User's Guide to DIRICHLET

John Bound

DIRICHLET is an Excel-based program for fitting the Dirichlet[i] model to data recording individual purchases of a number of entities, usually brands, in a particular product category. It was written by Zane Kearns (Kearns 2002). The purpose of this Guide is to supplement the instructions given in the program itself, which explain how to enter data and fit the model. Some theoretical points are discussed in the Notes at the end. A comprehensive theoretical background is given in the book 'Repeat Buying' by Professor A.S.C. Ehrenberg[ii].

Keywords: Dirichlet, repeat purchase, consumer panel data

1. Introduction

This Guide covers some of the more practical aspects of using the Dirichlet Model to get more out of your consumer panel data. We start by considering why you should want to fit a Dirichlet distribution to your data at all.

2. Why it is useful to fit a Dirichlet distribution to your data?

The regular input to brand management from consumer panels is market size, and brand share, penetration and rate of purchase data for the whole market and for individual brands. These top-line data are also broken down for sub-markets such as different areas, types of outlet, and demographic groupings of purchasers. Such analyses received and examined period by period, and compared over time, are the most common use for panel data, but make no use of the essential feature of panels, which is that panels track individual purchasers over time. This feature enables individual and average rates of penetration, purchase and brand duplication[iii] over defined time periods to be seen.

Such analyses are normally provided by panel operators in various formats to supplement the period by period reports. It is a great help in analysing these to have some established benchmarks with which to compare the tabulated results. Many of these measures vary greatly between brands in ways evidently related to brand share. In particular, smaller brands

in general may be expected to show lower loyalty measure predictions. For any one brand on the contrary the predictions vary very little from one period to another.

In a wide range of markets in many countries the Dirichlet model works very well in providing such benchmarks (see Section 10). The predictions for the various measures for each brand will not of course agree exactly with the tabulated observations. For some of these deviations knowledge of the particular market may sometimes suggest a cause, but we suggest that analysis of data for other periods should be carried out to see if these deviations are in fact reproducible before too much reliance is placed on marketing interpretations. There is at present no general theory to help in this. There are also some measures which, as noted in Section 5, the Dirichlet model generally does not predict so well.

3. About the Dirichlet model and DIRICHLET program

The Dirichlet Model[iv] incorporates a few basic and believable assumptions[v] about the way people buy. The justification of the model is, we stress, not just the plausibility of the assumptions, but the observed fact that its predictions have been found to fit observed data. The fact that the assumptions seem close to reality helps us to understand the underlying patterns of consumer buying behaviour.

The Dirichlet Model has indeed been around for many years. Originally, the calculations had to be done by hand, and only a small number of big corporations were able to invest in it. In 1988 DOS-based software[vi] made the calculations easier to do, either on a mainframe or on a PC, and the model became more widely used. Now newer Excel-based software means that anyone with summary consumer or similar panel data can readily make use of the model. The Dirichlet software is available on Open General Licence[vii] as an Excel Workbook. There is a worksheet containing instructions for operating the program. Some worksheets used in calculation are hidden.

There are two major aspects of the Dirichlet Model which we should like to make clear from the beginning. The first is that it does not predict brand share, but uses data about individual brand shares and numbers of purchase occasions together with total category data as inputs to make predictions for the behaviour of such a brand of that share in such a product category. The second is that the model makes the assumption of a stationary market. By this we mean a

market in which peoples' purchase patterns remain the same. This is not to say that their purchases are always the same from one period to another, but that the pattern behind their purchases is constant. The total market and brand shares are thus also assumed to be constant. Although this assumption of a stationary market is never exactly true, it seems usually to be true enough to make the model fit well.

4. What you need for the Dirichlet Model

Data needs are simple[viii]. To fit the model for the chosen product category in a selected period you need for the product category in total:

1. The proportion (that is percentage divided by 100) of the population buying the category at all (B).
2. The average number of purchase occasions of the product category recorded for those in the population who purchase it all (W).

and for one or more individual brands:

1. The proportion of the population buying each brand (I) at all (b_I).
2. The average number of purchases of each brand by those who buy it at all (w_I).

Distinguish the penetration of the brand among the population as a whole as used above from the relative penetration of the brand among the buyers of the product category. Note also that the number of purchase data are for the number of purchase occasions[ix], not for amounts of expenditure, or volume of product.

If the data were exactly distributed in accordance with the Dirichlet distribution, data for the total product category and one brand only would be needed, but in practice the data for a number of brands are needed. It may, as we shall see, give a better fit if some brands are missed out from the fitting. The brand shares may be input instead of the individual brand penetrations, but the program always requires the rate of purchase for the individual brands (w_I) and for the whole product category (W). The program will calculate and input penetration if the rates of purchase and brand share are entered in the column provided in place of penetration.

Observed figures for the input period corresponding with the norms produced by the Dirichlet Model are also needed to examine deviations from the norms. It is helpful if the brands include an 'All other' so that their shares add to 100 per cent. It is also very helpful in seeing the patterns if the brands are entered in order of brand share.

5. What the Dirichlet Model can do

When these data for any product category are supplied to DIRICHLET, the program produces a number of predicted purchasing, repurchase and duplication measures for the time period of the data supplied. Predictions for other time periods may also be produced, a useful facility which is explained below.

The default predicted norms shown in a Brand Performance Audit table are currently for each brand:

1. Penetration of purchasers (percentage buying the brand at all).
2. Percentage buying the brand once and five times in the designated period.
3. Average number of purchases of the brand per buyer of the brand.
4. Average number and distribution of the numbers of purchases of the category by buyers of the brand.
5. Share of category requirement (this follows from the average number of purchases).
6. Percentage of sole buyers. (A sole buyer is one who buys only one brand in the category in the period).
7. Rate of purchase of sole buyers.
8. Percentage repeat buying period to period.

Other estimates are available from detailed tables of results. Only one set of these will be discussed here, the tables of Duplication of Purchase or Brand Duplication.

The program provides no measures of goodness of fit[x]. The product category penetration and rate of purchase are used in fitting, and the predictions for these therefore equal the observed values. The predictions for the penetration of individual brands are also used in fitting, but do not in practice all fit quite exactly[xi]. The brand shares are predicted exactly. For the rate

of purchase and the other predictions for each brand there are normally deviations, mostly small and irregular.

Some more substantial deviations may be the result of particular unusual situations, or non-stationarity beyond that which can be coped with by the model. Others may be the result of a poor fit which can be improved by changing some of the fitting parameters, as described in Section 5 below.

There are some measures which the Dirichlet habitually gives not so good a fit[xii]. For these the observed values tend to be:

Less than the predicted number of:

1. Very frequent buyers. This seems to be because few grocery products are bought more than once a week.
2. Repeat buyers from one 13-week period to another by 5 to 10 percentage points. This suggests that even in stationary markets there is a 'leaky bucket': the regular buyers of a brand are gradually replaced by others.
3. 'Medium' buyers, as opposed to 'heavy' and 'light', of a product category.

Greater than the predicted number of:

1. Average annual purchases for some market leaders by one purchase or so. This may be because large brands suffer less from out-of-stock situations.

2. Annual purchases rates of 100% loyal buyers of brands by one or two purchases. There is no obvious explanation, but it is often found.

The Dirichlet Model has the remarkable property that it can produce estimates for any other time period as well as for the input time period. So if data are input for 12 weeks, by simply changing the T value input in DIRICHLET from the default 1, estimates for 13 or 52 or any other number of weeks may easily be produced. Such analysis is valuable in 'what-if' situations such changes in brand share. Although these are by definition non-stationary markets, the model still seems to work in its predictions.

Another useful property is that if brands are aggregated, as for example all the brands of a particular manufacturer, or divided, as for example into the individual pack sizes of a brand, and the appropriate data input, the fitted model and predictions remain unchanged. This also enables hypothetical new brands to be studied. Brand shares may be altered to include the planned share of the new brand, and predictions re-run using already established model parameters.

6. What the Dirichlet Model cannot do

The Dirichlet Model is a map, and not an itinerary. So we repeat that penetrations or market shares (which are closely related to one another) are input, not output. The Dirichlet Model then tells us how a brand of that size may be expected to behave in terms of the measures predicted, given the input information about both the total category and the other brands. If the market is changing the Dirichlet Model may not work at all, although it seems resistant to such changes[xiii].

7. How to work the DIRICHLET program

You need the summary data, a computer with Microsoft Excel, and the DIRICHLET program workbook (see attached Dirichlet VB.xls). A worksheet of instructions in the program tells you which buttons to press. It is desirable to save a working version of the DIRICHLET workbook under some other name when starting a particular job to avoid changes made causing inconvenience later, although the program offers facilities for entering a number of new and separate sets of input data in the one workbook.

This section gives advice on how to select data and use some of the options. Section 8 shows the outputs derived from a set of real data.

7.1 Data entry

To repeat from Section 2: for the chosen product category in a selected period you need for the category in total and for at least one brand:

1. The proportion (that is percentage divided by 100) of the population buying the category at all (B). Note again that these are absolute penetrations, that is the penetration among the whole population and not just the buyers of the product category[xiv].
2. The average number of purchases of the product category made by those who purchase it all (W).
3. The proportion of the population buying each brand (I) at all. (b_I).
4. The average number of purchases of each brand by those who buy it at all (w_I).

To see the deviations of the norms from the observed, the observed figures may be conveniently inserted in the results worksheets after the norms are calculated. The program shows automatically as 'observed' those figures input for the fitting.

Up to thirty brands may be analysed, but normally fewer than a dozen are required. As noted above, the additive property of the Dirichlet means that brand variants may be grouped together, or an 'all other' category created.

As observed in Section 4, it is a help if the brand shares analysed are arranged in order of size and convenient also if the shares total 100%. Inspection of the data before input will then normally show the penetrations for the various brands declining much in line with brand shares, and the rates of purchase very much the same for all brands, but declining somewhat with brand share. If this is not so, the Dirichlet distribution will not fit at all well. Aberrant brands may however be omitted in fitting, as described below, and if small in share will affect upset the results little. The effects of sampling fluctuations in small samples may thus be avoided.

The length of period chosen for the input data should be great enough to allow a number of purchases of the category by the average buyer, and to take in any seasonal fluctuations. Since the model assumes a stationary market, the average of several such periods may be usefully taken if data for a longer period are available, so reducing sampling and other fluctuations. Non-typical periods such as Christmas may be excluded.

The necessary summary data[xv] can normally be readily produced by panel operators[xvi], and require only entering into to the data input sheet and the observed data if required into the output sheets of DIRICHLET.

If data for the penetration of individual brands are not available, but the brand shares and rates of purchases are available there is as previously noted an option to enter brand shares rather than penetrations.[xvii]

7.2 Time period for Predictions

The default time period for predictions is the period for the entered data, and this may often be all that is needed. If predictions for other periods are required, the DIRICHLET input parameter T may be altered from the default value 1 to whatever fraction or multiple of 1 represents the desired period. If the input data refer to 13 weeks, and predictions are required for a year, t is set as 52/13 = 4. If the predictions are required for 4 weeks, T is set at 4/13 = .308. A space is provided for recording whether the numbers apply to weeks or months etc.

7.3 The S Parameter

The other option provided is for the value of S to be used. The parameter S is one of the few[xviii] which define a particular Dirichlet Distribution and has no simple meaning. It appears to be some measure of diversity of purchasing behaviour.

DIRICHLET calculates a value of S separately for each brand for which data are entered[xix]. It then offers a choice between two methods of calculation for an overall value of S to be used in producing predictions for all brands. The first is the old-established method described in Repeat Buying. This method averages the estimates of S for the individual brands weighted by brand share, to give an overall estimate of S, which is then used in fitting.

The other option is for weighting the individual estimates of S for each brand so as to minimise the sum of the squares of the differences between in the estimated and observed values of the proportion of buyers of each brand. We have no experience of the relative merits of these methods.

Small or atypical brands, or "All Other" categories, may give S values out of line with those of the other brands. The option is provided of excluding such brands in the calculation of the overall estimate of S. It is usually desirable however to use several brands to estimate S. Any desired value for S may also be put in manually. The proof of the pudding is in the eating: whatever estimate of S which gives the best fit in overall judgement, is best. Experience has helpfully shown that in general, the precise value of S used makes little difference to the predictions, as is illustrated in Table 8 below.

8. An example of results for established brands

The example we use is of some data on the US Instant Coffee market in 1992 published by Hallberg[xx]. We describe the fitting process and comparisons with the observed data. This market had shrunk greatly in the preceding ten years, and Hallberg gave data for a selection of brands based on their earlier importance. The fit is not remarkably good, so serves as a practical example of what may be expected in a not very stationary market, though the poor fit is not necessarily due to that.

We start by supplying data as in the following table. Capital letters as before denote the total category and small letters the individual brand. B indicates the penetration for the category, and the b values for the penetration for the individual brands. These are entered as proportions between 0 and 1, but then appear as percentages. Similarly, W is the rate of sale for the category, and the w values are for the individual brands. We also indicate by the column of Y's that we wish every brand to be used in the calculation of the estimate of S. If any brand ism not be used, the Y opposite its name is omitted. It is helpful to enter the brands in order of size, preferably brand share, but using b is much the same. This shows up variations from the usual patterns of association with brand size. The 'Other brands' are treated as a single brand in the data. A small brand like Brim could be included in 'Other brands' as described above in Section 6.

The market share column is used only as an alternative to entry of penetration data. We also enter in the spaces provided that our data are for 52 weeks, and by leaving unchanged the default value of T as 1 indicate that we require predictions for 52 weeks initially. We may also enter a description of the job. The market share column is used only as an alternative to entry of penetration data. We also enter in the spaces provided that our data are for 52 weeks, and by

leaving unchanged the default value of T as 1 indicate that we require predictions for 52 weeks initially. We may also enter a description of the job.

Table 1. Data entry table

DATA ENTRY				
Product Category and Brand Information				
Category Total	b **31%**	w **5.0**		
Brand	b	w	Market Share	Use To Est.
Folgers	11%	3.2		Y
Maxwell House	10%	3.3		Y
Tasters Choice	9%	2.8		Y
Other brands	8%	3.0		Y
Nescafe	6%	2.7		Y
Sanka	5%	3.0		Y
Maxim	0%	4.5		Y
High Point	1%	2.6		Y
Brim	0%	2.1		Y

This table itself is instructive. As the penetrations diminish, the rates of purchase (w) slightly. This is the common pattern. The first results are on the input data page, and show the estimates of S for the brand and their average according to the method of averaging chosen. The columns look rather like this:

Table 2. Estimates of S are examined.

Brand	S^	Weighted S^
Folgers	.9	.2
Maxwell House	.8	.2
Tasters Choice	1.4	.2
Other brands	1.0	.2
Nescafe	1.3	.1
Sanka	.8	.1
Maxim	.1	.0
High Point	1.1	.0
Brim	2.1	.0
Average as estimate of S		**1.0**

These estimates of S are all much the same, except for Brim. When weighted by the small shares of this brand this estimate has little influence, so we need not worry about this. From raw data not shown here we see in fact that there are only 27 buyers for Maxim, 34 for High

Point and 17 for Brim out of a total sample of about 2,500 buyers. It seems likely therefore that the recorded purchasing of all these brands may not be typical. But of course, this is only speculation. However, we accept the offered value of S. Of course as we show below we can readily do it all again with a different S value if we want to see if that improves the overall fit.

We now look at the Brand Performance Audit table which gives the principal results in a convenient comparative form for each brand. We have first taken for our example in Table 3 the results for brand share and penetration. The brand share follows directly from the input data.

Table 3. Results for penetrations and percentages buying

	Brand Share	Penetration		% Buying			
				Once		Five +	
	O	O*	**P**	O	**P**	O	**P**
Any instant	100.0	31	**31**	28	**31**	37	**34**
Folgers	23.5	11	**12**	46	**46**	18	**19**
Maxwell House	21.5	10	**11**	40	**47**	20	**19**
Tasters Choice	16.8	9	**9**	48	**48**	18	**18**
Other brands	15.7	8	**8**	49	**48**	16	**17**
Nescafe	11.0	6	**6**	49	**49**	12	**16**
Sanka	9.3	5	**5**	45	**50**	19	**16**
Maxim	0.9	0	**1**	31	**53**	23	**14**
High Point	0.8	1	**0**	59	**53**	16	**14**
Brim	0.3	0	**0**	61	**53**	15	**14**
Average Brand	11.1	6	**6**	48	**49**	17	**16**

O = Observed P = Predicted *These Observed figures were used in the fitting

The predicted penetrations are close to the observed data, but not precisely the same. The averaged S value used for the predictions cannot give an exactly correct estimate of penetration for each brand. The differences are usually very small Maxim shows up with fewer than predicted buying once, and more than predicted buying five or more times. The percentages of buyers of each brand once (sole buyers) and five times are given[xxi].

We turn now to the next section of the Brand profiles. Table 4 shows the rates of purchase[xxii] and repurchase.

Table 4. The rates of purchase and repurchase

	Purchases per Buyer				Share of category requirements		100% loyal				Repeat buying 52 weeks	
	of the brand		Of the category				Percentage		Rate of purchase			
	O*	**P**	O	**P**	O	**P**	O	**P**	O	**P**	O	**P**
Purchasers of Any brand	*5.0*	***5.0***	*5.0*	***5.0***	*100*	***100***	*100*	***100***	*5.0*	***5.0***	*N/A*	***N/A***
Folgers	3.2	**3.2**	6.4	**6.6**	50	**48**	40	**36**	3.6	**2.5**	N/A	**66**
Maxwell House	3.3	**3.1**	6.9	**6.7**	49	**47**	31	**35**	3.8	**2.5**	N/A	**66**
Tasters Choice	2.8	**3.0**	6.4	**6.8**	44	**44**	41	**33**	2.9	**2.4**	N/A	**65**
Other Brands	3.0	**3.0**	7.1	**6.8**	42	**44**	36	**32**	3.4	**2.3**	N/A	**64**
Nescafe	2.7	**2.9**	7.3	**7.0**	37	**41**	20	**30**	3.8	**2.2**	N/A	**63**
Sanka	3.0	**2.8**	7.9	**7.1**	38	**40**	43	**30**	3.2	**2.2**	N/A	**63**
Maxim	4.5	**2.7**	8.3	**7.3**	54	**36**	30	**26**	4.0	**2.1**	N/A	**60**
High Point	2.6	**2.7**	5.5	**7.3**	47	**36**	51	**26**	2.6	**2.1**	N/A	**60**
Brim	2.1	**2.7**	5.0	**7.4**	42	**36**	28	**26**	2.9	**2.1**	N/A	**60**
Average Brand	3.0	**2.9**	6.8	**7.0**	44.8	**41.4**	35.7	**30.4**	2.3	**2.3**	N/A	**63**

*These Observed figures were used in the fitting

Here we see that the average number of purchases of the brand is less for smaller brands, again as usual following the Double Jeopardy pattern. Share of category requirements follows the same pattern: we note that no brand supplies more than half its total customer requirements, that only about one-third of the buyers of each brand are 100% loyal, and these buy somewhat less than the average buyer of the brand. The predictions for sole buyers do not fit well: the model predicts fewer.

The general lesson from Table 3 and Table 4 is that for brands with the largest numbers buying at all tend to have more purchases per buyer, a higher share of category requirements, and a smaller percentage of their buyers are 100% loyal. The smaller brands suffer therefore not only from having fewer buyers, but these buyers purchase less often, and are less loyal. This is the familiar story of Double Jeopardy for small brands. Not all the brands follow this pattern neatly, although the average brand is fairly close.

When we come to the repeat buying figures for the 52 weeks, we have a prediction, but since there are only 52 weeks data, we have no observed data to compare. We do however have 13 week repurchase figures, so all we have to do is to re-run the model using T = .25 to give repurchase predictions for 13 weeks. Table 5 compares these with the observed figures.

Table 5. Repeat buying over 13 week periods.

	Observed	**Predicted**
Folgers	37	**48**
Maxwell House	39	**48**
Tasters Choice	34	**47**
Other Brands	33	**48**
Nescafe	29	**45**
Sanka	34	**45**
Maxim	31	**42**
High Point	53	**42**
Brim	18	**42**
Average Brand	34	**45**

The fit is here is not very good at all, even bearing in mind the general tendency noted in Section 5 for the Dirichlet to over-estimate observed repurchase. Perhaps the market is not stationary, and a study of trends in market size might confirm this. The observed repeat buying percentages are generally smaller for the smaller brands, in line with the theoretical trend. Those concerned with smaller brands should not therefore be concerned unless, as for Brim, their figure is way out of line.

The final set of results considered is the Brand Duplication of Purchase tables. These are often helpful, pointing to possible sub-markets. We look at the conditional repurchase figures predicted in the 'Tables' worksheet. These show the percentages of the purchasers of a brand who in a particular period make purchases of specified other brands.

Predictions are shown for the conditional duplications. In these the theoretical figures are constant in each column[xxiii]. When the observed figures are compared with the predictions a sub-pattern of deviations is sometimes seen. Groups of brands have a higher than expected duplication between one another indicating possible functional differences between this group and other remaining brands[xxiv].

Table 6. Predicted duplication of purchase

	Who also buy								
	Folgers	Maxwell House	Tasters Choice	Other	Nescafe	Sanka	Maxim	High Point	Brim
Purchasers of									
Folgers		26	22	21	15	14	1	3	0
Maxwell House	29		22	21	15	14	1	3	0
Tasters Choice	29	26		21	15	14	1	3	0
Other	29	26	22		15	14	1	3	0
Nescafe	29	26	22	21		14	1	3	0
Sanka	29	26	22	21	15		1	3	0
Maxim	29	26	22	21	15	14		3	0
High Point	29	26	22	21	15	14	1		0
Brim	29	26	22	21	15	14	1	3	
Average Brand	29	26	22	21	15	14	1	3	0

The figures on each line show the percentage of buyers of that brand buying the brand named in the column. The 100% figures for those buying the brand itself are omitted. Now let us look at the actual observed purchasing duplication.

Table 7. Observed duplication of purchase

	Who also buy								
	Folgers	Maxwell House	Tasters Choice	Other	Nescafe	Sanka	Maxim	High Point	Brim
Purchases of									
Folgers		31	24	21	21	12	0	1	0
Maxwell House	35		26	28	27	13	1	1	1
Tasters Choice	29	28		20	23	11	2	1	1
Other	31	34	24		27	10	1	1	0
Nescafe	38	43	33	36		17	1	1	1
Sanka	28	27	22	17	22		1	1	0
Maxim	13	27	46	12	14	17		0	0
High Point	25	19	9	11	14	5	0		2
Brim	39	24	20	8	0	5	0	5	
Average Brand	30	29	26	19	19	11	1	1	1

The averages for each column agree reasonably well, although the individual duplication predictions vary quite widely from the observed. There is no sign of a group of brands which all have higher than expected duplication, so the market appears non-partitioned, an important conclusion. As noted in Section 3 above, if the market is segmented, the Dirichlet assumptions do not apply, but in practice the model works.

9. An example of the effect of changing the value of S used for fitting.

The removal of some atypical brands from the calculation of S, or the use of some value derived from earlier work makes surprisingly little difference to the predictions. Let us put in an arbitrary value of S =2 in place of the fitted 1.04 and see what that does to the 52 weeks predictions.

Table 8. The effect on predictions of changing S from 1.04 to 2

	Penetration			% Buying Once		
	Obser-ved	**Pred. for S = 1.04**	**Pred. for S = 2**	Obser-ved	**Pred. for S = 1.04**	**Pred. for S = 2**
Folgers	11	**12**	**13**	46	**46**	**49**
Maxwell House	10	**11**	**13**	40	**47**	**49**
Tasters Choice	9	**9**	**10**	48	**48**	**52**
Other	8	**8**	**10**	49	**48**	**52**
Nescafe	6	**6**	**7**	49	**49**	**55**
Sanka	5	**5**	**6**	45	**50**	**55**
Maxim	0	**1**	**1**	31	**53**	**60**
High Point	1	**0**	**1**	59	**53**	**60**
Brim	0	**0**	**0**	61	**53**	**60**
Average Brand	6	**6**	**7**	48	**49**	**55**

The changes, although substantial, do not destroy the pattern. The change due to the exclusion of any one brand would not be nearly as great. Other measures show similar scales of change. The choice of how precisely to calculate of S is not in practice a great problem.

10. Experience of applications

The predictability of the patterns of buyer behaviour for brands – how many consumers buy at all, how often they buy, what else they buy – has been well documented over many years for frequently bought product categories such as groceries. Recent extensions have covered behaviour relating to product variants such as size, flavour and form, less frequently bought categories such as cars and subscription markets such as credit cards and bank accounts, where consumers may use the product frequently but rarely change the brand they use, even for those with an annual renewal.

In all these instances, the patterns of choice are well described by a single model of choice behaviour, the Dirichlet. This parsimonious model assumes consumers choose from a small portfolio of the available options (split loyalty), with on-going as-if fixed and random propensities to choose any one entity (e.g. brand X six times out of ten). Propensities differ greatly from consumer to consumer but are highly predictable in total and can be summarised using standard distributions. As we have seen above the theory predicts a large number of performance measures for various time periods.

The simplified conclusion from this work is that the over-riding determinant of how many customers an entity has (penetration) and how loyal they are (frequency) is brand size (share). Frequency usually varies much less than penetration (the well documented Double Jeopardy phenomenon). Major deviations from this are very rare, e.g. almost no particular brands or variants have much higher than expected loyalty from a smaller than expected customer base. Even the duplications between variants are pretty much as predicted by share alone, although there are sometimes partitions or groupings of brands that duplicate more with each other. Mostly partitions are category specific and so need to be established empirically. Against this background of knowledge of other markets, patterns of usage in your own market can quickly be assimilated.

Notes.

[i] The model is more exactly described as the Negative Binomial Distribution-Dirichlet. The distribution of the total number of purchases in the product category is assumed to be a Negative Binomial.

[ii] Ehrenberg, A.S.C., (1972, New edition 1988), Repeat Buying, Charles Griffin & Co. Ltd., London, Oxford University Press, New York. This is now out of print but the complete text is available at http://www.empgens.com/index.html

[iii] Brand Duplication tables show the percentages of the purchasers of a brand who in a particular period make purchases of specified other brands. Brand switching tables show the percentages of the purchasers of a brand who in a following period or on a different purchasing occasion make purchases of the same or other specified brands.

[iv] Other related models such as the Empirical Dirichlet Model may also be fitted, and Repeat Buying describes these. In practice they are seldom needed. One possible need for the Empirical Dirichlet we may note. The NBD-Dirichlet Model is not appropriate for the prediction of 'two-purchase' data, such as arise when the most recent and the previous purchase are the only two purchases considered. The reason is that the model assumes the distribution of the number of purchases in the product category to follow the Negative Binomial Distribution. Two purchase data do not. The Empirical Dirichlet Model, where some other distribution is specified, could be used instead. This topic is beyond the scope of this Guide

[v] There are five assumptions in the NBD-Dirichlet model. The first two relate to buying the product category.

1. Each consumer is assumed to buy the product category at some steady long-term rate, and these rates to be distributed between individuals following a smooth Gamma type distribution. The Gamma typically has many light buyers and few heavy ones, unless the mean is very high.
2. A given consumer's specific purchases of the category are assumed to be spread over time independently of when the previous purchase was made and in a Poisson distribution.

The third and fourth assumptions are about brand choice.

3. Heterogeneity in consumers' brand-choice probabilities is assumed to follow a smooth Beta distribution of the multivariate Dirichlet type.
4. On any one category purchase occasion a consumer is assumed to choose a brand as if randomly with his or her own fixed brand choice probabilities. in a zero-order multinomial Multinomial distribution of brand choice.

The fifth assumption is about the relationship between product category buying and brand choice.

5. Purchase incidence and brand choice are assumed to be independent.

The theory is completely set out in :

Goodhardt, G. J., Ehrenberg, A.S.C. and Chatfield, C. (1984), "The Dirichlet: A Comprehensive Model of Buying Behaviour", *Journal of the Royal Statistical Society Series A,* **147**, 621-655.

[vi] The BUYER program originally developed by Professor Mark Uncles of the University of New South Wales can use unsummarised panel data to produce predictions either for the Dirichlet or the Empirical-Dirichlet model. The Dirichlet predictions are however not quite the same as those produced by DIRICHLET. DIRICHLET uses less truncation of the tails of distributions.

[vii] The DIRICHLET program was written Dr. Zane Kearns formerly of the Marketing Department of Massey University, New Zealand. [No URL available at the moment]

[viii] The three input measures considered here are brand share, penetration and rate of purchase. Given penetration (B) and rate of purchase (W) for the total category, the mean number of purchases among those buying the product category at all (M) is determined. Knowledge of B and either W or M is always required. Fitting the model requires input of B, M, and for at least one brand the volume of the individual brand (m), and its penetration (b). For an individual brand there are three combinations of summary data which will produce m and b.

1. The penetration (b) of the individual brand and the rate of purchase (w) of the brand. This is the default input provided for in DIRICHLET
2. Brand share and rate of purchase (w). Given M, brand share defines m. From m and w the value of b is fixed. This is the alternative data entry form in Dirichlet.
3. Brand share and penetration (b). A separate calculation may be made of m and then of w. The default entry of b and w may then be used.

[ix] The purchase occasion is not necessarily the same as the number of purchases. A purchaser might buy two or more packs of the product at one time, but this would amount to one purchase occasion. Consumer panel operators distinguish the two values. Treatment of two identical purchases made on the same day may however vary between panel operators. We write 'purchases' in this paper for brevity.

[x] The correlation may be calculated between the expected and observed values in any of the output tables, and will give a measure of goodness of fit for that table. That enables

comparison of the fit of that measure for the particular choices of options. Since there are many output tables of varying importance, an overall measure of fit for many measures is not helpful.

[xi] The fitting method used by Dirichlet follows that set out in Repeat Buying, and is in outline as follows. For each brand successively the category data and the input data for the individual brand are used to estimate a value for S such that the prediction of penetration for that brand is exact. These estimates for S are then combined and an overall value of S applied to all the data. The predictions of penetration for each brand using this overall estimate are then not exact but very close to the observed. For each brand the volume and the brand share are known inputs. The predicted penetration and the brand volume are then used to calculate a rate of purchase of the brand.

The brand share is thus exactly predicted. The penetration is very closely predicted, and the rate of purchase rather less so. Other measures are predicted with varying exactness.

Other methods of fitting may be used, but do not have these properties. The discussion of these methods is beyond the scope of this paper.

[xii] See Ehrenberg, A.S.C., Uncles, M.D., and Goodhardt, G.J., (2004), "Understanding Brand Performance Measures: Using Dirichlet Benchmarks", Journal *of Business Research*, 57, 1307-1325.

[xiii] The assumptions of the Dirichlet Model imply that the market concerned is unsegmented (with no special groupings of consumers) and non-partitioned (with no special clusters of brands). In practice these restrictions do not need to apply closely for the model to fit satisfactorily.

[xiv] In practice the accuracy of the of estimate of the category penetration is not material, providing that the individual brand penetrations are similarly scaled. That is to say it is the relative penetrations which are important.

[xv] Data may also be used from one-time surveys, and the Juster scale for the probability of purchase appears to give usable estimates of brand penetrations.

[xvi] If only unsummarised panel data are available, it may be necessary to aggregate large numbers of stock-keeping units and generally to clean the data. The BUYER program described in Note 6 or other software may be then used to produce the summary data.

[xvii] See Note 8.

[xviii] The parameters which describe a Dirichlet distribution are M, A and S. B and W are the penetration and rate of purchase of the product category and enable the calculation of the parameter M (the mean rate of purchase of the category by purchasers of the category,

and the NBD parameter A. The parameter S is derived from M, A, b the penetration of one band, and m, the mean number of purchases of the product by that brand.

[xix] See Note 11.

[xx] Hallberg (1992)

[xxi] If these numbers are inappropriate for categories with either very frequent or very infrequent purchases, the DIRICHLET output also gives the figures for every number of purchases from 0 to 60 for the product category. In this example, a predicted 91 per cent of purchases of Instant Coffee are accounted for by those buying the category five or fewer times, so there can be few purchasers of individual brands more than five times.

[xxii] As set out in Note 10 above, the rate of purchase is therefore not exactly predicted.

[xxiii] If more decimal places are shown there may in fact be seen a slight increase in the results for the smaller brands, in contradiction to the usual Double Jeopardy pattern. There is a theoretical explanation for this apparent anomaly. See Repeat Buying page 355 on the point

[xxiv] There are techniques for analysis of this kind using the observed duplication figures and the 'Duplication Law' for fitting. These are not only simple but in fact more useful than the Dirichlet predictions. The topic is outside the scope of this Guide, but discussed in.[JAB note reference].

John Bound is a Visiting Research Associate, The Ehrenberg Centre for Research in Marketing, London South Bank University.

Email: John.A.Bound@marketingscience.info