МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ

ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ

ВЫСШЕГО ОБРАЗОВАНИЯ

«НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Кафедра систем сбора и обработки данных

**КУРСОВОЙ ПРОЕКТ**

по дисциплине: Компьютерные технологии моделирования и анализа данных

на тему: Проверка гипотезы о виде распределения

Вариант №26

Факультет: ФПМИ

Группа: ПММ-21

Выполнил: Сухих А.С.

Проверил: д.т.н., профессор Лемешко Борис Юрьевич

Дата выполнения: 09.01.23

Отметка о защите:

Новосибирск 2022

**Введение**

Цель работы: Знакомство с современными тенденциями развития аппарата прикладной математической статистики и состоянием программного обеспечения задач статистического анализа. Освоение методов статистического моделирования как средства исследования и развития аппарата прикладной математической статистики. Исследование особенностей методов проверки статистических гипотез. Закрепление навыков проведения самостоятельных исследований.

Наиболее распространенным законом распределения данных, наблюдаемых в реальном мире, является закон нормального распределения. Принадлежность выборки нормальному закону часто может являться требованием для применения различных методов математической статистики. Определить эту принадлежность можно с помощью проверки на нормальность с помощью различных критериев. В данной работе будет исследовано применение критерия Дэвида–Хартли–Пирсона, основанного на отношении размаха выборки к её выборочному стандартному отклонению.

**Постановка задачи**

Согласно варианту №26 необходимо выполнить проверку гипотезы о нормальности распределения по критерию Дэвида–Хартли–Пирсона. В данном критерии рассматривается отношение размаха выборки к выборочному стандартному отклонению и его статистика имеет вид

,

где *R = xmax - xmin* — размах выборки,

— несмещенная оценка дисперсии

Критерий двусторонний: гипотеза о нормальности распределения отвергается, если *U < Uα/2* или *U > U1-α/2*. Вывод о справедливости гипотезы можно сделать, сравнив вычисленную статистику с таблицей процентных точек, представленной в Приложении 1.

Также необходимо вычислить достигаемый уровень значимости статистики критерия. Поскольку для данного критерия неизвестна предельная статистика достигаемый уровень значимости будет вычислен методом Монте-Карло:

1. Вычислить *S = S(Xn)* — статистику критерия по выборке.

2. Установить *m =* 0.

3. Сгенерировать выборку *Yn* при верной гипотезе H0.

4. Вычислить значения *S*(*Yn*).

5. Если *S(Yn)* > *S(Xn)*, то *m = m+1.*

6. Повторять шаги 3-5 *N* раз.

Оценка достигаемого уровня значимости

Количество повторений N определяется из выражения

,

где γ — доверительная вероятность, tγ — квантиль стандартного нормального распределения, ε — погрешность моделирования, p — вероятность попадания в доверительную область.

**Аналитический обзор**

Критерий Дэвида-Хартли-Пирсона был предложен в 1954 году как результат совместного исследования. На момент исследования данного критерия уже были представлены несколько исследований касаемо стьюдентизированного диапазона являющегося отношением размаха выборки размера *n* генеральной совокупности со стандартным отклонением σ к независимой среднеквадратичной оценке *σ*. Также уже был представлен критерий Гири (1933).

Однако в исследовании нового критерия фокус был направлен на изучение иной статистики, связанной с отношением размаха выборки к стандартному отклонению, при этом обе величины отношения вычислялись по одной и той же выборке из *n* наблюдений. Так, ранее исследованное отношение могло быть использовано для быстрого анализа отклонения, а статистика нового критерия, которая зависела только от конкретной выборки, позволяла выявлять неоднородность данных и отклонение от нормального распределения.

Предпосылкой к появлению данного критерия стала переписка между одним из исследователей критерия и доктором Джозефом Берксоном, проводившим рутинную проверку данных, сравнивая размах и оценки стандартного отклонения *σ.* В результате переписки начались эмпирические исследования связи между оценками для определения стандартной ошибки разницы между ними. Также за несколько лет до этого в 1946 году Г. А. Бейкер показал возможность использования такого отношения как критерий однородности, продемонстрировав на эксперименте с искусственной выборкой, что это соотношение будет значительно зависеть от формы родительской совокупности.

**Результаты исследований**

С помощью языка программирования Python было разработано программное обеспечение для проверки гипотезы о соответствии нормальному распределению с применением критерия Дэвида-Хартли-Пирсона. Программное обеспечение представляет собой консольное приложение с возможностью генерации выборки согласно нормальному закону распределения с заданным объемом выборки и параметрами сдвига и масштаба и возможностью проверки выборки на нормальность.

Исследования были проведены на примере выборок, смоделированных в соответствии с нормальным законом, а также на примере реальных данных с различными объемами выборок.

Во всех исследованиях ошибка первого рода задаётся равной 0,05, а количество повторений в методе Монте-Карло равно 16600, что обеспечивает относительную погрешность моделирования 0,01.

Результаты тестирования программы на смоделированных выборках показаны в таблице 1. Были смоделированы выборки:

* стандартное нормального распределения (сдвиг 0, масштаб 1)
* нормальное со сдвигом 3 и масштабом 1
* нормальное со сдвигом 0 и масштабом 5

Каждая выборка представлена объемом n=100 и n=1000. Проверяется гипотеза о соответствии стандартному нормальному распределению.

Таблица 1.

|  |  |  |  |
| --- | --- | --- | --- |
| Выборка | Статистика U | p-value | Гипотеза отклоняется? |
| Нормальное N(0,1), n=100 | 5.04374 | 0.90651 | не отклоняется |
| N(3,1), n=100 | 5.10267 | 0.80578 | не отклоняется |
| N(0,5), n=100 | 4.92366 | 0.89819 | не отклоняется |
| N(0,1), n=1000 | 6.55909 | 0.79723 | не отклоняется |
| N(3,1), n=1000 | 6.24 | 0.65783 | не отклоняется |
| N(0,5), n=1000 | 6.86053 | 0.40747 | не отклоняется |

Проверить справедливость гипотезы также можно с помощью таблицы процентных точек (приложение 1). Для n=100 при вероятности ошибки первого рода 0,05 достоверной областью является интервал значений (4,206; 6,112). Как видно в таблице 1 все значения для n=100 попадают в данный интервал и, следовательно, гипотеза не отклоняется.

По приведенным выше результатам видно, что критерий для всех выборок не отверг гипотезу о нормальности выборки, что верно, так как все выборки сгенерированы согласно нормальному закону. Однако критерий не смог отличить выборки с измененными параметрами сдвига или масштаба от стандартного нормального распределения даже на сравнительно больших (n=1000) объемах выборки.

Данный критерий присутствует также в программе ISW. Результаты проверки тех же выборок на нормальность в ISW приведены в таблице 2.

**Список использованных источников**

1. David H. A., Hartley H. O., Pearson E. S. The distribution of the ratio, in a single normal sample, of range to standard deviation //Biometrika. – 1954. – Т. 41. – №. 3/4. – С. 482-493.

2. Кобзарь А. И. Прикладная математическая статистика. – 2006.

3. Лемешко Б.Ю., Рогожников А.П. Исследование особенностей и мощности некоторых критериев нормальности // Метрология. 2009. № 4. – С. 3-24.

4. Лемешко Б. Ю., Постовалов С. Н., Лемешко С. Б. Компьютерные технологии анализа данных и исследования статистических законномерностей. – 2007.

**ПРИЛОЖЕНИЯ**

Приложение 1. Таблица процентных точек статистики критерия Дэвида-Хартли-Пирсона для некоторых объемов выборки.

