Исследование распределений статистик и мощности критериев однородности в случае больших массивов данных

**Оглавление**

[**Введение** 3](#_Toc515205657)

[**1.** **Критерии проверки однородности законов распределения** 8](#_Toc515205658)

[**1.1.** **Общая постановка** 8](#_Toc515205659)

[**1.2.** **Критерий Смирнова** 8](#_Toc515205660)

[**1.3.** **Критерий Лемана-Розенблатта** 9](#_Toc515205661)

[**1.4.** **Критерий Андерсона-Дарлинга** 10](#_Toc515205662)

[**2.** **Исследование распределений статистик критериев однородности на данных ограниченной точности** 12](#_Toc515205663)

[**2.1.** **Исследование распределений статистик** 12](#_Toc515205664)

[**2.2.** **Выводы** 17](#_Toc515205665)

[**3.** **Исследование мощностей критериев однородности на данных ограниченной точности** 18](#_Toc515205666)

[**3.1.** **Исследование мощностей критериев** 18](#_Toc515205667)

[**3.2.** **Выводы** 28](#_Toc515205668)

[**Заключение** 30](#_Toc515205669)

[**Список литературы** 31](#_Toc515205670)

[**Приложение А. Программные модули** 33](#_Toc515205671)

# **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 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 чаще всего 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 тоже становится большим. Становится интересно, можно ли руководствоваться данными по исследованию 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 мощностей критериев однородности подробно рассматривались в работах [8 - 11].

**Цель и задачи исследований.** Целью данной диссертационной работы является разработка математического и алгоритмического обеспечения для исследования критериев однородности на данных ограниченной точности.

Для достижения сформулированной цели были поставлены и решены следующие задачи:

* исследование распределения статистик критериев однородности: Андерсона-Дарлинга, Лемана-Розенблатта, Смирнова на данных ограниченной точности;
* сравнительный анализ распределения статистик критериев с предельными функциями распределения;
* сравнительный анализ мощности вышеизложенных критериев на данных ограниченной точности и сравнение с мощностями, полученными по данным без ограничения точности;
* разработка программного обеспечения для анализа схожести распределений статистик критериев с предельными распределениями и для вычисления мощностей критериев.

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

**Научная новизна** диссертационной работы заключается в следующем:

* 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.

**Основные положения, выносимые на защиту.** На защиту выносятся следующие результаты:

* /////////

**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 описании методики исследования данных ограниченной точности;
* в исследовании распределения статистик критериев (проверка близости к предельной функции распределения);
* в вычислении мощности данных критериев на данных ограниченной точности и сравнение с мощностями на данных без округления;
* в реализации полученных результатов в 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. Разработанное программное обеспечение может применяться во многих сферах для решения прикладных задач, связанных с выявлением однородности данных.

**Структура работы.** Диссертация состоит из введения, 5 глав основного содержания, заключения, списка литературы и 3 приложений. Основная часть содержания изложена на … страницах, включая … рисунков, … таблиц и списка литературы из … 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 разработанной программы на языке программирования Python.

# **Критерии проверки однородности законов распределения**

## **Общая постановка**

При анализе случайных ошибок средств измерений, при статическом 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 упорядоченные по не убыванию выборки размером  и  :

 и .

Для определенности обычно полагают, что , но это совсем необязательно. Проверяется гипо­теза о том, что обе выборки извлечены из одной и той же генеральной сово­купности, т. е. :  при любом .

## **Критерий Смирнова**

Критерий Смирнова - это двухсторонний тест с нулевой гипотезой о том, что из одного и того же непрерывного распределения извлекаются 2 независимых выборки. Критерий однородности Смирнова предложен в работе [1]. Предполагается, что функции распределения  и  являются непрерывными. Статистика критерия Смирнова измеряет расстояние ме­жду эмпирическими функциями распределения, построенными по выборкам [1]



На практике, значение статистики  рекомен­дуется вычислять в соответствии с соотношениями [8]:

,

,

.

Если гипотеза  справедлива, то при неограниченном увеличении объемов выборок [12] , т. е. статис­тика



в пределе подчиняется распределению Колмогорова  [12] с функцией распределения

.

Однако при ограниченных значениях  и  случайные величины  и  являются дискретными, и множество их возможных значений представляет собой решетку с шагом , где  – наименьшее общее кратное  и  [12]. Условное распределение  статистики  при верности гипотезы  медленно сходится к  и имеет существенное отличие от него при малых значениях  и .

Гладкость распределения статистики сильно зависит от величины . Поэтому предпочтительнее применять критерий, когда объемы выборок  и  не равны и представляют собой вза­имно простые числа. В таких случаях наименьшее общее кратное  и  максимально и равно , а распределе­ние статистики больше напоминает непрерывную функцию распределения.

## **Критерий Лемана-Розенблатта**

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

,

где  – эмпирическая функция распределения, построенная по вариационному ряду объединения двух выборок. Статистика  используется в форме [12]

|  |  |
| --- | --- |
| , | (1.1) |

где  – порядковый номер (ранг) ;  – порядковый номер (ранг)  в объе­диненном вариационном ряде.

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

.

Функция распределения  имеет вид [12]:



,

где  – модифицированные функции Бесселя вида

.

В отличие от критерия Смирнова распределение статистики  быстро сходится к предельному  [12]

## **Критерий Андерсона-Дарлинга**

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

.

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

|  |  |
| --- | --- |
| , | (1.2) |

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

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



.

# **Исследование распределений статистик критериев однородности на данных ограниченной точности**

## **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.

Можно ли использовать критерии, если данные ограничены, подчиняются ли статистики, вычисленные по таким данным предельным законам распределения заданных критериев однородности.

В таблицах ниже (когда будут номера у таблиц, надо их тут указать!) представлены значения расстояний между эмпирическими и предельными функциями распределения статистик, рассчитанные по метрике Колмогорова. Зададимся величиной, при котором будем считать, что распределение статистик подчиняется предельному закону распределения, равной 0.05.

**АД**

**\* -** Среднее число различных значений в выборках

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | N=16600, n: 10, m: 10 | |  | |  | |
|  | Расстояние по Колмогорову | \* |  |  |  |  |
| 0 | 0.7307337944425589 | 5.0 |  |  |  |  |
| 1 | 0.09732733270929703 | 28 |  |  |  |  |
| 2 | 0.05093170506557959 | 47 |  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 200, 200 | | n: 500, m: 500 | | n: 1000, m: 1000 | |
|  |  |  |  |  |  |  |
| 0 | 0.9999999999999677 | 7.0 | 1.0000000010906982 | 7.0 | 49783007492.5392 | 8.0 |
| 1 | 0.3177606900344735 | 50.0 | 0.5580592698629411 | 59.0 | 0.7813174951763587 | 64.5 |
| 2 | 0.016657660678184527 | 241.0 | 0.02413591555767497 | 377.0 | 0.026778729840812754 | 442.0 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | n: 2000, m: 2000 | | n: 5000, m: 5000 | |  | |
|  |  |  |  |  |  |  |
| 0 | 7.325497982177374e+24 | 8.0 | 1.6864794476131036e+86 | 9 |  |  |
| 1 | 0.9406704541308208 | 66.5 | 0.9984727668727141 | 72.5 |  |  |
| 2 | 0.041710861946732514 | 510.0 | 0.08155058951357014 | 576.5 |  |  |

Где-то между n=m=2000 и n=m=5000 в округлении до 2ух знаков расстояние становится большим чем 0.05.

В округлении до целых, при n=m <= 500 расстояние было около единицы, на 1000 и больше функция распределения статистик трудно назвать схожей с предельным распределением. Хотя и при меньших размерах выборок, расстояние от предельного, равное единицы, это большая разница между функциями.

Как и предполагалось, расстояние растет с увеличением размерностей выборок. Расстояние, не превышающее величину 0.05 достигается только при очень маленьких значениях размерностей выборок при округлении до одного, двух знаков после запятой.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 30, 30 | | 30, 40 | | 30, 50 | |
|  |  |  |  |  |  |  |
| 0 | 0.971939404837819 | 5.0 | 0.9613735405573581 | 6.0 | 0.9513233366841691 | 6.0 |
| 1 | 0.12327215961403043 | 31.5 | 0.09601900915710787 | 33.5 | 0.09355808370624932 | 34.5 |
| 2 | 0.026233336804103197 | 55.5 | 0.01749731042612679 | 63.0 | 0.017520768482571808 | 71.0 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | n: 500, m: 1000 | | n: 500, m: 2000 | | n: 500, m: 5000 | |
|  |  |  |  |  |  |  |
| 0 | 0.9999999999999912 | 7.5 | 0.9999999999931858 | 8.0 | 0.9999999999999996 | 8.0 |
| 1 | 0.45527807797431663 | 60.5 | 0.3901980362227492 | 63.5 | 0.3489378555792969 | 68.5 |
| 2 | 0.009735145378913068 | 422.0 | 0.012344253974853281 | 465.0 | 0.009978816566083515 | 532.5 |

Что интересно, при различных размерностях выборок, с увеличением размерности второй выборки при зафиксированном значении размерности первой, расстояния оказываются меньшими, чем когда размерности двух выборок одинаковые. Возможно надо еще потестировать при n=1000 и m=2000, 5000.

Для графиков важно подписать оси (ГРАФИКИ вставить нормальные)

**Лем-Роз**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | N=16600, n=20, m=30 | | 30, 30 | | 30, 40 | |
|  | Расстояние по Колмогорову | \* |  |  |  |  |
| 0 | 0.2233477694432367 | 6.0 |  |  |  |  |
| 1 | 0.02697723285461584 | 27.5 |  |  |  |  |
| 2 | 0.021285450329402766 | 47.5 |  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 200, 200 | | n: 500, m: 500 | | n: 1000, m: 1000 | |
|  |  |  |  |  |  |  |
| 0 | 0.11536710664079695 | 7.0 | 0.11339391175792744 | 7.0 | 0.11519912345224 | 8.0 |
| 1 | 0.006769699360278936 | 49.0 | 0.006149829639210902 | 56.5 | 0.005304291721869081 | 62.5 |
| 2 | 0.006126600306398777 | 243.5 | 0.0034988541215064117 | 369.5 | 0.005246511650933039 | 448.5 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | n: 2000, m: 2000 | | n: 5000, m: 5000 | |  | |
|  |  |  |  |  |  |  |
| 0 | 0.11978353880223283 | 9.0 | 0.12067961230049604 | 9.0 |  |  |
| 1 | 0.010790795109519302 | 67.5 | 0.005362467256544123 | 72.5 |  |  |
| 2 | 0.008234065731409945 | 510.5 | 0.005361824969545181 | 578.5 |  |  |

Судя по результатам, 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. Интересно и то, что при округлении до одного знака результаты оказываются не многим хуже, чем при округлении до двух знаков.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 30, 30 | | 30, 40 | | 30, 50 | |
|  |  |  |  |  |  |  |
| 0 | 0.11695639111559329 | 6.0 | 0.13034441570224864 | 5.5 | 0.5475105174962243 | 6.0 |
| 1 | 0.0173855133189037 | 30.5 | 0.015110378505309996 | 33.0 | 0.0312338093564187 | 34.5 |
| 2 | 0.01861528625271719 | 55.5 | 0.014258353406632063 | 63.0 | 0.01397052387315445 | 71.5 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | n: 500, m: 1000 | | n: 500, m: 2000 | | n: 500, m: 5000 | |
|  |  |  |  |  |  |  |
| 0 | 0.999999999813874 | 8.0 | 1.0000000000000002 | 8.0 | 1.0 | 9.0 |
| 1 | 0.34727665584086514 | 60.0 | 0.9425134239782521 | 63.0 | 0.9999999753969714 | 70.0 |
| 2 | 0.009420167926454087 | 418.0 | 0.034882791447205796 | 469.0 | 0.24467005638422118 | 535.0 |

Судя по результатам данной таблицы для критерия Лемана-Розенблатта не наблюдается приближения распределения статистик к предельному закону при различных размерностях выборок в сравнении с результатами, полученными при одинаковых размерностях выборок. Некоторое уменьшение расстояния наблюдается при малых размерностях выборок.

**Смирнов**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | N=16600, 200, 200 | | n: 500, m: 500 | | n: 1000, m: 1000 | |
|  | Расстояние по Колмогорову | \* |  |  |  |  |
| 0 | 0.540092677385176 | 7.0 | 0.5297165835765115 | 7.0 | 0.5244438293905644 | 8.0 |
| 1 | 0.2033904740447694 | 50.5 | 0.19198287602519692 | 57.0 | 0.18685736339206777 | 62.5 |
| 2 | 0.09244206408266703 | 246.5 | 0.07452621478141924 | 368.5 | 0.06584598231754063 | 449.0 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | n: 2000, m: 2000 | | n: 5000, m: 5000 | |  | |
|  |  |  |  |  |  |  |
| 0 | 0.5320172493751892 | 9.0 | 0.521783875767813 | 9.0 |  |  |
| 1 | 0.18373768330822304 | 67.0 | 0.18056920086908662 | 72.0 |  |  |
| 2 | 0.06747756952701311 | 507.0 | 0.06094983259899789 | 580.5 |  |  |

Что интересно, наблюдается уменьшение расстояния с ростом размерностей выборок при одинаковых размерах обеих выборок, в отличие от других критериев. Но тем не менее, заданное расстояние между функциями распределения не достигается.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | n=30, m=30 | | 30, 40 | | 30, 50 | |
|  |  |  |  |  |  |  |
| 0 | 0.5646230429354729 | 5.5 | 0.511252231412543 | 6.0 | 0.5089345868104809 | 6.0 |
| 1 | 0.2557694018270209 | 31.0 | 0.1635153609583173 | 33.0 | 0.16659448091948403 | 33.5 |
| 2 | 0.17625132953786432 | 56.0 | 0.08060662373472738 | 64.0 | 0.08069086646165269 | 73.0 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 70, 100 | | 70, 110 | | 70, 130 | |
|  |  |  |  |  |  |  |
| 0 | 0.5072942563445924 | 7.0 | 0.5090442970635045 | 7.0 | 0.5041810957534603 | 6.5 |
| 1 | 0.15735952708057133 | 44.0 | 0.15996048792424555 | 43.5 | 0.15479612518033248 | 46.0 |
| 2 | 0.0575481039293369 | 136.5 | 0.052490608406173256 | 141.0 | 0.047139872823409645 | 156.5 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | n: 500, m: 1000 | | n: 500, m: 2000 | | n: 500, m: 5000 | |
|  |  |  |  |  |  |  |
| 0 | 0.5248443717187758 | 8.0 | 0.5217873150933013 | 8.0 | 0.5186468004121724 | 8.0 |
| 1 | 0.1856612053783412 | 61.5 | 0.17521460213913198 | 64.5 | 0.17020949944498526 | 68.5 |
| 2 | 0.06638792869242294 | 421.5 | 0.05844190731832288 | 468.0 | 0.0516987519130242 | 533.5 |

Наблюдается схожая картина с аналогичными исследованиями критерия Андерсона-Дарлинга. При различных размерностях выборок, с увеличением размерности второй выборки при зафиксированном значении размерности первой, расстояния оказываются меньшими, чем когда размерности двух выборок одинаковые.

## **Выводы**

# **Исследование мощностей критериев однородности на данных ограниченной точности**

## **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  соответствовала принад­лежность выборок одному и тому же стандартному нормальному закону распределения с плотностью



и параметрами сдвига  и масштаба .

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

В частности, при альтернативе сдвига в случае конкурирующей гипотезы  вторая вы­борка соответствовала нормальному закону с параметром сдвига  и па­раметром масштаба , в случае конкурирующей гипотезы  – нормальному закону с параметрами  и .

При изменении масштаба в случае конкурирующей гипотезы  вторая вы­борка соответствовала нормальному закону с параметрами  и , в случае конкурирующей гипотезы  – нормальному за­кону с параметрами  и .

В случае конкурирующей гипотезы  вторая вы­борка соответствовала логистическому закону с плотностью



и параметрами  и .

В таблицах 3.1 – 3.4 представлены рассчитанные оценки мощностей критерия однородности Смирнова (, - вероятность ошибки второго рода). Значения представлены относительно конкурирующих гипотез  для различных значений размерности выборок. Значения оценок мощности также представлены в зависимости от различных значений заданных уровней значимости (вероятностей ошибок первого рода): .

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Уровень значимости | *n* = 500 | | *n* = 1000 | | *n* = 2000 | |
| Относительно альтернативы | | | | | | |
| 0.1 | 0.37759036 | | 0.61150602 | | 0.86777108 | |
| 0.05 | 0.27240964 | | 0.48054217 | | 0.78259036 | |
| 0.025 | 0.17825301 | | 0.36036145 | | 0.68692771 | |
| Относительно альтернативы | | | | | | |
| 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 | | | | | | |
| 0.1 | 0.18698795 | | 0.30054217 | | 0.5503012 | |
| 0.05 | 0.10283133 | | 0.17337349 | | 0.35524096 | |
| 0.025 | 0.05295181 | | 0.09108434 | | 0.21150602 | |
| Относительно альтернативы | | | | | | |
| 0.1 | 0.99915663 | | 1.0 | | 1.0 | |
| 0.05 | 0.995 | | 1.0 | | 1.0 | |
| 0.025 | 0.9803012 | | 1.0 | | 1.0 | |
| Относительно альтернативы | | | | | | |
| 0.1 | | 1.0 | | 1.0 | | 1.0 |
| 0.05 | | 1.0 | | 1.0 | | 1.0 |
| 0.025 | | 0.99993976 | | 1.0 | | 1.0 |

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Уровень значимости | *n* = 500 | | *n* = 1000 | | *n* = 2000 | |
| Относительно альтернативы | | | | | | |
| 0.1 | 0.36650602 | | 0.58277108 | | 0.84006024 | |
| 0.05 | 0.25457831 | | 0.45409639 | | 0.74162651 | |
| 0.025 | 0.1686747 | | 0.34259036 | | 0.63343373 | |
| Относительно альтернативы | | | | | | |
| 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 | | | | | | |
| 0.1 | 0.19174699 | | 0.29825301 | | 0.53759036 | |
| 0.05 | 0.10668675 | | 0.16843373 | | 0.32584337 | |
| 0.025 | 0.05668675 | | 0.09686747 | | 0.19668675 | |
| Относительно альтернативы | | | | | | |
| 0.1 | 0.99987952 | | 1.0 | | 1.0 | |
| 0.05 | 0.99801205 | | 1.0 | | 1.0 | |
| 0.025 | 0.98849398 | | 1.0 | | 1.0 | |
| Относительно альтернативы | | | | | | |
| 0.1 | | 1.0 | | 1.0 | | 1.0 |
| 0.05 | | 1.0 | | 1.0 | | 1.0 |
| 0.025 | | 0.99987952 | | 1.0 | | 1.0 |

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Уровень значимости | *n* = 500 | | *n* = 1000 | | *n* = 2000 | |
| Относительно альтернативы | | | | | | |
| 0.1 | 0.38722892 | | 0.61409639 | | 0.86939759 | |
| 0.05 | 0.2796988 | | 0.4886747 | | 0.78349398 | |
| 0.025 | 0.18921687 | | 0.37343373 | | 0.68716867 | |
| Относительно альтернативы | | | | | | |
| 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 | | | | | | |
| 0.1 | 0.20349398 | | 0.31307229 | | 0.5676506 | |
| 0.05 | 0.11168675 | | 0.1813253 | | 0.36668675 | |
| 0.025 | 0.05445783 | | 0.09993976 | | 0.2226506 | |
| Относительно альтернативы | | | | | | |
| 0.1 | 0.99975904 | | 1.0 | | 1.0 | |
| 0.05 | 0.99783133 | | 1.0 | | 1.0 | |
| 0.025 | 0.98692771 | | 1.0 | | 1.0 | |
| Относительно альтернативы | | | | | | |
| 0.1 | | 1.0 | | 1.0 | | 1.0 |
| 0.05 | | 1.0 | | 1.0 | | 1.0 |
| 0.025 | | 0.99993976 | | 1.0 | | 1.0 |

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Уровень значимости | *n* = 500 | | *n* = 1000 | | *n* = 2000 | |
| Относительно альтернативы | | | | | | |
| 0.1 | 0.38349398 | | 0.60240964 | | 0.86566265 | |
| 0.05 | 0.26933735 | | 0.47475904 | | 0.77963855 | |
| 0.025 | 0.17783133 | | 0.37198795 | | 0.68168675 | |
| Относительно альтернативы | | | | | | |
| 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 | | | | | | |
| 0.1 | 0.19271084 | | 0.29361446 | | 0.55801205 | |
| 0.05 | 0.10590361 | | 0.16144578 | | 0.35295181 | |
| 0.025 | 0.05084337 | | 0.09313253 | | 0.21662651 | |
| Относительно альтернативы | | | | | | |
| 0.1 | 0.99933735 | | 1.0 | | 1.0 | |
| 0.05 | 0.99578313 | | 1.0 | | 1.0 | |
| 0.025 | 0.98331325 | | 1.0 | | 1.0 | |
| Относительно альтернативы | | | | | | |
| 0.1 | | 1.0 | | 1.0 | | 1.0 |
| 0.05 | | 1.0 | | 1.0 | | 1.0 |
| 0.025 | | 0.99993976 | | 1.0 | | 1.0 |

Выводы по смирнову сравнивая округления

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.

В таблицах 3.5 – 3.8 представлены рассчитанные оценки мощностей критерия однородности Лемана-Розенблатта. Значения представлены относительно конкурирующих гипотез  для различных значений размерности выборок. Значения оценок мощности также представлены в зависимости от различных значений заданных уровней значимости: .

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Уровень значимости | *n* = 500 | | *n* = 1000 | | *n* = 2000 | |
| Относительно альтернативы | | | | | | |
| 0.1 | 0.44198795 | | 0.67542169 | |  | |
| 0.05 | 0.32457831 | | 0.55638554 | |  | |
| 0.025 | 0.23144578 | | 0.44066265 | |  | |
| Относительно альтернативы | | | | | | |
| 0.1 | 1.0 | | 1.0 | |  | |
| 0.05 | 1.0 | | 1.0 | |  | |
| 0.025 | 1.0 | | 1.0 | |  | |
| Относительно альтернативы | | | | | | |
| 0.1 | 0.19174699 | | 0.32325301 | |  | |
| 0.05 | 0.095 | | 0.16078313 | |  | |
| 0.025 | 0.0460241 | | 0.07295181 | |  | |
| Относительно альтернативы | | | | | | |
| 0.1 | 0.99993976 | | 1.0 | |  | |
| 0.05 | 0.9996988 | | 1.0 | |  | |
| 0.025 | 0.99662651 | | 1.0 | |  | |
| Относительно альтернативы | | | | | | |
| 0.1 | | 1.0 | | 1.0 | |  |
| 0.05 | | 1.0 | | 1.0 | |  |
| 0.025 | | 1.0 | | 1.0 | |  |

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Уровень значимости | *n* = 500 | | *n* = 1000 | | *n* = 2000 | |
| Относительно альтернативы | | | | | | |
| 0.1 | 0.41403614 | | 0.63331325 | |  | |
| 0.05 | 0.29801205 | | 0.51680723 | |  | |
| 0.025 | 0.21626506 | | 0.41283133 | |  | |
| Относительно альтернативы | | | | | | |
| 0.1 | 1.0 | | 1.0 | |  | |
| 0.05 | 1.0 | | 1.0 | |  | |
| 0.025 | 1.0 | | 1.0 | |  | |
| Относительно альтернативы | | | | | | |
| 0.1 | 0.20506024 | | 0.32228916 | |  | |
| 0.05 | 0.10710843 | | 0.18042169 | |  | |
| 0.025 | 0.05403614 | | 0.09746988 | |  | |
| Относительно альтернативы | | | | | | |
| 0.1 | 0.99993976 | | 1.0 | |  | |
| 0.05 | 0.99921687 | | 1.0 | |  | |
| 0.025 | 0.99590361 | | 1.0 | |  | |
| Относительно альтернативы | | | | | | |
| 0.1 | | 1.0 | | 1.0 | |  |
| 0.05 | | 1.0 | | 1.0 | |  |
| 0.025 | | 1.0 | | 1.0 | |  |

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Уровень значимости | *n* = 500 | | *n* = 1000 | | *n* = 2000 | |
| Относительно альтернативы | | | | | | |
| 0.1 | 0.44271084 | | 0.67626506 | |  | |
| 0.05 | 0.31939759 | | 0.56018072 | |  | |
| 0.025 | 0.2246988 | | 0.4463253 | |  | |
| Относительно альтернативы | | | | | | |
| 0.1 | 1.0 | | 1.0 | |  | |
| 0.05 | 1.0 | | 1.0 | |  | |
| 0.025 | 1.0 | | 1.0 | |  | |
| Относительно альтернативы | | | | | | |
| 0.1 | 0.20156627 | | 0.32439759 | |  | |
| 0.05 | 0.09614458 | | 0.17204819 | |  | |
| 0.025 | 0.04560241 | | 0.08379518 | |  | |
| Относительно альтернативы | | | | | | |
| 0.1 | 1.0 | | 1.0 | |  | |
| 0.05 | 0.99975904 | | 1.0 | |  | |
| 0.025 | 0.99578313 | | 1.0 | |  | |
| Относительно альтернативы | | | | | | |
| 0.1 | | 1.0 | | 1.0 | |  |
| 0.05 | | 1.0 | | 1.0 | |  |
| 0.025 | | 0.99993976 | | 1.0 | |  |

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Уровень значимости | *n* = 500 | | *n* = 1000 | | *n* = 2000 | |
| Относительно альтернативы | | | | | | |
| 0.1 | 0.43662651 | | 0.68433735 | |  | |
| 0.05 | 0.32295181 | | 0.57246988 | |  | |
| 0.025 | 0.23054217 | | 0.45692771 | |  | |
| Относительно альтернативы | | | | | | |
| 0.1 | 1.0 | | 1.0 | |  | |
| 0.05 | 1.0 | | 1.0 | |  | |
| 0.025 | 1.0 | | 1.0 | |  | |
| Относительно альтернативы | | | | | | |
| 0.1 | 0.18722892 | | 0.32692771 | |  | |
| 0.05 | 0.09283133 | | 0.17385542 | |  | |
| 0.025 | 0.04620482 | | 0.08391566 | |  | |
| Относительно альтернативы | | | | | | |
| 0.1 | 0.99993976 | | 1.0 | |  | |
| 0.05 | 0.99945783 | | 1.0 | |  | |
| 0.025 | 0.99644578 | | 1.0 | |  | |
| Относительно альтернативы | | | | | | |
| 0.1 | | 1.0 | | 1.0 | |  |
| 0.05 | | 1.0 | | 1.0 | |  |
| 0.025 | | 1.0 | | 1.0 | |  |

Выводы по Лем-Роз сравнивая округления

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.

В таблицах 3.9 – 3.12 представлены рассчитанные оценки мощностей критерия однородности Андерсона-Дарлинга. Значения представлены относительно конкурирующих гипотез  для различных значений размерности выборок. Значения оценок мощности также представлены в зависимости от различных значений заданных уровней значимости: .

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Уровень значимости | *n* = 500 | | *n* = 1000 | | *n* = 2000 | |
| Относительно альтернативы | | | | | | |
| 0.1 | 0.44415663 | | 0.68849398 | | 0.92072289 | |
| 0.05 | 0.32921687 | | 0.5736747 | | 0.86259036 | |
| 0.025 | 0.23885542 | | 0.45939759 | | 0.78813253 | |
| Относительно альтернативы | | | | | | |
| 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 | | | | | | |
| 0.1 | 0.28060241 | | 0.52831325 | | 0.86084337 | |
| 0.05 | 0.15331325 | | 0.33686747 | | 0.71427711 | |
| 0.025 | 0.07783133 | | 0.19481928 | | 0.53518072 | |
| Относительно альтернативы | | | | | | |
| 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 | | | | | | |
| 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 в зависимости от объемов выборок (n = m) на данных, округленных до целых чисел

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Уровень значимости | *n* = 500 | | *n* = 1000 | | *n* = 2000 | |
| Относительно альтернативы | | | | | | |
| 0.1 | 0.50120482 | | 0.70789157 | | 0.90626506 | |
| 0.05 | 0.35042169 | | 0.56963855 | | 0.82349398 | |
| 0.025 | 0.23222892 | | 0.44138554 | | 0.72560241 | |
| Относительно альтернативы | | | | | | |
| 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 | | | | | | |
| 0.1 | 0.02536145 | | 0.01253012 | | 0.00445783 | |
| 0.05 | 0.00915663 | | 0.00457831 | | 0.00144578 | |
| 0.025 | 0.00373494 | | 0.00156627 | | 0.00042169 | |
| Относительно альтернативы | | | | | | |
| 0.1 | 0.00018072 | | 0.0 | | 0.0 | |
| 0.05 | 0.0 | | 0.0 | | 0.0 | |
| 0.025 | 0.0 | | 0.0 | | 0.0 | |
| Относительно альтернативы | | | | | | |
| 0.1 | | 6.02409639e-05 | | 0.0 | | 0.0 |
| 0.05 | | 6.02409639e-05 | | 0.0 | | 0.0 |
| 0.025 | | 0.0 | | 0.0 | | 0.0 |

**Мощность сильно упала на 3,4 гипотезах (где менялся параметр масштаба)**

**Также сильно упала на 5 альтернативе**

**Странно: в H3 мощность уменьшается с увеличением размерности выборок**

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Уровень значимости | *n* = 500 | | *n* = 1000 | | *n* = 2000 | |
| Относительно альтернативы | | | | | | |
| 0.1 | 0.59698795 | | 0.81349398 | | 0.96349398 | |
| 0.05 | 0.45831325 | | 0.70861446 | | 0.92909639 | |
| 0.025 | 0.34277108 | | 0.59650602 | | 0.87771084 | |
| Относительно альтернативы | | | | | | |
| 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 | | | | | | |
| 0.1 | 0.13222892 | | 0.1663253 | | 0.21831325 | |
| 0.05 | 0.06638554 | | 0.08614458 | | 0.12156627 | |
| 0.025 | 0.03289157 | | 0.04463855 | | 0.06614458 | |
| Относительно альтернативы | | | | | | |
| 0.1 | 0.99927711 | | 1.0 | | 1.0 | |
| 0.05 | 0.99584337 | | 1.0 | | 1.0 | |
| 0.025 | 0.98216867 | | 1.0 | | 1.0 | |
| Относительно альтернативы | | | | | | |
| 0.1 | | 1.0 | | 1.0 | | 1.0 |
| 0.05 | | 1.0 | | 1.0 | | 1.0 |
| 0.025 | | 1.0 | | 1.0 | | 1.0 |

**Все еще слабо распознает распределения в гипотезе Н3**

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Уровень значимости | *n* = 500 | | *n* = 1000 | | *n* = 2000 | |
| Относительно альтернативы | | | | | | |
| 0.1 | 0.50349398 | | 0.7473494 | | 0.95036145 | |
| 0.05 | 0.38662651 | | 0.64 | | 0.91433735 | |
| 0.025 | 0.28807229 | | 0.53777108 | | 0.86638554 | |
| Относительно альтернативы | | | | | | |
| 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 | | | | | | |
| 0.1 | 0.29319277 | | 0.50246988 | | 0.83975904 | |
| 0.05 | 0.16210843 | | 0.31186747 | | 0.69596386 | |
| 0.025 | 0.08355422 | | 0.18084337 | | 0.52138554 | |
| Относительно альтернативы | | | | | | |
| 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 | | | | | | |
| 0.1 | | 1.0 | | 1.0 | | 1.0 |
| 0.05 | | 1.0 | | 1.0 | | 1.0 |
| 0.025 | | 1.0 | | 1.0 | | 1.0 |

**В Н3 на n=m=500 мощность получилась выше, чем на данных без округления**

Выводы по АД сравнивая округления

На данных, округленных до одного и двух знаков после запятой по альтернативе H1 мощность оказалась выше чем на данных без округления.

## **Выводы**

Выводы по всем критериям, сравнивая одинаковые округления

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 H1наибольшую 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, который на данных ограниченных до целых чисел, абсолютно был не способен распознать распределения в гипотезах H1-H3.

# **Заключение**

# **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. Stist. 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**