RT NO. UMTA-MA-06-0049-78-8

UMTA/TSC Evaluation Series

Sample Design for Discrete Choice Analysis of Travel Behavior

Final Report
July 1978

Service and Methods Demonstration Program

U.S. DEPARTMENT OF TRANSPORTATION
Urban Mass Transportation Administration
and Transportation Systems Center

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
UMTA-MA-06-0049-78-8		

4. Title and Subtitle	5. Report Date
SAMPLE DESIGN FOR DISCRETE CHOICE ANALYSIS OF TRAVEL BEHAVIOR: THE STATE OF THE ART	July 1978
	6. Performing Organization Code
7. Author(s) Steven R. Lerman and Charles F. Manski	8. Performing Organization Report No. DOT-TSC-UMTA-78-31
9. Performing Organization Name and Address Cambridge Systematics, Inc.* 238 Main Street Cambridge MA 02142	10. Work Unit No. (TRAIS) UM827/R8712 11. Contract or Grant No. DOT-TSC-1083
12. Sponsoring Agency Name and Address U.S. Department of Transportation Urban Mass Transportation Administration Office of Transportation Planning Management and Demonstrations Washington DC 20590	13. Type of Report and Period Covered Final Report August 1977-March 1978 14. Sponsoring Agency Code

15. Supplementary Notes *Subcontractor to:
Multisystems, Inc.†
1050 Massachusetts Ave.
Cambridge MA 02138

†Under contract to:
U.S. Department of Transportation
Reseach and Special Programs Administration
Transportation Systems Center
Cambridge MA 02142

16. Abstract

A review of the state of the art in designing samples for discrete choice analysis of traveller behavior is presented. The basic discrete choice analysis framework is reviewed. It is assumed that any sample used is drawn by a process termed stratified sampling, in which the analyst partitions the population based on attributes and choices made, and then selects the fraction of observations taken within each stratum and the total sample size. Observations within strata are drawn at random. Two related problems, determining the distribution of the attributes in the population and estimating the choice probabilities conditional on the attributes, are explored. Various procedures for solving both these problems are detailed. The role of experimentation in extending the range of attributes in the population is explored.

17. Key Words	18. Distribution Statement
Travel Demand Modeling Sample Design Discrete Choice Analysis	DOCUMENT IS AVAILABLE TO THE U.S. PUBLIC THROUGH THE NATIONAL TECHNICAL INFORMATION SERVICE, SPRINGFIELD, VIRGINIA 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	70	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

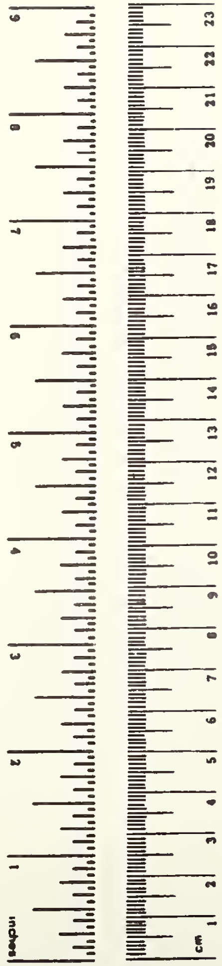
PREFACE

The comments of Carla Heaton, the project technical monitor, and Howard Slavin are gratefully acknowledged. Both these representatives of the Transportation Systems Center made numerous useful suggestions for modifications of the draft report. All errors and omissions are, of course, the responsibility of the authors.

METRIC CONVERSION FACTORS

Approximate Conversions to Metric Measures

Symbol	When You Know	Multiply by	To Find	Symbol
		LENGTH		
in	inches	2.5	centimeters	cm
ft	feet	30	centimeters	cm
yd	yards	0.9	meters	m
mi	miles	1.6	kilometers	km
		AREA		
in^2	square inches	6.5	square centimeters	cm^2
ft^2	square feet	0.09	square meters	m^2
yd^2	square yards	0.8	square meters	m^2
mi^2	square miles	2.6	square kilometers	km^2
	acres	0.4	hectares	ha
		MASS (weight)		
oz	ounces	28	grams	g
lb	pounds	0.45	kilograms	kg
	short tons (2000 lb)	0.9	tonnes	t
		VOLUME		
tsp	teaspoons	5	milliliters	ml
Tbsp	tablespoons	15	milliliters	ml
fl oz	fluid ounces	30	milliliters	ml
c	cups	0.24	liters	l
pt	pints	0.47	liters	l
qt	quarts	0.95	liters	l
gal	gallons	3.8	liters	l
ft^3	cubic feet	0.03	cubic meters	m^3
yd^3	cubic yards	0.76	cubic meters	m^3
		TEMPERATURE (exact)		
°F	Fahrenheit temperature	5/9 (after subtracting 32)	Celsius temperature	°C



Approximate Conversions from Metric Measures

Symbol	When You Know	Multiply by	To Find	Symbol
		LENGTH		
mm	millimeters	0.04	inches	in
cm	centimeters	0.4	inches	in
m	meters	3.3	feet	ft
m	meters	1.1	yards	yd
km	kilometers	0.6	miles	mi
		AREA		
cm^2	square centimeters	0.16	square inches	in^2
m^2	square meters	1.2	square yards	yd^2
km^2	square kilometers	0.4	square miles	mi^2
ha	hectares (10,000 m^2)	2.5	acres	
		MASS (weight)		
g	grams	0.035	ounces	oz
kg	kilograms	2.2	pounds	lb
t	tonnes (1000 kg)	1.1	short tons	
		VOLUME		
ml	milliliters	0.03	fluid ounces	fl oz
l	liters	2.1	pints	pt
l	liters	1.06	quarts	qt
l	liters	0.26	gallons	gal
m^3	cubic meters	35	cubic feet	ft^3
m^3	cubic meters	1.3	cubic yards	yd^3
		TEMPERATURE (exact)		
°C	Celsius temperature	9/5 (then add 32)	Fahrenheit temperature	°F

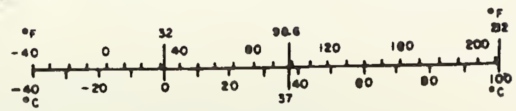


TABLE OF CONTENTS

GLOSSARY OF SYMBOLS

t	denotes a decision-maker
i	denotes an alternative
z_{ti}	vector of variables characterizing decision-maker t and alternative i
T	population of decision-makers
C	choice set
z_t	matrix consisting of vectors z_{ti} for all $i \varepsilon C$
Z_o	collection of attribute matrices z_t faced by all decision makers in T
Z	complete attribute space, of which Z_o is a subset
$f(i,z)$	joint generalized probability density of i and z in population
$P(i\|z)$	the choice model predicting probability i is chosen given z
$p(z)$	generalized probability density of z in population
θ^*	a vector of unknown parameters
θ	an estimate of θ^*
$\tilde{p}(z)$	the generalized probability density of z after a policy change
$Q(i)$	expected fraction of population choosing alternative i
$\tilde{Q}(i)$	the expected fraction of population choosing i after a policy change
$(CxZ)_b$	the b-th subset of CxZ
B	the number of subsets of CxZ
H_b	the fraction of the sample drawn from the b-th subset of CxZ
H	the vector $(H_1, H_2, \ldots H_b, \ldots H_B)$
N	the total sample size
N_b	the number of observations drawn from the b-th subset of CxZ

T_b	the sub-population of T in the b-th subset of CxZ
(i_n, z_n)	an observation from the population corresponding to the n-th observation drawn from the b-th subset of CxZ
F_b	the fraction of the population who are members of T_b
F	the vector $(F_1, F_2, \ldots F_b, \ldots F_B)$
L	the likelihood of a stratified sample
L_r	the likelihood of a random sample
L_e	the likelihood of an exogenous sample
L_c	the likelihood of a choice-based sample
Z_b	the b-th subset of Z defined for an exogenous sample
C_b	the b-th subset of C defined for a choice-based sample
g(z)	the sample distribution of z in an exogenous sample
y(z)	a function mapping old attribute values into new ones
z_k	the k-th entry in the attribute vector z
$\hat{Q}(i)$	an estimate of $\tilde{Q}(i)$ from a sample enumeration
$\Psi(z)$	cumulative distribution of attributes in the population
$\Psi(z\|b)$	cumulative distribution of attributes in b-th subpopulation
$\hat{\Psi}(z\|b)$	empirical distribution in subsample drawn from $(CxZ)_b$
x(z)	any vector valued function of the attribute vector z
E(x)	the expected value of x
$E(x\|b)$	the expected value of x in the b-th subpopulation
$\|C\|$	the number of alternatives in the choice set
z_m	the population median for z
z_o	any particular value of z
Y	the dependent variable in a linear model
β	a vector of parameters in a linear model

ε	a disturbance term
$V(\varepsilon)$	variance of ε
σ^2	an unknown scalar multiplier of variance of ε in the linear model
G	a known matrix, where $V(\varepsilon) = \sigma^2 G$ in the linear model
D	a function which is strictly increasing in each of its arguments
$z_i - z_j$	the difference between the attribute vectors for the i-th and j-th alternative
ϕ^*	a vector of parameters with same number of entries as $z_i - z_j$
γ^*_i, γ^*_j	alternative specific constants for the i-th and j-th alternatives respectively
U_i	the random utility of the i-th alternative
I	an identity matrix of dimension $\|C\|-1$
$V(z)$	the expected sampling variance of attributes across alternatives
z_{ij}	$z_i - z_j$
Z_1	the post-experimental range of z
θ_1^*	a subset of θ^*, the parameter vector
$\tilde{f}(i,z)$	the post-experimental distribution of z in the population.

EXECUTIVE SUMMARY

In the past five to ten years, significant advances have been made in the development of discrete choice models for travel demand analysis. Discrete choice models represent the choices of individuals among alternatives such as modes of travel, auto types and destinations. As these models (such as multinomial logit) continue to be applied in practice, there is a growing need for a coherent theory of how data should be used in discrete choice analysis and for practical guidance in the collection of such data.

The problems associated with designing samples are exemplified by the Urban Mass Transportation Administration's Service and Methods Demonstration Program. Under this program, changes in the transportation services provided in an urban area are made, and the resulting shifts in level of service and user response are monitored. Data is collected in such experiments in order to evaluate the impacts of the changes and to generalize the results to other situations. The problem of how to collect useful data in a cost-effective manner is critical in such evaluation efforts.

This paper is an effort to synthesize the state of the art in sample design for one major aspect of evaluating transportation system change, traveller behavior. The paper focuses on discrete choice analysis of travellers' decisions. It incorporates recent published and unpublished theoretical findings as well as some new results. In addition, it addresses the practical concerns which arise in designing and using data samples in travel demand analysis.

A basic assumption of discrete choice analysis is that within any population, it is possible to characterize any individual by both a list of attributes such as income, auto ownership, travel time, and costs by various modes, etc., and an actual choice of a discrete alternative. Throughout the paper, it is presumed that the primary motivation for sampling is to learn something about the characteristics of the population, the attributes of the alternatives individuals face, and the choices they make. A central hypothesis in discrete choice analysis is that there is a causal link, in which the probability of each individual choosing any particular alternative (termed a choice probability) depends on his/her attributes; changes in the attributes will therefore change the choice probabilities.

Given these assumptions, the goal of any particular data collection scheme is fairly clear. The analyst seeks to learn about (1) the distribution of attributes in the population, and (2) the choice probabilities for the population. For example, the most traditional approach of survey data collection for transportation planning has been the home interview survey. This provides estimates of the distribution of attributes such as income, auto ownership, household size, age, sex, race, etc., in the metropolitan area. This data, along with level of service estimates (typically derived from skim trees), provides a relatively complete estimate of the distribution of attributes in the population. This data can then also be used to infer choice probabilities for the population using models such as multinomial logit, in which the probability of any member of the population selecting an alternative (e.g. driving alone, carpooling, and using transit) depends on the attribute values.

The home interview survey, however, is just one of a number of sampling strategies of potential use in inferring both the distribution of attributes and choice probabilities. Given this, the paper considers four interrelated questions:

1. What different sampling strategies for discrete choice analysis exist?

2. How can different sampling strategies be used to estimate the distribution of attributes in the population?

3. How can different sampling strategies be used to estimate choice probabilities?

4. What is the role of experimentation in improving travel demand analysis?

Each of these questions is considered below.

1. What different sampling strategies exist?

The review deals with a very broad class of sampling strategies (termed stratified sampling), which includes as special cases random sampling, exogeneous sampling and choice-based sampling. In stratified sampling, the data sample is assumed to be obtained by the following four steps:

a) Divide the entire population into groups based on both their attributes and the decisions made.

b) Choose how many people are to be sampled from each group.

c) Within each group, sample the preset number of people at random.

d) For each person, observe his/her attributes and the choice he/she made.

It is important to note that the strata into which the population is divided can be defined by both attributes and choices. In mode choice, for example, one stratification might be to split the population into high and low income travellers; another might be to define transit users and highway users as distinct groups. Stratifications based on combinations of attributes and choices are also feasible.

The first, and most widely used special case of general stratified sampling is random sampling, in which the entire population is a single stratum. The second is exogenous sampling, (e.g., home interview surveys), in which the stratification is based solely on attributes, not on actual choices. The third is choice-based sampling (e.g., on-board surveys), in which the stratification is based on choices but not attributes.

It is important to stress that in stratified sampling, one can control two items, the definition of the strata, and the size of the sample within each stratum. The analyst does not control which decision-makers are actually sampled in each stratum, since these are drawn at random.

2. How can different sampling strategies be used to infer the distribution of attributes in the population?

There are two general approaches to using stratified samples to estimate the distribution of attributes in the population. The simple, less general method is to constrain the sample design such that the fraction of observations in each stratum equals the corresponding population fraction. In this case, the resulting stratified sample can be used as "representative" of the population.

The second more general approach involves use of a simple probability statement to solve for the population attribute distribution as a weighted sum of the attribute distributions within the strata. The weights in this case are the fraction of the population in each stratum.

In both these approaches, the analyst must know the share of population in each stratum. At least four approaches to determining these shares are available.

a) Use existing data sources which yield direct information on strata shares (such as the census for geographical-based stratifications)

b) Use a random sample (such as a telephone survey) to estimate the share of the population in each stratum.

c) Use published statistics and solve a set of linear equations derived from probability theory for population shares.

d) Estimate the population fractions simultaneously with the choice model.

Each of these techniques has both strengths and weaknesses described in the paper, but some may not be applicable to all situations.

3. <u>How can different sampling strategies be used to estimate the choice probabilities</u>?

In almost all cases of practical interest, the problem of determining choice probabilities reduces to one of estimating, or calibrating, the parameters of some model. For example, in mode choice analysis, the choice probabilities might be represented by the multinomial logit model, and the coefficients of the model would have to be estimated.

A number of significant theoretical advances have been made in this area. Perhaps the most significant conclusion is that, <u>under certain technical restrictions, any stratified sample can be used to estimate discrete choice models</u>. This includes random, exogenous, choice-based, and mixed sample designs.

Given that any stratified sample may be used, a related question is how to select the "best" sample design. Major observations in this area include the following:

a) The existing literature on sample design is based on the classical criterion of minimizing the variance of parameter estimates.

b) Mathematical results have been difficult to achieve and are still limited.

c) Prior information about the shares of the population in the strata and the distribution of attributes in the population can improve parameter estimates.

d) The "best" sample design (by the classical criterion) depends on the true parameter values, which are obviously unknown a priori. This contrasts with the standard linear regression model, in which an optimal sample design can be easily derived and does not depend on the actual model parameters.

e) Limited Monte Carlo tests suggest that for binary choice models estimated with choice-based samples, it is advantageous to make sample shares close to 1/2, and that prior knowledge of the share of the population choosing each alternative is very valuable.

f) If one wants to choose a sample to test the hypothesis that a particular model as a whole is more informative than a model in which the choice probabilities for every individual equal the corresponding population shares, the best sample includes decision-makers facing widely disparate alternatives. This result does not apply to samples designed to test other hypotheses.

4. What is the role of experimentation in improving travel demand forecasts?

A significant problem in discrete choice analysis is that under some conditions it is impossible to estimate certain parameters. This situation (termed non-identification), can arise in four important ways:

a) The alternatives everyone faces are homogeneous along some attribute. For example, in analyzing taxi users' choice of which cab company to call for service, it would be impossible to determine the effect of fare differences (with corresponding variations in service quality) across taxi operators. Due to local regulatory policy, all companies provide roughly homogeneous service at the same price.

b) An attribute of a particular alternative is constant for all decision-makers. For example, in most cities, transit systems do not offer any demand responsive services, while the auto mode is by its very nature demand responsive. In this case, it would be impossible to estimate mode choice probabilities for route deviation bus service.

c) Two attributes are perfectly correlated. For example, in many small cities where congestion is minimal, taxi fare (based on meters) may for all practical purposes be perfectly correlated with in-vehicle travel time. Car times and operating costs would be similarly correlated. Thus, it would be impossible to distinguish between the effects of time and cost in a choice model.

d) Alternatives are unavailable. For example, some cities do not have any transit service at all, and the demand for such service would be impossible to determine.

It is important to point out that carefully thought out experiments can be used to create situations in which previously unestimable parameters of discrete choice models can be estimated. Many of the current Service and Methods Demonstration projects serve precisely this function. By changing current attributes for a small group within a larger population and estimating a discrete choice model, it is then feasible to forecast how the entire population will respond to area-wide implementation.

Experiments can also provide a way to achieve greater confidence in parameter estimates. In many situations, some parameters are technically identified, but the amount of variation in the data is too low to make precise parameter estimates feasible.

Some Practical Considerations

All of the above discussion is based on theoretical results. It is important to emphasize that given the current state of the art, there is no general rule for selecting the best sample design for discrete choice analysis problems. In fact, given the analytic intractability of many of the sample design problems, it is unlikely that a rule for optimal sample design will be found in the near future. However, some general practical guidelines can be proposed:

a) It is obvious that the first concern in designing a sample must be to assure that a model can be estimated at all (i.e., that the model parameters are identified in the sample).

b) The duration of experiments must be carefully considered. Existing discrete choice models are for the most part static in structure, and any dynamic effects in the period between implementation and response cannot be reflected.

c) The classical statistical framework should not be applied dogmatically. In most cases, the analyst has more information than classical statistical analysis presumes. For example, in most cases, some a priori statement about the sign and magnitude of certain parameters will be possible. This information should be used if only in an intuitive way.

d) A particular sample may be used for estimation of both the attribute distribution and the choice probabilities. A "good" sample design must balance these uses.

e) In many cases, the idealized stratified sample may be difficult to obtain. In particular, stratified sampling requires the ability to identify the stratum to which a decision maker belongs and the ability to draw at random from each stratum. The last requirement is often violated in common survey practices such as roadside interviews, on-board surveys, and mailback questionnaires.

It is important to note that many of the results reported here are quite recent, and that further work will undoubtedly resolve some of the questions raised in the report. Discrete choice analysis is still a quickly growing area of knowledge, and further work on sample design problems will hopefully make more precise statements about alternative sampling strategies possible. In particular, further work in classical sample design analysis, non-classical sample design criteria (e.g., use of Bayesian analysis), further Monte Carlo studies, and a broader base of actual experience with different stratified sampling rules should yield greater insight into sample design for discrete choice analysis.

1. INTRODUCTION

This paper summarizes recent advances in the theory of sample design for discrete choice analysis and discusses the relevance of these advances for the practice of travel demand forecasting. The objective of the summary is to provide a general framework for analyzing existing data, and designing new samples.

The issues that arise in designing useful, cost-effective samples are exemplified by the Urban Mass Transportation Administration's Service and Methods Demonstration Program. The types of transportation system changes introduced under this program influence both the performance of the transportation system and the population's response to that system. In evaluating such projects, a relatively large base of data must be collected, and the cost and accuracy of the resulting samples may critically influence the success or failure of the entire evaluation effort.

Moreover, in many demonstrations, the impact of the transportation system change is confounded with the effect of other, exogenous changes which occur over the demonstration period. This is particularly the case when the population's response to the system change is being measured. For this reason, the data collection strategy must often provide an adequate base to support a multivariate analysis of traveller response. Discrete choice analysis has enormous potential for providing evaluations of population response in such situations.

In developing this review, an effort is made to recognize the distinction between the idealizations imposed by formal theory, and the real world issues arising in practice. Discrete choice models rest on a set of assumptions about the population and its behavior. In interpreting the theory of sample design, one must keep in mind that this theory relates to the idealized world of formal analysis. The application of sample design theory to actual travel demand forecasting problems is, if anything, an art.

Our first task, undertaken in Section 2, will be set to out the idealized probability model assumed in formal discrete choice analysis. The analyst will usually attempt to specify a model which accurately represents the actual population of interest, but at the same time is simple enough to provide a useful tool for forecasting. In most applied contexts, however, the analyst's knowledge of the actual population will not suffice to totally specify a satisfactory probability model and its parameters a priori. From this, the purpose of data collection emerges, that is to allow one to learn more about the population and consequently to improve one's ability to forecast how that population will respond to transportation system changes.

Our second task, addressed in Section 3, will be to describe some of the alternative sampling rules that can be used in data collection. Attention will be focused on rules in which the population is stratified in some way, and observations are then drawn at random within stratifications. This wide class of sample designs includes almost all currently used methods in transportation planning. For example, a home interview survey is typically performed by sampling randomly from the entire relevant population; on-board surveys are random samples from the stratum of transit users.

For the purposes of this review, it is useful to separate the travel demand analysis process into two phases. First, there is a <u>population description phase</u> in which one formally characterizes the decision making population and the travel alternatives its members face. This characterization can include distributions of socioeconomic attributes such as income, auto ownership, or household size, as well as level of service variables such as time and cost. Second, there is a choice modelling phase in which one specifies a model of travel behavior (e.g. logit or probit), and estimates its unknown parameters. A need for data samples may appear in both of the phases. Sections 4 and 5 respectively, consider in detail the sample design problem that arises in the two phases of travel demand analysis. In these sections the known theoretical results on sample design are collected, and those respects in which sampling must remain an art are articulated.

Usually the travel demand forecasting process draws its data from observations of travel behavior under whatever travel environment happens to prevail at the time of data collection. Sometimes, however, the existing travel environment does not contain a range of attributes sufficiently varied to permit inference of travel behavior to proceed. In such circumstances, an additional phase may usefully be added to the analysis process. This is an experimentation phase in which the existing travel environment of a subset of the population is artifically modified so as to create the variation in travel alternatives needed to support behavioral modelling. In Section 6, we examine issues in the design of such experiments and their role in the forecasting process. Directions for future research are indicated in Section 7.

The theoretical results on sample design reported in this paper are drawn from a number of sources. In particular, we draw heavily from Lerman, Manski and Atherton (1975), Manski and Lerman (1977), Manski and McFadden (1977), and Cosslett (1977). Some of the work presented here is new and has not previously been reported.

2. PROBABILITY MODEL

The probability model underlying modern discrete choice analysis of travel behavior has been laid out in a general form in Manski and McFadden (1977) and is summarized here.

It is assumed that an idealized decision making population, representing the actual population of interest, has been defined. Each member of this idealized population faces a common, finite set of travel alternatives. Let T designate the population and C the choice set. With each decision maker $t \epsilon T$ and alternative $i \epsilon C$ there is associated a vector z_{ti} which characterizes the decision maker and the alternative. Let $z_t = (z_{ti}, i \epsilon C)$ be the matrix of attributes characterizing decision maker t's choice set and let $Z_o = (z_t, t \epsilon T)$ be the collection of attribute matrices faced by the various decision makers in T. Finally, let Z denote the attribute space, in which Z_o is a subset of Z. That is, Z is a collection of attribute matrices including at least those currently faced by decision makers. Usually, Z will be defined so as to encompass attributes that might be found among the population in the future as well as those included in Z_o.[1]

In actual applications, the definition of T (the decision-making population), C (the choice set) and Z (the attribute space) varies considerably. For example, in an analysis of mode choice for work trips, T may include all workers travelling on a particular day to or from their place of employment. The choice set modes such as driving alone, transit, and carpooling. The attribute space may include times, costs, etc., for each mode, socioeconomic characteristics such as income and auto ownership, as well as functions of both these

[1] An integral part of the sample design problem is to decide what attributes of decision makers and alternatives should be obtained in the data collection process. Thus, the structure of the attribute space Z is under the potential control of the analyst. This aspect of sample design will not be discussed in this paper. Instead it will be assumed that a structure for Z has somehow been chosen.

types of attributes.

Using this notation, the basic probabilistic assumption is that the frequency distribution of choices (i), and attribute matrices (z) in the actual population can be characterized by a generalized probability density

$$(1) \quad f(i, z) \equiv P(i|z)\, p(z)$$

defined over C x Z.

In discrete choice analysis, the decomposition of the joint density f(i,z) into the product of the conditional probability P(i|z) and the marginal density p(z) is of particular importance. In discrete choice analysis, P(i|z) is not simply a conditional probability; it is rather the probabilistic prediction of a behavioral model describing how a decision maker with associated attributes z would select among the alternatives in C. For the above reason, P(i|z) is often termed a "choice probability". In most applications, the behavioral model generating P(i|z) is a priori specified by the analyst to be a member of a parametric family, implying that P(i|z) itself is known up to this family. For example, one might assume that the choice probabilities have the conditional logit form, where

$$P(i|z) = \frac{e^{z_{ti}\theta^*}}{\sum_{j \varepsilon C} e^{z_{ti}\theta^*}}$$

where θ* is a vector of unknown parameters. In this case, the choice probability may be written as P(i|z, θ*).

Consider now four basic assumptions of the above, general model. One seemingly restrictive assumption, namely that all members of T face the same choice set C, is actually innocuous. To see this, observe that the attributes z_t can vary with each individual t. Also, observe that if some alternative j is "unavailable" to a decision maker t, this fact can be reflected in the value taken by z_t, and we can set $P(j|z) = 0$ in this case. For example, the alternative of driving alone is for obvious reasons generally assumed to be unavailable to travellers without access to an automobile. This can be reflected in the choice model $P(j|z)$, where j = driving alone, by defining one attribute in z to be auto availability, and defining $P(j|z) = 0$ when this attribute is zero.[1]

In most applications, the decision making population of interest is relatively large. It then becomes analytically convenient, and basically innocuous, to let the idealized population T be infinite so distinctions between sampling with and without replacement can be ignored. With this second assumption of T possibly infinite, it is natural to characterize the distribution of attributes in the population by a generalized probability density p(z).[2]

As Manski and McFadden (1977) emphasize, the application of discrete choice analysis does require one crucial assumption not imposed in the general statistical analysis of discrete data. This is the postulate (typically derived from some behavioral theory) that the probability $P(i|z)$ reflects a "causal" link between the independent variables z and the choice of any alternative $i \varepsilon C$. Moreover, it is implicit in the use

1. At the extreme, some alternative may be currently available to no decision maker in T. Nevertheless, it would still be desirable to formally include such an alternative within the choice set if the alternative might become available in the future.

2. A generalized probability density is simply a mixture of a probability distribution assigning positive probability to a finite set of points in Z and an ordinary probability density function over Z. This assumption, like that of infinite T, can generally be accepted without concern.

of the model to make forecasts that this link will continue to hold even if the joint distribution of (i,z) pairs in the population is changed.

This third assumption provides the basis for the use of discrete choice analysis as a tool in travel demand forecasting. Under the behavioral postulate, changes in transportation policy may modify people's travel environments, as expressed in the attribute distribution p(z), but do not change their behavior, as expressed in the choice probabilities $P(i|z)$. Thus, if the choice probabilities are known to the analyst, and if the effect of a policy change on the attribute distribution can be determined, the effect of that policy change on travel choices can be predicted.

For example, assume that the initial attribute distribution is p(z), that a proposed policy change will modify it to $\tilde{p}(z)$, and that we wish to predict the effect of the policy change on the fraction of the population making various travel choices. In more concrete terms, p(z) could include the distribution of travel cost in the population before some pricing change, and $\tilde{p}(z)$ would be the same distribution after the change. For any alternative $i \varepsilon C$, the expected fraction initally choosing i is, by definition, $Q(i) \equiv \int_Z P(i|z)\, p(z)\, dz$. Given the behavorial postulate, the predicted post-policy fraction is:

$$\tilde{Q}\,(i) = \int_Z P(i|z)\, \tilde{p}(z)\, dz.$$

A fourth assumption implicit in discrete choice analysis as currently practiced deserves comment. The probability model we have set out is static, that is, it describes the population's travel environment and behavior in a manner which ignores time. Recently, researchers have begun to develop discrete choice models which explain sequences of choices over time, and attempt to realistically incorporate dynamic aspects of behavior.[1] Because research on dynamic choice analysis is still in its infancy, and because no corresponding literature on sample design has yet developed, this paper confines its attention to the less general but quite rich models of static choice analysis.

With the probability model of discrete choice analysis now laid out and interpreted, the formal objectives of data collection can now be expressed. These are first to learn the form of the attribute distribution $p(z)$, and second to learn the choice probabilities $P(i|z)$. Where, as is usual, the choice probabilities are a priori given the parametric form $P(i|z, \theta^*)$, the second objective reduces to one of estimating θ^*. The design of samples meeting the above objectives will be examined as soon as we have, in the next section, introduced a class of sampling rules suitable for investigation.

---

1 See for example, Heckman (1977).

3. CLASS OF STRATIFIED SAMPLING RULES

Within the general model developed above, one can view the process of sampling as drawing observations on individuals (as described by their attributes z) and their respective choices $i \epsilon C$. Observations are (i,z) pairs, from which the probabilities p(z) and $P(i|z)$ may be learned. For example, in a typical application, an observation might consist of a traveller with a known mode choice among carpooling, driving alone, and transit, along with associated attributes such as times and costs on each mode, as well as the traveller's socioeconomic characteristics.

In describing alternative sampling rules, we shall first present a relatively abstract theoretical development and then illustrate that theory with a brief example. Finally, we shall consider the most useful practical cases within the class of rules to be considered.

The existing theoretical literature on sample design in discrete choice analysis essentially assumes that such observations are drawn in the following manner:[1]

First, the analyst partitions the set C x Z, consisting of all possible choice-attribute pairs, into a collection of B mutually exclusive and exhaustive subsets $(C \times Z)_b$, b = 1,....B. Such a partitioning is conventionally termed a "stratification" of C x Z.

Second, the analyst selects a set of sampling fractions $(H_b, b = 1, \ldots B)$ such that $\sum_{b=1}^{B} H_b = 1$, and a sample size N.[2]

Then, for each b=1,...B, a total of $N_b = H_b \cdot N$ decision makers are independently drawn, at random, from T_b, the sub-population of T defined by $T_b = (t \epsilon T : (i_t, z_t) \epsilon (C \times Z)_b)$.

[1] See Manski and McFadden (1977), for a formal presentation.

[2] The sampling fractions H_b may be set directly or may themselves be determined by an auxiliary exogenous process. In the former case, we speak of a "single stage" stratification; in the latter case, a "multi-stage" one. Cluster sampling is one type of multi-stage process.

Finally, for each sampled decision maker, the associated choice-attribute pair is observed. Thus, a sample (i_n, z_n), $n = 1, \ldots N_b$, $b = 1, \ldots B$ of such pairs is produced.[1]

Consider again the example of a simple mode choice model in which travellers choose among ride sharing, driving alone and carpooling. Furthermore, assume that time, cost, income and auto ownership are the only relevant variables in the model. In this example, the set C x Z would be all feasible combinations of modes and their attributes.

In Figure 1, the modes in C are rows and the attributes Z are columns.

<u>Display of C x Z</u>

Z - attributes	Time			Cost			Income	Auto Ownership
C - mode used	mode 1	mode 2	mode 3	mode 1	mode 2	mode 3		
mode 1 - Carpooling								
mode 2 - Drive Alone								
mode 3 - Transit								

1 It would be notationally more proper to write (i_{n_b}, z_{n_b}), $n_b = 1, \ldots N_b$, $b = 1, \ldots B$. For simplicity, the subscript b on n is omitted.

Thus, Z consists of eight separate components, seven of which are for all practical purposes continuous and one of which, auto ownership, is discrete. Any particular observation of an (i,z) pair in $C \times Z$ would consist of the mode chosen also the eight modal and socioeconomic attributes of the sampled individual.

The simplest possible stratification of $C \times Z$ would be to define only one stratum consisting of all of $C \times Z$. The corresponding sample would be drawn randomly from this single strata.

A second stratification rule might define three income groups, as follows:

$(C \times Z)_1$ = all (i,z) pairs with income $\leqslant$ \$7500

$(C \times Z)_2$ = all (i,z) pairs with income between \$7500 and \$15,000

$(C \times Z)_3$ = all (i,z) pairs with income $\geqslant$ \$15,000

Still a third possible stratification might be to define modal users groups, such as:

$(C \times Z)_1$ = all carpool users

$(C \times Z)_2$ = all drive alone users

$(C \times Z)_3$ = all transit users.

Further examples can include mixtures of the above such as follows:

$(C \times Z)_1$ = all transit users with income $\leq$ \$7500

$(C \times Z)_2$ = everyone else

In each of these examples, B would equal the number of strata, and T_b, b = 1, ... B would be the relevant subsets of the population.

The class of stratified sampling rules in general and the most relevant special cases in particular offer an enormous range of sample design possibilities to the analyst. Of course, many interesting rules lie outside the stratified class. Some attention has recently been given to samples created by mixing stratified sub-samples of different types. In particular, the so-called "enriched" samples, in which a sample of users of one alternative in C and a random sample are combined, have been studied by McFadden (1977), and Cosslett (1978a).

There may in some circumstances be advantages to using sampling strategies in which the sample size is determined as part of, rather than prior to the actual data collection. While sampling with so-called "informative stopping rules" has been analyzed in many statistical contexts (in particular, see DeGroot (1970)), no research on the use of such rules in discrete choice analysis has been performed.

The discussion of sample design in this paper concerns itself exclusively with sampling rules of the stratified class. Before introducing those special cases of this class, which have been found most useful for applications, three general observations are in order.

First, it is important to distinguish what aspects of the sampling process the analyst does and does not control. What he <u>does</u> control is the stratification $(C \times Z)_b$, $b = 1,\ldots B$ and the number of decision makers N_b to be drawn from each sub-population T_b. What he does not control are the identities of the decision makers then drawn. These drawings are to be independent and at random. Thus, the sample likelihood is:

$$(2) \quad L = \prod_{b=1}^{B} \prod_{n=1}^{N_b} \frac{f(i_n, z_n)}{F_b} \cdot H_b$$

where F_b is the fraction of the population T who are members of T_b. To see that equation (2) is the sample likelihood, consider the likelihood of any observation drawn via a stratified sampling rule. This is the probability that the stratum $(C \times Z)_b$ containing the observation is selected, times the conditional likelihood of drawing the observed (i,z) pair out of this stratum. The former probability is the sampling fraction H_b. The latter conditional likelihood is $\frac{f(i,B)}{F_b}$. Since observations are drawn independently, equation (2) follows. We note for later use that the population fraction F_b can be expressed as an integral of the joint distribution f(i,z), over the subset $(C \times Z)_b$, that is:

$$F_b = \int_{(C \times Z)_b} f(i,z)d(i,z)$$

Second, one should understand why samples produced by stratified sampling rules yield information about $P(i|z)$ and $p(z)$. The reason, very simply, is that the sample likelihood (2) is a function of the density values $f(i_n,z_n)$, $n=1,\ldots N_b$, $b=1,\ldots B$, and of the population fractions F_b, $b=1,\ldots B$; these values are, of course, functions of P and p through equation (1). Note that the sample likelihood also depends on the stratification imposed, on the sampling fraction H_b, b=1,...B, and on the sample size N. The sample design theory discussed in this paper can usefully be viewed as the study of how these control variables should be chosen, so as to yield likelihoods with desirable properties.

Our third observation is an operational one. In order to apply a stratified sampling rule, the analyst must have a viable procedure for sampling at random within each of the sub-populations T_b, b = 1,...B. This requirement is sometimes difficult to meet, either because there are problems in effectively separating the various sub-populations from one another, or because a suitable mechanism for selecting decision makers independently at random is elusive. These operational concerns in sampling will be discussed further in Section 4.1.1.

Among the class of all stratified sampling rules, three types are of particular applied interest. These are:

3.1.1. Random Sampling: The stratification of C x Z is the trivial case in which the entire population is a single stratum. In this case,

$$B = 1 \text{ and } (C \times Z)_1 = C \times Z.$$

Then $F_1 = \int_{C \times Z} f\,(i,z)d(i,z) = 1$, and $H_1 = 1$, so the sample likelihood (2) reduces to

$$(3) \quad L_r = \prod_{n=1}^{N} P(i_n|z_n)\; p(z_n).$$

3.1.2. Exogenous Sampling: The analyst partitions the attribute space Z into mutually exclusive and exhaustive subsets Z_b, b = 1,...B, and lets $(C \times Z)_b$ = $C \times Z_b$. That is, the pair (i,z) is included in stratum $(C \times Z)_b$ if and only if $z \varepsilon Z_b$ and the identity of i is not used in defining the stratum. Then

$F_b = \int_{C \times Z_b} f(i,z)d(i,z) = \int_{Z_b} p(z)dz$ and the sample likelihood becomes:

$$(4)\ L_e = \prod_{b=1}^{B} \prod_{n=1}^{N_b} \frac{P(i_n|z_n)\cdot p(z_n)H_b}{\int_{Z_b} p(z)\,dz}$$

This case corresponds to the example of income stratification given above, where the subsets Z_b are the three income classes and choice of mode does not affect the stratum to which a member of the population belongs.

3.1.3. Choice Based Sampling: The analyst partitions the choice set C into mutually exclusive and exhaustive subsets C_b, $b = 1,\ldots B$ and lets $(C \times Z)_b = C_b \times Z$. That is, (i,z) belongs to $(C\times Z)_b$ if and only if $i\varepsilon C_b$. Then $F_b = \int_Z \sum_{i\varepsilon C_b} f(i,z)d(i,z) = \int_Z \left(\sum_{i\varepsilon C_b} P(i|z)\right) p(z)dz$ and the resulting sample likelihood is:

$$(5)\ L_c = \prod_{b=1}^{B} \prod_{n=1}^{N_b} \frac{P(i_n|z_n)\ p(z_n)}{\int_Z \left(\sum_{i\varepsilon C_b} P(i|z)\right)p(z)\ dz} \cdot H_b \cdot$$

This case corresponds to the example of stratification by modal user groups given above, in which attributes of the modes or decision-makers do not enter into the strata definitions.

Let us examine these three types of sampling rules. Random sampling, the simplest, is a fully specified rule. That is, once the analyst is committed to random sampling, he exercises no further control over the data collection process.

In exogenous sampling, the analyst, through stratification of Z and selection of the sample fractions H_b, partially controls the sample attribute distribution $\frac{p(z)}{\int_{Z_b} p(z)dz} \cdot H_b \cdot$

Let g(z) designate this sample distribution. Then the exogenous sampling likelihood can be written in the familiar form

$$(6)\ L_e = \prod_{b=1}^{B} \prod_{n=1}^{N_b} P(i_n|z_n)g(Z_n)$$

It has been common, although not strictly correct, to assert that under exogenous sampling, the analyst has full control over the sample distribution g(z), and to represent the sample design problem as one of selecting among alternatives such as distribution.[1] As the above indicates, however, exogenous sampling really offers more limited design possibilities, in that only H_b and the stratification can be controlled, not the entire distribution g(z). In particular, the drawing of observations out of each stratum is done at random, and hence is not under the analyst's control.

Note that the exogenous sampling likelihood reduces to the random sampling one if the analyst sets $H_b = F_b = \int_{Z_b} p(z)\, dz$, all b = 1, ...B or, alternatively, samples so that the density of g will be g(z) = p(z), for all $z\varepsilon Z$. Finally, we remark that the transportation home interview survey is often cited as an example of exogenous sampling. While this example is often apt, it is not always proper. In particular, if the choice being analyzed is that of residential location, then the geographic stratification used in home interview surveys is choice based rather than exogenous.

Choice based sampling rules give the analyst control over the frequencies with which the various alternatives in C appear in the sample. The most refined form of choice based sampling is that in which each alternative in C defines a separate stratum. In this case, B is the choice set size and b=1, ...B indexes the alternatives in C. For this stratification, the choice based sampling likelihood may be written as follows:

$$(7)\ L_c = \prod_{i\varepsilon c} \prod_{n=1}^{N_i} \frac{P(i\ z_n)p(z_n)}{\int_Z P(i\ z)p(z)dz} \cdot H_i.$$

The choice based sampling likelihood, like the exogenous sampling one, reduces

[1] See, for example, Lerman, Manski and Atherton (1975).

to that of random sampling in special cases. Specifically, this occurs whenever

$$H_b = F_b = \int_Z \left(\sum_{i \in C_b} P(i|z) \right) p(z)dz \quad \text{all } b = 1,\ldots B.$$

On board and roadside surveys are often cited as examples of choice based sampling in transportation. As indicated before, however, even home interview surveys may in some contexts be choice based.

It is of interest to observe that the sample likelihood associated with random sampling may be achieved using other stratified sampling rules as well. In particular, it is easy to see that if the analyst choose any stratification and sample composition such that $H_b = F_b$, all $b=1,\ldots B$, then the general stratified sampling likelihood (2) reduces to the random sampling one (3). Since the sample likelihood embodies all information in the data sample, all stratified rules satisfying the vector condition $H = F$ where $H=(H_b, b=1,\ldots B)$ and $F = (F_b, b=1,\ldots B)$ are statistically equivalent.

4. SAMPLE DESIGN FOR DESCRIPTION OF ATTRIBUTE DISTRIBUTION

The preliminaries for our discussion of sample design in discrete choice analysis have now been completed, and we turn in this section to the problem of learning the attribute distribution characterized by the density $p(z)$.[1] Organizationally, it is convenient to first treat this problem within the idealized world of theory, then address the practical concerns that arise from incompleteness of the theory and divergences between the idealized and real worlds. This same sequence of presentation will be followed in the next section when the problem of learning the choice probabilities is discussed.

4.1 THEORETICAL RESULTS

Let us first recall the reason why we should like to know the attribute distribution, and clarify the sense in which this distribution is to be learned.

The travel demand forecasting process requires knowledge of the present attribute distribution, in order to determine the distribution that would prevail after a hypothesized policy or environmental shift has occurred. Let us be precise. In discrete choice analysis, a policy shift is simply a function changing each decision-maker's present attribute value into some new value. If $p(z)$ is the current attribute density and if $y(z)$ is the function mapping old into new attribute values, then clearly p and y together determine $\tilde{p}$, the post-policy attribute density. For example, the attribute vector z might include as its k^{th} entry, z_k, the transit fare, which might be 25¢ currently.

---

[1] A technical note is required here. The distribution of attributes in the population may be expressed through a cumulative distribtution function or through its derivative, the probability density function. While the likelihood of an observation is defined in terms of the density function, it turns out that for travel demand analysis, it is the distribution function, not the density, that must be learned. See Section 4.1.1 for further elaboration of this point.

If the only policy change being evaluated were a doubling of fare, then $y(z_k)$ would reduce to $y_k = 2z_k$, and the corresponding post-policy distribution of the entire attribute vector, $\tilde{p}(y)$, could readily be derived.

Once $\tilde{p}$ and the (time invariant) choice probabilities $P(i|z)$ are known, all consequences of the policy shift can be determined. For example, the new expected share of the population choosing alternative i is defined by

$$\tilde{Q}(i) = \int_Z P(i|z)\tilde{p}(z)dz$$

Because the attribute distribution simply describes the existing travel environment and is not derived from any causal model, it is generally assumed in discrete choice analysis that one knows little, if anything, about the form of p(z) a priori. In particular, unlike the behaviorally derived choice probabilities, the attribute density is usually not specified to be a member of any parametric family. Thus, learning the attribute distribution means learning the whole distribution function, not merely some parameters characterizing this function.[1]

How then may we learn the attribute distribution be learned. Two approaches, both of which are correct in theory and useful in practice shall be discussed.

4.1.1. The Representative Sample Approach

The simpler but less powerful of the two approaches is as follows. Select a sampling rule such that the likelihood of observing any attribute value z on each draw is p(z). Then draw an actual sample of decision-makers according to this rule. Use the resulting distribution of z values in the sample as an estimate of the attribute distribution in the population.

---

[1]Note that it will not, in general, be sufficient to learn lower order moments of p(z), say its mean and covariance matrix. The values we should like to forecast, such as $\tilde{Q}(i)$, depend on the entire density $\tilde{p}$ and hence on all of p.

The theoretical basis for the above procedure is the fact that the empirical distribution of z in the sample will in general converge to the population distribution as the sample size increases.[1] Hence, in large enough samples, the sample distribution of z will be appropriately "close" to the population distribution. The procedure is useful in practice because once the sample is drawn, the analyst can simply treat the sample of decision-makers as if it were the whole population, and produce forecasts using this sample. This is the so-called "random sample enumeration" method of forecasting. To see how this works, consider the problem of forecasting the expected share of the population choosing some alternative i after a policy change. This share, it will be recalled, is $\tilde{Q}(i) = \int_Z P(i|z)\, \tilde{p}(z)\, dz$. Given a sample of N individuals, $\tilde{Q}(i) = \frac{1}{N}\sum_{n=1}^{N} P(i|z_n)$. This estimate will be consistent as long as the sampling rule used satisfies the property described above.[2]

We now must specify sampling rules which do have the desired property that the likelihood of observing z is p(z).

Consider the set of stratified sampling rules in which $H_b = F_b$, all b=1, ...B. These are rules for which the fraction of the sample in each stratum equals the share of the population in that stratum. All rules meeting the condition, it will be recalled, yield the random sampling likelihood in which the likelihood of any (i,z) observation is $P(i|z)p(z)$. Clearly, the marginal likelihood of any z observation is $\sum_{i\varepsilon C} P(i|z)p(z) = p(z)$. Hence, all stratified rules satisfying the H=F conditions are appropriate for learning p(z).

Since the set of sampling rules satisfying the H = F conditions are sta-

[1] See Rao, Section 6f.1, 1973.

[2] Note that construction of the estimate $\tilde{Q}(i)$ only requires the empirical distribution of sample points. It does not require one to estimate the density p(z).

tistically equivalent, the analyst's choice among such rules can, with one strong caveat, be based on considerations of relative sampling costs. The caveat is that, given any stratification H_b, the sampling fraction in stratum b is known and under the control of the analyst, but F_b, the share of the population in stratum b, may not be known. Thus, to devise a sample with H = F, we must select a stratification for which the values of F_b are a priori known. This is certainly a non-trivial requirement, particularly as the F values are themselves generally functions of the unknown p(z) density. (The one circumstance in which F is trivially known, is the case of random sampling where B=1 and F_1=1). Nevertheless, the requirement can often be met in practice. This point will be discussed at length in Section 4.2.2.

4.1.2 Representative Sub-Sample Approach

It is significant that the sampling rules satisfying the H = F condition are not the only ones from which the attribute distribution may be learned. In fact, any stratified rule can be used as long as H_b is made positive whenever F_b is positive. To see this, let $(C \times Z)_b$, b=1,...B be an arbitrary stratification, let $\overline{\Psi}(z)$ be the cumulative attribute distribution in T, and let $\overline{\Psi}(z|b)$ designate the cumulative attribute distribution among decision makers in stratum b. Consider the identity

$$(8) \quad \Psi(z) = \sum_{b=1}^{B} \Psi(z|b) \cdot F_b$$

and observe that in stratified sampling, N_b observations are drawn at random from T_b. The empirical cumulative distribution of z in these N_b observations, designated $\hat{\Psi}(z|b)$, is therefore a consistent estimate of the sub-population attribute distribution $\Psi(z|b)$. From (8), it then follows that $\sum_{b=1}^{B} \hat{\Psi}(z|b) \cdot F_b$ is a consistent estimate of the population attribute distribution $\Psi(z)$.

What this result implies is that consistent estimates of the distribution of socioeconomic characteristics--such as income and auto ownership, as well as level of service variables--such as time and cost, can be recovered from stratified samples as long as the population shares in each stratum are known. For example, suppose one had a roadside interview, and an on-board survey (choice-based samples for the car and transit modes respectively). Using the data from these surveys as empirical distributions, and a priori knowledge of mode shares, equation (8) can be used to estimate the attribute distribution for the entire population.

To see how the above approach may be used in practice, consider again the problem of forecasting the post-policy aggregate share $\tilde{Q}(i)$. Observe that $\tilde{Q}(i) = \sum_{b=1}^{B} \tilde{Q}(i|b) \cdot F_b$ where $\tilde{Q}(i|b)$ is the aggregate share choosing i among the sub-population T_b. Note that $\tilde{Q}(i|b)$ may be consistently estimated by $\hat{Q}(i|b) = \frac{1}{N_b} \sum_{n=1}^{N_b} P(i|z_n)$. Hence, $\tilde{Q}(i)$ may be estimated by $\hat{Q}(i) = \sum_{b=1}^{B} \hat{Q}(i|b) \cdot F_b$.

Four remarks should be made about the above procedure. First, given any stratification, prior knowledge of the F values is required to implement the procedure. Second, random sampling falls within the class of procedures as we may simply set B=1. Third, as long as the shares of the population in each stratum are known, the use of the above procedure does not require knowledge of the choice process; it is based on a simple probability identity which does not involve $P(i|z)$. Fourth, for a given total sample size, the $\sum_{b=1}^{B} \hat{\Psi}(z|b) \cdot F_b$ estimates resulting from different stratifications and sample compositions are not, in general, statistically equivalent.[1] Unfortunately, there exists very little theory to help one select among alternative designs. We shall, however, offer some heuristic guidance on this question.

[1] This point is discussed in Section 4.2.3,

4.2 PRACTICAL CONCERNS

The foregoing theoretical discussion of sampling to learn the attribute distribution is incomplete as a basis for selecting sample designs in practice. First, we have as yet said nothing about the feasibility of implementing stratified sampling rules. Second, given any implementable rule, we have not indicated how the requirement for prior knowledge of the F values, necessary for estimation of the attribute distribution, can be met. Third, we have thus far offered no assistance to the analyst in selecting among those rules which are implementable and whose associated F values can be determined. These practical concerns in sample design are addressed below.

4.2.1 Problem of Implementation[1]

Given a population T, the process of selecting a stratification T_b, b=1, ...B and then sampling at random within each T_b appears deceptively simple. Actually, implementation of this process always requires careful thought and often some theoretical compromise. The reason is that in order to sample at random within a sub-population, one must first be able to isolate this sub-population for purpose of sampling. In practice, such isolation is sometimes difficult to achieve.

Two examples will serve to illustrate the point. Let the population of interest be the set of all people potentially making trips within a metropolitan area. First consider stratification based on place of residence. A home interview survey can easily isolate and sample from the sub-population of potential

[1] The discussion in this section pertains to the problem of estimating the choice probabilities as well as to that of estimating the attribute distribution.

trip-makers who are residents of the area. It is however far more difficult to isolate and sample from the remaining sub-population, namely non-residents.

Now consider the same population, and let the stratification be based on the mode used in trip-making on a given day. Transit users will be relatively easy to isolate because of their physical proximity in transit vehicles and stations. Automobile users will generally be more difficult to isolate as a group. Then, of course, there is the sub-population who make no trip on the given day. Non-trip making residents may be isolated through a home interview survey on that day. How non-trip making non-residents can be sampled is not clear.

The above examples are fairly typical of the practical difficulties that may arise in isolating sub-populations. There do exist some situations in which no practical way can be found to sample from some sub-population. The non-trip making non-residents in the above examples may be such a case. In these situations, the analyst may do one of two things, neither very palatable. First, he may ignore the problematic sub-population, that is define the population T so as to exclude it. Second, he may assume that the attribute distribution in this sub-population is identical to that in some "similar" sub-population which can be sampled.

One further warning should be given to conclude this discussion. When a means of isolating and sampling from a sub-population has been found, care must still be taken to ensure that the sample is drawn at random. How this essential requirement can be satisfied in practice must be determined on a case-by-case basis, but some potential problems can at least be highlighted.

Consider, for example, the following three surveys:

1. an on-board survey of passengers on selected bus routes;
2. a roadside interview at various points in the city;
3. a mailback survey sent to a random selection of households.

In the first survey, the relevant sub-population might be all transit riders. However, the need to choose which routes to survey makes achievement of random drawings of transit users difficult. Some routes may have a high percentage of elderly users, while others may attract primarily workers. Furthermore, if a sample is taken on a single day, some transit users may be interviewed more than once, and such individuals are likely to have very different characteristics than the rest of the sub-population.

The same problems arise in the second example, where the objective would presumably to be to draw randomly from all auto users.

In the third example, the high rejection rate generally associated with mailback surveys makes attainment of random drawing extremely difficult. It is often unlikely that people who choose to respond to mailback questionnaires have the same attribute distribution as the population as a whole.

4.2.2 Determination of Sub-Population Sizes

Given a population stratification there exist at least four distinct ways one might determine the sub-population sizes F_b, b=1, ...B:

(a) direct measurement; (b) estimation from a random sample; (c) solution of a set of linear equations; and (d) estimation with the choice model. These four approaches are described in more detail.

a. Direct Measurement: For some stratifications, the sub-population sizes may be measured directly. Two examples will suffice to illustrate this approach. First, let the population be the set of residents of an SMSA and consider stratification by location of their residence within the SMSA. If each sub-population contains the residents of an integral number of Census tracts, Census population data will provide relatively accurate measure of the sub-population sizes at given points in time. Second, let the population be the set of individuals making work trips within the SMSA and consider a stratification by mode. Rush hour transit fare and highway cordon counts might then provide adequate measures of transit and auto usage on work trips. (Such counts cannot be perfect measures as some trips made during rush hour are not work trips, and some work trips are made at times other than rush hour.)

b. Estimation From a Random Sample: If one draws a random sample of decision makers from T, then the sample distribution of (i,z) pairs is a consistent estimator of the population distribution f(i,z). It follows that for any stratification $(C \times Z)_b$, b=1,...B, the fraction of the random sample who belong to each stratum is a consistent estimate of F_b. Thus, given any stratified rule, the associated F values can be estimated if an auxiliary random sample is available or can be drawn.

One important issue that must be highlighted is that, the cost of a random survey designed solely to determine values of F_b, b=1,...B should not be compared with the costs for random surveys to determine an empirical attribute distribution. The former only requires information from each respondent sufficient to identify the stratum to which he/she belongs. This typically will consist of a small set of socioeconomic characteristics and/or the actual choice $i \varepsilon C$ made. The latter survey requires the full set of attributes, typically including level of service for each alternative in the choice set.

c. Solution from Set of Linear Equations: Recall the identity

(8) $\Psi(z) \equiv \sum_{b=1}^{B} \Psi(z|b) \cdot F_b$

introduced earlier in this section. Let $x(z)$ be any vector valued function of z. Then, letting E designate the expectation operation, it follows from (8) that

$$(9) \qquad E(x) = \sum_{b=1}^{B} E(x|b) \cdot F_b .$$

Imagine that the values $E(x)$ and $E(x|b)$, $b=1, \ldots B$ were known. Then the vector equation (9) plus the identity $\sum_{b=1}^{B} F_b \equiv 1$ would form a set of linear equations in the unknown parameters F_b, $b=1, \ldots B$. In particular, if the x vector has at least $|C|-1$ (where $|C|$ denotes the number of alternatives in C) components, then, given usual linear independence conditions, this set of equations could be uniquely solved for the F_b values.

Observe now that if a stratified sample is drawn, the sample mean of x among those decision-makers belonging to T_b is a consistent estimate for $E(x|b)$. As for the population mean $E(x)$, these values are, for many x functions, available from published sources. For example, if the population is the set of residents of an SMSA, often Census tables will provide the mean of income, age, education and similar socio-economic and demographic variables. If the population is the set of automobile owners in a state, statewide registration figures may provide mean vehicle age, type, etc. Clearly, the key to determining F by solving equations of the form (9), is to search out functions $x(z)$ for which published population means are available. If one is imaginative, this search will often be successful.

As an example, return to the simple three mode case, and suppose one knew from Census data that the average income and auto ownership in the population were \$11,300 and .94 respectively. Suppose further that one had an on-board survey and roadside interview that provided the following estimated expected values:

Mode	Average Income	Average Auto Ownership
1. Carpool Users	\$10,000	.8
2. Drive Alone Users	17,000	1.4
3. Transit Users	6,000	.6

In this case, the three modal user groups would be the relevant strata (corresponding in this case to a choice-based stratification). The equations implied by (9) would be:

$$11{,}300 = 10{,}000\ F_1 + 17{,}000\ F_2 + 6{,}000\ F_3$$

$$.94 = .8\ F_1 + 1.4\ F_2 + .6\ F_3$$

$$1 = F_1 + F_2 + F_3$$

The resulting solution implied by this would be $F_1 = .50$, $F_2 = .30$, and $F_3 = .20$.

It should be noted that the above procedure for determining F can be implemented using population medians rather than means. We indicate here only the simplest case. Let z be a scalar, let $x(z) = z$, and let z_m be the population median of z, assumed known from published sources. Then it follows from (8) that

$$(10)\quad \frac{1}{2} = \Psi(z_m) = \sum_{b=1}^{B} \Psi(z_m|b)\cdot F_b$$

For each b=1,...B, the quantity $\Psi(z_m|b)$ is consistently estimated by the fraction of sampled decision makers belonging to T_b whose z value lies below z_m. Using this estimate in (10), a linear equation in F results.

We must also point out that when one attempts to solve for F using eq. (9), the replacement of $E(x|b)$ by a sample estimate implies that the equations to be solved are no longer exact. It is also likely that more than the minimal set of $|C|-1$ estimated expected values will be available, in which case the problem of solving (9) becomes one of finding some "best fit" values of F_b according to some criterion (e.g. least squares). For both of these reasons, the values of F that emerge as solutions will themselves only be estimates of the true F values. This comment of course continues to apply if eq. (10) is used rather than eq. (9).

Finally, we note that the procedure of solving for F described here, differs in an important way from the direct measurement and random sample estimation methods described earlier. That is, the present procedure determines F only after the stratified sample has been drawn, while the others do so prior to the drawing. This fact represents a drawback to the present procedure because given a stratification, prior knowledge of F can be useful in selecting the sample composition H_b, b=1, ...B. Why this is so will be made clear in Section 4.2.3.

d. <u>Estimation with the Choice Model</u>: Given a stratified sample and having specified a parametric form for the choice probabilities

$P(i|z,\theta^*)$, it is usually theoretically possible to jointly estimate the parameter vector θ^* and the population fractions F.[1] This approach is like (c) above in that it determines F only after a sample has been drawn. It differs from (a), (b), and (c), in that consistency of the F estimates obtained here depends on the correctness of the parametric model assumed for the choice probabilities. In contrast, the earlier approaches make no use of the choice probabilities whatsoever.

4.2.3. Selection Among Alternative Designs

Among the class of all possible stratifications of a population, some will not be implementable because relevant sub-populations can not be isolated or sampled from at random. Others will not be useable for estimation of the attribute distribution because the sub-population sizes F cannot be determined. Still, in most applied contexts there are likely to exist many feasible stratifications, and for each of these a set of alternative possible sample compositions. How then should the analyst select among these?

Unfortunately, relatively little guidance can presently be given. The choice among sample designs depend of course both on the relative costs and quality of the approximations to p associated with different designs. Sampling costs can only be determined on a case by case basis, so let us concentrate on the quality of approximation issue.

---

[1] See Manski and McFadden(1977) for details. Exceptions to the result are that F cannot be estimated in this way if the sampling is exogenous or if the choice model has the conditional logit form with alternative specific constants.

Consider again the identity

$$(8) \qquad \Psi(z) \equiv \sum_{b=1}^{B} \Psi(z|b) \cdot F_b$$

which forms the basis for stratified sampling estimation of p. On the basis of (8), some useful heuristic statements about the accuracy of alternative designs can be made.

First observe that given a sub-sample size N_b, the empirical distribution of z values among the N_b observations approximates $\Psi(z|b)$ best when this conditional density has all its mass concentrated on a single z point--that is when the sub-population T_b is homogenous in z. This suggests that a good stratification is one that separates the population into groups which are relatively homogenous in z. In general, internal homogeneity of the sub-populations can be enhanced by increasing their number, that is by more finely partitioning C x Z. Note, however, that given a fixed total sample size, the larger B is, the fewer observations that can be drawn from each sub-population; hence, the less accurate is each empirical z distribution as an estimate of the sub-population distribution. Thus, in selecting among stratifications, there is a tension between the desire for internal homogeneity of each sub-population and the need for an adequate number of observations per sub-population.

Assume now that a stratification has some how been selected. The analyst must then select a sample composition. Holding the total sample size fixed, two directives for choosing a composition can be given. First, observe that the influence of each conditional distribution $\Psi(z|b)$ on the population distribution $\Psi(z)$ increases directly with the value of F_b. This suggests that, all else equal, the larger F_b is, the larger the sample fraction H_b should be. Second, recall that the more homogeneous T_b is in z, the fewer observations are needed to achieve any accuracy in estimating $\Psi(z|b)$. This suggests that all else equal, the more homogeneous T_b is, the smaller H_b should be.

Finally, consider the choice of a sample size holding the sample composition fixed. It is easy to show that if N_b observations are drawn from sub-population T_b, the empirical distribution $\hat{\Psi}(z|b)$ is, when evaluated at any value z_o, a binomial random variable with mean $\Psi(z_o)$ and standard error $(\Psi(z_o)(1-\Psi(z_o))^{\frac{1}{2}}/\sqrt{N_b}$. Thus, increasing accuracy may always be obtained by taking larger samples but as measured by the standard error, accuracy increases only as $\frac{1}{\sqrt{N}}$. No general guidance can be offered as to how large a sample is "large enough." This question must be dealt with on a case by case basis.

The reader familiar with classical sampling theory will recognize that the above heuristic statements are extrapolations of well known results on optimal (i.e., minimum variance) sampling for estimation of a population mean. Here, interest is in estimating the entire density p(z), not simply the mean E(z). Hence, the classical results offer guidance but do not apply directly.

There is one further respect, ignored in the above discussion, in which the present sampling problem may differ from the classical one. That is, in practice we may only have estimates of the sub-population sizes F_b, not the true values. In contrast, the classical literature always assumes that the true values are available. The consequences of using F estimates in stratified sampling estimation of p(z) have not yet been explored.

5. SAMPLE DESIGN FOR CHOICE MODEL ESTIMATION

We now turn to the problem of learning the choice probabilities $P(i|z)$. As mentioned earlier, the literature on discrete choice analysis universally assumes that these probabilities are a priori specified up to the value of a finite parameter vector. As before, we will let θ^* designate the true value of this vector and let the choice probabilities be written as $P(i|z,\theta^*)$. It is furthermore generally assumed that the analyst has no prior knowledge of θ^*. Section 5.1 summarizes the theoretical literature on the estimation of θ^* given a sample of (i,z) observations and on the design of samples to be used in such estimation. Section 5.2 then assesses the practical implications of the existing theory for travel demand analysis.

5.1 THEORETICAL RESULTS

We first briefly review the quite comprehensive literature on choice model estimation given a sample design. We then present the contrastingly small set of available results relevant to the problem of selecting among alternative designs.

5.1.1. Estimation Given a Sampling Rule

From the perspective of sample design, the most important proven theoretical result on choice model estimation is certainly the following: <u>Subject to certain technical conditions, any stratified sampling rule provides a basis for consistent estimation of θ^*.</u>

The above finding emerges from the intensive investigation of maximum likelihood and related methods of estimation of θ^* made by Manski and McFadden (1977), and extended by Cosslett (1977). Maximum likelihood estimation rests, of course, on a re-interpretation of the sample likelihood as a function of those of its determinants which are unknown to the analyst. Consider then the stratified sampling likelihood originally given in equation (2), and rewritten here as

$$(2') \qquad L = \prod_{b=1}^{B} \prod_{n=1}^{N_b} \frac{P(i_n|z_n,\theta^*)\ p(z_n)}{\int_{(C \times Z)_b} P(j|y,\theta^*)p(y)d(j,y)} \cdot H_b.$$

If the attribute density p(z) is a priori known, a maximum likelihood estimate for θ^* is obtained by treating L as a function of θ^*, and finding the maximum of this function. Less transparently, if p(z) is unknown, the likelihood may be treated as a function of the pair of unknowns θ^* and p, and maximized jointly over all possible θ^* values and all possible attribute densities p. The above maximizations may be carried out in an unconstrained manner. If the sub-population sizes F_b, b = 1,...B are known, the constraints $F_b = \int_{(C \times Z)_b} P(j|y,\theta^*)p(y)d(j,y)$, b = 1,...B can be imposed.

It is shown in Manski and McFadden (1977), and Cosslett (1977), that subject to technical conditions, all of the above variants of the maximum likelihood method yield consistent estimates for θ^* whatever the stratified sampling rule used to generate the data. Among the required technical conditions, there are two of practical importance, one a condition on the probability model and the other a restriction on the sampling process.

The probability model condition is that θ^* be identified in the population. Roughly, this means that there must exist no vector of parameters other than θ^* which yield exactly the same choice probabilities as θ^* for all (i,z) pairs. Clearly, if there did exist some θ always yielding the same choice probabilities as θ^*, the sample likelihood would always be identical when evaluated at θ and θ^*. Hence, identification in the population is a necessary condition for consistent estimation of θ^* by _any_ estimation method, and _any_ sampling rule.[1]

Assume now that θ^* is identified in the population. Then the sampling process condition is that θ^* also be identified in the sample. That is, for a given stratification and sample composition, there must not exist a $\theta \neq \theta^*$ such that, for all possible samples, the likelihood (2') is identical when evaluated at θ^* and at θ.[2] In practice, if θ^* is identified in the population, it will generally also be identified in the sample as the latter condition can fail only in atypical sampling processes.

Maximum likelihood estimators are, of course, not the only methods available for estimation of θ^*. The reader interested in alternative approaches is referred to Amemiya (1975), Manski (1975), and Manski and Lerman (1977), for presentation of some alternative methods suitable in certain contexts. Because of their generality of application and their classical asymptotic efficiency properties, however, maximum likelihood estimators do occupy a special place in the literature. This special place will be evident when we next discuss selection among sample designs, since the available theoretical results relevant to this question all presume that maximum likelihood methods will be used in estimation.

5.1.2. Selection Among Sample Designs

We have earlier stated that there exists only a small set of theoretical results relevant to the problem of selecting a sample design for choice model estimation. Before describing these results, it will be useful to explain why significant findings in this area have been so difficult to achieve.

---

[1] Failure of this condition constitutes an important motivation for experimentation. See Section 6 for a discussion.

[2] Identification in the sample does not preclude the possibility that in some realizations of the sampling process, a unique maximum likelihood estimate for θ^* will not exist.

The traditional statistical measure of the precision of an (asymptotically) unbiased estimator is the (asymptotic) variance matrix of its parameter estimates. Maximum likelihood estimates of the choice model parameters θ^* are, in general, asymptotically unbiased. Moreover, holding the sample design and one's prior knowledge of p and F fixed, maximum likelihood estimates for θ^* are asymptotically efficient, in the sense that the asymptotic variance matrix of any other asymptotically unbiased estimator must exceed that of maximum likelihood by a positive semi-definite matrix. From the above facts, a natural strategy for statistically comparing alternative sample designs emerges. That is, for given sample size and informational conditions, examine how the maximum likelihood asymptotic variance matrix changes as a function of the sampling rule used.[1] A statistically "good" sampling rule can then be defined as one yielding a "small" variance matrix, where smallness of the matrix can be measured by its trace, largest eigenvalue or some other statistic.

The asymptotic variance matrices of maximum likelihood estimates for θ^* obtained under alternative exogenous and choice based sampling designs, and various informational conditions are given in Manski and McFadden (1977).[2] Inspection of these matrices reveals the following:

(i) For given prior information on p and F used in estimation, the relative precision of alternative sampling rules depends on the unknown value of θ^*. In particular, no design is uniformly best across the possible values of θ^*. This implies that the optimal sampling rule will depend on the true value of θ^*, which, if we knew in the first place, would obviate the need for sampling altogether.

---

[1] We note that under all stratified sampling rules, the asymptotic standard error of θ^* estimates decrease with sample size at the rate $1/\sqrt{N}$.

[2] While these authors do not present variance matrices for stratified designs other than choice based and exogenous ones, the findings reported below extend to such designs directly.

(ii) For a given value of θ^*, the relative precision associated with alternative sampling rules depends on what prior information on p(z) and F is used in estimation. In particular, knowledge of either the distribution of attributes, or the shares of the population in each strata will decrease the variance of the parameter estimates, except that knowledge of p alone is valueless if exogenous sampling is used.

(iii) For given θ^* and p-F information, the maximum likelihood asymptotic variance matrix is, except in some special cases, an analytically complicated function of the sampling rule. Hence, a ranking of rules by precision is usually difficult to achieve.

These three facts constitute the source of the problems researchers attempting to statistically compare alternative sample designs have faced. It should be noted that these problems are not peculiar to choice model estimation. In fact, they apparently arise in all non-linear modelling contexts. The reader familiar with the strong findings of classical sampling theory and expecting similar results here should recall that the classical theory presumes the relatively simple linear model $Y = x\beta + \varepsilon$, $E(\varepsilon|x) = 0$, $V(\varepsilon) = \sigma^2 G$ with G known and β and σ^2 to be estimated. The classical results do not apply outside this model, for example, even if the only change is to make G unknown.[1]

With the above as prelude, the theoretical results on sample design that have been achieved are now described. We first present the available analytical results, then the findings of some Monte Carlo experiments.

a. <u>Analytical Results</u>: The primary available analytical results concern the relative estimation precision obtained under alternative exogenous sampling designs when neither p nor F is known, and when the choice probabilities have the form

$$\text{(11)} \qquad P(i|z,\theta^*) = D(\ (z_i - z_j)\varpi^* + (\gamma_i^* - \gamma_j^*),\ j \in C)$$

[1]The classical results for the linear model are given in Rao (1973).

where $\theta^* = (\phi^*, \gamma_j^*, j\varepsilon C)$ and where the function D is strictly increasing in each of its arguments. In this case, each γ_j^*, $j\varepsilon C$ is an alternative specific constant, or dummy variable. All random utility models for which the utilities U_i are defined to be $U_i = z_i\phi^* + \gamma_i^* + \varepsilon_i$, where $(\varepsilon_i, i\varepsilon C)$ are independent and identically distributed disturbances, yield choice probabilities satisfying (11). In particular, the conditional logit model is a member of this class.

Given the above assumptions, Manski (1978), investigates the maximum likelihood asymptotic variance matrix for θ^* estimates under the hypothesis $\phi^* = 0$. It turns out that under this hypothesis, this usually complex matrix takes on a relatively simple form. In particular, if the choice set contains two alternatives or if $\gamma_j^* = 0$, all $j\varepsilon C$, then the asymptotic variance matrix becomes $\frac{\alpha}{N}\begin{pmatrix} V(z)^{-1} & 0 \\ 0 & I \end{pmatrix}$ where $\alpha > 0$, N is the sample size, I is the $|C| - 1$ dimensional identity matrix, and $V(z) = E(\sum_{i\varepsilon C} \sum_{j\varepsilon C} (z_i - z_j)'(z_i - z_j))$ is the expected sampling variance in attributes across alternatives in the choice set.

It follows from the above that, given the assumptions imposed, a good exogenous sampling design is one in which the decision-makers drawn face widely disparate alternatives within their choice sets, in the sense that the expected $\sum_{i\varepsilon C} \sum_{j\varepsilon C} (z_i - z_j)'(z_i - z_j)$ is "large". This is a rather intuitive result, particularly as it is analogous to the classical result for the standard linear model. Since the assumptions made in the present case are so stringent, however, the practical usefulness of the result should be questioned.

Actually, the work described here does have one important application. In practice, one sometimes is interested in determining whether a given choice model is informative, in the sense that it non-trivially explains some part of choice behavior. In the context of models of the form (11), a specification is informative if and only if $\phi^* \neq 0$. Given this, one may wish to design a sample suitable for the purpose of testing against the null hypothesis $\phi^* \neq 0$. For this purpose, the Manski (1978) results describe the characteristics of a good sample design.

Having offered a legitimate application, we now give a word of caution. It is tempting to extrapolate from the special case described here and conclude that in exogenous sampling a large attribute variance among the alternatives in a choice set is always desirable. This, however, is not the case.

In particular, consider the simple one parameter binary choice logit model

$$P(i|z,\theta^*) = e^{\theta^* z_{ij}} \Big/ 1 + e^{\theta^* z_{ij}} \quad \text{where} \quad z_{ij} = z_i - z_j.$$

The asymptotic variance of the exogenous sampling maximum likelihood estimate for θ^* can then be shown to be $\frac{1}{N}\left(E \dfrac{z_{ij}^2 \cdot e^{\theta^* z_{ij}}}{(1 + e^{\theta^* z_{ij}})^2} \right)^{-1}$. Inspection of the operand in the above expression reveals that if $\theta^* \neq 0$, then

$$\lim_{|z_{ij}| \to \infty} \frac{z_{ij}^2 \cdot e^{\theta^* z_{ij}}}{(1 + e^{\theta^* z_{ij}})^2} = 0$$

from which it follows that the value of $|z_{ij}|$ minimizing the variance of the θ^* estimate is finite. On the other hand, if $\theta^* = 0$, then in accord with this discussion of this section, we find that

$$\frac{z_{ij}^2 \cdot e^{\theta^* z_{ij}}}{(1 + e^{\theta^* z_{ij}})^2} = \frac{z_{ij}^2}{4}$$

implying that the variance of the θ^* estimate is an everywhere decreasing function of $|z_{ij}|$.

To close out this discussion, we call attention to a simple result on choice based sampling which highlights the role of prior information on $p(z)$ and F in determining the relative efficiency associated with alternative designs. It can easily be shown that if $p(z)$ is known, then the best choice based sampling design is one in which all observations are drawn from a single sub-population T_i. That is, the best design has all observations drawn from users of a single alternative. (Which alternative it is best to draw from depends on the value of θ^*, however.) On the other hand, if $p(z)$ is not known, then the best design must satisfy the condition $H_j > 0$, all $j \varepsilon C$. In fact, a design not meeting this requirement will often not even suffice to identify θ^* in the sample.[1]

b. <u>Monte Carlo Findings</u>: Cosslett (1978b) has been using Monte Carlo experiments to study the relative estimation precision associated with alternative choice based sampling designs in the context of single parameter binary choice models. If the choice set contains the two alternatives (i,j) and sampling is choice based, the analyst's control variable for sample design is the sample fraction H_i. Assuming that the choice probabilities have the logit, probit, or arctan form and that $p(z)$ is unknown, Cosslett examines how the asymptotic variance of θ^* changes as a function of H_i and of one's prior knowledge of F_i. While Cosslett's work is still ongoing and while Monte Carlo findings cannot be conclusive, two interesting tentative findings can be cited.

First, it appears that when F is not known, good designs are ones which place H_i close to $\frac{1}{2}$. This conclusion is quite strong in the logit and probit models, less so in the arctan one. On the other hand, when F is known, it appears optimal to oversample the rare alternative, that is to set $H_i > \frac{1}{2}$ if $F_i < \frac{1}{2}$.

---

[1]The first result occurs because in the case of p known, the information matrix (i.e., the inverse of the variance matrix of the estimates) is linear in the H_b values. On the other hand, when p is unknown, this matrix turns out to be linear in $\frac{1}{H_b}$.

Second, the usefulness of knowledge of F seems evident. Holding the sample design fixed, such knowledge reduces the variance of the θ^* estimates substantially in Cosslett's experiments.

5.2 PRACTICAL CONCERNS

We have seen that the existing theoretical literature on choice model estimation is very successful in offering the analyst methods for estimating θ^*.[1] The literature is, contrariwise, very weak in providing guidance on how one should select among alternative sample designs. Relatively few results are available and it appears that relatively little can be learned. To place the literature in appropriate applied perspective, the basic assumptions of existing theory must first be understood and interpreted.

Two assumptions characterize the existing estimation theory. First, the analyst is presumed able to a priori specify the functional form of the choice model; the only estimation problem is associated with the value of the parameter vector θ^*. This assumption is extremely useful because the data requirements for parametric analysis are considerably smaller than for non-parametric analysis. Second, one is presumed to have no prior knowledge of θ^* whatsoever. This assumption is standard in classical statistical analysis.

From the perspective of applications, the foregoing two idealized assumptions stand in interesting contrast. On the one hand, it is undoubtedly overly optimistic to suppose that in practice one can correctly place the choice probabilities in a known parametric family. Behavioral theory and empirical observation will usually let one put some structure on the choice probabilities,

---

[1] Although it should be noted that not all the estimators discussed above have been programmed in available econometric software.

but not the very exacting structure implied by a parameterization. On the other hand, having specified a parametric family, it will often be too pessimistic to assert that the analyst is totally ignorant about θ^*. Loosely, θ^* characterizes tastes and we usually do have some prior knowledge of what people's tastes are.

With the above as background, three general comments relevant to sampling practice can be offered.

First, in designing a sample for choice model estimation, concern with estimability, that is the ability of the design to support estimation of θ^* at all, should dominate worry about estimation precision. This advice is given for a very simple reason. That is, estimability is a necessary requirement before precision can even become an issue.

Second, to the extent that one does become concerned with the relative precision of alternative designs, the classical statistical framework assumed in the existing theoretical literature should be applied sensibly rather than dogmatically. If the analyst has prior information about the value of θ^*, he should use it in selecting among designs.[1] If he views his choice model as only an imperfect approximation of reality, he should recognize that theoretical results comparing designs can themselves hold at most approximately. If the classical measure of estimation precision, that is the asymptotic variance matrix, differs from the measure he feels most desirable, he should understand that a classical ranking of designs may not be most appropriate.

Third, it should be understood that the problems of describing p and estimating θ^*, while formally distinct, are nevertheless related in various ways. For one thing, prior knowledge of F is both necessary to describe p and useful in estimating θ^*. For another, the F values can in theory

[1] A formal framework for incorporating such prior information is provided by Bayesian analysis. See Section 7 for further discussion.

be estimated along with θ^* and then used in describing p. (Such joint estimation of F and θ^* is possible except when the sampling is exogenous or the choice model has the conditional logit form. See Manski and McFadden, 1977.) Perhaps most important, the same data sample is often used both to describe p and estimate θ^*. When such dual use is intended, the sample design selected must be suitable for both objectives.

6. DESIGN OF EXPERIMENTS

Consider a proposed policy whose effect, if implemented, would be to create a travel environment which differs in some way from that currently faced by decision makers. To forecast the impact of such a policy on travel behavior, the choice probabilities which would prevail under the new travel environment must of course be known. However, in the absence of historical experience, these probabilities may not be a priori known, nor estimable from data on current travel choices.

For example, suppose one were interested in forecasting how the introduction of buses with wheelchair lifts into an area without such buses would influence transit usage. In a mode choice model, any parameters relating to the desirability of buses with lifts would not be estimable from pre-introduction data. In such contexts, where the choice probabilities cannot be inferred from existing data, it may be useful to subject a subset of the decision making population to the policy of interest, observe their subsequent behavior and then infer the required choice probabilities from these observations. That is, it may be useful to perform an experiment.

Within the framework of discrete choice analysis, an experiment, like a permanent policy change, can be viewed formally as a function $y(z)$ changing each decision maker's present attribute value into some new value.

By combining information about the population's behavior under the pre-experiment attributes z and their behavior under the post-experiment attributes $y(z)$, the set of choice probabilities for any attribute vector in the original range of attributes Z_o or in the post-experiment range $Z_1 = (y(z), z \varepsilon Z_o)$ can, in theory, be inferred.

It is natural to ask how one should design an experiment so that it will be informative regarding the consequences of policies of interest. This important question has not previously been addressed in the discrete choice literature. To begin what should eventually be an extensive investigation, Section 6.1 offers a few simple theoretical results relevant to the design of experiments. Section 6.2 then discusses some practical concerns that arise in experimentation.

Before proceeding, we should point out that experimentation has uses beyond the discrete choice applications discussed here. In particular, experiments may be used to determine what consequences a policy change would actually have for the attribute distribution. That is, when the function $y(z)$ associated with a policy change is unknown, experimentation may enable one to learn this function. Many of the experiments carried out in UMTA's Service and Methods Demonstration Program have this objective. Experiments performed for such purposes will not be discussed further in this paper.

6.1 THEORETICAL RESULTS

The role of experiments within the existing theory of choice model estimation can be easily described. Assume as usual that the choice probabilities $(P(i|z), z\varepsilon Z)$ have been placed in a parametric family indexed by θ^*. Let there exist a proposed policy of interest which would map Z_o onto an attribute set Z_1 and assume that the choice probabilities $(P(i|z), z\varepsilon Z_1)$ can themselves be parametrized by θ^*. Consider now a situation in which some subset θ_1^* of θ^* is not identified in the present population described by $f(i,z) = P(i|z,\theta^*)p(z)$ but is identified in the post policy population described by $\tilde{f}(i,z) = P(i|z, \theta^*)\tilde{p}(z)$. Hence, forecasting post policy behavior requires that θ_1^* be known, but θ_1^* cannot be inferred from observations of present choices. From this, the objective of experimentation emerges, that is to modify the present attribute distribution in such a way that θ_1^* is identified in the post-experiment population.

The above discussion is quite abstract. Let us therefore describe some simple examples of applied importance. Let $z_{ti} = (z_{oti}, z_{1ti})$, $\theta^* = (\phi_o^*, \phi_1^*, \gamma_i^*, i\varepsilon C)$ and assume that the choice probabilities $P(i|z)$ are derived from the random utility model $U_{ti} = \gamma_i^* + \phi_o^* \cdot z_{oti} + \phi_1^* \cdot z_{1ti} + \varepsilon_{ti}$ where $(\varepsilon_{ti}, i\varepsilon C)$ has some given probability distribution. Consideration of four problems within this context will serve to indicate some of the uses of experiments.

(i) Assume that for each $t\varepsilon T$, $z_{1ti} = z_{1tj}$ for all $i, j\varepsilon C$. That is, alternatives are, for each decision-maker, completely homogenous along the z_1 attribute. Clearly, ϕ_1^* is not then identified. If we wish to forecast behavior under a policy which makes alternatives heterogenous along the z_1 attribute, knowledge of ϕ_1^* is necessary. For example, local regulatory policy tends to create taxi fares which are uniform across operators in a given area. In a model of choice of which taxi company people call for service, a coefficient (corresponding to ϕ_1^*) for the effect of taxi fare (corresponding to z_{1t}^*) could not be estimated.

(ii) Assume that for all $s,t\varepsilon T$ and all $i\varepsilon C, z_{1ti} = z_{1si}$. That is, the z_1 attributes are constant across decision-makers. In this case the composite parameters $(\gamma_i^* + \phi_1^* z_{1ti})$, $i\varepsilon C$ may be identified, but not the γ_i^*'s and ϕ_1^* separately. If we wish to forecast behavior under a policy which makes the z_1 attributes heterogenous across decision-makers or one which simply changes the z_1 values uniformly, the γ_i^* and ϕ_1^* parameters must be known. An example of this is if z_{1t} were a dummy variable in a mode choice model which indicated whether or not a mode was characterized as demand responsive. Since most urban areas do not offer demand responsive service, and the auto mode by its very nature is demand responsive, it would be impossible to distinguish the effect on utility of an automobile constant from that of a dummy variable describing whether or not a mode was demand responsive. Only a composite automobile/demand responsive constant could be estimated.

iii) Assume that for each $t\varepsilon T$ and $i\varepsilon C$, $z_{oti} = \alpha z_{1ti} + \beta$ for some α and β. That is, the z_o and z_1 attributes are perfectly correlated. Now $\phi_o^*\alpha + \phi_1^*$ may be identified but not ϕ_o^* and ϕ_1^* separately. If we wish to forecast behavior under a policy which makes z_o and z_1 less than perfectly correlated or one which simply changes the α constant, the values ϕ_o^* and ϕ_1^* must somehow be learned. One example of this might occur in small cities with metered taxis and zonal bus fares. In such cases, it is conceivable that fares by taxi and bus would be perfectly correlated with in-vehicle travel times; car operating costs and times would be similarly correlated. In this case, it would be impossible to estimate separate time and cost effects; only a composite coefficient for the sum of cost and in-vehicle time could be estimated.

iv) Assume a choice $k\varepsilon C$ is entirely unavailable to a population, and $P(k|z)$ would always therefore be zero. In this case, if $z_{1ti} = z_{1tj}$ for all $i,j\varepsilon C$, $i, j\neq k$, then ϕ_1^* and γ_k^* would not be estimable. For example, some small cities may not have any transit service. In this case, the parameters of all the transit specific variables would not be estimable, including the transit constant. Some parameters, such as that for generic travel time, however, could be estimated from existing choices among carpooling, driving alone and taxi.

Appropriately designed experiments can solve each of the above problems. For example, one might in each case subject a subset of the decision making population to a scaled down version of the intended policy. However, such a strong correspondence between experiment and policy is not necessary. Any modification of the present attribute distribution which renders the relevant parameters identified will suffice.

Return now to our abstract discussion of experimental design. Given a set of experiments each of which can identify θ_1^*, it remains to ask how one should select among this set. This question has a very simple formal interpretation. Each experiment one can conduct produces some new attribute distribution. Given this, the problem of selecting a good experiment becomes one of selecting a good attribute distribution. The latter problem is obviously closely related to the sample design problem treated in Section V but has not itself been investigated thus far.

6.2 PRACTICAL CONCERNS

In most respects, the practical concerns that arise in designing an experiment are analogous to those that arise in designing a sample for choice model estimation. The experimental design problem does, however, have one aspect that does not appear when sampling. That is, a duration for the experiment must be selected.

If it could be shown that people react quickly to changes in their travel environments, the duration question would be of little consequence. However, there exists ample anecdotal evidence that adjustments to new conditions occur only slowly. Moreover, the adjustments people make when they believe a change is temporary may differ from those they make when they think one is permanent.

Existing discrete choice theory, being static, obviously can offer no guidance on the selection of an experiment's duration. The nascent dynamic choice theory cited in Section 1. could potentially provide such guidance but will require considerable development before it becomes useful in practice.

A second practical issue arises when none of the four theoretical conditions for non-identification discussed in Section 6.1 holds entirely, but one or more is nearly true. That is, while it may be theoretically possible to estimate all the parameters of interest, some estimates may be extremely unreliable. In such situations, it might prove useful to conduct an experiment which increases the range of an attribute or makes a particular alternative available to more decision makers.

For example, even in a fixed fare system, there may be some variation in transit fares due to people having to transfer to make certain trips (assuming, of course, that transfers are not free). However, while this variation in transit fare may theoretically identify a parameter for transit fare in a mode choice model, the reliability of the parameter estimate may be extremely low. In this case, a zone fare experiment on selected routes would be a useful way to improve predictions of more extensive fare policy changes.

To conclude this section, we should comment on the substantial difference between the use we have advocated for experiments and a more traditional view of experimentation. Traditionally an experiment is performed to test for the existence of an "effect" when a single factor in the environment is changed. No explicit model is assumed and proper inference requires that the factor of interest be the only one that changes over the duration of the experiment. Otherwise, the effect may be "confounded".[1] Here, in contrast, an

[1] In practice, confounding is very difficult to avoid in transportation experiments as the real world environment in which such experiments are conducted does not permit ceteris paribus conditions.

experiment is a mechanism for learning the values of parameters in a formal model. There is no restriction on the number of factors changing during the experiment, whether by design or otherwise. Proper inference from the experiment requires only that the changes that do occur either be observed or, if unobserved, satisfy appropriate statistical conditions so that the assumed probabilistic choice model continues to describe behavior.

7. DIRECTIONS FOR FUTURE RESEARCH

A reading of this paper indicates that the state of the art of sample design for discrete choice analysis is advanced in some respects and primitive in others. It is important to note that many of the results reported here are quite recent and that further work will undoubtedly resolve some of the questions raised in the report. Discrete choice analysis is still a quickly growing area of knowledge and future work on sample design problems will hopefully make more precise statements about alternative sampling strategies possible. A large set of useful directions for future research may be enumerated.

With regard to estimation of the attribute distribution, research in three areas would seem particularly productive. First, there is a need for a better understanding of the relative merits of the various methods for determining the population shares F and of the implications of using estimated F values in estimating the attribute distribution. Second, more formal statistical criteria for comparing alternative stratified sample designs should be developed. Third, ways to use various forms of prior knowledge of the attribute distribution in the estimation process should be researched. In particular, while interested in the current attribute distribution, one often has available knowledge of this distribution at some past time. Duguay, Jung, and McFadden (1976), have developed an interesting but ad hoc approach to updating such past attribute distributions using available aggregate data on current conditions. Work aimed at assessing the properties of their method would be useful.

In the area of choice model estimation, extensions of the classical type analytical work of Manski (1978) and Monte Carlo tests of Cosslett (1978b) would be of some use. Of potentially greater value, however, would be work aimed at replacing the classical sample design framework assumed in the existing discrete choice literature with a more powerful one. In particular, the Bayesian approach offers such a framework. Adoption of this approach would provide a means of incorporating prior information on θ^* into both the sample design and choice model estimation process. Additionally, the statistical decision theory aspect of Bayesian analysis offers a wide variety of sampling strategies outside of the stratified sampling rules applied to date.[1]

The field of experimentation offers some of the most interesting challenges for future research. In particular, there presently exists no theoretical basis for selection among alternative experimental designs. While we have previously stated that the experimental design problem seems similar to that of sample design, the exact relation between the two problems is not clear. A second important issue in experimentation regards the selection of a duration for the experiment. Consideration of this question ultimately leads one to be concerned with the dynamics of choice behavior and thus beyond the static discrete choice framework assumed in this paper.

As a final direction for future work, recall that the attribute space Z is assumed a priori defined in this paper but that its structure is actually under the control of the analyst through his decisions to collect data on some attributes but not others. Research aimed at providing guidance to aid the analyst in these decisions could prove quite useful.

[1]For an extensive treatment of the Bayesian approach, see DeGroot (1970).

References

1. T. Amemiya, "The Maximum Likelihood, the Minimum Chi-Square, and the Non-Linear Weighted Least Squares Estimators in the General Qualitative Response Model", Journal of the American Statistical Association, Vol. 71, 1976, pp. 347-351.

2. S. Cosslett, "Choice Based Sampling and Disaggregate Demand Forecasting", December, 1977, unpublished.

3. S. Cosslett, "Efficient Choice Model Estimation Using Enriched Samples", 1978a, unpublished.

4. S. Cosslett, unpublished notes, 1978b.

5. M. DeGroot, Optimal Statistical Decisions, McGraw-Hill, 1970.

6. S. Duguay, W. Jung and D. McFadden, "SYNSAM: A Methodology for Synthesizing Household Transportation Survey Data," 1976, unpublished.

7. J. Heckman, "Models for Testing the Hypothesis of True State Dependence Against that of Spurious State Dependence", July, 1977, unpublished.

8. S. Lerman, C. Manski and T. Atherton, "Alternative Sampling Procedures for Disaggregate Choice Model Estimation", Transportation Research Record No. 592, 1975, pp 24-28.

9. C. Manski, "Maximum Score Estimation of the Stochastic Utility Model of Choice", Journal of Econometrics, Vol. 3, No. 3, 1975, pp 205-228.

10. C. Manski and S. Lerman, "The Estimation of Choice Probabilities from Choice Based Samples", Econometrica, Vol. 45, No. 8, 1977, pp 1977-1988.

11. C. Manski and D. McFadden, "Alternative Estimators and Sample Designs for Discrete Choice Analysis", February 1977, unpublished.

12. C. Manski, "Optimal Sample Design Under the Hypothesis $\theta^* = 0$", (technical note), January 1978.

13. D. McFadden, "Recent Advances in Travel Demand Forecasting", 1977, unpublished.

14. C.R. Rao, Linear Statistical Inference and its Applications, 2nd Edition, Wiley, 1973.

Report of Inventions Appendix

Although a diligent review of the work performed under this contract has revealed that no new innovation, discovery, or invention of a patentable nature was made, this report summarizes recent advances in the theory of sample design for discrete choice analysis and presents some theoretical results and practical guidelines which are new and have not been previously reported. For example, Section 4 on sample design for description of the attribute distribution concludes with some heuristic guidance for selecting among alternative sampling strategies. Section 5 presents guidance on sample design for choice model estimation based on new analytical results and the findings of recent Monte Carlo experiments. Section 6 presents some novel ideas on the role and design of experiments to learn about the values of probabilistic choice model estimation.

☆U.S. GOVERNMENT PRINTING OFFICE: 1978—700-446/189

170 copies

U.S. DEPARTMENT OF TRANSPORTATION
Research and Special Programs Administration
Transportation Systems Center
KENDALL SQUARE, CAMBRIDGE, MA. 02142