Calcutta Statistical Association Bulletin
Vol. 60, March & June 2008, Nos. 237-238

STOCHASTIC MODELING OF BUYING BEHAVIOUR OF INDIAN CUSTOMERS

PULAKESH MAITI
Economics Research Unit
Indian Statistical Institute, Kolkata, India

ABSTRACT : A collaborative research on a problem of common interest and of immediate concern, which Hindustan Lever Limited (HLL) - a multinational company was facing or likely to face in the near future, was undertaken at the Indian Statistical Institute (ISI), Kolkata. The problem was to explain the purchase behaviour of frequently bought branded consumer products using stochastic models. For this, the panel data, after being coded to ensure anonymities were supplied to the ISI and on the basis of the available data, modeling of the buying behaviour was made. To begin with, some descriptive measures were calculated to understand the data and, finally, *'Dirichlet' multinomial model* was used for explaining the buying behaviour of the customers in the specific segment with respect to the specific group of commodities. Because of not-so-wide coverage of the data and not-well-validated assumptions on the underlying distributions, the results failed to reveal much of the consumer behaviour pattern.

Keywords and phrases : Stochastic modeling, Descriptive measures, Buying behaviour.

AMS (2000) Subject Classification : 62P20.

1. Introduction and Preliminaries

A consumer's buying behaviour is characterised by its distinctiveness with regard to (i) the choice(s) of a brand or a group of brands, (ii) adherence or non-adherence to a particular brand or "group of

brands", (iii) frequencies of purchasing occasions within a given period (iv) quantities and/or units purchased on each occasion, and (v) total purchases of a brand or "specified group of brands," during a given period. The buying behaviour may be influenced by too many factors. To mention a few, even under similar conditions with respect to underlying tastes and preferences, income of the buyer, prices of the products and competing goods (or relative prices), experience of previous usage, effects of promotional activities like advertising and offering discounts among others, words of mouth effects are some such factors. Whether or not and to what extent the consumers react to the changes in these factors also depend on the nature of the commodity being considered; for example whether it is a "necessary" or a "luxury" item i.e., on the elasticity of demand. No doubt there will be variation in total and/or average purchases of a brand or group of specified brands purchased across the households/consumers, but even for the same household, there would be variations in the frequencies of purchasing occasions in different time periods.

In this paper an attempt is made to describe the behaviour of the buyers of some fast moving consumer goods (FMCG) with regard to the above characteristics. The data were provided by Hindustan Lever Limited (HLL) and a project was undertaken at the Indian Statistical Institute, Kolkata. First, we summarized the data using a few descriptive measures. Then a model of purchasing frequencies, popularly known as "Dirichlet" multinomial model was tried to the data. Before concluding this section, we describe the nature and coverage of available data.

The data on brand-wise amount of purchase (in grammes) were available on three products viz., detergent bar, detergent powder, and toilet soap for each of 1699 households but for some households data were missing for some months. The complete month-wise data for one full calendar year from January to December were available for 1153 households. However, there was no information on the purchase frequency of the households month-wise. This panel data set also provided information on (a) household identification number, (b) brand code, (c) month number and (d) quantity of product purchased. For the sake of completeness, it recorded *no purchase* of a particular brand as 0 grammes. The data also provided demographic particulars viz., family size, income class etc.

We refer to the Project Report which is available with the author for most of the technical derivations and detailed tables and computations.

We only present two tables in the Appendix.

2. Some Relevant Concepts

For the sake of completeness, we introduce and explain some concepts specific to the understanding of the consumer behaviour.

2.1 Penetration

Market penetration or simply penetration of a brand is defined as the proportion of buyers buying a particular brand atleast once during a given period of time. It normally increases with the length of time, though less than proportionally.

2.2 Repeat buying

It represents the proportion of buyers purchasing a given brand at least once during the given period out of all those who have purchased the same brand at least once previous to the period under consideration. Actually some of the previous buyers may make zero purchases of the brand during the given period, some may buy it once, twice and so on. All those who buy once or more will indicate, as a proportion of the previous buyers, repeat buying.

2.3 Sole buyers

The buyers who buy only one particular brand of a product, and no other brand during the given period are called 'sole' buyers. These buyers may in this sense, be regarded as 'loyal' to the brand under consideration. The proportion of sole buyers is the ratio of the buyers who buy the particular brand under consideration to the total number of buyers and is normally higher, the shorter is the period of analysis.

2.4 Duplicate buyers

The proportion of buyers of the brand buying another brand of the same product, at least once, during the given period at time.

2.5 Purchase frequency per buyer

The (average) purchase frequency of a brand per buyer is defined as the ratio of total number of purchases of the brand to the total number of buyers during a given period of time.

Thus

$$\text{Average purchase frequency} = \sum i \cdot f_i / \sum f_i$$

where i = number of purchases and f_i = number of buyers making i number of purchases.

3. The Dirichlet Model : A Model of Buying Behaviour

One of the simplest statistical models which has become quite popular in the literature in describing purchase behaviour of the consumers is the Ehrenberg model, popularly known as Dirichlet model. It tries to describe, for a stationary and unsegmented market, the manner in which consumers, buying some consumer product, behave in their purchasing activity.

This model describes the number of purchases of different brands out of a given number of brands in a product class. In other words, it tries to combine both purchase incidence and brand choice behaviour of the buyer in the same model.

There are four basic components of the Dirichlet model which capture the buying situation. The model description is as follows.

(a) The purchase incidence distribution

For a given buyer say i, the distribution of the number of purchases n_i made by the i^{th} buyer in the product group during a given time

period T follows a Poisson distribution with μT, the average number of purchases *i.e.*,

$$P(n_i \mid \mu_i, T) = \frac{e^{-\mu_i T}(\mu_i T)^{n_i}}{n_i!}; \; n_i = 0, 1, 2, \ldots, \infty \qquad (3.1)$$

(b) **Mean-purchasing rate distribution**

Each buyer makes an average number of purchases during a given period which, varies across the buyers. This variation in mean purchasing rate has been described by assuming μ_i to follow a Gamma distribution as

$$G(\mu_i \mid M, K) = e^{-\mu_i K/M} \mu_i^{K-1} (K/M)^K / \Gamma(K); \; \mu_i > 0. \qquad (3.2)$$

where M and K are the parameters of the distribution, with M representing the rate of purchase of the product per unit of time.

From the above two assumptions on n_i and μ_i, one gets the compound distribution of the number of purchases from the product class by the i^{th} buyer as the Negative Binomial Distribution (NBD), written as

$$P(n_i/\mu_i) = \binom{n_i + K - 1}{n_i} p^{n_i} (1-p)^K, \qquad (3.3)$$

$$\text{where } p = (MT)/(MT + K).$$

(c) **Distribution of number of purchases of a particular brand**

Given n_i, the number r_{ig} of purchases of the g^{th} brand $(g = 1, 2, \ldots, G)$ made by the i^{th} buyer in the product group has a multinomial distribution with different brand choice probabilities $p_i = (p_{i1}, p_{i2}, \ldots, p_{iG})$ *i.e*,

$$M(r_i = (r_{i1}, r_{i2}, \ldots, r_{iG})/n_i, p_i) = \frac{n_i!}{\prod_g^G r_{ig}!} \prod_{j=1}^{G} p_{ig}^{r_{ig}}$$

The assumption of the multinomial distribution follows from the fact that the stochastic buying behaviour at the individual level tends to be stationary, though irregular.

(d) **Distribution of brand choice probabilities**

Finally, it is assumed that brand choice probabilities $p_i = (p_{i1}, p_{i2}, \ldots, p_{iG})$ are fixed over time but vary over the buyers according to the Dirichlet distribution $D(p/\alpha)$ with parameter $\alpha = (\alpha_1, \ldots, \alpha_G)$.

Hence, given the assumptions from (a) to (d), the unconditional distribution of $r = (r_{i1}, r_{i2}, \ldots, r_{iG})$ during a given period of length T is given by the compounding of the four distributions stated above and is given by

$$M(r \mid p, n) \underset{p}{\wedge} D(p \mid \alpha) \underset{n}{\wedge} p(n \mid \mu) \underset{\mu}{\uparrow} G(\mu \mid M, T, K).$$

Remark 3.1 It may be noted that the summary measures defined in Section 2 above can be expressed as different functions of model parameters; for example, penetration of the market share of the g^{th} brand can be expressed as $\alpha_g / \sum_{i=1}^{G} \alpha_i$.

Remark 3.2 It may be noted that the basic input of the model is the **number of purchases** made during a given period along with information on number of distinct brands procured on each of occasion.

Remark 3.3 Another point of importance is to ascertain if the households under consideration can be taken to be single decision making units.

4. An empirical Regularity - A Prelude to the Model Verification

Let us mention at the very outset that the data supplied were not suitable for checking the appropriateness of the Dirichlet model.

The available data on the quantity (in grammes) purchased of each brand by the households had to be converted into number of units, as the basic unit of analysis in the Dirichlet model is the number of units purchased. This required knowledge about the standard size of each brand. In the absence of such information and also for the sake of

rendering flexibility to our analysis, we used alternative sizes viz., 125 grammes, 250 grammes, 500 grammes and 1000 grammes with proper rounding off to obtain the number of purchasing units. In fact, we used the figures for the total quantity purchased divided by (alternative) standard sizes as the number of purchases made by the households. By this procedure, we generated four sets of data on the number of purchases of each brand made by each household in each of the 12 months. Also combining all brands together, similar figures for the product as a whole were also obtained.

4.1 Unit of Time for Statistical Analysis

When one talks of purchase behaviour, it is necessary to define the length of the period (denoted by T in our model description), because the observed behaviour may very well depend on the duration of time considered. For the purpose of our analysis, we have taken one month, three months and one year as alternatives. Thus 4 different sizes and 3 different lengths of reference duration resulted in 12 combinations of data sets for statistical analysis.

To begin with, the product detergent bar was taken up in search of empirical regularity, if any, in the given data. However, while supplying the data, the fourteen brands were coded as

100, 300, 400, 800, 900, 1100, 1400, 1500, 1700, 1800, 1900, 5300, 5400, 5500

and then the data were given to the ISI.

4.2 On the choice of the preferred data set

Though the exercises of model verification were carried out using all the twelve different data sets [arising out of 4 different sizes and 3 different durations] and similar results were obtained in all the cases, the data set with size 1000 grammes was preferred because of the following reason:

It may be observed that the theoretical model on the number of purchases is unimodal, but the empirical distributions for the above mentioned alternatives, except for the one with size 1000 grammes, appeared to be bimodal. This justified the use of the data set with size

1000 grammes (Ehrenberg, 1959).

4.3 Descriptive Statistics

In our search for any empirical regularity in the data which can be captured in terms of a model, certain descriptive measures were computed to obtain more insight into the data. To begin with we tried to identify the "**most preferred brand**" and "**preferred combination of brands**".

The Criteria used for identification were market share and brand loyalty.

(a) Considering the above criteria and through the analysis made, the '**most preferred brand**' appeared to be the '**brand 800**' with highest market share. According to the market shares we then grouped the brands as follows:

I : (800, 1100, 1400, 1900);

II : (300, 5300, 5500);

III : (1800, 100, 900).

(b) The preliminary analysis revealed the existence of sole as well as duplicate buyers and also the phenomenon of repeat buying.

(c) It may be observed that the estimates of the probabilities are rather high.

Remark 4.1

All the patterns similar to those in (a) through (c) were found even when the analysis was made using the data on number of units of purchases for all the twelve data sets.

5. Model verification, Conclusions and Recommendations

Frequency chi-square tests for goodness of fit were used for verifying the models. In calculating the expected frequencies, different estimation methods were adopted for different parameters involved in the different components of the distributions. The details are available

in the report. It turned out that the distributional assumptions were not validated by the available data.

It may be mentioned, however, that the data collected and supplied to us by the company are, strictly speaking, not appropriate for the verification exercises. The Dirichlet model requires "**the number of purchases**" made by the households every month. What was available was the "**quantity of purchases**" and what was done therefore, was to convert, somewhat artificially, the quantity of purchase data into number of purchases by assuming alternative sizes.

Based on these data the verification exercises were carried out, but the results were negative. The exercise of validity of the assumptions of Negative Binomial Distribution, the multinomial distribution and the multivariate - β for the brand choice probability were done separately. In all these cases, the appropriateness of the distributional assumptions for the Indian data was not beyond questions. We must hasten to add, however, that the results could have been different if the proper data were available.

A second point of importance is to ascertain if the households under consideration can be taken to be single decision making units. There may be households within households and the existence of such households may make a difference. Thus, there may be 'nuclear' households where the decision on to what to buy, when to buy and what quantities to buy are taken by only one person, presumably the head of the household. As contrasted to this, there are households consisting of joint families or households where the parents live with married son(s). In such cases, there may be divergences in the choice of the brand(s) among the members of the households. From the point of view of statistical modeling, such multiple decision making units perhaps need to be treated differently.

Acknowledgement : The author acknowledges Professors J. K. Ghosh, Late Robin Mukherjee, M. Pal, Dr. D. Roy, all from the Indian Statistical Institute, Kolkata and Dr. S. Roy, former general manager, Business Research and Corporate planning, HLL and presently Economic Adviser Tata Group, Tata Services Limited, Mumbai, for their participation in the project. The author is also thankful to the referee for his valuable and constructive suggestions.

REFERENCES

Goodhart, G. J., Ehrenberg A. S. C. and Chatfield C. (1984). The Dirichlet : A Comprehensive model of buying behaviour, *JRSSA* **147** 624 - 655.

Grahn, G. L. (1969). The negative binomial distribution model of repeat purchase loyalty : An Empirical investigation", *J. Marketing Res.* **6** 72 - 78.

Ehrenberg, A. S. C. (1959). The Pattern of Consumer Purchases, *Appl. Statist.* **8** 26 - 41.

Engel, J. E., Kollat, D. T. and Blackwell, R. D. (1968). Consumer behaviour. New York, Holt, Rienhart and Winston.

Johnson, N. L. and Kotz, S. (1969). Discrete Distributions. Boston, Houghton Miffin.

Patil, G. P. and Joshi, S. W. (1968). A Dictionary and Bibliography of Discrete Distributions. Edinburgh.

Consumer Behaviour and Business Research (2000). Indian Statistical Institute, Kolkata, India.

APPENDICES

A 1 : Market share of different brands according to quantities purchased and corresponding standard errors

Brand code	Market share	Standard error	Brand code	Market share	Standard error	Brand code	Market share	Standard error
(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
100	0.0089	0.0008	300	0.0277	0.0010	400	0.0027	0.0004
800	0.3400	0.0052	900	0.0050	0.0006	1100	0.2729	0.0045
1400	0.0950	0.0025	1500	0.0007	0.0001	1700	0.0022	0.0003
1800	0.0219	0.0010	1900	0.1560	0.0036	5300	0.0283	0.0012
5400	0.0078	0.0022	5500	0.0308	0.0017	—	—	—

A 2 : Probability that a household purchasing a brand would purchase it in the next month

Brand code	Prob. of Purchase	Standard error	Brand code	Prob. of Purchase	Standard error	Brand code	Prob. of Purchase	Standard error
(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
100	0.5022	0.0531	300	0.5668	0.0193	400	0.5227	0.1146
800	0.7929	0.0067	900	0.2893	0.0957	1100	0.7694	0.0148
1400	0.6975	0.0127	1500	0.8030	0.0835	1700	0.4790	0.0564
1800	0.5414	0.0255	1900	0.6659	0.0095	5300	0.5538	0.0021
5400	***	***	5500	0.6290	0.0329	—	—	—