Аннотация

Объем работы – 46 страниц, состоит из введения, 3 глав основного содержания, заключения, списка литературы и приложения с программным кодом, включая 8 рисунков, 28 таблиц и списка литературы из 16 источников.

Объект исследования – критерии однородности Смирнова, Андерсона-Дарлинга, Лемана-Розенблатта, данные ограниченной точности, оценки мощности критериев.

Цель работы – исследование распределений статистик и мощности критериев однородности Смирнова, Лемана-Розенблатта, Андерсона-Дарлинга на данных ограниченной точности.

В результате работы были проведены исследования распределения статистик. Показано влияние близости функции распределения статистики к функции предельного распределения от размерности выборок и количества уникальных значений в объединенной выборке. Исследованы оценки мощности критериев на данных ограниченной точности.

Практическая ценность работы заключается в полученных результатах исследования критериев однородности на данных ограниченной точности. Сформулированы рекомендации по использованию критериев Андерсона-Дарлинга, Лемана-Розенблатта, Смирнова на данных ограниченной точности. Разработанная программа может применяться во многих сферах для решения прикладных задач, связанных с выявлением однородности данных.

Оглавление

Введение

Современное состояние и актуальность темы исследования.

В прикладных исследованиях довольно часто возникает необходимость выяснить, имеют ли различия генеральные совокупности, из которых взяты две независимые выборки. В математической статистике данная задача формулируется как проверка гипотезы об однородности законов распределения вероятностей. Необходимость проверки данных гипотез появляется в различных ситуациях, когда хотят удостовериться в неизменности (или напротив в изменении) статистических свойств некоторого объекта или процесса после целенаправленного изменения фактора или факторов (методики, технологии и т.д.), неявным образом влияющих на исследуемый объект. Иными словами, проверяется изменение или наоборот сохранение статистических показателей объекта или процесса до некоторого оказанного воздействия и после с течением времени. Например, надо выяснить, влияет ли способ упаковки некоторых деталей на заводе на их потребительские качества через год после хранения. Или другой пример применения исследований однородности: в маркетинге важно выделить сегменты потребительского рынка.

В случае если установлена однородность двух выборок, то вполне вероятно группировка сегментов, из которых они взяты, в один. В последующем это позволит, например, воплотить в жизнь по отношению к ним схожую рекламную политику (проводить одни и те же маркетинговые процедуры и т.п.). В случае если же установлено отличие, то поведение потребителей в двух сегментах различно, объединять эти сегменты невозможно, и могут понадобиться различные рекламные компании, своя для каждого из этих сегментов.

Для решения данной задачи широко используются критерии однородности. Критерии однородности призваны определить, взяты ли две (или более) выборки из одного распределения вероятностей. На данный момент

существуют множество таких критериев. Критерий однородности Смирнова предложен в работе [1] и рассмотрен в работах [2, 3]. В русскоязычной литературе трудно найти упоминания о критерии Андерсона-Дарлинга. Тем не менее, критерий однородности Андерсона-Дарлинга был подробно рассмотрен в работах [4, 5]. На ряду с критерием Смирнова на практике частое применение находит критерий Лемана-Розенблатта [6, 7].

На практике всегда приходится иметь дело с данными ограниченной точности. Зачастую, это целые числа, или данные с одним, двумя знаками после запятой. При больших объемах выборок, количество повторений в выборках становится большим. По сути, в этом случае мы имеем дело с некоторой дискретной случайной величиной. Становится интересно, можно ли руководствоваться результатами исследования критериев однородности для таких выборок. Подчиняются ли статистики критериев предельным распределениям, и при каких объемах выборок можно реально пользоваться этими предельными распределениями статистик критериев. Исследования распределений статистик и мощности критериев однородности подробно рассматривались в работах [8 - 11].

Цель и задачи исследований. Целью данной диссертационной работы является исследование распределений статистик и мощности критериев однородности на данных ограниченной точности.

Для достижения сформулированной цели были поставлены и решены следующие задачи:

- разработка программы для исследования методами статистического моделирования распределений статистик критериев и вычисления мощности критериев однородности;
- исследование распределений статистик критериев однородности Андерсона-Дарлинга, Лемана-Розенблатта, Смирнова на данных ограниченной точности;
- сравнительный анализ распределений статистик критериев с предельными функциями распределения;

– сравнительный анализ мощности рассматриваемых критериев на данных ограниченной точности.

Методы исследования. Для решения поставленных задач использовались методы статистического анализа, теории вероятности, математической статистики и компьютерного моделирования.

Научная новизна диссертационной работы заключается:

– в результатах исследований распределений статистик по данным ограниченной точности;
– в результатах исследования мощности критериев однородности на данных ограниченной точности и в сравнительном анализе с мощностями, полученными по выборкам без ограничений на точность.

Достоверность и обоснованность научных положений, рекомендаций и выводов подтверждается:

– корректным применением математического аппарата и методов статистического моделирования для исследования свойств и распределений статистик критериев;
– совпадением результатов статистического моделирования с известными теоретическими результатами.

Личный творческий вклад автора заключается:

– в формулировании этапов исследования распределений статистик рассматриваемых критериев однородности на данных ограниченной точности;
– в исследовании распределений статистик критериев однородности (проверка близости к предельной функции распределения);
– в вычислении мощности критериев на данных ограниченной точности и сравнение с мощностями на данных без округления;
– в реализации рассматриваемых критериев однородности на языке разработки программного обеспечения Python.

Практическая ценность и реализация результатов работы.

Полученные в работе результаты могут быть использованы в прикладных задачах статистического анализа в задачах по выявлению однородности. Сформулированы рекомендации по использованию критериев Андерсона-Дарлинга, Лемана-Розенблатта, Смирнова на данных ограниченной точности. Разработанная программа может применяться во многих сферах для решения прикладных задач, связанных с выявлением однородности данных.

Структура работы. Диссертация состоит из введения, 3 глав основного содержания, заключения, списка литературы и приложения с программным кодом. Основная часть содержания изложена на 10-44 страницах, включая 8 рисунков, 28 таблиц и списка литературы из 16 источников.

Краткое содержание работы. В первой главе представлены основные определения, необходимые теоретические выкладки, используемые в работе, формулируются задачи исследования.

Во второй главе исследуются распределения статистик критериев Смирнова, Андерсона-Дарлинга, Лемана-Розенблатта, полученные на данных ограниченной точности.

В третьей главе исследуются мощности вышеизложенных критериев.

В заключении приводится перечень основных результатов исследований, в приложении представлены фрагменты исходных текстов

1. Критерии проверки однородности законов распределения

1.1. Гипотеза об однородности распределений

При анализе случайных ошибок средств измерений, при статистическом управлении качеством процессов часто возникают вопросы решения задачи проверки гипотез о принадлежности двух выборок случайных величин одной и той же генеральной совокупности. Такая задача, естественно, возникает при проверке средств измерений, когда пытаются убедиться в том, что закон распределения случайных ошибок измерений не претерпел существенных изменений с течением времени.

Задача проверки однородности двух выборок формулируется следующим образом. Пусть имеются две выборки: $X_1, X_2, ..., X_n$ из распределения $F(x)$ и $Y_1, Y_2, ..., Y_m$ из распределения $G(x)$. Обозначим упорядоченные по неубыванию элементы выборок следующим образом:

$$x_1 \le x_2 \le ... \le x_m \text{ и } y_1 \le y_2 \le ... \le y_n.$$

Для определенности обычно полагают, что $m \le n$, но это совсем необязательно. Проверяется гипотеза о том, что обе выборки извлечены из одной и той же генеральной совокупности, т. е.

$$H_0: F(x) = G(x)$$

при любом x.

1.2. Критерий Смирнова

Критерий Смирнова – это правосторонний критерий проверки нулевой гипотезой о том, что из одного и того же непрерывного распределения извлекаются 2 независимые выборки. Критерий однородности Смирнова предложен в работе [1]. Предполагается, что функции распределения $F(x)$ и $G(x)$ являются непрерывными. Статистика критерия Смирнова измеряет расстояние между эмпирическими функциями распределения, построенными по выборкам [1]

$$D_{m,n} = \sup_x \left| G_m(x) - F_n(x) \right|.$$

На практике значение статистики $D_{m,n}$ рекомендуется вычислять в соответствии с соотношениями [8]:

$$D^{+}_{m,n} = \max_{1\le r\le m}\left[\frac{r}{m} - F_n(x_r)\right] = \max_{1\le s\le n}\left[G_m(y_s) - \frac{s-1}{n}\right],$$

$$D^{-}_{m,n} = \max_{1\le r\le m}\left[F_n(x_r) - \frac{r-1}{m}\right] = \max_{1\le s\le n}\left[\frac{s}{n} - G_m(y_s)\right],$$

$$D_{m,n} = \max\left(D^{+}_{m,n}, D^{-}_{m,n}\right).$$

Если гипотеза H_0 справедлива, то при неограниченном увеличении объемов выборок [12] $\lim\limits_{m\to\infty} P\left\{\sqrt{\dfrac{mn}{m+n}}D_{m,n} < S\right\} = K(S)$, т. е. статистика

$$S_{\text{C}} = \sqrt{\frac{mn}{m+n}}D_{m,n}$$

в пределе подчиняется распределению Колмогорова $K(S)$ [12] с функцией распределения

$$K(s) = \sum_{k=-\infty}^{\infty}(-1)^k e^{-2k^2s^2}\ .$$

Однако при ограниченных значениях m и n случайные величины $D^{+}_{m,n}$ и $D_{m,n}$ являются дискретными, и множество их возможных значений представляет собой решетку с шагом $1/k$, где k – наименьшее общее кратное m и n [12]. Условное распределение $G\left(S_{\text{C}}\middle|H_0\right)$ статистики S_{C} при верности гипотезы H_0 медленно сходится к $K(S)$ и имеет существенное отличие от него при малых значениях m и n.

Гладкость распределения статистики сильно зависит от величины k. Поэтому предпочтительнее применять критерий, когда объемы выборок m и n не равны и представляют собой взаимно простые числа. В таких случаях наименьшее общее кратное m и n максимально и равно $k = mn$, а распределе-

ние статистики больше напоминает непрерывную функцию распределения.

1.3. Критерий Лемана-Розенблатта

Критерий однородности Лемана–Розенблатта представляет собой критерий типа ω^2. Критерий предложен в работе [13] и исследован в [14]. Статистика критерия имеет вид [12]

$$T = \frac{mn}{m+n} \int_{-\infty}^{\infty} \left[G_m(x) - F_n(x) \right]^2 dH_{m+n}(x),$$

где $H_{m+n}(x) = \frac{m}{m+n} G_m(x) + \frac{n}{m+n} F_n(x)$ – эмпирическая функция распределения, построенная по вариационному ряду объединения двух выборок. Статистика T используется в форме [12]

$$T = \frac{1}{mn(m+n)} \left[n \sum_{i=1}^{n} \left(r_i - i \right)^2 + m \sum_{j=1}^{m} \left(s_j - j \right)^2 \right] - \frac{4mn-1}{6(m+n)}, \qquad (1.1)$$

где r_i – порядковый номер (ранг) y_i; s_j – порядковый номер (ранг) x_j в объединенном вариационном ряде.

В [15] было показано, что статистика (1.1) в пределе распределена как $a1(t)$:

$$\lim_{\substack{m\to\infty \\ n\to\infty}} P\{T < t\} = a1(t).$$

Функция распределения $a1(t)$ имеет вид [12]:

$$a1(t) = \frac{1}{\sqrt{2s}} \sum_{j=0}^{\infty} \frac{\Gamma(j+1/2)\sqrt{4j+1}}{\Gamma(1/2)\Gamma(j+1)} \exp\left\{ -\frac{(4j+1)^2}{16t} \right\} \times$$

$$\times \left\{ I_{-\frac{1}{4}}\left[\frac{(4j+1)^2}{16t} \right] - I_{\frac{1}{4}}\left[\frac{(4j+1)^2}{16t} \right] \right\},$$

где $I_{-\frac{1}{4}}(\cdot), I_{\frac{1}{4}}(\cdot)$ – модифицированные функции Бесселя вида

$$I_\nu(z)=\sum_{k=0}^{\infty}\frac{\left(\frac{z}{2}\right)^{\nu+2k}}{\Gamma(k+1)\Gamma(k+\nu+1)},\ |z|<\infty,\ |\arg z|<\pi\,.$$

В отличие от критерия Смирнова распределение статистики T быстро сходится к предельному $a1(T)$ [12].

1.4. Критерий Андерсона-Дарлинга

Двухвыборочный критерий Андерсона–Дарлинга (критерий однородности) рассмотрен в работе [16]. Статистика критерия определяется выражением

$$A^2=\frac{mn}{m+n}\int_{-\infty}^{\infty}\frac{[G_m(x)-F_n(x)]^2}{(1-H_{m+n}(x))H_{m+n}(x)}dH_{m+n}(x).$$

Для выборок непрерывных случайных величин выражение для этой статистики принимает простой вид [16]

$$A^2=\frac{1}{mn}\sum_{i=1}^{m+n-1}\frac{\left(M_i(m+n)-mi\right)^2}{i(m+n-i)},\qquad(1.2)$$

где M_i – число элементов первой выборки, меньших или равных i-му элементу вариационного ряда объединенной выборки.

Предельным распределением статистики (1.2) при справедливости проверяемой гипотезы H_0 является то же самое распределение $a2(t)$ [16], которое является предельным для статистики критерия согласия Андерсона–Дарлинга [12]. Функция распределения $a2(t)$, имеет вид [12]

$$a2(t)=\frac{\sqrt{2\pi}}{t}\sum_{j=0}^{\infty}(-1)^j\frac{\Gamma(j+1/2)(4j+1)}{\Gamma(1/2)\Gamma(j+1)}\exp\left\{-\frac{(4j+1)^2\pi^2}{8t}\right\}\times$$

$$\times\int_0^{\infty}\exp\left\{\frac{t}{8(y^2+1)}-\frac{(4j+1)^2\pi^2y^2}{8t}\right\}dy\,.$$

1.5. Выводы

В данной главе были представлены основные понятия и определения, описывающие процесс вычисления статистик критериев однородности Смирнова, Андерсона-Дарлинга и Лемана-Розенблатта. Описаны предельные распределения, которым подчиняются распределения статистик данных критериев.

В соответствие с поставленной целью в работе необходимо выполнить:

1) исследование распределений статистик критериев однородности Андерсона-Дарлинга, Лемана-Розенблатта, Смирнова на данных ограниченной точности;
2) сравнительный анализ распределений статистик критериев с предельными функциями распределения при различных объемах выборок;
3) сравнительный анализ мощности рассматриваемых критериев на данных ограниченной точности и сравнение с мощностью критериев по данным без ограничения точности;
4) разработку программы для исследования методами статистического моделирования распределений статистик критериев и вычисления мощности критериев однородности.

2. Исследование распределений статистик критериев однородности на данных ограниченной точности

2.1. Исследование распределений статистик

Так как цель исследования заключается в исследовании распределения статистик на данных ограниченной точности, нужно моделировать такие данные. Значения моделируемых выборок ограничивались до целого числа, до одного, двух знаков после запятой: сначала генерируется выборка заданного размера и производится округление значений.

Целью данной главы является проведение исследования, с целью выяснить, можно ли использовать критерии, если данные получены с ограниченной точностью, подчиняются ли статистики, вычисленные по таким данным, соответствующим предельным законам распределения рассматриваемых критериев однородности.

Зададимся величиной расстояния, равной 0.05, при котором будем считать, что распределение статистик все еще подчиняется предельному закону распределения.

Обозначим некоторые величины для таблиц с результатами исследований:

- количество выборок $N = 16600$,
- $\rho = \sup_x |F_n(x) - F(x)|$ - расстояние между эмпирическими и предельными функциями распределения статистик критерия в метрике Колмогорова.

2.2. Исследование распределения статистики критерия Смирнова

В таблицах 2.1, 2.2 исследования проводились на сгенерированных данных, обе выборки, в которых, подчинялись стандартному нормальному закону распределения с плотностью

$$f(x) = \frac{1}{\theta_1\sqrt{2\pi}} \exp\left\{-\frac{(x-\theta_0)^2}{2\theta_1^2}\right\}$$

и параметрами сдвига $\theta_0 = 0$ и масштаба $\theta_1 = 1$.

Таблица 2.1 – Результаты для критерия однородности Смирнова, округление до 2 знаков, n=m, выборки из нормального закона распределения с параметрами $\theta_0 = 0, \theta_1 = 1$.

n, m	ρ	среднее число различных значений в объединенной выборке
200, 200	0.09	246.5
500, 500	0.07	368.5
1000, 1000	0.07	449.0
2000, 2000	0.07	507.0
5000, 5000	0.06	580.5
10000, 10000	0.05	626.5

На рисунке 2.1 представлена графическая иллюстрация результатов, представленных в таблице 2.1.

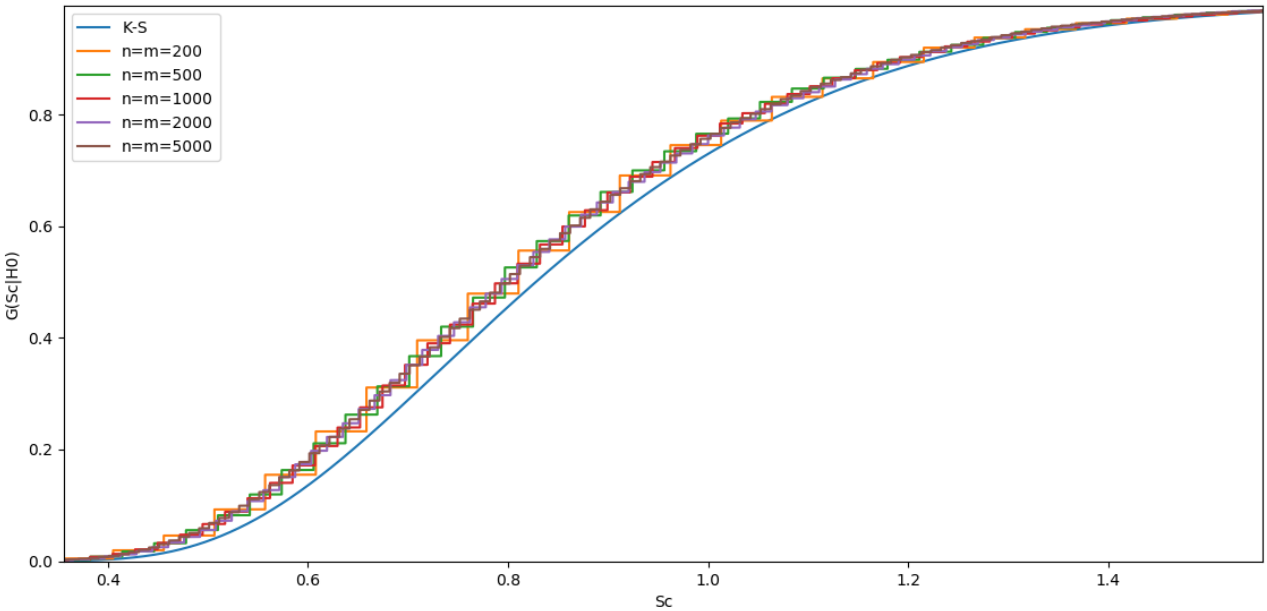

Рисунок 2.1 – Распределения статистики критерия Смирнова при справедливости H0 в зависимости от m и n, округление до 2 знаков

Как видно из таблицы 2.1, наблюдается уменьшение расстояния с ростом объема выборок при одинаковых размерах обеих выборок, в отличие от других критериев. В связи с этим были проведены дополнительные исследования критерия Смирнова на объемах выборок 10000. Даже при таких размерах моделируемых выборок расстояние имеет тенденцию к уменьшению. И тем не менее, заданное расстояние 0.05 между функциями распределения начинает достигаться лишь при объеме выборок 10000.

В таблице 2.2 обе выборки также принадлежали стандартному нормальному закону распределения, но при различных объемах выборок.

Таблица 2.2 – Результаты для критерия однородности Смирнова, округление до 2 знаков, $n \neq m$, выборки из нормального закона распределения с параметрами $\theta_0 = 0, \theta_1 = 1$.

n, m	ρ	среднее число различных значений в объединенной выборке
500, 500	0.07	368.5
500, 1000	0.07	421.05
500, 2000	0.06	468.0
500, 5000	0.05	533.5

Суммируя результаты по таблице 2.2, можно заметить, что при различных объемах выборок, с увеличением объема второй выборки и при зафиксированном значении объема первой, расстояния оказываются меньшими, чем когда объемы двух выборок одинаковые (табл. 2.1).

В предыдущих исследованиях было замечено, что расстояния между эмпирической функцией распределения и предельной функцией распределения статистики критерия оказывались неприемлемо большими на данных ограниченных до целых чисел и до одного знака. Это могло быть связанно с большим количеством повторений в выборке. Поэтому, для данных, ограниченных до целых чисел и одного знака, были проведены исследования на данных с большим количеством уникальных значений при тех же объемах выборок, что и в исследовании на данных ограниченных до двух знаков. С этой целью, выборки генерировались из распределения, с большей дисперсией, чем стандартное нормальное. Величина дисперсии подбиралась эмпирическим путем, чтобы ее величина была максимально приближена к единице и, чтобы расстояние не превышало 0.05.

Таблица 2.3 – Результаты для критерия однородности Смирнова, округление до одного знака, n=m, выборки из нормального закона распределения с параметрами $\theta_0 = 0, \theta_1 = 50$.

n, m	ρ	среднее число различных значений в объединенной выборке
200, 200	0.08	360.5
500, 500	0.05	772.5
1000, 1000	0.04	1228.5
2000, 2000	0.04	1719.5
5000, 5000	0.03	2243.0
10000, 10000	0.03	2546.0

Таблица 2.4 – Результаты для критерия однородности Смирнова, округление до целых, n=m, выборки из нормального закона распределения с параметрами $\theta_0 = 0, \theta_1 = 100$.

n, m	ρ	среднее число различных значений в объединенной выборке
200, 200	0.1	243.5
500, 500	0.08	368.5
1000, 1000	0.08	448.5
2000, 2000	0.07	508.5
5000, 5000	0.06	578.0
10000, 10000	0.06	628.0

Для данных, ограниченных до одного знака и до целых (табл. 2.3, 2.4), при увеличении количества уникальных значений, за счет увеличения дисперсии закона распределения моделируемых выборок, в объединенной выборке расстояния становятся схожими с результатами, полученными на данных, ограниченных до двух знаков (табл. 2.1).

В силу особенностей критерия Смирнова, упомянутых в главе 1, было необходимо провести исследования на выборках, размеры которых представляются как взаимно простые числа. Объемы выборок подбирались с максимальной схожестью объемов выборок из предыдущих исследований.

Таблица 2.5 – Результаты для критерия однородности Смирнова, округление до двух знаков, объемы выборок взаимно простые, выборки из нормального закона распределения с параметрами $\theta_0 = 0, \theta_1 = 1$.

n, m	ρ	среднее число различных значений в объединенной выборке
199, 201	0.05	248.5
499, 501	0.06	375.5
999, 1001	0.05	455.5
1999, 2001	0.06	507.5
4999, 50001	0.06	573.5
9999, 10001	0.06	624.0

Для взаимно простых n и m расстояния от функции распределения статистик до предельного не имеют существенных отличий в сравнении с предыдущими исследованиями из табл. 2.1.

2.3. Исследование распределения статистики критерия Лемана-Розенблатта

В таблицах ниже (2.6-2.12) представлены значения расстояний между эмпирическими и предельными функциями распределения статистик, рассчитанные по метрике Колмогорова для критерия Лемана-Розенблатта.

В таблицах 2.6, 2.7 обе выборки также принадлежали стандартному нормальному закону распределения, но при различных объемах выборок.

Таблица 2.6 – Результаты для критерия однородности Лемана-Розенблатта, округление до 2 знаков, n=m, выборки из нормального закона распределения с параметрами $\theta_0 = 0, \theta_1 = 1$.

n, m	ρ	среднее число различных значений в объединенной выборке
200, 200	0.01	243.5
500, 500	0.01	369.5
1000, 1000	0.01	448.5
2000, 2000	0.01	510.5
5000, 5000	0.01	578.5

На рисунке 2.2 представлена графическая иллюстрация результатов, представленных в таблице 2.6.

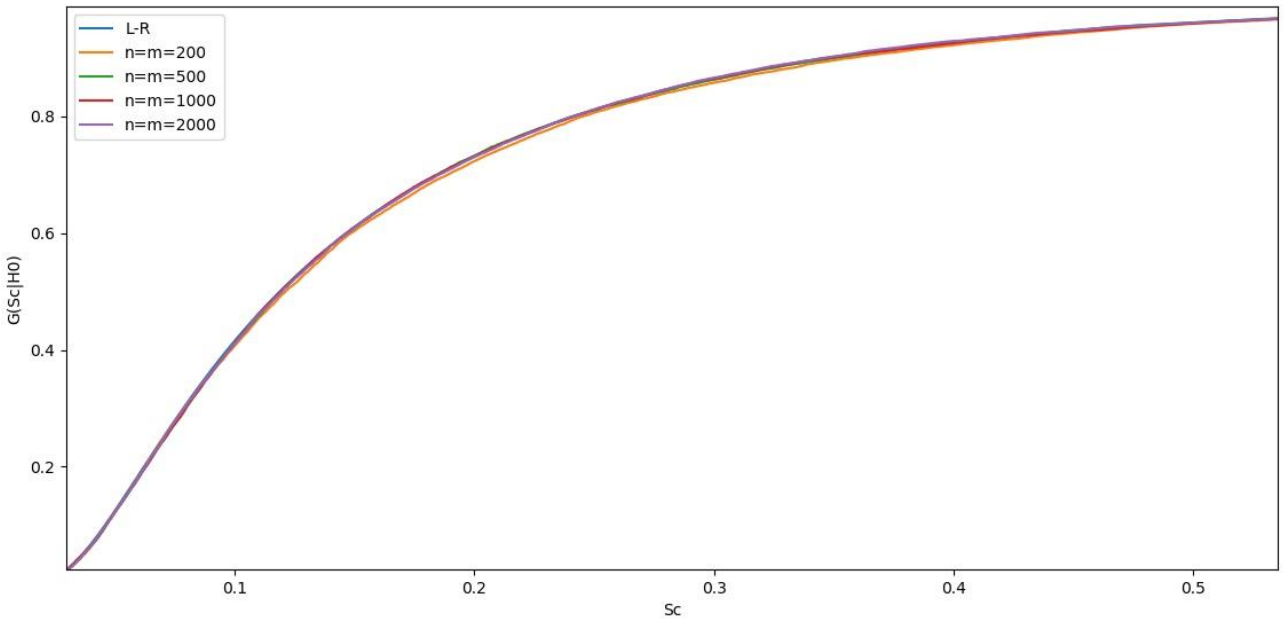

Рисунок 2.2 – Распределения статистики критерия Лемана-Розенблатта при справедливости H0 в зависимости от m и n, округление до 2 знаков

Таблица 2.7 – Результаты для критерия однородности Лемана-Розенблатта, округление до 1 знака, n=m, выборки из нормального закона распределения с параметрами $\theta_0 = 0, \theta_1 = 1$.

n, m	ρ	среднее число различных значений в объединенной выборке
200, 200	0.01	49.0
500, 500	0.01	56.5
1000, 1000	0.01	62.5
2000, 2000	0.01	67.5
5000, 5000	0.01	72.5

Судя по результатам из таблиц 2.6 и 2.7, распределение статистик для критерия Лемана-Розенблатта довольно близко располагается с предельным распределением. Для выборок, округленных до двух и одного знаков, выполняется условие не превышения расстояния в 0.05.

В таблицах 2.8-2.11 обе выборки также принадлежали стандартному нормальному закону распределения, но при различных объемах выборок.

Таблица 2.8 – Результаты для критерия однородности Лемана-Розенблатта, округление до 2 знаков, $n \neq m$, выборки из нормального закона распределения с параметрами $\theta_0 = 0, \theta_1 = 1$, при малых объемах выборок.

n, m	ρ	среднее число различных значений в объединенной выборке
30, 30	0.02	55.5
30, 40	0.01	63.0
30, 50	0.01	71.5

Таблица 2.9 – Результаты для критерия однородности Лемана-Розенблатта, округление до 2 знаков, $n \neq m$, выборки из нормального закона распределения с параметрами $\theta_0 = 0, \theta_1 = 1$, при больших объемах выборок.

n, m	ρ	среднее число различных значений в объединенной выборке
500, 500	0.01	369.5
500, 1000	0.01	418.0
500, 2000	0.03	469.0
500, 5000	0.24	535

Таблица 2.10 – Результаты для критерия однородности Лемана-Розенблатта, округление до 1 знака, $n \neq m$, выборки из нормального закона распределения с параметрами $\theta_0 = 0, \theta_1 = 1$, при малых объемах выборок.

n, m	ρ	среднее число различных значений в объединенной выборке
30, 30	0.02	30.5
30, 40	0.02	33.0
30, 50	0.03	34.5

Таблица 2.11 – Результаты для критерия однородности Лемана-Розенблатта, округление до 1 знака, $n \neq m$, выборки из нормального закона распределения с параметрами $\theta_0 = 0, \theta_1 = 1$, при больших объемах выборок.

n, m	ρ	среднее число различных значений в объединенной выборке
500, 500	0.01	56.5
500, 1000	0.35	60.0
500, 2000	0.94	63.0
500, 5000	0.99	70.0

Судя по результатам данных таблиц для критерия Лемана-Розенблатта не

наблюдается приближения распределения статистик к предельному закону при различных объемах выборок в сравнении с результатами, полученными при одинаковых объемах выборок. Для таблицы 2.11 эти выводы проявляются в наибольшей степени.

Для данных, ограниченных до целых чисел, были проведены исследования на данных с большим количеством уникальных значений при тех же объемах выборок, что и в исследовании на данных ограниченных до двух и одного знаков. С этой целью, выборки генерировались из распределения, с большей дисперсией, чем стандартное нормальное.

Таблица 2.12 – Результаты для критерия однородности Лемана-Розенблатта, округление до целых, n=m, выборки из нормального закона распределения с параметрами $\theta_0 = 0, \theta_1 = 10$.

n, m	ρ	среднее число различных значений в объединенной выборке
200, 200	0.01	51.0
500, 500	0.01	57.0
1000, 1000	0.01	63.5
2000, 2000	0.01	67.5
5000, 5000	0.01	69.0

Для данных, ограниченных до целых, при увеличении количества уникальных значений, за счет увеличения дисперсии закона распределения моделируемых выборок, в объединенной выборке расстояния становятся схожими с результатами, полученными на данных, ограниченных до одного и двух знаков.

2.4. Исследование распределения статистики критерия Андерсона-Дарлинга

В таблицах ниже (2.13-2.18) представлены значения расстояний между эмпирическими и предельными функциями распределения статистик, рассчитанные по метрике Колмогорова для критерия Андерсона-Дарлинга.

В таблицах 2.13-2.16 исследования проводились на сгенерированных

данных, обе выборки, в которых, подчинялись стандартному нормальному закону.

Таблица 2.13 – Результаты для критерия однородности Андерсона-Дарлинга, округление до 2 знаков, n=m, выборки из нормального закона распределения с параметрами $\theta_0 = 0, \theta_1 = 1$.

n, m	ρ	среднее число различных значений в объединенной выборке
200, 200	0.02	241.0
500, 500	0.02	377.0
1000, 1000	0.03	442.0
2000, 2000	0.04	510.0
5000, 5000	0.08	576.5

Как видно из таблицы, с увеличением объемов выборок расстояние между эмпирической функцией распределения и предельной функцией распределения статистики критерия увеличивалось. По результатам, представленным в таблице 2.13, видно, что между $n = m = 2000$ и $n = m = 5000$ расстояние становится большим чем 0.05 на данных, округленных до двух знаков.

На рисунке 2.3 представлена графическая иллюстрация результатов, представленных в таблице 2.3.

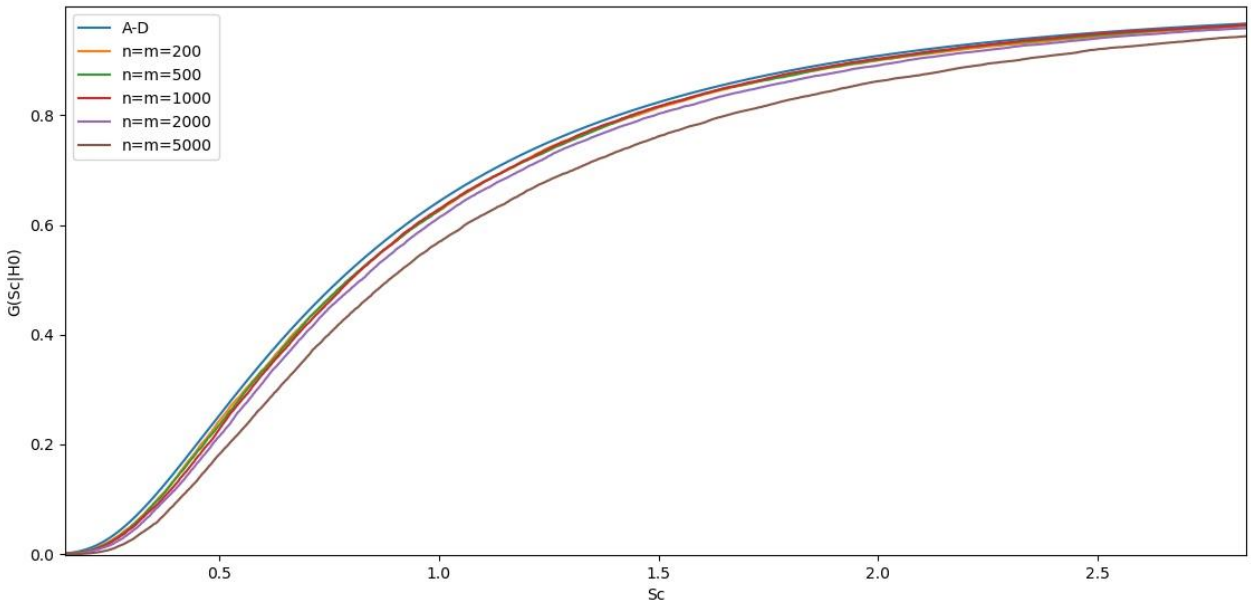

Рисунок 2.3 – Распределения статистики критерия Андерсона-Дарлинга при справедливости H0 в зависимости от m и n, округление до 2 знаков

При округлении до целых и до одного знака после запятой наблюдалась такая же тенденция увеличения расстояния с увеличением объемов выборок.

Но величина расстояния была около единицы и около 0.5 соответственно, что является показателем, что функции распределения лежат далеко друг от друга.

В таблице 2.14 обе выборки также принадлежали стандартному нормальному закону распределения, но при различных объемах выборок.

Таблица 2.14 – Результаты для критерия однородности Андерсона-Дарлинга, округление до 2 знаков, $n \neq m$, выборки из нормального закона распределения с параметрами $\theta_0 = 0, \theta_1 = 1$.

n, m	ρ	среднее число различных значений в объединенной выборке
500, 500	0.02	377.0
500, 1000	0.01	422.0
500, 2000	0.01	465.0
500, 5000	0.01	532.5

Из результатов таблицы 2.14 наблюдается схожая картина с аналогичными исследованиями критерия Смирнова. При различных объемах выборок, с увеличением объема второй выборки при зафиксированном значении объема первой, расстояния оказываются меньшими, чем когда объемы двух выборок одинаковые (табл. 2.13).

Для данных, ограниченных до целых чисел и одного знака, были проведены исследования на данных с большим количеством уникальных значений при тех же объемах выборок, что и в исследовании на данных ограниченных до двух знаков. С этой целью, выборки генерировались из распределения, с большей дисперсией, чем стандартное нормальное.

Таблица 2.15 – Результаты для критерия однородности Андерсона-Дарлинга, округление до 1 знака, n=m, выборки из нормального закона распределения с параметрами $\theta_0 = 0, \theta_1 = 10$.

n, m	ρ	среднее число различных значений в объединенной выборке
200, 200	0.02	249.0
500, 500	0.02	374.5
1000, 1000	0.02	442.5
2000, 2000	0.04	503.5
5000, 5000	0.09	579.0

Таблица 2.16 – Результаты для критерия однородности Андерсона-Дарлинга, округление до целых, n=m, выборки из нормального закона распределения с параметрами $\theta_0 = 0, \theta_1 = 80$.

n, m	ρ	среднее число различных значений в объединенной выборке
200, 200	0.01	221.0
500, 500	0.03	321.0
1000, 1000	0.04	374.0
2000, 2000	0.06	421.5
5000, 5000	0.12	475.0

Анализируя результаты, представленные в таблицах для критерия Андерсона-Дарлинга, можно заметить тенденцию, что при уменьшении отношения числа различных значений в объединенной выборке к общему объему объединенной выборки, увеличивается расстояние между распределениями эмпирической функции распределения статистик и предельным распределением.

2.5. Выводы

Суммируя полученные результаты исследования распределения статистик по всем критериям на данных ограниченной точности, можно сделать следующие выводы:

- для критерия Смирнова наблюдается сходимость распределения статистики к предельному распределению статистики при увеличении объема выборок. Также, было замечено, что с увеличением объема второй выборки и при зафиксированном значении объема первой, расстояния оказываются меньшими, чем когда объемы двух выборок одинаковые. Для данных, ограниченных до одного знака и до целых, при увеличении количества уникальных значений, за счет увеличения дисперсии закона распределения моделируемых выборок, в объединенной выборке расстояния

становятся схожими с результатами, полученными на данных, ограниченных до двух знаков.

- для критерия Лемана-Розенблатта была замечена следующая особенность, что при зафиксированном распределении генерируемых выборок и при различных $n = m$ $(200, 500, \ldots)$ не меняется расстояние между эмпирическим распределением статистики и предельным. Расстояние меняется лишь при изменении дисперсии распределения выборок.
- для критерия Андерсона-Дарлинга можно заметить тенденцию, что при уменьшении числа различных значений в объединенной выборке к общему объему объединенной выборки, увеличивается расстояние между эмпирической функцией распределения статистик и предельным распределением. При более точном исследовании можно попытаться получить предельное соотношение числа уникальных элементов в объединенной выборке к общему числу элементов при заданном расстоянии между эмпирической функцией распределения и предельной функцией распределения статистики критерия Андерсона-Дарлинга.

3. Исследование мощностей критериев однородности на данных ограниченной точности

3.1. Исследование мощностей критериев

Очевидно, что при проверке любой статистической гипотезы предпочтительней использовать наиболее мощный критерий. Статистическая мощность в математической статистике является показателем вероятности отклонения основной (или нулевой) гипотезы при проверке статистических гипотез в случае, когда нулевая гипотеза неверна (верна альтернативная гипотеза).

Мощность критериев проверки однородности исследовалась в случае ряда альтернатив. Для определенности проверяемой гипотезе H_0 соответствовала принадлежность выборок одному и тому же стандартному нормальному закону распределения с плотностью

$$f(x) = \frac{1}{\theta_1\sqrt{2\pi}} \exp\left\{-\frac{(x-\theta_0)^2}{2\theta_1^2}\right\}$$

и параметрами сдвига $\theta_0 = 0$ и масштаба $\theta_1 = 1$.

При всех конкурирующих гипотезах первая выборка всегда соответствовала стандартному нормальному закону, а вторая – некоторому другому.

В частности, при альтернативе сдвига в случае конкурирующей гипотезы H_1 вторая выборка соответствовала нормальному закону с параметром сдвига $\theta_0 = 0.1$ и параметром масштаба $\theta_1 = 1$, в случае конкурирующей гипотезы H_2 – нормальному закону с параметрами $\theta_0 = 0.5$ и $\theta_1 = 1$.

При изменении масштаба в случае конкурирующей гипотезы H_3 вторая выборка соответствовала нормальному закону с параметрами $\theta_0 = 0$ и $\theta_1 = 1.1$, в случае конкурирующей гипотезы H_4 – нормальному закону с параметрами $\theta_0 = 0$ и $\theta_1 = 1.5$.

В случае конкурирующей гипотезы H_5 вторая выборка соответствовала

логистическому закону с плотностью

$$f(x)=\exp\left\{-\frac{(x-\theta_0)}{\theta_1}\right\}\Bigg/\left[1+\exp\left\{-\frac{(x-\theta_0)}{\theta_1}\right\}\right]^2$$

и параметрами $\theta_0=0$ и $\theta_1=1$.

На рисунках ниже (3.1 – 3.5) представлены графики теоретических распределений, из которых генерировались выборки для исследования мощностей.

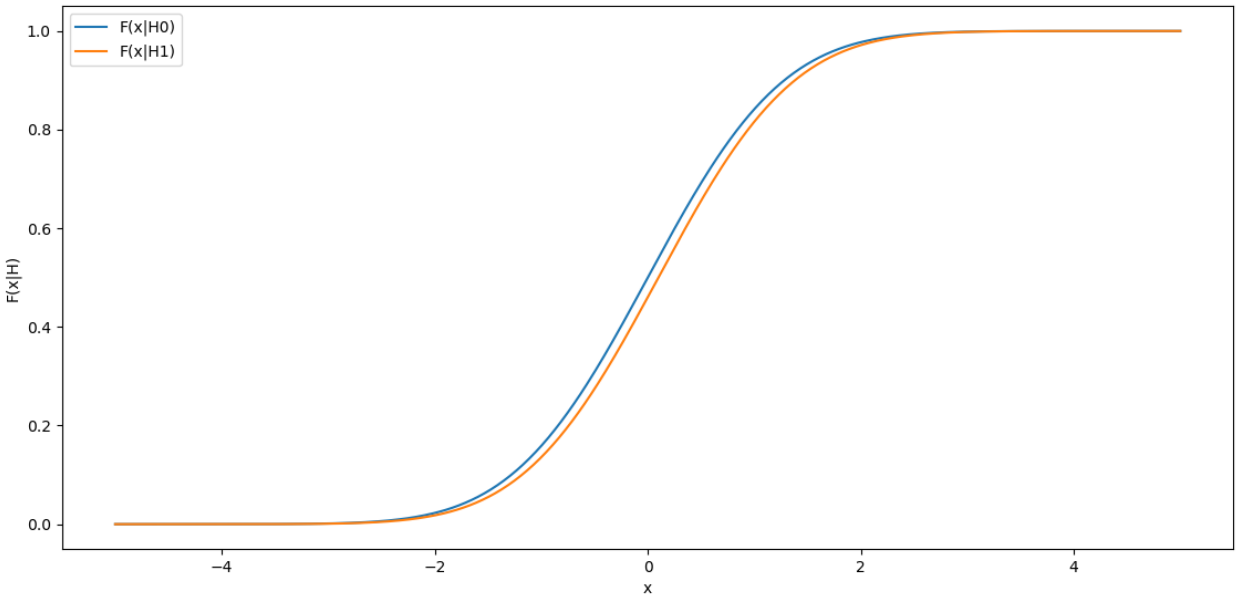

Рисунок 3.1 – Графики функций распределения законов, из которых генерировалась вторая выборка в гипотезах $H0, H1$.

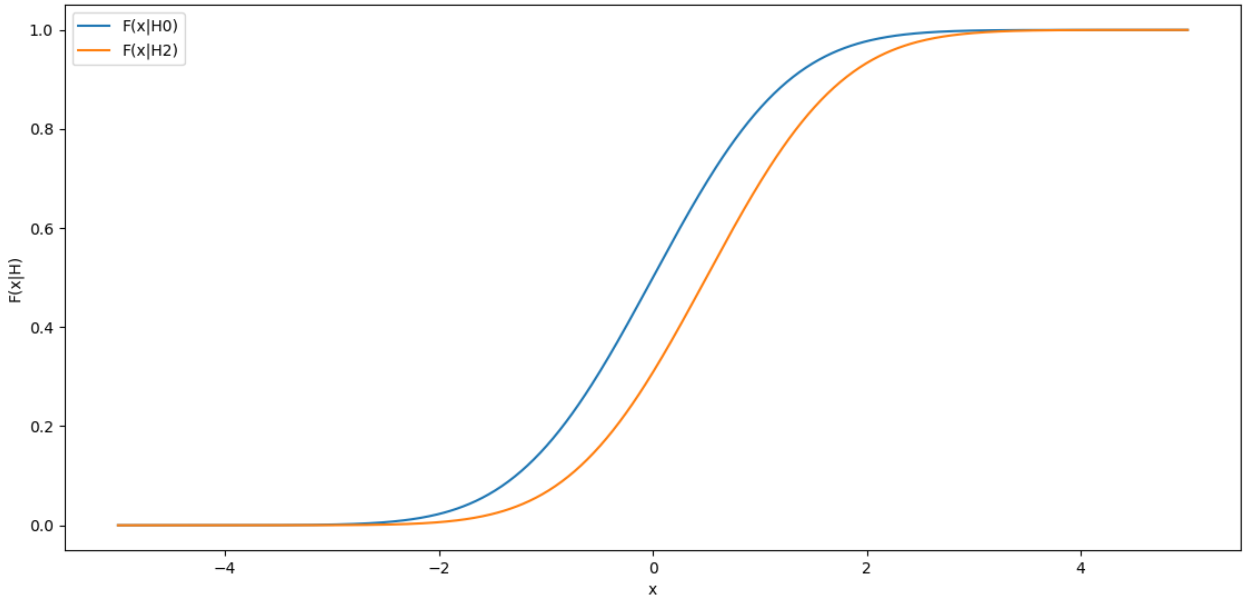

Рисунок 3.2 – Графики функций распределения законов, из которых генерировалась вторая выборка в гипотезах $H0, H2$.

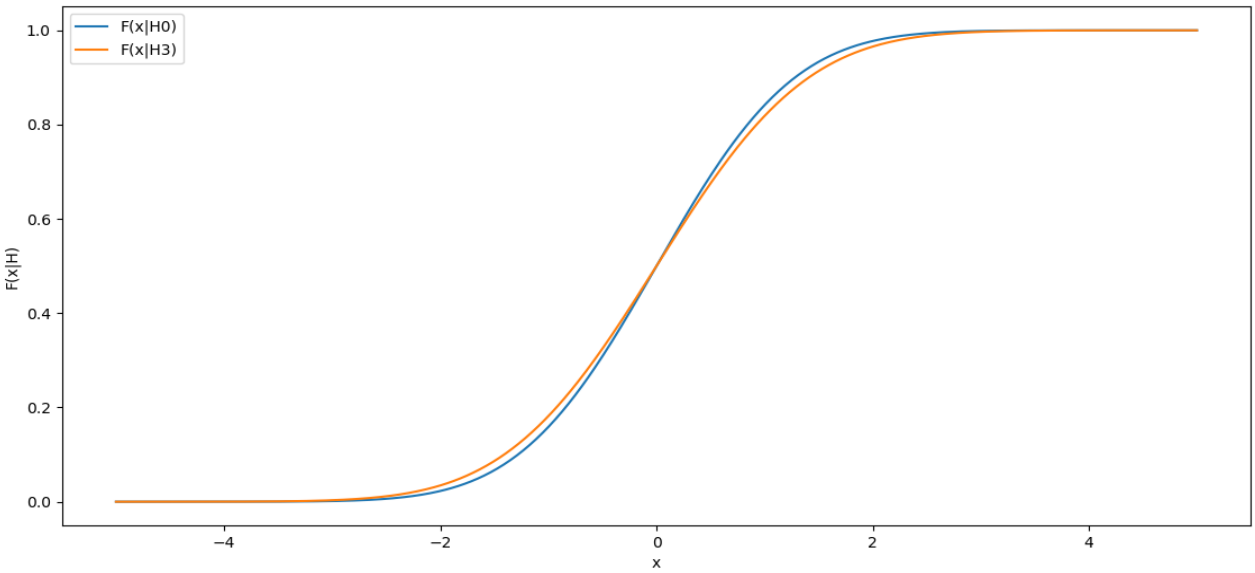

Рисунок 3.3 – Графики функций распределения законов, из которых генерировалась вторая выборка в гипотезах $H0, H3$.

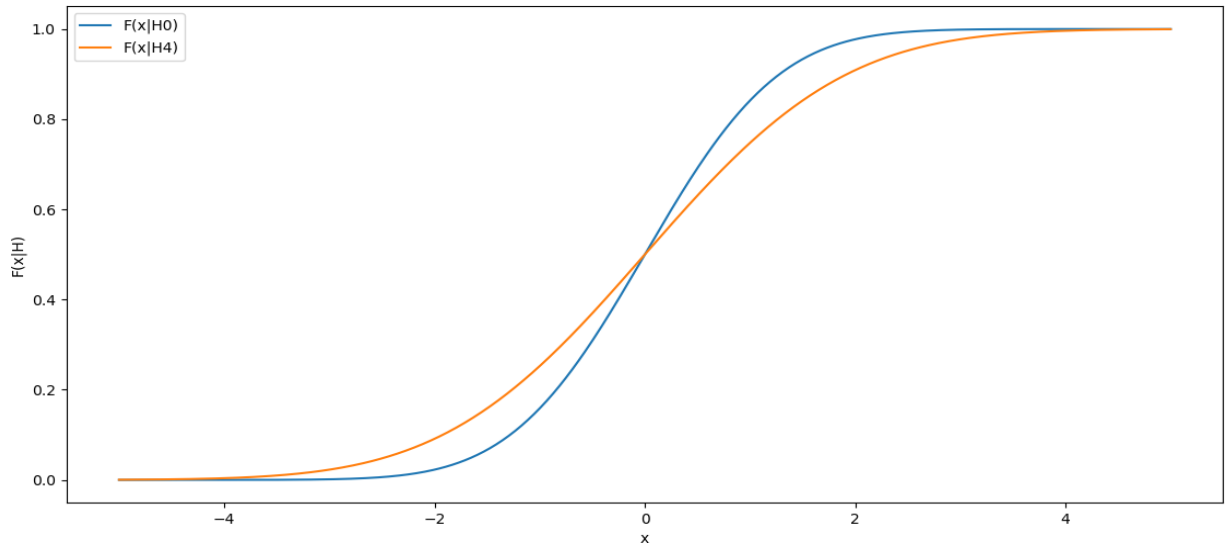

Рисунок 3.4 – Графики функций распределения законов, из которых генерировалась вторая выборка в гипотезах $H0, H4$.

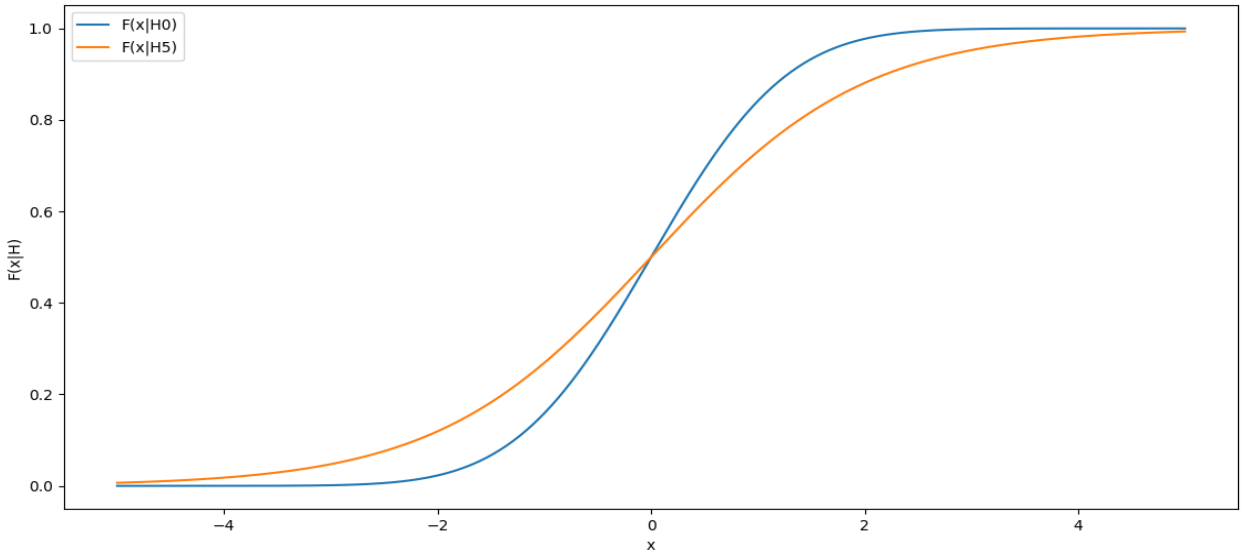

Рисунок 3.5 – Графики функций распределения законов, из которых генерировалась вторая выборка в гипотезах $H0, H5$.

3.2. Исследование мощности критерия Смирнова

В таблицах 3.1 – 3.4 представлены рассчитанные оценки мощностей критерия однородности Смирнова ($1-\beta$, где β - вероятность ошибки второго рода). Значения представлены относительно конкурирующих гипотез $H_1 - H_5$ для различных значений объемов генерируемых выборок. Значения оценок мощности также представлены в зависимости от различных значений заданных уровней значимости (вероятностей ошибок первого рода): $\alpha = 0.1, 0.05, 0.25.$

Таблица 3.1 – Мощность критерия Смирнова относительно гипотез $H_1 - H_5$ в зависимости от объемов выборок $(n = m)$ на данных без округления

Уровень значимости α	n = 500	n = 1000	n = 2000
Относительно альтернативы H_1			
0.1	0.38	0.61	0.87
0.05	0.27	0.48	0.78
0.025	0.18	0.36	0.69
Относительно альтернативы H_2			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0
Относительно альтернативы H_3			
0.1	0.19	0.30	0.55
0.05	0.10	0.17	0.36
0.025	0.05	0.09	0.21
Относительно альтернативы H_4			
0.1	0.99	1.0	1.0
0.05	0.99	1.0	1.0
0.025	0.98	1.0	1.0
Относительно альтернативы H_5			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	0.99	1.0	1.0

Таблица 3.2 – Мощность критерия Смирнова относительно гипотез $H_1 - H_5$ в зависимости от объемов выборок $(n = m)$ на данных, округленных до целых чисел

Уровень значимости α	n = 500	n = 1000	n = 2000
Относительно альтернативы H_1			
0.1	0.37	0.58	0.84
0.05	0.25	0.45	0.74
0.025	0.17	0.34	0.63
Относительно альтернативы H_2			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0
Относительно альтернативы H_3			
0.1	0.19	0.30	0.54
0.05	0.11	0.17	0.33
0.025	0.06	0.10	0.20
Относительно альтернативы H_4			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	0.99	1.0	1.0
Относительно альтернативы H_5			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0

Таблица 3.3 – Мощность критерия Смирнова относительно гипотез $H_1 - H_5$ в зависимости от объемов выборок $(n = m)$ на данных, округленных до 1 знака после запятой

Уровень значимости α	n = 500	n = 1000	n = 2000
Относительно альтернативы H_1			
0.1	0.39	0.61	0.87
0.05	0.28	0.49	0.78
0.025	0.19	0.37	0.69
Относительно альтернативы H_2			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0

Продолжение таблицы 3.3

Относительно альтернативы H_3			
0.1	0.20	0.31	0.57
0.05	0.11	0.18	0.37
0.025	0.05	1.0	0.22
Относительно альтернативы H_4			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	0.99	1.0	1.0
Относительно альтернативы H_5			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0

Таблица 3.4 – Мощность критерия Смирнова относительно гипотез $H_1 - H_5$ в зависимости от объемов выборок $(n = m)$ на данных, округленных до 2 знаков после запятой

Уровень значимости α	n = 500	n = 1000	n = 2000
Относительно альтернативы H_1			
0.1	0.38	0.60	0.87
0.05	0.27	0.47	0.78
0.025	0.18	0.37	0.68
Относительно альтернативы H_2			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0
Относительно альтернативы H_3			
0.1	0.19	0.29	0.56
0.05	0.11	0.16	0.35
0.025	0.05	0.09	0.27
Относительно альтернативы H_4			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	0.98	1.0	1.0
Относительно альтернативы H_5			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0

Суммируя результаты, полученные по таблицам 3.1 – 3.4, можно сказать, что на округленных данных мощность получалась выше для гипотезы $H3$ почти во всех случаях.

3.3. Исследование мощности критерия Лемана-Розенблатта

В таблицах 3.5 – 3.8 представлены рассчитанные оценки мощностей критерия однородности Лемана-Розенблатта. Значения оценок мощности представлены относительно конкурирующих гипотез $H_1 - H_5$ для различных значений объемов выборок, также в зависимости от различных значений заданных уровней значимости: $\alpha = 0.1, 0.05, 0.25$.

Таблица 3.5 – Мощность критерия Лемана-Розенблатта относительно гипотез $H_1 - H_5$ в зависимости от объемов выборок $(n = m)$ на данных без округления

Уровень значимости α	n = 500	n = 1000	n = 2000
Относительно альтернативы H_1			
0.1	0.44	0.68	0.91
0.05	0.32	0.56	0.85
0.025	0.23	0.44	0.78
Относительно альтернативы H_2			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0
Относительно альтернативы H_3			
0.1	0.19	0.32	0.62
0.05	0.10	0.16	0.41
0.025	0.05	0.07	0.23
Относительно альтернативы H_4			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0
Относительно альтернативы H_5			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0

Таблица 3.6 – Мощность критерия Лемана-Розенблатта относительно гипотез $H_1 - H_5$ в зависимости от объемов выборок $(n = m)$ на данных, округленных до целых чисел

Уровень значимости α	n = 500	n = 1000	n = 2000
Относительно альтернативы H_1			
0.1	0.41	0.63	0.88
0.05	0.30	0.52	0.81
0.025	0.22	0.41	0.73
Относительно альтернативы H_2			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0
Относительно альтернативы H_3			
0.1	0.21	0.32	0.60
0.05	0.18	0.18	0.40
0.025	0.05	0.10	0.25
Относительно альтернативы H_4			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0
Относительно альтернативы H_5			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0

Таблица 3.7 – Мощность критерия Лемана-Розенблатта относительно гипотез $H_1 - H_5$ в зависимости от объемов выборок $(n = m)$ на данных, округленных до 1 знака

Уровень значимости α	n = 500	n = 1000	n = 2000
Относительно альтернативы H_1			
0.1	0.44	0.68	0.91
0.05	0.32	0.56	0.84
0.025	0.22	0.45	0.77
Относительно альтернативы H_2			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0

Продолжение таблицы 3.7

Относительно альтернативы H_3			
0.1	0.20	0.32	0.62
0.05	0.10	0.17	0.40
0.025	0.05	0.08	0.24
Относительно альтернативы H_4			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0
Относительно альтернативы H_5			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0

Таблица 3.8 – Мощность критерия Лемана-Розенблатта относительно гипотез $H_1 - H_5$ в зависимости от объемов выборок $(n = m)$ на данных, округленных до 2 знаков после запятой

Уровень значимости α	n = 500	n = 1000	n = 2000
Относительно альтернативы H_1			
0.1	0.44	0.68	0.91
0.05	0.32	0.57	0.85
0.025	0.23	0.46	0.77
Относительно альтернативы H_2			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0
Относительно альтернативы H_3			
0.1	0.19	0.33	0.61
0.05	0.09	0.17	0.40
0.025	0.05	0.08	0.22
Относительно альтернативы H_4			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0
Относительно альтернативы H_5			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0

Суммируя результаты, полученные по оценкам мощностей критерия Лемана-Розенблатта, можно сказать, что на округленных данных мощность получалась выше для гипотезы $H3$ почти во всех случаях.

3.4. Исследование мощности критерия Андерсона-Дарлинга

В таблицах 3.9 – 3.12 представлены рассчитанные оценки мощностей критерия однородности Андерсона-Дарлинга. Значения представлены относительно конкурирующих гипотез $H_1 - H_5$. Значения оценок мощности также представлены в зависимости от различных значений заданных уровней значимости: $\alpha = 0.1, 0.05, 0.25$.

Таблица 3.9 – Мощность критерия Андерсона-Дарлинга относительно гипотез $H_1 - H_5$ в зависимости от объемов выборок $(n = m)$ на данных без округления

Уровень значимости α	n = 500	n = 1000	n = 2000
Относительно альтернативы H_1			
0.1	0.44	0.69	0.92
0.05	0.33	0.57	0.86
0.025	0.24	0.46	0.79
Относительно альтернативы H_2			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0
Относительно альтернативы H_3			
0.1	0.28	0.53	0.86
0.05	0.15	0.34	0.71
0.025	0.08	0.19	0.54
Относительно альтернативы H_4			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0
Относительно альтернативы H_5			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0

Таблица 3.10 – Мощность критерия Андерсона-Дарлинга относительно гипотез $H_1 - H_5$ в зависимости от объемов выборок $(n = m)$ на данных, округленных до целых чисел

Уровень значимости α	n = 500	n = 1000	n = 2000
Относительно альтернативы H_1			
0.1	0.50	0.71	0.91
0.05	0.35	0.57	0.82
0.025	0.23	0.44	0.73
Относительно альтернативы H_2			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0
Относительно альтернативы H_3			
0.1	0.03	0.01	0.00
0.05	0.01	0.00	0.00
0.025	0.00	0.00	0.00
Относительно альтернативы H_4			
0.1	0.0	0.0	0.0
0.05	0.0	0.0	0.0
0.025	0.0	0.0	0.0
Относительно альтернативы H_5			
0.1	0.0	0.0	0.0
0.05	0.0	0.0	0.0
0.025	0.0	0.0	0.0

Из таблицы 3.10 видно, что критерий Андерсона-Дарлинга оказывается смещенным (мощность меньше задаваемого уровня значимости) в случае альтернатив с пересечением функций распределения (рис. 3.3-3.5).

Таблица 3.11 – Мощность критерия Андерсона-Дарлинга относительно гипотез $H_1 - H_5$ в зависимости от объемов выборок $(n = m)$ на данных, округленных до 1 знака после запятой

Уровень значимости α	n = 500	n = 1000	n = 2000
Относительно альтернативы H_1			
0.1	0.60	0.81	0.96
0.05	0.46	0.71	0.93
0.025	0.34	0.60	0.88

Продолжение таблицы 3.11

Относительно альтернативы H_2			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0
Относительно альтернативы H_3			
0.1	0.13	0.17	0.22
0.05	0.07	0.09	0.12
0.025	0.03	0.045	0.07
Относительно альтернативы H_4			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	0.98	1.0	1.0
Относительно альтернативы H_5			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0

Таблица 3.12 – Мощность критерия Андерсона-Дарлинга относительно гипотез $H_1 - H_5$ в зависимости от объемов выборок $(n = m)$ на данных, округленных до 2 знаков после запятой

Уровень значимости α	n = 500	n = 1000	n = 2000
Относительно альтернативы H_1			
0.1	0.50	0.75	0.95
0.05	0.39	0.64	0.91
0.025	0.29	0.54	0.87
Относительно альтернативы H_2			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0
Относительно альтернативы H_3			
0.1	0.29	0.50	0.84
0.05	0.16	0.31	0.70
0.025	0.08	0.18	0.52
Относительно альтернативы H_4			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0

Продолжение таблицы 3.12

Относительно альтернативы H_5			
0.1	1.0	1.0	1.0
0.05	1.0	1.0	1.0
0.025	1.0	1.0	1.0

По данным, полученным из таблиц исследования мощности критерия Лемана-Розенблатта, также, можно заметить, что на данных, округленных до одного и двух знаков после запятой по альтернативе $H1$ мощность оказалась выше чем на данных без округления.

3.5. Выводы

Суммируя полученные результаты оценки мощностей по всем критериям на данных ограниченной точности, можно сделать следующие выводы:

- На данных, округленных до целых чисел, как наиболее мощный критерий себя показал критерий Лемана-Розенблатта по всем предложенным альтернативам, кроме гипотезы $H1$, где наибольшую мощность продемонстрировал критерий Андерсона-Дарлинга;
- На данных, округленных до одного знака после запятой, на предложенных альтернативах оказалось трудно явно определить наиболее мощный критерий. По альтернативной гипотезе $H1$ наибольшую мощность, как и на данных, округленных до целых, показал критерий Андерсона-Дарлинга. По альтернативе $H3$ наибольшую мощность проявил критерий Смирнова. По всем остальным гипотезам наиболее мощным оказался критерий Лемана-Розенблатта;
- На данных, округленных до двух знаков после запятой, наибольшую мощность по всем представленным альтернативам показал критерий Андерсона-Дарлинга;
- На данных, округленных до целых, критерий Андерсона-Дарлинга оказался смещенным.

Заключение

В соответствии с целю данной работы получены следующие основные результаты:

1) Разработана программа для проведения исследований с помощью методов имитационного моделирования распределений статистик и мощности критериев однородности в случае данных ограниченной точности.
2) В результате исследования распределений статистик показано, что:
   – для критерия Андерсона-Дарлинга расстояние между эмпирической функцией распределения статистики и предельным уменьшается с ростом отношения числа различных значений в объединенной выборке к объему объединенной выборки;
   – для критерия Лемана-Розенблатта распределения статистики остаются близкими к предельному закону при равных объемах выборок, однако при $n \neq m$ расстояние между эмпирической функцией распределения статистики и предельным увеличивается с ростом объема объединенной выборки;
   – для критерия Смирнова наблюдается медленная сходимость распределения статистики к предельному закону при увеличении объемов выборок.
3) На данных ограниченной точности наибольшую мощность среди рассмотренных критериев показали критерии Андерсона-Дарлинга и Лемана-Розенблатта. Однако в случае округления наблюдений в выборках до целых критерий Андерсона-Дарлинга оказался смещенным относительно конкурирующих гипотез с пересечением функций распределения.

Обобщая полученные результаты, можно сделать вывод о предпочтительности использования критерия Лемана-Розенблатта при равных объемах выборок $n = m$.

Список литературы

1) Смирнов Н.В. Оценка расхождения между эмпирическими кривыми распределения в двух независимых выборках / Н.В. Смирнов // Бюллетень МГУ, серия А. – 1939. – Т.2. №2. – С.3-14.

2) Massey, F. J. The Kolmogorov-Smirnov Test for Goodness of Fit. / F. J. Massey/ Journal of the American Statistical Association. Vol. 46, No. 253, 1951, pp. 68–78.

3) Miller, L. H. Table of Percentage Points of Kolmogorov Statistics. / L. H. Miller / Journal of the American Statistical Association. Vol. 51, No. 273, 1956, pp. 111–121.

4) Anderson T. W. Asymptotic theory of certain «goodness of fit» criteria based on stochastic processes / T. W. Anderson, D. A. Darling // Ann. Math. Statist. — 1952. — V. 23. — P. 193—212.

5) Anderson T. W. A test of goodness of fit / T. W. Anderson, D. A. Darling // J. Amer. Statist. Assoc., 1954. — V. 29. — P. 765—769.

6) Lehman S. Exact and approximate distributions for the Wilcoxon statistic with ties // Journal of the American Statistical Association. 1961. Vol. 56. – P. 293-988.

7) Scholz F.W., Stephens M.A. K-Sample Anderson–Darling Tests // Journal of the American Statistical Association. 1987. Vol. 82. No. 399. – P. 918-924.

8) Лемешко Б.Ю. Критерии проверки гипотез об однородности. Руководство по применению / Б.Ю. Лемешко. – М: ИНФРА–М, 2016. – 207 с.

9) Лемешко Б. Ю. О сходимости распределений статистик и мощности критериев однородности Смирнова и Лемана–Розенблатта / Б. Ю. Лемешко, С. Б. Лемешко // Измерительная техника. – 2005. – № 12. – С. 9–14.

10) Lemeshko B. Yu. Statistical distribution convergence and homogeneity test power for Smirnov and Lehmann–Rosenblatt tests / B. Yu. Lemeshko,

S. B. Lemeshko // Measurement Techniques – 2005. – Vol. 48, № 12. – P. 1159–1166.

11) Lemeshko B. Y. Application of Homogeneity Tests: Problems and Solution / B. Y. Lemeshko, I. V. Veretelnikova, S. B. Lemeshko, A. Y. Novikova // In: Rykov V., Singpurwalla N., Zubkov A. (eds) Analytical and Computational Methods in Probability Theory. ACMPT 2017. Lecture Notes in Computer Science. : monograph. - Cham : Springer, 2017. - 10684. - P. 461-475.

12) Большев Л. Н. Таблицы математической статистики / Л. Н. Большев, Н. В. Смирнов. – М. : Наука, 1983. – 416 с.

13) Lehmann E. L. Consistency and unbiasedness of certain nonparametric tests / E. L. Lehmann // Ann. Math. Statist. – 1951. – Vol. 22, № 1. – P. 165–179.

14) Newman D. The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation // Biometrika. 1939. Vol. 31. No.1/2. – P. 20-30.

15) Rosenblatt M. Limit theorems associated with variants of the von Mises statistic / M. Rosenblatt // Ann. Math. Statist. – 1952. – Vol. 23. – P. 617–623.

16) Pettitt A.N. A two-sample Anderson-Darling rank statistic // Biometrika. 1976. Vol. 63. No.1. P. 161-168.

Приложение А. Программные модули

AndersonDarling.py:

```
from scipy import stats
import numpy as np
import scipy.integrate as integrate
from math import sqrt, pi
from mpmath import nsum, inf, exp, gamma
from IHaveStatistic import IHaveSatistic


class AndersonDarlingCriteria:
    #сравнение с теоретическим: X2 - 'norm'
    def SciPyResult(self, X1, X2):
        return stats.anderson(X1, X2)

    def Result2SamplesBykSamp(self, X1, X2):
        return stats.anderson_ksamp([X1, X2])

    def Result2Samples(self ,X1, X2):
        m = len(X1)
        n = len(X2)
        N = m + n
        X1.sort()
        X2.sort()
        dataAll = np.concatenate((X1, X2))
        dataAll.sort()
        dataAll = dataAll.tolist()

        sum = 0
        sumOfX1 = 0
        for i in range(len(dataAll)-1):
            elem = dataAll[i]
            for j in range(sumOfX1, m):
                if (X1[j] > elem):
                    break
                else:
                    sumOfX1 += 1

            sum += pow((sumOfX1*N - m*(i+1)), 2) / ((i+1)*(N-i-1))

        return IHaveSatistic(sum/(m*n))

    @staticmethod
    def GetStatisticDistribution(statistics):
        def a2(jj):
            j = float(jj)
            temp = pow((4.0*j + 1), 2)
            temp2 = 8*stat
            res = gamma(j + 0.5)*(4.0*j + 1)/(gamma(0.5)*gamma(j + 1.0))
            res *= exp(-temp*pi*pi/temp2)
```

```
            res *= integrate.quad(lambda y: exp((stat / (8*(y*y+1))) - temp*pi*pi*y*y/temp2), 0,
np.inf)[0]
            return pow(-1, j) * res

        result = []
        for index, stat in enumerate(statistics):
            # if (index % 100 == 0):
            #     print(index)
            sum = float(nsum(lambda j: a2(j), [0, inf]))

            st = sqrt(2*pi) / stat
            result.append(st*sum)
        return result
```

LehmanRosenblatt.py:

```
import numpy as np
from Helper import Helper
from IHaveStatistic import IHaveSatistic
from scipy.special import iv
from mpmath import nsum, inf, exp, gamma
from math import sqrt
import math


class LehmanRosenblattCriteria:
    def Result2Samples(self, X1, X2):
        m = len(X1)
        n = len(X2)
        sum0 = m + n
        X1.sort()
        X2.sort()
        dataAll = np.concatenate((X1, X2))
        dataAll.sort()
        dataAll = dataAll.tolist()

        sum1 = 0
        for i, elem in enumerate(X1):
            sum1 += pow(Helper.GetRang(dataAll, elem) - i, 2)

        sum2 = 0
        for i, elem in enumerate(X2):
            sum2 += pow(Helper.GetRang(dataAll, elem) - i, 2)

        stat = (n * sum2 + m * sum1)/(m*n*sum0) - (4*m*n - 1) / (6 * sum0)
        return IHaveSatistic(stat if stat != 0 else 1E-15)#иногда статистика получается равна 0, не
должно быть такого

    @staticmethod
    def GetStatisticDistribution(statistics):
        def a1(jj):
```

```
            j = float(jj)
            el = (4 * j + 1)*(4 * j + 1) / 16.0 / stat
            temp = gamma(j + 0.5)*sqrt(4.0*j + 1)/(gamma(0.5)*gamma(j + 1.0))
            bessel = (iv(-0.25, el) - iv(0.25, el))
            return temp * exp(-el) * bessel if not math.isnan(bessel) else 0.0

        result = []
        for stat in statistics:
            sum = float(nsum(lambda j: a1(j), [0, inf]))

            st = (1 / sqrt(2 * stat))
            result.append(st*sum)
        return result
```

Smirnov.py:

```
from scipy import stats

class SmirnovCriteria:
    def SciPyResult(self, X1, X2):
        return stats.kstest(X1, X2)

    def Result2Samples(self, X1, X2):
        return stats.ks_2samp(X1, X2)
```

Helper.py:

```
from decimal import *


class Helper:
    @staticmethod
    def GetLastIndexOf(list, value):
        return len(list) - list[::-1].index(value) - 1

    '"Получить ранг элемента вариационного ряда"'
    @staticmethod
    def GetRang(list, value):
        first = list.index(value)
        last = first
        for elem in list[first+1:]:
            if elem == value:
                last += 1
                continue
            break
        return (first + last) / 2
```

```
    @staticmethod
    def RoundingArray(array, digitCount):
        # getcontext().rounding = ROUND_HALF_UP
        result = []
        for elem in array:
            result.append(round(Decimal(elem), digitCount))
        return result
```

PowerCalculateHelper.py:

```
import scipy.stats as stats
from pandas import Series as ser
import numpy as np
from statsmodels.distributions.empirical_distribution import ECDF
from Helper import Helper
import matplotlib.pyplot as plt
import LehmanRosenblatt as lr

class PowerCalculateHelper:
    @staticmethod
    def CalculateStats(n, m, N, criteria, digit):
        SH0 = []
        SH1 = []
        SH2 = []
        SH3 = []
        SH4 = []
        SH5 = []

        for i in range(N):
            rvs = stats.norm.rvs(loc=0, scale=1, size=n)
            x1_H0 = rvs if digit == '-' else Helper.RoundingArray(rvs, digit)
            rvs = stats.norm.rvs(loc=0, scale=1, size=m)
            x2_H0 = rvs if digit == '-' else Helper.RoundingArray(rvs, digit)
            SH0.append(criteria.Result2Samples(x1_H0, x2_H0).statistic)

            rvs = stats.norm.rvs(loc=0, scale=1, size=n)
            x1_H1 = rvs if digit == '-' else Helper.RoundingArray(rvs, digit)
            rvs = stats.norm.rvs(loc=0.1, scale=1, size=m)
            x2_H1 = rvs if digit == '-' else Helper.RoundingArray(rvs, digit)
            SH1.append(criteria.Result2Samples(x1_H1, x2_H1).statistic)

            rvs = stats.norm.rvs(loc=0, scale=1, size=n)
            x1_H2 = rvs if digit == '-' else Helper.RoundingArray(rvs, digit)
            rvs = stats.norm.rvs(loc=0.5, scale=1, size=m)
            x2_H2 = rvs if digit == '-' else Helper.RoundingArray(rvs, digit)
            SH2.append(criteria.Result2Samples(x1_H2, x2_H2).statistic)

            rvs = stats.norm.rvs(loc=0, scale=1, size=n)
```

```
            x1_H3 = rvs if digit == '-' else Helper.RoundingArray(rvs, digit)
            rvs = stats.norm.rvs(loc=0, scale=1.1, size=m)
            x2_H3 = rvs if digit == '-' else Helper.RoundingArray(rvs, digit)
            SH3.append(criteria.Result2Samples(x1_H3, x2_H3).statistic)

            rvs = stats.norm.rvs(loc=0, scale=1, size=n)
            x1_H4 = rvs if digit == '-' else Helper.RoundingArray(rvs, digit)
            rvs = stats.norm.rvs(loc=0, scale=1.5, size=m)
            x2_H4 = rvs if digit == '-' else Helper.RoundingArray(rvs, digit)
            SH4.append(criteria.Result2Samples(x1_H4, x2_H4).statistic)

            rvs = stats.norm.rvs(loc=0, scale=1, size=n)
            x1_H5 = rvs if digit == '-' else Helper.RoundingArray(rvs, digit)
            rvs = stats.logistic.rvs(loc=0, scale=1, size=m)
            x2_H5 = rvs if digit == '-' else Helper.RoundingArray(rvs, digit)
            SH5.append(criteria.Result2Samples(x1_H5, x2_H5).statistic)

        # Вычисление мощностей
        print(PowerCalculateHelper.CalculatePower(SH0, SH1, [0.1, 0.05, 0.025]))
        print(PowerCalculateHelper.CalculatePower(SH0, SH2, [0.1, 0.05, 0.025]))
        print(PowerCalculateHelper.CalculatePower(SH0, SH3, [0.1, 0.05, 0.025]))
        print(PowerCalculateHelper.CalculatePower(SH0, SH4, [0.1, 0.05, 0.025]))
        print(PowerCalculateHelper.CalculatePower(SH0, SH5, [0.1, 0.05, 0.025]))


@staticmethod
    def CalculatePower(statsH0, statsH1,  alphas, criteriaSide = None):
        quantiles = ser(statsH0).quantile(np.ones(len(alphas)) - alphas).values

        ecdf = ECDF(statsH1)

        possibilites = ecdf(quantiles)

        print()

        return np.ones(len(alphas)) - possibilites
```