高等学校计算机科学与技术教材

- 原理与技术的完美结合
- 教学与科研的最新成果
- 语言精炼，实例丰富
- 可操作性强，实用性突出

试验数据及图像计算机处理

刘蓉生　主编

清华大学出版社 · 北京交通大学出版社

高等学校计算机科学与技术教材

试验数据及图像计算机处理

刘蓉生 主 编

清 华 大 学 出 版 社
北京交通大学出版社

·北京·

内容简介

运用数理统计的方法处理试验数据是自然与社会科学研究中常用的方法。为了使学习者能更好地掌握SPSS软件，本书介绍了统计学中的一些重要概念和方法，重点介绍了应用性很强的统计推断方法和回归方法及其计算机实现。Microcal Origin 6.1是在国外工程技术人员和大学生中广泛使用的科技数据分析及作图软件，本书介绍了该软件的基本使用方法。计算机图像处理部分主要介绍了计算机图像处理的硬件系统、计算机图像处理的技术基础、图像的浏览管理及截取技术和图像的编辑与处理技术等内容。

书中涉及了与试验数据及图像处理有关的六大主流软件，充分掌握这些软件必将大大提高读者的工作、研究水平及其效率。

本书可作为高等学校工科类相关专业的教材，也可作为科研人员的参考书。

图书在版编目（CIP）数据

试验数据及图像计算机处理／刘蓉生主编．—北京：清华大学出版社；北京交通大学出版社，2005.9
（高等学校计算机科学与技术教材）
ISBN 7－81082－468－6

Ⅰ．试…　Ⅱ．刘…　Ⅲ．统计分析－软件包，SPSS－高等学校－教材　Ⅳ．C819

中国版本图书馆CIP数据核字（2005）第097589号

责任编辑：黎　丹
出 版 者：清华大学出版社　邮编：100084　电话：010－62776969　http://www.tup.com.cn
　　　　　北京交通大学出版社　邮编：100044　电话：010－51686414　http://press.bjtu.edu.cn
印 刷 者：北京东光印刷厂
发 行 者：新华书店总店北京发行所
开　　本：185×260　印张：19　字数：475千字
版　　次：2005年9月第1版　2005年9月第1次印刷
书　　号：ISBN 7－81082－468－6/C·11
印　　数：1～4 000册　定价：26.00元

本书如有质量问题，请向北京交通大学出版社质监组反映。对您的意见和批评，我们表示欢迎和感谢。
投诉电话：010－51686043，51686008；传真：010－62225406；E-mail：press@center.bjtu.edu.cn。

前　言

从古到今，大概没有哪一项技术与现今人类的生活及工作的关系比计算机更密切，更没有哪一项技术的发展速度可以与计算机相比。笔者从事材料领域教学和研究约 30 年，深感计算机软硬件技术的飞跃发展对自然科学工作者和社会科学工作者带来的便利，因而要由衷地说一句："生活在现代社会的自然科学工作者或社会科学工作者真是太幸福了！"

现在回过头来看，如果没有现代计算机的软硬件技术，人们要做研究将会困难重重。不用说研究之前的调研、资料查询及试验过程获取数据的过程，仅就面对大量试验数据的建表、计算、作曲线、分析，直至最后的撰写论文或研究报告，如果没有计算机和有关的软件，那么工作将非常麻烦，有时甚至是寸步难行。

根据笔者的经验，本书采用了对于自然科学工作者和社会科学工作者处理试验数据和图像最常用的方法和技术，选用当今最热门的相关软件，其目的旨在帮助人们充分利用计算机来更好地完成自己的研究工作，从试验数据中获取更多的信息，以便编写出高质量的、图文并茂的论文及研究报告。

本书由"统计学基本知识"、"统计软件 SPSS 10.0 for Windows"、"数据分析及科技作图软件 Microcal Origin 6.1"和"计算机图像处理"四大部分组成。

统计学基本知识部分

运用数理统计的方法处理试验数据是自然科学与社会科学研究中常用的方法。为了使学习者能更好的掌握 SPSS 软件，在第 1 篇首先介绍了统计学中的一些重要概念和方法，并贯彻"不求系统性，但求实用性"的基本编写思路，重点介绍应用性很强的统计推断方法和回归方法。

统计软件 SPSS 10.0 for Windows 部分

介绍目前世界广泛应用的三大统计软件之一的 SPSS，该部分是对第一部分所选统计方法的计算机实现。介绍 SPSS 的专著已不少，因此本书在介绍 SPSS 时，也尽量围绕第一部分所选的内容进行，主要包括：统计描述、均值检验、分布检验、方差分析、回归分析及常用的统计图。

Microcal Origin 6.1 部分

Origin 是一个在国外工程技术人员和大学生中广泛使用的科技数据分析及作图软件，据笔者所见，国内还没有专门介绍该软件的专著及章节。笔者根据对该软件使用的体会，介绍了该软件的基本使用方法。

计算机图像处理部分

主要介绍了"计算机图像处理的硬件系统"、"计算机图像处理的技术基础"、"图像的浏览、管理及截取技术"、"图像的编辑与处理技术"等内容。这是一个应用十分广泛的领域，除人们的日常生活外，在许多专业领域，如材料、医疗、刑侦、新闻、艺术等都大量采用数字图像。可以预料，在不久的将来，人们终将会摒弃暗室成像技术，因为数字图像易于

编辑处理且其方法简单，甚至在用“Word”编写图文文档时，“Word”自带的图像处理工具就能把文档中的图像提高到一个新的水平。

最后，本书第4篇由成都航空职业技术学院龚宏编写，其余部分由刘蓉生编写。

刘蓉生

2005年9月

目　录

第 1 篇　统计学基本知识

本篇要点

- ☞ 统计学基本概念
- ☞ 两种常用的概率分布
- ☞ 随机误差的分布及实验数据精密度评定
- ☞ 随机变量函数的分布
- ☞ 统计推断及其基本方法
- ☞ 回归分析简介及一元线性回归分析
- ☞ 曲线回归及多项式回归
- ☞ 多元线性回归分析

第 1 章　统计学基本概念

1.1　频率、概率和贝努里定理

频率与概率都是用来表征某个事件在一定条件下发生与否的可能性大小的量，但二者具有不同的概念。

1. 频率（Frequency）

设在 n 次试验中，某随机事件 A 出现次数为 v，则称 $f(A)=\dfrac{v}{n}$ 为事件 A 出现的频率。频率具有不稳定性。例如，抛硬币实验，若抛有限次数，则每次实验正反面出现的次数都不一样，其频率也不一样。

2. 概率（Probability）

反映随机事件在某个条件下发生与否的可能性大小的客观定值。上述抛硬币实验历史上已有人做过，其结果见表 1－1。

表 1-1　抛硬币试验

试验者	试验次数 n	出现正面次数 v	频率$\left(\frac{v}{n}\right)$
蒲　丰	4 040	2 048	0.506 9
皮尔逊	12 000	6 016	0.501 6
皮尔逊	24 000	12 012	0.500 5

可见，随着实验次数 n 的增加，频率趋向 0.5，该值即是抛硬币试验的概率值。这暗示了频率与概率的关系，即有如下定理。

3. 贝努里定理（Bernoulli Theorem）

设 v 为 n 次试验中事件 A 出现的次数，P 是事件 A 在每次试验中出现的概率，于是对任意小的正数 ε（$\varepsilon>0$）均有下列关系成立。

$$\lim_{n\to\infty}P\left\{\left|\frac{v}{n}-P\right|<\varepsilon\right\}=1$$

这说明在 $n\to\infty$ 时，频率$\frac{v}{n}$与概率 P 之间的差小于任意小正数 ε 的概率趋近于 1。也就是说，当试验无限地做下去时，事件 A 发生的频率将与它的概率趋于一致。这就是所谓频率稳定性的由来。这个定理是频率与概率之间的联系桥梁。由此，人们实际上往往用频率来代替概率。

从频率的角度考虑概率具有更普遍的意义，在讨论随机变量分布规律时，常用频率分布直方图的面积来表示随机变量在某个范围内出现的频率。当试验次数无限增加时，则用光滑曲线代替曲边梯形，其曲线包围的面积即表示概率，如图 1-1 所示。

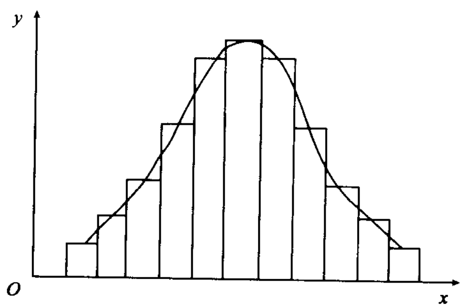

图 1-1　频率直方图与概率分布图

概率具有如下几个基本性质。

性质 1　任一事件 A 的概率 $P(A)$ 必介于 0 和 1 之间，即 $0\leqslant P(A)\leqslant 1$。

证　因为 v 永远在 0 和 n 之间，所以频率$\frac{v}{n}$必然在 0 和 1 之间。当 n 不断增加时，频率近似于概率，故 $0\leqslant P(A)\leqslant 1$。

性质 2　必然事件 U 的概率等于 1，即 $P(U)=1$。

证　因为对于必然事件，$v=n$，其频率$\frac{v}{n}=1$，故有 $P(U)=1$。

性质 3　不可能事件 V 的概率等于 0，$P(V)=0$。

证　对于不可能事件，$v=0$，其频率$\frac{v}{n}=0$，故有 $P(V)=0$。

如果任一事件 A 发生的概率很小，则称该事件为“小概率事件”。从实验观点看，可以认为小概率事件在一次试验中基本不会出现；反之，对于大概率事件，则可认为在一次试验中总会出现。

1.2　随机变量及其概率分布

1. 随机事件（Random Event）

从试验角度讲，通常把一次独立试验的结果称为一个事件。当进行试验时，如果某一事件在试验后有时发生，有时不发生，则称此事件为随机事件。为了从量的方面去讨论随机事件，便引出了随机变量的概念。

2. 随机变量（Random Variable）

随机事件数量性的表征称为随机变量。由个别试验结果所得到的每个实数值都是一个随机变量。例如，投掷正立方体，其随机变量的取值分别为 1，2，…，6，并且随机变量是以 1/6 的概率来取得每个数值的。

一般设 ξ 为随机变量，x 为 ξ 的观测值，x_1，x_2，…，x_n 表示 ξ 的 1，2，…，n 个观测值。

随机变量分为离散型随机变量（Discrete type random variable）和连续型随机变量（Random variable of continuous type）两种。ξ 在某区间内取有限多个数据者称为离散型随机变量；ξ 在某区间内可取任意多个数值者称为连续型随机变量。前者如投掷正立方体，后者如一批电视机的寿命。

需要指出的是，同一事件往往有不同的随机变量。例如，将废品钢锭看作一随机事件，讨论废品钢锭的根数是一种随机变量，讨论废品钢锭某一成分的含量又是一种随机变量。

3. 随机变量的概率分布

随机变量 ξ 的取值 x_i 与它出现的概率 P_i 之间的关系称为随机变量的概率分布（Probability Distribution）。设某连续型随机变量与其概率有如图 1-2 的关系，则分别称：

曲线——ξ 的概率分布曲线（Probability Distribution Curve）；

曲线方程 $y=\varphi(x)$——ξ 的概率分布密度函数（Probability Distribution Density Function）

阴影面积——ξ 的概率分布函数（Probability Distribution Function），即

$$F(x) = P\{a < x < b\} = \int_a^b \varphi(x)\mathrm{d}x$$

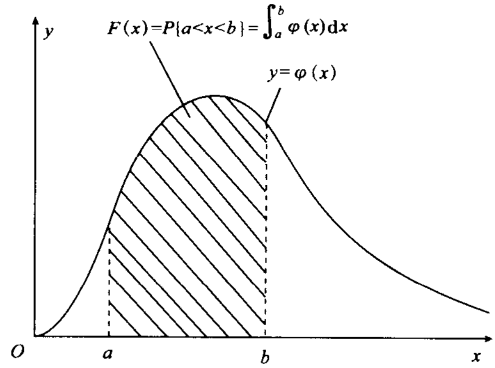

图 1-2 密度函数与分布函数

可知，分布函数 $F(x)$ 实际是连续型随机变量 ξ 在某个区间出现的概率，是 ξ 最完善的描述。当知道一个随机变量的分布函数时，不仅知道该变量取哪些值，还可知道它是以什么样的概率取这些值的。

1.3 随机变量的数字特征

如上所述，若已知随机变量的分布函数，则不仅能知道这个随机变量可取得哪些值，而且还能知道它以什么概率来取得这些值。但有时没有必要知道那么多详细情况，只要知道几个特征值就行了，这些特征值分别为数学期望、中值和方差。

1. 数学期望（Mathematical Expectation）

表征随机变量 ξ 在数轴上取值的集中位置，说明 x_i 值大多数在哪里，数学期望可看成是平均值概念在随机变量方面的推广。

（1）离散型随机变量的数学期望

离散型随机变量的数学期望为

$$E(\xi) = \sum_i x_i P_i \tag{1-1}$$

式中：x_i——离散型随机变量的值；

P_i——该随机变量取值发生的概率。

证 设试验次数 N 很大，某离散型随机变量取值为 x_1，x_2，…，x_n，相应出现的次数分别为 m_1，m_2，…，m_n，其平均值为 $\bar{x}$，则有

$$\begin{aligned}\bar{x} &= \frac{m_1x_1 + m_2x_2 + \cdots + m_nx_n}{N} \\ &= \frac{m_1}{N}x_1 + \frac{m_2}{N}x_2 + \cdots + \frac{m_n}{N}x_n \\ &= f_1x_1 + f_2x_2 + \cdots + f_nx_n\end{aligned}$$

当 N 很大时，频率 f_i 近似地等于值 x_i 发生的概率 P_i，因此有

$$E(\xi) = \bar{x} = P_1x_1 + P_2x_2 + \cdots + P_nx_n = \sum_i x_iP_i \tag{1-2}$$

可见，离散型随机变量的数学期望是以概率为权的加权平均值，且当试验次数越多，此式越正确。

（2）连续型随机变量的数学期望

连续型随机变量的数学期望为

$$E(\xi) = \int_{-\infty}^{+\infty} x\varphi(x)\mathrm{d}x$$

证　设 $\varphi(x)$ 在区间（$-\infty$，$+\infty$）上有定义，在（x，$x+\Delta x$）区间上 ξ 可取任何值，其出现的概率为 $\varphi(x)\Delta x$，近似地数学期望为

$$E(\xi) = \sum_i x\varphi(x)\Delta x$$

当 $\Delta x \to 0$ 时，上式为

$$E(\xi) = \int_{-\infty}^{+\infty} x\varphi(x)\mathrm{d}x$$

符号 $E(\xi)$ 可看作是施加在 ξ 上的一种运算。

数学期望的几何意义见图1-3。在图1-3中，$\varphi(x)\mathrm{d}x$ 表示在任一 x 处的微面积，该微面积对纵坐标轴的静矩为 $x\varphi(x)\mathrm{d}x$，故整个曲线所包围的面积对纵坐标轴的静矩为

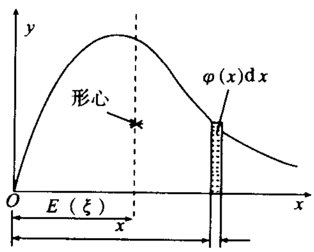

图1-3　数学期望的几何意义

$$J = \int_0^{+\infty} x\varphi(x)\mathrm{d}x$$

将静矩除以总面积，即得到面积的形心位置为

$$形心位置 = \frac{\int_0^{+\infty} x\varphi(x)\mathrm{d}x}{\int_0^{+\infty} \varphi(x)\mathrm{d}x} = \int_0^{+\infty} x\varphi(x)\mathrm{d}x$$

可见，$E(\xi)$ 表示分布曲线所包围面积的形心位置。

2. 中值（Medium Value）

平分频率曲线所包围面积的垂直线的横坐标称为中值，以符号 x_{50} 表示之。因为分布曲线所包围的面积为1，所以中值垂线左右两部分面积各等于1/2。根据这一性质，当已知随机变量的密度函数 $\varphi(x)$ 时，可按以下条件求出中值 x_{50} 的大小。

$$\int_{-\infty}^{x_{50}} \varphi(x)\mathrm{d}x = \frac{1}{2}$$

$$\int_{x_{50}}^{\infty}\varphi(x)\,\mathrm{d}x = \frac{1}{2}$$

数学期望与中值是两个不同的概念。当分布曲线左右两部分对称时，则数学期望与中值两者重合；当分布曲线不对称时，若曲线高峰向左偏斜，则数学期望一般大于中值。

3. 方差（Variance）

方差是表征随机变量ξ围绕数学期望的离散程度，也就是衡量随机变量分布的分散程度的特征值。方差的定义式为

$$\mathrm{Var}(\xi) = E[\xi - E(\xi)]^2 \tag{1-3}$$

即随机变量与其理论均值$E(\xi)$之差的平方的数学期望。

离散型随机变量的方差为

$$\mathrm{Var}(\xi) = \sum_i [x_i - E(\xi)]^2 P_i \tag{1-4}$$

其中：$E(\xi)$——ξ的数学期望；

P_i——ξ值为x_i发生的概率。

连续型随机变量的方差为

$$\mathrm{Var}(\xi) = \int_{-\infty}^{+\infty}[x - E(\xi)]^2\varphi(x)\,\mathrm{d}x \tag{1-5}$$

可以从对一批数据分散程度的描述来理解式(1-3)。设一批数据为x_1，x_2，…，x_n，每个数据的残差为$\delta_i = x_i - \bar{x}$，取其平方和并用数据个数n去除，则有

$$\frac{\sum (x_i - \bar{x})^2}{n} \tag{1-6}$$

而符号$\frac{\sum}{n}$是平均值的概念，$\bar{x}$也是平均值。对于随机变量，均值可理解为数学期望。可见，式(1-3)与式(1-6)具有相同的形式。

方差也可从几何意义来理解，如图1-3，若将通过形心的垂线作为曲线所包围面积的中心线，则$x - E(\xi)$表示任一微面积$\varphi(x)\,\mathrm{d}x$与中心线的距离。按照力学中惯性矩的定义，该距离的平方$[x - E(\xi)]^2$和微面积$\varphi(x)\,\mathrm{d}x$乘积的积分等于整个面积对中心线的惯性矩。可见，按照式(1-5)，方差可看作是分布曲线所包围面积的中心惯性矩。随机变量距中心线的分布越远，惯性矩就越大，即方差越大。

数学期望和方差均为常量，二者存在如下关系。

$$\mathrm{Var}(\xi) = E(\xi^2) - [E(\xi)]^2 \tag{1-7}$$

证

$$\begin{aligned}\mathrm{Var}(\xi) &= \int_{-\infty}^{+\infty}[x - E(\xi)]^2\varphi(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{+\infty}\{x^2 - 2xE(\xi) + [E(\xi)]^2\}\varphi(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{+\infty}x^2\varphi(x)\,\mathrm{d}x - 2E(\xi)\int_{-\infty}^{+\infty}x\varphi(x)\,\mathrm{d}x + \\ &\quad [E(\xi)]^2\int_{-\infty}^{+\infty}\varphi(x)\,\mathrm{d}x\end{aligned}$$

因为$\int_{-\infty}^{+\infty}\varphi(x)\mathrm{d}x=1$,并由式(1-2)将“$E(\ )$”施加在$\xi^2$上运算,当随机变量为$\xi^2$时,其取值为$x^2$, 于是有

$$E(\xi^2)=\int_{-\infty}^{+\infty}x^2\varphi(x)\mathrm{d}x$$

这样方差表达式可简化为

$$\begin{aligned}\mathrm{Var}(\xi)&=E(\xi^2)-2E(\xi)E(\xi)+[E(\xi)]^2\\&=E(\xi^2)-[E(\xi)]^2\end{aligned}$$

最后, 给出标准差的概念, 即方差的平方根$\sqrt{\mathrm{Var}(\xi)}$, 称为随机变量 ξ 的标准差(Standard Deviation)。

第 2 章　两种常用的概率分布

2.1　随机变量的正态分布

正态分布是数理统计学中最重要的分布，应用非常广泛，如材料性能实验、产品寿命研究、化学成分、测量误差、医学检测等，都可用正态分布加以处理。

1. 正态分布的概率分布密度函数

正态分布的概率分布密度函数为

$$\varphi(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{2-1}$$

可知，除随机变量 ξ 的取值 x 外，还有两个参量 μ 和 σ^2。随机变量遵从正态分布，简记为 $x\sim N(\mu,\ \sigma^2)$。

2. 正态分布的概率分布函数

正态分布的概率分布函数为

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]dx,\ x\in(-\infty,+\infty)$$

随机变量取值 x 发生在区间 $(a,\ b)$ 的概率为

$$F(x)=P\{a<x<b\}=\int_a^b\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]dx$$

3. 正态分布的概率分布曲线

正态分布的概率分布曲线具有如下特征（参见图 2－1）。

① 曲线的纵坐标恒为非负值；

② 观测值在平均值附近出现的机会最多，故曲线存在一个高峰；

③ 曲线有一个对称轴，对称轴两侧的发生概率相等；

④ 对称轴两边曲线上相同位置上各有一拐点。