Measuring an equivalence of purchase intervals using a revised Gini index

Abe, Hiroyasu
Doshisha University, Graduate School of Culture and Information Science
Kyoto (610-0394),Japan
E-mail: dik0001@mail4.doshisha.ac.jp

Terada, Yoshikazu
Osaka University, Graduate School of Engineering Science, Division of Mathematical Science, Department of Systems Innovation
Osaka (560-8531),Japan
E-mail: terada@sigmath.es.osaka-u.ac.jp

Yadohisa, Hiroshi
Doshisha University, Department of Culture and Information Science
Kyoto (610-0394),Japan
E-mail: hyadohis@mail.doshisha.ac.jp

1. Introduction

In recent years, one-to-one marketing has attracted much attention among various marketing disciplines(Peppers and Martha, 1993). The core principle of marketing is to establish and maintain interaction between a company and its customers by knowing their characteristics and needs. To do so, companies must collect data about customers' behavior, such as their purchasing histories, and then analyze and evaluate important features presented by the data. One-to-one marketing employs several indices for evaluating customer data. For example, Abe (2005) proposed a behavioral model to predict the probability that an individual customer will remain a loyal consumer.

This paper proposes a new method–the adjusted Gini index (AGI)–for using the Gini index, proposed by Gini (1912), to evaluate customers' purchasing interval. Simulation results demonstrate that AGI is a reliable indicator for evaluating the regularity of a customer's purchases.

2. Measuring the equivalence of purchase intervals

The definitions and notation used to measure purchasing equivalence are as follows. Let x_{ij} be the jth purchase interval data of customer i. It is represented as

$$x_{ij} = t_{i,\,j+1} - t_{i,\,j} \quad (j = 1,\ \dots,\ n_i - 1)\,,$$

where $t_{i,\,j}$ is the dates of the jth purchase by customer i, and $n_i\,(\geq 3)$ is the number of purchases for customer i.

We can measure the equivalence of purchase intervals for customer i by calculating the Gini index GI_i for the purchase interval data using

$$GI_i = 1 - 2\sum_{j=1}^{n_i-1}\int_{\frac{j-1}{n_i-1}}^{\frac{j}{n_i-1}} f_{ij}\,(p)\,dp,$$

$$f_{ij}\,(p) = \frac{x_{i,\,(j)}}{\bar{x}_i}\left(p - \frac{j}{n_i-1}\right) + \frac{\sum_{k=1}^{j} x_{i,\,(k)}}{(n_i-1)\,\bar{x}_i} \quad \left(p \in \left[\frac{j-1}{n_i-1},\ \frac{j}{n_i-1}\right]\ ;\ j = 1,\ 2,\ \dots,\ n_i - 1\right)$$

where $x_{i,\,(j)}$ are the order statistics of x_{ij} and

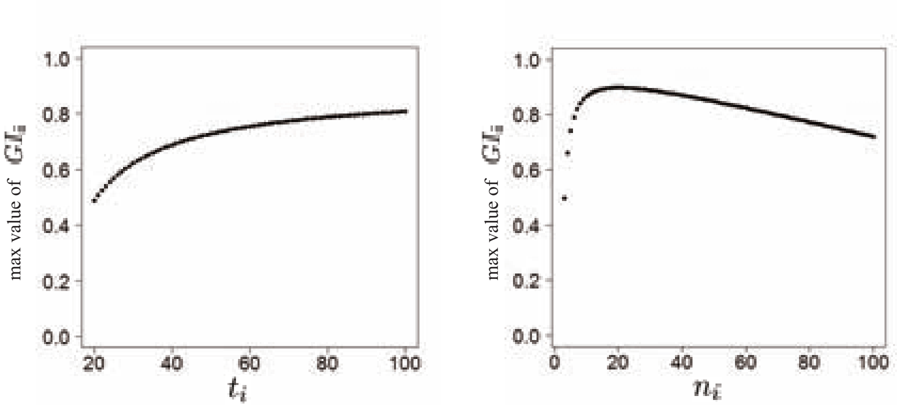
$$\bar{x}_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i - 1} x_{ij}.$$

The continuous function $f_i\,(p)\quad (p \in [0,\ 1])$ obtained by combining the linear function $f_{ij}\,(p)$ for $(j = 1,\ 2,\ \ \ldots,\ n_i - 1)$ is the Lorenz curve.

GI_i represents the uniformity of the purchase interval for a customer i. If GI_i is closer to 0, the customer tends to purchase at regular intervals; if it is closer to 1, the customer tends to purchase at irregular intervals.

However, GI_i poses a problem: its maximum value depends on the time elapsed between the customer's first and most recent purchase $t_i = t_{i,\,n_i} - t_{i,\,1}$, and n_i. The maximum GI_i is the value of Gini index when $n_i - 2$ purchase interval data are 1 and the only one data is $t_i - (n_i - 2)$. This problem is caused by the feature of the purchase interval data, i.e., $x_{ij} \neq 0$.

Figure 1: Relationship between the maximum value of GI_i and t_i and n_i



The left panel of Figure 1 shows the relationship between t_i and the maximum value of GI_i when $n_i = 10$. The right panel shows the relationship between n_i and the maximum value of GI_i when $t_i = 365$. The graph implies that the maximum value of GI_i depends on a customer i.

Therefore, GI_i needs to be modified by the maximum value of the Gini index $GI_{\max}\,(i)$, calculated as

$$GI_{\max}\,(i) = 1 - \frac{(n_i - 2)\,(n_i - 1) + t_i}{t_i\,(n_i - 1)}.$$

AGI_i is defined as

$$AGI_i = \frac{GI_i}{GI_{\max}\,(i)}.$$

3. Time series of the equivalence

A problem in evaluating the equivalence of purchase intervals is that we cannot measure their local equivalence by calculating the Gini index using all historical data. For example, if a customer makes a purchase every day during the first month, AGI_i calculated for that month is 1. However, if the customer makes only one purchase during the second month, AGI_i calculated for the two months is 0. Although this is an extreme example, it indicates the need of calculating AGI_i using data for the appropriate time frame.

In calculating local AGI_i, we define the time frame which is shorter than the observing period, and then we calculate AGI_i using data in the time frame with moving the time frame between the first and last time of the observing period. With this method, we can derive the time-series AGI_i.

Purchase data between t_s and t_f were observed and F was selected as the time frame for calculating AGI_i; AGI_i is calculated at $t_s + F,\ t_s + F + 1,\ \ldots,\ t_f$. Let T denote the endpoint of the calculation period, let $n_i^T\ (\geq 3)$ denote the number of times customer i makes a purchase during the period, and let $J_i^T = \left\{\ell_1,\ \ell_2,\ \ldots,\ \ell_{n_i^t}\right\}$ denote the ordinal numbers regarding purchase in the time frame.

We can then write AGI_i^T, which is denotes AGI_i at T as

$$AGI_i^T = \frac{GI_i^T}{GI_{\max}^T(i)}$$

where

$$GI_i^T = 1 - 2\sum_{j=1}^{n_i^T-1} \int_{\frac{j-1}{n_i^T-1}}^{\frac{j}{n_i^T-1}} f_{ij}^T(p)\, dp \qquad GI_{\max}^T(i) = 1 - \frac{\left(n_i^T - 2\right)^2 + t_i^T}{t_i^T\left(n_i^T - 1\right)}$$

$$f_{ij}^T(p) = \frac{x_{i,\,(\ell_j)}}{\bar{x}_i^T}\left(p - \frac{j}{n_i^T - 1}\right) + \frac{\sum_{k=1}^{j} x_{i,\,(\ell_k)}}{\left(n_i^T - 1\right)\bar{x}_i^T} \quad \left(p \in \left[\frac{j-1}{n_i^T - 1},\ \frac{j}{n_i^T - 1}\right]\ ; j = 1,\ 2,\ \ldots,\ n_i^T - 1\right).$$

Here, $x_{i,\,(\ell_j)}$ are the order statistics of $x_{i,\,\ell_j}$ $\left(j = 1,\ 2,\ \ldots,\ n_i^T - 1\right)$ and $\bar{x}_i^T$ is the mean of the purchase interval during the time frame, given as

$$\bar{x}_i^T = \frac{1}{n_i^T - 1}\sum_{j=1}^{n_i^T-1} x_{i,\,\ell_j}.$$

4. Application 1

This section shows the validity of AGI_i in measuring the equivalence of purchase intervals by using simulation. We compare AGI_i of regular purchase interval data with that of irregular data. If AGI_i derived from the former data is large and that derived from the latter data is small, it can be concluded that AGI_i is valid.

First, by assuming that regular and irregular purchase interval data follow a normal and an exponential distribution, respectively, we generate the data by using normal and exponential random number generations, respectively.

Then, we generate one set of random numbers, the sum of which is less than 365. That is, a customer's purchase is observed in the period of 365. If a negative number is generated by the normal random number, the value is rejected. Second, we consider one set of random numbers to be the purchase data of one customer and repeat the generation until the number of sets is 3000, rejecting data sets that contain fewer than 10 entries. Finally, we calculate AGI_i for each set and estimate its kernel density.

This experimental design has four simulation patterns. The expected value and the variance are different for both normal distribution and exponential distribution from which the random numbers are generated. That is, they differ with respect to the length of the purchase interval. The parameter values of the normal and exponential distributions are set such that the expected values of both are 40, 30, 20, and 10. The variance values of the normal distribution are set as $(80/3)^2$, $(20)^2$, $(40/3)^2$, and $(20/3)^2$, respectively. (Variances of the exponential distribution are fixed when its expected values are fixed.)

Figure 2: Comparison of density estimates for AGI_i

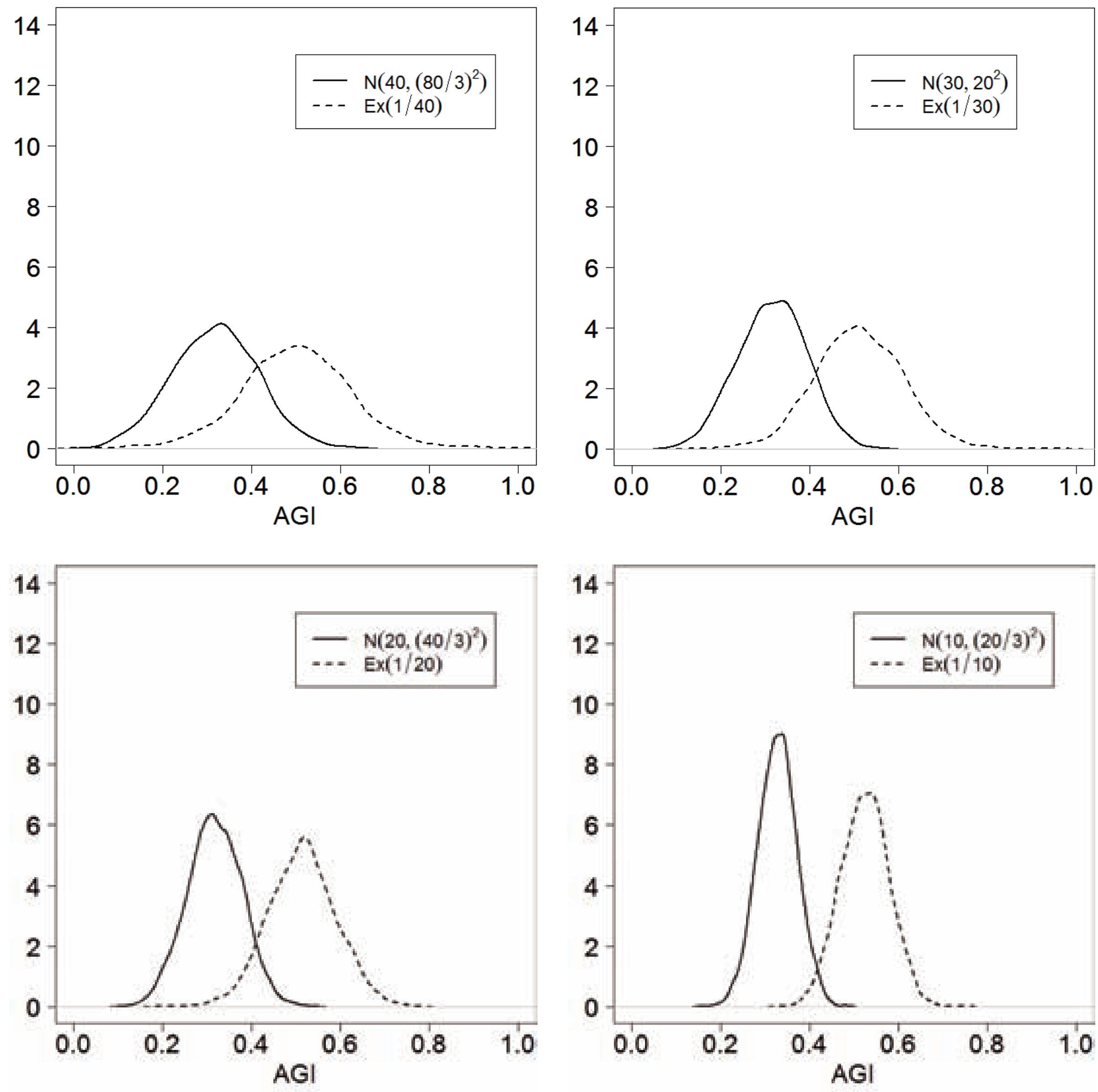


Figure 2 shows the result of the kernel density estimate for AGI_i in the four simulation patterns. From this graph, we observe that in any simulation pattern AGI_i calculated from the exponentially-distributed purchase interval is larger than that calculated from the normally-distributed interval. This indicate that AGI_i is small when a customer purchases at regular intervals and is large when the customer purchases at irregular intervals. This result shows that AGI_i is valid for measuring the equivalence of purchase intervals.

4. Application 2

A simulation similar to that in Application 1 shows the validity of AGI_i^T in measuring the time-series equivalence of purchase intervals.

First, we obtain simulation data by random number generation as in Application 1. However, in this simulation we create only 1000 sets of purchase interval data and establish the observed period as 547. Then, we calculate AGI_i^T with each set of purchase interval data. Here, we fix the value of F as 7 times the mean of the purchase interval data in each set. Moreover, if fewer than five data points appear during the time frame, we do not calculate AGI_i^T. In this manner, we produce 1000 sets of AGI_i^T compensating for the missing values.

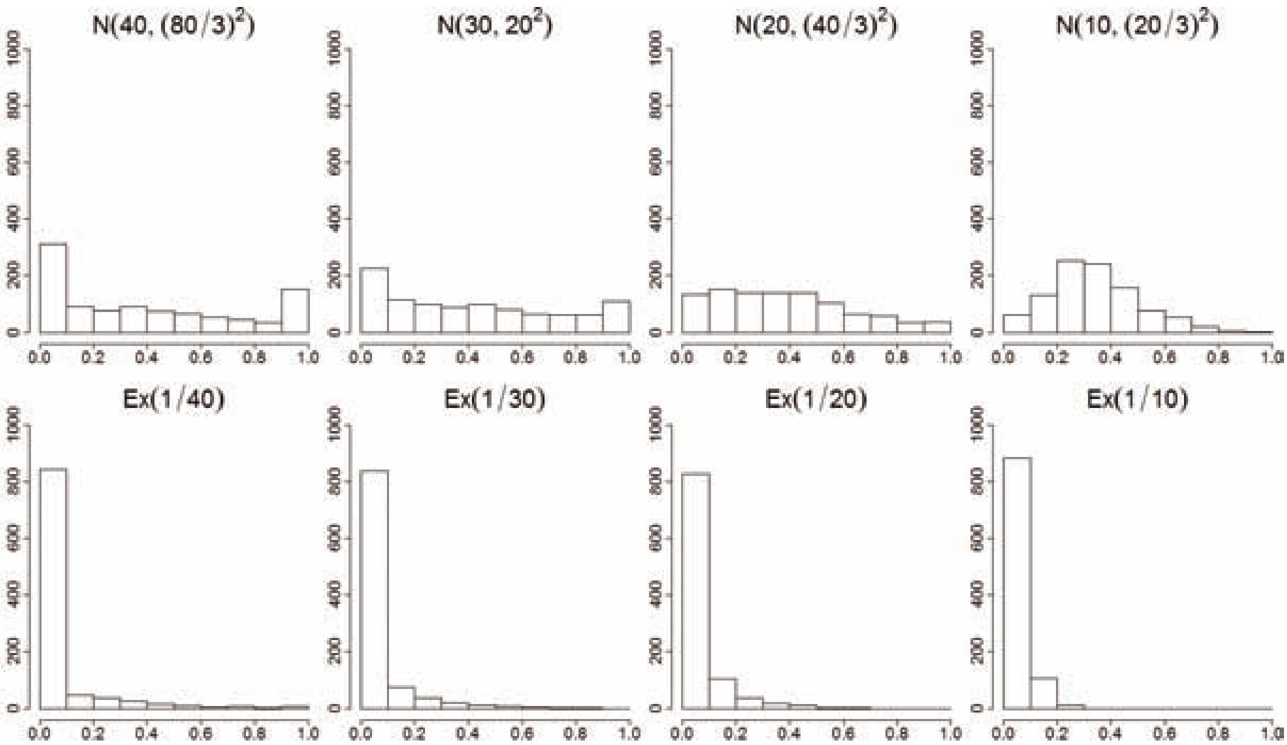
Because we cannot observe all graphs for checking the validity of AGI_i^T, we define R_i^τ as

$$R_i^\tau = \frac{\sum_{T=t_s+F}^{t_f} B_i^T}{\sum_{T=t_s+F}^{t_f} A_i^T}, \text{ where } A_i^T = \begin{cases} 0 & (n_i^T < 5) \\ 1 & (n_i^T \geq 5) \end{cases} \text{ and } B_i^T = \begin{cases} 0 & (AGI_i^T > \tau) \\ 1 & (AGI_i^T \leq \tau) \end{cases},$$

to count the number of times AGI_i^T indicates that customer i purchases at regular intervals. $\sum_{T=t_s+F}^{t_f} A_i^T$ denotes the number of times AGI_i^T can be calculated. $\sum_{T=t_s+F}^{t_f} B_i^T$ denotes the number of times the calculated AGI_i^T is lower than τ (τ is optional). A customer's purchases are considered to be regular when AGI_i^T is lower than τ. R_i^τ ranges from 0 to 1. If R_i^τ is closer to 0, it indicate that customer i purchases at irregular intervals throughout the observed period. If R_i^τ is closer to 1, it indicates that customer i purchases at regular intervals throughout the observed period.

We compare R_i^τ calculated using normally distributed random numbers with that calculated using exponentially distributed random numbers in the four patterns in which the parameters of the density functions differ. Here, we set $\tau = 0.3$.

Figure 3: Comparison of the histogram of $R_i^{0.3}$



The four upper graphs in Figure 3 show the histogram of $R_i^{0.3}$ for normally distributed data, and the four lower graphs show the histogram of $R_i^{0.3}$ for exponentially distributed data. $R_i^{0.3}$ for normally distributed data have disparate values, but most $R_i^{0.3}$ for exponentially distributed data have values closer to 0. This indicates that AGI_i^T of a customer who purchases at irregular intervals is small; hence, it may be said that the customer purchases at irregular intervals during any period.

Figure 4 and 5 show examples of purchase timing data and AGI_i^T. The broken lines represent the timing of purchases and the solid lines represent AGI_i^T. In Figure 4, purchase intervals where $R_i^{0.3} = 1$ are almost regular, and AGI_i^T is small in any period for which it is calculated. In Figure 5, purchase intervals where $R_i^{0.3} = 0$ are irregular and AGI_i^T is large. From the examples, we observe that AGI_i^T effectively represents the equivalence of purchase intervals.

Figure 4: Example of purchase timing data and AGI_i^T ***where*** $R_i^{0.3} = 1$

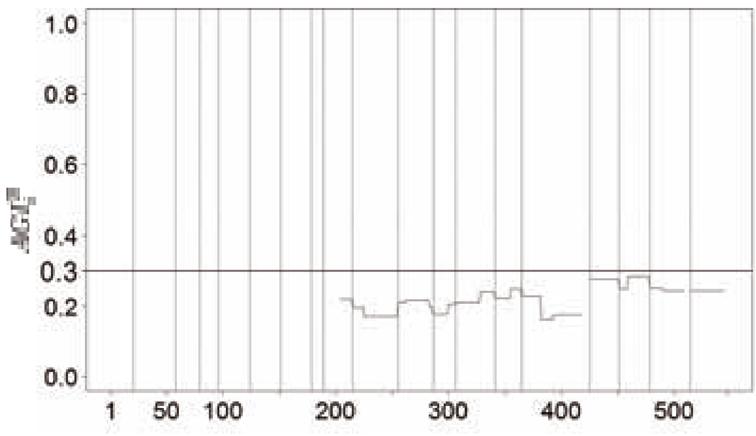
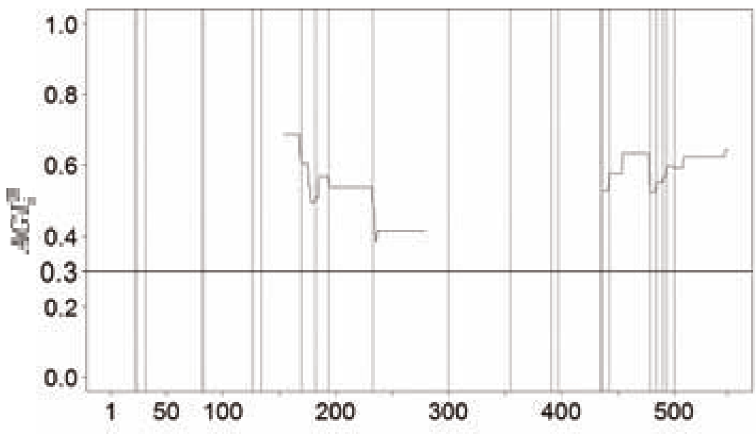


Figure 5: Example of purchase timing data and AGI_i^T ***where*** $R_i^{0.3} = 0$



5. Conclusion and outlook

This paper has proposed method using the Gini index to evaluate the regularity with which customers make purchases, and has confirmed its validity by simulation. The results show that the equivalence of purchase interval can be measured using an adjusted Gini index AGI_i and the time series of the adjusted Gini index AGI_i^T.

Using the measuring method, we can identify loyal customers by the regularity of their purchases and implement appropriate marketing actions. Our future task is to discuss the marketing actions.

REFERENCES

Gini, C. (1912), Variabilita e mutabilita, Studi Economico-Giuridici dell'Universita di Cagliari, 3, 1-158.

Makoto, A. (2005), Counting Your Customers One by One : An Individual Level RF Analysis Based on Consumer Behavior Theory, Discussion Paper F series (CIRJE-F-408), UT Repository, The University of Tokyo.

Peppers, D. and Martha, R. (1993), The One to One Future, Currency Doubleday.