Санкт-Петербургский Политехнический университет
Петра Великого
Физико-механический институт
Высшая школа прикладной математики и вычислительной физики

Лабораторная работа

по дисциплине "Многомерный статистический анализ"
на тему "Построение и обоснование модели закона распределения исследуемой случайной величины"
вариант 12

Выполнил студент
гр.5030102/90401 Кунгуров Ф.А.
Руководитель:
Доцент, к.ф.-м.н. Павлова Л. В.

Санкт-Петербург
2023

Содержание

1 Постановка задачи

Дана выборка $\{X_i\}_{i=1}^n, X_i \in \mathbb{R}, n = 60$. Требуется построить и обосновать модель закона распределения исследуемой случайной величины.

2 План

Чтобы восстановить распределение по выборке, нужно сделать следующие шаги:

- Посчитать выборочные характеристики (среднее, дисперсию, коэффициент асимметрии и эксцесса)
- Построить эмпирическую функцию распределения(а также 0.9 и 0.95 доверительные интервалы для теоретической функции распределения) и гистограмму, по которым выдвинуть гипотезу о семействе, к которому принадлежит исследуемая случайная величина.
- Проверить гипотезу с помощью хи-квадрат теста Фишера.
- После определения семейства найти параметры распределения методом максимального правдоподобия.
- Сравнить гистограмму с графиком плотности вероятности полученного в предыдущем пункте гипотетического распределения, а также эмпирическую функцию распределения с теоретической.

3 Выборочные характеристики и гистограмма

Вычисление выборочных характеристик и построение гистограммы позволяет исследователю провести первичный анализ выборки и получить представление о выборке без особых затрат.

Значения основных выборочных характеристик для данной выборки:

- Выборочное среднее $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i = 1.6$
- Выборочная несмещенная дисперсия $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2 = 2.17$
- Выборочный несмещенный коэффициент эксцесса (куртосис Фишера) $G_2 = \frac{n(n-1)}{(n-1)(n-2)(n-3)}\frac{\sum_{i=1}^n (X_i-\bar{X})^4}{(S^2)^2} - 3\frac{(n-1)^2}{(n-2)(n-3)} = 1.33$

- Выборочный несмещенный коэффициент асимметрии
  $G_1 = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{(S^2)^{1.5}} = 1.21$

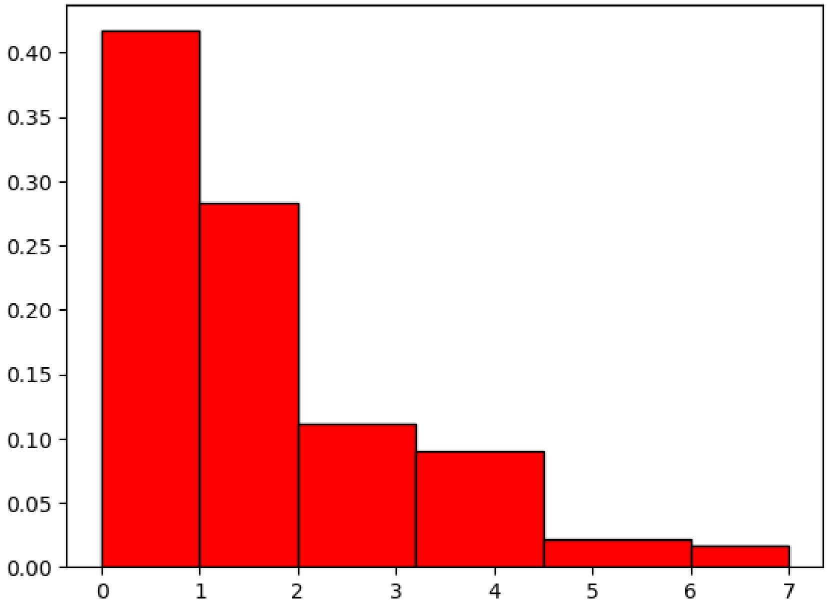


Рис. 1: Нормированная гистограмма

Похоже на экспоненциальное распределение...

4 Эмпирическая функция распределения и доверительные интервалы

Эмпирическая функция распределения $F_n(x)$ - ступенчатая функция, определяемая следующим образом:

$$F_n(x) = \sum_{i=1}^{n} I(x \leq X_i)$$

$I(x)$– функция-индикатор. Для построения доверительного интервала используется теорема Колмогорова, дающая оценку для скорости сходимости эмпирической функции к теоретической:

$$P\{n^{0.5} \sup_{t \in \mathbb{R}} |F_n(t) - F(t)|\} \xrightarrow{n} K(u)$$

$K(u)$– функция распределения Колмогорова. Используя данную оценку, можно с помощью квантилей распределения Колмогорова u_γ вычислить границы доверительного интервала с доверительной вероятностью γ для функции распределения:

$$\max\{0, F_n(t) - t^{-0.5} u_\gamma\} \leq F(t) \leq \min\{1, F_n(t) + t^{-0.5} u_\gamma\}$$

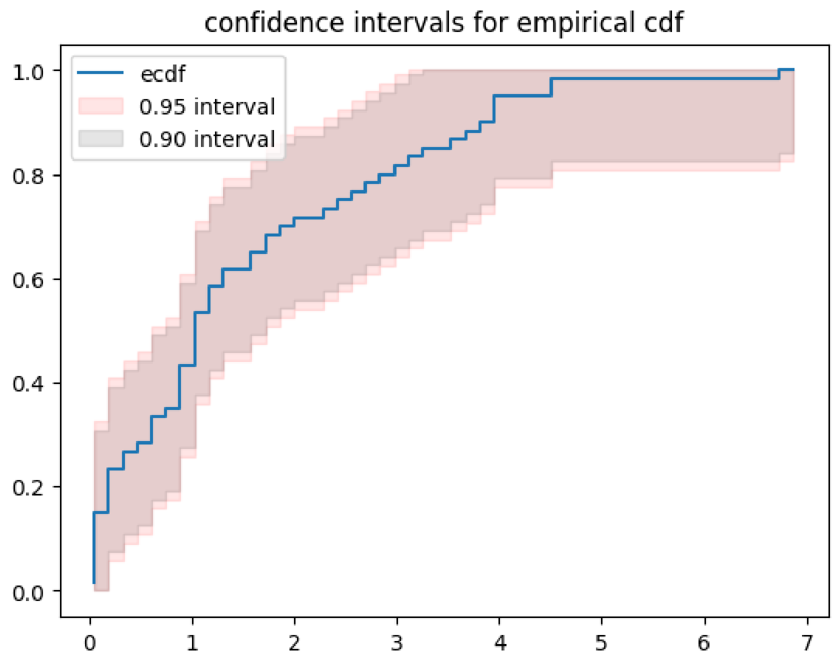


Рис. 2: Эмпирическая функция распределения с доверительными интервалами с доверительными вероятностями 0.9 и 0.95

5 Хи-квадрат тест Фишера

5.1 Описание теста

Тест позволяет проверить гипотезу о принадлежности выборки некоторому семейству законов распределения - $H_0 : F_n(t) \in \mathcal{F}_\theta = \{F(t;\theta)|\theta \in \Theta\}$, $\Theta \subset \mathbb{R}^r$– пространтво параметров семейства распределений.

В тесте используется X_n^2 статистика:

$$X_n^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i}$$

После группирования данных на N интервалов вычисляются O_i– фактическое кол-во попавших в i-ый интервал наблюдений и $E_i = np_i$ - ожидаемое кол-во попаданий в i-ый интервал в предположении H_0.

Если нулевая гипотеза верна, то $X_n^2 \xrightarrow{n} X^2 \sim \chi^2_{N-r-1}$. Тогда, выбрав уровень значимости α, H_0 принимается, если значение статистики не превосходит $1-\alpha$ квантиль распределения хи-квадрат $t_{1-\alpha,N-r-1}$, в противном случае - принимается альтернатива $H_1 : F_n(t) \notin \mathcal{F}_\theta$

5.2 Практические соображения к использованию теста

Чтобы вычислить p_i, нужно знать параметры распределения θ', к которому принадлежит выборка в предположении H_0 - тогда p_i - интеграл от плотности вероятности $f_{\theta'}$ по i-му интервалу. Искать θ' можно как решение оптимизационной задачи:

$$\theta' = \arg\min_{\theta \in \Theta} \sum_{i=1}^{N} \frac{(O_i - E_i(\theta))^2}{E_i(\theta)}$$

Также рекомендуется брать такое разбиение выборки, что в каждый интервал попало бы не менее 5 наблюдений.

6 Тест на экспоненциальное распределение

6.1 Хи-квадрат тест

Чтобы разбить выборку на интервалы с заданным минимальным кол-вом попаданий, была взята равномерная сетка

$$\{x_i\}_{i=0}^{l},\ x_0 = \min_i X_i,\ x_l = \max_i X_i$$

Уровень значимости α принят равным 0.05 - стандартное значение. Получившееся значение статистики значительно меньше критического значения:

$$X_n^2 = 4.1 \leq 9.49 = t_{0.95,N-r-1}$$

Значит, нулевая гипотеза принимается.

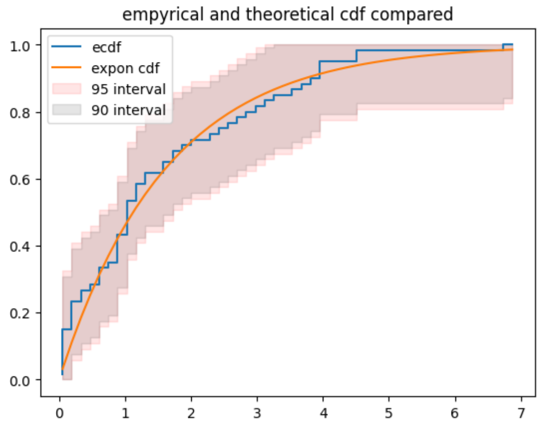
7 Оценка параметров

Оценка параметра экспоненциального распределения - интенсивности λ методом максимального правдоподобия совпадают с выборочными смещенными оценками. В нашем случае:

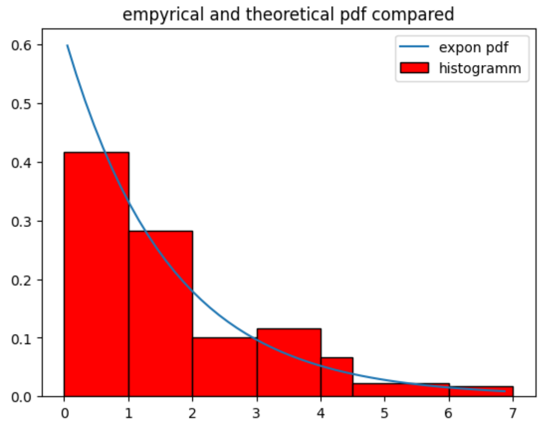
- $\lambda = \frac{1}{\hat{\mu}} = 0.618$

8 Сравнение гипотетического распределения с выборочным

Теперь можно построить совместные графики функций распределения и плотности вероятности для теоретической и эмпирических распределений и проверить, насколько хорошо получившееся распределение соответствует выборке:



(a) Эмпирическая и гипотетическая функция распределения



(b) Гистограмма и гипотетическая функция плотности

9 Вывод

Анализируя результаты, можно сделать выводы:

- Распределение было действительно экспоненциальным (или гамма-распределением с параметром равному единице),в том числе потому что p-value получилось большим уровня значимости (0.39 >> 0.05)

- Гистограмма в общем приемлемо приближает функцию плотности вероятности, но данные смещены вправо из-за небольшого размера выборки (о чем можно было судить еще и по коэффициенту асимметрии)