第3章

关于总体分布的检验和独立性检验

学习要点

1. 理解多元中心极限定理。
2. 掌握拟合优度检验的方法。
3. 掌握一些常用的正态性检验方法。
4. 掌握独立性检验的方法。
5. 综合应用本章所学方法完成实际应用问题的总体分布和独立性检验。

3.1 拟合优度检验

3.1.1 多项分布的中心极限定理

设总体 X 服从 k 项分布

$$P\{X=x_j\}=p_j, \qquad j=1,2,\cdots,k \tag{3.1}$$

对 X 做 n 次简单抽样, 记 x_j 出现的次数为 n_j（称为**频数**）, 它是一个服从 Bernoulli 分布的随机变量, 不难得到 $E(n_j)=np_j, D(n_j)=np_j(1-p_j)$。$x_j$ 出现的**频率**为 n_j/n, 记

$$Y^{(n)}=(Y_1^{(n)},Y_2^{(n)},\cdots,Y_k^{(n)})^{\mathrm{T}}, \qquad Y_j^{(n)}=\sqrt{n}\left(\frac{n_j}{n}-p_j\right), \qquad j=1,2,\cdots,k \tag{3.2}$$

我们想知道当 $n\to\infty$ 时随机变量 $Y^{(n)}$ 的极限分布。

设 $X_1,X_2,\cdots,X_n$ 是来自总体的简单样本, 对每个 X_i 定义随机向量 $\eta^{(i)}=(\eta_1^{(i)},\eta_2^{(i)},\cdots,\eta_k^{(i)})^{\mathrm{T}}$ 如下:

如果 $X_i=x_j$, 则定义

$$\eta_j^{(i)}=\frac{1}{\sqrt{n}}, \qquad \eta_{j'}^{(i)}=0, \quad \forall\, j'\neq j$$

则有

$$Y^{(n)}=\sum_{i=1}^{n}\eta^{(i)}-\sqrt{n}p \tag{3.3}$$

其中 $p=(p_1,p_2,\cdots,p_k)^{\mathrm{T}}$。注意到 $\eta^{(i)}$ 的矩母函数为

$$M_{\eta^{(i)}}(t)=E\left[\mathrm{e}^{t^{\mathrm{T}}\eta^{(i)}}\right]=\sum_{j=1}^{k}p_j\mathrm{e}^{\frac{t_j}{\sqrt{n}}} \tag{3.4}$$

因此

$$M_{Y^{(n)}}(t)=\mathrm{e}^{-\sqrt{n}t^{\mathrm{T}}p}\prod_{i=1}^{n}M_{\eta^{(i)}}(t)=\mathrm{e}^{-\sqrt{n}t^{\mathrm{T}}p}\left(\sum_{j=1}^{k}p_j\mathrm{e}^{\frac{t_j}{\sqrt{n}}}\right)^n \tag{3.5}$$

注意到

$$\mathrm{e}^{\frac{t_j}{\sqrt{n}}}=1+\frac{t_j}{\sqrt{n}}+\frac{t_j^2}{2n}+o\left(\frac{1}{n}\right) \tag{3.6}$$

因此有

$$\begin{aligned}\sum_{j=1}^{k}p_j\mathrm{e}^{\frac{t_j}{\sqrt{n}}}&=\sum_{j=1}^{k}p_j+\frac{1}{\sqrt{n}}\sum_{j=1}^{k}p_jt_j+\frac{1}{2n}\sum_{j=1}^{k}p_jt_j^2+o\left(\frac{1}{n}\right)\\&=1+\frac{1}{\sqrt{n}}t^{\mathrm{T}}p+\frac{1}{2n}\sum_{j=1}^{k}p_jt_j^2+o\left(\frac{1}{n}\right)\end{aligned} \tag{3.7}$$

再注意到

$$\ln(1+u)=u-\frac{1}{2}u^2+o(u^2),\qquad u\to 0$$

因此有

$$\begin{aligned}\ln\left(\sum_{j=1}^{k}p_j\mathrm{e}^{\frac{t_j}{\sqrt{n}}}\right)&=\ln\left(1+\frac{1}{\sqrt{n}}t^{\mathrm{T}}p+\frac{1}{2n}\sum_{j=1}^{k}p_jt_j^2+o\left(\frac{1}{n}\right)\right)\\&=\frac{1}{\sqrt{n}}t^{\mathrm{T}}p+\frac{1}{2n}\sum_{j=1}^{k}p_jt_j^2+o\left(\frac{1}{n}\right)\\&\quad-\frac{1}{2}\left(\frac{1}{\sqrt{n}}t^{\mathrm{T}}p+\frac{1}{2n}\sum_{j=1}^{k}p_jt_j^2+o\left(\frac{1}{n}\right)\right)^2+o\left(\frac{1}{n}\right)\\&=\frac{1}{\sqrt{n}}t^{\mathrm{T}}p+\frac{1}{2n}\sum_{j=1}^{k}p_jt_j^2-\frac{1}{2n}t^{\mathrm{T}}pp^{\mathrm{T}}t+o\left(\frac{1}{n}\right)\end{aligned} \tag{3.8}$$

由此得到

$$\left(\sum_{j=1}^{k}p_j\mathrm{e}^{\frac{t_j}{\sqrt{n}}}\right)^n=\exp\left\{n\ln\left(\sum_{j=1}^{k}p_j\mathrm{e}^{\frac{t_j}{\sqrt{n}}}\right)\right\}$$

$$= \exp\left\{\sqrt{n}t^{\mathrm{T}}p + \frac{1}{2}\sum_{j=1}^{k} p_j t_j^2 - \frac{1}{2}t^{\mathrm{T}}pp^{\mathrm{T}}t + o(1)\right\} \tag{3.9}$$

$$M_{Y^{(n)}}(t) = \exp\left\{\frac{1}{2}\sum_{j=1}^{k} p_j t_j^2 - \frac{1}{2}t^{\mathrm{T}}pp^{\mathrm{T}}t + o(1)\right\} \tag{3.10}$$

因此有

$$\begin{aligned}\lim_{n\to\infty} M_{Y^{(n)}}(t) &= \exp\left\{\frac{1}{2}\sum_{j=1}^{k} p_j t_j^2 - \frac{1}{2}t^{\mathrm{T}}pp^{\mathrm{T}}t\right\} \\ &= \exp\left\{\frac{1}{2}t^{\mathrm{T}}\left(\mathrm{diag}(p) - pp^{\mathrm{T}}\right)t\right\}\end{aligned} \tag{3.11}$$

其中 $\mathrm{diag}(p)$ 表示以向量 p 中的元素作主对角元素的对角矩阵。由此可以看出, 当 $n \to \infty$ 时, $Y^{(n)}$ 的极限分布是 k 维正态分布, 均值向量为 0, 协方差矩阵为

$$\Sigma = \mathrm{diag}(p) - pp^{\mathrm{T}} = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_k \\ -p_2p_1 & p_2(1-p_2) & \cdots & -p_2p_k \\ \vdots & \vdots & \ddots & \vdots \\ -p_kp_1 & -p_kp_2 & \cdots & p_k(1-p_k) \end{pmatrix} \tag{3.12}$$

需要指出的是, Σ 不是满秩矩阵, 因为它的 k 个列向量之和为 0。

综上所述, 我们证明了如下中心极限定理。

定理 3.1 设总体 X 服从 k 项分布 (3.1), $Y^{(n)}$ 和 $Y_j^{(n)}$ 由式 (3.2) 给出, 则当 $n \to \infty$ 时, k 维随机向量 $Y^{(n)}$ 的极限分布是均值向量为 0 的正态分布, 其协方差矩阵 Σ 由式 (3.12) 给出。

3.1.2 拟合优度检验

设总体 X 服从 k 项分布式 (3.1), 其中参数 $p_j, j = 1, 2, \cdots, k$ 未知, 需要检验如下假设:

$$H_0: \qquad p_j = p_j^o, \qquad j = 1, 2, \cdots, k \tag{3.13}$$

对 X 做 n 次简单抽样, 设 x_j 出现的频数为 n_j, 频率为 n_j/n。根据大数定律, 当 $n \to \infty$ 时, 频率 n_j/n 将以概率 1 趋于概率 p_j。但由于实际样本容量是有限的, 频率 n_j/n 与概率 p_j 总是有偏差, 因此不能简单地通过看 n_j/n 与 p_j^o 是否相等判断原假设 H_0 是否成立。为了解决这种假设检验问题, Pearson 提出了下列统计量[47]:

$$K^2 := \sum_{j=1}^{k} \frac{\left(\frac{n_j}{n} - p_j\right)^2}{\frac{p_j}{n}} = \sum_{j=1}^{k} \frac{(n_j - np_j)^2}{np_j} \tag{3.14}$$

Pearson 认为这个统计量是实际抽样数据对理论分布的拟合优度的度量, 因此这类假设检验称为**拟合优度检验 (goodness-of-fit test)**。利用式 (3.2) 中定义的变量 $Y^{(n)}$ 和 $Y_j^{(n)}$ 还可以将 K^2 表示为

$$K^2 := \sum_{j=1}^{k} \left(\frac{Y_j^{(n)}}{\sqrt{p_j}} \right)^2 \tag{3.15}$$

研究统计量 K^2 的极限分布时需要用到这种表示。

记

$$Z_j^{(n)} = \frac{Y_j^{(n)}}{\sqrt{p_j}}, \qquad Z^{(n)} = (Z_1^{(n)}, Z_2^{(n)}, \cdots, Z_k^{(n)})^{\mathrm{T}} \tag{3.16}$$

则 $Z^{(n)}$ 的极限分布是 $\mathcal{N}(0, Q)$, 其中

$$Q = \begin{pmatrix} 1-p_1 & -\sqrt{p_1 p_2} & \cdots & -\sqrt{p_1 p_k} \\ -\sqrt{p_2 p_1} & 1-p_2 & \cdots & -\sqrt{p_2 p_k} \\ \vdots & \vdots & \ddots & \vdots \\ -\sqrt{p_k p_1} & -\sqrt{p_k p_2} & \cdots & 1-p_k \end{pmatrix} \tag{3.17}$$

记 $v_1 = (\sqrt{p_1}, \sqrt{p_2}, \cdots, \sqrt{p_k})^{\mathrm{T}}$, 则有 $Q = I - v_1 v_1^{\mathrm{T}}$

引理 3.1 设 $v_1 = (\sqrt{p_1}, \sqrt{p_2}, \cdots, \sqrt{p_k})^{\mathrm{T}}$, 矩阵 Q 由式 (3.17) 定义, 则存在 $v_2, v_3, \cdots, v_k$, 使得 $v_1, v_2, v_3, \cdots, v_k$ 构成正交规范向量组, 且有

$$Q = \sum_{j=2}^{k} v_j v_j^{\mathrm{T}} \tag{3.18}$$

从而 Q 的秩为 $k-1$。

证明 将 v_1 扩充为 $\mathbb{R}^k$ 的规范正交基 $v_1, v_2, \cdots, v_k$, 则有

$$\begin{aligned} Q &= I - v_1 v_1^{\mathrm{T}} = \sum_{j=1}^{k} v_j v_j^{\mathrm{T}} - v_1 v_1^{\mathrm{T}} = \sum_{j=2}^{k} v_j v_j^{\mathrm{T}} \\ &= V \begin{pmatrix} 0 & 0 \\ 0 & I_{n-1} \end{pmatrix} V^{\mathrm{T}} \end{aligned} \tag{3.19}$$

其中 $V = (v_1, v_2, \cdots, v_n)$ 是正交矩阵, I_{n-1} 是 $n-1$ 阶单位矩阵。式 (3.19) 就是 Q 的特征分解。由此可以看出, Q 有 $k-1$ 个特征值为 1, 一个特征值为 0, 因此 $\mathrm{rank}(Q) = k-1$。 □

定理 3.2 在由式 (3.13) 定义的原假设 H_0 成立的条件下, 当样本容量 n 充分大时, 由式 (3.14) 定义的统计量 K^2 近似服从 $\chi^2(k-1)$ 分布。

证明　设 $Z_j^{(n)}$ 及 $Z^{(n)}$ 由式 (3.16) 定义, 则当 n 充分大时, $Z^{(n)}$ 近似服从 $\mathcal{N}(0,Q)$ 分布, 且

$$K^2=\sum_{j=1}^{k}\left(Z_j^{(n)}\right)^2 \tag{3.20}$$

设 $v_1=(\sqrt{p_1},\sqrt{p_2},\cdots,\sqrt{p_k})^{\mathrm{T}}$, 根据引理 3.1, 存在 $v_2,v_3,\cdots,v_k$, 使得 $v_1,v_2,\cdots,v_k$ 构成正交规范向量组, 且式 (3.19) 成立。设 V 是式 (3.19) 中最后一个等号右边的正交矩阵, Λ 是其中的分块矩阵, 令 $\eta=V^{\mathrm{T}}Z^{(n)}$, 则 η 近似服从 $\mathcal{N}(0,\Lambda)$, 可见随机向量 η 的第一个分量是 0, 第 $2\sim k$ 个分量是近似独立的、服从标准正态分布的随机变量, 从而 $\|\eta\|_2^2$ 近似服从自由度为 $k-1$ 的卡方分布 $\chi^2(k-1)$。又因为 V 是正交矩阵, 因此有

$$K^2=\|Z^{(n)}\|_2^2=\|V^{\mathrm{T}}Z^{(n)}\|_2^2=\|\eta\|_2^2$$

由此推出 K^2 近似服从 $\chi^2(k-1)$ 分布。　□

定理 3.2 只说了当样本容量 n 充分大时, Pearson 统计量 K^2 近似服从 $\chi^2(k-1)$ 分布, 但并没有说明 n 到底要多大才行。实践表明, 在大多数应用场景中, 当 $n\geqslant 50$ 时, 用卡方分布近似代替 K^2 的分布检验效果良好。

例 3.1　将一颗骰子连续掷 400 次, 各种点数出现的次数如表 3.1 所示, 试问这颗骰子是否质地均匀。

表 3.1　掷 100 次骰子结果统计

点　　数	1	2	3	4	5	6
出现次数	61	72	91	59	64	53

解　用 X 表示随机掷骰子掷出的点数, 如果这颗骰子质地均匀, 则各种点数出现的概率是一样的, 因此本题实际上是要检验下列原假设:

$$H_0:\quad p_j=P\{X=j\}=\frac{1}{6},\qquad j=1,2,3,4,5,6 \tag{3.21}$$

根据定理 3.2, 在原假设成立的条件下, Pearson 统计量

$$K^2=\sum_{j=1}^{6}\frac{(n_j-np_j)^2}{np_j}$$

近似服从自由度为 $\nu=6-1=5$ 的卡方分布 $\chi^2(5)$, 利用表 3.1 中的数据计算得到 $K^2=13.58$, 取显著性水平为 $\alpha=0.05$, 用 MATLAB 计算自由度为 5 的卡方分布的分位点, 得到 $\chi_\alpha^2(5)=11.0705$, 由于 $K^2>\chi_\alpha^2(5)$, 因此拒绝原假设, 即认为这个骰子质地不均匀, 各种点数出现的概率不平等。

计算过程的 MATLAB 代码如下:

```
function Li3_1()
%%这个函数实现了例3.1中Pearson卡方分布检验的计算
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
N=400;                          %%掷骰子的总次数
n=[61,72,91,59,64,53];          %%各种点数出现的次数
k=size(n,2)
p=(1/6)*ones(1,6);              %%原假设成立时各种点数出现的概率
K2=sum(((n-N*p).^2)./(N*p))     %%计算Pearson统计量
a=0.05                          %%显著性水平
Ka=chi2inv(1-a,k-1)             %%计算显著性水平为α的分位点
end
```

例 3.2 某服务台最近 100 天每天接待的售后服务次数统计如表 3.2 所示, 请检验该服务台每天接待的售后服务次数 X 是否服从参数为 $\lambda = 1$ 的泊松分布（取显著性水平为 $\alpha = 0.05$）。

表 3.2 服务台每天接待的售后服务次数统计

接待的售后服务次数	0	1	2	3	4	5	$\geqslant 6$
天数	22	37	20	13	6	2	0

解 依题意, 需要检验下列假设

$$P\{X = j\} = \frac{\mathrm{e}^{-1}}{j!}, \qquad j = 0, 1, 2, \cdots \tag{3.22}$$

但这样做存在一个问题, 就是 X 的可能取值有无穷多个, 而样本容量又是有限的, 导致很多像 $\{X = 7\}$ 这样的基本事件因概率太小而没有被观察到。为了解决这个问题, 我们将一些基本事件合并, 将 X 的所有可能取值重新划分为表 3.3 所示的 4 个互斥的事件, 使得这些事件发生的概率相差不太大。

表 3.3 重新划分的 4 个互斥事件

事　件	X 的取值范围	概率 p_j	理论频数 np_j	实际频数 n_j
A_1	$X = 0$	0.3679	36.79	22
A_2	$X = 1$	0.3679	36.79	37
A_3	$X = 2$	0.1839	18.39	20
A_4	$X \geqslant 3$	0.0803	8.03	21

其中 p_j 为事件 A_j 的理论概率, n_j 为事件 A_j 的频数, $n = n_1 + n_2 + n_3 + n_4$, np_j 为事件 A_j 的理论频数。需要检验的原假设为

$$H_0: \quad P(A_j) = p_j, \qquad j = 1, 2, 3, 4 \tag{3.23}$$

Pearson 统计量为

$$K^2 = \sum_{j=1}^{4} \frac{(n_j - np_j)^2}{np_j}$$

根据定理 3.2, K^2 近似服从自由度为 $\nu = 4-1 = 3$ 的卡方分布 $\chi^2(3)$, 将表 3.3 中的数据代入 K^2 的表达式, 计算得到 $K^2 = 27.0370$, 用 MATLAB 计算自由度为 3 的卡方分布的分位点, 得到 $\chi^2_\alpha(3) = 7.8147$, 由于 $K^2 > \chi^2_\alpha(3)$, 因此拒绝原假设, 即认为该服务台每天接待的售后服务次数 X 不服从参数为 $\lambda = 1$ 的泊松分布。

计算过程的 MATLAB 代码如下:

```
function Li3_2()
%%这个函数实现了例3.2中泊松分布的Pearson卡方检验的计算
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
N=100;                              %%样本容量
n=[22,37,20,21];                    %%4个事件的实际频数
k=size(n,2)
p=[0.3679,0.3679,0.1839,0.0803];          %%4个事件的理论概率
K2=sum(((n-N*p).^2)./(N*p))         %%计算Pearson统计量
a=0.05                              %%显著性水平
Ka=chi2inv(1-a,k-1)                 %%计算显著性水平为α的分位点
end
```

3.1.3 理论分布中含有未知参数的拟合优度检验

现在考虑如下问题: 设总体 X 只有 k 个可能取值, 需要用抽样数据检验下列原假设

$$H_0: \quad P\{X = x_j\} = p_j(\theta), \qquad j = 1, 2, \cdots, k \tag{3.24}$$

其中 $\theta = (\theta_1, \theta_2, \cdots, \theta_r)$ 是未知参数。

由于 p_j 依赖于参数 θ, 因此需要先估计参数 θ 的值才能确定 p_j 的值。参数 θ 的估计通常采用极大似然估计法。设 $\widehat{\theta}$ 是 θ 的极大似然估计量, 则 p_j 的估计量为

$$\widehat{p_j} = p_j(\widehat{\theta}) \tag{3.25}$$

因此 Pearson 统计量为

$$K^2 = \sum_{j=1}^{k} \frac{(n_j - n\widehat{p_j})^2}{n\widehat{p_j}} \tag{3.26}$$

Fisher (1924) 证明了在一定的条件下, 若原假设 (3.24) 成立, 则当样本容量 $n \to \infty$ 时, 统计量 (3.26) 的极限分布是自由度为 $\nu = k-1-r$ 的卡方分布 $\chi^2(k-1-r)$。

例 3.3 从 2013 年 1 月 1 日至 2013 年 7 月 2 日在全世界范围内监测到里氏 4 级及 4 级以上的地震共计 203 次, 相继两次地震的间隔天数统计情况如下:

2.8234, 0.4556, 0.9500, 0.0187, 4.9367, 0.6577, 0.1879, 2.1782, 1.4361, 0.1734, 2.0078, 0.8055, 5.1404, 0.0653, 0.8923, 0.2695, 3.8726, 1.3405, 0.4399, 1.1258, 2.1235, 0.3277, 1.2999, 1.9503, 2.8688, 0.8859, 0.4791, 0.1755, 3.2988, 1.3525, 2.0089, 1.0948, 0.1129, 1.9937, 0.2992, 0.0903, 0.5078, 0.1184, 0.4289, 0.6545, 0.9721, 0.6151, 0.0882, 0.1695,

0.0027, 0.0006, 0.0167, 0.0039, 0.0179, 0.0062, 0.0058, 0.0041, 0.0080, 0.0015, 0.0274, 0.0136, 0.0387, 0.0534, 0.1023, 0.1019, 0.0608, 0.0834, 0.1262, 0.1529, 0.0410, 0.1026, 0.1672, 0.0259, 0.2123, 0.0174, 0.0542, 0.1448, 0.4581, 0.3300, 0.5298, 1.0555, 0.9554, 0.0016, 1.3616, 2.3574, 2.1154, 6.0860, 4.2432, 0.7769, 0.5642, 1.5924, 0.1715, 1.0159, 0.2445, 0.0657, 0.6286, 0.0038, 1.0987, 0.3581, 0.3088, 0.3233, 1.5390, 1.2992, 0.1046, 0.0235, 0.0304, 2.4831, 0.1588, 0.3591, 0.3831, 1.6332, 0.2873, 2.5222, 4.1730, 2.7920, 0.1693, 2.4935, 1.4002, 4.6063, 1.9748, 0.0580, 1.3601, 3.2862, 3.8311, 0.0644, 0.6454, 0.5678, 0.9427, 2.2946, 0.2697, 1.2136, 0.5476, 1.0824, 0.1819, 0.2975

试问相继两次地震的间隔天数 X 是否服从指数分布?

解 由于题目并未给出指数分布的参数 θ, 因此需要先用极大似然估计法估计参数 θ。参数为 θ 的指数分布的密度函数为

$$f(x;\theta)=\begin{cases}\dfrac{1}{\theta}\mathrm{e}^{-x/\theta}, & x>0\\ 0, & x\leqslant 0\end{cases} \tag{3.27}$$

因此对数似然函数为

$$L(\theta):=\ln\prod_{i=1}^{n}f(x_i;\theta)=-n\ln\theta-\frac{1}{\theta}\sum_{i=1}^{n}x_i \tag{3.28}$$

其导数为

$$L'(\theta)=-\frac{n}{\theta}+\frac{1}{\theta^2}\sum_{i=1}^{n}x_i \tag{3.29}$$

令 $L'(\theta)=0$, 解得 θ 的极大似然估计量为

$$\widehat{\theta}=\frac{1}{n}\sum_{i=1}^{n}x_i=\overline{x} \tag{3.30}$$

将样本数据代入上述表达式, 计算得到 $\widehat{\theta}=0.8951$。

接下来需要将 X 的取值范围划分成若干互不相交的区间, 使得 X 落在各个区间的概率相差不要太大。区间划分情况如表 3.4 所示。

表 3.4 相继两次地震间隔天数统计

X 的取值范围	$0\leqslant X<0.2$	$0.2\leqslant X<0.4$	$0.4\leqslant X<0.6$	$0.6\leqslant X<0.8$
理论概率 $\widehat{p}_j$	0.2002	0.1601	0.1281	0.1024
理论频数 $n\widehat{p}_j$	40.45	32.35	25.87	20.69
实际频数 n_j	70	34	15	8
X 的取值范围	$0.8\leqslant X<1$	$1\leqslant X<1.4$	$1.4\leqslant X<1.8$	$X\geqslant 1.8$
理论概率 $\widehat{p}_j$	0.0819	0.1179	0.0754	0.1339
理论频数 $n\widehat{p}_j$	16.55	23.82	15.24	27.04
实际频数 n_j	10	22	10	33

记样本容量为 n, 第 j 个区间的实际频数为 n_j, X 落在第 j 个区间的理论概率为 $\widehat{p}_j$, 则 Pearson 统计量为

$$K^2 = \sum_{j=1}^{8} \frac{(n_j - n\widehat{p}_j)^2}{n\widehat{p}_j} \tag{3.31}$$

由于分布中有一个未知参数使用了极大似然估计, 因此统计量 K^2 近似服从自由度为 $\nu = 8-1-1 = 6$ 的卡方分布 $\chi^2(6)$。将表 3.4 中的数据代入 K^2 的表达式计算得到 $K^2 = 39.8720$, 取显著性水平为 $\alpha = 0.05$, 用 MATLAB 计算自由度为 6 的卡方分布的分位点, 得到 $\chi^2_\alpha(6) = 12.5916$, 由于 $K^2 > \chi^2_\alpha(6)$, 因此拒绝原假设, 即认为相继两次大于或等于里氏 4 级的地震的间隔天数 X 不服从指数分布。

本例题用到的地震时间间隔数据已输入至一个名为“Li3-3timeIntervalData”的 MATLAB 变量中, 并保存在文件“里氏 4 级以上地震时间间隔数据.mat”中, 使用时只要载入该文件即可。计算过程的 MATLAB 代码如下:

```
function Li3_3()
%%这个函数实现了例3.3中地震时间间隔数据是否服从指数分布的Pearson卡方检验的
%%计算
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
load('里氏4级以上地震时间间隔数据.mat');%%导入数据,其中有两个变量,本程序用
                                          %%到的变量为Li3_3timeIntervalData
N=size(Li3_3timeIntervalData,1);          %%样本容量
theta=mean(Li3_3timeIntervalData)         %%参数theta的极大似然估计
p=zeros(1,8);                             %%用于保存理论概率
p(1)=1-exp(-0.2/theta);
p(2)=1-exp(-0.4/theta)-(1-exp(-0.2/theta));
p(3)=1-exp(-0.6/theta)-(1-exp(-0.4/theta));
p(4)=1-exp(-0.8/theta)-(1-exp(-0.6/theta));
p(5)=1-exp(-1/theta)-(1-exp(-0.8/theta));
p(6)=1-exp(-1.4/theta)-(1-exp(-1/theta));
p(7)=1-exp(-1.8/theta)-(1-exp(-1.4/theta));
p(8)=exp(-1.8/theta)
Np=N*p                                    %%理论频数
n=zeros(1,8);                             %%用于保存实际频数
n(1)=sum(Li3_3timeIntervalData<0.2);
n(2)=sum((Li3_3timeIntervalData>=0.2)&(Li3_3timeIntervalData<0.4));
n(3)=sum((Li3_3timeIntervalData>=0.4)&(Li3_3timeIntervalData<0.6));
n(4)=sum((Li3_3timeIntervalData>=0.6)&(Li3_3timeIntervalData<0.8));
n(5)=sum((Li3_3timeIntervalData>=0.8)&(Li3_3timeIntervalData<1));
n(6)=sum((Li3_3timeIntervalData>=1)&(Li3_3timeIntervalData<1.4));
n(7)=sum((Li3_3timeIntervalData>=1.4)&(Li3_3timeIntervalData<1.8));
n(8)=sum(Li3_3timeIntervalData>=1.8)
k=size(n,2)
K2=sum(((n-Np).^2)./(Np))      %%计算Pearson统计量
```

```
a=0.05                          %%显著性水平
r=1;                            %%参数的个数
Ka=chi2inv(1-a,k-1-r)           %%计算自由度为k-1-r的卡方分布显著性水平为α的
                                %%分位点
end
```

3.2 正态性检验

本节讨论总体正态性检验的问题, 即检验总体分布是否为正态分布。下面将以上证指数 (000001)2021 年 8 月 31 日至 2022 年 9 月 30 日期间 264 个交易日的收益率作为案例数据, 介绍各种检验正态性的方法。具体数据如下。

0.0045, 0.0065, 0.0084, −0.0043, 0.0111, 0.0150, −0.0004, 0.0049, 0.0027, 0.0033, −0.0143, −0.0017, −0.0135, 0.0019, 0.0040, 0.0038, −0.0080, −0.0084, 0.0054, −0.0185, 0.0090, 0.0067, −0.0001, −0.0125, 0.0042, −0.0010, 0.0040, −0.0012, 0.0070, −0.0017, 0.0022, −0.0034, 0.0076, −0.0034, −0.0099, −0.0124, 0.0082, −0.0008, −0.0110, −0.0020, 0.0081, −0.0101, 0.0020, 0.0024, −0.0042, 0.0115, 0.0018, −0.0016, −0.0033, 0.0044, −0.0047, 0.0112, 0.0061, 0.0020, 0.0010, −0.0024, −0.0056, −0.0004, 0.0003, 0.0036, −0.0009, 0.0094, −0.0050, 0.0016, 0.0117, 0.0097, −0.0018, 0.0040, −0.0053, −0.0038, 0.0075, −0.0117, −0.0107, 0.0087, −0.0007, 0.0057, −0.0070, −0.0006, 0.0039, −0.0092, 0.0061, 0.0057, −0.0020, −0.0103, −0.0025, −0.0018, 0.0039, −0.0073, 0.0084, −0.0118, −0.0096, 0.0058, 0.0079, −0.0033, −0.0009, −0.0092, 0.0004, −0.0262, 0.0066, −0.0179, −0.0097, 0.0201, 0.0067, 0.0079, 0.0017, −0.0066, −0.0099, 0.0050, 0.0057, 0.0006, 0.0065, 0, −0.0096, 0.0092, −0.0171, 0.0062, 0.0032, 0.0076, −0.0013, −0.0009, −0.0097, −0.0219, −0.0238, −0.0113, 0.0121, 0.0041, −0.0264, −0.0508, 0.0342, 0.0139, 0.0111, 0.0008, 0.0019, 0.0034, −0.0064, −0.0118, 0.0007, −0.0033, 0.0194, −0.0044, 0.0093, 0.0002, −0.0143, 0.0047, −0.0264, 0.0145, −0.0083, 0.0121, −0.0045, −0.0049, −0.0005, −0.0135, −0.0229, 0.0023, −0.0527, −0.0145, 0.0246, 0.0058, 0.0238, 0.0068, −0.0218, 0.0009, 0.0105, 0.0075, −0.0012, 0.0095, −0.0034, 0.0065, −0.0025, 0.0036, 0.0159, 0.0001, −0.0244, 0.0118, 0.0050, 0.0023, 0.0060, 0.0118, −0.0013, 0.0042, 0.0127, 0.0017, 0.0068, −0.0076, 0.0141, −0.0090, 0.0102, 0.0050, −0.0061, 0.0095, −0.0004, −0.0026, −0.0120, 0.0161, 0.0089, 0.0087, 0.0088, −0.0141, 0.0110, −0.0032, 0.0052, −0.0004, −0.0144, 0.0027, −0.0025, −0.0127, −0.0097, 0.0009, −0.0008, −0.0165, 0.0154, 0.0004, 0.0077, −0.0100, −0.0006, −0.0060, 0.0083, −0.0005, 0.0021, −0.0090, 0.0021, −0.0229, −0.0071, 0.0080, 0.0118, 0.0031, 0.0032, −0.0054, 0.0159, −0.0015, −0.0002, 0.0005, 0.0045, −0.0046, −0.0060, 0.0060, −0.0005, −0.0188, 0.0096, −0.0031, 0.0014, −0.0042, −0.0078, −0.0054, 0.0005, 0.0042, 0.0135, 0.0009, −0.0033, 0.0081, 0.0005, −0.0081, −0.0117, −0.0232, −0.0035, 0.0022, −0.0017, −0.0027, −0.0066, −0.0121, 0.0139, −0.0159, −0.0013, −0.0055

以上数据已输入至一个名为“SHR”的 MATLAB 变量中, 并保存在文件“上证指数收益率.mat”中, 使用时只需要将这个文件载入即可。

3.2.1　图示法

1. 直方图

将 X 的取值范围均匀地划分成若干区间, 然后统计每个区间的频数, 并画出条形图展示, 这就是直方图。

例 3.4　画出本节开头给出的上证指数收益率数据的直方图, 并根据图形判断该数据是否来自正态总体。

解　样本数据的最小值为 -0.0527, 最大值为 0.0342, 因此将区间 $[-0.055, 0.055]$ 等分为 22 个长度为 0.005 的小区间, 统计每个小区间的频数, 并绘制成直方图, 如图 3.1 所示。

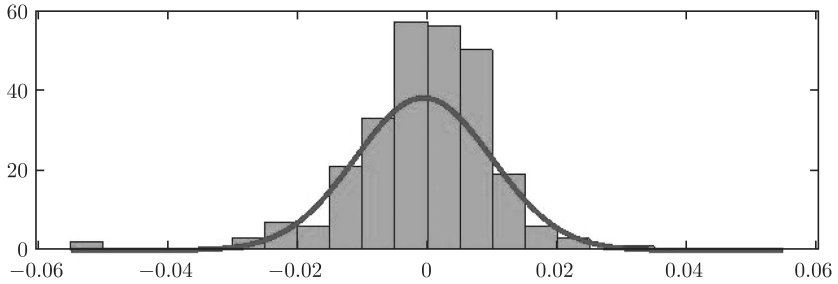

图 3.1　上证指数收益率直方图

为了与正态分布对比, 我们计算出样本数据的均值 $\hat{\mu} = -0.000583$, 标准差 $\hat{\sigma} = 0.0105$, 相应的正态分布密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi}\hat{\sigma}} \mathrm{e}^{-\frac{(x-\hat{\mu})^2}{2\hat{\sigma}^2}}$$

图 3.1 中的曲线便是这个概率密度函数的图像。从图 3.1 可以看出, 样本数据在均值附近的频率明显高于相应正态分布的概率, 且远离均值的异常值出现的频率也高于正态分布的概率, 因此我们认为样本数据不是来自正态总体的。

例 3.4 的计算和绘图过程的 MATLAB 代码如下:

```
function Li3_4()
%%这个函数实现了例3.4中绘制上证指数收益率直方图的操作
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
load('上证指数收益率.mat'); %%导入数据,上证指数收益率数据保存在变量SHR中
mu=mean(SHR);                %%计算样本均值
s=std(SHR);                  %%计算样本标准差
edges=(-0.055:0.005:0.055);     %%划分区间的分割点
histogram(SHR,edges);           %%绘制直方图
hold on
x=(-0.055:0.0005:0.055)';
y=(1/(sqrt(2*pi)*s))*exp(-(x-mu).^2./(2*s^2));
plot(x,y,'blue');
hold off
```

```
end
```

2. 经验分布函数

设 $x_1, x_2, \cdots, x_n$ 是来自总体 X 的样本数据, 为了估计总体 X 的分布函数 $F(x)$, 定义

$$F_n(x) := \frac{\sharp\{x_i:\ x_i \leqslant x, i = 1, 2, \cdots\}}{n} \tag{3.32}$$

其中 $\sharp\{x_i:\ x_i \leqslant x, i = 1, 2, \cdots\}$ 表示集合 $\{x_i:\ x_i \leqslant x, i = 1, 2, \cdots\}$ 中的元素个数, 称 $F_n(x)$ 为**经验分布函数 (empirical distribution function)**。按照上面的定义, 对于给定的实数 x, $F_n(x)$ 的值就是样本数据 $x_1, x_2, \cdots, x_n$ 中小于或等于 x 的那部分所占的比例。

如果把这些样本数据按照从小到大的顺序排序, 最小的记为 $x_{(1)}$, 第二小的记为 $x_{(2)}$, 以此类推, 最大的记为 $x_{(n)}$, 则称 $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$ 为**顺序统计量 (order statistics)**。

利用顺序统计量可以将经验分布函数表示成下列形式:

$$F_n(x) = \begin{cases} 0, & x < x_{(1)} \\ \dfrac{k}{n}, & x_{(k)} \leqslant x < x_{(k+1)}, \quad k = 1, 2, \cdots, n-1 \\ 1, & x \geqslant x_{(n)} \end{cases} \tag{3.33}$$

不难发现 $F_n(x)$ 是一个分段阶梯函数。

根据 Glivenko-Cantelli 引理（附录 C 定理 C.4）, 当样本容量 $n \to \infty$ 时, 经验分布函数 $F_n(x)$ 将以概率 1 一致收敛域总体 X 的分布函数 $F(x)$, 即有

$$P\left\{\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = 0\right\} = 1 \tag{3.34}$$

关于顺序统计量和经验分布的更多知识可参阅附录 C。

3. *P-P* 图

如果需要检验总体 X 是否服从某个给定的理论分布, 可以利用抽样数据 $x_1, x_2, \cdots, x_n$ 求出经验分布函数 $F_n(x)$, 然后将 $F_n(x)$ 与给定的理论分布函数 $F(x)$ 进行对比, 以判断总体 X 是否服从给定的理论分布。

如何检验经验分布函数 $F_n(x)$ 和理论分布函数 $F(x)$ 是否一致呢? ***P-P* 图 (probability-probability plot)** 是一种直观的图示检验方法。

P-P 图的绘制方法如下: 对于每个样本数据 x_i, 分别以 $F(x_i)$、$F_n(x_i)$ 为横坐标和纵坐标在二维坐标平面上绘制点 $(F(x_i), F_n(x_i))$, 这 n 个点构成一个散点图, 同时画出连节点 $O(0, 0)$ 和 $A(1, 1)$ 的直线段, 便得到了 *P-P* 图。

如何从散点图判断总体 X 是否服从给定的理论分布呢? 如果总体 X 服从给定的理论分布, 则经验分布 $F_n(x)$ 应与理论分布 $F(x)$ 很接近, 从而点 $(F(x_i), F_n(x_i)), i = 1, 2, \cdots, n$ 应落在线段 OA 附近。如果总体 X 不服从给定的理论分布, 则经验分布 $F_n(x)$ 应与理论分布 $F(x)$ 有显著的偏差, 从而点 $(F(x_i), F_n(x_i)), i = 1, 2, \cdots, n$ 会偏离线段 OA。

例 3.5　用 P-P 图判断本节开头给出的上证指数收益率数据是否服从正态分布。

解　首先求出收益率数据的样本均值 $\widehat{\mu}$ 和样本标准差 $\widehat{\sigma}$ 为

$$\widehat{\mu} = -0.000583, \qquad \widehat{\sigma} = 0.0105$$

然后对数据进行标准化变换

$$z_i = \frac{x_i - \widehat{\mu}}{\widehat{\sigma}}, \quad i = 1, 2, \cdots, n \tag{3.35}$$

如果总体 X 服从正态分布, 则标准化的数据 $z_i, i = 1, 2, \cdots, n$ 应服从标准正态分布, 因此只要画 P-P 图检验 $z_i, i = 1, 2, \cdots, n$ 是否服从标准正态分布即可。

标准化的上证指数收益率数据的 P-P 图如图 3.2 所示。可以看出, 经验分布 $F_n(x)$ 与理论分布 $F(x)$（标准正态分布）有显著偏差, 因此我们认为上证指数收益率不服从正态分布。

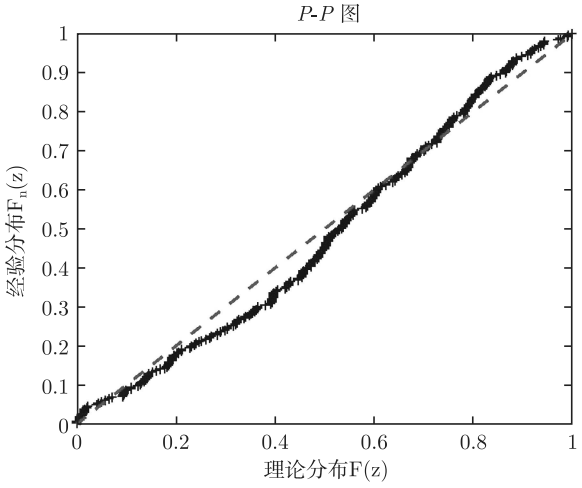

图 3.2　标准化的上证指数收益率数据 P-P 图

例 3.5 的计算和绘图过程的 MATLAB 代码如下:

```
function Li3_5()
%%这个函数实现了例3.5中画P-P图检验上证指数收益率是否服从正态分布的操作
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
load('上证指数收益率.mat'); %%导入数据，上证指数收益率数据保存在变量SHR中
mu=mean(SHR);               %%计算样本均值
s=std(SHR);                 %%计算样本标准差
z=(SHR-mu)/s;               %%数据标准化
z=sort(z);                  %%从小到大排序
n=size(z,1);
Fn=(1:n)'/n;                %%z对应的经验分布值
F=normcdf(z);               %%z对应的标准正态分布函数值
plot(F,Fn,'+','LineWidth',1,'MarkerSize',...
```

```
6,'MarkerFaceColor','k','MarkerEdgeColor','k');
hold on
plot([0;1],[0;1],'LineStyle','--','Color','b','LineWidth',2);
xlabel('理论分布F(z)');                    %% 横轴名称
ylabel('经验分布F_n(z)');                  %% 纵轴名称
title('P-P图');                            %% 添加图的标题
hold off
end
```

4. *Q-Q* 图

设某个随机变量 X 的分布函数为 $F(x)$, 我们知道 $F(x)$ 的值域为 $[0,1]$, 对于任意给定的 $p\in[0,1]$, 记

$$Q(p):=\inf\{x\in\mathbb{R}:\ F(x)\geqslant p\} \tag{3.36}$$

称为随机变量 X 的（或分布函数 $F(x)$ 的）**p-分位数 (p-quantile)**。如果 $F(x)$ 是严格单调增加的连续函数, 则不难发现 Q 是 F 的反函数, 即有 $Q(p)=F^{-1}(p)$。

例如, 设 Φ 是标准正态分布函数, 通过计算知道 $\Phi(-0.674489750196082)=0.25$, 因此标准正态分布的 0.25-分位数是

$$Q(0.25)=\Phi^{-1}(0.25)=-0.674489750196082$$

Q-Q 图 (quantile-quantile plot) 是另外一种检验总体分布是否为某个给定的理论分布的图示方法。设 $x_1,x_2,\cdots,x_n$ 是来自总体 X 的抽样数据, $F_n(x)$ 是经验分布函数, 记

$$p_i=F_n(x_i),\qquad i=1,2,\cdots,n$$

为了比较 $F_n(x)$ 与理论分布 $F(x)$, 我们计算 $F(x)$ 的 p_i-分位数

$$y_i=Q(p_i),\qquad i=1,2,\cdots,n$$

然后在二维平面直角坐标系中绘制出点集 $\{(y_i,x_i):\ i=1,2,\cdots,n\}$ 的散点图, 这就是所谓的 *Q-Q* 图。

例 3.6　用 *Q-Q* 图判断本节开头给出的上证指数收益率数据是否服从正态分布。

解　首先将数据标准化, 设标准化后的数据为 $z_i, i=1,2,\cdots,n$。为了方便计算经验分布函数, 将标准化的数据从小到大排序, 设排序后的数据为

$$z_{(1)}\leqslant z_{(2)}\leqslant z_{(3)}\leqslant\cdots\leqslant z_{(n)}$$

根据经验分布函数的定义得 $F_n(z_{(k)})=k/n, n=1,2,\cdots,n$, 为了使 *Q-Q* 图更合理, 通常取

$$y_{(k)}=Q\left(\frac{k-0.375}{n+0.25}\right)=\Phi^{-1}\left(\frac{k-0.375}{n+0.25}\right),\qquad k=1,2,\cdots,n \tag{3.37}$$

然后绘制点集 $\{(y_{(k)}, z_{(k)}):\ k = 1, 2, \cdots, n\}$ 的散点图。为了对比, 我们在同一坐标系中画出直线 $z = y$, 如果上证指数的收益率服从正态分布, 则标准化数据 $z_i, i = 1, 2, \cdots, n$ 应服从标准正态分布, 从而点集 $\{(y_{(k)}, z_{(k)}):\ k = 1, 2, \cdots, n\}$ 应落在直线 $z = y$ 附近。

标准化的上证指数收益率数据的 Q-Q 图如图 3.3 所示, 其中的虚线是直线 $z = y$。可以看出, 样本分位数与标准正态分布的分位数有显著偏差, 对于很小概率 p 对应的分位数, 样本分位数明显小于标准正态分布的对应分位数, 对很大的概率 p 对应的分位数, 样本分位数明显大于标准正态分布的对应分位数, 这说明上证指数收益率分布具有拖尾特性, 远离均值的收益率出现的概率明显高于正态分布。因此我们认为上证指数收益率不服从正态分布。

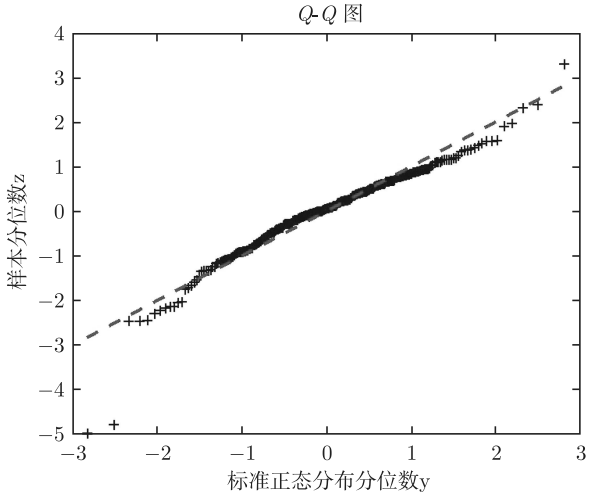

图 3.3　标准化的上证指数收益率数据 Q-Q 图

标准化的上证指数收益率 Q-Q 图的计算和绘图过程的 MATLAB 代码如下:

```
function Li3_6A()
%%这个函数实现了例3.6中绘制标准化收益率数据Q-Q图的操作
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
load('上证指数收益率.mat'); %%导入数据, 上证指数收益率数据保存在变量SHR中
mu=mean(SHR);               %%计算样本均值
s=std(SHR);                 %%计算样本标准差
z=(SHR-mu)/s;               %%数据标准化
z=sort(z);                  %%从小到大排序
n=size(z,1);
Fn=((1:n)'-0.375)/(n+0.25); %%z对应的经验分布修正值
y=norminv(Fn);              %%Fn对应的标准正态分布分位数
plot(y,z,'+','LineWidth',1,'MarkerSize',6,'MarkerFaceColor',...
'k','MarkerEdgeColor','k');
hold on
plot([y(1);y(n)],[y(1);y(n)],'LineStyle','--','Color','b',...
'LineWidth',2);
xlabel('标准正态分布分位数y');              %% 横轴名称
```

```
ylabel('样本分位数z');                    %% 纵轴名称
title('Q-Q图');                          %% 添加图的标题
hold off
end
```

也可以不对上证指数收益率数据 $x_i, i=1,2,\cdots,n$ 进行标准化，而是直接将其按照从小到大的顺序排列:

$$x_{(1)} \leqslant x_{(2)} \leqslant x_{(3)} \leqslant \cdots \leqslant x_{(n)}$$

并利用式 (3.37) 计算 $y_k, k=1,2,\cdots,n$, 然后绘制点集

$$\left\{(y_{(k)}, x_{(k)}):\ k=1,2,\cdots,n\right\} \tag{3.38}$$

的散点图。这样得到的 Q-Q 图具有下列性质:

如果总体 $X \sim N(\mu,\sigma^2)$, 则点集 $\{(y_{(k)}, x_{(k)}):\ k=1,2,\cdots,n\}$ 应落在直线 $x=\sigma y+\mu$ 附近。至于原因，请读者自行分析。

未经标准化的上证指数收益率数据的 Q-Q 图如图 3.4 所示，其中的虚线是直线

$$x = \widehat{\sigma} y + \widehat{\mu}$$

其中 $\widehat{\mu}$ 是样本均值，$\widehat{\sigma}$ 是样本标准差。可以看出，点集 $\{(y_{(k)}, x_{(k)}):\ k=1,2,\cdots,n\}$ 与直线有显著的偏离，因此我们认为上证指数收益率不服从正态分布。

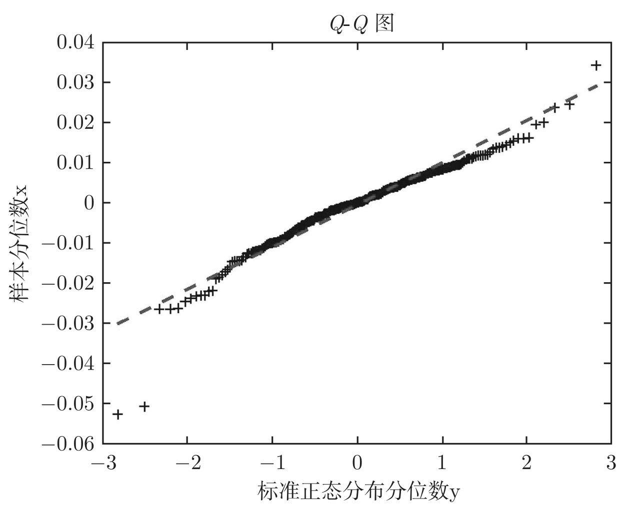

图 3.4　未经标准化的上证指数收益率数据 Q-Q 图

未经标准化的上证指数收益率 Q-Q 图的计算和绘图过程的 MATLAB 代码如下:

```
function Li3_6B()
%%这个函数实现了例3.6中绘制收益率数据Q-Q图的操作(不进行标准化)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
load('上证指数收益率.mat'); %%导入数据, 上证指数收益率数据保存在变量SHR中
```

```
mu=mean(SHR);                    %%计算样本均值
s=std(SHR);                      %%计算样本标准差
x=sort(SHR);                     %%从小到大排序
n=size(x,1);
Fn=((1:n)'-0.375)/(n+0.25); %%x对应的经验分布修正值
y=norminv(Fn);                   %%Fn对应的标准正态分布分位数
plot(y,x,'+','LineWidth',1,'MarkerSize',6,'MarkerFaceColor',...
'k','MarkerEdgeColor','k');
hold on
plot([y(1);y(n)],[s*y(1)+mu;s*y(n)+mu],'LineStyle','--',...
'Color','b','LineWidth',2);
xlabel('标准正态分布分位数y');                  %% 横轴名称
ylabel('样本分位数x');                          %% 纵轴名称
title('Q-Q图');                                 %% 添加图的标题
hold off
end
```

3.2.2 拟合优度检验

上文介绍的直方图、*P*-*P* 图和 *Q*-*Q* 图虽然直观, 但缺乏定量标准, 主观性强, 且不便于程序自动化执行, 因此在许多领域使用受限。为了解决这个问题, 统计学家提出了许多定量化的正态性检验方法, 下面选择几种常用的方法进行介绍。

首先是拟合优度检验, 这种方法的用途不限于正态性检验, 已在 3.1.1 节、3.1.2 节进行了详细介绍。下面通过一个例子说明如何用拟合优度检验法检验总体是否服从正态分布。

例 3.7 用拟合优度检验法检验本节开头给出的上证指数收益率数据是否服从正态分布。

解 由于正态分布的均值 μ 和标准差 σ 未知, 因此需要计算这两个参数的极大似然估计。利用附录 B 中的例 B.1 的结论得

$$\widehat{\mu}=\overline{x}=-0.000583, \qquad \widehat{\sigma}=\sqrt{\frac{n-1}{n}s^2}=0.010474 \tag{3.39}$$

接下来需要将收益率数据的取值范围划分成若干互不相交的区间, 并用正态分布 $N(\widehat{\mu},\widehat{\sigma}^2)$ 估计收益率落在每个区间的理论概率和理论频数。区间划分情况如表 3.5 所示。

记样本容量为 n, 第 j 个区间的实际频数为 n_j, 收益率落在第 j 个区间的理论概率为 $\widehat{p}_j$, 则 Pearson 统计量为

$$K^2=\sum_{j=1}^{11}\frac{(n_j-n\widehat{p}_j)^2}{n\widehat{p}_j} \tag{3.40}$$

由于正态分布中有 2 个未知参数使用了极大似然估计, 因此统计量 K^2 近似服从自由度为 $\nu=11-1-2=8$ 的卡方分布 $\chi^2(8)$。将表 3.5 中的数据代入 K^2 的表达式, 计算得到

$K^2 = 21.5537$, 取显著性水平为 $\alpha = 0.05$, 用 MATLAB 计算自由度为 8 的卡方分布的分位数, 得到 $\chi^2_\alpha(8) = 15.5037$, 由于 $K^2 > \chi^2_\alpha(8)$, 因此拒绝原假设, 即认为所给上证指数收益率数据不服从正态分布。

表 3.5　上证指数收益率统计

取值范围	$(-\infty, -0.025)$	$[-0.025, -0.02)$	$[-0.02, -0.015)$	$[-0.015, -0.01)$
理论概率 $\widehat{p}_j$	0.0099	0.0220	0.0525	0.1000
理论频数 $n\widehat{p}_j$	2.6057	5.8103	13.8485	26.3902
实际频数 n_j	5	7	6	21
取值范围	$[-0.01, -0.005)$	$[-0.005, 0)$	$[0, 0.005)$	$[0.005, 0.01)$
理论概率 $\widehat{p}_j$	0.1523	0.1856	0.1808	0.1409
理论频数 $n\widehat{p}_j$	40.2112	48.9931	47.7323	37.1861
实际频数 n_j	32	58	56	50
取值范围	$[0.01, -0.015)$	$[0.015, 0.02)$	$[0.02, +\infty)$	
理论概率 $\widehat{p}_j$	0.0877	0.0437	0.0247	
理论频数 $n\widehat{p}_j$	23.1647	11.5381	6.5198	
实际频数 n_j	19	6	4	

实现本例计算过程的 MATLAB 代码如下:

```
function Li3_7()
%%这个函数实现了例3.7中上证指数收益率数据是否服从正态分布的拟合优度检验的
%%计算
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
load('上证指数收益率.mat'); %%导入数据，上证指数收益率数据保存在变量SHR中
N=size(SHR,1);              %%样本容量
mu=mean(SHR)                %%均值的极大似然估计
s=sqrt((N-1)/N)*std(SHR)                %%标准差的极大似然估计
p=zeros(1,11);                          %%用于保存理论概率
p(1)=normcdf(-0.025,mu,s);
p(2)=normcdf(-0.02,mu,s)-normcdf(-0.025,mu,s);
p(3)=normcdf(-0.015,mu,s)-normcdf(-0.02,mu,s);
p(4)=normcdf(-0.01,mu,s)-normcdf(-0.015,mu,s);
p(5)=normcdf(-0.005,mu,s)-normcdf(-0.01,mu,s);
p(6)=normcdf(0,mu,s)-normcdf(-0.005,mu,s);
p(7)=normcdf(0.005,mu,s)-normcdf(0,mu,s);
p(8)=normcdf(0.01,mu,s)-normcdf(0.005,mu,s);
p(9)=normcdf(0.015,mu,s)-normcdf(0.01,mu,s);
p(10)=normcdf(0.02,mu,s)-normcdf(0.015,mu,s);
p(11)=1-normcdf(0.02,mu,s)
Np=N*p                                  %%理论频数
n=zeros(1,11);                          %%用于保存实际频数
n(1)=sum(SHR<-0.025);
n(2)=sum((SHR>=-0.025)&(SHR<-0.02));
n(3)=sum((SHR>=-0.02)&(SHR<-0.015));
```

```
n(4)=sum((SHR>=-0.015)&(SHR<-0.01));
n(5)=sum((SHR>=-0.01)&(SHR<-0.005));
n(6)=sum((SHR>=-0.005)&(SHR<0));
n(7)=sum((SHR>=0)&(SHR<0.005));
n(8)=sum((SHR>=0.005)&(SHR<0.01));
n(9)=sum((SHR>=0.01)&(SHR<0.015));
n(10)=sum((SHR>=0.015)&(SHR<0.02));
n(11)=sum(SHR>=0.02)
k=size(n,2)
K2=sum(((n-Np).^2)./(Np))      %%计算Pearson统计量
a=0.05                         %%显著性水平
r=2;                           %%参数的个数
Ka=chi2inv(1-a,k-1-r)          %%计算自由度为k-1-r的卡方分布显著性水平为α的
                               %%分位点
end
```

3.2.3 Kolmogorov-Smirnov 检验

设 $x_1, x_2, \cdots, x_n$ 是来自总体 X 的抽样数据, $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$ 是其顺序统计量, $F_n(x)$ 是经验分布函数。我们想要检验总体 X 是否服从某个指定的理论分布 $F(x)$, 即检验原假设

$$H_0: \quad \text{总体}X\text{的分布函数为}F(x) \tag{3.41}$$

为了度量 $F_n(x)$ 与 $F(x)$ 的偏差, Kolmogorov 定义了统计量

$$K := \sqrt{n}\sup_{x\in\mathbb{R}}|F_n(x) - F(x)| \tag{3.42}$$

在原假设 H_0 成立的条件下, 如果分布函数 $F(x)$ 是连续的, Kolmogorov[48] 证明了当 $n \to \infty$ 时统计量 K 的极限分布为

$$P\{K \leqslant x\} = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}\mathrm{e}^{-2k^2x^2} = \frac{\sqrt{2\pi}}{x}\sum_{k=1}^{\infty}\mathrm{e}^{-(2k-1)^2\pi^2/(8x^2)} \tag{3.43}$$

这个分布称为 **Kolmogorov 分布**。

计算 Kolmogorov 统计量涉及在实数集 $\mathbb{R}$ 上取上确界的运算, 貌似不可行, 但实际上经验分布函数 $F_n(x)$ 是形如式 (3.33) 的阶梯函数, 而分布函数又是在 $[0,1]$ 上取值的单调非减函数, 不难证明

$$\begin{aligned}\sup_{x\in\mathbb{R}}|F_n(x) - F(x)| &= \max\left\{|F_n(x_{(k)}) - F(x_{(k)})|, |F_n(x_{(k-1)}) - F(x_{(k)})| : \ k = 1, 2, \cdots, n\right\} \\ &= \max\left\{\frac{k}{n} - F(x_{(k)}), F(x_{(k)}) - \frac{k-1}{n} : \ k = 1, 2, \cdots, n\right\}\end{aligned} \tag{3.44}$$

因此有

$$K := \sqrt{n}\max\left\{\frac{k}{n} - F(x_{(k)}), F(x_{(k)}) - \frac{k-1}{n} : \ k = 1, 2, \cdots, n\right\} \tag{3.45}$$

为了实现检验, Smirnov[49] 制定了 Kolmogorov 分布的分位数表, 对于给定的显著性水平 α, 可查到相应的分位数 K_α, 当由样本数据计算得到的统计量 $K > K_\alpha$ 时, 拒绝原假设, 即认为总体 X 的分布函数不是 $F(x)$。

Kolmogorov-Smirnov 检验的实际计算过程比较烦琐, 可直接调用 MATLAB 提供的函数 kstest()。

例 3.8 用 MATLAB 函数 kstest() 检验本节开头给出的上证指数收益率数据是否服从均值 $\mu = -0.000583$、标准差 $\sigma = 0.010494$ 的正态分布。

解 先对数据进行变换:

$$y_i = \frac{x_i - \mu}{\sigma}, \qquad i = 1, 2, \cdots, n$$

如果原来的数据 $\{x_i: i = 1, 2, \cdots, n\}$ 服从正态分布 $N(\mu, \sigma^2)$, 则变换后的数据 $\{y_i: i = 1, 2, \cdots, n\}$ 服从标准正态分布, 因此只需要检验变换后的数据是否服从标准正态分布。我们取显著性水平为 $\alpha = 0.05$, 进行 Kolmogorov-Smirnov 检验。

下面是实现本例的 Kolmogorov-Smirnov 检验的 MATLAB 代码:

```
function Li3_8()
%%这个函数实现了例3.8中上证指数收益率数据的Kolmogorov-Smirnov检验的计算
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
load('上证指数收益率.mat'); %%导入数据,上证指数收益率数据保存在变量SHR中
mu=-0.000583;
s=0.010494;
y=(SHR-mu)/s;                %%数据变换
n=size(y,1)                  %%样本容量
a=0.05                       %%显著性水平
[h,p]=kstest(y,'Alpha',a)    %%Kolmogorov-Smirnov检验
                             %%原假设H0: y来自标准正态总体
%%输出参数: h--逻辑变量, h=1代表拒绝原假设, h=0代表接受原假设
%%           p--概率值,可以理解为能拒绝原假设的最小显著性水平
end
```

运行程序后屏幕输出计算结果为 $h = 1, p = 0.0467$。其中 $h = 1$ 表示拒绝原假设, 即认为变换后的数据 $\{y_i : i = 1, 2, \cdots, n\}$ 不服从标准正态分布, 从而原始收益率数据 $\{x_i : i = 1, 2, \cdots, n\}$ 不服从均值 $\mu = -0.000583$、标准差 $\sigma = 0.010494$ 的正态分布; $p = 0.0467$ 表示能够拒绝原假设的最小显著性水平为 0.0467。

需要指出的是, 虽然 Kolmogorov-Smirnov 检验拒绝了原始收益率数据 $\{x_i : i = 1, 2, \cdots, n\}$ 服从均值 $\mu = -0.000583$、标准差 $\sigma = 0.010494$ 的正态分布的假设, 但并没有拒绝这个数据服从其他均值和标准差的正态分布, 换句话说, Kolmogorov-Smirnov 检验的结果依赖于指定分布的参数, 必须保证分布参数精确, 检验结论才有效。