# 食品安全体系的抽样理论研究

李 兵[1],陈国华[1],杨涤尘[1],朱 宁[2]

(1.湖南人文科技学院数学与应用数学系,湖南娄底 417000;2.桂林电子科技大学数学与计算科学学院,广西桂林 541004)

**摘要**　介绍了多层抽样方法,给出了分层抽样方案的理论无偏估计及试验设计检验的模型和统计量。探讨了膳食暴露模型、污染物分布模型和风险评估模型的建立,最后,提出了抽样设计方法。

**关键词**　暴露模型;污染模型;分层抽样;试验设计;风险评估模型

**中图分类号**　TS207.7　　**文献标识码**　A　　**文章编号**　0517 - 6611(2009)22 - 10336 - 02

## 1　问题的提出

我国是一个拥有13亿人口的发展中国家,每天都在消费大量食品,这批食品是由成千上万的食品加工厂、不可计数的小作坊、几亿农民生产出来的,并且经过较多的中间环节和长途运输后才为广大群众所消费,近年来我国经济发展迅速而环境治理相对落后,以至环境污染形势十分严峻。随着我国进出口贸易额的迅速增加,加上某些国外媒体的炒作,对外食品贸易中的矛盾日益尖锐,因此,建立包括食品卫生安全保障体系在内的公共安全应急机制是关系国计民生的重大而迫切的任务。

笔者拟从膳食暴露模型、污染物分布模型、风险评估模型三方面探讨食品安全模型的建立,并根据抽样设计原理提出抽样设计方法,以期为相关研究提供借鉴。

## 2　模型的理论基础

**2.1　多层抽样方案与设计原理**　方案与设计:独立实践 $r$ ($\geqslant 2$)层设计为 $P(\cdot)$ 和 $r(\geqslant 2)$ 阶追加设计为 $P+(\cdot)$,相应的一阶样本为 $S^{(1)}$,一阶追加样本为 $S^{(1)+}$。相应的终极抽样设计为 $P^{\otimes}(\cdot)$,一阶终极样本为 $S^{\otimes(1)}=S^{(1)}\cup S^{(1)+}$。

定理:在方案与设计的追加设计下,样本总量 $Y_D$ 的无偏估计为 $\bar{Y}_{D^+}=\sum_{\alpha\in S_b^{(1)+}}\frac{\bar{Y}_\alpha^{(1)+}}{\pi_\alpha^{(1)+}}$。其中,$\pi_\alpha^{(1)+}=P(\alpha\in S^{(1)+})$。

证明:文献[1]给出了两阶分层抽样样本总量 $Y_D$ 的无偏估计 $\bar{Y}_{\pi^\otimes}=\sum_{\alpha\in S_\otimes^{(1)+}}\frac{\bar{Y}_\alpha^{(1)+}}{\pi_\alpha^{(\otimes)+}}$,应用相似的证明过程很容易得出上述定理的结论,即 $E=(\bar{Y}_\alpha^{(1)+}\mid S^{(1)+})=Y_\alpha^{(1)}$。利用上述定理的方法和结论可以构造总体总量 $Y$ 的复合估计 $\bar{Y}(\bar{\alpha}_{opt}\mid\bar{\theta}_{opt})$,从而得出对总体的估计。

**2.2　试验设计原理**　设因子 $A$ 有 $a$ 个水平,因子 $B$ 有 $b$ 个水平,因子 $C$ 有 $c$ 个水平,从每个因子各取一个水平,共构成 $abc$ 个水平组合,每个试验重复 $n$ 次,共进行 $abcn$ 次。建立如下模型:

$$Y=\mu+X_1+X_2+X_3+(X_1X_2)+(X_2X_3)+(X_1X_3)+(X_1X_2X_3)+\varepsilon$$

式中,$X_1=(\tau_1,\tau_2,\cdots,\tau_a)$,$X_2=(\beta_1,\beta_2,\cdots,\beta_b)$,$X_3=(\gamma_1,\gamma_2,\cdots,\gamma_c)$,$\varepsilon\sim N(0,\Sigma)$。

方差分析问题要检验7个假设:$H_{01}:\tau_i=0,\cdots,H_{07}:(\tau\beta\gamma)_{ijk}=0$。其中,$i=1,\cdots,\alpha;j=1,\cdots,b;k=1,\cdots,c$。

由初等代数运算可以证明总偏差平方的分解公式:[2]

$SS_T=SS_A+SS_B+SS_C+SS_{AB}+SS_{AC}+SS_{BC}+SS_{ABC}+SS_E$。其中,$SS_T,SS_A,SS_B,SS_C,SS_{AB},SS_{AC},SS_{BC},SS_{ABC},SS_E$ 分别称为总偏差平方和。

## 3　模型的构建

**3.1　膳食暴露决定论与概率模型**　概率分析在暴露评估中有两个优点:①它允许模型考虑所有暴露的分布情况,包括从小到大所有模式和所有百分位数;②它包括了对暴露结果不确定性和参数灵敏性的综合分析。因为概率模型所提供的信息是暴露全部分布,因此暴露模型能够推断出不同方案对各部分的影响。概率分布还有利于某些风险利益分析。文献[3]提供的点估计不能给暴露模型或风险管理者提供摄入量在整个群体中的发生过程。但是在进行概率分析以前,暴露模型必须具备良好的信息来源(抽样数据),并处理好定量暴露评估的不确定性。对此,笔者提出了多层抽样方案方法,并进行了一个试验设计,对所选抽样数据进行优化以达到信息量的准确,从而总结出其概率分布密度函数,进而确定出各个分布之间的依赖性与相关性。

**3.2　污染物分布模型**　污染物分布模型是根据农药、化工等污染行业的污染物排放数据和食品卫生安全监测部门日常对水、农贸市场和大宗食品中污染物的抽查数据以及进出口口岸的检测数据来估计各类食物中污染物的含量。由于中国居民消费的食品种类比其他国家复杂得多,包括主食、肉类、蔬菜、水果、水、饮料、各种调味剂和经过加工的食品,细分将达数千种以上,在实际调查过程中进行如此详细地分类,其调查工作量过大,而如果随意粗糙地进行分类,则将影响调查的精度,因此,需要根据污染物分布模型的数据合理设计抽样调查中食物的分类办法[4]。笔者首先建立主成分模型以简化数据并进行分类,设 $X_1,X_2,\cdots,X_p$ 为调查的居民消费品种类的 $p$ 个随机变量。取50~100个样本为含有污染物的量或者是某种污染物的量。运用主成分分析建立第 $i$ 个样本主成分:

$$Y_i=e_i^{-T}x=\bar{e}_{i1}x_1+\bar{e}_{i2}x_2+\cdots+\bar{e}_{ip}x_p\qquad(i=1,\cdots,p)$$

由于消费食品种类过多,为了简化其变量的个数,笔者采用主成分分析法[5-6]。在综合评价中,应用主成分分析法既可消除各指标间不同量纲的影响,又可以消除由于指标相关性带来的信息重叠,特别是克服了评价中人为确定权重系数的问题。在主成分分析中,选取主成分的标准是选取累计

**基金项目**　国家自然科学基金资助(10361003);湖南省教育厅资助科研项目(07C389);湖南人文科技学院教改课题(RKJGY0947)。

**作者简介**　李兵(1983 - ),男,山东潍坊人,硕士,助教,从事多元统计分析研究。

**收稿日期**　2009-04-13

贡献率达到 75% ~80% 左右的因子[7],这样大大简化了消费品种的数目,然后使用聚类分析对其进行分类。根据实际情况每类中抽出 3 ~5 个进行分析,这样可大大减少工作量。考虑某种污染物对总体的影响呈对数 - 正态分布的趋势,即存在如下的对数关系:

$$\ln y = \sum_{i=1}^{p} x_i$$

式中,$x_i$ 表示某种污染物对总体的影响。就实际情况而言,某种污染物对其影响较小,则可以得出 $y$ 近似服从一个对数 - 正态分布(图 1)。

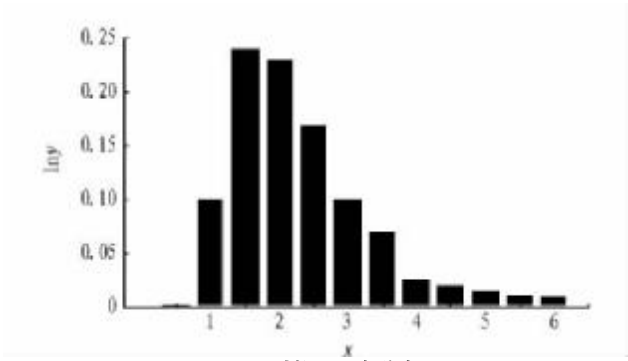


**图 1　对数 - 正态分布**

**3.3　风险评估模型**　根据前两个模型所提供的数据,计算得出全国或某地区人群某些污染物每天摄入量的 99.999% 的右分位点(把每个人每天某种污染物摄入量看成是一个随机变量),从而能够对某一时刻食品安全风险作出评估。根据分层抽样原则进行分层,进行试验设计找到一个合适的、准确描述中国国情的样本。污染分布模型通过主成分分析及石阶平博士给出的部分数据,得出近似服从一个 $m = 2.0$、$s = 1.0$ 的对数分布[8]。评估风险以 99.999% 的右分位点来描述,给出了 99.999% 的右分位点的概率值 $1 - P(x > c) = 0.999\ 99$ 所确定出的 $c$ 值。从而确定出污染的风险的评估标准。

## 4　抽样设计方法

**4.1　设计原则**　调查抽样设计按照科学、效率、便利、连贯的原则。首先作为一项全国性的大规模抽样调查,整体抽样方案必须是严格的概率(随机)抽样。要求样本既对全国有代表性,也要对部分有条件的省(自治区、直辖市)有代表性。其次,抽样方案必须要保证有较高的效率,即在同样工作量的条件下,方案设计应该使得调查的精度尽可能高,也即使目标量估计的抽样误差尽可能少。再次,抽样方案必须有较强的可操作性,不仅便于具体抽样的实施,而且也要求便于后期的数据处理[9-10]。科学与便利是其中的主要考虑原则。

**4.2　调查抽样的各阶划分**　对于一项全国性的抽样调查项目,要求样本对全国而言具有代表性(对全国目标量进行估计)。综合考虑各个因素,方案决定采用区、县作为第一阶抽样单元,为解决全国样本对省代表性不足、不能对省级目标量进行估计的问题,通过对有此要求的省及追加样本一起构成省级样本,用以对省级目标量进行估计。

根据全国行政区划资料,全国共有 1 689 个县级行政单位,437 个县级市,227 个地级市,这些地级市及北京、天津、上海、重庆 4 个直辖市总共包含了 737 个市辖区,上述即构成了第一阶抽样的抽样总体。

为了便于调查之后的资料分类、汇总及提高精度,需要将全国所有区、县进行分层,分层主要按照以下两种标志进行:地域;收入差距。

**4.2.1**　地域分层。我国幅员辽阔,各地经济、社会与文化的地域差异极大,因此,首先将所有区、县按照所在省的地理位置分成 3 大层。第一大层(东部地区):上海(1)、北京(2)、天津(3)、广东(4)、辽宁(5)、浙江(6)、江苏(7)、福建(8)、山东(9)。第二大层(中部地区):黑龙江(10)、河北(11)、吉林(12)、海南(13)、湖北(14)、山西(16)、湖南(17)、河南(18)、广西(19)、安徽(20)。第三大层(西部地区):内蒙古(21)、陕西(25)、宁夏(26)、甘肃(28)、青海(29)、新疆(15)、重庆(22)、四川(24)、云南(27)、贵州(30)、西藏(31)。省后括号中的数据为编号。

**4.2.2**　经济差别分层。同一大层的各市辖区与所属城市的规模,在城市中的地理位置和居民成分构成(非农业人口与农业人口的比例)有较大差异,各县也因经济、文化发达程度不同而有较大差异。笔者将各大层次所有区、县分成 6 类:一类区、二类区、三类区,一类县、二类县、三类县。具体标准有两个。

(1)区类别的划分标准。东部地区与中部地区:非农人口在总人口中的比例大于或等于 80% 为一类区,小于 80% 为二类区。西部地区:非农人口在总人口中的比例大于或等于 70% 为一类区,小于 70% 为二类区。由于农村域中一类区中的非农业人口总数很少,故农村域在抽样时,市辖区不再分层。

(2)县类别的划分标准。东部地区:人均 GDP 在 6 000 元以上为一类县;6 000 元以下 4 000 元以上为二类县,4 000 元以下为三类县。中部地区:人均 GDP 在 5 000 元以上为一类县;5 000 元以下 3 000 元以上为二类县,3 000 元以下为三类县。西部地区:人均 GDP 在 3 500 元以上为一类县;3 500 元以下 2 000 元以上为二类县,2 000 元以下为三类县。根据上述 2 个标准,利用相关资料划分出的全国区、县如表 1 所示。

**表 1　全国区县划分**

| 地域 | 一类区 | 二类区 | 三类区 | 一类县 | 二类县 | 三类县 | 合计 |
|---|---|---|---|---|---|---|---|
| 东部地区 | 125 | 154 | 169 | 84 | 78 | 89 | 699 |
| 中部地区 | 159 | 175 | 185 | 149 | 300 | 229 | 1 197 |
| 西部地区 | 57 | 73 | 97 | 188 | 237 | 242 | 894 |

## 5　结论

多层抽样并追加抽样的全面考虑,可以更加准确的描述中国的情况;对于全国所有区、县的分层,有利于资料分类、汇总和提高评价精度;对变化性和不确定性进行全面分析,其可靠性可以从分析过程和结果中得到。

## 参考文献

[1] 秦怀振. 抽样调查中若干理论与实践问题的研究[M]. 北京:中国统计出版社,2003.
[2] 王松桂. 线性模型引论[M]. 北京:科学出版社,1987.
[3] 罗祎,陈冬东,唐英章,等. 论食品安全暴露评估模拟模型[J]. 食品科技,2007(2):39 -42.
[4] 陈天金,魏益民,潘家荣. 食品中铅对人体危害的风险评估[J]. 中国食物与营养,2007(2):15 -18.
[5] 高惠璇. 应用多元统计分析[M]. 北京:北京大学出版社,2005.
[6] 范金城,梅长林. 数据分析[M]. 北京:科学出版社,2002.
[7] 陈希孺,王松桂. 近代回归分析[M]. 合肥:安徽教育出版社,1987.
[8] 石阶平. 食品安全风险评估与管理——加拿大的经验和中国的实践[Z]. 食品安全协调司,2007.
[9] 李金昌. 抽样估计精度问题研究[M]. 北京:中国物价出版社,2000.
[10] 袁建国. 抽样检验原理与应用[M]. 北京:中国计量出版社,2002.