第 9 讲　MATLAB 在数理统计中的应用

司守奎
烟台市，海军航空大学
Email：sishoukui@163.com

9.1　一些常用的统计量和统计图

9.1.1　统计量

假设有一个容量为 n 的样本（即一组数据），记作 $x=(x_1,x_2,\cdots,x_n)$，需要对它进行一定的加工，才能提出有用的信息，用作对总体（分布）参数的估计和检验。**统计量**就是加工出来的、反映样本数量特征的函数，它不含任何未知量。

下面我们介绍几种常用的统计量。

1.表示位置的统计量—算术平均值和中位数

算术平均值（简称均值）描述数据取值的平均位置，记作 $\overline{x}$，

$$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i\ . \tag{1}$$

中位数是将数据由小到大排序后位于中间位置的那个数值。

MATLAB 中 mean(x)返回 x 的均值，median(x)返回中位数。

2.表示变异程度的统计量—标准差、方差和极差

标准差 s 定义为

$$s=\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2\right]^{\frac{1}{2}}. \tag{2}$$

它是各个数据与均值偏离程度的度量，这种偏离不妨称为变异。

方差是标准差的平方 s^2。

极差是 $x=(x_1,x_2,\cdots,x_n)$ 的最大值与最小值之差。

MATLAB 中 std(x)返回 x 的标准差，var(x)返回方差，range(x)返回极差。

你可能注意到标准差 s 的定义(2)中，对 n 个 $(x_i-\overline{x})$ 的平方求和，却被 $(n-1)$ 除，这是出于无偏估计的要求。若需要改为被 n 除，MATLAB 可用 std(x,1)和 var(x,1)来实现。

标准差函数 std 的调用格式为

s=std(X,flag,dim) %flag=0,表示除以 n-1,flag=1,表示除以 n;dim 表示维数，dim 的默认值为 1，表示逐列求标准差，dim=2 表示逐行求标准差。

nanstd 表示忽略不确定数 NaN 剩下数据的标准差。类似的命令还有 nanmin，nanmax，nanmean，nanvar，nanmedian。

3.中心矩、表示分布形状的统计量—偏度和峰度

随机变量 x 的 r 阶**中心矩**为 $E(x-Ex)^r$。

随机变量 x 的偏度和峰度指的是 x 的标准化变量 $(x-Ex)/\sqrt{Dx}$ 的三阶中心矩和四阶中心矩：

$$\nu_1=E\left[\left(\frac{x-E(x)}{\sqrt{D(x)}}\right)^3\right]=\frac{E\left[(x-E(x))^3\right]}{(D(x))^{3/2}},$$

$$\nu_2=E\left[\left(\frac{x-E(x)}{\sqrt{D(x)}}\right)^4\right]=\frac{E\left[(x-E(x))^4\right]}{(D(x))^2}.$$

偏度反映分布的对称性，$\nu_1>0$ 称为右偏态，此时数据位于均值右边的比位于左边的多；$\nu_1<0$ 称为左偏态，情况相反；而 ν_1 接近 0 则可认为分布是对称的。

峰度是分布形状的另一种度量，正态分布的峰度为 3，若 ν_2 比 3 大得多，表示分布有沉重的尾巴，说明样本中含有较多远离均值的数据，因而峰度可以用作衡量偏离正态分布的尺

度之一。

MATLAB 中 moment(x,order)返回 x 的 order 阶中心矩，order 为中心矩的阶数。skewness(x)返回 x 的偏度，kurtosis(x)返回峰度。

在以上用 MATLAB 计算各个统计量的命令中，若 x 为矩阵，则作用于 x 的列，返回一个行向量。

4.协方差和相关系数

$\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$ 和 $\boldsymbol{y}=(y_1,y_2,\cdots,y_n)$ 的协方差

$$\mathrm{cov}(\boldsymbol{x},\boldsymbol{y})=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1},$$

其中 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$， $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$ 。

$\boldsymbol{x}$ 和 y 的相关系数

$$\rho_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}.$$

MATLAB中协方差和相关系数的函数使用格式如下：

cov(X)　%X为向量时，求X的方差；X为矩阵时，每一列作为一个变量的取值，求协方差矩阵。

corrcoef(X)　%计算矩阵 X 的列向量之间的相关系数矩阵。

MATLAB 中数据标准化的一种命令是

z=zscore(x)

它的变换公式为

$$z_i=\frac{x_i-\bar{x}}{s},$$

其中 $\bar{x}$ 和 s 分别为 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$ 的均值和标准差。

例 9.1　学校随机抽取 100 名学生，测量他们的身高和体重，所得数据如表 9.1 所示。试分别求身高的均值、中位数、标准差、方差、极差、二阶中心矩、三阶中心矩、偏度和峰度；计算身高与体重的协方差、相关系数；计算数据标准化以后，身高和体重的协方差矩阵。

表 9.1　100 名学生身高和体重数据

身高	体重	身高	体重	身高	体重	身高	体重	身高	体重
172	75	169	55	169	64	171	65	167	47
171	62	168	67	165	52	169	62	168	65
166	62	168	65	164	59	170	58	165	64
160	55	175	67	173	74	172	64	168	57
155	57	176	64	172	69	169	58	176	57
173	58	168	50	169	52	167	72	170	57
166	55	161	49	173	57	175	76	158	51
170	63	169	63	173	61	164	59	165	62
167	53	171	61	166	70	166	63	172	53
173	60	178	64	163	57	169	54	169	66
178	60	177	66	170	56	167	54	169	58
173	73	170	58	160	65	179	62	172	50
163	47	173	67	165	58	176	63	162	52
165	66	172	59	177	66	182	69	175	75
170	60	170	62	169	63	186	77	174	66
163	50	172	59	176	60	166	76	167	63

172	57	177	58	177	67	169	72	166	50
182	63	176	68	172	56	173	59	174	64
171	59	175	68	165	56	169	65	168	62
177	64	184	70	166	49	171	71	170	59

首先把表 9.1 的数据保存在纯文本文件 data91.txt 中。

```
clc, clear
a=load('data91.txt');
h=a(:,[1:2:end]); h=h(:); %提取身高数据并转换为列向量
w=a(:,[2:2:end]); w=w(:); %提取体重数据并转换为列向量
m=mean(h), me=median(h), s=std(h), v1=var(h), ra=range(h)
mm=moment(h,2),v2=var(h,1) %二阶中心矩和方差（除以样本容量）的比较
mmm=moment(h,3), sk=skewness(h), k=kurtosis(h)
c1=dot(h-mean(h),w-mean(h))/(length(h)-1)  %计算协方差
c2=cov(h,w)  %计算协方差阵
rr1=dot(h-mean(h),w-mean(w))/norm(h-mean(h))/norm(w-mean(w)) %计算相关系数
rr2=corrcoef(h,w)  %计算相关系数阵
bh=zscore(h);bw=zscore(w);  %数据标准化
rr3=cov(bh,bw)  %标准化数据的协方差就是相关系数矩阵
```

9.1.2 统计图

1.频数表及直方图

计算数据频数并且画直方图的命令为

h=histogram(X,nbins)

它将区间[min(X),max(Y)]等分为 nbins 份，统计在每个左闭右开小区间（最后一个小区间为闭区间）上数据出现的频数并画直方图。

h=histogram(X,edges)

它将根据 edges 作为区间端点，统计在每个左闭右开小区间（最后一个小区间为闭区间）上数据出现的频数并画直方图。

例 9.2（续例 9. 1） 画出身高和体重的直方图，并统计从最小体重到最大体重，步长间隔为 5 的小区间上，数据出现的频数。

```
clc, clear
a=load('data91.txt');
h=a(:,[1:2:end]); h=h(:); %提取身高数据并转换为列向量
w=a(:,[2:2:end]); w=w(:); %提取体重数据并转换为列向量
subplot(121), histogram(h), title('身高的直方图') %新版直方图命令，2014A 没有该命令
subplot(122), histogram(w), title('体重的直方图')
minh=min(h), maxh=max(h)
edge=[minh:5:maxh],edge=unique([edge,maxh])
figure,hh=histogram(h,edge)  %生成 histogram 对象
N1=hh.Values %在各小区间数据出现的频数
L=length(edge); %小区间端点的个数
for i=1:L-2
    N2(i)=sum(h>=edge(i) & h<edge(i+1)); %编程统计数据在各小区间出现的频数
end
N2(L-1)=sum(h>=edge(L-1) & h<=edge(L))  %N1 和 N2 是相等的
```

直方图如图 9.1 所示。

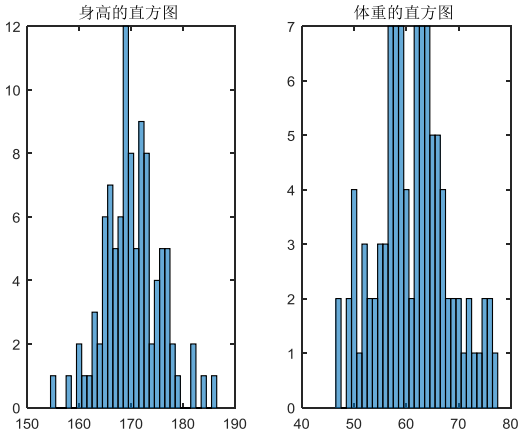

图 9.1　身高和体重的直方图

从直方图上可以看出，身高的分布大致呈中间高、两端低的钟形；而体重则看不出什么规律。要想从数值上给出更确切的描述，需要进一步研究反映数据特征的所谓“统计量”。直方图所展示的身高的分布形状可看作正态分布，当然也可以用这组数据对分布作假设检验。

2.箱线图

先介绍样本分位数。

定义 9.1　设有容量为 n 的样本观测值 $x_1, x_2, \cdots, x_n$，样本 p 分位数（$0 < p < 1$）记为 x_p，它具有以下的性质：（1）至少有 np 个观测值小于或等于 x_p；（2）至少有 $n(1-p)$ 个观测值大于或等于 x_p。

样本 p 分位数可按以下法则求得。将 $x_1, x_2, \cdots, x_n$ 按自小到大的次序排列成 $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$。

$$x_p = \begin{cases} x_{([np]+1)}, & \text{当} np \text{不是整数,} \\ \dfrac{1}{2}[x_{(np)} + x_{(np+1)}], & \text{当} np \text{是整数.} \end{cases}$$

特别，当 $p = 0.5$ 时，0.5 分位数 $x_{0.5}$ 也记为 Q_2 或 M，称为样本中位数，即有

$$x_{0.5} = \begin{cases} x_{([\frac{n}{2}]+1)}, & \text{当} n \text{是奇数,} \\ \dfrac{1}{2}[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}], & \text{当} n \text{是偶数.} \end{cases}$$

当 n 是奇数时中位数 $x_{0.5}$ 就是 $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ 这一数组最中间的一个数；而当 n 是偶数时中位数 $x_{0.5}$ 就是 $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ 这一数组中最中间两个数的平均值。

0.25 分位数 $x_{0.25}$ 称为第一四分位数，又记为 Q_1；0.75 分位数 $x_{0.75}$ 称为第三四分位数，又记为 Q_3。$x_{0.25}$，$x_{0.5}$，$x_{0.75}$ 在统计中是很有用的。

例 9.3　设有一组容量为 18 的样本值如下（已经过排序）

122　126　133　140　145　145　149　150　157
162　166　175　177　177　183　188　199　212

求样本分位数：$x_{0.25}$，$x_{0.5}$。

解　（1）因为 $np = 18 \times 0.25 = 4.5$，$x_{0.25}$ 位于第 $[4.5]+1=5$ 处，即有 $x_{0.25} = 145$。

（2）因为 $np=18\times0.5=9$， $x_{0.5}$ 是这组数中间两个数的平均值，即有

$$x_{0.5}=\frac{1}{2}(157+162)=159.5.$$

计算的 MATLAB 程序如下

```
clc, clear
x0=[122  126  133  140  145  145  149  150  157
162  166  175  177  177  183  188  199  212];
x0=x0'; x0=x0(:);
y=quantile(x0,[0.25 0.5])  %求两个样本分位数
yy=median(x0) %再求样本中位数
```

下面介绍箱线图。

数据集的箱线图是由箱子和直线组成的图形，它是基于以下 5 个数的图形概括：最小值 Min，第一四分位数 Q_1，中位数 M，第三四分位数 Q_3 和最大值 Max。它的作法如下：

（1）画一水平数轴，在轴上标上 Min，Q_1，M，Q_3，Max。在数轴上方画一个上、下侧平行于数轴的矩形箱子，箱子的左右两侧分别位于 Q_1，Q_3 的上方，在 M 点的上方画一条垂直线段，线段位于箱子内部。

（2）自箱子左侧引一条水平线直至最小值 Min；在同一水平高度自箱子右侧引一条水平线直至最大值 Max。这样就将箱线图做好了，如图 9.2 所示。箱线图也可以沿垂直数轴来作。自箱线图可以形象地看出数据集的以下重要性质。

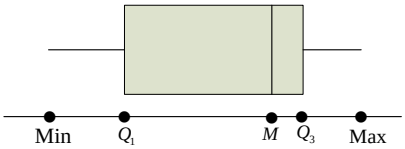

图 9.2　箱线图示意图

i）中心位置：中位数所在的位置就是数据集的中心。

ii）散步程度：全部数据都落在 $[\mathrm{Min},\mathrm{Max}]$ 之内，在区间 $[\mathrm{Min},Q_1]$，$[Q_1,M]$，$[M,Q_3]$，$[Q_3,\mathrm{Max}]$ 的数据个数各占 1/4。区间较短时，表示落在该区间的点较集中，反之较为分散。

iii）关于对称性：若中位数位于箱子的中间位置。则数据分布较为对称。又若 Min 离 M 的距离较 Max 离 M 的距离大，则表示数据分布向左倾斜，反之表示数据向右倾斜，且能看出分布尾部的长短。

例 9.4　下面分别给出了 25 个男子和 25 个女子的肺活量（以升计，数据已经过排序）。

女子组 2.7　2.8　2.9　3.1　3.1　3.1　3.2　3.4　3.4　3.4　3.4　3.4　3.5　3.5　3.5　3.6　3.7　3.7　3.7　3.8　3.8　4.0　4.1　4.2　4.2

男子组 4.1　4.1　4.3　4.3　4.5　4.6　4.7　4.8　4.8　5.1　5.3　5.3　5.3　5.4　5.4　5.5　5.6　5.7　5.8　5.8　6.0　6.1　6.3　6.7　6.7

画图的 MATLAB 程序如下：

```
clc, clear
a=[2.7  2.8  2.9  3.1  3.1  3.1  3.2  3.4  3.4  3.4  3.4  3.4  3.5  3.5  3.5  3.6
3.7  3.7  3.7  3.8  3.8  4.0  4.1  4.2  4.2]';
b=[4.1  4.1  4.3  4.3  4.5  4.6  4.7  4.8  4.8  5.1  5.3  5.3  5.3  5.4  5.4  5.5
5.6  5.7  5.8  5.8  6.0  6.1  6.3  6.7  6.7]';
name={'女子','男子'};
boxplot([a,b],name)
```

画出的箱线图见图 9.3。

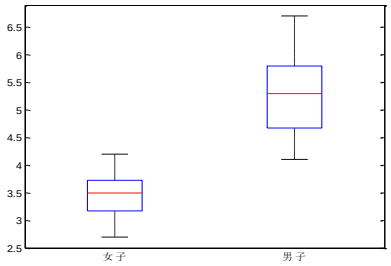

图 9.3　箱线图

箱线图特别适用于比较两个或两个以上数据集的性质，为此，我们将几个数据集的箱线图画在同一个图形界面上。例如在例 9.4 中可以明显地看到男子的肺活量要比女子大，男子的肺活量较女子的肺活量分散。

在数据集中某一个观察值不寻常地大于或小于该数集中的其它数据，称为疑似异常值。疑似异常值的存在，会对随后的计算结果产生不适当的影响。检查疑似异常值并加以适当的处理是十分重要的。

第一四分位数 Q_1 与第三四分位数 Q_3 之间的距离：Q_3-Q_1 记为 IQR，称为四分位数间距。若数据小于 $Q_1-1.5IQR$ 或大于 $Q_3+1.5IQR$，就认为它是疑似异常值。

3.经验分布函数

设 $X_1,X_2,\cdots,X_n$ 是总体 F 的一个样本，用 $S(x)$，$-\infty<x<\infty$ 表示 $X_1,X_2,\cdots,X_n$ 中不大于 x 的随机变量的个数。定义经验分布函数 $F_n(x)$ 为

$$F_n(x)=\frac{1}{n}S(x),\quad -\infty<x<\infty.$$

对于一个样本值，那么经验分布函数 $F_n(x)$ 的观察值是很容易得到的（$F_n(x)$ 的观察值仍以 $F_n(x)$ 表示）。

一般，设 $x_1,x_2,\cdots,x_n$ 是总体 F 的一个容量为 n 的样本值。先将 $x_1,x_2,\cdots,x_n$ 按自小到大的次序排列，并重新编号。设为

$$x_{(1)}\le x_{(2)}\le\cdots\le x_{(n)}.$$

则经验分布函数 $F_n(x)$ 的观察值为

$$F_n(x)=\begin{cases}0, & 若x<x_{(1)},\\ \dfrac{k}{n}, & 若x_{(k)}\le x<x_{(k+1)},k=1,2,\cdots,n-1,\\ 1, & 若x\ge x_{(n)}.\end{cases}$$

对于经验分布函数 $F_n(x)$，格里汶科（Glivenko）在 1933 年证明了，当 $n\to\infty$ 时 $F_n(x)$ 以概率 1 一致收敛于分布函数 $F(x)$。因此，对于任一实数 x，当 n 充分大时，经验分布函数的任一个观察值 $F_n(x)$ 与总体分布函数 $F(x)$ 只有微小的差别，从而在实际上可当作 $F(x)$ 来使用。

例 9.5（续例 9. 1）根据表 9. 1 的数据，计算身高的经验分布函数并画出经验分布函数的图形。

首先计算 $F_n(h_i)$ 在每个每个互异点 h_i（总共 25 个点）的值，计算结果列在表 9.2 中。画出经验分布函数 $F_n(h)$ 的图形，如图 9.4。

表 9.2　身高数据经验分布

h_i	155	158	160	161	162	163	164	165	166

$F_n(h_i)$	0.01	0.02	0.04	0.05	0.06	0.09	0.11	0.17	0.24
h_i	167	168	169	170	171	172	173	174	175
$F_n(h_i)$	0.29	0.35	0.47	0.55	0.6	0.69	0.77	0.79	0.83
h_i	176	177	178	179	182	184	186		
$F_n(h_i)$	0.88	0.93	0.95	0.96	0.98	0.99	1		

计算及画图的 MATLAB 程序如下：

```
clc, clear
a=load('data91.txt');
h=a(:,[1:2:end]); h=h(:); %提取身高数据并转换为列向量
[fh,hh,n]=cdfcalc(h) %计算经验分布函数的取值，注意函数的取值 fh 比自变量的取值 hh 多了一个，fh 的第一个 0 是无用的
xlswrite('data95.xls',[hh,fh(2:end)]') %为了做表方便，把数据写到 Excel 文件中
cdfplot(h) %画经验分布函数的图形
xlabel('h'), ylabel('F(h)'), title('')
```

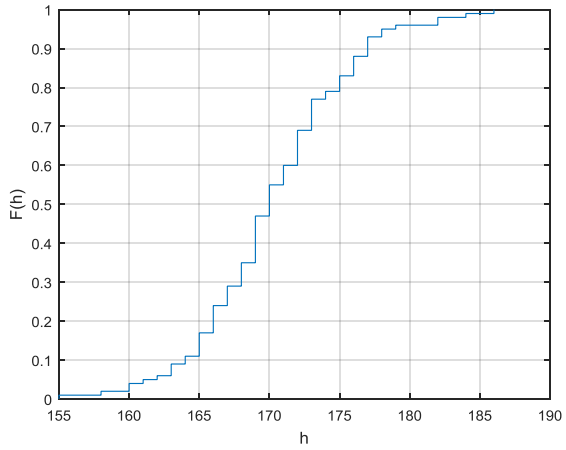

图 9.4　身高数据经验分布图

4.Q-Q 图

Q-Q 图是 Quantile-quantile Plot 的简称，是检验拟合优度的好方法，目前在国外被广泛使用，它的图示方法简单直观，易于使用。

对于一组观察数据 $x_1, x_2, \cdots, x_n$，利用参数估计方法确定了分布模型的参数 θ 后，分布函数 $F(x;\theta)$ 就知道了，现在我们希望知道观测数据与分布模型的拟合效果如何。如果拟合效果好，观测数据的经验分布就应当非常接近分布模型的理论分布，而经验分布函数的分位数自然也应当与分布模型的理论分位数近似相等。Q-Q 图的基本思想就是基于这个观点，将经验分布函数的分位数点和分布模型的理论分位数点作为一对数组画在直角坐标图上，就是一个点，n 个观测数据对应 n 个点，如果这 n 个点看起来像一条直线，说明观测数据与分布模型的拟合效果很好，以下我们给出计算步骤。

判断观测数据 $x_1, x_2, \cdots, x_n$ 是否来自于分布 $F(x)$，Q-Q 图的计算步骤如下：

（1）将 $x_1, x_2, \cdots, x_n$ 依大小顺序排列成：$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$；

（2）取 $y_i = F^{-1}((i-1/2)/n)$，$i = 1,2,\cdots,n$；

（3）将 $(y_i, x_{(i)})$，$i = 1,2,\cdots,n$，这 n 个点画在直角坐标图上；

（4）如果这 n 个点看起来呈一条 45° 角的直线，从 $(0,0)$ 到 $(1,1)$ 分布，我们就相信 $x_1, x_2, \cdots, x_n$ 拟合分布 $F(x)$ 的效果很好。

例 9.6（续例 9.1） 表 9.1 中的身高数据，如果它们来自于正态分布，求该正态分布的参数，试画出它们的 Q-Q 图，判断拟合效果。

解 （1）采用矩估计方法估计参数的取值。先从所给的数据算出样本均值和标准差

$$\bar{x}=143.7738,\quad s=5.9705,$$

正态分布 $N(\mu,\sigma^2)$ 中参数的估计值为 $\hat{\mu}=143.7738$，$\hat{\sigma}=5.9705$。

（2）画 Q-Q 图

i）将观测数据记为 $x_1,x_2,\cdots,x_{100}$，并依从小到大顺序排列为

$$x_{(1)}\le x_{(2)}\le\cdots\le x_{(100)}.$$

ii）取 $y_i=F^{-1}((i-1/2)/n)$，$i=1,2,\cdots,100$，这里 $F^{-1}(x)$ 是参数 $\mu=143.7738$，$\sigma=5.9705$ 的正态分函数的反函数。

iii）将 $(y_i,x_{(i)})$（$i=1,2,\cdots,100$）这 100 个点画在直角坐标系上，如图 9.5。

iv）这些点看起来接近一条 45° 角的直线，说明拟合结果较好。

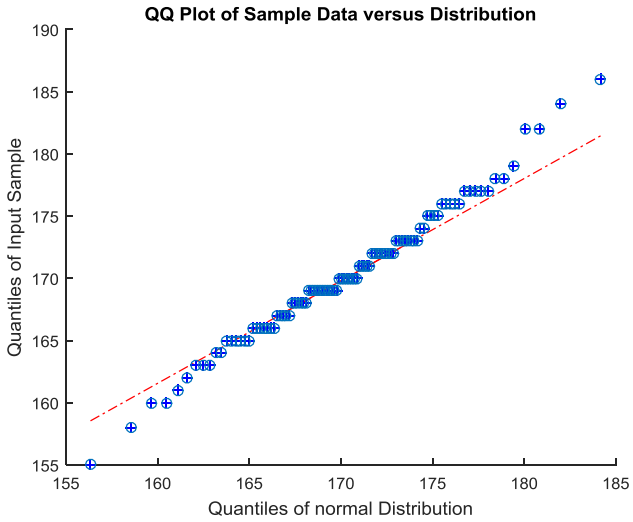

图 9.5 Q-Q 图

计算及画图的 MATLAB 程序如下

```
clc, clear, close all
a=textread('data91.txt');
h=a(:,[1:2:end]); h=h(:);
xbar=mean(h), s=std(h) %求均值和标准差
pd=makedist('normal','mu',xbar,'sigma',s) %定义正态分布
qqplot(h,pd)  %MATLAB 工具箱直接画 Q-Q 图
%下面不利用工具箱画 Q-Q 图
sa=sort(h); %把 a 按照从小到大排列
n=length(h); pi=([1:n]-1/2)/n;
yi=norminv(pi,xbar,s)' %计算对应的 yi 值
hold on, plot(yi,sa,'o') %重新描点画 Q-Q 图
```

9.2 统计中的概率分布

9.2.1 分布函数、密度函数和分位数

随机变量的特性完全由它的（概率）分布函数或（概率）密度函数来描述。设有随机变量 X，其分布函数定义为 $X\le x$ 的概率，即 $F(x)=P\{X\le x\}$。若 X 是连续型随机变量，则其密度函数 $p(x)$ 与 $F(x)$ 的关系为

$$F(x)=\int_{-\infty}^{x}p(x)dx.$$

定义 9.2 α 分位数

对于$0<\alpha<1$，使某分布函数$F(x)=\alpha$的x，称为这个分布的α分位数，记作x_α。

定义 9.3　上α分位数

若随机变量X的分布函数为$F(x)$，对于$0<\alpha<1$，若$\tilde{x}_\alpha$使得$P\{X>\tilde{x}_\alpha\}=\alpha$，则称$\tilde{x}_\alpha$为这个分布的上$\alpha$分位数。

9.2.2　MATLAB 统计工具箱中的概率分布

MATLAB 统计工具箱中有29种连续型概率分布，7种离散型概率分布，5种多元分布。在MATLAB命令窗口运行doc stats↵，打开超文本帮助后，再打开下一级目录Probability Distributions，就可以找到上述三种概率分布。

表9.3列举了MATLAB工具箱一些常用的概率分布名称。

表 9.3　MATLAB 工具箱常用概率分布命令字符

名称	二项分布	泊松分布	几何分布	离散均匀分布	连续均匀分布	指数分布
命令字符	bino	poiss	geo	unid	unif	exp
名称	正态分布	χ^2 分布	t 分布	F 分布	Γ 分布	多元正态分布
命令字符	norm	chi2	t	f	gam	mvn

MATLAB工具箱对每一种一维分布都提供5类函数，其命令字符见表9.4。

表 9.4　MATLAB 工具箱函数命令字符

函数	概率密度	分布函数	分布函数的反函数	均值与方差	随机数生成
命令字符	pdf	cdf	inv	stat	rnd

当需要一种分布的某一类函数时，将以上所列的分布命令字符与函数命令字符接起来，并输入自变量（可以是标量、数组或矩阵）和参数就行了，如：

```
p=tpdf(x,n)    %自由度为n的t分布的概率密度函数在x处的取值。
F=tcdf(x,n)    %自由度为n的t分布的分布函数在x处的取值。
x=tinv(p,n)    %自由度为n的t分布的p分位数为x，即分布函数在x处的值为p。
[M,V]=tstat(n)    %计算自由度为n的t分布的均值M，方差V。
R = trnd(V,[m,n])    %产生自由度为V的m×n的t分布随机数矩阵。
```

对部分分布还有参数估计的函数fit，例如

```
[mu,sigma,muci,sigmaci]=normfit(data,alpha)    %计算正态分布数据data的均值估计值mu，标准差估计值sigma，置信水平为100(1-alpha)%的均值和标准差的置信区间。
```

例 9.7　画出均值参数$\lambda=3$的泊松分布的概率分布图形和分布函数图形。

```
clc, clear, lambda=3;
x0=0:20;
subplot(121), plot(x0,poisspdf(x0,lambda),'*'), title('概率分布图形')
subplot(122), fplot(@(x)poisscdf(x,lambda),[0,20]),title('分布函数图')
```

例 9.8　设$X\sim N(0,1)$，标准正态分布的上α分位数记作z_α。试计算几个常用的z_α的值，并画出$z_{0.1}$的示意图。

计算得到几个常用的z_α的值见表9.5。

表 9.5　标准正态分布的上α分位数的值

α	0.001	0.005	0.01	0.025	0.05	0.10
z_α	3.0902	2.5758	2.3263	1.9600	1.6449	1.2816

$z_{0.1}$的示意图见图9.6。

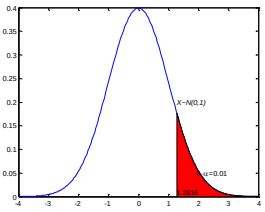

图 9.6　$z_{0.1}$ 的示意图

计算及画图的MATLAB程序如下：

```
clc, clear
alpha=[0.001 0.005 0.01 0.025 0.05 0.10];
za=norminv(1-alpha)
fplot(@(x)normpdf(x),[-4,4])   %画标准正态分布的概率密度曲线
x0=[za(end):0.01:4]; y0=normpdf(x0); %计算密度函数值
xx0=[x0,4,za(end)]; yy0=[y0,0,0]; %构造多边形顶点的x,y坐标
hold on, fill(xx0,yy0,'r') %多边形填充
text(1.9,0.05,'\leftarrow\alpha=0.01') %标注
text(za(end),0.01,num2str(za(end)))   %标注
text(1.2,0.2,'\it X~N(0,1)')   %标注
```

例 9.9　画对数正态分布的概率密度函数图形。

对数正态分布 Y 的概率密度函数为

$$g(y)=f(\ln y)\frac{1}{y}=\frac{1}{\sqrt{2\pi}\sigma y}e^{-\frac{(\ln y-\mu)^2}{2\sigma^2}},\quad y>0. \tag{3}$$

其中 $f(x)$ 为标准正态分布的密度函数。

（2）对数正态分布的分布函数为

$$G(y)=\Phi\left(\frac{\ln y-\mu}{\sigma}\right),\quad y>0. \tag{4}$$

这里 $\Phi(\cdot)$ 是标准正态分布（$\mu=0$，$\sigma=1$）的分布函数。

对数正态分布的概率密度函数图见图 9.7。

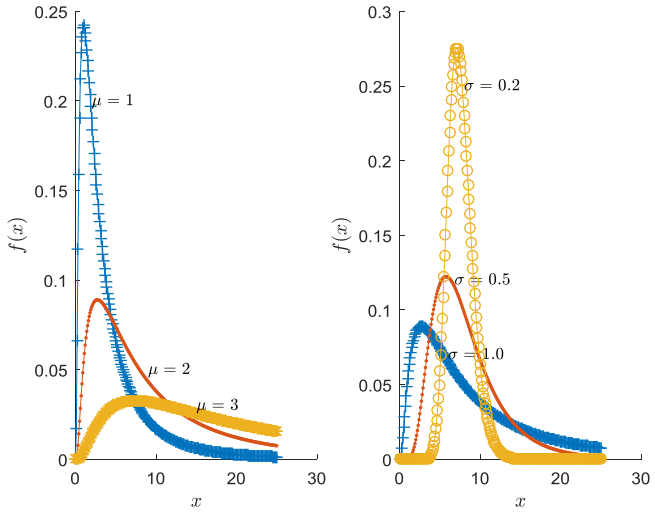

图 9.7　对数正态分布的概率密度函数图：左图为 $\sigma=1$，右图为 $\mu=2$

画图的 MATLAB 程序如下

```
clc, clear, close all
x0=0:0.1:25;
```

```
subplot(121), hold on
plot(x0,lognpdf(x0,1,1),'+-')
text(2,0.2,'$\mu=1$','Interpreter','Latex')
plot(x0,lognpdf(x0,2,1),'.-')
text(9,0.05,'$\mu=2$','Interpreter','Latex')
plot(x0,lognpdf(x0,3,1),'*-')
text(15,0.03,'$\mu=3$','Interpreter','Latex')
xlabel('$x$','Interpreter','Latex')
ylabel('$f(x)$','Interpreter','Latex')
subplot(122), hold on
plot(x0,lognpdf(x0,2,1.0),'+-')
text(6,0.07,'$\sigma=1.0$','Interpreter','Latex')
plot(x0,lognpdf(x0,2,0.5),'.-')
text(6.8,0.12,'$\sigma=0.5$','Interpreter','Latex')
plot(x0,lognpdf(x0,2,0.2),'o-')
text(8,0.25,'$\sigma=0.2$','Interpreter','Latex')
xlabel('$x$','Interpreter','Latex')
ylabel('$f(x)$','Interpreter','Latex')
```

定义 9.4　如果随机变量 X 的概率密度函数为

$$f(x)=\frac{x^{\alpha-1}}{\beta^{\alpha}}\frac{e^{-\frac{x}{\beta}}}{\Gamma(\alpha)},\quad x>0,\alpha>0,\quad \beta>0\,, \tag{5}$$

则称 X 服从参数为 (α,β) 的伽玛分布，记为 $X\sim \mathrm{Gamma}(\alpha,\beta)$，这时 α 称为形状参数，β 称为尺度参数。

注：$\Gamma(\alpha)=\int_0^{\infty}t^{\alpha-1}e^{-t}dt$，当 α 是正整数时，$\Gamma(\alpha)=(\alpha-1)!$。

伽玛函数的另一个重要而且常用的性质是下面的递推公式

$$\Gamma(\alpha+1)=\alpha\Gamma(\alpha),\quad \alpha>0\,.$$

例 9.10　画伽玛分布的概率密度函数图形。

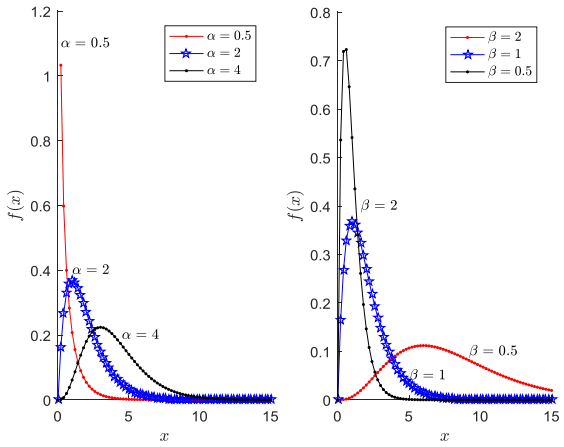

图 9.8　伽玛分布的概率密度函数图：左图为 $\beta=1$，右图为 $\alpha=4$

图 9.8 画出了伽玛分布的概率密度函数。

```
clc, clear, close all
x0=0:0.2:15;
%beta 参数数学上和 MATLAB 是倒数关系
subplot(121), hold on
plot(x0,gampdf(x0,0.5,1),'-r.')
text(0.2,1.1,'$\alpha=0.5$','Interpreter','Latex')
plot(x0,gampdf(x0,2,1),'-bp')
```

```
text(1,0.4,'$\alpha=2$','Interpreter','Latex')
plot(x0,gampdf(x0,4,1),'.k-')
text(4.5,0.2,'$\alpha=4$','Interpreter','Latex')
h=legend('$\alpha=0.5$','$\alpha=2$','$\alpha=4$');
set(h,'Interpreter','Latex')
xlabel('$x$','Interpreter','Latex')
ylabel('$f(x)$','Interpreter','Latex')
subplot(122), hold on
plot(x0,gampdf(x0,4,2),'-r.')
text(1.6,0.4,'$\beta=2$','Interpreter','Latex')
plot(x0,gampdf(x0,2,1),'-bp')
text(5,0.05,'$\beta=1$','Interpreter','Latex')
plot(x0,gampdf(x0,2, 0.5),'.k-')
text(9.2,0.1,'$\beta=0.5$','Interpreter','Latex')
h=legend('$\beta=2$','$\beta=1$','$\beta=0.5$');
set(h,'Interpreter','Latex')
xlabel('$x$','Interpreter','Latex')
ylabel('$f(x)$','Interpreter','Latex')
```

9.3 假设检验

统计推断的另一类重要问题是假设检验问题。在总体的分布函数完全未知或只知其形式但不知其参数的情况，为了推断总体的某些性质，提出某些关于总体的假设。例如，提出总体服从泊松分布的假设，又如对于正态总体提出数学期望等于 μ_0 的假设等。假设检验就是根据样本对所提出的假设做出判断：是接受还是拒绝。这就是所谓的假设检验问题。

9.3.1 参数检验

1.单个总体 $N(\mu,\sigma^2)$ 均值 μ 的检验

原假设（或零假设）为：$H_0:\mu=\mu_0$。

备选假设有三种可能：

$H_1:\mu\neq\mu_0$；$H_1:\mu>\mu_0$；$H_1:\mu<\mu_0$。

（1）σ^2 已知，关于 μ 的检验（Z 检验）

在 MATLAB 中 Z 检验法由函数 ztest 来实现，命令为

[h,p,ci]=ztest(x,mu,sigma,alpha,tail)

其中输入参数 x 是样本，mu 是 H_0 中的 μ_0，sigma 是总体标准差 σ，alpha 是显著性水平 α（alpha 缺省时设定为 0.05），tail 是对备选假设 H_1 的选择：H_1 为 $\mu\neq\mu_0$ 时用 tail=0（可缺省）；H_1 为 $\mu>\mu_0$ 时用 tail=1；H_1 为 $\mu<\mu_0$ 时用 tail=-1。输出参数 h=0 表示接受 H_0，h=1 表示拒绝 H_0，p 表示在假设 H_0 下样本均值出现的概率，p 越小 H_0 越值得怀疑，ci 是 μ 的置信区间。

例 9.11 某车间用一台包装机包装糖果。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 公斤。某日开工后为检验包装机是否正常，随机地抽取它所包装的糖 9 袋，称得净重为（公斤）：

0.497　0.506　0.518　0.524　0.498　0.511　0.520　0.515　0.512

问机器是否正常？

总体 σ 已知，$x\sim N(\mu,0.015^2)$，μ 未知。于是提出假设 $H_0:\mu=\mu_0=0.5$ 和 $H_1:\mu\neq0.5$。

MATLAB 实现如下：

```
x=[0.497  0.506  0.518  0.524  0.498  0.511  0.520  0.515  0.512];
[h,p,ci]=ztest(x,0.5,0.015)
```

求得 h=1，p=0.0248，说明在 0.05 的水平下，可拒绝原假设，即认为这天包装机工作不正常。

（2）σ^2 未知，关于 μ 的检验（t 检验）

在 MATLAB 中 t 检验法由函数 ttest 来实现，命令为

[h,p,ci]=ttest(x,mu,alpha,tail)

例 9.12　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)?

按题意需检验

$$H_0:\mu\le\mu_0=225,\quad H_1:\mu>225,$$

取 $\alpha=0.05$。MATLAB 实现如下：

```
clc, clear
x=[159  280  101  212  224  379  179  264
   222  362  168  250  149  260  485  170]; x=x(:);
[h,p,ci]=ttest(x,225,0.05,1)
```

求得 h=0，p=0.2570，说明在显著水平为 0.05 的情况下，不能拒绝原假设，认为元件的平均寿命不大于 225 小时。

2.两个正态总体均值差的检验（t 检验）

还可以用 t 检验法检验具有相同方差的 2 个正态总体均值差的假设。在 MATLAB 中由函数 ttest2 实现，命令为：

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

与上面的 ttest 相比，不同处只在于输入的是两个样本 x,y（长度不一定相同）。

例 9.13　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得率,试验是在同一平炉上进行的。每炼一炉钢时除操作方法外,其它条件都可能做到相同。先用标准方法炼一炉,然后用建议的新方法炼一炉,以后交换进行,各炼了 10 炉,其得率分别为

1 标准方法　78.1　72.4　76.2　74.3　77.4　78.4　76.0　75.6　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 $N(\mu_1,\sigma^2)$ 和 $N(\mu_2,\sigma^2)$, μ_1,μ_2,σ^2 均未知，问建议的新方法能否提高得率?(取 $\alpha=0.05$。)

（i）需要检验假设

$$H_0:\mu_1-\mu_2=0,\quad H_1:\mu_1-\mu_2<0.$$

(ii)MATLAB 实现

```
clc, clear
x=[78.1  72.4  76.2  74.3  77.4  78.4  76.0  75.6  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,p,ci]=ttest2(x,y,0.05,-1)
```

求得 h=1,p=2.2126×10^{-4}。表明在 $\alpha=0.05$ 的显著水平下，可以拒绝原假设，即认为建议的新操作方法较原方法优。

9.3.2　非参数检验

1.分布拟合检验

在实际问题中，有时不能预知总体服从什么类型的分布，这时就需要根据样本来检验关于分布的假设。下面介绍 χ^2 检验法

若总体 X 是离散型的，则建立待检假设 H_0：总体 X 的分布律为 $P\{X=x_i\}=p_i$，$i=1,2,\cdots$。

若总体 X 是连续型的，则建立待检假设 H_0：总体 X 的概率密度为 $f(x)$。

可按照下面的五个步骤进行检验：

（1）建立待检假设 H_0：总体 X 的分布函数为 $F(x)$。

（2）在数轴上选取 $k-1$ 个分点 $t_1,t_2,\cdots,t_{k-1}$，将数轴分成 k 个区间：$(-\infty,t_1)$，$[t_1,t_2)$，…，$[t_{k-2},t_{k-1})$，$[t_{k-1},+\infty)$，令 p_i 为分布函数 $F(x)$ 的总体 X 在第 i 个区间内取值的概率，设 m_i 为 n 个样本观察值中落入第 i 个区间上的个数，也称为组频数。

（3）选取统计量 $\chi^2=\sum_{i=1}^{k}\frac{(m_i-np_i)^2}{np_i}$，如果 H_0 为真，则 $\chi^2\sim\chi^2(k-1-r)$，其中 r 为分布函数 $F(x)$ 中未知参数的个数。

（4）对于给定的显著性 α，确定 χ^2_α，使其满足 $P\{\chi^2(k-1-r)>\chi^2_\alpha\}=\alpha$，并且依据样本计算统计量 χ^2 的观察值。

（5）作出判断：若 $\chi^2<\chi^2_\alpha$，则接受 H_0；否则拒绝 H_0，即不能认为总体 X 的分布函数为 $F(x)$。

例 9.14　检查了一本书的 100 页，记录各页中印刷错误的个数，其结果见**表** 9.6。

表 9.6　印刷错误数据表

错误个数 f_i	0	1	2	3	4	5	6	≥7
含 f_i 个错误的页数	36	40	19	2	0	2	1	0

问能否认为一页的印刷错误的个数服从泊松分布（取 $\alpha=0.05$）。

解　记一页的印刷错误数为 X，按题意需在显著性水平 $\alpha=0.05$ 下检验假设

H_0：X 的分布律为

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

因参数 λ 未知，应先根据观察值，用矩估计法来求 λ 的估计（在 H_0 下）。可知 λ 的矩估计值为 $\hat{\lambda}=\bar{x}=1$。在 X 服从泊松分布的假设下，X 的所有可能取得的值为 $\Omega=\{0,1,2,\cdots\}$，将 Ω 分成如表 9-7 左起第一栏所示的两两不相交的子集：A_0,A_1,A_2,A_3,A_4，接着根据估计式

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

计算有关概率的估计，计算结果列表见表 9.7。

表 9.7　χ^2 检验过程数据表

A_i	f_i	$\hat{p}_i$	$n\hat{p}_i$	$f_i^2/(n\hat{p}_i)$
$A_0:\{X=0\}$	36	0.3679	36.79	35.2270
$A_1:\{X=1\}$	40	0.3679	36.79	43.4901
$A_2:\{X=2\}$	19	0.1839	18.39	19.6302
$A_3:\{X\geq 3\}$	5	8.0291	8.03	3.1133
				$\Sigma=101.4606$

今 $\chi^2=101.4606-100=1.4606$，因估计了一个参数，$r=1$，只有 4 组，故 $k=4$，$\alpha=0.05$，$\chi^2_\alpha(k-r-1)=\chi^2_{0.05}(2)=5.9915>1.4606=\chi^2$，故在显著性水平 $\alpha=0.05$ 下接受假设 H_0，即认为样本来自泊松分布的总体。

计算的 MATLAB 程序如下

```
clc, clear, n=100;
f=0:7; num=[36  40  19  2  0  2  1  0];
lamda=dot(f,num)/100
pi=poisspdf(f,lamda)
[h,p,st]=chi2gof(f,'ctrs',f,'frequency',num,'expected',n*pi,'nparams',1) %调用工具箱
k2=chi2inv(0.95,st.df)  %求临界值
```

例 9.15　在一批灯泡中抽取 300 只作寿命试验，其结果如表 9.8。

表 9.8　寿命测试数据表

寿命 t（h）	$0 \le t \le 100$	$100 < t \le 200$	$200 < t \le 300$	$t > 300$
灯泡数	121	78	43	58

取 $\alpha = 0.05$，试检验假设

H_0：灯泡寿命服从指数分布

$$f(t) = \begin{cases} 0.005e^{-0.005t}, & t \ge 0, \\ 0, & t < 0. \end{cases}$$

解　本题是在显著性水平 $\alpha = 0.05$ 下，检验假设：H_0：灯泡寿命 X 服从指数分布，其概率密度为

$$f(t) = \begin{cases} 0.005e^{-0.005t}, & t \ge 0, \\ 0, & t < 0. \end{cases}$$

在 H_0 为真的假设下，X 可能取值的范围为 $\Omega = [0,+\infty)$。将 Ω 分成互不相交的 4 个部分：A_1, A_2, A_3, A_4 如表 9-9。以 A_i 记事件 $\{X \in A_i\}$。若 H_0 为真，X 的分布函数为

$$F(t) = \begin{cases} 1 - e^{-0.005t}, & t \ge 0, \\ 0, & t < 0. \end{cases}$$

得知

$$p_i = P(A_i) = P\{a_i < X \le a_{i+1}\} = F(a_{i+1}) - F(a_i),\ i = 1,2,3,4.$$

计算结果列于表 9.9。

表 9.9　χ^2 检验过程数据表

A_i	f_i	$\hat{p}_i$	$n\hat{p}_i$	$f_i^2/(n\hat{p}_i)$
$A_1: 0 \le t \le 100$	121	0.3935	118.05	124.0237
$A_2: 100 < t \le 200$	78	0.2387	71.61	84.9602
$A_3: 200 < t \le 300$	43	0.1447	43.41	42.5939
$A_4: t > 300$	58	0.2231	66.93	50.2615
				$\Sigma = 301.8393$

今 $\chi^2 = 1.8393$。由 $\alpha = 0.05$，$k = 4$，$r = 0$ 知

$$\chi_\alpha^2(k - r - 1) = \chi_{0.05}^2(3) = 7.8393 > 1.8393 = \chi^2.$$

故在显著性水平 $\alpha = 0.05$ 下，接受假设 H_0，认为这批灯泡寿命服从指数分布，其概率密度为

$$f(t) = \begin{cases} 0.005e^{-0.005t}, & t \ge 0, \\ 0, & t < 0. \end{cases}$$

计算的 MATLAB 程序如下

```
clc, clear
edges=[0:100:300 inf]; bins=[50 150 250 inf]; %定义区域的边界和中心
num=[121 78 43 58]; %
pd=makedist('exp',200)   %定义指数分布
[h,p,st]=chi2gof(bins,'Edges',edges,'cdf',pd,'Frequency',num)
k2=chi2inv(0.95,st.df)   %求临界值
```

例 9.16　下面表 9.10 给出了随机选取的某大学 200 名一年级学生一次数学考试的成绩。试取 $\alpha=0.1$ 检验数据来自正态总体 $N(60,15^2)$。

表 9.10　学生分数统计数据

分数 x	$20\le x\le 30$	$30<x\le 40$	$40<x\le 50$	$50<x\le 60$
学生数	5	15	30	51
分数 x	$60<x\le 70$	$70<x\le 80$	$80<x\le 90$	$90<x\le 100$
学生数	60	23	10	6

解　本题要求在显著性水平 $\alpha=0.1$ 下检验假设

H_0：数据 X 来自正态总体，$X\sim N(60,15^2)$，

即需检验 X 的概率密度为

$$f(x)=\frac{1}{15\sqrt{2\pi}}e^{-\frac{(x-60)^2}{2\times 15^2}},\quad -\infty<x<+\infty.$$

将在 H_0 下 X 可能取值的区间 $(-\infty,+\infty)$ 分为6个两两不相交的小区间 $A_1,A_2,\cdots,A_6$（分法见表9.11）。用 A_i 记事件“X 的观测值落在 A_i 内”，以 f_i（$i=1,2,\cdots,6$）记样本观察值 $x_1,x_2,\cdots,x_{200}$ 中落在 A_i 的个数，记 $p_i=P\{X\in A_i\}$。计算结果列于表9.11。

表 9.11　χ^2 检验过程数据表

A_i	f_i	$\hat{p}_i$	$n\hat{p}_i$	$f_i^2/(n\hat{p}_i)$
$A_1:(-\infty,40]$	20	0.0912	18.2422	21.9271
$A_2:(40,50]$	30	0.1613	32.2563	27.9016
$A_3:(50,60]$	51	0.2475	49.5015	52.5439
$A_4:(60,70]$	60	0.2475	49.5015	72.7251
$A_5:(70,80]$	23	0.1613	32.2563	16.3999
$A_6:(80,+\infty)$	16	0.0912	18.2422	14.0334
				$\Sigma=205.5309$

因此 $\chi^2=5.5309$。因 $\alpha=0.1$，$k=6$，$r=0$，有

$$\chi_\alpha^2(k-r-1)=\chi_{0.1}^2(5)=9.2364>5.5309=\chi^2.$$

故在显著性水平 $\alpha=0.1$ 下接受假设 H_0，即认为考试成绩的数据来自正态总体 $N(60,15^2)$。

```
clc, clear,alpha=0.1;
edges=[-inf 20:10:100 inf]; %区间的边界
x=[25:10:95]; %区间的中心
num=[5 15 30 51 60 23 10 6];
```

```
pd=@(x)normcdf(x,60,15); %定义正态分布的分布函数
[h,p,stats]=chi2gof(x,'cdf',pd,'Edges',edges,'Frequency',num)
k2=chi2inv(1-alpha,stats.df) %求临界值
```

例 9.17（续例 9. 1）用 χ^2 检验法检验身高的数据是否服从正态分布。

解　采用矩估计方法估计参数的取值。先从所给的身高数据算出样本均值和标准差

$$\bar{x}=143.7738,\quad s=5.9705,$$

本题是在显著性水平 $\alpha=0.05$ 下，检验假设：H_0：身高的观测数据服从正态分布 $N(143.7738,5.9705)$。检验的过程我们这里就不给出了，利用 MATLAB 的计算结果是可以接受假设 H_0，认为这些数据是来自正态总体。

计算的 MATLAB 程序如下：

```
clc, clear
a=load('data91.txt');
h=a(:,[1:2:end]); h=h(:); %提取身高数据并转换为列向量
mu=mean(h); s=std(h);
pd=@(x)normcdf(x,mu,s); %定义正态分布的分布函数
[h,p,stats]=chi2gof(h,'cdf',pd,'Nparams',2)
```

2.其他非参数检验方法

MATLAB 工具箱还有如下一些非参数检验方法，我们就不一一介绍了。

```
jbtest(x,alpha)   %正态总体的拟合优度检验—Jarque-Bera检验。
lillietest(x,Name,Value)   %正态总体的拟合优度检验—Lilliefors检验。
adtest(x,Name,Value)   %Anderson-Darling检验。
kstest(x, Name,Value)   %一个样本的Kolmogorov-Smirnov检验。
kstest2(x1,x2,Name,Value)      %两个样本的 Kolmogorov-Smirnov 检验
```

9.4　方差分析

方差分析实际上是多个总体的假设检验问题。下面我只给出单因素方差分析。

设因素 A 有 s 个水平 $A_1,A_2,\cdots,A_s$，在水平 A_j（$j=1,2,\cdots,s$）下，进行 n_j（$n_j\geq 2$）次独立试验，得出表 9.12 所列结果。

表 9.12　方差分析数据表

	A_1	A_2	$\cdots$	A_s
试验批号	X_{11}	X_{12}	$\cdots$	X_{1s}
	X_{21}	X_{22}	$\cdots$	X_{2s}
	$\vdots$	$\vdots$		$\vdots$
	$X_{n_1 1}$	$X_{n_2 2}$	$\cdots$	$X_{n_s s}$
样本总和 $T_{\bullet j}$	$T_{\bullet 1}$	$T_{\bullet 2}$	$\cdots$	$T_{\bullet s}$
样本均值 $\bar{X}_{\bullet j}$	$\bar{X}_{\bullet 1}$	$\bar{X}_{\bullet 2}$	$\cdots$	$\bar{X}_{\bullet s}$
总体均值	μ_1	μ_2	$\cdots$	μ_s

其中 X_{ij} 表示第 j 个等级进行第 i 次试验的可能结果，记 $n=n_1+n_2+\cdots+n_s$，

$$\bar{X}_{\bullet j}=\frac{1}{n_j}\sum_{i=1}^{n_j}X_{ij},\quad T_{\bullet j}=\sum_{i=1}^{n_j}X_{ij},\quad \bar{X}=\frac{1}{n}\sum_{j=1}^{s}\sum_{i=1}^{n_j}X_{ij},\quad T_{\bullet\bullet}=\sum_{j=1}^{s}\sum_{i=1}^{n_j}X_{ij}=n\bar{X}.$$

1.方差分析的假设前提

1° 对变异因素的某一个水平，例如第 j 个水平，进行实验，把得到的观察值 $X_{1j},X_{2j},\cdots,X_{n_jj}$ 可以看成是从正态总体 $N(\mu_j,\sigma^2)$ 中取得的一个容量为 n_j 的样本，且 μ_j,σ^2 未知。

2° 对于表示 s 个水平的 s 个正态总体的方差认为是相等的；

3° 由不同总体中抽取的样本相互独立。

2.统计假设

提出待检假设 H_0： $\mu_1=\mu_2=\cdots=\mu_s=\mu$。

3.检验方法

设 $S_T=\sum_{j=1}^{s}\sum_{i=1}^{n_j}(X_{ij}-\overline{X})^2=\sum_{j=1}^{s}\sum_{i=1}^{n_j}X_{ij}^2-\frac{T_{\bullet\bullet}^2}{n}$，

$S_E=\sum_{j=1}^{s}\sum_{i=1}^{n_j}(X_{ij}-\overline{X}_{\bullet j})^2=\sum_{j=1}^{s}\sum_{i=1}^{n_j}X_{ij}^2-\sum_{j=1}^{s}\frac{T_{\bullet j}^2}{n_j}$， $S_A=S_T-S_E$，

若 H_0 为真，则检验统计量 $F=\frac{(n-s)S_A}{(s-1)S_E}\sim F(s-1,n-s)$，对于给定的显著性水平 α，查表确定临界值 F_α， 使得 $P\left\{\frac{(n-s)S_A}{(s-1)S_E}>F_\alpha\right\}=\alpha$，依据样本值计算检验统计量 F 的观察值，并与 F_α 比较，最后下结论：若检验统计量 F 的观察值大于临界值 F_α，则拒绝原假设 H_0；若 F 的值小于 F_α， 则接受 H_0。

例 9.18　设有如表 9.13 所列的 3 个组 5 年保险理赔额的观测数据。试用方差分析法检验 3 个组的理赔额均值是否有显著差异（取显著性水平 $\alpha=0.05$， 已知 $F_{0.05}(2,12)=3.88$）。

表 9.13　**保险理赔额观测数据**

	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$
$j=1$	98	93	103	92	110
$j=2$	100	108	118	99	111
$j=3$	129	140	108	105	116

解　用 X_{jt} 来表示第 j 组第 t 年的理赔额，其中 $j=1,2,3$， $t=1,2,\cdots,5$。假设所有的 X_{jt} 相互独立且服从 $N(m_j,s^2)$ 分布，即对应于每组均值 m_j 可能不相等，但是方差 $s^2>0$ 是相同的。

记 $\overline{X}_{j\bullet}=\frac{1}{5}\sum_{t=1}^{5}X_{jt}$， $\overline{X}=\frac{1}{15}\sum_{j=1}^{3}\sum_{t=1}^{5}X_{jt}$，

$S_A=\sum_{j=1}^{3}5(\overline{X}_{j\bullet}-\overline{X})^2$， $S_E=\sum_{j=1}^{3}\sum_{t=1}^{5}(X_{jt}-\overline{X}_{j\bullet})^2$.

提出原假设 H_0： $m_1=m_2=m_3$， H_1： m_1,m_2,m_3 不全相等。

若 H_0 为真，则检验统计量 $F=\frac{12S_A}{2S_E}\sim F(2,12)$，对于给定的显著性水平 α，及临界值 $F_\alpha(2,12)$，依据样本值计算检验统计量 F 的观察值，并与 $F_\alpha(2,12)$ 比较，最后下结论：若检验统计量 F 的观察值大于临界值 $F_\alpha(2,12)$，则拒绝原假设 H_0；若 F 的值小于

$F_\alpha(2,12)$，则接受 H_0。

这里求得 $S_A=1056.53$ ，自由度为 2 ， $S_E=1338.8$，自由度为 12。于是 $F=4.73$，这与临界值 $F_{0.05}(2,12)=3.8853$ 比较起来数值过大了。我们的结论是这些数据表明每组的平均理赔不全相等。

计算的 MATLAB 程序如下：

```
clc, clear
a=[98    93   103  92   110
100  108  118  99   111
129  140  108  105  116];
p=anova1(a')   %进行单因素方差分析
fws=finv(0.95,2,12)   %求上 0.05 分位数
```

9.5 回归分析

9.5.1 线性回归分析

例 9.19 某种商品的需求量 y，消费者平均收入 x_1 以及商品价格 x_2 的统计数据如表 9.14。求 y 关于 x_1, x_2 的回归方程 $y=b_0+b_1x_1+b_2x_2$。

表 9.14

x_{1i}	1000	600	1200	500	300	400	1300	1100	1300	300
x_{2i}	5	7	6	6	8	7	5	4	3	9
y	100	75	80	70	50	65	90	100	110	60

利用 MATLAB 软件求得回归方程为 $y=111.6918+0.0143x_1-7.1882x_2$。

计算的 MATLAB 程序如下

```
clc, clear
a=[1000  600  1200      500  300  400  1300      1100      1300      300
5    7    6    6    8    7    5    4    3    9
100 75   80   70   50   65   90   100  110  60];
x=[ones(10,1),a([1:2],:)']; y=a(3,:)';
[b,bint,r,rint,st]=regress(y,x)
```

注：本例中线性回归模型主要是演示 MATLAB 命令的使用，从输出结果看，由于参数 b_1 的区间估计包含 0 点，所以变量 x_1 是不显著的。如果建立线性回归模型，是不能使用变量 x_1 的。

也可以用函数 fittype 与 fit 来拟合参数 b_0, b_1, b_2。

```
clc, clear
a=[1000  600  1200      500  300  400  1300      11001300      300
5    7    6    6    8    7    5    4    3    9
100 75   80   70   50   65   90   100  110  60];
x12=a([1,2],:)'; y=a(3,:)';
f=fittype('b0+b1*x1+b2*x2','independent',{'x1','x2'})
f0=fit(x12,y,f,'Start',rand(1,3))
```

9.5.2 多元二项式回归

统计工具箱提供了一个作多元二项式回归的命令rstool，它产生一个交互式画面，并输出有关信息，用法是

rstool(X,Y,model,alpha)，

其中alpha为显著性水平 α （缺省时设定为0.05），model可选择如下的4个模型（用字符串输入，缺省时设定为线性模型）

linear(线性)： $y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m$；

purequadratic(纯二次)： $y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m+\sum_{j=1}^{m}\beta_{jj}x_j^2$；

interaction（交叉）： $y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m+\sum_{1\le j<k\le m}\beta_{jk}x_jx_k$；

quadratic(完全二次)： $y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m+\sum_{1\le j\le k\le m}\beta_{jk}x_jx_k$.

$[y,x_1,\cdots,x_m]$的n个独立观测数据记为$[b_i,a_{i1},\cdots,a_{im}]$，$i=1,\cdots,n$，Y，XX 分别为$n$维列向量和$n\times m$矩阵，这里

$$\mathrm{Y}=\begin{bmatrix}b_1\\ \vdots\\ b_n\end{bmatrix},\quad \mathrm{XX}=\begin{bmatrix}a_{11} & \cdots & a_{1m}\\ \vdots & & \vdots\\ a_{n1} & \cdots & a_{nm}\end{bmatrix}.$$

注：（1）这里多元二项式回归中，数据矩阵 XX 与线性回归分析中的数据矩阵 X 是有差异的，后者的第一列为全 1 的列向量。

（2）在完全二次多项式回归中，二次项系数的排列次序是先交叉项的系数，最后是纯二次项的系数。

例 9.20（续例 9.19）根据表 9.14 的数据，在关于x_1,x_2的 linear(线性)、purequadratic(纯二次)、interaction（交叉）、quadratic(完全二次)的 4 个模型中，根据剩余标准差指标选择一个最好的模型。

运行如下 MATLAB 程序：

```
clc, clear
a=[1000  600  1200      500  300  400  1300      11001300      300
5     7     6     6     8     7     5     4     3     9
100  75    80    70    50    65    90    100  110  60];
x12=a([1,2],:)'; y=a(3,:)';
rstool(x12,y)
```

把计算结果利用左下角的“Export”按钮输出到 MATLAB 工作空间，如图 9.9 所示。依次选择图 9.9 左下角的第 2 个下拉框中的其他三个模型，把计算结果到输出到工作空间，在工作空间中比较 rmse(剩余标准差)，选择最好的模型。

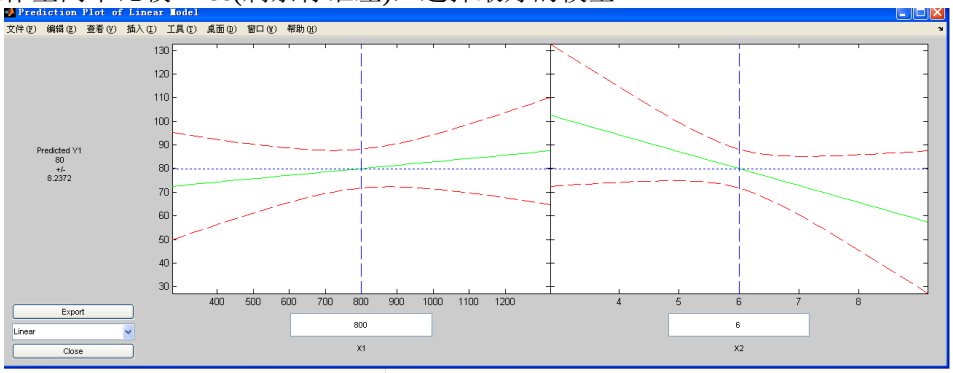

图 9.9　rstool 的图形界面

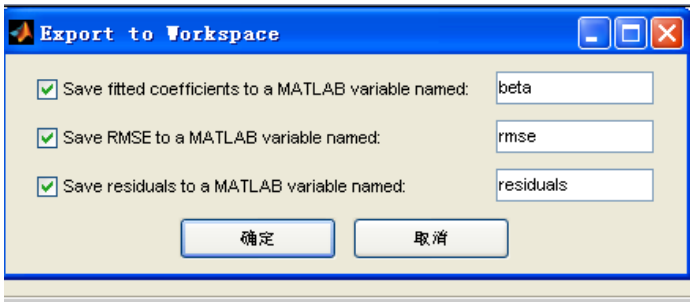

图 9.10　输出结果到工作空间

计算得到线性模型的剩余标准差为 7.2133，纯二次项模型的剩余标准差为 4.5362，交叉项模型的剩余标准差为 7.5862，完全二次项模型的标准差为 4.4179，所有我们选择完全二次项模型，所求的完全二次项模型为

$$y=-106.6095+0.3261x_1+21.299x_2-0.02x_1x_2-0.0001x_1^2-0.7609x_2^2.$$

9.5.3　非线性回归

例 9.21（续例 9.19）使用表 9.14 的数据，建立非线性回归模型

$$y=\frac{\beta_1 x_2}{1+\beta_2 x_1+\beta_3 x},$$

并求 $x_1=10$，$x_2=20$ 时 y 的预测值。

运行结果不稳定，我们就不给出答案了。

计算的 MATLAB 程序如下：

```
clc, clear, format long g
a=[1000  600  1200      500  300  400  1300      11001300      300
5    7    6    6    8    7    5    4    3    9
100  75   80   70   50   65   90   100  110  60];
x12=a([1,2],:)'; y=a(3,:)';
yx=@(beta,x)beta(1)*x(:,2)./(1+beta(2)*x(:,1)+beta(3)*x(:,2));
[beta,r,j]=nlinfit(x12,y,yx,rand(1,3))
yhat=nlpredci(yx,[10,20],beta,r,'jacobian',j)
format
```

使用 fittype 和 fit 命令拟合的结果每次也不一样。

计算的 MATLAB 程序如下：

```
clc, clear, format long g
a=[1000  600  1200      500  300  400  1300      11001300      300
5    7    6    6    8    7    5    4    3    9
100  75   80   70   50   65   90   100  110  60];
x12=a([1,2],:)'; y=a(3,:)';
yx=@(b1,b2,b3,x1,x2)b1*x2./(1+b2*x1+b3*x2);
yx=fittype(yx,'independent',{'x1','x2'})
yt=fit(x12,y,yx,'Start',rand(1,3))
yhat=yt(10,20) %计算预测值
format
```

注：非线性问题一般来说都比较困难的，除非指定什么特殊算法，经常会碰到运行结果不稳定的情况。碰到结果不稳定时，试试用 Lingo 编程，并用全局求解器进行求解，但运行时间很长。

习题 9

9.1　（99 年全国大学生数学建模竞赛 A 题）自动化车床管理

一道工序用自动化车床连续加工某种零件，由于刀具损坏等原因该工序会出现故障，其中刀具损坏故障占 95%, 其它故障仅占 5%。工序出现故障是完全随机的, 假定在生产任一零件时出现故障的机会均相同。工作人员通过检查零件来确定工序是否出现故障。现积累有 100 次刀具故障记录，故障出现时该刀具完成的零件数如表 9.15。现计划在刀具加工一定件数后定期更换新刀具。

已知生产工序的费用参数如下：故障时产出的零件损失费用 f=200 元/件；进行检查的费用 t=10 元/次；发现故障进行调节使恢复正常的平均费用 d=3000 元/次(包括刀具费)；未发现故障时更换一把新刀具的费用 k=1000 元/次。

1)假定工序故障时产出的零件均为不合格品，正常时产出的零件均为合格品, 试对该工序设计效益最好的检查间隔（生产多少零件检查一次）和刀具更换策略。

2)如果该工序正常时产出的零件不全是合格品，有 2%为不合格品；而工序故障时产出的零件有 40%为合格品，60%为不合格品。工序正常而误认有故障仃机产生的损失费用为 1500 元/次。对该工序设计效益最好的检查间隔和刀具更换策略。

3）在 2)的情况, 可否改进检查方式获得更高的效益。

表 9.15　100 次刀具故障记录(完成的零件数)

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	649	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

9.2　将冰晶放入一容器内，容器内维持规定的温度（–5℃）和固定的湿度。观察自冰晶放入的时刻开始计算的时间 T（以 s 计）和晶体生长的轴向长度 A（以 μm 计），得到 43 对观察数据如表 9.16 所示。

表 9.16　晶体生长的观察数据

T	50	60	60	70	70	80	80	90	90	90	95	100	100	100	105
A	19	20	21	17	22	25	28	21	25	31	25	30	29	33	35
T	105	110	110	110	115	115	115	120	120	120	125	130	130	135	135
A	32	30	28	30	31	36	30	36	25	28	28	31	32	34	25
T	140	140	145	150	150	155	155	160	160	160	165	170	180		
A	26	33	31	36	33	41	33	40	30	37	32	35	38		

设题目符号回归模型所要求的条件。

（1）画出散点图。

（2）求线性回归方程 $\hat{A}=\hat{a}+\hat{b}T$ 。

（3）检验（2）中所建立的回归模型。

9.3　在 7 个不同实验室中测量某种扑尔敏药片的扑尔敏有效含量（以 mg 计）。得到的结果（Lab 表示实验室）见表 9.17。

表 9.17　扑尔敏有效含量数据

Lab1	Lab2	Lab3	Lab4	Lab5	Lab6	Lab7
4.13	3.86	4.00	3.88	4.02	4.02	4.00
4.07	3.85	4.02	3.88	3.95	3.86	4.02

4.04	4.08	4.01	3.91	4.02	3.96	4.03
4.07	4.11	4.01	3.95	3.89	3.97	4.04
4.05	4.08	4.04	3.92	3.91	4.00	4.10
4.04	4.01	3.99	3.97	4.01	3.82	3.81
4.02	4.02	4.03	3.92	3.89	3.98	3.91
4.06	4.04	3.97	3.90	3.89	3.99	3.96
4.10	3.97	3.98	3.97	3.99	4.02	4.05
4.04	3.95	3.98	3.90	4.00	3.93	4.06

（1）画出各实验室测量结果的箱线图。

（2）设各样本分别来自正态总体 $N(\mu_i,\sigma^2)$， $i=1,2,\cdots,7$，各样本相互独立。试取显著水平 $\alpha=0.05$ 检验各实验室测量的扑尔敏的有效含量的均值是否有显著差异。