*Мат. VI международной конференции "Актуальные проблемы электронного приборостроения" АПЭП-2002. Новосибирск, 2002. - Т.6. - С.32-35.*

# РАСПРЕДЕЛЕНИЯ СТАТИСТИК КОРРЕЛЯЦИОННОГО АНАЛИЗА ПРИ ОТКЛОНЕНИИ МНОГОМЕРНОГО ЗАКОНА ОТ НОРМАЛЬНОГО[1]

Б.Ю. Лемешко, С.С. Помадин

Новосибирский государственный технический университет
Новосибирск, Россия. e-mail: headrd@fpm.ami.nstu.ru, ser@fpm.ami.nstu.ru

**Аннотация.** Методами статистического моделирования исследуются распределения различных статистик корреляционного анализа. Строятся модели распределений статистик при наблюдаемых законах, отличающихся от многомерного нормального.

В различных приложениях статистического анализа многомерных данных одну из ключевых позиций занимают задачи корреляционного анализа. В процессе решения этих задач вычисляются оценки коэффициентов и матриц парной, частной и множественной корреляции; проверяются различные статистические гипотезы относительно параметров многомерного распределения и коэффициентов корреляции; выявляется наличие и характер взаимосвязи величин, взаимозависимости величин при устранении влияния некоторой совокупности других или зависимости одной случайной величины от группы величин. На основании результатов корреляционного анализа может делаться вывод о наличии и характере функциональной зависимости или о предпочтительности для описания исследуемого объекта регрессионной модели того или иного вида.

В основе классического аппарата корреляционного анализа лежит предположение о принадлежности наблюдаемого случайного вектора многомерному нормальному закону. Базируясь на этом, получены предельные распределения статистик, используемых в корреляционном анализе. На практике предпосылки классического корреляционного анализа выполняются далеко не всегда. Естественно, возникает вопрос о справедливости выводов, получаемых на основании классического аппарата, при нарушении предположения о нормальности.

Ответить на эти вопросы, опираясь только на аналитические методы, чрезвычайно сложно из-за нетривиальности задач. Поэтому в основу проводимого исследования положена развиваемая методика компьютерного анализа статистических закономерностей. С ее помощью проводились исследования [1] распределений статистик, связанных с проверкой гипотез классического корреляционного анализа [2].

Подчеркнем, что все распределения рассмотренных в [1-2] статистик являются предельными при наблюдении многомерного нормального закона. Нельзя сказать заранее, что произойдет с предельными распределениями этих статистик и насколько будут справедливы выводы, формулируемые на основании решения задач классического корреляционного анализа, если наблюдаемый закон отличается от многомерного нормального.

Одной из проблем компьютерного моделирования распределений статистик корреляционного анализа является задача генерирования последовательностей псевдослучайных векторов по законам, “заданным образом” отличающимся от многомерного нормального. Удачный подход к решению данной проблемы был предложен в [3].

Проведенные в [1,4] исследования показали, что распределения ряда статистик, вычисляемых в корреляционном анализе при проверке гипотез о векторе математических ожиданий и о парных, частных и множественных коэффициентах корреляции, при существенном от-

[1] Работа выполнена при поддержке Российского фонда фундаментальных исследований (проект № 00-01-00913)

личии наблюдаемого закона от нормального *незначимо отличаются от предельных распределений, полученных в классическом случае*. Результаты моделирования распределений рассматриваемых статистик в случае принадлежности многомерных величин законам, отличающимся от нормального, показали, что эмпирические распределения статистик очень хорошо согласуются с предельными законами, полученными в предположении о нормальности многомерного случайного вектора: нет, например, оснований для отказа от использования в качестве предельных в соответствующих случаях распределений $\chi^2$–, $F$ – или нормального.

$\chi^2$ – распределение представляет собой частный случай гамма–распределения, $F$ – распределение Фишера – частный случай бета–распределения 2-го рода. Гамма–распределения, бета–распределения 2-го рода и нормальные распределения всегда оказываются хорошими моделями, описывающими эмпирические распределения соответствующих статистик корреляционного анализа, получаемые в результате моделирования. Если, например, действительно $\chi^2$ – распределение является предельным распределением некоторой статистики и в том случае, когда нарушается предположение о нормальности наблюдаемой многомерной величины, и мы для выравнивания эмпирического распределения статистики каждый раз будем использовать гамма–распределение, оценивая его параметры по выборке статистики, то модель гамма–распределения с параметрами, полученными усреднением по множеству экспериментов, должна привести нас к соответствующему $\chi^2$ – распределению.

В данной работе предпринята такая попытка уточнения моделей распределений некоторых статистик корреляционного анализа усреднением этих моделей по совокупности смоделированных эмпирических распределений статистик. Модели строились при объемах выборок статистик в 1000 наблюдений. Параметры моделей усреднялись по нескольким десяткам экспериментов. В основе датчика моделирования многомерных случайных величин, "заданным образом" отличающихся от нормального [3], лежит экспоненциальное семейство распределений с плотностью

$$f(x) = \frac{\lambda}{2\sqrt{2}\theta_1 \Gamma(1/\lambda)} \exp\left( -\left( \frac{|x-\theta_0|}{\sqrt{2}\theta_1} \right)^{\lambda} \right).$$

С помощью параметра $\lambda$ можно регулировать "удаление" генерируемого многомерного закона от нормального, делая его более плосковершинным по сравнению с нормальным при $\lambda > 2$ и более островершинным при $0 < \lambda < 2$. При $\lambda = 2$ моделируется многомерное нормальное распределение.

Предельным распределением статистики, используемой при проверке гипотезы о равенстве вектора математического ожидания некоторому номинальному, является $\chi^2_m$ –распределение, где $m$ – размерность многомерного вектора. Это соответствует гамма–распределению с плотностью $f(x) = \frac{1}{\sigma^{\theta}\Gamma(\theta)} x^{\theta-1} e^{-x/\sigma}$ с параметром формы $\theta = m/2$ и масштаба $\sigma = 2$. В таблице 1 представлены усредненные по 50 смоделированным выборкам значения параметров гамма–распределения при $m$ = 3, а на рис. 1 – соответствующие функции распределения статистик. Очевидно, что значения параметров сходятся к значениям 2 и 1,5 соответственно, а функции распределения практически совпадают (рис.1).

Таблица 1.

| Параметры гамма–распределения | $\lambda = 1$ | $\lambda = 2$ | $\lambda = 5$ | $\lambda = 10$ |
|---|---|---|---|---|
| $\sigma$ | 2,0157 | 2,0000 | 1,9892 | 1,9644 |
| $\theta$ | 1,4737 | 1,5000 | 1,5137 | 1,5302 |

При проверке аналогичной гипотезы при неизвестной ковариационной матрице предельным распределение статистики является $F_{m,n-m}$ – распределение. Данному случаю при

размерности вектора $m = 3$ и объеме выборки $n = 30$ соответствует бета–распределение 2-го рода, плотность которого имеет вид $f(x) = \dfrac{\theta_2}{B(\theta_0, \theta_1)} \dfrac{[\theta_2 (x-\mu)]^{\theta_0 - 1}}{[1+\theta_2 (x-\mu)]^{\theta_0+\theta_1}}$, с масштабным параметром $\theta_2 = \dfrac{n-m}{m}$, параметрами формы $\theta_0 = \dfrac{m}{2}$ и $\theta_1 = \dfrac{n-m}{2}$. Представленные в таблице 2 усредненные по 50 смоделированным выборкам значения параметров бета–распределения (при $m = 3$ и $n = 30$) показывают аналогичную картину сходимости.

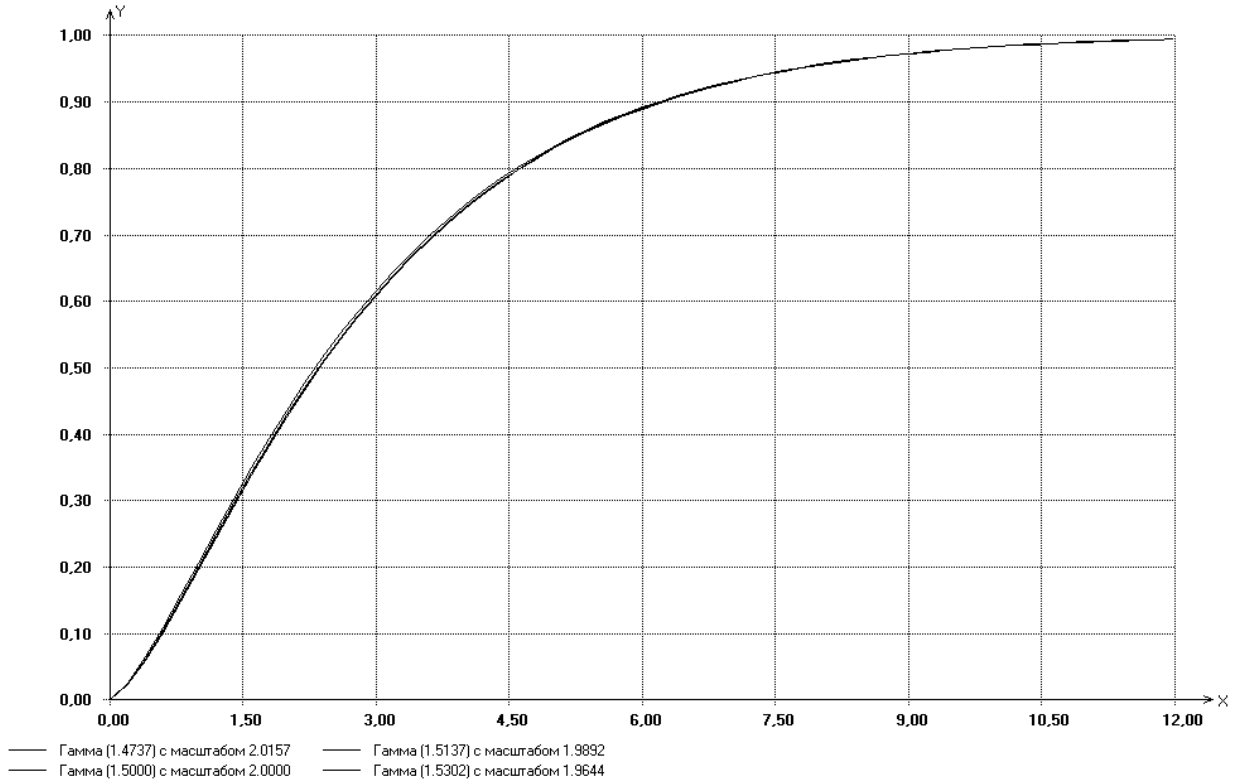


Рис. 1.

Таблица 2.

| Параметры бета–распределения | $\lambda = 1$ | $\lambda = 2$ | $\lambda = 5$ | $\lambda = 10$ |
|---|---|---|---|---|
| $\theta_2$ | 8,8628 | 9,0000 | 9,0619 | 9,1576 |
| $\theta_0$ | 1,5636 | 1,5000 | 1,4861 | 1,4627 |
| $\theta_1$ | 13,7685 | 13,5000 | 13,4401 | 13,3474 |

Предельным распределением статистик, используемых при проверке гипотез о парном и частном коэффициентах корреляции на равенство их определенному значению, в классическом случае является стандартное нормальное распределение. Таблицы 3 и 4 иллюстрируют сходимость к стандартному нормальному закону распределений исследуемых статистик при многомерных законах, существенно отличающихся от нормального: таблица 3 – в случае проверки гипотез о парном коэффициенте корреляции, а таблица 4 – в случае проверки гипотез о частном коэффициенте корреляции. Как и в предыдущих случаях, усреднение осуществлялось по 50 выборкам статистик.

Предельным распределением статистики, вычисляемой при проверке гипотезы о равенстве нулевому значению множественного коэффициента корреляции, в классическом случае является $F_{m-l,n-m+l-1}$ – распределение (см. [2,5,6]). Как сходятся к данной ситуации рас-

пределения этой же статистики (при $m = 3$, $n = 30$, $l=2$) в случае наблюдаемых законах, отличающихся от многомерного нормального, иллюстрирует таблица 5.

Таблица 3.

| Параметры нормального закона | $\lambda = 1$ | $\lambda = 2$ | $\lambda = 5$ | $\lambda = 10$ |
|---|---|---|---|---|
| $\mu$ | 0,0296 | 0,0000 | 0,0183 | 0,0150 |
| $\sigma$ | 1,0183 | 1,0000 | 0,9851 | 0,9927 |

Таблица 4.

| Параметры нормального закона | $\lambda = 1$ | $\lambda = 2$ | $\lambda = 5$ | $\lambda = 10$ |
|---|---|---|---|---|
| $\mu$ | 0,0070 | 0,0000 | 0,0154 | 0,0051 |
| $\sigma$ | 1,0263 | 1,0000 | 0,9923 | 0,9821 |

Таблица 5.

| Параметры бета–распределения | $\lambda = 1$ | $\lambda = 2$ | $\lambda = 5$ | $\lambda = 10$ |
|---|---|---|---|---|
| $\theta_2$ | 27,9500 | 28,0000 | 28,0089 | 27,9893 |
| $\theta_0$ | 0,4910 | 0,5000 | 0,4969 | 0,5044 |
| $\theta_1$ | 13,6056 | 14,0000 | 13,8450 | 14,0621 |

Таким образом, проведенные в данной работе исследования распределений статистик корреляционного анализа в случае многомерных законов, отличающихся от нормального в достаточно широких пределах, позволяют с бóльшей уверенностью утверждать, что распределения рассматриваемых статистик хорошо описываются предельными распределениями, полученными в классическом корреляционном анализе в предположении о нормальности наблюдаемого вектора. Это существенно расширяет сферу корректного применения методов классического корреляционного анализа в приложениях.

ЛИТЕРАТУРА


1. Лемешко Б.Ю., Помадин С.С. Исследование распределений статистик корреляционного анализа при отклонении многомерного закона от нормального // Тр. V международной конференции “Актуальные проблемы электронного приборостроения” АПЭП-2000. Новосибирск, 2000. – Т. 7. – С. 184-187.
2. Андерсон Т. Введение в многомерный статистический анализ. – М.: Физматгиз, 1963. – 500 с.
3. Лемешко Б.Ю., Помадин С.С. Один подход к моделированию псевдослучайных векторов с “заданными” числовыми характеристиками по законам, отличным от нормального // Материалы МНТК “Информатика и проблемы телекоммуникаций”. Новосибирск, 2002.
4. Лемешко Б.Ю., Помадин С.С., Кузьменко С.В. Программное обеспечение компьютерного исследования статистических закономерностей в задачах корреляционного анализа // Материалы МНТК “Информатика и проблемы телекоммуникаций”. – Новосибирск, 2001. – С. 79.
5. Кендалл М., Стьюарт А. Многомерный статистический анализ и временные ряды. – М.: Наука, 1976. – 736 с.
6. Лемешко Б.Ю. Корреляционный анализ многомерных наблюдений случайных величин: Программная система. – Новосибирск: Изд-во НГТУ, 1995. – 39 с.