NOTICE: This is the author's version of a work that was accepted for publication in Australasian Marketing Journal. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in:

Trinh, Giang (2014), "Predicting variation in repertoire size with the NBD model", ***Australasian Marketing Journal,*** **vol. 22, pp. 111-116. DOI: 10.1016/j.ausmj.2014.01.002**

Predicting variation in repertoire size with the NBD model

Abstract

This paper proposes and validates the negative binomial distribution (NBD) to predict the variation in repertoire size (the number of brands purchased by a consumer in a specific time period) within a category. From a managerial perspective, the variation is crucial for brand managers who would like to know the nature and intensity of competition that is facing their brands. Empirical findings across multiple datasets from UK consumer packaged goods demonstrate that the NBD model predicts variations in repertoire size very well in different time periods (1 year, 18 months and 3 years), and different buyer groups (light and heavy category buyers). The paper then suggests a simple method to predict those who are brand exclusively loyal, those who are loyal to a few brands, and those who are diversified to buy multiple brands, based on very little information such as category and brand penetrations, using the parameters of the NBD model.

Keywords: Repertoire size; variation; stochastic process; NBD model.

1. Introduction

Repertoire size (the number of brands purchased by a consumer in a specific time period) is an important concept in marketing (Banelis et al., 2013; Bogomolova and Grudinina, 2011; Colombo and Jiang, 2002; Dawes, 2008; Sharp et al., 2002; Stern and Hammond, 2004). It is well known that consumers do not buy all brands in a given market, rather they only purchase from a smaller relevant set of brands (Martin, 2008; Sharp et al., 2002). Therefore, repertoire size is an important factor in evaluating competition between brands. The larger the repertoire size, the stronger the competition between brands (Colombo and Jiang, 2002; Banelis et al., 2003). While many studies have examined the average repertoire size across different product categories (e.g. Andrews and Currim, 2004; Banelis et al., 2013; Banelis, 2008; Banelis et al., 2005; Banelis et al., 2003; Chrysochou et al., 2011; Chrysochou and Krystallis, 2010; Colombo and Jiang, 2002; Dawes, 2008; Massy et al., 1968; Sharp et al., 2002; Wierenga, 1974), there is a lack of studies that investigate the variation in repertoire size within a category. For example, Banelis et al. (2013) compared the average repertoire size across dozens categories and found that time frame, number of brands available and category purchase rate affect the average repertoire size. The authors also found that the average repertoire size is 2 and this does not vary much across categories. However, despite the difference in category purchase rate, how much consumers vary from the average repertoire size still remains unknown. The variation in repertoire size is important as it allows us to identify how many consumers are exclusive to one brand (e.g. repertoire size of 1 brand); how many consumers are loyal to a few brands (e.g. repertoire size of 2 or 3 brands); and how many consumers are diversified to buy many brands (e.g. repertoire size of 4+ brands). It is not possible to identify these consumer groups if we only look at the

average repertoire size. From a managerial perspective, the variation is crucial for brand managers who would like to know the nature and intensity of competition facing their brands.

This paper therefore proposes a model to predict the variation in repertoire size. The proposed model is a combination of the Poisson distribution and the gamma distribution (otherwise known as the negative binomial distribution or the NBD model). The NBD model is one of the most well known models for count data (Winkelmann, 2008). The NBD model was initially used by Greenwood and Yule (1920) to model accidents and was first applied to marketing science by Ehrenberg (1959) to model consumer purchasing behaviour. Ehrenberg made two assumptions when forming the model: Firstly, purchases of a given consumer in successive time periods follow a Poisson distribution. This implies that the variance of purchases within individual consumers is "as if" random over time (i.e. Poisson process). Secondly, the mean rates of purchasing of different consumers in the long run differ and their distribution is a gamma distribution. In other words, the variance of mean purchases across different consumers is measured by a gamma distribution. Following these assumptions, the frequency of consumers making 0, 1, 2, 3,…x purchases in a given time period can be modelled by the negative binomial distribution.

Since the original article by Ehrenberg (1959), the NBD model has been successfully applied across numerous situations including different brands in different categories, different time periods, and different countries (Chatfield et al., 1966; Ehrenberg, 1988; Ehrenberg, 1996). Recently, the NBD has also been applied to different types of behaviour such as gambling (Mizerski et al., 2004; Lam and Mizerski, 2009) and consumption of mobile phone services (Lee et al., 2011).

As stochastic changes in brand buying behaviour have been well modelled by the NBD model, it is reasonable to assume that a stochastic process also captures changes in repertoire size. Although stochastic changes in repertoire size have not been directly reported in the literature, previous research on actual buying behaviour has shown that heavy buyers have a greater repertoire size than light buyers (Banelis, 2008; Banelis et al., 2003; Colombo and Jiang, 2002). On the other hand, these buyer groups are not stable. Many heavy buyers this year become light buyers next year and vice versa, as a result of the "regression to the mean" effect (Ehrenberg, 1988; Romaniuk and Wight, 2011). Consequently, we would also expect repertoire size to be dynamic both across different consumers and within a given individual consumer.

2. Mathematical expression

With the NBD model of repertoire size we assume that

(1) The repertoire size *x* of a given consumer follows a Poisson distribution with parameter λ.

$$f(x)_{poisson} = \frac{\exp(-\lambda)\lambda^x}{x!} \qquad (1)$$

with mean

$$E[x] = \lambda$$

(2) The mean repertoire size λ of different consumers follows a gamma distribution.

$$f(\lambda;k,a) = \lambda^{k-1}\frac{\exp(-\lambda/a)}{a^k\Gamma(k)} \qquad (2)$$

where *k* and *a* are the shape and scale parameters of the gamma distribution, respectively

with mean

$$E[\lambda] = ka \qquad (3)$$

and variance

$$\text{var}[\lambda] = ka^2 \qquad (4)$$

Combining (2) and (1), the probability density function of repertoire size *x* is

$$\begin{aligned} f(x)_{NBD} &= \int_0^\infty f(x)_{poisson} f(\lambda;k,a)d\lambda \\ &= (1+a)^{-k} \frac{\Gamma(x+k)}{x!\Gamma(k)} \left(\frac{a}{1+a}\right)^x \end{aligned} \qquad (5)$$

with mean

$$E[x] = ka \qquad (6)$$

and variance

$$\text{var}[x] = ka(1+a) \qquad (7)$$

3. Parameter estimation and measure of fit

To estimate the parameters *a* and *k* of the NBD model we use the maximum likelihood procedure proposed by Rungie and Goodhardt (2004).

The likelihood function is

$$L = \sum n_x * \log(p_x) \qquad (8)$$

To measure the fit of the NBD model we compare the observed variance and the theoretical variance (Anscombe, 1950; Ehrenberg, 1959; Chatfield et al., 1966). A small discrepancy between the observed variance and the theoretical variance indicates the model is a good fit.

In addition to variance discrepancy, we use Theil's U coefficient of equality, as it is easy to understand and interpret (Wu and Chen, 2000a; 2000b; Fader and Hardie, 2002). It ranges from 0 to 1, with scores approaching 0 indicating a good fit to the model.

The U coefficient is calculated as.

$$U = \frac{\sqrt{\sum_{x=0}^{n}(A_x - P_x)^2}}{\sqrt{\sum_{x=0}^{n}A_x^{\,2}} + \sqrt{\sum_{x=0}^{n}P_x^{\,2}}} \qquad (9)$$

where A_x and P_x are actual and predicted values, respectively.

4. Data

The information used to fit the NBD model is household consumer data for a three year period from the Taylor Nelson Sofres (TNS) Superpanel database. The panel consists of 35,369 telephone-owning households across the UK. It is drawn only from full-time residents. The sample is demographically and regionally balanced in order to represent the UK population. The data were collected from panel participants twice weekly via electronic terminals in the home, with purchases being recorded using home-scanning technology (TNS, 2012). As such it represents a very large, valid and reliable data source. We analysed four fast moving consumer goods (FMCG) categories: toothpaste, toilet tissue, instant coffee and tea bags. Table 1

shows the descriptive statistics of the four categories. For reason of data stability, we include only brands that are available over the three year period (brands that are withdrawals or new entrants are not included in the analysis), as well as brands that are available in the geographical coverage of the panel (London, Midlands, North East, Yorkshire, Lancashire, South, Scotland, East England, Wales, and South West).

Table 1 Descriptive statistics of the four categories

Category	# of brands	Category purchase rate	Average repertoire size
Toothpaste	41	8.6	2.00
Instant coffee	34	16.2	2.34
Tea bags	59	14.9	2.35
Toilet tissue	40	17.6	2.92

As we can see from table 1, the average repertoire size increases from 2.00 to 2.92 as category purchase rate increases from 8.6 to 17.6. The number of brands available also affects repertoire size. For example, despite a smaller category purchase rate for tea bags than instant coffee (14.9 versus 16.2), the average repertoire size is as large as that of instant coffee (2.35 and 2.34, respectively). This might be explained by the reason that tea bags category has more brands available than instant coffee (59 brands versus 34 brands).

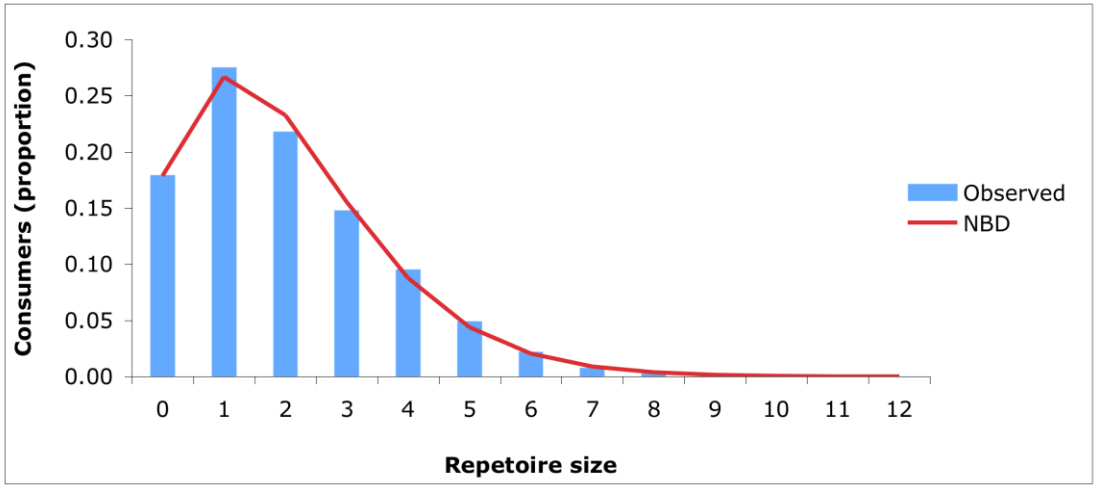
5. Results

5.1. Model fit to repertoire size in four product categories

Figure 1 shows the typical fit of the NBD model to the number of brands bought in the toothpaste category in a three year period. As we can see, the NBD model fits the observed data very well. The good fit is also shown by a very small variance

discrepancy between the predicted and observed values, and a very small Theil's U coefficient of equality (0.02 and 0.02, respectively). The results show that while the average repertoire size is 2, only 22% of consumers have a repertoire size of 2. About 28% of consumers have a repertoire size of 1, and 33% have a repertoire size of 3 or more.

Figure 1 Variation in repertoire size (Toothpaste)



Results for the other categories are shown in figures 2, 3 and 4. Overall, the majority of consumers seem to have noticeably more or less than the average repertoire size (40% have a repertoire size of 3 or more and 20% have a repertoire size of 1). While those consumers who have a repertoire size of 2 occur in many cases, they account for less than 25% of the total consumers. So we see here, the variation is not symmetric at the mean but skewed to the right.

Figure 2 Variation in repertoire size (Instant coffee)

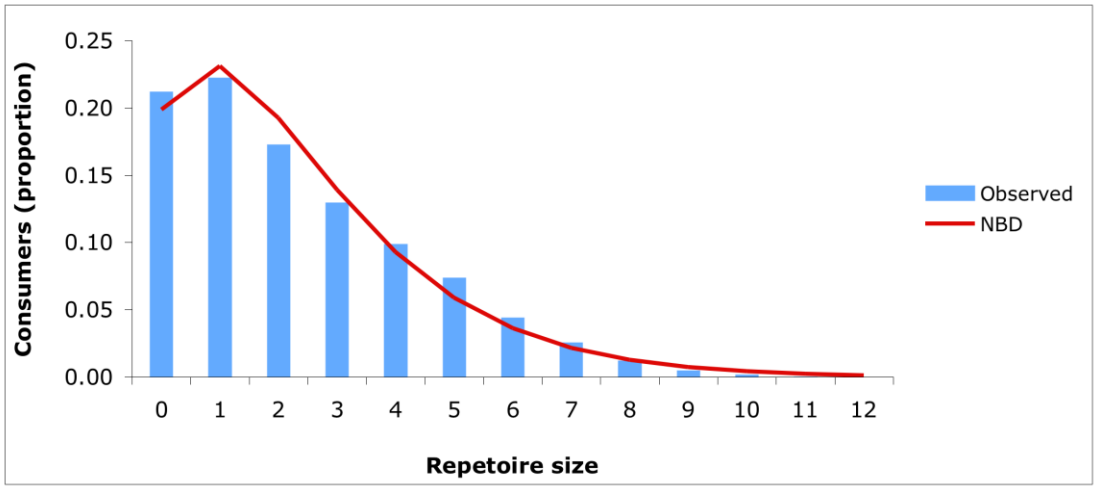


Figure 3 Variation in repertoire size (Toilet tissue)

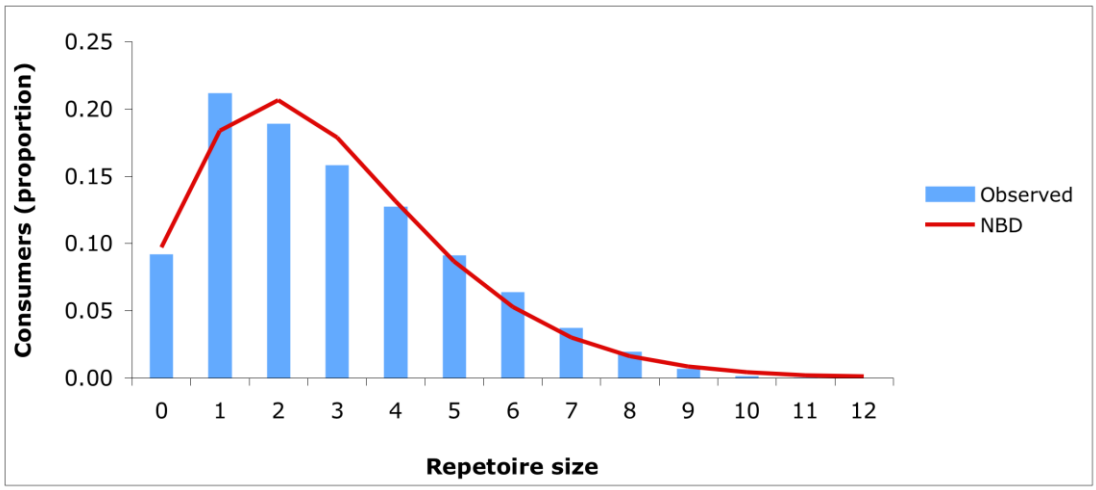


Figure 4 Variation in repertoire size (Tea bags)

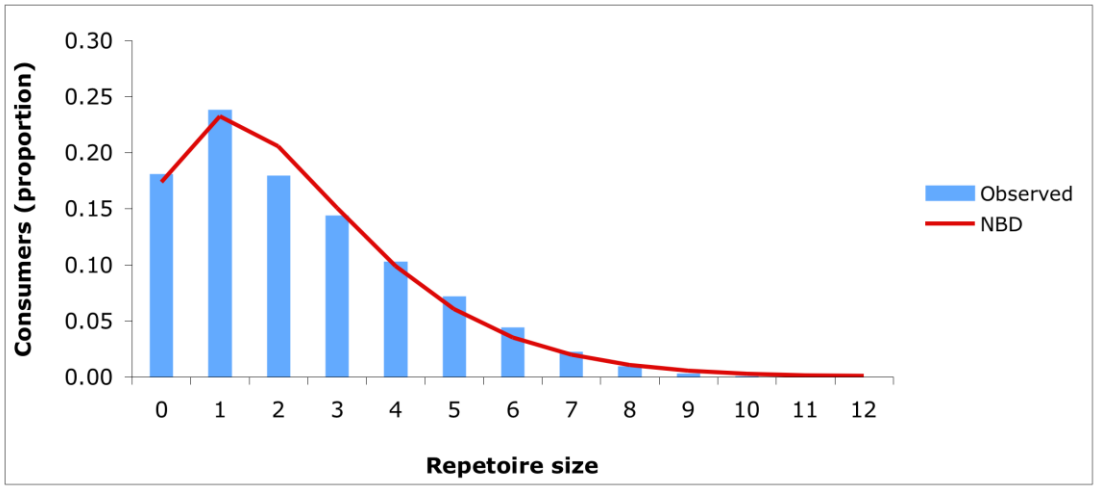


Figure 5 represents the general fit of the model over multiple categories by comparing the observed variance to the predicted variance. Again, the model's predicted variances closely match the observed variances.

Figure 5 Actual and estimated variances of four categories.

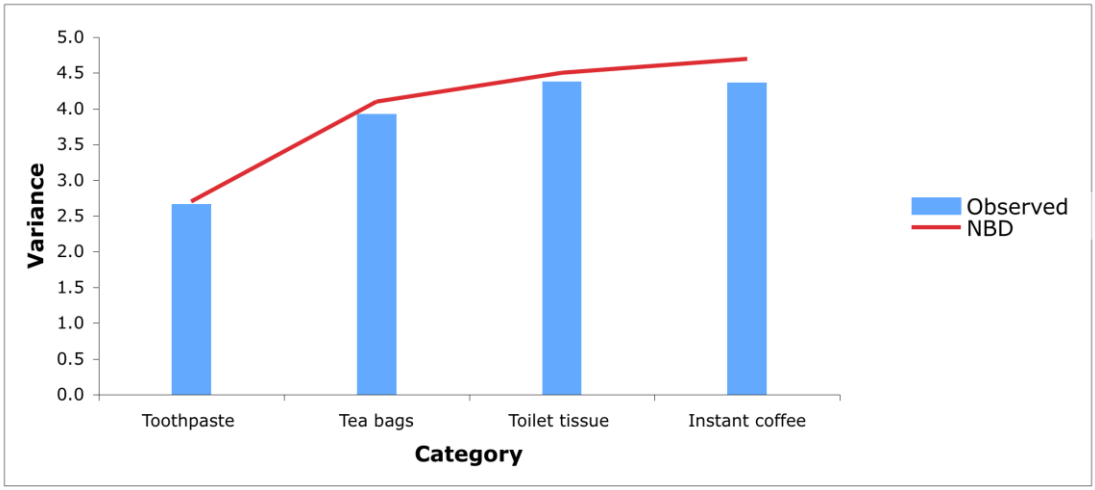


Table 2 shows Theil's U values for four product categories. On average, Theil's U coefficient is 0.04 and ranges from 0.02 to 0.06, which indicates that the NBD model gives accurate predictions.

Table 2 Theil's U coefficient of equality

Category	Theil's U
Toothpaste	0.02
Instant coffee	0.04
Tea bags	0.04
Toilet tissue	0.06

With the NBD model of repertoire sizes, the mean of the distribution is an important measure of loyalty in a particular market. Specifically, the higher the mean,

the lower the level of loyalty in that market. However, it should be noted that the mean of the distribution only partly describes loyalty in a given market. The variance of the distribution should also be considered as it shows how repertoire size varies within and across consumers. For example, tea bags and instant coffee have quite similar means (2.34 and 2.35, respectively) but markedly different variances (3.92 and 4.70, respectively). The higher variance of instant coffee suggests that there is more chance for a consumer to change his/her repertoire size; and repertoire size is more varied across the population in this category. In contrast, the lower variance of tea indicates that a consumer tends to stick to his/her repertoire size; and repertoire size is less varied across the population in this category. Analysis of variance is therefore crucial for marketers who wish to choose a particular market to promote their brand.

5.2. Variation in repertoire size across light and heavy category buyers

Previous research has noted that repertoire size increases when category purchase rate increases (Banelis et al., 2013). This suggests that heavy category buyers have a larger repertoire size than light category buyers. However, how repertoire size varies within each buyer group is not known. Does repertoire size vary more among heavy or light category buyers? Answering this question might shed light on choosing a category buyer class to promote a new brand.

Figures 6 and 7 show examples of the fit of the NBD model to repertoire size for light category buyers (those who purchase less than the average category purchase rate) and heavy category buyers (those who purchase more than the average category purchase rate) in the toilet tissue category. As we can see, the model fits each category buyer class very well. Consistent with previous findings in observed data, the NBD model captures the phenomenon that heavy category buyers have a larger repertoire size than light category buyers, on average. However, if we look at the

disaggregate level, not all the heavy category buyers have a larger repertoire size than the light category buyers. Many heavy category buyers have a small repertoire size (e.g. 1, 2, 3 brands) whereas many light category buyers have a large repertoire size. These results suggest that, while the average repertoire size is a good measure at the aggregate level, individuals do behave differently from each other, hence the use of a model such as the NBD that captures the variation across different consumers will provide a better representation of consumer behavior.

Figure 6. Variation in repertoire size (light category buyers – toilet tissue)

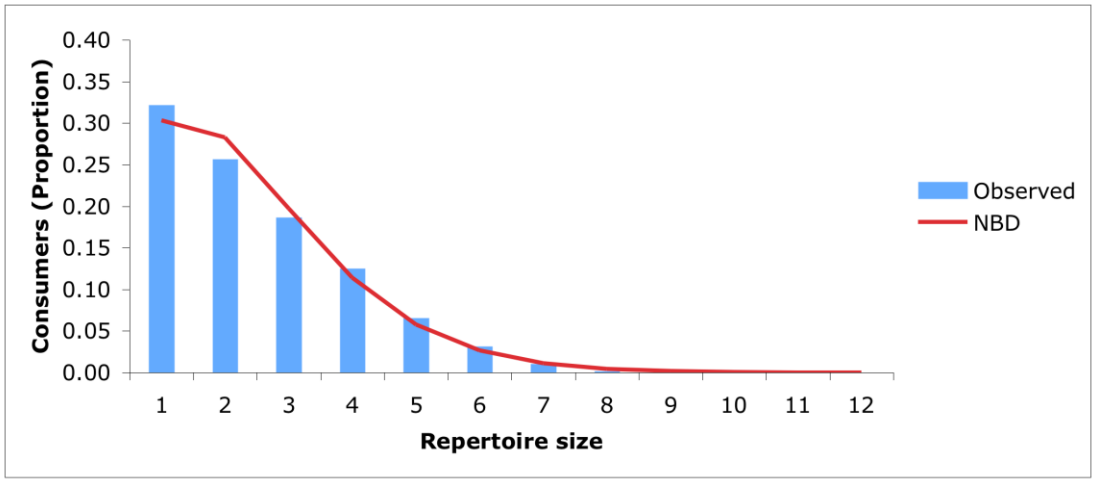
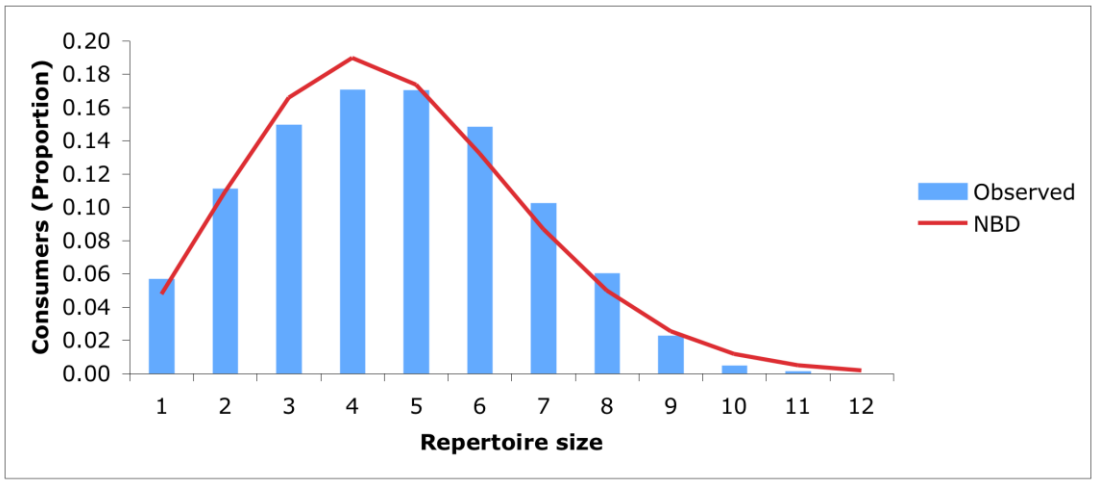


Figure 7. Variation in repertoire size (heavy category buyers – toilet tissue)



Furthermore, the NBD model accurately predicts the variance of repertoire size within each buyer group. Table 3 shows that heavy category buyers have considerably larger variations in repertoire size than light category buyers. This suggests that heavy category buyers might be a potential group to promote for trials of a new brand.

Table 3 Estimated variance of repertoire sizes among light and heavy category buyers

Category	Light buyers	Heavy buyers
Toothpaste	2.22	3.81
Tea bags	2.24	4.04
Instant coffee	2.33	4.32
Toilet tissue	2.72	4.60

5.3. Variation in repertoire size across different time frames

Previous research suggested that the longer the time frame the larger the average repertoire size (Banelis et al., 2013). However, whether or not the variance of repertoire size also increases as time increases is not known. We therefore apply the NBD model to predict the variance of repertoire size in different time frames. Figures 8 and 9 show examples of the fit of the NBD model to repertoire size in the instant coffee category for one year and 18 months. Similarly to the fit for three year period data, the model shows very good fit to one year and 18 month data.

Figure 8. Variation in repertoire size (one year instant coffee)

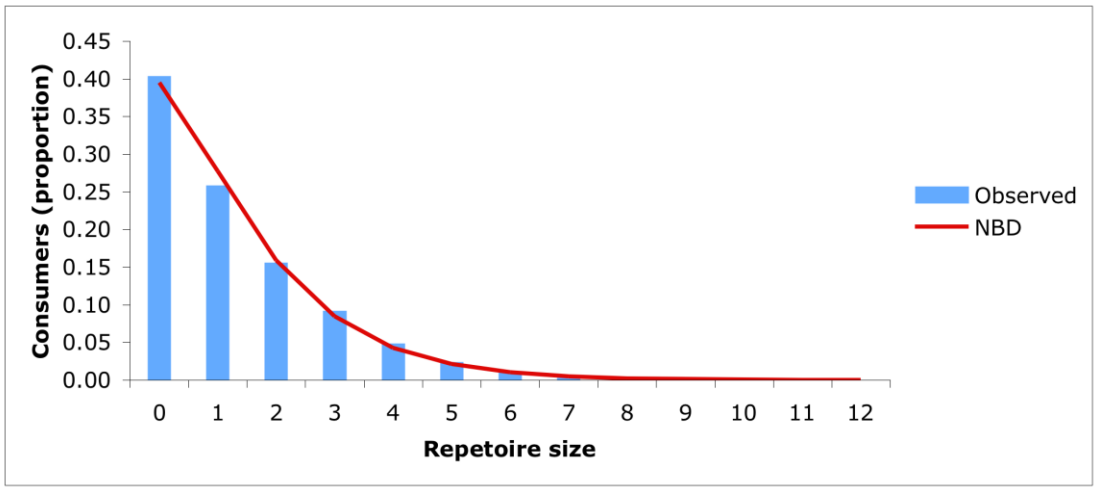


Figure 9. Variation in repertoire size (18 months instant coffee)

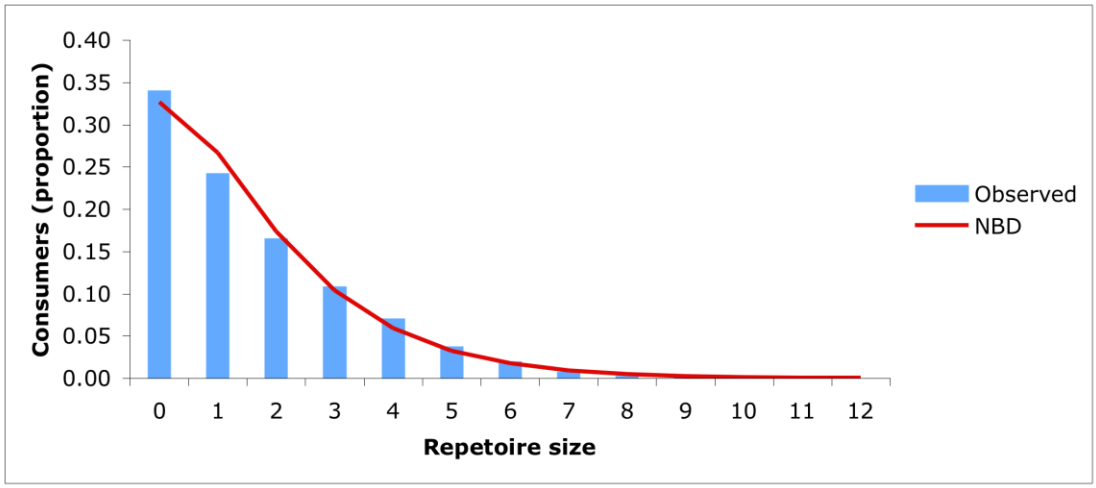


Table 3 reports the variance of repertoire size predicted by the NBD model for three different time periods (1 year, 18 months and 3 years). As we can see from the table, across all four categories, the variances of repertoire size increase as time increases from 1 year to 18 months, then to 3 years. So we see here, consumers not only increase their repertoire size but also vary in their repertoire size more as time increases.

Table 3 Estimated variance of repertoire sizes across different time periods

Category	1 year	18 months	3 years
Toothpaste	1.8	2.3	2.7
Toilet tissue	3.5	4.2	4.5
Instant coffee	2.3	3.1	4.7
Tea bags	2.2	2.9	4.1

6. Conclusion, implication and future research

This paper proposes and validates the NBD model to predict the variation in repertoire size. Empirical findings demonstrate that the model fits the observed data very well in multiple categories, different time frames and different buyer groups. Several practical implications can be drawn from these findings.

First, with very little information such as brand and category penetration, marketers could use the NBD to predict the number of exclusively loyal buyers and the number of multi-brands buyers. This can be used to indicate the intensity of competition that is facing their brands. Following Gerald Goodhardt's proposition, Banelis et al. (2003) showed that the mean repertoire size can be calculated based on brand penetration according to the following equation.

$$E[x] = \sum_{j=1}^{n} b_j \quad (10)$$

where b_j is brand j penetration and n is the number of brands

If information is known about brand penetration and category penetration, the two parameters *a* and *k* of the NBD model can be estimated by the *mean and zero* method (Ehrenberg, 1988) using the following equations:

$$\sum_{j=1}^{n} b_j = E[x] = ka \qquad (11)$$

$$1 - \text{category penetration} = p_{x=0} = \frac{1}{(1+a)^k} \qquad (12)$$

where $p_{x=0}$ is the probability of buyers who have a repertoire size of zero.

Then, we can use the estimated parameters *a* and *k* to predict those who are exclusively loyal buyers ($p_{x=1}$), those who are loyal to a few brands (e.g. $p_{x=2,3}$), and those who are diversified to buy multiple brands (e.g. $p_{x=4+}$) using equation (5).

Previous research has used the NBD-Dirichlet model to predict loyalty metrics such as share of category requirement and exclusively loyal buyers at the brand level (Goodhardt et al., 1984; Ehrenberg et al., 2004). In this paper, we extend the application of the NBD to predict loyalty at the category level. The prediction at the category level is useful for brand managers as it shows the nature of exclusive and diversified loyalty that their brands are facing.

Secondly, brand managers could use the NBD model to predict changes in repertoire size over time using the distribution of changes of the NBD model (Trinh, 2013). For example, if a consumer is exclusively loyal to only one brand in one period, what is the probability that he/she is still exclusively loyal to only one brand in the next period? What is the probability that the consumer is loyal to a few brands or converts to be loyal to multi-brands in the next period? Consequently, the NBD

model could be used as a benchmark to evaluate any real change in repertoire size (as opposed to stochastic change), caused by marketing activity such as sales promotions and advertising.

Lastly, as heavy category buyers show a greater variation in their repertoire size than light category buyers, managers of a new brand might consider this group as a primary target to launch the brand as it will have more chance to be included into their repertoire.

For future research, one could extend the univariate NBD model to a bivariate NBD model to examine differences in heterogeneity in repertoire sizes between different consumer segments (e.g. do older consumers have a smaller repertoire size than younger consumers?); or between different product categories (e.g. if a consumer has a large repertoire size in one category, does she/he also have a large repertoire size in another other category?) The bivariate NBD model can be used to answer these questions. The correlation of the bivariate NBD model will indicate how consistent repertoire sizes are across segments or categories without the results being dominated by differences in the means of repertoire sizes. The higher the correlation, the higher the consistency in repertoire sizes. The second direction of future research is to develop a stochastic model with the NBD component of repertoire size to predict changes in repertoire compositions. There have been a number of attempts to predict changes in consumers purchasing from one period to the next period - how many heavy buyers stay heavy, and how many become light buyers and vice versa, as well as brand switching behaviour (Goodhardt and Ehrenberg, 1967; Ehrenberg, 1988; Trinh et al., 2013). Yet it would be more useful if one could model and predict not only the changes in purchase size and repertoire size, but also the brands that are included or excluded from the consumer's repertoire from one period to another.

Acknowledgement

The author would like to thank an anonymous reviewer for his/her helpful comments and suggestions that lead to a significantly improved article.

References

Andrews, R., Currim, I., 2004. Behavioural differences between consumers attracted to shopping online versus traditional supermarkets: implications for enterprise design and marketing strategy, *International Journal of Internet Marketing and Advertising* **1** (1), 38-61.

Anscombe, F. J., 1950. Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika **36***, 358–382.

Banelis, M., Riebe, E., Rungie, C., 2013. Empirical evidence of repertoire size. *Australasian Marketing Journal* **21** (1) 59–65.

Banelis, M. (2008) *Understanding consumers' repertoire sizes*. PhD thesis. University of South Australia.

Banelis, M., Rungie, C.M., Colombo, R., 2003. Empirical verification of two expected portfolio size formulae. *Australian and New Zealand Marketing Academy Conference Proceedings*. Adelaide, Australia.

Banelis, M., Rungie, C.M., Riebe, E., Meyer-Waarden, L., 2005. Do higher spending households buy a greater variety of brands? : An application of repertoire regression. *Marketing Bulletin* **16**, 1-7.

Bogomolova, S., Grudinina, O., 2011. Under the marketers' radar: Commonly ignored triggers for brand repertoire changes. *Journal of Marketing Management* **27**, 1378-1403.

Chatfield, C., Ehrenberg, A.S.C., Goodhardt, G.J., 1966. Progress on a simplified model of stationary purchasing behaviour. *Journal of the Royal Statistic Society. A* **129** (3), 317-367.

Chrysochou, P., Lockshin, L., Habenschuss, S., Trinh, G., 2011. Does behaviour of heavy and light wine buyers differ?. *6th International Conference of the Academy of Wine Business Research BEM Proceedings*, Bordeaux, France.

Chrysochou, P., Krystallis, A., 2010. Repertoire and frequency of consumption in wine: Are heavy buyers more loyal to product attributes?. *39th European Academy of Marketing Conference Proceedings*, Copenhagen, Denmark.

Colombo, R., Jiang, W., 2002. Portfolio size of brand purchases in Dirichlet markets. *Australia and New Zealand Marketing Academy Conference Proceedings*, Melbourne, Australia.

Dawes, J., 2008. Regularities in buyer behaviour and brand performance: The case of Australian beer. *Journal of Brand Management* **15**, 198-208.

Ehrenberg, A.S.C., 1959. The pattern of consumer purchases. *Applied Statistic* **8**, 26-41.

Ehrenberg, A.S.C., 1988. *Repeat buying*. (2nd ed.). London: Griffin. (reprinted in *J. Empirical Generalisations in Marketing Science* **5**, 392–770, available from http://www.empgens.com/).

Ehrenberg, A.S.C., 1996. Towards an integrated theory of consumer behaviour *International Journal of Market Research* **38** (4), 395-427.

Ehrenberg, A.S.C., Uncles, M.D., & Goodhardt, G.J., 2004. Understanding brand performance measures: using Dirichlet benchmarks. *Journal of Business Research*, **57**, 1307-1325.

Fader, P.S., Hardie, B.G.S., 2002. A note on an integrated model of customer buying behavior, *European Journal of Operational Research* **139**, 682–687.

Goodhardt, G. J. and Ehrenberg, A. S. C., 1967. Conditional trend analysis: A breakdown by initial purchasing level. *Journal of Marketing Research*., **4**, 155-161.

Goodhardt, G.J, Ehrenberg, A.S.C. and Chatfield, C., 1984. The Dirichlet: a comprehensive model of buying behavior. *Journal of Royal Statistical Society Series A*., **147**, 621– 55.

Greenwood, M., Yule, G.U., 1920. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society* **83** (2), 255-279.

Lam, D., Mizerski, R., 2009. An investigation into gambling purchases using the NBD and NBD–Dirichlet models. *Marketing Letters* **20**, 263-276.

Lee, R., Rungie, C., Wright, M., 2011. Regularities in the consumption of a subscription service. *Journal of Product & Brand Management* **20** (3), 182-189.

Massy, W.F., Frank, R.E., Lodahl, T., 1968. *Purchasing behavior and personal attributes*. Philadelphia: University of Pennsylvania Press.

Martin, N., 2008. *Habit: the 95% of behavior marketers ignore*. Upper Saddle River, NJ: Pearson Education.

Mizerski, D., Miller, R., Mizerski, K., Lam, D., 2004. The stochastic nature of purchasing a state's lottery products. *Australasian Marketing Journal* **12** (3), 56-69.

Romaniuk, J., Wight, S., 2011. The stability and sales contribution of heavy buyers. *Working paper*, Ehrenberg-Bass Institute, University of South Australia.

Rungie, C., Goodhardt, G., 2004. Calculation of theoretical brand performance measures from the parameters of the Dirichlet model. *Marketing Bulletin* **15**, 1-19.

Stern, P., Hammond, K., 2004. The relationship between customer loyalty and purchase incidence. *Marketing Letters* **15**, 5-19.

Sharp, B., Wright, M., Goodhardt, G., 2002. Purchase loyalty is polarised into either repertoire or subscription patterns. *Australasian Marketing Journal* **10**, 7-20.

TNS., 2012. http://superpanel.tns-global.com/superpanel/

Trinh, G., 2013. *Modelling changes in buyer purchasing behaviour*. PhD thesis. University of South Australia.

Trinh, G., Rungie, C., Wright, M., Driesener, C. and Dawes, J. 2013. Predicting future purchases with the Poisson lognormal model. *Marketing Letters* http://link.springer.com/article/10.1007%2Fs11002-013-9254-1#

Wierenga, B., 1974. *An investigation of brand choice processes*. Rotterdam: Rotterdam University Press.

Winkelmann, R., 2008. *Econometric analysis of count data*. (5th ed.). Berlin: Springer-Verlag.

Wu, C., Chen, H.-L., 2000a. Counting your customers: Compounding customer's in-store decisions, interpurchase time, and repurchasing behavior. *European Journal of Operational Research* **127** (1), 109–119.

Wu, C., Chen, H.-L., 2000b. A consumer purchasing model with learning and departure behaviour. *The Journal of Operational Research Society* **51** (5), 583-591.