统计方法应用标准化丛书

数据的统计处理和解释

张尧庭 编著

中国标准出版社

数据的统计处理和解释

张尧庭　编著

中国标准出版社

图书在版编目(CIP)数据

数据的统计处理和解释/张尧庭编著. —北京：中国标准出版社，1996.6

（统计方法应用标准化丛书/成平主编）

ISBN 7-5066-1213-5

Ⅰ.数… Ⅱ.张… Ⅲ.①统计资料-数据处理②统计资料-分析 Ⅳ.C812

中国版本图书馆 CIP 数据核字(96)第 00895 号

中国标准出版社出版

北京复兴门外三里河北街 16 号

邮政编码：100045

电　　话：68522112

中国标准出版社秦皇岛印刷厂印刷

新华书店北京发行所发行　各地新华书店经售

*

开本 880×1230　1/32　印张 6⅛　字数 172 千字

1997 年 3 月第一版 1997 年 3 月第一次印刷

*

印数 1—2 000　定价 13.00 元

*

标 目 299—01

丛书编委会

总　序

当今世界，由于地区化、集团化经济的发展，贸易竞争日益激烈，产品质量的竞争已成为贸易竞争的最重要的因素。在这种形势下，各企业都深刻地感到不提高产品质量就没有出路，不能生存；产生了强烈的提高产品质量的紧迫感。我国政府有关部门也正在制定质量振兴计划，以迅速提高我国产品的市场竞争能力。提高产品质量，一要依靠技术进步，二要加强科学管理。有人说，三分技术，七分管理，这是很有道理的。GB/T 19000—ISO 9000 族标准的发布为我国企业进行科学的质量管理提供了保证。这一系列标准提出了建立质量体系的一系列要求，并将统计技术也作为要求提出来，可见统计技术是科学质量管理的重要手段，要贯彻质量管理和质量保证标准离不开运用统计技术。

迄今为止，我国已正式颁布了 77 项数理统计方法标准，涉及到数据处理、质量控制图、抽样检验、产品可靠性等方面。这些标准都是由全国统计方法应用标准化技术委员会负责制定、审查的。该技术委员会集中了一批享有声望的数理统计专家。多年来，他们在数理统计应用于质量管理方

面，做了大量的研究和推广工作。为了更好地宣传、推广统计技术，他们编写了这套《统计方法应用标准化丛书》。这套丛书包括下列四个分册：

《数据的统计处理和解释》

《生产过程质量控制》

《产品质量抽样检验》

《可信性工程（可靠性、维修性、维修保障性）》。

这套丛书深入浅出地阐明了在质量管理工作中，如何使用统计方法标准，并介绍了通过使用统计方法标准，提高产品质量、降低产品成本的有效途径。它的出版，无疑对于建立科学的质量体系有着十分重要的指导意义。

这套丛书避免了高深的数学推导，以实用性为主，内容十分丰富，理论上既严谨又通俗易懂，具有可读性、可操作性，是广大科技人员、管理人员掌握数理统计技术的一套好书。

叶柏林

1995.10.15

前　言

宣传、推行统计方法应用国家标准是提高我国产品质量、改进企业管理的重要方面。为了能对这方面标准有一个比较全面的了解，组织出版一套《统计方法应用标准化丛书》，系统地介绍有关的标准和基本原理。本书就是这套丛书的基础部分。

本书的重点是两个内容：介绍统计方法的一些基本概念和原理；解释有关的统计方法国家标准。第一部分的内容是了解统计标准所必须的，但又不能是一本高校的统计教材，所以写法上与普通教材不同，着重从实际问题背景去理解统计的概念和原理，不是着重于数学的推理。一般具有中学数学知识的人可以看懂大部分内容，如果知道一些微积分的知识，对一些公式的理解就会更好。所以本书的第 1 章至第 4 章的部分内容不仅对理解本书介绍的标准有用，对于理解其他的统计标准也是会有帮助的。

本书的另一个内容是给一些统计标准作出说明，为什么标准这样定，它的适用范围是什么，如何使用，如何决定样本量的大小等，对这些问题，

都作了原则的分析和示例的解释，希望能对大家有真正的帮助。

限于水平和能力，本书是否能达到上述要求还需由读者去判定，欢迎提出批评和意见。

张尧庭

1995 年 6 月

目　录

第1章

基本概念

1.1 引　言

标准化是现代工厂企业管理上的一项重要工作，它对于保证产品质量、改善经营管理起着不可替代的作用。无论是制定标准还是推行标准，都要与大量的资料、数据打交道，怎样从大量的资料、数据中提取所要的信息，从而能作出比较正确的判断，这都离不开统计。

统计的功能不只是记录过去的状况，作一些简单的整理、归纳，让人们对研究对象的总的轮廓有一个了解，更重要的是分析这些资料、数据为什么会不一样？应从哪些方面去探索形成这些差异的原因，揭示研究对象随时间的推移相应的变化规律，在这个基础上对未来的状况作出估计和判断，帮助人们作出正确的判断和决策。所以统计不是只能描述、记录历史，而是分析历史、对现在和将来根据数据提供的信息来作判断。它是一种认识客观事物规律的手段。

就以打靶的记录为例。如果我们观察了一个射手100次打靶的结果，其中72次中靶，28次未中；中靶的72次弹着点的坐标记录下来是：

$$(x_1,y_1),(x_2,y_2),\cdots,(x_{72},y_{72})$$

很明显，一般说来，72个中靶的位置不会是一样的。首先就要问：这些(x_i,y_i)，$i=1,2,\cdots,72$为什么有差异？如何从数据中分析出造成这些差异的原因？粗略地说，有两个原因造成了这100次打靶结果的差异：

(i) 这个射击手的打靶技术。如果技术高，弹着点的密集程度

高；如果技术差，弹着点就分散。

(ii) 这个射击手打靶时是否瞄准中心？如未瞄准，弹着点就会偏离中心，以致有的脱靶。

很容易看出，原因(i)需要长期训练后才能改进，而原因(ii)只要调整瞄准的中心位置就可改进。前者称为随机因素，也就是种种难以确定的因素造成的对射击手的影响，后者称为系统因素，这是比较明确的一个对结果有确定影响的因素。如果对这两者分析清楚了，就可以判断脱靶的原因大致是什么。怎样能从这 100 个打靶结果中分析出上述判断呢？这就需要统计分析，统计分析的重要作用也就在这些方面。

从生产过程来看，每一道工序都在对产品进行加工，其中一道工序加工质量不好，就会影响后续的工序，甚至造成大量次品或废品。所以常常在生产时隔一定时间抽查若干件产品，检验这道工序加工质量是否合格，是否有变坏的趋势；应否调整，如何调整。这也要根据生产中历史资料以及现在的抽检结果作出判断，而统计分析也能提供较好的处理方法。

根据生产、管理中的实际需要，介绍一些统计的基本概念和常用的方法，有助于读者理解一些国家标准(或国际标准)制定时所依据的原理，掌握正确使用这些标准的方法。这就是本书的目的。我们不采用对统计方法方面的国家标准进行逐个介绍或逐个说明的方法，因为每个标准本身都已有这些内容。我们的目标是为使用这些标准的人提供一份便于自学、通俗易懂的教材，通过自学，理解和掌握有关的标准。这一章着重介绍一系列基本的概念，这些概念与每个统计标准都有联系，是理解各个统计标准必备的基础。

1.2 指标的分类

我们把研究对象中观察、记录的内容称为指标，有些内容是描述属性的，有些内容是可以通过工具来测量的，大致上可以分为下面的四类：

(i) 计量的。如物体的长度、重量、室内的温度、湿度等等。这一

类指标的特点是：原则上它的取值可以是实数某一区间内的任一个值，通常称这类指标是连续型的。这种指标的统计分析与具有密度的连续随机变量的分布相联，在通常的统计教科书中占有主要的地位。

(ii) 计数的。如一批产品中的不合格品个数、一个居民区内拥有的电冰箱台数、固定资产在亿元以上企业的个数等等。这一类指标的特点是：它的取值范围是整数或自然数，通常称这类指标是离散型的计数指标(或计件指标)。这种指标的统计分析与离散的随机变量的分布相联，在通常教科书中不占重要地位，偶而论述到它。

(iii) 有序的。这是一种属性指标可以反映程度深浅的，例如颜色的深浅、味觉的鲜美程度、手感的丰满程度、弹性好不好等等，可以比较它们的顺序，但无法量化，这一类指标称为有序的。有时人们为了方便，对有序指标往往也用量化的符号与术语来表示，例如医生体检时，表示砂眼的程度用"+"号的个数来描述，工厂的产品质量用一等品、二等品、三等品来标记，这里的一、二、三已没有量的意义，只是一个顺序的标记，所以这类指标不能认为是计数的指标，在使用统计方法时必须要认识这种差别。

(iv) 名义的。这是单纯的属性指标，例如给产品编号，给公民以身份证号等等，这些号只是一个代码，它的值和大小顺序一般不表示任何意义。这一类指标称为名义指标。然而不要把名义值与名义指标混同起来。一个电阻的名义值是10Ω(欧姆)，但它的实际值可能是10.04Ω，名义值是产品规格上标定的值，这个值是有意义的，它不是代码。

这四大类指标可以归为两大类：

定量的：计量、计数；

定性的：有序、名义。

目前的标准大部分都是与定量指标有关的，只有极少数似乎涉及定性指标，实际上也还是定量指标的问题，所以我们这本书也是以定量指标作为研究的主要对象。

指标的分类可以从不同的角度去考虑，例如生产中往往要考虑可控指标与不可控的指标，从心理学上往往要考虑是主观的指标还

是客观的指标等等。我们上面介绍的分类主要是从统计分析处理的方法来考虑的，不同的指标往往处理的方法不一样，这些是必须注意的。

1.3 指标的分布

当我们想了解一批产品的性能时，性能指标对每个产品并不是一样的。例如一批灯泡，它的使用寿命随每个灯泡而改变，由于生产这批灯泡的原材料、工艺、操作人员的条件是一样的，所以它们的寿命都相近，集中在某一个值的附近，描述这种状况的工具就是寿命指标的分布。如何获得指标的分布呢？通常有两个途径：一种是从实测数据出发，整理后拟合一条分布曲线作为指标的分布；一种是从理论出发，推导出分布曲线应该是一个什么样的函数。当然最理想的是：理论上已给出分布曲线的函数类型，然后根据实测数据去确定这类函数中的参数。下面我们通过一些例子来说明这些。

先解释一下什么是指标的分布。以灯泡的寿命为例，一批灯泡的寿命用 T 表示，在这批灯泡中，$T\leqslant 3000$ 小时所占的比例用 $P(T\leqslant 3000)$ 表示，或写成 $F_T(3000)=P(T\leqslant 3000)$，于是：

$$F_T(x)=P(T\leqslant x) \tag{1-1}$$

就表示“寿命 T 小于或等于 x”在这批灯泡中所占的比例，$F_T(x)$ 是 x 的函数，它就称为这批灯泡寿命 T 的分布函数。它是平时累计的百分比的更精确、更一般的术语。例如我们已知混凝土块的抗压强度分组的百分比如下表，表中第一列表示强度分组的值，第二列表示各组相应的百分比，第三列表示累计的百分比。

如果用 T 表示混凝土块的抗压强度，则 T 的分布在表 1-1 中只给出了一部分的值，实际上表 1-1 的第三列是 $F_T(x)$ 在 14 个点上的值，即：

$F_T(18.95)=1\%$，$F_T(20.95)=2.75\%$，

$F_T(22.95)=8.25\%$，$F_T(24.95)=23.75\%$，

$F_T(26.95)=45.50\%$，$F_T(28.95)=71.25\%$，

$F_T(30.95)=86.75\%$，$F_T(32.95)=94.75\%$，

$F_T(34.95)=97.50\%$，$F_T(36.95)=98.50\%$，

$F_T(38.95)=99.25\%$，$F_T(40.95)=99.75\%$，

$F_T(42.95)=99.75\%$，$F_T(44.95)=100\%$

对于 $F_T(x)$ 的函数形式，如何确定其他 x 相应的 $F_T(x)$ 的值，表 1-1 都没有给出。数理统计正是要回答这些问题，当然首先要了解指标分布 $F_T(x)$ 表示的是什么意思。从这个例子可以看出，对每个指标 T，客观上存在一个指标的分布 $F_T(x)$，我们通过实际观察了解它在一部分点上的值。

表 1-1　混凝土块抗压强度表

抗压强度 N/mm² (1)	各组占的百分比，% (2)	累计百分比，% (3)	抗压强度 N/mm² (1)	各组占的百分比，% (2)	累计百分比，% (3)
16.95～18.95	1.00	1.00	32.95～34.95	2.75	97.50
18.95～20.95	1.75	2.75	34.95～36.95	1.00	98.50
20.95～22.95	5.50	8.25	36.95～38.95	0.75	99.25
22.95～24.95	15.50	23.75	38.95～40.95	0.50	99.75
24.95～26.95	21.75	45.50	40.95～42.95	0.0	99.75
26.95～28.95	26.25	71.25	42.95～44.95	0.25	100.00
28.95～30.95	15.00	86.75	合　　计	100.00	
30.95～32.95	8.00	94.75			

自然会想到，如果我们观察的混凝土块更多，分组后的百分比就会改变，累计的百分比也会改变，是否 $F_T(x)$ 也会改变呢？从理论上看，这批混凝土块的抗压强度客观上有一个分布 $F_T(x)$，它不随观察资料的多少而变化，它的存在是确定无疑的。我们表上所得的、根据实测数据算出的累计百分比只是反映了被观察到的这一部分抗压强度的分布，它当然随我们观察到的资料多少而改变，它与客观存在的 $F_T(x)$ 并不完全相同。如果我观察了很多，自然可以认为观察到的累计百分比与客观存在的真实的 $F_T(x)$ 几乎相同。从这里我们就能体会到 $F_T(x)$ 的确是累计百分比的更深入的一个概念。

描述指标的分布状况，由式(1-1)这样来定义的分布是很自然

的，然而，进一步的讨论将会发现还有其他的描述分布的方法，它们比式(1-1)的形式更便于讨论和作数学上的处理，这些我们将在第二章中详细论述。这里主要是要有一个分布有概念，理解分布和比例的关系。

1.4 数据与取样

当我们想了解产品的某项指标时，通常的办法就是进行测试，有些指标是便于测试的，有的是很难测试的，有的甚至是破坏性的。例如灯泡的使用寿命、电视机的开关次数、混凝土的抗压强度等等，测试一件就破损一件，不可能逐一测试，而且还只能测试很小一部分。从考察的全体对象中抽取一小部分来测试，这一小部分就称为样本，从样本所得的测试结果就成了我们掌握的数据。统计方法标准的绝大部分都涉及数据处理，这些数据又都是取样得来的，数据就是反映样本信息的资料。很明显，样本提供的信息仅仅是所要考察的全体对象的一部分，它既反映、又不完全反映全体对象的状况，我们的目的恰恰是要了解总的情况，所以搞清楚样本在多大的程度上能较好地反映总的状况，就是很重要的。在这一小节，我们逐步地引入一些概念，来把其中的关系叙述清楚。

我们把研究考察的全部对象称为总体，总体中的一个成员称为个体，从总体中抽查的部分个体称为样本。例如我们要了解这一班学生的英语水平，全班有 50 名学生，这就是问题要考察的全部对象，他们组成了一个“总体”，每个学生就是一个“个体”，如从中抽查了三个学生，这三个学生就成了“样本”。现在的问题是：如何选择三个学生能反映这个班学生的英语水平。如果选取三个班上最好的学生，测试的成绩并不反映这个班学生的成绩，应该说是反映了这个班的最高水平；同样如果选取三个最差的学生来测试，所得的成绩反映的是这个班的最差水平。如果想了解这个班的学生上、中、下三种水平的情况，那就应该取一个好的、一个中等的、一个差的去参加测验。可见对于不同的要求，应有不尽相同的取样方法，如果我们想了解这个班的学生的英语水平的整体状况，既不是最好，也不是中等或最差的，而

是要了解整个的状况，又不是对50名都测验，而是只选一部分，这就发生了问题，应如何选取一部分呢？这一部分的比例应多大呢？现在来用以前的概念把它叙述清楚。只有把问题弄清楚了，才能找到正确的解决办法。

首先要明确问题所考察的总体实际上是与特定的指标有关的，本例中就是学生的英语水平，用什么指标来反映英语水平呢？就是考试成绩（这里不考虑试题有问题等其他因素，正如同检查产品技术指标是否合格时不考虑测试仪器、人员等因素一样），因此实际上要考察的就是50个英语测试的成绩，也就是说，我们想从这50个客观存在的英语成绩中抽查一部分，以这一部分成绩来了解这50个的状况。50个客观存在的成绩用$x_1, x_2, \cdots, x_{50}$来表示，选取的一部分，就是这$x_1, x_2, \cdots, x_{50}$中的某几个。这样我们就可以理解到总体实际上是许多数据形成的一个特定的集合，而样本是这个集合抽查部分数据形成的一个子集合，关键是抽取的样本要能代表总体的状况。什么是样本的"代表性"呢？怎样选取才能使样本具有所要的"代表性"呢？现在设想要检查的产品数目很大，例如大批的电子元件，检查的指标是电阻值，总体就是电阻值形成的一个集合，而样本是抽查一部分所形成的子集合。如果我们把总体电阻值y的分布$F_y(x)$这一概念引入，即客观上"$y \leqslant x$"的电阻在总体中占的比例是确定的，用$P(y \leqslant x)$表示这个比例，$F_y(x)=P(y \leqslant x)$就是y的分布，它是x的函数，随x而改变，$F_y(x)$很好地描述了电阻值y的分布状况。因此取样的"代表性"就可以认为是取出的样本的分布应和总体分布$F_y(x)$很相似，很接近。很明显，只取"好的"或者只取"坏的"，都无法反映出$F_y(x)$的分布，只有按所占的比例取样，才能使样本的分布与总体分布很相似。按比例取样就是$P(y \leqslant x)$大时，样本中电阻值小于或等于x的比例也应大；$P(y \leqslant x)$小时，样本中相应的这部分比例也应小。问题是我们根本就不知道$F_y(x)(=P(y \leqslant x))$这个函数，怎么能实现按比例取样呢？这里随机化就发挥了很大的作用。如果我们随机地从总体中抽查一个电阻，设它的值为ξ，于是"$\xi \leqslant x$"发生的可能性（也就是概率）就由"$y \leqslant x$"所占的比例$P(y \leqslant x)$决定，所谓"随机"就是