Working Paper Series, N.12, 2010

Latent Class Analysis for Marketing Scales Development

Francesca Bassi
Department of Statistical Sciences
University of Padua
Italy

Department of Statistical Sciences
University of Padova
Italy

Abstract: This paper examines the Dirichelet model describing consumer behaviour. The model estimates brand performance measures in the case of repeat purchases over a set of brands. The Dirichelet model relies on some assumptions such as stationarity and the fact that the market is unsegmented. Its formulation derives from a combination of the Negative Binomial and the Dirichelet distributions. Various estimation methods have been proposed. The original one is an iterative procedure based on the method of moments and requires as inputs only aggregated quantities, such as brand penetrations and average purchase rates. There is also an estimation method based on likelihood maximization which requires raw individual or household panel data. The method of moments deserves attention, since raw panel data are frequently not available to researchers and/or enterprises. In this paper, the Dirichelet model is used to analyze the Italian beer market as a by-product of the main objective, which is to compare two estimation procedures available on-line for the method of moments: one based on an Excel Workbook and the other written in R.

Keywords: Dirichelet model, consumer behaviour, estimation, market segmentation

UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI SCIENZE STATISTICHE

Final Version (2010-10-15)

Contents

The Dirichelet model: analysis of a market and comparison of estimation procedures.

Francesca Bassi
Department of Statistical Sciences
University of Padua
Italy

Abstract: This paper examines the Dirichelet model describing consumer behaviour. The model estimates brand performance measures in the case of repeat purchases over a set of brands. The Dirichelet model relies on some assumptions such as stationarity and the fact that the market is unsegmented. Its formulation derives from a combination of the Negative Binomial and the Dirichelet distributions. Various estimation methods have been proposed. The original one is an iterative procedure based on the method of moments and requires as inputs only aggregated quantities, such as brand penetrations and average purchase rates. There is also an estimation method based on likelihood maximization which requires raw individual or household panel data. The method of moments deserves attention, since raw panel data are frequently not available to researchers and/or enterprises. In this paper, the Dirichelet model is used to analyze the Italian beer market as a by-product of the main objective, which is to compare two estimation procedures available on-line for the method of moments: one based on an Excel Workbook and the other written in R.

Keywords: Dirichelet model, consumer behaviour, estimation, market segmentation

1. Introduction

The Dirichelet model describes how frequently-bought branded consumer products are purchased when the market is stationary and unsegmented. It was developed by Goodhardt, Ehrenberg and Chatfield (1984) and in the following years was shown to be applicable to many product categories and to have substantial uses, particularly with regard to analysis of brand performance measures.

In this paper, the Dirichelet model is applied not only to describe the Italian beer market but, more importantly, to compare results obtained estimating model parameters with two software packages available on-line: an Excel-based one, written by Kearns (2002) and that developed by Chen (2008) using programming language R.

2. The Dirichelet model

The Dirichelet model describes patterns of repeat purchases of brands within a product category. It models simultaneously the counts of the number of purchases of each brand over a period of time, so that it describes purchase frequency and brand choice at the same time. It assumes that consumers have an experience of the product category, so that they are not influenced by previous purchase and marketing strategies; for this reason, consumer characteristics and marketing-mix instruments are not included in the model. As the market is assumed to be stationary, these effects are already incorporated in each brand market share which influences other brand performance indexes calculated by the model. The market is also assumed to be unsegmented.

Let us consider a sample of n consumers making purchases in a market with g brands. The specification of the Dirichelet model derives from the following assumptions:

1) The number of purchases of each brand j, with $j=1,\ldots,g$, $r_1,\ldots, r_g$, made by the i-th consumer over a succession of purchases, can be modelled by a multinomial distribution with parameters r, $p_1,\ldots, p_g$:

$$P(r_1,...,r_g) = r!\prod_{j=1}^{g}\left(\frac{p_j^{r_j}}{r_j!}\right)$$

where r is the total number of purchases in the product category.
2) The probabilities p_j vary among individuals according to a Dirichelet distribution with parameters $\alpha_1,\ldots,\alpha_g$:

$$f(p_1,..., p_{g-1} \mid \alpha_1,...,\alpha_g) = \frac{\Gamma(\alpha_1 + ... + \alpha_g)}{\Gamma(\alpha_1)...\Gamma(\alpha_g)} p_1^{\alpha_1 - 1}...p_{g-1}^{\alpha_{g-1}-1}(1 - p_1 - ... - p_{g-1})^{\alpha_{g-1}}$$

3) Successive purchases by the ith consumer are independent. The number of purchases n_i made by the ith consumer in each of a succession of equal non-overlapping periods of length T, follows a Poisson distribution with mean $\mu_i T$.
4) Mean purchasing rates vary between individuals according to a Gamma distribution with parameters m and k.
5) Customers' brand-choice probabilities and average-purchase-frequencies are distributed independently over the population.

From assumptions 1-5, it follows that: (i) the number of purchases of the product category made by all individuals in a certain time period follows a Negative-Binomial distribution with mean mT and exponent k; (ii) the number of purchases an individual makes of each of the g brands in a period of time T is given by the following expression, which Goodhardt, Ehrenberg and Chatfield (1984) called the NDB-Dirichelet model:

$$f_{k,m,\alpha_1,...,\alpha_g}(r_1,...,r_g) = f(r \mid m,k) f_{\alpha_1,...,\alpha_g}(r_1,...,r_g \mid r_1 + ... + r_g = r) =$$
$$\frac{(k + r - 1)!}{r!(k-1)!}\left(\frac{k}{m+k}\right)^k \left(1 - \frac{k}{m+k}\right)^r \frac{\Gamma(\alpha_1 + ... + \alpha_g)k!}{\Gamma\left(\sum_{j=1}^{g}\alpha_j + r\right)} \prod_{j=1}^{g}\frac{\Gamma(\alpha_j + r_j)}{r_j!\Gamma(\alpha_j)}$$

The above authors proposed an iterative method for model estimation which requires summary statistics as input values, such as brand penetrations b_j and average purchase rates m_j. The Dirichelet model has been used for many years. Originally, the calculations had to be done by hand, and later with DOS-based software; at the present time various tools are freely available on-line, for example, software developed as an Excel Workbook by Kearns (2000), with User's Guide written by Bound (2009). Another estimation procedure freely available is that composed with programming language R by Chen (2008).

The iterative estimation method needs very simple input data:

(i) with reference to the category, the proportion of people buying product at least once (b) and the average number of purchase occasions recorded for those in the population who purchase the product;

(ii) with reference to the various brands, the proportion of the population buying brand at least once (b_j) – or, alternatively, market shares - and the number of purchases of each brand by those who buy the brand at least once.

When the above data are supplied, the method produces a series of brand performance measures, both for the time period of the data supplied and for other time periods, such as penetration, the percentage of customers buying the brand once and five times, average number of purchases of the brand and of the category per buyer of that brand, share of category requirements, percentage of sole buyers, and percentage of customers repeat buying from period to period.

To activate the model, $g+2$ quantities need to be estimated: m, k, $\alpha_1,\ldots,\alpha_g$. With the g observed per capita purchase rates m_j, the iterative estimation procedure calculates the category purchase rate as $m=\sum_{j=1}^{g} m_j$ and equates the theoretical and observed market shares:

$$\frac{\alpha_j}{\sum_{j=1}^{g}\alpha_j}=\frac{m_j}{\sum_{j=1}^{g}m_j}$$

Parameter k is calculated by fitting an NDB model to the distribution of purchases of the product category.

Both software types considered in this paper (one based on the Excel Workbook and the other in R), estimate the parameters following the method of Goodhardt, Ehrenberg and Chatfield (1984).

Rungie (2003) describes the use of likelihood theory to estimate the parameters of the Dirichelet model, providing an alternative to the standard procedure based on the method of zeros and ones and on marginal moments. The likelihood approach to estimation is more efficient and is well suited to the extensions of the Dirichelet model, e.g., its development into a generalized model, with the inclusion of covariates such as marketing mix variables and consumers' characteristics (Rungie and Goodhardt 2004). In order to write the likelihood function, the data should be in the form of joint frequencies, like those contained in a contingency table with n rows, representing the number of consumers, and g columns, for the number of brands.

Alternatively, the iterative procedure proposed by Goodhardt, Ehrenberg and Chatfield (1984) is computationally easy to use, quick, and requires only aggregated data as input, as access to original panel data is not necessary. Raw panel data cannot always be used since panel operators who measure sales and household consumption provide information only in some aggregate format such as market share, penetration, and average purchase rate with reference to the various brands (Wright et al. 2000). In these situations, the only way to estimate the Dirichelet model is to use the traditional method. Dirichelet modelling also remains a successful and influential approach, and is increasingly being used to provide norms against which brand performance can be interpreted (see, among others, Uncles et al. 1995; Bhattacharya 2000; Ehrenberg et al. 2000).

For the above reasons, it becomes interesting to compare estimation results obtained by applying the various available software to perform iterative estimations.

From the viewpoint of practical applications, the Dirichelet model is useful for various objectives. Estimated values can be used to provide norms for stationary markets, to supply baselines for interpreting change (i.e., non-stationary situations) without having to match the results against a control sample, to help strategic decision-making, and to understand the nature of markets.

3. The data and the Italian beer market

The data used here refer to monthly purchases of 9 brands of beer (Moretti, Heineken, Nastro Azzurro, Dreher, Tuborg, Beck's, Stella Artois, Bud, Kronembourg) by Italian families in the period from August 2001 to July 2004. For each month, we also know the number of families buying each brand, product category, brand market shares, brand and product average purchase rate, and average purchase frequency.

Figure 1 shows average purchases of beer in litres for the 9 brands and the product category. The market shows a clear seasonal pattern, with consumption increasing in summer.

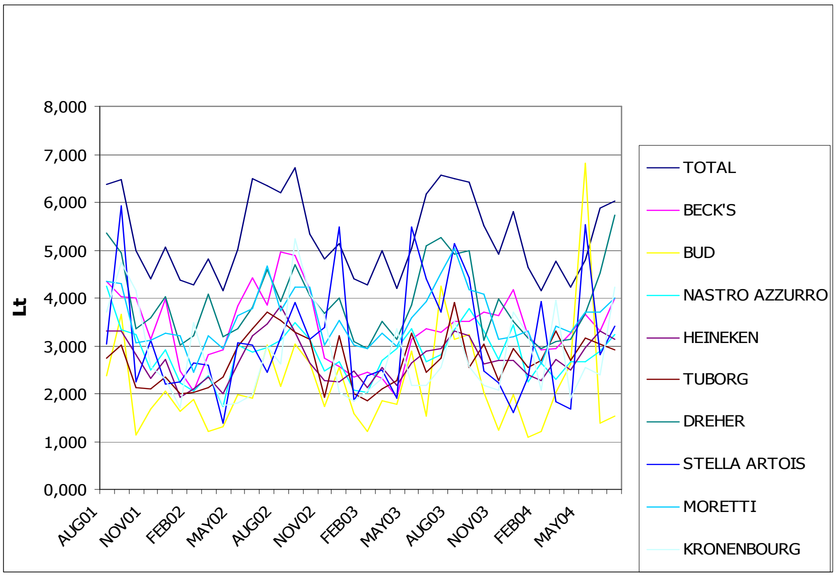

Figure 1. Average purchases per household in lt, August 2001–July 2004

In the last 15 years, Italy's beer market has shown interesting changes in both supply and demand. Total consumption has increased, although average per capita consumption is still substantially lower than in many other European countries such as Greece, Spain and, of course, Germany. Consumption is also linked to warm weather, unlike the situation in Northern Europe, where consumption is distributed throughout the year. Production is concentrated, with a few large groups producing over three-quarters of the total product. Instead, the market is characterized by a quite high number of competing brands. In this paper, the 9 most popular brands are examined and Table 1 lists their average market shares over the study period.

Brand	Moretti	Dreher	Heineken	Beck's	Tuborg	Nastro Azzurro	Kronembourg	Stella Artois	Bud
MS %	12.76	9.09	7.82	3.51	3.27	3.24	1.53	1.17	0.90

Table 1. Average market shares (MS) per brand, August 2001-July 2004

	Δ consumption/litres.			Δ no. of families buying at least once in year		
	Year 1 to year 2	Year 2 to year 3	Year 1 to year 3	Year 1 to year 2	Year 2 to year 3	Year 1 to year 3
Moretti	4,885,310	7,795,561	12,680,871	274,000	810,000	1,084,000
Dreher	1,420,470	-406,283	1,014,187	-3,000	243,000	240,000
Heineken	1,609,209	4,105,277	5,714,486	133,000	-45,000	88,000
Beck's	2,633,307	2,124,514	4,757,821	217,000	389,000	606,000
Tuborg	1,514,408	-762,317	752,091	342,000	103,000	445,000
Nastro Azzurro	759,365	-11,626,250	-10,866,885	-151,000	260,000	109,000
Kronembourg	291,167	-1,344,356	-1,053,189	145,000	-85,000	60,000
Stella Artois	-647,861	-1,111,657	-1,759,518	-132,000	-38,000	-170,000
Bud	439,028	-1,287,586	-848,558	-84,000	-69,000	-153,000
Total	12,904,403	-2,513,097	10,391,306	741,000	1,568,000	2,249,000
Category	33,514,730	18,466,037	51,391,306	408,000	1,576,000	1,984,000

Table 2. Market evolution for 9 brands in the three-year study period

Table 2 compares market evolution for the 9 brands in the three-year analysis. Due to the nature of the available data, year 1 goes from August 2001 to July 2002, year 2 from August 2002 to July 2003, and year 3 from August 2003 to July 2004. Consumption of the product category increased in the period, whereas consumption of our group of brands decreased. Nastro Azzurro, Stella Artois, Bud and Kronembourg contributed to this negative result, whereas Moretti is the brand which most increased its consumption. Regarding the number of families buying the product in the study period our set of brands performed better than the category; Moretti, again, showed the greatest increase.

4. Results

With data on beer consumption in Italy, the Dirichelet model was estimated with two types of softwares applying the method of moments. Both estimate parameters *m*, *k* and *S*. One software application, written in programming language R and called the Dirichelet Package (DP) (Chen 2008), requires as input data category (*b*) and brand (b_j) penetration, category purchase rate (*m*) and brand market shares. The other, based on an Excel Workbook (EW) (Kearns 2002), requires as input data category (*b*) and brand (b_j) penetration, and average purchase frequency for category (*w*) and the various brands (w_j); if brand penetrations are not available, market shares can be used. Table 3 lists estimated parameters *m*, *k* and *S* for the three years with the two types of software and shows that estimated parameters for the NBD part of the model are the same, whereas differences occur in estimating *S*. This result may be due to outlier values for parameters α_j because of the presence in the market of atypical brands (see Table 4; the values for brand Nastro Azzurro in year 3). Bound (2009) suggested excluding such brands when estimating the overall value of *S*[1].

Year 1	Dirichelet Package	Excel Workbook
M	18.48	18.50
k	0.36	0.36
S	0.69	0.90
Year 2		
m	20.03	20.00
k	0.38	0.38
S	0.76	0.90
Year 3		
m	20.87	20.90
k	0.50	0.50
S	0.70	1.60

Table 3. Dirichelet model estimates with the two types of software

Parameters *k* and *S* are characteristics of the product class and may be interpreted as reflecting consumers' heterogeneity. In this market, low *k* values indicate that purchase frequencies vary greatly among buyers, whereas high *S* values mean that purchase probabilities do not differ greatly for the various brands.

Both software types make predictions of the market behaviour estimating some brand performance measures.

The DP estimates category (*b*) and brand (b_j) penetration, average purchase frequency per brand (w_j), average purchase frequency per category per buyers of the brand (w_{Pj}), average number of purchases per brand and its distribution by buyers of the brand, brand penetration and average

[1] The EW calculates a value of *S* separately for each brand so that the prediction of penetration for that brand is exact.

These estimates are then combined and an overall value of *S* is applied to all data purchase frequency among category buyers with a specific frequency range and duplication measures.

	Year 1	Year 2	Year 3
Moretti	0.77	0.79	0.73
Dreher	0.69	0.69	0.71
Heineken	0.60	0.69	0.41
Beck's	0.57	0.62	0.69
Tuborg	1.02	1.34	1.44
Nastro Azzurro	1.84	1.56	28.75
Kronembourg	1.46	2.00	2.35
Stella Artois	1.77	1.68	2.30
Bud	1.31	0.88	1.05

Table 4. Estimated α_j for the 9 brands in three-year study period.

The EW estimates category (b) and brand (b_j) penetration, average purchase frequency per brand (w_j), average purchase frequency per category per buyer of the brand (w_{Pj}), percentage buying the brand once and five times, percentage of sole buyers, rate of purchase of sole buyers, percentage of repeat buying from period to period and duplication measures.

Table 5 lists some estimation results with reference to our market. The parameters estimated with the two types of software are compared with observed values, and the results confirm that Nastro Azzurro is quite atypical in this market, especially in the third year of observation. Following Bound's (2009) suggestion, the model was re-estimated excluding this brand (Table 6).

	b_j								
	Year 1			Year 2			Year 3		
	Observed	DP	EW	observed	DP	EW	observed	DP	EW
Moretti	0.22	0.22	0.24	0.24	0.24	0.25	0.27	0.27	0.36
Dreher	0.15	0.15	0.17	0.15	0.16	0.17	0.16	0.16	0.23
Heineken	0.13	0.13	0.15	0.14	0.14	0.15	0.14	0.17	0.23
Beck's	0.06	0.07	0.07	0.07	0.08	0.09	0.09	0.09	0.13
Tuborg	0.07	0.06	0.07	0.09	0.07	0.07	0.09	0.07	0.09
Nastro Azzurro	0.11	0.07	0.07	0.11	0.08	0.08	0.12	0.03	0.04
Kronembourg	0.04	0.03	0.03	0.05	0.03	0.03	0.04	0.02	0.03
Stella Artois	0.04	0.02	0.04	0.03	0.02	0.02	0.03	0.02	0.02
Bud	0.03	0.02	0.02	0.03	0.02	0.03	0.02	0.02	0.03
	w_j								
Moretti	10.94	11.27	10.39	11.31	10.74	10.74	11.08	11.18	8.41
Dreher	10.90	10.76	9.85	11.34	10.10	10.10	10.45	10.40	7.48
Heineken	11.47	10.65	9.74	11.20	9.97	9.97	13.08	10.43	7.52
Beck's	11.11	10.13	9.21	11.25	9.50	9.50	10.10	9.94	6.95
Tuborg	8.52	10.08	9.15	10.05	9.41	9.41	6.99	9.80	6.77
Nastro Azzurro	6.55	10.16	9.24	10.11	9.47	9.47	2.05	9.58	6.52
Kronembourg	6.81	9.87	8.94	9.79	9.16	9.16	5.22	9.57	6.51
Stella Artois	6.30	9.84	8.91	9.74	9.11	9.11	5.20	9.52	6.46
Bud	7.19	9.82	8.90	9.75	9.13	9.13	7.72	9.57	6.51

Table 5. Penetration and frequency of purchase by brand: observed, and estimated with DP and with EW, for three-year study period, with brand Nastro Azzurro

	b_j								
	Year 1			Year 2			Year 3		
	observed	DP	EW	observed	DP	EW	observed	DP	EW
Moretti	0.22	0.22	0.22	0.24	0.23	0.24	0.27	0.27	0.28
Dreher	0.15	0.15	0.16	0.15	0.16	0.17	0.16	0.16	0.17
Heineken	0.13	0.14	0.15	0.14	0.14	0.15	0.14	0.17	0.18
Beck's	0.06	0.07	0.07	0.07	0.08	0.08	0.09	0.09	0.10
Tuborg	0.07	0.06	0.06	0.09	0.07	0.07	0.09	0.07	0.07
Kronenbourg	0.04	0.03	0.03	0.05	0.03	0.03	0.04	0.02	0.02
Stella Artois	0.04	0.02	0.02	0.03	0.02	0.02	0.03	0.02	0.02
Bud	0.03	0.02	0.02	0.03	0.02	0.02	0.02	0.02	0.02
	w_j								
Moretti	10.94	11.27	10.85	11.31	11.34	11.05	11.08	11.18	10.81
Dreher	10.90	10.76	10.32	11.34	10.72	10.42	10.45	10.40	10.02
Heineken	11.47	10.65	10.22	11.20	10.59	10.29	13.08	10.43	10.05
Beck's	11.11	10.13	9.70	11.25	10.12	9.83	10.10	9.94	9.56
Tuborg	8.52	10.08	9.64	10.05	10.05	9.74	6.99	9.80	9.41
Kronembourg	6.81	9.87	9.43	9.79	9.79	9.49	5.22	9.57	9.47
Stella Artois	6.30	9.84	9.41	9.74	9.74	9.44	5.20	9.52	9.14
Bud	7.19	9.82	9.40	9.75	9.75	9.46	7.72	9.53	9.47

Table 6. Penetration and frequency of purchase by brand: observed, and estimated with DP and with EW, for three-year study period, without brand Nastro Azzurro

A comparison of Tables 5 and 6 immediately shows that the EW procedure is less robust to the presence of atypical brands in the market, where estimates obtained with the DP are only marginally affected when brand Nastro Azzurro is omitted.

Table 7 contains some other brand performance measures that help deeper analysis of the Italian beer market.

	Year 1			Year 2			Year 3		
	100% loyal	Repeat buying %	w_{Pj}	100% loyal	Repeat buying %	w_{Pj}	100% loyal	Repeat buying %	w_{Pj}
Moretti	5.73	84.04	32.20	5.57	84.45	34.11	6.69	85.49	31.20
Dreher	5.45	83.37	32.80	5.24	83.68	34.83	6.22	84.42	31.97
Heineken	5.39	83.22	32.93	5.18	83.52	35.02	6.24	84.47	31.95
Beck's	5.11	82.51	33.59	4.94	82.91	35.63	5.95	83.75	32.48
Tuborg	5.09	82.43	33.65	4.90	82.79	35.73	5.85	83.52	32.63
Kronembourg	4.97	82.12	33.93	4.78	83.45	36.08	5.72	83.16	32.90
Stella Artois	4.96	82.09	33.97	4.75	82.38	36.14	5.69	83.10	32.96
Bud	4.95	82.07	33.99	4.76	82.41	36.13	5.70	83.12	32.94

Table 6. Percentage of consumers 100% loyal, percentage of consumers who repeat purchase in the period and average frequency of category purchase by buyers of the brand, for three-year study period, without brand Nastro Azzurro

The first evidence emerging from Tables 5, 6 and 7 is that the Italian beer market is segmented. Two segments can be identified: one composed of brands Tuborg, Kronembourg, Stella Artois and Beck's, which show estimated penetrations lower than observed ones, estimated average purchase frequencies of the brand higher than observed ones, the lowest percentages of loyal customers, and the lowest percentages of repeat buying but the highest average purchase frequency of the category. This segment may be defined as a mass consumption, one with low loyalty and many light buyers. The second segment is composed of brands Moretti, Dreher, Heineken and Beck's. They show estimated penetrations equal to or higher than observed ones, estimated purchase frequencies for the brand lower than observed ones, the highest percentages of loyal customers, the highest percentages of repeat buying and the lowest averages purchase frequencies for the category, and may be defined as a niche market segment with many heavy buyers.

5. Some concluding remarks

Application of the Dirichelet model to the Italian beer market shows that it is segmented into at least two parts, massive consumption on one hand and a niche, in which consumers behave quite differently. Many applications of the Dirichelet model have shown that, even when the market is not quite steady, or when some clustering occurs, the model mostly still holds and it provides useful benchmarks (see, for example, Ehrenberg et al. 2004). In this paper, I show again how the model can be used to assess how existing brands are performing.

The model is very parsimonious, at least when the method of moments is used for parameters estimation. In this case only a few numerical inputs are needed, typically penetrations and average purchase frequencies of the category and the various brands.

Results obtained with two available types of software for the methods of moments are compared here. The software based on Excel Workbook turns out to be less robust to the presence of atypical brands on the market. Lack of robustness does not affect estimation of the parameters of the NDB component of the model but, as it does affect all other parameters, it is advisable to eliminate such brands when conducting analysis, for reliable results. The software written with in R is more robust because estimation of parameter *S* of the Dirichelet distribution is made directly, whereas the Excel procedure estimates *S* as a function of estimated α_js.

It would be interesting at this point to compare parameters estimated by maximizing the loglikelihood function with those presented here. However, this exercise, would require raw panel data which are currently not available.

References

Bhattacharya, C.B. (2000). Is your brand's loyalty too much, too little, or just right? International Journal of Research in Marketing, 14, 421-435.

Bound, J. (2009). User's Guide to DIRICHELET. Marketing Bulletin, Technical Note 1.

Chen, F. (2008). The Dirichelet Package.

Ehrenberg, A.S.C., Barnard, N.R., Sharp, B. (2000). Decision models or descriptive models? International Journal of Research in Marketing, 17, 147-158.

Ehrenberg, A.S.C., Uncles, M.D., Goodhardt, G.J. (2002). Understanding brand performance measures: using Dirichelet benchmarks. Journal of Business Research, 57, 1307-1325.

Goodhardt, G.J., Ehrenberg, A.S.C., Chatfield, C. (1984). The Dirichelet: a comprehensive model of buying behaviour. Journal of the Royal Statistical Society, Series A, 147 (5), 621-655.

Kearns, Z. (2002). Dirichelet No Solver Software, Massey University, New Zealand.

Rungie, C. (2003). How to estimate the parameters of the Dirichelet model using likelihood theory in Excel. Marketing Bulletin, 14, Technical Note 3.

Rungie, C., Goodhardt, G. (2004). Calculation of the theoretical brand performance measures from the parameters of the Dirichelet model. Marketing Bulletin, 15, Technical Note 2.

Uncles, M., Ehrenberg, A..S.C., Hammond, K. (1995). Patterns of buyer behaviour: regularities, models, and extensions, Marketing Science, 14, G71-G78.

Wright, M., Sharp, A., Sharp, B. (2000). Market statistics for the Dirichelet model: using the Juster scale to replace panel data. International Journal of Research in Marketing, 19, 81-90

Working Paper Series
Department of Statistical Sciences, University of Padua

You may order copies of the working papers from by emailing to wp@stat.unipd.it
Most of the working papers can also be found at the following url: http://wp.stat.unipd.it