第九章 模型的选择

All models are wrong, but some models are usefull.

在建模过程的最后阶段，我们需要从众多的备选模型中选择一个“最优”模型。

一般来说，对模型的筛选将经历如下的过程：

- 被选模型与实际数据图形上的直观比较和筛选
- 用统计学方法对模型分布函数与经验分布函数进行检验（如 χ^2 拟合优度检验、K-S 检验、Anderson-Darling 检验等）

由一定的标准进行模型选择（常用主观判断法和和评分法）。

但是，无论选择哪个模型，都是对实际情况的一种近似。所有的模型都存在一定的问题，并不存在绝对最优的模型。

一、直观选择

1、密度函数与分布函数的图像比较

例 9.1 由 10 只试验老鼠组成的样本，其死亡时间（以天为单位）为：3，4，5，7，7，8，10，10，10，12。假定适合的生存模型为指数分布。试用极大似然估计估计指数分布参数，并用图像对比法进行模型的筛选。

解：设指数分布的分布函数为：$F(x)=1-\exp(-x/\theta)$，用极大似然估计可以得到指数分布的参数为：$\hat{\theta}=7.6$，$\hat{F}(x)=1-\exp(-x/7.6)$。将样本经验分布函数与估计函数画在同一个图中，如下：

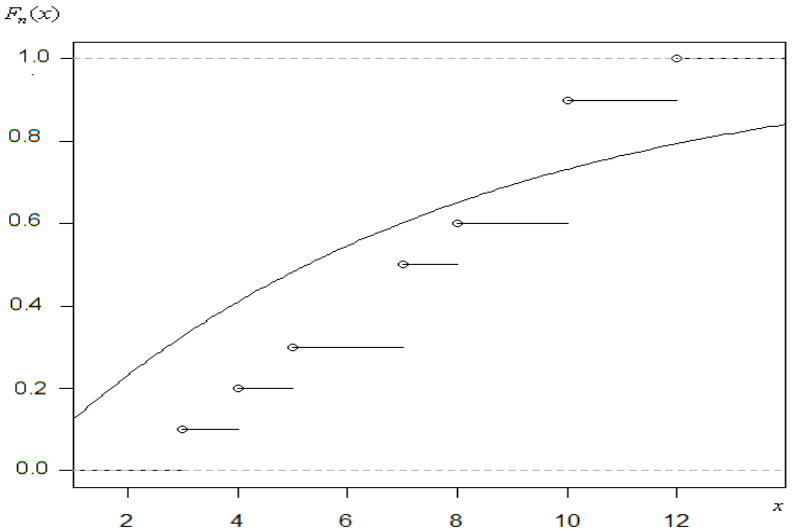


图 9.1 小鼠死亡时间的经验分布函数图像

由图看出，用指数分布来拟合小鼠的生存 函数并不合适。在 x 较小时，拟合值大于样本值。当 x 较大时，拟合值小于样本值。

当模型的分布函数和经验分布函数很接近时，很难从图像上分辨出细微的差别。可以直接画出两个函数差值的图像。也就是说，如果 $F_n(x)$ 和 $F(x)$ 分别表示经验分布函数和由模型得到的分布函数，画出 $F(x)-F_n(x)$ 的图像即可。

```
rm(list = ls())
x<-c(3,4,5,7,7,8,10,10,10,12)
n<-length(x)

f<-function(p,x,n){
  a=-sum(x)/p-n*log(p)   #似然函数值
```

```
  return(a)
}
xmax <- optimize(f,c(1,100),tol = 0.0001, maximum = TRUE,x,n)
$maximum
xmax

## [1] 7.600004

Fnx<-ecdf(x)
plot.ecdf(Fnx)
curve(pexp(x,1/xmax),xlim =c(0,14),ylab = "Fn(x)",add=TRUE)

pexp(unique(x),1/xmax) #计算模型得到的分布函数的分布密度

## [1] 0.3261425 0.4092223 0.4820592 0.6019000 0.6509817 0.73
17374 0.7938078

cumsum(table(x)/length(x)) #计算原始的分布密度
```

```
##   3   4   5   7   8  10  12
## 0.1 0.2 0.3 0.5 0.6 0.9 1.0

plot(pexp(unique(x),1/xmax)-cumsum(table(x)/length(x)),ylab =
 "F(x)-Fn(x)",xlab = "x")
abline(h=0)
```

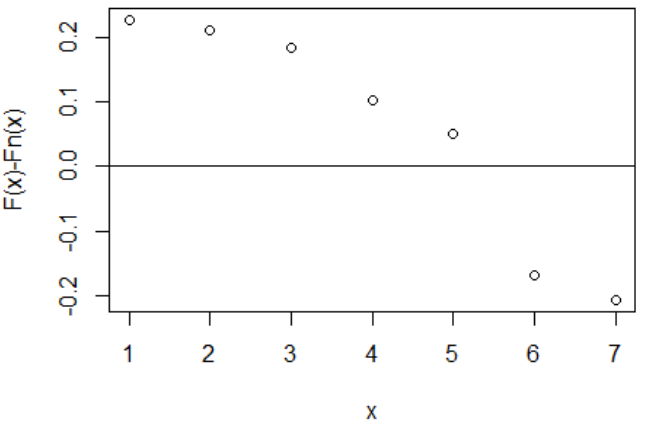


图 9.2 $F(x)-F_n(x)$

2、 $p-p$ 图，

根据变量的经验分布与指定分布的累积分布函数之间的关系所绘制的图形。通过 $p-p$ 图可以检验数据是否符合指定的分布。其步骤如下：首先将观测值排序 $x_1 \le \cdots \le x_n$，再对每个值构造坐标 $(F_n(x_j), F^*(x_j))$，最后将每个坐标对应的点画在 $(F_n(x), F^*(x))$ 的平面上。当数据符合指定分布时， $p-p$ 图中各点近似呈一条 45^o 直线。

例 9.1 续 由 10 只试验老鼠组成的样本，其死亡时间（以天为单位）为：3，4，5，7，7，8，10，10，10，12。画出指数分布的 $p-p$ 图。

解：我们用 $x_1, \cdots, x_{10}$ 来表示上面各值。以观测值 $x_{(2)} = 4$ 为例，经验分布

值为 $F_{10}(4)=2/11=0.1818$，另一个坐标的值是

$$F^{*}(4)=1-\exp(-4/7.6)=0.4092$$

这就得到 $p-p$ 图中的一个点（0.1818,0.4092）。类似的，可以得到所有的点，其图形由图 9.3 所示。

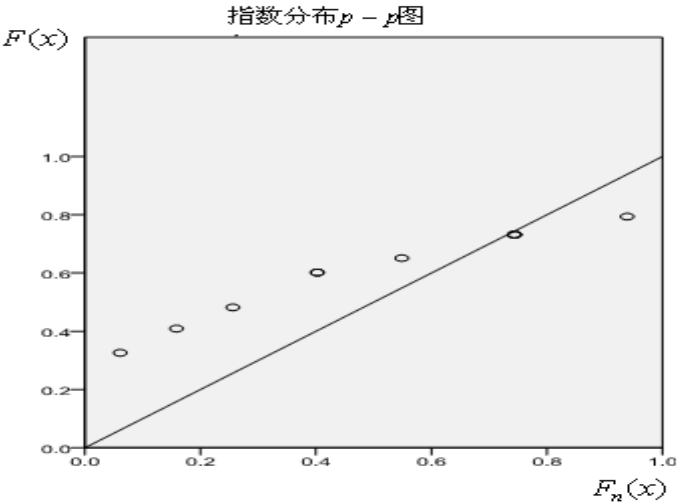


图 9.3 例 1 的数据 $p-p$ 图

从图 9.3 可以看出，指数分布不适合描述该数据的分布。

```
##例 9.1 续，画 p-p 图
a<-pexp(unique(x),1/xmax) #纵坐标
b<-cumsum(table(x)/(length(x)+1)) #横坐标
plot(b,a,xlim=c(0,1),ylim = c(0,0.8))
abline(0,1)  #对角线
```

3、*Q-Q* 图

Q-Q 图是用样本数据的经验分位数与所指定分布的分位数之间的关系曲线来进行检验的。下图是例 1 的数据的 *Q-Q* 图。

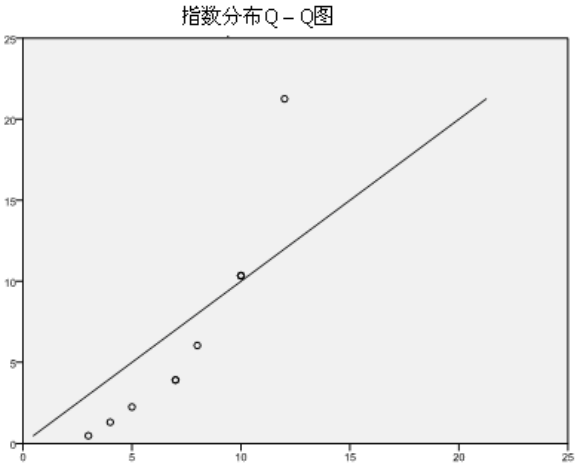


图 9.4 例 9.1 数据的 Q-Q 图

```
#画qq 图
c<-qexp(b,1/xmax) #纵坐标，理论数据
plot(unique(x),c,xlim = c(0,25),ylim = c(0,25))
abline(0,1)
```

由图像可见 $F_n(x)$ 和 $F^*(x)$ 的差距在 ± 0.25 之间，差距较大，因此并不适合用指数分布来拟合样本数据，与例子结论一致。

4、平均剩余寿命函数图

平均剩余寿命函数考虑的是数据在尾部的情况，其定义为：

$$e(d) = E[X - d \mid X > d] \qquad (10.2.3)$$

如果平均剩余寿命函数随 d 递增，那么在变量取值较大处的期望结

果会很大，因此概率向右移，说明其尾部相比那些平均剩余生命函数递减或增速较慢的模型更厚。反之，如果平均剩余寿命函数随 d 递减，说明 X 的分布是轻尾分布。这里通过样本平均剩余寿命函数图 $\left(d,\hat{e}(d)\right)$ 观察样本数据的尾部特征。使用经验估计 $\hat{e}(d)$ 来代替 $e(d)$，有：

$$\hat{e}(d)=\frac{\sum_{i=1}^{n}\max(X_i-d,0)}{\sum_{i=1}^{n}I_{\{X_i>d\}}}$$

如果平均剩余寿命函数图呈现上升的趋势，说明样本的损失分布是一个明显的厚尾分布；而如果呈现下降的趋势则是轻尾分布；指数分布的平均剩余寿命函数图近似为一条水平的直线。

例 9.2 画例 9-1 的平均剩余寿命函数图

```
lam <- 1/7.6
stopifnot(lam > 0)
set.seed(271)
fx <- rexp(n, rate = lam)
mean_excess_plot(fx)
```

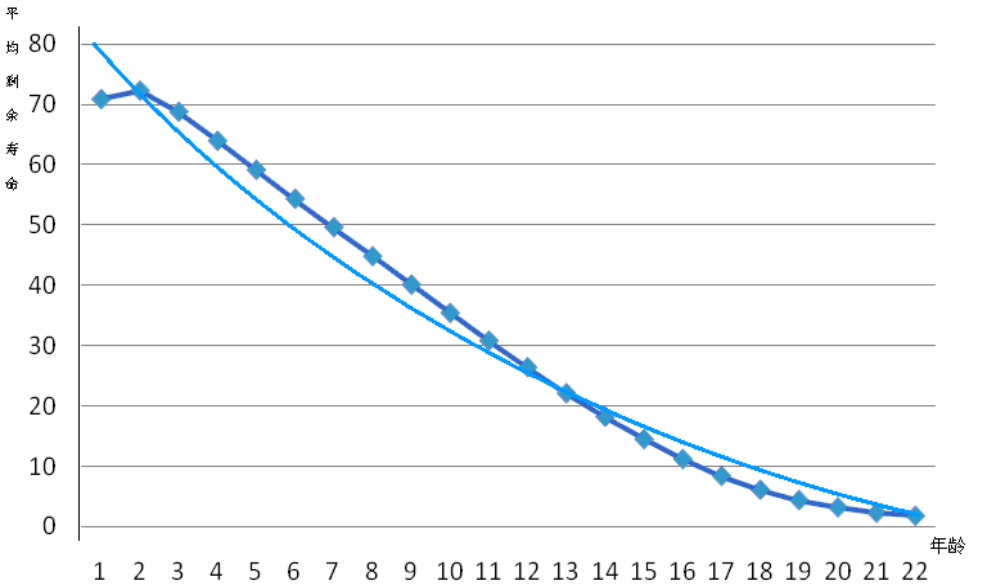


图 9.5 平均期望寿命图像

比较平均剩余生命图发现，指数分布并不适合描述该模型。指数分布图像是水平的的，而由观察可知，剩余寿命图像刚开始是上凸，然后向下倾斜，与指数分布图像不符，□

二、分布的拟合优度检验

根据实际经验或某种认识，有时可以认为总体的分布函数是一个已知的函数 $F_0(x)$，为检验这种认识是否正确，考虑以下假设检验问题：

H_0：$F(x)=F_0(x)$，H_1：$F(x)\neq F_0(x)$。

其中 F_0 分布形式已知，参数 θ 未知。通常把上述假设检验问题称为分布拟合优度检验。

方法：

χ^2 拟合检验

K－S 检验

Aderson-Darling test

Likelihood ratio test

1、χ^2 拟合检验

H_0： $F(x)=F_0(x;\theta)$， F_0 分布形式已知，参数 $\theta=(\theta_1,...,\theta_r)$ 未知。

H_1： $F(x)\neq F_0(x;\theta)$

先把观察记录按大小分组，如分为 m 组，每组包括组上限，即分组为 $(C_{i-1},C_i],i=1,\cdots,m$。左端点是开区间，右端点是闭区间，其中，$C_0=0$， C_m 的理论值是 $+\infty$，但在实际问题中，C_m 通常是某有限的数值，用所选择的分布模型计算出“理论平均值”：

$$E_0=nF(C_0,\hat{\theta})\,,$$

$$E_j=n[F(C_j,\hat{\theta})-F(C_{j-1},\hat{\theta})]\,,\quad j=1,\cdots,m-1$$

$$E_m = \mathrm{n}[1 - F(C_m,,\hat{\theta})]$$

其中 n 为样本容量，$\hat{\theta}$ 是未知参数的极大似然估计值。构建统计量

$$Q = \sum_{j=1}^{m} \frac{(n_j - E_j)^2}{E_j}$$

当观测值 n 的数量充分大时，统计量 Q 的分布会收敛于自由度为 m-k-1 的 χ^2 分布，所以可以将这个 χ^2 分布作为 Q 的分布的近似。即 $Q \sim \chi^2(m-k-1)$，k 为未知参数的个数。如果计算得到的统计量 $Q > \chi_\alpha^2(m-k-1)$，α 一般等于 0.05，则拒绝原假设，即认为选择的分布不可以拟合总体的分布。

在 χ^2 拟合检验中，一定要注意满足样本容量 n 要足够大，以及 E_i 不太小这两个条件。一般要求样本容量 n 不小于 50，每组的观察值个数不少于 5 个，否则需要将个数较少的组合并，使得 E_i 满足这个要求。

例 9.3：设某保险人经营某车辆险，对过去所发生的 1000 次理赔情况作了记录，平均理赔额是 2200 元，又按赔付金额分作 5 档，各档中的记录次数如下：

赔付	0-1000	1000-2000	2000-3000	3000-4000	4000-5000	5000 以上
次数	200	300	250	150	100	5

试用 χ^2 拟合检验判断是否能用指数分布来拟合个体理赔额的分布。

解：先设个别理赔额 X 服从指数分布，使用矩估计或最大似然估计可以估计出 $\hat{\lambda}=2200$。

下面计算 E_i， $E_i=nP_\theta(X\in(C_{i-1},C_i])=n(F(C_i;\theta)-F(C_{i-1};\theta))$

例如，在 2000-3000 组内的 E_3 为

$$1000\times\int_{2000}^{3000}\frac{1}{\lambda}e^{-x/\lambda}dx=e^{-2000/\lambda}-e^{-3000/\lambda}$$

$$=1000\times0.1472=147.2$$

类似的，可以计算得到其它组的平均次数：

$$E_1=365.3, E_2=231.8, E_4=93.4, E_5=59.3, E_6=103$$

因此，χ^2 统计量的值为

$$Q=\frac{(200-365.3)^2}{365.3}+\frac{(300-231.8)^2}{231.8}+\cdots+\frac{(5-103)^2}{103}=322.13$$

χ^2 分布的自由度为 6－1－1＝4，查表得，$\chi^2_{0.05}(4)=14.9$，因此

$Q \geqslant \chi^2_{0.05}(4)=14.9$，故应拒绝原假设，即选择指数分布不恰当。

例 9.4: 请用 χ^2 拟合检验指数分布是否可以用来拟合 data set B（truncated at 50）

27	82	115	126	155	161	243	294	340	384
457	680	855	877	974	1193	1340	1884	2558	3476

解：看图比较

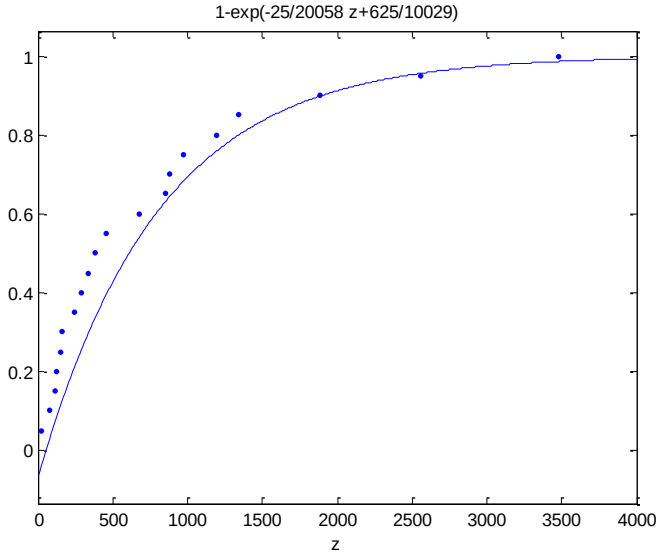

1-exp(-25/20058 z+625/10029)
1
0.8
0.6
0.4
0.2
0
0
500
1000
1500
2000
2500
3000
3500
4000
z

图 9.7

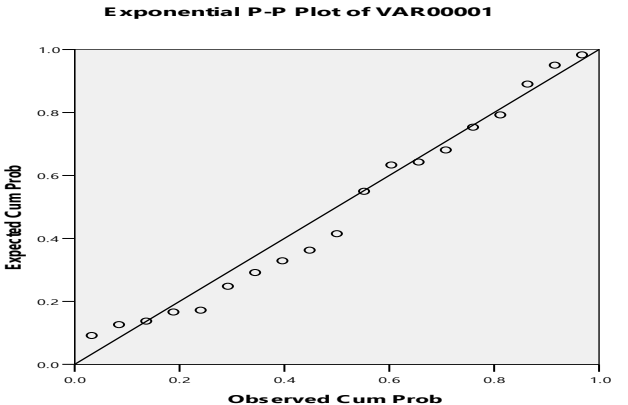


图 9.8

将数据分为 6 档，计算相应的 χ^2 值，得到下表：

Range	$\hat{p}$	Expected	Observed	χ^2
50-150	0.1172	2.227	3	0.2687
150-250	0.1035	1.966	3	0.5444
250-500	0.2087	3.964	4	0.0003
500-1000	0.2647	5.029	4	0.2105
1000-2000	0.2180	4.143	3	0.3152
2000-∞	0.0880	1.672	2	0.0644
合计	1	19	19	1.4034

自由度等于 4（6-1-1），5%的临界值为 9.4877，p 值为 0.8436。因此指数分布可以用来拟合 data set B（truncated at 50）。

例 9.5:下表是某保险公司统计的每天发生理赔事故的次数，请问是否可以用 Poisson 分布来拟合。

No. of claim/day	Observed no. of days

0	47
1	97
2	109
3	62
4	25
5	16
6	4
7	3
8	2
9+	0

解：

$$L = \prod_{k=0}^{\infty} p_k^{n_k}, l = \sum_{k=0}^{\infty} n_k \ln p_k$$

$$p_k = \frac{e^{-\lambda}\lambda^k}{k!},$$

$$l = \sum_{k=0}^{\infty} n_k(-\lambda + k\ln\lambda - \ln k!) = -\lambda n + \sum_{k=0}^{\infty} k n_k \ln\lambda - \sum_{k=0}^{\infty} n_k \ln k!$$

其中 $n = \sum_{k=0}^{\infty} n_k$ 。对 l 求微分得到

$$\frac{dl}{d\lambda} = -n + \sum_{k=0}^{\infty} k n_k \frac{1}{\lambda}$$

极大似然估计值为

$$\hat{\lambda} = \frac{\sum_{k=0}^{\infty} k n_k}{n} = \overline{x} = \frac{742}{365} = 2.0329$$

Claim/day	Observed	Expected	Chi Square
0	47	47.8	0.01
1	97	97.2	0.00
2	109	98.8	1.06
3	62	66.9	0.36
4	25	34	2.39
5	14	13.8	0.34
6	4	4.7	0.10
7+	5	1.8	5.66
Totals	365	365	9.93

注意最后一列的计算 $E_{7+} = n\hat{p}_k = n(1-\hat{p}_0-\hat{p}_1-\cdots-\hat{p}_{k-1})$。

χ^2 统计量为 9.93，自由度为 6 的 5%的临界值为 12.59，p 值为 0.1277，无法拒绝原假设，即认为 poisson 分布是可以用来描述该数据的。

例9.6 A particular line of business has three types of claims. The historical probability and the number of claims for each type in the current year are:

Type	Historical Probability	Number of Claims in Current Year
A	0.2744	112
B	0.3512	180
C	0.3744	138

You test the null hypothesis that the probability of each type of claim in the current year is the same as the historical probability.
Calculate the chi-square goodness-of-fit test statistic.

解：There are 430 observations. The expected counts are 430(.2744) = 117.99, 430(.3512) =151.02, 430(.3744) = 160.99.

The test statistic is

$$\frac{(112-117.99)^2}{117.99}+\frac{(180-151.02)^2}{151.02}+\frac{(138-160.99)^2}{160.99}=9.15.$$

课堂练习

23. For a sample of 15 losses, you are given:

(i)

Interval	Observed Number of Losses
(0, 2]	5
(2, 5]	5
(5, ∞)	5

(ii) Losses follow the uniform distribution on $(0,\theta)$.

Estimate θ by minimizing the function $\sum_{j=1}^{3}\frac{(E_j-O_j)^2}{O_j}$, where E_j is the expected number of losses in the jth interval and O_j is the observed number of losses in the jth interval.

(A) 6.0

(B) 6.4

(C) 6.8

(D) 7.2

(E) 7.6

Question #23
Key: E

Assume that $\theta > 5$. Then the expected counts for the three intervals are $15(2/\theta) = 30/\theta, 15(3/\theta) = 45/\theta$, and $15(\theta - 5)/\theta = 15 - 75/\theta$ respectively. The quantity to minimize is

$$\frac{1}{5}\left[(30\theta^{-1} - 5)^2 + (45\theta^{-1} - 5)^2 + (15 - 75\theta^{-1} - 5)^2\right].$$

Differentiating (and ignoring the coefficient of 1/5) gives the equation $-2(30\theta^{-1} - 5)30\theta^{-2} - 2(45\theta^{-1} - 5)45\theta^{-2} + 2(10 - 75\theta^{-1})75\theta^{-2} = 0$. Multiplying through by θ^3 and dividing by 2 reduces the equation to $-(30 - 5\theta)30 - (45 - 5\theta)45 + (10\theta - 75)75 = -8550 + 1125\theta = 0$ for a solution of $\hat{\theta} = 8550/1125 = 7.6$.

2、K—S 检验

这个检验常用于总体是连续分布的情况。假设 X 的分布是一个已知的连续分布函数 F_0，Y_1，…，Y_n 是独立同分布 F 的随机变量，但分布 F 未知，如果有 Y_1，…，Y_n 的一个样本观测值 y_1，…，y_n，考虑 Y_1，…，Y_n 是否与 X 同分布。所以假设检验问题还是：

H_0：$F(x)= F_0(x)$，H_1：$F(x)\neq F_0(x)$。

这个检验的思想是：虽然 Y_1，…，Y_n 的分布未知，但根据大样本理论 Y_1，…，Y_n 的经验分布函数 F_e 在某种意义下收敛于其真实的分布，所以可以把 F_n 与所假设的分布函数 F_0 作比较，看它们是否吻合。如果它们不能很好的吻合，就拒绝 H_0，即未知的真实分布函数 F 不是由 F_0 给定的。由于经验分布 $F_n(x)$和分布 $F_0(x)$都是 x 的函数，所以要比较两者的差异需要一个合适的度量。Kolmogorov-Smirnov 提出一个最简单的度量，就是用 $F_e(x)$和 $F_0(x)$在垂直方向上的最大距离作为统计量。如果

已知一个样本观测值 y_1，…，y_n，则 Kolmogorov-Smirnov 检验统计量为

$$D = \max_{i=1,\cdots,n}\{\left|F_n(y_{i-1}) - F(y_i,\hat{\theta})\right|, \left|F_n(y_i) - F(y_i,\hat{\theta})\right|\}$$

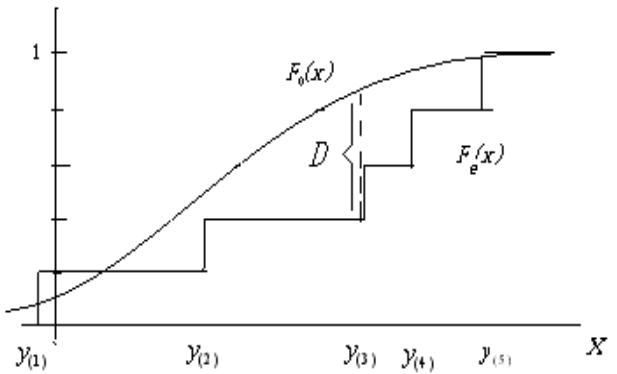


图 9.9 给定样本时 Kolmogorov-Smirnov 检验统计量

注意图中的 $F_e(x)$ 就是 $F_n(x)$

若存在 truncated or censored 情况，则检验统计量为

$$D = \max_{t \le x \le u} \left| F_n(x) - F(x, \hat{\theta}) \right|$$

其中 t 为左 truncated value(可以为 0), u 为右 censored value （可为无穷）

下表为不同显著性水平下的临界值：

显著性水平（α）	临界值
0.2	$1.07/\sqrt{n}$
0.1	$1.22/\sqrt{n}$
0.05	$1.36/\sqrt{n}$
0.01	$1.63/\sqrt{n}$

当 D 大于临界值时，拒绝原假设。

例 9.7：假设 data set B　truncated at 50。若用指数分布来拟合这个数据，估计它们的参数，并使用 K-S 检验来说明指数分布是否为合适的分布。

27	82	115	126	155	161	243	294	340	384
457	680	855	877	974	1193	1340	1884	2558	3476

解：对于数据集 B，truncated 后只有 19 个观察值，极大似然估计值为 $\hat{\theta}=802.32$，分布 $F(x,\hat{\theta})$ 为

$$F^*(x)=\frac{1-e^{-x/802.32}-(1-e^{-50/802.32})}{1-(1-e^{-50/802.32})}=1-e^{-(x-50)/802.32}$$

将经验分布函数和估计分布画图如下：

x	$F^*(x)$	$F_n(x-)$	$F_n(x)$	Maximum difference
82	0.0391	0.000	0.0526	0.0391

115	0.0778	0.0526	0.1053	0.0275
126	0.0904	0.1053	0.1579	0.0675
155	0.1227	0.1579	0.2105	0.0878
161	0.1292	0.2105	0.2632	0.134
243	0.2138	0.2632	0.3158	0.102
294	0.2622	0.3158	0.3684	0.1062
340	0.3033	0.3684	0.4211	0.1178
384	0.3405	0.4211	0.4737	0.1332
457	0.3979	0.4737	0.5263	0.1284
680	0.5440	0.5263	0.5789	0.0349
855	0.6433	0.5789	0.6316	0.0644
877	0.6433	0.6316	0.6842	0.0409
974	0.6839	0.6842	0.7368	0.0529
1193	0.7594	0.7368	0.7895	0.0301
1340	0.7997	0.7895	0.8421	0.0424

1884	0.8983	0.8421	0.8947	0.0562
2558	0.9561	0.8947	0.9474	0.0614
3476	0.9860	0.9474	1.000	0.0386

D 统计量为 0.134, 5%的临界值为$1.36/\sqrt{19}=0.3120$，0.134<0.3120, 说明指数分布是合适的分布。

当$n\to\infty$时，若$F^*(x)$的函数形式完全给定，Y的近似分布为

$$P(Y > x) = 2\sum_{j=1}^{\infty}(-1)^{j+1}e^{-2j^2x^2}$$

若 $F^*(x)$ 的形式已知，参数由数据估计，偏差度量 $Y = \sqrt{n}D_n$ 将取得与前面结果不同的概率值，在这种情况下则需要对进行修正。例如，假定一个指数模型，θ 由数据估计而得，则

$$Y^* = (D_n - \frac{0.2}{n})(\sqrt{n} + 0.26 + \frac{0.5}{\sqrt{n}})$$

为一个比未修正 $Y = \sqrt{n}D_n$ 更好的估计。关于这些修正的具体方法和临界值表，可见 Stephens, M.(1986)的论文。

例9.8： You are given:

(i) A sample of claim payments is:

29 64 90 135 182

(ii) Claim sizes are assumed to follow an exponential distribution.

(iii) The mean of the exponential distribution is estimated using the method of moments.

Calculate the value of the Kolmogorov-Smirnov test statistic.

(A) 0.14

(B) 0.16

(C) 0.19

(D) 0.25

(E) 0.27

Key:E

X	$F_n(x)$	$F_n(x^-)$	$F_0(x)$	$\lvert F_n(x)-F_0(x)\rvert$	$\lvert F_n(x^-)-F_0(x)\rvert$
29	0.2	0	0.252	0.052	0.252
64	0.4	0.2	0.473	0.073	0.273
90	0.6	0.4	0.593	0.007	0.193
135	0.8	0.6	0.741	0.059	0.141
182	1.00	0.8	0.838	0.162	0.038

where:

$\hat{\theta} = \bar{x} = 100$ and $F_0(x) = 1 - e^{-x/100}$.

The maximum value from the last two columns is 0.273.

```
library(goftest)
library(MASS)
library(VGAM)
```

```r
## Read in data and get number of claims.

claim_lev <- read.csv("Data/CLAIMLEVEL.csv", header
 = TRUE)

  nrow(claim_lev)  # 6258

  claim_data <- subset(claim_lev, Year == 2010);

  length(unique(claim_data$PolicyNum))  # 403 uniqu
e policyholders

# Inference assuming a gamma distribution

fit.gamma_2 <- glm(Claim ~ 1, data = claim_data, fam
ily = Gamma(link = log))

  summary(fit.gamma_2, dispersion = gamma.dispersio
n(fit.gamma_2))
```

```r
# Kolmogorov-Smirnov # the test statistic is "D"

ks.test(claim_data$Claim, "pgamma", shape = alpha, s
cale = theta)

   fit.pareto <- vglm(Claim ~ 1, paretoII, loc = 0, d
ata = claim_data)

   summary(fit.pareto)

   ks.test(claim_data$Claim, "pparetoII", loc = 0, s
hape = exp(coef(fit.pareto)[2]),        scale = exp
(coef(fit.pareto)[1]))
```

3、Anderson-Darling 检验

H_0：$F(x)=F_0(x;\theta)$，F_0 分布形式已知，参数 $\theta=(\theta_1,...,\theta_r)$ 未知。

H_1：$F(x)\neq F_0(x;\theta)$

检验统计量为

$$A^2=n\int_t^u \frac{[F_n(x)-F^*(x)]^2}{F^*(x)[1-F^*(x)]}f^*(x)dx$$

注意当 x 接近于 t 或 u 时，分母很小，从而权重较大，因此这个统计量更加看重尾部的估计。

这个积分很难计算，但是对于个体数据来说，其中 uncensored 数据点是 $t=y_0<y_1<\cdots<y_{k+1}=u$，则积分变为求和：

$$A^2 = -nF^*(u) + n\sum_{j=0}^{k-1}[1-F_n(y_j)]^2\{\ln(1-F^*(y_j)) - \ln(1-F^*(y_{j+1}))\}$$

$$+n\sum_{j=1}^{k}F_n(y_j)^2[\ln F^*(y_{j+1}) - \ln F^*(y_j)]$$

若 u=∞，则第一个求和项的最后一项等于 0。

Anderson-Darling 统计量的 10%，5%和 1%的临界值为 1.933，2.492，3.857.

例 9.9： data set B

j	Y_j	$F^*(x)$	$F_n(x-)$	Summand
0	50	0.000	0.000	0.0399
1	82	0.0391	0.0526	0.0388
2	115	0.0778	0.1053	0.0126
3	126	0.0904	0.1579	0.0332
4	155	0.1227	0.2105	0.007
5	161	0.1292	0.2632	0.0904
6	243	0.2138	0.3158	0.0501
7	294	0.2622	0.3684	0.0426
8	340	0.3033	0.4211	0.0389
9	384	0.3405	0.4737	0.0601

10	457	0.3979	0.5263	0.1490
11	680	0.5440	0.5789	0.0897
12	855	0.6433	0.6316	0.0099
13	877	0.6433	0.6842	0.0407
14	974	0.6839	0.7368	0.0758
15	1193	0.7594	0.7895	0.0403
16	1340	0.7997	0.8421	0.0994
17	1884	0.8983	0.8947	0.0592
18	2558	0.9561	0.9474	0.0308
19	3476	0.9860	1.0	0.0141
20	∞	1.0	1.0	0
合计				1.0226

例如：$0.0399=\ln\frac{1-0}{1-0.0391}$，

$0.0126=(1-0.1053)^2[\ln(1-0.0778)-\ln(1-0.0904)]+0.1053^2[\ln(0.0904)-\ln(0.0778)]$

Anderson-Darling 统计量等于-19(1)+19(1.0226)=0.4292，小于 5%临界值 2.492，因此指数分布可以做为拟和分布。

```
library(goftest)


# Anderson-Darling  # the test statistic is "An"

   ad.test(claim_data$Claim, "pgamma", shape = alpha, scale = theta)
```

```
   ad.test(claim_data$Claim, "pparetoII", loc = 0, s
hape = exp(coef(fit.pareto)[2]),
           scale = exp(coef(fit.pareto)[1]))
```

三种统计量的比较

注意：当样本量变大（例如扩大 2 倍时），样本的值变化不大，但是

- K-S 检验统计量没有多大的变化，但是临界值将变小，
- Anderson-Darling 统计量将发生较大的变化，而临界值没有变化

因此，使用这两种方法，在样本量扩大时，容易导致拒绝原假设。

例9.10：If the proposed model is appropriate, which of the following tends to zero as the sample size goes to infinity?
(A) Kolmogorov-Smirnov test statistic
(B) Anderson-Darling test statistic
(C) Chi-square goodness-of-fit test statistic
(D) Schwarz Bayesian adjustment
(E) None of (A), (B), (C) or (D)

Key:A

4、似然比（likelihood ratio）检验

问题：当分布族 A（未知参数个数为 k_0）是分布族 B（未知参数个数为 k_1，$k_1>k_0$）的特殊情形时（例如指数分布是 gamma 分布的特例），且两种分布可以用来拟合某个数据集时，这时该选择哪种分布呢。

假设：

H_0：样本来自分布 A

H_1：样本来自分布 B

似然比检验：设$L(\theta)$表示参数为θ的似然函数，设θ_0为原假设中分布 A 的极大似然估计值，$L_0 = L(\theta_0)$为极大似然函数值，设θ_1是对立假设中分布 B 的极大似然估计值，$L_1 = L(\theta_1)$为极大似然函数值。则统计量

$$T = 2\ln(L_1 / L_0) = 2(\ln L_1 - \ln L_0)$$

服从参数为$k_1 - k_0$的χ^2分布，若 T 大于置信水平为α临界值 c，则拒绝原假设，即分布 B 更适合用来描述观察数据集

例 9.11：假设已用 gamma 分布来拟合数据集 B（原始数据集），请用似然比检验该数据集的均值是否等于 1200。

解：构造检验假设如下：

H_0: gamma with $\mu = 1200$

H_1: gamma with $\mu \neq 1200$

经计算，未受约束的 gamma 分布的极大似然估计值为 $\hat{\alpha} = 0.55616, \hat{\theta} = 2561$，对数极大似然值为 $\ln L_1 = -163.293$。在原假设下，$\alpha\theta = 1200$，从而 $\theta = 1200/\alpha$，这时只要一个未知参数，重新计算极大似然估计值得到 $\hat{\alpha} = 0.54955, \hat{\theta} = 2183.6$，对数极大似然值为 $\ln L_0 = -162.466$。似然比检验统计量 $T = 2(-162.293 + 162.466) = 0.346$，

一个自由度的 5%的临界值为 3.8415。由于 0.346<3.8415，因此无法拒绝原假设。

例9.12: You fit a Pareto distribution to a sample of 200 claim amounts and use the likelihood ratio test to test the hypothesis that $\alpha = 1.5$ and $\theta = 7.8$.

You are given:

(i) The maximum likelihood estimates are $\hat{\alpha} = 1.4$ and $\hat{\theta} = 7.6$.

(ii) The natural logarithm of the likelihood function evaluated at the maximum likelihood estimates is -817.92.

(iii) $\Sigma \ln(xi+7.8) = 607.64$

Determine the result of the test.

(A) Reject at the 0.005 significance level.

(B) Reject at the 0.010 significance level, but not at the 0.005 level.

(C) Reject at the 0.025 significance level, but not at the 0.010 level.
(D) Reject at the 0.050 significance level, but not at the 0.025 level.
(E) Do not reject at the 0.050 significance level.
Key:C

The likelihood function is $L(\alpha,\theta)=\prod_{j=1}^{200}\frac{\alpha\theta^{\alpha}}{(x_j+\theta)^{\alpha+1}}$ and its logarithm is

$l(\alpha,\theta)=200\ln(\alpha)+200\alpha\ln(\theta)-(\alpha+1)\sum_{i=1}^{200}\ln(x_i+\theta)$. When evaluated at the hypothesized values of 1.5 and 7.8, the loglikelihood is –821.77. The test statistic is 2(821.77–817.92) = 7.7. With two degrees of freedom (0 free parameters in the null hypothesis versus 2 in the alternative), the test statistic falls between the 97.5th percentile (7.38) and the 99th percentile (9.21).

二、选择合适的分布

当存在多种分布可以用来拟合理赔数据的分布时，精算师应该选择一个“最优”的分布作为理赔数据的分布。一般来说，选择的标准有

1、 最大似然函数值（越大越好）；

2、 拟合检验中 χ^2 统计量 Q 的值。（越小越好）；

3、 拟合检验中 p 值， $p = P(\chi^2 > Q)$，（越大越好）；

4、 K-S 检验统计量（越小越好）。

5、 Anderson-Darling 检验统计量（越小越好）。

在选择最优分布时还要注意以下两点

1. 简洁原则与精确原则的权衡，上述几种方法，除了卡方 p 值的方法外，其余方法在节俭原则都有缺陷。
2. 要根据实际情况来选择，如尾部估计要精细，或分布的适用性等。

方法：

1. 主观判断法
2. 评分法

3. 表 9.1：评分方法及判断标准

方法	分数依据	判断标准
似然函数法	极大似然函数值	越大越好
χ^2 拟合优度检验法	χ^2 统计量的值	越小越好
χ^2 拟合优度 p 值法	p 值： $p = P(\chi^2 > Q)$	越大越好
K-S 检验法	K-S 检验统计量	越小越好
Anderson-Darling 检验法	Anderson-Darling 统计量	越小越好

例 9.13：用几种分布对某分组数据进行分布拟合，得到结果如下：

分布类型	参数个数	NLL	χ^2	P－value
指数分布	1	548.72	81.02	<0.0001
逆指数分布	1	520.27	49.06	<0.0001
对数正态分布	2	498.29	4.51	0.8744
逆高斯分布	2	502.26	12.95	0.1648
佩尔托分布	2	499.31	6.37	0.7028
逆佩尔托分布	2	500.09	7.52	0.5831
Loglogistist 分布	2	499.93	7.51	0.5847
Gamma 分布	2	507.84	16.38	0.0372
逆 gamma 分布	2	509.80	26.77	0.008

Weibull 分布	2	501.63	8.16	0.4183
逆 weibull 分布	2	506.72	20.27	0.0163
Paralogistic 分布	2	499.79	7.30	0.6055
逆 Paralogistic 分布	2	500.01	7.58	0.5767
Burr 分布	3	498.41	4.79	0.7793
逆 burr 分布	3	499.01	5.33	0.7220
广义 pareto 分布	3	498.62	5.00	0.7580

从中选择一个最合适的分布类型。

如果使用 p 值为检验标准，显然对数正态分布是最合适的分布。
若用似然值作为选择标准，

- 在两参数分布中，对数正态分布的似然值最大，为-498.29；
- 在三参数分布中，Burr 分布的似然值最大，为－498.41。

参数个数的选择：

- 似然比检验
- Schwarz Bayesian 准则（SBC）

SBC 准则衡量模型时将对数似然函数值减 $(r/2)\ln n$，即 $SBC = \ln L - (r/2)\ln n$，其中 r 是待估参数的个数，n 是样本容量。这样一来，只有对数似然函数增加 $0.5\ln n$ 才能增加一个参数，样本容量越大，要求的似然函数增量就越大，但这个要求的增量并非随样本容量比例增长。

- AIC 信息准则之类的判定准则。

AIC 信息量准则（ Akaike information criterion）也是权衡所估计模型的

复杂度和此模型拟合数据的优良性的一种标准。在一般的情况下，AIC 可以表示为 $AIC = 2r - 2\ln L$ 。在 AIC 准则下，通常选择模型应是 AIC 值最小的那一个。

例 9.14：请比较 weibull 分布和指数分布哪个更适合于描述数据集 B

Table 16.14 Results for Example 16.11.

	B truncated at 50		B censored at 1,000	
Criterion	Exponential	Weibull	Exponential	Weibull
K–S*	0.1340	0.0887	0.0991	0.0991
A–D*	0.4292	0.1631	0.1713	0.1712
χ^2	1.4034	0.3615	0.5951	0.5947
p-value	0.8436	0.9481	0.8976	0.7428
Loglikelihood	−146.063	−145.683	−113.647	−113.647
SBC	−147.535	−148.628	−115.145	−116.643
	C			
χ^2	61.913	0.3698		
p-value	10^{-12}	0.9464		
Loglikelihood	−214.924	−202.077		
SBC	−217.350	−206.929		

*K–S and A–D refer to the Kolmogorov–Smirnov and Anderson–Darling test statistics, respectively.

。

对于以上的结果，可以发现韦伯分布都有一定的优势。因此在备选模型为指数分

布和韦伯分布时，我们理应选择韦伯分布。

例 9.15 我国某保险公司 1996 年的 35072 辆投保车辆索赔次数统计结果如表第 1、2 列所示，试分析索赔次数的分布。

索赔次数 k	保单数 n_k	$k\dfrac{n_k}{n_{k-1}}$
0	27141	
1	5789	4.69
2	1443	8.02
3	457	9.47
4	155	11.79
5	56	13.84
6	27	12.44
7	2	94.5
8	1	16
9	1	9
≥ 10	0	
总计	35072	

解：首先可以按照近似公式(4.3.20)计算 $k\dfrac{n_k}{n_{k-1}}$ 的值如表 4-3 第 3 列所示，然后绘出其图形，见图 9.6。

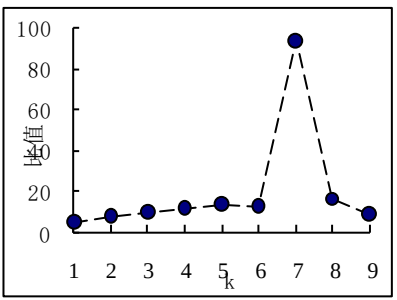


图 9.6　函数 $k\dfrac{n_k}{n_{k-1}}$ 的曲线

从图 9.6 可以看出，除了 $k=7$ 这个点之外，其余各点近似呈一条直线，并且斜率为正，因此可以考虑用负二项分布来拟合索赔次数。当然，斜率为正也必须通过假设检验才能确定。

我们初步判断可以采用负二项分布拟合，但是不能排除泊松分布。因为斜率近似为 0。

首先估计泊松分布，由表的数据可以算出样本均值和方差分别是 $\bar{x}=0.3176, s^2=0.4913$，以样本均值作为泊松分布的参数，可以得出实际观测数和拟合频数比较，如表 9.2 所示，

表 9.2 实际观测值和用 P(0.3176)拟合结果比较

	车辆数

索赔次数	观测值	拟合值	误差
0	27141	25528.69	1612.31
1	5789	8107.91	-2318.91
2	1443	1287.54	155.46
3	457	136.31	320.69
4	155	10.82	144.18
5	56	0.69	55.31
6	27	0.04	26.96
7	2	0	2
8	1	0	1

9	1	0	1
≥ 10	0	0	0
总计	35072	35072	0

其次，采用二元结构模型，用两个泊松分布的混合再次拟合，结果见表 9.3：

表 9.3 实际观测值、利用一元结构函数以及二元结构函数拟合结果的比较

索赔次数	车辆数				
	观察值	拟合值		误差	
		$N \sim P(0.3176)$	$N \sim 0.8876P(0.1719) + 0.1124P(1.4694)$	一元结构函数	二元结构函数
0	27141	25528.69	27120.34	1612.31	20.66

1	5789	8107.91	5838.70	-2318.91	-49.7
2	1443	1287.54	1366.37	155.46	76.63
3	457	136.31	501.74	320.69	-44.74
4	155	10.82	177.12	144.18	-22.12
5	56	0.69	51.81	55.31	4.19
6	27	0.04	12.68	26.96	14.32
7	2	0	2.66	2	-0.66
8	1	0	0.49	1	0.51
9	1	0	0.08	1	0.92
≥ 10	0	0	0.01	0	-0.01

总计	35072	35072	35072	0	0

从表 9.3 可以看出，用混合泊松分布拟合的误差明显降低。

最后，结构函数采用伽玛函数，参数分别为α 和θ，有

$$\begin{cases} E(N)=\alpha\theta=\bar{x} \\ Var(N)=\alpha\theta+\alpha\theta^2=s^2 \end{cases}$$

因此α 和θ 的矩估计值分别为

$$\begin{cases} \hat{\alpha}=\bar{x}^2/(s^2-\bar{x})=0.5807 \\ \hat{\theta}=(s^2-\bar{x})/\bar{x}=0.5469 \end{cases}$$

即负二项分布的参数估计值分别为

$$\hat{r} = \hat{\alpha} = 0.5807$$

$$\hat{p} = \frac{1}{1+\hat{\theta}} = 0.6464$$

再次拟合的结果见表 9.4：

表 9.4　实际观测值、利用一元结构函数、二元结构函数以及伽玛结构函数拟合结果的比较

索赔次数	观察值	车辆数					
		拟合值			误差		
		$N \sim P(0.3176)$	$N \sim 0.8876P(0.1719) + 0.1124P(1.4694)$	$N \sim NB(0.5807, 0.6464)$	一元结构函数	二元结构函数	伽玛结构函数

0	27141	25528.69	27120.34	27223.0	1612.31	20.66	-82
1	5789	8107.91	5838.70	5589.2	-2318.91	-49.7	199.8
2	1443	1287.54	1366.37	1561.8	155.46	76.63	-118.8
3	457	136.31	501.74	475.0	320.69	-44.74	-18
4	155	10.82	177.12	150.3	144.18	-22.12	4.7
5	56	0.69	51.81	48.7	55.31	4.19	7.3
6	27	0.04	12.68	16.0	26.96	14.32	11
7	2	0	2.66	5.3	2	-0.66	-3.3
8	1	0	0.49	1.8	1	0.51	-0.8
9	1	0	0.08	0.6	1	0.92	0.4

≥10	0	0	0.01	0.3	0	-0.01	-0.3
总计	35072	35072	35072	35072	0	0	0

从表 9.4 可以看出，二元混合泊松分布对前三个点的拟合效果较好，而负二项分布则对尾部的拟合效果较好。

```
rm(list = ls())
k<-c(0:9)
nk<-c(27141,5789,1443,457,155,56,27,2,1,1)
n=length(nk)
p<-rep(NA,n-1)
for(i in 1:n-1){
  p[i]=k[i+1]*nk[i]/nk[i+1]
}
p=round(p,2)

plot(k[-1],p,type = "b",lty=2,col="blue")
```

```r
#假设服从泊松分布
lambda=round(crossprod(k,nk)/sum(nk),4)  #估计 Lamda 的值，即为均值
nk1<-rep(NA,n) #存放泊松分布的拟合值
nk2<-rep(NA,n) #存放泊松分布拟合后的误差
for(i in 1:n){
  nk1[i]=round(dpois(k[i],lambda = lambda)*sum(nk),2)
  nk2[i]=nk[i]-nk1[i]
}

data1=data.frame("索赔次数"=c(0:9),"观测值"=nk,"拟合值"=nk1,"误差"=nk2)
data1

#假设服从混合泊松分布
a=0.8876
```

```r
lambda1=0.1719
lambda2=1.4694

nk_1<-rep(NA,n) #存放泊松分布的拟合值
nk_2<-rep(NA,n) #存放泊松分布拟合后的误差
for(i in 1:n){
  nk_1[i]=round((a*dpois(k[i],lambda = lambda1)+((1-a)*dpois
(k[i],lambda = lambda2)))*sum(nk),2)
  nk_2[i]=nk[i]-nk_1[i]
}

data2=data.frame("索赔次数"=c(0:9),"观测值"=nk,"拟合值 1"=nk1,"
拟合值 2"=nk_1,"误差 1"=nk2,"误差 2"=nk_2)
data2
```

```
#假设服从负二项分布
x_=0.3176 #已知均值
ssquare=0.4913 #已知方差
alpha=round(x_^2/(ssquare-x_),4)
theta=round((ssquare-x_)/x_,4)
r=alpha
p=1/(1+theta) #负二项分布的参数
nk_3<-rep(NA,n) #存放泊松分布的拟合值
nk_4<-rep(NA,n) #存放泊松分布拟合后的误差
for(i in 1:n){
  nk_3[i]=round(dnbinom(k[i],r,p)*sum(nk),2)
  nk_4[i]=nk[i]-nk_3[i]
}

data3=data.frame("索赔次数"=c(0:9),"观测值"=nk,"泊松分布拟合值"
=nk1,"混合泊松分布拟合值"=nk_1,"负二项分布拟合值"=nk_3,"泊松分布
```

```
误差"=nk2,"混合泊松分布误差"=nk_2,"负二项分布误差"=nk_4)
data3
```

例 9.16 下面事故数据，总共有 67856 个机动车事故保单。试选择一个合适的分布描述该数据。

表 9.5 机动车事故保单数据

索赔次数	**0**	**1**	**2**	**3**	**4**
频数	63232	4333	271	18	2

解：由表格可以看出，索赔次数的分布比较集中，只包括 0，1，2，3，4 五种情况，并且索赔次数为 0 次的保单数最多，占总保单数的 93.19%。而且索赔次数越

高，保单数越少。再具体研究索赔次数数据的特征。几个相关的统计量如下：

表 9.6 索赔次数的相关统计量

最小值	最大值	均值	方差
0	4	0.07275701	0.07739737

从上面数字特征可以看出，索赔次数数据的方差要大于均值，在常用的（a,b,0）分布类中，负二项分布具有这种性质，而且负二项分布具有两个参数，因此比泊松分布在拟合上会更加灵活。

通过索赔次数的频数表可以看出，索赔次数为 0 次的保单数最多，占总保单数的 93.19%。考虑到零膨胀现象，因此下面对零点的概率做适当的调整，使它正

好等于一个特定值，同时调整其他非零点上的概率，使所有概率值和等于 1。

表 9.3 对泊松、负二项、零修正的泊松、零修正的负二项作综合的比较。从表中可以看出，从拟合优度检验的 p 值来看，负二项分布的效果是最好的。从似然值上来看，零修正的负二项分布的似然值最大，效果有一定提高。但是从 SBC 准则看，从实际拟合值上来看，负二项分布是最优选择。

表 9.7 四种不同分布的比较

索赔数	观测值	用于拟合的分布			
		泊松	负二项	零修正的泊松	零修正的负二项
0	63232	63094.323	62333.100	63230.000	63230.000

1	4333	4590.544	4321.293	4321.707	4330.303
2	271	166.998	276.273	286.220	270.228
3	18	4.050	17.209	12.637	18.945
4+	2	0.075	1.125	5.436	6.524
参数		λ=0.072757	r=1.1560670 p=0.9407908	λ=0.132457	r=0.4594147 p=0.9144807
卡方		177.153919	0.82193845	5.29	3.1881106
自由度		3	2	3	2
p 值		<0.01	0.6630	0.1520	0.3635

对数似然	-18101.5	-18049.68	-18052.2	-18049.47
SBC	-18107.1	-18060.8	-18063.3	-18066.2

Table 7.1 Hossack et al. [48] data.

No. of claims	Observed frequency
0	565,664
1	68,714
2	5,177
3	365
4	24
5	6
6+	0

The mean, variance, and third central moment are 0.1254614, 0.1299599, and 0.1401737, respectively. For these numbers,

$$\frac{\mu_3 - 3\sigma^2 + 2\mu}{(\sigma^2 - \mu)^2/\mu} = 7.543865.$$

From among the Poisson – binomial, negative binomial, Polya – Aeppli, Neyman Type A, and Poisson – ETNB distributions, only the latter is appropriate. For this distribution, an estimate of r can be obtained from

$$7.543865 = \frac{r+2}{r+1},$$

resulting in $r = -0.8471851$.

Table 16.15 Results for Example 16.12

No. of claims	Observed frequency	Fitted distributions: Negative binomial	Poisson–inverse Gaussian	Poisson–ETNB
0	565,664	565,708.1	565,712.4	565,661.2
1	68,714	68,570.0	68,575.6	68,721.2
2	5,177	5,317.2	5,295.9	5,171.7
3	365	334.9	344.0	362.9
4	24	18.7	20.8	29.6
5	6	1.0	1.2	3.0
6+	0	0.0	0.1	0.4
Parameters		$\beta = 0.0350662$ $r = 3.57784$	$\lambda = 0.123304$ $\beta = 0.0712027$	$\lambda = 0.123395$ $\beta = 0.233862$ $r = -0.846872$
Chi square		12.13	7.09	0.29
Degrees of freedom		2	2	1
p-value		<1%	2.88%	58.9%
−Loglikelihood		251,117	251,114	251,109
SBC		−251,130	−251,127	−251,129

R 代码

```
#加载 MASS、survival、fitdistrplus、gamlss.dist 包
library(MASS)
library(survival)
library(fitdistrplus)
library(gamlss.dist)

#理赔次数观察值
claim = c(0, 1, 2, 3, 4)
#样本数的观察值
freq = c(63232, 4333, 271, 18, 2)
#把理赔次数观察值整理为一个向量
num = rep(claim, freq)

#用泊松分布拟合数据，应用极大似然法估计参数
fit1 = fitdist(num, "pois", method = "mle")
```

#查看 fit1 的参数及模型拟合度
summary(fit1)

##
Fitting of the distribution ' pois ' by maximum likelihood
Parameters :
estimate Std. Error
lambda 0.07275702 0.001035288
Loglikelihood: -18101.5 AIC: 36205 BIC: 36214.13

#用负二项分布拟合数据，应用极大似然法估计参数
fit2 = fitdist(num, "nbinom", method = "mle")
#查看 fit2 的参数及模型拟合度
summary(fit2)

##其中 size 表示的是参数 r，mu 表示的是均值，即 r β

Fitting of the distribution ' nbinom ' by maximum likelihood

Parameters :

	estimate	Std. Error
size	1.15881641	0.143342065
mu	0.072759355	0.001067318

Loglikelihood: -18049.68 AIC: 36103.36 BIC: 36121.61

Correlation matrix:

	size	mu
size	1.000000e+00	- 1.230828e-05
mu	-1.230828e-05	1.000000e+00

#用零修正泊松分布拟合数据，应用极大似然法估计参数

fit3 = fitdist(num, "ZAP", method = "mle", start = list(mu = 0.1, sigma = 0.9))

#查看 fit3 的参数及模型拟合度

summary(fit3)

##mu 为均值，在零调整泊松分布中与 λ 相等，sigma 表示为零点的概率大小
Fitting of the distribution ' ZAP ' by maximum likelihood
Parameters :

	estimate	Std. Error
mu	0.13239997	0.0074038604
sigma	0.9318645	0.0009670657

Loglikelihood: -18052.2 AIC: 36108.4 BIC: 36126.65
Correlation matrix:

	mu	sigma
mu	1.000000e+00	-6.511999e-12
sigma	-6.511999e-12	1.000000e+00

#用零修正负二项分布拟合数据，应用极大似然法估计参数
fit4 = fitdist(num, "ZANBI", method = "mle", start = list(mu = 0.1, sigma = 0.1, nu = 0.9))
#查看 fit4 的参数及模型拟合度

summary(fit4)

##mu 为零调整负二项分布的均值，sigma 为其方差，nu 表示为零点的概率大小

Fitting of the distribution ' ZANBI ' by maximum likelihood

Parameters :

	Estimate	Std. Error
mu	0.04332845	0.0329219659
sigma	2.14631769	2.4220094599
nu	0.93185786	0.0009671525

Loglikelihood: -18049.46 AIC: 36104.93 BIC: 36132.3

Correlation matrix:

	mu	sigma	nu
mu	1.000000e+00	-9.969517e-01	-2.123954e-09
sigma	-9.969517e-01	1.000000e+00	2.130448e-09
nu	-2.123954e-09	2.130448e-09	1.000000e+00

表 7—21 列出了 178 张保单的损失额不超过 200 的索赔(以千元为单位)。

表 7—21

损失范围	保单数	损失范围	保单数
1~5	3	41~50	19
6~10	12	51~75	28
11~15	14	76~100	21
16~20	9	101~125	15
21~25	7	126~150	10
26~30	7	151~200	15
31~40	18		

另外还有 22 张保单损失超过 200 的索赔如下:206,219,230,235,241,272,283,286,312,319,385,427,434,555,562,584,700,711,869,980,999,1506。请选择合适的分布拟合该数据。

【解】先画出这组数据的直方图(见图 7—2)。

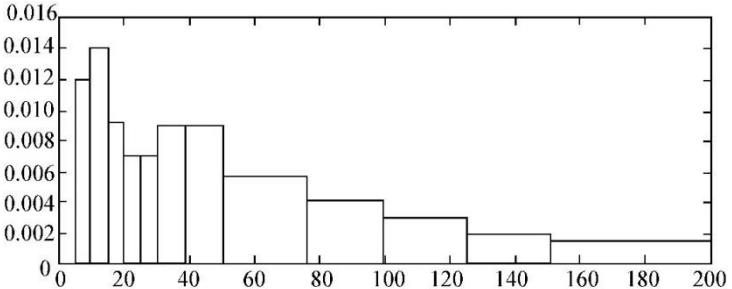


图 7—2　直方图

从直方图（图 7-2）可看出,潜在的分布有一个非零的众数点。为了检查尾部的情况,我们计算了一些点的经验平均超额函数(见表 7—22),近似为常数,因此指数模型比较合适。平均超额函数接近常数,故可以用指数分布近似。

表 7—22

损失	平均超额函数
200	314
300	367
400	357
500	330
600	361
700	313
800	289
900	262

我们首先选择一个分两段的模型,200(以千元为单位)之前为经验模型,之后为指数模型。

$$f(x,\theta)=\begin{cases}F_{178}(x), & x \le 200\\ af(x), & x > 200\end{cases}$$

至少有两种方法确定指数模型。一种方法是限制参数,使损失超过 200 的概率为 11%(22/200),即

$$\int_{200}^{\infty} \frac{1}{\theta} e^{-x/\theta} dx = 11\%$$

由指数分布的分布函数知 $e^{-200/\theta}=0.11$,从而 $\theta=90.61$。因此

$$f(x)=\frac{1}{90.61} e^{-x/90.61}, \quad x>200$$

另一种方法是独立于 11%的要求估计指数模型,然后乘上密度函数使其超过 200 的区域为 0.11。利用极大似然估计,得

$$L=\prod_{i=1}^{22} \frac{1}{\theta} \exp(-x_i/\theta) \Rightarrow \theta=314$$

$$\int_{200}^{\infty} af(x;314)dx=11\% \Rightarrow a=0.11/\int_{200}^{\infty} f(x;314)dx=0.208$$

$$f(x)=0.208\times\frac{1}{314}e^{-x/314}=0.000662344e^{-x/314}$$

由此得到的参数估计为 θ=314。位于 200 以下的部分经验分布在每个观测点的概率为 1/200,则两种方法确定的密度函数如图 7—3 所示。

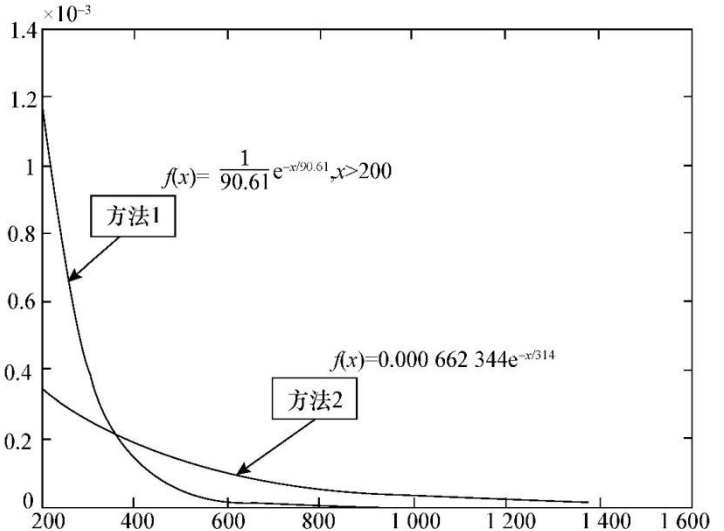


图 7—3　指数分布的密度函数

通过比较两种方法的分布函数与经验分布函数(见图 7—4 和图 7—5)可以看出,方法 2 的拟合效果较好,因此选择后一个模型。

其次,选择单参数模型。由直方图可知,如果选择单参数模型,则应该考虑非零众数。因为数据都在 1000 左右,区间可以设为 0.5~5.5,5.5~10.5,等等。可以考虑对数正态、威布尔、伽玛和混合模型(包括指数分布)等分布,对数正态分布明显是最合适的(利用 SBC 准则)。参数 $\hat{\mu}$=4.0626, $\hat{\sigma}$=1.1466。χ^2 拟合优度检验(将超过 200 的观测值单放在一组)的统计量为 7.77,p 值为 0.73。图 7—6 将对数正态模型和经验模型进行了比较,说明拟合效果很好。

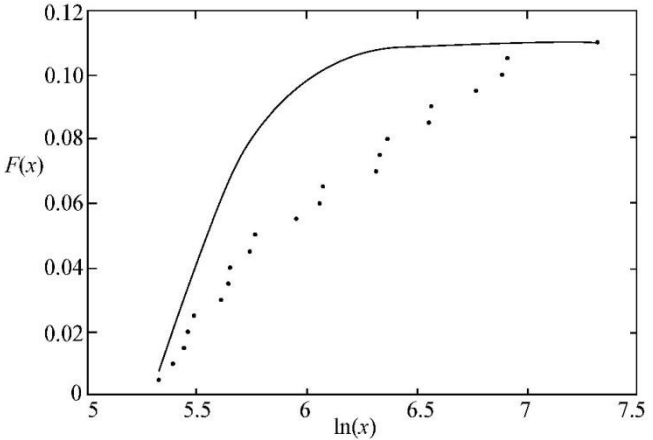


图 7—4　方法 1 的分布函数与经验分布函数

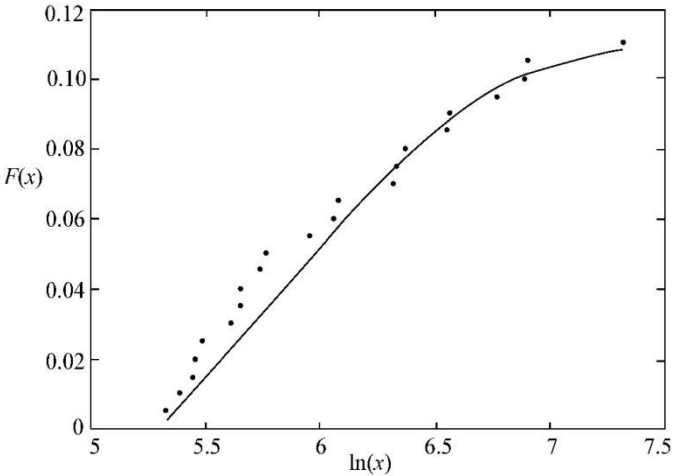


图 7—5　方法 2 的分布函数与经验分布函数

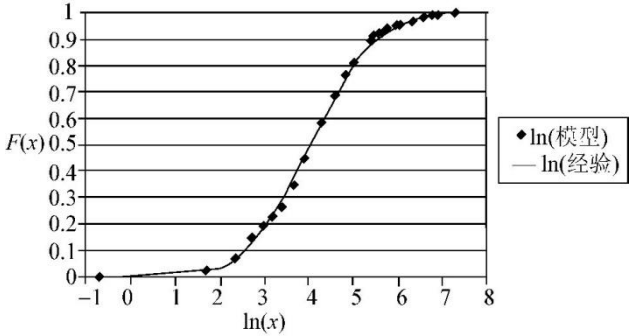


图 7—6　对数正态分布函数与经验分布函数

#引用 fitdist、MASS、survival，actuar 包
library(fitdistrplus)
library(MASS)

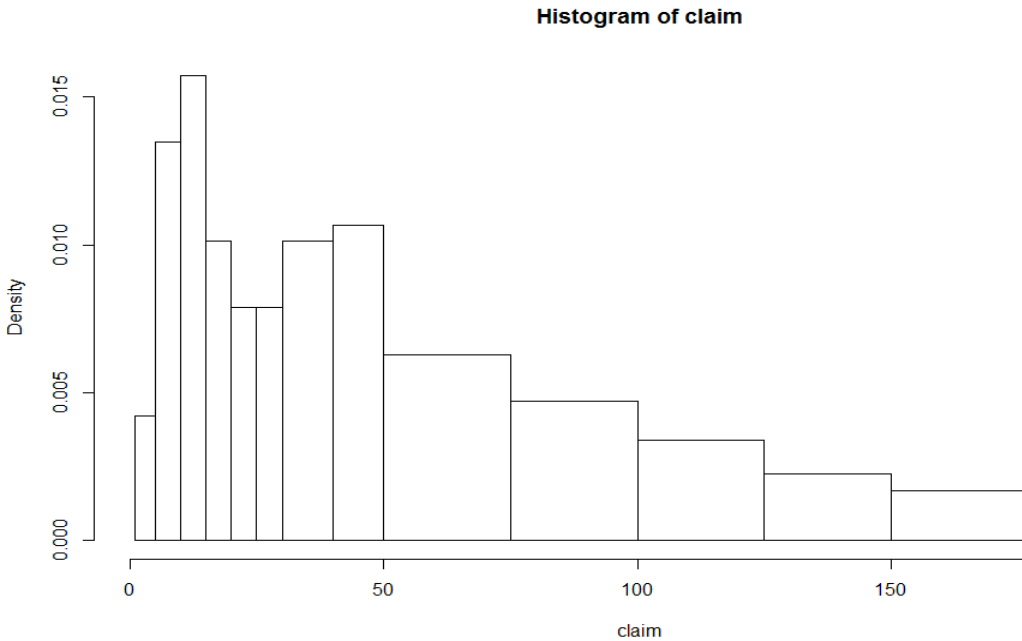
```
library(survival)
library(actuar)

#录入小于 200 的删失保单的数据
x = c(rep(1, 3), rep(6, 12), rep(11, 14), rep(16, 9), rep(21, 7), rep(26, 7), rep(31, 18), rep(41, 19), rep(51, 28), rep(76, 21), rep(101, 15), rep(126, 10), rep(151, 15))

#使用 actuar 包里的 grouped.data 函数，将向量 x 按照自定义的区间端点进行归类
claim = grouped.data(x, breaks = c(1, 5, 10, 15, 20, 25, 30, 40, 50, 75, 100, 125, 150, 200))

#绘制端点自定义的不等距直方图
hist(claim)

##
```


Histogram of claim
Density
0.000
0.005
0.010
0.015
0
50
100
150
claim

```
#进行单一模型的选择
#录入删失数据和具体数据，left 为左端点，right 为右端点
left = c(rep(1, 3), rep(6, 12), rep(11, 14), rep(16, 9), rep(21, 7), rep(26, 7), rep(31, 18), rep(41, 19), rep(51, 28), rep(76, 21), rep(101, 15), rep(126, 10), rep(151, 15), 206, 219, 230, 235, 241, 272, 283, 286, 312, 319, 385, 427, 434, 555, 562, 584, 700, 711, 869, 980, 999, 1506)
right = c(rep(5, 3), rep(10, 12), rep(15, 14), rep(20, 9), rep(25, 7), rep(30, 7), rep(40, 18), rep(50, 19), rep(75, 28), rep(100, 21), rep(125, 15), rep(150, 10), rep(200, 15), 206, 219, 230, 235, 241, 272, 283, 286, 312, 319, 385, 427, 434, 555, 562, 584, 700, 711, 869, 980, 999, 1506)

#将 left 和 right 向量合并成数据集 data1
data1 = data.frame(left, right)

#使用 mledist 函数，分别用对数正态分布、威布尔分布和伽玛分布去拟合数据集
```

data1，由三个模型的 loglikehood 可以判断对数正态拟合较好

```
mledist(data1, "lnorm")$loglik
##[1] -677.9834
mledist(data1, "weibull")$loglik
##[1] -695.8474
mledist(data1, "gamma")$loglik
##[1] -699.8303
```

#选定由对数正态分布来拟合分布，得到其最大似然的参数估计值

```
mledist(data1, "lnorm")$loglik
## meanlog          sdlog
   4.062646      1.148227
```

自主性研究作业：

对数据集 claimleve 的 claim 数据进行如下分析

（1）描述性分析，画直方图

（2）用对数正态、帕累托、指数、伽马、广义 bata 分布来拟合，与经验分布函数进行图像对比，画 pp 图和 qq 图

（3）分别进行拟合优度检验

（4）选择最优的索赔强度分布模型