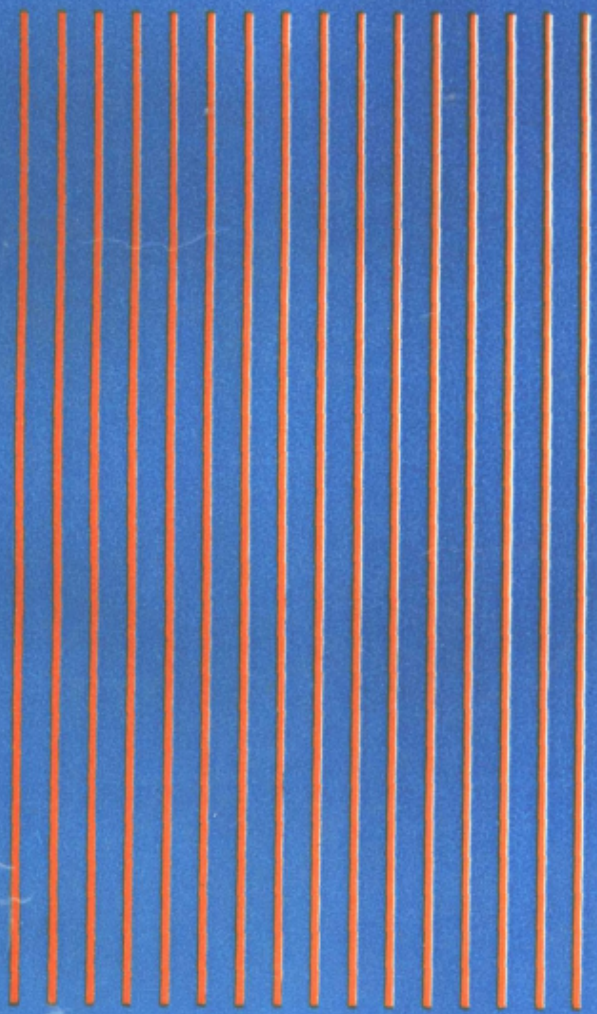
应用统计实例选

江泽培　成　平
严士健　吴　荣　主编



南开大学出版社

责任编辑：裴志明
封面设计：丁沙铃

ISBN 7-310-00956-8
O·98 定价：32.00元

应用统计实例选

江泽培　成平
严士健　吴荣　主编

南开大学出版社

内容提要

本书是由南开数学所组织编选的应用统计的论文集，精选了反映近年来我国高等院校、科研单位、实际应用部门的统计应用成果的论文13篇。

这些论文涉及到社会经济、人口、工业产品和质量控制、工程、林业、医学、天文、地质等十多个方面，引用了数理统计的各种方法和技术，如多元分析方法、抽样调查技术、时间序列分析方法、可靠性统计、试验设计、回归分析和估计理论等等。这些专题都是近年来我国实际部门所提出的亟需研究的课题。每篇专题论文各有一个主题，作者从问题的提出到问题的解决，从收集数据到建立模型，以及应用统计方法与技术进行成功的处理和分析，直到获得有效的结果，都做了系统扼要的阐述。

本书可作为高等院校有关专业教材，也可作为从事实际应用的统计工作者的参考书。

应用统计实例选

江泽培　成平
严士健　吴荣　主编

南开大学出版社出版
（天津八里台南开大学校内）
邮编：300071　电话：23508542
新华书店天津发行所发行
天津宝坻第二印刷厂印刷

1997年2月第1版　1997年2月第1次印刷
开本：787×1092 1/16　印张：14.5　插页：2
字数：380千　印数：1—1000

ISBN 7-310-00956-8
O·98　定价：32.00元

前　言

南开数学所于1988年9月至1989年6月举办了概率学术年。在学术年期间，召开了“应用概率统计研讨会”，有20多位国内学者应邀做了应用研究的报告。会后成立了编辑小组，约请专家撰稿，组织编选了这本《应用统计实例选》。

本书收集了13篇应用统计的专题论文，它们涉及到社会经济、人口、工业产品和质量控制、工程、林业、医学、天文、地质等十多个方面，引用了数理统计的各种方法和技术，如多元分析方法、抽样调查技术、时间序列分析方法、可靠性统计、试验设计、回归分析和估计理论等等。这些专题都是近年来我国实际部门所提出的亟需研究的课题。每篇专题论文各有一个主题，作者从问题的提出到问题的解决，从收集数据到建立模型，以及应用统计方法与技术进行成功的处理和分析，直到获得有效的结果，都做了系统扼要的阐述。

这是我国出版的第一本《应用统计实例选》。当前，高等院校与研究单位为培养应用统计的人才，亟需设置“统计咨询”课程，本书可选作教学参考书。它还可以作为从事实际应用的统计工作者的参考书。希望本书的出版有助于提高对应用统计的方法与成果的鉴赏水平，对我国应用统计工作的发展起积极推动作用。

张尧庭、谢衷洁、杨振海实际担负了全书的编选工作，张润楚为本书的编辑出版做了大量的工作，曹建毅为本书绘制了大量的墨图，谨向他们表示衷心的感谢。

南开数学所概率统计学术年组织委员会

一九九六年十二月

目 录

全国专业技术人员现状抽样调查的设计与分析①

冯士雍　董莉明
（中国科学院系统科学所）

1 概述

1987年11月至12月，由原国家科委科技干部局主持，在全国范围内进行了一次大规模的专业技术人员现状抽样调查。此次调查的主要目的是了解我国专业技术人员队伍的分布结构、广大专业技术人员的工作环境、生活条件和健康情况、发挥作用情况以及了解他们对科技人员管理体制改革的要求和愿望，从而为国家制定有关的方针政策提供切实可靠的依据，以建立一个保证广大专业技术人员充分发挥作用的良好社会与工作环境，更好地为经济建设服务。

调查采用对专业技术人员个人的问卷调查形式。问卷除了编号以及有关该专业技术人员所属单位的性质、行业分类及隶属关系等背景材料外，对专业技术人员个人情况及意愿的问题分以下四个部分：

1)基本情况。包括性别、年龄、民族、政治面貌、工龄、学历、学位、专业职称及职务等13个问题；

2)环境条件情况。包括家庭人口、经济收入、居住条件、本人健康状况、出国留学进修、脱产学习、从事科技或新技术开发活动的经费、仪器设备图书资料与助手配备情况，单位领导用人及对工作支持程度等23个问题；

3)发挥作用情况。包括承担课题及所起作用、工作量饱满程度、单位用人管理、本人担任领导职务以及对技术工作感到忧虑与烦恼等16个问题；

4)科技人员管理体制改革等情况。包括所在单位实行专业技术职务聘任制及职称评定情况、本人兼职、工作调动及专业变动情况以及对这些问题的态度和对社会保险、户籍制度与住房制度改革的意见等27个问题。

全部共计79个问题，约460个选择项。

根据全国各省、市、自治区及中央各部门的统计，至1986年底全国专业技术人员(不包含中小学教师)共有1003万人。我们没有可能也没有必要对每个专业技术人员进行调查。因此调查采用抽样调查方法。受主持单位委托，我们承担了此项目的总体抽样设计并给出相应的数据处理公式。根据抽样方案，调查采用四阶抽样方法，即全国抽省(直辖市、自治区)，在抽中的样本省(市、自治区)中抽市(地区或州)，在样本市(地区或州)中抽基层单位以及在抽中的基层单位中抽个人的四阶抽样。考虑到调查的组织实施的便利以及尽可能地提高调查的精度，我们

① 国家自然科学基金7860013部分资助项目。

在前三阶抽样中反复采用了分层技术，同时在第二阶与第三阶抽样中采用效率较高的不等概率抽样。

鉴于上海、天津两市已在这以前进行了类似的调查，但调查内容与本次调查内容不一致，因此直辖市一层由北京市作代表。考虑到直辖市情况的特殊性，北京市的调查无论从抽样设计到问卷都是独立进行的，但在设计中又考虑到全国汇总的需要，因而是与全国方案相协调的。关于北京地区专业技术人员现状调查的详细情况见[1]。

调查规模即样本量，既受时间与经费的限制，又直接影响调查的精度。在设计时，我们要求对于总体比例型目标量，对一般样本省（自治区）在置信度为95%下最大绝对误差不超过2%，对较小省（自治区）不超过3%。由此确定在第一阶抽样中抽取的辽宁、江苏、广东、江西、湖北、四川、青海、新疆、贵州等九省（自治区）中抽取29000人，加上北京地区的5000人共34000人进行调查。总的抽样比为3.39‰。问卷回收率达到100%，有效率也达99%以上。

所有有效问卷在录入计算机后，按本文给出的方法进行数据处理。主要结果见[2]，其中若干重要内容已摘要公布于《人民日报》(1988年5月25日头版)。

下面我们将详细介绍此项调查的抽样方案及样本抽取情况；总体目标量的估计及其方差估计公式；调查结果的精度分析以及调查所得的主要结论。

2　抽样方案及样本抽取情况

2.1　基本考虑

对于一项全国性大规模的抽样调查，抽样方案的设计必须从几个方面综合考虑。首先是调查的组织与实施形式，此次调查由原国家科委科技干部局主持，但具体调查实施必须通过各省市自治区科干局来组织。另外中央各部门也集中了相当数量的专业技术人员，但其分布与构成对于不同部门相差很大。在三个直辖市，特别是北京市尤为集中，而在一般省、自治区则较为分散。因此我们决定首先按省、市、自治区抽样而不按部门抽样。中央各部门所属的专业技术人员按单位所在地区参加统一抽样。当然在抽样过程中，只要有可能，中央及地方所属单位可作不同层处理（例如北京地区的抽样即是首先分为中央所属单位及市属单位两大层，详见[1]）。由于省（直辖市、自治区）仍是一个很大的范围，因此为了使样本相对集中，同时又为使样本具有良好的代表性，故采用省内抽市（地区、自治州），市（地，州）内抽基层单位，在基层单位内抽人的多阶抽样。为了提高调查的精度，考虑到不同地域社会经济发展程度以及单位行业性质之间的差异，故在前几阶抽样中都采用了分层技术。为了提高效率，减小抽样误差，又在各层内抽样采用与地区或单位规模（专业技术人员数）成比例的不等概率抽样。为此在各阶抽样中都准备了完整而可靠的抽样框资料。在确定样本量时，则是结合经费、时间以及精度多方面的考虑来定。在精度方面，既保证了全国汇总的要求，又在一定程度上满足调查省、市、自治区汇总的需要。总之，抽样方案是在综合了方法的科学性与实际的可行性两个方面而制定的。

2.2　第一阶抽样

第一阶抽样是在全国抽省、直辖市或自治区。对于直辖市，正如前已所述，已确定调查北京市。全国当时的其余26个省（自治区）按国民经济生产总值及专业技术人员力量，划分为以下

三类，以每类作为层：

1)沿海地区：包括辽宁、河北、山东、江苏、浙江、福建、广东(含海南)共七省；

2)内地地区：包括黑龙江、吉林、山西、河南、安徽、江西、湖北、湖南、陕西及四川十省；

3)边远地区：包括内蒙古、甘肃、宁夏、青海、新疆、广西、贵州、云南及西藏共九省区；

根据需要与可能，特别是各省实施调查的条件，兼顾各方面的代表性，在上述三类地区中每类确定三个省(自治区)进行调查。进行调查的九个省(自治区)是：

1)沿海地区：辽宁、江苏、广东；

2)内地地区：江西、四川、湖北；

3)边远地区：青海、新疆、贵州。

2.3 第二阶抽样

第二阶抽样是在第一阶抽样中抽取的九省(自治区)中按分层不等概率抽样方法抽取市(地或州)。具体方法是：

1)省会及计划单列市为必抽市，即每个都为自我代表层；

2)将省(自治区)内其他市(地或州)按经济发达的程度分为若干类。一般分为三类，即Ⅰ类较发达地区，Ⅱ类一般地区及Ⅲ类较不发达地区。地区分类也可按各省(自治区)原有自然分类为准。例如四川是按其地域经济区划分的。

将上述分类中的每类都作为层处理。层内若只有3个或3个以下市(地、州)，则按与专业技术人员数成比例的概率抽样(PPS抽样)方法抽取一个市(地、州)；若层内有4个或4个以上的市(地区或州)，则按不放回的与专业技术人员数成比例的概率抽样(πPS抽样)抽取2个市(地、州)。抽样结果见表3。下面举例说明上述两种抽样方法实施的具体步骤。

1)PPS抽样。若层内只包含3个或3个以下的市，则按与专业技术人员数M_i成正比的概率抽取一个样本市。设层内专业技术人员总数为M_0，则第i个市(地、州)被抽到的概率为$z_i = M_i/M_0$。下面以湖北省Ⅱ类地区抽取样本市的过程来说明PPS抽样的实施步骤：

1°将该地区所包含的黄岗、郧阳与鄂西三地(州)以及专业技术数M_i列成表1，并计算累计专业技术人员数。层内专业技术人员总数$M_0 = 45860$。

表1 湖北省Ⅱ类地区PPS抽样过程

序号i	市(地、州)名称	专业技术人员数M_i	累计专业技术人员数	随机数
1	黄　岗	19271	19271	
2	郧　阳	11695	30930	
3	鄂　西	14930	45860	37225

2°用随机数发生器产生1—45860之间的(离散)均匀分布的一个随机数。此处为37225，它介于30930与45860之间，故与之相应的第3个单元即鄂西自治州被抽为样本市(地、州)。

2)πPS抽样。当层内包含4个或4个以上的市，则按不放回的与专业技术人员数M_i成比例的概率抽取2个样本市。具体地说，我们采用Brewer方法[3]。两个样本市的抽取方法如下：

设层内的专业技术人员总数为$M_0 = \sum_{i=1}^{N} M_i$，其中N为该层内的市(地、州)数，令

$$z_i = M_i/M_0, i = 1,2,\cdots,N \tag{1}$$

则第一个样本市以正比于

$$\frac{z_i(1-z_i)}{1-2z_i}$$

的概率抽取，亦即以下述概率抽取：

$$\frac{2z_i(1-z_i)}{(1-2z_i)\left(1+\sum_{j=1}^{N}\frac{z_j}{1-2z_j}\right)} \tag{2}$$

设第 i 市被抽中，则在剩下的 $N-1$ 个市中，以概率

$$\frac{z_j}{1-z_i}, j \neq i \tag{3}$$

抽取第 2 个样本市。可以证明按上述方法，第 i 市入样(两次抽样中有一次被抽中)概率

$$\pi_i = 2z_i, i = 1,2,\cdots,N \tag{4}$$

从而严格地与规模大小 M_i 成正比。

下面我们以湖北省 I 类地区为例说明上述 Brewer 方法的具体实施过程。

该地区共包含 5 个市，$M_i, z_i, y_i = z_i(1-z_i)/(1-2z_i)$ 列于表 2 中。

表 2　湖北省 I 类地区 πPS 抽样过程

序号	市	M_i	z_i	$y_i=\frac{z_i(1-z_i)}{1-2z_i}$	累计 y_i	随机数 r_1	再累计 z_i	随机数 r_2
1	黄石	14780	0.4217	1.5573	1.5573	0.5450		
2	沙门	6992	0.1995	0.2657	1.8230		0.1995	
3	荆门	3250	0.0927	0.1032	1.9262		0.2922	
4	十堰	3570	0.1018	0.1148	2.0410		0.3940	0.3202
5	宜昌	6460	0.1843	0.2381	2.2791		0.5783	
合计		35052						

y 的总和为 2.2791，用随机数发生器产生一个〔0,2.2791〕区间上均匀分布的随机数 r_1，此处为 0.5450。由于 $0 < r_1 < 1.5573$，故与之对应的黄石市被抽为第 1 个样本市。第 2 个样本市则是在剩下的 4 个市中按与 M_i(亦即与 z_i) 成正比的概率抽取，故对这 4 个市累计 z_i，总和为 0.5783，产生一个〔0,0.5783〕上均匀分布的随机数字 r_2，由于 $z_3 < r_2 < z_4$，故相应的十堰市被抽中为第 2 个样本市。

2.4　样本量的确定与分配

样本量取决于对调查目标量的估计精度要求以及所用抽样方案的设计效应。要求的精度高(估计量误差小)，则所需的样本量就大。设计效应(deff)则是确定抽样方法所得的估计量的方差与相同样本量时简单随机抽样同一估计量方差之比。在此项调查中，由于采用问卷形式，故绝大多数目标量以比例形式出现，例如对一个问题给出某种特定回答的人的比例。对于简单随机抽样，在置信水平 $1-\alpha$ 下，为使比例 P 的估计的绝对误差不超过给定值 d 时，所需的样本量 n 有如下公式(若抽样比 f 很小，从而可不考虑有限总体修正系数 $1-f$)：

$$n = U_\alpha^2 P(1-P)/d^2, \tag{5}$$

其中U_α是标准正态分布的双侧分位数。我们取$\alpha = 5\%$，即置信水平为95%。此时$U_\alpha = 1.96$，对较大的样本省，要求$d = 2\%$，对较小的样本省，要求$d = 3\%$。代入(5)式，计算得n分别为2401与1667。对设计效应我们估计为1.8，这样计算样本量分别需要4322及1921。于是我们确定每个样本省实际样本量为2000—4500。具体数字按各省抽取的样本市(地、州)的多少而定。为简单起见，我们定一般样本市调查500人，计划单列市及某些省会城市调查1000人。各省(自治区)调查人数也见表3。

表3 各样本省(自治区)抽取的样本市(地、州)及其样本量

省(自治区)	Ⅰ类地区	Ⅱ类地区	Ⅲ地区	样本市(地、州)数	样本量
江苏	常州	南京*，扬州	盐城，连云港	5	3000
青海	西宁，海东	海北州	果洛州	4	2000
江西	南昌，景德镇	萍乡，宜春	上饶，赣州	6	3000
贵州	贵阳，遵义	黔南	毕节，黔东南	5	2500
湖北	武汉*，黄石，十堰	襄樊，孝感	鄂西	6	3500
广东	广州*，佛山	江门	茂名，韶关	5	3000
辽宁	沈阳*，大连*，鞍山	丹东，辽阳	盘锦，朝阳	7	4500
新疆	乌鲁木齐，石河子	阿克苏，喀什	博尔塔拉	5	3000
四川	重庆*，万县(川东)	成都*，乐山(川中)	内江(川南)	甘孜(川西北)，攀枝花(川西南)	7	4500
总计				50	29000

注1 凡打*号的城市调查1000人，其它市(地、州)调查500人。

注2 北京地区根据[1]共调查5000人，故全国总计调查34000人。

2.5 第三阶抽样

第三阶抽样是在每个被抽中的样本市(地、州)中对基层单位的抽样。大城市每个抽100个单位，其它市(地、州)每个抽50个单位，抽样方法是分层不等概率系统抽样。

首先将样本市(地、州)按行政管理系统或其它任何比较方便的形式分层(包括市属县)，可以适当地进行归并或分析。例如我们可以将某市按中央各部门所属单位、省直单位(必要的话每个还可细分)、计委、建委、经委、文教卫生、司法、党政群、人民团体、所属区县级单位等系统分层。按每层中的专业技术人员数的比例分配各层应抽的基层单位数。

在层内，将所有基层单位按自然顺序排列，根据每个单位的专业技术人员数进行不等概率系统抽样抽取基层单位。

2.6 第四阶抽样

第四阶抽样，亦即最后一阶抽样是在每个被抽中的基层单位中，按等概率系统抽样(即等距抽样)或简单随机抽样抽取10人作实际问卷调查。注意，必须在符合这次调查范围的专业技

术人员中抽取。

3　总体目标量的估计及其方差估计

3.1　总体目标量的分类

根据问卷设计，专业技术人员现状调查的所有指标，即总体目标量可以分成以下四类：

1）总体总量 Y。即对某个指标 y，全国或分省、市（加上适当的下标）的总和。例如专业技术人员中想调动工作的人数以及承担国家七五攻关、863 高技术及自然科学基金项目的专业技术人员总数等。

2）总体平均数 $\bar{Y}$。总体总值按所统计范围内专业技术人员数的平均数。例如月平均工资、平均每户的居住面积等。

3）总体比例 P。按某种方式分类（例如性别、年龄组以及对某个问题持特定回答的人）在全体人员中所占的比例。例如女性比例，受聘专业技术职务后工资提高了三级以上者所占的比例等。易见它是相应指标值 y 只取 0、1 值的总体平均数。

由于 Y 与 $\bar{Y}$ 只相差一个已知常数，故上述三类目标量在数据处理时，实质上是相同的。

4）总体两个目标量总量或平均数之比 R，即 $R=Y/X=\bar{Y}/\bar{X}$，其中 Y,X 都需要从样本中估计。对 R 的处理与对 P 的相应公式有很大的不同。

3.2　记号

为表达方便起见，将下文中常用的记号作如下的规定：

下标 h 表示省内层的编号；i 表示层内市（地、州）的编号；j 表示市（地、州）内系统（小层）的编号；k 表示基层单位的编号；l 表示同一基层单位内被调查的专业技术人员的编号。

$y,x,\cdots$ 表示不同的指标量，加上有关人员的编号即表示该人员的指标值。

Y（或 X）加上适当的下标表示指标 y（或 x）在一个省内对指定下标范围内的（总体）总和。例如以

$$Y_{hij}=\sum_k\sum_l y_{hijkl}$$

表示某省 h 层 i 市 j 系统内所有专业技术人员指标 y 的总和；以

$$Y_{hi}=\sum_j Y_{hij}=\sum_j\sum_k\sum_l y_{hijkl}$$

表示某省 h 层 i 市所有专业技术人员指标 y 的总和；以

$$Y_h=\sum_i Y_{hi},Y=\sum_h Y_h$$

分别表示省内 h 层以及该省所有专业技术 y 的总和；以

$$\bar{Y}_{hij},\bar{Y}_{hi},\bar{Y}_h,\bar{Y}$$

分别表示指标 y 在相应编号范围内所有技术人员指标 y 的平均数。

对另一个指标 X，X_{hij},X_{hi},X_h 及 X 以及 $\bar{X}_{hij},\bar{X}_{hi},\bar{X}_h,\bar{X}$ 的意义与 y 的相应量意义相同。

上述所有有关总体的目标量上添加“^”号表示相应目标量的估计量。例如 $\hat{Y}_{hi}$ 是 Y_{hi} 的估计量；

$V(\cdot),S(\cdot),c(\cdot)$ 分别表示估计量的方差，标准差及变异系数的估计；

M 加上适当的下标表示在抽样时所依据的专业技术人员数(1986 年底)；N 加上相应的下标表示在数据处理时采用的专业技术人员数(1987 年底)；

为了公式表达的简洁起见，在不会发生混淆的情形，层的编号 h 常被省略。

3.3 样本市(地、州)目标量的估计及其方差估计

根据抽样方案，对省内某层第 i 样本市(地、州)内的抽样即第三阶抽样是在市内抽取基层单位。方法是先按行政管理系统(或其它形式划分的，通称系统)分层，在系统层内按单位中的专业技术人员数成比例的不等概率系统抽样抽取基层单位。而第四阶抽样即是在每个被抽中的基层单位内按等概率的系统抽样(即等距抽样)或简单随机抽样抽取相同数量 ($m = 10$) 的人员。由于系统层内按单位的实际抽样比很小，不放回抽样与放回抽样差别不大(但效率前者稍高，即实际抽样方差前者比后者稍小)，而单位在系统层内的排列顺序以及单位内人员的排列顺序都可看作与调查指标量无关。因此为简化起见，我们用放回的按与单位大小成比例的概率抽取基层单位，再在每个被抽中的基层单位内，用简单随机抽样抽取相同样本量的二阶抽样来作为在一个样本市系统层内上述实际抽样中第三、四两阶抽样的近似。而按照这种简化的抽样模型，所得的样本是自加权的。

设系统层内抽取了 n 个基层单位，则该系统层内指标 y 的总体平均数 $\overline{Y}_{ij}$ 的估计值应为[3]：

$$\hat{\overline{Y}}_{ij} = \frac{1}{n}\sum_{k=1}^{n}\overline{y}_{ijk} = \frac{1}{n}\sum_{k=1}^{n}\frac{1}{m}\sum_{l=1}^{m}y_{ijkl} \tag{6}$$

而相应的总量 Y_{ij} 的估计为

$$\hat{Y}_{ij} = N_{ij}\hat{\overline{Y}}_{ij} \tag{7}$$

$\hat{\overline{Y}}_{ij}$ 及 $\hat{Y}_{ij}$ 都是无偏的。而 $\hat{\overline{Y}}_{ij}$ 及 $\hat{Y}_{ij}$ 的方差估计分别为

$$V(\hat{\overline{Y}}_{ij}) = \frac{1}{n(n-1)}\sum_{k=1}^{n}(\overline{y}_{ijk} - \overline{Y}_{ij})^2 \tag{8}$$

$$V(\hat{Y}_{ij}) = N_{ij}^2 V(\hat{\overline{Y}}_{ij}) = \frac{N_{ij}^2}{n(n-1)}\sum_{k=1}^{n}(\overline{y}_{ijk} - \overline{Y}_{ij})^2 \tag{9}$$

这里由于对单位的抽样比很小，故有限总体修正系数忽略不计。

对于另一个指标 x，按照上述完全相同的步骤，可计算 $\hat{\overline{X}}_{ij},\hat{X}_{ij},V(\hat{\overline{X}}_{ij})$ 及 $V(\hat{X}_{ij})$。对于比值型的目标量 R_{ij}：

$$R_{ij} = \frac{Y_{ij}}{X_{ij}} = \frac{\overline{Y}_{ij}}{\overline{X}_{ij}} \tag{10}$$

我们用以下的估计公式

$$\hat{R}_{ij} = \frac{\hat{Y}_{ij}}{\hat{X}_{ij}} \tag{11}$$

为推导 $\hat{R}_{ij}$ 的方差估计公式，我们对(11)式进行 Taylor 级数展开，取其线性项，求其方差，得

$$V(\hat{R}_{ij}) \approx \left(\frac{\partial\hat{R}_{ij}}{\partial\overline{Y}_{ij}}\right)^2 V(\hat{Y}_{ij}) + \left(\frac{\partial\hat{R}_{ij}}{\partial\overline{X}_{ij}}\right)^2 V(\hat{X}_{ij})$$

$$+ 2\cdot\frac{\partial\hat{R}_{ij}}{\partial\overline{Y}_{ij}}\cdot\frac{\partial\hat{R}_{ij}}{\partial\overline{X}_{ij}}\mathrm{Cov}(\hat{Y}_{ij},\hat{X}_{ij})$$

$$= \hat{R}_{ij}^2\left(\frac{V(\hat{Y}_{ij})}{\hat{Y}_{ij}^2} + \frac{V(\hat{Y}_{ij})}{\hat{X}_{ij}^2} - 2\frac{\mathrm{Cov}(\hat{Y}_{ij},\hat{X}_{ij})}{\hat{Y}_{ij}\hat{X}_{ij}}\right) \tag{12}$$

上式中的 $V(\hat{Y}_{ij})$,$V(\hat{X}_{ij})$ 可用(9)式获得估计。为估计 $\mathrm{Cov}(\hat{Y}_{ij},\hat{X}_{ij})$,令

$$u_{ijkl} = y_{ijkl} + x_{ijkl} \tag{13}$$

则

$$\hat{U}_{ij} = \hat{Y}_{ij} + \hat{X}_{ij}$$

$$V(\hat{U}_{ij}) = V(\hat{Y}_{ij}) + V(\hat{X}_{ij}) + 2\mathrm{Cov}(\hat{Y}_{ij},\hat{X}_{ij})$$

所以

$$\mathrm{Cov}(\hat{Y}_{ij},\hat{X}_{ij}) = \frac{1}{2}(V(\hat{V}_{ij}) - V(\hat{Y}_{ij}) - V(\hat{X}_{ij}))$$

故 $\mathrm{Cov}(\hat{Y}_{ij},\hat{X}_{ij})$ 可按下式估计:

$$\mathrm{Cov}(\hat{Y}_{ij},\hat{X}_{ij}) = \frac{1}{2}(V(\hat{U}_{ij}) - V(\hat{Y}_{ij}) - V(\hat{X}_{ij}))$$

于是 $V(\hat{R}_{ij})$ 的估计可采用以下形式:

$$V(\hat{R}_{ij}) = \hat{R}_{ij}^2\left(\frac{V(\hat{Y}_{ij})}{\hat{Y}_{ij}^2} + \frac{V(\hat{X}_{ij})}{\hat{X}_{ij}^2} + \frac{V(\hat{Y}_{ij}) + V(\hat{X}_{ij}) - V(\hat{U}_{ij})}{\hat{Y}_{ij}\hat{X}_{ij}}\right) \tag{14}$$

其中 $V(\hat{U}_{ij})$ 是对指标 U 按(9)式计算的。

对系统层内的目标量进行估计并获得相应的方差估计后,样本市(地、州)目标量的估计及其方差估计可按分层抽样公式得到。令

$$W_{ij} = \frac{N_{ij}}{N_i} \tag{15}$$

为层权,则

$$\hat{\bar{Y}}_i = \sum_j W_{ij}\hat{\bar{Y}}_{ij} \tag{16}$$

$$\hat{Y}_i = \sum_j \hat{Y}_{ij} \tag{17}$$

$$\hat{R}_i = \sum_j W_{ij}\hat{R}_{ij} \tag{18}$$

相应的方差估计分别为

$$V(\hat{\bar{Y}}_i) = \sum_j W_{ij}^2 V(\hat{\bar{Y}}_{ij}) \tag{19}$$

$$V(\hat{Y}_i) = N_i^2 V(\hat{\bar{Y}}_i) \tag{20}$$

$$V(\hat{R}_i) = \sum_j W_{ij}^2 V(\hat{R}_{ij}) \tag{21}$$

3.4 样本省目标量的估计及其方差估计

根据抽样方法,样本省中的计划单列市及省会是必抽城市,因此它们中的每一个都单独成层(自我代表层)。其余的市(地、州)一般分为三层(四川分为五层),在每层中按 PPS 抽样抽一个市,或按 Brewer 方法抽两个市。下面我们分别就这三种情况进行讨论。

1)若 i 市是计划单列市或省会,此时该市按 3.3 有关的估计公式也就是所在层的估计公式。即

$$\hat{\bar{Y}}_h = \hat{\bar{Y}}_{hi}, \hat{Y}_h = \hat{Y}_{hi}, \hat{R}_h = \hat{R}_{hi} \tag{22}$$

$$V(\hat{\bar{Y}}_h) = V(\hat{\bar{Y}}_{hi}), V(\hat{Y}_h) = V(\hat{Y}_{hi}), V(\hat{R}_h) = V(\hat{R}_{hi}) \tag{23}$$

2)若 i 市是从层内按 PPS 抽样抽出的唯一样本市，则根据 Hansan-Hurwitz 公式，层总量 Y_h 与 X_h 的估计为：

$$\hat{Y}_h = \frac{M_h}{M_{hi}}\hat{Y}_{hi}, \hat{X}_h = \frac{M_h}{M_{hi}}\hat{X}_{hi} \tag{24}$$

其中 M_{hi} 及 M_h 是抽样时(1986 年底) 第 i 市及该层的专业技术人员数。而 $\bar{Y}_h$ 及 $\bar{X}_h$ 的估计为

$$\hat{\bar{Y}}_h = \frac{\hat{Y}_h}{N_h}, \hat{\bar{X}}_h = \frac{\hat{X}_h}{N_h} \tag{25}$$

这里的 N_h 是该层 1987 年底的专业技术人员数。

对比值型目标量 R_h，采用以下估计公式

$$\hat{R}_h = \frac{\hat{Y}_h}{\hat{X}_h} \tag{26}$$

上述 $Y_h(X_h)$ 及 $\bar{Y}_h(\bar{X}_h)$ 的估计都是无偏的。为推导方差估计公式，根据二阶抽样方差估计，我们有

$$\begin{aligned} V(\hat{\bar{Y}}_h) &= V_1(E_2(\hat{\bar{Y}}_{hi})) + E_1(V_2(\hat{\bar{Y}}_{hi})) \\ &= V_1(\hat{\bar{Y}}_h) + \sum_i \frac{M_{hi}}{M_h} V_2(\hat{\bar{Y}}_{hi}) \\ &= \sum_i \frac{M_{hi}}{M_h}(\bar{Y}_{hi} - \bar{Y}_h)^2 + \sum_i \frac{M_{hi}}{M_h} V_2(\hat{\bar{Y}}_{hi}) \end{aligned} \tag{27}$$

式中求和是对层内所有(实际至多只有三个)市(地、州)求的，E_1, V_1 是对市(地、州) 这一阶抽样求期望与方差，而 E_2, V_2 是对市内进一步抽样求期望与方差。由于我们在层内实际只抽取一个样本市，故严格地说(27) 式前一项不能单从这个唯一样本市的数据获得估计。根据抽样方案，省内的市(地、州) 是按其经济发达的程度分类为层的。因此层内差别，亦即 $\bar{Y}_{hi}$ 的差别一般不是很大，因此我们将(27) 式中前一项分量忽略不计。①而第二项分量可用下式估计：

$$\left(\frac{M_h}{M_{hi}}\right)^2 V(\hat{\bar{Y}}_{hi}) \tag{28}$$

其中 $V(\hat{\bar{Y}}_{hi})$ 即 3.3 中关于样本市的方差估计，于是我们有

$$V(\hat{\bar{Y}}_h) = \left(\frac{M_h}{M_{hi}}\right)^2 V(\hat{\bar{Y}}_{hi}) \tag{29}$$

$$V(\hat{Y}_h) = N_h^2 V(\hat{\bar{Y}}_h) \tag{30}$$

对于比值型目标量 R_h 的估计 $\hat{R}_h$ 的方差估计 $V(\hat{R}_h)$ 的推导采用与 $V(\hat{R}_{hi})$ 完全类似的公式，即

$$V(\hat{R}_h) = \hat{R}_h^2\left(\frac{V(\hat{Y}_h)}{\hat{Y}_h^2} + \frac{V(\hat{X}_h)}{\hat{X}_h^2} + \frac{V(\hat{Y}_h) + V(\hat{X}_h) - V(\hat{U}_h)}{\hat{Y}_h\hat{X}_h}\right) \tag{31}$$

其中 $V(\hat{U}_h)$ 是以 $\hat{U}_h = \hat{X}_h + \hat{Y}_h$ 代替(30) 式中的 $\hat{Y}_h$ 而得。

---

① 我们曾对某些样本省，比较过那些层内按 Brewer 方法抽取二个样本市(地、州) 获得的 $\hat{\bar{Y}}_{h1}$ 与 $\hat{\bar{Y}}_{h2}$，发现一般差别都不大，按下面一小段中给出的此种情况的精确方差估计，亦即按(34) 式计算的层内方差估计 $V(\hat{\bar{Y}}_h)$ 的具体数字，若折算成对 $\hat{\bar{Y}}_h$ 的估计方差，并不比按(30) 式计算的相应量大很多(一般约高 20% 左右)，这表明将(27) 式第一项忽略所造成的影响并不很大。

从理论上说，只要有各样本市 $\bar{Y}_{hi}$ 的估计量 $\hat{\bar{Y}}_{hi}$，则根据多种重抽样方法(例如随机分组或平衡半样本等分法) 也可给出对(27) 式满意的估计方法。但在本项目中，限于当时计算条件没有实现。

3)若样本市是从层内按 Brewer 方法 πPS 抽样抽出，将抽得的两个样本市的编号简记为 1 和 2，则根据 Hurwitz-Thompson 公式，Y_h 的无偏估计量为：

$$\hat{Y}_h = \frac{1}{2}\left(\frac{\hat{Y}_{h1}}{z_{h1}} + \frac{\hat{Y}_{h2}}{z_{h2}}\right) \tag{32}$$

式中

$$z_{h1} = \frac{M_{h1}}{M_h}, z_{h2} = \frac{M_{h2}}{M_h} \tag{33}$$

$\hat{Y}_h$ 的方差估计采用 Yates-Grandy-Sen 估计，此时有

$$V(\hat{Y}_h) = \frac{4z_{h1}z_{h2} - A_h}{4A_h}\left(\frac{\hat{Y}_h 1}{z_{h1}} - \frac{\hat{Y}_{h2}}{z_{h2}}\right)^2 \tag{34}$$

其中

$$A_h = \frac{2z_{h1}z_{h2}(1 - z_{h1} - z_{h2})}{(1 - 2z_{h1})(1 - 2z_{h2})D_h} \tag{35}$$

又

$$D_h = \frac{1}{2}\left(1 + \sum^{i} \frac{z_{hi}}{1 - 2z_{hi}}\right) \tag{36}$$

上式中的求和是对层内所有市(地，州)求的。

$\bar{Y}_h, R_h$ 的估计及其方差估计公式与 2)中的相应公式完全相同，这里不再重复。

得到省内各层的有关公式后，各样本省的相应估计及其方差估计即可根据分层公式方便给出。具体公式与(15)—(21)式完全类似。

3.5 全国目标量的估计及其方差估计

从样本省的结果进行全国汇总时，仍采用分层抽样公式。其中直辖市一层用北京市的调查结果〔1〕，注意由于北京地区的调查对象包括中小学教师，因此在汇总时，作了必要的技术处理。因此本文中关于北京地区的结果不包括中小学教师，这一点与〔1〕中的结果不同，具体公式从略。

4 调查结果的精度分析

本项调查，采用派调查员面访的形式。问卷回收率达到 100%，有效率也达 99%以上。因此在分析中可以排除由于不回答造成的误差。

在抽样设计中，我们要求在置信水平 95%下，关于比例型目标量，估计的绝对误差对样本省不超过 2%(或 3%)。而抽样调查的实际精度必须按照上节中给出的方差估计公式计算。在获得各目标量估计的方差估计 V(·)后，根据大样本下估计量的渐近正态性，可换算成该项目标量估计实际达到的最大绝对误差 d*。对于置信水平为 1−α 时，

$$d^* = u_\alpha \sqrt{V(\cdot)} = u_\alpha S(\cdot) \tag{37}$$

其中 $S(\cdot)$ 是相应估计量的标准差估计。

估计量的精度还可用一定置信水平下的最大相对误差 r 或用估计量的变异系数 $cV(\cdot)$ 来表示。r 与 $cV(\cdot)$ 的关系为

$$r = u_\alpha cV(\cdot) \tag{38}$$

作为说明，我们在表4及表5中分别列出了全国及北京、江苏、四川、贵州(各类地区各取一个省、市代表)有关专业技术人员承担课题、获奖情况及流动情况若干指标的估计及其相应的标准差估计与变异系数的估计。从表中可以看出，对省(直辖市)的目标量估计的变异系数绝大多数都在7.5%以内，对全国的都在5%以内。换言之，对省(直辖市)目标量估计的最大相对误差在置信水平为95%时绝大多数在15%以内，而对全国则都在10%以内。

表4 专业技术人员承担课题及获奖总人数的估计与精度

		自然科学基金 863高技术 七五攻关	上级下达 指令性项目	横向课题	自选课题 及其它	获省、市、部委 以上奖人数
全国	$\hat{Y}$	351311	1353629	366322	1490434	1313036
	$S(\hat{Y})$	14979	29501	14684	27102	23846
	$cV(\hat{Y})$ (%)	4.26	2.18	4.01	1.82	1.82
北京	$\hat{Y}$	34581	59593	29126	74725	74866
	$S(\hat{Y})$	3550	3920	2522	4445	3874
	$cV(\hat{Y})$ (%)	10.27	6.07	8.66	5.95	5.17
江苏	$\hat{Y}$	16679	88942	20447	79761	75052
	$S(\hat{Y})$	1909	3641	2287	3696	3253
	$cV(\hat{Y})$ (%)	11.44	4.09	11.19	4.63	4.33
四川	$\hat{Y}$	25179	112040	32460	117189	112473
	$S(\hat{Y})$	2298	4336	2514	4400	4348
	$cV(\hat{Y})$ (%)	9.13	3.87	7.74	3.75	3.87
贵州	$\hat{Y}$	4058	28036	3579	29401	23027
	$S(\hat{Y})$	752	2145	591	1530	1578
	$cV(\hat{Y})$ (%)	18.45	7.65	16.51	5.20	6.85