文章编号：1002-0268（2005）04-0099-04

居民出行调查中的抽样技术研究

王京元，王 炜，程 琳
（东南大学交通学院，江苏 南京 210096）

摘要：居民出行调查的数据是城市交通规划的基础，它的准确度直接影响着规划的质量。恰当的调查方法和抽样率对保证数据精度和节省费用至关重要。目前，我国绝大多数城市在进行居民出行调查时，根据国外经验来确定抽样率，理论依据不充分，无法估计调查数据的精度和预测数据的准确性。运用抽样技术理论，首先对居民出行调查中用的抽样方法和抽样率的计算做理论上的阐述，然后对常用的抽样率计算公式进行推导，并讨论其他相关问题，最后给出计算实例。

关键词：居民出行调查；抽样方法；抽样率
中图分类号：U491 文献标识码：A

Study on Sampling Techniques of Trip Survey

WANG Jing-yuan, WANG Wei, CHENG Lin
(Transportation College of Southeast University, Jiangsu Nanjing 210096, China)

Abstract: The data of trip survey is the basis of urban traffic planning and its accuracy is crucial. The proper survey method and sampling rate are important for the survey data precision. The sampling rate is often based on the foreign experiences in China at present. The precision of survey data and the accuracy of the forecasting data can't be estimated for the absence of theoretic basis. The theories are given for sampling methods and the sampling rate calculation and the formulae are deduced by applying the sampling theory. Some other aspects concerned are discussed and the applications are explicit. The example illustrates the applications.

Key words: Trip survey; Sampling method; Sampling rate

0 引言

随着交通的发展，各种层次的交通规划越来越受到相关部门的重视。居民出行调查是为了了解现状交通的发生、吸引及其分布情况，把握 1 天内人的行动全体的综合性调查。交通调查的数据是城市交通规划的基础数据，利用其提供的信息，分析交通现状，并预测未来规划范围内的交通时空分布，为近期的交通治理、规划决策以及新一轮的综合交通战略规划服务，它的准确度直接影响着规划的质量。

抽样方法和抽样率直接关系到调查数据的精度。在调查目的和抽样方法确定的情况下，仅从保证数据精度的要求出发，抽样率一般与以下 3 个因素有关：总体参数的变异性、精度的要求（抽样误差和置信度）和总体的大小[1]。

自 20 世纪 80 年代起，国内若干大城市陆续开展了居民出行调查工作。而在确定抽样率时，绝大多数城市是根据国外经验和做法，并借鉴其他城市的取值来确定，随意性较大；很少考虑根据调查目标和调查方法来计算确定。这种理论依据不充分的做法无法估计调查数据的精度和预测数据的准确性。鉴于此，本文将运用抽样技术理论，从保证数据精度的要求出发，通过误差分析对抽样率的确定进行讨论。其他相关问题在另文中进行讨论。

收稿日期：2004-02-19

作者简介：王京元（1977—），男，山东淄博人，博士研究生，研究方向为交通运输规划与管理.

1 常用抽样方法及抽样率的计算

不同的调查目标对应不同的统计量，对样本的数量要求不同。常用的调查目标有：均值、总值、两个均值的比率或两个总体的比率、属于某一特定类的单位所占的比例。下面以均值为例，仅从调查精度对样本量要求的角度出发，用误差分析的理论对其计算进行讨论[2]。

对精度的要求通常以允许绝对误差(绝对误差限) d 或允许相对误差(相对误差限) r 来表示。

即对统计量 $\overline{Y}$ 及它的估计 $\bar{y}$，以绝对误差限表示为

$$P(|\bar{y}-\overline{Y}|\leqslant d)=1-\alpha \tag{1}$$

或以相对误差限表示为

$$P\left\{\frac{|\bar{y}-\overline{Y}|}{\overline{Y}}\leqslant r\right\}=1-\alpha \tag{2}$$

在对总体未做任何假定的情况下，$\bar{y}$ 的精确分布很难求得，已经证明，当 n 增大时，具有有限标准差的任何总体样本均值的分布趋于正态分布(这个结果是关于无限总体的)，这时绝对误差限

$$d=t\sqrt{V(\bar{y})} \tag{3}$$

而相对误差限

$$r=t\frac{\sqrt{V(\bar{y})}}{\overline{Y}}=tC_v(\bar{y}) \tag{4}$$

式中，t 为标准正态分布的双侧 α 分位数。在实际问题中，当 $\overline{Y}$ 未知时，可以用其估计量代替。

因此，根据对 d 或 r 的要求以及 $1-\alpha$ 所对应的 t 可推算出所需的样本量。

1.1 简单随机抽样

从一个简单随机样本所得的均值 $\bar{y}$ 的方差是

$$\sigma_{\bar{y}}^2=V(\bar{y})=E(\bar{y}-\overline{Y})^2=\frac{\sigma^2}{n}\frac{(N-n)}{N-1}=\frac{S^2}{n}\frac{(N-n)}{N-1}=\frac{S^2}{n}(1-f)=\frac{\gamma^2S^2}{n} \tag{5}$$

式中，$f=n/N$ 是抽样率。

在概率论中已经证明，从一个无限总体中抽取一个含量为 n 的随机样本，样本平均数的方差 $\sigma_{\bar{y}}^2=\frac{\sigma^2}{n}$。当总体有限时，唯一变化的是引入因子 $\gamma^2=(N-n)/(N-1)$。$(N-n)/(N-1)$和$\sqrt{(N-n)/(N-1)}$分别称为方差和标准差的有限总体校正系数(fpc)。假如 n/N 仍然是低的，fpc 就接近于 1，这样总体含量对样本均值的标准差就没有直接的影响。

根据不同精度的要求，将 $\sigma_{\bar{y}}$带入 d 或 r 的表达式，可得到不同条件下所需的样本量的计算公式。

(1)若要使 $\sigma_{\bar{y}}=\sqrt{\sigma_{\bar{y}}^2}\leqslant\delta$($\delta$事先给定)，则要求

$$n=\frac{NS^2}{N\delta+S^2} \tag{6}$$

(2)若要使 $P(|\bar{y}-\overline{Y}|\leqslant d)=1-\alpha$($\alpha$，$d$ 事先给定)，则要求

$$n=\frac{Nt^2S^2}{Nd^2+t^2S^2}=\frac{NS^2}{N\left(\frac{d}{t}\right)^2+S^2}=\frac{\gamma^2t^2S^2}{d^2} \tag{7}$$

(3)若要使 $P\left\{\frac{|\bar{y}-\overline{Y}|}{\overline{Y}}\leqslant r\right\}=1-\alpha$($\alpha$，$r$ 事先给定)，则要求

$$n=\frac{Nt^2S^2}{Nr^2\overline{Y}^2+t^2S^2}=\frac{NC^2}{N\left(\frac{r}{t}\right)^2+C^2}=\frac{\gamma^2t^2S^2}{r^2\overline{Y}^2} \tag{8}$$

式中，$C=\frac{S}{\overline{Y}}$为总体变异系数。

(4)若要使 $C<\upsilon$(即相对精度高于指定的指标 υ)，则要求

$$n=\frac{NC^2}{N\upsilon^2+C^2} \tag{9}$$

在实际应用中，通常对(7)式和(8)式进行如下处理。先计算 $n_0=\frac{t^2S^2}{d^2}$或 $n_0=\frac{t^2S^2}{r^2\overline{Y}^2}=\left(\frac{t}{r}\right)^2C^2$，则 $n=\frac{n_0}{1+\frac{n_0}{N}}$，如果$\frac{n_0}{N}<0.05$，则就取 n_0。

公式中的 S^2、C^2 是未知量，在实际工作中，可根据过去的资料或先进行少量抽样，对其进行预估，从而粗略地确定 n。

1.2 分层抽样

分层抽样就是将母体分为若干互不重复的类型(层次)，然后在每个层中分别独立的进行抽样。如果每层都是简单随机抽样，则称为分层随机抽样，所得的样本称为分层随机样本。

假如共分为 L 层，用下标 h($h=1, 2, \cdots, L$)表示层号，关于 h 层的记号如下：单元总数 N_h；样本单元总数 n_h；层权 W_h；抽样率 f_h；总体方差 S_h^2；样本方差 s_h^2。

下面重点讨论的是如何确定总的样本数 n 和每层的样本量n_h，记 $w_h=n_h/n$。

同样的样本量在各层中不同的分配(n_h 由调查者选取)，产生的标准差是不同的。根据分层随机抽样理论，可求得 n 的一般公式(10)和(11)，对其进行变换可得任何一种分配情况的总样本量

$$n=\frac{\sum\frac{W_h^2s_h^2}{w_h}}{V+\frac{1}{N}\sum W_hs_h^2} \tag{10}$$

其中，V 为总体平均值的方差，又根据 $V=\left(\frac{d}{t}\right)^2$，可得

$$n=\frac{\sum\frac{W_h^2s_h^2}{w_h}}{\left(\frac{d}{t}\right)^2+\frac{1}{N}\sum W_hs_h^2}=\frac{\sum\frac{W_h^2s_h^2}{w_h}}{\left(\frac{r\bar{Y}}{t}\right)^2+\frac{1}{N}\sum W_hs_h^2} \tag{11}$$

在居民出行调查中，各层经常采用相同的抽样率，即 $\frac{n_h}{N_h}=\frac{n}{N}$，$\frac{n_h}{n}=\frac{N_h}{N}$（$w_h=W_h$），也就是说按各层单元数占总体单元数的比例（各层的层权）进行分配，称作按比例分配的分层随机抽样。

由式（10）可得按比例分配的分层随机抽样的样本量为

$$n=\frac{\sum W_hs_h^2}{V+\frac{1}{N}\sum W_hs_h^2}=\frac{Nt^2\sum W_hs_h^2}{Nd^2+t^2\sum W_hs_h^2} \tag{12}$$

同样，若 fpc 可以忽略不计，则可记 $n_0=\frac{1}{V}\sum W_hs_h^2$；若不可以忽略，则 $n=\frac{n_0}{1+\frac{n_0}{N}}$。

若将 $\sum W_hs_h^2$ 看作广义的方差，则由式（7）和式（12）可以看出，简单随机抽样和分层抽样的抽样率计算公式可用统一的公式表示。

2　其他常用抽样目标抽样率的计算

以上是以均值或总值为控制目标来确定抽样率，在居民出行调查中还经常用如下统计量作为控制目标。下面仅给出样本量的计算公式。

2.1　比例调查

调查研究总体中某个特征的单元占总体的比例 P，并对其做出估计。

2.1.1　简单随机抽样

$$n=\frac{t^2\frac{PQ}{d^2}}{1+\frac{1}{N}\left(\frac{t^2PQ}{d^2}-1\right)}=\frac{Nt^2PQ}{(N-1)d^2+t^2PQ}=\frac{\gamma^2t^2PQ}{d^2} \tag{13}$$

或

$$n=\frac{t^2\frac{Q}{r^2P}}{1+\frac{1}{N}\left(\frac{t^2Q}{r^2P}-1\right)}=\frac{Nt^2Q}{(N-1)r^2P+t^2Q}=\frac{\gamma^2t^2Q}{r^2P} \tag{14}$$

式中，$Q=1-P$，$\gamma^2=\frac{N-n}{N-1}\approx\frac{N-n}{N}$ 称作有限总体的校正系数（fpc）。

在实际应用中，通常先计算 $n_0=t^2\frac{PQ}{d^2}$ 或 $n_0=t^2\frac{Q}{r^2P}$。

2.1.2　按比例分配的分层随机抽样

假设 $N_h-1\approx N_h$，则总样本量为

$$n=\frac{\sum W_hP_hQ_h}{V+\frac{\sum W_hP_hQ_h}{N}}=\frac{Nt^2\sum W_hP_hQ_h}{Nd^2+t^2\sum W_hP_hQ_h} \tag{15}$$

同理简化计算

$$n_0=\frac{\sum W_hP_hQ_h}{V}$$

若将 $\sum W_hP_hQ_h$ 看作广义的 PQ，则由式（13）和式（15）可以看出，简单随机抽样和分层抽样的抽样率计算公式可用统一的公式表示。

2.2　比率调查

比率估计不同于比例估计。比率涉及总体两个变量指标，这两个指标（包括分母）都需要通过样本进行估计；而比例估计涉及的总体大小是已知的（即样本容量），不需要估计。

比率估计是有偏估计，当样本量足够大时，估计的偏倚趋于零，因此，比率估计要有足够的样本量才能保证估计的有效性。

从严格意义上来讲，出行方式结构（各方式的出行次数和全方式总的出行次数均未知）、出行目的结构、各年龄组的人均出行次数以及各职业的人均出行次数均属于比率调查。但考虑到比率估计的复杂性，在居民出行调查中通常将比率估计当作比例估计处理，下面将对此做进一步讨论。

3　居民出行调查中常用的抽样率公式

3.1　日本的抽样率计算公式

日本继上世纪 60 年代以来陆续开展的居民出行调查工作为其积累了丰富的经验，其关于居民出行调查抽样率选取有以下经验公式[3]

$$r=t\left[\frac{1}{N-1}\frac{1-f}{f}\frac{Q}{P}\right]^{\frac{1}{2}} \tag{16}$$

式中，N 为一般代表城市的出行总量（人次）；P 为样本的某项特征或特征的某个部分在母集团中的出现概率（期望值），例如，不同目的、交通方式的出行比例，或某两小区之间的出行交换量占总出行量的比例，最好通过预调查或相关调查确定；r 为容许误差，根据调查

精度要求确定，一般不超过 20%。

将上述公式变换，得到的即是上面简单随机抽样的比例估计样本公式(14)

$$n=\frac{Nt^2Q}{(N-1)r^2P+t^2Q}$$

在此需要说明两点：(1)从上面对公式中符号的解释可以看出，此公式可用来计算如出行方式结构和目的结构以及出行交换量比例的抽样率，而严格上讲，这些项目应当按比率估计处理，这里将其按比例估计来近似处理。(2)以上总体的单位是人次，而不是抽样单位人，所以在计算时，可先计算出所要抽取的人次，再除以人均出行次数即可得到要调查的人。

3.2 美国的抽样率计算公式

$$n=\left(\frac{t}{r}\right)^2C^2=\frac{t^2S^2}{d^2}=\frac{t^2S^2}{r^2\overline{Y}^2} \qquad (17)$$

根据上面所讲的抽样率计算公式，可以将此公式看作简单随机抽样和分层抽样在忽略 fpc 时的均值抽样率计算公式，如果考虑 fpc 的影响，就是 $n=\frac{Nt^2\sigma^2}{Nd^2+t^2\sigma^2}$。对于不同的抽样方法，方差的表达式不同，分层抽样时方差的表达式[4]应为 $\sigma^2=\frac{1}{N}(\sigma_1^2N_1+\sigma_2^2N_2+\cdots+\sigma_n^2N_n)$，($\sigma_i^2$ 为各层内部的方差)，整理后为 $\sigma^2=\sum W_k s_k^2$。这一公式在交通书籍有关抽样调查的部分很常见。

4 抽样率设计

4.1 设计抽样率

上面公式计算的是理论最小抽样率，而在实际调查时采用的是设计抽样率。理论最小抽样率与设计抽样率之间的差异主要表现在表格的回收率和回收表格的有效率，设计抽样率与理论抽样率的关系如下

$$f=k_1k_2f_d=kf_d;\ f_d=\frac{f}{k_1k_2}=\frac{f}{k}$$

式中，f 为理论抽样率；k_1 为表格回收率；k_2 为回收表格有效率；f_d 为设计抽样率；k 为表格有效回收率。

这样计算出的设计抽样率只能够保证数据精度的最低要求，还要考虑到其他的不可见因素（如无回答现象和伪劣数据等）以及替补样本的需要，对设计抽样率适当扩大，为最终推荐采用的抽样率。

4.2 抽样率敏感性分析

根据抽样率的计算原理，要计算抽样率先要预先确定公式中的其他指标（例如方差和离散系数），这些指标往往通过小规模的调查以前的调查、或借鉴其他城市的数据来确定，必然存在一定的误差，从而影响抽样率的计算精度。分析这些因素的变化对抽样率精度的影响程度，就是敏感性分析。

例如，以交通大区之间的公交 OD 交换量为调查目标确定抽样率，精度要求为：交通大区之间的公共交通出行总量在 95%的置信度下，相对误差不超过 10%。此时，抽样率只与公交方式结构比例及居民出行总量有关，假设居民出行总量和公交方式比例结构分别在可能的范围内上下浮动一定的数值，则计算抽样率必然会发生一定的变化。

分析在这种情况抽样率的变化情况。如果抽样率的变化比例与居民出行总量以及公交方式结构比的变化比例在同一个量级上，认为敏感性一般，能够满足精度的要求；否则，认为比较敏感。在敏感性较强的情况下要保证调查的精度，则要大幅度提高抽样率，但如果抽样率太高，致使调查费用太高和后续工作太重，或根本无法开展正常的调查工作，则应考虑对调查方案进行调整，甚至取消此项调查。

5 计算实例

东南大学交通学院与徐州市交巡警支队于 2003 年 9 月 23 日在徐州市共同开展了大规模的城市人口出行调查。将调查范围划分为 69 个交通小区，共有人口 98.88 万，小区平均人口规模为 1.43 万。

本次调查发放居民调查表 15 000 份，回收 13 271 份，其中有效表格 11 282 份，设计抽样率为 1.52%，实际抽样率为 1.14%；回收率为 88.47%；有效率为 85%；有效回收率为 75.21 %。据统计，徐州市区居民（全部被调查者）人均出行次数为 3.39 次/（人·日）。

现将上述资料作为一次预调查的数据，对抽样率重新进行计算。

（1）以人均出行次数为调查目标

精度要求：人均出行次数在 95%的置信度下相对误差不大于 10%，即 $t=1.96$，$r=0.1$。采用简单随机抽样，整个调查范围的离散系数为 0.87，得样本量 n 为 291 人，理论抽样率仅为 0.03%，设计抽样率为 0.04%。若以小区为单位，小区的平均离散系数为 0.73，用简单随机抽样公式得到，每个小区所需的样本量为 205 人，共 14 145 人，理论抽样率为 1.43%，设计抽样率为 1.90%。

（2）以三大出行方式比例为调查目标

以简单随机抽样公式对以调查范(下转第 107 页)

表 4　理论计算结果检验

	东	南	西	北	总量
观测近似值/ $pcu \cdot h^{-1}$	528	416	500	376	1820
理论计算值/ $pcu \cdot h^{-1}$	550	473	516	397	1936
绝对误差/ $pcu \cdot h^{-1}$	22	57	16	21	116
相对误差/ %	4.17	13.70	3.20	5.59	6.37

注：观测近似值是对高峰时段（15*min*）饱和流量的观测值进行扩样得到。

从检验结果可见，理论计算值与观测所得近似值间的相对误差均在 15%以内，通行能力总量值的相对误差仅为 6.37%，说明前述计算方法具有较高的精度。同时还可以发现，各进口观测流量值均不同程度地小于理论计算值；同时南北两进口的相对误差大于东西两进口，其中南进口相对误差达到 13.7%，明显大于其它三个进口。原因在于：（1）该交叉口内仍存在一定流量的非机动车，机动车的行驶不可避免的要受到非机动车的影响而导致各进口通行能力的下降；（2）在非机动车遵循机动车信号放行规则的情况下，对向非机动车流主要干扰本面直行、左转机动车流，而本面非机动车流主要干扰本面右转机动车流。该交叉口南北进口非机动车流量较东西进口要大，因此相对误差也较大。北进口的非机动车流量最大，对南进口的直行、左转车流产生较大影响而导致相对误差明显偏高。

4　结语

对单向交通条件下信号交叉口通行能力的计算问题进行了探讨，给出不同相位、不同单向交通组织形式下通行能力的计算公式。通过实例计算比较，证实给出的计算公式具有较高精度，可以作为今后各交通标准及规范制定的参考。

参考文献：

[1] Transportation Research Board Highway Capacity Manual 2000 [R]. Washington, D.C., USA: National Research Council, 2000.
[2] 中国公路学会《交通工程手册》编委会. 交通工程手册 [M]. 人民交通出版社, 1998
[3] 李岚. 城市道路单向交通设置方法研究 [D]. 长春：吉林大学, 2001
[4] 李荣波. 关于单向交通通行能力的研讨 [J]. 中国市政工程, 1998 (3)：1－6
[5] 裴玉龙. 道路通行能力 [M]. 哈尔滨：哈尔滨建筑大学, 1996.
[6] 东南大学交通学院. 温岭市城市道路交通管理规划 [R]. 南京：东南大学出版社, 2003.

（上接第 102 页）

围为整体和小区为单位的抽样率分别进行计算（精度要求同上），结果如表 1。

表 1　三大出行方式所需的样本量和抽样率

出行方式	步行	自行车	公交车
出行量比例/ %	21.80	54.71	14.69
理论抽样率 *	2.76% (396) / 0.042% (407)	0.66% (94) / 0.01% (94)	4.40% (630) / 0.07% (658)

*分子是以小区为单位，分母是以整个调查范围为整体；括号内为样本量。

6　结论

（1）对居民出行调查中用到的抽样方法和抽样率的计算做了理论上的阐述，并对常用的抽样率计算公式进行了推导。

（2）不同的调查目的对应着不同的抽样方法，根据精度的要求确定样本量的大小。

（3）实际调查时要考虑其他影响因素对理论抽样率进行修正，得到设计抽样率，即为实际应用的抽样率。

（4）对抽样率进行敏感度分析，可得知调查方案是否可行，为方案调整提供依据。

目前，我国在进行居民出行调查确定抽样率时多是凭经验，并且很少得到人们的重视，研究者更是寥寥无几。希望通过本文的初步尝试，引起人们的注意，有更多的交通工作者和统计学者来共同从事这方面的研究。

参考文献：

[1] 毛保华，曾会欣，袁振洲. 交通规划模型及其应用 [M]. 北京：中国铁道出版社, 1999.
[2] W.G. 科克伦，著. 张尧庭，吴辉，译. 抽样技术 [M]. 北京：中国统计出版社, 1985.
[3] 北京市城市交通综合调查 [R]. 北京：北京城市规划设计研究院, 2002.
[4] 王炜，过秀成. 交通工程学 [M]. 南京：东南大学出版社, 2000.