A CONTRIBUTION TO THE ECONOMETRIC THEORY OF UNOBSERVED COMPONENTS

Robert E. Hall

Department of Economics
Massachusetts Institute of Technology

January 1973

The research reported here was carried out while the author was on a leave supported by the Social Science Research Council. Support was also provided by the National Science Foundation.

Statistical models of unobserved components seem destined for an increasing role in econometric work. Especially in cross-sections, the difference in the values of the left-hand variables among observations with identical values of the right-hand variables are sufficiently large to justify careful analysis of the apparently random component of the behavior under study. The simple characterization of randomness implicit in the stochastic specification of the regression model seems inadequate when the right-hand variables in a problem account for only a small portion of the dispersion of the left-hand variable. Many recent authors have sought to attribute part of the randomness in their samples to variations within the population of characteristics that are not observed. For example, Griliches (1972) assigns part of the dispersion of earnings conditional on education to the unobserved differences in ability of individuals with equal amounts of education. McFadden (1972) hypothesizes a distribution of tastes within the population to explain choices of modes of transportation by individual commuters. The present paper takes up the following question: what can be discovered about the underlying distribution of characteristics from the observed body of data? Are the assumptions about the distributions of unobserved characteristics made by previous authors verifiable, or must they be accepted on pure faith?

A general statistical model suitable for this discussion is the following

$$y = h(x, \theta, u)$$

where y is the scalar left-hand variable, assumed to be qualitative (taking on only a finite number of integer values[1]), x is a vector of observed

---

[1] If the left-hand variable is continuous, y can be defined by a set of intervals of values of the variable.

characteristics, θ is the unobserved characteristic, and u is a disturbance whose distribution may depend on x and θ. Apart from the presence of θ, this would be a regression model if the distribution of u did not depend on x and θ; in the qualitative case especially, however, this dependence is critical. Our discussion concerns the untangling of the separate effects of θ and u, where the role of x is subsidiary, so for the rest of the paper we consider the case of sampling from a population whose members are observationally identical, where it is appropriate to suppress x:

$$y = h(\theta,u)$$

All observations from the same individual are assumed to correspond to the same θ, but each one involves a new drawing from the distribution of u. Finally, we assume prior knowledge of $h(\theta,u)$ and of the distribution of u. The last assumption should become more plausible as the discussion progresses.

Models of unobserved components are particularly important in the study of the distribution of income. The major theme of the most influential recent work on income distribution, Christopher Jencks' book, Inequality (1972), is exactly that observed differences among individuals account for very little of the dispersion of income among them: "Neither family background, cognitive skill, educational attainment, nor occupational status explains much of the variation in men's incomes. Indeed, when we compare men who are identical in all these respects, we find only 12 to 15 percent less inequality than among random individuals. How are we to explain these variations among men who seem to be similarly situated?" (p. 227). Jencks replies that unmeasured differences in motivation, ability, and especially luck account for the bulk of the dispersion in income. His discussion is limited by his failure to distinguish between unobserved differences among individuals, on the one hand, and differences in the experience of the same individual at different points in time, on the other. In the context of measuring income, this distinction is familiar to economists in Milton Friedman's notion of the permanent and transitory components of measured income. Jencks alludes briefly to the distribution of permanent income (footnote 1, p. 233) but the distinction has no role in his discussion.

The class of statistical models studied here provides a general framework for separating the two sources of the apparently random differences among individuals at a point in time. Systematic differences among individuals are indexed by the random variable θ, and differences in the experiences of a single individual by the random variable u. Friedman's model is a special case of the general model in which θ and u are simply added together

$$y = \theta + u$$

Here θ is permanent income and u is transitory income. If y is observed for a few successive years, then it is tempting to estimate permanent income for the kth individual as the average income over the years:

$$\hat{\theta} = \frac{1}{T}\sum_{t=1}^{T} y_t$$

The difficulty is that the distribution of $\hat{\theta}$ among the members of the population has more dispersion than the distribution of θ. This problem arises most critically in Jencks' data, where T is 1, but even where T is 3 or 4 one does not know how much the distribution of $\hat{\theta}$ tells about the distribution of θ. What is needed, and what this paper supplies, is a method for extracting as much reliable information as possible about the distribution of θ.

The Sex Composition of Families

The details of the problem considered in this paper can best be introduced through an example. Suppose that we observe a large number of apparently identical families with two children, and suppose further that a fraction ϕ_1 of the families have no girls, ϕ_2 have one girl, and ϕ_3 have two girls. Each family has a probability θ that a given child will be a girl. In terms of the general model just given, if y is the number of girls in a family,

$$y = h(\theta,u)$$

$$= u, \text{ where } u \text{ is binomial of order 2 with parameter } \theta.$$

If all families have the same θ, then ϕ will be the common binomial distribution:

$$\phi_1 = (1-\theta)^2; \quad \phi_2 = 2\theta(1-\theta); \quad \text{and} \quad \phi_3 = \theta^2$$

If θ varies among families, then ϕ will no longer be binomial.[2] The problem treated in this paper is the estimation of the distribution of θ given only the observed distribution ϕ. Suppose, first, that there is a finite set of alternative propensities to bear girls, say $\theta_1,\ldots,\theta_N$, and that a fraction p_1 of the population has propensity θ_1, p_2 has propensity θ_2, and so forth. Then the relation between these and the observed ratios is

$$\phi_1 = \sum_{j=1}^{N}(1 - \theta_j)^2 p_j$$

$$\phi_2 = 2\sum_{j=1}^{N}\theta_j(1 - \theta_j)p_j$$

$$\phi_3 = \sum_{j=1}^{N}\theta_j^2 p_j$$

---

[2] This possibility has been discussed in the literature on mathematical demography (for example, Goodman (1961) and Weiler (1959)). The problem of estimating the underlying distribution does not seem to have been studied by demographers. My treatment of the sex composition of families is only

A matrix form of this is convenient; let ϕ and p be vectors of length M ($M = 3$ in this case) and N, respectively, and let A be a matrix with M rows and N columns, each of whose columns is the probability associated with one of the values of θ. Then

$$\phi = Ap$$

What can be deduced about p given the known A and the observed ϕ? This is the basic question answered by this paper.

The question becomes interesting only when N is larger than M. Then the equation above together with the requirement that p is a probability ($p_j \geq 0$, $j = 1, \ldots, N$ and $\Sigma p_j = 1$) defines a set, $S(A,\phi)$, of alternative values of p that are compatible with the observed ϕ. The paper characterizes the solution set in a number of ways, the most important being an identification theorem giving necessary and sufficient conditions that there exist some vector ϕ for which a given probability, say p_1, is bounded strictly above zero.

A more realistic hypothesis is that θ obeys a continuous distribution described by a density function, $f(\theta)$. Then the observed probabilities are

$$\phi = \int_{-\infty}^{\infty} a(\theta)\, f(\theta) d\theta$$

where $a(\theta)$ is a vector-valued function giving the set of probabilities in a unit with parameter θ; in the example of the sex composition of families,

$$a(\theta) = \begin{bmatrix} (1-\theta)^2 \\ 2\theta(1-\theta) \\ \theta^2 \end{bmatrix}$$

It would be asking too much to expect to extract any information about $f(\theta)$ at a single point. Information can be obtained about the cumulative distribution

---

an example and does not consider other aspects of the problem, especially the effects of the efforts of parents to influence the composition through stopping rules. On this, see Ben Porath and Welch (1972).

of θ at a finite set of points, however. Let such a set be $\bar{\theta}_1, \ldots, \bar{\theta}_{N-1}$. Further, let

$$\underline{a}_{i,j} = \min_{\bar{\theta}_{j-1} \leq \theta \leq \bar{\theta}_j} a_i(\theta)$$

$$\bar{a}_{i,j} = \max_{\bar{\theta}_{j-1} \leq \theta \leq \bar{\theta}_j} a_i(\theta)$$

where $\bar{\theta}_0 = -\infty$ and $\bar{\theta}_N = +\infty$. Now

$$\int_{-\infty}^{\infty} a_i(\theta)\ f(\theta)\ d\theta \geq \sum_{j=1}^{N} \underline{a}_{i,j} \int_{\bar{\theta}_{j-1}}^{\bar{\theta}_j} f(\theta) d\theta$$

and

$$\int_{-\infty}^{\infty} a_i(\theta)\ f(\theta)\ d\theta \leq \sum_{j=1}^{N} \bar{a}_{i,j} \int_{\bar{\theta}_{j-1}}^{\bar{\theta}_j} f(\theta)\ d\theta$$

Let $p_j = \int_{\bar{\theta}_{j-1}}^{\bar{\theta}_j} f(\theta)\ d\theta$; this is just the probability that a unit drawn at random will have a θ between $\bar{\theta}_{j-1}$ and $\bar{\theta}_j$. Then we have the system of inequalities,

$$\underline{A}p \leq \phi \leq \bar{A}p$$

where $\underline{A}$ and $\bar{A}$ are the M by N matrices of values of $\underline{a}_{ij}$ and $\bar{a}_{ij}$. Again, the paper characterizes the set of solutions, $S(\underline{A}, \bar{A}, \phi)$ and presents necessary and sufficient conditions for the identifiability of p_1.

The problem of drawing inferences about the density, $f(\theta)$, given only the finite set of parameters describing it, ϕ, is a generalization of the problem associated with the Chebysheff inequality. The latter takes two parameters, the first and second moments of $f(\theta)$, and gives a lower bound on $\int_{\mu-t}^{\mu+t} f(\theta)d\theta$ where μ is the first moment of $f(\theta)$. Our generalization permits arbitrary definitions of the intervals for which the bounds are calculated, allows any number of parameters, each of which may be an arbitrary weighting of $f(\theta)$, and gives both upper and lower bounds. It reduces to the Chebysheff inequality for intervals that are symmetric about the mean of $f(\theta)$, and where the parameters, ϕ, are equivalent to the first and second moments of $f(\theta)$. The distribution of the sex composition of families is such a case, where the mean of $f(\theta)$ is $\frac{1}{2}\phi_2 + \phi_3$ and the second moment is[3] ϕ_3. Note that ϕ has only two independent parameters, since the three ϕ_i must sum to one.

---

[3] Steven Shavell pointed out the relation between ϕ and the moments of $f(\theta)$. In the case of a population of families with M-1 children, ϕ is equivalent to the first M-1 moments of $f(\theta)$. In other cases, however, there may not be any way to recover the moments from ϕ, so we do not pursue this point here.

The Discrete Case

The solution set, $S(A,\phi)$ is a convex polyhedron.[4] As such, it can be fully described by its finite set of vertices, $v^{(1)}, \ldots, v^{(R)}$, in that $S(A,\phi)$ is the convex hull of its vertices.

$$S(A,\phi) = < v^{(1)}, \ldots, v^{(R)} >$$

Each vertex is an extreme value of the set of probabilities consistent with the observed distribution, ϕ. From the vertices we can compute bounds on the individual probabilities:

$$\underline{p}_j = \min_{k=1,\ldots,R} v_j^{(k)}$$

$$\bar{p}_j = \max_{k=1,\ldots,R} v_j^{(k)}$$

The set of probabilities that fall within these bounds understates the information available about p, because it encloses $S(A,\phi)$. Still, the individual bounds are a compact way to summarize the information in the set of vertices, since there may be a large number of them.

The computation of the vertices is in principle straightforward. Each vertex has at most M non-zero values, so each is a solution to a set of M equations in M unknowns,

$$\hat{A}\,\hat{v} = \phi$$

where $\hat{A}$ is a square matrix consisting of M columns selected from the N columns of A, and $\hat{v}$ is the vector of corresponding non-zero elements of the vertex, v. There are

$$\frac{N!}{(N-M)!\,M!}$$

---

[4] For economists, Gale (1960), chapter 2, provides all the algebraic background

ways to choose $\hat{A}$. Not all of these give rise to vertices, however, since some of the solutions may have negative elements. The set of vertices can be obtained by generating all the possible solutions in a systematic way and selecting those that are non-negative. The pivot method provides an efficient way to do this by exchanging one column for another at each step.

What conditions are required to assure that this procedure yields any useful information about p? This is the problem of identification. It is clear that p is not identified in the usual econometric sense, which would require that exact knowledge of ϕ imply exact knowledge of p. Here we must settle for a weaker notion of identifiability. The natural definition seems to be that there be some possible value of the observed ϕ such that the probability under consideration, say p_1, is definitely positive, that is, that units with $\theta = \bar{\theta}_1$ must exist in the population. To make this more formal, let $a^{(1)}, \ldots, a^{(N)}$ be the columns of A. The set of possible values of ϕ implied by the model is just the convex hull of these columns. Then p_1 is identifiable if

$$\max_{\phi \epsilon < a^{(1)}, \ldots, a^{(N)}>} \left[\min_{p \epsilon S(A,\phi)} p_1 \right] > 0$$

The probability p_1 will be identifiable if the column $a^{(1)}$ has something to say about ϕ that is independent of the other columns. One's first thought is that identifiability depends on a condition on the rank of submatrices of A, that $a^{(1)}$ should not be a linear combination of less than M other columns, but this is not so. Suppose

$$A = \begin{bmatrix} \frac{1}{2} & \frac{1}{3} & \frac{1}{6} & \frac{2}{3} \\ 0 & \frac{1}{3} & \frac{2}{3} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{6} & 0 \end{bmatrix}$$

---

necessary to understand what follows. A more extensive treatment and a complete list of references appears in Valentine (1964), Part 12.

and $\phi = \begin{bmatrix} \frac{1}{2} \\ 0 \\ \frac{1}{2} \end{bmatrix}$

This has the unique solution

$$p = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

and so p_1 is identified despite the linear dependence

$$a^{(1)} = 2a^{(2)} - a^{(3)}$$

This example shows that non-negativity of the elements of p has identifying power that can make up for linear dependence of $a^{(1)}$ on other columns.

The true condition for identifiability of p_1 is a refinement of the requirement that $a^{(1)}$ not be a linear combination of less than M other columns. The refinement is that a linear combination is permissable if at least one of its coefficients is negative. This is stated formally in the

Identification Theorem for the Discrete Case

A necessary and sufficient condition for the identifiability of p_1 is that $a^{(1)}$ be a vertex of $< a^{(1)},\ldots, a^{(N)} >$.

Proof:

(i) Necessity. Suppose that $a^{(1)}$ is not a vertex. Then there is a vector

$$\lambda = \begin{bmatrix} -1 \\ \lambda_2 \\ \vdots \\ \lambda_N \end{bmatrix}$$

with $\lambda_i \geq 0$, $i = 2,\ldots, N$, and $\sum_{i=2}^{N} \lambda_i = 1$, such that

$$A\lambda = 0$$

Now consider a solution, p, to

$$Ap = \phi$$

Let p^* be defined by

$$p^* = p + p_1\lambda$$

Note that $p^* \geq 0$ and p_1^* is zero. Now

$$Ap^* = Ap + p_1 A\lambda$$

$$= Ap$$

$$= \phi$$

Thus for each solution, p, there corresponds another solution, p^*, such that $p_1^* = 0$, so p_1 is not identifiable.

(ii) Sufficiency. We need to exhibit a vector ϕ such that all solutions to $Ap = \phi$ have $p_1 > 0$. Consider $\phi = a^{(1)}$. Since $a^{(1)}$ is not a non-negative combination of the other columns of A,

$$p = \begin{bmatrix} 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$

is the unique solution, and it has $p_1 > 0$.

Application to the Distribution of Family Compositions

We return briefly to the example considered at the beginning of the paper. Our purpose is to show that the hypotheses of the theorem on identification for the discrete case can be met in practical work and to indicate the tightness of the bounds on p that are obtained for alternative values of ϕ.

The question of identification in the problem of family composition can be answered quickly and unambiguously. Identifiability holds for any value of N, provided only that $\theta_1, \ldots, \theta_N$ are distinct. To prove this, it suffices to show that no column of A can be written as a convex combination of any three other columns (see Gale, 1960, Theorem 2.11, p. 50). Consider such a combination with coefficients λ_1, λ_2 and $1 - \lambda_1 - \lambda_2$. Then

$$(1-\theta)^2 = \lambda_1(1-\bar{\theta}_1)^2 + \lambda_2(1-\bar{\theta}_2)^2 + (1-\lambda_1-\lambda_2)(1-\bar{\theta}_3)^2$$

$$2\theta(1-\theta) = 2\lambda_1\bar{\theta}_1(1-\bar{\theta}_1) + 2\lambda_2\bar{\theta}_2(1-\bar{\theta}_2) + 2(1-\lambda_1-\lambda_2)\bar{\theta}_3(1-\bar{\theta}_3)$$

$$\theta^2 = \lambda_1\bar{\theta}_1^2 + \lambda_2\bar{\theta}_2^2 + (1-\lambda_1-\lambda_2)\bar{\theta}_3^2$$

The solution to this system is

$$\lambda_1 = \frac{(\bar{\theta}_2-\theta)(\bar{\theta}_3-\theta)}{(\bar{\theta}_2-\bar{\theta}_1)(\bar{\theta}_3-\bar{\theta}_1)}$$

$$\lambda_2 = \frac{(\bar{\theta}_1-\theta)(\bar{\theta}_3-\theta)}{(\bar{\theta}_1-\bar{\theta}_2)(\bar{\theta}_3-\bar{\theta}_2)}$$

$$1-\lambda_1-\lambda_2 = \frac{(\bar{\theta}_1-\theta)(\bar{\theta}_2-\theta)}{(\bar{\theta}_3-\bar{\theta}_1)(\bar{\theta}_3-\bar{\theta}_2)}$$

Without loss of generality we take $\bar{\theta}_1 < \bar{\theta}_2 < \bar{\theta}_3$. Then it is impossible for all three of these coefficients to be non-negative: If $\theta < \bar{\theta}_1$, $\lambda_2 < 0$. If $\bar{\theta}_1 < \theta < \bar{\theta}_2$,

$1-\lambda_1-\lambda_2<0$. If $\bar{\theta}_2 < \theta < \bar{\theta}_3$, $\lambda_1 < 0$. Finally, if $\theta > \bar{\theta}_3$, $\lambda_2 < 0$. Thus no convex combination is possible. No matter how many values of θ are considered in the discrete problem of family composition, there is a ϕ such that the probability assigned to a given one of them is definitely positive.

Now consider the following observed distribution of compositions: $\phi_1 = 0.26$, $\phi_2 = 0.48$, and $\phi_3 = 0.26$. This distribution shows more dispersion than the binomial, and must therefore correspond to a distribution of θ that is not concentrated at a point. If we consider $\bar{\theta}_j = 0.3, 0.4, 0.5, 0.6, 0.7$, then the vertices of $S(A,\phi)$ are the following:

	p_1 $\theta = .3$	p_2 $\theta = .4$	p_3 $\theta = .5$	p_4 $\theta = .6$	p_5 $\theta = .7$
$v^{(1)}$	0	0.5	0	0.5	0
$v^{(2)}$	0.17	0	0.5	0.33	0
$v^{(3)}$	0.13	0	0.75	0	0.13
$v^{(4)}$	0	0.33	0.5	0	0.17

The individual bounds are $0 \leq p_1 \leq 0.17$, $0 \leq p_2 \leq 0.5$, $0 \leq p_3 \leq 0.75$, $0 \leq p_4 \leq 0.5$, and $0 \leq p_5 \leq 0.17$. This ϕ does not identify any individual p_j in the sense defined earlier, but it does identify certain combinations of them: $0.5 \leq p_2 + p_3 \leq 0.83$, $0.5 \leq p_3 + p_4 \leq 0.83$, and $0.75 \leq p_2 + p_3 + p_4 \leq 1.0$.

The following distribution of sexes of the first two children of American families is reported by Ben Porath and Welch (1972):

$$\phi_1 = 0.262, \ \phi_2 = 0.497, \text{ and } \phi_3 = 0.241$$

The mean of the distribution of θ, $\frac{1}{2}\phi_2 + \phi_3$, is 0.489, so there is a slight tendency for families to have more boys than girls. With the same set of $\bar{\theta}_j$

as in the previous example, the vertices of $S(A, \phi)$ are

	p_1 $\theta = .3$	p_2 $\theta = .4$	p_3 $\theta = .5$	p_4 $\theta = .6$	p_5 $\theta = .7$
$v^{(1)}$	.02	.06	.92	0	0
$v^{(2)}$	0	.13	.85	.02	0
$v^{(3)}$	.08	.12	.87	0	.01

Thus $0 \leq p_1 \leq .08$, $.06 \leq p_2 \leq .13$, $.87 \leq p_3 \leq .92$, $0 \leq p_4 \leq .02$ and $0 \leq p_5 \leq .01$. The data are consistent with only a very small proportion of families with $\theta = 0.3$, 0.6, or 0.7. In fact, the bound on $p_2 + p_3$ is very sharp: $0.98 \leq p_2 + p_3 \leq .99$.

The Continuous Case

The continuous case requires

$$\underline{A}p \leqq \phi \leqq \bar{A}p$$

It bears a close connection to the discrete case in the following way: Any p in the set of solutions to the continuous case, $S(\underline{A}, \bar{A}, \phi)$ can be portrayed as the solution to a particular discrete problem characterized by a pair of vectors that play the roles of probabilities, and any solution to such a discrete problem is also a solution to the continuous problem. This is formalized in a

Lemma

$$S\ (\underline{A}, \bar{A}, \phi) = \bigcup_{\psi \varepsilon H(\tilde{\phi})} S\ (\tilde{A}, \psi)$$

where $\tilde{A} = \begin{bmatrix} \underline{A} \\ -\bar{A} \\ v \end{bmatrix}$, $v = [1 \ldots 1]$, $\tilde{\phi} = \begin{bmatrix} \phi \\ -\phi \\ 1 \end{bmatrix}$ and

$$H(\tilde{\phi}) = \{\psi \mid \psi \leqq \tilde{\phi} \quad \text{and } \psi_{2M+1} = 1\}$$

Proof:

(i) Consider $p \varepsilon S(\underline{A}, \bar{A}, \phi)$, so $\tilde{A}p \leqq \tilde{\phi}$. Let $\psi = \tilde{A}p$; $\psi \varepsilon H(\tilde{\phi})$. Then $p \varepsilon S(\tilde{A}, \psi)$, as required.

(ii) Consider $p \varepsilon S(\tilde{A}, \psi)$ for some $\psi \varepsilon H(\tilde{\phi})$. Then $\tilde{A}p = \psi \leqq \tilde{\phi}$, and $p \varepsilon S(\underline{A}, \bar{A}, \phi)$, as required.

The set of solutions, $S(\underline{A}, \bar{A}, \phi)$, is again a convex polyhedron and can be represented most compactly as the convex hull of its vertices. Each vertex is a solution to the system

$$[\ \tilde{A} \quad -I\ ] \begin{bmatrix} p \\ \psi \end{bmatrix} = 0$$

satisfying the constraints $p \geq 0$ and $\psi \leq \tilde{\phi}$, and with no more than 2M + 1 non-zero elements in p and ψ together, say k from p, $\hat{p}$, and 2M + 1 - k from ψ, $\hat{\psi}$. Then $\hat{p}$ and $\hat{\psi}$ must satisfy

$$\begin{bmatrix} \hat{A} & 0 \\ A^* & -I \end{bmatrix} \begin{bmatrix} \hat{p} \\ \hat{\psi} \end{bmatrix} = \begin{bmatrix} \hat{\phi} \\ 0 \end{bmatrix}$$

where $\hat{\phi}$ consists of the k elements of $\tilde{\phi}$ corresponding to the elements of ψ not included in $\hat{\psi}$, and $\hat{A}$ and A^* contain the appropriate elements of $\tilde{A}$. Since the system is block-triangular, it has the recursive solution

$$\hat{p} = \hat{A}^{-1}\hat{\phi}$$
$$\hat{\psi} = A^*\hat{p}$$

$\hat{p}$ contains the non-zero elements of a vertex of $S(\underline{A}, \bar{A}, \phi)$ if $\hat{p} \geq 0$ and the elements of $\hat{\psi}$ are no larger than the corresponding elements of $\tilde{\phi}$. Again, the vertices of the set of all solutions can be calculated by generating systematically the solutions for all possible choices of the elements of $\hat{p}$ and $\hat{\psi}$.

The problem of identification arises more acutely in the continuous case than in the discrete case. In addition to the uncertainty about p inherent in the discrete case, further uncertainty is introduced by the lack of knowledge about where the probability is distributed inside each interval. In what follows we give necessary and sufficient conditions for the identifiability of p_1. Although these conditions are rather more complicated than those for the discrete case, they are basically of the same form and can be verified for a given problem in a finite number of computational steps. We begin by defining the set of all possible observable probabilities compatible with a given pair of matrices $\underline{A}$ and $\bar{A}$:

$$\Phi = \{ \phi \mid \text{there exists } p \text{ such that } \underline{A}\, p \leqq \phi \leqq \bar{A}\, p;\ p \geqq 0,\ \phi \geqq 0 \text{ and } \Sigma p_i = \Sigma \phi_i = 1 \}$$

It is possible to show that Φ is a convex polyhedron and thus can be represented as the convex hull of a finite set of vertices:

$$\Phi = \langle \phi^{(1)}, \ldots, \phi^{(Q)} \rangle$$

Second, define B as the convex hull of the columns of the stacked matrix $\tilde{A}$ other than its first column:

$$B = \langle \tilde{a}^{(2)}, \ldots, \tilde{a}^{(N)} \rangle$$

Then we have the

<u>Identification Theorem for the Continuous Case</u>: p_1 is identifiable if and only if $H(\tilde{\phi}^{(k)}) \cap B$ is empty for at least one vertex of Φ, $\phi^{(k)}$.

<u>Proof</u>: (i) Necessity. Suppose there is no such $\phi^{(k)}$. Then for every k, $H(\tilde{\phi}^{(k)}) \cap B$ contains a $\psi^{(k)} = \tilde{A}p^{(k)}$ and $p_1^{(k)} = 0$. By the Lemma, $p^{(k)} \varepsilon S(\underline{A}, \bar{A}, \phi^{(k)})$. Now consider an arbitrary $\phi \varepsilon \Phi$. It can be written $\phi = \sum_{k=1}^{Q} \lambda_k \phi^{(k)}$. Let $p = \Sigma \lambda_k p^{(k)}$. Then $p \varepsilon S(\underline{A}, \bar{A}, \Sigma \lambda_k \phi^{(k)})$ and $p \varepsilon S(\underline{A}, \bar{A}, \phi)$. But $p_1 = 0$, and identifiability cannot hold.

(ii) Sufficiency. Consider a $\phi^{(k)}$ such that $H(\tilde{\phi}^{(k)}) \cap B$ is empty, and consider any $\psi \varepsilon H(\tilde{\phi}^{(k)})$. Suppose there were a $p \varepsilon S(\tilde{A}, \psi)$ with $p_1 = 0$. Then $Ap = \psi$, and $\psi \varepsilon B$, a contradiction. Thus for any $p \varepsilon \bigcup_{\psi \varepsilon H(\tilde{\phi}^{(k)})} S(\tilde{A}, \psi)$, $p_1 > 0$, and p_1 is identifiable.

The following is a much simpler necessary condition:

Corollary : p_1 is identifiable only if $H(\tilde{a}^{(1)}) \cap B$ is empty.

Proof: Suppose, on the contrary, that $H(\tilde{a}^{(1)}) \cap B$ is not empty. Then there is a vector λ with $\lambda_1 = 0$, $\Sigma\lambda_j = 1$ such that $\tilde{a}^{(1)} \geqq \tilde{A}\lambda$. Now consider an arbitrary $\phi\varepsilon\Phi$. By hypothesis there is a $\psi\varepsilon H(\tilde{\phi})$ and a p such that $\psi = \tilde{A}p$; $p_j \geqq 0$ and $\Sigma p_j = 1$. Let $p_1' = 0$ and $p_j' = p_1\lambda_j + p_j$, $j > 1$. Define $\psi' = \tilde{A}p' = p_1(\tilde{A}\lambda - \tilde{a}^{(1)}) + \tilde{A}p$. Thus $\psi' \leqq \psi$, so $\psi'\varepsilon H(\tilde{\phi})$. But $p'\varepsilon S(\tilde{A},\psi')$, and p_1 is not identifiable.

Further Application to the Sex Composition of Families

For the distribution of the number of girls among the first two children reported by Ben Porath and Welch ($\phi_1 = 0.262$; $\phi_2 = .497$; $\phi_3 = .241$), $S(\underline{A}, \bar{A}, \phi)$ has the 23 vertices shown in Table 1. Here we have used six intervals along the θ axis: $\bar{\theta}_0 = 0.1$, $\bar{\theta}_1 = 0.3$, $\bar{\theta}_2 = 0.4$, $\bar{\theta}_3 = 0.5$, $\bar{\theta}_4 = 0.6$, $\bar{\theta}_5 = 0.7$, and $\bar{\theta}_6 = 0.9$. The bounds on the individual probabilities are as follows:

$0 \leq \text{Prob}\ [0.1 \leq \theta \leq 0.3] \leq .037$

$0 \leq \text{Prob}\ [0.3 \leq \theta \leq 0.4] \leq .150$

$0 \leq \text{Prob}\ [0.4 \leq \theta \leq 0.5] \leq .963$

$0 \leq \text{Prob}\ [0.5 \leq \theta \leq 0.6] \leq .963$

$0 \leq \text{Prob}\ [0.6 \leq \theta \leq 0.7] \leq .150$

$0 \leq \text{Prob}\ [0.7 \leq \theta \leq 0.9] \leq .038$

Although none of the individual probabilities is identified by this value of ϕ, certain sums are identified, for example,

$.850 \leq \text{Prob}\ [0.4 \leq \theta \leq 0.6] \leq 1.000$

$.962 \leq \text{Prob}\ [0.3 \leq \theta \leq 0.7] \leq 1.000$

$.037 \leq \text{Prob}\ [0.1 \leq \theta \leq 0.5] \leq 1.000$

These results tend to confirm the earlier impression that the observed distribution of the sex composition of families is not consistent with a great deal of dispersion in the underlying probabilities of having girls.

TABLE 1 Vertices of $S(\underline{A}, \bar{A}, \phi)$ for BPW data

p_1	p_2	p_3	p_4	p_5	p_6
.037	0	0	.963	0	0
0	.056	0	.944	0	0
0	0	.109	.891	0	0
.037	0	.963	0	0	0
0	0	.850	0	.150	0
0	0	.963	0	0	.037
0	.100	.900	0	0	0
.004	0	.090	.906	0	0
0	.023	.060	.918	0	0
0	0	.128	.850	.022	0
0	0	.120	.873	0	.007
.023	.060	.918	0	0	0
0	.150	.800	.050	0	0
0	.133	.850	0	.017	0
0	.124	.870	0	0	.007
0	.094	0	.851	.056	0
0	0	.283	.567	.150	0
0	.082	0	.901	0	.017
0	0	.200	.762	0	.038
.022	.060	.918	0	0	0
0	.128	.850	.022	0	0
0	.120	.873	0	.007	0
0	.114	.884	0	0	.002

Parametric Restrictions on $f(\theta)$

The previous analysis makes no assumptions about the form of the density of the unobserved component, $f(\theta)$. Most investigators are likely to be willing to make more or less restrictive assumptions about $f(\theta)$ in order to tighten the results by ruling out implausible densities. If $f(\theta)$ is thought to depend on K parameters, say β, then the relation between the observed ϕ and the underlying distribution of individual types is

$$\phi = \int_{-\infty}^{\infty} a(\theta)f(\theta,\beta)d\theta$$

If K is less than or equal to the number of independent observed data points, M-1, and if $f(\theta,\beta)$ is sufficiently well behaved so that the relation can be inverted to get β as a function of ϕ, then this is a perfectly conventional statistical model and none of the problems considered earlier in this paper will arise.[5] But if K exceeds M-1, and if $f(\theta,\beta)$ is linear in β (not, of course, in θ), then the problem is very much like the discrete problem considered earlier.

Suppose, therefore, that $f(\theta,\beta) = g(\theta)\beta$. Then

$$\phi = \int_{-\infty}^{\infty} a(\theta)g(\theta)\beta d\theta$$

If we define the matrix A by

$$A_{ij} = \int_{-\infty}^{\infty} a_i(\theta)g_j(\theta)d\theta ,$$

then the matrix form of the model is

$$\phi = A\beta ,$$

[5] Steven Shavell has pursued this line of thought in the problem of sex composition by assuming that $f(\theta,\beta)$ is a beta distribution, for which the necessary inversion is straightforward.

exactly the form of the discrete problem considered earlier. Each column of A can be scaled so that it sums to one. If, in addition, non-negativity of $g(\theta)\beta$ requires non-negativity of β, then the set of vectors of parameters, β, is precisely the set $S(A,\phi)$ discussed earlier.

The step function is one natural choice for $g(\theta)$:

$$g_j(\theta) = 1 \quad \text{if } \bar{\theta}_{j-1} \leq \theta < \bar{\theta}_j$$
$$= 0 \quad \text{otherwise}$$

For this case, the scaled matrix A is

$$A_{ij} = \frac{1}{\bar{\theta}_j - \bar{\theta}_{j-1}} \int_{\bar{\theta}_{j-1}}^{\bar{\theta}_j} a_i(\theta)d\theta \quad ,$$

$$\text{and } f(\theta) = \frac{\beta_j}{\bar{\theta}_j - \bar{\theta}_{j-1}} \quad \text{if } \bar{\theta}_j \leq \theta < \bar{\theta}_j$$
$$= 0 \quad \text{otherwise}$$

$$\text{Thus Prob } [\bar{\theta}_{j-1} \leq \theta < \bar{\theta}_j] = \int_{\bar{\theta}_{j-1}}^{\bar{\theta}_j} \frac{\beta_j}{\bar{\theta}_j - \bar{\theta}_{j-1}} d\theta = \beta_j \ ,$$

so each β_j has the same interpretation as the p_j in the continuous problem; it is the fraction of the population with θ falling between $\bar{\theta}_{j-1}$ and $\bar{\theta}_j$.

For the problem of the sex composition of families with two children,

$$a^{(j)} = \frac{1}{\bar{\theta}_j - \bar{\theta}_{j-1}} \begin{bmatrix} \frac{1}{3}((1 - \bar{\theta}_{j-1})^3 - (1 - \bar{\theta}_j)^3) \\ \bar{\theta}_j^2 - \bar{\theta}_{j-1}^2 - \frac{2}{3}(\bar{\theta}_j^3 - \bar{\theta}_{j-1}^3) \\ \frac{1}{3}(\bar{\theta}_j^3 - \bar{\theta}_{j-1}^3) \end{bmatrix}$$

Again we examine the set of distributions consistent with the data of Ben Porath and Welch, this time using a somewhat different set of values of $\bar{\theta}_j$: 0.1, 0.4, 0.47, 0.49, 0.51, 0.6, 0.9. The set has the following 8 vertices:

β_1	β_2	β_3	β_4	β_5	β_6
.014	.108	0	.878	0	0
.019	0	.283	.698	0	0
.008	0	.840	0	.152	0
0	.106	.704	0	.190	0
0	.243	0	.662	.096	0
0	.193	0	.798	0	.008
0	0	.735	.249	0	.017
0	0	.903	0	.086	.011

The following bounds on individual probabilities and sums of pairs of probabilities are implied:

$0 \leqq \text{Prob } [0.1 \leqq \theta < 0.4] \leqq .019$

$0 \leqq \text{Prob } [0.4 \leqq \theta < 0.47] \leqq .243$

$0 \leqq \text{Prob } [0.47 \leqq \theta < 0.49] \leqq .903$

$0 \leqq \text{Prob } [0.49 \leqq \theta < 0.51] \leqq .878$

$0 \leqq \text{Prob } [0.51 \leqq \theta < 0.6] \leqq .190$

$0 \leqq \text{Prob } [0.6 \leqq \theta < 0.9] \leqq .017$

$$0 \leq \text{Prob } [0.1 \leq \theta < 0.47] \leq .243$$

$$.108 \leq \text{Prob } [0.4 \leq \theta < 0.49] \leq .903$$

$$.662 \leq \text{Prob } [0.47 \leq \theta < \text{o}.51] \leq .984$$

$$.086 \leq \text{Prob } [0.49 \leq \theta < 0.6] \leq .878$$

$$0 \leq \text{Prob } [0.51 \leq \theta < 0.9] \leq .190$$

No more than 3.6% of all families have probabilities of having girls below 0.4 or above 0.6, and at most 33.8% and perhaps as few as 1.6% have probabilities below 0.47 or above 0.51.

Sampling

So far, the observed probabilities, ϕ, have been taken as known. In practical work, we have only a set of frequencies for a finite sample, say $\hat{\phi}$, which we regard as a random variable. If the total sample contains n individuals, the probability of a given sample, $y_1,\ldots,y_n$, is

$$\text{Prob }[y_1,\ldots,y_n] = \prod_{k=1}^{n} \phi_{y_k}$$
$$= \phi_1^{Y_1} \phi_2^{Y_2} \ldots \phi_M^{Y_M}$$

where Y_1 is the number of individuals with $y_k = 1$, y_2 the number with $y_k = 2$, and so forth ($\Sigma Y_i = n$). This probability may be interpreted as the likelihood of ϕ, in which case it is convenient to deal with the log-likelihood,

$$\log L(\phi) = \sum_{i=1}^{M} Y_i \log \phi_i$$

It is well known that the likelihood achieves its unique maximum when each element of ϕ is set equal to the corresponding sample frequency:

$$\hat{\phi}_i = \frac{Y_i}{n}$$

The properties of this estimator are discussed by Rao (1965, pp. 291-295). In the parametric case just discussed,

$$\phi = A\beta$$

so the log-likelihood of β is

$$\log L(\beta) = \sum_{i=1}^{m} Y_i \log(A\beta)_i$$

Let Φ be the set of possible values of ϕ; Φ is the convex hull of the columns of A. Then maximum likelihood estimation of β involves two cases: First, if

$\hat{\phi}\epsilon\Phi$, then the likelihood of β attains its maximum for any β that implies the observed frequencies, $\hat{\phi}$, that is, the set $S(A,\hat{\phi})$. The vertices of this set give the extreme values of β for which the likelihood attains its maximum, and from the vertices individual bounds on the maximum likelihood estimates of each β_j can be calculated. Second, sampling variation may cause $\hat{\phi}$ to lie outside Φ. In this case $\hat{\phi}$ cannot be taken as the set of observed frequencies, but must instead be the value of ϕ that maximizes $L(\phi)$ within Φ. There does not seem to be any simple way to solve this problem in non-linear programming. Fortunately, if ϕ is in the interior of Φ, the probability that the second case will arise approaches zero as the sample size increases.

In addition to finding the set of maximum likelihood estimates of β, it is also useful to indicate the potential effect of sampling errors in $\hat{\phi}$ on the set of β's. If the sampling variability of $\hat{\phi}$ can be characterized by a confidence region, R, in Φ, then the set

$$R^* = \bigcup_{\phi\epsilon R} S(A,\phi)$$

takes simultaneous account of the uncertainty about p caused by the sampling variability of $\hat{\phi}$ and that caused by the nature of the relation between ϕ and p. R^* is not a confidence region in the usual sense, however, because a probability on Φ does not induce a probability on the set of all p's. If R is a convex polyhedron with vertices $r^{(1)},\ldots, r^{(m)}$, then

$$R^* = < S(A,r^{(1)}),\ldots, S(A,r^{(m)}) >.$$

Concluding Remarks

This paper has presented three methods for estimating the distribution of an unobserved component from the distribution of an observed statistic. The first method, based on the assumption that the unobserved component has a discrete distribution, has little intrinsic interest, but is mathematically identical to the third method. The second method makes no assumptions about the distribution of the unobserved component, and therefore gives the weakest results. It is probably of more theoretical than practical interest. Our study of it demonstrates that exact but not complete knowledge of the distribution can be obtained, so the assumptions of previous authors can, in fact, be tested. Finally, the third method imposes prior knowledge about the smoothness of the distribution of the unobserved component. This is done through the conventional device of confining the distribution to a family indexed by a limited number of parameters. The third method is probably of greatest potential usefulness in practical work.

References

Yoran Ben Porath and Finis Welch, "Uncertain Quality: Sex of Children and Family Size" Rand Publication R-1117- NIH/RF, Aug. 30, 1972, Revised October 1972

David Gale, The Theory of Linear Economic Models, New York, McGraw-Hill, 1960

Leo Goodman, "Some Possible Effects of Birth Control on the Human Sex Ratio" Annals of Human Genetics, Vol. 25, 1961, pp. 75-81

Zvi Griliches, research in progress.

Christopher Jencks, *et al.*, Inequality: A Reassessment of the Effect of Family and Schooling in America, New York, Basic Books 1972.

Daniel McFadden, A Behavioral Model of Modal Choice, unpublished, 1972

C. Radakrishna Rao, Linear Statistical Inference and Its Applications, New York, Wiley, 1965

F. A. Valentine, Convex Sets, New York, McGraw Hill, 1964.

H. Weiler, "Sex Ratio and Birth Control" American Journal of Sociology, Vol. 65, 1959, pp. 289-299.