Анализ страховых выплат в Дании: оценка распределения Парето

Аннотация

В данной работе проводится анализ страховых выплат по пожарам в Дании на основе датасета danish. Цель исследования — оценка распределения убытков и параметров распределения Парето, которые описывают данные. Используются методы визуализации (гистограммы, CCDF), оценка параметров методом максимального правдоподобия и бутстрэп-анализ. Полученные результаты показывают, что параметр формы $\theta \approx 1.27$, что подтверждает наличие степенного распределения. Построены доверительные интервалы для параметров и математического ожидания. Результаты важны для оценки рисков и расчета тарифов в страховании. Предлагаются направления для дальнейших исследований, включая смеси распределений и байесовские методы.

Введение

В страховой и финансовой практике одной из ключевых задач является оценка рисков крупных убытков. Такие убытки, как правило, описываются распределениями с тяжелыми хвостами. Оценка распределения страховых выплат позволяет вычислить математическое ожидание ущерба, которое является важным параметром для страховых компаний при расчете тарифов, резервов, стратегий управления рисками, а также служит основой для прогнозирования средних выплат и планирования финансовой устойчивости компании.

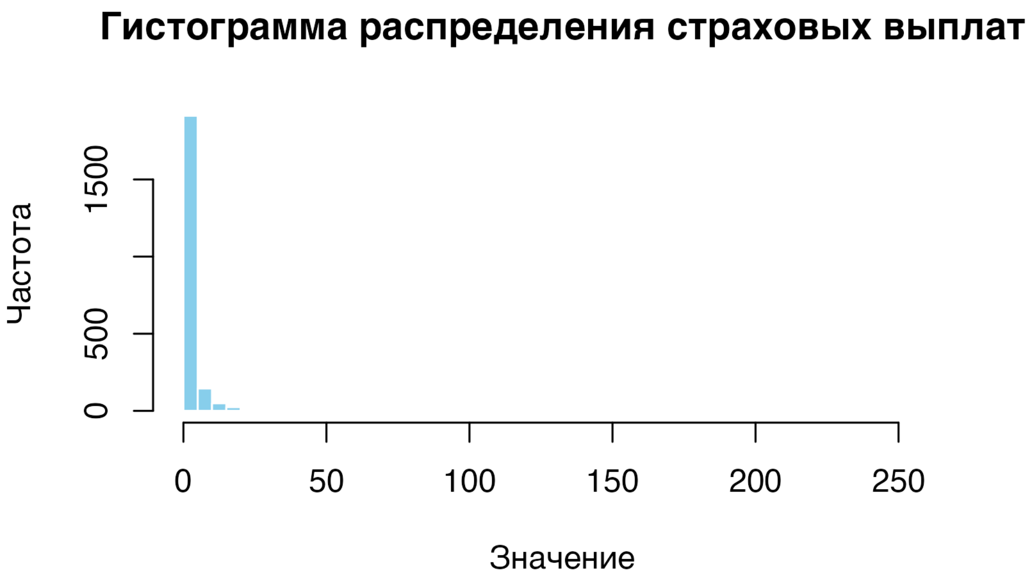
Основная часть

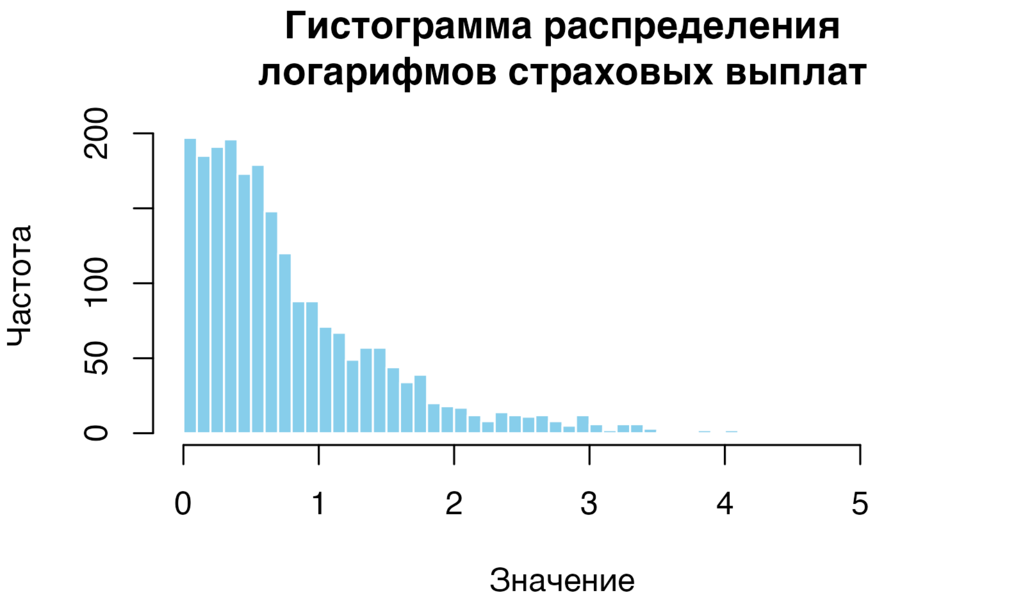
Для анализа страховых выплат используется датасет danish, доступный в пакете evir. Сначала загружаем данные и сохраняем их в локальную переменную.

```
library(evir)

data(danish, package="evir")
x <- danish
```

Построим распределение страховых выплат, а также гистограмму распределения их логарифмов. Этот шаг необходим для предварительного анализа и выбора подходящего класса распределений, описывающих данные.





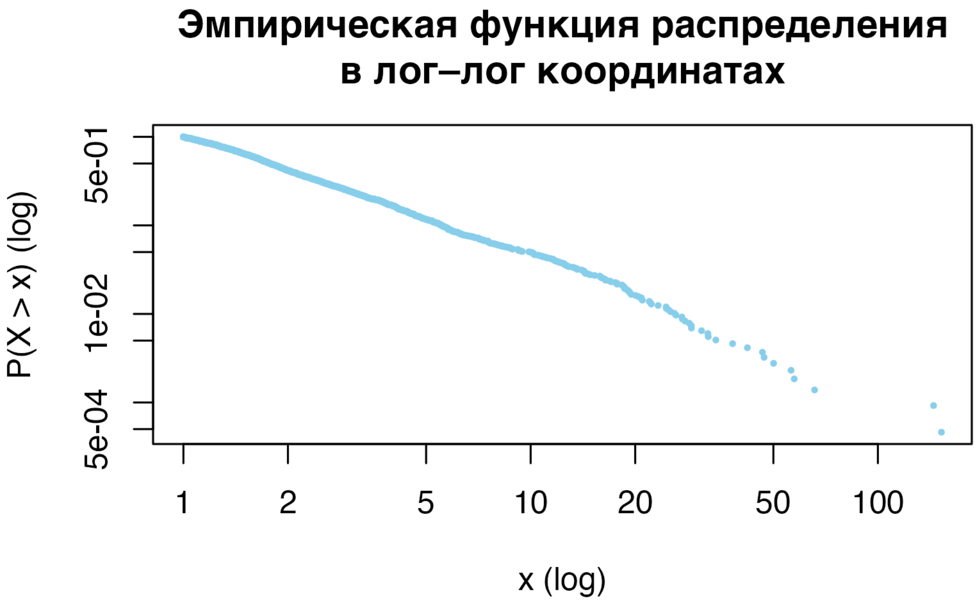
На первой гистограмме видно наличие тяжелого длинного хвоста в распределении страховых выплат. После логарифмирования данные приобретают вид, близкий к экспоненциальному распределению. Это указывает на то, что исходные данные подчиняются степенному распределению.

Чтобы проверить гипотезу о степенном распределении данных, изобразим эмпирическую функцию распределения превышений (CCDF) , $\overline{F}(x) = 1 - F(x) = P(X > x)$ показывает вероятность того, что случайная величина превысит значение x, на логарифмических координатных осях. Численно для выборки, упорядоченной по возрастанию ($x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$), эмпирическая оценка CCDF задается: $\widehat{\overline{F}}(x_{(i)}) = 1 - \frac{i}{n}, \quad i = 1, \ldots, n-1$. Таким образом, для каждой точки $x_{(i)}$ вероятность превышения оценивается долей наблюдений, больших данного значения.

```r
xt <- sort(x)
n <- length(x)

#функция CCDF
ccdf <- 1 - (1:n)/n
xt <- xt[-n]
ccdf <- ccdf[-n]

plot(xt, ccdf,
     log = "xy",
     pch = 20,
     cex = 0.5,
     main = 'Эмпирическая функция распределения\nв лог-лог координатах',
     col = "skyblue",
     xlab = "x (log)", ylab = "P(X > x) (log)")
```



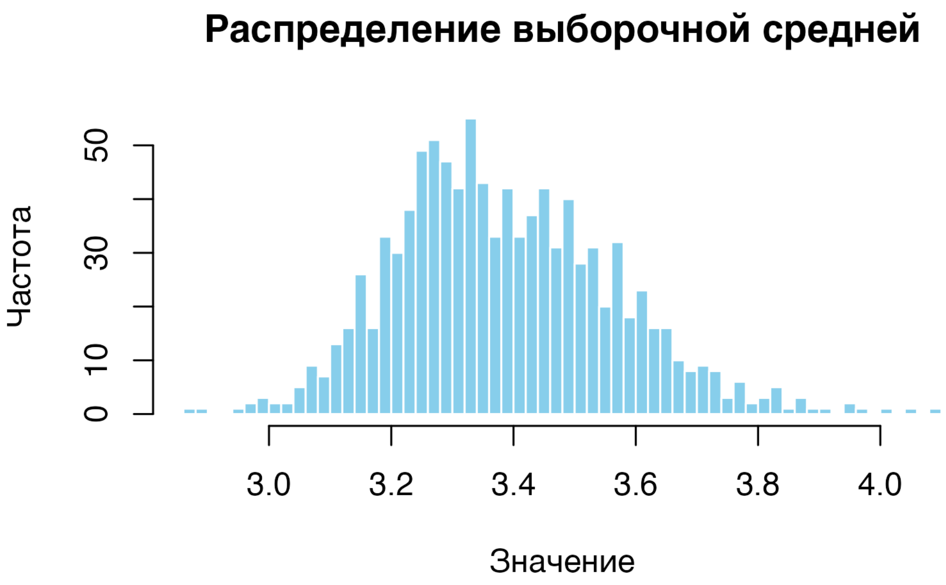
Данное распределение в логарифмических осях дало линейную зависимость, что следовательно, можно предположить, что данное распределение — распределение Парето. Теперь мы знаем семейство нашего расцделения. Функция распредления $p(x) = \frac{\theta \cdot u^{\theta}}{x^{\theta+1}} \cdot I\{x > u\}$, $\theta > 0, u = x_{xmin}$ - фиксированный сдвиг.

Следующий этап - оценка математического ожидания ущерба. Для распределений с тяжелыми хвостами, такими как Парето, важно учитывать, что выборочная средняя может сильно колебаться и быть чувствительной к редким крупным выплатам. Поэтому сначала построим распределение среднего по бутстрэп-выборкам, чтобы визуализировать его вариативность и оценить устойчивость среднего значения.

Для этого сгенерируем 1000 независимых бутстрэп-выборок объемом равным исходной выборке. Для каждой бутстрэп-выборки вычислим выборочное среднее. В результате получим эмпирическое распределение 1000 значений средних, которое позволяет оценить поведение статистики $\overline{X}$.

```r
set.seed(42)

B <- 1000
mean_boot <-  replicate(
  B, {
    xb <- sample(x, n, replace = TRUE)
    mean(xb)}
  )
hist_plot(mean_boot, breaks = 60, name = 'Распределение выборочной средней')
```



Гистограмма распределения выборочного среднего имеет куполообразную форму, однако она не симметрична: длинный правый хвост. Это свидетельствует о том, что среднее значение выборки подвержено влиянию крупных выбросов и не может считаться нормальным. Следовательно, применение классического z-теста для оценки математического ожидания не является корректным для данного набора данных.

Вместо прямого использования выборочного среднего необходимо перейти к оценке математического ожидания, используя формулу математического ожидания $\mathbb{E}[X] = \frac{\theta\, x_{min}}{\theta - 1} = \frac{\theta\, u}{\theta - 1}$ для $\theta > 1$. Чтобы получить аналитическую оценку ущерба, найдем $\widehat{\theta}_{MLE}$.

Функция правдоподобия для распределения Парето с функцией плотности $p(x \mid \theta) = \frac{\theta\, u}{x^{\theta+1}}$:

$$L(\theta \mid X) = \prod_{i=1}^{n} \frac{\theta\, u^{\theta}}{x_i^{\theta+1}} = \theta^n \frac{u^{n\theta}}{\prod_{i=1}^{n}(x_i)^{\theta+1}} = \frac{\theta^n\, u^{n\theta}}{\left[\prod_{i=1}^{n} x_i\right]^{\theta+1}}$$

Логарифм функции правдободобия:

$$\log L(\theta \mid X) = \log\left(\frac{\theta^n\, u^{n\theta}}{\left[\prod_{i=1}^n x_i\right]^{\theta+1}}\right) =$$

$$= \log\theta^n + \log u^{n\theta} - \log\left(\prod_{i=1}^n x_i\right)^{\theta+1} =$$

$$= n\,\log\theta + n\theta\,\log u - (\theta+1)\sum_{i=1}^n \log x_i$$

Производная логарифма функции правдободобия по θ:

$$\frac{\partial}{\partial\theta}\log L(\theta \mid X) = \frac{n}{\theta} + n\log u - \sum_{i=1}^n \log x_i = \frac{n}{\theta} - \sum_{i=1}^n (\log x_i - \log u)$$

Максимизируем функцию правдоподобия, приравнивая производную к нулю:

$$\frac{n}{\theta} - \sum_{i=1}^n (\log x_i - \log u) = 0 \quad\Rightarrow\quad \hat\theta_{\text{MLE}} = \frac{n}{\sum_{i=1}^n (\log x_i - \log u)} = \frac{n}{\sum_{i=1}^n \log\frac{x_i}{u}}$$

$$\frac{\hat\theta_{\text{MLE}} - \theta}{\sigma} \xrightarrow{d} \mathcal{N}(0,1)$$

Дисперсия ММП оценки равна информации Фишера: $Var(\hat\theta_{MLE}) = \frac{1}{I_n(\hat\theta_{MLE})}$

Информация Фишера: $I_n(\theta) = nI_1(\theta)$

$$I_1(\theta) = \mathbb{E}[(\frac{\partial}{\partial\theta}\log L(\theta \mid X))^2] = \mathbb{E}[(\frac{1}{\theta} - \log\left(\frac{x_1}{u}\right)^2] = \text{Var}(\log\frac{x}{u})$$

$$I_n(\theta) = \mathbb{E}[(\frac{\partial}{\partial\theta}\log L(\theta \mid X))^2] \tag{1}$$

$$I_n(\theta) = -\mathbb{E}[(\frac{\partial^2}{\partial\theta^2}\log L(\theta \mid X))] \tag{2}$$

$$\text{Var}(\log x) = \mathbb{E}[\log^2 x] - (\mathbb{E}[\log x])^2 = \frac{2}{\theta^2} - \left(\frac{1}{\theta}\right)^2 = \boxed{\frac{1}{\theta^2}}$$

$$\mathbb{E}[\log x] = \int_1^\infty \log x \cdot p(x \mid \theta)\, dx = \int_1^\infty \frac{\log \frac{x}{u} \cdot \theta}{x^{\theta+1}}\, dx = \theta \int_1^\infty \frac{\log \frac{x}{u}}{x^{\theta+1}}\, dx =$$

$$= \left| \begin{matrix} u = \log \frac{x}{u}, \quad du = \frac{u}{x} \cdot \frac{1}{u} dx = \frac{1}{x} dx \\ dv = \frac{1}{x^{\theta+1}} dx, \quad v = \int \frac{1}{x^{\theta+1}} dx = -\frac{1}{\theta x^\theta} \end{matrix} \right| =$$

$$= uv - \int v\, du = -\theta \frac{\log \frac{x}{u}}{\theta x^\theta} \Big|_1^\infty + \theta \int_1^\infty \frac{1}{\theta x^{\theta+1}}\, dx =$$

$$= \int_1^\infty \frac{1}{x^{\theta+1}}\, dx = -\frac{1}{\theta x^\theta}\Big|_1^\infty = \boxed{\frac{1}{\theta}}$$

$$\mathbb{E}[\log^2 x] = \int_1^\infty \log^2(x) p(x \mid \theta)\, dx = \theta \int_1^\infty \frac{\log^2 \frac{x}{u}}{x^{\theta+1}}\, dx$$

$$= \left| \begin{matrix} u = \log^2 \frac{x}{u}, \quad du = \frac{2 \log \frac{x}{u} \cdot u}{x} \frac{1}{u} dx = \frac{2 \log \frac{x}{u}}{x}\, dx \\ dv = \frac{\theta}{x^{\theta+1}} dx, \quad v = \int \frac{\theta}{x^{\theta+1}} dx = -\frac{\theta}{\theta x^\theta} = -\frac{1}{x^\theta} \end{matrix} \right|$$

$$= uv - \int v\, du = -\frac{\log^2 \frac{x}{u}}{x^\theta} \Big|_1^\infty + \int_1^\infty \frac{2 \log \frac{x}{u}}{x^{\theta+1}}\, dx$$

$$= 2 \int_1^\infty \frac{\log \frac{x}{u}}{x^{\theta+1}}\, dx = \boxed{\frac{2}{\theta^2}}$$

$$\boxed{I_n(\theta) = n \cdot I_1(\theta) = n \cdot \mathrm{Var}(\log x) = n \cdot \frac{1}{\theta^2} = \frac{n}{\theta^2}}$$

$$\theta \sim \mathcal{N}\left(\hat{\theta}_{\text{MLE}}, \frac{1}{I(\hat{\theta}_{\text{MLE}})}\right)$$

Отсюда следует, что:

$$\mathbb{P}\left(-q_{1-\alpha/2} < (\hat{\theta}_{MLE} - \theta) \cdot \sqrt{n I_1(\theta)} < q_{1-\alpha/2}\right) \approx 1 - \alpha$$

$$\mathbb{P}\left(\hat{\theta}_{\text{MLE}} - q_{1-\alpha/2}\frac{1}{\sqrt{nI_1(\theta)}} < \theta < \hat{\theta}_{\text{MLE}} + q_{1-\alpha/2}\frac{1}{\sqrt{nI_1(\theta)}}\right) \approx 1-\alpha$$

Возьмем уровень значимости $\alpha = 0.05$ (соответствующий 95% уровню доверия). Тогда $\theta \in \hat{\theta}_{\text{MLE}} \pm z_{1-0.025} \cdot \frac{1}{\sqrt{nI_1(\theta)}}$

```
u <- min(x)
theta_mle <- n / sum(log(x / u))
cat('Точечная оценка theta:', theta_mle)
```

```
Точечная оценка theta: 1.270729
```

```
theta_I <- n / theta_mle^2
alpha <- 0.05
z_left <- qnorm(alpha / 2)
z_right <- qnorm(1 - alpha / 2)

g_theta_mle <- 1 + 1 / (theta_mle - 1)
g_prob_theta <- -1 / (theta_mle - 1)^2
theta_left <- theta_mle+ z_left  / sqrt(theta_I)
theta_right <- theta_mle + z_right / sqrt(theta_I)
cat('Доверительный интервал theta: [', theta_left,':', theta_right, ']')
```

```
Доверительный интервал theta: [ 1.217226 : 1.324231 ]
```

Математическое ожидание будем оценивать как $\mathbb{E}[X] = \frac{\hat{\theta}u}{\hat{\theta}-1} = u + \frac{u}{\hat{\theta}-1}$. Поскольку в данном случае математическое ожидание является монотонным преобразованием оценки $\hat{\theta}$, т.е. $\mathbb{E}[X] = g(\hat{\theta})$. Тогда $\frac{g(\hat{\theta})-g(\theta)}{\sigma\cdot|g'(\hat{\theta})|} \xrightarrow{d} \mathcal{N}(0,1)$, где $Var(\hat{\theta}) = \sigma^2 = \frac{1}{I(\hat{\theta})}$, а $g'(\hat{\theta}) = -\frac{u}{(\hat{\theta}-1)^2}$

$$g(\theta) \xrightarrow{d} \mathcal{N}\left(g(\hat{\theta}),\ \frac{g'(\hat{\theta})}{I(\hat{\theta})}\right)$$

$$g(\theta) \in g(\hat{\theta}_{\text{MLE}}) \pm z_{1-\frac{\alpha}{2}} \cdot \frac{g'(\hat{\theta})}{\sqrt{nI_1(\theta)}}$$

$$\mathbb{E}[X] \in g(\hat{\theta}_{\text{MLE}}) \pm z_{1-\frac{\alpha}{2}} \cdot \frac{g'(\hat{\theta})}{\sqrt{nI_1(\theta)}}$$

```
g_theta_mle <- u + u / (theta_mle - 1)
g_prob_theta <- -u / (theta_mle - 1)^2
expect_left <- g_theta_mle + z_left * abs(g_prob_theta) / sqrt(theta_I)
expect_right <- g_theta_mle + z_right * abs(g_prob_theta) / sqrt(theta_I)
cat('Точечная оценка математического ожидания:', g_theta_mle, '\n')
```

```
Точечная оценка математического ожидания: 4.693736
```

```r
cat('Доверительный интервал мат. ожидание: [', expect_left,':', expect_right, ']')
```

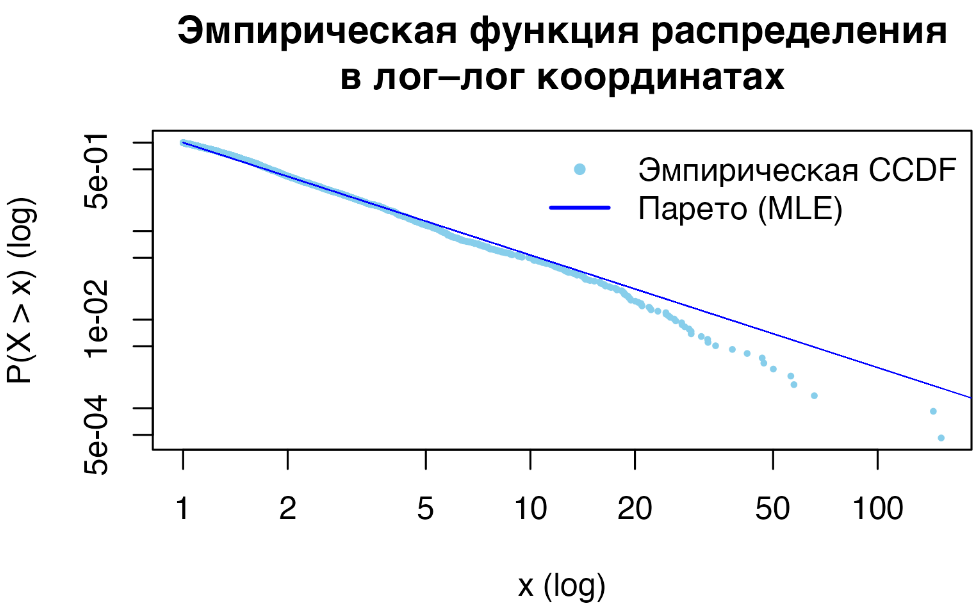
```
Доверительный интервал мат. ожидание: [ 3.963769 : 5.423703 ]
```

Полученное значение параметра $\hat{\theta}$ позволяет построить теоретическую функцию распределения превышений (CCDF) для модели Парето.

```r
plot(xt, ccdf,
     log = "xy",
     pch = 20,
     cex = 0.5,
     main = 'Эмпирическая функция распределения\nв лог-лог координатах',
     col = "skyblue",
     xlab = "x (log)", ylab = "P(X > x) (log)")


u <- min(x)
ccdf_pareto <- (u / x)^theta_mle

lines(x, ccdf_pareto, col = "blue", lwd = 0.5) # теоретичская ccdf
legend("topright",
       legend = c("Эмпирическая CCDF", "Парето (MLE)"),
       col = c("skyblue", "blue"),
       pch = c(20, NA),
       lty = c(NA, 1),
       lwd = c(NA, 2),
       bty = "n")
```



Заключение

Анализ эмпирической функции распределения превышений (CCDF) в лог–лог координатах показывает, что данные по страховым выплатам демонстрируют степенное поведение в диапазоне значений $x \in [1, 20]$, где эмпирическая кривая хорошо аппроксимируется прямой, соответствующей распределению Парето с $\widehat{\theta}_{MLE} \approx 1.27$. Это свидетельствует о том, что модель Парето с одним параметром формы θ адекватно описывает основную часть распределения. Однако при значениях $x > 20$ наблюдается отклонение эмпирической кривой вниз относительно теоретической линии, что указывает на факт, что экстремальные убытки подчиняются другому закону.

В качестве перспективного направления дальнейшего исследования предлагается рассмотреть модели смесей распределений.Кроме того, перспективным представляется применение байесовского подхода к оценке параметров и математического ожидания ущерба.