中国民用航空飞行学院硕士研究生

应用数理统计

实验讲义

高等数学教研室 曾艳编

2019 年 12 月修订

目　录

附：Matlab 2007b PLP（注册码）
18-41519-34649-39940-00621-01988-02577-01245-51575-44112-12966-44686-37374-43430-36283-64095-18584-34803-54175-05965-54469-56859-47170-56703-00300-00857-63903-48349-07297-57752-37962-48933-62342-43508-41646-31266-38461-54713-50260-57403-18654-13756-59612-18880

经验分布函数 ecdf——Empirical cumulative distribution function
概率密度函数 pdf——Probability density functions
累积分布函数 cdf——Cumulative distribution functions
随机数 rnd——Random numbers
样本标准差 std——Standard deviation
相关系数 corrcoef——Correlation coefficients
协方差 cov——Covariance
最大似然估计 mle——Maximum likelihood estimates
置信区间 ci——Confidence intervals
临界值 cv——Critical Value
皮尔逊 χ^2 拟合检验 chi2gof——Chi-square goodness-of-fit test
Q-Q 图 qqplot——Quantile-quantile plot
多项式拟合 polyfit——Polynomial curve fitting
多项式求值 polyval——Polynomial evaluation
方差分析 anova——analysis of variance
多重比较 multcompare——Multiple comparison

绪言　MATLAB 基础知识

一、MATLAB 软件简介

1967 年美国 Mathwork 公司推出了、基于矩阵运算的“Matrix Laboratory”(缩写为 MATLAB)的交互式软件包. MATLAB 既是一种直观、高效的计算机语言, 同时又是一个科学计算平台. 它为数据分析和数据可视化、算法和应用程序开发提供了最核心的数学和高级图形工具. 根据它提供的 600 多个数学和工程函数, 工程技术人员和科学工作者可以在它的集成环境中交互或编程以完成各自的计算. MATLAB 一般用于线性代数、概率统计、图像处理、样条分析、信号处理、小波分析、振动理论、神经网络、自动控制、系统识别、算法优化和财政金融等各个方面.

不过, MATLAB 作为一种新的计算机语言, 要想运用自如, 充分发挥它的威力, 也需要系统的学习. 但由于使用 MATLAB 编程运算与人进行科学计算的思路和表达方式完全一致, 所以不像学习其他高级语言如 Basic、Fortan 和 C 语言等那样难于掌握. 下面的内容均是基于 MATLAB7.5 版本.

1．MATLAB 的主要功能

(1) 数值计算功能(Numeric)
(2) 符号计算功能(Symblic)
(3) 图形和可视化功能 (Graphic)
(4) MATLAB 的活笔记本功能(Notebook)
(5) 可视化建模和仿真功能(Simulink)

2．MATLAB 的工作环境

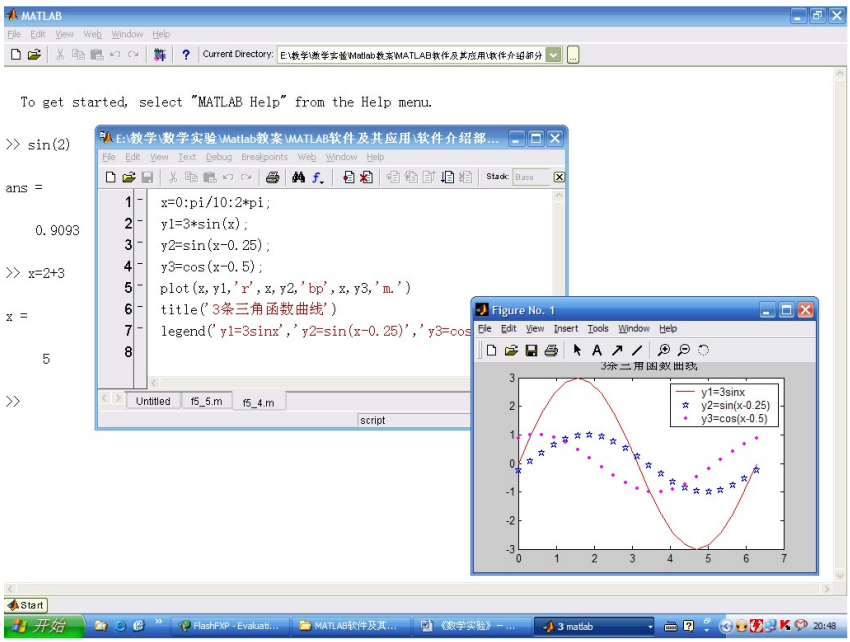


图 0-1　MATLAB 6.*x* 的命令窗、文本编辑窗、图形窗、菜单栏和工具栏

MATLAB 的工作环境主要包括:

- 【Command Window】命令窗口;
- 【File Editor】文本编辑窗口;
- 【Figure Window】图形窗口.

MATLAB 7.5 还包含几个辅助视窗, 组成其“桌面系统”. 它们分别为:

- 【Workspace】工作台窗口;
- 【Command History】指令历史纪录窗口;
- 【Current Directory】当前目录选择窗口.

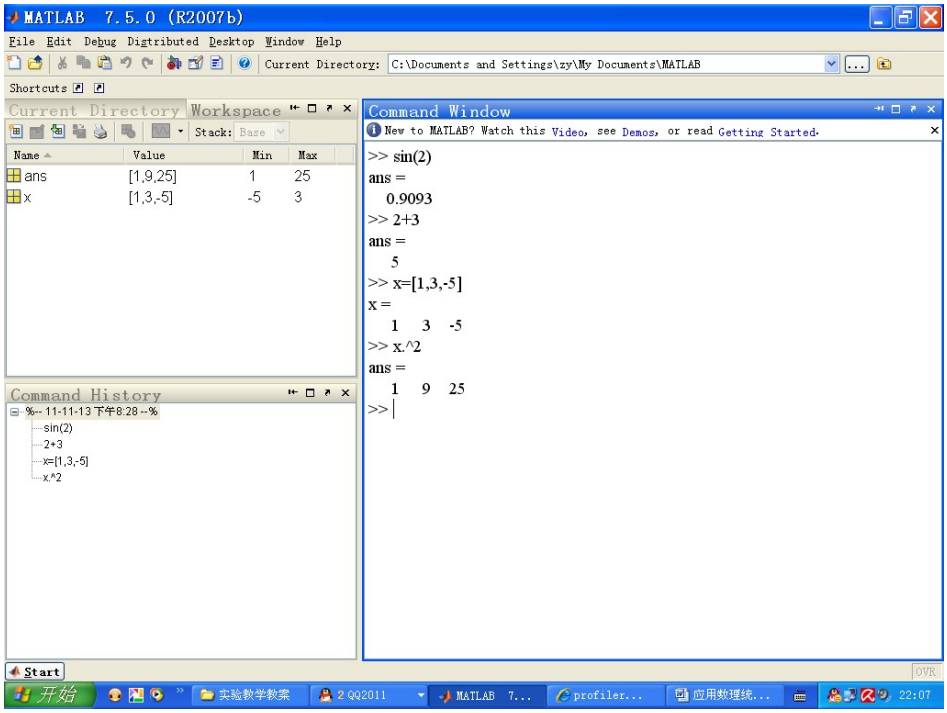


图 0-2　MATLAB 7.5 的桌面系统和命令窗口

3．MATLAB 的工作原理

(1) 语言结构:

MATLAB 语言 = 窗口命令 + m 文件

(2) 窗口命令:

在 MATLAB 命令窗口中输入的 MATLAB 语句, 并直接执行它们完成相应的运算、绘图等.

(3) m 文件:

在 MATLAB 文本编辑窗口中用 MATLAB 语句编写的磁盘文件, 扩展名为“.m”.

二、MATLAB 入门

1．数据的格式与输入、提取

由于 MATLAB 是基于矩阵的运算, 所以在数值计算中, 所有数据都是以矩阵的形式输入, 且对非向量型矩阵, 如不作特殊说明, 都是列优先. 另外, 数可看作是 1×1 的矩阵.

【例 0-1】输入矩阵 $A=\begin{bmatrix}1 & -2 & 3 & -6\\ 4 & 0 & -5 & 2\end{bmatrix}$ 与其转置阵 $A'=\begin{bmatrix}1 & 4\\ -2 & 0\\ 3 & -5\\ -6 & 2\end{bmatrix}$, 并提取 $a_{23},[a_{12},a_{13},a_{14}]$.

```
>>A=[1,-2,3,6;4,0,-5,2]  %矩阵内换行用分号“;”, 同一行内的数据用逗号“,”或“空格”隔开
    A =   1   -2    3    6
          4    0   -5    2
>>B=A'                   %转置用单引号“’”
    B =   1    4
         -2    0
          3   -5
          6    2
>>a23=A(2,3)             %提取A阵第2行第3列元素
    a23 =   -5
>> A(1, 2:4)             %提取A阵第1行第2-4列元素, 2:4表示以2为起点4为终点步长为1的向量
    ans =   -2    3    6
```

2．基本库函数

(1) 基本初等函数:

- 幂函数:

  sqrt(x) — 开平方 $\sqrt{x}$　　x^u — 幂函数 x^u

- 指数函数:

  exp(x) — 以 e 为底的指数 e^x　　a^x — 指数函数 a^x

- 对数函数:

  log(x) — 自然对数 $\ln x$　　log10(x) — 以 10 为底的对数 $\lg x$

- 三角函数与反三角函数:

  sin(x) — 正弦函数　　cos(x) — 余弦函数

  tan(x) — 正切函数　　cot(x) — 余切函数

  sec(x) — 正割函数　　csc(x) — 余割函数

  asin(x) — 反正弦函数　　acos(x) — 反余弦函数

  atan(x) — 反正切函数　　acot(x) — 反余切函数

  asec(x) — 反正割函数　　acsc(x) — 反余割函数

(2) 常用基本函数:

  abs(x)—取绝对值　　round(x)—四舍五入法取整

  ceil(x)—向右取整　　floor(x)—向左取整

  sum(x)—求和　　prod(x)－求积

max(x)—最大值　　min(x)—最小值

length(x)—矩阵行数与列数中的最大值　　size(x)—矩阵的行数与列数

注意: (1) 由于 MATLAB 是基于矩阵的运算, 所以上面的 x 均表示矩阵, 数可看作是 1×1 的矩阵.

(2) 对非向量型矩阵, 如不作特殊说明, 都是列优先.

3．数学运算符及特殊字符

数组的算术运算符: + - .* ./ .\ .^

矩阵的算术运算符: + - * / \ ^

关系运算符: < <= > >= == ~=

逻辑运算符: & 与; | 或; ~ 非

三种运算的顺序依次为: 算术运算、关系运算、逻辑运算.

pi	数学常数, 即3.1415926535897....
eps	系统的浮点 (Floating-ponit) 精确度. 在PC机上, 它等于 2^{-52}
Inf	正无穷大, 定义为 $\frac{1}{0}$
ans	计算结果的默认变量名
NaN	不定值, 由 Inf/Inf或0/0等运算产生

4．命令行的编写

随时输入指令并按回车键, 即时给出结果;

在指令最后不用任何符号并按回车键, 将显示最后结果;

在指令最后用“;”并按回车键, 将只计算但不显示最后结果.

同时输入几条指令时, 用“,”或“;”隔开.

【例 0-2】数学运算符、特殊字符与基本库函数的应用

```
>>3*(-5), 2/5, [1 2 3].*[2 4 5], [1 2 3]./[2 4 5], [2,4,5].^2
    ans =   -15
    ans =    0.4000
    ans =     2     8    15
    ans =    0.5000    0.5000    0.6000
    ans =     4    16    25
>> sin(pi/4), log(exp(1))
    ans =    0.7071
    ans =     1
```

5．变量与表达式

在MATLAB中, 把由下标表示次序的标量的集合称为矩阵或数组. MATLAB是基于矩阵运算的, 因此其基本数据结构只有一个: 矩阵. 一个数也是矩阵, 只不过它是1行×1列的矩阵. MATLAB中的变量可用来存放数据, 也可用来存放向量或矩阵, 并进行各种运算.

(1)变量命名的规则为:

• 变量名是要区分大小写字母的;
• 第一个字符必须是英文字母;
• 字符间不可留空格;
• 最多只能有31个字符 (只能有英文字母、数字和下连字符) .

(2)表达式由变量名、运算符和函数名等组成. 如x/sin(x), 其中x为变量名, /为运算符, sin为函数名.

(3)MATLAB语句有两种最常见形式:

1) 表达式;

2) 赋值语句: 变量 = 表达式.

【例 0-3】赋值语句的使用

```
>> x=1; y=x/sin(x)
   y =    1.1884
>> x=[pi/6,pi/4,pi/3,pi/2]; sin(x)
   ans =0.5000   0.7071   0.8660   1.0000
>> x=0:0.1:2*pi; y=sin(x); plot(x,y)
```

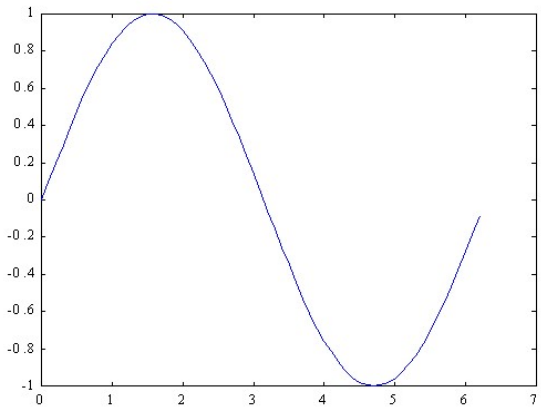


图 0-3　y=sin(x)的曲线图

6．m 文件的建立、编写、保存与调用

(1) 进入文本编辑窗口的方式: 在菜单栏“File”下直接点击“新建…”进入文本编辑窗口.

(2) m 文件的分类与格式:

① 命令文件: 由一系列 MATLAB 语句组成, 运行时将自动执行一系列命令直至给出最后结果, 而不交互地等待键盘输入. 命令文件定义的变量为全局变量, 存放于内存.

② 函数文件: 第一行必须包含“function”, 主要功能是建立一个函数. 函数文件定义的变量为局部变量.

function 因变量名= 函数名(自变量名)

注意: 函数文件要求函数名和文件名相同, 且函数名、文件名与变量名的命名规则一样.

(3) 退出文本编辑窗口: 录入完毕, 存盘退出文本编辑窗口则可.

【例 0−4】求 $f(x)=\lg\sqrt{(x-5)^2+(x-10)^2}$ 分别在 $x=0,x=5,x=10$ 处的函数值.

1)利用命令文件完成任务

• 在文本编辑窗口中编写命令文件example0_4_1.m:

```
x=[0,5,10];
y=log10(sqrt((x-5).^2+(x-10).^2))        %显示运行结果
```

• 在命令窗口中运行命令文件example0_4_1.m:

```
>>  example0_4_1;
    y =    1.0485    0.6990    0.6990
```

2)利用函数文件完成任务

• 在文本编辑窗口中编写函数文件example0_4_2.m:

```
function y= example0_4_2(x)
y=log10(sqrt((x-5).^2+(x-10).^2));
```

• 在命令窗口中调用函数文件example0_4_2.m:

```
>>x=[0,5,10];  y= example0_4_2(x);
    y =    1.0485    0.6990    0.6990
```

【例 0-5】已知 $x_1=-2,\ x_2=3,\ x_3=1$, 而 $\begin{cases} z_1=3x_1^2 \\ z_2=x_2+x_3 \end{cases}$, $\begin{cases} y_1=z_1+z_2 \\ y_2=z_1-z_2 \end{cases}$, 试求 y_1, y_2 的值.

1)利用命令文件完成任务

• 在文本编辑窗口中编写命令文件example0_5_1.m:

```
x1=-2;x2=3;x3=1;
z1=3*x1^2;
z2=x2+x3;
y1=z1+z2
y2=z1-z2
```

• 在命令窗口中运行命令文件example0_5_1.m:

```
>> example0_5_1
    y1 =    16
    y2 =     8
```

2)利用函数文件完成任务

• 在文本编辑窗口中编写函数文件example0_5_2.m:

```
function [y1, y2]= example0_5_2(x1, x2, x3)
z1=3*x1^2;
z2=x2+x3;
y1=z1+z2
y2=z1-z2
```

• 在命令窗口中调用函数文件example0_5_2.m:

```
>> [y1, y2]=example0_5_2(-2, 3, 1)
    y1 =    16
    y2 =     8
```

7．MATLAB 的在线帮助

(1) 从菜单栏上的“help”进入

(2) 其它命令窗口帮助

clc —— 清除显示屏上的内容

clear —— 清除内存变量和函数

what —— 列出当前目录下的m、mat、mex文件

who —— 列出当前工作空间 (Workspace) 的变量名

8．路径的设置

在保存 m 文件时, MATLAB 的默认位置是 C:\Program Files\MATLAB71\work. 如果用户将编写的 m 文件保存在 E:\experiment 目录下, 则从 MATLAB 窗口的“File”菜单中单击子菜单“Save As…”, 选择 E:\experiment, 再输入本 m 文件的文件名, 按“保存”键返回则可.

第一章　数理统计的基本概念

一、直方图与经验分布函数图的绘制

hist(A,n) —— 对矩阵A按列作统计频数直方图, n为条形图的条数

ni=hist(A,n)—— 对矩阵A按列得各划分区间内的统计频数

注意: 当A为向量时, 上述所有命令直接作用在向量上, 而不是列优先.

[Fn,x0]=ecdf(x) —— 得到样本x的经验分布函数值Fn, 当x中有m个不同的数 (记为向量x0) 时, 则Fn的个数为m+1个

ecdfhist(Fn,x0, m) —— 绘制数据x的频率(密度)直方图, 其中Fn与x0是由ecdf函数得到的样本x的经验分布函数值Fn与分段点x0, m为条形的个数, m的默认值为10

cdfplot(x) —— 绘制样本x的经验分布函数图

例如:

```
>> x = [6  4  5  3  6  8  6  7  3  4];
>> [Fn,x0]=ecdf(x)
      Fn =   0   0.2000   0.4000   0.5000   0.8000   0.9000   1.0000
      x0 =   3        3        4        5        6        7        8
>> cdfplot(x)
```

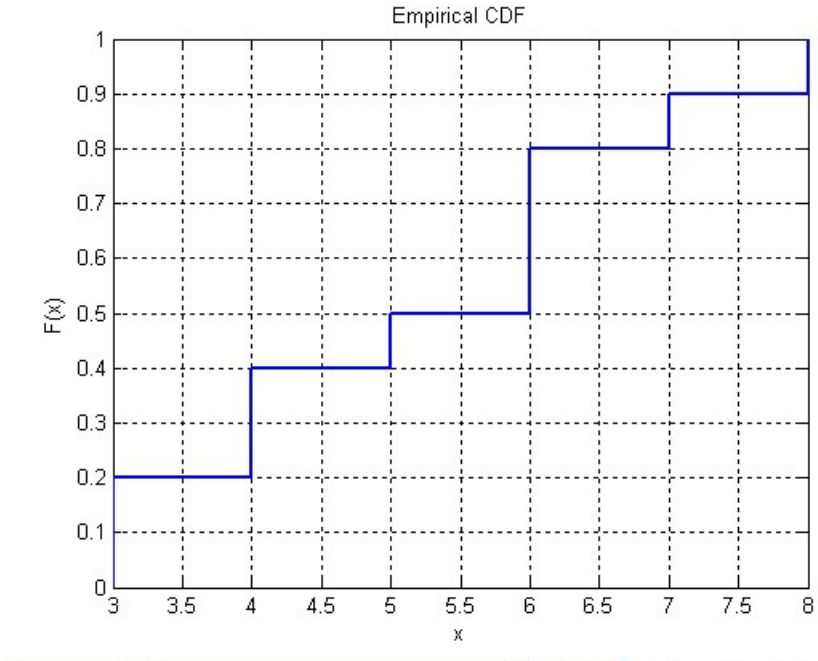


图 1-1　经验分布函数图

【⋆例1.6(P_{10})】在齿轮加工中, 齿轮的径向综合误差 $\Delta F_i''$ 是个随机变量, 今对200件同样的齿轮进行测量, 测得 $\Delta F_i''$ 的数值 (mm) 如下, 求作 $\Delta F_i''$ 的频率密度直方图, 并作出 $\Delta F_i''$ 的经验分布函数图形.

16 25 19 20 25 33 24 23 20 24 25 17 15 21 22 26 15 23 22 24
20 14 16 11 14 28 18 13 27 31 25 24 16 19 23 26 17 14 30 21
18 16 18 19 20 22 19 22 18 26 26 13 21 13 11 19 23 18 24 28
13 11 25 15 17 18 22 16 13 12 13 11 09 15 18 21 15 12 17 13
14 12 16 10 08 23 18 11 16 28 13 21 22 12 08 15 21 18 16 16
19 28 19 12 14 19 28 28 28 13 21 28 19 11 15 18 24 18 16 28
19 15 13 22 14 16 24 20 28 18 18 28 14 13 28 29 24 28 14 18
18 18 08 21 16 24 32 16 28 19 15 18 18 10 12 16 26 18 19 33
08 11 18 27 23 11 22 22 13 28 14 22 18 26 18 16 32 27 25 24
17 17 28 33 16 20 28 32 19 23 18 28 15 24 28 29 16 17 19 18

• 编写命令文件example1_6.m:

```
F=[16 25 19 20 25 33 24 23 20 24 25 17 15 21 22 26 15 23 22 24....
   20 14 16 11 14 28 18 13 27 31 25 24 16 19 23 26 17 14 30 21....
   18 16 18 19 20 22 19 22 18 26 26 13 21 13 11 19 23 18 24 28....
   13 11 25 15 17 18 22 16 13 12 13 11 09 15 18 21 15 12 17 13....
   14 12 16 10 08 23 18 11 16 28 13 21 22 12 08 15 21 18 16 16....
   19 28 19 12 14 19 28 28 28 13 21 28 19 11 15 18 24 18 16 28....
   19 15 13 22 14 16 24 20 28 18 18 28 14 13 28 29 24 28 14 18....
   18 18 08 21 16 24 32 16 28 19 15 18 18 10 12 16 26 18 19 33....
   08 11 18 27 23 11 22 22 13 28 14 22 18 26 18 16 32 27 25 24....
   17 17 28 33 16 20 28 32 19 23 18 28 15 24 28 29 16 17 19 18];
%(1) ----------下面作频数直方图
figure(1)
hist(F,8)
title('频数直方图');
xlabel('齿轮的径向综合误差(mm)');
%(2) ----------下面作频率(密度)直方图
[Fn,x0]=ecdf(F);
figure(2)
ecdfhist(Fn,x0,8);
title('频率(密度)直方图');
xlabel('齿轮的径向综合误差(mm)');
%(3) ----------下面作经验分布函数图
figure(3)
cdfplot(F)
title('经验分布函数图');
xlabel('齿轮的径向综合误差(mm)');
```

• 运行命令文件example1_6.m:

```
>> example1_6
```

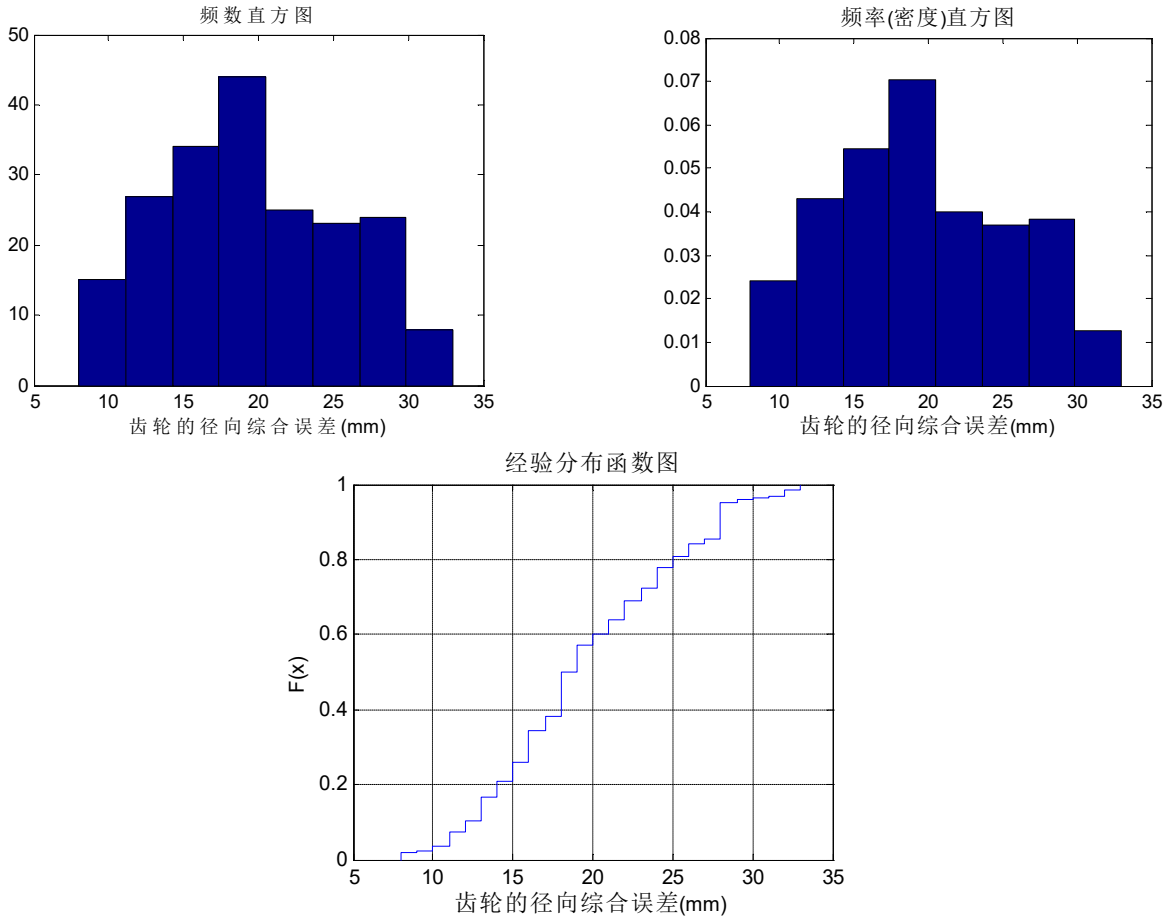


图1-2

二、常见的概率分布

表 1-1　常用概率分布及代码

连续型分布				离散型分布	
分布名称	代码	分布名称	代码	分布名称	代码
连续均匀分布	unif	χ^2 分布	chi2	二项分布	bino
指数分布	exp	非中心 χ^2 分布	ncx2	离散均匀分布	unid
正态分布	norm	F 分布	f	几何分布	geo
多维正态分布	mvn	非中心 F 分布	ncf	超几何分布	hyge
对数正态分布	logn	t 分布	t	负二项分布	nbin
β 分布	beta	非中心 t 分布	nct	泊松分布	poiss
γ (Gamma)分布	gam	多维 t 分布	mvt		
Rayleigh 分布	rayl	I 型极值分布	ev		
Weibull 分布	wbl	广义极值分布	gev		

注意: MATLAB 中的指数分布的概率密度函数是 $f(x)=\begin{cases}\dfrac{1}{\theta}e^{-\frac{x}{\theta}}, & x>0\\ 0, & x\le 0\end{cases}$.

三、MATLAB 为常见分布提供的五类函数

1) 概率密度函数(分布名+pdf)

2) (累积)分布函数(分布名+cdf)

3) 逆(累积)分布函数——左分位点(分布名+inv)

4) 随机数发生器(分布名+rnd)

5) 均值和方差(分布名+stat)

1．概率密度函数

表 1-2　概率密度函数(pdf)

函数名称	函数说明	调用格式
normpdf	正态分布	Y=normpdf (X, mu, sigma)
chi2pdf	χ^2 分布	Y=chi2pdf (X, N)
tpdf	t 分布	Y=tpdf (X, N)
fpdf	F 分布	Y=fpdf (X, N1, N2)

注意: Y=normpdf (X, mu, sigma)的 sigma 是指标准差σ, 而非σ^2.

【补例 1-1】 绘制标准正态分布 $N(0,1)$ 的概率密度图.

```
x= -4:0.1:4;
y=normpdf(x,0,1);
plot(x,y)
title('N(0,1)的概率密度曲线图')
```

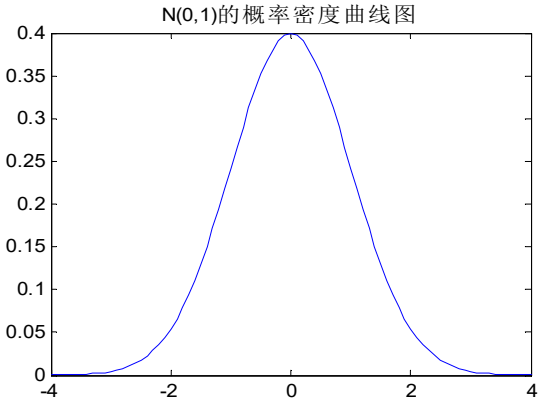


图 1-3　标准正态分布的概率密度图

2．累积分布函数

表 1-3　累积分布函数(cdf)

函数名称	函数说明	调用格式
normcdf	正态分布	P=normcdf (X, mu, sigma)
chi2cdf	χ^2 分布	P=chi2cdf (X, N)
tcdf	t 分布	P=tcdf (X, N)
fcdf	F 分布	P=fcdf (X, N1, N2)

【补例 1-2】求服从标准正态分布的随机变量落在区间[-2, 2]上的概率.

```
>> P=normcdf (2,0,1)-normcdf(-2,0,1)
    ans =    0.9545
```

3．逆累积分布函数 (用于求左分位数)

表 1-4　逆累积分布函数(inv)

函数名称	函数说明	调用格式
norminv	正态分布	X=norminv (P, mu, sigma)
chi2inv	χ^2 分布	X=chi2inv (P, N)
tinv	t 分布	X=tinv (P, N)
finv	F 分布	X=finv (P, N1, N2)

【★例 1.13(P_{22})】求下列左分位数:

(i) $u_{0.9}$; (ii) $t_{0.25}(4)$; (iii) $F_{0.1}(14,10)$; (iv) $\chi^2_{0.025}(50)$.

```
>> u_alpha=norminv(0.9,0,1)
    u_alpha =    1.2816
>> t_alpha=tinv(0.25,4)
    t_alpha =   -0.7407
>> F_alpha=finv(0.1,14,10)
    F_alpha =    0.4772
>> X2_alpha=chi2inv(0.025,50)
    X2_alpha =   32.3574
```

4．随机数发生函数[1]

表 1-5　随机数发生函数(rnd)

函数名称	函数说明	调用格式
normrnd	正态分布	R=normrnd(mu, sigma, m, n)
chi2rnd	χ^2 分布	R=chi2rnd(N, m, n)
trnd	t 分布	R=trnd(N, m, n)
frnd	F 分布	R=frnd(N1, N2, m, n)

5．均值和方差

表 1-6　常见分布的均值和方差函数(stat)

函数名称	函数说明	调用格式
unifstat	连续均匀分布: $\mu=\frac{a+b}{2}$, $\sigma^2=\frac{(b-a)^2}{12}$	[M,V]=unifstat (A, B)

[1] 这是一个随机数发生器，常用于蒙特卡罗(Monte Carlo)方法中. Monte Carlo 方法是一种随机抽样或统计试验方法，是一种基于“随机数”的计算方法，属于计算数学的一个分支，是 20 世纪 40 年代中期为了适应当时原子能事业的发展而发展起来的. 它是由数学家冯·诺伊曼用世界赌城——摩纳哥首都 Monte Carlo 来命名的，冯·诺伊曼是美国二次世界大战期间研制原子弹的“曼哈顿计划”的主持人之一.

expstat	指数分布: $\mu=\theta$, $\sigma^2=\theta^2$	[M,V]=expstat (theta)
normstat	正态分布: $\mu=\mu$, $\sigma^2=\sigma^2$	[M,V]=normstat (mu, sigma)
chi2stat	χ^2 分布: $\mu=n$, $\sigma^2=2n$	[M,V]=chi2stat (N)
tstat	t 分布: $\mu=0\,(n\geq 2)$, $\sigma^2=\dfrac{n}{n-2}\,(n\geq 3)$	[M,V]=tstat (N)
fstat	F 分布: $\mu=\dfrac{n_2}{n_2-2}\,(n_2\geq 3)$ $\sigma^2=\dfrac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}\,(n_2\geq 5)$	[M,V]=fstat (N1, N2)
binostat	二项分布: $\mu=np$, $\sigma^2=np(1-p)$	[M,V]=binostat (N, p)
poisstat	泊松分布: $\mu=\lambda$, $\sigma^2=\lambda$	[M,V]=poisstat (LAMBDA)

注意: 表 1-6 中, 如果省略调用格式左边的[M, V], 则只计算出均值.

四、常用的统计量

表 1-7　常用统计量

函数名称	函数说明	调用格式
mean	样本均值	m=mean(X)
range	样本极差	y=range(X)
std	样本标准差	y=std(X) , y=std(X, 1)
var	样本方差	y=var(X), y=var(X, 1)
moment	任意阶中心矩	m=moment(X, order)
cov	协方差矩阵	C=cov(X), C=cov(X, Y)
corrcoef	相关系数阵	R=corrcoef (X)

说明:

(1) y=std(X) ——计算 X 中每列数据的标准差, 其中 $\text{std}(X)=\sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$.

y=std(X, 1) ——与 std(X)相比, 差异仅为 $\text{std}(X)=\sqrt{\dfrac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$.

(2) y=var(X) ——计算 X 中每列数据的方差, 其中 $\text{var}(X)=\dfrac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$.

y=var(X, 1) ——与 var(X)相比, 差异仅为 $\text{var}(X)=\dfrac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$.

(3) C＝cov(X) ——返回一个协方差矩阵. 如果 X 为 n×m 的矩阵, 则 C 为 m×m 的矩阵. 其中, 第 j 列与第 k 列两个样本的协方差 $c_{jk}=\dfrac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_{\cdot j})(x_{ik}-\bar{x}_{\cdot k})$.

C＝cov(X,1) ——与 cov(X)相比, 差异仅为 $c_{jk}=\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_{\cdot j})(x_{ik}-\bar{x}_{\cdot k})$.

(4) var(X)=diag(cov(X)), std(X)=sqrt(diag(cov(X))), 其中 diag(X)表示取方阵 X 中的对角元素.

注意: 若 X 为向量, 则表示一维随机变量的样本; 若 X 为 n×m 的矩阵, 则 X 表示总体为 m 维随机向量 $(X_1,\cdots,X_m)$ 的样本数据矩阵, 其中第 j 列的 n 个数据表示指标 X_j 的样本.

【★例1.6(P10)续】在齿轮加工中, 齿轮的径向综合误差 $\Delta F_i''$ 是个随机变量, 今对200件同样的齿轮进行测量, 测得 $\Delta F_i''$ 的数值 (mm) 如下. 若假设该批数据来自正态分布 $N(\mu,\sigma^2)$, 试求出 μ,σ^2 的点估计, 并在同一图中作出 $\Delta F_i''$ 的频率密度直方图与该正态分布的概率密度图.

16 25 19 20 25 33 24 23 20 24 25 17 15 21 22 26 15 23 22 24
20 14 16 11 14 28 18 13 27 31 25 24 16 19 23 26 17 14 30 21
18 16 18 19 20 22 19 22 18 26 26 13 21 13 11 19 23 18 24 28
13 11 25 15 17 18 22 16 13 12 13 11 09 15 18 21 15 12 17 13
14 12 16 10 08 23 18 11 16 28 13 21 22 12 08 15 21 18 16 16
19 28 19 12 14 19 28 28 28 13 21 28 19 11 15 18 24 18 16 28
19 15 13 22 14 16 24 20 28 18 18 28 14 13 28 29 24 28 14 18
18 18 08 21 16 24 32 16 28 19 15 18 18 10 12 16 26 18 19 33
08 11 18 27 23 11 22 22 13 28 14 22 18 26 18 16 32 27 25 24
17 17 28 33 16 20 28 32 19 23 18 28 15 24 28 29 16 17 19 18

• 编写命令文件example1_6_2.m:

```
%书P10例1.6续: 作“频率密度直方图”与“概率密度图”的程序, 划分为8个区间
F=[16 25 19 20 25 33 24 23 20 24 25 17 15 21 22 26 15 23 22 24....
   20 14 16 11 14 28 18 13 27 31 25 24 16 19 23 26 17 14 30 21....
   18 16 18 19 20 22 19 22 18 26 26 13 21 13 11 19 23 18 24 28....
   13 11 25 15 17 18 22 16 13 12 13 11 09 15 18 21 15 12 17 13....
   14 12 16 10 08 23 18 11 16 28 13 21 22 12 08 15 21 18 16 16....
   19 28 19 12 14 19 28 28 28 13 21 28 19 11 15 18 24 18 16 28....
   19 15 13 22 14 16 24 20 28 18 18 28 14 13 28 29 24 28 14 18....
   18 18 08 21 16 24 32 16 28 19 15 18 18 10 12 16 26 18 19 33....
   08 11 18 27 23 11 22 22 13 28 14 22 18 26 18 16 32 27 25 24....
   17 17 28 33 16 20 28 32 19 23 18 28 15 24 28 29 16 17 19 18];
%(1)下面作频率(密度)直方图
[Fn,x0]=ecdf(F);
ecdfhist(Fn, x0, 8);
hold on
%(2)下面作正态分布的概率密度图
mu=mean(F);              %估计均值mu
sigma=std(F);            %估计标准差sigma
x=min(F)-10:0.1:max(F)+10;
```

```
y=normpdf(x, mu, sigma);
plot(x,y,'r', 'LineWidth', 2)
title('频率(密度)直方图与概率密度图');
xlabel('齿轮的径向综合误差(mm)');
hold off
```

• 运行命令文件example1_6_2.m:

```
>> example1_6_2
```

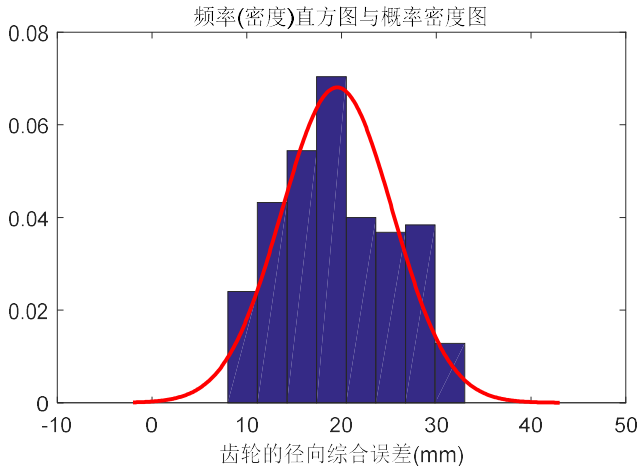


图1-4

作业: $P_{26\text{-}28}$ 1.1, 1.12

第二章　参数估计

参数估计包含两种常用方式: 点估计和区间估计.

Matlab 统计工具箱给出了常用概率分布中参数的点估计 (采用最大似然估计法－MLE)与区间估计, 另外还提供了部分分布的对数似然函数的计算功能.

一、统计工具箱中的参数估计

1．利用 mle 函数对概率分布中的参数进行估计

格式: [phat,pci] = mle(data,'distribution',dist, 'alpha', alpha,'ntrials',n)

功能: 根据样本数据 data, 给出由 dist 指定分布的参数的 MLE 估计与区间估计.

说明:

(1) phat 与 pci 中的“p”为分布中的参数, 可表示多个参数;

(2) phat 为参数的最大似然估计值, pci 为参数的置信水平为$1-\alpha$的置信区间, 其中α由 alpha 的取值确定;

(3) dist 为总体分布的指定类型, 其取值为表 1-1 中的代码;

(4) 当 dist 为'norm'时, phat 中的参数“p”是指μ,σ, 而非μ,σ^2.

(5) alpha 为显著性水平, 取值在 0－1 之间. $\alpha=0.05$是默认值, 此时本选项可省略;

(6) 'ntrials', n 只在 dist 为 bino (二项分布$b(n,p)$) 时才选用, n 为$b(n,p)$中的 n, 此时待估参数为 p.

【★习题 2.3(P_{66})】使用一测量仪器对同一值进行了 12 次独立测量, 其结果为 (单位: mm)

232.50　232.48　232.15　232.52　232.53　232.30

232.48　232.05　232.45　232.60　232.47　232.30

并设测量值$X\sim N(\mu,\sigma^2)$, 试求μ,σ^2的最大似然估计与区间估计($\alpha=0.05$).

(1) 问题分析:

设总体X——测量值, 且$X\sim N(\mu,\sigma^2)$.

今抽得一容量为12的样本, 本问题是求参数μ,σ^2的最大似然估计与置信水平为0.95的区间估计.

(2) 问题求解:

μ,σ^2的最大似然估计分别为: $\hat{\mu}_{MLE}=\bar{X},\ \hat{\sigma}^2_{MLE}=\dfrac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$;

μ的置信水平为$1-\alpha$的置信区间为: $\left[\bar{X}-\dfrac{S_*}{\sqrt{n}}t_{1-\alpha/2}(n-1),\ \bar{X}+\dfrac{S_*}{\sqrt{n}}t_{1-\alpha/2}(n-1)\right]$;

σ^2的置信水平为$1-\alpha$的置信区间为: $\left[\dfrac{(n-1)S_*^2}{\chi^2_{1-\alpha/2}(n-1)},\ \dfrac{(n-1)S_*^2}{\chi^2_{\alpha/2}(n-1)}\right]$.

• 编写命令文件exercise2_3.m:

```
x=[232.50, 232.48, 232.15, 232.52, 232.53, 232.30, 232.48, 232.05, 232.45, 232.60,
232.47, 232.30];
[hat,ci]=mle(x,'distribution','norm')
```

• 运行命令文件exercise2_3.m:

```
>> exercise2_3
    hat =   232.4025   0.1598
    ci =    232.2965   0.1182
            232.5085   0.2834
```

(3) 问题结果 :

$\hat{\mu}_{MLE}=232.4025,\quad \hat{\sigma}^2_{MLE}=0.1598^2=0.0255$,

μ 的置信区间为$[232.2965, 232.5085]$,

σ^2 的置信区间为$[0.1182^2, 0.2834^2]=[0.0140, 0.0803]$.

【★例 2.31(P66)】对一批产品, 欲通过抽样检查其不合格率. 今抽取产品 100 件, 发现不合格品有 4 件, 求不合格率的 0.95 的双侧置信区间.

(1) 问题分析:

设总体 $X=\begin{cases}1, & \text{本产品为不合格品}\\ 0, & \text{本产品为合格品}\end{cases}$, 即 $X\sim b(1,p)$.

今抽得一容量为100的样本$(x_1=\cdots=x_4=1, x_5=\cdots=x_{100}=0)$, 本问题即要求参数 p 的双侧置信区间.

(2) 问题求解:

选用下列两种方法计算:

①利用P65$(n+u^2_{1-\alpha/2})p^2-(2n\overline{X}+u^2_{1-\alpha/2})p+n\overline{X}^2\le 0$近似计算(独立同分布中心极限定理);

②利用Matlab中二项分布参数估计函数mle计算 (借助F分布).

• 编写命令文件example2_31.m:

```
alpha=0.05;
x=[ones(1,4),zeros(1,96)];
n=length(x);
%(1)---------- 利用中心极限定理近似计算
u=norminv(1-alpha/2,0,1);
a=n+u^2;
b=-(2*n*mean(x)+u^2);
c=n*mean(x)^2;
p=[(-b-sqrt(b^2-4*a*c))/(2*a),(-b+sqrt(b^2-4*a*c))/(2*a)]      %一元二次方程求根公式
%(2) ----------利用mle计算
[phat,pci]=mle(x,'distribution','bino','ntrials',1)
```

• 运行命令文件example2_31.m:

```
>> example2_31
    p =   0.0157   0.0984                    %近似计算结果
```

phat = 0.0400

pci = 0.0110 0.0993　　%mle 计算结果

(3) 问题结果:

①利用中心极限定理得到 p 的近似估计区间为[0.0157, 0.0984];

②利用 mle 得到 p 的估计区间为[0.0110, 0.0993].

2．当 mle 无效时, 自己编程对概率分布中的参数进行估计

【习题 2.22(P_{69})】随机地从一批零件中抽取 16 个, 测得长度 (单位: cm) 为:

2.14 2.10 2.13 2.15 2.13 2.12 2.13 2.10

2.15 2.12 2.14 2.10 2.13 2.11 2.14 2.11

设零件长度的分布为正态的, 试求总体均值的 90%的置信区间:

①若 $\sigma = 0.01$(cm);　　②若 σ 未知.

(1) 问题分析:

设总体X——零件长度, 则 $X \sim N(\mu, \sigma^2)$. 本问题是求参数 μ 的置信区间.

(2) 问题求解:

①若 $\sigma = 0.01$: μ 的置信水平为 $1-\alpha$ 的置信区间为 $\left[\bar{X} - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}, \bar{X} + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}\right]$.

- 编写函数文件zestimate.m:

```
function muci=zestimate(x,sigma,alpha)
n=length(x);
xhat=mean(x);
u_alpha=norminv(1-alpha/2,0,1);
delta1=sigma/sqrt(n)*u_alpha;
muci=[xhat-delta1,xhat+delta1];
```

- 调用函数文件zestimate.m:

```
>> x=[2.14, 2.10, 2.13, 2.15, 2.13, 2.12, 2.13, 2.10, 2.15, 2.12, 2.14, 2.10, 2.13, 2.11, 2.14,
2.11];
>> sigma=0.01;
>> alpha=0.1;
>> muci=zestimate(x,sigma,alpha)
   muci = 2.1209 2.1291
```

②若 σ 未知: μ 的置信水平为 $1-\alpha$ 的置信区间为 $\left[\bar{X} - \frac{S_*}{\sqrt{n}} t_{1-\alpha/2}(n-1), \bar{X} + \frac{S_*}{\sqrt{n}} t_{1-\alpha/2}(n-1)\right]$.

- 编写命令文件exercise2_22_2.m:

```
x=[2.14, 2.10, 2.13, 2.15, 2.13, 2.12, 2.13, 2.10, 2.15, 2.12, 2.14, 2.10, 2.13, 2.11, 2.14, 2.11];
 [phat,pci]= mle(x,'distribution','norm','alpha',0.1);
 muci=pci(:,1)
```

- 运行命令文件exercise2_22_2.m:

```
>> exercise2_22_2
    muci = 2.1175 2.1325
```

(3) 问题结果:

①当$\sigma=0.01$时, μ的置信水平为 0.90 的置信区间为[2.1209, 2.1291];

②当σ未知时, μ的置信水平为 0.90 的置信区间为[2.1175, 2.1325].

【★例 2.28(P63)】设 A、B 两种牌号的灯泡寿命(小时)相互独立, 且其寿命分别服从$N(\mu_1,\sigma_1^2)$, $N(\mu_2,\sigma_2^2)$. 随机选取 A 种灯泡 5 只, B 种灯泡 7 只, 做灯泡寿命试验, 算得$\bar{x}_A=1000$, $\bar{x}_B=980$, $s_{*A}^2=784$, $s_{*B}^2=1024$.试求$\mu_1-\mu_2$的置信水平为 0.99 的置信区间, 并说明两种灯泡的寿命是否有明显差异, 其中假设$\sigma_1^2=\sigma_2^2=\sigma^2$.

(1) 问题分析:

设总体X与Y——牌号A与B的灯泡寿命, 则$X\sim N(\mu_1,\sigma_1^2), Y\sim N(\mu_2,\sigma_2^2)$. 本问题是求$\mu_1-\mu_2$的置信区间.

(2) 问题求解:

因$\sigma_1^2=\sigma_2^2$未知, 故$\mu_1-\mu_2$的置信水平为$1-\alpha$的置信区间为$\left[(\bar{X}-\bar{Y})\pm t_{1-\alpha/2}(n_1+n_2-2)\cdot S_w\cdot\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right]$.

• 编写命令文件example2_28.m:

```
xmean=1000;  ymean=980;   xs2=784;   ys2=1024;
n1=5;   n2=7;
alpha=0.01;
t_alpha=tinv(1-alpha/2, n1+n2-2);
Sw=sqrt(((n1-1)*xs2+(n2-1)*ys2)/(n1+n2-2));
delta1=t_alpha*Sw*sqrt(1/n1+1/n2);
mu12ci=[(xmean-ymean)-delta1, (xmean-ymean)+delta1]
```

• 运行命令文件example2_28.m:

```
>> example2_28.m
    mu12ci =   -36.5315   76.5315
```

(3) 问题结果:

$\mu_1-\mu_2$的置信水平为 0.99 的置信区间为[-36.5315, 76.5315].

作业: P67-69 2.24, 2.25

第三章　假设检验

假设检验分为两种: 参数假设检验与非参数假设检验.

一、正态总体参数的假设检验

（一）检验问题、检验统计量及拒绝域

表 3-1 的说明:

对一个正态总体 $X \sim N(\mu,\sigma^2)$, 抽取样本 $X_1, X_2, \cdots, X_n$;

对两个正态总体 $X \sim N(\mu_1,\sigma_1^2)$, $Y \sim N(\mu_2,\sigma_2^2)$, 且 X 与 Y 独立, 分别抽取样本 $X_1, X_2, \cdots, X_{n_1}$ 与 $Y_1, Y_2, \cdots, Y_{n_2}$.

表 3-1　正态总体参数的假设检验

总体	H_0	H_1	剩余参数	检验统计量	H_0 的否定域	Matlab 函数
一个正态总体	$\mu=\mu_0$ $\mu\le\mu_0$ $\mu\ge\mu_0$	$\mu\ne\mu_0$ $\mu>\mu_0$ $\mu<\mu_0$	σ^2 已知	$u=\dfrac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$	$\lvert u\rvert > u_{1-\alpha/2}$ $u > u_{1-\alpha}$ $u < u_{\alpha}$	ztest
			σ^2 未知	$t=\dfrac{\bar{X}-\mu_0}{S_*/\sqrt{n}}$	$\lvert t\rvert > t_{1-\alpha/2}(n-1)$ $t > t_{1-\alpha}(n-1)$ $t < t_{\alpha}(n-1)$	ttest
	$\sigma^2=\sigma_0^2$ $\sigma^2\le\sigma_0^2$ $\sigma^2\ge\sigma_0^2$	$\sigma^2\ne\sigma_0^2$ $\sigma^2>\sigma_0^2$ $\sigma^2<\sigma_0^2$	μ 已知	$\chi^2=\dfrac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma_0^2}$	$\chi^2 < \chi^2_{\alpha/2}(n)$ 或$\chi^2 > \chi^2_{1-\alpha/2}(n)$ $\chi^2 > \chi^2_{1-\alpha}(n)$ $\chi^2 < \chi^2_{\alpha}(n)$	无
			μ 未知	$\chi^2=\dfrac{(n-1)S_*^2}{\sigma_0^2}$	$\chi^2 < \chi^2_{\alpha/2}(n-1)$ 或$\chi^2 > \chi^2_{1-\alpha/2}(n-1)$ $\chi^2 > \chi^2_{1-\alpha}(n-1)$ $\chi^2 < \chi^2_{\alpha}(n-1)$	vartest
两个正	$\mu_1=\mu_2$ $\mu_1\le\mu_2$ $\mu_1\ge\mu_2$	$\mu_1\ne\mu_2$ $\mu_1>\mu_2$ $\mu_1<\mu_2$	σ_1^2,σ_2^2 已知	$u=\dfrac{(\bar{X}-\bar{Y})}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}$	$\lvert u\rvert > u_{1-\alpha/2}$ $u > u_{1-\alpha}$ $u < u_{\alpha}$	无

态总体			$\sigma_1^2=\sigma_2^2$ 未知	$t=\dfrac{(\bar{X}-\bar{Y})}{S_w\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$	$\lvert t\rvert > t_{1-\alpha/2}(n_1+n_2-2)$ $t > t_{1-\alpha}(n_1+n_2-2)$ $t < t_{\alpha}(n_1+n_2-2)$	ttest2
	$\sigma_1^2=\sigma_2^2$	$\sigma_1^2\neq\sigma_2^2$	μ_1,μ_2 已知	$F=\dfrac{\sum_{i=1}^{n_1}(X_i-\mu_1)^2\Big/n_1}{\sum_{j=1}^{n_2}(Y_j-\mu_2)^2\Big/n_2}$	$F < F_{\alpha/2}(n_1,n_2)$ 或$F > F_{1-\alpha/2}(n_1,n_2)$ $F > F_{1-\alpha}(n_1,n_2)$ $F < F_{\alpha}(n_1,n_2)$	无
	$\sigma_1^2\le\sigma_2^2$ $\sigma_1^2\ge\sigma_2^2$	$\sigma_1^2>\sigma_2^2$ $\sigma_1^2<\sigma_2^2$	μ_1,μ_2 未知	$F=\dfrac{S_{1*}^2}{S_{2*}^2}$	$F < F_{\alpha/2}(n_1-1,n_2-1)$ 或$F > F_{1-\alpha/2}(n_1-1,n_2-1)$ $F > F_{1-\alpha}(n_1-1,n_2-1)$ $F < F_{\alpha}(n_1-1,n_2-1)$	vartest2
配对数据检验 (P-T 检验)	$\mu_D=0$ $\mu_D\le 0$ $\mu_D\ge 0$	$\mu_D\neq 0$ $\mu_D>0$ $\mu_D<0$	σ_D^2 未知	$t=\dfrac{\bar{D}-0}{S_{D*}/\sqrt{n}}$	$\lvert t\rvert > t_{1-\alpha/2}(n-1)$ $t > t_{1-\alpha}(n-1)$ $t < t_{\alpha}(n-1)$	ttest
	n 对相互独立的观测 $(X_i,Y_i), i=1,\cdots,n$, $D_i=X_i-Y_i\sim N(\mu_D,\sigma_D^2)$.					

其中, $S_w^2=\dfrac{(n_1-1)S_{1*}^2+(n_2-1)S_{2*}^2}{n_1+n_2-2}$.

（二）参数假设检验函数

表 3-2　统计工具箱中的参数假设检验函数 (test)

	函数名称	函数说明	调用格式
正态总体的参数检验	ztest	单样本均值 μ 的 z 检验 (总体服从正态分布, σ^2 已知)	[h,p,ci,zval] = ztest(x,mu0,sigma,alpha,tail)
	ttest	单样本均值 μ 的 t 检验 (总体服从正态分布, σ^2 未知)	[h,p,ci,stats] = ttest(x,mu0,alpha,tail)
	ttest2	双样本均值差 $\mu_1-\mu_2=0$ 的 t 检验 (两个总体均服从正态分布, $\sigma_1^2=\sigma_2^2$ 未知)	[h,p,ci,stats] = ttest2(x,y,alpha,tail)
	vartest	单样本方差 σ^2 的 χ^2 检验 (总体服从正态分布, μ 未知)	[h,p,ci,stats] = vartest(x,sigma02,alpha,tail)
	vartest2	双样本方差比 $\sigma_1^2/\sigma_2^2=1$ 的 F 检验 (两个总体均服从正态分布, μ_1,μ_2 未知)	[h,p,ci,stats] = vartest2(x,y,alpha,tail)

注意:

(1) [h,p,ci,zval] = ztest(x,mu0,sigma,alpha,tail)中的 sigma 是指标准差 σ, 而非 σ^2;

(2) [h,p,ci,stats] = vartest(x,sigma02,alpha,tail)中的 sigma02 是指 σ_0^2, 而非 σ_0.

（三）参数假设检验函数的格式说明及例题

1．[h,p,ci,zval] = ztest(x,mu0,sigma,alpha,tail)

功能: 方差已知时, 对单正态总体均值 μ 与实数 μ_0 的关系进行 Z 检验. 并可通过指定 tail 的值来控制备择假设的类型. tail 的取值及表示意义如下:

tail='both' 表示 $H_1:\mu\neq\mu_0$ (缺省值) ;

tail='right'表示 $H_1:\mu>\mu_0$;

tail= 'left'表示 $H_1:\mu<\mu_0$. (原假设则为 $H_0:\mu\geq\mu_0$)

• 输出变量含义:

h——如果 h=0, 则接受 H_0; 如果 h=1, 则拒绝 H_0 而接受备择假设 H_1;

p——在当前样本下拒绝 H_0 的最小显著性水平, 即犯第 I 类错误的最小概率 $P(\text{拒绝}H_0\,|\,H_0\text{为真})$;

ci——均值 μ 的置信水平为 $1-\alpha$ 的置信区间;

zval——Z 统计量 $Z=\dfrac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$ 的观测值.

α 是我们设定的显著性水平, p 是最小显著性水平, 因此当 p<α 时, 则拒绝 H_0;

当 μ_0 落入 ci 中, 则接受 H_0.

【★例3.4(P76)】一台包装机装洗衣粉, 额定标准重量为500g, 根据以往经验, 包装机的实际装袋重量服从正态分布 $N(\mu,\sigma^2)$, 其中 $\sigma=15$ g, 为检验包装机工作是否正常, 随机抽取9袋, 称得洗衣粉净重数据如下 (单位: g) :

497　506　518　524　488　517　510　515　516

若取显著性水平 $\alpha=0.01$, 问这包装机工作是否正常?

(1) 问题分析:

设总体X——每袋洗衣粉的重量, $X\sim N(\mu,\sigma^2)$, 且 $\sigma^2=15^2$ 已知.

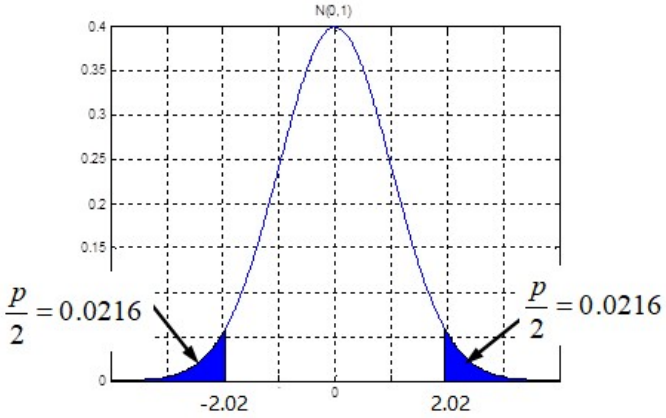
今抽得一容量为9的样本, 本问题是检验假设: $H_0:\mu=\mu_0$,　$H_1:\mu\neq\mu_0$ $(\mu_0=500)$.

(2) 问题求解:

选取检验统计量 $Z=\dfrac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$, 则 H_0 的拒绝域为 $|Z|>u_{1-\alpha/2}$.

```
>> x=[497,506,518,524,488,517,510,515,516];
>> [h,p,ci,zval]=ztest(x,500,15,0.01,'both')
    h =    0                    %接受 H0: μ = μ0 = 500
    p =    0.0432               %H0为真条件下|Z| = |(X̄ - μ0)/(σ/√n)| > u1-α/2 成立的最小
                                 的α 值 (参照书 P84 例 3.7)
    ci =  497.2320  522.9903    %σ 已知时 μ 的置信水平为 0.99 的双侧置信区间
```

zval = 2.0222　　%统计量 $Z=\dfrac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$ 的值.



(3) 问题结果:

由于 h = 0 (或 $p=0.0432>0.01=\alpha$), 故接受 $H_0:\mu=\mu_0$, 即认为包装机工作正常.

2. [h,p,ci,stats] = ttest(x,mu0,alpha,tail)

功能: 方差未知时, 对单正态总体均值进行 t 检验.

• 输出变量含义:

stats 包含三个结果:

tstat——t 统计量 $t=\dfrac{\bar{X}-\mu_0}{S_*/\sqrt{n}}$ 的值;

df——t 分布的自由度;

sd——样本标准差 S_*.

【★例 3.5(P79)】某部门对当前市场的价格情况进行调查. 以鸡蛋为例, 所抽查的全省 20 个集市上, 售价分别为 (单位: 元/500 克)

3.05, 3.31, 3.34, 3.82, 3.30, 3.16, 3.84, 3.10, 3.90, 3.18,
3.88, 3.22, 3.28, 3.34, 3.62, 3.28, 3.30, 3.22, 3.54, 3.30.

已知往年的平均售价一直稳定在 3.25 元/500 克左右, 在显著性水平 $\alpha=0.025$ 下, 能否认为全省当前的鸡蛋售价明显高于往年?

(1) 问题分析:

设总体X——每500克的鸡蛋价格, $X\sim N(\mu,\sigma^2)$, 且 σ^2 未知.

今抽得一容量为20的样本, 本问题是检验假设: $H_0:\mu=\mu_0$, $H_1:\mu>\mu_0$ $(\mu_0=3.25)$.

(2) 问题求解:

选取检验统计量 $t=\dfrac{\bar{X}-\mu_0}{S_*/\sqrt{n}}$, 则 H_0 的拒绝域为 $t>t_{1-\alpha}(n-1)$.

```
>> x=[3.05,3.31,3.34,3.82,3.30,3.16,3.84,3.10,3.90,3.18,...
      3.88,3.22,3.28,3.34,3.62,3.28,3.30,3.22,3.54,3.30];
>> [h,p,ci,stats]=ttest(x,3.25,0.025,'right')
    h =    1
    p =    0.0114
```

ci = 3.2731 Inf

stats = tstat: 2.4763 df: 19 sd: 0.2691

(3) 问题结果:

由于 h = 1, 故接受 $H_1:\mu>\mu_0$, 即认为全省当前的鸡蛋售价明显高于往年.

3．[h,p,ci,stats] = ttest2(x,y,alpha,tail)

功能: 两方差相等但未知时, 对两个正态总体均值关系进行 t 检验. 并可通过指定 tail 的值来控制备择假设的类型. tail 的取值及表示意义如下:

tail='both' 表示 $H_1:\mu_1\neq\mu_2$ (缺省值) ;

tail='right'表示 $H_1:\mu_1>\mu_2$;

tail= 'left'表示 $H_1:\mu_1<\mu_2$. (原假设则为 $H_0:\mu_1\geq\mu_2$)

• 输出变量含义:

stats 包含三个结果:

tstat——t 统计量 $t=\dfrac{(\bar{X}-\bar{Y})}{S_w\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$ 的值;

df——t 分布的自由度;

sd——两样本的合并标准差 $S_w=\sqrt{\dfrac{(n_1-1)S_{1*}^2+(n_2-1)S_{2*}^2}{n_1+n_2-2}}$.

【★例 3.6(P_{81})】某工厂生产某种电器材料. 要检验原来使用的材料与一种新研制的材料的疲劳寿命有无显著性差异, 各取若干样品, 做疲劳寿命试验, 所得数据如下 (单位: 小时) :

原材料: 40, 110, 150, 65, 90, 210, 270

新材料: 60, 150, 220, 310, 380, 350, 250, 450, 110, 175

一般认为, 材料的疲劳寿命服从对数正态分布, 并可以假定原材料疲劳寿命的对数 $\ln X$ 与新材料疲劳寿命的对数 $\ln Y$ 有相同的方差, 即可设 $\ln X\sim N(\mu_1,\sigma^2)$, $\ln Y\sim N(\mu_2,\sigma^2)$. 在显著性水平 $\alpha=0.05$ 下, 能否认为两种材料的疲劳寿命没有显著性差异?

(1) 问题分析:

设总体X——原材料的疲劳寿命, $\ln X\sim N(\mu_1,\sigma_1^2)$;

设总体Y——新材料的疲劳寿命, $\ln Y\sim N(\mu_2,\sigma_2^2)$, 且 $\sigma_1^2=\sigma_2^2=\sigma^2$ 未知.

今从两个总体中各抽得一样本, 本问题是检验假设: $H_0:\mu_1=\mu_2$, $H_1:\mu_1\neq\mu_2$.

(2) 问题求解:

选取检验统计量 $t=\dfrac{(\overline{\ln X}-\overline{\ln Y})}{S_w\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$, 则 H_0 的拒绝域为 $|t|>t_{1-\alpha/2}(n_1+n_2-2)$.

```
>> x=[40,110,150,65,90,210,270];
>> y=[60,150,220,310,380,350,250,450,110,175];
>> [h,p,ci,stats]=ttest2(log(x),log(y),0.05,'both')
```

```
h =    0
p =   0.0626
ci =  -1.3095   0.0379
stats =    tstat: -2.0116      df: 15      sd: 0.6414
```

(3) 问题结果:

由于 h = 0, 故接受 $H_0:\mu_1=\mu_2$, 即认为两种材料的疲劳寿命没有显著差异.

【例 4 (浙大四版 P187)】做以下的实验以比较人对红光或绿光的反应时间(以 s 计). 实验在点亮红光或绿光的同时, 启动计时器, 要求受试者见到红光或绿光点亮时, 就按下按钮, 切断计时器, 这就能测得反应时间. 测量的结果如下表:

红光(x)	0.30	0.23	0.41	0.53	0.24	0.36	0.38	0.51
绿光(y)	0.43	0.32	0.58	0.46	0.27	0.41	0.38	0.61

问能否认为人们对红光的反应要比绿光快? (显著性水平取 $\alpha=0.05$)

(1) 问题分析(本题是配对数据检验):

设 $D_i=X_i-Y_i\ \ (i=1,\cdots,8)$ 是来自正态总体 $N(\mu_D,\sigma_D^2)$ 的样本, μ_D,σ_D^2 均未知. 本问题是检验假设: $H_0:\mu_D\geq 0,\quad H_1:\mu_D<0$.

(2) 问题求解:

选取检验统计量 $t=\dfrac{\bar{D}-\mu_0}{S_*/\sqrt{n}}$, 则 H_0 的拒绝域为 $t<t_\alpha(n-1)$.

```
>> x=[0.30 0.23 0.41 0.53 0.24 0.36 0.38 0.51];
>> y=[0.43 0.32 0.58 0.46 0.27 0.41 0.38 0.61];
>> d=x-y;
>> [h,p,ci,stats] = ttest(d,0,0.05,'left')
    h =    1
    p =   0.0270
    ci =     -Inf   -0.0113
    stats =    tstat: -2.3113      df: 7      sd: 0.0765
```

(3) 问题结果:

由于h = 1, 故接受 $H_1:\mu_D<0$, 即认为人对红光的反应要比绿光快.

4. [h,p,ci,stats] = vartest(x,sigma02,alpha,tail)

功能: 均值未知时, 对单正态总体方差进行 χ^2 检验. 并可通过指定 tail 的值来控制备择假设的类型. tail 的取值及表示意义如下:

tail='both' 表示 $H_1:\sigma^2\neq\sigma_0^2$ (缺省值);

tail='right'表示 $H_1:\sigma^2>\sigma_0^2$;

tail= 'left'表示 $H_1:\sigma^2<\sigma_0^2$. (原假设则为 $H_0:\sigma^2\geq\sigma_0^2$)

• 输入变量含义:

sigma02——此处表示 σ_0^2 而不是 σ_0.

• 输出变量含义:

stats 包含两个结果:

chisqstat——χ^2 统计量 $\chi^2=\dfrac{(n-1)S_*^2}{\sigma_0^2}$ 的值;

df——分布的自由度.

【习题 8.11(浙大四版 P_{220})】一种混杂的小麦品种, 株高的标准差为 $\sigma_0=14$, 经提纯后随机抽取 10 株, 它们的株高(以 cm 计)为

90, 105, 101, 95, 100, 100, 101, 105, 93, 97

考察提纯后群体是否比原群体整齐? 取显著性水平 $\alpha=0.01$, 并设小麦株高服从.

(1) 问题分析:

设总体X——小麦株高, , 且未知.

今抽得一容量为10的样本, 本问题是检验假设: .

(2) 问题求解:

选取检验统计量, 则的拒绝域为.

```
>> x=[90,105,101,95,100,100,101,105,93, 97];
>> [h,p,ci,stats] = vartest(x,14^2,0.01,'left')
    h =     1
    p =  8.6935e-004
    ci =        0  104.4590
    stats =     chisqstat: 1.1128          df: 9
```

(3) 问题结果:

由于 h = 1, 故接受, 即认为提纯后群体比原群体整齐.

5．[h,p,ci,stats] = vartest2(x,y,alpha,tail)

功能: 两均值未知时, 对两个正态总体方差比进行 F 检验. 并可通过指定 tail 的值来控制备择假设的类型. tail 的取值及表示意义如下:

tail='both' 表示 $H_1:\sigma_1^2\neq\sigma_2^2$ (缺省值) ;

tail='right'表示 $H_1:\sigma_1^2>\sigma_2^2$;

tail= 'left'表示 $H_1:\sigma_1^2<\sigma_2^2$. (原假设则为 $H_0:\sigma_1^2\geq\sigma_2^2$)

• 输出变量含义:

stats 包含三个结果:

fstat——F 统计量 $F=\dfrac{S_{1*}^2}{S_{2*}^2}$ 的值;

df1——F 分布的第一个自由度;

df2——F 分布的第二个自由度.

【例 2(浙大四版 P_{185})】用两种方法(A 和 B)测定冰自 $-0.72^\circ C$ 转变为 $0^\circ C$ 的水的融化热(以 cal/g 计). 测得以下的数据:

方法 A: 79.98 80.04 80.02 80.04 80.03 80.03 80.04 79.97 80.05 80.03 80.02 80.00 80.02

方法 B: 80.02 79.94 79.98 79.97 79.97 80.03 79.95 79.97

设这两个样本相互独立, 且分别来自正态总体 $N(\mu_1,\sigma_1^2)$ 和 $N(\mu_1,\sigma_2^2)$, $\mu_1,\mu_2,\sigma_1,\sigma_2$ 均未知.

试首先检验假设

$$H_0:\sigma_1^2=\sigma_2^2,\quad H_1:\sigma_1^2\neq\sigma_2^2$$

其次, 如果接受 $H_0:\sigma_1^2=\sigma_2^2$, 则在方差齐性的基础上再检验假设 (显著性水平都取 $\alpha=0.05$)

$$H_0:\mu_1\le\mu_2,\quad H_1:\mu_1>\mu_2.$$

(1) 问题分析:

设总体X——第一种方法的融化热, $X\sim N(\mu_1,\sigma_1^2)$;

设总体Y——第二种方法的融化热, $Y\sim N(\mu_2,\sigma_2^2)$, 且 $\mu_1,\mu_2,\sigma_1,\sigma_2$ 均未知.

今从两个总体中各抽得一样本, 本问题是:

①先进行方差齐性检验: $H_0:\sigma_1^2=\sigma_2^2,\quad H_1:\sigma_1^2\neq\sigma_2^2$;

②若接受 $H_0:\sigma_1^2=\sigma_2^2$, 则再检验: $H_0:\mu_1\le\mu_2,\quad H_1:\mu_1>\mu_2$.

(2) 问题求解:

①选取检验统计量 $F=\dfrac{S_{1*}^2}{S_{2*}^2}$, 则 H_0 的拒绝域为 $F<F_{\alpha/2}(n_1-1,n_2-1)$或$F>F_{1-\alpha/2}(n_1-1,n_2-1)$.

②选取检验统计量 $t=\dfrac{(\bar{X}-\bar{Y})}{S_w\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$, 则 H_0 的拒绝域为 $t>t_{1-\alpha}(n_1+n_2-2)$.

```
>> x=[79.98 80.04 80.02 80.04 80.03 80.03 80.04 79.97 80.05 80.03 80.02 80.00 80.02];
>> y=[80.02 79.94 79.98 79.97 79.97 80.03 79.95 79.97];
>> [h,p,ci,stats] = vartest2(x,y,0.05,'both')        %方差齐性检验
    h =    0
    p =    0.3938
    ci =   0.1251   2.1053
    stats =    fstat: 0.5837    df1: 12    df2: 7
>> [h,p,ci,stats] = ttest2(x,y,0.05,'right')         %均值差检验
    h =    1
    p =    0.0013
    ci =   0.0211    Inf
    stats =    tstat: 3.4722    df: 19    sd: 0.0269
```

(3) 问题结果:

①由于h = 0, 故接受 $H_0:\sigma_1^2=\sigma_2^2$, 即认为两总体方差相等;

②由于 h = 1, 故接受 $H_1:\mu_1>\mu_2$, 即认为方法 A 比方法 B 测得的融化热要大.

二、非正态总体参数的假设检验

（一）检验问题、检验统计量及拒绝域

表 3-3　非正态总体参数的假设检验

总体	H_0	H_1	检验统计量	H_0 的否定域
(0-1)	$p = p_0$ $p \le p_0$ $p \ge p_0$	$p \ne p_0$ $p > p_0$ $p < p_0$	近似: $u = \dfrac{\sum_{i=1}^{n} X_i - np_0}{\sqrt{np_0(1-p_0)}}$	$\lvert u \rvert > u_{1-\alpha/2}$ $u > u_{1-\alpha}$ $u < u_{\alpha}$
$\pi(\lambda)$	$\lambda = \lambda_0$ $\lambda \le \lambda_0$ $\lambda \ge \lambda_0$	$\lambda \ne \lambda_0$ $\lambda > \lambda_0$ $\lambda < \lambda_0$	近似: $u = \dfrac{\sum_{i=1}^{n} X_i - n\lambda_0}{\sqrt{n\lambda_0}}$	$\lvert u \rvert > u_{1-\alpha/2}$ $u > u_{1-\alpha}$ $u < u_{\alpha}$
$\exp(\theta)$	$\theta = \theta_0$ $\theta \le \theta_0$ $\theta \ge \theta_0$	$\theta \ne \theta_0$ $\theta > \theta_0$ $\theta < \theta_0$	精确: $\chi^2 = 2\dfrac{1}{\theta_0}\sum_{i=1}^{n} X_i$	$\chi^2 < \chi^2_{\alpha/2}(2n)$ 或 $\chi^2 > \chi^2_{1-\alpha/2}(2n)$ $\chi^2 > \chi^2_{1-\alpha}(2n)$ $\chi^2 < \chi^2_{\alpha}(2n)$
			近似: $u = \dfrac{\sum_{i=1}^{n} X_i - n\theta_0}{\sqrt{n\theta_0^2}}$	$\lvert u \rvert > u_{1-\alpha/2}$ $u > u_{1-\alpha}$ $u < u_{\alpha}$

注意: 表 3-3 中的指数分布的概率密度函数为 $f(x) = \begin{cases} \dfrac{1}{\theta} e^{-\frac{x}{\theta}}, & x > 0 \\ 0, & x \le 0 \end{cases}$.

（三）例题

【★例 3.8(P_{88})】(续例 3.1)某厂家向一百货商店长期供应某种货物, 双方根据厂家的传统生产水平写出质量标准: 若次品率超过 3%, 百货商店拒收该批货物. 今有一批货物, 随机抽 43 件检验, 发现次品 2 件, 用假设检验方法, 给出该批商品的验收方案及检验结果(设 $\alpha = 0.25$).

(1)问题分析:

设总体 $X = \begin{cases} 1, & \text{本产品为次品} \\ 0, & \text{本产品为合格品} \end{cases}$, $X \sim b(1, p)$.

今抽得一容量为 43 的样本, 本问题是检验假设: $H_0: p \le p_0$, $H_1: p > p_0$ $(p_0 = 0.03)$.

(2) 问题求解:

选取检验统计量 $u = \dfrac{\sum_{i=1}^{n} X_i - np_0}{\sqrt{np_0(1-p_0)}}$, 则 H_0 的拒绝域为 $u > u_{1-\alpha}$.

• 编写命令文件example3_8.m:

```
x=[ 1,1 zeros(1,41)];
```

```
n=length(x);
p0=0.03;
alpha=0.25;
u=(sum(x)-n*p0)/sqrt(n*p0*(1-p0));
u_alpha=norminv(1-alpha, 0,1);
[u, u_alpha]
```

• 运行命令文件example3_8.m:

```
ans =    0.6347    0.6745
```

(3) 问题结果:

由于 $u=0.6347<u_{0.75}=0.6745$, 故不拒绝 H_0, 从而接收该批产品.

【⋆例 3.9(P_{89})】下列数据是某十字路口在一分钟内车辆到达的时刻(单位:秒), 假设没有 2 辆以上的车在完全同一时刻到达.

2.7, 5.5, 7.5, 11.4, 14.1, 14.4, 14.8, 15.6, 16.7, 19.6, 20.7, 23.3,
24.5, 24.7, 27.6, 29.9, 31.1, 31.8, 33.7, 36.5, 37.4, 42.2, 44.1, 44.3,
46.6, 46.9, 47.7, 50.2, 50.3, 50.6, 50.9, 52.5, 55.4, 58.4, 58.6.

根据以往统计, 在每天 8:00 - 8:01 这段时间通过该十字路口的车流量为 0.5 辆/秒, 试以这些数据(单位:秒)检验以往的结论 (设 $\alpha=0.1$).

(1)问题分析:

设总体 X ——1秒内通过此路口的车辆数, $X\sim\pi(\lambda)$.

今抽得一容量为60的样本, 本问题是检验假设: $H_0:\ \lambda=\lambda_0$, $H_1:\ \lambda\neq\lambda_0$ $(\lambda_0=0.5)$.

(2) 问题求解:

选取检验统计量 $u=\dfrac{\sum_{i=1}^{n}X_i-n\lambda_0}{\sqrt{n\lambda_0}}$, 则 H_0 的拒绝域为 $|u|>u_{1-\alpha/2}$.

• 编写命令文件example3_9.m:

```
x=[0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 3, 1, 1, 0, 0, 1, ...
   1, 0, 0, 1, 2, 0, 0, 1, 0, 1, 0, 2, 0, 1, 0, 0, 1, 1, 0, 0,...
   0, 0, 1, 0, 2, 0, 2, 1, 0, 0, 4, 0, 1, 0, 0, 1, 0, 0, 2, 0];
n=length(x);
lamda0=0.5;
alpha=0.1;
u=(sum(x)-n*lamda0)/sqrt(n*lamda0);
u_alpha=norminv(1-alpha/2, 0,1);
[u, u_alpha]
```

• 运行命令文件example3_9.m:

```
ans =    0.9129    1.6449
```

(3) 问题结果:

由于 $|u|=0.9129<u_{0.95}=1.6449$, 故不拒绝 H_0, 即认为车流量为 0.5 辆/秒.

【★例 3.10(P_{90} 续例 3.9)】下列数据是某十字路口在一分钟内车辆到达的时刻(单位:秒), 假设没有 2 辆以上的车在完全同一时刻到达.

2.7, 5.5, 7.5, 11.4, 14.1, 14.4, 14.8, 15.6, 16.7, 19.6, 20.7, 23.3,
24.5, 24.7, 27.6, 29.9, 31.1, 31.8, 33.7, 36.5, 37.4, 42.2, 44.1, 44.3,
46.6, 46.9, 47.7, 50.2, 50.3, 50.6, 50.9, 52.5, 55.4, 58.4, 58.6.

一般地每相邻两辆车到达路口的时间间隔服从指数分布且相互独立, 其中 λ 为平均每秒的车流量, 用例 3.9 的数据及指数总体参数检验方法检验是否有 $\lambda=\lambda_0=0.5$ ($\alpha=0.1$).

(1)问题分析:

设总体 Y ——两车间隔时间, $Y\sim\exp(\theta)$, 显然 $\theta=\dfrac{1}{\lambda}$.

今抽得一容量为34的样本, 本问题是检验假设: $H_0:\ \theta=\theta_0,\quad H_1:\ \theta\neq\theta_0\quad(\theta_0=2)$.

(2) 问题求解:

(精确法) 选取检验统计量 $\chi^2=2\dfrac{1}{\theta_0}\sum_{i=1}^{n}Y_i$, 则 H_0 的拒绝域为 $\chi^2<\chi^2_{\alpha/2}(2n)$ 或 $\chi^2>\chi^2_{1-\alpha/2}(2n)$.

(近似法) 选取检验统计量 $u=\dfrac{\sum_{i=1}^{n}Y_i-n\theta_0}{\sqrt{n\theta_0^2}}$, 则 H_0 的拒绝域为 $|u|>u_{1-\alpha/2}$.

• 编写命令文件example3_10.m:

```
T=[2.7, 5.5, 7.5, 11.4, 14.1, 14.4, 14.8, 15.6, 16.7, 19.6, 20.7, 23.3,...
   24.5, 24.7, 27.6, 29.9, 31.1, 31.8, 33.7, 36.5, 37.4, 42.2, 44.1, 44.3,...
   46.6, 46.9, 47.7, 50.2, 50.3, 50.6, 50.9, 52.5, 55.4, 58.4, 58.6];
y=diff(T)
n=length(y);
theta0=2;
alpha=0.1;
%(1)精确法
chi2=2/theta0*sum(y)
chi2_cv=chi2inv([alpha/2, 1-alpha/2], 2*n)
%(2)近似法
u=(sum(y)-n*theta0)/sqrt(n*theta0^2);
u_alpha=norminv(1-alpha/2, 0,1);
[u, u_alpha]
```

• 运行命令文件example3_10.m:

(精确法) ans = 55.9000 50.0202 88.2502

(近似法) ans = -1.0376 1.6449

(3) 问题结果:

(精确法) 由于 $\chi^2=55.9000$, $\chi^2_{\alpha/2}(2n)=50.0202<\chi^2=55.9000<\chi^2_{1-\alpha/2}(2n)=88.2502$, 故不拒绝 H_0, 即认为车流量为 0.5 辆/秒.

(近似法) 由于$|u|=1.0376<u_{0.95}=1.6445$, 故不拒绝$H_0$, 即认为车流量为 0.5 辆/秒.

三、非参数假设检验

（一）各检验方法的功能与优缺点

表 3-4　常用非参数假设检验

检验法	功能	总体 X 类型	总体参数	优点	缺点
正态概率纸检验	判断单总体 $X\sim N(\mu,\sigma^2)$ (函数: normplot)	一维, 连续型	未知	快速粗略地估计出总体的某些数字特征	定性分析而非定量分析
皮尔逊 χ^2 拟合检验	检验单总体 $X\sim F_0(x;\theta)$ (函数: chi2gof)	一维或多维, 离散或连续型	已知或未知	利用实际频数与理论频数的差异构造检验统计量. 可以用于全样本, 也可用于截尾样本, 还可用于成群数据.	由于分组处理样本的观测值, 因而很容易犯第 II 类错误.
	检验两个总体的独立性 (函数: crosstab)	离散型	未知		
柯尔莫哥洛夫检验	检验单总体 $X\sim F_0(x;\theta)$ (函数: kstest)	一维, 连续型	已知	利用经验分布函数与理论分布函数的差异构造检验统计量. 与 Pearson χ^2 检验相比, 当总体为一维且理论分布 $F_0(x;\theta)$ 完全已知时, 柯尔莫哥洛夫检验优于 χ^2 检验.	柯尔莫哥洛夫检验的适用范围不如 χ^2 检验广. 特别当理论分布含未知参数时, 该检验难度较大, 目前只对正态分布、指数分布和 I 型极值分布给出了结果.
	检验单总体 $X\sim F_0(x;\theta)$ (函数: lillietest)	正态、指数和 I 型极值	未知		
偏度峰度检验	判断单总体 $X\sim N(\mu,\sigma^2)$ (函数: jbtest)	一维, 连续型	未知	适用于 $n\geq 100$	
斯米尔诺夫检验	检验两个总体是否同分布 (函数: kstest2)	一维, 连续型	未知	采用与柯尔莫哥洛夫检验类似的方法, 借助经验分布函数构造检验统计量	
秩和检验	检验两个总体是否同分布 (函数: ranksum)	一维, 连续型	未知	利用样本混合排序后的秩和构造检验统计量, 方法简单	只利用了样本数据的排序, 而没有利用样本数据本身
W 检验	判断单总体 $X\sim N(\mu,\sigma^2)$ (无已知函数)	一维, 连续型	未知	适用于小样本($3\leq n\leq 50$)	
D 检验	判断单总体 $X\sim N(\mu,\sigma^2)$ (无已知函数)	一维, 连续型	未知	适用于大样本 ($50\leq n\leq 1000$)	

（二）非参数假设检验函数

表 3-5　统计工具箱中的非假设检验函数 (test)

	函数名称	函数说明	调用格式
数值检验	chi2gof	皮尔逊 χ^2 拟合检验 (H_0: 样本来自指定分布)	[h,p,stats] = chi2gof(x,name1,val1,name2,val2,...)
	crosstab	皮尔逊 χ^2 拟合检验 (H_0: X 与 Y 独立)	[table,chi2,p,label] = crosstab(col1,col2)
	kstest	单样本分布的 Kolmogorov-Smirnov 检验 (H_0: 样本来自'cdf '指定的连续分布)	[h,p,ksstat,cv] = kstest(x,cdf,alpha,tail)
	lillietest	单样本正态分布、指数分布、极值分布的 Lilliefors 检验 (H_0: 样本来自'distr'指定的分布)	[h,p,lstat,cv]= lillietest(x,alpha,distr)
	jbtest	单样本正态分布的偏度峰度检验 (H_0: 样本来自正态分布) ($n \geq 100$)	[h,p,jbstat,cv]=jbtest(x,alpha)
	kstest2	双样本同分布的 Kolmogorov-Smirnov 检验 (H_0: 两样本来自同一连续分布)	[h,p,ks2stat] = kstest2(x,y,alpha,tail)
	ranksum	双样本同分布的 Wilcoxon 秩和检验 (H_0: 两样本来自同一连续分布)	[p,h,stats] = ranksum(x,y,alpha)
绘图检验	normplot	单样本正态分布概率纸检验 (H_0: 样本来自正态分布)	normplot(x)
	qqplot	画双样本同分布检验的分位数—分位数图 (简称 qq 图) (H_0: 两样本来自同一分布)	qqplot(x,y)

（三）非参数假设检验函数的格式说明及例题

1．[h,p,stats] = chi2gof(x,name1,val1,name2,val2,...)

功能: 用来作皮尔逊 χ^2 拟合检验(Chi-square goodness-of-fit test), 检验样本 x 是否服从指定的分布. 通过可选的成对出现的参数名(如 name1)与参数值(如 val1)来控制区间初始划分、原假设中的理论分布、显著性水平等. 各参数名与参数值如表 3-6、表 3-7 与表 3-8 所列.

• 输入变量含义如(1)(2)(3)所示:

(1) 划分初始区间并决定观测频数

表 3-6 chi2gof 函数控制区间划分的参数与参数值列表

参数名	参数值	说明
'nbins'	正整数, 默认值为 10	分组(或区间)个数
'ctrs'	向量	指定各区间的中点

'edges'	向量	指定各区间的边界

注意: 表 3-6 中 3 个参数不能同时指定, 一次最多只能调用其中的一个参数.

(2) 指定理论分布并决定理论频数

表 3-7 chi2gof 函数控制原假设理论分布的参数与参数值列表

参数名	参数值	说明
'cdf'	形如{'累积分布函数', 参数 1 值, ...}的元胞数组	指定原假设中的分布. '累积分布函数'可选'normcdf'、'poisscdf'等
'expected'	向量	指定各区间的理论频数
'nparams'	正整数	指定分布中待估参数的个数, 它确定了 χ^2 分布的自由度

注意: 表 3-7 中'cdf'与'expected'选项不能同时出现. 如果缺失参数项'nparams', 则认为未知参数个数为'cdf'或'expected'分布类型本身所含参数个数(如正态分布含两个参数); 如果分布中参数已知, 则必须申明'nparams'值为 0.

(3) 其它参数

表 3-8 chi2gof 函数控制检验的其它参数与参数值列表

参数名	参数值	说明
'emin'	非负整数, 默认值为 5	指定各区间对应的最小理论频数. 初始分组中, 理论频数小于这个值的区间将和相邻区间合并. 如果指定为 0, 将不进行区间合并 (比如对离散有限型分布, 不必合并区间).
'frequency'	与 x 等长的向量	指定 x 中各元素出现的频数, 此处 x 不是样本(见 help)
'alpha'	(0,1)上的实数, 默认值为 0.05	指定检验的显著性水平 α

• 输出变量含义:

stats 包含五个结果:

chi2stat—— χ^2 统计量 $\chi^2=\sum_{i=1}^{k}\frac{(n_i-n\hat{p}_i)^2}{n\hat{p}_i}$ 的值;

df—— χ^2 分布的自由度(k-m-1);

edges——合并后区间边界;

O——每个区间的观测频数;

E——每个区间的理论频数.

【★例 3.15(P_{104})】某厂宣称自己产品的合格率为 99%, 检验人员从该厂的一批产品中抽查了 100 件, 发现有两件次品. 在 $\alpha=0.1$ 下, 能否据此断定该厂谎报合格率.

(1) 问题分析:

设 $X=\begin{cases}0, \text{次品}\\ 1, \text{合格品}\end{cases}$, 本问题是检验X是否服从 $b(1,p)=b(1,0.99)$ 的分布, 即检验假设:

$$H_0: P(X=0)=0.01, \quad P(X=1)=0.99.$$

(2) 问题求解:

本总体变量只有两个取值, 即 $k=2$. 由于参数 $p=0.99$ 已知, 故选取检验统计量 $\chi^2=\sum_{i=1}^{2}\frac{(n_i-np_i)^2}{np_i}$, 则 H_0 的拒绝域为 $\chi^2>\chi^2_{1-\alpha}(k-1)=\chi^2_{1-\alpha}(1)$.

• 编写命令文件 example3_15.m:

```
x=[ 0,0,ones(1,98)];
n=length(x);
%法一: 用'expected'指定理论频数
[h,p,stats] = chi2gof(x,'ctrs',[0,1],'expected',n*[0.01,0.99],'nparams',0,'emin',0,'alpha',0.1) %注意'ctrs'与'expected'值的顺序必须一致
%法二: 用'cdf'指定理论频数
[h,p,stats] = chi2gof(x,'ctrs',[0,1],'cdf',{'binocdf',1,0.99},'nparams',0,'emin',0,'alpha',0.1)
```

• 运行命令文件example3_15.m (两种方法的结果一样):

```
>> example3_15
    h =     0
    p =    0.3149
    stats =     chi2stat: 1.0101
                df: 1
                edges: [-0.5000 0.5000 1.5000]
                O: [2 98]
                E: [1 99]
```

(3) 问题结果:

由于 h = 0, 故接受 H0, 即不能断定该厂谎报合格率.

【★例 3.16(P_{106})】1991 年某校工科研究生有 60 名以数理统计作为学位课, 考试成绩如下:

93	75	83	93	91	85	84	82	77	76	77	95	94	89	91
88	86	83	96	81	79	97	78	75	67	69	68	84	83	81
75	66	85	70	94	84	83	82	80	78	74	73	76	70	86
76	90	89	71	66	86	73	80	94	79	78	77	63	53	55

试问考试成绩是否服从正态分布 ($\alpha=0.1$).

(1) 问题分析:

设X——数理统计考试成绩, 本问题是检验假设:

$$H_0: X\sim N(\mu,\sigma^2).$$

(2) 问题求解:

本总体是连续型随机变量, 故需划分初始区间. 又由于参数 μ,σ^2 未知, 因此需要先求得

$\hat{\mu}_{MLE}=\bar{X}$, $\hat{\sigma}_{MLE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2}$, 再选取检验统计量 $\chi^2=\sum_{i=1}^{2}\frac{(n_i-n\hat{p}_i)^2}{n\hat{p}_i}$, 则 H_0 的拒绝域为 $\chi^2>\chi^2_{1-\alpha}(k-m-1)$.

• 编写命令文件 example3_16.m:

```
x=[93  75  83  93  91  85  84  82  77  76  77  95  94  89  91....
   88  86  83  96  81  79  97  78  75  67  69  68  84  83  81....
   75  66  85  70  94  84  83  82  80  78  74  73  76  70  86....
   76  90  89  71  66  86  73  80  94  79  78  77  63  53  55];
mu_hat=mean(x)
sigma_hat=std(x,1)
[h,p,stats] = chi2gof(x,'edges',[-Inf,60,70,80,90,Inf],
            'cdf',{'normcdf',mu_hat,sigma_hat},'nparams',2,'alpha',0.1)
```

• 运行命令文件example3_16.m:

```
>> example3_16
    h =     0
    p =    0.4760
    stats =     chi2stat: 0.5081
                df: 1
                edges: [-Inf 70 80 90 Inf]
                O: [8 20 21 11]
                E: [8.8270  20.9245  21.1317  9.1169]
```

(3) 问题结果:

由于 h = 0, 故接受 H0, 即认为学生成绩服从正态分布.

2．[table,chi2,p,label] = crosstab(x,y)

功能: 列联表的独立性检验. x 表示总体 X 的观测值向量, y 表示总体 Y 的观测值向量.

• 输出变量含义:

table——由(x, y)整理得到的列联表;

chi2——皮尔逊 χ^2 统计量 $\chi^2=n\left(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}}-1\right)$ 的值;

p——p 值, 即拒绝 H_0 的最小显著性水平;

label——第一列表示总体 X 的取值, 第二列表示总体 Y 的取值.

【★例 3.17(P_{110})】某研究所推出一种感冒特效新药,为证明其疗效, 选择 200 名患者为志愿者. 将他们等分为两组, 分别不服药或服药, 观察三日后痊愈的情况, 得出下列数据:

是否服药 \ 是否痊愈	未痊愈者数	痊愈者数	合计
未服药者数	52	48	100
服药者数	44	56	100
合计	96	104	200

问新药是否疗效明显? ($\alpha = 0.25$)

(1) 问题分析:

设 $X = \begin{cases} 0, \text{未服药} \\ 1, \text{服药} \end{cases}$, $Y = \begin{cases} 0, \text{未痊愈} \\ 1, \text{痊愈} \end{cases}$. 则本问题是检验服药与否与是否痊愈是否独立, 即检验假设:

$$H_0 : P(X = i, Y = j) = P(X = i)P(Y = j), \quad i, j = 0,1.$$

(2) 问题求解:

本问题是列联表的独立性检验, 故选取检验统计量 $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - n\hat{p}_{i\cdot}\hat{p}_{\cdot j})^2}{n\hat{p}_{i\cdot}\hat{p}_{\cdot j}}$, 其中 $\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n}$, $\hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}$, 则 H_0 的拒绝域为 $\chi^2 > \chi^2_{1-\alpha}\left((r-1)(s-1)\right)$, 其中r、s分别为X与Y的可能取值个数 .

• 编写命令文件example3_17.m:

```
%法一: 如果列联表已知, 自己编写程序检验独立性
display('法一: 自己编程的结果--------------------------')
nij=[52,48;44,56];
[r,s]=size(nij);
nj=sum(nij,1);              %各列求和, 得行向量
ni=sum(nij,2);              %各行求和, 得列向量
n=sum(ni);                  %所有数据和
chi2=n*(sum(sum(nij.^2./(ni*nj)))-1)          %利用矩阵的乘法与点除
chi2_alpha=chi2inv(0.75,(r-1)*(s-1))          %拒绝域的临界点
if chi2>chi2_alpha
   h=1
else
   h=0
end
%法二: 如果列联表未知, 调用 crosstab 函数得列联表、及独立性检验
display('法二: crosstab 函数的结果--------------------------')
x=[zeros(1,52),zeros(1,48),ones(1,44),ones(1,56)];      %本题是由列联表反找出两个总体的样本观测值
y=[zeros(1,52),ones(1,48),zeros(1,44),ones(1,56)];
[table,chi2,p,label] = crosstab(x,y)
```

• 运行命令文件example3_17.m:

```
>> example3_17
    法一: 自己编程的结果--------------------------
    chi2 =    1.2821
    chi2_alpha =    1.3233
```

```
h =  0
法二: crosstab 函数的结果--------------------------
table =    52    48
           44    56
chi2 =    1.2821
p =    0.2575
label =     '0'    '0'
            '1'    '1'
```

(3) 问题结果:

法一中 h = 0, 法二中 p=0.2575>0.25, 故两种方法均接受 H0, 即认为这种感冒新药并无明显疗效.

3．[h,p,ksstat,cv] = kstest(x,cdf,alpha,tail)

功能: 对假设“H_0：样本 x 的总体服从由 cdf 指定的分布 $F_0(x)$”进行显著水平为 α 的 Kolmogorov-Smirnov 检验. 其中 cdf 是一个两列的矩阵, 第一列为 x 值, 第二列为相应的理论累积分布函数值 $F_0(x)$.并

可通过指定 tail 的值来控制备择假设的类型. tail 的取值及表示意义如下:

tail='unequal' 表示 $H_1: F(x) \neq F_0(x)$ (缺省值);

tail='larger'表示 $H_1: F(x) > F_0(x)$;

tail= 'smaller'表示 $H_1: F(x) < F_0(x)$. (原假设则为 $H_0: F(x) \geq F_0(x)$)

注意: (1) cdf 一定是一个两列矩阵, 且分布中参数已知;

(2) 如果 cdf=[ ], kstest( )将使用标准正态分布.

• 输出变量含义:

cv——判断是否拒绝 H_0 的临界值.

【★例 3.18(P_{114})】对一台设备进行寿命试验, 记录 10 次无故障工作时间, 并从小到大排列得

420, 500, 920, 1380, 1510, 1650, 1760, 2100, 2300, 2350

问此设备的无故障工作时间 X 的分布是否服从 $\theta = 1500$ 的指数分布 ($\alpha = 0.05$)?

(1) 问题分析:

设总体X—— 一台设备无故障工作时间.

今抽得一容量为 10 的样本, 本问题是检验假设: $H_0: X \sim \exp(1500)$, 即

$$f(x) = \begin{cases} \dfrac{1}{1500} e^{-\frac{x}{1500}}, & x > 0 \\ 0, & x \leq 0 \end{cases}.$$

(2) 问题求解:

选取检验统计量 $D_n = \sup\limits_{-\infty < x < +\infty} |F_n(x) - F_0(x)|$, 则 H_0 的拒绝域为 $D_n > D_{n,\alpha}$.

```
>> x=[420, 500, 920, 1380, 1510, 1650, 1760, 2100,2300, 2350];
>> [h,p,ksstat,cv]=kstest(x,[x',expcdf(x',1500)],0.05,'unequal')
    h =     0
```

p = 0.2654
ksstat = 0.3015
cv = 0.4093

(3) 问题结果:

由于 h = 0, 故接受 H0, 即认为此设备的无故障工作时间 X 服从$\theta=1500$的指数分布.

4．[h,p,lstat,cv]= lillietest(x,alpha,distr)

功能: 用于检验样本 x 是否服从指定的分布 distr, 这里分布的参数都是未知的, lillitest 会自己根据样本作出估计, 然后使用 Kolmogorov-Smirnov 检验法进行检验. “指定的分布”包含正态分布、指数分布和 I 型极值分布.

【⋆例 3.19(P_{116})】对 8 个产品进行强度试验, 所得强度数据取自然对数后为

0.25, 0.53, 0.88, 1.22, 1.76, 2.44, 3.41, 4.90

问这批强度数据是否来自对数正态分布($\alpha=0.2$)?

(1) 问题分析:

设总体Y——产品的强度.

今抽得一容量为8的样本, 本问题是检验假设: $H_0: X=\ln Y \sim N(\mu,\sigma^2)$.

(2) 问题求解:

选取检验统计量 $\hat{D}_n=\sup\limits_{-\infty<x<+\infty}\left|F_n(x)-\hat{F}_0(x)\right|$, 则 H_0 的拒绝域为 $\hat{D}_n>\hat{D}_{n,\alpha}$.

```
>> x=[0.25, 0.53, 0.88, 1.22, 1.76, 2.44, 3.41, 4.90];
>> [h,p,lstat,cv]=lillietest(x,0.2,'norm')
    h =    0
    p =    0.5000
    lstat =    0.1710
    cv =    0.2387
```

(3) 问题结果:

由于 h = 0, 故接受 H0, 即认为这批强度数据来自对数正态分布.

5．[h,p,jbstat,cv] = jbtest(x,alpha)

功能: 对假设“H_0: 样本 x 的总体服从正态分布 (未指定均值和方差) ”进行显著水平为α的 Jarque-Bera 检验. 此检验基于 x 的偏度与峰度. 对于真实的正态分布, 样本偏度应接近于 0, 样本峰度应接近于 3. Jarque-Bera 检验通过 χ^2 统计量来判定样本偏度和峰度是否与它们的期望值显著不同.

• 输出变量含义:

jbstat——检验统计量 $\chi^2=n\left(\frac{J^2}{6}+\frac{B^2}{24}\right)$的值,

其中 $J=\frac{1}{n}\sum\limits_{i=1}^{n}\left(\frac{X_i-\bar{X}}{S}\right)^3$ 为样本偏度, $B=\frac{1}{n}\sum\limits_{i=1}^{n}\left(\frac{X_i-\bar{X}}{S}\right)^4-3$ 为样本峰度;

cv——判断是否拒绝 H_0 的临界值.

【⋆例 3.20(P117)】在 20 天内, 从维尼纶正常生产时生产报表上看到的维尼纶纤度 (表示纤维粗细程度的一个量) 的情况, 有如下 100 个数据:

1.36,	1.49,	1.43,	1.41,	1.37,	1.40,	1.32,	1.42,	1.47,	1.39
1.41,	1.36,	1.40,	1.34,	1.42,	1.42,	1.45,	1.35,	1.42,	1.39
1.44,	1.42,	1.39,	1.42,	1.42,	1.30,	1.34,	1.42,	1.37,	1.36
1.37,	1.34,	1.37,	1.37,	1.44,	1.45,	1.32,	1.48,	1.40,	1.45
1.39,	1.46,	1.39,	1.53,	1.36,	1.48,	1.40,	1.39,	1.38,	1.40
1.36,	1.45,	1.50,	1.43,	1.38,	1.43,	1.41,	1.48,	1.39,	1.45
1.37,	1.37,	1.39,	1.45,	1.31,	1.41,	1.44,	1.44,	1.42,	1.47
1.35,	1.36,	1.39,	1.40,	1.38,	1.35,	1.42,	1.43,	1.42,	1.42
1.42,	1.40,	1.41,	1.37,	1.46,	1.36,	1.37,	1.27,	1.37,	1.38
1.42,	1.34,	1.43,	1.42,	1.41,	1.41,	1.44,	1.48,	1.55,	1.37

要求在显著性水平 $\alpha = 0.01$ 下检验假设

$$H_0 : F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right); \quad H_1 : F(x) \neq \Phi\left(\frac{x-\mu}{\sigma}\right)$$

其中 F(x)为纤度的分布函数, Φ(·) 为标准正态分布函数.

(1) 问题分析:

设总体X——维尼纶纤度.

今抽得一容量为100的样本, 本问题是检验假设: $H_0 : X \sim N(\mu, \sigma^2)$.

(2) 问题求解:

①(皮尔逊 χ^2 拟合检验) 选取检验统计量 $\chi^2 = \sum_{i=1}^{2} \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i}$, 其中未知参数 μ, σ^2 的估计量分别是 $\hat{\mu}_{MLE} = \bar{X}$ 与 $\hat{\sigma}_{MLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$, 则 H_0 的拒绝域为 $\chi^2 > \chi^2_{1-\alpha}(k-m-1)$.

②(Lilliefors检验) 选取检验统计量 $\hat{D}_n = \sup_{-\infty<x<+\infty} \left|F_n(x) - F_0(x; \hat{\mu}, \hat{\sigma}^2)\right|$, 则 H_0 的拒绝域为 $\hat{D}_n > \hat{D}_{n,\alpha}$.

③(偏度峰度检验) 选取检验统计量 $\chi^2 = n\left(\frac{J^2}{6} + \frac{B^2}{24}\right)$, 则 H_0 的拒绝域为 $\chi^2 > \chi^2_{1-\alpha}(2)$.

• 编写命令文件 example3_20.m:

```
x=[1.36,  1.49,  1.43,  1.41,  1.37,  1.40,  1.32,  1.42,  1.47,  1.39...
   1.41,  1.36,  1.40,  1.34,  1.42,  1.42,  1.45,  1.35,  1.42,  1.39...
   1.44,  1.42,  1.39,  1.42,  1.42,  1.30,  1.34,  1.42,  1.37,  1.36...
   1.37,  1.34,  1.37,  1.37,  1.44,  1.45,  1.32,  1.48,  1.40,  1.45...
   1.39,  1.46,  1.39,  1.53,  1.36,  1.48,  1.40,  1.39,  1.38,  1.40...
   1.36,  1.45,  1.50,  1.43,  1.38,  1.43,  1.41,  1.48,  1.39,  1.45...
   1.37,  1.37,  1.39,  1.45,  1.31,  1.41,  1.44,  1.44,  1.42,  1.47...
   1.35,  1.36,  1.39,  1.40,  1.38,  1.35,  1.42,  1.43,  1.42,  1.42...
   1.42,  1.40,  1.41,  1.37,  1.46,  1.36,  1.37,  1.27,  1.37,  1.38...
```

```
    1.42,  1.34,  1.43,  1.42,  1.41,  1.41,  1.44,  1.48,  1.55,  1.37];
%法一(皮尔逊卡方检验)
mu=mean(x);
sigma=std(x,1);        %此处mu, sigma用的正态分布的最大似然估计
[h,p,stats] = chi2gof(x,'cdf',{'normcdf',mu,sigma},'nparams',2,'alpha',0.01)
%法二 (Lilliefors检验)
 [h,p,lstat,cv]= lillietest(x,0.01,'norm')          %lillietest.m中的临界值非常精确
%法三(偏度峰度检验)
 [h,p,jbstat,cv]= jbtest(x,0.01)                    %jbtest.m中的临界值非常精确
```

• 运行命令文件example3_20.m:

```
>> example3_20
①(皮尔逊卡方检验)
    h =    0
    p =   0.6146
    stats =    chi2stat: 1.8016
               df: 3
               edges: [1.2700 1.3540 1.3820 1.4100 1.4380 1.4660 1.5500]
               O: [12 22 23 20 13 10]
               E: [14.5405 17.4792 22.8373 21.2965 14.1741 9.6724]
②(Lilliefors 检验)
    h =    0
    p =   0.0434
    lstat =   0.0904
    cv =   0.1037
③(偏度峰度检验)
    h =    0
    p =   0.2847
    jbstat =   1.9681
    cv =   12.5067
```

(3) 问题结果:

由于三种方法都是 h = 0, 故不论是皮尔逊 χ^2 检验, Lilliefors 检验, 还是偏度峰度检验, 在显著性水平 $\alpha = 0.01$ 下, 都接受 H0, 即认为维尼纶纤度服从正态分布.

6．[h,p,ksstat] = kstest2(x,y,alpha,tail)

功能: 对假设“H_0: 两个样本 x 与 y 来自同一连续分布”进行显著水平为 α 的 Kolmogorov-Smirnov 检验. 并可通过指定 tail 的值来控制备择假设的类型. tail 的取值及表示意义如下:

tail='unequal' 表示 $H_1 : F_1(x) \neq F_2(x)$ (缺省值) ;

tail='larger'表示 $H_1 : F_1(x) > F_2(x)$;

tail= 'smaller'表示 $H_1: F_1(x) < F_2(x)$. (原假设则为 $H_0: F_1(x) \geq F_2(x)$)

注意: 对于大容量的样本来说, p-值将很精确, 一般来说, 当两样本容量满足 $(n_1 * n_2)/(n_1 + n_2) \geq 4$ 时, p-值即可认为是精确的.

7. [p,h,stats] = ranksum(x,y,alpha)

功能: 对假设“H0: 两个样本 x 与 y 来自同一连续分布”进行显著水平为 α 的 Wilcoxon 秩和检验.

注意: 此函数输出变量的第一个变量是 p 而非 h, 这是与表 3-4 中其它函数所不同的地方.

【★例 3.25(书 P128)】以下是两个地区所种小麦的蛋白质含量检验数据:

地区 1: 12.6, 13.4, 11.9, 12.8, 13.0

地区 2: 13.1, 13.4, 12.8, 13.5, 13.3, 12.7, 12.4

问两地区小麦的蛋白质含量有无显著性差异 ($\alpha = 0.05$)?

(1) 问题分析:

设总体X——第一个地区小麦的蛋白质含量;

总体Y——第二个地区小麦的蛋白质含量;

今从X与Y中分别抽得容量为5与7的两个样本, 本问题是检验假设: $H_0: F(x) = G(x)$.

(2) 问题求解:

①(Smirnov检验－斯米尔诺夫检验)选取检验统计量 $D_{n_1,n_2} = \sup_{-\infty<x<+\infty} \left|F_{n_1}(x) - G_{n_2}(x)\right|$, 则 H_0 的拒绝域为 $D_{n_1,n_2} > D_{n,\alpha}$.

②(Wilcoxon秩和检验)选取检验统计量 $T = \begin{cases} T_1, & n_1 \leq n_2 \\ T_2, & n_1 > n_2 \end{cases}$, 则 H_0 的拒绝域为 $T < T_\alpha^{(1)}$ 或 $T > T_\alpha^{(2)}$.

解: 本问题分别采用斯米尔诺夫检验与秩和检验来检验假设: $H_0: F_1(x) = F_2(x)$.

```
>> x=[12.6,13.4,11.9,12.8,13.0];
>> y=[13.1,13.4,12.8,13.5,13.3,12.7,12.4];
>> [h,p,ksstat] = kstest2(x,y,0.05)          %斯米尔诺夫检验
     h =    0
     p =    0.7065
     ksstat =    0.3714
>> [p,h,stats] = ranksum(x,y,0.05)           %秩和检验
     p =    0.4066
     h =    0
     stats =    ranksum: 27
```

(3) 问题结果:

由于两种方法都有 h = 0, 故不论是斯米尔诺夫检验还是秩和检验, 都接受 H0, 即认为两地区小麦的蛋白质含量无显著差异.

8. normplot(x)

功能: 绘出 x 中数据的正态检验概率图. 如果 x 是一个矩阵, 则对每一列绘出一条线. 图中样本数据用符号‘＋’来表示, 叠加在数据上的实线是数据的第一个与第三个四分位点之间的连线. 如果数据是来自一个正态分布, 则‘＋’线近似地在一直线上.

注意: 中间的点离直线位置的偏差不能过大, 两头的点的偏差可以允许大一些.

【补例 3-1】试用正态概率纸、Lilliefors 检验及偏度峰度检验分别检验下面这批数据是否来自正态分布 ($\alpha = 0.05$) .

0.5200	0.9703	1.0336	1.0767	1.1583	1.4525	1.5492	1.5788	1.7976	1.8729
2.0430	2.1025	2.2447	2.4279	2.5194	2.5414	2.5825	2.7086	2.7243	2.7410
2.8060	2.8803	2.9088	2.9124	2.9664	3.0561	3.1853	3.2789	3.3776	3.5871
3.6029	3.6094	3.6480	3.8045	3.8330	3.8334	4.4338	4.4844	4.7712	4.8263
4.8879	5.2141	5.2460	5.2908	5.3068	5.3151	5.3991	5.4015	5.5083	5.6991
5.7038	5.7758	5.8448	5.8800	5.9195	5.9206	5.9511	6.1380	6.1872	6.3015
6.3183	6.3209	6.3515	6.4242	6.5356	6.5763	6.6056	6.6876	6.7917	6.8239
7.0577	7.1367	7.1706	7.2543	7.2703	7.8775	7.8989	7.9055	8.1488	8.2000
8.2126	8.2694	8.2851	8.4294	8.4497	8.5119	8.6095	8.6570	8.7644	8.8447
8.9524	9.1165	9.4678	9.6167	10.2459	10.4512	10.6135	10.8366	11.5103	11.8165

(1) 问题分析:

今抽得来自总体X的一容量为100的样本, 本问题是检验假设: $H_0 : X \sim N(\mu, \sigma^2)$.

(2) 问题求解:

①(正态概率纸检验) 如果数据来自正态分布, 则由样本给出的‘＋’近似地在一直线上.

②(Lilliefors检验) 选取检验统计量 $\hat{D}_n = \sup_{-\infty<x<+\infty} \left| F_n(x) - F_0(x; \hat{\mu}, \hat{\sigma}^2) \right|$, 则 H_0 的拒绝域为 $\hat{D}_n > \hat{D}_{n,\alpha}$.

③(偏度峰度检验) 选取检验统计量 $\chi^2 = n\left(\frac{J^2}{6} + \frac{B^2}{24}\right)$, 则 H_0 的拒绝域为 $\chi^2 > \chi^2_{1-\alpha}(2)$.

• 编写命令文件 addexample3_1.m:

```
%几种正态分布检验方法的比较
x=[0.5200  0.9703  1.0336  1.0767  1.1583  1.4525  1.5492  1.5788  1.7976  1.8729...
2.0430  2.1025  2.2447  2.4279  2.5194  2.5414  2.5825  2.7086  2.7243  2.7410...
2.8060  2.8803  2.9088  2.9124  2.9664  3.0561  3.1853  3.2789  3.3776  3.5871...
3.6029  3.6094  3.6480  3.8045  3.8330  3.8334  4.4338  4.4844  4.7712  4.8263...
4.8879  5.2141  5.2460  5.2908  5.3068  5.3151  5.3991  5.4015  5.5083  5.6991...
5.7038  5.7758  5.8448  5.8800  5.9195  5.9206  5.9511  6.1380  6.1872  6.3015...
6.3183  6.3209  6.3515  6.4242  6.5356  6.5763  6.6056  6.6876  6.7917  6.8239...
7.0577  7.1367  7.1706  7.2543  7.2703  7.8775  7.8989  7.9055  8.1488  8.2000...
8.2126  8.2694  8.2851  8.4294  8.4497  8.5119  8.6095  8.6570  8.7644  8.8447...
8.9524  9.1165  9.4678  9.6167  10.2459 10.4512 10.6135 10.8366 11.5103 11.8165];
normplot(x)                                  %正态概率纸检验
[h_lillitest,p,lstat,cv]= lillietest(x,0.05)  %Lillifors检验
```

[h_jbtest,p,jbstat,cv]=jbtest(x,0.05)　　%偏度峰度检验

• 运行命令文件 addexample3_1.m:

>> addexample3_1

①(正态概率纸检验)

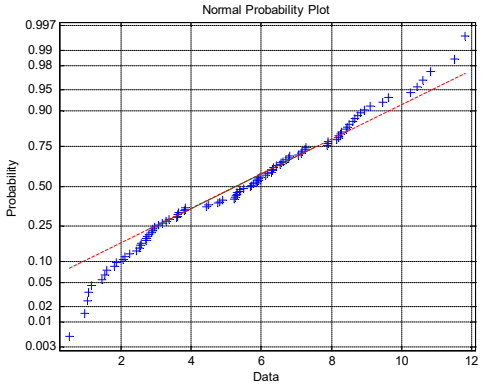


图 3-1　正态概率纸图

②(Lilliefors 检验)

h_lillitest =　1

p =　0.0410

lstat =　0.0910

cv =　0.0890

③(偏度峰度检验)

h_jbtest =　0

p =　0.1372

jbstat =　3.0874

cv =　5.4314

(3) 问题结果:

①由正态概率纸图可以认为这批数据来自正态分布;

②由于 h_lillitest = 1, 故 Lilliefors 检验认为这批数据不是来自正态分布;

③由于 h_jbtest = 0, 故偏度峰度检验认为这批数据来自正态分布.

说明: 实际上本题数据来自 $\chi^2(6)$.

9．qqplot(x,y)

功能: 绘出两样本的分位数-分位数图. 图中样本数据用符号‘＋’来表示, 叠加在数据上的实线是各分布的第一个与第三个四分位点之间的连线. 如果两个样本来源于**同一类型**(此“类型”是指该类型任意两个分布的 α 分位数具有线性关系，如均匀分布、指数分布、正态分布等，但威布尔分布则没该性质)的分布, 则‘＋’线近似地在一直线上.

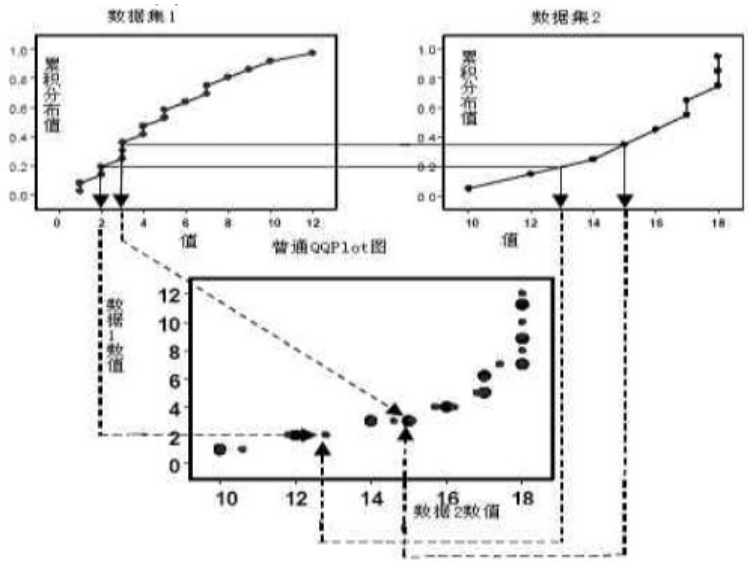


图 3-2　两样本 q-q 图映射原理

qqplot(x) 绘出样本 x 的分位数-标准正态分布的理论分位数图. 如 x 来自正态分布, 则‘＋’线近似地在一直线上.

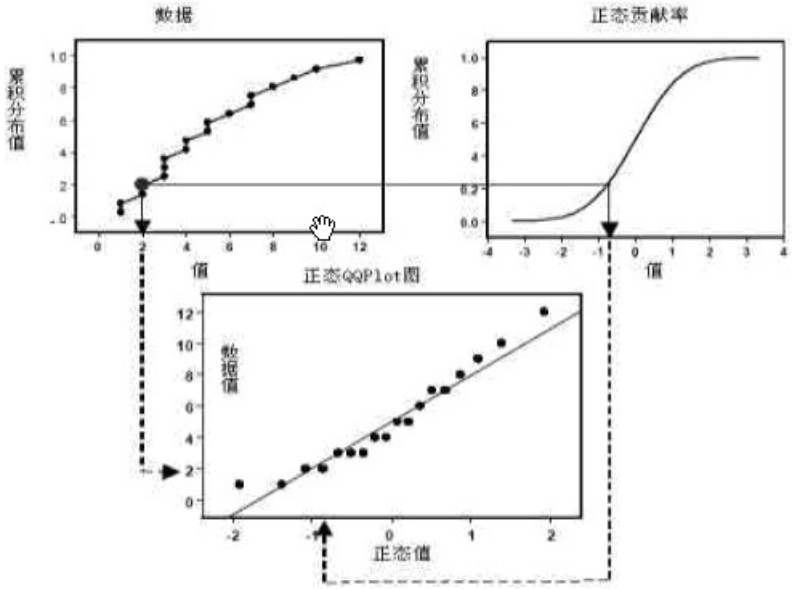


图 3-3　单样本 q-q 图映射原理

【补例 3-2】设 $X \sim N(0,1)$, $Y \sim N(0.5,2^2)$, $Z \sim U[0,2]$. 试用 matlab 从 X、Y、Z 中随机产生出样本容量分别为 50、100、200 的三个样本, 分别记为 x、y、z, 并作出下列四个图 :

①y 与正态分布的理论数据之间的 q-q 图;

②z 与正态分布的理论数据之间的 q-q 图;

③x−y 的 q-q 图;

④x−z 的 q-q 图.

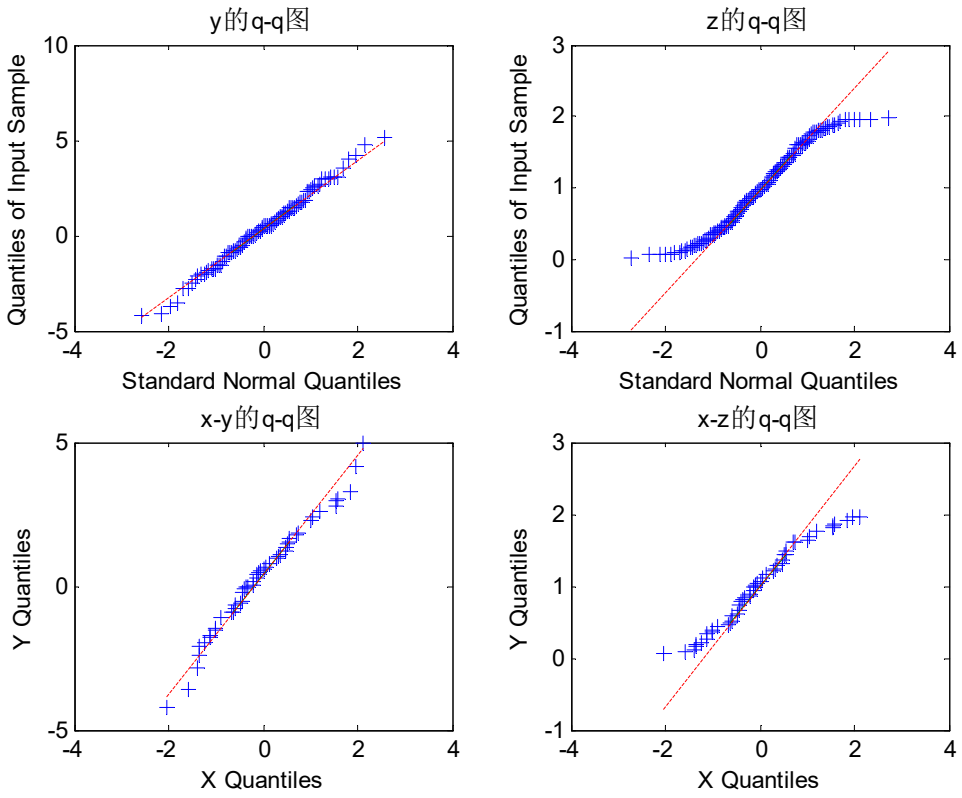


图 3-4　三个样本各种组合的 q-q 图

①从 y 的 q-q 可以推断出此数据来自正态分布总体;

②从 z 的 q-q 可以推断出此数据来自非正态分布总体;

③从 x-y 的 q-q 图可以推断出两组数据来自同一类型分布总体;

④从 x-z 的 q-q 图可以推断出两组数据不是来自同一类型分布总体.

作业: $P_{131\text{-}134}$ 3.4, 3.8, 3.12, 3.13, 3.15, 3.19, 3.21, 3.22, P_{99} 例 3.14

第四章　回归分析

一、线性回归模型

线性模型是指因变量与一个或多个自变量之间的关系可由式(4.1)形式表示:

$$y=\beta_0+\beta_1x_1+\cdots+\beta_px_p+\varepsilon \quad (\varepsilon\sim N(0,\sigma^2)) \tag{4.1}$$

当取得 n 组独立的观测值 $(y_i;x_{i1},\cdots,x_{ip}),i=1,2,\cdots,n$, 则由(4.1)有

$$\begin{cases} y_1=\beta_0+\beta_1x_{11}+\cdots+\beta_px_{1p}+\varepsilon_1 \\ \quad\vdots \\ y_i=\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip}+\varepsilon_i \text{ 且 } \varepsilon_i \overset{i.i.d.}{\sim} N(0,\sigma^2) \quad (i=1,\cdots,n) \\ \quad\vdots \\ y_n=\beta_0+\beta_1x_{n1}+\cdots+\beta_px_{np}+\varepsilon_n \end{cases} \text{——样本模型} \tag{4.2}$$

若记 $Y=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n\times1}$, $X=\begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}_{n\times(p+1)}$, $\beta=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}_{(p+1)\times1}$, $\varepsilon=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{n\times1}$,

其中,

Y——表示 n×1 的因变量观测值向量;

X——表示 n×(p+1)的由自变量决定的设计矩阵, 且要求 X 列满秩, 即 rank(X)=p+1;

β——表示(p+1)×1 的未知参数向量;

ε——表示 n×1 的误差向量, 各分量相互独立且通常服从正态分布.

则(4.2)可写成

$$Y_{n\times1}=X_{n\times(p+1)}\beta_{(p+1)\times1}+\varepsilon_{n\times1} \quad (\varepsilon\sim N(0_{n\times1},\sigma^2I_n)) \quad \text{——矩阵模型} \tag{4.3}$$

其中 I_n 为 n 阶单位阵.

二、回归分析中研究的主要问题

1、确定因变量 y 对自变量 $x_1,x_2,\cdots,x_p$ 的回归方程:

(1) β 的 LS 估计为 $\hat{\beta}=(X'X)^{-1}X'Y$;

(2) y 对 $x_1,x_2,\cdots,x_p$ 的回归方程为: $\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_px_p$.

2、对求得的回归方程的可信度(显著性)进行检验:

(1)检验假设 $H_0:\beta_1=\beta_2=\cdots=\beta_p=0$;

(2)检验统计量: $F=\dfrac{U/p}{Q_e/(n-p-1)}\overset{H_0:\beta_1=\cdots=\beta_p=0\text{成立}}{\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim}F(p,n-p-1)$;

(3) $H_0:\beta_1=\beta_2=\cdots=\beta_p=0$ 的拒绝域: $F>F_{1-\alpha}(p,n-p-1)$.

3、判断自变量 $x_j(j=1,2,\cdots,p)$ 对 y 有无显著的影响:

(1)检验假设 $H_{0j}:\beta_j=0 \quad (j=1,\cdots,p)$;

(2)检验统计量: $T_j=\dfrac{\hat{\beta}_j\big/\sqrt{c_{jj}}}{\sqrt{Q_e/(n-p-1)}}\overset{H_{0j}成立}{\sim\sim\sim\sim} t(n-p-1)$,

其中 c_{jj} 为矩阵 $(X'X)^{-1}$ 的第 $j+1$ 个对角元;

(3) $H_{0j}:\beta_j=0 \quad (j=1,\cdots,p)$ 的拒绝域: $|T_j|>t_{1-\alpha/2}(n-p-1)$.

4、利用所求得的回归方程进行预测和控制:

令 $(x_{01},x_{02},\cdots,x_{0p})$ 为 $(x_1,x_2,\cdots,x_p)$ 的一组固定值, 设 $y_0,y_1,\cdots,y_n$ 相互独立. 则

(1) y_0 的预测值:

$$\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1x_{01}+\cdots+\hat{\beta}_px_{0p}. \tag{4.4}$$

(2) y_0 的置信水平为 $1-\alpha$ 的预测区间:

$$[\hat{y}_0\pm\delta(x_0)] \tag{4.5}$$

其中, $\delta(x_0)=\hat{\sigma}_e\cdot t_{1-\alpha/2}(n-p-1)\cdot\sqrt{1+\dfrac{1}{n}+\sum_{i=1}^{p}\sum_{j=1}^{p}c_{ij}(x_{0i}-\overline{x}_i)(x_{0j}-\overline{x}_j)}$.

特别地, 当 n 较大而 x_{0j} 接近于 $\overline{x}_j$ 时, (4.4)近似为

$$[\hat{y}_0-\hat{\sigma}_e\cdot t_{1-\alpha/2}(n-p-1),\ \hat{y}_0+\hat{\sigma}_e\cdot t_{1-\alpha/2}(n-p-1)] \tag{4.6}$$

其中 c_{jj} 为矩阵 $(X'X)^{-1}$ 的第 $j+1$ 个对角元, $\hat{\sigma}_e=\sqrt{Q_e/(n-p-1)}$.

5、β_j 的区间估计和 σ^2 的无偏估计与区间估计:

(1) β_j 的区间估计: $\dfrac{(\hat{\beta}_j-\beta_j)\big/\sqrt{c_{jj}}}{\sqrt{Q_e/(n-p-1)}}\sim t(n-p-1)\Rightarrow[\hat{\beta}_j\pm\hat{\sigma}_e\cdot\sqrt{c_{jj}}\cdot t_{1-\alpha/2}(n-p-1)]$;

(2) σ^2 的无偏估计: $\hat{\sigma}^2=\dfrac{Q_e}{n-p-1}\triangleq\hat{\sigma}_e^2$;

σ^2 的区间估计: $\dfrac{Q_e}{\sigma^2}\sim\chi^2(n-p-1)\Rightarrow\left[\dfrac{Q_e}{\chi^2_{1-\alpha/2}(n-p-1)},\ \dfrac{Q_e}{\chi^2_{\alpha/2}(n-p-1)}\right]$.

三、回归分析函数

（一）回归分析函数

表 4-1　线性模型部分分析函数

回归分析	regress	多元线性回归	[b,bint,r,rint,stats] = regress(Y,X,alpha)
	stepwise	逐步回归分析的交互式环境	stepwise(X,Y,inmodel,penter,premove)
多项式拟合	polyfit	多项式拟合	p = polyfit(x,y,n)
	polyval	多项式求值	y_fit = polyval(p,x)

（二）回归分析函数的格式说明及例题

1．[b,bint,r,rint,stats] = regress(Y,X,alpha)

功能: 通过求解线性模型 (4.3), 返回 β 的最小二乘估计与区间估计, 残差的点估计与区间估计, 以及其他检验结果. 其中

Y——表示 n×1 的观测向量;

X——表示 n×(p+1)的设计矩阵;

b——表示 β 的最小二乘估计 $\hat{\beta}=(X'X)^{-1}X'Y$;

bint——表示 β 的置信水平为100(1−α)% 的估计区间, 是(p+1)×2 的矩阵;

r——表示残差向量, 是 n×1 的矩阵;

rint——表示残差向量的置信水平为100(1−α)% 的估计区间, 是 n×2 的矩阵;

stats——包含四个结果, 分别是回归分析中的 (复) 相关系数的平方 R^2 统计量 $R^2=\dfrac{U}{L_{yy}}=\dfrac{U}{Q_e+U}$、F 统计量 $F=\dfrac{U/p}{Q_e/(n-p-1)}$、p 值、以及误差方差估计 $\hat{\sigma}^2=\dfrac{Q_e}{n-p-1}$.

【★例 4.1-4.4(P138-159)】为研究温度 x 对某个化学过程的生产量 y 的影响, 收集到如下数据:

x	-5	-4	-3	-2	-1	0	1	2	3	4	5
y	1	5	4	7	10	8	9	13	14	13	18

①利用最小二乘法求 y 对 x 的回归方程, 作回归直线图与观测数据的散点图;

②在正态分布下利用 F 检验法检验回归方程效果是否显著 ($\alpha=0.05$);

③求出回归系数 β_0,β_1 的置信区间 ($\alpha=0.05$);

④取 $x_0=3$, 求 y_0 的预测值与置信水平为 $1-\alpha=0.95$ 的预测区间.

(1) 问题分析:

假设y与x有如下线性关系:

$$y=\beta_0+\beta_1x+\varepsilon\quad(\varepsilon\sim N(0,\sigma^2))$$

今得到11对相互独立的观测数据, 试利用这批数据完成题目要求的4个任务.

(2) 问题求解:

①记 $Y=\begin{pmatrix}y_1\\y_2\\\vdots\\y_n\end{pmatrix}=\begin{pmatrix}1\\5\\\vdots\\18\end{pmatrix}_{n\times1}$, $X=\begin{pmatrix}1&-5\\1&-4\\\vdots&\vdots\\1&5\end{pmatrix}_{n\times(p+1)}$, $\beta=\begin{pmatrix}\beta_0\\\beta_1\end{pmatrix}_{(p+1)\times1}$, 此处 $n=11, p=1$, 则

β 的 LS 估计为 $\hat{\beta}=(X'X)^{-1}X'Y$, 从而 y 对 x 的回归方程为 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1x$;

利用 $\hat{Y}=X\hat{\beta}$, 即 $\begin{pmatrix}\hat{y}_1\\\hat{y}_2\\\vdots\\\hat{y}_n\end{pmatrix}=\begin{pmatrix}1&-5\\1&-4\\\vdots&\vdots\\1&5\end{pmatrix}\times\begin{pmatrix}\hat{\beta}_0\\\hat{\beta}_1\end{pmatrix}$ 得到 $y_i\ (i=1,\cdots,n)$ 的拟合值 $\hat{y}_i$, 从而利用 $(x_i,\hat{y}_i)$ $(i=1,\cdots,n)$ 作出回归直线图、利用 (x_i,y_i) $(i=1,\cdots,n)$ 作出观测数据的散点图.

②对回归方程进行显著性检验:

对回归方程的显著性检验, 相当于检验假设 $H_0:\beta_1=0$;

检验统计量为: $F=\dfrac{U/p}{Q_e/(n-p-1)}\overset{H_0:\beta_1=0\text{成立}}{\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim}F(1,n-p-1)$;

$H_0:\beta_1=0$ 的拒绝域为: $F>F_{1-\alpha}(1,n-p-1)$.

若 $F>F_{1-\alpha}(1,n-p-1)$, 则认为回归方程是显著的.

③ β_0,β_1 的置信区间分别为: $[\hat{\beta}_j\pm\hat{\sigma}_e\cdot\sqrt{c_{jj}}\cdot t_{1-\alpha/2}(n-p-1)]\,(j=0,1)$,

其中 c_{jj} 为矩阵 $(X'X)^{-1}$ 的第 $j+1$ 个对角元, $\hat{\sigma}_e=\sqrt{Q_e/(n-p-1)}$.

④取 $x_0=3$, y_0 的预测值: $\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1x_0$;

y_0 的置信水平为 $1-\alpha$ 的预测区间: $\left[\hat{y}_0\pm\hat{\sigma}_e\cdot t_{1-\alpha/2}(n-p-1)\cdot\sqrt{1+\dfrac{1}{n}+\dfrac{(x_0-\bar{x})^2}{L_{xx}}}\right]$,

其中 $L_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2$.

• 编写命令文件 example4_1.m:

```
alpha=0.05;
x=[-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5]';          %x 为列向量 n*1
n=length(x);
X=[ones(n,1),x];
Y=[1, 5, 4, 7, 10, 8, 9, 13, 14, 13, 18]';          %Y 为列向量 n*1
plot(x,Y,'b.')                                      %作散点图
hold on
[b,bint,r,rint,stats] = regress(Y,X,alpha)          %利用最小二乘法进行线性回归
disp('(1)----------------------------------------')  %第(1)小题
beta_hat=b                                          %回归系数 beta 的 LS 估计
Y_hat=X*beta_hat;                                   %观测向量 Y 的拟合值
plot(x,Y_hat,'r')                                   %作回归直线
legend('散点图','回归直线')
hold off
disp('(2)----------------------------------------')  %第(2)小题
F_equation=stats(1,2:3)                  %回归方程显著性检验中 F 统计量的值与 p 值
disp('(3)----------------------------------------')  %第(3)小题
beta_ci=bint                                        %回归系数 beta 的区间估计
disp('(4)----------------------------------------')  %第(4)小题
x0=3;
y0_hat=[1,x0]*beta_hat                              %y0 的点估计
sigma_hat=sqrt(stats(1,4));
```

```
t_alpha=tinv(1-alpha/2,n-2);
Lxx=(x-mean(x))'*(x-mean(x));
delta=sigma_hat*t_alpha*sqrt(1+1/n+(x0-mean(x))^2/Lxx);
y0_ci=[y0_hat-delta,y0_hat+delta]                    %y0 的区间估计
```

• 运行命令文件 example4_1.m:

```
>> example4_1
    (1)----------------------------------------
    beta_hat =    9.2727    1.4364
    (2)----------------------------------------
    F_equation =   96.1798    0.0000
    (3)----------------------------------------
    beta_ci =    8.2250   10.3204
                 1.1050    1.7677
    (4)----------------------------------------
    y0_hat =   13.5818
    y0_ci =    9.8188   17.3449
```

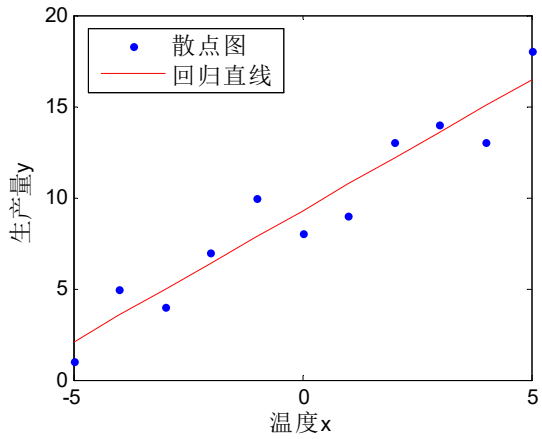


图 4-1　观测数据的散点图与回归直线图

(3) 问题结果:

①回归方程为 $\hat{y}=9.2727+1.4364x$, 回归直线与观测数据散点图见图 4-1;

②由于 F 统计量 F=96.1798> $F_{1-\alpha}(n-p-1)=F_{0.95}(9)=5.1174$, 样本落入 H_0 的拒绝域中, 故拒绝原 假设 $H_0:\beta_1=0$, 说明回归方程显著;

或由于 p=0.0000< $\alpha=0.05$, 故拒绝原假设 $H_0:\beta_1=0$, 说明回归方程显著;

③ β_0 的置信区间为[8.225, 10.3204], β_1 的置信区间为[1.1050, 1.7677];

④当 $x_0=3$, $\hat{y}_0=13.5818$, y_0 的置信水平为 0.95 的预测区间为[9.8188, 17.3449].

【★例 4.5-4.7($P_{178\text{-}185}$)】在平炉炼钢中, 由于矿石与炉气的氧化作用, 铁水的总含碳量在不断降低, 一炉钢在冶炼初期总的去碳量 y 与所加的两种矿石的量 x_1, x_2 及熔化时间 x_3 有关. 经实测某号平炉的 49 组数据如下表所列:

编号	x_1	x_2	x_3	y	编号	x_1	x_2	x_3	y
1	2	18	50	4.3302	26	9	6	39	2.7066

2	7	9	40	3.6485	27	12	5	51	5.6314
3	5	14	46	4.4830	28	6	13	41	5.8152
4	12	3	43	5.5468	29	12	7	47	5.1302
5	1	20	64	5.4970	30	0	24	61	5.3910
6	3	12	40	3.1125	31	5	12	37	4.4533
7	3	17	64	5.1182	32	4	15	49	4.6569
8	6	5	39	3.8759	33	0	20	45	4.5212
9	7	8	37	4.6700	34	6	16	42	4.8650
10	0	23	55	4.9536	35	4	17	48	5.3566
11	3	16	60	5.0060	36	10	4	48	4.6098
12	0	18	49	5.2701	37	4	14	36	2.3815
13	8	4	50	5.3772	38	5	13	36	3.8746
14	6	14	51	5.4849	39	9	8	51	4.5919
15	0	21	51	4.5960	40	6	13	54	5.1588
16	3	14	51	5.6645	41	5	8	100	5.4373
17	7	12	56	6.0795	42	5	11	44	3.9960
18	16	0	48	3.2194	43	8	6	63	4.3970
19	6	16	45	5.8076	44	2	13	55	4.0622
20	0	15	52	4.7306	45	7	8	50	2.2905
21	9	0	40	4.6805	46	4	10	45	4.7115
22	4	6	32	3.1272	47	10	5	40	4.5310
23	0	17	47	2.6104	48	3	17	64	5.3637
24	9	0	44	3.7174	49	4	15	72	6.0771
25	2	16	39	3.8946					

设 y 与 x_1, x_2, x_3 之间有线性关系

$$\begin{cases} y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \quad i = 1, \cdots, 49 \\ \varepsilon_i \sim N(0, \sigma^2), \text{且} \varepsilon_1, \cdots, \varepsilon_{49} \text{相互独立} \end{cases}$$

①【例 4.5】求 y 与 x_1, x_2, x_3 的回归方程, 并检验回归方程的显著性($\alpha = 0.01$).

②【例 4.6】在不剔除不显著变量的前提下, 求回归系数 $\beta_1, \beta_2, \beta_3$ 的置信区间 ($1-\alpha = 0.95$);

③【例 4.7】在不剔除不显著变量的前提下, 若取 $(x_{01}, x_{02}, x_{03}) = (5,10,50)$, 求 y_0 的置信水平为 95%的预测区间;

④【例 4.5】检验①所得回归方程中各回归系数的显著性. 如有不显著的变量, 请剔除之并求剔除不显著的变量之后的回归方程($\alpha = 0.01$).(本小题程序较难, 可不看, 例 4-4 之后我们将借用逐步回归函数 stepwise, 通过手动实现“只出不进”法, 从而淘汰掉所有不显著自变量.)

(1) 问题分析:

假设y与 x_1, x_2, x_3 有如下线性关系:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \quad (\varepsilon \sim N(0, \sigma^2))$$

今得到49组相互独立的观测数据, 试利用这批数据完成题目要求的4个任务.

(2) 问题求解:

①第一问: 求回归方程.

记 $Y=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}=\begin{pmatrix} 4.3302 \\ 3.6485 \\ \vdots \\ 6.0771 \end{pmatrix}_{n\times 1}$, $X=\begin{pmatrix} 1 & 2 & 18 & 50 \\ 1 & 7 & 9 & 40 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 4 & 15 & 72 \end{pmatrix}_{n\times(p+1)}$, $\beta=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$, 此处 $n=49$, $p=3$, 则

β 的 LS 估计为 $\hat{\beta}=(X'X)^{-1}X'Y$, 从而 y 对 x_1,x_2,x_3 的回归方程为:

$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\hat{\beta}_3x_3$;

第二问: 对回归方程进行显著性检验.

对回归方程的显著性检验, 相当于检验假设 $H_0:\beta_1=\beta_2=\beta_3=0$;

检验统计量为: $F=\dfrac{U/p}{Q_e/(n-p-1)}\overset{H_0:\beta_1=\beta_2=\beta_3=0\text{成立}}{\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim}F(p,n-p-1)$;

$H_0:\beta_1=\beta_2=\beta_3=0$ 的拒绝域为: $F>F_{1-\alpha}(p,n-p-1)$.

若 $F>F_{1-\alpha}(p,n-p-1)$, 则认为回归方程是显著的.

② β_1,β_2,β_3 的置信区间分别为: $[\hat{\beta}_j\pm\hat{\sigma}_e\cdot\sqrt{c_{jj}}\cdot t_{1-\alpha/2}(n-p-1)]\ (j=1,2,3)$;

其中 c_{jj} 为矩阵 $(X'X)^{-1}$ 的第 $j+1$ 个对角元, $\hat{\sigma}_e=\sqrt{Q_e/(n-p-1)}$.

③取 $(x_{01},x_{02},x_{03})=(5,10,50)$, y_0 的预测值: $\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1x_{01}+\hat{\beta}_2x_{02}+\hat{\beta}_3x_{03}$;

由于 (x_{01},x_{02},x_{03}) 与 $\bar{x}$ 接近, 且 $n=49$ 较大, 故 y_0 的置信水平为 $1-\alpha$ 的预测区间取近似区间为:

$$\left[\hat{y}_0\pm\hat{\sigma}_e\cdot t_{1-\alpha/2}(n-p-1)\right].$$

④对问题①中所求回归方程中的各回归系数进行显著性检验, 用“只出不进法”依次剔除最不显著变量, 其中最不显著变量的标准如下:

计算回归方程中各回归系数的 t 统计量的值 $t_j=\dfrac{\hat{\beta}_j/\sqrt{C_{jj}}}{\sqrt{Q_e/(n-p_0-1)}}\ (j=1,\cdots,p_0)$ (其中 p_0 为当前回归方程中自变量的个数), 满足 $|T_j|<t_{1-\alpha/2}(n-p_0-1)$ 中的最小 t_j 所对的自变量即为最不显著的变量.

• 编写命令文件 example4_5_0.m:

```
%A 为观测数据阵, 共 49 行 4 列.
A=[2      18      50      4.3302
          中间数据略
   4      15      72      6.0771];
```

• 编写命令文件 example4_5.m:

```
x=A(:,1:3);
[n,p]=size(x);                          %n为观测数据组数, p为自变量个数
X=[ones(n,1),x];
```

```
Y=A(:,4);
alpha14=0.01;                                  %第(1)(4)小题的显著性水平
alpha23=0.05;                                  %第(2)(3)小题的显著性水平
disp('(1)----------------------------------------')    %第(1)小题: 回归方程与回归方程的显著性检验
[b,bint,r,rint,stats]=regress(Y,X,alpha14);
beta_hat=b                                     %回归系数beta的LS估计
F_equation=stats(1,2:3)                        %回归方程的显著性检验中的F统计量的值与p值
disp('(2)----------------------------------------')    %第(2)小题: 回归系数的置信区间
[b,bint,r,rint,stats]=regress(Y,X,alpha23);
beta_ci=bint(2:p+1,:)                          %回归系数beta的置信区间
disp('(3)----------------------------------------')    %第(3)小题: y0的预测区间
x0=[5,10,50];
y0_hat=[1,x0]*b;
sigma_hat =sqrt(stats(1,4));
t_alpha=tinv(1-alpha23/2,n-p-1);
%因为x0=[5,10,50]与xi的平均值[5.2857,11.7959,49.2041]接近,且n=49较大,故求近似区间
y0_ci=[y0_hat-sigma_hat*t_alpha,y0_hat+sigma_hat*t_alpha]          %y0的预测区间
disp('(4)----------------------------------------')    %第(4)小题: 使用“只出不进法”剔除不显著变量
p0=p;                                          %用于记录删除某个自变量后剩余的自变量个数
for j=1:p
   disp('------------------------')
   p0
   [b,bint,r,rint,stats]=regress(Y,X,alpha14);     %再次建立回归方程
   beta_hat=b
   sigma2_hat=stats(1,4);                      %误差方差sigma2的估计
   C=inv(X'*X);                                %矩阵(X'X)的逆矩阵
   Cjj=diag(C);                                %C阵中对角线上所有元素
   tj=beta_hat(2:p0+1)./sqrt(Cjj(2:p0+1))./sqrt(sigma2_hat)   %对各自变量显著性检验时, 各tj统计量的值
   t0=tinv(1-alpha14/2, n-p0-1)                %t检验的临界点
   if  min(abs(tj))<t0                         %表示有不显著的自变量
      jj=find(tj>min(abs(tj)));                %从当前所有自变量中将最不显著自变量所在的列标剔除后,剩余的自变量的列标
      x=x(:,jj);
      X=[ones(n,1),x];                         %从设计阵中剔除掉最不显著自变量所在列
      p0=p0-1;
```

```
    else
        break                                   %如果所有变量均显著, 则跳出for循环语句
    end
end
```

• 依次运行命令文件 example4_5_0.m 与 example4_5.m:

```
>> example4_5_0
>> example4_5
    (1)----------------------------------------
    beta_hat =    0.6952    0.1606    0.1076    0.0359
    F_equation =    7.7011          0.0003
    (2)----------------------------------------
    beta_ci =    0.0392    0.2821
                 0.0322    0.1829
                 0.0147    0.0572
    (3)----------------------------------------
    y0_ci =    2.7359    6.0069
    (4)----------------------------------------
    ------------------------
    p0 =     3
    beta_hat =    0.6952    0.1606    0.1076    0.0359
    tj =    2.6634    2.8760    3.4012
    t0 =    2.6896
    ------------------------
    p0 =     2
    beta_hat =    2.5150    0.0233    0.0364
    tj =    1.0977    3.2386
    t0 =    2.6870
    ------------------------
    p0 =     1
    beta_hat =    2.6475    0.0393
    tj =    3.5883
    t0 =    2.6846
```

(3) 问题结果:

①回归方程为 $\hat{y}=0.6952+0.1606x_1+0.1076x_2+0.0359x_3$, 由于 F 检验法的 $p=0.0003<0.01=\alpha$, 故在 $\alpha=0.01$ 下, 认为此回归方程显著;

② β_1 的置信区间为[0.0392, 0.2821], β_2 的置信区间为[0.0322, 0.1829], β_3 的置信区间为[0.0147, 0.0572];

③ 当 $(x_{01},x_{02},x_{03})=(5,10,50)$, y_0 的置信水平为 0.95 的近似预测区间为[2.7359, 6.0069];

④ 第一次线性回归得回归方程为: $\hat{y}=0.6952+0.1606x_1+0.1076x_2+0.0359x_3$, 经 t 检验, 由于 $t_1=2.6634<2.6896=t_0=t_{0.995}(45)$, 故在 $\alpha=0.01$ 下 x_1 不显著;

剔除 x_1 进行第二次线性回归, 得回归方程为: $\hat{y}=2.515+0.0233x_2+0.0364x_3$, 经 t 检验, 由于 $t_2=1.0977<2.6870=t_0=t_{0.995}(46)$, 故在 $\alpha=0.01$ 下 x_2 不显著;

剔除 x_2 进行第三次线性回归, 得回归方程为: $\hat{y}=2.6475+0.0393x_3$, 经 t 检验, 由于 $t_3=3.5883>2.6846=t_0=t_{0.995}(47)$, 故在 $\alpha=0.01$ 下 x_3 显著.

利用“只出不进法”, 在显著性水平 $\alpha=0.01$ 下, 最终的线性回归方程为 $\hat{y}=2.6475+0.0393x_3$.

【★例 4.8-4.9(P190-192)】出钢时所用的盛钢水的钢包, 由于钢水对耐火材料的浸蚀, 容积不断增大. 我们希望找出使用次数 x 与增大的容积 y 之间的关系. 试验数据列于下表. 试分别选用

模型 1: 双曲线 $1/y=a+b/x$;　　模型 2: 倒指数曲线 $y=ae^{b/x}$

进行非线性回归分析, 并对回归方程进行显著性检验; 然后作出散点图与回归曲线图; 如果所求回归方程均显著, 试比较哪个回归模型更优. ($\alpha=0.01$)

x	2	3	4	5	6	7	8	9
y	6.42	8.20	9.58	9.50	9.70	10.00	9.93	9.99
x	10	11	12	13	14	15	16	
y	10.49	10.59	10.60	10.80	10.60	10.90	10.76	

(1) 问题分析:

首先作出散点图, 发现y是关于x的增函数, 但增加速度随x的增大而减缓, 且据实际背景知, y最终逼近于一个定值, 故选用下列两个模型进行非线性回归分析:

模型 1: 双曲线 $1/y=a+b/x$;　　模型 2: 倒指数曲线 $y=ae^{b/x}$

今有15对相互独立的观测数据, 需利用这批数据完成下列3个任务:

①对模型1, 通过变量代换, 将非线性转化为线性模型, 求得一元线性回归方程, 检验此方程的显著性, 并绘出原始数据散点图与非线性回归方程的曲线图;

②对模型2, 完成对模型1一样的工作;

③如果两个方程均显著, 则选出最优回归模型.

(2) 问题求解:

①对模型 1:

首先, 求回归方程.令 $y1=\frac{1}{y}, x1=\frac{1}{x}$, 得到线性模型 $y1=a1+b1\cdot x1+\varepsilon 1$, 且设 $E(\varepsilon 1)=0$, $D(\varepsilon 1)=\sigma 1^2$. 类似例 4-1, 利用最小二乘法求得回归系数, 得到

一元线性回归方程: $\hat{y}1=\hat{a}1+\hat{b}1\cdot x1$　　(4.7)

一元双曲线回归方程: $\frac{1}{\hat{y}}=\hat{a}1+\hat{b}1\cdot\frac{1}{x}$　　(4.8)

其次, 对回归方程(4.7)进行显著性检验, 相当于检验假设 $H_{01}:\hat{b}1=0$.

取检验统计量 $F1=\frac{U/p}{Q_e/(n-p-1)}\overset{H_{01}:\hat{b}_1=0\text{成立}}{\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim}F(p,n-p-1)$ (其中 $n=15, p=1$), 则

$H_{01}:\hat{b}1=0$ 的拒绝域为: $F1>F_{1-\alpha}(p,n-p-1)$.

若 $F1>F_{1-\alpha}(p,n-p-1)$, 则认为回归方程(4.7)是显著的.

最后, 利用(4.8)得到 y 的拟合值, 绘出原始数据的散点图与双曲线回归方程的曲线图.

②对模型 2, 完全类似模型 1 处理.

③如果两个方程均显著, 选取残差平方和 $Q_e=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=e'e$ 小的那个回归方程所对的原非线性回归方程为最优方程.

• 编写命令文件 example4_8.m:

```
x=[2:16]';
y=[6.42, 8.20, 9.58, 9.50, 9.70, 10.00, 9.93, 9.99, 10.49, 10.59, 10.60, 10.80, 10.60, 10.90, 10.76]';
n=length(x);
alpha=0.01;
plot(x,y,'ko')                          %作散点图
hold on
disp('(1)双曲线模型 1/y=a+b/x----------')          %(1)双曲线 1/y=a+b/x 模型
y1=1./y;
x1=1./x;
X1=[ones(n,1),x1];
[b1,b1int,r1,r1int,stats1]=regress(y1,X1,alpha);
b1                                      %y1=a1+b1*x1 中 a1, b1 的值
ab1=b1                                  %双曲线 1/y=a+b/x 中 a,b 的值
F1=stats1(1,2:3)                        %对线性回归方程(4.7)进行显著性检验的 F 值与 p 值
y1_fit=X1*b1;
y_fit=1./y1_fit;
plot(x,y_fit,'r')                       %绘拟合双曲线 1/y=a+b/x
Qe1=(y-y_fit)'*(y-y_fit)                %计算残差平方和
disp('(2)倒指数曲线 y=a*exp(b/x)----------')      %(2)倒指数曲线 y=a*exp(b/x)模型
y2=log(y);
x2=1./x;
X2=[ones(n,1),x2];
[b2,b2int,r2,r2int,stats2]=regress(y2,X2,alpha);
b2                                      %y2=c2+b2*x2 中 c2, b2 的值
ab2=[exp(b2(1,1)),b2(2,1)]              %倒指数曲线 y=a*exp(b/x)中 a,b 的值
F2=stats2(1,2:3)                        %对线性回归方程进行显著性检验的 F 值与 p 值
y2_fit=X2*b2;
```

```
y_fit=exp(y2_fit);
plot(x,y_fit,'b-.')                    %绘拟合倒指数曲线 y=a*exp(b/x)
hold off
Qe2=(y-y_fit)'*(y-y_fit)               %计算残差平方和
legend('散点图','双曲线 1/y=a+b/x 拟合图','倒指数曲线 y=a*exp(b/x)拟合图')
disp('(3)选择最优非线性回归模型-----------------------------------')    %(3)最优非线性回归模型的选择
if Qe1<=Qe2
   disp('选择 1/y=a+b/x')
else
    disp('选择 y=a*exp(b/x)')
end
```

• 运行命令文件 example4_8.m:

```
>> example4_8
```

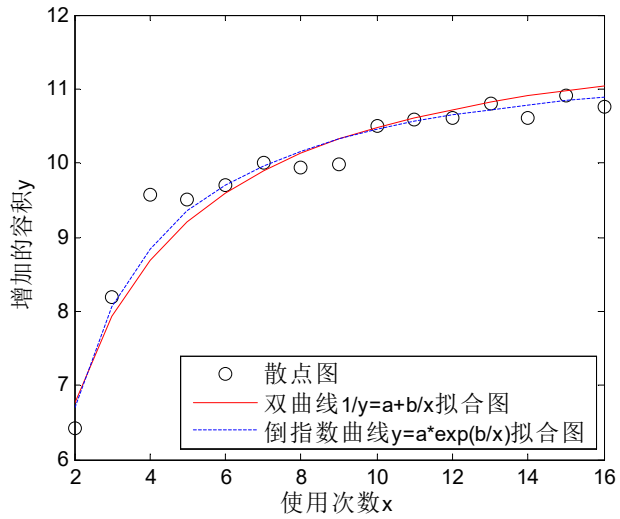


图 4-2　观测数据的散点图与回归曲线图

```
(1)双曲线模型 1/y=a+b/x------------------------------------------------------------
b1 =    0.0823    0.1312
ab1 =    0.0823    0.1312
F1 =  196.2270    0.0000
Qe1 =    1.4396
(2)倒指数曲线 y=a*exp(b/x)---------------------------------------------------
b2 =    2.4578   -1.1107
ab2 =   11.6789   -1.1107
F2 =  303.1896    0.0000
Qe2 =    0.8913
(3)选择最优非线性回归模型-----------------------------------------------------
选择 y=a*exp(b/x)
```

(3) 问题结果:

①对双曲线模型 1, 线性回归方程为 $\hat{y}1=0.0823+0.1312\cdot x1$, 由于 F 检验法的 $p=0.0000<0.01=\alpha$, 故在 $\alpha=0.01$ 下, 认为此回归方程显著; 双曲线回归方程为 $\hat{y}=\dfrac{x}{0.0823x+0.1312}$.

②对倒指数曲线模型, 线性回归方程为 $\hat{y}2=2.4578-1.1107\cdot x2$, 由于 F 检验法的 $p=0.0000<0.01=\alpha$, 故在 $\alpha=0.01$ 下, 认为此回归方程显著; 双曲线回归方程为 $\hat{y}=11.6789e^{-\frac{1.1107}{x}}$.

③由于 $Q_e1=1.4396>0.8913=Q_e2$, 所以模型 2 倒指数曲线比模型 1 双曲线更优.

2．stepwise(X,y,inmodel,penter,premove)

功能: 进行变量 y 与 X 的逐步回归分析. X 是从设计矩阵中去掉第一列后剩下的 n×p 矩阵. 向量 inmodel 的值是包含在初始模型中的矩阵 X 列的标识, 若初始回归方程中不包含任何自变量, 则 inmodel 替换为[ ]; 若包含 x_2,x_3 且它们位于 X 中的第 2、3 列的话, 则 inmodel 替换为[2,3]. penter 表示将变量移入模型时的显著性水平, 默认值为 0.05; 而 premove 表示将变量移出模型时的显著性水平, 默认值为 0.1, 且 penter≤premove.

显示窗口分三个子栏, 其中:

➤ 第一个子栏对模型中变量项的增加或移除加以控制, 并以蓝色(或红色) 字图分别表示在(或不在) 当前模型中的变量, 同时显示当前模型或下一个模型的回归系数、t 检验统计量的值与 p 值 (蓝字为当前模型, 红字为下一个模型). 另外, Coeficients with Error Bars 为各自变量系数 β_j 的区间估计, 黑色线条是置信水平为 1-penter 的置信区间, 红色线条是置信水平为 1-premove 的置信区间.

➤ 第二个子栏中各项的含义:

intercept——当前回归方程中的截距 $\hat{\beta}_0$

RMSE——均方根误差 $\hat{\sigma}_e=\sqrt{Q_e/(n-p-1)}$

R-square——复相关系数的平方(或叫多元决定系数) $R^2=\dfrac{U}{U+Q_e}$

Adj R-sq——表示调整后的多元决定系数 Adjusted $R^2=1-(1-R^2)\dfrac{n-1}{n-p-1}$

F——方程显著的检验统计量 $F=\dfrac{U/p}{Q_e/(n-p-1)}$

p——在当前 F 值下拒绝“H_0: 当前回归方程不显著”的最小的显著性水平

【例 9.4.1(中山大学《概率论与数理统计》P229)】某种水泥在凝固时放出的热量 y 与水泥中下列 4 种化学成份有关：

x_1: $3CaO\cdot Al_2O_3$ 的成份 (%) ;

x_2: $3CaO\cdot SiO_2$ 的成份 (%) ;

x_3: $4CaO\cdot Al_2O_3\cdot Fe_2O_3$ 的成份 (%) ;

x_4: $2CaO \cdot SiO_2$ 的成份 (%) .

下表为测得的 13 组数据, 试用逐步回归分析建立 y 关于这些因子的“最优”回归方程.

编号	x_1	x_2	x_3	x_4	y
1	7	25	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.5
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

• 编写命令文件 zhongshanP230.m:

```
x=[7 25 6 60;
   1 29 15 52;
   11 56 8 20;
   11 31 8 47;
   7 52 6 33;
   11 55 9 22;
   3 71 17 6;
   1 31 22 44;
   2 54 18 22;
   21 47 4 26;
   1 40 23 34;
   11 66 9 12;
   10 68 8 12];
y=[78.5 74.3 104.3 87.6 95.5 109.2 102.7 72.5 93.1 115.9 83.8 113.3 109.4]';
stepwise(x,y,[ ],0.1, 0.1)
```

• 运行命令文件 zhongshanP230.m:

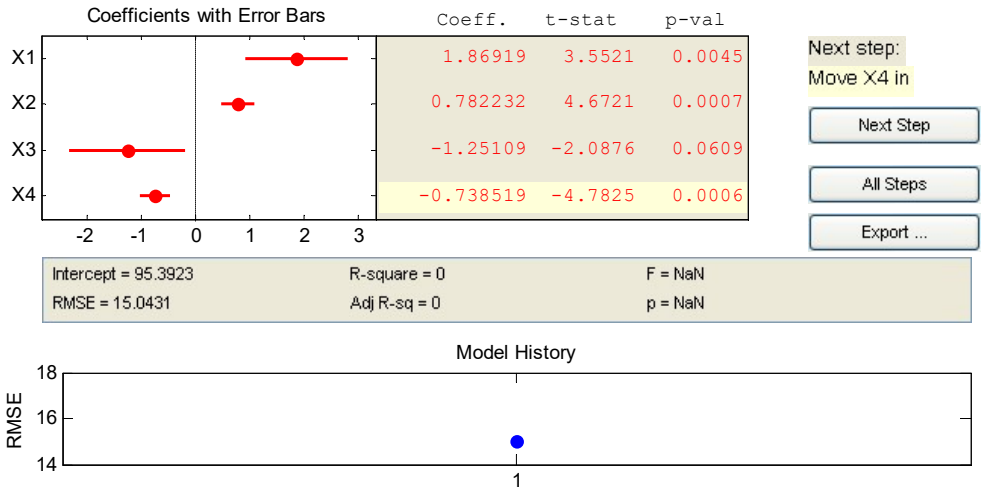


图 4-3 (1)　逐步回归第一步

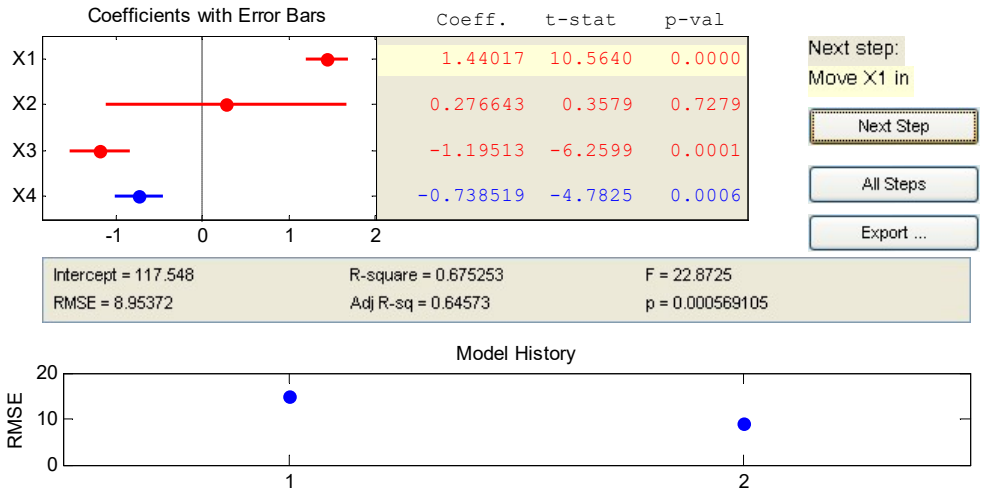


图 4-3 (2)　逐步回归第二步

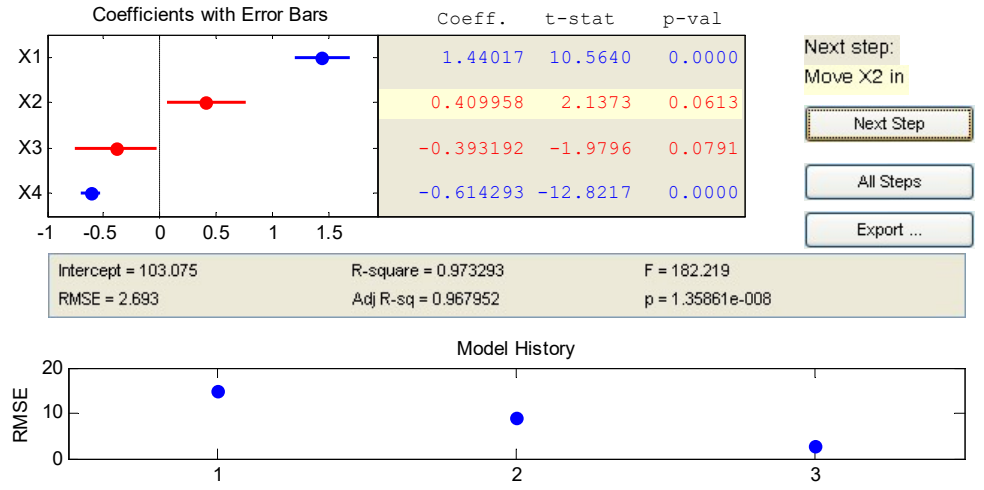


图 4-3 (3)　逐步回归第三步

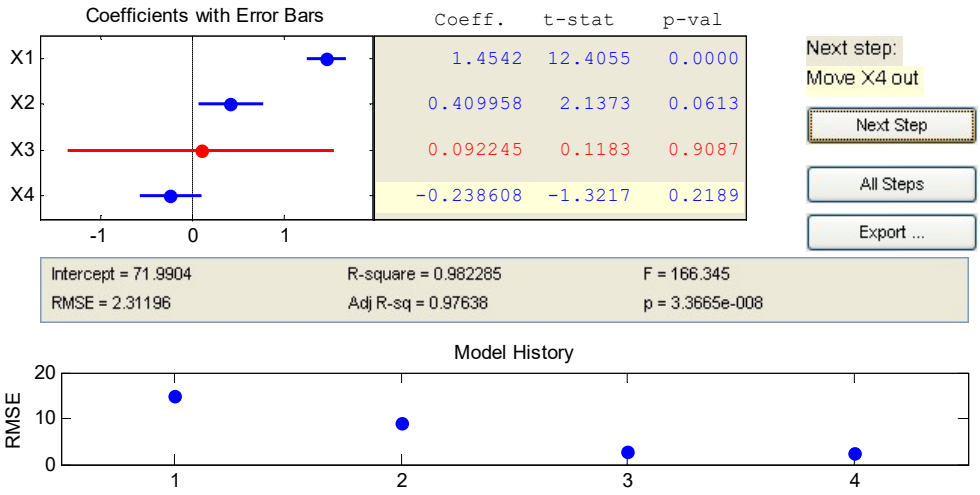


图 4-3 (4)　逐步回归第四步

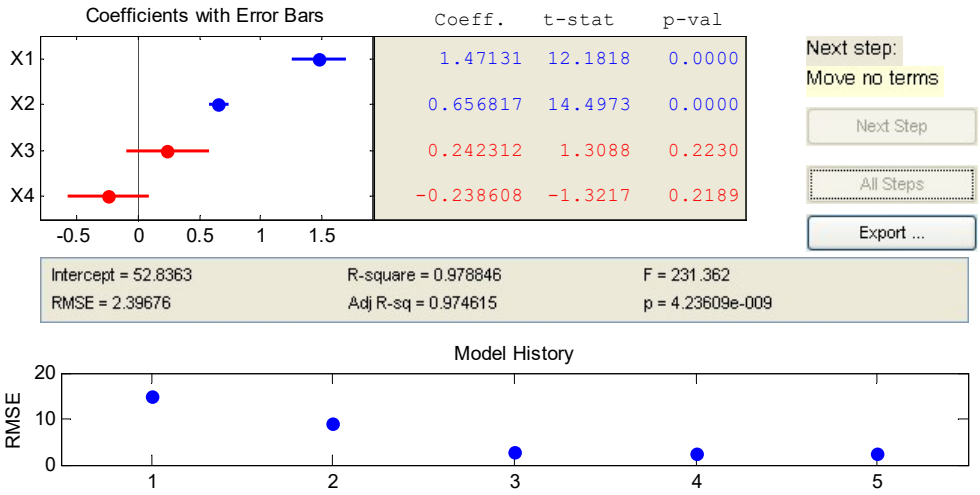


图 4-3 (5)　逐步回归第五步

可见, 在 $\alpha=0.1$ 下, 逐步回归所得的最优回归方程为 $\hat{y}=52.8363+1.4731x_1+0.6568x_2$.

同理可得, 在 $\alpha=0.05$ 下, 逐步回归所得的最优回归方程为

$$\hat{y}=103.075+1.44017x_1-0.614293x_4;$$

在 $\alpha=0.01$ 下, 逐步回归所得的最优回归方程为

$$\hat{y}=103.075+1.44017x_1-0.614293x_4.$$

【★例 4.5-4.7 续(P178-185)】在平炉炼钢中, 由于矿石与炉气的氧化作用, 铁水的总含碳量在不断降低, 一炉钢在冶炼初期总的去碳量 y 与所加的两种矿石的量 x_1,x_2 及熔化时间 x_3 有关. 经实测某号平炉的 49 组数据如下表所列(表略):

④请借用 stepwise 函数, 手动实现“只出不进”法, 以淘汰掉所有不显著自变量 ($\alpha=0.01$).

• 运行命令文件 example4_5_0.m 及下面的命令行:

```
x=A(:,1:3);
Y=A(:,4);
alpha=0.01;
stepwise(x,Y,[1,2,3],alpha,alpha)        %将 x1,x2,x3 均选入初始回归方程中
```

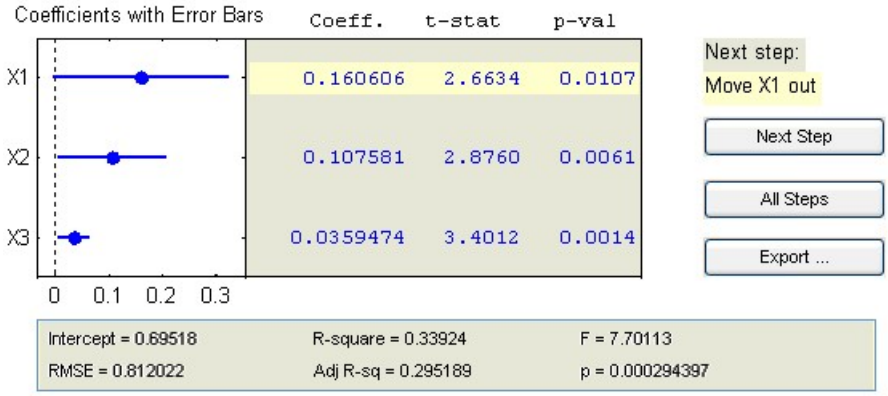


图 4-4 (1)　“只出不进”第一步

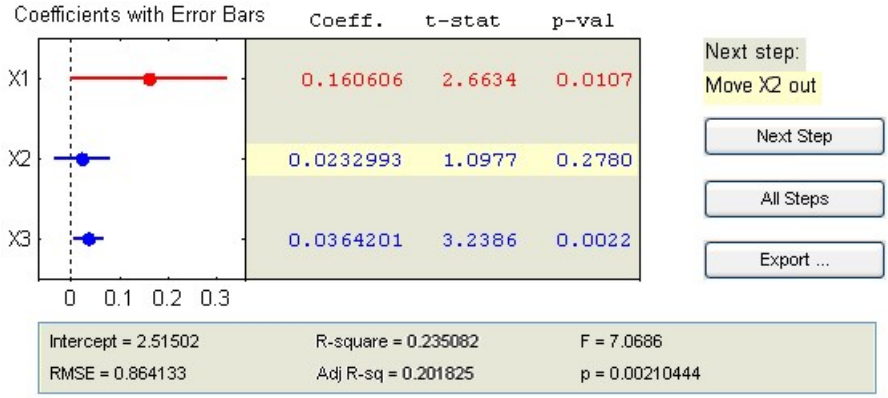


图 4-4 (2)　“只出不进”第二步

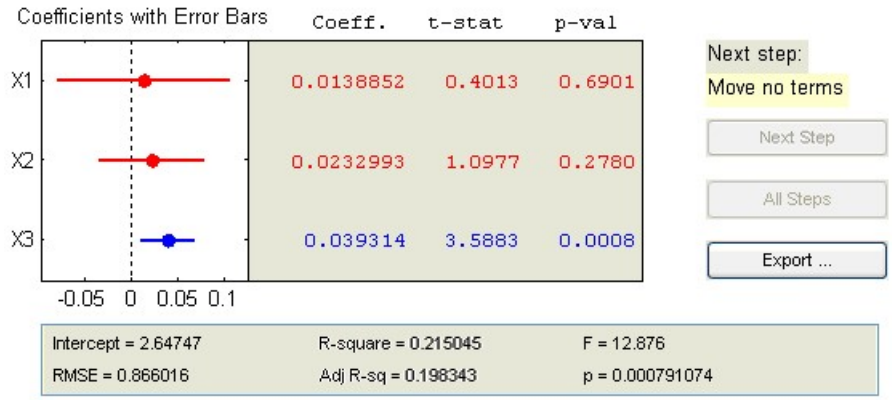


图 4-4 (3)　“只出不进”第三步

可见, 在 $\alpha = 0.01$ 下, 利用“只出不进法”得到的回归方程为 $\hat{y} = 2.6475 + 0.0393x_3$.

3．多项式拟合 (polyfit)、多项式求值 (polyval)等函数.

p=polyfit(x,y,n):

功能: 在最小二乘意义下, 将向量 x,y 进行 n 次多项式拟合, 并返回多项式的系数 p.

y=polyval(p,x):

功能: 返回给定系数 p 的多项式在 x 处的函数值 y.

★p 的第一个元素为 x^n 的系数, 最后一个元素为常数项, 这与 regress 命令中的输出 b 的顺序刚好相反.

【★例 4.10(P_{194})】某种半成品在生产过程中的废品率 y(%)与它所含的某种化学成分 x(0.01%)有关, 现将试验所得的 16 组数据记录如下:

编号	y	x	编号	y	x
1	1.30	34	9	0.44	40
2	1.00	36	10	0.56	41
3	0.73	37	11	0.30	42
4	0.90	38	12	0.42	43
5	0.81	39	13	0.35	43
6	0.70	39	14	0.40	45
7	0.60	39	15	0.41	47
8	0.50	40	16	0.60	48

求 y 对 x 的回归方程.

1、方法一: 利用polyfit函数求解.

(1) 问题分析:

首先作出这16对数据 (x_i, y_i) 的散点图, 从图4-4可以看到, y先随x的增加而降低, 当x超过一定值后, y又随x的增加而增加. 故考虑用抛物线拟合y与x的关系, 即要确定的回归方程为 $y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$.

(2) 问题求解:

• 编写命令文件 example4_10_1.m:

```
x=[34,36,37,38,39,39,39,40,40,41,42,43,43,45,47,48];
y=[1.30,1.00,0.73,0.90,0.81,0.70,0.60,0.50,0.44,0.56,0.30,0.42,0.35,0.40,0.41,0.60];
plot(x,y,'r*')                    %作观测数据的散点图
hold on
p=polyfit(x,y,2)                  %拟合多项式的系数向量[beta2,beta1,beta0]
y_fit=polyval(p,x);               %y 的拟合值
Qe=(y-y_fit)*(y-y_fit)'           %残差平方和
plot(x,y_fit,'k')                 %作拟合曲线图
hold off
```

• 运行命令文件 example4_10_1.m:

```
>> example4_10
    p =    0.0092   -0.8097   18.2642
    Qe =    0.1304
```

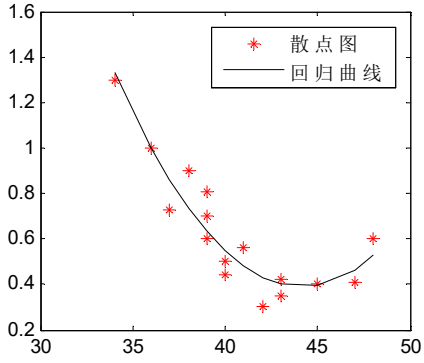


图 4-5　观测数据的散点图与回归曲线图

(3) 问题结果:

所求回归方程为 $\hat{y}=0.0092x^2-0.8097x+18.2642$, 此时残差平方和为 $Q_e=0.1304$; 观测数据散点图与回归曲线图见图 4-4.

2、方法二: 利用线性回归分析函数regress求解.

(1) 问题分析:

首先作出这16对数据 (x_i,y_i) 的散点图, 从图4-4可以看到, y先随x的增加而降低, 当x超过一定值后, y又随x的增加而增加. 故考虑y与x具有如下关系:

$$y=\beta_0+\beta_1x+\beta_2x^2+\varepsilon \quad (\varepsilon\sim N(0,\sigma^2)) \tag{4.10}$$

试确定回归方程 $y=\hat{\beta}_0+\hat{\beta}_1x+\hat{\beta}_2x^2$, 并检验此方程的显著性.

(2) 问题求解:

①求回归方程, 并作图.

记 $x_1=x, x_2=x^2$, 则模型(4.10)可转化为二元线性回归模型:

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\varepsilon \quad (\varepsilon\sim N(0,\sigma^2)) \tag{4.11}$$

从而数据扩展为下表:

编号	y	$x_1=x$	$x_2=x^2$	编号	y	$x_1=x$	$x_2=x^2$
1	1.30	34	1156	9	0.44	40	1600
2	1.00	36	1296	10	0.56	41	1681
3	0.73	37	1369	11	0.30	42	1764
4	0.90	38	1444	12	0.42	43	1849
5	0.81	39	1521	13	0.35	43	1849
6	0.70	39	1521	14	0.40	45	2025
7	0.60	39	1521	15	0.41	47	2209
8	0.50	40	1600	16	0.60	48	2304

记 $Y=\begin{pmatrix}y_1\\y_2\\\vdots\\y_n\end{pmatrix}=\begin{pmatrix}1.30\\1.00\\\vdots\\0.60\end{pmatrix}_{n\times1}$, $X=\begin{pmatrix}1&34&1156\\1&36&1296\\\vdots&\vdots&\vdots\\1&48&2304\end{pmatrix}_{n\times(p+1)}$, $\beta=\begin{pmatrix}\beta_0\\\beta_1\\\beta_2\end{pmatrix}_{(p+1)\times1}$, 此处 $n=16, p=2$, 则 β 的 LS 估计为 $\hat{\beta}=(X'X)^{-1}X'Y$, 从而 y 对 x_1,x_2 的线性回归方程为 $y=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2$;

利用 $\hat{Y}=X\hat{\beta}$ 得到 $y_i\ (i=1,\cdots,n)$ 的拟合值 $\hat{y}_i$, 从而利用 $(x_i,y_i)\ (i=1,\cdots,n)$ 作出观测数据的散点图、利用 $(x_i,\hat{y}_i)\ (i=1,\cdots,n)$ 作出回归曲线图.

②对回归方程进行显著性检验.

对回归方程的显著性检验, 相当于检验假设 $H_0:\beta_1=\beta_2=0$;

检验统计量为: $F=\dfrac{U/p}{Q_e/(n-p-1)}\overset{H_0:\beta_1=\beta_2=0\text{成立}}{\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim}F(p,n-p-1)$;

$H_0:\beta_1=\beta_2=0$ 的拒绝域为: $F>F_{1-\alpha}(p,n-p-1)$.

若 $F>F_{1-\alpha}(p,n-p-1)$, 则认为回归方程是显著的.

• 编写命令文件 example4_10_2.m:

```
x=[34,36,37,38,39,39,39,40,40,41,42,43,43,45,47,48]';
Y=[1.30,1.00,0.73,0.90,0.81,0.70,0.60,0.50,0.44,0.56,0.30,0.42,0.35,0.40,0.41,0.60]';
plot(x,Y,'r*')                          %作观测数据的散点图
hold on
X=[ones(16,1),x,x.^2];
[b,bint,r,rint,stats]=regress(Y,X,0.05);
beta_hat=b                              %回归系数 beta 的 LS 估计
F_equation=stats(1,2:3)                 %回归方程的显著性检验中的 F 统计量的值与 p 值
Y_hat=X*beta_hat;                       %观测向量 Y 的拟合值
plot(x,Y_hat,'k')                       %作拟合曲线图
legend('散点图','回归曲线')
hold off
```

• 运行命令文件 example4_10.m:

```
>> example4_10_2
    beta_hat =   18.2642   -0.8097    0.0092
    F_equation =   48.2196    0.0000
```

(3) 问题结果:

所求回归方程为 $\hat{y}=18.2642-0.8097x_1+0.0092x_2$, 即 $\hat{y}=18.2642-0.8097x+0.0092x^2$, 观测数据散点图与回归曲线图见图 4-4;

由于 F 检验法的 $p=0.0000<0.01=\alpha$, 故在 $\alpha=0.01$ 下, 认为此回归方程显著.

作业: $P_{196\text{-}200}$ 4.4, 4.8, 4.11

第五章　方差分析

一、方差分析模型与假设检验方法

（一）单因素方差分析

1．单因素方差分析数学模型

$$\begin{cases} X_{ij}=\mu_i+\varepsilon_{ij} \\ \varepsilon_{ij}\sim N(0,\sigma^2)\text{且相互独立} \end{cases} \quad (i=1,\cdots,r;\quad j=1,\cdots,n_i) \tag{5.1}$$

$(\mu_i$与σ^2未知)

2．假设检验

检验假设: $H_0:\mu_1=\mu_2=\cdots=\mu_r$

(1)总偏差平方和的分解式:

记: $\bar{X}_i=\dfrac{1}{n_i}\sum_{j=1}^{n_i}X_{ij}$ ——第 i 个总体的样本均值

$\bar{X}=\dfrac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i}X_{ij}$ ——样本的总均值

$S_T=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\bar{X})^2$ ——总偏差平方和, 反映样本波动程度的大小

$S_e=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2$ ——误差平方和 (组内平方和)

$S_A=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{X}_i-\bar{X})^2$ ——因素平方和 (组间平方和)

则 $$S_T=S_e+S_A \tag{5.2}$$

(2)检验统计量与拒绝域:

检验统计量: $F=\dfrac{S_A/(r-1)}{S_e/(n-r)}\overset{H_0\text{成立}}{\sim\sim\sim\sim}F(r-1,n-r)$

H_0的拒绝域: $F>F_{1-\alpha}(r-1,n-r)$

（二）双因素等重复试验方差分析

1．双因素等重复试验方差分析的数学模型

$$\begin{cases} X_{ijk}=\mu+\alpha_i+\beta_j+\gamma_{ij}+\varepsilon_{ijk} & \\ \sum_{i=1}^{r}\alpha_i=0,\quad \sum_{j=1}^{s}\beta_j=0,\quad \sum_{i=1}^{r}\gamma_{ij}=\sum_{j=1}^{s}\gamma_{ij}=0 & \begin{matrix}(i=1,\cdots,r;\\ j=1,\cdots,s;\\ k=1,\cdots,t)\end{matrix} \\ \varepsilon_{ijk}\sim N(0,\sigma^2)\text{且相互独立} & \end{cases} \tag{5.3}$$

$(\mu,\ \alpha_i,\ \beta_j,\ \gamma_{ij}$与$\sigma^2$未知)

2．假设检验

检验假设: $H_{01}:\alpha_1=\alpha_2=\cdots=\alpha_r=0$

$H_{02}:\beta_1=\beta_2=\cdots=\beta_s=0$

$H_{03}:\gamma_{ij}=0\quad (i=1,\cdots,r;\ j=1,\cdots,s)$

(1)总偏差平方和的分解式:

记: $\bar{X}=\frac{1}{rst}\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}X_{ijk}$，　$\bar{X}_{ij\cdot}=\frac{1}{t}\sum_{k=1}^{t}X_{ijk}$，　$\bar{X}_{i\cdot\cdot}=\frac{1}{st}\sum_{j=1}^{s}\sum_{k=1}^{t}X_{ijk}$，　$\bar{X}_{\cdot j\cdot}=\frac{1}{rt}\sum_{i=1}^{r}\sum_{k=1}^{t}X_{ijk}$

$S_T=\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}(X_{ijk}-\bar{X})^2$ ——总偏差平方和

$S_e=\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}(X_{ijk}-\bar{X}_{ij\cdot})^2$ ——误差平方和

$S_A=\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}(\bar{X}_{i\cdot\cdot}-\bar{X})^2$ ——因素 A 的平方和

$S_B=\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}(\bar{X}_{\cdot j\cdot}-\bar{X})^2$ ——因素 B 的平方和

$S_{A\times B}=\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}(\bar{X}_{ij\cdot}-\bar{X}_{i\cdot\cdot}-\bar{X}_{\cdot j\cdot}+\bar{X})^2$ ——交互作用 A×B 的平方和

则
$$S_T=S_e+S_A+S_B+S_{A\times B} \tag{5.4}$$

(2)检验统计量与拒绝域:

◆H_{01}的检验统计量: $F_A=\frac{S_A/(r-1)}{S_e/\left(rs(t-1)\right)}\overset{H_{01}\text{成立}}{\sim\sim\sim\sim}F\left(r-1,rs(t-1)\right)$

H_{01}的拒绝域: $F_A>F_{1-\alpha}\left(r-1,rs(t-1)\right)$;

◆H_{02}的检验统计量: $F_B=\frac{S_B/(s-1)}{S_e/\left(rs(t-1)\right)}\overset{H_{02}\text{成立}}{\sim\sim\sim\sim}F\left(s-1,rs(t-1)\right)$

H_{02}的拒绝域: $F_B>F_{1-\alpha}\left(s-1,rs(t-1)\right)$;

◆H_{03}的检验统计量: $F_{A\times B}=\frac{S_{A\times B}/\left((r-1)(s-1)\right)}{S_e/\left(rs(t-1)\right)}\overset{H_{03}\text{成立}}{\sim\sim\sim\sim}F\left((r-1)(s-1),rs(t-1)\right)$

H_{03}的拒绝域: $F_{A\times B}>F_{1-\alpha}\left((r-1)(s-1),rs(t-1)\right)$.

（三）双因素无交互作用的方差分析

1．双因素无交互作用方差分析的数学模型

$$\begin{cases} X_{ij}=\mu+\alpha_i+\beta_j+\varepsilon_{ij} \\ \sum_{i=1}^{r}\alpha_i=\sum_{j=1}^{s}\beta_j=0 \qquad (i=1,\cdots,r;\ j=1,\cdots,s) \\ \varepsilon_{ij}\sim N(0,\sigma^2)\text{且相互独立} \\ (\mu,\ \alpha_i,\ \beta_j\text{与}\sigma^2\text{未知}) \end{cases} \tag{5.5}$$

2．假设检验

检验假设: $H_{01}:\alpha_1=\alpha_2=\cdots=\alpha_r=0$

$H_{02}:\beta_1=\beta_2=\cdots=\beta_s=0$

(1)总偏差平方和的分解式:

记: $\bar{X}=\frac{1}{rs}\sum_{i=1}^{r}\sum_{j=1}^{s}X_{ij}$，　$\bar{X}_{i\cdot}=\frac{1}{s}\sum_{j=1}^{s}X_{ij}$，　$\bar{X}_{\cdot j}=\frac{1}{r}\sum_{i=1}^{r}X_{ij}$

$S_T=\sum_{i=1}^{r}\sum_{j=1}^{s}(X_{ij}-\bar{X})^2$ ——总偏差平方和

$S_A=\sum_{i=1}^{r}\sum_{j=1}^{s}(\bar{X}_{i\cdot}-\bar{X})^2$ ——因素 A 的平方和

$S_B=\sum_{i=1}^{r}\sum_{j=1}^{s}(\bar{X}_{\cdot j}-\bar{X})^2$ ——因素 B 的平方和

$S_{A\times B}=\sum_{i=1}^{r}\sum_{j=1}^{s}(X_{ij}-\bar{X}_{i\cdot}-\bar{X}_{\cdot j}+\bar{X})^2\triangleq S_e$——误差平方和

则　$S_T=S_e+S_A+S_B$　(5.6)

(2)检验统计量与拒绝域:

◆H_{01} 的检验统计量: $F_A=\frac{S_A/(r-1)}{S_e/\left((r-1)(s-1)\right)}\overset{H_{01}\text{成立}}{\sim\sim\sim\sim}F\left(r-1,(r-1)(s-1)\right)$

H_{01} 的拒绝域: $F_A>F_{1-\alpha}\left(r-1,(r-1)(s-1)\right)$;

◆H_{02} 的检验统计量: $F_B=\frac{S_B/(s-1)}{S_e/\left((r-1)(s-1)\right)}\overset{H_{02}\text{成立}}{\sim\sim\sim\sim}F\left(s-1,(r-1)(s-1)\right)$

H_{02} 的拒绝域: .

注意: 方差分析要求指标观测值必须满足: 独立、正态、方差齐性.

二、方差分析函数

（一）方差分析函数

表 5.1　方差分析部分分析函数

函数分类	函数名称	函数说明	调用格式
方差分析	anova1	单因素方差分析	[p,table,stats] = anova1(X,group)
	anova2	等重复(均衡)的双因素方差分析	[p,table,stats] = anova2(X,reps)
	multcompare	均值或其它估计量的多重比较	c = multcompare (stats,param1,val1,param2,val2,...)

（二）方差分析函数的格式说明及例题

1．[p,table,stats] = anova1(X,group)

功能: 进行均衡 (也叫等重复) 或不均衡 (也叫不等重复) 的单因素 r 个水平试验的方差分析. 默认返回两幅图表, 第一幅为标准方差分析表, 第二幅为各水平数据的盒形 (box) 图. 如果盒形图的中心线差别很大, 表示各水平下指标 X_i 的均值 μ_i 差异很大, 从而对应的 F 值很大, 相应的概率 p-值就小.

• 输入变量含义:

①X——如果 X 是矩阵, 如 m×r 矩阵, 则表示因素 A 包含 r 个水平, 第 i 列表示水平 A_i 下的样本, 每个样本包含 m 个相互独立的观察值(如果数据缺失, 用 NaN 补齐). 此时检验数据 X 中各列(各水平下指标)的均值是否相等.

如果 X 是向量, 如 1×n 的行向量, 则必须通过 group 来指明每个数据来自哪个水平. 此时检验各水平下指标的均值是否相等.

②group——用于分类, 指明观测数据来自于哪个类. group 可以是如下几种: 数值向量, 字符数组, 字符阵列, 分类数组, 逻辑向量.

如果 X 是矩阵, 如 m×r 矩阵, 则 group 是 1×r 向量, 指明 X 中各列(各水平)的名称. 若省略 group, 则默认值为[1,2,---,r].

如果 X 是向量, 如 1×n 的行向量, 则 group 必须是一个 1×n 的行向量, 用以指明 X 中每个数据来自哪个水平.

• 输出变量含义:

p——p 值, 即拒绝 $H_0:\mu_1=\mu_2=\cdots=\mu_r$(或 $H_0:\alpha_1=\alpha_2=\cdots=\alpha_r=0$)的最小的显著性水平. 对给定的显著性水平 α, 如果 $p\le\alpha$, 则拒绝 H_0;

table——以单元数组的形式返回方差分析表;

stats——利用 stats 可接下来进行多重比较检验. 用户可以将 stats 结构作为输入利用 multcompare 函数进行这种检验. stats 包含六个结果:

gnames: 各水平 A_i 的名称, 是 r 维向量;

n: 各水平 A_i 下的样本容量, 是 r 维向量;

source: 是'anova1';

means: 各水平下指标 X_i 的样本均值 $\bar{X}_i$, 是 r 维向量;

df: 误差平方和 S_e 的自由度;

s: 误差平方和 S_e 的均方开方, 即 $\sqrt{\frac{S_e}{n-r}}$.

【★例 5.3(P209)】对六种不同的农药在相同的条件下分别进行杀虫试验, 试验结果见下表. 问杀虫率是否因农药的不同而有显著的差异? ($\alpha = 0.01$)

试验号 \ 农药	A1	A2	A3	A4	A5	A6
1	87	90	56	55	92	75
2	85	88	62	48	99	72
3	80	87			95	81
4		94			91	

(1) 问题分析:

这里杀虫率是试验指标, 农药品种为因素, 这是一个单因素六水平试验. 我们用 $X_1, X_2, \cdots, X_6$ 分别表示六种不同农药的杀虫率, 即六个总体, 从而得到单因素方差分析模型:

$$\begin{cases} X_{ij} = \mu_i + \varepsilon_{ij} \\ \varepsilon_{ij} \sim N(0,\sigma^2) \text{且相互独立} \end{cases} \quad (i = 1,\cdots,r; \quad j = 1,\cdots,n_i)$$

(μ_i与σ^2未知)

问杀虫率是否因农药的不同而有显著的差异, 即假设检验:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_r .$$

此处, $r = 6, n_1 = 3, \cdots, n_6 = 3, n = \sum_{i=1}^{6} n_i = 18$.

(2) 问题求解:

选取检验统计量 $F = \frac{S_A/(r-1)}{S_e/(n-r)}$, 则 $H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$ 的拒绝域为 $F > F_{1-\alpha}(r-1, n-r)$. 若 $F > F_{1-\alpha}(r-1, n-r)$, 则认为杀虫率因农药的不同而差异显著.

• 编写命令文件 example5_3.m:

```
display('格式 1:----------------------------------------------------')
X=[87,90,56,55,92,75;
  85,88,62,48,99,72;
  80,87,NaN,NaN,95,81;
  NaN,94,NaN,NaN,91,NaN];
[p,table,stats] = anova1(X,{'农药 A1','农药 A2','农药 A3','农药 A4','农药 A5','农药 A6'});
display('格式 2:----------------------------------------------------')
x=[87,85,80,...
   90,88,87,94,...
   56,62,...
   55,48,...
   92,99,95,91,...
   75,72,81];
```

```
group={'农药 A1','农药 A1','农药 A1',...
    '农药 A2','农药 A2','农药 A2','农药 A2',...
    '农药 A3','农药 A3',...
    '农药 A4','农药 A4',...
    '农药 A5','农药 A5','农药 A5','农药 A5',...
    '农药 A6','农药 A6','农药 A6'};
[p,table,stats] = anova1(x,group);
```

• 运行命令文件 example5_3.m:

```
>> example5_3
```

ANOVA Table

Source	SS	df	MS	F	Prob>F
Groups	3794.5	5	758.9	51.16	1.11906e-007
Error	178	12	14.833		
Total	3972.5	17			

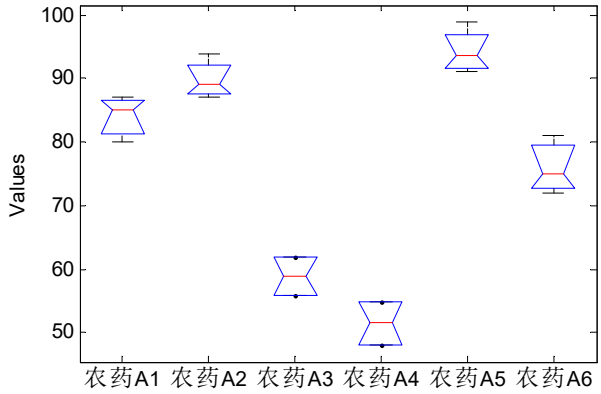


图 5-1(a)　各样本数据的盒形图

(3) 问题结果:

由于 $F = 51.16 > 5.0643 = F_{1-0.01}(5,12)$, 故拒绝 $H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$, 即认为不同的农药的杀虫率有显著的差异.

2． c = multcompare(stats,param1,val1,param2,val2,...)

功能: 在方差分析出现拒绝 H_0 的情况下, 需对均值或其它估计量进行多重比较. 利用方差分析命令中输出的 stats 结构所含信息进行多重比较检验. 此命令结果还显示一个交互式图形. 图中每组的均值用一个符号和符号周围的区间表示. 如果两个均值的区间不交叠, 说明它们显著不同; 如果两个区间交叠, 则说明它们不是显著不同的. 可以用鼠标选中任何一组, 图中其它任何与之显著不同的组将会高亮度显示.

• 输入变量含义:

①stats——方差分析函数 anova1 等的输出 stats 结构.

②各参数与参数值说明如下:

Parameter Name	Parameter Values
'alpha'	显著性水平 α

'displayopt'	'on'显示多重比较的交互式图形 (默认值) 'off'不显著多重比较的交互式图形
'ctype'	指定多重比较检验所选用的方法, 可取'hsd' (默认值)、'lsd'、'bonferroni'、'dunn-sidak'、'scheffe'
'dimension'	是一个向量(或数), 用于指明求 group 中哪些因素的边缘均值
'estimate'	指明 stats 来自何函数, 可取'anova1' (可省略)、'anova2'、'anovan' (可省略)、'aoctool'、'friedman' (可省略)、'kruskalwallis' (可省略)

• 输出变量含义:

c——返回成对比较的结果矩阵 c. 矩阵 c 的每一行对应一对均值比较, 每行包含 5 个数据. 第一、二数据表示比较组的编号, 第四个数据表示被比较组的均值差的点估计值, 第三、五个数据则表示均值差的置信区间.

【★例 5.3 续(P_{209})】请问例 5.3 中哪些水平间的差异是显著的? ($\alpha = 0.01$)

```
>>c = multcompare(stats,'alpha',0.01)
c =   1.0000    2.0000  -18.4403   -5.7500    6.9403
      1.0000    3.0000    9.8322   25.0000   40.1678
      1.0000    4.0000   17.3322   32.5000   47.6678
      1.0000    5.0000  -22.9403  -10.2500    2.4403
      1.0000    6.0000   -5.5665    8.0000   21.5665
      2.0000    3.0000   16.3605   30.7500   45.1395
      2.0000    4.0000   23.8605   38.2500   52.6395
      2.0000    5.0000  -16.2490   -4.5000    7.2490
      2.0000    6.0000    1.0597   13.7500   26.4403
      3.0000    4.0000   -9.1155    7.5000   24.1155
      3.0000    5.0000  -49.6395  -35.2500  -20.8605
      3.0000    6.0000  -32.1678  -17.0000   -1.8322
      4.0000    5.0000  -57.1395  -42.7500  -28.3605
      4.0000    6.0000  -39.6678  -24.5000   -9.3322
      5.0000    6.0000    5.5597   18.2500   30.9403
```

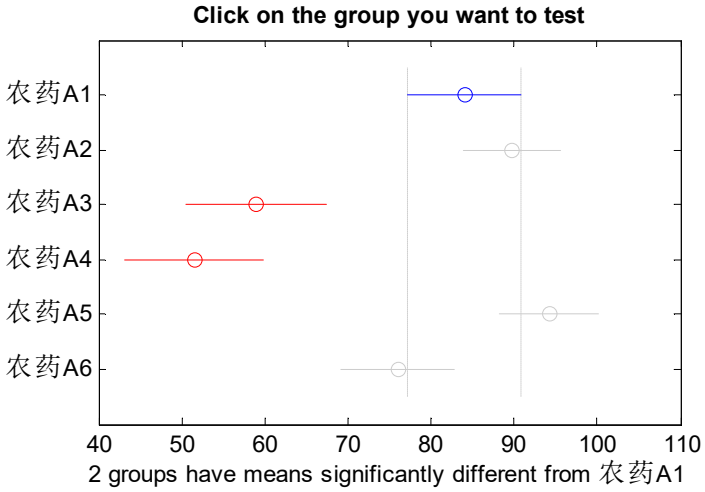


图 5-1(b)　多重比较的交互式图

从矩阵 c 可看出与图 5-1(b)都可看出, 在 $\alpha = 0.01$ 下, 水平 1-3, 1-4, 2-3, 2-4, 2-6, 3-5, 3-6, 4-5, 4-6, 5-6 各对之间差异是显著的. 显然, 第五种农药的杀虫率最高, 平均为 94.25%.

3．[p,table,stats] = anova2(X,reps)

功能: 进行均衡 (也叫等重复) 的双因素方差分析 (对于不均衡设计, 使用函数 anovan). X 为一矩阵. 比如下面这个矩阵, 其中行因素 A 有三个水平, 列因素 B 有两个水平, 每种水平组合下有两个观察值 (重复试验次数 reps=2). 下标分别表示行、列和每个水平对中的观察值.

$$\begin{array}{c} \\ A=1\left\{\right. \\ A=2\left\{\right. \\ A=3\left\{\right. \end{array} \begin{array}{c} \begin{array}{cc} B=1 & B=2 \end{array} \\ \left[\begin{array}{cc} x_{111} & x_{121} \\ x_{112} & x_{122} \\ x_{211} & x_{221} \\ x_{212} & x_{222} \\ x_{311} & x_{321} \\ x_{312} & x_{322} \end{array}\right] \end{array}$$

• 原假设说明:

设因素 A 有 r 个水平, 因素 B 有 s 个水平, 等重复试验了 t 次, 因素 A 的各水平效应记为 $\alpha_i\ (i=1,\cdots,r)$,因素 B 的各水平效应记为 $\beta_j\ (j=1,\cdots,s)$, 因素 A 与 B 的交互效应记为 $\gamma_{ij}\ (i=1,\cdots,r,\ j=1,\cdots,s)$, 则

当 reps=1 (缺省值) 时, anova2 返回的向量 p 中包含如下两个零假设的概率值:

$H_{0A}:\ \alpha_1=\alpha_2=\cdots=\alpha_r=0$;

$H_{0B}:\ \beta_1=\beta_2=\cdots=\beta_s=0$;

当 reps＞1 时, 还有一个原假设如下:

$H_{0AB}:\ \gamma_{ij}=0\ (i=1,\cdots,r,\ j=1,\cdots,s)$.

【★例 5.5(P_{224})】某农业研究所对三种小麦种子 A_1, A_2, A_3 和四种农肥 B_1, B_2, B_3, B_4 在相同的试验田里做试验, 结果列于下表, 表中数据是小麦的亩产量 (单位: kg) . 问小麦种子和农肥以及它们的交互作用对小麦产量有无显著的影响? ($\alpha = 0.01$)

种子＼农肥	B_1	B_2	B_3	B_4
A_1	173, 172	174, 176	177, 179	172, 173
A_2	175, 173	178, 177	174, 175	170, 171
A_3	177, 175	174, 174	174, 173	169, 169

(1) 问题分析:

这里小麦的亩产量是试验指标, 小麦种子为因素 A, 农肥为因素 B, 这是一个双因素 3×4 水平试验, 且为等重复的试验. 我们用 $X_{ij}\ (i=1,\cdots,3;\ j=1,\cdots,4)$ 表示水平组合 (A_i, B_j) 下的亩产量, 从而得到双因素方差分析模型:

$$\begin{cases} X_{ijk}=\mu+\alpha_i+\beta_j+\gamma_{ij}+\varepsilon_{ijk} & \\ \sum_{i=1}^{r}\alpha_i=0,\quad \sum_{j=1}^{s}\beta_j=0,\quad \sum_{i=1}^{r}\gamma_{ij}=\sum_{j=1}^{s}\gamma_{ij}=0 & (i=1,\cdots,r;\ j=1,\cdots,s;\ k=1,\cdots,t) \\ \varepsilon_{ijk}\sim N(0,\sigma^2)\text{且相互独立} & \end{cases}$$

$(\mu,\ \alpha_i,\ \beta_j,\ \gamma_{ij}$与$\sigma^2$未知)

问小麦种子、农肥、以及它们的交互作用对小麦产量有无显著的影响, 即假设检验:

H_{0A}: $\alpha_1=\cdots=\alpha_r=0$;

H_{0B}: $\beta_1=\cdots=\beta_s=0$;

$H_{0A\times B}$: $\gamma_{ij}=0\ (i=1,\cdots,r,\ j=1,\cdots,s)$.

此处, $r=3, s=4, t=2$.

(2) 问题求解:

对 H_{0A}, 选取检验统计量 $F_A=\dfrac{S_A/(r-1)}{S_e/\left(rs(t-1)\right)}$, 则 H_{0A} 的拒绝域为 $F_A>F_{1-\alpha}\left(r-1,rs(t-1)\right)$. 若 $F_A>F_{1-\alpha}\left(r-1,rs(t-1)\right)$, 则认为因素 A(小麦品种)对小麦亩产量的影响显著;

对 H_{0B}, 选取检验统计量 $F_B=\dfrac{S_B/(s-1)}{S_e/\left(rs(t-1)\right)}$, 则 H_{0B} 的拒绝域为 $F_B>F_{1-\alpha}\left(s-1,rs(t-1)\right)$. 若 $F_B>F_{1-\alpha}\left(s-1,rs(t-1)\right)$, 则认为因素 B(农肥)对小麦亩产量的影响显著;

对 $H_{0A\times B}$, 选取检验统计量 $F_{A\times B}=\dfrac{S_{A\times B}/\left((r-1)(s-1)\right)}{S_e/\left(rs(t-1)\right)}$, 则 $H_{0A\times B}$ 的拒绝域为 $F_{A\times B}>F_{1-\alpha}\left((r-1)(s-1),rs(t-1)\right)$. 若 $F_{A\times B}>F_{1-\alpha}\left((r-1)(s-1),rs(t-1)\right)$, 则认为小麦品种与农肥的交互作用对小麦亩产量的影响显著.

• 编写命令文件 example5_5.m:

```
X=[173,174,177,172;
   172,176,179,173;
   175,178,174,170;
   173,177,175,171;
   177,174,174,169;
   175,174,173,169];
t=2;    %t 表示试验重复次数
[p,table,stats] = anova2(X,t);
```

• 运行命令文件 example5_5.m:

```
>> example5_5
```

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	90.833	3	30.2778	33.03	0
Rows	8.083	2	4.0417	4.41	0.0367
Interaction	51.917	6	8.6528	9.44	0.0006
Error	11	12	0.9167		
Total	161.833	23			

(3) 问题结果:

因为 $p_A = 0.0367 > \alpha = 0.01$, 所以不拒绝 H_{0A}, 即小麦种子对产量的影响不显著;

因为 $p_B = 0 < \alpha = 0.01$, 所以拒绝 H_{0B}, 即肥料对产量的影响高度显著;

因为 $p_{A\times B} = 0.0006 < \alpha = 0.01$, 所以拒绝 $H_{0A\times B}$, 即种子与肥料的交互作用对产量的影响高度显著.

由于 (A_1, B_3) 的产量最高, 因此应把 A_1 和 B_3 搭配起来生产.

【★例 5.6(P_{231})】四个工人轮流在三台机床上加工某种零件, 下表中的数据表示他们在相同规定时间内加工出的合格品件数. 问四个工人的技术是否有显著差异; 又三台机床的性能是否有显著差异? ($\alpha = 0.05$)

工人 \ 机床	B_1	B_2	B_3
A_1	8	3	7
A_2	10	4	8
A_3	6	5	6
A_4	8	4	7

(1) 问题分析:

这里合格品件数是试验指标, 工人技术为因素 A, 机床性能为因素 B, 这是一个双因素 4×3 水平试验, 且为无重复的试验. 我们用 X_{ij} $(i = 1,\cdots,4; j = 1,\cdots,3)$ 表示水平组合 (A_i, B_j) 下的合格品件数, 从而得到双因素无交互作用的方差分析模型:

$$\begin{cases} X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \\ \sum_{i=1}^{r} \alpha_i = \sum_{j=1}^{s} \beta_j = 0 \qquad (i = 1,\cdots,r;\ j = 1,\cdots,s) \\ \varepsilon_{ij} \sim N(0, \sigma^2) \text{且相互独立} \end{cases}$$

$(\mu,\ \alpha_i,\ \beta_j$ 与 σ^2 未知)

问工人技术、机床性能对合格品件数有无显著的影响, 即假设检验:

$$H_{0A}:\ \alpha_1 = \cdots = \alpha_r = 0;$$

$$H_{0B}:\ \beta_1 = \cdots = \beta_s = 0.$$

此处, $r = 4, s = 3$.

(2) 问题求解: 记 $S_e = S_{A\times B}$,

对 H_{0A}，选取检验统计量 $F_A=\dfrac{S_A/(r-1)}{S_e/\big((r-1)(s-1)\big)}$，则 H_{0A} 的拒绝域为 $F_A>F_{1-\alpha}\big(r-1,(r-1)(s-1)\big)$. 若 $F_A>F_{1-\alpha}\big(r-1,(r-1)(s-1)\big)$, 则认为因素 A(工人技术)对合格品件数的影响显著;

对 H_{0B}，选取检验统计量 $F_B=\dfrac{S_B/(s-1)}{S_e/\big((r-1)(s-1)\big)}$，则 H_{0B} 的拒绝域为 $F_B>F_{1-\alpha}\big(s-1,(r-1)(s-1)\big)$. 若 $F_B>F_{1-\alpha}\big(s-1,(r-1)(s-1)\big)$, 则认为因素 B(机床性能)对合格品件数的影响显著.

• 编写命令文件 example5_6.m:

```
X=[8,3,7;
   10,4,8;
    6,5,6;
    8,4,7];
[p,table,stats] = anova2(X);
```

• 运行命令文件 example5_6.m:

```
>> example5_6
```

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	34.6667	2	17.3333	14.18	0.0053
Rows	4.6667	3	1.5556	1.27	0.3654
Error	7.3333	6	1.2222		
Total	46.6667	11			

因为 $p_A=0.3654>\alpha=0.05$, 所以不拒绝 H_{0A}, 即这四个工人的技术没有显著差别;

因为 $p_B=0.0053<\alpha=0.05$, 所以拒绝 H_{0B}, 即这三台机床的性能存在显著差异.

作业: $P_{232\text{-}234}$ 5.1, 5.5, 5.8