

PPS和简单随机抽样估计效率的实证检验

周庆元

(四川大学政治学院,成都 610064)

摘要:抽样调查理论揭示,PPS抽样的效率要优于简单随机抽样的结果。文章以我国年各地区主要行业全社会固定资产投资额的均值和总量估计为例,验证小样本条件下两者的抽样效率。结果显示,无论是抽样估计结果效率、抽样误差还是95%置信度条件下的区间估计,PPS估计的效率均高于简单随机抽样。

关键词:PPS估计;简单随机抽样估计;固定资产投资

中图分类号:O212 **文献标识码:**A **文章编号:**1002-6487(2014)01-0014-03

0 引言

在统计抽样调查理论发展的过程中,相对于简单随机抽样而言,PPS抽样估计是抽样调查估计理论质的飞跃。关于PPS抽样调查方法的理论研究,一些学者借助于数理方法,给出了严密的推导和证明。从整体上看,现有研究侧重统计抽样调查方法的改进。如蔡择林^[1]讨论了PPS分层抽样时,样本单位数的设计问题,并给出了一个方便有效的样本分配规则。马慧敏,方国斌^[2]利用Monte Carlo方法对改进的PPS抽样进行模拟,并对改进的PPS抽样设计效应进行了评价。然而,相对于方法改进的研究而言,对不同抽样估计方法进行对比的研究成果较少,现有各种方法的对比研究虽然涉及面较广,但是不够深入。如张荷^[3]观对整群比估计与PPS抽样方法进行了比较,给出了整群比估计优于PPS抽样的条件。李战江和闫在在^[4]在PPS抽样方案和有效回简单随机抽样方案的情形下构造了这两种估计方法并精确比较了它们的优劣。俞纯权^[5]对PPS抽样下子总体参数的估计进行了说明。卢静莉^[6]等人则考察了多元简单随机抽样与二相抽样的最优分配问题。总体来看,对各种抽样估计方法的效率对比研究明显不足,PPS抽样估计和简单随机抽样估计的对比研究更是不多。它们两者之间的对比研究,倪加勋,张勇^[7]从理论上做了论证,并选择辅助变量进行了实证检验,Nibia, Aires^[8]也只是给出了一个理论上的说明。但是现有研究中,很少有人专门对两种抽样估计方法从实践的角度实证分析并加以比较和验证。理论毕竟是理论,只有能通过实践检验的理论才是具有社会现实意义的理论。为此,本文拟对最为常用和实用的两种抽样估计方法从实践的角度进行对比和实证检验。

1 PPS估计与简单随机抽样估计

1.1 PPS抽样及估计

1.1.1 PPS抽样

PPS估计是较为实用的一种不等概率抽样方法。若总体中第*i*个单元的规模度量为 M_i ,则该总体的总规模度量为 $M_0 = \sum_{i=1}^N M_i$,那么该单元的抽选概率应 $Z_i = \frac{M_i}{M_0}$,这种不等概率抽样就称之为与规模大小成比例的抽样,即为PPS抽样。

根据研究或实际工作需要,可以选用不同的PPS抽样方法,具体包括累积总和法、分裂法、拉希里方法、规模累积等距抽选法等。考虑到数据特点,为便于说明问题,本文主要采用规模累积等距抽选法。

规模累积等距抽选方法为:设总体单元数为 N ,其规模度量分别为 $M_1, M_2, \dots, M_N$,将规模度量按代码法进行累积,直至 $M_0 = \sum_{i=1}^N M_i$ 。假设所要抽取样的本容量为 n ,则先

求出等距抽样的间隔的数值,即 $K = \frac{M_0}{n}$,然后在1~ K 之间按照随机等概率抽取一个值,假定为 r ,那么 r 所在的单元代码区间相应的单元就是被抽中的单元。此后,同理可得,每间隔 K 个度量值,即 $r+K, r+2K, r+3K, \dots, r+(n-1)K$ 等数值所在的单元代码区间的相应单元,也就是被抽中的单元。

1.1.2 PPS抽样估计

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:

$$Y_{HH}^{\Delta} = \frac{1}{n} \sum_{i=1}^n \frac{Y_i}{Z_i}$$

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

作者简介:周庆元(1979-),男,江苏东台人,博士研究生,讲师,研究方向:社会主义经济理论。

$$\hat{V}(Y_{HH}^A) = \frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{y_i}{z_i} - Y_{HH}^A \right)^2$$

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:

$$\frac{\Lambda}{\bar{Y}_{PPS}} = \frac{1}{Nn} \sum_{i=1}^n \frac{Y_i}{Z_i}$$

方差的估计量为:

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

1.2 简单随机抽样及估计

1.2.1 简单随机抽样

简单随机抽样又称纯随机抽样,设总体有N个单元,从中抽取n个样本容量,使得每个样本被抽中的概率相等,这种抽样即为简单随机抽样。

简单随机抽样可以分为重复抽样和不重复抽样,其抽样概率是相等的,每个总体单位被抽中的概率都为 $\frac{1}{N}$,其随机样本的产生可以通过随机数表、计算机或计算器等方法产生。

1.2.2 简单随机抽样估计

①总体均值的估计

简单随机抽样中,估计总体均值通常用样本均值来估计,即:

$$\frac{\Lambda}{\bar{Y}} = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

其抽样方差为: $\hat{V}\left(\frac{\Lambda}{\bar{Y}}\right) = \frac{S_y^2}{n} \left(1 - \frac{n}{N}\right)$

②总体总量的估计

当N在抽样时已知时,总体总量的估计就较为简单,它的估计量为总体均值乘以总体单位数N,即: $\bar{Y} = N\bar{y}$

抽样方差为: $\hat{V}(\bar{Y}) = \frac{N^2 s_y^2}{n} \left(1 - \frac{n}{N}\right)$

2 基于中国固定资产投资的实证检验

2.1 实证数据及实证过程说明

抽样调查理论揭示,PPS抽样的效率要优于简单随机抽样的结果。本文以我国2010年各地区主要行业分的全社会固定资产投资额估计全国全社会固定资产投资均值和总值为例,验证总体和样本都较小的情况下两者的抽样效率。本文将估计我国全社会固定资产投资规模和各省(市)平均规模为例实证检验PPS抽样和简单随机抽样的效率。所用数据为2010年我国内地31个省市的全社会固定资产投资总额的数据,全部数据来源于《中国统计年鉴—2011》。

具体来说,本文实证检验PPS估计和简单随机抽样估计的效率,分为三个步骤来实现。首先对2010年全国内地31个省市的固定资产投资额进行PPS抽样估计,其次进行简单随机抽样估计,再次计算2010年全国内地31个

省市的固定资产投资额真实的均值及总值,结合真实的均值和总值对比分析PPS和简单随机抽样估计的效率。实证过程中选取样本容量为总体单位量的30%。

2.2 全社会固定资产投资的PPS抽样估计

PPS估计的抽样有很多中,其中,规模累积等距抽样是较为稳定和客观的方法,所以本文使用规模累积等距抽样选取样本。为此,首先对全社会固定资产投资规模计算累积规模、规模代码区间及规模比重。全社会固定资产投资规模、累积规模、代码区间以及规模比重如表1所示:

表1 各省市全社会固定资产投资规模及起止代码区间及规模比重

	规模	累积规模	代码区间	$Z_i = \frac{M_i}{M_0}$
北京	5403.0	5403	0 ~ 5403	0.019910
天津	6278.1	11681	5404 ~ 11681	0.023135
河北	15083.4	26764	11682 ~ 26764	0.055583
山西	6063.2	32827	26765 ~ 32827	0.022343
内蒙古	8926.5	41754	32828 ~ 41754	0.032895
辽宁	16043.0	57797	41755 ~ 57797	0.059120
吉林	7870.4	65667	57798 ~ 65667	0.029003
黑龙江	6812.6	72480	65668 ~ 72480	0.025105
上海	5108.9	77589	72481 ~ 77589	0.018827
江苏	23184.3	100773	77590 ~ 100773	0.085436
浙江	12376.0	113149	100774 ~ 113149	0.045606
安徽	11542.9	124692	113150 ~ 124692	0.042536
福建	8199.1	132891	124693 ~ 132891	0.030214
江西	8772.3	141663	132892 ~ 141663	0.032327
山东	23280.5	164944	141664 ~ 164944	0.085790
河南	16585.9	181530	164945 ~ 181530	0.061120
湖北	10262.7	191793	181531 ~ 191793	0.037819
湖南	9663.6	201457	191794 ~ 201457	0.035611
广东	15623.7	217081	201458 ~ 217081	0.057574
广西	7057.6	224139	217082 ~ 224139	0.026008
海南	1317.0	225456	224140 ~ 225456	0.004853
重庆	6688.9	232145	225457 ~ 232145	0.024649
四川	13116.7	245262	232146 ~ 245262	0.048336
贵州	3104.9	248367	245263 ~ 248367	0.011442
云南	5528.7	253896	248368 ~ 253896	0.020374
西藏	462.7	254359	253897 ~ 254359	0.001705
陕西	7963.7	262323	254360 ~ 262323	0.02935
甘肃	3158.3	265481	262324 ~ 265481	0.011639
青海	1016.9	266498	265482 ~ 266498	0.003747
宁夏	1444.2	267942	266499 ~ 267942	0.005322
新疆	3423.2	271365	267943 ~ 271365	0.012615

从上表中得知,31个省市投资规模累计和为271365,按照30%抽取样本,需要抽取9个样本,等距抽样的间隔 $K = \frac{M_0}{n} = \frac{271365}{9} \approx 30152$,依据随机数表在1~30152中产生一个随机代码,作为第一个抽样代码。在Excel中产生得到一个随机数为0.859615738。乘以K=30152得到第一个规模区间的抽取代码为25919,相应的样本单位为河北,样本单位的Y值为15083。按照规模等距抽样方法给各初始代码分别加K、2K、……8K得到抽样单位的所有代码、相应的抽样单位及样本单位的Y值如表2所示:

表2中的数据为9个被抽中的样本,为了估算2010年我国社会固定资产投资,对表2中的样本数据进行PPS估计,以下为PPS抽样估计过程。

(1)固定资产投资总额的估计

表2 抽样代码及抽样结果

代码	对应抽样单位	Y值	权重
25919	河北	15083	0.055583
56071	辽宁	16043	0.059120
86223	江苏	23184	0.085436
116375	安徽	11543	0.042536
146527	山东	23281	0.085790
176679	河南	16586	0.061120
206831	广东	15624	0.057574
236983	四川	13117	0.048336
267135	宁夏	1444	0.005322

依据表2的数据得总体总量的估计量为:

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

$$= \frac{1}{9} \left(\frac{15083}{0.055583} + \frac{16043}{0.059120} + \cdots + \frac{1444}{0.005322} \right)$$

$$= 271362.72$$

总体总量的估计量的抽样误差为:

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

$$= \frac{1}{9 \times (9-1)} \left[\left(\frac{15083}{0.055583} - 271362.72 \right)^2 + \cdots + \left(\frac{1444}{0.005322} - 271362.72 \right)^2 \right]$$

$$= 22.94$$

因此可以根据公式 $\hat{Y} \pm Z_{\frac{\alpha}{2}} \sqrt{V(\hat{Y})}$ 得出 95% 的置信条件下平均固定资产投资总额的置信区间为:

$$\hat{Y} \pm Z_{\frac{\alpha}{2}} \sqrt{V(\hat{Y})} = 271362.72 \pm 2 \times \sqrt{22.94} \approx 271362.72 \pm 9.58$$

(2) 固定资产投资均值的估计

依据表2的数据得全社会固定资产投资的均值估计量为:

$$\frac{\Lambda}{Y}_{PPS1} = \frac{1}{Nn} \sum_{i=1}^n \frac{y_i}{Z_i}$$

$$= \frac{1}{31 \times 9} \left(\frac{15083}{0.055583} + \frac{16043}{0.05912} + \cdots + \frac{1444}{0.005322} \right)$$

$$= 8753.64$$

抽样误差为:

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

$$= \frac{1}{31^2 \times 9 \times (9-1)}$$

$$\left(\frac{15083}{0.055583} - 271362.72 \right)^2 + \cdots + \left(\frac{1444}{0.005322} - 271362.72 \right)^2$$

$$= 0.02387$$

因此可以根据公式 $\frac{\Lambda}{Y} \pm Z_{\frac{\alpha}{2}} \sqrt{V(\frac{\Lambda}{Y})}$ 得出 95% 的置信条件下平均固定资产投资置信区间为:

$$\frac{\Lambda}{Y} \pm Z_{\frac{\alpha}{2}} \sqrt{V(\frac{\Lambda}{Y})} = 8753.64 \pm 2 \times \sqrt{0.02387} \approx 8753.64 \pm 0.309$$

2.3 全社会固定资产投资的简单随机抽样估计

为了简单随机抽样的实施,先对我国31个省市分别给以编号,以1~31为编号,编制一个抽样代码,即为抽样框,编制抽样框的结果如表3所示:

表3 中国31个省市简单随机抽样抽样框

北京	1	上海	9	湖北	17	云南	25
天津	2	江苏	10	湖南	18	西藏	26
河北	3	浙江	11	广东	19	陕西	27
山西	4	安徽	12	广西	20	甘肃	28
内蒙古	5	福建	13	海南	21	青海	29
辽宁	6	江西	14	重庆	22	宁夏	30
吉林	7	山东	15	四川	23	新疆	31
黑龙江	8	河南	16	贵州	24		

用Excel在1~31中产生9个随机数为:05、12、23、03、17、07、14、22、31,即内蒙古、安徽、四川、河北、湖北、吉林、江西、重庆、新疆九个省被抽选为样本,抽取结果如表4所示:

表4 中国31个省市简单随机抽样结果

代码	05	12	23	03	17	07	14	22	31
省份	内蒙古	安徽	四川	河北	湖北	吉林	江西	重庆	新疆
固定资产投资规模	8927	11543	13117	15083	10263	7870	8772	6689	3423

(1) 固定资产投资均值及均值误差的估计

固定资产投资估计量为:

$$\frac{\Lambda}{Y} = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

$$= \frac{1}{9} (8927 + 11543 + \cdots + 3423) = \frac{85687}{9} = 9521$$

均值方差估计量为:

$$\hat{V}(\bar{y}) = \frac{s^2}{n} \left(1 - \frac{n}{N} \right) = \frac{1}{n} \left(1 - \frac{n}{N} \right) s^2$$

$$= \frac{1}{n} \left(1 - \frac{n}{N} \right) \times \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

$$= \frac{1}{9} \left(1 - \frac{9}{31} \right) \times \frac{1}{9-1} [(8927-9521)^2 + \cdots + (3423-9521)^2]$$

$$= 959559$$

以95%的置信度可以得到均值的区间为:

$$\bar{y} \pm Z_{\frac{\alpha}{2}} \sqrt{\hat{V}(\bar{y})} \approx 9521 \pm 2 \times 980$$

(2) 固定资产投资总值总量及总值误差的估计

总值总量: $\hat{Y} = N\bar{y} = 31 \times 9521 = 295151$

总体方差:

$$V(\hat{Y}) = V(N\bar{y}) = N^2 V(\bar{y}) = 31^2 \times 959559 = 922136199$$

总体区间为:

$$\hat{Y} \pm Z_{\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{Y})} = 295151 \pm 2 \times \sqrt{922136199} \approx 295151 \pm 2 \times 30367$$

3 实证结果及结论

为了分析PPS和简单随机抽样的效率,本文计算真实的全社会固定资产投资总额和均值作为参照,通过计算,并整合以上的实证结果如表5所示。

由此可以看出,无论是均值估计还是总量估计,PPS估计的效果都要好于简单随机抽样的估计效果。具体而言,PPS抽样均值的估计值和真实值的系统性差异只有0.06,均值抽样方差只有0.02387,而简单随机抽样与真实值相差767.3,抽样方差达到959559。从总量估计看,PPS

基于改进模拟退火算法的模糊回归参数估计

彭宇文¹, 郭莉莎¹, 毛超²

(1.湖南师范大学 商学院,长沙 410000;2.湖南大学 工商管理学院,长沙 410082)

摘要:文章讨论了自变量、因变量以及回归系数均为模糊数的线性和非线性回归模型,提出了一种与模糊隶属度相关的回归效果评价方法,并设计了模拟退火算法进行优化求解。将隶属度为1时各模糊数取值以及各模糊数的边界进行回归产生的系数作为初始解,采用多层邻域操作产生新解,通过多准则终止运算,并且利用增加记忆装置记录了每次运算的最优解。算例和对比分析的结果表明,该方法收敛速度快、回归效果较好,适合处理较复杂的模糊回归模型。

关键词:模糊回归分析;模糊数;模拟退火算法;参数估计

中图分类号: O212.4 **文献标识码:** A **文章编号:** 1002-6487(2014)01-0017-03

0 引言

在现实生活中,统计对象的某些特性使得一些统计数据往往难以准确的记录或收集。例如,在测量河流水位高低时,会由于波浪使水位在一定范围内波动,而不是停留在某个具体的数值上,样本数据具有一定的模糊性。在研究对象的观测值具有模糊性时,模糊集理论成为重要的方法。Zadeh(1965)首次提出了模糊集理论^[1],Tanaka等(1982)在此基础上,提出了模糊数据回归问题。模糊回归在许多领域得到了广泛的应用^[2]。

国内外对模糊回归的研究包括对线性和非线性回归模型的研究,总体上可分为三类:第一类是变量之间关系是具有模糊性,即需求解的回归系数是模糊的;第二类是因变量、自变量中的部分或者全部具有模糊性,即变量本身具有模糊性;第三类是变量与系数均具有模糊性。由于

第三类较复杂,大多数研究集中在第一类和第二类。

第一类和第二类模糊回归问题是第三类模糊回归问题的特殊情形,因此,探索更具一般性的模糊回归方法具有重要意义。本文讨论第三类模糊回归问题的参数估计,其变量与系数均具有模糊性的回归模型。

1 模糊回归模型

设因变量 $\tilde{Y}$ 为模糊变量,是某一组模糊自变量 $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n$ 的函数,其中: $f: \mathbb{R}^n \rightarrow E$, $\tilde{Y}_i = f(\tilde{x}_{i1}, \tilde{x}_{i2}, \dots, \tilde{x}_{in})$, i 为观测值的编号, n 为因变量个数。对于模糊线性回归,其目标是对下述模糊线性回归模型的模糊参数进行估计,以使得回归得出的因变量值与样本观测值的“最匹配”:

$$\tilde{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \tilde{x}_{i1} + \hat{\beta}_2 \tilde{x}_{i2} + \dots + \hat{\beta}_n \tilde{x}_{in} + u_i \quad (1)$$

非线性回归的目标是对下述模糊非线性模型的参数

作者简介: 彭宇文(1976-),男,湖南双峰人,博士,副教授,研究方向:数理统计,农业经济。

表5 全社会固定资产估计值及真实值

	均值估计量	均值方差	均值区间估计	总体估计量	总体方差	总体区间估计
真实值	8753.7	—	—	271365	—	—
PPS估计结果	8753.64	0.02387	8753.64 ± 0.309	271362.72	22.94	271362.72 ± 9.58
简单随机估计结果	9521	959559	9521 ± 2 × 1960	295151	922136199	295151 ± 60734

抽样总量估计值与真实值之间相差只有2.28,抽样方差只有22.94,而简单随机抽样估计量与真实值相差到达23786,方差达到922136199。因此,无论是抽样估计结果效率、抽样误差还是95%置信度条件下的区间估计,PPS估计的效率均高于简单随机抽样。在实际调研分析过程中,可以根据研究需要,充分既定利用条件,选用信度更高的方法。

参考文献:

[1]蔡泽林.PPS分层抽样时样本单位数的设计[J].湖北师范学院学报

(自然科学版),1996,(6).

[2]马慧敏,方国斌.改进PPS抽样设计效应的Monte Carlo模拟[J].统计与决策,2009,(11).

[3]张荷观.整群比估计与PPS抽样的比较[J].无锡轻工大学学报,2001,(7).

[4]李战江,闫在在. PPS估计与比估计的比较[J].内蒙古农业大学学报(自然科学版),2005,(4).

[5]俞纯权. PPS抽样下子总体参数的估计[J]. 统计与决策,2006,(18).

[6]卢静莉,闫在在,丁昌江. 多元简单随机抽样与二相抽样的最优分配[J]. 统计与决策,2010,(19).

[7]倪加勤,张勇. 调查内比估计效果的数据模拟分析——关于简单抽样设计与PPS系统抽样设计的比较[J].统计研究,2005,(11).

[8]Nibia,Aires. Comparisons between Conditional Poisson Sampling and Pareto π ps Sampling Designs[J]. Journal of Statistical Planning and Inference, 2000,(88).

(责任编辑/亦民)