PPS 抽样方法在我国农村居民生活消费支出估计中的应用

李晓玲，雷　恒，陈伟伟
（西北师范大学 经济管理学院，甘肃　兰州　730070）

〔摘要〕由于金融危机引起出口增长受阻，国内投资增长缓慢，城镇居民消费预期恶化且收入分配差距过大，农村有效需求的扩大成为备受关注的问题。因此，有必要通过市场调查，了解和掌握我国农村居民的消费状况。应用 PPS 抽样方法对我国农村居民生活消费支出进行估计，可以得到相关数据，进而为制定有效的产业策略提供参考依据。

〔关键词〕PPS 抽样；Hansen-Hurvitz 估计量；消费支出
〔中图分类号〕F224.0　　〔文献标识码〕A　　〔文章编号〕1671-6671（2009）03-0049-03

一、引言

目前，“三农”问题已经成为中央政府及有关各部门和理论界极为关注的热点问题。消费是经济的原动力，消费、投资和净出口被誉为拉动经济增长的“三架马车”，其中消费的作用是最重要的。当前，我国消费市场的形势并不乐观。由于金融危机引起出口增长受阻，国内投资增长缓慢，城镇居民消费预期恶化且收入分配差距过大，在这种情况下，人们自然地将增加有效需求的注意力转向了农村。

根据当前的经济形势，为扩大国内需求、改善民生、拉动消费、带动生产、促进经济平稳较快发展，2008 年 11 月，国务院决定在全国推广“家电下乡”活动。2009 年中国农村经济绿皮书指出：“扩大国内需求，最大潜力在农村”。因此，有必要通过市场调查，了解和掌握我国农村居民的消费状况，进而为制定有效的产业策略提供参考依据。[1]为此，本文采用不等概率抽样(PPS 抽样)方法，对全国 31 个省、自治区、直辖市的农村居民家庭平均每人生活消费支出进行估计。

二、PPS 抽样及其估计量

1.PPS 抽样概述

PPS 抽样法（Probability Proportionate to Size Sampling)又称按规模大小成比例的概率抽样或按容量比例概率抽样(PPS)法。它是多项抽样的一种，多项抽样是一种不等概抽样，设 Z_1、Z_2…，Z_n 是一组概率，按这组概率对总体中的 N 个单元进行放回抽样，每次抽中第 i 个单元的概率为 Zi，独立地进行这样的抽样 n 次，则这种不等概抽样为多项抽样。尤其在每个单元有说明其大小或规模的度量 Mi，则 Zi 可取 Zi=Mi/Mo，这时，每个单元在每次抽选中入样的概率与其单元规模的大小成比例。这种抽样方法被称为 PPS 抽样。[2]

由于所面对的总体有可能差异不大，也有可能差异非常大，当总体单元之间差异不大时，各单元具有一定的代表性。这时使用简单随机抽样得到的估计值是精确有效的。但是当单元之间差异非常大时，使用简单随机抽样抽出的样本所估计的估计值误差极大，这时有必要考虑使用不等概随机抽样方法，即赋予各单元一个不同的入样概率，使大样本的入样概率大，小样本的入样概率小，从而提高估计量的估计精度。因此，在抽样时对样本大的单元赋予一个较大的入样概率，推算时给予一个较大的权；对待样

〔收稿日期〕2009－07－20
〔作者简介〕李晓玲（1985－），女，甘肃白银人，西北师范大学经济管理学院数量经济学硕士研究生。

本较小的单元赋予一个较小的入样概率，推算时赋予一个较小的权。加入辅助信息从而提高了抽样策略的统计效率，与简单随机抽样甚至与分层抽样相比，能显著地减少抽样误差，从而使估计更为精确有效。[3]

2.PPS 抽样的实施方法

PPS抽样的实施方法主要有累积总和法、拉希里方法、规模累积等距抽选的方法、分裂法。本文拟采用规模累积等距抽选的方法。

规模累积等距抽选方法的基本原理是：设总体单元数为N，其规模度量分别为M_1，M_2，…M_n，假定M_i都是整数，且有$\sum_{i=1}^{n} M_i = M_0$，这样总共有M_0个代码，每个总体单元都有一个代码的范围，其中第i个单元相应地有M_i个代码。若欲抽取的样本容量为n，则先求得等距抽样的间隔$K=\frac{M_0}{n}$，然后在1 ~ K之间随机等概率抽取一个数，假设为r，则r所在的单元代码区间相应的单元即为被抽中的单元。以后每隔K个度量值，即：r + K，r + 2K，r + 3K……，r +（n-1）K等数字所在的单元代码区间的相应单元，即为被抽中的单元。这种抽样方法的特点是当所有单元的度量Mi<K时，它是不重复的抽样；当某个M_i>2K时，则第i个单元肯定会被重复抽中。这种方法抽取样本比较容易，每个单元的被抽中概率与M_i的大小成比例。[4]

3.Hansen-Hurvitz 估计量

（1）总体总量的估计

1943年，汉森和赫维茨对PPS抽样提出了估计总体总量的估计量为：

$$\hat{Y}_{HH}=\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{Z_i}$$

其中yi为入样的第i个单元的变量值，zi为第i个单元根据其规模大小的入样概率。通常情况下若以该单元包含的元素单位为度量时，$Z_i=\frac{M_i}{M_0}$，其中，$\hat{Y}_{HH}$是总体总量的一个无偏估计量。[4]

方差估计量为：

$$v(\hat{Y}_{HH})=\frac{1}{n(n-1)}\sum_{i=1}^{n}(\frac{y_i}{z_i}-\hat{Y}_{HH})^2$$

（2）总体均值的估计

总体均值是将总体总量除以总体单元数，因此在汉森—赫维茨估计量的基础上除以N，即得到总体均值的估计量为：[4]

$$\hat{\bar{Y}}_{PPS}=\frac{1}{Nn}\sum_{i=1}^{n}\frac{y_i}{z_i}$$

方差的估计量为：

$$\hat{v}(\hat{\bar{Y}}_{PPS})=\frac{1}{N^2n(n-1)}\sum_{i=1}^{n}(\frac{y_i}{z_i}-\hat{\bar{Y}})^2$$

三、数据的采集并处理

选取2006年全国各地区农村居民家庭平均每人生活消费支出，以2006年的消费额作为规模Mi，并进行累积，如表1所示：

表1 2006年全国各地区农村居民家庭平均每人生活消费支出

地区	2006年人均消费支出(Mi)	∑Mi	代码范围	地区	2006年人均消费支出(Mi)	∑Mi	代码范围
北 京	5 724	5 724	1 ~ 5 724	湖 北	2 732	59 957	57 226 ~ 59 975
天 津	3 341	9 065	5 725 ~ 9 065	湖 南	3 013	62 970	59 958 ~ 62 970
河 北	2 495	11 560	9 066 ~ 11 560	广 东	3 885	66 855	62 971 ~ 66 855
山 西	2 253	13 813	11 561 ~ 13 813	广 西	2 413	69 268	66 856 ~ 69 268
内蒙古	2 771	16 584	13 814 ~ 16 584	海 南	2 232	71 500	69 269 ~ 71 500
辽 宁	3 066	19 650	16 585 ~ 19 650	重 庆	2 205	73 705	71 501 ~ 73 705
吉 林	2 700	22 350	19 651 ~ 22 350	四 川	2 395	76 100	73 706 ~ 76 100
黑龙江	2 618	24 968	22 351 ~ 24 968	贵 州	1 627	77 727	76 101 ~ 77 727
上 海	8 006	32 974	24 969 ~ 32 974	云 南	2 195	79 922	77 728 ~ 79 922
江 苏	4 135	37 109	32 975 ~ 37 109	西 藏	2 002	81 924	79 923 ~ 91 924
浙 江	6 057	43 166	37 110 ~ 43 166	陕 西	2 181	84 105	81 925 ~ 84 105
安 徽	2 420	45 586	43 167 ~ 45 586	甘 肃	1 855	85 960	84 106 ~ 85 960
福 建	3 591	49 177	45 587 ~ 49 177	青 海	2 178	88 138	85 961 ~ 88 138
江 西	2 676	51 853	49 178 ~ 51 853	宁 夏	2 246	90 384	88 139 ~ 90 384
山 东	3 143	54 996	51 854 ~ 54 996	新 疆	2 032	92 416	90 385 ~ 92 416
河 南	2 229	57 225	54 997 ~ 57 225				

将 M_0 = 92 416 除以样本量 n = 15，得抽样间隔 K= 6 161.07，在 1~K 之间抽一随机数，假设为 R = 3 204，处于北京的代码范围，因此北京被作为抽中的样本，其余的样本代码：3 204 + 6 161.07 = 9 365.07，9 365.07 + 6 161.07 = 15 526.14，21 687.21，27 848.28，34 009.35，40 170.42，46 331.49，52 492.56，58 653.63，64 814.7，70 975.77，77 136.84，83 297.91，89 458.98。所以被抽中的样本分别为：北京、河北、内蒙古、吉林、上海、江苏、浙江、福建、山东、湖北、广东、海南、贵州、陕西、宁夏。

四、估计量的计算

1.总体总量的估计

这 15 个省市被抽选的概率为：$Z_i=\frac{M_i}{M_0}$，分别为：北京 0.0619、河北 0.0270、内蒙古 0.0300、吉林 0.0292、上海 00866、江苏 0.0447、浙江 0.0655、福建 0.0389、山东 0.0340、湖北 0.0296、广东 0.0420、海南 0.0242、贵州 0.0176、陕西 0.0236、宁夏 0.0243，用这 15 个样本省市来估计 2007 年的全社会消费额，采用汉森——赫维茨估计量，由公式得：

$$\hat{Y}_{HH}=\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{Z_i}$$

$$=\frac{1}{15}\left(\frac{6\,399}{0.0619}+\frac{3\,065}{0.0292}\cdots\cdots+\frac{2\,528}{0.0243}\right)$$

$$=\frac{1\,575\,036.83}{15}=105\,002.46$$

故估计推断，这 31 个省市的农村居民的生活消费支出为 105 002.46 元。

抽样的方差：

$$\hat{v}(\hat{Y}_{HH})=\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{z_i}-\hat{Y}_{HH}\right)^2$$

$$=\frac{1}{15\times4}\left[\left(\frac{6\,399}{0.0619}-105\,002.6\right)^2+\left(\frac{3\,065}{0.0292}-105\,002.6\right)^2+\cdots\cdots+\left(\frac{2\,528}{0.0243}-105\,002.6\right)^2\right]$$

$$=\frac{1}{15\times14}\times88\,008\,466.13=82\,141\,235.05$$

置信度为 95%的置信区间为：

$$\hat{Y}_{HH}\pm Z_{\frac{\alpha}{2}}\sqrt{\hat{v}(\hat{Y}_{HH})}$$

即：

105 002.46 ±2×9 063.18=105 002.46±18 126.36

所以，置信区间为（86 876.10 ~ 123 128.82）。2007 年，这 31 省市的实际农村居民消费支出为 104652，位于置信区间之内。

2.总体均值的估计为：

$$\hat{\bar{Y}}_{PPS}=\frac{1}{Nn}\sum_{i=1}^{n}\frac{y_i}{z_i}=\frac{1}{31}\times105002.46=3387.18$$

方差的估计量为：

$$\hat{v}(\hat{\bar{Y}}_{PPS})=\frac{1}{N^2n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{z_i}-\hat{\bar{Y}}\right)^2$$

$$\frac{1}{31^2}\times82141235.05=85474.75$$

所以总体均值的抽样标准误为：

$$\hat{v}(\hat{\bar{Y}}_{HH})=\sqrt{85474.75}=292.36$$

五、小结

通过对 2006 年全国各地区农村居民人均生活消费支出数据的估计，2007 年全国各地区农村居民实际人均生活消费支出位于置信区间之内。由此可见，不等概率抽样虽然在实施方面较简单随机抽样复杂，但是对差异总体较大的总体单元进行抽样估计会更为精确有效。

参考文献：

[1] 钟子良.对 PPS 法在上海市居民家庭消费行为和意向调查中应用的评估[J].上海统计，2000，(6)：24-27.
[2] 李培军.不等概率抽样估计的原理与应用[J].辽宁师范大学学报（自然科学版），2004，(12)：385-388.
[3] 俞纯权.抽样下子总体参数的估计[J].统计与决策，2006，(9)：14-15.
[4] 倪家勋.抽样调查[M].桂林：广西师范大学出版社，2002，(12)：161-190.

The Use of PPS Sampling Method in Estimating of China's Rural Residents' Consumption Expenditure

LI Huan-ling, LEI Heng, CHEN Wei-wei
(College of Economics and Management, Northwest Normal University, Lanzhou 730070, China)

Abstract: Because economic crisis hinders export increase, internal investment increases slowly, China's rural residents' consumption expectation is getting worse and the distance between the rich and the poor is expending, which draw all the people's attention to expending rural effective requirement. Therefore, it is necessary to know and get control of the consumption situation of the rural residents. Through using PPS Sampling Method to estimate the consumption expenditure of rural residents' life, we can get the corresponding data and provide reference basis for formulating effective business strategies.

Key words: PPS sampling; Hansen-Hurvitz estimator; consumption expenditure　　〔责任编辑：杨晓丹〕