应用概率统计 第一卷
第二期 1985 年 11 月

Chinese Journal of Applied Probability
and Statistics Vol. 1. No. 2. Nov. 1985

全国粮食农药污染调查抽样方案的设计与数据处理方法

全国粮食农药污染调查技术组
冯士雍 程翰生 汪仁官 (执笔)

§1 引 言

根据国务院指示，1984 年 3 月至 10 月由国家环境保护局与商业部、农牧渔业部共同组织了一次全国粮食受六六六与滴滴涕农药污染情况的大规模抽样调查. 调查的目的是对全国各省、直辖市、自治区(除西藏、台湾外)1983 年生产的主要粮食(小麦，早、中、晚稻及玉米)中六六六和滴滴涕残留量的超标率、检出未超标率及未检出率与平均残留量作出全面而精确的估计. 作为调查技术组的成员，我们承担了制定粮食采样点分布方案(以下称抽样方案)及提出相应数据处理方法的工作. 现将所用的抽样方案、目标量的估计与精度公式及其理论依据报告如下.

§2 抽 样 方 案

(1) 采样点的确定 由于作为调查对象的粮食是一种散料，因此在制定具体抽样方案前需确定基本抽样单元. 我们选取乡(公社)级粮库作为基本抽样单元，称为采样点. 对每个被抽中的采样点，根据粮食品种及不同的存贮方式，按规定的方法，采取有代表性的样品，经充分混和后，分取一公斤样品作为试样送检. 这份试样完全作为相应采样点此种粮食的代表. 例如若试样中六六六含量超标，则相应采样点的该种粮食都按六六六含量超标计算.

(2) 抽样方案的类型 调查采用分层两级不等概率随机抽样法，将 28 个省(市、自治区)作为层，全部进行抽查. 每层中第一级抽样(省抽采样县)采用与县的该种粮食产量成比例的概率无放回的抽取方法(详见(4)). 第二级抽样(即采样县中抽采样点)则采用简单随机抽样，每个采样县抽取数目相同(8 个)的采样点. 采用这个方案的原因是：样本代表性好，实施方便，并有可能采用简单的数据处理方法. 调查结果表明，所得估计量的精度较高.

(3) 采样县与采样点数的确定 样本量大小取决于调查精度和调查费用(工作量)之间的平衡. 由于全国规模的粮食农药污染调查还是首次进行，缺少现成的污染程度及差异的有关资料，加之因时间紧迫不允许事先作试验性调查，因此只得从控制总工作量的前提下考虑样本量的大小，并对各层(省)作合理的分配.

从总的工作量考虑，各种粮食的采样点数以控制在 5000—6000 个为宜，即在所调查的粮

本文 1985 年 5 月 14 日收到.

食种类中平均每亿斤粮食取一个样. 为保证全国及各省的调查精度，以及不使产量高的省工作量过大，我们采用各省每种粮食的采样县数(以及采样点数)与该省的这种粮食的产量的平方根成正比的原则确定. 各省每种粮食按其产量所分配的采样县数见表1.

表 1

实际产量(亿斤)	所需采样县数	实际产量(亿斤)	所需采样县数
0.3 以下	0	110 以上—132	11
0.3—2	1	132 以上—156	12
2 以上—6	2	156 以上—182	13
6 以上—12	3	182 以上—210	14
12 以上—20	4	210 以上—240	15
20 以上—30	5	240 以上—272	16
30 以上—42	6	272 以上—306	17
42 以上—56	7	306 以上—342	18
56 以上—72	8	342 以上—380	19
72 以上—90	9	380 以上—	20
90 以上—110	10		

在制定方案时，尚缺各省(市、自治区)1983年分粮食种类的产量数据，因此在方案制定过程中都用1982年的相应数据代替. 实际抽取的采样县及采样点数按粮食种类划分，见表2.

表 2

粮食种类	实际采样县数	实际采样点数
早稻	136	1088
中稻	52	416
晚稻	146	1168
小麦	159	1272
玉米	141	1128
合计	634	5072

方案还允许各省根据具体情况及需要，自设补充的采样县. 但从这些采样点得到的数据在处理时不与按随机原则确定的数据混合. 加上补充采样县，实际采样县总数为679个，采样点数为5432个.

(4) 各省采样县的具体抽取方式

省内抽取采样县是按各县该种粮食的产量大致成正比的不等概率随机抽样办法抽取的. 具体抽取步骤是，首先根据该省这种粮食的总产量在表1中查得所需采样县数，按各县的产量赋予每个县以与其产量成正比的代码个数(例如每0.1亿斤一个代码)，代码按全省各县级单位的自然顺序统一编号. 若代码总数为 d，则利用计算机产生1到 d 的(离散)均匀分布随机数，与所产生的随机数代码相应的县就作为抽中的采样县. 直至所需的采样县数满足为止.

在抽取过程中，若一个县被抽取到两次或多于两次，则仍作为一个采样县处理. 而以后面的随机数所代表的县依次递补. 显然，实际采用的抽取方法是无放回的抽样方法. 每次抽取时，每个当时还未被抽中的县被抽中为采样县的概率为该县的产量对未被抽中县的总产量的比. 即若令 Y_{hi} 为 h 省第 i 个县的产量，$Y_h=\sum_{i=1}^{N_h} Y_{hi}$ 为全省总产量，设前 $k-1$ 次抽中的采样县为

$i_1, i_2, \cdots, i_{k-1}$，则第 k 次抽中 i_k 县的概率为

$$P\{\text{第 } k \text{ 次抽中 } i_k \text{ 县} \mid \text{前 } k-1 \text{ 次抽到 } i_1, i_2, \cdots i_{k-1} \text{ 县}\}$$

$$= \frac{Y_{h,i_k}}{Y_h - \sum_{l=1}^{k-1} Y_{h,i_l}}, \quad i_k \neq i_1, i_2, \cdots, i_{k-1}. \tag{1}$$

§3 对目标量的估计及其精度公式

调查数据的统计计算是根据粮食样品分析所得的六六六和滴滴涕残留量数据，结合1983年的实际产量，计算每种粮食按各采样县、各省及全国每种农药残留量的超标率、检出未超标率、未检出率和平均残留量的估计量和它们的精度。鉴于各种率的估计公式与精度公式对不同粮食、不同农药都是相同的，而对平均残留量也只须作少许变化就能采用同样的公式，因此下面仅以一种粮食一种农药的超标率为例给出有关的计算公式.

(1) 记号　本节中涉及的各种主要记号的含义如下：

a) 编号　省编号 h, $h=1, 2, \cdots, L(L=28)$；县编号 i, h 省中实际县数为 N_h，而采样县数为 n_h；县中采样点的编号为 j, $j=1, 2, \cdots, 8$.

b) 1983年该种粮食的产量用 Y 表示，1982年产量用 Y' 表示. 特别，Y_{hij} 表示 h 省 i 县 j 点的1983年产量，$Y_{hi\cdot}$ 为 h 省 i 县的产量，$Y_{h\cdot\cdot}$ 为 h 省的总产量，$Y_{\cdots}$ 为全国总产量. 若在 Y 上打上“′”号，则表示相应的1982年产量.

c) 真实超标率记为 p，相应的估计量记为 $\hat{p}$.

d) $\lambda_{hij} = \begin{cases} 1, & \text{若 } h \text{ 省 } i \text{ 县 } j \text{ 点的粮食样品分析结果超标;} \\ 0, & \text{否则.} \end{cases}$

注：若令 λ_{hij} 为该点粮食样品分析结果的农药残留量，则所有计算公式中的 p 即为平均残留量.

(2) 县超标率的估计　h 省 i 县的真实超标率

$$p_{hi} = \frac{h \text{ 省 } i \text{ 县(该种粮食 1983 年) 产量中的超标部分}}{h \text{ 省 } i \text{ 县(该种粮食 1983 年) 总产量}}$$

$$= \frac{\sum_j \lambda_{hij} Y_{hij}}{\sum_j Y_{hij}} = \frac{\sum_j \lambda_{hij} Y_{hij}}{Y_{hi\cdot}}. \tag{3}$$

式中的求和是对县中的所有点进行的. (3)式是一个比值. 由于在一个采样县中取的采样点数(8个)相对于一般县中的总点数(乡或公社的粮库数)比例较大，p_{hi} 可以用采样点的数值估计：

$$\hat{p}_{hi} = \frac{\sum_{j=1}^{8} \lambda_{hij} Y_{hij}}{\sum_{j=1}^{8} Y_{hij}}. \tag{4}$$

$E(\hat{p}_{hi})$ 与 p_{hi} 的偏差也甚小，以下我们将这个偏差忽略不计，即假定

$$E(\hat{p}_{hi}) \approx p_{hi}. \tag{5}$$

(3) 省超标率的估计及方差计算　省超标率可表成

$$p_h = \sum_{i=1}^{N_h} p_{hi} \frac{Y_{hi\cdot}}{Y_{h\cdot\cdot}} \tag{6}$$

根据各省中采样县的抽取方式，即无放回的与产量有关的概率抽样(见公式(1)，以下简称无放回抽样)，省超标粮食数 $Y_{h..}p_h$ 的估计可用 Murthy(1957)的公式

$$Y_{h..}\hat{p}_h=\frac{\sum_{i=1}^{n_h}P(S|i)\hat{p}_{hi}Y_{hi.}}{P(S)}$$

从而

$$\hat{p}_h=\frac{\sum_{i=1}^{n_h}P(S|i)\hat{p}_{hi}Y_{hi.}}{P(S)Y_{h..}} \tag{7}$$

其中 $P(S)$ 是 h 省中按无放回抽样抽到特定样本 S(大小为 n_h)的无条件概率，$P(S|i)$ 是在抽样中已知第一个抽到第 i 县而获得特定样本 S 的条件概率. 由于对固定的 i, $\sum_S P(S|i)=1$. (其中求和是对所有大小为 n_h 的样本求的. 下同)，因此

$$E(\hat{p}_h)=E_1E_2(\hat{p}_h)\approx E_1\left[\frac{\sum_{i=1}^{n_h}P(S|i)p_{hi}Y_{hi.}}{P(S)Y_{h..}}\right]=\sum_S\sum_{i=1}^{n_h}P(S|i)p_{hi}Y_{hi.}/Y_{h..}$$

$$=\sum_{i=1}^{N_h}p_{hi}Y_{hi.}/Y_{h..}=p_h \tag{8}$$

从而 $\hat{p}_h$ 是近似无偏的(这里的近似仅是由于(5)式引起的).

在无放回抽样情形，$\hat{p}_h$ 的方差估计为

$$v(\hat{p}_h)=\frac{\sum_{i=1}^{n_h}\sum_{j<i}[P(S)P(S|i,j)-P(S|i)P(S|j)]Y_{hi.}Y_{hj.}(\hat{p}_{hi}-\hat{p}_{hj})^2}{[P(S)Y_{h..}]^2} \tag{9}$$

式中 $P(S|i,j)$ 为在前两个抽到第 i 县和第 j 县(不考虑其次序)情况下，抽到特定样本 S 的条件概率. [严格地说，为使 $v(\hat{p}_h)$ 是 $V(\hat{p}_h)$ 的无偏估计，(9)式还应添加一项与第二级(即是抽点)抽样效应有关的小量，参见 Cochran(1977)第十一章].

公式(7)与(9)的计算量非常大，若用它们处理调查所得的所有数据有困难. 因此我们寻求替代办法.

将上述无放回抽样按与产量成正比的概率有放回抽样处理. 即在省抽县过程的 n_h 次抽样中，每个县每次被抽到的概率都为 $Y_{hi.}/Y_{h..}$，则 p_h 可用 $\hat{p}_{hi}$ 的算术平均数估计(Cochran(1977)第十一章)：

$$\hat{p}_h=\frac{1}{n_h}\sum_{i=1}^{n_h}\hat{p}_{hi} \tag{10}$$

它是 p_h 的无偏估计(若不考虑 $\hat{p}_{hi}$ 对 p_{hi} 的偏差). 而 $\hat{p}_h$ 的方差的无偏估计为：

$$v(\hat{p}_h)=\frac{1}{n_h(n_h-1)}\sum_{i=1}^{n_h}(\hat{p}_{hi}-\hat{p}_h)^2 \tag{11}$$

参见[1]第十一章，作为有放回抽样的方差估计(11)比无放回抽样的方差估计(9)为大，也即(11)式给出 $\hat{p}_h$ 的方差估计的一个上限.

在应用上述公式时，还有一个问题需要考虑. 在制定省抽县的方案时，我们采用的是概率与每个县的 1982 年产量 $Y'_{hi.}$ 成比例，而不是与 1983 年的实际产量 $Y_{hi.}$ 成比例. 下面我们证明，只要假定 1983 年的污染程度与 1982 年的近似相等(粮食农药残留量主要与土壤残留的农药量及当年农药施用量有关，因此六六六与滴滴滴污染程度在它们完全停止使用前相邻两年间的变化不会很大)，也即假定

$$p_{hi} \approx p'_{hi}, \quad i=1, 2, \cdots, N_h; \quad p_h \approx p'_h \tag{12}$$

则上面的结论，例如 $\hat{p}_h$ 的无偏性亦近似成立．这是因为：

$$E(\hat{p}_h) = E_1 E_2\left(\frac{1}{n_h}\sum_{i=1}^{n_h}\hat{p}_{hi}\right) = \frac{1}{n_h} E_1\left[\sum_{i=1}^{n_h} E_2(\hat{p}_{hi})\right] \approx \frac{1}{n_h} E_1\left(\sum_{i=1}^{n_h}\hat{p}_{hi}\right)$$

$$= \frac{1}{n_h}\sum_{i=1}^{n_h} E_1(p_{hi}) = E_1(p_{h1}) \approx E_1(p'_{h1}) = \sum_{i=1}^{N_h} p'_{hi}\frac{Y'_{hi\cdot}}{Y'_{h\cdot\cdot}} = p'_h \approx p_h$$

这里样本值与总体值用了相同的记号，实际期望号 E_1 下的 p_{hi} 和 p'_{hi} 为样本 S 的第 i 个样的值．

表 3 是两组数据分别按精确的无放回抽样公式(7)，(9)与按有放回近似公式(10)，(11)的比较．

表 3

	超标率的估计 $\hat{p}_h$		超标率的标准差 $\sqrt{V(\hat{p}_h)}$	
	按式(7)	按式(10)	按式(9)	按式(11)
样本 1($n_h=5$)	10.03%	10.14%	4.40%	4.89%
样本 2($n_h=7$)	6.26%	6.17%	2.66%	2.96%

从表中可以看到近似公式(10)和(11)与精确公式(7)、(9)相差甚微．因此为计算方便起见，我们实际采用的是按有放回抽样的近似公式．

(4) 全国超标率的估计与方差公式

按分层抽样公式，从各省超标率的估计 $\hat{p}_h$ 及其方差估计 $v(\hat{p}_h)$ 可得全国超标率

$$p = \sum_{h=1}^{L} p_h Y_{h\cdot\cdot}/Y_{\cdot\cdot\cdot} \tag{13}$$

的估计 $\hat{p}$ 及其方差 $V(\hat{p})$ 的估计 $v(\hat{p})$ 如下：

$$\hat{p} = \sum_{h=1}^{L} W_h \hat{p}_h \tag{14}$$

$$v(\hat{p}) = \sum_{h=1}^{L} W_h^2 v(\hat{p}_h) \tag{15}$$

其中层权

$$W_h = Y_{h\cdot\cdot}/Y_{\cdot\cdot\cdot} \tag{16}$$

是各省产量对全国总产量的比．只要 $\hat{p}_h$, $v(\hat{p}_h)$ 是 p_h, $V(\hat{p}_h)$ 的无偏估计，则 $\hat{p}$, $v(\hat{p})$ 分别是 p, $V(\hat{p})$ 的无偏估计．

参 考 文 献

[1] Cochran, W. G., *Sampling Techniques*, 3rd ed. John Wiley & Sons, 1977.

[2] Murthy, M. N., Ordered and unordered estimators in sampling without replacement. *Sankhya*, **18** (1957), 379—390.

SAMPLING AND DATA ANALYSIS OF NATIONWIDE GRAIN PESTICIDE CONTAMINATION SURVEY IN CHINA

THE TECHNICAL GROUP OF NATIONWIDE GRAIN PESTICIDE CONTAMINATION SURVEY WRITTEN BY

FENG SHIYONG CHENG HANSHEN WANG RENGUAN

(Inst. of Systems Science, Academia Sinica) *(Beijing University)*

In 1984, a nationwide survey on grain pesticide contamination was engaged in China. The main purpose of the survey is to estimate the average residues level and the percentages over the standard level of 666 (hexachloro-cyclohexane) and DDT (dichloro-diphenyl-trichloroethane) of various kinds of grain produced in 1983. A stratified two-stage sampling plan was designed for the survey using the provinces as strata, counties as primary units in each stratum and grain storehouses of townships (or communes) as subunits. Draw a sample of counties with probabilities (approximately) proportional to the grain production of counties without replacement and subunits with simple random sampling.

The formulae for estimating the percentages over the standard level (or other characteristics) and their approximate variances are given for each province and the whole country using the stratified two-stage sampling with probability proportional to size with replacement. The results are compared with the accurate estimates according to the actual plan using Murthy's estimate.