Санкт-Петербургский государственный университет

Кафедра системного программирования

Группа 23.Б11-мм

Реализация критериев для распределения Вейбулла

Усольцев Данил Вячеславович

Отчёт по учебной практике
в форме «Решение»

Научный руководитель:
доцент кафедры системного программирования, к. ф.-м. н. Гориховский В. И.

Санкт-Петербург
2025

Оглавление

Введение

Распределение Вейбулла является одним из самых распространенных и широко используемых в статистике. Оно позволяет моделировать поведение случайных величин в различных областях, от инженерии до медицины. Уникальность этого распределения заключается в его гибкости — при определенных значениях параметров оно может описывать различные типы распределений, что делает его универсальным инструментом для анализа данных.

Важнейшей задачей в применении распределения Вейбулла является проверка его пригодности для описания данных. Для этого разработаны критерии, которые помогают оценить, насколько сильно данное распределение близко к эмпирическим данным. Реализация таких критериев позволяет проводить более точный и обоснованный анализ.

Современные исследования предлагают различные критерии для проверки гипотезы о принадлежности данных распределению Вейбулла. Эти критерии, основанные на статистических методах, позволяют глубже понять характер данных и проверять их соответствие заданной модели. В данной работе ставится задача реализовать такие критерии, основанные на материалах научных статей и перенести уже реализованные на языке R в библиотеку на Python.

1. Постановка задачи

Целью работы является реализация критериев для распределения Вейбулла в пакете на Python. Для её выполнения были поставлены следующие задачи:

1. Поиск научных статей, где описаны различные критерии проверки распределения Вейбулла и примеры их реализации, а также уже реализованных критериев в R-библиотеках;
2. Реализация критериев в Python-пакете;
3. Разработка методики валидации и её проведение;
4. Проведение сравнения алгоритмов по метрикам, таким как: уровень значимости и производительность;

Справочные материалы

Теоретическая функция распределения для распределения Вейбулла (**CDF**) имеет следующий вид:

$$F(x, k, \lambda, a) = \left[1 - e^{-\left(\frac{x}{\lambda}\right)^k}\right]$$

Функция плотности вероятности (**PDF**) выглядит как:

$$f(x; k, \lambda; \alpha) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k}$$

Эмпирическая функция распределения (ECDF) определяется как доля наблюдений в выборке, которые меньше или равны заданному значению $\boldsymbol{x}$:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(X_i \leq x\right),$$

где: n - размер выборки, X_i - значения выборки, $\mathbb{I}\left(X_i \leq x\right)$ - функция, которая равна 1 , если $X_i \leq x$, и 0 в противном случае.

2. Обзор

Существуют различные библиотеки для проведения статистического анализа распределения Вейбулла, однако в осноном они реалиованы на языке программирования R, с которым знаком далеко не каждый. Поскольку с Python знакомо большее количество пользователей, имеет смысл адаптировать критерии проверки распределения Вейбулла именно для этого языка программирования. В рамках данной работы будет проведена реализация таких критериев на Python, что позволит облегчить их использование и расширить аудиторию потенциальных пользователей.

2.1. Критерий Колмогорова-Смирнова

Критерий оценивает максимальное отклонение между эмпирической функцией распределения (ECDF) и теоретической функцией распределения (CDF):

$$D_n = \max_x |F_n(x) - F(x)|$$

где $F_n(x)$ - эмпирическая CDF, $F(x)$ - теоретическая CDF.

Критерий Колмогорова-Смирнова универсален и прост в реализации. Его основное преимущество — чувствительность к отклонениям в центральной части распределения. Однако он менее эффективен на краях и плохо масштабируется для многомерных данных. Также критерий может быть слишком строгим для больших выборок или недостаточно эффективным для малых.

2.2. Критерий Андерсона-Дарлинга

Критерий увеличивает чувствительность в хвостах распределения. Его статистика определяется как:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}\left[(2i-1)\left(\ln F\left(X_{(i)}\right) + \ln\left(1 - F\left(X_{(n+1-i)}\right)\right)\right)\right]$$

2.3. Критерий хи-квадрат

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

где: O_i - число наблюдений в i-м интервале, E_i - ожидаемое число наблюдений в i-м интервале, $k-$ количество интервалов.

2.4. Критерий Крамера-Мизеса

Критерий учитывает отклонения во всех частях распределения, а не только в экстремумах, а также менее чувствителен к экстремальным значениям по сравнению с другими критериями

$$W^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left[F\left(X_i\right) - \frac{2i-1}{2n}\right]^2$$

,

2.5. Критерий Смита и Бейна

Статистика Z для теста выглядит так:

$$Z_2 = n(1 - R_{SB}^2)$$

$$R_{SB}^2 = \frac{\left[\sum_{i=1}^{n}(\ln X_i - \ln \bar{X})(c_i - \bar{c}_n)\right]^2}{\sum_{i=1}^{n}(\ln X_i - \ln \bar{X})^2 \sum_{i=1}^{n}(c_i - \bar{c}_n)^2}$$

X_i — упорядоченные статистики (порядковые значения выборки).

$c_i = \ln[-\ln(1-p_i)]$, где p_i — оценки ранговых вероятностей:

$p_i = \frac{i}{n+1}, \quad i = 1, \ldots, n$

$\ln X_i$ — логарифмы упорядоченных значений выборки.

$\ln \bar{X}$ — среднее логарифмов значений X_i^*.

$\bar{c}_n$ — *среднее значение* c_i*:*

$$\bar{c}_n = \frac{1}{n} \sum_{i=1}^{n} c_i$$

2.6. Критерий Эванса, Джонсона и Грина

Этот тест основан на аналогичной статистике R_{SB}^2, но использует другие веса M_i:

$$R_{EJG}^2 = \frac{\left[\sum_{i=1}^{n} \left(\ln X_i - \ln \bar{X}\right) M_i\right]^2}{\sum_{i=1}^{n} \left(\ln X_i - \ln \bar{X}\right)^2 \sum_{i=1}^{n} \left(M_i - \bar{M}_n\right)^2},$$

где M_i определяется как:

$$M_i = \frac{1}{\hat{\beta}_n} \left[-\ln \left(1 - \frac{i - 0.3175}{n + 0.365} \right) \right].$$

$\hat{\beta}_n$ — оценка параметра формы распределения Вейбулла. $\bar{M}_n$ - среднее значение M_i.

2.7. Статистика Шапиро и Брейна

$$S_{B_n} = \frac{\left(\frac{1}{n} \sum_{i=1}^{n} \left[0.6079 w_{n+i} - 0.257 w_i\right] \ln X_i\right)^2}{\frac{1}{n} \sum_{i=1}^{n} \left(\ln X_i - \ln \bar{X}\right)^2}.$$

$\ln X_i$ - логарифмы упорядоченных значений выборки.

$\ln \bar{X}$ - среднее значение логарифмов выборки.

w_{n+i} и w_i - специальные веса, которые корректируют стандартное отклонение и задаются следующим образом:

$$w_i = \ln \frac{n+1}{n-i+1} \quad \forall i \in \{1, \ldots, n-1\}$$

$$w_n = n - \sum_{i=1}^{n-1} w_i$$

$$w_{n+i} = w_i \left(1 + \ln w_i\right) - 1 \quad \forall i \in \{1, \ldots, n-1\}$$

$$w_{2n} = 0.4228n - \sum_{i=1}^{n-1} w_{n+i}$$

n - размер выборки.

2.8. Статистика Озтюрка и Корукоглу

$$OK_n = \frac{\ln 2(n-1) \sum_{i=1}^{n} \left[0.6079 w_{n+i} - 0.257 w_i\right] \ln X_i^*}{\sum_{i=1}^{n} (2i - n - 1) \ln X_i^*}$$

Для улучшения теста Озтюрк и Корукоғлу предложили стандартизировать статистику OK_n чтобы сделать её более чувствительной к отклонениям.

$$OK_n^* = \frac{OK_n - 1 - 0.13/\sqrt{n} + 1.18/n}{0.49/\sqrt{n} - 0.36/n}.$$

2.9. Статистика Манна, Шойера и Фертига

Нормализованные интервалы E_i определяются следующей формулой:

$$\forall i \in \{1, \ldots, n\}, E_i = \frac{\ln X_i - \ln X_{i-1}}{\mathbb{E}\left[\ln X_i - \ln X_{i-1}\right]}$$

E_i - нормализованный интервал между логарифмами упорядоченных значений.

Сама статистика выглядит следующим образом:

$$MSF_n = \frac{\sum_{j=\left\lfloor \frac{n}{2} \right\rfloor + 2}^{n} E_j}{\sum_{j=2}^{n} E_j}$$

2.10. Статистика Тику и Сайна

$$TS_n = \frac{2\sum_{i=2}^{n-1}(n-i)E_i}{(n-2)\sum_{i=2}^{n}E_i}$$

.

E_i - нормализованный интервал между упорядоченными значениями выборки.

2.11. Статистика О'Рейли и Стивенса

$$Z_j = \frac{\sum_{i=2}^{j}E_i}{\sum_{i=2}^{n}E_i}, j = 2, \ldots, n-1$$

2.12. Статистика Lockhart, O'Reilly и Stephens

$$LOS_n = 2 - n - \frac{1}{n-2}\sum_{i=2}^{n-1}\left[(2(i-n)+1)\ln(1-X_i) - (2i-3)\ln X_i\right]$$

2.13. Модифированная статистика Крамера-Мизеса

Статистика DS_n представляет собой модификацию критерия Крамера-Мизеса и выглядит следующим образом:

$$DS_n = n\int_0^{\infty}\frac{(F_n(x)-F(x))^2}{F(x)}\,dF(x)$$

В отличие от статистики Крамера-Мизеса, модицированная статистика более чувствительна к отклонениям в хвостах распределения.

Статистика Крамера-Мизеса хорошо работает для центральных областей распределения, но недостаточно уделяет внимания хвостам.

Модифицированная статистика справедливо распределяет внимание между центральной частью и хвостами, что делает её более уни-

версальной.

2.14. Статистика Liao-Shimokawa

$$L_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\max\left[\frac{i}{n} - F\left(x_i; \hat{\theta}\right), F\left(x_i; \hat{\theta}\right) - \frac{i-1}{n}\right]}{\sqrt{F\left(x_i; \hat{\theta}\right)\left[1 - F\left(x_i; \hat{\theta}\right)\right]}}$$

где n — объём выборки, $x_1, x_2, ..., x_n$ — упорядоченные по возрастанию элементы выборки.

Данная статистика содержит в себе идеи Колмогорова-Смирнова, Крамера-Мизеса, Андерсона-Дарлинга статистик.

Её основные преимущества состоят в том, что статистика Ляо и Шимокавы хорошо работает с небольшими объемами данных, позволяет выделять хвостовые отклонения и относительно проста в реализации.

2.15. Статистика Watson‘a

Критерий Ватсона — это модификация критерия Крамера-Мизеса. Его основная цель — устранить зависимость от сдвигов данных, делая его более устойчивым к смещению и особенно полезным для анализа данных на круге или в других циклических пространствах (например, для угловых данных).

Критерий выглядит следующим образом:

$$U_n^2 = \sum_{i=1}^{n} \left(F\left(x_i, \theta\right) - \frac{i - 0,5}{n} \right)^2 - n \left(\frac{1}{n} \sum_{i=1}^{n} F\left(x_i, \theta\right) - 0,5 \right)^2 + \frac{1}{12n},$$

где n — объём выборки, $x_1, x_2, ..., x_n$ — упорядоченные по возрастанию элементы выборки.

2.16. Статистика, основанная на информации Кульбака-Лейблера

$$KL_{m,n} = -\frac{1}{n}\sum_{i=1}^{n}\ln\left[\frac{n}{2m}\left(X_{i+m} - X_{i-m}\right)\right] - \frac{1}{n}\sum_{i=1}^{n}X_i + \frac{1}{n}\sum_{i=1}^{n}e^{X_i}$$

.

где X_i — это выборка, полученная с использованием оценок моментов.

2.17. Статистика, использующая преобразование Лапласа

$$\hat{v}_n(s) = \sqrt{n}\left(\frac{1}{n}\sum_{j=1}^{n}e^{-Y_j s} - \Gamma(1-s)\right),$$

где $\boldsymbol{s}$ - это параметр, который может принимать значения из некоторого интервала J.

где Y_j — это выборка, полученная с использованием оценок моментов.

Рекомендуется выбирать интервал J как подмножество [-2.5; 0.49].

2.18. Статистика Cabana и Quiroz

$$\hat{C}_{Q_n} = \hat{v}_n^t A^{-1}\hat{v}_n$$

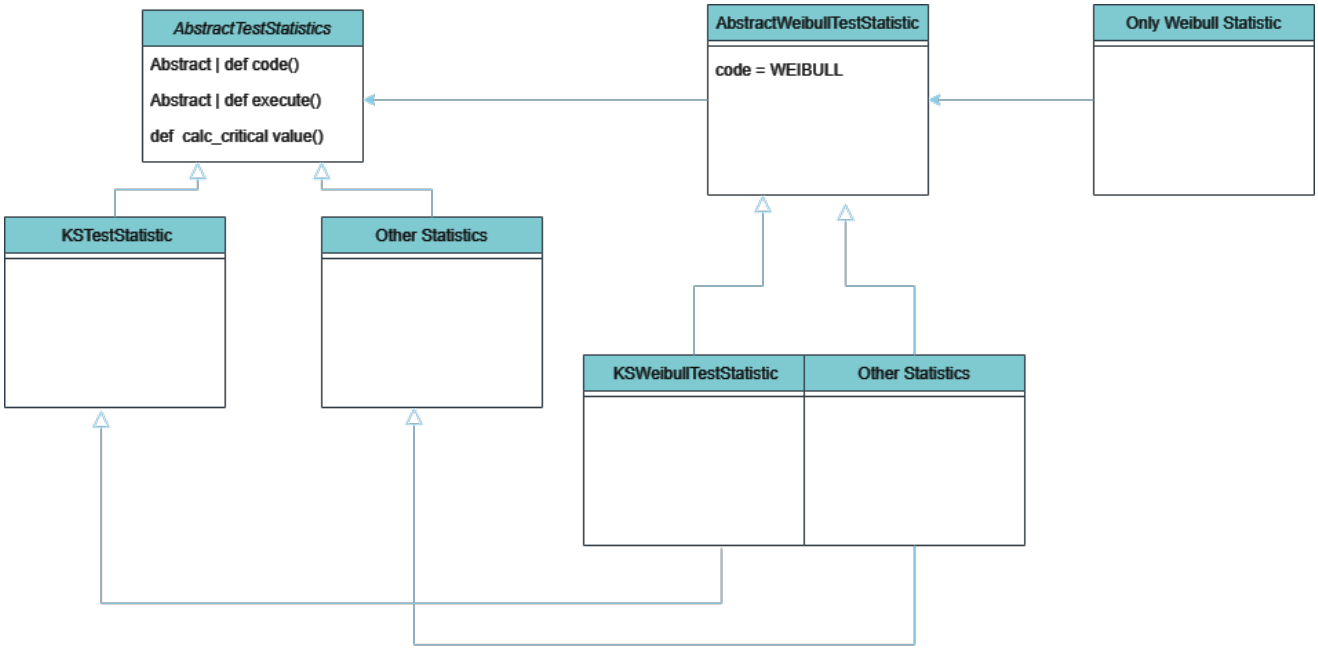
где $\hat{v}_n^t$ - вектор оценок, полученных из выборки.

$S = \{-0.1, 0.02\}$, $A = \begin{bmatrix} 1.59 & 0.91 \\ 0.91 & 0.53 \end{bmatrix}$

3. Описание решения

В библиотеке на языке R уже реализовано некоторое количество критериев, которые были рассмотрены в обзоре. Некоторые из них были уже реализованы в нашем пакете: Колмогорова-Смирнова, Андерсона-Дарлинга, хи-квадрат, Крамера-Мизеса, Смита и Бейна, Эванса, Джонсона и Грина, Шапиро и Брейна, Озтюрка и Корукоглу, Манна-Шойера и Фертига, Тику и Сайна, О'Рейли и Стивенса, Lockhart-O'Reilly и Stephens. В рамках своей работы я перенес из R-пакета следующие критерии: Liao-Shimokawa, Watson'a, Кульбака-Лейблера, Cabana и Quiroz. Остальные критерии были взяты из научных статей и с их помощью реализованы.

3.1. Архитектура



- Для обеспечения универсальности и уменьшения дублирования кода, критерии, которые применимы к нескольким типам распределений (например, *Колмогорова-Смирнова, Крамера-Мизеса*), реализованы в отдельном модуле *common.py*.

- Пакет также поддерживает работу с другими распределениями, такими как нормальное, экспоненциальное и т.д. Для каждого из этих распределений могут быть реализованы свои критерии.

3.2. Особенности реализации

- В нашем пакете распределение Вейбулла характеризуется двумя параметрами: масштабом (λ) и формой (k).

- Некоторые критерии используют дополнительные параметры для вычисления статистик. Для обеспечения удобства использования задаются стандартные значения для этих параметров. Это позволяет пользователю пакета использовать критерии без необходимости вручную настраивать все параметры.

- Для работы с распределением Вейбулла был разработан абстрактный класс *AbstractWeibullTestStatistic*, который служит базовым классом для всех критериев, специфичных для этого распределения.

- На основе класса *AbstractWeibullTestStatistic* были реализованы критерии, представленные в обзоре.

```
class AbstractTestStatistic(ABC):
    @staticmethod
    @abstractmethod
    def code() -> str:
        raise NotImplementedError("Method is not implemented")

    @abstractmethod
    def execute_statistic(self, rvs, **kwargs) -> float or float64:
        raise NotImplementedError("Method is not implemented")

    def calculate_critical_value(self, rvs_size, sl) -> Optional[float] or
    ↪  Optional[float64]:
        return None
```

4. Валидация

Основная цель методики валидации — убедиться, что перенос статистик с языка R на Python выполнен корректно и результаты вычислений находятся в допустимых пределах.

4.1. Описание данных

Для проверки реализованных критериев были выбраны тестовые данные, состоящие из 50 значений:

0.2215	0.4445	0.8579	1.5711	0.7127
0.6289	2.5255	0.9512	1.1088	0.4055
2.1205	0.4848	0.4887	0.9066	0.5249
0.1320	0.6222	1.2585	0.0996	1.1450
0.4176	1.7122	1.1898	0.3077	0.6360
1.5284	0.2197	0.2458	0.2303	1.2696
0.4827	0.5641	0.4107	2.0854	0.4701
0.2559	0.7593	0.6767	0.2414	0.8581
0.1762	1.3629	0.8801	0.2669	0.2269
1.2188	0.3703	1.5553	0.2566	0.7727

4.2. Методика валидации

Методика валидации заключается в сравнении значений статистик, рассчитанных с помощью Python и R. Основные шаги:

- Используются одинаковые наборы данных в R и Python
- Расчет значения у статистик в R-пакете
- Рассчет значения у реализованных статистик в Python-пакете
- Сравнение значений

4.3. Результаты

Были получены следующие результаты:

Критерий	**Python**	**R**
Лиао-Шимокавы	1.2425	1.1551
Уотсона	0.2345	0.1110
Кульбака-Лейблера	0.3231	0.1742
Cabana и Quiroz	0.0106	0.0113

Результаты статистик для критериев Уотсона и Кульбака-Лейблера имеют более значительные расхождения по сравнению с другими статистиками из-за особенности реализации. При этом расхождения не критичны, что позволяет считать реализации корректными для практического применения.

5. Сравнение алгоритмов

5.1. Уровень значимости

Было сгенерировано $N = 10000$ выборок, подчиняющихся нулевой гипотезе H_0 (закон распределения Вейбулла). Для каждой выборки применены реализованные критерии, и зафиксировано количество случаев, когда H_0 была отвергнута при уровне значимости $\alpha = 0.05$

Критерий	**Кол-во отвергнутых выборок**
Лиао-Шимокавы	505
Уотсона	498
Кульбака-Лейблера	529
Cabana и Quiroz	466
Мод. Крамер вон Мизес	509
Преобразование Лапласа	502

Проверка уровня значимости подтвердила, что каждый критерий в равной степени соблюдает установленное значение

5.2. Производительность

Тест на производтельность производился на следующем устройстве:
Ноутбук: Xiaomi Redmibook15
Процессор: i3-1115G4 3.00GHz
Оперативная память: 8 GB

Размер выборки (n)	**n=50**	**n=100**	**n=400**
Лиао-Шимокава	13.47	13.55	17.82
Уотсон	20.55	21.13	29.11
Кульбак-Лейблер	13.55	20.06	72.12
Cabana и Quiroz	6.71	7.12	8.66
Мод. Крамер вон Мизес	18.13	29.65	111.04
Преобразование Лапласа	9.71	11.34	18.58

Заключение

В течение семестра были достигнуты следующие результаты:

- Были изучены научные статьи, в которых описаны различные критерии проверки распределения Вейбулла. Проведен анализ существующих реализаций критериев в R-библиотеках, таких как EWGoF

- Реализованы следующие критерии проверки распределения Вейбулла: Критерий Ляо-Шимокава, Статистика Ватсона, Критерий Кульбака-Лейблера, Статистика Cabana-Quiroz, Модифицированная статистка Крамера-Мизеса, Статистика, использующая преобразование Лапласа

- Разработана методика валидации и в дальнейшем была применена для оценки качества реализации критериев

- Проведено сравнение критериев

С реализацией можно ознакомиться на GitHub https://github.com/PySATL/pysatl-experiment.

Список литературы

[1] Mahdi Doostparast. *Goodness-of-fit tests for weibull populations on the basis of records*. https://www.researchgate.net/publication/51948640_Goodness-of-fit_tests_for_weibull_populations_on_the_basis_of_records.

[2] Meryam Krit. *Goodness-of-fit tests in reliability : Weibull distribution and imperfect maintenance models*. https://theses.hal.science/tel-01126901/document.

[3] Min Liao and Toshiyuki Shimokawa. *A New Goodness-of-Fit Test for Type-1 Extreme-Value and 2-Parameter Weibull Distributions with Estimated Parameters*. https://doi.org/10.1080/00949659908811965.

[4] Критерий согласия Ватсона. https://ru.wikipedia.org/wiki/Критерий_согласия_Ватсона.

[5] Gholamhossein Yari, Alireza Mirhabibi and Abolfazl Saghafi. *Estimation of the Weibull parameters by Kullback-Leibler divergence of Survival functions*. https://www.naturalspublishing.com/files/published/16vf9f1n913359.pdf.

[6] Alejandra Cabana, Adolfo J. Quiroz. *Using the Empirical Moment Generating Function in Testing for the Weibull and the Type I Extreme Value Distributions*. https://link.springer.com/article/10.1007/BF02595411.

[7] Meryam Krit. *Goodness-of-fit tests for the Weibull distribution based on the Laplace transform*. http://www.numdam.org/item/JSFS_2014__155_3_135_0/.