包研科 编著

数据分析教程

清华大学出版社

包研科 编著

数据分析教程

清华大学出版社
北京

内 容 简 介

本书作为工科“概率论与数理统计”课程后续统计数据分析类课程的教材.内容包括：非参数统计推断、方差分析和多元线性分析的基本内容；多维正态总体的推断问题、主成分分析、典型相关分析、聚类分析、判别分析、偏最小二乘回归分析和 Logistic 回归分析；MATLAB 数据处理方法.

本书可作为工科硕士研究生“统计分析与应用”课程的基础教材，也可作为对统计数据分析有较高要求的本科各专业高年级学生的选修教材，还可作为管理、科研和工程技术人员的参考读物.

图书在版编目（CIP）数据

数据分析教程 / 包研科编著. — 北京：清华大学出版社，2011.9
ISBN 978-7-302-26596-2

Ⅰ.①数… Ⅱ.①包… Ⅲ.①统计数据—统计分析(数学)—高等学校—教材
Ⅳ.①O212.1

中国版本图书馆 CIP 数据核字(2011)第 175502 号

责任编辑：庄红权　洪　英
责任校对：刘玉霞
责任印制：李红英

出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编**：100084
社　总　机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：清华大学印刷厂
装 订 者：三河市新茂装订有限公司
经　　销：全国新华书店
开　　本：185×260　　**印　张**：16.75　　**字　数**：406 千字
版　　次：2011 年 9 月第 1 版　　**印　次**：2011 年 9 月第 1 次印刷
印　　数：1～4000
定　　价：29.00 元

产品编号：043690-01

前　言

对客观事物变量之间相互关系的研究，一直是人类在社会生产实践和科学研究中所探讨的重要课题之一. 从数学手段上讲，微分方程模型与统计模型是揭示变量关系的两大重要工具. 然而，两者在问题的背景、解决问题的原理和解决问题的方法上都存在区别. 通常，微分方程是基于物理学原理建立的，用于描述精确的物理过程；而统计推断则是基于观测数据，对随机现象进行的定量分析. 在绝大多数不存在或尚找不到严格的客观物理规律的场合中，统计数据分析是研究变量之间相互关系的主要方法，尤其在信息科学与管理科学中发挥着重要的作用，也在具有复杂系统背景的气象、水文、地质、地震、医疗、经济等领域得到了广泛的应用，它是现代管理、科研和工程技术人员必备的数学技术.

目前，在本科生及工科硕士研究生的培养方案中，一般分设数理统计课程和多元统计分析课程. 数理统计课程侧重于单变量的统计推断理论，多元统计分析课程则侧重于多变量的数据分析技术，这种知识结构难以使学生形成一个完整的认知体系. 随着统计分析软件的发展，工程数据分析的问题与需求日益增多，促进学生在有限的时间内领悟统计数据分析的基本思想方法、掌握核心概念和分析技术、获得初步的实践能力是本书的编写动机，作者多年的数理统计课程和多元统计分析课程的教学经验构成了本书的编写基础. 本书具有如下特点.

(1) 在工程数据分析问题中，面对的往往是多维复杂数据集合，对多元统计分析技术的需求日益增强. 然而，由于统计推断与数据分析技术的脱节，使得学习者的统计学知识与技能难以应对实际问题，甚至是一些常规的问题. 为解决这一矛盾，本书在融合数理统计的基本思想方法、多元统计分析的基本内容和形成统一的课程内容方面进行了积极的尝试.

(2) 选择哪些具体知识作为教材的支撑内容、如何把握知识的处理方式以控制内容的难易是编著教材与学术著作的区别所在. 概括地讲，统计数据分析技术主要由样本数据的拟合技术和随机误差的评价与控制技术构成. 本书在样本数据的拟合技术方面，兼顾学生的代数学基础、知识的历史与传统和工程应用的实际，以线性最小二乘法为主线展示教学内容；在随机误差的评价与控制技术方面，出于同样的理由突出了均方误差分析技术.

(3) 在数据处理软件的选择上，出于对工科专业数学计算的多样需求、软件的通用性和一举兼得等方面的考虑，本书同步介绍了 MATLAB 数据处理的基本方法，结合书中附录可收到 MATLAB 入门的功效.

本书反映了作者对统计类课程改革的思考和实践，可作为工科硕士研究生统计分析与应用课程的基础教材，也可作为对统计数据分析有较高要求的本科各专业高年级学生的选修教材，还可作为管理、科研和工程技术人员的参考读物.

郭嗣琮

2011 年 7 月于辽宁工程技术大学

前奏曲·缘

包研科

写于《数据分析教程》付梓之日

（一）

今生
数据让我们结缘
先人睿智
后生求索
传承
铸就我生命机玄

（二）

教室
你求知目光专注
激荡心绪
飞扬神采
回眸
却见你俏媚凝蹙

（三）

心动
解读你质惑询然
描摹先知
秀我博学
奈何
难觅你欢颜灿烂

（四）

期盼
机巧我语言神功
小品生动
相声幽默
同学
请原谅我的无能

（五）

师好
如莺莺鹂歌欢鸣
叫我动容
令吾坚持
谢谢
让我们挽手前行

（六）

起舞
和数据分析旋律
青春搏击
年少奇志
击鼓
拙著与君共徜徉

目　录

1 数据分析常用的概率分布

数据分析是以样本的实测数据为客观事实基础，以抽样分布为理论依据对总体的概率分布特征和相关事物之间的联系、未来发展趋势等进行科学推断的数学理论与方法.

概率分布对于数据分析的重要性，不仅仅是提供了基本的数学模型.更重要的是随机现象间的复杂联系蕴涵在概率分布间的联系之中.从概率分布之间的联系理解概率分布的意义，对学习概率分布的基础知识，并将其应用于实际问题的数据分析之中是非常有益的.

本章介绍在数据分析的理论与应用中常用的总体概率分布和抽样分布.

1.1 计数与计数值的概率分布

1.1.1 二项分布

研究一个事件最简单的方法是对其在 n 次重复试验中发生的次数计数.计数值为非负整数值变量，其概率分布通常用概率函数描述.

在实际应用中，一种重要的试验类型是**二项试验**，即在试验过程中对样本空间 S 仅做两种模式的认知：将需要特别关注的某些样本点(具有共同性质)归为一类，记为事件 A；剩余的样本点归为另一类，记为事件 $\overline{A}$.

定义 1.1 称变量

$$I_A=\begin{cases}1, & A\text{ 发生},\\ 0, & \overline{A}\text{ 发生},\end{cases}$$

为事件 A 的**示性函数**.

显然 I_A 是一个随机变量，可以理解为事件 A 发生的“记录器”.

定义 1.2 设随机变量 X 表示在一次试验中事件 A 发生的次数，若概率函数

$$P(X=x)=p^x(1-p)^{1-x},\quad x=0,1,$$

其中，$p=P(X=1)=P(A)$，则称 X 服从参数为 p 的**伯努利**(Bernoulli)**分布**，记为 $X\sim B(1,p)$.

若重复进行 n 次试验，记第 k 次试验中事件 A 的示性函数为

$$X_k=\begin{cases}1, & \text{第 }k\text{ 次试验 }A\text{ 发生},\\ 0, & \text{第 }k\text{ 次试验 }\overline{A}\text{ 发生},\end{cases}\quad k=1,2,\cdots,n,$$

则

$$X=\sum_{k=1}^{n}X_k$$

可以理解为 n 次重复试验中事件 A 发生的“计数器”.

显然,随机变量 X 表示 n 次重复试验中事件 A 发生的次数,其概率分布同各次试验是否相互独立有关.

一般情况下,前次试验的结果对后续试验中事件 A 发生的概率会产生影响,用条件概率描述即 $\forall i<j$,有

$$P(X_j=1 \mid X_i=1) \neq P(X_j=1 \mid X_i=0),$$

此时,X 的概率分布定义如下.

定义 1.3 设非空集合 S 有 N 个元素,$A \subset S$ 包含 M 个元素. 随机变量 X 表示 n 次二项试验中事件 A 发生的次数,若 X 的概率函数为

$$P(X=x)=\binom{M}{x}\binom{N-M}{n-x}\Big/\binom{N}{n}, \quad x=0,1,\cdots,\min\{n,M\},$$

则称 X 服从参数为 n,N,M 的**超几何分布**,记为 $X \sim Hg(n,N,M)$.

在理论中,人们更重视“前次试验的结果不影响后续试验中事件 A 发生的概率”的独立二项试验,即 $\forall i<j$,有

$$P(X_j=1 \mid X_i=0)=P(X_j=1 \mid X_i=1)=P(A),$$

通常称 n 次独立的二项试验为 n **重伯努利试验**.

定义 1.4 设在 n 重伯努利试验中,若 X 的概率函数为

$$P(X=x)=\binom{n}{x}p^x(1-p)^{n-x}, \quad x=0,1,\cdots,n$$

其中,$p=P(A)$,则称 X 服从参数为 n,p 的**二项分布**,记为 $X \sim B(n,p)$.

显然,伯努利分布是二项分布在 $n=1$ 时的特殊情形.

容易证明,若 $X \sim B(n,p)$,则 $E(X)=np$,$\mathrm{Var}(X)=np(1-p)$.

二项分布的参数 p 对其分布形态的影响如图 1.1 所示. 设 $n=20$,分别取 $p=0.10$,$p=0.25$,$p=0.50$,画出 $x=0,1,2,\cdots,20$ 各点处的概率.

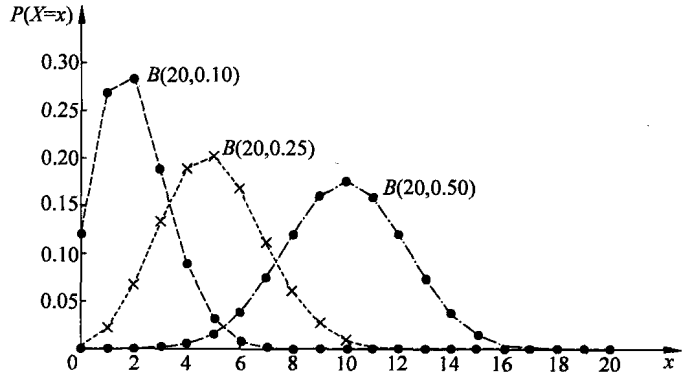

图 1.1 参数 p 不同取值的二项分布

由图 1.1 可知,当 p 值较小时分布是偏态的(右尾长);随着 p 值的增大,分布呈现出关于均值 np 值的对称形态;若 p 值继续增大,则分布将又变成偏态的(左尾长).

因此,对于确定的 p 值,在应用中往往要求 n 值充分大,使 $np \geqslant 5$ 或 $n(1-p) \geqslant 5$,此时

可按对称分布处理.

二项分布关于参数 n 具有再生性.

定理 1.1 设 $X_k \sim B(n_k,p)$,$k=1,2,\cdots,m$ 且相互独立,则 $X=\sum_{k=1}^{m}X_k \sim B\left(\sum_{k=1}^{m}n_k,p\right)$.

特别地,当 $X_k \sim B(1,p)$,$k=1,2,\cdots,m$ 且相互独立时,有 $X=\sum_{k=1}^{m}X_k \sim B(m,p)$.

概率分布的再生性在理论与应用中均发挥着重要的作用.

下面给出在二项分布的理论与应用中具有重要作用的两个概率极限定理.

定理 1.2(伯努利大数定律) 设事件 A 在一次试验中发生的概率 $P(A)=p$,在 n 次独立重复试验中发生的频数 $X_n \sim B(n,p)$,$\forall \varepsilon>0$,则有

$$\lim_{n\to+\infty} P\left(\left|\frac{X_n}{n}-p\right|\geqslant\varepsilon\right)=0.$$

定理 1.2 表明,随着试验次数 n 的增大,事件 A 发生的频率 $\frac{X_n}{n}$ 与概率 p 之间的偏差 $\left|\frac{X_n}{n}-p\right|$ 大于预先给定精度 ε 的可能性越来越小,小到可以忽略不计. 这是在应用中由统计方法确定事件概率的理论基础.

定理 1.3(De Moivre-Laplace 中心极限定理) 设 $X_n \sim B(n,p)$,$\forall x\in\mathbb{R}$,有

$$\lim_{n\to+\infty} P\left(\frac{X_n-np}{\sqrt{np(1-p)}}\leqslant x\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt.$$

定理 1.3 表明,若 $X_n \sim B(n,p)$,则当 n 充分大时,近似地 $X_n \sim N(np,np(1-p))$.

定理 1.3 亦称二项分布的正态近似定理,在应用中只需 $n\geqslant\frac{5}{\min\{p,1-p\}}$ 便可得到较好的近似效果.

若 $X_n \sim B(n,p)$,则

$$P(X_n\leqslant k)=\beta \tag{1-1}$$

刻画了抽样调查问题中三个基本的特征数值 n、k、β 间的关系. 根据定理 1.3,式(1-1)可以近似地转化为

$$\Phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right)=\beta. \tag{1-2}$$

其中,$y=\Phi(x)$ 是标准正态分布的分布函数. 根据式(1-2)能够方便地解决已知 n,k,β 中的两个值确定另一个值的计算问题.

定理 1.1、定理 1.2、定理 1.3 的证明分别参见文献[1]的第 156 页、第 211 页、第 231 页.

例 1.1(抽样模型) 一批产品共有 N 件,其中 M 件是不合格品. 从中随机取出 n 件产品,求恰好取出了 m 件不合格品的概率.

解 设 X 表示取出的 n 件产品中不合格品的个数.

(1) 若是不放回抽样(如对产品进行有损伤检验),则 $X \sim Hg(n,N,M)$,恰好取出了 m 件不合格品的概率

$$P(X=m)=\binom{M}{m}\binom{N-M}{n-m}\Big/\binom{N}{n},\quad m=0,1,\cdots,\min\{n,M\}.$$

(2) 若是有放回抽样，则 $X\sim B\left(n,\frac{M}{N}\right)$，恰好取出了 m 件不合格品的概率

$$P(X=m)=\binom{n}{m}\left(\frac{M}{N}\right)^m\left(1-\frac{M}{N}\right)^{n-m},\quad m=0,1,\cdots,n.$$

二项分布是抽样调查计数问题的基本概率模型. 在应用中，当抽样比 $\frac{n}{N}\leqslant\frac{1}{10}$ 时，即可用二项分布 $B\left(n,\frac{M}{N}\right)$ 替代超几何分布 $Hg(n,N,M)$ 进行有关概率的计算.

1.1.2 泊松分布

泊松(Poisson)分布是一种重要的计数值分布，是稀有事件(正常情况下事件发生的可能性很小)发生次数的概率模型.

定理 1.4(泊松定理) 设 $\lim\limits_{n\to+\infty}np_n=\lambda$，$0<p_n<1$，则

$$\lim_{n\to+\infty}\binom{n}{x}p_n^x(1-p_n)^{n-x}=\frac{\lambda^x}{x!}e^{-\lambda}.$$

对定理 1.4 可以做如下解读：设 X_n 表示在 n 次独立试验中事件 A 发生的次数，由伯努利大数定律，$P(A)\approx\frac{X_n}{n}=p_n$，近似地 $X_n\sim B(n,p_n)$，则

$$P(X_n=x)\approx\binom{n}{x}p_n^x(1-p_n)^{n-x},\quad x=0,1,\cdots,n.$$

定理 1.4 表明，当 $\lim\limits_{n\to+\infty}E(X_n)=\lambda$，即事件 A 发生次数的均值与试验次数 n 的增大无关而稳定于常数 λ 时，$P(X_n=x)\approx\frac{\lambda^x}{x!}e^{-\lambda}$，$x=0,1,\cdots,\infty$.

容易证明 $\sum\limits_{x=0}^{\infty}\frac{\lambda^x}{x!}e^{-\lambda}=1$，这表明由定理 1.4 得到了一种新的概率分布.

定义 1.5 设 X 是一非负整数值随机变量，若 X 的概率函数为

$$P(X=x)=\frac{\lambda^x}{x!}e^{-\lambda},\quad x=0,1,\cdots,\infty,$$

则称 X 服从参数为 λ 的**泊松分布**，记为 $X\sim P(\lambda)$.

容易证明，若 $X\sim P(\lambda)$，则 $E(X)=\lambda$，$\mathrm{Var}(X)=\lambda$.

泊松分布的参数 λ 对其分布形态的影响如图 1.2 所示. 分别取 $\lambda=0.5$，$\lambda=3.0$，$\lambda=7.0$，画出 $x=0,1,2,\cdots,\infty$ 各点处的概率.

由图 1.2 可知，当 λ 值较小时分布是偏态的；随着 λ 值的增大，分布呈现出关于均值 λ 的对称形态. 在应用中，当 $\lambda\geqslant5$ 时可按对称分布处理.

泊松分布关于参数 λ 具有再生性.

定理 1.5 设 $X_k\sim P(\lambda_k)$，$k=1,2,\cdots,m$ 且相互独立，则 $X=\sum\limits_{k=1}^{m}X_k\sim P\left(\sum\limits_{k=1}^{m}\lambda_k\right)$.

定理 1.4 和定理 1.5 的证明分别参见文献[1]的第 95 页和第 115 页.

应用中，在某个试验单元(约定的单位试验时间与空间)上观察某一随机事件 A，X 表示事件 A 发生的次数，当事件 A 满足下列假定

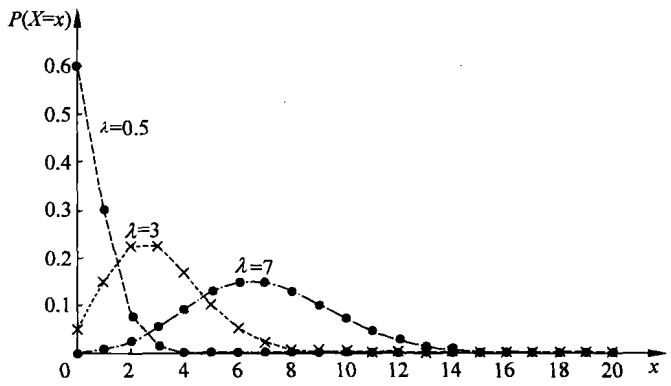

图 1.2 参数 λ 不同取值的泊松分布

(1) **独立性** 在不重叠的试验单元上事件 A 的发生是相互独立的;即在前一个试验单元事件 A 发生与否不改变后一试验单元事件 A 发生的概率.

(2) **平稳性** 在任何等测度的试验单元(时间长度、空间的面积或体积相等)上,事件 A 发生的期望次数 $E(X)=\lambda$ 是一常数;或者说事件 A 发生的平均次数只与试验单元的测度有关,而与试验单元的位置无关.

(3) **普通性** 在任何一个时刻点最多可观测到事件 A 的一次发生;即在一个瞬间事件 A 不会出现爆发性发生.

此时,可以认为 X 服从参数为 λ 的泊松分布,λ 是事件 A 的强度(试验单元上事件 A 发生的平均次数).

在二项分布的应用问题中,若分布参数 p 很小,如 $p=0.001$,用正态分布近似二项分布需 $n\geqslant\dfrac{5}{\min\{p,1-p\}}=5\,000$ 时才能得到较好的近似效果,显然这是一个不容易满足的条件. 研究表明,当参数 p 很小且 $1\leqslant np<5$ 时,根据泊松定理用泊松分布近似二项分布便能得到较好的效果.

例 1.2(α 粒子数模型) 1910 年,卢瑟福(Rutherford)等人就"放射性物质放射出的 α 粒子数的概率分布"问题进行了著名的实验研究. 观察体积一定的小块放射性物质在 7.5 s 的时间区间内、在指定区域上记录到的质点数(α 粒子在感光材料上的影像),观察重复进行了 2 608 次,表 1.1 是整理出的实验数据与分析. 其中,泊松分布的参数 λ 按其理论意义"试验单元事件发生的平均次数",用 $\lambda\approx\dfrac{1}{N}\sum\limits_{k=0}^{10}kN_k=3.87$ 估算.

表 1.1 卢瑟福实验数据与泊松分布理论计算值对比表

粒子数 k	观察到的频数 N_k	实际发生的频率 $p_k^*=\dfrac{N_k}{N}$	按泊松分布计算出的概率 $p_k=\dfrac{\lambda^k}{k!}e^{-\lambda}$
0	57	0.022	0.021
1	203	0.078	0.081
2	383	0.147	0.156
3	525	0.201	0.201