基于理论贝叶斯信度模型的案例分析及其 R 实现

摘要：信度模型是保险公司用于估计未来费率的众多模型之一，其根据个体风险的索赔经验来调整费率。本文基于理论贝叶斯信度模型的假设，寻找数据并对其进行 R 语言环境下的信度计算。

关键词：信度理论，贝叶斯方法，R 语言

1. 信度保费的统计性质

从贝叶斯决策观点来看，信度保费为线性贝叶斯估计，在形式上为个体均值与聚合均值的加权平均，并且结构参数可以由样本来估计，得到经验贝叶斯估计。信度保费的有如下统计性质:

（1）估计的相合性。Schmidt（1991）提出信度估计的相合性问题，并证明了经典信度理论中贝叶斯保费与信度保费都是风险保费的相合估计，相关证明参考文献 5。

（2）估计的稳健性。保险公司的数据中偶尔会出现外来数据，如超大数据、输入错误。但有时又无法找到合适的方法将数据分离，这时一般要求给出的估计量具有稳健性。信度保费公式中包含样本均值统计量，而均值是非稳健统计量，因此信度估计也是不稳健估计。

（3）经验贝叶斯具有渐进最优性。由于单合同信度保费估计中仍然存在结构参数，在多合同信度模型中，可以用样本对结构参数进行估计，将信度估计中的结构参数用其估计量代替，这时称该估计为经验贝叶斯估计。将结构参数代入后的经验贝叶斯信度估计是否能与单合同的信度估计非常接近，这是经验贝叶斯渐进最优问题，相关证明参考文献 6。

（4）线性马尔可夫性。Witting（1988）讨论了信度保费的线性马尔科夫性，并证明了其满足线性马尔科夫性的充分必要条件。

2. 贝叶斯信度假设及计算步骤

2.1 贝叶斯信度假设

（1）假设风险参数θ服从特定分布。

（2）假设样本数据在给定风险参数情形下相互独立，即随机变量 $X_j|\theta$ 相互独立且服从同一分布。并假设$X_j|\theta$服从特定分布。

（3）信度公式为信度因子的线性表达式：$Z \times \bar{X} + (1 - Z) \times \mu$，其中$\bar{X}$ 为个体均值，$\boldsymbol{\mu}$ 为聚合均值。

2.2 贝叶斯信度计算步骤

（1）确定样本数据$X_j|\theta$的分布形式。

（2）确定样本分布的参数θ的分布形式。

（3）计算未知参数θ的后验分布。

（4）使用贝叶斯估计方法中的平方损失函数得出参数θ估计值，该估计值为信度因子。

3. 案例数据分析

3.1 拟合样本数据分布

我们使用 R 的 CASdatasets 程序包中的”usautoBI”——2002 年美国汽车损险数据集中的 LOSS 赔付金额作为样本数据。操作思路如下：

（1）加载“actuar”和“fitdistrplus”程序包

（2）使用以上程序包和 MLE 方法，以及相应函数对数据进行简单分布的拟合（Gamma, Pareto, Weibull, lognormal, exponential 等分布）

（3）根据 AIC 和 BIC 指标选取拟合情况最好的简单分布。

首先对数据进行分析，如图 1 所示，数据的 99%分位点为 67.82，为方便之后的分布拟合，选取 0 至 70 的数据进行后续操作。

```
> x1<-usautoBI$LOSS
> quantile(x1, 90:100/100)
       90%        91%        92%        93%        94%        95%        96%
   8.07650    8.78568    9.89212   10.84966   12.64006   15.50865   18.97160
       97%        98%        99%       100%
  28.03231   40.70586   67.82299 1067.69700
> quantile(x1, 0.995)
    99.5%
125.3998
```

图 1

图 2 为索赔金额在范围[0，70]的直方图，可以看出大多数索赔金额集中在[0,10]之间。

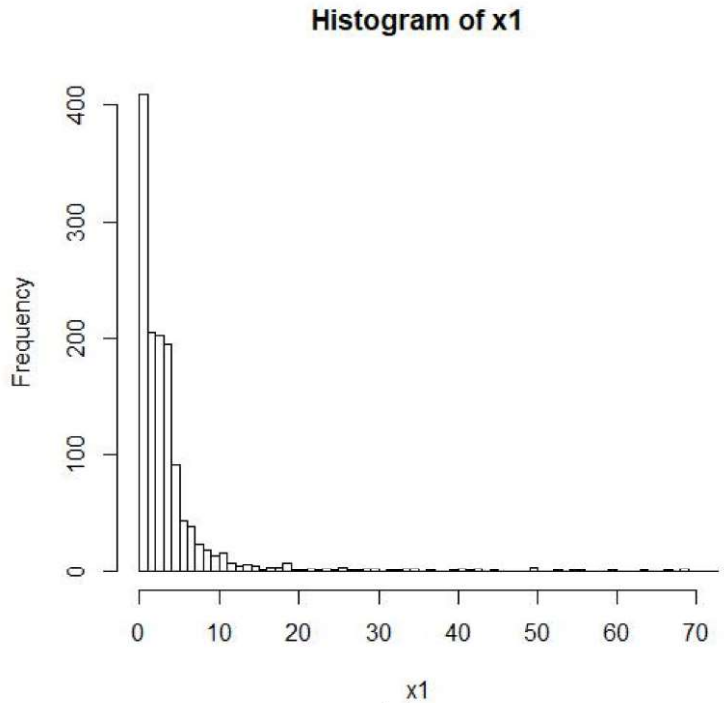


图 2

根据直方图所展示的数据分布特征，选取长尾分布并使用“fitdist”函数对数据进行拟合。分别是如下六种分布：正态分布（Normal）、伽马分布（Gamma）、对数正态分布（Lognormal）、威布尔分布（Weibull）、帕雷托分布（Pareto）、指数分布（Exponential）。

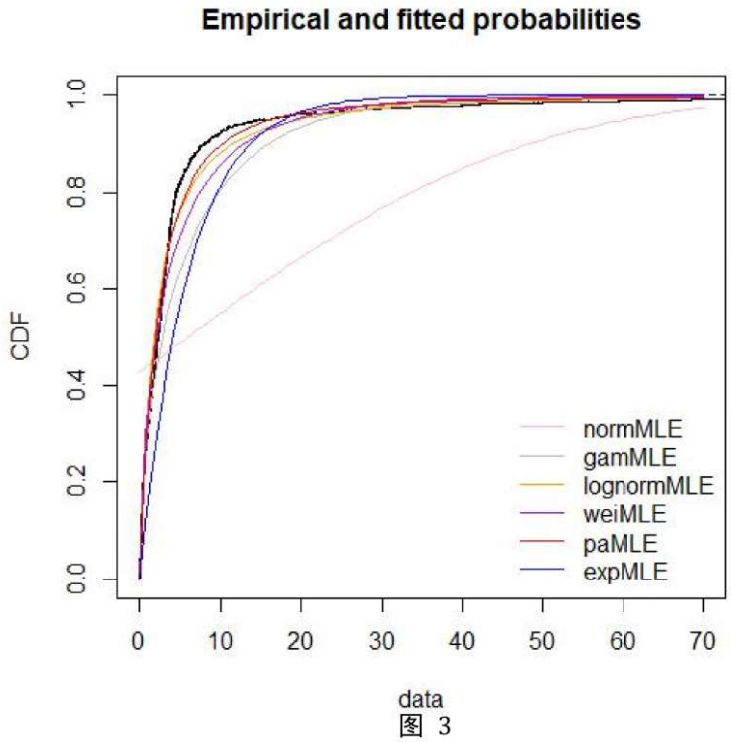


图 3

图 3 为以上六种分布函数的拟合结果和实证分布函数的对比。

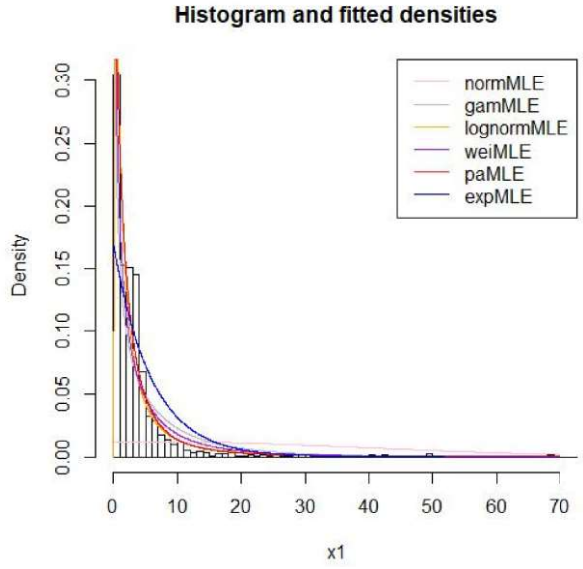


图 4

图 4 为以上六种密度函数的拟合结果和实证密度函数的对比。

根据图 3 和图 4 可以看出 Pareto 分布（如红色曲线所示）的拟合情况较其他分布好。

对各分布进行适应性检验结果如图 5 所示，得出 Pareto 分布的 AIC 和 BIC 最小，综合以上结果，本报告选取 Pareto 分布为样本数据分布。图 6 显示 Pareto 分布的估计参数，shape = 1.910784，scale = 4.357720。

```
Chi-squared statistic:  17410.66 NaN NaN NaN NaN NaN
Degree of freedom of the Chi-squared distribution:  11 11 11 11 11 12
Chi-squared p-value:  0 NaN NaN NaN NaN NaN
   the p-value may be wrong with some theoretical counts < 5
Chi-squared table:
      obscounts theo normMLE theo gamMLE theo lognormMLE theo weiMLE theo paMLE theo expMLE
<= 0          0    574.43244     0.00000         0.00000     0.00000    0.00000     0.00000
<= 1        408     15.92107   416.05617       473.28259   473.25285  437.04127   207.19133
<= 2        204     15.99310   161.85835       245.98458   190.11127  251.84664   175.15532
<= 3        202     16.05083   114.71660       142.42649   125.84033  158.58282   148.07274
<= 4        194     16.09409    89.11943        93.32355    91.91508  106.44641   125.17767
<= 5         91     16.12277    72.26895        65.86399    70.59975   74.98647   105.82265
<= 6         43     16.13679    60.09569        48.86697    55.96928   54.86011    89.46031
<= 7         38     16.13610    50.80716        37.59380    45.36296   41.37524    75.62792
<= 8         23     16.12072    43.46187        29.73102    37.37979   31.99511    63.93430
<= 9         18     16.09068    37.50615        24.03176    31.20305   25.26463    54.04876
<= 10        13     16.04607    32.58759        19.77290    26.32153   20.30757    45.69172
<= 11        16     15.98700    28.46832        16.51051    22.39791   16.57436    38.62684
<= 12         7     15.91364    24.98041        13.95942    19.20034   13.70821    32.65435
> 12          83    572.95471   208.07333       128.65242   150.44585  107.01115   178.53608

Goodness-of-fit criteria
                                normMLE   gamMLE lognormMLE   weiMLE    paMLE   expMLE
Akaike's Information Criterion 13187.43 6942.452   6345.768 6592.228 6295.842 7463.047
Bayesian Information Criterion 13197.83 6952.853   6356.169 6602.629 6306.243 7468.247
```

图 5

```
> summary(CpaMLE)
Fitting of the distribution ' pareto ' by maximum likelihood
Parameters :
      estimate Std. Error
shape 1.910784         NA
scale 4.357720         NA
Loglikelihood:  -3145.921   AIC:  6295.842   BIC:  6306.243
```

图 6

3.2 估计先验分布及似然分布

由于仅研究保险人限定赔付范围内的赔付额分布，因此假设 Scale 参数已知且固定，下求 Shape 参数的先验分布。

对于 Pareto 索赔分布，通常取 Gamma 分布作为风险参数 shape 的先验分布，原因如下：

（1）Shape 的取值具有连续性和非负性，适用于 Gamma 分布

（2）Gamma 分布中包含形状参数和尺度参数，是一个比较大的分部指数族，当取不同的参数值时可退化为指数，卡方分布等常用分布。

（3）根据资料，Gamma 分布是 Pareto 分布的共轭先验分布，使得风险参数具有较好的估计性质。

报告将假设先验分布为 Gamma 分布，使用 MLE 方法估计 shape 分布的超参数。

由于此时在 shape 已知的条件下，来自边际分布 f（x|shape）的混合样本已经确定——Pareto 分布，且先验分布形式已知——Gamma 分布，则根据 II 型最大似然法，如果先验密度函数族的形式已知，仅其中的超参数未知，则先验密度函数族可表示如下：

$$\Psi = \{\pi(a|a,b),\ (a,b) \in \Lambda\}$$

其中Λ 是超参数集，ML-Ⅱ先验要求寻找这样的超参数满足以下函数关系式：

$$m(x|\hat{a},\hat{b}) = sup_{(a,b)\in\Lambda}\{x|a,b\} = sup_{(a,b)\in\Lambda}\left\{\prod_{i=1}^{n} m(x_i|a,b)\right\}$$

$$m(x|a,b) = \int_C f(x|\alpha)\times\pi(\alpha|a,b)\,d\alpha$$

如图 7 所示，建立 Gamma 分布的参数(a,b)可能的取值矩阵，并设定循环次数为 1000。

```
set.seed(1)
times<-1000  #循环次数为times=1000，可增加次数以提高精确度
a<-rep(0,times)
b<-rep(0,times)
for (i in 1:times){
  a[i]=sample(100:10000,1)/1000
  b[i]=sample(100:10000,1)/1000
  i=i+1
}
ffmul<-cbind(a,b)
```

图 7

对每一组参数(a,b)，计算由 1340 个样本值得出的边际密度函数的似然函数。其中先验分布密度函数、似然函数、以及后验分布密度函数表达式如下：

$$\boldsymbol{Prior}: \quad p(\alpha) = \frac{b^a}{\Gamma(a)}\times\alpha^{a-1}\times e^{-b\alpha}\ ,\ \alpha>0$$

$$\boldsymbol{Likelihood}: \quad f(x_i) = \frac{\alpha\times\lambda^{\alpha}}{(\lambda+x_i)^{\alpha+1}}\ ,\ x_i>0$$

$$m(\underline{x}|\alpha) = \prod_{i=1}^{n} f(x_i) = \frac{\alpha^n\times\lambda^{n\times\alpha}}{\left(\prod_{i=1}^{n}(\lambda+x_i)\right)^{\alpha+1}}$$

图 8 所示为 R 代码计算过程。图 9 所示为求解似然函数最大值过程。

```
f<-rep(1,times)
ff<-rep(1,1340)
for (k in 1:times) {
  a<-ffmul[k,1]
  b<-ffmul[k,2]
  for (i in 1:1340){
    f1<-function(t)
      t^(a-1)*exp(-t)
    integrate(f1, 0, Inf)
    gam1<-integrate(f1, 0, Inf)$value
    f2<-function(m)
      (m*4.357720^m/(4.357720+x1[i])^m+1)*(b^a/gam1)*(m^(a-1))*exp(-b*m)
    integrate(f2,0,10)
    gam2<-integrate(f2,0,10)$value
    ff[i]<-log(gam2)
    i=i+1
  }
  sum<-cumsum(ff[1:1340])
  f[k]<- sum[1340]
  k=k+1
  }
f
```

图 8

```
> abf<-cbind(ffmul,f)
> max(abf[,3])
[1] 905.8147
> maxrow<-which(abf[,3]==max(abf[,3]))
```

图 9

计算结果如图 10 所示，形状参数 Shape 服从 Gamma(9.424 ,2.389) 分布。

```
abf[maxrow,]
     a         b         f
9.4240    2.3890 905.8147
```

图 10

3.3 计算未知参数的后验分布

根据后验正比于先验与样本分布的乘积，得出后验分布表达式:

$$p(\alpha|\underline{x}) = p(\alpha) \times m(\underline{x}|\alpha) = \frac{\alpha^n \times \lambda^{n\times\alpha}}{\left(\prod_{i=1}^{n}(\lambda + x_i)\right)^{\alpha+1}} \times \frac{b^a}{\Gamma(a)} \times \alpha^{a-1} \times e^{-b\alpha}$$

化简易知后验分布符合 Gamma 分布：

$$Shape = n + a - 1$$

$$Scale = b - n \times \log\lambda + \sum_{i=1}^{n} \log(\lambda + x_i)$$

代入计算数值 a=9.4240，b=2.3890，λ=4.357720 后得出后验分布为

$$Gamma(1348.424\ ,\ 703.6357\ )$$

3.4 计算预期赔付额

根据 Gamma 分布的期望计算公式，计算得出下一年的预计赔付额为该期望估计值。

$$\mathrm{E(X)} = \frac{shape\ parameter}{scale\ parameter}$$

计算结果如图 11 所示，为 1.916367（thousand dollars）。

```
> EX<-postshape/postscale
> EX
       a
1.916367
```

图 11

可以看到计算所得的期望与第一步拟合样本分布所得的参数值非常接近，进一步说明了贝叶斯方法估计的合理性。

3.5 模型局限性

由于 R 程序本身无法输出分布函数等的具体表达式，因此本报告所采取的模型存在以下局限：

（1）仅采用抽样方法模拟出了 1000 个超参数 a，b 的可能取值的组合，该步骤并未覆盖超参数的所有可能取值，限制了 a，b 的实际范围，从而产生误差；

（2）根据 ML-Ⅱ方法，积分求解 Likelihood 时，α的取值范围为（0，10）而非（0，+∞），该限制是为了避免积分项出现无穷取值从而干扰计算，但该取值范围也可能存在一定的偏差。经过调整选择α>1 的值（根据 Pareto distribution 的性质）仍存在积分出现无穷值的情况。

以上步骤皆有可能使计算结果偏离实际先验参数α的分布。

4. 贝叶斯信度保费的不足

（1）经典的信度理论假设风险之间是相互独立的（且经验信度理论认为在时间分量上的索赔是条件独立的），然而实际中存在许多风险相依的情形，比如夫妻寿命相关。且同一保单在不同时间的索赔可能具有相依性。

（2）信度保费中同时含有先验信息和样本信息。然而 Young.V.R 认为样本分布中的参数容易由样本进行估计，但先验分布本身或先验分布的超参数很难估计，她提出用半参数的方法对先验分布进行估计。

（3）经典信度理论建立的是净保费的估计模型，是因为它使用了平方损失函数。然而在实际运用中，净保费原理不能满足正的安全负荷性，即保险人收取的保费至少要大于风险的期望值，保险公司将根据自身对风险态度选取合适的保费原理。(Geber 1980)

（4）在实际计算中，根据贝叶斯方法求得的参数结果不一定能够直接转化成信度公式的的形式（如样本平均数在分母上的情况等）。很难直接从结果中分离出信度因子。

参考文献

[1]温利民,张美,程子红,章溢.帕累托索赔分布中风险参数的经验贝叶斯估计[J].应用概率统计,2015,31(03):225-237.

[2]温利民. 风险保费的信度估计及其统计推断[D].华东师范大学,2010.

[3]王宏洲.引信的可靠性评定——贝叶斯方法和经验贝叶斯方法[J].现代引信,1989(02):11-19+10.

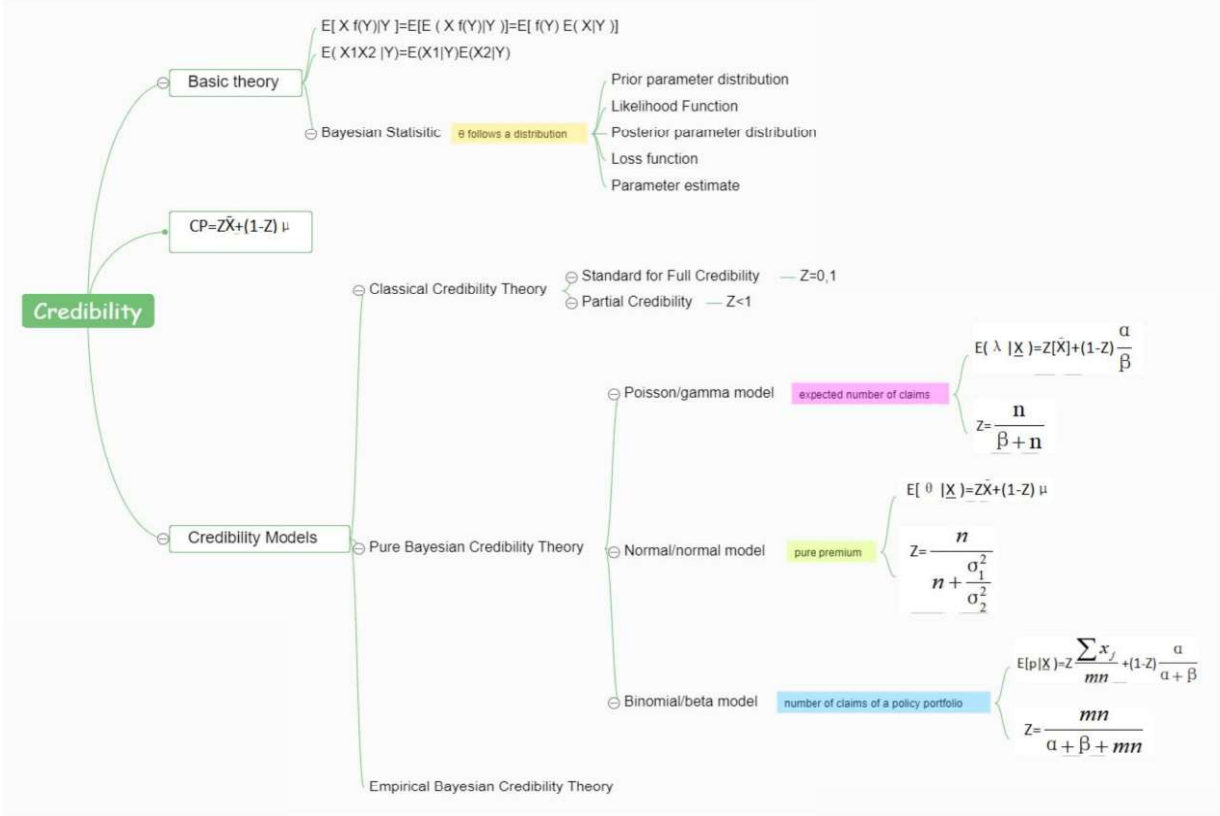
[4]黄长全.2017.贝叶斯统计及其 R 实现[M].清华大学出版社

[5]Pan M , Wang R , Wu X . On the consistency of credibility premiums regarding Esscher principle[J]. Insurance Mathematics and Economics, 2008, 42(1):119-126.

[6] 王立春, 韦来生. 渐近最优的经验贝叶斯决策[J]. 应用数学, 2006, 19(2):356-362.

附录

1. 思维导图



2. R code

#第一步：选取数据并进行拟合

#为方便计算，所选取的数据应尽量服从简单分布

#==========

```
library("CASdatasets", lib.loc="~/R/win-library/3.5")
library("fitdistrplus", lib.loc="~/R/win-library/3.5")
library("actuar", lib.loc="~/R/win-library/3.5")

#1.读取数据

data(usautoBI)
View(usautoBI)
x1<-usautoBI$LOSS
quantile(x1, 90:100/100)
quantile(x1, 0.995)
hist(x1,xlim = c(0,70), breaks = 1000)
```

#2.拟合分布

##注：mle 的良好性质：无偏、渐进正态、依概率收敛为真值、最有效(无偏估计量中方差最小)

#------以下选用 mle 进行估计-----#

```
##normal
CnormMLE <- fitdist(x1,"norm",method = "mle")
##gamma
CgamMLE <- fitdist(x1,"gamma",method = "mle")
##lognormal
ClognormMLE <- fitdist(x1,"lnorm",method="mle")
##Weibull
CweiMLE <- fitdist(x1, "weibull", method = "mle")
##pareto
CpaMLE <- fitdist(x1, "pareto", method = "mle", start = list(shape=2,scale=2), lower = 0)
##exponential
CexpMLE <- fitdist(x1, "exp", method = "mle")
```

#------对比以上 6 种基本分布------#

```
#对比分布函数
txt <- c("normMLE","gamMLE", "lognormMLE", "weiMLE", "paMLE", "expMLE")
cdfcomp(list(CnormMLE,CgamMLE, ClognormMLE, CweiMLE, CpaMLE, CexpMLE),
            datapch = ".", legendtext = txt,
            xlim = c(0, 70), ylim = c(0, 1), fitlty = 1,
            fitcol = c("pink","grey", "orange", "purple", "red", "blue"),
            main = "Empirical and fitted probabilities")

#对比密度函数
txt <- c("normMLE","gamMLE", "lognormMLE", "weiMLE", "paMLE", "expMLE")
hist(x1, breaks = 1000, prob = T, xlim = c(0, 70),
        main = "Histogram and fitted densities")
x <- seq(0, 70, 0.01)
y1 <- dnorm(x,mean = CnormMLE$estimate[1],sd = CnormMLE$estimate[2])
y2 <- dgamma(x, shape = CgamMLE$estimate[1], rate = CgamMLE$estimate[2])
y3 <- dlnorm(x, meanlog = ClognormMLE$estimate[1], sdlog =    ClognormMLE$estimate[2])
y4 <- dweibull(x, shape = CweiMLE$estimate[1], scale = CweiMLE$estimate[2])
y5 <- dpareto(x, shape = CpaMLE$estimate[1], scale = CpaMLE$estimate[2])
y6 <- dexp(x, rate = CexpMLE$estimate[1])
lines(x, y1, type = 'l', col = "pink", lwd = 1.5)
```

```
lines(x, y2, type = 'l', col = "grey", lwd = 1.5)
lines(x, y3, type = 'l', col = "orange", lwd = 1.5)
lines(x, y4, type = 'l', col = "purple", lwd = 1.5)
lines(x, y5, type = 'l', col = "red", lwd = 1.5)
lines(x, y6, type = 'l', col = "blue", lwd = 1.5)
legend('topright', leg = txt, lty = rep(1, 5), lwd = rep(1.5, 5),
          col = c("pink","grey", "orange", "purple", "red", "blue"))


#3.适应性检验

gofstat(list(CnormMLE,CgamMLE, ClognormMLE, CweiMLE, CpaMLE, CexpMLE),
            discrete = TRUE, chisqbreaks = c(0:12),
            fitnames=c("normMLE","gamMLE", "lognormMLE", "weiMLE", "paMLE", "expMLE"))

##pareto 分布的 AIC, BIC 最小，因此选取 pareto 分布为样本分布

summary(CpaMLE)

#shape = 1.910784，scale = 4.357720

##根据实际情况，shape 为形状参数，scale 为尺度参数
 #由于仅研究保险人限定赔付范围内的赔付额分布，因此 scale 已知且固定，下求 shape 参数的先验分布


#=============================================================================

#第二步：进行 shape 参数先验分布的估计

#=============================================================================

#pareto 模型在非寿险精算领域常用于描述再保险或具有免赔额保险的索赔分布，进一步说明拟合分布具有现实意义
#对于 pareto 索赔分布，通常取 Gamma 分布作为风险参数 shape 的先验分布，原因如下：
#（1）shape 的取值具有连续性和非负性，适用于 Gamma 分布；
#（2）Gamma 分布中包含形状参数和尺度参数，是一个比较大的分部指数族，当取不同的参数值时可退化为指数，卡方分布等常用分布
#（3）根据资料，Gamma 分布是 pareto 分布的共轭先验分布，使得风险参数具有较好的估计性质

#已知先验分布为 Gamma 分布，使用 MLE 估计 shape 分布的超参数
##原理：由于此时在 shape 条件下，来自边际分布 f（x|shape）的混合样本已经确定，且先验分布形式已知，则：
```

#根据Ⅱ型最大似然法：可确定边际分布的密度函数公式（见 word）

```
#建立 gamma 分布的参数 a，b 的可能取值矩阵
times<-1000   #循环次数为 times=1000，可增加次数以提高精确度
a<-rep(0,times)
b<-rep(0,times)
set.seed(111)
for (i in 1:times){
   a[i]=sample(100:10000,1)/1000
   b[i]=sample(100:10000,1)/1000
   i=i+1
}
ffmul<-cbind(a,b)

#对每一组 a，b 的取值，求解 1340 个样本值得出的边际密度函数的 log-likelihood
f<-rep(1,times)
ff<-rep(1,1340)
for (k in 1:times) {
   a<-ffmul[k,1]
   b<-ffmul[k,2]
   for (i in 1:1340){
      f1<-function(t)
         t^(a-1)*exp(-t)
      integrate(f1, 0, Inf)
      gam1<-integrate(f1, 0, Inf)$value
      f2<-function(m)
         (m*4.357720^m/(4.357720+x1[i])^m+1)*(b^a/gam1)*(m^(a-1))*exp(-b*m)
      integrate(f2,0,10)
      gam2<-integrate(f2,0,10)$value
      ff[i]<-log(gam2)
      i=i+1
   }
   sum<-cumsum(ff[1:1340])
   f[k]<- sum[1340]
   k=k+1
   }
f

#对应出每组参数取值的 log-likelihood 取值并求出最大值
abf<-cbind(ffmul,f)
max(abf[,3])
maxrow<-which(abf[,3]==max(abf[,3]))
abf[maxrow,]
```

##a=9.4240，b=2.3890，此时 f 取最大值 905.8147
##根据该结果，shape 参数符合分布为 Gamma（shape=9.4240，scale=2.3890）

#==========

#第三步：根据先验分布和样本数据，计算后验分布

#==========

#根据后验正比于先验与样本分布的乘积，得出后验分布表达式（见 word）
##由于先验分布为 Gamma 分布，后验分布仍为 Gamma 分布，参数计算如下：

```
n<-1340   #样本容量
a<-abf[maxrow,1]   #先验分布超参数 1
b<-abf[maxrow,2]   #先验分布超参数 2
y<-4.357720   #混合样本分布 scale 参数

#post shape 计算
postshape<-n+a-1
postshape #1348.424

xx<-rep(0,1340)
for (i in 1:1340) {
   xx[i]<-log(y+x1[i])
}

#post scale 计算
postscale<-b-n*log(y)+sum(xx[1:1340])
postscale   #703.6357
```

##最终后验分布的形式为 Gamma(1348.424，703.6357)

#==========

#第四步：求解后验分布的期望，即为下一年的未来预计赔付额

#==========

#根据 Gamma 分布的期望公式计算如下：

```
EX<-postshape/postscale
EX
```