

概率统计建模方法

第1章 概率方法建模简介

第5章 马氏链模型

第2章 数据统计描述和分析

第6章 时间序列模型

第3章 方差分析

第7章 主成分分析及应用

第4章 回归分析

第8章 判别分析简介及应用

Shandong University

School of Mathematics

Chenjianliang

第1章 概率方法建模简介

自然界中的现象总的来说可以概括为两大现象：

确定性现象和随机现象

在确定性现象中可以忽略随机因素的影响，

在随机现象中必须考虑随机因素的影响。

确定性离散模型，主要使用差分方程方法、层次分析方法以及比较简单的图的方法和逻辑方法等方法建立模型；

确定性连续模型，主要使用微积分、微分方程及其稳定性、变分方法等方法建立模型；

随机性模型，是指研究的对象包含有随机因素的规律，以概率统计为基本数学工具，其结果通常也是在概率意义下表现出来。随机因素的影响可以用概率、平均值（即数学期望）等的作用来体现。

§ 1 概率论基础知识

一、几个重要计算概率公式

1. 加法公式

$$P(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i) - \sum_{1 \leq i < j \leq n} P(A_i A_j) + \sum_{1 \leq i < j < k \leq n} P(A_i A_j A_k) + \cdots + (-1)^{n-1} P(\bigcap_{i=1}^n A_i)$$

2. 条件概率公式

$$P(A|B) = \frac{P(AB)}{P(B)}$$

3. 乘法公式

若 $P(A_1 A_2 \cdots A_{n-1}) > 0$ 则

$$P(A_1 A_2 \cdots A_n) = P(A_1)P(A_2 | A_1) \cdots P(A_n | A_1 A_2 \cdots A_{n-1})$$

4. 全概率公式

A_1, A_2, \dots 为 Ω 的完备事件组，则

$$P(B) = \sum_{i=1}^{\infty} P(B | A_i) \cdot P(A_i)$$

5. 贝叶斯(Bayes)公式

A_1, A_2, \dots 为 Ω 的完备事件组，则，

$$P(A_i | B) = \frac{P(B | A_i) \cdot P(A_i)}{\sum_{i=1}^{\infty} P(B | A_i) \cdot P(A_i)}$$

二、常见概率分布及其数字特征

1. 二项分布 $X \sim B(n, p)$

$$P_k = P\{X = k\} = C_n^k p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n$$

$$EX = np, \quad DX = np(1-p)$$

2. 泊松分布 $P(\lambda)$

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

3. 负二项分布（巴斯卡分布）

记 $C_k = \{\text{第 } r \text{ 次成功发生在第 } k \text{ 次试验上}\}$ ，则其概率为

$$f(k; r, p) = C_{k-1}^{r-1} p^r q^{k-r}$$

$r=1$ 时即为几何分布（也称为等待分布）---无记忆性

4. 区间 $[a,b]$ 上的均匀分布记作 $U[a,b]$

$$\text{密度函数 } f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a,b] \\ 0, & x \notin [a,b] \end{cases} \quad \begin{cases} EX = \frac{a+b}{2} \\ DX = \frac{(b-a)^2}{12} \end{cases}$$

5. 指数分布(具有无记忆性)

$$\text{密度函数 } f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}, EX = \frac{1}{\lambda}, DX = \frac{1}{\lambda^2}$$

6. 正态分布 $N(\mu, \sigma^2)$

$$\text{密度函数 } f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in R$$

$$EX = \mu, DX = \sigma^2$$

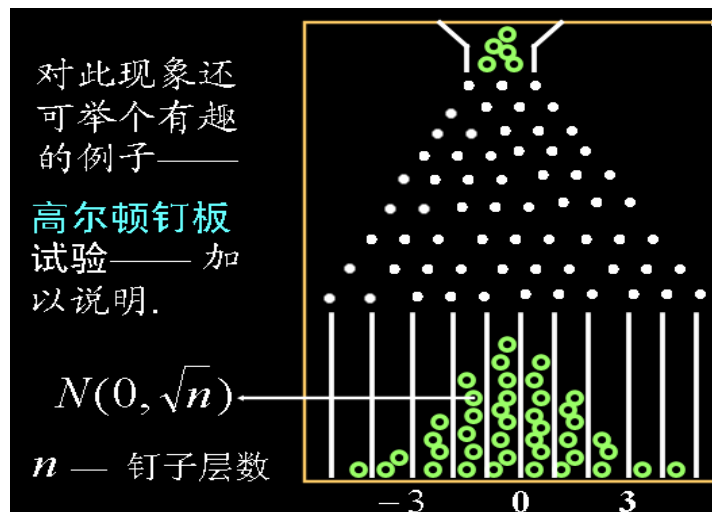
还有一些分布，例如伽玛分布、威布尔分布、贝塔分布等。

三、林德贝格-勒维中心极限定理

设 X_1, \dots, X_n 独立同分布，且有有限的期望和方差，则

$$\frac{\sum_{k=1}^n X_k - E\left(\sum_{k=1}^n X_k\right)}{\sqrt{\sum_{k=1}^n DX_k}} \stackrel{\text{近似}}{\sim} N(0,1)$$

例如：高尔顿钉板试验



§ 2 概率方法建模实例分析

实例一、报童的策略问题

1. 问题描述

报童每天清晨从报站批发报纸零售，晚上将未卖完的报纸退回。设每份报纸的批发价为 b ，零售价为 a ，退回价为 c ，且设 $a > b > c$ ，因此报童每售出一份报纸赚 $(a-b)$ ，退回一份赔 $(b-c)$ 。若批少了不够买就会少赚，若批多了买不完就赔钱，报童如何确定每天批发报纸的数量，才能获得最大收入？

2. 分析

显然应根据需求量来确定批发量。一种报纸的需求量是一随机变量。假定报童通过自己的实践经验或其它方式掌握了需求量的随机规律，即在他的销售范围内每天报纸的需求量为 $X = x$ 份的概率为 $P(x)$ ，则通过 $P(x)$ 和 a, b, c 就可建立关于批发量的优化模型。

3. 数学模型

设每天批发量为 n ，因需求量 x 是随机的，因此 x 可以小于、等于或大于 n ，从而报童每天的收入也是随机的，作为优化模型的目标函数，应考虑他长期（半年、一年等）卖报的日平均收入。据概率论中的大数定律，这相当于报童每天收入的期望值（以下简称平均收入）。

3. 数学模型

设报童每天批发进 n 份报纸时的平均收入为 $S(n)$ ，若某天需求量 $x \leq n$ ，则他售出 x 份，退回 $(n-x)$ 份；若这天需求量 $x > n$ ，则 n 份报纸全部卖出。因需求量为 x 的概率为 $P(x)$ ，故平均收入为

$$S(n) = \sum_{x=0}^n [(a-b)x - (b-c)(n-x)]P(x) + \sum_{x=n+1}^{\infty} (a-b)nP(x)$$

所需考虑的问题变为：

当 $P(x)$ 及 a, b, c 已知时，求使 $S(n)$ 达到最大值的 n 。

4. 模型求解

为便于分析和计算，同时考虑需求量 x 的取值和 n 都相当大，故将 x 视为连续变量，这时概率 $P(x)$ 转化为概率密度函数 $f(x)$ （或分布函数）， $S(n)$ 的表达式变为

$$S(n) = \int_0^n [(a-b)x - (b-c)(n-x)] f(x) dx + \int_n^\infty (a-b)n f(x) dx$$

令 $\frac{dS(n)}{dn} = 0$ 得

$$(a-b)n f(n) - \int_0^n (b-c) f(x) dx - (a-b)n f(n) + \int_n^\infty (a-b) f(x) dx$$

$$= -(b-c) \int_0^n f(x) dx + (a-b) \int_n^\infty f(x) dx = 0$$

解得 $\frac{\int_0^n f(x) dx}{\int_n^\infty f(x) dx} = \frac{a-b}{b-c} \dots\dots\dots (1)$

因此使报童日平均收入达最大值的批发量 n 应满足(1)式

$$\because \int_0^{\infty} f(x)dx = 1 \quad \text{故 (1) 式又可写为} \quad \int_0^n f(x)dx = \frac{a-b}{a-c}$$

$$\text{令 } P_1 = \int_0^n f(x)dx, \quad P_2 = \int_n^{\infty} f(x)dx$$

则当批发进 n 份报纸时, P_1 是需求量 x 不超过 n 的概率, 即卖不完的概率, P_2 是需求量 x 超过 n 的概率, 即卖完的概率, 由(1)知批发的报纸份数 n 应使得卖不完与卖完的概率之比等于卖出一份赚的钱 $(a-b)$ 与退回一份赔的钱 $(b-c)$ 之比

综上所述, 当每份报纸赚钱与赔钱之比越大时, 报童批发进的报纸份数应该越多.

下面采用随机离散的单时期存贮模型方法求解。

随机离散的单时期存贮模型方法求解

设报童每天售报数量是一个离散型随机变量，设销售量 x 的概率分布 $P(x)$ 为已知，每张报纸的销售价为 a 元，成本为 b 元（ $a > b$ ）。如果报纸当天卖不出去，第二天就降低处理，设处理价为 c 元（ $c < b$ ）。问报童每天最好准备多少份报纸？

这个问题就是要确定报童每天报纸的订货量 n 为何值时，使赢利的期望值最大或损失的期望值最小？

解法一 计算损失的期望值最小

设售出报纸数量为 x ，其概率为 $P(x)$ 为已知， $\sum_{x=0}^{+\infty} P(x) = 1$

设报童订购报纸数量为 n ，这时的损失有两种

(1) 当供大于求 ($n \geq x$) 时，报纸因当天不能售完，第二天

需降价处理，损失的期望值 $\sum_{x=0}^{+\infty} (b-c)(n-x)P(x)$

(2) 当供不应求 ($n < x$) 时，因缺货而失去销售机会，损失的

期望值 $\sum_{x=n+1}^{+\infty} (a-b)(x-n)P(x)$

故总损失的期望值

$$L(n) = (b-c) \sum_{x=0}^{+\infty} (n-x)P(x) + (a-b) \sum_{x=n+1}^{+\infty} (x-n)P(x)$$

要从上式中决定 n 的值，使 $L(n)$ 最小

由于报纸订购的份数 n 只能取整数，需求量 x 也只能取整数，即都是离散变量，所以不能用微积分的方法求 $L(n)$ 式的极值，为此用差分法，设报童每天订购的报纸的最佳批量为

n^* ,必有

$$\begin{cases} L(n^*) \leq L(n^* + 1) \\ L(n^*) \leq L(n^* - 1) \end{cases}$$

同时成立。故将上两不等式联立求解可得最佳批量 n^*

由第一式得

$$\begin{aligned} (b-c) \sum_{x=0}^n (n-x)P(x) + (a-b) \sum_{x=n+1}^{+\infty} (x-n)P(x) \\ \leq (b-c) \sum_{x=0}^{n+1} (n+1-x)P(x) + (a-b) \sum_{x=n+2}^{+\infty} (x-n-1)P(x) \end{aligned}$$

简化后得

$$(a-c) \sum_{x=0}^n P(x) - (a-b) \geq 0 \quad \text{即} \quad \sum_{x=0}^n P(x) \geq \frac{a-b}{a-c}$$

由第二式得
$$(b-c) \sum_{x=0}^n (n-x)P(x) + (a-b) \sum_{x=n+1}^{+\infty} (x-n)P(x)$$

$$\leq (b-c) \sum_{x=0}^{n-1} (n-1-x)P(x) + (a-b) \sum_{x=n}^{+\infty} (x-n+1)P(x)$$

简化后得
$$(a-c) \sum_{x=0}^{n-1} P(x) - (a-b) \leq 0 \quad \text{即} \quad \sum_{x=0}^{n-1} P(x) \leq \frac{a-b}{a-c}$$

由
$$\sum_{x=0}^{n-1} P(x) \leq \frac{a-b}{a-c} \leq \sum_{x=0}^n P(x)$$

可以确定最佳订购批量 n^* (称为**临界值**), 满足

$$\sum_{x=0}^{n^*-1} P(x) \leq \frac{a-b}{a-c} \leq \sum_{x=0}^{n^*} P(x)$$

解法二 计算盈利的期望值最大: 这时也可以分两种情况

(1) 供大于求($n \geq x$)时, 这时只能售出 x 份报纸, 故可赚 $(a-b)x$ 。未售出的报纸降价处理后, 每份损失 $(b-c)$, 共损失为 $(b-c)(n-x)$ 。因此, 赢利的期望值为

$$\sum_{x=0}^n [(a-b)x - (b-c)(n-x)] P(x)$$

(2) 当供不应求($n < x$)时, 这时只有 n 份报纸可供销售, 故无

滞销损失。因此, 盈利的期望值为 $\sum_{x=n+1}^{+\infty} (a-b)nP(x)$

故总盈利的期望值为

$$C(n) = \sum_{x=0}^n [(a-b)x - (b-c)(n-x)] P(x) + \sum_{x=n+1}^{+\infty} (a-b)nP(x)$$

故最佳订购批量 n^* 应满足
$$\begin{cases} C(n^*) \geq C(n^* + 1) \\ C(n^*) \geq C(n^* - 1) \end{cases}$$

同时成立。故将上两不等式联立求解可得最佳批量 n^*

由第一式得
$$\sum_{x=0}^n [(a-b)x - (b-c)(n-x)]P(x) + \sum_{x=n+1}^{+\infty} (a-b)nP(x)$$
$$\geq \sum_{x=0}^{n+1} [(a-b)x - (b-c)(n+1-x)]P(x) + \sum_{x=n+2}^{+\infty} (a-b)(n+1)P(x)$$

简化后得
$$(a-b) \left[1 - \sum_{x=0}^n P(x) \right] - (b-c) \sum_{x=n+1}^{+\infty} (a-b)nP(x) \leq 0$$

即
$$\sum_{x=0}^n P(x) \geq \frac{a-b}{a-c}$$

由第二式得 $\sum_{x=0}^{n-1} P(x) \leq \frac{a-b}{a-c}$

综合得 $\sum_{x=0}^{n-1} P(x) \leq \frac{a-b}{a-c} \leq \sum_{x=0}^n P(x)$

可以确定最佳订购批量 n^* (称为**临界值**), 满足

$$\sum_{x=0}^{n^*-1} P(x) \leq \frac{a-b}{a-c} \leq \sum_{x=0}^{n^*} P(x)$$

与解法一的结论一致。

由上面的两种解法可看出，尽管报童问题中损失最小的期望值与盈利最大的期望值是不同的，但确定最佳订购批量的条件是相同的，即无论从哪一方面来考虑，最佳订购批量是一个确定数值。另外，本模型有一个严格的约定，即两次订货之间没有联系，都看作独立的一次订货，这也是单时期模型的含义。这种存贮策略也可以称为**定期定量订货**。

类似的问题 设某货物的需求量在**17**件至**26**件之间，已知需求量 x 的概率分布如下表

需求量 x	17	18	19	20	21	22	23	24	25	26
概率 $P(x)$	0.12	0.18	0.23	0.13	0.10	0.08	0.05	0.04	0.04	0.03

并知其成本为每件**5**元，售价为每件**10**元，处理价为每件**2**元。问应进货多少，能使总利润的期望值最大？

解 此题属于单时期需求是离散随机变量的存贮模型

已知 $b=5$, $a=10$, $c=2$, 得
$$\sum_{x=17}^{n-1} P(x) \leq \frac{10-5}{10-2} = 0.625 \leq \sum_{x=17}^n P(x)$$

因为 $P(17)=0.12$, $P(18)=0.18$, $P(19)=0.23$, $P(20)=0.13$

所以 $P(17)+P(18)+P(19)=0.53 < 0.625$

$$P(17)+P(18)+P(19)+P(20)=0.66 > 0.625$$

故最佳订货批量 $n^*=20$ (件)

实例二、轧钢问题

1. 问题描述

用连续热轧方法生产钢材一般要经过两道工序，第一道是**热轧**(粗轧)，形成钢材的雏形；第二道是**冷轧**(精轧)，得到最后成品。由于受设备、环境等方面随机因素的影响，钢材经热轧再冷却后的长度大致服从正态分布，其均值可在轧制过程中由轧机调整，而其均方差由设备精度决定，无法随意改变。冷轧后把多出规定长度的部分切除，但若热轧后钢材长度已经比规定长度短，则整根钢材报废。冷轧设备精度很高，轧出的成品材可以认为完全符合规定长度要求。

根据轧制工艺要求，要在**成品材规定长度 l** 和热轧后钢材长度的均方差 σ 已知的条件下，确定热轧后钢材长度的均值 m ，使得当轧机调整到 m 进行热轧，在通过冷轧以得到成品材时的浪费最少。

2. 问题分析

设热轧后的钢材长度为 x ，则 x 服从正态分布 $N(m, \sigma^2)$

密度函数
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}, (-\infty < x < \infty)$$

其中 $\sigma > 0$ 已知， m 待定

当成品材的规定长度 l 给定后，记 $x \geq l$ 的概率为 P ，即

$$P = P(x \geq l) = \int_l^{\infty} f(x) dx$$

轧制过程中的浪费包括以下两部分：

- (1) 冷轧后钢材的长度超出成品材规定长度时造成的浪费；
- (2) 热轧后钢材的长度达不到成品材规定长度时造成的整根钢材报废。

由于 $P(x \geq l) + P(x < l) = 1$ ，故第一部分的浪费增加（或减少）时，第二部分的浪费减少（或增加）。我们的目的是找出一个最佳的均值，使得两部分的浪费综合起来最小。

这是一个优化模型，建模的关键是选择恰当的目标函数，并用已知的和待定的量 l, σ, m 将其表示出来。自然地，可能会想到把上述两部分的浪费加起来作为目标函数，于是可得总的浪费长度为

$$W = \int_{-\infty}^l x f(x) dx + \int_l^{\infty} (x - l) f(x) dx$$

$$\because \int_{-\infty}^{\infty} f(x) dx = 1, \int_{-\infty}^{\infty} x f(x) dx = m$$

$$\therefore W = m - l P \quad \dots\dots\dots (2)$$

W 是每热轧一根钢材所浪费的长度，假设共热轧了 N 根钢材(一般 N 很大)，所用钢材总长为 mN ，而 N 根钢材中可以轧出成品材的只有 PN 根，成品总长为 lPN ，故浪费的长度为 $(mN-lPN)$ ，平均每根浪费长度为 $(mN-lPN)/N=m-lP$ ，此式与(2)是一致的，但以 W 作为目标函数是否合理呢？

轧钢的最终产品是成品材，浪费的多少不应以每热轧一根钢材的平均浪费量来衡量，而应用每得到一根成品材所浪费的平均长度来衡量，故应以每得到一根成品材所浪费钢材的平均长度作为目标函数。

3. 数学模型

热轧 N 根钢材，得到 PN 根成品材，浪费的总长度为 $(mN-lPN)$ ，因此目标函数为

$$J_1 = \frac{(mN - lPN)}{PN} = \frac{m}{P} - l$$

由于 l 为已知常数，故等价地取目标函数

$$J(m) = \frac{m}{P(m)} \dots\dots\dots (3)$$

这里 $P(m)=P(x \geq l)$ 是 m 的函数。要求出 m ，使 $J(m)$ 达最小

4. 模型求解

设 $F(x)$ 为 $N(m, \sigma^2)$ 的分布函数, $\Phi(x)$ 和 $p(x)$ 为 $N(0,1)$ 的分布函数和密度函数, 则

$$J(m) = \frac{m}{P(m)} = \frac{m}{1 - F(l)} = \frac{m}{1 - \Phi\left(\frac{l-m}{\sigma}\right)}$$

$$\text{令 } \mu = \frac{m}{\sigma}, \lambda = \frac{l}{\sigma} \dots\dots\dots (4)$$

$$\text{则(3)可表示为 } J(\mu) = \frac{\sigma \mu}{1 - \Phi(\lambda - \mu)} \quad \text{进一步令}$$

$$z = \lambda - \mu \dots\dots\dots (5)$$

$$\text{则(3)也可表示为 } J(z) = \frac{\sigma(\lambda - z)}{1 - \Phi(z)}$$

利用微分求函数极值，易知极值点(最优值) z^* 应满足

$$\lambda - z = \frac{1 - \Phi(z)}{p(z)} \triangleq G(z) \quad \dots\dots\dots (6)$$

$\Phi(x)$ 和 $p(x)$ 的值可查表得到，从(6) 解得 z^* ，代入
(5),(4)得 m 的最优值 m^* ，易知当 $z < 0$ 时 $\frac{dG(z)}{dz} < 0$

故(6)仅有唯一负根 z^* ，再进一步地可证得 $\left. \frac{d^2 J(z)}{dz^2} \right|_{z^*} > 0$

因此 z^* 使得 $J(z)$ 取得极小值 $J(z^*)$

5. 举例

设要轧制长为 $l=2$ 米的成品钢材，由热轧设备等因素决定的热轧冷却后钢材长度的均方差 $\sigma=0.1$ 米。问这时钢材长度的均值应调整到多少才能使浪费最少？

解 由于 $\lambda = \frac{l}{\sigma} = 20$ 解方程(6)得 $z^* = -2.1$

代入(5)及(4)得 $\mu=22.1$, $m^* = 2.21$

即最佳均值应调整为2.21米，又可算出

$P(m^*)=0.9821$ ，故每一根成品材浪费钢材平均长度为

$$J_1 = \frac{m^*}{P(m^*)} - l = 0.25 \text{ 米}$$

6. 模型评注

模型中假定热轧后钢材的长度达不到成品材规定长度时造成的整根钢材报废，实际上，当热轧后钢材的长度达不到成品材规定长度时，还可留着需要短一些的成品材使用或降级使用，此时模型的建立和求解更复杂。

第2章 数据统计描述和分析

数理统计研究的对象是受随机因素影响的数据，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;

总体中的每一个基本单位称为个体.

从总体中随机产生的若干个个体的集合称为**样本或子样**

统计的任务是由样本推断总体.

1.2 频数表和直方图

将数据取值范围划分为若干个区间, 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，当样本容量较大时可借助**Matlab**这样的软件。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**，它是一个包含有个数据的矩阵。

为了得到需要的**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); %%取出身高数据（5列）  
high=high(:); %%转换身高数据成1列  
weight=data(:,2:2:10); %%取出体重数据（5列）  
weight=weight(:) %%转换体重数据成1列
```

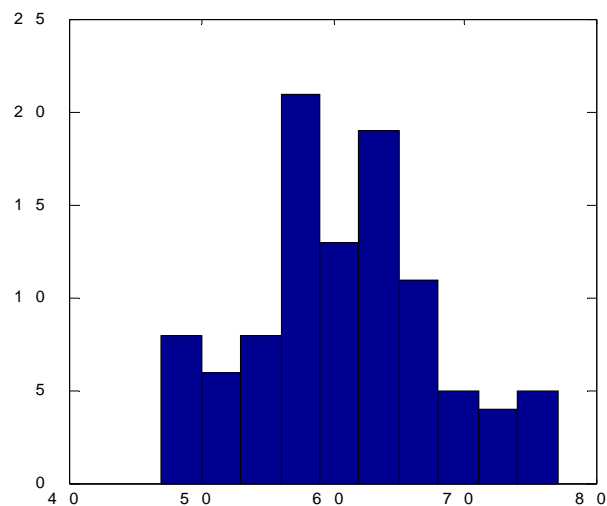
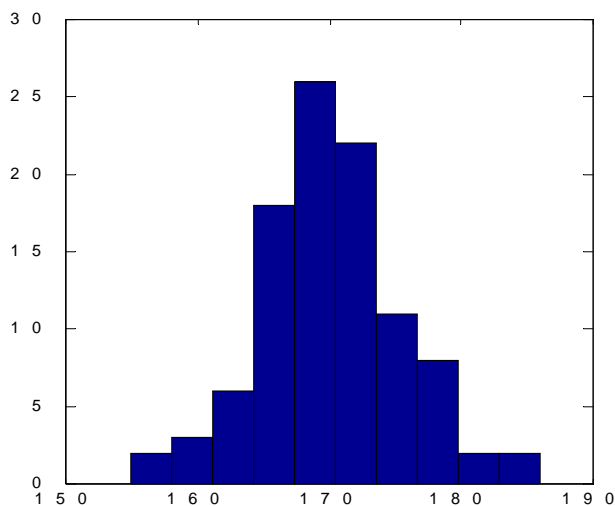
(ii) 作频数表及直方图用 **hist** 命令实现

用法是： **[N,X] = hist(Y,M)**

作数组（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)作数组**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(:);**
- **%绘制直方图**
- **[n1,x1]=hist(high)**
- **[n2,x2]=hist(weight)**
- **subplot(1,2,1) %%绘制第一个直方图**
- **hist(high)**
- **subplot(1,2,2) %%并排绘制第二个直方图**
- **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**、**c**、**g**、**t**出现的频数

1. **aggcacggaaaaacgggaataacggaggaggacttggcacggcattacacggagg**
2. **cggaggacaaacgggatggcgggtattggagggtggcggactgttcggggga**
3. **gggacgggatacggattctggccacggacgggaaaggaggacacggcggacataca**
4. **atggataacgggaaacaaaccagacaaacttcggtagaaatacagaagctta**
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);
```


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**、**c**、**g**、**t** 四个字符。每一行包含**a**、**c**、**g**、**t** 的频数及小计、含其它字符小计如下

f = 19 10 21 5 55 55

10 7 24 8 49 49

16 12 21 5 54 54

24 9 10 8 51 51

14 13 24 4 55 55

a、c、g、t、合计、含其它字符合计

he = 83 51 100 30 264 264

pinshu.txt

19	10	21	5	55	55
10	7	24	8	49	49
16	12	21	5	54	54
24	9	10	8	51	51
14	13	24	4	55	55

1.3 统计量 统计量是加工出来的、反映样本数量特征的函数，它不含任何未知量,是样本的函数。

常用的统计量:

平均值 $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

中位数 $Median = \begin{cases} x_{(\frac{n}{2})}, & n \text{ 是奇数时} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}, & n \text{ 是偶数时} \end{cases}$

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

Matlab中**mean(x)**返回x的均值，**median(x)**返回中位数；若标准差的定义中将分母(n-1)改为n，可用**std(x,1)**和**var(x,1)**来实现。

极差 $R_n = x_{(n)} - x_{(1)}$

表示分布形状的统计量的

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

偏度反映分布的对称性， $\nu_1 > 0$ 右偏态； $\nu_1 < 0$ 左偏态

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

峰度用作衡量偏离正态分布的尺度之一，正态分布峰度为**3**，若峰度比**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。

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)
```

运行结果:

jun_zhi = 170.2500 61.2700

zhong_wei_shu = 170 62

biao_zhun_cha = 5.4018 6.8929

ji_cha = 31 30

pian_du = 0.1545 0.1380

feng_du = 3.5573 2.6644

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

正态分布 $N(\mu, \sigma^2)$

卡方分布 $\chi^2(n)$

t 分布 $t(n)$

F 分布 $F(m, n)$

Matlab统计工具箱4个概率分布命令:

norm - 正态分布; **chi2** - 卡方分布; **t** - t 分布; **f** - F 分布

对每一种分布都提供5类函数:

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

例如

p = normpdf(x,mu,sigma)

表示均值**mu**、标准差**sigma**的正态分布在**x**的密度函数值
(**mu=0**, **sigma=1**时可缺省)。

p = tcdf(x,n) 是自由度为 **n** 的 **t** 分布在**x** 处的分布函数值。

x=chi2inv(p,n) 自由度为 **n** 的卡方分布的 **p** 分位数。

x=chi2inv(0.9,10) 得 **x = 15.9872**

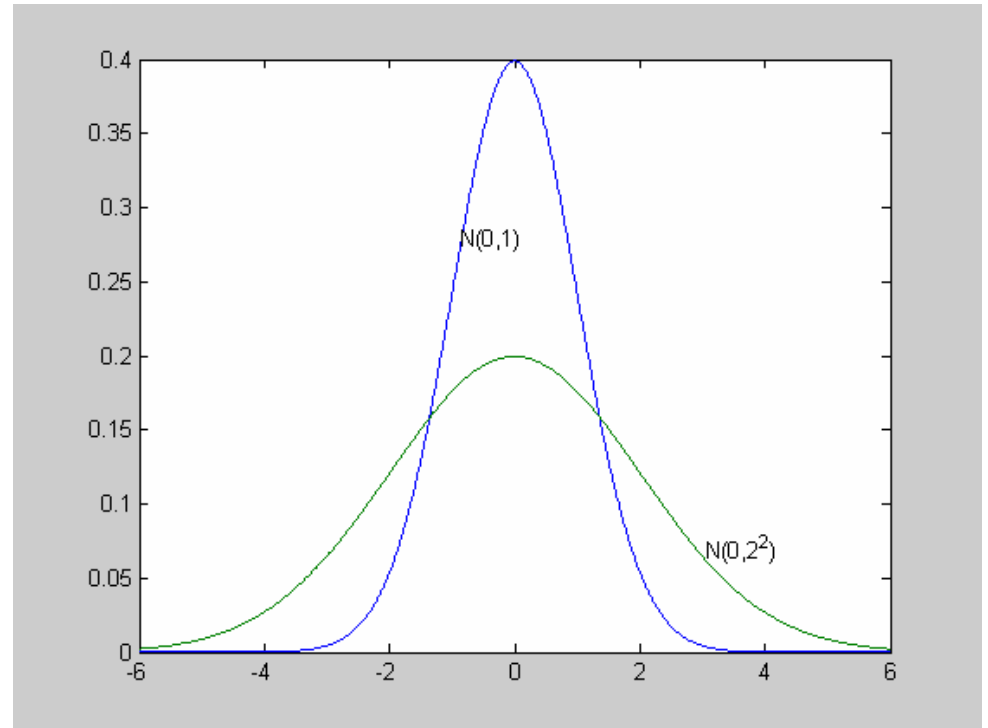
p=chi2cdf(15.9872,10) 得 **p = 0.9000**

[m,v] = fstat(n1,n2)

是**F**分布(自由度**n1,n2**)的均值 **m** 和方差 **v**

分布的密度函数图形可以用这些命令作出，例如

```
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)')
```



1.5 正态总体统计量的分布

单个正态总体情形 $X \sim N(\mu, \sigma^2)$

$$1). \bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

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

$$3). \frac{\bar{x} - \mu}{s / \sqrt{n}} \sim t(n-1)$$

1.5 正态总体统计量的分布

两个正态总体情形 $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$

$$4). \frac{(\bar{x} - \mu_1) - (\bar{y} - \mu_2)}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}} \sim N(0, 1)$$

$$5). \frac{(\bar{x} - \mu_1) - (\bar{y} - \mu_2)}{s \sqrt{1/n_1 + 1/n_2}} \sim t(n_1 + n_2 - 2)$$

$$\text{其中 } s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$6). \frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$

§ 2 参数估计

2.1 点估计

用样本统计量确定总体参数的一个数值。

$$\widehat{EX} = \bar{x} \quad , \quad \widehat{DX} = s^2$$

点估计的两种常用方法

矩估计 $EX^k = \frac{1}{n} \sum_{i=1}^n X_i^k$

极大似然估计 $\hat{\theta} : L(\hat{\theta}) = \sup_{\theta \in \Theta} \prod_{i=1}^n f(x_i; \theta)$

评价估计优劣的标准有无偏性、有效性、一致性(相合性)等

无偏性 $E\hat{\theta} = \theta$

有效性 对无偏估计 $\hat{\theta}_1, \hat{\theta}_2$ 若 $D\hat{\theta}_1 \leq D\hat{\theta}_2$
则称 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 有效。

一致性 $\forall \varepsilon > 0, \lim_{n \rightarrow \infty} P\{|\hat{\theta}_n - \theta| > \varepsilon\} = 0$

$$\Leftrightarrow \lim_{n \rightarrow \infty} E(\hat{\theta}_n) = \theta, \lim_{n \rightarrow \infty} D(\hat{\theta}_n) = 0$$

2.2 区间估计

$$P\{\hat{\theta}_1 < \theta < \hat{\theta}_2\} = 1 - \alpha$$

置信区间: $[\hat{\theta}_1, \hat{\theta}_2]$

置信度(置信水平): $1 - \alpha$

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

单个正态总体情形 $X \sim N(\mu, \sigma^2)$ μ 的置信区间

方差 σ^2 已知 $\left[\bar{x} - \mu_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + \mu_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right]$

方差 σ^2 未知 $\left[\bar{x} - \frac{s}{\sqrt{n}} \cdot t_{\alpha/2}(n-1), \bar{x} + \frac{s}{\sqrt{n}} \cdot t_{\alpha/2}(n-1) \right]$

单个正态总体情形 $X \sim N(\mu, \sigma^2)$ σ^2 的置信区间

μ 已知 $\left[\frac{\sum_{i=1}^n (x_i - \mu)^2}{\chi_{\alpha/2}^2(n)}, \frac{\sum_{i=1}^n (x_i - \mu)^2}{\chi_{1-\alpha/2}^2(n)} \right]$

μ 未知 $\left[\frac{(n-1)s^2}{\chi_{\alpha/2}^2(n-1)}, \frac{(n-1)s^2}{\chi_{1-\alpha/2}^2(n-1)} \right]$

两个正态总体情形 $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$

方差均已知时 $\mu_1 - \mu_2$ 的置信区间

$$\left[\bar{x} - \bar{y} \pm \mu_{\alpha/2} \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} \right]$$

方差均未知但相等时 $\mu_1 - \mu_2$ 的置信区间

$$\left[\bar{x} - \bar{y} \pm t_{\alpha/2}(m+n-2) S_w \sqrt{\frac{1}{m} + \frac{1}{n}} \right]$$

$$\text{其中 } S_w^2 = \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}$$

两个正态总体情形 $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$

1) μ_1 、 μ_2 均已知时方差比 σ_1^2 / σ_2^2 的置信区间

$$\left[\frac{\sum_{i=1}^m (x_i - \mu_1)^2 / m}{\sum_{i=1}^n (y_i - \mu_2)^2 / n} \cdot \frac{1}{F_{\frac{\alpha}{2}}(m, n)}, \frac{\sum_{i=1}^m (x_i - \mu_1)^2 / m}{\sum_{i=1}^n (y_i - \mu_2)^2 / n} \cdot \frac{1}{F_{1-\frac{\alpha}{2}}(m, n)} \right]$$

2) μ_1 、 μ_2 均未知时方差比 σ_1^2 / σ_2^2 的置信区间

$$\left[\frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{\alpha/2}(m-1, n-1)}, \frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{1-\alpha/2}(m-1, n-1)} \right]$$

2.3 参数估计的Matlab实现

[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 α （**alpha**缺省时设定为**0.05**），返回总体均值 μ 和标准差 σ 的点估计**mu**和**sigma**，及总体均值 μ 和标准差 σ 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

3.1 单个总体 $N(\mu, \sigma^2)$ 均值 μ 的检验

原假设（零假设） $H_0 : \mu = \mu_0$

备选假设 $H_1 : \mu \neq \mu_0$;或 $H_1 : \mu > \mu_0$;或 $H_1 : \mu < \mu_0$

σ^2 已知，均值 μ 的检验

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

`tail=0`(缺省)，**1**，**2** 分别对应于上述三种备选假设之一

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 μ_0 的置信区间

例3 某车间用一台包装机包装糖果。包的袋装糖重是一个随机变量，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 $x \sim N(\mu, 0.015^2)$, $H_0: \mu = \mu_0 = 0.5 \leftrightarrow H_1: \mu \neq 0.5$

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**, **ci=[0.5014 0.5210]**

在**0.05**水平下可拒绝原假设，即认为这天包装机工作不正

常

σ^2 已知 **`[h,p,ci]=ztest(x,mu,sigma,alpha,tail)`**

σ^2 未知检验 μ **`[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,现测得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)?

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 $h=0$, $p=0.2570$, $ci=[198.2321 \quad \text{Inf}]$

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)

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

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

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

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同方差的正态总体,均值和方差均未知,问建议的新方法能否提高得率?(取检验水平**0.05**)

解 $H_0 : \mu_1 - \mu_2 = 0 \quad \Leftrightarrow \quad H_1 : \mu_1 - \mu_2 < 0$

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 $h=1$, $p=2.2126 \times 10^{-4}$, $ci = (-Inf \quad -1.9000]$

表明在**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

H_0 : 总体 x 的分布为 $F(x) \longleftrightarrow H_1$: 总体 x 的分布不是 $F(x)$

1. 卡方检验法

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - n \hat{p}_i)^2}{n \hat{p}_i} \underset{H_0 \text{ 成立时}}{\overset{\text{近似}}{\sim}} \chi^2(k - r - 1)$$

k 是分组数(分点 $x_i, i = 0, 1, \dots, k$), r 是未知参数个数(用点估计如极大似然估计等代替), 要求样本容量 n 不少于**50**, 每个 $n p_i$ 最好在**5**以上, 若某些组中落入的数据过少, 可以适当并组。

$$\hat{p}_i = F(x_i) - F(x_{i-1}) \quad \chi^2 > \chi_{1-\alpha}^2(k - r - 1) \text{ 时拒绝 } H_0$$

例6 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>=154))] %各区间上出现的频数
```

```
mu=mean(x),sigma=std(x)    %均值和标准差
fendian=[135,138,142,146,150,154] %区间的分点
p0=normcdf(fendian,mu,sigma)%分点处分布函数值
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**,

$$\chi_{0.1}^2(7-2-1) = \chi_{0.1}^2(4) = 7.7794$$

故在水平**0.1**下接受**H₀**，即认为数据来自正态分布总体

例7 一道工序用自动化车床连续加工某种零件，由于刀具损坏等会出现故障.故障是完全随机的，并假定生产任一零件时出现故障机会均相同.工作人员是通过检查零件来确定工序是否出现故障的.现积累有**100**次故障纪录，故障出现时该刀具完成的零件数如下,试观察该刀具出现故障时完成的零件数属于哪种分布？

459	362	624	542	509	584	433	748	815	505
612	452	434	982	640	742	565	706	593	680
926	653	164	487	734	608	428	1153	593	844
527	552	513	781	474	388	824	538	862	659
775	859	755	49	697	515	628	954	771	609
402	960	885	610	292	837	473	677	358	638
699	634	555	570	84	416	606	1062	484	120
447	654	564	339	280	246	687	539	790	581
621	724	531	512	577	496	468	499	544	645
764	558	378	765	666	763	217	715	310	851

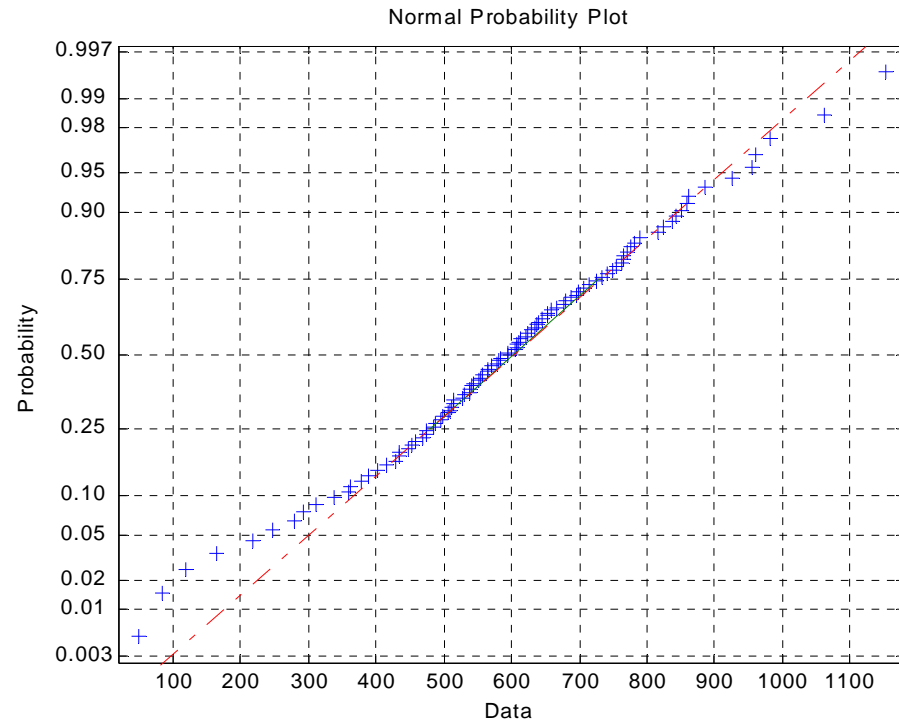
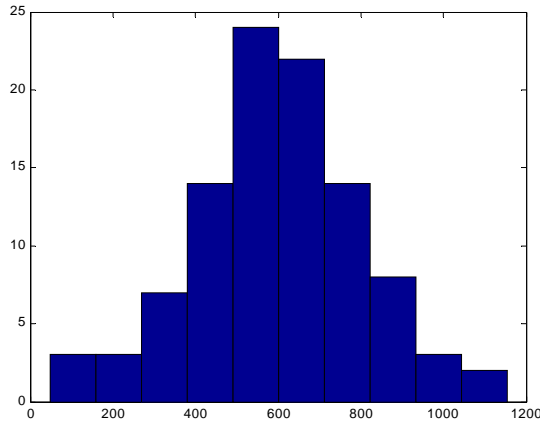
clc % 以下数据输入

```
x=[459 362 624 542 509 584 433 748 815 505 612  
452 434 982 640 742 565 706 593 680 926 653 164  
487 734 608 428 1153 593 844 527 552 513 781 474  
388 824 538 862 659 775 859 755 49 697 515 628  
954 771 609 402 960 885 610 292 837 473 677 358  
638 699 634 555 570 84 416 606 1062 484 120 447  
654 564 339 280 246 687 539 790 581 621 724 531  
512 577 496 468 499 544 645 764 558 378 765 666  
763 217 715 310 851];
```

hist(x,10) % 作频数直方图

normplot(x) %分布正态性检验，刀具寿命服从正态分布

[muhat,sigmahat,muci,sigmaci] = normfit(x) %参数估计



估计出该刀具的

均值 $\mu_{\text{hat}} = 594$ ，方差 $\sigma_{\text{hat}} = 204.1301$

均值0.95的置信区间为 [553.4962, 634.5038]

方差0.95的置信区间为 [179.2276, 237.1329]

假设检验:

已知刀具的寿命服从正态分布, 现在方差未知的情况下, 检验其均值 m 是否等于**594**.

计算结果: **$h = 0$, $sig = 1$, $ci = [553.4962, 634.5038]$** .

检验结果:

1. 布尔变量 **$h=0$** ,

表示不拒绝零假设. 说明提出的假设寿命均值**594**合理.

2. **95%**的置信区间为 **$[553.5, 634.5]$** ,

它完全包括**594**, 且精度很高.

3. **sig -值为1**, 远超过**0.5**, 不能拒绝零假设

2. 偏度、峰度检验（Jarque-Bera检验）

基于数据样本的偏度和峰度，评价给定数据服从未知均值和方差正态分布的假设是否成立。函数：**jbtest**

调用格式为： **$[h,p,JBSTAT,CV]=jbtest(x,alpha)$**

以**alpha** (默认**0.05**)显著水平对数据**x**进行**Jarque-Bera**检验

返回值：**h=0** 接受**x**服从正态分布的假设；**h=1** 拒绝该假设

检验值**p**，检验统计量值**JBSTAT**和临界值**CV**

对上例： **$[h,p,JBSTAT,CV]=jbtest(x)$** 得结果：

$h = 0$, $p = 0.6913$, $JBSTAT = 0.7384$, $CV = 5.9915$

刀具寿命近似服从正态分布

注：**Jarque-Bera**检验不能用于小样本检验

3. Wilcoxon秩和检验

命令格式: **[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8 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 $H_0: \mu_A = \mu_B \leftrightarrow H_1: \mu_A \neq \mu_B$

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=0.8282**, **h=0**, 表明两样本总体均值相等的概率为**0.8282**, 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.

4. 中位数检验

(1) 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**。

例9 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，试验是在同一个平炉上进行的。每炼一炉钢，除操作方法外其它条件都尽可能做到相同。先用标准方法炼一炉，然后用建议的新方法炼一炉，以后交替进行，各炼**10**炉，其钢的得率分别为：

标准方法 **78.1 72.4 76.2 74.3 77.4 78.4 76.0 75.5 76.7 77.3**

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，均值和方差都未知。问建议的新操作方法是否能提高钢的得率？

解 Matlab实现如下：

```
x=[78.1 72.4 76.2 74.3 77.4 78.4 76.0 75.5 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];
```

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

结果： $p = 0.0020$ $h = 1$ 有显著差异，但无法判断优劣。

若用两样本t检验： $h = ttest2(x,y,0.05,-1)$ 可得

$h = 1$ ，知新方法得钢率比标准方法高

(2) 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。

练习1

下面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 $\alpha = \mathbf{0.05}$ ）？

练习2

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
斯诺特格拉斯	0.209	0.205	0.196	0.210	0.202
马克·吐温	0.229	0.235	0.217		
斯诺特格拉斯	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 $\alpha = 0.05$ ）？

第3章 方差分析

方差分析：通过观测数据对因素的影响大小作出合理推断。

方差分析种类：

- ◆ 单因素方差分析
- ◆ 两因素方差分析
 - ▲ 无交互作用的两因素方差分析
 - ▲ 有交互作用的两因素方差分析
- ◆ 三因素方差分析

一、单因素方差分析

方差分析的目的是在众多因素中找出有显著影响的因素，为此需要做试验，试验中可以变化的、影响试验指标的因素称为因素，用大写字母 A 、 B 、 C 、.....表示，因素在试验中所取的不同状态称为水平。

因素 A 的 r 个不同水平用 A_1, \dots, A_r 表示。

方差分析是检验同方差的若干正态总体均值是否相等的一种统计分析方法。

设因素A有 r 个不同水平 A_1, \dots, A_r , 在 A_i 下试验结果 $X_i \sim N(\mu_i, \sigma^2)$, $i=1, \dots, r$ 。在 A_i 下做 $n_i (\geq 2)$ 次试验, 相当于从总体 X_i 中抽取了一组样本 X_{i1}, \dots, X_{in_i} , 他们相互独立, 故方差分析模型为:

$$\begin{cases} X_{ij} \sim N(\mu_i, \sigma^2) \\ X_{ij} \text{ 相互独立} & , i=1, 2, \dots, r \quad ; j=1, 2, \dots, n_i \\ \mu_i, \sigma^2 \text{ 未知} \end{cases}$$

检验问题: $H_0: \mu_1 = \dots = \mu_r \leftrightarrow H_1: \mu_1, \dots, \mu_r$ 不全相同

若拒绝 H_0 , 则表示因素A显著, 否则为不显著。

单因素方差分析的试验指标数据可列成下表形式

因素水平	总体	样本观测数据
A_1	$X_1 \sim N(\mu + \alpha_1, \sigma^2)$	$X_{11}, X_{12}, \dots, X_{1n_1}$
A_2	$X_2 \sim N(\mu + \alpha_2, \sigma^2)$	$X_{21}, X_{22}, \dots, X_{2n_2}$
\vdots	\vdots	\vdots
A_r	$X_r \sim N(\mu + \alpha_r, \sigma^2)$	$X_{r1}, X_{r2}, \dots, X_{rn_r}$

平方和分解公式

$$S_T = S_e + S_A$$

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

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

$$S_A = \sum_{i=1}^r n_i \cdot (\bar{X}_i - \bar{X})^2$$
 因素偏差平方和

检验统计量：
$$F = \frac{S_A / (r-1)}{S_e / (n-r)}$$

拒绝域：
$$W = \{ (x_1, \dots, x_n) : F > F_\alpha(r-1, n-r) \}$$

若 $F > F_\alpha(r-1, n-r)$, 认为因素取不同水平对试验指标的影响显著。

具体来说

$F > F_{0.01}(r-1, n-r)$ 认为因素的影响高度显著，用**表示；

$F_{0.05} < F \leq F_{0.01}(r-1, n-r)$ 认为因素的影响显著，用*表示；

$F_{0.1} < F \leq F_{0.05}(r-1, n-r)$ 认为因素有一定影响，用(*)表示；

$F \leq F_{0.1}(r-1, n-r)$ 认为因素的影响不显著。

单因素方差分析表

来源	平方和	自由度	均方和	F 比	临界值	显著性
因素A	S_A	$r-1$	$S_A / (r-1)$	$F = \frac{S_A / (r-1)}{S_e / (n-r)}$	F_α	
误差e	S_e	$n-r$	$S_e / (n-r)$			
总和	S_T	$n-1$				

具体计算 S_T 、 S_A 和 S_e 时可用变形的公式：

$$n = \sum_{i=1}^r n_i$$

$$S_T = \sum_{i=1}^r \sum_{j=1}^{n_i} X_{ij}^2 - n (\bar{X})^2$$

$$S_A = \sum_{i=1}^r n_i (\bar{X}_i)^2 - n \cdot (\bar{X})^2$$

$$S_e = S_T - S_A$$

未知参数的点估计

(1) $\hat{\mu}_i = \bar{X}_i$ 、 $\hat{\mu} = \bar{X}$ 、 $\hat{\alpha}_i = \bar{X}_i - \bar{X}$ 分别是
 μ_i 、 μ 、 α_i 的无偏估计, $i = 1, \dots, r$

(2) $\hat{\sigma}^2 = \frac{S_e}{n-r}$ 是 σ^2 的无偏估计

未知参数的区间估计

(1) μ_i 的置信度为 $1-\alpha$ 置信区间:

$$\left[\bar{X}_i \pm t_{\alpha/2}(n-r) \sqrt{\frac{S_e}{n_i(n-r)}} \right], i=1,2,\dots,r$$

(2) σ^2 的置信度为 $1-\alpha$ 置信区间:

$$\left[\frac{S_e}{\chi^2_{\alpha/2}(n-r)}, \frac{S_e}{\chi^2_{1-\alpha/2}(n-r)} \right]$$

例1 某厂家为考察某种家电的广告内容对其销售量的影响，在其他条件尽量不变的情况下，设计了三种不同内容的广告：**广告 A_1 强调安装方便性；**
广告 A_2 强调能耗经济性； **广告 A_3 强调低噪性。**在广告被广泛宣传后，按寄回的广告上的订购数计算，一年四个季度的销售量见下表：

广告 季度	A_1	A_2	A_3
1	163	184	206
2	176	198	191
3	170	179	218
4	185	190	224

问:哪种广告引起的销售量最多?

消费者最看重的是此种家电的哪种性能?

解 显然 $r=3$, $n_1=n_2=n_3=n_4=4$,

$$n = n_1 + n_2 + n_3 + n_4 = 12$$

方差分析表

来源	平方和	自由度	均方和	F 比	临界值
因素A	2668.17	2	1334.09	10.93	$F_{0.05}(2,9)=$ 4.16 < $F=10.93$
误差e	1098.4	9	122.06		
总和	3766.67	11			

即认为广告内容的不同对销售量的影响是很大的。

Matlab统计工具箱中单因素方差分析的命令是**anova**

各组数据个数相等(均衡数据)时用法: **p=anova(x)**

返回值**p**是一个概率, 当 **$p > \alpha$** 时接受**H0**,

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

各组数据个数不相等时用法: **p=anova1(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 **i** (**$i=1, \dots, r$**))

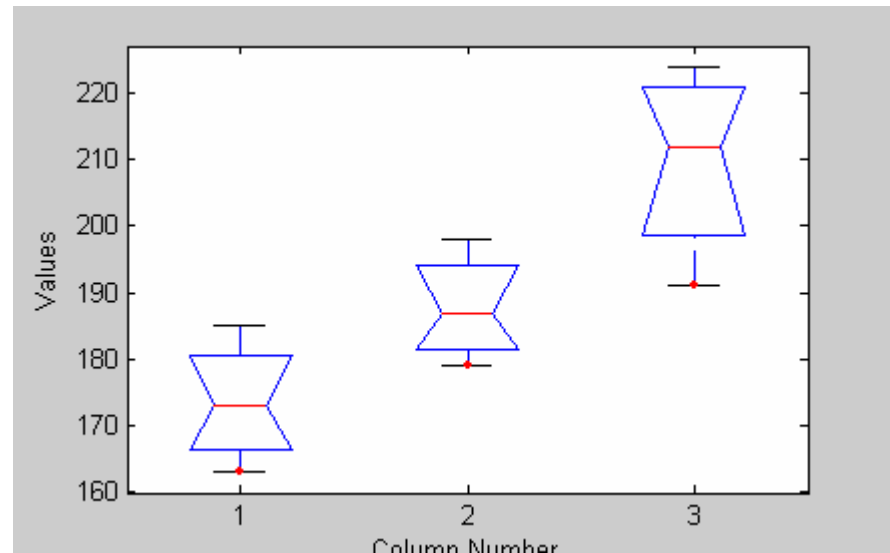
编写程序如下

```
x=[ 163    184    206  
    176    198    191  
    170    179    218  
    185    190    224 ];
```

```
p=anova1(x)
```

运行结果

ANOVA Table					
Source	SS	df	MS	F	Prob>F
Columns	2668.17	2	1334.08	10.93	0.0039
Error	1098.5	9	122.06		
Total	3766.67	11			



求得 $p=0.0039<0.05$ ，故拒绝 H_0 ，即认为广告内容的不同对销售量的影响是很大的。

对不同广告引起的销售量的估计： $\hat{\mu}_{A_i} = \overline{X}_i$

$$\hat{\mu}_{A_1} = \frac{694}{4} = 173.5, \hat{\mu}_{A_2} = \frac{754}{4} = 187.75, \hat{\mu}_{A_3} = \frac{839}{4} = 209.75$$

95%的置信区间： $\left[\overline{X}_i \pm t_{\alpha/2}(n-r) \sqrt{\frac{S_e}{n_i(n-r)}} \right], 1 \leq i \leq r$

$$\hat{\mu}_{A_1} : [161, 186], \hat{\mu}_{A_2} : [175, 200], \hat{\mu}_{A_3} : [197, 222]$$

由此可见，广告A₃引起的销售量最多，今后应多宣传噪音低的优良性，同时进一步进行工艺的改革以降低噪音。

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 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 ];
```

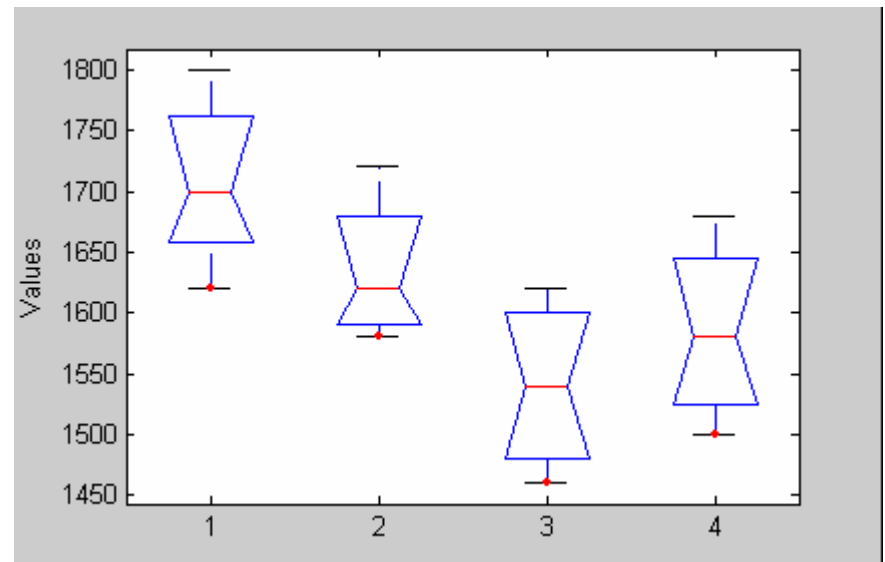
```
x=[x(1:4),x(16),x(5:8),x(9:11),x(12:15)];
```

```
g=[ones(1,5),2*ones(1,4),3*ones(1,3),4*ones(1,4)];
```

```
p=anova1(x,g)
```

Source	SS	df	MS	F	Prob>F
Groups	62820	3	20940	4.06	0.0331
Error	61880	12	5156.67		
Total	124700	15			

求得 $0.01 < p = 0.0331 < 0.05$,
 所以几种工艺制成的灯泡寿命在显著水平 $\alpha = 0.01$ 下无显著差异，但在显著水平 $\alpha = 0.05$ 下有显著差异



二、两因素方差分析

因素A 取 r 个不同水平 A_1, \dots, A_r ;

因素B取 s 个不同水平 B_1, \dots, B_s ;

(A_i, B_j) 组合下的试验结果 $X_{ij} \sim i.i.d. N(u_{ij}, \sigma^2)$

两因素方差分析模型

$$\left\{ \begin{array}{l} \mu = \frac{1}{r \cdot s} \sum_{i=1}^r \sum_{j=1}^s \mu_{ij} \\ \mu_{i \cdot} = \frac{1}{s} \sum_{j=1}^s \mu_{ij} \quad , \quad i = 1, \dots, r \\ \mu_{\cdot j} = \frac{1}{r} \sum_{i=1}^r \mu_{ij} \quad , \quad j = 1, \dots, s \end{array} \right. \quad \left\{ \begin{array}{l} \alpha_i = \mu_{i \cdot} - \mu \quad , \quad i = 1, \dots, r \\ \beta_j = \mu_{\cdot j} - \mu \quad , \quad j = 1, \dots, s \end{array} \right.$$

(一) 无交互作用的两因素方差分析

1. 数学模型 $\mu_{ij} = \mu + \alpha_i + \beta_j$

只需对 (A_i, B_j) 的每个组合做一次试验,

试验结果 X_{ij}

因素A \ 因素B	B_1, B_2, \dots, B_s	平均值 $\bar{X}_{i\cdot}$
A_1	$X_{11}, X_{12}, \dots, X_{1s}$	$\bar{X}_{1\cdot}$
A_2	$X_{21}, X_{22}, \dots, X_{2s}$	$\bar{X}_{2\cdot}$
...
A_r	$X_{r1}, X_{r2}, \dots, X_{rs}$	$\bar{X}_{r\cdot}$
平均值 $\bar{X}_{\cdot j}$	$\bar{X}_{\cdot 1}, \bar{X}_{\cdot 2}, \dots, \bar{X}_{\cdot s}$	\bar{X}

对此模型的假设检验有两个

$$\begin{cases} H_{0A} : \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0 \leftrightarrow H_{1A} : \alpha_1, \alpha_2, \cdots, \alpha_r \text{ 不全为零} \\ H_{0B} : \beta_1 = \beta_2 = \cdots = \beta_s = 0 \leftrightarrow H_{1B} : \beta_1, \beta_2, \cdots, \beta_s \text{ 不全为零} \end{cases}$$

$$\bar{x}_{i\bullet} = \frac{1}{s} \sum_{j=1}^s x_{ij} \quad , \quad \bar{x}_{\bullet j} = \frac{1}{r} \sum_{i=1}^r x_{ij}$$

2. 平方和分解公式: $S_T = S_e + S_A + S_B$

$$\left\{ \begin{array}{l} S_T = \sum_{i=1}^r \sum_{j=1}^s x_{ij}^2 - \frac{1}{n} \left(\sum_{i=1}^r \sum_{j=1}^s x_{ij} \right)^2, \quad n = r \cdot s \\ S_A = s \cdot \sum_{i=1}^r (\bar{x}_{i\bullet} - \bar{x})^2 = \frac{1}{s} \sum_{i=1}^r \left(\sum_{j=1}^s x_{ij} \right)^2 - \frac{1}{n} \left(\sum_{i=1}^r \sum_{j=1}^s x_{ij} \right)^2 \\ S_B = r \cdot \sum_{j=1}^s (\bar{x}_{\bullet j} - \bar{x})^2 = \frac{1}{r} \sum_{j=1}^s \left(\sum_{i=1}^r x_{ij} \right)^2 - \frac{1}{n} \left(\sum_{i=1}^r \sum_{j=1}^s x_{ij} \right)^2 \\ S_e = \sum_{i=1}^r \sum_{j=1}^s (x_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{x})^2 = S_T - S_A - S_B \end{array} \right.$$

3. 检验统计量

$$\begin{cases} F_A = \frac{S_A / (r-1)}{S_e / ((r-1)(s-1))} \sim F(r-1, (r-1)(s-1)) \\ F_B = \frac{S_B / (s-1)}{S_e / ((r-1)(s-1))} \sim F(s-1, (r-1)(s-1)) \end{cases}$$

4. 给定检验水平 α 时, H_{0A} , H_{0B} 的拒绝域

$$\begin{cases} W_A = \{ (x_1, \dots, x_n) : F_A > F_\alpha(r-1, (r-1)(s-1)) \} \\ W_B = \{ (x_1, \dots, x_n) : F_B > F_\alpha(s-1, (r-1)(s-1)) \} \end{cases}$$

5. 无交互作用的两因素方差分析表

来源	平方和	自由度	均方和	<i>F</i> 比	临界值	显著性
因素 <i>A</i>	S_A	$r-1$	$MSA = \frac{S_A}{r-1}$	$F_A = \frac{MSA}{MSE}$	$F_{\alpha}(r-1, n_e)$	
因素 <i>B</i>	S_B	$s-1$	$MSB = \frac{S_B}{s-1}$	$F_B = \frac{MSB}{MSE}$	$F_{\alpha}(s-1, n_e)$	
误差 <i>e</i>	S_e	n_e	$MSE = \frac{S_e}{n_e}$			
总和	S_T	$n-1$				

表中 $n = rs, n_e = (r-1)(s-1)$

$F_A > F_{\alpha}(r-1, n_e)$ 时拒绝*H*_{0*A*} ; $F_B > F_{\alpha}(s-1, n_e)$ 时拒绝*H*_{0*B*}

例3 设四名工人分别操作机床甲、乙、丙各一天，生产同种产品，其日产量统计如下（单位：件）

问工人的不同和机床的不同在日产量上是否有显著差异？（假定四名工人对这三台机器的熟悉情况是一样的）设($\alpha=0.01$)

工人 机床	张三	李四	王五	赵六
甲	53	47	57	45
乙	56	50	63	52
丙	45	47	54	42

解 把工人看作因素A，它有四个水平，把机床看作因素B，它有3个水平，由题意（假定四名工人对这三台机器的熟悉情况是一样的）知，因素A和B间无交互作用，且

$$r = 4, s = 3, n = r \cdot s = 12;$$

$$\frac{1}{n} \left(\sum_{i=1}^r \sum_{j=1}^s x_{ij} \right) = \frac{1}{12} \times 611^2 = 31110.08$$

$$S_T = \sum_{i=1}^r \sum_{j=1}^s x_{ij}^2 - \frac{1}{n} \left(\sum_{i=1}^r \sum_{j=1}^s x_{ij} \right)^2 = 31515 - 31110.08 = 404.92$$

$$S_A = \frac{1}{s} \sum_{i=1}^r \left(\sum_{j=1}^s x_{ij} \right)^2 - \frac{1}{n} \left(\sum_{i=1}^r \sum_{j=1}^s x_{ij} \right)^2 = \frac{1}{3} \times 94094 - 31110.08 = 239.59$$

$$S_B = \frac{1}{r} \sum_{j=1}^s \left(\sum_{i=1}^r x_{ij} \right)^2 - \frac{1}{n} \left(\sum_{i=1}^r \sum_{j=1}^s x_{ij} \right)^2 = \frac{1}{4} \times 124989 - 31110.08 = 137.17$$

$$S_e = S_T - S_A - S_B = 404.92 - 239.59 - 137.17 = 28.16$$

方差分析表

来源	平方和	自由度	均方和	F 比	临界值	显著性
因子A	239.59	3	79.86	17.03	$F_{0.1}(3,6)=9.78$	**
因子B	137.17	2	68.59	14.62	$F_{0.1}(2,6)=10.9$	**
误差e	28.16	6	4.69			
总和	404.92	11				

结果表明:工人的不同和机床的不同在日产量上有非常显著的差异

(二) 有交互作用的两因素方差分析

1. 数学模型为 $\mu_{ij} \neq \mu + \alpha_i + \beta_j$

对 (A_i, B_j) 的每个组合至少做 $t (\geq 2)$ 次试验, 试验结果 X_{ijk}

$$1 \leq i \leq r, 1 \leq j \leq s, 1 \leq k \leq t$$

称 $r_{ij} = \mu_{ij} - \mu - \alpha_i - \beta_j$ 为因素 A 的第 i 个水平与因素 B

的第 j 个水平的交互效应, 这时 $\mu_{ij} = \mu + \alpha_i + \beta_j + r_{ij}$

有交互作用的方差分析模型:

$$\begin{cases} X_{ijk} = \mu + \alpha_i + \beta_j + r_{ij} + \varepsilon_{ijk}, \varepsilon_{ijk} \sim i.i.d N(0, \sigma^2) \\ \sum_{i=1}^r \alpha_i = \sum_{j=1}^s \beta_j = 0, \sum_{i=1}^r r_{ij} = \sum_{j=1}^s r_{ij} = 0, 1 \leq i \leq r, 1 \leq j \leq s \\ \mu, \alpha_i, \beta_j, r_{ij}, \sigma^2 \text{ 未知} \end{cases}$$

对此模型的假设检验有三个

$$\left\{ \begin{array}{l} H_{0A} : \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0 \quad \leftrightarrow \quad H_{0A} : \alpha_1, \alpha_2, \cdots, \alpha_r \text{不全为零} \\ H_{0B} : \beta_1 = \beta_2 = \cdots = \beta_s = 0 \quad \leftrightarrow \quad H_{0B} : \beta_1, \beta_2, \cdots, \beta_s \text{不全为零} \\ H_{0AB} : r_{ij} = 0, 1 \leq i \leq r, 1 \leq j \leq s \quad \leftrightarrow \quad H_{1AB} : r_{ij} \text{不全为零} \end{array} \right.$$

$$\text{记} \left\{ \begin{array}{l} \bar{X} = \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^t X_{ijk} \quad (n = r \cdot s \cdot t) \\ \bar{X}_{ij\cdot} = \frac{1}{t} \sum_{k=1}^t X_{ijk} \quad (i = 1, 2, \cdots, r ; j = 1, 2, \cdots, s) \\ \bar{X}_{i..} = \frac{1}{st} \sum_{j=1}^s \sum_{k=1}^t X_{ijk} \quad (i = 1, 2, \cdots, r) \\ X_{\cdot j\cdot} = \frac{1}{rt} \sum_{i=1}^r \sum_{k=1}^t X_{ijk} \quad (j = 1, 2, \cdots, s) \end{array} \right.$$

2. 平方和分解公式: $S_T = S_e + S_A + S_B + S_{AB}$

$$\left\{ \begin{array}{l} S_T = \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^t (X_{ijk} - \bar{X})^2 \\ S_A = st \sum_{i=1}^r (\bar{X}_{i..} - \bar{X})^2 \\ S_B = rt \sum_{j=1}^s (\bar{X}_{.j.} - \bar{X})^2 \\ S_{AB} = t \sum_{i=1}^r \sum_{j=1}^s (\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X})^2 \end{array} \right.$$

3. 检验统计量

$$\left\{ \begin{array}{l} F_A = \frac{S_A / (r-1)}{S_e / (r s(t-1))} \sim F(r-1, r s(t-1)) \\ F_B = \frac{S_B / (s-1)}{S_e / (r s(t-1))} \sim F(s-1, r s(t-1)) \\ F_{AB} = \frac{S_{AB} / ((r-1)(s-1))}{S_e / (r s(t-1))} \sim F((r-1)(s-1), r s(t-1)) \end{array} \right.$$

4. 给定检验水平 α 时, H_{0A} , H_{0B} 的拒绝域

$$\left\{ \begin{array}{l} W_A = \{(x_1, \dots, x_n) : F_A > F_\alpha(r-1, r s(t-1))\} \\ W_B = \{(x_1, \dots, x_n) : F_B > F_\alpha(s-1, r s(t-1))\} \\ W_{AB} = \{(x_1, \dots, x_n) : F_{AB} > F_\alpha((r-1)(s-1), r s(t-1))\} \end{array} \right.$$

5. 有交互作用的两因素方差分析表

表中 $n = rst$, $n_{AB} = (r-1)(s-1)$, $n_e = rs(t-1)$

来源	平方和	自由度	均方和	F比	临界值	显著性
因素A	S_A	$r-1$	$MSA = \frac{S_A}{r-1}$	$F_A = \frac{MSA}{MSE}$	$F_\alpha(r-1, n_e)$	
因素B	S_B	$s-1$	$MSB = \frac{S_B}{s-1}$	$F_B = \frac{MSB}{MSE}$	$F_\alpha(s-1, n_e)$	
$A \times B$	S_{AB}	n_{AB}	$MSAB = \frac{S_{AB}}{n_{AB}}$	$F_{AB} = \frac{MSAB}{MSE}$	$F_\alpha(n_{AB}, n_e)$	
误差e	S_e	n_e	$MSE = \frac{S_e}{n_e}$			
总和	S_T	$n-1$				

例4 在某化工厂生产中为了提高收率，选了三种不同浓度，四种不同温度做试验。在同一浓度和温度组合下各做两次试验，其收率数据如下计算表所列（数据均已减**75**）。试在显著性水平($\alpha = 0.01$)下检验不同浓度、不同温度以及它们之间的交互作用对收率有无显著影响。

解：先计算 $X_{ij\bullet}$, $X_{i\bullet\bullet}$, $X_{\bullet j\bullet}$, $\sum_i X_{i\bullet\bullet}^2$, $\sum_j X_{\bullet j\bullet}^2$, $\sum_i \sum_j X_{ij\bullet}^2$, $\sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^t X_{ijk}^2$, 得计算表如下, 且 $r=3, s=4, t=2, n=rst=24$

浓度 A	温度 B				$X_{i\bullet\bullet}$	$X_{i\bullet\bullet}^2$
	B_1	B_2	B_3	B_4		
A_1	14, 10 (24)	11, 11 (22)	13, 9 (22)	10, 12 (22)	90	8100
A_2	9, 7 (16)	10, 8 (18)	7, 11 (18)	6, 10 (16)	68	4624
A_3	5, 11 (16)	13, 14 (27)	12, 13 (25)	14, 10 (24)	92	8464
$X_{\bullet j\bullet}$	56	67	65	62	$\sum_{ijk} X_{ijk} = 250$	$\sum_j X_{i\bullet\bullet}^2$ = 21188
$X_{\bullet j\bullet}^2$	3136	4489	4225	3844	$\sum_j X_{\bullet j\bullet}^2 = 15694$	

$$\sum_i \sum_j \sum_k X_{ijk}^2 = 2752 \qquad \frac{1}{24} (\sum_i \sum_j \sum_k X_{ijk})^2 = 2604.1667$$

$$S_A = \frac{1}{8} \times 21188 - 2604.1667 = 44.3333$$

$$S_B = \frac{1}{6} \times 15694 - 2604.1667 = 11.5000$$

$$S_{AB} = \frac{1}{2} \times 5374 - 2604.1667 - 44.3333 - 11.5000 = 27.0000$$

$$S_e = S_T - S_A - S_B - S_{AB} = 65.0000$$

得方差分析表如下

方差分析表

来源	平方和	自由度	均方和	<i>F</i> 比	临界值	显著性
因素 <i>A</i>	44.333	2	22.167	4.09	3.89	*
因素 <i>B</i>	11.500	3	3.833	0.708	3.49	
交互 <i>A</i> × <i>B</i>	27.000	6	4.5	0.831	3.00	
误差 <i>e</i>	65.000	12	5.417	/	/	
总和	147.833	23	/	/	/	

表明:只有因素*A*(浓度)作用显著，温度和交互作用不显著。
故应控制好浓度。

例5 一个超市将一种商品采用**3**种不同的包装，放在**3**个不同的货架上作销售试验，希望检验不同的包装、不同货架对销售量是否有显著影响，交互作用显著，随机地抽取**3**天的销售量作样本，取检验水平 $\alpha=0.05$ ，其观测结果如下表：

因素A \ 因素 B	包装1	包装2	包装3
货架1	5 6 4	6 8 7	4 3 5
货架2	7 8 8	5 5 5	3 6 4
货架3	3 2 4	6 6 5	8 9 6

方差分析表

来源	平方和	自由度	均方和	F 比	临界值	显著性
因子A	0.96	2	0.48	0.45	3.55	
因子B	3.18	2	1.59	1.49	3.55	
交互 $A \times B$	61.27	4	15.32	14.31	2.93	*
误差 e	19.33	18	1.07	/	/	
总和	84.74	26	/	/	/	

结果表明:货物的包装及放的货架这两个因素对销售量的影响不显著，但两者的交互作用对销售量的影响显著.

Matlab实现: 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”)有不只一个的观测值，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

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

```

clc,clear
x0=[5 6 4 6 8 7 4 3 5
     7 8 8 5 5 5 3 6 4
     3 2 4 6 6 5 8 9 6];
x1=x0(:,1:3:7);x2=x0(:,2:3:8);x3=x0(:,3:3:9);
for i=1:3
    x(3*i-2,:)=x1(i,:);
    x(3*i-1,:)=x2(i,:);
    x(3*i,:)=x3(i,:);
end
p=anova2(x,3)

```

ANOVA Table					
Source	SS	df	MS	F	Prob>F
Columns	2.2963	2	1.1481	1.11	0.352
Rows	0.5185	2	0.2593	0.25	0.7815
Interaction	63.2593	4	15.8148	15.25	0
Error	18.6667	18	1.037		
Total	84.7407	26			

求得 **p=0.352 0.7815 0**，表明货物的包装及放的货架这两个因素试验均值相等的概率不是小概率，故可接受均值相等假设。但两者交互作用显著的

三、三因素方差分析

因素 A 取 r 个不同水平 A_1, \dots, A_r ;

因素 B 取 s 个不同水平 B_1, \dots, B_s ;

因素 C 取 t 个不同水平 C_1, \dots, C_t ;

(A_i, B_j, C_k) 组合下重复 q 次试验,

试验结果 $X_{ijkl} \sim i.i.d. N(u_{ijk}, \sigma^2)$

平方和分解公式:

$$S_T = S_e + S_A + S_B + S_C + S_{AB} + S_{AC} + S_{BC} + S_{ABC}$$

$$\left\{ \begin{array}{l} S_T = \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^t \sum_{l=1}^q (X_{ijkl} - \bar{X})^2 \\ S_A = s t q \sum_{i=1}^r (\bar{X}_{i...} - \bar{X})^2 \\ S_B = r t q \sum_{j=1}^s (\bar{X}_{\cdot j \cdot \cdot} - \bar{X})^2 \\ S_C = r s q \sum_{k=1}^t (\bar{X}_{\cdot \cdot k \cdot} - \bar{X})^2 \end{array} \right.$$

$$\left\{ \begin{array}{l} S_{AB} = tq \sum_{i=1}^r \sum_{j=1}^s (\overline{X}_{ij..} - \overline{X}_{i...} - \overline{X}_{.j..} + \overline{X})^2 \\ S_{AC} = sq \sum_{i=1}^r \sum_{k=1}^t (\overline{X}_{i.k.} - \overline{X}_{i...} - \overline{X}_{..k.} + \overline{X})^2 \\ S_{BC} = rq \sum_{j=1}^s \sum_{k=1}^t (\overline{X}_{.jk.} - \overline{X}_{.j..} - \overline{X}_{..k.} + \overline{X})^2 \end{array} \right.$$

$$\left\{ \begin{array}{l} S_{ABC} = q \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^t \\ \quad (\overline{X}_{ijk\cdot} - \overline{X}_{ij\cdot\cdot} - \overline{X}_{i\cdot k\cdot} - \overline{X}_{\cdot jk\cdot} \\ \quad + \overline{X}_{i\cdot\cdot\cdot} + \overline{X}_{\cdot j\cdot\cdot} + \overline{X}_{\cdot\cdot k\cdot} + \overline{X})^2 \\ \\ S_e = \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^t \sum_{l=1}^q (X_{ijkl} - \overline{X}_{ijk\cdot})^2 \end{array} \right.$$

$$n = rstq - 1$$

$$n_{AB} = (r - 1)(s - 1)$$

$$n_{AC} = (r - 1)(t - 1)$$

$$n_{BC} = (s - 1)(t - 1)$$

$$n_{ABC} = (r - 1)(s - 1)(t - 1)$$

$$n_e = rst(q - 1)$$

有交互作用的三因素方差分析表

来源	平方和	自由度	均方和	F 比	临界值
因素A	S_A	$r-1$	$MSA = \frac{S_A}{r-1}$	$F_A = \frac{MSA}{MSE}$	$F_\alpha(r-1, n_e)$
因素B	S_B	$s-1$	$MSB = \frac{S_B}{s-1}$	$F_B = \frac{MSB}{MSE}$	$F_\alpha(s-1, n_e)$
因素C	S_C	$t-1$	$MSC = \frac{S_C}{t-1}$	$F_C = \frac{MSC}{MSE}$	$F_\alpha(t-1, n_e)$
$A \times B$	S_{AB}	n_{AB}	$MSAB = \frac{S_{AB}}{n_{AB}}$	$F_{AB} = \frac{MSAB}{MSE}$	$F_\alpha(n_{AB}, n_e)$
$A \times C$	S_{AC}	n_{AC}	$MSAC = \frac{S_{AC}}{n_{AC}}$	$F_{AC} = \frac{MSAC}{MSE}$	$F_\alpha(n_{AC}, n_e)$
$B \times C$	S_{BC}	n_{BC}	$MSBC = \frac{S_{BC}}{n_{BC}}$	$F_{BC} = \frac{MSBC}{MSE}$	$F_\alpha(n_{BC}, n_e)$
$A \times B \times C$	S_{ABC}	n_{ABC}	$MSABC = \frac{S_{ABC}}{n_{ABC}}$	$F_{ABC} = \frac{MSABC}{MSE}$	$F_\alpha(n_{ABC}, n_e)$
误差e	S_e	n_e	$MSE = \frac{S_e}{n_e}$		
总和	S_T	$n-1$			

例6 某集团为研究销售点所在地理位置、销售点处的广告和销售点的装潢这三个因素对商品销售量的印象程度，选了三个位置（如市中心黄金地段、非中心地段、城乡结合部），两种广告形式，两种装潢档次再四个城市进行了搭配试验。

用 A_1, A_2, A_3 表示三种位置， B_1, B_2 代表两种广告形式， C_1, C_2 表示装潢档次，它们分别称为 A 、 B 、 C 三种因素。每个组合在四个城市的销售量的统计数据如下：

城市号 水平组合	1	2	3	4
$A_1 B_1 C_1$	955	967	960	980
$A_1 B_1 C_2$	927	949	950	930
$A_1 B_2 C_1$	905	930	910	920
$A_1 B_2 C_2$	855	860	880	875
$A_2 B_1 C_1$	880	890	895	900
$A_2 B_1 C_2$	860	840	850	830
$A_2 B_2 C_1$	870	865	850	860
$A_2 B_2 C_2$	830	850	840	830
$A_3 B_1 C_1$	875	888	900	892
$A_3 B_1 C_2$	870	850	847	965
$A_3 B_2 C_1$	870	863	845	855
$A_3 B_2 C_2$	821	842	832	848

问：哪种组合对销售量的影响显著，即何种组合对增加销售量效果最好，位置、广告、装潢这三个因素中哪一个对销售量影响最大？

销售量三因素试验方差分析表(取检验水平0.05)

来源	平方和	自由度	均方和	F 比	临界值
因素A	41596.74	2	20798.37	175.04	3.26
因素B	14666.42	1	14666.42	123.43	4.11
因素C	13306.68	1	13306.68	111.99	4.11
$A \times B$	4010.72	2	2005.36	16.88	3.26
$A \times C$	275.7	2	137.85	1.16	3.26
$B \times C$	18.12	1	18.12	0.15	4.11
$A \times B \times C$	1170.79	2	585.39	4.93	3.26
误差 e	4277.5	36	118.82	/	
总和	79322.67	47	/	/	

查临界值表得到结论：

$$F_A > F_{0.05}(2,36)=3.26$$

$$F_B > F_{0.05}(1,36)=4.11$$

$$F_C > F_{0.05}(1,36) =4.11$$

说明销售点的位置对销售量影响最显著；

广告形式对销售量影响较显著；

装潢对销售量影响也显著；

$F_{AB} > F_{0.05}(2,36)=3.26$ 说明销售点的位置与广告形式的组合作用对销售量的交互影响显著；
 $F_{AC} < F_{0.05}(2,36)=3.26$
 $F_{BC} < F_{0.05}(1,36) =4.11$

而销售点的位置与广告形式、广告形式与装潢档次的组合作用对销售量的交互影响并不显著；

$F_{ABC} > F_{0.05}(4,12)=3.26$ 说明三种因素的组合作用对销售量的交互影响显著。

因此，销售点位置、广告形式和装潢档次的共同作用效果最好，而销售点位置、广告形式和装潢档次的组合效果最差。

练习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。参考**Matlab**程序见下面。

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

```
x=[ 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];
p=anova1(x)
```


练习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。参考**Matlab**程序见下面。

化肥		A1	A2	A3	A4
品种	B1	173, 172, 173	174, 176, 178	177, 179, 176	172, 173, 174
	B2	175, 173, 176	178, 177, 179	174, 175, 173	170, 171, 172
	B3	177, 175, 176	174, 174, 175	174, 173, 174	169, 169, 170

```
clc,clear
```

```
x0=[173 172 173 174 176 178 177 179 176 172 173 174  
    175 173 176 178 177 179 174 175 173 170 171 172  
    177 175 176 174 174 175 174 173 174 169 169 170];
```

```
x1=x0(:,1:3:10);x2=x0(:,2:3:11);x3=x0(:,3:3:12);
```

```
for i=1:3
```

```
    x(3*i-2,:)=x1(i,:);
```

```
    x(3*i-1,:)=x2(i,:);
```

```
    x(3*i,:)=x3(i,:);
```

```
end
```

```
p=anova2(x,3)
```

ANOVA Table					
Source	SS	df	MS	F	Prob>F
Columns	125	3	41.6667	33.33	0
Rows	13.1667	2	6.5833	5.27	0.0127
Interaction	68.8333	6	11.4722	9.18	0
Error	30	24	1.25		
Total	237	35			

求得 **p=0 0.0127 0**,品种、化肥及二者的交互作用在显著水平 $\alpha =0.05$ 下对小麦产量均有显著差异,但在显著水平 $\alpha =0.01$ 下化肥对小麦产量无显著差异

第4章 回归分析

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（一个或多个）自变量之间的一个函数，使这个函数对该组数据拟合得最好。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，要作的工作就是由数据用最小二乘法（不用最小一乘法）计算函数中的待定系数。

简单地说，回归分析就是对拟合问题作的统计分析。

回归分析在一组数据的基础上研究这样几个问题

- 建立因变量 y 与自变量 x_1, \dots, x_m 间的回归模型(经验公式);
- 对回归模型的可信度进行检验;
- 判断每个自变量 x_i 对 y 的影响是否显著;
- 诊断回归模型是否适合这组数据;
- 利用回归模型对 y 进行预报或控制

§ 1 多元线性回归

1.1 模型

多元线性回归模型
$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon \\ \varepsilon \sim N(0, \sigma^2) \quad \text{其中 } \sigma \text{ 未知} \end{cases}$$

n 个独立观测数据 $(y_i; x_{i1}, \cdots, x_{im}), (i = 1, \cdots, n; n > m)$

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

记
$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{pmatrix}, Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}$$

多元线性回归模型可表示为

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

1.2 参数估计

误差平方和 $Q(\beta) = \sum_{i=1}^n \varepsilon_i^2 = (Y - X\beta)^T (Y - X\beta)$

利用最小二乘法 (求使 $Q(\beta)$ 达最小的 β)可求得

最小二乘估计 $\hat{\beta} = (X^T X)^{-1} X^T Y$

得多元线性回归方程 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m$

拟合值 $\hat{Y} = X\hat{\beta}$

残差向量(拟合误差) $e = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = Y - \hat{Y} = \begin{pmatrix} y_1 - \hat{y}_1 \\ \vdots \\ y_n - \hat{y}_n \end{pmatrix}$

残差平方和 (或剩余平方和) $S_{\text{残}} = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$

1.3 统计分析

平方和分解公式 $S = \sum_{i=1}^n (y_i - \bar{y})^2 = S_{\text{残}} + S_{\text{回}}$

且 $\frac{S}{\sigma^2} \sim \chi^2(n-1)$.其中 $S_{\text{残}} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$, $S_{\text{回}} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$

1. $E\hat{\beta} = \beta$

2. $\hat{\beta} \sim N(\beta, \sigma^2(X^T X)^{-1})$

3. $\frac{S_{\text{残}}}{\sigma^2} \sim \chi^2(n-m-1) \Rightarrow \hat{\sigma}^2 = \frac{S_{\text{残}}}{n-m-1}$ 是 σ^2 的无偏估计

4. $\frac{S_{\text{回}}}{\sigma^2} \sim \chi^2(m) \Rightarrow \hat{\sigma}^2 = \frac{S_{\text{回}}}{m}$ 也是 σ^2 的无偏估计

1.4 回归模型的假设检验

检验问题:

$$H_0: \beta_j = 0 (j = 1, \dots, m) \quad \leftrightarrow \quad \beta_j (j = 1, \dots, m) \text{ 不全为 } 0$$

检验统计量:
$$F = \frac{S_{\text{回}} / m}{S_{\text{残}} / (n - m - 1)} \sim F(m, n - m - 1)$$

判断:

当 $F > F_{\alpha}(m, n - m - 1)$ 时拒绝 H_0 即认为回归模型显著

相关系数
$$R^2 = \frac{S_{\text{回}}}{S}, \quad (0 \leq R \leq 1, R \text{ 越大越好})$$

1.5 回归系数的假设检验

检验问题: $H_0^{(j)} : \beta_j = 0 \leftrightarrow H_1^{(j)} : \beta_j \neq 0$, $(j=1, \dots, m)$

检验统计量: $\because \hat{\beta}_j \stackrel{H_0^{(j)} \text{成立时}}{\sim} N(\beta_j, \sigma^2 c_{jj})$

$$\therefore t_j = \frac{\hat{\beta}_j / \sqrt{c_{jj}}}{\sqrt{S_{\text{残}} / (n - m - 1)}} \stackrel{H_0^{(j)} \text{成立时}}{\sim} t(n - m - 1)$$

其中 c_{jj} 是 $(X^T X)^{-1}$ 对角线上第 $j+1$ 个元素

判断: 当 $|t_j| > t_{\frac{\alpha}{2}}(n - m - 1)$ 时拒绝 $H_0^{(j)}$

说明 x_j 的作用显著

1.6 回归系数的区间估计

对置信水平 $1-\alpha$, β_j 的置信区间:

$$\left[\hat{\beta}_j \pm t_{\frac{\alpha}{2}}(n-m-1) s \sqrt{c_{jj}} \right]$$

$$\text{其中 } s = \sqrt{\frac{S_{\text{残}}}{n-m-1}}$$

1.7 利用回归模型进行预测

对给定的 $x_0 = (x_{01}, \dots, x_{0m})$

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \dots + \hat{\beta}_m x_{0m}$$

预测区间 $\left[\hat{y}_0 - u_{\frac{\alpha}{2}} s, \hat{y}_0 + u_{\frac{\alpha}{2}} s \right]$

对 y_0 的区间估计方法可用于给出已知数据残差

$$e_i = y_i - \hat{y}_i, (i = 1, \dots, n) \quad \text{的置信区间,}$$

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。

1.8 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：**b=regress(Y,X)**

这里 Y, X 为数组矩阵， b 为回归系数估计值 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_m$

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

这里 Y, X 同上， α 为显著性水平(缺省时设为**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 ，第二个是 F ，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.

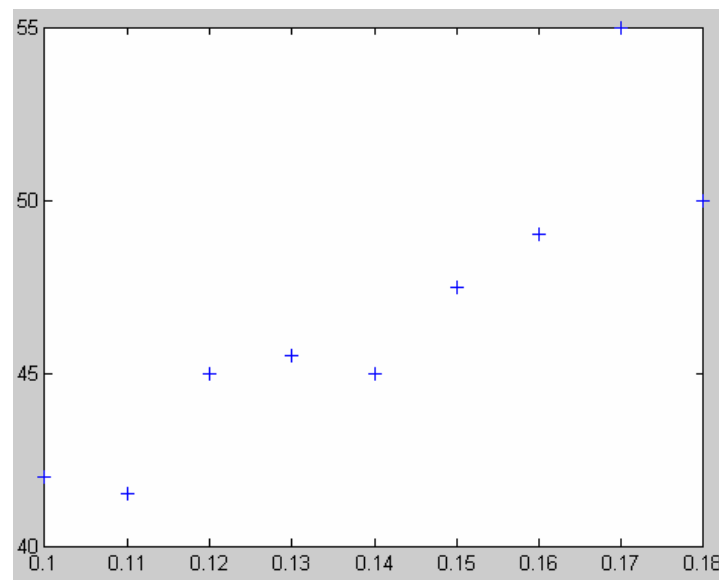
例1 合金的强度与其中的碳含量有比较密切的关系，今从生产中收集了一批数据如下表

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 $y=b_0+b_1x$

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 $b = 27.4722 \quad 137.5000$

$bint = 18.6851 \quad 36.2594 \quad \leftrightarrow \quad \hat{b}_0$ 的置信区间

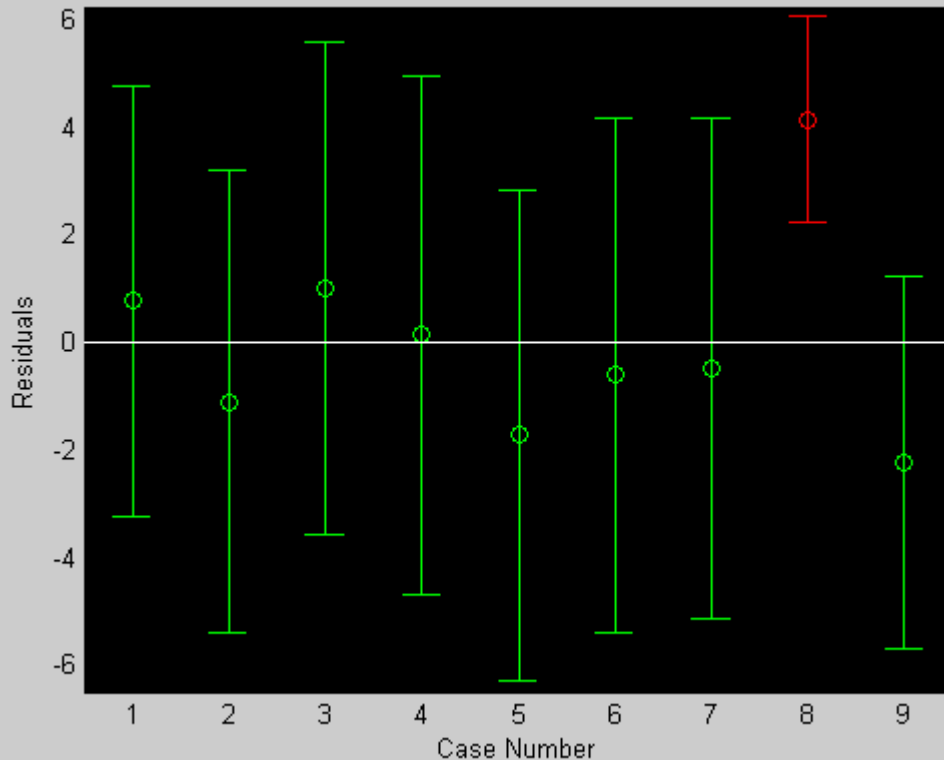
$75.7755 \quad 199.2245 \quad \leftrightarrow \quad \hat{b}_1$ 的置信区间

$stats = 0.7985 \quad 27.7469 \quad 0.0012$

$\hat{b}_0 = 27.4722, \hat{b}_1 = 137.5000, R^2 = 0.7985, F = 27.7469, p = 0.0012$

可知模型成立

Residual Case Order Plot



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

$b = 30.7820 \quad 109.3985$

$bint = 26.2805 \quad 35.2834$

$76.9014 \quad 141.8955$

$stats = 0.9188 \quad 67.8534 \quad 0.0002$

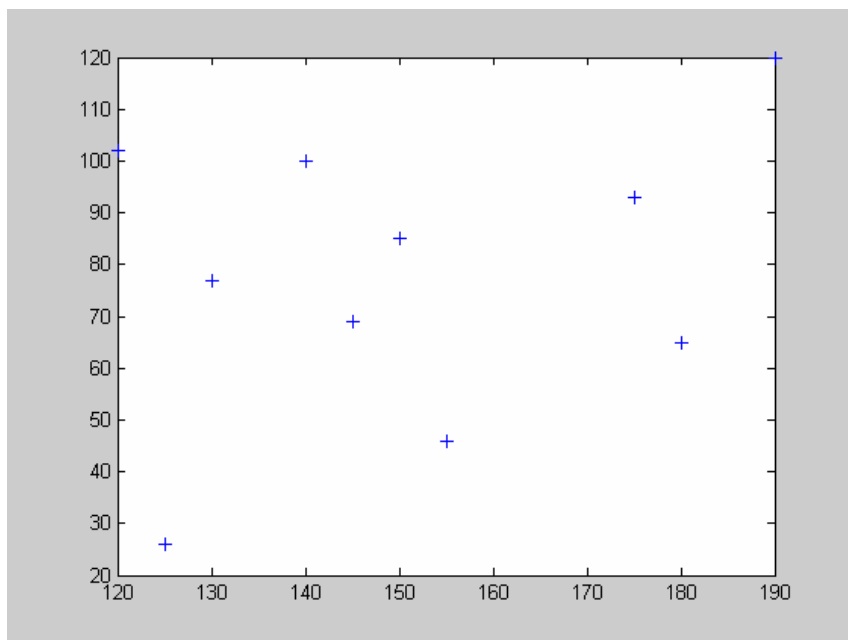
应该用修改后的这个结果

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

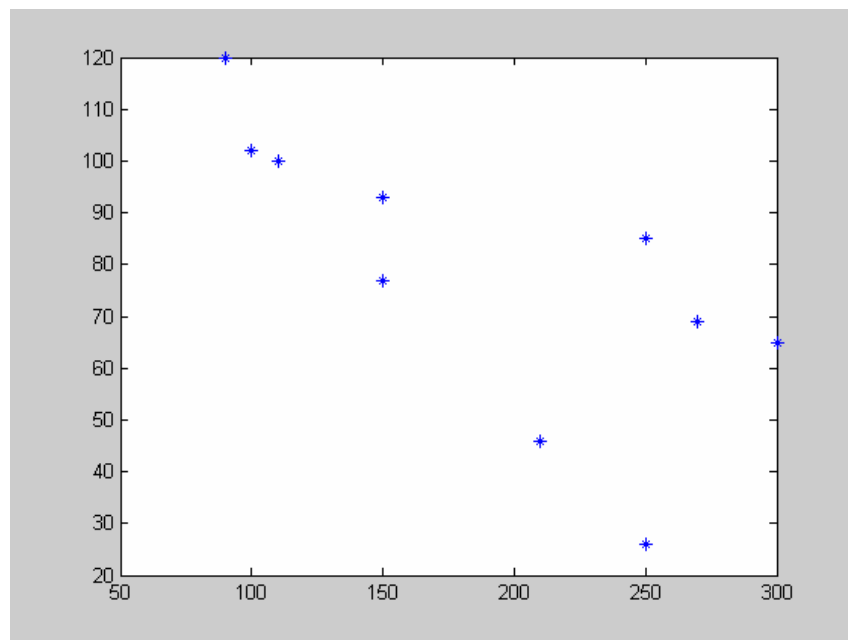
x_1	120	140	190	130	155	175	125	145	180	150
x_2	100	110	90	150	210	150	250	270	300	250
y	102	100	120	77	46	93	26	69	65	85

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



y 关于 x_1 的散点图



y 关于 x_2 的散点图

设回归模型为 $y = b_0 + b_1x_1 + b_2x_2$

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

得到 $b = 66.5176 \quad 0.4139 \quad -0.2698$
 $b_{int} = -32.5060 \quad 165.5411$
 $\quad -0.2018 \quad 1.0296$
 $\quad -0.4611 \quad -0.0785$
 $stats = 0.6527 \quad 6.5786 \quad 0.0247$

可以看出结果不是太好：

$p=0.0247$ ，取 $\alpha=0.05$ 时回归模型可用，但取 $\alpha=0.01$ 则模型不能用； $R^2=0.6527$ 较小

\hat{b}_0, \hat{b}_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。

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, 或者用线性模型的效果不太好, 就可以选用多项式回归.

1. 一元多项式回归 命令:polyfit

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.

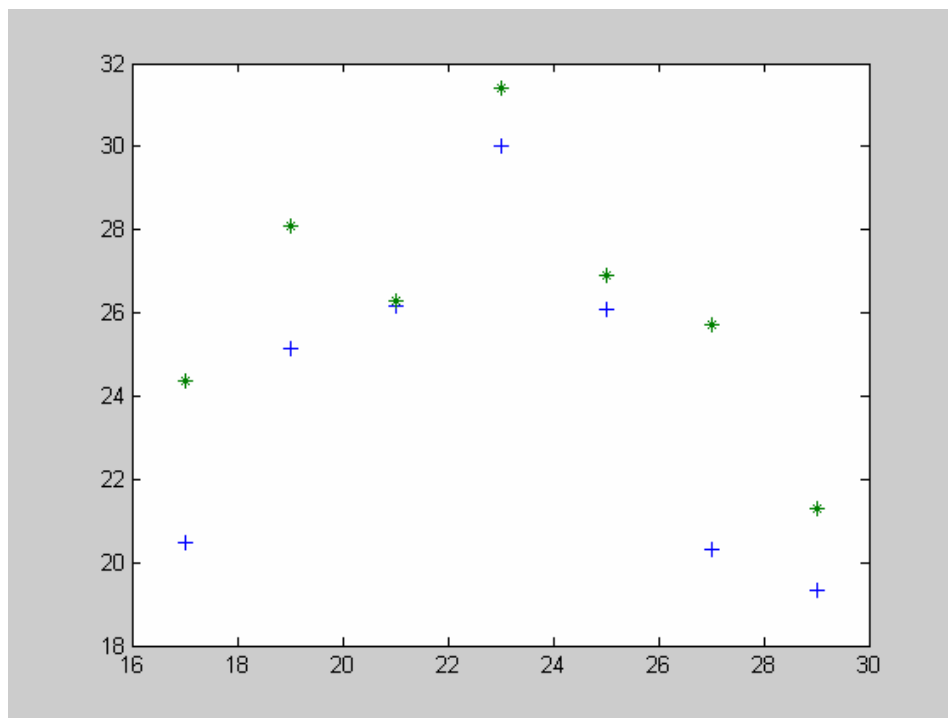
解 画出散点图先

```
x=[17 19 21 23 25 27 29]
```

```
y1=[20.48 25.13 26.15 30.0 26.1 20.3 19.35]
```

```
y2=[24.35 28.11 26.3 31.4 26.92 25.7 21.3]
```

```
plot(x , y1 , '+', x , y2 , '*')
```



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：

$$y = a_0 + a_1x + a_2x^2$$

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];  
[a,s]=polyfit(x0,y0,2); a
```

得到 $a = -0.2003 \quad 8.9782 \quad -72.2150$

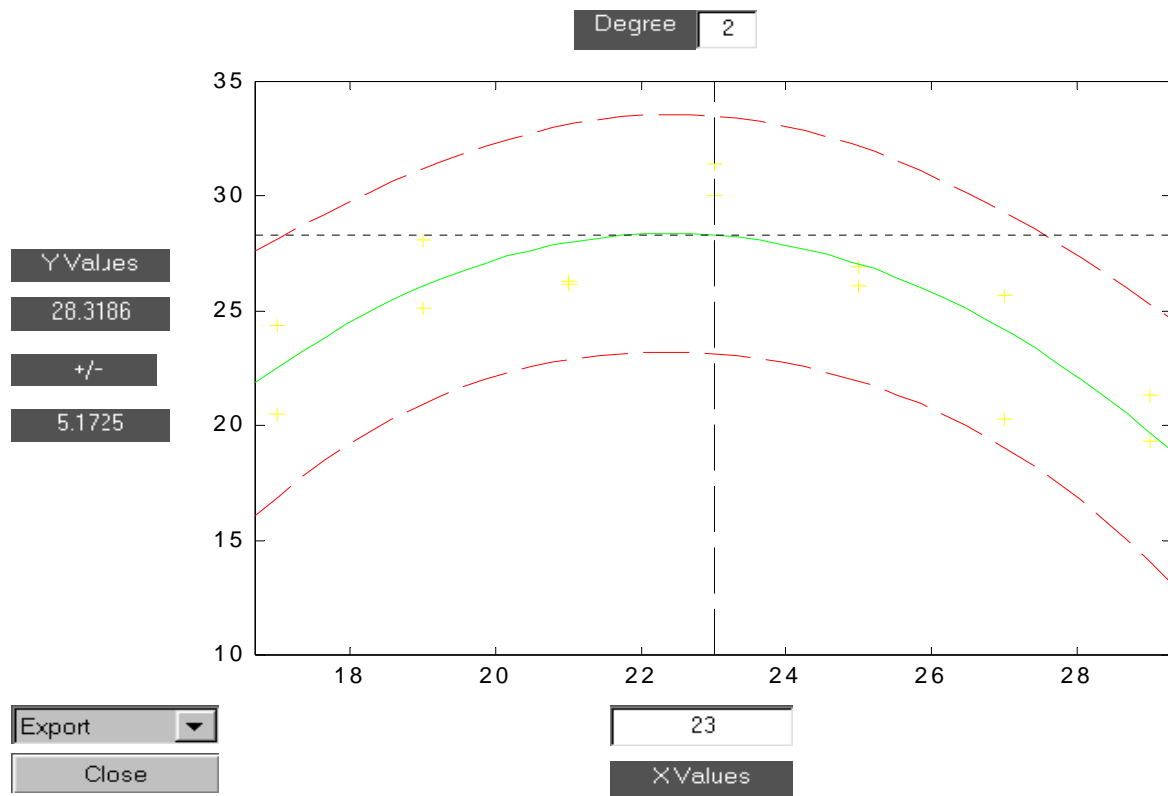
即 $a_2 = -0.2003$, $a_1 = 8.9782$, $a_0 = -72.2150$

上面的 **s** 是一个数据结构，用于计算函数值，如

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

得到 **y** 的拟合值及其置信区间半径 **delta**

```
y    = 22.5243  26.0582  27.9896  28.3186  27.0450  
      24.1689  19.6904  22.5243  26.0582  27.9896  
      28.3186  27.0450  24.1689  19.6904  
delta = 5.6275  5.1195  5.1195  5.1725  5.1195  
      5.1195  5.6275  5.6275  5.1195  5.1195  
      5.1725  5.1195  5.1195  5.6275
```



用`polytool(x0,y0,2)`，可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，两侧红线是 y 的置信区间。可用鼠标移动图中的十字线来改变图下方的 x 值，也可在窗口内输入，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，可输出回归系数等。这个命令的用法与下面介绍的**rstool**相似

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 α （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) $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$

purequadratic(纯二次) $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{j=1}^m \beta_{jj} x_j^2$

interaction(交叉) $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{1 \leq j \neq k \leq m} \beta_{jk} x_j x_k$

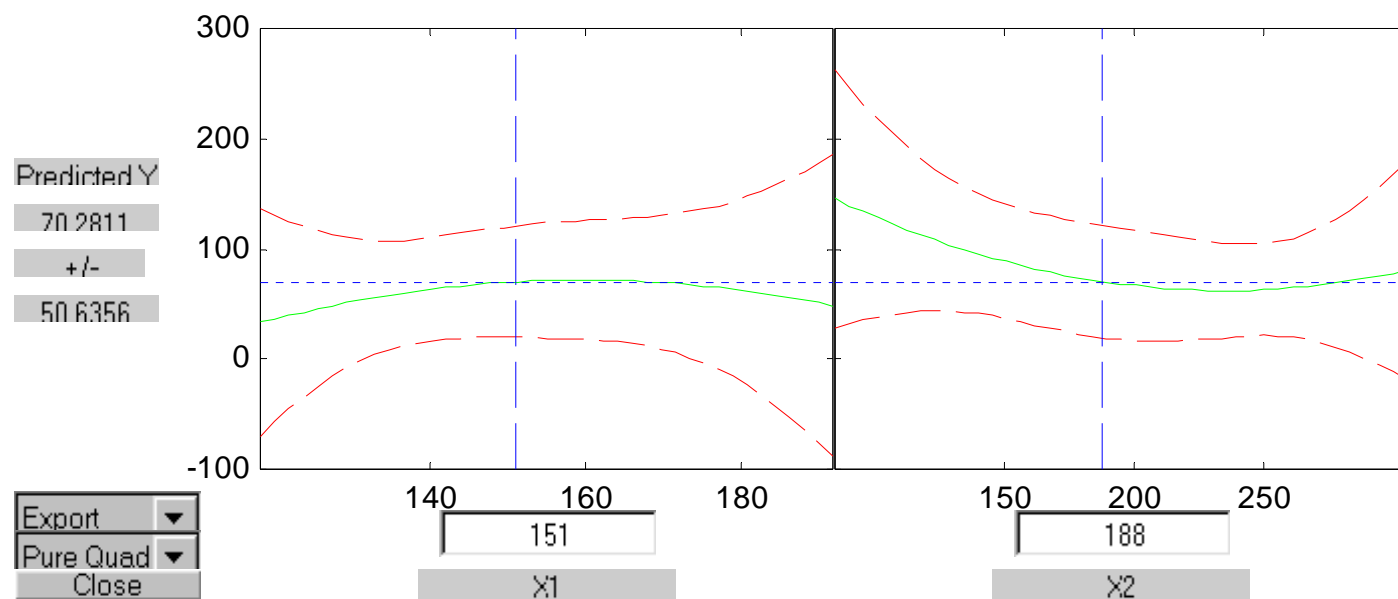
quadratic(完全二次) $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{1 \leq j, k \leq m} \beta_{jk} x_j x_k$

对例2 商品销售量与价格问题，选择纯二次模型，即

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2$$

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')
```



得到上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，用这种画面可回答例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)。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，可以比较它们的剩余标准差，发现模型**rmse=16.6436**最小.

§ 2 非线性回归和逐步回归

2.1 非线性回归

非线性回归是指因变量 y 对回归系数 β_1, \dots, β_m (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 求经验公式 $y = a + bx^2$ ，使它与下表所示的数据拟合

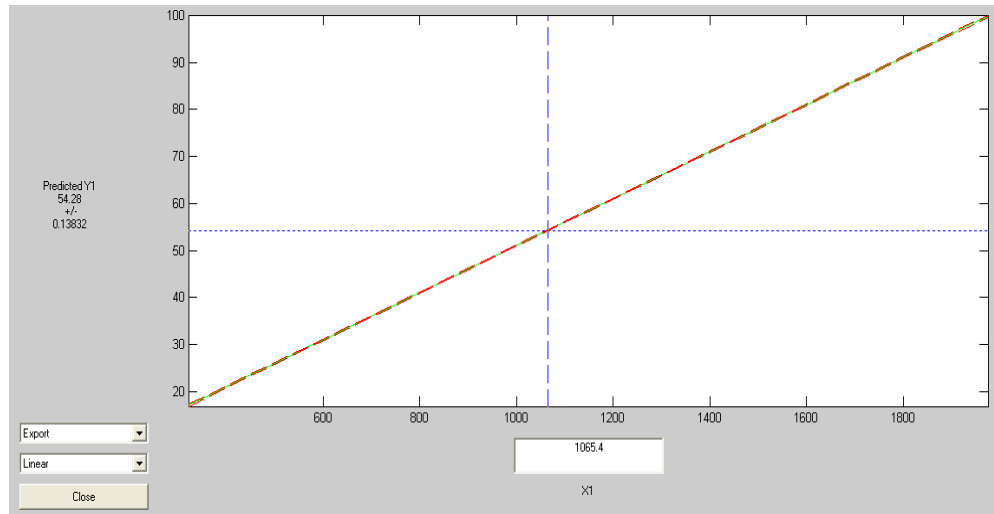
x	19	25	31	38	44
y	19.0	32.3	49.0	73.3	97.8

解 作 $x = x^2$ 变换，再用最小二乘法

```
x=[19  25  31  38  44]';
```

```
y=[19.0  32.3  49.0  73.3  97.8]';
```

```
rstool(x=x.^2,y,'linear');
```



在左下角的**Export**列表框中选择**All**传送参数，键入**beta** 得结果

beta = 0.9726 0.0500

说明：图形是关于自变量 $x = x^2$ 的，故显示为直线。

若使用插值与拟合方法

```
x=[19 25 31 38 44]';
```

```
y=[19.0 32.3 49.0 73.3 97.8]';
```

```
r=[ones(5,1),x.^2];
```

```
ab=r\y
```

```
x0=19:0.1:44;
```

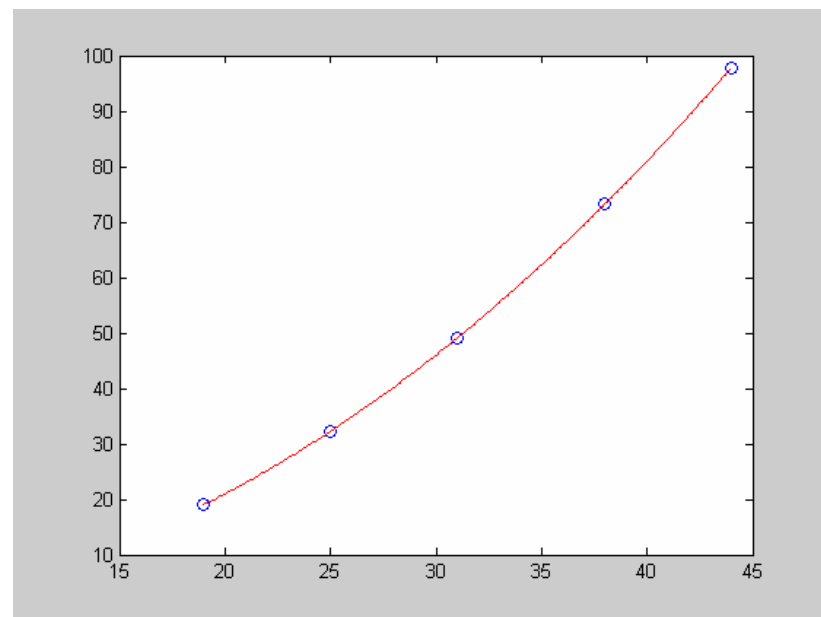
```
y0=ab(1)+ab(2)*x0.^2;
```

```
plot(x,y,'o',x0,y0,'r')
```

得结果 **ab = 0.9726 0.0500**

即 **a = 0.9726** ， **b=0.0500**

由此可以看出，两种方法结论一致。



例5 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

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

其中 β_1, \dots, β_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，试由此确定参数 β_1, \dots, β_5 ,并给出其置信区间

β_1, \dots, β_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)

运行结果

回归系数及其置信区间

betaa =

0.0628	-0.0377	0.1632
0.0400	-0.0312	0.1113
0.1124	-0.0609	0.2857
1.2526	-0.7467	3.2519
1.1914	-0.7381	3.1208

y的预测值

yhat =

8.4179
3.9542
4.9109
-0.0110
2.6358
14.3402
2.5662
4.0385
13.0292
8.3904
-0.0216
11.4701
3.4326

置信区间半径

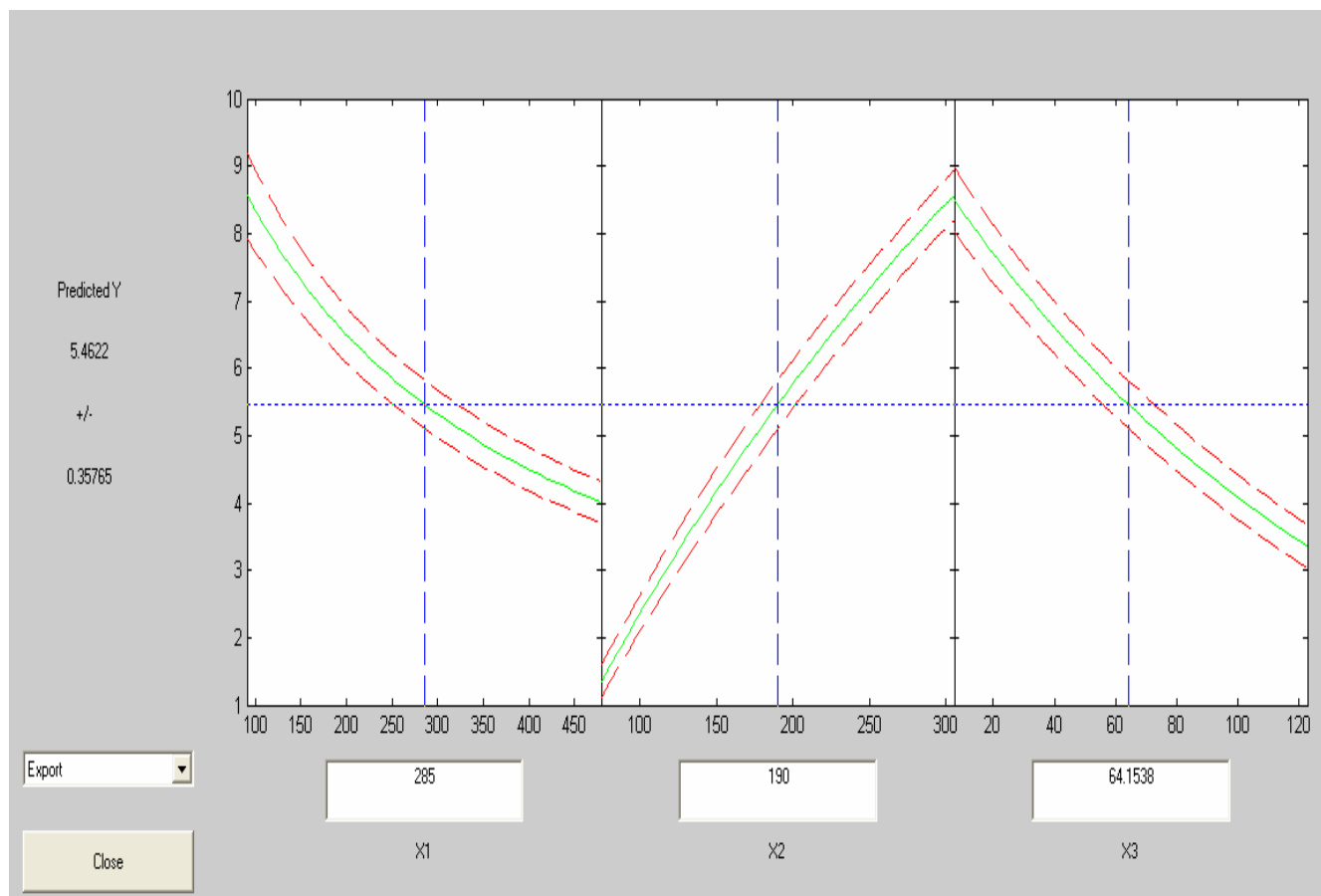
delta =

0.2805
0.2474
0.1766
0.1875
0.1578
0.4236
0.2425
0.1638
0.3426
0.3281
0.3699
0.3237
0.1749

y的预测值及其置信区间的半径，置信区间为 $yhat \pm delta$
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), rmse`**

可看到以下画面，并传出剩余标准差 **`rmse= 0.1933`**



2.2 逐步回归

实际问题中影响因变量的因素可能很多，我们希望能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 $S=\{x_1, \dots, x_m\}$ ，从中选出一个子集 $S_1 \subset S$ ，设 S_1 中有 l 个自变量($l=1, \dots, m$)，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

$$s^2 = \frac{Q}{n-l-1}$$

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

1. 要适当地选定引入变量的显著性水平 α_{in} 和剔除变量的显著性水平 α_{out} . 显然 α_{in} 越大, 引入的变量越多; α_{out} 越大, 剔除的变量越少。
2. 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命令是 **stepwise**, 它提供了一个交互式画面, 通过这个工具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)

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（缺省时设定为空），**alpha**为显著性水平。

Stepwise Regression 窗口显示回归系数及其置信区间、和其它一些统计量的信息。其中点表示回归系数的值，点两边的 (实或虚)水平直线段表示其置信区间；虚线表示该变量的拟合系数与0无显著差异，实线表示有显著差异；绿色的线表明当前在模型中的变量，红色的线表明从模型中移去的变量。点击一条直线会改变其状态。在这个窗口中有**Export**按钮，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。

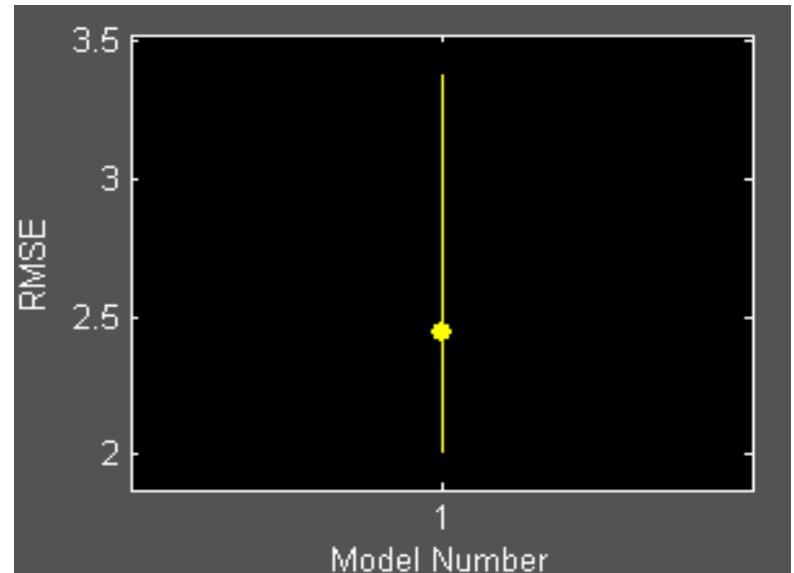
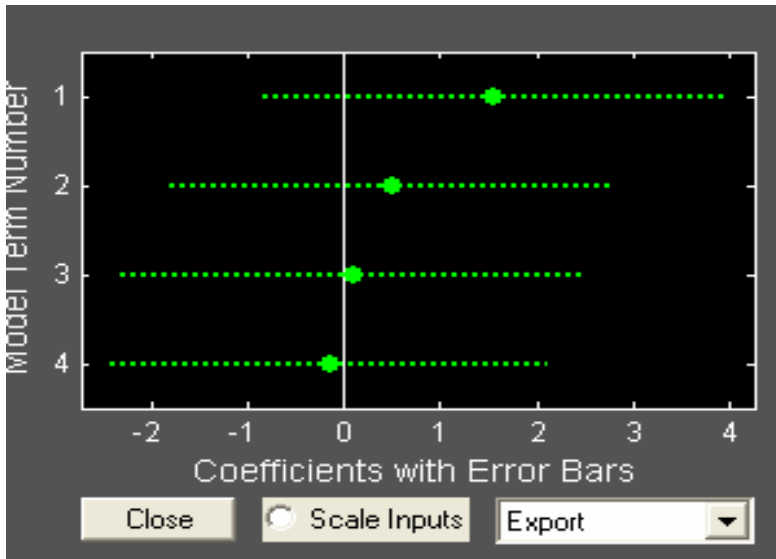
例6 水泥凝固时放出的热量 y 与水泥中 4 种化学成分 x_1, x_2, x_3, x_4 有关, 今测得一组数据如下, 试用逐步回归来确定一个线性模型.

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)
meanx=mean(x); % 计算 x 的平均值
meany=mean(y); % 计算 y 的平均值
```

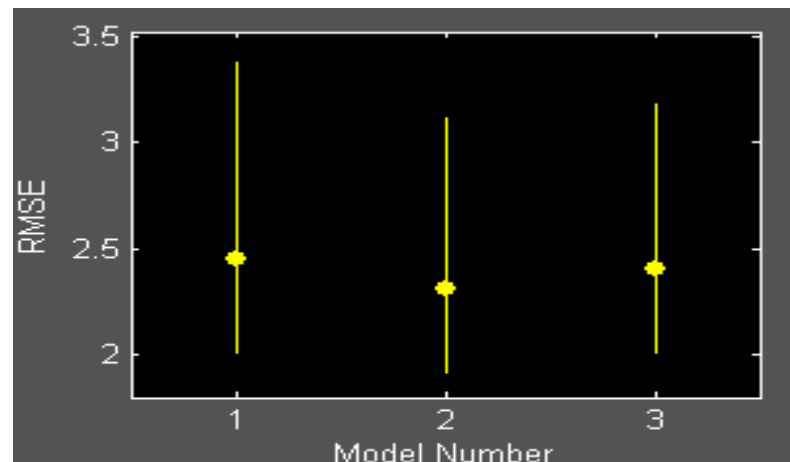
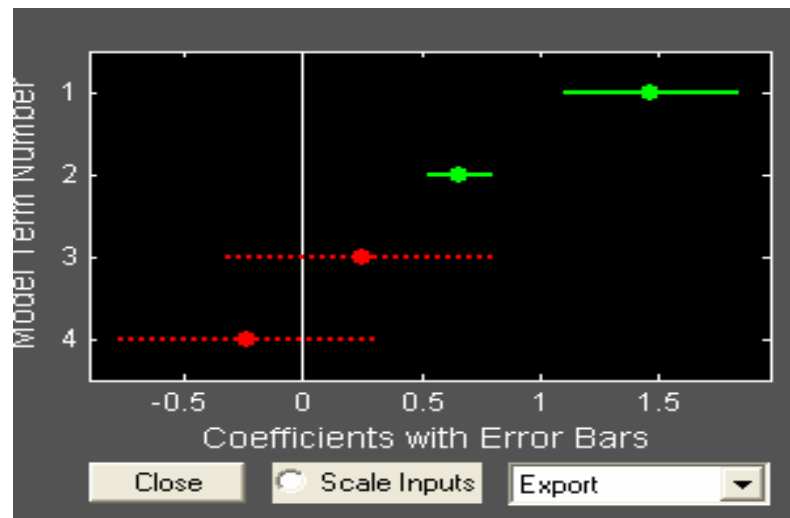
得到如下图形界面



Column #	Parameter	Confidence Intervals	
		Lower	Upper
1	1.551	-0.8319	3.934
2	0.5102	-1.806	2.826
3	0.1019	-2.313	2.517
4	-0.1441	-2.413	2.125
RMSE		F	P
2.446		111.5	4.756e-007
R-square			
0.9824			

Buttons: Close, Help

四条直线均为虚线，说明模型的显著性不好，从图中可以看出， x_3 ， x_4 显著性最差，点击第 3，4 两条虚线，移去这两个变量后的统计结果如下



Column #	Parameter	Confidence Intervals	
		Lower	Upper
1	1.468	1.1	1.836
2	0.6623	0.5232	0.8013
3	0.25	-0.3235	0.8236
4	-0.2365	-0.7746	0.3015
RMSE		F	P
2.406		229.5	4.407e-009
R-square			
0.9787			

Close Help

表中的 x_3 , x_4 两行用红色显示, 表明它们已移去。同时直线1和直线2变为实线。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, 但统计量 F 的值明显增大, 因此新的回归模型更好一些。在 **Export** 中选择 **All-beta, betaci, in, out** 输入命令

$$\text{meany-beta}(1) * \text{meanx}(1) - \text{beta}(2) * \text{meanx}(2)$$

$$(b_0 = \bar{y} - 1.4683\bar{x}_1 - 0.6623\bar{x}_2)$$

求出常数项 **b0=52.5773**

得最终的模型为

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

第5章 马氏链模型

§ 1 随机过程的基本概念

随机试验的结果有多种可能性，可以用一个随机变量(或随机向量)来描述。在很多情况下，人们需要对随机现象进行多次观测，甚至连续不断地观测其的变化过程。这就要研究无限多个，即一族随机变量。随机过程研究的是随机现象变化过程的概率规律性。

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 $t \in T$, ξ_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。 ξ_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。

本章主要介绍一类特殊的随机序列—马尔可夫链

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 ξ_n 表示第 n 次检验结果，则 ξ_n 是一个随机变量。不断检验，得到一系列随机变量 ξ_1, ξ_2, \dots ，记为 $\{\xi_n, n=1, 2, \dots\}$ 。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”， $j = 1, \dots, m$ 。则 $\{\xi_n, n=1, 2, \dots\}$ 是一个随机序列，其状态空间 $E = \{1, 2, \dots, m\}$.

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， $\{\xi_t, t \in [0, \infty]\}$ 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, \dots, t_n ，相应的随机变量 $\xi_{t_1}, \dots, \xi_{t_n}$ 的联合分布函数记为

$$F_{t_1 \dots t_n}(x_1, \dots, x_n) = P\{\xi_{t_1} \leq x_1, \dots, \xi_{t_n} \leq x_n\} \quad \dots\dots (1)$$

由 n 及 $t_i (i=1, \dots, n)$ 的任意性，(1) 式给出了一族分布函数，

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

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。

§ 2 马尔可夫链

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。例如，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。描述这类随机现象的数学模型称为马氏链模型。

马氏链模型描述的是一类重要的随机动态系统(过程)的模型，这种系统的特点是

- 1) 系统在每个时期所处的状态是随机的；
- 2) 从某一时期到下一时期的状态按一定概率转移；
- 3) 下一时期状态只取决于本时期状态和转移概率：
即已知现在，将来与过去无关(无后效性)。

定义2 设 $\{\xi_n, n=1,2,\dots\}$ 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 $i,j,i_k \in E (k=1,\dots, n-1)$ ，有

$$\begin{aligned} &P\{\xi_{n+m} = j \mid \xi_n = i, \xi_{n-1} = i_{n-1}, \dots, \xi_1 = i_1\} \\ &= P\{\xi_{n+m} = j \mid \xi_n = i\} \quad \dots\dots\dots (2) \end{aligned}$$

则称 $\{\xi_n, n=1,2,\dots\}$ 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 $\{\xi_n, n=1,2,\dots\}$ 具有马氏性，即 $\{\xi_n, n=1,2,\dots\}$ 是一个马尔可夫链.

定义3 设 $\{\xi_n, n=1,2,\dots\}$ 是一个马氏链。如果等式(2)右边的条件概率与 n 无关, 即

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

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)式称为时齐性。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。

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 $\{\xi_n, n=1,2,\dots\}$, 称以 m 步转移概率 $p_{ij}(m)$ 为元素的矩阵 $P(m)=(p_{ij}(m))$ 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:

(1) 对一切 $i, j \in E, 0 \leq p_{ij}(m) \leq 1$;

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

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

用马尔可夫链来描述实际问题时, 首先要确定其状态空间及参数集合, 然后确定其一步转移概率。关于这一概率的确定, 可由问题的内在规律得到, 也可由过去经验给出, 还可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现故障，现每隔**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
111011011010111101110111101111110011011111100111

解 设 $X_n (n=1, \dots, 97)$ 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='111001001111111001111011111100111111110001101101';  
a2='111011011010111101110111101111110011011111100111';  
a=[a1 a2];  
f00=length(findstr('00',a))  
f01=length(findstr('01',a))  
f10=length(findstr('10',a))  
f11=length(findstr('11',a))
```

如果把上述数据序列保存到纯文本文件**data.txt**中，存放在**Matlab**下的**work**子目录中，编程如下

```
clc,clear  
format rat  
fid=fopen('data.txt','r');  
a=[];  
while (~feof(fid))  
    a=[a fgetl(fid)];  
end
```

```
for i=0:1  
    for j=0:1  
        s=[int2str(i),int2str(j)];  
        f(i+1,j+1)=length(findstr(s,a));  
    end  
end  
fs=sum(f');  
for i=1:2  
    f(i,:)=f(i,:)/fs(i);  
end  
f
```

求得**96**次状态转移的情况是：

$0 \rightarrow 0$:8次 ; $0 \rightarrow 1$:18次 ; $1 \rightarrow 0$:18次 ; $1 \rightarrow 1$:52次

因此一步转移概率可用频率近似地表示为

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

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

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

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

一步转移矩阵为

$$\begin{pmatrix} \frac{4}{13} & \frac{9}{13} \\ \frac{9}{35} & \frac{26}{35} \end{pmatrix}$$

例5 设一随机系统状态空间，记录观测系统所处状态如下

n_{ij} $i \backslash j$	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 $p_{ij} = \frac{n_{ij}}{n_i}$

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
```

运行结果

$$f = \begin{pmatrix} 4 & 4 & 1 & 1 \\ 3 & 2 & 4 & 2 \\ 4 & 4 & 2 & 1 \\ 0 & 1 & 4 & 2 \end{pmatrix}$$

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

例6 (带反射壁的随机徘徊问题) 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 $q (=1-p)$ 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 ξ_n 表示徘徊 n 步后的质点位置。 $\{\xi_n, n=1,2,\dots\}$ 是一个马尔可夫链，其状态空间 $E = \{1,2,\dots,s\}$ ，写出转移矩阵 P 。

解

$$P\{\xi_0 = i\} = \begin{cases} 1, & \text{当 } i = 2 \text{ 时} \\ 0, & \text{当 } i \neq 2 \text{ 时} \end{cases} \quad p_{1j} = \begin{cases} q, & \text{当 } j = 1 \text{ 时} \\ p, & \text{当 } j = 2 \text{ 时} \\ 0, & \text{其它} \end{cases}$$

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

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

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 $\{\xi_n, n = 1, 2, \dots\}$ 是一个马尔可夫链，其状态空间 $E = \{1, 2, \dots\}$ ，则对任意正整数 m, n 和 $i, j \in E$ 有

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

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 $P^{(n)} = P^{(0)} \cdot P^n$ 。此式即为马尔科夫预测模型。

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， $\{\xi_n, n=1, 2, \dots\}$ 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下表给出。求顾客第四次购买各家味精的概率。

		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 $P^{(0)} = [0.2 \quad 0.4 \quad 0.4]$

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

Matlab程序:

```
p0=[0.2 0.4 0.4];
```

```
p=[ 0.8 0.1 0.1
```

```
0.5 0.1 0.4
```

```
0.5 0.3 0.2];
```

```
p3=p0*p^3
```

运行结果为

p3 = 0.7004 0.1360 0.1636

即顾客第四次购买各家味精的概率为

$$P^{(3)} = P^{(0)} P^3 = [0.7004 \quad 0.136 \quad 0.1636]$$

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：

$$\text{转移矩阵 } P = \begin{pmatrix} 0.5 & 0.5 \\ 0.7 & 0.3 \end{pmatrix} \quad \lim_{n \rightarrow \infty} P^n = \begin{pmatrix} \frac{7}{12} & \frac{5}{12} \\ \frac{7}{12} & \frac{5}{12} \end{pmatrix}$$
$$\text{又若取 } u = \begin{pmatrix} \frac{7}{12} & \frac{5}{12} \end{pmatrix}$$

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：

(1) P 有唯一的不动点向量 W ， W 的每个分量为正。

(2) 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(信息传播问题) 一条新闻在 $a_1, a_2, \dots, a_n, \dots$ 等人中间传播，传播方式是： 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，研究消息的真实程度如何？

解 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：

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

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

$$\lim_{n \rightarrow \infty} p_{ij}(n) = \pi_j \text{ (不依赖于 } i \text{)} \quad \text{或}$$

$$\lim_{n \rightarrow \infty} P(n) = \lim_{n \rightarrow \infty} P^n = \begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_j & \cdots \\ \pi_1 & \pi_2 & \cdots & \pi_j & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \pi_1 & \pi_2 & \cdots & \pi_j & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

则称此链具有遍历性。又若 $\sum_j \pi_j = 1$ ，则同时称

$\pi = (\pi_1, \pi_2, \cdots)$ 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 $\{\xi_n, n = 1, 2, \dots\}$ 的状态空间为 $E = \{a_1, \dots, a_N\}$, $P = (p_{ij})$ 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 $p_{ij}^{(m)} > 0$, $i, j = 1, \dots, N$, 则此链具有遍历性; 且有极限分布 $\pi = (\pi_1, \dots, \pi_N)$, 它是方程组

$$\pi = \pi P \text{ 即 } \pi_j = \sum_{i=1}^N \pi_i p_{ij}, j = 1, \dots, N \text{ 的满足条件}$$

$$\pi_j > 0, \sum_{j=1}^N \pi_j = 1 \text{ 的唯一解。}$$

例9 据例7中给出的一般顾客购买三种味精倾向的转移矩阵，预测经长期多次购买之后，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
```

也可用转移矩阵 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。编程如下

```
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
```

运行结果

$x =$

[Inf, 5/7, NaN]

[Inf, 11/84, Inf]

[Inf, 13/84, Inf]

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

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**5/7**，*B*、*C*两厂产品分别占有市场的 **11/84**，**13/84**

2.4 吸收链

例如：若马氏链的转移矩阵

$$P = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{pmatrix} 0.3 & 0.3 & 0 & 0.4 \\ 0.2 & 0.3 & 0.2 & 0.3 \\ 0 & 0.3 & 0.3 & 0.4 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix}$$

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：

$$P = \begin{pmatrix} I_r & 0 \\ R & S \end{pmatrix} \dots\dots\dots (4)$$

其中 I_r 为 r 阶单位阵， 0 为 $r \times s$ 零阵， R 为 $s \times r$ 矩阵， S 为 $s \times s$ 矩阵

$$P^n = \begin{pmatrix} I_r & 0 \\ Q & S^n \end{pmatrix} \dots\dots\dots (5)$$

(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 $F = (I_r - S)^{-1}$ ，则 F 称为基矩阵。

对具有标准形式(4)转移矩阵的吸收链，可证明以下定理：

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， $C = (1 \ 1 \ \dots \ 1)^T$ ，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 $B = FR = (b_{ij})$ ，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，若甲队赢得**1**分，则甲队的总分将增加**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 **a_0, a_1, a_2, a_3, a_4** ，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

$$P = \begin{matrix} & \begin{matrix} a_0 & a_1 & a_2 & a_3 & a_4 \end{matrix} \\ \begin{matrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{matrix} & \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1-p & 0 & p & 0 & 0 \\ 0 & 1-p & 0 & p & 0 \\ 0 & 0 & 1-p & 0 & p \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{matrix}$$

变形得

$$P = \begin{matrix} & a_0 & a_4 & a_1 & a_2 & a_3 \\ \begin{matrix} a_0 \\ a_4 \\ a_1 \\ a_2 \\ a_3 \end{matrix} & \left[\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1-p & 0 & 0 & p & 0 \\ 0 & 0 & 1-p & 0 & p \\ 0 & p & 0 & 1-p & 0 \end{array} \right] & \dots\dots & (6) \end{matrix}$$

将**(6)**式改记为标准形式

$$P = \begin{bmatrix} I_2 & 0 \\ R & S \end{bmatrix} \quad \text{其中}$$

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

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

其中 $q = 1 - p$

因 a_2 是初始状态，据定理5，甲队获得1，2，3分的平均次数

$$\frac{q}{1-2pq}, \frac{1}{1-2pq}, \frac{p}{1-2pq}$$

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

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

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

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)
```

结果如下

f =

$$[(-1+p*q)/(-1+2*p*q), \quad -p/(-1+2*p*q), \quad -p^2/(-1+2*p*q)]$$

$$[\quad -q/(-1+2*p*q), \quad -1/(-1+2*p*q), \quad -p/(-1+2*p*q)]$$

$$[\quad -q^2/(-1+2*p*q), \quad -q/(-1+2*p*q), \quad (-1+p*q)/(-1+2*p*q)]$$

n =

$$[-(1-p*q+p+p^2)/(-1+2*p*q)]$$

$$[\quad -(q+1+p)/(-1+2*p*q)]$$

$$[(-q^2-q-1+p*q)/(-1+2*p*q)]$$

b =

$$[(-1+p*q)/(-1+2*p*q)*q, \quad -p^3/(-1+2*p*q)]$$

$$[\quad -q^2/(-1+2*p*q), \quad -p^2/(-1+2*p*q)]$$

$$[\quad -q^3/(-1+2*p*q), \quad (-1+p*q)/(-1+2*p*q)*p]$$

§ 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。估计其转移概率如下表所示，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？

转移概率		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

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 $\{X_n, n=1, 2, \dots\}$ 是一个马尔可夫链，其转移矩阵 P 由上表给出。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

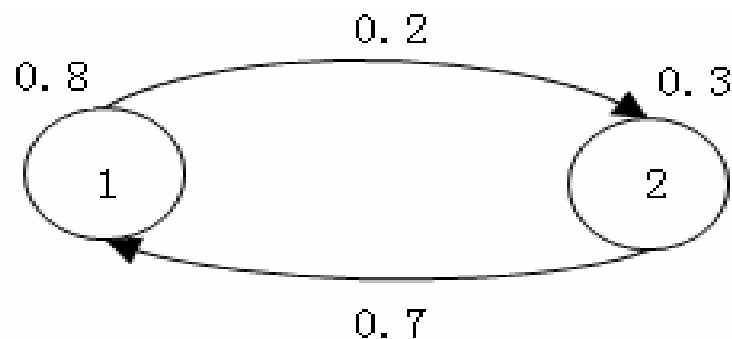
$$\begin{cases} p_1 = 0.2p_1 + 0.8p_2 + 0.1p_3 \\ p_2 = 0.8p_1 + 0.3p_3 \\ p_3 = 0.2p_2 + 0.6p_3 \\ p_1 + p_2 + p_3 = 1 \end{cases}$$

得极限概率 $p_1 = 17/41, p_2 = 16/41, p_3 = 8/41$

由计算看出，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**17/41, 16/41, 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。

例12（健康与疾病问题） 人的健康状态随着时间的推移会随机地发生转变，保险公司要对投保人未来的健康状态作出估计，以制订保险金和理赔金的数额。人的健康状况分为健康和疾病两种状态，设对特定年龄段的人，今年健康、明年保持健康状态的概率为**0.8**，而今年患病、明年转为健康状态的概率为**0.7**，若某人投保时健康，求**10**年后他仍处于健康状态的概率状态与状态转移。

解 状态 $X_n = \begin{cases} 1, & \text{第 } n \text{ 年健康} \\ 2, & \text{第 } n \text{ 年疾病} \end{cases}$



状态概率 $a_i(n) = P(X_n = i), i = 1, 2, n = 0, 1, \dots$

转移概率 $p_{ij} = P(X_{n+1} = j | X_n = i)$, $i, j = 1, 2$, $n = 0, 1, \dots$

$$p_{11} = 0.8, \quad p_{12} = 1 - p_{11} = 0.2, \quad p_{21} = 0.7, \quad p_{22} = 1 - p_{21} = 0.3$$

X_{n+1} 只取决于 X_n 和 p_{ij} , 与 X_{n-1}, \dots 无关, 状态转移具有无后效性。

$$a_1(n+1) = a_1(n)p_{11} + a_2(n)p_{21}, \quad a_2(n+1) = a_1(n)p_{12} + a_2(n)p_{22}$$

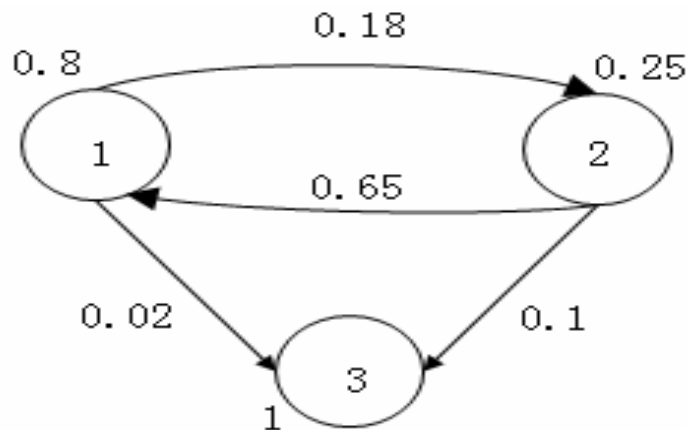
给定 $a(0)$, 预测 $a(n)$, $n=1, 2, \dots$

	n	0	1	2	3	∞
投保时健康	$a_1(n)$	1	0.8	0.78	0.778	7/9
	$a_2(n)$	0	0.2	0.22	0.222	2/9
投保时疾病	$a_1(n)$	0	0.7	0.77	0.777	7/9
	$a_2(n)$	1	0.3	0.23	0.223	2/9

$n \rightarrow \infty$ 时状态概率趋于稳定值，稳定值与初始状态无关。

如果再增加一种死亡状态，则

$$X_n = \begin{cases} 1 & \text{健康} \\ 2 & \text{疾病} \\ 3 & \text{死亡} \end{cases}$$



$$p_{11}=0.8, p_{12}=0.18, p_{13}=0.02$$

$$p_{21}=0.65, p_{22}=0.25, p_{23}=0.1$$

$$p_{31}=0, p_{32}=0, p_{33}=1$$

$$a_1(n+1) = a_1(n)p_{11} + a_2(n)p_{21} + a_3(n)p_{31}$$

$$a_2(n+1) = a_1(n)p_{12} + a_2(n)p_{22} + a_3(n)p_{32}$$

$$a_3(n+1) = a_1(n)p_{13} + a_2(n)p_{23} + a_3(n)p_{33}$$

设投保时处于健康状态，预测 $a(n)$, $n=1,2,\dots$

n	0	1	2	3	...	50	...	∞
$a_1(n)$	1	0.8	0.757	0.7285	...	0.1293	...	0
$a_2(n)$	0	0.18	0.189	0.1835	...	0.0326	...	0
$a_3(n)$	0	0.02	0.054	0.0880	...	0.8381	...	1

不论初始状态如何，最终都要转到状态**3**；一旦 $a_1(k)=$
 $a_2(k)=0$ ， $a_3(k)=1$ ，则对于 $n>k$ ， $a_1(n)=0$ ， $a_2(n)=0$ ， $a_3(n)=1$ ，
 即从状态**3**不会转移到其它状态。

例13（钢琴销售的存贮策略） 钢琴销售量很小，商店的库存量不大以免积压资金。一家商店根据经验估计，平均每周的钢琴需求为1架。现采用如下存贮策略：每周末检查库存量，仅当库存量为零时，才订购3架供下周销售；否则，不订购。估计在这种策略下失去销售机会的可能性有多大，以及每周的平均销售量是多少？

解 问题分析： 顾客的到来相互独立，需求量近似服从泊松分布，参数由需求均值为每周1架确定，由此计算需求概率。存贮策略是周末库存量为零时订购3架 → 周末的库存量可能是0, 1, 2, 3，周初的库存量可能是1, 2, 3。用马氏链描述不同需求导致的周初库存状态的变化。动态过程中每周销售量不同，失去销售机会（需求超过库存）的概率不同。可按稳态情况（时间充分长以后）计算失去销售机会的概率和每周的平均销售量。

模型建立：记 D_n 表示第 n 周需求量，则 $D_n \sim P(1)$,

$$P(D_n = k) = \frac{1}{k!} e^{-1}, (k = 0, 1, 2, \dots)$$

D_n	0	1	2	3	> 3
P	0.368	0.368	0.184	0.061	0.019

S_n 表示第 n 周初库存量(状态变量), $S_n \in \{1, 2, 3\}$,

状态转移规律

$$S_{n+1} = \begin{cases} S_n - D_n, & D_n < S_n \\ 3, & D_n \geq S_n \end{cases}$$

$$p_{11} = P(S_{n+1} = 1 | S_n = 1) = P(D_n = 0) = 0.368$$

$$p_{12} = P(S_{n+1} = 2 | S_n = 1) = 0$$

$$p_{13} = P(S_{n+1} = 3 | S_n = 1) = P(D_n \geq 1) = 0.632$$

.....

$$p_{33} = P(S_{n+1} = 3 | S_n = 3) = P(D_n = 0) + P(D_n \geq 3) = 0.448$$

$$\text{状态转移阵 } P = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix} = \begin{bmatrix} 0.368 & 0 & 0.632 \\ 0.368 & 0.368 & 0.264 \\ 0.184 & 0.368 & 0.448 \end{bmatrix}$$

$$\text{状态概率 } a_i(n) = P(S_n = i), i = 1, 2, 3$$

马氏链的基本方程

$$a(n+1) = a(n)P, P = \begin{bmatrix} 0.368 & 0 & 0.632 \\ 0.368 & 0.368 & 0.264 \\ 0.184 & 0.368 & 0.448 \end{bmatrix}$$

已知初始状态，可预测第 n 周初库存量 $S_n = i$ 的概率

正则链 $\Leftrightarrow \exists N, P^N > 0, P^2 > 0$ 可知此马氏链为正则链

稳态概率分布 w 满足 $wP = w$

$$\Rightarrow w = (w_1, w_2, w_3) = (0.285, 0.263, 0.452)$$

$$n \rightarrow \infty \quad \text{状态概率} \quad a(n) = (0.285, 0.263, 0.452)$$

模型求解

(1) 估计在这种策略下失去销售机会的可能性

D	0	1	2	3	> 3
P	0.368	0.368	0.184	0.061	0.019

$$w = (0.285, 0.263, 0.452)$$

n 充分大时 $P(S_n=i)=w_i$ ，第 n 周失去销售机会的概率

$$\begin{aligned} P(D_n > S_n) &= \sum_{i=1}^3 P(D_n > i | S_n = i) P(S_n = i) \\ &= P(D > 1)w_1 + P(D > 2)w_2 + P(D > 3)w_3 \\ &= 0.264 \times 0.285 + 0.080 \times 0.263 + 0.019 \times 0.452 \\ &= 0.105 \quad \text{长期看失去销售机会可能性约 } \mathbf{10.5\%} \end{aligned}$$

(2) 估计这种策略下每周的平均销售量

n 充分大时 $P(S_n=i)=w_i$ ，第 n 周平均售量

$$\begin{aligned} R_n &= \sum_{i=1}^3 \left[\sum_{j=1}^i j P(D_n = j, S_n = i) + i P(D_n > i, S_n = i) \right] \\ &= \sum_{i=1}^3 \left[\sum_{j=1}^i j P(D_n = j | S_n = i) + i P(D_n > i | S_n = i) \right] P(S_n = i) \\ &= 0.632 \times 0.285 + 0.896 \times 0.263 + 0.977 \times 0.452 \\ &= 0.857 \quad \text{长期看，每周的平均销售量为 } \mathbf{0.857(架)} \end{aligned}$$

思考：为什么这个数值略小于每周平均需求量1(架)？

敏感性分析

当平均需求在每周**1 (架)** 附近波动时, 最终结果有多大变化?

设 D_n 服从均值为 λ 的波松分布

$$P(D_n = k) = \lambda^k e^{-\lambda} / k!, \quad (k = 0, 1, 2, \dots)$$

状态转移阵 $P = \begin{bmatrix} e^{-\lambda} & 0 & 1 - e^{-\lambda} \\ \lambda e^{-\lambda} & e^{-\lambda} & 1 - (1 + \lambda)e^{-\lambda} \\ \lambda^2 e^{-\lambda} / 2 & \lambda e^{-\lambda} & 1 - (\lambda + \lambda^2 / 2)e^{-\lambda} \end{bmatrix}$

第 n 周(n 充分大)失去销售机会的概率 $P = P (Dn > Sn)$

λ	0.8	0.9	1.0	1.1	1.2
P	0.073	0.089	0.105	0.122	0.139

平均需求增长(减少)时，失去销售机会的概率将增长(减少)

第6章 时间序列模型

时间序列是按时间顺序排列的、随时间变化且相互关联的数据序列。分析时间序列的方法构成数据分析的一个重要领域，即时间序列分析。

时间序列根据所研究的依据不同，可有不同的分类

1. 按研究对象多少分：一元时间序列和多元时间序列；
2. 按时间连续性分：离散时间序列和连续时间序列；
3. 按序列的统计特性分：平稳时间序列和非平稳时间序列；
4. 按时间序列分布规律分：高斯型和非高斯型时间序列。

如果一个时间序列的概率分布与时间 t 无关，则称该序列为严格的（狭义的）平稳时间序列。

如果序列的一、二阶矩存在，且对任意时刻 t 满足：

（1）均值为常数

（2）协方差为时间间隔 τ 的函数。

则称该序列为宽平稳时间序列，也叫广义平稳时间序列。以后所研究的时间序列主要是宽平稳时间序列。

§ 1 确定性时间序列分析方法概述

时间序列预测技术就是通过对预测目标自身时间序列的处理，来研究其变化趋势的。一个时间序列往往是以下几类变化形式的叠加或耦合。

(1) **长期趋势变动**。是指时间序列朝着一定的方向持续上升或下降，或停留在某一水平上的倾向，它反映了客观事物的主要变化趋势。

(2) **季节变动**。

(3) **循环变动**。通常是指周期为一年以上，由非季节因素引起的涨落起伏波形相似的波动。

(4) **不规则变动**。通常它分为突然变动和随机变动。

通常用 T_t 表示长期趋势项, S_t 表示季节变动趋势项, C_t 表示循环变动趋势项, R_t 表示随机干扰项。常见的确定性时间序列模型有以下几种类型:

加法模型 $y_t = T_t + S_t + C_t + R_t$

乘法模型 $y_t = T_t \cdot S_t \cdot C_t \cdot R_t$

混合模型 $y_t = T_t \cdot S_t + R_t$, $y_t = S_t + T_t \cdot C_t \cdot R_t$

其中 y_t 是观测目标的观测记录, $E(R_t) = 0$, $E(R_t^2) = \sigma^2$

如果在预测时间范围以内, 无突然变动且随机变动的方差 σ^2 较小, 并且有理由认为过去和现在的演变趋势将继续发展到未来时, 可用一些经验方法进行预测, 具体方法如下:

1.1 移动平均法

设观测序列为 y_1, \dots, y_T , 取移动平均的项数 $N \leq T$

一次移动平均值计算公式

$$\begin{aligned} M_t^{(1)} &= \frac{1}{N} (y_t + y_{t-1} + \dots + y_{t-N+1}) \\ &= \frac{1}{N} (y_{t-1} + \dots + y_{t-N}) + \frac{1}{N} (y_t - y_{t-N}) \\ &= M_{t-1}^{(1)} + \frac{1}{N} (y_t - y_{t-N}) \end{aligned}$$

二次移动平均值计算公式

$$\begin{aligned} M_t^{(2)} &= \frac{1}{N} (M_t^{(1)} + \cdots + M_{t-N+1}^{(1)}) \\ &= M_{t-1}^{(2)} + \frac{1}{N} (M_t^{(1)} - M_{t-N}^{(1)}) \end{aligned}$$

当预测目标的基本趋势是在某一水平上下波动时，可用一次移动平均方法建立预测模型：

$$\hat{y}_{t+1} = M_t^{(1)} = \frac{1}{N} (\hat{y}_t + \cdots + \hat{y}_{t-N+1}), t = N, N+1, \cdots$$

其预测标准误差

$$S = \sqrt{\frac{\sum_{t=N+1}^T (\hat{y}_t - y_t)^2}{T - N}}$$

最近 N 期序列值的平均值作为未来各期的预测结果。一般 N 取值范围： $5 \leq N \leq 200$ 。当历史序列的基本趋势变化不大且序列中随机变动成分较多时， N 的取值应较大一些。否则 N 的取值应小一些。在有确定的季节变动周期的资料中，移动平均的项数应取周期长度。选择最佳 N 值的一个有效方法是，比较若干模型的预测误差。均方预测误差最小者为好。

当预测目标的基本趋势与某一线性模型相吻合时，常用二次移动平均法，但序列同时存在线性趋势与周期波动时，可用趋势移动平均法建立预测模型：

$$\hat{y}_{T+m} = a_T + b_T m, \quad m = 1, 2, \dots$$

其中
$$a_T = 2M_T^{(1)} - M_T^{(2)}, \quad b_T = \frac{2}{N-1}(M_T^{(1)} - M_T^{(2)})$$

例1 某企业1月~11月份的销售收入时间序列如下表所示。

取 $N=4$ ，试用简单一次滑动平均法预测第12月份的销售收入，并计算预测的标准误差。

月份 t	1	2	3	4	5	6
销售收入 y_t	533.8	574.6	606.9	649.8	705.1	772.0
月份 t	7	8	9	10	11	12
销售收入 y_t	816.4	892.7	963.9	1015.1	1102.7	

解：先计算 $M_t^{(1)} = \frac{1}{4}(y_t + y_{t-1} + \cdots + y_{t-3}), t = 4, 5, \cdots, 11$

$$\because \hat{y}_{t+1} = M_t^{(1)}, t = 4, 5, \cdots, 11 \quad \therefore \hat{y}_{12} = M_{11}^{(1)} = 993.6$$

预测的标准误差 $S = \sqrt{\frac{\sum_{t=5}^{11} (\hat{y}_t - y_t)^2}{11-4}} = 150.5$

Matlab程序

```
y=[533.8 574.6 606.9 649.8 705.1 772.0 816.4  
892.7 963.9 1015.1 1102.7];  
temp=cumsum(y);    % 求累积和  
mt=(temp(4:11)-[0 temp(1:7)])/4  
y12=mt(end)  
ythat=mt(1:end-1);  
fangcha=mean((y(5:11)-ythat).^2);  
sigma=sqrt(fangcha)
```

结果

temp = 1.0e+003 *

0.5338 1.1084 1.7153 2.3651 3.0702 3.8422
4.6586 5.5513 6.5152 7.5303 8.6330

mt = 591.2750 634.1000 683.4500 735.8250 796.5500
861.2500 922.0250 993.6000

y12 = 993.6000

ythat = 591.2750 634.1000 683.4500 735.8250 796.5500
861.2500 922.0250

fangcha = 2.2654e+004

sigma = 150.5121

1.2 指数平滑法

一次移动平均实际上认为最近 N 期数据对未来值影响相同，都加权 $1/N$ ；而 N 期以前的数据对未来值没有影响，加权为 0 。但二次及更高次移动平均数的权数却不是 $1/N$ ，且次数越高，权数的结构越复杂，但永远保持对称的权数，即两端项权数小，中间项权数大，不符合一般系统的动态性。一般说来历史数据对未来值的影响是随时间间隔的增长而递减的。所以，更切合实际的方法应是对各期观测值依时间顺序进行加权平均作为预测值。指数平滑法可满足这一要求，而且具有简单的递推形式。

设观测序列为 y_1, \dots, y_T , α 为加权系数, $0 < \alpha < 1$, 一次指数平滑公式为:

$$S_t^{(1)} = \alpha y_t + (1 - \alpha) S_{t-1}^{(1)} = S_{t-1}^{(1)} + \alpha (y_t - S_{t-1}^{(1)}) \quad \dots\dots\dots (1)$$

假定历史序列无限长, 则有

$$S_t^{(1)} = \alpha y_t + (1 - \alpha) [\alpha y_{t-1} + (1 - \alpha) S_{t-2}^{(1)}] = \dots = \alpha \sum_{j=0}^{\infty} (1 - \alpha)^j y_{t-j} \quad \dots\dots\dots (2)$$

表明 $S_t^{(1)}$ 是全部历史数据的加权平均, 加权系数分别为

$$\alpha, \alpha(1 - \alpha), \alpha(1 - \alpha)^2, \dots \quad \text{显然} \quad \sum_{j=0}^{\infty} \alpha(1 - \alpha)^j = \frac{\alpha}{1 - (1 - \alpha)} = 1$$

由于加权系数序列呈指数函数衰减, 加权平均又能消除或减弱随机干扰的影响, 所以**(2)**称为一次指数平滑.

类似地有

二次指数平滑公式 $S_t^{(2)} = \alpha S_t^{(1)} + (1 - \alpha)S_{t-1}^{(2)} \quad \dots\dots (3)$

三次指数平滑公式 $S_t^{(3)} = \alpha S_t^{(2)} + (1 - \alpha)S_{t-1}^{(3)} \quad \dots\dots (4)$

P 次指数平滑公式 $S_t^{(P)} = \alpha S_t^{(P-1)} + (1 - \alpha)S_{t-1}^{(P)} \quad \dots\dots (5)$

利用指数平滑公式可以建立指数平滑预测模型。原则上说，不管序列的基本趋势多么复杂，总可以利用高次指数平滑公式建立一个逼近很好的模型，但计算量很大。因此用的较多的是几个低阶指数平滑预测模型。

1) 一次指数平滑预测 $\hat{y}_{t+1} = S_t^{(1)}$

2) 线性趋势预测模型—**Brown**单系数线性平滑预测(二次指数平滑预测) $\hat{y}_{t+m} = a_t + b_t m, m = 1, 2, \dots$

$$\text{其中 } a_t = 2S_t^{(1)} - S_t^{(2)}, b_t = \frac{\alpha}{1-\alpha} (S_t^{(1)} - S_t^{(2)})$$

3) 二次曲线趋势预测模型—*Brown*单系数二次式平滑预测

$$\hat{y}_{t+m} = a_t + b_t m + \frac{1}{2} c_t m^2, m = 1, 2, \dots$$

$$\text{其中} \begin{cases} a_t = 3S_t^{(1)} - 3S_t^{(2)} + S_t^{(3)} \\ b_t = \frac{\alpha}{2(1-\alpha)^2} [(6-5\alpha)S_t^{(1)} - 2(5-4\alpha)S_t^{(2)} + (4-3\alpha)S_t^{(3)}] \\ c_t = \frac{\alpha^2}{2(1-\alpha)^2} [S_t^{(1)} - 2S_t^{(2)} + S_t^{(3)}] \end{cases}$$

指数平滑预测模型以时刻 t 为起点，综合历史序列信息，对未来进行预测。

选择合适的加权系数 α 是提高预测精度的关键环节。

据经验， α 的取值范围一般以 **0.1~0.3** 为宜。

α 值愈大，加权系数序列衰减速度愈快，所以 α 取值大小起着控制参加平均的历史数据个数的作用。 α 值愈大意味着采用的数据愈少。因此可得到选择 α 值的一些基本准则。

(1) 如果序列的基本趋势比较稳，预测偏差由随机因素造成，则 α 值应取小一些，以减少修正幅度，使预测模型能包含更多历史数据的信息。

(2) 如果预测目标的基本趋势已发生系统地变化，则 α 值应取得大一些。这样，可以偏重新数据的信息对原模型进行大幅度修正，以使预测模型适应预测目标的新变化。

由于指数平滑公式是递推计算公式，必须确定初始值

$$S_0^{(1)}, S_0^{(2)}, S_0^{(3)}$$

可以取前**3~5**个数据的算术平均值作为初始值。

例2 下表数据是某股票在**8**个连续交易日的收盘价，试用一次指数平滑法预测第**9**个交易日的收盘价（初始值 $S_0^{(1)}=y_1$, $\alpha=0.4$ ）

时间 t	1	2	3	4	5	6	7	8
价格 y_t	16.41	17.62	16.15	15.54	17.24	16.83	18.14	17.05

解 由于 $S_0^{(1)} = y_1$

$$S_t^{(1)} = \alpha y_t + (1 - \alpha) S_{t-1}^{(1)}, t = 1, 2, \dots, 8$$

求得 $\hat{y}_9 = S_8^{(1)} = 17.18$

预测标准误差
$$S = \sqrt{\frac{\sum_{t=2}^8 (\hat{y}_t - y_t)^2}{7}} = 0.96$$

Matlab 程序

```
alpha=0.4;  
y=[16.41 17.62 16.15 15.54 17.24 16.83 18.14 17.05];  
s1(1)=y(1);  
for i=2:8  
    s1(i)=alpha*y(i)+(1-alpha)*s1(i-1);  
end  
yhat9=s1(end)  
sigma=sqrt(mean((s1(1:end-1)-y(2:end)).^2))
```

运行结果

s1 =16.4100 yhat9 = 17.1828 sigma = 0.9613

例3 上例中用两次指数平滑法预测第9个交易日的收盘价

$$(S_0^{(1)} = S_0^{(2)} = y_1, \alpha = 0.4)$$

解 $a_8 = 2S_8^{(1)} - S_8^{(2)} = 17.38, b_8 = \frac{\alpha}{1-\alpha}(S_8^{(1)} - S_8^{(2)}) = 0.13$

$$\therefore \hat{y}_9 = a_8 + b_8 = 17.51$$

$$\text{取 } \hat{y}_{t+1} = a_t + b_t = (2S_t^{(1)} - S_t^{(2)}) + \frac{\alpha}{1-\alpha}(S_t^{(1)} - S_t^{(2)})$$

$$= S_t^{(1)} + \frac{1}{1-\alpha}(S_t^{(1)} - S_t^{(2)})$$

预测标准误差 $S = \sqrt{\frac{\sum_{t=1}^8 (\hat{y}_t - y_t)^2}{8-2}} = 1.21$

Matlab 程序

```
clc,clear
alpha=0.4;
y=[16.41 17.62 16.15 15.54 17.24 16.83 18.14 17.05];
s1(1)=y(1);
for i=2:8
    s1(i)=alpha*y(i)+(1-alpha)*s1(i-1);
end
s2=y(1);
for i=2:8
    s2(i)=alpha*s1(i)+(1-alpha)*s2(i-1);
end
a8=2*s1(8)-s2(8)
b8=alpha/(1-alpha)*(s1(8)-s2(8))
yhat9=a8+b8
yhat(1)=y(1)
for i=2:8
    yhat(i)=s1(i-1)+1/(1-alpha)*(s1(i-1)-s2(i-1));
end
temp=sum((yhat-y).^2);
sigma=sqrt(temp/6)
```

运行结果:

a8 =17.3801

b8 = 0.1315

yhat9 =17.5116

yhat =16.4100

sigma =1.2054

预测结果不如
一次指数平滑
法预测的预测
结果。

例4 (商品销售量预测问题)某商品前**5**年销售量见表。现希望据前**5**年的统计数据预测第**6**年起该商品在各季度中的销售量.

<div>年份 季度</div>	第一年	第二年	第三年	第四年	第五年
1	11	12	13	15	16
2	16	18	20	24	25
3	25	26	27	30	32
4	12	14	15	15	17

从表中可看出，该商品在前**5**年相同季节里的销售量呈增长趋势，而在同一年中销售量先增后减，第一季度的销售量最小，而第三季度的销售量最大。预测该商品以后的销售情况，据本例中数据的特征，可用回归分析方法按季度建立四个经验公式，分别用来预测以后各年同一季度的销售量.

例如，如果认为第一季度的销售量大体按线性增长，可设

$$y_t^{(1)} = at + b$$

由 **Matlab**

```
x=[[1:5]',ones(5,1)] ; y=[11 12 13 15 16]' ; z=x\y
```

运行结果 $z = \begin{matrix} 1.3000 \\ 9.5000 \end{matrix}$

求得 $a = z(1) = 1.3, b = z(2) = 9.5$

根据 $y_t^{(1)} = 1.3t + 9.5$ 预测第六年起第一季度销售量

$$y_6^{(1)} = 17.3, y_7^{(1)} = 18.6$$

由于数据较少，用回归分析效果不一定好。

如果认为销售量并非逐年等量增长，而是按前一年或前几年同期销售量的一定比例增长的，则可建立相应的差分方程模型。仍以第一季度为例，为简单起见不再引入上标，以 y_t 表示第 t 年第一季度的销售量，

建立差分公式： $y_t = a_1 y_{t-1} + a_2$ 或 $y_t = a_1 y_{t-1} + a_2 y_{t-2} + a_3$ 等

上述差分方程中的系数不一定能使所有统计数据吻合，较为合理的办法是用最小二乘法求一组总体吻合较好的数据。以建立二阶差分方程 $y_t = a_1 y_{t-1} + a_2 y_{t-2} + a_3$ 为例，选取 a_1, a_2, a_3 使下式最小：

$$\sum_{i=3}^5 [y_t - (a_1 y_{t-1} + a_2 y_{t-2} + a_3)]^2$$

Matlab 程序

```
y0=[11 12 13 15 16]';  
y=y0(3:5) ; x=[y0(2:4),y0(1:3),ones(3,1)];  
z=x\y
```

求得 $a_1 = z(1) = -1, a_2 = z(2) = 3, a_3 = z(3) = -8$

所求二阶差分方程 $y_t = -y_{t-1} + 3y_{t-2} - 8$

据这一方程可迭代求出以后各年第一季度销售量的预测值:

$$y_6 = 21, y_7 = 19, \dots$$

虽然这一差分方程恰好使前**5**年第一季度所有统计数据吻合，但这只是一个巧合，凭直觉，第六年估计值明显偏高，而第七年销售量预测值甚至小于第六年销售量。不难看出，如分别对第一季度建立一差分方程，则据统计数据拟合出的系数可能会相差很大，但对同一种商品，这种差异应该很小，故应据统计数据建立一个共用于各个季度的差分方程。为此将季度编号为 $t=1,2,\dots,20$ ，令

$$y_t = a_1 y_{t-4} + a_2 \quad \text{或} \quad y_t = a_1 y_{t-4} + a_2 y_{t-8} + a_3$$

等，利用全体数据拟合，可得到最好的系数。以二阶差分方程为例，求 a_1, a_2, a_3 ，使得下式最小：

$$Q(a_1, a_2, a_3) = \sum_{i=9}^{20} [y_t - (a_1 y_{t-4} + a_2 y_{t-8} + a_3)]^2$$

Matlab 程序

```
y0=[11 16 25 12 12 18 26 14 13 20 27 15 15 24 30 15 16 25 32 17]';  
y=y0(9:20);  
x=[y0(5:16),y0(1:12),ones(12,1)];  
z=x\y
```

求得 $a_1 = z(1) = 0.8737, a_2 = z(2) = 0.1941, a_3 = z(3) = 0.6957$

得二阶差分方程

$$y_t = 0.8737 \cdot y_{t-4} + 0.1941 \cdot y_{t-8} + 0.6957, (t \geq 21)$$

据此可求得第六和第七年第一季度销售量的预测值为

$$y_{21} = 17.5869, y_{25} = 19.1676$$

例5 (投资额与国民生产总值和物价指数问题) 建立投资额模型，研究某地区实际投资额与国民生产总值 (**GNP**) 及物价指数 (**PI**) 的关系，根据对未来**GNP**及**PI**的估计，预测未来投资额.

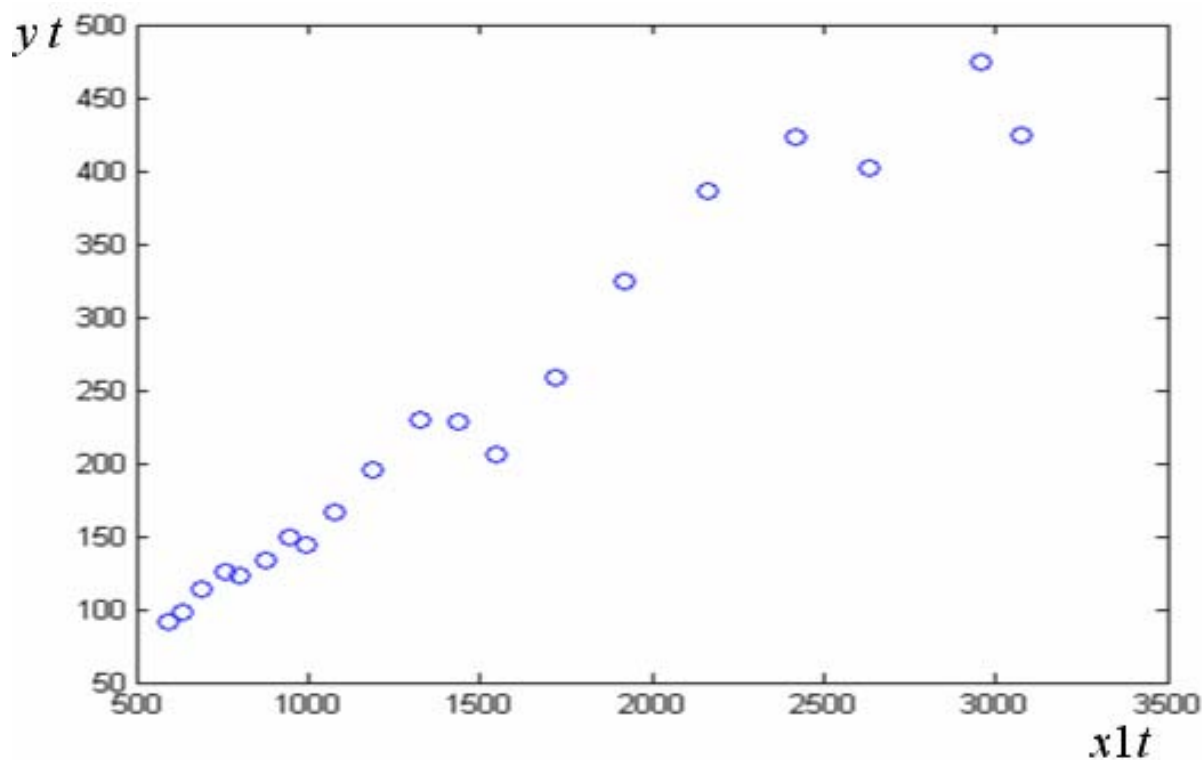
该地区连续**20**年的统计数据

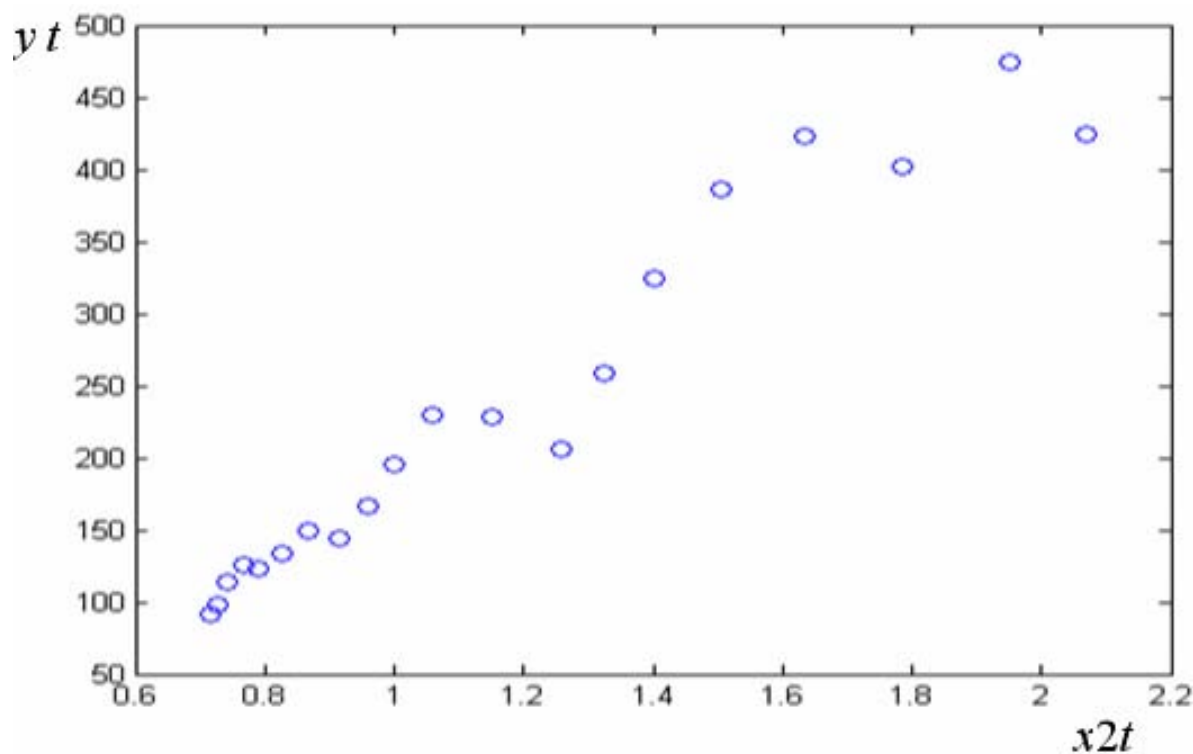
年份 序号	投资额	国民生 产总值	物价 指数	年份 序号	投资额	国民生 产总值	物价 指数
1	90.9	596.7	0.7167	11	229.8	1326.4	1.0575
2	97.4	637.7	0.7277	12	228.7	1434.2	1.1508
3	113.5	691.1	0.7436	13	206.1	1549.2	1.2579
4	125.7	756.0	0.7676	14	257.9	1718.0	1.3234
5	122.8	799.0	0.7906	15	324.1	1918.3	1.4005
6	133.3	873.4	0.8254	16	386.6	2163.9	1.5042
7	149.3	944.0	0.8679	17	423.0	2417.8	1.6342
8	144.2	992.7	0.9145	18	401.9	2631.7	1.7842
9	166.4	1077.6	0.9601	19	474.9	2954.7	1.9514
10	195.0	1185.9	1.0000	20	424.5	3073.0	2.0688

分析:许多经济数据在时间上有一定的滞后性。以时间为序的数据，称为时间序列，时间序列中同一变量的顺序观测值之间存在自相关，若采用普通回归模型直接处理将会出现不良后果，需诊断并消除数据的自相关性，建立新的模型。

基本回归模型

t ~ 年份, y_t ~ 投资额, $x1_t$ ~ GNP, $x2_t$ ~ 物价指数





投资额与 **GNP**及物价指数间均有很强的线性关系

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t$$

$\beta_0, \beta_1, \beta_2$ 为回归系数, $\varepsilon_t \sim i.i.d.N(0, \sigma^2)$ 是随机误差

基本回归模型的结果与分析

参数	参数估计值	置信区间
β_0	322.7250	[224.3386 421.1114]
β_1	0.6185	[0.4773 0.7596]
β_2	-859.4790	[-1121.4757 -597.4823]
$R^2 = 0.9908 \quad F = 919.8529 \quad p = 0.0000$		

$$\hat{y}_t = 322.725 + 0.6185x_{1t} - 859.479x_{2t}$$

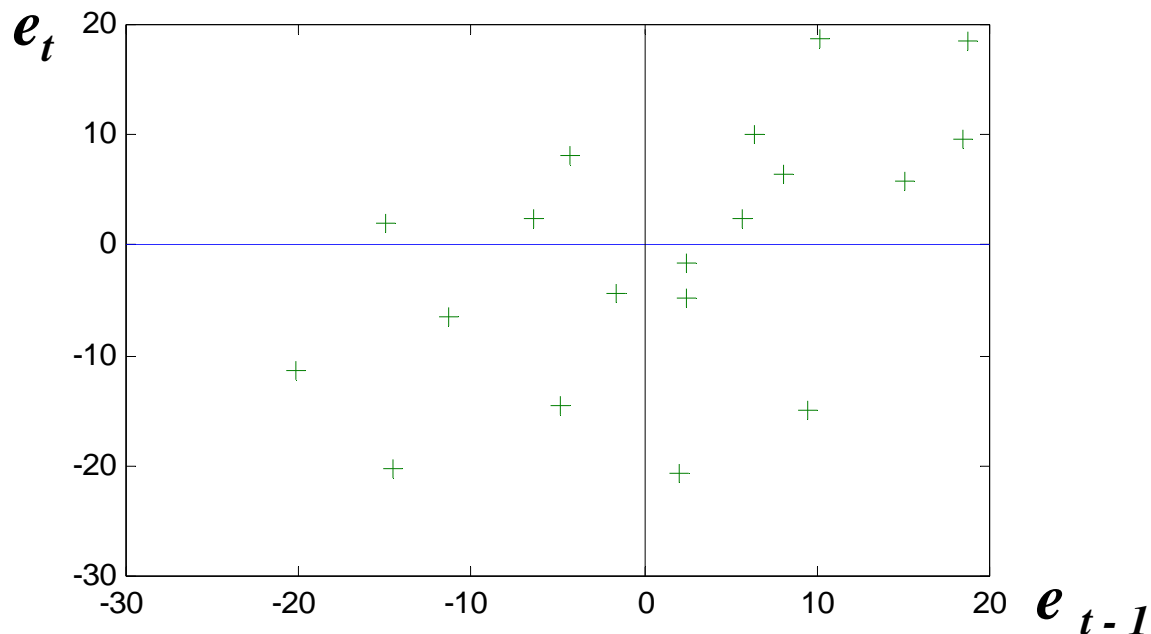
模型优点： **$R^2=0.9908$** ，拟合度高；剩余标准差 **$s=12.7164$**

模型缺点：没有考虑时间序列数据的滞后性影响，可能忽视了随机误差存在自相关；如果存在自相关性，用此模型会有不良后果.

下面进行自相关性的定性诊断---采用残差诊断法

模型残差 $e_t = y_t - \hat{y}_t$ 其中 e_t 为随机误差 ε_t 的估计值.

在MATLAB
工作区中输
出



判断依据:大部分点落在第1, 3象限: ε_t 存在正自相关;

大部分点落在第2, 4象限: ε_t 存在负自相关。

本例判断: 基本回归模型的随机误差项 ε_t 存在正自相关。

自回归性的定量诊断---DW检验

自回归模型: $y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t$, $\varepsilon_t = \rho \varepsilon_{t-1} + u_t$

$\beta_0, \beta_1, \beta_2$ 为回归系数 ρ ($|\rho| \leq 1$) 表示自相关系数

$\rho = 0$: 不相关; $\rho > 0$: 正自相关; $\rho < 0$: 负自相关

用DW统计量估计 ρ , 用广义差分法消除自相关性

DW统计量与DW检验

$$\hat{\rho} = \sum_{t=2}^n e_t e_{t-1} / \sum_{t=2}^n e_t^2$$

$$DW = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=2}^n e_t^2} \underset{n \text{ 较大}}{\approx} 2 \left[1 - \frac{\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=2}^n e_t^2} \right] = 2(1 - \hat{\rho})$$

$$-1 \leq \hat{\rho} \leq 1 \Rightarrow 0 \leq DW \leq 4 \quad ; \quad \hat{\rho} = 1 \Rightarrow DW = 0$$

$$\hat{\rho} = -1 \Rightarrow DW = 4 \quad ; \quad \hat{\rho} = 0 \Rightarrow DW = 2$$

由此可知随机误差不相关。由检验水平、样本容量、回归变量数目和 DW 分布表得检验临界值 d_L 和 d_U ，由 DW 值的大小确定自相关性

0	d_L	d_U	2	$4-d_U$	$4-d_L$	4
正自相关	不能确定	无自相关	不能确定	负自相关		

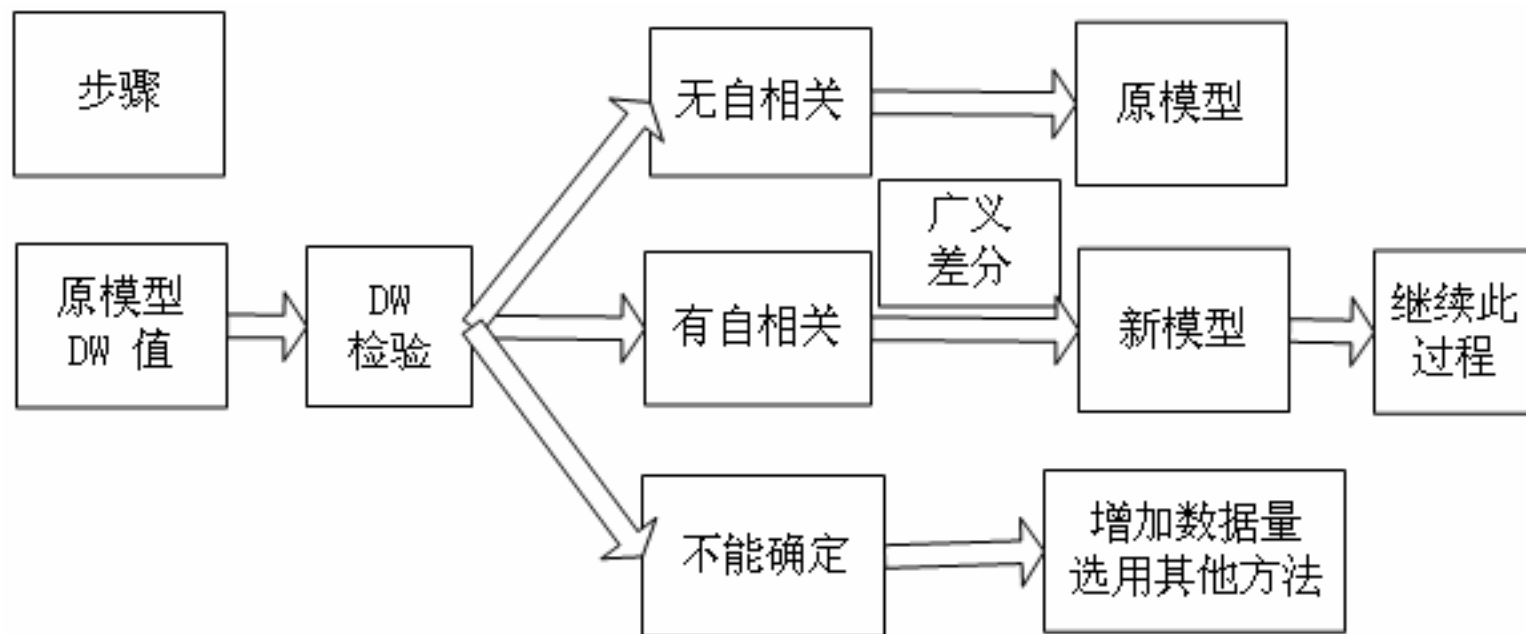
广义差分变换 $DW = 2(1 - \hat{\rho}) \Rightarrow \hat{\rho} = 1 - \frac{DW}{2}$

原模型 $y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t, \quad \varepsilon_t = \rho \varepsilon_{t-1} + u_t$

变换 $y_t^* = y_t - \rho y_{t-1}, \quad x_{it}^* = x_{it} - \rho x_{i,t-1}, \quad i = 1, 2$

新模型 $y_t^* = \beta_0^* + \beta_1 x_{1t}^* + \beta_2 x_{2t}^* + u_t, \quad \beta_0^* = \beta_0(1 - \rho)$

以 β_0^* , β_1 , β_2 为回归系数的普通回归模型



投资额新模型的建立

$$DW = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=2}^n e_t^2}$$

样本容量 $n = 20$ ，回归变量数目 $k = 3$ ， $\alpha = 0.05$ ，查表得临界值 $d_L = 1.10, d_U = 1.54$ ，原模型残差 e_t ，

$DW_{old} = 0.8754 < 1.10 = d_L$ 知原模型有正自相关

$$\hat{\rho} = 1 - DW / 2 = 0.5623$$

作变换 $y_t^* = y_t - 0.5623y_{t-1}$ ， $x_{it}^* = x_{it} - 0.5623x_{i,t-1}$ ， $i = 1, 2$

$$y_t^* = \beta_0^* + \beta_1 x_{1t}^* + \beta_2 x_{2t}^* + u_t$$

由数据 $y_t^*, x_{1t}^*, x_{2t}^*$ 估计系数 $\beta_0^*, \beta_1, \beta_2$

参数	参数估计值	置信区间
β_0^*	163.4905	[1265.4592 , 2005.2178]
β_1	0.6990	[0.5751 , 0.8247]
β_2	-1009.0333	[-1235.9392 , -782.1274]
$R^2 = 0.9772$, $F = 342.8988$, $p = 0.0000$		

总体效果良好。

剩余标准差 $s_{new} = 9.8277 < s_{old} = 12.7164$

新模型的自相关性检验

计算新模型残差 e_t ，得 $DW_{new} = 1.5751$ ，

样本容量 $n = 19$ ，回归变量数目 $k = 3$ ， $\alpha = 0.05$ ，

查临界值表 $d_L = 1.08$ ， $d_U = 1.53$ ，

$d_U < DW_{new} < 4 - d_U$ ，故知新模型无自相关性

新模型 $\hat{y}_t^* = 163.4905 + 0.699x_{1t}^* - 1009.033x_{2t}^*$

还原为原始变量得一阶自回归模型

$$\hat{y}_t = 163.4905 + 0.5623y_{t-1} + 0.699x_{1t} - 0.3930x_{1,t-1} - 1009.0333x_{2t} + 567.3794x_{2,t-1}$$

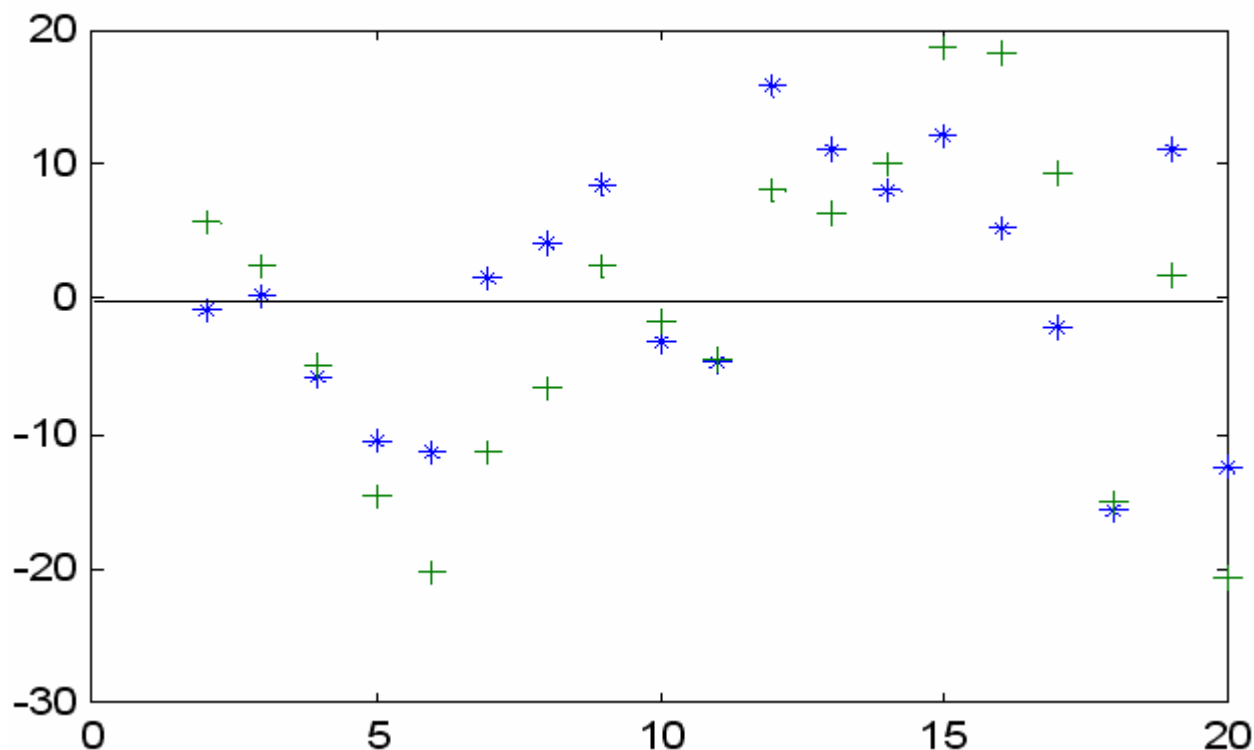
模型结果比较

基本回归模型 $\hat{y}_t = 322.725 + 0.6185x_{1t} - 859.479x_{2t}$

一阶自回归模型

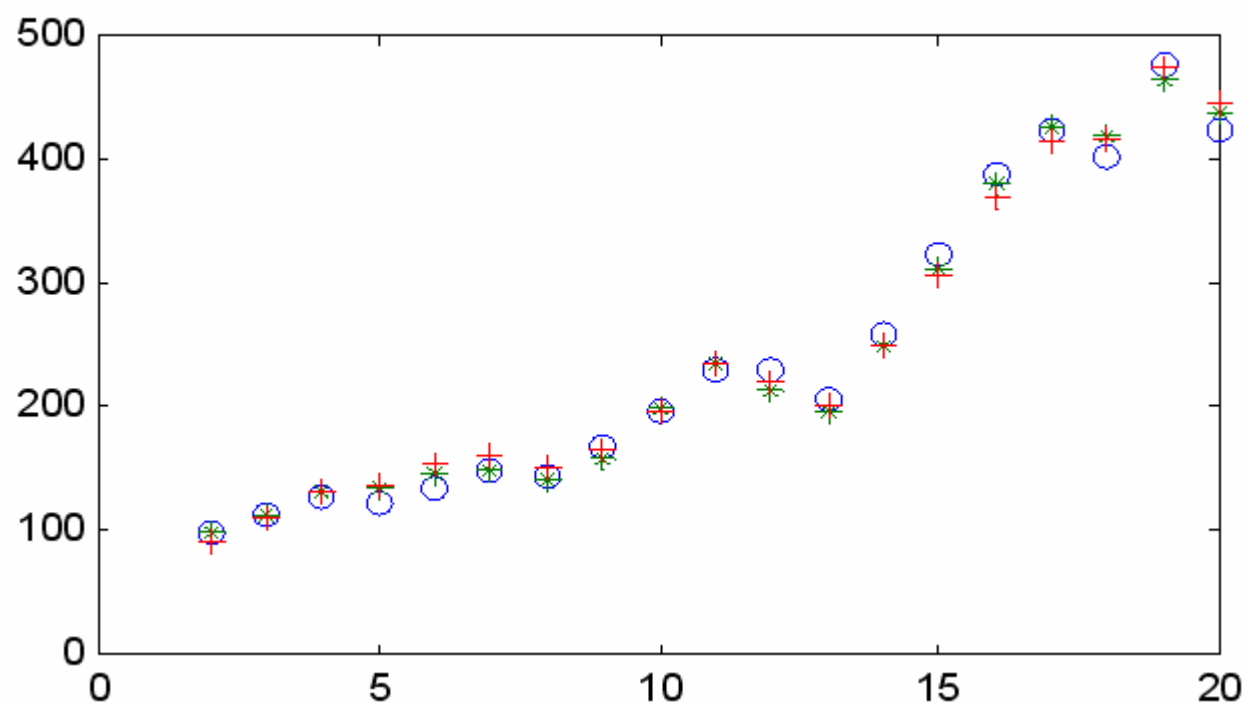
$$\hat{y}_t = 163.4905 + 0.5623y_{t-1} + 0.699x_{1t} - 0.3930x_{1,t-1} - 1009.0333x_{2t} + 567.3794x_{2,t-1}$$

残差图比较 (新模型 $e_t \sim *$, 原模型 $e_t \sim +$)



一阶自回归
模型残差 e_t 比
基本回归模
型要小

拟合图比较 (新模型 $\hat{y}_t \sim *$, 原模型 $\hat{y}_t \sim +$)



投资额预测 对未来投资 y_t 作预测，需先估计出未来的国民生产总值 x_{1t} 和物价指数 x_{2t}

年份 序号	投资额	国民生 产总值	物价 指数	年份 序号	投资额	国民生 产总值	物价 指数
1	90.9	596.7	0.7167	18	401.9	2631.7	1.7842
2	97.4	637.7	0.7277	19	474.9	2954.7	1.9514
3	113.5	691.1	0.7436	20	424.5	3073.0	2.0688

设已知 $t=21$ 时， $x_{1t} = 3312$ ， $x_{2t} = 2.1983$ ，

基本回归模型 $\hat{y}_t = 485.6720$

一阶自回归模型 $\hat{y}_t = 469.7638$

\hat{y}_t 较小是由于 $y_{t-1} = 424.5$ 过小所致.

§ 2 * 平稳时间序列模型

这里平稳是指宽平稳，其特性是序列的统计特性不随时间平移而变化，即均值和协方差不随时间的平移而变化。

下面自回归模型（**Auto Regressive Model**）简称**AR**模型，移动平均模型（**Moving Average Model**）简称**MA**模型，自回归移动平均模型（**Auto Regressive Moving Average Model**）简称**ARMA**模型。下面的 X_t 为零均值（即中心化处理的）平稳序列。

(1) 一般自回归模型 $AR(n)$

假设时间序列 X_t 仅与 $X_{t-1}, X_{t-2}, \dots, X_{t-n}$ 有线性关系，而在 $X_{t-1}, X_{t-2}, \dots, X_{t-n}$ 已知条件下， X_t 与 $X_{t-j} (j = n+1, n+2, \dots)$ 无关， a_t 是一个独立于 $X_{t-1}, X_{t-2}, \dots, X_{t-n}$ 的白噪声序列，

$$a_t \sim N(0, \sigma^2) \quad X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \dots + \varphi_n X_{t-n} + a_t$$

$$\text{或者 } a_t = X_t - \varphi_1 X_{t-1} - \varphi_2 X_{t-2} - \dots - \varphi_n X_{t-n}$$

可见 $AR(n)$ 系统的响应 X_t 具有 n 阶动态性。 $AR(n)$ 模型通过把 X_t 中的依赖于 $X_{t-1}, X_{t-2}, \dots, X_{t-n}$ 的部分消除掉后，使得具有 n 阶动态性的序列 X_t 转化为独立的序列 a_t 。因此拟合 $AR(n)$ 模型的过程也就是使相关序列独立化的过程。

(2) 移动平均模型 $MA(m)$

如果一个系统在 t 时刻的响应 X_t ，与其以前时刻 $t-1, t-2, \dots$ 的响应 X_{t-1}, X_{t-2}, \dots 无关，而与其以前时刻 $t-1, t-2, \dots, t-m$ 进入系统的扰动 $a_{t-1}, a_{t-2}, \dots, a_{t-m}$ 存在着一定的相关关系，那么这一类系统为 $MA(m)$ 系统。

$$X_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots - \theta_m a_{t-m}$$

(3) 自回归移动平均模型

一个系统，如果它在时刻 t 的响应 X_t ，不仅与其以前时刻的自身值有关，而且还与其以前时刻进入系统的扰动存在一定的依存关系，那么，这个系统就是自回归移动平均系统。

ARMA(n, m)模型：

$$X_t - \varphi_1 X_{t-1} - \cdots - \varphi_n X_{t-n} = a_t - \theta_1 a_{t-1} - \cdots - \theta_m a_{t-m}$$

对于平稳系统来说，由于AR、MA、ARMA(n, m)模型都是ARMA($n, n-1$)模型的特例，我们以ARMA($n, n-1$)模型为一般形式来建立时序模型。

§ 3* ARMA模型的特性

格林函数就是描述系统记忆扰动程度的函数.

3.1 AR(1)系统的格林函数

$$AR(1): X_t - \varphi_1 X_{t-1} = a_t$$

$$\Rightarrow X_t = \sum_{j=0}^{\infty} \varphi_1^j a_{t-j} = \sum_{j=0}^{\infty} G_j a_{t-j} = \sum_{k=-\infty}^t G_{t-k} a_k, (G_0 = \varphi_1^0 = 1)$$

记忆函数(也叫格林函数) $G_j = \varphi_1^j$

格林函数的意义

- (1) G_j 是前 j 个时间单位以前进入系统的扰动 a_{t-j} 对系统现在行为(响应)影响的权数。
- (2) G_j 客观地刻画了系统动态响应衰减的快慢程度。
- (3) 对于一个平稳系统来说, 在某一时刻由于受到进入系统的扰动 a_t 的作用, 离开其平衡位置(即平均数—零), G_j 描述系统回到平衡位置的速度, φ_1 的值较小, 速度较快; φ_1 的值较大, 回复速度就较慢

3.2 ARMA(2,1)系统的格林函数

ARMA(2,1)模型 $X_t - \varphi_1 X_{t-1} - \varphi_2 X_{t-2} = a_t - \theta_1 a_{t-1}$

解为 $X_t = \sum_{j=0}^{\infty} G_j a_{t-j}$

格林函数的隐式

$$G_0 = 1, G_1 = \varphi_1 - \theta_1, G_j = \varphi_1 G_{j-1} + \varphi_2 G_{j-2}, (j = 3, 4, \dots)$$

格林函数的显式 $G_j = \frac{\lambda_1 - \theta_1}{\lambda_1 - \lambda_2} \lambda_1^j + \frac{\lambda_2 - \theta_1}{\lambda_2 - \lambda_1} \lambda_2^j$

$$\lambda_1 = \frac{\varphi_1 + \sqrt{\varphi_1^2 + 4\varphi_2}}{2}, \lambda_2 = \frac{\varphi_1 - \sqrt{\varphi_1^2 + 4\varphi_2}}{2}$$

是特征方程 $\lambda^2 - \varphi_1 \lambda - \varphi_2 = 0$ 的特征根

3.3 逆函数

X_t 的“逆转形式”是一个无穷阶的自回归模型:

$$X_t = \sum_{j=1}^{\infty} I_j X_{t-j} + a_t$$

系数函数 $I_j (I_0 = 1)$ 称为逆函数

对于**AR (2) 模型** $X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + a_t$

$$I_1 = \varphi_1, I_2 = \varphi_2, I_j = 0, j = 3, 4, \dots$$

对于**MA(1)模型** $X_t = \sum_{j=1}^{\infty} (-\theta_1^j X_{t-j}) + a_t$

$$I_j = -\theta_1^j, |\theta_1| < 1$$

§ 4 时间序列建模的基本步骤

1. 数据的预处理：数据的剔取及提取趋势项；
2. 取 $n=1$ ，拟合 $ARMA(2,1)$ 模型
 - (1) 拟合 $AR(3)$ 模型用最小二乘法拟合出系数。
 - (2) 估计 $ARMA(2,1)$ 模型 参数的初始值。
 - (3) 以(2)中得到的为初始值，利用非线性最小二乘法得到的终值及置信区间，并且求出残差平方和(RSS)；
3. $n=n+1$ ，拟合 $ARMA(2n,2n-1)$ 模型：其基本步骤与2类似；
4. 用 F 准则检验模型的适用性。若 F 检验显著，则转入第2步；若 F 检验不显著，转入第5步；

对于ARMA模型的适用性检验的实际就是对 a_t 的独立性检验。检验 a_t 的独立性的一个简便而有效的办法是拟合更高阶的模型。若更高阶模型的残差平方和有明显减少，就意味着现有模型的 a_t 不是独立的，因而模型不适用；若更高阶模型的残差平方和没有明显减少，同时更高阶模型中的附加参数的值也很小（其置信区间包含0），则可认为该模型是适用的。具体的检验准则如下：

设有模型 $ARMA(n_1, m_1)$ 和 $ARMA(n_2, m_2)$, $n_2 > n_1, m_2 > m_1$

设 $A_0 = ARMA(n_1, m_1)$ 模型残差 a_t 的平方和，

$A_1 = ARMA(n_2, m_2)$ 模型残差 a_t 的平方和。

N 是采集数据的数目，则检验准则为 $F = \frac{A_1 - A_0}{s} \bigg/ \frac{A_0}{N - \gamma} \sim F(s, N - \gamma)$

其中 $\gamma = n_2 + m_2, s = n_2 + m_2 - (n_1 + m_1)$

若这样得到的 F 值超过由 F 分布查表所得的在 **5 %** 置信水平上的 $F(s, N-r)$ 值, 那么由 $ARMA(n_1, m_1)$ 模型改变为 $ARMA(n_2, m_2)$ 时, 残差平方和的改善是显著的, 因而拒绝关于模型 $ARMA(n_1, m_1)$ 的适用性假设; F 值低于查表所得之值, 就可以认为在该置信水平上这个模型是适用的.

- 5. 检查 $\varphi_{2n}, \theta_{2n-1}$ 的值是否很小, 其置信区间是否包含零。若不是, 则适用的模型就是 $ARMA(2n, 2n-1)$**
- 若 $\varphi_{2n}, \theta_{2n-1}$ 很小, 且其置信区间包含零,
则拟合 $ARMA(2n-1, 2n-2)$

6. 利用 F 准则检验模型 $ARMA(2n,2n-1)$ 和 $ARMA(2n-1,2n-2)$, 若 F 值不显著, 转入第7步; 若 F 值显著, 转入第8步;
7. 舍弃小的 MA 参数, 拟合 $m < 2n-2$ 的模型 $ARMA(2n-1,m)$, 并用 F 准则进行检验。重复这一过程, 直到得出具有最小参数的适用模型为止;
8. 舍弃小的 MA 参数, 拟合 $m < 2n-1$ 的模型 $ARMA(2n,m)$, 并用 F 准则进行检验。重复这一过程, 直到得出具有最小参数的适用模型为止.

第7章 主成分分析及应用

主成分分析是一种多变量分析方法，被誉为质量管理的新工具之一，也称为矩阵数据分析法。**基本思想是**:用较少几个不相关的变化量，能综合原始多个变量的绝大部分信息，即通过变量替换方法把原来相关的变量变为不相关的若干新变量，这对分析数据带来很大方便。

主成分分析是通过对一组变量的几个线性组合来解释这组变量的方差和协方差结构，以达到数据的压缩和数据的解释的目的。

主成分分析试图在力保数据信息丢失最少的原则下，对这种多变量的截面数据表进行最佳综合简化，也就是说，对高维变量空间进行降维处理。

很显然，识辨系统在一个低维空间要比在一个高维空间容易得多。

在力求数据信息丢失最少的原则下，对高维的变量空间降维，即研究指标体系的少数几个线性组合，并且这几个线性组合所构成的综合指标将尽可能多地保留原来指标变异方面的信息。这些综合指标就称为主成分。要讨论的问题是：

(1) 基于相关系数矩阵还是基于协方差矩阵做主成分分析。当分析中所选择的经济变量具有不同的量纲，变量水平差异很大，应该选择基于相关系数矩阵的主成分分析。

(2) 选择几个主成分。主成分分析的目的是简化变量，一般情况下主成分的个数应该小于原始变量的个数。关于保留几个主成分，应该权衡主成分个数和保留的信息。

(3) 如何解释主成分所包含的经济意义。

例1： 我们知道生产服装有很多指标，比如袖长、肩宽、身高等十几个指标，服装厂生产时，不可能按照这么多指标来做，怎么办？一般情况，生产者考虑几个综合的指标，象标准体形、特形等。

例2： 企业经济效益的评价，它涉及到很多指标。例百元固定资产原值实现产值、百元固定资产原值实现利税，百元资金实现利税，百元工业总产值实现利税，百元销售收入实现利税，每吨标准煤实现工业产值，每千瓦时电力实现工业产值，全员劳动生产率，百元流动资金实现产值等，我们要找出综合指标，来评价企业的效益。

第一节 主成分分析的基本思想

在经济问题的研究中，我们常常会遇到影响此问题的很多变量，这些变量多且又有一定的相关性，因此我们希望从中综合出一些主要的指标，这些指标所包含的信息量又很多。这些特点，使我们在研究复杂的问题时，容易抓住主要矛盾。

那么怎样找综合指标？

若有一些指标 X_1, \dots, X_p ，取综合指标即它们的线性组合 F ，当然有很多，我们希望线性组合 F 包含很多的信息，即 $Var(F)$ 最大，这样得到 F 记为 F_1 ，然后再找 F_2 ， F_1 与 F_2 无关，以此类推，我们找到了一组综合变量 F_1, \dots, F_m ，这组变量基本包含了原来变量的所有信息且相互独立。

第二节 主成分分析的数学模型及几何解释

1. 数学模型

设样本资料阵为

样本资料阵为

$$x = (x_1, \dots, x_p) = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}$$

综合指标为

[illegible]

简写为

简写为 $F_i = a_{1i}x_1 + a_{2i}x_2 + \cdots + a_{pi}x_p, 1 \leq i \leq m$

要求

(1) 每个主成分的系数平方和为1。即

$$a_{1i}^2 + a_{2i}^2 + \cdots + a_{pi}^2 = 1$$

(2) F_i, F_j 不相关 $Cov(F_i, F_j) = 0, (i \neq j, i, j = 1, \cdots, m)$

(3)主成分的方差依次递减，重要性依次递减，即

$$Var(F_1) \geq Var(F_2) \geq \cdots \geq Var(F_m)$$

F_1 是 X_1, \dots, X_p 的线性函数中方差最大。

依次类推.....

2. 主成分的几何意义

设有 n 个样品，每个样品有两个观测变量 X_1 ， X_2 二维平面的散点图。 n 个样本点，无论沿着 X_1 轴方向还是 X_2 轴方向，都有较大的离散性，其离散程度可以用 X_1 或 X_2 的方差表示。

当只考虑一个时，原始数据中的信息将会有较大的损失。若将坐标轴旋转一下。

$$\begin{cases} F_1 = X_1 \cos \theta + X_2 \sin \theta \\ F_2 = -X_1 \sin \theta + X_2 \cos \theta \end{cases}$$

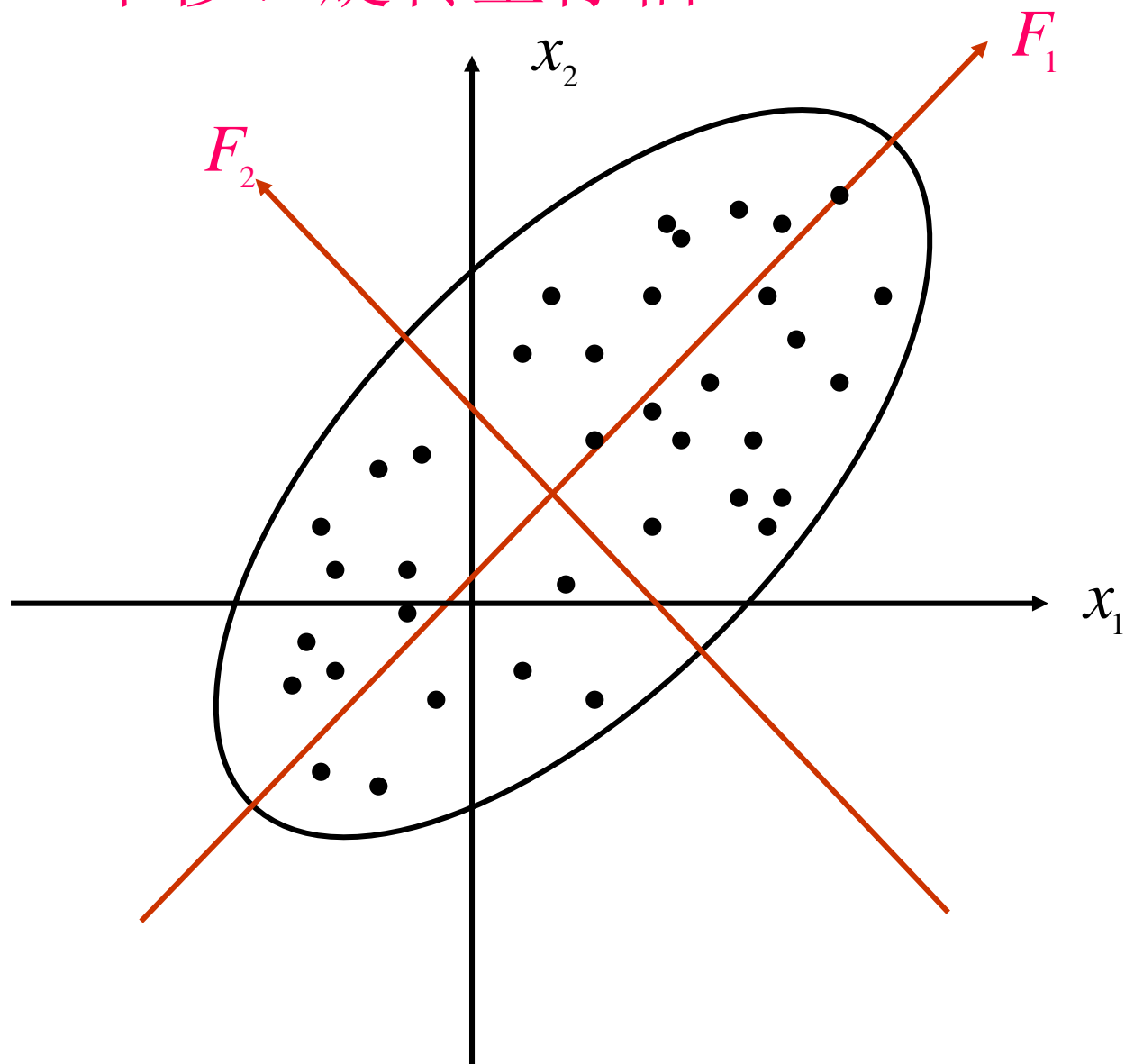
$$\text{即} \quad \begin{pmatrix} F_1 \\ F_2 \end{pmatrix} = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = UX$$

且 U 是正交矩阵, 即 $U'U = I$

则 n 个样品在 F_1 轴的离散程度最大(方差最大), 变量 F_1 代表了原始数据的绝大部分信息, 即使不考虑 F_2 , 信息损失也不多。而且 F_1 , F_2 不相关。只考虑 F_1 时, 二维降为一维。

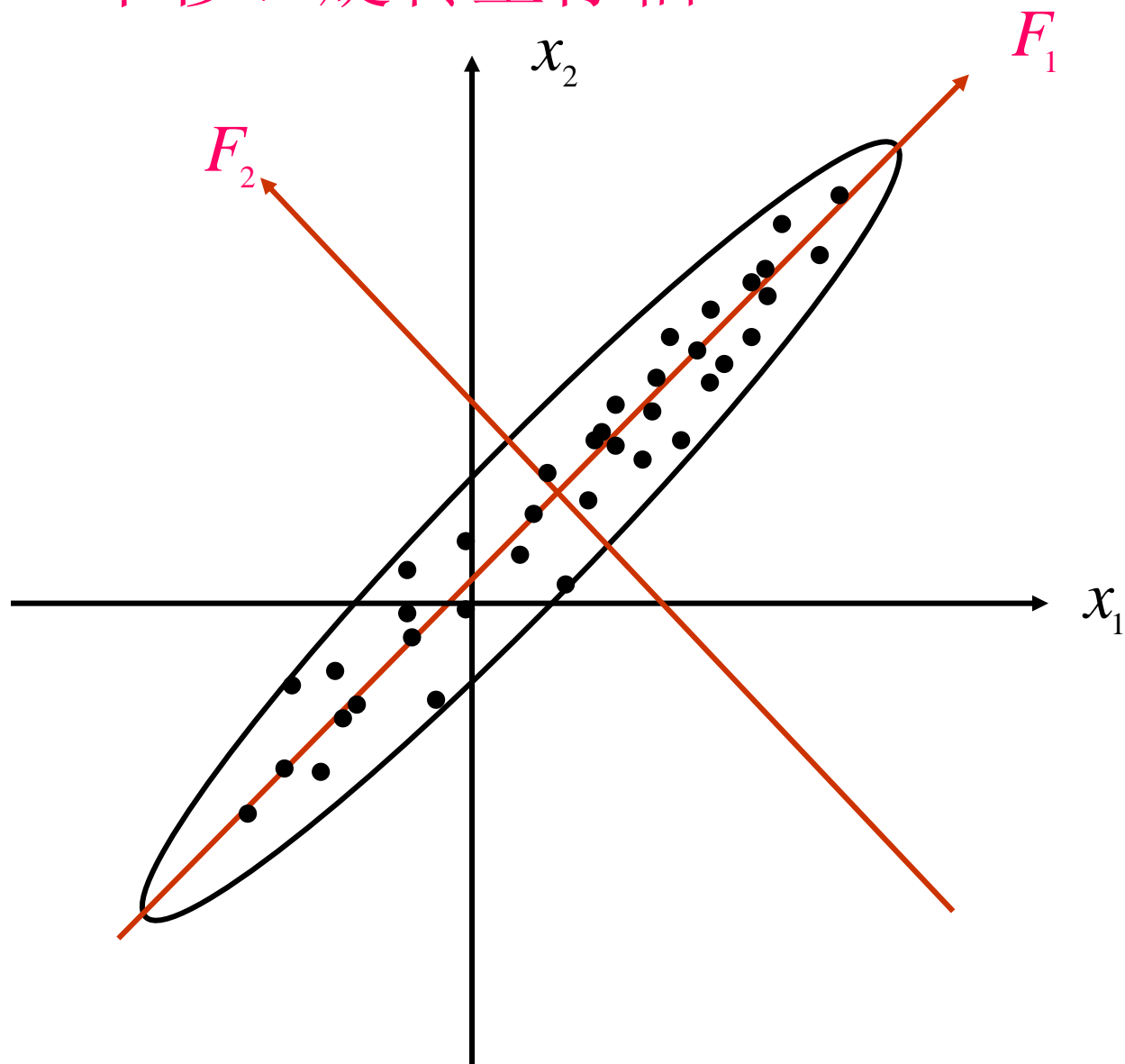
平移、旋转坐标轴

主成分分析的几何解释



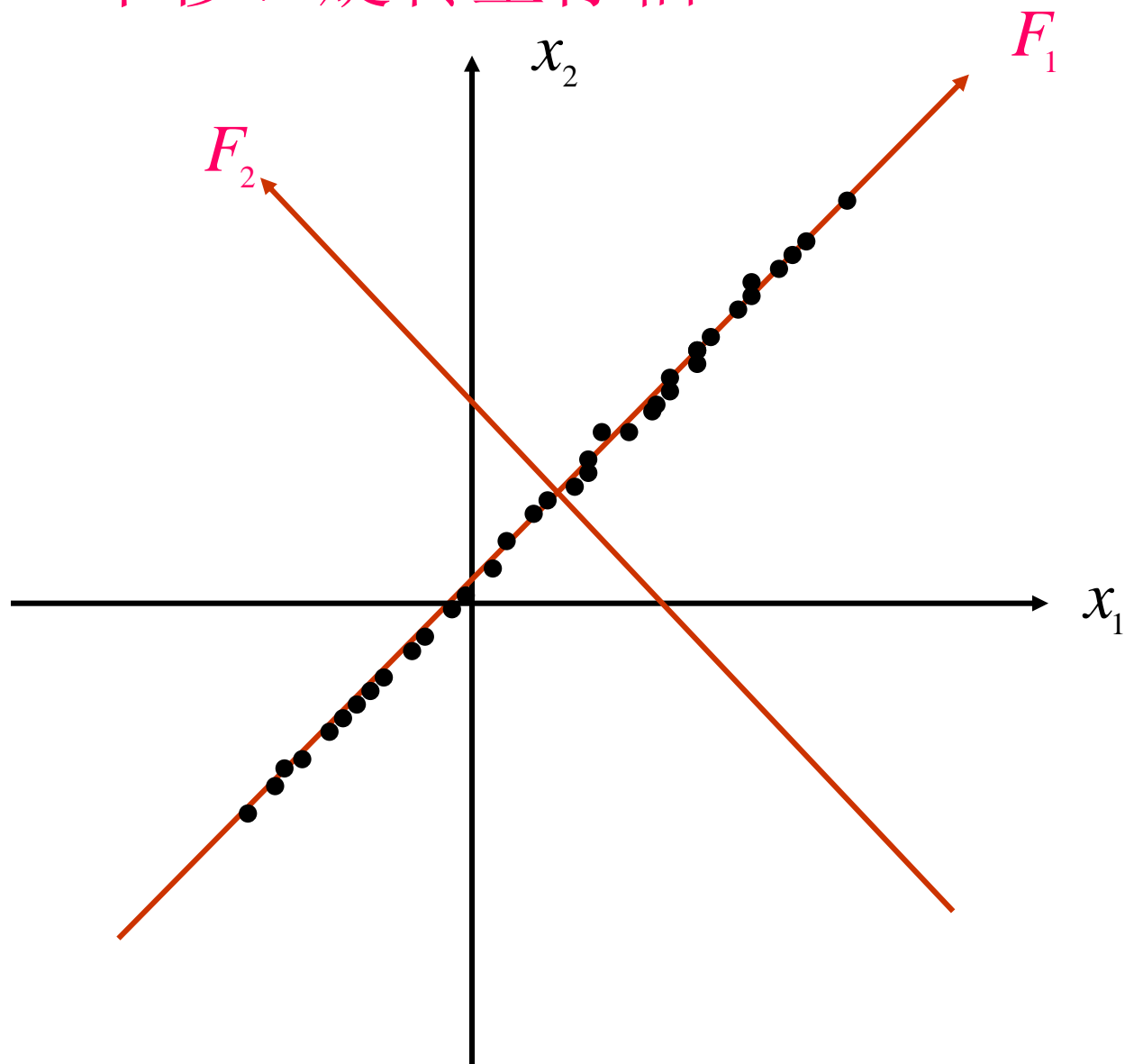
平移、旋转坐标轴

主成分分析的几何解释



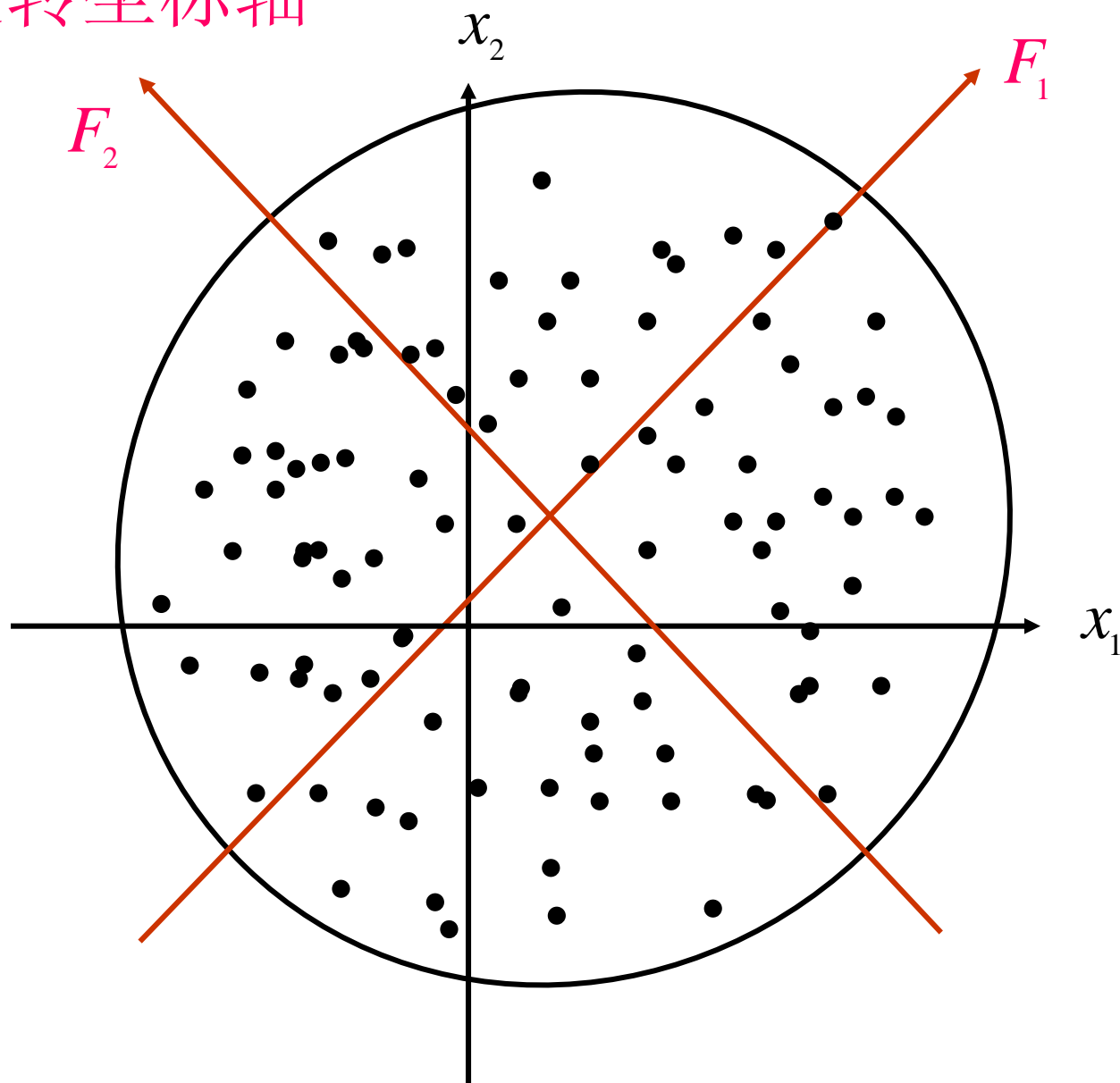
平移、旋转坐标轴

主成分分析的几何解释



平移、旋转坐标轴

主成分分析的几何解释



旋转变换的目的是为了使得 n 个样品点在 F_1 轴方向上的离散程度最大，即 F_1 的方差最大。变量 F_1 代表了原始数据的绝大部分信息，在研究某经济问题时，即使不考虑变量 F_2 也无损大局。经过上述旋转变换原始数据的大部分信息集中到 F_1 轴上，对数据中包含的信息起到了浓缩作用。

F_1 , F_2 除了可以对包含在 X_1 , X_2 中的信息起着浓缩作用之外, 还具有不相关的性质, 这就使得在研究复杂的问题时避免了信息重叠所带来的虚假性。二维平面上的个点的方差大部分都归结在 F_1 轴上, 而 F_2 轴上的方差很小。 F_1 和 F_2 称为原始变量 x_1 和 x_2 的综合变量。 F_1 简化了系统结构, 抓住了主要矛盾。

第三节 主成分的推导及性质

定理1 若 A 是 P 阶实对称阵，则一定可以找到正交阵

$$U, \text{ 使 } U^{-1}AU = \text{diag}(\lambda_1, \dots, \lambda_p)$$

其中 $\lambda_1, \dots, \lambda_p$ 是 A 的特征根

定理2 若上述 A 的特征根所对应的单位特征向量为

$$\mu_1, \dots, \mu_p, U \triangleq (\mu_1, \dots, \mu_p) = (\mu_{ij})_{p \times p}$$

则不同特征根对应的特征向量正交，即

$$\mu_i' \mu_j = 0$$

1. 主成分的推导

记 $X = (x_1, \dots, x_p)$, 设 $F = a_1x_1 + \dots + a_px_p = a'X$

找到系数 a ($a'a = 1$), 使 $Var(F)$ 最大, 即

$Var(F) = Var(a'X) = a'Var(X)a = a'\Sigma a$ 最大

$$\Sigma_x = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix}$$

定理3 设 Σ 的特征根为 $\lambda_1 \geq \cdots \geq \lambda_p > 0$

对应的标准正交基为 μ_1, \cdots, μ_p

则 $Var(F) = \lambda_1$, $a = \mu_1$, 记 $F_1 = \mu_1' x$, F_1 与 F_2 无关

$F_2 = \mu_2' x$ 余此类推, 称 F_1 为第一主成分,

F_2 为第二主成分,, F_p 为第 p 主成分

写为矩阵形式: $F = U'X$

$$U = (\mu_1, \dots, \mu_p) = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix}$$

$$X = (X_1, X_2, \dots, X_p)'$$

精度分析

- 1) 贡献率：第 i 个主成分的方差在全部方差中所占比重 $\lambda_i / \sum_{i=1}^p \lambda_i$ ，称为**贡献率**，反映了原来 P 个指标多大的信息，有多大的综合能力。
- 2) 累积贡献率：前 m 个主成分共有多大的综合能力，用这 m 个主成分的方差和在全部方差中所占比重 $\sum_{i=1}^m \lambda_i / \sum_{i=1}^p \lambda_i$ 来描述，称为**累积贡献率**。

进行主成分分析的目的之一是希望用尽可能少的主成分 F_1, F_2, \dots, F_m ($m \leq p$) 代替原来的 P 个指标。到底应该选择多少个主成分，在实际工作中，主成分个数的多少取决于能够反映原来变量80%以上的信息量为依据，即当累积贡献率 $\geq 80\%$ 时的主成分的个数就足够了。最常见的情况是主成分为2到3个。

2. 样本主成分的导出

设有 n 个样品， P 个指标，得到的原始资料矩阵为

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

$$\text{记 } S_{ij} = \frac{1}{n-1} \sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_i)', 1 \leq i, j \leq p$$

$$\bar{x}_i = \frac{1}{n} \sum_{k=1}^n x_{ki}, (1 \leq i \leq p) \quad , \quad r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}} \sqrt{S_{jj}}}$$

$$S = (S_{ij})_{p \times p}, R = (r_{ij})_{p \times p}$$

称 S 为**样本协方差**，是总体协方差 Σ 的无偏估计；

称 R 为**样本相关矩阵**，是总体相关矩阵的估计。

为避免指标差异和量纲的不同，用 R 代替 Σ

3. 主成分的性质

1) 均值 $E(U' X) = U' \mu$

2) 方差为所有特征根之和

$$\sum_{i=1}^p \text{Var}(F_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_p^2$$

说明主成分分析把 P 个随机变量的总方差分解成为 P 个不相关的随机变量的方差之和。

协方差矩阵 Σ 的对角线上的元素之和等于特征根之和。

例1 应收账款是指企业因对外销售产品、材料、提供劳务及其它原因，应向购货单位或接受劳务的单位收取的款项，包括应收销货款、其它应收款和应收票据等。出于扩大销售的竞争需要，企业不得不以赊销或其它优惠的方式招揽顾客，由于销售和收款的时间差，于是产生了应收款项。应收款赊销效果的好坏，不仅依赖于企业的信用政策，还依赖于顾客的信用程度。由此，评价顾客的信用等级，了解顾客的综合信用程度，做到“知己知彼，百战不殆”，对加强企业的应收账款管理大有帮助。某企业为了了解其客户的信用程度，采用西方银行信用评估常用的5C方法，5C的目的是说明顾客违约的可能性。

1、**品格**（用 X_1 表示），指顾客的信誉，履行偿还义务的可能性。企业可以通过过去的付款记录得到此项。

2、**能力**（用 X_2 表示），指顾客的偿还能力。即其流动资产的数量和质量以及流动负载的比率。顾客的流动资产越多，其转化为现金支付款项的能力越强。同时，还应注意顾客流动资产的质量，看其是否会出现存货过多过时质量下降，影响其变现能力和支付能力。

3、**资本**（用 X_3 表示），指顾客的财务势力和财务状况，表明顾客可能偿还债务的背景。

4、**附带的担保品**（用 X_4 表示），指借款人以容易出售的资产做抵押。

5、**环境条件**（用 X_5 表示），指企业的外部因素，即指非企业本身能控制或操纵的因素。

首先并抽取了10家具有可比性的同类企业作为样本，
又请8位专家分别给10个企业的5个指标打分，然后分别计算企业5个指标的平均值，如表

企业 指标	1	2	3	4	5	6	7	8	9	10
1	76.5	81.5	76	75.8	71.7	85	79.2	80.3	84.4	76.5
2	70.6	73	67.6	68.1	78.5	94	94	87.5	89.5	92
3	90.7	87.3	91	81.5	80	84.6	66.9	68.8	64.8	66.4
4	77.5	73.6	70.9	69.8	74.8	57.7	60.4	57.4	60.8	65
5	85.6	68.5	70	62.2	76.5	70	69.2	71.7	64.9	68.9

总方差 = 485.31477778

协方差矩阵的特征根

指标	特征根	差值	贡献率	累积贡献率
第1主成分	410.506	367.242	0.845854	0.84585
第2主成分	43.264	22.594	0.089146	0.93500
第3主成分	20.670	12.599	0.042591	0.97759
第4主成分	8.071	5.266	0.016630	0.99422
第5主成分	2.805	.	0.005779	1.00000

特征向量

	第1主成分	第2主成分	第3主成分	第4主成分	第5主成分
X1	0.468814	-.830612	0.021406	0.254654	-.158081
X2	0.484876	0.329916	0.014801	-.287720	-.757000
X3	0.472744	-.021174	-.412719	-.588582	0.509213
X4	0.461747	0.430904	-.240845	0.706283	0.210403
X5	0.329259	0.122930	0.878054	-.084286	0.313677

第一主成份的贡献率为 **84.6%**，第一主成份 $Z_1=0.469X_1+0.485X_2+0.473X_3+0.462X_4+0.329X_5$ 的各项系数大致相等，且均为正数，说明第一主成份对所有的信用评价指标都有近似的载荷，是对所有指标的一个综合测度，可以作为综合的信用等级指标，可以用来排序。将原始数据的值中心化后，代入第一主成份 Z_1 的表示式，计算各企业的得分，并按分值大小排序：

序号	1	2	3	4	5	6	7	8	9	10
得分	3.16	13.6	-9.01	35.9	25.1	-10.3	-4.36	-33.8	-6.41	-13.8
排序	4	3	7	1	2	8	5	10	6	9

在正确评估了顾客的信用等级后，就能正确制定出对其的信用期、收帐政策等，这对于加强应收帐款的管理大有帮助。

例2 基于相关系数矩阵的主成分分析。对美国纽约上市的有关化学产业的三个证券和石油产业的2个证券做了**100**周的收益率调查。下表是其相关系数矩阵。

- 1) 利用相关系数矩阵做主成分分析。
- 2) 决定要保留的主成分个数，并解释意义。

1	0.577	0.509	0.0063	0.0037
0.577	1	0.599	0.389	0.52
0.509	0.599	1	0.436	0.426
0.387	0.389	0.436	1	0.523
0.462	0.322	0.426	0.523	1

Eigenvalues of the Correlation Matrix(相关系数矩阵特征值)

	Eigenvalue	Difference	Proportion	Cumulative
F 1	2.85671	2.04755	0.571342	0.57134
F2	0.80916	0.26949	0.161833	0.73317
F3	0.53968	0.08818	0.107935	0.84111
F4	0.45150	0.10855	0.090300	0.93141
F5	0.34295	.	0.068590	1.00000

Eigenvectors(特征向量)

	F1	F2	F3	F4	F5
X1	0.463605	-.240339	-.611705	0.386635	-.451262
X2	0.457108	-.509305	0.178189	0.206474	0.676223
X3	0.470176	-.260448	0.335056	-.662445	-.400007
X4	0.421459	0.525665	0.540763	0.472006	-.175599
X5	0.421224	0.581970	-.435176	-.382439	0.385024

例3 各地区居民消费情况主成分分析

一、模型假设

- 1、假设构成全国31个省市自治区的消费情况的因数只有：食品、衣着、家庭设备用品及服务、医疗保健和个人用品、交通和通信、娱乐教育文化和居住这七个因素。
- 2、假设各数据有效完整。

原始数据 2001年全国各地区消费情况指数(表1)

地区	食品 X1	衣着 X2	家庭设备用 品及服务X3	医疗保健 和个人用 品X4	交通和 通信X5	娱乐教育 文化X6	居住X7
北 京	101.5	100.4	97.0	98.7	100.8	114.2	104.2
天 津	100.8	93.5	95.9	100.7	106.7	104.3	106.4
河 北	100.8	97.4	98.2	98.2	99.5	103.6	102.4
山 西	99.4	96.0	98.2	97.8	99.1	98.3	104.3
内 蒙 古	101.8	97.7	99.0	98.1	98.4	102.0	103.7
辽 宁	101.8	96.8	96.4	92.7	99.6	101.3	103.4
吉 林	101.3	98.2	99.4	103.7	98.7	101.4	105.3
黑 龙 江	101.9	100.0	98.4	96.9	102.7	100.3	102.3
上 海	100.3	98.9	97.2	97.4	98.1	102.1	102.3
江 苏	99.3	97.7	97.6	101.1	96.8	110.1	100.4

原始数据 2001年全国各地区消费情况指数(表2)

地区	食品 X1	衣着 X2	家庭设备用 品及服务 X3	医疗保健 和个人用 品X4	交通和 通信 X5	娱乐教育 文化X6	居住X7
浙 江	98.7	98.4	97.0	99.6	95.6	107.2	99.8
安 徽	99.7	97.7	98.0	99.3	97.3	104.1	102.7
福 建	97.6	96.5	97.6	102.5	97.2	100.6	99.9
江 西	98.0	98.4	97.1	100.5	101.4	103.0	99.9
山 东	101.1	98.6	98.7	102.4	96.9	108.2	101.7
河 南	100.4	98.6	98.0	100.7	99.4	102.4	103.3
湖 北	99.3	96.9	94.0	98.1	99.7	109.7	99.2
湖 南	98.6	97.4	96.4	99.8	97.4	102.1	100.0
广 东	98.2	98.2	99.4	99.3	99.7	101.5	99.9
广 西	98.5	96.3	97.0	97.7	98.7	112.6	100.4

原始数据

2001年全国各地区消费情况指数(表3)

地区	食品 X1	衣着 X2	家庭设备用品及服务 X3	医疗保健和个人用品X4	交通和通信 X5	娱乐教育文化X6	居住X7
海 南	98.4	99.2	98.1	100.2	98.0	98.2	97.8
重 庆	99.2	97.4	95.7	98.9	102.4	114.8	102.6
四 川	101.3	97.9	99.2	98.8	105.4	111.9	99.9
贵 州	98.5	97.8	94.6	102.4	107.0	115.0	99.5
云 南	98.3	96.3	98.5	106.2	92.5	98.6	101.6
西 藏	99.3	101.1	99.4	100.1	103.6	98.7	101.3
陕 西	99.2	97.3	96.2	99.7	98.2	112.6	100.5
甘 肃	100.0	99.9	98.2	98.3	103.6	123.2	102.8
青 海	102.2	99.4	96.2	98.6	102.4	115.3	101.2
宁 夏	100.1	98.7	97.4	99.8	100.6	112.4	102.5
新 疆	104.3	98.7	100.2	116.1	105.2	101.6	102.6

二、变量设置

- 1、 X 为表示31个省市自治区的各地区的消费情况所列出的数值对应矩阵（为方便起见，设 X 代表的矩阵已对数据作了标准化）。
- 2、 R 为 X 的相关系数矩阵。
- 3、 P 是 R 的特征根。
- 4、 A 是 R 的特征根 P 相应的单位特征向量。
- 5、 F 及其相应的向量是几个主成分

三、模型建立

第一步:建立变量(即观测指标)的相关系数矩阵 R

第二步:求 R 的特征根 P 及相应的单位特征向量 A 、累计贡献率:对 R 这个 $7*7$ 的矩阵, 容易算出它的特征根 P 及对应的特征向量 A 和累计贡献率。如下所示:

第三步: 写出主成分 F

相关系数矩阵 R

相关系数	$X(1)$	$X(2)$	$X(3)$	$X(4)$	$X(5)$	$X(6)$	$X(7)$
$X(1)$	1.00000	0.22961	0.30836	0.19029	0.37884	0.03790	0.56277
$X(2)$	0.22961	1.00000	0.33902	0.00774	0.12501	0.14806	-0.17597
$X(3)$	0.30836	0.33902	1.00000	0.37043	-0.10841	-0.43431	0.19191
$X(4)$	0.19029	0.00774	0.37043	1.00000	0.07893	-0.16740	-0.00785
$X(5)$	0.37884	0.12501	-0.10841	0.07893	1.00000	0.33637	0.18266
$X(6)$	0.03790	0.14806	-0.43431	-0.16740	0.33637	1.00000	-0.08397
$X(7)$	0.56277	-0.17597	0.19191	-0.00785	0.18266	-0.08397	1.00000

规格化特征向量

相关系数	因子1	因子2	因子3	因子4	因子5	因子6	因子7
$X(1)$	0.57588	0.269254	-0.094613	-0.114975	0.154721	-0.607293	-0.425382
$X(2)$	0.22735	0.08165	0.73686	-0.38054	0.00627	-0.13027	0.48668
$X(3)$	0.49459	-0.39904	0.21154	-0.13801	-0.02005	0.58513	-0.43541
$X(4)$	0.33041	-0.23171	0.14134	0.80676	0.28909	-0.08335	0.27529
$X(5)$	0.23489	0.55633	0.06848	0.31944	-0.69457	0.21468	0.00945
$X(6)$	-0.16746	0.60825	0.23108	0.09506	0.60513	0.35890	-0.21108
$X(7)$	0.42388	0.16864	-0.57032	-0.24696	0.20843	0.30000	0.52171

相应的特征向量

序号	特征值	贡献率%	累计贡献率%
1	2.02238	28.89119	28.89119
2	1.66972	23.85316	52.74436
3	1.26358	18.05110	70.79546
4	0.91979	13.13988	83.93534
5	0.53063	7.58038	91.51572
6	0.34394	4.91348	96.42920
7	0.24996	3.57080	100.00000

从上表看，前四个特征值累计贡献率已达**83.94%**，说明前四个主成分基本包含了全部指标具有的信息，所以取前四个特征值，它们对应的特征向量为：

(0.57588,0.22735,0.49459,0.33041,0.23489,-0.16746,0.42388)

(0.269254,0.08165,-0.39904,-0.23171,0.55633,0.60825,0.16864)

(-0.094613,0.73686,0.21154,0.14134,0.06848,0.23108,-0.57032)

(-0.114975,-0.38054,-0.13801,0.80676,0.31944,0.09506,-0.24696)

所以前四个主成分为：

$$\text{第一主成分: } F_1 = 0.57588 X_1 + 0.22735 X_2 + 0.49459 X_3 + 0.33041 X_4 + 0.23489 X_5 - 0.16746 X_6 + 0.42388 X_7$$

$$\text{第二主成分: } F_2 = 0.269254 X_1 + 0.08165 X_2 - 0.39904 X_3 - 0.23171 X_4 + 0.55633 X_5 + 0.60825 X_6 + 0.16864 X_7$$

$$\text{第三主成分: } F_3 = -0.094613 X_1 + 0.73686 X_2 + 0.21154 X_3 + 0.14134 X_4 + 0.06848 X_5 + 0.23108 X_6 - 0.57032 X_7$$

$$\text{第四主成分: } F_4 = -0.114975 X_1 - 0.38054 X_2 - 0.13801 X_3 + 0.80676 X_4 + 0.31944 X_5 + 0.09506 X_6 - 0.24696 X_7$$

四、模型分析

- 由上述三个主成分的表达式，可以得到：
- 在第一主成分的表达式中，除第6个指标外其他每项指标的系数都差不大，可以把第一主成分看成是六个指标共同刻画的反映消费结构的综合指标。
- 在第二主成分的表达式中，第五、六项指标的影响较大，可以看成是交通和通信、娱乐教育文化的综合影响。
- 在第三主成分的表达式中，第二项影响特别大，可以单独看成是衣着的影响。
- 在第四主成分的表达式中，第四项影响特别大，可以单独看成是医疗保健和个人用品的影响。

Matlab 中主成分分析的有关函数

1.princomp

功能：主成分分析

格式：PC=princomp(X)

[PC,SCORE,latent,tsquare]=princomp(X)

说明：[PC,SCORE,latent,tsquare]=princomp(X)对数据矩阵X进行主成分分析，给出各主成分(PC)、所谓的Z-得分(SCORE)、X的方差矩阵的特征值(latent)和每个数据点的HotellingT2统计量(tsquare)。

2.pcacov

功能：运用协方差矩阵进行主成分分析

格式：PC=pcacov(X)

[PC,latent,explained]=pcacov(X)

说明：[PC,latent,explained]=pcacov(X)通过协方差矩阵X进行主成分分析，返回主成分(PC)、协方差矩阵X的特征值(latent)和每个特征向量表征在观测量总方差中所占的百分数(explained)。

3.pcares

功能：主成分分析的残差

格式：`residuals=pcares(X,ndim)`

说明：`pcares(X,ndim)`返回保留X的ndim个主成分所获的残差。注意，ndim是一个标量，必须小于X的列数。而且，X是数据矩阵，而不是协方差矩阵。

4.barttest

功能：主成分的巴特力特检验

格式：`ndim=barttest(X,alpha)`

`[ndim,prob,chisquare]=barttest(X,alpha)`

说明：巴特力特检验是一种等方差性检验。

`ndim=barttest(X,alpha)`是在显著性水平alpha下，给出满足数据矩阵X的非随机变量的n维模型，ndim即模型维数，它由一系列假设检验所确定，ndim=1表明数据X对应于每个主成分的方差是相同的；ndim=2表明数据X对应于第二成分及其余成分的方差是相同的。

第8章 判别分析简介及应用

判别分析是根据多指标来判断个体所属类别的一种多元统计分析方法，本质是利用多指标进行综合判断，根据变量取舍情况又分为多组判别和逐步判别。目前，在经济、气象、地质、冶金、生物、农业和医学等需要处理多元数据的诸多领域得到广泛应用。

常用逐步判别法有：马氏距离判别法、广义平方距离法、最大后验概率判别法、*Bayes* 准则判别法、*Fisher*判别法等。

一、逐步判别分析原理

对于一个多元数据矩阵，在数据库领域表现为多字段**二维表**。假设有来自 G 个母体的 n 个已知分类样本，每个样本有 m 个变量，则在数据库中加上样品标识和已知分类，共有 $m+2$ 个字段、 n 条记录。每个样本被看着是 m 维欧氏空间 R 上的一个点，每个母体都是 R 中的一个子空间 R_g ，这些子空间是互相排斥的，组成了 R 。需要找出一个办法，即找到判别函数，把空间 R 划分为 G 个子空间 $R_g(g=1,2,\dots,G)$ 。已知的样本有了空间归属和函数，就能对未知归属的样本进行判别，确定其归属，即判别归类或预测。

人们总是希望用较少的变量去划分空间 R ，因为采集数据记录时，字段越少越好，成本越低。这就需要衡量每个变量参与划分 G 个母体的能力。这就需要 F 检验，给出引入变量的 F 值和剔除变量的 F 值，作为引入和剔除变量的门限值。在一个母体内样本间的差异应当较小，不同母体的样本差异应当较大。根据 $Wilks$ 准则，组内离差越小、组间离差越大，越有利于 G 个母体的分类。通过计算组间离差 B 和组内离差 W ，然后进行 F 检验，就可以确定变量的取舍。逐步引入和剔除，最终得到区分能力较大的变量组合。

求得区分能力显著的 k 个变量组合后，计算判别系数，最终建立 G 个子空间的判别函数：

$$y_g(X) = \ln q_g + c_{0g} + \sum_{i \in k} c_{ig} x_i$$

其中 q_g 是第 g 组的先验概率，一般采用样品频率代替（ $q_g = n_g/n$ ）。 c_{ig} 是判别系数。对于某一待判别归属的样本，如果

$$y_g^*(X) = \max \{y_g(X)\}$$

则把该样本划归 g^* 类。也就是把样本中变量的观察值分别代入 G 个判别函数，哪个函数取值最大，就划归那一组。

可以将用来建模(求得判别函数)的原始数据回代到各组判别函数，求得样品的归属，与原来实际分类对比，以确定判别函数的准确度。一般回判效果都很好，正判率**85%**以上。

判别分析求得判别函数，就是建立了模型。目前的多元统计软件都是判别分析和待判别样品一并进行，这就造成模型的应用离不开建模样本。而有些单位或个人又不愿公布原始数据，这给模型的应用造成困难。实际上，只要模型中 q_g 和 cig 齐全，简单的编程就可解决问题。

二、蠓虫分类问题

(一). 问题提出

生物学家试图对两类蠓虫(Af 与 Apf)进行鉴别,依据的资料是蠓虫的触角和翼长,已经测得9只Af和6只Apf的数据(触角长度用 x 表示,翼长用 y 表示)

Af 数据

x 1.24 1.36 1.38 1.38 1.38 1.40 1.48 1.54 1.56

y 1.72 1.74 1.64 1.82 1.90 1.70 1.82 1.82 2.08

Apf 数据

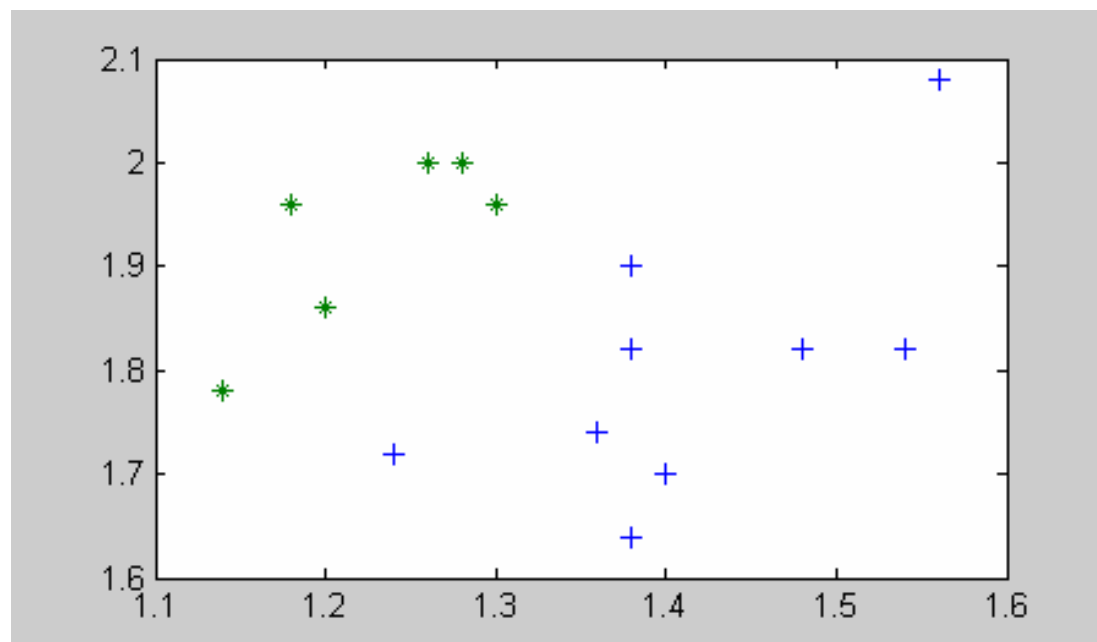
x 1.14 1.18 1.20 1.26 1.28 1.30

y 1.78 1.96 1.86 2.00 2.00 1.96

Af 数据: +

Apf数据: *

翼
长



触 角

需要解决的问题

- (1) 如何凭借原始资料(**A_f**和**A_p**的已知数据被称之为学习样本)制定一种方法,正确区分两类蠓虫;
- (2) 依据确立的方法,对未知类别的三个样本:
(1.24,1.80) ,**(1.28,1.84)** ,**(1.40,2.04)**加以识别。
- (3) 更一般的问题:设有 k 个类别 G_1, \dots, G_k ,对任意一个样品 $x \in G_i (i=1, \dots, k)$,其指标 X (一般 p 维)的值可测。现给定一个由已知所属类别的一些样品 x_1, \dots, x_n 组成的“学习样本”,要求对一个来自这 k 个类别的某样品 x ,据其指标 X 的值作出所属类别的判别。

(二). 问题解决

1. 距离判别模型

距离判别理论中，通常采用重心距离来定义类与类之间的距离,对待判样品进行判别.

在蠓虫分类中, $k = 2, G_1 = \text{Af}, G_2 = \text{Apf}$, 指标 $X = (A, W)^T$ 是二维的, 其中 A 为触角长, W 为翼长, 样品 x 的指标值为 $(a, w)^T$, 学习样品共 **15** 个, 其中 **9** 个属于 **Af**, **6** 个属于 **Apf**。

距离判别模型中，把每个样品 $x=(a,w)^\tau$ 视为二维空间中的一点，算出**9**个**Af**和**6**个**Apf** 样品点集合的中心

$$(a_{1i}, w_{1i}) \in Af, i = 1, \dots, 9$$

$$(a_{2i}, w_{2i}) \in Apf, i = 1, \dots, 6$$

$$\bar{x}_{Af} = (a_1, w_1)^\tau \triangleq \left(\frac{1}{9} \sum_{i=1}^9 a_{1i}, \frac{1}{9} \sum_{i=1}^9 w_{1i} \right)^\tau$$

$$\bar{x}_{Apf} = (a_2, w_2)^\tau \triangleq \left(\frac{1}{6} \sum_{i=1}^6 a_{2i}, \frac{1}{6} \sum_{i=1}^6 w_{2i} \right)^\tau$$

对于给定的样品 $x=(a,w)^T$ ，称 x 与 \bar{x}_{Af} 之间的距离为 x 距**Af**类的距离；称 x 与 \bar{x}_{Apf} 之间的距离为 x 距**Apf**类的距离。若 x 与**Af**的距离小于与**Apf**的距离，则判定 $x \in Af$ ，否则 $x \in Apf$.

一般地，平面上两点 $A(a_1,a_2),B(b_1,b_2)$ 的欧氏 *Euclidean* 距离定义为：

$$d(A,B) = \sqrt{\sum_{i=1}^2 (a_i - b_i)^2}$$

这种欧氏距离的缺陷：

- (1) 欧氏距离与单位选取有关；
- (2) 从概率角度看，用欧氏距离描述随机点之间距离并不好，因它忽略了随机变量的统计性质。

当然还有其它各种形式的距离，如

***Block*(绝对距离):** $\sum_i |x_i - y_i|$

***Chebychev*距离:** $\text{Max}_i |x_i - y_i|$

***Minkowski*距离:** $\left(\sum_i (x_i - y_i)^q \right)^{\frac{1}{q}}$

***Lance* 和 *Williams*距离:** $\frac{1}{p} \sum_i \frac{|x_i - y_i|}{x_i + y_i}$

下面介绍一种马氏距离(*Mahalanobis*: 马哈朗诺比斯)

马氏距离(Mahalanobis: 马哈朗诺比斯)

设总体 p 维总体 G 的均值为 $\mu = (\mu_1, \dots, \mu_p)^\tau$

协方差矩阵为非奇异矩阵 $V_{p \times p}$

则 p 维样品的 x 到总体 G 的马氏距离为

$$d_M(x, G) = \sqrt{(x - \mu)^\tau V^{-1} (x - \mu)}$$

马氏距离的好处:

- 1) 可以克服变量之间的相关性干扰;
- 2) 可以消除各变量量纲的影响。

当 μ 、 V 未知时，用样本均值和样本方差代替

$$\bar{x} = (\bar{x}_1, \dots, \bar{x}_p)^T, \text{ 其中 } \bar{x}_i = \frac{1}{n} \sum_{j=1}^n x_{ij} \quad (i = 1, \dots, p)$$

$$S = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})(x_j - \bar{x})^T, \text{ 其中 } x_j = (x_{1j}, \dots, x_{pj})^T$$

利用马氏距离进行判别的准则:

若 $d_M(x, G_1) < d_M(x, G_2)$, 则判断 $x \in G_1$;

若 $d_M(x, G_1) > d_M(x, G_2)$, 则判断 $x \in G_2$;

若 $d_M(x, G_1) = d_M(x, G_2)$, 则判断 $x \in G_1$ 或 $x \in G_2$

到蠓虫分类问题，据学习样本数据，可求得

$$\bar{x}_{Af} = \begin{pmatrix} 1.413 \\ 1.804 \end{pmatrix}, S_{Af} = \begin{pmatrix} 0.00975 & 0.00813 \\ 0.00813 & 0.01688 \end{pmatrix}$$

$$\bar{x}_{Apf} = \begin{pmatrix} 1.223 \\ 1.927 \end{pmatrix}, S_{Apf} = \begin{pmatrix} 0.0044 & 0.0042 \\ 0.0042 & 0.0078 \end{pmatrix}$$

设任给一蠓虫 $x = (a, w)^T$ ，则它到 Af 和 Apf 的马氏距离分别为

$$d_M(x, Af) = \sqrt{171.40a^2 + 99.47w^2 - 165.89aw - 185.11a - 125.79w + 245.41}$$

$$d_M(x, Apf) = \sqrt{467.63a^2 + 263.79w^2 - 502.36aw - 175.27a - 402.26w + 495.06}$$

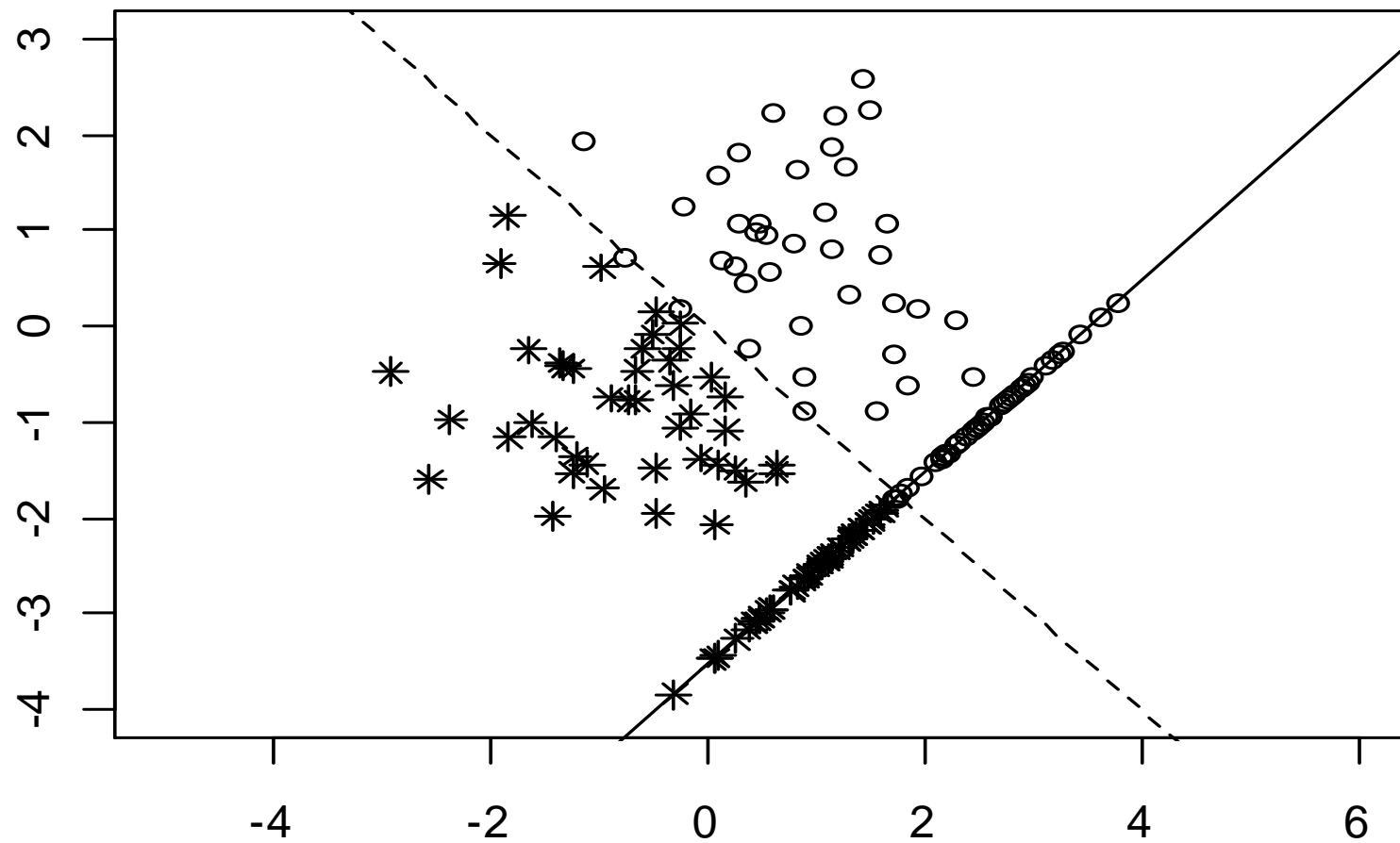
当 $x = (1.24, 1.80)$ 时， $d_M(x, Af) = 2.24 < d_M(x, Apf) = 2.34$ ，故判断为 $x \in Af$

当 $x = (1.28, 1.84)$ 时， $d_M(x, Af) = 1.98 < d_M(x, Apf) = 2.45$ ，故判断为 $x \in Af$

当 $x = (1.40, 2.04)$ 时， $d_M(x, Af) = 2.40 < d_M(x, Apf) = 2.82$ ，故判断为 $x \in Af$

2. *Fisher*判别模型 *Fisher*判别法(先进行投影)

- 所谓**Fisher**判别法，就是一种先投影的方法。
- 考虑只有两个（预测）变量的判别分析问题。
- 假定这里只有两类。数据中的每个观测值是二维空间的一个点。见图（下一张幻灯片）。
- 这里只有两种已知类型的训练样本。其中一类有**38**个点（用“○”表示），另一类有**44**个点（用“*”表示）。按照原来的变量（横坐标和纵坐标），很难将这两种点分开。
- 于是就寻找一个方向，也就是图上的虚线方向，沿着这个方向朝和这个虚线垂直的一条直线进行投影会使得这两类分得最清楚。可以看出，如果向其他方向投影，判别效果不会比这个好。
- 有了投影之后，再用前面讲到的距离远近的方法来得到判别准则。这种首先进行投影的判别方法就是**Fisher**判别法。



在距离模型中，是将二维样品 $x=(a,w)^T$ 变为1维的距离 d_M 来作判断。*Fisher*的思想也是将多维样品测量值 x 变换为1维的测量值 y ，并据 y 进行判别。

具体做法:先引入一个与样品有相同维数的待定向量 u ，再将 y 取为 x 坐标（或分量）的线性组合 $y = u^T x$ 。 u 的选取要使同一类别产生的 y 尽量聚拢，不同类别产生的 y 尽量拉开，这样可将样品 x 到 G 的距离定义为

$$L(x, G) = |y - \bar{y}| = |\mu^T (x - \bar{x})|, \bar{x} \text{ 是 } G \text{ 的中心}$$

并由样品 x 距各类别的距离大小判别 x 的所属类别。

记 $x_j^{(i)}$ 为第 i 类第 j 个样品 ($i = 1, \dots, k ; j = 1, \dots, n_i$)

[illegible]

$$\bar{x}^{(i)} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_j^{(i)} \text{ 为第 } i \text{ 类的中心, } \bar{x} = \frac{1}{k} \sum_{i=1}^k \bar{x}^{(i)} \text{ 为总的中心}$$

$$\mathbf{y}_j^{(i)} = \mathbf{u}^\tau \mathbf{x}^{(i)} \quad , \quad \overline{\mathbf{y}}^{(i)} = \mathbf{u}^\tau \overline{\mathbf{x}}^{(i)} \quad , \quad \overline{\mathbf{y}} = \mathbf{u}^\tau \overline{\mathbf{x}}$$

为选取满足要求的 u ,先计算 y 的总离差平方和

$$\begin{aligned} Z &= \sum_{i=1}^k \sum_{j=1}^{n_i} \left(y_j^{(i)} - \bar{y} \right)^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} \left[\left(y_j^{(i)} - \bar{y}^{(i)} \right) - \left(\bar{y}^{(i)} - \bar{y} \right) \right]^2 \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} \left(y_j^{(i)} - \bar{y}^{(i)} \right)^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} \left(y_j^{(i)} - \bar{y} \right)^2 \triangleq E_0 + B_0 \end{aligned}$$

E_0 刻画同类别产生的 y 值的离散程度;

B_0 刻画不同类别产生的 y 值的离散程度。

记 $S_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(x_j^{(i)} - \bar{x}^{(i)} \right) \left(x_j^{(i)} - \bar{x}^{(i)} \right)^\tau$ 则有

$$E_0 = u^\tau \left(\sum_{i=1}^k (n_i - 1) S_i \right) u \triangleq u^\tau E u, \text{ 其中 } E = \sum_{i=1}^k (n_i - 1) S_i$$

类似地有

$$E_0 = u^\tau \left\{ \sum_{i=1}^k \left[n_i (x^{(i)} - \bar{x})(x^{(i)} - \bar{x})^\tau \right] \right\} u \triangleq u^\tau B u$$

$$\text{其中 } B = \sum_{i=1}^k \left[n_i (x^{(i)} - \bar{x})(x^{(i)} - \bar{x})^\tau \right]$$

根据对 u 的不同要求, *Fisher* 提出所选择的 u 应使下式定义的判别效率达极大

$$\varphi(u) = \frac{u^\tau B u}{u^\tau E u} = \frac{B_0}{E_0}$$

因为这时 B_0 达到极大而 E_0 达到极效，从而用这样的 u 作线形组合 $y = u^T x$ ，就能使同类别产生的 y 值尽可能聚拢，不同类别产生的 y 值尽可能拉开。

由于 $\varphi(u)$ 达极大的 u 不唯一，可以差任一常数倍，故可增加一个约束条件，将求 u 问题化为如下优化问题：

$$\begin{cases} \max \varphi(u) = u^T B u \\ s.t. \quad u^T E u = 1 \end{cases}$$

用拉格朗日乘数法求解，令 $L(u) = u^{\tau} B u - \lambda(u^{\tau} E u - 1)$

$$\text{由} \begin{cases} \frac{d L(u)}{d u} = 0 \\ u^{\tau} E u = 1 \end{cases} \Rightarrow (E^{-1} B) u = \lambda u$$

说明所求 u 的正好是矩阵 $E^{-1} B$ 的特征方程

$$|E^{-1} B - \lambda I| = 0$$

对应于最大特征值 λ^* 且满足 $u^{\tau} E u = 1$ 的特征向量 u^*

回到蠓虫分类问题，利用所给的学习样本数据算得

$$E = \begin{pmatrix} 0.100 & 0.086 \\ 0.086 & 0.174 \end{pmatrix} \quad B = \begin{pmatrix} 0.135 & 0.134 \\ 0.134 & 0.058 \end{pmatrix}$$

$$\lambda^* = 1.37 \quad u^* = (2.930, 0.258)^T$$

对任一样品 $x = (a, w)^T$ ，有

$$L(x, Af) = |2.93a + 0.258w - 4.605|$$

$$L(x, Apf) = |2.93a + 0.258w - 4.080|$$

当 $x = (1.24, 1.80)$ 时 $L(x, Af) = 0.508, L(x, Apf) = 0.017 \Rightarrow x \in Apf$

当 $x = (1.28, 1.84)$ 时 $L(x, Af) = 0.380, L(x, Apf) = 0.145 \Rightarrow x \in Apf$

当 $x = (1.40, 2.04)$ 时 $L(x, Af) = 0.023, L(x, Apf) = 0.548 \Rightarrow x \in Af$

与距离判别法结论相比，只有第三个相同，前两个却相反，说明两种方法均有误判的情况。

3. *Bayes*判别模型

距离判别和*Fisher*判别模型中均不涉及各类别(总体)的分布, 只要求均值、方差和协方差存在(即二阶矩均存在) 即可, 虽然使用方便, 但不能计算误判概率及因误判而引起的损失.

*Bayes*判别模型依据各类别分布的信息, 从考虑因误判而引起的损失最小角度出发, 建立判别准则, 是当前应用较为广泛的一种判别模型。

为简便起见，这里只考虑两个类型的情形。

设已知总体 G_1 和 G_2 的概率密度函数为 $f_1(x)$ 和 $f_2(x)$ ，仍将一个 p 维样品 x 视为 R^p 空间中的一个点。这时建立一个判别准则，相当于把整个 R^p 空间作一个划分，分为互不相交的两个集合 D_1 和 D_2 ($D_1 \cup D_2 = R^p, D_1 D_2 = \Phi$),若 $x_i \in D_i$,则判断 $x_i \in G_i (i=1,2)$.

这里把既可判 $x_1 \in G_1$ ，又可判 $x_2 \in G_2$ 的样品明确归入某一类。

属于 G_1 误判为 G_2 的概率为 $P_1(D) = \int_{D_2} f_1(x) dx$

属于 G_2 误判为 G_1 的概率为 $P_2(D) = \int_{D_1} f_2(x) dx$

若不考虑误判引起的损失或两类误判引起的损失相等时

使 $P_1(D) + P_2(D)$ 达最小的 D 是最佳判别

若两类误判引起的损失不相等，应以平均损失最小为佳

设两类误判引起的损失分别为 C_1 和 C_2 ，并设样品 x 属于 G_1 和 G_2 的先验概率分别为 p_1 和 p_2 ($p_1 + p_2 = 1$), 则判别 $D = (D_1, D_2)$ 因误判引起的平均损失为

$$C_1 p_1 P_1(D) + C_2 p_2 P_2(D)$$

$$= C_1 p_1 \int_{D_2} f_1(x) dx + C_2 p_2 \int_{D_1} f_2(x) dx$$

$$= \int_{D_1} (C_2 p_2 f_2(x) - C_1 p_1 f_1(x)) dx + C_1 p_1$$

由此可见要是误判引起的平均损失取最小值，当且仅当

将满足 $C_1 p_1 f_1(x) > C_2 p_2 f_2(x)$ 的 x 归入 D_1

将满足 $C_1 p_1 f_1(x) < C_2 p_2 f_2(x)$ 的 x 归入 D_2

将满足 $C_1 p_1 f_1(x) = C_2 p_2 f_2(x)$ 的 x 随意归入 D_1 或 D_2

定义 $D=(D_1, D_2)$ 为

$$D_1 = \{x : C_1 p_1 f_1(x) \geq C_2 p_2 f_2(x)\}$$

$$D_2 = \{x : C_1 p_1 f_1(x) < C_2 p_2 f_2(x)\}$$

用这样的 D 作判别的模型就是 **Bayes** 模型

回到蠓虫分类问题，不妨设指标 $X=(A,w)^T$ 在 **Af** 和 **Apf** 中分别服从正态分布：

$$N(\mu^{(1)}, \Sigma^{(1)}) \text{ 和 } N(\mu^{(2)}, \Sigma^{(2)})$$

其中 $\mu^{(1)}, \Sigma^{(1)}, \mu^{(2)}, \Sigma^{(2)}$ 可由学习样本提供的 Af 和 Apf 的数据算得的 \bar{x}_{Af}, S_{Af} 和 \bar{x}_{Apf}, S_{Apf} 分别代替

这样就可写出 **Af** 和 **Apf** 的概率密度如下

$$\begin{aligned}
f_{Af}(a, w) &= \frac{1}{2\pi \begin{vmatrix} 0.00975 & 0.00813 \\ 0.00813 & 0.01688 \end{vmatrix}^{\frac{1}{2}}} \\
&\exp \left\{ -\frac{1}{2} (a-1.413, w-1.804) \begin{pmatrix} 0.00975 & 0.00813 \\ 0.00813 & 0.01688 \end{pmatrix}^{-1} \begin{pmatrix} a-1.413 \\ w-1.804 \end{pmatrix} \right\} \\
&= \frac{1}{0.01985\pi} \exp \left\{ - \begin{pmatrix} 85.7000a^2 + 49.7366w^2 - 82.9453a w \\ -92.5549a - 62.8945w + 122.7041 \end{pmatrix} \right\}
\end{aligned}$$

$$\begin{aligned}
f_{Apf}(a, w) &= \frac{1}{2\pi \begin{vmatrix} 0.0044 & 0.0042 \\ 0.0042 & 0.00781 \end{vmatrix}^{\frac{1}{2}}} \\
&\exp \left\{ -\frac{1}{2}(a-1.223, w-1.927) \begin{pmatrix} 0.0044 & 0.0042 \\ 0.0042 & 0.00781 \end{pmatrix}^{-1} \begin{pmatrix} a-1.223 \\ w-1.927 \end{pmatrix} \right\} \\
&= \frac{1}{0.00817\pi} \exp \left\{ - \begin{pmatrix} 233.8130a^2 + 131.8945w^2 - 251.1798aw \\ -87.8828a - 201.1278w + 247.5276 \end{pmatrix} \right\}
\end{aligned}$$

假设自然界中**Af**与**Apf**的蠓虫数量相等，可取
 $p_1=p_2=1/2$ ，又若两类误判引起的损失相同，可取
 $C_1=C_2=C$ ，计算得

$$D_1=\{(a,w):H(a,w) \leq 0, \quad D_2=\{(a,w):H(a,w) > 0$$

$$\text{其中 } H(a,w) = -148.113a^2 - 82.1579w^2 + 168.2345aw \\ - 4.6721a + 138.2334w - 123.9359$$

$$\text{经计算得 } H(1.24,1.80) = 0.0061 > 0$$

$$H(1.28,1.84) = -0.1630 < 0, \quad H(1.40,2.04) = -0.2128 < 0$$

可判断

$$(1.24,1.80) \in Apf, \quad (1.28,1.84) \in Af, \quad (1.40,2.04) \in Af$$

设**Af**为传粉益虫，**Apf**为某种疾病的载体，则有理由认为把**Apf**误判为**Af**的危害比把**Af**误判为**Apf**的危害严重，这时可选取适当的 C_1 ， C_2 使 $C_1 < C_2$ 。例如在 $p_1 = p_2 = 1/2$ 假设下，可取 $C_1 = 1$ ， $C_2 = 2$ ，这时有

$$D_1 = \{(a, w) : H^*(a, w) \leq 0, \quad D_2 = \{(a, w) : H^*(a, w) > 0$$

$$\text{其中 } H^*(a, w) = -148.113a^2 - 82.1579w^2 + 168.2345aw \\ - 4.6721a + 138.2334w - 123.9359$$

$$\text{经计算得 } H^*(1.24, 1.80) = 1.3533 > 0$$

$$H^*(1.28, 1.84) = 0.5302 > 0, \quad H^*(1.40, 2.04) = 0.4804 > 0$$

可判断 $(1.24, 1.80)$ ， $(1.28, 1.84)$ ， $(1.40, 2.04)$ 均属于**Apf**

假设学习样本中的**15**只蠓虫是随机捕获的，且自然界中每只蠓虫有相同的机会被捕获时，则可采用学习样品中**A_f**与**A_{pf}**的数量作为***p*1**与***p*2**的比值，即取***p*1=0.6**，***p*2=0.4**。