APPENDIX C

CALCULATIONS FOR MULTI-BRAND BUYING

C.1 Calculations and Tabulations

In this Appendix we show how to calculate and model the measures of multi-brand buying discussed in Chapters 9 through 13. The same structure as Appendix A is followed, starting in § C.2 to C.4 with the estimation of the basic parameters of the theoretical model, here the Dirichlet; then showing how to use these in calculating the theoretical multi-brand statistics like w_p, b_s and duplication percentages $b_{X.Y.}$ in § C.5 to C.8; and finally outlining the appropriate tabulations of observed data in § C.9 to C. 12. The theoretical treatment of time-periods is somewhat different from Appendix A and a discussion of this is reserved until § C.8.

The theoretical calculations are best done with a computer, using for example the software on the floppy disc mentioned in the Preface. However, by working through a simple example on paper the principles will become clearer. We will use the same hypothetical illustration as in Table A20 of Appendix A which is reproduced as Table C.7 in § C.9.

There are three parts to dealing with the theoretical model before we can calculate values of interest such as w_p. First we have to input values for the distribution of purchases of the product-field as a whole, then estimate the structural parameter S, and thirdly estimate the Dirichlet probabilities. All these steps involve relatively cumbersome numerical iterative calculations and approximations, since the measures we want do not have explicit algebraic formulae in the Dirichlet model. (The difficulties are not due to statistical sampling considerations – it is not for example a question of how to derive maximum likelihood estimates.)

C.2 The Product-Field Distribution

Two methods of dealing with the distribution of purchases of the product-field are available for the Dirichlet model, as noted in Chapter 13. One is the NBD-Dirichlet. This assumes that the distribution of product-field purchases is close to an NBD and estimates this from two summary statistics like B and W in some chosen base-period. Values for any other time-period of length T can then be estimated merely by substituting TA and TM for A and M in what follows (see Section C.8).

If the NBD assumption for the total product-field does not apply, and if the observed distribution is available, this can be used as input for the Empirical-Dirichlet model. This is simpler but less powerful because it does not provide predictions for other time-periods, and it has greater data requirements.

The NBD-Dirichlet. In the NBD-Dirichlet model, B, the percentage buying the product (ie. buying *any* brand) in the time-period, and W, the average purchase frequency of the product per buyer in that period, are used to generate a theoretical NBD. This is fed into subsequent calculations.

We assume that $B = .1$ or 10% and $W = 2.8$ (as in the numerical example of Table C7 later, where for a sample of 200, $B = 20/200 = .1$ and $W = 56/20 = 2.8$, and hence the mean number of purchases per household is $M = 56/200 = .28$ (or $B\,W = .28$)). Following the procedure in Section A.2 we derive the NBD parameter A = 4.559 and hence also the value of the related parameter $K = M/A = .0614$ (here we have obtained A directly rather than by interpolation). This now enables us to follow the guidelines in Section A.6 to calculate the NBD proportion P_n of households who make n purchases of the product-field, where n is any whole number.

$$P_n = \left(\frac{A}{1 + A}\right)\left(1 - \frac{A - M}{An}\right) P_{n-1}, \quad \text{for } n = 1, 2, 3, \ldots$$

$$= \left(\frac{4.559}{5.559}\right)\left(1 - \frac{4.559 - .28}{4.559n}\right) P_{n-1},$$

$$= .8201 \left(1 - \frac{.9386}{n}\right) P_{n-1}.$$

Given that $P_0 = 1 - B = .9$, we have

$$P_1 = .8201\ (1 - .9386)\ .90000 = .04532 \text{ or } 4.5\%,$$
$$P_2 = .8201\ (1 - .4693)\ .04532 = .01972 \text{ or } 2.0\%,$$
$$P_3 = .8201\ (1 - .3129)\ .01972 = .01111 \text{ or } 1.1\%,$$

and so forth (the full distribution is shown later in Table C1). We will use these theoretical proportions buying 1, 2, 3 etc. times in the calculation of the full Dirichlet model.

Dealing with an Infinite Series. However, we need to note that since the NBD is an infinite distribution it is strictly impossible to calculate all these proportions numerically. We therefore have to curtail the calculations at some finite point, n^* say, and develop approximate estimates for the tail of the NBD distribution. For hand calculations, one can generate the individual P values for say .99 of all households, so that we have to deal with a small remainder term of .Ol. When using computer routines one can easily go further for still greater accuracy, eg. to leave a very small tail of only .0001%. *(These tails are numerically more important than they might seem because they consist of very heavy buyers).*

An approximation procedure, suggested by Professor G. J. Goodhardt, is to divide the sum of the tail probabilities over two values n' and $n'+1$, determining purchase proportions $P_{n'}$ and $P_{n'+1}$ such that the weighted sum $n'P_{n'} + (n'+1)P_{n'+1}$ equals the total purchases accounted for in the exact NBD by those buying more than n^* times.

Suppose we have chosen to cut off at a tail of 1%. Calculating numeric values of P_n as above shows that this tail is reached at about $n^* = 6$ and purchases up to this level amount to .18843; ie.

$$\sum_{n=0}^{6} P_n = .99097,$$

and

$$\sum_{n=1}^{6} nP_n = .18843.$$

We represent the tail of the distribution by two proportions at n' and $n'+1$ by solving the two equations

$$\begin{aligned} P_{n'} + P_{n'+1} &= 1 - \sum_{n=0}^{n^*} P_n = P_R \\ &= 1 - .99097 \\ &= .00903, \end{aligned}$$

$$\begin{aligned} n'P_{n'} + (n'+1)P_{n'+1} &= M - \sum_{n=1}^{n^*} nP_n = Q_R \\ &= .28 - .18843 \\ &= .09157, \end{aligned}$$

where M is again the mean number of purchases of the product-field per household. Substantively this means .00903 people make .09157 purchases in the tail.

Multiply P_R by n' and subtract from Q_R to give

$$P_{n'+1} = Q_R - n'P_R$$

Divide by P_R to give

$$\frac{P_{n'+1}}{P_R} = \frac{Q_R}{P_R} - n'$$

If Q_R/P_R is an integer (which is unlikely) put $n' = Q_R/P_R$ and then $P_{n'+1} = 0$, $P_{n'} = P_R$. Otherwise, n' must be the integral part of Q_R/P_R. Then

$$P_{n'+1} = P_R \times (\text{non-integral part of } Q_R/P_R).$$

Thus, in our example where $P_R = .00903$ and $Q_R = .09157$:

$$\begin{aligned} Q_R/P_R &= 10.14064, \text{ so } n' = 10, \\ P_{n'+1} &= P_{11} = .00903 \times .14064 = .00127, \\ P_{n'} &= P_{10} = .00903 - .00127 = .00776. \end{aligned}$$

The cumulative number of purchases at $P_{n'+1}$ should equal M and it is worthwhile doing this calculation as an arithmetic check

$$\sum_{n=1}^{n'+1} nP_n = .28000 = M.$$

The complete results from this example are shown in Table Cl.

C.3 The Dirichlet Parameter $\hat{S}$

The Dirichlet parameter $\hat{S}$ can be estimated using the ***average*** brand. But in practice we usually form a separate estimate $\hat{S}_j$ for each brand j, and then obtain the overall Dirichlet estimate as a ***weighted average*** across

Table C1. The Estimated Proportions Buying the Product-Field n Times

n	Proportion Buying P_n	Cumulative Proportion ΣP_n	Purchases nP_n	Cumulative Purchases ΣnP_n
0	.90000	.90000	.00000	.00000
1	.04532	.94532	.04532	.04532
2	.01972	.96504	.03944	.08476
3	.01111	.97615	.03333	.11809
4	.00697	.98312	.02788	.14597
5	.00464	.98776	.02320	.16917
6	.00321	.99097	.01926	.18843
		....		
10	.00776	.99873	.07760	.26603
11	.00127	1.00000	.01397	.28000

brands $j = 1,\ldots, g$ (or for a sub-set of brands g^*). This provides a diagnostic check on cases where the model fails to fit a particular brand, and these might be left out of the weighted average.

Each brand is analysed in turn. Here we describe the calculations for brand X in the example of Table A20 (or Table C7 later), the basic statistics for which are shown in Table C2. For simplicity, the subscript X will be dropped in this section.

Table C2. Input Statistics

	W	B as %	B	P_0	M
Product Field:	2.800	10.0	.100	.900	.280
	w	b as %	b	p_0	m
Brand X:	1.923	6.5	.065	.935	.125

The estimation of $\hat{S}$ for brand X has to be by iteration, as once again there is no explicit algebraic formula. We begin with a more or less arbitrary starting value, S', usually guessed at from previous experience of relevant $\hat{S}$ values, say $S' = 2$, or we start with $S' = 1$. The aim is to find an estimate such that the predicted number of non-buyers of the brand, p_0', is equal (or close to) the observed number p_0.

We estimate non-buyers of the brand separately for each frequency n of buying the product-field, by multiplying the probability P_n in Table Cl by a provisional estimate of the probability $p'_{(0|n)}$ of *not buying* brand X, as given by the model, conditional on having made n product purchases. These are **summated** up to n^*, plus the terms for n' and $n'+1$, ie.

$$p'_0 = \Sigma\ \{P_n p'_{(0|n)}\}$$

where p' stands for provisional estimates, using the assumed Dirichlet value S'.

The procedure is to get a good estimate of p'_0 using the product-field proportions which we have found already. First p'_0 is obtained using P_0 and P_1 only, then using P_2 also, and so on through the whole distribution. In order to estimate the values of $p'_{(0|n)}$ for each iteration it is convenient to work with two terms, c' and d' (neither of these is meaningful but they help us to do our calculations).

These terms need to be adjusted for each P_n, using the observed m and M from Table C2 and the assumed working value of $S' = 2$. First, for P_0 and P_1

$$\begin{aligned} c' &= S' - ((m \times S')/M) \\ &= 2 - ((.125 \times 2)/.28) \\ &= 1.10714, \end{aligned}$$

and

$$\begin{aligned} d' &= c'/S' \\ &= 1.10714/2 \\ &= .55357. \end{aligned}$$

We now use the P_0 and P_1 to derive a first estimate of p'_0 for $S' = 2$

$$\begin{aligned} P'_0 &= P_0 + (P_1 \times d') \\ &= .9 + (.04532 \times .55357) \\ &= .92509. \end{aligned}$$

Next we use P_2 as well, still keeping $S' = 2$. The value of d' has to be revised and the sum for p'_0 is adjusted. Thus

$$\begin{aligned}
\text{New } d' &= \text{Old } d' \times (c' + (n-1))/(S' + (n-1)) \\
&= .55357\ (1.10714 + (2-1))/(2 + (2-1)) \\
&= .38882, \\
\text{New } p'_0 &= \text{Old } p'_0 + (P_2 \times d') \\
&= .92509 + (.01972 \times .38882) \\
&= .93276.
\end{aligned}$$

And for $n = 3$ we have

$$\begin{aligned}
\text{New } d' &= \text{Old } d' \times (c' + (n-1))/(S' + (n-1)) \\
&= .38882 \times (1.10714 + (3-1))/(2 + (3-1)) \\
&= .30203,
\end{aligned}$$

$$\begin{aligned}
\text{New } p'_0 &= \text{Old } p'_0 + (P_3 \times d') \\
&= .93276 + (.O1111 \times .30203) \\
&= .93612.
\end{aligned}$$

And so on for all P_n up to n' and $n' + 1$. The final estimated value of p'_0 obtained in this cycle of calculations is .94093. This differs by .00593 from the observed value of $p_0 = .935$ for brand X in Table C2. To get closer a new iteration is made.

If the estimated p'_0 is greater than the observed p_0 (as here), we use a larger starting value S'', say twice the old value of S', ie. $S'' = 2 \times 2 = 4$. If p'_0 is less than p_0, *we* use a smaller S'', say half the previous one.

The first estimate of p''_0 for P_0 and P_1 but now with $S'' = 4$ will in fact be the same as that for $S' = 2$, since the new d'' here is the same as before. The calculations are

$$\begin{aligned}
c'' &= S'' - ((m \times S'')/M) \\
&= 4 - ((.125 \times 4)/.28) \\
&= 2.21429,
\end{aligned}$$

and

$$\begin{aligned}
d'' &= c''/S'' \\
&= 2.21429/4 \\
&= .55357.
\end{aligned}$$

Using P_0 and P_1 to get the first estimate in the second iteration

$$\begin{aligned} P_0'' &= P_0 + (P_1 \times d'') \\ &= .9 + (.04532 \times .55357) \\ &= .92509, \end{aligned}$$

the same as p_0' for P_0 and P_1.

But the subsequent estimates of p_0'' (using P_2, P_3 etc) will be closer to the observed p_0. Thus, as before, if we now also use P_2 to upgrade our first value of p_0'' we obtain

$$\begin{aligned} \text{New } d'' &= \text{Old } d'' \times (c'' + (n-1))/(S'' + (n-1)) \\ &= .55357 \times (2.21429 + (2-1))/(4 + (2-1)) \\ &= .35587, \end{aligned}$$

$$\begin{aligned} \text{New } p_0'' &= \text{Old } p_0'' + (P_2 \times d'') \\ &= .92509 + (.01972 \times .35587) \\ &= .93211. \end{aligned}$$

All subsequent values of p_0'', taking account of P_3, P_4, . . ., $P_{n'}$, $P_{n'+1}$, are calculated in the same way, revising d'' and p_0'', each time by a new d'' and a new p_0''. At the end of this iteration $p_0'' = .93812$, which is slightly. closer to .935.

For the third and subsequent iterations, the new S value can be obtained from the two preceding ones (here $S' = 2$ and $S'' = 4$) by interpolation or extrapolation, depending on whether the latest estimated p_0 value is too high or too low. Thus, as we converge on p_0, the estimate of S' is improved:

S'	p_0'	$p_0' - p_0$
2	.94093	.00593
4	.93812	.00312
8	.93628	.00128
.	.	.
.	.	.
.	.	.

Further iterations are best done on a computer, where some iteration criterion like .0001, or a limit on the n-umber of cycles, can be used as a cut-off point. In the present example, convergence to within .0005 is reached after 5 cycles, with an estimated S' for brand X of 12.592.

Estimates for brands Y and Z are found in exactly the same way. The final table of results, using the brand suffix j again is shown in Table C3.

Table C3. Estimates of the Dirichlet Parameter for each Brand

Brand	$\hat{S}_j$	m_j	m_j/M
X	12.592	.125	.4464
Y	27.291	.090	.3214
Z	36.241	.065	.2321

The three values of $\hat{S}_j$ here differ markedly – from 12 to 36 – whereas with a well-fitting model they should all be similar. The variation is because our small hypothetical buying pattern is unrealistic and does not follow a Dirichlet too well, but this does not detract from its illustrative value.

The overall Dirichlet parameter $\hat{S}$ is a weighted average of these $\hat{S}_j$, obtained by taking the market shares of m_j into account:

$$\hat{S} = \sum_{g^*} (\hat{S}_j m_j/M) / \sum_{g^*} (m_j/M).$$

In forming such an estimate one can drop the estimates for one or more brands if they look irregular, like the low 12.592 for brand X here perhaps, and estimate $\hat{S}$ from the $g^* = 2$ brands Y and Z and then check the fit of the model for all g brands, ie. including X here. More generally, we can fit the Dirichlet to any selection of g* brands.

If we use the full set of g brands that make up the product-field (ie. $g^* = g$), the denominator in the above equation is 1 and $\hat{S}$ is merely the sum of the g $\hat{S}_j$s weighted by the market share of each brand. Numerically, the weighted estimate of $\hat{S}$ for our three brands is

$$\hat{S} = \frac{(12.592 \times 4464 + 27.291 \times .3214 + 36.241 \times .2321)}{(4464 + .3214 + .2321)}$$

$$= 22.8062/1 = 22.8.$$

The values of the Dirichlet parameters $\hat{\alpha}_j$ (see Chapter 13, § 13.3) are given by $\hat{\alpha}_j = \hat{S}(m_j/M)$ and here are

$$\hat{\alpha}_X = 22.8062 \times 4464 = 10.1807$$
$$\hat{\alpha}_Y = 22.8062 \times .3214 = 7.3299$$
$$\hat{\alpha}_Z = 22.8062 \times .2321 = 5.2933$$

which add up to $\hat{S} = 22.8$, as above.
We can also define the quantity $\hat{\beta}_j = \hat{S} - \hat{\alpha}_j$, which will be required in the next section.

Possible Short-Cuts. With suitable software the above calculations are less burdensome than they might seem. We have therefore not explored possible short-cuts very much so far. One set of possibilities is to curtail the calculations of the NBD in Section C.2 earlier and to iterate less often in estimating $\hat{S}$. Yardsticks for evaluating such short-cuts are how much the answers differ from the more accurate estimates, whether in terms of the $\hat{S}$ values or the $\hat{\alpha}_j$, or in terms of the estimates of behavioural statistics like b, w, w_p, b_s and the duplication patterns that are discussed in Sections C.5 to C.8.

Another possible short-cut where accuracy and adequacy needs to be explored is to estimate a single $\hat{S}$ for the average brand, using the above procedure. It should be particularly useful in cases where earlier analyses of the product-field have already shown the model to fit well; in these cases it is largely unnecessary to examine every brand.

C.4 The Dirichlet Proportions for Brand X

Having somewhat laboriously estimated the main Dirichlet parameter $\hat{S}$ we still cannot directly calculate any of the derived statistics like b and w for a given brand since there are no explicit algebraic formulae relating them to $\hat{S}$ (or to the other inputs to the model, ie. B, W and market shares). Instead, we now have to calculate a matrix of proportions for all those making n purchases of the product and r purchases of a brand, as set out schematically in Table C4. From these numerical values, the various statistics for brand X can then be tabulated at last, as will be discussed in C.5 to C.8.

In Table C4 the proportions buying in the second line are for the whole product-field, as presented in Section C.2 (the numerical values in our example being $P_0 = .90000$, $P_1 = .04532$, etc). The body of the table

Table C4. Matrix of Dirichlet Proportions for a Single Brand

		Purchases of the Product								
Numbers:		0	1	2	3		n^*	n'	$n'+1$	Total
Proportions:		P_0	P_1	P_2	P3	…	P_{n^*}	$P_{n'}$	$P_{n'+1}$	1.0
	0	p_{00}	p_{01}	p_{02}	PO3	…	p_{0n^*}	$p_{0n'}$	$p_{0(n'+1)}$	p_0
	1		p_{11}	p_{12}	P13	…	p_{1n^*}	$p_{1n'}$	$p_{1(n'+1)}$	p_1
	2			p_{22}	P23	…	p_{2n^*}	$p_{2n'}$	$p_{2(n'+1)}$	p_2
Purchases	.					…		.	.	
of the						…		.	.	.
Single Brand	.					…			.	.
	$r'-1$						$p_{(r'-1)n^*}$	$p_{(r'-1)n'}$	$p_{(r'-1)(n'+1)}$	$p_{r'-1}$
	r'							$p_{r'n'}$	$p_{r'(n'+1)}$	$p_{r'}$
	$r'+1$								$p_{(r'+1)(n'+1)}$	$p_{r'+1}$

Notes: $n = 0,1,2,\ldots, n'-1, n', n'+1$ for the Product Field
$r = 0,1,2\ldots, r'-1, r', r'+1$ for the single Brand
All entries below the leading diagonal are identically equal to zero.

represents the proportion of the population making n purchases of the product and r purchases of brand X.

The row labelled "0" for purchases of X are the proportions not buying the brand for each group making $n = 0,1,2,$ etc purchases of the product. We compute these values by the recurrence formula

$$p_{0n} = p_{0(n-1)} \times \frac{(\hat{\beta} + n - 1)}{(\hat{\alpha} + \hat{\beta} + n - 1)} \times \frac{P_n}{P_{(n-1)}} \quad \text{for } n = 1,\ldots,n' + 1.$$

The starting value of p_{00} is equal to P_0 (ie. .90000). Putting into this expression our numerical values of $\hat{\alpha}, \hat{\beta}, P_n$ and p_{00} from the last section we have:

$$p_{01} = .9 \times \frac{(12.6255 + 1 - 1)}{(22.8062 + 1 - 1)} \times \frac{.04532}{.9}$$

$$= .9 \times .55360 \times .05036$$

$$= .02509,$$

$$p_{02} = .02509 \times \frac{(12.6255 + 2 - 1)}{(22.8062 + 2 - 1)} \times \frac{.01972}{.04532}$$

$$= .02509 \times .57235 \times .43513$$

$$= .00625.$$

And so on to build up the whole row for non-buyers of the brand. For the next row, and the rest of the matrix, entries to the left of the leading diagonal are identically equal to zero. The remaining entries are found from the recurrence formula

$$p_{rn} = \frac{(n - r + 1)}{r} \times \frac{(\hat{\alpha} + r - 1)}{(\hat{\beta} + n - r)} \times p_{(r-1)n}, \quad \text{for } r = 1,\ldots, r' + 1,$$

where the starting values of p_{1n} are obtained from the proportions of p_{0n} in the first row as:

$$p_{11} = \frac{1 - 1 + 1}{1} \times \frac{(10.1807 + 1 - 1)}{(12.6255 + 1 - 1)} \times .02509$$

$$= 1.0 \times .80636 \times .02509$$

$$= .02023,$$

$$p_{12} = \frac{2-1+1}{1} \times \frac{(10.1807 + 1 - 1)}{(12.6255 + 2 - 1)} \times .00625$$

$$= 2.0 \times .74718 \times .00625$$

$$= .00934.$$

And so on for all n. For the next row, $r = 2$, the leading diagonal is:

$$p_{22} = \frac{2-2+1}{2} \times \frac{(10.1807 + 2 - 1)}{(12.6255 + 2 - 2)} \times .00934$$

$$= .5 \times .88556 \times .00934$$

$$= .00414.$$

Again, values for all n are obtained from the recurrence formula. The complete matrix is shown in Table C5.

Some arithmetic checks should be made at this stage

$$\sum_{n=0}^{n'+1} p_n = 1.0,$$

$$\sum_{r=0}^{r'+1} p_r = 1.0,$$

$$\sum_{r=1}^{r'+1} rp_r = .125 = BW(\hat{\alpha}/\hat{S}),$$

$$BW(\hat{\alpha}/\hat{S}) = .1 \times 2.8 \times (10.1807/22.8062)$$
$$= .125$$

and for each n,

$$\sum_{r=0}^{r'+1} p_{rn} = P_n.$$

Table C5. Dirichlet Proportions for Brand X

(**In** Base Period)

		Purchases of the Product									
Numbers		0	1	2	3	4	5	6	10	11	Total
Proportions		.90000	.04532	.01972	.01111	.00697	.00464	.0032 1	.00775	.00127	0.99999
	0	.90000	.02509	.00625	.00208	.00079	.00032	.00014	.00007	.00001	.93475
	1		.02023	.00934	.00433	.00205	.00100	.00050	.00031	.00004	.03780
	2			.00413	.00356	.00236	.00143	.00083	.00077	.00010	.01318
	3				.001 14	.00140	.00119	.00087	.00127	.00017	.00604
Purchases	4					.00037	.00057	.00058	.00157	.00023	.00332
of	5						.00013	.00024	.00152	.00024	.00213
Brand X	6							.00005	.00116	.00021	.00142
	7								.00068	.00015	.00083
	8								.00030	.00008	.00038
	9								.00009	.00003	.00012
	10								.00001	.00001	.00002
	11									.00000	.00000

C.5 Single-Brand Measures of Buyer Behaviour

The standard measures of buyer behaviour can now be calculated from the estimates of $\hat{S}, \hat{\alpha}_j$ and the proportions in Table C5. They are of two types: (i) single-brand measures, which are illustrated in this section, and (ii) measures of multi-brand purchasing, which relate one brand to the whole product-field (Section C.6) or to individual competing brands (Section C.7).

The theoretical frequency distribution of purchases, for a single brand in the analysis period, is derived from the final column in Table C4, ie. p_0, $p_1, \ldots, p_{r'+1}$. In our example (Table C5) an estimated 93.5% of households do not buy brand X, 3.8% buy it once, 1.3% buy it twice, and so forth. From this we calculate that of those buying brand X in the base period, 58% are estimated to have bought the brand once, and that such once-only buyers contributed to 30% of sales.

The theoretical value of the brand's penetration is:

$$\hat{b} = 1 - p_0$$

$$= 1 - .93475$$

$$= .06525 \text{ or } 6.5\%.$$

The estimate of the average number of purchases per buyer, $\hat{w}$, comes either from observed values of B, W and the market share $(\hat{\alpha}/\hat{S})$, or alternatively from the numerical sum of rp_r in the Table C5:

$$\hat{w} = BW(\hat{\alpha}/\hat{S})/\hat{b} = \sum_{r=1}^{r'+1} rp_r \;/\hat{b}$$

$$= .125/.06525$$

$$= 1.916 \text{ in either case.}$$

These values are close to the "observed" $b = 6.5\%$ and w = 1.923 in Table C2.

Other buying statistics for a single brand can also be estimated, such as the incidence of period-by-period repeat-buying, including estimates conditional on being light, medium or heavy buyers in the first period (see Chapter 3). In practice we tend however to use NBD estimates here as they are simpler to obtain and more robust to any non-stationarities in other parts of the market.

C.4 Total Product Usage and Sole Buying

The Dirichlet model also enables us to estimate how brand Xis bought in relation to other purchases of the product. One measure is w_p, the total product usage made by buyers of brand X:

$$\hat{w}_p = \{1(P_1 - p_{01}) + 2(P_2 - p_{02}) + \ldots + (n'+1)(P_{n'+1} - p_{0(n'+1)})\}/\hat{b}$$

$$= \{1(.04532-.02509) + 2(.01972-.00625) + 3(.01111-.00208) + \ldots\}/.06525$$

$$= \{.02023 + .02694 + .02709 + \ldots\} \text{ I}.06525$$

$$= .2309/.06525$$

$$= 3.521, \text{ or } 3.5 \text{ approximately.}$$

Another measure is the proportion of panel-members who only buy brand X and no other brand in the analysis period. The estimate of how many of these sole buyers there are is:

$$\hat{b}_s = p_{11} + p_{22} + \ldots + p_{(r'+1)(n'+1)}$$

$$= .02023 + .00413 + .00114 + \ldots$$

$$= .02606, \text{ or } 2.6\%.$$

Expressed as a percentage of all buyers of the brand:

$$\hat{b}_s / \hat{b} = .02606/.06525$$

$$= .3994 \text{ or about } 40\%.$$

The average number of purchases made by sole buyers in the period is

$$\hat{w}_s = \{1 \times p_{11} + 2 \times p_{22} + \ldots + (n'+1) \times p_{(r'+1)(n'+1)}\}/\hat{b}_s$$

$$= \{.02023 + .00826 + .00342 + \ldots\}/.02606$$

$$= 1.3216, \text{ or about } 1.3.$$

Theoretical results of all the measures so far for all three brands in the example are shown in Table C6.

Table C6. Theoretical Measures of Buyer Behaviour

Brand	Market Share $(m_j/\Sigma m)\%$	Penetration $\hat{b}_j\%$	Average Purchase Frequency $\hat{w}_j$	Total Usage $\hat{w}_{pj}$	Incidence of Sole Buyers $\hat{b}_{sj}\%$	Average Rate of Sole Buying $\hat{w}_{sj}$
X	45	6.5	1.9	3.5	40	1.3
Y	32	5.3	1.7	3.8	33	1.2
Z	23	4.3	1.5	4.1	28	1.2

C.7 **Duplication of Purchase**

The proportion of individuals who buy a pair of brands, X and Y, is estimated from the Dirichlet by calculating their separate and combined penetrations in the base-period, as noted in Chapter 13. This is done by forming a composite brand $(X+Y)$ - adding together their separate shares to give $\hat{\alpha}_X + \hat{\alpha}_Y$ = 10.1807 + 7.3299 = 17.5106. This composite parameter is used to revise the matrix of Dirichlet proportions, following the procedure set out in Section C.4. From the new matrix a composite penetration figure, $\hat{b}_{(X+Y)}$, is obtained (.0880 in the present case). The proportion, $\hat{b}_{XY}$, buying both brands X and Y is then given by

$$\begin{aligned} \hat{b}_{XY} &= \hat{b}_X + \hat{b}_Y - \hat{b}_{(X+Y)} \\ &= .0653 + .0531 - .0880 \\ &= .0304. \end{aligned}$$

Conditional proportions can also be calculated:

$$\begin{aligned} \hat{b} &= \hat{b}_{XY}/\hat{b}_Y = .0304/.0531 = ,573 \text{ or } 57\%, \\ \hat{b}_{Y|X} &= \hat{b}_{YX}/\hat{b}_X = .0304/.0653 = .466 \text{ or } 47\%. \end{aligned}$$

In Chapter 10.5 it was seen how the average levels of duplication are systematically related to a brand's penetration. For this purpose an empirical coefficient **D** was derived. Here we find the theoretical Duplication Coefficient for the two brands ***X*** and Y as:

$$\hat{D}_{XY} = \hat{b}_{XY}/(\hat{b}_X \times \hat{b}_Y)$$

$$= .0304/(.0653 \times .0531)$$

$$= 8.767.$$

This is abnormally high compared with our experience of real-life cases, and is due once more to the artificiality of our small example.

So far we have estimates for the proportion of duplicate buyers and the Duplication Coefficient for a pair of brands. The calculations have in principle to be repeated for all other pairs of brands to obtain the full picture. The remaining answers in our three-brand example are:

$$\hat{b}_{XZ} = \hat{b}_X + \hat{b}_Z - \hat{b}_{(X+Z)}$$

$$= .0653 + .0425 - .0826$$

$$= .0252,$$

$$\hat{b}_{YZ} = \hat{b}_Y + \hat{b}_Z - \hat{b}_{(Y+Z)}$$

$$= .0531 + .0425 - .0739$$

$$= .0217.$$

(Note that the table is symmetric, i.e. $\hat{b}_{XY} = \hat{b}_{YX}$, $\hat{b}_{ZX} = \hat{b}_{XZ}$, and $\hat{b}_{ZY} = \hat{b}_{YZ}$). Complete tables of (i) the proportions buying, and (ii) the conditional percentages are respectively:

Absolute Proportions	Brands *X*	Y	*Z*
X	–	.0304	.0252
Y	.0304	–	.0217
Z	.0252	.0217	–

and

Conditional Proportions	% also buying		
	X	y	Z
Buyers of			
X		47	39
Y	57	–	41
Z	59	51	–
Average	58	49	40

The theoretical entries down each column increase slightly with decreasing penetration (from X to Z). In principle this contradicts the Duplication of Purchase Law $b_{X|Y} = Db_X$ of Chapter 10 which expects them to be constant (with the same D for all pairs of brands). But it is in line with the patterns already observed empirically in 1972 and before (p. 178). The Dirichlet is therefore more realistic than the Duplication of Purchase Law, but the difference is generally small.

In parallel with this, the pairwise duplication coefficients predicted by the Dirichlet are also not constant, although all of them are close to the average of about 9.2:

Duplication Coefficients	X	Y	Z		
X		8.8	9.1		
Y	8.8	–	9.6		
Z	9.1	9.6	–		
Average	9.0	9.2	9.4	=	9.2

Short-Cuts. Calculating all the pairwise duplications for a large number of brands requires extensive computing; for g brands there are $((g \times g) - g)/2$ pairs, which means 45 pairs for 10 brands, or 190 pairs for 20 brands. In practice, calculating the average D from the observed data as in Chapter 10 remains a useful simplification where these data exist.

Two theoretical short-cuts, which approximate the Dirichlet itself, could be:

(i) To calculate the g Ds for the *average* brand with each of the g specific brands, along the lines already set out.

(ii) To calculate the Ds for extreme pairs of brands – if L_1 *and* L_2 and S_1 and S_2 are the two largest and two smallest pairs, then calculate Ds for L_1L_2, L_1S_2, L_2S_1, and S_1S_2. This will cover the range of possible values. To simplify one can approximate to the required single value by a (possibly weighted) average.

In practice, it is easier to use a computer routine.

Other Multi-Brand Statistics. To illustrate briefly the wider range of statistics that can be estimated from the Dirichlet and which show how buyers combine their purchases, we note first the average number of brands bought per buyer of the product

$$\sum_{j=1}^{g} \hat{b}_j/B \quad = \quad .0653 + .0531 + .0425 \ / \ .100$$

$$= 1.609.$$

Second, using the sole-buying figures in Table C5, the proportion of the total sample buying only one brand is

$$\sum_{j=1}^{g} \hat{b}_{sj} \quad = \quad .0261 + .0174 + .0120$$

$$= \quad .0555, \text{ or } 5.6\%.$$

Alternatively, we divide this sum by B to get the proportion of product buyers who buy only one brand, ie. $.0555/.100 = .555$, or about 55%.

C.8 Different Length Time Periods

Theoretical measures of buyer behaviour have so far been presented for some chosen base period. Now we shall illustrate the calculations for other time spans, of length T relative to the "unit" length of the chosen base period (where T can be greater or less than 1).

The proportions along the first row of Table C4 now depend on TA, TM and K (where A, A4 and K come from the existing product-field NBD in the base period). We predict B and W for length T using standard NBD theory

(Appendix A.4 and A.5). In our numerical example the unit length is 12 weeks, so for $T = 2$ (ie. 24 weeks) TB and TW are

$$TB = 1 - (1 - TA)^{-K}$$

$$= 1 - (1 - (2 \times 4.559))^{-.0614}$$

$$= 1 - .8675$$

$$= .1325,$$

$$TW = TM/TB$$

$$= .56 / .1325$$

$$= 4.2264.$$

This gives an estimate for P_0 of .8675 which is used in a complete re-run of the NBD model with the new parameters:

$$P_0 = .8675,$$

$$2A = 2 \times 4.559 = 9.118,$$

$$2M = 2 \times .28 = .56,$$

$$k = 2M/2A = .0614 \text{ (constant)}.$$

These values are fitted into the earlier recursive formula

$$P_n = \left(\frac{2A}{1 + 2A}\right)\left(1 - \frac{2A - 2M}{2An}\right) P_{n-1},$$

so that

$$P_1 = .901166\,(1 - .93858)\,.86750 = .04802$$

$$P_2 = .901166\,(1 - .46929)\,.04802 = .02296$$

$$P_3 = .901166\,(1 - .31286)\,.02296 = .01422$$

$$\vdots \qquad \vdots$$

With this new distribution of product purchases the cycle starts again from Section C.5. The distribution P_n, $n = 0, 1, ..., n' + 1$, replaces line one in Table C4 and, along with the original estimates of $\hat{S}$ and $\hat{\alpha}_j$ (ie. 22.8062, and 10.1807, 7.3299 and 5.2933 respectively), the body of the matrix is revised. (The values of the brand choice parameters $\hat{S}$ and $\hat{\alpha}_j$ in the Dirichlet are unaffected by T).

The standard measures of buyer behaviour for brand X in the double period are

Measure	Single Period	Double Period
b_X	.0653	.0917
w_X	1.92	2.48
w_{PX}	3.54	4.91

For two periods of equal unit length several aspects of repeat-buying are now predictable, such as the incidence of repeat-buyers and new buyers, and their rates of buying. These measures are based on the penetration in a single period in relation to that in a double period. Thus, the proportion buying X in both periods is:

$$\hat{b}_{X12} = 2 \times \hat{b}_{X1} - \hat{b}_{X2}$$

$$= 2 \times .0653 - .0917$$

$$= .0389.$$

The proportion of new buyers over the same time span is:

$$\text{new } \hat{b}_{X2} = \hat{b}_{X1} - \hat{b}_{X12}$$

$$= .0653 - .0389$$

$$= .0264.$$

Finally, the incidence of repeat-buying is defined as:

$$\hat{b}_{X12}/\hat{b}_X = .0389/.0653 \times 100$$

$$= 59.6\%$$

which compares with 56.1% calculated from conventional NBD theory (using data from the first 12 weeks alone – Appendix A.8). Dirichlet and NBD period-to-period levels of repeat-buying are generally similar, and in this case the difference is just 3.5%.

At this stage the difference between NBD and Empirical versions of the Dirichlet model becomes important. The Empirical version cannot be used directly to make predictions over different time spans. Instead, a new product field distribution and new values of $\hat{S}$ and $\hat{\alpha}_j$ for T must be tabulated from raw data, if these are available. This means the base-period is re-defined at T and the cycle starts from Section C.4 again. Under stationary conditions K, $\hat{S}$ and $\hat{\alpha}_j$ ought to be constant – it is good to check this. Data requirements for the Empirical version are heavy and the model itself is inelegant; it is however useful to fit both versions and then compare the resultant statistics.

C.9 Multi-Brand Tabulations

We now consider the hand tabulation of observed data to obtain the multi-brand buying statistics discussed in Chapters 9 and 10 and in theoretical terms earlier in this Appendix. We use again the hypothetical 12-week example in Table A20 which is reproduced here in Table C7 with various summary statistics.

In this section we discuss tabulating the penetration B and average purchase frequency W for the product field, and in Section C. 10, w_p, the total number of purchases of the product-field per buyer of a given brand. In Section C.11 and C. 12 we consider tabulations of sole buyers and of duplicate buyers.

The first two summary columns in Table C7 show whether a panel-member bought the product (ie. any brand) at least once in the 12-week period and how often they bought it – eg. 11 times for panel-member 1. There are 20 panel-members (out of 200) who each made at least one purchase of some brand in the 12 weeks so that $B = 20/200 = .1$ or 10%. They bought 56 times, so that $W = 56/20 = 2.8$ and $M = 56/200 = .28$, as noted earlier.

The frequency distribution of product-field purchases, ie. one 11, one 6, one 5, three 4s, three 3s, two 2s and nine 1s (adding up to 56 as a check), can be derived from the Total column. This can be used as direct input for the Empirical-Dirichlet model, or to check the assumptions of an NBD.

As a numerical check it is worthwhile to count the number of purchases each week (at the foot of each column) and to check that the total (56)

Table C7. The Buying Pattern of Table A20 with Summary Statistics

	Purchases in Week:												The Product		Brands					
															At least once			Totals		
Panel Member	1	2	3	4	5	6	7	8	9	10	11	12	At least once	Total	*X*	*Y*	*Z*	*X*	*Y*	*Z*
1	*X*	*X*	*Y*	*X*	*X*	*X*	*Y*	*X*	.	*Y*	*X*	*Y*	1	11	1	1	.	7	4	.
2	.	.	*X*	.	*Z*	*X*	.	.	.	*Z*	*X*	*X*	1	6	1	.	1	4	.	2
3	.	.	*Z*	.	.	*Z*	.	*X*	*X*	.	*Z*	.	1	5	1	.	1	2	.	3
4	.	.	.	.	*Y*	.	*Y*	*Y*	.	*X*	.	.	1	4	1	1	.	1	3	.
5	*Y*	*X*	.	.	*X*	.	.	.	.	.	.	*Y*	1	4	1	1	.	2	2	.
6	*X*	.	.	.	.	.	.	*Z*	*Z*.	*Y*	.	.	1	4	1	1	1	1	1	2
7	.	*Z*	*X*	.	.	.	.	.	.	.	.	*X*	1	3	1	.	1	2	.	1
8	.	*Y*	.	*Y*	.	.	.	.	*X*	.	.	.	1	3	1	1	.	1	2	.
9	*Z*	.	.	.	.	.	*X*	.	.	.	*Y*	.	1	3	1	1	1	1	1	1
10	.	.	.	*X*	.	*Y*	.	.	.	.	.	.	1	2	1	1	.	1	1	.
11	.	.	.	*Y*	.	.	*X*	.	.	.	.	.	1	2	1	1	.	1	1	.
12	.	.	.	.	.	.	.	.	.	*X*	.	.	1	1	1	.	.	1	.	.
13	*Y*	.	.	.	.	.	.	.	.	.	.	.	1	1	.	1	.	.	1	.
14	.	.	.	.	.	.	.	*Y*	.	.	.	.	1	1	.	1	.	.	1	.
15	.	.	*Z*	.	.	.	.	.	.	.	.	.	1	1	.	.	1	.	.	1
16	.	.	.	.	.	.	.	.	*Y*	.	.	.	1	1	.	1	.	.	1	.
17	.	.	.	.	*Z*	.	.	.	.	.	.	.	1	1	.	.	1	.	.	1
18	.	.	.	.	.	.	.	.	.	.	*Z*	.	1	1	.	.	1	.	.	1
19	.	.	.	.	.	.	.	.	.	*Z*	.	.	1	1	.	.	1	.	.	1
20	.	.	.	.	.	*X*	.	.	.	.	.	.	1	1	1	.	.	1	.	.
21–200	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
The Product																				
Buying	5	4	5	4	5	5	4	5	4	6	5	4	20	.	13	11	9	-	-	-
Total Product	5	4	5	4	5	5	4	5	4	6	5	4	.	56	49	36	25	-	-	-

agrees with the sum of the row totals. No such cross-check is possible for the number of buyers each week.

C.10 Tabulating Product Rates of Buying

To get w_p for brand X, we have to count the total number of purchases of the product by the 13 panel-members who bought X at least once in the 12 weeks. This is tricky by hand, probably needing two fingers as we move down the column, one for the Total Product and another for buying X at least once. It is easy to make a mistake, and there is no check other than by repeating the count (preferably starting from the bottom the second or third time, to avoid possibly making the same mistakes as before). The final w_ps for the three brands are:

$$\text{Brand A} = 49/13 = 3.8, \ \text{B} = 35/11 = 3.2, \ \text{C} = 25/9 = 2.8.$$

C.11 Tabulating Sole Buyers

In Table C7 we see by visual inspection that panel-members 12 and 20 are 100%-loyal or "sole" buyers of X in the 12-week period, and similarly that there are sole buyers of Y (13, 14 and 16), and of Z. In the shorter period, weeks 1-4, panel-members 2 and 10 are sole buyers of X (but they buy other brands in subsequent weeks).

Tabulating sole buyers of a given brand, X say, by hand requires seeing whether a panel-member

(i) bought brand X at all, and
(ii) bought no other brand

in the chosen analysis period, and doing a gate-count accordingly. Numerically, some $b_s = 2/200 = .01$ (or 1 %) of the sample only bought X in the 12 weeks, or $b_s/b = 2/13 = .15$ (or 15%) of X buyers were sole buyers.

If the number of purchases made by sole buyers is to be tabulated as well, a separate gate count of all numbers of purchase occasions should be kept, which for brand X happens to give 6 and hence $w_s = 6/6 = 1.0$ (w_s is usually low, but not as low as this). The results for all three brands are:

	b_s/b	w_s
X	15	1.0
Y	27	1.0
Z	44	1.0

The marked variation of b_s/b is an atypical feature of our artificial example.

Sole-buyers' rate of buying the product in an analysis period necessarily equals their rate of buying the brand. This provides a very handy computational short-cut for identifying sole buyers: they are simply any panel-member for whom $w = w_p$.

C. 12 Tabulating Duplication Tables

Tabulating duplication tables by hand is a substantial task if the number of brands is at all large. For three brands, as in our example, separate counts are needed for just three combinations of duplicated buying, X and Y, Y and Z, Z and X. For 10 brands, $(10 \times 9)/2 = 45$ separate counts would be needed.

To obtain the number of duplicated buyers of brands X and Y from Table C7, we look down the columns for buying each brand at least once towards the right of the table and count that 8 panel-members were buying both brands, using gate counts:

X & Y	Y & Z	X & Z
8	2	5

Note how tallying for brands X and Z is more of a mental strain because they are "separated" in Table C7; this would be worse for more than three brands.

The tallies give a duplication table for b_{XY} of the symmetrical form

	Brands		
Absolute values	X	Y	Z
X	13	8	5
Y	8	11	2
Z	5	2	9

There is no indirect check on counting the duplications, so that all one can do is repeat the exercise.

To get figures like $b_{Y.X}$ ie. the percentage of buyers of brand X who also buy Y, we divide the $b_{XY} = 8$ by $b_X = 13$, to give $b_{Y.X} = 8/13 = .6154$, or 62%. The full duplication table is

Conditional proportions	Percentage also buying X	Y	Z
Buyers of			
X	(100)	62	38
Y	73	(loo)	18
Z	56	22	(100)
Average	65	42	28

With real life examples the figures in each column (other than the 100%) are usually more uniform.

Average Purchases for Duplicate Buyers. To tabulate the average purchase rates of duplicate buyers, like $w_{X.Y}$ and $w_{Y.X}$, we run through the last columns of Table C7, doing separate counts of their purchases:

Buyers of X & Y Who Purchase: X	Y
7	4
1	3
2	2
1	1
1	2
1	1
1	1
1	1
15	15

This gives $w_{X.Y} = 15/8 = 1.9$, and $w_{Y.X} = 15/8 = 1.9$ as well. (The exact equality of the two averages is accidental, though the figures are usually fairly similar). Note that this type of tabulation is tricky to do correctly when working direct from the summary of Table C7, and so it may be quicker in the long run to write the figures out as above.

The complete results for the $w_{X.Y}$ (and w_X) rates of buying are given by the following non-symmetric table:

Buyers of	Average Purchases of X	Y	Z
X	(1.92)	1.9	1.8
Y	1.9	(1.64)	1.5
Z	2.0	1.0	(1.44)
Average	2.0	1.5	1.7

The D - Coefficient. The Duplication Coefficient can be estimated as the ratio of the average of the six duplications for all brands to their average penetration, eg.

(73 + 56 + 62 + 22 + 38 + 18)/6 = 45 divided by (6.5 + 5.5 + 4.5)/3 = 5.5, so that D = 4515.5 = 8.2

(which in our artificial example is abnormally high; usually it tends to lie in the range 1 to 2).

An alternative form of estimation is to take the average of all the ratios of $b_{Y.X}/b_Y$, calculated for all pairs of brands. Where exceptional duplications occur one can calculate a normative value of D for the more homogeneous subset of duplications.