电子计算机教学丛书

数理统计程序设计与使用

曾秋成　著

SHU LI TONG JI
CHENG XU SHE JI
YU SHI YONG

安徽教育出版社

电 子 计 算 机 教 学 丛 书

数理统计程序设计与使用

曾秋成 著

安 徽 教 育 出 版 社

数理统计程序设计与使用

安徽教育出版社出版

（合肥市跃进路1号）

安徽省新华书店发行　安徽新华印刷厂印刷

*

开本：787×1092　1/16　印张：18.75　字数：480,000

1988年4月第1版　1988年4月第1次印刷

印数：5,500

ISBN7-5336-0149-1/G·730

平装定价：4.00元

精装定价：7.00元

内容简介

本书介绍常用数理统计方法的程序设计及其使用，它的内容除经典方法外，还有国内外行之有效的最新方法，基本上适应于科研、生产、教学等各方面的实际需要。全书共分9章33节，提供程序31个，可完成分布拟合检验、方差分析、正交试验分析、一元及多元回归、属性统计检验以及质量管理与极值估计等100种以上的统计分析与运算。所提供的程序均可独立使用，并具有多种功能，其中有20个还可绘出分布图与相关图等。

为配合本书使用，本社还出版一套适用于PC—1500型袖珍机的软件。两者配合，无论对院校教学还是对实际工作来说，都具有实用价值。

前　言

数理统计方法愈来愈广泛地应用于科学技术、工农业生产、国民经济、政法、财贸、国防建设等各个领域，成为这些领域的一种重要的分析手段和数学工具。数理统计方法也愈来愈引起高等院校、中等专业学校的普遍重视，目前国内愈来愈多的学校为非数学专业的学生开设数理统计的必修课或选修课。如何结合实际工作和教学的需要，有针对性地选择行之有效的数理统计方法，并把它们编成计算机语言，使之易于在计算机上进行分析和运算，是广大读者感兴趣的问题。作者根据多年的实际工作经验，从实际需要出发，参考国内外有关文献，撰写了这本《数理统计程序设计与使用》，供读者参考，其目的在于推广数理统计方法及计算机应用技术。

本书是一本以实用为主的参考书，因此，书中只着重介绍数理统计方法和这些方法如何在电子计算机上实现。尽量避免对各统计方法的理论探讨。本书除介绍应用价值较大的经典统计方法外，还注意对近几十年来国内外出现的一些行之有效的新方法的引进，力求能较全面地反映这些新的成果。

本书在内容处理方面，对一些计算量小而单独编写程序意义并不大的经典统计方法，能综合的就综合，能代替的就代替，对那些计算步骤较多而处理起来很烦杂的经典统计方法，尽量编成程序由计算机来完成，以减少使用时的失误；对一些统计理论比较成熟的新方法新成果，着重让实际工作者能准确理解和应用。总之，本书本着便于实用的原则进行编排。

本书共9章33节，提供可独立使用的程序31个，书中提供的程序所能完成的统计运算都具有多功能，功能少则几种，多则十几种；全部程序可进行100多种统计运算。由于程序的多功能性，不可避免地在若干程序的某些功能上有部分交叉或重复，在处理这类问题时，本书尽可能使这种交叉具有各自的特点，使之能在实用中有不同的用途，不同的效果。

本书在数据的输入、程序的操作、计算结果的表达等方面，力求简便、适用和功能完整。在所提供的31个程序中，有20个具有绘图功能，它能将计算结果和数据关系用相关图、分布图等图形表现出来，这可说是本书的一大特色。

本书选择SHARP公司生产的PC—1500型袖珍计算机作为基本机型，用它及其基本附件(不包括配用CE—158接口后可扩充的部件)作为本书提供程序的支持硬件，这自然是考虑到这种袖珍机为目前国内较普及而拥有量最多的个人微机的缘故。虽然，它的内存有限，显示窗十分简单，CE—150打印机的幅面狭窄，影响程序中对话语句的完整性，不便使用汉字等，在应用上有一定的局限性，但它操作简单，便于携带及在工作现场和家庭使用，对非专业的微机使用者来说更有其实用价值，便于推广。

本书的全部程序都采用BASIC语言编写，它的计算部份基本上可直接置入其它用磁盘操作的微机(如IBM—PC、长城0520、Apple—Ⅱ等)中。绘图部份的程序，只要确定了相应的绘图软件后，便可简便地进行移植。

为有利于推广本书介绍的统计方法及所提供的程序，已由安徽教育出版社审定出版适用于PC—1500机的程序软件，装成两盒普通盒式磁带，委托马鞍山钢铁公司钢铁研究所监制并经销发行。此外，该所还监制发行了适用于IBM—PC机和紫金—Ⅱ机的汉字操作程序软

盘(计算部份)，供用户使用。

数理统计内容丰富，方法甚多，发展迅速，限于篇幅和作者水平，在方法选取上难免挂一漏万，在内容处理及程序构思方面也难免有不妥之处，敬请读者批评指正。

在本书撰写及软件制作过程中，自始至终得到马鞍山钢铁公司钢铁研究所的领导和有关同志大力支持。书稿完成后，又蒙合肥工业大学卢树铭副教授帮助审阅，并提出不少宝贵的修改意见，在此一并致谢。

曾秋成

1986年12月

目　　录

第一章　分布拟合的柯尔莫哥洛夫检验

§1.综述

本章所包括的四个程序，都是对连续变量型统计数据的分布作理论分布曲线拟合检验的．这些理论分布曲线是：正态分布(normal distribution)及对数正态分布(log-normal distribution)、指数分布(exponential distribution)、威布尔分布(Weibull distribution)、Γ分布(gamma distribution)．对于实测得到的统计数据，无论是已分组处理的大样本还是未分组的大样本、小样本，这四个程序都能适用．对这三种数据条件，各程序都设计了三个数据入口：

1.已分组的大样本数据入口“A”；

2.未分组的大样本，需要自动分组计算的数据入口“S”；

3.小样本直接计算的数据入口“Z”．

第一种入口采用 INPUT 语句键入有关数据信息．第二、第三种入口采用READ语句读入，数据置于 DATA语句中．

对于第二种需要自动分组计算的 N个原始数据y_1，y_2，……，y_N，使用下列公式计算分组的组间距(J)

$$J=\frac{y_{\max}-y_{\min}}{1+3.322\lg N} \tag{1—1}$$

$y_{\max}$＝数据中的最大值；$y_{\min}$＝数据中的最小值．

当$J\geqslant1$时，取实用组距c为$J+0.5$的整数部份；

当$J<1$时，取实用组距c为J的两位整小数(计算中显示“Error c＝0”说明J值太小而超界)．

以第一组的组下限$Z(0)=y_{\min}-\frac{c}{2}$为分组的起点，可以计算出分组组数($M$)，各组的数据频数$F_1$，$F_2$，$\cdots F_M$等．

除正态分布定义域是$(-\infty,\infty)$外，其他三种分布的定义域均为$(0, \infty)$，且这三种分布曲线存在一个非负的起点；一般地说，这三种分布都是非对称的，定义域靠起点端的一半的密度函数大于定义域另一半的密度函数．分组时总是把最靠起点端的一组定义为第一组，依次排列下去．需要程序自动分组的数据，分组以后如果不能满足上面定义，则程序具有自动对组序的“反置”处理．反置后相应的组间距将是原组间距的相反数(即对原组间距取负号)．

对不分组计算的小样本数据(一般在$N=5\sim30$左右)，它的实测频率的计算是将数据先按数值自大到小重新排列以后再用下列公式算出：

威布尔分布用公式

$$p_N(x)=\frac{m-0.32}{N+0.36} \tag{1—2}$$

Γ 分布用公式

$$p_N(x)=\frac{m}{N+1} \tag{1—3}$$

其中m为按大小排列后每个数据的秩序数.

这样的选择，只是为了与人们对以上诸公式的应用习惯相一致. 正态分布和指数分布的参数计算则不需先确定实测频率.

特征数计算

对特征数计算，这些程序都是分两步进行的. 首先，计算其常见的统计特征值，其次，作拟合优度检验. 常用统计特征值计算包括平均数、标准偏差、方差(二阶中心矩)等；其中已分组的大样本(由“A”进入的)全部按分组后的简化计算方法进行. 对数据作简化的公式是

$$u_i=\frac{x_i-x(1)}{c} \tag{1—4}$$

式中 u_i=简化后的数据第i组的组中值；x_i=原始数据第i组的组中值；

$x(1)$=原始数据第一组的组中值；c=组距.

各项特征数的计算公式：

原始数据的平均数为

$$x_c=(\sum F_i u_i/M)c+x(1) \tag{1—5}$$

式中 F_i=第i组的组频数，M=分组数；

原始数据的方差(二阶中心距)为

$$M_2=c^2(\sum u_i^2/M-(\sum F_i u_i/M)^2) \tag{1—6}$$

原始数据的标准偏差为

$$S=\sqrt{M_2} \tag{1—7}$$

对未分组的大样本和小样本数据(由“S”或“Z”进入的)，则采用直接计算，公式为

平均数 $$x_c=\sum y_i/N \tag{1—8}$$

方差 $$M_2=\frac{1}{N-1}\sum(y_i-x_c)^2 \tag{1—9}$$

标准偏差 $$S=\sqrt{M_2} \tag{1—10}$$

关于拟合优度的检验，如果通过拟合检验接受$H_0:F_0(x)-F(x)=0$以后，将给出所适合的这种理论分布的有关特征参数，其各理论分布的特征参数计算公式列于各程序的计算方法说明中.

这四个程序的分布曲线的拟合优度检验，我们都采用柯尔莫哥洛夫(Колмогоров)检验. 这不仅因为柯尔莫哥洛夫检验的功效较高，也因为这种检验能给出理论分布的函数值. 有了理论分布函数值，就可以在计算中输出，用于对比；也可以通过绘制实测分布(实线)和理论分布(虚线)阶梯图，给出一种直观性好的比较拟合情况的图形来.

拟合优度检验所取的显著性水平，通常为$\alpha=0.10$；$\alpha=0.05$；$\alpha=0.01$三种供选用；在无要求时，程序自动置$\alpha=0.05$；做其他显著性水平下的拟合优度检验可通过给定“K_0”值来达到. 表1—1给出不同显著性水平下的柯尔莫哥洛夫函数值：

表1—1　　各种显著性水平下的柯尔莫哥洛夫函数值($p(K_O)=1-K(K_O)=\alpha$)

显著性水平	$\alpha=0.0$	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.10$	$\alpha=1.0$
K_O　　值	大于2.51	1.62	1.36	1.22	小于0.3

柯尔莫哥洛夫检验是在给定的某个显著性水平下，检验统计假设

$$H_0:\quad F_0(x)-F(x)=0$$

能否成立和可信．式中

$F_0(x)$＝某个完全确定的理论分布函数；$F(x)$＝被实测的这个分布函数．

该检验所使用的统计量为

$$D_N=\sup_{-\infty<x<\infty}|F_N(x)-F_0(x)| \tag{1—11}$$

式中，$F_N(x)$＝实测分布函数中抽取的容量为N的样本的经验分布．

由于$F_N(x)$和$F_0(x)$都是x的单调非降函数，求$|F_N(x)-F_0(x)|$的上确界，就是将N个个体按x值由小到大次序排列后，以这个新的子样的次序统计量X_i所计算出的$F_0(X_i)$与其经验分布值之间的最大偏差δ_i，

$$\delta_i=\max\left\{F_0(X_i)-\frac{m-1}{N},\ \frac{m}{N}-F_0(X_i)\right\} \tag{1—12}$$

取δ_i值最大的一个作为统计量D_N，做柯尔莫哥洛夫检验．显然，D_N值愈小，$F_0(x)$与$F_N(x)$愈接近，若统计假设H_0成立，则根据柯尔莫哥洛夫定理知$\sqrt{N}D_N$的分布趋近于柯尔莫哥洛夫分布$K(K_O)$．于是，检验H_0能否接受，其做法是在给定的显著性水平（α）条件下，从柯尔莫哥洛夫分布表中查找$K_{O\alpha}$值，将其与$\sqrt{N}D_N$相比较，若$\sqrt{N}D_N<K_{O\alpha}$，则接受H_0；若$\sqrt{N}D_N\geqslant K_{O\alpha}$，则否定$H_0$．

为了提高小样本拟合检验的功效，并希望能减少检验中发生第二类错误的概率，Finklestein 和 Schafar 1971年提出以

$$S_n^*=\sum_{i=1}^{m}\delta_i \tag{1-13}$$

作为统计量的费—斯改进方法，并制出了不同显著性水平下的S_n^*临界值表．至于检验计算的统计量S_n^*在本程序中会同时打印出来，这可供参考．

频数直方图与累频数阶梯图的绘制

数据的频率(频数)直方图的绘制是从“D”入口键入，由程序自动完成．若直方图的纵坐标给定的频率(频数)最大组长为40mm(程序中坐标值等于200)；横坐标给定的分组数为M＝80mm(程序中坐标值为－400)，则每组直方图的幅宽等于$80/M$；图形面积为$(40\times80)\text{mm}^2$．若分布曲线拟合检验结果H_0被接受，则程序在同一坐标系中自动描绘出相应的密度函数曲线．这时密度曲线的横坐标与频率(频数)直方图完全一致；而密度函数最大值的纵坐标则定在频率(频数)最大组与次大组值之间．这样，所绘出的密度函数曲线尺寸是不会超界的．在本程序中一条密度函数曲线，是将其化成许多段折线描绘的．

累频率(累频数)阶梯图的绘制是由“X”口键入程序自动完成的．理论分布函数的阶梯图与实测累频率(累频数)阶梯图是采用完全相同的坐标单位绘制在同一个图中，这样就能更形象，更直观地看出理论分布与实测分布的拟合效果．

实测分布的累频率计算，采用下列两式：

$$F_N(x)=1-\frac{m-1}{N} \tag{1—14}$$

$$F_N(x)=1-\frac{m}{N} \tag{1—15}$$

前者用短虚线(兰色)描出，后者用长虚线(红色)描出，理论分布的阶梯一般都是由实线(黑色)绘制的. 用小样本数据做直接拟合检验，用本程序同样可以绘出累频率(累频数)阶梯图，但不能绘出频率(频数)直方图.

四个程序的绘图程序结构基本相同，其框图如图1—1和图1—2.

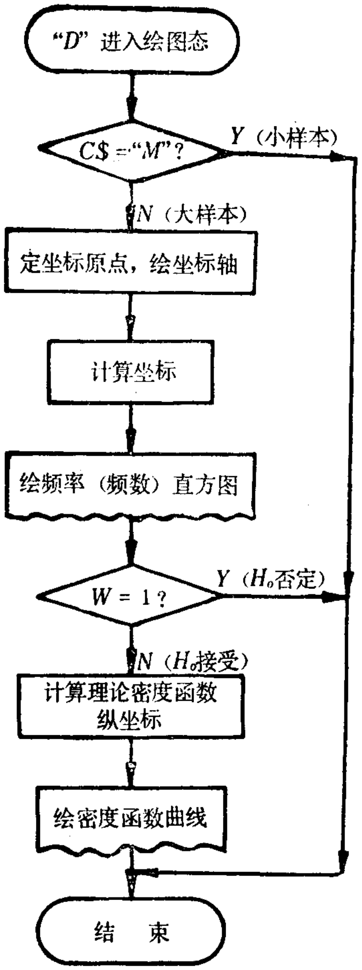


图1—1 频率(频数)直方图绘制过程

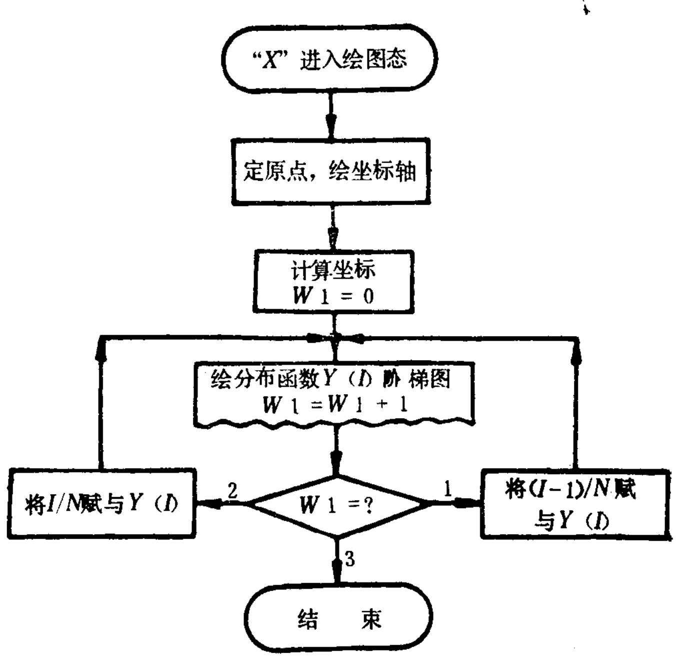


图1—2 分布函数阶梯图绘制过程

§2.正态(对数正态)分布

正态分布的密度函数为

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{1—16}$$

μ = 正态随机变量X的期望;

$\sigma^2=X$ 的方差.

正态分布函数

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\frac{-(x-\mu)^2}{2\sigma^2}}\,dx \tag{1—17}$$

的密度函数的一阶导数为

$$f'(x)=\frac{1}{\sigma\sqrt{2\pi}}\;\frac{-(x-\mu)}{\sigma^2}\,f(x) \tag{1—18}$$

令 $f'(x)=0$，有 $x-\mu=0$； $x=\mu$. 此时概率密度达到最大，最大值为

$$f(\mu)=\frac{1}{\sigma\sqrt{2\pi}}$$

若令 $$u=\frac{x-\mu}{\sigma}$$

则得到标准化随机变量u的密度函数。

$$f(u)=\frac{1}{\sqrt{2\pi}}\,e^{\frac{-u^2}{2}} \tag{1—19}$$

分布函数

$$F(u)=\frac{1}{\sqrt{2\pi}}\int e^{\frac{-u^2}{2}}\,du \tag{1—20}$$

以柯尔莫哥洛夫检验作正态分布拟合检验，计算它的分布函数 $F_0(x)$时，本程序采用了哈斯廷(Hasting)近似公式：

$$F_0(x)\cong(1-(1+a_1x+\cdots\cdots a_6t^6)^{-16})/2 \qquad (x\geqslant0) \tag{1—21}$$

并安排在子程序(1400～1500句)中供调用.

对变量X的样本做对数正态分布的拟合检验，只要设$\ln X=x$，对x做正态分布的拟合检验即可. 做对数正态分布拟合检验时，我们视实测数据x为原数据X的对数值，这样就有

$$\overline{X}=\exp\left(\overline{x}+\frac{1}{2}\sigma^2\right) \tag{1—22}$$

$$\sigma_X=\exp\left(\overline{x}+\frac{1}{2}\sigma^2\right)(\exp\sigma^2-1)^{\frac{1}{2}} \tag{1—23}$$

式中，$\overline{x}$＝实测数据x的平均值；σ^2＝实测数据x的方差；

$\overline{X}$＝原数据的平均值；　　$\sigma_X{}^2$＝原数据的方差.

正态(对数正态)分布拟合检验计算部份的程序框图如图1—3所示.

计算所需的输入信息是：

一、对已分组数据做正态(对数正态)分布拟合时，启动“A”后全部信息通过 INPUT语句输入. 这些信息是：

1. “$M=$” 已分组数据的分组数；　　(简单变量)
2. “$F=$” 各组数据的组频数；　　(数组变量)
3. “$c=$” 分组组距；　　(简单变量)
4. “$X(1)=$” 第一组的组中值；　　(简单变量)
5. “*Normal*？” 做正态分布拟合时答 “Y”,

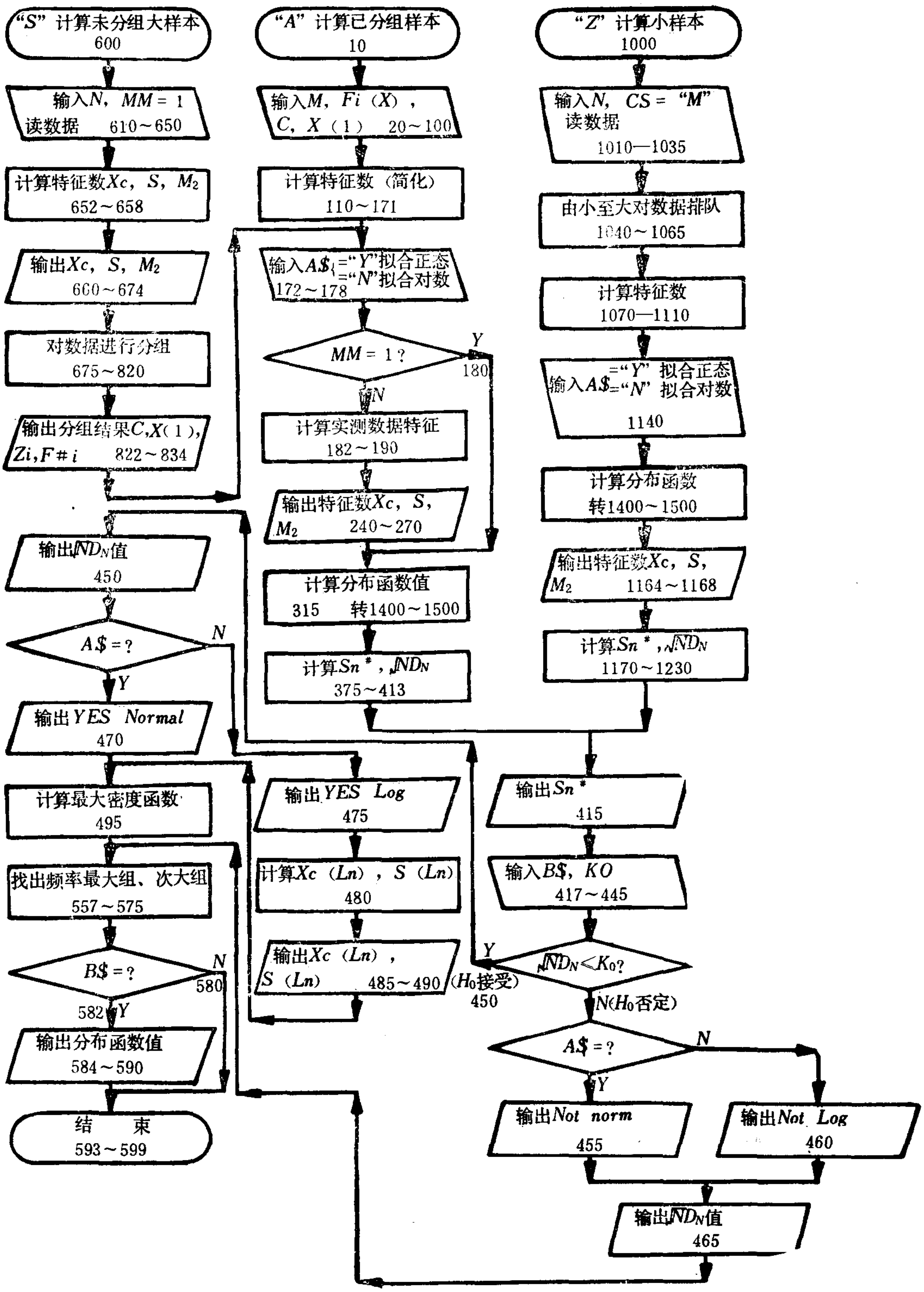


图1—3 正态(对数正态)分布曲线拟合程序框图

做对数正态分布拟合时答“N”；　　　　　　(逻辑变量)

6.　“*Print Formation?*” 要打印理论分布函数答“Y”，

不打印理论分布函数答“N”；　　　　(逻辑变量)

7.　“$K_O=$” 给定的柯尔莫哥洛夫变量值　　　　　　　　(简单变量)

二、对于需要程序分组的实测数据，通过置数据于程序后(1655句以后)的DATA语句中，启动“*S*”；另由INPUT语句输入以下信息：

“$N=$”实测样本的容量，　　　　　　　　　　　　　(简单变量)

以后输入的信息与5、6、7同．

三、对于直接拟合的小样本，用“Z”启动；其他信息输入过程与二所述的完全相同．

计算过程中所输出的信息是：

X_c＝实测数据的平均值

S＝实测数据的标准差

M_2＝实测数据的方差

S_n^*＝费—斯(Finklestein Schafar)统计量．

若H_0被接受，则输出如下的信息：

K_O＝计算得到的柯尔莫哥洛夫统计量$\sqrt{N}\,D_N$

“*yes Normal distribution*”　　(若5为“Y”)

“*yes Logarithmic distribution*”　　(若5为“N”)

$x_c(\ln)$＝原数据X的均值　　(若5为“N”)

$S(\ln)$＝原数据X的标准差　　(若5为“N”)

若H_0被否定，则输出如下的信息：

“*Not Normal distribution*”　　(若5为“Y”)

“*Not Logarithmic distribution*”　　(若5为“N”)

K_0＝计算得到的柯尔莫哥洛夫统计量$\sqrt{N}D_N$．

(对6答“Y”)还会输出：

$F(i)$＝第i组的理论分布函数值　　　　　($i=1, 2, \cdots M$)

需程序分组的数据，在完成分组以后会打印出分组结果：

C＝分组组距

$X(1)$＝第一组组中值

$Z(i)$＝第i组组下限　　　　　　($i=1, 2, \cdots M$)

$F_*(i)$＝第i组组频数　　　　　　($i=1, 2, \cdots M$)

$Z(i+1)$＝第$i+1$组组下限　　　　($i=1, 2, \cdots M$)

程序清单参见174～178页．

〔例1—1〕试对下列20个数据：.985，.955，.945，.96，.99，.985，.975，.975，.925，1.04，.98，.975，.98，.905，.99，.965，.955，.98，.895，.945，作正态分布拟合检验；取$\alpha=0.05$．

〔解法一〕 如果按小样本直接计算，$N=20$．将数据置入程序后的DATA语句中(见程序清单1700，1710语句)．按表1—2操作键盘．

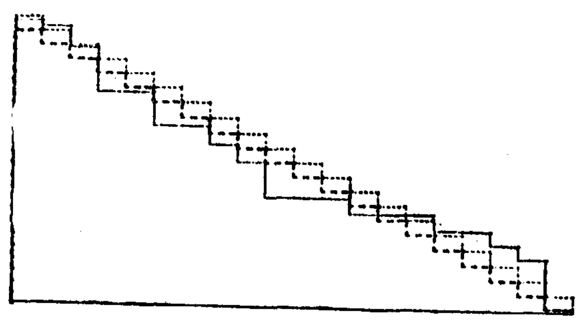
表1—2

序	键 操 作		按键后的显示	备 注
1	*DEF*	*Z*	*N* =	开始计算
2	20	*ENTER*	BUSY *N* = 20	
3			*Normla?*	
4	*Y*	*ENTER*	BUSY *Normal? Y*	
5			*Print Formation?*	
6	*N*	*ENTER*	*KO* =	
7		*ENTER*	BUSY *KO* = >	结 束
8	*DEF*	*X*	BUSY > >	绘分布阶梯图 绘图结束

打印结果如下：

```
Xc= 0.9655
S= 3.243076354E-02
M2= 0.001055

Sn#= 1.71595657
Ko= 7.338333676E-0
1
Yes normal distrib
ution; a= 05

  --Q.C--
```

绘图结果如下：



〔**解法二**〕 按由程序分组的大样本方法计算，数据置入程序后，按表1—3操作键盘。

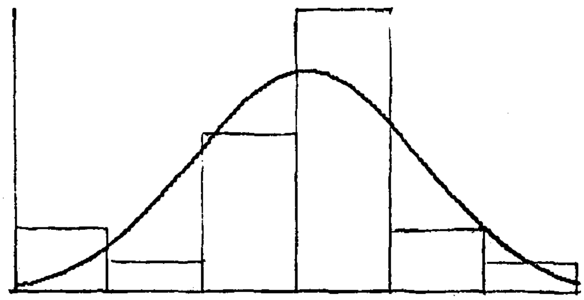
表1—3

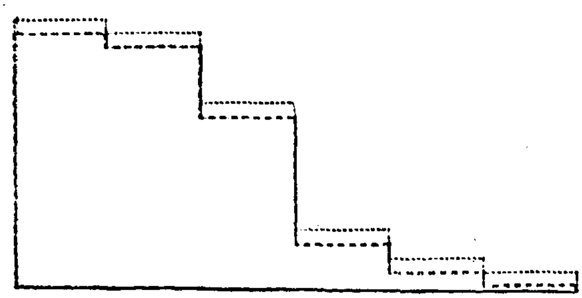
序	键 操 作		按键后的显示	备 注
1	*DEF*	*S*	*N* =	开始计算
2	20	*ENTER*	BUSY *N* = 20	
3			*Normal?*	
4	*Y*	*ENTER*	BUSY *Normal? Y*	
5			*Print Formation?*	
6	*N*	*ENTER*	*KO* =	
7	1.36	*ENTER*	BUSY *K* = 1.36 >	结 束
8	*DEF*	*D*	BUSY > >	绘频率直方图 绘完
9	*DEF*	*X*	BUSY > >	绘分布阶梯图 绘完

打印结果如下：

```
Xc= 0.9655
S= 3.248876354E-02
M2= 0.001055

C= 0.027
X(1)= 0.895

Z(1)= 0.9315
F#(1)= 2
Z(2)= 0.9285
F#(2)= 1
Z(3)= 0.9355
F#(3)= 5
Z(4)= 0.9625
F#(4)= 9
Z(5)= 0.9935
F#(5)= 2
Z(6)= 1.0165
F#(6)= 1
Z(7)= 1.0435

Sn#= 0.41583662
Ko= 5.062630497E-0
1
Yes normal distrib
ution; a=.05

  --Q.C--
```

绘图结果如下：





〔**例1—2**〕 表1—4是已经分组的一批钢材强度数据，试做正态分布拟合检验；取显著性水平$\alpha=0.05$；请打印出各组分布函数值来.

表1—4 810个钢样的强度分组

强度(公斤/毫米²)	～25.2	～26.2	～27.2	～28.2	～29.2	～30.2	～31.2
频数	2	5	15	50	88	124	181
强度(公斤/毫米²)	～32.2	～33.2	～34.2	～35.2	～36.2	～37.2	Σ
频数	137	98	62	30	14	4	810

〔**解**〕 在所分的13组数据中，分组组距为$c=1$（公斤/毫米²），第一组组中值为$x(1)=24.7$(公斤/毫米²)；计算时，将数据置入程序后，按表1—5操作键盘.