质量数据分析技术

王敏华　曾其勇　著

中国质检出版社
中国标准出版社

质量数据分析技术

王敏华　曾其勇　著

中国质检出版社
中国标准出版社
·北　京·

图书在版编目（CIP）数据

质量数据分析技术/王敏华，曾其勇著. —北京：中国质检出版社，2016.5
ISBN 978-7-5026-4269-3

Ⅰ.①质… Ⅱ.①王… ②曾… Ⅲ.①数据处理 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字（2016）第 053926 号

内 容 提 要

本书系统地介绍了质量数据收集的目的和方法、质量数据的统计特征、质量数据分布规律、测量系统分析的相关概念和分析方法、质量管理常用数据分析技术、常规控制图和其他控制图分析技术、过程能力分析方法、抽样检验方案的制定、质量检验数据的分析、常用质量数据统计分析软件等质量数据统计分析技术方面的知识。

本书可作为质量专业技术人员、从事质量检验和质量管理工作人员的参考用书，也可作为大专院校相关专业的参考教材。

中国质检出版社
中国标准出版社 出版发行
北京市朝阳区和平里西街甲 2 号（100029）
北京市西城区三里河北街 16 号（100045）
网址 www.spc.net.cn
总编室：（010）68533533 发行中心：（010）51780238
读者服务部：（010）68523946
中国标准出版社秦皇岛印刷厂印刷
各地新华书店经销
*
开本 787×1092 1/16 印张 12 字数 283 千字
2016 年 5 月第一版 2016 年 5 月第一次印刷
*
定价：**32.00** 元

前言
Preface

大数据技术已经成为互联网时代主流技术，企业对数据的应用，是大数据行业发展的引擎。在大数据技术时代，如何推动大数据在质量管理工程中的发展和应用，是摆在质量管理工程人员面前的一项实质性工作。推动质量数据统计分析技术的发展与科研创新的有机结合，对质量领域大数据的使用，有利于提高质量管理水平，为企业的科学决策提供坚实基础。

数据本身不产生价值，当数据经过挖掘、归类和分析，能够给企业决策提供帮助之后，才具有价值。而质量领域的大数据分析涉及质量数据的统计分析技术，通过科学地应用质量数据统计分析方法，提升产品质量水平，产品在市场上才会有较强的竞争能力。

本书系统地介绍了质量数据收集的目的和方法、质量数据的统计特征、质量数据分布规律、测量系统分析的相关概念和分析方法、质量管理常用数据分析技术、常规控制图和其他控制图分析技术、过程能力分析方法、抽样检验方案的制定、质量检验数据的分析、常用质量数据统计分析软件等质量数据统计分析技术方面的知识。本书共分 9 章，内容有：质量数据基础知识、测量系统分析、质量管理常用数据分析方法、控制图分析技术、相关分析与回归分析、其他质量数据分析方法、过程能力分析、质量检验数据分析技术、常用质量数据统计分析软件简介。其中第二章和第九章由中国计量大学曾其勇撰写；第一章、第三章、第四章、第五章、第六章、第七章和第八章由中国计量大学王敏华撰写。全书由中国计量大学王敏华负责统稿。

本书力求有一定的专业深度，理论与实践相联系，把知识性、实用性和系统性结合起来，在撰写过程中，除了引用作者多年在质量数据统计分析技术方面的研究成果之外，还参考了不少质量数据统计分析与质量管理方面的著作及学术论文，在此对所有给予我们帮助的专家表示衷心的感谢！此外还要感谢华立仪表集团股份有限公司为本书提供案例和数据。

由于国内外质量数据统计分析技术不断发展，新方法、新经验层出不穷，质量数据统计分析方面的文献、著作目不暇接，将这些成果消化吸收并纳入到本书中，确实有较大的困难，所以本书可能还存在许多有待完善的地方，衷心希望读者朋友提出宝贵意见。

著者

2016 年 2 月

目录
Contents

第一章

质量数据基础知识

第一节　收集质量数据的目的和方法

一、什么是质量数据

质量数据是指某个质量指标的质量特性值，如一批灯泡寿命、一批轴承的长度、一批炮弹的射程等。质量指标呈现多种多样性，质量数据在质量管理中无处不在。在相同的生产技术条件下生产出来的一批产品，其质量特性数据由于受到操作者、设备、材料、方法、环境等多种因素的影响而存在着一定的差异，但当生产过程处于正常状态时，其质量数据的波动服从一定的统计规律。在质量管理过程中，若不收集数据、无数据的定量分析，也就没有明确的质量概念，就没有科学的统计质量控制理论。在质量工程实践中，强调“一切用数据说话”，就是要尽可能用数据来反映事实，利用数理统计方法分析波动规律，区分正常波动与异常波动，进而控制异常波动；进行相关分析和回归分析以确定最佳工艺参数；采用抽样检验方法得出产品质量信息，作出符合实际的结论和正确的判断。因此，从生产过程中客观地获取有用的数据，进行科学的分析和整理，掌握过程的质量状况，是非常重要的，它是我们针对具体问题采取行动的基础。数据是质量管理的依据，如果没有数据或者没有对其进行定量分析，就不能找到研究对象的客观规律。因此，研究数据的统计规律具有极其重要的意义。

在质量管理数据分析过程中，特别关注三个方面的问题：一是数据的集中位置；二是数据的分散程度；三是数据的分布规律。

二、质量数据收集的目的

收集质量数据的主要目的是：

（1）掌握和了解现场质量状况，分析质量问题；

（2）对过程质量进行分析，判断过程是否处于受控状态；

（3）在质量改进过程中，选择最佳的工艺参数；

（4）对产品质量进行评价和验收。

三、质量数据的分类和特点

根据质量数据的特点，质量数据可分为计数值和计量值两大类。

1. 计数值

当质量特性值只能取一组特定的数值，如0，1，2，…，而不能取这些数值之间的其他数值时，这样的特性值称为计数值。它属于离散型变量。计数值可进一步分为计件值和计点值。计件值是指产品按件检查时所产生的属性，如一批产品中的合格品数、废品数等；计点值是指每件产品中质量缺陷的个数，如棉布上的疵点数，铸件上的砂眼数等。

2. 计量值

当质量特性值可以取某一范围内的任何一个可能的数值时，这样的特性值称之为计量值。它属于连续型变量。用各种计量工具测量得到的数据通常是计量值，如长度、位移、温度等。不同类型的质量特性值其统计规律是不同的，因而采用的控制方法也不同。

四、质量数据的收集方法

质量数据的收集通过全数检验和抽样检验获得。为了更好地理解全数检验和抽样检验的概念，我们需要知道总体与样本的相关概念。

（一）总体与样本

1. 总体

GB/T 3358.2—2009《统计学词汇及符号　第2部分：应用统计》中关于总体的定义是：所研究个体/单位产品的全体。

总体的表现形式如下：

（1）总体可以是真实的、有限的，如一批产品的寿命。

（2）总体也可以是虚构的，如基于某一概率分布的总体。

（3）总体可以是一个包括未来产出的，正在进行中过程的结果。

（4）总体可由可区分的物体组成，也可由散装材料组成。

通常总体的单位数用 N 来表示。

2. 个体

个体是指：能被单独描述和考虑的一个事物。

若总体是一批产品，则个体相当于单位产品。如一个分立的物品、一定量的散料等。

3. 抽样单元

抽样单元是指：将总体进行划分后的每一部分。

抽样单元可以包含一个或多个个体。

4. 样本

样本是指：由一个或多个抽样单元构成的总体的子集。

从总体中抽取部分个体所组成的集合称为样本。样本中所包含的抽样单元的数目称为

样本量。样本量也称为样本大小或样本容量，用 n 表示。

满足下面两个条件的样本为随机样本。

（1）随机性；

（2）独立性。

（二）全数检验

全数检验即100%检验，按GB/T 3358.2—2009《统计学词汇及符号　第2部分：应用统计》，100%检验是指：对所考虑的产品集合内每个单位产品被选定的特性都进行的检验。

全数检验就是对总体中的全部个体逐一观察、测量、试验或估量，从而获得对总体质量水平评价结论的方法。

（三）抽样检验

抽样检验是指从所考虑的产品集合中抽取若干单位产品进行的检验。

常见的抽样方法有以下几种。

1. 简单随机抽样

简单随机抽样又称纯随机抽样、完全随机抽样，是对总体不进行任何分类，完全按随机的方法抽样而获取样本的方法。如将全部产品编号，同时将这些编号写在纸条上，纸条充分混合后采用随机方式抽取，根据纸条的编号再找出对应的产品，这样逐一抽取直到所需的产品数。按这种方式抽样，总体中每个个体都有同等的机会被抽入样本，这样得到的样本称简单随机样本。

2. 分层抽样

分层抽样又称分类或分组抽样，是将总体按某一特性分为若干组，然后在每组内随机抽取样品组成样本的方法。

3. 等距抽样

等距抽样又称机械抽样、系统抽样，是将总体 N 按某一特性排队编号后均分为 n 组，这时每组有 $K = N/n$ 个个体，然后在第一组内随机抽取第一件样品，以后每隔一定距离（K 个）选取其余样品，直到抽出 n 件产品组成样本。

4. 整群抽样

整群抽样一般是将总体按自然存在的状态分为若干群，并从中抽取一群或多群，这些选中的群内所有产品组成样本，这种抽样方法称为整群抽样。如对原材料质量进行检测，可按原包装箱或原包装盒为群随机抽取，对选中的箱或盒做全数检验。

由于数据的随机性表现在群间，而整群抽样时样品集中，分布不均匀，因此，代表性差，产生的抽样误差也大，同时在有周期性变动时，也应注意避免系统偏差。

5. 多阶段抽样

多阶段抽样又称多级抽样。以上四种抽样方法的共同特点是整个过程中只有一次随机抽样，因而统称为单阶段抽样。但是当总体很大时，很难一次抽样完成预定的目标。多阶

段抽样是将各种单阶段抽样方法结合起来使用，通过多次随机抽样来实现的抽样方法。如在钢材和水泥等产品质量检验时，可以对总体按不同批次分为 w 群，从中随机抽取 w 群，而后在其中选的 w 群中的 M 个个体中随机抽取 m 个个体，这就是整群抽样与分层抽样相结合的二阶段抽样，它的随机性表现在群间和群内共有两次。

第二节　质量数据的统计特征

在质量管理过程中，我们需要研究总体的均值、方差（或标准差）是多少？其中，均值是代表平均水平或数据集中程度的量，故也称为平均值，方差和标准差则是表示总体离散程度的量。这些量都是质量数据的统计特征值。通常，总体均值用 μ 表示，总体方差用 σ^2 表示，总体标准差则用 σ 表示。由于实际生产过程产品数量很多，因此，要得到总体的均值和方差不现实，尤其当某一产品质量特性需采用破坏性检验才能得到时，这时就通过抽取若干个产品组成样本，计算样本的数据来近似估计总体的对应参数。用样本描述质量状况的量也分为两类，一类是代表数据集中程度的量，主要有样本平均值和样本中位数；另一类是表示数据离散程度的量，主要有样本极差、样本方差和样本标准差。

一、表示数据集中程度的量

1. 样本平均值 $\bar{X}$

设从某总体中抽取一个样本量为 n 的样本，得到 n 个质量数据，分别为 $X_1, X_2, \cdots, X_n$（X_1，X_2，…，X_n 表示随机变量，其对应的观测值用 $x_1, x_2, \cdots, x_n$ 表示。但在质量数据统计分析领域，不像概率统计那样区分大小写），则它们的平均值为：

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{1-1}$$

样本平均值也称样本均值，在质量统计中，常常用样本均值来估计总体均值。

2. 样本中位数

样本中位数是将样本数据按由小到大顺序排列，若数据个数 n 为奇数时，位于中间位置的那个数称为样本中位数，通常用 $\tilde{X}$ 或 Me 表示。若 n 为偶数时，样本中位数为两个中间位置数据的平均值。即：

设 $X_1, X_2, \cdots, X_n$ 为从某总体中抽取的一个样本，若将样本的数据从小到大排列为 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$，称 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ 为有序样本，则样本中位数为：

$$\tilde{X} = \begin{cases} X_{\left(\frac{n+1}{2}\right)}, \text{当 } n \text{ 为奇数} \\ \dfrac{1}{2}\left[X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}\right], \text{当 } n \text{ 为偶数} \end{cases} \tag{1-2}$$

在现场产品质量控制中，对生产的产品随机抽若干个进行观察，若计算其平均数较麻烦，只要看中位数的大小就可知道其大致水平如何了。

例 1-1　为了得到某一零件的直径，抽取 5 件产品进行测量，得到如下测量数据：15.09，15.29，15.15，15.07，15.21（单位：mm），计算样本的均值和样本的中位数。

解：

样本均值为：

$$\bar{X} = \frac{1}{5}(15.09 + 15.29 + 15.15 + 15.07 + 15.21) = 15.162\ (\text{mm})$$

将数据 15.09，15.29，15.15，15.07，15.21 按从小到大的顺序排列为：

15.07，15.09，15.15，15.21，15.29

则样本中位数 $\bar{X} = 15.15$ (mm)。

总体均值可用样本均值或样本中位数估计，本例中，用这两种方法估计总体均值得出的结果不完全一致，但差别不是太大。一般情况下，用样本中位数估计总体均值时误差会大些，因为在计算样本中位数时采用了简单的计算方法，只用到有序样本中的一个或两个数据。

二、表示数据离散程度的量

1. 样本极差

样本极差是样本数据中最大值 X_{max} 与最小值 X_{min} 的差值，用 R 表示。

$$R = X_{max} - X_{min} \tag{1-3}$$

2. 样本方差

样本 $X_1, X_2, \cdots, X_n$ 偏离样本均值的平方的平均值称为样本方差，其数学符号是 s^2。样本方差表示 n 个质量数据 $X_1, X_2, \cdots, X_n$ 的分散程度，其计算公式为：

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \tag{1-4}$$

在统计质量控制中，用样本方差来估计总体方差。当计算精度要求较高时，我们可以用 s^2 来表示数据的离散程度。

3. 样本标准差

样本标准差用 s 来表示，其计算公式为：

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} \tag{1-5}$$

例 1-2　计算例 1-1 中 5 个数据的样本极差、样本方差和样本标准差。

解：样本极差：

$$R = X_{max} - X_{min} = 15.29 - 15.07 = 0.22\ (\text{mm})$$

样本方差：

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

$$= \frac{(15.09-15.162)^2 + (15.29-15.162)^2 + (15.15-15.162)^2 + (15.07-15.162)^2 + (15.21-15.162)^2}{5-1}$$

$$= \frac{(-0.072)^2 + (0.128)^2 + (-0.012)^2 + (-0.092)^2 + (0.048)^2}{4} = 0.00812$$

样本标准差：

$$s=\sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}}=\sqrt{0.00812}=0.0901$$

注意：总体方差可用样本方差来估计，但总体的标准差不能直接用样本标准差或样本极差来估计，总体标准差的无偏估计由 $\hat{\sigma}=\frac{\bar{R}}{d_2}$ 或 $\hat{\sigma}=\frac{\bar{s}}{c_4}$ 得到，其中 $\bar{R}$ 和 $\bar{s}$ 分别表示样本极差的均值和样本标准差的均值，参数 d_2 和 c_4 由本书第四章给出。

第三节　质量数据分布规律

质量数据的分布有多种形式，本节将介绍几种比较重要的分布。

一、离散型分布

（一）二项分布

假如一批产品分成合格品和不合格品，若不合格品率为 p，每次抽取一个，观察后放回去，下次再取一个，共抽取 n 次，在 n 次试验中，不合格品的个数 X 是一个随机变量，它可以取 0，1，2，…，n 共 $n+1$ 个可能值。一般地，n 次随机试验组成的随机现象，满足条件：

（1）重复进行 n 次随机试验。

（2）n 次试验间相互独立。

（3）每次试验仅有两个可能结果：A 发生或不发生。

（4）每次试验事件 A 发生的概率为一常数 p。

在 n 次随机试验中，事件 A 发生的次数 X 为一随机变量，其概率函数为：

$$P(X=k)=\binom{n}{k}p^kq^{n-k},\quad k=0,1,2,\cdots,n \tag{1-6}$$

式中，$q=1-p$，$\binom{n}{k}=\frac{n(n-1)(n-2)\cdots(n-k+1)}{k(k-1)(k-2)\cdots1}=\frac{n!}{(n-k)!k!}$。在这里 $P(X=k)$ 的值恰好是二项式 $(q+px)^n$ 展开式中第 $k+1$ 项 x^k 的系数，故称为二项分布。二项分布中包含两个参数 n 和 p，我们将参数为 n、p 的二项分布记作 $b(n,p)$。当随机变量 X 服从参数为 n、p 的二项分布时，记作 $X\sim b(n,p)$。

二项分布的数学期望 $E(X)$ 和方差 $V(X)$（随机变量 X 的方差有时记为 $Var(X)$）分别为：

$$E(X)=np$$

$$V(X)=npq$$

例 1-3　在一大批产品中，不合格品率为 0.1。现从成品中随机取出 6 个，记 X 为 6 个成品中的不合格品数，求：

（1）6 个全是合格品的概率；

（2）二项分布 b（6，0.1）的均值、方差与标准差。

解：（1）6 个全是合格品的概率为：

$$P(X=0)=\binom{6}{0}\times 0.1^{0}\times(1-0.1)^{6-0}$$

$$=1\times 1\times 0.9^{6}=0.5314$$

（2）均值为：$E(X)=np=6\times 0.1=0.6$

方差为：$V(X)=np(1-p)=6\times 0.1\times 0.9=0.54$

标准差为：$\sigma(X)=\sqrt{np(1-p)}=\sqrt{0.54}=0.73$

（二）超几何分布

例 1-4　某批产品共有 20 件，其中有 8 件产品为优质品，今采用不重复的方式从该批中任选 5 件，被选到的优质品数 X 是一个随机变量，求 X 的分布。

解：X 可以取 0，1，2，3，4，5 这 6 个值，相应的概率函数为：

$$P(X=k)=\frac{\binom{8}{k}\binom{12}{5-k}}{\binom{20}{5}},\quad k=0,1,2,3,4,5$$

在例 1-4 中如果优质品只有 4 件，那么 X 只取 0，1，2，3，4 这 5 个值，其他不变。将这一问题一般化，设 N 个元素分为两类，有 M 个属于第一类，$N-M$ 个属于第二类。从中按不重复抽样取 n 个，若 X 表示这 n 个中第一类元素的个数，则 X 的分布称为超几何分布。其概率函数为：

$$P(X=k)=\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{M}{n}},k=0,1,2,\cdots,\min(n,M)\tag{1-7}$$

超几何分布的数学期望和方差分别为：

$$E(X)=n\cdot\frac{M}{N}\cdot V(X)=n\cdot\frac{M}{N}\cdot\frac{N-M}{N}\cdot\frac{N-n}{N-1}$$

当 N 很大，n 相对于 N 较小时，可用二项分布近似代替超几何分布进行计算。其中 $p=M/N$，即有

$$P(X=k)=\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}\approx\binom{n}{k}p^{k}(1-p)^{n-k}$$

用二项分布近似代替超几何分布计算可对例 1-4 做这样的解释：若采用不重复抽取方式，则在选出的 5 件中所选的优质品数服从超几何分布。若采用重复抽取方式，由于每次抽样相互独立，且每次抽到优质品的可能性为 $p=0.4$，则选出的 5 件产品中优质品数服从二项分布。但当产品总数很多（即 N 很大），抽取的产品数 n 相对较小，这样产品被重复抽到的可能性很小，重复抽样和不重复抽样没多大差别，故可用二项分布近似代替超几何分布。

（三）泊松分布

若随机变量 X 的概率函数是：

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}\quad(k=0,1,2,\cdots\cdots) \tag{1-8}$$

其中，$\lambda>0$，为常数，则称 X 服从参数为 λ 的泊松（Poisson）分布。

零件、铸造表面一定大小面积内的砂眼数服从泊松分布。

泊松分布的数学期望和方差相等，都是 λ，即：

$$E(X)=\lambda \qquad V(X)=\lambda$$

通常在 n 比较大，p 较小时，用泊松分布近似代替二项分布公式，即：

$$\binom{n}{k}p^kq^{n-k}\approx\frac{\lambda^k}{k!}e^{-\lambda}$$

其中，$\lambda=np$。实际工作中当 $n\geqslant 20$，$p\leqslant 0.05$ 时，便可用上述近似计算公式。泊松分布的方便之处在于计算简捷。

例 1-5　一大批产品的废品率为 $p=2\%$，现按不重复方式随机抽取 100 个产品进行检验，求恰有 1 个废品的概率。

解：由于是不重复方式抽样，因而 100 个产品中出现的废品数服从超几何分布，但因产品数 N 很大，可用二项分布近似计算：

$$P(X=1)=\binom{100}{1}\times 0.02\times 0.98^{99}=0.27065$$

另外，由于 n 比较大，p 很小，可用泊松分布近似代替二项分布计算。其中 $\lambda=np=2$，由式（1-8）得：

$$P(X=1)=\frac{2^1}{1!}e^{-2}=0.27067$$

可见，在满足一定的条件下用泊松分布近似代替二项分布计算时，相对误差不大，如本例中相对误差不超过万分之一。

二、连续型分布

（一）正态分布

正态分布是最常见的也是最重要的一种分布。它是连续型分布，常用于描述测量误差及射击命中点与靶心距离的偏差等现象。这类分布都具有“中间大，两头小”的特点。在正常情况下，这种量都可以看成由许多微小的、独立的随机因素作用的总后果，而每一种因素都不能起压倒一切的主导作用。具有这种特点的随机变量，一般都可以认为服从正态分布。另外，如一批零件的长度为随机变量，分布的特点是长度在某一范围（平均值附近）内的零件数最多，较长的和较短的数量较少，也服从正态分布。

1. 正态分布的概率密度

正态分布的概率分布密度函数为：

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty \tag{1-9}$$

其中，σ、μ 为常数，并且 $\sigma > 0$。正态分布记为 $N(\mu,\sigma^2)$。

正态分布的数学期望和方差分别为：

$$E(X) = \mu \qquad V(X) = \sigma^2$$

因而正态分布概率密度中的两个参数 μ 和 σ 分别是随机变量的期望值和标准差。若已知正态分布的数学期望和方差，可以完全确定正态分布。正态分布密度函数图形见图 1-1。

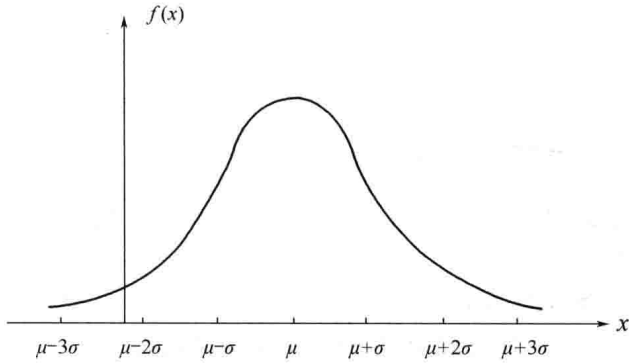
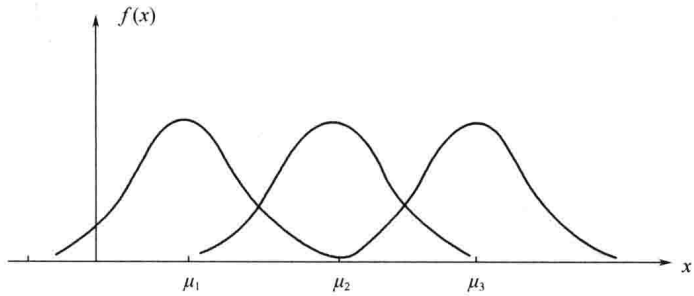


图 1-1　正态分布密度函数

正态分布的概率密度曲线具有下述特点：

（1）曲线关于直线 $x = \mu$ 对称，在 $x = \mu$ 处 $f(x)$ 达到最大值。

（2）当 σ 固定不变时，曲线的形状不变，对称中心取决于 μ 的值。μ 值越大，曲线越往右移；μ 值越小，曲线越往左移。因而 μ 是反映正态分布的中心位置和相应随机变量取值集中位置的参数。如图 1-2 所示。

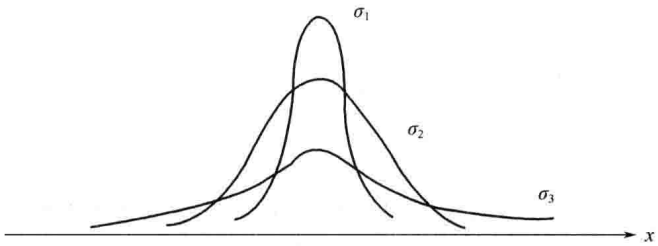


$\mu_1 < \mu_2 < \mu_3$

图 1-2　σ 固定，μ 值变化对正态分布密度函数的影响

（3）当 μ 固定不变时，曲线的对称位置不变，但其形状取决于 σ 的值。σ 值越大，曲线越平坦，且在 $x = \mu$ 处曲线越低，表明数据较分散；σ 值越小，曲线越陡，且在 $x = \mu$ 处

曲线越高，表明随机变量取值越集中于 $x = \mu$ 附近。因而 σ 是反映正态分布分散程度的参数。如图 1-3 所示。



$\sigma_1 < \sigma_2 < \sigma_3$

图 1-3　μ 固定，σ 值变化对正态分布密度函数的影响

2. 标准正态分布

特别地，当 $\mu = 0, \sigma = 1$ 时：

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

称为标准正态分布的概率密度函数，记为 $f_0(x)$。当随机变量 X 服从标准正态分布时，记为 $X \sim N(0,1)$。

标准正态分布概率密度的性质：

(1) $f_0(-x) = f_0(x)$，即 $f_0(x)$ 的图形关于 y 轴对称；

(2) $f_0(x)$ 在 $(-\infty, 0)$ 内单调上升，在 $(0, +\infty)$ 内单调下降，在 $x = 0$ 处达到最大值：$f_0(x) = 0.3989$；

(3) x 轴是曲线 $f_0(x)$ 的水平渐近线。

$f_0(x)$ 的图形如图 1-4 所示。

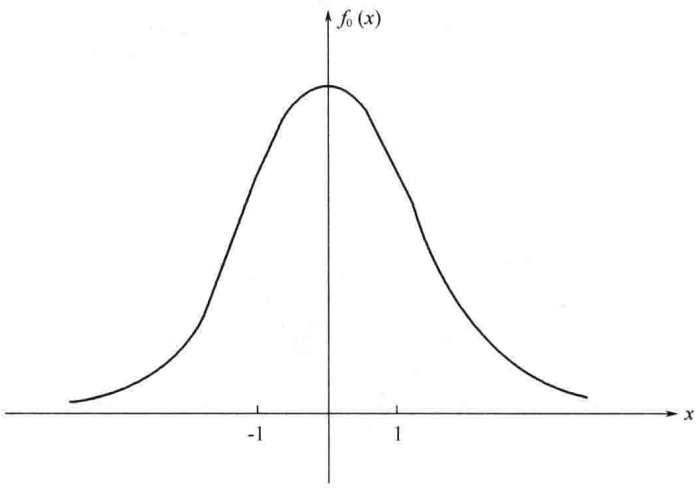


图 1-4　标准正态分布概率密度函数

3. 一般正态分布与标准正态分布的关系

(1) 一般正态分布与标准正态分布密度函数的关系

一般正态分布与标准正态分布密度函数之间满足：

$$f(x) = \frac{1}{\sigma} f_0\left(\frac{x-\mu}{\sigma}\right)$$

（2）一般正态分布与标准正态分布分布函数的关系

设一般正态分布和标准正态分布的分布函数分别为 $\Phi(x)$ 和 $\Phi_0(x)$，则两者之间存在关系：

$$\Phi(x) = \Phi_0\left(\frac{x-\mu}{\sigma}\right)$$

由此可见，一般正态分布的分布函数可通过标准正态分布函数求出。本书后附表 1 给出了 $x > 0$ 时标准正态分布函数 $\Phi_0(x)$ 值。

同时我们得到：若 $X \sim N(\mu,\sigma^2)$，则 $\dfrac{X-\mu}{\sigma} \sim N(0,1)$

例 1-6　设 $X \sim N(0,1)$，求 $P(X \leqslant 2)$，$P(X \leqslant -2)$，$P(|X| \leqslant 2)$，$P(-2 < X \leqslant 1)$。

解：由附表 1 可直接查得 $P(X \leqslant 2) = \Phi_0(2) = 0.97725$

由标准正态分布的对称性可得

$$P(X \leqslant -2) = P(X > 2) = 1 - P(X \leqslant 2) = 1 - \Phi_0(2) = 0.02275$$

所以 $\Phi_0(-2) = 1 - \Phi_0(2) = 1 - 0.97725 = 0.02275$

$$P(|x| \leqslant 2) = P(-2 \leqslant X \leqslant 2) = \Phi_0(2) - \Phi_0(-2) = 2\Phi_0(2) - 1 = 0.9545$$

$$\begin{aligned} P(-2 < X \leqslant 1) &= \Phi_0(1) - \Phi_0(-2) \\ &= \Phi_0(1) - [1 - \Phi_0(2)] \\ &= 0.8413 - [1 - 0.97725] \\ &= 0.81855 \end{aligned}$$

若 $X \sim N(0,1)$，则有如下结论：

（1）$P(X \leqslant x) = \begin{cases} \Phi_0(x) & x \geqslant 0 \\ 1 - \Phi_0(x) & x < 0 \end{cases}$

（2）$P(|X| \leqslant x) = 2\Phi_0(x) - 1$（当 $x > 0$ 时）

（3）$P(a < X \leqslant b) = \Phi_0(b) - \Phi_0(a)$

例 1-7　设 $X \sim N(5,4)$，求 $P(|X-5| < 2)$ 及 $P(X \leqslant 9)$。

解：因为 $X \sim N(5,4)$，则 $\mu = 5, \sigma^2 = 4$，故 $\sigma = 2$，因而 $\dfrac{X-5}{2} \sim N(0,1)$

所以，$P(|X-5| < 2) = P\left(\left|\dfrac{X-5}{2}\right| < 1\right) = 2\Phi_0(1) - 1 = 0.6826$

$$P(X \leqslant 9) = P\left(\frac{X-5}{2} \leqslant \frac{9-5}{2}\right) = \Phi_0(2) = 0.97725$$

（二）均匀分布

均匀分布在两端点 a 和 b 之间有一个平坦的概率密度函数，见图 1-5。它的全称是“在区间 (a,b) 上的均匀分布”，常记为 $U(a,b)$。这里“均匀”是指随机变量落在区间 (a,b) 内任一点的机会是均等的，从而在相等的小区间上的概率相等。

均匀分布的概率密度为：