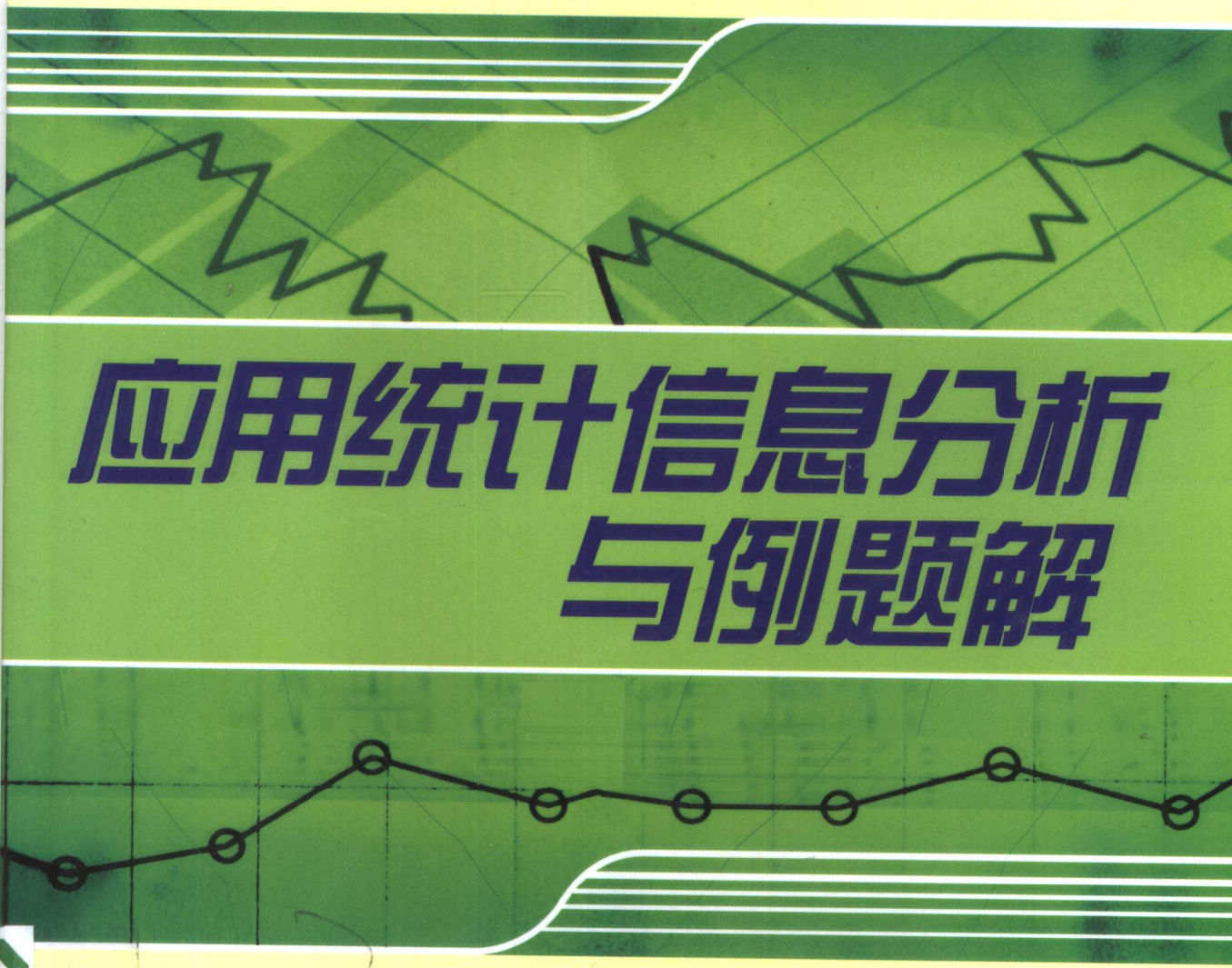

应用统计信息分析与例题解

勒中鑫 编著

国防工业出版社
National Defense Industry Press

应用统计信息分析与例题解

勒中鑫　编著

国防工业出版社

·北京·

图书在版编目（CIP）数据

应用统计信息分析与例题解 / 勒中鑫编著. —北京：国防工业出版社，2006.1
ISBN 7-118-03943-8

Ⅰ.应... Ⅱ.勒... Ⅲ.统计分析-信息处理 Ⅳ.C81

中国版本图书馆 CIP 数据核字（2005）第 057961 号

国防工业出版社出版发行
（北京市海淀区紫竹院南路 23 号）
（邮政编码 100044）
国防工业出版社印刷厂印刷
新华书店经售
*
开本 787×1092 1/16 印张 16¾ 382 千字
2006 年 1 月第 1 版 2006 年 1 月北京第 1 次印刷
印数:1—4000 册 定价:30.00 元

（本书如有印装错误，我社负责调换）
国防书店：（010）68428422 发行邮购：（010）68414474
发行传真：（010）68411535 发行业务：（010）68472764

引　言

本书主要是介绍各种实用统计信息的分析方法。内容有数理统计与估计，统计假设检验，回归信息分析，多元相关与特征分解，聚类与自组织分析，模式分析与识别，生存数据分析与辨识。

本书特点是通俗、精辟、实用。

所谓应用，主要是指数据统计信息分析方法在自然科学和社会科学领域的各类应用。其中包括电子、航天、机械产品寿命试验，火箭燃料与推进器试验，矿物、生物成分分析，喷漆工艺不匀数，铸件中气泡数，布匹上疵点数，洗衣粉包装量，目标距离误差测量，海拔高度与积雪深度关系，展览会参观人数估计，事故分析等。事物都是发展的，实用统计信息分析方法也随着潜在的多种多样的新事物新内容的发展而发展，数据输入计算机，利用计算机可自动或智能地进行分析与识别。

本书是将统计理论、应用例题与 Matlab、SPSS 有机地结合起来，从单纯解题的观点来看，应用 SPSS 在一些方面还是蛮好的，但为了理解解题过程中数学概念，本书侧重应用了 Matlab。书中涉及的许多知识，过去只是一些数学工作者或者分析专家所掌握，一般人员只是使用而不知其所以然。本书的出版，使一些过去只在某些专门领域应用的知识与方法，成为更多人所掌握，在更多更广领域推陈出新，真正成为“知识经济”，获得更多更好的成果，同时也将多维变换讲清楚了，这就是本书的宗旨。书中例题都做了详解，其中提出的主因子图像、主相关图像与众所周知的主分量图像一块更深入、更全面地反映了统计理论在图像分析中的应用，概念清晰、方法新颖、求解可行。近年来在统计理论中，结合神经网络的运用，使统计理论的计算、使用得到了重要的推动和发展。本书选用的神经网络函数正是为了这样一个目的。

本书中使用 Matlab、SPSS，这是统计工具的应用。通过大量统计函数的例题或经典例题解，不仅能使读者深入地全面地理解统计理论的应用、统计函数使用方法与范围，而且对从事统计业务的读者，也有助于他们更灵活更实效地运用统计函数创出新业绩。因此，本书对从事统计科学教学与研究的各类技术人员、专业教师、研究生和高年级学生，是一本有益于提高教学水平与学习水平的实用教材，有重要参照价值。

由于水平有限，难免有不当之处，请各位读者指正。

于信息工程大学

内 容 简 介

本书主要是介绍各种实用统计信息的分析方法。内容有数理统计与估计，统计假设检验，回归信息分析，多元相关与特征分解，聚类与自组织分析，模式分析与识别，生存数据分析与辨识。

本书特点是通俗、精辟、实用。

本书中使用 Matlab、SPSS，这是统计工具的应用。通过大量统计函数的例题或经典例题解，不仅能使读者深入地、全面地理解统计理论的应用、统计函数使用方法与范围，而且对从事统计业务的读者，有助于他们更灵活、更实效地运用统计函数创出新业绩。因此，本书对从事统计科学教学与研究的各类技术人员、专业教师、研究生和高年级学生，是一本有益于提高教学水平与学习水平的实用教材，有重要参考价值。

目　录

第1章　数理统计与估计 …… 1

1.1　数理统计基本概念 …… 1

1.1.1　总体与样本 …… 1

1.1.2　统计量与样本矩 …… 1

1.2　抽样分布与基本概率计算 …… 3

1.2.1　正态分布 …… 3

1.2.2　χ^2 分布与偏 χ^2 分布 …… 6

1.2.3　t 分布与偏 t 分布 …… 8

1.2.4　F 分布与偏 F 分布 …… 9

1.2.5　Weibull 分布 …… 10

1.3　点估计与估计量的评估标准 …… 11

1.3.1　矩估计 …… 11

1.3.2　极大似然估计 …… 12

1.3.3　最优无偏估计与有效性 …… 14

1.3.4　相合估计 …… 16

1.3.5　充分性与完备性 …… 17

1.4　区间估计 …… 17

1.4.1　正态总体均值与方差的区间估计 …… 17

1.4.2　0－1 分布参数的置信区间估计 …… 20

1.4.3　泊松分布参数的区间估计 …… 22

1.4.4　大样本均值参数的区间估计 …… 23

1.4.5　分布参数的单边置信限 …… 24

1.5　贝叶斯估计 …… 24

1.5.1　贝叶斯验后分布 …… 24

1.5.2　最大风险最小化的基本概念 …… 24

1.5.3　验前分布与贝叶斯最大后验估计 …… 26

1.5.4　贝叶斯区间估计 …… 29

第2章　统计假设检验 …… 32

2.1　正态总体参数检验 …… 32

2.1.1　参数检验概念与几个检验方法 …… 32

2.1.2　正态总体非配对数据的均值差检验 …… 37

2.1.3　正态相关数据对的均值差检验 …… 38

2.1.4 正态方差检验 …… 39
2.2 总体参数一致性分析 …… 40
2.2.1 单因素均值与方差检验 …… 40
2.2.2 等重复与不等重复双因素分析 …… 44
2.2.3 其他总体参数检验 …… 49
2.3 基于似然函数的检验与一致性 …… 52
2.3.1 似然比检验 …… 52
2.3.2 Neyman – Pearson 检验 …… 54
2.3.3 一致最优势检验与样本容量确定 …… 54
2.4 序统计量与统计容忍区间 …… 58
2.4.1 序统计量及其分布 …… 58
2.4.2 统计容忍区间 …… 58
2.5 总体假设检验 …… 60
2.5.1 正态概率图与偏度峰度检验 …… 60
2.5.2 Pearsonχ^2 检验 …… 61
2.5.3 柯尔莫哥洛夫与斯米尔诺夫检验 …… 63
2.5.4 秩和检验 …… 66
2.5.5 符号检验 …… 67
2.5.6 分类检验 …… 68
第3章 回归信息分析 …… 72
3.1 线性回归统计模型基本性质 …… 72
3.1.1 线性回归统计模型与最小二乘解 …… 72
3.1.2 最小二乘解的基本性质 …… 73
3.1.3 回归方程计算与相关 …… 74
3.1.4 回归分析检验 …… 78
3.2 重要线性回归类型 …… 80
3.2.1 回归变量的不同组合 …… 80
3.2.2 后向与前向逐步回归 …… 82
3.2.3 多元线性逐步回归 …… 82
3.2.4 共线、岭回归与特征筛选法 …… 87
3.2.5 加权最小二乘法 …… 90
3.2.6 虚拟变量回归 …… 93
3.2.7 主分量回归 …… 95
3.2.8 Logistic 回归 …… 98
3.3 多项式回归 …… 101
3.3.1 多项式直接回归 …… 101
3.3.2 按多元线性回归处理 …… 103
3.3.3 二次响应曲面预测 …… 104
3.4 非线性回归 …… 105

3.5 图像预测与复原条件回归 …… 109
第4章 多元相关与特征分解 …… 112
4.1 主分量 …… 112
4.1.1 主分量定义与性质 …… 112
4.1.2 结合例题分析透视主分量内涵 …… 113
4.1.3 多光谱图像与视频图像的主分量分析 …… 120
4.2 主因子 …… 121
4.2.1 主因子分析模型 …… 122
4.2.2 主因子算法与得分 …… 123
4.2.3 基于方差增大方向的正交矩阵旋转算法 …… 128
4.2.4 主因子图像及其意义 …… 136
4.3 主相关 …… 138
4.3.1 主相关信息分析模型与得分 …… 138
4.3.2 主相关信息分析实例 …… 139
4.3.3 显著性检验与方差 …… 144
4.3.4 图像主相关分析 …… 145
4.4 主对应 …… 147
4.4.1 主对应分析模型与举例 …… 147
4.4.2 主对应对图像处理的启示 …… 161
第5章 聚类与自组织分析 …… 163
5.1 距离及其聚类 …… 163
5.1.1 距离定义 …… 163
5.1.2 聚类分析与举例 …… 164
5.2 相似性聚类 …… 172
5.2.1 样本余弦与相关 …… 172
5.2.2 样本余弦聚类与距离方法的关系 …… 172
5.3 有序聚类 …… 175
5.3.1 有序聚类基本思路 …… 175
5.3.2 有序聚类用于预测 …… 179
5.4 神经网络自组织分析与图像编码 …… 181
第6章 模式分析与识别 …… 183
6.1 距离判别 …… 183
6.1.1 两总体判别 …… 183
6.1.2 线性化协方差矩阵 …… 184
6.2 Fisher 判别 …… 186
6.2.1 Fisher 准则 …… 186
6.2.2 Fisher 算法举例 …… 187
6.3 贝叶斯判别 …… 190
6.3.1 贝叶斯后验概率与风险 …… 190

6.3.2 贝叶斯判别与几种方法的比较 …… 191
6.4 逐步变量筛选判别 …… 196
6.4.1 逐步变量筛选概念 …… 196
6.4.2 逐步判别基本思路和计算 …… 196
6.5 模式训练判别 …… 200
6.5.1 感知器算法 …… 200
6.5.2 梯度算法 …… 202
6.5.3 最小均方误差算法 …… 202
6.5.4 势函数算法 …… 203
6.6 神经网络判别与图像分类 …… 206
第7章 生存数据分析与辨识 …… 209
7.1 生存数据分析问题 …… 209
7.1.1 生存函数 …… 209
7.1.2 危险率函数与生存函数关系 …… 210
7.1.3 生存数据推演 …… 211
7.2 生存分布几个非参数检验 …… 218
7.2.1 Mantel 检验 …… 218
7.2.2 Cox – F 检验 …… 221
7.2.3 Mantel – Haenszel 检验 …… 222
7.2.4 Kruskal – Wallis 检验 …… 223
7.2.5 Pearson – χ^2 检验 …… 225
7.3 几个应用于生存的理论分布与估计 …… 226
7.3.1 指数分布参数估计 …… 227
7.3.2 对数正态分布参数估计 …… 230
7.3.3 Γ 分布参数估计 …… 233
7.3.4 Gehan – Siddiqui 加权最小二乘法 …… 236
7.4 生存分布的参数检验 …… 240
7.4.1 指数分布的似然比检验与 Cox – F 检验 …… 240
7.4.2 Thoman – Bain 检验 …… 242
7.4.3 Rao – F 检验 …… 243
7.5 生存数据变量特性辨识 …… 244
7.5.1 指数模型 …… 246
7.5.2 风险比较与 Fisher 判决 …… 253
7.5.3 逻辑回归模型 …… 256
参考文献 …… 257

第 1 章　数理统计与估计

数理统计的观测数据，通常只能是所研究对象全体中的一部分。如一批灯管只能用部分灯管试验而获得寿命的可靠性数据，一批炮弹也只能取几枚炮弹试验而获得射程等性能数据。由这样一些数据对该批灯管或炮弹全体进行推测的结论必然含有不确定性。因为抽取的数据不包含每个灯管或炮弹的全部信息。数理统计就是利用统计理论辨识这种不确定性的程度，对研究的灯管或炮弹等对象给出反映规律性的实际估计。参数估计是根据抽取的一维或多维样本对各均值、方差及其分布特征参数作出合理的点或可信的区间估计。

1.1　数理统计基本概念

先说明几个常用术语。

1.1.1　总体与样本

在数理统计中通常将研究对象的全体称为总体。总体是基本单元的集合，总体中每个基本单元为个体。如总体为 100 万只灯管，则其中每个灯管就是个体。实际中人们关心的是研究对象的某个或某些数值指标和这些数值指标的分布。如灯管寿命、钢筋强度都是一数值指标，学生身高和体重是两数值指标，或视为两个总体。对抽取的每个个体，所观测到的某个或某些数值指标是这样或那样的一些数值，为一随机变量或随机向量，总体分布就是指人们关心的某个或某些数值指标的分布状态。总体与数值指标可能取值的全体组成的集合等同。灯管寿命总体可想象为该批生产的灯管寿命取值的全体所组成的集合。总体为无限个元素组成时，数值指标就是连续型随机变量。

对某个研究对象一次抽样得到的观测值$(x_1,x_2,\cdots,x_n)$，显然它是完全确定的，但它又是随每次抽样而不同。因此，将$(x_1,x_2,\cdots,x_n)$写为

$$\boldsymbol{x}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n)$$

在不同的观测中，容量 n 的样本 $\boldsymbol{x}$ 得到不同实现，每个分量 $\boldsymbol{x}_i$ 按随机变量取值 x_i，有$(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n)\leftrightarrow(x_1,x_2,\cdots,x_n)$。$\boldsymbol{x}$ 可能取值的全体属于 n 维样本空间，或为其中一子集，所以一个样本的观测值 $x=(x_1,x_2,\cdots,x_n)$ 是该样本空间中一点。简单随机样本是每个分量 $\boldsymbol{x}_i$ 是相互独立的随机变量，$(x_1,x_2,\cdots,x_n)$可看做 n 次的一组观测值，每个分量观测结果不影响其他观测结果，也不受其他观测结果的影响，每个分量 $\boldsymbol{x}_i$ 与所研究的总体具有独立同分布(i.i.d)。

1.1.2　统计量与样本矩

统计量是由 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$ 构造的、不包含任何未知参量的可测函数 $g(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n)$。

当 $x_1, x_2, \cdots, x_n$ 服从于正态分布 $N(\mu, \sigma^2)$，μ 已知，σ^2 未知，则 $\sum_{i=1}^{n}(x_i-\mu)^2$ 是统计量，$\sum_{i=1}^{n}x_i/\sigma$ 不是统计量。下面列举的是样本 $x_1, x_2, \cdots, x_n$ 的一些描述性的统计量。

样本均值
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i \tag{1.1}$$

样本方差
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{1.2}$$

样本 k 阶(原点)矩
$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i)^k \tag{1.3}$$

样本 k 阶中心矩 $q_k = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^k$(2 阶中心矩或记为 q^2) (1.4)

$$m_1 = \bar{x}, \quad q^2 = \frac{n-1}{n}s^2$$

令 μ 为总体均值，σ^2 为总体方差，α_k 为总体 k 阶矩，μ_k 为总体 k 阶中心矩，都是总体期望值，依据定义，有

$$Ex \equiv \mu, Dx = E(x-\mu)^2 \equiv \sigma^2, Ex^k \equiv \alpha_k, E(x-\mu)^2 \equiv \mu_k$$

其中 Ex、Dx 或加一括号。常数 c 的期望 $E(c)=c, E(cx)=cE(x)$，两个随机变量的期望 $E(x+y)=E(x)+E(y)$，相互独立时 $E(xy)=E(x)E(y)$，$D(c)=0, D(cx)=c^2D(x)$，$D(x+y)=D(x)+D(y)$(对有限多个相互独立的随机变量都成立)。从总体中抽得的一简单随机样本，在 2 阶矩存在时，样本均值

$$E(\bar{x})=\mu, \quad D(\bar{x}) = \sigma^2/n$$

因为
$$E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right) = \frac{1}{n}\sum_{i=1}^{n}E(x_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

$$D(\bar{x}) = D\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right) = \frac{1}{n^2}D\left(\sum_{i=1}^{n}x_i\right) = \frac{1}{n}D(x_i) = \frac{\sigma^2}{n}$$

例题 1.1 假定 $x=[x_1, x_2, \cdots, x_5]_{4\times 5}$ 的取值为表 1.1 所列。试估计每列均值、方差、3 阶矩、3 阶中心矩及峰度。

表 1.1 假设数据

x_1	x_2	x_3	x_4	x_5
4.30	6.10	6.50	9.30	9.50
7.80	7.30	8.30	8.70	8.80
3.20	4.20	8.60	7.20	11.4
6.50	4.10	8.20	10.1	7.80

解 调用函数，样本均值(每列的)为

$$\text{mean}(x) = [5.4500, 5.4250, 7.9000, 8.8250, 9.3750]$$

样本方差为

$$s^2 = (\text{std}(x))^2 = [4.3367, 2.4092, 0.9000, 1.5025, 2.3092]$$

样本 3 阶矩为

m_3 = (sum(x.^3))./4 = [215.3630,189.7517,508.4590,716.6023,873.7358]

3 阶中心矩为

q_3 = moment(x,3) = [0.3060,0.6837, -0.5775, -0.5283,1.0522]

峰度(未减 3)为

kurtosis(x) = [1.3965,1.3708,1.4.581,1.9031,1.9236]

对总体 x 的分布函数或数字特征的估计或推测时,统计量是重要的。在数理统计中,对小样本容量要确定一个统计量的精确分布比较难,但还是可找出的,总体 x 为正态分布就是精确分布的一范例。统计量的精确分布难以表达时,通常求它在 $n\to\infty$ 时的极限分布,这是对大样本容量的一种有效的研究方法。小样本与大样本是相对的,没有明确界限。大量应用正态总体,使其统计量的精确分布的数学分析比较容易,同时在许多统计领域所遇到的总体,以正态分布函数描述具有很高的逼近程度。

总体 x 的离散随机变量只有有限个或可数个可能值,其概率密度函数是观测到的某特定值的概率。连续随机变量的概率是区域间隔内概率密度函数的积分,在整个区域概率密度函数的积分为 1。记随机变量 x 的分布函数

$$F(x) = p(x < x) \quad (0 \leqslant F(x) \leqslant 1)$$

当 $x_1 < x_2$ 时,则 $F(x_1) \leqslant F(x_2)$。随机变量 x 的概率密度函数是分布函数的导数,有

$$f(x) = F'(x)$$

分布函数 $F(x)$ 的特征函数

$$\phi(t) = \int_{-\infty}^{+\infty} e^{jxt} dF(x) \quad 或 \quad \phi(t) = \int_{-\infty}^{+\infty} e^{jxt} f(x)dx \equiv Ee^{jxt} \tag{1.5}$$

特征函数是完全刻画分布函数的一重要参数,有时比分布函数更实用。

当随机变量 x 存在 n 阶矩 $Ex^k, k=1,2,\cdots,n$ 时,则随机变量 x 的特征函数 $\phi(t)$ 存在 k 阶导数 $\phi^{(k)}(t), k=1,2,\cdots,n$,有

$$Ex^k = (-j)^k \phi^{(k)}(t)\big|_{t=0}$$

如果 $x_1, x_2, \cdots, x_n$ 为 n 个独立随机变量,$y = \sum_{i=1}^{n} x_i$,则 y 的特征函数

$$\phi_y(t) = \prod_{i=1}^{n} \phi_{x_i}(t)$$

注:计算中依据 Matlab 规定,向量与矩阵元素间逗号表示分开,矩阵元素间分号表示换行,右上角'表示矩阵转置,* 表示矩阵相乘,上角^表示矩阵乘方,.^表示矩阵元素乘方等。形式上,书中公式按理论分析写出,计算时按 Matlab 形式,如书中 Cov、ln、$\log_{10}$ 分别与 cov、log、log10 是等价的。

1.2 抽样分布与基本概率计算

在数理统计中,有许多分布函数,下面介绍几种分布函数。

1.2.1 正态分布

正态密度函数称正态分布,表示为

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-(x-\mu)^2/2\sigma^2)$$

记为 $x \sim N(\mu,\sigma^2)$。由正态分布定义可知,样本均值 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ 服从正态分布

$$\bar{x} \sim N(\mu,\sigma^2/n), E(\bar{x}) = \mu, D(\bar{x}) = \sigma^2/n$$

$E(\bar{x})$ 与总体均值相等,$D(\bar{x})$ 只是总体方差的 n 分之一。n 越大,形状越向总体均值 μ 集中。令 $x_1,x_2,\cdots,x_n$ 的变换为 $y = Ax$,式中系数矩阵 $A = [\alpha_{ij}]_{q\times n}$。$x_1,x_2,\cdots,x_n$ 服从正态,则 $y_1,y_2,\cdots,y_q$ 也服从正态。$x_1,x_2,\cdots,x_n$ 独立同分布于 $N(\mu,\sigma^2)$,则 $y_1,y_2,\cdots,y_p$ 的均值、方差、协方差分别为

$$E(y_i) = a_{i1}E(x_1) + a_{i2}E(x_2) + \cdots + a_{in}E(x_n) = \mu\sum_{k=1}^{n} a_{ik}$$

$$D(y_i) = a_{i1}^2D(x_1) + a_{i2}^2D(x_2) + \cdots + a_{in}^2D(x_n) = \sigma^2\sum_{k=1}^{n} a_{ik}^2$$

$$\mathrm{Cov}(y_i,y_j) = \sigma^2\sum_{k=1}^{n} a_{ik}a_{jk}$$

式中协方差定义为 $\mathrm{Cov}(y_i,y_j) = E((y_i - E(y_i))(y_j - E(y_j)))$。如果 $\mathrm{Cov}(y_i,y_j) = 0$,则两个随机变量是相互独立的。$\mathrm{Cov}(y_i,y_j) \neq 0$,有

$$D(y_i + y_j) = D(y_i) + D(y_j) + 2\mathrm{Cov}(y_i,y_j)$$

$$\mathrm{Cov}(y_i,y_j) = E(y_i,y_j) - E(y_i)E(y_j)$$

y_i 与 y_j 相关,其相关系数定义为

$$r(y_i,y_j) = \frac{\mathrm{Cov}(y_i,y_j)}{\sqrt{D(y_i)}\sqrt{D(y_j)}} \tag{1.6}$$

根据协方差定义,有

$$\mathrm{Cov}(y_i,y_j) = \mathrm{Cov}(y_j,y_i)$$

$$\mathrm{Cov}(ay_i,by_j) = ab\cdot\mathrm{Cov}(y_i,y_j) \quad (a、b\text{ 为常数})$$

$$\mathrm{Cov}(y_i + y_j,y_k) = \mathrm{Cov}(y_i,y_k) + \mathrm{Cov}(y_j,y_k)$$

如果 $A = [a_{ij}]$ 是一 $n\times n$ 的正交矩阵,且 $\mu = 0$,则 $y_1,y_2,\cdots,y_n$ 也是相互独立且同服从于 $N(0,\sigma^2)$ 分布的随机变量。即独立、零均值的正态随机变量 $x_1,x_2,\cdots,x_n$,通过正交变换后获得的随机变量 $y_1,y_2,\cdots,y_n$ 也是独立、零均值的正态随机变量,y 与 x 同分布。而经 $y = (x-\mu)/\sigma$ 变换,则 $y_1,y_2,\cdots,y_n$ 相互独立,且分布为 $y \sim N(0,1)$。

正态分布 $x \sim N(\mu,\sigma^2)$ 调用 Matlab 函数的仿真程序段为

```
t = [-3:0.1:5];
p1 = normpdf(t,0,0.95);
p2 = normpdf(t,0,1.5);
p3 = normpdf(t,1.2,2.3);
plot(t,p1);hold on
plot(t,p2);hold on
plot(t,p3)
```

就得正态密度分布如图 1.1 所示。$\boldsymbol{x} \sim N(1.2, 2.3^2)$ 的概率 $p(0 < \boldsymbol{x} \leqslant 1.6)$，调用函数，有

$$p(0 < \boldsymbol{x} \leqslant 1.6) = \text{normcdf}(1.6, 1.2, 2.3) - \text{normcdf}(0, 1.2, 2.3) = 0.2681$$

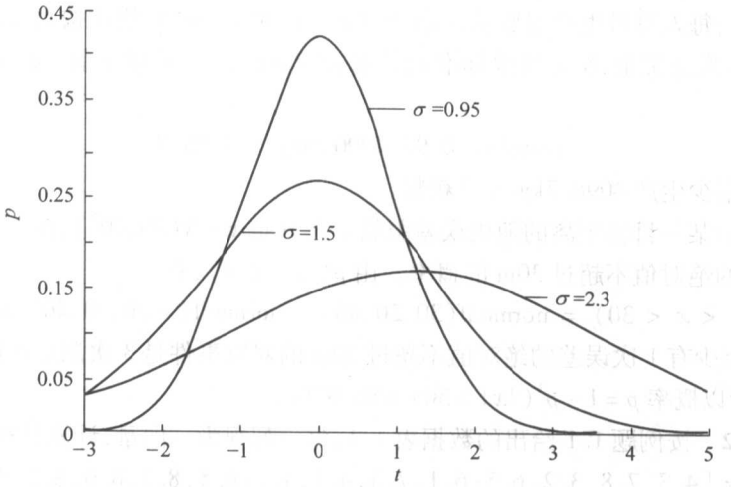

图 1.1　正态密度函数

在 $\boldsymbol{x} \sim N(1, 2^2)$ 中

$$p(1.47 < \boldsymbol{x} \leqslant 5.2) = \text{normcdf}(5.2, 1, 2) - \text{normcdf}(1.47, 1, 2) = 0.3892$$

$p(1.47 < \boldsymbol{x} \leqslant 5.2)$ 调用函数，有

$$p = \text{normspec}([5.2, 1.47], 1, 2) = 0.3892$$

同时绘得图 1.2。

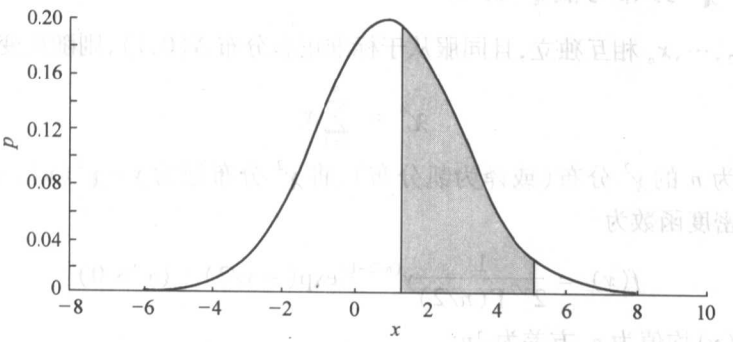

图 1.2　$p(1.47 < \boldsymbol{x} \leqslant 5.2)$ 的概率

对标准正态分布函数

$$p(\boldsymbol{x} \leqslant x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)\,\mathrm{d}u$$

$x = 1.6497$，则 $p(x \leqslant 1.6497) = \text{mormcdf}(1.6497, 0, 1) = 0.9505$。逆运算应用 norminv 函数，有

$$\text{norminv}(0.9505, 0, 1) = 1.6497$$

$$p(\boldsymbol{x} > 1.6497) = 1 - \text{normcdf}(1.6497, 0, 1) = 0.0495$$

某工厂生产的某批钢管，经统计其直径服从正态分布，$\boldsymbol{x} \sim N(2.05, 0.1^2)$，直径的合

格品标定为(2 ±0.2)cm,则产品合格率为

$p(1.8 < x \leqslant 2.2)$ = normcdf(2.2,2.05,0.1) − normcdf(1.8,2.05,0.1) = 92.7%

已知过去每人每月生产量服从正态分布 $x \sim N(4000,60^2)$,假定按每人每月生产的最高额5%的人发放奖金,奖金发放标准就是求 $p(x > z) = 0.05$ 的 z 值。调 norminv 函数,有

$$\text{norminv}(0.95,4000,60) = 4098.7$$

即每人每月最少生产4098.7kg才可获奖。

假定测量某一目标距离的随机误差服从正态分布 $x \sim N(20,40^2)$,在3次测量中至少有1次误差的绝对值不超过30m的概率。由 $p(|x| < 30)$,有

$p(-30 < x < 30)$ = normcdf(30,20,40) − normcdf(−30,20,40) = 0.4931

3次测量中至少有1次误差的绝对值不超过30m的对立事件是3次测量中误差绝对值都超过30m,所以概率 $p = 1 - p^3(|x| > 30) = 86.97\%$。

例题 1.2 按例题1.1给出的数据表1.1,每一列视为一向量,计算其相关量。

解 $x = [4.3,7.8,3.2,6.5;6.1,7.3,4.2,4.1;6.5,8.3,8.6,8.2;9.3,8.7,7.2,10.1;9.5,8.8,11.4,7.8]$,调用函数,相关矩阵为

$$\text{corrcoef}(x) = \begin{bmatrix} 1.0000 & 0.8034 & 0.7943 & 0.6194 \\ 0.8034 & 1.0000 & 0.8273 & 0.8912 \\ 0.7943 & 0.8273 & 1.0000 & 0.5485 \\ 0.6194 & 0.8912 & 0.5485 & 1.0000 \end{bmatrix}$$

1.2.2 χ^2 分布与偏 χ^2 分布

设 $x_1, x_2, \cdots, x_n$ 相互独立,且同服从于标准正态分布 $N(0,1)$,则随机变量

$$\chi^2 = \sum_{i=1}^{n} x_i^2$$

服从自由度为 n 的 χ^2 分布(或译为凯分布),将 χ^2 分布记为 $y \sim \chi^2(n)$,于是随机变量 $y \sim \chi^2(n)$ 的密度函数为

$$f(y) = \frac{1}{2^{n/2}\Gamma(n/2)} y^{(n/2-1)} \exp(-y/2) \quad (y > 0)$$

否则为0。$f(y)$ 均值为 n,方差为 $2n$。

对自由度 n 为1、4、7、10、15、20时,调用函数,有

```
y = 0:0.1:35;
f1 = chi2pdf(y,1);
f2 = chi2pdf(y,4);
f3 = chi2pdf(y,10);
f4 = chi2pdf(y,20);
```

绘得 $f(y)$ 分布,如图1.3所示。

$y \sim \chi^2(n)$ 分布的特征函数 $\phi(t) = \int_0^{\infty} e^{jyt} f(y)\,dy$,用符号积分有

$$\phi(t) = \int_0^{\infty} e^{jyt} f(y)\,dy =$$

int(exp(j * y * t) * 2^(- n/2)/gamma(n/2) * exp(- y/2) * y^(0.5 * n - 1),y,0,inf)

得(1 - j * 2 * t)^(- 1/2 * n),即 $\phi(t) = (1 - j2t)^{-n/2}$。

调用函数，计算 $\phi(t)$ 的导数，就可求得随机变量 y 的均值 Ey 与方差 Dy，即

$Ey = -j\phi'(0)$ = (- j) * diff((1 - 2 * j * t)^(- 1/2 * n)) | t = 0 =

(1 - 2 * j * 0)^(- 1/2 * n) * n/(1 - 2 * j * 0) = n

$Ey^2 = (-j)^2\phi''(t)$ = (- j)^2 * diff((1 - 2 * j * t)^(- 1/2 * n),2) | t = 0 =

((1 - 2 * j * 0)^(- 1/2 * n) * n^2/(1 - 2 * j * 0)^2 +

2 * (1 - 2 * j * 0)^(- 1/2 * n) * n/(1 - 2 * j * 0)^2) = n^2 + 2 * n

$Dy = Ey^2 - (Ex)^2 = 2n$

自由度为 18 时

$$p(y \leqslant 30 \mid 18) = \text{chi2cdf}(30,18) = 0.9626$$

当 $x_1, x_2, \cdots, x_n$ 相互独立、同服从于正态分布 $N(\mu_i, \sigma^2)$ 时，$i = 1,2,\cdots,n$，则随机变量

$$x_\delta^2 = \sum_{i=1}^{n} x_i^2/\sigma^2$$

的分布是自由度为 n 偏参数为 $\delta = \sqrt{\sum_{i=1}^{n}\mu_i^2/\sigma^2}$ 的偏 χ^2 分布（非中心分布），应用 ncx2pdf 函数可绘出偏 χ^2 分布。如 y = 0 : 0.1 : 35、自由度为 4、偏参数为 8 时，调用 ncx2pdf(y,4,8)，可算得 χ_8^2，即图 1.3 中所示的粗曲线，χ_8^2 的最大值位于 y = 8.9。

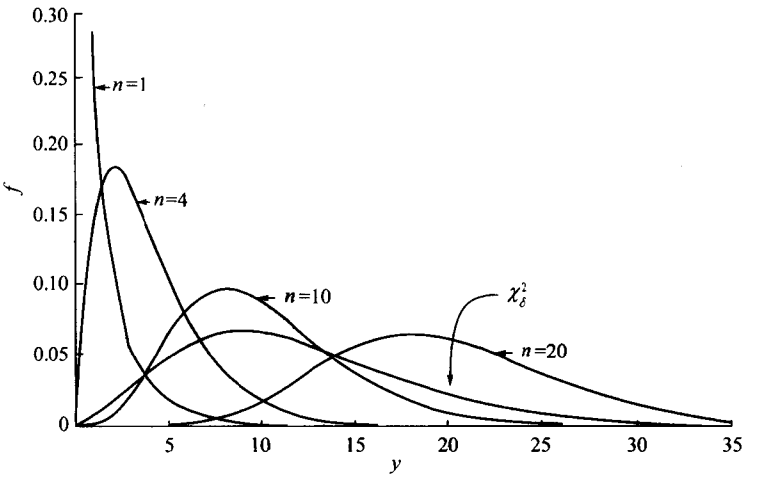

图 1.3 χ^2 与 χ_δ^2 的密度分布

例题 1.3 设随机变量 v 服从具有自由度 k 的 $v \sim \chi^2(k)$，求 $\eta = \sqrt{v/k}$ 的分布。

解 由题意知 $\eta = \sqrt{v/k}$，$v = k\eta^2 = s(\eta)$，$k > 0$，$v > 0$，$\eta > 0$，$s'(\eta) = 2k\eta$，有

$$f(s(\eta)) = \frac{1}{2^{k/2}\Gamma(k/2)}(k\eta^2)^{(k/2-1)}e^{-k\eta^2/2}$$

于是 η 的分布

$$\psi(\eta) = f(s(\eta)) \cdot 2k\eta = \frac{2(k/2)^{k/2}}{\Gamma(k/2)} y^{k-1} e^{-ky^2/2} \quad (\eta > 0)$$

$$\psi(\eta) = 0 \quad (\eta \leqslant 0)$$

1.2.3 t 分布与偏 t 分布

设 $x \sim N(0,1)$，$y \sim \chi^2(n)$，且 x 与 y 相互独立，随机变量

$$t = \frac{x}{\sqrt{y/n}}$$

服从自由度为 n 的 $t \sim t(n)$。t 的概率密度函数

$$f(t) = \frac{\Gamma((n+1)/2)}{(n\pi)^{1/2}\Gamma(n/2)}(1 + t^2/n)^{-(n+1)/2}$$

对自由度 n 为 1、4、8、30 时，调用函数，取 $t = -5:0.1:8$，有

```
f1 = tpdf(t,1);
f2 = tpdf(t,4);
f3 = tpdf(t,30);
plot(t,f1);hold on
plot(t,f2);hold on
plot(t,f3);hold on
```

得随机变量 t 的概率密度分布图 1.4。

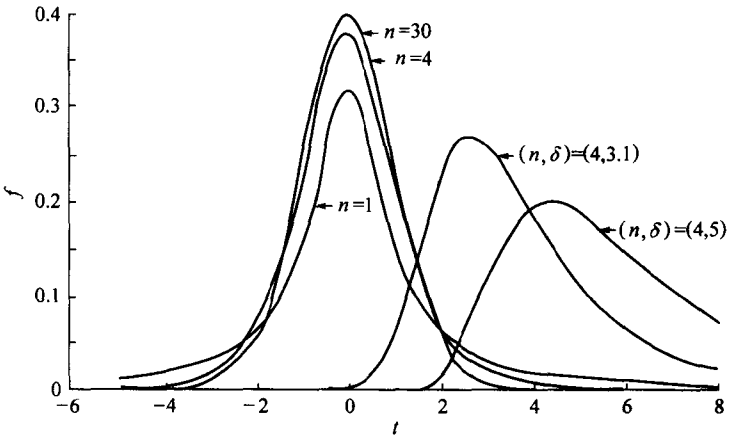

图 1.4 t 与 t_δ 的密度分布

$t \sim t(n)$ 的计算，自由度为 11 时，有

$$p(t \leqslant 1.7 \mid 11) = \text{tcdf}(1.7,11) = 0.9414$$

$$p((0.9 \mid 11) < t \leqslant 1.7 \mid 11) = \text{tcdf}(1.7,11) - \text{tcdf}(0.9,11) = 0.1351$$

x 与 y 相互独立，$x \sim N(\mu,\sigma^2)$，$y \sim \chi^2(n)$，随机变量

$$t_\delta = \frac{x}{\sqrt{y/n}}$$

所服从的分布称为自由度为 n、偏参数为 $\delta = \mu/\sigma$ 的偏 t 分布（非中心 t 分布）。图 1.4 所