# 第八章　统计方法建模

数理统计研究的对象是受随机因素影响的数据，以下数理统计就简称统计，统计是以概率论为基础的一门应用学科。

数据样本少则几个，多则成千上万，人们希望能用少数几个包含其最多相关信息的数值来体现数据样本总体的规律。描述性统计就是搜集、整理、加工和分析统计数据，使之系统化、条理化，以显示出数据资料的趋势、特征和数量关系。它是统计推断的基础，实用性较强，在统计工作中经常使用。

面对一批数据如何进行描述与分析，需要掌握参数估计和假设检验这两个数理统计的最基本方法。

我们将用 Matlab 的统计工具箱(Statistics Toolbox)来实现数据的统计描述和分析。

## §1 统计的基本概念

1.1 总体和样本

**总体**是人们研究对象的全体，又称母体，如工厂一天生产的全部产品（按合格品及废品分类），学校全体学生的身高。

总体中的每一个基本单位称为个体，个体的特征用一个变量（如 $x$ ）来表示，如一件产品是合格品记 $x=0$，是废品记 $x=1$；一个身高 170（cm）的学生记 $x=170$。

从总体中随机产生的若干个个体的集合称为样本，或子样，如 $n$ 件产品，100 名学生的身高，或者一根轴直径的 10 次测量。实际上这就是从总体中随机取得的一批数据，不妨记作 $x_1,x_2\cdots x_n$，$n$ 称为样本容量。

简单地说，统计的任务是由样本推断总体。

1.2 频数表和直方图

一组数据（样本）往往是杂乱无章的，作出它的频数表和直方图，可以看作是对这组数据的一个初步整理和直观描述。

将数据的取值范围划分为若干个区间，然后统计这组数据在每个区间中出现的次数，称为频数，由此得到一个频数表。以数据的取值为横坐标，频数为纵坐标，画出一个阶梯形的图，称为直方图，或频数分布图。

若样本容量不大，能够手工作出频数表和直方图，当样本容量较大时则可以借助 Matlab 这样的软件了。让我们以下面的例子为例，介绍频数表和直方图的作法。

例 1 学生的身高和体重

学校随机抽取 100 名学生，测量他们的身高和体重，所得数据如表

| 身高 体重 | 身高 体重 | 身高 体重 | 身高 体重 | 身高 体重 |
|---|---|---|---|---|
| 172 75 | 169 55 | 169 64 | 171 65 | 167 47 |
| 171 62 | 168 67 | 165 52 | 169 62 | 168 65 |
| 166 62 | 168 65 | 164 59 | 170 58 | 165 64 |
| 160 55 | 175 67 | 173 74 | 172 64 | 168 57 |
| 155 57 | 176 64 | 172 69 | 169 58 | 176 57 |
| 173 58 | 168 50 | 169 52 | 167 72 | 170 57 |
| 166 55 | 161 49 | 173 57 | 175 76 | 158 51 |
| 170 63 | 169 63 | 173 61 | 164 59 | 165 62 |
| 167 53 | 171 61 | 166 70 | 166 63 | 172 53 |
| 173 60 | 178 64 | 163 57 | 169 54 | 169 66 |

| | | | | |
|---|---|---|---|---|
| 178 60 | 177 66 | 170 56 | 167 54 | 169 58 |
| 173 73 | 170 58 | 160 65 | 179 62 | 172 50 |
| 163 47 | 173 67 | 165 58 | 176 63 | 162 52 |
| 165 66 | 172 59 | 177 66 | 182 69 | 175 75 |
| 170 60 | 170 62 | 169 63 | 186 77 | 174 66 |
| 163 50 | 172 59 | 176 60 | 166 76 | 167 63 |
| 172 57 | 177 58 | 177 67 | 169 72 | 166 50 |
| 182 63 | 176 68 | 172 56 | 173 59 | 174 64 |
| 171 59 | 175 68 | 165 56 | 169 65 | 168 62 |
| 177 64 | 184 70 | 166 49 | 171 71 | 170 59 |

（i） 数据输入

数据输入通常有两种方法，一种是在交互环境中直接输入，如果在统计中数据量比较大，这样作不太方便；另一种办法是先把数据写入一个纯文本数据文件 data.txt 中，格式如例 1 的表格，有 20 行、10 列，数据列之间用空格键或 Tab 键分割，该数据文件 data.txt 存放在 matlab\work 子目录下，在 Matlab 中用 load 命令读入数据，具体作法是：

load data.txt

这样在内存中建立了一个变量 data，它是一个包含有 $20\times10$ 个数据的矩阵。

为了得到我们需要的 100 个身高和体重各为一列的矩阵，应做如下的改变：

```
high=data(:,1:2:9);high=high(:)
weight=data(:,2:2:10);weight=weight(:)
```

（ii）作频数表及直方图

用 hist 命令实现，其用法是：

[N,X] = hist(Y,M)

数组（行、列均可）Y 的频数表。它将区间[min(Y),max(Y)]等分为 M 份（缺省时 M 设定为 10），N 返回 M 个小区间的频数，X 返回 M 个小区间的中点。

hist(Y,M)

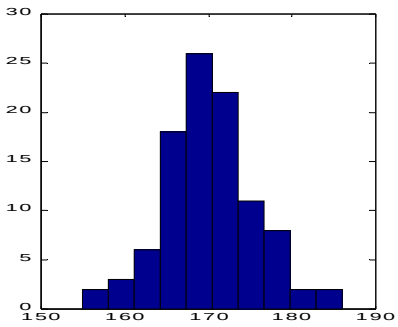
数组 Y 的直方图。

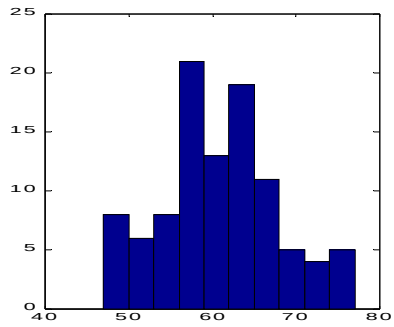
对于例 1 的数据，编写程序如下：

```
load data.txt;
high=data(:,1:2:9);high=high(:);
weight=data(:,2:2:10);weight=weight(:);
[n1,x1]=hist(high)
%下面语句与hist命令等价
%n1=[length(find(high<158.1)),...
%   length(find(high>=158.1&high<161.2)),...
%   length(find(high>=161.2&high<164.5)),...
%   length(find(high>=164.5&high<167.6)),...
%   length(find(high>=167.6&high<170.7)),...
%   length(find(high>=170.7&high<173.8)),...
%   length(find(high>=173.8&high<176.9)),...
%   length(find(high>=176.9&high<180)),...
%   length(find(high>=180&high<183.1)),...
%   length(find(high>=183.1))]
[n2,x2]=hist(weight)
subplot(1,2,1)
hist(high)
subplot(1,2,2)
```

```
hist(weight)
```

计算结果略，直方图如下图所示：




从直方图上可以看出，身高的分布大致呈中间高、两端低的钟形；而体重则看不出什么规律。要想从数值上给出更确切的描述，需要进一步研究反映数据特征的所谓“统计量”。直方图所展示的身高的分布形状可看作正态分布，当然也可以用这组数据对分布作假设检验。

例 2 统计下列五行字符串中字符 a、g、c、t 出现的频数

1.aggcacggaaaaacgggaataacggaggaggacttggcacggcattacacggagg
2.cggaggacaaacgggatggcggtattggaggtggcggactgttcgggga
3.gggacggatacggattctggccacggacggaaaggaggacacggcggacataca
4.atggataacggaaacaaaccagacaaacttcggtagaaatacagaagctta
5.cggctggcggacaacggactggcggattccaaaaacggaggaggcggacggaggc

解　把上述五行复制到一个纯文本数据文件 shuju.txt 中，放在 matlab\work 子目录下，编写如下程序：

```
clc
fid1=fopen('shuju.txt','r');
i=1;
while (~feof(fid1))
data=fgetl(fid1);
a=length(find(data==97));
b=length(find(data==99));
c=length(find(data==103));
d=length(find(data==116));
e=length(find(data>=97&data<=122));
f(i,:)=[a  b  c  d  e  a+b+c+d];
i=i+1;
end
f
he=[sum(f(:,1))  sum(f(:,2))  sum(f(:,3))  sum(f(:,4))...
    sum(f(:,5))  sum(f(:,6))]
fid2=fopen('pinshu.txt','w');
fprintf(fid2,'%8d  %8d  %8d  %8d  %8d  %8d\n',f');
fclose(fid1);fclose(fid2);
```

我们把统计结果最后写到一个纯文本文件 pinshu.txt 中，在程序中多引进了几个变量，是为了检验字符串是否只包含 a、g、c、t 四个字符。

1.3 统计量

假设有一个容量为$n$的样本（即一组数据），记作，需要对它进行一定的加工，才能提出有用的信息，用作对总体（分布）参数的估计和检验。**统计量**就是加工出来的、反映样本数量特征的函数，它不含任何未知量。

下面我们介绍几种常用的统计量。

（i）表示位置的统计量—算术平均值和中位数

算术平均值（简称均值）描述数据取值的平均位置，记作$\bar{x}$，

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

**中位数**是将数据由小到大排序后位于中间位置的那个数值。

Matlab 中 mean(x)返回 x 的均值，median(x)返回中位数。

（ii）表示变异程度的统计量—标准差、方差和极差

**标准差**$s$定义为

$$s = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{\frac{1}{2}} \qquad (2)$$

它是各个数据与均值偏离程度的度量，这种偏离不妨称为变异。

**方差**是标准差的平方$s^2$。

**极差**是的最大值与最小值之差。

Matlab 中 std(x)返回 x 的标准差，var(x)返回方差，range(x)返回极差。

你可能注意到标准差$s$的定义（2）中，对$n$个$(x_i - x)$的平方求和，却被$(n-1)$除，这是出于无偏估计的要求。若需要改为被$n$除，Matlab 可用 std(x,1)和 var(x,1)来实现。

（iii）中心矩、表示分布形状的统计量—偏度和峰度

随机变量$x$的$r$阶**中心矩**为$E(x-Ex)^r$。

随机变量$x$的偏度和峰度指的是$x$的标准化变量的三阶中心矩和四阶中心矩：

$$\nu_1 = E\left[\left(\frac{x - E(x)}{\sqrt{D(x)}}\right)^3\right] = \frac{E\left[(x - E(x))^3\right]}{(D(x))^{3/2}},$$

$$\nu_2 = E\left[\left(\frac{x - E(x)}{\sqrt{D(x)}}\right)^4\right] = \frac{E\left[(x - E(x))^4\right]}{(D(x))^2}.$$

偏度反映分布的对称性，$\nu_1 > 0$称为右偏态，此时数据位于均值右边的比位于左边的多；$\nu_1 < 0$称为左偏态，情况相反；而$\nu_1$接近 0 则可认为分布是对称的。

峰度是分布形状的另一种度量，正态分布的峰度为 3，若$\nu_2$比 3 大得多，表示分布有沉重的尾巴，说明样本中含有较多远离均值的数据，因而峰度可以用作衡量偏离正态分布的尺度之一。

Matlab 中 moment(x,order)返回 x 的 order 阶中心矩，order 为中心矩的阶数。skewness(x)返回 x 的偏度，kurtosis(x)返回峰度。

在以上用 Matlab 计算各个统计量的命令中，若 x 为矩阵，则作用于 x 的列，返回一个行向量。

对例 1 给出的学生身高和体重，用 Matlab 计算这些统计量，程序如下：

```
clc
load data.txt;
high=data(:,1:2:9);high=high(:);
weight=data(:,2:2:10);weight=weight(:);
shuju=[high weight];
jun_zhi=mean([high weight])
zhong_wei_shu=median(shuju)
biao_zhun_cha=std(shuju)
ji_cha=range(shuju)
pian_du=skewness(shuju)
feng_du=kurtosis(shuju)
```

统计量中最重要、最常用的是均值和标准差，由于样本是随机变量，它们作为样本的函数自然也是随机变量，当用它们去推断总体时，有多大的可靠性就与统计量的概率分布有关，因此我们需要知道几个重要分布的简单性质。

### 1.4 统计中几个重要的概率分布

#### 1.4.1 分布函数、密度函数和分位数

随机变量的特性完全由它的（概率）分布函数或（概率）密度函数来描述。设有随机变量 $X$，其分布函数定义为 $X \le x$ 的概率，即。若 $X$ 是连续型随机变量，则其密度函数 $p(x)$ 与 $F(x)$ 的关系为

分位数是下面常用的一个概念，其定义为：对于 $0<\alpha<1$，使某分布函数 $F(x)=\alpha$ 的 $x$，成为这个分布的 $\alpha$ 分位数，记作 $x_\alpha$。

我们前面画过的直方图是频数分布图，频数除以样本容量 $n$，称为频率，$n$ 充分大时频率是概率的近似，因此直方图可以看作密度函数图形的（离散化）近似。

#### 1.4.2 统计中几个重要的概率分布

（i）正态分布

正态分布随机变量 $X$ 的密度函数曲线呈中间高两边低、对称的钟形，期望（均值）$EX=\mu$，方差 $DX=\sigma^2$，记作，$\sigma$ 称均方差或标准差，当 $\mu=0,\sigma=1$ 时称为标准正态分布，记作 $X \sim N(0,1)$。正态分布完全由均值 $\mu$ 和方差 $\sigma^2$ 决定，它的偏度为 0，峰度为 3。

正态分布可以说是最常见的（连续型）概率分布，成批生产时零件的尺寸，射击中弹着点的位置，仪器反复量测的结果，自然界中一种生物的数量特征等，多数情况下都服从正态分布，这不仅是观察和经验的总结，而且有着深刻的理论依据，即在大量相互独立的、作用差不多大的随机因素影响下形成的随机变量，其极限分布为正态分布。

鉴于正态分布的随机变量在实际生活中如此地常见，记住下面 3 个数字是有用的：

68%的数值落在距均值左右 1 个标准差的范围内，即

$$P\{\mu-\sigma \le X \le \mu+\sigma\}=0.68;$$

95%的数值落在距均值左右 2 个标准差的范围内，即

$$P\{\mu-2\sigma\le X\le\mu+2\sigma\}=0.95;$$

99.7%的数值落在距均值左右 3 个标准差的范围内，即

$$P\{\mu-3\sigma\le X\le\mu+3\sigma\}=0.997.$$

（ii）$\chi^2$ 分布(Chi square)

若为相互独立的、服从标准正态分布 $N(0,1)$ 的随机变量，则它们的平方和 $Y=\sum_{i=1}^{n}X_i^2$ 服从 $\chi^2$ 分布，记作 $Y\sim\chi^2(n)$，$n$ 称自由度，它的期望 $EY=n$，方差 $DY=2n$。

（iii）$t$ 分布

若 $X\sim N(0,1)$，$Y\sim\chi^2(n)$，且相互独立，则 $T=\dfrac{X}{\sqrt{Y/n}}$ 服从 $t$ 分布，记作 $T\sim t(n)$，$n$ 称自由度。$t$ 分布又称学生氏(Student)分布。

$t$ 分布的密度函数曲线和 $N(0,1)$ 曲线形状相似。理论上 $n\to\infty$ 时，，实际上当 $n>30$ 时它与 $N(0,1)$ 就相差无几了。

（iv）$F$ 分布

若 $X\sim\chi^2(n_1)$，$Y\sim\chi^2(n_2)$，且相互独立，则 $F=\dfrac{X/n_1}{Y/n_2}$ 服从 $F$ 分布，记作 $F\sim F(n_1,n_2)$，$(n_1,n_2)$ 称自由度。

1.4.3 Matlab 统计工具箱(Toolbox\Stats)中的概率分布

Matlab 统计工具箱中有 20 种概率分布，这里只对上面所述 4 种分布列出命令的字符：

Norm 正态分布；　chi2 $\chi^2$ 分布；

t $t$ 分布　　f $F$ 分布

工具箱对每一种分布都提供 5 类函数，其命令的字符是：

pdf 概率密度；　cdf 分布函数；　inv 分布函数的反函数；

stat 均值与方差；　rnd 随机数生成

当需要一种分布的某一类函数时，将以上所列的分布命令字符与函数命令字符接起来，并输入自变量（可以是标量、数组或矩阵）和参数就行了，如：

p=normpdf(x,mu,sigma)　均值 mu、标准差 sigma 的正态分布在 $x$ 的密度函数(mu=0，sigma=1 时可缺省)。

p=tcdf(x,n)　$t$ 分布（自由度 n）在 x 的分布函数。

x=chi2inv(p,n)　$\chi^2$ 分布（自由度 n）使分布函数 F(x)=p 的 x(即 P 分位数)。

[m,v]=fstat(n1,n2)　$F$ 分布（自由度 n1,n2）的均值 m 和方差 v。

几个分布的密度函数图形就可以用这些命令作出，如：

```
x=-6:0.01:6;y=normpdf(x);z=normpdf(x,0,2);
plot(x,y,x,z),gtext('N(0,1)'),gtext('N(0,2^2)')
```

分布函数的反函数的意义从下例看出：

```
x=chi2inv(0.9,10)
x =  15.9872
```

如果反过来计算，则

```
P=chi2cdf(15.9872,10)   P =   0.9000
```

1.5 正态总体统计量的分布

用样本来推断总体，需要知道样本统计量的分布，而样本又是一组与总体同分布的随机变量，所以样本统计量的分布依赖于总体的分布。当总体服从一般的分布时，求某个样本统计量的分布是很困难的，只有在总体服从正态分布时，一些重要的样本统计量（均值、标准差）的分布才有便于使用的结果。另一方面，现实生活中需要进行统计推断的总体，多数可以认为服从（或近似服从）正态分布，所以统计中人们在正态总体的假定下研究统计量的分布，是必要的与合理的。

设总体，$x_1,x_2,\cdots,x_n$ 为一容量 $n$ 的样本，其均值 $\bar{x}$ 和标准差 $s$ 由式（1）、（2）确定，则用 $\bar{x}$ 和 $s$ 构造的下面几个分布在统计中是非常有用的。

$$\bar{x} \sim N(\mu, \frac{\sigma^2}{n}) \quad 或 \quad \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \tag{3}$$

$$\frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t(n-1) \tag{4}$$

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1). \tag{5}$$

设有两个总体和，及由容量分别为 $n_1$，$n_2$ 的两个样本确定的均值 $\bar{x},\bar{y}$ 和标准差 $s_1,s_2$，则

$$\frac{(\bar{x}-\mu_1)-(\bar{y}-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}} \sim N(0,1) \tag{6}$$

$$\frac{(\bar{x}-\mu_1)-(\bar{y}-\mu_2)}{\sqrt{s^2/n_1+s^2/n_2}} \sim t(n_1+n_2-2) \tag{7}$$

其中 $s^2=\dfrac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$，

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F(n_1-1,n_2-1) \tag{8}$$

对于（7）式，假定 $\sigma_1=\sigma_2$，但它们未知，于是用 $s$ 代替。在下面的统计推断中我们要反复用到这些分布。

## §2 参数估计

利用样本对总体进行统计推断的一类问题是参数估计，即假定已知总体的分布，通常是，估计参数的分布，如 $\mu,\sigma^2$。参数估计分点估计和区间估计两种。

2.1 点估计

点估计是用样本统计量确定总体参数的一个数值。评价估计优劣的标准有无偏性、最小方差性、有效性等，估计的方法有矩法、极大似然法等。

最常用的是对总体均值 $\mu$ 和 $\sigma^2$ 方差（或标准差 $\sigma$）作点估计。让我们暂时抛开评价标准，当从一个样本按照式（1）、（2）算出样本均值 $\bar{x}$ 和方差 $s^2$ 后，对 $\mu$ 和 $\sigma^2$（或 $\sigma$）一个自然、合理的点估计显然是（在字母上加^表示它的估计值）

$$\hat{\mu}=\bar{x},\ \hat{\sigma}^2=s^2\ \ \hat{\sigma}=s \qquad (9)$$

2.2 区间估计

点估计虽然给出了待估参数的一个数值，却没有告诉我们这个估计值的精度和可信程度。一般地，总体的待估参数记作$\theta$（如$\mu,\sigma^2$），由样本算出的$\theta$的估计量记作$\hat{\theta}$，人们常希望给出一个区间$[\hat{\theta}_1,\hat{\theta}_2]$，使$\theta$以一定的概率落在此区间内。若有

，

$$0<\alpha<1 \qquad (10)$$

则$[\hat{\theta}_1,\hat{\theta}_2]$称为$\theta$的置信区间，$\hat{\theta}_1,\hat{\theta}_2$分别称为置信下限和置信上限，$1-\alpha$称为置信概率或置信水平，$\alpha$称为显著性水平。

给出了置信区间$[\hat{\theta}_1,\hat{\theta}_2]$和置信水平$1-\alpha$的估计，称为$\theta$的区间估计。置信区间越小，估计的精度越高；置信水平越大，估计的可信程度越高。但是这两个指标显然是矛盾的，通常是在一定的置信水平下使置信区间尽量小。通俗地说，区间估计给出了点估计的误差范围。

2.3 参数估计的Matlab实现

Matlab统计工具箱中，有专门计算总体均值、标准差的点估计和区间估计的函数。对于正态总体，命令是

[mu,sigma,muci,sigmaci]=normfit(x,alpha)

其中x为样本（数组或矩阵），alpha为显著性水平$\alpha$（alpha缺省时设定为0.05），返回总体均值$\mu$和标准差$\sigma$的点估计mu和sigma，及总体均值$\mu$和标准差$\sigma$的区间估计muci和sigmaci。当x为矩阵时返回行向量。

Matlab统计工具箱中还提供了一些具有特定分布总体的区间估计的命令，如expfit，poissfit，gamfit，你可以从这些字头猜出它们用于哪个分布，具体用法参见帮助系统。

## §3 假设检验

统计推断的另一类重要问题是假设检验问题。在总体的分布函数完全未知或只知其形式但不知其参数的情况，为了推断总体的某些性质，提出某些关于总体的假设。例如，提出总体服从泊松分布的假设，又如对于正态总体提出数学期望等于$\mu_0$的假设等。假设检验就是根据样本对所提出的假设做出判断：是接受还是拒绝。这就是所谓的假设检验问题。

3.1 单个总体$N(\mu,\sigma^2)$均值$\mu$的检验

原假设（或零假设）为：$H_0:\mu=\mu_0$。

备选假设有三种可能：

$H_1:\mu\neq\mu_0$；$H_1:\mu>\mu_0$；$H_1:\mu<\mu_0$。

3.1.1 $\sigma^2$已知，关于$\mu$的检验（$u$检验）

在Matlab中$u$检验法由函数ztest来实现，命令为

[h,p,ci]=ztest(x,mu,sigma,alpha,tail)

其中输入参数x是样本，mu是$H_0$中的$\mu_0$，sigma是总体标准差$\sigma$，alpha是显著性水平$\alpha$（alpha缺省时设定为0.05），tail是对备选假设$H_1$的选择：$H_1$为$\mu\neq\mu_0$时用tail=0（可缺省）；$H_1$为$\mu>\mu_0$时用tail=1；$H_1$为$\mu<\mu_0$时用tail=-1。输

出参数 h=0 表示接受 $H_0$，h=1 表示拒绝 $H_0$，p 表示在假设 $H_0$ 下样本均值出现的概率，p 越小 $H_0$ 越值得怀疑，ci 是 $\mu_0$ 的置信区间。

例 3　某车间用一台包装机包装糖果。包得的袋装糖重是一个随机变量，它服从正态分布。当机器正常时，其均值为 0.5 公斤，标准差为 0.015 公斤。某日开工后为检验包装机是否正常，随机地抽取它所包装的糖 9 袋，称得净重为（公斤）：

0.497　0.506　0.518　0.524　0.498　0.511　0.520　0.515　0.512

问机器是否正常？

解　总体 $\sigma$ 已知，，$\mu$ 未知。于是提出假设和 $H_1:\mu\neq 0.5$。

Matlab 实现如下：

```
x=[0.497   0.506   0.518   0.524   0.498...
  0.511  0.520  0.515  0.512];
[h,p,ci]=ztest(x,0.5,0.015)
```

求得 h=1，p=0.0248，说明在 0.05 的水平下，可拒绝原假设，即认为这天包装机工作不正常。

3.1.2　$\sigma^2$ 未知，关于 $\mu$ 的检验（$t$ 检验）

在 Matlab 中 $t$ 检验法由函数 ttest 来实现，命令为

[h,p,ci]=ttest(x,mu,alpha,tail)

例 4　某种电子元件的寿命 $x$（以小时计）服从正态分布，$\mu,\sigma^2$ 均未知. 现得 16 只元件的寿命如下：

159　280　101　212　224　379　179　264

222　362　168　250　149　260　485　170

问是否有理由认为元件的平均寿命大于 225（小时）？

解　按题意需检验

$$H_0:\mu\le\mu_0=225,\quad H_1:\mu>225,$$

取 $\alpha=0.05$。Matlab 实现如下：

```
x=[159   280   101   212   224   379   179   264 ...
  222   362   168   250   149   260   485   170];
[h,p,ci]=ttest(x,225,0.05,1)
```

求得 h=0，p=0.2570，说明在显著水平为 0.05 的情况下，不能拒绝原假设，认为元件的平均寿命不大于 225 小时。

3.2　两个正态总体均值差的检验（$t$ 检验）

还可以用 $t$ 检验法检验具有相同方差的 2 个正态总体均值差的假设。在 Matlab 中由函数 ttest2 实现，命令为：

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

与上面的 ttest 相比，不同处只在于输入的是两个样本 x,y（长度不一定相同），而不是一个样本和它的总体均值；tail 的用法与 ttest 相似，可参看帮助系统。

例 5 在平炉上进行一项试验以确定改变操作方法的建议是否会增加钢的得率,试验是在同一平炉上进行的。每炼一炉钢时除操作方法外,其它条件都可能做到相同。先用标准方法炼一炉,然后用建议的新方法炼一炉,以后交换进行,各炼了 10 炉,其得率分别为

1° 标准方法　78.1　72.4　76.2　74.3　77.4　78.4　76.0　75.6　76.7　77.3

2° 新方法　　79.1　81.0　77.3　79.1　80.0　79.1　79.1　77.3　80.2　82.1

设这两个样本相互独立且分别来自正态总体$N(\mu_1,\sigma^2)$和$N(\mu_2,\sigma^2)$，$\mu_1,\mu_2,\sigma^2$均未知，问建议的新方法能否提高得率?(取$\alpha=0.05$。)

解　(i) 需要检验假设

，.

(ii)Matlab 实现

```
x=[78.1  72.4  76.2  74.3  77.4  78.4  76.0  75.6  76.7
77.3];
y=[79.1  81.0  77.3  79.1  80.0  79.1  79.1  77.3  80.2
82.1];
[h,p,ci]=ttest2(x,y,0.05,-1)
```

求得 h=1, p=2.2126×$10^{-4}$。表明在$\alpha=0.05$的显著水平下，可以拒绝原假设，即认为建议的新操作方法较原方法优。

3.3　分布拟合检验

在实际问题中，有时不能预知总体服从什么类型的分布，这时就需要根据样本来检验关于分布的假设。下面介绍$\chi^2$检验法和专用于检验分布是否为正态的“偏峰、峰度检验法”。

3.3.1　$\chi^2$检验法

$H_0$：总体$x$的分布函数为$F(x)$,

$H_1$：总体$x$的分布函数不是$F(x)$.

在用下述$\chi^2$检验法检验假设$H_0$时，若在假设$H_0$下$F(x)$的形式已知，但其参数值未知，这时需要先用极大似然估计法估计参数，然后作检验。

$\chi^2$检验法的基本思想如下：将随机试验可能结果的全体$\Omega$分为$k$个互不相容的事件$(\sum_{i=1}^{k} A_k=\Omega, A_iA_j=\Phi, i\neq j, i,j=1,2,\cdots,k)$。于是在假设$H_0$下，我们可以计算$p_i=P(A_i)$（或$\hat{p}_i=\hat{P}(A_i)$），$i=1,2,...,k$。在$n$次试验中，事件$A_i$出现的频率$f_i/n$与$p_i$（$\hat{p}_i$）往往有差异，但一般来说，若$H_0$为真，且试验的次数又甚多时，则这种差异不应该很大。基于这种想法，皮尔逊使用

$$\chi^2=\sum_{i=1}^{k}\frac{(f_i-np_i)^2}{np_i}\left(\text{或}\chi^2=\sum_{i=1}^{k}\frac{(f_i-n\hat{p}_i)^2}{n\hat{p}_i}\right) \qquad (11)$$

作为检验假设$H_0$的统计量。并证明了以下定理。

定理　若$n$充分大，则当$H_0$为真时(不论$H_0$中的分布属什么分布)，统计量（11）总是近似地服从自由度为$k-r-1$的$\chi^2$分布，其中$r$是被估计的参数的个数。

于是，若在假设$H_0$下算得（11）有

在显著性水平$\alpha$下拒绝$H_0$，否则就接受。

注意：在使用$\chi^2$检验法时，要求样本容量$n$不小于50,以及每个$np_i$都不小于5,而且$np_i$最好是在5以上。否则应适当地合并$A_i$，以满足这个要求。

例6　下面列出了84个伊特拉斯坎（Etruscan）人男子的头颅的最大宽度（mm），试检验这些数据是否来自正态总体（取$\alpha=0.1$)。

141　148　132　138　154　142　150　146　155　158
150　140　147　148　144　150　149　145　149　158
143　141　144　144　126　140　144　142　141　140
145　135　147　146　141　136　140　146　142　137
148　154　137　139　143　140　131　143　141　149
148　135　148　152　143　144　141　143　147　146
150　132　142　142　143　153　149　146　149　138
142　149　142　137　134　144　146　147　140　142
140　137　152　145

解　编写Matlab程序如下：

```
clc
x=[141  148  132  138  154  142  150  146  155  158 ...
150  140  147  148  144  150  149  145  149  158 ...
143  141  144  144  126  140  144  142  141  140 ...
145  135  147  146  141  136  140  146  142  137 ...
148  154  137  139  143  140  131  143  141  149 ...
148  135  148  152  143  144  141  143  147  146 ...
150  132  142  142  143  153  149  146  149  138 ...
142  149  142  137  134  144  146  147  140  142 ...
140  137  152  145];
min(x),max(x)    %求数据中的最小数和最大数
hist(x,8)        %画直方图
fi=[length(find(x<135)),...
   length(find(x>=135&x<138)),...
   length(find(x>=138&x<142)),...
   length(find(x>=142&x<146)),...
   length(find(x>=146&x<150)),...
   length(find(x>=150&x<154)),...
   length(find(x>=152))]       %各区间上出现的频率
mu=mean(x),sigma=std(x)      %均值和标准差
fendian=[135,138,142,146,150,152]    %区间的分点
p0=normcdf(fendian,mu,sigma)          %分点处分布函数的值
p1=diff(p0)                                %中间各区间的概率
p=[p0(1),p1,1-p0(6)]                      %所有区间的概率
chi=(fi-84*p).^2./(84*p)
chisum=sum(chi)                           %皮尔逊统计量的值
x_a=chi2inv(0.9,4)      %chi2分布的0.9分位数
```

求得皮尔逊统计量chisum=1.9723，$\chi^2_{0.1}(7-2-1)=\chi^2_{0.1}(4)=7.7794$，故在水平0.1下接受$H_0$，即认为数据来自正态分布总体。

3.3.2　偏度、峰度检验（留作习题1）

3.4　其它非参数检验

Matlab还提供了一些非参数方法。

3.4.1　Wilcoxon秩和检验

在Matlab中，秩和检验由函数ranksum实现。命令为：

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

其中x，y可为不等长向量，alpha为给定的显著水平，它必须为0和1之间的数量。p返回产生两独立样本的总体是否相同的显著性概率，h返回假设检验的结果。如果x和y的总体差别不显著，则h为零；如果x和y的总体差别显著，则h为1。如果p接近于零，则可对原假设质疑。

例7　某商店为了确定向公司 $A$ 或公司 $B$ 购买某种产品，将 $A,B$ 公司以往各次进货的次品率进行比较，数据如下所示，设两样本独立。问两公司的商品的质量有无显著差异。设两公司的商品的次品的密度最多只差一个平移，取 $\alpha=0.05$ 。

$A$：7.0　3.5　9.6　8.1　6.2　5.1　10.4　4.0　2.0　10.5

$B$：5.7　3.2　4.2　11.0　9.7　6.9　3.6　4.8　5.6　8.4　10.1　5.5　12.3

解　分别以 $\mu_A$、$\mu_B$ 记公司 $A$、$B$ 的商品次品率总体的均值。所需检验的假设是

$$H_0:\mu_A=\mu_B,\quad H_1:\mu_A\neq\mu_B.$$

Matlab实现如下：

```
a=[7.0  3.5  9.6  8.1  6.2  5.1  10.4  4.0  2.0  10.5];
b=[5.7  3.2  4.2  11.0  9.7  6.9  3.6  4.8  5.6  8.4  10.1
5.5  12.3];
[p,h]=ranksum(a,b)
```

求得p=0.8041，h=0，表明两样本总体均值相等的概率为0.8041，并不很接近于零，且h=0说明可以接受原假设，即认为两个公司的商品的质量无明显差异。

3.5　中位数检验

在假设检验中还有一种检验方法为中位数检验,在一般的教学中不一定介绍,但在实际中也是被广泛应用到的。在Matlab中提供了这种检验的函数。函数的使用方法简单，下面只给出函数介绍。

3.5.1　signrank函数

signrank　Wilcoxon符号秩检验

[p,h]=signrank(x,y,alpha)

其中p给出两个配对样本x和y的中位数相等的假设的显著性概率。向量x，y的长度必须相同，alpha为给出的显著性水平，取值为0和1之间的数。h返回假设检验的结果。如果这两个样本的中位数之差几乎为0，则h=0；若有显著差异，则h=1。

3.5.2　signtest函数

signtest　符号检验

[p,h]= signtest(x,y,alpha)

其中p给出两个配对样本x和y的中位数相等的假设的显著性概率。x和y若为向量，二者的长度必须相同；y亦可为标量，在此情况下，计算x的中位数与常数y之间的差异。alpha和h同上。

习 题

1. 试用偏度、峰度检验法检验例6中的数据是否来自正态总体（取$\alpha=0.1$）。

2. 下面列出的是某工厂随机选取的20只部件的装配时间（分）：9.8,10.4,10.6,9.6,9.7,9.9,10.9,11.1,9.6,10.2,10.3,9.6,9.9,11.2,10.6,9.8,10.5,10.1,10.5,9.7。设装配时间的总体服从正态分布，是否可以认为装配时间的均值显著地大于10（取$\alpha=0.05$）？

3.下表分别给出两个文学家马克．吐温(Mark Twain)的八篇小品文及斯诺特格拉斯（Snodgrass）的10篇小品文中由3个字母组成的词的比例。

| | |
|---|---|
| 马克．吐温 | 0.225 0.262 0.217 0.240 0.230 0.229 0.235 0.217 |
| 斯诺特格拉斯 | 0.209 0.205 0.196 0.210 0.202 0.207 0.224 0.223 0.220 0.201 |

设两组数据分别来自正态总体，且两总体方差相等。两样本相互独立，问两个作家所写的小品文中包含由3个字母组成的词的比例是否有显著的差异（取$\alpha=0.05$）？

## §4 方差分析

我们已经作过两个总体均值的假设检验，如两台机床生产的零件尺寸是否相等，病人和正常人的某个生理指标是否一样。如果把这类问题推广一下，要检验两个以上总体的均值彼此是否相等，仍然用以前介绍的方法是很难做到的。而你在实际生产和生活中可以举出许多这样的问题：从用几种不同工艺制成的灯泡中，各抽取了若干个测量其寿命，要推断这几种工艺制成的灯泡寿命是否有显著差异；用几种化肥和几个小麦品种在若干块试验田里种植小麦，要推断不同的化肥和品种对产量有无显著影响。

可以看到，为了使生产过程稳定，达到优质、高产，需要对影响产品质量的因素进行分析，找出有显著影响的那些因素，除了从机理方面进行研究外，常常要作许多试验，对结果作分析、比较，寻求规律。用数理统计分析试验结果、鉴别各因素对结果影响程度的方法称为方差分析（Analysis Of Variance），记作 ANOVA。

人们关心的试验结果称为**指标**，试验中需要考察、可以控制的条件称为**因素或因子**，因素所处的状态称为**水平**。上面提到的灯泡寿命问题是单因素试验，小麦产量问题是双因素试验。处理这些试验结果的统计方法就称为单因素方差分析和双因素方差分析。

### 4.1 单因素方差分析

只考虑一个因素 $A$ 对所关心的指标的影响，$A$ 取几个水平，在每个水平上作若干个试验，试验过程中除 $A$ 外其它影响指标的因素都保持不变（只有随机因素存在），我们的任务是从试验结果推断，因素 $A$ 对指标有无显著影响，即当 $A$ 取不同水平时指标有无显著差别。

$A$ 取某个水平下的指标视为随机变量，判断 $A$ 取不同水平时指标有无显著差别，相当于检验若干总体的均值是否相等。

#### 4.1.1 数学模型

设 $A$ 取 $r$ 个水平 $A_1,A_2,\cdots,A_r$，在水平 $A_i$ 下总体 $x_i$ 服从正态分布 $N(\mu_i,\sigma^2)$，

$i=1,\cdots,r$，这里$\mu_i,\sigma^2$未知，$\mu_i$可以互不相同，但假定$x_i$有相同的方差。又设在每个水平$A_i$下都作了$n$次独立试验，即从中抽取容量为$n$的样本，记作，$x_{ji}$服从$N(\mu_i,\sigma^2)$，且相互独立。将这些数据列成下表（单因素试验数据表）的形式：

| | $A_1$ | $A_2$ | $\cdots$ | $A_r$ |
|---|---|---|---|---|
| 1 | $x_{11}$ | $x_{12}$ | … | $x_{1r}$ |
| 2 | $x_{21}$ | $x_{22}$ | … | $x_{2r}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $n$ | $x_{n1}$ | $x_{n2}$ | $\cdots$ | $x_{nr}$ |

将第$i$列称为第$i$组数据。判断$A$的$r$个水平对指标有无显著影响，相当于要作以下的假设检验

；不全相等

由于$x_{ji}$的取值既受不同水平$A_i$的影响，又受$A_i$固定下随机因素的影响，所以将它分解为

$$x_{ji}=\mu_i+\varepsilon_{ji},\ i=1,\cdots,r,\quad j=1,\cdots,n \tag{1}$$

其中，且相互独立。记

$$\mu=\frac{1}{r}\sum_{i=1}^{r}\mu_i,\ \alpha_i=\mu_i-\mu,\quad i=1,\cdots,r \tag{2}$$

$\mu$是总均值，$\alpha_i$是水平$A_i$对指标的效应。由（1）、（2）模型可表为

$$\begin{cases} x_{ji}=\mu+\alpha_i+\varepsilon_{ji} \\ \sum_{i=1}^{r}\alpha_i=0 \\ \varepsilon_{ji}\sim N(0,\sigma^2),\ i=1,\cdots,r,\ j=1,\cdots,n \end{cases} \tag{3}$$

原假设为（以后略去备选假设）

（4）

4.1.2 统计分析

记

$$\bar{x}_i=\frac{1}{n}\sum_{j=1}^{n}x_{ji},\ \bar{x}=\frac{1}{r}\sum_{i=1}^{r}\bar{x}_i=\frac{1}{rn}\sum_{i=1}^{r}\sum_{j=1}^{n}x_{ji} \tag{5}$$

$\bar{x}_i$是第$i$组数据的组平均值，$\bar{x}$是总平均值。考察全体数据对$\bar{x}$的偏差平方和

$$S=\sum_{i=1}^{r}\sum_{j=1}^{n}(x_{ji}-\bar{x})^2 \tag{6}$$

经分解可得

$$S = \sum_{i=1}^{r} n(\bar{x}_i - \bar{x})^2 + \sum_{i=1}^{r}\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)^2$$

记

$$S_A = \sum_{i=1}^{r} n(\bar{x}_i - \bar{x})^2 \qquad (7)$$

$$S_E = \sum_{i=1}^{r}\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)^2 \qquad (8)$$

则

$$S = S_A + S_E \qquad (9)$$

$S_A$ 是各组均值对总方差的偏差平方和，称为组间平方和；$S_E$ 是各组内的数据对均值偏差平方和的总和。$S_A$ 反映 $A$ 不同水平间的差异，$S_E$ 则表示在同一水平下随机误差的大小。

对 $S_E$ 和 $S_A$ 作进一步分析可得

（10）

$$ES_A = (r-1)\sigma^2 + \sum_{i=1}^{r} n\alpha_i^2 \qquad (11)$$

当 $H_0$ 成立时

（12）

可知若 $H_0$ 成立，$S_A$ 只反映随机波动，而若 $H_0$ 不成立，那它就还反映了 $A$ 的不同水平的效应 $\alpha_i$。单从数值上看，当 $H_0$ 成立时，由（10）、（12）对于一次试验应有

$$\frac{S_A/(r-1)}{S_E/[r(n-1)]} \approx 1$$

而当 $H_0$ 不成立时这个比值将远大于 1。当 $H_0$ 成立时，该比值服从自由度 $n_1 = r-1$，$n_2 = r(n-1)$ 的 $F$ 分布，即

$$F = \frac{S_A/(r-1)}{S_E/[r(n-1)]} \sim F(r-1, r(n-1)) \qquad (13)$$

为检验 $H_0$，给定显著性水平 $\alpha$，记 $F$ 分布的 $1-\alpha$ 分位数为，检验规则为

时接受 $H_0$，否则拒绝。

以上对 $S_A, S_E, S$ 的分析相当于对组间、组内等方差的分析，所以这种假设检验方法称方差分析。

4.1.3 方差分析表

将试验数据按上述分析、计算的结果排成下表的形式，称为单因素方差分析表。

| 方差来源 | 平方和 | 自由度 | 平方均值 | $F$ 值 | 概率 |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| 因素 $A$ | $S_A$ | $r-1$ | $\bar{S}_A = \dfrac{S_A}{r-1}$ | $F_{1-p}$ | $p>\alpha$ |
| 误差 | $S_E$ | $r(n-1)$ | $\bar{S}_E = \dfrac{S_E}{r(n-1)}$ | | |
| 总和 | $S$ | $rn-1$ | | | |

最后一列给出的概率 $p>\alpha$ 相当于。

方差分析一般用的显著性水平是：取 $\alpha=0.01$，拒绝 $H_0$，称因素 $A$ 的影响（或 $A$ 各水平的差异）非常显著；取 $\alpha=0.01$，不拒绝 $H_0$，但取 $\alpha=0.05$，拒绝 $H_0$，称因素 $A$ 的影响显著；取 $\alpha=0.05$，不拒绝 $H_0$，称因素 $A$ 无显著影响。

4.1.4 Matlab 实现

Matlab 统计工具箱中单因素方差分析的命令是 anoval，用法为：

p=anoval(x)

返回值 p 是一个概率，当 P$>\alpha$ 时接受 $H_0$，x 为 $n\times r$ 的数据矩阵（如上面的单因素试验数据表形式），x 的每一列是一个水平的数据。另外，还给出一个方差表和一个 Box 图。

例 1 为考察 5 名工人的劳动生产率是否相同，记录了每人 4 天的产量，并算出其平均值，如下表。你能从这些数据推断出他们的生产率有无显著差别吗？

| 工人<br>天 | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ |
|---|---|---|---|---|---|
| 1 | 256 | 254 | 250 | 248 | 236 |
| 2 | 242 | 330 | 277 | 280 | 252 |
| 3 | 280 | 290 | 230 | 305 | 220 |
| 4 | 298 | 295 | 302 | 289 | 252 |
| 平均产量 | 269.00 | 292.25 | 264.75 | 280.50 | 240.00 |

解 编写程序如下：

```
x=[256     254     250     248     236
   242     330     277     280     252
   280     290     230     305     220
   298     295     302     289     252];
p=anoval(x)
```

求得，故接受 $H_0$，即 5 名工人的生产率没有显著差异。方差表对应于上面的单因素方差分析表的 $1\sim5$ 列，$F=2.262$ 是 $F(4,15)$ 分布的 $1-p$ 分位数，可以验证

fcdf(2.262,4,15)=0.8891=1-p

Box 图反映了各组数据的特征。

注：接受 $H_0$，是将 5 名工人的生产率作为一个整体进行假设检验的结果，并不表明取其中 2 个工人的生产率作两总体的均值检验时，也一定接受均值相等的假设。实际上，读者可以用 ttest2 对本题作 $H_0:\mu_2=\mu_5$ 的检验，看看会得到什么结果。

**非均衡数据的方差分析**

上面所讨论的情况是 $r$ 个样本的容量即各组数据个数相等，称为均衡数据。若各组数据个数不等，称非均衡数据。非均衡数据的方差分析，其数学模型和统计分析的思路

和方法与上面一样。

anoval 也能处理非均衡数据，与处理均衡数据的区别仅在于数据输入的不同：

p=anoval(x,group)

x 为数组，从第 1 组到第 $r$ 组数据依次排列；group 为与 x 同长度的数组，标志 x 中数据的组别（在与 x 第 $i$ 组数据相对应的位置处输入整数）。

例 2　用 4 种工艺生产灯泡，从各种工艺制成的灯泡中各抽出了若干个测量其寿命，结果如下表，试推断这几种工艺制成的灯泡寿命是否有显著差异。

| 工艺<br>序号 | $A_1$ | $A_2$ | $A_3$ | $A_4$ |
|---|---|---|---|---|
| 1 | 1620 | 1580 | 1460 | 1500 |
| 2 | 1670 | 1600 | 1540 | 1550 |
| 3 | 1700 | 1640 | 1620 | 1610 |
| 4 | 1750 | 1720 | | 1680 |
| 5 | 1800 | | | |

解　编写程序如下：

```
x=[1620    1580    1460    1500
   1670    1600    1540    1550
   1700    1640    1620    1610
   1750    1720    1680    1800];
x=[x(1:4),x(16),x(5:8),x(9:11),x(12:15)];
g=[ones(1,5),2*ones(1,4),3*ones(1,3),4*ones(1,4)];
p=anoval(x,g)
```

求得 0.01<p=0.0331<0.05，所以几种工艺制成的灯泡寿命有显著差异。

4.1.6　多重比较

在灯泡寿命问题中，为了确定哪几种工艺制成的灯泡寿命有显著差异，我们先算出各组数据的均值：

| 工艺 | $A_1$ | $A_2$ | $A_3$ | $A_4$ |
|---|---|---|---|---|
| 均值 | 1708 | 1635 | 1540 | 1585 |

虽然 $A_1$ 的均值最大，但要判断它与其它几种有显著差异，尚需作两总体均值的假设检验。用 ttest2 检验的结果如下：

| 原假设 | $\mu_1=\mu_2$ | $\mu_1=\mu_3$ | $\mu_1=\mu_4$ |
|---|---|---|---|
| $h$ | 0 | 1 | 1 |
| $p$ | 0.1459 | 0.0202 | 0.0408 |

即 $A_1$ 与 $A_3, A_4$ 有显著差异（$\alpha=0.05$），但与 $A_2$ 无显著差异，要想进一步比较优劣，应增加试验数据。

以上作的几个两总体均值的假设检验，是多重比较的一部分。一般多重比较要对所有 $r$ 个总体作两两对比，分析相互间的差异。根据问题的具体情况可以减少对比次数。

## 4.2　双因素方差分析

如果要考虑两个因素 $A,B$ 对指标的影响，$A,B$ 各划分几个水平，对每一个水平组合作若干次试验，对所得数据进行方差分析，检验两因素是否分别对指标有显著影响，或者还要进一步检验两因素是否对指标有显著的交互影响。

4.2.1　数学模型

设 $A$ 取 $r$ 个水平 $A_1,A_2,\cdots,A_r$，$B$ 取 $s$ 个水平 $B_1,B_2,\cdots,B_s$，在水平组合 $(A_j,B_i)$ 下总体 $x_{ij}$ 服从正态分布 $N(\mu_{ij},\sigma^2)$，$i=1,\cdots,s$，$j=1,\cdots,r$。又设在水平组合 $(A_j,B_i)$ 下作了 $t$ 个试验，所得结果记作 $x_{ijk}$，$x_{ijk}$ 服从 $N(\mu_{ij},\sigma^2)$，$i=1,\cdots,s$，$j=1,\cdots,r$，$k=1,\cdots,t$，且相互独立。将这些数据列成下表的形式：

| | $A_1$ | $A_2$ | $\cdots$ | $A_r$ |
|---|---|---|---|---|
| $B_1$ | $x_{111}\cdots x_{11t}$ | $x_{121}\cdots x_{12t}$ | $\cdots$ | $x_{1r1}\cdots x_{1rt}$ |
| $B_2$ | $x_{211}\cdots x_{21t}$ | $x_{221}\cdots x_{22t}$ | $\cdots$ | $x_{2r1}\cdots x_{2rt}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $B_s$ | $x_{s11}\cdots x_{s1t}$ | $x_{s21}\cdots x_{s2t}$ | $\cdots$ | $x_{sr1}\cdots x_{srt}$ |

将 $x_{ijk}$ 分解为

，

$$i=1,\cdots,s,\quad j=1,\cdots,r,\quad k=1,\cdots,t \qquad (14)$$

其中，且相互独立。记

$$\mu=\frac{1}{rs}\sum_{i=1}^{s}\sum_{j=1}^{r}\mu_{ij},\quad \mu_{i\bullet}=\frac{1}{r}\sum_{j=1}^{r}\mu_{ij},\quad \beta_i=\mu_{i\bullet}-\mu$$

$$\mu_{\bullet j}=\frac{1}{s}\sum_{i=1}^{s}\mu_{ij},\quad \alpha_j=\mu_{\bullet j}-\mu, \qquad (15)$$

$\mu$ 是总均值，$\alpha_j$ 是水平 $A_j$ 对指标的效应，$\beta_i$ 是水平 $B_i$ 对指标的效应，$\gamma_{ij}$ 是水平 $A_j$ 与 $B_i$ 对指标的交互效应。模型表为

$$\begin{cases} x_{ijk}=\mu+\alpha_i+\beta_j+\gamma_{ij}+\varepsilon_{ijk} \\ \sum_{j=1}^{r}\alpha_j=0,\sum_{i=1}^{s}\beta_i=0,\sum_{i=1}^{s}\gamma_{ij}=\sum_{j=1}^{r}\gamma_{ij}=0 \\ \varepsilon_{ijk}\sim N(0,\sigma^2),i=1,\cdots,s,j=1,\cdots,r,k=1,\cdots,t \end{cases} \qquad (16)$$

原假设为

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

4.2.2　无交互影响的双因素方差分析

如果根据经验或某种分析能够事先判定两因素之间没有交互影响，每组试验就不必重复，即可令 $t=1$，过程大为简化。

Matlab 实现

统计工具箱中用 anova2 作双因素方差分析。命令为

p=anova2(x,reps)

其中 x 不同列的数据表示单一因素的变化情况，不同行中的数据表示另一因素的变化情况。如果每种行—列对（“单元”）有不止一个的观测值，则用参数 reps 来表明每个“单元”多个观测值的不同标号，即 reps 给出重复试验的次数 $t$。下面的矩阵中，列因素有 3 种水平，行因素有两种水平，但每组水平有两组样本，相应地用下标来标识：

$$\begin{bmatrix} x_{111} & x_{121} & x_{131} \\ x_{112} & x_{122} & x_{132} \\ x_{211} & x_{221} & x_{231} \\ x_{212} & x_{222} & x_{232} \end{bmatrix}$$

例 3　一火箭使用了 4 种燃料，3 种推进器作射程试验，每种燃料与每种推进器的组合各发射火箭 2 次，得到结果如下：

| | $B_1$ | $B_2$ | $B_3$ |
|---|---|---|---|
| $A_1$ | 58.2, 52.6 | 56.2, 41.2 | 65.3, 60.8 |
| $A_2$ | 49.1, 42.8 | 54.1, 50.5 | 51.6, 48.4 |
| $A_3$ | 60.1, 58.3 | 70.9, 73.2 | 39.2, 40.7 |
| $A_4$ | 75.8, 71.5 | 58.2, 51.0 | 48.7, 41.4 |

试在水平 0.05 下，检验不同燃料（因素 $A$）、不同推进器（因素 $B$）下的射程是否有显著差异？交互作用是否显著？

解　编写程序如下：

```
clc,clear
x0=[58.2,52.6  56.2,41.2   65.3,60.8
49.1,42.8   54.1,50.5   51.6,48.4
60.1,58.3   70.9,73.2   39.2,40.7
75.8,71.5   58.2,51.0   48.7,41.4];
x1=x0(:,1:2:5);x2=x0(:,2:2:6);
for i=1:4
   x(2*i-1,:)=x1(i,:);
   x(2*i,:)=x2(i,:);
end
p=anova2(x,2)
```

求得 p=0.0035　0.0260　0.001，表明各试验均值相等的概率都为小概率，故可拒绝均值相等假设。即认为不同燃料（因素 $A$）、不同推进器（因素 $B$）下的射程有显著差异，交互作用也是显著的。

## 习　题

1. 将抗生素注入人体会产生抗生素与血浆蛋白质结合的现象，以致减少了药效。下表列出 5 种常用的抗生素注入到牛的体内时，抗生素与血浆蛋白质结合的百分比。试在

水平 $x_{111}\cdots x_{11t}$ 下检验这些百分比的均值有无显著的差异。设各总体服从正态分布，且方差相同。

| 青霉素 | 四环素 | 链霉素 | 红霉素 | 氯霉素 |
|---|---|---|---|---|
| 29.6 | 27.3 | 5.8 | 21.6 | 29.2 |
| 24.3 | 32.6 | 6.2 | 17.4 | 32.8 |
| 28.5 | 30.8 | 11.0 | 18.3 | 25.0 |
| 32.0 | 34.8 | 8.3 | 19.0 | 24.2 |

2. 为分析 4 种化肥和 3 个小麦品种对小麦产量的影响，把一块试验田等分成 36 小块，对种子和化肥的每一种组合种植 3 小块田，产量如下表所示（单位公斤），问品种、化肥及二者的交互作用对小麦产量有无显著影响。

| 化肥 | | $A_1$ | $A_2$ | $A_3$ | $A_4$ |
|---|---|---|---|---|---|
| 品种 | $B_1$ | 173, 172, 173 | 174, 176, 178 | 177, 179, 176 | 172, 173, 174 |
| | $B_2$ | 175, 173, 176 | 178, 177, 179 | 174, 175, 173 | 170, 171, 172 |
| | $B_3$ | 177, 175, 176 | 174, 174, 175 | 174, 173, 174 | 169, 169, 170 |

## §5 回归分析

前面我们讲过曲线拟合问题。曲线拟合问题的特点是，根据得到的若干有关变量的一组数据，寻找因变量与（一个或几个）自变量之间的一个函数，使这个函数对那组数据拟合得最好。通常，函数的形式可以由经验、先验知识或对数据的直观观察决定，要作的工作是由数据用最小二乘法计算函数中的待定系数。从计算的角度看，问题似乎已经完全解决了，还有进一步研究的必要吗?

从数理统计的观点看，这里涉及的都是随机变量，我们根据一个样本计算出的那些系数，只是它们的一个（点）估计，应该对它们作区间估计或假设检验，如果置信区间太大，甚至包含了零点，那么系数的估计值是没有多大意义的。另外也可以用方差分析方法对模型的误差进行分析，对拟合的优劣给出评价。简单地说，回归分析就是对拟合问题作的统计分析。

具体地说，回归分析在一组数据的基础上研究这样几个问题：

（i）建立因变量 $y$ 与自变量 $x_1,x_2,\cdots,x_m$ 之间的回归模型（经验公式）；

（ii）对回归模型的可信度进行检验；

（iii）判断每个自变量对 $y$ 的影响是否显著；

（iv）诊断回归模型是否适合这组数据；

（v）利用回归模型对 $y$ 进行预报或控制。

### 5.1 多元线性回归

回归分析中最简单的形式是 $y=\beta_0+\beta_1 x$， $x,y$ 均为标量， $\beta_0,\beta_1$ 为回归系数，称一元线性回归。它的一个自然推广是 $x$ 为多元变量，形如

(1)

$m\geq 2$，或者更一般地

$$y=\beta_0+\beta_1 f_1(x)+\cdots+\beta_m f_m(x) \qquad (2)$$

其中，是已知函数。这里$y$对回归系数是线性的，称为多元线性回归。不难看出，对自变量$x$作变量代换，就可将（2）化为（1）的形式，所以下面以（1）为多元线性回归的标准型。

5.1.1 模型

在回归分析中自变量是影响因变量$y$的主要因素，是人们能控制或能观察的，而$y$还受到随机因素的干扰，可以合理地假设这种干扰服从零均值的正态分布，于是模型记作

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon \\ \varepsilon \sim N(0,\sigma^2) \end{cases} \tag{3}$$

其中$\sigma$未知。现得到$n$个独立观测数据，，由（3）得

$$\begin{cases} y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} + \varepsilon_i \\ \varepsilon_i \sim N(0,\sigma^2), \quad i = 1,\cdots,n \end{cases} \tag{4}$$

记

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \tag{5}$$

，

（4）表为

$$\begin{cases} Y = X\beta + \varepsilon \\ \varepsilon \sim N(0,\sigma^2) \end{cases} \tag{6}$$

5.1.2 参数估计

用最小二乘法估计模型（3）中的参数$\beta$。

由（4）式这组数据的误差平方和为

$$Q(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = (Y - X\beta)^T (Y - X\beta) \tag{7}$$

求$\beta$使$Q(\beta)$最小，得到$\beta$的最小二乘估计，记作$\hat{\beta}$，可以推出

（8）

将$\hat{\beta}$代回原模型得到$y$的估计值

（9）

而这组数据的拟合值为$\hat{Y} = X\hat{\beta}$，拟合误差$e = Y - \hat{Y}$称为残差，可作为随机误差$\varepsilon$的估计，而

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{10}$$

为残差平方和（或剩余平方和），即$Q(\hat{\beta})$。

5.1.3 统计分析

不加证明地给出以下结果：

（i）$\hat{\beta}$ 是 $\beta$ 的线性无偏最小方差估计。指的是 $\hat{\beta}$ 是 $Y$ 的线性函数；$\hat{\beta}$ 的期望等于 $\beta$；在 $\beta$ 的线性无偏估计中，$\hat{\beta}$ 的方差最小。

（ii）$\hat{\beta}$ 服从正态分布

（11）

（iii）对残差平方和 $Q$，，且

$$\frac{Q}{\sigma^2} \sim \chi^2(n-m-1) \qquad (12)$$

由此得到 $\sigma^2$ 的无偏估计

$$s^2 = \frac{Q}{n-m-1} = \hat{\sigma}^2 \qquad (13)$$

$s^2$ 是剩余方差（残差的方差），$s$ 称为剩余标准差。

（iv）对 $Y$ 的样本方差 $S = \sum_{i=1}^{n} (y_i - \bar{y})^2$ 进行分解，有

$$S = Q + U, \quad U = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad (14)$$

其中 $Q$ 是由（10）定义的残差平方和，反映随机误差对 $y$ 的影响，$U$ 称为回归平方和，反映自变量对 $y$ 的影响。

5.1.4 回归模型的假设检验

因变量 $y$ 与自变量 $x_1,\cdots,x_m$ 之间是否存在如模型（1）所示的线性关系是需要检验的，显然，如果所有的 $|\hat{\beta}_j|\ (j=1,\cdots,m)$ 都很小，$y$ 与 $x_1,\cdots,x_m$ 的线性关系就不明显，所以可令原假设为

当 $H_0$ 成立时由分解式（14）定义的 $U,Q$ 满足

$$F = \frac{U/m}{Q/(n-m-1)} \sim F(m,n-m-1) \qquad (15)$$

在显著性水平 $\alpha$ 下有 $1-\alpha$ 分位数，若，接受 $H_0$；否则，拒绝。

**注意** 拒绝 $H_0$ 只说明 $y$ 与 $x_1,\cdots,x_m$ 的线性关系不明显，可能存在非线性关系，如平方关系。

还有一些衡量 $y$ 与 $x_1,\cdots,x_m$ 相关程度的指标，如用回归平方和在样本方差中的比值定义

$$R^2 = \frac{U}{S} \qquad (16)$$

$R \in [0,1]$ 称为相关系数，$R$ 越大，$y$ 与 $x_1,\cdots,x_m$ 相关关系越密切，通常，$R$ 大于 0.8（或 0.9）才认为相关关系成立。

5.1.5 回归系数的假设检验和区间估计

当上面的 $H_0$ 被拒绝时，$\beta_j$ 不全为零，但是不排除其中若干个等于零。所以应进

一步作如下$m$个检验$(j=1,\cdots,m)$：

$$H_0^{(j)}:\beta_j=0$$

由（11）式，，$c_{jj}$是$(X^TX)^{-1}$对角线上的元素，用$s^2$代替$\sigma^2$，由（11）~（13）式，当$H_0^{(j)}$成立时

$$t_j=\frac{\hat{\beta}_j/\sqrt{c_{jj}}}{\sqrt{Q/(n-m-1)}}\sim t(n-m-1) \qquad (17)$$

对给定的$\alpha$，若，接受$H_0^{(j)}$；否则，拒绝。

(17)式也可用于对$\beta_j$作区间估计（$j=0,1,\cdots,m$），在置信水平$1-\alpha$下，$\beta_j$的置信区间为

$$[\hat{\beta}_j-t_{1-\frac{\alpha}{2}}(n-m-1)s\sqrt{c_{jj}},\hat{\beta}_j+t_{1-\frac{\alpha}{2}}(n-m-1)s\sqrt{c_{jj}}] \qquad (18)$$

其中$s=\sqrt{\frac{Q}{n-m-1}}$。

5.1.6 利用回归模型进行预测

当回归模型和系数通过检验后，可由给定的预测$y_0$，$y_0$是随机的，显然其预测值（点估计）为

$$\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1x_{01}+\cdots+\hat{\beta}_mx_{0m} \qquad (19)$$

给定$\alpha$可以算出$y_0$的预测区间（区间估计），结果较复杂，但当$n$较大且$x_{0i}$接近平均值$\bar{x}_i$时，$y_0$的预测区间可简化为

（20）

其中$u_{1-\frac{\alpha}{2}}$是标准正态分布的$1-\frac{\alpha}{2}$分位数。

对$y_0$的区间估计方法可用于给出已知数据残差$e_i=y_i-\hat{y}_i\ (i=1,\cdots,n)$的置信区间，$e_i$服从均值为零的正态分布，所以若某个$e_i$的置信区间不包含零点，则认为这个数据是异常的，可予以剔除。

5.1.7 Matlab 实现

Matlab 统计工具箱用命令 regress 实现多元线性回归，用的方法是最小二乘法，用法是：

b=regress(Y,X)

其中 Y,X 为按（5）式排列的数据，b 为回归系数估计值$\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_m$。

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

这里 Y,X 同上，alpha 为显著性水平（缺省时设定为 0.05），b,bint 为回归系数估计值和它们的置信区间，r,rint 为残差（向量）及其置信区间，stats 是用于检验回归模型的统计量，有三个数值，第一个是$R^2$（见（16）式），第二个是$F$（见（15）式），第 3 个是与$F$对应的概率$p$，$p<\alpha$拒绝$H_0$，回归模型成立。

残差及其置信区间可以用 rcoplot(r,rint)画图。

例 1 合金的强度 $y$ 与其中的碳含量 $x$ 有比较密切的关系，今从生产中收集了一批数据如下表：

| $x$ | 0.10 | 0.11 | 0.12 | 0.13 | 0.14 | 0.15 | 0.16 | 0.17 | 0.18 |
|---|---|---|---|---|---|---|---|---|---|
| $y$ | 42.0 | 41.5 | 45.0 | 45.5 | 45.0 | 47.5 | 49.0 | 55.0 | 50.0 |

试先拟合一个函数 $y(x)$，再用回归分析对它进行检验。

解　先画出散点图：

```
x=0.1:0.01:0.18;
y=[42,41.5,45.0,45.5,45.0,47.5,49.0,55.0,50.0];
plot(x,y,'+')
```

可知 $y$ 与 $x$ 大致上为线性关系。

设回归模型为

$$y=\beta_0+\beta_1 x \qquad (21)$$

用 regress 和 rcoplot 编程如下：

```
clc,clear
x1=[0.1:0.01:0.18]';
y=[42,41.5,45.0,45.5,45.0,47.5,49.0,55.0,50.0]';
x=[ones(9,1),x1];
[b,bint,r,rint,stats]=regress(y,x);
b,bint,stats,rcoplot(r,rint)
```

得到

b =27.4722  137.5000

bint =18.6851  36.2594

75.7755  199.2245

stats =0.7985  27.7469  0.0012

即 $\hat{\beta}_0=27.4722$，$\hat{\beta}_0$ 的置信区间是 [18.6851,36.2594]，$\hat{\beta}_1$ 的置信区间是 [75.7755,199.2245]；$R^2=0.7985$，$F=27.7469$，$p=0.0012$。

可知模型（21）成立。

观察命令 rcoplot(r,rint)所画的残差分布，除第 8 个数据外其余残差的置信区间均包含零点，第 8 个点应视为异常点，将其剔除后重新计算，可得

b =30.7820  109.3985

bint =26.2805  35.2834

76.9014  141.8955

stats =0.9188  67.8534  0.0002

应该用修改后的这个结果。

例 2 某厂生产的一种电器的销售量 $y$ 与竞争对手的价格 $x_1$ 和本厂的价格 $x_2$ 有关。下表是该商品在 10 个城市的销售记录。

| $x_1$ 元 | 120 | 140 | 190 | 130 | 155 | 175 | 125 | 145 | 180 | 150 |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_2$ 元 | 100 | 110 | 90 | 150 | 210 | 150 | 250 | 270 | 300 | 250 |
| $Y$ 个 | 102 | 100 | 120 | 77 | 46 | 93 | 26 | 69 | 65 | 85 |

试根据这些数据建立 $y$ 与 $x_1$ 和 $x_2$ 的关系式，对得到的模型和系数进行检验。若某市本

厂产品售价 160（元），竞争对手售价 170（元），预测商品在该市的销售量。

解　分别画出 $y$ 关于 $x_1$ 和 $y$ 关于 $x_2$ 的散点图，可以看出 $y$ 与 $x_2$ 有较明显的线性关系，而 $y$ 与 $x_1$ 之间的关系则难以确定，我们将作几种尝试，用统计分析决定优劣。

设回归模型为

（22）

编写如下程序：

```
x1=[120  140  190  130  155  175  125  145  180  150]';
x2=[100  110   90  150  210  150  250  270  300  250]';
y=[102  100   120  77   46   93   26   69   65   85]';
x=[ones(10,1),x1,x2];
[b,bint,r,rint,stats]=regress(y,x);
b,bint,stats
```

得到

```
b =66.5176   0.4139  -0.2698
bint =-32.5060  165.5411
       -0.2018   1.0296
       -0.4611  -0.0785
stats =0.6527   6.5786   0.0247
```

可以看出结果不是太好：$p=0.0247$，取 $\alpha=0.05$ 时回归模型（22）可用，但取 $\alpha=0.01$ 则模型不能用；$R^2=0.6527$ 较小；$\hat{\beta}_0,\hat{\beta}_1$ 的置信区间包含了零点。下面将试图用 $x_1,x_2$ 的二次函数改进它。

5.1.8 多项式回归

如果从数据的散点图上发现 $y$ 与 $x$ 呈较明显的二次（或高次）函数关系，或者用线性模型（1）的效果不太好，就可以选用多项式回归。

5.1.8.1 一元多项式回归

一元多项式回归可用命令 polyfit 实现。

例 3　将 17 至 29 岁的运动员每两岁一组分为 7 组，每组两人测量其旋转定向能力，以考察年龄对这种运动能力的影响。现得到一组数据如下表：

| 年　龄 | 17 | 19 | 21 | 23 | 25 | 27 | 29 |
|---|---|---|---|---|---|---|---|
| 第一人 | 20.48 | 25.13 | 26.15 | 30.0 | 26.1 | 20.3 | 19.35 |
| 第二人 | 24.35 | 28.11 | 26.3 | 31.4 | 26.92 | 25.7 | 21.3 |

试建立二者之间的关系。

解　数据的散点图明显地呈现两端低中间高的形状，所以应拟合一条二次曲线。

选用二次模型

（23）

编写如下程序：

```
x0=17:2:29;x0=[x0,x0];
y0=[20.48  25.13  26.15  30.0  26.1   20.3  19.35...
     24.35  28.11  26.3   31.4  26.92  25.7  21.3];
[p,s]=polyfit(x0,y0,2); p
```

得到

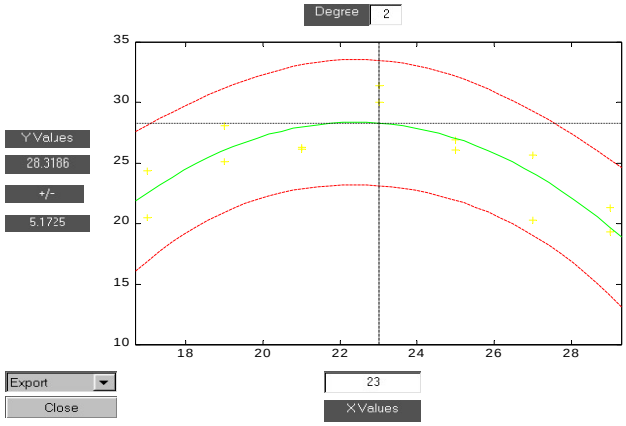
```
p =-0.2003   8.9782  -72.2150
```

即 $a_2=-0.2003$，$a_1=8.9782$，。

上面的s是一个数据结构，用于计算其它函数的计算，如

```
[y,delta]=polyconf(p,x0,s);y
```

得到$y$的拟合值，及预测值$y$的置信区间半径delta。




用polytool(x0,y0,2)，可以得到一个如上图的交互式画面，在画面中绿色曲线为拟合曲线，它两侧的红线是$y$的置信区间。你可以用鼠标移动图中的十字线来改变图下方的$x$值，也可以在窗口内输入，左边就给出$y$的预测值及其置信区间。通过左下方的Export下拉式菜单，可以输出回归系数等。这个命令的用法与下面将介绍的rstool相似。

5.1.8.2 多元二项式回归

统计工具箱提供了一个作多元二项式回归的命令rstool，它也产生一个交互式画面，并输出有关信息，用法是

rstool(x,y,model,alpha)

其中输入数据x,y分别为$n\times m$矩阵和$n$维向量，alpha为显著性水平$\alpha$（缺省时设定为0.05），model由下列4个模型中选择1个（用字符串输入，缺省时设定为线性模型）：

linear(线性)：

purequadratic(纯二次)：$y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m+\sum_{j=1}^{m}\beta_{jj}x_j^2$

interaction（交叉）：$y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m+\sum_{1\le j\ne k\le m}\beta_{jk}x_jx_k$

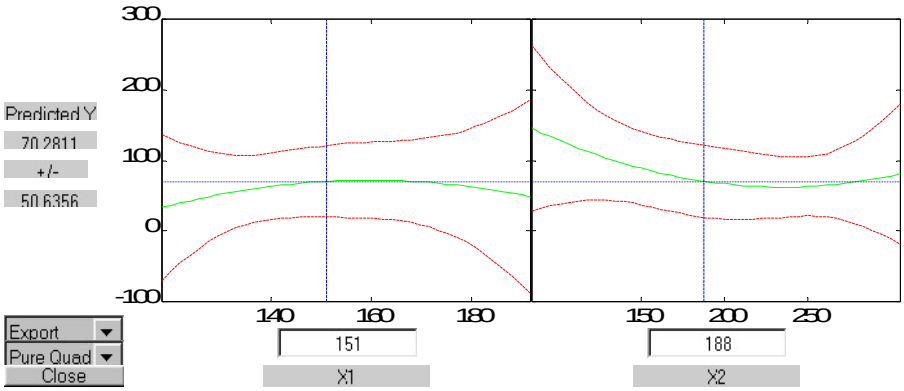
quadratic(完全二次)：$y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m+\sum_{1\le j,k\le m}\beta_{jk}x_jx_k$

我们再作一遍例2 商品销售量与价格问题，选择纯二次模型，即

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_{11}x_1^2+\beta_{22}x_2^2 \qquad (24)$$

编程如下：

```
x1=[120  140  190  130  155  175  125  145  180  150]';
x2=[100  110   90  150  210  150  250  270  300  250]';
y=[102  100   120  77   46   93   26   69   65   85]';
x=[x1 x2];
rstool(x,y,'purequadratic')
```




得到一个如图所示的交互式画面，左边是 $x_1$（=151）固定时的曲线 $y(x_1)$ 及其置信区间，右边是 $x_2$（=188）固定时的曲线 $y(x_2)$ 及其置信区间。用鼠标移动图中的十字线，或在图下方窗口内输入，可改变 $x_1, x_2$。图左边给出 $y$ 的预测值及其置信区间，就用这种画面可以回答例2提出的“若某市本厂产品售价160（元），竞争对手售价170（元），预测该市的销售量”问题。

图的左下方有两个下拉式菜单，一个菜单Export用以向Matlab工作区传送数据，包括beta(回归系数)，rmse（剩余标准差），residuals(残差)。模型（24）的回归系数和剩余标准差为

beta =-312.5871　7.2701　-1.7337　-0.0228　0.0037

rmse =16.6436

另一个菜单model用以在上述4个模型中选择，你可以比较以下它们的剩余标准差，会发现以模型（24）的rmse=16.6436最小。

## 5.2 非线性回归和逐步回归

本节介绍怎样用Matlab统计工具箱实现非线性回归和逐步回归。

### 5.2.1 非线性回归

非线性回归是指因变量 $y$ 对回归系数 $\beta_1,\cdots,\beta_m$（而不是自变量）是非线性的。Matlab统计工具箱中的nlinfit，nlparci，nlpredci，nlintool，不仅给出拟合的回归系数，而且可以给出它的置信区间，及预测值和置信区间等。下面通过例题说明这些命令的用法。

例4 在研究化学动力学反应过程中，建立了一个反应速度和反应物含量的数学模型，形式为

$$y=\frac{\beta_4 x_2-\dfrac{x_3}{\beta_5}}{1+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3}$$

其中 $\beta_1,\cdots,\beta_5$ 是未知的参数，$x_1, x_2, x_3$ 是三种反应物（氢，$n$ 戊烷，异构戊烷）的含量，$y$ 是反应速度。今测得一组数据如下表，试由此确定参数 $\beta_1,\cdots,\beta_5$，并给出其置信区间。$\beta_1,\cdots,\beta_5$ 的参考值为（0.1，0.05，0.02，1，2）。

| 序号 | 反应速度 $y$ | 氢 $x_1$ | $n$ 戊烷 $x_2$ | 异构戊烷 $x_3$ |
|---|---|---|---|---|
| 1 | 8.55 | 470 | 300 | 10 |

| | | | | |
|---|---|---|---|---|
| 2 | 3.79 | 285 | 80 | 10 |
| 3 | 4.82 | 470 | 300 | 120 |
| 4 | 0.02 | 470 | 80 | 120 |
| 5 | 2.75 | 470 | 80 | 10 |
| 6 | 14.39 | 100 | 190 | 10 |
| 7 | 2.54 | 100 | 80 | 65 |
| 8 | 4.35 | 470 | 190 | 65 |
| 9 | 13.00 | 100 | 300 | 54 |
| 10 | 8.50 | 100 | 300 | 120 |
| 11 | 0.05 | 100 | 80 | 120 |
| 12 | 11.32 | 285 | 300 | 10 |
| 13 | 3.13 | 285 | 190 | 120 |

解　首先，以回归系数和自变量为输入变量，将要拟合的模型写成函数文件huaxue.m：

```
function yhat=huaxue(beta,x);
yhat=(beta(4)*x(2)-x(3)/beta(5))./(1+beta(1)*x(1)+...
beta(2)*x(2)+beta(3)*x(3));
```

然后，用nlinfit计算回归系数，用nlparci计算回归系数的置信区间，用nlpredci计算预测值及其置信区间，编程如下：

```
clc,clear
x0=[ 1        8.55       470       300        10
 2        3.79       285       80         10
 3        4.82       470       300        120
 4        0.02       470       80         120
 5        2.75       470       80         10
 6        14.39      100       190        10
 7        2.54       100       80         65
 8        4.35       470       190        65
 9        13.00      100       300        54
 10       8.50       100       300        120
 11       0.05       100       80         120
 12       11.32      285       300        10
 13       3.13       285       190        120];
x=x0(:,3:5);
y=x0(:,2);
beta=[0.1,0.05,0.02,1,2];  %回归系数的初值
[betahat,f,j]=nlinfit(x,y,'huaxue',beta);  %f,j是下面命令用的信息
betaci=nlparci(betahat,f,j);
betaa=[betahat,betaci]    %回归系数及其置信区间
[yhat,delta]=nlpredci('huaxue',x,betahat,f,j)
%y的预测值及其置信区间的半径，置信区间为yhat±delta。
```

用nlintool得到一个交互式画面，左下方的Export可向工作区传送数据，如剩余标准差等。使用命令

```
nlintool(x,y,'huaxue',beta)
```

可看到画面，并传出剩余标准差rmse= 0.1933。

5.2.2 逐步回归

实际问题中影响因变量的因素可能很多，我们希望从中挑选出影响显著的自变量来建立回归模型，这就涉及到变量选择的问题，逐步回归是一种从众多变量中有效地选择重要变量的方法。以下只讨论线性回归模型（1）式的情况。

变量选择的标准，简单地说就是所有对因变量影响显著的变量都应选入模型，而影响不显著的变量都不应选入模型，从便于应用的角度应使模型中变量个数尽可能少。

若候选的自变量集合为，从中选出一个子集$S_1 \subset S$，设$S_1$中有$l$个自变量$(l=1,\cdots,m)$，由$S_1$和因变量$y$构造的回归模型的误差平方和为$Q$，则模型的剩余标准差的平方$s^2 = \dfrac{Q}{n-l-1}$，$n$为数据样本容量。所选子集$S_1$应使$s$尽量小，通常回归模型中包含的自变量越多，误差平方和$Q$越小，但若模型中包含有对$y$影响很小的变量，那么$Q$不会由于包含这些变量在内而减少多少， 却因$l$的增加可能使$s$反而增大，同时这些对$y$影响不显著的变量也会影响模型的稳定性，因此可将剩余标准差$s$最小作为衡量变量选择的一个数量标准。

逐步回归是实现变量选择的一种方法，基本思路为，先确定一初始子集，然后每次从子集外影响显著的变量中引入一个对$y$影响最大的，再对原来子集中的变量进行检验，从变得不显著的变量中剔除一个影响最小的，直到不能引入和剔除为止。使用逐步回归有两点值得注意，一是要适当地选定引入变量的显著性水平$\alpha_{in}$和剔除变量的显著性水平$\alpha_{out}$，显然，$\alpha_{in}$越大，引入的变量越多；$\alpha_{out}$越大，剔除的变量越少。二是由于各个变量之间的相关性，一个新的变量引入后，会使原来认为显著的某个变量变得不显著，从而被剔除，所以在最初选择变量时应尽量选择相互独立性强的那些。

在Matlab统计工具箱中用作逐步回归的是命令stepwise，它提供了一个交互式画面，通过这个工具你可以自由地选择变量，进行统计分析，其通常用法是：

stepwise(x,y,inmodel,alpha)

其中x是自变量数据，y是因变量数据，分别为$n \times m$和$n \times 1$矩阵，inmodel是矩阵x的列数的指标，给出初始模型中包括的子集（缺省时设定为全部自变量），alpha为显著性水平。

stepwise命令产生三个图形窗口：Stepwise Table，Stepwise History，Stepwise Plot。

Stepwise Table窗口中列出了一个统计表，包括回归系数及其置信区间，模型的统计量（RMSE R-square,F,p等，其含义与regress,rstool相同）。你可以通过这些统计量的变化来确定模型。

Stepwise History窗口显示RMSE的值及其置信区间。

Stepwise Plot窗口，显示回归系数及其置信区间，绿色表明在模型中的变量，红色表明从模型中移去的变量，两边有虚线或实线，虚线表示该变量的拟合系数与零无显著差异，实线则表明有显著差异。在这个窗口中还有Scale Inputs和Export按钮。

按下Scale Inputs表明对于输入数据的每列进行正态化处理，使其标准差为1。点击Export产生一个菜单，表明了要传送给Matlab工作区的参数，它们给出了统计计算的一些结果。

下面通过一个例子说明stepwise的用法。

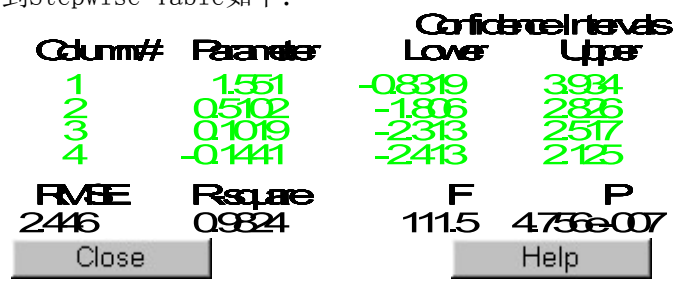
例5　水泥凝固时放出的热量 $y$ 与水泥中4种化学成分 $x_1,x_2,x_3,x_4$ 有关，今测得一组数据如下，试用逐步回归来确定一个线性模型

| 序号 | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|---|
| 1 | 7 | 26 | 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 | 95.9 |
| 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 |

编写程序如下：

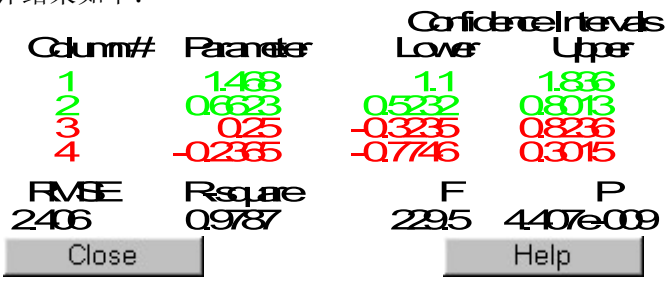
```
clc,clear
x0=[1      7     26     6     60     78.5
 2      1     29     15    52     74.3
 3      11    56     8     20     104.3
 4      11    31     8     47     87.6
 5      7     52     6     33     95.9
 6      11    55     9     22     109.2
 7      3     71     17    6      102.7
 8      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];
x=x0(:,2:5);
y=x0(:,6);
stepwise(x,y)
```

得到Stepwise Table如下：




可以看出，$x_3,x_4$ 不显著，移去这两个变量（程序为stepwise(x,y,[1,2])）

后的统计结果如下：




这个表中的 $x_3, x_4$ 两行用红色显示，表明它们已移去。

从新的统计结果可以看出，虽然剩余标准差 $s$（RMSE）没有太大的变化，但是统计量 $F$ 的值明显增大，因此新的回归模型更好一些。使用前面的回归分析方法可以求出最终的模型为

$$y = 52.5773 + 1.4683x_1 + 0.6623x_2$$

## 习　题

1. 某人记录了21天每天使用空调器的时间和使用烘干器的次数，并监视电表以计算出每天的耗电量，数据见下表，试研究耗电量（KWH）与空调器使用的小时数(AC)和烘干器使用次数(DRYER)之间的关系，建立并检验回归模型，诊断是否有异常点。

| 序号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| KWH | 35 | 63 | 66 | 17 | 94 | 79 | 93 | 66 | 94 | 82 | 78 |
| AC | 1.5 | 4.5 | 5.0 | 2.0 | 8.5 | 6.0 | 13.5 | 8.0 | 12.5 | 7.5 | 6.5 |
| DRYER | 1 | 2 | 2 | 0 | 3 | 3 | 1 | 1 | 1 | 2 | 3 |
| 序号 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | |
| kWH | 65 | 77 | 75 | 62 | 85 | 43 | 57 | 33 | 65 | 33 | |
| AC | 8.0 | 7.5 | 8.0 | 7.5 | 12.0 | 6.0 | 2.5 | 5.0 | 7.5 | 6.0 | |
| DRYER | 1 | 2 | 2 | 1 | 1 | 0 | 3 | 0 | 1 | 0 | |

2. 在一丘陵地带测量高程，$x$ 和 $y$ 方向每隔100米测一个点，得高程如下表，试拟合一曲面，确定合适的模型，并由此找出最高点和该点的高程。

| $y$ \ $x$ | 100 | 200 | 300 | 400 |
|---|---|---|---|---|
| 100 | 636 | 697 | 624 | 478 |
| 200 | 698 | 712 | 630 | 478 |
| 300 | 680 | 674 | 598 | 412 |
| 400 | 662 | 626 | 552 | 334 |

3. 一矿脉有13个相邻样本点，人为地设定一原点，现测得各样本点对原点的距离 $x$，与该样本点处某种金属含量 $y$ 的一组数据如下，画出散点图观测二者的关系，试建立合适的回归模型，如二次曲线、双曲线、对数曲线等。

| $x$ | 2 | 3 | 4 | 5 | 7 | 8 | 10 |
|---|---|---|---|---|---|---|---|
| $y$ | 106.42 | 109.20 | 109.58 | 109.50 | 110.00 | 109.93 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 110.49 | | | | | | |
| $x$ | 11 | 14 | 15 | 16 | 18 | 19 |
| $y$ | 110.59 | 110.60 | 110.90 | 110.76 | 111.00 | 111.20 |

## §6 马氏链模型

6.1 随机过程的概念

一个随机试验的结果有多种可能性，在数学上用一个随机变量（或随机向量）来描述。在许多情况下，人们不仅需要对随机现象进行一次观测，而且要进行多次，甚至接连不断地观测它的变化过程。这就要研究无限多个，即一族随机变量。随机过程理论就是研究随机现象变化过程的概率规律性的。

定义 1 设 $\{\xi_t, t\in T\}$ 是一族随机变量，$T$ 是一个实数集合，若对任意实数 $t\in T, \xi_t$ 是一个随机变量，则称 $\{\xi_t, t\in T\}$ 为随机过程。

$T$ 称为参数集合，参数 $t$ 可以看作时间。$\xi_t$ 的每一个可能取值称为随机过程的一个状态。其全体可能取值所构成的集合称为状态空间，记作 $E$。当参数集合 $T$ 为非负整数集时，随机过程又称随机序列。本章要介绍的马尔可夫链就是一类特殊的随机序列。

例 1 在一条自动生产线上检验产品质量，每次取一个，“废品”记为 1，“合格品”记为 0。以 $\xi_n$ 表示第 $n$ 次检验结果，则 $\xi_n$ 是一个随机变量。不断检验，得到一列随机变量 $\xi_1, \xi_2, \cdots$，记为。它是一个随机序列，其状态空间 $E=\{0,1\}$。

例 2 在 $m$ 个商店联营出租照相机的业务中（顾客从其中一个商店租出，可以到 $m$ 个商店中的任意一个归还），规定一天为一个时间单位，“$\xi_t = j$”表示“第 $t$ 天开始营业时照相机在第 $j$ 个商店”，$j=1,2,\cdots,m$。则是一个随机序列，其状态空间。

例 3 统计某种商品在 $t$ 时刻的库存量，对于不同的 $t$，得到一族随机变量，是一个随机过程，状态空间 $E=[0,R]$，其中 $R$ 为最大库存量。

我们用一族分布函数来描述随机过程的统计规律。一般地，一个随机过程 $\{\xi_t, t\in T\}$，对于任意正整数 $n$ 及 $T$ 中任意 $n$ 个元素 $t_1,\cdots,t_n$ 相应的随机变量 $\xi_{t_1},\cdots,\xi_{t_n}$ 的联合分布函数记为

$$F_{t_1\cdots t_n}(x_1,\cdots,x_n)=P\{\xi_{t_1}\le x_1,\cdots,\xi_{t_n}\le x_n\} \qquad (1)$$

由于 $n$ 及 $t_i(i=1,\cdots,n)$ 的任意性，（1）式给出了一族分布函数。记为

$$\{F_{t_1\cdots t_n}(x_1,\cdots,x_n), t_i\in T, i=1,\cdots,n; n=1,2,\cdots\}$$

称它为随机过程 $\{\xi_t, t\in T\}$ 的有穷维分布函数族。它完整地描述了这一随机过程的统计规律性。

6.2 马尔可夫链

6.2.1 马尔可夫链的定义

现实世界中有很多这样的现象：某一系统在已知现在情况的条件下，系统未来时刻的情况只与现在有关，而与过去的历史无直接关系。比如，研究一个商店的累计销售额，如果现在时刻的累计销售额已知，则未来某一时刻的累计销售额与现在时刻以前

的任一时刻累计销售额无关。上节中的几个例子也均属此类。描述这类随机现象的数学模型称为马氏模型。

定义 2 设是一个随机序列，状态空间 $E$ 为有限或可列集，对于任意的正整数 $m,n$，若，有

$$P\{\xi_{n+m}=j\,|\,\xi_n=i,\xi_{n-1}=i_{n-1},\cdots,\xi_1=i_1\}=P\{\xi_{n+m}=j\,|\,\xi_n=i\} \quad (2)$$

则称为一个马尔可夫链（简称马氏链），（2）式称为马氏性。

事实上，可以证明若等式（2）对于 $m=1$ 成立，则它对于任意的正整数 $m$ 也成立。因此，只要当 $m=1$ 时（2）式成立，就可以称随机序列具有马氏性，即是一个马尔可夫链。

定义 3 设是一个马氏链。如果等式（2）右边的条件概率与 $n$ 无关，即

$$P\{\xi_{n+m}=j\,|\,\xi_n=i\}=p_{ij}(m) \quad (3)$$

则称为时齐的马氏链。称 $p_{ij}(m)$ 为系统由状态 $i$ 经过 $m$ 个时间间隔（或 $m$ 步）转移到状态 $j$ 的转移概率。（3）称为时齐性。它的含义是：系统由状态 $i$ 到状态 $j$ 的转移概率只依赖于时间间隔的长短，与起始的时刻无关。本章介绍的马氏链假定都是时齐的，因此省略“时齐”二字。

6.2.2 转移概率矩阵及柯尔莫哥洛夫定理

对于一个马尔可夫链，称以 $m$ 步转移概率 $p_{ij}(m)$ 为元素的矩阵为马尔可夫链的 $m$ 步转移矩阵。当 $m=1$ 时，记 $P(1)=P$ 称为马尔可夫链的一步转移矩阵，或简称转移矩阵。它们具有下列三个基本性质：

（i）对一切 $i,j\in E$，；

（ii）对一切 $i\in E$，$\sum_{j\in E}p_{ij}(m)=1$；

（iii）对一切 $i,j\in E$，$p_{ij}(0)=\delta_{ij}=\begin{cases}1, & \text{当}i=j\text{时}\\ 0, & \text{当}i\neq j\text{时}\end{cases}$。

当实际问题可以用马尔可夫链来描述时，首先要确定它的状态空间及参数集合，然后确定它的一步转移概率。关于这一概率的确定，可以由问题的内在规律得到，也可以由过去经验给出，还可以根据观测数据来估计。

例 4 某计算机机房的一台计算机经常出故障，研究者每隔 15 分钟观察一次计算机的运行状态，收集了 24 小时的数据（共作 97 次观察）。用 1 表示正常状态，用 0 表示不正常状态，所得的数据序列如下：

1110010011111110011110111111001111111110001101101
1110111011010111101110111101111110011011111100111

解 设为第 $n$ 个时段的计算机状态，可以认为它是一个时齐马氏链，状态空间 $E=\{0,1\}$，编写如下 Matlab 程序：

```
a1='1110010011111110011110111111001111111110001101101';
a2='1110111011010111101110111101111110011011111100111';
a=[a1 a2];
f00=length(findstr('00',a))
```

```
f01=length(findstr('01',a))
f10=length(findstr('10',a))
f11=length(findstr('11',a))
```

求得 96 次状态转移的情况是：

$0\to 0$，8 次；　　$0\to 1$，18 次；

$1\to 0$，18 次；　　$1\to 1$，52 次，

因此，一步转移概率可用频率近似地表示为

$$p_{00}=P\{X_{n+1}=0\,|\,X_n=0\}\approx\frac{8}{8+18}=\frac{4}{13}$$

$$p_{01}=P\{X_{n+1}=1\,|\,X_n=0\}\approx\frac{18}{8+18}=\frac{9}{13}$$

$$p_{10}=P\{X_{n+1}=0\,|\,X_n=1\}\approx\frac{18}{18+52}=\frac{9}{35}$$

$$p_{11}=P\{X_{n+1}=1\,|\,X_n=1\}\approx\frac{52}{18+52}=\frac{26}{35}$$

例 5　设一随机系统状态空间 $E=\{1,2,3,4\}$，记录观测系统所处状态如下：

4 3 2 1 4 3 1 1 2 3
2 1 2 3 4 4 3 3 1 1
1 3 3 2 1 2 2 2 4 4
2 3 2 3 1 1 2 4 3 1

若该系统可用马氏模型描述，估计转移概率 $p_{ij}$。

解　首先将不同类型的转移数 $n_{ij}$ 统计出来分类记入下表

$i\to j$ 转移数 $n_{ij}$

| | 1　2　3　4 | 行和 $n_i$ |
|---|---|---|
| 1 | 4　4　1　1 | 10 |
| 2 | 3　2　4　2 | 11 |
| 3 | 4　4　2　1 | 11 |
| 4 | 0　1　4　2 | 7 |

各类转移总和 $\sum_i\sum_j n_{ij}$ 等于观测数据中马氏链处于各种状态次数总和减 1，而行和 $n_i$ 是系统从状态 $i$ 转移到其它状态的次数，$n_{ij}$ 是由状态 $i$ 到状态 $j$ 的转移次数，则 $p_{ij}$ 的估计值 $p_{ij}=\dfrac{n_{ij}}{n_i}$。计算得

$$\hat{P}=\begin{bmatrix}2/5 & 2/5 & 1/10 & 1/10\\ 3/11 & 2/11 & 4/11 & 2/11\\ 4/11 & 4/11 & 2/11 & 1/11\\ 0 & 1/7 & 4/7 & 2/7\end{bmatrix}$$

Matlab 计算程序如下：

```
format rat
clc
a=[4   3   2   1   4   3   1   1   2   3 ...
```

```
   2   1   2   3   4   4   3   3   1   1 ...
   1   3   3   2   1   2   2   2   4   4 ...
   2   3   2   3   1   1   2   4   3   1];
for i=1:4
   for j=1:4
      f(i,j)=length(findstr([i j],a));
   end
end
f
ni=(sum(f'))'
for i=1:4
   p(i,:)=f(i,:)/ni(i);
end
p
```

例 6（带有反射壁的随机徘徊）如果在原点右边距离原点一个单位及距原点 $s(s>1)$ 个单位处各立一个弹性壁。一个质点在数轴右半部从距原点两个单位处开始随机徘徊。每次分别以概率 $p(0<p<1)$ 和 $q(q=1-p)$ 向右和向左移动一个单位；若在+1 处，则以概率 $p$ 反射到 2，以概率 $q$ 停在原处；在 $s$ 处，则以概率 $q$ 反射到 $s-1$，以概率 $p$ 停在原处。设 $\xi_n$ 表示徘徊 $n$ 步后的质点位置。是一个马尔可夫链，其状态空间，写出转移矩阵 $P$ 。

解 $P\{\xi_0 = i\} = \begin{cases} 1, & \text{当} i = 2\text{时} \\ 0, & \text{当} i \neq 2\text{时} \end{cases}$

$$p_{1j} = \begin{cases} q, & \text{当} j = 1\text{时} \\ p, & \text{当} j = 2\text{时} \\ 0, & \text{其它} \end{cases}$$

$$p_{sj} = \begin{cases} p, & \text{当} j = s\text{时} \\ q, & \text{当} j = s-1\text{时} \\ 0, & \text{其它} \end{cases}$$

$$p_{ij} = \begin{cases} p, & \text{当} j - i = 1\text{时} \\ q, & \text{当} j - i = -1\text{时} (i = 2,3,\cdots,s-1) \\ 0, & \text{其它} \end{cases}$$

因此，$P$ 为一个 $s$ 阶方阵，即

$$P = \begin{bmatrix} q & p & 0 & \cdots & 0 & 0 \\ q & 0 & p & \cdots & 0 & 0 \\ 0 & q & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & q & 0 & p \\ 0 & 0 & 0 & 0 & q & p \end{bmatrix}。$$

定理 1 （柯尔莫哥洛夫—开普曼定理）设是一个马尔可夫链，其状态空间 $E=\{1,2,\cdots\}$，则对任意正整数 $m,n$ 有

$$p_{ij}(n+m)=\sum_{k\in E}p_{ik}(n)p_{kj}(m)$$

其中的 $i,j\in E$ 。

定理 2 设 $P$ 是一个马氏链转移矩阵（$P$ 的行向量是概率向量），$P^{(0)}$ 是初始分布行向量，则第 $n$ 步的概率分布为

$$P^{(n)}=P^{(0)}P^{n}\text{。}$$

例 7 若顾客的购买是无记忆的，即已知现在顾客购买情况，未来顾客的购买情况不受过去购买历史的影响，而只与现在购买情况有关。现在市场上供应 $A$、$B$、$C$ 三个不同厂家生产的 50 克袋状味精，用“$\xi_n=1$”、“$\xi_n=2$”、“$\xi_n=3$”分别表示“顾客第 $n$ 次购买 $A$、$B$、$C$ 厂的味精”。显然，是一个马氏链。若已知第一次顾客购买三个厂味精的概率依次为 0.2，0.4，0.4。又知道一般顾客购买的倾向由表 2 给出。求顾客第四次购买各家味精的概率。

表 2

| | | 下 次 购 买 | | |
|---|---|---|---|---|
| | | $A$ | $B$ | $C$ |
| 上次购买 | $A$ | 0.8 | 0.1 | 0.1 |
| | $B$ | 0.5 | 0.1 | 0.4 |
| | $C$ | 0.5 | 0.3 | 0.2 |

解　第一次购买的概率分布为

转移矩阵 $P=\begin{bmatrix}0.8 & 0.1 & 0.1\\ 0.5 & 0.1 & 0.4\\ 0.5 & 0.3 & 0.2\end{bmatrix}$

则顾客第四次购买各家味精的概率为

$$P^{(4)}=P^{(1)}P^{3}=\begin{bmatrix}0.7004 & 0.136 & 0.1636\end{bmatrix}\text{。}$$

6.2.3 转移概率的渐近性质—极限概率分布

现在我们考虑，随 $n$ 的增大，$P^n$ 是否会趋于某一固定向量？先考虑一个简单例子：

转移矩阵 $P=\begin{bmatrix}0.5 & 0.5\\ 0.7 & 0.3\end{bmatrix}$，当 $n\to+\infty$ 时，

$$P^{(n)}\to\begin{bmatrix}\frac{7}{12} & \frac{5}{12}\\ \frac{7}{12} & \frac{5}{12}\end{bmatrix}$$

又若取 $u=\begin{bmatrix}\frac{7}{12} & \frac{5}{12}\end{bmatrix}$，则 $uP=u$，$u^T$ 为矩阵 $P^T$ 的对应于特征值 $\lambda=1$ 的特征（概率）向量，$u$ 也称为 $P$ 的不动点向量。哪些转移矩阵具有不动点向量？为此我们

给出正则矩阵的概念。

定义 4　一个马氏链的转移矩阵 $P$ 是正则的，当且仅当存在正整数 $k$，使 $P^k$ 的每一元素都是正数。

定理 3　若 $P$ 是一个马氏链的正则阵，那么：

（i）$P$ 有唯一的不动点向量 $W$，$W$ 的每个分量为正。

（ii）$P$ 的 $n$ 次幂 $P^n$（$n$ 为正整数）随 $n$ 的增加趋于矩阵 $\overline{W}$，$\overline{W}$ 的每一行向量均等于不动点向量 $W$。

例 8 信息的传播　一条新闻在等人中间传播，传播的方式是 $a_1$ 传给 $a_2$，$a_2$ 传给 $a_3$，…如此继续下去，每次传播都是由 $a_i$ 传给 $a_{i+1}$。每次传播消息的失真概率是 $p$，$0<p<1$，即 $a_i$ 将消息传给 $a_{i+1}$ 时，传错的概率是 $p$，这样经过长时间传播，第 $n$ 个人得知消息时，消息的真实程度如何？

设整个传播过程为随机转移过程，消息经过一次传播失真的概率为 $p$，转移矩阵

$$P=\begin{array}{c} \\ 假 \\ 真 \end{array}\begin{array}{c} \begin{array}{cc} 假 & 真 \end{array} \\ \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix} \end{array}$$

$P$ 是正则矩阵。又设 $V$ 是初始分布，则消息经过 $n$ 次传播后，其可靠程度的概率分布为 $V\cdot P^n$。

一般地，设时齐马氏链的状态空间为 $E$，如果对于所有 $i,j\in E$，转移概率 $p_{ij}(n)$ 存在极限

$$\lim_{n\to\infty} p_{ij}(n)=\pi_j,\quad（不依赖于 i）$$

或

$$P(n)=P^n\xrightarrow[(n\to\infty)]{}\begin{bmatrix} \pi_1 & \pi_2 & \cdots & \pi_j & \cdots \\ \pi_1 & \pi_2 & \cdots & \pi_j & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \pi_1 & \pi_2 & \cdots & \pi_j & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix},$$

则称此链具有遍历性。又若 $\sum_j \pi_j=1$，则同时称为链的极限分布。

下面就有限链的遍历性给出一个充分条件。

定理 4　设时齐（齐次）马氏链的状态空间为，$P=(p_{ij})$ 是它的一步转移概率矩阵，如果存在正整数 $m$，使对任意的 $a_i,a_j\in E$，都有

$$p_{ij}(m)>0,$$

则此链具有遍历性；且有极限分布，它是方程组

$$\pi=\pi P \quad 或即 \quad \pi_j=\sum_{i=1}^{N}\pi_i p_{ij},\quad j=1,\cdots,N$$

的满足条件

$$\pi_j > 0,\ \sum_{j=1}^{N} \pi_j = 1$$

的唯一解。

例 9 根据例 7 中给出的一般顾客购买三种味精倾向的转移矩阵，预测经过长期的多次购买之后，顾客的购买倾向如何？

解　这个马氏链的转移矩阵满足定理 4 的条件，可以求出其极限概率分布。为此，解下列方程组：

$$\begin{cases} p_1 = 0.8p_1 + 0.5p_2 + 0.5p_3 \\ p_2 = 0.1p_1 + 0.1p_2 + 0.3p_3 \\ p_3 = 0.1p_1 + 0.4p_2 + 0.2p_3 \\ p_1 + p_2 + p_3 = 1 \end{cases}$$

编写如下的 Matlab 程序：

```
format rat
p=[0.8 0.1 0.1;0.5 0.1 0.4;0.5 0.3 0.2];
a=[p'-eye(3);ones(1,3)];
b=[zeros(3,1);1];
p_limit=a\b
```

或者利用求转移矩阵 $P$ 的转置矩阵 $P^T$ 的特征值 1 对应的特征(概率)向量，求得极限概率。编写程序如下：

```
p=[0.8 0.1 0.1;0.5 0.1 0.4;0.5 0.3 0.2];
p=sym(p');
[x,y]=eig(p)
for i=1:3
   x(:,i)=x(:,i)/sum(x(:,i));
end
x
```

求得 $p_1 = \frac{5}{7}$，$p_2 = \frac{11}{84}$，$p_3 = \frac{13}{84}$。

这说明，无论第一次顾客购买的情况如何，经过长期多次购买以后，$A$ 厂产的味精占有市场的 $\frac{5}{7}$，$B,C$ 两厂产品分别占有市场的 $\frac{11}{84}$，$\frac{13}{84}$。

6.2.4 吸收链

马氏链还有一种重要类型—吸收链。

若马氏链的转移矩阵为

$$P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0.3 & 0.3 & 0 & 0.4 \\ 0.2 & 0.3 & 0.2 & 0.3 \\ 0 & 0.3 & 0.3 & 0.4 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array},$$

$P$ 的最后一行表示的是，当转移到状态 4 时，将停留在状态 4，状态 4 称为吸收状态。

如果马氏链至少含有一个吸收状态，并且从每一个非吸收状态出发，都可以到达

某个吸收状态，那么这个马氏链被称为吸收链。

具有 $r$ 个吸收状态，$s(s=n-r)$ 个非吸收状态的吸收链，它的 $n\times n$ 转移矩阵的标准形式为

$$P=\begin{bmatrix} I_r & O \\ R & S \end{bmatrix} \qquad (4)$$

其中 $I_r$ 为 $r$ 阶单位阵，$O$ 为 $r\times s$ 零阵，$R$ 为 $s\times r$ 矩阵，$S$ 为 $s\times s$ 矩阵。从（4）得

$$P^n=\begin{bmatrix} I_r & O \\ Q & S^n \end{bmatrix} \qquad (5)$$

（5）式中的子阵 $S^n$ 表示以任何非吸收状态作为初始状态，经过 $n$ 步转移后，处于 $S$ 个非吸收状态的概率。

在吸收链中，令 $F=(I-S)^{-1}$，则 $F$ 称为基矩阵。

对于具有标准形式（即(4)式）转移矩阵的吸收链，可以证明以下定理：

定理 5 吸收链的基矩阵 $F$ 中的每个元素，表示从一个非吸收状态出发，过程到达每个非吸收状态的平均转移次数。

定理 6 设 $N=FC$，$F$ 为吸收链的基矩阵，，则 $N$ 的每个元素表示从非吸收状态出发，到达某个吸收状态被吸收之前的平均转移次数。

定理 7 设，其中 $F$ 为吸收链的基矩阵，$R$ 为（4）式中的子阵，则 $b_{ij}$ 表示从非吸收状态 $i$ 出发，被吸收状态 $j$ 吸收的概率。

例 10 智力竞赛问题　甲、乙两队进行智力竞赛。竞赛规则规定：竞赛开始时，甲、乙两队各记 2 分，在抢答问题时，如果甲队赢得 1 分，那么甲队的总分将增加 1 分，同时乙队总分将减少 1 分。当甲（或乙）队总分达到 4 分时，竞赛结束，甲（或乙）获胜。根据队员的智力水平，知道甲队赢得 1 分的概率为 $p$，失去 1 分的概率为 $1-p$，求：（i）甲队获胜的概率是多少？（ii）竞赛从开始到结束，分数转移的平均次数是多少？（iii）甲队获得 1、2、3 分的平均次数是多少？

**分析**　甲队得分有 5 种可能，即 0、1、2、3、4，分别记为状态，其中 $a_0$ 和 $a_4$ 是吸收状态，$a_1,a_2$ 和 $a_3$ 是非吸收状态。过程是以 $a_2$ 作为初始状态。根据甲队赢得 1 分的概率为 $p$，建立转移矩阵：

$$P=\begin{array}{c} \\ a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{array}\begin{array}{c} \begin{array}{ccccc} a_0 & a_1 & a_2 & a_3 & a_4 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1-p & 0 & p & 0 & 0 \\ 0 & 1-p & 0 & p & 0 \\ 0 & 0 & 1-p & 0 & p \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array} \qquad (6)$$

将（6）式改记为标准形式：

$$P=\begin{bmatrix} I_2 & O \\ R & S \end{bmatrix}$$

其中

$$R=\begin{bmatrix}1-p & 0\\ 0 & 0\\ 0 & p\end{bmatrix},\quad S=\begin{bmatrix}0 & p & 0\\ 1-p & 0 & p\\ 0 & 1-p & 0\end{bmatrix},$$

计算

$$F=(I_3-S)^{-1}=\frac{1}{1-2pq}\begin{bmatrix}1-pq & p & p^2\\ q & 1 & p\\ q^2 & q & 1-pq\end{bmatrix}$$

其中 $q=1-p$ 。

因为 $a_2$ 是初始状态，根据定理 5，甲队获得 1，2，3 分的平均次数为 $\frac{q}{1-2pq}$，$\frac{1}{1-2pq}$，$\frac{p}{1-2pq}$。又

$$N=FC=\frac{1}{1-2pq}\begin{bmatrix}1-pq & p & p^2\\ q & 1 & p\\ q^2 & q & 1-pq\end{bmatrix}$$

$$=\frac{1}{1-2pq}\begin{bmatrix}1+2p^2 & 2 & 1+2p^2\end{bmatrix}$$

根据定理 6，以 $a_2$ 为初始状态，甲队最终获胜的分数转移的平均次数为 $\frac{2}{1-2pq}$。

又因为

$$B=FR=\frac{1}{1-2pq}\begin{bmatrix}(1-pq)p & p^3\\ q^2 & p^2\\ q^3 & (1-pq)p\end{bmatrix}$$

根据定理 7，甲队最后获胜的概率 $b_{22}=\frac{p^2}{1-2pq}$。

Matlab 程序如下：

```
syms p q
r=[q,0;0,0;0,p];
s=[0,p,0;q,0,p;0,q,0];
f=(eye(3)-s)^(-1);f=simple(f)
n=f*ones(3,1);n=simple(n)
b=f*r;b=simple(b)
```

6.3 马尔可夫链的应用

应用马尔可夫链的计算方法进行马尔可夫分析，主要目的是根据某些变量现在的情况及其变动趋向，来预测它在未来某特定区间可能产生的变动，作为提供某种决策的依据。

例 11（服务网点的设置问题）为适应日益扩大的旅游事业的需要，某城市的甲、

乙、丙三个照相馆组成一个联营部，联合经营出租相机的业务。游客可由甲、乙、丙三处任何一处租出相机，用完后，还在三处中任意一处即可。估计其转移概率如表 3 所示：

| | | 还 相 机 处 | | |
|---|---|---|---|---|
| | | 甲 | 乙 | 丙 |
| 租相机处 | 甲 | 0.2 | 0.8 | 0 |
| | 乙 | 0.8 | 0 | 0.2 |
| | 丙 | 0.1 | 0.3 | 0.6 |

今欲选择其中之一附设相机维修点，问该点设在哪一个照相馆为最好？

解　由于旅客还相机的情况只与该次租机地点有关，而与相机以前所在的店址无关，所以可用 $X_n$ 表示相机第 $n$ 次被租时所在的店址；“$X_n=1$”、“$X_n=2$”、“$X_n=3$”分别表示相机第 $n$ 次被租用时在甲、乙、丙馆。则是一个马尔可夫链，其转移矩阵 $P$ 由表 3 给出。考虑维修点的设置地点问题，实际上要计算这一马尔可夫链的极限概率分布。

转移矩阵满足定理 4 的条件，极限概率存在，解方程组

$$\begin{cases} p_1 = 0.2p_1 + 0.8p_2 + 0.1p_3 \\ p_2 = 0.8p_1 + 0.3p_3 \\ p_3 = 0.2p_2 + 0.6p_3 \\ p_1 + p_2 + p_3 = 1 \end{cases}$$

得极限概率 $p_1 = \dfrac{17}{41}$，$p_2 = \dfrac{16}{41}$，$p_3 = \dfrac{8}{41}$。

由计算看出，经过长期经营后，该联营部的每架照相机还到甲、乙、丙照相馆的概率分别为 $\dfrac{17}{41}$、$\dfrac{16}{41}$、$\dfrac{8}{41}$。由于还到甲馆的照相机较多，因此维修点设在甲馆较好。但由于还到乙馆的相机与还到甲馆的相差不多，若是乙的其它因素更为有利的话，比如，交通较甲方便，便于零配件的运输，电力供应稳定等等，亦可考虑设在乙馆。

## 习　题

1. 在英国，工党成员的第二代加入工党的概率为 0.5，加入保守党的概率为 0.4，加入自由党的概率为 0.1。而保守党成员的第二代加入保守党的概率为 0.7，加入工党的概率为 0.2，加入自由党的概率为 0.1。而自由党成员的第二代加入保守党的概率为 0.2，加入工党的概率为 0.4，加入自由党的概率为 0.4。求自由党成员的第三代加入工党的概率是多少？在经过较长的时间后，各党成员的后代加入各党派的概率分布是否具有稳定性？

2. 社会学的某些调查结果指出：儿童受教育的水平依赖于他们父母受教育的水平。调查过程是将人们划分为三类：$E$ 类，这类人具有初中或初中以下的文化程度；$S$ 类这类人具有高中文化程度；$C$ 类，这类人受过高等教育。当父或母（指文化程度较高者）是这三类人中某一类型时，其子女将属于这三种类型中的任一种的概率由下面给出

$$
\begin{array}{cc} & \text{子女} \\ & \begin{array}{ccc} E & S & C \end{array} \\ \begin{array}{c} \text{父} \\ \text{或} \\ \text{母} \end{array} & \begin{array}{c} E \\ S \\ C \end{array}\begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.4 & 0.4 & 0.2 \\ 0.1 & 0.2 & 0.7 \end{bmatrix} \end{array}
$$

问：（i）属于$S$类的人们中，其第三代将接受高等教育的概率是多少？

（ii）假设不同的调查结果表明，如果父母之一受过高等教育，那么他们的子女总可以进入大学，修改上面的转移矩阵。

（iii）根据（ii）的解，每一类型人的后代平均要经过多少代，最终都可以接受高等教育？

3. 色盲是$X-$链遗传，由两种基因$A$和$a$决定。男性只有一个基因$A$或$a$，女性有两个基因$AA$、$Aa$或$aa$，当基因为$a$或$aa$时呈现色盲。基因遗传规律为：男性等概率地取母亲的两个基因之一，女性取父亲的基因外又等概率地取母亲的两个基因之一。由此可知，母亲色盲则儿子必色盲但女儿不一定。试用马氏链研究：

（i）若近亲结婚，其后代的发展趋势如何？若父亲非色盲而母亲色盲，问平均经多少代，其后代就会变为全色盲或全不色盲，两者的概率各为多少？

（ii）若不允许双方均色盲的人结婚，情况会怎样？