Mark Lett (2017) 28:491–507
https://doi.org/10.1007/s11002-017-9442-5

The Pareto rule for frequently purchased packaged goods: an empirical generalization

Baek Jung Kim[1] · **Vishal Singh**[1] · **Russell S. Winer**[1]

Published online: 18 October 2017

Abstract Many markets have historically been dominated by a small number of best-selling products. The Pareto principle, also known as the 80/20 rule, describes this common pattern of sales concentration. Several papers have provided empirical evidence to explain the Pareto rule, although with limited data. This article provides a comprehensive empirical investigation on the extent to which the Pareto rule holds for mass-produced and distributed brands in the consumer-packaged goods (CPG) industry. We used a rich consumer panel dataset from A.C. Nielsen with 6 years of purchase histories from over 100,000 households. Our analysis utilizes a large number of potential factors such as brand attributes, category attributes, and consumer purchase behavior to explain variation in the Pareto ratio at the brand level across products. Our main conclusion is that the Pareto principle generally holds across a wide variety of CPG categories with the mean Pareto ratio at the brand level across product categories of .73. Several variables related to consumer purchase behavior (e.g., purchase frequency and purchase expenditure) are found to be positively correlated with the Pareto ratio. In addition, niche brands are more likely to have a higher Pareto ratio. Finally, brand/category size, promotion variables, change-of-pace brands, and market competition variables are negatively correlated with the Pareto ratio.

Keywords Pareto rule · Frequently purchased products · Empirical generalization

✉ Baek Jung Kim
bkim2@stern.nyu.edu

Vishal Singh
vsingh@stern.nyu.edu

Russell S. Winer
rwiner@stern.nyu.edu

[1] Marketing Department, Stern School of Business, New York University, 40 W. 4th Street, New York, NY 10012, USA

1 Introduction

The Pareto principle (i.e., the 80/20 rule) explains that, in many events, 80% of consequences come from 20% of the causes. This phenomenon was first observed by Italian economist Vilfredo Pareto in 1906. He observed that 20% of the pea pods in his garden contained 80% of the peas and applied this observation to find that 80% of the land in Italy was owned by 20% of the population.

The Pareto rule also has been studied in marketing in two streams of literature. One stream of literature suggests a theoretical framework to explain the concentration by using a negative binomial distribution (NBD) of purchase frequencies (Morrison and Schmittlein 1981, 1988). Schmittlein et al. (1993) provide some empirical results supporting the NBD explanation for the Pareto rule.

In particular, the authors of these studies argue that this parsimonious model (i.e., NBD) can be used to predict a variety of market statistics such as the distribution of purchase frequencies across households, the average number of purchases per buyer, and the market-penetration level, which all follow the Pareto rule. It also predicts how these quantities will vary depending on the duration of the time period being considered.

Another stream of literature includes empirical studies that look for evidence of the Pareto rule in different product categories by using consumer panel data. A study by Twedt (1964) was the first study in marketing to show how the product category-level sales is concentrated in "heavy users" of the product category. Another study by Schmittlein et al. (1993) investigated the proportion of total category-level sales coming from the heaviest category users (not at the brand level). Other empirical studies (Brynjolfsson and Smith 2000; Brynjolfsson et al. 2003; Anderson 2006; Brynjolfsson et al. 2011) examine the Internet's "long tail" phenomenon, which describes how sales of niche products can grow to take over a larger share of the market than might otherwise have in a purely bricks-and-mortar world.

In contrast to the previous literature, this paper aims to provide robust empirical evidence of the Pareto rule by using consumer panel data over 6 years and investigates the following research questions: (1) To what extent does the Pareto rule (Pareto principle) hold in the consumer-packaged goods (CPG) industry? (2) How does it vary across product categories/brands? (3) Which potential factors (brand/product attributes or consumer purchase behavior) explain the variation in Pareto distributions?

2 Empirical method

2.1 Data description

The data used in this paper, provided by A.C. Nielsen, consists of details of grocery purchases from retail outlets. The data was collected by households on A.C. Nielsen's panel via in-home optical scanners. The data includes the purchase histories of approximately 100,000 demographically balanced households spanning all 50 US states. The range of the data is from 2004 to 2009 and households remained on the panel for an average of 3 years. We observed purchase histories of approximately 18,000 households in 22 product categories for the entire 6-year period.

Specifically, the database contains information on store information, product description (brand name, size, etc.), number of units purchased, price paid, and indicators for any promotion coupon usage. The database also includes a large number of household demographics, such as age, gender, household composition, income, and education.

2.2 Pareto ratio measures

One of the main objectives in this study is to find empirical evidence of the Pareto rule at the brand level, and thus we examine the proportion of total sales that is driven by the top 20% of consumers for each brand. However, first, we need to define a "brand." A.C. Nielsen defines each brand on the basis of flavor, packaging, size, fat content, and so forth within a product category. For example, in the regular soft drinks category, regular (non-diet) Coca-Cola and regular Cherry Coca-Cola are defined as different brands. In this study, our approach was to create a new brand by merging all original brands, regardless of flavor, size, and fat content, instead of using an original definition of the brand as defined by A.C. Nielsen. In other words, we used the "umbrella" or "family" brand name rather than the different line extension brands.[1] For example, in the regular soft drinks category, we considered regular Coca-Cola and regular Cherry Coke as the same brand even though Nielsen considers them to be different. Next, having defined the unit of analysis, we created a measure of the Pareto rule. Since we were interested in the degree to which the proportion of sales from the top 20% of consumers of total sales at the brand level, we used a simple measure of the Pareto rule at the brand level as follows[2]:

$$Pareto\ ratio = \frac{Aggregated\ Dollar\ Sales\ from\ \text{Top}\ 20\%\ of\ Consumers}{Aggregated\ Dollar\ Sales\ from\ Total\ Consumers} \quad (1)$$

This measure indicates what proportion of brand-level dollar sales comes from the top 20% of consumers, and we call this the Pareto ratio (PR).

Rather than using all of the Nielsen data available, we applied two criteria to reduce the observations analyzed. First, we selected the top 22 product categories from the data based on category-level sales ranking. These selected categories include a wide variety of CPG items.[3] Second, since both PR (brand and category) would be positively affected by the inclusion of households who had rarely purchased anything during the observational period, we excluded those

[1] For a robustness check, we conducted the same analyses on the basis of the original definition created by A.C. Nielsen and the results were not significantly changed (please refer to the footnote 6 for details).

[2] As a second robustness check, we conducted analyses using a volume definition of the Pareto ratio in addition to the dollar sales definition. Using a volume definition, the mean of the Pareto ratio is .72 compared to the Pareto ratio measured by dollars (.73).

[3] The list of product categories used in the analyses is the following: cigarettes, carbonated soft drinks, low-calorie soft drinks, toilet tissue, nutritional supplements (vitamin), cookies, ice cream (bulk), canned soup, candy-chocolate, wine, ground and whole bean coffee, yogurt, bottled water, liquid detergent, frozen pizza, potato chips, fruit drinks (canned), light beer, paper towels, orange juice, cheese, and cereal (ready to eat).

households that made purchases less than five times during the sample period.[4] After applying these criteria, our final sample included 238 brands[5] across 22 product categories from approximately 18,000 households.

2.3 Pareto ratio summary statistics

The overall distributions of the PR at the brand and product category level are displayed in Fig. 1.

The upper panel in Fig. 1 provides a box plot of the PR at the brand level within each product category. The *x*-axis indicates the proportion of sales from the top 20% of consumers at the brand level (i.e., PR), and the *y*-axis indicates a product category. The black bar in each box indicates the mean of the PR within the product category. As can be seen, the overall mean of the PR at the brand level is .73[6] and the standard variation is .07. We observe several interesting patterns in this analysis. First, as shown in the upper panel of Fig. 1, the average PR (at the brand level within the product category) of all 22 product categories is between .65 and .90. In addition, more than half of the total product categories show the PR greater than .70. Although the PR is not precisely .80, this result implies that the sales revenues of most brands in these 22 product categories depend heavily on the top 20% of consumers. Second, the PR (at the brand level) varies within the product category. "Light Beer" category shows the largest variance (i.e., a standard deviation is .07) and "Orange Juice," "Pizza-Frozen," "Detergent," and "Toilet Tissue" categories have the smallest variance (i.e., a standard deviation is .02). This result implies that some brand-level attributes within the product category (e.g., market share, niche versus change-of-pace brand, purchase (promotion) frequency/expenditure of the brand, and so on) are potential factors to explain variation of the PR. Third, the average PR (at the category level) varies across product categories. Cigarettes have the highest average PR (i.e., .89) and detergents have the lowest average PR (i.e., .64). Interestingly, cigarettes, light beer, and soft drinks, which can be considered to be hedonic (or even addictive) product categories, have a higher average PR than other product categories. Similarly, this variation shows that some category-level attributes (e.g., purchase frequency/expenditure or promotion frequency/expenditure) could be important factors to explain a variation of the PR across all product categories. The lower panel in Fig. 1 provides a histogram of the PR of all brands across the 22 product categories.

2.4 Explaining the variation of Pareto ratio

The variation of the PR both within each product category and across product categories suggests the following question: Can this variation be related to brand and product category

[4] Households with at least five purchases in a category were labeled as "category users" and were included in the analyses. We also did some sensitivity analysis on the threshold to define a category of users by altering a cutoff point (i.e., 1, 3, 5, and 10). The Pareto ratio slightly increases under the cutoff point at 1 and 3, but the results from 5 and 10 are the same.

[5] Additionally, we did not include the brands with less than 1% market share because including those brands in our analysis would inflate the PR, which would be higher than the current number. In addition, in the CPG market, there are many small brands that did not exist during our whole data collection period (i.e., 2004–2010), so focusing on main brands (which have existed in the market over the long term) would show clearer patterns of the PR.

[6] The mean of PR at the brand level based on the original definition of A.C. Nielsen is .65, which is not significantly different.

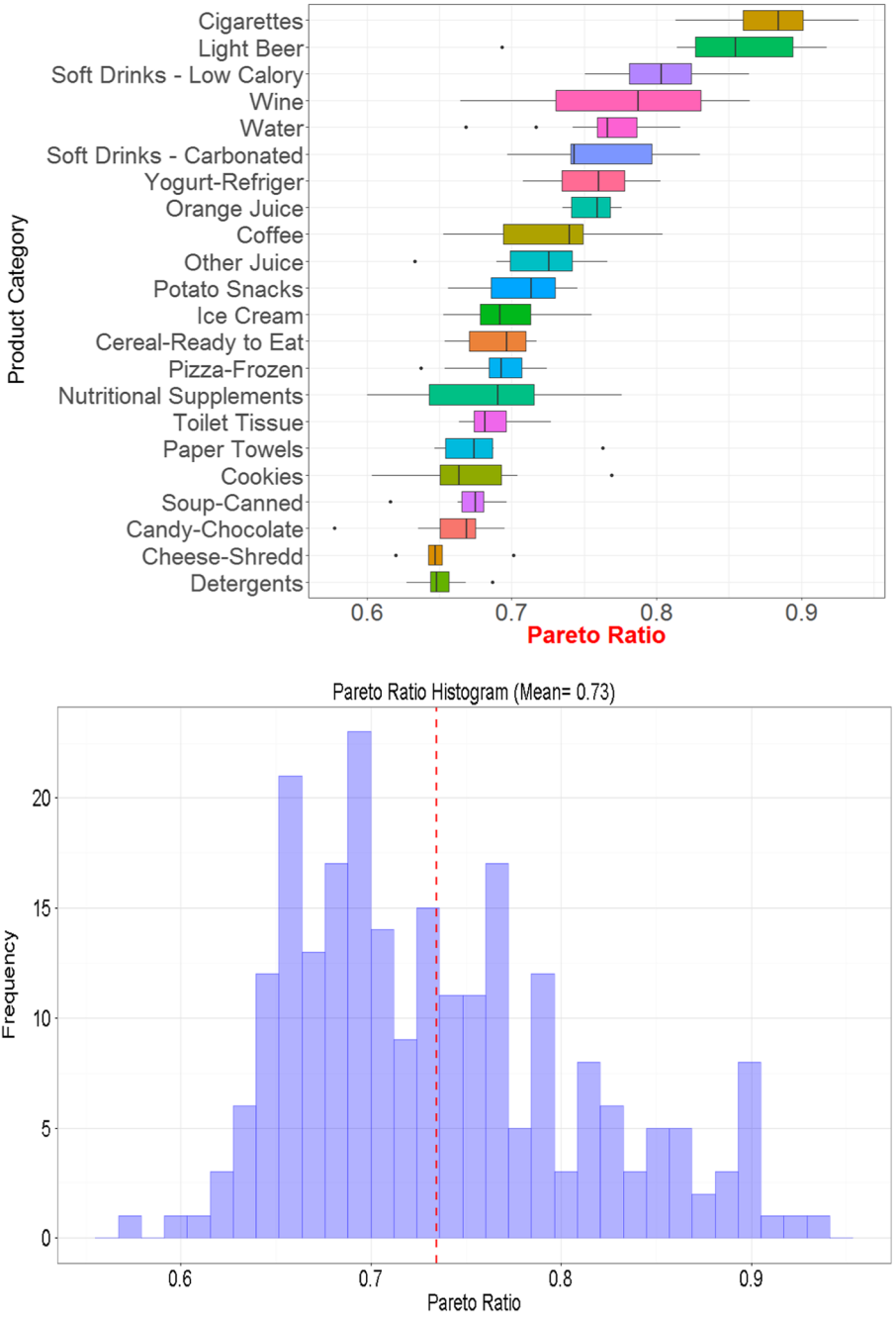

Fig. 1 The distribution of the Pareto ratio at the brand level

characteristics? In previous literature (McPhee 1963; Raj 1985; Ehrenberg 1988; Kahn et al. 1988; Ehrenberg et al. 1990), the authors examine brand attributes, consumer purchase behavior, marketing mix variables (e.g., promotion), and market competition to explain the

effect of marketing on brand loyalty. Although the PR is not the same concept as brand loyalty, in some sense, those two concepts are similar because the PR indicates the concentration of a brand's sales coming from the top 20% of consumers who are more likely to be behaviorally loyal.[7] Therefore, we chose to investigate brand/category attributes, consumer purchase behaviors, marketing mix variables, and market competition as potential factors that affect the variation of the PR's distribution.

2.4.1 Brand-/category-level size

In the extant brand loyalty literature, it has been argued (e.g., Ehrenberg 2000) that in the CPG industry, an increase in a brand's sales is not typically due to purchases by loyal customers but rather from newly acquired consumers. Because current loyal consumers of CPGs have already reached the maximum of their consumption capacity, they cannot increase their purchasing level significantly. This implies that if brands of CPGs hope to increase market share (or penetration), they must increase the number of new consumers. As a result, brands having a higher market share (or penetration) are more likely to have a large proportion of non-loyal consumers which leads to a decrease in the PR. Thus, the hypothesis regarding the brand/category size is as follows:

- Hypothesis 1: Brand market share (or penetration) is negatively related to the Pareto ratio.

2.4.2 Brand-/category-level purchase frequency and expenditure

In the previous literature on brand loyalty, there are two conflicting arguments about how purchase frequency (i.e., how often do consumers go on shopping trips during an observation period?) or expenditure (i.e., how much money do consumers spend on a given shopping trip?) influences brand loyalty.

McPhee (1963) argues that within a product category, a small brand has a higher proportion of infrequent consumers and their purchase frequency and expenditure is much smaller than that of a more popular brand. This is called the "double jeopardy effect." Ehrenberg et al. (1990) make a similar argument that a frequently purchased brand tends to have a higher brand penetration while a rarely purchased brand tends to have a lower brand penetration due to the double jeopardy effect. In other words, a frequently (rarely) purchased brand is more likely to have many loyal customers (non-

[7] In this paper, the top 20% of consumers was determined by an amount of total dollar expenditure for a certain brand during the observation period. We defined the top 20% of consumers by sorting all consumers of each brand on the basis of total dollar expenditure for each brand. The top 20% of consumers could be either frequent shoppers or large basket shoppers since total dollar expenditure is a function of basket size and purchase frequency. For example, total dollar expenditure could be higher if consumers frequently purchase the brand although expenditure per shopping trip might be small. On the other hand, this number could also be higher if the basket size is large, meaning that expenditure per shopping trip is large although consumers rarely purchase a particular brand. Thus, it could be controversial to define the top 20% of consumers by their behavioral loyalty. However, interestingly, in our dataset, consumers having a higher purchase frequency were also more likely to have a larger basket size. We concluded that the top 20% of consumers, based on the total dollar expenditure, were behaviorally loyal consumers and we will investigate correlates (i.e., that have been studied in previous literature to see the effects of those on brand loyalty) to see how those affect the PR.

loyal customers) and this could lead to having the higher (lower) PR. This leads to the following hypothesis:

- Hypothesis 2: Purchase frequency and expenditure are positively related to the Pareto ratio.

2.4.3 Brand-/category-level purchase frequency and expenditure with promotion

In previous literature related to promotions, Gupta (1988) argues that the main effect of a sales promotion is from non-loyal customers switching due to the temporary price reduction. Other studies (Dodson et al. 1978; Neslin and Shoemaker 1989) found that the repurchase probability of consumers induced by price promotions is significantly lower than that for people who purchased without a deal. For this reason, brands that have frequent promotions are more likely to have a large proportion of non-loyal consumers. Thus, the PR of frequently promoted brands tends to be lower than other infrequently promoted brands. The same logic can be applied to purchase expenditure with a promotion variable:

- Hypothesis 3: Promotion purchase frequency and expenditure are negatively related to the Pareto ratio.

2.4.4 Brand attribute: niche versus change-of-pace brands

Kahn et al. (1988) set a definition of "niche" and "change-of-pace" brands using a constant noted by Ehrenberg (1972):

$$\begin{aligned} &\textit{Ehrenberg constant} \\ &\quad = \textit{Annual Purchase Frequency of Brand} \,(\text{per } \textit{Household}) \times \\ &\qquad (1 - \textit{Annual Brand Penetration}) \end{aligned} \tag{2}$$

According to Ehrenberg's constant, niche brands are the most frequently purchased brands by a small number of people in a certain product category (i.e., niche brands have the highest Ehrenberg constant within a product category). This definition implies that consumers of niche brands are more likely to be loyal and the PR of niche brands could be higher than that for other brands.

On the contrary, change-of-pace brands are the least frequently purchased brands by a large number of consumers in a certain product category (i.e., change-of-pace brands have the lowest Ehrenberg constants within a product category). Similar to the case of a niche brand, this definition implies that consumers of the change-of-pace brands are more likely to be non-loyal and the PR of these brands will be lower than that of the other brands:

- Hypothesis 4: Niche brands are more likely to have a higher Pareto ratio and change-of-pace brands are more likely to have a lower Pareto ratio.

2.4.5 Market competition

Although there is not much in the literature explaining a correlation between market competition and brand loyalty, we can hypothesize that brands in competitive product categories are more likely to have non-loyal consumers because consumers would be given constant inducements to switch brands. This could be due to either promotion (i.e., Hypothesis 4), advertising, channel policies, or other competitive moves. Thus, many consumers could be non-loyal, and a large proportion of non-loyal consumers leads to a lower PR than that of other brands in the less competitive product category. This leads to another hypothesis:

- Hypothesis 5: The level of competition in a product category is negatively related to the Pareto ratio.

2.5 Measures of independent variables

Brand size was measured in two ways. One measure was market share. Another measure was brand-specific penetration. This variable was calculated by dividing the total number of consumers who purchased the brand at least once by the total number of consumers who purchased the applicable category at least once in that year. Because each brand had six observations (i.e., from 2004 to 2009), we simply took an average of those six observations and called this number the annual brand-specific penetration. An annual category-specific penetration was also calculated in the same manner; however, the denominator in this case was the total number of consumers in the panel for that year.

Purchase frequency was measured by dividing the total annual purchase frequency of the brand by the total number of consumers who purchased the brand at least once in that year. We called this measure the annual average purchase frequency. Purchase expenditure was measured in the same manner; however, the numerator was the total annual sales of the brand instead of purchase frequency.

Promotion purchase frequency was measured by taking an average of the proportion of all consumers' purchase frequency on a promotion; this number was the annual average promotion purchase frequency. Promotion purchase expenditure was measured in the same manner.

Competition was measured in two ways. One measure was the total number of brands in the product category. Another measure was by the Herfindahl Hirschman Index (HHI), which aggregates a square of the market share of all brands in a product category. A high HHI indicates low competition in the market.

Classifications of a niche brand and change-of-pace brand were measured by Ehrenberg statistics. As defined in the previous section, a brand with a higher Ehrenberg statistic is a niche brand, and a brand with a lower statistic is a change-of-pace brand.

A description of all the measurements is summarized in Table 1.

3 Empirical analysis and results

3.1 Regression analysis

The discussion above and Fig. 1 demonstrated that there is a large variation of the PR at the brand level within a product category and across product categories. In addition, we hypothesize several potential factors, including brand-/category-level attributes, to be the source of the variation. To test the above five hypotheses, we ran a brand-level regression model where the dependent variable was the PR for each brand in 22 product categories and the independent variables were the brand-/category-level attributes for the 238 brands from the 22 product categories.[8,9,10] The regression model specification is as follows:

$$\begin{aligned} &Pareto\ \ ratio \\ &= \alpha + \beta_1 Brand\ Size + \beta_2 Brand{-}Purchase\ Frequency + \beta_3 Brand{-}Purchase\ Expenditure \\ &+\beta_4 Brand{-}Promotion\ Frequency + \beta_5 Brand{-}Promotion\ Expenditure + \beta_6 Niche + \beta_7 Change{-}of \\ &-Pace\ \ + \gamma_1 Category{-}Size + \gamma_2 Category{-}Purchase\ Frequency + \gamma_3 Category{-}Purchase\ Expenditure \\ &+\gamma_4 Category{-}Promotion\ Frequency + \gamma_5 Category{-}Promotion\ Expenditure \\ &+\gamma_6 Category\text{-}Competition + \varepsilon \end{aligned} \tag{3}$$

Because OLS regression can give biased coefficients when the dependent variable is the proportion or fraction, previous literature (Ferrari and Cribari-Neto 2004) suggests using either the generalized linear model (GLM) with logit transformation or beta regression, assuming the dependent variable follows the beta distribution and is distributed on the (0, 1) interval.

By using this method, we can deal with the continuous dependent variable that lies between 0 and 1 through a regression structure. For a robustness check, we ran both GLM and a beta regression and the results were consistent, although the sizes of the coefficients were a little different. Thus, hereafter, we only report the results from the beta regression.

Second, we discretized all continuous independent variables into three groups (high, medium, and low) based on quintiles. We did this to better understand any non-linear relationships between the independent and the dependent variables.[11] Because the

[8] For the robustness check, we conducted the same analysis with larger samples including 16,000 brands from 100 product categories. The results were consistent with the results from smaller samples.

[9] We also ran a regression with different samples, separated by an observation period, to see the difference between samples. Since there might be a difference between consumers who had been on the panel for 6 years and for 1 year, we checked the robustness of our regression results by separating observations with the sample length. First, we ran a regression with observations only from 2004 to 2006 and from 2007 to 2009, respectively. Next, we ran a regression of the total observations from 2004 to 2009, and then compared the results between those three regressions. The results are consistent with the prior findings.

[10] We also looked at the correlations between the independent variables. For both brand- and category-level attributes, most variables are not correlated with each other except the purchase (and promotion) frequency and expenditure. As we mentioned in footnote 7, in our data, the purchase (and promotion) frequency and expenditure are highly correlated, so to deal with multicollinearity, we ran multiple regressions with and without those correlated variables.

[11] We also repeated the same analyses with continuous independent variables. In order to capture any non-linear relationships, we included quadratic terms of the continuous independent variables such as market share, brand penetration, purchase frequency/expenditure, and promotion frequency/expenditure. The results are consistent with the discretized ones (except the coefficient of the niche-brand dummy became insignificant).

Table 1 Summary description of measures

Variable	Description
Brand/category size	Brand level: annual penetration (= no. of brand users/no. of product category users), Market share (= brand total sales/product category total sales) Category level: annual penetration (= no. of product category users/no. of total users in the sample)
Purchase frequency	Brand level: annual avg. purchase frequency (per consumer) for each brand Category level: annual avg. purchase frequency (per consumer) for each product category
Purchase expenditure	Brand level: annual avg. purchase expenditure (per consumer) for each brand Category level: annual avg. purchase expenditure (per consumer) for each product category
Purchase frequency with promotion	Brand level: annual avg. purchase frequency with promotion (per consumer) for each brand Category level: annual avg. purchase frequency with promotion (per consumer) for each product category
Purchase expenditure with promotion	Brand level: annual avg. purchase expenditure with promotion (per consumer) for each brand Category level: annual avg. purchase expenditure with promotion (per consumer) for each product category
Market competition	HHI for each product category
Niche Change-of-pace brands	Ehrenberg's statistics for each brand within a product category

reference group is the middle quintile, the way of interpreting the coefficients would be as a contrast between the high- (or low-) and middle-quintile group.

Table 2 shows the regression results with only brand-level variables. Because the dependent variable, the PR, is measured at the brand level, we first investigated whether only brand-level attributes can explain its variation. To control for heterogeneity across product categories, we included category fixed effects dummy variables (coefficients not shown). Additionally, to find the best-fit model to explain the PR and to deal with endogeneity, we sequentially included variables from the exogenous set and observed the changes of the direction and significance of the coefficients. For example, because promotion variables could be endogenous, we ran a regression, including and excluding promotion variables, and then saw how the coefficients of other variables are changed. As shown in Table 2, columns 4 and 5, the coefficients are stable across the different specifications.

Table 3 shows the results of regression with both brand- and category-level variables. In this regression, we included category-level attributes as independent variables to control for category-specific effects instead of using category-specific dummy variables. We looked to see if the brand-level coefficients were stable under the category-level attribute specification by comparing Tables 2 and 3.

In addition, we again ran a regression with a different specification by including and excluding several endogenous variables to see the change of significance and direction of the other variables' coefficients. The results were consistent across the different model specifications. In particular, the direction

Table 2 Beta-regression results with brand-level variables and category fixed effects

	Dependent variable Pareto ratio					
	(1)	(2)	(3)	(4)	(5)	(6)
Market share (high)	*− 0.084***	*− 0.174****	*− 0.124****	*− 0.124****	*− 0.128****	
	(0.035)	*(0.031)*	*(0.032)*	*(0.031)*	*(0.032)*	
Market share (low)	*− 0.085***	− 0.008	− 0.004	− 0.0002	0.001	
	(0.034)	(0.029)	(0.028)	(0.027)	(0.027)	
Penetration (high)						*− 0.146****
						(0.033)
Penetration (low)						− 0.039
						(0.030)
Purchase frequency (high)		*0.204****	*0.175****	*0.152****	*0.135****	*0.154****
		(0.044)	*(0.043)*	*(0.042)*	*(0.042)*	*(0.043)*
Purchase frequency (low)		*− 0.288****	*− 0.251****	*− 0.220****	*− 0.213****	*− 0.210****
		(0.032)	*(0.031)*	*(0.031)*	*(0.032)*	*(0.031)*
Purchase expenditure (high)				*0.193****	*0.267****	*0.266****
				(0.049)	*(0.062)*	*(0.061)*
Purchase expenditure (low)				− 0.038	− 0.042	− 0.044
				(0.035)	(0.036)	(0.035)
Promotion frequency (high)					− 0.046	− 0.047
					(0.030)	(0.030)
Promotion frequency (low)					0.020	0.029
					(0.030)	(0.030)
Promotion expenditure (high)					− 0.024	− 0.029
					(0.034)	(0.034)
Promotion expenditure (low)					*− 0.123**	*− 0.107**
					(0.063)	*(0.063)*
Niche			*0.124****	*0.123****	*0.119****	*0.122****
			(0.039)	*(0.038)*	*(0.037)*	*(0.037)*
Change of pace			*− 0.121****	*− 0.115****	*− 0.111****	*− 0.106****
			(0.036)	*(0.035)*	*(0.035)*	*(0.034)*
Observations	238	238	238	238	238	238
R^2	0.770	0.842	0.853	0.862	0.866	0.867
Log likelihood	444.799	488.231	499.160	507.169	510.952	512.558

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

and significance of the brand-level variables' coefficients are the same as the results from Table 2 and suggest that our results are robust under each different model specification and potential endogenous variables.

Lastly, Table 4 shows the marginal effects of the independent variables based on the parameter estimates from Table 3.

Table 3 Beta-regression results with brand- and category-level variables

	Dependent variable Pareto ratio				
	(1)	(2)	(3)	(4)	(5)
Market share (high)	*− 0.126****	*− 0.127****	*− 0.141****	*− 0.148****	*− 0.130****
	(0.036)	*(0.034)*	*(0.038)*	*(0.034)*	*(0.035)*
Market share (low)	0.0002	0.001	− 0.009	0.005	− 0.001
	(0.031)	(0.029)	(0.033)	(0.030)	(0.029)
Purchase frequency (high)	*0.180****	*0.171****	*0.174****	*0.176****	*0.168****
	(0.045)	*(0.042)*	*(0.046)*	*(0.043)*	*(0.043)*
Purchase frequency (low)	*− 0.175****	*− 0.190****	*− 0.188****	*− 0.201****	*− 0.196****
	(0.037)	*(0.034)*	*(0.038)*	*(0.034)*	*(0.034)*
Purchase expenditure (high)	*0.307****	*0.275****	*0.309****	*0.287****	*0.286****
	(0.073)	*(0.067)*	*(0.076)*	*(0.069)*	*(0.068)*
Purchase expenditure (low)	− 0.010	− 0.045	0.002	− 0.050	− 0.045
	(0.039)	(0.037)	(0.041)	(0.038)	(0.037)
Promotion frequency (high)	*− 0.094****	*− 0.052**	*− 0.101****	*− 0.057**	*− 0.057**
	(0.033)	*(0.031)*	*(0.034)*	*(0.032)*	*(0.032)*
Promotion frequency (low)	0.009	0.008	0.016	0.002	0.006
	(0.035)	(0.032)	(0.036)	(0.033)	(0.033)
Promotion expenditure (high)	− 0.030	− 0.019	*− 0.064**	− 0.012	− 0.010
	(0.037)	(0.036)	*(0.038)*	(0.037)	(0.037)
Promotion expenditure (low)	*− 0.119**	− 0.086	*− 0.123**	*− 0.113**	− 0.103
	(0.072)	(0.067)	*(0.075)*	*(0.068)*	(0.067)
Niche	*0.112***	*0.119****	*0.108***	*0.105***	*0.120****
	(0.044)	*(0.041)*	*(0.046)*	*(0.041)*	*(0.041)*
Change of pace	*− 0.118***	*− 0.116****	*− 0.122****	*− 0.115****	*− 0.112****
	(0.041)	*(0.038)*	*(0.043)*	*(0.038)*	*(0.038)*
Category					
Purchase frequency (high)	*0.383****	*0.410****	*0.350****	*0.325****	*0.367****
	(0.048)	*(0.055)*	*(0.050)*	*(0.049)*	*(0.050)*
Purchase frequency (low)	*− 0.187****	*− 0.237****	*− 0.225****	*− 0.273****	*− 0.236****
	(0.042)	*(0.042)*	*(0.042)*	*(0.041)*	*(0.041)*
Purchase expenditure (high)	0.015	*0.314***	0.092	0.155	*0.228****
	(0.077)	*(0.122)*	(0.077)	(0.112)	*(0.114)*
Purchase expenditure (low)	*− 0.105***	*− 0.177****	− 0.015	− 0.057	*− 0.121***
	(0.041)	*(0.053)*	(0.038)	(0.039)	*(0.048)*
Promotion frequency (high)		*− 0.182****		*− 0.137****	*− 0.180****
		(0.044)		*(0.040)*	*(0.044)*
Promotion frequency (low)		− 0.043		*0.064**	0.013
		(0.048)		*(0.035)*	(0.042)
Promotion expenditure (high)		*− 0.128****		*− 0.166****	*− 0.103***
		(0.044)		*(0.043)*	*(0.043)*

Table 3 (continued)

	Dependent variable Pareto ratio				
	(1)	(2)	(3)	(4)	(5)
Promotion expenditure (low)		*− 0.222***		0.001	− 0.119
		(0.103)		(0.078)	(0.095)
Number of brands (high)	*0.163****	*0.112****			0.058
	(0.045)	*(0.049)*			(0.041)
Number of brands (low)	*− 0.139****	*− 0.196****			*− 0.101**
	(0.048)	*(0.066)*			*(0.055)*
HHI (high)	*0.085***	*0.121***	− 0.004	0.025	
	(0.043)	*(0.048)*	(0.038)	(0.040)	
HHI (low)	0.001	− 0.024	*0.069**	− 0.003	
	(0.041)	(0.041)	*(0.037)*	(0.038)	
Penetration (high)	− 0.057	− 0.045	*− 0.173****	*− 0.231****	*− 0.138****
	(0.045)	(0.070)	*(0.040)*	*(0.042)*	*(0.060)*
Penetration (low)	0.065	− 0.097	0.043	− 0.073	− 0.089
	(0.057)	(0.060)	(0.059)	(0.060)	(0.060)
Observations	238	238	238	238	238
R^2	0.819	0.844	0.804	0.838	0.840
Log likelihood	469.951	490.105	459.410	484.825	486.884

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

3.2 Summary of results

First, as shown in Table 4, we found that the brands with the largest share and the categories with the highest penetration rates have the lowest PRs. Thus, Hypothesis 1 is supported. Interestingly, the relationships are non-linear. Specifically, high market share brands are likely to have a 3% lower PR than middle market share brands; however, low market share brands are not significantly different from middle market share brands. This result is consistent with Ehrenberg's (2000) suggestion that because current loyal CPG consumers have already reached maximum consumption capacity, they cannot increase their level of consumption. Therefore, higher market share brands have a large number of consumers with a high proportion of non-loyal consumers. For this reason, the PR tends to decrease as market share increases, and this finding is stronger for the higher market share brands.

Second, we found that purchase frequency and expenditure at both the brand and category levels are significantly positively related to the PR supporting Hypothesis 2. As an illustration, high purchase frequency brands are likely to have a 3% higher PR, and low purchase frequency brands are likely to have a 4% lower PR than the middle purchase frequency brands. Also, high purchase expenditure brands are likely to have a 5% higher PR than the middle purchase expenditure brands. This result corresponds with the double jeopardy effect and shows that incremental purchase frequency of brands is driven by the loyal consumers.

Third, we found that the PR decreases with promotion purchase frequency only for high-quintile brands and increases with promotion purchase expenditure only for low-quintile brands. In terms of promotion purchase frequency and expenditure variables, we only reported the results from the regression analysis, because the marginal effects of the promotion purchase frequency and expenditure are not significant (i.e., only high promotion purchase frequency brands are marginally significant at the .1 level) although the regression coefficients are significant. This is because the delta method was used to obtain the standard error of the marginal effect, meaning that the standard error for the marginal effect of one independent variable relates to the entire variance-covariance matrix from the estimation together with the corresponding entries from the Jacobian. Thus, many variables were included in the calculation beyond just the standard error of the coefficient of that variable, and these other variables could cause it to be greater than .05, even when the coefficient standard error was smaller than .05.

As shown in Tables 2 and 3, the coefficient of the promotion purchase frequency of high-quintile brands is negative and significant, indicating that high promotion purchase frequency brands are more likely to have the lower PR than middle-quintile brands. However, the low-quintile brands are insignificantly different from the middle-quintile brands meaning that promotion purchase frequency is non-linear with respect to the PR.

This result corresponds to Gupta's (1988) argument that the main effect of a sales promotion is due to brand switching because many non-loyal consumers switch their choices due to the temporal price promotions. In other words, high promotion purchase frequency brands tend to have a large proportion of non-loyal consumers and this leads to a lower PR than other brands.

In contrast, as shown in Tables 2 and 3, the coefficient of the promotion purchase expenditure of low-quintile brands is negative and marginally significant at the .1 level, indicating that low promotion purchase expenditure brands are more likely to have a lower PR than middle-quintile brands. However, high-quintile brands are insignificantly different from the middle-quintile brands meaning that promotion purchase expenditure is non-linear with respect to the PR. Thus, we can conclude that Hypothesis 3 is partially supported.

Fourth, we found that niche brands have a higher PR and change-of-pace brands have a lower PR compared to other brands. Specifically, niche brands tend to have a 2% higher PR and change-of-pace brands tend to have a 2% lower PR than other brands. Because niche brands are the most frequently purchased brands by a small number of consumers in a certain product category, these brands are more likely to have a high proportion of loyal consumers, and this leads to an increase in the PR. However, change-of-pace brands are the least frequently purchased brands by a large number of consumers in a certain product category; therefore, change-of-pace brands tend to have a high proportion of non-loyal consumers and this leads to a decrease in the PR. Thus, we can conclude that Hypothesis 4 is supported.

Fifth, we observed that brands in a slightly competitive market are more likely to have a higher PR than brands where the market exhibits high competition (i.e., positive and significant coefficient of the high-quintile HHI variable). As stated previously, this is because competition leads each brand to have more competitive marketing activity with a goal of acquiring new consumers. As a result, brands in a high-competition market tend to have a large proportion of non-loyal consumers, and this contributes to a lower PR; therefore, we can conclude that Hypothesis 5 is supported.

Table 4 Marginal effects with brand- and category-level variables from beta-regression

	Marginal effects	*p* value
Market share (high)	− *0.03***	0.00
Market share (low)	0.00	0.87
Purchase frequency (high)	*0.03***	0.00
Purchase frequency (low)	− *0.04***	0.00
Purchase expenditure (high)	*0.05***	0.00
Purchase expenditure (low)	− 0.01	0.19
Promotion frequency (high)	− *0.01**	0.08
Promotion frequency (low)	0.00	0.95
Promotion expenditure (high)	0.00	0.74
Promotion expenditure (low)	− 0.02	0.10
Niche	*0.02***	0.01
Change of pace	− *0.02***	0.00
Category		
Penetration (high)	− *0.05***	0.00
Penetration (low)	− 0.01	0.23
Purchase frequency (high)	*0.06***	0.00
Purchase frequency (low)	− *0.05***	0.00
Purchase expenditure (high)	0.03	0.16
Purchase expenditure (low)	− 0.01	0.15
Promotion frequency (high)	− *0.03***	0.00
Promotion frequency (low)	*0.01**	0.07
Promotion expenditure (high)	− *0.03***	0.00
Promotion expenditure (low)	0.00	0.99
HHI (high)	0.00	0.52
HHI (low)	0.00	0.94

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

4 Discussion

In this paper, we have examined the following three research questions: (1) To what extent does the Pareto principle hold in the CPG industry? (2) How does this principle vary across product categories/brands? (3) Which potential factors (brand/product attributes and consumer purchase behavior) explain the variation of the Pareto distribution?

First, we have provided evidence that the Pareto rule approximately holds in the CPG industries. The mean of the PR of each brand across product categories is .73, which is close to .80, which is what we expected. Second, we found that consumer purchase behaviors regarding brands (e.g., purchase frequency and purchase expenditure) are positively correlated with the PR. Additionally, niche brands are more likely to have a higher PR. On the other hand, brand/category size, promotion variables, and market competition variables are negatively correlated with the PR. In addition, change-of-pace brands are more likely to have a lower PR.

The results from this analysis have important managerial implications. Although CPGs are mass products that many people use, the Pareto rule still approximately holds. In other words, companies need to understand the proportion of loyal and heavy customers and how this varies across product categories and brands. For example, in the carbonated-soda product category, the average of the PR is .77. However, major brands such as Coca-Cola and Pepsi have a higher PR (i.e., Coca-Cola .79; Pepsi .80) but smaller brands such as 7 Up and A&W have the lower PR (i.e., 7 Up .71; A&W .70) than the average. If the brand has a high (or low) PR and companies know the brand's features, their customers, and the market, then efficient marketing strategies for the brand can be established to capitalize on either retention of current loyal customers or acquisition of new brand customers. In previous marketing literature about brand loyalty, there are equivocal arguments regarding whether companies focus on retaining current loyal consumers or acquiring new consumers (Rosenberg and Czepiel 1983; Ehrenberg 1972, 1988; Krishnamurthi and Raj 1991; Aaker 1991; Shin and Sudhir 2010). Of course, from the companies' perspectives, they could focus on both retention and acquisition strategies. However, given the importance of marketing efficiency today, many brand managers may have to make decisions on which to focus due to limited resources. As in our example, major brands such as Coca-Cola and Pepsi could choose efficient marketing strategies (i.e., either retention or acquisition) compared with smaller brands such as 7 Up and A&W based on results from the PR analysis incorporating other brand- and category-level attributes.

In terms of future research, while, in this paper, we have mainly focused on finding an empirical evidence to show the "Pareto rule" holds in frequently purchased product categories, it would be interesting to further analyze the data to see why this phenomenon is observed. Second, the PR at the manufacturer, UPC, and store levels can be investigated. For example, we can analyze store-level data to see whether the top 20% of UPCs in a certain product category can explain the 80% of product category sales for each store. In addition, how this ratio varies across product categories and different types of stores can also be studied.

References

Aaker, D. A. (1991). *Managing brand equity*. New York: The Free Press.

Anderson, C. (2006). *The long tail: why the future of business is selling less of more*. New York: Hachette Books.

Brynjolfsson, E., & Smith, M. (2000). Frictionless commerce? A comparison of internet and conventional retailers. *Management Science, 46*(4), 563–585.

Brynjolfsson, E., Hu, Y. J., & Smith, M. D. (2003). Consumer surplus in the digital economy: estimating the value of increased product variety at online booksellers. *Management Science, 49*(11), 1580–1596.

Brynjolfsson, E., Hu, Y. J., & Simester, D. (2011). Goodbye Pareto principle, hello long tail: the effect of search costs on the concentration of product sales. *Management Science, 57*(8), 1373–1386.

Dodson, J. A., Tybout, A. M., & Sternthal, B. (1978). Impact of deals and deal retraction on brand switching. *Journal of Marketing Research, 15*(1), 72–81.

Ehrenberg, A. S. C (1972). *Repeat-buying : theory and applications*. Amsterdam: American Elsevier.

Ehrenberg, A. S. C. (1988). Repeat-buying: facts, theory and applications, 2nd ed. Edward Arnold, London; Oxford University Press, New York. Reprinted in the Journal of Empirical Generalisations in Mark Science, 2000, 5, 392–770 (www.empgens.com).

Ehrenberg, A. (2000). Repeat buying. *Journal of Empirical Generalisations in Marketing Science, 5*(2).

Ehrenberg, A. S. C., Goodhardt, G. J., & Patrick Barwise, T. (1990). Double jeopardy revisited. *The Journal of Marketing, 54*(3), 82–91.

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics, 31*(7), 799–815.

Gupta, S. (1988). Impact of sales promotions on when, what, and how much to buy. *Journal of Marketing Research, 25*(4), 342–355.

Kahn, B. E., Kalwani, M. U., & Morrison, D. G. (1988). Niching versus change-of-pace brands: using purchase frequencies and penetration rates to infer brand positionings. *Journal of Marketing Research, 25*(4), 384–390.

Krishnamurthi, L., & Raj, S. P. (1991). An empirical analysis of the relationship between brand loyalty and consumer price elasticity. *Marketing Science, 10*(2), 172–183.

McPhee, W. N. (1963). *Formal theories of mass behaviour*. New York: The Free Press of Glencoe.

Morrison, D. G., & Schmittlein, D. C. (1981). Predicting future random events based on past performance. *Management Science, 27*(9), 1006–1023.

Morrison, D. G., & Schmittlein, D. C. (1988). Generalizing the NBD model for customer purchases: what are the implications and is it worth the effort? *Journal of Business and Economic Statistics, 6*(2), 145–159.

Neslin, S., & Shoemaker, R. W. (1989). An alternative explanation for lower repeat rates after promotion purchases. *Journal of Marketing Research, 26*(2), 205–213.

Raj, S. P. (1985). Striking a balance between brand "popularity" and brand loyalty. *Journal of Marketing, 49*(1), 53–59.

Rosenberg, L., & Czepiel, J. (1983). A marketing approach for consumer retention. *Journal of Consumer Marketing, 1*(1), 45–51.

Schmittlein, D. C., Cooper, L. G., & Morrison, D. G. (1993). Truth in concentration in the land of (80/20) laws. *Marketing Science, 12*(2), 167–183.

Shin, J., & Sudhir, K. (2010). A customer management dilemma: when is it profitable to reward one's own customers? *Marketing Science, 29*(4), 671–689.

Twedt, D. W. (1964). How important to marketing strategy is the heavy users? *Journal of Marketing, 28*(1), 71.