SQC-1

统计质量控制

STATISTICAL QUALITY CONTROL

数据收集和整理

Data Collection and Processing

陈国铭　主编
杨丽春　　编

中国石化出版社

SQC-1

统计质量控制

STATISTICAL QUALITY CONTROL

数据收集和整理

Data Collection and Processing

陈国铭　主编

杨丽春　　编

中国石化出版社

（京）新登字048号

内 容 提 要

本分册作为《统计质量控制》丛书的第一册，着重于对质量特性数据的产生、特点和各种初步整理方法的介绍。其中，数据列表整理法，图形整理法及特征值法都是数据处理过程中的最常用和最基本的手法。本书还介绍了有关异常数据的判断方法、测量值误差的概念、表示方法及有效数字的运算问题。

本书以方法介绍为主，并力求通俗易懂，适用面广，可供企业的各级质量管理、质量检验及工程技术人员阅读。

SQC-1

统计质量控制

STATISTICAL QUALITY CONTROL

数据收集和整理

Data Collection and Processing

陈国铭 主编

杨丽春 编

*

中国石化出版社出版发行

（北京朝阳区太阳宫路甲1号 邮政编码：100029）

煤炭工业出版社印刷厂排版

中国纺织出版社印刷厂印刷

新华书店北京发行所经销

*

850×1168毫米 大32开本 4 3/4 印张 126千字 印 1—6400

1995年3月北京第1版 1995年3月北京第1次印刷

ISBN 7-80043-554-7/O·019 定价：5.50元

中国石油化工总公司质量管理协会组织编写

生产技术顾问：张德义
统计技术审核：王经涛

主　编：陈国铭
副主编：张祖荫　郭耀曾
编　委（按姓氏笔划）：李世英　陈国铭　杨丽春
张祖荫　饶上建　郭耀曾　崔廷铨
其他编辑校核人员：万　涛　刘秋萍　吕巧云
邱以玲　田从金

序　言

为了适应国际贸易往来和经济合作的要求，国际标准化组织经过十多年的努力，于1986年和1987年相继正式发布ISO8402《质量——术语》标准和ISO9000质量管理和质量保证系列标准，将世界多年质量管理的经验进行了标准化。ISO9000系列标准的基本点是要求企业在生产过程中建立有效的质量保证体系，并对质量体系中相互关联、相互作用的若干要素进行有效的控制。在过程质量控制中，科学、有效方法之一就是数理统计方法。因此在ISO9000系列标准的各个模式中以及质量管理和质量体系要素指南中都要求在市场分析、产品设计、工序控制、性能评定、数据分析等方面广泛使用统计技术，其范围包括实验设计、方差分析、显著性检验、累积和控制图、抽样检验等技术。因此，研究学习统计质量控制技术对于贯彻ISO9000质量保证系列标准，提高科学管理水平是非常必要的。

回顾世界质量管理的发展史，可以看出，数理统计技术在质量管理中发挥了重要作用。从19世纪末到现在，质量管理在历史上经过了检验质量管理、统计质量控制和全面质量管理三个阶段。单纯检验质量管理的严重缺点：一是只能从产品中发现和挑出废品，事前预防功能不强，二是由于检验人员的差错，即使全数检验也可能漏检或错检；三是至关重要的破坏性试验不可能全数进行。产品是生产出来的，单靠检验是不能防止产生废品的。1924年美国贝尔研究所的休哈特（W. A. Shewhart）运用数理统计的原理提出了控制生产过程中的“6σ”方法，即后来发展的质量控制图和预防缺陷的概念。与此同时，同属贝尔研究所的道奇（H. E. Dodge）和罗米格（H. G. Romig）联合提出了在破坏性试验情况下采用的“抽样检验表”。二次大战初期，美国大批民用品转入军

工生产，由于事先无法控制废品而不能满足交货期要求，又由于军工生产多属破坏性试验，全数检验不可能也不允许。美国国防部为了解决这一难题，邀集休哈特、道奇、罗米格以及美国材料与试验协会、美国标准协会、美国机械工程师协会等有关人员研究，于1941～1942年先后公布一系列“美国战时质量管理标准”，要求各公司普遍实行统计质量控制方法，结果半年内取得显著成效。后来统计质量控制取得了很大发展。

我国自从1978年从日本引进全面质量管理十多年取得了显著成效。纵观我国的质量管理发展历史，是由检验质量管理直跃全面质量管理，对数理统计方法的运用远不是像当年美国那样深入广泛，不少决策、设计、科研、生产、销售、服务部门在提出问题、解决问题、检查结果时有些人还不习惯于进行科学的数理解析。

为了普及数理统计基本知识并在生产实际中发挥作用，我们组织石化行业中具有实践经验的质量管理专家编写了这套《统计质量控制》系列丛书。本书共分十册，第一册是数据收集和整理，第二册概率和数理统计基础，第三册估计和检验，第四册控制图，第五册方差分析，第六册实验设计，第七册相关和回归分析，第八册抽样检验，第九册统计方法应用演示50例，第十册数表。

数理统计方法就是通过对生产实践中大量数据的收集、整理、解析，研究生产实际中的内在规律的数学方法。和目前国内其它有关数理统计的书籍相比，本系列丛书的显著特点；一是它不同于一般的数学教科书，特别突出了实际应用，因此在编写中尽量减少不必要的公式推导，是一本实用性较强的书籍；第二个特点是书中列举了大量社会和生产（特别是石油化工生产）实例，文章从实例引出理论，又从理论回到实例，便于读者理解和应用，适合于工业企业特别是石油化工等流程型行业设计、研究、生产、销售、辅助等系统技术人员和管理干部学习参考；第三是语言既通俗易懂，又有一定深度和广度，既可用于中等水平人员学习应用，又可适用于高等水平技术人员研究参考。

为了更好应用本书，建议学习中注意几点：一是随着计算机的高度发展，许多数理统计方法可完全不需用手工计算，即可很快得出结果，已经掌握了统计方法的人可直接借助计算机，但对于初学之人，还是先用手算为好，防止知其然而不知其所以然，不利于在实践中灵活运用；二是对于现场技术人员，不要去深究公式推导，只要求会实际灵活运用；三是统计方法只提供解决问题的手段，必须和固有技术相结合才能解决问题，因此要使读者学会用数学的思维考虑专门技术问题；四是质量管理所用的方法不限于数理统计方法，还包括许多其它方法，如价值分析（VA）、生产工学（IE）、操作研究（OR）、价值工程（VE）、可靠性工程（RE）等，本书这次没有列入，读者可根据需要深入研究，灵活运用。

侯荣生

1995 年 1 月

目　录

1　数据与质量管理 …… 1
1.1　基于数据的管理 …… 1
1.2　质量数据的特点 …… 2
1.3　质量数据要求 …… 4
2　数据的分类 …… 6
2.1　根据使用目的分类 …… 6
2.2　根据数量化尺度的分类 …… 6
3　数据的收集 …… 10
3.1　个体、总体和样本 …… 10
3.2　数据的收集过程 …… 12
3.3　数据的收集方法 …… 13
4　数据的列表整理法 …… 25
4.1　列表整理概述 …… 25
4.2　顺序及权数整理 …… 25
4.3　数据的分层法 …… 27
4.4　频数整理 …… 29
5　数据的图形整理法 …… 34
5.1　直方图 …… 34
5.2　频数分布曲线 …… 42
5.3　排列图 …… 44
5.4　相关图 …… 48
6　数据的特征值 …… 58
6.1　特征值 …… 58
6.2　数据中心位置特征值 …… 58
6.3　数据离散程度特征值 …… 64

6.4 无量纲的特征值 …… 70
7 数据的简化计算与近似计算…… 79
7.1 数据的线性变换 …… 79
7.2 近似计算 …… 85
8 异常数据 …… 103
8.1 异常数据的概念…… 103
8.2 异常值检验的显著性水平…… 103
8.3 已知总体标准差 σ 时的异常值判断…… 105
8.4 未知总体标准偏差场合的异常值判断（Ⅰ）…… 110
8.5 未知总体标准偏差场合的异常值判断（Ⅱ）…… 115
8.6 异常值判断准则…… 118
8.7 各种判断方法的选择及异常值的处理…… 121
9 误差 …… 126
9.1 误差的概念…… 126
9.2 误差分类…… 128
9.3 误差的来源…… 130
9.4 误差的表示法…… 130
9.5 准确度、精密度与误差的关系…… 131
9.6 有效数字的运算规则…… 134
本册使用符号…… 137
习题…… 139

1 数据与质量管理

1.1 基于数据的管理

在企业的质量管理活动中，不论是决策分析，还是质量控制，无处不遇大量的数据，也无处不需大量的数据。也就是说，任何企业的质量管理活动都不是凭主观想象做出结论或决策的，而是基于反映客观实际的数据的一系列科学管理活动。

当然，我们在现场所取得的原始数据或称信息，在未经任何处理前，很难直接作为控制和决策依据，必须要将取得的原始数据进行整理和数理统计解析，通过数据所表现的客观现象，分析事物的本质，为质量控制和决策提供科学可靠的依据。

取得原始数据以后，通常可按照以下三个步骤来处理。

第一步，将原始数据进行初步整理。初步整理的方法可归纳为三种：(1) 数据重新排列、组合、整理的结果常以“表”来体现，可简称列表整理法。(2) 用某种图示表现数据，使整理的结果一目了然，可简称为图形整理法。(3) 进行一些简单的数学运算，计算出某些数字特征值，可简称为数字特征整理法。这些数据的初步整理方法能从不同的角度或深度，将数据的分布状态大致描述或描绘出来，以至有时可直接为一些简单的质量问题提供决策依据。

第二步，对数据进行数理解析，数理解析是在大致了解数据分布状态的基础上，应用概率统计的原理与方法，进行更深入的统计分析，以确定出具有规律性、指导性的东西。这一步骤所处理的数据不仅指原始数据，还指中间数据，如某些特征值、统计量。

第三步，统计推断。统计推断工作则要对以上第二步所得到的统计规律结论，依概率统计原理，在确定的可信程度下，对所

要研究的事物做出推断或预测。

由此可见，质量控制和决策是一个运用数理统计方法，从数据的收集入手，并对其进行初步整理、数理解析、统计推断的过程。因此，数据是质量管理的基础，质量管理是依据数据的管理。

1.2 质量数据的特点

在这里，我们把反映某产品的某项质量指标的原始数据称为质量特性数据(也简称为质量数据)，如一批尿素缩二脲含量数据、丙烯水分含量数据、纤维强度数据、汽油干点、辛烷值数据等等，都可以被称为质量数据。

为了整理质量数据，我们应当对质量数据共有的特点有所了解。质量数据有什么共同的特点呢？让我们先对表 1-1 中的一组质量数据做个观察。

表 1-1 尿素单包重量数据表 （单位：kg）

40.10	40.12	40.12	40.12	40.10	40.02	40.12	40.18	40.06	40.16
40.14	40.18	40.18	40.10	40.16	40.12	40.16	40.08	40.16	40.20
40.18	40.12	40.10	40.18	40.16	40.18	40.16	40.16	40.14	40.20
40.14	40.16	40.12	40.16	40.14	40.14	40.16	40.16	40.18	40.14
40.16	40.16	40.14	40.20	40.20	40.16	40.14	40.14	40.16	40.14
40.22	40.22	40.18	40.22	40.22	40.28	40.22	40.26	40.24	40.16
40.20	40.22	40.24	40.22	40.18	40.22	40.24	40.22	40.26	40.22
40.24	40.30	40.22	40.22	40.26	40.14	40.10	40.12	40.08	40.14
40.14	40.14	40.12	40.12	40.12	40.14	40.12	40.12	40.12	40.12
40.20	40.14	40.20	40.14	40.14	40.14	40.16	40.12	40.12	40.06

表 1-1 中所记录的数据是某化肥厂在一次尿素质量行检中，对随机抽取的 100 袋尿素测得的单包重量数据。这组质量数据具有什么特点呢？显然，它们是一组不完全相等的数据，但稍加观察，你会发现它们也不是杂乱无章、无章可循的。首先，数据总是在一定的范围内波动着，其次是大量地数据位于波动范围的中心附近。如表 1-1 中的 100 个数据在区间（40.015，40.285）内波动，居中的一小段区间（40.12，40.18）上，集中了百分之六十多的数据，而这个小区间的长度只是整个波动区间长度的四分之一。再观察表 1-2 这组数据，这是某维尼纶厂从生产正常的 20 天

内取得的维尼纶纤度数据。

表 1-2　维尼纶纤度数据表　（单位：mg/m）

1.36	1.49	1.43	1.41	1.37	1.40	1.32	1.42	1.47	1.39
1.41	1.36	1.40	1.34	1.42	1.42	1.45	1.35	1.42	1.39
1.44	1.42	1.39	1.42	1.42	1.30	1.34	1.42	1.37	1.36
1.37	1.34	1.37	1.37	1.44	1.45	1.32	1.48	1.40	1.45
1.39	1.46	1.39	1.53	1.36	1.48	1.40	1.39	1.38	1.40
1.36	1.45	1.50	1.43	1.38	1.43	1.41	1.48	1.39	1.45
1.37	1.37	1.39	1.45	1.31	1.41	1.44	1.44	1.42	1.47
1.35	1.36	1.39	1.40	1.38	1.35	1.42	1.43	1.42	1.42
1.42	1.40	1.41	1.37	1.46	1.36	1.37	1.27	1.37	1.38
1.42	1.34	1.43	1.42	1.41	1.41	1.44	1.48	1.55	1.37

这批纤度数据在区间（1.265，1.555）内波动，且分布在区间（1.355，1.445）内的数据达百分之七十，居中的这个区间长度仅占波动区间的三分之一，也即，百分之三十的数据分散在区间（1.265，1.355）和区间（1.455，1.555）上。对以上两批质量数据所共有的这两个明显特征，我们称之为质量数据的两大特性——波动性与规律性。

1）质量数据的波动性

质量数据的波动性是指质量数据的不等同性，“波动”不仅意喻一批数据在某个值的上下随机变化，还意喻着数据变化的幅度不大。质量数据都具有这种波动性，这已是大家所司空见惯的。事实上，由于产品质量的波动是必然的，而质量数据作为产品质量的客观反映，其波动也是必然的。

数据的波动可根据引起波动的原因而分成两种类型：一种叫正常波动，另一种叫异常波动。

正常波动是由偶然性原因和难以避免的原因造成的产品质量波动。这类原因在生产过程中大量地存在，表现为无方向性地、反复经常性地对产品某质量特性产生着影响。这类原因虽大量地存在，但对产品质量所造成的影响往往较小。如机器的轻微机械振动，操作者动作上的微小差异，空气温度、湿度的微小变化等，常是引起产品质量正常波动、同时又难以避免的原因。正是由于导

致正常波动的原因是大量的、不易确定和难以消除的，因此，一般情况下，正常波动在控制的前提下被允许存在。

异常波动是因系统性原因或可以避免的原因而造成的产品质量波动。这类原因在生产过程中并非大量地存在，表现为具有方向性或周期性地、突然而至地对产品质量产生影响。这类原因虽少，但对产品质量造成的影响往往较大。如设备出现故障，操作者违反操作规程、原材料性质变化等。由于导致异常波动的原因是少量地，并且常带有方向性或周期性等特征，使得这类原因比较容易被查明。一般情况下，异常波动在生产过程中不允许存在，一旦出现，必须立即查明原因，消除异常波动。

2）质量数据的规律性

质量数据的规律性是指质量数据的分布状态具有一定的规律。如前面的两组数据，都具有“中间多，两边少”的分布规律，绝大多数的质量数据都具有这样的分布规律，正是因为质量数据有某种规律可循、才使得质量数据有了可分析性和可研究性。事实上，大多数的质量数据分布规律呈现正态分布（*normal distribution*）或近似于正态分布。

总之，了解和认识质量数据的特性，对于我们理解数据处理的原理、掌握数据整理分析方法都大有益处。

1.3 质量数据要求

既然质量数据是质量管理的基础，是客观质量的反映，那么，质量数据必须尽可能真实地反映客观事实，也即质量数据要真实可靠。不可靠的数据往往导致不可靠、甚至是完全错误的结论。没有数据，无法进行科学的分析与决策，这固然不好，但由不可靠的数据进行分析决策可能更糟。

数据的可靠性依赖于抽样、实验或检测方法与技术。因此，首先要科学合理地确定抽样方法，按照已定的抽样方案进行抽样，抽样量以及抽样点要符合规定要求，减小抽样误差（*sampling error*）。然后，要保证对样品的测量分析具有一定的准确度和精密度。这是保证数据可靠性最重要的环节。

准确度（*accuracy*）也称正确度或准确性。它是指实际测量所得的结果与被测对象的真实计量值的接近程度。

精密度（*precision*）也称为精度或重现性。它是指同一个实验的重复测定值之间彼此相近的程度。

要减小实验或测量误差，就需要准确度好且精密度高的实验或测量方法。

另外，数据的可靠性还一定程度地依赖于数据的完整性。对所有的检测结果要如实地记录，包括那些看上去不正常的数据，这种数据往往给我们提供了更重要的信息，不能随便丢弃。记录的同时可根据数据的用途，按照数据的修约规则或特殊要求进行修约，并将抽样方式和时间、测试方法和时间、所用测试仪器、分析人员、地点等必要的事项记录于原始数据表上。

以上是对质量管理统计方法原始数据的基本要求，其目的就是要保证原始数据能尽可能真实地反映客观事实，为下面的分析、推断打下坚实可信的基础。

2 数据的分类

2.1 根据使用目的的分类

当我们为着不同的目的去取数据，所用的抽样方法及数据量也随之不同，根据使用目的，数据大致可以分成以下几种：

1）以掌握问题点为目的而取数据。如建立质量计划，进行质量改进活动等，需要掌握质量现状，把握存在的问题点，而收集过去的数据和现在的新数据。

2）为工程解析取数据。一切的产品质量和工程质量不是检验出来的，而是生产出来的，因此，生产的全过程都会对质量产生影响，而不同的作业条件会产生不同的质量特性值，为了进行工程解析，所取数据必须反映出这种因果关系，把握了它的因果关系，就可能进行有效地工程解析。

3）为检验而收集数据。这是质量检验经常遇到的，目的是从全体中抽取部分，通过检测得数据，以判定全体是否合乎规定的标准。显然，这种数据必须和判定标准所要求的项目相一致。

4）为调整操作而收集数据。为了得到好的质量，常需要变换各种影响因素，通过调节，以决定最佳生产条件或最低成本。

5）为了控制而收集数据。为了维持其稳定的管理状态，需要及时不断地获得现状信息，一旦发现异常可立即采取措施消除异常，达到控制的目的。

6）为了记录而收集数据。没有明确的目的或暂时无明确目的，预想将来可能会需要这些数据而进行的收集。

2.2 根据数量化尺度的分类

不论以什么目的收集数据，只要是能够数量化的，即可以用数值表示的数据，要规定衡量的标准尺度，用这个尺度来度量各个测定量的大小，表示特性值的数据均可以数量化，构成尺度的

数值大致分为两类，计数值数据和计量值数据。

1）计数值数据

计数值指不能连续取值，只能计算个数的数值。例如，不合格品件数、非计划停工次数、输气管上砂眼的个数、布匹上的疵点数等，都是计数值，它们的每一次取值只可能是零或自然数。

计数值的特点是非连续性，在任何两个计数值之间不可能插入无穷多个数值，否则将出现不能表达原意义的数值。如，非计划停工次数 1（次）与 4（次）之间，最多只能插入 2（次）和 3（次）两个数值，再想插入任何不同于 2 和 3 的数值如 2.5，则不能表达停工次数的含义，因为停工次数不可能为 2.5 次。

计数值还可以再分为计件值和计点值。

计件值如上面提到的不合品件数，非计划停工次数。计件值数据一般服从于二项分布（*binomial distribution*）或超几何分布（*hypergeometric distribution*）。

计点值数据如砂眼个数、疵点数等。计点值数据一般服从于泊松分布（*Poisson distribution*）。这里提到的几种分布，我们将在第二册中给予介绍。

2）计量值数据

计量值是指可以在某个区间上连续取值的数值。也就是说，只要测量仪器的精度能达到，计量值可以是某区间上的任何一个实数。

例如，单包重量、塔顶温度、滤网使用寿命、弹力丝的伸度、纤度等等，都属计量值。

计量值数据一般服从或近似服从于正态分布。

计量值的特点是，在任何两个计量值之间还可以插入无穷多个数值。如在滤网使用寿命数据的 2160（*h*）与 2160.5（*h*）之间，插入 2160.1，2160.2，2160.3，2160.4 以及插入 2160.11，2160.12，……等等，都是有实际意义的。

3）关于比率值和百分数

比率值和百分数的结果多表现为小数形式，从形式上看它们

都象是计量值。它们究竟属于计数值数据还是计量值数据，却应由计算比率值或百分数的商式的分子来确定。即当商式的分子是计数值数据时，所求得比率值或百分数值为计数值，当商式的分子是计量值数据时，所求得的比率值或百分数就为计量值。

例如，当某产品的合格品率定义为：

$$\frac{\text{合格品件数}}{\text{检验产品总件数}}\times 100\%$$

时，这个合格品率值为计数值，因为分子为合格品件数，属计数值数据。又如，某添加剂含量定义为添加剂数量与单位体积之比，即：

$$\text{含量}=\frac{\text{数量}}{\text{单位体积}}$$

则这个含量数据为计量值数据。

例 2-1 测得三个 1kg 石油酸样品含水分别为 7g、5g、7.5g，问含水率分别是多少？它们是计量值数据还是计数值数据？

解：

由三个样品的含水率分别为：

$$\frac{7}{1000}\times 100\%=0.7\%$$

$$\frac{5}{1000}\times 100\%=0.5\%$$

$$\frac{7.5}{1000}\times 100\%=0.75\%$$

各商式的分子均是含水量，即计量值，所以含水率分别为 0.7%，0.5%，0.75%，且均是计量值。

例 2-2 为检验某批产品单包重量合格率，测得 500 袋中有 7 袋为不合格的袋重。试问袋重不合格品率为多少？这个值是计量值还是计数值？

解：

$$\text{袋重不合格品率}=\frac{7}{500}\times 100\%=1.4\%$$