文章编号: 1001 - 9081(2011) S1 - 0185 - 03

基于 B/S 架构的复杂抽样调查统计推断系统

罗智超[1,2],管河山[3],曹礼华[4]

(1. 厦门大学 王亚南经济研究院,福建 厦门 361005;　2. 厦门大学 教育部计量经济学重点实验室,福建 厦门 361005;
3. 南华大学 经济管理学院,湖南 衡阳 421001;　4. 深圳供电局,广东 深圳 518001)

(zhichao. luo@ gmail. com)

摘　要: 基于 B/S 构架的复杂抽样调查统计推断系统采用跨平台设计,与业务系统及数据库无缝链接。系统提供随机抽样、分层抽样、Neyman 分层抽样、不等概率抽样(PPS) 、多阶段抽样等常用抽样算法。系统用户根据研究目标自定义组合抽样方法,系统根据样本和总体属性及用户抽样方案自动推算统计推断结果及置信区间,实现抽样调查、统计推断的自动化与系统化。

关键词: 抽样调查; 统计推断; B/S 构架

中图分类号: TP311; C81　　**文献标志码:** A

Statistic inference system for complex sample survey based on B/S framework

LUO Zhi-chao[1,2], GUAN He-shan[3], CAO Li-hua[4]

(1. *The Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen Fujian* 361005, *China*;
2. *Key Laboratory of Econometrics University of Ministry of Education, Xiamen University, Xiamen Fujian* 361005, *China*;
3. *School of Economics and Management, Nanhua University, Hengyang Hunan* 421001, *China*;
4. *Shenzhen Electric Power Bureau, Shenzhen Guangdong* 518001, *China*)

Abstract: A statistic inference system of complex sample survey based on B/S framework used cross platform construction, and can interconnect business system and database seamlessly. The system includes usual sampling methods, such as random sampling, stratified sampling, Neyman stratified sampling, Unequal Probability Sampling (PPS) and multi-stage sampling. Users can choose user-defined sampling method; system will calculate and inference the population score based on the information input into the system including population detail, sample size, sample methods and etc. Such system can make sampling and inference process automatically and systematically.

Key words: sample survey; statistics inference; B/S framework

0　引言

抽样调查已广泛应用于社会、经济、科技、自然等领域,是研究人员获取统计数据的重要手段之一。研究人员通常使用通用统计软件的抽样模块来进行抽样调查分析。传统的操作方式是先将业务数据导入到统计软件,再通过通用统计软件的抽样模块进行抽样,最后将调查结果数据导入到统计软件进行统计推断。这种方式有以下几个缺点: 1) 该方式对具体操作人员的统计知识及软件编程知识要求较高,任何一个环节的一个代码失误都将引起推断偏误; 2) 该方式无法实时获得复杂抽样涉及的许多辅助变量,而这些变量常常需要实时查询业务数据库才能获取; 3) 该方式处理数据记录在百万级别以上的业务数据时效率较低,需要手动操作并且浪费大量的时间用于数据传输; 4) 该方式的低效率手动操作方法不适合于周期性抽样调查业务。而抽样调查方案与企业业务系统的无缝链接能够解决以上问题,因此必将成为企业决策支持系统建设应用的一个新的需求方向。文献[1]将抽样调查算法与国家旅游局旅游调查系统相结合,文献[2-3]将抽样调查算法与电力系统客户及稽查业务系统相结合。本文以深圳供电局营业稽查抽样决策支持系统中的复杂抽样调查统计推断系统为例介绍基于业务系统底层数据库设计的抽样统计推断算法和系统,实现抽样统计推断的自动化与系统化。

1　系统概述

深圳供电局营业稽查抽样决策支持系统主要功能为每个月从深圳供电局下属 6 个区局上一个月的 4 项业务近 200 万笔业务数据中抽取大约 6000 个样本,再根据样本数据进行实地检查录入样本得分,系统根据抽样方案及样本得分推断各个区局各项业务的得分,及整个供电局各项业务的总体得分及置信区间。该系统自 2008 年 1 月正式运行以来,已实现 34 批次,近 20 万抽样样本的调查研究。

复杂抽样调查统计推断系统主要包括以下模块: 1) 业务数据库接口模块。考虑到系统的扩展性与通用性,该模块设置了待抽样业务的数据库表及字段名,一旦系统抽样业务类型发生变化只要修改相应的库表字段名即可。2) 抽样样本量估算模块。样本量的估算取决于业务量、置信度、误差率、调查成本等信息,该模块可以动态调整以上信息,保证样本量的估计满足抽样需求的变化。3) 抽样方案设定模块。根据不同业务类型设定相应的抽样方法及样本分配方法。4) 总体推断模块。根据抽样方案的设置对总体进行均值估计和比率估计,并计算相应的置信区间[3]。

复杂抽样调查统计推断系统主要流程为: 1) 初始化抽样调查成本用于估算样本量; 2) 设置抽样框时间范围、业务范围及部分范围,计算待抽样总体数; 3) 根据不同业务选择样

收稿日期: 2010 - 11 - 12; 修回日期: 2011 - 02 - 15。　基金项目: 国家自然科学基金资助项目(70971113) 。

作者简介: 罗智超(1977 -) ,男,福建建阳人,工程师,硕士,主要研究方向: 计量经济、统计建模;　管河山(1981 -) ,男,湖南衡阳人,讲师,博士,主要研究方向: 数据挖掘、统计建模;　曹礼华(1967 -) ,男,湖南郴州人,工程师,硕士,主要研究方向: 电力稽查、电力管理。

本分配方法及抽样方法;4) 录入样本调查得分;5) 计算样本得分并推断总体得分。

2 系统抽样及推断算法

复杂抽样调查统计推断系统包括了常用的样本分配、抽样及推断算法,用户可以根据业务特性选择不同的组合。主要包括:样本量估算、随机抽样、分层抽样、Neyman 分层抽样、不等概率抽样、多阶段抽样等常用抽样算法。

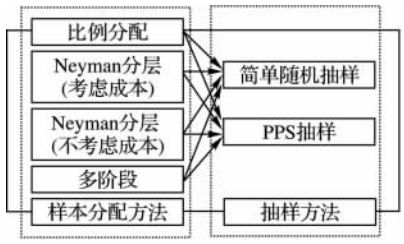

图 1 样本分配及抽样方法

2.1 样本量估计算法

系统提供三种样本估算方法供用户选择:1) 基于抽样成本估算;2) Neyman 分层(不考虑成本)估算;3) Neyman 分层(考虑成本)估算。实际调查过程中由于调查成本投入的限制,基本上无法满足理想状态下的 Neyman 分层(不考虑成本)样本量估算方法,而采用 Neyman 分层(考虑成本)估算方法[4-5]。

2.1.1 基于抽样成本的样本量估计

计算抽样的总工作量(例如,每个月投入 2 个工作人员进行调查),记为 T;计算平均调查时间,记为 t,总抽样样本总量为 $n = T/t$。

2.1.2 基于 Neyman 分层的样本量估计

不考虑成本的 Neyman 分层样本量估计,需要两个输入参数:某置信度下的 t 值(置信区间默认为 95%),相对误差 r(默认为 10%)。此时 Neyman 分层对样本量的计算可以根据不同 r 和 t 来调整,通常是调整 r 的取值。Neyman 对某个业务(例如,业扩业务)的抽样样本量 n_i 的计算公式如下:

$$n_i = \frac{\left(\sum_{h=1}^{L} w_h s_h\right)^2}{\left(\frac{r\bar{Y}}{t}\right)^2 + \frac{\sum_{h=1}^{L} w_h s_h^2}{N}} \tag{1}$$

考虑成本的 Neyman 分层样本量估计,假定费用满足 $C = c_0 + \sum_{h=1}^{L} c_h n_h$,其中:$c_0$ 为基本的费用,c_h 是该业务下对第 h 层进行单位抽样的成本,C 为总成本。费用作为输入参数,直接影响到抽样的样本量。Neyman 对某个业务(第 i 种业务)的抽样样本量 n_i 的计算公式如下:

$$n_i = (C - c_0)\frac{\sum_{h=1}^{L} w_h s_h}{\sum_{h=1}^{L} w_h s_h c_h} \tag{2}$$

2.2 样本分配方法

系统样本分配步骤如下:1) 采用 Neyman 分层方法将总样本在不同业务之间进行分配;2) 将各个业务中分得的样本在不同业务部门中进行分配,采用的方法也是 Neyman 分层方法。

2.3 抽样算法

系统提供两种抽样方法:随机抽样和不等概率抽样。不等概率抽样又分为有放回的和不放回的两种抽样方法。有放回的抽样主要有 PPS 法,不放回的抽样主要有 πPS 法。系统采取了 PPS 法,样本如果重复将重新抽取。不等概率实现的方式有两种:代码法和拉希里法。系统采取了拉希里法[7-8]。

不等概率抽样,需要事先确定一个辅助变量 M_i 作为单元度量。例如,针对供电局不同业务,系统设计的辅助变量如表 1 所示。

表 1 不等概率抽样辅助变量列表

业务类型	辅助变量
业扩报装	业扩容量
抄核收	用电量
用电检查	窃电量
计量	计量容量

不等概率抽样拉希里法的算法设计。

输入:层的所有样本,度量 M_i;

输出:层的抽样样本($n_{i,h}$ 个样本);

步骤 1 对度量进行必要的转换,使其变为正数,比如乘以 10 或 100 等。

步骤 2 取出最大的度量,记为 M:

$$M = \max_{1 \leqslant i \leqslant N}\{M_i\} \tag{3}$$

其中 N 为该层的样本总量(跟前文中 N 有所区别)。

步骤 3 同时对区间 [1,M] 和 [1,N] 分别生成一个随机数,分别记为 m 和 i,如果 $M_i \geqslant m$,则第 i 个样本被抽中。

步骤 4 重复步骤 3,直到得到事先规定的抽样个数 $n_{i,h}$,停止抽样。

2.4 总体推算算法

根据样本分配及样本抽样的不同组合有不同的总体推算算法。下面简要介绍几个比较有代表性的推算算法:分层随机抽样总体均值推算算法、分层不等概率(PPS)抽样总体比率推算、多阶段抽样总体均值估计[9-10]。

2.4.1 分层随机抽样总体均值推算

总体均值:

$$\bar{Y} = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{N_h} Y_{h,i} = \sum_{h=1}^{L} W_h \bar{Y}_h \tag{4}$$

总体均值的估计量:

$$\bar{\bar{y}}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h = \frac{1}{N}\sum_{i=1}^{L} N_h \bar{y}_h \tag{5}$$

该估计量的方差为:

$$v(\bar{\bar{y}}_{st}) = \sum_{h=1}^{L}\frac{W_h^2 S_h^2}{n_h} - \frac{\sum_{h=1}^{L} W_h S_h^2}{N} \tag{6}$$

该方差的估计量:

$$v(\bar{\bar{y}}_{st}) = \sum_{h=1}^{L}\frac{W_h^2 s_h^2}{n_h}(1 - f_h) = \sum_{h=1}^{L}\frac{W_h^2 s_h^2}{n_h} - \frac{\sum_{h=1}^{L} W_h s_h^2}{N} \tag{7}$$

其中 $f_h = n_h/N_h$。

2.4.2 分层不等概率(PPS)总体比率推算

抽样,样本总个数为 n。

总体(层),样本总数为 N(此时的总体是指某个层)。

抽样样本,第 i 个单元的标志值为 y_i,$i = 1,2,\cdots,n$。

总体(层),第 i 个单元的辅助变量:M_i,($i = 1,2,\cdots,N$)。

总体(层),辅助变量的总值: $M_0 = \sum_{i=1}^{N} M_i$。

总体(层),第 i 个单元被抽中的概率 $z_i = \frac{M_i}{M_0}$, $\sum_{i=1}^{N} z_i = 1$。

抽样样本,第 i 个样本单元。令 $q_i = \frac{y_i}{Nz_i}, i = 1, 2, \cdots, n$。

总体均值的估计,就是对 q 进行估计:

$$\bar{q} = \frac{1}{n}\sum_{i=1}^{n} q_i \tag{8}$$

$\bar{q}$ 的估计方差:

$$s^2(\bar{q}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(p_i - \bar{q})^2 = \frac{1}{n(n-1)}\left(\sum_{i=1}^{n} p_i^2 - n\bar{q}^2\right) \tag{9}$$

2.4.3　多阶段抽样总体均值估计

多阶段抽样的估计分三步:

1) 估计二级单元的均值和方差;

2) 估计一级单元的均值和方差;

3) 估计总体的均值和方差。

多阶段抽样需要对各级单元做一次均值和方差的估计,可通过递推的方式获得。

3　系统实现

3.1　系统开发环境

系统开发采用 Microsoft. Net C#技术,B/S 架构,数据库选择 Oracle9。

3.2　系统运行环境

选择计算机软硬件时,既要考虑计算机运行速度、容量、操作灵活性等,又要考虑费用、新老系统的兼容性、企业的实力与实际需求[11]。系统服务器配置决定了抽样算法执行的时间,因此在 CPU 和内存方面推荐高配置以节省抽样时间。

硬件环境　服务器端最低配置至强 CPU 2.4 GHz,2 GB 内存,160 GB 硬盘; 客户端计算机 CPU Pentium 800 MHz 以上,512 MB 内存,80 GB 硬盘。

软件环境　服务器操作系统: Windows 2003 Advanced Server; 客户机端操作系统: Windows2000 / XP Internet Explorer 6.0。

3.3　系统主界面实现

图 2 为系统主界面。

图 2　系统主界面

3.4　系统样本抽样框设置

用户可以通过抽样框设置选择样本时间范围、业务部门、方差计算时间范围(为 Neyman 抽样提供方差值),并可以实时查看样本框设置时间范围内的业务量统计。

3.5　系统抽样方法设定实现

系统抽样方法根据业务类型可以选择不同的抽样方法。

3.6　系统总体推断实现

系统根据抽样成本设置、抽样框设置、抽样方法选择等参数根据第 2 章提供的推断算法,自动推断出总体得分,并提供均值推断、比例推断以及相应的置性区间。

图 3　样本任务初始化样例

图 4　样本抽样框设置样例

图 5　抽样方法设定

图 6　总体推断结果样例

4　结语

复杂抽样推断系统可以与用户业务系统及业务数据库无缝链接,并根据用户业务需要设定复杂抽样算法实现统计抽样推断的自动化,解决通用统计软件抽样模块无法解决的问题。在抽样框设计方面,用户可以结合实际业务选择需要调查的时间范围、业务类型及调查部门。在抽样算法方面,囊括了目前主流的随机抽样、分层抽样、Neyman 分层抽样、不等概率抽样、多阶段抽样等方法,并可以根据业务特性选择适合的抽样方法。在总体推断方面,根据抽样方法实现自动化,降低手动操作比例,提升效率。在软件构架设计方面,基于 B/S 的构架大大提升了操作人员的工作的便捷性,无论是抽样阶段还是调查阶段的数据录入都可以随时随地进行,甚至在现场调查时还可以使用 PDA 之类的手持终端进行数据录入与查询。基于 B/S 的复杂抽样推断系统使用户的抽样调查活动周期化、非专业人员操作自动化成为可能,很大程度上提升了抽样效率,降低了调查成本。　(下转第 206 页)

系统能够不依赖用户使用的图形设备实现图像显示。因此，需要快速构建可视化过程时可以采取直接图形系统，如果需要对图像做进一步操作处理，采用对象图形系统。

3 应用实例

基于 IDL 开发的海洋环境微波遥感数据处理系统可以处理微波辐射计、高度计、散射计与 SAR 等数据，实现海洋风场、海表温度、海面高度、波高等海洋环境参数反演与处理、数据格式转换与部分信息的可视化等功能。海洋环境微波遥感数据处理系统主界面如图 2 所示。在基于 SAR 影像的海洋风场反演子系统中，采用了直接图形系统与对象图形系统两种方法，如 SAR 影像显示部分采用的是对象图形法，界面如图 3 所示。基于 SAR 影像反演出的海洋风场显示部分采用的是直接图形法，界面如图 4 所示。

图 2 海洋环境微波遥感数据处理系统 V1.0

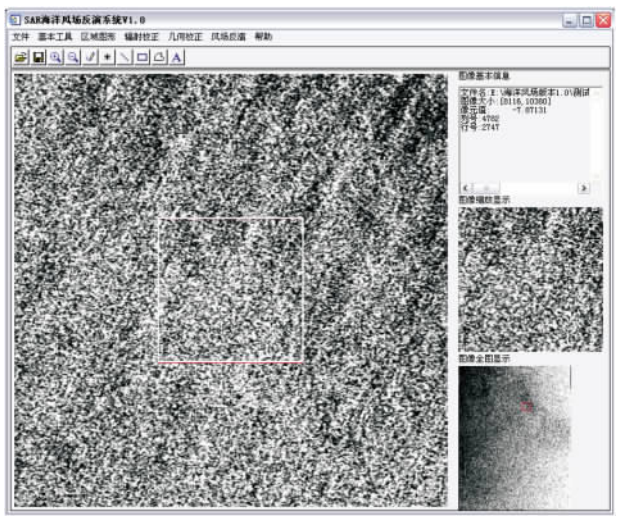

图 3 SAR 影像显示

4 结语

IDL 语言具有面向数组，支持科学数据格式的读写和高级的图像处理能力，使得应用程序的开发更加方便快捷。科研人员可以借助 IDL 语言快速实现研究的算法与处理数据；应用软件开发人员可以选择基于 IDL 开发平台开发出相应的应用软件。基于 IDL 开发的海洋环境微波遥感数据处理系统可以作为独立运行的系统发布，也可以作为子系统集成到其他系统中去，同时留有接口，后继的研究成果可以快速集成到系统中，满足不断扩展的要求。

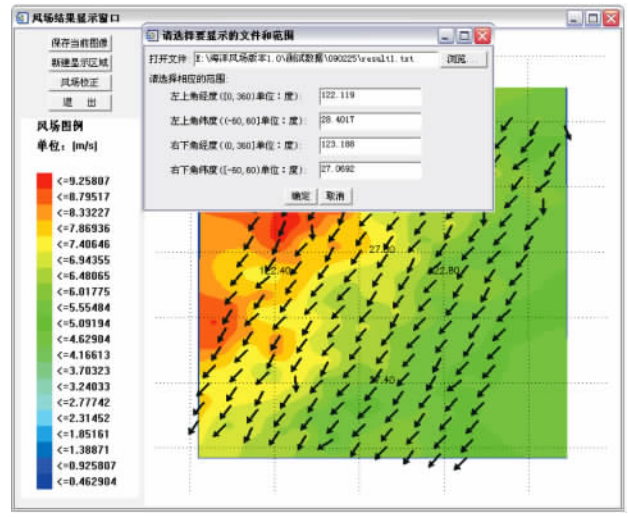

图 4 海洋风场显示

参考文献:

[1] 邵芸，张风丽，田维，等. 海洋环境微波遥感应用研究进展[J]. 遥感学报，2009，13(增刊)：154－159.

[2] 张杰，黄卫民，纪永刚，等. 中国海洋微波遥感研究进展[J]. 海洋科学进展，2004，22(增刊)：157－165.

[3] IDL6.3_6.4 白皮书[EB/OL]. [2010－04－15]. http://www.imagetek.info.com/download/IDL/book/IDL6.3_6.4 白皮书.pdf.

[4] 朱玉伟，拾兵，黄勇，等. 基于 IDL 与 VB 的遥感数据提取[J]. 计算机应用研究，2006，23(3)：181－182.

[5] 路文海. IDL 语言处理 HDF 格式遥感数据的研究[J]. 海洋信息，2006(3)：6－7.

[6] 阚瑗珂，朱利东，张瑞军，等. 基于 IDL 和. NET 的可视化程序设计[J]. 计算机应用研究，2007，24(9)：149－151.

[7] 郭建文，冯敏，尚庆生，等. IDL 在分布式 GIS 系统中的应用研究[J]. 计算机应用研究，2007，24(5)：220－222.

[8] 汤泉，牛铮. 基于 IDL 与 ENVI 二次开发的遥感系统开发方法[J]. 计算机应用，2008，28(S1)：270－272.

[9] 王姓，江南，胡斌. 基于 IDL 语言的太湖蓝藻水华遥感监测信息系统设计[J]. 遥感应用，2010(2)：59－64.

[10] 韩培友. IDL 可视化分析与应用[M]. 西安：西北工业大学出版社，2006.

[11] 闫殿武. IDL 可视化工具入门与提高[M]. 北京：机械工业出版社，2003.

(上接第 187 页)

参考文献:

[1] 于曦. 基于 C/S 模式的旅游抽样调查系统的设计与实现[D]. 成都：电子科技大学，2006.

[2] 罗智超，管河山，曹礼华. 电力营业稽查的分类预测方法研究[J]. 统计与决策，2010(10)：60－62.

[3] 罗智超，吴育青. 基于电力细分业务抽样调查的客户满意度模型[J]. 广东电力，2010，23(11)：64－66，101.

[4] 吕国英. 抽样调查管理系统开发平台的设计[J]. 山西大学学报，2004，27(4)：49－51.

[5] 李金昌. 应用抽样技术[M]. 北京：科学出版社，2007.

[6] 刘卫江，白磊，景泉. 改进型分层抽样技术及性能研究[J]. 计算机工程与应用，2007，43(8)：114－117.

[7] 李金昌. 不等概率抽样若干问题探讨[J]. 统计研究，1999(S1)：146－148.

[8] 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，2004，27(4)：385－388.

[9] 刘建华，金水高. 复杂抽样调查总体特征量及其方差的估计[J]. 中国卫生统计，2008，25(4)：377－379.

[10] 陈丹萍，赵耐青，林燧恒. 分层整群随机抽样数据的不同分析方法及结果比较[J]. 中国卫生统计，2010，27(2)：122－124.

[11] 陈锦玲. 基于 C/S 架构的电话营销访问系统分析与实现[J]. 计算机应用，2010，30(S1)：317－320.