01

统计分析

01 描述性统计

- 直方图
- 均值
- 标准差 ...
- 偏度
- 峰度

01 直方图 (histogram)

- hist(x)
- hist(x, m)、
- histfit(x, m)　　% 带正态拟合的直方图

01 描述性统计量

- sum(x)　　%　和
- mean(x)　　%　均值
- std(x)　　%　标准差
- var(x)　　%　方差
- sort(x)　　%　顺序统计量
- median(x)　%　中位数
- skewness(x)　%　偏度，正态是0
- kurtosis(x)　%　峰度，正态是3

01 随机数（random number）

- 均匀分布随机数
- 正态分布随机数
- 指数分布随机数
- 卡方分布随机数
- t分布随机数
- F分布随机数
- 离散分布随机数

01 均匀分布的随机数

- rand(n)　　% [0, 1]区间上
- rand(m, n)　　% [0, 1]区间上
- unifrnd(a, b, m, n)　% [a, b]区间上

01 正态分布的随机数

- randn(n)　　% N(0, 1)
- randn(m, n)　　% N(0, 1)
- normrnd(a, b, m, n)　　% N(a, b^2)
- x=randn(m, n); x=a+b*x %等价

01 指数分布的随机数

- exprnd (lambda)　　% 1个随机数

- exprnd (lambda, m, n)

$$f(x)=\frac{1}{\lambda}\exp\left\{-\frac{1}{\lambda}x\right\},\quad x>0.$$

01 卡方分布的随机数

- chi2rnd (df)

$$\chi^2(df)$$

- chi2rnd (df, m, n)

01 t分布的随机数

- trnd (df)
- trnd (df, m, n)

01 二项分布的随机数

- binornd (N, p)
- binornd (N, p, m, n)

$$B(N,p): \quad P(X=k)=C_N^k p^k (1-p)^{N-k}.$$

01 Poisson分布的随机数

- poissrnd (lambda)
- poissrnd (lambda, m, n)

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \cdots$$

01 离散型分布的随机数

- 以标准的均匀分布U作为模拟变量
- 若　　U ≤0.20，则X值x1；
- 若0.20<U ≤ 0.35, 则X值x2；
- 若0.35<U ≤ 0.60, 则X值x3；
- 若0.60<U ≤ 1，　则X值x4 .

$$设X \sim \begin{pmatrix} x_1 & x_2 & x_3 & x_4 \\ 0.20 & 0.15 & 0.25 & 0.40 \end{pmatrix},$$

```
clear
n=5000;
for i=1:n
    u=rand(1);
    if u<=0.2
        x(i)=1;
    elseif u<=0.35
        x(i)=2;
    elseif u<=0.6
        x(i)=3;
    else
        x(i)=4;
    end
end                    % sum(x==1)/n
```

01 单样本与两样本的t检验

- 单样本的t检验
- 两样本的t检验
- 检验的水平
- 检验的功效（势）

01 单样本的t检验

设总体的分布为 $N(\mu,\sigma^2)$，从总体中抽取容量为n的样本，要检验的问题是

$$H_0:\mu=\mu_0,\qquad H_1:\mu\neq\mu_0,$$

设总体的方差未知，则使用的是单样本t检验：

$$t=\frac{\overline{x}-\mu_0}{s/\sqrt{n}}\overset{H_0}{\sim}t(n-1),$$

取检验的水平为 α，则检验的拒绝域为：

$$|t|>t_{\alpha/2}(n-1).$$

01 单样本t检验的Matlab实现：

```
h=ttest(x, mu0)
    %  x是样本;
    %  muo缺省时为0 ;
    %  h输出值0和1，分别表示接受和拒绝H0 .
```

01 单样本t检验的Matlab实现：

点头教育

```
[h, sig, ci, stats]=ttest(x, mu0, alpha, tail)
%  alpha: 显著性水平, 缺省时为0.05.
%  tail: 取0表示双侧检验(可缺省); 取-1或1表示单侧检验,  其中-1对应H1:
mu<mu0,  1对应H1: mu>mu0.
%  h输出值0和1, 分别表示接受和拒绝H0 .
%  sig: 检验的p-值, sig<0.05等价于h=1.
%  ci输出置信区间, stats输出统计量的值和自由度.
```

01 两样本的t检验

设有两个总体 $N(\mu_1,\sigma^2)$和 $N(\mu_2,\sigma^2)$， 分别从这两个总体中抽取容量为n1和 n2的样本，要检验的问题是

$$H_0:\mu_1=\mu_2,\quad H_1:\mu_1\neq\mu_2,$$

设总体的方差未知，则使用的是两样本t检验：

$$t=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{(n_1-1)s_1^2+(n_1-1)s_1^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

取检验的水平为α，则检验的拒绝域为：

$$|t|>t_{\alpha/2}(n_1+n_2-2).$$

$$\overset{H_0}{\sim} t(n_1+n_2-2),$$

01 两样本t检验的Matlab实现：

```
h=ttest2(x, y)
    %  x，y是样本;
    %  h输出值0和1，分别表示接受和拒绝H0 .
```

01 两样本t检验的Matlab实现：

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

% alpha: 显著性水平, 缺省时为0.05.

% tail: 取0表示双侧检验(可缺省); 取-1或1表示单侧检验,

其中-1对应H1: mu1<mu2, 1对应H1: mu1>mu2.

% h输出值0和1, 分别表示接受和拒绝H0 .

% sig: 检验的p-值, sig<0.05等价于h=1.

% ci输出置信区间, stats输出统计量的值和自由度.

01 检验的水平

在零假设成立下，重复执行检验过程，考察零假设被拒绝的概率，这就是犯第一类错误的概率，即检验的实际水平。

01 检验的水平

```
clear
n=20;
N=10000;
mu0=0;
for i=1:N
        x=randn(1, n);
        a(i)=ttest(x,mu0);
end
sum(a)/N                          %    t检验的实际水平
```

01 检验的功效（势，power）

在备择假设成立下，重复执行检验过程，考察零假设被拒绝的概率，这就是不犯第二类错误的概率，即检验的功效。

检验的功效越高，检验就越好。

01 检验的功效（势，power）

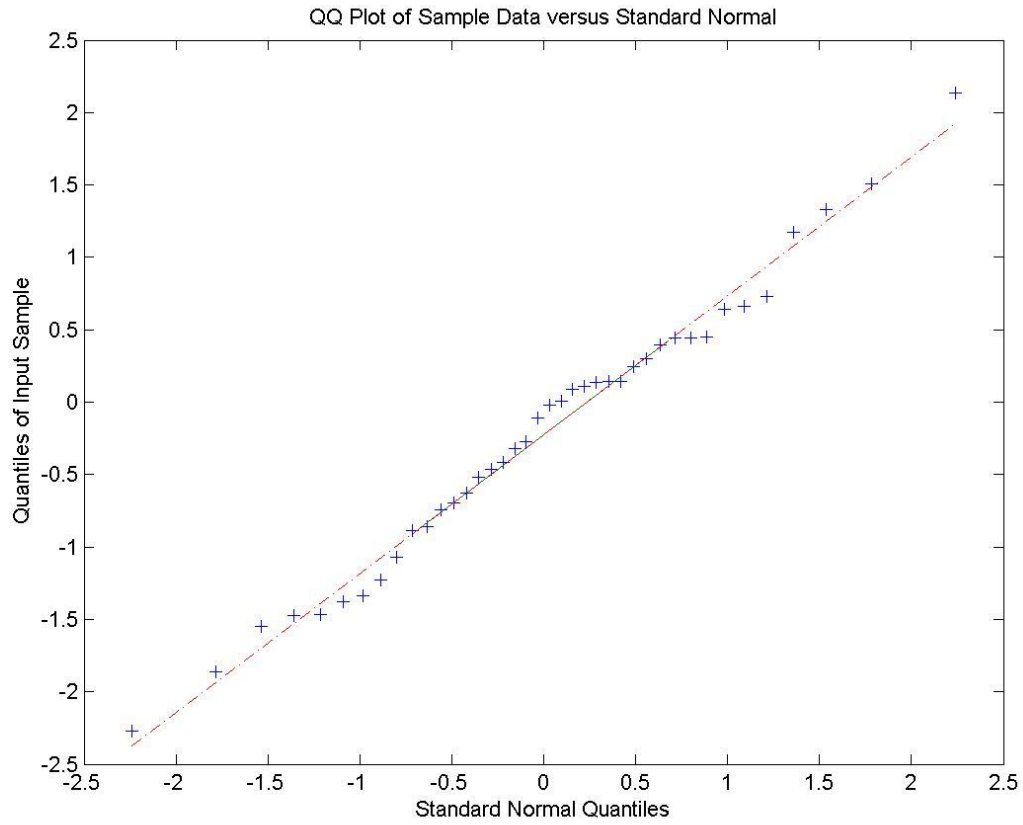
```
clear
n=20;
N=10000;
mu0=0.5;
for i=1:N
      x=randn(1, n);
      a(i)=ttest(x,mu0);
end
sum(a)/N                          %    功效
```

01 正态性检验

- Q-Q图
- Kolmogorov-Smirov检验
- Lilliefors检验

01 Q-Q图 （quantile-quantile）

```
clear
n=40;
x=randn(1,n);
qqplot(x)
```



Q-Q图： 第i个点的纵坐标是排序的样本观测值 $x_{(i)}$，横坐标是

$$\Phi^{-1}\left(\frac{i-0.5}{n}\right), \qquad i=1,2,\cdots,n.$$

理论直线的方程为： $y=\text{mean}+\text{std}\cdot x.$

01 Kolmogorov-Smirov检验

检验的统计量是：

$$D_n = \sqrt{n}\sup_{t\in R}\left|F_n(t)-F(t)\right|,$$

$$= \sqrt{n}\max_{1\le i\le n}\left\{\left|\frac{i-1}{n}-F(x_{(i)})\right|,\left|\frac{i}{n}-F(x_{(i)})\right|\right\},$$

其中$F(x)$是待检验的分布函数。

01 Kolmogorov-Smirov检验

Matlab中的命令：

[h, p]=kstest(x, [ ], alpha, tail)

检验样本是否服从标准的正态分布，其中

alpha: 检验的水平

tail: 检验的类型，0表示双侧检验，-1和1是单侧检验

h: 取值0和1，分别表示接受和拒绝零假设

p: 检验的p-值

简单用法：h=kstest(x)

01 Kolmogorov-Smirov检验

```
clear
n=30;
N=5000;
for i=1:N
    x=randn(1, n);
    h=kstest(x);
    if h==1
        a(i)=1;
    else
        a(i)=0;
    end
end
sum(a)/N                    % 结果是什么?
```

01 Kolmogorov-Smirov检验

```
clear
n=80;
N=5000;
for i=1:N
    x=trnd(1, 1, n);  %样本来自于t(1)
    a(i)= kstest(x);
end
sum(a)/N                    %结果是什么?
```

01 Lilliefors检验

用来检验数据是否具有与样本相同均值和方差的正态分布。

Matlab中的命令：

h=lillietest(x)

[h,p]=lillietest(x, [ ], alpha, tail)

01 Lilliefors检验

```
clear
n=30;
N=5000;
for i=1:N
    x=randn(1, n)+2;
    a(i)= lillietest(x);
end
sum(a)/N                    % ?
```

01 方差分析(analysis of variance)

两因素方差分析的Matlab实现：、

p=anova1(x)

% x是样本观测值构成的矩阵，每一列为一个水平

% p是检验的p-值.

% 输出结果除p-值外，还有方差分析表以及箱形图。

```
clear
x=[34    37    34    36
     36    36    37    34
     34    35    35    37
     35    37    37    34
     34    37    36    35];
p=anova1(x)  % 1是数字
```

01 方差分析(analysis of variance)

单因素方差分析的Matlab实现：

p=anova2(x, 1)

% 括号中的1表示每个水平组合下只有一次观测，此时不考虑交互效应。

p=anova2(x, m)

% 括号中的m表示每个水平组合下有m次重复观测。

The number of rows must be a multiple of reps.

Columns: 列因素

Rows: 行因素

Interaction: 交互作用

```
clear
x=[ 215 145 160
    196 174 203
    209 150 185
    228 178 193
    148 121 144
    156 114 147
    135 127 138
    164 145 120 ];
p=anova2(x,4)
```

01 Lilliefors检验

```
clear
n=30;
N=5000;
for i=1:N
    x=randn(1, n)+2;
    a(i)= lillietest(x);
end
sum(a)/N                    % ?
```

01 直方图（histogram）

子函数

在MATLAB语言中，与其他的程序设计语言类似，也可以定义子函数，以扩充函数的功能。在函数文件中题头中所定义的函数为主函数，而在函数体内定义的其他函数均被视为子函数。子函数只能被主函数或同一主函数下其他的子函数所调用。

局部函数

在MATLAB语言中将放置在目录private下的函数称为局部函数，这些函数只能被private目录的父目录中函数调用，而不能被其他的目录的函数调用。

01 回归分析

- 一元线性回归
- 多元线性回归

01 一元线性回归分析

- 设x为自变量，y为因变量，考虑y对x的线性回归。
- 采用最小二乘估计法。
- 一元线性回归分析的Matlab实现：

```
b=polyfit(x, y, 1)
  %  x是自变量的样本观测值
  %  y是因变量的样本观测值
```

01 一元线性回归分析的Matlab实现：

```
x=[0.7608 -0.9291 -0.4007 -0.1267 -0.4829 -0.6075 -0.7594 -1.3627
0.4069 0.4236];
y=[1.7159 -1.2786 -0.3744  0.5986 -0.6290 -1.0179 -1.0921 -2.6222
1.6396 1.8324];
plot(x,y,'*')          % 散点图
a=polyfit(x,y,1)   %  a(1)是一次项系数，  a(2)是截距项
yy=polyval(a,x);
hold on
plot(x,yy)
```

01 多元线性回归分析

设x1, x2, …, xk为自变量，y为因变量，考虑y对x1, x2, …, xk的线性回归:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k .$$

假设进行了n次观测.

多元线性回归分析的Matlab实现：

```
b=regress(y, X)
```

或

```
[b, bint, r, rint, stats]=regress(y, X)
```

y： 因变量的观测值所构成的列向量.

X：自变量的取值构成的矩阵，

$$X=\begin{pmatrix}1 & x_{11} & \cdots & x_{k1}\\ 1 & x_{12} & \cdots & x_{k2}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & x_{1n} & \cdots & x_{kn}\end{pmatrix}$$

X各列(除第1列外)就是自变量的n次取值。

```
x=[0.7608 -0.9291 -0.4007 -0.1267 -0.4829 -0.6075 -0.7594 -1.3627  0.4069
0.4236]’;
y=[1.7159 -1.2786 -0.3744  0.5986 -0.6290 -1.0179 -1.0921 -2.6222  1.6396
1.8324]’;
n=length(x);
X=[ones(n, 1), x];
[b, bint, r, rint, stats]=regress(y, X)
```