CHAPTER 7

THE NBD THEORY

7.1. The Nature of the Theory

The NBD theory provides an integrative model for the various aspects of repeat-buying, under stationary or "no-trend" conditions. The basic concept in the theory was outlined in § 4.5. It is that the purchases of a brand made by each potential buyer are "as if" random over time and independent of each other (i.e. a Poisson process). Each consumer has his own average frequency of purchase, and these average purchase frequencies (i.e. the means of the Poissons) follow a Gamma distribution with parameter k. This is set out more fully in § 7.2 of this chapter.

Predictions about various aspects of buying behaviour follow from this formulation. These have generally been verified for a great.variety of different frequently-bought products and brands, as has been illustrated in Part II.

One particular deduction from the model is that the frequency of buyers making 0, 1, 2, 3, etc. purchases in a given time-period should follow a Negative Binomial Distribution. This distribution is discussed in § 7.3.

Purchasing patterns in more than one time-period can be represented in the model's more general form as a *m&i-variate* NBD (§ 7.4). One aspect of the Poisson-Gamma formulation is that the Gamma parameter k should be constant over different length time-periods (§ 7.5). This provides the basis for inter-relating penetration levels and average purchase frequencies in such periods. The formulae for the incidence of repeat-buyers from one period to another which also follow are treated in § 7.6, together with the more general form of "conditional trend analysis" in which repeat-buying by light, medium and heavy buyers is tracked.

NBD's for different brands or pack-sizes can in theory be aggregated under certain conditions which are set out in § 7.7. A general discrepancy problem in fitting theoretical frequency distributions occurs for items which an appreciable proportion of people buy more or less regularly once a week, there being then a theoretical excess of still heavier buyers (§ 7.8). This is the "variance discrepancy" problem which links up with the existence of a minimum inter-purchase interval for many products (§ 7.9).

7.2. The Compound Poisson Model

A theoretical formulation of buyer behaviour which leads to a Negative Binomial Distribution for the frequency distribution of purchases in any given time-period is the stochastic model which was introduced in §4.5 of Chapter 4. Thus to analyse stationary purchasing behaviour over successive (equal) periods of time, the following two-parameter model of a compound Poisson form was postulated:

(i) Purchases of a given consumer at successive points in time can be regarded as independent drawings from a Poisson distribution.

(ii) The average rates of purchasing of different consumers in the long run differ, their distribution being Gamma with exponent k and mean m.

It is not necessary to assume that this model holds indefinitely over time, but only that in any time-period one wishes to consider, the purchasing data should behave as if it did. The postulation is not unreasonable on a priori grounds (except for very short time-periods – see §7.9), but the real question is the empirical validity and the practical utility of specific deductions made from it.

One basic deduction is that the distribution of the number of purchases r made by different consumers in any given time-period should be an NBD. Thus in a time-period of "unit" length, the Poisson assumption says that the probability that a consumer with a long-run mean purchasing frequency μ would make r purchases ($r \geqslant 0$) is

$$\frac{e^{-\mu}\mu^r}{r!} .$$

Multiplying this by the Gamma-distribution probability that the consumer has a mean purchasing rate of value μ in the long-run, namely

$$\frac{e^{-\mu/a}\mu^{k-1}}{a^k \Gamma(k)}$$

(where $a = m/k$), and integrating over all μ (i.e. over all consumers), gives

$$\int_0^\infty \frac{e^{-\mu/a}\mu^{k-1}}{a^k\Gamma(k)} \frac{e^{-\mu}\mu^r}{r!}\, d\mu = (1+a)^{-k} \frac{\Gamma(k+r)}{\Gamma(r+1)\,\Gamma(k)} \left(\frac{a}{1+a}\right)^r .$$

This is the term for the probability p_r of r occurrences (i.e. purchases) in an NBD with exponent k and mean $m = ak$, as set out in § 7.3 below.

This formulation has had previous practical applications, e.g. in the study of accident statistics, in certain ecological birth and death and contagious processes, and in some operational research theory [see for example Greenwood and Yule 1920, Irwin 1964, Kemp 1970].

The fact that the NBD tends to give a good fit to observed distributions in a single time-period has already been discussed in Part II and lends support to the Poisson-Gamma model. But the real support for the model comes from the extent to which many *other* deductions from it also hold. The point is that a "compound Poisson" type of model is only one of several theoretical stochastic processes which can lead to a negative binomial distribution for the distribution of occurrences in a given time-period, as discussed for example by Anscombe [1950]. One of the alternatives is, in terms of buying behaviour, that purchasing occasions are still distributed as a Poisson distribution, but that the distribution is the same for all consumers, and that the amounts bought per occasion are distributed as a logarithmic series distribution. This model has been discussed by Williamson and Bretherton [1964] in the context of industrial purchasing, and has been suggested by Professor Cramer [1965] as a possible model for consumer goods purchasing as analysed here. A mathematical derivation of this model was sketched in earlier by Quenouille [1949]. However, this model does not appear applicable here. Thus, it is inconsistent with general experience to suppose that different consumers' average purchasing patterns are the same, whilst in many product-fields, the amount bought per purchase occasion is virtually 1 (see Table 3.3 in Chapter 3). More mathematically, this alternative model requires that the parameter $a = m/k$ should remain constant for all different lengths of analysis periods and that the parameter k should vary, whilst in practice the opposite is found to be so, as is discussed in § 7.5 below.

A potentially more open question occurs within the confines of the compound Poisson model itself. This relates to an old controversy in accident statistics, where there is an observed tendency for some people to have more accidents than others. The question is whether this is because having had an accident makes it more likely for that person to have another (i.e. "contagion" or "learning"), or that people differ in some inherent "proneness" to have accidents. Related discussions for other buyer behaviour models between regarding consumers as inherently differing from each other (population heterogeneity or "proneness") or as all being initially the same but conditioned by their differing stochastic experiences ("learning" or "contagion") are summarised by Massy et al. [1970]. The evidence in analysing empirical pur-

chasing behaviour under stationary conditions points clearly to the Poisson-Gamma formulation in general, and the heterogeneity version in particular, as providing a variety of successful descriptions.

7.3. **The Negative Binomial Distribution**

The background and technicalities of the Negative Binomial Distribution itself are briefly as follows. It is a two-parameter discrete distribution for certain probabilities p_r of observing any non-negative integer r, where

$$p_r = \left(1+\frac{m}{k}\right)^{-k} \frac{\Gamma(k+r)}{\Gamma(r+1)\,\Gamma(k)} \left(\frac{m}{m+k}\right)^{r}.$$

The two parameters are usually expressed as the mean m and the exponent k, with the ratio $a = m/k$ also being a convenient function of the two parameters.

The probabilities p_r arise in expanding an expression of the binomial form

$$\left(1-\frac{m}{m+k}\right)^{-k} \quad \text{or} \quad \left(1-\frac{a}{1+a}\right)^{-k},$$

with a *negative* exponent $-k$. Thus $\{1-(a/1+a)\}^{-k} = 1 + k(a/1+a) + k(k+1)(a/1+a)^2/2 + \ldots$. and multiplying by $(1+a)^{-k}$ to make the probabilities sum to 1, *we* have $p_0 = (1+a)^{-k}$, $p_1 = (1+a)^{-k}k(a/1+a)$, etc.

A general statement of the NBD was probably first given by Montmort in 17 14, within a year of Bernouilli's derivation in 1713 of the much better-known positive binomial distribution obtained from expanding an expression like $(p+q)^n$. (Unlike the positive binomial – which can refer to the proportion of times an event with probability p occurs in n independent trials – the mathematical derivation of the NBD from a binomial expansion appears to have no direct physical meaning for our purposes here.) The general history of the negative binomial distribution has been summarised in reviews of Gurland [1954], Bartko [1961] and Boswell and Patil [1970] – see also Patil and Joshi [1968] and Johnson and Kotz [1969].

One way of expressing the NBD is through its probability generating function

$$(1 + a - au)^{-k}$$

where u is a dummy variable. This is a particularly convenient approach when dealing with the *multivariate* NBD in the next section, and the

univariate case is therefore worth setting out here. Thus expanding this expression in powers of u gives the probability p_r of r occurrences as the coefficient of u^r, i.e.

$$(1 + a - au)^{-k} = (1+a)^{-k}\left(1 - \frac{au}{1+a}\right)^{-k}$$

$$= (1+a)^{-k}\left\{1 + k\left(\frac{a}{1+a}\right)u + \frac{k(k+1)}{2.1}\left(\frac{a}{1+a}\right)^2 u^2 + \ldots\right.$$

$$\left. + \frac{\Gamma(k+r)}{\Gamma(r+1)\,\Gamma(k)}\left(\frac{a}{1+a}\right)^r u^r + \ldots\right\}.$$

Reading off the coefficient of u^r gives $p_0 = (1+a)^{-k}$, $p_1 = ka(1+a)^{-k-1}$, and so on.

The mean of the NBD is

$$\sum_r rp_r = (1+a)^{-k} \sum_r \frac{r\Gamma(k+r)}{\Gamma(r+1)\,\Gamma(k)}\left(\frac{a}{1+a}\right)^r$$

$$= (1+a)^{-k} \sum_r \frac{k\Gamma(k+r)}{\Gamma(r)\,\Gamma(k+1)}\left(\frac{a}{1+a}\right)^r$$

$$= ka\left\{(1+a)^{-k-1} \sum_r \frac{\Gamma(k+r)}{\Gamma(r)\,\Gamma(k+1)}\left(\frac{a}{1+a}\right)^{r-1}\right\}$$

$$= ka = m,$$

since the terms in brackets are an NBD expression with parameter (k+l), which sums to unity. Other moments, and in particular the formulae for the variance mentioned in Chapter 4,

$$m(1+a) \quad \text{or} \quad m(1+m/k),$$

follow similarly by working through the expression $\Sigma(r-m)^2 p_r$.

As discussed in § 4.2 of Chapter 4, the parameter k can in general be best estimated from the observed mean m and the proportion of non-buyers p_0 by solving the expression

$$p_0 = (1+a)^{-k} \quad \text{or} \quad (1+m/k)^{-k},$$

subject to the condition for the existence of an NBD that $m \geqslant -\ln p_0$.

7.4. **The Multivariate** NBD

The NBD can be generalised to more than one time-period in terms of a very powerful reformulation as a *multivariate* NBD, as noted by

G.J. Goodhardt. Instead of dealing with one time-period, this deals with any number $i = 1$ to t time-periods of varying lengths T_i. The probability generating function (*p.g.f.*) of the distribution of people' making r_i purchases in the ith period of length T_i, making r_j purchases in the jth period of length T_j, etc., can be written as

$$\{1 + a \sum_{i=1}^{t} T_i(1-u_i)\}^{-k},$$

where the u_i are dummy variables, m is the average amount bought in a time-period of some convenient "unit" length, and a is again m/k, in terms of the negative exponent $-k$.

The co-efficient of

$$(u_i)^{r_i}(u_j)^{r_j} \ldots$$

in the expansion of the p.g.f. in powers of the dummy variables u_i, u_j, etc. gives the proportion of the population who make r_i purchases in period i, r_j purchases in period j, etc. All the specific formulae for repeat-buying and for different length time-periods discussed in this book then follow. For example, the proportion of consumers who are "lapsed" buyers (i.e. who buy in the first but not in the second of two equal periods) comes from putting $t = 2$, $T_1 = T_2$, and then adding the co-efficients of $(u_1)^{r_1}(u_2)^0$ over all non-zero values of r in the first period.

The general multivariate p.g.f. clearly simplifies to the p.g.f. of the univariate NBD in a single period of length T

$$\{1 + aT(1-u)\}^{-k},$$

or to the p.g.f. $\{1 + a - au\}^{-k}$ for "unit" length time-period as given in § 7.3. More generally, using arguments similar to those used by Feller [1957] in discussing so-called "generalised" distributions, the multivariate model can be shown to yield an NBD when partitioned in quite a number of different ways. In particular, if the $i = 1$ to t time-periods are divided into the first s and the remaining $t-s$, then the conditional distribution of $(r_{s+1}, r_{s+2}, \ldots r_t)$ purchases in periods $s+1$ to t, given $(r_1, \ldots r_s)$ purchases in the first s periods is itself a multivariate NBD with parameters m' and k' given by

$$m' = (k + \sum_{i=1}^{s} r_i)\{m/(m+k)\}, \qquad k' = (k + \sum_{i=1}^{s} r_i),$$

where m is again the mean in a time-period of some "unit" length. This typifies the great simplicity of the multivariate NBD in terms of its partitioning and additivity.

The practical applications of the NBD to other applied areas (such as accident statistics) have in the past been mostly restricted to dealing with the frequency distribution in a single time-period. The basic mathematical properties of the bivariate NBD were however reported by Arbous and Kerrich [1951], but they did not explicitly note the NBD properties of their results, nor develop their possible practical applications. An explicit treatment of the multivariate NBD was provided by Bates and Neyman [1952], with some illustrative empirical applications, and had also already been tackled in Lundberg's pioneer work [Lundberg 1940]. Some practical concern with pairs of time-periods and with the bivariate frequency distribution of accidents has been shown by Cresswell and Frogatt [1963], but the development was not taken very far.

7.5. The Theoretical Constant *k* in Different Length Time-Periods

The properties of the parameter k in the NBD expression $\{1+a\Sigma T_i(1-u_i)\}^{-k}$ are basic to the NBD theory. As already noted, the two parameters of the NBD model are usually denoted firstly by the exponent k, and secondly either by m, the mean number of the purchases made by all consumers in the population in a time-period of some convenient (but arbitrary) unit length, or by the quantity $a = m/k$.

Under stationary conditions, the mean m in equal time-periods must be the same, because of the defmition of stationarity (viz. no trend in the level of aggregate sales). The numerical value of m in any given time-period is therefore proportional to its length. Denoting by m_T the mean in a period of length T relative to the "unit" time-period, we have under stationary conditions that

$$m_T = Tm .$$

The parameter m or, more generally, m_T, therefore acts as a scale-factor, reflecting the length of the analysis-period.

In contrast, the parameter k in the theoretical Poisson-Gamma model is constant for different lengths of analysis-period (e.g. 4 weeks, 8 weeks, 24 weeks, etc.). Thus k is the parameter of the Gamma-distribution which describes the long-run differences in the average purchasing rates of different consumers. Most of the repeat-buying formulae derived from the NBD model depend explicitly on k being constant in this way.

The empirical behaviour of estimates of the parameter k for different lengths of time-periods is therefore of crucial importance to the validity of the NBD model. For any given (near-stationary) brand, the value of k_T for a period of length T can be estimated from the observed proportion b_T of consumers who buy and the mean number of purchases m_T per consumer in the time-period by solving the equation

$$b_T = 1 - \left(\frac{k_T + m_T}{k_T}\right)^{-k_T},$$

as discussed in § 4.2 of Chapter 4 and Appendix A. If the NBD model holds, the values of k_T for the different points should then be the same, i.e. independent of T.

Table 7.1 shows a typical example of the constancy of k for a typical brand in a certain product-field X [Chatfield et al. 1966] in the 4-, 8-, 12-, and 24-weekly periods which were analysed. It also contrasts this constancy of k with the increase in the observed amount m_T bought, which is sixfold and therefore *pro rata* to T, with the corresponding increase by a ratio of about 2 in the observed proportions b_T of the population buying in each period, and with the increase by a factor of 3 in the observed average frequency of purchase w_T bought *per buyer* (i.e. m_T/b_T).

The values of k for all other relevant analyses available in the same product-field (i.e. for 14 brands or pack-sizes for which near-stationary data from four weeks up to 24 weeks could be analysed) also showed

Table 7.1. The Virtual Constancy of the NBD Parameter k for Increasing Lengths of Time-Period, and the Increase of other Observed Statistics

(A typical brand, and the average value of k for 14 other brands or pack-sizes in product-field X)

	Length of Period (in Weeks)				Ratio of 24-Week to 4-Week values
	4	8	12	24	6
A typical brand:					
Mean amount bought per household, m_T	.12	.24	.36	.72	6
Percentage of buyers, $100b_T$	4	5	6	8	2
Average amount per buyer, $m_T/b_T = w_T$	3	5	6	9	3
NBD parameter k_T, shown as $10k_T$	.20	.19	.21	.23	1
Average k_T for 14 brands, as $10k_T$	.57	.49	.48	.54	1

no trend in k, as is shown by the average values of k_T in the last line of the table. The same result has since been seen in many other product-fields.

Since k_T is constant under stationary conditions and $m_T = Tm$, the parameter $a_T = m_T/k$ acts as a scale-parameter like m_T itself, i.e. a_T is proportional to T. The formulae relating results in different length time-periods in §4.8 of Chapter 4 are founded on this simple basis. Firstly, we have that

$$b_T = 1 - (1+a_T)^{-k} = 1 - (1+aT)^{-k} ,$$

in terms of $a = m/k$ and k for the "unit" length period. Secondly, since $m_T = Tm$, the earlier formula for w_T, the average purchase frequency *per* buyer, follows immediately from $w_T = m_T/b_T$, i.e. $w_T = Tm/\{1-(1+aT)^{-k}\}$.

7.6. Repeat-Buying in Two Periods

One of the deductions from the NBD model in Chapter 4 concerned repeat-buying behaviour from one time-period to the next under stationary conditions. Two basic features are the proportion b_R of consumers who buy the brand or pack-size in both periods, and the average number of purchases m_R (expressed on a "per consumer" basis) made in each time-period by such "repeat-buyers". Corresponding results for "new" and for "lapsed" buyers are obtained by substraction from the values of b and m for *all* buyers in a single period.

The terminology of repeat, "new" and "lapsed" buyers here is defined in terms of any pair of equal time-periods I and II according to whether a consumer buys in both periods (repeat-buying), or in the Period II only or in Period I only, as was set out in Table 4.4 of Chapter 4. None of these definitions depend in any way on what the consumers in question did in still earlier or in later periods; thus the so-called "new" buyers may have bought the brand prior to Period I (and generally will in fact have done so), and the "lapsed" buyers may well buy the brand again after Period II.

The quantities b_R and m_R concerning repeat-buying behaviour in two time-periods can then be related to (or predicted from) knowledge of purchasing behaviour in a single time-period, and in particular from the average number of purchases m in the first time-period and the proportion b of consumers who buy this item at all in that first time-period. More specifically, the values of b_R and of m_R can be expressed

by formulae which only involve m and the NBD parameter k, the latter being calculated from the values of b and m in the single time-period according to the usual formula $b = 1 - (1+m/k)^{-k}$ or $1 - (1+a)^{-k}$.

Thus according to the result for b_T for $T = 2$ at the end of the preceding section, the proportion of the population buying in the combined period is

$$b_2 = 1 - (1+2a)^{-k}$$

in terms of the first period's values of a and k, and since

$$b_R = 2b - b_2 ,$$

we have

$$b_R = 1 - 2(1+a)^{-k} + (1+2a)^{-k} .$$

The proportions b_N and b_L of "new" and "lapsed" buyers are under stationary conditions correspondingly given by $b - b_R$, so that

$$b_N = b_L = (1+a)^{-k} - (1+2a)^{-k} .$$

The average frequency of purchase m_N (or m_L) by "new" (or lapsed) buyers can easily be calculated by using the Poisson assumption. Thus a consumer with mean purchasing rates μ will be a non-buyer in the first period with Poisson probability $e^{-\mu}$. Given the independence property of the model, his purchases in the next period will still have an expected value μ. Integrating across the Gamma-distribution of μ for all consumers, the average purchasing rate m_N by "new" buyers (on a "per consumer" basis) is

$$\int e^{-\mu}\{(1/a)^k e^{-\mu/a} \mu^{k-1}/\Gamma(k)\}\mu \, d\mu = \frac{ak}{(1+a)^{k+1}} \int \left(\frac{1+a}{a}\right)^{k+1} \frac{e^{-\mu(1+a)/a} \mu^k \, d\mu}{\Gamma(k+1)}$$

where the integral is a Gamma-distribution with parameters $(1+a)/a$ and $(k+1)$, which therefore sums to 1. In terms of the mean purchasing rate m of all consumers, we therefore have

$$m_N = \frac{m}{(1+a)^{k+1}} .$$

Correspondingly, the average purchasing rate m_R by repeat-buyers (again on a "per consumer" basis) is

$$m_R = m - m_N = m\{1 - (1+a)^{-k-1}\} .$$

Conditional Trend Analysis. For more detailed cross-analyses of buying behaviour from one period to another, we consider all the consumers who made exactly r purchases in Period I. The "conditional" distribution of their purchases in Period II then turns out to be itself a negative binomial, with mean $(k+r)/\{a/(1+a)\}$ and with the k type of parameter of "exponent" taking the value $(k+r)$. (This follows from the multivariate results in § 7.4.) The proportion of the buyers of r purchases in the first period who buy in the next period is for example

$$1 - \{1 + (a/1+a)\}^{-k-r}.$$

This "conditional" type of analysis was used in Table 3.10 in Chapter 3 and in the case-histories in § 5.6 of Chapter 5 and § 6.2 of Chapter 6. The formulae for the incidence of "new" and "lapsed" buyers and for their average rates of purchasing which have already been given can also be deduced as special cases (i.e. with $r = 0$) of this more general result.

The simple conditional distribution for any value of r in two equal time-periods generalises further for two unequal periods, of unit and relative length T respectively. Thus the purchases in the second period (of relative length T) made by those consumers who buy r units in the first period still follow an NBD, with mean $(k+r)[aT/(1+a)]$ and exponent $(k+r)$, where a again refers to the unit-length period. The formulae for varying time-periods set out in § 7.5 of this chapter can be regarded as special cases of these conditional distributions.

7.7. The Combination of the NBD's

Two major kinds of aggregation (or dis-aggregation) problems occur in analysing repeat-buying behaviour by means of the NBD model. One is how repeat-buying patterns in sub-groups of the population compare with those in the population as a whole. The other is how the repeat-buying patterns for individual items such as a particular pack-size or brand compare with those for ***combinations*** of such items, e.g. all the pack-sizes of a brand combined, or different brands combined into some product-class total.

For subgroups of the population such as consumers in specific age-groups or in different geographical regions, the question is whether the NBD will give a good fit for such sub-groups when it does so for the

population as a whole. It can be shown that in theory, different negative binomial distributions, variously weighted and added together, will almost never combine into an exact negative binomial distribution.

In practice however, negative binomial distributions for sub-groups seem to combine into a negative binomial distribution for the population as a whole, the fit in all cases of course being approximate rather than absolutely exact. This is illustrated in Table 7.2 for a breakdown by household size, the socio-economic breakdown which usually produces the sharpest differences in purchasing levels – here an average of .7 purchases per household of 5 or more persons, and only .2 purchases per 1-person household, and the parameters of the fitted distribution differ accordingly. However, the distributions of the purchases made by households' of sizes 1, 2, 3, 4 and 5+ respectively, and by all households together, are all virtually negative binomials, as is illustrated by the fact that the observed and theoretical estimates of the standard deviations, s and σ, in each size-group agree well.

Table 7.2. Negative Binomial Parameters in Different Size-of-Household Groups for a Certain Brand

Parameters (Rounded)	Size of Household Groups					Total Sample
	1	2	3	4	5+	
Mean m	.2	.4	.5	.6	.7	.5
NBD Parameters:						
$10k$	.3	.4	.6	.8	.7	.6
$a = m/k$	6	10	9	8	10	9
Standard Deviations:						
Observed s	1.2	2.2	2.5	2.3	2.9	2.4
Theoretical σ	1.2	2.2	2.3	2.3	2.7	2.2

The explanation of this paradox between theory and practice lies mainly in the fact that even for something with as wide a range of purchasing levels m as a breakdown by size-of-household, the differences between the distributions for the subgroups are small compared with the scatter within each subgroup. Furthermore, the breakdown by size-of-household illustrated here reflects the apparently common case where the extreme groups are small. Thus households containing 2, 3 or 4 people forrn the great majority of all households, and the distributions for these three sub-groups (i.e. their means) are relatively similar.

(Combining sizes 1 and 5+ only, for example, might well not lead to a good negative binomial fit, but such a combination would not usually be of practical interest.)

Turning to the second type of aggregation mentioned above, we consider the fit of the distribution for an individual brand or pack-size on the one hand and for combinations of such items (e.g. the whole product-group) on the other. We have a theoretical result which can be derived from the probability generating function of the NBD in § 7.3, namely that if x and y are two negative binomial variables with parameters (m_x, k_x) and (m_y, k_y), then their sum $(x+y)$ will follow a negative binomial distribution under two conditions, namely if and only if the two variables are independent and if $a_x = a_y$, where $a = m/k$ as usual. The parameters of the aggregated NBD will be $(m_x + m_y)$, $(k_x + k_y)$. Even though one may expect the result to hold approximately when a_x and a_y are not exactly equal, this condition would seem to impose a marked constraint on aggregation possibilities. However, some powerful empirical findings described in Chapter 10 show that these conditions do in fact tend to be fulfilled to quite a close degree of approximation under a wide range of circumstances, as is discussed further in § 11.4 of Chapter 11.

7.8. The Variance Discrepancy

The basic finding in developing the NBD model was that the observed distribution of purchases in any single period could generally be well fitted by the negative binomial distribution. Table 3.4 in Chapter 3 gave some very recent examples, and Table 4.1 in Chapter 4 one of the earliest.

When fitting by using the proportion of buyers b (or $p_0 = 1 - b$) and the mean m as the observed data, the degree to which the standard deviations (or variances) of the observed and theoretical distributions agree can be used as a measure of fit, as discussed in § 4.2 of Chapter 4. Fig. 7.1 summarises in this sense the earliest results [Ehrenberg 1959] amounting to some 150 varied cases (different products, brands, time-periods, etc.).

For standard deviations up to 1 or 2, the agreement between the observed standard deviations and the theoretical values $\sqrt{\{m(1+a\}}$ of the fitted NBD was clearly good. However, a failure to fit becomes apparent for larger values of the standard deviation, where the theoretical value is generally higher than the observed one. A striking feature of

the data is that these discrepancies in the standard deviation (or variances) are themselves extremely regular and systematic. This "variance discrepancy" – i.e. a failure to fit under certain circumstances – has therefore been known from the earliest stage of the work and has been confirmed in many other cases since.

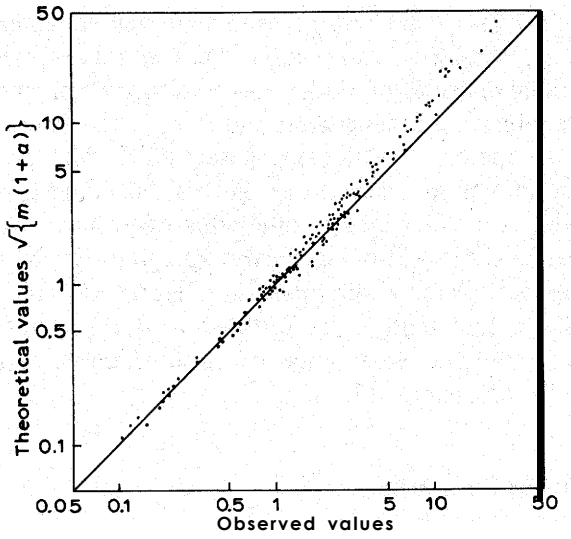

Fig. 7.1. Comparison of 'theoretical' and 'observed' values for the standard deviation of the frequency distributions of consumer purchases.

If we denote the observed variance by s^2 and the theoretical variance by $\sigma^2 = m(1+a)$, Fig. 7.1 shows that the numerical differences $\sigma - s$ increase rapidly and systematically with increasing values of s. One possibility in trying to describe the variance discrepancy $\sigma^2 - s^2$ is therefore to fit some suitable mathematical equation to the values of σ^2 and s^2 themselves. However, this is not the only possibility. Thus $(\sigma^2 - s^2)$ is correlated not only with s^2, but also with the average rate of buying m, and with the length of the analysis-period T. And there could be various interactions between all these factors.

To account for all these possible relationships, the NBD model provided a starting-point (despite its failure to provide an exact fit here).

Firstly there was the effect of the length of the analysis-period T. According to the NBD model, $\sigma^2 = m(1+m/k)$ and so for a stationary brand the expression

$$\frac{\sigma^2 - m}{m^2} = \frac{1}{k}$$

should be constant irrespective of the length of the analysis-period, since k itself is independent of the length of time-period. In studying various brands and pack-sizes in a product-field where the variance discrepancy generally occurred, it was found [Chatfield et al. 1966] that the corresponding expression $(s^2 - m)/m^2$ for the observed variance s^2 was also approximately constant. (For example, the value of $(s^2 - m)/m^2$ for all the items analysed averaged at about 21 in the 4-weekly periods and at 22 in the 24-weekly periods.) We therefore have that the two expressions $(s^2 - m)/m^2$ and $(\sigma^2 - m)/m^2$ tend each to be independent of the length of time-period analysed.

Secondly it was found empirically that these two quantities were related in such a way that the differences between their square roots were approximately constant, irrespective of the brand and of the values of s or σ (and of course by now irrespective of the length of time-period). The numerical value of these differences averages at a little less than 1 for the product-field in question. Taking the value to be 1 as a rough but convenient approximation, we have

$$\sqrt{\{(\sigma^2 - m)/m^2\}} - \sqrt{\{(s^2 - m)/m^2\}} \doteqdot 1 .$$

This equation held with a mean deviation of about .5 for the data in question, where the observed variances ranged from less than .1 to about 50.

This relationship can be further simplified. In consumer purchasing data m is generally low, and often very low, compared with s^2 (especially for longer time-periods). Ignoring therefore the term m in both $(s^2 - m)$ and $(\sigma^2 - m)$, the above relationship simplifies to the difference between the observed and theoretical co-efficients of variation s/m and a/m of the observed distribution of purchases and the fitted NBD,i.e.

$$a/m - s/m \doteqdot 1.$$

For the initial data this relationship again held with a mean deviation of about .5. The relationship can also be conveniently expressed in the form

$$\sigma - s \doteqdot m .$$

In some subsequent work on the variance discrepancy it appeared that a closer fit might generally be obtained by something like

$$\sigma - s \doteqdot m/2 ,$$

and more systematic study is still required. However, whatever the precise numerical values should be, this type of relationship is simple, and simple to use. For example, it helps to resolve the apparent conflict between the two kinds of results that have occurred — one where the NBD appears to give a good fit and $\sigma \doteqdot s$ (as in the example of Table 4.1 and the smaller values represented in Fig. 7.1), and the others where it does *not give* a good fit and $\sigma > s$ (as shown for example by the larger values in Fig. 7.1). Both cases can now be subsumed by the "discrepancy" formula $\sigma - s \doteqdot m$ (or $\sigma - s \doteqdot m/2$), since this also covers the case $\sigma \doteqdot s$, namely when m is very small compared with s.

All this however does little more than cope descriptively with the variance discrepancy as such. It does not provide a more successful theoretical distribution to fit to observed data subject to the variance discrepancy. However, further insight into the nature of the discrepancyhas been provided by a number of other checks.

One negative finding concerns the assumption of a Poisson distribution for each consumer's purchases over time. This is known not to apply in very short time-periods (as discussed in § 4.9 of Chapter 4 and also §7.9 below). But any failure of this assumption to hold in longer time-periods could not in fact account for the variance discrepancy. Thus the component of the NBD variance $m(1+a)$ which is due to the Poisson distribution amounts to m, and this is generally small compared with the term ma and therefore small compared with the variance discrepancy itself, especially in longer time-periods. (The $\sigma - s \doteqdot m$ formula implies that in terms of *variances,* the discrepancy would be of the order of m^2. A detailed analysis [Chatfield and Goodhardt 1972] of one possible alternative formulation — replacing the *Poisson* distribution by an *Erlang* distribution (a Poisson with every other reading censored) as suggested by Herniter [1969] — has shown no effective difference from the NBD in the resulting distribution of purchases for individual brands.)

Other possible factors which have led to negative conclusions [as discussed elsewhere, eg. Chatfield et al. 1966] are the variable marketing mix, non-stationarity in the data, the occurrence of excessively heavy buyers, and errors of measurement such as under-recording of very frequently-purchased items or relative over-claiming of infrequent

purchases. Errors of measurement for example seems an unlikely explanation, given the variety of different types of product-fields covered (*see* Table 4.2 and the discussion in § 1.2 of Chapter 1), and with data collected by somewhat different techniques.

Another finding is that the fit of the repeat-buying formulae from one period to another as discussed in § 7.6 does not seem to be systematically affected by whether or not the variance discrepancy occurs within each time-period. In other words, the repeat-buying patterns appear to be largely self-compensating in this respect. A *theoretical* implication is that the fit of the repeat-buying formulae is not necessarily a very rigorous test of the underlying rationale and assumptions of the NBD model. The *practical* consequence is however that the variance discrepancy is usually of relatively little concern in applications of the theory to period-to-period repeat-buying. It only arises when dealing directly with the frequency of purchase as such (which includes the "conditional" type of analyses in § 7.6).

7.9. Shelving

A feature of consumer purchasing for grocery products which was noted in the earliest analyses is that purchase frequencies tend to cluster or "bunch" at or near a number equal to the number of weeks in the analysis-period. Typically, the observed frequencies at or near this point exceed the values of the theoretical distribution that has been fitted. This has already been briefly commented on in § 4.3 of Chapter 4.

"Bunching" occurs because a small number of consumers report exceptionally regular purchasing, for example one virtually every week. This is largely due to a few consumers marginally adjusting their purchasing behaviour (and possibly their actual consumption habits) to fit into such a neat purchase-per-week pattern along the lines already indicated in § 4.9 of Chapter 4. (It could also be due to regular ***reporting,*** rather than to regular purchasing, i.e. an error of measurement or bias in the data collection procedure.)

The numerical incidence of such very regular reporting of virtually one purchase per week is usually relatively small, the vast majority of buyers reporting much more sporadic and irregular purchasing patterns. In early work it seemed that bunching of this kind would have little numerical effect on the fit of the NBD in general and on the variance discrepancy in particular. Thus if the bunched purchases were "smoothed",

i.e. distributed on either side of their peaked value, the mean and the variance of the observed distribution would hardly be affected.

This earlier analysis of the bunching phenomenon was however too superficial. Thus it was not just a case of there being a few too many purchasers buying at or near the critical frequency (e.g. equal to the number of weeks in the analysis-period), but that there are consistently too few buyers buying *more* frequently than that. The observed distribution contains a *shelf-like* discontinuity. Since for most goods there are in any case very few frequent buyers, the actual number of buyers involved in the "shelving" phenomenon is usually small. But since it is the *heavy* buyers in the tail of the distribution who are missing, the effect in terms of their sales importance is quite large (as illustrated by the bracketed figures in Table 2.1 of Chapter 2), and so is the effect on the variance. This is illustrated in Table 7.3 where there is a theoretical excess of 1.5% of sample supposedly buying more often than once a week (2.1% theoretical versus .6% observed), but the variances (or standard deviations) differ quite markedly.

The "shelving" effect occurs because for many different products there is an upper limit to the number of times people will generally buy

Table 7.3. A Typical Example of "Shelving"

(The observed frequency distribution of purchases of a certain brand in 4 weeks, and the NBD fitted to the mean and the proportion of zeros)

No. of Purchases in 4 Weeks	Observed	NBD
	%	%
0	79.3	(79.3)*
1	8.4	10.6
2	5.1	4.5
3	3.0	2.3
4	3.6	1.3
5	.2	.8
6	.4	.5
7+	0	.8
Total Sample	100%	100%
Mean m	.45	(.45)*
Variance	1.13	1.54
Standard Deviation	1.06	1.24

* Used in fitting.

it in a given time-period. This upper limit is mainly a matter of *purchasing* habits (i.e. a tendency to buy at most once a week, say on a Friday), but it may in some cases also be influenced by *usage* habits (i.e. a need to use up one purchase before another one is made). For many household products (e.g. many packaged foods, soaps and toiletries), the effective minimum inter-purchase time is the week *. In Table *7.3, we* typically have very few people buying more than 4 times in 4 weeks, and a corresponding higher "shelf" of observed frequencies at or just below 4 purchases. In contrast, the distribution shown in Table 4.1 in Chapter 4 showed no variance discrepancy and no shelving because there were only 3 households buying anything like as often as once a week (about 0.15% of the sample, or 1% of all buyers, compared with something like 7% of the sample, or a third of all buyers, in Table 7.3).

There are then few if any consumers who buy more often than the number of "minimum" time-periods in the analysis-period, e.g. more than 12 times in a 12-week quarter, or 24 times in a 24-week half-year. The converse of this "shortage" of very frequent buyers – which only makes itself manifest if appreciable numbers of buyers *do* buy as often as once a week – is the much more general failure of the NBD repeat-buying model to fit for very short time-periods, i.e. at or near the minimum inter-purchase interval, as has already been discussed in § 4.9 of Chapter 4.

The general conclusion therefore is that the NBD model only breaks down near the boundaries, i.e. for very short time-periods and for relatively large numbers of purchases. Despite this limited failure, it seems that a fundamental reformulation of the theory is required. One possibility which may deal with the problems more fully is analysis in terms of purchasing *periods,* not purchasing occasions, as is touched on in Chapter 11.

For the present, the NBD theory (and the LSD approximation) provide the best working tool available. It tends to give a good fit in the "middle range" of cases, i.e. for relatively long time-periods (say 4 weeks or more for grocery products), and for brands or other items which relatively few consumers buy as regularly as every week.

* With more frequently-bought products like petrol, cigarettes, milk and bread, the situation is as yet less clear but there is little doubt that the basic problems discussed in this section also apply.

7.10. Summary

The NBD theory states that under stationary, no-trend conditions, a consumer's purchases over time are effectively like an independent random drawing from a Poisson distribution, the mean of the distribution for different consumers – their average long-run frequencies of purchase – being distributed according to a Gamma-distribution with exponent k.

This stochastic formulation leads to an NBD for the distribution of purchases in any given time-period, and a multivariate NBD with probability generating function $\{1+a\Sigma T_i(1-u_i)\}^{-k}$ for any number $i = 1$ to t time-periods of lengths T_i. This multivariate distribution partitions readily, in that for any sub-group defined by their purchases in one or more periods, their purchases in *other* periods are again NBD.

The model has two parameters, the overall mean m in a given time-period and the exponent k. Under stationary conditions, the mean m (or the alternative parameter $a = m/k$) varies directly with the length of the analysis-period, whilst the exponent k remains the same. From this, formulae for penetration-growth over time and for various aspects of repeat-buying from one period to another can readily be deduced.

Different NBD's can be aggregated if they are independent and have the same values of a. These two conditions tend in practice to be approximately fulfilled, as is discussed in Chapter 11.

In time-periods near the minimum inter-purchase time which tend to operate for many products – such as a week for many grocery products – repeat-buying is of a different form, almost no one buying more than once in any one such period. This shows up as the "variance discrepancy" or "shelving" effect, where the NBD overstates the number of people who buy more often than 10 times in 10 weeks say, but this is only noticeable for products where an appreciable proportion of consumers *do* buy as often as about once every week.

CHAPTER 8

THE LSD THEORY*

8.1. An Approximation to the NBD

As already noted in Chapter 4, the LSD theory of repeat-buying is based on the Logarithmic Series Distribution and gives a very close approximation to the NBD theory in certain parameter ranges (roughly, a value of the NBD exponent k less than 0.2).

The Logarithmic Series Distribution is simpler than the NBD, having only one parameter rather than two. The corresponding LSD repeat-buying formulae tend therefore to be more convenient than the NBD ones. The LSD formulae can often be further simplified, in terms of approximate expressions in the average frequency of purchase per buyer w**.

The crucial feature of the LSD is that it concerns only the distribution of *buyers*: non-buyers can be dealt with separately, as is shown in §8.2. The Logarithmic Series Distribution itself is described in §8.3, and a simple formula for the *cumulative* distribution in §8.4. The LSD formula together with simplifying approximations for penetration growth and for period-to-period repeat-buying are developed in §§ 8.5 and 8.6. An underlying stochastic model is outlined in § 8.7.

8.2. The Number of Non-Buyers

One possible explanation of the "variance discrepancy" discussed in §7.8 – i.e. the partial failure of the NBD to fit in certain cases – might have lain with the number of *non-buyers* ***. One early question was therefore whether the *non-zero* part of an observed buying distribution subject to the variance discrepancy might not be successfully fitted by a

* This chapter describes the more mathematical parts of the LSD theory.

** These simplifications are important conceptually and for ***ad hoc*** applications; for routine use, calculations tend to be computerised and the more general NBD formulae are to be preferred.

*** This number is determined by the definition of the population of potential buyers, which is often partly arbitrary. For example, in analysing petrol buying, should the "population" be all adults, or all drivers (including wives, friends and children and – for the wife's car – husbands), or car ***owners*** and the like, or people who have ***ever*** bought petrol, or who ***might*** ever do so, or what?

zero-truncated NBD. If successful, this would then lead to a new "estimated" (i.e. artificial) number of zeros and give a good fit to the *full* NBD.

Table 8.1. **NBD's** with Varying Numbers of Notional "Non-Buyers", fitted to the Observed **Purchases** for a **2,000-household Sample (Table** 4.1)

(388 buyers making 1272 purchases)

	Size of notional "sample"						
	750	1,000	1,500	2,000 *	5,000	10,000	20,000
Observed buyers	388 *	388 *	388 *	388 *	388 *	388 *	388 *
Notional "non-buyers"	362	612	1,112	1,612 *	4,612	9,612	19,612
Proportion of buyers, b	0.51	0.39	0.26	0.19 *	0.07	0.04	0.02
Proportion of "non-buyers", $1-b$	0.49	0.61	0.74	0.81 *	0.43	0.96	0.98
Av. no. of purchases:							
– per buyer, w	3.3	3.3	3.3	3.3 *	3.3	3.3	3.3
– per "informant", m	1.7	1.3	0.85	0.63 *	0.25	0.13	0.06
Once-only buyers:							
"Observed"	164 *	164 *	164 *	164 *	164 *	164 *	164 *
"Theoretical"	136	145	153	156	162	164	165
Standard deviations:							
"Observed" s	3.2	2.9	2.4	2.1 *	1.4	1.0	0.7
"Theoretical" σ	2.8	2.6	2.3	2.0	1.4	1.0	0.7
k	0.48	0.29	0.16	0.11	0.04	0.02	0.01
$m/k = a$	3.5	4.3	5.1	5.5	6.2	6.5	6.6

* Directly observed.

Investigation of this possibility [Chatfield et al. 1966] quickly led to quite the contrary result. Neither the cause nor the cure of the variance discrepancy lay with the observed number of non-buyers. Furthermore, it was found that the particular value of the proportion of non-buyers was often altogether irrelevant to the fit of the NBD, i.e. even in those cases where the fit of the full distribution was good.

Thus for any distribution of purchases which is fitted well by an NBD, it is found that an NBD will still give a good fit if the number of zeros (i.e. non-buyers) is arbitrarily increased. An illustration is given in Table 8.1. This is based on the observed data in Table 4.1 of Chapter 4 (reproduced in full in Table 8.2). In Table 8.1 the number of zeros and hence the total sample size are varied from a total "sample" of *750* up to *20,000,* but the actual number of *buyers* is kept fixed at the observed level of 388.

For numbers of "non-buyers" which are smaller than the observed 1612, the fit of the NBD progressively worsens. This is illustrated in Table 8.1 by the increasing discrepancies between such indices as the "observed" and "theoretical" standard deviations (as measures of overall fit), and the "observed" and "theoretical" numbers of once-only buyers (which for the sample of 750 are quite marked). For *increasing* numbers of zeros however, these indices show that the fit of the NBD's remains good.

For increasing numbers of zeros, the parameters m and k of the fitted NBD decrease, but in such a way that their ratio m/k tends towards a steady value, as is implied by the last line of Table 8.1 (where m/k tends towards a limiting value of about *6.7).* This reflects a general mathematical theorem [Fisher et al. 1943] that as the parameters k and m of an NBD are made to tend to zero in such a way that their ratio m/k – usually denoted by the quantity a – tends towards a non-zero limit, the non-zero part of the NBD tends towards the so-called logarithmic series distribution or LSD. This means for practical purposes that the non-zero part of an observed NBD can be approximated to by a logarithmic series distribution, if the NBD parameter k is low enough.

The strict condition for the LSD to give the same results as the non-zero part of an NBD is that the NBD parameter $k = 0$. In practice, the value of k for individual brands or pack-sizes in consumer purchasing data are often about 0.1 or less, and this is "low enough" for the LSD to be quite a good approximation to the NBD. (For heavily aggregated data, such as total product-field purchases, the values of k are often substantially higher and the LSD will differ from the zero-truncated NBD and *not* give such a good fit.) The degree of numerical agreement between the LSD and NBD results in fact depends on both the NBD parameters or, in other words, on both b and w. In practice, a good LSD fit is obtained for data where b, the proportion of buyers, is less than about .2 for any w greater than about 1.5, and for $b < .4$ as long as $w \geqslant 4$. Table 8.2 illustrates the goodness of fit of both the LSD and the NBD to the earlier data of Table 4.1, where k was .1 1. (This

Table 8.2. **An Example of the Fit of the LSD and NBD to Observed Data**

(26-week data for a 2,000 household sample, from Table 4.1)

Number of Purchases	The Number of Buyers		
	LSD	Observed	NBD
0	(1,612)	1,612	(1,612)*
1	165	164	157
2	72	71	74
3	42	47	44
4	27	28	29
5	19	17	20
6	14	12	15
7	10	12	11
8	8	3	8
9	6	7	6
10	4	6	5
11	3	3	4
12	3	3	3
13	2	5	2
14	2	0	2
15	2	0	1
16	1	0	1
17	1	2	1
18	1	0	1
19	1	0	1
20	1	1	1
21+	4	5**	2

Proportion of non-buyers: $p_0 = .806$*
Proportion of buyers: $b = 1 - p_0 = .194$

Av. no. of purchases *per household:* $m = .636$*
Av. no. of purchases *per buyer:* $w = 3.3$ ***

Standard deviations of the non-zero distribution
Observed: = 3.8
Truncated NBD: = 3.6
LSD: = 3.8

* Used in fitting the NBD.
** Actual values 22, 22, 25, 26, 26.
*** Used in fitting the LSD.

was the earliest published data for the NBD, and the fact that the LSD gives if anything a better fit than the NBD itself is of no general significance.)

It follows that much consumer purchasing data which can be successfully described by the two-parameter NBD can *also* be described by the one-parameter LSD, together with a quite separate parameter, namely (1-b), the proportion of the population who are non-buyers. The latter does not directly enter into the LSD itself, whereas in the NBD model, the proportion of non-buyers (1 $-b$) and the mean m operate in an interrelated sort of way and lead to relatively complex formulae involving both statistics (e.g. $b = 1 - (1+a)^{-k}$, where $a = m/k$). The LSD formulation therefore gives scope for considerable simplification. The general properties of the LSD have been discussed by Anscombe [1950], Patil [1962] and Patil et al. [1964]. Patil and Bildikar [1964] have also discussed the mathematical background of stochastic models tending to logarithmic series distributions.

8.3. The **Logarithmic Series Distribution**

The LSD concerns the frequency distribution of *purchases,* i.e. how many buyers make 1, 2, 3 etc. purchases, excluding the non-buyers or zeros (or dealing with them separately).

The probabilities of the logarithmic series distribution are best expressed in terms of a certain parameter q. Thus the LSD is a particular statistical distribution of the probabilities p'_r of making r purchases in the given time-period (for $r \geqslant 1$), where

$$p'_r = \frac{-q^r}{r \ln (1-q)} .$$

(The logarithm "ln" is to base e – a table is given in Appendix B.)*

For $r \geqslant 1$,

$$\sum p'_r = 1, \quad \text{since} \quad \sum \frac{q^r}{r} = -\ln (1-q) .$$

The probabilities p'_r are therefore essentially the series of terms in the expansion of the logarithmic expression $-\ln (1-q)$ in powers of q, i.e.

* In terms of the proportion p_r of all *consumers* who make r purchases, we have $p_r = bp'_r$ for $r \geqslant 1$, and $p_0 = 1-b$.

$q, q^2/2, q^3/3$ *etc*. This derivation has no direct physical meaning here, but the name of the distribution stems from it.

The mean of the LSD is w, the average number of purchases per buyer. This is given in terms of the LSD parameter q by

$$w = \frac{-q}{(1-q)\ln(1-q)}.$$

In terms of the mean m of the full frequency distribution including buyers (i.e. the per capita type of measure of sales) and the proportion of buyers b, *we* have

$$w = m/b .$$

As far as the LSD theory is concerned, w operates independently of the proportion of buyers b in the population as a whole *.

The average rate of purchasing per buyer is clearly a meaningful measure of consumer behaviour in its own right. Furthermore, various early studies of the NBD had also suggested that it was a useful statistic in technical terms; thus within certain ranges of the parameters, it was found that many NBD formulae depend not so much on the varying combination of the two statistics m and b as simply on their ratio $m/b = w$ [e.g. Brace 1959, Ehrenberg 1963]. The LSD now brings this out explicitly. In addition, w has subsequently been found to be relatively constant for different brands or pack-sizes (as is discussed in Chapter 10 and was illustrated in Table 3.2).

To carry out calculations in the LSD theory (such as calculating the values of p'_r, or the repeat-buying formulae later), we need the value of q for the observed data. This is usually best calculated from the observed value w. The above equation between w and q cannot however be solved explicitly for q in terms of a given value of w, but *tables* can be constructed for reading off values of q from given values of w [e.g. Patil 1942, Kamat and Wani 1964, Williamson and Bretherton 1964]. A simple example was given as Table 2.2 in Chapter 2, and a more extensive one is given in Table B. 1 in Appendix B.

The values of q range from 0 (for w = 1) to 1 (for infinite w), but change very slowly relative to w for values of w which are at all large,

* This applies to the LSD formula&for any given brand, and is not to be confused with the way in which w and b may vary together empirically from one brand to another (see § 3.2 in Chapter 3 and §§ 10.2 and 11.5).

so that care must be taken over rounding-off and interpolation. This difficulty can largely be overcome by a reformulation of the parameter q as $q/(1-q)$, which is also more generally useful in many of the LSD calculations. This expression $q/(1-q)$ is usually denoted by a*. Thus

$$a = \frac{q}{(1-q)} \quad \text{or} \quad q = \frac{a}{1+a},$$

and w can hence be expressed in terms of a as

$$w = \frac{a}{\ln(1+a)}.$$

The latter equation still cannot be solved for a directly in terms of w, but values of a can be read off from a suitable table (such as Table B.3 in Appendix B) **. The equation $w = a/\ln(1+a)$ is more nearly linear between w and a than the corresponding relation between w and q and is therefore easier to use, e.g. in interpolating [Chatfield 1969]. (Thus for any $w < 20$, one can calculate a to the same number of significant figures, for a desired level of accuracy; this is not true for q, which varies very little for higher values of w and therefore requires more significant places then.)

Additional simplification arises because numerical analysis has shown that a direct algebraic approximation exists within the parameter range of about $2 < w < 20$ or so (which contains much of the consumer purchasing data that has so far been analysed). Thus a can be expressed to a very close degree of approximation (i.e. to within about 2%) as the following *explicit* function of w,

$$a \doteq 2.46(w-1)^{1.23},$$

and q is similarly given in terms of w as

$$q \doteq (w-1.4)/(w-1.15).$$

The usefulness of the parameter a is however not only that it makes certain formulae and arithmetical operations simpler than when using q,

* This is equivalent to the limit of the NBD ratio m/k *in* §*8.2.*

** As G.J. Goodhardt has noted, the same table can be used for reading off a from w as was used for calculating the NBD parameter a from the observed value of the ratio $m/\ln p_0$ (see §4.2 in Chapter 4).

but more importantly, that it relates, as already mentioned, to the NBD parameter a ($a = m/k$). Indeed, the LSD a is numerically equal to the NBD a to just the extent to which the LSD and NBD give equivalent results anyway. (They are strictly the same only when $k \equiv 0$.) In general, the LSD a has the basic property that in a time-period of relative length T, the corresponding value a_T is directly proportional to a, i.e.

$$a_T = Ta,$$

just as occurred in the NBD theory (§7.5 of Chapter 7). From this simple property, many repeat-buying results can be readily deduced, as is shown in §§8.5 and 8.6 below.

Since the LSD is a one-parameter distribution, its variance, σ^2 say, depends on its mean w, but the relationship cannot be written out explicitly. In terms of q, the variance σ^2 of the LSD is given by

$$\sigma^2 = \frac{-q\{1+q/\ln(1-q)\}}{(1-q)^2\ln(1-q)}$$

which simplifies somewhat on using *two* versions of the parameter, such as q and w, or a and w,

$$\sigma^2 = \{w/(1-q)\} - w^2 = w(1+a-w).$$

This of course refers to the variance of the distribution of *buyers*. The corresponding NBD formula – for the variance of the zero-truncated distribution – is

$$w(1+a_N+m-w),$$

where a_N now represents the value of the parameter a in fitting the NBD (which will differ – at least slightly – from the LSD value of a for the same data).

The ratio of the LSD and NBD standard deviation of zero-truncated distributions for different parameters b and w indicates the degree to which the two distributions give the same results. Illustrative values of the ratio $\sqrt{\{(1+a-w)/(1+a_N+m-w)\}}$ are given in Table 8.3 for various values of b and w [*see* also Chatfield 1970]. As is to be expected from the discussion in §8.2, the degree of agreement depends mainly on the value of b. When b is small, the agreement is very close, but when b reaches .4 or so, the two standard deviations begin to differ by 15% or more.

Table 8.3. The Ratio of the LSD to the Zero-Truncated NBD Standard Deviation

(For various possible values of b and w) *

$\frac{\sqrt{(1+a-w)}}{\sqrt{(1+a_N+m-w)}}$	Proportion of Buyers, b						
	.01	.05	.1	.2	.4	.6	.a
Av. Purchases per Buyer, w							
1.1	1.00	1.00					
1.3	1.00	1.01	1.03	1.1			
1.5	1.00	1.02	1.03	1.1	1.2		
2.0	1.00	1.02	1.03	1.1	1.2	1.3	
5.0	1.00	1.01	1.03	1.1	1.1	1.3	1.5
10.0	1.00	1.01	1.03	1.1	1.1	1.3	1.5
20.0	1.00	1.01	1.03	1.1	1.1	1.3	1.5
Average	1.00	1.01	1.03	1.1	1.1	1.3	1.5

* Data for which $w < -\ln(1-b)/b$ cannot be fitted by an NBD.

A major simplification to the variance formula can also be achieved by numerical approximation, using the "displaced" coefficient of variation $\sigma/(w-1)$, i.e. the standard deviation of the distribution of purchases relative to (w-l) rather than its mean w. This expression is found to vary only very slowly with w. Indeed, in the typical range of w for most purchasing data, namely $1.5 < w < 20$, we have that the quantity $\sigma/(w-1)$ is virtually constant at about 1.7 or so,

$$\frac{\sigma}{(w-1)} \doteq 1.7 ,$$

even though σ^2 itself varies then from about 1 to 400.

8.4. The Importance of Heavy Buyers

A particularly simple formula in the LSD theory is that for the cumulative proportion of total sales (or total purchases) accounted for by heavier buyers. Thus as already noted in Chapters 2 and 4, and illustrated in Tables 2.1 and 3.5, the proportion of total purchases of an item accounted for by those buyers who buy it more than r times in the analysis-period is a very regular quantity which in LSD theory is simply

$$q^r.$$

The derivation is as follows. If a proportion p_i of consumers make i purchases, then the purchases on a *per informant* basis by buyers who buy more than r times is $\sum ip_i$, summed over $i > r$. Next, the total purchases of the item by *all* consumers, again on a *per informant* basis, are of course m, the mean of the full distribution, which also equals bw in terms of w, the mean of the distribution of buyers. The proportion of total purchases accounted for by buyers making more than r purchases can therefore be written algebraically as

$$\frac{1}{m}\sum_{i>r} ip_i \quad , \text{ or } \quad \frac{1}{bw}\sum_{i>r} ip_i \quad , \text{ or } \quad \frac{1}{w}\sum_{i>r} ip_i' \, ,$$

in terms of the proportion $p_i' = p_i/(1-p_0)$ of buyers making i purchases.

In the NBD theory the sum $\sum ip_i$ for $i > r$ can only be obtained *numerically*, by calculating all the individual NBD probabilities p_i for $i = 1$ up to $i = r$, multiplying each by its value of i, summing, and taking the sum away from m. For the LSD however, the corresponding expression simplifies algebraically to q^r, a very much simpler result *.

Thus,

$$\frac{1}{w}\sum_{i>r} ip_i' = \frac{-1}{w\ln(1-q)}\sum_{i>r} iq^i/i \quad , \text{ since } p_i' = -q^i/i \ln(1-q)$$

$$\sum \frac{(1-q)}{q} q^{r+1}{}_{i\geq 0} \, q^i \, , \quad \text{since} \quad w = -q/(1-q)\ln(1-q)$$

$$= q^r \, , \quad \text{since} \quad \sum q^i = (1-q)^{-1} \text{ when } i \geq 0 \, .$$

For an item bought at least 3 times per average buyer in a given period, we have w = 3 and $q = .85$ (from Table 2.2). In the LSD theory, the proportion of total sales accounted for by people who buy more than once (viz. $r = 1$) would therefore be estimated at .85 or 85%, the proportion accounted for by people who buy more than twice is $.85^2 = .72$ or 72%, more than 4 times say is about 50% more than 8 times about 25% and so on.

In terms of having an explicit formula for this particular aspect of buyer behaviour, the LSD theory therefore differs markedly from the NBD, and having an explicit formula is important for practical work.

* For r = 0, $q^0 = 1$, i.e. buyers buying more often than zero times (i.e. at least once) naturally account for 100% of sales.

However, neither the LSD nor the NBD yields a simple formula for such things as how *many* buyers make more than r purchases: in either model, this has still to be computed arithmetically in each specific case.

8.5. Varying Lengths of Time-Periods

Having discussed the LSD for analysing purchasing behaviour in any *single* time-period, we now start to consider a stochastic LSD model which interrelates stationary purchasing behaviour in time-periods of different lengths. The "sample base" of the LSD – i.e. the total number of buyers in the period, as measured by b – will vary with the length of the period, as do also the descriptive statistics of the LSD such as its mean w, its variance σ^2 and its parameter q. None of these statistics are however directly proportional to the length T of the time-period in question.

To start from a simpler point, we therefore note that in the stationary NBD model, the mean m_T in a period T times as long as some "unit" period with mean m is simply proportional to T, i.e. $m_T = Tm$. The NBD parameter $a = m/k$ also varies directly with T, i.e. $a_T = Ta$, since the exponent k remains constant.

The only parameter in the LSD model with comparable properties is the LSD version of a, which was defined in § 8.3 as $a = q/(1-q)$. This has the same property as the NBD a of being directly proportional to T, i.e. $a_T = Ta$. (This can be seen by using the Fisherian derivation of the LSD from the NBD model, letting k and m tend to zero whilst $m/k = a$ tends to a non-zero limit.) Because of this simple proportionality property, the parameter $a = q/(1-q)$ is a particularly convenient variant of the basic LSD parameter q*.

From the equation $a_T = Ta$ it follows that the other versions of the LSD parameter, viz. q_T and w_T, vary with T as follows:

$$q_T = \frac{Tq}{1+(T-1)q}, \qquad w_T = \frac{Ta}{\ln(1+Ta)}.$$

*One other "invariant" result is the "displaced" coefficient of variation $\sigma/(w-1)$ introduced in §8.3. Thus whilst the variance of LSD in different length time-periods varies in an apparently complex manner, the approximate result that $\sigma/(w-1) \doteqdot 1.7$ means that this expression is independent of T, i.e. $\sigma_T/(w_T-1) \doteqdot 1.7$, for any T. However, no further analytic use of this finding has yet been made.

For the means w_T and w in different time-periods, we therefore have

$$\frac{w_T}{w} = \frac{T\ln(1+a)}{\ln(1+Ta)} .$$

The right-hand side of this expression itself depends on w since $w = a/\ln(1+a)$, but it varies rather slowly. Indeed, by using the ratio of the "displaced" means $(w_T - 1)$ and (w-l) instead of w_T/w, the relationship between w_T in time T and w in "unit" time can be reduced to a virtual constant. Thus for $1.5 < w < 20$ – the range of w mainly found in consumer purchasing data – we have by numerical analysis that

$$\frac{w_T - 1}{w - 1} \doteqdot T^{0.82} .$$

This ratio is now independent of a and is simply a function of T. For $T = 2$, $(w_2 - 1)/(w-1)$ is 1.76, and values of $T^{0.82}$ for some other commonly occurring values of T were given in Table 4.13.

The varying proportions of buyers b_T in time-periods of varying length T can also be dealt with within the framework of the LSD model, even though the incidence of buyers (or, better perhaps, the incidence of *non-buyers)* is, of course, outside the LSD as such. But for any given population of consumers, if the proportion of non-buyers (1-b) in "unit" time-period is known, this tells us about the corresponding proportion $(1 - b_T)$ in time-period T.

We know (by definition) that under stationary conditions the mean amount m_T bought per *informant* varies directly with T, i.e. as $m_T = Tm$. Since the proportion of buyers is given by $b_T \equiv m_T/w_T$,

$$b_T = \frac{Tm}{w_T} = \frac{Tw}{w_T} b .$$

So

$$\frac{b_T}{b} = \frac{Tw}{wT} = \frac{\ln(1+Ta)}{\ln(1+a)} ,$$

$$\doteqdot \frac{Tw}{1+(w-1)T^{0.82}} ,$$

using the w_T/w results and the $(w_T - 1)$ form of approximation above.

8.6. Repeat-Buying in Two Equal Periods

Applying the results for b_T to the special case of a period T twice the length of a "unit" period with b buyers, and noting that under stationary conditions $b_2 = 2b - b_R$, where b_R is the proportion of the population buying in both "unit" periods, we obtain the LSD repeat-buying formulae already set out in Chapters 2 and 4. Thus in terms of the LSD a,

$$\frac{b_R}{b} = 2 - \frac{\ln(1+2a)}{\ln(1+a)},$$

or in terms of q,

$$\frac{b_R}{b} = 1 + \frac{\ln(1+q)}{\ln(1-q)},$$

$$= \frac{\ln(1-q^2)}{\ln(1-q)},$$

or, by numerical approximation for $w > 2$,

$$\doteqdot 2(w-1)/(2.3w-1).$$

These expressions also determine the proportion of buyers in the first period who "lapse" in the second period (or conversely, the proportion of "new" buyers), in that under stationary conditions

$$\frac{b_L - b_N}{b} = 1 - \frac{b_R}{b} \doteqdot \frac{\ln(1+q)}{\ln(1-q)} \doteqdot \frac{\ln(1+2a)}{\ln(1+a)}.$$

The numerical values of these LSD expressions (which depend only on w through a or q) differ very little from the NBD ones (which depend on b as well as on w or $m = bw$), other than for low values of w or high values of b. This is illustrated in Table 8.4 for the proportion of repeat-buyers.

LSD expressions for various *rates* of buying can be derived using the Fisherian derivation of the LSD from the NBD. We note that in §7.6 of Chapter 7, m_R, the average number of purchases by repeat-buyers expressed on a *per informant* basis, was given in the NBD theory as

$$m_R = m\{1-(1+a)^{-k-1}\}.$$

Table 8.4. The Ratio of the LSD to the NBD Formulae for the Incidence of Repeat-Buyers, b_R/b

(For various possible values of b and w) *

$\frac{\ln(1-q^2)/\ln(1-q)}{\{1-2(1+a)^{-k}+(1+2a)^{-k}\}/\{1-(1+a)^{-k}\}}$	Proportion of Buyers, b						
	.01	.05	.1	.2	.4	.6	.8
Av. Purchases per Buyer, w							
1.1	1.00	.96					
1.3	1.00	.98	.96	.92			
1.5	1.00	.98	.97	.93	.9		
2.0	1.00	.99	.98	.95	.9	.8	
5.0	1.00	.99	.99	.97	.9	.9	.8
10.0	1.00	1.00	.99	.98	1.0	.9	.9
20.0	1.00	1.00	.99	.98	1.0	.9	.9
Average	1.00	.99	.98	.96	.9	.9	.9

* Data for which $w < -\ln(1-b)/b$ cannot be fitted by an NBD.
Bold figures denote deviations greater than 10%.

We therefore have first of all that the proportion of total sales accounted for by repeat-buyers is given by

$$\frac{m_R}{m} = 1 - (1+a)^{-k-1},$$

$$= 1 - \frac{1}{(1+a)^{k+1}}.$$

As $k \to 0$,

$$\frac{m_R}{m} \to 1 - \frac{1}{(1+a)},$$

$$= \frac{a}{1+a}$$

$$= q.$$

This is the LSD result already used in §2.3 of Chapter 2, and is very much simpler than the corresponding NBD expression.

Next, to obtain w_R itself, i.e. the average frequency of purchase per repeat-buyer, we use this result, i.e. $m_R = mq$ and the earlier one that $b_R = b \ln(1-q^2)/\ln(1-q)$ and therefore have

$$w_R = \frac{m_R}{b_R} = \frac{mq}{b \ln(1-q^2)/\ln(1-q)} .$$

Since $w = -q/(1-q)\ln(1-q)$, this simplifies to

$$w_R = \frac{-q^2}{(1-q)\ln(1-q^2)} ,$$

or by numerical approximation in the range $1.5 < w < 20$, to

$$\doteqdot 1.23\, w .$$

The rates of purchasing w_L or w_N of lapsed or "new" buyers are given by breaking down total sales bw as

$$bw = b_L w_L + b_R w_R = b_N w_N + b_R w_R ,$$

so that for w_L say, dividing through by b and using the above results for $b_L/b, b_R/b, w_R$ and w in terms of q,

$$\frac{-q}{(1-q)\ln(1-q)} = \frac{\ln(1+q)}{\ln(1-q)} w_L + \frac{\ln(1-q^2)}{\ln(1-q)} \frac{-q^2}{(1-q)\ln(1-q^2)} ,$$

from which

$$w_L = (w_N) = \frac{q}{\ln(1+q)} .$$

The value of the expression $q/\ln(1+q)$ is about 1.3 5 for $w = 2$, and tends to a maximum of $1/\ln 2 = 1.443$ as w increases indefinitely and q tends to 1. For $w > 2$, one can therefore take

$$w_L = w_N \doteqdot 1.4, \text{ an (approximate) constant.}$$

Except for low values of w and high values of b, the numerical values given by the LSD formulae for w_R and w_L or w_N differ little from the NBD ones, as is illustrated in Tables 8.5 and 8.6.

Table 8.5. The ratio of the LSD to the NBD Formulae for the Average Purchase Frequency per Repeat-Buyer, w_R

(For various possible values of b and w)

$\frac{-q^2/(1-q)\ln(1-q^2)}{m\{1-(1+a)^{-k-1}\}/\{1-2(1+a)^{-k}+(1+2a)^{-k}\}}$	Proportion of Buyers, b						
	.01	.05	.1	.2	.4	.6	.8
Av. purchases per buyer, w							
1.1	1.00	1.02					
1.3	1.00	1.01	1.03				
1.5	1.00	1.01	1.02	1.05	1.1		
2.0	1.00	1.01	1.02	1.04	1.1	1.2	
5.0	1.00	1.01	1.01	1.02	1.0	1.1	1.1
10.0	1.00	1.00	1.01	1.02	1.0	1.1	1.1
20.0	1.00	1.00	1.01	1.02	1.0	1.1	1.1
Average	1.00	1.01	1.02	1.03	1.0	1.1	1.1

Bold figure denotes deviation greater than 10%.

Table 8.6. The Numerical **Values** of the LSD and NBD Formulae for the Average Purchase Frequency per "New" or per "Lapsed" Buyer, w_N or w_L

(For various possible values of b and w)

NBD formula: $m/(1+a)^{k+1}$	LSD		NBD, for b =						
	Approx. $q/\ln(1+q)$		.01	.05	.1	.2	.4	.6	.8
Av. purchases per buyer, w									
1.1	1.4	1.1	1.1	1.1					
1.3	1.4	1.2	1.2	1.2	1.2				
1.5	1.4	1.3	1.3	1.3	1.3	1.3	1.4		
2.0	1.4	1.3	1.3	1.3	1.3	1.4	1.4	1.6	
5.0	1.4	1.4	1.4	1.4	1.4	1.4	1.5	1.6	1.8
10.0	1.4	1.4	1.4	1.4	1.4	1.5	1.5	1.6	1.7
20.0	1.4	1.4	1.4	1.4	1.4	1.5	1.5	1.6	1.7
Average	1.4	1.3	1.3	1.3	1.3	1.4	1.5	1.6	1.7

Bold figures differ by more than .1 from 1.4.

Conditional Analysis. A more detailed "conditional" form of analysis of repeat-buying was illustrated in Table 3.10 of Chapter 3 and discussed in § 7.6 of Chapter 7. Here one examines the incidence of repeat-buying by the specific frequency of buying (1, 2, 3, etc. purchases) in the previous period. Examining this in theoretical LSD terms leads

directly back to the NBD theory, rather than being of an LSD form. [The mathematics has been studied rigorously by Patil and Bildikar 1964.]

In general, the NBD theory for two unequal time-periods of "unit" length and length T shows that the distribution of purchases in the second period made by buyers who made r purchases in the first period is an NBD with a mean of $(k+r)\{aT/(1+a)\}$ – where a refers to the "unit" period – and the exponent is $(k+r)$, which is independent of T. Now given that the non-zero part of the NBD in this first period is closely approximated to by an LSD, i.e. for small or zero k, this conditional NBD in the second period reduces approximately to an NBD with mean $r\{aT/(1+a)\}$ and exponent r. This is *not* equivalent to an LSD, since the exponent is not small.

8.7. **An Underlying Stochastic Model**

The properties of the LSD model have so far been obtained from those of the earlier NBD model by the Fisherian limiting process, but they can also be derived directly from an underlying stochastic model as such. As shown by Chatfield, this can be set up along roughly similar lines to the NBD stochastic model which was summarised in § 7.2 of Chapter 7.

Firstly, it is supposed that there is some (unspecified) proportion of "never-buyers", i.e. people in the population who *never* buy the item.

Secondly, as in the NBD model, it is supposed that purchases of any one buyer in successive time-periods follow a Poisson distribution with a certain long-run average, μ say.

Thirdly, it is supposed that the long-run average rates of purchasing μ of different "buyers" in the market follow a *truncated* I'-distribution, i.e., that the frequency of any particular value μ is given by

$$(ce^{-\mu/a}/\mu)\, d\mu\,, \quad \text{for} \quad \delta \leqslant \mu \leqslant \infty.$$

Here δ is some very small number *, c is a constant chosen so that

$$\int_{\delta}^{\infty} (ce^{-\mu/a}/\mu)\, d\mu = 1\,,$$

and a is a parameter of the distribution.

* The quantity δ must, however, be positive since $\int_0^{\infty} (e^{-\mu/a}/\mu)\, d\mu$ *does* not exist. It could refer to a very low rate of purchasing, i.e. less than once in a very long time-period of length $1/\delta$.

The difference from the NBD stochastic model is the postulation of a definite group of never-buyers, who are distinct from the people directly covered by the truncated r-distribution and who *all* have a positive long-run rate of buying $\mu \geqslant 6$. At an "intuitive" level, this postulation seems at least as acceptable as the NBD stochastic model itself, in which the long-run mean rates of *all* members of the population are assumed to follow a full r-distribution from zero to infinity, with the frequency of consumers with a long-run mean rate of purchasing at zero (i.e. the frequency of never-buyers) being itself strictly zero.

It is, of course, not necessary to assume that customers' purchasing behaviour actually follows this stochastic model in the long-run. As for the NBD model, it is only necessary to suppose that in any time-period or periods being analysed, the purchases behave *as if* they were a random sample from the values generated by such a model. The various LSD formulae given earlier can then be deduced from the model.

In particular, the probability p_r of a buyer making r purchases in a given time-period is given by taking the Poisson probability

$$e^{-\mu}\mu^r/r!$$

of a particular consumer with long-run mean μ of buying r units in the time-period and integrating over all "buyers" in the truncated Γ-distribution, i.e.

$$p_r = c \int_0^\infty (e^{-\mu}\mu^r/r!)\,(e^{-\mu/a}/\mu)\,d\mu$$

$$= \{c/r!\,(1+1/a)^r\} \int_0^\infty e^{(1+1/a)\mu}\,\{(1+1/a)\,\mu\}^{r-1} d\{(1+1/a)\mu\}$$

$$\doteq [c/\{r!\,(1+1/a)^r\}]\Gamma(r), \quad \text{for } r \geqslant 1, \text{ since } \delta \text{ is very small,}$$

$$= c/\{(1+1/a)^r r\}$$

$$= cq^r/r$$

$$= qp_{r-1}(r-1)/r, \quad \text{with } q = a/(1+a).$$

If $\Sigma p_r \equiv 1$ for $r \geqslant 1$, we must have

$$p_1 = -q/\ln(1-q),$$

and hence the probability p_r of r purchases being made in a given time-period is

$$p_r = -q^r / r \log(1-q), \quad r \geqslant 1,$$

i.e. the logarithmic series distribution.

Next, for any consumer with long-run mean rate of purchasing μ per "unit" time-period, the (Poisson) probability of making r purchases in a period T times as long as this unit period is

$$e^{-T\mu}(T\mu)^r/r!.$$

For the population of all consumers who in the long-run are buyers, i.e. those who are included in the F-distribution, the frequency p_{rT} of r purchases being made in a period T is therefore

$$p_{rT} = c_T \int_{\delta}^{\infty} \left\{ \frac{e^{-T\mu}(T\mu)^r}{r!} \right\} \left\{ \frac{e^{-\mu/a}}{\mu} \right\} d\mu,$$

which can also be written (by writing $T\mu = \mu$) as

$$p_{rT} = c_T \int_{\delta/T}^{\infty} (e^{-\mu}\mu^r/r!)\,(e^{-\mu/Ta}/\mu)\, d\mu$$

$$\doteqdot -q_T^r / \{r \log(1-q_T)\},$$

i.e. again an LSD, but with parameter $q_T = Ta/(1+Ta)$. The parameters of the LSD's in periods of unit length and length T are therefore related as

$$a_T = Ta, \text{ and } \quad q_T = Tq/\{1+(T-1)q\}.$$

The other LSD formulae given earlier then follow.

8.8. Summary

The LSD model is not an alternative of the NBD, but a special case which can give virtually the same results in a simpler form. This applies

in a certain range of parameter values, which for many purposes is broadly when the penetration b is relatively small, say less than .2, or less than .4 as long as the average purchase frequency per buyer, w, is greater than about 1.5 or 2.

The conceptual importance of the LSD is that it shows how repeat-buying behaviour is largely independent of the precise definition of the population of *potential* buyers (as long as it is large compared with the numbers of *actual* buyers), and that it depends only on a single parameter, q say. Since q is a function of w, the LSD theory brings out very clearly how repeat-buying behaviour depends largely on w, the average frequency of purchase per buyer.

Being a one-parameter distribution, the LSD theory leads to much simpler formulae for penetration growth and repeat-buying than the NBD. Additionally, there are a number of numerical approximations to these formulae which are simpler still.