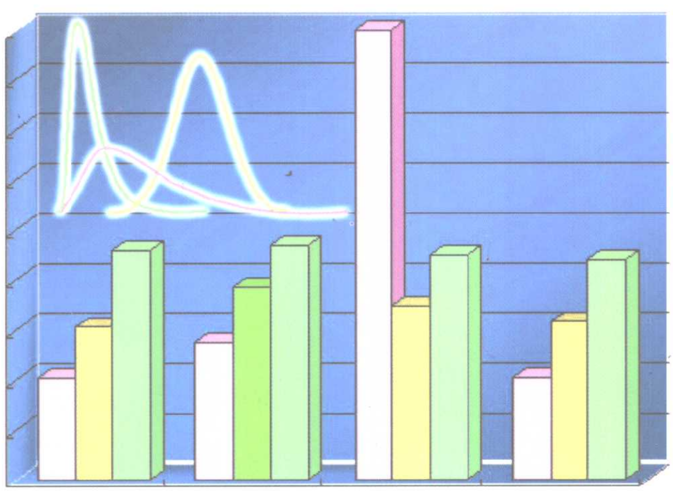
MATLAB语言应用系列书

MATLAB数理统计(6.x)

陈桂明 戚红雨 潘 伟 编著



科学出版社

MATLAB语言应用系列书

MATLAB数理统计(6.x)

陈桂明　戚红雨　潘　伟　编著

科学出版社

2002

内 容 简 介

本书结合数据统计概念、理论和应用，以 MATLAB 6.1 为对象，系统地介绍了统计工具箱中的概率分布、估计、假设检验、多变量统计、聚类分析、试验设计、线性和非线性模型以及在数理统计中的应用等内容。本书重点是运用 MATLAB 统计工具箱介绍统计分析研究中的各种概念、理论、方法、算法及其实现。

本书内容安排合理，理论结合实际，并列举了大量作者总结的应用实例，书中讲述的各种统计理论和方法浅显易懂，并能在实际生活中找到应用对象。

本书可作为高等学校理科、工科、文科及管理学科等有关专业师生的参考教材及自学用书，对从事上述领域工作的广大科技工作者和开发应用人员具有重要的参考价值。

图书在版编目(CIP)数据

MATLAB 数理统计(6.x)，陈桂明等编著．—北京：科学出版社，2002
(MATLAB 语言应用系列书)
ISBN 7-03-010095-6

Ⅰ.M... Ⅱ.陈... Ⅲ.数理统计-计算机辅助计算-软件包，MATLAB(6.x) Ⅳ.TP391.75

中国版本图书馆 CIP 数据核字(2002)第 005410 号

科学出版社 出版
北京东黄城根北街16号
邮政编码：100717
http://www.sciencep.com
双青印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2002 年 3 月第 一 版　　开本：787×1092 1/16
2002 年 3 月第一次印刷　　印张：16
印数：1—4 000　　字数：375 000

定价：28.00 元

(如有印装质量问题，我社负责调换〈环伟〉)

前　言

MATLAB 语言近年来在全国各高校、科研院所中风行一时，由于其科学计算方面的独特优势而深受大学生、研究生及科研人员喜爱。人们在学习新课程时用它来加深理解，熟悉算法；在进行课程设计时用它来进行计算、试验和仿真；在科研中用它来解决复杂的问题。

在各工科院校中，几乎都开设概率论与数理统计课程，而书本上的理论往往不容易理解，而且计算又十分繁杂，MATLAB 统计工具箱几乎包括了数理统计方面的所有概念、理论、方法和算法，本书运用 MATLAB 统计工具箱介绍统计分析研究中的各种概念、理论、方法、算法及其实现。读者可以通过本书的学习，很快掌握所学内容，并十分方便地通过列举实例，在计算机上进行计算，从而进一步加深理解；在实际科研工作中，应用 MATLAB 进行统计分析，可以大大减轻计算工作量。

本书的特点在于将不易理解的理论和难于手工运算的方法用 MATLAB(6.1 版)语言的函数来进行介绍，通过运行 MATLAB 程序来了解这些理论和过程。

同时 MATLAB 强大的图形功能，使得概念、过程和结果可以直观地展现在读者面前。

本书在内容安排上，理论结合实际，用浅显易懂的例子来说明复杂难懂的统计理论和方法。

本书可作为理科、工科、管理学科和文科等学科有关专业的大学生、研究生和教师，以及有关的科研人员作为学习数理统计课和研究有关问题的教材或参考书。

全书内容共分 10 章，第 1 章主要介绍数理统计中一些基本概念和各种重要的概率分布，并说明了在 MATLAB 中这些分布函数的应用及其实际意义。第 2 章详细阐述了各类分布随机变量的数字特征的含义及其计算方法。第 3 章介绍了各种参数估计方法的概念以及在 MATLAB 中的实现方法和函数。第 4 章叙述了如何应用 MATLAB 实现假设检验。第 5 章直观地展示了 MATLAB 各种统计绘图函数及工具的用法。第 6 章介绍了常用线性模型的 MATLAB 函数，包括方差分析、回归分析、多项式拟合、鲁棒回归和非参数方法等内容。第 7 章论述了非线性回归模型几个主要函数的应用。第 8 章论述了常见的几种聚类分析方法和 MATLAB 函数。第 9 章阐述了主成分分析、多变量方差分析和线性判别分析方法。第 10 章论述了常用的几种试验设计的概念和方法，以及用 MATLAB 实现试验设计的方法。

由于作者水平所限，错误与不当之处在所难免，恳请广大读者批评指正。作者电子邮箱：chenguiming@yahoo.com

作者

2002 年 1 月

目　　录

第 1 章　随机变量及其概率分布

1.1　随机变量及其概率分布

1.1.1　随机变量

首先介绍数理统计中的几个重要概念。

(1) 试验

对一定条件下发生或出现对象进行的观察，通常叫做“试验”。条件实现一次就是一次试验。这里的“试验”一词，包含诸如“实验”、“观察”及“测量”等概念。

(2) 随机试验

如果试验可以在相同的条件下重复进行，而且每次试验的可能结果不少于一个，且试验之前不能确定哪一个结果会出现，而只能明确试验结果可能在某个范围内，那么我们就称这个试验为随机试验。

(3) 样本空间

一个随机试验的一切可能结果所组成的集合称为该随机试验的样本空间。

(4) 随机事件

在随机试验中，对每一次试验可能出现也可能不出现，而在大量重复试验中却具有某种规律性的某一结果，称为一个随机事件，简称事件。

(5) 概率

对于一个随机事件 A，用区间[0,1]上的一个数 P(A)来表示该事件出现可能性大小的尺度，这个数 P(A)就称为随机事件 A 的概率。

(6) 随机变量

与随机事件相对应，在一次随机试验中可能取这个或那个数值，而不能预先断定取什么值的变量叫随机变量。随机变量常用大写字母 $X,Y,Z,\cdots$ 代表。随机变量可分为离散型和非离散型两类。离散型随机变量是指它的全部可能取到的值为有限个或为可列无限个；否则称为非离散型随机变量。非离散型随机变量中最重要的是连续型随机变量。

1.1.2　随机变量的概率分布

1. 分布函数

设 X 是一个随机变量，称

$$F(x) = P\{X \leqslant x\},\ -\infty < x < +\infty \tag{1.1}$$

为随机变量 X 的概率分布函数，简称为随机变量的分布函数(又称为累积分布函数)。

分布函数的性质：

(1) $F(x)$是 x 的不减函数，即对于任意的 $x_2 > x_1$，有

$$F(x_2) \geqslant F(x_1)$$

(2) $0 \leqslant F(x) \leqslant 1$，且有

$$\lim_{x \to -\infty} F(x) = 0, \text{记为 } F(-\infty) = 0$$
$$\lim_{x \to +\infty} F(x) = 1, \text{记为 } F(+\infty) = 1$$

(3) $F(x)$在任一点 x_0 处右连续，即

$$\lim_{x \to x_0+0} F(x) = F(x_0)$$

2. 离散型随机变量及其分布

离散型随机变量只可能取有限个或可列个值。设离散型随机变量 X 的所有可能取值为 $x_1, x_2, \cdots, x_k, \cdots$，而相应于取这些值的概率为 $P\{X=x_1\}=p_1, P\{X=x_2\}=p_2, \cdots, P\{X=x_k\}=p_k, \cdots$。一般地，写为

$$P\{X = x_k\} = p_k, k = 1,2,\cdots \tag{1.2}$$

则式(1.2)称为离散型随机变量 X 的概率分布(也叫离散随机变量的概率密度)。这个概率分布应该满足条件

(1) $0 \leqslant p_k \leqslant 1, k=1,2,\cdots$

(2) $\sum_{0}^{\infty} p_k = 1$

根据分布函数的定义，离散型随机变量 X 的分布函数为

$$F(x) = P\{X \leqslant x\} = \sum_{x_i \leqslant x} p_i \tag{1.3}$$

3. 连续型随机变量及其分布

设随机变量 X 的分布函数为 $F(x)$，如果存在非负函数 $f(x)$使得对于任意连续型随机变量 X，$F(x)=P\{X \leqslant x\}=\int_{-\infty}^{x} f(x)\mathrm{d}x$，其中 $f(x)$称为 X 的概率密度函数或分布密度函数，简称为概率密度或分布密度函数。

概率密度函数 $f(x)$具有下列性质：

(1) $f(x) \geqslant 0$

(2) $\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1$

4. 逆分布函数

逆(累积)分布函数返回给定显著概率条件下假设检验的临界值，实际上是分布函数的逆函数。

对于连续变量分布，计算分布函数的返回值介于 0 和 1 之间的概率，对这些概率进行逆累积分布计算，则会得到那些原始的值。对于离散分布，分布函数与逆分布函数之间的关系更复杂。

逆分布函数可用来计算各种分布的分位数，在进行假设检验和计算置信区间时很有用。

1.1.3 总体与样本

数理统计讨论问题的出发点是试验(观察)数据，数理统计的任务是研究如何有效地

搜集、整理和分析受随机因素影响的试验(观察)数据,并对所观察问题的整体特性做出推测和判断。

试验有全面试验和抽样试验(或称局部试验)两种形式。对研究的对象全部逐个进行试验叫做全面试验,如人口普查、产品质量的逐个检查都是全面试验。从研究对象中随机抽取一小部分逐个试验叫做抽样试验。全面试验不但费时费工,不经济,而且对破坏性的试验,如灯泡的寿命试验及金属材料的强度试验等,也不允许进行全面试验。所以,抽样试验是最基本的试验,数理统计中的试验均为抽样试验,而总体与样本则是抽样试验中最常用的两个基本概念。

(1) 总体

由研究对象的某个数量指标的值的全体组成的集合(数集)叫做总体(或母体),总体中的每一个元素(实数)叫做个体,总体中的个体总数叫做总体的容量。

(2) 样本

从总体 X 中随机抽取的一部分个体叫做总体的一个样本(或子样),样本中的每一个个体叫做样品,通常用 X_i 表示第 i 个样品,样本中的样品总数叫做样本容量,抽取样本的过程叫做抽样。

每一个样品都有二重性。一方面,抽样前样品的取值具有随机性,它是一个随机变量;另一方面,抽样后的样品是个确定的实数值,并称为样品值。同理,容量为 n 的样本$(X_1, X_2, \cdots, X_n)$也有两重性,抽样前容量为 n 的样本是一个 n 维随机变量,抽样后容量为 n 的样本是一个确定的 n 维实数组$(x_1, x_2, \cdots, x_n)$,并称之为样本值。

(3) 简单随机样本

设$(X_1, X_2, \cdots, X_n)$是总体 X 的一个样本,若 $X_1, X_2, \cdots, X_n$ 相互独立,且每一个样品 $X_i (i=1,2,\cdots,n)$都与总体 X 有相同的分布,则称$(X_1, X_2, \cdots, X_n)$为总体 X 的一个简单随机样本,简称样本。在数理统计中所提及的样本均指简单随机样本。

(4) 统计量

设总体 X 的一个样本为$(X_1, X_2, \cdots, X_n)$,$g(x_1, x_2, \cdots, x_n)$为一个 n 元连续函数,如果函数 $g(X_1, X_2, \cdots, X_n)$中不含任何未知参数,则称 $g(X_1, X_2, \cdots, X_n)$是一个统计量。

统计量是一个随机变量。对确定的样本值$(x_1, x_2, \cdots, x_n)$可得统计量 $g(X_1, X_2, \cdots, X_n)$的观察值 $g(x_1, x_2, \cdots, x_n)$,叫做统计值。

常见的统计量是样本均值与样本方差。设总体的一个样本为$(X_1, X_2, \cdots, X_n)$,则

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1.4}$$

$$S^2 = \frac{1}{n-1}\sum_{n=1}^{n}(X_i - \overline{X})^2 \tag{1.5}$$

分别叫做样本均值和样本方差。

对确定的样本值$(x_1, x_2, \cdots, x_n)$,则

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.6}$$

$$s^2 = \frac{1}{n-1}\sum_{n=1}^{n}(x_i - \bar{x})^2 \tag{1.7}$$

分别是样本均值和样本方差的观察值。

(5) 抽样分布

统计量是样本的已知连续函数,它应有确定的分布,统计量的分布叫做抽样分布。

1.1.4 常见的一些重要概率分布

MATLAB 的统计工具箱中提供了 20 种常见的概率分布,分为连续分布和离散分布,见表 1.1。

表 1.1 概率分布分类表

连续随机变量分布	连续统计量分布	离散随机变量分布
β 分布	χ^2 分布	二项分布
连续均匀分布	非中心 χ^2 分布	离散均匀分布
γ(Gamma)分布	F 分布	几何分布
对数正态分布	非中心 F 分布	超几何分布
正态分布	t 分布	负二项分布
Rayleigh 分布	非中心 t 分布	泊松分布
指数分布		
Weibull 分布		

1. 连续随机变量的分布

(1) β 分布

β 分布描述了一类独特的曲线,它们只在(0,1)区间非零。该函数更一般的形式为区间的两个端点可以改变,赋以其它数值。β 分布的分布函数与不完全 β 函数相同。

β 分布与 t 分布之间存在函数关系。如果 Y 服从自由度为 n 的 t 分布,那么下面的变换将产生服从 β 分布的 X。

$$X = \frac{1}{2} + \frac{1}{2}\frac{Y}{\sqrt{n+Y^2}} \tag{1.8}$$

如果 Y 服从 $t(n)$ 分布,那么 X 服从 $\beta(\frac{n}{2},\frac{n}{2})$ 分布。

定义

β 分布的分布密度函数为

$$y = f(x|a,b) = \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}I_{(0,1)}(x) \tag{1.9}$$

其中 $B(\cdot)$ 为 β 函数,β 函数定义为 $B(a,b)=\int_0^1 u^{a-1}(1-u)^{b-1}\mathrm{d}u$。指示函数 $I_{(0,1)}$ 保证了

只有介于 0 和 1 之间的 x 才有非零的概率。

(2) 连续均匀分布

连续均匀分布(又称矩形分布)的概率密度函数在它的两个参数 a(最小)和 b(最大)之间为常数。标准均匀分布($a=0,b=1$)是 β 分布的一个特例(将两个参数均置为 1)。

均匀分布适于表示数值的圆整误差分布。

定义　均匀分布的概率分布密度函数为

$$y = f(x|a,b) = \frac{1}{b-a}I_{[a,b]}(x) \tag{1.10}$$

分布函数为

$$p = F(x|a,b) = \frac{x-a}{b-a}I_{[a,b]}(x) \tag{1.11}$$

(3) γ(Gamma)分布

γ 分布是基于两个参数的一族曲线。χ^2 分布和指数分布均是固定了其中一个参数而只有一个参数的 γ 分布的子类。γ 分布与非完全 γ 函数具有下列关系:

$$\Gamma(x|a,b) = \mathrm{gammanic}\left(\frac{x}{b},a\right) \tag{1.12}$$

当 $b=1$ 时,两个函数是相同的。

当 a 很大时,γ 分布近似于正态分布,而且 γ 分布还有一个优点就是只有正实数才有分布密度值。

定义　γ 分布的分布密度函数为

$$y = f(x|a,b) = \frac{1}{b^a\Gamma(a)}x^{a-1}\mathrm{e}^{-\frac{x}{b}} \tag{1.13}$$

其中 $\Gamma(\cdot)$为 γ 函数,定义为 $\Gamma(z)=\int_0^\infty u^{z-1}\mathrm{e}^{-u}\mathrm{d}u$。

(4) 指数分布

指数分布是 γ 分布的一个特例($a=1$)。

指数分布具有非常良好的特性,它在描述时间上随机发生的事件时很有应用价值。指数分布主要的应用领域是寿命研究。

定义　指数分布的分布密度函数为

$$y = f(x|\mu) = \frac{1}{\mu}\mathrm{e}^{-\frac{x}{\mu}} \tag{1.14}$$

(5) 正态分布

正态分布又叫高斯分布,是一种很重要的分布,实际上许多重要的随机变量都服从或近似服从正态分布。正态分布具有两个参数,第一个参数 μ 是分布的均值,第二个参数 σ 是分布的标准差。若随机变量 X 服从正态分布,常记作 $X\sim N(\mu,\sigma^2)$。在标准正态分布(记作 $N(0,1)$)中,$\mu=0,\sigma=1$。

正态分布可用作二项分布的连续近似。它的另外一个用途是,由中心极限定理可知,当被研究的随机变量可以表示成个数充分多的相互独立(或相依关系很弱)的随机变量的和,而且每个随机变量对总和的影响都很微小时,则总和随机变量必然服从或近似服从正态分布。

定义　正态分布的分布密度函数为

$$y=f(x|\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{1.15}$$

(6) 对数正态分布

正态分布与对数正态分布密切相关。如果随机变量 X 服从参数为 μ 和 σ^2 的对数正态分布,那么 $\ln X$ 服从参数为 μ 和 σ^2 的正态分布。服从对数正态分布的随机变量必须为正数,因为 $\ln X$ 只有在随机变量 X 为正时才存在。经济学家在描述收入分布时常使用对数正态分布。

定义 对数正态分布的分布密度函数为

$$y=f(x|\mu,\sigma)=\frac{1}{x\sigma\sqrt{2\pi}}e^{\frac{-(\ln x-\mu)^2}{2\sigma^2}} \tag{1.16}$$

(7) Weibull 分布

Waloddi Weibull 于 1939 年提出了一种适于描述材料断裂应力分布情况的分布,后来这种分布就以他的名字命名,称为 Weibull 分布。现在 Weibull 分布的应用还包括可靠性和寿命分析建模。Weibull 分布在这些方面应用时比指数分布要更灵活一些。

定义 Weibull 分布的分布密度函数为

$$y=f(x|a,b)=abx^{b-1}e^{-ax^b}I_{(0,\infty)}(x) \tag{1.17}$$

(8) Rayleigh 分布

Rayleigh 分布是 Weibull 分布的一个特例。如果 Weibull 分布的参数是 A 和 B,那么参数为 b 的 Rayleigh 分布等价于参数为 $A=1/(2b^2)$和 $B=2$ 的 Weibull 分布。

如果一个物体在 x 和 y 方向的两个速度分量都服从均值为 0 的正态分布,而且它们的方差相等且相互独立,那么这个物体单位时间的位移服从 Rayleigh 分布。

定义 Rayleigh 分布的分布密度函数为

$$y=f(x|b)=\frac{x}{b^2}e^{\frac{-x^2}{2b^2}} \tag{1.18}$$

2. 连续统计量的分布

(1) χ^2 分布

设 $X_1,X_2,\cdots,X_n$ 是 n 个相互独立的的标准正态随机变量,则它们的平方和

$$\chi^2=X_1^2+X_2^2+\cdots+X_n^2 \tag{1.19}$$

叫做自由度为 n 的 χ^2 变量,它的分布叫做自由度为 n 的 χ^2 分布,记做 $\chi^2\sim\chi^2(n)$。

χ^2 分布具有可加性,若 $X\sim\chi^2(n)$,$Y\sim\chi^2(m)$,且 X 与 Y 相互独立,则

$$X+Y\sim\chi^2(n+m) \tag{1.20}$$

χ^2 分布是 γ 分布的特例,它是 γ 分布密度函数关系式中参数 b 取 2 时的情况。

χ^2 分布是很重要的一种分布,因为它在正态抽样定理中有着重要应用。如果在一个含有 n 个样本的样本集中,所有的样本都服从方差为 σ^2 的正态分布,且 S^2 为样本方差,那么有$\frac{(n-1)S^2}{\sigma^2}$服从 $\chi^2(n-1)$分布。

定义 χ^2 分布的分布密度函数为

$$y=f(x|n)=\frac{x^{(n-2)/2}e^{-x/2}}{2^{n/2}\Gamma(n/2)} \tag{1.21}$$