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**

# **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**