Incorporating frequency estimation on discrete choice

Aku-Ville Lehtimäki
Aalto University School of Business
Outi Somervuori
Aalto University School of Business & DT Helsinki Oy

Cite as:
Lehtimäki Aku-Ville, Somervuori Outi (2019), Incorporating frequency estimation on discrete choice. *Proceedings of the European Marketing Academy*, 48th, (8654)

Paper presented at the 48th Annual EMAC Conference, Hamburg, May 24-27, 2019.

Incorporating frequency estimation on discrete choice

Abstract

Discrete choice models are often used to estimate preference shares. However, less if not all attention is paid to the purchase frequency. This information is equally essential for demand estimates. In this paper we present a method with theoretical background how to survey purchase frequency and model the purchase frequency behavior, which is followed by an example. Incorporating frequency estimation into discrete choice research will allow researchers to make better predictions of potential demand.

Keywords: Purchase frequency, demand estimation, exponential distribution

The track this paper is intended for: "Consumer Behavior"

Discrete choice models mainly concentrate on estimating preference shares. In a typical conjoint a purchase situation is a static one, and only one product concept is chosen. However, in real world the customer may buy more than one piece of several products. This angle has been subject to volumetric conjoint studies (e.g. Kim, Allenby & Rossi, 2004). However, often purchase frequency also plays a crucial role on demand estimates. Some products are consumed daily, while others more seldom. For example, most of us purchase, say, food weekly, while we go to movies, say, a few times a year. This logic is very difficult if not impossible to capture in any discrete choice analysis. In this paper, we present a method to measure this frequency. The respondents are surveyed when their last purchase was, and segments based on purchase frequency are determined using the exponential distribution and the theory of finite mixture models. If the next visit is known (it is happening at the time of the survey), the respondents can be assigned into these segments.

1. Surveying the usage frequency

The respondents may in general underestimate or overestimate the frequencies, if they are asked directly how often they purchase a product or use a service. However, an alternative way to estimate usage frequency is to ask, "How long ago was the last time?" and offer ordinal responses. This technique yields more reliable answers than open answers since it is considerably easier for the respondent to remember the last instance than to estimate frequency.

The ordinal alternatives can be, for example, one day, week, month, year etc. ago. Additional alternatives, not part of the actual ordinal scale, such as over a year ago and never can also be offered.

The answers are transformed into the same scale: One month becomes 30 days and one year 365 days, for example. If intervals are used then their midpoint is a reasonable choice to represent them meaning, for example, "within two days" becomes 1 day and "within a week" becomes 3,5 days.

If the survey is carried out just when the customer is purchasing a product or using the service, the average interval between visits can be calculated since we know the current time and previous time. However, the marketing research takes usually place in the context where this is not the case, and the time until next visit is left open. This problem can, however, be solved using in the model we propose.

2. Finite mixture model

In general, the finite mixture model (FMM) assumes the data result from not just one but many random processes with different parameters (see e.g. McLachlan & Peel, 2000). The model itself is therefore a sum of products of a function and corresponding proportion parameter i.e.:

$$\sum_{k=1}^{K} \pi_k g_k \tag{1}$$

Where πk refers to the proportion, and gk is any function we choose to represent the origin of the data. K is the number of segments, which can be decided upon based on the data, since the segmentation is model based, and in general different models can be compared.

3. Exponential distribution

The exponential distribution describes time between events in Poisson point process. Here the event refers to purchase or visit. Its point density function (PDF) and cumulative density function (CDF) can be defined as:

$$f(x) = \lambda \exp(-\lambda x) \Leftrightarrow F(x) = \int_0^x \lambda \exp(-\lambda t)\, dt = 1 - \exp(-\lambda x) \tag{2}$$

Where λ refers to the INVERSE of average time between events i.e. visits or usage in general.

4. The next visit is known

If we know when the respondent is going to visit next time, the interval between two visits is known. This interval can naturally be assumed coming from an exponential distribution, since in general the exponential distribution measures the waiting times between events in a Poisson point process, and we are interested how long the customers wait after a visit until they visit again.

However, when the waiting times between visits are known, we may define the function g to be the Poisson distribution point density function and we get the following form for the finite mixture model:

$$\sum_{k=1}^{K} \pi_k \lambda_k \exp(-\lambda_k x_j) \tag{3}$$

Where xj is the answer provided by the jth respondent. Since this FMM PDF is indeed a PDF, a likelihood function can be constructed, and subsequently maximum likelihood estimates for the parameters πk and λk can be calculated.

5. The next visit is unknown – observations are right-censored[1]

When the respondent is asked about the time of her previous visit, but the next visit does not take place immediately, and hence we do not know the waiting time: The observation is right-censored. However, what we do know is that the waiting time is at least the time between the last visit and the answer.

In this situation, no likelihood function can be constructed. However, the PDF can be replaced with the corresponding CDF. The function g is now:

$$g_k(z) = 1 - \exp(-\lambda_k z) \tag{4}$$

The notation z is used instead of x, since we are now focusing on the accumulation of the answers. The corresponding FMM is therefore:

$$\sum\nolimits_{k=1}^{K} \pi_k (1 - \exp[-\lambda_k z]) \tag{5}$$

Now this expression is the expected cumulation at point z. What we want to achieve naturally is to compare it to the observed corresponding CDF values, and make this difference as non-existent as possible by choosing the best values for πk and λk.

6. Determining the segment membership of a respondent

Since the exponential distribution has only one parameter, the probability of an observation coming from any of the probability distributions can be calculated scaling the corresponding PDF with all the possible PDFs and having the value of the observation as their argument i.e.:

$$p_k = f_k(x_i) / \sum\nolimits_{k=1}^{K} f_k(x_i) \tag{6}$$

The PDFs can be weighted with prior information, and the segment membership can be subsequently decided with, for example, a maximum a posteriori rule: Choosing the segment with the highest probability. However, while the parameters can be estimated using right-censored data, a right-censored observation does not qualify as an argument in the formulation as such. Determining the segment membership is therefore more cumbersome for right-censored data.

7. A small example with real data

[1] A similar model has been used previously to infer how much the respondents copy copyrighted material from different sources (Malmberg, 2015).

The following data come from a real survey. The panel respondents were asked about what the last time was they visited a restaurant, and therefore the next visit was unknown. The frequency and observed cumulative percentage distributions are presented in the following table.

	Frequency	**Obs. cum. %**	**Exp. cum. %**	**Residual**
Within 2 days	90	19.52 %	18.86 %	<0.0001
3 to 6 days ago	118	45.12 %	46.33 %	0.0001
1 to 2 weeks ago	87	63.99 %	61.83 %	0.0005
3 to 4 weeks ago	48	74.40 %	80.46 %	0.0037
Within 2 months	72	90.02 %	84.97 %	0.0026
Within 1 year	30	96.53 %	99.99 %	0.0012
Within 2 years	16	100.00 %	100.00 %	<0.0001
Total	461			0.0081

A model with two segments were chosen and the expected cumulative percentage distribution is based on that. The residual represents the squared difference between observed and expected relative cumulative distributions. The parameters λ1, λ2 and π are .5590, .0476 and .3723 in this respect. In other words, about 37 % of the respondents belong to the first segments and 63 % belong to the second segment. The average time between visits is 1/.5590 = 1,79 days in the first segment and 1/.0476 = 21 days in the second segment.

References

Kim. J.. Allenby. G. M.. & Rossi . P. E.. Volumetric Conjoint Analysis (May 2004). Available at SSRN: https://ssrn.com/abstract=552862 or http://dx.doi.org/10.2139/ssrn.552862

Malmberg J.-O. (2015) Peruskoulut ja lukiot - Julkaisujen valokopiointi ja skannaus – internetaineistojen tulostaminen ja tallentaminen 2014, 19.2.2015: https://www.kopiosto.fi/kopiosto/Tutkimus/oppilaitokset/fi_FI/oppilaitokset/_files/95315738322487613/default/Pelu2014_raportti_Lopullinen.pdf, retrieved 16 September 2018

McLachlan. G. J.. & Peel. D. (2006). *Finite mixture models*. New York. New York: John Wiley & Sons.