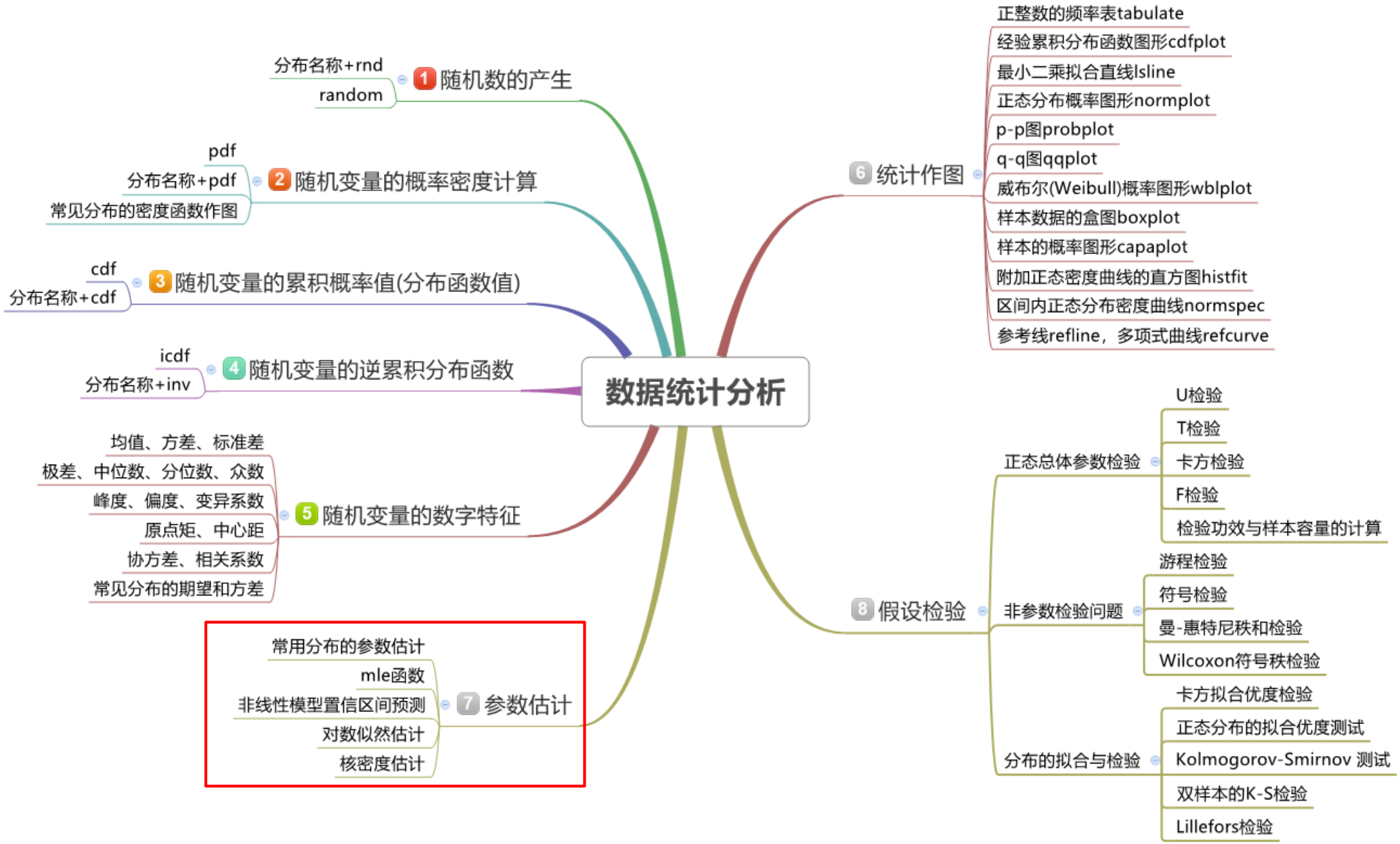
信陽師範学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

第10章 数据统计分析

讲授人：牛言涛

日期：2020年4月9日

第10章 数据统计分析知识点思维导图



10.7 参数估计

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

- 统计学有两大主要分支，分别是描述性统计学和推断统计学。
  - 描述性统计学用于描述和概括数据的特征以及绘制各类统计图表。
  - 总体数据，往往因为数据量太大而难以被获取，所以就有了通过较小的样本数据推测总体特性的推断统计学，如参数估计、假设检验等。
- 在很多实际的问题中，为了进行某些统计判断，需要确定总体所服从的分布，通常根据问题的实际背景或适当的统计方法可以判断总体分布的类型，但是总体分布中往往含有未知参数，需要用样本观测数据进行估计。
- 例如：学生的某门课程的考试成绩通常服从正态分布$N(\mu,\sigma^2)$，其中μ和σ是未知参数，就需要用样本观测数据进行估计，这就是所谓的参数估计，它是统计推断的一种重要形式。

1. 常用分布的参数估计函数表

信阳师范学院
数学与统计学院

调 用 形 式	函 数 说 明
[PHAT, PCI]= binofit (X, N, alpha)	二项分布，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
[Lambdahat, Lambdaci]= poissfit (X, alpha)	泊松分布，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
[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
[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
[ahat,bhat,ACI,BCI]=unifit(X, alpha)	均匀分布，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
[muhat,muci] = expfit(X,alpha)	指数分布，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
[phat,pci] = gamfit(X,alpha)	γ分布，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
[phat,pci] = wblfit(X,alpha)	韦伯分布，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
[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，pl为试验总次数。

1. 常用分布的参数估计函数表

例1：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$N(\mu,\sigma^2)$，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
```

由结果可知，金球测定的μ估计值为6.6782，置信区间为[6.6750，6.6813]；

σ的估计值为0.0039，置信区间为[0.0026，0.0081]。

泊球测定的μ估计值为6.6640，置信区间为[6.6611，6.6669]；

σ的估计值为0.0030，置信区间为[0.0019，0.0071]。

2. 参数估计—mle函数

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

利用mle函数进行极大似然参数估计

- mle(观察值, 'distribution', '分布名称')
- mle(观察值, 'pdf', 自定义分布pdf, 'start', 猜测的分布参数值)： pdf是分布的概率密度函数，格式是f(X,theta)，前面是X，后面跟参数值。
- [phat,pci] = mle(data,...,name1,val1,name2,val2,...)
- [phat,pci] = mle(data,'pdf',pdf,'cdf',cdf,'start',start,...)
- [...] = mle(data,'logpdf',logpdf,'logsf',logsf,'start',start,...)
- [...] = mle(data,'nloglf',nloglf,'start',start,...)

2. 参数估计—mle函数

例2：从某厂生产的滚珠中随机抽取10个，测得滚珠的直径（单位：mm）如下：

15.14 14.81 15.11 15.26 15.08 15.17 15.12 14.95 15.05 14.87

若滚珠直径服从正态分布$N(\mu,\sigma^2)$，μ和σ为未知，求μ、σ的极大似然估计和置信水平为90%的置信区间。

```
>> x = [15.14,14.81,15.11,15.26,15.08,15.17,15.12,14.95,15.05,14.87];
>> [muhat,sigmahat,muci,sigmaci] = normfit(x,0.1)
>> [mu_sigma,mu_sigma_ci] = mle(x,'distribution','norm','alpha',0.1)
muhat =
   15.0560
sigmahat =
    0.1397
muci =
   14.9750  15.1370
sigmaci =
    0.1019   0.2298
mu_sigma =
   15.0560    0.1325
mu_sigma_ci =
   14.9750    0.1019
   15.1370    0.2298
```

2. 参数估计—mle函数

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

例3：已知总体X的密度函数为 $f(x;\theta)=\begin{cases}\theta x^{\theta-1}, & 0<x<1\\ 0, & otherwise\end{cases}$

其中θ是未知参数。现从总体X中随机抽取容量为20的样本，得样本观测值如下：

0.7917 0.8448 0.9802 0.8481 0.7627 0.9013 0.9037 0.7399 0.7843 0.8424
0.9842 0.7134 0.9959 0.6444 0.8362 0.7651 0.9341 0.6515 0.7956 0.8733

试根据以上样本观测值求参数θ的最大似然估计和置信水平为95%的置信区间。

```
>> x = [0.7917,0.8448,0.9802,0.8481,0.7627, 0.9013, 0.9037, 0.7399,
0.7843, 0.8424, 0.9842, 0.7134, 0.9959, 0.6444, 0.8362, 0.7651, 0.9341,
0.6515, 0.7956, 0.8733];
>> PdfFun = @(x,theta) theta*x.^(theta-1).*(x>0 & x<1);
>> [phat,pci] = mle(x(:),'pdf',PdfFun,'start',1)
```

```
phat =
    5.1502
pci =
    2.8931
    7.4073
```

2. 参数估计—mle函数

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

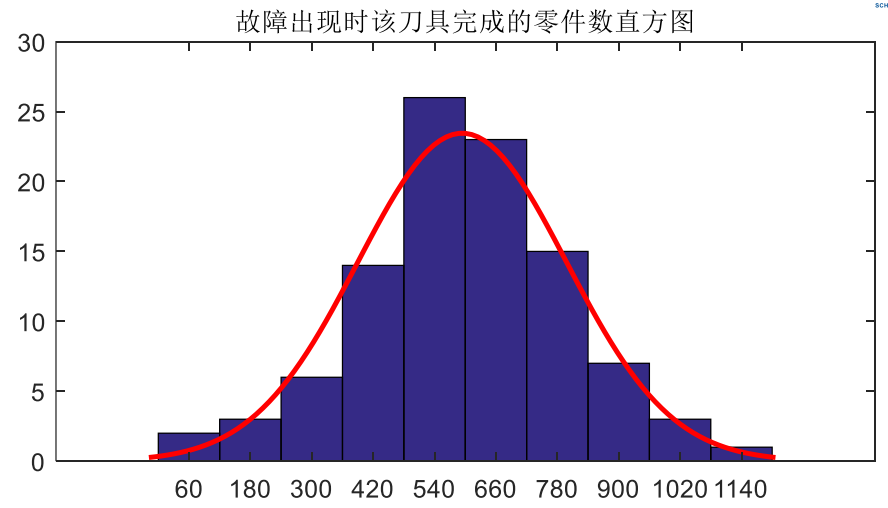
例4：一道工序用自动化车床连续加工某种零件，由于刀具损坏等会出现故障。故障是完全随机的，并假定生产任一零件时出现故障机会均相同。工作人员是通过检查零件来确定工序是否出现故障的。现积累有100次故障纪录，故障出现时该刀具完成的零件数如下：

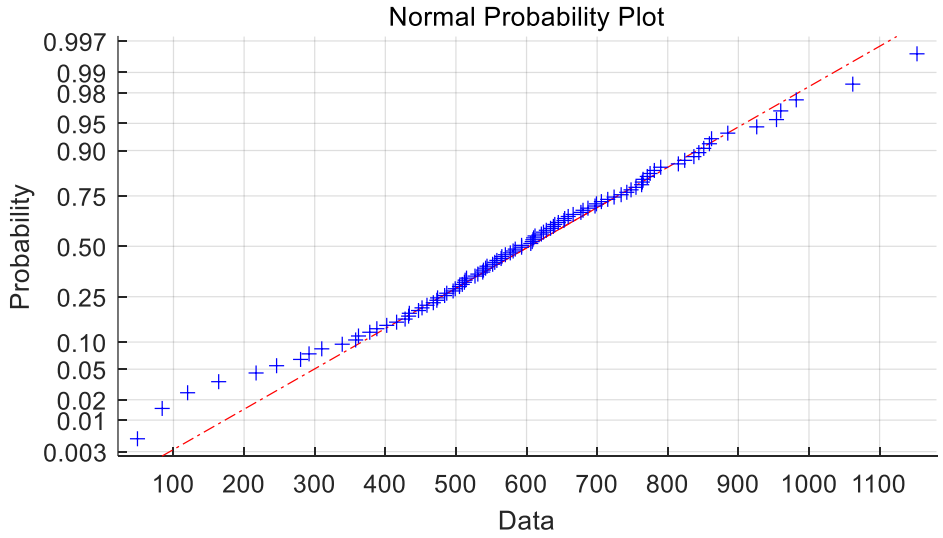
459	362	624	542	509	584	433	748	815	505
612	452	434	982	640	742	565	706	593	680
926	653	164	487	734	608	428	1153	593	844
527	552	513	781	474	388	824	538	862	659
775	859	755	49	697	515	628	954	771	609
402	960	885	610	292	837	473	677	358	638
699	634	555	570	84	416	606	1062	484	120
447	654	564	339	280	246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

试观察该刀具出现故障时完成的零件数属于哪种分布.

2. 参数估计—mle函数

```
>> kd = xlsread('knifedata.xlsx');
>> kd = kd(:);  %数据转换为一列
>> subplot(2,1,1)
>> histfit(kd)   %从直方图可以看出近似服从正态分布
>> title('故障出现时该刀具完成的零件数直方图')
>> subplot(2,1,2)
%近似服从正态分布，存在一定的厚尾特性，些许左偏
>> normplot(kd)
>> [phat,pci] = mle(kd,'distribution','norm','alpha',0.05)
phat =
  594.0000  203.1069
pci =
  553.4962  179.2276
  634.5038  237.1329
```





2. 参数估计—mle函数

例5：某校60名学生的一次考试成绩如下:

1) 计算均值、标准差、极差、偏度、峰度，画出直方图;

2) 检验分布的正态性;

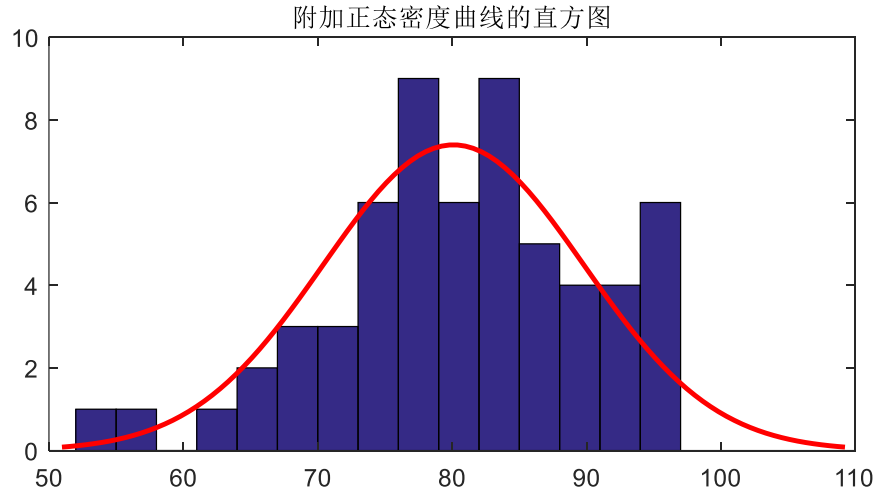
3) 若检验符合正态分布，估计正态分布的参数并检验参数。

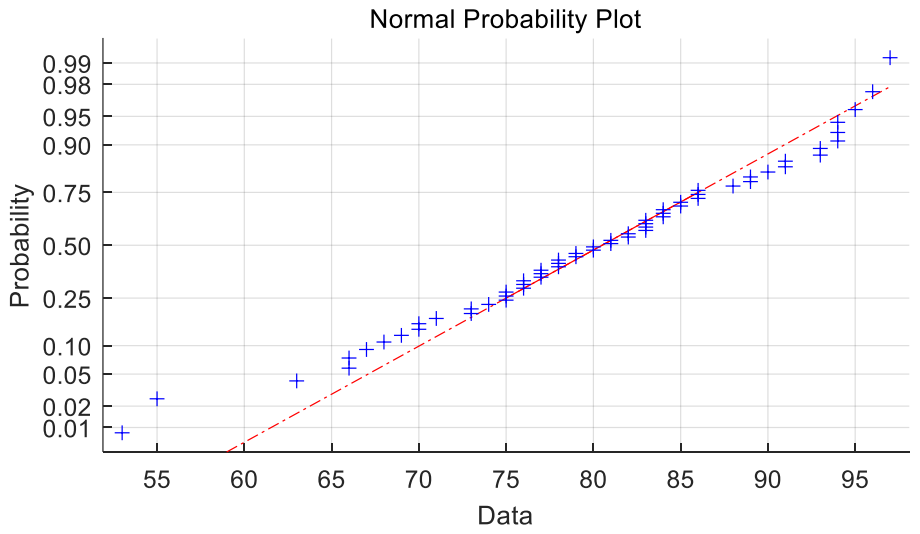
93	75	83	93	91	85	84	82	77	76
77	95	94	89	91	88	86	83	96	81
79	97	78	75	67	69	68	84	83	81
75	66	85	70	94	84	83	82	80	78
74	73	76	70	86	76	90	89	71	66
86	73	80	94	79	78	77	63	53	55

```
>> score = [93 75 83 93 91 85 84 82 77 76 77 95 94 89 91 88 86 83 96 81 79 97 78 75 67 69 68 84 83 81 75 66 85
70 94 84 83 82 80 78 74 73 76 70 86 76 90 89 71 66 86 73 80 94 79 78 77 63 53 55];
>> score = score (:);
>> stats = [mean(score ),var(score ),range(score ),skewness(score ),kurtosis(score )]
stats =  80.1000   94.2949   44.0000   -0.4682    3.1529
```

2. 参数估计—mle函数

```
>> subplot(2,1,1);
>> histfit(score,15)  %近似服从正态分布，存在严重左偏
>> title('附加正态密度曲线的直方图')
>> subplot(2,1,2);
>> normplot(score) %近似服从正态分布，略有厚尾特性
>> [phat,pci] = mle(score,'distribution','norm','alpha',0.05)
phat =
   80.1000    9.6293
pci =
   77.5915    8.2310
   82.6085   11.8436
```





3. 非线性模型置信区间预测

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

高斯—牛顿法的非线性最小二乘数据拟合，函数 nlinfit

- [beta,r,J] = nlinfit(X,y,FUN,beta0) %返回在FUN中描述的非线性函数的系数。FUN形式$\hat{y} = f(\beta, X)$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$\hat{y}$ 。beta为拟合系数，r为残差，J为Jacobi矩阵，beta0为初始预测值。
- 说明：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。

非线性模型的参数估计的置信区间，函数nlparci

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。

3. 非线性模型置信区间预测

例6：调用MATLAB提供的数据文件reaction.mat，进行非线性参数估计和置信区间

```
>> load reaction  %加载数据
>> [beta,resids,J] = nlinfit(reactants,rate,@hougen,beta);  %非线性参数估计
betafit =
    1.2526  0.0628  0.0400  0.1124  1.1914
>> ci = nlparci(beta,resids,J)
ci =  %参数置信区间
   -0.7467    3.2519
   -0.0377    0.1632
   -0.0312    0.1113
   -0.0609    0.2857
   -0.7381    3.1208
```

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

变量 - reactants

beta | reactants | rate

13x3 double

	1	2	3
1	470	300	10
2	285	80	10
3	470	300	120
4	470	80	120
5	470	80	10
6	100	190	10
7	100	80	65
8	470	190	65
9	100	300	54
10	100	300	120
11	100	80	120
12	285	300	10
13	285	190	120

变量 - rate

beta | reactants | rate

13x1 double

	1	2	3
1	8.5500		
2	3.7900		
3	4.8200		
4	0.0200		
5	2.7500		
6	14.3900		
7	2.5400		
8	4.3500		
9	13		
10	8.5000		
11	0.0500		
12	11.3200		
13	3.1300		

变量 - beta

beta | reactants | rate

5x1 double

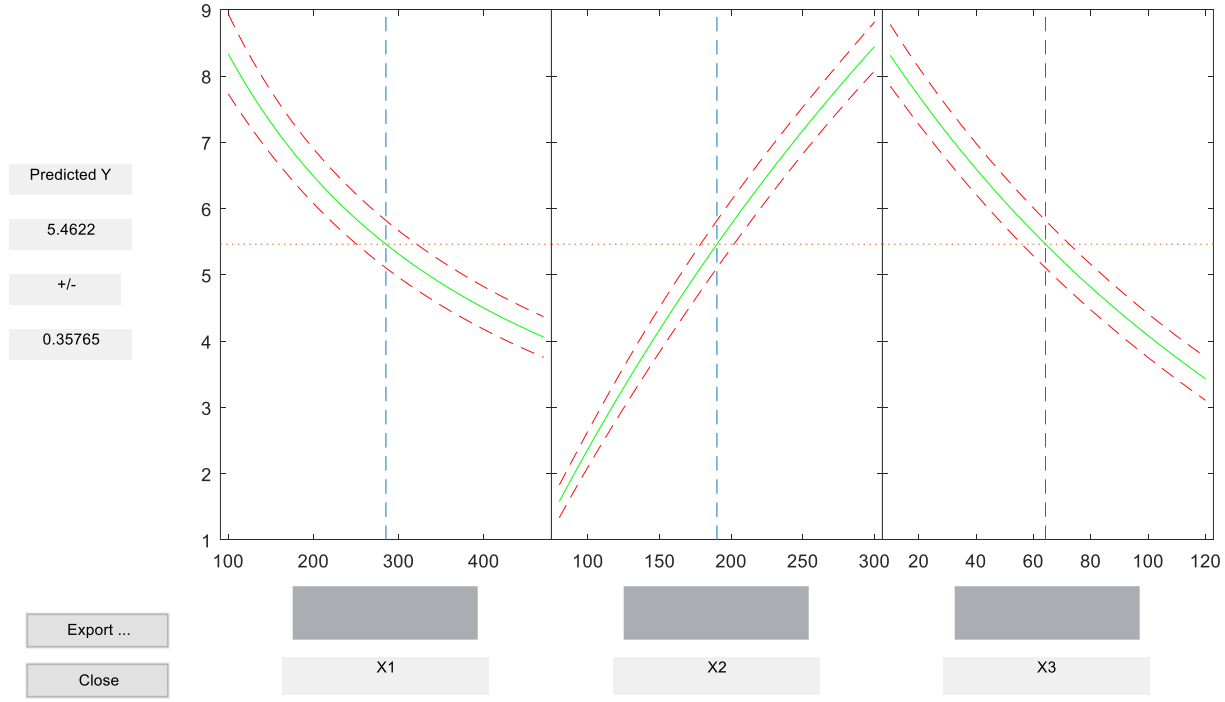
	1	2	3
1	1		
2	0.0500		
3	0.0200		
4	0.1000		
5	2		

3. 非线性模型置信区间预测

- 非线性拟合和显示交互图形，函数nlintool
  - 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)×100%

load reaction

nlintool(reactants,rate,@hougen,beta)



3. 非线性模型置信区间预测

非线性模型置信区间预测，函数nlpredci

- [ypred,delta] = 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， 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,delta] = 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。

3. 非线性模型置信区间预测

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

例7：炼钢厂出钢时所用盛钢水的钢包，由于钢水对耐火材料的侵蚀，容积不断增大，我们希望找出使用次数与增大容积之间的函数关系，实验数据如表：

使用次数x	2	3	4	5	6	7	8	9
增大容积y	6.42	8.2	9.58	9.5	9.7	10	9.93	9.99
使用次数x	10	11	12	13	14	15	16	
增大容积y	10.49	10.59	10.6	10.8	10.6	10.9	10.76	

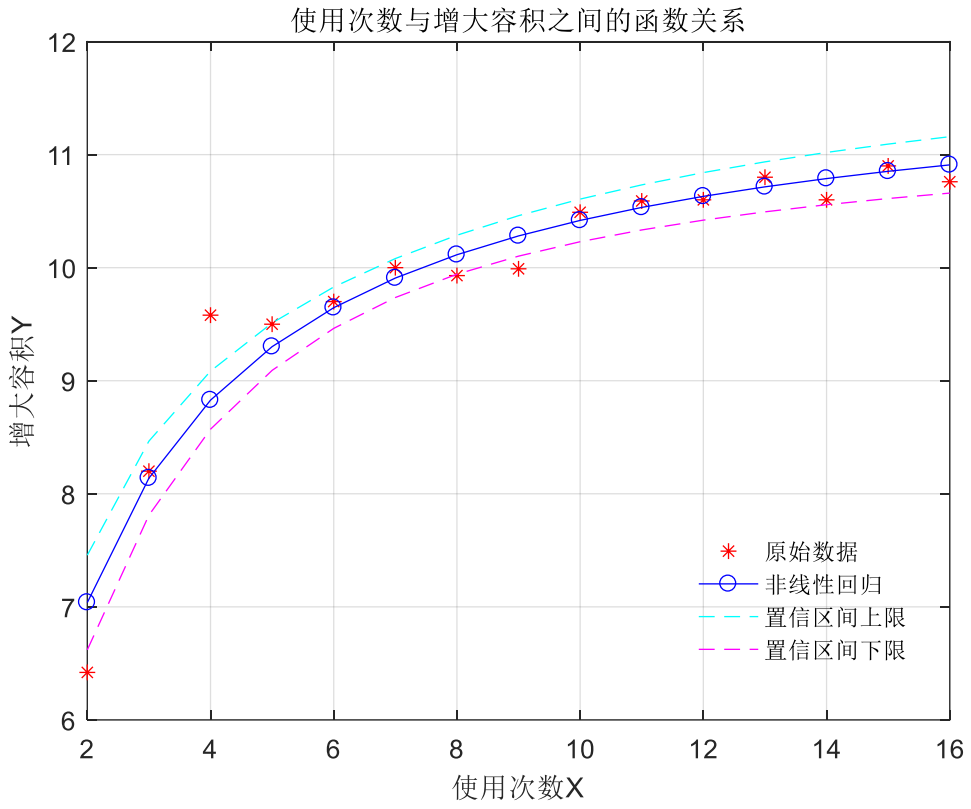
(1) 建立非线性回归模型$\boldsymbol{1/y = a + b/x}$;

(2) 预测钢包使用$\boldsymbol{x0 = 17}$次后增大的容积$\boldsymbol{y0}$;

(3) 计算回归模型参数的95%的置信区间。

3. 非线性模型置信区间预测

```
x = [2:16];
y = [6.42 8.2 9.58 9.5 9.7 10 9.93 9.99 10.49 10.59 10.6 10.8 10.6 10.9 10.76];
b0 = [0.1,0.1];  %参数的初始估计
fun = @(b,x)x./(b(1)*x+b(2));  %模型
[beta,r,J] = nlinfit(x,y,fun,b0)   %非线性参数估计
ypred = nlpredci(fun,17,beta,r,J,0.05)     %预测ypred  = 10.9599
ci = nlparci(beta,r,J)   %置信区间
[ypredt,delta] = nlpredci(fun,x,beta,r,J) %根据建立的模型进行预测
plot(x,y,'r*',x,ypredt,'-ob')  %绘制测量数据
grid on;  hold on
plot(x,ypredt+delta,'c--');  plot(x,ypredt-delta,'m--')
legend('原始数据','非线性回归','置信区间上限','置信区间下限');
legend('boxoff')
title('使用次数与增大容积之间的函数关系')
nlintool(x,y,fun,beta)
```



3. 非线性模型置信区间预测

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

例8：为了分析X射线的杀菌作用，用200千伏的X射线来照射细菌，每次照射6分钟用平板计数法估计尚存活的细菌数，照射次数记为t，照射后的细菌数y如表所示：

t	1	2	3	4	5	6	7	8
y	352	211	197	160	142	106	104	60
t	9	10	11	12	13	14	15	
y	56	38	36	32	21	19	15	

(1) 给出y与t的回归模型 $y = ae^{bt}$

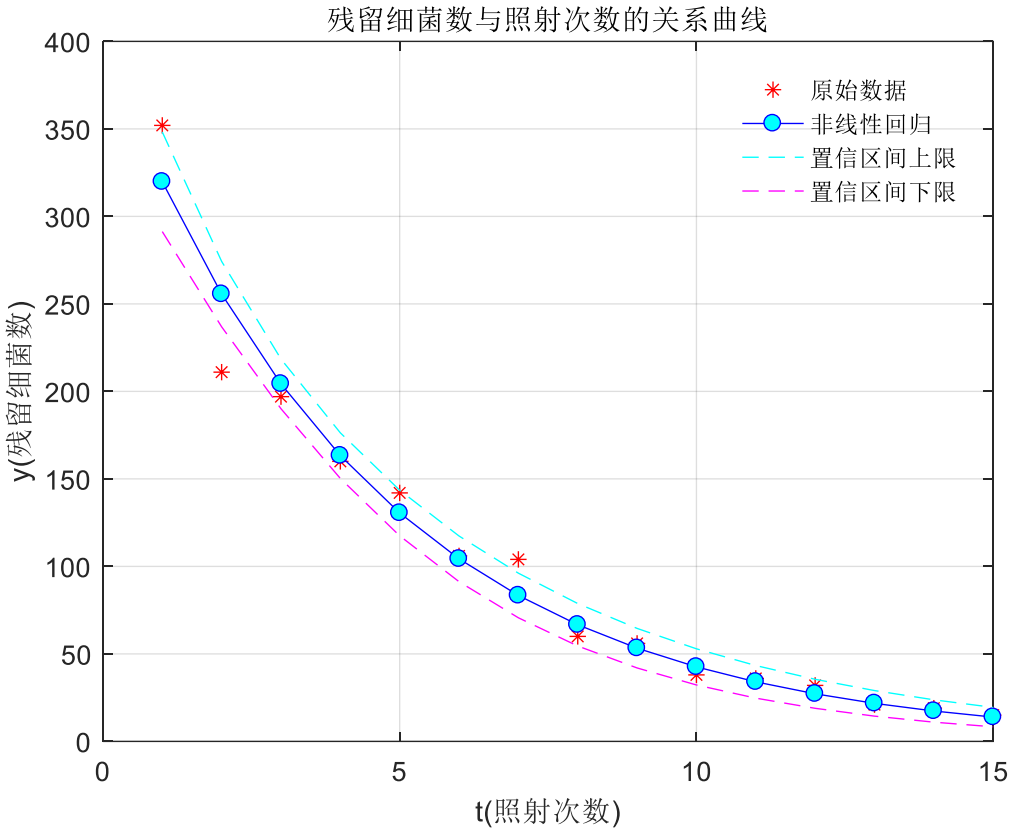
(2) 在同一坐标系内做出原始数据与拟合结果的散点图

(3) 预测t=16时残留的细菌数

(4) 根据问题实际意义选择多项式函数是否合适?

3. 非线性模型置信区间预测

```
t = 1:15;  y = [352 211 197 160 142 106 104 60 56 38 36 32 21 19 15];
beta0 = [148,-0.2];
fun = @(b,t)b(1)*exp(b(2)*t);
[beta,r,J] = nlinfit(t,y,fun,beta0)
ypred16 = nlpredci(fun,16,beta,r,J)    %ypred  = 11.1014
ci = nlparci(beta,r,J)   %beta参数的置信区间
[ypredt,delta] = nlpredci(fun,t,beta,r,J);   %绘制预测值的置信区间
plot(t,y,'r*');
hold on;grid on
plot(t, ypredt,'b-o','MarkerFaceColor','c')
xlabel('t(照射次数)'),ylabel('y(残留细菌数)');
plot(t,ypredt+delta,'c--'); plot(t,ypredt-delta,'m--')
legend('原始数据','非线性回归','置信区间上限','置信区间下限');
legend('boxoff')
title('残留细菌数与照射次数的关系曲线')
```



4. 对数似然函数

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

负β分布的对数似然函数，函数betalike

- 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β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。**

4. 对数似然函数

负γ分布的对数似然估计，函数Gamlike

- logL=gamlike(params,data) %返回由给定样本数据data确定的γ分布的参数为params（即[a，b]）的负对数似然函数值
- [logL,info]=gamlike(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 gamlike是γ分布的最大似然估计函数。因为gamlike返回对数似然函数值，故用fmins函数将gamlike最小化后，其结果与最大似然估计是相同的。

4. 对数似然函数

负正态分布的对数似然函数，函数normlike

- 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[mu，sigma])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。

4. 对数似然函数

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

威布尔分布的对数似然函数，函数weiblike

- logL = weiblike(params,data) %返回由给定样本数据data确定的、威布尔分布的、参数为params(即[a，b])的对数似然函数值。
- [logL,info]=weiblike(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 威布尔分布的负对数似然函数定义为**

$$-\log L = -\log \prod_{i=1}^{n} f(a,b \mid x_i) = -\sum_{i=1}^{n} \log f(a,b \mid x_i)$$

5. 非参数估计

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

1.参数估计方法

- 假定样本集符合某一概率分布，然后根据样本集拟合该分布中的参数，例如：似然估计，混合高斯等，由于参数估计方法需要加入主观先验知识，往往很难拟合出真实分布的模型；

2.非参数估计

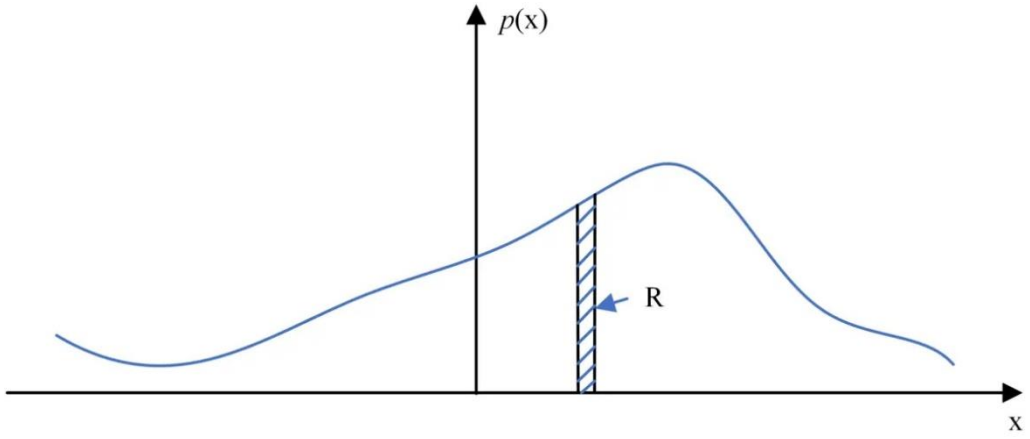
- 非参数估计并不加入任何先验知识，而是根据数据本身的特点、性质来拟合分布，这样能比参数估计方法得出更好的模型。
- 核密度估计就是非参数估计中的一种，由Rosenblatt (1955)和Emanuel Parzen(1962)提出，又名Parzen窗（Parzen window）。Ruppert和Cline基于数据集密度函数聚类算法提出修订的核密度估计方法。

5. 非参数估计

https://www.jianshu.com/p/249e5ff97c04

如右图，对于一个未知的概率密度函数$p(x)$，某一个随机变量x落在区域R里的概率可以表示成：$P=\int_R p(x)dx$

如果R足够窄，可以用P来表示$p(x)$的一个平均后的结果。



假设现在有n个样本，且都服从独立同分布，那么n个样本中的 k 个落在区域 R 中的概率可以表示成公式：$P_k=C_n^k P^k(1-P)^{n-k}$

由公式可得到k的期望$E(X)=np$。根据大数定理，当n足够大时，可以认为$P\approx k/n$。

假设n足够大，R足够小，并且假设$p(x)$是连续的，那么可以得到：$\int_R P(x)dx\approx p(x)\cdot V$

其中x是区域R中的一个点，V是R的面积(体积)，结合上述，得：

$$p(x)\cdot V=\int_R P(x)dx=P=\frac{E(k)}{n}\Rightarrow p(x)=\frac{E(k)}{nV}\approx\frac{k}{nV}$$

因此，某一小区域内的概率密度函数就可以用上述公式表示了。

5. 非参数估计

$$p(x) \approx \frac{k}{nV} = \frac{k/n}{V} \approx \frac{P}{V} = \frac{\int_R p(x)dx}{\int_R dx}$$

- 显然估计的这个概率密度是一个平滑的结果，即当V选择的越大，估计的结果和真实结果相比就越平滑；因此需要把V设置的小一点，然而如果把V选择的过小，也会出现问题：
  - 太小的V会导致这块小区域里面没有一个点落在里面，因此该点的概率密度为0；
  - 另外，假设刚好有一个点落在了这个小区域里，由于V过于小，计算得到的概率密度可能会趋近于无穷。两个结果对于我们来说都是没有太大意义。
- 从实际的角度来看，我们获取的数据量一定是有限的，因此体积不可能取到无穷小。使用非参数概率密度估计有以下两方面限制，且是不可避免的：
  - 在有限数据下，使用非参数估计方法计算的概率密度一定是真实概率密度平滑后的结果。
  - 在有限数据下，体积趋于无穷小计算的概率密度没有意义。

5. 非参数估计

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

从理论的角度来看，我们希望知道如果有无限多的采样数据，那么上述两个限制条件应该怎样克服？假设使用下面的方法来估计点x处的概率密度：构造一系列包含x的区域R_1，R_2，…，R_n，其中R_1中包含一个样本，R_n中包含n个样本．则：

$$p_n(x)=\frac{k_n/n}{V_n}$$

其中$p_n(x)$表示第n次估计结果，如果要求$p_n(x)$能够收敛到$p(x)$，则需要满足下面三个条件：

$$\lim_{n\to\infty}V_n=0,\quad \lim_{n\to\infty}k_n=\infty,\quad \lim_{n\to\infty}\frac{k_n}{n}=0$$

6. 核密度估计

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

假设R_n是一个d维的超立方体(hypercube)，且其边长为h，那么可以用公式表示V_n：$V_n = h_n^d$

然后再定义一个窗函数（window function）：

$$\varphi(u) = \begin{cases} 1, & |u_i| \le 1/2 \\ 0, & otherwise \end{cases}$$

φ定义了一个以圆点为中心的单位超立方体，这样就可以用φ来表示体积V内的样本个数:

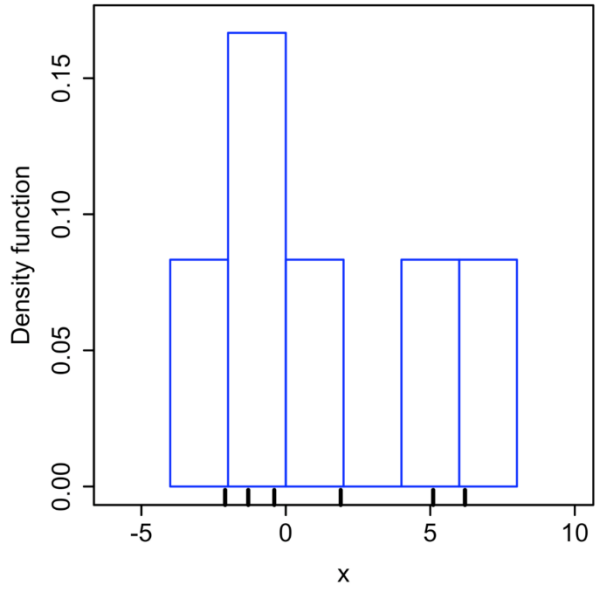
$$k_n = \sum_{i=1}^{n} \varphi\left(\frac{x - x_i}{h_n}\right)$$

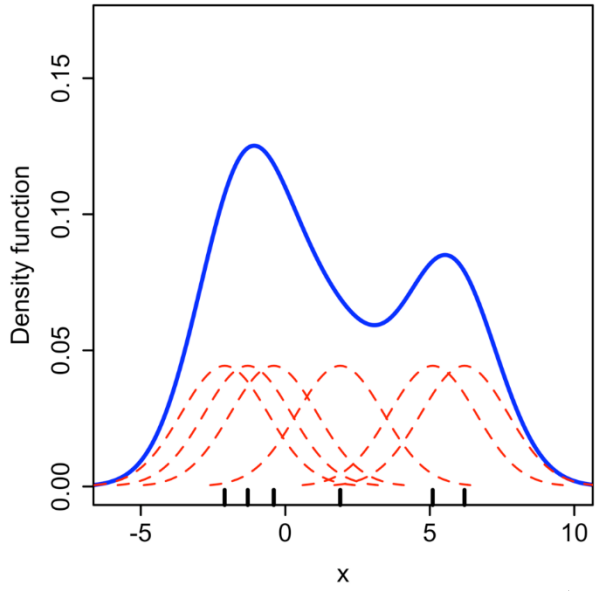
有了k_n和V_n，直接把他们带入公式，可以得到parzen窗法的计算公式: $$p_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi\left(\frac{x - x_i}{h_n}\right)$$

我们发现这个φ不仅可以是上述的单位超立方体的形式，只要其满足约束 $\varphi(x) \ge 0,\ \int \varphi(u)du = 1$ 就可以，因此也就出现了各种各样更能表现样本属性的窗函数，比如用的非常多的高斯窗。

6. 核密度估计

某一点的概率密度是其他样本点在这一点的概率密度分布的平均值，如图：





上面一句话可以这样解释，定义核函数：$\delta(x)=\frac{1}{V_n}\varphi\left(\frac{x}{h_n}\right)$

那么某一点x的概率密度可以用如下函数来表示：

$$p_n(x)=\frac{1}{n}\sum_{i=1}^{n}\delta_n(x-x_i)$$

从公式可以看出，当h_n很大的时候，$\delta_n(x)$就是一个 矮胖的函数，公式将每个样本点在点x处的贡献取平均之后，点x处的概率密度就是一个非常平滑的结果；

当h_n太小的时候，$\delta_n(x)$就是一个高瘦的函数，公式将每个样本点在点x处的贡献取平均之后，点x处的概率密度就是一个受噪声影响非常大的值，因此估计的概率密度平滑性就很差，反而和真实值差的很远。

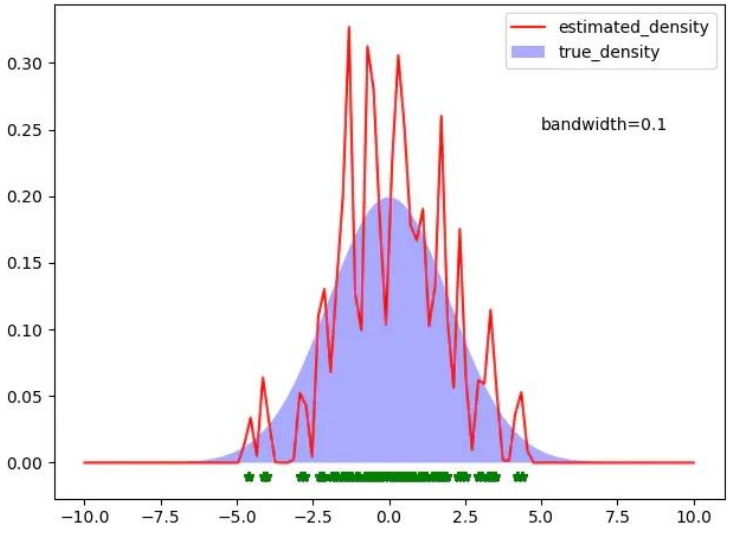
这两点和总结的两点缺陷正好吻合.

6. 核密度估计

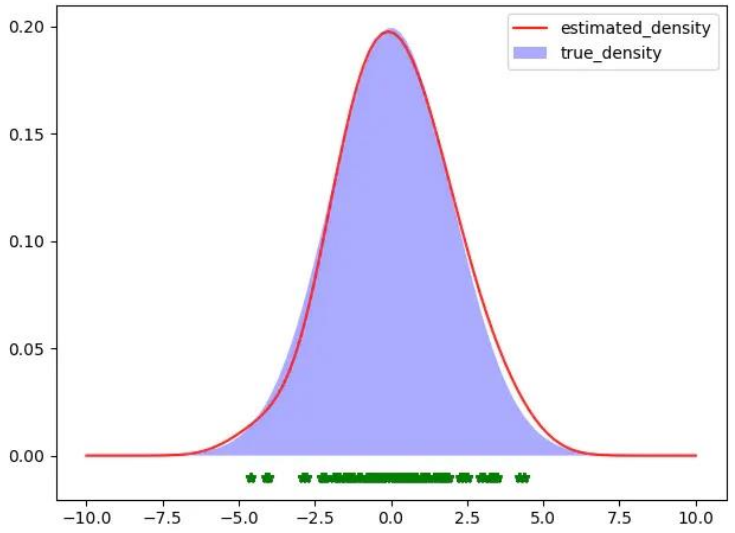
信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

仿真实验：生成了均值是0，方差是2的服从高斯分布的数据，分别使用bandwidth为0.1, 1, 5三个值进行估计 。

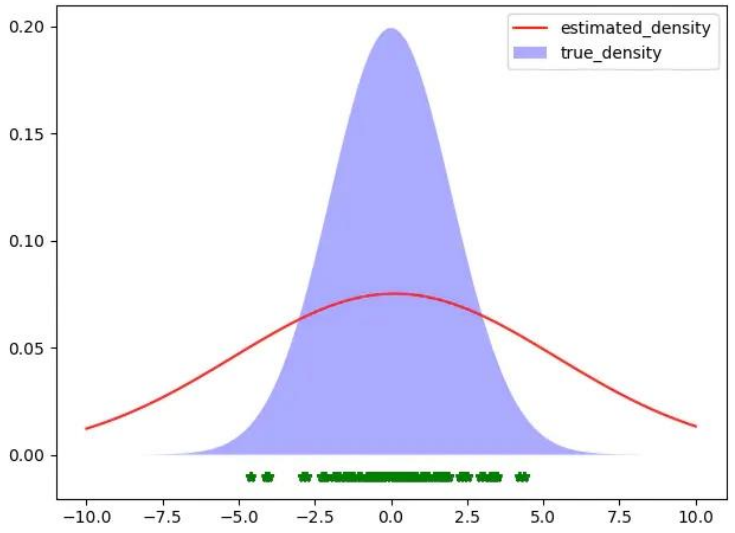
bandwidth = 0.1



bandwidth = 1



bandwidth = 5



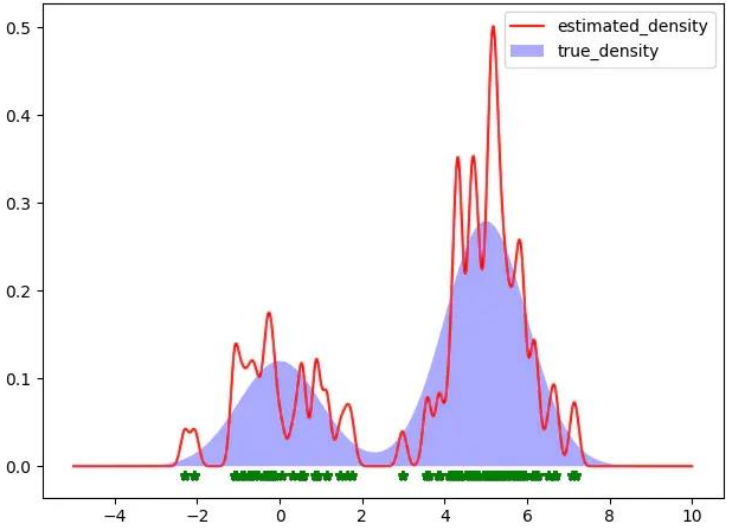
可以很明显得看到估计的概率密度是如何受到bandwidth影响的，当bandwidth选择的太小，则估计的密度函数受到噪声影响很大，这种结果是不能用的；当bandwidth选择过大，则估计的概率密度又太过于平滑。总之，无论bandwidth过大还是过小，其结果都和实际情况相差的很远，因此合理地选择bandwidth是很重要的。

6. 核密度估计

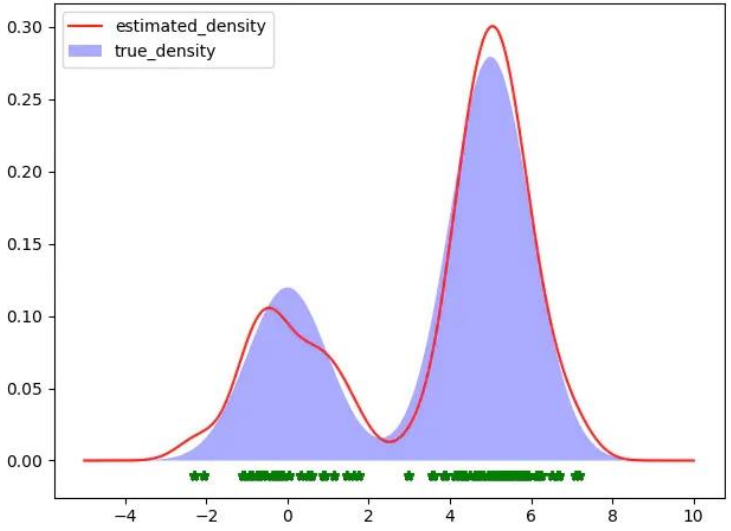
信阳师范学院
数学与统计学院

仿真实验：生成了100个服从高斯混合分布的数据，分别是均值为0、方差为1以及均值为5、方差为1的两个高斯混合模型，两者相互独立。

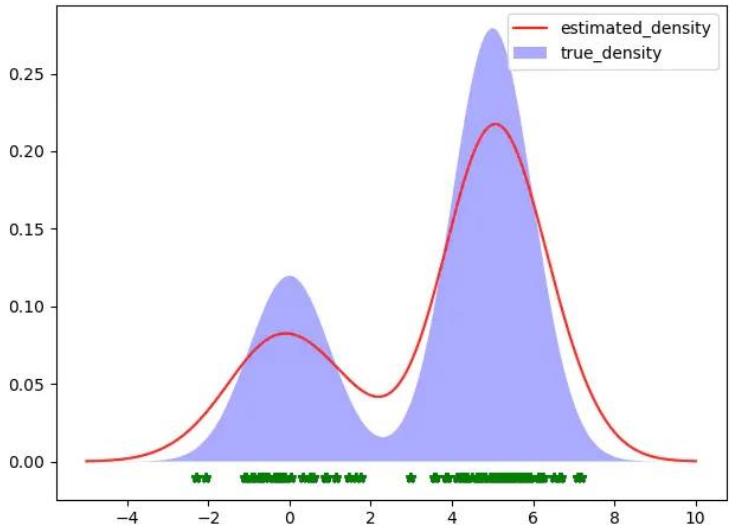
bandwidth = 0.1



bandwidth = 1



bandwidth = 5



可以很明显得看到估计的概率密度是如何受到bandwidth影响的，当bandwidth选择的太小，则估计的密度函数受到噪声影响很大，这种结果是不能用的；当bandwidth选择过大，则估计的概率密度又太过于平滑。总之，无论bandwidth过大还是过小，其结果都和实际情况相差的很远，因此合理地选择bandwidth是很重要的。

6. 核密度估计

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

- 所谓核密度估计，就是采用平滑的峰值函数(“核”)来拟合观察到的数据点，从而对真实的概率分布曲线进行模拟。
- 核密度估计（Kernel density estimation），是一种用于估计概率密度函数的非参数方法，为独立同分布F的n个样本点，设其概率密度函数为f，核密度估计为以下：

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

$K(.)$为核函数（非负、积分为1，符合概率密度性质，并且均值为0）。$h>0$为一个平滑参数，称作带宽(bandwidth)或窗口。$K_h(x)=1/h\,K(x/h)$为缩放核函数(scaled Kernel)。

5. 核密度估计

核密度估计需要指定核函数和窗宽，但是取不同的核函数对核密度估计影响不大。

（1）常用的核函数：Uniform（或Box）、Triangle、Epanechnikov、Quaritic、Triweight、Gaussian、Cosinus

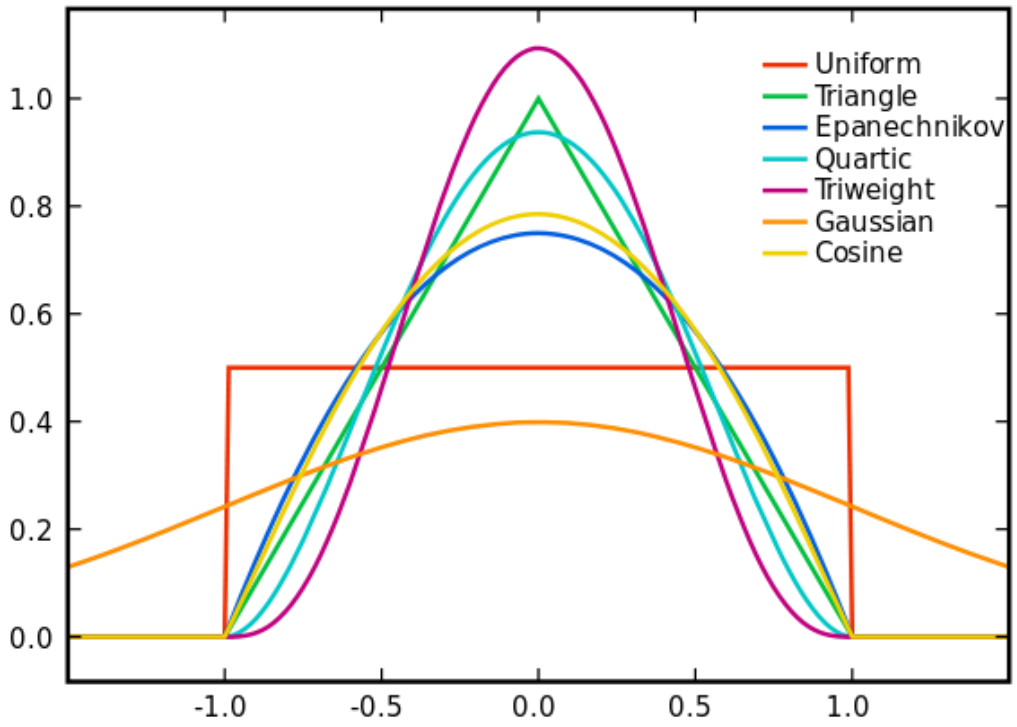
（2）窗宽对核密度估计的影响

- 窗宽会影响光滑程度，如果窗宽h取较大的值，图形较为光滑，但同时也丢失了数据所包含的一些信息；
- 如果窗宽取值较小，则图像是不光滑的曲线，但它能反映出每个数据所包含的信息。

（3）核密度估计的MATLAB实现

- ksdensity函数，用来求核密度估计。

5. 核密度估计



$$(Uniform)\,K(u)=\frac{1}{2}\quad Support:|u|\le 1$$

$$(Triangular)\,K(u)=(1-|u|)\quad Support:|u|\le 1$$

$$(Epanechnikov)\,K(u)=\frac{3}{4}(1-u^2)\quad Support:|u|\le 1$$

$$(Quartic)\,K(u)=\frac{15}{16}(1-u^2)^2\quad Support:|u|\le 1$$

$$(Triweight)\,K(u)=\frac{35}{32}(1-u^2)^2\quad Support:|u|\le 1$$

$$(Tricube)\,K(u)=\frac{70}{81}(1-|u|^3)^3\quad Support:|u|\le 1$$

$$(Gaussian)\,K(u)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2};\ (Cosine)\,K(u)=\frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right)\quad Support:|u|\le 1$$

5. 核密度估计

- [f,xi]=ksdensity(x)：求样本观测向量x的核密度估计，xi是在x取值范围内等间隔选取的100个点构成的向量，f是与xi相对应的核密度估计值向量。所用的核函数是Gaussian核函数，窗宽也是默认值。
- [f,xi]=ksdensity(x,pts)：根据样本观测向量x计算pts处的核密度估计值f，xi和f是等长的向量。

`[f,xi] = ksdensity(x)` returns a probability density estimate, `f`, for the sample data in the vector or two-column matrix `x`. The estimate is based on a normal kernel function, and is evaluated at equally-spaced points, `xi`, that cover the range of the data in `x`. `ksdensity` estimates the density at 100 points for univariate data, or 900 points for bivariate data.

`ksdensity` works best with continuously distributed samples.

`[f,xi] = ksdensity(x,pts)` specifies points (`pts`) to evaluate `f`. Here, `xi` and `pts` contain identical values.

`[f,xi] = ksdensity( ___ ,Name,Value)` uses additional options specified by one or more name-value pair arguments in addition to any of the input arguments in the previous syntaxes. For example, you can define the function type `ksdensity` evaluates, such as probability density, cumulative probability, survivor function, and so on. Or you can specify the bandwidth of the smoothing window.

`[f,xi,bw] = ksdensity( ___ )` also returns the bandwidth of the kernel smoothing window, `bw`. The default bandwidth is the optimal for normal densities.

`ksdensity( ___ )` plots the kernel smoothing function estimate.

`ksdensity(ax, ___ )` plots the results using axes with the handle, `ax`, instead of the current axes returned by `gca`.

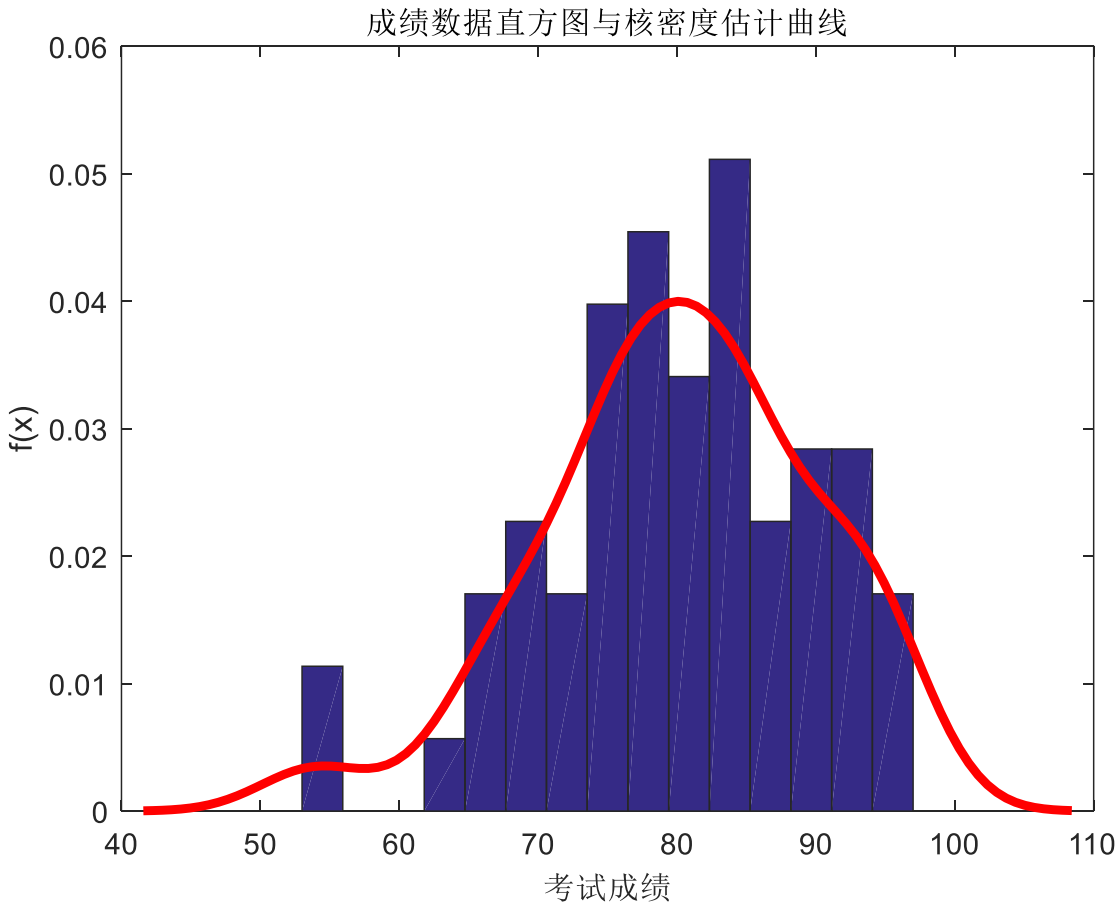
5. 核密度估计

[……]=ksdensity（……，param1，val1，param2，val2，….）

参数名	参数值	说明	参数名	参数值	说明
censoring	与x等长的逻辑向量	指定哪些项是截尾观测，默认是没有截尾	support	unbounded	指定密度函数的支撑集为全体实数集，默认情况
kernel	normal	指定用Gaussian（高斯或正态）核函数，默认情况		positive	指定密度函数的支撑集为正实数集
	box	指定用Uniform核函数		包含两个元素的向量	指定密度函数的支撑集的上下限
	triangle	指定用Triangle核函数	function	pdf	指定对密度函数进行估计
	epanechnikov	指定用Epanechnikov核函数		cdf	指定对累积分布函数估计
	函数句柄或函数名	自定义核函数		icdf	指定对逆概率分布函数估计
npoints	正整数	指定xi中包含的等间隔点的个数，默认100		survior	指定对生存函数进行估计
weights	与x等长的向量	指定x中元素的权重		cumhazard	指定对累积危险函数进行估计

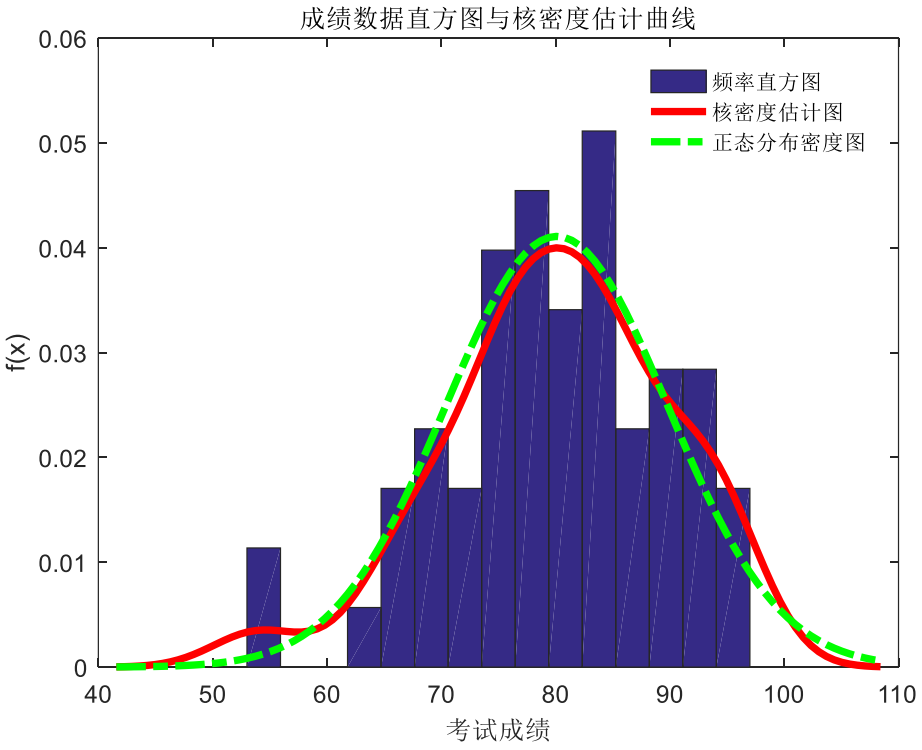
5. 核密度估计

```
score = [93 75 83 93 91 85 84 82 77 76 77 95 94 89 91 88 86 83 96 81 79 97 78 75 67 69 68 84 83 81 75 66
85 70 94 84 83 82 80 78 74 73 76 70 86 76 90 89 71 66 86 73 80 94 79 78 77 63 53 55]';
%调用ecdf函数计算xc处的经验分布函数值f_ecdf
[f_ecdf,xc]=ecdf(score);
%绘制成绩直方图，直方图对应17个小区间
ecdfhist(f_ecdf,xc,15);
hold on;
xlabel('考试成绩');  ylabel('f(x)');
%调用ksdensity函数进行核密度估计
[f_ks1,xi1,u1]=ksdensity(score);
plot(xi1,f_ks1,'r','linewidth',3);  %绘制核密度估计图
title('成绩数据直方图与核密度估计曲线')
```



5. 核密度估计

```
>> ms=mean(score); %均值
>> ss=std(score);  %方差
%计算xi1处的正态分布密度函数值，正态分布的均值是ms，方差是ss
>> f_norm=normpdf(xi1,ms,ss);
%绘制正态分布密度函数图，并设置线条颜色为红色点画线，宽3
>> plot(xi1,f_norm,'g-.','linewidth',3);
>> legend('频率直方图','核密度估计图','正态分布密度图');
>> legend('boxoff')
>> u1  %查看默认的窗宽u1
u1 =
    3.8084
```

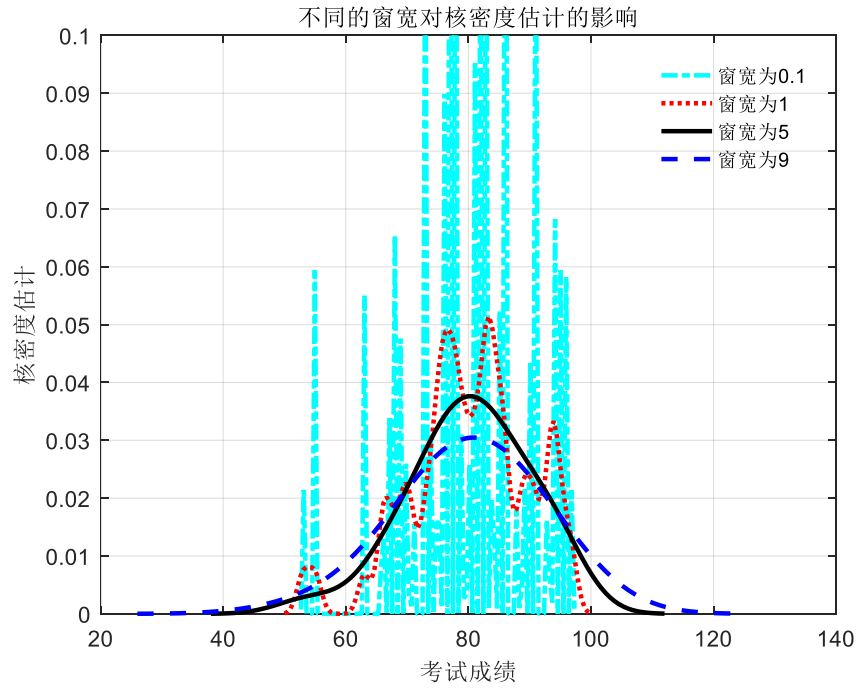


在默认窗宽下，利用Gaussian核函数求出的密度曲线与N(ms, ss)分布的密度曲线非常接近，与总成绩的频率直方图附和的也很好。

5. 核密度估计

固定核函数为Gaussian核函数，让窗宽进行变动，观察不同的窗宽对核密度估计的影响。

```
%设置窗宽分别为0.1, 1, 5, 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);
%分别绘制不同窗对应的核密度估计图，他们对应不同的线型和颜色
plot(xi1,f_ks1,'c-.','linewidth',2);
hold on; grid on
xlabel('考试成绩');  ylabel('核密度估计');
plot(xi2,f_ks2,'r:',xi3,f_ks3,'k',xi4,f_ks4,'b--','linewidth',2);
legend('窗宽为0.1','窗宽为1','窗宽为5','窗宽为9');
legend('boxoff')
axis([20,140,0,0.1])
title('不同的窗宽对核密度估计的影响')
```

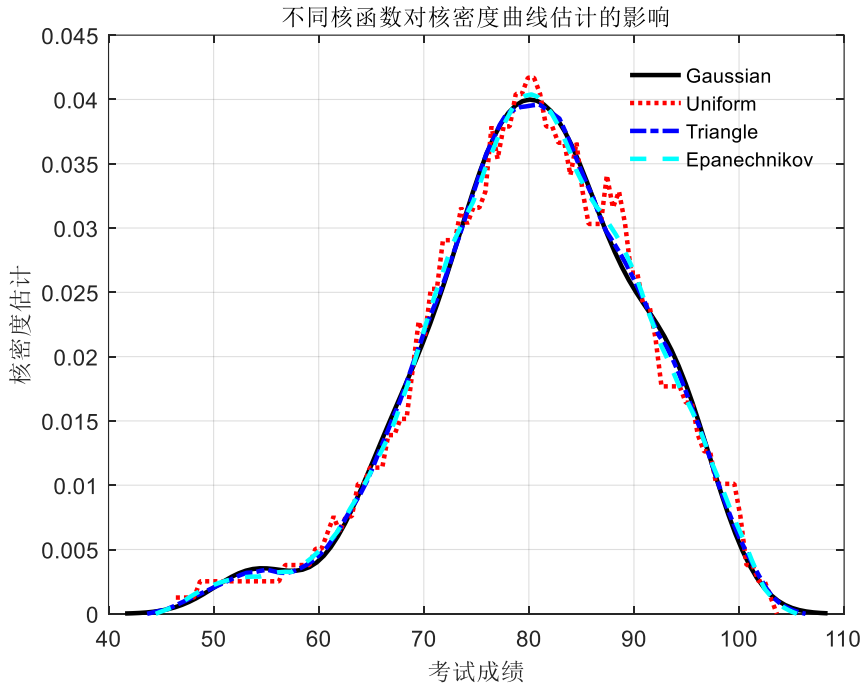


由图可以发现，不同的窗宽下，核密度估计曲线形状差距比较大，对于比较小的窗宽值，核密度估计曲线比较曲折，光滑性很差，但是反映了较多的细节；对于比较大的窗宽值，核密度估计曲线比较光滑，但是掩盖了许多细节。

5. 核密度估计

固定窗宽为默认的最佳窗宽，让核函数变动，观察不同核函数对核密度曲线估计的影响。

```
%设置核函数分别为Gaussian，Uniform，Triangle和Epanechnikov
%调用ksdensity函数进行核密度检验
[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');
%分布绘制不同核函数所对应的核密度估计图
plot(xi1,f_ks1,'k','linewidth',2);
hold on; grid on
plot(xi2,f_ks2,'r:',xi3,f_ks3,'b-.',xi4,f_ks4,'c--','linewidth',2);
xlabel('考试成绩');  ylabel('核密度估计');
legend('Gaussian','Uniform','Triangle','Epanechnikov');
legend('boxoff')
title('不同核函数对核密度曲线估计的影响')
```



通过上图可以看出，不同的核函数对核密度估计的影响不大，就光滑性而言，Gaussian和Epanechnikov核函数对应的光滑性较好，Triangle次之，Uniform最差，在应用中，一般用Gaussian核函数。


01
IDEA
02
LEARNING
03
CREATIVE
INSPIRATION
EDUCATION
BUSINESS
PHOTO BOOK

感谢聆听