第3章 统计估计

统计估计是统计推断的主要内容，包括两个方面的任务。

(1) 变量的分布形态未知，根据样本数据对变量的分布形态做出推测(估计)。

(2) 变量的分布形态已知，即已知其概率分布函数(或概率分布律，或概率密度函数)的数学表达式，但是某些参数(或数字特征)未知，根据样本数据对未知的参数(或未知参数的函数)做出估计。

3.1 点估计

点估计(point estimation)是用样本统计量来估计总体参数，因为样本统计量为数轴上某一点值，估计的结果也以一个点的数值表示，所以称为点估计。点估计和区间估计属于总体参数估计问题。何为总体参数统计，当在研究中从样本获得一组数据后，如何通过这组信息，对总体特征进行估计，也就是如何从局部结果推论总体的情况，称为总体参数估计。

点估计也称定估计，它是以抽样得到的样本指标作为总体指标的估计量，并以样本指标的实际值直接作为总体未知参数的估计值的一种推断方法。

点估计的方法有矩估计法、顺序统计量法、最大似然法、最小二乘法等。

3.1.1 矩估计

在统计学中，矩是指以期望为基础而定义的数字特征，一般分为原点矩和中心矩。

设 X 为随机变量，对任意正整数 k，称 $E(X^k)$ 为随机变量 k 阶原点矩，记为

$$m^k = E(X^k)$$

当 $k=1$ 时，$m_1=E(X)=\mu$。可见一阶原点矩为随机变量 X 的数学期望。我们把 $c_k=E[X-E(X)^k]$ 称为以 $E(X)$ 为中心的 k 阶中心矩。

显然，当 $k=2$ 时，$c_2=E[X-E(X)^2]=\sigma^2$。可见，二阶中心矩为随机变量 X 的方差。

【例 3-1】 已知某种灯泡的寿命 $X\sim N(\mu,\sigma^2)$，其中，μ、σ^2 都是未知的，今随机取得 4 只灯泡，测得寿命(单位：小时)为 1052、1453、1367、1650，试估计 μ 和 σ。

解：因为 μ 为全体灯泡的平均寿命，$\bar{x}$ 为样本的平均寿命，很自然地会想到用 $\bar{x}$ 去估计 μ；同理用 S 去估计。由于

$$\bar{x}=\frac{1}{4}(1502+1453+1367+1650)=1493$$

$$S^2=\frac{(1502-1493)^2+(1453-1493)^2+(1367-1493)^2+(1650-1493)^2}{4-1}=14\,068.7$$

$$S=118.61$$

因此，μ 和 σ 的估计值分别为 1493 小时及 118.61 小时。

矩估计简便、直观、比较常用，但是矩估计法也有其局限性。首先，它要求总体的 k 阶原点矩存在，如果不存在则无法估计；其次，矩估计法不能充分地利用估计时已掌握的有关总体分布形式的信息。

通常设 θ 为总体 X 的待估计参数，一般用样本 $X_1,X_2,\cdots,X_n$ 构成一个统计量 $\hat{\theta}=\hat{\theta}(X_1,X_2,\cdots,X_n)$来估计 θ，则称 $\hat{\theta}$ 为 θ 的估计量。对于样本的一组数值 $x_1,x_2,\cdots,x_n$，估计量 $\hat{\theta}$ 的值$\hat{\theta}(x_1,x_2,\cdots,x_n)$称为 θ 的估计值。于是点估计即是寻求一个作为待估计参数 θ 的估计量 $\hat{\theta}(x_1,x_2,\cdots,x_n)$的问题。但是必须注意，对于样本的不同数值，估计值是不相同的。

如在上例中，分别用样本平均数和样本修正方差来估计总数数学期望和总体均方差，即有

$$\hat{\mu}=\hat{\mu}(X_1,X_2,\cdots,X_n)=\frac{1}{n}\sum_{i=1}^{n}X_i=\overline{X}$$

$$\hat{\sigma}=\hat{\sigma}(X_1,X_2,\cdots,X_n)=\sqrt{\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}}=S$$

其中对应于给定的估计值 $\mu=\bar{x}=1493\text{h}$，$\hat{\sigma}=S=118.61\text{h}$。

3.1.2 极大似然估计

极大似然估计方法(Maximum Likelihood Estimate，MLE)也称为最大概率似然估计或最大似然估计，是求估计的另一种方法，1821 年首先由德国数学家 C. F. Gauss(高斯)提出，但是这个方法通常被归功于英国的统计学家 R. A. Fisher(罗纳德·费希尔)，他在 1922 年的论文 *On the mathematical foundations of theoretical statistics*，*reprinted in Contributions to Mathematical Statistics*(by R. A. Fisher，1950，J. Wiley & Sons，New York)中再次提出了这个思想，并且首先探讨了这种方法的一些性质，极大似然估计这一名称也是费希尔给的。这是一种目前仍然得到广泛应用的方法。

下面分别讨论 X 为离散型和连续型随机变量时总体中某些参数 θ 的最大似然估计。

设总体 X 是离散型随机变量，其分布律 $P\{X=x\}=p\{x;\theta\}$，其中 θ 是未知参数，如果取得样本观测值为 $x_1,x_2,\cdots,x_n$，则表示随机事件 $X_1=x_1,X_2=x_2,\cdots,X_n=x_n$ 发生了。考虑 n 个事件 $X_1=x_1,X_2=x_2,\cdots,X_n=x_n$ 的交点的概率，注意到 $X_1,X_2,\cdots,X_n$ 的独立性，即有

$$\begin{aligned} L(\theta) &= P\{X_1=x_1,X_2=x_2,\cdots,x_n=x_n\} \\ &= P\{X_1=x_1\}P\{X_2=x_2\}\cdots P\{X_n=x_n\} \\ &= p\{x_1;\theta\}p\{x_2;\theta\}\cdots p\{x_n;\theta\} \\ &= \prod_{i=1}^{n} p\{x;\theta\} \end{aligned} \tag{3-1}$$

函数 $L(\theta)$称为似然函数，对于已给定的 $x_1,x_2,\cdots,x_n$，它是未知参数 θ 的函数。

按极大似然估计法的直观想法是：若抽样的结果得到样本观测值 $x_1,x_2,\cdots,x_n$，则应当这样选取参数 $L(\theta)$的值，使这组样本观测值出现的可能性最大，也就是使似然函数 $L(\theta)$达到最大值，从而求得参数 θ 的估计值 $\hat{\theta}$，利用极大似然估计法求得的参数估计值称为极大似然估计值。

极大似然估计值的问题，就是求似然函数 $L(\theta)$的最大值问题，这个问题可以通过解下面的方程

$$\frac{\mathrm{d}L}{\mathrm{d}\theta}=0 \tag{3-2}$$

来解决。因为 $\ln L$ 是 L 的增函数，所以 $\ln L$ 与 L 在 θ 的同一值处取得最大值。因此，也可将方程(3-2)换成下面的方程

$$\frac{\mathrm{d}\ln L}{\mathrm{d}\theta}=0 \tag{3-3}$$

解方程(3-2)或式(3-3)得到的 $\hat{\theta}$ 就是参数 θ 的最大似然估计值，而从后一方程求解往往比较方便，式(3-3)称为对数似然方程。

【例 3-2】 设有甲、乙两个布袋，甲袋中有 99 个白球和 1 个黑球，乙袋中有 1 个白球和 99 个黑球。由于某种原因已不能识别哪一个是甲袋，哪一个是乙袋。你能否用统计的方法识别出来？

解：下面对这一问题进行数学描述与分析。

不妨设变量 X 表示袋中的白球数，则 $X\sim\begin{pmatrix}1 & 99\\ p & 1-p\end{pmatrix}$，$p$ 是未知的分布参数，其取值依赖于变量 X 代表的是甲袋中的白球数还是乙袋中的白球数。显然，变量 X 代表的是甲袋中的白球数与 $p=99/100$ 是等价的，变量 X 的代表是乙袋中的白球数与 $p=1/100$ 是等价的。

可以通过抽样(任取一袋，从该袋中任取一球，观察其颜色)的方法来确定 $p=99/100$ 还是 $p=1/100$。

设事件 A 表示“取出的一袋为甲袋”，事件 B 表示“从袋子中取出的是白球”，则

$$P(A)=0.5,\quad P(B\mid A)=99/100,\quad P(B\mid \overline{A})=1/100$$

假定取出的是白球。在已知取出的是白球的条件下，判断该球来自甲袋还是乙袋的问题，可由贝叶斯公式，通过比较概率 $P(B|A)$和 $P(\overline{A}|B)$的大小来做出判断。由于在一

次试验中大概率事件容易发生，因此，若 $P(A|B)>P(\overline{A}|B)$，则该球来自甲袋；如果 $P(A|B)<P(\overline{A}|B)$，收该球来自乙袋。

因为

$$P(A \mid B)=\frac{P(AB)}{P(B)}=\frac{P(A)P(B \mid A)}{P(A)P(B \mid A)+P(\overline{A})P(B \mid \overline{A})},$$

$$P(\overline{A} \mid B)=\frac{P(\overline{A}B)}{P(B)}=\frac{P(\overline{A})P(B \mid \overline{A})}{P(A)P(B \mid A)+P(\overline{A})P(B \mid \overline{A})}$$

这两个式子的分母相同，分子中 $P(A)=P(\overline{A})$，故其大小取决于 $P(B|A)$ 和 $P(B|\overline{A})$ 的大小，而 $P(B|A)$ 和 $P(B|\overline{A})$ 的取值恰好等于变量 X 的分布参数 p 的两个可能的取值。这说明参数的取值同逆概率 $P(B|A)$ 和 $P(B|\overline{A})$ 之间的大小是相互决定的，即 $p=99/100$ 等价于 $P(A|B)>P(\overline{A}|B)$，$p=1/100$ 等价于 $P(A|B)>P(\overline{A}|B)$。

通过计算可知，$P(A|B)>P(\overline{A}|B)$，因此 $p=99/100$，即现在取出的这一袋是甲袋。

概括这里的思想方法，就可以得到极大似然估计法的数学原理——大概率原理：大概率事件在一次试验中容易发生。或者说，在一次试验中已经发生的事件具有较大的概率，而变量的分布参数有助于关于该变量的大概率事件的发生。

【例 3-3】 设 $X\sim N(\mu,\sigma^2)$，求 μ 和 σ^2 的极大似然估计。

解：正态样本 $N(\mu,\sigma^2)$ 的密度函数是 $\frac{1}{\sqrt{2\pi}\sigma}\cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$，则似然函数为

$$L(\mu,\sigma^2)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\cdot e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}=\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}}\cdot e^{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}$$

将其取对数，并令关于 μ、σ^2 的一阶导数为零，则得

$$\frac{\partial \ln L(\mu,\sigma^2)}{\partial \mu}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0$$

$$\frac{\partial \ln L(\mu,\sigma^2)}{\partial \sigma^2}=-\frac{n}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2=0$$

解此关于 μ，σ^2 的方程组，得驻点

$$\mu=\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\quad \sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

又可求得对数似然函数的二阶导函数矩阵是非正定矩阵，因此驻点处即为似然函数的极大值点处，并将 μ 的样本表达式代入 σ^2 的驻点表达式，得 μ 与 σ^2 的极大似然估计为

$$\hat{\mu}=\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\quad \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

表 3-1 所示函数的返回值为数据向量 x 的参数最大似然估计值，以及置信度为 $(1-a)\times100\%$ 的置信区间。a 的默认值为 0.05，即置信度为 95%。

表 3-1 参数估计函数

函数名	调用格式	函数说明
binofit	phat = binofit(x,n)	二项分布的概率最大似然估计
	[phat,pci] = binofit(x,n)	置信度为 95%的参数估计和置信区间
	[phat,pci] = binofit(x,n,alpha)	返回水平 a 的参数估计和置信区间

续表

函数名	调 用 格 式	函 数 说 明
poissfit	lambdshat = poissfit(data)	泊松分布的参数的最大似然估计
	[lambdahat,lambdaci] = poissfit(data)	置信度为95%的参数估计和置信区间
	[lambdahat,lambdaci] = poissfit(data,alpha)	返回水平 a 的参数估计和置信区间
normfit	[muhat,sigmahat] = normfit(data)	正态分布的最大似然估计，置信度为95%
	[muhat,sigmahat,muci,sigmaci] = normfit(data,alpha)	返回水平 a 的期望、方差值和置信区间
betafit	phat = betafit(data)	返回 β 分布参数 a 和 b 的最大似然估计
	[phat,pci] = betafit(data,alpha)	返回最大似然估计值和水平 a 的置信区间
unifit	[ahat,bhat] = unifit(data)	均匀分布参数的最大似然估计
	[ahat,bhat,ACI,BCI] = unifit(data)	置信度为95%的参数估计和置信区间
	[ahat,bhat,ACI,BCI] = unifit(data,alpha)	返回水平 a 的参数估计和置信区间
expfit	muhat = expfit(data)	指数分布参数的最大似然估计
	[muhat,muci] = expfit(data)	置信度为95%的参数估计和置信区间
	[muhat,muci] = expfit(data,alpha)	返回水平 a 的参数估计和置信区间
gamfit	phat = gamfit(data)	r 分布参数的最大似然估计
	[phat,pci] = gamfit(data)	置信度为95%的参数估计和置信区间
	[phat,pci] = gamfit(data,alpha)	返回最大似然估计值和水平 a 的置信区间
wblfit	parmhat = wblfit(data)	韦伯分布参数的最大似然估计
	[parmhat,parmci] = wblfit(data)	置信度为95%的参数估计和置信区间
	[parmhat,parmci] = wblfit(data,alpha)	返回水平 a 的参数估计及其区间估计
mle	[phat=mle('dist', data)	分布函数名为 dist 的最大似然估计
	[phat, pci]=mle('dist', data)	置信度为95%的参数估计和置信区间
	[phat, pci]=mle('dist', data, alpha)	返回水平 a 的最大似然估计值和置信区间
	[phat, pci]=mle('dist', data, alpha, p1)	仅用于二项分布，p1 为试验总次数

【例 3-4】 随机产生 100 个服从正态分布 $N(2,0.5^2)$ 的样本数据 X，并用这些数据估计总体 $N(\mu,\sigma^2)$ 中的参数 μ、σ，求出参数的最大似然估计值和置信水平为 99% 的置信区间。

分析：随机产生的 100 个数据可视为总体中抽出容量为 100 的样本，样本的观测值就是这具体的 100 个数据，可用命令 normfit(X, alpha)求出参数 μ、σ 的估计。

其 MATLAB 代码编程如下。

```
>> clear all;
X = normrnd(2,0.5,100,1);              % 产生 100 个样本数据
[muhat,sigmahat,muci,sigmaci] = normfit(X,0.01)
```

运行程序，输出如下。

```
muhat =
    2.0240
sigmahat =
    0.4343
muci =
    1.9099
```

```
    2.1380
sigmaci =
    0.3665
    0.5298
```

说明：参数 μ、σ 的估计最大似然值分别为 2.0240、0.4343，参数 μ，σ 的置信水平为 99%的置信区间分别为[1.9099，2.1380]、[0.3665，0.5298]。这一估计结果和总体 $N(\mu,\sigma^2)$中的参数真实数值 $\mu=2$，$\sigma=0.5$ 是非常接近的。

可以概括出求极大似然估计值的一般步骤如下。

(1) 明确变量的分布律和密度函数；

(2) 写出似然函数 $L(\theta)$；

(3) 求似然函数 $L(\theta)$的最大值点，得 $\hat{\theta}_{MLE}$；

(4) 应用问题中，将样本数据代入 $\hat{\theta}_{MLE}$，求出具体的估计值。

值得注意的是，求解对数似然方程组是在假定其可导并且导数变号的基础上，如果不满足这一条件，需针对似然函数 $L(\theta_1,\theta_2,\cdots,\theta_k)$的单调性，利用极大似然估计的基本原理直接进行 $L(\theta_1,\theta_2,\cdots,\theta_k)$的最大值问题的讨论。

极大似然估计量有一个简单而有用的性质：设 θ 的函数 $g=g(\theta)$是 Θ 上的实值函数，且有唯一反函数。如果 $\hat{\theta}$ 是 θ 的极大似然估计量，则 $g(\hat{\theta})$也是 $g(\theta)$的极大似然估计量。这个性质称为极大似然估计的不变性。根据这一性质可以使一些复杂结构的参数的极大似然估计问题简单化。

极大似然估计法是在变量分布类型已知的情况下使用的一种参数估计法。一般地，用极大似然法所得的估计的性质比用矩估计法所得的要好，故通常多用极大似然法。

在 MATLAB 中，提供了 mle 函数进行极大似然估计，函数的调用格式如下。

```
phat = mle(data)
[phat,pci] = mle(data)
[…] = mle(data,'distribution',dist)
[…] = mle(data,…,name1,val1,name2,val2,…)
[…] = mle(data,'pdf',pdf,'cdf',cdf,'start',start,…)
[…] = mle(data,'logpdf',logpdf,'logsf',logsf,'start',start,…)
[…] = mle(data,'nloglf',nloglf,'start',start,…)
[phat, pci] = mle(data, 'distribution', dist, 'alpha', a, 'ntrials', n)
```

其中，输出参数 phat 是指定分布的参数的极大似然估计值(多参数时为行向量)，pci 是参数的区间估计的置信上限和下限(与参数对应的二维列向量，可以缺省)。输入参数 data 是样本数据向量(不可缺省)。引用参数'distribution'及其取值 dist 设置变量的分布类型(应用中 dist 要用具体的分布名称字符串替换并用单引号引起)，二者要成对出现(可以同时缺省，缺省时分布类型默认为正态分布)。引用参数'alpha'及其取值 a 设置区间估计的显著性水平，二者成对出现(可以同时缺省，缺省时默认为 0.05，即置信水平为 0.95)。引用参数'ntrials'及其取值 n 仅在分布类型为二项分布时引用(对于其他分布可以缺省)，设置二项分布中试验的次数。

dist 的取值包括：Beta，Bernoulli，Binomial，Discrete Uniform，Exponential，Extreme Value，Gamma，Geometric，Lognormal，Negative Binomial，Normal，Poisson，Rayleigh，

Uniform,Weibull。

【例 3-5】 引用常数的测定值服从均值为 μ、标准差为 σ 的正态分布。某人在实验中使用金球测定引力常数,6 次测定观察值为：6.683,6.681,6.676,6.678,6.679,6.672。试用极大似然估计法对未知参数 μ 和 σ 做出估计。

其 MATLAB 代码编程如下：

```
>> clear all;
x = [6.683,6.681,6.676,6.678,6.679,6.672];
phat = mle(x,'distribution','norm','alpha',0.05)
```

运行程序,输出如下：

```
phat =
    6.6782    0.0035
```

即金球测定的 μ 估计值为 6.6782,σ 的估计值为 0.0035。其实,此例计算中 mle 函数的调用可以简化为 p=mle(x)。

3.1.3 顺序统计量

1. 统计量法的定义

设$\varsigma_1,\varsigma_2,\cdots,\varsigma_n$是总体$\varsigma$的样本,将其按大小排列为$\varsigma_1^* \leqslant \varsigma_2^* \leqslant \cdots \leqslant \varsigma_n^*$,则称$\varsigma_1^*,\varsigma_2^*,\cdots,\varsigma_n^*$为顺序统计量。

明显地,ς_1^*与ς_n^*分别为样本的最小值与最大值。称$\bar{\varsigma}=\begin{cases}\varsigma_{k+1}^*, & n=2k+1\\ \dfrac{\varsigma_k^*+\varsigma_{k+1}^*}{2}, & n=2k\end{cases}$为样本中位数。

样本中位数的取值规则为：将样本值 $x_1,x_2,\cdots,x_n$ 从小至大排成 $x_1^* \leqslant x_2^* \leqslant \cdots \leqslant x_n^*$,当 $n=2k+1$ 时,$\bar{\varsigma}$取居中的数据 x_{k+1}^*为其观测值；当 $n=2k$ 时,$\bar{\varsigma}$取居中的两个数据的平均值$\dfrac{\varsigma_k^*+\varsigma_{k+1}^*}{2}$为其观测值,中位数$\bar{\varsigma}$带来了总体$\varsigma$取值的平均数的信息,因此用$\bar{\varsigma}$估计总体$\varsigma$的数学期望是合适的。

用样本中位数$\bar{\varsigma}$估计总体ς的数学期望的方法称为数学期望的顺序统计量估计法。

顺序统计量估计法的优点是计算简便,且不易受个别异常数据的影响。如果一组样本值某一数据异常(如过小或过大),则这个异常数据可能是总体ς的随机性造成的,也可能是受外来干扰造成的(如工作人员粗心,记录错误),当原因属于后者,用样本平均值$\bar{x}$估计 $E(x)$显然受到影响,但用样本中位数$\bar{\varsigma}$估计 $E(x)$时,由于一个(甚至几个)异常的数据不易改变中位数取值,所以估计值不易受到影响。即称 $R=\varsigma_n^*-\varsigma_1^*$ 为样本极差。

由于样本极差带来总体样本取值离散程度的信息,因此可以用 R 作为对总体ς的标准差 σ 的估计(R 与 σ 量纲相同)。用样本极差对总体ς的标准差做估计的方法称为极差估计法。

极差估计法的优点是计算简便，但不如用 S 可靠，n 越大两者可靠的程度差别越大，这时一般不用极差估计。

2. 顺序统计量法主要适用范围

顺序统计量法主要适用于正态总体，当总体不是正态分布，但是连续型且分布密度对称时，也常用样本中位数来估计总体的期望。

3.1.4 最小二乘法

在科学实验数据处理中，往往要根据一组给定的实验数据$(x_i,y_i)(i=0,1,\cdots,m)$，求出自变量 x 与因变量 y 的函数关系 $y=s(x,a_0,\cdots,a_n)(n<m)$，这是 a_i 为待定参数，由于观测数据总有误差，且待定参数 a_i 的数量比给定数据点的数量少(即 $n<m$)，因此它不同于插值问题。这类问题不要求 $y=s(x)=s(x,a_0,\cdots,a_n)$通过点$(x_i,y_i)(i=0,1,\cdots,m)$，而只要求在给定点 x_i 上的误差 $\delta_i=s(x_i)-y_i(i=0,1,\cdots,m)$的平方和 $\sum_{i=0}^{m}\delta_i^2$ 最小。当 $S(x)\in\mathrm{span}\{\varphi_0,\varphi_1,\cdots,\varphi_n\}$时，即

$$s(x)=a_0\varphi_0(x)+a_1\varphi_1(x)+\cdots+a_n\varphi_n(x) \tag{3-4}$$

这里 $\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x)\in C[a,b]$是线性无关的函数族，假定在$[a,b]$上给出一组数据$\{(x_i,y_i),i=0,1,\cdots,m\}$，$a\leqslant x_i\leqslant b$ 以及对应的一组权$\{\rho_i\}_0^m$，这里 $\rho_i>0$ 为权系数，要求 $s(x)=\mathrm{span}\{\varphi_0,\varphi_1,\cdots,\varphi_n\}$使 $I(a_0,a_1,\cdots,a_n)$最小，其中

$$I(a_0,a_1,\cdots,a_n)=\sum_{i=0}^{m}\rho_i[s(x_i)-y_i]^2 \tag{3-5}$$

这就是最小二乘逼近，得到的拟合曲线为 $y=s(x)$，这种方法称为曲线拟合的最小二乘法。

式(3-5)中 $I(a_0,a_1,\cdots,a_n)$实际上是关于 $a_0,a_1,\cdots,a_n$ 的多元函数，求 I 的最小值就是求多元函数 I 的极值，由极值必要条件，可得

$$\frac{\partial I}{\partial a_k}=2\sum_{i=0}^{m}\rho_i[a_0\varphi_0(x_i)+a_1\varphi_1(x_i)+\cdots+a_n\varphi_n(x_i)-y_i]\varphi_k(x_i),\quad k=0,1,\cdots,n \tag{3-6}$$

根据内积定义引入相应带权内积记号

$$\begin{cases}(\varphi_j,\varphi_k)=\sum_{i=0}^{m}\rho_i\varphi_j(x_i)\varphi_k(x_i)\\(y,\varphi_k)=\sum_{i=0}^{m}\rho_i y_i\varphi_k(x_i)\end{cases} \tag{3-7}$$

则(3-6)可改写为

$$(\varphi_0,\varphi_k)a_0+(\varphi_1,\varphi_k)a_1+\cdots+(\varphi_n,\varphi_k)a_n=(y,\varphi_k),\quad k=0,1,\cdots,n$$

这是关于参数 $a_0,a_1,\cdots,a_n$ 的线性方程组，用矩阵表示为

$$\begin{bmatrix}(\varphi_0,\varphi_0) & (\varphi_0,\varphi_1) & \cdots & (\varphi_0,\varphi_n)\\(\varphi_1,\varphi_0) & (\varphi_1,\varphi_1) & \cdots & (\varphi_1,\varphi_n)\\\vdots & \vdots & & \vdots\\(\varphi_n,\varphi_0) & (\varphi_n,\varphi_1) & \cdots & (\varphi_n,\varphi_n)\end{bmatrix}\begin{bmatrix}a_0\\a_1\\\vdots\\a_n\end{bmatrix}=\begin{bmatrix}(y,\varphi_0)\\(y,\varphi_1)\\\vdots\\(y,\varphi_n)\end{bmatrix} \tag{3-8}$$

式(3-8)称为法方程。当$\{\varphi_j(x);j=0,1,\cdots,n\}$线性无关,且在点集$X=\{x_0,x_1,\cdots,x_m\}(m\geqslant n)$上至多只有$n$个不同零点,则称$\varphi_0,\varphi_1,\cdots,\varphi_n$在$X$上满足Haar条件,此时式(3-8)的解存在唯一。记式(3-8)的解为

$$a_k=a_k^*,\quad k=0,1,\cdots,n$$

从而得到最小二乘拟合曲线

$$y=s^*(x)=a_0^*\varphi_0(x)+a_1^*\varphi_1(x)+\cdots+a_n^*\varphi_n(x) \tag{3-9}$$

可以证明对$\forall(a_0,a_1,\cdots,a_n)^{\mathrm{T}}\in R^{n+1}$,有

$$I(a_0^*,a_1^*,\cdots,a_n^*)\leqslant I(a_0,a_1,\cdots,a_n)$$

故式(3-9)得到的$s^*(x)$即为所求的最小二乘解。它的平方误差为

$$\|\delta\|_2^2=\sum_{i=0}^{m}\rho_i[s^*(x_i)-y_i]^2 \tag{3-10}$$

均方误差为

$$\|\delta\|_2=\sqrt{\sum_{i=0}^{m}\rho_i[s^*(x_i)-y_i]^2}$$

在最小二乘逼近中,若取$\varphi_k(x)=x^k(k=0,1,\cdots,n)$,则$s(x)\in\mathrm{span}\{1,x,\cdots,x^n\}$,表示为

$$s(x)=a_0+a_1x+\cdots+a_nx^n \tag{3-11}$$

此时关于系数$a_0,a_1,\cdots,a_n$的方程(3-8)是病态方程,通常当$n\geqslant 3$时都不直接取$\varphi_k(x)=x^k$作为基。

3.1.5 点估计的优良性准则

样本统计量,如样本均值,样本标准差S,样本成数如何用于对相应总体参数μ、σ和p的点估计值。直观上,这些样本统计量对相应总体参数的点估计值是很有吸引力的。然而,在用一个样本统计量作为点估计量之前,统计学应检验说明这些样本统计量是否具有某些与好的点估计量相联系的性质。本节讨论点估计量的性质:无偏性、有效性和一致性。

1. 无偏性

设$\hat{\theta}=\hat{\theta}(X_1,X_2,\cdots,X_n)$的数学期望等于$\theta$,即

$$E(\hat{\theta})=\theta$$

则称$\hat{\theta}$是参数θ的无偏估计量;如果样本观测值为$x_1,x_2,\cdots,x_n$,则称$\hat{\theta}(x_1,x_2,\cdots,x_n)$为参数$\theta$的无偏估计值。

在科学技术中$E(\hat{\theta})-\theta$称为以$\hat{\theta}$作为θ的估计的系统误差,无偏估计的实际意义就是无系统误差。

无偏性是对估计量的一个最重要、最常见的要求,它的实际意义在于,当这个估计量经常使用时,在多次重复的平均意义下,给出了接近于真值θ的估计,在此应当指出,同一个参数θ的无偏估计量不是唯一的。例如,有$E(X_i)=\mu$,这表明任一样本的每一分量X_i

($i=1,2,\cdots,n$)都是总体均值 μ 的无偏估计量，在参数 θ 的许多无偏估计中，当然是以对 θ 的平均偏差较小者为好，即较好的估计量应当有尽可能小的方差。因此，便有了第二个评选标准。

下面列举出关于无偏性的几个重要结论。

(1) 无论变量 X 服从何种分布，样本的 k 阶原点矩 $A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k(i=1,2,\cdots,n)$ 是变量 X 的 k 阶原点矩 $E(X^k)$ 的无偏估计。自然，$\overline{X}$ 是 $E(X)$ 的无偏估计。

(2) 无论变量 X 服从何种分布，样本(修正)方差 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$ 是变量 X 的方差 σ^2 的无偏估计。

(3) 样本方差(二阶中心矩)B_2 不是变量方差 σ^2 的无偏估计，但是 $\lim\limits_{n\to\infty}E(B_2)=\sigma^2$，所以 B_2 是 σ^2 的渐近无偏估计。

(4) 样本标准差 $S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2}$ 不是变量 X 的标准差 σ 的无偏估计。但是，在变量的正态性假设下，可将样本标准差修正为 $\hat{\sigma}_S=C_nS$，$\hat{\sigma}_S$ 是 σ 的无偏估计，其中 $C_n=\sqrt{\frac{n-1}{2}}\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}$ 称为正态标准差的无偏系数。由于 $\lim\limits_{n\to\infty}C_n=1$，所以 S 是 σ 的渐近无偏估计。

无偏性准则是对估计量的一个基本要求。无偏性估计的统计意义是指估计量不产生系统性的偏差。例如，用样本均值 $\overline{X}$ 作为变量均值 μ 的估计时，由于 $\overline{X}$ 是随机变量，因此在一次估计中 μ 的实现值与其真值之间存在偏差 $\overline{X}-\mu$。这种偏差是随机的，虽无法说明一次估计所产生的偏差，但是对同一统计问题大量重复使用 $\overline{X}$ 估计 μ 时，实际产生的偏差 $\overline{X}-\mu$ 随机地在 0 的周围波动，不会产生系统的 $\overline{X}$ 偏大(小)于 μ 的情况。

渐近无偏差是指估计量存在系统性的偏差，但是这种系统性偏差随着样本容量的增加而趋向于消失。

【例 3-6】 设总体 $X\sim X^2(n)$，$X_1,X_2,\cdots,X_{20}$ 为来自总体的简单随机样本，想要估计总体均值 μ(注意 n 未知)，比较以下三个点估计量的好坏：$\hat{\mu}_1=101X_1-100X_2$，$\hat{\mu}_2=\frac{1}{2}(X_{10}+X_{11})$，$\hat{\mu}_3=\overline{X}$。

实例中给出了利用 MSE 评价点估计量的随机模拟方法。由于 $X^2(n)$ 的总体均值为 n，因此可以先取定一个固定值，例如 $n=\mu_0=5$，然后在这个参数已知且固定的总体中抽取容量为 20 的样本，分别用样本值依照三种方法分别计算估计值，看看哪种方法误差大，哪种方法误差小。一次估计的比较一般不能说明问题，正如低手射击也可能命中 10 环，高手射击也可能命中 9 环，如果连续射击 10 000 次，比较总环数，多者一定是高手。同理，如果抽取容量为 20 的样本 $N=10\,000$ 次，分别计算：$\mathrm{MSE}(\hat{\mu}_i)\approx\frac{1}{N}\sum_{k=1}^{n}[\hat{\mu}_i(k)-\mu_0]^2$，值小者为好。

其 MATLAB 代码编程如下：

```
>> clear all;
N = 10000;
m = 5;n = 20;
mse1 = 0;mse2 = 0;mse3 = 0;
for k = 1:N
    x = chi2rnd(m,1,n);
    m1 = 101 * x(1) - 100 * x(2);
    m2 = median(x);
    m3 = mean(x);
    mse1 = mse1 + (m1 - m)^2;
    mse2 = mse2 + (m2 - m)^2;
    mse3 = mse3 + (m3 - m)^2;
end
mse1 = mse1/N
mse2 = mse2/N
mse3 = mse2/N
```

运行程序，输出如下：

```
mse1 =
   1.9716e + 05
mse2 =
    0.9909
mse3 =
   9.9087e - 05
```

2. 有效性

一个未知参数 θ 的估计量 $\hat{\theta}$ 仅有无偏性是不够的。因为一方面，无偏性仅反映估计量在参数真值周围波动，而没有反映出“集中”的程度；另一方面，一个参数的无偏估计量可能不止一个，对于数学期望 μ，样本均值 $\overline{X}$ 是它的无偏估计量，样本的第一个观测值 X_1 也是它的无偏估计量(因 $E(X_1)=\mu$)，那么哪个更好呢？仅有无偏性一个标准是不能确定的。一个自然的想法是进一步比较它们的方差，方差越小，表示 $\hat{\theta}$ 越集中在 θ 的附近，从这个意义上讲方差越小的无偏估计量越好。

设 $\hat{\theta}_1$，$\hat{\theta}_2$ 都是 θ 的无偏估计量，若 $D(\hat{\theta}_1)<D(\hat{\theta}_2)$，则称 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 有效。

【例 3-7】 试比较总体期望 μ 的 2 个无偏估计量 $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ 及 $\hat{\alpha}=X_1$ 的有效性。

设总体方差为 σ^2，则

$$D(\overline{X})=\frac{1}{n}\sigma^2,\quad D(\hat{\alpha})=D(X_1)=\sigma^2$$

显然 $D(\overline{X})<D(\hat{\alpha})$，故 $\overline{X}$ 较 $\hat{\alpha}$ 有效。

可以证明，无偏估计 $\hat{\theta}$ 的方差 $D(\hat{\theta})$有一个非零的下界，即最小方差 $D_0(\hat{\theta})$，它等于

$$D(\hat{\theta})\geqslant\frac{1}{nE\left(\left(\frac{\partial\ln f(X,\theta)}{\partial\theta}\right)^2\right)}=D_0(\hat{\theta})$$

其中 $f(X,\theta)$ 为总体分布的概率密度函数。如果 $\ln f(X,\theta)$ 存在关于 θ 的二阶偏导数，则可证明

$$D(\hat{\theta}) \geqslant \frac{1}{-nE\left(\left(\frac{\partial^2 \ln f(X,\theta)}{\partial \theta^2}\right)^2\right)} = D_0(\hat{\theta})$$

若 $\hat{\theta}$ 满足 $E(\hat{\theta})=\theta, D(\hat{\theta})=D_0(\theta)$，则称 $\hat{\theta}$ 为 θ 的方差一致最小无偏估计量，称为 UMVUE 估计。

可以证明 $\overline{X}$ 和 S^2 是总体均值 μ 和方差 σ^2 的 UMVUE 估计，$\overline{X}$ 和 S^2 使用率很高，一是由于 μ 和 σ^2 是很重要且常用的总体参数，二是 μ 和 σ^2 有很好的统计性质。

3. 一致性

无偏性准则和均方误差准则是在样本容量 n 固定的情形下讨论估计量优劣的。设变量 $X \sim F(x)$，$F_n(x)$ 为样本的经验分布函数，由 Гливенко 定理，得

$$P\left\{\lim_{n\to\infty} \sup_{-\infty<x<+\infty} |\hat{F}_n(x) - F(x)| = 0\right\} = 1$$

当样本容量 n 趋向于无穷时，样本的经验分布函数以概率 1 一致收敛于变量的分布函数。也即是说，当样本容量 n 趋向于无穷时，样本中包含的关于变量分布的信息不断增加，以致充分到可以将变量分布刻画到任意精确的程度。因此，有理由要求，一个“好的”估计量，当样本容量 n 趋向于无穷时，在一定的数学意义下收敛于被估参数。

设 $\hat{\theta}(X_1, X_2, \cdots, X_n)$ 为参数 θ 的估计量，如果对任意的 $\varepsilon>0$，有

$$\lim_{n\to\infty} P\{|\hat{\theta}-\theta| \geqslant \varepsilon\} = 0$$

而且这对 θ 的一切可能取的值都成立，则称 $\hat{\theta}$ 是参数 θ 的一个一致性估计。

一致性准则是对一个估计量最基本的要求。它说明，随着样本容量的增大，一个“好的”估计量 $\hat{\theta}$ 应该越来越靠近参数 θ 的真值，使绝对偏差 $|\hat{\theta}-\theta|$ 较大的概率越来越小。如果一个估计量没有一致性，那么，不论样本取多大，我们也不可能把未知参数估计到预定的精度。这种估计量显然是不可取的。

下面给出一致性估计的几个重要结论。

(1) 一致性估计具有不变性。即当 $\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_k$ 分别是 $\theta_1, \theta_2, \cdots, \theta_k$ 的一致性估计时，如果 $g(\theta_1, \theta_2, \cdots, \theta_k)$ 为连续函数，则 $g(\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_k)$ 是 $g(\theta_1, \theta_2, \cdots, \theta_k)$ 的一致性估计。

(2) 样本的 k 阶原点矩 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ 是变量 X 的 k 阶原点矩 $E(X^k)$ 的一致性估计，因此样本均值 $\overline{X}$ 是变量均值 μ 的一致性估计。

(3) 样本的二阶中心矩 $B_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$ 是变量 X 的方差 σ^2 的一致性估计。

(4) 样本方差 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$ 是变量的方差 σ^2 的一致性估计，样本标准差 $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2}$ 是变量的标准差 σ 的一致性估计。

(5) 事件发生的频率是其概率的一致性估计。

(6) 极大似然估计量往往具有一致性。

3.2 区间估计

上一节讨论了参数点估计,它是用样本算得的一个值去估计未知参数,但是,点估计值仅仅是未知参数的一个近似值,它没有反映出这个近似值的误差范围,使用起来把握不大,区间估计(interval estimation)正好弥补了点估计的这个缺陷。

3.2.1 区间估计简介

区间估计就是以一定的概率保证估计包含总体参数的一个值域,即根据样本指标和抽样平均误差推断总体指标的可能范围。它包括两部分内容:一是这一可能范围的大小;二是总体指标落在这个可能范围内的概率。区间估计既说清估计结果的准确程度,又同时表明这个估计结果的可靠程度,所以区间估计是比较科学的。

用样本指标来估计总体指标,要达到100%的准确而没有任何误差,几乎是不可能的,所以在估计总体指标时就必须同时考虑估计误差的大小。从人们的主观愿望上看,总是希望花较少的钱取得较好的效果,也就是说希望调查费用和调查误差越小越好。但是,在其他条件不变的情况下,缩小抽样误差就意味着增加调查费用,它们是一对矛盾。因此,在进行抽样调查时,应该根据研究目的和任务以及研究对象的标志变异程度,科学确定允许的误差范围。

区间估计必须同时具备3个要素。即具备估计值、抽样极限误差和概率保证程度3个基本要素。

抽样误差范围决定抽样估计的准确性,概率保证程度决定抽样估计的可靠性,二者密切联系,但同时又是一对矛盾,所以,对估计的精确度和可靠性的要求应慎重考虑。

3.2.2 区间估计的含义

区间估计就是根据样本来确定统计量 $\underline{\theta}(X_1,X_2,\cdots,X_n)$ 和 $\bar{\theta}(X_1,X_2,\cdots,X_n)$,使

$$P(\underline{\theta}(X_1,X_2,\cdots,X_n)<\theta<\bar{\theta}(X_1,X_2,\cdots,X_n))=1-\alpha \tag{3-12}$$

其中,$(\underline{\theta},\bar{\theta})$ 为 θ 的置信区间,$1-\alpha$ 称为此置信区间的置信度,$\underline{\theta}$ 和 $\bar{\theta}$ 分别称为置信下限和置信上限。

显然,置信区间是一个随机区间,式(3-12)的含义是:如果反复抽样多次(每次取样本容量都是 n),在每次取样下,对样本的观察值 $x_1,x_1,\cdots,x_n$,就得到一个区间 $\underline{\theta}(X_1,X_2,\cdots,X_n)$,$\bar{\theta}(X_1,X_2,\cdots,X_n)$,每个这样的区间要么包含 θ 的真值,要么不包含 θ 的真值,按伯努利大数定理,在这样多的区间中,大约有 $100(1-\alpha)\%$ 的区间包含未知参数 θ,而不包含 θ 的区间约占 $100\alpha\%$。例如,若 $\alpha=0.01$,反复抽样1000次,则得到的1000个区间中不包含 θ 真值的约仅有10个。通常 α 给得较小,这样式(3-12)的概率就较大。因此,置信区

间的长度的平均 $E(\bar{\theta}-\underline{\theta})$表达了区间估计的精确性；置信度 $1-\alpha$ 表达了区间估计的可靠性，它是区间估计的可靠概率，而显著性水平 α 表达了区间估计的不可靠概率。

置信度 $1-\alpha$ 一般要根据具体问题的要求来选定，并要注意：α 越小，$1-\alpha$ 越大，即区间$(\bar{\theta}-\underline{\theta})$包含 θ 真值的可信度越大，但区间也越长，亦即估计的精确度就越差；反之，提高估计的精确度则会增大误判风险 α，即$(\bar{\theta}-\underline{\theta})$不包含 θ 真值的概率会增大。从后面推出的置信区间公式可看出，如果其他条件不变，增大样本容量 n，可以缩短置信区间的长度，从而提高精度，但增大样本容量往往不现实。因此，通常是根据不同类型的问题，先确定一个较大的置信概率 $1-\alpha$，在这一前提下，寻找精度尽可能高的区间估计。如果对 $\alpha=0.05$，

$$P\left[-1.96<\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}<1.96\right]=0.95,\quad P\left[-1.75<\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}<2.33\right]=0.95$$

比较两个置信区间$\left(\overline{X}-\frac{\sigma}{\sqrt{n}}u_{0.025},\overline{X}+\frac{\sigma}{\sqrt{n}}u_{0.025}\right)$和$\left(\overline{X}-\frac{\sigma}{\sqrt{n}}u_{0.01},\overline{X}+\frac{\sigma}{\sqrt{n}}u_{0.04}\right)$，前者的区间长度 $2u_{0.025}\times\frac{\sigma}{\sqrt{n}}=3.92\times\frac{\sigma}{\sqrt{n}}$比后者的区间长度$(u_{0.04}+u_{0.01})\times\frac{\sigma}{\sqrt{n}}=4.08\times\frac{\sigma}{\sqrt{n}}$短，置信区间越短表示估计的精度越高。由经验知，当 n 固定时，在给定的 $1-\alpha$ 下，对称区间的长度最短。

3.2.3 区间估计的基本思想

对于给定值 $\alpha(0<\alpha<1)$为得到满足 $P(\bar{\theta}<\theta<\underline{\theta})=1-\alpha$ 的统计量 $\underline{\theta}(X_1,X_2,\cdots,X_n)$和 $\bar{\theta}(X_1,X_2,\cdots,X_n)$，将随机区间$(\underline{\theta},\bar{\theta})$包含 θ 的概率 $P(\bar{\theta}<\theta<\underline{\theta})=1-\alpha$，转化成某随机变量 $W(X_1,X_2,\cdots,X_n;\theta)$落在区间$(a,b)$上的概率

$$P(a<W(X_1,X_2,\cdots,X_n;\theta)<b)=1-\alpha$$

然后通过解不等式 $a<W(X_1,X_2,\cdots,X_n;\theta)<b$ 得到

$$\underline{\theta}(X_1,X_2,\cdots,X_n)<\theta<\bar{\theta}(X_1,X_2,\cdots,X_n)$$

为实现这个目的，我们所要找的函数 $W(X_1,X_2,\cdots,X_n;\theta)$必须满足两个条件：

(1) 仅是样本 $X_1,X_2,\cdots,X_n$ 和待估计参数θ 的函数，而不再含有其他未知参数；

(2) (a,b)必须是确定的。为此要求 $W(X_1,X_2,\cdots,X_n;\theta)$的分布已知。

3.2.4 区间估计的方法

在实际抽样调查中，区间估计根据给定的条件不同，有两种估计方法：

(1) 给定极限误差，要求对总体指标做出区间估计；

(2) 给定概率保证程度，要求对总体指标做出区间估计。

【例 3-8】 某企业对某批电子元件进行检验，随机抽取 100 只，测得平均耐用时间为 1000h，标准差为 50h，合格率为 94%，求：

(1) 以耐用时间的允许误差范围 $\Delta x=10\text{h}$,估计该批产品平均耐用时间的区间及其概率保证程度。

(2) 以合格率估计的误差范围不超过 2.45%,估计该批产品合格率的区间及其概率保证程度。

(3) 试以 95%的概率保证程度,对该批产品的平均耐用时间做出区间估计。

(4) 试以 95%的概率保证程度,对该批产品的合格率做出区间估计。

解:求(1)的计算步骤如下。

① 求样本指标。

$$\bar{x}=100\text{h},\quad \sigma=50\text{h}$$

$$\mu_x=\frac{\sigma}{\sqrt{n}}=\frac{50}{\sqrt{100}}=5\text{h}$$

② 根据给定的 $\Delta x=10\text{h}$,计算总体平均数的上、下限。

下限: $\bar{x}-\Delta x=1000-10=990\text{h}$

上限: $\bar{x}+\Delta x=1000+10=1010\text{h}$

③ 根据 $t=\frac{\Delta x}{\mu_x}=\frac{10}{5}=2$,由概率表得 $F(t)=95.45\%$,由计算结果,估计该批产品的平均耐用时间在 990～1010h 之间,有 95.45%的概率保证程序。

求(2)的计算步骤如下。

① 求样本指标。

$$p=94\%$$

$$\sigma_p^2=p(1-p)=0.94\times0.06=0.0564$$

$$\mu_p=\sqrt{\frac{p(1-p)}{n}}=\sqrt{\frac{0.0564}{100}}=2.38\%$$

② 根据给定的 $\Delta p=2.45\%$,求总体合格率的上、下限。

下限: $p-\Delta p=94\%-2.45\%=91.55\%$

上限: $p+\Delta p=94\%+2.45\%=96.45\%$

③ 根据 $t=\frac{\Delta p}{\mu_p}=\frac{2.45\%}{2.38\%}=1.03$,查概率表得 $F(t)=69.70\%$。

由以上计算结果,估计该批产品的合格率在 91.55%～96.45%之间,有 69.70%的概率保证程度。

求(3)的计算步骤如下。

① 求样本指标。

$$\bar{x}=1000(\text{h}),\quad \sigma=50\text{h}$$

$$\mu_x=\frac{\sigma}{\sqrt{n}}=\frac{50}{\sqrt{100}}=5\text{h}$$

② 根据给定的 $F(t)=95\%$,查概率表得 $t=1.96$。

③ 根据 $\Delta x=t\times\mu_x=1.96\times5=9.8$,计算总体平均耐用时间的上、下限。

下限: $\bar{x}-\Delta x=1000-9.8=990.2\text{h}$

上限: $\bar{x}+\Delta x=1000+9.8=1009.8\text{h}$

所以，以95%的概率保证程度估计该批产品的平均耐用时间在990.2～1009.8h之间。

求(4)的计算步骤如下。

① 求样本指标。

$$p = 94\%$$

$$\sigma_p^2 = p(1-p) = 0.94 \times 0.06 = 0.0564$$

$$\mu_p = \sqrt{\frac{p(1-p)}{n}} = 2.37\%$$

$$\Delta p = t \times \mu_p = 1.96 \times 2.37\% = 0.046$$

② 计算总体平均耐用时间的上、下限。

下限：　　$p-\Delta p=94\%-4.6\%=89.4\%$

上限：　　$p+\Delta p=94\%+4.6\%=98.6\%$

所以，以95%的概率保证程度估计该批产品的合格率在89.4%～98.6%之间。

1. 单正总体均值的置信区间

对正态总体均值 μ 的区间估计分为两种情形：方差 σ^2 已知和未知。

1) σ^2 为已知时，均值 μ 的置信区间

以样本均值 $\overline{X}$ 作为 μ 的一个点估计，由正态分布公式可知

$$U = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

由正态分布的分位点知

$$P(|U| < u_{\frac{\alpha}{2}}) = 1-\alpha$$

即

$$P\left[\left|\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right| < u_{\frac{\alpha}{2}}\right] = 1-\alpha$$

或

$$P\left(\overline{X}-\frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}} < \mu < \overline{X}+\frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}}\right) = 1-\alpha$$

故

$$\left(\overline{X}-\frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}}, \overline{X}+\frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}}\right) \tag{3-13}$$

为 μ 的置信度 $1-\alpha$ 的置信区间。

【例 3-9】 设1.1,2.2,3.3,4.4,5.5为来自正态总体 $N(\mu,2.3)^2$ 的简单随机样本，求 μ 的置信水平为95%的置信区间。

其MATLAB代码编程如下：

```
>> clear all;
x = [1.1 2.2 3.3 4.4 5.5];
n = length(x);
m = mean(x);
```

```
c = 2.3/sqrt(n);
d = c * norminv(0.975);
a1 = m - d;
b1 = m + d;
[a1,b1]
```

运行程序，输出如下：

```
ans =
    1.2840    5.3160
```

2）σ^2 为未知时，均值 μ 的置信区间

这时，自然会想到以样本标准差 S 代替总体均方差 σ，即知选取统计量

$$T=\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}}\sim t(n-1)$$

对给定的数 α，由

$$P(|T|<t_{\frac{\alpha}{2}}(n-1))=1-\alpha$$

查概率表得 $t_{\frac{\alpha}{2}}(n-1)$，解不等式得 $\overline{X}-\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)<\mu<\overline{X}+\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)$，即 μ 的 $1-\alpha$ 的置信区间为

$$\left(\overline{X}-\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1),\overline{X}+\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)\right)$$

简记为

$$\overline{X}\pm\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1) \tag{3-14}$$

在实际问题中，很难找到一种情况，其总体均值未知，但方差已知。通常情况下，均值和方差都要通过样本进行估计，因此式(3-14)比式(3-13)更实用。

【例 3-10】 数据同例 3-9，求 σ^2 未知，均值 μ 的置信区间。

其 MATLAB 代码编程如下：

```
>> clear all;
x = [1.1 2.2 3.3 4.4 5.5];
n = length(x);
m = mean(x);
S = std(x);
dd = S * tinv(0.975,4)/sqrt(n);
a2 = m - dd;
b2 = m + dd;
[a2,b2]
```

运行程序，输出如下：

```
ans =
    1.1404    5.4596
```

3）单正态方差的区间估计

设总体 $X\sim N(\mu,\sigma^2)$，$X_1,X_2,\cdots,X_n$ 是 X 的样本，求 σ^2 的 $1-\alpha$ 置信区间。由 χ^2 分

布知选取统计量

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

对给定的 α，取 χ^2 分布分位点 $\chi^2_{\frac{\alpha}{2}}(n)$ 和 $\chi^2_{1-\frac{\alpha}{2}}(n)$，使

$$\left(\chi^2_{1-\frac{\alpha}{2}}(n-1) < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\frac{\alpha}{2}}(n-1)\right) = 1-\alpha$$

从而得到 σ^2 的 $1-\alpha$ 置信区间为

$$\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}}, \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}}\right)$$

【例 3-11】 数据同例 3-9，求以下 σ^2 的置信区间。

其 MATLAB 代码编程如下：

```
>> clear all;
x = [1.1 2.2 3.3 4.4 5.5];
n = length(x);
c1 = chi2inv(0.025,4);
c2 = chi2inv(0.975,4);
T = (n - 1) * var(x);
a3 = T/c2;
b3 = T/c1;
[a3,b3]
```

运行程序，输出如下：

```
ans =
    1.0859   24.9784
```

4）两正态总体均值差的置信区间

当方差已知时，设 $X_1, X_2, \cdots, X_m \sim N(\mu_1, \sigma_1^2)$，$Y_1, Y_2, \cdots, Y_m \sim N(\mu_2, \sigma_2^2)$，两样本独立，此时 $\mu_1 - \mu_2$ 的置信区间为

$$\left(\overline{X} - \overline{Y} - u_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}, \overline{X} - \overline{Y} + u_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}\right)$$

在此已经知道 $u_{\frac{\alpha}{2}}$ 可用 norminv(0.975)求得。

当方差未知但相等时，此时 $\mu_1 - \mu_2$ 的置信区间为

$$\left(\overline{X} - \overline{Y} - t_{\frac{\alpha}{2}}C, \overline{X} - \overline{Y} + t_{\frac{\alpha}{2}}C\right)$$

其中，$C = \sqrt{\frac{1}{m} + \frac{1}{n}}\sqrt{\frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}}$，而 $t_{\frac{\alpha}{2}}$ 依照自由度 $m+n-2$ 计算。

5）两正态总体方差比的置信区间

查自由度为$(m-1, n-1)$的 F 分布临界值表使得，

$$P(c_1 < F < c_2) = 1-\alpha$$

则$\frac{\sigma_1^2}{\sigma_2^2}$的置信区间为$\left(\frac{\left(\frac{S_1^2}{S_2^2}\right)}{c_2}, \frac{\left(\frac{S_1^2}{S_2^2}\right)}{c_1}\right)$。

【例 3-12】 设两台车床加工同一零件，各加工 8 件，长度的误差如下。

A：−0.12　−0.80　−0.05　−0.04　−0.01　0.05　0.07　0.21

B：−1.50　−0.80　−0.40　−0.10　0.20　0.61　0.82　1.24

求方差比的置信区间。

其 MATLAB 代码编程如下：

```
>> clear all;
x = [ - 0.12 - 0.80 - 0.05 - 0.04 - 0.01 0.05 0.07 0.21];
y = [ - 1.50 - 0.80 - 0.40 - 0.10 0.20 0.61 0.82 1.24];
v1 = var(x);
v2 = var(y);
c1 = finv(0.025,7,7);
c2 = finv(0.975,7,7);
a4 = (v1/v2)/c1;
b4 = (v1/v2)/c1;
[a4,b4]
```

运行程序，输出如下：

```
ans =
    0.5720    0.5720
```

方差比小于 1 的概率至少达到 95%，说明车床 A 的精度明显高。

2. 单侧置信区间

上面的讨论中，对于未知参数 θ，给出两个统计量 $\underline{\theta}$ 和 $\bar{\theta}$，得到 θ 的置信区间为 $(\underline{\theta},\bar{\theta})$ 的形式。但在有些实际应用中，常常只关心参数的上限或下限。例如，对于设备、元件的寿命来说，我们只关心平均寿命 θ 至少是多少（θ 的“下限”）；与之相反，在考虑化学药品中杂质含量时，我们关心的却是平均杂质含量 θ' 最多是多少（θ' 的“上限”）。这就引出了单侧置信区间的概念。

对于给定值 $\alpha(0<\alpha<1)$，若由样本 $X_1,X_2,\cdots,X_n$ 确定的统计量 $\underline{\theta}(X_1,X_2,\cdots,X_n)$ 满足对任意 θ 有

$$P(\theta>\underline{\theta})=1-\alpha$$

则称随机区间 $(\underline{\theta}(X_1,X_2,\cdots,X_n),+\infty)$ 是 θ 的置信水平为 $1-\alpha$ 的下侧置信区间，称 $\underline{\theta}(X_1,X_2,\cdots,X_n)$ 是置信水平为 $1-\alpha$ 的置信下限。

又如果统计量 $\bar{\theta}(X_1,X_2,\cdots,X_n)$ 满足对任意 θ 有

$$P(\theta<\bar{\theta})=1-\alpha$$

则称随机区间 $(-\infty,\bar{\theta}(X_1,X_2,\cdots,X_n))$ 是 θ 的置信水平为 $1-\alpha$ 的上侧置信区间，称 $\bar{\theta}(X_1,X_2,\cdots,X_n)$ 是置信水平为 $1-\alpha$ 的单侧置信上限。

【例 3-13】 从一批灯泡中随机地抽取 5 只做寿命试验，其寿命如下（单位：h）

1050　1100　1120　1250　1280

已知这批灯泡寿命 $X\sim N(\mu,\sigma^2)$，求平均寿命 μ 的置信度为 95%的单侧置信下限

$$t=\frac{\overline{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$$

对于给定置信度 $1-\alpha$，有

$$P\left\{\frac{\overline{X}-\mu}{S/\sqrt{n}}<t_{\alpha}(n-1)\right\}=1-\alpha$$

即

$$P\left\{\mu>\overline{X}-t_{\alpha}(n-1)\frac{S}{\sqrt{n}}\right\}=1-\alpha$$

可得 μ 的置信度为 $1-\alpha$ 的单侧置信下限为

$$\overline{X}-t_{\alpha}(n-1)\frac{S}{\sqrt{n}}$$

由所得数据计算，有

$$\bar{x}=1160,\quad s=99.75,\quad n=5,\quad \alpha=0.05$$

查表得 $t_{0.05}(4)=2.14$，从而平均寿命 μ 的置信度为 95%的置信下限为

$$\bar{x}-t_{\alpha}(n-1)\frac{s}{\sqrt{n}}=1064.56$$

也就是说，该批灯泡的平均寿命至少在 1064.56h 以上，可靠程度为 95%。

其 MATLAB 代码编程如下：

```
>> clear all;
x = [1050,1100,1120,1250,1280];
N = length(x);
muEST = mean(x)
muLOWER = muEST - tinv(0.95,N - 1) * sqrt(var(x)/N)
```

上述指令的运行结果是：

```
muEST =
        1160
muLOWER =
  1.0649e + 003
```

计算结果表明，这批灯泡的平均寿命约为 1160h，以 95%的概率保证这批灯泡的平均寿命不低于 1065h。

3.2.5 区间估计函数

在 MATLAB 的函数工具箱中，也提供了相关函数用于实现区间估计，下面分别对这些函数给予介绍。

1. nlinfit 函数

在 MALTAB 中，提供了 nlinfit 函数用于求解高斯-牛顿法的非线性最小二乘数据拟合。函数的调用格式如下。

beta = nlinfit(X,y,fun,beta0)：返回在 fun 中描述的非线性函数的系数。fun 为用户提供的形如 $\hat{y}=f(\beta,x)$ 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,r,J,COVB,mse] = nlinfit(X,y,fun,beta0)：同时返回的 beta 为拟合系数，r 为残差，J 为 jacobi 矩阵，COVB 为评估的协方差矩阵，mse 为误差的方差。输入参数 beta0 为初始预测值。

[…] = nlinfit(X,y,fun,beta0,options)：指定控制参数后返回值。参数 options 包括 MaxIter、TolFun、TolX、Display、DerivStep 等。

当 X 为矩阵时，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@，则表示函数的句柄。

【例 3-14】 使用 nlinfit 函数求高斯-牛顿法的非线性最小二乘数据拟合。

其 MATLAB 代码编程如下：

```
>> clear all;
S = load('reaction');
X = S.reactants;
y = S.rate;
beta0 = S.beta;
% 利用 nlinfit 函数求非线性最小二乘数据拟合
beta = nlinfit(X,y,@hougen,beta0)
```

运行程序，输出如下：

```
beta =
    1.2526
    0.0628
    0.0400
    0.1124
    1.1914
```

2. nlparci 函数

在 MATLAB 中，提供了 nlparci 函数用于求解非线性模型的参数估计的置信区间。函数的调用格式为

ci = nlparci(beta,resid,'covar',sigma)：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，resid 为残差，sigma 为协方差矩阵系数。

ci = nlparci(beta,resid,'jacobian',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，resid 为残差，J 为 Jacobian 矩阵。

ci = nlparci(…,'alpha',alpha)：返回 100×(1−alpha)%的置信区间。

【例 3-15】 利用 nlparci 函数求求非线性模型 $y_j=\partial_1+\partial_2\exp(-\partial_3 x_j)+\varepsilon_j$ 的参数估计的置信区间。

```
>> clear all;
% 用函数句柄表示模型
mdl = @(a,x)(a(1) + a(2) * exp( - a(3) * x));
rng(9845,'twister')                    % 可重复性
a = [1;3;2];
```

```
x = exprnd(2,100,1);                    %指数分布
epsn = normrnd(0,0.1,100,1);            %正态分布
y = mdl(a,x) + epsn;
%数据拟合模型为随机的
a0 = [2;2;2];
[ahat,r,J,cov,mse] = nlinfit(x,y,mdl,a0);
ahat
%检查在95%的置信区间是否[1 3 2]是使用雅可比参数 nlparci
ci1 = nlparci(ahat,r,'Jacobian',J)
%使用的协方差参数
ci2 = nlparci(ahat,r,'covar',cov)
```

运行程序,输出如下:

```
ahat =
    1.0153
    3.0229
    2.1070
ci1 =
    0.9869   1.0438
    2.9401   3.1058
    1.9963   2.2177
ci2 =
    0.9869   1.0438
    2.9401   3.1058
    1.9963   2.2177
```

3. nlintool 函数

在 MATLAB 中,提供了 nlintool 函数用于求解非线性拟合并显示交互图形。函数的调用格式如下。

nlintool(X,y,fun,beta0):返回数据(X,y)的非线性曲线的预测图形,它用两条红色曲线预测全局置信区间。beta0 为参数的初始预测值,默认值为 0.05,即置信度为 95%。

nlintool(X,y,fun,beta0,alpha):将置信度设置为(1−alpha)×100%。

nlintool(X,y,fun,beta0,alpha,'xname','yname'):给 X 和 y 的变量分别赋予变量名 xname 和 yname。

【例 3-16】 使用 nlintool 函数求非线性拟合并显示交互图形。

其 MATLAB 代码编程如下:

```
>> clear all;
>> load reaction
nlintool(reactants,rate,@hougen,beta,0.01,xn,yn)
```

运行程序,效果如图 3-1 所示。

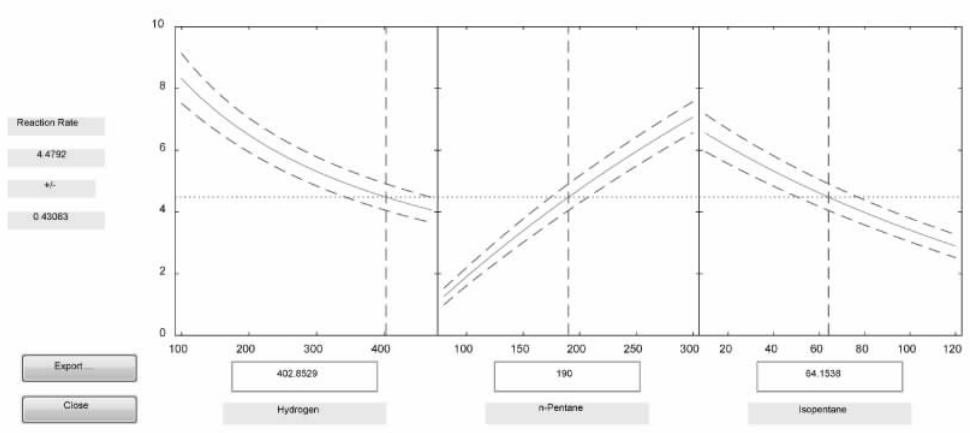

图 3-1　非线性拟合交互图形

4. nlpredci 函数

在 MATLAB 中，提供了 nlpredci 函数用于求解非线性模型置信区间预测。函数的调用格式如下。

[ypred,delta] = nlpredci(modelfun, x, beta, resid, 'jacobian', J)：返回预测值 ypred，fun 与前面相同，beta 为给出的适当参数，resid 为残差，J 为 Jacobi 矩阵，x 为非线性函数中的独立变量的矩阵值。返回的 delta 为非线性最小二乘法估计的置信区间长度的一半。当 resid 长度超过 beta 的长度，并且 J 的列满秩时，置信区间的计算才是有效的，[ypred－delta，ypred＋delta]为置信度，是 95%的不同步置信区间。

[ypred,delta] = nlpredci(modelfun,x,beta,resid,'covar',sigma)：参数 sigma 为协方差矩阵系数。

[…] = nlpredci(…,param1,val1,param2,val2,…)：设置多个参数的名称及其对应的值，参数包括：'alpha'、'mse'、'predopt'和'simopt'。

【例 3-17】 使用 nlpredci 函数求非线性最小二乘预测置信区间。

其 MATLAB 代码编程如下：

```
>> clear all;
S = load('reaction');
X = S.reactants;
y = S.rate;
beta0 = S.beta;
[beta,R,J] = nlinfit(X,y,@hougen,beta0);
[ypred,delta] = nlpredci(@hougen,mean(X),beta,R,'Jacobian',J)
```

运行程序，输出如下：

```
ypred =
    5.4622
```

```
delta =
    0.1921
```

3.3 参数估计实例

下面通过一个实例来演示怎样利用参数估计实现工程实际领域中的应用。

【例 3-18】 分别使用金球和铂球测定引力常数(单位：$10^{-11}\mathrm{m}^3\cdot\mathrm{kg}^{-1}\cdot\mathrm{s}^{-2}$)。

(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 $N(\mu,\sigma^2)$,试就(1)、(2)两种情况分别求 μ 的置信水平为 0.9 的置信区间,并求 σ 的置信水平为 0.9 的置信区间。

其 MATLAB 代码编程如下：

```
>> clear all;
data1 = [6.683 6.681 6.676 6.678 6.679 6.672];
alpha = 0.1;
[muhat1,sigmahat1,muci1,sigmaci1] = normfit(data1,alpha)
[phat1,pci1] = mle(data1,'distribution','normal','alpha',alpha)
data2 = [6.661 6.661 6.667 6.667 6.664];
[muhat2,sigmahat2,muci2,sigmaci2] = normfit(data2,alpha)
[phat2,pci2] = mle(data1,'distribution','normal','alpha',alpha)
```

运行程序,输出如下：

```
muhat1 =
    6.6782
sigmahat1 =
    0.0039
muci1 =
    6.6750
    6.6813
sigmaci1 =
    0.0026
    0.0081
phat1 =
    6.6782    0.0035
pci1 =
    6.6750    0.0026
    6.6813    0.0081
muhat2 =
    6.6640
sigmahat2 =
    0.0030
muci2 =
    6.6611
```

```
    6.6669
sigmaci2 =
    0.0019
    0.0071
phat2 =
    6.6782    0.0035
pci2 =
    6.6750    0.0026
    6.6813    0.0081
```

3.4 核密度估计

核密度估计(Kernel Density Estimation)是在概率论中用来估计未知的密度函数,属于非参数检验方法之一,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。

3.4.1 核密度估计的概述

对于一组关于 X 和 Y 观测数据$\{(x_i,y_i)\}_{i=1}^n$,假设它们存在关系 $y_i=m(x_i)+\varepsilon_i$,通常我们的目的在于估计 $m(x)$的形式。在样本数量有限的情况下,我们无法准确估计 $m(x)$的形式。这时,可以采用非参数方法,在非参数方法中,并不假定也不固定 $m(x)$的形式,仅假设 $m(x)$满足一定的光滑性,函数在每一点的值都由数据决定。显然,由于随机扰动的影响数据有很大的波动,极不光滑,因此要去除干扰使图形光滑。

最简单最直接的方法就是取多点平均,也就是每一点 $m(x)$的值都由离 x 最近的多个数据点所对应的 y 值的平均值得到。显然,如果用来平均的点越多,所得的曲线越光滑。当然,如果用 n 个数据点来平均,则 $m(x)$为常数,这时它最光滑,但失去了大量的信息,拟合的残差也很大。所以说,这就存在了一个平衡的问题,也就是说,要决定每个数据点在估计 $m(x)$的值时要起到的作用问题。直观上,和 x 点越近的数据对决定 $m(x)$的值所起作用越大,这就需要加权平均。因此,如何选择权函数来光滑及光滑到何种程度即是我们这里所关心的核心问题。

3.4.2 核密度估计的形式

对于数据 $x_1,x_2,\cdots,x_n$,核密度估计的形式如下。

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

这是一个加权平均,而核函数(kernal function)$K()$为一个权函数,核函数的形状和值域控制着用来估计 $f(x)$在点 x 的值时所用数据点的个数和利用的程度,直观来看,核密度估计的好坏依赖于核函数和带宽 h 的选取。通常考虑的核函数关于原点对称且其

积分为1，下面四个函数为最为常用的权函数：

Uniform：

$$\frac{1}{2}I(|t|\leqslant 1)$$

Epanechikov：

$$\frac{3}{4}I(1-t^2)I(|t|<1)$$

Quartic：

$$\frac{15}{16}(1-t^2)I(|t|<1)$$

Gaussian：

$$\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}t^2}$$

对于均匀核函数，$K\left(\frac{x-x_i}{h}\right)=\frac{1}{2}I\left(\left|\frac{x-x_i}{h}\right|\leqslant 1\right)$用作密度函数，则只有$\frac{x-x_i}{h}$的绝对值小于1(或者说离$x$的距离小于带宽$h$的点)才用来估计$f(x)$的值，不过所有起作用的数据的权重都相同。

对于高斯函数，由$f_h'(x)$的表达式可看出，如果x_i离x越近，$\frac{x-x_i}{h}$越接近于零，这时密度值$\phi\left(\frac{x-x_i}{h}\right)$越大，因为正态密度的值域为整个实轴，所以所有的数据都用来估计$f_h'(x)$的值，只不过离x点越近的点对估计的影响越大，当h很小的时候，只有特别接近x的点才起较大作用，随着h增大，则远一些的点的作用也随之增加。

如果使用形如Epanechikov和Quartic核函数，不但有截断(即离x的距离大于带宽h的点则不起作用)，并且起作用的数据它们的权重也随着与x的距离增大而变小。一般说来，核函数的选取对和核估计的好坏的影响远小于带宽h的选取。

3.4.3 带宽的选取

带宽值的选择对估计量$f_h'(x)$的影响很大，如果h太小，那么密度估计偏向于把概率密度分配得太局限于观测数据附近，致使估计密度函数有很多错误的峰值，如果h太大，那么密度估计就把概率密度贡献散得太开，这样会光滑掉f的一些重要特征。

所以，要想判断带宽的好坏，必须了解如何评价密度估计量$f_h'(x)$的性质。通常使用积分均方误差MSE(h)，作为判断密度估计量好坏的准则。

$$\mathrm{MSE}(h)=\mathrm{AMISE}(h)+o\left(\frac{1}{nh}+h^4\right)$$

其中，

$$\mathrm{AMISE}(h)=\frac{\int K^2(x)\mathrm{d}x}{nh}+\frac{h^2\sigma^2\int[f''(x)]^2\mathrm{d}x}{4}$$

称作渐进均方积分误差。要最小化 AMISE(h),必须把 h 设在某个中间值,这样可以避免 $f_h'(x)$有过大的偏差或过大的方差。关于 h 最小化 AMISE(h)表明最好是精确地平衡 AMISE(h)中偏差项和方差项的阶数,显然最优的带宽是

$$h = \left(\frac{\int K^2(x)\mathrm{d}x}{n\sigma^2 \int [f''(x)]^2 \mathrm{d}x} \right)^{\frac{1}{5}} \tag{3-15}$$

以下是几种常用的选择方法。

1. 拇指法

为简便起见,定义 $R(g) = \int g^2(z)\mathrm{d}z$,针对最小化 AMISE 得到的最优带宽中含有未知量 $R(f'')$,Silverman 提出一种初等的方法——Rule of Thumb(拇指法则,即根据经验的方法)。

把 f 用方差和估计方差相匹配的正态密度替换,这就等于用$\frac{R(\phi')}{\sigma^5}$估计 $R(f'')$,其中 ϕ 为标准正态密度函数,如果取 K 为高斯密度核函数,而 σ 使用样本方差 $\hat{\sigma}$,Silverman 拇指法则得到 $h = \left(\frac{4}{3n}\right)^{\frac{1}{5}}\hat{\sigma}$。

2. Plug-in 法

该方法即代入法,其考虑在最优带宽中使用某适当的估计$\hat{R}(f'')$来代替$R(f'')$,在众多的方法中,最简单且最常用的即是 Sheather and Jones 在 1991 年所提出的$\hat{R}(f'') = R(\hat{f}'')$,而$\hat{f}''$的基于核的估计量为

$$\hat{f}''(x) = \frac{\partial^2}{\partial x^2}\left\{ \frac{1}{nh_0}\sum_{i=1}^{n} L\left(\frac{x - x_i}{h}\right)\right\} = \frac{1}{h^3 n}\sum_{i=1}^{n} L''\left(\frac{x - x_i}{h}\right)$$

其中 h_0 为带宽,L 为用来估计 f''的核函数。在对其平方并对 x 积分后即可得到 $R(\hat{f}'')$。估计 f 的最优带宽和估计 f''或$R(f'')$的最优带宽是不同的。根据理论上以及经验上的考虑,Sheather and Jones 建议用简单的拇指法则计算带宽 h_0,该带宽用来估计 $R(f'')$,最后通过式(3-15)来计算带宽 h。

3.4.4 核密度估计的 MATLAB 实现

在 MATLAB 工具箱中,也提供了 ksdensity 函数用于实现核密度估计的相关函数,下面给予介绍。

在 MATLAB 统计工具箱中提供了 ksdensity 函数用于求核密度估计。其调用格式如下。

[f,xi] = ksdensity(x):求样本观测值向量 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 相应的核密度估计值向量。这里所用的核

函数为 Gaussian 核函数，所用的窗宽是样本容量的函数。

f = ksdensity(x,xi)：根据样本观测值向量 x 计算 xi 处的核密度估计值 f，xi 和 f 是等长的向量。

ksdensity(…)：不返回任何输出，此时在当前坐标系中绘制出核密度函数图。

ksdensity(ax,…)：不返回任何输出，此时在句柄值 ax 对应的坐标系中绘制出核密度函数图。

[f,xi,u] = ksdensity(…)：同时返回窗宽 u。

[…] = ksdensity(…,'Name',value)：通过可选的成对出现的参数名及参数值来控制核密度估计。可用的参数名及参数值如表 3-2 所示。

表 3-2 ksdensity 函数支持的参数名及对应的参数值

参数名	参数值	说　明
'censoring'	与 x 等长逻辑向量	指定哪些项是截尾观测，默认是没有截尾
'kernel'	'normal'	指定用 Gaussian(高斯或正态)核函数，为默认情况
	'box'	指定用 Uniform 核函数
	'triangle'	指定用 Triangle 核函数
	'epanechnikov'	指定用 Epanechnikov 核函数
	函数句柄或函数名，如 @normpdf 或'normpdf'	自定义核函数
'npoints'	正整数	指定 xi 中包含的等间隔点的个数，默认值为 100
'support'	'unbounded'	指定密度函数的支撑集为全体实数集，是默认情况
	'positive'	指定核密度函数的支撑集为正实数集
	包含两个元素的向量	指定密度函数的支撑集为上下限
'weights'	与 x 等长的向量	指定 x 中元素的权重
'width'	正实数	指定窗宽，默认值是由式(3-15)得到的最佳窗宽。取较小的窗宽，能反映较多的细节
'function'	'pdf'	指定对密度函数进行估计
	'cdf'	指定对累积分布函数进行估计
	'icdf'	指定对逆概率分布函数进行估计
	'survivor'	指定对生存函数进行估计
	'cumhazard'	指定对累积危险函数进行估计

【例 3-19】 利用 ksdensity 函数对给定的随机数据进行核密度估计。

其 MATLAB 代码编程如下：

```
>> clear all;
rng default
x = [randn(30,1); 5 + randn(30,1)];
[f,xi] = ksdensity(x);
figure
plot(xi,f);
```

运行程序,效果如图 3-2 所示。

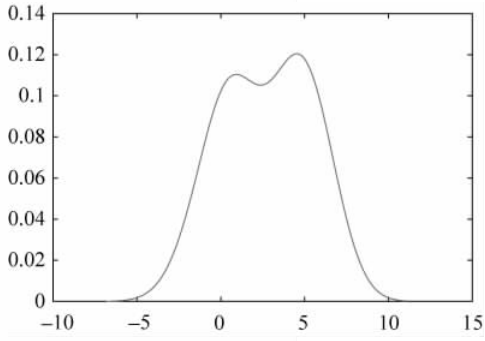

图 3-2　核密度估计曲线图

【例 3-20】 核密度估计的案例分析。

其 MATLAB 代码编程如下:

```
>> % 对 MATLAB 自带的数据绘制其直方图
clear all;
cars = load('carsmall','MPG','Origin');
MPG = cars.MPG;
hist(MPG)                                  % 效果如图 3-3 所示
xlabel('样本');ylabel('直方图');
set(get(gca,'Children'),'FaceColor',[.8 .8 1])
% 绘制不同的窗宽核密度估计曲线,效果如图 3-4 所示
[f,x,u] = ksdensity(MPG);
plot(x,f)
title('MPG 的核密度估计')
hold on
[f,x] = ksdensity(MPG,'width',u/3);
plot(x,f,':r');
[f,x] = ksdensity(MPG,'width',u * 3);
plot(x,f,'--g');
legend('默认窗宽','默认 1/3 窗宽','默认 3 倍窗宽');
xlabel('x');ylabel('f');
hold off
% 设置不同核密度估计参数,绘制相应曲线,效果如图 3-5 所示
hname = {'normal' 'epanechnikov' 'box' 'triangle'};
colors = {'r' 'b' 'g' 'm'};
for j = 1:4
    [f,x] = ksdensity(MPG,'kernel',hname{j});
    plot(x,f,colors{j});
    hold on;
end
legend(hname{:});
xlabel('x');ylabel('f');
hold off
```

```
% 比较密度估计,显示了从不同的原产地国家的汽车省油分布曲线,效果如图 3-6 所示
Origin = cellstr(cars.Origin);
I = strcmp('USA',Origin);
J = strcmp('Japan',Origin);
K = ~(I|J);
MPG_USA = MPG(I);
MPG_Japan = MPG(J);
MPG_Europe = MPG(K);
[fI,xI] = ksdensity(MPG_USA);
plot(xI,fI,':b')
hold on
[fJ,xJ] = ksdensity(MPG_Japan);
plot(xJ,fJ,'r')
[fK,xK] = ksdensity(MPG_Europe);
plot(xK,fK,'--g')
legend('USA','Japan','Europe')
xlabel('样本');ylabel('核密度估计')
hold off
```

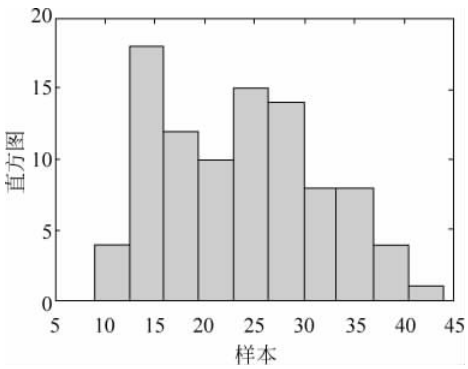

图 3-3　数据频率直方图

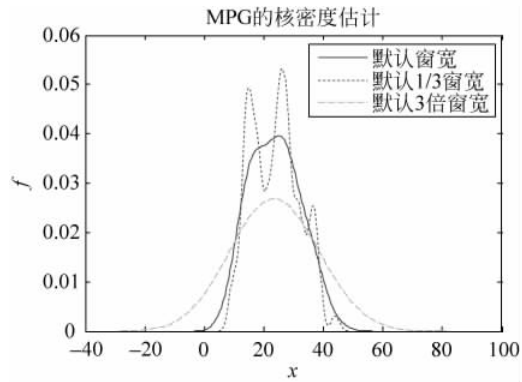

图 3-4　不同窗宽下的核密度估计曲线

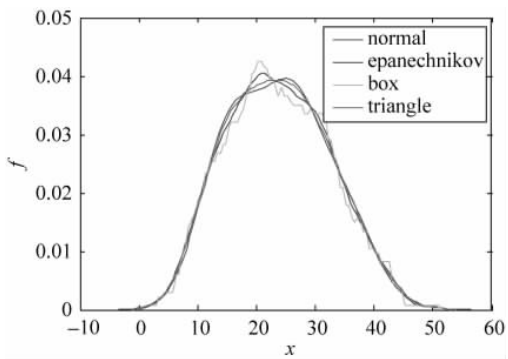

图 3-5　不同参数设置核密度估计曲线

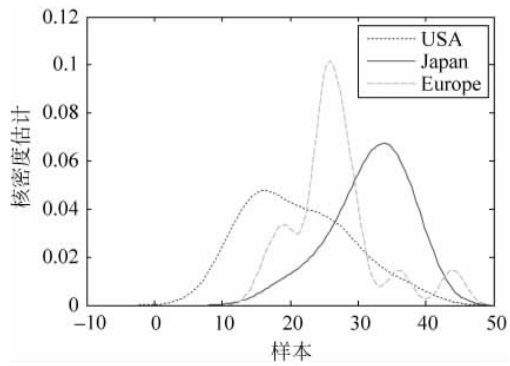

图 3-6　各个国家汽车省油核密度估计曲线图

3.5 统计作图

用图形表达样本数据的统计特征具有生动、直观的特点，MATLAB 提供了多种常用的统计图形函数来完成统计图绘制。

3.5.1 直方图

直方图又称质量分布图，它是表示资料变化情况的一种主要工具。用直方图可以解析出资料的规则性，比较直观地看出产品质量特性的分布状态，对于资料分布状况一目了然，便于判断其总体质量分布情况。在制作直方图时，牵涉统计学的概念，首先要对资料进行分组，因此如何合理分组是其中的关键问题。按组距相等的原则进行的两个关键数位是分组数和组距，是一种几何形图表，它是根据从生产过程中收集来的质量数据分布情况，画成以组距为底边、以频数为高度的一系列连接起来的直方型矩形图。

在 MATLAB 中，提供了 hist 函数用于实现绘制直方图。函数的调用格式如下。

hist(x)：表示把矩阵 x 中的数据等距地划分为 10 个区间进行统计，并将每一区间内的数据个数作为返回值矢量的元素，最后画出 10 个柱形，如果 x 是矩阵，则将矩阵 x 的每一列进行统计。

hist(x,nbins)：表示 nbins 是一个常量且指定了统计的区间个数。

hist(x,xbins)：表示 x 是要统计的数据，nbins 为一个矢量，矢量 xbins 的长度指定了统计的区间数，并以该矢量的各元素为中心进行统计。

hist(ax,__)：表示在 ax 指定的坐标系中画出直方图。

counts = hist(__)：表示只返回数据的频数。

[counts,centers] = hist(__)：表示返回矢量 counts 和 centers，分别表示频数和各个区间的位置。

【例 3-21】 利用函数 hist 绘制 randn 概率分布图。

其 MATLAB 代码编程如下：

```
>> clear all;
x = randn(1000,1);
subplot(3,1,1)
xbins1 = -4:4;
hist(x,xbins1)
subplot(3,1,2)
xbins2 = -2:2;
hist(x,xbins2)
subplot(3,1,3)
xbins3 = [-4 -2.5 0 0.5 1 3];
hist(x,xbins3)
```

运行程序，效果如图 3-7 所示。

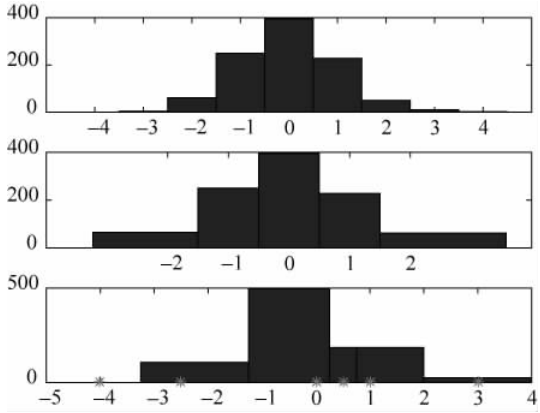

图 3-7　随机数据的直方图

3.5.2　频数表

在观察值个数较多时，为了解一组同质观察值的分布规律和便于指标的计算，可编制频数分布表，简称频数表。频数表是统计描述中经常使用的基本工具之一。

1. 频数分布的特征

由频数表可看出频数分布的两个重要特征：集中趋势(Central Tendency)和离散程度(Dispersion)。身高有高有矮，但多数人身高集中在中间部分组段，以中等身高居多，此为集中趋势；由中等身高到较矮或较高的频数分布逐渐减少，反映了离散程度。对于数值变量资料，可从集中趋势和离散程度两个侧面去分析其规律性。

2. 频数分布的类型

频数分布有对称分布和偏态分布之分。对称分布是指多数频数集中在中央位置，两端的频数分布大致对称。偏态分布是指频数分布不对称，集中位置偏向一侧，若集中位置偏向数值小的一侧，称为正偏态分布；集中位置偏向数值大的一侧，称为负偏态分布，如冠心病、大多数恶性肿瘤等慢性病患者的年龄分布为负偏态分布。临床上正偏态分布资料较多见。不同的分布类型应选用不同的统计分析方法。

3. 频数表的用途

频数表可以揭示资料分布类型和分布特征，以便选取适当的统计方法；便于进一步计算指标和统计处理；便于发现某些特大或特小的可疑值。

4. MATLAB 实现

在 MATLAB 中，提供了 tabulate 函数用于绘制频数表。函数的调用格式如下。

tbl = tabulate(x)：表示对矢量 x 中的数据绘制频数表，返回值 tbl 的第 1 列是矢

量 x 中的唯一值，第 2 列是每一个值出现的次数，第 3 列是每一个值出现的百分比例，如果 x 是一个数值型数组，则 tbl 是一个数值型矩阵，如果 x 的每一个元素都是非负整数，则 tbl 包含 0 到不包含在 x 中的从 1 到 max(x)的整数；如果 x 是一个分类变量、字符数组或字符串单元数组，则 tbl 是一个单元数组。

tabulate(x)：表示不返回频数表。

例如，在命令窗口中输入：

```
>> clear all;
tbl = tabulate([1 2 4 4 3 4])
```

运行程序，输出如下：

```
tbl =
    1.0000    1.0000   16.6667
    2.0000    1.0000   16.6667
    3.0000    1.0000   16.6667
    4.0000    3.0000   50.0000
```

3.5.3 箱形图

箱形图可以比较清晰地表示数据的分布特征，MATLAB 提供了 boxplot 函数来绘制箱形图，它由 5 个部分组成：

(1) 箱形上、下横线为样本的 25%和 75%分位数，箱形顶部和底部的差值为内四分位极值。

(2) 箱形中间的横线为样本的中值，如果该横线没在箱形中央，则说明存在偏度。

(3) 箱形向上或向下延伸的直线称为“触须”，如果没有异常值，样本的最大值为上触须的顶部，样本最小值为下触须的底部。默认情况下，距离箱形顶部或底部大于 1.5 倍同四分极值的值为异常值。

(4) 图中顶部的加号表示该处数据为一异常值，该值的异常可能是输入错误、测量失误或系统误差引起的。

(5) 箱形两侧的 V 形槽口对应于样本中值的置信区间。默认情况下，箱形图没有 V 形槽口。

boxplot 函数的调用格式如下。

boxplot(X)：对 X 中的每列数据绘制一个箱形图。

boxplot(X,notch)：当 notch=1，得到一个有凹口的盒子图；notch=0，得到一个矩形箱形图。

boxplot(X,notch,'sym')：'sym'为标记符号，缺省符号为“+”。

boxplot(X,nocth,'sym',vert,whis)：参数 vert 控制箱形图水平放置还是垂直放置。当 vert=0 时，箱形图水平放置；当 vert=1 时(缺省)，箱形图垂直放置；whis 定义虚线的长度，为内四分位间距(IQR)的函数(缺省情况为 15 * IQR)。如果 whis=0，则 box 图用'sym'规定的标记显示“箱子”外所有的数据。

【例 3-22】 根据参数的设置不同，绘制对应的样本的盒子图。

其 MATLAB 代码编程如下：

```
>> clear all;
% 产生正态分布的样本
% 样本长度
N = 1024;
x1 = normrnd(5,1,N,1);
x2 = normrnd(6,1,N,1);
x = [x1 x2];
% 参数
figure(1);
sym1 = '* ';
notch1 = 1;                              % 凹口
boxplot(x,notch1,sym1);
figure(2);
notch2 = 0;                              % 矩形
boxplot(x,notch2);
figure(3);
vert = 0;                                % 水平
boxplot(x,notch1,'+ ',vert);
```

运行程序，效果如图 3-8～图 3-10 所示。

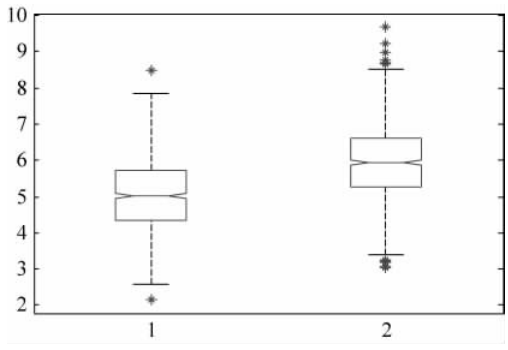

图 3-8　垂直、带凹口的盒子图

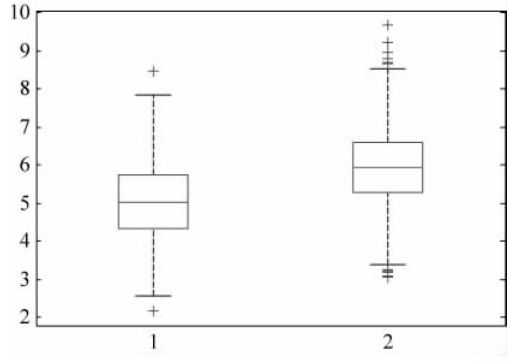

图 3-9　垂直、矩形的盒子图

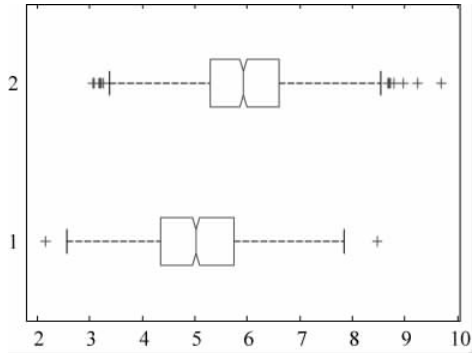

图 3-10　水平、带凹口的盒子图

3.5.4 经验累加分布图

称函数

$$F_n(x)=\begin{cases}0, & x\leqslant x_{(1)}\\ \sum\limits_{i=1}^{i} f_k, & x_{(i)}\leqslant x<x_{(i+1)},\quad i=1,2,\cdots,l-1\\ 1, & x\geqslant x_{(l)}\end{cases}$$

为样本分布函数(或经验分布函数)。经验分布函数图是阶梯状图,反映了样本观测数据的分布情况。

在MATLAB统计工具箱中提供了cdfplot函数用于绘制样本经验分布函数。可以把经验分布函数图和某种理论分布函数图叠放在一起,以对它们之间的区别。函数的调用格式如下。

cdfplot(X):表示绘制由矢量X指定的数据经验累加分布函数图,经验累加函数的定义是在x点处的值定义为X中小于等于x的数的比例。

h = cdfplot(X):表示绘制统计图的同时返回一个指向该曲线的一个句柄h。

[h,stats] = cdfplot(X):除了返回句柄外还返回一个结构体stats,该结构体包含域:min最小值、max最大值、mean样本平均值、median样本中值(50%的位置),以及std样本标准方差。

【例3-23】 在同一图中绘制经验分布函数及理论正态分布函数图。

其MATLAB代码编程如下:

```
>> clear all;
rng default;                              % 设置重复性
y = evrnd(0,3,100,1);
[h,stats] = cdfplot(y)
hold on
x = -20:0.1:10;
f = evcdf(x,0,3);
plot(x,f,'m:')
legend('经验分布曲线','理论上分布曲线','Location','NW')
```

运行程序,输出如下,效果如图3-11所示。

```
h =
  Line (具有属性):
              Color: [0 0.4470 0.7410]
          LineStyle: '-'
          LineWidth: 0.5000
             Marker: 'none'
         MarkerSize: 6
    MarkerFaceColor: 'none'
              XData: [1x202 double]
              YData: [1x202 double]
              ZData: [1x0 double]
```

```
显示所有属性
stats =
       min: -10.5349
       max: 4.4659
      mean: -2.0350
    median: -1.5294
       std: 3.7186
```

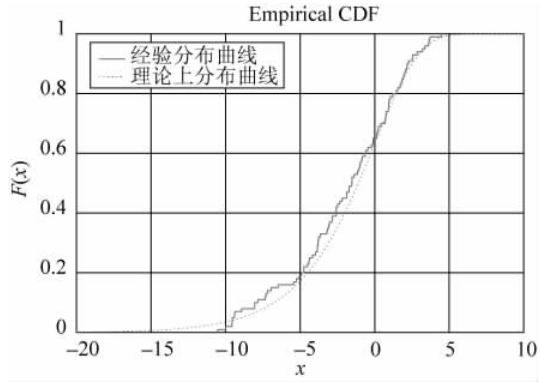

图 3-11　经验累积分布函数图

3.5.5　误差条图

误差条图通常用于统计或科学数据，显示潜在的误差或相对于系列中每个数据标志的不确定程度。误差条图可以用标准差（平均偏差）或标准误差来表示，它们的区别如下。

（1）概念不同。标准差是离均差平方和平均后的方根，标准误差是标准误差定义为各测量值误差的平方和的平均值的平方根；

（2）用途不同。标准差与均数结合估计参考值范围，计算变异系数，计算标准误差等。标准误差用于估计参数的可信区间，进行假设检验等；

（3）它们与样本含量的关系不同。当样本含量 n 足够大时，标准差趋向稳定；而标准误等随 n 的增大而减小，甚至趋于 0。

误差条形图类型的序列具有三个 Y 值。虽然可以手动将这些值分配给每个点，但在大多数情况下，是从其他序列中的数据来计算这些值。Y 值的顺序十分重要，因为值数组中的每个位置都表示误差条形图上的一个值。在 MATLAB 统计工具箱中提供了 errorbar 函数用于绘制误差条图。函数调用格式如下。

errorbar(Y,E)：表示绘制 Y，以及对 Y 的每个元素绘制误差条，误差条的上半部分和下半部分都是长为 E(i)的对称条。

errorbar(X,Y,E)：X、Y 与 E 必须有相同的大小，如果 X 与 Y 是矢量，则误差条以(X(i),Y(i))为中心，上下各长为 E(i)的线段条，如果 X 与 Y 是矩阵，则误差条是以(X(i,j),Y(i,j))为中心，上下各长为 E(i,j)的线段条。

errorbar(X,Y,L,U)：表示由 L 和 U 指定误差条的上下界，在此 X、Y、L、U 必须有相同的长度，如果 X 是矢量，则在误差条是以(X(i),Y(i))为中心，下长为 L(i)上长为 U(i)的线段条，如果 X 是矩阵则误差条是以(X(i,j),Y(i,j))为中心，下长为 L(i,j)上长为 U(i,j)的线段条。

errorbar(…,LineSpec)：表示由字符串 LineSpec 指定的线颜色、线类型来绘制误差条图。

errorbar(ax,…)：指定当前指定的坐标轴 ax 上绘制误差条图。

h = errorbar(…)：表示返回误差条对象的一个句柄 h。

【例 3-24】 对所给定数据绘制其误差条图。

其 MATLAB 代码编程如下：

```
>> clear all;
X = 0:pi/10:pi;
Y = sin(X);
E = std(Y) * ones(size(X));
errorbar(X,Y,E)
xlabel('数据');ylabel('误差条图');
```

运行程序，效果如图 3-12 所示。

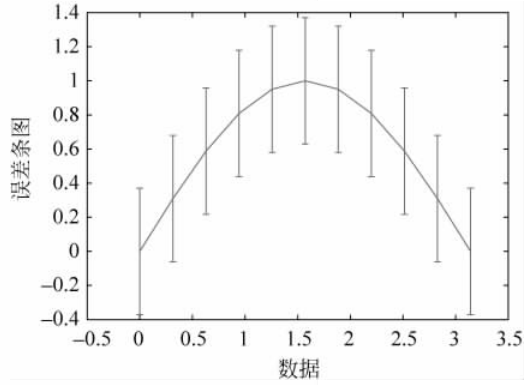

图 3-12 误差条图

3.5.6 交互等值线图

交互等值线图即是指既可用代码实现绘图也可以通过手工来实现绘图。在 MATLAB 统计工具箱中提供了 fsurfht 函数用于绘制交互等值线图。函数的调用格式如下。

fsurfht(fun,xlims,ylims)：表示生成由变量 fun 指定的交互式等值线图，x 轴的限制由 xlims=[xmin,xmax]来指定，y 轴的限制由 ylims=[ymin,ymax]来指定。

fsurfht(fun,xlims,ylims,p1,p2,p3,p4,p5)：表示允许函数 fun 提供 5 个选项参数，函数 fun 的前面两个变量分别为 x 轴变量和 y 轴变量。

图中有垂直参照线和水平参照线，两者的交点对应于当前点的 x 值和 y 值，可以通过拖拉这些带点的白色参考线来查看计算的 z 值(在图形上方)。另外，也可以通过在 x

轴和 y 轴的文本框中输入 z 值来得到指定的 z 值。

【例 3-25】 绘制由 gas. mat 文件提供数据的高斯似然函数图形。

其 MATLAB 代码编程如下：

```
function z = gauslike(mu,sigma,p1)
n = length(p1);
z = ones(size(mu));
for i = 1:n
z = z . * (normpdf(p1(i),mu,sigma));
end
```

调用 fsurfht 函数绘制高斯似然函数图形，代码为

```
>> clear al;
load gas
fsurfht('gauslike',[112 118],[3 5],price1)      % 求似然函数图形
mumax = mean(price1)                            % 求 price1 的均值
sigmamax = std(price1) * sqrt(19/20)            % 求 price1 的标准差
```

运行程序，输出如下，效果如图 3-13 所示。

```
mumax =
  115.1500
sigmamax =
    3.7719
```

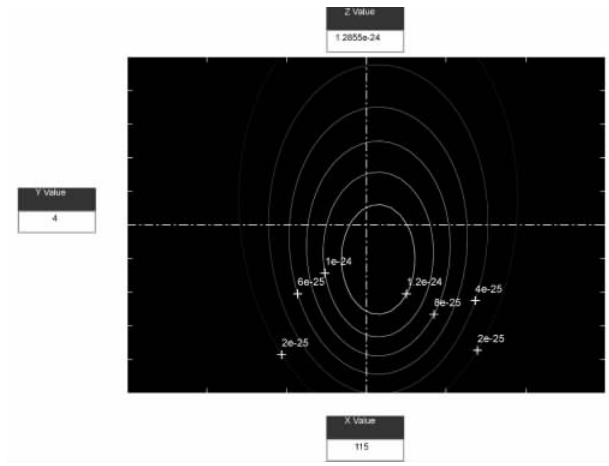

图 3-13 交互等值线图

3.5.7 散点图

散点图(Scatter Diagram)是指在回归分析中，数据点在直角坐标系平面上的分布图。其是表示因变量随自变量而变化的大致趋势，据此可以选择合适的函数对数据点进行拟合。在 MATLAB 中，提供了 gscatter 函数用于实现散点图的绘制。函数的调用格式

如下。

gscatter(x,y,group)：表示创建 x 和 y 的散点图，用 group 进行分组，其中 x 和 y 是矢量，且它们具有相同的大小，group 可以是矢量、字符串数组或字符串单元数组，具有相同 group 值的点分在一组，在图中用相同的标记和颜色来表示，另外，group 可以是包含一些分组变量（如[G1,G2,G3]）的单元数组。

gscatter(x,y,group,clr,sym,siz)：表示指定每组的颜色、标记类型和大小，默认是，clr='bgrcmyk'，sym 是可以被函数 plot 识别的字符串数组，其默认值为“.”，siz 是数组大小组成的矢量，其默认值由'DefaultLineMarkerSize'属性指定。

gscatter(x,y,group,clr,sym,siz,doleg)：表示由 doleg 指定是否在图中显示图例，当 deleg='on'时，表示显示图例，当 doleg='off'时表示不显示图例，默认值为'on'。

gscatter(x,y,group,clr,sym,siz,doleg,xnam,ynam)：表示由 xnam 和 ynam 指定 x 轴和 y 轴的名称，如果 x 和 y 的输入为简单的变量名，而且 xnam 和 ynam 被忽略，则函数 gscatter 用变量名标示坐标轴。

h=gscatter(…)：表示返回图中直线的句柄数组。

【例 3-26】 比较三种不同类型汽车的重量和里程数。

其 MATLAB 代码编程如下：

```
>> clear all;
 % 装载数据
load carsmall
 % 比较不同类型汽车的重量和里程数
gscatter(Weight,MPG,Model_Year,'','xos');
xlabel('重量');
ylabel('里程数');
```

运行程序，效果如图 3-14 所示。

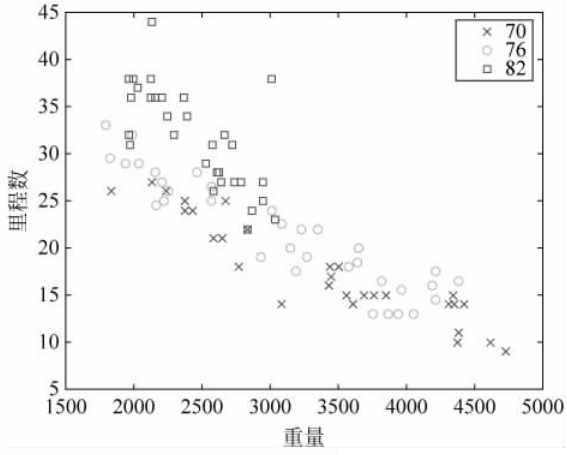

图 3-14　三种不同类型汽车的重量和里程数散度图

由图 3-14 可以看出，1982 年生产的汽车的里程数和重量明显区别于其他两种汽车。

3.5.8 最小二乘拟合线

最小二乘法(又称最小平方法)是一种数学优化技术。它通过最小化误差的平方和寻找数据的最佳函数匹配。利用最小二乘法可以简便地求得未知的数据,并使得这些求得的数据与实际数据之间误差的平方和为最小。最小二乘法还可用于曲线拟合。其他一些优化问题也可通过最小化熵或最大化熵用最小二乘法来表达。

在 MATLAB 中,提供了 lsline 函数用于添加最小二乘拟合线。函数的调用格式如下。

lsline:表示在当前轴中每一直线对象上添加最小二乘直线。

lsline(ax):在指定的坐标轴 ax 中添加最小二乘拟合线。

h=lsline(__):返回直线对象的句柄 h。

【例 3-27】 利用 lsline 函数绘制最小二乘拟合线。

其 MATLAB 代码编程如下:

```
>> clear all;
x = 1:10;
rng default;                                    %设置重复性
figure;
y1 = x + randn(1,10);
scatter(x,y1,25,'b','*')
hold on
y2 = 2*x + randn(1,10);
plot(x,y2,'mo')
y3 = 3*x + randn(1,10);
plot(x,y3,'rx:')
```

运行程序,在一个图形中得到 3 条拟合曲线,效果如图 3-15 所示。

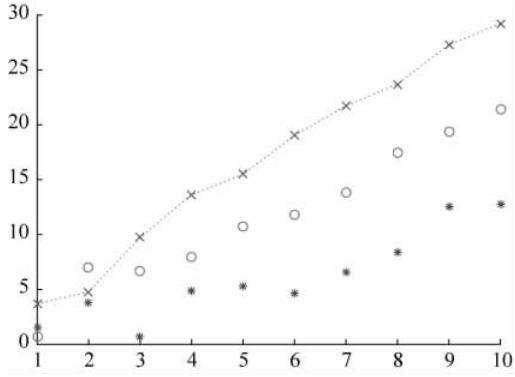

图 3-15 三条拟合线

在拟合曲线上添加最小二乘线,在命令窗口中输入:

```
>> lsline
```

运行程序，效果如图 3-16 所示。

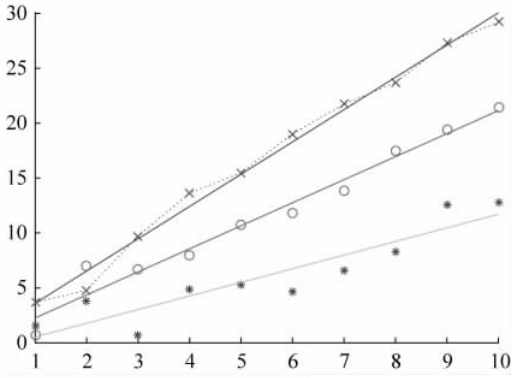

图 3-16　添加最小二乘直线

3.5.9　正态概率图

正态概率图用于检查一组数据是否服从正态分布，是实数与正态分布数据之间函数关系的散点图。如果这组实数服从正态分布，正态概率图将是一条直线。通常，概率图也可以用于确定一组数据是否服从任一已知分布，如二项分布或泊松分布。

在 MATLAB 统计工具箱中提供了 normplot 函数用于绘制图形化正态性检验的正态概率图。函数调用格式如下。

h = normplot(X)：显示数据 X 的正态概率图，如果 X 为矩阵，则为 X 的每一列生成一条直线，该图中的样本数据用图形标记“+”显示，并在图中添加 X 中每列数据 1/4 和 3/4 处的连线，该线可以看做样本次序统计量的稳健性直线拟合，它可帮助评价数据的线性特征，如果数据源于正态分布，则图形呈现直线形，否则为曲线。

【例 3-28】 利用 normplot 函数对给定的正态数据绘制概率图。

其 MATLAB 代码编程如下：

```
>> clear al;
 %生成正态分布数据
M = 100;N = 1;
x = normrnd(0,1,M,N);
 %生成均匀分布
y = rand(M,N);
z = [x,y];
 %绘制正态概率图
h = normplot(z);
xlabel('数据');ylabel('概率');
title('正态概率图');
legend('正态分布数据','均匀分布数据');
grid on;
```

运行程序，效果如图 3-17 所示。

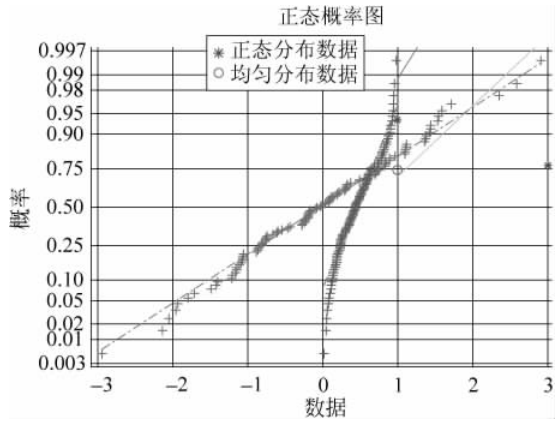

图 3-17　正态概率分布图

在正态概率图中有三个图形元素："+"号表示每一个样本点数值的经验概率；实线连接了数据的第 25 个和第 75 个百分点，表示一个线性拟合；点画线将实线延伸到样本的两端。

在正态概率图中，如果所有的样本点都在直线附近，则假设样本服从正态分布是合理的；否则，如果样本不是正态分布的，则"+"号构成了一条曲线。通过观察图 3-17 中的两种不同分布样本的概率图可以验证这一点。

3.5.10　QQ 图

由两个样本的分位数绘制成的效果图称为 QQ 图，QQ 图亦称为"分位数图"。在 MATLAB 中提供了 qqplot 函数用于绘制 QQ 图，其调用格式如下。

qqplot(X)：显示一个分位数——分位数图。如果绘制分位数图的样本 X 源于正态分布，则绘制的 QQ 图近似于直线。

qqplot(X,Y)：显示两个样本的分位数——分位数。如果两个样本来源于同一分布，那么，图中的曲线为直线。如果 X 与 Y 为乱阵，则为它们的每列数据绘制单独的曲线。图中样本数据以"+"符号表示，并将位于第一分位数和第三分位数间的数据拟合绘制成一条线(这是两个样本顺序统计量的鲁棒性拟合)。此线外推到样本数据的两端，以帮助用户评估数据的线性程度。

qqplot(X,Y,pvec)：函数可在 pvec 矢量中规定分位数。

h=qqplot(X,Y,pvec)：返回线段的句柄值 h。

【例 3-29】　绘制样本的 QQ 图。

其 MATLAB 代码编程如下：

```
>> clear all;
 %生成正态分布数据
M = 100;N = 1;
x = normrnd(0,1,M,N);
 %生成均匀分布
y = rand(M,N);
z = [x,y];
 %绘制 QQ 图
```

```
subplot(221);
h1 = qqplot(z);
xlabel('标准正态样本');ylabel('输入样本');title('QQ 图');
legend('正态分布数据','均匀分布数据');
grid on;
% 生成两个正态分布样本
x = normrnd(0,1,100,1);
y = normrnd(0.5,2,50,1);
subplot(222)
h2 = qqplot(x,y);
xlabel('输入样本 x');ylabel('输入样本 y ');title('QQ 图');
grid on;
% 生成两个不同分布的样本
x = normrnd(5,1,100,1);
y = weibrnd(2,0.5,100,1);
subplot(223)
h3 = qqplot(x,y);
xlabel('输入样本 x');ylabel('输入样本 y');title('QQ 图');
grid on;
subplot(224)
% 生成一个正态分布的样本
x = normrnd(10,1,100,1);
subplot(224)
qqplot(x);
xlabel('输入样本 x');ylabel('输入样本 x');title('QQ 图');
grid on;
```

运行程序,效果如图 3-18 所示。

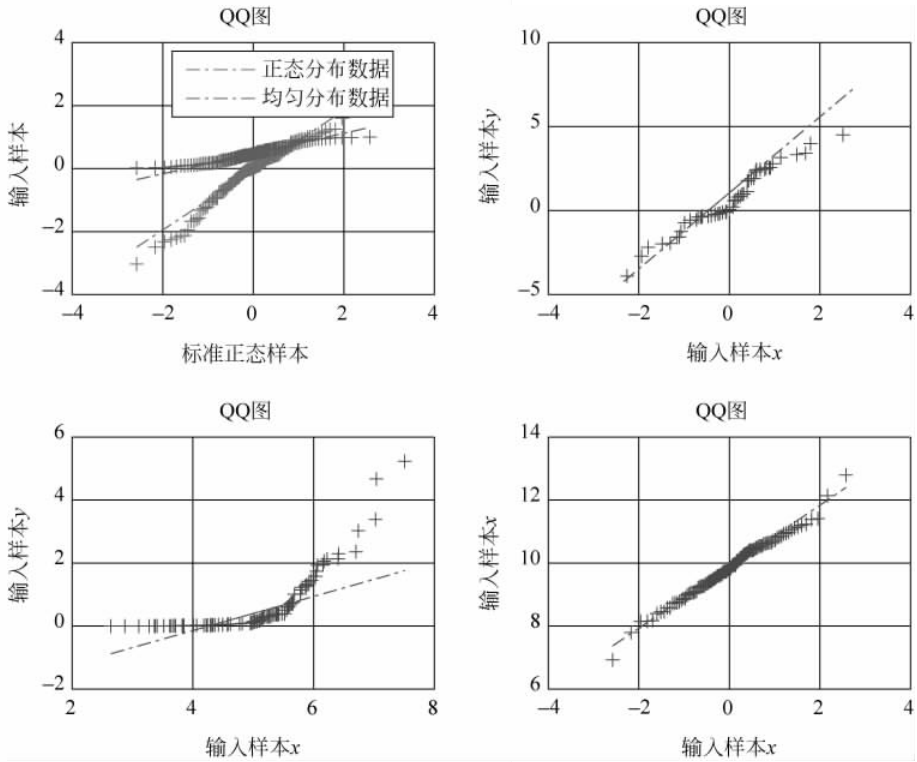

图 3-18　QQ 图

3.5.11 帕累托图

帕累托图又叫排列图、主次图，是按照发生频率大小顺序绘制的直方图，表示有多少结果是由已确认类型或范畴的原因所造成。它是将出现的质量问题和质量改进项目按照重要程度依次排列而采用的一种图表。可以用来分析质量问题，确定产生质量问题的主要因素。按等级排序的目的是指导如何采取纠正措施，项目班子应首先采取措施纠正造成最多数量缺陷的问题。从概念上说，帕累托图与帕累托法则一脉相承，该法则认为相对来说数量较少的原因往往造成绝大多数的问题或缺陷。

帕累托法则往往称为二八原理，即百分之八十的问题是百分之二十的原因所造成的。帕累托图在项目管理中主要用来找出产生大多数问题的关键原因，用来解决大多数问题。

在帕累托图中，不同类别的数据根据其频率降序排列的，并在同一张图中画出累积百分比图。帕累托图可以体现帕累托原则：数据的绝大部分存在于很少类别中，极少剩下的数据分散在大部分类别中。这两组经常被称为“至关重要的极少数”和“微不足道的大多数”。

帕累托图能区分“微不足道的大多数”和“至关重要的极少数”，从而方便人们关注于重要的类别。帕累托图是进行优化和改进的有效工具，尤其应用在质量检测方面。

在 MATLAB 的统计工具箱中提供了 pareto 函数用于绘制帕累托图。函数调用格式如下。

pareto(Y)：将矢量 Y 中的每个元素按元素数值递减顺序绘成直方条，并以其 Y 中的索引号进行标记。各直方条上方的折线显示累积频率。

pareto(Y,names)：以字符串 names 中的名称对 Y 中相应的元素所绘的直方条进行标记。

pareto(Y,X)：根据给定的 X 值对直方条进行标记。

H = pareto(…)：返回帕累托图的句柄值 H。

【例 3-30】 根据给定的一组生产工的生产情况，绘制帕累托图。

其 MATLAB 代码编程如下：

```
>> clear all;
 % 给定生产力
codelines = [200 120 555 608 1024 101 57 687];
 % 生产工名
coders = {'Fred','Ginger','Norman','Max','Julia','Wally','Heidi','Pat'};
pareto(codelines, coders)                          % 绘制帕累托图
title('生产力制帕累托图')
xlabel('数据');ylabel('效果图');
```

运行程序，效果如图 3-19 所示。

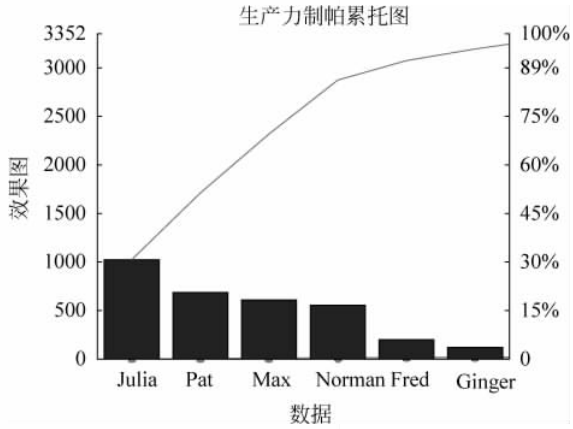

图 3-19　帕累托图

3.5.12　频率直方图

将样本观测值 $x_1, x_2, \cdots, x_n$ 从小到大排序并去除多余的重复值，得到 $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$。适当选取略小于 $x_{(1)}$ 的数 a 与略大于 $x_{(n)}$ 的数 b，将区间 (a,b) 随意分为 k 个不相交的小区间，记第 i 个小区间为 I_i，其长度为 h_i。把样本观测值逐个分到各区间内，并计算样本观测值落在各区间内的频数 n_i 及频率 $f_i = \frac{n_i}{n}$。在 x 轴上截取各区间，并以各区间为底，以 n_i 为高作小矩形，就得到频数直方图；如果以 $\frac{f_i}{h_i}$ 为高作小矩形，就得到频率直方图。

在 MATLAB 统计工具箱中提供了 ecdfhist 函数用于绘制频率直方图。其调用格式如下。

n = ecdfhist(f,x)：其中参数 f 为给定的经验分布函数，x 为给定的样本值。

n = ecdfhist(f,x,m)：m 为划分的区间数，其为一个标量。

n = ecdfhist(f,x,c)：c 也为一个标量，用于指定中心频率。

ecdfhist(…)：绘制频率直方图。

【例 3-31】 绘制随机分布的频率直方图。

其 MATLAB 代码编程如下：

```
>> clear all;
y = exprnd(10,50,1);                          % 随机故障次数
d = exprnd(20,50,1);                          % 随机丢失次数
t = min(y,d);                                 % 最低次数
censored = (y>d);                             % 观察是否受失败
 % 计算经验分布并绘制频率直方图
[f,x] = ecdf(t,'censoring',censored);
ecdfhist(f,x)
set(get(gca,'Children'),'FaceColor',[.8 .8 1])
```

```
hold on
 % Superimpose a plot of the known population pdf
xx = 0:.1:max(t);
yy = exp( - xx/10)/10;
plot(xx,yy,'r - ','LineWidth',2)
hold off
```

运行程序,效果如图 3-20 所示。

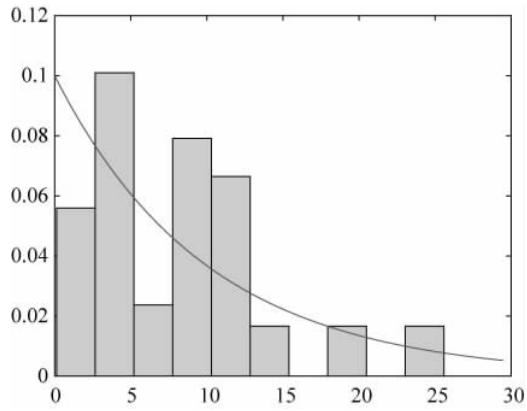

图 3-20　随机数据的频率直方图