

MATLAB统计分析与应用

——数据分布拟合与检验

主讲人：谢中华

ssstudy.com
科学软件学习网

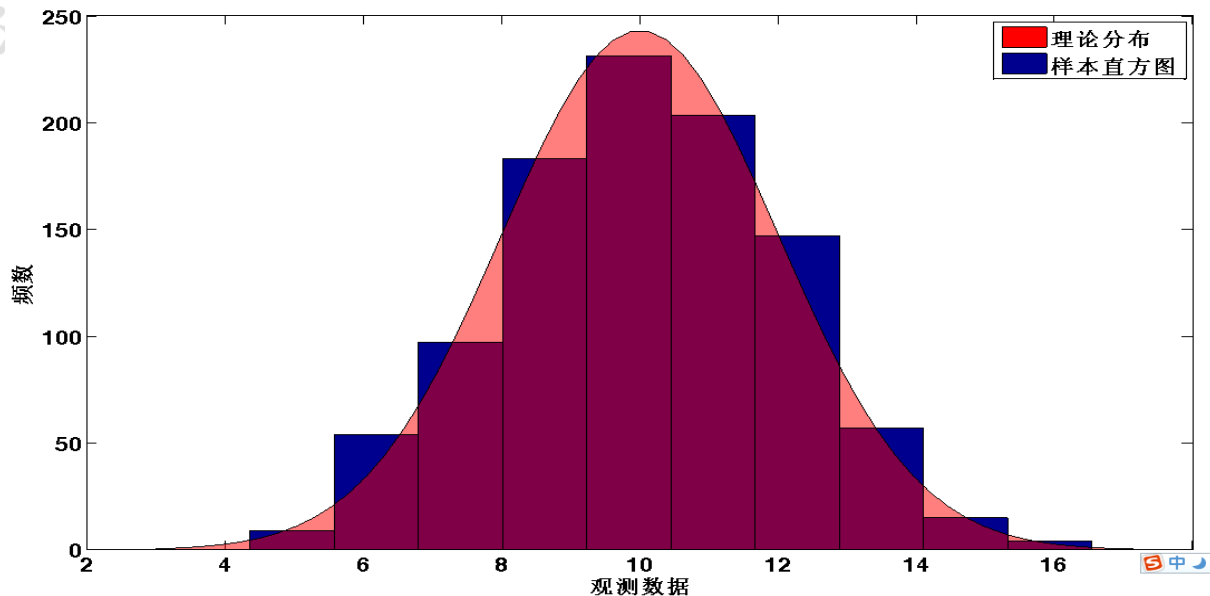
主要内容

- 分布的检验
- 一维核密度估计
- 多维核密度估计

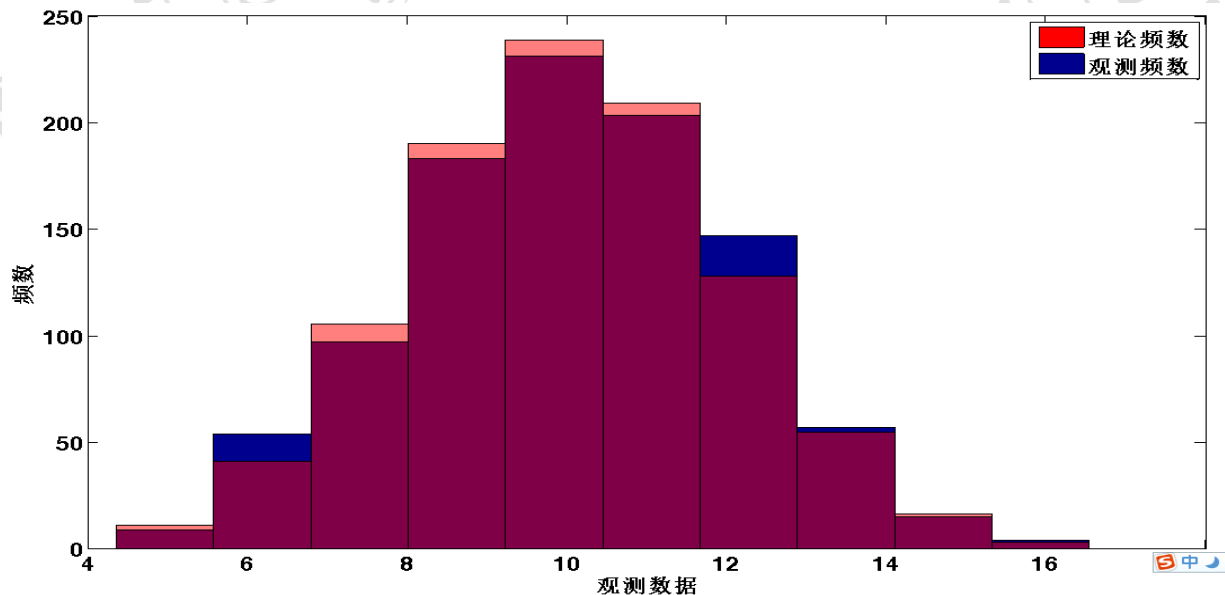
第一节 分布的检验

一、卡方拟合优度检验

1. 分布拟合图示



1. 分布拟合图示（续）



2. 卡方拟合优度检验的数学原理

构造统计量

$$\chi^2 = \sum_{i=1}^r \frac{(n_i - np_i)^2}{np_i}$$

实际频数

理论（期望）频数

当统计量的观测值超过某个临界值时可认为数据不服从理论分布。

3. chi2gof函数

调用格式：

`h = chi2gof(x)`

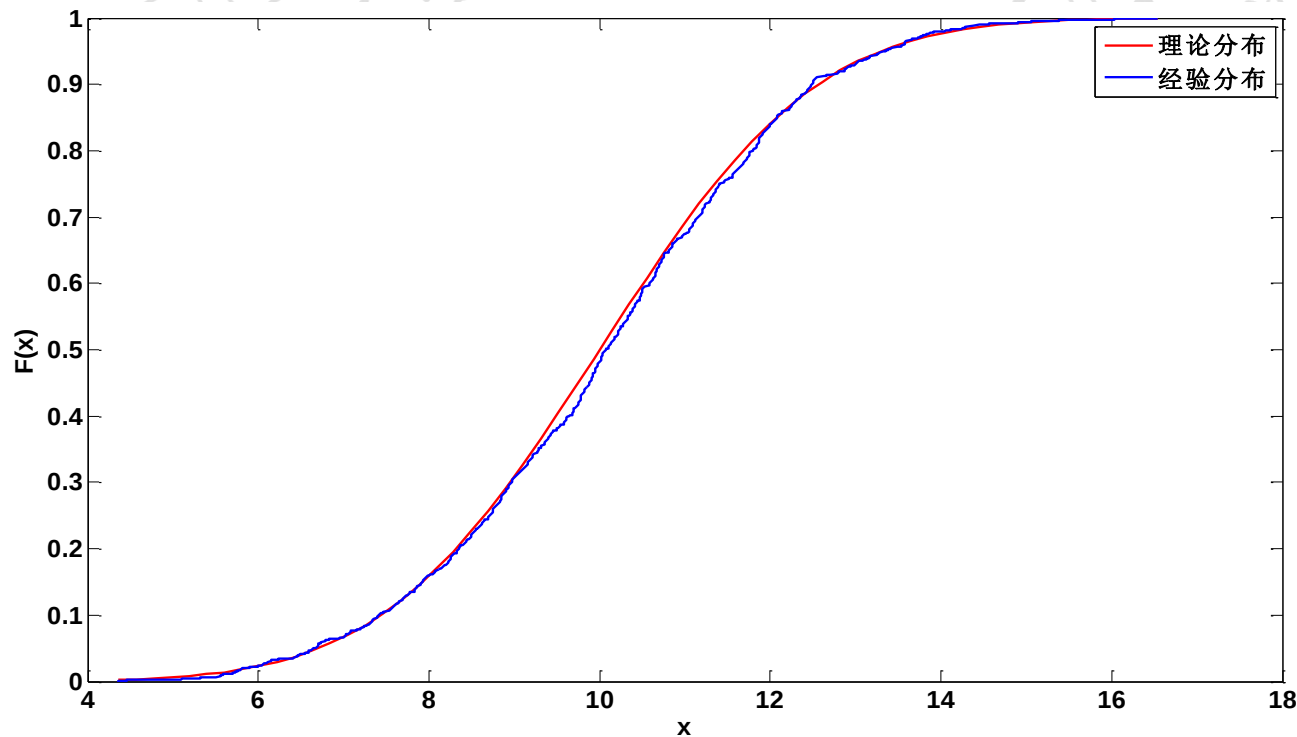
`[h,p] = chi2gof(...)`

`[h,p,stats] = chi2gof(...)`

`[...] = chi2gof(X,'Name',value)`

二、Kolmogorov-Smirnov检验

1. 分布拟合图示



2. Kolmogorov检验

$$H_0 : F(x) = F_0(x; \theta_0), \quad H_1 : F(x) \neq F_0(x; \theta_0)$$

检验统计量

$$D_n = \sup_{-\infty < x < +\infty} |F_n(x) - F_0(x; \theta_0)|$$

拒绝域

$$W = \{D_n \geq D_{n,\alpha}\}$$

3. Lilliefors检验

$$H_0 : F(x) = F_0(x; \theta), \quad H_1 : F(x) \neq F_0(x; \theta)$$

检验统计量

$$\Delta = \sup_{-\infty < x < +\infty} |F_n(x) - F_0(x; \hat{\theta})|$$

拒绝域

$$W = \{\Delta \geq \Delta_\alpha\}$$

4. Smirnov检验

$$H_0 : F(x) = G(x),$$

$$H_1 : F(x) \neq G(x)$$

检验统计量

$$D_{n_1, n_2} = \sup_{-\infty < x < +\infty} |F_{n_1}(x) - G_{n_2}(x)|$$

拒绝域

$$W = \{ D_{n_1, n_2} > D_{n_1, n_2, \alpha} \}$$

5. kstest函数

调用格式：

`h = kstest(x)`

`h = kstest(x,CDF)`

`h = kstest(x,CDF,alpha)`

`h = kstest(x,CDF,alpha,type)`

`[h,p,ksstat,cv] = kstest(...)`

6. kstest2函数

调用格式：

`h = kstest2(x1,x2)`

`h = kstest2(x1,x2,alpha,type)`

`[h,p] = kstest2(...)`

`[h,p,ks2stat] = kstest2(...)`

7. lillietest函数

调用格式：

`h = lillietest(x)`

`h = lillietest(x,alpha)`

`h = lillietest(x,alpha,distr)`

`[h,p] = lillietest(...)`

`[h,p,kstat] = lillietest(...)`

`[h,p,kstat,critval] = lillietest(...)`

`[h,p,...] = lillietest(x,alpha,distr,mctol)`

【例10.1-1】现有某两个班的某门课程的考试成绩，如下表

序号	班名	学号	姓名	平时成绩	期末成绩	总成绩	备注
1	60101	6010101	陈亮	0	63	63	
2	60101	6010102	李旭	0	73	73	
3	60101	6010103	刘鹏飞	0	0	0	缺考
4	60101	6010104	任时迁	0	82	82	
5	60101	6010105	苏宏宇	0	80	80	
...	...	...	...	...	...	...	...
49	60102	6010225	曾亦可	0	90	90	
50	60102	6010226	张阳	0	85	85	
51	60102	6010227	张梓涵	0	92	92	

试根据以上数据（`examp4_3_1.xls`），推断总成绩数据所服从的分布。

%*****调用kstest函数进行正态性检验*****

```
banji = xlsread('examp4_3_1.xls','Sheet1','B2:B52');  
score = xlsread('examp4_3_1.xls','Sheet1','G2:G52');  
banji = banji(score > 0);  
score = score(score > 0);  
cdf = [score, normcdf(score, 79, 10.1489)];
```

% 调用kstest函数，检验总成绩是否服从由cdf指定的分布

```
[h,p,ksstat,cv] = kstest(score,cdf)
```

%*调用kstest2函数检验两个班的总成绩是否服从相同的分布*

```
score1 = score(banji == 60101);  
score2 = score(banji == 60102);
```

% 调用kstest2函数检验两个班的总成绩是否服从相同的分布

```
[h,p,ks2stat] = kstest2(score1,score2)
```

%*****分别绘制两个班的总成绩的经验分布图*****

figure;

F1 = cdfplot(score1); % 绘制60101班总成绩的经验分布函数图

set(F1,'LineWidth',2,'Color','r')

hold on

F2 = cdfplot(score2); % 绘制60102班总成绩的经验分布函数图

set(F2,'LineStyle','-','LineWidth',2,'Color','k')

legend('60101班总成绩的经验分布函数','60102班总成绩的经验分布函数',...

'Location','NorthWest')

%*****调用lillietest函数进行分布的检验*****

[h,p,kstat,critval] = lillietest(score)

第二节 一维核密度估计

一、经验密度函数

1. 频率频数分布表

设 $x_1, x_2, \dots, x_n \in [a, b]$ 是取自总体 X 的样本，将样本取值区间分为 k 个子区间：

$$a = t_0 < t_1 < \dots < t_{k-1} < t_k = b, [t_0, t_1], (t_1, t_2], \dots, (t_{k-1}, t_k]$$

列出如下频率频数分布表

区间	$[t_0, t_1]$	$(t_1, t_2]$	...	$(t_{k-1}, t_k]$	总计
频数	n_1	n_2	...	n_k	n
频率	$f_1 = \frac{n_1}{n}$	$f_2 = \frac{n_2}{n}$	...	$f_k = \frac{n_k}{n}$	1

2. 经验密度函数

定义 2.1 (经验密度函数) 令

$$\hat{f}_n(x) = \begin{cases} \frac{f_i}{h_i} = \frac{n_i}{nh_i}, & x \in I_i, i = 1, 2, \dots, k \\ 0, & \text{其他} \end{cases}$$

称 $\hat{f}_n(x)$ 为样本的**经验密度函数**，它可以作为总体密度函数的一个非参数估计，它所对应的函数图形就是频率直方图。

$\hat{f}_n(x)$ 中的 $h_i = t_i - t_{i-1}$, $i = 1, \dots, k$ 表示每个区间的长度，称为**窗宽**或**带宽** (bandwidth)，它决定了经验密度函数的形状。

二、核密度估计

1. Parzen窗密度估计法

定义以原点为中心，半径为 0.5 的邻域函数（或 Parzen 窗函数）为

$$H(u) = \begin{cases} 1, & |u| \leq \frac{1}{2} \\ 0, & \text{其他} \end{cases}$$

则当第 i 个样本点 x_i 落入以 x 为中心，以 $h/2$ 为半径的邻域内时， $H(\frac{x-x_i}{h})=1$ ，否则 $H(\frac{x-x_i}{h})=0$ ，因此落入这个邻域内总的样本点数为 $\sum_{i=1}^n H(\frac{x-x_i}{h})$ ，于是点 x 处的密度函数的估计值为

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n H\left(\frac{x-x_i}{h}\right).$$

2. 核密度估计的一般定义

定义 2.2（核密度估计） 设 $x_1, x_2, \dots, x_n$ 是取自一元连续总体 X 的样本，在任意点 x 处的总体密度函数 $f(x)$ 的核密度估计定义为

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right).$$

其中 $K(\cdot)$ 称为**核函数（kernel function）**， h 称为窗宽。

3. 常用核函数

核函数名称	核函数表达式
Uniform (或 Box)	$\frac{1}{2} I(u \leq 1)$
Triangle	$(1 - u) I(u \leq 1)$
Epanechnikov	$\frac{3}{4} (1 - u^2) I(u \leq 1)$
Quaritic	$\frac{15}{16} (1 - u^2)^2 I(u \leq 1)$
Triweight	$\frac{35}{32} (1 - u^2)^3 I(u \leq 1)$
Gaussian	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} u^2\right)$
Cosinus	$\frac{\pi}{4} \cos\left(\frac{\pi}{2} u\right) I(u \leq 1)$

4. 窗宽对核密度估计的影响

核密度估计 $\hat{f}_h(x)$ 中窗宽 h 的取值会影响到 $\hat{f}_h(x)$ 的光滑程度。如果 h 取较大的值，将有较多的样本点对 X 处的密度估计产生影响，并且距 X 较近的点与较远的点对应的核函数值差距不大，此时 $\hat{f}_h(x)$ 的图像是较为光滑的曲线，但同时也丢失了数据所包含的一些信息；如果 h 取较小的值，只有很少的样本点对 X 处的密度估计产生影响，并且距 X 较近的点与较远的点对应的核函数值差距比较大，此时 $\hat{f}_h(x)$ 的图像是不光滑的折线，但它能反映出每个数据所包含的信息。

5. 如何选择最佳窗宽

$$\text{MISE}(\hat{f}_h) = E \left[\int \left\{ \hat{f}_h(x) - f(x) \right\}^2 dx \right],$$

MISE (mean integrated squared error) 是关于窗宽 h 的函数，求它的最小值点，可以得出最佳窗宽的估计值。

$$\hat{h} = \left\{ \frac{\int [K(x)]^2 dx}{\sigma_k^4 \int [f''(x)]^2 dx} \right\}^{\frac{1}{5}} n^{-\frac{1}{5}}.$$

$$\hat{h} = \left(\frac{4}{3} \right)^{\frac{1}{5}} \sigma n^{-\frac{1}{5}} \approx 1.06 \sigma n^{-\frac{1}{5}}.$$

三、一维核密度估计的MATLAB实现

1. **ksdensity**函数的调用格式

`[f,xi] = ksdensity(x)`

`f = ksdensity(x,xi)`

`ksdensity(...)`

`ksdensity(ax,...)`

`[f,xi,u] = ksdensity(...)`

`[...] =`

`ksdensity(...,param1,val1,param2,val2,...)`与参

数值列表见下一页。

四、核密度估计的案例分析

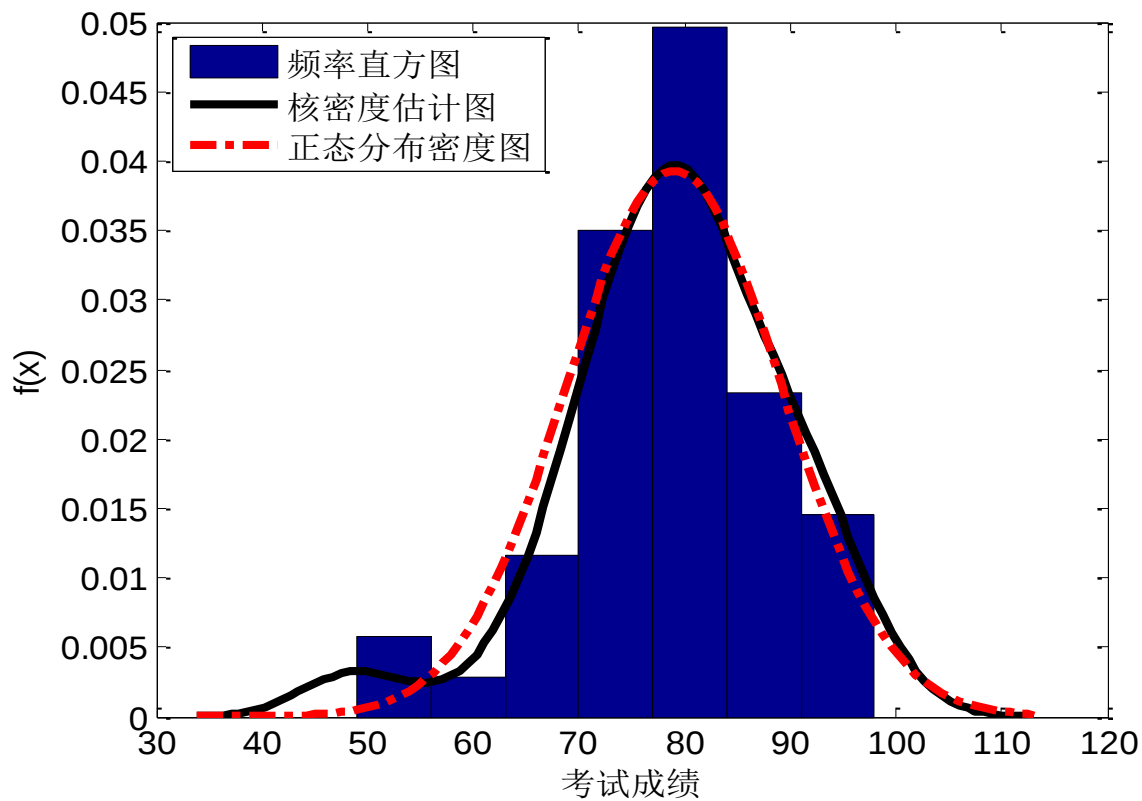
【例10.2-1】现有某两个班的某门课程的考试成绩，如下表

序号	班名	学号	姓名	平时成绩	期末成绩	总成绩	备注
1	60101	6010101	陈亮	0	63	63	
2	60101	6010102	李旭	0	73	73	
3	60101	6010103	刘鹏飞	0	0	0	缺考
4	60101	6010104	任时迁	0	82	82	
5	60101	6010105	苏宏宇	0	80	80	
...	...	...	...	...	...	...	...
49	60102	6010225	曾亦可	0	90	90	
50	60102	6010226	张阳	0	85	85	
51	60102	6010227	张梓涵	0	92	92	

试根据以上数据（ examp4_3_1.xls ），求总成绩数据的核密度估计。

1. 总成绩数据的核密度估计

```
>> score =  
xlsread('examp4_3_1.xls','Sheet1','G2:G52');  
>> score = score(score > 0);  
>> [f_ecdf, xc] = ecdf(score);  
% 绘制频率直方图  
>> figure;  
>> ecdfhist(f_ecdf, xc, 7);  
>> hold on;  
>> xlabel('考试成绩'); ylabel('f(x) ');  
>> [f_ks1,xi1,u1] = ksdensity(score);  
>> plot(xi1,f_ks1,'k','linewidth',3)
```



2. 窗宽对核密度估计的影响

% 设置窗宽分别为0.1, 1, 5和9, 调用
ksdensity函数进行核密度估计

```
>> [f_ks1,xi1] = ksdensity(score,'width',0.1);  
>> [f_ks2,xi2] = ksdensity(score,'width',1);  
>> [f_ks3,xi3] = ksdensity(score,'width',5);  
>> [f_ks4,xi4] = ksdensity(score,'width',9);
```

```
>> figure; % 新建图形窗口
```

```
% 分别绘制不同窗宽对应的核密度估计图
```

```
>> plot(xi1,f_ks1,'c-.','linewidth',2);
```

```
>> hold on;
```

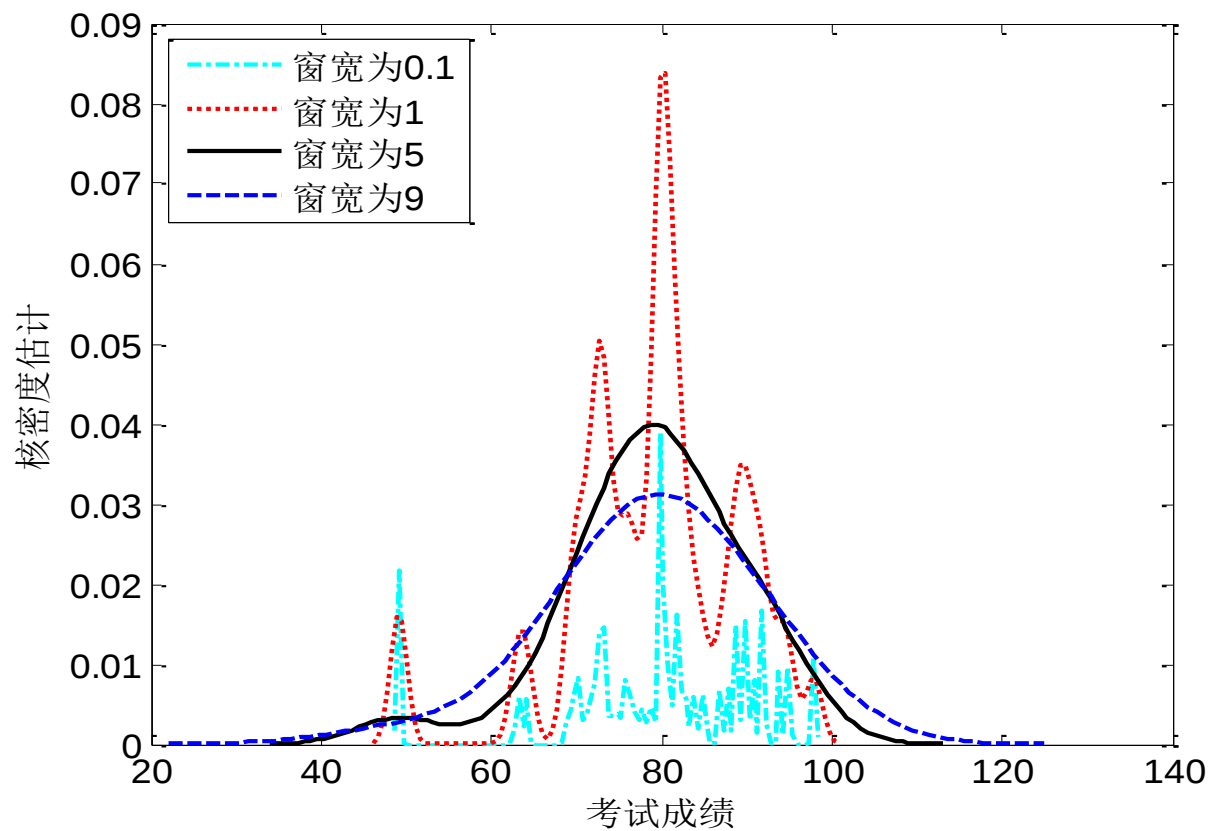
```
>> xlabel('考试成绩'); ylabel('核密度估计');
```

```
>> plot(xi2,f_ks2,'r:','linewidth',2);
```

```
>> plot(xi3,f_ks3,'k','linewidth',2);
```

```
>> plot(xi4,f_ks4,'b--','linewidth',2);
```

```
>> legend('窗宽为0.1','窗宽为1','窗宽为5',...  
'窗宽为9','Location','NorthWest');
```



3. 核函数对核密度估计的影响

% 设置不同核函数

```
>> [f_ks1,xi1] = ksdensity(score,'kernel','normal');  
>> [f_ks2,xi2] = ksdensity(score,'kernel','box');  
>> [f_ks3,xi3] = ksdensity(score,'kernel','triangle');  
>> [f_ks4,xi4] = ksdensity(score,'kernel','epanechnikov');
```

```
>> figure; % 新建图形窗口
```

```
% 分别绘制不同核函数对应的核密度估计图
```

```
>> plot(xi1,f_ks1,'k','linewidth',2)
```

```
>> hold on
```

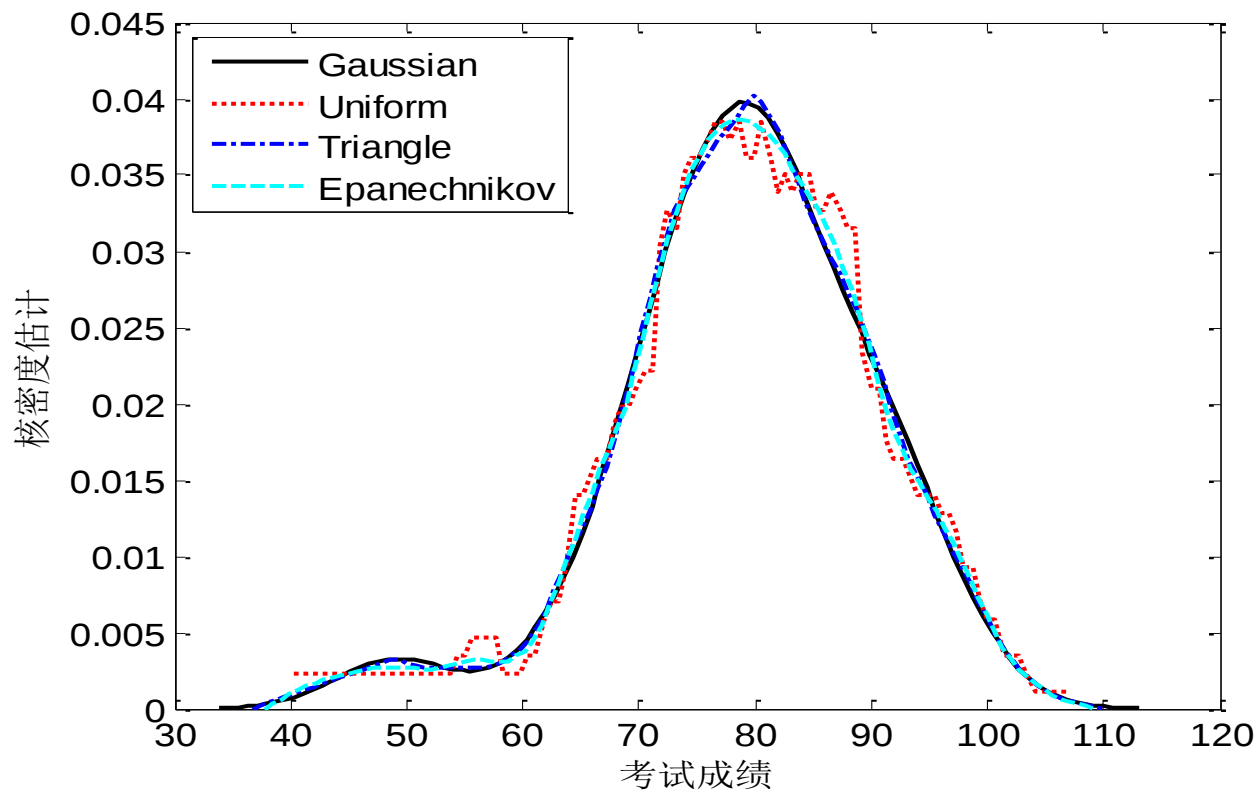
```
>> xlabel('考试成绩'); ylabel('核密度估计');
```

```
>> plot(xi2,f_ks2,'r:','linewidth',2)
```

```
>> plot(xi3,f_ks3,'b-.','linewidth',2)
```

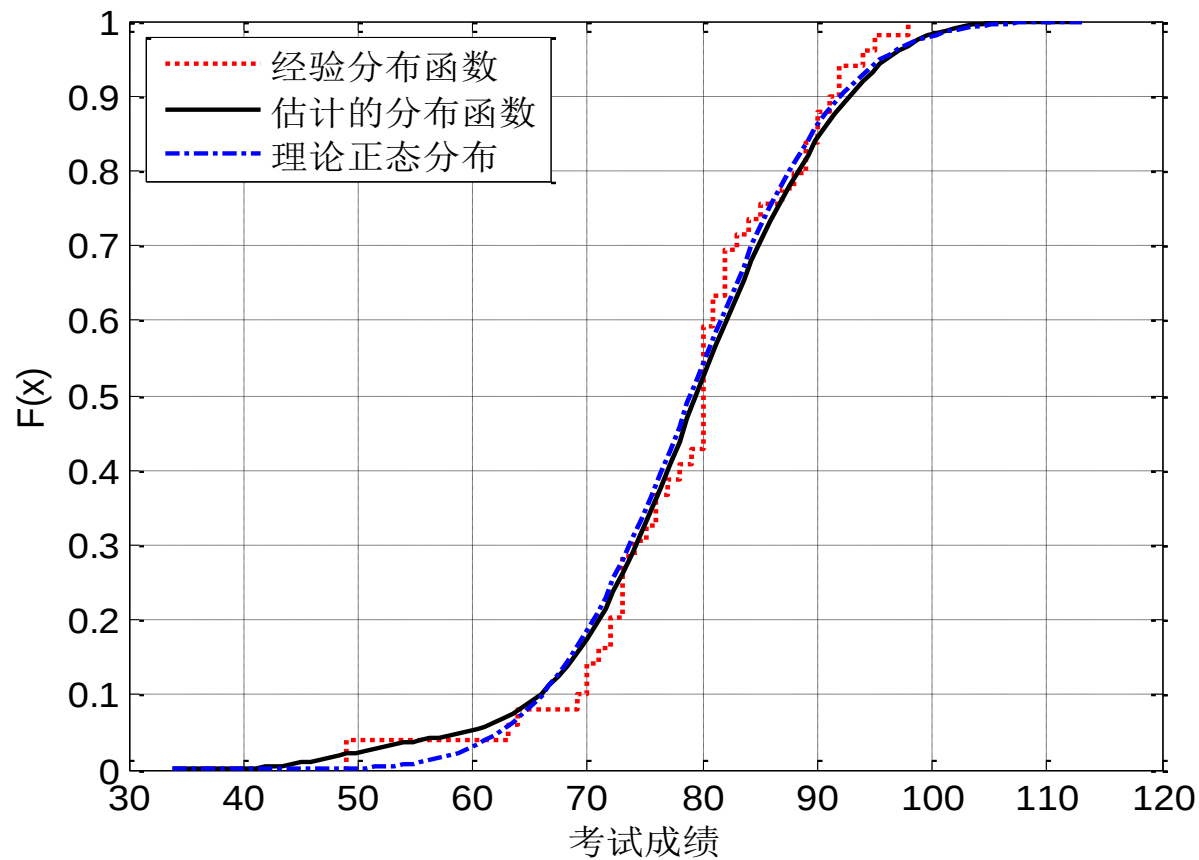
```
>> plot(xi4,f_ks4,'c--','linewidth',2)
```

```
>> legend('Gaussian' , 'Uniform' , 'Triangle' , ...  
'Epanechnikov','Location','NorthWest');
```



4. 累积分布的核估计

```
>> figure;  
>> [h,stats] = cdfplot(score);  
>> set(h,'color','r','LineStyle','-','LineWidth',2);  
>> hold on  
>> title('');  
>> xlabel('考试成绩');  
>> ylabel('F(x)');  
>> [f_ks, xi] = ksdensity(score,'function','cdf');  
>> plot(xi,f_ks,'k','linewidth',2);  
>> y = normcdf(xi,stats.mean,stats.std);  
>> plot(xi,y,'b-','LineWidth',2);  
>> legend('经验分布函数','估计的分布函数','理论正态分布',  
'Location','NorthWest');
```



Thank You