第三次全国统计科学讨论会
论 文 选 编

（三）

数理统计的应用和抽样调查的推广分册

中国统计学会秘书处

1983年12月

目　录

论 抽 样 调 查

广东省计委　陈应中

一、抽 样 调 查

抽样调查的目的是获得一个规模有限的、能够代表总体（全及总体，下同）的标本总体（以下简称样本）。样本是总体的缩影。

抽样调查是非全面调查中用来推算总体的最完善、最有科学根据的一种调查方法。它从总体各单位中抽出一定数量进行调查，以其结果推算总体的综合指标的。

（一）论据。抽样调查是数理统计的前提。数理统计的频数分布、平均数、差异程度、显著性检验、假设检验、方差分析、回归分析、极值分析等，就是以抽样理论为前提的。因此也可以说数理统计实质上就是解决如何从样本来了解或研究总体的问题。

抽样调查计算样本数量与抽样误差是根据大数定律和概率论推演出来的。大数定律又名平均数定律。即在总体的各个体中，普遍存在着本质相同的某种现象的规律性，这种同类现象的规律性，只有在大量现象中，个体现象的偶然差异，由于平均结果，才相互抵消，它的规律性便在一定的数量和质量上表现出来。这就是恩格斯说的："在表面上是偶然性起作用的地方，这种偶然性始终是受内部的隐蔽着的规律支配的，而问题只在于发现这些规律"（恩格斯：《路德维希费尔巴哈和德国古典哲学的终结》，人民出版社。第二卷第468页）。

按概率论是一个数，分子是某事件可能发生的机会数，分母是全部可能发生的总机会数。这个总机会数也可以说是某事件可能发生的机会数的总体。

总体是同性质的个体组成的。抽样调查是通过样本总体反映总体的。但是不是个别样本能代表总体，任何个别样本的特征都反映不了总体的特征的。只有样本总体的综合特征：平均数（或比例关系）才能反映总体相应的综合特征的。在综合指标的数目足够多数时（一般超过30个数目的），个别现象的偶然性才会经过平均而互相抵消，客观现象的规律才充分表现出来。这也是恩格斯说的："必然的东西，通过无数的偶然性而给自己开辟道路"（马恩选集第五卷第468页）。这种用平均数办法去估计总体，是最准确的估计量，其他估计量，如比率法：用抽样总体与总体的比率，来求出总体估计量；回归法：用回归方程，来求抽样总体对总体的回归估计量，都不如平均数估计量精确。这是现代所普遍应用的方法，也是本文所商讨的估计量的内容。

（二）总体结构和离差。

根据大数定律正态分布定理，总体的综合指标（平均数或成数）必将落在抽样总体综合指标（平均数或成数）$\pm u$的范围内。用公式表示：总体平均$\bar{Y}_{总}$ = 抽样总体平均$\bar{Y}_{n} \pm u$。这就是说，由于抽样总体平均不可能恰恰等于总体平均，总有大于或小于总体

平均的，这就出现离差（σ），离差大小取决于总体内涵大小和差异程度。在差异程度相等情况下，总体划得愈小，也就是把总体划成几个群体，各群体内部的差异程度就愈小，差异程度愈小，方差也愈小，方差愈小，抽样误差（u）也愈小。用公式表示：设总体有15个单位，为便于说明起见，按等差级数排列成1，2，3，4，5，6，15，则：

$$\overline{Y}=\frac{1+2+3\cdots+15}{15}$$

$$\sigma=\sqrt{\frac{(8-1)^2+(8-2)^2+(8-3)^2+(8-4)^2\cdots(8-5)^2}{15}}$$

$$=\sqrt{\frac{49+36+25+16+9+4+1+1+4+9+16+25+36+49}{15}}$$

$$=\sqrt{\frac{98+72+50+32+18+8+2}{15}}=\sqrt{\frac{280}{15}}=\sqrt{18.6}$$

如果分为三个类型，则第一类型从1到5，

$$\sigma=\sqrt{\frac{(3-1)^2+(3-2)^2+(3-4)^2+(3-5)^2}{5}}$$

$$=\sqrt{\frac{4+1+1+4}{5}}=\sqrt{\frac{10}{3}}=\sqrt{2}$$

第二类型从6到10，

$$\sigma=\sqrt{\frac{(6-8)^2+(11-8)^2+(a-8)^2+(10-8)^2}{15}}=\sqrt{2}$$

第三类型$a=\sqrt{2}$

都比按总体抽样的方差小三倍多。

广东农产量抽样调查，就是根据这个道理，把省一级粮食产量总体，从省到生产队分为四级：省、县、社队、生产队。这样划分结果，就把全省粮食每造亩产，由最低500斤到最高1000斤相差500斤的总体，划小到按生产队的600～650，650～700，700～750，750～800，800～850，850～900，900～950，950～1000八个相差50斤的小总体。也就是把总体内部的差距缩小到十分之一。此外，为减少总体内部的差异程度，还在选取样本时，和老农、有经验的干部，就样本所在的土地类型，稻谷品种和经营管理水平等，在本类型内有无代表性，进行田间考察。经过这几个过程，就使样本更富有代表性了。

但是仅仅计算抽样平均离差是不够的，它往往不足以反映总体内部的不同结构的不同离差。总体这种不同结构，往往表现在不同群体上。因此引起离差的场合是多种多样的：一由于总体内部不同差异程度引起的；二总体各单位或多或少的变异度相同，但群体所包含的基本单位数不同引起的；三各群体中抽取的样本平均数不同引起的，解决这种由于总体内部不同离差对抽样的影响，应当根据不同情况，分别处理。

要减少总体各单位的差异程度对样本的影响，要按两种最常见的误差，分别处理。一种只能靠分类抽样、不等概率抽样或平均数的平均抽样加以处理；另一种要看情况而定，凡是总体内的每一群体都有相同的变异度，但不是所有群体的变异度都相同的，这

要在每一群体中抽取一个基本单位样本，就能得到所需要的数据，因而群体可以分得更少，估计量就会更精确。

用公式表示：$\bar{Y}n = \frac{1}{M}\sum_{i=1}^{n}\bar{M}_{Ni}$

$\bar{Y}n$代表平均数估计量；

M代表群体；

$\bar{M}_{Ni}$代表总体中第i个群体数值；

凡是各群体间具有高度的同类相关，群体中的基本单位非常相似，用两层抽样较好，第一层抽群体，第二层再从所抽群体中的基本单位随机抽选；或用单一层不等概率整群抽样。如果不能用不等概率抽样，就要使群体的平均数涵量尽可能划小，有时还必须作人为的划分，使群体内涵（基本单位数）近于相等。还有：每一群体都是随机样本，群体间的变异度只许在随机误差的范围以内。如果各群体的内涵相等又较小，例如都不超过10个，可用等概率整群抽样，可以得到准确的估计量。

最后，各群体中所抽取的样本的平均数不同发生的误差。处理这种数据决不能用算术平均数公式来求估计量。这种做法是把各群体所占比重等同起来，而必须根据总体中的基本单位的不同平均程度分别处理。

设以$\bar{Y}$代表估计平均数。$\bar{X}_1$，$\bar{X}_2$，$\bar{X}_3$，……$\bar{X}n$代表各类型平均。n_1，n_2，n_3…n_n代表各类型数。

则：$\bar{Y} = \frac{1}{N}\Sigma(n_1\bar{X}_1 + n_2\bar{X}_2 + n_3\bar{X}_3 + \cdots\cdots n_n\bar{X}_n)$

二、纯随机抽样与非纯随机抽样

在抽样调查的认识方面，有个很突出而普遍的见解，认为抽样调查只能在纯随机范围内进行，一点也离不开随机。这个问题不解决，抽样调查就难于发展。因此，首先要着重商讨随机性问题，也就是纯随机抽样与非纯随机抽样问题。

（一）随机性问题。所谓纯随机抽样，就是总体中的每一单位都有可以计算的、不等于“0”的被抽中的概率。能够在一定可信程度下推断总体、测定抽样误差，对总体进行定量分析。它的缺点：一是遇到总体各单位的差异大时，就要很大样本；二是存在着因机遇变动带来的不稳定性。即从同一总体中，按同样的样本容量几次抽样，可能有几种不同的甚至很大的差异程度；三是纯随机抽样没有充分利用人对客观事物的已有认识，不利于发挥人的主观能动性。

非纯随机抽样不是这样。调查单位的取得，不是仅仅根据随机原则；还根据人对客观事物的已有认识、根据具体情况，主动地、深入地认识事物，按有限随机原则，选出有代表性的样本。例如平均数的平均抽样和按估计量标识排队等距抽样，就不是按纯随机抽样：使总体每一单位都有被抽选的机会；只是某一特定单位有被抽样的机会。因而是“有限制的随机”。因此，对“随机”问题，不能强调得太绝对化了，只要可以求得

推算全面的更正确的数据，又可以大大缩小抽样误差，就应当使用。任何科学都是在发展中前进，在发展中引导出新的规律；绝不能墨守陈规地沿袭过去的一套老办法。否则科学就不会有新的发展了。

根据广东经验，在抽取同等数量的样本条件下，来划分准确程度：那么，随机性最低的，如划类造点，准确程度最高；随机程度最高的，如纯随机抽样，连起点定在那里，也是随机的，在同等样本条件下，准确程度最低。

为什么说平均数的平均抽样的随机程度最低呢？因为这个抽样调查的抽选是按总体各单位排队和分组，抽取各组的组中值的。这样，不是组中值单位就没有抽中机会了。因而不是按纯随机原则的。可是它利用了现有统计资料，了解总体有关情况，采取更简便、抽样误差更小办法，抽选那些最富有代表性的单位进行调查，因而它是一种比较好的抽样调查方法。

等距抽样的随机性大于划类选点，特别是按无关标识排队的，因为它接近于总体的每一单位都有被抽中的机会。按有关标识排队，其随机性少于按无关标识排队，特别其第一次抽样从组中值抽起的。因为这样抽样，接近于平均数的平均抽样了。还有，无论那一种抽样，除划类选点外，遇到差异程度很大时，误差就也很大。如广西的玉米产区靖西15公社。南北片的亩产太悬殊。按等距抽样抽很大的比重，达47%，全年计算，误差还不大，但是分季节计算，早玉米误差9.87%，晚玉米误差9.65%。把15公社分为南北两类后，误差大降。所以遇到类似情况，就应当采取划类选点办法，把抽样误差降下来。

（二）随机性与代表性。以纯随机为主，还是以代表性为主，这是纯随机与非纯随机抽样的根本分歧，是随机问题的关键。如果只顾随机原则，那么只有纯随机抽样了。如果既要富有代表性，又要有限的随机，那么只有划类选点了。按划类选点抽样，由于类型是客观存在的事物内在联系，是科学的分类。类型内的组中值是客观事物围绕在平均数周围集结起来的实际形态，是平均数定律的必然结果。因而这种抽样是科学的切合实际的。但是现在对划类选点等还有许多不同的看法。需加以探讨。

有的对划类选点的起点选在亩产中心的一个田丘上，除中心地块外，其他没有被抽的机会；上下两头突大突小的极少数亩也没有被抽的机会；遇到中心点有两个地块亩产相同，就会参什主观意图；遇到调查地块多数是双数，就要除一部分地块不调查；只顾了抽样地块的平均单产在调查地块的单位之间的均匀分布，但播种面积（或收获面积）越大的，没有越大的可能被抽中机会。这些意见，除播种面积越大的，应加权便其在实际上被抽中的机会也加权而增大的：认为不从组中值抽起的见解是背离代表性原则的。如果不从组中值抽起，则第一个样本大于或小于平均数的，整个样本就偏高或偏低，缺乏代表性了。这是在解决样本代表性与随机原则时，把随机原则绝对化了。

有的认为类型的组距规定，会参什个人意见不合随机原则。当然认识客观规律有个过程，分组的组距也是这样。但是我们是辩证唯物论者，而不是不可知论者，终究会使组距划分得恰到好处的。

除此外，其他论点有的属于主观能动性问题；有的属于客观事物的特性（如双数）

需特殊处理，而不是什么违背随机原则问题。

（三）随机原则问题还要进一步澄清。上面事实说明：随机原则问题是普遍存在问题，必须进一步搞清楚。我国长期以来都过分强调随机原则，在抽样调查上，只限于等概率抽样——随机抽样，这样，就把自己手足缚死了。现在必须把现代的各种抽样调查方法，结合我国实际，独立自主地加以发展提高，而不能墨守成规，故步自封，把划类选点、不等概率抽样等其他方法，摒之于门外。

根据最近上海财经学院翻译的，联合国统计局出版的《抽样调查理论基础》一书，概括了现代抽样调查，除等概率抽样的，还发展到不等概率抽样和平均数平均抽样等种类型的二十几种不同的抽样方法。并说："以前的统计课本提醒统计人员，关于使用'平均数的平均'的危险性，现代抽样理论则不拘泥于数条"。这说明在国际上抽样调查从理论到实践都大大发展了。我国在五十年代就已经开展了划类选点的平均数平均的抽样调查，也已经有近三十年历史了。现在这个方法已被全国许多省份所广泛采用。最近我们到南海县平洲公社夏西大队调查。据植保员陈众说，平州公社20个大队中是以粮产中等水平的夏西大队为代表队的。按大队产量分片选中等生产队四个，分别代表四片，在每片中选中等户，用中等户的亩产平均，户数加权推算公社总户。计两个人用3—4天工夫就搞好了。因为实行包产到户后，算不出生产队总产量了，只能靠选中等水平的生产队的分片平均亩产来推算总产。这个材料说明，在依靠按户取得材料后，平均数的平均抽样方法显得更重要了。

根据"不塞不流，不止不行"道理，还要分清随机性与主观抽样的原则区别。按随机性是一回事，主观抽样又是另一回事。不能把这两种不同性质的东西混淆了。如果混淆了，就混淆了科学与反科学的界限。很明显，主观抽样是市侩行为，把某些商业意图，上下其手，以达到其不可告人营私舞弊的目的。例如在彩票中标号码中的弄虚作假，在博奕的六面掷子的一面贯钻，使其偏重而出现有利于赌博的企图，等等，都属于主观行为。这些主观的卑鄙行为，怎么能够和科学的、不同方式的抽样方法等量齐观呢？

三、排队标识与抽样

（一）排队标识要和估计量标识一致。这样，抽样结果，就能够达到所要求的估计量；如果不一致，两者脱节了，就得不到预期的目的。例如估计水稻或小麦产量，就不能用粮食这个标识排队，而必须用水稻或小麦标识排队。因为水稻、小麦、玉米、高粱等几种农作物的产量之间并不存在着确定的依存关系。因而粮食作物单产这个指标，不能同时确切地反映这几种农作物各自单产的高低情况，同时对抽样调查排队和计算误差作用也不大，应按各种农作物的各自近三年的平均亩产这个标识排队。也就是按常年产量排队。如果前三年遇到灾害，则应当改按当年预计产量排队。此外，农作物的品种不同，产量也各异，还要按不同品种的不同标识分别排队。如广东南海县种的六种早稻面积，1982年亩产由731斤到878斤，相差二成，1983年比1982年增产，由13%到66%。相差二倍多。针对着增产幅度大的连片分品种种植特点，按每一品种不同面积抽取不同比

例的田块，实割实测。该县平洲公社西二队，杂优面积占总面积46.9%，所抽样本面积占总样本44.4%；桂潮面积占40%，样本占40%；糯谷面积占10.3%，样本占11%。这样，分品种按亩数加权推算平均亩产910.6斤，接近于实产。如果品种分散，产量不一，则要打破田块和品种界限，按原来土名地段等距均匀放样实测，再按亩数加叉平均推算总产。凡地段面积不及50面的，抽 6 ～10个样本，50～100亩的抽10～15个；100亩以上的抽15～20个。这种做法，可以克服品种差异带来的推算困难，还有利于推广良种。

广西曾经按全年粮食产量这个有关标识排队，在全省86县中抽取20%左右的县对粮食产量有代表性，但对品种、分季节的，都没有代表性，比容许误差 2 %超过很多，有的超过几倍。改按稻谷分早晚季节标识排队后也抽取20%县，抽17个样本，方差60斤，抽样误差

$$U=\sqrt{\frac{60}{17}\left(1-\frac{17}{86}\right)}=\pm 1.8\text{斤}$$

按T = 2，3.5斤，0.76%，误差较小。所以按估计量的标识排队，是排队时必须遵守的一个原则。

西安统计学校郑人杰同志对在农产量抽样调查按甚么标识抽选样本。主张按1963年出版的《农产量抽样调查》提出的“按亩产量而不应当按面积”。标识排队，因为累计面积等距抽样。由于亩产的高低和面积的大小决不会成正比例。因而按累计面积的落点来分组，就会出现串组和跨组的现象，影响抽样效果。只有在亩产的变异与面积的变异基本一致的情况下，才能使样本单位在亩产不同的类型内均匀分布。否则低产田多而面积大的，则偏低；反之则偏高。

而亩产分组等距抽样，则把抽样标识和排队标识结合起来。一是亩产分组，上下组段明确，不会串组和跨组。二是各组组距相等，每个类型组内各单位亩产不致过分悬殊，既不会把亩产差距增大、属于不同生产水平类型的单位归并在一个组，也不会把亩产相同或相近的单位分割在两个组，保证了样本的均匀分布。三则抽样调查单位完全根据组平均数决定，既不受调查单位面积大小影响，又不受第一组抽中单位的位置的制约。由最接近组平均亩产的样本单位组成的样本总体，分布较均匀，代表性较好，方差最小。

（二）各种抽样方法的比较。在随机性方面，上节已经比较过，这里只从别方面加以评述。

（ 1 ）在代表性方面，如果所要求的代表性相同，则划类选点可以抽较少的样本；

（ 2 ）在样本数目方面，如果限定了样本数目，那么划类选点可以选出最有代表性的样本；

（ 3 ）在推算方面，如果不但推算总体，还要推算各类型，也以划类选点切当；

（ 4 ）在重复抽样和不重复抽样方面，重复抽样就是抽样后将样本放回重抽；不重复抽样就是抽样后不放回样本。

重复抽样的抽样误差

$$Ux=\sqrt{\frac{O^2}{N}}\left(Up=\sqrt{\frac{p(1-p)}{N}}\right)$$

不重复抽样的抽样误差

$$Ux=\sqrt{\frac{O^2}{n}\left(1-\frac{n}{N}\right)}\quad\left(Up=\sqrt{\frac{p(1-p)}{N}\left(1-\frac{n}{N}\right)}\right)$$

按$\left(1-\frac{n}{N}\right)$永远小于1，所以重复抽样误差永远小于不重复抽样误差。

下面就划类选点、等距抽样和定点调查等加以较系统论述。

（三）划类选点是科学的抽样理论与群众的实践经验相结合的产物，是我国统计工作者通过实践创造出来的。划类选点又名类型抽样。上面已经提到，是将总体单位按其属性特征分为若干类型，以便于适应各地区之间的气候、雨量、地理、土址、耕作技术、遭受灾害、科学种田等的差异性大等特点。由于划分类型后，各类型中的各单位的共同性增大了，差异性缩小了，因而能够抽出更有代表性的样本。这种办法，在总体各单位之间差异程度较大，而在各群体中又有互相一致的情况下，尤其适用。各类型抽取的样本，可以按类型的大小比例抽，也可以不按比例，按代表性抽。这种抽样调查办法不但对县有代表性，对各社、队也有代表性，并能推算出各类型产量，能够得到群众和各级领导的支持，便于开展工作。

（四）等距抽样。分为两种，一种按无关标态排队，如按电话本上的花名册、按地理上位置、按一定图式，顺序等距抽取样本单位。这种抽样，由于总体各部位都能在一定程度上被包括在样本中，较能保证样本在总体中的均匀分布，可以比纯随机抽样抽取较少的样本获得较为可靠的代表性；还有一种是按有关标志排队。经过排队后，总体各单位按性质相同、数量接近的排在一起，等于划分了类型。可以大大缩小各类型集团内部的差异程度。因而可以缩小抽样误差。

（五）定点抽样。这个办法便于组织领导，统一使用调查力量，便于开展工作，连续观察农情；便于培训干部，可以得到基层和群众的支持；还可节省选点的人力。定点后，至少三年不变，三年后再每年轮换三分之一。固定点只定到生产队，生产队以下的地块、样本不能固定。固定会影响抽样的精确。

如果要在一套点内，进行两项以上的调查，并要求都具有足够的代表性的话，那就要按随机抽样或按无关标志（如地理位置或县的编号顺序）排队，等距抽选一套，然后针对各种不同的调查标识，进行排队和抽样。这样抽选的调查点的数目就得增多。

四、多层抽样与户抽样

（一）单层（阶段，下同）抽样与多层抽样。

单层抽样，就是从县直接抽样本。多层抽样，就是省抽县、县抽公社，公社抽大队或生产队，大队或生产队抽地块，地块抽样本实割实测。

在多层抽样中，必须充分注意样本在一个县内自然条件不同的地域中的分布。在一个县内的大队之间的自然条件的差异程度不如公社间的差异程度那么明显，那么突出。

公社一般是按地形区划分设的。如按山地、丘陵、平原、湖泊分。因而县抽大队不如抽公社，更能使样本单位在不同自然条件的地域中均匀分布。即使各个公社的产量构成不同，但是不论高产区、中产区、低产区，都有亩产较高的大队，也有亩产较低的大队。高产区公社，一般是亩产高而面积大的大队多，低产区的公社则刚刚相反，一般是亩产较低而面积较大的大队多。如果县抽公社，一是按亩产面积排队等距抽样，大队面积愈大的，抽中的机会愈大，对公社的代表性愈强；二是在每个类型内，每公社参加排队的大队数不同，按亩产面积排队等距抽样，参加排队的单位数愈多，被抽中的单位数也多，对公社的代表性也愈强。而这种抽法又没有违背随机原则。同时县抽的公社的代表性强，公社抽生产队（或大队）和生产队抽地块，其代表性也将同样增强了。

但是县抽公社的应该考虑公社力量的强弱问题。公社要及时在插秧后分品种掌握面积数字。收割时根据品种在各大队的分布情况，到品种比较集中的大队实测，按各大队各自的平均亩产，用亩数加叉推算公社产量。这种方法调查面广，工作简便，能反映各主要品种产量，但样本数目要大，才足以代表总体。

在多层抽样中还有一种是以行政区的上年总产量为总体，按调查点的亩产平均分类，等距抽样的。这样，总产量多的类、样本也比例地增多，反之减少。这办法有一定的代表性。毛弊一在决定总产量的两个因素：单产和面积的趋向不会相同，影响总产有大有小，二是平均总产不够或稍超过的就要四舍五入，使各类型内各单位抽中机会不均等，同时还缺乏代表性，只有平均产量等于类型产量的才有代表性。但这是少见的，还有，多层抽样要因地制宜。如新疆的沙湾县，11公社，62大队，222生产队。采用四层抽样：县、社、大队（或生产队）、地块、样本，由于第一，二层的社、队较少，总体单位少，抽出样本代表性差，第三、四层由于距离远，范围大，交通不便，工作困难。在这样具体条件下，就不如用三层抽样：县抽生产队，生产队抽地块，地块抽样本实割实测。如果第二层的面积少，地块少，也可以由生产队直接抽样本实割实测。沙湾县抽中的8个生产队中，有3个生产队就是用三层抽样的，抽样误差为－1.9%。新疆准备在全省都推广这三层抽样办法。

（二）按户抽样。由于农村实行责任制包产到户，过去以生产队集体经营的，大面积的田间科学管理的，品种布局比较集中的，生产比较平衡的，品种之间和田块之间的产量差异程度比较缩少的，因而对抽样调查所选田块和样本实割后的产量也比较接近平均数的；改变为品种繁多、布局分散、往近搭配、肥瘦兼有，田块另碎插花、产量不平衡，这就不能再简单地按过去办法抽样，必须有所改变。其中按户抽样就是一种办法。同时，这种改变，由于生产责任制是调整农业经济、调动农民生产积极性，发展生产力的，将长期坚持下去。因而按户调查这种改变也必然随之长期化。但是无论怎么改变，仍然必须紧紧抓住生产队这个环节。它现在和将来仍然是粮食生产在面积上的安排和调节者。在按户调查中，还可以利用生产队包产到户的“定产”平均亩产排队，等距抽取7～9户（户数在30户以下的，也可抽5～6户）。可以分田块抽样或按品种比例抽样实测推算平均户产量，然后按户数加叉推算队产量。这种做法其工作量比过去一般要大

一倍。

联产计酬责任制后，也还有仍旧以大队为单位，用不同方法，取得总产量的。湛江地区的做法就是这样，由生产队自报公议，推算全大队产量。海南地区也以大队为单位，按一般办法排队抽样的。

在公社抽选大队或生产队，以南海的平州和盐步两个公社为例，1983年早稻以大队为单位开展队队实割实测，依靠队干和植保员，进行目测估产，等距抽生产队（或选上、中、下平均产量的生产队），生产队每队选三个地段田（或分品种），共抽50～80样本，推算大队总产量。盐步公社12个大队实测结果亩产889.5和公社掌握的实产881斤相差0.37%。公社抽大队这一级，还适合于县社领导决策的要求。

1981年，贵州省贵阳市对花溪区调查，对全区大队采用1980年各大队上报统计年报的粮食亩产，从低到高排队，等距抽样，抽取七个大队，154户。由于没有按户材料，但“包干到户”时，是按人均田土好坏搭配的特点，所以确定按户排队人口累计等距抽选调查户。用按户平均产数449斤乘粮食作物播种面积22.4012万亩，得总产1058亿斤，与年报产量9338万斤接近。

五、抽样调查的推算、方案和样本数目

（一）抽样结果的推算。用抽样总体平均数×总体单位。这就是以抽样总体的点来推算总体的面。用抽样总体平均数来代表总体平均数。这种算法比较简便。是现在常用的办法。有差额U，应该加以估计。公式如下：

（抽样总体平均$\bar{X}\pm U$）$\times N=$ 总产量Y_N

（二）抽样调查方案

要把抽样调查方案搞好。只有具体设计一套完整的抽样调查方案，才有可能使抽样调查做到及时、准确、可靠、有效，而又节省人、财、物，在制订方案时，要全面考虑下列各方面因素，为取得抽样调查成功创造条件。一般的应包括下列几个主要内容：

第一，根据调查目的，确运调查对象和范围；

第二，确定对调查结果准确性的要求；

第三，确定必要的抽样数目n；

第四，设计有关指标的确切涵义和计算方法；

第五，选择适当的抽样调查方法和具体办法以及实地调查有关问题、调查结果的推算等。

（三）样本数目。在抽样调查方案中，确定抽样数目n是个关键问题。它取决于下列五个因素：

（1）差异程度O。差异程度大的要抽多，小的可以抽少；

（2）抽样误差极限（容许误差）△愈小的，抽样数目要愈多；

（3）把握程度要求t愈大的（置信区间的置信概率），抽样数目也愈多；

（4）总体单位数愈多的，抽样数目也应相对地增多。

（5）抽样方法不同，也影响抽样数目。这里O代表抽样离差，t代表可靠程度，△代表准确性要求的极限误差。

在重复抽样时，

$$n=\frac{t^2O^2}{\triangle^2}\left(或=\frac{t^2p(1p)}{\triangle^2}\right)。$$

不重复抽样时，

$$n=\frac{t^2O^2}{n}\left(1-\frac{n}{N}\right)=\frac{t^2O^2}{n}\times\frac{N-n}{N},$$

$$\triangle=t\sqrt{\frac{O^2}{n}\left(1-\frac{n}{N}\right)}$$

从而$\triangle^2nN=t^2O^2-t^2O^2n$

$\triangle^2nN+t^2O^2n=t^2O^2N$

$n(\triangle^2N+t^2O^2)=t^2O^2N$

$$\therefore\quad n=\frac{t^2O^2N}{\triangle^2N+t^2O^2}$$

如果是成数的n，则将O^2代以$p(1-p)$

$$\therefore\quad n=\frac{t^2p(1p)N}{\triangle^2N+t^2p(1-p)}$$

例：总体N＝100生产队

亩产的平均误差极限，绝对数△＝10斤，

均方差O＝8.18　t＝2.34

代O上式：$n=\frac{(2.34)^2\times100\times(8.16)^2}{100\times(10)^2\times(2.34)^2\times(8.18)^2}=3.52$

要得到准确的，可信程度为t＝1的，需要调查三到四个生产队。

如果是整群抽样，其误差取决于各群间平均数的差异程度，按不重复抽样：

以r代表整群数，$Or^2=\frac{\Sigma(\overline{X}r-Xr)^2}{nr}$

$$\overline{X}r=tr\sqrt{\frac{Or^2(Nr-nr)}{nr\quad Nr}}$$

Nr代表总群数，

nr代表抽样的群数，

Or^2代表各群间的方差。

（四）抽样调查必须大发展

对抽样调查的发展前途，做充分的估计，以适应“四化”的需要，是大有好处的，可以避免“急时抱佛脚”。

赵紫阳同志在农业上贯彻执行“依靠政策和科学，加快农业的发展”（赵《当前的经济形势和今后经济建设的方针》，人民出版社第16页）中提出：党中央国务院不但需要对各省（区）粮食产量需要及时准确地了解，而且对棉、油、麻、丝、茶、糖、菜、

烟、果、药、杂等各项经济作物和其他农付产品的普遍增长也需要及时了解。这就显得农产量抽样调查以及农村经济抽样调查在当前更加重要了。

由于抽样调查只要抽查有限的单位就能推断和取得全面资料，因而可以大大减少调查及整理资料的工作量，灵活简便，节省人、财、物，资料及时，取得事半功倍的效果，所以马列主义领导者都重视抽样调查。列宁在1921年给中央统计局的信中，建议搞抽样调查。列宁说："目前这种企业还不多（指集体企业，作者注），……以后这些企业为数很多的时候，则用抽样方法详细调查其中的五分之一或十分之一。"（列宁全集第33卷15页），周总理对抽样调查十分重视。1962年，鉴于当时粮食等按报表层层上报的数字不可靠也指示要搞抽样调查。1963年成立的全国农产量调查总队，就是根据周总理这个指示的。我们知道我国当前农业生产受自然因素影响很大；比较分散；国家、集体和个人三者之间还存在着一定的矛盾，出现了虚报或瞒报。有的省反映：1954年虚报了30亿斤粮食。同时，从收割到入库的统计上报时间相隔三个月，预报统计又有很大出入。据广西反映，1978年，估产增产几亿斤，实际减产几亿斤。1980年在全区的93%县市开展抽样调查后，数字就准确了。

这些事实和论据，说明了抽样调查必须有更大的发展，才能适应"四化"建设的需要。而抽样调查科学的发展，又是抽样调查大发展的前提与条件。近两年来，收到国内许多专家学者的论文，说明了我们统计工作者正在为"四化"建设而努力。谨在此表示感谢。

关于有关标识排队等距抽样方法的再探讨

广东省统计局　龚鉴尧

我的《关于有关标识排队等距抽样方法的探讨》论文发表后，在统计学界引起了很大的兴趣和反响。许多同志认为，《探讨》一文"提出了新的见解"，用有关标识排队等距抽样方法抽选出来的样本，"完全符合优样本原则"；这种方法的形成与发展，是我国统计工作者"长期应用富有成果的划类选典方法"和"吸收新发展的概率抽样法所结出的丰盛果实"。在一些新近出版的统计教材和著作，如陕西财经学院、暨南大学、中央财政金融学院、四川财经学院等合编的高等财院试用教材《统计原理与经济统计》，天津财经学院编写的《社会经济统计学原理》，宋元村、黄玉喜编著的《数理统计学》等，都肯定了有关标识排队等距抽样方法；中国人民大学、北京经济学院、天津财经学院和国家统计局等主编的《社会经济统计学原理电视讲座学习材料》中，还对这种方法作了详细论述。目前，这种方法在我国农产量调查、农民和职工家计调查、人口调查等

许多领域，正在被越来越多地推广应用。但是，也有一些同志不同意这种方法，认为它“不符合随机原则”，“缺乏充分科学根据”；有同志甚至写信给我说，这种方法是否科学，站得住，关系着抽样调查在我国的推广应用和农产量抽样调查等工作的前途，要求我进一步加以澄清。因此，我想确有进一步探讨的必要，以就正于关心抽样调查的同志们。

一、关于有关标识排队等距抽样方法的随机性问题

有的同志认为，有关标识排队等距抽样，在实践上可能是一种好办法，但在理论上似乎还缺乏充分的科学根据，它“不符合随机原则”，只能算做是一种“主观抽样法”。

我认为，这种对有关标识排队等距抽样方法的认识是不正确的。我在《抽样法浅说》和《探讨》论文中都曾经说过：“任何抽样方法的首要之点，就是必须消除一切产生偏误的根源。要做到这一点，最简单的可靠方法是：或者抽取样本的方法完全随机，或者在提高正确性而又不致给调查结果带来偏误的情况下限制随机。”有关标识排队等距抽样方法，在我看来，就是这样一种符合随机原则而在一定程度上限制随机的方法。事实上，在实际工作中，采用完全随机办法是很少的，因为完全随机样本的抽样误差比较大，需要抽取的样本单位数比较多，因而所需的人力和费用也比较大，不经济。而类型抽样、等距抽样，以及任何多阶段抽样方法，它们都使随机性在不同程度上受到了限制，但却比完全随机抽样方法更为有效：即或是在抽取同样大小的样本时具有较小的抽样误差，或是在具有同样大小的抽样误差的情况下需要的样本单位数要少；同时，也更便于组织实施。这在理论上和实践上都是已经被证明了的，不需再多作论证。

有些同志对抽样调查，片面地理解了随机性，认为随机性越大越好，因此把完全随机抽样看做是最理想的方法，其实是不正确的。因为现代统计科学和抽样技术的发展，越来越多地并不是采用简单随机抽样法，而是提倡尽可能地搜集辅助信息和利用已知信息，采用那些能提高正确性而又不致给调查结果带来偏误，限制随机而又不损害随机原则，同时更加便于组织和经济有效的方法。

有同志说，无关标识排队等距抽样是限制随机的办法，但在抽取样本时，可以任意一点为起点，虽然限制了随机，却无损于随机原则，这种方法是可以赞成的；而有关标识排队等距抽样，从组中点抽取样本单位，一旦抽样数目确定后，整个样本也就确定了，只能抽出一套样本，其它单位再没有被抽中的机会，这样大的限制，是否还符合随机原则？

我们说，任何等距抽样方法，包括无关标识排队和有关标识排队在内，只要抽样数目和起点一经确定，就只能有一套样本。问题在于，无关标识排队，是以任意点为起点，在抽样起点确定以前，它有可能抽中距离内的任何一套样本，比如说，总体单位数为$N=100$，抽样数目为$n=10$，抽样距离$=\frac{N}{n}=\frac{100}{10}=10$，则有10套可能样本；而有关标识排队，我们通常使用的办法，是从组中点开始，只能抽取一套样本。从这个意义上

说，有关标识排队等距抽样的确比无关标识排队等距抽样的随机性小。但这只能说它的随机性更加受到限制，而不是“不符合随机原则”。

应当说明的一点是：有关标识排队等距抽样，由于总体单位是按高低顺序排列的，通常对抽样调查单位的具体做法，都是在计算出抽样距离后，以距离的中间一个单位——组中点为起点，这样抽出来的样本，应当说是很理想的样本。而从其它起点抽取的样本，往往会出现偏高偏低的情况。这从直观上就可以看到它的好处，简便易行，比较接近我国习惯上所用的划类选点的做法，易于为调查人员和广大基层干部、群众所接受。因此，我们过去在谈到有关标识排队抽样时，一般都只谈到这种办法。

采用组中点为起点，不仅在实践上是有效的，而且在理论上也是可行的。潘孝瑞、许刘俊等同志的文章，对此都已作过充分的论证。

由于我们在五十年代就对有关标识排队等距抽样方法进行过研究和试点调查，六十年代初期采用这种方法进行了全国规模的农产量抽样调查，当时我曾经编写过一本《抽样调查方法介绍》的小册子，着重介绍这种方法，广为流传；加上长期闭关锁国，对国外抽样技术的新发展了解很少，也没有看到过有谈到这种方法的。因此，我和许多同志都认为：这种方法是我国独创的；国际统计学界似乎没有人论证过，可能对这种方法是持否定态度的。直到去年，北京农业大学刘宗鹤教授寄了一本W.G柯克恋写的《抽样方法》（英文本，1977年第三版）给我，我才注意到，在柯克恋的著作中，也谈到了这种方法，并且持肯定的态度。不过，他不是叫做“有关标识排队”，而是叫做“直线趋势总体”；并认为当总体是呈这种“直线趋势”排列时，采用中心位置样本比随机位置样本会更为正确，而且要比普通等距抽样（即无关标识排队）优越。可见，我的关于有关标识排队等距抽样方法的观点，在国际统计学界，也并非孤立的。所不同的是，柯克恋是把这种“直线趋势总体”作为调查中可能遇到的一种特殊现象来对待，而我们则是自觉地把被研究总体改造成为“直线趋势总体”。不论这种方法是否我们创始的，但是，我深信就世界范围来说，对这种方法实践应用最多，最有经验，最有发言权的，仍然是我们国家。这是适合我国具体情况的一种行之有效的抽样方法。

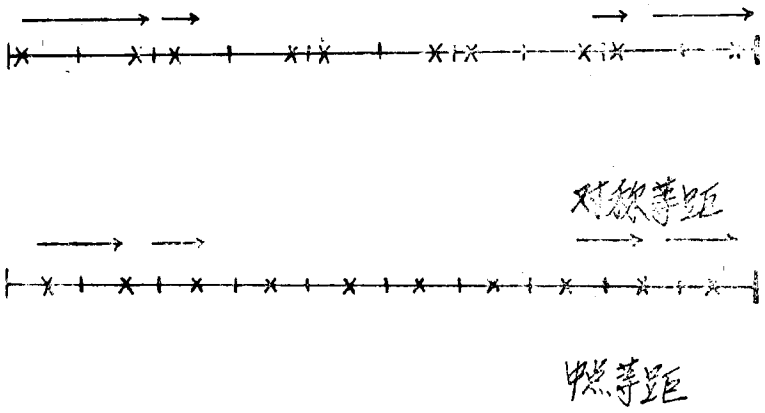
二、改进或增强有关标识排队等距抽样方法随机性的办法

对有关标识排队等距抽样方法，能否不用组中点样本，而用任意起点随机样本，以改进或增强它的随机性呢？我认为，是可以的。在去年八月中国统计学会和广东省统计学会联合召开的“抽样调查在农业方面的应用”科学讨论会上，我和刘宗鹤同志等，都曾经提到过一些别的方法，譬如说：采用对称等距法，就可以和无关标识排队一样，以任意一点为起点抽取样本，并获得良好的无偏误推算数。例如，有100个综合商店，打算采用有关标识排队等距抽样方法，抽取其中10个商店来调查其营业状况，包括购、销、盈利等情况。根据有关材料，已知其投资额大小的排队次序，在求出抽样距离

$$\left(R=\frac{N}{n}=\frac{100}{10}=10\right)$$

后，即可采用对称等距方法，从第一个等距中的任意一点开始，对称地等距抽出样本单位。假定其起点为第2号商店，即第一个抽中单位，第二个抽中单位应为第二个距离中的倒数第二个单位，即第19号商店，第三个抽中单位为22号商店，第四个单位为第39号商店，以下抽中的商店为第42号，第59号，第62号，第79号，第82号，第99号。如以第7号为起点，则以下抽中的商店为第14号，第27号，第34号，……依此类推。

对称等距与中点等距的抽点方法示意如下：



中点等距的每个样本单位的距离都是相等的；而对称等距实际是两两等距。

再一种办法是采用顺逆排列交替等距抽样法。仍用上例，已知N＝100，n＝10，R＝10，则可排列如下式：

划分等距数	分样本									
	1	2	3	4	5	6	7	8	9	10
1	1	2	3	4	5	6	7	8	9	10
2	20	19	18	17	16	15	14	13	12	11
3	21	22	23	24	25	26	27	28	29	30
4	40	39	38	37	36	35	34	33	32	31
5	41	42	43	44	45	46	47	48	49	50
6	60	59	58	57	56	55	54	53	52	51
7	61	62	63	64	65	66	67	68	69	70
8	80	79	78	77	76	75	74	73	72	71
9	81	82	83	84	85	86	87	88	89	90
10	100	99	98	97	96	95	94	93	92	91

从上表可以看出，用顺逆排列交替等距抽样法，以第一个距离中的任一点为起点，每个样本单位的距离完全相等，抽出的10套分样本，实际上同前面用对称等距法所抽出

的结果，是完全一致的。用这种方法，如果根据已知辅助信息，对每个单位不仅有次序上的了解，而且还有了数量上的了解，例如，在前面的例子中，已经知道每个综合商店的投资额，则这种排队方法的好处，可以更加明显地看出来。下面是根据各商店已知投资额按顺逆交替法排队的各套样本的情况：

划分等距数	按投资额（万元）顺逆交替排列的分样本									
	1	2	3	4	5	6	7	8	9	10
1	5.50	5.50	5.75	5.81	5.88	5.89	6.00	6.04	6.07	6.09
2	6.49	6.47	6.40	6.40	6.40	6.38	6.32	6.30	6.11	6.11
3	6.54	6.55	6.55	6.62	6.63	6.70	6.73	6.73	6.77	6.79
4	7.09	7.09	7.02	7.01	6.92	6.92	6.92	6.91	6.84	6.80
5	7.10	7.10	7.10	7.14	7.15	7.15	7.15	7.20	7.21	7.24
6	7.40	7.40	7.38	7.35	7.32	7.30	7.25	7.25	7.25	7.24
7	7.45	7.50	7.53	7.59	7.60	7.62	7.62	7.70	7.70	7.71
8	8.00	7.98	7.98	7.97	7.90	7.85	7.82	7.80	7.80	7.74
9	8.04	8.05	8.09	8.09	8.10	8.19	8.20	8.28	8.28	8.28
10	8.62	8.60	8.50	8.50	8.47	8.42	8.36	8.36	8.34	8.30
小　计	72.23	72.24	72.30	72.48	72.37	72.42	72.35	72.57	72.37	72.40
平　均	7.22	7.22	7.23	7.25	7.24	7.24	7.23	7.26	7.24	7.24

以上两种有关标识排队等距抽样的办法，和无关排识排队一样，都可以任意一点为起点，比中点等距更具有随机性，但抽样数目最好取偶数，否则就不能完全对称。

三、有关标识排队等距抽样和类型抽样的相同和不同之处

有关标识排队等距抽样和类型抽样有什么相同和不同的地方？为什么说这种方法是把等距抽样和类型抽样结合起来，兼有二者的优点，为什么不把它干脆归入类型抽样呢？

首先，让我们来看看有关标识排队等距抽样方法和类型抽样方法有什么相同之处。我认为，它们的相同之处，最基本的有以下三点：

（一）划分类型和有关标识排队，都需要事先搜集了解总体的有关资料，利用总体的已知信息和辅助信息。

（二）通过对总体的划类和按有关标识次序排队，都可以大大缩小总体的差异程度，增强样本对总体的代表性，提高调查结果的准确程度。

（三）类型抽样和有关标识排队等距抽样，都是属于限制随机性的抽样方法。

有关标识排队等距抽样方法和类型抽样方法的不同之处是：

（一）划分类型与抽样距离做法不一样。类型抽样通常是将总体单位按其属性特征分为若干类型。然后在各类型中抽取样本单位。例如，在进行工业企业抽样调查时，将