粮农组织

经济和社会发展
文　集

统计数据的质量

作者；S.S.扎科维奇

（联合国粮农组织统计处）

中国农业科学院科技文献信息中心
根据其同
联合国粮食及农业组织的协议出版

中　国
农业科技出版社
北京 1988

联 合 国
粮食及农业组织

统计数据的质量

责任编辑　段道怀

中国农业科技出版社出版（北京海淀区白石桥路30号）

新华书店北京发行所发行　各地新华书店经售

北京海淀区东华印刷厂印刷

开本：787×1092毫米1/16　印张：15　字数380千字

1988年12月第一版　1988年12月第一次印刷

印数：1-3000册　定价：7.60元

ISBN 7-80026-115-8/S·84

目　录

第一章　几个基本的概念

1.1 误差的定义

在进行一次调查之前，必须先确定下面这些因素：概念和定义，收集数据的方法，在表述回答时所使用的单位，制表程序，调查计划，问题的提法，等等。我们把所有这些因素统称为“采用的工作法”。因此，采用的工作法指的是在调查中需要收集哪些数据以及如何收集这些数据，等等。

采用的工作法是按照调查的目标制定的。由于工作法提出了一套固定的概念、定义和程序以及调查所包括的一些做法，因此把它具体明确下来之后就能够判断所做的工作是否符合要求。当然，有时候只能在理论上做到这一点。

根据采用的工作法的概念，可以给真值的概念下一个定义。真值就是在一次具体的调查工作中正确实施工作法所应当获得的结果。真值是进行具体的调查工作所获得的理想的结果；如果绝对按照采用的工作法进行这种调查工作，这种结果是一定能获得的。

真值有几种。第一种是某一给定的总体单位特性的单项真值。应用所说的工作法获得给定单位的特性值，结果就产生了单项真值。如果在一次人口普查中要了解各户户主在最近一次生日的年龄，那么，不管这些户主是否知道这项真值，也不管他们在普查中给了什么答复，这项真值实际上就是这些户主的周岁数。某户以公顷为单位的土地总面积的真值，就是这户的各块土地面积按其最接近的整数进行舍入取整的真值总和。由此可见，采用的工作法固定下来之后，真值就成为一个固定的量了。

在某些情况下，洞察真值的含义并不容易。例如，种植意图之类的“意图”调查就是这样。然而，这方面的困难以及确定真值的实际困难，都不会妨碍这个概念的使用，因为不用真值这个概念便难说明统计误差的实质。

除了单项真值以外，我们还将谈到总数、平均数、比率、相关系数的真值和其它的统计量。这些概念的含义是很清楚的。

为了确定总体总数的真值，我们用符号X_i来表示总体的第i个单位某特性的真值。假设这个总体的单位总数等于N，那么，就这个特性来说，总体总数的真值就是：

$$X=\sum_{i}^{N}X_i$$

其它统计量的真值定义是显然的。

显然，在实际调查工作中未必总能了解到所有单位的单项直值。实际了解到的结果称为“调查值”。如果总体第i个单位同一特性真值仍为前面所说的X_i，则调查值要用符号Z_i表示。我们按照真值的定义类推，便可区分单项调查值和各种统计尺度的调查值。因此总体总数的调查值显然为：

$$Z=\sum_{i}^{N}Z_i \qquad (1.2)$$

至于其它的统计尺度的调查值，我们可用变数Z 和统计理论上人所共知的公式来确定。

通过真值和调查值，我们现在定义误差为调查值与相应的真值之差。因此，第 i 个单位的单项误差为[1]：

$$di=Z_i-X_i \tag{1.3}$$

单项误差可能是正的也可能是负的。如果调查值与相应的真值相等，即$Z_i=X_i$ 或 $di=0$，我们认为Z_i是准确的。另一方面，如果$di\neq 0$，则Z_i便是不准确的。

下面还有一些概念。根据公式（1.3），我们得出

$$Z_i=X_i+d_i \tag{1.4}$$

和

$$\sum_i^N Z_i=\sum_i^N X_i+\sum_i^N d_i$$

或

$$Z=X+D \tag{1.5}$$

数量D叫做“偏差”。显然，如果 $D=0$，那么某特性总体总数的调查值就与相应的真值相等。在这种情况下，Z就是准确的或“无偏的”。反之，如果$D\neq 0$，则Z 就是“有偏的”。

应用公式（1.3）或（1.4），可以很容易地确定其它统计尺度的偏差。

从实用的观点来看，应当高度重视单项误差的频率分布。如果正负误差地分布在零附近，对总数和平均数的估计就是无偏的。然而，在许多情况下，有些误差结构类型是：正误差或负误差居于支配地位。我们把这种情况叫做系统误差。根据可能有系统的误差的数据得出的总数和平均数通常是有偏差的。因此，偏差是所有误差的净结果。

从公式（1.5）中还可以清楚地看出偏差是有符号的。在$Z>X$的情况下，偏差是正的。而在$Z<X$时，偏差是负的。当偏差是正的时，我们说调查总数Z夸大或过高地估计了总数的真值。如果偏差是负的，那就是缩小或低估了真值。

偏差的大小及其符号的大小并非在所有的研究中都同样重要。数据的使用者主要关心偏差的大小。然而，在某些误差分析中，偏差的正负号也许极为重要。

上述误差定义仅指简单的误差，在本书以后的章节中，将较广义地使用误差的概念。因此，如果实际采用的程序与采用的工作法规定不一致，我们也把它称为误差，虽然这种误差的结果不是以单项误差的形式表现出来的。例如，如果未按规定进行抽样，那么我们就说它是抽样中的误差。抽样中的单位数据可能是准确的，但是，根据抽样得出的对总数和其它统计尺度的估计可能容易产生偏差。这类误差的结果，将成为最后调查结果中的偏差。这就是程序也可称为有偏选择程序的原因。

在1.3节中，将按照不同的标准对误差进行分类，并给业已分类的各类误差定一个单独名称。这将有助于了解在某一具体事例中出现了何种误差。

1.2 误差出现在哪里？

在准备一次统计调查时，第一步通常是制订调查计划和确定将使用的基本概念和定义。在准备调查的第一阶段，可能容易发生后果严重的误差：一些后来发现对正确了解调查对象

(1) 关于误差的定义和使用公式(1.3)的理由，请详见第3章。

很重要的特性可能被遗忘；一些定义下得不适当，结果未把一些单位包括在调查范围之内；有些概念的定义容易引起误解，等等。

当然，在以后的工作阶段也有发生误差的其它许多可能性。例如，在拟定调查表时，一些问题的提法可能引起误解；调查表的内容格式可能难以填写；调查表太大，可能使填表人对问题的回答写错地方。

对实地工作的指示是发生误差的另一个重要原因。如果不把任务充分解释清楚，工作人员便可能各行其是。如果解释过长则可能使他们感到迷惑不解，结果还是各人按照各人的想法行动。如果指示意思不明确，模棱两可，就很可能引起各种误解。

在制图方面，计数地区或其它地区单位的划分可能有漏洞，因此有些总体单位被遗漏，有些总体单位处于界线上，根本不清楚它们属于哪一边。如果不用图而用文字来说明计数地区的界线，也可能出现同样的问题。

至于计数员，有些人选不合适，有一些人可能没经过适当的培训。这两种情况都会导致各种误差。由于主观和客观上的一些原因，计数员可能遗漏一些单位；他们也可能把界线上的一些单位计算两次或多次；他们打电话询问某户时，如果家里没人，就可能不再打电话去问了，这样有些数据就遗漏了；在有些情况下，他们可能在工作中掺杂一些自己的想法和意见；有时他们使人们回答问题的角度，与采用的工作法规定的角度完全不同。他们的举止可能会造成一种紧张的气氛，使得人们拒绝回答问题；如果计数员是领计件工资的话，他们可能会不适当地加快工作速度，从而忽视工作质量；等等。

造成困难的另一个原因在于回答问题的人有自己的想法，而他们的想法可能在许多方面与调查者的意图不同。尽管通常采取了预防措施，但还是不能完全消除回答问题者个性的影响。他们有时害羞，有时害怕，有时想提高个人的声誉，因而更改其应给出的回答。

进行调查时的总的环境也可能促使误差发生。众所周知，如果人们认为某项调查的数据将用于非统计的目的，则在这样的气氛中进行的调查便可能出现严重的差错。某些回答问题的人认为，他们提供不准确的情况有利于维护他们的尊严和利益。

在开始处理数据时，还可能发生许多新的误差。在编辑时，必须对几百份调查表中的许多问题进行质量检验。在这样大量的工作中，即使是受过最好训练的工作人员也会出错。一部分有差错的数据就是在最好的自动化机器上产生的。在编码、打孔等其它数据处理阶段，同工业中的大规模生产一样，也会发生差错。

由此可见，任何统计工作都难免发生误差。误差是普遍存在的，尤如身影跟随人一样，只要你一做统计工作，就免不了要出错。但是，不能因此感到悲观。一般地说，误差出现的次数有时比预计的要多，因此误差的影响也很可能比原先想象的要大。然而，现有的资料也表明，可以采取一些措施来控制误差出现的次数。因此有理由对此抱乐观态度。

和在其它类似的领域中一样，要采取有效的减少误差的措施，就必须对统计工作中遇到的各类误差有一个透彻的了解。统计人员应当知道误差是在什么情况下产生的，它们产生的原因是什么，它们对各种统计尺度会产生些什么影响，可以采用什么手段和方法来提高数据质量，等等。只有全面了解了每个误差问题的各个方面，才能着手制定有效地减少误差的措施。

1.3 误差的分类[2]

在我们试图将本书所要谈的误差进行分类之前，先谈一下调查统计准备工作开始时所发生的误差，是有益的。一种情况是调查计划不够具体，例如漏掉一些重要的特征数，结果是收集到的数据不能提供所需要的情况。如果采用不适当的制表程序等，也会出现同样的情况。虽然这些误差可能是很严重的，但是它们基本上不属于统计理论的范畴。制定统计调查计划及确定所要调查的总体的定义，是统计数据使用者和统计人员的共同责任。因此，本书不讨论这类误差。由于同样原因，本书也不讨论印刷差错，因为保证印刷质量是印刷厂的责任。本书所讨论的完全是应由统计人员负责的误差。换句话说，本书将只论述统计人员应当研究和评价的那些误差。

广义的误差可以分为三大类：

（1）由于准备工作不充分而造成的误差；

（2）数据收集阶段发生的误差；

（3）数据处理中发生的误差。

这个分类方法有一些不足之处。例如，第一类和第二类有些重复。因此，如果实地工作人员所了解到的有关某人的收入情况不准确的话，那可能是由于概念和定义不明确、指示不全面和采访方法不对头等缘故。换句话说，很难确定这种误差是属于第一类还是属于第二类，或者同时属于这两类。然而，从实际出发，最好是把第一类误差单独作为一类。这样的分类可以提醒统计人员要细心，因为在调查工作的准备阶段就可能发生误差的。

第一类误差又分为有偏程序和有偏工具。如果一种程序的反复使用，会使调查结果出现偏差，这样的程序便称为有偏程序。根据个人判断的抽象就是一例。我们将分别讨论三种有偏程序：测量、抽象和估计程序。

如果一种工具，即使按照采用的工作法正确使用，也使结果出现偏差的话，这种工具就叫做有偏工具。在统计工作中可能出现的各种有偏工具中，我们在这里将对随机数字表，调查表，登记表和指示进行讨论。测量仪器也可能是有偏工具。使用长度不精确的绳子进行测量，会导致对面积或长度的估计出现偏差。对于测量仪器，本书将不加讨论，因为它们不是专门的统计工具。

第二类误差分为清单误差、遗漏数据和回答误差或观测误差。

清单误差在数字性普查和抽样调查中都会发生。在数字性普查中，数字经过整理，汇总成一些总体单位清单。如果在清单中，有些单位被遗漏而有些单位又出现两次或多次，这样发生的误差就称为清单误差。漏掉的单位称为遗漏；把某些单位列两次或多次这样的误差常称为重复。另一种属于这一类的典型误差是列举不存在的单位。

清单误差又称为包括范围误差或计数不完整误差。由清单误差造成的调查估计数偏差，称为清单偏差。

有一种特别的误差，我们称它为分类错误。这类误差在人口普查中常常遇到，因为在人

(2) 关于各类误差的详细论述，请参阅：W.E.戴明《论调查中出现的误差》(载1944年《美国社会学评论》第九卷第359—369页)；W.E.戴明《关于抽样的理论》(1950年纽约约翰·威利公司出版)；M.H.汉森、W.N.赫维茨和W.G.马道合著《抽样调查的方法和理论》(1953年纽约约翰·威利公司出版)；L.基希《抽样调查》(1965年纽约约翰·威利公司出版)；P.C.马哈拉诺比斯《印度统计研究所关于统计抽样的新实验》(1946年《皇家统计学会报》第109卷第326—378页)；F·耶茨《普查和调查的抽样法》(1960年伦敦查理士·格里劳公司第三版)。

口普查中，不仅要列出人口，而且还要分门别类地列，如常住居民、暂时在外居民和暂住居民。把一名常住居民列为暂住居民，这就是分类错误。既然常住居民和暂时在外居民都是某一地区的常住人口或居民人数。所以分类错误的严重性是显而易见的。

当然，分类错误主要是由于有关单位的数据不准确造成的。

普查中的另一类清单误差是把一些单位列在错误的计数地区。这种错误在于，有关单位是列入了，但是没有按照普查指示列入应当列入的地区，如果把一个应当计入第一计数区的人计入第二计数区，对第一计数区来说这是一个遗漏，对第二计数区来说就称为错列。

在抽样调查中，常常把新的单位清单作为以后抽样的一个框框。例如，在一次农田抽样调查中，在初期选择阶段可能选择若干公社或村庄作为样本。然后指示计数员就初期选择的单位中的所有农户编制清单，并从中挑选百分之十的农户进行的采访。在这样的清单中出现的上述误差（即遗漏或重复），我们也称之为清单误差。

遗漏数据是主要在抽样调查中遇到的一种特殊误差。如果由于某些原因而得不到有关抽样中的某个单位的材料，则遗漏数据的情况就会产生。例如，粮食消费调查可能以户为抽样单位。在调查的时候，可能有些家中没有人。这样，这些单位的数据就会遗漏。另一个常见的例子发生在以农作物收获量为依据的单产调查中。当计数人员来了解收获量时，某些列为抽样的农田可能已经收获完了。因此有关该农田的情况就会遗漏。

拒绝回答是产生遗漏数据的一种特殊情况。当通过采访调查或书信调查与有关人联系时。他可能不愿为调查提供回答。

在统计文献中，遗漏数据通常是放在叙述拒绝回答或样本不全的章节中论述的。“拒绝回答”一语，适用于采访调查，尽管它不能完全表达观测调查的情况。

概括地说，回答误差或观测误差指的是单项真值与相应的调查之间的差异，而不涉及造成这一差异的原因。如果某一农户回答说他们的土地总面积是八公顷，而按照地籍数据计算他们的土地总面积却是七公顷。那么，这一户的回答中就有回答误差。如果计数人员所计算的某一时间内的进港船只数与进港的实际船只数不符，我们就把这种情况称为观测误差。

很显然，回答误差可能是正的，也可能是负的。如果系统地出现回答误差（如使用有偏差的衡器称重所得出的结果等），那么计算出来的数量（如总数）也会受回答偏差的影响。

数据处理过程中发生的误差也可分成若干类，例如编辑、编码、打孔、制表等误差。显然，分类的数目可能因处理数据所使用的技术和设备不同而不同。这里所举的几类，是使用标准的机械制表设备时可能出现的情况。

从以上的论述中便可以清楚地了解到数据处理过程中的误差的含义。

1.4 误差的相对性

统计误差有相对的意义。这可能是它们最重要的特征。在一次采访调查中得到一个回答或在一次观测调查中记录下来的一个数字是否是一个误差，只有从采用的工作法来判断。例如，如果按照地籍测量某户的农田总面积是五点四公顷，而按照土地普查却是五公顷，那么即使指示要求回答精确到第一位小数，普查的数据也仍然被认为是不精确的。但是，如果指示确定舍去小数，那就没有误差了。同样，如果在一次人口普查中一个人说他生于1933年4月20日，而按照他的出生证实际上却生于1931年5月18日（假设这个日期是正确的），那么从他的出生日期和周岁数来看，回答都是不准确的。然而，对制表来说年龄数据是以五年为

单位的年龄组分类，这个回答就是准确的，因为根据他的回答，他按照出生证也是属于这一组的。

记住统计误差的这一特性是有用的。为了减少误差数，有时候也可以对采用的工作法进行修改。

1.5 有偏的估计程序

有些定义需要扩大以适应在抽样调查中遇到的具体情况。

首先假设一个n个单位的样本是取自N个总体单位的简单重复随机样本。还假设为n个单位中的每个单位都提供了要了解的特性的真值。然后根据样本数据，用公式$\bar{X}=\frac{1}{n}\sum_{i}^{n}X_i$来估算总体平均数$\bar{X}$的真值。

由于仅使用了n个单位的样本，估计数$\bar{X}$通常是与$\bar{X}$不同的。事实上，即使从同一个总体中选择所有可能有的n个单位所有样本，并从每个样本中计算出$\bar{X}$，估计数$\bar{X}$将以正态分布的形式围绕$\bar{X}$而变化。这种分布的一个重要特点是：所有可能的估计平均数的算术平均数等于$\bar{X}$。另一种表达这个意思的方法是：估计数$\bar{X}$的期望值等于$\bar{X}$即$E_{\bar{x}}=\bar{X}$。在这个特定的情况下，我们说$\bar{x}$是$\bar{X}$的无偏估计。事实上不管使用什么统计尺度，只要我们计算出数量U的样本估计数u，我们就认为：Eu=U，u就是U的无偏估计数。

从上述情况已很清楚地看出无偏估计数的优点。平均数的每个无偏估计可能与$\bar{x}$有或多或少的差异。但是，我们知道$\bar{x}$的平均数等于$\bar{X}$。$\bar{x}$在$\bar{X}$附近变化的范围用“标准误差”来度量。符号是$S_{\bar{x}}$。标准误差的平方$\sigma_{\bar{x}}^2$称为估计平均数的方差，并确定为$\sigma_{\bar{x}}^2=E(\bar{x}-\bar{x})^2$在重复简单随机抽样情况下，我们已知基本结果$\sigma_{\bar{x}}^2=\sigma_{\bar{x}}^2/n$，这表明估计数$\bar{x}$可能的变化范围首先取决于总体中x值的变化，其次取决于样本的大小。如果我们从同一总体中抽样，那么样本越大，估计数$\bar{x}$平均来说就越接近$\bar{X}$。

估计数$\bar{x}$在$\bar{X}$附近的平均变异数值也可用“精确度”这个词来表示。而精确度则用标准误差来衡量。标准误差越小，估计数$\bar{x}$越精确，反过来也是一样。

在其它统计尺度方面也使用这些术语。例如，假如在重复简单随机抽样中用公式

$$S_x^2=\frac{1}{n-1}\sum_{i}^{n}(x_i-\bar{x})^2$$

来估算总体方差σ_x^2，那么数量σ_x^2称为σ_x^2的无偏估计数，因为$E\sigma_x^2=\sigma_x^2$。估计数σ_x^2也可能受抽样变化的影响，因此它有自己的标准误差和精确度。

但是，并不是所有的样本估计数都属于无偏估计数的范畴。下面举例子加以说明。为了估计家畜头数，假设从M个地区的总体中抽取m个计数地区样本。在选定的m个计数地区中采访所有的农户，并获得每户家畜头数的准确数据。在这样的情况下，可用下面的公式来估算平均每户的家畜头数$\bar{\bar{x}}$：

$$\bar{\bar{x}}=\frac{1}{m}\sum_{i}^{m}\frac{1}{N_i}\sum_{j}^{N_i}x_{ij}=\frac{1}{m}\sum_{j}^{m}\bar{X}_i \qquad (1.6)$$

式中的符号的含义如下：

x_{ij}=i计数地区j户的家畜头数，

N_i=i计数地区的户数。

$\bar{\bar{X}}_i$ =i计算地区平均每户的家畜头数。

显然，公式（1.6）的结果是计数地区平均数的一个简单算术平均数。

公式（1.6）是用来估算$\bar{\bar{X}}$的。$\bar{\bar{X}}$的值为：

$$\left.\begin{aligned}\bar{\bar{X}} &= \frac{1}{N}\sum_{i}^{M}\sum_{j}^{N_i} x_{ij} \\ &= \frac{1}{M}\sum_{i}^{M}\frac{N_i}{\bar{N}}\bar{\bar{X}}_i\end{aligned}\right| \qquad (1.7)$$

式中的N代表总体的总户数，$\bar{N}=N/M$。

如果$E\bar{\bar{x}}$等于$\bar{\bar{x}}$，估计数$\bar{\bar{X}}$则被认为是无偏的。事实上

$$E\bar{\bar{x}} = \frac{1}{M}\sum_{i}^{M}\bar{\bar{X}}_i \qquad (1.8)$$

显然，公式（1.8）确定的$E\bar{\bar{x}}$不等于公式（1.7）确定的$\bar{\bar{X}}$。公式（1.8）是地区平均每户的简单算术平均数，而公式（1.7）是地区平均数的加权算术平均数。按照前面使用的术语，我们说$\bar{\bar{x}}$是$\bar{\bar{X}}$的有偏估计数。

如果我们用符号$\bar{\bar{D}}$表示每户数字相应的偏差大小，则：

$$\bar{\bar{D}} = E\bar{\bar{x}} - \bar{\bar{X}} \qquad (1.9)$$

在这个特定的情况下，$\bar{\bar{D}}$的偏差程度是：

$$\left.\begin{aligned}\bar{\bar{D}} &= \frac{1}{M}\sum_{i}^{M}\bar{\bar{X}}_i - \frac{1}{M}\sum_{i}^{M}\frac{N_i}{\bar{N}}\bar{\bar{X}}_i \\ &= \frac{1}{M}\sum_{i}^{M}\bar{\bar{X}}_i\left(1-\frac{N_i}{\bar{N}}\right)\end{aligned}\right| \qquad (1.10)$$

公式（1.10）表明如果各计数地区的户数和平均每户家畜头数不同，公式（1.6）中的估计数是有偏的。

我们可以看出，在抽样调查中，即使每个单项单位的数据是准确的，得出的估计数仍可能有偏差。

关于公式（1.6）的估计数，还应当指出，按照前面的假设，还可以使用无偏估计数$\bar{\bar{x}}'$。列式如下：

$$\bar{\bar{x}}' = \frac{1}{m\bar{N}}\sum_{i}^{m}\sum_{i}^{N_i} x_{ij} \qquad (1.11)$$

$E\bar{\bar{x}}$等于$\bar{\bar{X}}$。如果对选择的计数地区进行二次抽样并从第i选择计数地区抽取n_i个子样本单位，那么$\bar{\bar{X}}$的另一个估计数是：

$$\bar{\bar{x}}'' = \frac{1}{m\bar{N}}\sum_{i}^{m}\frac{N_i}{n_i}\sum_{i}^{n_i} x_{ij} \qquad (1.12)$$

这个估计数也是无偏的。

由此可见，可以用无偏估计数来消除上述有偏估计程序引起的偏差[3]。

1.6 有偏估计数对无偏估计数

在前一节中曾经假设一个简单随机样本的n个单位现有数据都是准确的。在本节中为了使用普通的例子，可允许有些单位的数据不准确。利用前面的符号和公式（1.4），平均数的调查值$\bar{Z}$估计为：

$$\left.\begin{aligned} E\bar{z} &= \bar{Z} \\ &= \bar{X} + \bar{D} \end{aligned}\right| \qquad (1.13)$$

和

$$\bar{D} = E\bar{z} - \bar{X}$$

$\bar{Z}$的方差当然是：

$$\left.\begin{aligned} \sigma_{\bar{z}}^2 &= E(\bar{z} - E\bar{z})^2 \\ &= \frac{1}{n}(\sigma_x^2 + \sigma_d^2 + \rho_{xd}\sigma_x\sigma_d) \end{aligned}\right| \qquad (1.14)$$

公式（1.14）在误差论中具有极大的重要性。从$\sigma_{\bar{z}}^2$的定义中可以看出，这个公式能够测定估计平均数$\bar{Z}$在其期望值$\bar{Z}$附近的变化情况。如果偏差$\bar{D}$大的话，则图1所示的情况在实践中就可能出现。

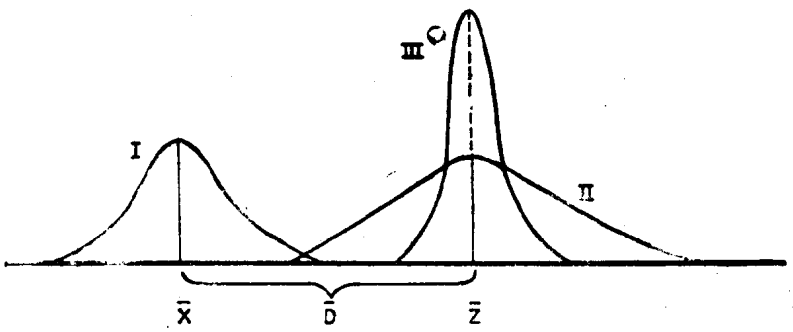

图1 在数据不准确时有偏估计平均数与无偏估计平均数之间可能出现的关系

在上图中，曲线Ⅰ表示单项真正值在总体的分布。这个分布的方差是σ_x^2。曲线Ⅱ表示调查值与方差σ_x^2的相应分布。这样，如果从曲线Ⅱ的分布中选择n个单位的样本，则估计数$\bar{z}$将按曲线Ⅲ的方式分布在$\bar{Z}$的附近。公式（1.14）中所确定的方差$\sigma_{\bar{z}}^2$是曲线Ⅲ中估计数$\bar{z}$分布在$\bar{Z}$附近的变异数。方差$\sigma_{\bar{z}}^2$表示出估计数$\bar{z}$（对$\bar{Z}$）的精确度，但它不表示调查的基本目的，即我们的估计数$\bar{z}$平均与$\bar{X}$相差多少（我们想知道的是这个数）。根据$\sigma_{\bar{z}}^2$的数值无法断定$\bar{X}$的位置。

事实上，我们是很想知道$\bar{z}$在$\bar{X}$周围变化的情况的。换句话说，我们需要一个$\bar{z}$在$\bar{X}$周围的变异数。这个新的数将称为均方差，用符号ζ^2来表示。

(3) 关于有偏估计数置信限度的概率论，必须非常小心地使用正态分布论。探讨这个问题的文献有：W.G.科克伦：《抽样技术》（纽约约翰·威廉公司出版，1963年第二版）；M.H.汉森、W.N.赫维茨和W.G.马道合著：《抽样调查的方法和理论》（1953年纽约约翰·威廉利公司出版）；基希：《抽样调查》（1965年纽约约翰·威廉公司出版）。

ζ^2的定义公式如下：

$$\left.\begin{aligned}\zeta_{\bar{z}}^2&=E(\bar{z}-\bar{X})^2\\&=E\left[(\bar{z}-\bar{Z})+(\bar{Z}-\bar{X})\right]^2\\&=\sigma_{\bar{z}}^2+\bar{D}^2\end{aligned}\right\}\qquad(1.15)$$

公式（1.15）中所确定的均方差是估计数$\bar{z}$在$\bar{X}$周围的变化数。均方差的方根称为均方根误差。与精确度相反，ζ^2所衡量的是估计数$\bar{z}$的准确度。因此，“准确度”一词指的是估计数量的真值，而精确度指的是估计数的期望值，从已用的公式来类推，我们说如果$\zeta_{\bar{z}}^2<\zeta_{\bar{z}}^2$，则估计数$\bar{z}$便比另一个估计数$\bar{z}$更准确。反之也是一样。必须明确区分精确度和准确度。一个估计数可能很精确但同时也可能不太准确。图1也许有助于了解这样的可能性。

根据前面的解释，读者可以把准确度的概念推广应用于其它统计尺度。

公式（1.15）对于抽样调查工作极为重要。$\zeta_{\bar{z}}^2$的值包括两项：一项是$\zeta_{\bar{z}}^2$，取决于样本的大小；另一项是$\bar{D}^2$，则与样本的大小无关。这表明在试图提高估计数$\bar{z}$的准确度的时候，增加样本量的决定在某些情况下可能是很不适宜的。图2所示，就属于这种情况。

在这个图中，Y轴表示示$\zeta_{\bar{z}}^2$的值。在x轴上表示出样本的大小n。曲线表示随着样本量的增加，$\zeta_{\bar{z}}^2$的值减少。如果在调查中了解到的$\zeta_{\bar{z}}^2$的值在C点上，而我们希望以减少$\sigma_{\bar{z}}^2$的办法来减少$\zeta_{\bar{z}}^2$，譬如说减少百分之二十五，其结果将使样本量增加到无法办的程度。另外有一个办法是可以使$\bar{D}$等于零或接近零，不过这个办法需要一些花费，但比较巧妙，它将自动导致$\zeta_{\bar{z}}^2$减少，这个方法可能比增加样本量花费少。

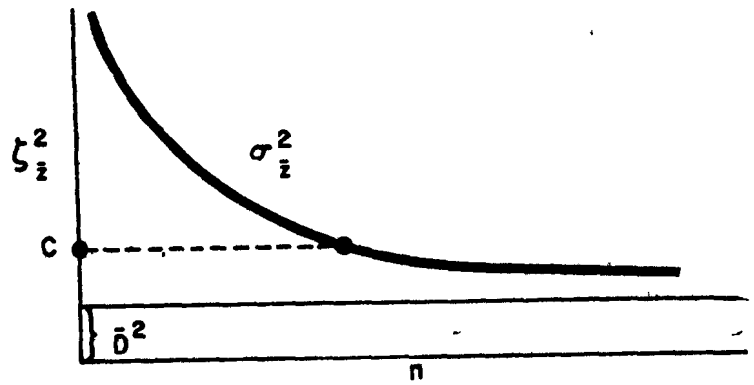

图2　增加样本量对均方差值的影响

这里针对数据不准确而确定的均方差、方差与偏差三者之间的关系在数据准确和有偏估计过程中也同样适用。

我们现在可以看出在有些情况下使用有偏估计方法之所以更为可取的原因。图3就说明了这样的情况。曲线A为算术平均数的无偏估计数分布情况。这个分布情况的算术平均数是$\bar{X}$曲线B为有偏估计数$\bar{z}$的分布情况。$\bar{Z}$是平均数。偏差的数值是$\bar{D}=\bar{Z}-\bar{X}$。从图中可以看出$\sigma_{\bar{x}}^2>\zeta_{\bar{z}}^2$如果情况是这样，我们认为有偏估计数$\bar{z}$则更为可取，因为与相应的无偏估计数$\bar{X}$相比，有偏估计数$\bar{z}$更有助于了解真值$\bar{X}$的位置。因此，如果几种方法在其它各方面都相同，那就应当选用均方差最小的那个方法。

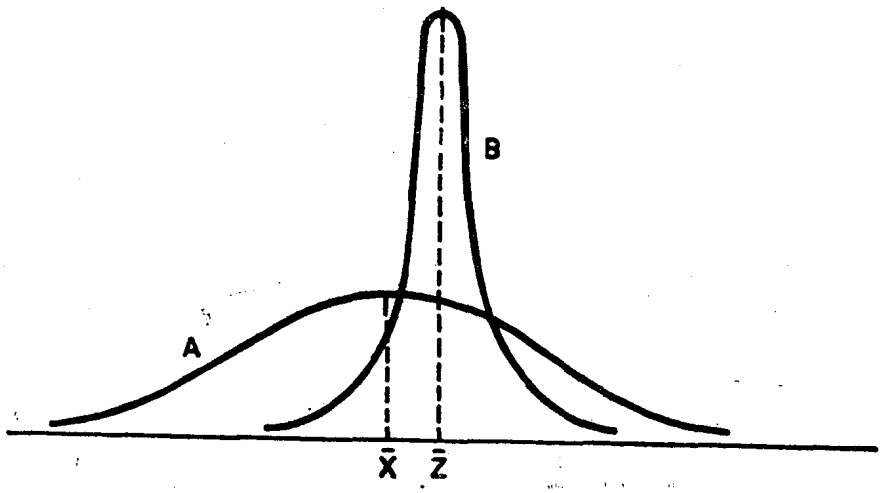

图3　说明有偏估计比无偏估计更为可取的一种情况

1.7　举例说明

现在举一些实例加以说明可能是有益的。第一个实例举自潘思写的一本小册子(4)。潘思研究了农作物单产的两种估计数。一种是公式（1.12）中的无偏加权算术平均数，另一种是公式（1.6）中的有偏简单算术平均数。各种农作物单产的这两种估计数的方差见表1。从这个表中可以看出，有偏估计数可以为相当大的偏差提供比较准确的数据。未加权平均数可产生与加权平均数相同的均方差，它的百分偏差等于$\sqrt{a^2-b^2}$。

表1　两种单产估计数的百分数标准误差(1)

作　　物		加权算术平均数	简单算术平均数
		(a)	(b)
小　麦	(1947—48年)	14.0	3.7
	(1948—49年)	10.0	2.5
棉　花	(1944—45年)	15.0	5.5
	(1945—46年)	14.0	6.9

(1)　V.G.潘思，同前.

第二个实例来自希腊农业普查中的一次质量检验。为了检验普查中报告的农田面积数据的准确度。对一批农田进行了抽样并对其各自的面积进行了测量。另外，有些地区还在地图上标出所报告的农田，而遗漏的农田也就查明了。检验结果见表2。和前面一样，我们用z表示普查中了解到的有关特征数值，x表示检验结果。我们可以看出，估计平均数$\bar{z}$是受回答偏差的影响的。关于农田面积的回答偏差估计在百分之十二以上，关于农田块数的回答偏差估计为百分之三十二左右。两种偏差都是负的。这意味着作为一种数据收集方法，报告对有关的特征数的估计偏低。

(4)　V.G.潘思：《农作物的单产估计》(联合国粮食及农业组织，罗马，1954年)。

表 2　希腊农业普查以及面积数据质量检验的一些结果①

特征数 \ 计量指标	$\bar{z}$	$\bar{x}$	$\bar{d}$	σ_x^2	σ_x^2	σ_d^2	ρ_{dx}
面积(按斯特雷马测量单位计算)	37.8	42.2	−4.4	920	680	507	−0.23
农田块数	6.6	8.7	−2.1	32.5	33.0	18.9	−0.39

① 国家统计局：普查中农民的报告，油印报告，1962年。

如果用两种方法（即报告和测量）来估计农田的平均面积（按斯特雷马计算），在抽样农田面积的估计方面可以达到的准确度，请见表3（根据表2的数据整理）。在以均方根误差来计算准确度时，我们使用了公式（1.15）。就测量来说，我们假设$\bar{D}$等于零。还假设报告和测量的质量不随样本的大小而变化。

表3 用报告和测量两种方法估计各块抽样农田平均面积时各种样本大小的百分数均方根误差的数值

方法	样本的大小			
	1	100	10,000	全面调查
报告	30.6	5.3	4.7	4.4
测量	26.1	2.6	0.26	0

表 3 的数据很清楚地表明，如果数据有误差会出现什么情况。用均方根误差来衡量的报告的准确度接近4.4斯特雷马的限度；这个限制是偏差的数值。即使进行一次全面调查，均方根误差的这个数值仍然不变。由此可见，根据由抽样测量采集的不大不小的样本得出的估计数，比根据报告全面调查所有农田得出的估计数更为准确。

根据表 2 的数字，还可以很容易地计算出用这两种方法获得一个相等的均方根误差需要多大的样本。

如果知道费用因素，例如报告和测量一块地的平均费用，那就有可能作进一步的推测。如果调查预算为C，c_1和c_2分别是通过报告和测量获得每块地的面积数据的平均费用，那就需要知道在预算范围内用哪一种方法获得的数据更为准确。

这个问题的答案是清楚的：按照提出的假设情况，应当选用导致均方根误差较小的那一种方法。用角标 1 代表报告，角标 2 代表测量，可列入用这两种方法进行调查的单位数大约是$n_1=\frac{C}{c_1}$和$n_2=\frac{C}{c_2}$。把n_1和n_2的这些值列入公式（1.15），就能得出这个问题的答案。在其它情况下要求的准确度可能是固定的，但问题是要确定哪一种数据收集方法需要的预算较少。在这种情况下，ζ的值是固定不变的。然后用公式（1.15）求出上述的n并利用算出的样本大小来确定调查预算。这样就可找到预算较少的那种方法了。另一个与此类似的实例请见大卫写的一篇论文(5)。

(5) M·大卫：《1959年获得福利援助的一些抽样家庭所报告的收入情况的有效性》，载《美国统计学会学报》1962年第57卷，第680—685页。