Санкт-Петербургский политехнический университет Петра Великого
Институт прикладной математики и механики
Высшая школа прикладной математики и вычислительной физики

Многомерный статистический анализ

Отчет по лабораторной работе

Тема: Оценка закона распределения

Выполнил:
студент гр. 3630102/60401
Камалетдинова Ю. А.

Проверил:
к. ф-м. н., доцент
Павлова Л. В.

Санкт-Петербург
2020

Содержание

Постановка задачи

Цель лабораторной работы — проанализировать выборку, построить и обосновать модель закона распределения, которой подчиняются данные.

1. Ход работы

Первоначально, до выдвигания гипотезы о распределении, необходимо визуализировать данные. Они представляют собой вещественнозначную выборку $\{x\}_{i=1}^{n}$ объема $n = 60$, все значения положительны. Границы выборки равны $x_{min} \sim 0.005$ и $x_{max} \sim 5.032$. Гистограмма и эмпирическая функция распределения представлены на рисунке 1.

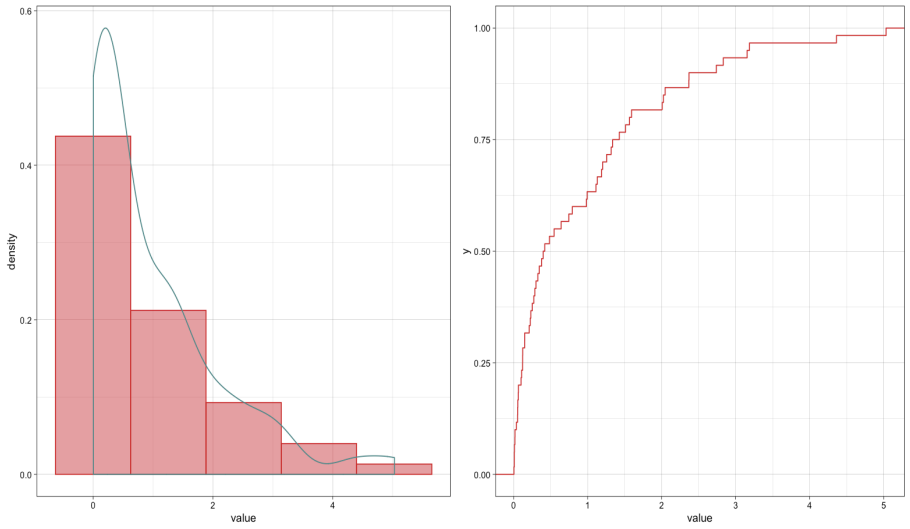

Рис. 1: Гистограмма и эмпирическая функция распределения набора данных

Гистограмма по выборке 1, очевидно, описывается асимметричным распределением. Вычислим некоторые статистки по формулам 1 - 4. По результатам из таблицы 1 можно сказать, что данные сильно положительно скошены ($\gamma_1 > 1$). Значение коэффициента эксцесса указывает на то, что перед нами распределение с тяжелыми хвостами.

$\overline{x}$	s	γ_1	γ_2
0.948	1.126	1.631	5.500

Таблица 1: Статистики по набору данных

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

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}} \tag{2}$$

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} \tag{3}$$

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3 \tag{4}$$

$$f(x;k) = \frac{1}{2^{(k/2)}\Gamma(k/2)} x^{k/2-1} e^{-x/2},\ k = 1, 2, \ldots \tag{5}$$

Критерий хи-квадрат требует, чтобы интервалы разбиения содержали не менее чем 5 наблюдений 2.

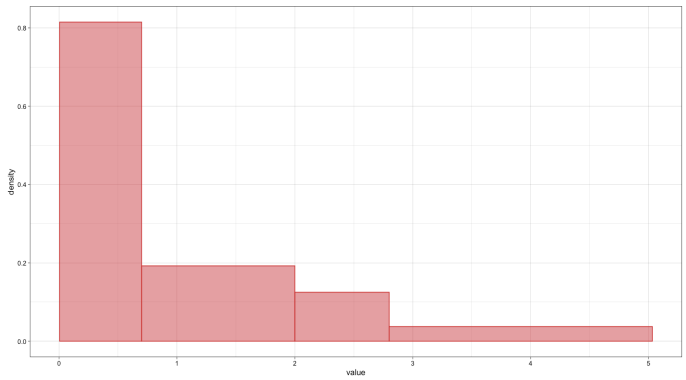

Рис. 2: Гистограммы с интервалами, где частоты попадания наблюдений ≥ 5

Для сравнения распределений применим критерий χ^2. Сложная гипотеза H_0 звучит так: наблюдения $\{x\}_{i=1}^n$ порождаются функцией $F(x,\theta),\ \theta \in \mathcal{R}^d$. Для этого посчитаем вероятности попадания в интервалы разбиения, зная истинные функции предполагаемых распределений, и минимизируем статистику χ^2 6 по параметру θ. Поскольку гистограмма по построенным данных явно несимметрична, будем проверять гипотезу для несимметричных распределений, а именно: логнормальное, экспоненциальное, гамма-распределение.

Критическое значение статистики χ^2 с 2 степенями свободы для правостороннего теста на уровне значимости 0.05 — 5.991, с 1 степенью свободы — 3.841. Приведем полученные значения статистки с оптимизированными параметрами распределения в таблице 2.

$$\chi^2 = \sum_{j=1}^{k} \frac{(n_j - E_j)^2}{E_j} \sim \chi^2_{k-d-1},\ k = 4,\ d = 1, 2 \tag{6}$$

Вид распределения	Статистика χ^2	Параметры
Lognorm	$1.667 \leq 3.841$	logmean = -0.41, logsd = 1.439
Exponential	$3.842 \leq 5.991$	rate = 0.942
Gamma	$1.066 \leq 3.841$	shape = 0.564, scale = 1.993

Таблица 2: Оптимизированные статистики для распределений

Таким образом, все статистики в таблице 2 меньше соответствующих критических значений критерия, и гипотезы не могут быть отклонены. Наименьшее значение статистики было получено в случае гамма-распределения. Посмтроим график функции плотности распределения с найденными параметрами и гистограмму данных 3.

2. Реализация

Для расчетов использовались библиотеки языка программирования **R**.

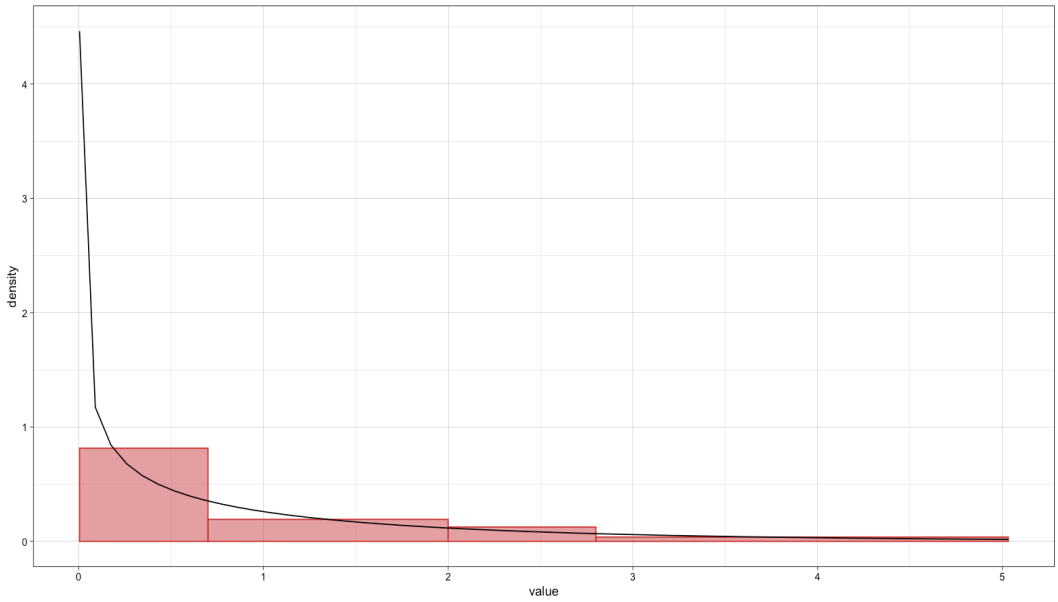

Рис. 3: Гистограммы данных и плотность гамма-распределения

Заключение

В ходе анализа выборки были выдвинуты и проверены сложные гипотезы о нескольких распределениях. Наиболее блинзким оказалось гамма-распределение с параметрами $k = 0.564, \theta = 1.993$.

Список литературы