多变量统计过程控制

张　杰　阳宪惠　编著
王桂增　审

化学工业出版社
工业装备与信息工程出版中心

多变量统计过程控制

张　杰　阳宪惠　编著
王桂增　审

化学工业出版社
工业装备与信息工程出版中心
·北　京·

（京）新登字 039 号

DK49/26

图书在版编目（CIP）数据

多变量统计过程控制/张杰，阳宪惠编著．—北京：化学工业出版社，2000.8
ISBN 7-5025-2881-4

Ⅰ.多… Ⅱ.①张…②阳… Ⅲ.多变量-数理统计-过程控制 Ⅳ.TP273

中国版本图书馆 CIP 数据核字（2000）第 28303 号

多变量统计过程控制

张 杰 阳宪惠 编著

王桂增 审

责任编辑：刘 哲

责任校对：马燕珠

封面设计：郑小红

*

化 学 工 业 出 版 社
工业装备与信息工程出版中心 出版发行

（北京市朝阳区惠新里 3 号 邮政编码 100029）

http://www.cip.com.cn

*

新华书店北京发行所经销

北京市管庄永胜印刷厂印刷

三河市东柳装订厂装订

开本 787×1092 毫米 1/16 印张 8¾ 字数 206 千字

2000 年 8 月第 1 版 2000 年 8 月北京第 1 次印刷

印 数：1—3000

ISBN 7-5025-2881-4/TP·269

定 价：19.00 元

前　言

随着全球化经济的进展，市场竞争变得越来越激烈，提高产品质量并保证产品质量的高度一致性是企业在竞争中获胜的一个重要手段，对生产过程的监控和优化是保障产品质量的重要途径。多变量统计控制是近年来发展起来的一种过程监控方法。通过对过程数据的多变量统计分析，可以从大量过程变量的变化中找出影响产品质量的主要原因，并进一步采取措施来提高产品质量和过程能力。通过对在线测量的大量过程变量进行实时统计分析，还可以及时地发现那些仅靠观测各单一变量难以发现的过程变化，从而及时采取控制措施来避免或减少不合格产品的出现。

传统的统计过程控制是基于单变量统计控制的方法。单变量统计控制只能检测单一测量变量的变化，而不能有效地提供关于多个变量相互作用的信息。在现代生产过程中，往往需要测量大量的过程变量。单变量统计过程控制难以在这类过程中有效地发挥作用。另外，单变量统计过程控制也不适用于间歇过程。随着快捷制造技术的推广，适用于生产小批量高附加值产品如医药和精细化工产品的间歇过程，已被越来越多地应用于工业生产中。多变量统计过程控制正是在这些应用需求的驱动下逐渐发展起来的。

多变量统计控制的成功应用能提高产品质量的一致性，提高生产过程的灵活性和能力，进而提高企业资产的有效利用率。通过对过程进行有效的监控，可以提高合格品率，从而降低由于再加工所引起的原材料和能源的消耗。通过对过程进行有效的监控，也能及早发现过程中的故障隐患，从而提高过程运行的安全性。

多变量统计过程控制在西方工业界的应用也刚刚开始，但它已被作为一种提高产品质量和过程运行能力，从而提高企业竞争能力的新技术而得到工业界的高度重视。随着我国改革开放的不断深入，国内企业所面临的竞争也会加剧，提高产品质量，保持产品质量的高度一致性，提高过程的运行能力和灵活性，将是企业在竞争中得以存活和发展的保障。

本书的编写出版得到国家 863 高技术项目基金的支持。清华大学王桂增教授在百忙中认真审阅了全文。空军指挥学院张自维同志、海军航空工程学院马智明同志参加了本书的部分编写、修改工作，徐昕女士为本书的录入付出了辛勤的努力，在此向他们致以诚挚地感谢。

本书的写作目的是介绍多变量统计过程控制技术及其最新进展。由于这一领域的研究成果还在不断出现，加之作者水平有限，掌握的资料不够全面，恳请读者对本书的缺点和不足之处予以批评指正。

作者

2000 年 2 月

内容提要

本书旨在介绍多变量统计过程控制技术及其最新发展动向，内容包括统计基础知识，主元分析、主元回归、部分最小二乘等多变量统计分析方法，单变量、多变量统计控制图，多变量统计分析方法在过程监控、过程故障诊断以及产品质量控制中的应用，统计过程控制中的非线性多变量统计分析方法，以及实验设计等。

全书图文并茂，内容深入浅出，可作为教材，也可作为有关工程技术人员、质量管理人员以及大学相关专业师生的参考书。

目　录

第 1 章　统计过程控制简介 …… 1
1.1　统计过程控制与多变量统计过程控制 …… 1
1.2　统计过程控制的发展与应用现状 …… 2
1.3　统计过程控制的类型 …… 3
1.4　统计过程控制的基本方法 …… 4
1.5　统计过程控制实施的几个阶段 …… 5
1.6　本书的内容安排 …… 6
第 2 章　统计基础知识 …… 7
2.1　总体与样本 …… 7
2.2　样本特征数与总体数字特征 …… 8
2.3　样本的频率分布 …… 10
2.4　常用统计分布 …… 11
2.5　参数估计 …… 14
2.6　假设检验 …… 20
第 3 章　主元分析 …… 24
3.1　主元分析简介 …… 24
3.2　主元分析方法简介 …… 24
3.3　主元的特性 …… 26
3.4　主元分析的几何解释 …… 28
3.5　主元图 …… 28
第 4 章　主元回归与部分最小二乘 …… 31
4.1　模型与模型参数 …… 31
4.2　多元线性回归 …… 31
4.3　主元回归 …… 34
4.4　部分最小二乘 …… 37
第 5 章　单变量统计过程控制 …… 41
5.1　过程变化及其描述 …… 41
5.2　分析过程变化的图形方法 …… 44
5.3　过程能力 …… 47
5.4　统计控制图 …… 49
第 6 章　多变量统计过程控制 …… 58
6.1　单变量统计过程控制的局限性 …… 58
6.2　主元模型 …… 59
6.3　多变量统计控制图 …… 59
6.4　控制限的确定 …… 60

6.5 多向主元分析…… 62
6.6 多变量统计过程控制的应用…… 64
第 7 章 基于主元分析的过程故障诊断 …… 70
7.1 故障诊断方法简述…… 70
7.2 故障诊断的特征方向法…… 71
7.3 连续搅拌反应器故障诊断实例…… 72
7.4 基于统计距离和角度的故障诊断…… 75
7.5 传感器的故障诊断与恢复…… 77
第 8 章 基于主元分析的过程控制 …… 80
8.1 主元分析与过程控制…… 80
8.2 基于主元回归与部分最小二乘的软传感器…… 80
8.3 主元控制器…… 84
第 9 章 非线性多变量分析方法 …… 86
9.1 主元曲线与主元曲面…… 86
9.2 自相关神经元网络…… 88
9.3 输入训练神经元网络…… 89
9.4 多项式非线性部分最小二乘…… 92
9.5 神经元网络非线性部分最小二乘…… 94
第 10 章 非线性多变量统计过程控制…… 97
10.1 非线性多变量统计控制图 …… 97
10.2 累积非线性主元图 …… 98
10.3 非线性统计过程控制在聚合反应过程中的应用…… 100
第 11 章 实验设计 …… 104
11.1 响应曲面方法…… 104
11.2 方差分析…… 109
11.3 因子设计…… 115
11.4 Taguchi 方法 …… 117
第 12 章 统计过程控制的发展方向 …… 120
附录…… 122
参考文献…… 131

第 1 章　统计过程控制简介

1.1　统计过程控制与多变量统计过程控制

18、19 世纪的工业革命改变了产品的生产方式，机器加工取代了手工操作，使得工厂可以借助机器大批量地生产产品。但由于工厂的最终产品往往是由不同工人在不同工序所加工的零部件组合形成的，使得通过大批量生产方式所形成的产品缺乏质量保障。随着工业技术的进一步发展，到了 20 世纪初，为满足大规模生产的需要，要求不同工人生产的各种零部件能达到高度的一致性和协调性。正是这种对产品质量高度一致性的需求，促成了统计过程控制即 SPC（Statistical Process Control）技术的出现。

统计过程控制开始于 70 多年前美国休哈特博士（W.A.Shewhart）的第一张质量控制图，因而从一开始，SPC 就被看作一种提高产品质量和生产效率的技术手段。从质量控制的角度来看，统计过程控制又被称为统计质量控制 SQC（Statistical Quality Control）。由于产品质量在现代工业中的重要地位，使统计过程控制已经在机械 、纺织、汽车、电子产品等离散制造业得到了广泛应用，并正逐渐向造纸、炼油、化工、食品等间歇工业和连续制造业渗透。

传统的统计过程控制以概率论和数理统计为基础，以提高产品质量水平为目标，采用统计控制图、统计描述、统计相关分析、实验设计、回归分析等方法，分析处理与产品质量相关的生产过程数据。其成功应用大都集中在离散制造业中。由于连续生产过程本身的复杂性，其产品质量往往涉及到具有相关关系的几十、甚至上百个变量，这些变量在一段时间的采样数据量之大，使得传统 SPC 在该领域的应用受到限制。

传统的统计过程控制采用单变量统计过程控制方法，只对生产过程中的一些重要指标单独地实施统计过程控制，比如为这些指标单独地建立 Shewhart 控制图。在统计过程控制的应用早期，由于受测量技术以及数据存储和分析技术的限制，人们只能测量生产过程中少数几个重要指标，并对这几个指标单独进行统计过程控制。这在某种程度上能够改进产品质量。但由于一些更重要的产品性能指标往往不能测量，只让所测量的少数几个重要指标分别保持在规定的范围内，并不能真正保证产品的高质量和高性能。

随着测量技术的发展，人们已经能够对越来越多的产品性能指标进行测量，同时用户对产品性能的定量要求也越来越严格。这就要求对许多产品性能指标和过程变量进行监视。如果需要监视的多个产品性能指标或多个过程变量之间存在相关关系，则仅靠分别对它们采用单变量统计过程控制，其结果往往不太可靠，需要采用变量关联图作进一步的监视。若某个装置的生产过程有 100 个测量变量，就需要监视 100 个变量趋势图，如果还想要监视变量间的关联图，这 100 个测量变量的两两关联图就有 4950 幅！过程操作人员很难同时监视这么多图形中变量的变化，需要引入多变量统计过程控制技术来改进对过程的监视。

随着近年来计算机系统、数据库系统的普及应用，使工厂拥有了相当丰富的生产数据资源，提出了采用多变量统计分析方法对大量测量控制数据、产品质量数据等进行处理的应用需求，目的是通过生产数据分析来揭示、反映过程的内在变化，为提高产品质量提供有用信

息，从而把数据资源的拥有优势转化为生产效益和产品质量优势。

将多变量统计分析方法融入传统的统计过程控制，形成了多变量统计过程控制 MSPC 的基本框架。它采用多元投影方法，将过程数据和质量数据从高维数据空间投影到低维特征空间，所得到的特征变量保留了原始数据的特征信息，摒弃了冗余信息，是一种高维数据分析处理的有效工具。在数据量大、数据维数高、变量间具有相关性的连续过程中，MSPC 主要用于实现统计质量控制、过程监控、生产数据的分析挖掘、故障诊断等。多变量统计分析方法包括有主元分析 PCA（Principal Component Analysis）、部分最小二乘 PLS（Partial Least Squares）、主元回归 PCR（Principal Component Regression）等。由于在实际的连续过程中，变量间的非线性关系普遍存在，由此又发展到把非线性多变量分析方法引入到多变量统计过程控制之中，致使今天的 MSPC 已经形成为一个具有众多研究热点的学科方向。可以认为，多变量统计过程控制是把主元分析、部分最小二乘等多元统计投影方法融入传统的统计过程控制而形成的、对存在多个相关变量的生产过程进行监控、分析、控制的方法与技术。

在统计过程控制中，取得数据十分重要。需要得到关于过程输入、输出、产品质量的数据，也需要得到过程运行情况的数据，统计过程控制正是通过对这些数据的统计方法分析，发现过程变化，并追寻引起变化的原因。基于事实和数据的决策与基于主观感觉的决策往往有很大的差别，对数据的简单统计分析往往就能获得较大的收益。通过对过程数据的统计分析，可以完成以下工作：

① 了解过程目前的运行状态，并预测可能出现的情况；

② 对现有过程所能达到的质量指标做出评估；

③ 告诉人们什么时候、什么条件下过程出现了异常，应该去寻找影响过程运行的异常因素；

④ 告诉人们过程的哪些方面可能出现了不正常情况；

⑤ 根据所了解的过程运行状况，进而改进过程及产品质量。

还应该指出的是：统计分析的结果需要用过程操作人员的经验、知识进行解释，需要根据过程的机理去理解、运用。因此数据的采集、统计分析计算、图表只是统计过程控制的一部分，更重要的是需要有管理、生产人员的参与。

统计过程控制是用来促进和保持企业健康生产的有效工具，但并不是医治生产中所有弊病的万能药。它可以帮助人们认识和了解工业过程中存在的问题，认识生产过程的内在特性、变化规律及寻找生产过程发生异常的原因。统计过程控制还用于对生产过程进行控制、对过程进行再设计，进而改进现有的生产过程。

1.2 统计过程控制的发展与应用现状

本世纪初，统计科学的发展为统计过程控制技术的出现奠定了基础。美国贝尔实验室的 Shewhart 博士在 1924 年 5 月绘出世界上第一张质量控制图。在随后的几年里，他和他的同事不断进行这方面的研究。1929 年，贝尔电话公司的道奇（H.F.Dodge）和罗米格（H.G.Romig）发表了论文《抽样检查方法》，提出用抽样检查代替全数检查的方法来保证产品的质量。1931 年，Shewhart 博士又发表了他的经典著作《工业产品质量的经济控制》。

第二次世界大战期间，统计过程控制在美国和英国得到广泛应用，当时主要应用在军工生产领域。二战时英美军队需要大量高质量的军需品。军需品生产商必须充分利用一切新技

术来保证大批量生产中的产品质量。SPC 被成功用于军工生产后，实现了在增加产量和降低成本的同时，明显提高产品质量的目标。

二战摧毁了日本经济，战后的日本工业处于百废待兴的局面，非常迫切需要学习英美的先进技术。1950 年，日本科学家和工程师协会邀请美国统计学家 Deming 博士在日本做了为期 8 天的关于质量控制的讲座。他讲解了采用 Deming 周期（计划 、行动、检查、修正）在统计中辨认变化的重要性，以及如何应用统计控制图。1954 年，另一位美国质量专家 Juran 博士又为日本很多大公司的中、高层管理人员讲 解了管理层的领导作用在质量系统中的重要性。这些思想与方法在战后复苏的日本得到了广泛应用，使其产品质量明显提高，所生产出的高质量低成本的电子、汽车等产品迅速占领了世界市场，一跃成为经济强国。

统计过程控制在日本的成功应用又引起了西方国家对这一技术的重新认识。美国通用电器公司的费根堡姆（A.V.Feigenbaum）和质量管理专家朱兰（J.M.Juran）又提出了全面质量管理 TQM（Total Quality Management）理论，把质量控制推向了新的高度，它与 SQC（Statistical Quality Control，统计质量控制）一起，被列为质量控制发展的两个阶段。

在近年来激烈的市场竞争中，为适应连续制造业寻求产品质量提高的需求，在传统统计过程控制和多变量统计分析方法研究的基础上逐渐形成了多变量统计过程控制。由于连续过程的特点及其相关数据处理的复杂性，使得多变量统计过程控制从理论方法到实际应用，都还有许多问题有待研究解决，因而多变量统计过程控制属于尚在发展过程中的新技术。

统计过程控制今天已得到广泛的工业应用。单变量统计过程控制方法（如 Shewhart 控制图、累积和图、指数加权平均图）以及一些质量控制和质量管理方法（如 Taguchi 方法等）已被工业界广为接受，多变量统计控制方法也开始进入工业应用。目前所报道的多变量统计控制在工业生产中的应用主要集中于北美和欧洲，这和多变量统计过程控制的研究主要集中在北美和欧洲有关。Miller 等（1993）报告了美国柯达公司应用多变量统计控制的情况，并提出了贡献图方法。Kosanovich 和 Piovoso（1995）报告了美国杜邦公司应用多变量统计控制的情况。Wise 等（1999）报告了将多向主元分析、三线性分解及平行因素分析应用于美国德克萨斯仪表公司的半导体蚀刻过程的情况。Martin 等（1999）报告了多变量统计过程控制 在欧洲的应用情况。这些应用基本上是通过多变量统计分析方法来对生产过程进行监控，及时找出产生不正常运行情况的原因或故障。通过及时排除这些故障，可以提高过程能力和产品质量的一致性。

关于多变量统计控制的商业软件也逐渐增多了。进行多变量统计分析的一个很好的软件包是美国特征向量研究公司（Eigenvector Research）出的在 MATLAB 软件下运行的 PLS 软件包（Wise 和 Gallagher，1998）。该公司的创办人 Wise 博士从 80 年代后期就从事多变量统计分析研究，并于 90 年代初推出了 PLS 软件包。经过 10 多年的不断更新，PLS 软件包已被广大工业界人士所接受。英国的 MDC 技术公司（MDC Technology）同英国纽卡斯尔大学化工系合作，于近期推出了 MSPC + 软件包（Hawkins 和 Wood，1999）。该软件融入了纽卡斯尔大学化工系在多变量统计控制方面的一些研究成果。瑞典的 UMETRICS 公司于最近推出了 SIMCA4000 软件包，该软件包可用于在线多变量统计过程控制。

1.3 统计过程控制的类型

统计过程控制主要针对过程的平均水平及过程的分散度进行控制，而过程的分散度往往是影响产品质量的主要因素。在生产过程中，产品质量受到以下五大因素的影响：原材料、

设备、操作方法、操作人员、环境。这些因素的变化往往会引起产品质量的波动。在统计过程控制中，我们将产品质量定义为过程输出，而将原材料、设备、操作方法、操作人员、环境等影响质量的因素称为过程输入。

引起质量波动的原因可分为偶然因素和系统因素两大类。由偶然因素造成的质量特征值的随机波动称为正常波动。当仅有偶然因素存在时，产品质量处于正常波动范围，可以认为生产过程处于受控状态；由系统因素造成的质量特征值的波动称为异常波动。当系统因素的影响使质量特征值偏离规定的范围时，则认为生产过程处于非受控状态（失控状态）。通过统计过程控制，判断出生产过程是否处于受控状态。当过程出现非受控状态时，再进一步找出异常因素并消除它们对过程的影响，达到提高产品质量的目的。

统计过程控制大致可分为以下两类。

① 筛选型统计过程控制。通过抽样检查检测过程输出，将不合格的产品筛选出来进行再加工，或作为低档产品降价出售，或作为废品。采用这种方法控制产品质量的过程被称为筛选型统计过程控制。

② 预防型统计过程控制。这是一种试图通过过程控制防止不合格产品产生的方法。通过运用过程变量的各种控制图、抽样检查生产原料等手段，监测并调整影响产品质量的各过程输入，预防不合格产品的发生。

统计过程控制又有在线与离线之分。通常把筛选型统计过程控制和预防型统计过程控制称为在线统计过程控制。

离线统计过程控制通过修改过程来减少或消除引起产品质量变化的因素，从而达到控制产品质量的目的。最好能将离线统计过程控制与产品设计、过程设计结合起来进行。这往往需要包括统计学家在内的具有不同专业技能的人共同参与才能完成。这类统计过程控制不太可能给出比较通用的规则，即对于不同的过程及背景，所应采用的离线统计过程控制方法往往不一样。在某些情况下，离线统计过程控制可能与在线统计控制关系不大。但在多数情况下，在线统计过程控制对改进产品质量起着关键的作用。离线统计过程控制很大程度上需要应用实验设计方法，如 Taguchi 方法。本书第 11 章将对实验设计方法作具体介绍。

Becknell（1987 年）报告了在福特汽车公司进行的关于节流阀体的实验。这些节流阀体由压铸而成。虽然产品质量在很多方面都满足要求，但部件常出现影响外观质量的气隙。通过研究讨论后，工程技术人员选出了七个可能影响产品外观质量的因素进行实验。这七个因素包括金属纯净度、固体杂质的尺寸、喷涂模式、强化度、靠模速度、模具温度以及金属温度。每个因素定为两个层次。通过实验找出了这些因素的优化组合。结果使能见气隙减少了 73%，年节约超过了大约 30 万美元。

1.4 统计过程控制的基本方法

当从一个过程采集的数据服从单一分布（通常是正态分布）并具有一些理想的特性，如产品性能指标符合规定时，说明这个过程处于受控状态，其均值、方差的大小及分布曲线形状保持一致。当过程中存在系统原因引起的变化，即这个过程处于失控状态时，质量数据的均值、方差及分布曲线的形状会发生变化。图 1.1 分别表示了均值、方差及分布曲线形状发生变化时的曲线。

图 1.2 为一个简单的 Shewhart 控制图。它按时间顺序将质量或过程测量数据的子组均值画在图上，同时画出上下控制限而形成。只要过程的变化保持在限度之内，就可以认为其

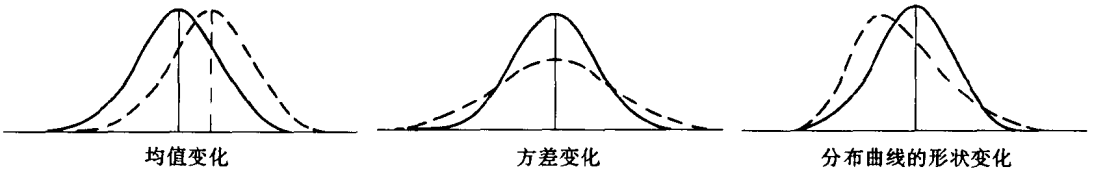

图 1.1　过程数据的均值、方差及分布曲线形状的变化

变化是正常的且过程在统计控制之下。假如画的点超出了任何控制限，变化就不正常了，反映有系统因素开始对过程起作用了。从 Shawhart 控制图能辨别出过程是否处于受控状态，可以通过它监视生产运行参数和质量指标的变化，分析生产过程状态。

图 1.2　简单的 Shewhart 控制图

在统计过程控制中，把过程生产出的产品能满足质量要求的能力称为过程能力，过程能力实际上表明了质量变化与生产规格相比较的符合程度。图 1.3 表明了受控状态下过程能力的区别。它在数据的分布曲线图上添加了对合格品所规定的技术指标的上下限。尽管这些过程都处于受控状态，但它们所表现出来的过程能力有较大差异。前面三条曲线代表了过程处于受控状态但过程能力不足，有一小部分产品超出了合格品的上限指标。后面三条曲线表现出过程具有充足的过程能力，所有的产品都分布在规定的界限之内。显然，后三条曲线的方差小于前三条曲线的方差。一般情况下，质量指标分布的方差越小，质量波动的范围越小，过程能力就越高。

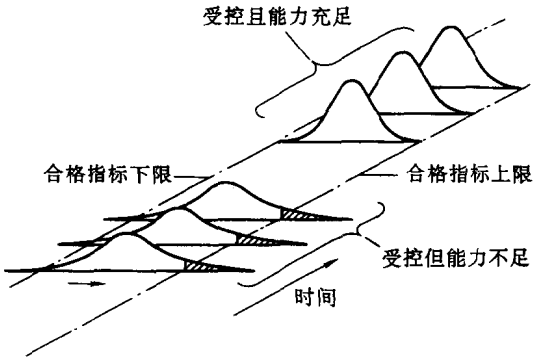

图 1.3　受控状态下过程能力的区别

1.5　统计过程控制实施的几个阶段

连续制造业的生产过程往往包括多个生产装置和自动控制回路，包括有物流的再循环以及多阶段的混合。通常一个典型的过程会有十几个描述产品质量的变量，大约有 200 到 500 个过程参数或变量，仅描述原材料质量与数量的变量就可多达 30 多个，且这些变量之间的关系往往不太明确。处理复杂系统的一个关键是将其分解为若干模块。

对于这类过程的统计过程控制一般要按以下三个阶段进行。

第一阶段：构画过程流程。

① 首先需要画出过程流程图，并标注组成过程的各个阶段。

② 研究过程中的数据流向与数据储存。

第二阶段：确定问题。

① 收集工程技术人员及用户对产品质量问题的看法。

② 确定重要的产品变量，不管它们是否被测量。

③ 收集关于这些变量的数据，并用移动平均图、累积和图以及过程能力分析等方法研究这些变量的数据。

④ 计算不合格产品的成本。

⑤ 通过与过程工程技术人员以及过程操作员的讨论来解释数据，确定生产过程的问题所在。

第三阶段：过程探索。

① 收集关于过程的信息，包括技术报告，过程操作人员的观点、推测等。

② 将过程分解为若干模块，并决定还需要哪些额外数据。

③ 通过质量控制或其他操作途径来采集数据。

④ 利用图表、累积和图、多变量回归，以及多变量统计分析方法来分析和解释数据。

⑤ 结合产品设计、过程设计和实验设计对生产过程进行实验，测试并建立经验模型或理论模型。

⑥ 选定统计过程控制图并确定采用哪些变量。

⑦ 实施统计过程控制。

在实施统计过程控制中，实验设计阶段往往是必不可少的，因为没有它我们往往不具备进行统计过程控制的数据与知识。而多变量统计分析方法，如主元分析等，用以追踪过程变化的原因。

1.6 本书的内容安排

本书第 2、3、4 章为基础部分，分别介绍统计基础知识、主元分析、主元回归与部分最小二乘。第 2 章介绍均值、方差等基本统计概念、常用统计分布，以及假设检验等。第 3 章和第 4 章介绍多变量统计分析方法，包括主元分析、主元回归与部分最小二乘等。这两章将为学习多变量统计控制打下基础。

第 5 章介绍传统的单变量统计过程控制，包括各种控制图及其应用。第 6 章介绍多变量统计过程控制，包括主元模型、多变量统计控制图、适用于间歇过程的多向主元分析及应用实例。多变量统计控制的一个重要应用是过程运行情况监控及过程故障诊断。第 7 章将专门介绍近几年提出来的应用主元分析进行过程故障诊断的方法。第 8 章介绍如何运用多变量统计分析方法进行产品质量控制。

由于很多比较复杂的实际过程，特别是间歇过程，往往具有很强的非线性，对这类过程采用非线性多变量统计分析方法将更为有效。第 9 章专门介绍非线性多变量分析方法，包括非线性主元分析和非线性部分最小二乘等。第 10 章介绍基于非线性多变量统计分析的统计过程控制。

在多变量统计过程控制中，数据的产生、采集及处理将是十分重要的。第 11 章介绍实验设计方法，如 Taguchi 方法等。

第 2 章　统计基础知识

数理统计是统计过程控制的基本理论基础，是通过样本来了解和判断总体的统计特性的科学方法。本章简要介绍数理统计的基础知识。

2.1　总体与样本

总体与样本是统计中最基本的概念。

(1) 总体

研究某个问题时，其对象的所有可能的观测结果称为总体，或把研究问题的全体称为总体。组成总体的每个元素称为个体。描述个体特性的指标称为个体的属性。例如，一批仪表为一个总体，而每块仪表为此总体中的一个个体，而量程、精度、使用寿命等指标为仪表的属性。

把仪表精度作为所研究问题的总体。每块仪表的精度是不完全相同的，从概率论的观点看，仪表精度是一个随机变量，可用随机变量 X 表示，这批仪表中每块仪表的精度值就是随机变量 X 的一个个具体的取值。也就是说，随机变量 X 是定义在总体上的一维随机变量。

对总体的研究主要从总体的分布特性和统计特性入手。总体分布特性的研究是指对总体(随机变量 X) 的分布规律所进行的研究。只要掌握了 X 的分布规律，就对 X 有了全面的了解和掌握。总体统计特性的研究是对随机变量 X 的期望值 (均值)、方差等进行的研究，这是从宏观上把握总体的一个简单、有效的方法，特别适用于对总体分布不太了解或对总体分布的了解要求不高的场合。

对总体分布特性的研究主要表现在两个方面。一是结构方面的研究，即对随机变量 X 的类型和分布规律所进行的研究，是属于正态分布、指数分布还是其他类型的分布。随机变量 X 的分布规律往往用 X 的概率密度 $p(x)$ 或分布函数 $F(x)$ 来描述。例如，正态随机变量 $X\sim N(\mu,\sigma^2)$ 的概率密度 $p(x)$ 为

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.1}$$

其中，σ 为均方差，μ 为均值。μ 和 σ 是正态随机变量 X 的两个重要的分布参数。分布函数 $F(x)$ 为

$$F(x)=\int_{-\infty}^{x}p(x)\mathrm{d}x=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x \tag{2.2}$$

根据经验和理论的研究，对某些实际问题的分布规律一般是了解的，例如测量误差多服从正态分布，产品寿命常服从威布尔分布或对数正态分布等。问题是对这些分布中的参数不太了解，若分布参数确定了，则分布规律就可确定。例如正态分布中当均值 (μ) 和方差 (σ^2) 确定以后，其概率密度和分布函数就可确定。因此，求取分布参数构成了对总体研究的另一个重要方面。

在统计控制过程中，主要通过对数据的分析和研究来推断总体特性。

（2）样本

统计中经常采用的方法是从总体中抽取一部分个体进行观察，然后根据得到的数据来推断总体的特性。被抽出的部分个体，称为总体的一个样本。若对总体 X 进行 n 次抽样观测，得到 n 个观测值 x_1、x_2、…、x_n，其中 x_i（$i=1, 2, \cdots, n$）表示第 i 次观测所得到的数据，称 x_1、x_2、…、x_n 为来自总体 X 的样本值，样本中样本值的个数称为样本容量。抽样的目的是通过对样本值的研究来推断总体的特性。

要使样本能尽量多地包含总体的有关信息，对样本最基本的要求是具有代表性，这就对样本的获取方法提出了一定的要求。如果总体中的每个个体被抽取的机会相等，且在抽取一个个样本后，总体的成分不变，则所得到的样本一定具有很好的代表性。也就是要求抽取样本时，一是要随机，二是要独立。为了保证在抽取一个个样本后总体的成分不变，对于有限总体来说，应该是有放回地抽取，即每次抽取后，测量有关指标，然后放回到总体，以便下次随机地抽取；对于无限总体（包括虽为有限总体但数量众多）来说，由于抽取样本后总体成分变化不大，故可不放回。用此方法得到的样本，称为简单随机样本。对于生产过程的测量值来说，可以认为总体是无限的，只要保证每次测量的独立性与随机性即可。

由于抽样的随机性与独立性，任意两次观测值一般是不同的，因此，每个观测值 x_i 都可以看作某个随机变量 X_i（$i=1, 2, \cdots, n$）的一次实现。样本值 x_1、x_2、…、x_n 为这些随机变量在一次抽取后的具体取值。

因为容量为 n 的样本可以看成由 n 个独立同分布的随机变量组成的随机向量，连续型随机向量其概率密度可用联合概率密度来描述。若总体 X 的概率密度为 $p(x)$，则（X_1, X_2, …, X_n）的联合概率密度为各个随机变量概率密度的乘积 $p(x_1)p(x_2)\cdots p(x_n)$。

2.2 样本特征数与总体数字特征

（1）统计量与顺序统计量

统计分析的基本依据就是样本。设 X_1、X_2、…、X_n 为来自总体 X 的一个样本，若构造的样本函数 $T=T(X_1, X_2, \cdots, X_n)$ 不含有总体分布中的任何未知参数，则称 T 为一统计量。

例如，总体 $X\sim N(\mu, \sigma^2)$，若期望值和方差（μ, σ^2）均未知，则

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \text{ 和 } S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

就是利用样本构成的两个最常用的统计量——样本均值和样本方差。而这两个统计量又被称为样本特征数。但 $\sum_{i=1}^{n}(X_i-\mu)^2$ 和 $\frac{\sum_{i=1}^{n}X_i}{\sigma^2}$ 就不是统计量，因为它们含有总体分布中的未知参数。

由于 x_1、x_2、…、x_n 是样本 X_1、X_2、…、X_n 的观察值，则 T（x_1, x_2, …, x_n）也可看成样本函数 $T=T(X_1, X_2, \cdots, X_n)$ 的一个观察值。为了描述统计量，常用到 T 的分布、T 的数字特征（均值和方差）等。不同的统计量一般有不同的分布和数字特征。

样本 X_1、X_2、…、X_n 按从小到大顺序重新排列为 $X_{n1}\leqslant X_{n2}\leqslant\cdots\leqslant X_{nn}$，则（$X_{n1}\leqslant X_{n2}\leqslant\cdots\leqslant X_{nn}$）称为顺序统计量。

(2) 样本与总体的数字特征

均值、方差、标准差等都是通常用以描述样本或总体统计特性的数字特征，表 2.1 描述了在样本与总体中这些统计特征之间的对应关系。

表 2.1 样本与总体的统计特征

名　称	样本特征数	总体数字特征
均值	$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$	$\mu=E(X)$(数学期望值)
方差	$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$	$\sigma^2=D(X)$
标准差	$S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$	$\sigma=\sqrt{D(X)}$

(3) 样本特征数

① 样本均值。当不需要精确掌握总体的分布，只要求了解样本数据的中心值及离散程度时，可直接利用样本值进行估计。样本均值可用来刻画样本数据的中心位置，并可直接利用样本均值来估计总体的平均值。样本均值的计算公式为

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{2.3}$$

② 中位数。中位数是刻画数据中心位置的另一个量。将样本值 x_1，x_2，…，x_n依照从小到大的顺序重新排列，记为 $x_{n1}\leqslant x_{n2}\leqslant\cdots\leqslant x_{nn}$，则数据中位数 M^* 定义为

$$\begin{aligned} M^*&=x_{\frac{n+1}{2}} && n\text{ 为奇数时}\\ M^*&=\frac{1}{2}\left(x_{\frac{n}{2}}+x_{\frac{n}{2}+1}\right) && n\text{ 为偶数时}\end{aligned} \tag{2.4}$$

③ 样本方差。根据样本利用以下公式可以估计样本方差

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{2.5}$$

当样本容量 n 较大时，有时也采用

$$S^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{2.6}$$

来计算样本方差。只要 n 比较大，两者之间只有微小的区别。

④ 样本标准差。样本标准差的计算公式为

$$S=\sqrt{S^2}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{2.7}$$

当采用统计描述来说明某组数据的特征时，往往还会用到下面的极差、分位数的概念。

⑤ 极差。数据的极差可用样本数据中的最大值与最小值之差表示，即

$$R=x_{\max}-x_{\min} \tag{2.8}$$

极差表明了样本数据的变化范围。

⑥ 四分位数。四分位数也是刻画数据离散程度的一种方法。将样本数据从小到大重新排序后，分成数目相等的四份，每份各占样本容量的 1/4。从左至右，依次处于第 1、第 2、

第3个分位点上的数据（或处于分位点两侧的数据的算术平均值），称为第1、第2、第3分位数，分别记为 q_1、q_2、q_3。显然，q_2 等于中位数。区间 $[q_1, q_3]$ 的数据量占样本容量的一半，表示该区间大小的值也能反映数据的集中程度。

方差、极差与四分位数都是刻画样本数据离散程度的量。

此外，样本特征数还有样本的 r 阶原点矩、r 阶中心矩等。

2.3 样本的频率分布

利用样本推断总体，这是数理统计的基本方法。总体的分布密度可直接利用样本的频率分布来估计，样本的频率分布能较完整地反映实验数据的变化规律，而求取样本频率分布的一种近似方法是分布密度的图解法——直方图法。

（1）样本分布密度的直方图

直方图是被广泛采用的用以得到样本频率分布的简便方法。其具体作法如下。

① 找出样本数据的最大值与最小值，求得极差

$$R=\max[x_i]-\min[x_i] \tag{2.9}$$

② 根据样本大小把样本值 x_1、x_2、…、x_n分为 m 组，并根据组数 m 和极差决定组距 c。令

$$t_0=\min[x_i], t_m=\max[x_i], t_0<t_1<t_2<\cdots<t_m$$

若按等距离分组，则

$$c=t_i-t_{i-1}=R/m \tag{2.10}$$

式中 m 的大小没有硬性规定，一般来说，当样本容量 n 较小时，m 应小些；反之，m 应取大些。另外，为精确起见，t_i 比样本值多取一位小数。

③ 数出样本值落入每个小区间的个数。把落入小区间 $[t_i-t_{i-1}]$ 的样本个数记为 v_i $(i=1, 2, \cdots, m)$。

④ 计算样本值落入各小区间的频率 f_i

$$f_i=v_i/n \quad (i=1,2,\cdots,m) \tag{2.11}$$

由于样本的抽取是独立的，由概率的统计定义可知，当样本容量 n 足够大且每个小区间足够小时，f_i 近似等于随机变量 X 落入小区间 $[t_i, t_{i-1}]$ 的概率，即

$$f_i\approx P\{t_{i-1}<X\leqslant t_i\} \quad (i=1,2,\cdots,m) \tag{2.12}$$

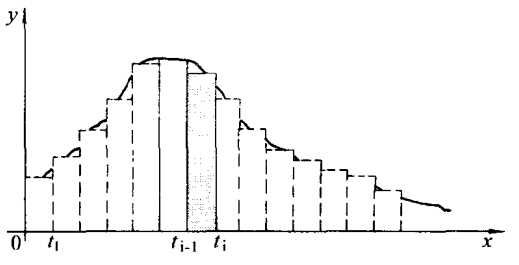

图 2.1 样本直方图

⑤ 画直方图。在 xy 平面上，以小区间 $[t_i, t_{i-1}]$ 的长度为底，以 $\frac{f_i}{(t_i-t_{i-1})}$ 为高作矩形，则得到直方图（见图2.1）。

由于每个小矩形的面积近似等于样本值落入小区间的概率，即

$$\frac{f_i}{t_i-t_{i-1}}(t_i-t_{i-1})=f_i\approx P(t_{i-1}<x\leqslant t_i) \tag{2.13}$$

故直方图就大致地描述了样本的概率密度曲线，而样本与总体是同分布的，因而直方图就大致描述了该样本所代表的总体的概率密度。当样本容量 n 足够大，等分的小区间非常小时，这种描述更接近 X 概率密度。