Санкт-Петербургский государственный университет

Григорьев Дмитрий Артемович

Выпускная квалификационная работа

Исследование эффективности некоторых непараметрических критериев проверки гипотез о разрывности функции регрессии и интенсивности пуассоновского процесса

Уровень образования: бакалавриат

Направление 01.03.02 «Прикладная математика и информатика»

Основная образовательная программа СВ.5004.2018 «Прикладная математика и информатика»

Профиль «Вычислительная стохастика и статистические модели»

Научный руководитель:

Профессор, кафедра статистического моделирования

д. ф.-м. н., профессор М. С. Ермаков

Рецензент:

Старший научный сотрудник,

Лаборатория статистических методов

ПОМИ РАН

к. ф.-м. н. В. Н. Солев

Санкт-Петербург

2022

Saint Petersburg State University
Applied Mathematics and Computer Science
Computational Stochastics and Statistical Models

Grigorev Dmitrii Artemovich

Graduation Project

Investigation of efficiency of some nonparametric tests for testing discontinuity of regression function or Poisson process intensity

Scientific Supervisor:
Professor, Department of Statistical
Modelling, D.Sc. M. S. Ermakov

Reviewer:
Senior research fellow, Laboratory of
Statistical Methods of POMI RASc
Ph.D. V. N. Solev

Saint Petersburg
2022

Оглавление

Введение

Задача о проверке непрерывности или разрывности функций в статистике имеет множество приложений как в случае распределений случайных величин, так и процессов и временных рядов. В класс рассматриваемых функций входят функция регрессии одной случайной величины на другую и плотность интенсивности пуассоновского процесса.

Задача о разрывности функции регрессии исследована в большом количестве работ, связанных с общественными науками. Впервые идея проверки гипотезы о разрывности появилась в психологии [1], но также нашла применение в исследованиях в экономической [2], образовательной [3], политической [4] и прочих сферах. В работе [5] для функции регрессии, представленной как условное математическое ожидание, получены асимптотические свойства статистики критерия проверки гипотезы о непрерывности.

Задача о проверке разрывности функции плотности интенсивности пуассоновского процесса входит в круг задач, имеющих отношение к обнаружению разладок (в англоязычной литературе *change-point analysis*) в процессах и временных рядах. Задачи такого рода встречаются в таких дисциплинах, как акустика [6], геномика [7], океанография [8] и т.д. Методы обнаружения сводят эту задачу к максимизации функции логарифмического правдоподобия [9, 10] или к статистической проверке гипотез [11, 12]. Насколько известно автору, непараметрическая проверка рассматриваемой гипотезы о разрывности плотности интенсивности пуассоновского процесса ещё не изучалась.

В рамках данной работы предложен и изучен в плане асимптотических свойств непараметрический критерий проверки гипотезы разрывности функции регрессии, построенный на основе оценки условной медианы, а также построена и изучена его модификация на основе теории перестановочных критериев для малых объёмов выборки. Также предложен и изучен непараметрический критерий проверки гипотезы о разрывности плотности интенсивности пуассоновского процесса. В основу критериев положено ядерное оценивание регрессии и плотности интенсивности, а также используется перестановочный подход построения критериев.

Цели работы:

- Изучить литературу о свойствах ядерных оценок и построении на их основе перестановочных критериев;
- Построить непараметрический критерий проверки гипотезы о разрывности функции регрессии на основе ядерной оценки условной медианы, изучить его асимптотические свойства аналитически и построить его перестановочную модификацию;
- Сравнить в моделировании построенный критерий с известным, представленным в литературе критерием, построенным на основе оценки условного матожидания;
- Построить непараметрический критерий проверки гипотезы о разрывности плотности интенсивности пуассоновского процесса с использованием ядерной оценки разрыва плотности интенсивности и аналитически изучить его асимптотические свойства;
- Для всех построенных критериев провести моделирование для валидации теоретических результатов.

Организация работы:

Глава 1 данной работы относится к построению и изучению критерия проверки гипотезы разрывности функции регрессии, построенного на основе оценки условной медианы, и его статистики. Раздел 1.1 содержит необходимые определения и обозначения, относящиеся к этой главе. Раздел 1.2 приводит постановку задачи проверки гипотезы о разрывности функции регрессии в виде условной медианы. Раздел 1.3 посвящен построению критерия проверки этой гипотезы, для которого в этом разделе сформулировано основное утверждение о предельном распределении и дисперсии его статистики. В разделе 1.4 приведены результаты моделирования на модельных данных, подтверждающего теоретические результаты, полученные в разделе 1.3. Перестановочный вариант критерия в обычной и стьюдентизированной формах, предназначенный для меньших объемов выборки, представлен в разделе 1.5 вместе с моделированием, демонстрирующим ситуации, когда они неравносильны. Раздел 1.6 посвящен сравнению построенного критерия с известным, представленным в литературе критерием, построенном на основе ядерной оценки условного матожидания, предложенным в литературе, на модельных

данных в терминах задачи о робастности их статистик. Моделирование показало, что статистика построенного критерия более устойчива к шуму в данных, чем аналогичная.

Глава 2 данной работы относится к построению и изучению критерия проверки гипотезы разрывности плотности интенсивности пуассоновского процесса и его статистики. Раздел 2.1 содержит необходимые определения, обозначения и известные результаты, относящиеся к теории пуассоновского процесса. В разделе 2.2 приведена постановка задачи проверки гипотезы о разрывности плотности интенсивности, а также вместе с построением критерия для проверки этой гипотезы представлено основное утверждение этой главы, формулирующее асимптотические свойства статистики построенного критерия, а именно предельное распределение и предельная дисперсия. В разделе 2.3 представлены результаты моделирования на модельных данных, которые подтверждают теоретические результаты раздела 2.2.

В приложении А представлены доказательства утверждений из разделов 1.3 и 2.2. Приложение Б содержит дополнительную теорию, относящуюся к моделированию, представленному в разделе 1.4. Приложение В содержит дополнительну теорию, относящуюся к главе 1.

Глава 1

Критерий проверки разрывности функции условной медианы

1.1. Необходимые определения

Вероятностная O-символика

Определение 1.1 (O-большое по вероятности)**.** Говорят, что случайные величины $\mathbf{X}_n$ являются O-большим для констант c_n по вероятности (сокращение: $\mathbf{X}_n = O_{\mathcal{P}}(c_n)$), если

$$\forall \varepsilon > 0 \; \exists M > 0 \; \exists N > 0 : \forall n \geq N \; \mathbb{P}\left(\left|\frac{\mathbf{X}_n}{c_n}\right| > M\right) < \varepsilon.$$

Определение 1.2 (o-малое по вероятности)**.** Говорят, что случайные величины $\mathbf{X}_n$ являются o-малым для констант c_n по вероятности (сокращение: $\mathbf{X}_n = o_{\mathcal{P}}(c_n)$), если

$$\forall \varepsilon > 0 \; \mathbb{P}\left(\left|\frac{\mathbf{X}_n}{c_n}\right| \geq \varepsilon\right) \xrightarrow{n\to\infty} 0.$$

Ядерное оценивание плотности

Самый известный метод оценивания плотности — это построение гистограммы. Гистограмма легко интерпретируется, так как явно показывает частоту отдельных наблюдений или группы наблюдений. Но данный метод обладает существенными недостатками: во-первых, форма гистограммы зависит от положения столбца частоты первой группы, во-вторых, гистограмма не непрерывна (см. [13]). Эти проблемы решает ядерная оценка плотности.

Пусть дана выборка $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ из распределения с плотностью $p(x)$.

Определение 1.3 (Ядро)**.** Функция $K(u)$ называется ядром, если

- K симметрична, неотрицательна,
- $\int_{\mathbb{R}} K(u)\,du = 1$,
- $\int_{\mathbb{R}} u^2 K(u)\,du < \infty$.

Определение 1.4 (Ядерная оценка плотности)**.** Ядерная оценка плотности $p(x)$ с ядром K имеет вид:

$$\hat{p}(x) = \frac{1}{nh}\sum_{k=1}^{n} K\left(\frac{x - \mathbf{X}_k}{h}\right),$$

где h – сглаживающий параметр (ширина окна).

1.2. Постановка задачи

Рассмотрим $(\mathbf{X}, \mathbf{Y}) \sim P$. Пусть $(\mathbf{X}_k, \mathbf{Y}_k)_{k=1}^n$ — выборка объема n из распределения P, заданного на $\mathbb{R}^2$. Рассматривается регрессия случайной величины $\mathbf{X}$ на случайную величину $\mathbf{Y}$ — некоторая функция, задаваемая распределением P. В частности, такой функцией являются функция условного матожидания и рассматриваемая в работе функция условной медианы $\mathbf{Y}$ при условии $\mathbf{X} = x$:

$$\mathbb{M}[\mathbf{Y} \mid \mathbf{X} = x] = \arg\min_a \mathbb{E}[|\mathbf{Y} - a| \mid \mathbf{X} = x].$$

Функция регрессии не обязана быть непрерывной. Это приводит к задаче статистической проверки гипотезы о непрерывности (или разрывности) функции регрессии в некоторой точке x. Не умаляя общности, достаточно рассматривать эту задачу в точке $x = 0$.

Сформулируем задачу проверки гипотезы непрерывности функции регрессии, основанной на условной медиане, на языке сравнения параметров двух распределений. Определим распределение P_1 как распределение пар $(\mathbf{X}_k, \mathbf{Y}_k)$, где $\mathbf{X}_k \geq 0$, и распределение P_2 — пар $(-\mathbf{X}_k, \mathbf{Y}_k)$, где $\mathbf{X}_k < 0$. Таким образом рассматривается смесь двух распределений P_1 и P_2. Обозначим как $\mathcal{P} = \{\lambda P_1 + (1-\lambda)P_2 \mid \lambda \in [0,1]\}$ множество всех возможных смесей этих двух распределений. Пусть в выборке $(\mathbf{X}_k, \mathbf{Y}_k)_{k=1}^n$ первые n_1 элементов взяты из распределения P_1 и остальные n_2 — из P_2, $n_1 + n_2 = n$. Заметим, что n_k вообще говоря случайны. Для удобства переопределим получившиеся выборки, а именно

$$(\mathbf{X}_{1i}, \mathbf{Y}_{1i}) = (\mathbf{X}_i, \mathbf{Y}_i),\ i = 1, \ldots, n_1$$

— выборка из P_1 и

$$(\mathbf{X}_{2i}, \mathbf{Y}_{2i}) = (-\mathbf{X}_{n_1+i}, \mathbf{Y}_{n_1+i}),\ i = 1, \ldots, n_2$$

— выборка из P_2. Пусть $\theta(P_k)$ — значение условной медианы в нуле для распределения P_k. Тогда, исходная задача сводится к задаче проверки гипотезы о равенстве парамет-

ров:

$$\mathbb{H}_0 : \theta(P_1) = \theta(P_2).$$

Для проверки этой гипотезы необходимо построить критерий.

1.3. Построение критерия

Для построения критерия необходимо привести статистику, на основе которой он строится. В первую очередь, необходимо оценить параметры P_k по выборке. Одним из подходов оценивания является непараметрический подход с использованием ядерного сглаживания (ядерное оценивание). Пусть K — ядро и h_n — ширина окна ядра, зависящая от объёма исходной выборки. Тогда с использованием этого подхода получается оценка условной медианы в нуле $\theta(P_k)$ для распределения P_k по выборке объёма n_k и она имеет вид:

$$\hat{\theta}_k = \arg\min_a \sum_{i=1}^{n_k} |\mathbf{Y}_{ki} - a| K\left(\frac{\mathbf{X}_{ki}}{h_n}\right).$$

С использованием таких оценок построена статистика T_n, имеющая вид:

$$T_n = \sqrt{nh_n}(\hat{\theta}_1 - \hat{\theta}_2).$$

Для построения критерия необходимо изучить свойства данной статистики. Данная статистика входит в класс статистик, для которых в работе [5] получен результат об их предельном распределении. В частности, необходимо существование функции $\psi_n(\cdot\,; P)$, которая обеспечивает линейную аппроксимацию оценок $\hat{\theta}_k$, и $\xi^2(P)$, которая определяет предельную дисперсию статистики, а также числа $\lambda \in (0,1) : \frac{n_1}{n} \xrightarrow{p} \lambda$. Тогда статистика, которая удовлетворяет этим требованиям, обладает асимптотическим нормальным распределением с нулевым средним и дисперсией σ^2:

$$\sigma^2 = \xi^2(P_1)/\lambda + \xi^2(P_2)/(1-\lambda).$$

Детали представлены в приложении В.

Конкретно для задачи с условной медианой в рамках данной работы были получены утверждение и следствие, описывающие асимптотические свойства статистики T_n:

Утверждение 1.1. Пусть

1. $n \to \infty$, $n_1/n \to \lambda$, $h_n \to 0$, $nh_n \to \infty$ и $h_n = o(n^{-\frac{1}{3}})$;

2. K ограничено, симметрично, с конечным вторым моментом;

3. Распределение $\mathbf{X}$, где $(\mathbf{X}, \mathbf{Y}) \sim P \in \mathcal{P}$, имеет плотность $f_{\mathbf{X}}$, ограниченную, отделенную от нуля, дифференцируемую дважды всюду, за исключением точки $x = 0$, причем ее производные ограничены;

4. Распределение $\mathbf{Y}$ при условии $\mathbf{X}$ имеет плотность $f_{Y|X}(y \mid x)$, ограниченную, отделенную от нуля по y при условии x, дифференцируемую по обоим аргументам всюду за исключением $x = 0$, причем соответствующие частные производные ограничены;

5. Распределение $\mathbf{Y}$ при условии $\mathbf{X}$ имеет функцию распределения $F_{Y|X}(y \mid x)$, дважды дифференцируемую по x всюду за исключением $x = 0$, причем соответствующие частные производные ограничены.

Обозначим $\mathbf{V} = (\mathbf{X}, \mathbf{Y})$. Тогда предположения B.1 выполняются для оценки $\hat{\theta}_n$ разрыва медианы $\theta(P)$ вместе с функциями

$$\psi_n(\mathbf{V}, P) = -K\left(\frac{\mathbf{X}}{h_n}\right) \frac{\mathbb{I}\{\mathbf{Y} < \theta(P)\} - F_{Y|X}(\theta(P) \mid \mathbf{X}; P)}{f_{Y|X}(\theta(P) \mid 0^+; P) f_X(0^+; P)\sqrt{h_n}/2};$$

$$\xi^2(P) = \int_0^\infty K^2(u)\, du \frac{1}{f^2_{Y|X}(\theta(P) \mid 0^+; P) f_X(0^+; P)}.$$

Здесь $f_X(0^+; P)$ — плотность распределения с.в. $\mathbf{X}$ относительно совместного распределения P пары $(\mathbf{X}, \mathbf{Y})$, $f_{Y|X}(y \mid 0^+; P)$ и $F_{Y|X}(y \mid 0^+; P)$ — плотность и функция распределения условного распределения $\mathbf{Y}$ при условии $\mathbf{X} = 0^+$ относительно P. 0^+ — сколь угодно близкое к нулю положительное число.

Доказательство этого утверждения представлено в приложении A.1.

Следствие 1.1. Статистика $T_n = \sqrt{nh_n}(\hat{\theta}_1 - \hat{\theta}_2)$, построенная на оценке условной медианы, асимптотически нормальна:

$$T_n - \sqrt{nh_n}(\theta(P_1) - \theta(P_2)) \xrightarrow{d} N(0, \sigma^2),$$

где

$$\sigma^2 = \int_0^\infty K^2(u)\, du \Big(\frac{1}{\lambda f^2_{Y_1|X_1}(\theta(P_1) \mid 0^+) f_{X_1}(0^+)} + \\ + \frac{1}{(1-\lambda) f^2_{Y_2|X_2}(\theta(P_2) \mid 0^+) f_{X_2}(0^+)} \Big).$$

Еще одно следствие получается из доказательства уверждения 1.1 при ослаблении условия на h_n:

Следствие 1.2. Пусть $h_n = n^{-\frac{1}{3}}$. Тогда асимптотическая нормальность статистики T_n сохраняется с той же предельной дисперсией, но появляется асимптотическое смещение B:

$$T_n - \sqrt{nh_n}(\theta(P_1) - \theta(P_2)) \xrightarrow{d} N(B, \sigma^2).$$

Исправленная статистика $T_n^{adj} = \sqrt{nh_n}(\hat{\theta}_1 - \hat{\theta}_2 - h_n(B(P_1) - B(P_2)))$, где

$$B(P_k) = -2\int_0^\infty uK(u)\,du \frac{\nabla_x F_{Y_k|X_k}(\theta(P_k) \mid 0^+)}{f_{Y_k|X_k}(\theta(P_k) \mid 0^+)},$$

обладает центрированным асимптотическим нормальным распределением:

$$T_n^{adj} - \sqrt{nh_n}(\theta(P_1) - \theta(P_2)) \xrightarrow{d} N(0, \sigma^2).$$

Здесь ∇_x обозначает операцию взятия производной по аргументу x.

Величины $B(P_k)$ можно брать как из соображений априорной информации об исследуемых распределениях, так и можно заменять их с помощью их состоятельных оценок $\widehat{B}(P_k)$, что не отразится на предельной дисперсии. Эти оценки можно строить с тем же подходом ядерного оценивания.

Перейдем непосредственно к построению критерия. Выберем уровень значимости $\alpha \in (0, 1)$. Тогда гипотеза $\mathbb{H}_0$ о непрерывности условной медианы в нуле отвергается, если $|T_n| > \Phi^{-1}_{\sigma^2}(1 - \alpha/2)$ (или то же с T_n^{adj}), где $\Phi_{\sigma^2}(x)$ — функция распределения нормального распределения $N(0, \sigma^2)$ и σ^2 взято из следствия 1.1. Так как критерий построен по статистике, то говорят, что T_n (или T_n^{adj}) является статистикой критерия проверки гипотезы непрерывности функции регрессии, построенный на основе (оценки) условной медианы в нуле.

Проверку непрерывности функции регрессии в нуле можно свести как к проверке гипотезы о непрерывности функции условной медианы в нуле, так и к проверке гипотезы о непрерывности функции условного матожидания в нуле, критерий для которой строится аналогично: в работе [5] для аналогичной статистики T_n, построенной на основе оценки условного матожидания в нуле

$$\hat{\theta}_k = \frac{\sum_{i=1}^{n_k} \mathbf{Y}_{ki} K\left(\frac{\mathbf{X}_{ki}}{h_n}\right)}{\sum_{i=1}^{n_k} K\left(\frac{\mathbf{X}_{ki}}{h_n}\right)}$$

получен результат об асимптотическом нормальном распределении с предельной дисперсией σ^2

$$\sigma^2 = 4\int_0^\infty K^2(u)\,du\left(\frac{\mathbb{D}(\mathbf{Y}_1 \mid \mathbf{X}_1 = 0^+)}{\lambda f_{\mathbf{X}_1}(0^+)} + \frac{\mathbb{D}(\mathbf{Y}_2 \mid \mathbf{X}_2 = 0^+)}{(1-\lambda) f_{\mathbf{X}_2}(0^+)}\right),$$

где $\mathbb{D}(\mathbf{Y}_i \mid \mathbf{X}_i = 0^+)$ — условная дисперсия второй компоненты распределения P_i при условии равенства первой сколь угодно близкой справа к нулю точке. Два таких критерия подвергнуты сравнению в разделе 1.6.

1.4. Моделирование для проверки результатов

1.4.1. Описание модели

Проведем моделирование для демонстрации работы полученных результатов о статистике критерия проверки гипотезы непрерывности условной медианы в нуле.

Рассмотрим следующую модель: Пусть $\mathbf{X} \sim (1-\lambda)N^+(0,1) + \lambda N^-(0,1)$, $\lambda \in (0,1)$ — смесь односторонних нормальных распределений; $\mathbf{Y} = f(\mathbf{X}) + \varepsilon$, где $f(x) = x$ и $\varepsilon \sim N(0,\sigma^2)$ с $\sigma^2 = 1$, если $\mathbf{X} \geq 0$, иначе $\sigma^2 = s^2$. Рассматриваются конфигурации $\lambda \in \{0.5, 0.25\}$ и $s^2 \in \{1, 5\}$.

В такой модели условная медиана как функция от x совпадает с $f(x)$:

$$\mathbb{E}[|\mathbf{Y}_i - a| \mid \mathbf{X}_i = x] = \int_{\mathbb{R}} |y-a| f_{Y_i|X_i}(y \mid x)\,dy = \frac{1}{\sqrt{2\pi}\sigma_i}\int_{\mathbb{R}} |y-a| e^{\frac{(y-f(x))^2}{2\sigma_i^2}}\,dy,$$

$$\arg\min_a \mathbb{E}[|\mathbf{Y}_i - a| \mid \mathbf{X}_i = x] = f(x),$$

поэтому проверять непрерывность $f(x)$ равносильно проверке непрерывности условной медианы.

В качестве ядра выбрано ядро Епанечникова

$$K(x) = 0.75(1-x^2)\mathbb{I}\{|x| \leq 1\},$$

так как оно оптимально для нашей задачи среди неотрицательных ядер. Оптимальность такого ядра рассмотрена в [14]. В качестве ширины окна ядра h_n выбрана такая, что она минимизирует $MISE(h)$ — Mean Integrated Square Error — определяемая для ядерной оценки $\hat{f}_h(x)$ плотности $f(x)$ следующим образом:

$$MISE(h) = \mathbb{E}\left(\int_{-\infty}^{\infty} (\hat{f}_h(x) - f(x))^2\,dx\right).$$

Такой ширины окна будет достаточно для наших задач.

К сожалению, прямое применение такого ядра и такой ширины окна приводит к неточностям в вычислениях. Это связано с

- проявлением краевых эффектов при ядерном оценивании: ядро при приближении к границе носителя плотности распределения захватывает все меньше наблюдений, что приводит к смещенности оценки плотности в этой области;
- ширина окна, минимизирующая $MISE$, не удовлетворяет требованию $h_n = O(n^{-1/3})$, что требуется в утверждении 1.1.

В качестве решения этих проблем предлагаются отзеркаливание ядерной оценки и исправление ширины окна. В приложении Б представлены описания этих методов с применением к выбранной модели.

1.4.2. Моделирование

В заданной модели проведено моделирование статистики T_n^{adj} с различными λ и s^2 с объемом выборки $n \in \{2500, 5000, 10000\}$ и числом повторений эксперимента $m = 1000$ для проверки соблюдения асимптотической нормальной распределенности и установления дисперсии на уровне предельной.

На рисунках 1.1a–1.1d представлены графики поведения оценки предельной дисперсии статистики T_n^{adj} при различных объемах выборки в случае подстановки теоретического значения поправки смещения. Можно увидеть, что в случаях (a) и (c) оценка установилась достаточно быстро, в то время как в случаях (b) и (d) должно потребоваться больше индивидов в выборке, чем $n = 10000$, но точное предельное значение остаётся в пределах 95% доверительного интервала. Все это согласуется с теоретическими результатами.

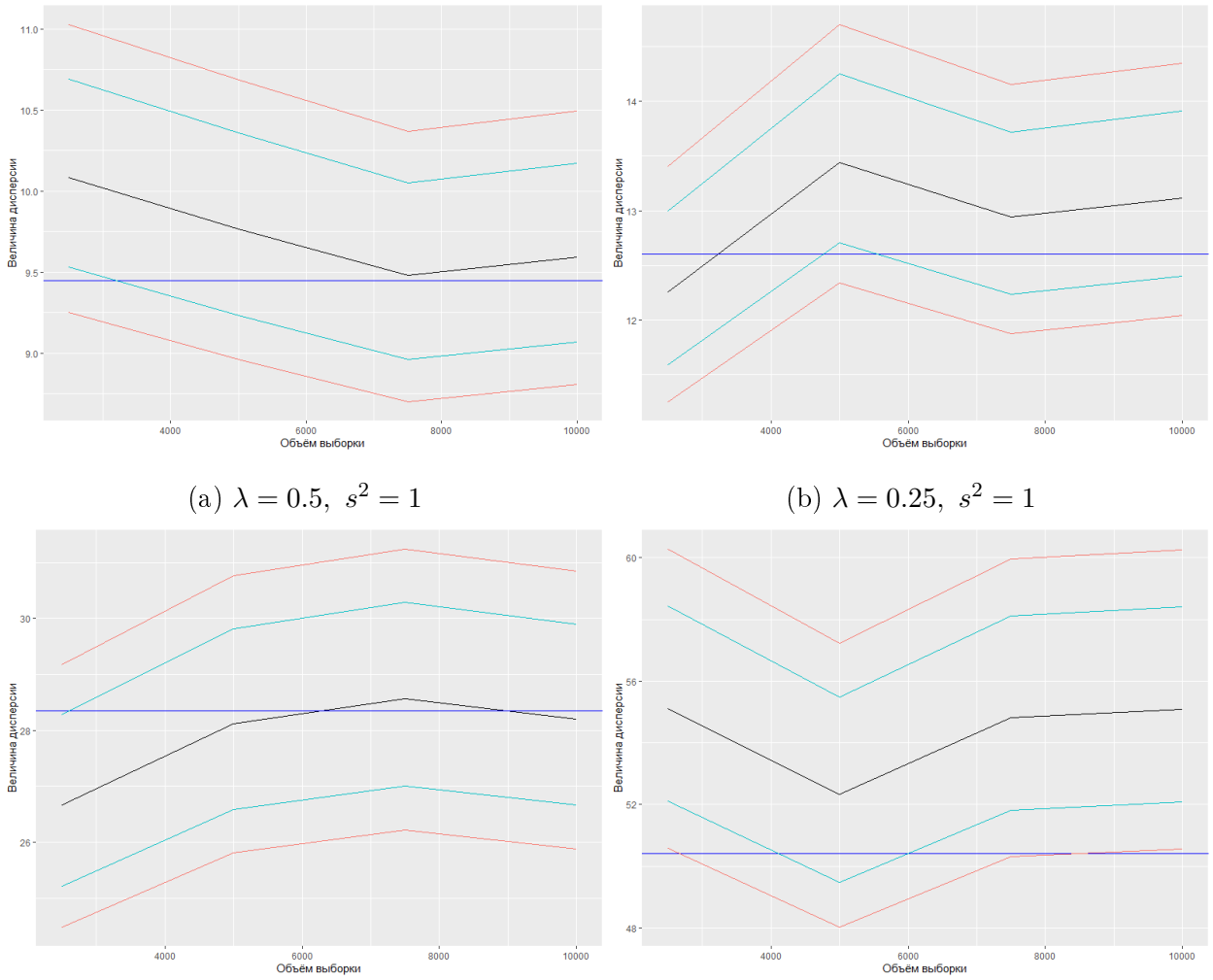
На рисунках 1.2a–1.2d представлены графики поведения оценки предельной дисперсии статистики T_n^{adj} при различных объемах выборки в случае подстановки состоятельной оценки поправки смещения, построенной с помощью ядерных оценок составляющих этой поправки (производной функции условного распределения $\nabla_x F_{Y_k|X_k}$ и плотности условного распределения $f_{Y_k|X_k}$). Можно увидеть, что оценка дисперсии здесь ведет себя так же, как в случае с точным значением поправки смещения. Здесь так же результаты согласуются с теоретическими.

На рисунках 1.3a–1.3d представлены гистограммы значений статистики T_n^{adj} при объеме выборки $n = 10000$ и плотности соответствующих предельных распределений. Как можно увидеть, гистограммы вполне соответстуют предельным распределениям, что согласуется с теоретическими результатами.

В таблице 1.1 представлены значения оценки вероятности ошибки первого рода α_I при проверке гипотезы $\mathbb{H}_0$. Как видно, оценки сближаются с уровнем значимости $\alpha = 0.05$ с ростом объема выборки, что отвечает асимптотичности построенного критерия.

σ_1^2	σ_2^2	q	$n = 2500$	$n = 5000$	$n = 10000$
1	1	0.5	0.063	0.062	0.052
1	1	0.25	0.041	0.065	0.046
5	1	0.5	0.039	0.048	0.044

Таблица 1.1: Оценка вероятности ошибки первого рода при разных объемах выборки.



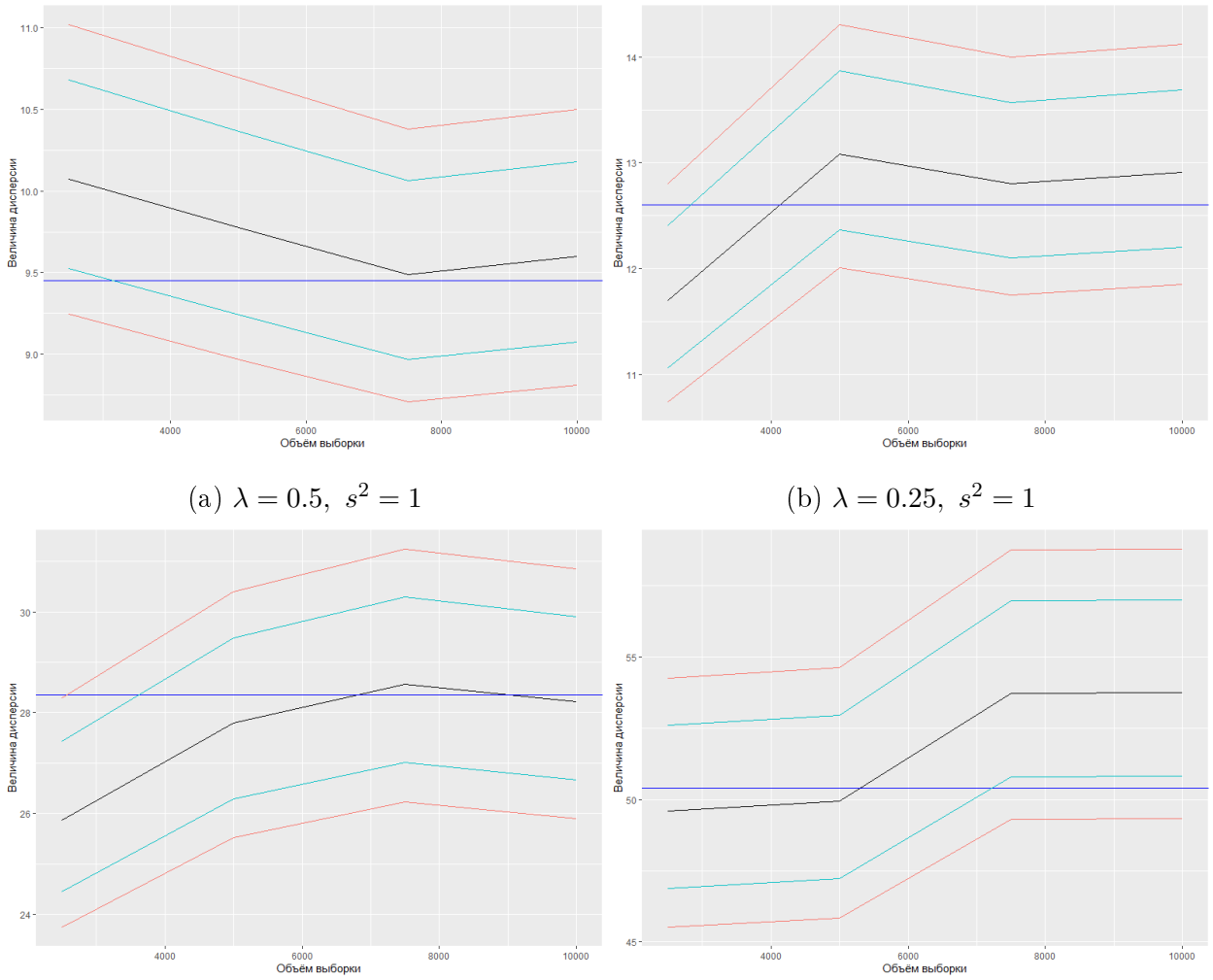
(a) $\lambda = 0.5,\ s^2 = 1$

(b) $\lambda = 0.25,\ s^2 = 1$

(c) $\lambda = 0.5,\ s^2 = 5$

(d) $\lambda = 0.25,\ s^2 = 5$

Рис. 1.1: Поведение оценки дисперсии статистики T_n^{adj} с ростом объема выборки в случае подстановки точного значения смещения $B(P_1) - B(P_2)$. Синяя горизонтальная прямая соответствует предельной дисперсиию. Линия черного цвета соответствует оценкам предельной дисперсии при разных объемах выборки. Бирюзовые и красные линии показывают соответственно 80% и 95% асимптотические доверительные интервалы для дисперсии.



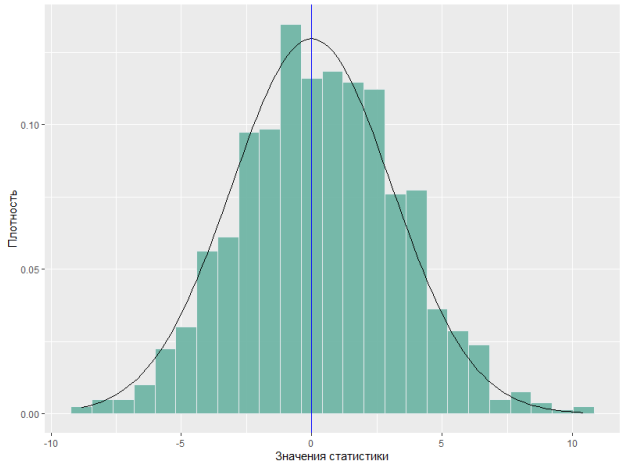
(a) $\lambda = 0.5,\ s^2 = 1$

(b) $\lambda = 0.25,\ s^2 = 1$

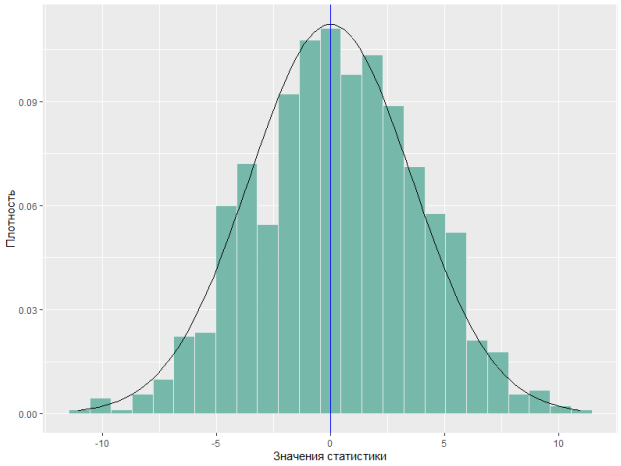
(c) $\lambda = 0.5,\ s^2 = 5$

(d) $\lambda = 0.25,\ s^2 = 5$

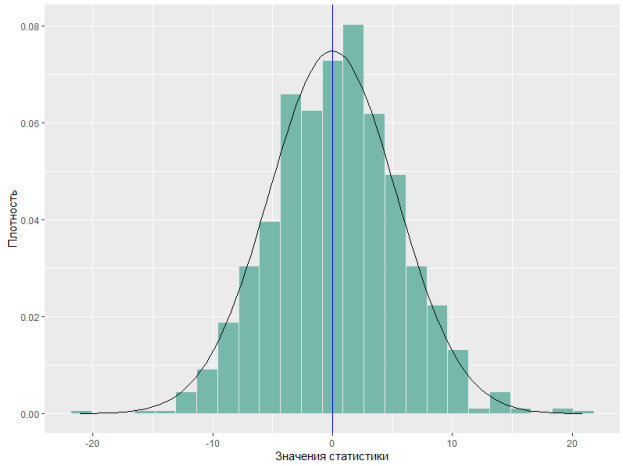
Рис. 1.2: Поведение оценки дисперсии статистики T_n^{adj} с ростом объема выборки в случае подстановки состоятельной оценки смещения $B(P_1) - B(P_2)$. Синяя горизонтальная прямая соответствует предельной дисперсиию. Линия черного цвета соответствует оценкам предельной дисперсии при разных объемах выборки. Бирюзовые и красные линии показывают соответственно 80% и 95% асимптотические доверительные интервалы для дисперсии.



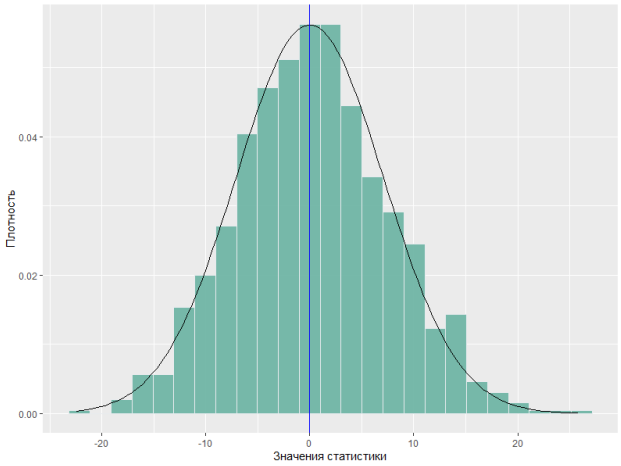
(a) $\lambda = 0.5,\ s^2 = 1$



(b) $\lambda = 0.25,\ s^2 = 1$



(c) $\lambda = 0.5,\ s^2 = 5$



(d) $\lambda = 0.25,\ s^2 = 5$

Рис. 1.3: Гистограммы значений статистики T_n^{adj} при объеме выборки $n = 10000$ и плотность предельного распределения. Синей вертикальной прямой показано предельное значение среднего. Здесь использована оцененная поправка смещения.

1.5. Перестановочный вариант критерия

Результаты моделирования 1.4.2, а именно графики 1.1a–1.1d и 1.2a–1.2d, показывают, что критерий проверки гипотезы о разрыве условной медианы, использующий критические значения асимптотического распределения статистики T_n, требует достаточно большого объема выборки. Для более малых объемов выборки можно предложить перестановочные критерий, использующий статистику T_n.

1.5.1. Построение критерия

Построим перестановочный критерий на основе статистики T_n критерия проверки гипотезы о разрыве условной медианы в нуле по схеме, представленной в [15] и [5]. Описание схемы представлено в приложении В. Вместе с ним построим и стьюдентизированный перестановочный критерий, использующий критические значения от значений статистики $S_n = T_n/\widehat{\sigma}_n$ при переставляемой выборке, где $\widehat{\sigma}_n$ строится по $\widehat{\xi_k^2}$:

$$\widehat{\sigma}_n^2 = \frac{n}{n_1}\widehat{\xi_1^2} + \frac{n}{n_2}\widehat{\xi_2^2}.$$

$\widehat{\xi_k^2}$ являются состоятельными оценками ξ_k^2, представленных в утверждении 1.1. Например, их можно строить с помощью ядерного оценивания.

В [5] показано, что нестьюдентизированный вариант критерия является точным в конечных выборках, если выполнена совершенная нулевая гипотеза[1] $P_1 = P_2$, в противном случае он перестаёт быть корректным, так как ошибка первого рода такого критерия не контролируется. В таком случае нужно использовать стьюдентизированный перестановочный критерий со статистикой S_n. Он в свою очередь контролирует ошибку первого рода асимптотически. Подробности представлены в приложении В. Конкретно для случая условной медианы эти эффекты продемонстрированы в моделировании.

Очевидно, пересчет статистики T_n для всех $n!$ перестановок — затратная по времени задача. В [15] предлагается пересчет статистики для произвольного подмножества множества всех перестановок объема N, что, естественно, сделает контроль ошибки первого рода асимптотическим для этого критерия при росте мощности подмножества перестановок.

[1] В англоязычной литературе *sharp null hypothesis*.

1.5.2. Моделирование

Предлагается провести моделирование, в рамках которого при малом объеме выборки сравниваются два перестановочных критерия со статистиками T_n^{adj} и S_n^{adj}[2] не только между собой, но и против классического t-критерия, отвергающего нулевую гипотезу $\mathbb{H}_0$, когда $|S_n^{adj}| > \Phi^{-1}(1-\alpha/2)$, где $\Phi(x)$ — функция распределения стандартного нормального распределения $N(0,1)$.

Пусть $\mathbf{X} \sim \lambda U(-1,0) + (1-\lambda)U(0,1)$, $\lambda \in (0,1)$ — смесь равномерных распределений; $\mathbf{Y} = f(\mathbf{X}) + \varepsilon$, где $f(x) = x$ и $\varepsilon \sim N(0,\sigma^2)$ с $\sigma^2 = 1$, если $\mathbf{X} \geq 0$, иначе $\sigma^2 = s^2$. Рассматриваются конфигурации $\lambda \in \{0.5, 0.3\}$ и $s^2 \in \{1, 5\}$.

Проверяется гипотеза о непрерывности $f(x)$ в нуле. В такой модели условная медиана как функция от x совпадает с $f(x)$ (по аналогичным соображениям), поэтому проверка непрерывности $f(x)$ равносильна проверке непрерывности условной медианы.

Проведено моделирование статистик T_n^{adj}, S_n^{adj}, посчитанные на $N = 500$ случайных перестановках с различными λ и s^2 с объемом выборки $n = 200$, числом повторений эксперимента $m = 500$. Рассматриваются три вышеупомянутых критерия, при этом перебираются различные значения ширины окна h: 0.1, 0.3 и 0.5 для выбора из них оптимального. В качестве ядра выбрано ядро Епанечникова.

Результаты моделирования представлены в таблице 1.2, где содержатся оценки вероятности ошибки первого рода. Оптимальным значением ширины окна оказалось $h = 0.1$. Как видно, t-критерий справляется плохо, что, естественно, следует из малого объема выборки и асимптотичности этого критерия. В то же время, при $s^2 = 1$ можно утверждать, что два перестановочных критерия справляются одинаково с проверкой гипотезы, но если $s^2 \neq 1$,то стьюдентизированный перестановочный критерий работает лучше обычного, так как в этом случае не выполняется совершенная нулевая гипотеза.

[2] То есть поправленные на смещение. В данном случае рассматриваются теоретические поправки на смещение.

		$h = 0.1$			$h = 0.3$			$h = 0.5$		
s^2	λ	$\widehat{\alpha}_I(\ T_n^{adj})$	$\widehat{\alpha}_I(S_n^{adj})$	$\widehat{\alpha}_I(t)$	$\widehat{\alpha}_I(T_n^{adj})$	$\widehat{\alpha}_I(S_n^{adj})$	$\widehat{\alpha}_I(t)$	$\widehat{\alpha}_I(T_n^{adj})$	$\widehat{\alpha}_I(S_n^{adj})$	$\widehat{\alpha}_I(t)$
1	0.5	0.062	0.053	0.030	0.102	0.068	0.030	0.280	0.284	0.150
5	0.5	0.110	0.039	0.018	0.154	0.054	0.014	0.260	0.138	0.020
1	0.3	0.058	0.032	0.016	0.098	0.080	0.030	0.262	0.244	0.130
5	0.3	0.189	0.072	0.032	0.268	0.090	0.026	0.344	0.164	0.026

Таблица 1.2: Значения оценок вероятности ошибки первого рода для перестановочных критериев со статистиками T_n^{adj}, S_n^{adj} и t-критерия при различных конфигурациях модели и значениях ширины окна h_n.

1.6. Сравнение критериев

1.6.1. Постановка сравнения

Рассмотрим задачу проверки гипотезы о разрывности функции регрессии $m(x)$ в точке $x = 0$. Проверяется гипотеза:

$$\mathbb{H}_0 : \ m(0^+) = m(0^-).$$

Эту гипотезу можно проверять как с помощью критерия, построенного по статистике T_n с использованием оценки условного матожидания, предложенного в [5], так и по статистике, использующей конструкцию оценки условной медианы. Имеет смысл сравнить два критерия в зависимости от данных, на которых гипотеза проверяется.

В качестве способа сравнения выбрано сравнение критериев в постановке задачи о робастности их статистик, рассматриваемой в [16]. Статистика T_n называется робастной, если она выдерживает наличие шума в данных. Одним из критериев устойчивости статистики к шуму является поведение ее дисперсии: статистика тем устойчивее, чем меньше ее дисперсия меняется при появлении шума. В рамках проверки гипотез это бы означало, что статистика будет реже попадать в теоретическую критическую область и реже отвергать верную нулевую гипотезу. Сравнение было проведено на модельных данных в двух случаях: с нормальным шумом, но с дисперсией, отклоняющейся от модели, и с шумом, имеющим распределение Коши, которое имеет тяжелые хвосты.

1.6.2. Моделирование

Рассмотрим следующую модель: Пусть $\mathbf{X} \sim N(0,1)$, $\mathbf{Y} = f(\mathbf{X}) + \varepsilon$, где $f(x) = x$ и $\varepsilon \sim N(0, \sigma^2)$ с $\sigma^2 = 1$.

В таблице 1.3 представлены значения оценок дисперсий статистики T_n для двух рассматриваемых конструкций при различных объемах выборки при повторе эксперимента $m = 1000$ раз. Видно, что при верности этой модели дисперсия статистики с условным матожиданием меньше, чем в случае условной медианы. Это показывает эффективность статистики, использующей условное матожидание, что согласуется с обычными средним и медианой.

Допустим, что наблюдения зашумлены: пусть $\mathbf{Y} = f(\mathbf{X}) + \varepsilon$, с вероятностью $(1-\alpha)$ $\varepsilon \sim N(0,1)$ и с вероятностью α $\varepsilon \sim N(0, s^2)$, где $s^2 > 0$ — дисперсия шума. Для этой ситуации проведено моделирование при различном уровне шума α и величины дисперсии шума s^2 при объеме выборки $n = 10000$ с числом повторов эксперимента $m = 1000$. Его результаты представлены в таблице 1.4. Из нее видно, что при повышении уровня шума и при повышении величины дисперсии шума дисперсия статистики, использующей оценку условного матожидания, меняется значительно, в то время как дисперсия статистики с оценкой условной медианы остается той же или меняется меньше в сравнении со случаем отсутствия шума.

Аналогично рассмотрим ситуацию с шумом Коши: с вероятностью α ε имеет распределение Коши. Результаты моделирования для данного случая представлены в таблице 1.5 с числом наблюдений $n = 10000$ с числом повторов эксперимента $m = 1000$. Из нее напрямую видно, что статистика, использующая конструкцию оценки условной медианы, робастна к шумным наблюдениям, что, конечно, связано с тем, что условное математическое ожидание не существует у распределения Коши.

Рассмотренные случаи показывают, что конструкция критерия с использованием оценки условной медианы обладает преимуществом при проверке гипотезы в случае, когда нет уверенности в правильности модели.

n	2500	5000	10000
$\hat{\sigma}^2_{n,\,mean}$	6.384	6.268	6.010
$\hat{\sigma}^2_{n,\,median}$	9.620	9.512	8.962

Таблица 1.3: Значения оценок дисперсии двух статистик при различных объемах выборки при отсутствии шума

s^2	2		5		10	
α	$\widehat{\sigma^2}_{n,\,mean}$	$\widehat{\sigma^2}_{n,\,med}$	$\widehat{\sigma^2}_{n,\,mean}$	$\widehat{\sigma^2}_{n,\,med}$	$\widehat{\sigma^2}_{n,\,mean}$	$\widehat{\sigma^2}_{n,\,med}$
0	5.894	8.962	5.894	8.962	5.894	8.962
0.01	6.079	8.988	6.282	9.035	6.619	9.075
0.05	6.209	9.704	7.126	9.954	8.628	10.311
0.1	6.895	10.256	8.634	10.868	11.533	11.152
0.2	6.865	10.672	10.338	12.066	16.283	12.753
0.5	9.977	13.497	19.592	19.395	35.530	23.426

Таблица 1.4: Значения оценок дисперсии двух статистик при нормальном шуме, $n = 10000$

α	$\widehat{\sigma^2}_{n,\,mean}$	$\widehat{\sigma^2}_{n,\,med}$
0	6.010	8.962
0.01	36.577	9.582
0.05	1861.970	9.736
0.1	10388.807	10.423
0.2	26640.35	10.415
0.5	307110.19	11.139

Таблица 1.5: Значения оценок дисперсии двух статистик при шуме Коши, $n = 10000$

Глава 2

Критерий проверки разрывности плотности интенсивности пуассоновского процесса

2.1. Необходимые определения

Определение 2.1 (Неоднородный пуассоновский процесс)**.** Пусть даны вероятностное пространство $(\Omega, \mathcal{F}, \mathbb{P})$ с σ-алгеброй $\mathcal{F}$ и полное сепарабельное метрическое пространство $(\mathcal{X}, \rho)$ вместе с борелевской σ-алгеброй $\mathcal{B} = \mathcal{B}(\mathcal{X})$. Пусть $\mathcal{M}$ — пространство σ-конечных мер на $(\mathcal{X}, \mathcal{B})$ и $\mathcal{M}_0$ — его подпространство целочисленных мер вида

$$\mathbf{X} = \sum_i \delta_{x_i}, \ x_i \in \mathcal{X}.$$

Пусть $\Lambda \in \mathcal{M}$. Тогда случайный элемент $\mathbf{X}$, определенный на $(\Omega, \mathcal{F}, \mathbb{P})$ и имеющий значения в $\mathcal{M}_0$ называется пуассоновским процессом с мерой интенсивности (интенсивностью) Λ, если

- для любого конечного набора дизъюнктных множеств $B_1, \ldots, B_m \in \mathcal{B}$ случайные величины $\mathbf{X}(B_1), \ldots, \mathbf{X}(B_m)$ независимы,
- $\mathbf{X}(B)$ имеет распределение Пуассона с параметром $\Lambda(B)$ для любого $B \in \mathcal{B}$ с конечной мерой $\Lambda(B)$.

Определение 2.2. Центрированный пуассоновский процесс π относительно пуассоновского процесса $\mathbf{X}$ с интенсивностью Λ определяется следующим образом:

$$\forall\, B \in \mathcal{B} \ \ \pi(B) = \mathbf{X}(B) - \Lambda(B).$$

Определение 2.3 (Стохастический интеграл)**.** Пусть $\mathbf{X}$ — процесс пуассона с интенсивностью Λ и f — измеримая (относительно Λ) ограниченная функция с конечным носителем. Тогда стохастический интеграл функции f относительно X задается следующим образом:

$$I(f) = \int_{\mathcal{X}} f(x)\, \mathbf{X}(dx) = \sum_i f(x_i).$$

В то же время определен интеграл относительно центрированного процесса π:

$$I^*(f) = \int_{\mathcal{X}} f(x)\, \pi(dx) = I(f) - \int_{\mathcal{X}} f(x)\, \Lambda(dx).$$

Свойства интеграла [17]

Пусть f суммируема с модулем относительно Λ. Тогда справедливы следующие факты:

1. $I(f)$ и $I^*(f)$ существуют,
2. $\mathbb{E}I(f) = \int_{\mathcal{X}} f(x)\,\Lambda(dx)$, $\mathbb{E}I^*(f) = 0$.

Определение 2.4 (Плотность интенсивности)**.** Пусть Λ_1 и Λ_2 — две конечные меры на $\mathcal{X}$, то есть $\Lambda_i(\mathcal{X}) < \infty,\ i = 1, 2$. Пусть Λ_2 абсолютно непрерывна относительно Λ_1. Тогда производная Радона-Никодима $\lambda(x) = \frac{d\Lambda_2}{d\Lambda_1}(x)$ называется плотностью интенсивности пуассоновского процесса с интенсивностью Λ_2 относительно Λ_1.

2.2. Построение критерия

Рассматривается пуассоновский процесс $\mathbf{X} = \xi_t$ на отрезке времени $\mathcal{X} = [-1, 1]$ с плотностью интенсивности $\lambda(t)$. По определению это означает, что для $[a, b] \subset [-1, 1]$ $\mathbf{X}([a, b])$ имеет распределение Пуассона с параметром $\int_a^b \lambda(t)\,dt$. Из этого процесса отбирается n реализаций $\mathbf{X}_1, \ldots, \mathbf{X}_n$, для которых известен набор времен событий:

$$\begin{array}{lcccc} \mathbf{X}_1: & T_{11}, & T_{12}, & \ldots & T_{1N_1}, \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}_n: & T_{n1}, & T_{n2}, & \ldots & T_{nN_n}, \end{array}$$

где $N_1, \ldots, N_n$ имеют распределение Пуассона с параметром $\int_{-1}^1 \lambda(t)\,dt$ и определяют число произошедших событий за рассматриваемое время.

Необходимо по выборке проверить гипотезу о непрерывности функции плотности интенсивности в точке $\tau = 0$ (не умаляя общности):

$$\mathbb{H}_0 : \lambda(0^+) = \lambda(0^-),$$

где 0^+ и 0^- обозначают сколь угодно близкие к нулю точки справа и слева соответственно.

Предлагается оценивать разрыв следующим образом: выборка из реализаций разбивается на две объемов $n_1 = \lceil \frac{n}{2} \rceil$ и $n_2 = n - n_1$, по каждой оценивается половина

разрыва плотности интенсивности так, как если бы мы брали выборки из двух процессов ξ_t^1 и ξ_t^2, из следующих тривиальных соображений:

$$\lambda(0^+) - \lambda(0^-) = \frac{1}{2}\left(\lambda(0^+) - \lambda(0^-)\right) + \frac{1}{2}\left(\lambda(0^+) - \lambda(0^-)\right).$$

Для удобства элементы выборки $\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}$ из первой половины обозначим следующим образом:

$$\mathbf{X}_{1i} = \mathbf{X}_i,\ i = 1, \ldots, n_1,$$

а из второй:

$$\mathbf{X}_{2i} = \mathbf{X}_{n_1+i},\ i = 1, \ldots, n_2.$$

Тогда оценка разрыва равнялась бы сумме оценок половины, которые имеют вид

$$\widehat{\theta}_k = \frac{1}{2}\left(\widehat{\lambda}(0^+) - \widehat{\lambda}(0^-)\right) = \frac{1}{n_k h}\sum_{i=1}^{n_k}\int_{-1}^{1} K\left(\frac{y}{h}\right)\operatorname{sign}(y)\,\mathbf{X}_{ki}(dy).$$

Дополнительно, выборку из второго набора развернем во времени для того, чтобы записать оценку разрыва как разность:

$$\widehat{\theta} = \widehat{\theta}_1 - \widehat{\theta}_2.$$

Итого, рассматривается статистика T_n вида $T_n = \sqrt{nh_n}(\widehat{\theta}_1 - \widehat{\theta}_2)$. Потребуем выполнение следующих предположений:

Предположение 2.1. Пусть

1. $n \to \infty,\ nh_n \to \infty,\ h_n = o(n^{-\frac{1}{3}})$,
2. Ядро K неотрицательно, ограничено, с конечным носителем $[-1, 1]$,
3. $\lambda(t)$ непрерывна и дважды дифференцируема на $[-1, 1]$ всюду за исключением точки $\tau = 0$, ее производные ограничены.

Тогда для статистики T_n выполняется следующее утверждение, которое сформулировано в немного более общей постановке:

Утверждение 2.1. Рассмотрим два пуассоновских процесса ξ_t^1 и ξ_t^2 (где второй идет в обратном времени) на отрезке времени $[-1, 1]$ с плотностями интенсивности $\lambda_1(t)$ и $\lambda_2(t)$. Рассмотрим оценки $\widehat{\theta}_k$ половины разрыва плотности интенсивности $\theta(\xi_t^k) = \theta_k = \frac{1}{2}\left(\lambda_k(0^+) - \lambda_k(0^-)\right)$, которые построены по тому же принципу, что выше. Пусть

предположения верны для обоих процессов. Пусть $\mathbf{X}_{k1}, \ldots, \mathbf{X}_{kn_k}$, $k = 1, 2$ — выборка из реализаций этих двух процессов, причем $n_1 + n_2 = n$ и $n_1/n \to 0.5$. Тогда статистика $T_n = \sqrt{nh_n}(\widehat{\theta}_1 - \widehat{\theta}_2)$ имеет асимптотическое нормальное распределение

$$T_n - \sqrt{nh_n}(\theta_1 - \theta_2) \xrightarrow{d} N(0, \sigma^2)$$

с предельной дисперсией

$$\sigma^2 = 4 \int_0^1 K^2(u)\, du \left(\lambda(0^+) + \lambda(0^-)\right).$$

Доказательство утверждения 2.1 представлено в приложении А.2. Если ослабить условие на h_n, то из доказательства этого утверждения можно получить следующее следствие:

Следствие 2.1. Пусть $h_n = n^{-\frac{1}{3}}$. Тогда асимптотическая нормальность статистики T_n сохраняется с той же предельной дисперсией, но появляется асимптотическое смещение B:

$$T_n - \sqrt{nh_n}(\theta_1 - \theta_2) \xrightarrow{d} N(B, \sigma^2).$$

Исправленная статистика $T_n^{adj} = \sqrt{nh_n}(\hat{\theta}_1 - \hat{\theta}_2 - h_n(B_1 - B_2))$, где

$$B_k = \int_0^1 uK(u)\, du(\lambda_k'(0^+) + \lambda_k'(0^-)),$$

обладает центрированным асимптотическим нормальным распределением:

$$T_n^{adj} - \sqrt{nh_n}(\theta_1 - \theta_2) \xrightarrow{d} N(0, \sigma^2).$$

В качестве поправки на смещение можно использовать как ее точное значение, исходя из априорной информации, так и ее состоятельную оценку.

Собственно, критерий проверки гипотезы $\mathbb{H}_0$ строится как и в разделе 1.3. Выберем уровень значимости $\alpha \in (0, 1)$. Тогда гипотеза $\mathbb{H}_0$ о непрерывности условной медианы в нуле отвергается, если $|T_n| > \Phi^{-1}_{\sigma^2}(1 - \alpha/2)$ (или то же с T_n^{adj}), где $\Phi_{\sigma^2}(x)$ — функция распределения нормального распределения $N(0, \sigma^2)$ и σ^2 взято из утверждение 2.1. Так как критерий построен по статистике, то говорят, что T_n (или T_n^{adj}) является статистикой критерия проверки гипотезы непрерывности плотности интенсивности пуассоновского процесса в момент времени $t = 0$ (в нуле).

2.3. Моделирование

Для проверки правильности теоретических результатов, проведено моделирование построенной статистики.

Рассмотрим следующую модель: пусть плотность интенсивности процесса имеет вид:

$$\lambda(t) = \frac{t+1}{2},\ t \in [-1, 1].$$

Проверяется гипотеза $\mathbb{H}_0$ о непрерывности ее в точке 0. Очевидно, что $\mathbb{H}_0$ верна. Эта гипотеза проверяется на уровне значимости $\alpha = 0.05$ с объемом выборки, варьирующимся от $n = 2500$ до $n = 20000$ с числом повторов эксперимента $m = 1000$. В качестве ядра рассматривается использовалось ядро Епанечникова $K(x) = 0.75(1 - x^2),\ x \in [-1, 1]$, величина h_n выбрана следующим образом: $h_n = n^{-\frac{1}{3}}$ (оптимально по [17]).

На рисунке 2.1 приведено поведение оценки предельной дисперсии. Как можно увидеть, сходимость к предельному значению медленная, но это объясняется скоростью сходимости к предельному распределению, определяемой величиной $\sqrt{nh_n} = n^{\frac{1}{3}}$. Тем не менее, результаты согласованы с теоретическими значениями.

На рисунках 2.2a и 2.2b представлены гистограммы значений статистики при m повторных испытаниях при объеме выборки $n = 20000$ без поправки на предельное смещение и с поправкой. В обоих случаях гистограмма соответствует предельному распределению, которое является нормальным с предельной дисперсией $\sigma^2 = 1.2$.

На рисунке 2.3 представлена гистограмма p-значений статистики при выполненной $\mathbb{H}_0$, вычисленных по предельному распределению при m испытаниях при объеме выборки $n = 20000$. Как видно, оно близко к равномерному, хотя и сходится к нему медленно, что снова объясняется скоростью сходимости.

В таблице 2.1 представлены оценки $\widehat{\alpha}_I$ вероятности ошибки первого рода α_I при разных объемах выборки и оценки их дисперсии, демонстрирующие их точность. Как можно увидеть, с ростом объема выборки $\widehat{\alpha}_I$ сближается с $\alpha = 0.05$, что согласуется с асимптотичностью построенного критерия.

n	2500	5000	10000	20000
$\widehat{\alpha_I}$	0.047	0.067	0.063	0.053
Оц. дисперсии $\widehat{\alpha_I}$	4.48×10^{-5}	6.26×10^{-5}	5.91×10^{-5}	5.02×10^{-5}

Таблица 2.1: Оценка вероятности ошибки первого рода и оценка ее дисперсии при разных объемах выборки.

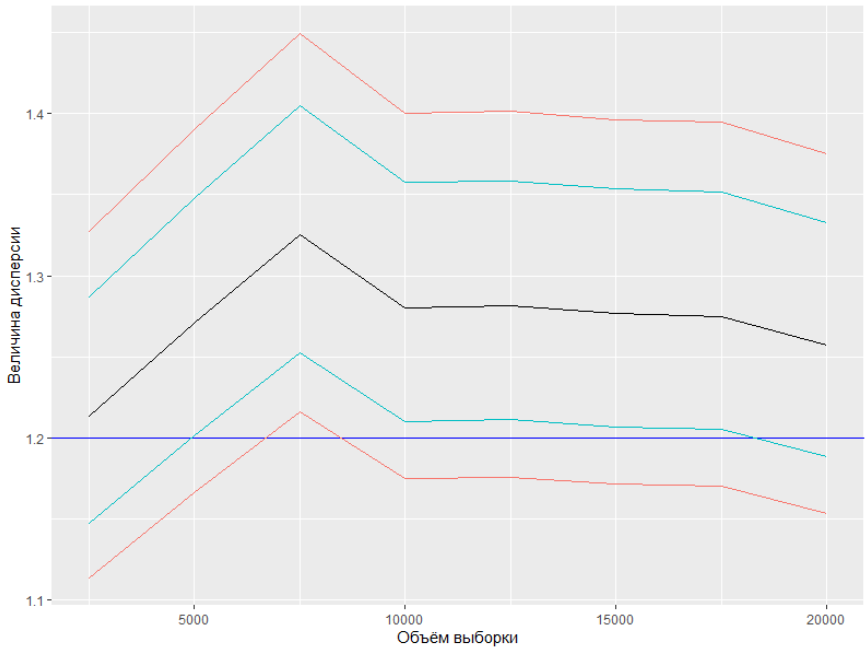
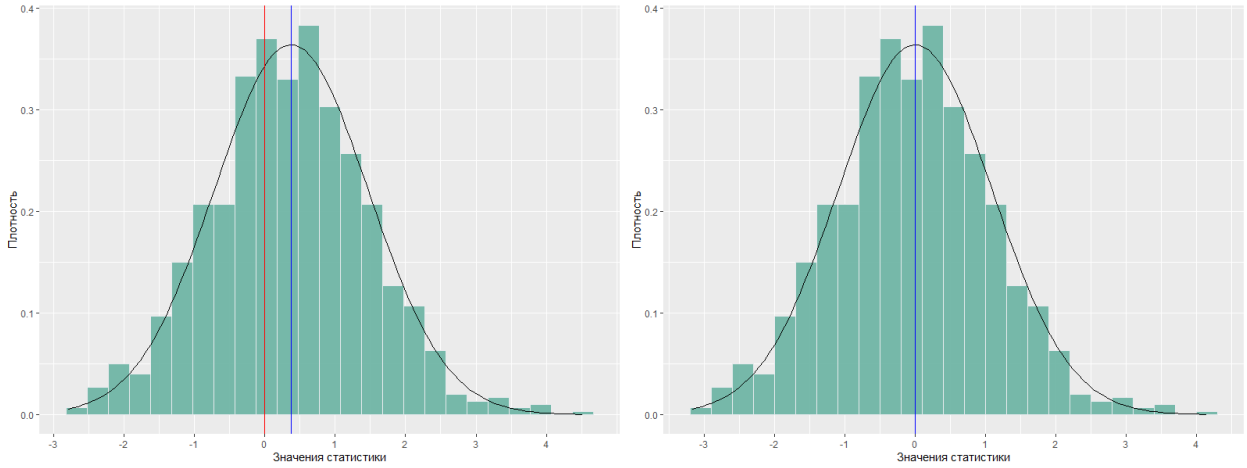


Рис. 2.1: Поведение оценки предельной дисперсии. Синяя горизонтальная прямая соответствует предельной дисперсии, которая в данном случае равна $\sigma^2 = 1.2$. Линия черного цвета соответствует оценкам предельной дисперсии при разных объемах выборки. Бирюзовые и красные линии показывают 80% и 95% асимптотические доверительные интервалы для дисперсии.



(a) (b)

Рис. 2.2: Гистограммы значений статистики и плотности предельных распределений. На рисунке (a) продемонстрирован случай без поправки на асимптотическое смещение, на рисунке (b) — с поправкой. Синей вертикальной прямой показано предельное значение среднего. Величина смещения равна $B_1 - B_2 = 0.375$, что отвечает среднему в случае (a).

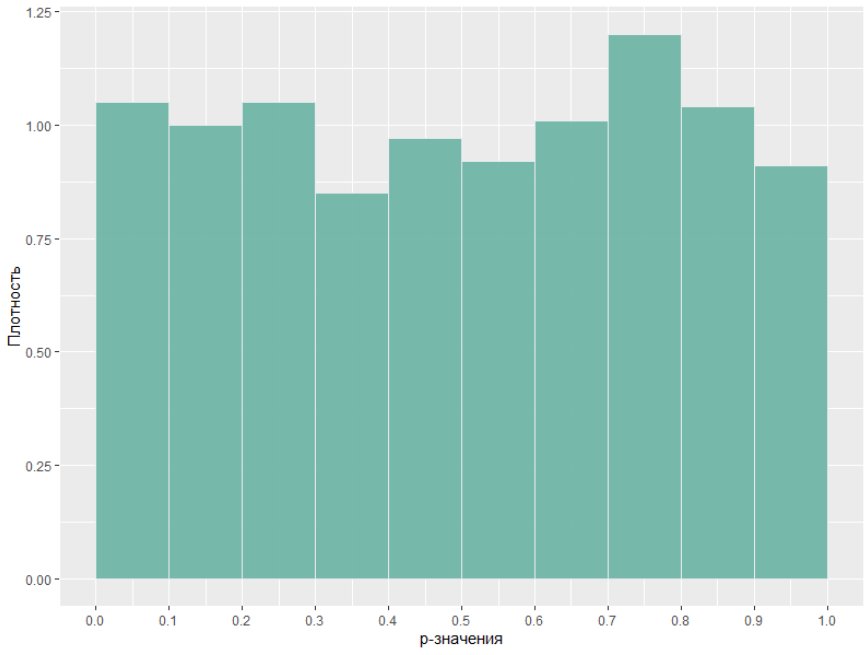


Рис. 2.3: Гистограмма p-значений статистики, посчитанные по предельному распределению.

Заключение

В результате работы были получены критерии проверки гипотезы о разрывности функции условной медианы и плотности интенсивности пуассоновского процесса, а также модификация первого в двух формах, основанная на перестановках и предназначенный для меньших объемов выборки. При этом

- Для двух построенных в работе критериев аналитически изучены асимптотические свойства;
- Для них же было проведено моделирование, которое подтвердило эти свойства;
- Обычный перестановочный критерий проверки гипотезы о разрывности функции регрессии, построенный на оценке условной медианы, работает корректно при соблюдении гипотезы равенства распределений;
- Построен стьюдентизированный перестановочный критерий проверки гипотезы о разрывности функции регрессии, построенный на оценке условной медианы, который корректен даже при нарушении гипотезы равенства распределений. Моделирование это подтверждает;
- Для двух критериев проверки гипотезы о разрывности функции регрессии, основанных на условном матожидании и условной медиане, проведено сравнение на модельных данных, показавшее неустойчивость первого к шуму, в отличие от второго.

Программный код, реализованный для моделирования, размещен на ресурсе Zenodo [18].

В дальнейшем планируется построить модификацию критерия проверки разрывности плотности пуассоновского процесса с использованием подхода перестановочных критериев, а также сопоставить полученный критерий проверки разрывности плотности пуассоновского процесса с известными методами обнаружения разладок в случае пуассоновского процесса.

Список литературы

1. Thistlethwaite D. L., Campbell D. T. Regression-discontinuity analysis: An alternative to the ex post facto experiment // Journal of Educational Psychology. — 1960. — Vol. 51, no. 2. — P. 309–317. — Access mode: https://doi.org/10.1037/h0044319.
2. Lemieux Th., Lee D. Regression Discontinuity Design in Economics // Journal of Economic Literature. — 2010. — 06. — Vol. 48. — P. 281–355.
3. Calcagno J., Long B. The Impact of Postsecondary Remediation Using a Regression Discontinuity Approach: Addressing Endogenous Sorting and Noncompliance. — 2008. — 08. — P. 47.
4. Skovron C., Titiunik R. A Practical Guide to Regression Discontinuity Designs in Political Science. — 2015.
5. Bertanha Marinho, Chung EunYi. Permutation Tests at Nonparametric Rates. — 2021. — Access mode: https://arxiv.org/abs/2102.13638.
6. An analysis of airport noise data using a non-homogeneous Poisson model with a change-point / Guarnaccia C., Quartieri J., Tepedino C., and Rodrigues E. // Applied Acoustics. — 2015. — Vol. 91.
7. Circular Binary Segmentation for the Analysis of Array-based DNA Copy Number Data / Olshen A., Venkatraman E.S., Lucito R., and Wigler M. // Biostatistics (Oxford, England). — 2004. — Vol. 5. — P. 557–72.
8. The Uncertainty of Storm Season Changes: Quantifying the Uncertainty of Autocovariance Changepoints / Nam C. F. H., Aston J. A. D., Eckley I. A., and Killick R. // Technometrics. — 2015. — Vol. 57, no. 2. — P. 194–206. — https://doi.org/10.1080/00401706.2014.902776.
9. Yao Y.C. Estimating the number of change-points via Schwarz' criterion // Statistics & Probability Letters. — 1988. — Vol. 6, no. 3. — P. 181–189. — Access mode: https://www.sciencedirect.com/science/article/pii/0167715288901186.
10. Lavielle M. Using penalized contrasts for the change-point problem // Signal Processing. — 2005. — Vol. 85, no. 8. — P. 1501–1510. — Access mode: https://www.sciencedirect.com/science/article/pii/S0165168405000381.
11. Bai J., Perron P. Estimating and testing linear models with multiple structural changes // Econometrica. — 1995. — Vol. 66. — P. 47–78.
12. Dette H., Wied D. Detecting relevant changes in time series models // Journal of the Royal Statistical Society. Series B (Statistical Methodology). — 2016. — Vol. 78, no. 2. — P. 371–394. — Access mode: http://www.jstor.org/stable/24775343 (online; accessed: 2022-05-05).

13. Gramacki A. Nonparametric Kernel Density Estimation and Its Computational Aspects. — 1 ed. — Poland : Springer, 2018. — P. 176. — ISBN: 978-3-319-71687-9. — Access mode: https://doi.org/10.1007/978-3-319-71688-6.
14. Zucchini W. Applied smoothing techniques, Part 1 Kernel Density Estimation. — 2003.
15. Lehmann E. L., Romano J. P. Testing statistical hypotheses. Springer Texts in Statistics. — Third ed. — New York : Springer, 2005. — P. xiv+784. — ISBN: 0-387-98864-5.
16. Huber P.J. Robust statistics. — Wiley New York, 1981.
17. Kutoyants Yu. A. Statistical Inference for Spatial Poisson Processes. — London : Springer New York, NY, 1998. — P. 17–18, 21, 28.
18. Grigorev D. Graduation project simulation code. — 2022. — May. — Access mode: https://doi.org/10.5281/zenodo.6569651.
19. Cline D. B. H, Hart J. D. Kernel Estimation of Densities with Discontinuities or Discontinuous Derivatives // Statistics. — 1991. — Vol. 22, no. 1. — P. 69–84. — https://doi.org/10.1080/02331889108802286.
20. Schuster E. F. Incorporating support constraints into nonparametric estimators of densities // Communications in Statistics-theory and Methods. — 1985. — Vol. 14. — P. 1123–1136.
21. Silverman B. W. Density Estimation for Statistics and Data Analysis. — London : Chapman & Hall, 1986.
22. Kheireddine S, Abdallah S., Yahia D. General method of boundary correction in kernel regression estimation // Afrika Statistika. — 2015. — 12. — Vol. 10. — P. 739–750.
23. Rosenblatt M. Remarks on Some Nonparametric Estimates of a Density Function // The Annals of Mathematical Statistics. — 1956. — Vol. 27, no. 3. — P. 832 – 837. — Access mode: https://doi.org/10.1214/aoms/1177728190.

Приложение А

Доказательства утверждений

А.1. Доказательство утверждения 1.1

В целом, доказательство этого утверждения аналогично доказательству утверждения D.2 в работе [5], поэтому некоторые моменты доказательства будут обосновываться ссылкой на страницы доказательства этого утверждения.

У величины h_n далее будем опускать индекс, но эта величина все равно зависит от n.

Пусть дана выборка $\mathbf{V}_1 = (\mathbf{R}_1, \mathbf{S}_1), \ldots, \mathbf{V}_m = (\mathbf{R}_m, \mathbf{S}_m)$ объема m из распределения $P \in \mathcal{P}$. Рассматривается условная медиана $\theta(P) = \arg\min_a \mathbb{E}[|\mathbf{S} - a| \mid \mathbf{R} = 0^+]$.

Оценка условной медианы $\mathbf{S}$ при условии $\mathbf{R} = x = 0^+$, как уже известно, имеет вид:

$$\hat{\theta} = \arg\min_a \sum_{k=1}^{m} |\mathbf{S}_k - a| K\left(\frac{\mathbf{R}_k}{h}\right),$$

что то же, что

$$\hat{\theta} = \arg\min_a \frac{1}{2} \sum_{k=1}^{m} |\mathbf{S}_k - a| \, K\left(\frac{\mathbf{R}_k}{h}\right).$$

Определим $\mathbf{Z}_m = \sqrt{mh}(\hat{\theta} - \theta(P))$. Необходимо получить линейную аппроксимацию для $\mathbf{Z}_m$.

Из определения следует, что

$$\mathbf{Z}_m = \arg\min_z \frac{1}{2} \sum_{k=1}^{m} \left|\mathbf{S}_k - \theta(P) - \frac{z}{\sqrt{mh}}\right| K\left(\frac{\mathbf{R}_k}{h}\right).$$

Функцию под $\arg\min$ обозначим как $L_m(z)$. Так как $L_m(0)$ не зависит от z,то

$$\mathbf{Z}_m = \arg\min_z (L_m(z) - L_m(0)).$$

Пусть $Q_m(z) = L_m(z) - L_m(0)$ и $\mathbf{U}_k = \mathbf{S}_k - \theta(P)$. Тогда

$$Q_m(z) = \frac{1}{2} \sum_{k=1}^{m} \left(\left|\mathbf{U}_k - \frac{z}{\sqrt{mh}}\right| - |\mathbf{U}_k| \right) K\left(\frac{\mathbf{R}_k}{h}\right).$$

По построению $\mathbf{Z}_m$ минимизирует $Q_m(z)$.

Получение аппроксимации $\mathbf{Z}_m$

Произведем квадратичную аппроксимацию функции $Q_m(z)$.

Пусть $D_i = \mathbb{I}\{\mathbf{U}_i < 0\} - \frac{1}{2}$ и $\mathbf{V}_i(z) = \frac{1}{2}\left|\mathbf{U}_i - \frac{z}{\sqrt{mh}}\right| - \frac{1}{2}|\mathbf{U}_i| - \frac{z}{\sqrt{mh}}\mathbf{D}_i$, а $\underline{\mathbf{R}}_m$ — вектор из величин $\mathbf{R}_i$. Тогда функцию $Q_m(z)$ можно записать в следующем виде:

$$Q_m(z) = \mathbb{E}_P(Q_m(z) \mid \underline{\mathbf{R}}_m) + \tag{A.1.1}$$

$$+ \frac{1}{\sqrt{mh}} \sum_{k=1}^{m} z(\mathbf{D}_k - \mathbb{E}_P[\mathbf{D}_k \mid \mathbf{R}_k]) K\left(\frac{\mathbf{R}_k}{h}\right) + \tag{A.1.2}$$

$$+ \sum_{k=1}^{m} (\mathbf{V}_i(z) - \mathbb{E}_P[\mathbf{V}_i(z) \mid R_i]) K\left(\frac{\mathbf{R}_i}{h}\right), \tag{A.1.3}$$

где $\underline{\mathbf{R}}_m = (\mathbf{R}_1, \ldots, \mathbf{R}_m)$, $\mathbf{D}_k = \mathbb{I}(\mathbf{U}_k < 0) - \frac{1}{2}$.

Для изучения $\mathbb{E}_P(Q_m(z) \mid \underline{\mathbf{R}}_m)$ введем функцию

$$M(t \mid r; P) = \frac{1}{2}\mathbb{E}_P(|\mathbf{U} + t| \mid \mathbf{R} = r).$$

Тогда

$$\begin{aligned}\mathbb{E}_P(Q_m(z) \mid \underline{\mathbf{R}}_m) &= \frac{1}{2}\sum_{k=1}^{m} \mathbb{E}_P\left[\left|\mathbf{U}_k - \frac{z}{\sqrt{mh}}\right| - |\mathbf{U}_k| \,\middle|\, \mathbf{R}_k\right] K\left(\frac{\mathbf{R}_k}{h}\right) \\ &= \sum_{k=1}^{m}\left[M\left(-\frac{z}{\sqrt{mh}}\middle|\mathbf{R}_k; P\right) - M\left(0 \mid \mathbf{R}_k; P\right)\right] K\left(\frac{\mathbf{R}_k}{h}\right)\end{aligned}$$

Функция M имеет производные:

$$\begin{aligned}\nabla_t M(t \mid r; P) &= \frac{1}{2} - F_{U|R}(-t \mid r; P), \\ \nabla_{t^2} M(t \mid r; P) &= f_{U|R}(-t \mid r; P), \\ \nabla_{t^3} M(t \mid r; P) &= -\nabla_u f_{U|R}(-t \mid r; P), \\ \nabla_{tr} M(t \mid r; P) &= -\nabla_r F_{U|R}(-t \mid r; P), \\ \nabla_{tr^2} M(t \mid r; P) &= -\nabla_{r^2} F_{U|R}(-t \mid r; P),\end{aligned}$$

так как существуют производные $\nabla_u f_{U|R}(u \mid r; P)$, $\nabla_r F_{U|R}(u \mid r; P)$, $\nabla_{r^2} F_{U|R}(u \mid r; P)$ по предположениям 3–5.

Разлагая в ряд Тейлора по t около 0 функцию M с остатком в форме Лагранжа,

приходим к выражению:

$$\mathbb{E}_P(Q_m(z) \mid \underline{\mathbf{R}}_m) = \sum_{k=1}^{m} \Big[\nabla_t M(0 \mid \mathbf{R}_k; P)\frac{-z}{\sqrt{mh}} + \frac{1}{2}\nabla_{t^2} M(0 \mid \mathbf{R}_k; P)\frac{z^2}{mh} -$$
$$- \frac{1}{6}\nabla_{t^3} M(q^* \mid \mathbf{R}_k; P)\frac{z^3}{(mh)^{3/2}}\Big] K\left(\frac{\mathbf{R}_k}{h}\right) =$$
$$= \frac{-z}{\sqrt{mh}} \sum_{k=1}^{m} \nabla_t M(0 \mid \mathbf{R}_k; P) K\left(\frac{\mathbf{R}_k}{h}\right) + \tag{A.2.1}$$
$$+ \frac{1}{2}\frac{z^2}{mh} \sum_{k=1}^{m} f_{U|R}(0 \mid r; P) K\left(\frac{\mathbf{R}_k}{h}\right) + \tag{A.2.2}$$
$$+ \frac{z^3}{6(mh)^{3/2}} \sum_{k=1}^{m} \nabla_u f_{U|R}(q^* \mid r; P) K\left(\frac{\mathbf{R}_k}{h}\right), \tag{A.2.3}$$

где q^* — точка между 0 и $\frac{-z}{\sqrt{mh}}$. Слагаемые из последнего равенства также нужно аппроксимировать.

Учитываем, что

$$\nabla_t M(0 \mid 0^+; P) = \frac{1}{2} - \mathbb{P}_P(\mathbf{S}_k - \theta(P) \le 0 \mid \mathbf{R}_k = 0^+) = 0.$$

Раскладываем в ряд Тейлора с остатком в форме Лагранжа под знаком матожидания функцию $\nabla_t M(0 \mid \mathbf{R}_k; P)$ по аргументу $\mathbf{R}_k$ около $\mathbf{R}_k = 0^+$, то есть справа от нуля, то же делаем и с $f_R(r; P)$:

$$\mathbb{E}_P\left[\frac{-z}{\sqrt{mh}} \sum_{k=1}^{m} \nabla_t M(0 \mid \mathbf{R}_k; P) K\left(\frac{\mathbf{R}_k}{h}\right)\right] = \mathbb{E}_P\left[\frac{-z}{\sqrt{mh}} \sum_{k=1}^{m} \nabla_t M(0 \mid 0^+; P) K\left(\frac{\mathbf{R}_k}{h}\right)\right] +$$
$$+ \mathbb{E}_P\left[\frac{-z}{\sqrt{mh}} \sum_{k=1}^{m} \nabla_{tr} M(0 \mid 0^+; P)\mathbf{R}_k K\left(\frac{\mathbf{R}_k}{h}\right)\right] +$$
$$+ \mathbb{E}_P\left[\frac{-z}{\sqrt{mh}} \sum_{k=1}^{m} \frac{1}{2}\nabla_{tr^2} M(0 \mid x^*; P)\mathbf{R}_k^2 K\left(\frac{\mathbf{R}_k}{h}\right)\right] =$$
$$= z\sqrt{mh} \int_0^{+\infty} \left[\nabla_r F_{U|R}(0 \mid 0^+; P)uh\right] K(u) f_R(uh; P)\, du =$$
$$= z\sqrt{mh}h\nabla_r F_{U|R}(0 \mid 0^+; P) \int_0^{+\infty} uK(u) f_R(uh; P)\, du +$$
$$+ \underbrace{\frac{z}{2}\sqrt{mh}h^2\nabla_{r^2} F_{U|R}(0 \mid x^*; P) \int_0^{+\infty} u^2 K(u) f_R(uh; P)\, du}_{O_{\mathcal{P}}(\sqrt{mh}h^2)} =$$
$$= z\sqrt{mh}h\nabla_r F_{U|R}(0 \mid 0^+; P) f_R(0^+; P) \int_0^{+\infty} uK(u)\, du +$$
$$+ \underbrace{z\sqrt{mh}h^2\nabla_r F_{U|R}(0 \mid 0^+; P) \int_0^{+\infty} u^2 K(u)\nabla_r f_R(x^{**}; P)\, du}_{O_{\mathcal{P}}(\sqrt{mh}h^2)} + O_{\mathcal{P}}(\sqrt{mh}h^2),$$

где использованы следующие факты:

- x^* — точка между $\mathbf{R}_k$ и 0^+ из разложения в ряд Тейлора с остатком в форме Лагранжа,
- x^{**} — точка между uh и 0^+ из разложения в ряд Тейлора с остатком в форме Лагранжа,
- $f_R(r;P)$, $\nabla_r f_R(r;P)$, $F_{U|R}(0 \mid r;P)$ и ее производные ограничены по r и P по предположениям 3 и 5,
- $\int_0^{+\infty} u^2 K(u)\,du < \infty$.

Доказательство того, что дисперсия A.2.1 есть $o_{\mathcal{P}}(1)$, аналогично тому, что приведено в доказательстве утверждения D2 в [5](см. стр. 62). Все это подтверждает то, что

$$(\text{A.2.1}) = z\sqrt{mh}h\nabla_r F_{U|R}(0 \mid 0^+;P) f_R(0^+;P) \int_0^{+\infty} uK(u)\,du + o_{\mathcal{P}}(1).$$

Рассмотрим теперь A.2.2. Аналогично покажем, что оно приближается некоторым числом. Под знаком матожидания разложим в ряд Тейлора $f_{U|R}(0 \mid r;P)$ по r около 0^+ с остатком в форме Лагранжа, а также $f_R(uh;P)$:

$$\begin{aligned}
&\mathbb{E}_P\left[\frac{z^2}{2mh}\sum_{k=1}^{m} f_{U|R}(0 \mid \mathbf{R}_k;P) K\left(\frac{\mathbf{R}_k}{h}\right)\right] = \frac{z^2}{2h}\mathbb{E}_P f_{U|R}(0 \mid \mathbf{R};P) K\left(\frac{\mathbf{R}}{h}\right) = \\
&= \frac{z^2}{2}\int_0^{+\infty} f_{U|R}(0 \mid uh;P) f_R(uh;P) K(u)\,du = \\
&= \frac{z^2}{2} f_R(0^+;P)\int_0^{+\infty} (f_{U|R}(0 \mid 0^+;P) + f_{U|R}(0 \mid x^*;P)uh) K(u)\,du + \\
&+ \frac{z^2 h}{2} f_R(x^{**};P)\int_0^{+\infty} (f_{U|R}(0 \mid 0^+;P) + f_{U|R}(0 \mid x^*;P)uh) uK(u)\,du = \\
&= \frac{z^2}{2} f_{U|R}(0 \mid 0^+;P) f_R(0^+;P) \underbrace{\int_0^{+\infty} K(u)\,du}_{=\frac{1}{2}} + o_{\mathcal{P}}(1).
\end{aligned}$$

где использованы следующие факты:

- x^* — точка между $\mathbf{R}$ и 0^+ из разложения в ряд Тейлора с остатком в форме Лагранжа,
- x^{**} — точка между uh и 0^+ из разложения в ряд Тейлора с остатком в форме Лагранжа,

- $f_R(r;P)$, $f_{U|R}(0 \mid r;P)$ и их производные ограничены по r и P по предположениям 3 и 5,
- $\int_0^{+\infty} u^2 K(u)\,du < \infty$ и ядро K симметрично.

Дисперсия A.2.2 также мала (см. док-во утверждения D2 в [5], стр. 63). Отсюда

$$(\text{A.2.2}) = \frac{z^2}{4} f_{U|R}(0 \mid 0^+;P) f_R(0^+;P) + o_{\mathcal{P}}(1).$$

Оттуда же (A.2.3) $= o_{\mathcal{P}}(1)$ (см. [5], стр. 63–64). Возвращаясь к разложению $Q_m(z)$ по слагаемым A.1.1, A.1.2 и A.1.3, введем следущие величины:

$$\begin{aligned}
B_0(P) &= \int_0^{+\infty} uK(u)\,du F_{U|R}(0 \mid 0^+;P) f_R(0^+;P),\\
M_m &= \frac{1}{\sqrt{mh}} \sum_{k=1}^{m} (\mathbf{D}_k - \mathbb{E}_P[\mathbf{D}_k \mid \mathbf{R}_k]) K\left(\frac{\mathbf{R}_k}{h}\right) + \sqrt{mh} h B_0(P),\\
Q_m^*(z) &= \frac{z^2}{4} \underbrace{f_{U|R}(0 \mid 0^+;P) f_R(0^+;P)}_{N} + zM_m,\\
r_m(z) &= Q_m(z) - Q_m^*(z),
\end{aligned}$$

откуда по построению $r_m(z) = o_{\mathcal{P}}(1)$ для любого фиксированного z и $Q_m^*(z)$ является квадратичной аппроксимацией $Q_m(z)$. Найдем точку минимума $Q_m^*(z)$:

$$Q_m^*(z) = \frac{z^2}{4} N + zM_m = \frac{1}{4} N \left(z + \frac{2M_m}{N}\right)^2 - \frac{M_m^2}{N},$$

что минимизируется при

$$z = z_m = -\frac{2M_m}{N} = -\frac{2}{N}\left(\frac{1}{\sqrt{mh}} \sum_{k=1}^{m} (\mathbf{D}_k - \mathbb{E}_P[\mathbf{D}_k \mid \mathbf{R}_k]) K\left(\frac{\mathbf{R}_k}{h}\right) + \sqrt{mh} h B_0(P)\right).$$

Отсюда же получаем поправку на смещение $B(P)$:

$$B(P) = \frac{-2B_0(P)}{N} = \frac{-2\int_0^{+\infty} uK(u)\,du}{f_{S|R}(\theta(P) \mid 0^+;P)} \nabla_r F_{S|R}(\theta(P) \mid 0^+;P)$$

По доказательству утверждения D2 в [5] $\mathbf{Z}_m = z_m + o_{\mathcal{P}}(1)$ (см. стр. 66). Все это приводит к тому, что

$$\begin{aligned}
&\sqrt{mh}(\widehat{\theta} - \theta(P) - hB(P)) =\\
&= \frac{1}{\sqrt{mh}} \sum_{k=1}^{m} \frac{-1}{f_{U|R}(0 \mid 0^+;P) f_R(0^+;P)/2} (\mathbf{D}_k - \mathbb{E}_P[\mathbf{D}_k \mid \mathbf{R}_k]) K\left(\frac{\mathbf{R}_k}{h}\right) + o_{\mathcal{P}}(1) =\\
&= \frac{1}{\sqrt{m}} \sum_{k=1}^{m} \underbrace{\frac{-(\mathbb{I}\{\mathbf{S}_k < \theta(P)\} - F_{S|R}(\theta(P) \mid \mathbf{R}_k;P))}{f_{S|R}(\theta(P) \mid 0^+;P) f_R(0^+;P)\sqrt{h}/2} K\left(\frac{\mathbf{R}_k}{h}\right)}_{\psi_n(\mathbf{V}_k;P)} + o_{\mathcal{P}}(1).
\end{aligned}$$

Получены функции ψ_n из предположения B.1. В то же время

$$\mathbb{E}_P(\mathbb{I}\{\mathbf{S}_k < \theta(P))K\left(\frac{\mathbf{R}_k}{h}\right) = \mathbb{E}_R K\left(\frac{\mathbf{R}_k}{h}\right)\mathbb{E}_S\left[(\mathbb{I}\{\mathbf{S}_k < \theta(P)) \mid \mathbf{R}_k\right] =$$
$$= \mathbb{E}_R K\left(\frac{\mathbf{R}_k}{h}\right) F_{S|R}(\theta(P) \mid \mathbf{R}_k; P),$$

что гарантирует, что $\forall P \in \mathcal{P}\ \mathbb{E}_P\psi_n(\mathbf{V}_k; P) = 0$.

Величина $\xi^2(P)$, имеющая вид:

$$\xi^2(P) = \int_0^\infty K^2(u)\,du\frac{1}{f^2_{U|R}(0 \mid 0^+; P)f_R(0^+; P)} = \int_0^\infty K^2(u)\,du\frac{1}{f^2_{S|R}(\theta(P) \mid 0^+; P)f_R(0^+; P)}$$

В точности аппроксимирует $\mathbb{D}\psi_n(\mathbf{V}_k; P)$ в том смысле, что,

$$\sup_{P\in\mathcal{P}}\left|\mathbb{D}\psi_n(\mathbf{V}_k; P) - \xi^2(P)\right| \to 0,$$

что следует из доказательства утверждения D2 в [5].

Таким образом, пункты предположения B.1 удовлетворены.

□

А.2. Доказательство утверждения 2.1

Лемма А.1 (F. Liese)**.** Пусть f суммируема относительно Λ со степенью $2p$. Тогда существует константа C_p, независящая от f и Λ, что $\forall A \in \mathcal{B}: \ \Lambda(A) < \infty$ справедливо неравенство:

$$\mathbb{E}\left(\int_A f(x)\,\pi(dx)\right)^{2p} \le C_p\left[\int_A f^{2p}(x)\,\Lambda(dx) + \left(\int_A f^2(x)\,\Lambda(dx)\right)^p\right].$$

Этап 1: линейная аппроксимация статистики

Распишем статистику следующим образом:

$$\begin{aligned}
&\sqrt{nh}(\widehat{\theta}_1 - \widehat{\theta}_2) - \sqrt{nh}(\theta_1 - \theta_2) = \sqrt{nh}(\widehat{\theta}_1 - \theta_1) - \sqrt{nh}(\widehat{\theta}_2 - \theta_2) = \\
&= \sqrt{\frac{n}{n_1}}\sqrt{n_1 h}\left(\frac{1}{n_1 h}\sum_{i=1}^{n_1}\int_{-1}^{1} K\left(\frac{y}{h}\right)\operatorname{sign}(y)\,\mathbf{X}_{1i}(dy) - \frac{1}{n_1 h}\sum_{i=1}^{n_1}\int_{-1}^{1} K\left(\frac{y}{h}\right)\operatorname{sign}(y)\lambda_1(y)\,dy + \right. \\
&\left. + \frac{1}{n_1 h}\sum_{i=1}^{n_1}\int_{-1}^{1} K\left(\frac{y}{h}\right)\operatorname{sign}(y)\lambda_1(y)\,dy - \frac{1}{2}(\lambda_1(0^+) - \lambda_1(0^-))\right) - \\
&- \sqrt{\frac{n}{n_2}}\sqrt{n_2 h}\left(\frac{1}{n_2 h}\sum_{i=1}^{n_2}\int_{-1}^{1} K\left(\frac{y}{h}\right)\operatorname{sign}(y)\,\mathbf{X}_{2i}(dy) - \frac{1}{n_2 h}\sum_{i=1}^{n_2}\int_{-1}^{1} K\left(\frac{y}{h}\right)\operatorname{sign}(y)\lambda_2(y)\,dy + \right. \\
&\left. + \frac{1}{n_2 h}\sum_{i=1}^{n_2}\int_{-1}^{1} K\left(\frac{y}{h}\right)\operatorname{sign}(y)\lambda_2(y)\,dy - \frac{1}{2}(\lambda_2(0^+) - \lambda_2(0^-))\right) = \\
&= \sqrt{\frac{n}{n_1}}\left(\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\underbrace{\frac{1}{\sqrt{h}}\left(\int_{-1}^{1} K\left(\frac{y}{h}\right)\operatorname{sign}(y)\,\mathbf{X}_{1i}(dy) - \int_{-1}^{1} K\left(\frac{y}{h}\right)\operatorname{sign}(y)\lambda_1(y)\,dy\right)}_{=\psi_{n1}(\mathbf{X}_{1i})}\right) + \\
&+ \sqrt{\frac{n}{n_1}}\underbrace{\sqrt{n_1 h}\left(\int_{-1}^{1}\frac{1}{h}K\left(\frac{y}{h}\right)\operatorname{sign}(y)\lambda_1(y)\,dy - \frac{1}{2}(\lambda_1(0^+) - \lambda_1(0^-))\right)}_{=\chi_1} - \\
&- \sqrt{\frac{n}{n_2}}\left(\frac{1}{\sqrt{n_2}}\sum_{i=1}^{n_2}\underbrace{\frac{1}{\sqrt{h}}\left(\int_{-1}^{1} K\left(\frac{y}{h}\right)\operatorname{sign}(y)\,\mathbf{X}_{2i}(dy) - \int_{-1}^{1} K\left(\frac{y}{h}\right)\operatorname{sign}(y)\lambda_2(y)\,dy\right)}_{=\psi_{n2}(\mathbf{X}_{2i})}\right) - \\
&- \sqrt{\frac{n}{n_2}}\underbrace{\sqrt{n_2 h}\left(\int_{-1}^{1}\frac{1}{h}K\left(\frac{y}{h}\right)\operatorname{sign}(y)\lambda_2(y)\,dy - \frac{1}{2}(\lambda_2(0^+) - \lambda_2(0^-))\right)}_{=\chi_2}.
\end{aligned}$$

Покажем далее, что $\frac{1}{\sqrt{n_k}}\sum_{i=1}^{n_k}\psi_{nk}(\mathbf{X}_{ki}),\ k = 1, 2$ имеют асимптотическое нормальное распределение, а $\chi_k,\ k = 1, 2$ являются $O_{\mathcal{P}}(\sqrt{n_k h}h)$. Структурно, $\psi_{nk}(\mathbf{X}_{ki})$ и χ_k похожи, потому далее будем доказывать все для $k = 1$, опустив индекс, относящийся к k, а n_1 обозначим как m.

Этап 2: $\chi = O_{\mathcal{P}}(\sqrt{mh}h)$

Покажем, что χ либо мало, либо обеспечивает асимптотическое смещение, поправку на которое нужно делать.

$$\chi = \sqrt{mh}\left(\int_{-1}^{1}\frac{1}{h}K\left(\frac{y}{h}\right)\operatorname{sign}(y)\lambda(y)\,dy - \frac{1}{2}(\lambda(0^+)-\lambda(0^-))\right) =$$
$$= \sqrt{mh}\left(-\int_{-h^{-1}}^{0}K(u)\lambda(uh)\,du + \frac{\lambda(0^-)}{2} + \int_{0}^{h^{-1}}K(u)\lambda(uh)\,du - \frac{\lambda(0^+)}{2}\right) =$$

Поскольку при больших n h^{-1} также велико, то пределы интегрирования по u можно считать варьирующимися от -1 до 1 ввиду того, что вне отрезка $[-1,1]$ $K \equiv 0$. Далее, раскладываем $\lambda(t)$ в ряд Тейлора с остатком в форме Лагранжа:

$$= \sqrt{mh}\left(-\int_{-1}^{0}K(u)(\lambda(0^-)+uh\lambda'(0^-)+\frac{1}{2}u^2h^2\lambda''(u^*))\,du + \frac{\lambda(0^-)}{2}\right.$$
$$\left. + \int_{0}^{1}K(u)(\lambda(0^+)+uh\lambda'(0^+)+\frac{1}{2}u^2h^2\lambda''(u^{**}))\,du - \frac{\lambda(0^+)}{2}\right) =$$
$$\sqrt{mh}h\underbrace{\int_{0}^{1}uK(u)\,du(\lambda'(0^+)+\lambda'(0^-))}_{=B} + O_{\mathcal{P}}(\sqrt{mh}h^2),$$

где использованы следующие факты

- точки $u^*,\ u^{**}$ — точки из представления остатка в форме Лагранжа,
- производная $\lambda''(t)$ ограничена,
- ядро K ограничено,
- $\int_0^1 K(u)\,du = \frac{1}{2}$, $\int_0^1 u^2K(u)\,du < \infty$.

Если $h = n^{-\frac{1}{3}}$, то необходима поправка на смещение B, т.к. в таком случае $\sqrt{nh}h \to 1$.

Этап 3: $\mathbb{E}\psi_n(\mathbf{X}_i)$ **и** $\mathbb{D}\psi_n(\mathbf{X}_i)$

По свойствам интеграла по мере процесса Пуассона:

$$\mathbb{E}\psi_n(\mathbf{X}_i) = \frac{1}{\sqrt{h}}\left(\int_{-1}^{1}K\left(\frac{y}{h}\right)\operatorname{sign}(y)\lambda(y)\,dy - \int_{-1}^{1}K\left(\frac{y}{h}\right)\operatorname{sign}(y)\lambda(y)\,dy\right) = 0.$$

Что касается дисперсии, то покажем, что

$$\xi^2 = \int_0^1 K^2(u)\,du(\lambda(0^+)+\lambda(0^-))$$

является пределом для $\mathbb{D}\psi_n(\mathbf{X}_i)$. По свойствам интеграла по мере процесса Пуассона

$$\begin{aligned}
&\mathbb{D}\psi_n(\mathbf{X}_i) - \xi^2 = \frac{1}{h}\int_{-1}^{1} K^2\left(\frac{y}{h}\right)\operatorname{sign}^2(y)\lambda(y)\,dy - \xi^2 = \\
&= \int_{-1}^{0} K^2(u)\lambda(uh)\,dy + \int_{0}^{1} K^2(u)\lambda(uh)\,dy - \xi^2 = \\
&= \int_{-1}^{0} K^2(u)(\lambda'(0^-) + uh\lambda(u^*))\,dy + \int_{0}^{1} K^2(u)(\lambda''(0^+) + uh\lambda(u^{**}))\,dy - \xi^2 = \\
&= O_{\mathcal{P}}(h),
\end{aligned}$$

где были использованы следующие факты

- точки $u^*,\ u^{**}$ — точки из представления остатка в форме Лагранжа,
- производная $\lambda'(t)$ ограничена,
- ядро K ограничено,
- $\int_0^1 uK^2(u)\,du < \infty$.

Таким образом мы имеем информацию о предельной дисперсии.

Этап 4: Асимптотическая нормальность

Докажем нормальность суммы $\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\psi_n(\mathbf{X}_i)$ по ЦПТ Ляпунова, где в качестве слагаемых рассматриваются $\frac{\psi_n(\mathbf{X}_i)}{\sqrt{m}}$. Для этого необходимо показать, что

$$\exists\delta > 0:\ m^{-\delta/2}\frac{\mathbb{E}|\psi_n(\mathbf{X}_i)|^{2+\delta}}{(\sqrt{\mathbb{D}\psi_n(\mathbf{X}_i)})^{2+\delta}} \to 0.$$

Во-первых, заметим, что т.к. $\mathbb{D}\psi_n(\mathbf{X}_i) - \xi^2 = O_{\mathcal{P}}(h)$, то $(\mathbb{D}\psi_n(\mathbf{X}_i))^{-1} = O_{\mathcal{P}}(1)$. Далее, займемся оставшейся частью нужного выражения. В процессе доказательства было получено, что

$$\int_{-1}^{1} K\left(\frac{y}{h}\right)\operatorname{sign}(y)\lambda(y)\,dy = O_{\mathcal{P}}(h),$$

поэтому далее будем опускать эту величину таким образом. Возьмем произвольное $\delta > 0$.

$$m^{-\delta/2}\mathbb{E}|\psi_n(\mathbf{X}_i)|^{2+\delta} = \frac{1}{(mh)^{\delta/2}h}\mathbb{E}\left|\int_{-1}^{1} K\left(\frac{y}{h}\right)\operatorname{sign}(y)\,\mathbf{X}_i(dy) - O_{\mathcal{P}}(h)\right|^{2+\delta}.$$

Для величины под знаком математического ожидания выполняется лемма Liese A.1, так как функции K и sign ограничены и поэтому $L_{2+\delta}$ суммируемы, потому существует

константа C_δ, независящая от подынтегральной функции в интеграле под матожиданием и такая, что

$$\mathbb{E}\left|\int_{-1}^{1} K\left(\frac{y}{h}\right)\operatorname{sign}(y)\,\mathbf{X}_i(dy) - O_{\mathcal{P}}(h)\right|^{2+\delta} \le C_\delta \left[\underbrace{\int_{-1}^{1} K^{2+\delta}\left(\frac{y}{h}\right)(\operatorname{sign}(y))^{2+\delta}\lambda(y)\,dy}_{=O_{\mathcal{P}}(h)} + \left(\underbrace{\int_{-1}^{1} K^{2}\left(\frac{y}{h}\right)(\operatorname{sign}(y))^{2}\lambda(y)\,dy}_{=O_{\mathcal{P}}(h)}\right)^{1+\delta/2}\right],$$

откуда

$$m^{-\delta/2}\mathbb{E}|\psi_n(\mathbf{X}_i)|^{2+\delta} = \frac{C_\delta}{(mh)^{\delta/2}h}(O_{\mathcal{P}}(h) + O_{\mathcal{P}}(h^{1+\delta/2})) \xrightarrow{p} 0,$$

так как $mh \to \infty$.

Таким образом, ЦПТ Ляпунова верна и тогда

$$\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\psi_n(\mathbf{X}_i) \xrightarrow{d} N(0,\xi^2).$$

Итог

В итоге, суммируя все результаты и то, что $n_1/n \to 1/2$, получается, что

$$\sqrt{nh_n}(\widehat{\theta}_1 - \widehat{\theta}_2) - \sqrt{nh_n}(\theta_1 - \theta_2) \xrightarrow{d} N(0,\sigma^2),$$

если $h = o(n^{-1/3})$, или если $h = n^{-1/3}$, то

$$\sqrt{nh_n}(\widehat{\theta}_1 - \widehat{\theta}_2) - \sqrt{nh_n}(\theta_1 - \theta_2) \xrightarrow{d} N(0,\sigma^2) + B_1 - B_2,$$

где $B_k = \int_0^1 uK(u)\,du(\lambda_k'(0^+) + \lambda_k'(0^-))$, а $\sigma^2 = \xi_1^2/0.5 + \xi_2^2/0.5$. Если же ξ_1 и ξ_2 представляют один и тот же процесс, то

$$\sigma^2 = 4\int_0^1 K^2(u)\,du(\lambda(0^+) + \lambda(0^-)).$$

□

Приложение Б

Описание методов, примененных в моделировании 1.4.2

Б.1. Отзеркаливание ядерной оценки плотности

Рассматривается ядро K с носителем $[-1, 1]$. Метод отзеркаливания рассматривался в работах [19, 20, 21] и он состоит в следующем. Пусть $\mathbf{X}_1, \ldots, \mathbf{X}_n$ — выборка из распределения с плотностью $f(x)$, что при $x < 0$ $f(x) = 0$. Расширим выборку зеркальными наблюдениями относительно $x = 0$: $\mathbf{X}_1, \ldots, \mathbf{X}_n, -\mathbf{X}_1, \ldots, -\mathbf{X}_n$ — и по ней оценим $f(x)$ так, как если бы оценивали $2f(x)$. Ядерная оценка плотности в таком случае приобретает вид:

$$\hat{p}(x) = \frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{x - \mathbf{X}_k}{h}\right) + K\left(\frac{x + \mathbf{X}_k}{h}\right).$$

Этот метод решает проблему краевых эффектов, что необходимо в моделировании.

Для ядерного оценивания функций от пары случайных величин $(\mathbf{X}, \mathbf{Y})$, к которым относятся условное матожидание $\mathbb{E}(\mathbf{Y} \mid \mathbf{X} = x)$, условную медиану $\mathbb{M}(\mathbf{Y} \mid \mathbf{X} = x)$, условную плотность $f_{Y|X}(y \mid x)$ и функцию распределения условного распределения $F_{Y|X}(y \mid x)$, также существуют методы отзеркаливания и их расширения. Один из них предложен в работе [22]. Идея оценивания аналогична стандартному отзеркаливанию с заменой $\mathbf{X}$ на $g_q(\mathbf{X})$:

$$\widehat{\mathbb{E}}(\mathbf{Y} \mid \mathbf{X} = x) = \frac{\sum_{k=1}^{n} \mathbf{Y}_k \left(K\left(\frac{x - g_1(\mathbf{X}_k)}{h_n}\right) + K\left(\frac{x + g_1(\mathbf{X}_k)}{h_n}\right)\right)}{\sum_{k=1}^{n} \left(K\left(\frac{x - g_2(\mathbf{X}_k)}{h_n}\right) + K\left(\frac{x + g_2(\mathbf{X}_k)}{h_n}\right)\right)},$$

где $g_q(x),\ q = 1, 2$ — некоторые неотрицательные возрастающие функции, удовлетворяющие ряду свойств, определяемых плотностями $f_X(x)$ и $f_{XY}(x, y)$. В рамках модели 1.4.1 эти функции вырождаются в обычную линейную: $g_q(x) = x$.

Б.2. Исправление ширины окна

В [23] получена аппроксимация h_n, которое минимизирует $MISE$ и она имеет порядок $O(n^{-1/5})$, что точно не удовлетворяет необходимому требованию. Предлагается исправить эту величину до требуемого порядка $O(n^{-1/3})$ домножением на некоторую

степень размера выборки. Покажем, что при больших объемах выборки такое исправление будет незначительным.

Лемма Б.1. Пусть K и f дважды дифференцируемы, интегрируемы с квадратом и их вторые производные интегрируемы с квадратом. Пусть $C(K, f)$ – константа для них. Рассмотрим оптимальную для величины $MISE$ ширину окна $h_{n1} = C(K, f)n^{-1/5}$ и исправленную ширину окна $h_n = h_{n1}n^{-\delta},\ \delta > 0$. Тогда $|MISE(h_{n1}) - MISE(h_n)| = O(n^{-4/5+\delta})$.

Доказательство. По утверждению об асимптотическом представлении $MISE$ в [23]:

$$MISE(h) = \frac{c_1(K)}{nh} + c_2(K, f)h^4 + o((nh)^{-1}) + o(h^4).$$

Здесь $c_1,\ c_2$ – константы для K и f (см. [23]). Тогда

$$MISE(h_n) - MISE(h_{n1}) = \frac{c_1(K)}{C(K, f)}n^{-4/5}(n^{\delta} - 1) + \frac{c_2(K, f)}{C^4(K, f)}n^{-4/5}(n^{-4\delta} - 1) + o(n^{-4/5})$$

Из всех слагаемых, получаемых после раскрытия скобок, только одно ухудшает асимптотику с $O(n^{-4/5})$ до $O(n^{-4/5+\delta})$. Поэтому эта разность есть $O(n^{-4/5+\delta})$. □

Для удовлетворения требованиям утверждения 1.1 предлагается взять $\delta = 2/15 \approx 0.133$, чтобы h_n было $O(n^{-\frac{1}{3}})$. Улучшение ширины окна приводит к уменьшению смещенности ядерной оценки, но в то же время приводит к увеличению ее дисперсии. В рамках моделирования смещение является более важным.

Приложение B

Теория, использованная в главе 1

B.1. Об асимптотической нормальности

Пусть P_1 и P_2 — два распределения, описывающие разные популяции и заданные на $\mathbb{R}$ или $\mathbb{R}^q$, $q \in \mathbb{N}$. Обозначим за $\theta(P_k)$, $k = 1, 2$, вещественнную характеристику распределений. Ставится задача проверки гипотезы:

$$\mathbb{H}_0 : \theta(P_1) = \theta(P_2).$$

Обозначим за $\mathbf{Z}_{k,i}$ выборку без возвращения из распределения P_k объема n_k, $k = 1, 2$, $n_1 + n_2 = n$, причем потребуем независимость этих наблюдений по k (то есть выборки из разных популяций берутся независимо). Также определим семейство распределений $\mathcal{P}$ как выпуклую комбинацию распределений P_k. Для характеристики $\theta(P_k)$ объявим $\hat{\theta}_k = \theta_{n_k}(\mathbf{Z}_{k,1}, \ldots, \mathbf{Z}_{k,n_k})$ как ее состоятельную оценку. Для функций θ_{n_k} выполено предположение:

Предположение B.1 ([5]). Пусть $\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_m$ — выборка из распределения $P \in \mathcal{P}$. Пусть $\hat{\theta} = \theta_m(\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_m)$ оценка характеристики $\theta(P)$. Допустим существование последовательности функций $\psi_n : \mathbb{R}^q \times \mathcal{P} \to \mathbb{R}$, функции $\xi : \mathcal{P} \to \mathbb{R}$ и чисел h_n таких, что $nh_n \to \infty$ и

$$\forall \varepsilon > 0 \ \sup_{P \in \mathcal{P}} \mathbb{P}_P \left\{ \left| \sqrt{mh_n}(\hat{\theta} - \theta(P)) - \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \psi_n(\mathbf{V}_i, P) \right| > \varepsilon \right\} \to 0, \tag{B.1}$$

$$\mathbb{E}_P \psi_n(\mathbf{V}_i, P) = 0 \ \forall P \in \mathcal{P}, \tag{B.2}$$

$$\sup_{P \in \mathcal{P}} |\mathbb{D}_P \psi_n(\mathbf{V}_i, P) - \xi^2(P)| \to 0, \tag{B.3}$$

Прокомментируем пункты предположения. Предположение B.1 требует линейной аппроксимации с помощью функций ψ_n от оценки $\hat{\theta}$ равномерно по всем смесям распределений. Предположение B.2 означает отсутствие смещения у ψ_n с подстановкой элементов выборки. Предположение B.3 обозначает наличие функций ξ_2, которые определяют предельную дисперсию ψ_n с подстановкой элементов выборки.

Отметим, что n_1 и n_2 вообще говоря случайны. Относительно этих величин дополнительно введем следующее предположение:

Предположение B.2 ([5]). Существует $0 < \lambda < 1$ такое, что $n_1/n \xrightarrow{p} \lambda$.

Рассматривается статистика T_n вида:

$$T_n = T_n(\overline{\mathbf{Z}}_n) = \sqrt{nh_n}(\hat{\theta}_1 - \hat{\theta}_2). \tag{B.4}$$

В [5] для статистик такого вида при соблюдении предположений B.1, B.2 доказана следующая теорема.

Теорема 1 ([5]). При выполнении перечисленных предположений статистика T_n асимптотически нормальна:

$$T_n - \sqrt{nh_n}(\theta(P_1) - \theta(P_2)) \xrightarrow{d} N(0, \sigma^2),$$

где $\sigma^2 = \xi^2(P_1)/\lambda + \xi^2(P_2)/(1-\lambda)$.

B.2. О построении перестановочного критерия

Двухсторонний перестановочный критерий строится на основе значений статистики T_n для всех перестановок n-элементного множества $\pi \in \mathbf{G}_n$, отвечающих перестановке наблюдений.

Пусть дана выборка $(\mathbf{Z}_i)_{i=1}^n$ из смеси двух распределений P_1 и P_2. Введем метки $(\mathbf{W}_i)_{i=1}^n$ со значениями в множестве $\{1, 2\}$, причем $\mathbf{W}_i = 1$, если $\mathbf{Z}_i$ имеет распределение P_1, иначе $\mathbf{W}_i = 2$. После определения меток, они фиксируются на своих местах. Можно считать, что первые n_1 наблюдений $\mathbf{Z}_1, \ldots, \mathbf{Z}_{n_1}$ имеют метку $\mathbf{W}_i = 1$. Полагая, что верна нулевая гипотеза

$$\mathbb{H}_0 : \theta(P_1) = \theta(P_2).$$

можем считать, что выборка взята из одного распределения. Статистика T_n также зависит от меток:

$$T_n = T_n(\overline{\mathbf{W}}_n, \overline{\mathbf{Z}}_n) = \sqrt{nh_n}(\hat{\theta}_1 - \hat{\theta}_2).$$

Обозначим $\overline{\mathbf{Z}}_n^\pi$ как результат перестановки наблюдений для некоторой перестановки $\pi \in \mathbf{G}_n$:

$$\overline{\mathbf{Z}}_n^\pi = \mathbf{Z}_{\pi(1)}, \ldots, \mathbf{Z}_{\pi(n_1)}, \mathbf{Z}_{\pi(n_1+1)}, \ldots, \mathbf{Z}_{\pi(n_2)}.$$

Алгоритм построения двухстороннего перестановочного критерия, предложенный в [5], имеет следующий вид:

1. Для каждой перестановки вычисляется $T_n^{\pi} = T_n(\overline{\mathbf{W}}_n, \overline{\mathbf{Z}}_n^{\pi})$;

2. Упорядочиваем их: $T_n^{(1)} \leq \cdots \leq T_n^{(n!)}$;

3. Задаем уровень значимости $\alpha \in (0,1)$;

4. Считаем следующие числа:

   а. $k_- = \lfloor n!\alpha/2 \rfloor$;

   б. $k_+ = n! - k_-$;

   в. $M_- = \#\left\{T_n^{(j)} : \ T_n^{(j)} < T_n^{(k_-)}\right\}$;

   г. $M_+ = \#\left\{T_n^{(j)} : \ T_n^{(j)} > T_n^{(k_+)}\right\}$;

   д. $M_0 = \#\left\{T_n^{(j)} : \ T_n^{(j)} = T_n^{(k_-)} \text{ или } T_n^{(j)} = T_n^{(k_+)}\right\}$;

   е. $a = (n!\alpha - (M_- + M_+))/M_0$;

5. Считаем тестовое значение φ:

$$\varphi(\overline{\mathbf{W}}_n, \overline{\mathbf{Z}}_n) = \begin{cases} 0, \text{ если } T_n^{(k_-)} < T_n < T_n^{(k_+)} \\ a, \text{ если } T_n = T_n^{(k_-)} \text{ или } T_n = T_n^{(k_+)} \\ 1, \text{ иначе} \end{cases}$$

Если $\varphi = 1$, гипотеза $\mathbb{H}_0$ отвергается; если $\varphi = 0$, то не отвергаем ее; если $\varphi = a$, то с вероятностью a отвергаем гипотезу $\mathbb{H}_0$.

Как видно из алгоритма, критическая область критерия определяется значениями $T_n^{(k_-)}$ и $T_n^{(k_+)}$. Для построенного критерия в [5] доказано, что при гипотезы равенства распределений $P_1 = P_2$, данный критерий является точным: $\alpha_I = \mathbb{E}\varphi(\overline{\mathbf{W}}_n, \overline{\mathbf{Z}}_n) = \alpha$.

К сожалению, из равенства $\theta(P_1) = \theta(P_2)$, которое полагает нулевая гипотеза, вообще говоря не следует выполнение гипотезы $P_1 = P_2$, потому в общем случае ошибка первого рода не контролируется в выборках конечного объема, что очень важно. В то же время, ошибка первого рода не контролируется и в асимптотическом случае [5]. Потому от статистики T_n переходят к стьюдентизированной статистике S_n:

$$S_n(\overline{\mathbf{W}}_n, \overline{\mathbf{Z}}_n) = \frac{T_n(\overline{\mathbf{W}}_n, \overline{\mathbf{Z}}_n)}{\widehat{\sigma}_n},$$

где $\widehat{\sigma}_n^2 = \frac{n}{n_1}\widehat{\xi_1^2} + \frac{n}{n_2}\widehat{\xi_2^2}$, $\widehat{\xi_k^2}$ — состоятельная оценка для ξ_k^2. При построении перестановочного критерия на основе S_n при верной нулевой гипотезе $\mathbb{H}_0$ ошибка первого рода контролируется асимптотически: $\alpha_I = \mathbb{E}\varphi(\overline{\mathbf{W}}_n, \overline{\mathbf{Z}}_n) \to \alpha$.