

一、随机变量的分布

name	Distribution	Input Parameter A	Input Parameter B	Input Parameter C	Input Parameter D
'beta' or 'Beta'	Beta Distribution	a	b	—	—
'bino' or 'Binomial'	Binomial Distribution	n: number of trials	p: probability of success for each trial	—	—
'birnbaumsaunders'	Birnbaum-Saunders Distribution	θ	γ	—	—
'burr' or 'Burr'	Burr Type XII Distribution	α : scale parameter	c: shape parameter	k: shape parameter	—
'chi2' or 'Chisquare'	Chi-Square Distribution	ν : degrees of freedom	—	—	—
'exp' or 'Exponential'	Exponential Distribution	μ : mean	—	—	—
'ev' or 'Extreme Value'	Extreme Value Distribution	μ : location parameter	σ : scale parameter	—	—
'f' or 'F'	F Distribution	$\nu1$: numerator degrees of freedom	$\nu2$: denominator degrees of freedom	—	—
'gam' or 'Gamma'	Gamma Distribution	a: shape parameter	b: scale parameter	—	—
'gev' or 'Generalized Extreme Value'	Generalized Extreme Value Distribution	k: shape parameter	σ : scale parameter	μ : location parameter	—
'gp' or 'Generalized Pareto'	Generalized Pareto Distribution	k: tail index (shape) parameter	σ : scale parameter	μ : threshold (location) parameter	—
'geo' or 'Geometric'	Geometric Distribution	p: probability parameter	—	—	—
'hn' or 'Half Normal'	Half-Normal Distribution	μ : location	σ : scale	—	—
'hyge' or 'Hypergeometric'	Hypergeometric Distribution	M: size of the population	K: number of items with the desired characteristic in the population	n: number of samples drawn	—
'inversegaussian'	Inverse Gaussian Distribution	μ	λ	—	—
'logistic'	Logistic Distribution	μ	σ	—	—
'loglogistic'	Loglogistic Distribution	μ	σ	—	—
'logn' or 'Lognormal'	Lognormal Distribution	μ	σ	—	—
'nakagami'	Nakagami Distribution	μ	ω	—	—
'nbin' or 'Negative Binomial'	Negative Binomial Distribution	r: number of successes	p: probability of success in a single trial	—	—
'ncf' or 'Noncentral F'	Noncentral F Distribution	$\nu1$: numerator degrees of freedom	$\nu2$: denominator degrees of freedom	δ : noncentrality parameter	—
'nct' or 'Noncentral t'	Noncentral t Distribution	ν : degrees of freedom	δ : noncentrality parameter	—	—
'ncx2' or 'Noncentral Chi-square'	Noncentral Chi-Square Distribution	ν : degrees of freedom	δ : noncentrality parameter	—	—
'norm' or 'Normal'	Normal Distribution	μ : mean	σ : standard deviation	—	—
'poiss' or 'Poisson'	Poisson Distribution	λ : mean	—	—	—
'rayl' or 'Rayleigh'	Rayleigh Distribution	b: scale parameter	—	—	—
'rician'	Rician Distribution	s: noncentrality parameter	σ : scale parameter	—	—
'stable'	Stable Distribution	α : first shape parameter	β : second shape parameter	γ : scale parameter	δ : location parameter
't' or 'T'	Student's t Distribution	ν : degrees of freedom	—	—	—
'tlocationscale'	t Location-Scale Distribution	μ : location parameter	σ : scale parameter	ν : shape parameter	—
'unif' or 'Uniform'	Uniform Distribution (Continuous)	a: lower endpoint (minimum)	b: upper endpoint (maximum)	—	—
'unid' or 'Discrete Uniform'	Uniform Distribution (Discrete)	N: maximum observable value	—	—	—
'wbl' or 'Weibull'	Weibull Distribution	a: scale parameter	b: shape parameter	—	—

33种分布

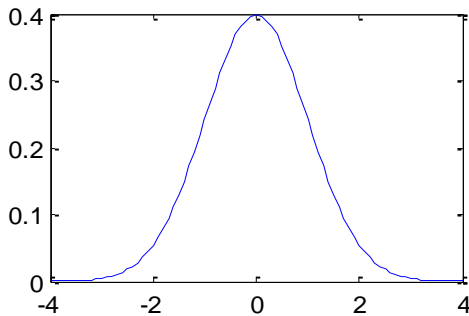
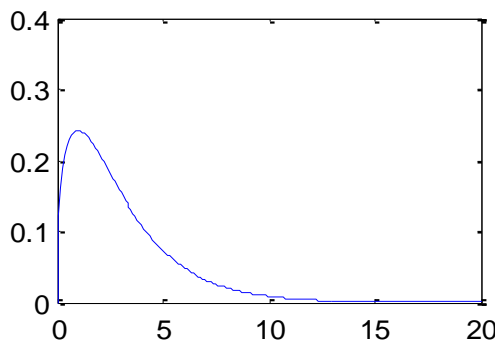
常用分布

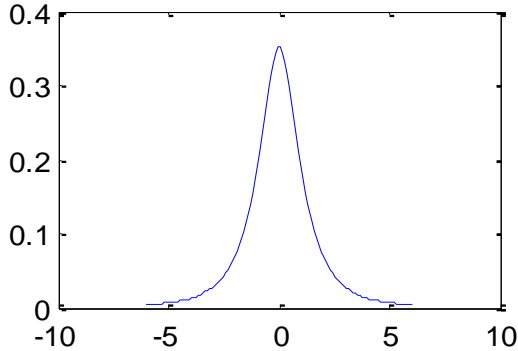
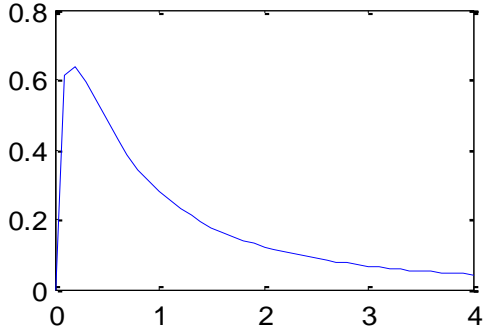
name	Distribution	Input Parameter A	Input Parameter B	Input Parameter C
'beta' or 'Beta'	Beta Distribution	a	b	—
'bino' or 'Binomial'	Binomial Distribution	n: number of trials	p: probability of success for each trial	—
'chi2' or 'Chisquare'	Chi-Square Distribution	v : degrees of freedom	—	—
'exp' or 'Exponential'	Exponential Distribution	μ : mean	—	—
'f' or 'F'	F Distribution	$v1$: numerator degrees of freedom	$v2$: denominator degrees of freedom	—
'gam' or 'Gamma'	Gamma Distribution	a: shape parameter	b: scale parameter	—
'geo' or 'Geometric'	Geometric Distribution	p: probability parameter	—	—
'logistic'	Logistic Distribution	μ	σ	—
'loglogistic'	Loglogistic Distribution	μ	σ	—
'logn' or 'Lognormal'	Lognormal Distribution	μ	σ	—
'nakagami'	Nakagami Distribution	μ	ω	—
'nbin' or 'Negative Binomial'	Negative Binomial Distribution	r: number of successes	p: probability of success in a single trial	—
'ncf' or 'Noncentral F'	Noncentral F Distribution	$v1$: numerator degrees of freedom	$v2$: denominator degrees of freedom	δ : noncentrality parameter
'nct' or 'Noncentral t'	Noncentral t Distribution	v : degrees of freedom	δ : noncentrality parameter	—
'ncx2' or 'Noncentral Chi-square'	Noncentral Chi-Square Distribution	v : degrees of freedom	δ : noncentrality parameter	—
'norm' or 'Normal'	Normal Distribution	μ : mean	σ : standard deviation	—
'poiss' or 'Poisson'	Poisson Distribution	λ : mean	—	—
'rayl' or 'Rayleigh'	Rayleigh Distribution	b: scale parameter	—	—
't' or 'T'	Student's t Distribution	v : degrees of freedom	—	—
'unif' or 'Uniform'	Uniform Distribution (Continuous)	a: lower endpoint (minimum)	b: upper endpoint (maximum)	—
'unid' or 'Discrete Uniform'	Uniform Distribution (Discrete)	N: maximum observable value	—	—
'wbl' or 'Weibull'	Weibull Distribution	a: scale parameter	b: shape parameter	—

常见分布函数表

name 的取值			函数说明
'beta'	或	'Beta'	Beta 分布
'bino'	或	'Binomial'	二项分布
'chi2'	或	'Chisquare'	卡方分布
'exp'	或	'Exponential'	指数分布
'f'	或	'F'	F 分布
'gam'	或	'Gamma'	GAMMA 分布
'geo'	或	'Geometric'	几何分布
'hyge'	或	'Hypergeometric'	超几何分布
'logn'	或	'Lognormal'	对数正态分布
'nbin'	或	'Negative Binomial'	负二项式分布
'ncf'	或	'Noncentral F'	非中心 F 分布
'nct'	或	'Noncentral t'	非中心 t 分布
'ncx2'	或	'Noncentral Chi-square'	非中心卡方分布
'norm'	或	'Normal'	正态分布
'poiss'	或	'Poisson'	泊松分布
'rayl'	或	'Rayleigh'	瑞利分布
't'	或	'T'	T 分布
'unif'	或	'Uniform'	均匀分布
'unid'	或	'Discrete Uniform'	离散均匀分布
'weib'	或	'Weibull'	Weibull 分布

四个重要的分布

分布	密度函数和图形	说明
正态分布 $N(\mu, \sigma^2)$	$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ 	$P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.68$ $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.95$ $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.997$
χ^2 分布 $\chi^2(n)$		若随机变量 X_1, X_2, \dots, X_n 相互独立, 都服从标准正态分布 $N(0, 1)$, 则随机变量 $Y = X_1^2 + X_2^2 + \dots + X_n^2$ 服从自由度为 n 的 χ^2 分布, 记为 $Y \sim \chi^2(n)$, Y 的均值为 n , 方差为 $2n$

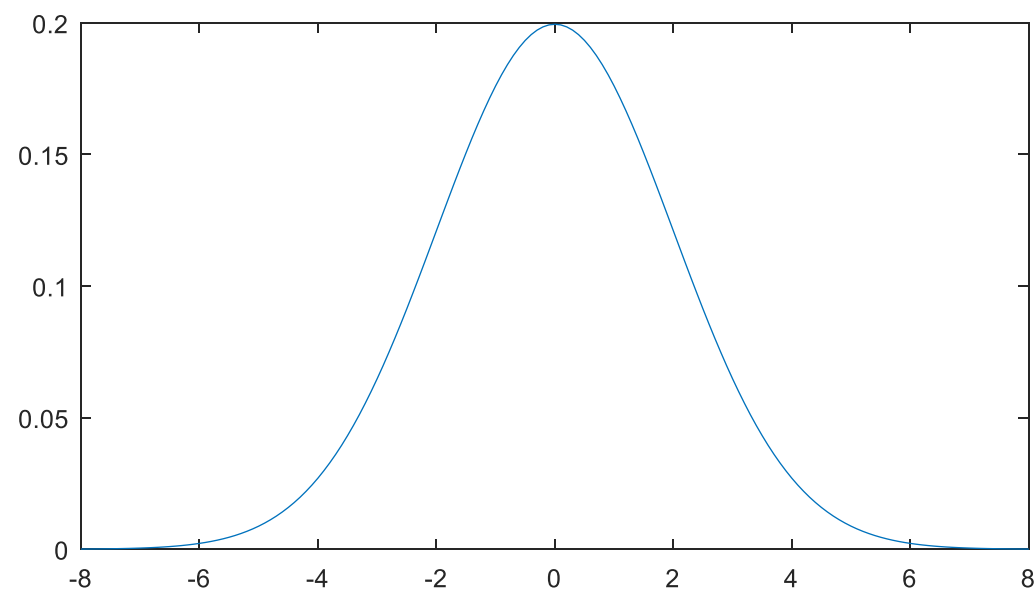
<p>t 分布</p> <p>t (n)</p>		<p>若 $X \sim N(0, 1)$, $Y \sim \chi^2(n)$, 且相互独立, 则随机变量</p> $T = \frac{X}{\sqrt{\frac{Y}{n}}}$ <p>服从自由度为 n 的 t 分布, 记为 $T \sim t(n)$.</p> <p>t 分布 t(20) 的密度函数曲线和 $N(0, 1)$ 的曲线形状相似. 理论上 $n \rightarrow \infty$ 时, $T \sim t(n) \rightarrow N(0, 1)$</p>
<p>F 分布</p> <p>F (n₁, n₂)</p>		<p>若 $X \sim \chi^2(n_1)$, $Y \sim \chi^2(n_2)$, 且相互独立, 则随机变量</p> $F = \frac{X/n_1}{Y/n_2}$ <p>服从自由度为 (n₁, n₂) 的 F 分布, 记作 $F \sim F(n_1, n_2)$</p>

二、密度函数、累计分布 等函数的计算

2.1 通用函数计算概率密度函数值

- 命令 通用函数计算概率密度函数值
- 函数 **pdf**
- 格式
 - $Y = \text{pdf}(\text{name}, x, A)$
 - $Y = \text{pdf}(\text{name}, x, A, B)$
 - $Y = \text{pdf}(\text{name}, x, A, B, C)$
- 说明
 - 返回在 $X=x$ 处、参数为A、B、C的概率密度值，对于不同的分布，参数个数可以不同；name为分布函数名，其取值见前表。


```
mu=0;  
  
sigma=2;  
  
x=-8:0.1:8;  
  
y = pdf( 'Normal' ,x,mu,sigma); %通用函数  
% 同上 y=normpdf(x,mu,sigma); %专用函数  
  
plot(x,y)
```



概率密度图

2.2 通用函数计算累积概率值

- 命令 通用函数cdf用来计算随机变量的概率之和（累积概率值）
- 函数 **cdf**
- 格式
 - **cdf('name', K, A)**
 - **cdf('name', K, A, B)**
 - **cdf('name', K, A, B, C)**
- 说明
 - 返回以name为分布、随机变量 $X \leq K$ 的概率之和的累积概率值，name的取值见表1“常见分布函数表”

```
mu=0;
```

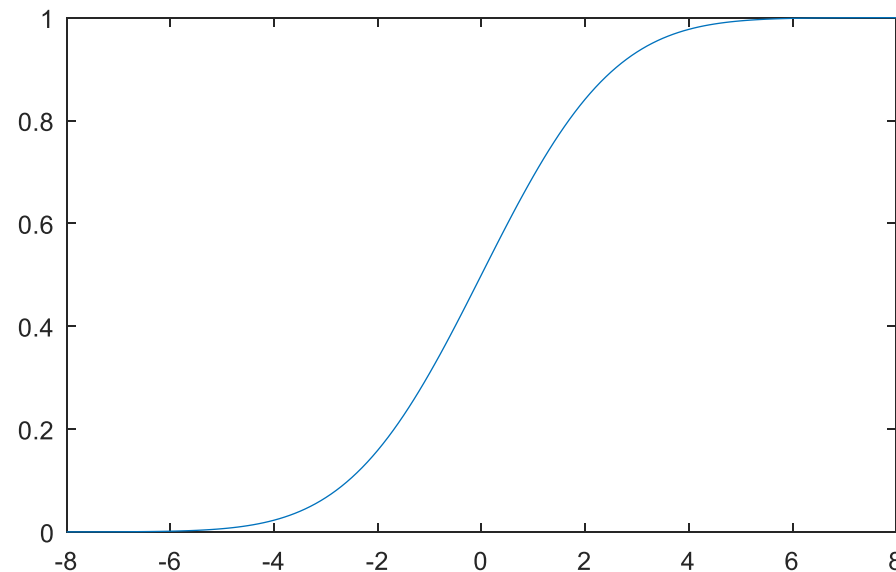
```
sigma=2;
```

```
x=-8:0.1:8;
```

```
y = cdf( 'Normal' ,x,mu,sigma); %通用
```

```
% y=normcdf(x,mu,sigma); %专用
```

```
plot(x,y)
```



累积概率图

2.3 通用函数计算逆累积分布函数值

- 命令 **icdf** 计算逆累积分布函数
- 格式
 - `icdf('name', P, A, B, C)`
 - 返回分布为name, 参数为A, B, C, 累积概率值为P的临界值
 - 如果 $P = \text{cdf}(\text{'name'}, x, A, B, C)$,
则 $x = \text{icdf}(\text{'name'}, P, A, B, C)$

%% 逆概率分布

mu=0;

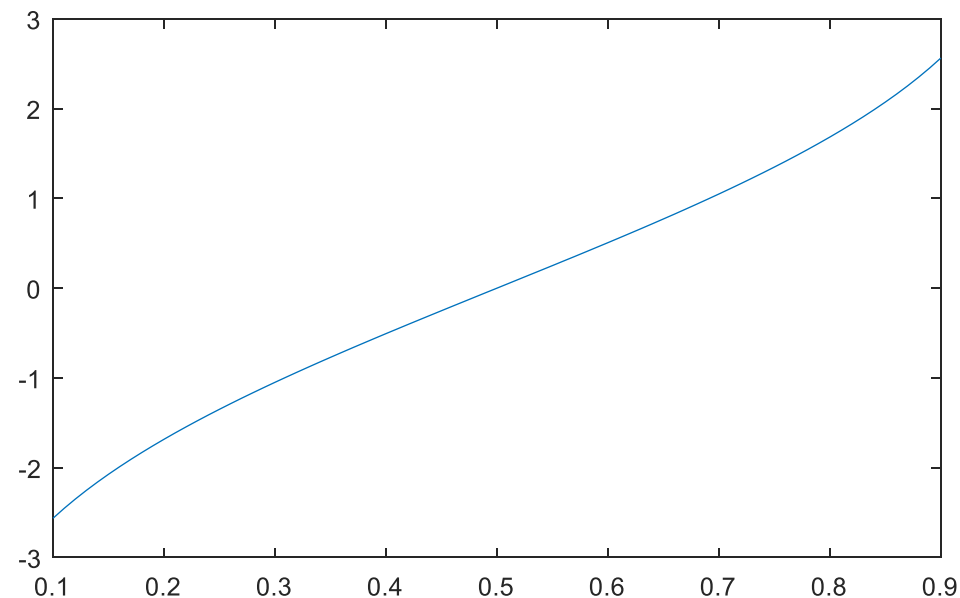
sigma=2;

p=0.1:0.01:0.9;

%x = icdf('Normal' ,p,mu,sigma); %通用

x=norminv(p,mu,sigma); %专用

plot(p,x)



逆概率分布图

三、 随机变量的数字特征

3.1 均值、中值

3.2 数据比较

3.3 期望

3.4 方差

3.5 常见分布的期望和方差

3.6 协方差与相关系数

3.1 平均值、中值

- 命令 利用mean求算术平均值

- 格式

- mean(X)
 - X为向量，返回X中各元素的平均值
- mean(A)
 - A为矩阵，返回A中各列元素的平均值构成的向量
- mean(A,dim)
 - 在给定的维数内的平均值

- 说明

- X为向量时，算术平均值的数学含义是 $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ ，即样本均值。
- nanmean(X)
 - X为向量，返回X中除NaN外元素的算术平均值。

- 命令 利用median计算中值（中位数）
- 格式
 - median(X)
 - X为向量，返回X中各元素的中位数。
 - median(A)
 - A为矩阵，返回A中各列元素的中位数构成的向量。
 - median(A,dim)
 - 求给出的维数内的中位数
 - nanmedian(X)
 - X为向量，返回X中除NaN外元素的中位数。

- 命令 利用geomean计算几何平均数
- 格式
 - $M = \text{geomean}(X)$
 - X 为向量，返回 X 中各元素的几何平均数。
 - $M = \text{geomean}(A)$
 - A 为矩阵，返回 A 中各列元素的几何平均数构成的向量。
- 说明
 - 几何平均数的数学含义是 $M = \left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}}$ ，
其中：样本数据非负，主要用于对数正态分布。

3.2 数据比较

- 命令 排序
- 格式
 - `Y=sort(X)`
 - X为向量，返回X按由小到大排序后的向量。
 - `Y=sort(A)`
 - A为矩阵，返回A的各列按由小到大排序后的矩阵。
 - `[Y,I]=sort(A)`
 - Y为排序的结果，I中元素表示Y中对应元素在A中位置。
 - `sort(A,dim)`
 - 在给定的维数dim内排序
- 说明
 - 若X为复数，则通过 $|X|$ 排序。

命令 求最大值与最小值之差（极差）

函数 **range**

格式

- $Y = \text{range}(X)$
 - X 为向量，返回 X 中的最大值与最小值之差。
- $Y = \text{range}(A)$
 - A 为矩阵，返回 A 中各列元素的最大值与最小值之差。

3.3 期望

- 命令 计算样本均值
- 函数 **mean**

设随机变量X的分布率为:

X	-2	-1	0	1	2
P	0.3	0.1	0.2	0.1	0.3

求 $E(X)$ 和 $E(X^2+1)$

3.4 方差

- 命令 求**样本**方差

- 函数 **var**
- $$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2$$

- 格式

- $D = \text{var}(X)$

- $\text{var}(X) = s^2$, 若 X 为向量, 则返回向量的样本方差。

- $D = \text{var}(A)$

- A 为矩阵, 则 D 为 A 的列向量的样本方差构成的行向量。

- $D = \text{var}(X, 1)$

- 返回向量 (矩阵) X 的简单方差 (即置前因子为 $1/n$ 的方差)

- $D = \text{var}(X, w)$

- 返回向量 (矩阵) X 的以 w 为权重的方差

- 命令 求标准差

- 函数 **std**

- 格式

$$std = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2}$$

- std(X)

- 返回向量（矩阵）X的样本标准差（置前因子为1/(n-1)）：

- std(X,1)

- 返回向量（矩阵）X的标准差（置前因子为1/n）

- std(X, 0)

- 与std (X)相同

- std(X, flag, dim)

- 返回向量（矩阵）中维数为dim的标准差值，

- 其中flag=0时，置前因子为1/(n-1)；否则置前因子为1/n。

- y = nanstd(X)

- 若X为含有元素NaN的向量，则返回除NaN外的元素的标准差，

- 若X为含元素NaN的矩阵，则返回各列除NaN外的标准差构成的向量。

- 命令 样本的偏斜度
- 函数 **skewness**
- 格式
 - $y = \text{skewness}(X)$
 - X 为向量，返回 X 的元素的偏斜度； X 为矩阵，返回 X 各列元素的偏斜度构成的行向量。
 - $y = \text{skewness}(X, \text{flag})$
 - $\text{flag}=0$ 表示偏斜纠正， $\text{flag}=1$ （默认）表示偏斜不纠正。
- 说明
 - 偏斜度样本数据关于均值不对称的一个测度，如果偏斜度为负，说明均值左边的数据比均值右边的数据更散；如果偏斜度为正，说明均值右边的数据比均值左边的数据更散，因而正态分布的偏斜度为 0；
 - 偏斜度的定义：
$$y = \frac{E(x - \mu)^3}{\sigma^3}$$
 - 其中： μ 为 x 的均值， σ 为 x 的标准差， $E(.)$ 为期望值算子

命令 均匀分布（连续）的期望和方差

函数 **unifstat**

格式

- $[M,V] = \text{unifstat}(A,B)$
 - A、B为标量时，就是区间上均匀分布的期望和方差
 - A、B也可作为向量或矩阵，则M、V也是向量或矩阵。

3.5 常见分布的期望和方差

命令 正态分布的期望和方差

函数 **normstat** (*stat 专用函数)

格式

- [M,V] = normstat(MU,SIGMA)

MU、SIGMA可为标量也可为向量或矩阵，

则M=MU， V=SIGMA2

命令 二项分布的均值和方差

函数 **binostat**

格式

- [M,V] = binostat(N,P)
 - N， P为二项分布的两个参数， 可为标量也可为向量或矩阵

函数名	调用形式	注 释
unifstat	[M,V]=unifstat (a, b)	均匀分布(连续)的期望和方差, M 为期望, V 为方差
unidstat	[M,V]=unidstat (n)	均匀分布(离散)的期望和方差
expstat	[M,V]=expstat (p, Lambda)	指数分布的期望和方差
nomstat	[M,V]=nomstat(mu,sigma)	正态分布的期望和方差
chi2stat	[M,V]=chi2stat (x, n)	卡方分布的期望和方差
tstat	[M,V]=tstat (n)	t 分布的期望和方差
fstat	[M,V]=fstat (n ₁ , n ₂)	F 分布的期望和方差
gamstat	[M,V]=gamstat (a, b)	γ 分布的期望和方差
betastat	[M,V]=betastat (a, b)	β 分布的期望和方差
lognstat	[M,V]=lognstat (mu, sigma)	对数正态分布的期望和方差
nbinstat	[M,V]=nbinstat (R, P)	负二项式分布的期望和方差
ncfstat	[M,V]=ncfstat (n ₁ , n ₂ , delta)	非中心 F 分布的期望和方差
nctstat	[M,V]=nctstat (n, delta)	非中心 t 分布的期望和方差
ncx2stat	[M,V]=ncx2stat (n, delta)	非中心卡方分布的期望和方差
raylstat	[M,V]=raylstat (b)	瑞利分布的期望和方差
Weibstat	[M,V]=weibstat (a, b)	韦伯分布的期望和方差
Binostat	[M,V]=binostat (n,p)	二项分布的期望和方差
Geostat	[M,V]=geostat (p)	几何分布的期望和方差
hygestat	[M,V]=hygestat (M,K,N)	超几何分布的期望和方差
Poisstat	[M,V]=poisstat (Lambda)	泊松分布的期望和方差

3.6 协方差与相关系数

命令 协方差

函数 **cov**

格式

- cov(X)
 - 求向量X的协方差
- cov(A)
 - 求矩阵A的协方差矩阵，该协方差矩阵的对角线元素是A的各列的方差，即：
 $\text{var}(A) = \text{diag}(\text{cov}(A))$ 。
- cov(X,Y)
 - X,Y为等长列向量，等同于cov([X Y])。

For two random variable vectors A and B , the covariance is defined as

$$\text{cov}(A, B) = \frac{1}{N-1} \sum_{i=1}^N (A_i - \mu_A)^* (B_i - \mu_B)$$

命令 相关系数

函数 **corrcoef** $\rho(A, B) = \frac{\text{cov}(A, B)}{\sigma_A \sigma_B} .$

格式

- `corrcoef(X,Y)`
 - 返回列向量X,Y的相关系数，等同于`corrcoef([X Y])`。
- `corrcoef (A)`
 - 返回矩阵A的列向量的相关系数矩阵

四、统计作图

- 4.1 经验累积分布函数图形
- 4.2 最小二乘拟合直线
- 4.3 绘制正态分布概率图形
- 4.4 样本数据的盒图
- 4.5 给当前图形加一条参考线
- 4.6 在当前图形中加入一条多项式曲线
- 4.7 附加有正态密度曲线的直方图
- 4.8 在指定的界线之间画正态密度曲线

4.1 经验累积分布函数图形

- 函数 `cdfplot`
- 格式
 - `cdfplot(X)`
 - 作样本 X （向量）的累积分布函数图形
 - `h = cdfplot(X)`
 - h 表示曲线的句柄
 - `[h,stats] = cdfplot(X)`
 - $stats$ 表示样本的一些特征

```
rng default; % For reproducibility
```

```
y = normrnd(0,3,100,1);
```

```
cdfplot(y)
```

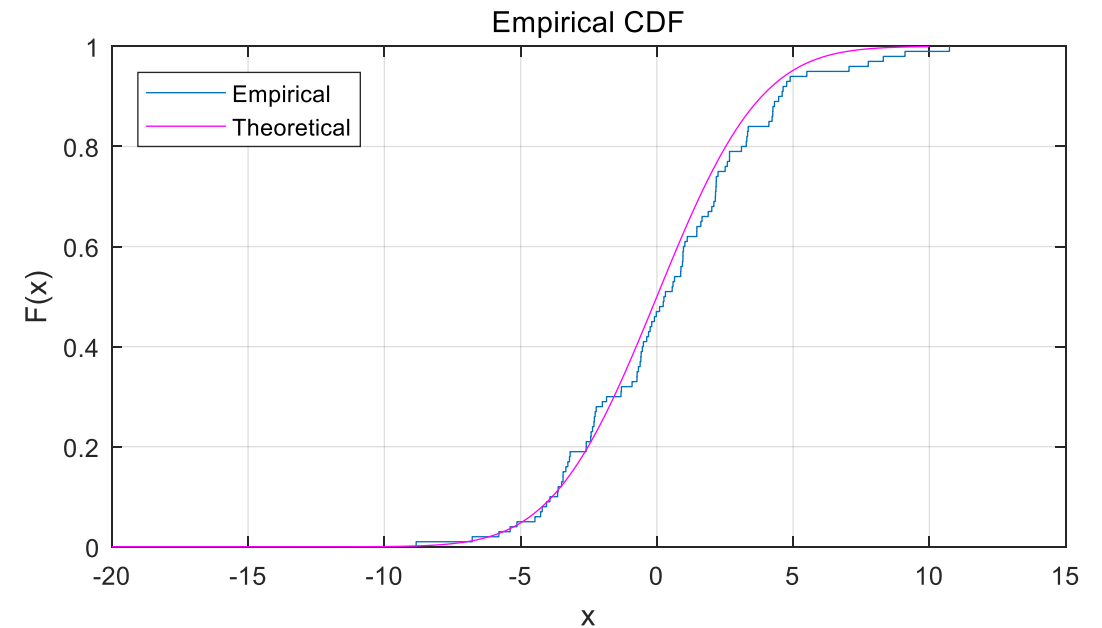
```
hold on
```

```
x = -20:0.1:10;
```

```
f = normcdf(x,0,3);
```

```
plot(x,f,'m')
```

```
legend('Empirical','Theoretical','Location','NW')
```



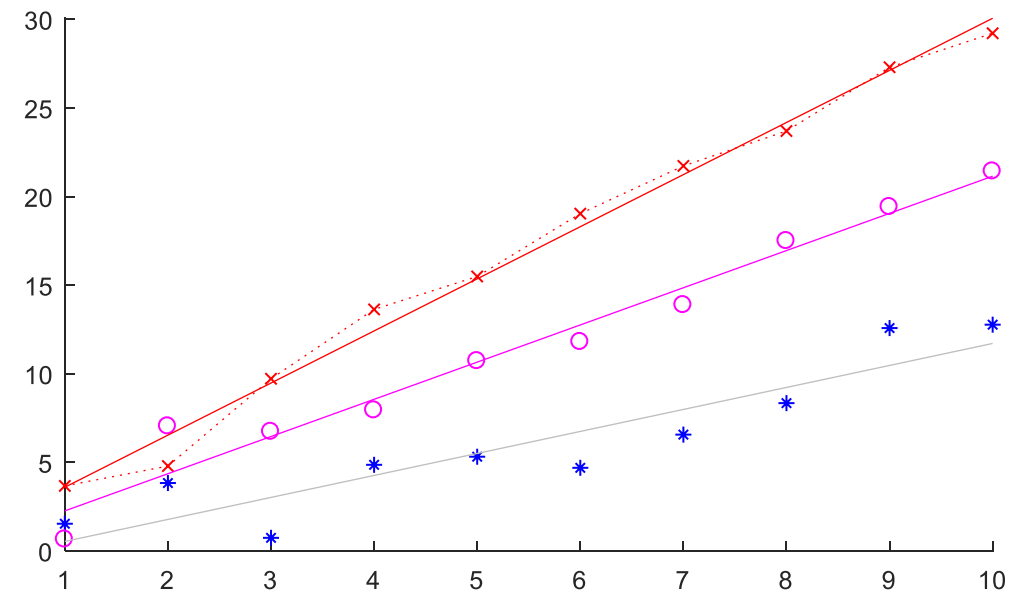
4.2 最小二乘拟合直线

函数 lsline

格式

- lsline
 - 最小二乘拟合直线
- h = lsline
 - h为直线的句柄

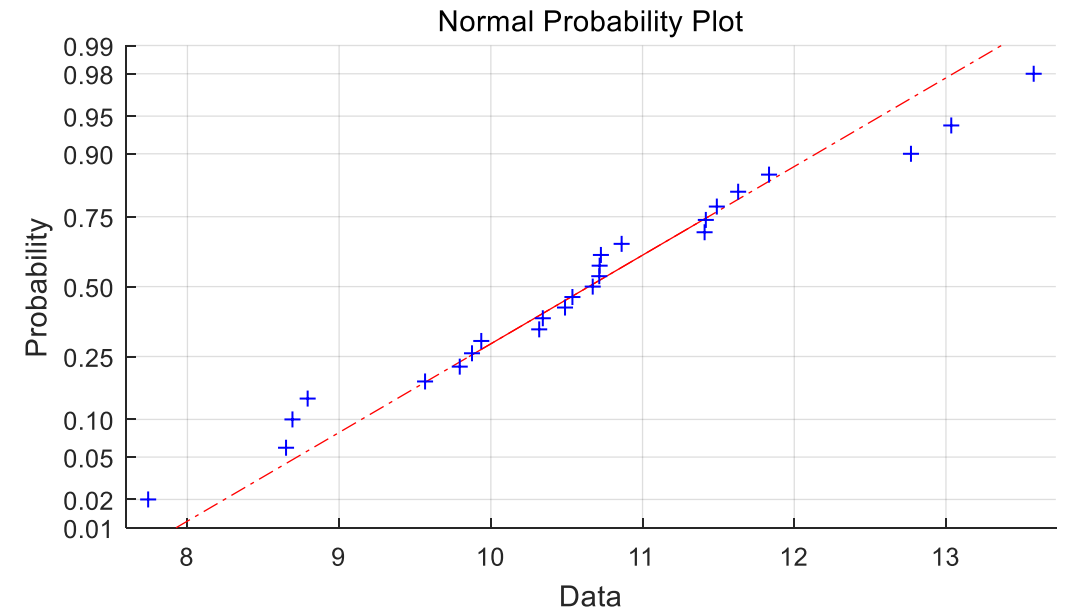
```
x = 1:10;  
rng default; % For reproducibility  
figure;  
y1 = x + randn(1,10);  
scatter(x,y1,25,'b','*')  
hold on  
y2 = 2*x + randn(1,10);  
plot(x,y2,'mo')  
y3 = 3*x + randn(1,10);  
plot(x,y3,'rx:')  
Isline
```



4.3 绘制正态分布概率图形

- 函数 `normplot` (推广: `*plot` 专用函数)
- 格式
 - `normplot(X)`
 - 若X为向量, 则显示正态分布概率图形,
 - 若X为矩阵, 则显示每一列的正态分布概率图形。
 - `h = normplot(X)`
 - 返回绘图直线的句柄
- 说明
 - 样本数据在图中用“+”显示;
 - 如果数据来自正态分布, 则图形显示为直线, 而其它分布可能在图中产生弯曲。

```
%%  
rng default; % For reproducibility  
x = normrnd(10,1,25,1);  
%Create a normal probability plot of the sample data.  
figure;  
normplot(x)
```



4.4 样本数据的盒图

函数 `boxplot`

格式

- `boxplot(X)`
 - 产生矩阵X的每一列的盒图和“须”图，“须”是从盒的尾部延伸出来，并表示盒外数据长度的线，如果“须”的外面没有数据，则在“须”的底部有一个点。
- `boxplot(X,notch)`
 - 当`notch=1`时，产生一凹盒图，`notch=0`时产生一矩箱图。
- `boxplot(X,notch,'sym')`
 - `sym`表示图形符号，默认值为“+”。
- `boxplot(X,notch,'sym',vert)`
 - 当`vert=0`时，生成水平盒图，`vert=1`时，生成竖直盒图（默认值`vert=1`）。
- `boxplot(X,notch,'sym',vert,whis)`
 - `whis`定义“须”图的长度，默认值为1.5，若`whis=0`则`boxplot`函数通过绘制`sym`符号图来显示盒外的所有数据值。

```
load carsmall
```

```
% Create a box plot of the miles per gallon (MPG)
```

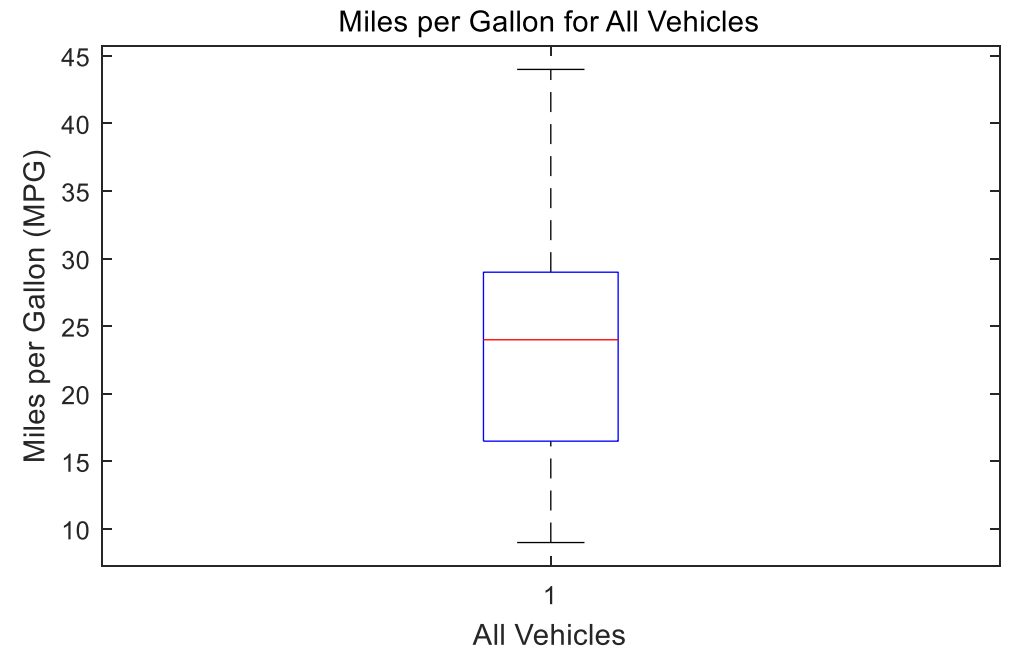
```
% measurements. Add a title and label the axes.
```

```
boxplot(MPG)
```

```
xlabel('All Vehicles')
```

```
ylabel('Miles per Gallon (MPG)')
```

```
title('Miles per Gallon for All Vehicles')
```



4.5 给当前图形加一条参考线

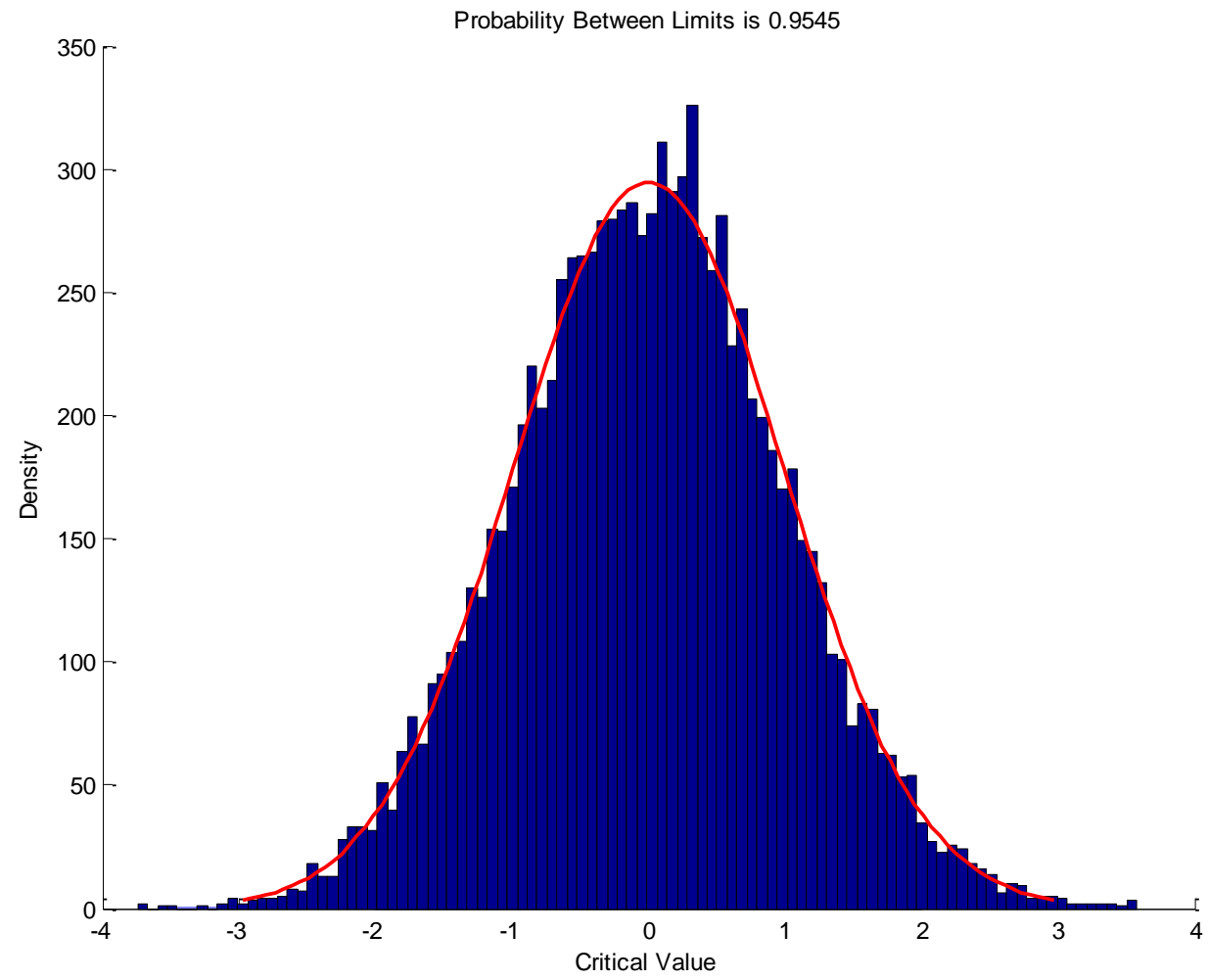
- 函数 `refline`
- 格式
 - `refline(slope,intercept)`
 - `slope`表示直线斜率, `intercept`表示截距
 - `refline(slope)`
 - `slope=[a b]`, 图中加一条直线: $y=b+ax$ 。

4.6 在当前图形中加入一条多项式曲线

- 函数 refcurve
- 格式
 - $h = \text{refcurve}(p)$
 - 在图中加入一条多项式曲线,
 - h 为曲线的环柄,
 - p 为多项式系数向量, $p=[p_1, p_2, p_3, \dots, p_n]$, 其中 p_1 为最高幂项系数。

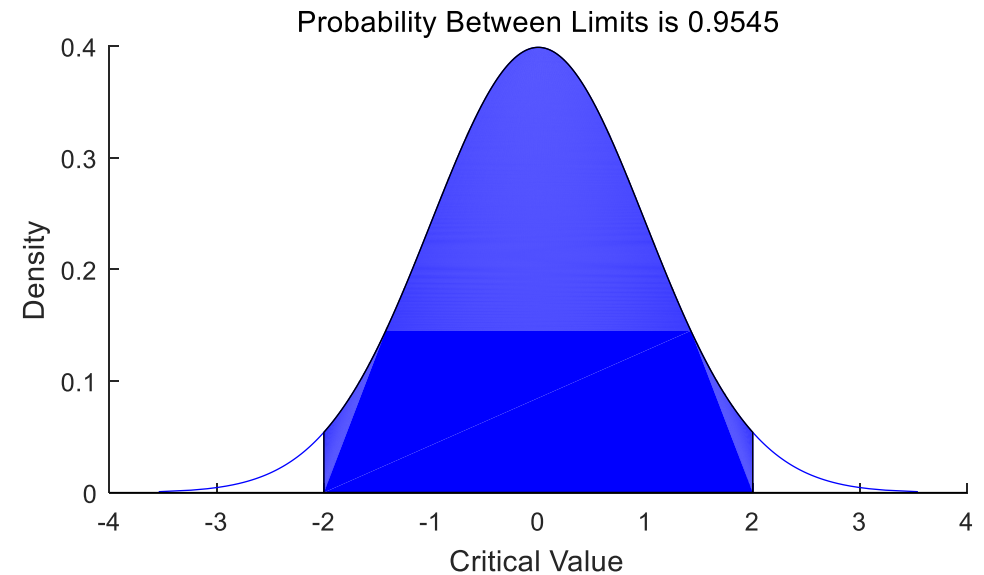
4.7 附加有正态密度曲线的直方图

- 函数 histfit
- 格式
 - histfit(data)
 - data为向量，返回直方图和正态曲线。
 - histfit(data,nbins)
 - nbins指定bar的个数，缺省时为data中数据个数的平方根。



4.8 在指定的界线之间画正态密度曲线

- 函数 `normspec`
- 格式
 - `p = normspec(specs,mu,sigma)`
 - specs指定界线,
 - mu,sigma为正态分布的参数
 - p为样本落在上、下界之间的概率。



`normspec([-2,2],0,1)`

五、 参数估计

5.1 常见分布的参数估计

5.2 非线性模型置信区间预测

5.3 对数似然函数

5.1 常见分布的参数估计

命令 β 分布的参数a和b的最大似然估计值和置信区间

函数 `betafit`

格式

- `PHAT=betafit(X)`
- `[PHAT,PCI]=betafit(X,ALPHA)`
 - PHAT为样本X的 β 分布的参数a和b的估计量
 - PCI为样本X的 β 分布参数a和b的置信区间，是一个 2×2 矩阵，其第1例为参数a的置信下界和上界，第2例为b的置信下界和上界，ALPHA为显著水平， $(1-\alpha) \times 100\%$ 为置信度。

- 随机产生100个 β 分布数据，相应的分布参数真值为4和3。求该样本的最大似然估计值和置信度为99%的置信区间。

```
>>X = betarnd (4,3,100,1);  
>>[PHAT,PCI] = betafit(X,0.01)  
PHAT =  
      3.9010      2.6193  
PCI =  
      2.5244      1.7488  
      5.2776      3.4898
```

命令 正态分布的参数估计

函数 normfit

格式

- `[muhat,sigmahat,muci,sigmaci] = normfit(X)`
- `[muhat,sigmahat,muci,sigmaci] = normfit(X,alpha)`
 - muhat, sigmahat分别为正态分布的参数 μ 和 σ 的估计值,
 - muci, sigmaci分别为置信区间, 其置信度为 $(1-\alpha)\times 100\%$;
 - alpha给出显著水平 α , 缺省时默认为0.05, 即置信度为95%。

- 有两组(每组100个元素)正态随机数据, 其均值为10, 均方差为2, 求95%的置信区间和参数估计值。

```
>>r = normrnd (10,2,100,2);  
>>[mu,sigma,muci,sigmaci] = normfit(r)  
mu =  
    10.1455    10.0527           %各列的均值的估计值  
sigma =  
    1.9072    2.1256           %各列的均方差的估计值  
muci =  
    9.7652    9.6288           %各列均值的置信区间  
    10.5258   10.4766  
sigmaci =  
    1.6745    1.8663           %各列均方差的置信区间  
    2.2155    2.4693
```

- 分别使用金球和铂球测定引力常数
 - (1) 用金球测定观察值为: 6.683 6.681 6.676 6.678 6.679 6.672
 - (2) 用铂球测定观察值为: 6.661 6.661 6.667 6.667 6.664
- 设测定值总体为, μ 和 σ 为未知。对(1)、(2)两种情况分别求 μ 和 σ 的置信度为0.9的置信区间。

```
X=[6.683 6.681 6.676 6.678 6.679 6.672];
Y=[6.661 6.661 6.667 6.667 6.664];
[mu,sigma,muci,sigmaci]=normfit(X,0.1) %金球测定的估计
[MU,SIGMA,MUCI,SIGMACI]=normfit(Y,0.1) %铂球测定的估计
```

命令 利用mle函数进行参数估计

函数 mle

格式

- `phat=mle('dist', X)`
 - `dist`为分布函数名, 如: `beta`(分布)、`binom` (二项分布) 等,
 - `X`为数据样本
 - 返回用`dist`指定分布的最大似然估计值
- `[phat, pci]=mle ('dist', X)`
 - 置信度为95%
- `[phat, pci]=mle ('dist', X, alpha)`
 - 置信度由`alpha`确定, `alpha`为显著水平 α , $(1-\alpha)\times 100\%$ 为置信度
- `[phat, pci]=mle ('dist', X, alpha, pl)`
 - 仅用于二项分布, `pl`为试验次数。

参数估计函数表

函数名	调用形式	函数说明
binofit	PHAT=binofit(X, N) [PHAT, PCI]=binofit(X,N) [PHAT, PCI]=binofit(X, N, ALPHA)	二项分布的概率的最大似然估计 置信度为 95% 的参数估计和置信区间 返回水平 α 的参数估计和置信区间
poissfit	Lambdahat=poissfit(X) [Lambdahat, Lambdaci]=poissfit(X) [Lambdahat, Lambdaci]=poissfit(X, ALPHA)	泊松分布的参数最大似然估计 置信度为 95% 的参数估计和置信区间 返回水平 α 的 λ 参数和置信区间
normfit	[muhat,sigmahat,muci,sigmaci]=normfit(X) [muhat,sigmahat,muci,sigmaci]=normfit(X, ALPHA)	正态分布的最大似然估计, 置信度为 95% 返回水平 α 的期望、方差值和置信区间
betafit	PHAT=betafit(X) [PHAT, PCI]=betafit(X, ALPHA)	返回 β 分布参数 a 和 b 的最大似然估计 返回最大似然估计值和水平 α 的置信区间
unifit	[ahat,bhat]=unifit(X) [ahat,bhat,ACI,BCI]=unifit(X) [ahat,bhat,ACI,BCI]=unifit(X, ALPHA)	均匀分布参数的最大似然估计 置信度为 95% 的参数估计和置信区间 返回水平 α 的参数估计和置信区间
expfit	muhat=expfit(X) [muhat,muci]=expfit(X) [muhat,muci]=expfit(X,alpha)	指数分布参数的最大似然估计 置信度为 95% 的参数估计和置信区间 返回水平 α 的参数估计和置信区间
gamfit	phat=gamfit(X) [phat,pci]=gamfit(X) [phat,pci]=gamfit(X,alpha)	γ 分布参数的最大似然估计 置信度为 95% 的参数估计和置信区间 返回最大似然估计值和水平 α 的置信区间
weibfit	phat=weibfit(X) [phat,pci]=weibfit(X) [phat,pci]=weibfit(X,alpha)	韦伯分布参数的最大似然估计 置信度为 95% 的参数估计和置信区间 返回水平 α 的参数估计及其区间估计
Mle	phat=mle('dist',data) [phat,pci]=mle('dist',data) [phat,pci]=mle('dist',data,alpha) [phat,pci]=mle('dist',data,alpha,p1)	分布函数名为 dist 的最大似然估计 置信度为 95% 的参数估计和置信区间 返回水平 α 的最大似然估计值和置信区间 仅用于二项分布, p1 为试验总次数

5.2 非线性模型置信区间预测

- 命令 高斯—牛顿法的非线性最小二乘数据拟合
- 函数 nlinfit
- 格式
 - `beta = nlinfit(X,y,FUN,beta0)`
 - 返回在FUN中描述的非线性函数的系数。
 - FUN为用户提供形如 $\hat{y} = f(\beta, X)$ 的函数，该函数返回已给初始参数估计值 β 和自变量X的y的预测值。
 - 若X为矩阵，则X的每一列为自变量的取值，y是一个相应的列向量。
 - 如果FUN中使用了@，则表示函数的柄。
 - `[beta,r,J] = nlinfit(X,y,FUN,beta0)`
 - beta为拟合系数，r为残差，J为Jacobi矩阵，beta0为初始预测值。

- 混凝土的抗压强度随养护时间的延长而增加，现将一批混凝土作成12个试块，记录了养护日期 x （日）及抗压强度 y （kg/cm²）的数据：

养护时间 $x=[2\ 3\ 4\ 5\ 7\ 9\ 12\ 14\ 17\ 21\ 28\ 56]$

抗压强度 $y=[35\ 42\ 47\ 53\ 59\ 65\ 68\ 73\ 76\ 82\ 86\ 99]+r$

r 为一个 $[-0.5, 0.5]$ 之间的随机向量

建立非线性回归模型，对得到的模型和系数进行检验。

%模型为： $y=a+k_1\exp(m*x)+k_2\exp(-m*x)$;

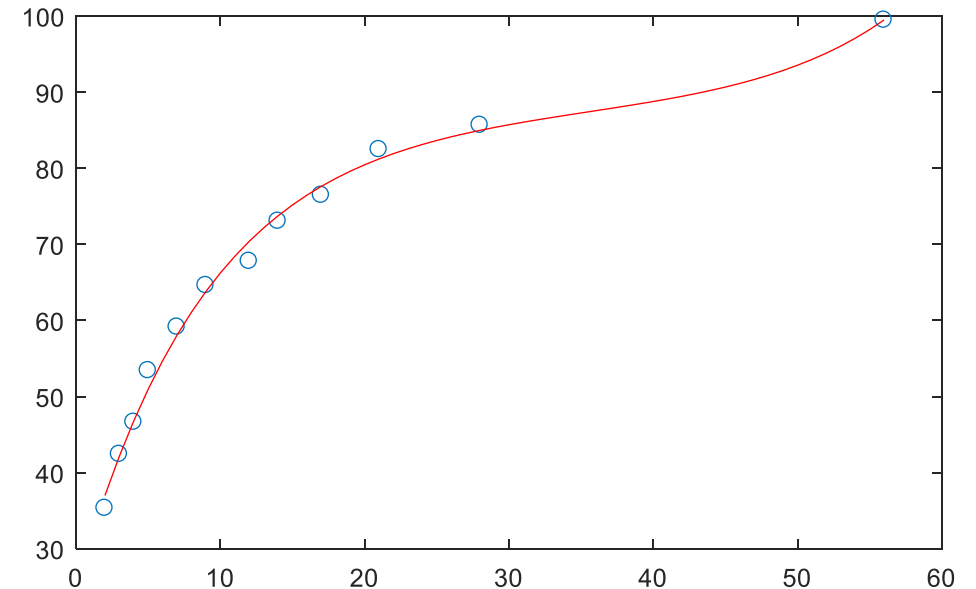
```

clc;clear;
x=[2 3 4 5 7 9 12 14 17 21 28 56];
y1=[35 42 47 53 59 65 68 73 76 82 86 99];
r=rand(1,12)-0.5;
y=y1+r
myfunc=@(beta,x) beta(1)+beta(2)*exp(beta(4)*x)+beta(3)*exp(-beta(4)*x);
beta=nlinfit(x,y,myfunc,[0.5 0.5 0.5 0.5]);
a=beta(1),k1=beta(2),k2=beta(3),m=beta(4)
%test the model
xx=min(x):max(x);
yy=a+k1*exp(m*xx)+k2*exp(-m*xx);
plot(x,y,'o',xx,yy,'r')

```

beta =

87.4409 0.0289 -62.6478 0.1080



命令 非线性模型的参数估计的置信区间

函数 nlparci (Nonlinear regression parameter confidence intervals)

格式

- `ci = nlparci(beta,r,J)`
 - 返回置信度为95%的置信区间,
 - `beta`为非线性最小二乘法估计的参数值,
 - `r`为残差,
 - `J`为Jacobian矩阵。
 - `nlparci`可以用`nlinfit`函数的输出作为其输入。

```

clc;clear;
x=[2 3 4 5 7 9 12 14 17 21 28 56];
y1=[35 42 47 53 59 65 68 73 76 82 86 99];
r=rand(1,12)-0.5;
y=y1+r
myfunc=@(beta,x) beta(1)+beta(2)*exp(beta(4)*x)+beta(3)*exp(-beta(4)*x);
[beta, resids, J]=nlinfit(x,y,myfunc,[0.5 0.5 0.5 0.5]);
ci = nlparci(beta, resids, J)

```

结果:

beta =

```

87.4409    0.0289   -62.6478    0.1080

```

ci =

```

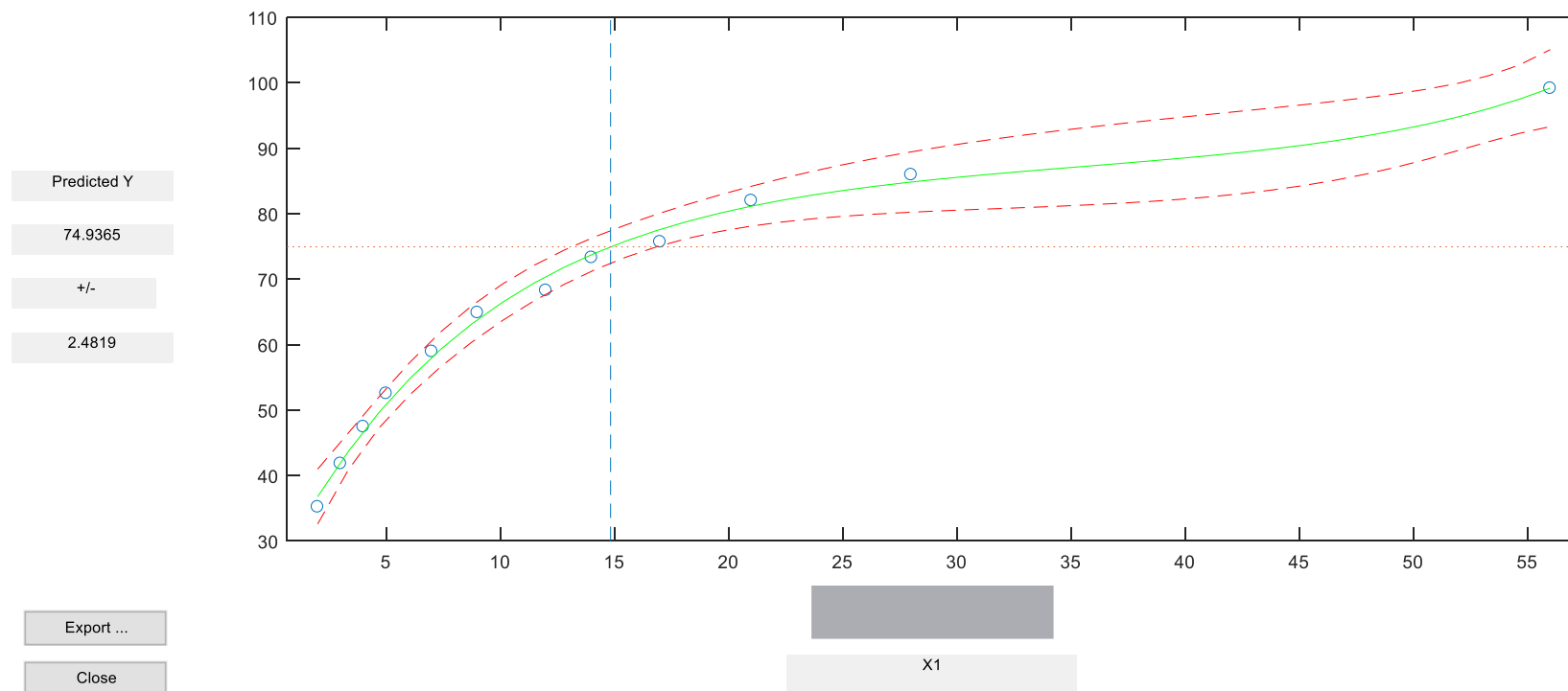
83.0937  91.6618
-0.0026  0.0567
-66.9701 -58.6733
0.0851   0.1333

```

- 命令 非线性拟合和显示交互图形
- 函数 nlintool
- 格式
 - nlintool(x,y,FUN,beta0)
 - 返回数据(x,y)的非线性曲线的预测图形, 它用2条红色曲线预测全局置信区间。
 - beta0为参数的初始预测值, 置信度为95%。
 - nlintool(x,y,FUN,beta0,alpha)
 - 置信度为 $(1-\alpha)\times 100\%$

%% 参数的区间估计

```
clc;clear;  
x=[2 3 4 5 7 9 12 14 17 21 28 56];  
y1=[35 42 47 53 59 65 68 73 76 82 86 99];  
r=rand(1,12)-0.5;  
y=y1+r  
myfunc=@(beta,x) beta(1)+beta(2)*exp(beta(4)*x)+beta(3)*exp(-beta(4)*x);  
nlintool(x,y,myfunc,[0.5 0.5 0.5 0.5]);
```



命令 非线性模型置信区间预测

函数 `nlpredci`

格式

- `ypred = nlpredci(FUN,inputs,beta,r,J)`
 - `ypred` 为预测值，`FUN`与前面相同，`beta`为给出的适当参数，`r`为残差，`J`为Jacobian矩阵，`inputs`为非线性函数中的独立变量的矩阵值。
- `[ypred,delta] = nlpredci(FUN,inputs,beta,r,J)`
 - `delta`为非线性最小二乘法估计的置信区间长度的一半，当`r`长度超过`beta`的长度并且`J`的列满秩时，置信区间的计算是有效的。`[ypred-delta,ypred+delta]`为置信度为95%的不同步置信区间。
- `ypred = nlpredci(FUN,inputs,beta,r,J,alpha,'simopt','predopt')`
 - 控制置信区间的类型，置信度为 $100(1-\alpha)\%$ 。
 - `'simopt' = 'on'` 或 `'off'` (默认值)分别表示同步或不同步置信区间。
 - `'predopt'='curve'` (默认值) 表示输入函数值的置信区间，
 - `'predopt'='observation'` 表示新响应值的置信区间。`nlpredci`可以用`nlinfit`函数的输出作为其输入。

续前例，在[15 25 45]处的预测函数值和置信区间一半宽度

```
clc;clear;  
x=[2 3 4 5 7 9 12 14 17 21 28 56];  
y1=[35 42 47 53 59 65 68 73 76 82 86 99];  
r=rand(1,12)-0.5;  
y=y1+r  
myfunc=inline('beta(1)+beta(2)*exp(beta(4)*x)+beta(3)*exp(-beta(4)*x)','beta','x');  
[beta, resids, J]=nlinfit(x,y,myfunc,[0.5 0.5 0.5 0.5]);  
[ypred, delta]=nlpredci(myfunc, [15 15 45], beta, resids, J)
```

ypred =

75.1180 75.1180 90.5652

delta =

1.5842 1.5842 3.9479

5.3 对数似然函数

- 命令 beta分布的对数似然函数
- 函数 Betalike
- 格式
 - `logL=betalike(params,data)`
 - 返回负 β 分布的对数似然函数，params为向量[a, b]，是 β 分布的参数，data为样本数据。
 - `[logL,info]=betalike(params,data)`
 - 返回Fisher逆信息矩阵info。如果params中输入的参数是极大似然估计值，那么info的对角元素为相应参数的渐近方差。
- 说明
 - betalike是 β 分布最大似然估计的实用函数。似然函数假设数据样本中所有的元素相互独立。因为betalike返回负对数似然函数，用fmins函数最小化betalike与最大似然估计的功能是相同的。

命令 正态分布的对数似然函数

函数 normlike

格式

- `logL=normlike(params,data)`
 - 返回由给定样本数据data确定的、负正态分布的、参数为params(即[μ , σ])的对数似然函数值。
- `[logL,info]=normlike(params,data)`
 - 返回Fisher逆信息矩阵info。如果params中输入的参数是极大似然估计值, 那么info的对角元素为相应参数的渐近方差。

六、 假设检验

- 6.1 已知方差，单个正态总体的均值 μ 的假设检验（U检验法）
- 6.2 未知，单个正态总体的均值 μ 的假设检验(t检验法)
- 6.3 两个正态总体均值差的检验（t检验）
- 6.4 两个总体一致性的检验——秩和检验
- 6.5 两个总体中位数相等的假设检验——符号秩检验
- 6.6 两个总体中位数相等的假设检验——符号检验
- 6.7 正态分布的拟合优度测试
- 6.8 正态分布的拟合优度测试
- 6.9 单个样本分布的 Kolmogorov-Smirnov 测试
- 6.10 两个样本具有相同的连续分布的假设检验

6.1 已知 σ^2 ，单个正态总体的均值 μ 的假设检验（U检验法）

- 函数 ztest

- $h = \text{ztest}(x, m, \text{sigma})$
 - x 为正态总体的样本， m 为均值， sigma 为标准差，显著性水平为0.05(默认值)
- $h = \text{ztest}(x, m, \text{sigma}, \alpha)$
 - 显著性水平为 α
- $[h, \text{sig}, \text{ci}, \text{zval}] = \text{ztest}(x, m, \text{sigma}, \alpha, \text{tail})$
 - sig 为观察值的概率，当 sig 为小概率时则对原假设提出质疑，
 - ci 为真正均值 μ 的 $1-\alpha$ 置信区间， zval 为统计量的值。
 - 若 $h=0$ ，表示在显著性水平 α 下，不能拒绝原假设；
 - 若 $h=1$ ，表示在显著性水平 α 下，可以拒绝原假设。

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

- 原假设

- 若 $\text{tail}=0$ ，表示备择假设： $H_0: \mu = \mu_0 = m$ （默认，双边检验）； $H_0: \mu_1 \neq \mu_0 = m$
- $\text{tail}=1$ ，表示备择假设： $H_1: \mu_1 > \mu_0 = m$ （单边检验）；
- $\text{tail}=-1$ ，表示备择假设： $H_2: \mu_1 < \mu_0 = m$ （单边检验）。

- 某车间用一台包装机包装葡萄糖，包得的袋装糖重是一个随机变量，它服从正态分布。当机器正常时，其均值为0.5公斤，标准差为0.015。某日开工后检验包装机是否正常，随机地抽取所包装的糖9袋，称得净重为（公斤）0.497, 0.506, 0.518, 0.524, 0.498, 0.511, 0.52, 0.515, 0.512；问机器是否正常？
- 总体 μ 和 σ 已知，该问题是当 σ^2 为已知时，在水平 $\alpha = 0.05$ 下，根据样本值判断 $\mu=0.5$ 还是 $\mu \neq 0.5$
- 原假设： $H_0: \mu = \mu_0 = 0.5$ 备择假设： $H_1: \mu \neq 0.5$ $H_2: \mu > 0.5$ $H_3: \mu < 0.5$

```
>> X=[0.497,0.506,0.518,0.524,0.498,0.511,0.52,0.515,0.512];
>> [h1,sig1,ci1,zval1]=ztest(X,0.5,0.015,0.05,0)
>> [h2,sig2,ci2,zval2]=ztest(X,0.5,0.015,0.05,1)
>> [h3,sig3,ci3,zval3]=ztest(X,0.5,0.015,0.05,-1)
```

结果

h = 1	h2 = 1	h3 = 0
sig = 0.0248	sig2 = 0.0124	sig3 = 0.9876
ci = 0.5014 0.5210	ci 2 = 0.5030 Inf	ci3 = -Inf 0.5194
zval = 2.2444	zval2 = 2.2444	zval3 = 2.2444

6.2 未知 σ^2 , 单个正态总体的均值 μ 的假设检验 (t检验法)

- 函数 ttest
- 格式
 - $h = \text{ttest}(x, m)$
 - x 为正态总体的样本, m 为均值 μ_0 , 显著性水平为0.05
 - 若 $h=0$, 表示在显著性水平 α 下, 不能拒绝原假设;
 - 若 $h=1$, 表示在显著性水平 α 下, 可以拒绝原假设。
 - $h = \text{ttest}(x, m, \alpha)$
 - α 为给定显著性水平
 - $[h, \text{sig}, \text{ci}] = \text{ttest}(x, m, \alpha, \text{tail})$
 - sig 为观察值的概率, 当 sig 为小概率时则对原假设提出质疑,
 - ci 为真正均值 μ 的 $1-\alpha$ 置信区间。
- 原假设 $H_0: \mu = \mu_0 = m$
 - $\text{tail}=0$, 表示备择假设: $H_0: \mu_1 \neq \mu_0 = m$ (默认, 双边检验) ;
 - $\text{tail}=1$, 表示备择假设: $H_1: \mu_1 > \mu_0 = m$ (单边检验) ;
 - $\text{tail}=-1$, 表示备择假设: $H_2: \mu_1 < \mu_0 = m$ (单边检验) 。

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

- 某种电子元件的寿命X（以小时计）服从正态分布， μ, σ^2 均未知。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）？

- σ^2 未知，在 $\alpha = 0.05$ 水平下检验假设：

$$H_0 : \mu \leq \mu_0 = 225 \quad H_1 : \mu > 225$$

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

结果

h = 0

sig = 0.2570

ci = 198.2321 Inf

6.3 两个正态总体均值差的检验 (t检验)

- 两个正态总体方差未知但等方差时，比较两正态总体样本均值的假设检验
- 函数 ttest2
 - $[h, sig, ci] = ttest2(X, Y)$
 - X, Y为两个正态总体的样本，显著性水平为0.05
 - 若h=0，表示在显著性水平alpha下，不能拒绝原假设；
 - 若h=1，表示在显著性水平alpha下，可以拒绝原假设。
 - $[h, sig, ci] = ttest2(X, Y, alpha)$
 - $[h, sig, ci] = ttest2(X, Y, alpha, tail)$
 - sig为当原假设为真时得到观察值的概率，当sig为小概率时则对原假设提出质疑，ci为真正均值 μ 的1-alpha置信区间。
- 原假设： $\mu_1 = \mu_2$ (为X为期望值，为Y的期望值)
 - tail=0，表示备择假设： $H_0: \mu_1 = \mu_2$ (默认，双边检验)；
 - tail=1，表示备择假设： $H_0: \mu_1 > \mu_2$ (单边检验)；
 - tail=-1，表示备择假设： $H_1: \mu_1 < \mu_2$ (单边检验)。

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}} \sim t(n_1 + n_2 - 2)$$

- 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产率，试验是在同一只平炉上进行的。每炼一炉钢时除操作方法外，其他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，其产率分别为

(1) 标准方法: 78.1 72.4 76.2 74.3 77.4 78.4 76.0 75.5 76.7 77.3

(2) 新方法: 79.1 81.0 77.3 79.1 80.0 79.1 79.1 77.3 80.2 82.1

设这两个样本相互独立，且分别来自正态总体

和 $N(\mu_2, \sigma^2)$ ， μ_1, μ_2, σ^2 均未知。问建议的新操作方法能否提高产率？（取 $\alpha=0.05$ ）

- 两个总体方差不变时，在 $\alpha=0.05$ 水平下检验假设： $N(\mu_1, \sigma^2)$

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 < \mu_2$$

```
>> 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];
>> [h,sig,ci]=ttest2(X,Y,0.05,-1)
```

6.4 两个总体一致性的检验——秩和检验

- 函数 ranksum

- 格式

- $p = \text{ranksum}(x, y, \alpha)$
 - x 、 y 为两个总体的样本，可以不等长，
 - α 为显著性水平(significance level)
 - P 为两个总体样本 X 和 Y 为一致的显著性概率，若 P 接近于0，则不一致较明显。
- $[p, h] = \text{ranksum}(x, y, \alpha)$
 - h 为检验结果， $h=0$ 表示 X 与 Y 的总体差别不显著；
 - $h=1$ 表示 X 与 Y 的总体差别显著
- $[p, h, \text{stats}] = \text{ranksum}(x, y, \alpha)$
 - stats 中包括：ranksum为秩和统计量的值以及 $zval$ 为过去计算 p 的正态统计量的值

- 某商店为了确定向公司A或公司B购买某种商品，将A和B公司以往的各次进货的次品率进行比较，数据如下：

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 4.1 11.0 9.7 6.9 3.6 4.8 5.6 8.4 10.1 5.5 12.3

设两样本独立。问两公司的商品的质量有无显著差异，取 $\alpha=0.05$ 。

- 设 μ_A ， μ_B 别为A、B两个公司的商品次品率总体的均值。在水平 $\alpha=0.05$ 下检验假设：

$$H_0 : \mu_A = \mu_B$$

$$H_1 : \mu_A \neq \mu_B$$

```
>> 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 4.1 11.0 9.7 6.9 3.6 4.8 5.6 8.4 10.1 5.5 12.3];
>> [p,h,stats]=ranksum(A,B,0.05)
```

结果：

p = 0.8282

h = 0

Stats.zval: -0.2171

stats.ranksum: 116

6.5 两个总体中位数相等的假设检验 ——符号秩检验

- 函数 signrank
- 格式
 - $p = \text{signrank}(X, Y, \alpha)$
 - X 、 Y 为两个总体的样本，长度必须相同，
 - α 为显著性水平，
 - P 两个样本 X 和 Y 的中位数相等的概率， p 接近于0则可对原假设质疑。
 - $[p, h] = \text{signrank}(X, Y, \alpha)$
 - h 为检验结果： $h=0$ 表示 X 与 Y 的中位数之差不显著， $h=1$ 表示 X 与 Y 的中位数之差显著。
 - $[p, h, \text{stats}] = \text{signrank}(x, y, \alpha)$
 - stats 中包括： signrank 为符号秩统计量的值以及 zval 为过去计算 p 的正态统计量的值。

6.6 两个总体中位数相等的假设检验 ——符号检验

- 函数 `signtest`
- 格式
 - `p=signtest(X, Y, alpha)`
 - X、Y为两个总体的样本，长度必须相同，
 - alpha为显著性水平，
 - P两个样本X和Y的中位数相等的概率，p接近于0则可对原假设质疑。
 - `[p, h]=signtest(X, Y, alpha)`
 - h为检验结果：h=0表示X与Y的中位数之差不显著，h=1表示X与Y的中位数之差显著。
 - `[p,h,stats] = signtest(X,Y,alpha)`
 - stats中sign为符号统计量的值

6.7 正态分布的拟合优度测试

- 函数 `jbtest`
- 格式
 - `H = jbtest(X)`
 - 对输入向量X进行Jarque-Bera测试，显著性水平为0.05。
 - H为测试结果，若H=0，则可以认为X是服从正态分布的；若H=1，则可以否定X服从正态分布。
 - X为大样本，对于小样本用lillietest函数
 - `H = jbtest(X,alpha)`
 - 在水平alpha下施行 Jarque-Bera 测试，alpha在0和1之间。
 - `[H,P,JBSTAT,CV] = jbtest(X,alpha)`
 - P为接受假设的概率值，P越接近于0，则可以拒绝是正态分布的原假设；
 - JBSTAT为测试统计量的值，CV为是否拒绝原假设的临界值。

6.8 正态分布的拟合优度测试

函数 `lillietest`

格式

- `H = lillietest(X)`
 - 对输入向量X进行lillietest测试，显著性水平为0.05。
 - H为测试结果，若H=0，则可以认为X是服从正态分布的；若H=1，则可以否定X服从正态分布。
- `H = lillietest(X,alpha)`
 - 在水平alpha进行lillietest测试，alpha在0.01和0.2之间。
- `[H,P,LSTAT,CV] = lillietest(X,alpha)`
 - P为接受假设的概率值，P越接近于0，则可以拒绝是正态分布的原假设；
 - LSTAT为测试统计量的值，CV为是否拒绝原假设的临界值。

```
rng default; % For reproducibility
X = normrnd(0, 3, 10000, 1);
alpha=0.05;
[H, P, LSTAT, CV] = lillietest(X, alpha)
```

```
结果
H =    0
P =    0.2758
LSTAT =    0.0071
CV =    0.0091
```

6.9 单个样本分布的 Kolmogorov-Smirnov 测试

- 函数 kstest
- 格式
 - $H = \text{kstest}(X)$
 - 测试向量X是否服从标准正态分布，测试水平为5%。
 - 原假设为X服从标准正态分布。若 $H=0$ 则不能拒绝原假设， $H=1$ 则可以拒绝原假设。
 - $H = \text{kstest}(X, \text{cdf})$
 - 指定累积分布函数为cdf的测试(cdf=[]时表示标准正态分布)，测试水平为5%
 - $H = \text{kstest}(X, \text{cdf}, \alpha)$
 - α 为指定测试水平
 - $[H, P, \text{KSSTAT}, \text{CV}] = \text{kstest}(X, \text{cdf}, \alpha)$
 - P为原假设成立的概率，KSSTAT为测试统计量的值，CV为是否接受假设的临界值。

```
load examgrades;  
x = grades(:, 1);  
test_cdf = makedist('tlocation', 'mu', 75, 'sigma', 10, 'nu', 1)  
[h, p] = kstest(x, 'CDF', test_cdf, 'Alpha', 0.05)
```

结果
h = 1
p = 0.0021

6.10 两个样本具有相同的连续分布的假设检验

函数 `kstest2`

格式

- `H = kstest2(X1,X2)`
 - 测试向量X1与X2是具有相同的连续分布，测试水平为5%。
 - 原假设为具有相同连续分布。测试结果为H，若H=0，表示应接受原假设；若H=1，表示可以拒绝原假设。这是Kolmogorov-Smirnov测试方法。
- `H = kstest2(X1,X2,alpha)`
 - alpha为测试水平
- `[H,P,KSSTAT] = kstest(X,cdf,alpha)`
 - 与指定累积分布cdf相同的连续分布，
 - P为假设成立的概率，
 - KSSTAT为测试统计量的值。

七、 方差分析

- 7.1 单因素方差分析
- 7.2 双因素方差分析

7.1 单因素方差分析

单因素方差分析是比较两组或多组数据的均值，它返回原假设——均值相等的概率

函数 `anova1`

- `p = anova1(X)`
 - `X`的各列为彼此独立的样本观察值，其元素个数相同，`p`为各列均值相等的概率值，若`p`值接近于0，则原假设受到怀疑，说明至少有一列均值与其余列均值有明显不同。
 - `anova1`函数产生两个图：标准的方差分析表图和盒图。
- `p = anova1(X,group)`
 - `X`和`group`为向量且`group`要与`X`对应
- `p = anova1(X,group,'displayopt')`
 - `displayopt=on/off`表示显示与隐藏方差分析表图和盒图
- `[p,table] = anova1(...)` % `table`为方差分析表
- `[p,table,stats] = anova1(...)` % `stats`为分析结果的构造

anova1分析前条件检验

(1) 正态性检验

`[h,p]=lillitest(x)` %x为单因素某水平的数据

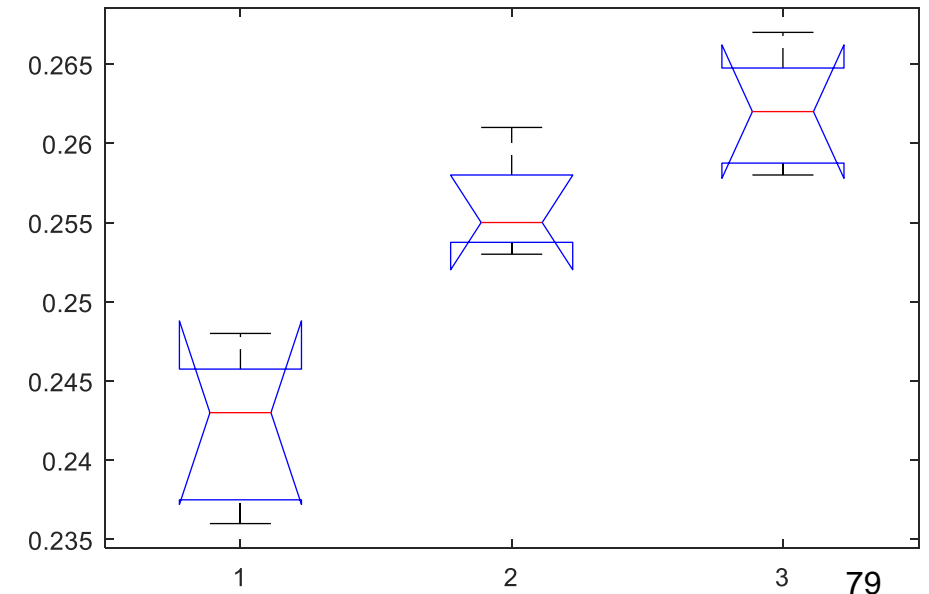
(2) 方差齐性检验(Bartlett test)

`[p,stats] = vartestn(A)` %A中各列为各水平下的数据

- 设有3台机器，用来生产规格相同的铝合金薄板。取样测量薄板的厚度，精确至‰厘米。得结果如下：
 机器1: 0.236 0.238 0.248 0.245 0.243
 机器2: 0.257 0.253 0.255 0.254 0.261
 机器3: 0.258 0.264 0.259 0.267 0.262
 检验各台机器所生产的薄板的厚度有无显著的差异？

```
>> X=[0.236 0.238 0.248 0.245 0.243; 0.257 0.253 0.255 0.254 0.261;  
0.258 0.264 0.259 0.267 0.262];  
>> [p,table,stats]=anova1(X') %均衡情形
```

Source	SS	df	MS	F	Prob>F
Columns	0.00105	2	0.00053	32.92	1.34305e-05
Error	0.00019	12	0.00002		
Total	0.00125	14			



- 建筑横梁强度的研究：3000磅力量作用在一英寸的横梁上来测量横梁的挠度，三种材料的测试强度是
钢筋：82 86 79 83 84 85 86 87;
合金1：74 82 78 75 76 77;
合金2：79 79 77 78 82 79
检验这些合金强度有无明显差异？

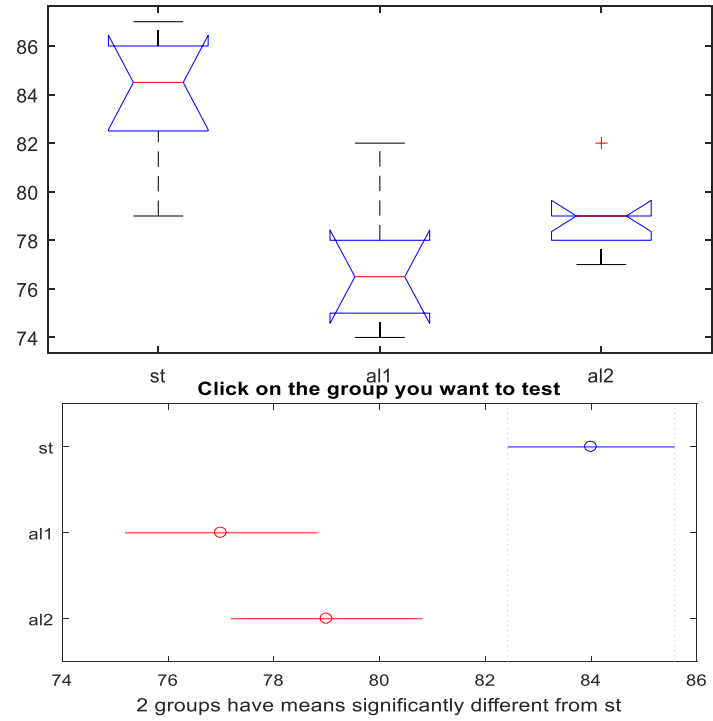
```
strength = [82 86 79 83 84 85 86 87 74 82 78 75 76 77 79 79 77 78 82 79];  
alloy = {'st','st','st','st','st','st','st','st', 'al1','al1','al1','al1','al1','al1','al2','al2','al2','al2','al2','al2'};  
[p,table,stats] = anova1(strength,alloy,'on') %此时容许不平衡  
c=multcompare(stats) %组组比较
```

Source	SS	df	MS	F	Prob>F

Groups	184.8	2	92.4	15.4	0.0002
Error	102	17	6		
Total	286.8	19			

C =

组序号	组	置信上限	组均值差	置信下限	检验p值
1.0000	2.0000	3.6064	7.0000	10.3936	0.0002
1.0000	3.0000	1.6064	5.0000	8.3936	0.0040
2.0000	3.0000	-5.6280	-2.0000	1.6280	0.3560



7.2 双因素方差分析

- 函数 `anova2`

- 格式

- `p = anova2(X, reps)`
- `p = anova2(X, reps, 'displayopt')`
- `[p, table] = anova2(...)`
- `[p, table, stats] = anova2(...)`

- 说明

- 执行平衡的双因素试验的方差分析来比较X中两个或多个列（行）的均值，不同列的数据表示因素A的差异，不同行的数据表示另一因素B的差异。如果行列对有多于一个的观察点，则变量reps指出每一单元观察点的数目，每一单元包含reps行，如右上图，则reps=2
- 其余参数与单因素方差分析参数相似。

A=1	A=2	
X ₁₁₁	X ₁₁₂	}B = 1
X ₁₂₁	X ₁₂₂	
X ₂₁₁	X ₂₁₂	}B = 2
X ₂₂₁	X ₂₂₂	
X ₃₁₁	X ₃₁₂	}B = 3
X ₃₂₁	X ₃₂₂	

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

推进器（B）		B1	B2	B3
燃料A	A1	58.2000	56.2000	65.3000
		52.6000	41.2000	60.8000
	A2	49.1000	54.1000	51.6000
		42.8000	50.5000	48.4000
	A3	60.1000	70.9000	39.2000
		58.3000	73.2000	40.7000
	A4	75.8000	58.2000	48.7000
		71.5000	51.0000	41.4000

- 考察推进器和燃料这两个因素对射程是否有显著的影响？

```

A=[ 58.2000    56.2000    65.3000
    52.6000    41.2000    60.8000
    49.1000    54.1000    51.6000
    42.8000    50.5000    48.4000
    60.1000    70.9000    39.2000
    58.3000    73.2000    40.7000
    75.8000    58.2000    48.7000
    71.5000    51.0000    41.4000];
[p,tbl,stats] = anova2(A,2)
[c1,m1]=multcompare(stats,'estimate','column')
[c2,m2]=multcompare(stats,'estimate','row')

```

Source	SS	df	MS	F	Prob>F

Columns	370.98	2	185.49	9.39	0.0035
Rows	261.68	3	87.225	4.42	0.026
Interaction	1768.69	6	294.782	14.93	0.0001
Error	236.95	12	19.746		
Total	2638.3	23			

stats =

包含以下字段的 struct:

```

source: 'anova2'
sigmasq: 19.7458
colmeans: [58.5500 56.9125 49.5125]
coln: 8
rowmeans: [55.7167 49.4167 57.0667 57.7667]
rown: 6
inter: 1
pval: 6.1511e-05
df: 12

```

c1 =

1.0000	2.0000	-4.2900	1.6375	7.5650	0.7469
1.0000	3.0000	3.1100	9.0375	14.9650	0.0041
2.0000	3.0000	1.4725	7.4000	13.3275	0.0153

m1 =

58.5500	1.5711
56.9125	1.5711
49.5125	1.5711

c2 =

1.0000	2.0000	-1.3168	6.3000	13.9168	0.1188
1.0000	3.0000	-8.9668	-1.3500	6.2668	0.9511
1.0000	4.0000	-9.6668	-2.0500	5.5668	0.8535
2.0000	3.0000	-15.2668	-7.6500	-0.0332	0.0489
2.0000	4.0000	-15.9668	-8.3500	-0.7332	0.0304
3.0000	4.0000	-8.3168	-0.7000	6.9168	0.9925

m2 =

55.7167	1.8141
49.4167	1.8141
57.0667	1.8141
57.7667	1.8141

```
load popcorn
popcorn
[p, tbl] = anova2(popcorn, 3)
```

popcorn =

5.5000	4.5000	3.5000
5.5000	4.5000	4.0000
6.0000	4.0000	3.0000
6.5000	5.0000	4.0000
7.0000	5.5000	5.0000
7.0000	5.0000	4.5000

```
p = 0.0000 0.0001 0.7462
tbl = 6×6 cell 数组
    'Source'      'SS'      'df'      'MS'      'F'      'Prob>F'
    'Columns'     [15.7500] [ 2]     [7.8750] [56.7000] [7.6790e-07]
    'Rows'        [ 4.5000] [ 1]     [4.5000] [32.4000] [1.0037e-04]
    'Interaction' [ 0.0833] [ 2]     [0.0417] [ 0.3000] [ 0.7462]
    'Error'       [ 1.6667] [12]     [0.1389] []         []
    'Total'       [ 22]     [17]     []         []         []
```

ANOVA Table					
Source	SS	df	MS	F	Prob>F
Columns	15.75	2	7.875	56.7	0
Rows	4.5	1	4.5	32.4	0.0001
Interaction	0.0833	2	0.04167	0.3	0.7462
Error	1.6667	12	0.13889		
Total	22	17			

多因素方差分析 (anovan)

[p,tbl,stats] = anovan(y,group)

```
%%  
y = [52.7 57.5 45.9 44.5 53.0 57.0 45.9 44.0]';  
g1 = [1 2 1 2 1 2 1 2];  
g2 = {'hi'; 'hi'; 'lo'; 'lo'; 'hi'; 'hi'; 'lo'; 'lo'};  
g3 = {'may'; 'may'; 'may'; 'may'; 'june'; 'june'; 'june'; 'june'};  
[p, tbl, stats] = anovan(y, {g1, g2, g3})  
%%
```

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F

X1	3.781	1	3.781	0.82	0.4174
X2	199.001	1	199.001	42.95	0.0028
X3	0.061	1	0.061	0.01	0.914
Error	18.535	4	4.634		
Total	221.379	7			

%% 三要素方差分析

```
y = [52.7 57.5 45.9 44.5 53.0 57.0 45.9 44.0]';  
g1 = [1 2 1 2 1 2 1 2];  
g2 = {'hi'; 'hi'; 'lo'; 'lo'; 'hi'; 'hi'; 'lo'; 'lo'};  
g3 = {'may'; 'may'; 'may'; 'may'; 'june'; 'june'; 'june'; 'june'};  
model=[1 0 0;0 1 0;0 0 1;1 1 0;0 1 1]; %主因素g1、 g2 、 g3和交叉g1g2、 g2g3设置  
[p, tbl, stats] = anovan(y, {g1, g2, g3}, 'model', model, 'varnames',{'g1','g2','g3'})  
%%
```

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F

g1	3.781	1	3.781	33.99	0.0282
g2	199.001	1	199.001	1788.78	0.0006
g3	0.061	1	0.061	0.55	0.5354
g1*g2	18.301	1	18.301	164.51	0.006
g2*g3	0.011	1	0.011	0.1	0.7806
Error	0.223	2	0.111		
Total	221.379	7			

8.1 数据的预处理

- 数据的平滑 【smooth;smooths;medfilt1】
- 数据的标准化变换 【zscore】
- 数据的极差归一化变换

9.1 聚类分析

- Q型聚类分析（样品聚类）
- R型聚类分析（变量聚类）
- 变量（如长度、等级、性别）
- 距离
- 相似系数（夹角余弦、相关系数等）
- 系统聚类法、K均值聚类法、模糊C均值聚类法

样品聚类法

地 区	食 品	衣 着	居 住	家庭设备用	医疗保健	交通和通信	教育文化娱	杂项商品和
北 京	4934.05	1512.88	1246.19	981.13	1294.07	2328.51	2383.96	649.66
天 津	4249.31	1024.15	1417.45	760.56	1163.98	1309.94	1639.83	463.64
河 北	2789.85	975.94	917.19	546.75	833.51	1010.51	895.06	266.16
山 西	2600.37	1064.61	991.77	477.74	640.22	1027.99	1054.05	245.07
内蒙古	2824.89	1396.86	941.79	561.71	719.13	1123.82	1245.09	468.17
辽 宁	3560.21	1017.65	1047.04	439.28	879.08	1033.36	1052.94	400.16
吉 林	2842.68	1127.09	1062.46	407.35	854.8	873.88	997.75	394.29
黑龙江	2633.18	1021.45	784.51	355.67	729.55	746.03	938.21	310.67
上 海	6125.45	1330.05	1412.1	959.49	857.11	3153.72	2653.67	763.8
江 苏	3928.71	990.03	1020.09	707.31	689.37	1303.02	1699.26	377.37
浙 江	4892.58	1406.2	1168.08	666.02	859.06	2473.4	2158.32	467.52
安 徽	3384.38	906.47	850.24	465.68	554.44	891.38	1169.99	309.3
福 建	4296.22	940.72	1261.18	645.4	502.41	1606.9	1426.34	375.98
江 西	3192.61	915.09	728.76	587.4	385.91	732.97	973.38	294.6
山 东	3180.64	1238.34	1027.58	661.03	708.58	1333.63	1191.18	325.64
河 南	2707.44	1053.13	795.39	549.14	626.55	858.33	936.55	300.19
湖 北	3455.98	1046.62	856.97	550.16	525.32	903.02	1120.29	242.82
湖 南	3243.88	1017.59	869.59	603.18	668.53	986.89	1285.24	315.82
广 东	5056.68	814.57	1444.91	853.18	752.52	2966.08	1994.86	454.09
广 西	3398.09	656.69	803.04	491.03	542.07	932.87	1050.04	277.43
海 南	3546.67	452.85	819.02	519.99	503.78	1401.89	837.83	210.85
重 庆	3674.28	1171.15	968.45	706.77	749.51	1118.79	1237.35	264.01
四 川	3580.14	949.74	690.27	562.02	511.78	1074.91	1031.81	291.32

具体程序： doc clusterdata

- X %数据x
- X=zscore(x)
- T=clusterdata(x,'linkage','average','maxclust',3)
- % 分类个数的确定（不一致系数:inconsistent）

样品聚类法

- %*****读取数据，并进行标准化*****
- [X,textdata] = xlsread('examp09_02.xls'); % 从Excel文件中读取数据
- X = zscore(X); % 数据标准化（减去均值，除以标准差）
- %*****调用clusterdata函数进行一步聚类*****
- obslabel = textdata(2:end,1); % 提取城市名称，为后面聚类做准备
- % 样品间距离采用欧氏距离，利用类平均法将原始样品聚为3类，Taverage为各观测的类编号
- Taverage = clusterdata(X,'linkage','average','maxclust',3);
- obslabel(Taverage == 1) % 查看第1类所包含的城市
- obslabel(Taverage == 2) % 查看第2类所包含的城市
- obslabel(Taverage == 3) % 查看第3类所包含的城市
- %***** 分步聚类 *****
- y = pdist(X); % 计算样品间欧氏距离，y为距离向量
- Z = linkage(y,'average') % 利用类平均法创建系统聚类树
- obslabel = textdata(2:end,1); % 提取城市名称，为后面聚类做准备
- % 绘制聚类树形图，方向从右至左，显示所有叶节点，用城市名作为叶节点标签，叶节点标签在左侧
- H = dendrogram(Z,0,'orientation','right','labels',obslabel); % 返回线条句柄H
- set(H,'LineWidth',2,'Color','k'); % 设置线条宽度为2，颜色为黑色
- xlabel('标准化距离（类平均法）') % 为X轴加标签
- inconsistent0 = inconsistent(Z,40) % 计算不一致系数，计算深度为40

程序结果

- $\text{ans} =$

• '北京'

• '上海'

- ans =

• '天津'

- '浙江'

• '广东'

ans =

'河北'

'山西'

‘内蒙古’

'辽宁'

‘吉林’

‘黑龙江’

‘江苏’

‘安徽’

‘福建’

'江西'

‘山东’

‘河南’

‘湖北’

‘湖南’

‘广西’

‘海南’

‘重庆’

· 四川 ·

‘ 貴 州 ’

‘云南’

‘西藏’

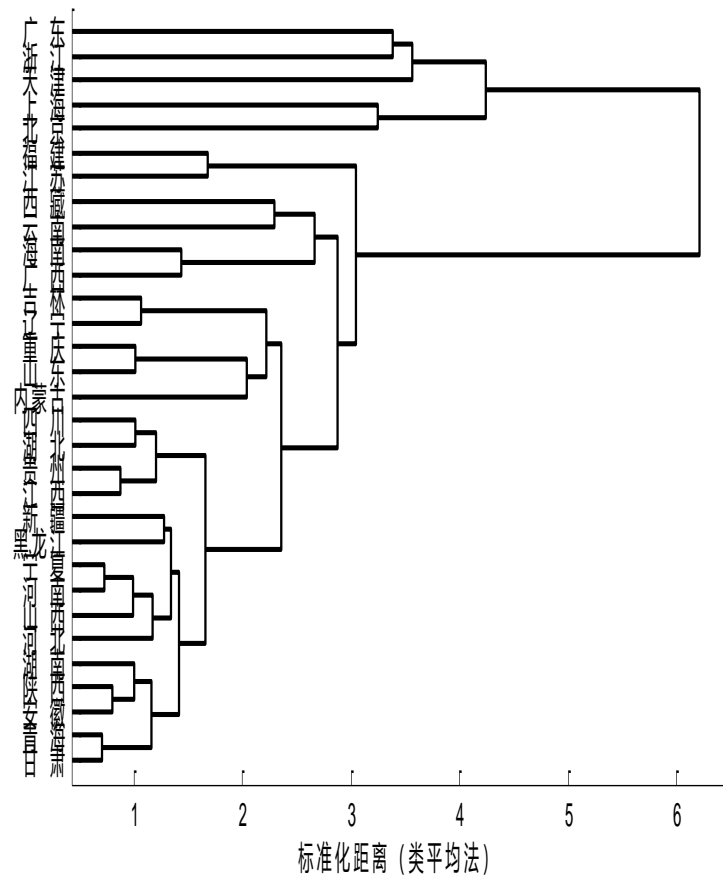
‘陕西’

‘甘肃’

‘青海’

‘宁夏’

‘新疆’



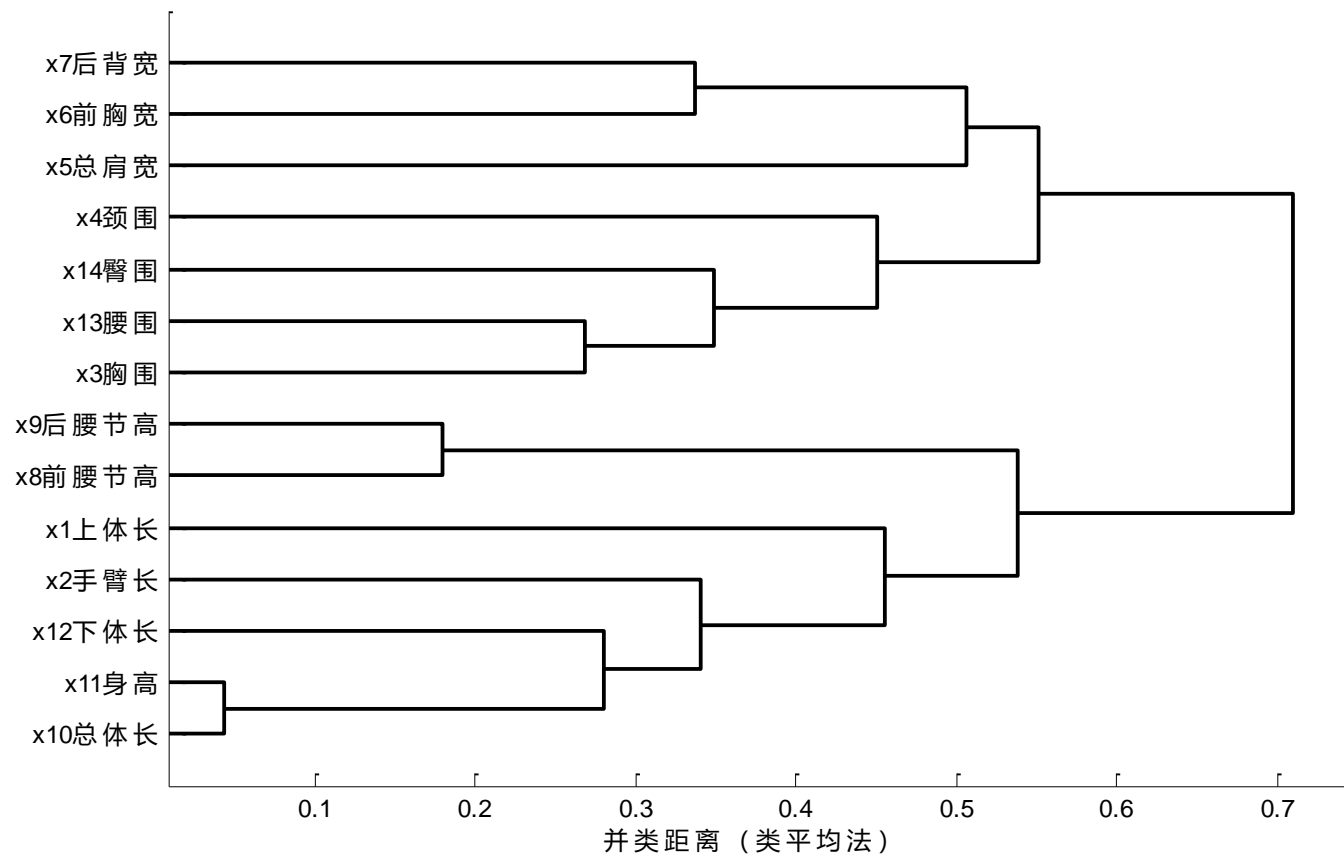
变量聚类案例

[illegible]

变量聚类程序

- %*****读取数据，并转为距离向量*****
- [X,textdata] = xlsread('examp09_03.xls'); % 从Excel文件中读取数据
- y = 1 - X(X~=1 & ~isnan(X))' % 提取X矩阵的不等于1和NaN的元素，并转为距离向量
- %*****调用linkage函数创建系统聚类树*****
- Z = linkage(y,'average') % 利用类平均法创建系统聚类树
- %***** 绘制聚类树形图 *****
- varlabel = textdata(2:end,1); % 提取变量名称，为后面聚类做准备
- % 作出聚类树形图，方向从右至左，显示所有叶节点，用变量名作为叶节点标签，叶节点标签在左侧
- H = dendrogram(Z,0,'orientation','right','labels',varlabel); % 返回线条句柄H
- set(H,'LineWidth',2,'Color','k'); % 设置线条宽度为2，颜色为黑色
- xlabel('并类距离（类平均法）') % 为X轴加标签

程序运行结果



判别分析 (统计)

Classify(sample,training,group,'method')

企业编号	组 别	x1	x2	x3	x4
1	1	-0.45	-0.41	1.09	0.45
2	1	-0.56	-0.31	1.51	0.16
3	1	0.06	0.02	1.01	0.4
4	1	-0.07	-0.09	1.45	0.26
5	1	-0.1	-0.09	1.56	0.67
6	1	-0.14	-0.07	0.71	0.28
7	1	0.04	0.01	1.5	0.71
8	1	-0.07	-0.06	1.37	0.4
9	1	0.07	-0.01	1.37	0.34
10	1	-0.14	-0.14	1.42	0.43
11	1	-0.23	-0.3	0.33	0.18
12	1	0.07	0.02	1.31	0.25
13	1	0.01	0	2.15	0.7
14	1	-0.28	-0.23	1.19	0.66
15	1	0.15	0.05	1.88	0.27
16	1	0.37	0.11	1.99	0.38
17	1	-0.08	-0.08	1.51	0.42
18	1	0.05	0.03	1.68	0.95
19	1	0.01	0	1.26	0.6
37	2	0.16	0.05	2.31	0.2
38	2	0.29	0.06	1.84	0.38
39	2	0.54	0.11	2.33	0.48
40	2	-0.33	-0.09	3.01	0.47
41	2	0.48	0.09	1.24	0.18
42	2	0.56	0.11	4.29	0.44
43	2	0.2	0.08	1.99	0.3
44	2	0.47	0.14	2.92	0.45
45	2	0.17	0.04	2.45	0.14
46	2	0.58	0.04	5.06	0.13
47	未判	-0.16	-0.1	1.45	0.51
48	未判	0.41	0.12	2.01	0.39
49	未判	0.13	-0.09	1.26	0.34
50	未判	0.37	0.08	3.65	0.43

MATLAB 程序

```
读取examp10_01.xls中数据，进行距离判别

%-----
%%

%*****读取数据*****

% 读取文件examp10_01.xls的第1个工作表中C2:F51范围的数据，即全部样本数据，包括未判企业
sample = xlsread('examp10_01.xls','', 'C2:F51');

% 读取文件examp10_01.xls的第1个工作表中C2:F47范围的数据，即已知组别的样本数据，
training = xlsread('examp10_01.xls','', 'C2:F47');

% 读取文件examp10_01.xls的第1个工作表中B2:B47范围的数据，即样本的分组信息数据，
group = xlsread('examp10_01.xls','', 'B2:B47');

obs = [1 : 50]'; % 企业的编号


%*****距离判别*****

% 距离判别，判别函数类型为mahalanobis，返回判别结果向量C和误判概率err
%[C,err] = classify(sample,training,group,'mahalanobis');
[C,err] = classify(sample,training,group,'quadratic');
[obs, C] % 查看判别结果
err % 查看误判概率
```

蠓虫(Af与Apf)进行鉴别

1. 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翅膀的长度,已经测得 9 只 Af 和 6 只 Apf 的数据,(触角长度用 x 表示,翅膀长度用 y 表示)具体数据为:
Af 类触角和翅膀长度

x	1.24	1.36	1.38	1.38	1.38	1.40	1.48	1.54	1.56
Y	1.27	1.74	1.64	1.82	1.90	1.70	1.82	1.82	2.08

← AF

Apf 类触角和翅膀长度数据

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

← APF

试对给定 $(x,y)=(1.24,1.80)$ 、 $(1.28,1.84)$ 、 $(1.40,2.04)$ 加以识别,即指出这三个样本数据代表 Af 还是 Apf?



待定

蠓虫(Af与Apf)进行鉴别程序

```
sample=[1.24 1.36 1.38 1.38 1.38 1.4 1.48 1.54 1.56 1.14 1.18 1.2 1.26 1.28 1.3
1.24 1.28 1.4
1.27 1.74 1.64 1.82 1.9 1.7 1.82 1.82 2.08 1.78 1.96 1.86 2 2 1.96 1.8
1.84 2.04]
training=[1.24 1.36 1.38 1.38 1.38 1.4 1.48 1.54 1.56 1.14 1.18 1.2 1.26 1.28 1.3
1.27 1.74 1.64 1.82 1.9 1.7 1.82 1.82 2.08 1.78 1.96 1.86 2 2 1.96]
group=[1 1 1 1 1 1 1 1 1 2 2 2 2 2 2]
obs = [1 : 18]
%*****距离判别*****
% 距离判别, 判别函数类型为mahalanobis, 返回判别结果向量C和误判概率err
%[C,err] = classify(sample,training,group,'mahalanobis');
[C,err] = classify(sample,training,group,'quadratic');
[obs, C] % 查看判别结果
err % 查看误判概率
```

ans =

1	1
2	1
3	1
4	1
5	1
6	1
7	1
8	1
9	1
10	2
11	2
12	2
13	2
14	2
15	2
16	2
17	2
18	1

err =

0

10.1 回归分析： 调用regress函数作一元线性回归

- %-----
- % 读取原始数据，调用regress函数作一元线性回归
- %-----
-
- %*****读取数据，绘制散点图*****
- ClimateData = xlsread('examp08_01.xls'); % 从Excel文件读取数据
- x = ClimateData(:, 1); % 提取ClimateData的第1列，即年平均气温数据
- y = ClimateData(:, 5); % 提取ClimateData的第5列，即全年日照时数数据
- plot(x, y, 'k.', 'Markersize', 15) % 绘制x和y的散点图
- xlabel('年平均气温(x)') % 给X轴加标签
- ylabel('全年日照时数(y)') % 给Y轴加标签

```
%*****计算相关系数*****
```

```
R = corrcoef(x, y) %计算x和y的线性相关系数矩阵R
```

```
%*****调用regress函数作一元线性回归*****
```

```
xdata = [ones(size(x, 1), 1), x]; % 在原始数据x的左边加一列1，即模型包含常数项
```

```
[b, bint, r, rint, s] = regress(y, xdata); % 调用regress函数作一元线性回归
```

```
yhat = xdata*b; % 计算y的估计值
```

```
% 定义元胞数组，以元胞数组形式显示系数的估计值和估计值的95%置信区间
```

```
head1 = {'系数的估计值','估计值的95%置信下限','估计值的95%置信上限'};
```

```
[head1; num2cell([b, bint])]
```

```
% 定义元胞数组，以元胞数组形式显示y的真实值、y的估计值、残差和残差的95%置信区间
```

```
head2 = {'y的真实值','y的估计值','残差','残差的95%置信下限','残差的95%置信上限'};
```

```
% 同时显示y的真实值、y的估计值、残差和残差的95%置信区间
```

```
[head2; num2cell([y, yhat, r, rint])]
```

```
% 定义元胞数组，以元胞数组形式显示判定系数、F统计量的观测值、检验的p值和误差方差的估计值
```

```
head3 = {'判定系数','F统计量的观测值','检验的p值','误差方差的估计值'};
```

```
[head3; num2cell(s)]
```

```
%*****绘制回归直线*****
```

```
plot(x, y, 'k.', 'Markersize', 15) % 画原始数据散点
```

```
hold on
```

```
plot(x, yhat, 'linewidth', 3) % 画回归直线
```

```
xlabel('年平均气温(x)') % 给X轴加标签
```

```
ylabel('全年日照时数(y)') % 给Y轴加标签
```

```
legend('原始散点','回归直线'); % 加标注框
```

%-----
% **剔除异常数据，重新调用regress函数作一元线性回归**
%-----

```
%*****残差分析*****  
figure % 新建一个图形窗口  
rcoplot(r,rint) %按顺序画出各组观测对应的残差和残差的置信区间  
  
%*****剔除异常数据，重新计算相关系数矩阵*****  
xt = x(y<3000 & y>1250); % 根据条件y<3000 & y>1250剔除异常数据  
yt = y(y<3000 & y>1250);  
figure % 新建一个空的图形窗口  
plot(xt, yt, 'ko') % 画剔除异常数据后的散点图  
xlabel('年平均气温(x)') % 为X轴加标签  
ylabel('全年日照时数(y)') % 为Y轴加标签  
Rt = corrcoef(xt, yt) % 重新计算相关系数矩阵
```


%*****重新调用regress函数作一元线性回归*****

xtdata = [ones(size(xt, 1), 1), xt]; % 在数据xt的左边加一列1

% 调用regress函数对处理后数据作一元线性回归

[b, bint, r, rint, s] = regress(yt, xtdata);

b % 显示常数项和回归系数的估计值

s % 显示判定系数、F统计量的观测值、p值和误差方差的估计值

ythat = xtdata*b; % 重新计算y的估计值

%*****绘制两次回归分析的回归直线*****

figure; % 新建一个图形窗口

plot(x, y, 'ko'); % 画原始数据散点

hold on; % 图形叠加

[xsort, id1] = sort(x); % 为了画图的需要将x从小到大排序

yhat-sort = yhat(id1); % 将估计值yhat按x排序

plot(xsort, yhat-sort, 'r--', 'linewidth', 3); % 画原始数据对应的回归直线，红色虚线

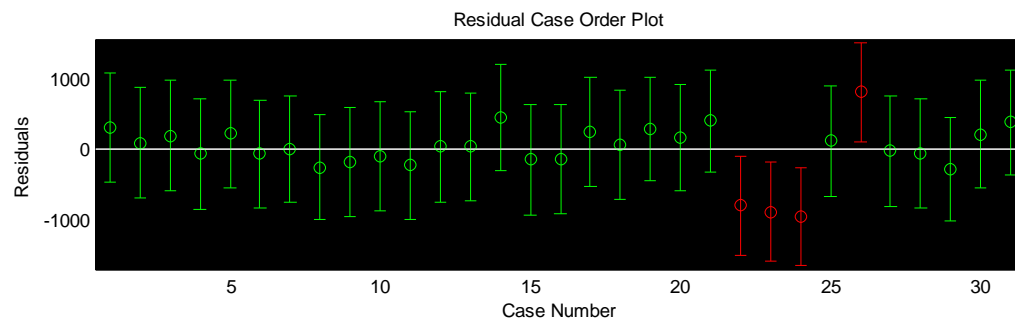
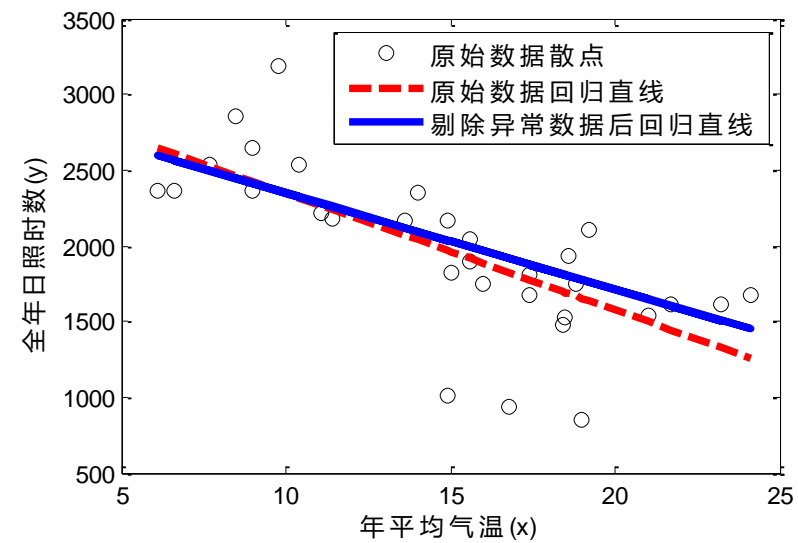
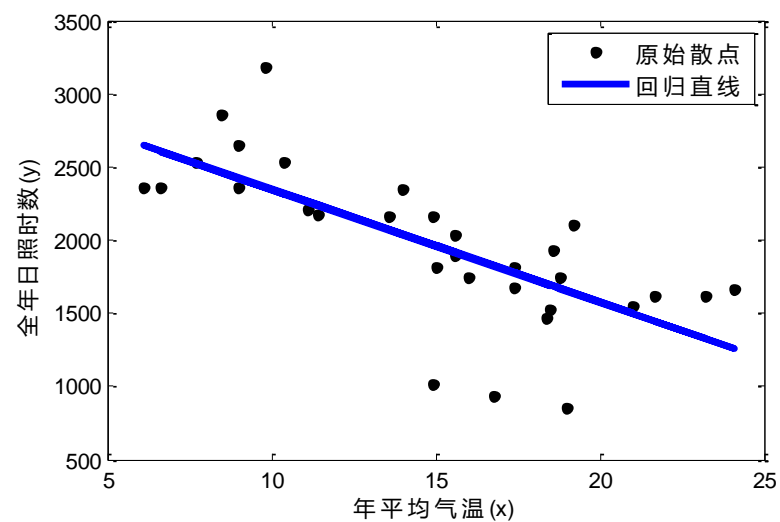
plot(xt, ythat, 'linewidth', 3); % 画剔除异常数据后的回归直线，蓝色实线

legend('原始数据散点', '原始数据回归直线', '剔除异常数据后回归直线') % 为图形加标注框

xlabel('年平均气温(x)'); % 为X轴加标签

ylabel('全年日照时数(y)'); % 为Y轴加标签

结果显示



交互式逐步回归分析

```
%-----  
%      读取原始数据， 调用stepwise函数作交互式逐步回归分析  
%-----  
  
% 从Excel文件examp08_03.xls中读取数值型数据  
xydata = xlsread('examp08_03.xls');  
y = xydata(:, 2); % 提取矩阵xydata的第2列数据， 即耗氧能力数据y  
X = xydata(:, 3:7); % 提取矩阵xydata的第3至7列数据， 即自变量观测值矩阵X  
  
inmodel = 1:5; % 初始模型中除了常数项， 还包含x1至x5等线性项  
stepwise(X,y,inmodel); % 交互式逐步回归分析
```

序号	y (ml/min. kg)	x1 (岁)	x2 (kg)	x3 (min)	x4 (次/min)	x5 (次/min)
1	44.6	44	89.5	6.82	62	178
2	45.3	40	75.1	6.04	62	185
3	54.3	44	85.8	5.19	45	156
4	59.6	42	68.2	4.9	40	166
5	49.9	38	89	5.53	55	178
6	44.8	47	77.5	6.98	58	176