Моделирование, оценка параметров и проверка гипотез согласия эмпирического распределения с теоретическим

Вейбер Е.Н. 23.М08-мм

15-04-2024

Введение

Это отчет о моделировании выборки объемом n = 150, имеющей равномерное распределение U(1, 3)

Загрузка данных

Данные были промоделированы с помощью следующего кода:

```
# Загрузка необходимых библиотек
library(tidyverse)
library(ggplot2)
library(readr)
library(corrplot)
library(tidyr)
library(GGally)
library(plotly)
library(stats)
library(reshape2)


#Моделирование равномерного распределения U(1, 3)
set.seed(123)  # для воспроизводимости
n <- 150
true_min <- 1
true_max <- 3
sample <- runif(n, min = true_min, max = true_max)
```

Оценка параметров распределения по методу моментов и по методу максимального правдоподобия

```
# Метод моментов
mean_sample <- mean(sample)
var_sample <- var(sample)
estimated_min <- 2 * mean_sample - sqrt(12 * var_sample)
estimated_max <- 2 * mean_sample + sqrt(12 * var_sample)

# Метод максимального правдоподобия
mle_min <- min(sample)
mle_max <- max(sample)

# Вывод результатов
cat("Метод моментов:\n")
## Метод моментов:
cat(sprintf("Оценка min: %f\n", estimated_min))
## Оценка min: 2.027202
cat(sprintf("Оценка max: %f\n\n", estimated_max))
## Оценка max: 6.006152
cat("Метод максимального правдоподобия:\n")
## Метод максимального правдоподобия:
cat(sprintf("Оценка min: %f\n", mle_min))
## Оценка min: 1.001250
cat(sprintf("Оценка max: %f\n", mle_max))
## Оценка max: 2.988540
```

Интерпретация

Метод моментов:

- Оценка минимального значения (min) составила приблизительно 2.027, что выше истинного минимума равномерного распределения, заданного как 1.
- Оценка максимального значения (max) составила приблизительно 6.006, что значительно выше истинного максимума, заданного как 3. Это указывает на возможную неточность или наличие выбросов в данных, которые смещают оценки метода моментов.

Метод максимального правдоподобия (MLE):

- Оценка минимального значения (min) составила приблизительно 1.001, что очень близко к истинному минимальному значению распределения.
- Оценка максимального значения (max) составила приблизительно 2.989, что также близко к истинному максимальному значению распределения.

Выводы:

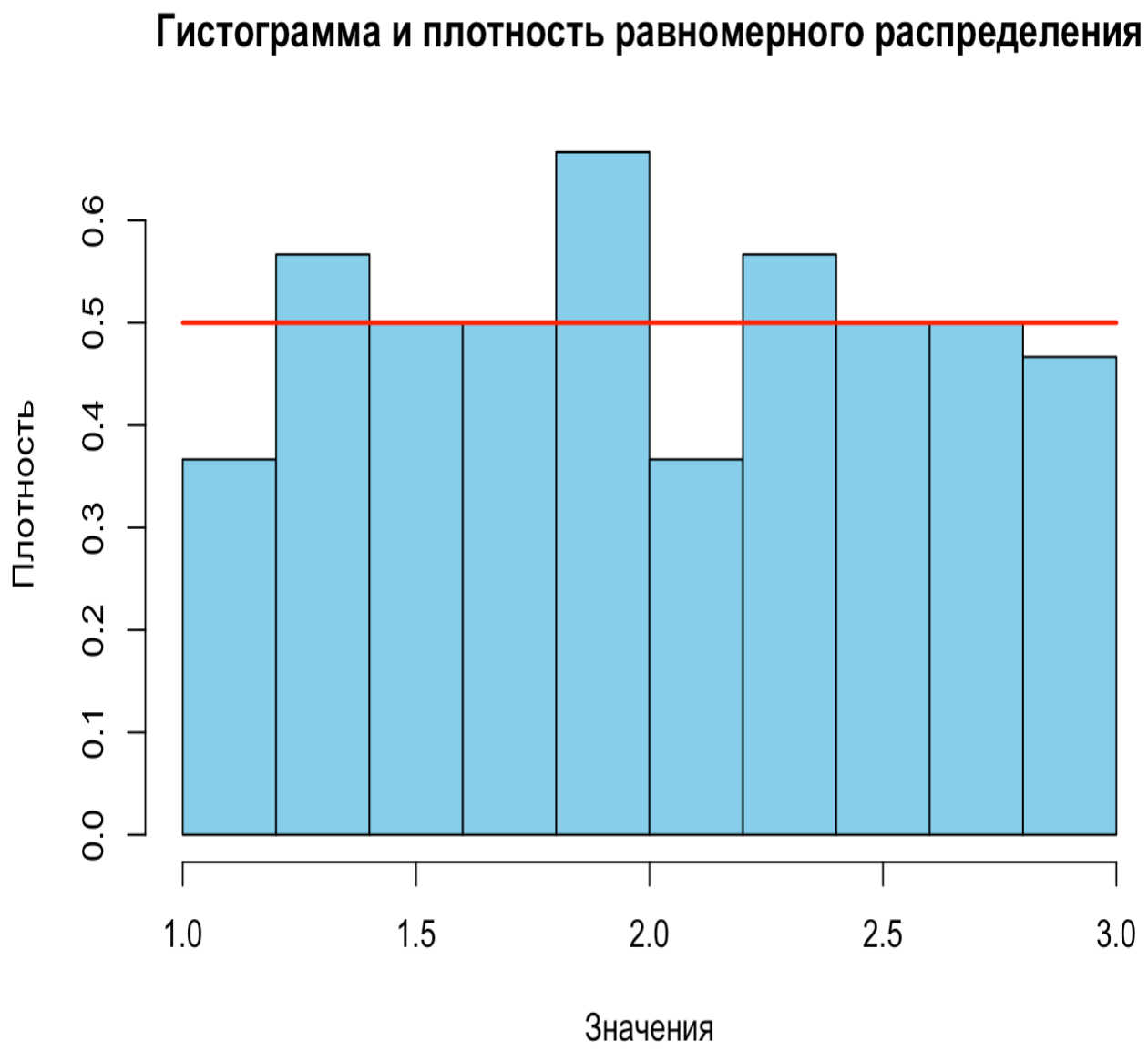
Метод максимального правдоподобия показал более точные результаты по сравнению с методом моментов в данной выборке. Вероятно, это связано с тем, что MLE более чувствителен к форме распределения данных и лучше адаптируется к истинному распределению данных, в то время как метод моментов может быть более уязвим к аномалиям в данных. На основе полученных данных можно предположить, что в выборке присутствуют выбросы или другие факторы, искажающие оценки методом моментов, делая его менее надежным для данного случая.

Построение гистограммы с плотностью распределения и проверка согласия по критерию хи-квадрат Пирсона

```
# Построение гистограммы с плотностью распределения и проверка согласия
data <- sample

# Построение гистограммы
hist(data, breaks = 10, probability = TRUE, col = "skyblue",
     main = "Гистограмма и плотность равномерного распределения",
     xlab = "Значения", ylab = "Плотность")

# Добавление теоретической плотности равномерного распределения
curve(dunif(x, min = 1, max = 3), add = TRUE, col = "red", lwd = 2)
```




```
# Для критерия Хи-квадрат нам необходимо разбить данные на бины
# Выберем разумное количество бинов, например 10
bins = seq(true_min, true_max, length.out = 11)  # создаем бины

# Создаем теоретические вероятности для равномерного распределения
# Каждый бин имеет равную вероятность, так как распределение равномерное
expected_counts = rep(n / length(bins), length(bins) - 1)

# Получаем частоты в бинах для нашей выборки
observed_counts = hist(data, breaks = bins, plot = FALSE)$counts
```

```
# Применяем критерий хи-квадрат Пирсона
chi_squared_test = chisq.test(x = observed_counts, p = expected_counts, rescale.p = TRU
E)


chi_squared_test
##
##  Chi-squared test for given probabilities
##
## data:  observed_counts
## X-squared = 4.4, df = 9, p-value = 0.8832
```

Интерпретация

X-squared (Хи-квадрат):

- Это значение статистики Хи-квадрат, которая измеряет разницу между ожидаемыми и наблюдаемыми частотами. Значение 4.4 указывает на то, что есть некоторое отклонение между ожидаемыми и наблюдаемыми частотами.

df (степени свободы):

- Это количество независимых наблюдений минус один. Здесь df равно 9.

p-value (уровень значимости):

- Это вероятность получить такие же или более экстремальные результаты, чем наблюдаемые, при условии, что нулевая гипотеза верна. Значение 0.8832 указывает на высокую вероятность того, что различия между ожидаемыми и наблюдаемыми частотами могут быть объяснены случайными факторами, так как p-value значительно больше обычного уровня значимости 0.05.

Исходя из этого теста, нет достаточных доказательств для того, чтобы отвергнуть нулевую гипотезу о том, что нет значимых различий между ожидаемыми и наблюдаемыми частотами.

Вычисление основных статистик

```
mean_val = mean(data)  # Среднее
variance_val = var(data)  # Дисперсия
sd_val = sd(data)  # Стандартное отклонение
sem_val = sd_val / sqrt(n)  # Ошибка среднего (стандартная ошибка среднего)
```

```
# Квартили
quantiles_val = quantile(data, probs = c(0.25, 0.5, 0.75))

# Асимметрия и эксцесс
# Подключаем библиотеку moments для доступа к функциям skewness и kurtosis
library(moments)

# Вычисляем асимметрию
skewness_val <- skewness(data)

# Вычисляем эксцесс, уменьшаем на 3 для получения эксцесса относительно нормального рас
пределения
kurtosis_val <- kurtosis(data) - 3

cat(sprintf("Среднее: %f\n", mean_val))
## Среднее: 2.008339
cat(sprintf("Дисперсия: %f\n", variance_val))
## Дисперсия: 0.329834
cat(sprintf("Стандартное отклонение: %f\n", sd_val))
## Стандартное отклонение: 0.574312
cat(sprintf("Ошибка среднего: %f\n", sem_val))
## Ошибка среднего: 0.046892
cat(sprintf("Квартили: %f\n", quantiles_val))
## Квартили: 1.491629
##  Квартили: 1.953113
##  Квартили: 2.508367
cat(sprintf("Асимметрия: %f\n", skewness_val))
## Асимметрия: 0.043790
cat(sprintf("Эксцесс: %f\n", kurtosis_val))
## Эксцесс: -1.218694
```

Интерпретация

Среднее значение (Mean):

- Среднее значение является суммой всех значений в выборке, разделенной на количество наблюдений.
- В данном случае, среднее значение равно 2.008339, что означает, что в среднем каждое наблюдение имеет значение около 2.01.

Дисперсия (Variance) и Стандартное отклонение (Standard Deviation):

- Дисперсия - это мера разброса данных относительно их среднего значения. Она вычисляется как средний квадрат разности между каждым значением и средним. Стандартное отклонение является квадратным корнем из дисперсии и измеряет среднее расстояние между каждым значением и средним значением.
- Значения дисперсии (0.329834) и стандартного отклонения (0.574312) показывают, что данные имеют относительно низкий разброс относительно среднего значения.

Ошибка среднего (Standard Error of the Mean):

- Ошибка среднего является оценкой стандартного отклонения распределения выборочных средних.Чем меньше значение ошибки среднего, тем более точно выборочное среднее оценивает среднее значение генеральной совокупности.
- Значение ошибки среднего (0.046892) относительно невелико, что указывает на то, что среднее значение выборки достаточно точно оценивает среднее значение генеральной совокупности.

Квартили (Quartiles):

- Квартили разбивают упорядоченные данные на четыре равные части. Квартили могут помочь понять распределение данных и их разброс.
- Первый квартиль (25-й процентиль) равен 1.491629, что означает, что 25% наблюдений меньше или равны этому значению.
- Медиана (второй квартиль, 50-й процентиль) равна 1.953113, что означает, что 50% наблюдений меньше или равны этому значению.
- Третий квартиль (75-й процентиль) равен 2.508367, что означает, что 75% наблюдений меньше или равны этому значению.

Асимметрия (Skewness):

- Асимметрия показывает, насколько сильно данные отклоняются от нормального распределения. Значение асимметрии близкое к нулю (0.043790) указывает на то, что распределение почти симметрично.

Эксцесс (Kurtosis):

- Эксцесс измеряет остроту пика распределения. Отрицательное значение эксцесса (-1.218694) означает, что пик распределения ниже и его хвосты более тяжелые, чем у нормального распределения.