一、参数估计

MATLAB 的统计工具箱中对一些常用的特殊分布，提供了相应的极大似然估计函数（以 fit 结尾），下面举例讲述极大似然估计以及它的区间估计。

例 1. 参数设总体 X 在[a, b]上服从均匀分布，a，b 未知。现有如下的一组样本值：

1.2，1.4，1.8，2.1，3.4，4.5,2.6,5.0,1.8,3.8,4.2,3.4,2.6,4.8,4.0

试求出 a，b 的估计量并给出 95%置信区间。

解：下面给出分别用专用函数 unifit 以及通用函数 mle 求解的过程

```
x = [ 1.2 1.4 1.8 2.1 3.4 4.5 2.6 5.0 1.8 3.8 4.2 3.4 2.6 4.8 4.0 ];
[ahat, bhat, aci, bci] = unifit(x)
[phat, pci] = mle(x, 'distribution', 'uniform')
```

```
ahat =
    1.2000
bhat =
     5
aci =
    0.3600
    1.2000
bci =
    5.0000
    5.8400
phat =
    1.2000    5.0000
pci =
    0.3600    5.0000
    1.2000    5.8400
```

由此解可得 a 的估计值为 1.2，置信区间为[0.36,1.2];b 的估计值为 5，置信区间为[5,5.84]。

例 2.一地质科学家为研究密执根湖湖滩地区的岩石成分，随机地自该地区取 100 个样品，每个样品有 10 块石子，记录了每个样品中属石灰石的石子数。

假设 100 次观察相互独立，并且由过去经验知，它们都服从参数为 n=10，p 的二项分布。p 是这地区一块石子是石灰石的概率，求 p 的极大似然估计值。数据如下标所示。

样品中属石灰石的样品数	0	1	2	3	4	5	6	7	8	9	10
观察到石灰石的样品个数	0	1	6	7	23	26	21	12	3	1	0

```
x=[1,2*ones(1,6),3*ones(1,7),4*ones(1,23),6*ones(1,21),7*ones(1,12),8*o
nes(1,3),9];
n = 10;
```

```
p = binofit(x, n);
p1 = mean(p)

p1 =
    0.4986
```

例 3. 已知下列数据为指数分布，求它们的点估计值和 97%的置信区间。数据为：

1, 6, 8, 25, 28, 15, 1, 3, 8

```
x = [1 6 8 25 28 15 1 3 8];
[parmhat, parmci] = expfit(x, 0.03)

parmhat =
   10.5556
parmci =
    5.6917
   25.2778
```

二、假设检验

假设检验是统计推断的基本问题之一，主要确定关于样本总体特征的判断是否合理的过程。按一定规则（检验准则）根据样本所做假设 H 是否成立，以决定是接受还是拒绝 H。建设检验的判断与结论是根据样本做出的，故具有“概率性”。

首先对几个必要的名词作解释。

零假设：即指初始判断。

显著性水平：是与支持对立假设二拒绝零假设的确定度有关的量。在小样本的前提下，不能肯定自己的结论，所以事先约定，如果观测到符合零假设的样本值的概率小于显著性水平 α，则拒绝零假设。典型的显著性水平 α =0.05。如果要减少犯错误的可能，可取更小的值。

P-值：在假定零假设为真的条件下，观测绘制定样本结果的概率值。如果 P-值小于 α，则拒绝零假设；反之则并非如此，如果 P-值大于 α，则并不能肯定就接受零假设。此时，只是没有足够的证据来拒绝零假设。

假设检验的输出包括置信区间。不严格来说，置信区间是那些包含真实的假设量的可选概率的值的范围。

每一种检验都有一个信噪比：

$$\mathrm{Z}=\frac{\bar{x}-\mu}{\sigma} \text{ 或} \mathrm{T}=\frac{\bar{x}-\mu}{S}$$

其中

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

信号为算术平均值与假设均值之差；噪声为指定的或估计得标准差。

如果零假设为真，则 Z 服从标准正态分布 N(0)。T 服从自由度为 v 的 T 分布，其中，v 等于数据个数减 1。

1．单个总体 N(μ,σ^2)均值 μ 的检验

•σ^2已知，关于 μ 的检验（Z 检验）

MATLAB 中 Z 检验用函数 ztest 来实现。

例 1．某糖厂有一台自动打包机做打包，额定标准每包质量为 100kg 包质量服从正态分布，且根据以往经验，其方差 $\sigma^2=(0.4)^2$。某天开工后，为检查打包机工作情况，随机地抽取 9 包，称得质量如下：

99	98.5	102.5	101	98	99	102	102.1	100.5

问这天打包机工作是否正常？设显著性水平位 0.05。

解：

原假设 *H0：　μ＝100；*

备择假设 *H1：　μ≠100。*

由于该样本服从正态分布且方差已知，因此可以用 z 统计量检验假设。具体通过下面的程序实现。

```
x = [ 99  98.5  102.5  101  98  99  102  102.1  100.5];
m = 100;
[h, sig, ci] = ztest(x, m, 0.4)
```

输出结果为：

```
h =
     1
sig =
    0.0303
ci =
  100.0276  100.5502
```

•σ^2未知，关于 μ 的检验（t 检验）

在 MATLAB 中用 ttest 来实现。

例 2.在某砖厂生产的一批砖中，随机地抽取 6 块进行抗断强度试验，测得结果（单位：kg/cm^2）如下：32.56，29.66，32.64，30.00，31.87，32,03；设砖的抗断强度服从正态分布，问这批砖的平均抗断强度是否不大于 32.50（kg/cm^2）？

解：

H0：　μ≤32.50

H1：　μ＞32.50

由于其为方差未知，检验均值大小的假设检验问题，故选择 t 统计量进行检验。具体用下列程序实现。

在 MATLAB 命令窗口中输入：

```
x = [32.56 29.66 32.64 30.00 31.87 32.03];
m = 32.5;
h = ttest(x, m, 0.05, 'right')
```

输出结果为：

```
h =
     0
```

由 h=0 可知在显著性水平为 0.05 的情况下，不能拒接原假设。即认为这批砖的平均抗断强度不大于 32.5(kg/cm^2)。

2.两个正态总体均值差的检验（t 检验）

在 MATLAB 中用 ttest2 来检验两个正态样本的均值是否相同。

例 3. 表 1 中分别给出两个文学家马克吐温的 8 篇小品文以及斯若特格拉斯的 10 篇小品文中由 3 个字母组成的单词的比例。

设两组数据分别来自正态总体，且两总体方差相等，但参数均未知。两样本相互独立。问两个作家所写的小品文中包含由 3 个字母组成的单词的比例是否有显著性差异。

表 1.两文学家中 3 个字母的单词比例

马克吐温	0.225	0.262	0.271	0.240	0.230	0.229	0.235	0.217		
斯若特格拉斯	0.209	0.205	0.196	0.210	0.202	0.207	0.224	0.223	0.220	0.201

解：

$$H0:\quad \mu_1 = \mu_2$$
$$H1:\quad \mu_1 \neq \mu_2$$

取 α =0.05.

采用两个正态总体均值差的检验，用 ttest2 计算如下：

```
x = [ 0.225 0.262  0.271  0.240  0.230  0.229  0.235  0.217];
y = [ 0.209 0.205  0.196  0.210  0.202  0.207  0.224  0.223  0.220
  0.201];
h = ttest2(x, y)

h =
     1
```

由 h=1 得，在显著性水平为 0.05 的情况下拒绝原假设。即两个作家的作品中包含 3 个字母的单词的比例有显著差异。

3.基于成对数据的检验（t 检验）

MATLAB 中的 ttest 的另一用法便是基于成对数据的检验。

例 4. 有两台光谱仪 Ix, Iy，用来测量材料中某种金属的含量，为鉴定他们的测量结果有无显著性差异，制备了 9 件试块（他们的成分、金属含量、均匀性等均各不相同），现在分别用这两台仪器对每一试块测量一次，得到 9 对观察值，如表 2 所示。

表 2. 两台光谱仪测定材料中金属含量

x(%)	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
y(%)	0.10	0.21	0.52	0.32	0.78	0.59	0.68	0.77	0.89

问：能否认为这两台仪器的测量结果有显著的差异（去 α=0.01）？

解：设 μ_D 为两组数据差的均值，可得如下假设

$$H0: \quad \mu_D = 0$$

$$H1: \quad \mu_D \neq 0$$

采用基于成对数据的 t 检验，用 ttest 计算如下：

```
x = [ 0.20  0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00 ];
y = [ 0.10  0.21 0.52 0.32 0.78 0.59 0.68 0.77 0.89 ];
h = ttest(x, y, 0.01)

h =
     0
```

故由 h=0 得，在显著性水平为 0.01 的情况下，不能拒绝原假设，即认为两台仪器的测量结果并无显著差异。

4. 正态总体方差的假设检验

• 单个总体的情况

样本来自单个服从正态分布总体，其均值、方差未知，检验其方差是否等于某一特定值。此类假设检验问题采用统计量 x^2 进行检验。

在 MATLAB 中用函数 vartest 实现。

例 5. 某厂生产一直比较稳定，长期以来，螺钉的直径服从方差为 0.0002（cm^2）的正态分布。今从产品中随机抽取 10 只进行测量，得螺钉直径的数据（单位：cm）如下：1.19, 1.21, 1.21, 1.18, 1.17, 1.20, 1.20, 1.17, 1.19, 1.18。问是否可以认为该厂生产的螺钉直径的方差为 0.0002（cm^2）？（取 α=0.05）

解：此题检验 H0：σ^2=0.0002，用 vartest 计算如下：

```
x = [ 1.19,1.21,1.21,1.18,1.17,1.20,1.20,1.17,1.19,1.18 ];
v = 0.0002;
h = vartest(x, v)

h =
     0
```

故在显著性水平 0.05 情况下接受原假设，即认为生产的螺钉直径的方差为 $0.0002cm^2$ 。

• 两个总体的情况

通常用 F 统计量检验两独立样本的方差是否相同。

MATLAB 中采用 vartest2 实现。

例 6.有两台机床加工同一种零件，这两台机床生产的零件尺寸服从正态分布。今从两台机床的零件中分别抽取 11 个和 9 个零件进行测量，得数据（单位：mm）如下：

甲机床：6.2,5.7,6.5,6.0,6.3,5.8,5.7,6.0,6.0,5.8,6.0;

乙机床：5.6,5.9,5.6,5.7,5.8,6.0,5.5,5.7,5.5 。

问甲机床的加工精度是否比乙机床的加工精度较差？（取 α =0.05）

解：

$$H0: \quad \sigma_1^2 = \sigma_2^2$$
$$H1: \quad \sigma_1^2 < \sigma_2^2$$

用 vartest2 计算如下

```
x = [ 6.2,5.7,6.5,6.0,6.3,5.8,5.7,6.0,6.0,5.8,6.0 ];
y = [ 5.6,5.9,5.6,5.7,5.8,6.0,5.5,5.7,5.5 ];
h = vartest2(x, y, 0.05, 'left')

h =
     0
```

故在显著性水平为 0.05 的情况下接受原假设，即认为甲机床的加工精度与乙机床加工精度相当。

5. 非参数检验

根据样本数据来估计总体的分布类型就是非参数检验问题。MATLAB 中提供了 jbtest、lillietest 这两个函数分别根据不同的方法来进行正态分布拟合的假设检验。

- h = jbtest(x)：对输入量 x 进行 Jarque-Bera 测试
- h = jbtest(x,alpha)：alpha 为显著性水平，默认值为 0.05。
- [h,p] = jbtest(…)：p 为观测值的概率，小于 alpha，则可以拒绝正态分布的原假设
- [h,p,jbstat] = jbtest(…)：jbstat 为测量统计量的值。
- [h,p,jbstat,critval] = jbtest(…)：critval 为是否拒绝原假设的临界值。

返回 h 为一个布尔值，当 h=1 是表示拒绝假设，h=0 是表示接受原假设。

例 7.从一批滚珠中随机抽取 40 个，测得它们的直径（单位：mm）为

14.8 15.0 14.9 15.1 15.6 15.8 15.7 14.3 14.7 14.6
15.1 15.3 15.0 14.7 14.6 14.2 15.3 15.6 15.5 14.9
15.2 14.6 15.3 14.7 15.0 15.2 14.6 15.3 14.3 14.4
15.7 14.3 15.9 14.6 15.0 15.1 14.7 14.6 14.8 15.8

是否可以认为这批滚珠的直径服从正态分布（α=0.05）？并求出总体的与方差的点估计。

```
x = [14.8 15.0 14.9 15.1 15.6 15.8 15.7 14.3 14.7 14.6 15.1 15.3 15.0
14.7 14.6 14.2 15.3 15.6 15.5 14.9 15.2 14.6 15.3 14.7 15.0 15.2 14.6
15.3 14.3 14.4 15.7 14.3 15.9 14.6 15.0 15.1 14.7 14.6 14.8 15.8];
[h, p, jbstat, critval] = jbtest(x, 0.05)
mu = mean(x)
sig2 = var(x)

h =
     0
p =
    0.2741
jbstat =
    1.6564
critval =
    4.7481
mu =
   14.9950
sig2 =
    0.2154
```

结果 h=0 表明，在置信区间水平α=0.05 下接受原假设，且 p=0.2729 表明接受假设的概率也很大，测试值 jbstat=1.6564 小于临界值 critval=4.7481，所以接受原假设。此时均值与方差的点估计分别为 14.9950 及 0.2154。

三、方差分析

任何事件总受多种因素的影响，但各个因素对事件的影响可能是不相同的，而且同一个因素对不同水平的影响也有可能不同。利用测量数据分析各个因素对该事件的影响，分析因素对该事物的影响是否显著，这种数据处理方法即为数理统计中的方差分析。

如果仅考虑某一因素 A 对事件的影响，在试验时让其他因素保持不变，只改变因素 A，这样的试验称之为单因素试验；如果考虑两个以上的因素 A，B 等对事件的影响，则称为双因素及多因素试验。A，B 等因素所处的状态称为水平。利用试验数据分析各因素对事件的影响是否显著的方法则相应地称之为单因素方差分析、双因素方差分析。

1．单因素试验的方差分析

单因素试验的方差分析是指试验中只有一个因素发生改变。MATLAB 中，单因素试验的方差分析课用函数 anoval 实现。

例 1. 将抗生素注入人体会产生抗生素与血浆蛋白结合的现象，以致降低了药效。表 3 列出 5 种常用的抗生素注入到牛的体内时，抗生素与血浆蛋白结合的百分比。试检验这些百分比的均值有无显著的差异。

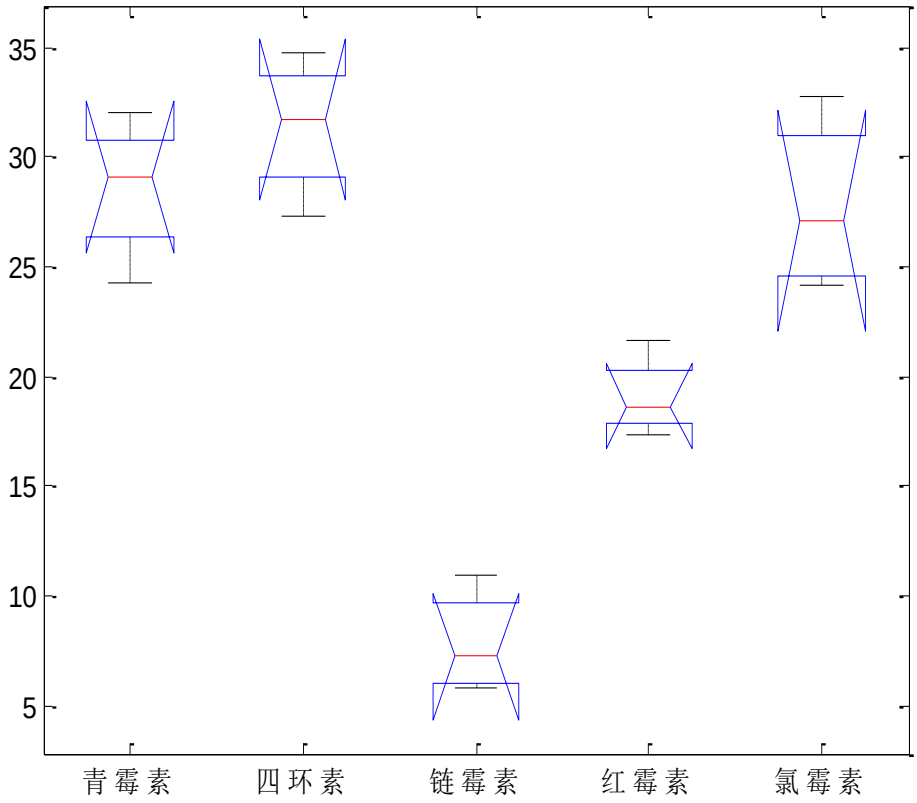
表 3. 各抗生素与血浆蛋白结合百分比

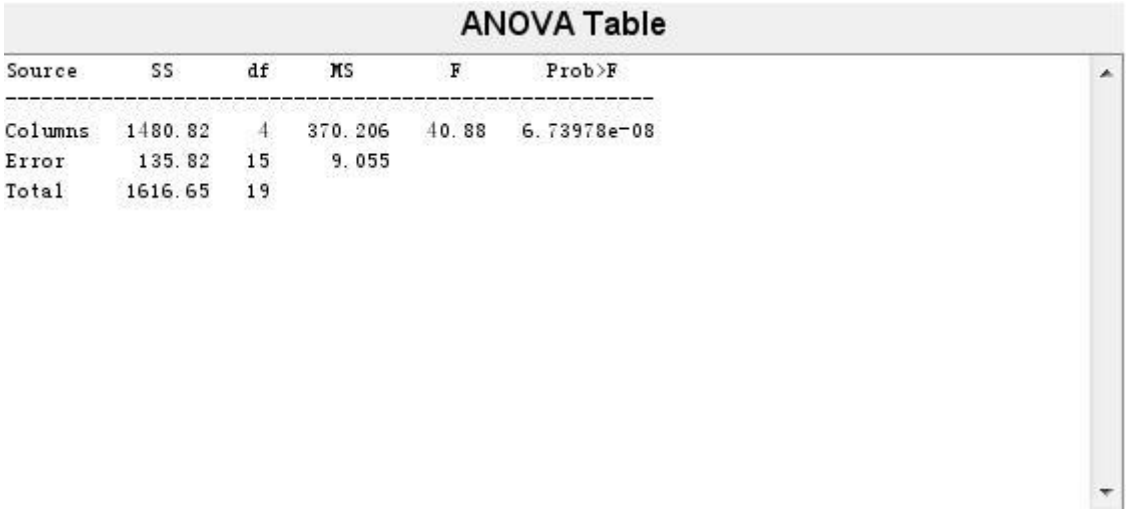
青霉素	四环素	链霉素	红霉素	氯霉素
29.6	27.3	5.8	21.6	29.2
24.3	32.6	6.2	17.4	32.8
28.5	30.8	11.0	18.3	25.0
32.0	34.8	8.3	19.0	24.2

解：

```
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];
group = ['青霉素';'四环素';'链霉素'; '红霉素';'氯霉素'];
p = anoval(x, group)

p =
    6.7398e-08
```





ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	1480.82	4	370.206	40.88	6.73978e-08
Error	135.82	15	9.055		
Total	1616.65	19			

2. 双因素试验的方差分析

通常用 F 统计量检验两独立样本的方差是否相同。MATLAB 中采用 vartest2 实现。

例 2. 在某种金属材料的生产过程中，对热处理温度（因素 B）与时间（因素 A）各取两个水平，产品强度的测试结果（相对值）如表 4 所示。

表 4. 不同温度（B）、时间（A）下产品强度

因素	B1	B2
A1	38.0 38.6	47.0 44.8
A2	45.0 43.8	42.4 40.8

在同一条件下每个试验重复两次。设各水平搭配下强度的总体服从正态分布且方差相同，各样本独立。问热处理温度、时间以及这两者的交互作用对产品强度是否有显著的影响（取 α =0.05）？

解：用 anova2 函数求解

```
x = [ 38.0 47.0; 38.6 44.8; 45.0 42.4; 43.8 40.8];
p = anova2(x, 2)

p =
    0.0340    0.3009    0.0024
```

双因素方差分析的 ANOVA 表格

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	11.52	1	11.52	10.02	0.034
Rows	1.62	1	1.62	1.41	0.3009
Interaction	54.08	1	54.08	47.03	0.0024
Error	4.6	4	1.15		
Total	71.82	7			

由结果知第一个 p 值代表列样本均值相同假设的 p 值，即反映了 B 因素（温度因素）的影响。由于 p(1)很小，故可得 B 因素对产品强度影响显著。

同理可得 A 因素（时间因素）对产品强度影响不显著（因 p(2)>0.05）,A、B 因素的交互作用影响更为显著（p(3)<0.01）。

四、正交试验分析

在科学研究与生产中，经常要做很多试验，这就存在着如何安排试验与如何分析试验结果的问题。试验安排得好，试验次数不用多，就可以得到满意的结果；安排得不好，次数再多，结果也往往不能令人满意。因此，合理安排试验时很值得研究的一个问题。正交设计法就是一种科学安排与分析多因素试验的方法，它主要是利用一套现成的规格化表——正交表来科学地挑选试验条件。

1．极差分析

极差分析法又叫直观分析法。他具有计算简便、直观形象、简单易懂等优点，是正交试验结果分析中最常用的方法。极差分析的方法简称为 R 法。

在 MATLAB 中没有提供专门的函数进行正交极差分析，下面通过编写自定义正交试验极差分析函数 zjsyjc.m 来进行分析，代码在 statistic 文件里。

例 1．某厂生产的油泵柱塞组合件存在质量不稳定、拉脱力波动大的问题。该组合件要求满足承受拉脱力大于 900kg。为了寻找最优工艺条件，提高产品质量，决定进行试验。根据经验，认为柱塞头的外径、高度、倒角、收口油压（分别记为 A，B，C，D）4 个因素对拉脱力可能有影响，因此决定在试验中考察这 4 个因素，并根据经验确定了各个因素的 3 种水平。试验结果如表 5 所示，试对其进行极差分析。

表 5.测量数据

编号	A	B	C	D	拉脱力数据
1	1	1	1	1	863
2	1	2	2	3	954
3	2	3	3	3	953
4	2	3	2	2	942
5	3	2	2	2	879
6	3	1	3	1	899
7	1	3	1	1	930
8	2	3	1	2	1320
9	3	2	3	3	912

```
s = [1 1 1 1 863;1 2 2 3 954;2 3 3 3 953;2 3 2 2 942;3 2 2 2 879;3 1 3
1 899;1 3 1 1 930;2 3 1 2 1320;3 2 3 3 912];
[r,ss] = zjsyjc(s,1)

    r =
            525        2383         349         449
              2           3           1           2
    ss =
           2747        1762        3113        2692
           3215        2745        2775        3141
           2690        4145        2764        2819
```

r 的第一行是每个因素的极差，反映的是该因素波动对整体质量波动的影响大小，从结果可看出，影响整体质量的大小顺序为 BADC。r 的第二行是相应因素的最优生产条件，在本题中选择的最大为最优，所以最优的生产条件为 $B_3A_2D_2C_1$ 。ss 的每一行是相应因素每个水平的数据和。

五、回归分析

1．一元多项式回归

如果从数据的散点图上发现 y 与 x 呈现较明显的二次（或高次）函数关系，则可以选多项式回归。

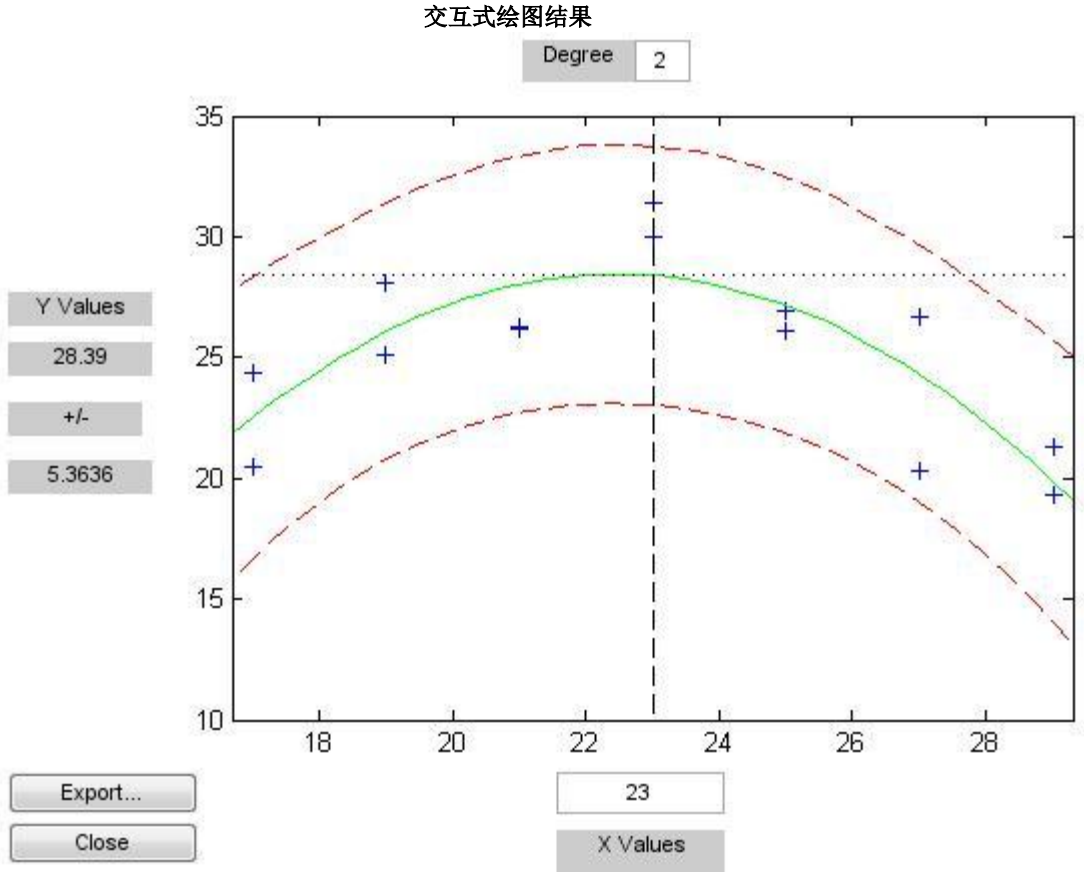
例 1．将 17~29 岁的运动员每两岁一组分为 7 组，每组两人测量其旋转定向能力，以及考察年龄对这种运动能力的影响。先得到一组数据如表 5 所示。

表 6.年龄与旋转定向能力数据

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

试建立两者之间的关系




```
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 26.7 21.3];
[p, s] = polyfit(x0, y0, 2);
p
[y, delta] = polyconf(p, x0, s);
y
polytool(x0, y0, 2)

p =
   -0.2003    8.9961  -72.5543
y =
  Columns 1 through 10
   22.4886   26.0582   28.0254   28.3900   27.1521   24.3118   19.8689
22.4886   26.0582   28.0254
  Columns 11 through 14
   28.3900   27.1521   24.3118   19.8689
```

2. 多元线性回归

在 MATLAB 统计工具箱中使用函数 regress 实现多元线性回归。

例 2．已知某湖泊八年来湖水中 COD 浓度实测值（y）与影响因素，如湖区工业产值（x1）、总人口数（x2）、捕鱼量（x3）、降水量（x4）的资料，建立污染物 y 的水质分析模型。

表 7.湖水浓度和影响因素数据表

X1	1.376	1.375	1.387	1.401	1.412	1.428	1.445	1.477
X2	0.450	0.475	0.485	0.500	0.535	0.545	0.550	0.575
X3	2.170	2.554	2.676	2.713	2.823	3.088	3.122	3.262
X4	0.8922	1.1610	0.5346	0.9589	1.0239	1.0499	1.1065	1.1387
y	5.19	5.30	5.60	5.82	6.00	6.06	6.45	6.95

```
x1 = [1.376 1.375   1.387   1.401   1.412   1.428   1.445   1.477];
x2 = [0.450 0.475   0.485   0.500   0.535   0.545   0.550   0.575];
x3 = [2.170 2.554   2.676   2.713   2.823   3.088   3.122   3.262];
x4 = [0.8922    1.1610  0.5346  0.9589  1.0239  1.0499  1.1065  1.1387];
y = [5.19   5.30    5.60    5.82    6.00    6.06    6.45    6.95];
x = [ones(8, 1), x1', x2', x3', x4'];
[b, bint, r, rint, stats] = regress(y', x);
b                           % 输出参数
bint                        % 各参数估计得置信区间
stats                       % 几个特殊统计量
```

```
b =
  -13.9849
   13.1920
    2.4228
    0.0754
   -0.1897
bint =
  -26.0019   -1.9679
    1.4130   24.9711
  -14.2808   19.1264
   -1.4859    1.6366
   -0.9638    0.5844
stats =
    0.9846   47.9654    0.0047    0.0123
```

所以，回归模型方程是：

$$\hat{y} = -13.9849 + 13.1920x_1 + 2.4227x_2 + 0.0754x_3 - 0.1897x_4$$

此外，由 stats 的值可知 R^2=0.9846，F=47.9654，P=0.0123。