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8日

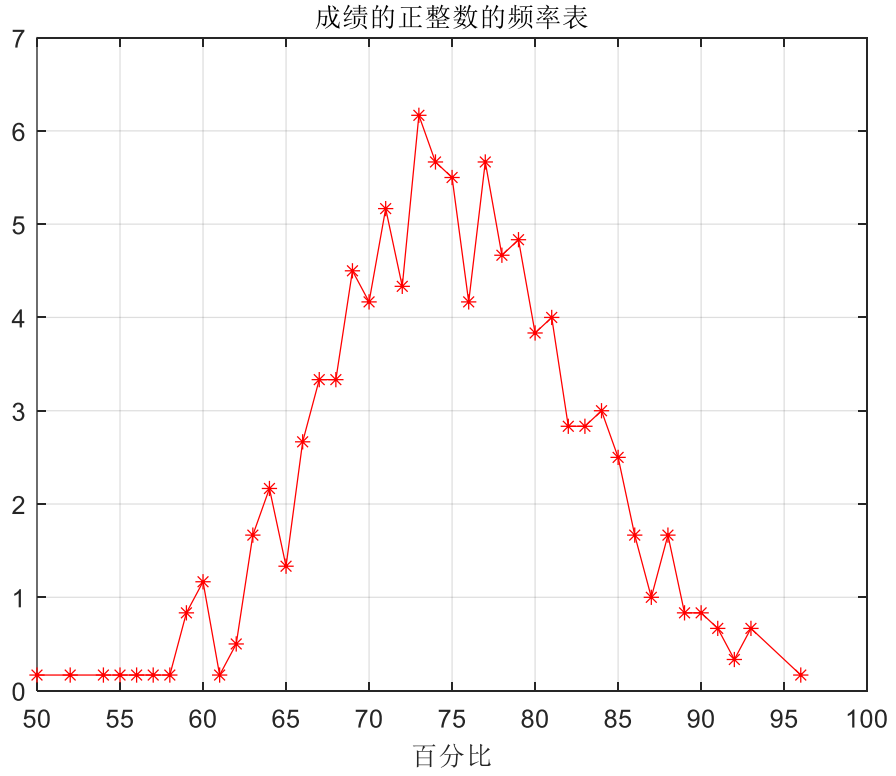
第10章 数据统计分析知识点思维导图

1. 正整数的频率表

- 正整数的频率表，函数 tabulate
  - 格式table = tabulate(X)　%X为正整数构成的向量，返回3列：第1列中包含X的值，第2列为这些值的个数，第3列为这些值的频率。

```
>> load examgrades
>> t = tabulate(grades(:))
%过滤占比为零的频率数据
>> K = find(t(:,3) > 0);
>> t = t(K,:); %取得非零正整数频率表
>> plot(t(:,1),t(:,3),'r-*')
>> grid on
>> title('成绩的正整数的频率表')
>> xlabel('成绩')
>> xlabel('百分比')
```

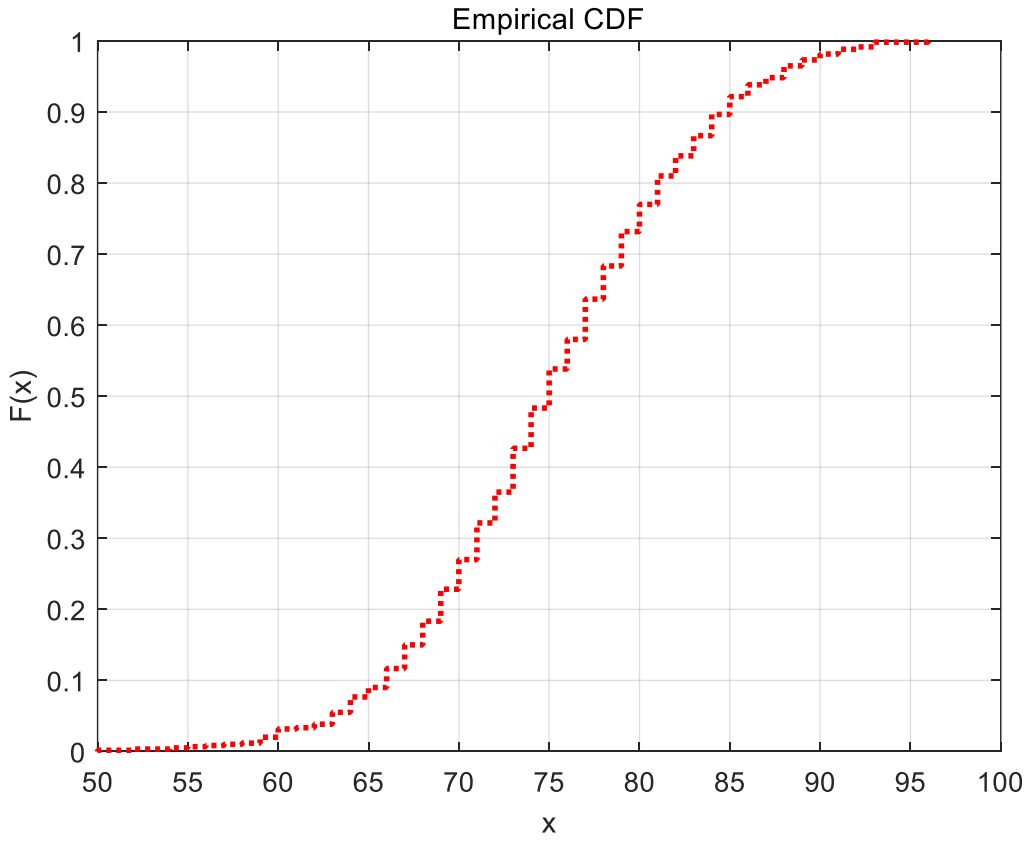
```
>> t = t(K,:)
t =
   50.0000    1.0000    0.1667
   52.0000    1.0000    0.1667
   54.0000    1.0000    0.1667
   55.0000    1.0000    0.1667
   56.0000    1.0000    0.1667
   57.0000    1.0000    0.1667
   58.0000    1.0000    0.1667
   59.0000    5.0000    0.8333
   60.0000    7.0000    1.1667
   61.0000    1.0000    0.1667
   62.0000    3.0000    0.5000
   63.0000   10.0000    1.6667
   64.0000   13.0000    2.1667
   65.0000    8.0000    1.3333
   66.0000   16.0000    2.6667
   67.0000   20.0000    3.3333
   68.0000   20.0000    3.3333
```



2. 经验累积分布函数图形

- 函数cdfplot
  - [h,stats] = cdfplot(X)：作样本X（向量）的累积分布函数图形，h表示曲线的环柄，stats表示样本的一些特征。

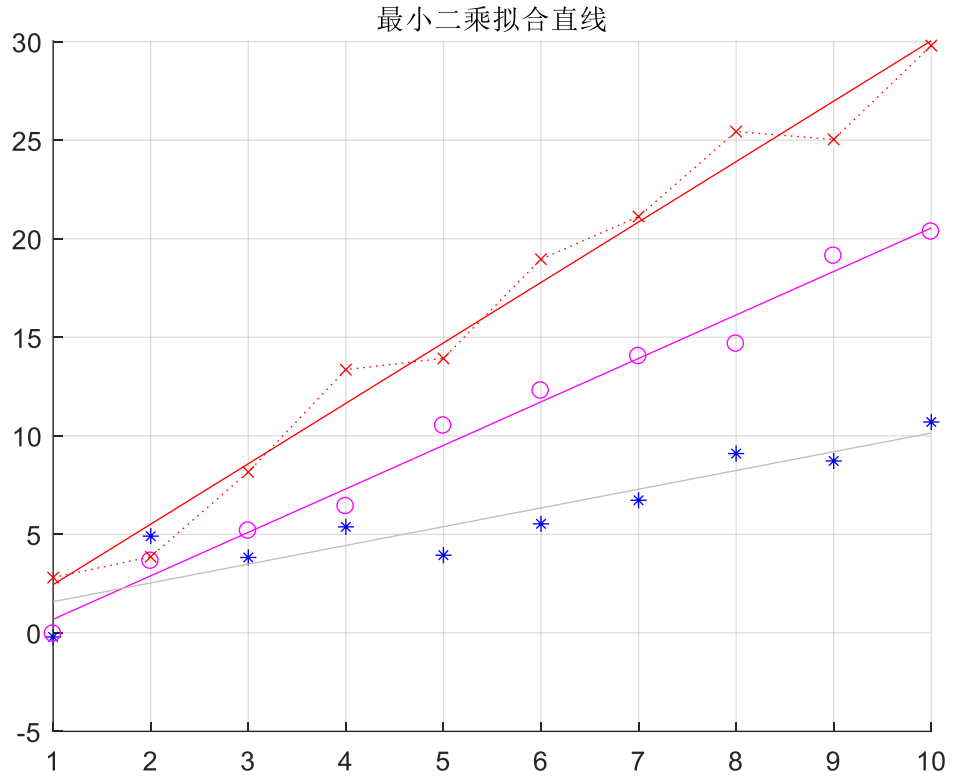
```
>> load examgrades
>> [h,stats]=cdfplot(grades(:))
stats =
  包含以下字段的 struct:
       min: 50
       max: 96
      mean: 75.0033
    median: 75
       std: 7.4008
>> set(h,{'linewidth','color','linestyle'},{2,'r',':'})
```



3. 最小二乘拟合直线

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

```
>> x = 1:10;
>> y1 = x + randn(1,10);
>> scatter(x,y1,25,'b','*')
>> hold on
>> y2 = 2*x + randn(1,10);
>> y3 = 3*x + randn(1,10);
>> plot(x,y2,'mo', x,y3,'rx:')
>> lsline
>> grid on
>> title('最小二乘拟合直线')
```



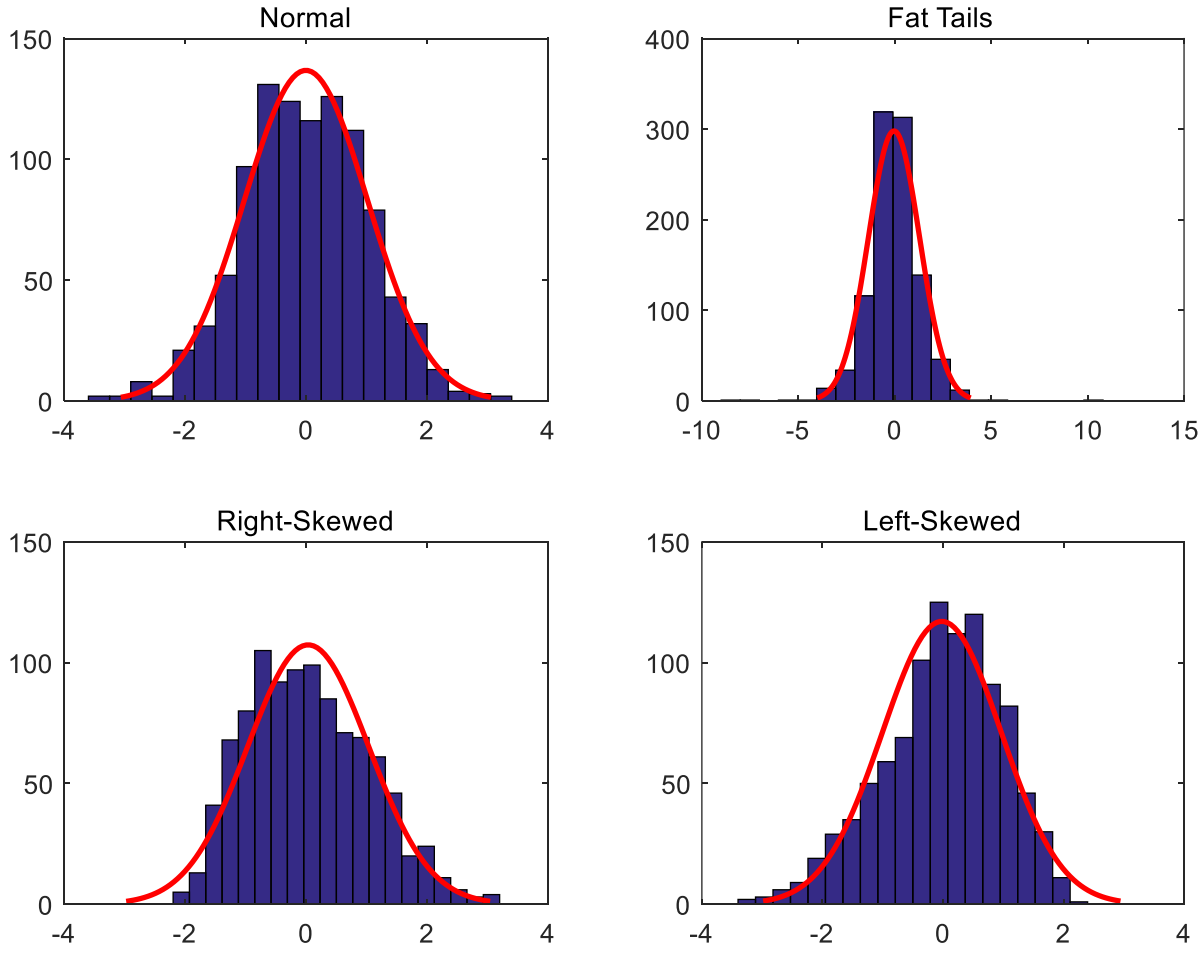
4. 正态分布概率图形

- 正态概率图用于正态分布的检验，正态分布的概率图描绘的是一条直线（参考线），每一个样本观测值对应上的一个'+'号，如果图中的'+'号都集中在参考线附近，说明样本观测数据近似服从正态分布；若果偏离参考线的'+'号越多，说明样本观测数据不服从正态分布。
- 函数normplot
  - 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：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。

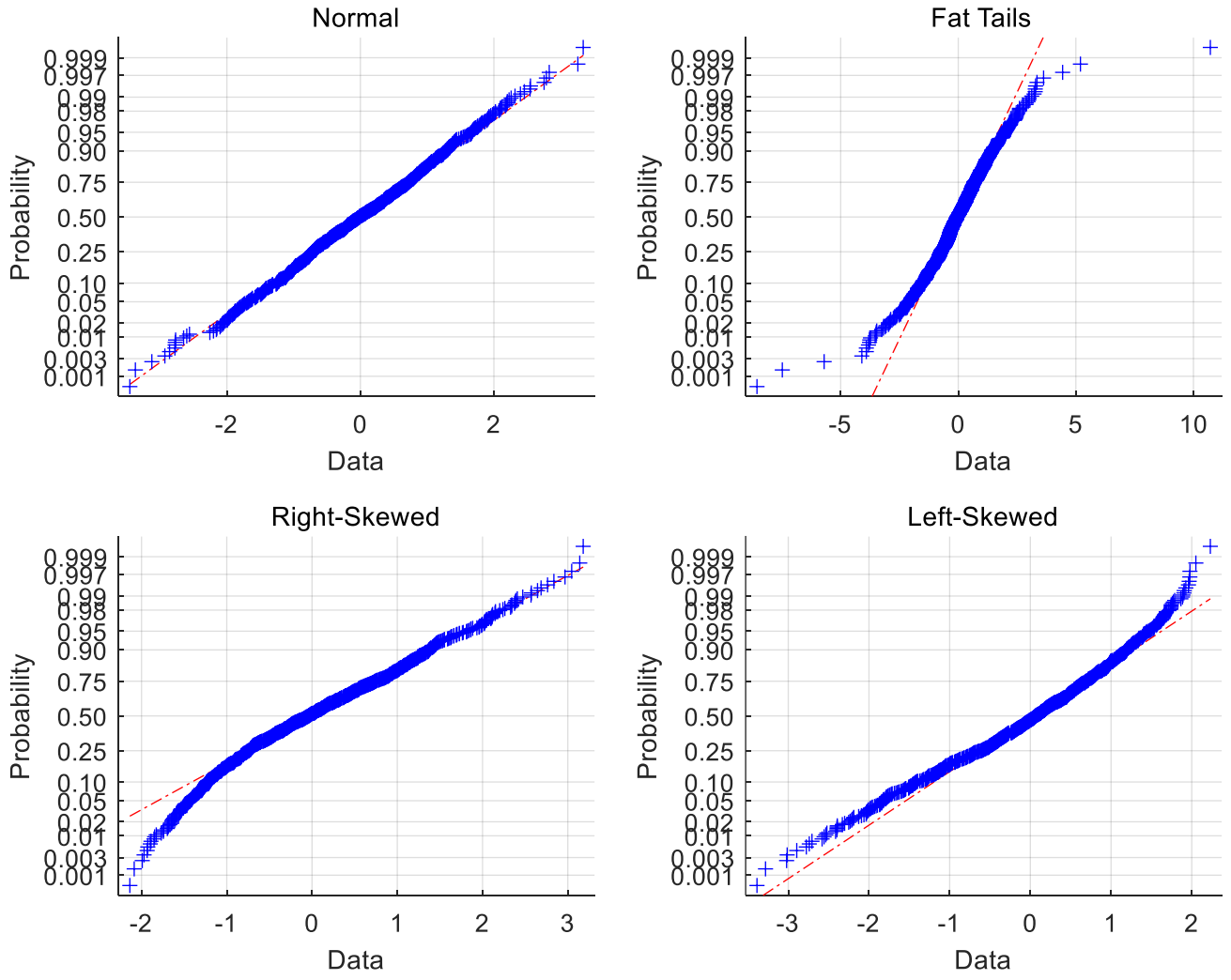
4. 正态分布概率图形

```
x1 = normrnd(0,1,[1000,1]); %标准正态分布
x2 = trnd(5,[1000,1]); % T分布的，厚尾
x3 = pearsrnd(0,1,0.5,3,[1000,1]); %产生右偏数据
x4 = pearsrnd(0,1,-0.5,3,[1000,1]); %产生左偏数据
subplot(2,2,1);histfit(x1,20);
title('Normal');
subplot(2,2,2);histfit(x2,20);
title('Fat Tails');
subplot(2,2,3);histfit(x3,20);
title('Right-Skewed');
subplot(2,2,4);histfit(x4,20);
title('Left-Skewed');
```



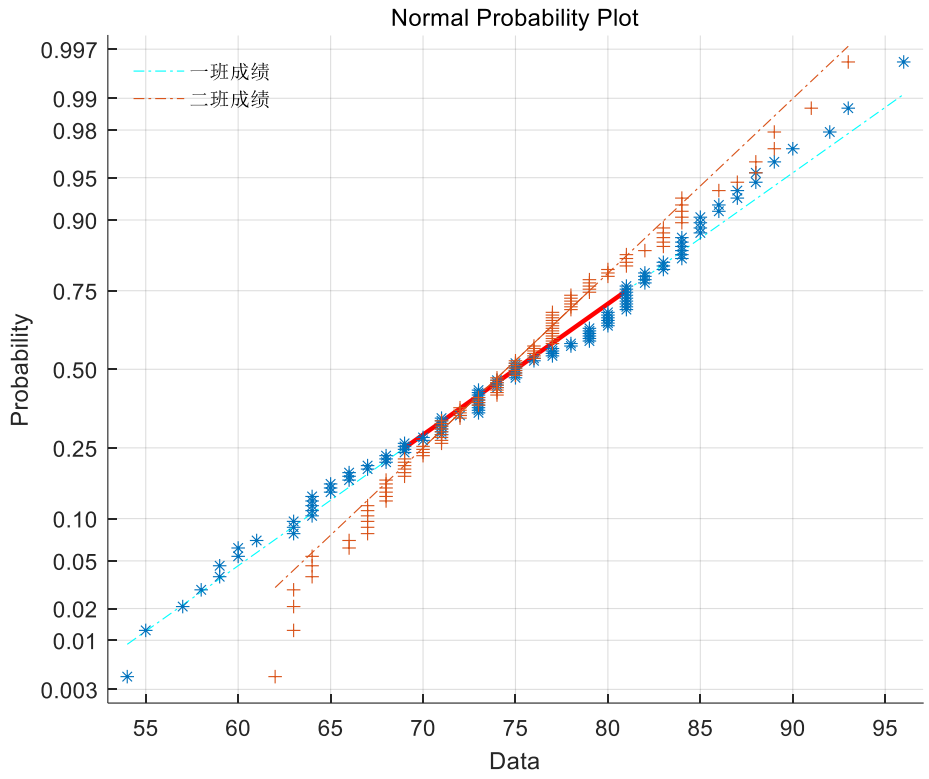
4. 正态分布概率图形

```
subplot(2,2,1); normplot(x1);
title('Normal')
subplot(2,2,2); normplot(x2);
title('Fat Tails')
subplot(2,2,3); normplot(x3);
title('Right-Skewed')
subplot(2,2,4); normplot(x4)
title('Left-Skewed')
```



4. 正态分布概率图形

```
>> load examgrades
>> h = normplot(grades(:,1:2))
>> h(1).Marker = '*';
>> h(3).Color = 'r';
>> h(3).LineWidth = 2;
>> h(5).Color = 'c';
>> legend({'一班成绩','二班成绩'},'Location','southeast')
>> legend('boxoff')
```



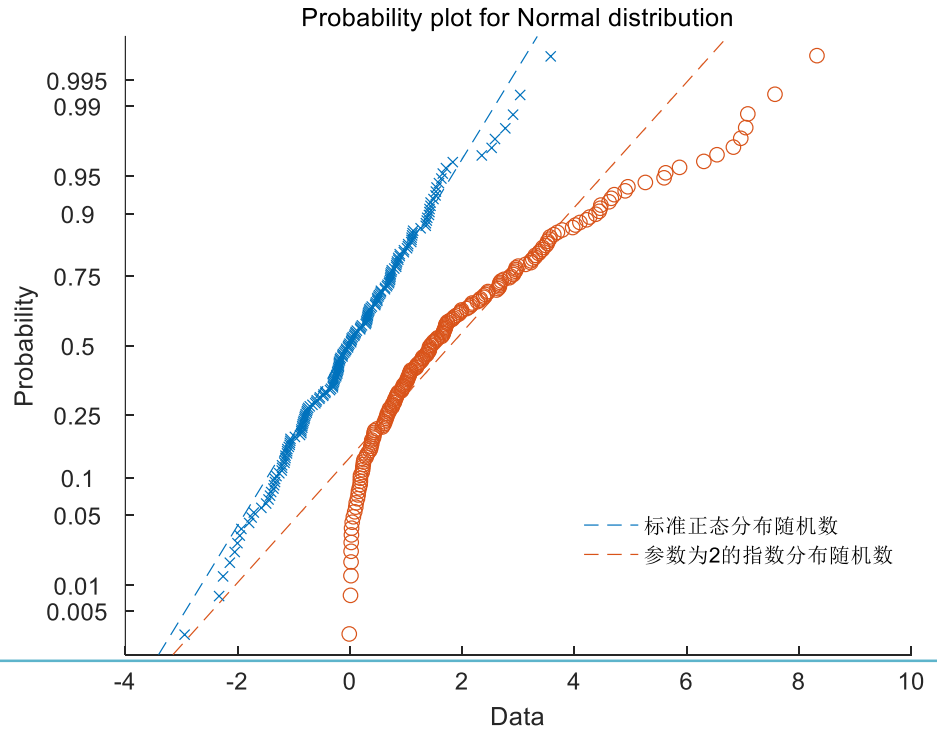
函数句柄h(1)和h(2)分别对应于正态分布和斜态分布的数据点。句柄h(3)和h(4)对应于适合样本数据的第二和第三四分位数线。句柄h(5)和h(6)对应于延伸到每组样本数据的最小值和最大值的外推线。

5. 绘制p-p图

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

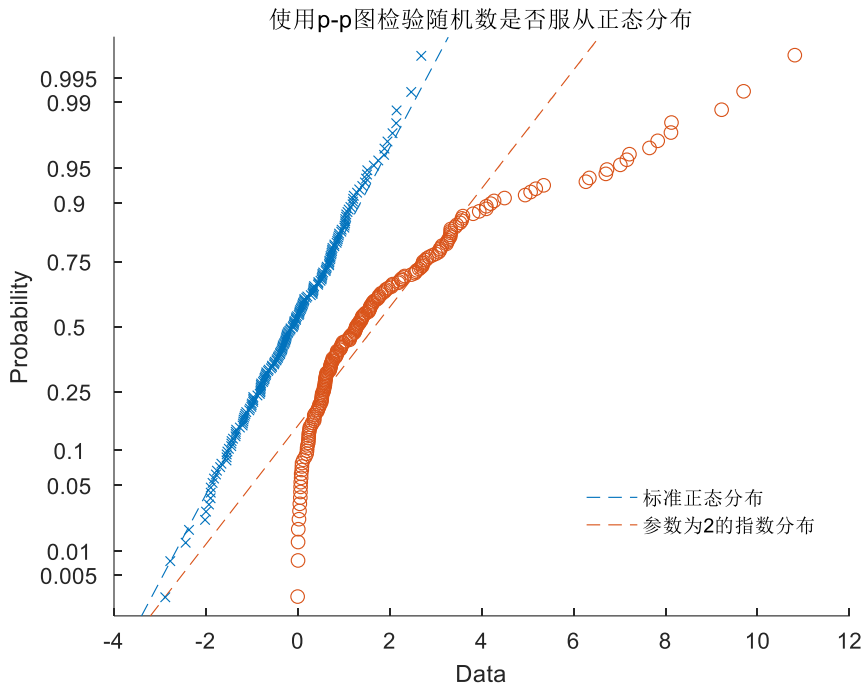
p-p图用来检测数据是否服从指定的分布。调用格式：

- probplot('name', x)：x是输入检验的数据，'name'指定检验哪种分布，name可以取
  - ✓ exponential：指数分布
  - ✓ extreme value或ev：极值分布
  - ✓ lognormal：对数分布
  - ✓ normal：正态分布
  - ✓ rayleigh：瑞利分布
  - ✓ weibull或wbl：韦伯分布

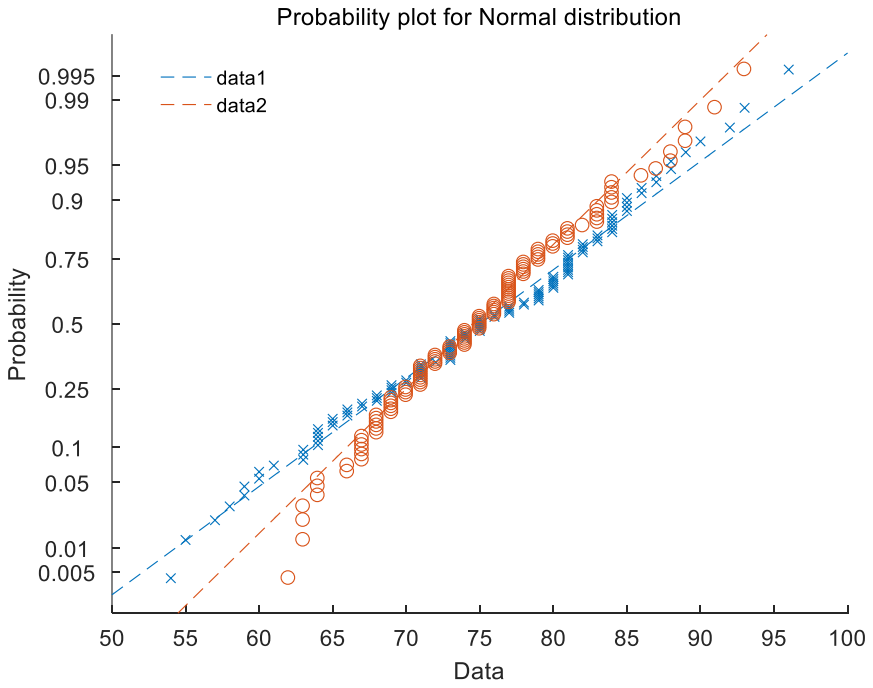


5. 绘制p-p图

```
>> R = normrnd(0,1,200,1);
>> E = exprnd(2,200,1);
>> probplot([R,E])
>> probplot('norm',[R,E])
>> legend('标准正态分布','参数为2的指数分布')
>> legend('boxoff')
>> title('使用p-p图检验随机数是否服从正态分布')
```



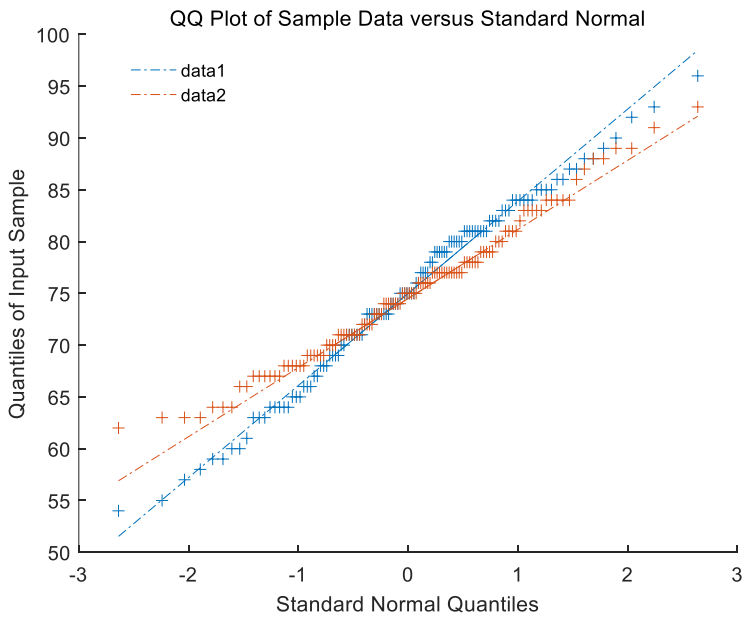

```
>> load examgrades
>> probplot(grades(:,1:2))
>> legend('data1','data2')
>> legend('boxoff')
```

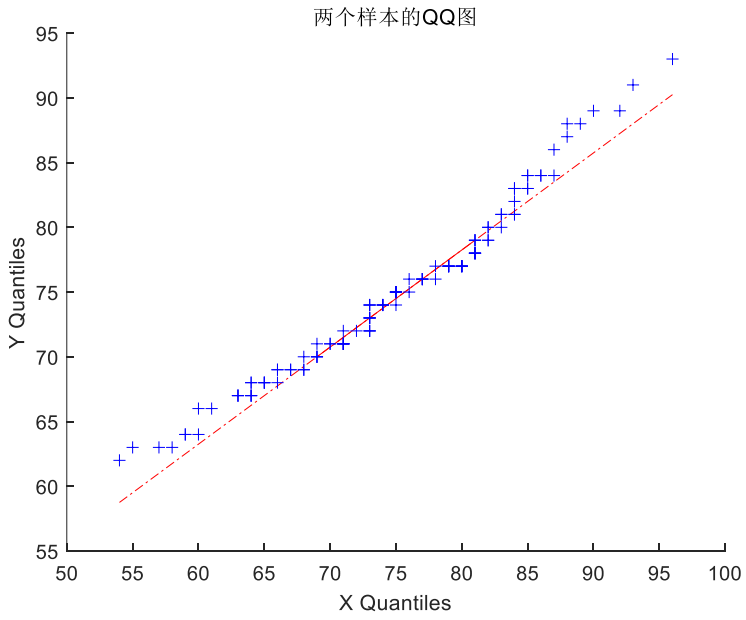


6. 绘制q-q图

- q-q图不仅能检验样本是否服从指定分布，还能检测两个样本是否服从相同的分布。调用格式：
  - qqplot(X)：makes an empirical QQ-plot of the quantiles of the data in the vector X versus the quantiles of a standard Normal distribution.
  - h = qqplot(X,Y,pvec)：可以画x, y他们各自分位数为横纵坐标的图，为QQ图专门一组指定分位数pvec

```
>> load examgrades
>> qqplot(grades(:,1:2))
>> legend('data1','data2')
>> legend('boxoff')
>> qqplot(grades(:,1),grades(:,2))
>> title('两个样本的QQ图')
```

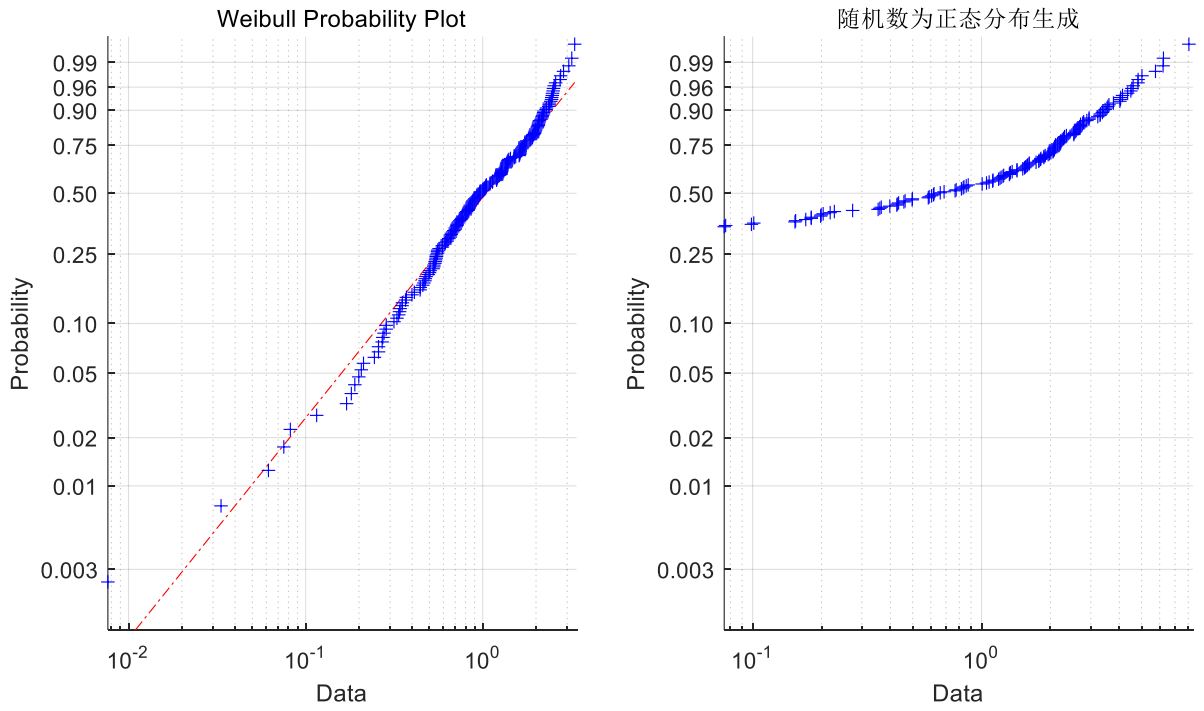




7. 绘制威布尔(Weibull)概率图形

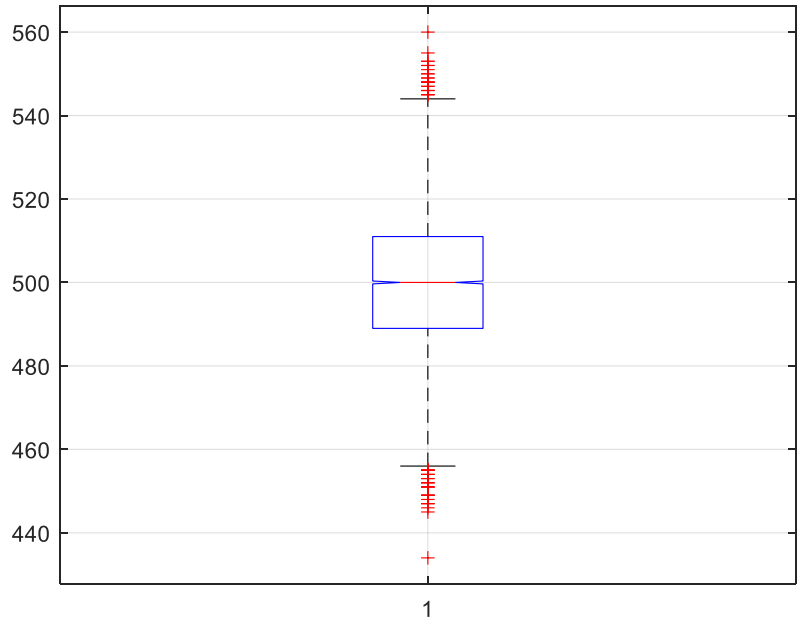
- 函数wblplot
  - h = wblplot(X)　%若X为向量，则显示威布尔(Weibull)概率图形，若X为矩阵，则显示每一列的威布尔概率图形。
  - 说明：绘制威布尔(Weibull)概率图形的目的是用图解法估计来自威布尔分布的数据X，如果X是威布尔分布数据，其图形是直线的，否则图形中可能产生弯曲。

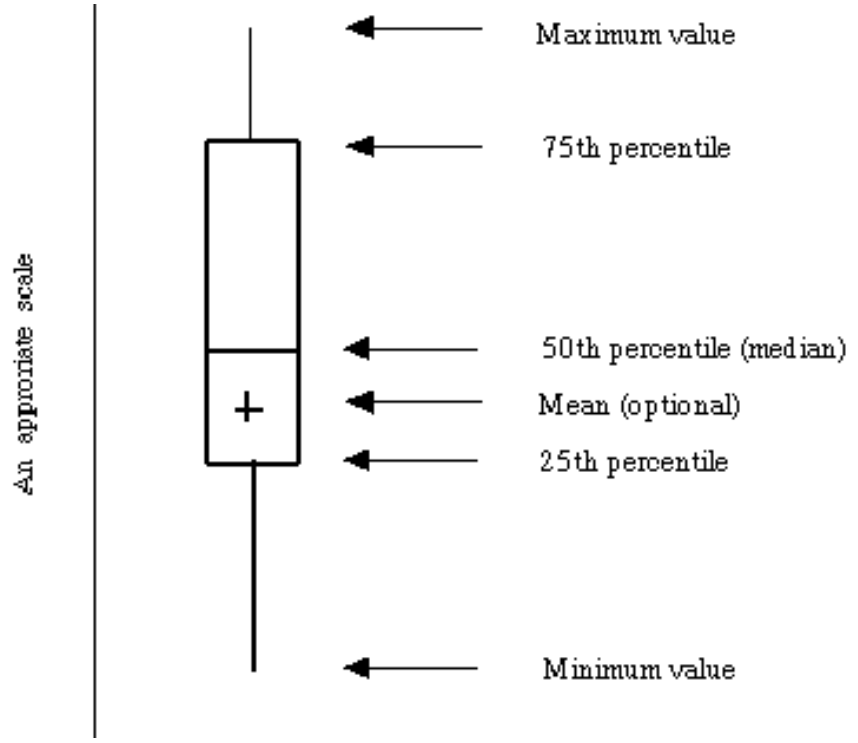
```
>> r = wblrnd(1.2,1.5,200,1);
>> subplot(1,2,1)
>> wblplot(r)
>> subplot(1,2,2)
>> R = normrnd(1,2,200,1);
>> wblplot(R)
>> title('随机数为正态分布生成')
```



8. 样本数据的盒图

盒图由五个数值点组成：最小值(min)，下四分位数(Q1)，中位数(median)，上四分位数(Q3)，最大值(max)。也可以往盒图里面加入平均值(mean)。下四分位数、中位数、上四分位数组成一个“带有隔间的盒子”。上四分位数到最大值之间建立一条延伸线，这个延伸线成为“胡须(whisker)”。





8. 样本数据的盒图

- 箱线图（盒图）需要用到统计学的四分位数（Quartile）的概念，所谓四分位数，就是把组中所有数据由小到大排列并分成四等份，处于三个分割点位置的数字就是四分位数。
  - 第一四分位数（Q1），又称“较小四分位数”或“下四分位数”，等于该样本中所有数值由小到大排列后第25%的数字。
  - 第二四分位数（Q2），又称“中位数”，等于该样本中所有数值由小到大排列后第50%的数字。
  - 第三四分位数（Q3），又称“较大四分位数”或“上四分位数”，等于该样本中所有数值由小到大排列后第75%的数字。
  - 第三四分位数与第一四分位数的差距又称四分位间距（InterQuartile Range，IQR）。

8. 样本数据的盒图

- 由于现实数据中总是存在各式各样“脏数据”，即“离群点”，于是为了不因这些少数的离群数据导致整体特征的偏移，将这些离群点单独汇出，而盒图中的胡须的两级修改成最小观测值与最大观测值。
- 经验：最大(最小)观测值设置为与四分位数值间距离为1.5个IQR(中间四分位数极差)。即
  - IQR = Q3-Q1，即上四分位数与下四分位数之间的差，也就是盒子的长度。
  - 最小观测值为min = Q1 - 1.5*IQR，如果存在离群点小于最小观测值，则胡须下限为最小观测值，离群点单独以点汇出。如果没有比最小观测值小的数，则胡须下限为最小值。
  - 最大观测值为max = Q3 +1.5*IQR，如果存在离群点大于最大观测值，则胡须上限为最大观测值，离群点单独以点汇出。如果没有比最大观测值大的数，则胡须上限为最大值。

8. 样本数据的盒图

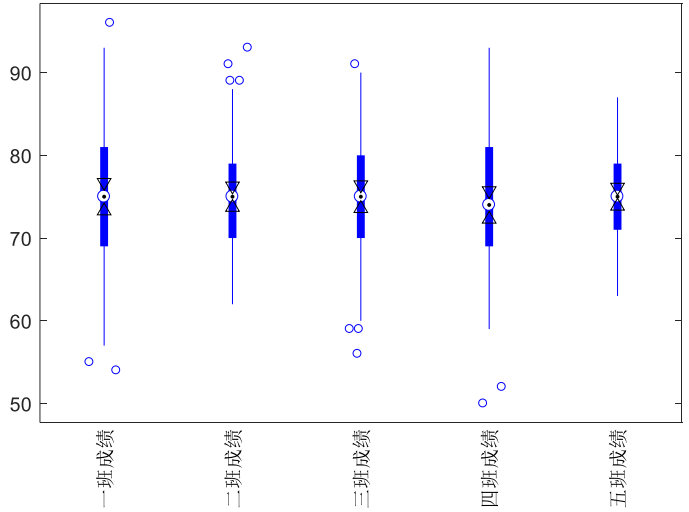
信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

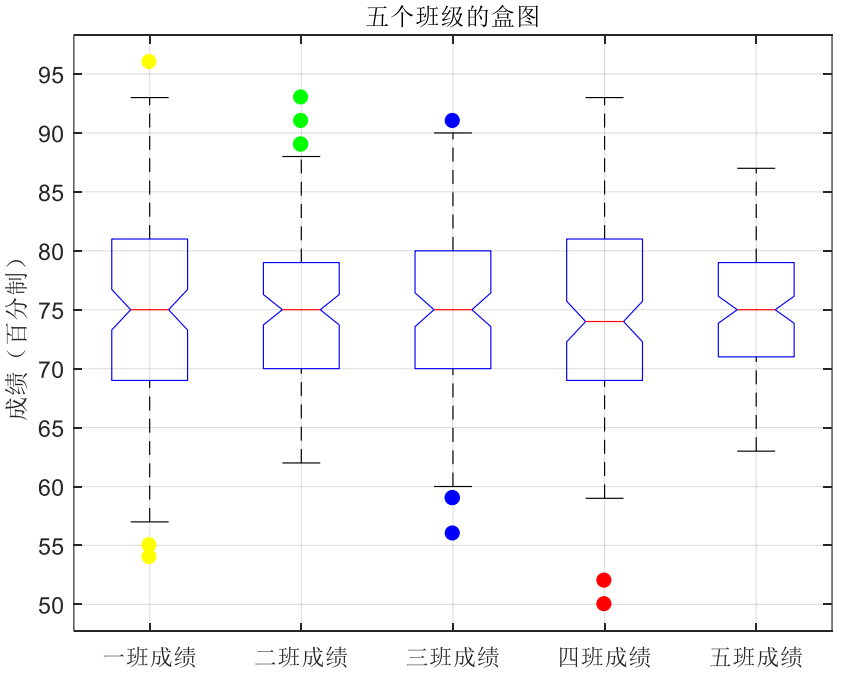
函数 boxplot

- 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',vert,whis) %sym表示图形符号，默认值为“+”，当vert=0时，生成水平盒图，vert=1时，生成竖直盒图（默认值vert=1），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。
- boxplot(___,Name,Value) 使用由一个或多个 Name,Value 对组参数指定的附加选项创建箱线图，如'PlotStyle'为'compact'，使用带实须线的窄实心箱绘制箱子。

8. 样本数据的盒图

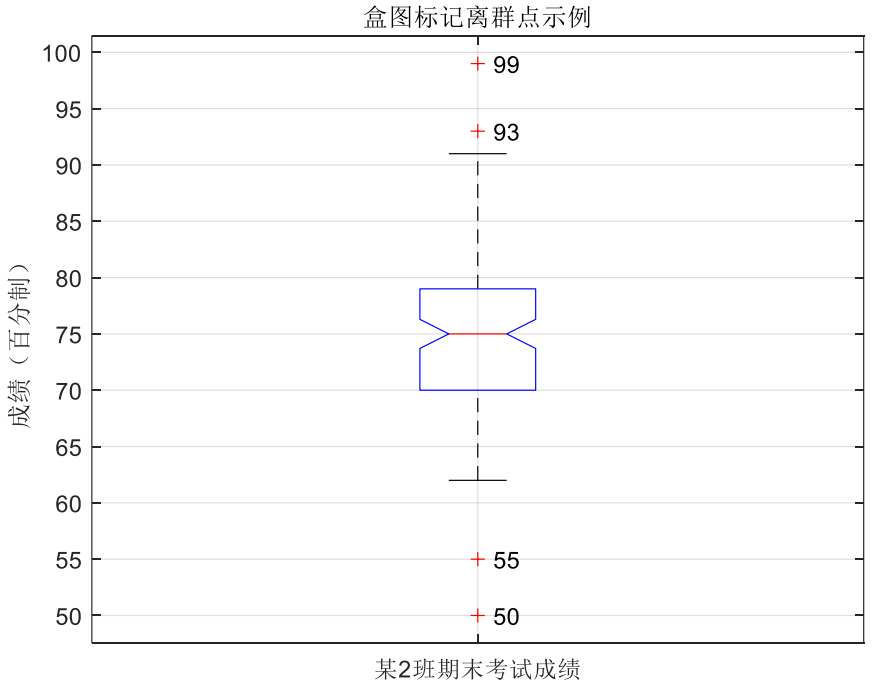
```
load examgrades
boxplot(grades,'Notch','on','Labels',{'一班成绩','二班成绩','三班成绩','四班成绩','五班成绩'},'Whisker',1)
h=findobj(gca,'tag','Outliers');
set(h,'Marker','o');
color = {'c','r','b','g','y'};
for i = 1:5
    set(h(i),{'MarkerEdgeColor','MarkerFaceColor'},{color{i},color{i}});
end
ylabel('成绩（百分制）')
grid on
title('五个班级的盒图')
```





8. 样本数据的盒图

```
s2 = grades(:,2);
s2(1) = 55;  s2(2) = 50; s2(3) = 99;  %设置几个异常值
pt = prctile(s2,[25,75]);    %箱形图上下界
p25 = pt(1);  p75 = pt(2); %获取Q1和Q3
upper = p75+ 1.5*(p75-p25); %最大观测值
lower = p25-1.5*(p75-p25); %最小观测值
upind = s2(s2 > upper); %超出最大值的离群点
lowind = s2(s2 < lower);  %小于最小值的离群点
ind =[upind;lowind];
index = sort(ind);
boxplot(s2,'Notch','on','Labels',{'某2班期末考试成绩'},...
    'Whisker',1.5,'outliersize',6)
hold on
rows = size(index,1);
for i =1:rows
    text(1+0.02,index(i,1),num2str(index(i,1)));
end
grid on;  title('盒图标记离群点示例')
ylabel('成绩（百分制）')
```



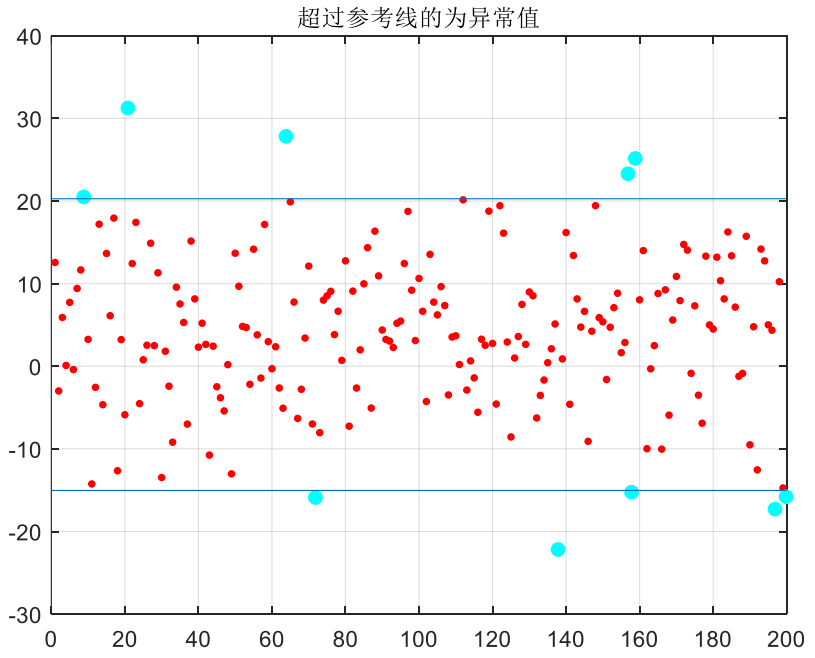
9. 给当前图形加一条参考线

信阳师范学院
数学与统计学院

当前图像添加参考系：函数refline

- 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：$y = b + ax$。

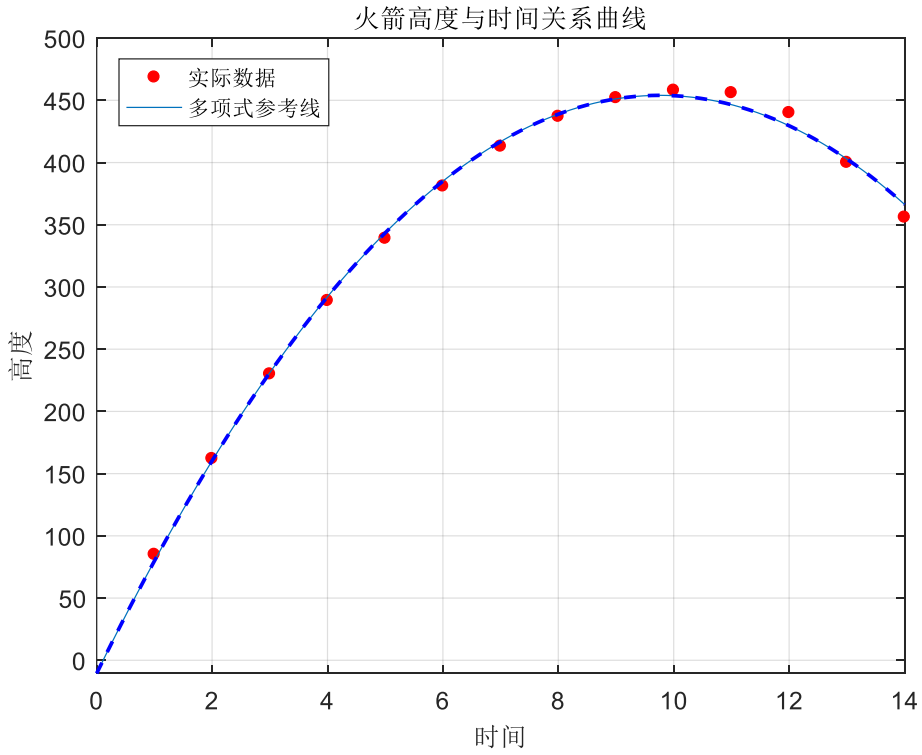
```
>> R = normrnd(5,9,200,1);
>> plot(R,'r.','MarkerSize',10)
>> grid on;  hold on
>> qt = quantile(R,[.025 .25 .50 .75 .975])
>> refline(0,qt(1))   %添加参考线
>> refline(0,qt(5))
>> indmin = find (R<qt(1));   %查找异常值
>> indmax = find (R>qt(5));   %查找异常值
>> plot(indmin,R(indmin),'co','MarkerFaceColor','c')
>> plot(indmax,R(indmax),'co','MarkerFaceColor','c')
```



10. 加入一条多项式曲线

- 函数refcurve
  - 格式h = 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=[p1,p2, p3,…,pn]，其中p1为最高幂项系数。
- 例：火箭的高度与时间图形，加入一条理论高度曲线。

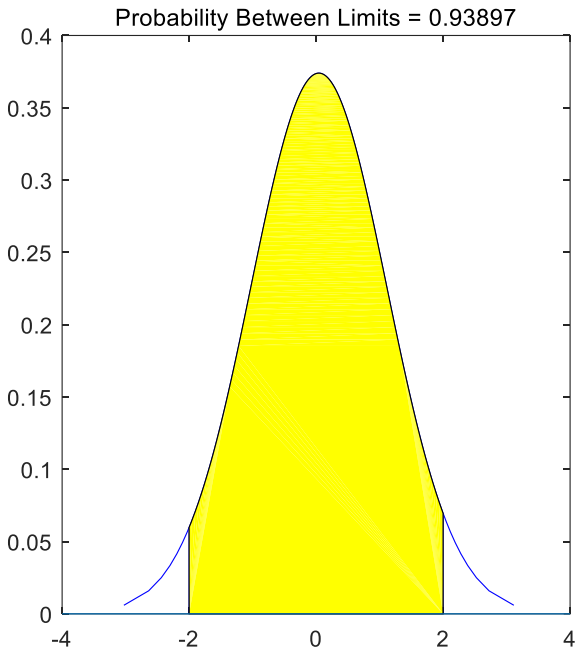
```
h = [85 162 230 289 339 381 413 437 452 458 456 440 400 356];
fh = @(b,x)b(1)*x.^2+b(2)*x + b(3);
[beta,r,J] = nlinfit(1:14,h,fh,[-1,1,1])
plot(h,'r.','MarkerSize',15);  grid on; hold on
fh = refcurve(beta);
set(fh,{'Color','LineWidth','LineStyle'},{'b',1.5,'--'})
legend('实际数据','多项式参考线')
```

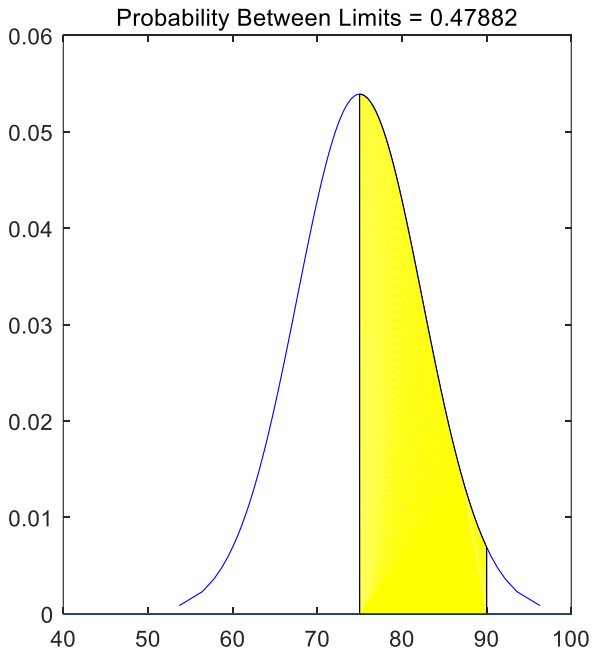


11. 样本的概率图形

- p=capaplot（data，specs）估计输入向量数据中观测值的均值和方差，并绘制结果T分布的pdf。数据中的观测值假定为正态分布。输出p是估计分布的新观测值在两个元素向量规格指定范围内的概率。specs中指定的上下限之间的分布部分在绘图中着色。

```
>> subplot(1,2,1)
>> data = normrnd (0,1,100,1);
>> p = capaplot(data,[-2,2])
p =
    0.9390
>> subplot(1,2,2)
>> p1 = capaplot(grades(:),[75,90])
p1 =
    0.4788
```

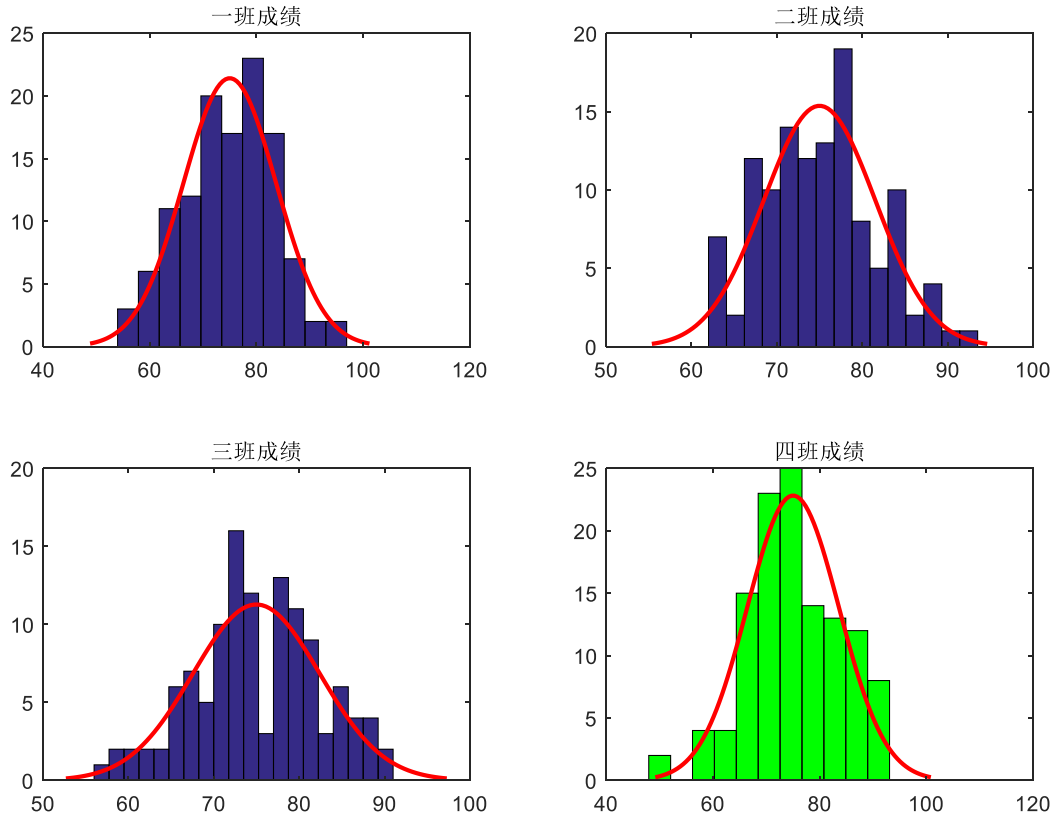




12. 附加正态密度曲线的直方图

- 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。

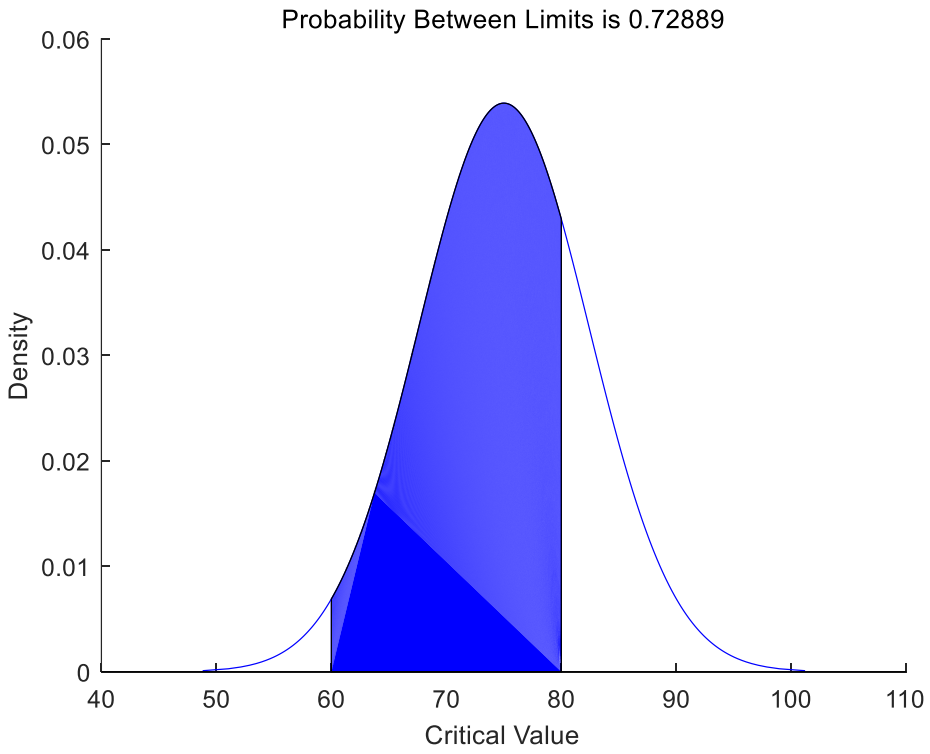
```
load examgrades
subplot(2,2,1); histfit(grades(:,1))
subplot(2,2,2); histfit(grades(:,2),15)
subplot(2,2,3); histfit(grades(:,3),20)
subplot(2,2,4); h =
histfit(grades(:,4));
h(2).Color = 'r';
h(1).FaceColor = 'g';
```



13. 区间内正态分布密度曲线

- 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。

```
>> load examgrades
>> mu = mean(grades(:));
>> sigma = std(grades(:));
>> p = normspec([60,80],mu,sigma)
p =
    0.7289
```




01
IDEA
02
LEARNING
03
CREATIVE
INSPIRATION
EDUCATION
BUSINESS
PHOTO BOOK

感谢聆听