The Annals of Mathematical Statistics
1968, Vol. 39, No. 2, 561-566

SMOOTHED ESTIMATES FOR MULTINOMIAL CELL PROBABILITIES[1]

BY JAMES M. DICKEY[2]

Yale University

1. Summary. Whittle's (1957), (1958) smoothed probability-mass-function estimate, determined by a Bayesian criterion, depends on a choice of the mean and covariance structure of a prior distribution on the unknown probabilities. A data-analytic choice is proposed (Equations (7) and (8)), based on natural stationarity assumptions (Equations (4)). Adaptations and modifications for multidimensional cells are given (Section 4) and some alternatives are mentioned (Section 5).

2. Preliminaries. Denote by X a generic cell of a sampled multinomial distribution with unknown cell probabilities p_X. Although finite, the range of X is here considered to be large, a subset of a finite dimensional real normed space, the cells X and Y being "close" if $\|X - Y\|$ is small.

Consider the problem of estimating p_X from the cell counts $n_x(\sum n_x = n.$, $n.$ fixed, say). The naive cell-frequency estimate $\hat{p}_X = n_X/n.$, may be too "rough," in the sense of a prior prejudice that p_X approximately equals p_{X+Z} for small increments Z in X. For example, the cells X may be located at the midpoints of grouping intervals, with the probabilities p_X integrals of an underlying smooth density function. A strict Bayesian estimate, $\hat{p}_X = (n_X + \alpha_X)/(n. + \alpha.)$, the posterior mean resulting from a Dirichlet prior distribution on the p_X's with parameters α_X (Wilks (1962)), is little better, merely expressing a prior prejudice against extreme values of p_X and giving no recognition to proximity. Unfortunately, the Dirichlet and closely related distributions (Carlson (1963)) are the only known distributions on the simplex of probabilities p_X for which the posterior mean, a ratio of high-order mixed moments, can be computed in practice.

Whittle (1957), (1958) has proposed the smoothed estimate,

$$\hat{p}_X = \sum_Y w_X(Y) n_Y/n., \tag{1}$$

based on weights $w_X(Y)$ chosen to minimize the prior expectation of the squared error $E(\hat{p}_X - p_X)^2$, where E indicates overall expectation under a prior distribution on the vector of unknown probabilities p_X, and where the weights $w_X(Y)$ are allowed to depend on $n.$. Whittle's criterion leads immediately to his system of linear equations for $w_X(Y)$,

$$\sum_Z (E(n_Y/n.)n_Z/n.)w_X(Z) = Ep_X p_Y, \tag{2}$$

Received 12 June 1967; revised 23 August 1967.

[1] This research was supported by the Army, Navy, Air Force, and NASA under a contract administered by the Office of Naval Research.

[2] Presently at the University of Southern California Medical School.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Mathematical Statistics.* www.jstor.org

where, according to an easy calculation,

$$E(n_Y/n.)n_Z/n. = (1 - n.^{-1})Ep_Yp_Z + n.^{-1}\delta_{Y,Z}Ep_Y, \tag{3}$$

$\delta_{Y,Z} = 1$, if $Y = Z$, and 0, if $Y \neq Z$.

Thus, the optimal weights $w_X(Y)$ are determined by the sample size and the first and second order moments of the prior distribution of the cell probabilities. For increasingly large samples, the optimal smoothed estimate $\hat{p}_X$ is easily seen by (2) and (3) to approach the appropriate cell frequency: as $n. \to \infty$, $w_X(Y) \to \delta_{X,Y}$. By straightforward matrix manipulations, $w_X(Y)Ep_X = w_Y(X)Ep_Y$. Equations (2) and (3) combined in various ways lead to Whittle's expressions for the optimal mean squared error,

$$\begin{aligned}\min_w E(\hat{p}_X - p_X)^2 &= Ep_X^2 - \textstyle\sum_Y w_X(Y)Ep_Xp_Y \\ &= n.^{-1}[w_X(X)Ep_X - \textstyle\sum_Y w_X(Y)Ep_Xp_Y] \\ &= (n. - 1)^{-1}[w_X(X)Ep_X - Ep_X^2].\end{aligned}$$

Expressions for the sampling and overall bias and variance are also easily obtained.

Whittle appears not to have asked whether his estimates $\hat{p}_X$ are mathematically probabilities, that is, are nonnegative and sum to unity. Leonard J. Savage has pointed out privately that $\sum \hat{p}_X \equiv 1$ (equivalently, $\sum_X w_X(Y) \equiv 1$). Since, if not, for the other such estimates, $\tilde{p}_X = \hat{p}_X - \bar{\hat{p}}. + 1/C$ (equivalently, $\tilde{w}_X(Y) = w_X(Y) - \bar{w}.(Y) + 1/C$), where C denotes the total number of cells,

$$\sum (\tilde{p}_X - p_X)^2 = \sum (\hat{p}_X - p_X)^2 - C(\bar{\hat{p}}. - 1/C)^2,$$

for every realization. Minimization of $E \sum (\tilde{p}_X - p_X)^2$ is equivalent to minimization of each $E(\tilde{p}_X - p_X)^2$.

The weights $w_X(Y)$, and hence the estimates $\hat{p}_X$, are not necessarily nonnegative. For example, with $C = 4$, if (u_0, u_1, u_2, u_3) is Dirichlet distributed with parameters identically $\frac{1}{4}$, and $p_X = \sum_Y a_{[X-Y]}u_Y$ with $(a_0, a_1, a_2, a_3) = (\frac{1}{2}, \frac{1}{4}, 0, \frac{1}{4})$ (the index $[X - Y]$ of $a_{[X-Y]}$ interpreted modulo 4), and if $n. = 17$, then the matrix of weights is circulant with first row $(29, 12, -5, 12)/48$. In this example, negative weights arise for $n.$ large (such as 17), for which cases, negative estimates $\hat{p}_X$ are rare. In the event of negative estimates $\hat{p}_X$, replacement of the offending values by zero and renormalization by division is recommended.

Whittle's criterion has met with less than universal acceptance in practice, partly because of the necessity of choosing whole functions Ep_Yp_Z and Ep_Y of Y and Z. The following weak-stationarity conditions, submitted as a realistic approximate description of many prior opinions, serve to meet this objection.

Assume for all X, Y, Z,

$$Ep_X = Ep_Y \equiv C^{-1}, \tag{4a}$$

$$Ep_Xp_Y = Ep_{X+Z}p_{Y+Z} = f(X - Y) = f(Y - X). \tag{4b}$$

This translation invariance of the first two moments is impossible without in-

variance under cyclic translations, that is, periodicity. For example, if one-dimensional $X = 0, 1, 2, \cdots, C-1$, then $Ep_X p_Y = Ep_{[X+Z]}\, p_{[Y+Z]}$, where $[X]$ is X modulo C, and hence $Ep_X p_Y = f(|X - Y|) = f(|X - Y \pm C|)$, the entries of a symmetric circulant matrix. As a proof, note that for X of any dimension,

$$\sum_Y Ep_X p_Y = Ep_X = C^{-1}, \tag{5}$$

constant in X, implying, for X and $X + Z$ in the body B of cells considered,

$$\sum_{Y, Y\varepsilon B} Ep_X p_Y = \sum_{Y, Y+Z\varepsilon B} Ep_{X+Z} p_{Y+Z},$$

the left- and right-hand members of which equal, respectively,

$$(\sum_{Y, Y\varepsilon B, Y+Z\varepsilon B} + \sum_{Y, Y\varepsilon B, Y+Z\notin B}) Ep_X p_Y$$

and

$$(\sum_{Y, Y+Z\varepsilon B, Y\varepsilon B} + \sum_{Y, Y+Z\varepsilon B, Y\notin B}) Ep_{X+Z} p_{Y+Z};$$

then by (4b),

$$Ep_X p(B_Z^-) = Ep_{X+Z} p(B_Z^+),$$

where

$$B_Z^- = \{Y\colon Y \varepsilon B,\ Y + Z \notin B\},$$
$$B_Z^+ = \{Y + Z\colon Y \notin B,\ Y + Z \varepsilon B\}.$$

For C large, these end effects are negligible, except to estimates $\hat{p}_X$ for cells X near the boundary of B (which could be much of B in high dimensions).

By Equations (3) and (2), the functions $En_X n_Y/n.^2$ and $w_X(Y)$ have similar invariance and periodicity properties,

$$E(n_X/n.)n_Y/n. = g(X - Y) = g(Y - X), \tag{6}$$

and

$$w_X(Y) = h(X - Y) = h(Y - X).$$

The known closed-form inverse of a circulant matrix is of interest for the one-dimensional case. If A has first row $(a_0, a_1, \cdots, a_{C-1})$, then the X, Yth entry of A^{-1} is $C^{-1}\sum_{Z=1}^{C} \lambda_Z^{-1} \exp[Z(Y - X)2\pi i/C]$, where $\lambda_Z = \sum_{W=1}^{C} a_{C-W} \exp[WZ2\pi i/C]$ (see Marcus and Minc (1964), page 66).

3. A procedure. Notice, from Equation (3), in the notation of (4b) and (6),

$$f(X - Y) = n.(n. - 1)^{-1} g(X - Y) - (n. - 1)^{-1}\delta_{X,Y}/C. \tag{7}$$

Suggestion. Estimate g from the data, a previous sample or even the object of the smoothing,

$$\hat{g}(Z) = C(Z)^{-1}\sum_X (n_X/n.)n_{X+Z}/n., \tag{8}$$

where $C(Z)$ indicates the number of cells X for which there exists the cell $X + Z$.

(Use the periodicity properties of g to determine $\hat{g}(Z)$ for $C(Z)$ small.) Choose the function $Ep_Xp_Y = f(X - Y)$ related to $g = \hat{g}$ by (7).

The use of (8) is, of course, recommended for the points X regularly spaced.

One might wish to choose g as a function fitted to $\hat{g}$, the smoothness of $\hat{g}$ helping to indicate its precision as an estimate. The overall mean and variance of $\hat{g}(Z)$ can be written,

$$E\hat{g}(Z) = g(Z)$$

and

$$\operatorname{Var} \hat{g}(Z) = \operatorname{Var} [E(\hat{g}(Z)|p_Y's)] + E[\operatorname{Var} (\hat{g}(Z)|p_Y's)],$$

where

$$E(\hat{g}(Z) \mid p_Y's) = (C(Z))^{-1}\sum_X [(1 - n.^{-1})p_Xp_{X+Z} + n.^{-1}\delta_{X,X+Z}p_X]$$

and

$$\begin{aligned}\operatorname{Var} (\hat{g}(Z) \mid p_Y's) = (C(Z))^{-2}(1 - n.^{-1})n.^{-2}[&- (4n. - 6)(\textstyle\sum_X p_Xp_{X+Z})^2 \\ &+ (2n. - 4)\textstyle\sum_X p_Xp_{X+Z}p_{X+2Z} + (n. - 2)\sum_X (p_Xp_{X+Z}^2 + p_X^2p_{X+Z}) \\ &+ \textstyle\sum_X (1 + \delta_{X,X+Z})p_Xp_{X+Z}].\end{aligned}$$

The concept of choosing moments of the prior distribution of the cell probabilities on the basis of a single set of observed cell counts could be viewed as an extension of the concept of an empirical Bayes procedure (Robbins (1964)). An empirical Bayes procedure is based on measurements of *independent* random variables having the sought prior distribution. The proposed choice of the mean $f(Z)$ for the prior distribution of p_Xp_{X+Z} is based on the measurements $(n_Yn_{Y+Z})/n.^2$ of the *mean-stationary* sequence of random variables p_Yp_{Y+Z}.

4. Multidimensional cells. In many applications, for example, classification problems, the multidimensionality of the cells X precludes any simple structure in the system (2) for the weights $w_X(Y)$. Even a purely numerical solution could easily be ruled out by an astonomical number of cells. We are led by the following additional symmetry condition on the prior distribution of the p_X's to a smaller, more usefully structured system. Reference is made to a given convenient norm $\|X\|$; for example, $\|X\|^2$ may be the usual sum of squared coordinates. The question of choosing a norm is a difficult one, essentially equivalent to choosing a joint transformation of variables.

Assume invariance of the prior second-order moment function f (and hence g) under "rotations" R. Namely, $\|RX\| = \|X\|$, for all X, implies

$$f(RX) = f(X), \quad \text{for all} \quad X. \tag{9}$$

The conditions (9) and (4a) yield

$$Ep_Xp_Y = f_1(\|X - Y\|), \qquad E(n_X/n.)n_Y/n.'' = g_1(\|X - Y\|),$$

and by (2)

$$w_X(Y) = h_1(\|X - Y\|),$$

all constant, in Y, on "shells" $0_{X,\rho}$,

$$0_{X,\rho} = \{Y\colon \|X - Y\| = \rho\},$$

$$0_{X,0} = X.$$

Consequently, each shell $0_{X,\rho}$ centered at X can be treated as a single cell, and the optimal weights $w_X(0_{X,\rho}) = h_1(\rho)$ satisfy a system of the form (2) — (3) with Ep_Y and Ep_Yp_Z replaced by

$$\mu(\rho) = Ep(0_{X,\rho}),$$

and

$$\varphi(\rho, \rho') = Ep(0_{X,\rho})\cdot p(0_{X,\rho'}).$$

It may be helpful in practice to group values of the radius ρ, say according to convenient intervals of its square $\rho^2 = \|X - Y\|^2$.

The second-order prior moments $\varphi(\rho, \rho')$ for the new cells could, in theory, be chosen by a version of (7) from the statistics,

$$\hat{\gamma}(\rho, \rho') = C^{-1}\textstyle\sum_X n_{0_{X,\rho}} n_{0_{X,\rho'}}/n.^2.$$

But, for any particular arguments, $\hat{\gamma}(\rho, \rho')$ could easily require the summation of an astronomical number C of terms. Any particular value of the previous version $\hat{g}(Z)$ (8) required a summation of at most $n.$ nonzero terms. To obtain a computationally feasible procedure, we consider a modification of the criterion of minimum mean squared error $E(\hat{p}_X - p_X)^2$, leading again to summations of at most $n.$ nonzero terms.

Modified Criterion. Let $\tilde{X}$ be a random vector distributed according to the unknown distribution p_X. Now, determine the weights w, under the constraint, $w_X(0_{X,\rho}) = h_1(\rho)$ independent of X, to minimize the overall expectation of the squared error of $\hat{p}_{\tilde{X}}$,

$$E(\hat{p}_{\tilde{X}} - p_{\tilde{X}})^2 = E\textstyle\sum p_X(\hat{p}_X - p_X)^2.$$

Then the optimal weights $h_1(\rho)$ satisfy a system of the form (2) — (3) with Ep_Y and Ep_Yp_Z replaced by

$$\tilde{\mu}(\rho) = Ep(0_{\tilde{X},\rho}) = \textstyle\sum_X Ep_Xp(0_{X,\rho}),$$

and

$$\begin{aligned}\tilde{\varphi}(\rho, \rho') &= Ep(0_{\tilde{X},\rho})p(0_{\tilde{X},\rho'})\\ &= \textstyle\sum_X Ep_Xp(0_{X,\rho})p(0_{X,\rho'}).\end{aligned}$$

Unbiased estimates of these prior moments are easily computed by correcting for bias in the analogous sums of products of cell frequencies,

$$\hat{\tilde{\mu}}(\rho) = n.(n. - 1)^{-1}\textstyle\sum_X n_X n_{0_{X,\rho}}/n.^2 - (n. - 1)^{-1}\delta_{0,\rho},$$

$$\begin{aligned}\hat{\tilde{\varphi}}(\rho, \rho') = {} & n.(n. - 1)^{-1}n.(n. - 2)^{-1}\textstyle\sum_X n_X n_{0_{X,\rho}} n_{0_{X,\rho'}}/n.^3\\ & - (n. - 2)^{-1}[\delta_{0,\rho}\hat{\tilde{\mu}}(\rho') + \delta_{0,\rho'}\hat{\tilde{\mu}}(\rho) + \delta_{\rho,\rho'}\hat{\tilde{\mu}}(\rho)]\\ & + 3(n. - 1)^{-1}(n. - 2)^{-1}\delta_{0,\rho}\delta_{\rho,\rho'}.\end{aligned}$$

For applications to practical classification problems, see the forthcoming paper, Dickey (1967).

5. Alternatives. A related criterion for choosing the weights w would be to minimize a chi-squared-like weighted sum of squares $\sum a_x(\hat{p}_X - n_X/n.)^2$ under some constraint such as $w_X(Y) = h(X - Y) = h(Y - X)$. This criterion leads to a system analogous to (2) with $En_Xn_Y/n.^2$ and Ep_Xp_Y replaced by $\hat{g}$-like quantities.

Whittle's smoothed estimate of a *density* function $p(x)$,

$$\hat{p}(x) = \int w(x, y)\, dF_{n.}(y),$$

a Stieltjes integral with respect to the empirical distribution $F_{n.}$, has weight function w satisfying an integral equation analogous to the system (2). A procedure to choose the required prior second moments $Ep(x)p(y)$ can be developed from the suggested procedure for cells X, located, say, at the midpoints of intervals $(x - \epsilon, x]$. Since,

$$E[p(x) - p(x - \epsilon)][p(y) - p(y - \epsilon)] = (\partial^2/\partial x\, \partial y)Ep_Xp_Y,$$

in case x is one-dimensional.

Other estimates of multinomial cell probabilities are based on special structural assumptions: Birch (1963); Bahadur (1961); Cornfield (1967).

Acknowledgment. I am grateful to Leonard J. Savage, Frederick Mosteller, Richard A. Olshen, and a referee for their help.

NOTE ADDED IN PROOF. D. V. Lindley (*J. Roy. Statist. Soc. Ser. B* **24** 286) in his discussion of a paper by C. M. Stein, G. E. P. Box and D. R. Cox (same journal **26** 218), and others have also proposed prior distributions based on peeking at the data. The methods of this paper will be generalized and extended to other smoothing problems in a later note.

REFERENCES

BAHADUR, R. R. (1961). A representation of the joint distribution of responses to n dichotomous items. *Studies in Item Analysis and Prediction*. Stanford University Press.

BIRCH, M. W. (1963). Maximum likelihood in three-way contingency tables. *J. R. Statist. Soc.* **25** 220–223.

CARLSON, B. C. (1963). Lauricella's hypergeometric function F_D. *J. Math. Analysis and Appl.* **7** 452–470.

CORNFIELD, JEROME (1967). Private communication.

DICKEY, JAMES M. (1967). Contributions to a practical Bayesian theory of classification. To appear.

MARCUS, MARVIN and MINC, HENRYK (1964). *A Survey of Matrix Theory and Matrix Inequalities*. Allyn and Bacon.

ROBBINS, HERBERT (1964). The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.* **35** 1–20.

WHITTLE, P. (1957). Curve and periodogram smoothing. *J. R. Statist. Soc.* **19** 38–47.

WHITTLE, P. (1958). On the smoothing of probability density functions. *J. R. Statist. Soc.* **20** 334–343.

WILKS, SAMUEL S. (1962). *Math. Statist.* Wiley, New York.