interpreted in terms of *locally* planar interpolation, in the same sense that a Gaussian random walk, viewed bilaterally, is locally linear. Again the level is arbitrary, in accordance with a definition of fertility based on variation. Mathematically, the scheme is an independent increments process, subject to the logical constraint that the increments on any loop sum to zero. This may be a useful interpretation in devising non-Gaussian versions. BH also discusses locally quadratic representations that do not attenuate peaks and troughs but suggests that these may overfit fertility.

Spatial scale (Section 8). Scale invariance, at least to a very good approximation, is a requirement for any genuine model of "fertility" and so it is natural to begin in continuous space, even if eventual interest is in discrete plots. However, it is less clear whether one should work in four-dimensional space–time, in

three-dimensional space or merely in two dimensions. Even in two dimensions, there has been rather limited success in devising plausible formulations that are amenable to integration. In geography, the issue of invariance is referred to as the modifiable areal unit problem and has a long history. In statistics, early references are Heine (1955), Whittle (1962) and Matérn (1986), first published (remarkably) in 1960. However, it seems extremely unlikely that any formulation can be thought of as an accurate model for variation in fertility without additional measurements of factors that influence the growth of the particular crop under study. These could be in the form of extensive soil measurements, such as pH levels, or the use of check plots of a standard variety, dispersed over the experimental site, as occurs in single-replicate early generation trials.

Fortunately, the sole purpose of variety trials is to compare varieties, not to assess spatial variation, which enters the formulation merely as a nuisance factor. With the benefit of some replication, it seems reasonable to expect that an approximate representation of "fertility" is generally adequate for statistical analysis. All the investigations that I know of support this view. Such investigations usually involve uniformity data to which dummy variety effects are added, so that the true values are known to be zero. An early example is in Besag and Kempton (1986). The findings typically suggest that the gains from spatial analysis in a badly designed experiment provide improvements commensurate with standard analysis and optimal design. This is not a reason to adopt poor designs but the simple fact is that, despite the efforts of statisticians, many experiments are carried out using nothing better than randomized complete blocks.

It is highly desirable that the representation of fertility is flexible but is also parsimonious because there are many variety effects to be estimated, with very limited replication. McCullagh's use of discrete approximations to harmonic functions in Section 8 fails on both counts: first, local maxima or minima cannot exist except (artificially) at plots on the edge of the trial; second, the degrees of freedom lost in the fit equals the number of such plots and is therefore substantial (in fact, four less in a rectangular layout because the corner plots are ignored throughout the analysis!).

Nevertheless, there is something appealing about the averaging property of harmonic functions, if only it were a little more flexible. What is required is a random effects (in frequentist terms) version and that is precisely the thinking behind the use of intrinsic autoregressions in BH and elsewhere. Indeed, such schemes fit McCullagh's discretized harmonic functions perfectly, except for edge effects (because BH embeds the array in a larger one to cater for such effects), and they also provide a good fit to more plausible fertility functions. For specific comments on the Mercer and Hall data, see below.

Of course, spatial scale remains an important issue for variety trials and indeed is discussed empirically in Section 2.3 and in the rejoinder of BH. For one-dimensional adjustment, the simplest plausible continuum process is Brownian motion with an arbitrary level, for which the necessary integrations can be

implemented rigorously in the analysis. In the example in BH, there is close agreement between the estimates from the discrete and continuous formulations (which are not quite consistent mathematically). In two-dimensional adjustment, one can experiment with splitting the plots and comparing the results obtained from the fertility priors at the two scales. This can be done rigorously via MCMC by treating the yields in each pair of half plots as unknown but summing to the observed value. The few examples I have tried again suggest close agreement but, of course, I would much rather see a sound mathematical justification of approximate closure under spatial aggregation. This might be available via an appropriate extension of McCullagh's harmonic processes.

Mercer and Hall data (Section 8.6). McCullagh does not draw a clear distinction between the purposes of analyzing data from uniformity trials and from genuine variety trials. He also comments on a purely illustrative analysis of mine from more than 25 years ago about which I wrote [Besag (1974)], "It cannot be claimed that the present auto-normal schemes have been successful in reflecting the overall probabilistic structure of the wheat plots process." The perhaps definitive discussion of the Mercer and Hall data is McBratney and Webster (1981), which uses the original records from Rothamsted Experimental Station to explain the characteristics noted by McCullagh, in terms of a previous ridge and furrow system. McCullagh's formulation includes fixed effects for rows and columns in addition to those for the harmonic approximation, so that more than 120 parameters are fitted. This type of approach does not seem well suited to variety trials. The BH formulation fits two parameters, one of which provides data-driven directional flexibility, which McCullagh does not allow. Although, after 90 years, the Mercer and Hall data are now very tired indeed and deserve a decent burial, it might be worth noting that the basic BH fit at least retains all of the peaks and troughs in McBratney and Webster's Table 1, though it is certainly not a physical model.

Spatial design (Section 8.7). McCullagh proposes the development of a theory of efficient harmonic designs for agricultural experiments. Such designs would be very close to those that exist already for more relevant spatial formulations. For a recent review, see Atkinson and Bailey (2001), especially Section 10.

Conclusion. Although McCullagh's paper makes many valuable points, I believe that the approach is too rigid and places an emphasis on pure mathematics that is inappropriate for applied statistics. The paper promotes a universality in statistical modeling that is seldom present or necessary in practice. The role of approximation and its interrelationship with sample size seem to be ignored. As regards spatial statistics, the paper rediscovers old problems but does not yet provide effective solutions. Nevertheless, I am glad to end on a positive note by agreeing that the generalized covariance function in Section 8.6, known as the

de Wijs model in geostatistics and dating back to the early 1950s, may be useful in developing more coherent spatial formulations for the analysis of agricultural field experiments.

Acknowledgments. I am grateful to Ted Harding for very helpful discussions on this topic and to Peter McCullagh for his patience in arguing about specific points.

REFERENCES

ATKINSON, A. C. and BAILEY, R. A. (2001). One hundred years of the design of experiments on and off the pages of *Biometrika*. *Biometrika* **88** 53–97.

BESAG, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **36** 192–236.

BESAG, J. E. (1975). Statistical analysis of non-lattice data. *The Statistician* **24** 179–195.

BESAG, J. E., GREEN, P. J., HIGDON, D. M. and MENGERSEN, K. L. (1995). Bayesian computation and stochastic systems (with discussion). *Statist. Sci.* **10** 3–66.

BESAG, J. E. and HIGDON, D. M. (1993). Bayesian inference for agricultural field experiments. *Bull. Internat. Statist. Inst.* **55** 121–136.

BESAG, J. E. and HIGDON, D. M. (1999). Bayesian analysis of agricultural field experiments (with discussion). *J. Roy. Statist. Soc. Ser. B* **61** 691–746.

BESAG, J. E. and KEMPTON, R. A. (1986). Statistical analysis of field experiments using neighbouring plots. *Biometrics* **42** 231–251.

BESAG, J. E. and KOOPERBERG, C. L. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82** 733–746.

BESAG, J. E., YORK, J. C. and MOLLIÉ, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Ann. Inst. Statist. Math.* **43** 1–59.

BREIMAN, L. (2001). Statistical modeling: the two cultures (with discussion). *Statist. Sci.* **16** 199–231.

BYERS, S. D. and BESAG, J. E. (2000). Inference on a collapsed margin in disease mapping. *Statistics in Medicine* **19** 2243–2249.

FAIRFIELD SMITH, H. (1938). An empirical law describing heterogeneity in the yields of agricultural crops. *J. Agric. Sci.* **28** 1–23.

FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A* **222** 309–368.

FISHER, R. A. (1928). *Statistical Methods for Research Workers*, 2nd ed. Oliver and Boyd, Edinburgh.

GILMOUR, A. R., CULLIS, B. R., SMITH, A. B. and VERBYLA, A. P. (1999). Discussion of paper by J. E. Besag and D. M. Higdon. *J. Roy. Statist. Soc. B* **61** 731–732.

HEINE, V. (1955). Models for two-dimensional stationary stochastic processes. *Biometrika* **42** 170–178.

KÜNSCH, H. R. (1987). Intrinsic autoregressions and related models on the two-dimensional lattice. *Biometrika* **74** 517–524.

MATÉRN, B. (1986). *Spatial Variation*. Springer, New York.

MCBRATNEY, A. B. and WEBSTER, R. (1981). Detection of ridge and furrow pattern by spectral analysis of crop yield. *Internat. Statist. Rev.* **49** 45–52.

PEARCE, S. C. (1976). An examination of Fairfield Smith's law of environmental variation. *J. Agric. Sci.* **87** 21–24.

RUE, H. and TJELMELAND, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scand. J. Statist.* **29** 31–49.

WHITTLE, P. (1962). Topographic correlation, power-law covariance functions, and diffusion. *Biometrika* **49** 305–314.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON
BOX 354322
SEATTLE, WASHINGTON 98195-4322
E-MAIL: julian@stat.washington.edu

DISCUSSION

BY PETER J. BICKEL

University of California, Berkeley

Peter McCullagh is to be congratulated for this formidable paper reexamining the questions of what is a statistical model and what is a parameter in terms of one of the major mathematical developments of the late twentieth century, the algebraic theory of categories.

I cannot but agree with the major points that McCullagh makes in connection with designed experiments which are at the base of his analysis.

1. Models must be embedded in suitable structures that permit extrapolation. A rather practical example of this issue not explicitly raised in this paper arises in well-designed clinical trials where permutation and rank tests assure us that any inference we make about the null hypothesis is valid without any assumptions on the nature of the subjects. This is, however, essentially irrelevant since we are interested not in the group on which the trials are being performed but a putative population of interest from which they were drawn. That is, one needs to talk about probability distributions referring to populations in the end rather than experimental randomization.
2. One must also consider a world of experiments in which to embed the experiment one performs. This seems an entirely correct and novel insight. That means one must consider collections of sets of probability distributions with specified allowable relations between them.
3. One must only consider parameters viewed as functions on the set of distributions to be relevant if they map properly in terms of the relations between experiments.

Unfortunately, I find the mathematics of category theory which McCullagh, I think necessarily, has to invoke, entirely unfamiliar. Since, in general, I have been

interested in data not from designed experiments, hopefully such mathematics is not too necessary.

Having said this I would like to comment on an example discussed by McCullagh, Box–Cox transformations, which Doksum and I discussed many years ago. As McCullagh notes, Hinkley and Runger validly criticized the analysis in which we concluded that one could not act as if the shape parameters were known in estimating regression parameters β, that is, that there was variance inflation due to the correlation between the estimates of the shape parameter λ and the estimate of β. Their critique was that statements about β in the absence of λ were meaningless. McCullagh gives a formal argument in his framework why β by itself is not a parameter when one takes into account that parameters must map appropriately under the mappings between experiments in which one changes units by scalar multiplication or taking powers.

There are two attractive choices among the parameters he lists as allowable, β/σ and (β, λ). The first has been considered by a number of authors, most recently by Chen, Lockhart and Stephens (2002). One reason for its attractiveness to me is that if one considers the more realistic semiparametric model,

$$a(Y) = \beta X + \varepsilon, \tag{6}$$

where a is an arbitrary monotone transformation and ε has a $\mathcal{N}(\mu, \sigma^2)$ distribution then β/σ is identifiable and estimable at the $n^{-1/2}$ rate while β is not identifiable. Bickel and Ritov (1997) discuss ways of estimating β/σ and a which is also estimable at rate $n^{-1/2}$ optimally and suggest approaches to algorithms in their paper.

The choice (β, λ) is of interest to me because its consideration is the appropriate response to the Hinkley–Runger critique. One needs to specify a joint confidence region for (β, λ) making statements such as "the effect magnitude β on the λ scale is consistent with the data."

The effect of lack of knowledge of λ on the variance of β remains interpretable.

It would be more attractive if McCullagh could somehow divorce the calculus of this paper from the language of functors, morphisms and canonical diagrams for more analysis-oriented statisticians such as myself.

REFERENCES

BICKEL, P. and RITOV, Y. (1997). Local asymptotic normality of ranks and covariates in the transformation models. In *Festschrift for Lucien Le Cam*: *Research Papers in Probability and Statistics* (D. Pollard, E. Torgersen and G. Yang, eds.) 43–54. Springer, New York.

CHEN, G., LOCKHART, R. A. and STEPHENS, M. A. (2002). Box–Cox tranformations in linear models: Large sample theory and tests of normality (with discussion). *Canad. J. Statist.* **30** 177–234.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
411 EVANS HALL
BERKELEY, CALIFORNIA 94720-3860
E-MAIL: bickel@stat.berkeley.edu

DISCUSSION

BY HANS BRØNS

University of Copenhagen

Peter McCullagh's paper is exciting, because it can be seen as the start of a new, long overdue discussion of the mathematical foundation of the theory of statistics. He rightly points out that only part of the thinking in theoretical statistics is formalized mathematically and tries to extend the existing theory of statistical modelling using modern abstract mathematical tools. This is a great philosophical challenge, but it is a formidable pedagogical task to communicate the results to the statisticians.

The paper contains beyond the abstract definition of the new extended concept of a statistical model a treasure trove of examples and counterexamples, but I shall concentrate on an analysis of the definition of models. McCullagh's idea is that parameter spaces and sample spaces which are usually treated as sets or measure spaces in applications have a richer structure defining what one could be tempted to call their "physical nature," which should be reflected in the models and in the choice of transformations between them. This is done by giving them an inner structure, symmetry for example, and by considering each model as a unit in a greater universe of models. To give this a mathematical expression, the much loved and much hated theory of categories is used.

McCullagh's categories are a little bushy. Any object in any category can be connected by an arrow with any other object in any of the categories considered. They are evidently assumed to be built on sets and can therefore be connected by functions; that is, they are all concrete categories [Mac Lane (1998), page 26]. A category $\mathcal{A}$ is *concrete*, if it is equipped with a faithful (injective on hom-sets) functor to the category **Set** of sets and functions. Let this forgetful functor be denoted by $U_{\mathcal{A}}$.

Several of the categories constructed by McCullagh are so-called comma categories [Mac Lane (1998), page 46]. Consider a diagram

$$\mathcal{B} \xrightarrow{T} \mathcal{A} \xleftarrow{S} \mathcal{C}$$

of three categories and two functors with the same codomain. The *comma category* from T to S, denoted (T, S) or $(T \downarrow S)$ as we shall do, has as objects triples (b, c, f), where b is an object in category $\mathcal{B}$, c an object in category $\mathcal{C}$, and $f : Tb \to Sc$ is an arrow in category $\mathcal{A}$. Arrows between objects (b, c, f) and (b', c', f') in the category $(T \downarrow S)$ are all pairs (g, g') consisting of an arrow $g : b \to b'$ in category $\mathcal{B}$ and an arrow $g' : c \to c'$ in category $\mathcal{C}$ such that the

diagram

$$\begin{array}{ccc} Tb & \xrightarrow{Tg} & Tb' \\ \downarrow{\scriptstyle f} & & \downarrow{\scriptstyle f'} \\ Sc & \xrightarrow{Sg'} & Sc' \end{array}$$

is commutative. Composition in $(T \downarrow S)$ is coordinatewise, and identities are pairs of identities.

By taking the first and second coordinates of the objects and arrows in the comma category it is provided with a pair of natural (forgetful) projection functors

$$\mathcal{B} \xleftarrow{\mathrm{Pr}_{\mathcal{B}}} (T \downarrow S) \xrightarrow{\mathrm{Pr}_{\mathcal{C}}} \mathcal{C}.$$

Let $\mathcal{D}$ be a category with object class $\mathrm{obj}(\mathcal{D})$ and let $D : \mathcal{D} \to (T \downarrow S)$ be a functor (if $\mathcal{D}$ is small enough D would be called a *diagram* over $\mathcal{D}$ in the comma category). For $d \in \mathrm{obj}(\mathcal{D})$ put $D(d) = (D_1(d), D_2(d), D_3(d))$ such that $D_3(d) : T(D_1(d)) \to S(D_2(d))$ in $\mathcal{A}$. By joggling the symbols it is seen that the family $(D_3(d))_{d \in \mathrm{obj}(\mathcal{D})}$ is a natural transformation $T \circ D \to S \circ D : \mathcal{D} \to \mathcal{A}$. The mapping $D \mapsto (\mathrm{Pr}_{\mathcal{B}} \circ D, \mathrm{Pr}_{\mathcal{C}} \circ D, (D_3(d))_{d \in \mathrm{obj}(\mathcal{D})})$ is a 1–1 correspondence between functors $D : \mathcal{D} \to (T \downarrow S)$ and triples $(D_{\mathcal{B}}, D_{\mathcal{C}}, \pi)$ consisting of a functor $D_{\mathcal{B}} : \mathcal{D} \to \mathcal{B}$, a functor $D_{\mathcal{C}} : \mathcal{D} \to \mathcal{C}$, and a natural transformation $\pi : T \circ D_{\mathcal{B}} \to S \circ D_{\mathcal{C}} : \mathcal{D} \to \mathcal{A}$. In this way a diagram in $(T \downarrow S)$ is identified with a natural transformation between certain functors.

In his most general definition of a statistical model McCullagh starts with three categories: (1) the category $\mathrm{cat}_{\mathcal{U}}$ of statistical units, (2) the category $\mathrm{cat}_{\mathcal{V}}$ of response scales, and (3) the category cat_{Ω} of covariate spaces. From these he first constructs the category $\mathrm{cat}_{\mathcal{D}}$ of designs as a comma category, in our notation:

$$\mathrm{cat}_{\mathcal{D}} = (U_{\mathrm{cat}_{\mathcal{U}}} \downarrow U_{\mathrm{cat}_{\Omega}})$$

and then the category $\mathrm{cat}_{\mathcal{S}}$ of sample spaces as a product category,

$$\mathrm{cat}_{\mathcal{S}} = \mathrm{cat}_{\mathcal{V}} \times \mathrm{cat}_{\mathcal{U}}^{\mathrm{op}}.$$

Finally, the basic category is the product category $\mathrm{cat}_{\mathcal{V}} \times \mathrm{cat}_{\mathcal{D}}^{\mathrm{op}}$. Let

$$\mathrm{cat}_{\mathcal{V}} \xleftarrow{\mathrm{Pr}_{\mathcal{V}}} \mathrm{cat}_{\mathcal{V}} \times \mathrm{cat}_{\mathcal{D}}^{\mathrm{op}} \xrightarrow{\mathrm{Pr}_{\mathrm{cat}_{\mathcal{D}}^{\mathrm{op}}}} \mathrm{cat}_{\mathcal{D}}^{\mathrm{op}}$$

be the diagram of projections. Since

$$\mathrm{cat}_{\mathcal{V}} \times \mathrm{cat}_{\mathcal{D}}^{\mathrm{op}} \xrightarrow{\mathrm{id}_{\mathrm{cat}_{\mathcal{V}}} \times \mathrm{Pr}_{\mathrm{cat}_{\mathcal{U}}^{\mathrm{op}}}} \mathrm{cat}_{\mathcal{V}} \times \mathrm{cat}_{\mathcal{U}}^{\mathrm{op}} = \mathrm{cat}_{\mathcal{S}},$$

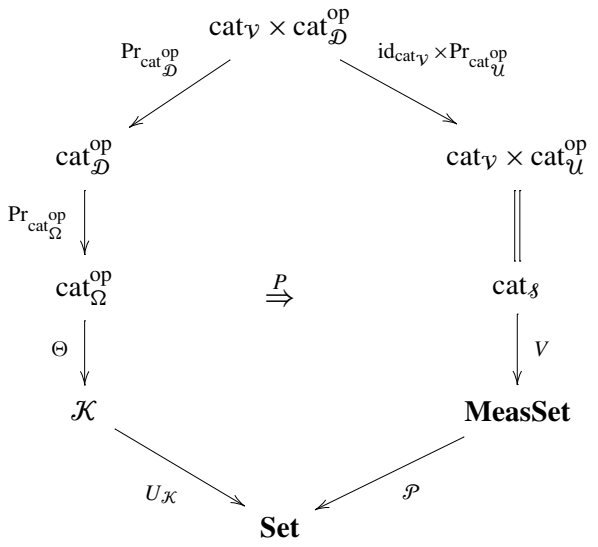
both $\mathrm{cat}_{\mathcal{D}}$ and $\mathrm{cat}_{\mathcal{S}}$ can be derived from the basic category.

The objects of $\mathrm{cat}_{\mathcal{S}}$ are the sample spaces $\mathcal{S}$ which are carriers of probability distributions with $\mathcal{P}(\mathcal{S})$ the set of probability distributions on $\mathcal{S}$. Sample spaces are therefore not just sets in general, but must be sets with some kind of measurability property by having an attached σ-algebra of subsets to support abstract probabilitiy measures, by having the structure of a locally compact topological space to be the basis for a normed Radon-measure, or by having some similar structure. Arrows between two such spaces should be measurable, proper or something similar, such that transformation of probability measures is defined. These properties should be part of the definition of the basic category, and they ensure the existence of a (faithful) functor V from $\mathrm{cat}_{\mathcal{S}}$ to the concrete category **MeasSet** of measurable sets and measurable mappings. If to an arrow $t : \mathcal{S} \to \mathcal{S}'$ in **MeasSet** we assign the mapping $\mathcal{P}(t) : \mathcal{P}(\mathcal{S}) \to \mathcal{P}(\mathcal{S}')$ that carries a probability measure on $\mathcal{S}$ into the t-transformed probability on $\mathcal{S}'$, we get a functor $\mathcal{P} : \mathbf{MeasSet} \to \mathbf{Set}$.

The parameter Θ in McCullagh's model is a contravariant functor on cat_{Ω} to a not very accurately specified concrete category $\mathcal{K}$, and the model is a natural transformation,

$$P : U_{\mathcal{K}} \circ \Theta \circ \mathrm{Pr}_{\mathrm{cat}_{\mathcal{D}}^{\mathrm{op}}} \circ \mathrm{Pr}_{\mathrm{cat}_{\mathcal{U}}^{\mathrm{op}}} \to \mathcal{P} \circ \Theta \circ (\mathrm{id}_{\mathrm{cat}_{\mathcal{V}}} \times \mathrm{Pr}_{\mathrm{cat}_{\mathcal{U}}^{\mathrm{op}}}) : \mathrm{cat}_{\mathcal{V}} \times \mathrm{cat}_{\mathcal{D}}^{\mathrm{op}} \to \mathbf{Set}$$

or in more detail,



Besides considering this overall structure McCullagh puts specific contraints on the arrows in some of the categories, constraints of the same sort as limiting factors to be surjective mappings in the multifactor analysis of variance models. Such restrictions tend to make the mathematical formalization heavier and are really not needed.

The simplest statistical models in McCullagh's paper are the (parametrized) statistical models, which are triples $(\Theta, \mathcal{S}, P)$ consisting of a parameter set, a

sample space, and a mapping $P : \Theta \to \mathcal{P}(\mathcal{S})$. The obvious choice of an arrow from a statistical model $(\Theta, \mathcal{S}, P)$ to another $(\Theta', \mathcal{S}', P')$ is a pair (ϕ, t) of a mapping $\phi : \Theta \to \Theta'$ and a measurable mapping $t : \mathcal{S} \to \mathcal{S}'$ such that the diagram

$$\begin{array}{ccc} \Theta & \xrightarrow{P} & \mathcal{P}(\mathcal{S}) \\ {\scriptstyle\phi}\downarrow & & \downarrow{\scriptstyle\mathcal{P}(t)} \\ \Theta' & \xrightarrow{P'} & \mathcal{P}(\mathcal{S})' \end{array}$$

is commutative. In this way a category **StatMod** is created, which is precisely the comma category $(\mathrm{id}_{\mathbf{Set}} \downarrow \mathcal{P})$. This category of statistical models can be interpreted as the sea in which statisticians maneuver or the universe in which they live their mathematical lives and can never escape. To include all statistical transformations Markov kernels must be allowed, but let us leave that aside.

Returning to McCullagh's general model we find, using the result above, that the natural transformation

$$P : \mathrm{id}_{\mathbf{Set}} \circ U_{\mathcal{K}} \circ \Theta \circ \mathrm{Pr}_{\mathrm{cat}_{\mathcal{D}}^{\mathrm{op}}} \circ \mathrm{Pr}_{\mathrm{cat}_{\mathcal{U}}^{\mathrm{op}}} \to \mathcal{P} \circ \Theta \circ (\mathrm{id}_{\mathrm{cat}_{\mathcal{V}}} \times \mathrm{Pr}_{\mathrm{cat}_{\mathcal{U}}^{\mathrm{op}}}) : \mathrm{cat}_{\mathcal{V}} \times \mathrm{cat}_{\mathcal{D}}^{\mathrm{op}} \to \mathbf{Set}$$

is equivalent to a functor

$$P : \mathrm{cat}_{\mathcal{V}} \times \mathrm{cat}_{\mathcal{D}}^{\mathrm{op}} \to (\mathrm{id}_{\mathbf{Set}} \downarrow \mathcal{P}),$$

or a "diagram" over the basic category in **StatModel**. McCullagh's models are therefore diagrams over a special kind of category in the category of statistical models. If we consider the complicated definition of the basic category and the difficulty with which McCullagh argues for the "scale of measurement," it is hard to see why only this special kind of category is included in the theory. The basic principle is that the elementary models considered are object values for the functor P and the arrows between them are arrow values of P. It seems natural to consider diagrams over arbitrary categories. This will then include inbeddings of subcategories, which seems to be very desirable.

McCullagh has some considerations on product sets, product categories, and repetitive models. His definition of a product category is not completely clear and not the usual one. Technically, the best way to treat structures of this kind is to use monoidal categories, monoidal functors and monoidal natural transformations, that is, categories with a tensor product, functors between two monoidal categories with a monoidal structure connecting the tensor products in the two categories and natural transformations respecting the monoidal structures of the monoidal functors they go between [see Mac Lane (1998), page 161]. In the category **Set** the Cartesian product is a tensor product, in **MeasSet** the Cartesian product of two measurable sets equipped with, for example, the product σ-algebra, is a tensor product, and the functor $\mathcal{P} : \mathbf{MeasSet} \to \mathbf{Set}$ has a monoidal structure from the mapping of two probabilities into the product probability. The monoidal structure

of the functors defining a comma category extends to a tensor product on the comma category, in our example producing the product of statistical models. The diagrams in **StatMod** should be over monoidal categories and should be monoidal functors, etc.

McCullagh's definition of a natural parameter or subparameter seems to be in accordance with the definition of the extended models. It is sad that he does not consider the consequences on the sample side. His main achievement is to have convincingly emphasized the importance of imposing further structure on the statistical models to give them more of a physical life and to have made a succesful attempt at formalizing this idea. We should be thankful for his drive and courage.

REFERENCE

MAC LANE, S. (1998). *Categories for the Working Mathematician*, 2nd ed. Springer, New York.

DEPARTMENT OF STATISTICS
AND OPERATIONS RESEARCH
UNIVERSITY OF COPENHAGEN
UNIVERSITETSPARKEN 5
DK-2100 COPENHAGEN
DENMARK
E-MAIL: hbrons@stat.ku.dk

DISCUSSION

BY D. A. S. FRASER AND N. REID

University of Toronto

This is an important and very original examination of statistical modelling, which is such a ubiquitous part of statistical theory and statistical practice that it is usually nearly overlooked. McCullagh provides a substantial examination and structuring of this serious foundational issue; indeed, the examination is both intensive and extensive. Foundational issues have not been prominent in statistics for some time, perhaps not since the shift away from the decision theoretic approach to the more pragmatic addressing of individual applied problems. It is gratifying now to see this substantial examination of the foundational issue of statistical modelling.

McCullagh motivates his development with a series of examples that display various logical or structural anomalies that can arise in the use and application of the examples. Most of the examples are extreme, but this is appropriate for highlighting potential difficulties.

His development works from the familiar definition of a model as a set of probability distributions on a sample space and a parameterized model as having

in addition a mapping from a parameter space into the set of distributions. The discussion focusses on how a model impacts on the context being examined, and how it might extend in applications to broader contexts. Category theory is used to provide the general framework for this.

One could reasonably question the familiar definition of a model in various simple ways. Do we wish to speak of just a set of possible response distributions as providing the core model of statistics? Why not perhaps the minimal step up to the perception of a black box mechanism, with the output being the response variable and an input dial giving the parameter value that determines the response distribution, or are we so influenced by set theory that we would view these as the same? One could argue that the two formulations are isomorphic, but the perception and accessibility and influence for the latter are quite different. The former describes an essentially unrelated set of possible statistical behaviors while the latter at least conceptualizes a behavioral pattern that is changed by the input dial [Fraser (1968a)].

The modelling should perhaps be more elaborate still. Most deterministic systems identify discrete component elements with linkages among these. The system is not modelled by just recording the set of pairs consisting of the determined outcome response with the corresponding input values. The components and the directions of causes and influences are modelled as well. Why should it be less for the statistical model? While the theory of graphical models does attempt to describe such linkages, much of applied statistics does not.

In many contexts there are identifiable sources for variation and error contribution. Shouldn't these individual sources, rather than the global effect, be directly modelled? The discussion in the paper does not address such more elaborate modelling. Does it matter when the response behavior is the same? It can matter in terms of what is known about the system when output values are observed. When the detailed model provides more information, one should certainly expect to obtain more for inference. An examination of statistical models that does not go beyond the pre-black box modelling to include specific elements of structure such as the error variables and other internal component elements can be viewed as short on the "extensive" side of coverage.

The attention given to component parameters is needed and very welcome. The artificiality of the regression parameters for the Box–Cox model has taken a long time to emerge with such clarity as it does here. It is also very interesting to have a mathematical explanation of the widely (but not uniformly) adopted rule in applied work that models with interactions and no corresponding main effects are not sensible.

In the context of transformation models there is some discussion of the nature of particular component parameters in Fraser [(1968b), page 65]. Consider a transformation model with parameter θ taking values in a transformation group and with identifiable error variable, say z; thus $y = \theta z$. In this transformation model setting what does it mean to say a component parameter φ is natural?

It is argued in Fraser [(1968b), page 65] that a component parameter is natural if it indexes the left cosets of a subgroup. The arguments are not dissimilar to those arguing in reverse that the regression parameter β is not a natural parameter for the Box and Cox model. Also, suppose for simplicity that the application of the group is unitary and that left and right invariant measures on the group are available. The model is invariant in the sense that $\tilde{y} = gy$ has distribution with parameter value $\tilde{\theta} = g\theta$. The application of the group to the parameter space would suggest the left invariant measure $d\mu(\theta)$ as the natural default prior measure for the parameter. But if we view the transformation θ as being relative to some initial error presentation then we might change that presentation and have $\bar{\theta} = \theta h$ applied to the modified error as $h^{-1}z$. We might then want invariance relative to the transformation $\bar{\theta} = \theta h$. This then suggests that we should use the right invariant measure $d\nu(\theta)$ as the natural default measure for the parameter, a default measure often preferred in the Bayesian context.

The weighing in of category theory leading to the definition of a natural parameter would seem to need to be qualified by an acceptance that a statistical model in an application is only providing some reasonable approximation to the reality under investigation. Thus, for example, in Section 6.1 it is noted that the coefficient of variation is not a natural component parameter; but in many applications it can be a very useful and informative parameter.

McCullagh shows how category theory provides a frame of reference for examining various simple but anomalous examples. But can category theory do more than provide a frame of reference? Can it provide a positive way of implementing and extracting anomalies in modelling? It is clear that the subject of category theory is difficult to master. Perhaps it would be better to extract simple and broadly based principles for modelling. Or will the approach need to be one of seeking the anomalous examples and then seeking the related principle.

The Cauchy example highlights an important point. McCullagh notes that the location parameter is not a natural parameter in the context of the linear fractional group. The concerns for the linear fractional group should perhaps go further. A typical transformation is not a mapping of the real line to itself, but can map a point to the point at infinity and vice versa; a rather severe lack of continuity. McCullagh (1992) used these transformations to illustrate the apparent nonuniqueness of the common ancillary statistic for location scale analysis. For this the proposed reference set from the use of the linear fractional group is the contour corresponding to the observed value of the ancillary, and will have multiple points with one or another coordinate at infinity, hardly a reference set with the reasonable continuity expected for applications. It would seem appropriate that such anomalies be avoided at the foundational assessment level for the model. Thus the use of the linear fractional group for assessing the location model should be rejected less for the anomalous nature of the location parameter and more for the fact that it does not provide mappings of the sample space to itself.

This is a very comprehensive examination of serious foundational issues for statistical theory and practice. We look forward to new examples of anomalies uncovered by category theory and also to a wider concern for foundational issues.

REFERENCE

FRASER, D. A. S. (1968a). A black box or a comprehensive model. *Technometrics* **10** 219–229.
FRASER, D. A. S. (1968b). *The Structure of Inference.* Wiley, New York.
MCCULLAGH, P. (1992). Conditional inference and Cauchy models. *Biometrika* **79** 247–259.

DEPARTMENT OF STATISTICS
UNIVERSITY OF TORONTO
100 ST. GEORGE STREET
TORONTO, ONTARIO M5S 3G3
CANADA
E-MAIL: reid@utstat.utoronto.ca

DISCUSSION

BY INGE S. HELLAND

University of Oslo

Theoretical statistics has since the beginning been founded on the idea that a statistical model is just a parametric family of probability measures. Now Peter McCullagh has given us a much richer model concept, and, importantly, this is strictly motivated from applied science considerations.

In fact this new model concept is so rich and at the same time so abstract that it will take time for us to absorb it and for the statistical community to assess its implications. Therefore I will limit these comments to the simplest situation, which is a model under symmetry, so that the categories involved are groups. The idea that it might be useful to supplement the choice of a statistical model—in the sense of a parametric class of measures—by the choice of a symmetry group, is an old one, albeit still somewhat controversial regarding its implications. One thing that is new here in McCullagh's approach is the systematic use of the concept of natural subparameter. I will expand a little on this concept, using parallel developments from Helland (2002).

Let a group G be defined on the parameter space Θ of a model. A measurable function ξ from Θ to another space Ξ is called a natural subparameter if $\xi(\theta_1) = \xi(\theta_2)$ implies $\xi(g\theta_1) = \xi(g\theta_2)$ for all $g \in G$.

For example, in the location and scale case the location parameter μ and the scale parameter σ are natural, while the coefficient of variation μ/σ is not natural (it is if the group is changed to the pure scale group). In general the parameter ξ is natural iff the level sets of the function $\xi = \xi(\theta)$ are transformed onto other

level sets by elements of the group G. Using this on the location and scale group, and making some smoothness assumptions, one can prove that ξ is natural iff $\xi = u(k_1\mu + k_2\sigma)$ for some 1–1 function u and some constants k_1 and k_2.

Thus the assumption that a given subparameter is natural is definitely a restriction. Yet McCullagh indicates through his examples that in general *statistical inference should be limited to natural subparameters*. I will give some additional arguments to support this rather radical statement.

First, ξ is a natural subparameter if and only a new group $\tilde{G}$ can be defined consistently on Ξ by $\tilde{g}\xi(\theta) = \xi(g\theta)$. This means that for natural subparameters, group-based inference can be done in a similar way as with the original parameter.

To illustrate this, consider a general and useful result on equality between confidence sets and credibility sets. Let $\xi(\theta)$ be a one-dimensional continuous natural parametric function, and let $\hat{\xi}_1(x)$ and $\hat{\xi}_2(x)$ be two equivariant estimators. [An estimator is called equivariant if $\hat{\xi}(gx) = \tilde{g}(\hat{\xi}(x))$ for all g, x.] Let the group G be locally compact, transitive and proper; that is, the actions of G on the sample space should be continuous, and the inverse image of compact sets under the mapping $(g, x) \to (gx, x)$ should be compact, a rather weak technical requirement. Define $C(x) = \{\theta : \hat{\xi}_1(x) \le \xi(\theta) \le \hat{\xi}_2(x)\}$, and let θ have a prior given by the right Haar measure from the group G. Then $C(x)$ is both a credibility set and a confidence set with the same associated probability/ confidence level.

The requirement that a subparameter should be natural may also help to resolve certain inconsistencies in statistical inference, in particular the marginalization inconsistency discussed in detail by Dawid, Stone and Zidek (1973). Their main problem is a violation of the plausible reduction principle: assume that a general method of inference, applied to data (y, z), leads to an answer that in fact depends on z alone. Then the same answer should appear if the same method is applied to z alone.

A Bayesian implementation of this principle runs as follows: assume first that the probability density $p(y, z \mid \eta, \zeta)$ depends on the parameter $\theta = (\eta, \zeta)$ in such a way that the marginal density $p(z \mid \zeta)$ only depends upon ζ. Then the following implication should hold: if (a) the marginal posterior density $\pi(\zeta \mid y, z)$ depends on the data (y, z) only through z, then (b) this $\pi(\zeta \mid z)$ should be proportional to $a(\zeta)p(z \mid \zeta)$ for some function $a(\zeta)$, so that it is proportional to a posterior based solely on the z data. For a proper prior $\pi(\eta, \zeta)$ this can be shown to hold with $a(\zeta)$ being the appropriate marginal prior $\pi(\zeta)$. Dawid, Stone and Zidek (1973) gave several examples where the implication above is violated by improper priors of the kind that we sometimes expect to have in objective Bayes inference.

For our purpose, the interesting case is when there is a transformation group G defined on the parameter space. Under the assumption that ζ is maximal invariant under G and making some regularity conditions, it is then first shown by Dawid, Stone and Zidek (1973) that it necessarily follows that $p(z \mid \eta, \zeta)$ only depends upon ζ, next (a) is shown to hold always, and finally (b) holds if and only if the prior is of the form $\nu_G(d\eta)\, d\zeta$, where ν_G is right Haar measure, and the measure

$d\zeta$ is arbitrary. Thus for this situation with such a prior not only does the reduction principle hold; we also have that the premises of the principle are automatically satisfied.

The strong assumption made above was that ζ is invariant. In a second class of examples Dawid, Stone and Zidek (1973) show that this assumption cannot be violated arbitrarily. In Helland (2002) we show that it is essentially enough to make the much weaker assumption that ζ is a natural subparameter.

Specifically, assume that ζ is natural, and let K be the subgroup of G given by $K = \{g : \zeta(g\theta) = \zeta(\theta) \text{ for all } \theta\}$. Then ζ is maximal invariant under K. Assume also that z is maximal invariant under the corresponding group acting on the sample space. Then using right Haar prior under G on the parameter space Θ, we have that any data (y, z) leads to a posterior of ζ proportional to the one obtained from only data z.

A simple example is provided by letting G be the location and scale group defined by $\mu \to a + b\mu$, $\sigma \to b\sigma$ $(b > 0)$, and then taking $\zeta = \mu$ or $\zeta = \sigma$ when the right invariant prior is $d\mu\, d\sigma/\sigma$. There is no marginalization paradox in this case.

In all such examples the choice of group is crucial. Some general requirements on this choice can be specified: (i) the class of probability distributions should be closed under the transformations in the group; (ii) if the problem is formulated in terms of a loss function, this should be unchanged when observations and parameters are transformed conformably by the group; (iii) the right Haar measure of the group should be chosen as the uninformative prior for the problem. A further requirement is that all parametric functions of interest should be natural subparameters with respect to the group. It is this last requirement which makes the location and scale group, not the group of real fractional linear transformations, the canonical choice for the Cauchy distribution case.

Statistical models can be reduced by keeping one subparameter constant and focusing on the remaining parameter. Reduction of statistical models is often useful in prediction settings. Every model reduction should be via a natural subparameter. We also know quite generally that the parameter along an orbit of the group can be estimated in an optimal way by a Pitman type estimator. Therefore, to be of use in estimation or prediction, model reduction should be limited to only concerning the orbit indices of the group. This was used in Helland (2001) to motivate model reduction in regression problems which gives a relationship to the chemometricians' partial least squares regression.

There is a situation resembling model reduction, however, which should not necessarily be done via natural subparameters: assume that $\psi \in \Psi$ is a superparameter that can be used to describe several experiments, among which a choice must be made. Assume also that a group is defined on Ψ. Then the parameters of the single experiments are not necessarily natural subparameters of ψ. It may namely be the case that the subparameter makes sense only relative to this particular subexperiment, not in any global way.

Consider as an example the following situation: assume that the model of two experiments on the same unit depends upon which experiment is carried out first, say that the parameter of the first experiment is $\theta_1(\psi, \alpha)$ if this is done first, otherwise $\theta_1(\psi, \beta)$, where there is a given relationship between α and β. Let the morphisms contain permutation of the order of the experiments. Then, in general, θ_1 will not be a natural subparameter.

It is tempting to approach quantum physics from a statistical point of view by saying that it contains a number of situations with similar properties. In fact, one can proceed quite a bit from this, but there are also very difficult questions in this approach, in particular in understanding the formal structure of quantum mechanics. It seems, however, that group representation theory may be of some help, combined with a systematic reduction of statistical models.

In working in such new directions it is very important that we now have a better characterization of what a statistical model is. The kind of precise link between common sense and formal theory which Peter McCullagh has demonstrated so forcefully in his paper, is indeed useful for several purposes, and it may set a standard for much theoretical work in statistics.

REFERENCE

DAWID, A. P., STONE, M. and ZIDEK, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **35** 189–233.

HELLAND, I. S. (2001). Reduction of regression models under symmetry. In *Algebraic Methods in Statistics and Probability* (M. Viana and D. Richards, eds.) 139–153. Amer. Math. Soc., Providence, RI.

HELLAND, I. S. (2002). Statistical inference under a fixed symmetry group. Available at http://www.math.uio.no/∼ingeh/.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF OSLO
P.O. BOX 1053 BLINDERN
N-0316 OSLO
NORWAY
E-MAIL: ingeh@math.uio.no

DISCUSSION

BY PETER J. HUBER

Klosters, Switzerland

McCullagh's paper is a plea for more realistic statistical models. He hopes to enforce sensibility of models through category theory, by imbedding some equivalent of "having a well-defined meaning" into the formal theory. I certainly agree with his aims and that it is necessary to pay attention to the inference domain

and that category theory may help you to do so. But I must confess that I stumbled over some telltale idiosyncrasies of his terminology. He variously denotes bad models as "absurd," "eccentric," "arbitrary" or "capricious." For me, a statistical model ordinarily is supposed to model some real-world situation, and as such it either is adequate, or it is not. His attributes do not apply, except to straw-man models concocted by a mathematical statistician. Correspondingly, he may go too far with his attempts to discuss sensibility of models in the abstract; in my opinion, constant reference to the underlying real-world situation is advisable and even necessary. If the latter is capricious, the model should faithfully render this capriciousness! I am thinking here, for example, of the distinction between integral and half-integral spin values in quantum mechanics, which to the uninitiated looks no less absurd or artificial than the even-odd distinction in McCullagh's Exercises 1, 2 and 4.

In the following, I shall concentrate on the category theory aspects of the paper. I should begin by recounting some bygone experiences of my own. More than forty years ago—I had just finished a thesis in category theory [Huber (1961)]—I went to Berkeley to learn statistics. There, I sat in Lucien Le Cam's courses on decision theory. He devoted much time, I believe an entire term, to the comparison of experiments in the style of his recently submitted, but as yet unpublished 1964 paper in the *Annals*. One should know that Le Cam loved abstractness and at that time was enthusiastic about Grothendieck's (1955) thesis. To avoid possible misunderstandings I should add that that thesis was in functional analysis, not in homological algebra [see Cartier (2001), page 392 for some of the background]. In Le Cam's terminology, an "experiment" is exactly the same as a "model" in McCullagh's sense, namely a parameter set Θ together with a function $P : \Theta \to \mathcal{P}(\mathcal{X})$, which assigns to each parameter point $\theta \in \Theta$ a probability distribution P_θ on the sample space $\mathcal{X}$. Le Cam, however, found it convenient to replace this classical notion of experiment by a weaker structure consisting of a vector lattice with a unit, E, together with a family indexed by Θ of positive normalized linear functionals on E [Le Cam (1964), pages 1420–1421]. With some restriction of generality, E may be thought of as the space of bounded observable random variables on $\mathcal{X}$, and to facilitate this interpretation, the set $\mathcal{X}$ was retained by Le Cam as a part of the structure. Moreover, since he wanted to handle sufficiency, simple pointwise maps would not do, and Le Cam had to replace them by randomized maps, or transition functions (see the example below). Anyway, things got pretty complex, and to sort them out in my own mind, I rephrased the fundamentals of Le Cam's theory in categorical language.

Example: *Sufficient statistics.* Let Experiment A consist in observing $X = (X_1, \ldots, X_n)$, where the X_i are i.i.d. $\mathcal{N}(\mu, \sigma^2)$, while Experiment B consists in observing two independent variables (U, V), where U is $\mathcal{N}(\mu, \sigma^2/n)$, and V/σ^2 is χ^2_{n-1}. There is a natural map from A to B [$U = \overline{X} = n^{-1} \sum X_i$, $V = S^2 =$

$\sum(X_i - X)^2]$. On the other hand, since the conditional distribution of $X = (X_1, \ldots, X_n)$ given $(\overline{X}, S^2)$ does not depend on (μ, σ^2), there is a randomized map (or transition function) from B to A, reconstructing a sample $Y = (Y_1, \ldots, Y_n)$ having the same stochastic properties as the original sample X for all values of (μ, σ). With suitably chosen definitions, these two morphisms from A to B and from B to A would be inverse to each other.

In order to deal with situations occurring in the analysis of variance, one had to consider randomized maps not only of the sample spaces, but also of the parameter spaces. Thus, morphisms in the category of statistical experiments really were suitably defined equivalence classes at least of pairs of transition functions, and built-in redundancies (vector lattices and sample spaces) complicated matters even further.

Categorical thinking, that is: thinking in terms of morphisms, functors and commutative diagrams, always helps to clarify the underlying structure of mathematical objects, to define structural isomorphisms properly, and in particular, to recognize morphisms that are natural (or canonical). Rephrasing statistical problems in categorical language had clarified my thinking about some of them in a crucial fashion. In particular, it had helped me to understand and properly formalize the concept of invariance of a statistical experiment within Le Cam's framework of topological vector lattices, and hence to formulate and prove the Hunt–Stein theorem in what I still believe to be its natural habitat; see the remarks by Brown [(1984), page 407].

For my own purposes I collected the categorized foundations of Le Cam's theory of statistical experiments in a memo which ultimately grew to about 10 to 12 pages, all of it definitions. But then, why did I not expand my little memo and prepare it for publication? There were two reasons: First, abstract categorical thinking had helped me only because I already was fluent in that language. Very few people would profit from a paper piling more abstraction on Le Cam's already very abstract setup, and I foresaw that I would have even more difficulties getting it accepted than Lucien had with his papers (he told us some anecdotes about his prolonged fights with editors and referees). Second, while my translation into categorese had streamlined the exposition, it had not added content to these particular problems. It did not unify different lines of thought (as I had been able to do in my thesis, for example), it did not lead to new theorems, except very indirectly, and did not even simplify the proof of a central statement such as the Blackwell–Sherman–Stein theorem. Apart from that, around that time I became engrossed in much more exciting robustness problems.

Now, how about McCullagh's paper? First, when we compare what he does now to what I did 40 years ago, it is quite illuminating to notice that basically the same situation can be categorized in more than one fashion. But I think the same comments, benefits and difficulties apply here too. It may be heartening to learn that such papers have become publishable in statistical journals now. McCullagh certainly is able to clarify thinking about models (but I must admit

that even I myself found it hard to get back into the categorical lingo after 40 years). On the other hand, also in his case, category theory does not seem to add content. Moreover, against McCullagh I doubt that categories have enough normative power to prevent one from doing absurd things. This may be the price we pay for the ability to model also capricious real-world situations.

REFERENCE

BROWN, L. D. (1984). The research of Jack Kiefer outside the area of experimental design. *Ann. Statist.* **12** 406–415.

CARTIER, P. (2001). A mad day's work: From Grothendieck to Connes and Kontsevich. The evolution of concepts of space and symmetry. *Bull. Amer. Math. Soc.* **38** 389–408.

GROTHENDIECK, A. (1955). Produits tensoriels topologiques et espaces nucléaires. *Mem. Amer. Math. Soc.* **16**.

HUBER, P. J. (1961). Homotopy theory in general categories. *Math. Ann.* **144** 361–385.

LE CAM, L. (1964). Sufficiency and approximate sufficiency. *Ann. Math. Statist.* **35** 1419–1455.

P.O. BOX 198
CH-7250 KLOSTERS
SWITZERLAND
E-MAIL: peterj.huber@bluewin.ch

DISCUSSION

BY RUDOLF KALMAN

ETH Zurich

"Model" and "modeling" are the most abused, distorted, misleading, self-serving terms in the contemporary scientific dictionary. Taking a hard look at this very soft and messy business is most commendable. So is the author's careful restriction of his investigation to statistical models, defined, as usual, to be families of probability distributions, and I am sure he would not object to the further qualification of mathematical statistical models. Indeed, the paper is a model of mathematical clarity.

It is very nice also to have attention focused on the technical mathematical questions of parametrization, motivated via the idea of a "natural map" first made precise by category theory, which is given a nicely organized summary in the Appendix, intended for the uninitiated enthusiast. While I cannot pretend to be an expert, has-been or to-be, as regards the examples in Section 2 around which the development of the paper is structured, common sense suggests that statistics cannot but benefit from a more careful look at the storehouse of current models. McCullagh's initiative should add much momentum to the constructive criticism of these models, serving as a substitute for real-world experiments,

which are not usually feasible in the problem area covered by the paper. In short, criticism = experimentation, and statistics, today still a nonempirical science, will be invigorated by a discussion of the kind of issues raised in the paper.

If I now come to more critical comments concerning "models" and "modeling" it should be clear that these do not apply specifically to McCullagh's work but refer to (it is sad that this has to be said) the mainstream statistical literature as a whole. My critique is that the currently accepted notion of a statistical model is not scientific; rather, it is a guess at what might constitute (scientific) reality without the vital element of feedback, that is, without checking the hypothesized, postulated, wished-for, natural-looking (but in fact only guessed) model against that reality. To be blunt, as far as is known today, there is no such thing as a concrete i.i.d. (independent, identically distributed) process, not because this is not desirable, nice, or even beautiful, but because Nature does not seem to be like that. (Historical aside: recall that physicists had thought at one time that ether was such a necessary, unavoidable, appealing, clear and beautiful concept that it must perforce exist; alas, all physicists now living had to learn that such argumentation cannot lead to good science.) As Bertrand Russell put it at the end of his long life devoted to philosophy, "Roughly speaking, what we know is science and what we don't know is philosophy." In the scientific context, but perhaps not in the applied area, I fear statistical modeling today belongs to the realm of philosophy.

To make this point seems less erudite, let me rephrase it in cruder terms. What would a scientist expect from statisticians, once he became interested in statistical problems? He would ask them to explain to him, in some clear-cut cases, the origin of randomness frequently observed in the real world, and furthermore, when this explanation depended on the device of a model, he would ask them to continue to confront that model with the part of reality that the model was supposed to explain. Something like this was going on three hundred years ago, for example, in discussions involving male and female births [Arbuthnott (1712)]. But in our times the idea somehow got lost when i.i.d. became the pampered new baby.

Without trying to fix the blame for all this (why not? someone may well be responsible, though certainly not the author) it is instructive to look at current popular hype. A random example is the recent book by Edward J. Beltrami [(1999), *What Is Random?*: *Chance and Order in Mathematics and Life*, 201 pages, Copernicus (a Springer imprint)]. In this the author, very properly, asks what randomness is in the real world; he notices that "probability" describes a certain kind of randomness, takes that as his definition of randomness and from then on speculates on the consequences, apparently forgetting that he did not get around to explaining or even to describing randomness in Nature. He merely filters through his own mind great thoughts of great thinkers about probability. Lightweight reading, a sophisticated joke, not scientific. It is ironic that a serious scientific publisher, Springer, should be hijacked into putting such shallow philosophical (in Russell's sense) stuff into a new series named after a scientific icon, a towering and independent intellect, an incorruptible critic of soft "applicable" knowledge.

REFERENCE

ARBUTHNOTT, J. (1712). An argument for Divine Providence, taken from the constant regularity observed in the births of both sexes. *Philos. Trans. Roy. Soc. London* **27** 186–190.

DEPARTMENT OF MATHEMATICS
ETH ZURICH
WEN B17 ZURICH
SWITZERLAND
E-MAIL: kalman@ufl.edu

DISCUSSION

BY STEVE PINCUS

Guilford, Connecticut

This is a very interesting and thought-provoking paper, which I enjoyed reading. While the notions of sound statistical models and parameters are essential to the core of a classical statistical framework, heretofore a broadly applicable, single (parsimonious) unifying conceptual approach and formalism had not been put forth, nor one addressing the caveats pointed out in the Introduction in the present paper. Professor McCullagh now successfully advances such a formalism, based on the natural algebraic methodology given by category theory, building on previous work on invariance and factorial models [McCullagh (2000)] and on quotient spaces and statistical models [McCullagh (1999)]. The author shows that a diverse range of sensible statistical applications, often considered as distinct settings, all can be handled within the framework of this formalism. The breadth of the range of the 12 examples in Section 2 reinforces the scope of the formalism. Some might argue that this unification could be achieved without introducing the technologic machinery of category theory or algebra. The crucial point, to my sensibilities, clarifying that the present (or similar) technology is required for a general formalism is the discussion of the naturality of subparameter specification, Section 4.5, with its implications to inference. The concept of a natural transformation is a central notion within group and category theory, but not ubiquitous throughout mathematics.

Insofar as cross-pollination between and from algebra to (probability and) statistics, there is a long and rich history that goes back at least half a century. Thus the utility of Professor McCullagh's approach is not unprecedented, although the extensive application at the level of generality of category theory may be so. The series of contributions to a variety of statistical problems, including topics as diverse as experimental design and card shuffling made by Persi Diaconis, Graham and Kantor (1983), Diaconis (1988) and Rosemary Bailey (1981, 1991), among others, exploiting group invariance and symmetries are probably well known

to the readers here. Notably, as well, the history also includes significant and influential contributions made by Ulf Grenander [e.g., Grenander (1963)] to the study of patterns, and by Harry Furstenberg (1963) to the study of products of random matrices (from which the Lyapunov spectra and related dynamical systems theory parameters are derived). I especially note these latter two mathematicians as each integrally features Lie group and Lie algebra theory throughout their efforts. Thus, disciples of these developments, already facile with the algebraic machinery utilized herein, may be able to make additional contributions atop those developed to date by the author.

I have two stylistic comments. First, the organization is fine for someone relatively familiar with category theory, but for other readers it may be considerably more challenging, as it leads heavily with this theory in Sections 3–5. Without a road map, the interested reader might give up before recognizing the "meat" of the approach. I'd strongly suggest that relatively early on, before diving headlong into Sections 3–5, the reader jump to Section 6 and find an example of particular interest, then go back and forth between the theory sections and the examples. One can then appreciate the formal developments "concretely" from which the abstractions and generalizations become much clearer. One illuminating example for me was the example in 6.2, Exercise 10 of Section 2, regarding inference concerning the correlation coefficient ρ derived from the standard linear regression model with one covariate.

The second (and cosmetic) comment is that I prefer the adjective ad hoc to *absurd* in describing, for example, the models of Section 2. I agree with the substance of the author's point, that there is no hope of a broad or natural theoretic general framework or basis for these models. Nonetheless, such models are of course used by well-intended yet (somewhat) naïve practitioners to produce, for example, figures of merit within specific contexts, and I see no need to embarass the target user.

My major concern is primarily a recognition of a limitation of the algebraic approach. Namely, the (category theory) technology utilized herein cannot be readily extended or adapted to address some spiritually related operational questions that are central to the core of statistical practice. To my sensibilities, this paper is probabilistic (rather than statistical) in spirit, albeit evidently applied to model types and problems classically considered by statisticians. I would probably be happier with the title "What is a parametrized probabilistic (or stochastic process) model?" This is not "bad," as I am a probabilist, but the formalism inferentially bears on processes and distributions, not short and moderate length sequences apart from their definition as possible initial segments from typical realizations of a process. The specific technical issue is that an algebraic approach has a considerable limitation, in the lack of a natural metric. This point is noted by Professor McCullagh in Section 8.3, albeit towards a very different consideration (the Besag–Higdon formulation of yield in agricultural field trials): "Algebra does not easily deal with approximations or inequalities. . . ." The presence of a metric

is not required in the present paper, and the lack thereof within category theory does not obviate any of the formal developments. But suppose we wish to consider which model *form* is better, given one or several (necessarily finite) realizations, among two proposed forms? As David Cox (1990) states, "In particular, choice of an appropriate family of distributions may be the most challenging phase of analysis." Most sensible means to resolve such a question, and there are many such, require a metric, and thus primarily algebraically (particularly category theory) based attempts would likely be artificial and inadequate. I do wish to point out if there is topological as well as algebraic structure in a formal model framework (e.g., a Lie group), then convergence defined by the topology can resolve the metric question in terms of some limiting or asymptotic behavior; thus some probabilistic, that is, infinite sequence comparisons can possibly be resolved within an algebraic framework. But for decidedly nonasymptotic data lengths, and statistical procedures based on them, many formal descriptions will still require nonalgebraic approaches.

Finally, also on a related topic, I would like to mention that for discrete state data, there is a means of assessing the "extent of randomness" or irregularity of a finite or infinite sequence that does not require a model-based (or stochastic process) setting, parametrized or otherwise, in the classical sense [Pincus and Singer (1996); Pincus and Kalman (1997)]. A formalism is developed strictly on a combinatorial basis, again for both finite and infinite length sequences. This could be used, for example, in both discrimination and classification contexts, although this is a separate discussion altogether.

REFERENCE

BAILEY, R. A. (1981). A unified approach to design of experiments. *J. Roy. Statist. Soc. Ser. A* **144** 214–223.

BAILEY, R. A. (1991). Strata for randomized experiments (with discussion). *J. Roy. Statist. Soc. Ser. B* **53** 27–78.

COX, D. R. (1990). Roles of models in statistical analysis. *Statist. Sci.* **5** 169–174.

DIACONIS, P. (1988). *Group Representations in Probability and Statistics*. IMS, Hayward, CA.

DIACONIS, P., GRAHAM, R. L. and KANTOR, W. M. (1983). The mathematics of perfect shuffles. *Adv. in Appl. Math.* **4** 175–196.

FURSTENBURG, H. (1963). Noncommuting random products. *Trans. Amer. Math. Soc.* **108** 377–428.

GRENANDER, U. (1963). *Probabilities on Algebraic Structures*. Wiley, New York.

MCCULLAGH, P. (1999). Quotient spaces and statistical models. *Canad. J. Statist.* **27** 447–456.

MCCULLAGH, P. (2000). Invariance and factorial models (with discussion). *J. Roy. Statist. Soc. Ser. B* **62** 209–256.

PINCUS, S. and KALMAN, R. E. (1997). Not all (possibly) "random" sequences are created equal. *Proc. Nat. Acad. Sci. U.S.A.* **94** 3513–3518.

PINCUS, S. and SINGER, B. H. (1996). Randomness and degrees of irregularity. *Proc. Nat. Acad. Sci. U.S.A.* **93** 2083–2088.

990 MOOSE HILL ROAD
GUILFORD, CONNECTICUT 06437
E-MAIL: stevepincus@alum.mit.edu

DISCUSSION

BY TUE TJUR

Copenhagen Business School

Peter McCullaghs article is an overwhelming cascade of interesting and provoking ideas, related to the vague and intutitive concept of "a model making sense." As stressed by the author, many of the conclusions coming out of this are already part of good statistical tradition. But I agree that it is interesting to investigate to what extent such traditions can be stated as consequences of a mathematically coherent models concept. What I think is missing, so far, is an application that results in a conclusion which is nontrivial or even surprising. But this is probably too much to ask for at the present stage of development.

Some statistical models are canonical. Canonical statistical models include, as a minimum, the linear normal models (regression and analysis of variance), the multiplicative Poisson models and the multinomial models derived from these by conditioning, and the logit linear models for binary data. Very few statisticians would probably dare to disagree here. A question closely related to the present paper is: *Why* are these models canonical? And what is—if anything—a canonical model?

A partial answer to this question is that a canonical model is a model which is constructed from other canonical models in a canonical way. Perhaps this statement is not of much help, but it nevertheless justifies a successful construction which I would like to reconsider here, since the author has only mentioned it in a brief remark (in Section 6.4). I am thinking of the models for ordinal data presented in McCullagh (1980).

Suppose we are dealing with a universe where the natural models for handling of binary responses are the logistic regression models. This could be some socioeconomic research area where peoples' attitudes to various features of brands or service levels are recorded on a binary scale, and the interest lies in the dependence of these attitudes on all sorts of background variables. How do we extend this universe to deal with ordered categorical responses, for example, on three-point positive/indifferent/negative scales? A natural requirement seems to be that if data are dichotomized by the (arbitrary) selection of a cutpoint (putting, for example, negative and indifferent together in a single category), then the marginal model coming out of this is a logistic regression model. This is, after all, just a way of recording a binary response, and even though it would hurt any statistician to throw away information in this way, it is done all the time on more invisible levels. Another natural requirement is that the parameters of interest—with the constant term as an obvious exception—should not depend on how the cutpoint is selected. It is easy to show that these two requirements are met by one and only one class of models for ordered responses, namely the models that can

be described by an underlying continuous linear position parameter model with logistic error distribution, where the responses are grouped by unknown cutpoints. The underlying continuous model can be regarded as a sort of limit of the discrete models as the number of ordered categories grows to infinity. A projective limit, to talk category language.

Another construction in the same spirit, though somewhat less successful from a theoretical point of view, is the construction of overdispersion models from simple generalized linear models. Is there a natural extension of logistic regression models for binomial data or multiplicative Poisson models that can be used in situations where the goodness-of-fit test shows that the model is incorrect, but point estimates of parameters and fitted values are, nevertheless, considered relevant and correct? Can we find a class of statistical models that extends the original generalized linear model by some scale parameter, in such a way that:

1. the original generalized linear model comes out as a special case when the scale parameter is fixed and equal to 1;
2. the maximum likelihood estimates in the model for the original "link-function-linear" parameters coincide with those of the original model (ignoring the overdispersion).

The answer to this question is no. At least for logistic regression it is obvious that no such model exists, because the bounded supports of the response distributions determined by the given binomial totals or indices do not allow for a freely varying scale parameter in the usual sense. Nevertheless, the hypothetical answer to this question is the driving force behind the development of the very useful methods for handling of generalized linear models with overdispersion as indicated by Nelder and Wedderburn (1972) and further developed by McCullagh and Nelder (1989). Thus, we have here the absurd situation that the potentially canonical—but unfortunately nonexisting—answer to a simple and canonical question results in a collection of very useful methods. The overdispersion models exist as perfectly respectable operational objects, but not as mathematical objects. My personal opinion [Tjur (1998)] is that the simplest way of giving these models a concrete interpretation goes via approximation by nonlinear models for normal data and a small adjustment of the usual estimation method for these models. But neither this, nor the concept of quasi-likelihood, answers the fundamental question whether there is a way of modifying the conditions (1) and (2) above in such a way that a meaningful theory of generalized linear models with overdispersion comes out as the unique answer.

It is tempting to ask, in the present context, whether it is a necessity at all that these models "exist" in the usual sense. Is it so, perhaps, that after a century or two people will find this question irrelevant, just as we find old discussions about existence of the number $+\infty$ irrelevant? If this is the case, a new attitude to statistical models is certainly required.

Statistics as practiced today suffers from many similar problems and paradoxes. In survival analysis we have Cox's likelihood, which undoubtably is a canonical thing, but unfortunately not a likelihood in any model that we know of. In basic statistical inference, paradoxes related to sequential situations and Birnbaum's paradox are among the problems we have learned to ignore and live with. This does not have much to do with Peter McCullaghs paper, but it illustrates the need for a clarification of basic concepts. A more restrictive definition of a statistical model than the usual one by arbitrary families of probability distributions on a sample space may turn out to be a necessary ingredient here.

Certain aspects of Peter McCullagh's exposition take us back to the roots, in the sense that many of the ideas that he expresses in terms of category theory are very similar to ingredients of the slightly old-fashioned way of expressing models in terms of populations and samples, rather than probability distributions and random variables. I would like to stress this by repeating a single point he has concerning regression analysis. Just to emphasize that category theory is not a necessary tool here, but only a convenient language, I will try to do it without this tool.

The standard textbook way of specifying a simple regression model is to say that we have observations y_i of independent normal random variables Y_i, $i = 1, \ldots, n$, with the same variance σ^2 and expectations of the form $\alpha + \beta x_i$, where $x_1, \ldots, x_n$ are known quantities. This is undoubtably a useful way of saying it, but it has the drawback that it does not specify the parameters that can be meaningfully estimated. Quantities like the sample size n and the average expectation $\alpha + \beta\bar{x}$ (or, more complicated, the correlation coefficient as suggested by the author) can be estimated without problems in this model, but this is irrelevant for most purposes because these quanties are related not only to the unknown parameters, but also to the design—in this case represented by the choice of the values $x_1, \ldots, x_n$ of the independent variable. A way of stating the model that takes this into account goes as follows. Take as the starting point an infinite population $\{(x_i, y_i)\}$ of x-values and corresponding y-values. Only the x-values are visible from the beginning, and what we do when we perform the experiment is actually to draw—not at random, but rather by some criterion for selection of a representative set of x-values—a sample from this population. Now, define a *parameter function* as a rule which to each such selection $x_1, \ldots, x_n$ of finitely many x-values assigns a parameter in the usual sense, that is, a function of the unknown distribution in the corresponding statistical model. This includes any function of $(\alpha, \beta, \sigma^2)$, and also, so far, expressions like n or $\alpha + \beta\bar{x}$. Here, a technical condition is required, stating that at least two distinct x-values should be present, but we will ignore this irrelevant detail in the following. Now, define a *meaningful* parameter function as a parameter function which is invariant under the formation of marginal models, that is, when a design is reduced by removal of some of the x's, the parameter associated with the marginal distribution should equal the parameter associated with the distribution in the model for the bigger sample. This reduces the possible parameter functions to the set of functions of $(\alpha, \beta, \sigma^2)$. A quantity like $\alpha + \beta\bar{x}$,

where $\bar{x}$ is the average of the x-values in the design we happened to select, is clearly not a meaningful parameter function in this sense. Moreover, this kind of set-up is exactly what is required if one wants a theory that does not allow for such absurdities as those suggested in "exercises" 1, 2, 3, 4 and 5 of Section 2.1.

REFERENCE

McCULLAGH, P. (1980). Regression models for ordinal data (with discussion). *J. Roy. Statist. Soc. Ser. B* **42** 109–142.

McCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd. ed. Chapman and Hall, London.

NELDER, J. A. and WEDDERBURN, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A* **135** 370–384.

TJUR, T. (1998). Nonlinear regression, quasi likelihood, and overdispersion in generalized linear models. *Amer. Statist.* **52** 222–227.

STATISTICS GROUP
COPENHAGEN BUSINESS SCHOOL
SOLBJERG PLADS 3
DK-2000 FREDERIKSBERG
DENMARK
E-MAIL: tt.mes@cbs.dk

REJOINDER

BY PETER McCULLAGH

University of Chicago

> Jourdain: And what are they, these three operations of the mind?
> Philosopher: The first, the second and the third.
> The first to conceive by means of premises;
> the second to reason by means of categories;
> and the third to draw conclusions by means of figures.
>
> Molière. *Le bourgeois gentilhomme.*

Major themes. The discussants cover a wide range of topics. Before responding individually, I will first address a few of the major themes.

Abstraction. The purpose of abstraction is to incorporate into the mathematics those essential relationships that are taken for granted in the domain of application. Yet that is seldom how it is seen from the outside. In my view, necessity is the mother of abstraction, so there can be no virtue in abstraction for its own sake. It follows that this paper should not have been written had I not judged the need to be clear and the proposed solution reasonably compelling. Unless the information

is suitably encoded, individual plots, pigs, varieties, treatments and subjects are, in mathematical terms, simply elements in a set. We can, and we must, do better than that. A neighbor relationship on a graph is a possible step in this direction, but this paper pushes that idea a little further by expecting each object (graph) to be embedded in other objects in such a way that relationships are preserved. Many of the discussants are clearly more comfortable with algebraic abstraction than I am, others evidently less so. To those floating in the abstract wing it may appear that my hesitant steps have not gone far enough. To others anchored in the concrete it may appear that I have fallen into an abyss of mindless nonsense. The tension between the concrete and the abstract is seldom comfortable, but I believe it is healthy and central to the art of statistical modelling. One does not have to look far in probability, statistics or physics to see that today's concrete foundations are yesterday's abstractions. On the other hand, it must ruefully be admitted that most of yesterday's abstractions are buried elsewhere.

Model versus analysis. Statistical models are usually encountered in connection with the statistical analysis of data, and for that reason many statisticians seem to regard the model and the analysis as inextricably intertwined concepts. The situation is quite different in classical applied mathematics, where it is accepted practice to discuss properties of the wave equation or Poisson's equation independently of any data that might (or might not) be collected. While it is difficult to discuss statistical models without also mentioning analysis, it seems to me that the two notions are logically distinct. As I have used the term in this paper, a model is divorced from any particular area of application in much the same way that the heat equation is not necessarily tied to heat transmission. Although I do not recommend the practice, one might talk of the Bradley–Terry model without indicating whether the intended application is to competition experiments or citation studies or transmission disequilibrium in genetics. In that sense, I am interested in what Tjur calls canonical models.

An analysis might well involve the comparison of two or more mutually incompatible models by methods that require numerical approximation of integrals using MCMC. Model checking is an essential component of every analysis, and this might well lead to the consideration of further models. An analysis of financial transaction data might well lead to the conclusion that Brownian motion is not consistent with the data but that an alternative long-tailed model is satisfactory. Thus, in any definition of the term "statistical model," it is essential to bear in mind that models must exist that are *not* compatible with any data likely to be collected. By comparison with models, statistical analysis is a complicated topic, and I have tried in the paper to say as little on the matter as I could hope to get away with. As the title suggests, the topic of the paper is narrow. So far as I can tell, it has nothing to say about computation, design or model checking, and very little to say about model selection, or Bayesian versus frequentist analysis.

Prediction. For purposes of prediction, it is often effective to use one model in one region and a different model in another region. One familiar example is the use of quantum mechanics for atomic-scale calculations, and Newtonian or relativistic mechanics at celestial scales. Despite its success in prediction, every physicist is well aware that this patchwork does not constitute a single model. A black box containing a list of rules that give effective predictions may be very useful, but this is not a model in the sense that I use the term.

Extension and embedding. Self-consistency forces the condition that $P_m\theta$ be the marginal distribution of $P_n\theta$ whenever m is a subobject in n. Here m, n are two sets of units, or more generally two designs, in which m is deemed to be an embedded subdesign in n. The only exceptions I can see to this condition involve systems in which the act of observation perturbs the system in an appreciable manner, as in quantum-mechanical systems and possibly a few social science applications. Julian Besag has hit the pig on the snout with his well-chosen example that a subset of a litter of pigs does not constitute a litter. In the formalism of the paper, this embedding must be coded in the category, so the category includes as objects all litters or sets of litters, but not sublitters or sets of sublitters. This is not a problem, but it must be spelled out explicitly by the morphisms. However, if the pigs were weaned, say 6–12 weeks old, it might be best to take a different view of the matter. The choice of category is essential to all that follows, but category theory itself has nothing to say on the matter. More likely than not, the choice is made on grounds of animal husbandry and statistical design. In principle, it is simply the consequences of this choice that are explored in the paper. However, the choice of category does have a constraining influence on the nature of the inferences that can subsequently be drawn.

Extrapolation. On macroscopic scales the inverse square law for gravitational or electrostatic forces is in good agreement with observation, but when extrapolated to the atomic level it gives predictions in clear violation of everyday experience. I doubt that Newton ever intended that his law should be applied at the angstrom level, but at the same time I do not view it as "reckless to contemplate" such extreme extrapolation. On the contrary, it was an essential step in the early development of quantum mechanics. Julian Besag seems inclined to say that the Newtonian model does not apply to atomic scales. In the sense that it does not work well, he is correct. I am inclined to say that the model does extend to atomic and subatomic scales and that it predicts a catastrophic collapse that does not agree with the observed facts. Provided that one does not confuse model with fact, there is no contradiction in entertaining and comparing models whose consequences may be absurd. To some extent, this difference in terminology is a matter of semantics, but I think there is a little more to it than that. If the inverse square law does not extend to atomic scales, I need to know where it stops and why it stops

there. So far as fertilizer levels in agricultural experiments are concerned, the models are more mundane and more numerous but the principle is the same. The potential to make falsifiable predictions is one of the properties that distinguishes a model from an algorithm.

Approximation. Approximations of various sorts are hard to avoid in statistical work. But an approximation must have a target; otherwise it might as well be exact. Provided there is a target, I see no objection in principle to approximations. Besag makes the point that there exist "schemes" that are perfectly adequate in practice as approximations to models, and I agree that the evidence supports this view. However, I do not agree that a statistical model, as a mathematical object, is necessarily an approximation to anything, any more than a vector space or a process is an approximation. In most applications several competing models vie for supremacy. The hope is that one of these is an adequate description of the observed facts, but the definition does not demand this or guarantee it.

Natural transformation. Natural transformations are more widespread in statistical work than I have indicated in the paper. Consider, for example, transformation of sample space(s) in the usual regression context. The condition of naturality amounts to the statement that transformation $g_n : \mathcal{R}^n \to \mathcal{R}^n$ followed by coordinate projection is the same as coordinate projection followed by transformation $g_{n-1} : \mathcal{R}^{n-1} \to \mathcal{R}^{n-1}$. The conclusion might alarm coordinate-free theorists, but comes as no great surprise to applied statisticians. Each natural transformation is a scalar transformation $g : \mathcal{R} \to \mathcal{R}$ acting componentwise, such as the link function in a generalized linear model. The only restriction is that g should be measurable. Most linear transformations are not natural because the components are not preserved.

For processes, the transformation $W = \Sigma^{-1}$ on covariance matrices is not natural because the inverse of the restriction is not the restriction of the inverse. Within the class of stationary Markov processes, noncommutativity is restricted to the boundary, so the transformation is "almost" natural. In that sense, the sequence of Gaussian Markov random fields with stationary interaction coefficients W_{ij} independent of the lattice size is an approximation to a process. This approximation may be perfectly adequate in practice, particularly if Besag's suggestions for boundary adjustments are incorporated.

Field trials. Julian Besag points out that the sole purpose of a variety trial is the comparison of varieties, not the study of fertility patterns. I agree. Even so, it is hard to see how anything useful can be said unless it is assumed that variety effects are proportional to plot area, or, if defined in ratio form, that variety effects are independent of plot area. Without admitting plots of various sizes the proposition cannot be stated, so it cannot be a consequence of other assumptions, such as additivity or lack of interference, defined for plots of standard

size. One may be inclined to regard the proposition as an matter of established agricultural fact, so much so that it does not count as an assumption. What makes it true, agriculturally speaking, is that agronomy knows something about plots and varieties that mathematics does not. Agronomy knows that variety v planted in a 2×2 plot is equivalent to the same variety planted in each of the individual subplots. On the scale of scientific insights this revelation may achieve a record low, but it not vacuous. First, it admits the existence of larger and smaller plots, and their relation to one another by subset inclusion. Second, it says something substantive about the relationship between varieties and plots that is true for most agricultural treatments but is not true in general. Remember that mathematics knows nothing about anything except mathematics, so mathematics must be instructed in the facts of rural life. In the formalism of the present paper these facts and relationships are encoded in the category and thereby made available explicitly. It seems that everyone does this instinctively, much like Molière's M. Jourdain who was delighted to learn that he had been speaking prose all his life.

Responses to individual discussants. Many of Julian Besag's remarks are concerned with the more complicated subject of statistical strategy, ranging from model construction and statistical analysis to model adequacy and numerical approximation. I do not dispute the importance or the subtlety of statistical strategy and model construction, but I view it as complementary to the topic of the paper. Even if we cannot agree on a definition, or even the need for one, I welcome Besag's remarks concerning the relevance of models in applications. It is also hard to disagree openly with his preference for models that have a physical justification, or what are sometimes called mechanistic or causal models. While this preference is understandable, it must be tempered with the knowledge that certain rather successful models, such as Maxwell's theory of electromagnetism, are hard to justify on such grounds. The universality that I demand of a model is only the universality that follows from the category of embeddings, and one can always opt for a smaller category if this seems sensible. One interpretation of definition [equation (1)] is that it states in an unfamiliar language only what is intuitively obvious. Thus, Besag is entirely correct in his assertion that the commutativity conditions amount to nothing more than common sense. But where I demand internal consistency for a range of plot sizes, it seems that Besag attaches little weight to this. Deep down, though, I'll bet he must! Aside from this point, I aim to show below that the extent of our disagreement is considerably less than I had originally thought.

I agree with essentially all of Peter Bickel's remarks, including the plea for expressing categorical ideas in more acceptable language. In the last section of his remarks, Tue Tjur goes some way in this direction, explaining what is meant by a natural parameter in regression. The connection between natural parameters and identifiable parameters is not an obvious one, so I wonder whether it is a lucky

coincidence that β/σ in the Box–Cox model is both natural and also identifiable in a larger model.

A bushy category is evidence, if any were needed, of a novice gardener at work, one who has acquired a few sharp tools but not yet mastered the principles of effective pruning. Brøns takes the formal definition, dissects it, extends it, reassembles it and reinterprets it with a dexterity I can admire but not emulate. The comma category StatMod is a very natural and familiar statistical object with the right properties and only those properties. The new proposal, to define a statistical model as a functor $\mathcal{C} \to (\mathrm{id}_{\mathrm{set}} \downarrow \mathcal{P})$ from an arbitrary category $\mathcal{C}$ into StatMod, achieves a degree of simplicity and generality through a new level of abstraction. The parameter space and the sample space remain accessible by composition with the natural projections, so nothing has been sacrificed. The pedagogical hurdles remain formidable, but Brøns's construction is right on the mark, and I thank him for it.

One difficulty in constructing a general theory of models is to determine where various conditions belong most naturally. In my view, no theory of linear models can be considered satisfactory unless, to each subrepresentation $\Theta' \subset \Theta$ there corresponds a quotient representation Θ/Θ' in a natural way. This is not a consequence of the extended definition, and is the reason for the additional surjectivity condition that I had included. If this is unnecessary in the general definition, I suspect it must be a part of the definition of linear models.

Peter Huber is right to point out that, in some circumstances at least, what appears to be absurd might well be real. After all, most of quantum mechanics appears absurd and yet we must suppose it to be real. At some stage, one must ask where the category comes from, and the answer here must depend on the application in ways that are not easy to describe. A very sparse category containing many objects but few morphisms is not very useful, but the inclusion of too many morphisms may lead to the exclusion of otherwise useful formulations. I was fascinated to read of Huber's earlier categorization of aspects of statistical theory, which seems to be connected with morphisms on the sample side.

Inge Helland brings a fresh quantum-mechanical perspective to the subject, and particularly to the matter of natural subparameters. I must confess that I was initially skeptical of the restriction to natural subparameters when I first encountered this in Helland (1999a), and I took an opposing view at the time, but I have to agree now that the arguments are convincing. The arguments in Fraser (1968b) are essentially the same. It is instructive to note that the objections raised in the paper to the correlation coefficient in regression do not apply to autocorrelation coefficients in time series or spatial processes.

Kalman has much to say about models and statistical practice, all negative. I'm sure glad he's not pinning the blame entirely on me! By the statement that there does not exist a concrete i.i.d. process in nature, I presume that he means that this mathematical model is not a satisfactory description of observed processes. If so, we need something better, most likely a family of non-i.i.d. processes.

History offers ample support for Kalman's assertion that scientists demand a physical or mechanical explanation, in our case an explanation for the origin of randomness in nature. But this demand does not persuade me that a mechanical explanation should necessarily be forthcoming. The existence of the ether was widely accepted in the nineteenth century, not for its beauty as Kalman claims, but for its supposed mechanical properties which were deemed necessary for the transmission of light. In 1839, the currently accepted mathematical model for the transmission of light was put forward by James MacCullagh (no relation), but rejected by physicists of the day because no mechanical explanation could be provided. Although the date is given as 1842 and the name is misspelled, other details provided by Feynman, Leighton and Sands [(1964) **2**, Section 1–5] are reliable. Two quotes from the 1839 paper suffice to illustrate my point. *Concerning the peculiar constitution of the ether, we know nothing and shall assume nothing except what is involved in the foregoing assumptions* (symmetry!). And later, by way of summary, *the reasoning which has been used to account for the* [*potential*] *function is indirect, and cannot be regarded as sufficient in a mechanical point of view. It is, however, the only kind of reasoning that we are able to employ, as the constitution of the luminiferous ether is entirely unknown.* Indeed it was, and remained so for much of the rest of the century, even after the same equations were resurrected as one leg of Maxwell's theory in 1864. Incidentally, the 1839 paper may be the first to employ the combination of partial derivatives now known as the curl operator, and to show that it transforms as a vector under coordinate rotation in $\mathcal{R}^3$.

Steve Pincus points out that the inferential formalism emphasizes processes and families of processes rather than families of distributions. This is true, but to some extent a matter of presentation. There is no objection in principle to considering a truncated category containing a finite number of objects, such as the category of injective maps on sets of size 12 and smaller. A process that is exchangeable relative to this category is not necessarily extendable to an infinitely exchangeable process, so inferences extending beyond 12 elements are inaccessible. The absence of a metric may look odd, but we must bear in mind that each sample space object must be a measure space with an attached σ-algebra, and possibly additional topological structure that I would rather not think about. One generic example is the category $\mathcal{C}$ in which the objects are finite-dimensional inner product spaces and the morphisms are injective linear maps $\varphi : \mathcal{X} \to \mathcal{X}'$ such that $\langle \varphi x,\ \varphi y\rangle = \langle x,\ y\rangle$. If $\mathcal{S}$ is the opposite, or dual, functor $\mathcal{C} \to \mathcal{C}$, the objects are implicitly assumed to be equipped with the usual Borel σ-algebra, and each φ is sent to the conjugate linear transformation $\varphi^* : \mathcal{X}' \to \mathcal{X}$.

These maps also preserve inner products, but in a different sense such that the induced map $(\ker\varphi^*)^\perp \cong \mathcal{X}'/\ker\varphi^* \to \mathcal{X}$ is an isomorphism of inner product spaces. A process relative to $\mathcal{C}$ is necessarily orthogonally invariant, or spherically symmetric, and is characterized by scale mixtures of normals [Kingman (1972)].

A similar characterization for symmetric and Hermitian matrix-valued processes has recently been obtained by Wichura (2001).

Fraser and Reid ask whether category theory can do more than provide a framework. My experience here is similar to Huber's, namely that category theory is well suited for this purpose but, as a branch of logic, that is all we can expect from it. Regarding the coefficient of variation, I agree that there are applications in which this is a useful and natural parameter or statistic, just as there are (a few) applications in which the correlation coefficient is useful. The groups used in this paper are such that the origin is either fixed or completely arbitrary. In either case there is no room for hedging. In practice, things are rarely so clear cut. In order to justify the coefficient of variation, it seems to me that the applications must be such that the scale of measurement has a reasonably well-defined origin relevant to the problem.

The Cauchy model with the real fractional linear group was originally used as an example to highlight certain inferential problems. I do not believe I have encountered an application in which it would be easy to make a convincing case for the relevance of this group. Nevertheless, I think it is helpful to study such examples for the light they may shed on foundational matters. The fact that the median is not a natural subparameter is an insight that casts serious doubt on the relevance of the group in "conventional" applications. To turn the argument around, the fact that the Cauchy model is closed under real fractional linear transformation is not, in itself, an adequate reason to choose that group as the base category. In that sense, I agree with a primary thesis of Fraser's *Structure of Inference* that the group supersedes the probability model.

Tjur's remarks capture the spirit of what I am attempting to do. In the cumulative logit model, it is clear intuitively what is meant by the statement that the parameter of interest should not depend on how the cutpoints are selected. As is often the case, what is intuitively clear is not so easy to express in mathematical terms. It does not mean that the maximum-likelihood estimate is unaffected by this choice. For that reason, although Tjur's second condition on overdispersion models has a certain appeal, I do not think it carries the same force as the first. His description of natural subparameters in regression is a model of clarity.

Conformal invariance and Markov random fields. Tue Tjur remarks that it would be more convincing if I could produce an example of a new model using category ideas. I agree, but this is not easy. It was in the hope of coming up with something different that I sought to explore the consequences for spatial models of using the category of conformal maps rather than the more conventional group of rigid motions. Section 8 of the paper is a bit of a disappointment, at least if judged by the suitability of the models for their intended application. Nevertheless, it may yet be worth pursuing the same line of argument using the subcategory of *invertible* conformal maps on planar domains. For simplicity of exposition, I take $\mathcal{D} = \mathcal{D}'$ to be the Riemann sphere, or the extended complex plane. This category

is a group in which each map is a fractional linear transformation of the form $\varphi : z \mapsto (az + b)/(cz + d)$ and inverse $w \mapsto (dw - b)/(a - cw)$, with complex coefficients a, b, c, d such that $ad - bc \neq 0$.

By the symbol $Y \sim \mathcal{W}_\sigma$ is meant the planar Gaussian process with covariance function $-\sigma^2 \log|z - z'|$. The principal departure from Section 8, is that Y is a function, not a measure. For the moment I pretend that it is defined pointwise, and the group acts by composition sending Y to $Y' = Y \circ \varphi$. The process is such that pointwise differences $Y(z_1) - Y(z_2)$ are zero-mean Gaussian with covariances

$$\begin{aligned}\operatorname{cov}\big(Y(z_1) - Y(z_2), Y(z_3) - Y(z_4)\big) &= \sigma^2 \log\left|\frac{(z_1 - z_4)(z_2 - z_3)}{(z_1 - z_3)(z_2 - z_4)}\right| \\ &= \sigma^2 \log\left|\frac{(\varphi z_1 - \varphi z_4)(\varphi z_2 - \varphi z_3)}{(\varphi z_1 - \varphi z_3)(\varphi z_2 - \varphi z_4)}\right|\end{aligned}$$

for each set of distinct points z_1, z_2, z_3, z_4. The second line, which follows from the invariance of the cross-ratio under fractional linear transformation, shows that the transformed process also satisfies $Y' \sim \mathcal{W}_\sigma$. The fact that variances are infinite is not a problem in practice because of regularization. Each regularized observation is an integrated contrast, $Y(\rho) = \int Y(z)\, d\rho$, in which $\rho = \rho^+ - \rho^-$ is a signed measure such that $\rho(\mathcal{D}) = 0$. Technically, Y is defined as a random linear functional on the L_2-space of contrast measures such that

$$|\rho|_w^2 = -\int_{\mathcal{D}^2} \log|z - w|\, d\rho(z)\, d\rho(w) < \infty,$$

whereas white noise is defined on the space such that

$$|\rho|^2 = \int_{\mathcal{D}^2} \delta(z, w)\, d\rho(z)\, d\rho(w) < \infty,$$

where $\delta(z, w)$ is Dirac's delta. Both are defined on the intersection, as are convolutions and mixtures. The group element φ sends ρ to $\rho\varphi^{-1}$, and Y to Y' such that $Y'(\rho) = Y(\rho\varphi^{-1})$, with the same distribution $Y \sim Y' \sim \mathcal{W}_\sigma$. This argument shows that logarithmic variograms are preserved under invertible conformal maps.

Both white noise and $\mathcal{W}_\sigma$ are Markov random fields in the sense that, for each closed contour, values in the interior and exterior are conditionally independent given the values on the contour (Matheron, 1971). Both processes are also conformal, but the similarity ends there. The set of conformal processes is also closed under addition of independent processes. Thus, the sum of white noise and $\mathcal{W}_\sigma$ is conformal but not Markov. Beyond convolutions of white noise and $\mathcal{W}_\sigma$, it appears most unlikely that there exists another conformal process with Gaussian increments. Whittle's (1954) family of stationary Gaussian processes has the Markov property [Chilès and Delfiner (1999)] but the family is not closed under conformal maps nor under convolution.

Ignoring variety effects, the Besag–Higdon model is a sum of white noise and a Markov random field on the lattice. Besag's remark that the variogram exhibits logarithmic behavior suggests that the fitted MRF is close to $\mathcal{W}_\sigma$ for some σ. This is certainly plausible since $\mathcal{W}_\sigma$ is Markov and can be well approximated by a low-order lattice scheme, in which the MRF coefficients in the interior are an approximation to the Laplacian. If this speculation is confirmed in applications, the fitted model is approximately a sum of white noise and $\mathcal{W}_\sigma$, which is in fact a conformal process. In principle, the relation between the MRF coefficients and σ (in $\mathcal{W}_\sigma$) can be determined. Thus, if the fitted MRF is close to $\mathcal{W}_\sigma$, or even to a linear deformation of $\mathcal{W}_\sigma$, the fitted Besag–Higdon process has a natural extension beyond the lattice to plots of arbitrary size and shape, in which case my criticism on these grounds does not apply.

At the time I wrote Section 8, I was aware of certain peculiarities of the de Wijs process connected with its interpretation as the Newtonian potential, or inverse of the Laplace operator. But I did not fully appreciate the Markov property, and I was unaware of its conformal invariance. So this revelation comes as a pleasant surprise, and helps me to understand better the connection with MRF lattice schemes in the analysis of field trials. With twenty years of first-hand experience in farming I thought I understood the meaning of fertility, but I knew of no plausible argument to suggest that fertility processes should be spatially Markov, and I failed to see why this proposition should be so readily accepted. A wide range of arguments suggested that fertility might be a compound process or sum of elementary processes, indicating that a model, as a set of processes, should be closed under convolution. From the present viewpoint of conformal transformation, this mechanistic reasoning is replaced by an argument of an entirely different sort in which the entire concept of fertility is absent. The set of conformal Gaussian processes is a convex cone of dimension two generated by white noise and $\mathcal{W}_\sigma$, in essence the random effects model that Besag suggests. Although the Besag–Higdon scheme is not closed under convolution, all of the conformal processes are already included at least as lattice approximations. It would be good to have this important subset explicitly labelled. This analysis suggests that two very different lines of argument may, in practice, lead to very similar conclusions.

The application of conformal transformations to agricultural processes in muddy fields may seem frivolous to the point of absurdity, but that is not how I see it. The driving thesis is that, whatever the generating process, it should look very much the same after any transformation that preserves the relevant aspects of the geometry. Conformal maps preserve Euclidean geometry locally with no angular distortion, which is the only argument for their relevance. However, Euclidean geometry is only the visible side of a corn field. Specific nonisotropic effects connected with topography, ploughing, harvesting and drainage are inevitable and may even dominate the nonspecific variations. To the extent possible, these are included in the category, and are carried along with the conformal maps. Even

with this adjustment, we should not delude ourselves by pretending that we have learned anything about the laws of biology or agricultural science by studying the axioms of algebra. To learn the laws of agriculture, we must study agricultural data, but it undoubtedly helps to do so in the light of potential models. The observed logarithmic behavior of variograms suggests that some field processes have a strong $\mathcal{W}_\sigma$ component, consistent with conformal processes, but we should not be surprised to find additional components that are not conformal.

Acknowledgments. I want to thank John Marden for arranging this discussion paper and also the discussants for their remarks. My colleagues Michael Stein, Steve Lalley, Norman Lebovitz, Steve Stigler and Sydney Webster have given advice freely on a range of points. But I am especially indebted to Julian Besag for his generous help and forthright remarks, which have helped to clarify my thinking on spatial processes and Markov random fields.

REFERENCE

CHILÈS, J.-P. and DELFINER, P. (1999). *Geostatistics*. Wiley, New York.

FEYNMAN, R. P., LEIGHTON, R. B. and SANDS, M. (1964). *The Feynman Lectures on Physics*. Addison-Wesley, Reading, MA.

FRASER, D. A. S. (1968b). *The Structure of Inference*. Wiley, New York.

HELLAND, I. S. (1999a). Quantum mechanics from symmetry and statistical modelling. *Internat. J. Theoret. Phys.* **38** 1851–1881.

KINGMAN, J. F. C. (1972). On random sequences with spherical symmetry. *Biometrika* **59** 492–494.

MACCULLAGH, J. (1839). An essay towards the dynamical theory of crystalline reflexion and refraction. *Trans. Roy. Irish Academy* **21** 17–50.

MATHERON, G. (1971). The theory of regionalized variables and its applications. *Cahiers du Centre de Morphologie Mathématique de Fontainbleu* **5**.

WHITTLE, P. (1954). On stationary processes in the plane. *Biometrika* **41** 434–449.

WICHURA, M. (2001). Some de Finetti type theorems. Preprint.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
5734 UNIVERSITY AVENUE
CHICAGO, ILLINOIS 60637-1514
E-MAIL: pmcc@galton.uchicago.edu