数学建模算法与应用

第7章　数理统计

数理统计研究的对象是受随机因素影响的数据，它是以概率论为基础的一门应用学科。数据样本少则几个，多则成千上万，人们希望能用少数几个包含其最多相关信息的数值来体现数据样本总体的规律。面对一批数据进行分析和建模，首先需要掌握参数估计和假设检验这两个数理统计的最基本方法，给定的数据满足一定的分布要求后，才能建立回归分析和方差分析等数学模型。

7.1　参数估计和假设检验

7.1.1　区间估计

例 7.1　有一大批糖果，现从中随机地取 16 袋，称得重量（以 g 计）如下：

506　508　499　503　504　510　497　512
514　505　493　496　506　502　509　496

设袋装糖果的重量近似地服从正态分布，试求总体均值 μ 的置信水平为 0.95 的置信区间。

解　μ的一个置信水平为$1-\alpha$的置信区间为$\left(\bar{X}\pm\frac{S}{\sqrt{n}}t_{\alpha/2}(n-1)\right)$，这里显著性水平$\alpha=0.05$，$\alpha/2=0.025$，$n-1=15$，$t_{0.025}(15)=2.1315$，由给出的数据算得$\bar{x}=503.75$，$s=6.2022$。计算得总体均值$\mu$的置信水平为 0.95 的置信区间为（程序略）

(500.4451,507.0549)

注：Matlab 命令 ttest 实际上是进行单个总体，方差未知的t检验，同时给出了参数的区间估计。

例 7.2 从一批灯泡中随机地取 5 只作寿命试验，测得寿命（以 h 计）为

1050　1100　1120　1250　1280

设灯泡寿命服从正态分布。求灯泡寿命平均值的置信区间为 0.95 的单侧置信区间。

解　这里显著性水平 $\alpha=0.05$，$n=5$，$t_\alpha(n-1)=2.1318$，$\bar{x}=1160$，$s=99.7497$，

寿命平均值 μ 的置信水平为 $1-\alpha$ 的单侧置信下限为

$$\underline{\mu}=\bar{X}-\frac{S}{\sqrt{n}}t_\alpha(n-1),$$

计算得所求的单侧置信下限为

$$\underline{\mu}=\bar{x}-\frac{s}{\sqrt{n}}t_\alpha(n-1)=1064.9.$$

例 7.3 分别使用金秋和铂球测定引力常数(单位: $10^{-11}m^3 \cdot kg^{-1} \cdot s^{-2}$)。

(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 $N(\mu,\sigma^2)$，μ,σ^2 均为未知，试就(1)，(2)两种情况分别求 μ 的置信度为 0.9 的置信区间，并求 σ^2 的置信度为 0.9 的置信区间。

解　（1）μ,σ^2均未知时，μ的置信度为 0.9 的置信区间为

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

这里$1-\alpha=0.9$，$\alpha=0.1$，$\alpha/2=0.05$，$n_1=6$，$n_2=5$，$n_1-1=5$，$n_2-1=4$。

$$\bar{x}_1=\frac{1}{6}\sum_{i=1}^{6}x_i=6.678,\quad \bar{s}_1^2=\frac{1}{5}\sum_{i=1}^{6}(x_i-\bar{x}_1)^2=0.15\times10^{-4},$$

$$\bar{x}_2=\frac{1}{5}\sum_{i=1}^{5}x_i=6.664,\quad \bar{s}_2^2=\frac{1}{4}\sum_{i=1}^{4}(x_i-\bar{x}_2)^2=0.9\times10^{-5},$$

$$t_{\alpha/2}(5)=2.0150,\quad t_{\alpha/2}(4)=2.1318.$$

代入得，用金球测定时，μ的置信区间是(6.675,6.681)，用铂球测定时，μ的置信区间为(6.661,6.667)。

（2）μ,σ^2均未知时，σ^2的置信度为 0.9 的置信区间为

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

这里$n_1-1=5$，$n_2-1=4$，$\alpha/2=0.05$，查表得

$\chi^2_{\alpha/2}(5)=11.071$，$\chi^2_{\alpha/2}(4)=9.488$，

$\chi^2_{1-\alpha/2}(5)=1.145$，$\chi^2_{1-\alpha/2}(4)=0.711.$

将这些值以及上面（1）中算得的s_1^2,s_2^2代入上面区间得

用金球测定时，σ^2的置信区间是

$(6.76\times10^{-6},6.533\times10^{-5})$，

用铂球测定时，σ^2的置信区间是

$(3.79\times10^{-6},5.065\times10^{-5})$。

例 7.4（续例 7.3）在例 7.3 中，设用金球和铂球测定时总体的方差相等，求两个测定值总体均值差的置信度为 0.90 的置信区间。

解　由题意知，总体均值差的置信度为 0.90 的置信区间为

$$\left(\bar{X}_1-\bar{X}_2\pm t_{\alpha/2}(n_1+n_2-2)S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right),$$

这里

$$S_w^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}.$$

此题中，$1-\alpha=0.90$，$\alpha=0.10$，$\alpha/2=0.05$；$n_1=6$，$n_2=5$，$n_1+n_2-2=9$，查表得$t_{\alpha/2}(9)=1.8331$。计算得，$s_w^2=1.233\times10^{-5}$，$s_w=\sqrt{s_w^2}=3.512\times10^{-3}$。

代入公式得总体均值差的置信度为 0.90 的置信区间为 $(0.010,0.018)$。

7.1.2 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）。

一般地，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

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

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1一致收敛于总体分布函数$F(x)$。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 7.5 下面列出了 84 个伊特拉斯坎（Etruscan）人男子的头颅的最大宽度（mm），计算经验分布函数并画出经验分布函数图形。

141	148	132	138	154	142	150	146	155	158
150	140	147	148	144	150	149	145	149	158
143	141	144	144	126	140	144	142	141	140
145	135	147	146	141	136	140	146	142	137
148	154	137	139	143	140	131	143	141	149
148	135	148	152	143	144	141	143	147	146
150	132	142	142	143	153	149	146	149	138
142	149	142	137	134	144	146	147	140	142
140	137	152	145						

解　首先把上面数据保存在纯文本文件 ex7_5.txt 中，计算经验分布函数$F_n(x)$在每个点x_i的值，计算结果保存在 Excel 文件。画出经验分布函数$F_n(x)$的图形，如图 7.1。

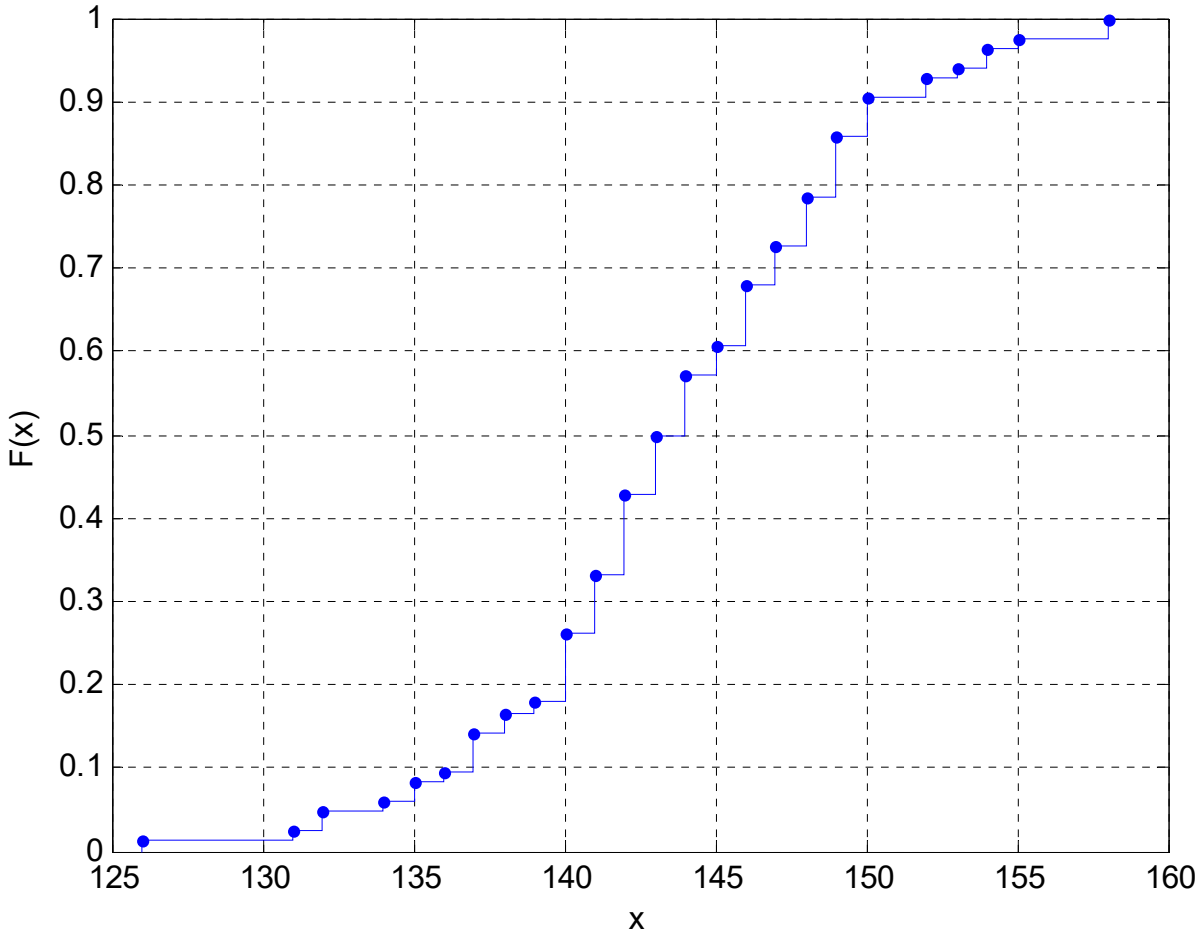

图 7.1　经验分布函数取值图

7.1.3 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，以下我们简单地给出计算步骤。

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 7.6（续 7.5）如果这些数据来自于正态总体，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

$$\bar{x} = 143.7738，\ s = 5.9705，$$

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 $x_1, x_2, \cdots, x_{84}$，并依从小到大顺序排列为

$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(84)}.$$

ii）取 $y_i = F^{-1}((i-1/2)/n)$，$i = 1,2,\cdots,84$，这里 是参数 $\mu = 143.7738$，$\sigma = 5.9705$ 的正态分布函数的反函数。

iii）将$(y_i, x_{(i)})$（$i = 1,2,\cdots,84$）这 84 个点画在直角坐标系上，如图 7.2。

iv）这些点看起来接近一条45°角的直线，说明拟合结果较好。

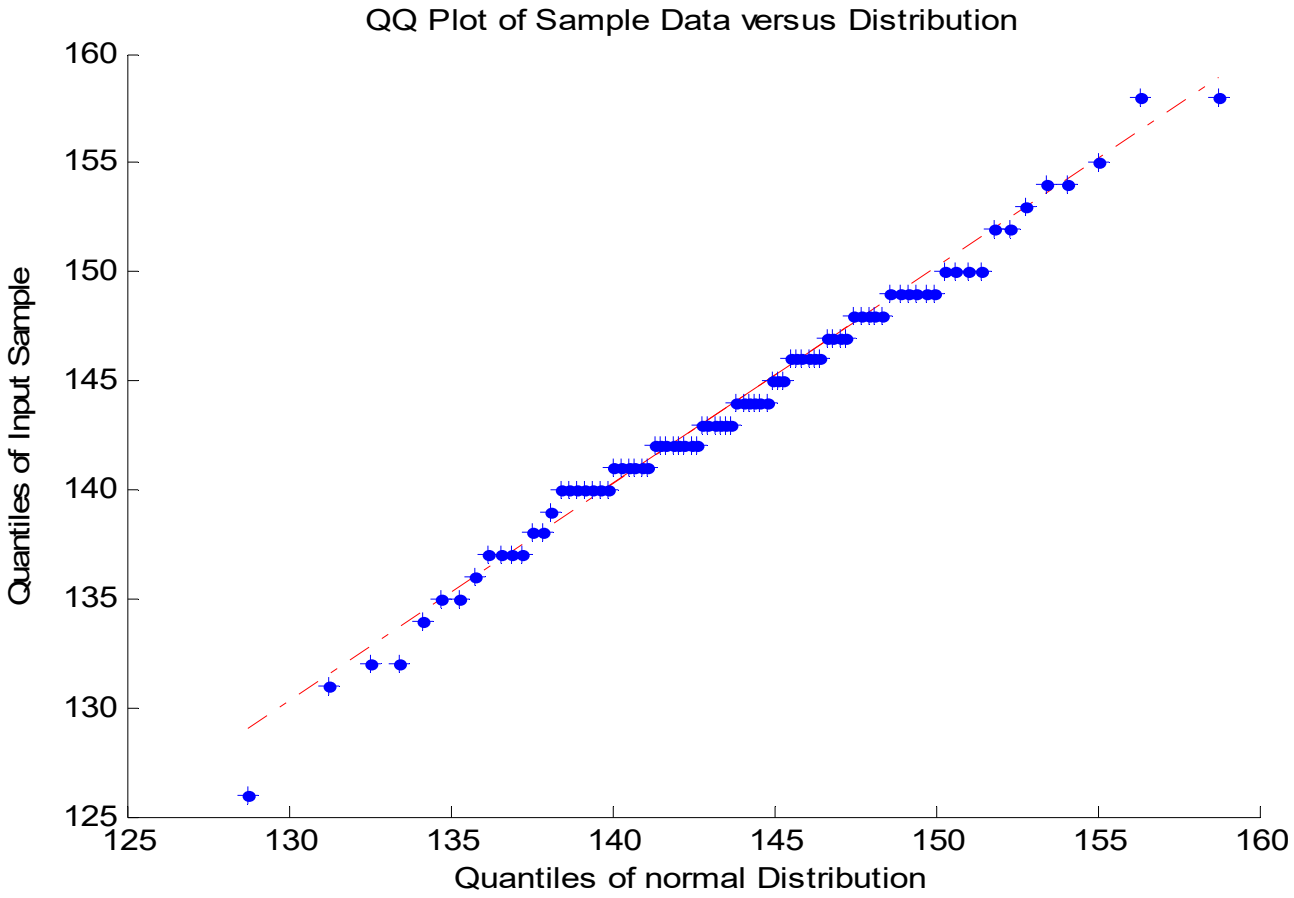
QQ Plot of Sample Data versus Distribution
Quantiles of Input Sample
Quantiles of normal Distribution
160
155
150
145
140
135
130
125
125
130
135
140
145
150
155
160

图 7.2　Q-Q 图

7.1.4　非参数检验

1．χ^2拟合优度检验

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

（5）作出判断：若$\chi^2<\chi_\alpha^2$，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 7.7　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 7.1。

表 7.1　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

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

因参数λ未知，应先根据观察值，用矩估计法来求λ的估计。可知λ的矩估计值为$\hat{\lambda} = \overline{x} = 1$。

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 7.2 左起第一栏所示的两两不相交的子集：A_0,A_1,A_2,A_3，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，计算结果列于表 7.2。

表 7.2　χ^2检验数据表

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	36.7879	35.2289
$A_1:\{X=1\}$	40	0.3679	36.7879	43.4925
$A_2:\{X=2\}$	19	0.1839	18.3940	19.6260
$A_3:\{X\geq 3\}$	5	0.0803	8.0291	3.1137
				$\Sigma=101.4611$

今$\chi^2 = 101.4611 - 100 = 1.4611$，因估计了一个参数，$r = 1$，只有 4 组，故$k = 4$，$\alpha = 0.05$，$\chi_\alpha^2(k - r - 1) = \chi_{0.05}^2(2) = 5.9915 > 1.4611 = \chi^2$，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。

例 7.8　在一批灯泡中抽取 300 只作寿命试验，其结果如表 7.3。

表 7.3　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

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

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 7.4。

以A_i记事件$\{X \in A_i\}$。若H_0为真，X的分布函数为

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

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 7.4。

表 7.4　χ^2检验数据表

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	118.0408	124.0334
$A_2 : 100 < t \le 200$	78	0.2387	71.5954	84.9776
$A_3 : 200 < t \le 300$	43	0.1447	43.4248	42.5794
$A_4 : t > 300$	58	0.2231	66.9390	50.2547
				$\Sigma = 301.845$

今$\chi^2 = 1.845$。由$\alpha = 0.05$，$k = 4$，$r = 0$知

$$\chi_\alpha^2(k-r-1) = \chi_{0.05}^2(3) = 7.8147 > 1.845 = \chi^2.$$

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

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

例7.9 表7.5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7.5 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

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

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 7.6）。用A_i记事件“X的观察值落在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 7.6。

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

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

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 7.6）。用A_i记事件“X的观察值落在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 7.6。

表 7.6　χ^2检验数据表

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
				$\Sigma = 205.5309$

例 7.10（续 7.5）试检验这些数据是否来自正态总体($\alpha = 0.1$)。

解　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

$$\overline{x} = 143.7738，\ s = 5.9705，$$

本题是在显著性水平$\alpha = 0.1$下，检验假设：H_0：头颅的最大宽度X服从正态分布$N(143.7738, 5.9705^2)$。样本观察值的最小值为 126，最大值为 158。将区间[126,158]分成互不相交的 7 个区间：$A_1, A_2, \cdots, A_7$如表 7.7 所示。

以 f_i $(i=1,2,\cdots,7)$ 记样本观察值落在 A_i 中的个数，以 A_i 记事件 $\{X \in A_i\}$。若 H_0 为真，X 服从正态分布 $N(143.7738, 5.9705^2)$，可以计算出

$$p_i = P(A_i) = P\{a_i < X \le a_{i+1}\}，\ i=1,2,\cdots,7.$$

计算结果列于表 7.7。

表 7.7　χ^2检验数据表

A_i	f_i	$\hat{p}_i$	$n\hat{p}_i$	$f_i^2/(n\hat{p}_i)$
$A_1: 126 \le t \le 135.6$	7	0.0855	7.1816	6.8230
$A_2: 135.6 < t \le 138.8$	7	0.1169	9.8204	4.9896
$A_3: 138.8 < t \le 142$	22	0.1808	15.1865	31.8703
$A_4: 142 < t \le 145.2$	15	0.2112	17.7409	12.6826
$A_5: 145.2 < t \le 148.4$	15	0.1864	15.6563	14.3712
$A_6: 148.4 < t \le 151.6$	10	0.1243	10.4375	9.5809
$A_7: 151.6 < t \le 158$	8	0.0950	7.9768	8.0233
				$\Sigma = 88.3408$

2．柯尔莫哥洛夫（Kolmogorov-Smirnov）检验

χ^2 拟合优度检验实际上是检验 $p_i = F_0(a_i) - F(a_{i-1}) = p_{i0}$ $(i = 1,2,\cdots,k)$ 的正确性，并未直接检验原假设的分布函数$F_0(x)$的正确性，柯尔莫哥洛夫检验直接针对原假设$H_0 : F(x) = F_0(x)$，这里分布函数$F(x)$必须是连续型分布。柯尔莫哥洛夫检验基于经验分布函数(或称样本分布函数)作为检验统计量，检验理论分布函数与样本分布函数的拟合优度。

设总体X服从连续分布，$X_1, X_2, \cdots, X_n$是来自总体X的简单随机样本，F_n为经验分布函数，根据大数定律，当n趋于无穷大时，经验分布函数$F_n(x)$依概率收敛总体分布函数$F(x)$。定义$F_n(x)$到$F(x)$的距离为

$$D_n = \sup_{-\infty < x < +\infty} \left| F_n(x) - F(x) \right|,$$

当n趋于无穷大时，D_n依概率收敛到 0。检验统计量建立在D_n基础上。

柯尔莫哥洛夫检验的步骤如下:

（1）原假设和备择假设

$$H_0: F(x) = F_0(x) \quad ,$$

$H_1: F(x) \neq F_0(x).$

（2）选取检验统计量

$$D_n = \sup_{-\infty < x < +\infty} \left| F_n(x) - F(x) \right|,$$

当H_0为真时，D_n有偏小趋势，则拟合得越好；当H_0不真时，D_n有偏大趋势，则拟合得越差。

Kolmogorov 定理 在$F_0(x)$为连续分布的假定下，当原假设为真时，$\sqrt{n}D_n$的极限分布为

$$\lim_{n\to\infty} P\{\sqrt{n}D_n \le t\} = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2t^2}\ ,\ \ t>0.$$

推导检验统计量的分布时，使用$\sqrt{n}D_n$比D_n方便。在显著性水平α下，一个合理的检验是：如果$\sqrt{n}D_n > k$，则拒绝原假设，其中k是合适的常数。

（3）确定拒绝域

给定显著性水平α，查D_n极限分布表，求出t_α满足

$$P\{\sqrt{n}D_n \geq t_\alpha\} = \alpha,$$

作为临界值，即拒绝域为$[t_\alpha, +\infty)$。

（4）作判断

计算统计量的观察值，如果检验统计量$\sqrt{n}D_n$的观察值落在拒绝域中，则拒绝原假设，否则不拒绝原假设。

注：对于固定的α值，我们需要知道该α值下检验的临界值。常用的是在统计量为D_n时，各个α值所对应的临界值如下：在$\alpha=0.1$的显著性水平下，检验的临界值是$1.22/\sqrt{n}$；在$\alpha=0.05$的显著性水平下，检验的临界值是$1.36/\sqrt{n}$；在$\alpha=0.01$的显著性水平下，检验的临界值是$1.63/\sqrt{n}$。这里n为样本的个数。当由样本计算出来的D_n值小于临界值时，说明不能拒绝零假设，所假设的分布是可以接受的；当由样本计算出来的D_n值大于临界值时，拒绝零假设，即所假设的分布是不能接受的。

例 7.11（续例 7.5）试用柯尔莫哥洛夫检验法检验这些数据是否服从正态分布（$\alpha = 0.05$）。

解　（1）假设H_0：$X \sim N(\mu,\sigma^2)$，H_1：X不服从$N(\mu,\sigma^2)$。

这里取μ和σ^2的估计值为

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{84} x_i = 143.7738,$$

$$\hat{\sigma}^2 = \frac{1}{84}\sum_{i=1}^{84}(x_i - \bar{x})^2 = 5.9705^2,$$

即H_0：$X \sim N(143.7738, 5.9705^2)$。

（2）$\alpha = 0.05$，拒绝域为$D_n \geq \frac{1.36}{\sqrt{n}}$，这里$n = 84$。

（3）计算经验分布函数值$F_n(x_i)$和理论分布函数值$F(x_i)$，并计算统计量$D_n=\sup_{x_i}\left|F_n(x_i)-F(x_i)\right|=0.0851$，由于$1.36/\sqrt{n}=0.1484$，所以$D_n<1.36/\sqrt{n}$，接受原假设，认为这些数据服从正态分布。

7.1.5 秩和检验

秩和检验可用于检验假设H_0：两个总体X与Y有相同的分布。

设分别从X、Y两总体中独立抽取大小为n_1和n_2的样本，设$n_1 \le n_2$，其检验步骤如下：

（1）将两个样本混合起来，按照数值大小统一编序，由小到大，每个数据对应的序数称为秩。

（2）计算取自总体X的样本所对应的秩之和，用T表示。

（3）根据n_1,n_2与水平α，查秩和检验表，得秩和下限T_1与上限T_2。

（4）如果$T \le T_1$或$T \ge T_2$，则否定假设H_0，认为X,Y两总体分布有显著差异。否则认为X、Y两总体分布在水平α下无显著差异。

秩和检验的依据是，如果两总体分布无显著差异，那么T不应太大或太小，以T_1和T_2为上、下界的话，则T应在这两者之间，如果T太大或太小，则认为两总体的分布有显著差异。

例 7.12　某涂漆原工艺规定烘干温度为 120℃，现欲将烘干温度提高到 160℃，为了考虑温度变化后是否对零件抗弯强度有明显影响，今用同一涂漆工艺加工了 15 个零件，其中 9 个在 120℃下烘干，6 个在 160℃下烘干，分别测得烘干后各零件的抗弯强度数值如表 7.8 所列。试讨论烘干温度对抗弯强度在水平$\alpha = 0.05$下是否有显著影响?

表 7.8　抗弯强度数值

120℃	41.5	42.0	40.0	42.5	42.0	42.2	42.7	42.1	41.4
160℃	41.2	41.8	42.4	41.6	41.7	41.3			

解　（1）15 个数据按自小到大的顺序排列结果如表 7.9 所列。

表 7.9　数据自小到大排序结果

秩号	1	2	3	4	5	6	7	8
120℃	40.0			41.4	41.5			
160℃		41.2	41.3			41.6	41.7	41.8
秩号	9	10	11	12	13	14	15	
120℃	42.0	42.1	42.2	42.3		42.5	42.7	
160℃					42.4			

（2）120℃下有 9 个数据：$n_2 = 9$，160℃下有 6 个数据，$n_1 = 6$，$n_1 < n_2$，所以

$$T = 2 + 3 + 6 + 7 + 8 + 13 = 39.$$

（3）对$\alpha = 0.05$，查秩和检验表得$T_1 = 33$，$T_2 = 63$。

（4）因为$33 < 39 < 63$，即$T_1 < T < T_2$，所以认为在两种不同的烘干温度下，零件的抗弯强度没有显著差异。

7.2 Bootstrap 方法

7.2.1 非参数 Bootstrap 方法

设总体的分布F未知，但已知有一个容量为n的来自分布F的数据样本，自这一样本按放回抽样的方法抽取一个容量为n的样本，这种样本称为 bootstrap 样本或称为自助样本。相继地，独立地自原始样本中取很多个 Bootstrap 样本，利用这些样本对总体F进行统计推断，这种方法称为非参数 Bootstrap 方法，又称自助法。

这一方法可以用于当人们对总体知之甚少的情况，它是近代统计中的一种用于数据处理的重要实用方法。这种方法的实现需要在计算机上作大量的计算，随着计算机威力的增长，它已成为一种流行的方法。

Bootstap 方法是 Efron 在 20 世纪 70 年代后期建立的。

1.估计量的标准误差的 Bootstrap 估计

在估计总体未知参数θ时，人们不但要给出θ的估计$\hat{\theta}$，还需指出这一估计$\hat{\theta}$的精度。通常我们用估计量$\hat{\theta}$的标准差$\sqrt{D(\hat{\theta})}$来度量估计的精度。估计量$\hat{\theta}$的标准差$\sigma_{\hat{\theta}}=\sqrt{D(\hat{\theta})}$也称为估计量$\hat{\theta}$的标准误差。

设$X_1, X_2, \cdots, X_n$是来自以$F(x)$为分布函数的总体的样本，θ是我们感兴趣的未知参数，用$\hat{\theta} = \hat{\theta}(X_1, X_2, \cdots, X_n)$作为$\theta$的估计量，在应用中$\hat{\theta}$的抽样分布常是很难处理的，这样，$\sqrt{D(\hat{\theta})}$常没有一个简单的表达式，不过我们可以用计算机模拟的方法来求得$\sqrt{D(\hat{\theta})}$的估计。

为此，自F产生很多容量为n的样本（例如B个），对于每一个样本计算$\hat{\theta}$的值，得$\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_B$，则$\sqrt{D(\hat{\theta})}$可以用

$$\hat{\sigma}_{\hat{\theta}}=\sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i-\bar{\theta})^2}$$

来估计，其中$\bar{\theta}=\frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i$。然而$F$常常是未知的，这样就无法产生模拟样本，需要另外的方法。

现在设分布F未知，$x_1, x_2, \cdots, x_n$是来自F的样本值，F_n是相应的经验分布函数。当n很大时，F_n接近F。我们用F_n代替上一段中的F，在F_n中抽样。在F_n中抽样，就是在原始样本$x_1, x_2, \cdots, x_n$中每次随机地取一个个体作放回抽样。如此得到一个容量为n的样本$x_1^*, x_2^*, \cdots, x_n^*$，这就是第一段中所说的 Bootstrap 样本。

用 Bootstrap 样本按上一段中计算估计 $\hat{\theta}(x_1, x_2, \cdots, x_n)$那样求出$\theta$的估计$\hat{\theta}^* = \hat{\theta}(x_1^*, x_2^*, \cdots, x_n^*)$，估计$\hat{\theta}^*$称为$\theta$的 Bootstrap 估计。相应地、独立地抽得$B$个 Bootstrap 样本，以这些样本分别求出$\theta$的相应的 Bootstrap 估计如下：

Bootstrap 样本 1　$x_1^{*1}, x_2^{*1}, \cdots, x_n^{*1}$，Bootstrap 估计$\hat{\theta}_1^*$；

Bootstrap 样本 2　$x_1^{*2}, x_2^{*2}, \cdots, x_n^{*2}$，Bootstrap 估计$\hat{\theta}_2^*$；

……

Bootstrap 样本B　$x_1^{*B}, x_2^{*B}, \cdots, x_n^{*B}$，Bootstrap 估计$\hat{\theta}_B^*$.

则$\hat{\theta}$的标准误差$\sqrt{D(\hat{\theta})}$，就以

$$\hat{\sigma}_{\hat{\theta}} = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i^* - \bar{\theta}^*)^2}$$

来估计，其中$\bar{\theta}^* = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^*$，上式就是$\sqrt{D(\hat{\theta})}$的 Bootstrap 估计。

综上所述得到求$\sqrt{D(\hat{\theta})}$的 Bootstrap 估计的步骤是：

1°自原始数据样本$x_1, x_2, \cdots, x_n$按放回抽样的方法，抽得容量为n的样本$x_1^*, x_2^*, \cdots, x_n^*$（称为 Bootstrap 样本）；

2°相继地、独立地求出B（$B \geq 1000$）个容量为n的 Bootstrap 样本，$x_1^{*i}, x_2^{*i}, \cdots, x_n^{*i}$，$i = 1,2,\cdots,B$。对于第$i$个 Bootstrap 样本，计算$\hat{\theta}_i^* = \hat{\theta}(x_1^{*i}, x_2^{*i}, \cdots, x_n^{*i})$，$i = 1,2,\cdots,B$（$\hat{\theta}_i^*$称为$\theta$的第$i$个 Bootstrap 估计）。

3°计算

$$\hat{\sigma}_{\hat{\theta}} = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i^* - \bar{\theta}^*)^2}$$，其中$\bar{\theta}^* = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^*$。

例 7.13 某种基金的年回报率是具有分布函数F的连续型随机变量，F未知，F的中位数θ是未知参数。现有以下的数据：

18.2　9.5　12.0　21.1　10.2

以样本中位数作为总体中位数θ的估计。试求中位数估计的标准误差的 Bootstrap 估计。多元线性回归分析的模型为

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon, \\ \varepsilon \sim N(0, \sigma^2), \end{cases} \qquad (15.20)$$

式中$\beta_0, \beta_1, \cdots, \beta_m, \sigma^2$都是与$x_1, x_2, \cdots, x_m$无关的未知参数，其中$\beta_0, \beta_1, \cdots, \beta_m$称为回归系数。

解　将原始样本自小到大排序，中间一个数为 12.0，得样本中位数为 12.0。

相继地、独立地在上述 5 个数据中，按放回抽样的方法取样，取$B=10$得到下述 10 个 Bootstrap 样本：

样本 1	9.5	18.2	12.0	10.2	18.2
样本 2	21.1	18.2	12.0	9.5	10.2
样本 3	21.1	10.2	10.2	12.0	10.2
样本 4	18.2	12.0	9.5	18.2	10.2
样本 5	21.1	12.0	18.2	12.0	18.2
样本 6	10.2	10.2	9.5	21.1	10.2
样本 7	9.5	21.1	12.0	10.2	12.0
样本 8	10.2	18.2	10.2	21.1	21.1
样本 9	10.2	10.2	18.2	18.2	18.2
样本 10	18.2	10.2	18.2	10.2	10.2

对以上每个 Bootstrap 样本，求得样本中位数分别为

$\hat{\theta}_1^* = 12.0$，$\hat{\theta}_2^* = 12.0$，$\hat{\theta}_3^* = 10.2$，$\hat{\theta}_4^* = 12.0$，$\hat{\theta}_5^* = 18.2$，

$\hat{\theta}_6^* = 10.2$，$\hat{\theta}_7^* = 12.0$，$\hat{\theta}_8^* = 18.2$，$\hat{\theta}_9^* = 18.2$，$\hat{\theta}_{10}^* = 10.2$，

则中位数估计的标准误差的 Bootstrap 估计

$$\hat{\sigma}_{\hat{\theta}} = \sqrt{\frac{1}{9}\sum_{i=1}^{10}(\hat{\theta}_i^* - \bar{\theta}^*)^2} = 3.4579.$$

本题中取$B = 10$，这只是为了说明计算方法，是不能实际运用的，在实际中应取$B \geq 1000$。

下面我们使用计算机进行抽样，取$B = 1000$，其中的一次运行结果

$$\hat{\sigma}_{\hat{\theta}} = \sqrt{\frac{1}{999}\sum_{i=1}^{1000}(\hat{\theta}_i^* - \overline{\theta}^*)^2} = 3.7311.$$

多元线性回归分析的模型为

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon, \\ \varepsilon \sim N(0, \sigma^2), \end{cases} \qquad (15.20)$$

式中$\beta_0, \beta_1, \cdots, \beta_m, \sigma^2$都是与$x_1, x_2, \cdots, x_m$无关的未知参数，其中$\beta_0, \beta_1, \cdots, \beta_m$称为回归系数。

2.估计量的均方误差的 Bootstrap 估计

设$X=(X_1,X_2,\cdots,X_n)$是来自总体F的样本，F未知，$R=R(X)$是感兴趣的随机变量，它依赖于样本X。假设我们希望去估计R的分布的某些特征。例如R的数学期望$E_F(R)$，就可以按照上面所说的三个步骤1°，2°，3°进行，只是在2°中对于第i个 Bootstrap 样本$x_i^*=(x_1^{*i},x_2^{*i},\cdots,x_n^{*i})$，计算$R_i^*=R_i^*(x_i^*)$代替计算$\theta_i^*$，且在3°中计算感兴趣的$R$的特征。

例如如果希望估计$E_F(R)$就计算

$$E_*(R^*)=\frac{1}{B}\sum_{i=1}^{B}R_i^*.$$

例 7.14 设金属元素铂的升华热是具有分布函数F的连续型随机变量，F 的中位数θ是未知参数，现测得以下的数据（以 kcal/mol 计）:

多元线性回归分析的模型为

$$\begin{cases} y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m+\varepsilon, \\ \varepsilon\sim N(0,\sigma^2), \end{cases} \qquad (15.20)$$

式中$\beta_0,\beta_1,\cdots,\beta_m,\sigma^2$都是与$x_1,x_2,\cdots,x_m$无关的未知参数，其中$\beta_0,\beta_1,\cdots,\beta_m$称为回归系数。

136.3 136.6 135.8 135.4 134.7 135.0 134.1
143.3 147.8 148.8 134.8 135.2 134.9 149.5
141.2 135.4 134.8 135.8 135.0 133.7 134.4
134.9 134.8 134.5 134.3 135.2

以样本中位数$M = M(X)$作为总体中位数θ的估计，试求均方误差$MSE = E[(M-\theta)^2]$的 Bootstrap 估计。

解　将原始样本自小到大排序，左起第 13 个数为 135.0，左起第 14 个数为 135.2，于是样本中位数为 $\frac{1}{2}(135.0+135.2)=135.1$。以 135.1 作为总体中位数$\theta$的估计，即$\hat{\theta}=135.1$。取$R=R(X)=(M-\hat{\theta})^2$，需估计$R(X)$的均值$E[(M-\hat{\theta})^2]$。

相继地、独立地抽取 10000 个 Bootstrap 样本如下：

样本 1，得样本中位数为 134.9，

……

样本 10000，得样本中位数为 135.2，

对于用第i个样本计算

$$R_i^* = R(x_i^*) = (M_i^* - \hat{\theta})^2 = (M_i^* - 135.1)^2,$$

$$i = 1, 2, \cdots, 10000.$$

即对于样本 1，$(M_1^* - 135.1)^2 = (134.9 - 135.1)^2 = 0.04$，

……

对于样本 10000，

$(M_{10000}^* - 135.1)^2 = (135.2 - 135.1)^2 = 0.01.$

用这 10000 个数的平均值

$$\frac{1}{10000}\sum_{i=1}^{10000}(M_i^* - 135.1)^2 = 0.071$$

近似$E[(M-\theta)^2]$，即得$MSE[(M-\theta)^2]$的 Bootstrap 估计为 0.071（其中的一次运行结果）。

3.Bootstrap 置信区间

下面介绍一种求未知参数θ的 Bootstrap 置信区间的方法。

设$X=(X_1,X_2,\cdots,X_n)$是来自总体F容量为n的样本，$x=(x_1,x_2,\cdots,x_n)$是一个已知的样本值。F中含有未知参数θ，$\hat{\theta}=\hat{\theta}(X_1,X_2,\cdots,X_n)$是$\theta$的估计量。现在来求$\theta$的置信水平为$1-\alpha$的置信区间。

相继地，独立地从样本$x=(x_1,x_2,\cdots,x_n)$中抽出B个容量为n的 Bootstrap 样本，对于每个 Bootstrap 样本求出θ的 Bootstrap 估计：$\hat{\theta}_1^*,\hat{\theta}_2^*,\cdots,\hat{\theta}_B^*$。将它们自小到大排序，得

$$\hat{\theta}_{(1)}^* \le \hat{\theta}_{(2)}^* \le \cdots \le \hat{\theta}_{(B)}^*.$$

取$R(X)=\hat{\theta}$，用对应的$R(X^*)=\hat{\theta}^*$的分布作为$R(X)$的分布的近似，求出$R(X^*)$的分布的近似分位数$\hat{\theta}^*_{\alpha/2}$和$\hat{\theta}^*_{1-\alpha/2}$使

$$P\{\hat{\theta}^*_{\alpha/2}<\hat{\theta}^*<\hat{\theta}^*_{1-\alpha/2}\}=1-\alpha,$$

于是近似地有

$$P\{\hat{\theta}^*_{\alpha/2}<\theta<\hat{\theta}^*_{1-\alpha/2}\}=1-\alpha.$$

记$k_1=\left[B\times\frac{\alpha}{2}\right]$，$k_2=\left[B\times(1-\frac{\alpha}{2})\right]$，

在上式中以$\hat{\theta}^*_{(k_1)}$和$\hat{\theta}^*_{(k_2)}$分别作为分位数$\hat{\theta}^*_{\alpha/2}$和$\hat{\theta}^*_{1-\alpha/2}$的估计，得到近似等式

$$P\{\hat{\theta}^*_{(k_1)} < \theta < \hat{\theta}^*_{(k_2)}\} = 1-\alpha.$$

于是由上式就得到θ的置信水平为$1-\alpha$的近似置信区间$(\hat{\theta}^*_{(k_1)}, \hat{\theta}^*_{(k_2)})$，这一区间称为$\theta$的置信水平为$1-\alpha$的Bootstrap 置信区间。这种求置信区间的方法称为分位数法。

例 7.15　有 30 窝仔猪出生时各窝猪的存活只数为

9　8　10　12　11　12　7　9　11　8　9　7　7
8　9　7　9　9　10　9　9　9　12　10　10　9　13
11　13　9

以样本均值$\bar{x}$作为总体均值μ的估计，以样本标准差s作为总体标准差σ的估计，按分位数法求μ以及σ的置信水平为 0.90 的 Bootstrap 置信区间。

解　相继地、独立地自原始样本数据用放回抽样的方法，得到 10000 个容量均为 30 的 Bootstrap 样本。

对每个 Bootstrap 样本算出样本均值 $\bar{x}_i^*$ $(i=1,2,\cdots,10000)$，将 10000 个 $\bar{x}_i^*$ 按自小到大排序，左起第 500 位为 $\bar{x}_{(500)}^* = 9.0333$，左起第 9500 位为 $\bar{x}_{(9500)}^* = 10.0667$。于是得 μ 的一个置信水平为 0.90 的 Bootstrap 置信区间为

$$(\bar{x}_{(500)}^*, \bar{x}_{(9500)}^*) = (9.0333, 10.0667).$$

对上述 10000 个 Bootstrap 样本的每一个算出标准差 s_i^* $(i=1,2,\cdots,10000)$，将 10000 个 s_i^* 按自小到大排序。左起第 500 位为 $s_{(500)}^*=1.4464$，左起第 9500 位为 $s_{(9500)}^*=2.0634$，于是得 σ 的一个置信水平为 0.90 的 bootstap 置信区间为

$$(s_{(500)}^*, s_{(9500)}^*)=(1.4464, 2.0634).$$

用非参数Bootstrap法来求参数的近似置信区间的优点是，不需要对总体分布的类型作任何的假设，而且可以适用于小样本，且能用于各种统计量（不限于样本均值）。

以上介绍的 Bootstrap 方法，没有假设所研究的总体的分布函数F的形式，Bootstrap 样本是来自已知的数据（原始样本），所以称之为非参数 Bootstrap 方法。

7.2.2 参数 Bootstrap 方法

假设所研究的总体的分布函数$F(x;\beta)$的形式已知，但其中包含未知参数β（β可以是向量）。现在已知有一个来自$F(x;\beta)$的样本

$$X_1, X_2, \cdots, X_n.$$

利用这一样本求出β的最大似然估计$\hat{\beta}$。在$F(x;\beta)$中以$\hat{\beta}$代替β得到$F(x;\hat{\beta})$，接着在$F(x;\hat{\beta})$中产生容量为n的样本

$$X_1^*, X_2^*, \cdots, X_n^* \sim F(x;\hat{\beta}).$$

这种样本可以产生很多个，例如产生B（$B \geq 1000$）个，就可以利用这些样本对总体进行统计推断，其做法与非参数 Bootstrap 方法一样。这种方法称为参数 Bootstrap 方法。

例 7.16　已知某种电子元件的寿命（以 h 计）服从威布尔分布，其分布函数为

$$F(x)=\begin{cases}1-e^{-(x/\eta)^{\beta}}, & x>0,\\ 0, & \text{其它},\end{cases}\quad \beta>0,\ \eta>0.$$

概率密度为

$$f(x)=\begin{cases}\dfrac{\beta}{\eta^{\beta}}x^{\beta-1}e^{-(x/\eta)^{\beta}}, & x>0,\\ 0, & \text{其它},\end{cases}$$

已知参数$\beta=2$。今有样本

142.84　97.04　32.46　69.14　85.67　114.43

41.76　163.07　108.22　63.28

（1）确定参数η的最大似然估计。

（2）对于时刻$t_0=50$，求可靠性$R(50)=1-F(50)=e^{-(50/\eta)^2}$的置信水平为0.95的Bootstrap单侧置信下限。

解 (1)设有样本$x_1, x_2, \cdots, x_n$,似然函数为(已将$\beta = 2$代入)

$$L = \prod_{i=1}^{n} \frac{2}{\eta^2} x_i e^{-(x_i/\eta)^2} = \frac{2^n}{\eta^{2n}} \left(\prod_{i=1}^{n} x_i \right) e^{-\left(\sum_{i=1}^{n} x_i^2\right)/\eta^2},$$

$$\ln L = n \ln 2 - 2n \ln \eta + \sum_{i=1}^{n} \ln x_i - \frac{1}{\eta^2} \sum_{i=1}^{n} x_i^2,$$

令$\frac{d}{d\eta} \ln L = 0$得

$$\frac{-2n}{\eta} + \frac{2}{\eta^3} \sum_{i=1}^{n} x_i^2 = 0,$$

$$\hat{\eta} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}.$$

以数据代入得η的最大似然估计为$\hat{\eta}=100.0696$。

（2）对于参数$\beta=2$，$\eta=\hat{\eta}=100.0696$，产生服从对应韦布尔分布的 5000 个容量为 10 的 Bootstrap 样本。

对于每个样本$x_1^{*i},x_2^{*i},\cdots,x_{10}^{*i}$，计算$\eta$的 Bootstrap 估计

$$\eta_i^*=\sqrt{\frac{\sum_{j=1}^{10}(x_j^{*i})^2}{10}}\quad .$$

将以上 5000 个η_i^*自小到大排列，取左起第 250（$[5000\times 0.05]=250$）位，得

$$\eta_{(250)}^* = 73.3758,$$

于是在$t=50$时，可靠性$R(50)$的置信水平为 0.95 的 Bootstrap 单侧置信下限为

$$e^{-(50/\hat{\eta}_{(250)}^*)^2} = 0.6286.$$

例 7.17　据 Hardy-Weinberg 定律，若基因频率处于平衡状态，则在一总体中个体具有血型M、MN、N的概率分别是$(1-\theta)^2, 2\theta(1-\theta), \theta^2$，其中$0<\theta<1$。据 1937 年对香港地区的调查有表 7.10 的数据。

表 7.10　血型的数据

血型	M	MN	N	
人数	342	500	187	共 1029

（1）求θ的最大似然估计$\hat{\theta}$；

（2）求θ的置信水平为 0.90 的 Bootstrap 置信区间。

解　分别记x_1, x_2, x_3为具有血型为M、MN、N的人数，记$x_1 + x_2 + x_3 = n$。似然函数为

$$L = [(1-\theta)^2]^{x_1}[2\theta(1-\theta)]^{x_2}(\theta^2)^{x_3} = 2^{x_2}\theta^{x_2+2x_3}(1-\theta)^{2x_1+x_2},$$

$$\ln L = x_2 \ln 2 + (x_2 + 2x_3)\ln\theta + (2x_1 + x_2)\ln(1-\theta).$$

令

$$\frac{d}{d\theta}\ln L = \frac{x_2 + 2x_3}{\theta} - \frac{2x_1 + x_2}{1-\theta} = 0,$$

解得

$$\hat{\theta} = \frac{x_2 + 2x_3}{2x_1 + 2x_2 + 2x_3} = \frac{x_2 + 2x_3}{2n}.$$

以数据$x_1=342$，$x_2=500$，$x_3=187$，$n=1029$，代入得到$\hat{\theta}=0.4247$。以$\hat{\theta}$代替θ，得到$(1-\theta)^2=0.3310$，$2\theta(1-\theta)=0.4887$，$\theta^2=0.1804$。于是血型的近似分布律见表 7.11。

表 7.11　血型的近似分布律

血型	M	MN	N
人数	0.3310	0.4887	0.1804

以表 7.11 为分布律产生 1000 个 Bootstrap 样本，从而得到θ的 1000 个 Bootstrap 估计$\hat{\theta}_1^*, \hat{\theta}_2^*, \cdots, \hat{\theta}_{1000}^*$，将这 1000 个数按自小到大的次序排列得到

$$\hat{\theta}_{(1)}^* \le \hat{\theta}_{(2)}^* \le \cdots \le \hat{\theta}_{(1000)}^*.$$

取$(\hat{\theta}_{(50)}^*, \hat{\theta}_{(950)}^*) = (0.4072, 0.4431)$为$\theta$的置信水平为 0.90 的 Bootstrap 置信区间。

7.3 方差分析

下面只给出单因素试验的方差分析，双因素试验和多因素试验的方差分析是类似的。

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 7.12 所列结果。

表 7.12　方差分析数据表

	A_1	A_2	...	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$T_{\bullet j}$	$T_{\bullet 1}$	$T_{\bullet 2}$	...	$T_{\bullet s}$
样本均值$\overline{X}_{\bullet j}$	$\overline{X}_{\bullet 1}$	$\overline{X}_{\bullet 2}$	...	$\overline{X}_{\bullet s}$
总体均值	μ_1	μ_2	...	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

$n = n_1 + n_2 + \cdots + n_s$，

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

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

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

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

$S_A=S_T-S_E$，多元线性回归分析的模型为

$$\begin{cases}y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m+\varepsilon,\\ \varepsilon\sim N(0,\sigma^2),\end{cases}\quad,\qquad(15.20)$$

式中$\beta_0,\beta_1,\cdots,\beta_m,\sigma^2$都是与$x_1,x_2,\cdots,x_m$无关的未知参数，其中$\beta_0,\beta_1,\cdots,\beta_m$称为回归系数。

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$F_\alpha$比较，最后下结论：若检验统计量$F$的观察值大于临界值$F_\alpha$，则拒绝原假设$H_0$；若$F$的值小于$F_\alpha$，则接受$H_0$。

例 7.18　设有某品牌的三台机器 A，B，C 生产同一产品，对每台机器观测 5 天。其日产量如表 7.13 所示，设各机器日产量服从正态分布，且方差相等，问三台机器的日产量有无显著差异（$\alpha = 0.05$）？

表 7.13　三台机器产量数据表

	A	B	C
1	41	65	45
2	48	57	51
3	41	54	56
4	49	72	48
5	57	64	48

解　设μ_1,μ_2,μ_3分别为 A，B，C 的平均日产量。

1°原假设H_0：$\mu_1=\mu_2=\mu_3$；H_1：μ_1,μ_2,μ_3不全相等。

2°当H_0为真时$F=\dfrac{(n-s)S_A}{(s-1)S_E}\sim F(s-1,n-s)$。

3°此题中，$n=n_1+n_2+n_3=15$，$s=3$，$\alpha=0.05$。

拒绝域为$F>F_\alpha(s-1,n-3)=F_\alpha(2,12)=3.8853.$

由题意列出方差分析表见表 7.14。

表 7.14　方差分析表

	A	B	C	$\sum$
1	41	65	45	
2	48	57	51	
3	41	54	56	
4	49	72	48	
5	57	64	48	
$T_{\bullet j}$	236	312	248	$T_{\bullet\bullet}=796$
$T_{\bullet j}^2$	55696	97344	61504	$\sum_{j=1}^{3}\frac{T_{\bullet j}^2}{n_j}=42908.8$
$\sum_{i=1}^{n_j}X_{ij}^2$	11316	19670	12370	$\sum_{i=1}^{5}\sum_{j=1}^{3}X_{ij}^2=43356$

$$S_T=\sum_{i=1}^{5}\sum_{j=1}^{3}X_{ij}^2-\frac{T_{\bullet\bullet}^2}{n}=43356-42241.07=1114.93,$$

$$S_E=\sum_{i=1}^{5}\sum_{j=1}^{3}X_{ij}^2-\sum_{j=1}^{3}\frac{T_{\bullet j}^2}{n_j}=43356-42908.8=447.2,$$

$$S_A=S_T-S_E=1114.93-447.2=667.73,$$

$$F=\frac{S_A/(s-1)}{S_E/(n-r)}=\frac{667.73/2}{447.2/12}=8.9589,$$

由于$F=8.9589>3.8853$。

4°结论：故拒绝H_0：即认为机器日产量存在显著差异。

7.4　回归分析

7.4.1　多元线性回归

1.模型

多元线性回归分析的模型为

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon, \\ \varepsilon \sim N(0, \sigma^2), \end{cases} \tag{7.1}$$

式中 $\beta_0, \beta_1, \cdots, \beta_m, \sigma^2$ 都是与 $x_1, x_2, \cdots, x_m$ 无关的未知参数，其中 $\beta_0, \beta_1, \cdots, \beta_m$ 称为回归系数。

现得到n个独立观测数据$[b_i, a_{i1}, \cdots, a_{im}]$，其中$b_i$为$y$的观察值，$a_{i1}, \cdots, a_{im}$分别为的观察值，$i = 1, \cdots, n$，$n > m$，由（7.1）得

$$\begin{cases} b_i = \beta_0 + \beta_1 a_{i1} + \cdots + \beta_m a_{im} + \varepsilon_i, \\ \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \cdots, n. \end{cases} \tag{7.2}$$

记

$$X = \begin{bmatrix} 1 & a_{11} & \cdots & a_{1m} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & a_{n1} & \cdots & a_{nm} \end{bmatrix}, Y = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}, \tag{7.3}$$

$\varepsilon = [\varepsilon_1, \cdots, \varepsilon_n]^T$，$\beta = [\beta_0, \beta_1, \cdots, \beta_m]^T$，

（7.1）表示为

$$\begin{cases} Y = X\beta + \varepsilon, \\ \varepsilon \sim N(0, \sigma^2 E_n), \end{cases} \tag{7.4}$$

其中E_n为n阶单位矩阵。

2.参数估计

模型（7.1）中的参数$\beta_0,\beta_1,\cdots,\beta_m$用最小二乘法估计，即应选取估计值$\hat{\beta}_j$，使当$\beta_j=\hat{\beta}_j$，$j=0,1,\cdots,m$时，误差平方和

$$Q=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}(b_i-\hat{b}_i)^2=\sum_{i=1}^{n}(b_i-\beta_0-\beta_1a_{i1}-\cdots-\beta_ma_{im}) \tag{7.5}$$

达到最小。

为此，令

$$\frac{\partial Q}{\partial \beta_j} = 0, \quad j = 0,1,2,\cdots,n,$$

得

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(b_i - \beta_0 - \beta_1 a_{i1} - \cdots - \beta_m a_{im}) = 0, \\ \dfrac{\partial Q}{\partial \beta_j} = -2\sum_{i=1}^{n}(b_i - \beta_0 - \beta_1 a_{i1} - \cdots - \beta_m a_{im}) a_{ij} = 0, \quad j = 1,2, \end{cases}$$

(7.6)

经整理化为以下正规方程组

$$\begin{cases} \beta_0 n + \beta_1 \sum_{i=1}^{n} a_{i1} + \beta_2 \sum_{i=1}^{n} a_{i2} + \cdots + \beta_m \sum_{i=1}^{n} a_{im} = \sum_{i=1}^{n} b_i, \\ \beta_0 \sum_{i=1}^{n} a_{i1} + \beta_1 \sum_{i=1}^{n} a_{i1}^2 + \beta_2 \sum_{i=1}^{n} a_{i1} a_{i2} + \cdots + \beta_m \sum_{i=1}^{n} a_{i1} a_{im} = \sum_{i=1}^{n} a_i \\ \vdots \\ \beta_0 \sum_{i=1}^{n} a_{im} + \beta_1 \sum_{i=1}^{n} a_{im} a_{i1} + \beta_2 \sum_{i=1}^{n} a_{im} a_{i2} + \cdots + \beta_m \sum_{i=1}^{n} a_{im}^2 = \sum_{i=1}^{n} c \end{cases} \tag{7.7}$$

正规方程组的矩阵形式为

$$X^T X \beta = X^T Y \tag{7.8}$$

当矩阵X列满秩时，X^TX为可逆方阵，式（7.8）的解为

$$\hat{\beta}=(X^TX)^{-1}X^TY. \qquad (7.9)$$

将$\hat{\beta}$代回原模型得到y的估计值

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_mx_m, \qquad (7.10)$$

而这组数据的拟合值为

$$\hat{b}_i=\hat{\beta}_0+\hat{\beta}_1a_{i1}+\cdots+\hat{\beta}_ma_{im}\ (i=1,\cdots,n),$$

记$\hat{Y} = X\hat{\beta} = [\hat{b}_1, \cdots, \hat{b}_n]^T$，拟合误差$e = Y - \hat{Y}$称为残差，可作为随机误差$\varepsilon$的估计，而

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (b_i - \hat{b}_i)^2 \qquad (7.11)$$

为残差平方和（或剩余平方和）。

3.统计分析

不加证明地给出以下结果

（1）$\hat{\beta}$是β的线性无偏最小方差估计；$\hat{\beta}$的期望等于β；在β的线性无偏估计中，$\hat{\beta}$的方差最小。

（2）$\hat{\beta}$服从正态分布

$$\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1}), \qquad (7.12)$$

记$(X^TX)^{-1}=(c_{ij})_{n\times n}$。

（3）对残差平方和Q，$EQ=(n-m-1)\sigma^2$，且

$$\frac{Q}{\sigma^2}\sim\chi^2(n-m-1). \qquad (7.13)$$

由此得到σ^2的无偏估计

$$s^2=\frac{Q}{n-m-1}=\hat{\sigma}^2. \qquad (7.14)$$

s^2是剩余方差（残差的方差），s称为剩余标准差。

（4）对总平方和$SST=\sum_{i=1}^{n}(b_i-\bar{b})^2$进行分解，有

$$SST=Q+U,\ U=\sum_{i=1}^{n}(\hat{b}_i-\bar{b})^2, \qquad (7.15)$$

其中$\bar{b}=\frac{1}{n}\sum_{i=1}^{n}b_i$，$Q$是由（7.5）定义的残差平方和，反映随机误差对$y$的影响，$U$称为回归平方和，反映自变量对$y$的影响。上面的分解中利用了正规方程组。

4.回归模型的假设检验

因变量y与自变量$x_1,\cdots,x_m$之间是否存在如模型（7.1）所示的线性关系是需要检验的，显然，如果所有的$|\hat{\beta}_j|$ $(j=1,\cdots,m)$都很小，y与$x_1,\cdots,x_m$的线性关系就不明显，所以可令原假设为

$$H_0:\beta_j=0,\quad j=1,\cdots,m.$$

当H_0成立时由分解式（7.15）定义的U,Q满足

$$F=\frac{U/m}{Q/(n-m-1)}\sim F(m,n-m-1), \qquad (7.16)$$

在显著性水平α下，对于上α分位数$F_\alpha(m,n-m-1)$，若$F < F_\alpha(m,n-m-1)$，接受H_0；否则拒绝。

注:接受H_0只说明y与$x_1,\cdots,x_m$的线性关系不明显，可能存在非线性关系，如平方关系。

还有一些衡量y与$x_1,\cdots,x_m$相关程度的指标，如用回归平方和在总平方和中的比值定义复判定系数

$$R^2 = \frac{U}{SST}. \qquad (7.17)$$

$R=\sqrt{R^2}$称为复相关系数，R越大，y与$x_1,\cdots,x_m$相关关系越密切，通常，R大于0.8（或0.9）才认为相关关系成立。

5.回归系数的假设检验和区间估计

当上面的H_0被拒绝时，β_j不全为零，但是不排除其中若干个等于零。所以应进一步作如下$m+1$个检验

$$H_0^{(j)}: \beta_j = 0, \quad j = 0,1,\cdots,m.$$

由（7.12）式，$\hat{\beta}_j \sim N(\beta_j, \sigma^2 c_{jj})$，$c_{jj}$是$(X^T X)^{-1}$中的第$(j,j)$元素，用$s^2$代替$\sigma^2$，由（7.12）~（7.14）式，当$H_0^{(j)}$成立时

$$t_j = \frac{\hat{\beta}_j / \sqrt{c_{jj}}}{\sqrt{Q/(n-m-1)}} \sim t(n-m-1). \qquad (7.18)$$

对给定的α，若$|t_j| < t_{\frac{\alpha}{2}}(n-m-1)$，接受$H_0^{(j)}$；否则拒绝。

（7.18）式也可用于对β_j作区间估计，在置信水平$1-\alpha$下，β_j的置信区间为

$$[\hat{\beta}_j - t_{\frac{\alpha}{2}}(n-m-1)s\sqrt{c_{jj}}, \hat{\beta}_j + t_{\frac{\alpha}{2}}(n-m-1)s\sqrt{c_{jj}}]$$

(7.19)

其中$s = \sqrt{\dfrac{Q}{n-m-1}}$。

6．利用回归模型进行预测

当回归模型和系数通过检验后，可由给定$[x_1,\cdots,x_m]$的取值$[a_{01},\cdots,a_{0m}]$预测y的取值b_0，b_0是随机的，显然其预测值（点估计）为

$$\hat{b}_0=\hat{\beta}_0+\hat{\beta}_1a_{01}+\cdots+\hat{\beta}_ma_{0m} \quad (7.20)$$

给定α可以算出b_0的预测区间(区间估计)，结果较复杂，但当n较大且a_{0i}接近平均值$\bar{x}_i$时，b_0的预测区间可简化为

$$[\hat{b}_0-z_{\frac{\alpha}{2}}s,\hat{b}_0+z_{\frac{\alpha}{2}}s], \quad (7.21)$$

其中$z_{\frac{\alpha}{2}}$是标准正态分布的上$\frac{\alpha}{2}$分位数。

对b_0的区间估计方法可用于给出已知数据残差$e_i = b_i - \hat{b}_i (i = 1, \cdots, n)$的置信区间，$e_i$服从均值为零的正态分布，所以若某个$e_i$的置信区间不包含零点，则认为这个数据是异常的，可予以剔除。

7.4.2　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)： $y=\beta_0+\beta_1 x_1+\cdots+\beta_m x_m$；

purequadratic(纯二次)：

$$y=\beta_0+\beta_1 x_1+\cdots+\beta_m x_m+\sum_{j=1}^{m}\beta_{jj}x_j^2;$$

interaction（交叉）：

$$y=\beta_0+\beta_1 x_1+\cdots+\beta_m x_m+\sum_{1\le j<k\le m}\beta_{jk}x_j x_k;$$

quadratic(完全二次)：

$$y=\beta_0+\beta_1 x_1+\cdots+\beta_m x_m+\sum_{1\le j\le k\le m}\beta_{jk}x_j x_k.$$

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

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

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 7.19　根据表 7.15（表略）某猪场 25 头育肥猪 4 个胴体性状的数据资料，试进行瘦肉量 y 对眼肌面积（x_1）、腿肉量(x_2)、腰肉量(x_3)的多元回归分析。

要求

（1）求y关于x_1, x_2, x_3的线性回归方程

$$y = c_0 + c_1 x_1 + c_2 x_2 + c_3 x_3,$$

计算c_0, c_1, c_2, c_3的估计值；

（2）对上述回归模型和回归系数进行检验（要写出相关的统计量）；

（3）试建立y关于x_1, x_2, x_3的二项式回归模型，并根据适当统计量指标选择一个较好的模型。

解　(1)记y, x_1, x_2, x_3的观察值分别为$b_i, a_{i1}, a_{i2}, a_{i3}$，$i = 1, 2, \cdots, 25$，

$$X = \begin{bmatrix} 1 & a_{11} & a_{12} & a_{13} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & a_{25,1} & a_{25,2} & a_{25,3} \end{bmatrix}, \quad Y = \begin{bmatrix} b_1 \\ \vdots \\ b_{25} \end{bmatrix}.$$

用最小二乘法求c_0, c_1, c_2, c_3的估计值，即应选取估计值$\hat{c}_j$，使当$c_j = \hat{c}_j$，$j = 0, 1, 2, 3$时，误差平方和

$$Q = \sum_{i=1}^{25} \varepsilon_i^2 = \sum_{i=1}^{25} (b_i - \hat{b}_i)^2 = \sum_{i=1}^{25} (b_i - c_0 - c_1 a_{i1} - c_2 a_{i2} - c_3 a_{i3})^2$$

达到最小。

为此，令

$$\frac{\partial Q}{\partial c_j}=0,\quad j=0,1,2,3,$$

得到正规方程组，求解正规方程组得c_0,c_1,c_2,c_3的估计值

$$[\hat{c}_0,\hat{c}_1,\hat{c}_2,\hat{c}_3]=(X^TX)^{-1}X^TY.$$

利用 Matlab 程序，求得

$$\hat{c}_0=0.8539,\quad \hat{c}_1=0.0178,\quad \hat{c}_2=2.0782,\quad \hat{c}_3=1.9396.$$

（2）因变量y与自变量x_1, x_2, x_3之间是否存在线性关系是需要检验的，显然，如果所有的$|\hat{c}_j|$ $(j=1,2,3)$都很小，y与x_1, x_2, x_3的线性关系就不明显，所以可令原假设为

$$H_0: c_j = 0, \quad j = 1,2,3. \qquad (7.22)$$

记 $m=3$ ， $n=25$ ， $Q=\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (b_i - \hat{b}_i)^2$ ，

$U=\sum_{i=1}^{n} (\hat{b}_i - \bar{b})^2$ ， 这里 $\hat{b}_i = \hat{c}_0 + \hat{c}_1 a_{i1} + \cdots + \hat{c}_m a_{im}$

$(i=1,\cdots,n)$， $\bar{b} = \frac{1}{n}\sum_{i=1}^{n} b_i$。

当H_0成立时统计量

$$F=\frac{U/m}{Q/(n-m-1)}\sim F(m,n-m-1),$$

在显著性水平α下，若

$$F_{1-\alpha/2}(m,n-m-1)<F<F_{\alpha/2}(m,n-m-1),$$

接受H_0；否则拒绝。

利用 Matlab 程序求得统计量$F=37.7453$，查表得上$\alpha/2$分位数$F_{0.025}(3,21)=3.8188$，因而拒绝（7.22）式的原假设，模型整体上通过了检验。

当（7.22）式的H_0被拒绝时，β_j不全为零，但是不排除其中若干个等于零。所以应进一步作如下$m+1$个检验

$$H_0^{(j)}: c_j = 0, \quad j = 0,1,\cdots,m, \qquad (7.23)$$

当$H_0^{(j)}$成立时

$$t_j = \frac{\hat{\beta}_j / \sqrt{c_{jj}}}{\sqrt{Q/(n-m-1)}} \sim t(n-m-1),$$

这里c_{jj}是$(X^TX)^{-1}$中的第(j,j)元素，对给定的α，若$|t_j| < t_{\frac{\alpha}{2}}(n-m-1)$，接受$H_0^{(j)}$；否则拒绝。

利用 Matlab 程序，求得统计量

$t_0 = 0.6223$，$t_1 = 0.6090$，$t_2 = 7.7407$，$t_3 = 3.8062$，

查表得上$\alpha/2$分位数$t_{0.025}(21) = 2.0796$。

对于（7.23）式的检验，在显著性水平$\alpha = 0.05$时，接受$H_0^{(j)}: c_j = 0$（$j = 0,1$），拒绝$H_0^{(j)}: c_j = 0$（$j = 2,3$），即变量x_1对模型的影响是不显著的。建立线性模型时，可以不使用x_1。

注： ⅰ）在 regress 的第 5 个返回值中，就包含F统计量的值，不需单独计算。

ⅱ）regress 的返回值中不包括t统计量的值，如果需要则要单独计算。由于假设检验和参数的区间估计是等价的，regress 的第 2 个返回值是各参数的区间估计，如果某参数的区间估计包含 0 点，则该参数对应的变量是不显著的。

（3）我们使用 Matlab 的用户图形界面解法求二项式回归模型。根据剩余标准差（rmse）这个指标选取较好的模型是完全二次模型，模型为

$$y = -17.0988 + 0.3611x_1 + 2.3563x_2 + 18.2730x_3 - 0$$

$$- 0.4404x_1x_3 - 1.2754x_2x_3 + 0.0217x_1^2 + 0.5025x_2^2$$

7.4.3 非线性回归

非线性回归是指因变量y对回归系数$\beta_1,\cdots,\beta_m$（而不是自变量）是非线性的。Matlab 统计工具箱中的命令nlinfit，nlparci，nlpredci，nlintool，不仅给出拟合的回归系数及其置信区间，而且可以给出预测值及其置信区间等。下面通过例题说明这些命令的用法。

例7.20　在研究化学动力学反应过程中，建立了一个反应速度和反应物含量的数学模型，形式为

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

其中$\beta_1,\cdots,\beta_5$是未知的参数，x_1,x_2,x_3是三种反应物(氢，n戊烷，异构戊烷）的含量，y是反应速度。今测得一组数据如表7.16所列，试由此确定参数$\beta_1,\cdots,\beta_5$，并给出其置信区间。$\beta_1,\cdots,\beta_5$的参考值为[0.1, 0.05, 0.02, 1, 2]。

表7.16　反应数据

序号	反应速度y	氢x_1	n戊烷x_2	异构戊烷x_3
1	8.55	470	300	10
2	3.79	285	80	10
3	4.82	470	300	120
4	0.02	470	80	120
5	2.75	470	80	10
6	14.39	100	190	10
7	2.54	100	80	65
8	4.35	470	190	65
9	13.00	100	300	54
10	8.50	100	300	120
11	0.05	100	80	120
12	11.32	285	300	10
13	3.13	285	190	120

解　首先，以回归系数和自变量为输入变量，将要拟合的模型写成匿名函数。然后，用 nlinfit 计算回归系数，用 nlparci 计算回归系数的置信区间，用 nlpredci 计算预测值及其置信区间。

用nlintool得到一个交互式画面，左下方的Export可向工作空间传送数据，如剩余标准差等。使用命令

nlintool(x,y,huaxue,beta0)

（注意这里 x、y、huaxue、beta0 必须在工作空间中，也就是说要把上面的程序运行一遍，再运行 nlintool）可看到画面，并向工作空间传送有关数据，例如剩余标准差 rmse=0.1933。

7.5 基于灰色模型和 Bootstrap 理论的大规模定制质量控制方法研究

7.5.1 引言

随着全球竞争的加剧和市场细分程度的提升，大规模定制生产方式越来越受到重视。大规模定制是在大规模生产的基础上，通过产品结构和制造过程的重组，运用现代信息技术、新材料技术、制造技术等一系列手段，以接近大规模生产的成本和速度，为单个顾客或小批量、多品种市场定制任意数量产品的一种生产方式。大规模定制生产过程质量控制与大批量生产过程比较具有以下新的特点：

（1）样本量较小，尤其是在定制化程度较高的情况下和生产的初级阶段；（2）样本数列往往具有时变性，不能简单假设其服从正态分布；（3）大规模定制生产模式要求灵活性和快速性，而传统的质量控制方法响应速度偏慢。因此，应用于大批量生产模式下的经典休哈特控制图便不再适用。

大规模定制生产尤其是新产品初期质量控制中的一个突出的问题是样本量不足，无法确定样本数据的统计分布，不能得到过程分布参数的真值，因此也无法构建出相应的控制图。上述问题的研究可以细分为两个方面：如何有效地拓展样本数量，以有助于分析样本数据的分布规律；如何获得统计量的分布，进而估计过程参数和构建控制图。

张炎亮，樊树海等研究了灰色神经网络模型在大规模定制生产中的应用；贺云花等、Raviwongse & Allada、杜尧研究了成组技术在柔性制造和小批量生产中的应用；Suykens & Vandewalle，孙林&杨世元分析了支持向量机模型的原理，研究了其在小批量及柔性生产质量预测中的应用；

吴德会研究了小批量生产中基于动态指数平滑模型的过程质量预测；李奔波对工序能力等级的判定方法进行了改进，对大规模定制环境下的单值控制图和中位数控制图进行了研究；王晶等将 Bootstrap 方法引入到多品种小批量生产的质量控制中，研究了基于该方法的控制图的构造和实施流程。

以上文献对可以用于大规模定制生产中的质量控制方法进行了研究，提出的灰色神经组合预测模型固然有很多优点，但在大规模定制实际生产中工序质量数据的样本量很小，无法获得足够的数据用于神经网络的训练，其训练效果及可信度难以保证。本文提出的基于灰色模型和 Bootstrap 的集成方法，在质量数据预测与统计推断方面具有优势，可以有效解决大规模定制生产中样本量小带来的研究局限性。

首先，灰色模型在极小样本量情况下进行质量数据预测具有独特的优势，预测效果也相对较好；其次，基于 Bootstrap 理论的统计推断方法通过重复抽样，能对未知分布的随机变量的分布参数进行较为精确的区间估计，为构建质量控制图提供依据。

7.5.2 基于灰色模型的大规模定制生产质量预测

灰色理论可充分开发并利用少量数据中的显信息和隐信息，根据行为特征数据找出因素本身或因素之间的数学关系，提取建模所需变量，通过建立离散数据的微分方程动态模型，了解系统的动态行为和发展趋势。灰色模型有以下优点：

（1）所需信息量较少（一般有 4 个以上数据即可建模）；（2）不需要知道原始数据分布的先验特征，通过有限次的生成，可将无规则分布（或服从任意分布）的任意光滑离散的原始序列转化为有序序列；(3) 可保持原系统特征，能较好地反映系统实际情况。因此，适用于大规模定制生产的质量预测分析。

由题意列出方差分析表见表 7.14。

灰色模型GM(1,1)是灰色系统理论中较常用的预测模型，基于该模型的质量指标预测建模步骤如下：

1）原始质量指标数列为

$$x^{(0)} = (x^{(0)}(1), x^{(0)}(2), \cdots, x^{(0)}(n)).$$

2） $x^{(1)}$是$x^{(0)}$的累加序列为

$$x^{(1)} = (x^{(0)}(1), \sum_{i=1}^{2} x^{(0)}(i), \cdots, \sum_{i=1}^{n} x^{(0)}(i)).$$

经过该处理，可使粗糙的原始离散数列变为光滑的离散数列。

3）建立基本预测模型GM(1,1)，其白化方程为

$$\frac{dx^{(1)}}{dt}+ax^{(1)}=b, \tag{7.24}$$

式（7.24）中，a,b为常系数，且符合

$$[a,b]^T=(B^TB)^{-1}B^TY, \tag{7.25}$$

$$B=\begin{bmatrix}-z^{(1)}(2) & 1\\ -z^{(1)}(3) & 1\\ \vdots & \vdots\\ -z^{(1)}(n) & 1\end{bmatrix},\ Y=\begin{bmatrix}x^{(0)}(2)\\ x^{(0)}(3)\\ \vdots\\ x^{(0)}(n)\end{bmatrix}, \tag{7.26}$$

式（7.26）中，$z^{(1)}(k)=0.5x^{(1)}(k)+0.5x^{(1)}(k-1)$。

4）对建立的GM(1,1)预测模型进行精度检验和评估。检验依据后验差比值c和小误差概率p两个指标，模型精度等级见表 7.17。其中c和p定义如下:

$$c = \frac{s_2}{s_1}, \tag{7.27}$$

$$p = P\left\{\left|q^{(0)}(k) - q\right| < 0.6475 s_1\right\}, \tag{7.28}$$

其中，$q^{(0)}$为残差序列，q为残差序列的均值，s_1为原始序列的标准差，s_2为残差序列的标准差。

表 7.17　灰色模型预测精度等级

等级精度	后验差比值c	小误差概率p
A（好）	≤ 0.35	≥ 0.95
B（合格）	$0.35 < c \leq 0.50$	$0.80 \leq p < 0.95$
C（勉强）	$0.50 < c \leq 0.65$	$0.70 \leq p < 0.80$
D（不合格）	> 0.65	< 0.70

如果精度不合要求，可以用残差序列建立GM(1,1)模型对原模型进行修正，以提高其精度，若GM(1,1)模型满足精度要求时，其还原数据与预测值见公式（7.29）和（7.30）。

$$\hat{x}^{(1)}(t+1)=[x^{(0)}(1)-b/a]e^{-at}+b/a, \qquad (7.29)$$

$$\hat{x}^{(0)}(t+1)=\hat{x}^{(1)}(t+1)-\hat{x}^{(1)}(t)=(1-e^{a})[x^{(0)}(1)-b/a]e^{-at} \qquad (7.30)$$

若要进一步提高预测精度，可采用GM(1,1)新陈代谢模型。首先采用原始序列建立一个GM(1,1)模型，按上述方法求出一个预测值，然后将该预测值补入已知数列中，同时去除一个最旧的数据；在此基础上再建立GM(1,1)模型，求出下一个预测值，以此类推，通过预测灰数的新陈代谢，逐个预测，依次递补，可以得到之后几期的数据，对原始数据数量进行有效扩充。

7.5.3　基于 Bootstrap 理论的过程质量分析

在大规模定制生产模式中，能采集到的质量数据十分有限，即便经过上述的灰色模型预测，样本量仍不能满足分布参数估计的要求。Bootstrap 方法可以通过重复抽样，获得一定规模的样本量，进而得到统计量的经验分布并进行区间估计。Bootstrap 理论由 Efron 于 1979 年提出，是一种新的增广样本统计方法。它的无先验性，以及计算过程中只需要有限的观测数据，使其可方便地应用于小样本数据处理。

1.Bootstrap 方法的数学描述

设$X=(x_1,x_2,\cdots,x_n)$是来自于某个未知总体F的样本，$R(X,F)$是总体分布F的某个分布特征。根据观测样本$X=(x_1,x_2,\cdots,x_n)$估计$R(X,F)$的某个参数（如均值，方差或分布密度函数等）。例如，设$\theta=\theta(F)$为总体分布F的某个参数，F_n是观测样本X的经验分布函数，$\hat{\theta}=\hat{\theta}(F_n)$是$\theta$的估计，记估计误差为

$$R(X,F)=\hat{\theta}(F_n)-\theta(F)\underline{\underline{\Delta}}T_n. \qquad (7.31)$$

由观测样本$X=(x_1,x_2,\cdots,x_n)$估计$R(X,F)$的分布特征，显然此时$R(X,F)$的均值和方差分别为$\theta(F)$估计误差的均值和方差。Bootstrap 方法的实质就是再抽样过程，通过对观测数据的重新抽样产生再生样本来模拟总体分布。计算$R(X,F)$分布特征的基本步骤如下：

1）根据观测样本$X=(x_1,x_2,\cdots,x_n)$构造经验分布函数F_n；

2）从F_n中抽取样本$X^*=(x_1^*,x_2^*,\cdots,x_n^*)$，称其为Bootstrap样本；

3）计算相应的Bootstrap统计量$R^*(X^*,F_n)$，其表达式为

$$R^*(X^*,F_n)=\hat{\theta}(F_n^*)-\hat{\theta}(F_n)\underline{\Delta}R_n. \qquad (7.32)$$

式(7.32)中，F_n^*是Bootstrap样本的经验分布函数；R_n是T_n的Bootstrap统计量；

4）重复 2)、3）B次，即可得到 Bootstrap 统计量 $R^*(X^*,F_n)$的B个可能取值,将统计量的值从小到大排列即为样本统计量的 Bootstrap 经验分布;

5)用$R^*(X^*,F_n)$的分布去逼近$R(X,F)$的分布,即用R_n的分布去近似T_n的分布，可得到参数$\theta(F)$的B个可能取值，即可统计求出参数θ的分布及其特征值。

采用上述 Bootstrap 方法作统计分析的目的在于获得所估计参数的置信区间。当置信水平为$1-\alpha$时，置信区间上限为经验分布的$1-\alpha/2$分位数，下限为经验分布的$\alpha/2$分位数。

由以上分析可知，Bootstrap 经验分布的一般特性如下：（1）经验分布集中在样本统计量T周围；（2）经验分布的均值是统计量T所有可能样本抽样分布的均值估计；（3）经验分布的标准差是统计量T的标准差估计；（4）经验分布的$\alpha/2$和$1-\alpha/2$分位数分别为$1-\alpha$置信水平下统计量T的置信区间的下限和上限。

2.基于 Bootstrap 的质量控制图分析

接下来分析采用 Bootstrap 方法对大规模定制生产过程进行质量控制的过程，见图 7.3，具体步骤为：4）对建立的

1）对原始数据重复抽样，得到一定数量的子样本；

2）对每个子样本计算相关的统计量；

3）将子样本的统计量按从小到大排序，得到 Bootstrap 经验分布；

4）根据控制图的控制限要求，上下限取 Bootstrap 经验分布的相应分位数，构建样本统计量控制图。

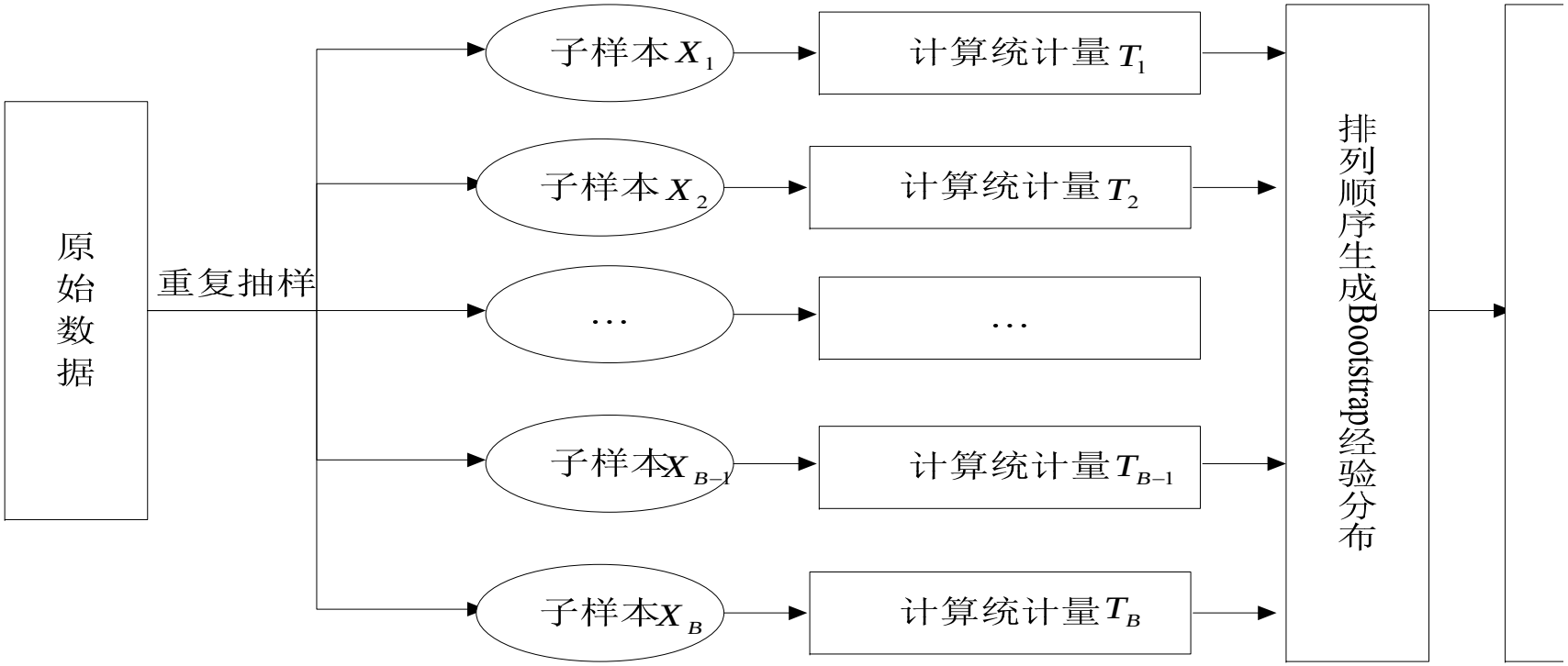

图 7.3　基于 Bootstrap 的质量控制图分析示意图

在具体实施中，需要考虑原始观测样本的样本量以及抽样次数。张湘平研究提出，要保证应用 Bootstrap 方法进行估计的有效性，至少要有 8 个观察值，相关文献中提出根据实际情况，观测样本越多越好；Efron 和 Tibshirani 研究提出重复抽样次数B一般取 1000~3000。

7.5.4 案例分析

某航空产品制造厂生产的一批框段根据技术指标及安装位置的差异性，其半径、形状、连接方式各不相同，共有 20 种，且每一种产品批生产数量都不超过 30 件，其生产方式可视为大规模定制模式。

其中一种框段（钣金零件）厚度要求为$\Phi 2.60^{+0.1}_{-0.1}$（mm），在生产初期采集了 10 件的加工数据，测得的质量数据为 2.5320，2.6470，2.6290，2.5840，2.6090，2.6010，2.5280，2.5630，2.6540，2.6190。根据本文的研究思路，对该零件的加工数据进行质量分析的流程如图 7.4 所示。

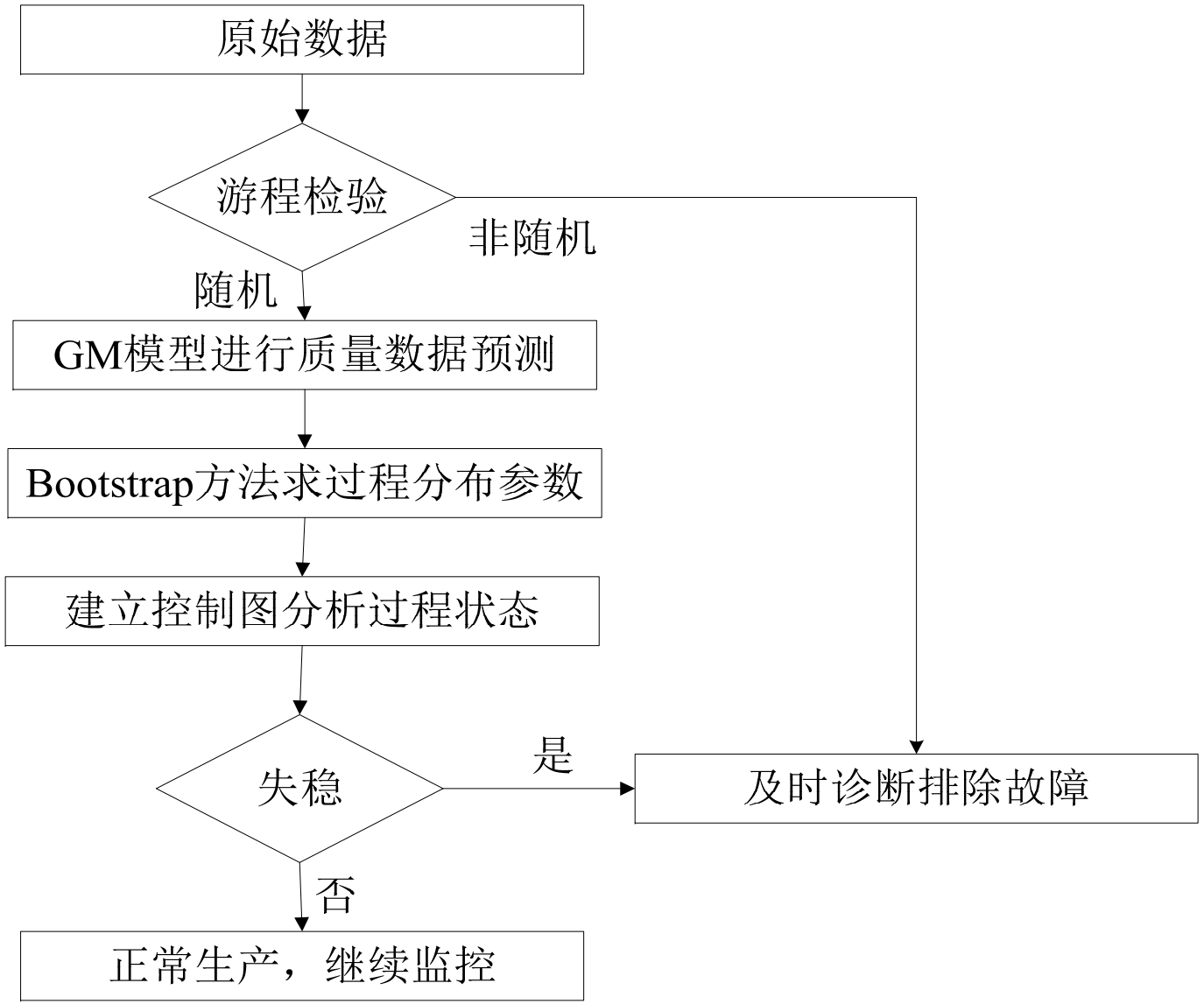
原始数据
游程检验
非随机
随机
GM模型进行质量数据预测
Bootstrap方法求过程分布参数
建立控制图分析过程状态
失稳
是
及时诊断排除故障
否
正常生产，继续监控

图 7.4　质量分析流程图

1）游程检验亦称“连贯检验”，是根据样本容量和游程的多少来判定一个给定样本序列是否排列随机的检验方法。采用该方法对初始样本数据的随机性进行检验。若排列随机，表明初始样本对应的生产过程状态正常，进行质量数据预测有意义。若排列非随机，应根据 5M1E 及时检查并纠正系统性质量因素。

本案例中，初始样本数列

$x^{(0)} = (2.5320, 2.6470, 2.6290, 2.5840, 2.6090, 2.5280, 2.56$

。

中位数 $Me = (2.6010 + 2.6090)/2 = 2.6050$，数列中样本取值若大于 Me 的记为“1”，小于 Me 的记为“0”，由此产生 0-1 数列“0110100011”。统计结果：3 个“0”游程，3 个“1”游程，共有 $U = 6$ 个游程；“1”的总个数 $n_1 = 5$，“0”的总个数 $n_2 = 5$，查游程检验临界值表可知，在 0.05 显著性水平上该样本序列排列是随机的，说明初始样本对应的生产过程状态是正常的。

2）采用灰色系统模型，用 Matlab 编程求解白化方程参数并求预测值，在此过程中检验模型精度。经计算得，后验差比值c为 0.5348，与精度等级表 12 比较，模型精度满足要求。

取之后 6 期的预测值

2.5904　2.5877　2.5850　2.5823　2.5797　2.5770，

扩充样本数量，得到新的样本数列$\tilde{x}^{(0)}$如下

2.5320，2.6470，2.6290，2.5840，2.6090，2.6010，2.5280，2.5630，2.6540，2.6190，2.5904，2.5877，2.5850，2.5823，2.5797，2.5770.

3）以新的样本序列$\tilde{x}^{(0)}$中数据为原始数据，按上节所述步骤，计算样本均值和样本极差的3σ控制限。依时间顺序将$\tilde{x}^{(0)}$分为 4 个子样本组，如表 7.18 所示。

表 7.18　框段厚度检测值（mm）

子样本	1	2	3	4
检测值	2.532	2.609	2.654	2.585
	2.647	2.601	2.619	2.5823
	2.629	2.528	2.5904	2.5797
	2.584	2.563	2.5877	2.577
均值$\bar{X}$	2.598	2.5753	2.6128	2.581
极差R	0.115	0.081	0.0663	0.008

在 Matlab 中用 Bootstrap 工具箱，基于样本数列$\tilde{x}^{(0)}$进行重复抽样，子样本容量为 4，抽取 1000 个子样本，计算子样本均值和极差。

4）将得到的 1000 个子样本的均值和极差按从小到大的顺序排列，则得到 Bootstrap 经验分布。当采用3σ控制图时，取显著性水平$\alpha = 0.27\%$，上下控制限分别为经验分布的$(1-\alpha/2)$分位数和$\alpha/2$分位数。因此，获得样本均值控制图的上下控制限分别为$LCL_{\bar{x}} = 2.5388$和$UCL_{\bar{x}} = 2.6390$，样本极差控制图的上下控制限分别为$LCL_R = 0.0017$和$UCL_R = 0.1260$。

注：由于是 bootstrap 抽样，每次的计算结果是不一样的。

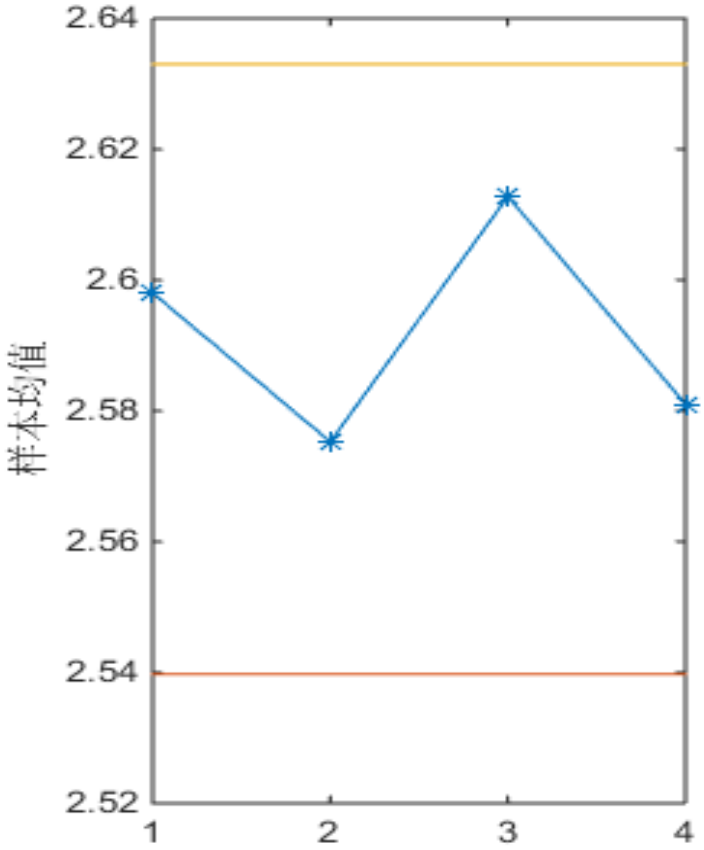

图 7.5　均值控制图

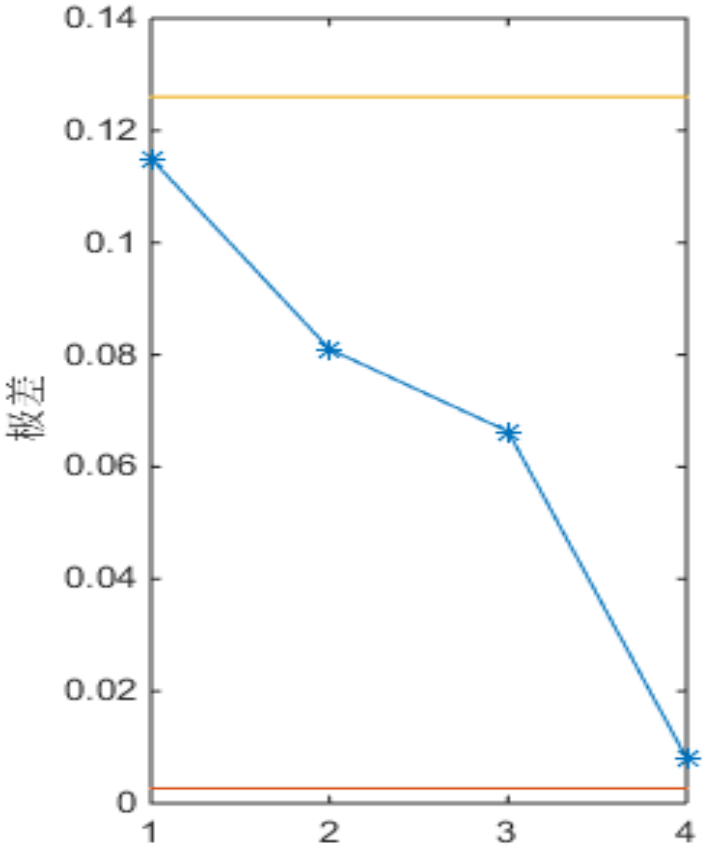

图 7.6　极差控制图

5）根据样本均值和样本极差控制图上下限和表 7.18 中的样本统计量取值，分别绘制样本均值控制图（图 7.5）和样本极差控制图（图 7.6），观察样本统计量数据点均在控制界限内，可以判断生产过程受控。

7.5.4 结论

鉴于灰色模型在极小样本量数据预测中的优势，采用灰色系统预测模型对样本数量进行扩展，再基于Bootstrap 统计推断方法得到统计量的经验分布，采用$\bar{X}$控制图和R控制图对大规模定制生产中的过程状态进行判断。通过对框段生产线大规模定制生产质量控制的案例研究，本文提出的新方法可用于大规模定制生产尤其是新型号生产初期的质量预测和控制。