Сибирский журнал
индустриальной математики
Июль–сентябрь, 2002. Том V, № 3(11)

УДК 519.237.5

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ НАБЛЮДЕНИЙ МНОГОМЕРНЫХ СЛУЧАЙНЫХ ВЕЛИЧИН ПРИ НАРУШЕНИИ ПРЕДПОЛОЖЕНИЙ О НОРМАЛЬНОСТИ*)

Б. Ю. Лемешко, С. С. Помадин

Для ряда статистик, используемых при проверке гипотез относительно наблюдаемых многомерных величин, показано, что в случае законов, отличающихся от многомерного нормального в достаточно широких пределах (более островершинных или более плосковершинных), значимого изменения предельных распределений статистик не происходит. Эмпирические распределения данных статистик по-прежнему хорошо описываются предельными законами, полученными в классическом корреляционном анализе в предположении о нормальности наблюдаемого вектора. Результаты расширяют сферу корректного применения методов классического корреляционного анализа в приложениях.

Введение

В различных приложениях статистического анализа многомерных величин одну из ключевых позиций занимают задачи корреляционного анализа. В процессе решения этих задач выявляются наличие и характер взаимосвязи величин, их взаимозависимости при устранении влияния совокупности других или зависимости одной случайной величины от группы величин, вычисляются оценки коэффициентов и матриц парной, частной и множественной корреляций, проверяются различные статистические гипотезы относительно параметров многомерного распределения и коэффициентов корреляции. На основании результатов корреляционного анализа может быть сделан вывод о наличии и характере функциональной зависимости или 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. Поэтому возникает вопрос о справедливости выводов, получаемых на основании классического аппарата, при нарушении основного предположения.

Целью данных исследований явилось стремление разобраться, что происходит с распределениями различных статистик корреляционного анализа в ситуациях, если наблюдаемый многомерный закон отличается от нормального. Ответить на поставленный вопрос, используя чисто аналитические методы, чрезвычайно затруднительно из-за нетривиальности возникающих задач. Поэтому в основу проводимых исследований положена развиваемая нами методика компьютерного анализа статистических закономерностей. Методика хорошо зарекомендовала себя при исследовании распределений статистик критериев согласия в случае проверки простых и сложных гипотез [1–6], при исследовании статистических свойств различных оценок [7, 8].

*) Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (код проекта 00–01–00913).

Для подтверждения работоспособности методики в случае многомерных величин было предусмотрено исследование эмпирических распределений статистик классического корреляционного анализа при наблюдении нормального закона. Соответствие в такой ситуации эмпирических распределений, получаемых в процессе моделирования, предельным классическим распределениям статистик должно было послужить доводом, подчеркивающим достоверность результатов в общем случае.

Очевидно, что многомерный нормальный закон далеко не всегда является наилучшей моделью для описания реально наблюдаемых многомерных случайных величин. Однако в литературных источниках очень трудно найти примеры использования в этих целях других математических моделей. Нас интересует вопрос, насколько корректны выводы, формируемые на основании конкретных процедур классического корреляционного анализа, если истинная модель многомерного закона в той или иной мере отличается от нормального, и как такое отличие влияет на распределения исследуемых статистик. Ключевым моментом для исследования распределений статистик корреляционного анализа при некоторых произвольных многомерных законах (отличающихся от нормального) является необходимость моделирования псевдослучайных векторов в соответствии с такими законами. Причем желательно иметь возможность моделирования псевдослучайных векторов по законам с «регулируемым удалением» от многомерного нормального, чтобы проследить соответствующие изменения распределений исследуемых статистик корреляционного анализа.

1. Моделирование многомерных случайных величин

1.1. Моделирование псевдослучайных нормальных векторов. Многомерное нормальное распределение случайного вектора $\bar{X} = \|X_1, X_2, \ldots, X_m\|^T$ размерности m полностью определяется вектором математических ожиданий $\bar{M} = [M_1, M_2, \ldots, M_m]^T$ и ковариационной матрицей $\Sigma = [\sigma_{ij}] = E[(X_i - M_i) \times (X_j - M_j)]$.

Функция плотности многомерного нормального закона имеет вид

$$f(\bar{X}) = \frac{1}{\sqrt{(2\pi)^m |\Sigma|}} e^{-(1/2)(\bar{X}-\bar{M})^T \Sigma^{-1} (\bar{X}-\bar{M})}.$$

Хорошо зарекомендовавший себя алгоритм генерирования псевдослучайных нормальных векторов подробно изложен в [9]. Пусть мы имеем совокупность случайных величин $\{Z_i\}$, $i = \overline{1, m}$, где Z_i подчиняется стандартному нормальному закону с параметрами (0, 1). Тогда вектор $\bar{X}$, распределенный по многомерному нормальному закону с параметрами $\bar{M}$ и Σ, получается через линейное преобразование вида

$$\bar{X} = A\bar{Z} + \bar{M}. \tag{1}$$

Обычно полагают, что A является нижней треугольной матрицей

$$A = \begin{bmatrix} a_{11} & 0 & \ldots & 0 \\ a_{21} & a_{22} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ a_{m1} & a_{m2} & \ldots & a_{mm} \end{bmatrix},$$

коэффициенты a_{ij} которой определяются рекуррентной процедурой

$$a_{ij} = \frac{\sigma_{ij} - \sum\limits_{k=1}^{j-1} a_{ik} a_{jk}}{\sqrt{\sigma_{jj} - \sum\limits_{k=1}^{j-1} a_{jk}^2}}, \quad 1 \le j \le i \le m. \tag{2}$$

1.2. Моделирование многомерных законов, отличных от нормального. Процедуру моделирования многомерных величин, распределенных по законам, отличных от нормального, с заданными математическим ожиданием и ковариационной матрицей предложено [10] реализовать в соответствии с описанным выше алгоритмом. При этом совокупность $\{Z_i\}$, $i = \overline{1,m}$, формируется уже не по стандартному нормальному закону, а в соответствии с некоторым одномерным законом распределения с нулевым математическим ожиданием и единичной дисперсией. Затем заданная матрица Σ раскладывается по формуле (2) и осуществляется преобразование (1). На выходе мы имеем некоторый многомерный закон, отличный от нормального закона, с известным математическим ожиданием, но, вообще говоря, с неизвестной ковариационной матрицей, так как ковариационная матрица смоделированного закона не совпадает с используемой при моделировании матрицей Σ.

Для моделирования различных совокупностей $\{Z_i\}$, $i = \overline{1,m}$, удобно использовать экспоненциальное семейство распределений с плотностью

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

где λ — параметр формы, так как оно охватывает целый класс симметричных распределений. Частными случаями данного закона являются распределение Лапласа ($\lambda = 1$), нормальное ($\lambda = 2$), предельными — распределение Коши ($\lambda \to 0$) и равномерное ($\lambda \to +\infty$). Рис. 1 иллюстрирует изменение функции плотности семейства экспоненциальных распределений при изменении параметра формы от 0,5 до 10. С помощью параметра формы λ мы можем задавать непрерывное «удаление» моделируемого (наблюдаемого) многомерного закона от нормального, делая его более плосковершинным по сравнению с нормальным при $\lambda > 2$ или более островершинным при $0 < \lambda < 2$. При $\lambda = 2$ будут формироваться псевдослучайные векторы $\bar{X}$ в соответствии с нормальным законом.

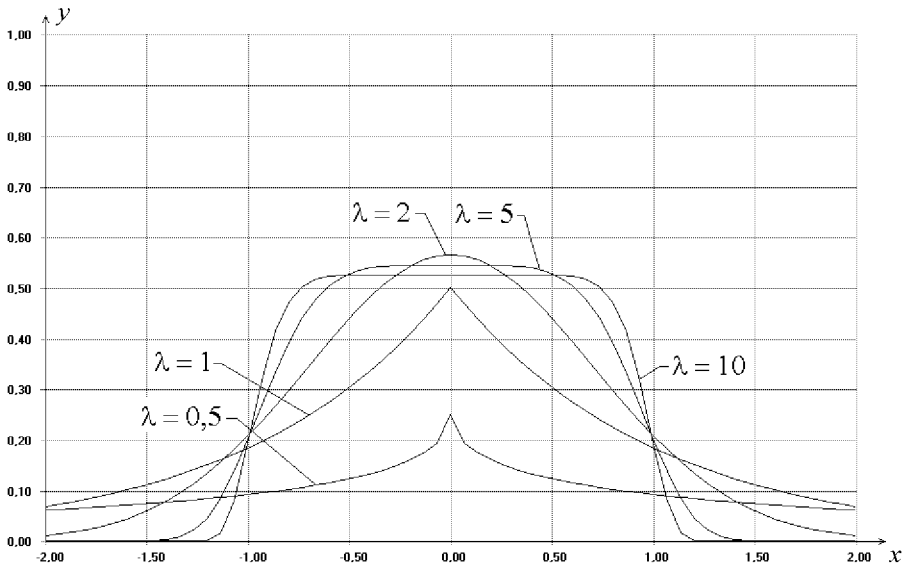


Рис. 1. Функции плотности экспоненциального семейства распределений при различных параметрах формы

К сожалению, такая процедура не позволяет нам моделировать многомерный закон с некоторой произвольной функцией распределения, с заданными математическим ожиданием и ковариационной матрицей и который находится на «заданном» расстоянии (определяемом в смысле некоторой меры) от многомерного нормального закона. Однако мы можем построить датчик, генерирующий псевдослучайные векторы по закону, отличающемуся от нормального (в соответствии с процессом моделирования), с известными математическим ожиданием и ковариационной матрицей. При этом вектор математического ожидания и ковариационная матрица определяются на основании исследования свойств полученного датчика (при заданных $\bar{M}$, Σ и λ). Для определения «истинной» ковариационной матрицы моделируемого многомерного закона нами использовались оценки максимального правдоподобия (ОМП), усредняемые по множеству проведенных экспериментов.

Таким образом, нами решалась задача не по моделированию закона с заданными математическим ожиданием и ковариационной матрицей, а задача по моделированию закона с математическим ожиданием и ковариационной матрицей, истинные значения которых уточнялись в процессе исследования многомерного датчика. Этого, вообще говоря, достаточно для целей настоящего исследования.

На рис. 2 приведен вид функций плотности, получаемых в случае моделирования двумерных векторов при $\lambda = 2$ (плотность нормального закона; слева) и при $\lambda = 10$ (справа). Как видим, во втором случае полученное плосковершинное распределение существенно отличается от нормального.

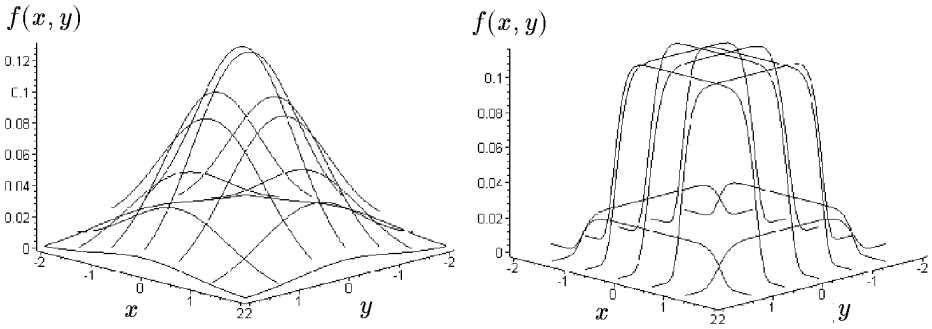


Рис. 2. Смоделированные плотности двумерного закона, построенного при значениях параметра формы $\lambda = 2$ (слева) и $\lambda = 10$ (справа)

2. Исследуемые задачи проверки статистических гипотез классического корреляционного анализа

Пусть $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_n$ — выборка из n наблюдений m-мерного случайного вектора; $\bar{M} = [M_i]_{i=1}^m$ — математическое ожидание случайного вектора $\bar{X}$; $\Sigma = [\sigma_{ij}]_{i,j=1}^m$ — ковариационная матрица случайного вектора $\bar{X}$; $\widehat{\bar{M}}$ и $\widehat{\Sigma}$ — ОМП математического ожидания и ковариационной матрицы

$$\widehat{\bar{M}} = \frac{1}{n}\sum_{i=1}^{n}\bar{X}_i, \quad \widehat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{X}_i - \widehat{\bar{M}})(\bar{X}_i - \widehat{\bar{M}})^T.$$

Основное множество задач проверки статистических гипотез в классическом корреляционном анализе ($\bar{X}$ принадлежит нормальному закону) касается проверки гипотез о векторе математического ожидания, ковариационной матрице, парных, частных и множественных коэффициентах корреляции [11]. Все эти задачи реализованы в системе [12].

2.1. Проверка гипотез о равенстве математического ожидания некоторому известному вектору. Проверяемая гипотеза H_0 имеет вид $\bar{M} = \bar{M}_0$, где $\bar{M}_0$ — номинальное значение вектора математических ожиданий. Ковариационная матрица может быть известной или неизвестной.

1. Ковариационная матрица Σ известна. В этом случае статистика

$$X_m^2 = n(\widehat{\bar{M}} - \bar{M}_0)^T \Sigma^{-1} (\widehat{\bar{M}} - \bar{M}_0) \tag{3}$$

при справедливой гипотезе H_0 в качестве предельного распределения $G\big(X_m^2|H_0\big)$ имеет χ_m^2-распределение с числом степеней свободы m [11].

2. Ковариационная матрица Σ неизвестна. В этом случае используется статистика

$$T^2 = \frac{n(n-m)}{m(n-1)} (\widehat{\bar{M}} - \bar{M}_0)^T \widehat{\Sigma}^{-1} (\widehat{\bar{M}} - \bar{M}_0), \tag{4}$$

которая при справедливости гипотезы H_0 в пределе подчиняется распределению Фишера с параметрами m и $n-m$: $G(T^2|H_0) = F_{m,n-m}$ [11].

2.2. Проверка гипотез о коэффициентах парной корреляции. Взаимозависимость двух компонент случайного вектора характеризуется парным коэффициентом корреляции $r_{ij} = \dfrac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$. Относительно парного коэффициента корреляции могут проверяться два вида гипотез H_0: о значимости корреляции ($r_{ij} = 0$) и равенстве коэффициента корреляции номинальному значению ($r_{ij} = r_0$).

В случае проверки гипотезы $r_{ij} = 0$ статистика

$$t = \frac{\sqrt{n-2}\,|\hat{r}_{ij}|}{\sqrt{1-\hat{r}_{ij}^2}}, \tag{5}$$

где $\hat{r}_{ij} = \dfrac{\widehat{\sigma}_{ij}}{\sqrt{\widehat{\sigma}_{ii}\widehat{\sigma}_{jj}}}$ — оценка парного коэффициента корреляции ($\widehat{\sigma}_{ij}$ — элементы ОМП ковариационной матрицы $\widehat{\Sigma}$), в качестве предельного распределения $G(t|H_0)$ имеет t_{n-2}-распределение Стьюдента с числом степеней свободы $n-2$ [11].

При проверке гипотезы $r_{ij} = r_0$ статистика

$$z_0 = \sqrt{n-3}\left(\frac{1}{2}\ln\left(\frac{1+\hat{r}_{ij}}{1-\hat{r}_{ij}}\right) - \frac{1}{2}\ln\left(\frac{1+r_0}{1-r_0}\right) - \left(\frac{r_0}{2(n-1)}\right)\right) \tag{6}$$

имеет в качестве предельного закона $G(z_0|H_0)$ стандартное нормальное распределение $N(0,1)$ [11].

2.3. Проверка гипотез о коэффициентах частной корреляции. В случае частных корреляций рассматриваются условные корреляции между двумя компонентами случайного вектора при фиксированных значениях некоторых других.

Представим случайный вектор $\bar{X}$ в следующем виде [11]:

$$\bar{X} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \end{bmatrix},$$

где $\bar{X}_1 = (X_1, X_2, \ldots, X_l)^T$, $\bar{X}_2 = (X_{l+1}, X_{l+2}, \ldots, X_m)^T$. Тогда вектор математических ожиданий и ковариационную матрицу представим в виде

$$\bar{M} = \begin{bmatrix} \bar{M}_1 \\ \bar{M}_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Если случайный вектор $\bar{X}$ подчиняется нормальному закону с вектором средних значений $\bar{M}$ и ковариационной матрицей Σ, то условное распределение подвектора $\bar{X}_1$ при известном $\bar{X}_2$ является нормальным с математическим ожиданием $\bar{M}_1 + B(\bar{X}_2 - \bar{M}_2)$ и ковариационной матрицей $\Sigma_{11\circ 2}$, где $B = \Sigma_{12}\Sigma_{22}^{-1}$, $\Sigma_{11\circ 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

ОМП для частного коэффициента корреляции определяется следующим соотношением:

$$\hat{r}_{ij\circ l+1,\dots,m} = \frac{\widehat{\sigma}_{ij\circ l+1,\dots,m}}{\sqrt{\widehat{\sigma}_{ii\circ l+1,\dots,m}\widehat{\sigma}_{jj\circ l+1,\dots,m}}},$$

где $\widehat{\sigma}_{ij\circ l+1,\dots,m}$ — элемент i-й строки и j-го столбца матрицы $\Sigma_{11\circ 2}$, l — число компонент в условном распределении, $2 \le l \le m$. В данном случае при оценке взаимозависимости между компонентами X_i и X_j случайной величины $\bar{X}$ исключается влияние компонент $X_{l+1}, X_{l+2}, \dots, X_m$.

При проверке гипотез H_0 вида $r_{ij\circ l+1,\dots,m} = 0$ и $r_{ij\circ l+1,\dots,m} = r_0$ используются те же самые статистики, что и для парного коэффициента корреляции. Но в данном случае в соответствующих соотношениях n заменяется на $n-m+l$.

Для проверки гипотезы $r_{ij\circ l+1,\dots,m} = 0$ вычисляется статистика

$$t = \frac{\sqrt{n-m+l-2}\,|\hat{r}_{ij\circ l+1,\dots,m}|}{\sqrt{1-\hat{r}_{ij\circ l+1,\dots,m}}}. \tag{7}$$

При этом предельным распределением статистики $G(t|H_0)$ является $t_{n-m+l-2}$-распределение Стьюдента с числом степеней свободы $n-m+l-2$ [11].

При проверке гипотезы $r_{ij\circ l+1,\dots,m} = r_0$ используется статистика

$$z_0 = \sqrt{n-3}\left(\frac{1}{2}\ln\left(\frac{1+\hat{r}_{ij\circ l+1,\dots,m}}{1-\hat{r}_{ij\circ l+1,\dots,m}}\right) - \frac{1}{2}\ln\left(\frac{1+r_0}{1-r_0}\right) - \left(\frac{r_0}{2(n-1)}\right)\right), \tag{8}$$

предельным распределением $G(z_0|H_0)$ которой является стандартное нормальное распределение $N(0,1)$ [11].

2.3. Проверка гипотезы о коэффициенте множественной корреляции. Множественный коэффициент корреляции является мерой зависимости компоненты многомерной случайной величины от некоторого множества компонент. Можно рассматривать корреляцию между одной компонентой случайного вектора и множеством всех остальных или каким-то подмножеством.

Если представить случайный вектор $\bar{X}$ в том виде, как это было показано выше, то ОМП множественного коэффициента корреляции между X_i, $i \le l$, и множеством компонент $X_{l+1}, X_{l+2}, \dots, X_m$ определяется соотношением

$$\hat{r}_{i\circ l+1,\dots,m} = \sqrt{\frac{\widehat{\sigma}_{(i)}\Sigma_{22}^{-1}\widehat{\sigma}_{(i)}^T}{\widehat{\sigma}_{ii}}},$$

где $\sigma_{(i)}$ — i-я строка матрицы Σ_{12}, σ_{ii} — элемент матрицы Σ_{11}.

Для проверки гипотезы $r_{i\circ l+1,\dots,m} = 0$ вычисляется статистика

$$F = \frac{n-m+l-1}{m-l}\,\frac{\hat{r}^2_{i\circ l+1,\dots,m}}{1-\hat{r}^2_{i\circ l+1,\dots,m}}, \tag{9}$$

предельным распределением $G(F|H_0)$ которой является $F_{m-l,n-m+l-1}$-распределение Фишера с параметрами $m-l$ и $n-m+l-1$ [11].

Еще раз подчеркнем, что все рассмотренные выше статистики имеют в качестве предельных указанные распределения лишь при наблюдении многомерного нормального закона. Как изменятся предельные распределения статистик, насколько будут справедливы выводы, формулируемые на основании решения классических задач корреляционного анализа, если наблюдаемый многомерный закон отличается от нормального, заранее сказать нельзя.

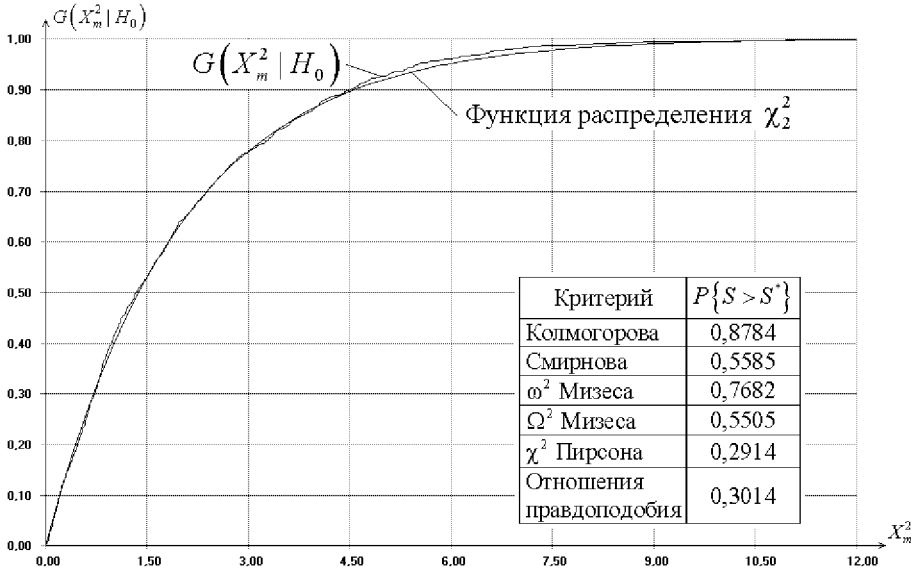
В настоящей работе продолжены исследования распределений статистик корреляционного анализа, начатые в [12–14].

3. Исследование распределений статистик критериев, используемых в корреляционном анализе

3.1. Исследование распределений статистик в случае принадлежности наблюдений нормальному закону. На первом этапе методами статистического моделирования исследовались распределения статистик корреляционного анализа при условии, что наблюдения принадлежат многомерному нормальному закону. Близость получаемых эмпирических распределений статистик, в данном случае известным предельным законам, является доводом в пользу надежности методики при анализе достоверности результатов последующих исследований.

Моделирование и исследование эмпирических распределений статистик классического корреляционного анализа показало, что они очень хорошо согласуются с соответствующими теоретическими предельными распределениями.

Например, на рис. 3 представлены эмпирическое распределение статистики X_m^2, определенной в (3), и соответствующее предельное χ_m^2-распределение при $m = 2$ и объеме выборки $n = 45$. В ходе исследований объемы выборок значений статистик, формируемых в результате моделирования, всегда задавались равными 1000. На рисунке отражены результаты проверки согласия эмпирического распределения с теоретическим предельным по критериям Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса, χ^2 Пирсона и отношения правдоподобия [3, 4]: по каждому из критериев приведен достигнутый уровень значимости $P\{S > S^*\} = 1 - G(S|H_0)$, где $G(S|H_0)$ — предельное распределение статистики S соответствующего критерия согласия при справедливости проверяемой гипотезы H_0, S^* — значение статистики критерия, вычисленное по анализируемой выборке.



Критерий	$P\{S > S^*\}$
Колмогорова	0,8784
Смирнова	0,5585
ω^2 Мизеса	0,7682
Ω^2 Мизеса	0,5505
χ^2 Пирсона	0,2914
Отношения правдоподобия	0,3014

Рис. 3. Распределение статистики X_2^2 при нормальном законе ($m = 2$) и объеме выборок $n = 45$

Исследование сходимости распределений статистик корреляционного анализа к предельным в зависимости от объема выборки n многомерного закона показало, что для тех статистик, параметры предельных распределений которых не зависят от объема выборки (статистики (3), (6) и (8)), эмпирические

распределения статистик оказываются близки к предельным уже при выборках сравнительно небольшого объема. Так, у статистики X_m^2 высокий достигаемый уровень значимости по критериям согласия наблюдается с объемов выборки $n = 30 \div 45$, а для статистики z_0 (как для парного коэффициента корреляции, так и для частного) — с $n = 100 \div 150$. Распределения статистик T_2, t (для парного и частного коэффициентов) и F, параметры предельных распределений которых зависят от объема выборки n, хорошо согласуются с предельными, начиная с объемов выборок $n = 15 \div 30$.

Существенного влияния размерности случайного вектора m на сходимость распределений соответствующих статистик к предельным при исследовании отмечено не было.

3.2. Исследование распределений статистик при отличающихся от нормального законах. Отметим еще раз особенность, связанную с моделированием псевдослучайных векторов с использованием экспоненциального семейства распределений с параметром формы λ, которую приходится учитывать при моделировании и исследовании рассматриваемых в данной работе статистик. Матрица Σ, задаваемая на этапе моделирования, отличается от ковариационной матрицы получаемого псевдослучайного вектора. Это является следствием того, что преобразуется случайный вектор, сформированный из компонент, распределенных по «ненормальному» закону (в нашем случае — по экспоненциальному семейству). Для исследования же распределений статистик корреляционного анализа нам необходимы выборки псевдослучайных векторов с известными (истинными) параметрами (математическим ожиданием и ковариационной матрицей), соответствующими проверяемой гипотезе. В качестве «истинной» ковариационной матрицы нами используется арифметическое среднее ее оценок максимального правдоподобия, получаемое по множеству выборок большого размера при неизменном векторе математических ожиданий. Мы не можем моделировать псевдослучайные векторы по «ненормальному» закону с заданной ковариационной матрицей, но можем моделировать с известной ковариационной матрицей. А этого достаточно для целей нашего исследования. Эту особенность приходится учитывать только при исследовании статистики (3), при вычислении которой используется известная ковариационная матрица. В выражения остальных статистик, исследуемых в данной работе, входит оценка ковариационной матрицы.

Исследования распределений статистик проводилось для многомерных законов, моделируемых с использованием рассмотренной процедуры при значениях параметра экспоненциального семейства $\lambda > 1$. Это ограничение обусловлено тем, что предельным случаем экспоненциального семейства при $\lambda \to 0$ является распределение Коши, которое представляет собой пример «патологического» распределения: не существует математического ожидания и дисперсия расходится. Поэтому в результате моделирования при параметре $\lambda < 1$ мы получаем закон с вырожденной ковариационной матрицей.

Распределения статистик корреляционного анализа при многомерных законах, отличающихся от нормального и моделируемых в соответствии с описанной процедурой, базирующейся на экспоненциальном семействе с параметром формы λ, определяющим вид закона, исследовались при различных объемах выборок n и различной размерности случайных величин m. На рис. 4–8 приведены примеры моделирования распределений исследуемых статистик с отражением соответствующих предельных распределений классических статистик. На рисунках представлены также значения достигнутых уровней значимости $P\{S > S^*\}$ по критериям Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса, χ^2 Пирсона и отношения правдоподобия при проверке согласия полученных в результате моделирования эмпирических распределений статистик с предельными распределениями классических статистик.

На рис. 4 показан вид распределения статистики T^2 при законе, смоделированном по параметру $\lambda = 5$. Высокие достигнутые уровни значимости по всем критериям согласия и визуальная близость распределения статистики T^2 и предельного в случае многомерного нормального закона F-распределения Фишера позволяют утверждать, что вид предельного распределения статистики значимо не изменился.

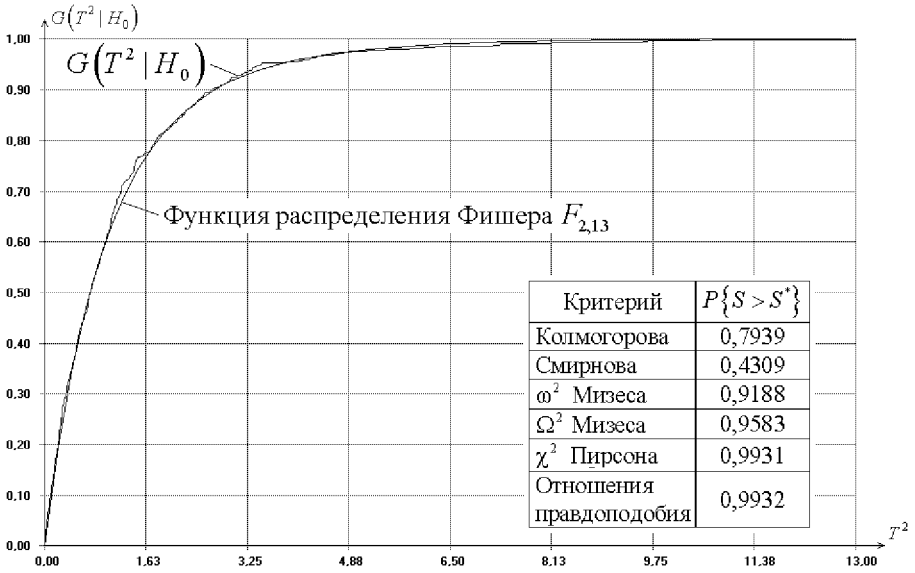


Рис. 4. Распределение статистики T^2 при многомерном законе ($m = 2$), построенном при параметре $\lambda = 5$, и объеме выборок $n = 15$

На рис. 5, 6 приведены результаты моделирования распределения статистики z_0, вычисляемой по формуле (6), при проверке гипотез о номинальном значении коэффициента парной корреляции в случае многомерных законов, моделируемых при параметрах нашей процедуры $\lambda = 5$ и $\lambda = 10$ соответственно. И здесь высокие достигнутые уровни значимости по всем критериям согласия свидетельствуют в пользу того, что вид предельного распределения статистики тот же, что и в классическом случае.

Результаты моделирования распределения статистики z_0, вычисляемой при проверке гипотез о частном коэффициенте корреляции, при законе, построенном по параметру $\lambda = 5$, отражены на рис. 7. На рис. 8 представлены результаты исследования распределения статистики F при многомерном законе, смоделированном с параметром $\lambda = 10$. В этих случаях также можно констатировать близость эмпирических распределений статистик к предельным, полученным в классическом корреляционном анализе.

Проведенные исследования распределений рассмотренных статистик корреляционного анализа показали, что в случае многомерных законов, достаточно существенно отличающихся от нормального (более островершинных или более плосковершинных, но симметричных), значимого изменения предельных распределений статистик не происходит. Это позволяет утверждать, что статистические выводы в исследованных задачах корреляционного анализа будут оставаться корректными и при нарушении предположений о нормальности наблюдаемого многомерного закона при условиях сохранения симметрии, существовании вектора математических ожиданий и невырожденности ковариационной матрицы.

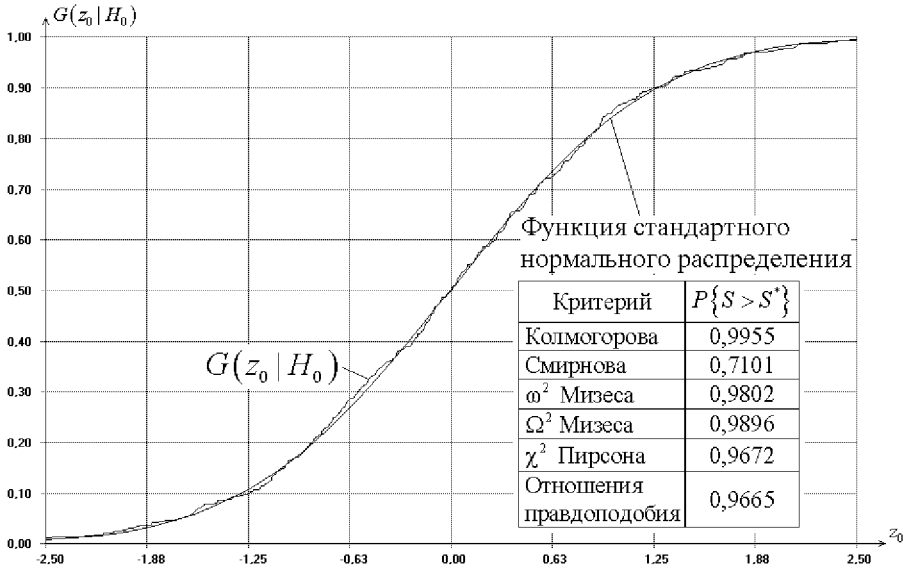


Рис. 5. Распределение статистики z_0 (для парного коэффициента корреляции) при многомерном законе ($m = 3$), построенном при параметре $\lambda = 5$, и объеме выборок $n = 100$

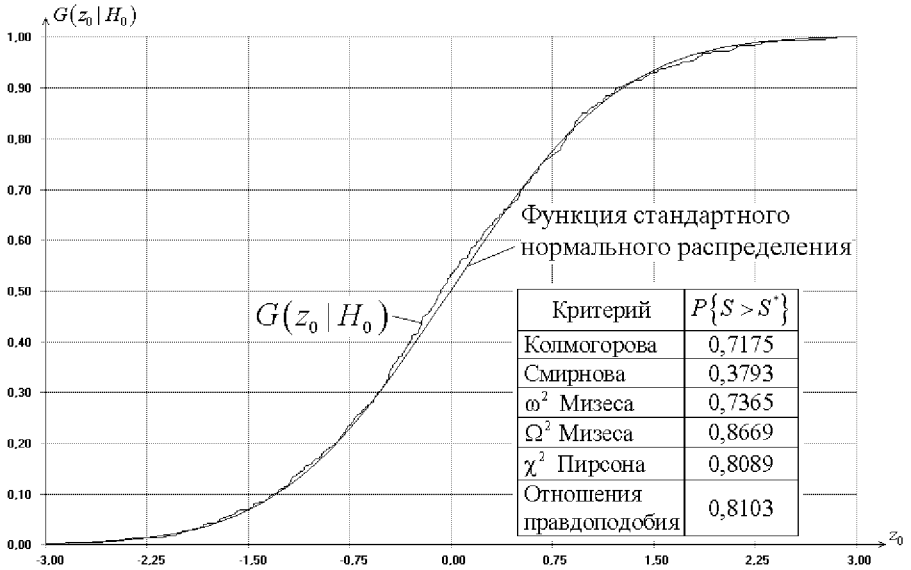
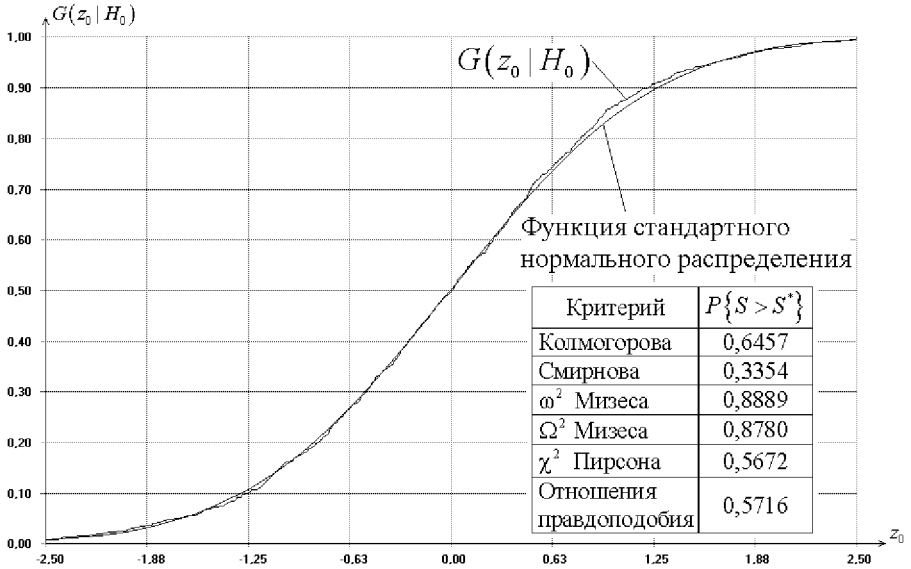
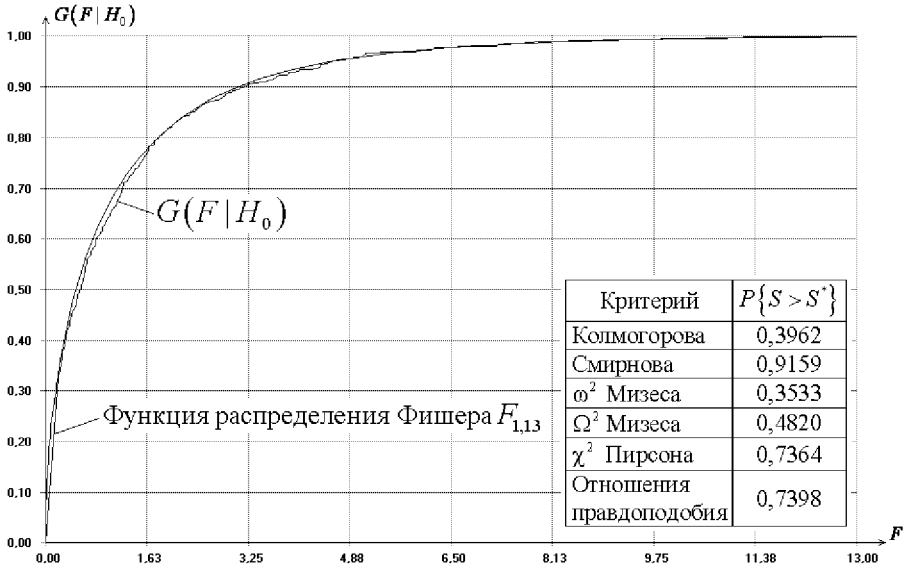


Рис. 6. Распределение статистики z_0 (для парного коэффициента корреляции) при многомерном законе ($m = 3$), построенном при параметре $\lambda = 10$, и объеме выборок $n = 100$



Критерий	$P\{S > S^*\}$
Колмогорова	0,6457
Смирнова	0,3354
ω^2 Мизеса	0,8889
Ω^2 Мизеса	0,8780
χ^2 Пирсона	0,5672
Отношения правдоподобия	0,5716

Рис. 7. Распределение статистики z_0
(для частного коэффициента корреляции) при многомерном законе ($m = 3$),
построенном при параметре $\lambda = 5$, и объеме выборок $n = 100$



Критерий	$P\{S > S^*\}$
Колмогорова	0,3962
Смирнова	0,9159
ω^2 Мизеса	0,3533
Ω^2 Мизеса	0,4820
χ^2 Пирсона	0,7364
Отношения правдоподобия	0,7398

Рис. 8. Распределение статистик и F
при многомерном законе ($m = 3$), построенном при параметре $\lambda = 10$,
и объеме выборок $n = 15$

4. Уточнение моделей распределений статистик корреляционного анализа при «ненормальных» законах

Как показано выше, распределения ряда статистик, вычисляемых в корреляционном анализе, при существенном отличии 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. Например, нет оснований для отказа от использования в качестве предельных в соответствующих случаях распределений χ^2, F или нормального.

χ^2-Распределение представляет собой частный случай гамма-распределения, F-распределение Фишера — частный случай бета-распределения второго рода. Гамма-распределения, бета-распределения второго рода и нормальные распределения всегда оказываются хорошими моделями, описывающими эмпирические распределения соответствующих статистик корреляционного анализа, получаемые в результате моделирования. Если, например, действительно χ^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 χ^2-распределению.

В данном случае мы попытались уточнить модели распределений некоторых статистик корреляционного анализа следующим образом. Моделировалась выборка интересующей нас статистики, как правило, объемом в 1000 наблюдений. Эмпирическое распределение статистики сглаживалось соответствующей моделью (гамма-, бета- или нормальным распределениями) с оцениванием ее параметров. Такой эксперимент повторялся несколько десятков раз. Параметры моделей усреднялись по всей совокупности экспериментов. Если вид модели соответствует предельному распределению статистики, то среднее арифметическое вектора параметров модели должно сходиться к истинному значению вектора параметров. Например, от модели гамма-распределения будем приходить к соответствующему ей частному случаю — χ^2-распределению, от бета-распределения — к соответствующему F-распределению и т. п.

Предельным распределением классической статистики, используемой при проверке гипотезы о равенстве вектора математического ожидания некоторому номинальному, является χ^2_m-распределение, где m — размерность многомерного вектора. Это соответствует гамма-распределению с плотностью $f(x) = \dfrac{1}{\sigma^\theta \Gamma(\theta)} x^{\theta-1} e^{-x/\sigma}$ с параметром формы $\theta = m/2$ и параметром масштаба $\sigma = 2$. В табл. 1 представлены усредненные по 50 смоделированным выборкам статистики значения параметров модели гамма-распределения, аппроксимирующие распределение статистики в случае многомерных законов величин, моделируемых при различных значениях параметра датчика λ ($\lambda = 2$ соответствует нормальному закону). Размерность моделируемых многомерных величин $m = 3$. Очевидно, что значения параметров в случае наблюдения нормального закона сходятся к значениям 2 и 1,5 соответственно. Это соответствует χ^2_3-распределению. На рис. 9 отражены соответствующие функции распределения статистики. Видно, что функции распределения статистики при законах, моделируемых с параметрами $\lambda = 1, 2, 5, 10$, практически совпадают.

Таблица 1

Параметры гамма-распределения	Параметр датчика			
	$\lambda = 1$	$\lambda = 2$	$\lambda = 5$	$\lambda = 10$
σ	2,0157	2,0094	1,9892	1,9644
θ	1,4737	1,4938	1,5137	1,5302

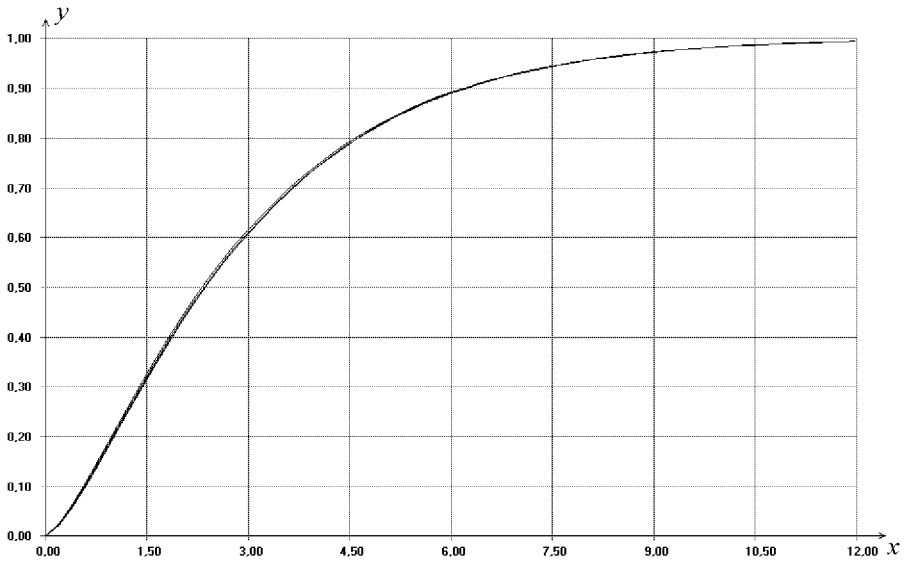


Рис. 9. Функции гамма-распределения с параметрами из табл. 1

При проверке аналогичной гипотезы при неизвестной ковариационной матрице предельным распределением статистики является $F_{m,n-m}$-распределение. Данному случаю при размерности вектора $m = 3$ и объеме выборки $n = 30$ соответствует бета-распределение второго 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 $\theta_2 = (n-m)/m$, параметрами формы $\theta_0 = m/2$ и $\theta_1 = (n-m)/2$. Представленные в табл. 2 усредненные по 50 смоделированным выборкам значения параметров бета-распределения (при $m = 3$ и $n = 30$) показывают аналогичную картину сходимости. Очевидно, что значения параметров бета-распределения в случае наблюдения нормального закона сходятся к значениям $\theta_0 = 1,5$, $\theta_1 = 13,5$, $\theta_2 = 9$, что соответствует F-распределению Фишера с числом степеней свободы 3 и 27. Различие (или совпадение) четырех бета-распределений, соответствующих табл. 2, такого же порядка, как у распределений на рис. 9.

Таблица 2

Параметры бета-распределения	Параметр датчика			
	$\lambda = 1$	$\lambda = 2$	$\lambda = 5$	$\lambda = 10$
θ_2	8,8628	8,9765	9,0619	9,1576
θ_0	1,5636	1,5062	1,4861	1,4627
θ_1	13,7685	13,5002	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. Табл. 3 и 4 иллюстрируют сходимость к стандартному нормальному закону распределений исследуемых статистик при многомерных законах, существенно отличающихся от нормального: табл. 3 — в случае проверки гипотез о парном коэффициенте корреляции, а табл. 4 — в случае проверки гипотез о частном коэффициенте корреляции. Как и в предыдущих случаях, усреднение осуществлялось по 50 выборкам статистик. Очевидно, что и в данном случае существенного различия между моделями распределений статистики нет.

Таблица 3

Параметры нормального закона	Параметр датчика			
	$\lambda = 1$	$\lambda = 2$	$\lambda = 5$	$\lambda = 10$
μ	0,0296	0,0885	0,0183	0,0150
σ	1,0183	1,0016	0,9851	0,9927

Таблица 4

Параметры нормального закона	Параметр датчика			
	$\lambda = 1$	$\lambda = 2$	$\lambda = 5$	$\lambda = 10$
μ	0,0070	0,0445	0,0154	0,0051
σ	1,0263	1,0021	0,9923	0,9821

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. При $m = 3$, $n = 30$, $l = 2$ это соответствует бета-распределению с параметрами $\theta_0 = 0,5$, $\theta_1 = 14$, $\theta_2 = 28$. Как сходятся к данной ситуации распределения этой же статистики в случае наблюдаемых законов, отличающихся от многомерного нормального, иллюстрирует табл. 5. Во всех случаях полученные модели распределений статистик очень близки.

Таблица 5

Параметры бета-распределения	Параметр датчика			
	$\lambda = 1$	$\lambda = 2$	$\lambda = 5$	$\lambda = 10$
θ_2	27,9500	27,9972	28,0089	27,9893
θ_0	0,4910	0,5007	0,4969	0,5044
θ_1	13,6056	14,0692	13,8450	14,0621

Заключение

Исследования эмпирических распределений статистик корреляционного анализа при наблюдении многомерного нормального закона показали, что они хорошо согласуются с теоретическими предельными распределениями, полученными в классическом корреляционном анализе, и подтвердили эффективность методики исследований.

Исследования распределений рассмотренных статистик корреляционного анализа в случае многомерных законов, отличающихся от нормального в достаточно широких пределах (более островершинных или более плосковершинных, но симметричных), показали, что значимого изменения предельных распределений статистик не происходит. Эмпирические распределения данных статистик по-прежнему хорошо описываются предельными законами, полученными в классическом корреляционном анализе в предположении о нормальности наблюдаемого вектора. Это существенно расширяет сферу корректного применения методов классического корреляционного анализа в приложениях.

Выводы не касаются задач проверки гипотез о ковариационных матрицах многомерного закона (гипотез H_0 вида $\Sigma = \Sigma_0$ при известном и неизвестном векторах математических ожиданий). Есть основания полагать, что предельные распределения статистик, используемых при проверке таких гипотез, существенно зависят от наблюдаемого многомерного закона. По крайней мере моделирование распределений аналогичных статистик в одномерном случае (при проверке гипотез вида $\sigma = \sigma_0$ при известном и неизвестном математических ожиданиях) показало, что предельные распределения этих статистик очень сильно зависят от наблюдаемого закона. В то время как на распределениях статистик, вычисляемых при проверке гипотез вида $\mu = \mu_0$ при известной и неизвестной дисперсиях, отклонения от нормальности наблюдаемого одномерного закона по сравнению с предыдущим случаем сказываются незначительно.

ЛИТЕРАТУРА

1. *Лемешко Б. Ю., Постовалов С. Н.* О распределениях статистик непараметрических критериев согласия при оценивании по выборкам параметров наблюдаемых законов // Заводская лаборатория. 1998, Т. 64, № 3. С. 61–72.
2. *Денисов В. И., Лемешко Б. Ю., Постовалов С. Н.* Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Методические рекомендации. Ч. I. Критерии типа χ^2. Новосибирск: Изд-во НГТУ, 1998.
3. *Лемешко Б. Ю., Постовалов С. Н.* Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Методические рекомендации. Ч. II. Непараметрические критерии. Новосибирск: Изд-во НГТУ, 1999.
4. *Лемешко Б. Ю., Постовалов С. Н.* О зависимости распределений статистик непараметрических критериев и их мощности от метода оценивания параметров // Заводская лаборатория. Диагностика материалов. 2001. Т. 67, № 7. С. 62–71.
5. *Лемешко Б. Ю., Постовалов С. Н.* Применение непараметрических критериев согласия при проверке сложных гипотез // Автометрия. 2001. Т. 2. С. 88–102.
6. *Лемешко Б. Ю., Постовалов С. Н.* Непараметрические критерии при проверке сложных гипотез о согласии с распределениями Джонсона // Докл. СО АН ВШ. 2002. № 1(5). С. 65–74.
7. *Лемешко Б. Ю., Гильдебрант С. Я., Постовалов С. Н.* К оцениванию параметров надежности по цензурированным выборкам // Заводская лаборатория. Диагностика материалов. 2001. Т. 67, № 1. С. 52–64.
8. *Лемешко Б. Ю., Чимитова Е. В.* Построение оптимальных L-оценок параметров сдвига и масштаба распределений по выборочным квантилям // Сиб. журн. индустр. математики. 2001. Т. 4, № 2(8). С. 166–183.

9. *Ермаков С. М., Михайлов Г. А.* Статистическое моделирование. М.: Наука, 1982.

10. *Лемешко Б. Ю., Помадин С. С.* Один подход к моделированию псевдослучайных векторов с «заданными» числовыми характеристиками по законам, отличным от нормального // Информатика и проблемы телекоммуникаций: Материалы конф. / Междунар. науч.-техн. конф. Новосибирск, 2002. С. 121–122.

11. *Андерсон Т.* Введение в многомерный статистический анализ. М.: Физматгиз, 1963.

12. *Лемешко Б. Ю., Помадин С. С., Французов А. В.* Статистическое моделирование распределений статистик, используемых в корреляционном анализе // Информатика и проблемы телекоммуникаций: Материалы конф. / Рос. науч.-техн. конф. Новосибирск, 2000. С. 101–102.

13. *Лемешко Б. Ю., Помадин С. С.* Исследование распределений статистик корреляционного анализа при отклонении многомерного закона от нормального // Актуальные проблемы электронного приборостроения: Тр. / 5 Междунар. конф. Т. 7. С. 184–187. Новосибирск, 2000.

14. *Лемешко Б. Ю., Помадин С. С., Кузьменко С. В.* Программное обеспечение компьютерного исследования статистических закономерностей в задачах корреляционного анализа // Информатика и проблемы телекоммуникаций: Материалы конф. / Междунар. науч.-техн. конф. Новосибирск, 2001. С. 79.

г. Новосибирск
Новосибирский гос. технический университет
E-mail: ser@fpm.ami.nstu.ru

Статья поступила 1 июля 2002 г.