前　言

随着计算机的发展与普及，数理统计已成为处理信息、进行决策的重要理论和方法。在科学研究中，用数理统计方法从数据中获取信息和判别初步规律，往往成为重大科学发现的先导。数理统计是数学方法与实际相结合应用最为广泛、最为重要的方式之一。因此，现代科研人员和工程技术人员应该具备数理统计的基础知识。而 MATLAB 则是一套高性能的数值计算和可视化软件，它集矩阵运算、数值分析、信号处理和图形显示于一体，构成了一个界面友好、使用方便的用户环境，是实现数据分析与处理的有效工具。

本书介绍了数理统计的基本原理、典型应用，以及使用作者开发的 MATLAB 程序代码进行实际数据分析的具体方法和步骤。全书共分 9 章，第 1 章概述了概率论基础；第 2～9 章依次介绍了描述性统计分析、参数估计、假设检验、方差分析、线性回归分析、曲线拟合分析、正交试验设计和判别分析的原理，同时给出了 MATLAB 程序源代码、.exe 程序应用实例。

本书是作者根据广大学生、科研人员、工程技术人员进行数据处理的需求而编写的，凝聚了作者近二十年来从事工科研究生、本科生数理统计和试验设计方法教学的经验以及参与工程研究项目和指导数学建模竞赛过程中的体会，是作者进行数理统计课程教学改革的研究成果。本教材具有以下特点：第一，注重数理统计的思想方法介绍。在阐述某一统计概念方法时，一般是从具体实例开始引出相关内容的客观背景，让学生带着实际问题去学习和思考。第二，注重应用性，数理统计是一门应用性很强的学科，其应用几乎遍及各个领域，成为解决实际问题的重要工具。因此，本教材充实了许多应用性内容，以适应读者解决实际问题的需要。第三，重视 MATLAB 应用对统计方法的简单性、实用性和可操作性。实际中，数据处理工作往往是庞大而繁琐的，使很多学生、科研人员、工程技术人员对此望而兴叹，感到无助。本书对每一章节的方法、例题都编制了 MATLAB 例题代码程序，并给出了源代码，对于想学习 MATLAB 语言编程的读者，可以通过本书学习、模仿、改写程序的源代码，提高自己的编程能力。对没有安装 MATLAB 的计算机，我们还提供了.exe 可执行程序，读者可按照使用说明在任何 Windows 操作系统中进行计算，操作方法简单、快捷，信息提示详尽。因此，本书不仅为教师教学提供了方便，为需要数据处理的读者（即便是对 MATLAB 知之甚少，或者对统计方法掌握得不够全面的读者）提供了可直接使用的平台，对从事程序开发的人员也具有重要的参考价值。本书配有一张程序光盘，并有独立发行的多媒体课件。程序盘中包含书中所有 MATLAB 例题源代码程序、可执行文件及其使用说明；多媒体课件涵盖理论教学课件、统计试验课件和.exe 案例分析，供读者选用。

本书不仅可作为本科生和工科研究生数理统计课程的基础教材、本科生相关专业的专业基础教材或选修教材及实验教材，也可作为科研人员、工程技术人员的工具书或参考读物。

本书是 2006 年版《数理统计与 MATLAB 工程数据分析》的全新修改。一是对教材内容进行了适度的增删与调整、重构编排，使段落层次更加清楚分明。如，增加了第 2 章“描述性统计分析”；删除了 2006 版第 4 章“非参数假设检验”；新编了第 4 章“假设检验”，详略有致地阐述了双边检验与单边检验、χ^2 拟合优度检验和独立性检验，增补了 p 值检验、误差统

第1章　概率论基础

数理统计是研究随机现象规律性的一门学科。它是以概率论为基础，研究如何以有效的方式获得、整理和分析受到随机性影响的数据，并以这些数据为依据，建立有效的数学模型，去揭示所研究问题的统计规律性。

数理统计的理论和方法已广泛应用于自然科学、技术科学、社会科学和人文科学等各个领域。随着计算机的发展和普及，数理统计已成为处理信息、进行决策的重要理论和方法。

数理统计研究的内容概括起来可分为两大类：其一是研究如何对随机现象进行观察、试验，以便更合理、更有效地获取观察资料的方法，即试验的设计和研究；其二是研究如何对所获得的有限数据进行整理、加工，并对所讨论的问题做出尽可能可靠、精确的判断，这就是统计推断问题。

1.1　概率的基本概念和性质

概率论是数理统计的基础，为此，我们先简要复习概率论的基本概念、性质与公式。

1.1.1　频率与概率

1. 随机事件与概率

自然界和人类社会中所发生的现象是多种多样的，但大致可分为两类：一类是确定性现象，即在一定条件下必然发生的现象，比如在标准大气压下“水加热至100℃时沸腾”等；另一类是随机现象，即在相同条件下重复进行某种试验，有多种可能的结果发生，而在试验或观察之前不能预知确切的结果。例如，投掷一枚均匀硬币，结果可能出现“正面”，也可能出现“反面”，掷前无法确定哪个结果会出现；远射一个目标可能击中也可能击不中，射前无法确定哪个结果会出现；从一袋小麦种子中任取10粒做发芽试验，试验的结果可能是有10，9，…，1发芽或全部不发芽，而试验前无法知道有几粒小麦种子会发芽，等等，这些都是随机现象。随机现象有两个特点：①在一次观察中，现象可能发生也可能不发生，即结果呈现不确定性；②在重复观察中，其结果具有统计规律性，例如，多次重复投掷硬币，出现“正”、“反”面的次数大致相同。概率论就是研究随机现象统计规律性的学科。

我们把具有以下几个特点的试验叫随机试验：①在相同的条件下可以重复进行；②试验的结果不止一个，所有结果事先明确知道；③进行一次试验前不能确定哪个结果会出现。随机试验通常以字母 E 表示。

随机试验中，可能出现也可能不出现的事情叫随机事件，用 A，B，C，…表示；每一个可能出现的结果，称为基本事件。例如随机试验“掷一只骰子，观察出现的点数”中“点数大于

(2) 各个体之间相互独立。

这种抽取样本的方法称为简单随机抽样,由此得到的样本称为简单随机样本,简称样本。

对一次具体的抽取,得到 n 个数值 $x_1,x_2,\cdots,x_n$,通常称之为样本观测值,简称样本值。

由简单随机样本的定义可知,来自总体 X 的一个样本 $X_1,X_2,\cdots,X_n$ 就是一组相互独立并且与总体同分布的随机变量。因此,若总体 X 的分布函数为 $F(x)$,密度函数为 $f(x)$,则样本$(X_1,X_2,\cdots,X_n)$的联合分布函数及联合密度函数分别为

$$F(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}F(x_i);\quad f(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}f(x_i)$$

例 2.1.4　设总体 $X\sim f(x)=\begin{cases}\theta x^{\theta-1}, & 0<x<1\\ 0, & \text{其他}\end{cases}$,$(X_1,X_2,\cdots,X_n)$为 X 的样本,则样本的分布为

$$f(x_1,x_2,\cdots x_n,\theta)=\prod_{i=1}^{n}\theta x_i^{\theta-1}=\theta^n\left(\prod_{i=1}^{n}x_i\right)^{\theta-1},\quad 0<x_i<1,1\leqslant i\leqslant n$$

当总体为离散型随机变量时,总体 X 的分布律为 $P(X=x_k)=p_k$,则样本$(X_1,X_2,\cdots,X_n)$的联合分布为

$$p(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}p(x_i)$$

例 2.1.5　设总体 $X\sim(0\text{-}1)$二点分布,$(X_1,X_2,\cdots,X_n)$为 X 的样本,求样本的分布。

解　因为 $P(X=1)=p,P(X=0)=1-p,0<p<1$,连续化可表示为

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

则样本的联合分布为

$$P(x_1,x_2,\cdots,x_n;p)=\prod_{i=1}^{n}p(x_i)=\prod_{i-1}^{n}p^{x_i}(1-p)^{1-x_i},\quad x_i=0,1$$

例 2.1.6　设总体 X 服从泊松分布,$X\sim\pi(\lambda)$,λ 未知,$(X_1,X_2,\cdots,X_n)$为 X 的样本,求样本的分布。

解　因为泊松分布的分布律为

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

连续化为

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

所以样本的联合分布

$$p(x_1,x_2,\cdots,x_n;\lambda)=\prod_{i=1}^{n}\frac{\lambda^{x_i}e^{-\lambda}}{x_i!}=\frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}(x_i)!}$$

2.1.3　统计量

我们知道样本是总体的代表和反映,是对总体进行统计分析和推断的依据,但在样本抽取后,样本所含的信息不能直接用于解决我们所要研究的问题,尚需进行“加工”、“提炼”,而

为样本方差；称

$$S=\sqrt{S^2}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2} \tag{2.2.3}$$

为样本标准差(均方差)。

对应的观察值为

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

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

3. 二阶中心距

称

$$S_n^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 \tag{2.2.4}$$

为二阶中心距。

方差、均方差、二阶中心距都是用来刻画数据的变异的度量值，是尺度参数。通常理论上采用样本方差 S^2 或样本标准差 S 来描述数据的变异度，因为 $E(S^2)=\sigma^2$，即样本方差 S^2 是总体方差 σ^2 的无偏估计量，而二阶中心距 S_n^2 与总体方差 σ^2 是有偏离的：$E(S_n^2)=\frac{n-1}{n}\sigma^2$。

4. 变异系数

样本方差的量纲与数据的量纲不一致，它是数据量纲的平方，而标准差的量纲与数据量纲一致。比较两个样本的变异度，由于单位不同或均数不同，不能单纯用标准差比较，而是用一个相对的百分数变异度来比较，这就是变异系数：

$$CV=100\times\frac{S}{\bar{x}}(\%) \tag{2.2.5}$$

用它可以对同一样本中的不同指标或不同样本中的同一指标进行比较，据 CV 的大小可以对指标的变异程度排序。

5. 样本矩

称

$$v_k=\frac{1}{n}\sum_{i=1}^{n}x_i^k,\quad k=1,2,3,\cdots$$

为样本 k 阶原点矩；称

$$u_k=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^k,\quad k=1,2,3,\cdots$$

为样本 k 阶中心矩。

设观测数据是由总体 X 中取出的样本,总体分布函数是 $F(x)$,当 X 为离散分布时,总体分布可由概率分布列刻画:

$$p_i = P(X = x_i), \quad i = 1,2,\cdots$$

总体分布为连续时,总体分布可由概率密度 $f(x)$ 刻画,连续分布中最重要的是正态分布,它的概率密度 $\varphi(x)$ 及分布函数 $\Phi(x)$ 分别为

$$\varphi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < +\infty; \quad \Phi(x) = \int_{-\infty}^{x} \varphi(t)\mathrm{d}t$$

具有正态分布的总体称为正态总体。

上述数据的数字特征即为样本的数字特征,与样本数字特征对应的是总体的数字特征,它们分别是:

总体均值	$\mu = E(X)$
总体方差	$\sigma^2 = \mathrm{var}(X)$
总体均方差	$\sigma = \sqrt{\mathrm{var}(X)}$
总体变异系数	$\gamma = \dfrac{\sigma}{\mu}$
总体偏度	$G_1 = \dfrac{\mu_3}{\sigma^3}$
总体峰度	$G_2 = \dfrac{\mu_4}{\sigma^4} - 3$

这里 $\mu_k = E(X-\mu)^k$ 为总体 k 阶中心矩。

根据统计学的结果,样本的数字特征是相应的总体数字特征的矩估计。当总体数字特征存在时,相应的样本数字特征是总体数字特征的相合估计,从而当 n 较大时,有

$$\mu \approx \overline{X}, \quad \sigma^2 \approx S^2, \quad \sigma \approx S, \quad \gamma \approx \mathrm{CV}, \quad G_1 \approx g_1, \quad G_2 \approx g_2$$

这里,特别要强调下列情况:当观测数据 $x_1, x_2, \cdots, x_n$ 是所要研究对象的全体时,数据的分布即总体分布。我们认为取得每一个观测数据 x_i 是等可能性的,即为 $\dfrac{1}{n}$,总体分布为离散均匀分布

$$P(X = x_i) = \frac{1}{n}, \quad i = 1,2,\cdots,n$$

对这种情况,数据数字特征即总体数字特征。许多实际数据属于这种情况,它更能体现数据分析的特点——让数据本身说话。实际上,我们也可以把这种情况看作取自定型模型的数据,而上述数字特征仍有相应的统计意义。

例 2.2.1　从 19 个杆塔上的普通盘形绝缘子测得该层电导率(μS)的数据如表 2.2.1 所示。

表 2.2.1　电导率数据表

8.98	8.00	6.40	6.17	5.39	7.27	9.08	10.40	11.20	8.57
6.45	11.90	10.30	9.58	9.24	7.75	6.20	8.95	8.33	

解　由式(2.2.1)～式(2.2.7)计算得表 2.2.2。

表 2.2.2　电导率特征值

均值	方差	标准差	极差	变异系数	偏度	峰度
8.4295	3.3029	1.8174	6.51	21.5598	0.11524	−0.69683

从运算结果看出，集中取值$\bar{x}=8.4295$，分散度 $s^2=3.3029$，$g_1=0.1152$ 向右微偏，g_1，g_2的绝对值较小，可以认为是来自正态总体的数据。

例 2.2.2　某电瓷厂的某种悬式绝缘子机电破坏负荷试验数据(单位：t)分组表示如表 2.2.3 所示。计算这批分组数据的均值、方差、变异系数、偏度、峰度。

表 2.2.3　绝缘子机电破坏负荷试验数据

组　段	组　中　值	组　频　数
5.5～6.0	5.75	4
6.0～6.5	6.25	3
6.5～7.0	6.75	15
7.0～7.5	7.25	42
7.5～8.0	7.75	49
8.0～8.5	8.25	78
8.5～9.0	8.75	50
9.0～9.5	9.25	31
9.5～10.0	9.75	5

解　对于分组数据，我们是将组中值当成各组段中实际数据的代表(它不一定是实际数据)，因此算得的各数字特征是原始数据的数字特征的近似，这里 $n=277$。由式(2.2.1)～式(2.2.7)计算得表 2.2.4。

表 2.2.4　绝缘子机电破坏负荷特征值

均值	方差	标准差	极差	变异系数	偏度	峰度
8.1002	0.62874	0.79293	4	9.7891	−0.38211	0.057169

从计算结果知，绝缘子机电破坏负荷集中取值$\bar{x}=8.1002$，分散度 $s^2=0.62874$，最大幅度 $R=4$，$g_1=-0.38211$ 向左微偏，g_1，g_2的绝对值较小，可以认为是来自正态总体的数据。

2.2.3　样本的其他特征值描述

上述数据的均值、方差、均方差等数字特征是总体相应特征值的一种矩估计，它更适合于来自正态分布的数据的分析。若总体的分布未知，或者数据严重偏态，有若干异常数据(极端值)，上述分析数据的方法不甚合适，而应计算中位数、分位数、三均值、极差等数据数字特征，计算上述特征值需要用到次序统计量。

设 $x_1,x_2,\cdots,x_n$是 n 个观测值，它可以理解为来自某总体的样本，将它们按数值由小到大记为

表 2.2.5　1952—1997 年我国人均国内总产值

年　份	人均生产总值	年　份	人均生产总值
1952	119	1975	327
1953	142	1976	316
1954	144	1977	339
1955	150	1978	379
1956	165	1979	417
1957	168	1980	460
1958	200	1981	489
1959	216	1982	525
1960	218	1983	580
1961	185	1984	692
1962	173	1985	853
1963	181	1986	956
1964	208	1987	1104
1965	240	1988	1355
1966	254	1989	1512
1967	235	1990	1634
1968	222	1991	1879
1969	243	1992	2287
1970	275	1993	2939
1971	288	1994	3923
1972	292	1995	4854
1973	309	1996	5576
1974	310	1997	6079

解　从表 2.2.5 的数据看，2287，2939，3923，4854，5575 与 6079 是异常值(特大值)。由改革开放的形势具体分析，这些特大值的出现是好事。由此可见，实际问题中必须结合问题背景对数据进行具体分析。异常值又称离群值，它们远离了 1952—1991 年人均国内生产总值数据的主要群体。

由式(2.2.8)～式(2.2.17)计算得表 2.2.6 和表 2.2.7。

表 2.2.6　1952—1997 年我国人均国内总产值其他特征值

中位数	下四分位点	上四分位点	四分位极差	三均值	下截断点	上截断点
313	216	956	740	449	−894	2066

表 2.2.7　1952—1997 年我国人均国内总产值特征值

均值	方差	标准差	极差	变异系数	偏度	峰度
965.48	2099532	1448.9764	5960	150.08	2.4149	5.248

从运算结果分析，由于偏度为 2.4149，数据分布的图形显著右偏；峰度为 5.248，数据分布的右端有许多极端值。又数据的标准差为 1448.9764，其数据甚至超过了均值 965.48，

解　资料为计数性的,每穗小穗数在 15～20 的范围内变动。把资料按小穗数加以归类,共分为 6 组,组与组相差为 1 小穗,称为组距 $\Delta x=1$。将资料归组整理就得到表 2.3.3 所示的频率分布表。

表 2.3.3　100 个麦穗每穗小穗数的频率分布表

每穗小穗数 x	15	16	17	18	19	20	总计
频数 m_i	6	15	32	25	17	5	100
频率 $f_i=\frac{m_i}{n}$	0.06	0.15	0.32	0.25	0.17	0.05	1

2. 试验指标为连续型

数据取值为一有限区间$[a,b)$,通常将$[a,b)$分成 $l(l<n)$个区间(一般是等间隔的),每个区间的长度$\frac{b-a}{l}$称为组距,则

$$a=a_0<a_1<a_2<\cdots<a_{l-1}<a_l=b$$

通常组数可以考虑取

$$l\approx 1.87(n-1)^{\frac{2}{5}}$$

表 2.3.4 给出了一些 l 值以供参考。

表 2.3.4　数据分组数的参考值

n	40～60	100	150	200	400	600	800	1000	1500	2000	5000	10000
l	6～8	7～9	10～15	16	20	24	27	30	35	39	56	74

组距 Δx=(样本最大观测值－样本最小观测值)/组数,各组区间端点为

$$a_0,a_0+\Delta x=a_1,a_0+2\Delta x=a_2,\cdots,a_0+l\Delta x=a_l$$

区间为

$$[a_0,a_1),[a_1,a_2),\cdots,[a_{l-1},a_l)$$

其中 a_0可略小于最小观测值,a_l可略大于最大观测值。

通常可用每组的组中值来代表该组的变量取值,组中值=(组上限+组下限)/2。

统计样本数据落入每个区间的个数——频数,并列出其频数频率分布表。

例 2.3.2　某炼钢厂生产 25MnSi 钢,由于各种随机因素的影响,各炉钢的含硅量 X 是有差异的。现在希望推断 X 的概率密度 $p(x)$。记录了 120 炉正常生产的 25MnSi 钢的含硅量(单位:%)如表 2.3.5 所示。

解　样本观测值的最大值和最小值分别为 $x_{(n)}=0.95$,$x_{(1)}=0.64$,$n=120$。取 $a=0.635$(略小于 $x_{(1)}$),$b=0.955$(略大于 $x_{(n)}$);取 $l=16$,得组距 $\Delta x=\frac{b-a}{l}=0.02$,频数频率分布表见表 2.3.6。

```
disp(['四分位极差: ',num2str(RS)]);
sss = 0.25 * ss25 + 0.5 * ss50 + 0.25 * ss75;
disp(['三均值: ',num2str(sss)]);
xjie = ss25 - 1.5 * RS;
disp(['下截断点: ',num2str(xjie)]);
sjie = ss75 + 1.5 * RS;
disp(['上截断点: ',num2str(sjie)]);
```

例 2.4.3 解例 2.2.4。

解 在 MATLAB 命令窗口中输入:

```
>> x = [119  142  144  150  165  168  200  216  218  185  173  181  208  240  254  235
222  243  275  288  292  309  310  327  316  339  379  417  460  489  525  580  692
853  956  1104  1355  1512  1634  1879  2287  2939  3923  4854  5576  6079];
>> fws(x)
```

运行后显示:

```
中位数: 313
下四分位数: 216
上四分位数: 956
四分位极差: 740
三均值: 449.5
下截断点: -894
上截断点: 2066
```

在 MATLAB 命令窗口中继续输入:

```
>> dts(x)
```

运行后显示:

```
均值: 965.4783
方差: 2099532.4773
标准差: 1448.9764
极差: 5960
变异系数: 150.0786
偏度: 2.4149
峰度: 5.248
```

2.4.2 用样本的分布描述总体的 MATLAB 编程实现

1. 直方图 MATLAB 程序代码

```
function sfpin(y)
y = y(:);
N = length(y);
L = floor(1.87 * (N - 1)^0.4);
[Y,X] = hist(y,L);
X = X(:)';
Y = Y(:)';
```

```
ind = find(Y == 0);
X(ind) = [];
Y(ind) = [];
xt1 = 1.5 * X(1) - X(2) * 0.5;
xtt = X(1:end - 1) * 0.5 + X(2:end) * 0.5;
xt2 = 1.5 * X(end) - X(end - 1) * 0.5;
X = [xt1,xtt,xt2];
n = sum(Y);
Y = Y/n;
xx = [X;X];yy = [Y;Y];
Xt = xx(:);Yt = [0;yy(:);0];
fill(Xt,Yt,'c')
hold on
x1 = (X(2:end) + X(1:end - 1))/2;
XX = [Xt';Xt'];
YY = [Yt';zeros(1,length(Yt))];
plot(x1,Y,' - k',Xt,Yt,' - k',XX,YY,' - k')
hold off
title('频率直方图')
```

2. 经验分布函数图形的 MATLAB 程序代码

```
function scdfplot(X)
X = X(:)';
X = sort(X);
n = length(X);
xsui = ones(size(X));
B = cumsum(xsui);
B = B/n;
xl = min(X) - (max(X) - min(X)) * 0.1;
xr = max(X) + (max(X) - min(X)) * 0.1;
x = [xl,X,xr];
y = [0,B,1];
h = stairs(x,y);
set(h,'linewidth',2,'color','k')
xlabel('x')
ylabel('F(x)')
grid on
axis([xl,xr, - 0.05,1.05])
title('经验分布函数')
```

3. QQ 图 MATLAB 程序代码

```
function qqs(y)
y = y(:)';
y = sort(y);
NNS = length(y);
x = norminv(((1:NNS) - 0.375)./(NNS + 0.25),0,1);
sigma = std(y);mu = mean(y);
xx = [min(x),max(x)];
yy = mu + sigma * xx;
```

运行后显示 QQ 图，如图 2.4.6 所示。

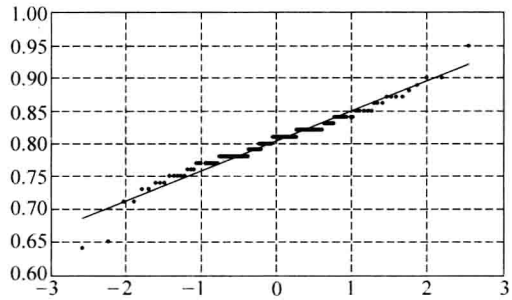

图 2.4.6　QQ 图

例 2.4.6　解例 2.2.4。

解　在命令窗口中输入：

```
>> x = [119,142,144,150,165,168,200,216,218,185,173,181,208,240,254,235,222,243,275,288,
292,309,310,327,316,339,379,417,460,489,525,580,692,853,956,1104,1355,1512,1634,1879,
2287,2939,3923,4854,5576,6079];
>> sfpin(x)
```

运行后显示频率直方图，如图 2.4.7 所示。

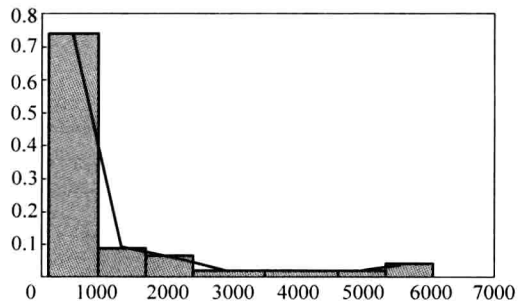

图 2.4.7　频率直方图

输入：

```
>> scdfplot(x)
```

运行后显示经验分布函数图，如图 2.4.8 所示。

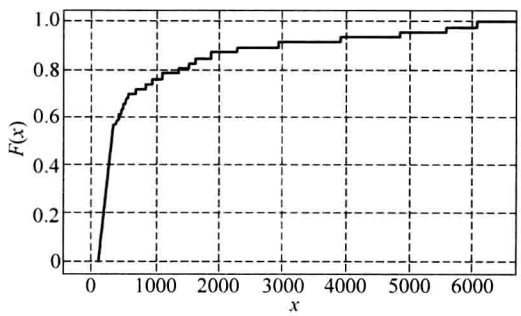

图 2.4.8　经验分布函数图

继续输入:

```
>> qqs(x)
```

运行后显示 QQ 图,如图 2.4.9 所示。

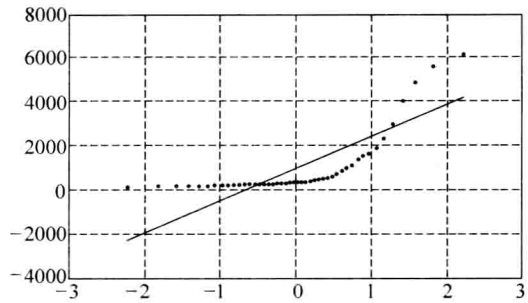

图 2.4.9 QQ 图

这与前面分析的结果一致。

2.4.3 MATLAB 代码综合分析实例

例 2.4.7 某公司对应聘人员进行能力测试,测试成绩总分为 150 分,下面是 50 位应聘人员的测试成绩(已经过排序),见表 2.4.1。

表 2.4.1 应聘人员测试成绩

64	67	70	72	74	76	76	79	80	81
82	82	83	85	86	88	91	91	92	93
93	93	95	95	95	97	97	99	100	100
102	104	106	106	107	108	108	112	112	114
116	118	119	119	122	123	125	126	128	133

解 在 MATLAB 命令窗口中输入:

```
>> x = [64,67,70,72,74,76,76,79,80,81,82,82,83,85,86,88,91,91,92,93,93,93,95,95,95,97,99,
100,100,102,104,106,106,107,108,108,112,112,114,116,118,119,119,122,123,126,128,133];
>> qqs(x)
```

运行后显示 QQ 图,如图 2.4.10 所示。

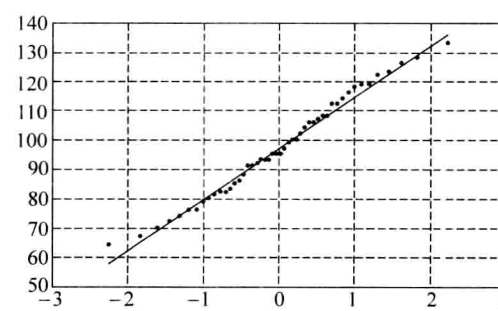

图 2.4.10 QQ 图

```
 64   67   70   72   74   76   76   79   80   81
 82   82   83   85   86   88   91   91   92   93
 93   93   95   95   95   97   97   99  100  100
102  104  106  106  107  108  108  112  112  114
116  118  119  119  122  123  125  126  128  133
```

把文件存为文件名："数据描述性分析.txt"。

(2) 启动应用程序

"描述性统计分析"应用程序启动后，生成两个窗口，后面的窗口形式如图 2.5.1 所示。

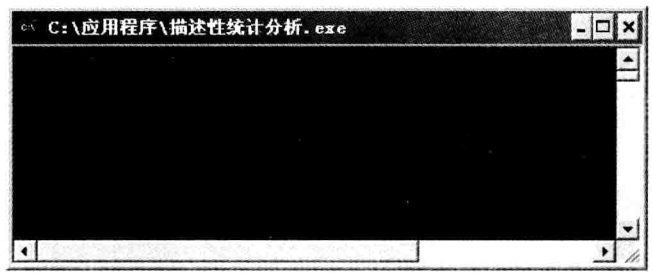

图 2.5.1　后面的窗口

前面的窗口形式如图 2.5.2 所示。

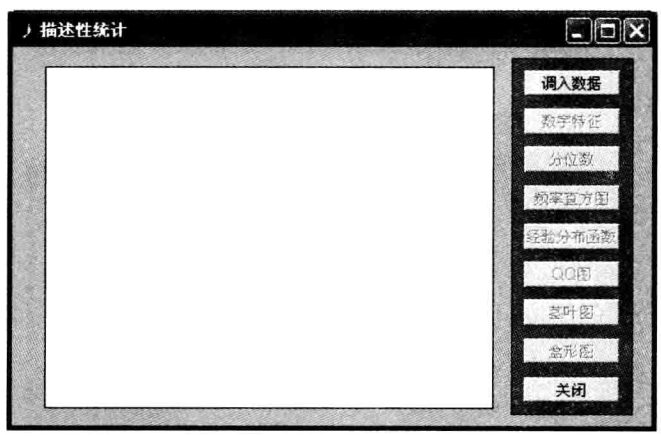

图 2.5.2　前面的窗口

此时各个选项按钮都不可用。

(3) 调入数据

单击"调入数据"按钮，打开"调入数据"对话框(图 2.5.3)，再查找到"数据描述性分析.txt"文件。

单击"打开"按钮。这时，各选项按钮变为可用，如图 2.5.4 所示。

(4) 进行数据分析

单击"数字特征"按钮，生成图 2.5.5。

单击"分位数"按钮，生成图 2.5.6。

图 2.5.3 “调入数据”对话框

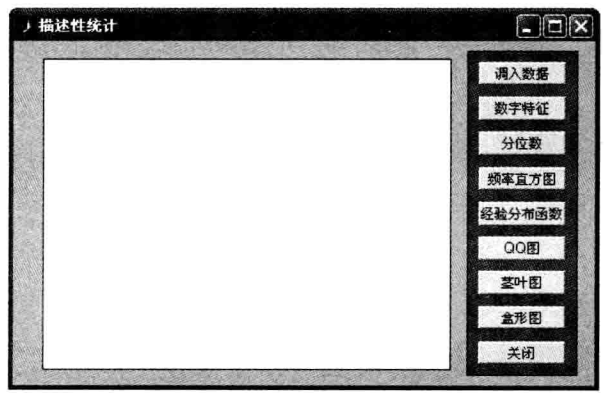

图 2.5.4 调入数据后的窗口

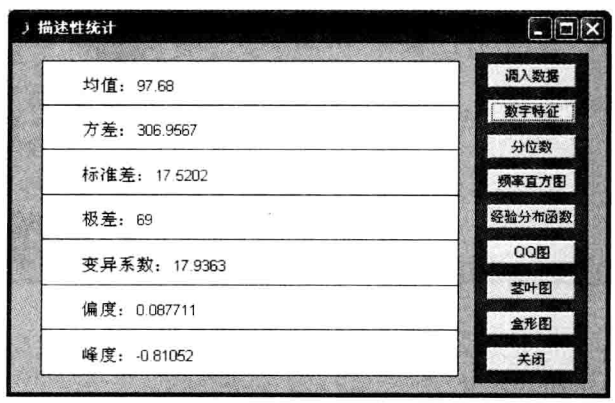

图 2.5.5 数字特征窗口

单击“频率直方图”按钮,生成图 2.5.7。

单击“经验分布函数”按钮,生成图 2.5.8。

单击“QQ 图”按钮,生成图 2.5.9。

单击“茎叶图”按钮,生成图 2.5.10。

单击“盒形图”按钮,生成图 2.5.11。

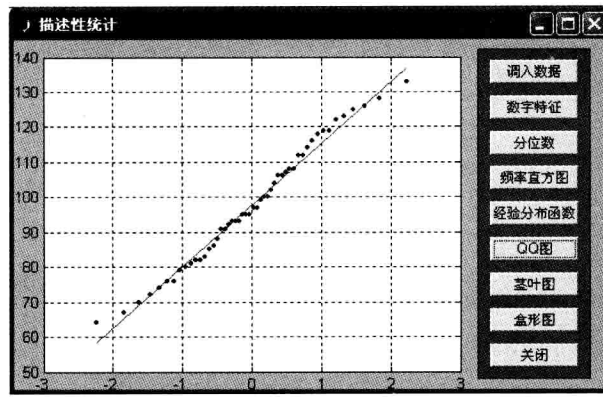

图 2.5.9 QQ 图窗口

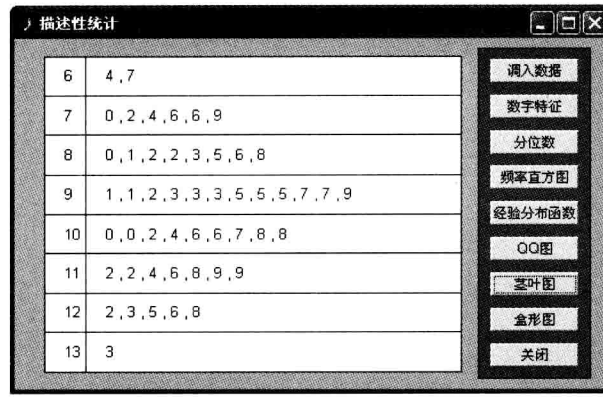

图 2.5.10 茎叶图窗口

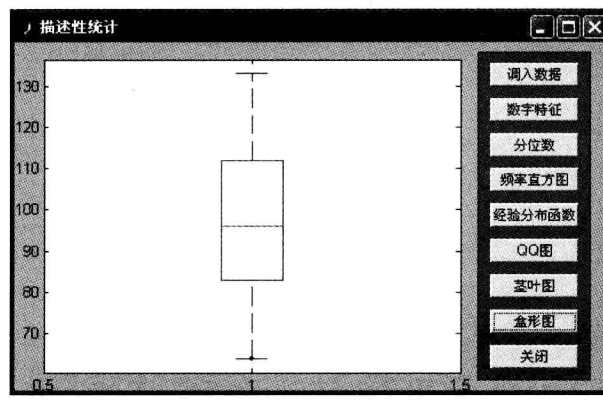

图 2.5.11 盒形图窗口

2.6　抽样分布

有很多统计推断是基于正态分布的假设的，以标准正态变量为基石而构造的统计量在实际中有广泛的应用，利用统计量时，需要知道它的分布。下面我们给出几种常用正态总体样本的均值和方差的分布，限于篇幅，不作理论上的推导。

2.6.1　U 分布(样本均值分布)

设 $X_1, X_2, \cdots, X_n$ 是来自总体 $N(\mu, \sigma^2)$ 的样本，则

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$U = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \tag{2.6.1}$$

标准正态分布的密度函数图形如图 2.6.1 所示。

下面介绍标准正态分布的分位点。

对于给定的 $\alpha(0<\alpha<1)$，称满足条件

$$P(U \geqslant u_\alpha) = \alpha$$

的点 u_α 为标准正态分布的上 α 分位点(也叫上侧分位点)，如图 2.6.2 所示。

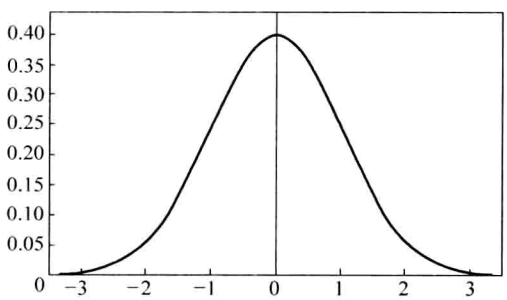

图 2.6.1　标准正态分布的密度函数

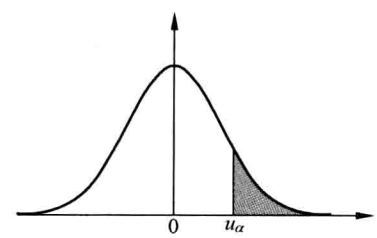

图 2.6.2　标准正态分布上侧分位点

满足条件

$$P(U \leqslant -u_\alpha) = \alpha$$

的点 $-u_\alpha$ 为标准正态分布的下 α 分位点(也叫下侧分位点)，如图 2.6.3 所示。

满足条件

$$P(|U| \geqslant u_{\frac{\alpha}{2}}) = \alpha$$

的点 $-u_{\frac{\alpha}{2}}$ 和 $u_{\frac{\alpha}{2}}$ 为标准正态分布的双侧 α 分位点(也叫双侧分位点)，如图 2.6.4 所示。

对于标准正态分布的上 α 分位点 u_α，由定义 $P(U \geqslant u_\alpha) = \alpha$ 知，对给定的 α，反查标准正态分布表 $\Phi(u_\alpha) = 1-\alpha$ 可得。见附表 1。

常用的 u_α 值如表 2.6.1 所示。