六种抽样模型◎智能化的样本量计算方法 适用不同抽样阶段◎满足不同层次抽样用户需求

空间抽样与统计推断

王劲峰 姜成晟 李连发 胡茂桂 著

高效·智能·专业

科学出版社
www.sciencep.com

空间抽样与统计推断

王劲峰　姜成晟　李连发　胡茂桂　著

科学出版社
北京

内 容 简 介

抽样方法广泛运用于资源环境和社会经济调查之中。相对于普查，抽样调查具有费用低、速度快等特点。本书介绍了经典抽样、考虑空间关系的空间抽样和 Kriging 估计等理论。结合具体案例，介绍了各主要抽样模型的实际运用步骤。阅读本书只需概率统计的基本知识即可。

本书可供地学和社会科学领域的学者在抽样调查、统计推断和监测网优化设计中参考使用。

图书在版编目(CIP)数据

空间抽样与统计推断/王劲峰等著. —北京：科学出版社，2009
ISBN 978-7-03-024450-5

Ⅰ. 空… Ⅱ. 王… Ⅲ. 空间测量-信息系统-抽样-统计分析
Ⅳ. P208

中国版本图书馆 CIP 数据核字(2009)第 060959 号

责任编辑：韩 鹏 朱海燕 赵 冰/责任校对：李奕萱
责任印制：钱玉芬/封面设计：王 浩

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

新蕾印刷厂 印刷
科学出版社发行 各地新华书店经销

*

2009 年 5 月第 一 版 开本：B5(720×1000)
2009 年 5 月第一次印刷 印张：12
印数：1—2 000 字数：229 000

定价：38.00 元

(如有印装质量问题，我社负责调换〈新欣〉)

前　言

抽样调查是用抽样数据推断调查对象的属性，这一技术已广泛运用于资源、环境、经济和社会等调查之中。相对于普查而言，抽样调查具有费用低、速度快和精度高等优点。一个好的抽样调查方案可以用较少的样本量获得较高精度的统计推断。

抽样调查分为精度后验抽样调查和精度先验抽样调查两大类。前者凭经验以某种方式（如简单随机、系统、分区等）抽取一定数量的样本；然后，据此推断总体的总量、均值；最后，对估值的精度进行检验。后者根据抽样理论，在获得调查对象的离散方差、比率等信息的前提下，计算区域调查所需样本量和估值精度之间的理论关系；据此，当给定精度要求时，估算调查所必需的样本量；当给定样本量后，计算估值精度，从而形成精度先验的抽样方案，并实施野外抽样；最后，计算得到估计值。精度先验抽样形成抽样调查的理论和技术，它能在外业之前对样本量和估值精度做出初步判断，降低抽样的不确定性。

精度先验的抽样调查可进一步划分为经典抽样和空间抽样。前者以 1977 年 Cochran，W. G. *Sampling Techniques*. John Wiley & Sons. 专著为代表，其理论建立在样本相互独立的假设之上。经典抽样可以用于空间分布对象的调查，虽然输入简单，较易使用，但效率较低。空间抽样调查考虑样本的空间相关性和空间异质性，效率较高。

本书试图对空间抽样的理论与实践进行归纳总结；介绍了作者提出的空间抽样三明治模型、异质表面均值 Kriging 模型、空间抽样最优决策 Trinity 理论，以及作者开发的空间抽样软件包 SSSI（网站地址：www. sssampling. com）等。读者可以据此完成以下任务：

（1）设计优化的抽样调查方案或监测网络（如环境、人口、经济和流行病等），计算最佳样点分布和密度，形成高效的空间抽样方案或监测网络；

（2）给已获取的样本或已存在的监测网络（如气象站、生态样方等）推荐最佳估值方法和监测改进建议（根据监测对象特点和样点分布）；

（3）对已发表的统计数据（如区域污染指数、区域社会经济指数等），评价其精度、可靠性（考察对象特征、样点分布、密度和估值方法）；

（4）一次抽样，多单元报告、多系统报告（按行政单元、自然单元和格网等）。

本研究得到国家自然科学基金、“863”计划、国家科技支撑、国家重大科技专项、中国科学院、国家留学基金、Marie Curie Fellowship 的支持；参与研究的还有

葛咏、曹志冬、马爱华、郭燕莎、姜新利、刘鑫、王娇娇和冯晓磊等；钟耳顺、宋关福在SSSI软件研发的地理信息系统部分给予了宝贵的帮助；李拴科、高小青帮助SSSI软件的平面设计；宋长青、周成虎、刘高焕、庄大方、王英杰、吴炳方、李强子、周清波、张科利、黎夏、董玉祥、唐守正、赵宪文、刘纪远、冯士雍、王道辰、闾国年、史文中、Robert Haining和George Christakos教授等对本书给予了指导、支持和帮助。在此一并表示衷心感谢。

王劲峰

2009年1月26日春节

目　录

第一篇　空间抽样原理

第二篇 Kriging 估计

第三篇 空间抽样软件包 SSSI 及案例

0 引　论

0.1 举　例

和顺县位于山西省中东部，由 $N=326$ 个行政村组成(图 0.1)，2005 年，抽取一定数量村庄的人口数，用以推断全县人口数量。

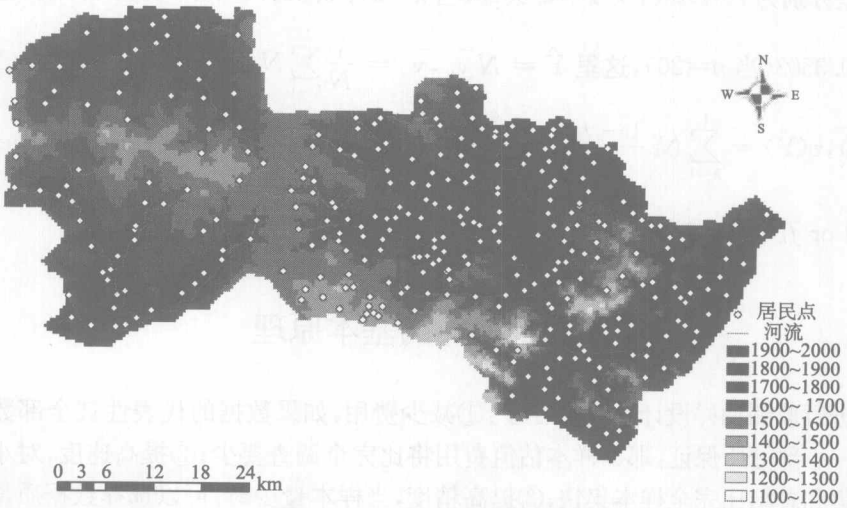

图 0.1　山西省和顺县居民点及高程

0.1.1 简单随机抽样

以行政村为最小抽样单元，分别随机抽取 $n=5,10,15$ 和 20 个村庄作为样本单元，调查各抽样村的人口数目；将样本值简单相加，乘以抽样率倒数(N/n)，即得全县人口总数。和顺县总体人口数估计的相对误差分别为 $r=t\times s(\hat{Y})/\hat{Y}=1.0512$(当 $n=5$)，0.6680(当 $n=10$)，0.4046(当 $n=15$)，0.2838(当 $n=20$)，这里 $\hat{Y}=N\bar{y}$，$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$，$s(\hat{Y})=\sqrt{v(\hat{Y})}$，$v(\hat{Y})=N^2v(\bar{y})$，$v(\bar{y})=\frac{1-f}{n}s^2$，$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$，$y_i$ 是第 i 个村的人口数，t 是 student-t 值，这里取 1.96 对于

95%的置信度。

0.1.2 分层随机抽样

人口数量的空间分异被发现与各村工、农业比例有关，为此首先根据 2001 年和顺县普查的工、农业产值比将和顺县所有 326 个村庄聚分两类。然后，在 2005 年通过人口抽样调查，推断和顺县人口总量。分别随机抽取 $n=5$，10，15，20，30，40 和 50 个村庄作为样本单元，这些样本单元按面积比例被分配到两个聚类层中，调查各抽样村的人口数，用以推断全县的人口数量。和顺县总体人口数估计的相对误差分别为 $r=t\times s(\hat{Y})/\hat{Y}=0.9629$(当 $n=5$)，0.5017(当 $n=10$)，0.3861(当 $n=15$)，0.3503(当 $n=20$)，这里 $\hat{Y}=N\bar{y}_{st}$，$\bar{y}_{st}=\frac{1}{N}\sum_{h=1}^{L}N_h\bar{y}_h$，$\bar{y}_h=\frac{1}{n_h}\sum_{i=1}^{n_h}y_{hi}$，$s(\hat{Y})=\sqrt{v(\hat{Y})}$，$v(\hat{Y})=\sum_{h=1}^{L}N_h^2\frac{1-f_h}{n_h}s_h^2=N^2v(\bar{y}_{st})$，$v(\bar{y}_{st})=\sum_{h=1}^{L}W_h^2\frac{1-f_h}{n_h}s_h^2$，$W_h=n_h/n=N_h/N$ or $f_h=n_h/N_h=n/N=f$，$s_h^2=\frac{1}{n_h-1}\sum_{i=1}^{n_h}(y_{hi}-\bar{y}_h)^2$。

0.2 抽样的基本原理

进行监测抽样设计的优点在于：①减少费用，如果数据的代表性被全部数据集中的一小部分所保证，那么样本估值费用将比完全调查要少；②提高速度，对小样本的收集和总结比完全样本集快；③提高精度，当样本量少时，可以选择数据质量更好的样本，并更加集中精力于少量样本采集处理以提高样本质量，基于高质量的小样本量估计，有可能比大样本估计精度更高。

0.2.1 抽样目的

从调查区域 A 中抽取 n 个样本单元，用于估计区域属性均值或总量、空间插值和绘制地图等不同目的，对应不同的抽样或监测网的误差评价指标。

目的 1：估计区域均值或总量，均值的估计误差可用区域均值方差来度量，即

$$\sigma_n^2=E\left[\frac{1}{n}\sum_{i=1}^{n}y_i-\frac{1}{A}\int_A y_i\mathrm{d}i\right]$$

式中：E 是数学期望；n 为样本单元数目；y_i 为第 i 个样本单元的属性值；$(1/A)\int_A y_i\mathrm{d}i$ 是区域 A 可观测的总体均值，它可以由 n 个样本单元的数学平均值

$(1/n)\sum_{i=1}^{n} y_i$ 来估计。这一内容形成抽样理论(Cochran,1977; Haining,2003; 王劲峰等,1999)。

目的 2:空间插值和绘制地图,则区域插值误差可用以下公式来度量,即

$$\sigma^2 = E(\sum_{i\neq j}^{n} w_i y_i - y_j)^2$$

未抽样点 y_j 的属性值用其他抽样点的加权平均 $\sum_{i\neq j}^{n} w_i y_i$ 来估计,使 σ^2 最小化的 $\{w_i\}$ 为权重。这一内容形成 Kriging 理论(Issaks and Srivatava,1989; Christakos,2005)。

目的 3:估计区域某些特征值,如离散方差、空间相关性、半变异函数,等等。其误差研究较少(Christakos,2005)。

0.2.2 抽 样 效 率

一个好的抽样方案,既要效率高,又要输入和计算尽量少,有较高的"性能价格比"。而这种性能价格比是随不同调查对象而变化的,需要灵活选择。

以上两类估值方差,可以进一步通过调节样本单元的空间抽取方式和空间分布位置$\{i\}$,使其达到最小,即抽样效率最高。即在给定样本量 n 的前提下,使区域均值方差 σ_n^2 或区域插值误差 σ^2 最小化;或者在给定区域均值方差 σ_n^2 或区域插值误差 σ^2 的前提下,使样本量 n 最小。

据 Rodriguez-Iturbe 等(1974),区域样本均值方差为

$$\sigma_n^2 = E[\frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{A}\int_A y_i \mathrm{d}i]^2 = \sigma_p^2 F(n)$$

式中:σ_p^2 是区域离散方差(dispersion variance),即 $E(y_i - Ey_i)^2$,可以通过样点估计;$F(n)$是方差降低因子,由下式估计

$$\begin{cases} F(n) = 1/n & \text{不考虑空间关联性} \\ F(n) = \{1 - E[r(y_i, y_j) \mid A]\}/n & \text{考虑空间关联性,随机抽样} \\ F(n) = \{1 - E[r(y_i, y_j) \mid A/k]\}/n & \text{考虑空间关联性,分层抽样} \end{cases}$$

式中:$E[\# \mid \cdot]$是"#"以"·"为条件的期望值;n 是抽样单元数目;k 是分层数目;$r(y_i, y_j)$是 y_i 和 y_j 之间的空间关联性。

根据上式,可以比较三种抽样策略的效率。令 n_r, n_s, n_0 分别表示考虑空间相关性条件下的随机抽样、考虑空间相关性条件下的分层抽样,以及不考虑空间相关性条件下的随机抽样的样本容量。假设给定样本均值精度 σ_n^2,则由上式得到

$$\frac{n_r}{n_0}=1-E[r(y_i,y_j)|A]$$

$$\frac{n_s}{n_0}=1-E[r(y_i,y_j)|(A/k)]$$

$$\frac{n_s}{n_r}=\frac{1-E[r(y_i.\,y_j)|(A/k)]}{1-E[r(y_i,y_j)|A]}$$

一般情况下，空间分布的属性距离越近，其间的相关性越强，即 $0<E[r(y_i,\ y_j)|A]<E[r(y_i,y_j)|(A/k)]$。因此，$n_s<n_r<n_0$，即在同样估值精度要求下，空间分层抽样所用的样本量最小，空间随机抽样次之，简单随机抽样所需样本量最多；类似地，给定样本量，容易得到空间分层抽样均值方差最小，空间随机抽样次之，简单随机抽样样本均值方差最大。

0.3 本书结构

本书内容包括空间抽样原理、Kriging 估计和空间抽样软件包 SSSI 及案例等三篇。第一篇的第 1 章经典抽样原理是空间抽样的起源，介绍了三种常用的抽样模型，包括简单随机抽样、系统抽样和分层抽样。空间抽样是在经典抽样的基础上发展起来的，考虑了空间相关性和空间异质性，包括空间简单随机抽样、空间系统抽样、空间分层抽样和空间三明治抽样四个模型，其中第 14 章为系统抽样，论述了通过等间距布样，按简单随机统计的简单的系统抽样方法；当样本按空间随机统计时，为空间系统抽样，特此说明，并且不再将“空间系统抽样”单独成章列出。空间三明治抽样模型可以用较少的样本对空间异质分布目标进行多单元报告。第二篇 Kriging 估计介绍基于 Kriging 理论的空间插值（第 7 章）和均值（第 8、9 章）推断方法，虽然需要较大的计算量和较多的参数输入，但具有无偏和最优的特点。第三篇空间抽样软件包 SSSI 及案例介绍了由王劲峰研究组研发的（中文网址：http://www.ssampling.com；英文网址：http://www.sssampling.org）空间抽样软件包“三明治空间抽样与统计推断软件包 SSSI”（sandwich spatial sampling and inference software），包括第一篇介绍的三种常见的经典抽样模型和四种常见的空间抽样模型，并附有使用说明和案例。表 0.1 叙述了本书各章的适用条件、特点和各章之间的关系。在经典抽样理论中（Cochran，1977），系统抽样模型考虑了样本之间的关联性（见 1.2），因此也适合于空间系统抽样（见 2.2）；抽样估计包括抽样获取数据和使用这些数据进行统计估计两部分，本书关于系统抽样的案例（第 14 章）涉及第一部分。

表 0.1 本书结构

均值估计	第一篇 第二篇 原理	第三篇 案例	模型	地表类型			均值		报告单元	
				静态	非静态	自相关	无偏	最优	单	群
抽样模型	第 1 章	第 13 章	简单随机	•			•		•	
	第 1 章	第 14 章	系统	•			•		•	
	第 1 章	第 15 章	分层		•			•—	•	
	第 2 章	第 16 章	空间简单随机			•			•	
	第 2 章	第 14 章	空间系统	•		•			•	
	第 2 章	第 17 章	空间分层		•	•		•—	•	
	第 3 章	第 18 章	空间三明治		•	•		•—	•	•
Kriging	第 7 章		经典 Kriging	•		•			•	
	第 8 章		静态 Mean-Kriging	•		•	•	•	•	
	第 9 章		非静态 Mean-Kriging		•	•	•	•	•	
核心技术	第 4 章		先验知识							
	第 5 章		单点属性							
	第 6 章		最优决策							
软件	第 10 章		SSSI 理论基础							
	第 11 章		SSSI 技术构架							
	第 12 章		SSSI 使用步骤							

注：• 为具有；•— 为趋向具有

具有空间分布的抽样越来越多地借助地理信息系统(GIS)技术。抽样理论中分层(stratification)是一种常用的技术，其英文“stratification”及其对应的中文“分层”在统计学中均是标准用语；地理信息系统中“layer”及其中文“分层”、“图层”也是标准用语。Stratification 与 layer 含义完全不同(前者指将总体划分为若干个次级总体，后者指存储于 GIS 的单一地理要素)，但在中文用同一词“分层”，无法分辨。本书使用这两个概念时，在可能产生误解的情形下，我们同时标注英文用词，以便区分。以下是本书用到的软件包：

空间抽样软件包 SSSI(中文版)：http://www.sssampling.com

空间抽样软件包 SSSI(英文版)：http://www.sssampling.org

空间统计软件包 GeoDA：https://www.geoda.uiuc.edu

空间贝叶斯建模软件包 GeoBUG：http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/geobugs.shtml

空间聚集探测软件包 Crimestat：http://www.icpsr.umich.edu/CRIMESTAT

空间扫描探测软件包 SatScan：http://www.satscan.org

第一篇　空间抽样原理

第一篇　空间构作原理

第1章　经典抽样原理

经典的抽样模型按抽样框来划分，主要有三种，即简单随机抽样、系统抽样和分层抽样。简单随机抽样不考虑空间关联，而系统和分层抽样主要在抽样框方面有所改进，一般较随机抽样精度有所提高。此外，估计某类事物所占总体的比例为目标的抽样，其计算公式中的离散方差可以用该比例值与其补数相乘得到，计算过程简单，在经典统计学中称为成数抽样或比例值抽样（Cochran，1977；冯士雍和施锡铨，1996；吴炳方，2000）。本章以森林遥感抽样调查为例，分别说明简单随机抽样、系统抽样和分层随机抽样三种模型。

1.1　简单随机抽样

简单随机抽样是经典方法中最基本、最简单的抽样模型，它不考虑空间关联，是其他抽样模型的基础。简单随机抽样模型根据调查对象总体的变化情况和用户希望抽样调查误差控制在某个范围之内，计算出样本量，然后根据样本量，从总体中随机抽取样本。对样本值调查以后，按照求解均值和方差的公式对抽样总体进行统计推断，下面将从计算样本量和统计推断两个方面介绍简单随机抽样模型计算方法。

1.1.1　刻度值样本容量 n 的计算方法

1）$n=\dfrac{\sigma^2}{V}$

式中：σ^2 是总体的方差，是由用户输入的；V 是希望得到的样本估值的方差，也就是抽样以后希望得到的总体估值的方差，也是由用户输入的。

因此在这种方法下，输入就是 σ^2 和 V。

适用条件，要求同时满足：

（1）样本无限大；

（2）有放回；

（3）抽样比很小（抽样比是样本和总体的比值）。

2）$n=\dfrac{n_0}{1+n_0/N}$，其中，$n_0=\dfrac{\sigma^2}{V}$