

신용카드 이용행태의 확률분포 도출과 활용

이찬경** · 노형봉**

*홍익대학교 대학원 경영학과

**홍익대학교 경영대학

Derivation and utilization of probability distribution of credit card usage behavior

Chan-Kyung Lee** · Hyung-Bong Roh**

*School of Management, Hongik University

**Department of Management, Hongik University

ABSTRACT

Purpose: To find out the appropriate probability distribution of credit card usage behavior by considering the relationship among income, expenditure and credit card usage amount. Such relationship is enabled by Korea's especially high penetration of credit card.

Method: Goodness-of-fit test and effect size statistic W were used to identify the distribution of income and credit card usage amount. A simulation model is introduced to generate the credit card transactions on individual user level.

Result: The three data sets for testing had either passed the chi-square test or showed low W values, meaning they follow the exponential distribution. And the exponential distribution turned out to fit the data sets well. The r values were very high.

Conclusion: The credit card usage behavior, denoted as the counts of users by usage amount band, follows the exponential distribution. This distribution is easy to manipulate, has a variety of applications and generates important business implications.

Keyword: Credit card usage, Probability distribution, Simulation

● Received 3 January 2018, 1st revised 22 February, accepted 23 February 2018

† Corresponding Author(mccoy2009@naver.com)

© 2018, The Korean Society for Quality Management

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0>) which permits unrestricted non-Commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. 서 론

신용카드업은 국내 민간 소비 시장에서 큰 비중을 차지하고 있다. 2014년 단기카드대출과 기업구매카드 실적을 제외한 신용카드 이용금액이 전체 민간 소비 지출에서 차지하는 비중은 60%에 달했다(CREFIA, 2014). 이것은 소비 지출의 60%가 신용카드로 이루어지고 있음을 의미한다. 이 같은 높은 신용카드 결제 비중은 호주(18%), 미국(28%), 캐나다(41%) 등을 크게 앞서는 것이다(Lee 2014).

국내 소비자들의 신용카드 보유비율과 이용건수도 세계 최상위권에 속한다. 2016년 우리나라의 신용카드 보유비율은 93%로서 독일(32%), 미국(72%), 캐나다(82%) 등을 상회하였다(BOK, 2014). 또한 월 평균 신용카드 이용건수는 16건으로 호주(7.9건), 미국(8건), 캐나다(9.4건)을 크게 앞섰다. 이처럼 국내 신용카드업은 높은 보급률과 세계적인 수준의 이용행태를 특징으로 한다.

이에 따라 소비자 경제활동에 대한 심도 있는 이해를 위해서는 신용카드의 이용행태, 특히 개인 소비자 단위의 이용행태에 대한 연구가 필수적인 요소가 되었다. 그러나 여러 가지 제약 때문에 가용한 데이터는 부족한 실정이다. 신용카드사들은 방대한 거래 데이터를 축적하고 있으나 일정 기간 동안의 매출 총액 정도만 외부에 공개하고 있고 개인 소비자의 이용 자료는 공개하지 않는다. 또 2011년 제정된 개인정보보호법에 의거하여 개인 식별 가능 자료는 원칙적으로 공개가 불가하다.

빅데이터 자원 공유를 기치로 삼아 개원한 서울시 빅데이터 캠퍼스는 다양한 정보를 제공하고 그 중에는 실제 자료에 기반을 둔 신용카드 추정매출액 데이터도 있다. 그러나 위와 같은 한계 때문에 데이터 집계 최소 단위는 개인 소비자가 아니라 지역 블록이다. 그 결과 개인 소비자의 신용카드 이용행태를 연구할 수 있는 공개 데이터는 없는 실정이다. 이런 실정은 국내·외 구별 없이 공통적인 것이어서 신용카드 이용행태 관련 논문은 전체 이용금액 만으로 연구 가능한 주제에 한정되어 있다. 예로서 신용카드 이용금액과 민간 소비 지출간의 시계열적 관계를 이용한 거시 경제 지표 연구나(Kim and Yeom, 2015) 소득공제 제도의 효과 분석(Song and Sung, 2012) 등이 그것이다.

과거 일부 연구자들은 개인별 신용카드 이용 자료를 확보하여 이용행태 속성을 모형화한 바 있다. Till and Hand(2003)는 1,977명의 신용카드 사용자가 10개월 동안 일으킨 139,326건의 거래를 분석하여 카드 연체 상태변화를 마코프 체인으로 묘사하였다. So and Thomas(2011)는 홍콩의 한 은행이 보유한 140만 카드의 60개월 이력 데이터를 기반으로 신용점수와 한도 변화를 마코프 의사결정 프로세스 모형으로 설명하였다. Lee et al.(2002)은 개별 소비자의 이탈 행태를 모형화 하였고 Jung et al.(2013)은 국내 한 카드사의 실제 데이터를 이용하여 일시불 비중, 할부개월수별 분포, 평균 잔고 등의 지표들을 계산하여 제시하였다.

그러나 이들의 연구 주제는 잔고, 상환, 이체 등 신용관리 중심의 이용행태에 치중되어 있고 민간 소비 지출 분석에 직접 연관된 사용실적 중심의 이용행태를 모형화 하려는 시도는 제한적이었다. 그 이유는 과거에는 신용카드를 금융상품으로만 인식했기 때문이다. 최근에는 신용카드를 일상생활에 필수적인 지불 결제 수단으로 보는 시각이 증가했고, 그 결과 사용실적 중심의 이용행태가 중요해졌다. 그런 상황 속에서 실제 데이터가 공개되지 않는 현실을 고려해 볼 때 이용행태의 속성을 확률분포로 묘사하고 모형화하는 연구는 그 중요성이 매우 크다고 볼 수 있다.

그러나 신용카드사들은 여전히 데이터마이닝 위주의 분석 성향을 유지하고 있다. Lee et al.(2014)은 국내 굴지의 한 카드사의 빅데이터 활용에 관한 과거 이력과 현황을 조사하였는데 이용실적을 향상시키기 위한 수많은 방법론이 적용되고 있었다. 모든 혜택을 단일 카드에 담은 One카드 전략, 고객의 특정 행동을 신호로 전개되는 이벤트 마케팅, 미리 정해진 규칙에 의해 실행되는 실시간 프로모션 등 다양한 차원의 데이터 활용 사례가 있었다. 그러나 확률분포 모형이나 시뮬레이션이 활용된 사례는 없었다. 이로부터 카드사들의 분석방법론은 2000년대 CRM 시대에 정립된

데이터베이스 마케팅 방법론의 영향과 거래 데이터를 전수 확보하고 있다는 사실 때문에 여전히 데이터마이닝 분야에 치중하고 있음을 알 수 있다.

전술한 바와 같이 개인별 신용카드 이용실적 자료를 구하기 어렵고, 신용관리 중심의 이용행태가 주요 관심 주제였고, 데이터마이닝이 분석 방법론의 지배적인 위치를 차지하고 있는 점 등으로 인해 확률분포를 기반으로 한 이용행태 분석을 다룬 논문은 거의 찾아보기 힘든 실정이다. 하지만 확률분포 모형은 데이터마이닝과 차별화되는 고유한 효용이 있다. 확률분포를 통해 소비자들의 속성을 설명할 수 있고, 실제 이용행태를 묘사하는 시뮬레이션 모형도 구축할 수 있으며, 궁극적으로 신용카드 서비스품질 관리의 도구로도 활용이 가능하기 때문이다. 특히 파라미터 변동으로 이용행태의 변화를 묘사함으로써 예측과 시나리오 분석을 효과적으로 수행할 수 있다.

따라서 본 연구에서는 개인별 신용카드 이용행태의 속성을 설명하고 묘사할 수 있는 확률분포를 찾고 그 활용분야를 탐색하고자 한다. 가용한 개인별 데이터는 존재하지 않으나, 민간 소비 지출의 절반 이상을 신용카드가 차지하고 있는 특수성 때문에 참고 가능한 자료들이 존재한다. 또한 선행연구들의 결과를 조합하고 국내 실정에 맞게 변용하여 분포의 도출에 활용하였다.

구체적인 연구 대상은 개인 소비자 단위의 사용실적 중심의 신용카드 이용행태이다. 이용행태에는 여러 가지가 있는데 데이터베이스 마케팅의 기본요소인 RFM, 즉 최근 이용시점(recency), 이용빈도(frequency), 이용금액(monetary)으로 정의하는 것이 합리적이다. 또한 개인 소비자 단위이므로 분석 대상 거래는 (1) 개인 신용카드 거래, (2) 신용판매 거래 (일시불+ 할부), (3) 일정 기간 동안의 거래로 정의하기로 한다.

범위를 개인 신용카드로 한정하였으므로 법인카드 거래는 배제하였다. 법인카드는 사전 계약 B2B 거래, 계약 가맹점 중심 거래 등 인위적인 이용행태가 포함되므로 배제하였다. 거래 범위를 신용판매로 한정한 것은 민간 소비 지출 내에서의 행태만을 고려하였기 때문이다. 그 외의 거래, 즉 현금서비스와 카드로는 속성상 대출에 해당하므로 일반 소비지출에서 제외한다. 또한 일정 기간 동안의 거래로만 한정된 것은 이탈의 영향을 배제한 것이다. 즉 개별 사용자 단위로 보면 매 기간마다 카드 이용 가능 상태에 변동이 있을 수 있지만, 일정 시점 내에서 특정 사용자 집단의 총수는 큰 변동이 없고, 고정 집단으로 간주할 수 있기 때문이다.

종합적으로, 본 연구의 목적은 신용카드 이용금액별 사용자 수의 확률분포를 찾는 것이다. 이용금액별 사용자 수는 RFM이 집적된 최종 결과물이기 때문에 이의 확률분포를 확인하는 것은 곧 이용행태를 모형화 하는 것이다. 확인된 확률분포의 파라미터는 소비자의 신용카드 이용행태를 묘사하는 값이고 이 값의 변화를 통해 총 이용금액에 미치는 영향을 계량화할 수 있는데, 총 이용금액은 궁극적으로 비즈니스 의사결정의 최종 목표 중 하나이다.

본 논문은 다음과 같은 단계로 진행한다. 2장에서는 소득-소비지출-신용카드 이용금액간의 관계를 통해서 신용카드 이용금액의 확률분포를 제시한다. 또한 도출된 확률분포 모형을 정량적으로 검증한다. 3장에서는 모형의 활용방안이 제시된다. 마지막으로 4장(결론)에서는 이 분야의 연구자들과 신용카드사들이 확률분포 모형을 활용하는 것이 매우 유용함을 보인다.

2. 신용카드 이용금액의 확률분포

2.1. 국내 소득자료의 확률분포

우리나라는 신용카드 이용금액이 소비 지출의 절반 이상을 차지하고, 소비 지출의 원천은 소득이므로 소득-지출-신용카드 이용금액 간에 밀접한 관계가 존재할 것으로 예상된다. 이에 소득분포의 속성부터 살펴보고자 한다. 소득금액별 인원수의 확률분포(이하 소득분포)에 대한 연구는 예전부터 존재하였다. Pareto(1897)는 파레토 분포를 이용한 소득분포 모형을 수립하였고, Champernowne(1953)은 감마분포 모형을 제안하였다. 최근에는 Drăgulescu and Yakovenko(2001)가 미국의 소득자료를 이용하여 소득분포가 지수분포를 따름을 밝혔다.

지수분포는 다른 분포들과 달리 파라미터가 하나이므로 해석이 용이하다. 또한 수많은 자연현상을 설명하는 데 유용한 분포라는 사실이 밝혀져 있다. 특히 확률과정의 기본 모형 중 하나인 포아송 과정(Poisson Process)에서 지수분포는 일련의 사건들의 발생시각을 적절히 묘사하는 것으로 알려져 있다. 이처럼 지수분포는 다양한 장점을 갖고 있으므로 국내 소득분포 자료의 확률분포에 대한 탐구는 지수분포를 따르는지의 여부를 검증하는 것으로 한정하고자 한다.

지수분포의 파라미터 $\lambda (> 0)$ 는 분포의 모양을 결정하는데, 포아송 과정에서 사용될 때는 ‘단위 시간 동안의 평균 사건발생빈도’로 해석된다. 그래서 λ 를 rate parameter라고도 부른다. 지수분포의 확률분포함수를 살펴보면 다음과 같다.

$$\text{누적확률분포: } P(X \leq x) = 1 - e^{-\lambda x} \quad (x \geq 0) \quad (1)$$

$$\text{확률밀도함수: } P(X = x) = \lambda e^{-\lambda x} \quad (x \geq 0) \quad (2)$$

$$\text{평균과 분산: } E(X) = \frac{1}{\lambda}, V(X) = \frac{1}{\lambda^2} \quad (3)$$

국내 소득분포 자료는 국가통계포털(KOSIS)에서 제공하는 ‘근로소득 연말정산 신고 현황(4.2.4)’에서 참조할 수 있다. 이것은 소득금액대별 인원수로서 테이블 형태의 이산형 자료이다. 따라서 식(1)~식(3)에서 X = 소득금액대(income band)로 정의하는 것이 직관적이다. 그러나 지수분포는 연속형 확률분포이므로 λ 가 결정되고 분포 적합이 이루어지면 X = 소득금액(income)으로 볼 수 있다. 이를 신용카드 이용금액별 사용자 수 확률분포에 적용하면 X = 신용카드 이용금액으로 정의할 수 있다.

KOSIS 소득자료에는 연간 소득금액대별 인원수가 있고, 급여소득과 비과세소득이 별도 집계되어 있다. Drăgulescu and Yakovenko(2001)는 U.S. Census 데이터인 SIPP 공식자료 중 TPTOINC 항목, 즉 전체 소득 자료를 사용하였는데 KOSIS 자료의 급여소득과 비과세소득을 합친 것과 동등하다고 판단된다. KOSIS 자료는 금액대가 동일 간격이 아니므로 동 간격으로 변형하면 Table 1과 같다.

Table 1. Distribution of head count by income band (KOSIS, 2015)

Income Band (in 10mil. KRW)	Actual head count (O)	Income (in 10mil. KRW)
0~2	7,657,059	8,043,495
2~4	4,858,285	14,123,679
4~6	2,348,297	11,658,253
6~8	1,303,229	9,114,428
8~10	570,400	5,131,348
Above	596,124	9,023,670
Total	17,333,394	57,094,872

이것은 1,733만 명에 달하는 원 데이터를 요약 집계한 자료이다. 따라서 소득분포가 지수분포를 따르는지 밝히기 위해서는 카이제곱 통계량을 이용한 적합도 검정을 해야 한다. 카이제곱 통계량은 지수분포로부터 계산한 이론적인 인원수(E)와 실제 인원수(O)를 아래와 같이 연산한 통계량이다.

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad (4)$$

O_i = 실측값, E_i = 이론적 기댓값

그런데 적합도 검정은 개체수가 너무 큰 경우 효과적이지 못한데 (Bentler and Bonnet, 1980) 위 자료는 1,733만 명으로 개체수가 매우 크다. 따라서 실제 속성과 관계없이 귀무가설(H_0 : E와 O 사이에는 유의적인 차이가 없음)이 가각된다. Cohen(1988)은 이런 경우 효과크기(Effect Size) 척도인 W를 참고하는 것이 적절하다고 설명하고 있다. 효과크기는 개체수의 영향을 배제한 실질적인 차이를 계량하는 척도인데, W가 크면(작으면) 두 집단 간에 실질적인 차이가 크다(작다)는 의미이다.

$$W = \sqrt{\frac{\chi^2}{n}} \quad (5)$$

KOSIS 소득자료에서 비교해야 할 두 집단은 소득금액대별 실제 인원수(O)와 지수분포로 적합하여 산출한 기대인원수(E)이다. 카이제곱 통계량은 Table 2와 같이 계산된다. 식(5)에서 n = 개체수 = 17,333,394이다. 이 두 값으로부터 W를 계산할 수 있다. W가 작을수록 두 집단 간에 실질적인 차이가 적어 지수분포에 가깝다고 판정할 수 있다. W값의 해석에 절대적인 기준은 없으나, Cohen은 0.1을 약함(weak), 0.3을 중간(mediaum), 0.5 이상은 강함(strong)으로 판정할 수 있다고 제안하였다. 따라서 W가 0.1 이하이면 두 집단 간에 실질적인 차이가 아주 작기 때문에 지수분포를 따르지 않는다고 하기에는 설득력이 약하다고 볼 수 있다.

Table 2. Calculation of Chi-square statistic and W

Income Band (in 10mil. KRW)	Actual head count (O)	Expected head count (E)	$\frac{(O-E)^2}{E}$
0~2	7,657,059	8,193,627	35,138
2~4	4,858,285	4,320,437	66,956
4~6	2,348,297	2,278,134	2,161
6~8	1,303,229	1,201,243	8,659
8~10	570,400	633,406	6,267
Above	596,124	706,547	17,258
Total	17,333,394	17,333,394	136,438

λ	X^2	W	r
0.320	136,438	0.0887	0.996

KOSIS 자료의 실제 인당 평균 소득금액은 3.294(천만원)이므로 $\lambda = 1/E(X) = 0.304$ 이다. 만약 원 자료가 0.304를 파라미터로 하는 지수분포에 정확하게 적합되면 O와 E는 서로 같은 값을 갖게 되어 카이제곱 통계량 = 0이 될 것이다. 그러나 사회현상이 이론적인 분포를 완벽하게 따르는 경우는 거의 찾을 수 없다고 할 수 있다. 따라서 KOSIS 자료를 여러 가지 λ 값의 지수분포와 비교하는 민감도분석을 수행하였고, 그 결과 KOSIS 자료는 $\lambda = 0.320$ 일 때 W가 0.0887로 최소가 됨을 알 수 있다. 이것은 Cohen의 약함(weak) 기준을 하회한다. 따라서 Table 2의 O와 E간에 실질적인 차이는 크지 않고, 소득분포는 $\lambda = 0.320$ 의 지수분포와 가깝다고 결론을 내릴 수 있다.

주어진 자료가 지수분포에 적합한 지에 대한 척도가 카이제곱 통계량과 W라면, 반대로 지수분포가 주어진 자료를 얼마나 잘 설명하는지에 대한 척도도 있다. 그것은 상관계수 r로서 특정 확률분포 모형이 데이터와 얼마나 선형적 관계가 강한지를 의미한다. 다만 지수분포는 곡선이므로 실제인원수와 기대인원수에 자연로그를 취하여 계산한다.

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2 \Sigma(y - \bar{y})^2}} \quad (6)$$

$$x = \ln(\text{actual headcount}), y = \ln(\text{expected headcount})$$

식(6)에 따라 계산하면 $r = 0.996$, $p\text{-value} = 0.000$ 으로서 지수분포로 산출한 인원수와 소득분포 자료의 인원수가 서로 큰 상관관계를 가짐을 알 수 있다. 그만큼 KOSIS 소득분포 자료를 지수분포로써 적절히 설명할 수 있다는 의미이다.

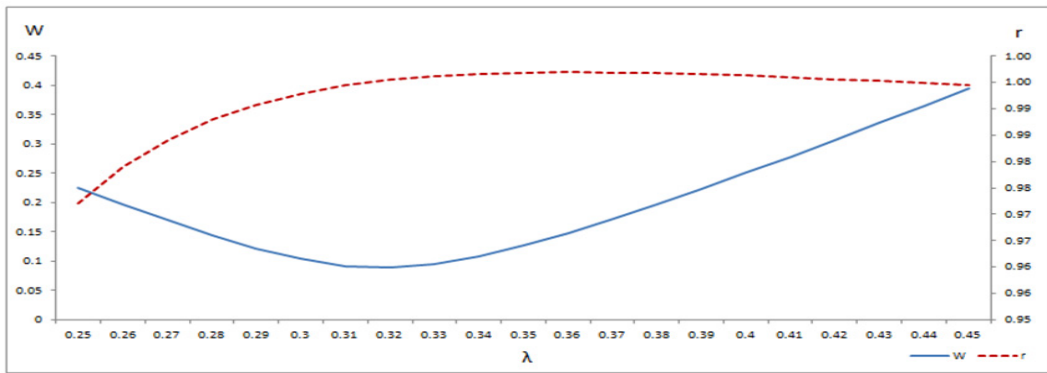


Figure 1. Distribution of W and r by λ in fitting KOSIS income data

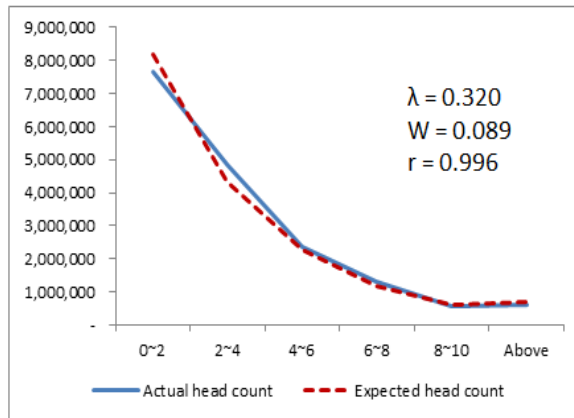


Figure 2. KOSIS Income distribution fitted by exponential distribution with $\lambda = 0.320$

λ 를 0.250부터 0.450까지 0.01 단위로 변경하면서 W와 r의 값을 관찰하는 민감도분석 결과, W가 최저가 되는 지점과 r이 최고가 되는 지점은 일치하지 않았다(Figure 1). W가 최저가 되는 지점은 평균 소득금액과 가까운 지점인 $\lambda = 0.320$ 이었으나 r이 최고가 되는 지점은 $\lambda = 0.360$ 이었다($r = 0.997$). 그러나 W는 최고-최소값 간에 차이가 큰 반면 r은 큰 차이가 없으므로 W값이 최소가 되는 지점의 λ 값을 선택하는 것이 바람직하고, 그에 따라 $\lambda = 0.320$ 을 지수분포의 최종 파라미터로 선택하였다. 적합의 정도는 Figure 2처럼 매우 가깝게 나타났다.

$\lambda = 0.320$ 에서 $r = 0.996$ 으로 매우 높은 선형적 상관관계를 유지하고 있다. 따라서 국내 소득분포 자료는 $\lambda = 0.320$ 의 지수분포에 잘 적합되며, 역으로 지수분포로써 충분히 설명할 수 있는 형태라는 결론을 내릴 수 있다. Figure 2에서 보듯 KOSIS 소득자료의 분포와 $\lambda = 0.320$ 의 지수분포를 비교하면 두 분포의 모양이 매우 비슷하게 나타난다. 이로써 국내 소득분포 자료 역시 미국의 사례와 마찬가지로 지수분포를 따른다고 볼 수 있다.

2.2 소득-지출-신용카드 이용금액의 관계

소득과 지출은 비례한다고 가정한다. 일찍이 Keynes(1936)는 소비자는 소득이 증가하면 지출을 증가시키는 경향이 있다고 주장하였다. 이를 바탕으로 경제학계는 ‘소득 = 지출 + 저축’의 관계식을 기본 가정으로 적용해 왔다. 이것은 오랜 전통을 지닌 가정이지만 세월이 지나면서 현재도 계속 유효한 가정을인지 검증해 볼 필요가 생겼다.

Diacon and Maha(2014)는 세계은행의 자료를 토대로 소득-지출-GDP간의 관계를 연구하였는데, 국가별 소득 수준에 따라 다른 경향이 나타났다. 그러나 우리나라가 속한 고소득 국가군에서는 소득과 지출이 대체로 비례하는 것으로 나타났고 이에 따라 Keynes의 가정은 여전히 유효하다고 결론 내릴 수 있다. KOSIS에서 제공하는 월평균 가계수지 자료(2006~2016)를 보아도 국내 소득과 소비지출 역시 서로 유사한 추세로 움직이는 것을 알 수 있다 (Figure 3).

소비지출과 신용카드 이용금액 역시 강한 비례관계를 보인다. 이 비례 관계는 세계 어디에서나 공통적인 것으로 추정되지만, 우리나라는 신용카드 사용금액이 민간 소비지출의 절반 이상을 차지하고 있기 때문에 그 비례관계는 더욱 더 뚜렷할 것으로 보인다. 여신금융협회의 자료(CREFIA, 2014)를 보면 2000년 ~ 2013년 민간소비지출과 신용카드 이용금액은 Figure 3과 같이 매우 유사한 추세로 변화하였음을 알 수 있다 ($r = 0.992$, $p\text{-value} = 0.000$).

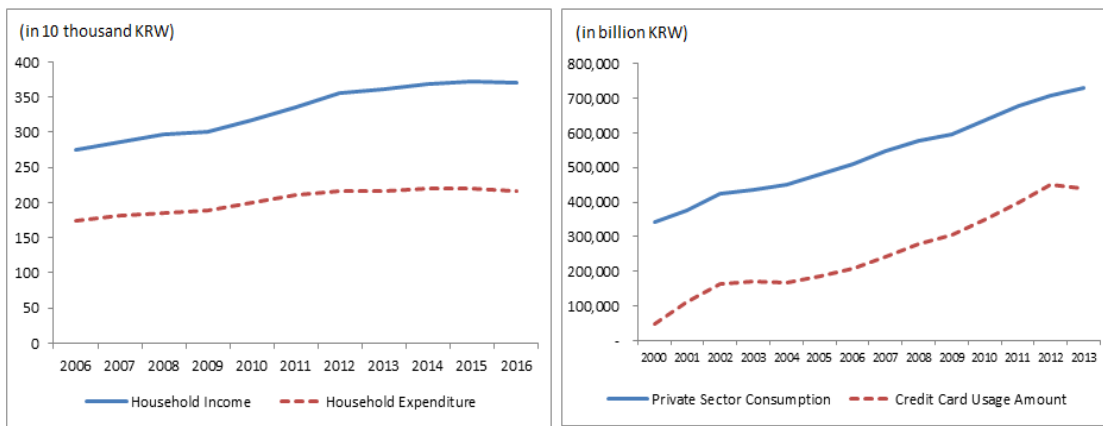


Figure 3. Proportional relationship among income, expenditure and credit card usage

2.3 신용카드 이용금액의 지수분포 적합

전술한 바와 같이 개인별 신용카드 이용금액 자료는 일반적으로 공개되지 않으나 참고할 만한 자료가 일부 존재하므로 그 각각에 대하여 살펴보기로 한다.

(1) UCI repository 데이터

머신 러닝 연구용 데이터를 제공하는 UCI repository에는 독일에서 공개한 실제 데이터(1천건)가 존재한다. 이 데이터에는 21개의 속성값이 있고 그 중 하나가 신용카드 한도금액(이하 신용한도)이다. 신용한도는 신용카드 이용금액의 상한이며 고객별로 상이하게 부여된다. 장기간 금융상품들의 이용 이력으로부터 산출된 신용점수에 근거하므로 개인 소비자별 이용행태의 기준이 된다. 또 월별로 부여되므로 신용한도 분포는 일정 기간 동안 발생한 신용판매 이용금액 분포의 바로미터가 될 수 있다.

이 데이터는 1천건으로 샘플 크기가 적당하므로 W값을 산출할 필요 없이 곧바로 카이제곱 통계량을 이용한 적합도 검정을 적용하였다. 세부 검정 절차는 RAC (Reliability Analysis Center) 에서 발행한 START Vol.10 (2003)을 참조하였다. 1천건 자료의 신용한도 평균은 3,271이다. 원 데이터에 단위는 명시되어 있지 않으나, EU 출범 이전의 자료이므로 단위는 독일 마르크(DM)로 추정된다.

Credit Limit (DM)	Actual head count	Expected head count	$\frac{(O-E)^2}{E}$
3,500	686	657	1.28
6,500	190	206	1.24
9,500	79	82	0.11
12,500	27	33	1.09
Above	18	22	0.73
Total	1,000	1,000	4.45

$$\chi^2 = 4.45 < \chi^2_{0.95,3} = 7.81$$

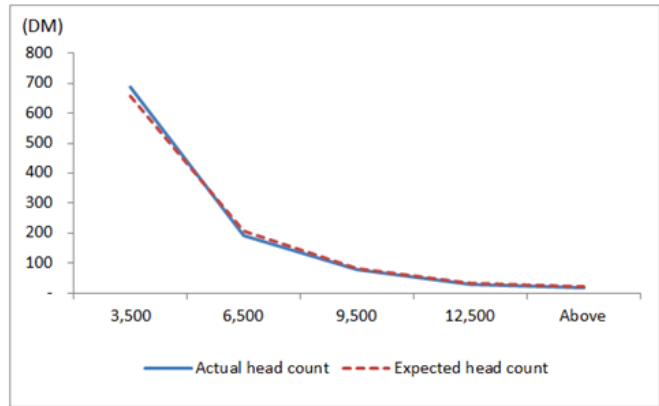


Figure 4. Fitting of UCI repository's German data to the exponential distribution

λ 의 초기값은 평균의 역수인 0.003 (1/3,271) 으로 산출하였다. Figure 4에서 보듯이 카이제곱 검정통계량은 4.45이고 이는 유의수준 5%에서의 기각역 하한 $\chi^2_{0.95,3} = 7.81$ 보다 작으므로 귀무가설(H_0 : O와 E 사이에는 유의적인 차이가 없다)를 기각할 수 없다. 즉, 신용한도의 분포는 $\lambda = 0.003$ 인 지수분포를 따르지 않는다고 볼 수 없으며 이에 따라 추가적인 민감도분석 없이 $\lambda = 0.003$ 을 최종 파라미터로 선택할 수 있다. 이때 카이제곱 분포의 자유도는

$$df = k (\text{범주의 수}) - 1 - s (\text{모수 추정 개수}) \tag{7}$$

로 정의된다. λ 라는 1개의 파라미터를 추정했으므로 $df = 5 - 1 - 1 = 3$ 이 되고, 그에 따라 유의수준 5% 하에서 귀무가설을 기각하기 위한 값은 7.81이 되었다.

(2) 서울시 빅데이터 캠퍼스 데이터

서울시 빅데이터 캠퍼스는 실제 신용카드 이용실적을 기반으로 한 추정매출액을 총 37,045개의 지역 블록단위별로 제공하고 있다. 추정매출액은 두 카드사가 개별적으로 집계한 지역별 매출액을 합산하여 평균한 후 일정한 가중치를 적용, 산출한 값이다. 비록 추정치이기는 하나 실제 매출액에 기반을 두어 매우 현실과 가까운 값이다.

이렇게 산출된 신용카드 추정매출액을 2016년 월별로 집계한 후 계절지수(=해당월 매출/연 평균 매출)를 산출해 보면 최소 0.96, 최대 1.10로서 그 차이가 크지 않다 (2월 지수 0.91은 제외함). 따라서 국내 신용카드 이용금액은 연간 고르게 발생한다는 것을 알 수 있다. 그럼에도 명절, 휴가 등으로 인해 계절성이 클 것으로 보이는 시기들을 피해 2016년 6월 매출액을 지역 블록별로 집계해 보았다. 그 결과가 Figure 5이다.

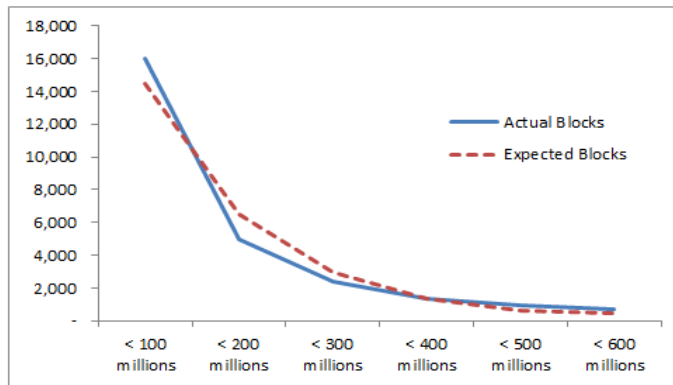


Figure 5. Credit card usage amount by regional block

일반적으로 지역별 매출은 개인별 이용금액과는 속성에 차이가 있다. 어떤 지역은 상업시설이 없어 카드 매출이 거의 발생하지 않는 반면 어떤 지역은 상업시설 초밀집 지구로서 높은 매출이 발생할 수 있다. 그러므로 지역별 편차는 개인별 편차보다 훨씬 크다. 또 자료 공급원인 카드사별로 지역 구분 기준이 다를 수도 있다.

이런 사정들을 감안하여 월 매출이 1천만 원 미만인 지역 블록 7,956개와 6억을 초과하는 2,743개 블록은 분포 산출 대상에서 제외하였다. 이렇게 양 극단값을 제거한 후 총 26,346개의 지역 블록에 대해 매출금액대별 블록수 분포를 산출해본 결과, 파라미터 $\lambda = 0.80$ 일 때 $W = 0.182$ 로 최소값이 된다. Cohen의 기준을 적용하면 중간 (medium)보다 약함(weak)에 더 가깝기는 하지만 앞서 살펴 본 소득분포 자료나 신용한도 데이터에 비하면 지수분포에의 적합성이 떨어진다.

그러나 매출금액대가 작은 곳에 많은 지역블록이, 매출금액대가 큰 곳에 적은 지역블록이 분포함으로써 전체적인 분포가 우하향의 L자형 곡선으로 나타나는 점은 실증되었고, 부수적으로 월별 지수가 큰 차이가 없다는 사실도 확인할 수 있었다. 이것은 어떤 집단의 특정 시기의 이용행태를 특정한 지수분포로 묘사할 수 있으면 그 집단의 이용행태는 계속 그 분포로써 묘사할 수 있음을 의미한다.

(3) 시뮬레이션 생성 데이터

마지막으로 시뮬레이션을 통해 생성한 개별 소비자 단위의 이용금액 데이터를 사용하여 지수분포 속성을 살펴보고자 한다. 공인된 신용카드 이용실적 시뮬레이션 모형은 존재하지 않으나, 선행연구 결과들을 활용하고 일부 항목을 국내 상황에 맞춰 보정하여 현실성 높은 모형을 구축하였다.

Schmittlein et al.(1987)은 포아송 과정을 기반으로 일반적인 거래의 빈도와 시점을 생성할 수 있는 모형을 수립하였다. Fan(2016)은 Schmittlein의 모형에 거래금액(신용카드 이용금액)을 생성할 수 있는 수식을 추가하여 신용카드 거래의 시뮬레이션 방법을 제시하였다. 그는 중국의 한 상업은행의 신용카드 매출 자료 44개월분을 이용하여 시뮬레이션 모형을 수립하고 데이터를 검증하였다.

그러나 중국의 신용카드 보급율이 2012년 기준 25%에 불과하므로(Heggestuen 2014) 국내 상황과 동떨어진 측면이 있다. 이러한 부분을 반영하여 최종적으로 정립한 신용카드 이용실적 시뮬레이션 모형은 Table 3과 같다.

Table 3. Credit card transaction model

Notation	Represents	Model	Value
i	card user index		1 to 50,000
T	limit of days		30
k	usage index for card user i		
λ_i	average daily number of usages for card user i	$\lambda_i \sim \text{Gamma}(\alpha_1, \beta_1)$	$\alpha_1=1.5$ $\beta_1=1.0$
$t_{i,k}$	time of k th usage for card user i	$t_{i,k} - t_{i,k-1} \sim \text{Exponential}(\lambda_i)$	
$U_{i,k}$	usage amount at $t_{i,k}$	$\ln(U_{i,k}) \sim \text{Normal}(\mu_i, \sigma_i^2)$	
μ_i	average of $U_{i,k}$	$\mu_i \sim \text{Normal}(\mu_u, \sigma_u^2)$	$\mu_u=9$ $\sigma_u^2=0.36$
σ_i^2	variance of $U_{i,k}$	$1/\delta_i \sim \text{Gamma}(\alpha_2, \beta_2)$ Set $\sigma_i^2 = \mu_i \delta_i$	$\alpha_2=15$ $\beta_2=1.0$

Modification of model

Stage	Process
1	Generate a random value following $\text{Gamma}(\alpha_2, \beta_2)$, which is C
2	$1/\delta = C$, so $\delta = 1/C$ ($\because \delta > 0$)
3	Set $\sigma_i^2 = \mu_i/C$

위 모형으로 시뮬레이션을 50회 시행하였다. 각 시행마다 5만명의 고객이 일으킨 신용카드 이용실적을 생성하였고 실험기간은 한 달로 정하였다($T = 30$). 서울시 빅데이터 캠퍼스의 자료에서도 보았듯이 국내 신용카드 이용금액은 월별 편차가 크지 않으므로 어느 한 달의 속성을 일반화할 수 있다.

λ_i 는 소비자 i 의 일평균 신용카드 이용건수이다. 한국은행의 지급수단 이용행태 조사결과 및 시사점(BOK, 2016)에 따르면 카드당 이용건수는 월 평균 16건이다. 이 값을 채택하면 5만개 λ_i 의 평균은 0.5 정도가 되어야 한다. 그러나 소비자들 가장 자주 쓰는 주 사용 카드만을 대상으로 한다면 이보다는 큰 값이어야 하고, 또한 λ_i 값이 너무 작으면 0의 매출이 대량 발생할 수 있다. 이런 상황을 종합적으로 고려해서 λ_i 의 평균값을 1.5로 결정하였다. λ_i 는 감마분포 $\text{Gamma}(\alpha_1, \beta_1)$ 을 따르는데 $\beta_1=1$ 로 놓으면 감마분포의 평균이 곧 α_1 이 되어 다루기 편리해진다. 따라서 $\alpha_1=1.5$, $\beta_1=1$ 을 적용하였다.

금융감독원 신용카드 영업실적(FSS, 2017)에 따르면 2016년 한 해 신용카드 이용실적은 596조원, 신용카드 발급매수는 9,564만 매이다. 이중 어느 정도를 실제 사용되는 카드로 보느냐에 따라 카드당 이용금액이 달라진다. 보수적으로 90%를 적용하면 카드당 월 평균 이용금액은 57만 원 정도로 계산된다. $U_{i,k}$ 의 수준을 결정하는 파라미터는 μ_u 와 σ_u^2 인데 μ_u 는 사전 모의실험 결과 9일 때 57만 원에 근접한 값이 나왔다. μ_u 의 퍼짐의 정도는 σ_u^2 로 표현되는데, 약 0.25 ~ 0.49일 때 이용금액들이 실제 신용카드 이용금액과 유사하였으므로 본 연구에서는 0.36을 사용하였다.

마지막으로 모형에서 신용카드 이용금액을 보정한 부분을 살펴보기로 한다. 예를 들어 사용자 i 의 평균 이용금액 ($U_{i,k}$)이 상대적으로 높은 수준이라면 μ_i 은 상대적으로 큰 값을 갖는다. 이 때 이용금액의 분산 σ_i^2 가 매우 작은 값이라고 가정하자. 이것은 $U_{i,k}$ 의 퍼짐의 정도가 작다는 의미이고 따라서 사용자 i 의 이용거래는 계속 높은 금액만 발생할 것이다. 이런 경향은 신용카드 보급률이 낮고, 신용카드가 부유층 위주로 보급되어 있는 중국에서는 적절할 수 있다. 그러나 우리나라와 같이 일상생활에서 신용카드를 현금보다 더 많이 사용하고 있는 상황에서 이용거래의 단가

는 낮은 금액과 높은 금액이 고르게 분포하는 것이 자연스럽다.

이에 따라 σ_i^2 가 μ_i 에 비례하는 것으로 모형을 수정하였다. 그렇게 하면 μ_i 값이 클 때 σ_i^2 도 커져서 작은 금액의 $U_{i,k}$ 가 출현 가능하다. 원 모형에서는 $1/\sigma_i^2$ 이 감마분포를 따른다고 가정되는데, 여기서는 새 변수 δ_i 을 도입하고 $1/\delta_i$ 이 α_2, β_2 를 파라미터로 하는 감마분포를 따르도록 하였다. $1/\delta_i$ 의 값이 C로 결정되면 $\sigma_i^2 = \mu_i/C$ 로 연산한다. 이렇게 수정한 모형을 이용하여 5만명 씩 50회 시뮬레이션을 한 결과 매 시행마다 Figure 6과 유사한 분포가 나타났다.

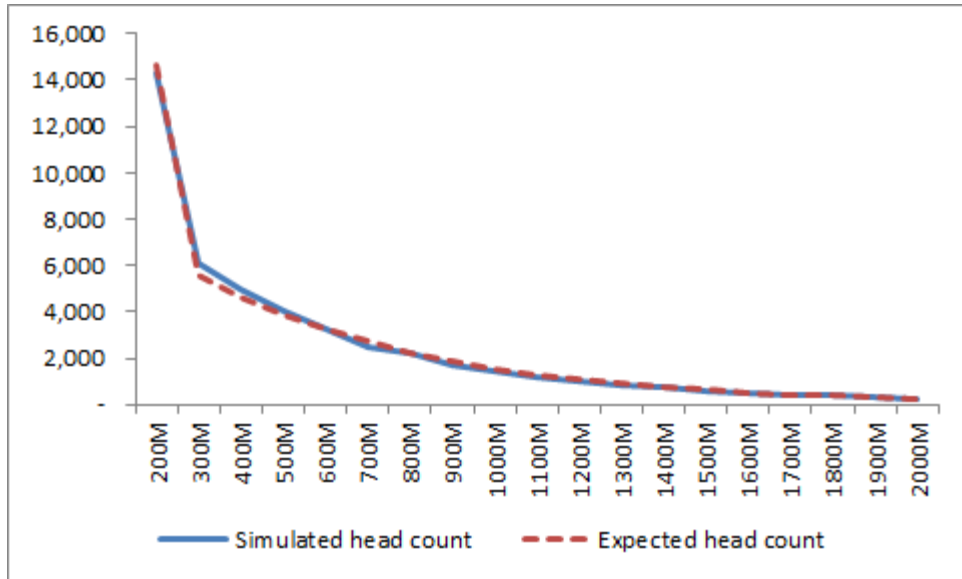


Figure 6. Fitting of simulated data to exponential distribution with $\lambda = 1/\text{average usage amount}$

소수의 무실적 매출(0의 매출)과 지나친 고금액 매출의 영향을 제거하기 위해 시행별로 상-하위 1.5%씩을 삭제하고 집계한 결과 각 시행마다 Figure 6처럼 지수분포에 가까운 분포를 나타냈고 평균 이용금액은 56만원으로 산출되어 사전 목표 57만원과 비슷한 수준이 되었다. W값은 0.053 ~ 0.071로서 50회 모두 Cohen이 제안한 약함(weak)의 기준인 0.1을 크게 밑돌았다. 이것은 이용금액대별 사용자 수 분포가 지수분포에 아주 가깝다는 의미이다. 한편 지수분포 적합의 척도인 r은 최저 0.994, 최고 0.999로 산출되어 이용금액대별 사용자수는 지수분포로 적절히 묘사될 수 있음이 입증되었다.

3. 지수분포의 활용

3.1 KPI로의 활용

신용카드 이용금액대별 사용자 수 자료를 지수분포로 적합시키면 많은 장점이 존재하는데, 그 중 하나는 지수분포는 단일 파라미터를 가지므로 분석 결과의 해석이 쉽고 복잡한 시나리오도 용이하게 분석 가능하다는 점이다. 또 하나 중요한 장점은 그 단일 파라미터가 바로 경영상의 주요 KPI (Key Performance Indicator)로 운용할 수 있다는 점이다.

2.1에서 소득분포에 대하여 역사적으로 다양한 모형이 제시되었음을 기술하였다. 이들 모형은 모두 L자형 곡선이므로 큰 차이는 없다. 그러나 곡선의 만곡률이 조금씩 다르다. 일부 집단의 이용금액별 사용자 수 분포는 지수분포보다 파레토 분포나 감마분포에 더 가까울 수도 있다. 그럼에도 지수분포를 대표 분포로 제시하는 것은 단일 파라미터 모형이므로 해석이 용이하기 때문이다.

지금까지 신용카드 이용금액별 사용자 수는 지수분포로 설명될 수 있음을 기술하였다. 지수분포와 관련된 식(1) ~ (3)에서 X는 ‘신용카드 이용금액’이므로 이 지수분포는 이용금액의 확률분포라고도 할 수 있다. 또한 이용금액의 평균 $E(X) = 1/\lambda$ 의 관계도 서술하였는데, X가 이용금액이므로 λ 는 이용금액과 이용자 수의 속성을 한꺼번에 나타낸다. 그리고 간접적으로 이용빈도와도 연관되어 있다. 따라서 λ 는 RFM을 모두 나타내는 파라미터이다.

이런 속성을 λ 를 KPI로 운용할 수 있는 여지를 갖게 한다. 즉, λ 는 해당 사용자 그룹의 종합적인 이용행태(RFM)를 나타내므로 경영 수준 지표, 목표 달성률 지표, 혹은 서비스품질 지표 등 다양한 KPI로 활용 가능하다. 다른 분포라면 멀티 파라미터의 수많은 조합이 발생하기 때문에 해석이 복잡하여 KPI로 운용하기는 어렵다.

3.2 카드상품별 이용행태의 묘사

지수분포에의 적합이 가져오는 또 하나의 장점은 다양한 카드 사용자 그룹별로 고유하게 분포를 설정할 수 있는 확장성이다. 지수분포로 특정 사용자 그룹의 이용행태를 설명할 수 있으므로 사용자들을 일정한 기준으로 세분한 집단 역시 지수분포로 설명 가능하다. 대표적인 사례로 카드상품별 이용행태를 지수분포로 적합할 수 있다. 다시 말해 카드상품별로 고유한 λ 를 갖는다. 마찬가지로 카드브랜드별 λ 도 존재하며 고객군별로도 상이한 λ 를 갖는다. 이런 속성은 카드사의 고객관리에 유용하게 활용될 수 있다. λ 값만 보면 해당 그룹의 주요 속성들을 한눈에 알아 볼 수 있기 때문이다.

다음으로 이용금액별 사용자 수 분포와 이용금액별 총이용금액 분포의 관계를 그림으로 비교함으로써 서로 어떻게 연관되어 있는지 밝히고자 한다. 그리고 그것을 통해 λ 를 좀 더 구체적으로 활용하는 방법을 예시하고자 한다. Figure 7의 왼쪽 그래프는 1만 명의 사용자를 보유한 가상의 A카드의 속성을 지수분포로 묘사한 것이다. 이 상품은 $\lambda = 0.2$ 의 지수분포를 따른다고 가정하자. 금액 단위를 10만원으로 설정했으므로 $\lambda = 0.2$ 는 평균 이용금액이 50만원 정도임을 의미한다.

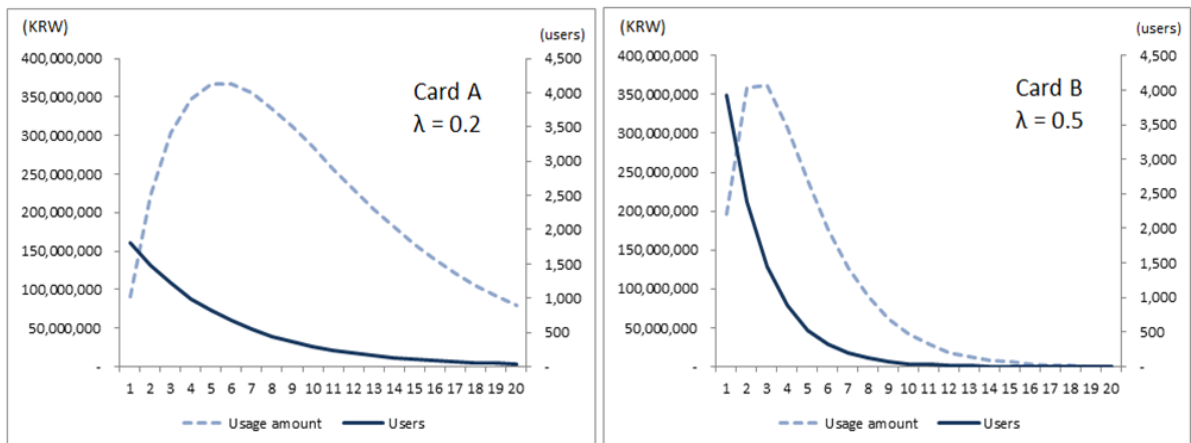


Figure 7. Distribution of users and usage amount at $\lambda = 0.2$ and 0.5 , respectively

각 이용금액대별로 중간값을 취하고 이것을 사용자 수와 곱하면 근사적인 총이용금액을 얻을 수 있다. 이렇게 얻은 각 이용금액대별 총이용금액을 이용하여 개인별 평균 이용금액을 구하면 493,400원인데 이론적인 평균값 50만원과 가깝다. 산출 결과 A카드는 월 150만 원 이상 신용카드를 쓰는 고객수가 전체의 4.98%이고 이들이 사용하는 총 이용금액의 비율은 18.53%에 달하는 것으로 나타났다.

오른쪽 그래프는 $\lambda = 0.5$ 의 지수분포를 따르는 가상의 B카드의 상황이다. 평균 이용금액이 20만 원 정도이다. 월 150만 원 이상 신용카드를 쓰는 고객수는 0.04%에 불과하고 이들의 총이용금액 비중도 0.32%에 그치고 있다. 이용금액의 수준이 낮으므로 금액 그래프의 정점이 A카드보다 왼쪽에 치우쳐 있다.

B카드는 A카드에 비해 실적이 저조하므로 당장 폐지해야 할 것 같지만 두 카드는 서로 다른 사용자 집단에 연관되어 있고 그들의 카드 보유 목적이 엄연히 다르므로 두 카드는 모두 존재의 의미가 있다. 예를 들어 A카드는 항공 마일리지 카드, B카드는 일반 적립 카드일 수 있는 것이다. 따라서 서로 다른 λ 를 갖는 카드상품들과 사용자 그룹들이 공존하고, 카드상품마다 고유의 λ 를 갖게 된다.

3.3 고객군별 이용행태의 묘사

카드사들은 대부분 VIP 마케팅을 하고 있다. VIP 고객군을 선정하여 전용 카드를 발급하거나 차별적인 혜택을 제공하는 것이다. 발렛 파킹, 프리미엄 매거진, 골프장 콘서트 예약 등 다른 고객군의 사용자들이 이용할 수 없는 혜택을 제공함으로써 고정 고객화 하려는 것이 VIP 마케팅의 전략이다. 따라서 그런 혜택을 향유하는 고객군의 이용행태는 다른 고객군의 행태와 차이가 있을 수밖에 없다.

어떤 사용자 그룹이 $\lambda = 0.05$ 의 속성을 갖고 있다면 평균 이용금액은 1,840만원에 달하고 월 5천만 원 이상 거래를 일으키는 사용자는 8.2%에 달한다. 그리고 이 최상위 사용자들의 총이용금액은 전체 이용금액의 22.5%를 차지한다. VIP 고객군의 평균 이용금액, 최상위 금액대의 사용자 비중, 총이용금액 비중 등은 매우 중요한 관리지표인데 이 주요 지표들이 바로 지수분포의 파라미터 λ 하나만을 이용하여 통합적으로 묘사될 수 있는 것이다.

3.4 마케팅 전략 수립에의 활용

3.2절에서 기술한 A카드는 상대적으로 우량 고객군이 사용하는 카드이고 B카드는 상대적으로 낮은 이용금액대의 고객들이 사용하는 카드이다. B카드의 인당 평균 이용금액은 204,065원에 그쳐 A카드의 평균 이용금액인 493,400원의 절반에도 미치지 못한다. 이 상황에서 B카드의 인당 평균 이용금액을 30만원까지 올리려면 $\lambda = 0.33$ 은 되어야 할 것이고, 따라서 그렇게 되기 위해서는 각 이용금액대별로 고객수가 어느 정도 변경되어야 하는지 산출해낼 수 있다.

지수분포의 장점은 도달해야 할 인당 평균 이용금액이 정해지면 그 값이 도출되도록 λ 를 변경하는 것만으로 각 이용금액대별 사용자 수 분포가 합리적으로 결정된다는 점이다. 어떤 금액대의 고객을 그 다음 금액대로 1명 올리는데 필요한 마케팅 예산이 경험적으로 알려져 있다면 총 마케팅 예산까지도 동시에 결정할 수 있다.

만일 지수분포의 속성을 모른다면 이용금액대별 고객 수를 정할 때 객관적 근거 없이 담당자의 주관적 판단에 의지할 가능성이 크다. 그 경우 고객 분포와 마케팅 예산이 현실과 동떨어지게 산출되거나 효율이 떨어지는 고객군에 많은 예산이 배정될 위험도 있다. 그러나 담당자가 이용금액대별 사용자 수의 지수분포 속성을 알고 있다면 합리적인 변화 수준을 예측할 수 있고 그 변화에 합당한 자원을 배분할 가능성이 매우 높아진다. 이처럼 지수분포는 단일 파라미터의 변동만으로 다양한 시나리오를 분석할 수 있는 장점이 있다.

4. 결 론

본 연구는 민간 소비 지출 시장에서 신용카드업이 차지하는 위치를 조명하고, 신용카드 분석의 새로운 길을 제시한 것에서 의의를 찾을 수 있다. 충분한 신용카드 이용실적 데이터가 공개되지 않는 현실 속에서 다양한 간접 자료와 시뮬레이션을 통해 개인별 신용카드 이용행태를 확률분포 모형으로 설명하고자 하였다. 신용카드 이용행태를 확률분포로 설명하려는 시도는 공개 데이터 확보의 어려움과 데이터마이닝에 중점을 두고 있는 카드사들의 분석 기조로 인해 거의 없는 실정이었다. 따라서 본 연구는 국내 신용카드 이용행태에 본격적인 확률분포 모형을 적용한 최초의 시도라고 볼 수 있다.

본 연구는 사용실적 중심의 이용행태를 RFM으로 규정하고 일정 기간 동안의 신용카드 이용금액별 사용자 수 분포, 즉 이용금액의 확률분포를 찾았다. 이 분포가 지수분포로 적합될 수 있음은 크게 두 단계로 밝혔다. 첫 단계로 민간 소비 지출에서 신용카드가 차지하는 비중이 현금 비중보다 높은 국내 결제시장의 특성을 이용하여 소득자료가 지수분포로 묘사됨을 보이고, 소득-지출-신용카드 이용금액간의 관계를 통해 지수분포 속성이 신용카드 이용금액에도 적용될 것임을 추론하였다.

두 번째 단계로는 UCI repository의 독일 신용카드 데이터, 서울시 빅데이터 캠퍼스의 지역블록별 카드 이용금액, 그리고 시뮬레이션 모형을 도입하여 신용카드 이용행태의 지수분포 속성을 실증하였다. 이렇게 수립된 지수분포 모형은 비즈니스 현장에서 KPI와 마케팅 전략 분석 도구로 활용될 수 있음을 예시하여 모형의 실질적인 효용을 제시하였다. 지수분포는 단일 파라미터의 변동을 통해 다양한 시나리오 분석이 가능하므로 카드사들에게도 유용하게 활용될 수 있을 것이다.

지수분포는 모형의 단순한 형태와 폭넓은 적용성으로 인해 오랫동안 널리 활용되어 왔다. 이것을 신용카드업으로 확장하면서 지수분포가 신용카드 이용행태를 묘사하는 데 크게 활용될 수 있고 효용도 적지 않음을 보였다. 정부의 빅데이터 활용 장려 기조가 계속되고 가용한 데이터가 많아지면 지수분포의 활용처가 더욱 다양하게 발굴될 수 있을 것이고, 그런 측면에서 본 연구는 후속 연구들의 합리적인 출발점이 될 것이다.

APPENDIX

1. 신용카드 이용실적 자료 생성을 위한 R코드 (N = 50,000)

```

` ` `` {r}
library(MASS)
custid <- seq(1,50000)
lambda <- rgamma(50000,1.5,1)
tr_mu <- rnorm(50000,9,0.6)
trinvsqsq <- rgamma(50000,15,1)
CTmodel <- data.frame(custid,lambda,tr_mu,trinvsqsq)
CTmodel$str_sg <- with(CTmodel,sqrt(tr_mu/trinvsqsq))
CTmodel$trinvsqsq <- NULL
for (i in 1:50000) {

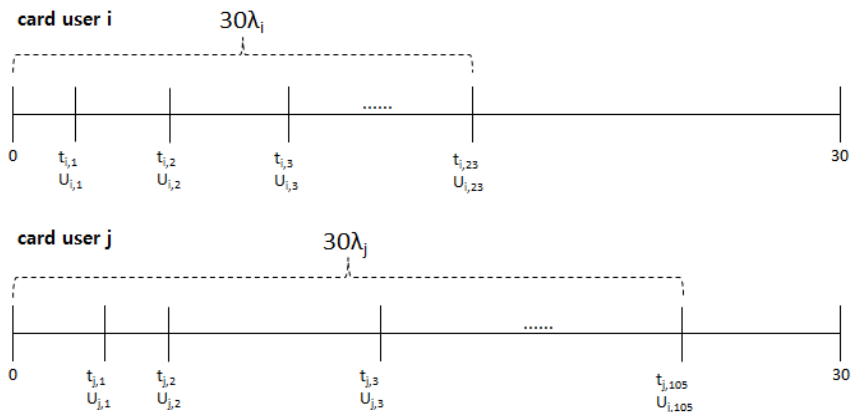
```

```

ptime <- cumsum(rexp(300,CTmodel$lambda[i]))
p_yn <- ifelse(ptime>30,0,1)
tr_amt <- exp(rnorm(300,CTmodel$str_mu[i],CTmodel$str_sg[i]))
sales <- c(tr_amt*p_yn)
CTmodel$totalsales[i] <- sum(sales)
}
write.csv(CTmodel$totalsales, file="ctmodel.csv")
---
```

위 코드에서 여러 파라미터 값들은 2.3절에서 설명한 바와 같다. 다만 rnorm 함수의 두 번째 인자는 분산이 아니라 표준편차이므로 0.36 대신 0.6을 적용하였다. For 루프 아래 300은 고객별 최대 거래건수이다. $\lambda = 1.5$ 일 때 최대 거래건수를 300으로 설정하면 이 값을 넘어가는 사용자가 나타나는 경우는 사실상 없다. 이것은 시뮬레이션을 1회 수행할 수 있는 Rmarkdown 코드이다.

2. 신용카드 이용실적 자료 생성 시뮬레이션 모형의 개념도



두 명의 신용카드 사용자 i 와 j 가 한 달 동안 일으키는 카드 사용 시점과 금액을 도식화 한 것이다. 카드 사용은 포아송 프로세스를 따르므로 사용 시점간의 간격은 각각 λ_i, λ_j 를 파라미터로 하는 지수 분포를, 사용 건수는 $30\lambda_i, 30\lambda_j$ 를 파라미터로 하는 포아송 분포를 따른다.

3. 신용카드 이용실적 자료 생성을 위한 Excel Spreadsheet Modeling 화면

다음 장의 왼쪽 화면은 신용카드 거래 시점(이용 시점)이다. 거래(n)과 거래(n-1)간의 간격 시간은 지수분포를 따른다. 오른쪽 화면은 신용카드 거래 금액(구매금액)이다. 구매금액은 거래 시점과 연동된다. 거래 시점이 1개월을 넘어가면 구매금액은 0으로 조정된다. 각 사용자별로 1개의 값을 갖게 되는데 한 달 동안의 구매금액을 모두 합한 값(월 이용금액)이다.

λ	2.02	0.60	0.63	1.17	2.80	0.28	0.81	μ	10.15	9.43	9.30	9.15	9.96	10.23	8.67
								$1/\sigma^2$	24.37	15.04	10.94	17.47	21.04	22.18	12.88
								σ	0.65	0.79	0.92	0.72	0.69	0.68	0.82

	사용자1	사용자2	사용자3	사용자4	사용자5	사용자6	사용자7	사용자1	사용자2	사용자3	사용자4	사용자5	사용자6	사용자7	
거래 1	1.0	1.9	4.0	0.3	0.3	3.3	0.2	구매금액 1	68,500	3,200	10,900	17,800	15,800	20,800	1,500
거래 2	1.1	3.4	5.1	0.3	0.8	12.2	2.6	구매금액 2	18,700	28,400	20,500	14,000	12,500	57,000	4,500
거래 3	1.6	4.6	5.6	0.5	1.5	16.2	3.7	구매금액 3	29,600	4,200	29,800	8,100	48,000	17,700	5,700
거래 4	2.3	10.1	5.7	1.0	2.2	20.7	4.6	구매금액 4	21,000	3,900	3,800	1,400	8,400	63,300	6,800
거래 5	2.4	13.9	5.9	3.0	3.2	21.2	6.2	구매금액 5	39,700	2,600	14,200	7,700	36,700	19,700	7,000
거래 6	2.8	14.6	6.4	3.1	3.3	27.7	6.9	구매금액 6	13,200	39,300	13,500	14,800	38,100	33,000	2,400
거래 7	2.9	15.5	6.9	3.3	3.7		7.7	구매금액 7	19,400	5,800	3,000	19,600	13,900		4,100
거래 8	3.1	15.7	8.5	10.7	3.7		9.1	구매금액 8	16,500	17,800	12,500	16,600	32,000		3,200
거래 9	3.3	18.0	9.8	10.9	3.8		9.4	구매금액 9	8,200	18,700	8,500	11,800	24,200		8,200
거래 10	4.0	20.0	12.7	13.1	4.4		9.4	구매금액 10	23,100	6,500	14,400	10,900	17,200		6,600
거래 11	4.1	20.5	13.3	13.3	4.5		9.6	구매금액 11	26,400	12,100	4,400	11,500	5,300		3,700
거래 12	4.4	21.9	14.6	13.4	4.8		10.1	구매금액 12	25,700	7,400	9,000	14,600	27,300		2,400
거래 13	4.5	22.9	18.0	14.2	5.0		11.7	구매금액 13	18,700	10,500	2,900	16,600	21,000		7,500
거래 14	5.2	26.7	19.9	14.5	5.0		12.7	구매금액 14	17,300	28,500	9,200	15,300	47,300		8,700
거래 15	5.6	29.5	24.8	16.6	5.9		13.0	구매금액 15	28,400	25,700	17,500	7,600	14,600		2,900
거래 16	6.7		26.7	17.3	6.1		20.3	구매금액 16	70,600		1,200	11,700	70,500		10,200
거래 17	6.8		29.9	17.4	6.3		21.4	구매금액 17	35,200		10,600	4,800	15,200		4,100
거래 18	6.8			18.1	6.5		21.5	구매금액 18	9,800			7,300	15,300		19,300
거래 19	7.0			21.7	6.8		21.5	구매금액 19	32,400			3,300	71,200		4,500
거래 20	9.1			22.8	7.7		22.8	구매금액 20	9,100			4,900	5,500		5,100

REFERENCES

- Bank of Korea(BOK). 2016. "A research on payment method usage behaviors and its implications."
- Bentler, P.M., and Bonnet, Douglas G. 1980. "Significance tests and goodness of fit in the analysis of covariance structure." *Psychological Bulletin* 88(3):588-606.
- Champernowne, D.G. 1953. "A model of income distribution." *The Economic Journal* 63(250):18-351.
- Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*: Erlbaum Associates 2nd ed.
- Credit Finance Association(CREFIA). 2014. "Credit Card Business Report September."
- Diacon, Paula-Elena, and Maha, Liviu-George. 2014. "The relationship between Income, Consumption and GDP." *Procedia Economics and Finance* 23:1535~1543.
- Drăgulescu, A., and Yakovenko, V.M. 2001. "Evidence for the exponential distribution of income in the USA." *European Physical Journal B* 20: 585-589.
- Fan, Yuhao Jeremy. 2016. "Examining credit card consumption pattern." *Washington University*:4-12.
- Heggestuen, John. 2014. "Low Credit Card Usage Means Carrier Billing Could Catch on in Emerging Markets." *Business Insider Australia*.
- Jung, H.M., Cho, S.J., and Chung, Justin. 2013. "An Empirical Study on Credit Card Usage." *Journal of Economic Research* 18:103-124.
- Keynes, John Maynard. 1936. *The General Theory of Employment, Interest and Money*, Palgrave Macmillan.
- Kim, K.G., and Yeom, M.B. 2015. "A Study on consumption of credit cards as surrogate economic index and characteristics of local consumption." *Journal of Economic Research* 33(1):121-141.
- Lee, G.C., Jung, N.S., and Shin, G.S. 2002. "An Artificial Intelligence-based Data Mining Approach to Extracting Strategies for Reducing the Churning date in Credit Card Industry." *Journal of intelligent information systems* 8(2):15-35.
- Lee, Seung-Hwan. 2014. "Credit Card Payment Ratio is 51%." *Financial News*.
- Lee, Y.J., Lee, S.H., and Lee, J.S. 2014. "KB's marketing activities and big data utilization." *KBR* 18(1).
- Reliability Analysis Center. 2003. "The Chi Square: A large sample goodness-of-fit test." *START* 10(4).
- Schmittlein, David C., Morrison, Donald G., Colombo, Richard. 1987. "Counting Your Customers: Who are they and what will they do next?" *Management Science* 33(1): 3-11.
- So, Meko M.C., and Thomas, Lyn C. 2011. "Modeling the profitability of credit cards by Markov decision processes." *European Journal of operational research* 212:123-130
- Song, H.J., and Sung, M.J. 2012. "Analysis of the effect of credit card deduction." *Journal of Finance Research* 5(2):157-194.
- Till, Robert, and Hand, David. 2003. "Behavioural models of credit card usage." *Journal of Applied statistics*, 30(10):1201~1220.