Федеральное государственное бюджетное образовательное учреждение

высшего образования

«Новосибирский государственный технический университет»

Кафедра теоретической и прикладной информатики

**Отчет ПО ПРАКТИКЕ**

Производственная практика по получению профессиональных умений и опыта

(наименование практики в соответствии с учебным планом)

профессиональной деятельности в период с 01.09.2017 по 30.12.2017

Направление подготовки: Прикладная математика и информатика (01.04.02)

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

|  |  |
| --- | --- |
| Выполнил:  Студент Федосов Д. Н.  (Ф.И.О.)  Группа ПММ-61  Факультет ПМИ  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  подпись  «\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2017 г. | Проверил:  Руководитель от НГТУ Чимитова Е. В.  (Ф.И.О.)  Балл: \_\_\_\_\_\_\_\_\_\_, ECTS\_\_\_\_\_\_\_\_\_\_\_\_\_,  Оценка \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  «отлично», «хорошо», «удовлетворительно», «неуд.»  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  подпись  «\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2017 г. |

Новосибирск 2017

Оглавление

[Введение 3](#_Toc515515858)

[1. Критерии проверки однородности законов распределения 5](#_Toc515515859)

[1.1. Общая постановка 5](#_Toc515515860)

[1.2. Критерий Андерсона-Дарлинга 5](#_Toc515515861)

[2. Исследование распределений статистик критериев однородности на данных ограниченной точности 7](#_Toc515515862)

[2.1. Исследование распределений статистик 7](#_Toc515515863)

[Список литературы 8](#_Toc515515864)

# 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 работы является разработка математического и алгоритмического обеспечения для исследования критерия однородности Андерсона-Дарлинга на данных ограниченной точности.

Для достижения сформулированной цели были поставлены и решены следующие задачи:

* исследование распределения статистики критерия однородности: Андерсона-Дарлинга на данных ограниченной точности;
* сравнительный анализ распределения статистики критерия с предельной функцией распределения;

# 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 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.

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

В таблицах ниже (2.1-2.5) представлены значения расстояний между эмпирическими и предельными функциями распределения статистик, рассчитанные по метрике Колмогорова. 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,
* \* - среднее число различных значений в объединенной выборке,
* **** - расстояние между эмпирическими и предельными функциями распределения статистик критерия по метрике Колмогорова.

Таблица 2.1 – Результаты для критерия однородности Андерсона-Дарлинга, округление до 2 знаков, n=m.

|  |  |  |
| --- | --- | --- |
| **n, m** |  | **\*** |
| 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 |

По результатам, представленным в таблице 2.1, видно, что между n=m=2000 и n=m=5000 расстояние становится большим чем 0.05 на данных, округленных до двух знаков.

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.

Таблица 2.2 –

|  |  |  |
| --- | --- | --- |
| **n, m** |  | **\*** |
| 30, 30 | 0.03 | 55.5 |
| 30, 40 | 0.02 | 63.0 |
| 30, 50 | 0.02 | 71.0 |

Таблица 2.3 –

|  |  |  |
| --- | --- | --- |
| **n, m** |  | **\*** |
| 500, 500 | 0.02 | 377.0 |
| 500, 1000 | 0.01 | 422.0 |
| 500, 2000 | 0.01 | 465.0 |
| 500, 5000 | 0.01 | 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.

Дисперсия увеличена для нормального закона, чтобы получить больше различных значений в совместной выборке для данных округленных до целых чисел и одного знака после запятой.

Для округления 1 знака после запятой, дисперсия = 10.

Таблица 2.4 –

|  |  |  |
| --- | --- | --- |
| **n, m** |  | **\*** |
| 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 |

Для округления до целых, дисперсия = 80.

Таблица 2.5 –

|  |  |  |
| --- | --- | --- |
| **n, m** |  | **\*** |
| 200, 200 | 0.01 | 221 |
| 500, 500 | 0.03 | 321.0 |
| 1000, 1000 | 0.04 | 374.0 |
| 2000, 2000 | 0.06 | 421.5 |
| 5000, 5000 | 0.12 | 475.0 |

# 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.