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

**Оглавление**

[**Введение** 3](#_Toc514019298)

[**1.** **Критерии проверки однородности законов распределения** 5](#_Toc514019299)

[**1.1.** **Общая постановка** 5](#_Toc514019300)

[**1.2.** **Критерий Смирнова** 5](#_Toc514019301)

[**1.3.** **Критерий Лемана-Розенблатта** 6](#_Toc514019302)

[**1.4.** **Критерий Андерсона-Дарлинга** 7](#_Toc514019303)

[**2.** **Исследование распределения статистик** 9](#_Toc514019304)

[**3.** **Исследование мощностей критериев** 9](#_Toc514019305)

[**Заключение** 9](#_Toc514019306)

[**Список литературы** 9](#_Toc514019307)

[**Приложение А. Программные модули** 9](#_Toc514019308)

# **Введение**

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.

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 [1, 2]. Таким образом, целью данной работы является исследование критериев однородности на данных ограниченной точности. Исследования проводятся с использованием метода компьютерного моделирования.

# **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 [3]. 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 [4]:

,

,

.

Если гипотеза  справедлива, то при неограниченном увеличении объемов выборок [4] , т. е. статис­тика



в пределе подчиняется распределению Колмогорова  [4] 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  [4]. 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 [5] и исследован в [6]. Статистика критерия имеет вид [4]

,

где  – эмпирическая функция распределения, построенная по вариационному ряду объединения двух выборок. Статистика  используется в форме [4]

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

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

В [7] было показано, что статистика (1.1) в пределе распределена как :

.

Функция распределения  имеет вид [4]:



,

где  – модифицированные функции Бесселя вида

.

В отличие от критерия Смирнова распределение статистики  быстро сходится к предельному  [4]

## **Критерий Андерсона-Дарлинга**

Двухвыборочный критерий Андерсона–Дарлинга (критерий однородности) рассмотрен в работе [8]. Статистика критерия определяется выражением

.

Для выборок непрерывных случайных величин выражение для этой статистики принимает простой вид [8]

|  |  |
| --- | --- |
| , | (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  [8], которое является предельным для статистики критерия согласия Андерсона–Дарлинга [4]. Функция распределения , имеет вид [4]



.

# **Исследование распределения статистик**

**АД**

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

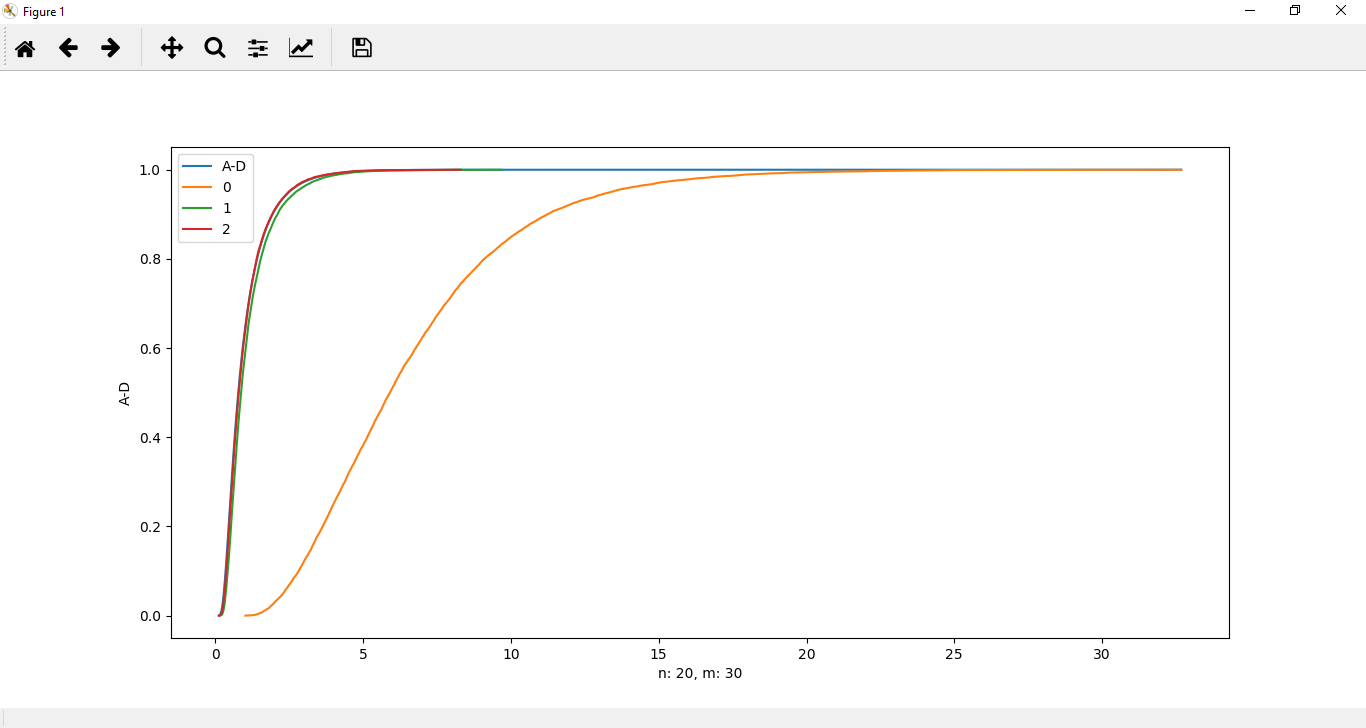
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | N=16600, n=20, m=30 | | 30, 30 | | 30, 40 | |
|  | Расстояние по Колмогорову | \* |  |  |  |  |
| 0 | 0.8865724614873727 | 5.0 | 0.971939404837819 | 5.0 | 0.9613735405573581 | 6.0 |
| 1 | 0.08431480230841298 | 28 | 0.12327215961403043 | 31.5 | 0.09601900915710787 | 33.5 |
| 2 | 0.022567759200884496 | 47 | 0.026233336804103197 | 55.5 | 0.01749731042612679 | 63.0 |

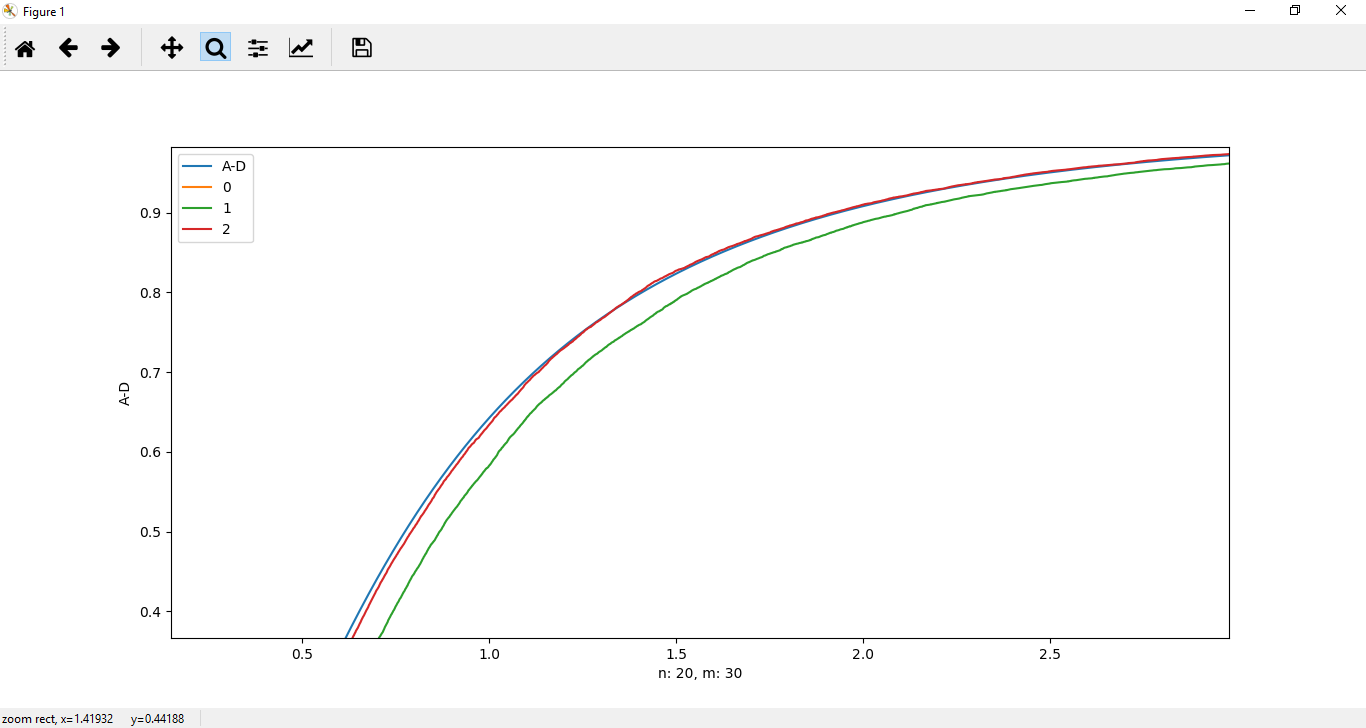
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 30, 50 | | 70, 100 | | 200, 200 | |
|  |  |  |  |  |  |  |
| 0 | 0.9513233366841691 | 6.0 | 0.9991038944859558 | 6.0 | 0.9999999999999677 | 7.0 |
| 1 | 0.09355808370624932 | 34.5 | 0.1368293454822288 | 41.5 | 0.3177606900344735 | 50.0 |
| 2 | 0.017520768482571808 | 71.0 | 0.014034539937959112 | 135.5 | 0.016657660678184527 | 241.0 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | n: 500, m: 500 | | n: 1000, m: 1000 | | n: 2000, m: 2000 | |
|  |  |  |  |  |  |  |
| 0 | 1.0000000010906982 | 7.0 | 49783007492.5392 | 8.0 | 7.325497982177374e+24 | 8.0 |
| 1 | 0.5580592698629411 | 59.0 | 0.7813174951763587 | 64.5 | 0.9406704541308208 | 66.5 |
| 2 | 0.02413591555767497 | 377.0 | 0.026778729840812754 | 442.0 | 0.041710861946732514 | 510.0 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | n: 5000, m: 5000 | |  | |  | |
|  |  |  |  |  |  |  |
| 0 | 1.6864794476131036e+86 | 9 |  |  |  |  |
| 1 | 0.9984727668727141 | 72.5 |  |  |  |  |
| 2 | 0.08155058951357014 | 576.5 |  |  |  |  |

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





**Лем-Роз**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | N=16600, n=20, m=30 | | 30, 30 | | 30, 40 | |
|  | Расстояние по Колмогорову | \* |  |  |  |  |
| 0 | 0.2233477694432367 | 6.0 | 0.11695639111559329 | 6.0 | 0.13034441570224864 | 5.5 |
| 1 | 0.02697723285461584 | 27.5 | 0.0173855133189037 | 30.5 | 0.015110378505309996 | 33.0 |
| 2 | 0.021285450329402766 | 47.5 | 0.01861528625271719 | 55.5 | 0.014258353406632063 | 63.0 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 30, 50 | | 70, 100 | | 200, 200 | |
|  |  |  |  |  |  |  |
| 0 | 0.5475105174962243 | 6.0 | 0.5587545268341145 | 6.5 | 0.11536710664079695 | 7.0 |
| 1 | 0.0312338093564187 | 34.5 | 0.02186134582883903 | 43.5 | 0.006769699360278936 | 49.0 |
| 2 | 0.01397052387315445 | 71.5 | 0.010353448928988152 | 139.0 | 0.006126600306398777 | 243.5 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | n: 500, m: 500 | | n: 1000, m: 1000 | | n: 2000, m: 2000 | |
|  |  |  |  |  |  |  |
| 0 | 0.11339391175792744 | 7.0 | 0.11519912345224 | 8.0 | 0.11978353880223283 | 9.0 |
| 1 | 0.006149829639210902 | 56.5 | 0.005304291721869081 | 62.5 | 0.010790795109519302 | 67.5 |
| 2 | 0.0034988541215064117 | 369.5 | 0.005246511650933039 | 448.5 | 0.008234065731409945 | 510.5 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | n: 500, m: 1000 | | n: 500, m: 2000 | |  | |
|  |  |  |  |  |  |  |
| 0 |  |  |  |  |  |  |
| 1 |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |

**Смирнов**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | N=16600, n=20, m=30 | | 30, 30 | | 30, 40 | |
|  | Расстояние по Колмогорову | \* |  |  |  |  |
| 0 | 0.5225715805763322 | 5.0 | 0.5646230429354729 | 5.5 | 0.511252231412543 | 6.0 |
| 1 | 0.1773358423198802 | 28.5 | 0.2557694018270209 | 31.0 | 0.1635153609583173 | 33.0 |
| 2 | 0.10245632424759099 | 46.0 | 0.17625132953786432 | 56.0 | 0.08060662373472738 | 64.0 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 30, 50 | | 70, 100 | | 200, 200 | |
|  |  |  |  |  |  |  |
| 0 | 0.5089345868104809 | 6.0 | 0.5072942563445924 | 7.0 | 0.540092677385176 | 7.0 |
| 1 | 0.16659448091948403 | 33.5 | 0.15735952708057133 | 44.0 | 0.2033904740447694 | 50.5 |
| 2 | 0.08069086646165269 | 73.0 | 0.0575481039293369 | 136.5 | 0.09244206408266703 | 246.5 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 70, 110 | | 70, 130 | |  | |
|  |  |  |  |  |  |  |
| 0 | 0.5090442970635045 | 7.0 | 0.5041810957534603 | 6.5 |  |  |
| 1 | 0.15996048792424555 | 43.5 | 0.15479612518033248 | 46.0 |  |  |
| 2 | 0.052490608406173256 | 141.0 | 0.047139872823409645 | 156.5 |  |  |

# **Исследование мощностей критериев**

# **Заключение**

# **Список литературы**

1. Лемешко Б.Ю. Критерии проверки гипотез об однородности. Руководство по применению / Б.Ю. Лемешко. – М: ИНФРА–М, 2016. – 207 с.
2. 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.
3. Смирнов Н. В. Оценка расхождения между эмпирическими кривыми распределения в двух независимых выборках / Н. В. Смирнов // Бюл. МГУ, Серия А. – 1939. – Т. 2, № 2. – С. 3–14.
4. Большев Л. Н. Таблицы математической статистики / Л. Н. Большев, Н. В. Смирнов. – М. : Наука, 1983. – 416 с.
5. Lehmann E. L. Consistency and unbiasedness of certain nonparametric tests / E. L. Lehmann // Ann. Math. Statist. – 1951. – Vol. 22, № 1. – P. 165–179.
6. 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.
7. Rosenblatt M. Limit theorems associated with variants of the von Mises statistic / M. Rosenblatt // Ann. Math. Statist. – 1952. – Vol. 23. – P. 617–623.
8. Pettitt A.N. A two-sample Anderson-Darling rank statistic // Biometrika. 1976. Vol. 63. No.1. P. 161-168.

# **Приложение А. Программные модули**