The Annals of Mathematical Statistics
1971, Vol. 42, No. 1, 141-146

RANDOMIZED MODELS AND THE DILUTION AND BIOASSAY PROBLEMS

BY D. A. S. FRASER AND R. L. PRENTICE

0. Summary. Randomization has been proposed [1] for the dilution-series problem with the purpose of producing a continuous random variable; the randomization is performed by selecting a value from the uniform distribution (0, 1]. This paper develops a structural model appropriate to a range of problems involving such randomization as part of the design. The model applies to the dilution series problem and to the bioassay problem, and produces a posterior distribution for the primary parameter and a marginal likelihood for other parameters.

1. Introduction. The concentration of organisms in a given solution is commonly assayed by taking several samples from the given solution, several from a solution obtained by 10^1-fold dilution, several from a solution obtained by 10^2-fold dilution, etc., and testing the samples for fertility by incubation with nutrient. As a variant Fisher [1] suggested a preliminary dilution by the factor 10^v where v is uniform (0, 1] and then samples and dilutions as just described. His purpose was to obtain a continuous variable. In fact his variant produces much more, it generates what randomization in design is intended to generate—objective error.

Under a process of continuous dilution there is a point at which the derived solution has a standard concentration of one organism per unit volume, or *zero* on a log scale. The identification on the dilution scale of this zero point in fact determines the concentration of the given solution. The dilution assay is then one of standardization—finding the zero point on a scale: the degree of dilution provides a difference scale and the assay is to determine the zero point.

This paper develops a structural model that embraces such problems. The model has a controllable variable that provides a difference scale for an input variable; and the input variable determines the distribution of a response variable, possibly with a nuisance parameter. The central inference problem is to determine the zero point of the input variable as located on the scale of the controllable variable. The design is discussed in Section 2, its feasibility in Section 3, and the analysis in Section 4. The results are applied to the dilution series problem in Section 5 and to the bioassay problem in Section 6.

2. The design. Let w be a real valued *input variable* controllable in part by an experimenter. Let e, with values in $\mathscr{Y}$, be a *response variable* that can be observed for any input w; let $f(e\colon w, \tau)$ be the density function for e relative to a measure μ on the Borel sets of a σ-compact metrizable topology on $\mathscr{Y}$; assume that f is continuous in (e, w).

Now suppose the input variable is available only on a difference scale and that the purpose of the experimenter is the standardization of this scale. Let x be the

Received February 25, 1970.

available *controllable variable* and let θ be the value (for this variable) at which the zero point of the input variable occurs. Then $x = \theta + w$.

The experimenter could choose values $x_1, \cdots, x_r$ for the controllable variable and observe the corresponding response values $y_1, \cdots, y_r$; these could be presented in a matrix,

$$\begin{bmatrix} x_1 \cdots x_r \\ y_1 \cdots y_r \end{bmatrix}.$$

He would be concerned with making inferences concerning θ and τ.

For theoretical purposes it is convenient to use an open ended sequence of x values. In practice the extreme values will have a predictable response and only a finite number of x values will be needed; this will be made precise in Section 3. It is also convenient to use symmetric sequences having a constant difference between adjacent values.

A *design sequence* (h) is a sequence $(\cdots, v-h, v, v+h, \cdots)$ of values for the controllable variable x; let v be the value in $(0, h]$. An *input sequence* is a sequence $(\cdots, u-h, u, u+h, \cdots)$ of values for the input variable w; let u be the value in $(0, h]$.

A design sequence $(\cdots, v-h, v, v+h, \cdots)$ produces an input sequence $(\cdots, v-h-\theta, v-\theta, \cdots)$; the input sequence, however, is unknown in location $(\text{mod}\, h)$ on its scale.

A *random design sequence* (h) is a design sequence (h) where v has a distribution on $(0, h]$. A *random input sequence* (h) is an input sequence (h) where u has a distribution on $(0, h]$.

A random design sequence with given distribution produces a random input sequence. The location distribution of this input sequence is typically unknown:

LEMMA. *A random design sequence with given distribution produces a random input sequence with known distribution if and only if the distribution of v is uniform.*

PROOF. The location u is the residue of $v-\theta$ modulo h. If the location distribution is known it is invariant of θ. As a distribution on the circle it must be invariant under rotation and hence must be uniform.

With a uniform distribution for v and hence for u the sequences will be called *randomized.* A randomized design sequence produces a randomized input sequence; the randomization produces a *known* distribution for input values.

The randomized input sequence and the corresponding responses can be presented in a matrix,

$$E = \begin{bmatrix} \cdots & u-h & u & u+h & \cdots \\ \cdots & e(u-h) & e(u) & e(u+h) & \cdots \end{bmatrix}.$$

The matrix E describes the experimental situation. Its distribution is known or known except for the nuisance parameter τ: u is uniform on $(0, h]$; the $e(w)$ are statistically independent with distribution $f(e\colon w, \tau)$.

The experimenter, however, has not fully identified the input variable; he records response values associated with values of the controllable variable x,

$$Y = \begin{bmatrix} \cdots & v-h & v & v+h & \cdots \\ \cdots & y(v-h) & y(v) & y(v+h) & \cdots \end{bmatrix}$$

$$= \begin{bmatrix} \cdots & u-h+\theta & u+\theta & u+h+\theta & \cdots \\ \cdots & e(u-h) & e(u) & e(u+h) & \cdots \end{bmatrix}$$

$$= [\theta] \begin{bmatrix} \cdots & u-h & u & u+h & \cdots \\ \cdots & e(u-h) & e(u) & e(u+h) & \cdots \end{bmatrix}$$

where $[\theta]$ applied to a matrix adds θ to each element in the first row.

The performance in the experimental situation is known or known except for a nuisance parameter τ. The experimenter, however, obtains his response values in association with θ *translates of the actual input values.*

3. Feasibility. The design sequences described in the preceding section require observations on a countable number of response variables. In contrast the usual approach to the applications to be examined involves observations on three to eight response variables. Consider now some truncation assumptions that effectively replace the countable number of variables by a finite number.

An input sequence (h) is *feasible* for the response model $f(e\colon w, \tau)$ if there are response atoms a_i ($i = 1, 2, a_1 \neq a_2$) such that $f(a_i\colon (-1)^i w, \tau)\mu(a_i)$ is monotone in w for w greater than some $W(\tau)$ and if

$$P_i((-1)^i w, h, \tau) = \textstyle\prod_{j=1}^{\infty} (f(a_i\colon (-1)^i(w+jh), \tau)\mu(a_i))$$

has limit 1 as $w \to \infty$. If an input sequence (h) is feasible then any input sequence is feasible: the monotonicity allows some power of $P_i((-1)^i w, h, \tau)$ with rearranged factors to dominate frpm below another P_i being checked.

Feasibility in effect says that responses from inputs outside a sufficiently large range are predictable; or more precisely, the probability for the predicted values outside a finite range approaches 1 as the end points are taken to $-\infty$ and $+\infty$.

Consider a randomized input sequence (h) and corresponding responses from $f(e\colon w, \tau)$. Assume that an input sequence (h) is feasible.

The sample space for u is $(0, h]$. The density function is h^{-1}.

For the sample space for the responses, let $e_j = e(u+jh)$ for u in $(0, h]$, and let $j_1 = \min\{j\colon e_j \neq a_1\}$, $j_2 = \max\{j\colon e_j \neq a_2\}$ where these exist. And let the space

$$S_{j_1 j_2} = (\mathcal{Y} - a_1) \times \mathcal{Y}^{j_2 - j_1 - 1} \times (\mathcal{Y} - a_2)$$

for $(e_{j_1}, \cdots, e_j, \cdots, e_{j_2})$ have the density $f(e\colon u+jh, \tau)$ re μ applied coordinate by coordinate and adjusted by the factor $P_1(u+j_1 h, h, \tau) P_2(u+j_2 h, h, \tau)$; let

$$S_M = \textstyle\bigcup_{-M \leqq j_1 < j_2 \leqq M} S_{j_1 j_2}$$

have the measure as defined on the components, and let $S = \lim S_M$. Feasibility ensures that the probability content of S_M approaches 1 as $M \to \infty$. In effect the response sequences with a_1's in the limit on the left and a_2's in the limit on the right are replaced by the truncated sequence and the probability for the truncated sequences embraces the probability for the omitted tails. The effective sample space is the union of the spaces for such truncated components.

The probability differential for the input sequence and the responses can then be presented as

$$f(E:\tau)\,dE = \prod_{-\infty}^{\infty}(f(e(u+jh):u+jh,\tau)\mu(e(u+jh)))h^{-1}\,du.$$

4. The analysis. A randomized design sequence generates a randomized input sequence. The randomized input sequence produces corresponding response values.

The distribution of the input sequence and corresponding responses is $f(E:\tau)\,dE$ as given in Section 3. For an appropriate application this describes the variables in the experimental situation.

The experimenter, however, has not fully identified the input variable and as a result he obtains Y. The outcome Y, however, derives from E in the experiment by a θ translation of the first row: $Y = [\theta]E$.

This model is a structural model with additional quantity τ (Fraser [3]).

As a translation variable let $r(Y) = \min\{v+jh\colon y(v+jh) \neq a_2\}$. Then

$$D(Y) = [r(Y)]^{-1}Y = \begin{bmatrix} \cdots & -h & 0 & h & \cdots \\ \cdots & d_{-h} & d_0 & d_h & \cdots \end{bmatrix}$$

is a reference point for the translation orbit (note $d_{jh} = a_1$ for all negative j).

The equation $Y = [\theta]E$ can be separated into between-orbit and within-orbit components:

$$D(E) = D(Y) \qquad r(E) = -\theta + r(Y).$$

The first equation shows that the characteristic $D(E)$ in the experimental situation can be identified; the second equation shows that characteristic $r(E)$ is totally inaccessible due to the absence of information concerning the quantity θ. The distribution that describes the inaccessible $r(E)$ is then the conditional distribution of the *variable* $r = r(E)$ given $D(E) = D(Y) = D$ obtained from the distribution in Section 3; it is

$$g(r:D,\tau)\,dr = k_\tau(D)\prod_{-\infty}^{\infty}f(d_{jh}:r+jh,\tau)\,dr$$

where

$$k_\tau^{-1}(D) = \int_{\infty}^{\infty}\prod_{-\infty}^{\infty}f(d_{jh}:t+jh,\tau)\,dt$$

exists (feasibility assumption) for all D having $d_{jh} = a_1$ for negative j and $= a_2$ for all j larger than some j_2.

The distribution $g(r\colon D,\tau)\,dr$ describes possible values for the inaccessible location $r(E)$. Each possible value for that $r(E)$ predicates a corresponding value for θ: $\theta = r(Y) - r(E)$. The predicated distribution describing the inaccessible θ is

$$g(r(Y)-\theta:D,\tau)\,d\theta.$$

Preliminary information concerning τ would be obtained from the marginal likelihood function

$$L(D:\tau) = R^{+}(D)k_{\tau}^{-1}(D(Y)).$$

5. The dilution series problem. A familiar applied problem involves a given solution containing living organisms of a single type distributed independently and at random through the solution; inferences are needed concerning the concentration of the living organism. Samples of a given size from the solution can be tested for fertility by incubation with nutrient; in a typical case these would all be fertile. A derived solution can be obtained by extreme dilution and samples of a given size from this derived solution can be tested for fertility; in a typical case these would all be sterile. The dilution series method is to take a series of dilutions (by factors of 1, 10, 100, $\cdots$ say) and to test samples at each dilution level hoping thereby to comfortably bracket the informative range where some samples are fertile and some are sterile. Fisher [1] has suggested a preliminary dilution by a factor 10^{v} where v is uniform $(0, 1]$. This preliminary dilution generates what would be a randomized design sequence (1) in Section 2.

Consider an arbitrary solution and a derived solution obtained by dilution. Let the dilution *factor* X be the final volume divided by the volume subject to dilution. And let the dilution *dosage* $x = \log X$ be the dilution factor reexpressed in log units.

Now consider a solution with an average of one organism per unit volume; call this a solution of *dilution strength* 0. Now consider a derived solution obtained by dilution dosage w; call this a solution of dilution *strength* w.

Now consider a given solution whose concentration is to be assayed. Let λ be the average number of organisms per unit volume. And let $\theta = \log \lambda$. A dilution dosage θ applied to this given solution produces a solution of strength 0; accordingly the dilution strength of the given solution is $-\theta$.

For a solution of strength w the average number of particles per unit volume is 10^{-w}. The probability that a random sample of unit volume is sterile is

$$P(w) = \exp\{-10^{-w}\} = 1-Q(w).$$

It is of interest that $P(w)$ can be viewed as a distribution function: Consider a unit volume in a solution subject to continuous and instantaneous dilution; let W be the strength of the solution at the instant that the unit volume becomes sterile; then $\Pr(W \leqq w) = P(w)$.

Let v be uniform $(0, h]$ and consider dilution factors 10^{v}, 10^{v+h}, 10^{v+2h}, $\cdots$. Except for missing negative dosages this is a randomized design sequence (h). Consider the application of this design to the given solution.

Suppose that n independent samples of unit volume are chosen from each derived solution and that the number e of sterile samples is observed. Then $\mathscr{Y} = \{0, 1, \cdots, n\}$.

$$f(e:w)\mu(e) = \binom{n}{e}P^{e}(w)Q^{n-e}(w),$$

and the feasibility follows easily with $a_1 = 0$, $a_2 = n$.

Let $y(v+jh)$ be the number of sterile samples at dilution 10^{v+jh}. Substitution in the expressions in Section 4 gives the distribution describing the error location $r(E)$ and the distribution describing the log-concentration θ. There is no nuisance parameter and hence no need for marginal likelihood.

6. The bioassay problem. In the bioassay problem a series of dosage levels is chosen for a drug under investigation. A sample of animals for each level is administered the corresponding dosage. The number of reactions (often "death") is observed. The purpose is one of standardization: finding the dosage level that produces a 50% reaction rate in animals, the LD50 (lethal dosage fifty percent).

An administration of drug is said to have administration *factor* X relative to some initial quantity of drug if the administration quantity is X times the initial amount, and is said to have an administration *dosage* x where $x = \log X$.

Now consider an LD50 administration of drug; call this an administration of *strength* 0. Now consider a derived administration obtained by an administration dosage w relative to the LD50 administration; call this an administration of strength w.

Now consider a reference quantity of the drug. Let λ be the quantity in reference units that produces the 50% reaction rate; and let $\theta = \log \lambda$. An administration θ relative to the reference quantity then produces an administration of strength 0; the strength of the reference amount is then $-\theta$.

For an administration strength w the probability of reaction is commonly represented by a normal or logistic distribution function $P(w\colon \tau) = 1 - Q(w\colon \tau)$:

$$P_N(w\colon\tau) = \int_{-\infty}^{\infty} (2\pi\tau^2)^{-\frac{1}{2}} \exp\{-w^2/2\tau^2\}\, dw,$$

$$P_L(w\colon\tau) = 1/(1+\exp\{-\tau w\}).$$

It is again of interest that $P(w\colon \tau)$ can be viewed as the distribution function of a variable W: an animal is chosen at random and subject to an administration strength that is increased continuously (with instantaneous effect) from $-\infty$ to $+\infty$; let W be the strength at which reaction occurs; then $\Pr(W \leqq w) = P(w\colon \tau)$.

Let v be uniform $(0, h]$ and consider administration factors 10^v, 10^{v+h}, $10^{n+2h}, \cdots$ relative to the reference quantity. Except for missing negative dosages this is a randomized design sequence (h).

Now suppose n animals are tested at each chosen dosage level and let e be the number of reactions. Then as with the dilution problem $\mathscr{Y} = \{0, 1, \cdots, n\}$ and

$$f(e\colon w, \tau)\mu(e) = \binom{n}{e} P^e(w\colon\tau) Q^{n-e}(w\colon\tau).$$

And the feasibility follows easily with $a_1 = 0$, $a_2 = n$.

Let $Y(v+jh)$ be the number of reaçtions at dosage $v+jh$. Substitution in the expressions in Section 4 gives the distribution describing the error location $r(E)$, the distribution describing the LD50 θ, and the marginal likelihood function for τ.

REFERENCES

[1] Fisher, R. A. (1935). The logic of inductive inference. *J. Roy. Statist. Soc.* **98** 39–54. Also Paper 26, Fisher [2].

[2] Fisher, R. A. (1950). *Contributions to Mathematical Statistics*. Wiley, New York.

[3] Fraser, D. A. S. (1968). *The Structure of Inference*. Wiley, New York.