Новосибирский Государственный Технический Университет

На правах рукописи

ПОМАДИН
Сергей Сергеевич

ИССЛЕДОВАНИЕ РАСПРЕДЕЛЕНИЙ СТАТИСТИК МНОГОМЕРНОГО АНАЛИЗА ДАННЫХ ПРИ НАРУШЕНИИ ПРЕДПОЛОЖЕНИЙ О НОРМАЛЬНОСТИ

Специальность 05.13.17 — теоретические основы информатики

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель
доктор технических наук
профессор Лемешко Б.Ю.

Новосибирск – 2004

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Современное состояние и актуальность темы исследований. Существует множество работ по многомерному статистическому анализу [13, 31, 36, 44, 45, 47, 94, 95, 108, 114], содержание которых указывает на актуальность и эффективность применения соответствующего математического аппарата в различных областях знаний, таких как экономика, биология и медицина. При этом в практике статистического анализа возникает существенно больше постановок задач, чем предлагается решений в классической математической статистике [101]. Разнообразие статистических гипотез, выдвигаемых в процессе статистического анализа в различных приложениях, оказывается существенно шире предлагаемого классическим аппаратом. Классический аппарат включает в себя ограниченный перечень задач проверки статистических гипотез, для которых найдены предельные распределения статистик, используемых в соответствующих критериях. Поэтому классические результаты оказываются применимыми при выполнении достаточно строгих предположений, которые на практике часто не имеют места.

С другой стороны, для обнаружения закономерных связей можно использовать аппарат анализа данных [53, 54, 63, 64], когда рассматриваемые объекты представляются как «черные ящики». В данном случае на анализируемые данные не накладываются какие–либо строгие ограничения. Но применение такого подхода обычно привязано к определенному классу задач, например, распознавание образов, и поэтому далеко не всегда удается использовать методы анализа данных в растущем множестве различных статистических задач.

Таким образом, можно говорить о наличии в математической статистике множества «пробелов», которые чаще всего связаны с проверкой разного рода статистических гипотез. В этом случае вопрос обычно упирается в необходимость нахождения предельного распределения статистики построенного критерия или распределения статистики при заданном объеме выборки. Как правило, нахождение предельного закона для статистики критерия проверки конкретной гипотезы аналитическими методами оказывается чрезвычайно

сложной задачей, а задач, требующих разрешения, — слишком много [124].

В большинстве случаев отсутствие необходимых теоретических результатов объясняется сложностью и трудоемкостью получения решений аналитическими методами. Можно констатировать, что количество и уровень сложности задач, выдвигаемых практикой, возрастают настолько быстро, что ресурсы человеческого интеллекта, его производительность просто не в состоянии обеспечить решение такого множества задач без создания и использования соответствующих вычислительных технологий.

Сегодня в связи с бурным развитием и внедрением персональных компьютеров, особую актуальность приобретает задача обеспечения высокого качества пакетов прикладных статистических программ. Несмотря на то, что, рынок насыщен различными пакетами программных систем статистического анализа [22, 115], реализуемые в них методы и алгоритмы сильно отстают от последних достижений в области статистических исследований. С одной стороны это объясняется, прежде всего, тем, что подробное описание последних результатов исследований очень сложно отыскать в литературных источниках, поэтому они остаются труднодоступными для разработчиков программного обеспечения. К сожалению, с другой стороны необходимо отметить и то, что в некоторых работах встречаются ошибки применения статистических методов [98], что также не облегчает быстрое внедрение новых методов в программные пакеты.

Перспективы программного обеспечения по статистическому анализу данных обсуждались в работах [27–30, 38], современные проблемы внедрения прикладной статистики поднимались в [100]. Расширяющиеся использование ЭВМ и их совершенствование в свою очередь отражается на развитии статистических методов и использовании статистических методов в приложениях [14, 32, 35, 42, 48, 56, 65, 104, 109, 116, 120].

Вышесказанное подчеркивает необходимость (а практика уже показывает возможность [61,67,81,82,86,89,90]) развития компьютерных методов исследования статистических закономерностей, компьютерных методов исследования свойств оценок и статистик различных критериев проверки статистических ги-

потез, построения вероятностных моделей для исследуемых закономерностей. Это позволяет с меньшими интеллектуальными затратами получать фундаментальные знания в области математической статистики, и, следовательно, осуществлять корректные статистические выводы при анализе данных в различных прикладных областях.

В последние годы при исследовании некоторых задач математической и прикладной статистики получено множество результатов, связанных с исследованием распределений статистик критериев согласия в случае проверки простых и сложных гипотез [84, 86–88], с исследованием статистических свойств различных оценок [69, 91], полученных как раз благодаря применению методов компьютерного моделирования. Накопленный опыт в данной области показал, что с использованием методов статистического моделирования и последующего анализа можно получать результаты по точности не уступающие аналитическим. Например, при оценивании параметров распределений некоторых законов в случаях проверки сложных гипотез с использованием методов статистического моделирования, когда наиболее часто применяют метод Монте–Карло [37, 49, 51, 52, 113], были получены таблицы процентных точек для предельных распределений статистик непараметрических критериев [5, 17, 23, 24, 117–119, 121]. В этой связи появилась обоснованная уверенность, что с использованием данного подхода можно закрывать многие существующие в прикладной статистике «пробелы», применяя относительно простой вычислительный и математический аппарат.

В различных приложениях статистического анализа многомерных случайных величин одну из ключевых позиций занимают задачи корреляционного анализа [122]. В процессе решения задач корреляционного анализа выявляется наличие и характер взаимосвязи величин, взаимозависимости величин при устранении влияния совокупности других или зависимости одной случайной величины от группы величин. Вычисляются оценки коэффициентов и матриц парной, частной и множественной корреляции, проверяются различные статистические гипотезы относительно параметров многомерного распределения и коэффициентов корреляции. На основании результатов корреляцион-

ного анализа может делаться вывод о наличии и характере функциональной зависимости или предпочтительности для описания исследуемого объекта регрессионной модели того или иного вида.

В основе существующего аппарата корреляционного анализа лежит *предположение о принадлежности наблюдаемого случайного вектора многомерному нормальному закону*. Базируясь на этом, получены предельные распределения статистик, используемых в критериях многомерного анализа [2, 16, 33, 57–59].

На практике, исследователь далеко не всегда имеет дело с нормальным законом [16, 94, 99]. Как правило, многие исследователи вообще не придают значения проверке этого важного предположения корреляционного анализа, либо они вынуждены «в силу обстоятельств» работать только с многомерными величинами, имеющим нормальное распределение, как это сделано в работах [31, 114]. Например, в нашей жизни достаточно мало экономических процессов, отклонения которых распределены по нормальному закону. Поэтому данное ограничение приводит к сужению области применения корреляционного анализа в экономике. Естественно, возникает вопрос о справедливости выводов, получаемых на основании результатов корреляционного анализа при нарушении основного предположения. В доступной литературе ответ на данный вопрос найден не был, хотя можно найти указания на робастность некоторых критериев, применяемых в многомерном анализе.

Целью данной диссертационной работы явилось стремление разобраться, что будет происходить с распределениями различных статистик корреляционного анализа, если наблюдаемый закон будет отличаться от многомерного нормального.

Немаловажен и такой аспект. Большинство наиболее весомых результатов в математической статистике имеет асимптотический характер. На практике же всегда имеют дело с ограниченными объемами наблюдений. И свойства используемых статистик в таких ситуациях порой существенно отличаются от асимптотических. Не являются исключением и предельные распределения статистик корреляционного анализа, которые получены для выборок многомерных величин с объемом $n \to \infty$ [2, 33, 57, 58]. На практике исследователю

важно знать конечные объемы выборок, начиная с которых можно пользоваться найденными предельными законами. Поэтому в процессе проводимых исследований можно оценить объемы выборок, которые могут быть рекомендованы как достаточные для принятия правильного решения по соответствующему критерию корреляционного анализа.

Очевидно, что ответить на поставленные вопросы, используя аналитические методы, чрезвычайно сложно из-за нетривиальности возникающих задач. Поэтому в основу проводимого исследования положена развиваемая на кафедре прикладной математики НГТУ методика компьютерного моделирования и анализа статистических закономерностей.

Цели и задачи исследований. Основной целью диссертационной работы является исследование поведения (предельных) законов распределений статистик многомерного анализа в случае принадлежности наблюдаемых случайных величин многомерным законам распределения, отличным от нормального.

Для достижения поставленной цели было предусмотрено решение следующих задач:

- исследование эмпирических распределений статистик корреляционного анализа в случае многомерного нормального закона для подтверждения теоретических результатов и выявления скорости сходимости распределений к соответствующим предельным;
- моделирование многомерных законов, отличных от нормального, с заданными вектором математических ожиданий, ковариационной матрицей и задаваемой мерой отклонения от нормального;
- исследование распределений статистик, используемых при проверке гипотез о векторе математических ожиданий и ковариационной матрице, в случае многомерных законов, отличающихся от нормального;
- исследование распределений статистик, используемых при проверке гипотез о парном, частном и множественном коэффициентах корреляции, в случае многомерных законов, отличающихся от нормального;
- исследование влияния способов группирования и количества интервалов на оценку корреляционного отношения, исследование критериев,

используемых при проверке гипотез о корреляционном отношении;

– исследование критериев проверки гипотез о математическом ожидании и дисперсии в одномерном случае при наблюдениях, не подчиняющихся нормальному закону.

Методы исследования. Для решения поставленных задач использовался аппарат теории вероятностей, математической статистики, вычислительной математики, математического программирования, статистического моделирования.

Научная новизна диссертационной работы заключается в:

– результатах исследования распределений статистик многомерного анализа данных при нарушении предположений о нормальном законе многомерных случайных величин;

– результатах исследования распределений статистик критериев, используемых при проверке гипотез о математическом ожидании и дисперсии, в случае принадлежности наблюдений семейству симметричных распределений;

– методе моделирования многомерных случайных величин по законам, заданным образом отличающихся от нормального.

Основные положения, выносимые на защиту.

1. Результаты исследования сходимости распределений статистик многомерного анализа к предельным распределениям в зависимости от объема выборки при наблюдаемом нормальном законе случайных векторов.
2. Подход и алгоритм моделирования многомерного закона распределения, отличающегося от нормального, с заданными вектором математических ожиданий и ковариационной матрицей.
3. Результаты исследований распределений статистик многомерного анализа для ситуаций, когда наблюдаемый многомерный закон отличается от нормального.
4. Результаты исследований распределений статистик критериев, используемых для проверки гипотез о математическом ожидании и дисперсии.

Практическая ценность и реализация результатов. Результаты исследования распределений статистик классического корреляционного анализа позволяют существенно расширить сферу корректного применения ряда критериев на многомерные законы, в достаточно широких пределах отличающиеся от нормального (более островершинных или более плосковершинных). Для законов такого вида показано, что распределения статистик, используемых в критериях проверки гипотез о векторе математических ожиданий и о нулевых значениях парного, частного и множественного коэффициентов корреляции, по–прежнему хорошо описываются классическими предельными распределениями. В случае других исследуемых критериев выявлена явная зависимость от наблюдаемого многомерного закона. Предложен метод моделирования многомерных случайных векторов с задаваемым параметром отклонения от многомерного нормального закона.

Апробация работы. Основные результаты исследований докладывались на Новосибирской межвузовской НТК «Интеллектуальный потенциал Сибири» (Новосибирск, 2000); Российской НТК «Информатика и проблемы телекоммуникаций» (Новосибирск, 2000, 2001, 2002, 2003, 2004); V международной конференции «Актуальные проблемы электронного приборостроения АПЭП-2000» (Новосибирск, 2000); Региональной НТК студентов, аспирантов, молодых ученых «Наука. Техника. Инновации» (Новосибирск, 2001); Всероссийской НТК «Информационные системы и технологии ИСТ-2001» (Нижний Новгород, 2001); VI международной конференции «Актуальные проблемы электронного приборостроения АПЭП-2002» (Новосибирск, 2002); Региональной конференции «Вероятностные идеи в науке и философии» (Новосибирск, 2003); всероссийской НТК «Информационные системы и технологии ИСТ-2004» (Нижний Новгород, 2004). Исследования по теме диссертации были поддержаны грантом Минобразования РФ (проект № А03-2.8-280), вошли составной частью в работы, поддержанные Российским фондом фундаментальных исследований (проект № 00-01-00913) и грантом Минобразования РФ (проект № Т02-3.3-3356).

Публикации. По теме диссертации опубликовано 16 печатных работ. Среди которых 8 публикаций отражают основные результаты исследований.

Структура работы. Диссертация состоит из введения, 6 глав основного содержания, включая 11 таблиц и 48 рисунков, заключения, списка использованных источников и приложения.

Краткое содержание работы. В первой главе представлен обзор проблем, связанных с встречающимися на практике многомерными наблюдениями, не подчиняющимися нормальному закону, и, как следствие, неприменимости ряда критериев многомерного анализа данных. Даются основные определения и теоремы, на которых базируется классический аппарат корреляционного анализа.

Во второй главе исследуются распределения классических статистик, используемых в критериях проверки гипотез о математических ожиданиях и дисперсиях, если наблюдаемый закон в той или иной мере отличается от нормального.

В третьей главе исследуются распределения статистик критериев, используемых при проверке гипотез о векторе математических ожиданий и ковариационной матрице, в случае многомерных законов, отличных от нормального.

В четвертой главе приводятся результаты исследования распределений статистик, применяемых в критериях проверки гипотез о парном, частном и множественном коэффициентах корреляции.

В пятой главе рассматриваются проблемы, связанные с вычислением оценки корреляционного отношения и влиянием различных способов группирования на получаемую оценку, исследуются критерии проверки гипотез о корреляционном отношении.

Во шестой главе дано краткое описание исследовательской программной системы и предлагается метод моделирования многомерных случайных величин с заданным «отклонением» от многомерного нормального закона. Показывается различие между моделируемым и многомерным нормальным законами.

ГЛАВА 1
ПОСТАНОВКА ЗАДАЧ ИССЛЕДОВАНИЯ

1.1. Основные понятия и определения

Введем для дальнейшего использования следующие обозначения:

$\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_n$ — 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 $\bar{X}$;

r_{ij} — парный коэффициент корреляции между компонентами X_i и X_j случайного вектора $\bar{X}$;

$r_{ij\cdot l+1,\ldots,m}$ — частный коэффициент корреляции между компонентами X_i и X_j случайного вектора $\bar{X}$ при исключении влияния компонент $X_{l+1}, \ldots, X_m$;

$r_{i\cdot l+1,\ldots,m}$ — множественный коэффициент корреляции между X_i и множеством компонент $X_{l+1}, \ldots, X_m$ случайного вектора $\bar{X}$;

ρ_{ij}^2 — корреляционное отношение компоненты X_i по X_j случайного вектора $\bar{X}$;

$\hat{\bar{M}}$ и $\hat{\Sigma}$ — несмещенные оценки максимального правдоподобия (ОМП) математического ожидания и ковариационной матрицы, которые вычисляются по следующим формулам

$$\hat{\bar{M}} = \frac{1}{n}\sum_{i=1}^{n} \bar{X}_i, \quad \hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n} \left(\bar{X}_i - \hat{\bar{M}}\right)\left(\bar{X}_i - \hat{\bar{M}}\right)^T;$$

$\hat{r}_{ij}$, $\hat{r}_{ij\cdot l+1,\ldots,m}$, $\hat{r}_{i\cdot l+1,\ldots,m}$ и $\hat{\rho}_{ij}^2$ — ОМП соответствующих величин, вычисляемых по формулам (4.1), (4.5), (4.8) и (5.2).

В диссертации рассматриваются различные выборочные оценки по моделируемым псевдослучайным величинам. Основным методом нахождения оценок является метод максимального правдоподобия для негруппированных данных. И только для вычисления оценки корреляционного отношения требуется группирование данных по одной из компонент случайного вектора.

Введем определение и рассмотрим используемые далее способы группирования для одномерных случайных величин [62].

Определение 1. Выборка называется негруппированной, если выборочные значения представляют собой индивидуальные значения наблюдений из области определения случайной величины:

$$x_1, \ldots, x_n,$$

где n — объем выборки.

Определение 2. Выборка называется группированной, если область определения случайной величины разбита на k непересекающихся интервалов граничными точками:

$$-\infty < x_{(1)} < \ldots < x_{(k-1)} < +\infty,$$

и зафиксированы количества наблюдений n_l, попавших в l-й интервал значений. Объем выборки $n = \sum_{l=1}^{k} n_l$.

Существует несколько способов разбиения области определения случайной величины на интервалы. Наиболее часто используют интервалы равной длины или равной частоты. Самым простым способом является равноинтервальное группирование (РИГ). Равночастотное группирование (РЧГ) подразумевает разбиение области определения так, чтобы частота попадания n_l в каждый интервал была одинаковой. В работе также применяется асимптотически оптимальное группирование (АОГ), где разбиение осуществляется по граничным точкам из таблиц асимптотически оптимального группирования для стандартной нормальной величины при оценивании параметра сдвига и масштаба. Более подробную информацию об асимптотически оптимальном группировании можно найти в [43], где приведены еще и таблицы АОГ для других одномерных законов.

В процессе исследований часто возникает задача проверки того, насколько хорошо эмпирическое распределение той или иной статистики согласуется с некоторым теоретическим распределением. При ее решении используются различные критерии согласия.

Определение 3. Гипотеза вида $H_0 : F(x) = F(x, \theta)$, где $F(x, \theta)$ — функция распределения вероятностей, с которой проверяется согласие наблюдаемой

выборки независимых одинаково распределенных величин $X_1, X_2, \ldots, X_n$ называется простой, если θ — известное значение параметра (скалярного или векторного).

Определение 4. Гипотеза вида $H_0 : F(x) \in \{F(x,\theta), \theta \in \Omega\}$ называется сложной, если в качестве значения неизвестного параметра θ используется его оценка $\hat{\theta}$, вычисленная по той же выборке, по которой проверяется гипотеза о согласии. Если оценка $\hat{\theta}$ вычислена по другой выборке, то гипотеза простая.

Проверка гипотезы о согласии эмпирического распределения с теоретическим осуществляется по следующей схеме [111, 112]. Для выбранного критерия вычисляется значение S^* статистики критерия S как некоторой функции от выборки и закона распределения, с которым проверятся согласие. Для используемых на практике критериев обычно известны предельные распределения $G(S|H_0)$ соответствующих статистик при условии истинности основной гипотезы H_0. Гипотеза о согласии не отвергается, если

$$P\{S > S^*\} = \int_{S^*}^{+\infty} g(S)dS > \alpha,$$

где α — заданный уровень значимости, $g(S)$ — плотность распределения $G(S|H_0)$. Вероятность $P\{S > S^*\}$ позволяет судить о степени согласия, так как по существу, представляет собой вероятность истинности основной гипотезы. В дальнейшем будем называть вероятность $P\{S > S^*\}$ — достигнутым уровнем значимости.

Задачи проверки статистических гипотез опираются на выборки независимых случайных величин. Случайность самой выборки предопределяет, что возможны и ошибки в результатах статистических выводов. С результатами проверки гипотез связывают ошибки двух видов: ошибка 1-го рода состоит в том, что отклоняется гипотеза H_0, когда она верна; ошибка 2-го рода состоит в том, что принимается гипотеза H_0, в то время как справедлива альтернативная гипотеза H_1. Величина α задает вероятность ошибки 1-го рода. Если гипотеза H_1 определена, то задание α определяет и вероятность ошибки 2-го рода β для используемого критерия проверки гипотез. Мощность критерия представ-

ляет собой величину $1-\beta$. Понятно, что чем выше мощность используемого критерия при заданном значении α, тем лучше критерий различает гипотезы H_0 и H_1. Особенно важно, чтобы используемый критерий хорошо различал близкие альтернативы.

Некорректное использование критериев согласия может приводить к необоснованному принятию или необоснованному отклонению проверяемой гипотезы. С рекомендациями по использованию критериев согласия можно ознакомиться в [43, 85, 111, 112].

1.2. Задачи корреляционного анализа

1.2.1. Критерии проверки гипотез о векторе математических ожиданий и ковариационной матрице

Важными статистическими задачами корреляционного анализа являются задачи проверки гипотез о том, что вектор математических ожиданий нормального распределения является данным вектором. Эти задачи могут быть рассмотрены в предположении, что ковариационная матрица Σ известна из ранее проводимых экспериментов, или неизвестна, тогда она должна быть оценена.

Критерии для проверки гипотез о векторе математических ожиданий, основываются на следующих двух теоремах [2–4, 18, 19, 25, 33, 59].

Теорема 1. Если проверяемая гипотеза для выборки объема n, взятой из совокупности с нормальным законом $N(\bar{M}, \Sigma)$, имеет вид $H_0 : \bar{M} = \bar{M}_0$ и ковариационная матрица Σ известна, тогда гипотеза H_0 не отклоняется с уровнем значимости α при выполнении неравенства

$$n(\hat{\bar{M}} - \bar{M}_0)^T \Sigma^{-1} (\hat{\bar{M}} - \bar{M}_0) \leq \chi^2_m(\alpha), \tag{1.1}$$

где распределение $F(x)$ левой части неравенства есть χ^2—распределение с m степенями свободы, и $\chi^2_m(\alpha)$ удовлетворяет равенству

$$P\{x \le \chi^2_m(\alpha)\} = \int\limits_0^{\chi^2_m(\alpha)} dF(x) = 1-\alpha. \tag{1.2}$$

Теорема 2. Когда ковариационная матрица Σ неизвестна и проверяется гипотеза $H_0 : \bar{M} = \bar{M}_0$ по выборке m—мерного случайного вектора объема n, полученной из совокупности с нормальным законом $N(\bar{M}, \Sigma)$, то гипотеза H_0 не отвергается для уровня значимости α, если

$$\frac{n(n-m)}{m(n-1)}(\hat{\bar{M}} - \bar{M}_0)^T \hat{\Sigma}^{-1}(\hat{\bar{M}} - \bar{M}_0) \le F_{m,n-m}(\alpha), \tag{1.3}$$

где распределение $F(x)$ левой части неравенства есть F—распределение Фишера с m и $n-m$ степенями свободы, и $F_{m,n-m}(\alpha)$ удовлетворяет равенству

$$P\{x \le F_{m,n-m}(\alpha)\} = \int\limits_0^{F_{m,n-m}(\alpha)} dF(x) = 1-\alpha. \tag{1.4}$$

Задачи проверки гипотез о ковариационной матрице имеют вид $H_0 : \Sigma = \Sigma_0$, где Σ_0 — номинальное значение ковариационной матрицы. Подразумевается, что вектор математических ожиданий будет оцениваться по исследуемой выборке. В случае, когда проверяется совместная гипотеза о векторе математических ожиданий и о ковариационной матрице, тогда гипотеза имеет вид $H_0 : \bar{M} = \bar{M}_0,\ \Sigma = \Sigma_0$. В корреляционном анализе для задач о ковариационных матрицах используют критерии, определяемые следующими теоремами [2, 11, 33].

Теорема 3. Если проверяемая гипотеза имеет вид $H_0 : \Sigma = \Sigma_0$ для m—мерных случайных векторов $\bar{X}_1, \ldots, \bar{X}_n$, подчиняющихся нормальному закону $N(\bar{M}, \Sigma)$, тогда отношение правдоподобия имеет вид

$$\lambda_1 = \left(\frac{e}{n}\right)^{\frac{mn}{2}} \left|B\Sigma_0^{-1}\right|^{\frac{n}{2}} e^{-\frac{1}{2}\operatorname{tr}\left(B\Sigma_0^{-1}\right)}, \tag{1.5}$$

где

$$B = \sum_{i=1}^{n} \left(\bar{X}_i - \hat{\bar{M}}\right)\left(\bar{X}_i - \hat{\bar{M}}\right)^T. \tag{1.6}$$

В этом случае распределение $F(x)$ статистики $-2\ln\lambda_1$ представляет собой χ^2—распределение с $m(m+1)/2$ степенями свободы. Гипотеза H_0 принимается с уровнем значимости α, когда выполняется условие

$$-2\ln\lambda_1 \le \chi^2_{m(m+1)/2}(\alpha), \tag{1.7}$$

где $\chi^2_{m(m+1)/2}(\alpha)$ удовлетворяет равенству

$$P\{x \le \chi^2_{m(m+1)/2}(\alpha)\} = \int\limits_0^{\chi^2_{m(m+1)/2}(\alpha)} dF(x) = 1-\alpha. \tag{1.8}$$

Теорема 4. Для проверки гипотезы $H_0 : \bar{M} = \bar{M}_0,\ \Sigma = \Sigma_0$ по выборке m—мерных случайных векторов $\bar{X}_1, \ldots, \bar{X}_n$, принадлежащих нормальному закону $N(\bar{M}, \Sigma)$, отношение правдоподобия имеет вид

$$\lambda_2 = \left(\frac{e}{n}\right)^{\frac{mn}{2}} \left|B\Sigma_0^{-1}\right|^{\frac{n}{2}} e^{-\frac{1}{2}\left[\mathrm{tr}\left(B\Sigma_0^{-1}\right)+n\left(\hat{\bar{M}}-\bar{M}_0\right)^T\Sigma_0^{-1}\left(\hat{\bar{M}}-\bar{M}_0\right)\right]}. \tag{1.9}$$

В этом случае распределение $F(x)$ статистики $-2\ln\lambda_2$ представляет собой χ^2—распределение с $m(m+1)/2+m$ степенями свободы. Гипотеза H_0 не отвергается при уровне значимости α, если

$$-2\ln\lambda_2 \le \chi^2_{m(m+1)/2+m}(\alpha), \tag{1.10}$$

где $\chi^2_{m(m+1)/2+m}(\alpha)$ определяется равенством

$$P\{x \le \chi^2_{m(m+1)/2+m}(\alpha)\} = \int\limits_0^{\chi^2_{m(m+1)/2+m}(\alpha)} dF(x) = 1-\alpha. \tag{1.11}$$

1.2.2. Критерии проверки гипотез о коэффициентах корреляции

В случае необходимости исследования взаимозависимости случайных величин применяют различные критерии корреляционного анализа, предназначенные для выявления характера статистической зависимости. В данной работе затрагиваются задачи корреляционного анализа, связанные с парной, частной и множественной корреляцией случайных величин.

Если требуется исследовать взаимозависимость двух величин, применяют критерии о парной корреляции, которые базируются на следующих теоремах [2, 10, 12, 33, 57, 58].

Теорема 5. Пусть $\bar{X}_1, \ldots, \bar{X}_n$ — независимые одинаково распределенные случайные величины с нормальным законом распределения $N(\bar{M}, \Sigma)$. Если проверяемая гипотеза имеет вид $H_0 : r_{ij} = 0$, тогда гипотеза H_0 не отвергается с уровнем значимости α при условии, что выполняется неравенство

$$-t_{n-2}(\alpha/2) \leq \frac{\sqrt{n-2}\,\hat{r}_{ij}}{\sqrt{1-\hat{r}_{ij}^2}} \leq t_{n-2}(\alpha/2), \tag{1.12}$$

где $\hat{r}_{ij}$ — ОМП парного коэффициента корреляции между компонентами вектора X_i и X_j, распределение $F(x)$ статистики $\frac{\sqrt{n-2}\,\hat{r}_{ij}}{\sqrt{1-\hat{r}_{ij}^2}}$ есть t—распределение Стьюдента с числом степеней свободы $n-2$, и $t_{n-2}(\alpha/2)$ удовлетворяет равенству

$$P\{-t_{n-2}(\alpha/2) \leq x \leq t_{n-2}(\alpha/2)\} = \int\limits_{-t_{n-2}(\alpha/2)}^{t_{n-2}(\alpha/2)} dF(x) = 1-\alpha. \tag{1.13}$$

Теорема 6. Если проверяется гипотеза $H_0 : r_{ij} = r_0$ по выборке $\bar{X}_1, \ldots, \bar{X}_n$ случайных векторов, распределенных по нормальному закону $N(\bar{M}, \Sigma)$, то гипотеза H_0 принимается с уровнем значимости α, если выполняется соотношение

$$-z(\alpha/2) \leq z_0 \leq z(\alpha/2), \tag{1.14}$$

где $z_0 = \sqrt{n-3}\left(\frac{1}{2}\ln\left(\frac{1+\hat{r}_{ij}}{1-\hat{r}_{ij}}\right) - \frac{1}{2}\ln\left(\frac{1+r_0}{1-r_0}\right) - \left(\frac{r_0}{2(n-1)}\right)\right)$, $\hat{r}_{ij}$ — ОМП парного коэффициента корреляции между компонентами вектора X_i и X_j, распределение $F(x)$ статистики z_0 есть стандартное нормальное распределение, и $z(\alpha/2)$ удовлетворяет

$$P\{-z(\alpha/2) \leq x \leq z(\alpha/2)\} = \int\limits_{-z(\alpha/2)}^{z(\alpha/2)} dF(x) = 1-\alpha. \tag{1.15}$$

Если нас интересует взаимозависимость двух величин при устранении воздействия остальных величин, то исследуется, так называемая, частная корреляция. Критерии проверки гипотез о частном коэффициенте корреляции вида $H_0 : r_{ij\cdot l+1,\ldots,m} = 0$ и $H_0 : r_{ij\cdot l+1,\ldots,m} = r_0$ базируются на тех же самых теоремах [2, 8, 33, 58], что и для парного коэффициента корреляции. Только в этом случае в соответствующих соотношениях n заменяется на $n-(m-l)$, где $l-$ число компонент случайного вектора в условном распределении X_i и X_j при фиксировании остальных.

Когда исследуется зависимость единственной величины от группы других, рассматривается множественная корреляция, и используют критерии проверки гипотез о множественной корреляции. В работе рассматривается критерий проверки гипотезы о значимости множественного коэффициента корреляции, базирующийся на следующей теореме [2, 9, 33, 58].

Теорема 7. Если проверяется гипотеза вида $H_0 : r_{i\cdot l+1,\ldots,m} = 0$ по выборке $m-$мерного случайного вектора объема n, полученной из совокупности с нормальным законом, тогда гипотеза H_0 принимается с уровнем значимости α, если справедливо следующее неравенство

$$\frac{n-m+l-1}{m-l}\frac{\hat{r}^2_{i\cdot l+1,\ldots,m}}{1-\hat{r}^2_{i\cdot l+1,\ldots,m}} \le F_{m-l,n-m+l-1}(\alpha), \tag{1.16}$$

где $\hat{r}_{i\cdot l+1,\ldots,m}-$ОМП множественного коэффициента корреляции. Распределение $F(x)$ левой части неравенства представляет собой $F-$распределение Фишера с m и $n-m+l-1$ степенями свободы, $F_{m-l,n-m+l-1}(\alpha)$ удовлетворяет равенству

$$P\{x \le F_{m-l,n-m+l-1}(\alpha)\} = \int_0^{F_{m-l,n-m+l-1}(\alpha)} dF(x) = 1-\alpha. \tag{1.17}$$

1.2.3. Критерии проверки гипотез о корреляционном отношении

В корреляционном анализе на основании соотношений между парным коэффициентом корреляции и корреляционным отношением можно судить о характере зависимости между компонентами случайного вектора.

Если требуется проверить гипотезы вида: о равенстве корреляционного отношения нулю $H_0 : \rho_{ij}^2 = 0$ или о равенстве корреляционного отношения квадрату коэффициента корреляции $H_0 : \rho_{ij}^2 = r_{ij}^2$ (критерий линейности регрессии X_i по X_j), применяют критерии о корреляционном отношении, которые базируются на следующих теоремах [58].

Теорема 8. Если проверяется гипотеза вида $H_0 : \rho_{ij}^2 = 0$ по выборке m—мерного случайного вектора объема n, полученной из совокупности с нормальным законом, тогда гипотеза H_0 принимается с уровнем значимости α, если справедливо следующее неравенство

$$\frac{n-k}{k-1}\frac{\hat{\rho}_{ij}^2}{1-\hat{\rho}_{ij}^2} \leq F_{k-1,n-k}(\alpha), \tag{1.18}$$

где $\hat{\rho}_{ij}^2$ — оценка корреляционного отношения. Распределение $F(x)$ левой части неравенства представляет собой F—распределение Фишера с $k-1$ и $n-k$ степенями свободы, $F_{k-1,n-k}(\alpha)$ удовлетворяет равенству

$$P\{x \leq F_{k-1,n-k}(\alpha)\} = \int_0^{F_{k-1,n-k}(\alpha)} dF(x) = 1-\alpha. \tag{1.19}$$

Теорема 9. В случае когда проверяется гипотеза вида $H_0 : \rho_{ij}^2 = r_{ij}^2$ по выборке $\bar{X}_1, \ldots, \bar{X}_n$ случайных векторов, распределенных по нормальному закону $N(\bar{M}, \Sigma)$, то гипотеза H_0 принимается с уровнем значимости α, если справедливо следующее неравенство

$$\frac{n-k}{k-2}\frac{\hat{\rho}_{ij}^2 - \hat{r}_{ij}^2}{1-\hat{\rho}_{ij}^2} \leq F_{k-2,n-k}(\alpha), \tag{1.20}$$

где $\hat{\rho}_{ij}^2$ и $\hat{r}_{ij}$ — соответственно оценка корреляционного отношения и ОМП парного коэффициента корреляции. Распределение $F(x)$ левой части неравенства представляет собой F—распределение Фишера с $k-2$ и $n-k$ степенями свободы, $F_{k-2,n-k}(\alpha)$ удовлетворяет равенству

$$P\{x \leq F_{k-2,n-k}(\alpha)\} = \int_0^{F_{k-2,n-k}(\alpha)} dF(x) = 1-\alpha. \tag{1.21}$$

Из приведенных теорем видно, что рассмотренные критерии и распределения статистик получены в предположении о нормальном законе наблюдаемого случайного вектора.

1.3. Цели исследования распределений статистик корреляционного анализа при нарушении предположения о нормальности

Как уже отмечалось, в основе аппарата классического корреляционного анализа лежит предположение о принадлежности наблюдаемого случайного вектора многомерному нормальному закону. Базируясь на этом, сформулированы теоремы о распределениях статистик, используемых в критериях классического корреляционного анализа.

На практике предпосылки классического корреляционного анализа выполняются далеко не всегда. Очевидно и то, что многомерный нормальный закон далеко не всегда является наилучшей моделью для описания реально наблюдаемых многомерных случайных величин [99]. Например, в работе [94] Леонов В. П. отмечает, что за последние 10 лет ему довелось провести достаточно детальный статистический анализ более 150 массивов данных из различных областей экспериментальной биологии и медицины, содержавших от 10 до 300 признаков и от 100 до нескольких десятков тысяч наблюдений. Получилось, что в 50-80% случаев количественные показатели биологических объектов не подчинялись нормальному распределению.

Поэтому центральное место нормального закона не стоит объяснять его универсальной применимостью. Нормальный закон — это один из многих типов распределения, правда, имеющий основание с относительно бóльшим удельным весом для применения на практике. Его основная привлекательность — это полнота теоретических исследований. В своих работах [93, 94] Леонов В. П. призывает специалистов в биомедицине уделять больше внимания проверке выборок наблюдений на нормальность. Так, проанализированные им работы указывают на то, что некоторые авторы «забывают» об осуществлении соответствующих проверок, и впоследствии интерпретируют

результаты некорректного применения классических критериев.

Что делать в случае, когда исследователь сталкивается с многомерным законом, который не является нормальным? Как использовать критерии корреляционного анализа? Или какой вид анализа применять в таком случае? Например, в работе [31] Айвазян С. А. предлагает два подхода для исследований наблюдений, которые не подчиняются многомерному нормальному закону. Первый подход заключается в использовании классических алгоритмов для получения первого начального приближения, а второй — в подборе такого преобразования, которое осуществило бы переход к многомерному нормальному закону. Оба способа очень тяжело реализуются в общем случае, да и исследователь должен быть весьма подготовлен в области статистического анализа, чтобы корректно видоизменять и интерпретировать наблюдаемые величины.

Поэтому с практической точки зрения интересен вопрос о степени корректности выводов, формируемых на основании применения конкретных процедур классического корреляционного анализа, в случае нарушения основного предположения. Насколько корректны будут выводы статистического анализа, если истинная модель многомерного закона в той или иной мере отличается от нормального, и как такое отличие влияет на распределения исследуемых статистик?

Настоятельная потребность в исследовании некоторых критериев корреляционного анализа на устойчивость или, наоборот, неустойчивость к отклонению многомерного закона от нормального проявилась давно. Например, А. Гейен [58] рассмотрел устойчивость коэффициента корреляции к отклонениям от двумерного нормального закона. Им было показано что, когда коэффициент корреляции равен нулю и, в частности, когда случайные величины независимы, критерий проверки гипотезы о нулевом значении коэффициента корреляции устойчив. Но при больших значениях этого коэффициента отклонения от нормальной теории становятся заметными.

В данной работе при помощи методов компьютерного моделирования и анализа закономерностей мы попытались определить границы применимости классического корреляционного анализа, ответить на вопрос, какие критерии

можно уверенно применять при отклонении многомерной выборки от нормального закона, а применение каких критериев требует строгого выполнения всех налагаемых условий.

Для подтверждения работоспособности методов компьютерного моделирования и исследования статистических закономерностей в случае многомерных величин в работе исследованы эмпирические распределения статистик классического корреляционного анализа в случае многомерного нормального закона. Эти исследования должны были подтвердить классические результаты и показать близость получаемых эмпирических распределений статистик, в данном случае, известным предельным законам. Соответствие в такой ситуации эмпирических распределений, получаемых в процессе моделирования, предельным классическим распределениям статистик должно послужить доводом, подчеркивающим достоверность результатов в общем случае.

1.4. Проблемы моделирования многомерных псевдослучайных величин

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.

Алгоритмы моделирования случайных векторов в случае нормального закона, а также для некоторых других частных случаев известны давно [51,52,106]. Эти алгоритмы позволяют достаточно быстро получать выборки случайных векторов произвольных объемов и при различных задаваемых параметрах: векторе математических ожиданий и ковариационной матрице.

Однако моделирование случайных векторов с произвольным распределением до сих пор остается нерешенной проблемой, так как реализация известных

общих подходов для решения этой задачи обычно приводит либо к непреодолимым практическим трудностям [51], либо огромным вычислительным затратам для получения больших объемов выборок, например, при использовании метода исключений.

Поэтому возникает потребность в разработке процедуры моделирования многомерных величин, распределенных по законам, отличным от нормального, с заданными математическим ожиданием и ковариационной матрицей, а для задач исследования критериев корреляционного анализа еще и с некоторой заданной мерой близости к многомерному нормальному закону.

В работе [60] Кирьяновым Б. Ф. предложен метод моделирования случайных векторов с произвольным, но одинаковым для всех координат одномерным законом распределения и с заданной ковариационной матрицей. Такой подход базируется на реализации системы линейных разностных уравнений со случайными коэффициентами. Однако, как отмечает сам автор, реализация указанных разностных уравнений приводит к корреляции между последовательно генерируемыми векторами, что во многих случаях недопустимо.

В данной работе предлагается процедура моделирования многомерных величин, распределенных по законам, отличным от нормального, с заданными математическим ожиданием и ковариационной матрицей [72]. Она базируется на подходе, используемом для нормальных случайных векторов [49, 51], и выборе «удобного» одномерного закона распределения для всех координат моделируемого вектора. В качестве одномерного закона используется семейство симметричных распределений (6.4).

К сожалению, реализованная процедура не позволяет моделировать многомерный закон с некоторой произвольной функцией распределения, 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. К тому же, на практике, при наблюдении выборок многомерных случайных векторов вставал бы вопрос об определе-

нии закона, которому они принадлежат. А покоординатный анализ сводится к одномерному случаю, который достаточно хорошо исследован и изучен.

Таким образом, на настоящем этапе исследований предложено направление решения задачи по моделированию закона с заданными математическим ожиданием и ковариационной матрицей с введением параметра в качестве меры различия между моделируемым и многомерным нормальным законами распределений.

1.5. Выводы

В данной главе диссертации рассмотрены некоторые критерии классического корреляционного анализа, связанные с проверкой гипотез о математическом ожидании, ковариационной матрице, парном, частном и множественном коэффициентах корреляции, из которых очевидна актуальность решения следующих задач:

- исследование эмпирических распределений статистик корреляционного анализа в случае многомерного нормального закона для выявления скорости их сходимости к соответствующим предельным распределениям;
- моделирование «удобным» способом многомерного закона, отличного от нормального;
- исследование распределений различных статистик классического корреляционного анализа в случае законов распределений, отличных от многомерного нормального.

ГЛАВА 2
ИССЛЕДОВАНИЕ КРИТЕРИЕВ ПРОВЕРКИ ГИПОТЕЗ О МАТЕМАТИЧЕСКИХ ОЖИДАНИЯХ И ДИСПЕРСИЯХ ПРИ ВЕРОЯТНОСТНЫХ ЗАКОНАХ, ОТЛИЧАЮЩИХСЯ ОТ НОРМАЛЬНОГО

При поверке измерительных приборов, в задачах контроля качества и в других приложениях часто возникает необходимость в проверке статистических гипотез о значении математического ожидания $H_0 : \mu = \mu_0$ или о значении дисперсии $H_0 : \sigma^2 = \sigma_0^2$. В основе применяемого классического аппарата проверки гипотез такого вида лежит предположение о принадлежности наблюдаемых данных (ошибок измерений) нормальному закону распределения. В то же время, не секрет, что ошибки измерений приборов и систем во многих случаях не удается удовлетворительно описать моделью нормального закона [97]. Необходимость проверки гипотез о математических ожиданиях и дисперсиях при нарушении предположений о нормальности наблюдаемого закона встречается во многих приложениях. Насколько корректно в этом случае применение классического аппарата проверки данных гипотез? Когда можно без боязни использовать классические критерии, а когда их применение является некорректным, и как следует поступать в данном случае?

В работе [58] обобщены теоретические исследования Бартлетта, Гири и Гейена, в которых рассматривались вопросы об устойчивости критериев проверки гипотез о математических ожиданиях по отношению к виду наблюдаемого закона и содержатся указания на существенную зависимость от вида закона критериев проверки гипотез о дисперсиях. Сведения, которые практик может почерпнуть из этого, сводятся к тому, что при нарушении нормальности нельзя использовать классические результаты для проверки гипотез о дисперсиях, а для проверки гипотез о математических ожиданиях, по-видимому, можно, но с долей осторожности.

Целью данной главы явилось стремление установить при помощи численных исследований, что происходит с распределениями классических стати-

стик, используемых в критериях проверки гипотез о математических ожиданиях и дисперсиях, если наблюдаемый закон в той или иной мере отличается от нормального; проверить, насколько будут корректны статистические выводы, базирующиеся на классических результатах, если нарушено предположение о нормальности; дать в руки исследователя необходимый математический аппарат, обеспечивающий корректность выводов при законах распределения, существенно отличающихся от нормального [76, 80, 107].

2.1. Классические критерии проверки гипотез о математических ожиданиях и дисперсиях

Пусть мы имеем выборку n случайных величин, распределенных по нормальному закону $\xi_1, \ldots, \xi_n \in N(\mu_{\text{ист}}, \sigma^2_{\text{ист}})$. В этом случае задачи проверки гипотез о математических ожиданиях и дисперсиях формулируются следующим образом.

1. В критерии проверки гипотез вида $H_0 : \mu = \mu_0$ при известной дисперсии $\sigma^2_{\text{ист}}$ используется статистика

$$T_1 = \frac{1}{n}\sum_{i=1}^{n}\xi_i, \tag{2.1}$$

которая при справедливости гипотезы H_0 подчиняется нормальному распределению: $G(T_1|H_0) = N(\mu_0, \sigma^2_{\text{ист}}/n)$ [123]. Проверяемая гипотеза H_0 отклоняется при больших отклонениях T_1 от μ_0.

2. Для проверки гипотезы $H_0 : \mu = \mu_0$ при неизвестной дисперсии $\sigma^2_{\text{ист}}$ используется статистика

$$T_2 = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}}\sqrt{n}, \tag{2.2}$$

где $\hat{\mu} = \frac{1}{n}\sum\limits_{i=1}^{n}\xi_i$, $\hat{\sigma}^2 = \frac{1}{n-1}\sum\limits_{i=1}^{n}(\xi_i - \hat{\mu})^2$. При справедливости H_0 статистика T_2 распределена как $G(T_2|H_0) = t_{n-1}$ — распределение Стьюдента [123].

3. Для проверки гипотезы вида $H_0 : \sigma^2 = \sigma_0^2$ при известном математическом

ожидании $\mu_{\text{ист}}$ вычисляется статистика

$$T_3 = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(\xi_i - \mu_{\text{ист}})^2, \tag{2.3}$$

условным распределением которой является $G(T_3|H_0) = \chi_n^2 -$ распределение [123].

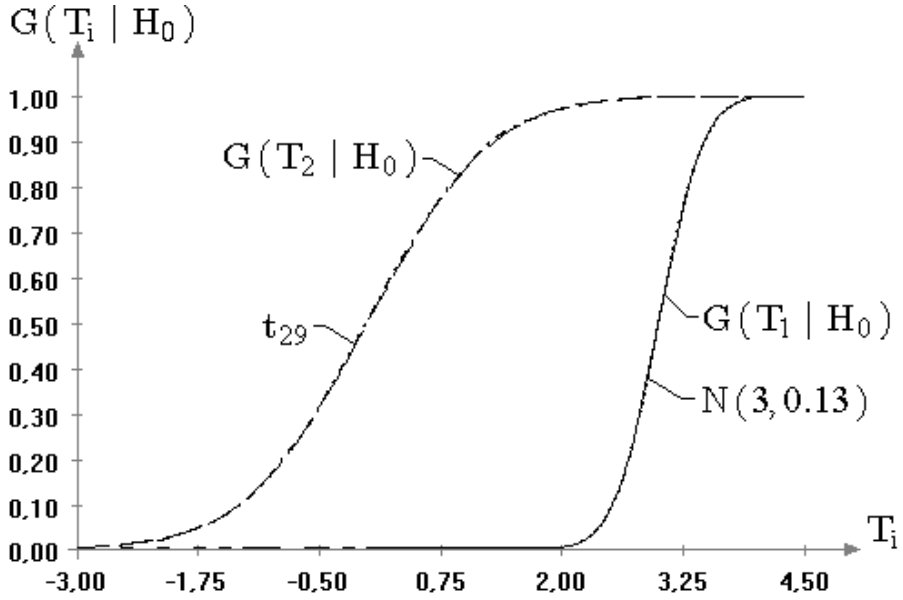
4. В критерии проверки гипотезы вида $H_0 : \sigma^2 = \sigma_0^2$ при неизвестном математическом ожидании $\mu_{\text{ист}}$ используется статистика

$$T_4 = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(\xi_i - \hat{\mu})^2, \tag{2.4}$$

подчиняющаяся $G(T_4|H_0) = \chi_{n-1}^2 -$ распределению [123].

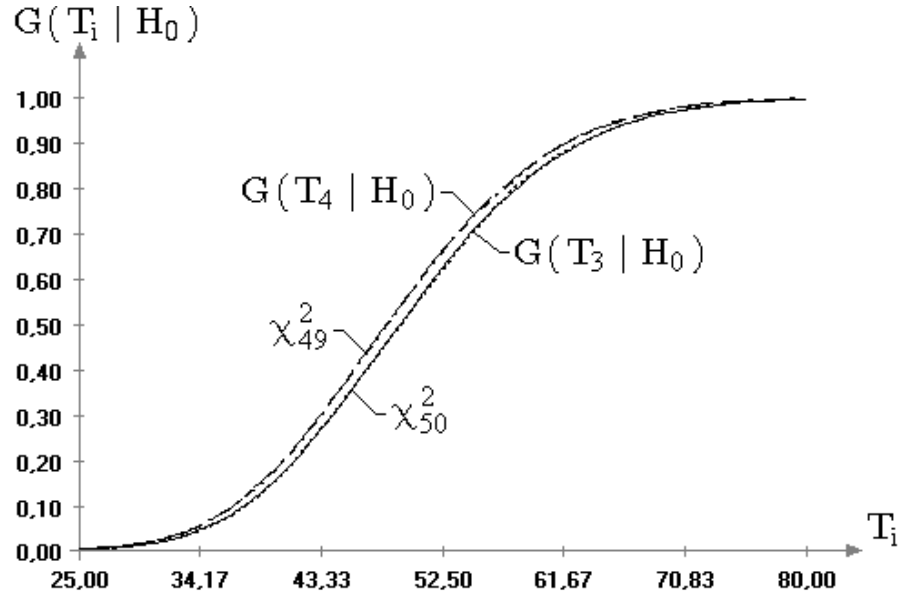
Для иллюстрации работоспособности применяемой методики исследований приведем результаты моделирования эмпирических распределений данных статистик в случае нормального закона регистрируемых наблюдений. В дальнейшем N указывает на объемы смоделированных выборок статистик рассматриваемых критериев.

В качестве примера рассмотрены распределения статистик T_1, T_2, T_3, T_4 при проверяемых гипотезах $H_0 : \mu = 3$ и $\sigma_{\text{ист}}^2 = 4$. На рис. 2.1 отражены полученные в результате моделирования эмпирические распределения статистик T_1, T_2 и теоретические распределения данных статистик при нормальности наблюдаемого закона. Видно, что смоделированные распределения статистик, используемых при проверке гипотез о значении математического ожидания, визуально совпадают со своими предельными законами: нормальным и $t_{n-1} -$ распределением Стьюдента. Количественной мерой близости полученных эмпирических распределений статистик и теоретических предельных служат достигнутые уровни значимости $P\{S > S^*\}$ по критериям согласия χ^2 Пирсона, Колмогорова, ω^2 Крамера–Мизеса–Смирнова, Ω^2 Андерсона–Дарлинга [111, 112], где $S -$ статистика соответствующего критерия согласия, $S^* -$ ее значение, вычисленное по конкретной выборке исследуемых статистик. Чем больше достигнутый уровень значимости, чем ближе он к 1, тем лучше согласуется эмпирическое распределение статистики с теоретическим.



	T_1	T_2
$P_{\chi^2} =$	0.42	0.67
$P_{\text{Колм}} =$	0.73	0.46
$P_{\omega^2} =$	0.87	0.41
$P_{\Omega^2} =$	0.88	0.47

Рис. 2.1. Эмпирические и теоретические функции распределения статистик T_1, T_2 при проверке гипотезы $H_0 : \mu = 3$ при известной ($\sigma^2_{\text{ист}} = 4$) и неизвестной дисперсии: $n = 30$; $N = 10000$



	T_3	T_4
$P_{\chi^2} =$	0.99	0.88
$P_{\text{Колм}} =$	0.75	0.91
$P_{\omega^2} =$	0.66	0.95
$P_{\Omega^2} =$	0.80	0.97

Рис. 2.2. Эмпирические и теоретические функции распределения статистик T_3 и T_4 при проверке гипотезы $H_0 : \sigma^2 = 4$ при известном ($\mu_{\text{ист}} = 3$) и неизвестном математическом ожидании: $n = 50$; $N = 10000$

Приведенные в таблице на рис. 2.1 значения достигнутых уровней значимости $P\{S > S^*\}$ для статистик T_1 и T_2 говорят об очень высокой близости полученных в результате моделирования эмпирических распределений статистик к предельным. Аналогичная картина наблюдается на рис. 2.2, где приведены результаты моделирования распределений статистик T_3, T_4, используемых в критериях проверки гипотез о значениях дисперсии.

2.2. Распределения статистик T_1, T_2, T_3, T_4 при нарушении предположений о нормальности

В работе [68] распределения статистик T_3, T_4 были исследованы в случае принадлежности наблюдаемых случайных величин распределениям экстремальных значений, логистическому и Лапласа. В данном случае рассмотрено распределение, более перспективное для описания ошибок измерений. Очень хорошей моделью для закона распределения ошибок конкретной измерительной системы иногда оказывается распределение из семейства с плотностью (6.4) и параметром формы λ, так как данное семейство охватывает широкий класс симметричных законов.

Далее будем рассматривать распределения статистик T_1, T_2, T_3, T_4 в случае принадлежности наблюдаемых случайных величин указанному семейству распределений $\xi_i \in f(x; \theta_0, \theta_1, \lambda)$, $i = \overline{1,n}$. Предельные распределения статистик T_1, T_2, T_3, T_4 известны только для частного случая этого семейства при $\lambda = 2$ (нормального закона).

Для статистик, вычисляемых по выборкам случайных величин $\xi_i \in f(x; \theta_0, \theta_1, \lambda)$, $i = \overline{1,n}$, распределенных по семейству (6.4) с параметром формы λ, введем обозначения $T_i(\lambda) = T_i$.

Результаты моделирования выборок статистик $T_1(\lambda)$ и $T_2(\lambda)$, где параметр λ изменялся в диапазоне от 1 до 10, показали, что значимого изменения предельных распределений статистик $T_1(\lambda)$ и $T_2(\lambda)$, используемых в критериях проверки гипотез о значениях математического ожидания (при известной и неизвестной дисперсии), не происходит.

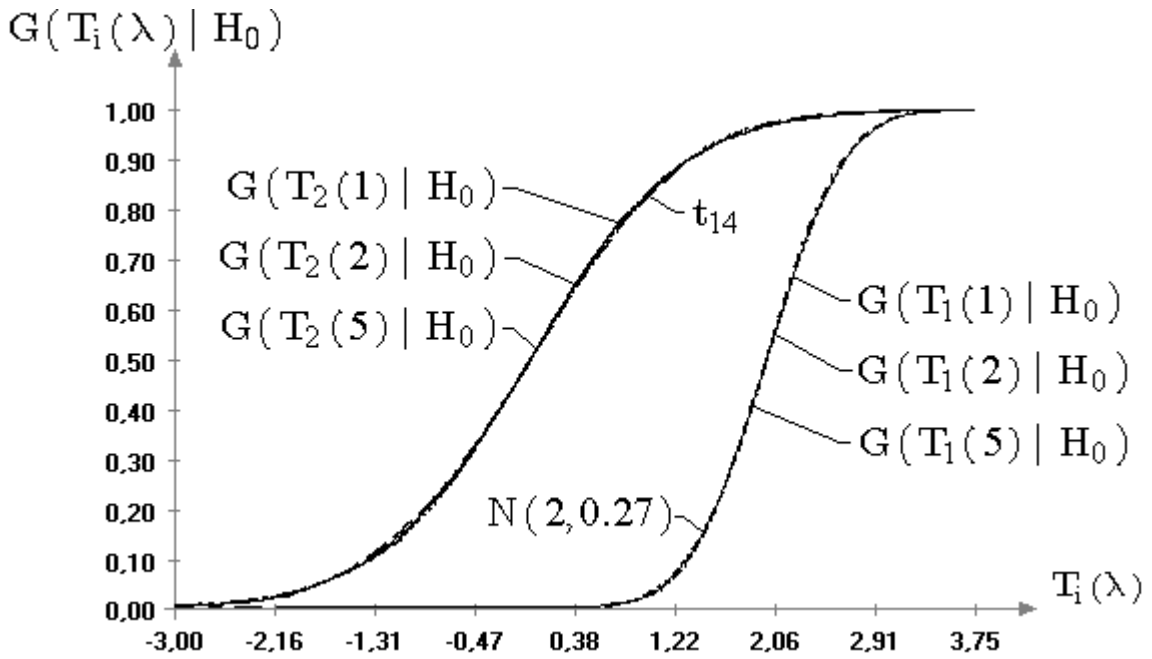


Рис. 2.3. Эмпирические и теоретические функции распределения статистик T_1 и T_2 при проверке гипотезы $H_0 : \mu = 2$ при известной ($\sigma^2_{\text{ист}} = 4$) и неизвестной дисперсии: $n = 15$; $N = 5000$

На рис. 2.3 в качестве примера представлены графики теоретических предельных, соответствующих классическому случаю, и полученных эмпирических функций распределения статистик $T_1(\lambda)$ и $T_2(\lambda)$ для объемов выборок $N = 5000$, используемых при проверке гипотезы $H_0 : \mu = 2$ при известной ($\sigma^2_{\text{ист}} = 4$) и неизвестной дисперсиях. Визуальная близость распределений статистик, построенных в случае принадлежности выборок семейству (6.4), к предельным (классическим) распределениям, полученным для нормального закона, позволяет отметить, что значимого изменения распределений статистик не произошло. Это же подтверждает применение критериев согласия для проверки значимости отклонений смоделированных эмпирических распределений статистик $T_1(\lambda)$ и $T_2(\lambda)$ от классических предельных распределений (при нормальном законе наблюдаемых величин). Достигнутые уровни значимости $P\{S > S^*\}$ представлены в таблице 2.1.

Результаты исследований распределений статистик T_1 и T_2 позволяют утверждать, что в случае отклонений наблюдаемого закона от нормального

Таблица 2.1

Значения достигнутых уровней значимости критериев согласия для примера на рис. 2.3

$T_1(1)$	$T_1(2)$	$T_1(5)$	$T_2(1)$	$T_2(2)$	$T_2(5)$
$P_{\chi^2} = 0.82$	$P_{\chi^2} = 0.64$	$P_{\chi^2} = 0.14$	$P_{\chi^2} = 0.52$	$P_{\chi^2} = 0.81$	$P_{\chi^2} = 0.17$
$P_{\text{Колм}} = 0.13$	$P_{\text{Колм}} = 0.97$	$P_{\text{Колм}} = 0.88$	$P_{\text{Колм}} = 0.76$	$P_{\text{Колм}} = 0.84$	$P_{\text{Колм}} = 0.52$
$P_{\omega^2} = 0.17$	$P_{\omega^2} = 0.93$	$P_{\omega^2} = 0.92$	$P_{\omega^2} = 0.46$	$P_{\omega^2} = 0.85$	$P_{\omega^2} = 0.54$
$P_{\Omega^2} = 0.16$	$P_{\Omega^2} = 0.88$	$P_{\Omega^2} = 0.81$	$P_{\Omega^2} = 0.36$	$P_{\Omega^2} = 0.82$	$P_{\Omega^2} = 0.56$

(при сохранении симметричности), использование классических предельных распределений для статистик T_1 и T_2 не нарушает корректности выводов статистического анализа при проверке гипотез вида $H_0 : \mu = \mu_0$.

В случае несимметричных законов наблюдаемых величин, например, при распределениях экстремальных значений, распределения статистик T_1 и T_2 претерпевают значимые изменения, которые можно заметить как визуально, так и с использованием критериев согласия. Соответствующий пример демонстрирует картина, представленная на рис. 2.4. Пример свидетельствует все-таки об ограниченной области устойчивости критериев проверки гипотез о математическом ожидании. В таблице на рисунке приведены достигнутые значения уровня значимости, которые свидетельствуют, что, не смотря на визуальную близость эмпирического распределения статистики к теоретическому, в данном случае гипотеза о нормальности статистики T_1 при уровне значимости $\alpha = 0.05$ должна быть отклонена.

В отличие от T_1 и T_2 распределения статистик T_3 и T_4, используемых в критериях проверки гипотез о дисперсии, как в случае известного математического ожидания, так и в случае неизвестного очень чувствительны к виду наблюдаемого закона распределения. Иллюстрацией к сказанному являются рисунки 2.5 и 2.6, на которых изображены графики эмпирических функций распределений статистик $T_3(\lambda)$ и $T_4(\lambda)$, смоделированных при семействе распределений (6.4) с параметром формы λ равным 1 и 10. На рисунках приведе-

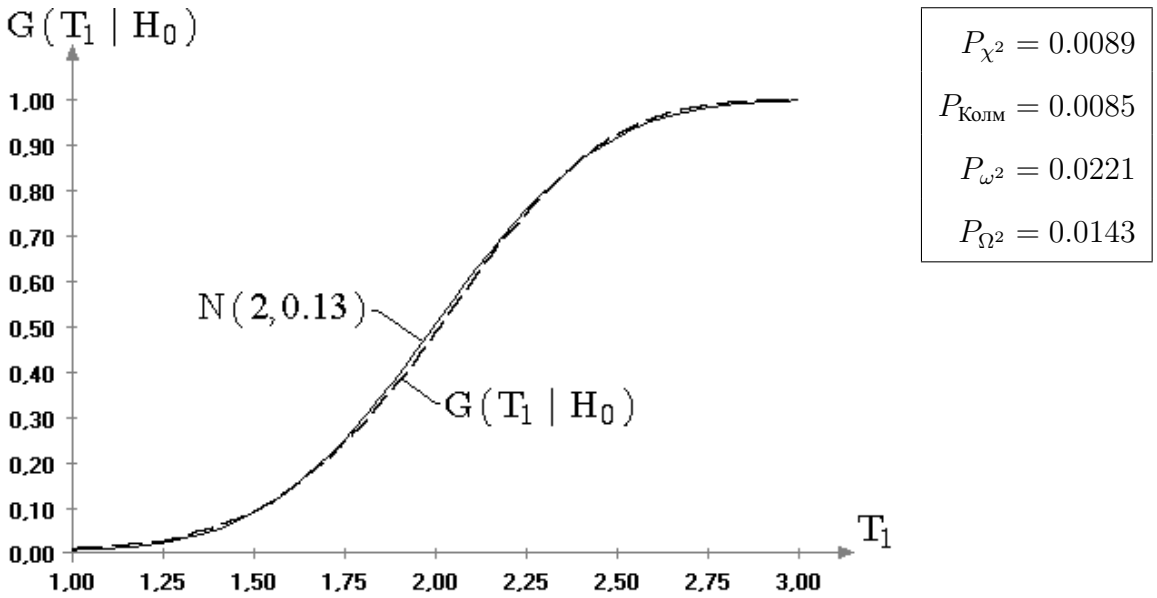


Рис. 2.4. Эмпирическая и теоретическая функции распределения статистики T_1, смоделированной по распределению минимального значения, при проверке гипотезы $H_0 : \mu = 2$ для известной дисперсии ($\sigma^2_{\text{ист}} = 4$): $n = 30$; $N = 5000$

ны также предельные распределения статистик T_3 и T_4 в случае нормального закона (χ^2_{30} и χ^2_{29} — распределения, соответственно).

Из представленной на рис. 2.5 картины очевидно, что распределения статистики $T_3(\lambda)$, смоделированные при выборках случайных величин, принадлежащих семейству распределений (6.4) с параметром формы не равным 2, существенно отличаются от предельного распределения, полученного для нормального закона. Аналогичную зависимость от вида наблюдаемого закона демонстрирует статистика $T_4(\lambda)$ при проверке гипотезы о значении дисперсии при неизвестном математическом ожидании (см. рис. 2.6).

Результаты проведенных исследований говорят о том, что распределения статистик, используемых при проверке гипотез о дисперсии (математическое ожидание известно или неизвестно), значимо отличаются от классических предельных при отклонениях наблюдаемого закона от нормального. Поэтому при использовании классических процедур для проверки гипотез о дисперсии целесообразно удостовериться в том, что наблюдаемый закон является нормальным, применяя соответствующие критерии проверки нормальности.

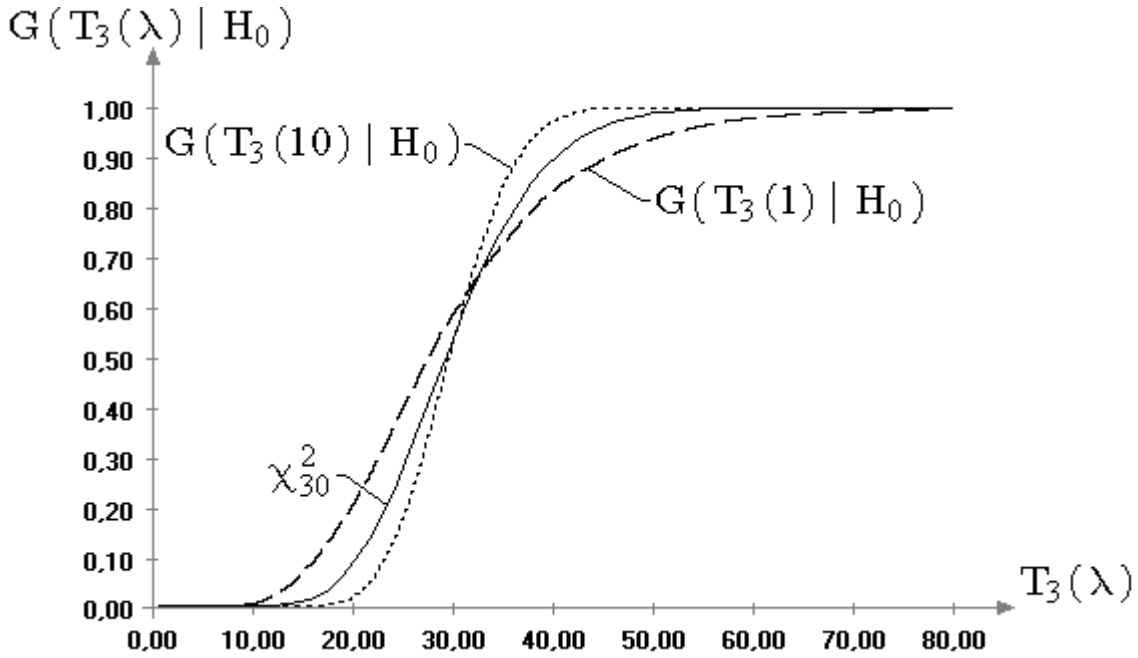


Рис. 2.5. Теоретическая и эмпирические функции распределения статистики T_3 при проверке гипотезы $H_0 : \sigma^2 = 4$ при известном ($\mu_{\text{ист}} = 3$) математическом ожидании: $n = 30$; $N = 5000$

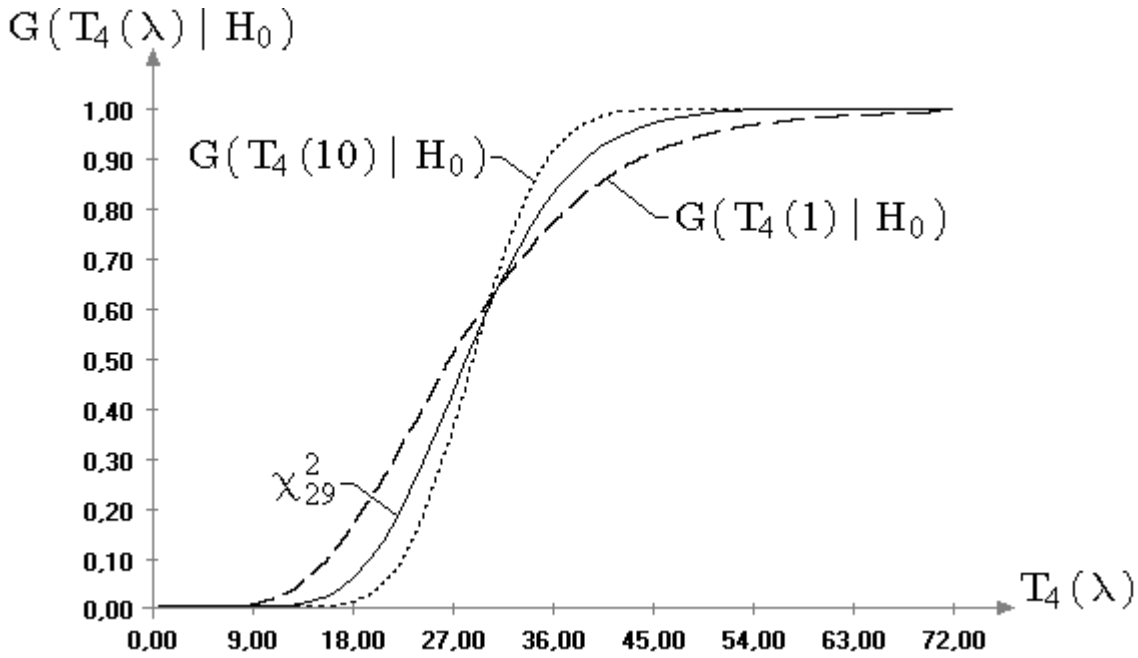


Рис. 2.6. Теоретическая и эмпирические функции распределения статистики T_4 при проверке гипотезы $H_0 : \sigma^2 = 4$ при неизвестном математическом ожидании: $n = 30$; $N = 5000$

Таким образом, приводимые результаты показывают, с одной стороны, высокую устойчивость к отклонениям от нормальности наблюдаемых величин критериев проверки гипотез о математических ожиданиях. А, с другой стороны, — неустойчивость критериев, используемых при проверке гипотез о дисперсиях. В то же время результаты подтверждают возможность построения моделей предельных распределений для статистик T_3 и T_4 при произвольных наблюдаемых законах случайных величин, что актуально для различных приложений задач статистического анализа данных.

Для построения приближенных моделей, наилучшим образом описывающих распределения статистик $T_3(\lambda)$ и $T_4(\lambda)$ при конкретных значениях λ и n, принципиальных трудностей нет. К сожалению, не удается построить аналитические модели распределений данных статистик с параметрами, зависящими от λ и n. Поэтому на основании результатов статистического моделирования были вычислены таблицы верхних процентных точек (квантилей) для ряда значений λ и n. Процентные точки рассчитывались по выборкам значений статистик достаточно больших объемов ($N = 100000$, $N = 150000$ и $N = 200000$), а затем усреднялись по ряду экспериментов.

Полученные процентные точки для статистик $T_3(\lambda)$ и $T_4(\lambda)$ при параметре формы λ семейства распределений (6.4), равном 1, 1.5, 3, 4, 5 и 10 приведены в таблицах 2.2 и 2.3 соответственно. Значения процентных точек при параметре формы $\lambda = 2$, приведенные в таблицах, соответствуют предельным распределениям статистик при нормальном законе наблюдаемых величин.

Таблица 2.2

Верхние процентные точки для статистики T_3^{λ} в случае принадлежности наблюдаемого закона семейству распределений (6.4) с параметром формы λ

		$\lambda = 1$	$\lambda = 1.5$	$\lambda = 2$	$\lambda = 3$	$\lambda = 4$	$\lambda = 5$	$\lambda = 10$
$n = 15$	$\alpha = 0.15$	22.94	21.45	20.64	19.76	19.39	19.18	18.81
	$\alpha = 0.1$	25.98	23.54	22.34	21.06	20.58	20.28	19.77
	$\alpha = 0.05$	31.38	26.98	25.01	23.08	22.41	21.96	21.22
	$\alpha = 0.025$	37.02	30.23	27.46	24.88	24.04	23.45	22.51
	$\alpha = 0.01$	44.36	34.40	30.59	27.03	26.00	25.21	24.02
$n = 30$	$\alpha = 0.15$	41.85	39.31	38.01	36.79	36.21	35.89	35.37
	$\alpha = 0.1$	45.97	42.09	40.26	38.60	37.84	37.41	36.71
	$\alpha = 0.05$	52.92	46.49	43.80	41.37	40.28	39.70	38.70
	$\alpha = 0.025$	59.56	50.59	46.97	43.80	42.47	41.72	40.46
	$\alpha = 0.01$	68.51	55.65	50.88	46.78	45.08	44.15	42.52
$n = 50$	$\alpha = 0.15$	65.86	62.02	60.30	58.77	58.02	57.60	56.91
	$\alpha = 0.1$	70.83	65.50	63.15	61.00	60.04	59.51	58.61
	$\alpha = 0.05$	78.47	70.91	67.51	64.42	63.10	62.36	61.17
	$\alpha = 0.025$	85.83	75.66	71.34	67.51	65.86	64.94	63.39
	$\alpha = 0.01$	95.36	81.92	76.15	71.24	69.22	67.98	66.05
$n = 100$	$\alpha = 0.15$	122.67	116.99	114.57	112.34	111.27	110.69	109.77
	$\alpha = 0.1$	129.31	121.54	118.47	115.47	114.13	113.38	112.15
	$\alpha = 0.05$	139.98	128.64	124.29	120.07	118.37	117.38	115.67
	$\alpha = 0.025$	149.80	135.17	129.33	124.27	122.14	120.90	118.78
	$\alpha = 0.01$	162.04	143.38	135.95	129.27	126.64	125.05	122.42

Таблица 2.3

Верхние процентные точки для статистики T_4^λ в случае принадлежности наблюдаемого закона семейству распределений (6.4) с параметром формы λ

		$\lambda = 1$	$\lambda = 1.5$	$\lambda = 2$	$\lambda = 3$	$\lambda = 4$	$\lambda = 5$	$\lambda = 10$
$n = 15$	$\alpha = 0.15$	21.49	20.16	19.40	18.65	18.34	18.14	17.82
	$\alpha = 0.1$	24.38	22.19	21.03	19.95	19.53	19.26	18.80
	$\alpha = 0.05$	29.52	25.46	23.65	21.92	21.34	20.94	20.27
	$\alpha = 0.025$	34.79	28.64	26.12	23.71	22.95	22.41	21.56
	$\alpha = 0.01$	41.88	32.70	29.25	25.85	24.94	24.22	23.09
$n = 30$	$\alpha = 0.15$	40.54	38.09	36.88	35.71	35.17	34.89	34.39
	$\alpha = 0.1$	44.53	40.83	39.11	37.49	36.79	36.42	35.73
	$\alpha = 0.05$	51.36	45.20	42.60	40.25	39.25	38.71	37.73
	$\alpha = 0.025$	57.85	49.20	45.74	42.70	41.41	40.77	39.51
	$\alpha = 0.01$	66.49	54.21	49.59	45.61	44.01	43.13	41.60
$n = 50$	$\alpha = 0.15$	64.62	60.90	59.24	57.70	56.99	56.59	55.92
	$\alpha = 0.1$	69.58	64.30	62.06	59.95	59.01	58.51	57.63
	$\alpha = 0.05$	77.18	69.65	66.39	63.35	62.07	61.37	60.17
	$\alpha = 0.025$	84.42	74.42	70.20	66.46	64.79	63.96	62.41
	$\alpha = 0.01$	93.75	80.63	74.94	70.21	68.13	66.98	65.05
$n = 100$	$\alpha = 0.15$	121.51	115.87	113.54	111.29	110.26	109.71	108.77
	$\alpha = 0.1$	128.08	120.45	117.35	114.43	113.11	112.39	111.15
	$\alpha = 0.05$	138.70	127.50	123.22	119.07	117.36	116.38	114.67
	$\alpha = 0.025$	148.27	134.04	128.29	123.18	121.07	119.87	117.79
	$\alpha = 0.01$	160.22	142.27	134.71	128.13	125.55	124.04	121.34

2.3. Выводы

Таким образом, численные исследования подтвердили теоретические результаты, приведенные в [58], а именно: устойчивость распределений статистик T_1 и T_2, используемых в критериях проверки гипотез о математических ожиданиях, к отклонениям наблюдаемого закона от нормального и неустойчивость распределений статистик T_3, T_4. Эмпирические распределения статистик $T_1(\lambda)$ и $T_2(\lambda)$ хорошо согласуются с предельными, полученными в предположении о нормальности наблюдаемого закона. Это позволяет на практике корректно применять классические результаты при наблюдаемых законах, существенно отличающихся от нормального. В частности, в таких ситуациях можно уверенно руководствоваться стандартом [39].

Полученные в данном разделе результаты подчеркивают общую закономерность: критерии, связанные с проверкой гипотез о математических ожиданиях устойчивы к отклонениям наблюдаемых величин от нормального закона. Это было показано при исследовании распределений статистик, используемых при проверке гипотез о векторе математических ожиданий многомерного закона распределения [74].

В то же время, как предполагалось [58], распределения статистик T_3 и T_4 очень существенно зависят от вида наблюдаемого закона. Если наблюдаемый закон значимо отличается от нормального, использование классических результатов для данных критериев недопустимо, так как такая попытка неизбежно приведет к некорректным выводам. В тех ситуациях, когда хорошей моделью для наблюдаемых случайных величин оказывается семейство симметричных распределений (6.4) с параметром формы λ, можно воспользоваться таблицами процентных точек, полученными в данной главе.

ГЛАВА 3
ИССЛЕДОВАНИЕ КРИТЕРИЕВ ПРОВЕРКИ ГИПОТЕЗ О ВЕКТОРЕ МАТЕМАТИЧЕСКИХ ОЖИДАНИЙ И КОВАРИАЦИОННОЙ МАТРИЦЕ

В данном разделе методами компьютерного моделирования исследуются распределения статистик критериев проверки гипотез о векторе математических ожиданий и ковариационной матрице при наблюдении случайных величин, подчиняющихся различным многомерным законам распределения [70, 71, 73–75, 78, 79, 83].

3.1. Классические критерии проверки гипотез о векторе математических ожиданий и ковариационной матрице

3.1.1. Проверка гипотез о векторе математических ожиданий

Одной из важных статистических проблем является проблема проверки гипотезы о том, что вектор среднего значения нормального распределения является данным вектором $H_0 : \bar{M} = \bar{M}_0$. Такая задача очень часто возникает на практике, когда, например, на основании наблюдений некоторого технологического процесса желают убедиться, что эти показатели равны номинальному значению $\bar{M}_0$, т.е. процесс протекает нормально, а отклонения наблюдаемых значений от номинальных объясняются лишь ошибками наблюдений (измерений). При решении этой задачи возможны две ситуации: ковариационная матрица Σ может быть известна из ранее проводимых экспериментов, или неизвестна, тогда в процессе вычислений для нее будет построена оценка.

Для проверки гипотезы $H_0 : \bar{M} = \bar{M}_0$ в зависимости от априорной информации могут использоваться различные критерии.

1. Ковариационная матрица Σ известна. В этом случае вычисляется статистика

$$X_m^2 = n \left(\hat{\bar{M}} - \bar{M}_0 \right)^T \Sigma^{-1} \left(\hat{\bar{M}} - \bar{M}_0 \right), \tag{3.1}$$

которая при справедливой гипотезе H_0 в качестве предельного распределения $G(X_m^2|H_0)$ имеет χ_m^2–распределение, с числом степеней свободы m [33].

2. Ковариационная матрица Σ неизвестна. Тогда в критерии проверки гипотезы используется статистика

$$T^2 = \frac{n(n-m)}{m(n-1)}(\hat{\bar{M}} - \bar{M}_0)^T \hat{\Sigma}^{-1}(\hat{\bar{M}} - \bar{M}_0), \tag{3.2}$$

которая при справедливости гипотезы H_0 в пределе подчиняется распределению Фишера с параметрами m и $n-m$: $G(T^2|H_0) = F_{m,n-m}$ [33].

3.1.2. Проверка гипотез о ковариационной матрице

Не менее важной задачей классического корреляционного анализа (вектор $\bar{X}$ принадлежит нормальному закону) является проверка гипотезы о ковариационной матрице $H_0 : \Sigma = \Sigma_0$, где Σ_0 — номинальное значение ковариационной матрицы. В этом случае подразумевается, что вектор математических ожиданий будет оцениваться по данной выборке. Если одновременно проверяется гипотеза и о векторе математических ожиданий, тогда проверяемая гипотеза имеет вид $H_0 : \Sigma = \Sigma_0,\ \bar{M} = \bar{M}_0$.

В критериях проверки данных гипотез используются следующие статистики.

1. Если проверяется гипотеза $H_0 : \Sigma = \Sigma_0$ (математическое ожидание $\bar{M}_0$ неизвестно), тогда вычисляется статистика

$$L_1 = -2\ln\lambda_1 = mn(\ln n - 1) - n\ln\left|B\Sigma_0^{-1}\right| + \operatorname{tr}\left(B\Sigma_0^{-1}\right), \tag{3.3}$$

где

$$B = \sum_{i=1}^{n}\left(\bar{X}_i - \hat{\bar{M}}\right)\left(\bar{X}_i - \hat{\bar{M}}\right)^T.$$

При справедливости гипотезы H_0 данная статистика имеет χ^2–распределение с числом степеней свободы $m(m+1)/2$: $G(L_1|H_0) = \chi^2_{m(m+1)/2}$ [33].

2. Если проверяется гипотеза $H_0 : \Sigma = \Sigma_0,\ \bar{M} = \bar{M}_0$, то используется статистика

$$L_2 = -2\ln\lambda_2 = mn(\ln n - 1) - n\ln\left|B\Sigma_0^{-1}\right| + \operatorname{tr}\left(B\Sigma_0^{-1}\right) + \\ + n\left(\hat{\bar{M}} - \bar{M}_0\right)^T \Sigma_0^{-1}\left(\hat{\bar{M}} - \bar{M}_0\right), \tag{3.4}$$

которая при справедливой гипотезе H_0 в качестве предельного распределения $G(L_2|H_0)$ имеет $\chi^2_{m(m+1)/2+m}$ – распределение, с числом степеней свободы $m(m+1)/2+m$ [33].

Подчеркнем, что рассмотренные выше статистики имеют в качестве предельных указанные распределения лишь при наблюдении многомерного нормального закона. Как изменятся предельные распределения статистик, если наблюдаемый многомерный закон отличается от нормального, заранее сказать нельзя.

3.2. Исследование распределений статистик критериев 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 основанием, подтверждающим корректность применения используемой методики при анализе достоверности результатов последующих исследований.

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

Например, на рис. 3.1 представлены полученное в результате моделирования эмпирическое распределение статистики X_m^2 (3.1) и соответствующее предельное χ_m^2 — распределение при проверке гипотезы $H_0 : M = M_0$ (ковариационная матрица Σ_0 известна) для размерности $m = 2$ и объеме выборки

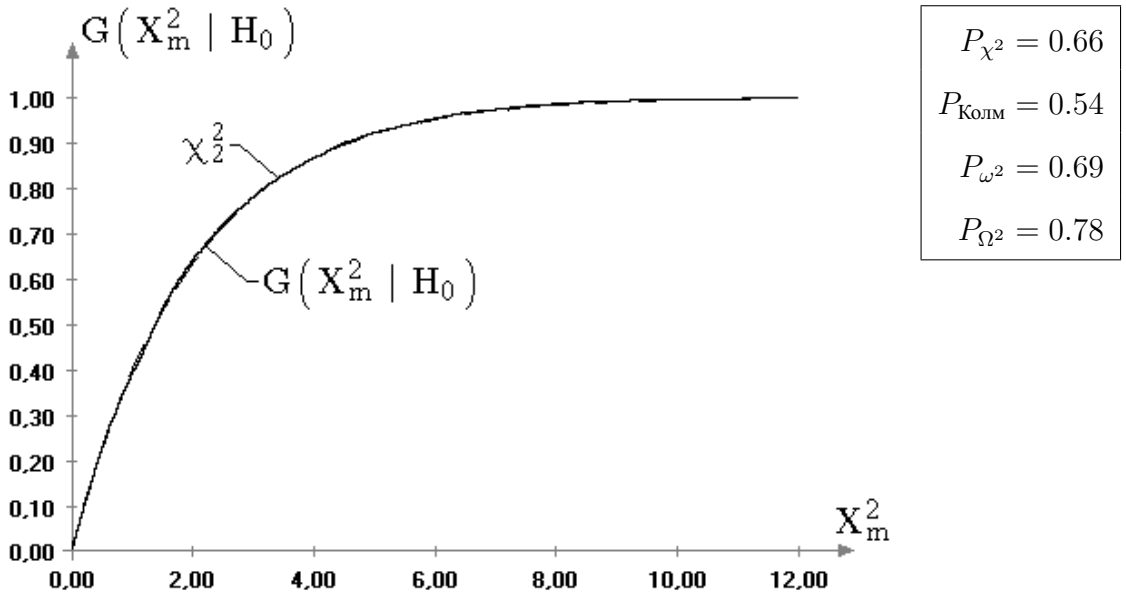


Рис. 3.1. Эмпирическая и теоретическая функции распределения статистики X_m^2 при проверке гипотезы $H_0 : M = M_0$ (ковариационная матрица известна): $m = 2$, $n = 30$

$n = 30$, где использовались

$$\bar{\Theta}_0 = \bar{M}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \Theta_1 = \Sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Рисунок дополнен таблицей, где отражены результаты проверки согласия эмпирического распределения с теоретическим предельным по критериям χ^2 Пирсона, Колмогорова, ω^2 и Ω^2 Мизеса [43, 85]: 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.

В ходе исследований объемы выборок значений статистик N, формируемых в результате моделирования, если не оговариваются явно, в данном разделе и далее предполагаются равными 5000.

На рис. 3.2 приведен пример, где отображены полученная в результате моделирования эмпирическая и теоретическая функции распределения статистики L_1, используемой для проверки гипотезы $H_0 : \Sigma = \Sigma_0$ (математическое ожидание неизвестно), где использовались следующие значения параметров

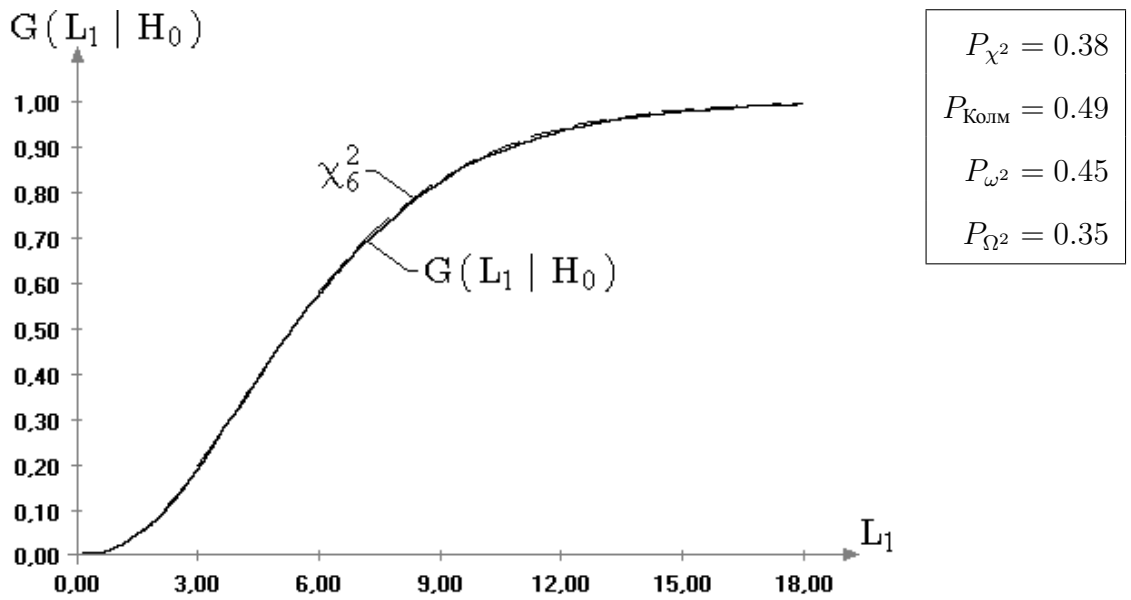


Рис. 3.2. Эмпирическая и теоретическая функции распределения статистики L_1 при проверке гипотезы $H_0 : \Sigma = \Sigma_0$ (математическое ожидание неизвестно): $m = 3$, $n = 100$

$m = 3$ и $n = 100$,

$$\Sigma_0 = \begin{bmatrix} 1 & 0.5 & 0 \\ 0.5 & 5 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Приведенные примеры подтверждают, что эмпирические распределения статистик, используемых в критериях проверки гипотез о векторе математических ожиданий и ковариационной матрице при наблюдении многомерного нормального закона распределения, действительно хорошо описываются соответствующими предельными законами, полученными в [33]. Аналогичная картина, подтверждающая очень хорошее согласие смоделированных эмпирических распределений статистик с классическими предельными, наблюдается и для распределений статистик T^2 (3.2) и L_2 (3.4).

Исследование сходимости распределений рассматриваемых статистик к предельным в зависимости от объема выборки n многомерного нормального закона показало, что для статистик [X_m^2 (3.1), L_1 (3.3) и L_2 (3.4)], параметры предельных распределений которых не зависят от объема выборки, эмпирические распределения статистик оказываются близки к предельным уже при

выборках сравнительно небольшого объема n. Так, у статистики X_m^2 высокий достигаемый уровень значимости по критериям согласия наблюдается, начиная с объемов выборки $n = 30 \div 45$, а для статистик L_1 и L_2 — с $n = 100 \div 150$.

Предельное распределение статистики T^2 зависит от объема выборки случайной величины n. Поэтому предельное распределение как бы «подстраивается» под объем выборки случайного вектора. Вследствие этого уже при малых объемах выборок $n \geq 30$ наблюдаются достаточно высокие достигаемые уровни значимости при проверке соответствия эмпирических распределений статистик предельным законам по критериям согласия.

Отметим, что при исследовании не было выявлено существенного влияния размерности случайного вектора m на сходимость распределений соответствующих статистик к предельным. Исследования проводились для размерности случайного вектора в диапазоне $m \leq 10$.

3.3. Исследование распределений статистик при законах, отличающихся от нормального

Далее проводились исследования распределений статистик для законов многомерных величин, моделируемых в соответствии с предложенной и описанной в главе 6 процедурой. Процедура моделирования опирается на семейство распределений (6.4) и позволяет генерировать псевдослучайные векторы, подчиняющиеся многомерным симметричным законам, более островершинным ($\lambda < 2$) или более плосковершинным ($\lambda > 2$) по сравнению с нормальным законом. Исследования были проведены при значениях параметра $\lambda \geq 1$. Это ограничение обусловлено тем, что предельным случаем семейства распределений (6.4) при $\lambda \to 0$ является распределение Коши, которое представляет собой пример «патологического» распределения: не существует математического ожидания и дисперсия расходится. Поэтому в результате моделирования псевдослучайных векторов при параметре $\lambda < 1$ мы получаем закон с ковариационной матрицей близкой к вырожденной.

Распределения статистик корреляционного анализа при многомерных за-

конах, отличающихся от нормального и моделируемых в соответствии с предлагаемой процедурой, базирующейся на семействе распределений (6.4) с параметром формы λ, определяющим вид закона, исследовались при различных объемах выборок n и различной размерности m случайных величин. Ниже приведены примеры моделирования распределений исследуемых статистик с отражением соответствующих предельных распределений классических статистик. На рисунках представлены значения достигнутых уровней значимости по критериям χ^2 Пирсона, Колмогорова, ω^2 и Ω^2 Мизеса при проверке согласия полученных в результате моделирования эмпирических распределений статистик с предельными распределениями классических статистик.

Для статистик, вычисляемых по выборкам псевдослучайных векторов, смоделированных с использованием параметра формы $\lambda \neq 2$, введем новые обозначения, где в скобках отразим зависимость распределения статистики от параметра λ. Например, для статистики X_m^2 будем использовать новое обозначение $X_m^2(\lambda)$.

На рис. 3.3 показан вид распределения статистики $X_m^2(\lambda)$ в случае закона, смоделированного при параметре $\lambda = 1$. Высокие достигнутые уровни значимости по всем критериям согласия и визуальная близость полученного эмпирического распределения статистики X_m^2 и предельного в случае многомерного нормального закона χ^2-распределения, позволяют утверждать, что вид предельного распределения статистики значимо не изменился. Аналогичная картина видна на рис. 3.4, где показаны эмпирическое распределение статистики $T^2(5)$ и предельное в классическом случае распределение Фишера.

Отметим, что при моделировании (6.6)–(6.7) многомерных величин по несимметричным одномерным законам (в качестве примеров рассматривалась принадлежность $\{Z_i\}, i = \overline{1,m}$, распределениям экстремальных значений) распределения статистик, используемых в критериях проверки гипотез о векторе математических ожиданий, по–прежнему хорошо описываются предельными распределениями, полученными в предположении о нормальности наблюдаемой выборки.

Проведенные исследования распределений статистик X_m^2 и T^2 показали,

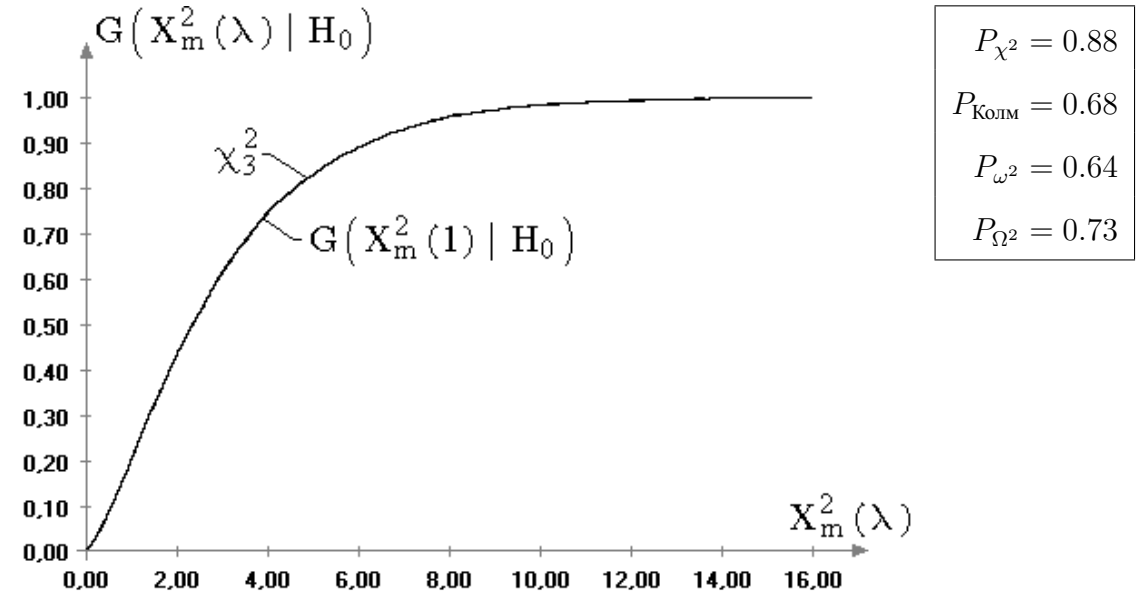


Рис. 3.3. Распределение статистики $X_m^2(1)$ и классическое предельное χ_3^2–распределение ($m = 3$, $n = 30$)

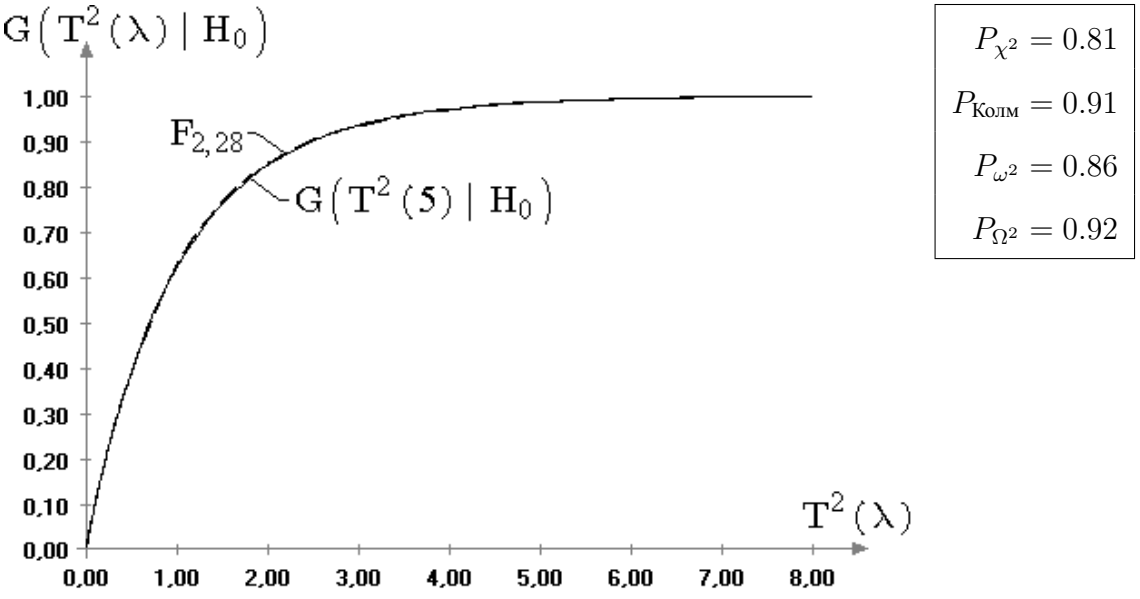


Рис. 3.4. Распределение статистики $T^2(5)$ и классическое предельное $F_{2,28}$–распределение ($m = 2$, $n = 30$)

что в случае многомерных законов, достаточно существенно отличающихся от нормального (более островершинных или более плосковершинных, и даже в случае многомерного закона, построенного по несимметричному одномерному распределению), значимого изменения предельных распределений статистик не происходит. Это позволяет утверждать, что статистические выводы, опирающиеся на классический аппарат в исследованных задачах корреляционного анализа о векторе математических ожиданий, будут оставаться корректными и при нарушении предположений о нормальности наблюдаемого многомерного закона при условии существования вектора математических ожиданий и невырожденности ковариационной матрицы.

В отличие от X_m^2 и T^2 распределения статистик L_1 и L_2, используемых в критериях проверки гипотез о ковариационной матрице, как в случае известного вектора математических ожиданий, так и в случае неизвестного, очень чувствительны к виду наблюдаемого закона распределения. Это хорошо видно на приведенных в качестве примера рисунках 3.5 и 3.6, на которых отображены графики эмпирических распределений статистик $L_1(\lambda)$, $L_2(\lambda)$ и предельные распределения статистик L_1, L_2 в случае нормального закона (χ_6^2 и χ_9^2- распределения, соответственно).

Так, из представленной на рис. 3.5 картины очевидно, что эмпирические распределения статистики $L_1(\lambda)$, смоделированные при значении параметра формы 1 и 10 семейства распределений (6.4), существенно отличаются от предельного распределения статистики L_1, полученного в случае принадлежности наблюдений многомерному нормальному закону. Аналогичную зависимость от вида наблюдаемого закона демонстрирует статистика $L_2(\lambda)$ при проверке гипотезы о ковариационной матрице и математическом ожидании $H_0 : M = M_0, \Sigma = \Sigma_0$ (см. рис. 3.6).

Результаты проведенных исследований говорят о том, что распределения статистик, используемых при проверке гипотез о ковариационной матрице, значимо отличаются от классических предельных при отклонениях наблюдаемого закона от многомерного нормального. Поэтому при использовании классических процедур для проверки гипотез о ковариационной матрице, так-

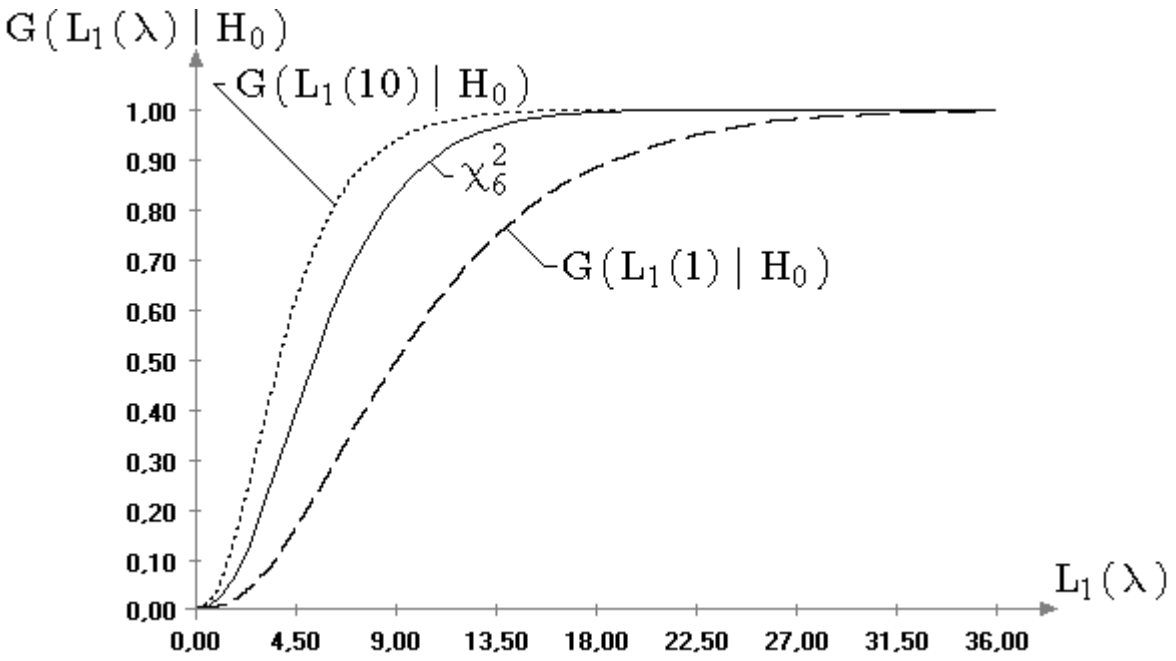


Рис. 3.5. Распределения статистик $L_1(1)$, $L_1(10)$ и предельное распределение статистики L_1: χ^2_6—распределение ($m = 3$, $n = 150$)

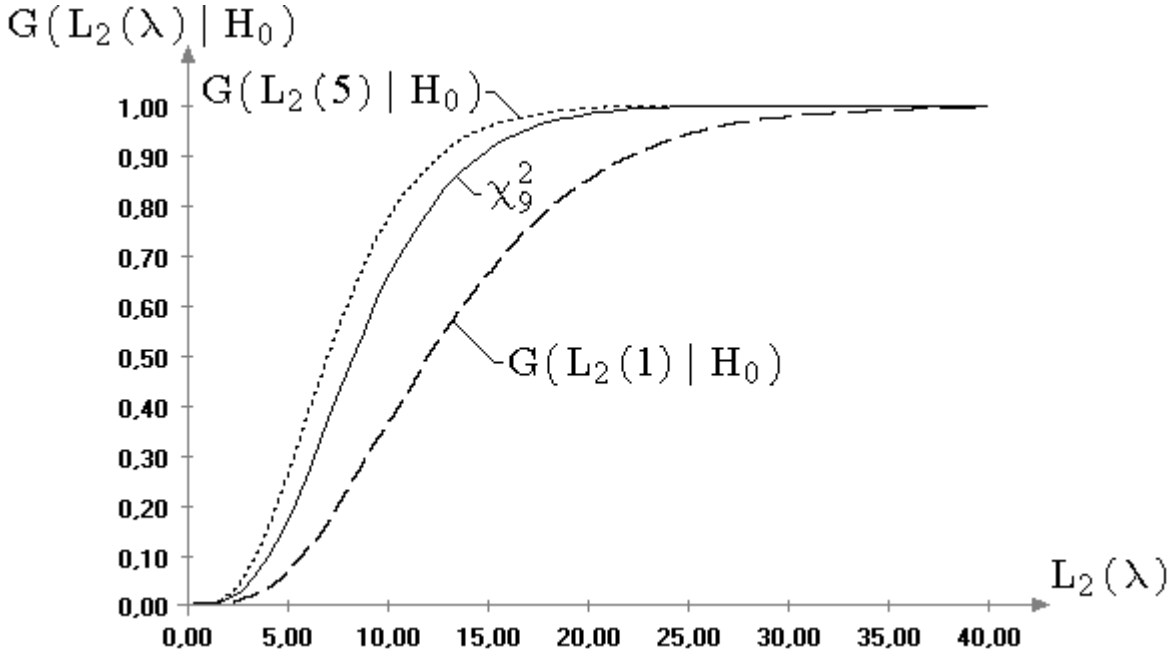


Рис. 3.6. Распределения статистик $L_2(1)$, $L_2(5)$ и предельное распределение статистики L_2: χ^2_9—распределение ($m = 3$, $n = 150$)

же как в одномерном случае при проверке гипотез о дисперсии, целесообразно удостовериться в том, что наблюдаемый закон является нормальным, применяя соответствующие критерии проверки нормальности.

Для проверки предположения об устойчивости статистик X_m^2 и T^2 к отклонению наблюдаемого закона от нормального, было проведено исследование распределений данных статистик в случае многомерного распределения Стьюдента (6.19). Напомним, что с ростом числа степеней свободы ($p \to \infty$) распределение Стьюдента стремится к нормальному закону.

На приведенном рис. 3.7 видно, что, действительно, эмпирическое распределение статистики X_m^2 в случае принадлежности наблюдаемой многомерной случайной величины распределению Стьюдента хорошо описывается χ^2—распределением. Здесь статистика X_m^2 была построена по распределению Стьюдента с числом степеней свободы $p = 15$ и следующих параметрах моделирования: $m = 3$, $n = 50$.

Отметим, что в случае принадлежности случайного вектора многомерному распределению Стьюдента статистика T^2 хорошо описывается классическим $F_{m,n-m}$ распределением, что отображено на рисунке 3.8.

При малых значениях степеней свободы $p \leq 5$ распределения статистик X_m^2 и T^2 претерпевают незначительные изменения, что сказывается на достигаемых уровнях значимости по критериям согласия. Предположительно, такое изменение распределений статистик обусловлено «утяжелением хвостов» распределения Стьюдента. При $p = 1$ распределение Стьюдента представляет собой распределение Коши. А ранее уже отмечалось изменение предельных распределений статистик X_m^2 и T^2 при многомерных законах, построенных по семейству распределений (6.4) с параметром формы $\lambda < 1$.

Полученные результаты для многомерного распределения Стьюдента не опровергают ранее сделанных предположений об устойчивости критериев проверки гипотез о векторе математических ожиданий к отклонению наблюдаемого многомерного закона от нормального. Распределения статистик критериев проверки гипотез о ковариационной матрице, как и ожидалось, сильно зависят от вида многомерного закона. Поэтому распределения статистик L_1 и

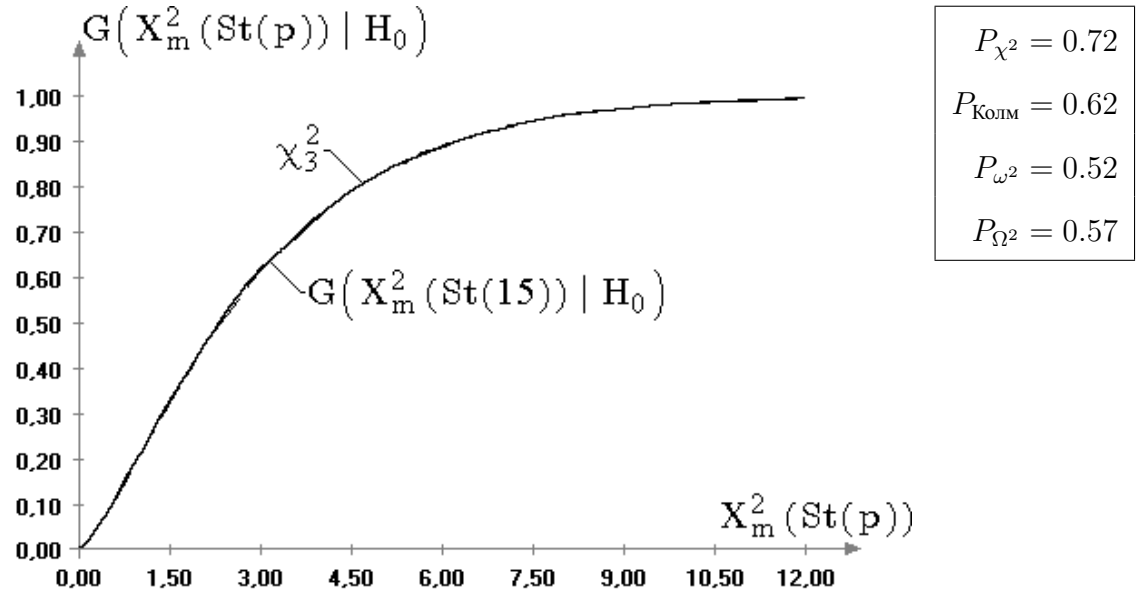


Рис. 3.7. Распределение статистики X^2_m, построенной по многомерному закону Стьюдента с числом степеней свободы $p = 15$, и классическое предельное χ^2_3–распределение ($m = 3$, $n = 50$)

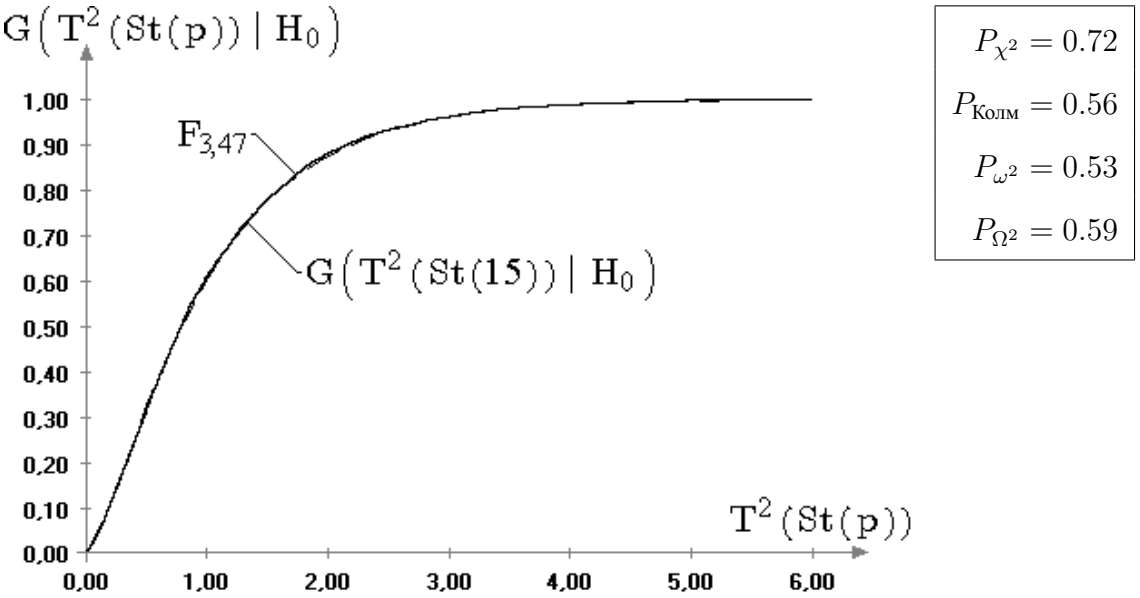


Рис. 3.8. Распределение статистики T^2, построенной по многомерному закону Стьюдента с $p = 15$ степенями свободы, и классическое предельное $F_{3,47}$–распределение ($m = 3$, $n = 50$)

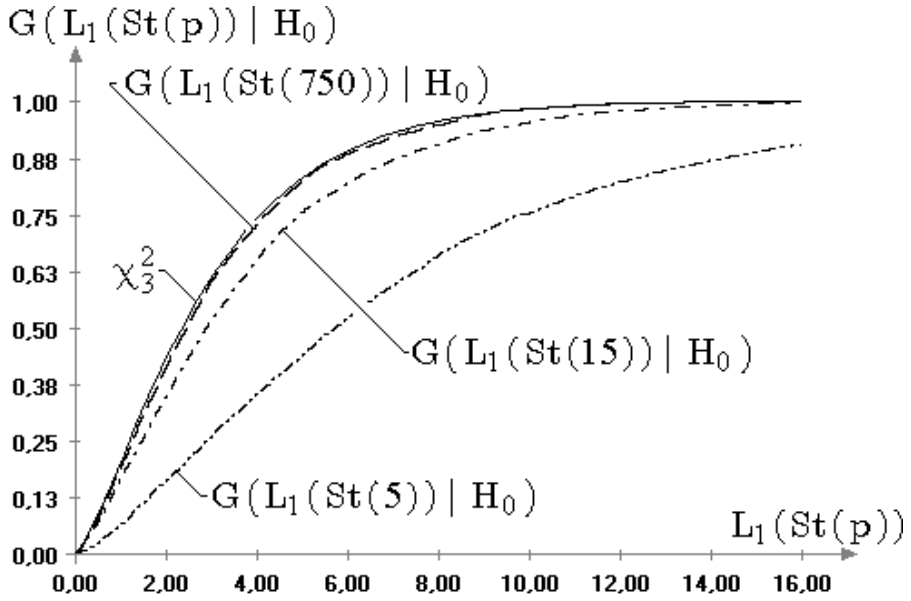


Рис. 3.9. Эмпирические распределения статистики L_1, построенной по многомерному закону Стьюдента с числами степеней свободы $p = 5$, $p = 15$ и $p = 750$, и классическое предельное χ^2_3—распределение ($m = 2$, $n = 150$)

L_2 стремятся к классическим предельным только при очень больших значениях числа степеней свободы $p > 750$ (когда распределение Стьюдента по виду очень близко к нормальному закону). В качестве примера на рисунке 3.9 показано, что распределение статистики L_1, моделируемой по двумерному закону Стьюдента при очень большом значении числа степеней свободы $p = 750$, уже достаточно хорошо описывается предельным классическим χ^2—распределением статистики (3.3).

3.4. Уточнение моделей распределений статистик рассматриваемых критериев

Как показано выше, распределения статистик, используемых в критериях проверки гипотез о векторе математических ожиданий, при существенном отличии наблюдаемого закона от нормального *незначимо отличаются от предельных распределений, полученных в классическом случае*. Результаты моделирования распределений статистик X^2_m и T^2 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 оснований для отказа от использования в качестве предельных в соответствующих случаях распределений χ^2 или Фишера.

Распределение χ^2 представляет собой частный случай гамма–распределения, F–распределение Фишера – частный случай бета–распределения 2-го рода. Если, например, действительно χ^2–распределение является предельным распределением некоторой статистики и в том случае, когда нарушается предположение о нормальности наблюдаемой многомерной величины, а 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.

Исходя из вышесказанного, мы попытались уточнить модели распределений статистик X_m^2 и T^2 следующим образом. Моделировалась выборка интересующей нас статистики, как правило, объемом в 5000 наблюдений. Эмпирическое распределение статистики сглаживалось соответствующей моделью (гамма–распределением, бета–распределением) 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 в соответствующем случае к ее частному случаю χ^2–распределению.

Предельным распределением классической статистики X_m^2 является χ_m^2–распределение (3.1), где m – размерность многомерного вектора. Это соответствует гамма–распределению с плотностью

$$f(x;\sigma,\theta) = \frac{x^{\theta-1}}{\sigma^{\theta}\Gamma(\theta)} e^{-\frac{x}{\sigma}}, \tag{3.5}$$

с параметром формы $\theta = m/2$ и параметром масштаба $\sigma = 2$.

Таблица 3.1

Оценки параметров выравнивающего гамма–распределения для статистики X_m^2, построенной по многомерным законам с различными λ ($m = 3$)

Параметры	$\lambda = 1$	$\lambda = 2$	$\lambda = 5$	$\lambda = 10$
σ	2.0368	2.0012	1.9892	1.9754
θ	1.4727	1.5019	1.5137	1.5164

Таблица 3.2

Оценки параметров выравнивающего бета–распределения для статистики T^2, построенной по многомерным законам с различными λ ($m = 3$ и $n = 30$)

Параметры	$\lambda = 1$	$\lambda = 2$	$\lambda = 5$	$\lambda = 10$
σ	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

В таблице 3.1 представлены усредненные по 50 смоделированным выборкам статистики X_m^2 значения параметров модели гамма–распределения, аппроксимирующего распределение статистики в случае законов многомерных величин, моделируемых при различных значениях параметра датчика λ ($\lambda = 2$ соответствует нормальному закону). В данном примере размерность моделируемых многомерных величин $m = 3$. Напомним, что значимого влияния размерности m на сходимость распределения статистики X_m^2 к предельному выявлено не было. Очевидно, что значения параметров в случае наблюдения нормального закона сходятся к значениям 2 и 1.5 соответственно, что соответствует χ_3^2–распределению. По крайней мере, нет оснований для отклонения данного предположения.

При проверке аналогичной гипотезы при неизвестной ковариационной матрице предельным распределение статистики T^2 является $F_{m,n-m}$–распределение. Данному случаю соответствует бета–распределение 2-го рода, плотность

Таблица 3.3

Оценки параметров гамма–распределений, используемых в качестве моделей распределений статистики $L_1(\lambda)$, построенной по многомерным законам с различными λ

	$\lambda=1$		$\lambda=2$		$\lambda=3$		$\lambda=4$		$\lambda=5$	
Параметры	σ	θ	σ	θ	σ	θ	σ	θ	σ	θ
$m=2$	4.21	1.46	2.00	1.50	1.68	1.49	1.45	1.50	1.48	1.43
$m=3$	3.83	2.71	2.00	3.00	1.74	2.97	1.70	2.85	1.69	2.77
$m=4$	3.58	4.43	2.00	5.00	1.80	4.99	1.78	4.84	1.75	4.80
$m=5$	–	–	2.00	7.50	1.84	7.50	1.87	7.08	1.85	7.00

которого имеет вид

$$f(x;\sigma,\theta_0,\theta_1)=\frac{1}{\sigma B(\theta_0,\theta_1)}\frac{\left[\dfrac{(x-\mu)}{\sigma}\right]^{\theta_0-1}}{\left[1+\dfrac{(x-\mu)}{\sigma}\right]^{\theta_0+\theta_1}}, \tag{3.6}$$

с масштабным параметром $\sigma=\frac{n-m}{m}$, параметрами формы $\theta_0=\frac{m}{2}$ и $\theta_1=\frac{n-m}{2}$.

Представленные в таблице 3.2 усредненные по 50 смоделированным выборкам статистики T^2 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.

Таким образом, уточнение моделей распределений статистик X_m^2 и T^2 еще раз подтверждает предположение об устойчивости соответствующих критериев к отклонению от нормальности.

В случае статистик L_1 и L_2, которые используются при проверке гипотез о ковариационной матрице, видна явная зависимость распределений данных статистик от вида наблюдаемого многомерного закона. Поэтому для распределений статистик L_1 и L_2 постарались найти подходящие аналитические модели

Таблица 3.4

Оценки параметров гамма–распределений, используемых в качестве моделей распределений статистики $L_2(\lambda)$, построенной по многомерным законам с различными λ

	$\lambda=1$		$\lambda=2$		$\lambda=3$		$\lambda=4$		$\lambda=5$	
Параметры	σ	θ	σ	θ	σ	θ	σ	θ	σ	θ
$m=2$	3.53	2.25	2.00	2.50	1.80	2.46	1.73	2.42	1.72	2.40
$m=3$	3.36	3.99	2.00	4.50	1.87	4.38	1.81	4.36	1.83	4.20
$m=4$	3.31	6.05	2.00	7.00	1.86	7.02	1.89	6.66	1.84	6.72
$m=5$	3.22	8.55	2.00	9.00	1.92	9.80	1.99	9.26	1.99	9.10

законов. К сожалению, как и в одномерном случае [76, 77], нам не удалось построить модели распределений данных статистик с параметрами, зависящими от λ. Поэтому на основании результатов статистического моделирования были найдены оценки параметров моделей законов, которые наилучшим образом (по критериям согласия) подходят для описания эмпирических распределений данных статистик. Оценки параметров распределений находились по выборкам значений статистик $L_1(\lambda)$ и $L_2(\lambda)$ достаточно больших объемов ($N = 5000$), а затем усреднялись по ряду экспериментов.

Полученные оценки параметров гамма–распределений, которые оказались наилучшими моделями для распределений статистик $L_1(\lambda)$ и $L_2(\lambda)$ при значениях параметра формы λ, равном 1, 3, 4 и 5, приведены в таблицах 3.3 и 3.4 соответственно. Значения параметров гамма–распределения при $\lambda = 2$, приведенные в таблицах, соответствуют предельным распределениям статистик при нормальном законе наблюдаемых величин.

Если наблюдается многомерный закон, отличный от нормального, а маргинальные функции плотности данного закона хорошо описываются семейством распределений (6.4), тогда при помощи таблиц 3.3 и 3.4 можно подобрать наилучшую модель для распределений статистик L_1 и L_2. Например, если в двумерном случае ковариационная матрица имеет диагональный вид, а

маргинальные функции распределения описываются семейством распределений (6.4) при параметре формы равным 1, тогда в качестве предельного закона распределения статистики L_1 можно использовать гамма—распределение с параметрами $\sigma = 4.21$ и $\theta = 1.46$.

3.5. Выводы

Исследования эмпирических распределений статистик, используемых в критериях проверки гипотез о векторе математических ожиданий и ковариационной матрице, при псевдослучайных величинах, подчиняющихся многомерному нормальному закону, показали, что они хорошо согласуются с теоретическими предельными распределениями, полученными в классическом корреляционном анализе, и подтвердили эффективность методики исследований.

Исследования распределений статистик X_m^2 и T^2 в случае многомерных законов, отличающихся от нормального в достаточно широких пределах (более островершинных или более плосковершинных), 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 при проверке гипотез о векторе математических ожиданий в приложениях. Аналогичная ситуация наблюдается в одномерном случае: на распределениях статистик, вычисляемых при проверке гипотез вида $H_0 : \mu = \mu_0$ при известной и неизвестной дисперсии, отклонения от нормальности наблюдаемого одномерного закона сказываются незначительно.

Используемые в критериях проверки гипотез о ковариационной матрице многомерного закона статистики L_1 и L_2 существенно зависят от наблюдаемого многомерного закона, что и подтвердили проведенные исследования. Это согласуется с полученными результатами при моделировании распреде-

лений аналогичных статистик в одномерном случае (при проверке гипотез вида $H_0 : \sigma = \sigma_0$ при известном и неизвестном математическом ожидании). Для распределений статистик L_1 и L_2 были найдены аналитические модели законов, описывающие распределения этих статистик при определенных значениях размерности m и параметре формы λ. При необходимости аналогичные аналитические модели могут быть построены для любых интересующих нас значений параметров m и λ.

ГЛАВА 4
ИССЛЕДОВАНИЕ КРИТЕРИЕВ ПРОВЕРКИ ГИПОТЕЗ О КОЭФФИЦИЕНТАХ КОРРЕЛЯЦИИ

В классическом корреляционном анализе на основании исследований парных, частных и множественных корреляций можно делать выводы о характере статистической зависимости. Когда требуется определить взаимозависимость двух величин, исследуется *парная корреляция*. В случае, если интересует взаимозависимость двух величин, когда устранено воздействие остальных величин, то исследуется, так называемая, *частная корреляция*. А когда требуется рассмотреть зависимость единственной величины от группы других, исследуют *множественную корреляцию*. В этой главе исследуется устойчивость критериев, используемых в задачах о выявлении характера статистической зависимости между двумя или большим числом случайных величин при наблюдении различных многомерных законов распределения [70, 71, 73–75, 79].

4.1. Классические критерии проверки гипотез о коэффициентах корреляции

4.1.1. Проверка гипотез о коэффициентах парной корреляции

Взаимозависимость двух компонент случайного вектора характеризуется парным коэффициентом корреляции r_{ij}. Он представляет собой меру тесноты линейной связи. Известно, что независимость двух случайных величин влечет равенство $r_{ij} = 0$, однако обратное утверждение в общем случае неверно. Что и представляет трудность интерпретации r_{ij} как коэффициента взаимозависимости в общем случае. Однако, оно справедливо для совместно нормальных величин. Коэффициент корреляции можно использовать в качестве некоторой меры взаимозависимости для нормального закона. Если известна оценка ковариационной матрицы $\hat{\Sigma}$, то оценка парного коэффициента корреляции может быть найдена в соответствии с выражением

$$\hat{r}_{ij} = \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}}. \tag{4.1}$$

В классическом корреляционном анализе относительно парного коэффициента корреляции могут проверяться два вида гипотез: о значимости коэффициента корреляции ($H_0 : r_{ij} = 0$) и о равенстве его номинальному значению ($H_0 : r_{ij} = r_0$).

1. В критерии проверки гипотезы $H_0 : r_{ij} = 0$ используется статистика

$$t = \frac{\sqrt{n-2}\,\hat{r}_{ij}}{\sqrt{1-\hat{r}_{ij}^2}}, \tag{4.2}$$

которая при справедливости гипотезы H_0 имеет в качестве предельного распределение Стьюдента с $n-2$ степенями свободы: $G(t|H_0) = t_{n-2}$ [33].

2. В случае проверки гипотезы $H_0 : r_{ij} = r_0$ вычисляется статистика

$$z_0 = \sqrt{n-3}\left(\frac{1}{2}\ln\left(\frac{1+\hat{r}_{ij}}{1-\hat{r}_{ij}}\right) - \frac{1}{2}\ln\left(\frac{1+r_0}{1-r_0}\right) - \left(\frac{r_0}{2(n-1)}\right)\right), \tag{4.3}$$

которая при справедливости гипотезы H_0 в качестве предельного распределения $G(z_0|H_0)$ имеет стандартный нормальный закон $N(0,1)$ [33].

В [58] выдвинуто предположение о том, что критерий некоррелированности ($H_0 : r_{ij} = 0$) можно строить без каких-либо предположений о нормальности исходного распределения.

Известно, что оценка для r_{ij} является смещенной, когда $0 < r_{ij}^2 < 1$, что видно из выражения [58]

$$E[\hat{r}_{ij}] = r_{ij}\left\{1 - \frac{1-r_{ij}^2}{2n} + O(n^{-2})\right\}.$$

Олкин и Прэтт [58] рекомендуют использовать несмещенную оценку в виде

$$\hat{r}_{ij}^{\text{н}} = \hat{r}_{ij}\left\{1 + \frac{1-\hat{r}_{ij}^2}{2(n-4)}\right\}. \tag{4.4}$$

4.1.2. Проверка гипотез о коэффициентах частной корреляции

Как ранее отмечалось, в случае двух нормальных или почти нормальных величин коэффициент корреляции между ними может быть использован в качестве меры взаимозависимости. Однако на практике при интерпретации «взаимозависимости» часто сталкиваются с трудностями, заключающимися в том, что, если одна величина коррелирована с другой, то это может быть всего лишь отражением того факта, что обе они коррелированы с некоторой третьей величиной или с совокупностью величин. Указанная возможность приводит к необходимости рассмотрения условных корреляций между двумя величинами при фиксированных значениях остальных величин. Это так называемые *частные корреляции*.

Если корреляция между двумя величинами уменьшается при фиксировании некоторой другой случайной величины, то это означает, что их взаимозависимость возникает частично через воздействие этой величины. Если же частная корреляция равна нулю или очень мала, то делается вывод, что их взаимозависимость целиком обусловлена этим воздействием. Наоборот, когда частная корреляция больше первоначальной корреляции между двумя величинами, то следует, что другие величины ослабляли связь, или, можно сказать, «маскировали» корреляцию. Но следует помнить, что даже в последнем случае нельзя предполагать наличие причинной связи, так как некоторая, совершенно отличная от рассматриваемых при анализе, величина может быть источником этой корреляции. Как при обычной корреляции, так и при частных корреляциях предположение о причинности должно всегда иметь внестатистические основания.

Представим случайный вектор $\bar{X}$ в следующем виде [33]:

$$\bar{X} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \end{bmatrix}, \text{где} \quad \bar{X}_1 = \begin{bmatrix} X_1 \\ \vdots \\ X_l \end{bmatrix}, \bar{X}_2 = \begin{bmatrix} X_{l+1} \\ \vdots \\ X_m \end{bmatrix},$$

а вектор математических ожиданий и ковариационную матрицу соответствен-

но в виде

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

Тогда если случайный вектор $\bar{X}$ подчиняется нормальному закону с вектором средних $\bar{M}$ 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 $\Sigma_{11\cdot 2}$, где $B = \Sigma_{12}\Sigma_{22}^{-1}$, $\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ [58].

ОМП для частного коэффициента корреляции определяется соотношением:

$$\hat{r}_{ij\cdot l+1,\ldots,m} = \frac{\hat{\sigma}_{ij\cdot l+1,\ldots,m}}{\sqrt{\hat{\sigma}_{ii\cdot l+1,\ldots,m}\hat{\sigma}_{jj\cdot l+1,\ldots,m}}}, \tag{4.5}$$

где $\hat{\sigma}_{ij\cdot l+1,\ldots,m}$ — элемент i-й строки и j-го столбца матрицы $\Sigma_{11\cdot 2}$, 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 $X_{l+1}, X_{l+2}, \ldots, X_m$.

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

1. В критерии проверки гипотезы $H_0 : r_{ij\cdot l+1,\ldots,m} = 0$ используется статистика

$$t^p = \frac{\sqrt{n-m+l-2}\;\hat{r}_{ij\cdot l+1,\ldots,m}}{\sqrt{1-\hat{r}^2_{ij\cdot l+1,\ldots,m}}}, \tag{4.6}$$

которая при справедливой гипотезе H_0 имеет в качестве предельного распределение Стьюдента с $n-m+l-2$ степенями свободы: $G(t^p|H_0) = t_{n-m+l-2}$ [33, 58].

2. В случае проверки гипотезы $H_0 : r_{ij\cdot l+1,\ldots,m} = r_0$ вычисляемая статистика

$$\begin{aligned} z_0^p = \sqrt{n-m+l-3}\left(\frac{1}{2}\ln\left(\frac{1+\hat{r}_{ij\cdot l+1,\ldots,m}}{1-\hat{r}_{ij\cdot l+1,\ldots,m}}\right) - \right. \\ \left. - \frac{1}{2}\ln\left(\frac{1+r_0}{1-r_0}\right) - \left(\frac{r_0}{2(n-m+l-1)}\right)\right), \end{aligned} \tag{4.7}$$

при справедливой гипотезе H_0 в качестве предельного распределения $G(z_0^p|H_0)$ имеет стандартное нормальное распределение $N(0,1)$ [33, 58].

4.1.3. 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.

Следует отметить, что множественный коэффициент корреляции $r_{i\cdot l+1,\ldots,m}$ случайной величины X_i относительно некоторого множества других случайных величин всегда не меньше, чем абсолютная величина любого парного коэффициента корреляции r_{ij} с таким же первичным индексом. Более того, множественный коэффициент корреляции никогда нельзя уменьшить путем расширения множества величин, относительно которых измеряется зависимость X_i.

Если коэффициент корреляции между X_i и множеством всех остальных компонент многомерной случайной величины равен нулю ($r_{i\cdot l+1,\ldots,m} = 0$), то все коэффициенты корреляции этой величины относительного любого подмножества также равны 0, т.е. величина X_i полностью некоррелирована со всеми остальными величинами.

С другой стороны, если $r_{i\cdot l+1,\ldots,m}$ относительно множества всех остальных компонент равен единице $r_{i\cdot l+1,\ldots,m} = 1$, то, по крайней мере, один из коэффициентов корреляции относительно некоторого подмножества компонент должен быть равен 1.

Надо отметить, что коэффициент корреляции, например, между X_1 и множеством всех остальных компонент является обычным коэффициентом корреляции между X_1 и условным математическим ожиданием $E[X_1|X_2,\ldots,X_m]$.

С учетом выше рассмотренного разбиения случайного вектора $\bar{X}$ ОМП множественного коэффициента корреляции между $X_i, i \leq l$ и множеством

компонент $X_{l+1}, X_{l+2}, \ldots, X_m$ определяется соотношением

$$\hat{r}_{i\cdot l+1,\ldots,m} = \sqrt{\frac{\hat{\sigma}_{(i)}\Sigma_{22}^{-1}\hat{\sigma}_{(i)}^T}{\hat{\sigma}_{ii}}}, \tag{4.8}$$

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

Для проверки гипотезы о значимости множественного коэффициента корреляции $H_0 : r_{i\cdot l+1,\ldots,m} = 0$ вычисляется статистика

$$F = \frac{n-m+l-1}{m-l}\frac{\hat{r}^2_{i\cdot l+1,\ldots,m}}{1-\hat{r}^2_{i\cdot l+1,\ldots,m}}, \tag{4.9}$$

предельным распределением $G(F|H_0)$ которой является $F_{m-l,n-m+l-1}$ — распределение Фишера с параметрами $m-l$ и $n-m+l-1$ [33, 58].

4.2. Исследование распределений статистик критериев для различных многомерных законов

4.2.1. В случае принадлежности наблюдений многомерному нормальному закону

Как и ранее в первую очередь при помощи статистического моделирования нами исследовались распределения статистик, используемых при проверке гипотез о различных коэффициентах корреляции, на подчиненность соответствующим предельным распределениям в случае многомерного нормального закона. Проведенные экспериментальные исследования подтвердили хорошее согласие между получаемыми эмпирическими распределениями статистик критериев о коэффициентах корреляции и соответствующими предельными законами.

В процессе исследования сходимости распределений статистик к предельным в зависимости от объема выборки n нами были оценены объемы выборок нормальных псевдослучайных векторов, начиная с которых наблюдается близость эмпирической и теоретической функций распределений статистик. Так, у статистик z_0 и z_0^p высокий достигаемый уровень значимости наблюдается, начиная с объемов выборки $n = 100 \div 150$, а для статистик t, t^p и F — с $n \geq 30$ (следствие зависимости предельных распределений данных статистик от n).

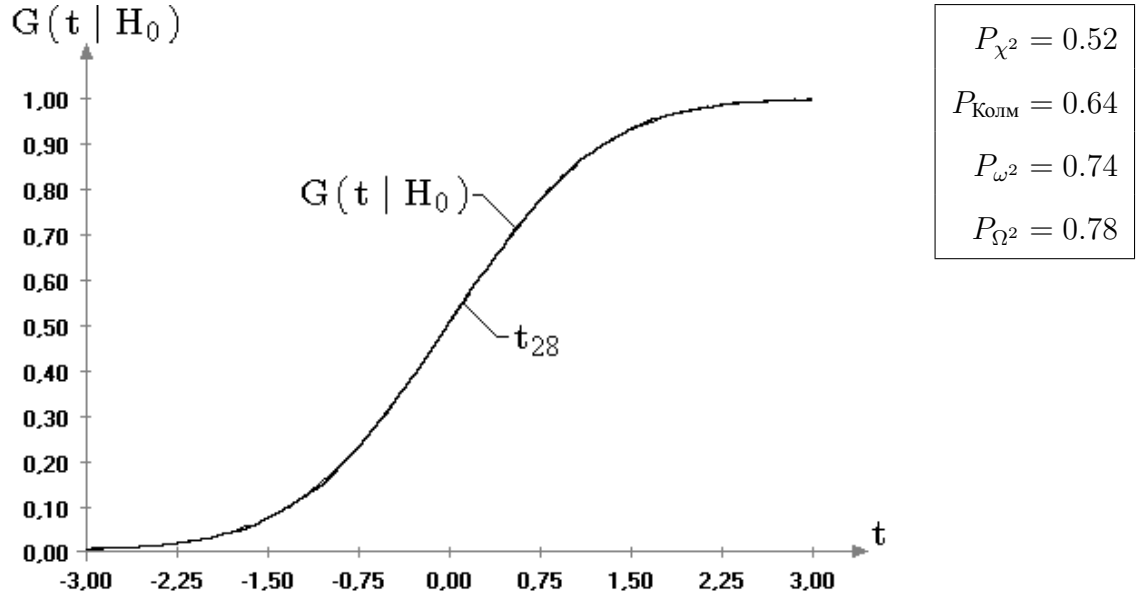


Рис. 4.1. Эмпирическая и теоретическая функции распределения статистики t (4.2) при проверке гипотезы $H_0 : r_{23} = 0$, построенная с использованием параметров моделирования (4.10): $n = 30$

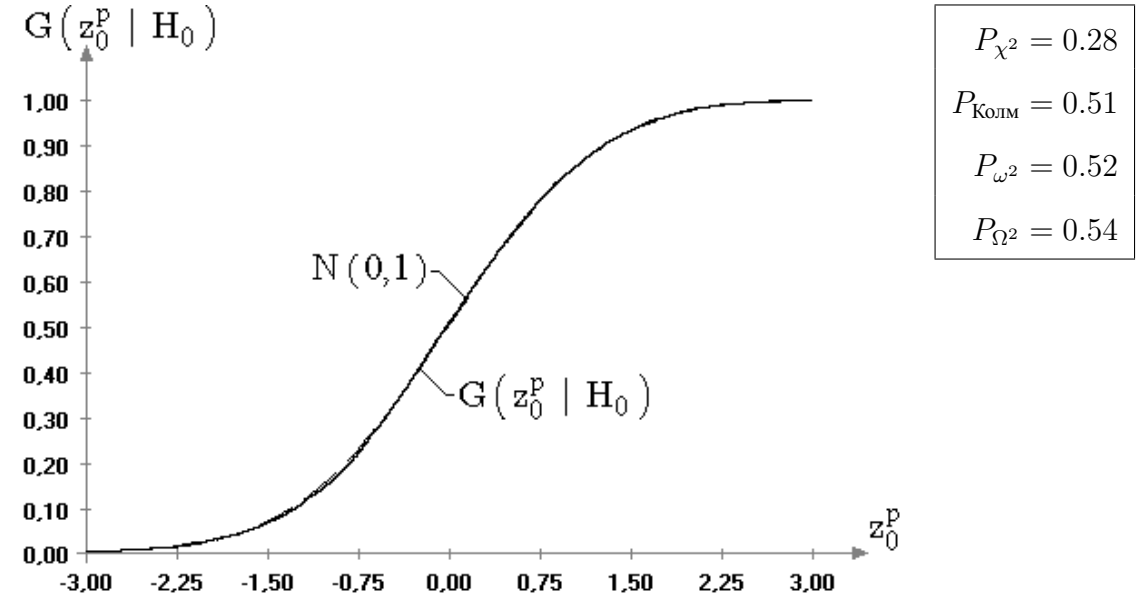


Рис. 4.2. Эмпирическая и теоретическая функции распределения статистики z_0^p (4.7) при проверке гипотезы $H_0 : r_{12 \cdot 3} = 0.21$, построенной с использованием параметров моделирования (4.10): $n = 100$

Продемонстрируем сказанное на двух примерах, со следующими наборами параметров моделирования

$$\bar{\Theta}_0 = \bar{M}_0 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \quad \Theta_1 = \Sigma_0 = \begin{bmatrix} 5 & 1 & 2.5 \\ 1 & 6 & 0 \\ 2.5 & 0 & 5 \end{bmatrix}, \tag{4.10}$$

$$\bar{\Theta}_0 = \bar{M}_0 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \quad \Theta_1 = \Sigma_0 = \begin{bmatrix} 5 & 0.5 & 2.5 \\ 0.5 & 6 & 1 \\ 2.5 & 1 & 5 \end{bmatrix}. \tag{4.11}$$

На рисунке 4.1 приведены в качестве примера полученная в результате моделирования эмпирическая и теоретическая функции распределения статистики t (4.2), используемой при проверке гипотезы о незначимости парного коэффициента корреляции ($H_0 : r_{23} = 0$). В данном случае при моделировании использовались следующие значения параметров: $m = 3$, $n = 30$, а $\bar{\Theta}_0$ и Θ_1 из (4.10). На основании достигнутых уровней значимости критериев согласия, приведенных на рисунке, и визуальной близости эмпирической и теоретической функций распределения статистики t можно судить о достаточности объемов выборок $n \geq 30$ случайных векторов для приемлемого согласия. Аналогичная картина наблюдается и при моделировании распределений статистики t^p (4.6).

Пример на рис. 4.2 демонстрирует близость между распределениями статистики z_0^p (4.7), построенными для многомерного нормального закона при моделировании с параметрами $m = 3$, $l = 2$, $n = 100$, $\bar{\Theta}_0$ и Θ_1 (4.11). Вновь наблюдается высокий достигаемый уровень значимости при проверке согласия между эмпирическим и теоретическим распределениями используемой статистики, начиная с объемов выборок $n \geq 100$. Полученные результаты моделирования статистики z_0 подтверждают общую картину, полученную при исследовании статистики z_0^p.

По результатам исследования распределений статистики F (4.9), используемой при проверке гипотезы о равенстве множественного коэффициента корреляции нулевому значению, моделируемых, например, с параметрами $m = 3$,

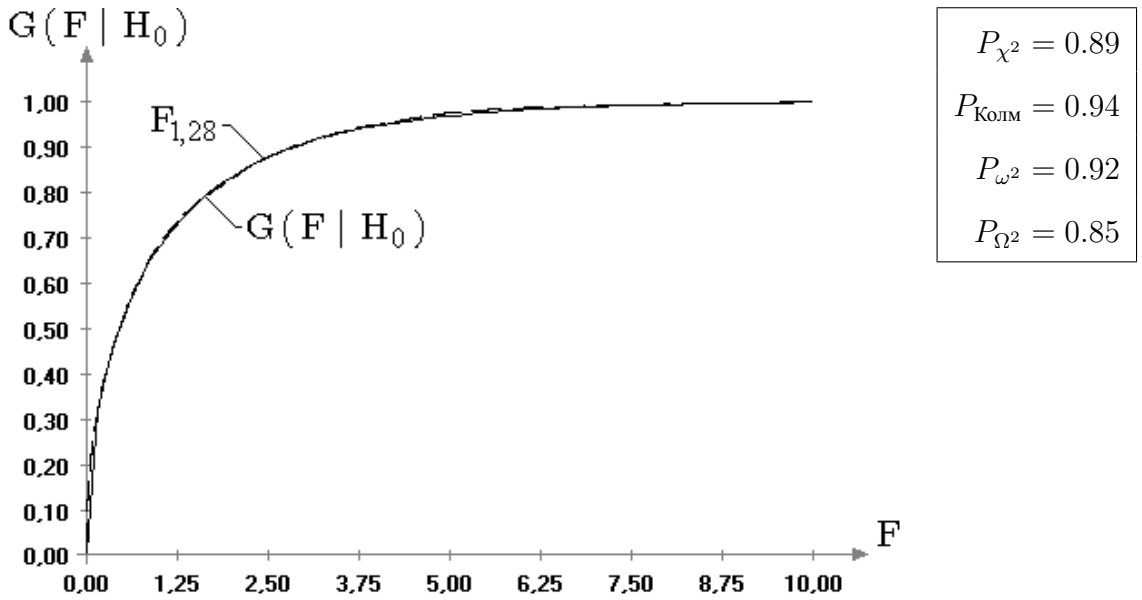


Рис. 4.3. Эмпирическая и теоретическая функции распределения статистики F (4.9) при проверке гипотезы $H_0 : r_{2\cdot3} = 0$, построенная с использованием параметров моделирования (4.10): $n = 30$

$l = 2$, $n = 30$, $\bar{\Theta}_0$ и Θ_1 (4.10), можно говорить о «достаточности» объемов выборок случайных векторов, начиная с $n \geq 30$. Результаты описанного эксперимента приведены на рис. 4.3.

Отметим, что при исследовании вновь не было выявлено существенного влияния размерности случайного вектора m и на сходимость распределений статистик данных критериев к соответствующим классическим предельным.

В работе [103] показано, что оценка парного коэффициента корреляции по формуле (4.1) не является устойчивой по отношению к нарушению предположения о нормальности распределения, из которого получена выборка для вычисления оценки. Различные робастные аналоги оценки коэффициента приведены во многих работах [1,7,20,21,42,102]. Например, одна из таких оценок имеет вид

$$\hat{r}_{ij} = \frac{m\left\{[X_{ki} - m\{X_{ki}\}][X_{kj} - m\{X_{kj}\}]\right\}}{\left(m\{[X_{ki} - m\{X_{ki}\}]^2\} m\{[X_{kj} - m\{X_{kj}\}]^2\}\right)^{1/2}}, \tag{4.12}$$

где $m\{X_{ki}\}_{k=1}^{n}$ — медиана псевдослучайных величин X_i.

Если использовать оценку (4.12) в статистике t (4.2), то наблюдается явное изменение предельного распределения статистики, что отражено на ри-

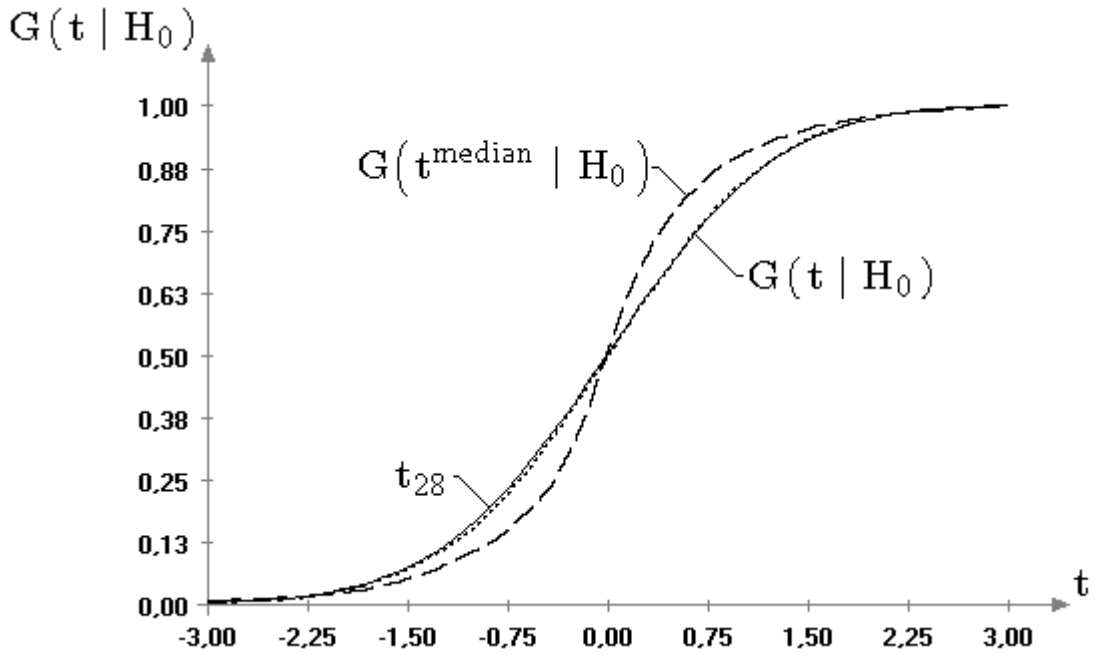


Рис. 4.4. Эмпирические и теоретическая функции распределения статистики t (4.2) при проверке гипотезы $H_0 : r_{12} = 0$, построенных с использованием оценок парного коэффициента корреляции по формулам (4.1) и (4.12): $n = 30$

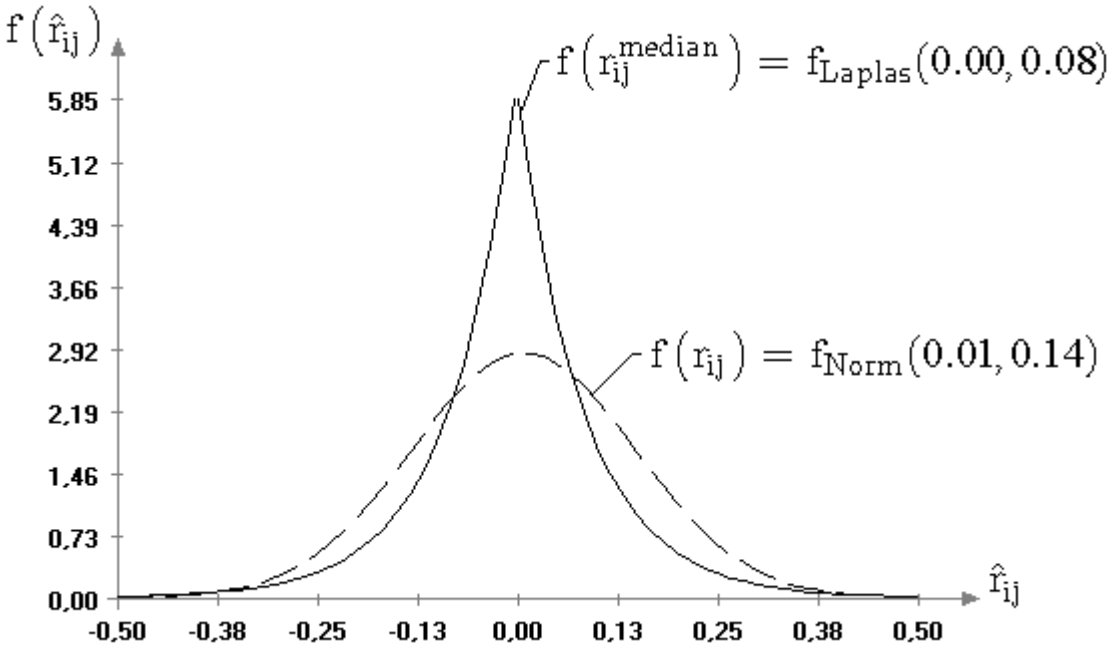


Рис. 4.5. Функции плотности оценок парного коэффициента корреляции, вычисляемого по формулам (4.1) и (4.12)

сунке 4.4. Такое изменение объясняется тем, что функция плотности распределения оценки (4.12) становится более «островершинной» (следствие робастности оценки). На рисунке 4.5 приведены функции плотности распределения оценок $\hat{r}_{ij}$ при $H_0 : r_{ij} = 0$, полученные в результате моделирования. Где для распределения оценки, вычисленной по формуле (4.1), лучше всего подходит нормальный закон с соответствующими параметрами сдвига и масштаба $f_{Norm}(0.01, 0.14)$, а для оценки (4.12) — распределение Лапласа $f_{Laplas}(0.00, 0.08)$. Это различие в распределениях оценок коэффициента парной корреляции и приводит к существенному уменьшению размаха предельного распределения статистики t (см. рис. 4.4).

Отсюда следует, что применяя критерии проверки гипотез о парном коэффициенте корреляции, следует использовать оценки по методам, указанным при построении критериев: в данном случае — по методу максимального правдоподобия.

4.2.2. В случае принадлежности наблюдений многомерным законам, моделируемым на основе семейства симметричных распределений (6.4)

Распределения статистик, используемых в критериях проверки гипотез о коэффициентах корреляции, исследовались при различных объемах выборок n и различной размерности случайных величин m на многомерных законах, моделируемых с использованием предложенной в данной работе процедуры. Ранее отмечалось, что в [58] выдвигалось предположение об устойчивости распределения статистики t (4.2) (критерий некоррелированности) к отклонениям от нормальности наблюдаемого закона. Там же была показана явная зависимость распределения статистики z_0 (4.3) от вида многомерного закона. Проверим эти предположения на моделируемых многомерных законах.

Приведем полученные в результате исследований примеры смоделированных эмпирических распределений статистик с отражением близости их к соответствующим предельным распределениям, полученным в предположении о нормальности выборки. Количественной мерой близости служат достигае-

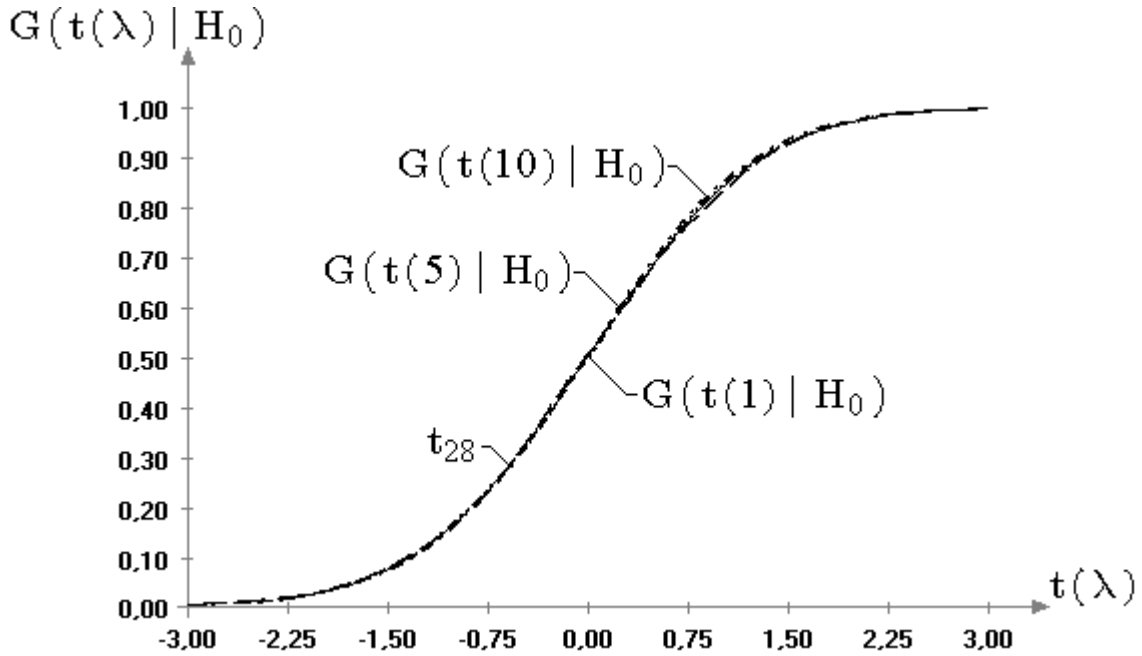


Рис. 4.6. Эмпирические распределения статистик $t(1)$, $t(5)$, $t(10)$ и классическое предельное t_{28}—распределение статистики (4.2) при проверке гипотезы $H_0 : r_{23} = 0$, где $n = 30$ и (4.10)

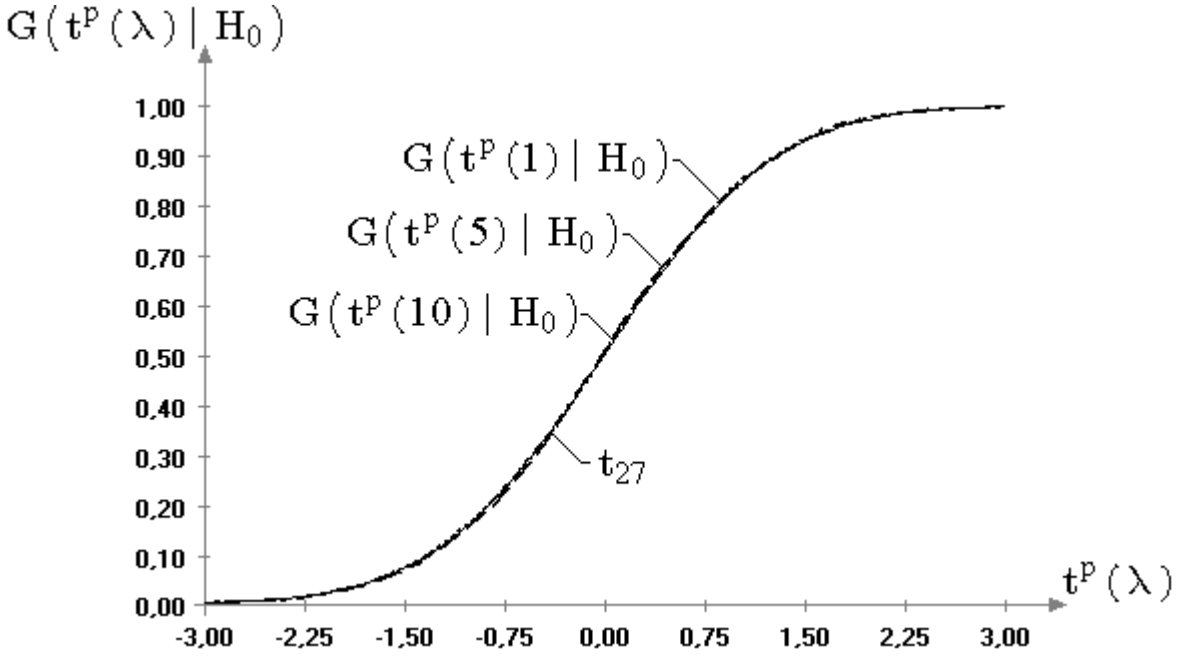


Рис. 4.7. Эмпирические распределения статистик $t^p(1)$, $t^p(5)$, $t^p(10)$ и классическое предельное t_{27}—распределение статистики (4.6) при проверке гипотезы $H_0 : r_{12 \cdot 3} = 0$, где $n = 30$, $m = 3$, $l = 2$ и (4.11)

мые уровни значимости по критериям согласия χ^2 Пирсона, Колмогорова, ω^2 и Ω^2 Мизеса. Чем ближе достигнутый уровень значимости к 1, тем лучше согласие эмпирического распределения с соответствующим теоретическим.

Из результатов приведенных на рисунках 4.6 и 4.7 следует, что нет оснований для отклонения предположений о том, что предельными распределениями статистик критериев проверки гипотез о равенстве парного и частного коэффициентов корреляции нулевому значению при наблюдении многомерных законов, построенных по одномерному закону из семейства распределений (6.4) при разных параметрах формы λ, являются соответствующие классические предельные распределения. Достигаемые уровни значимости по критериям согласия для результатов, отраженных на данных рисунках, сведены в таблицу 4.1. Результаты исследований показали, что распределения статистик (4.2) и (4.6) устойчивы к отклонениям многомерного закона от нормального.

Статистика (4.9), используемая при проверке гипотезы о равенстве нулю множественного коэффициента корреляции, также оказалась нечувствительна к отклонениям многомерного закона от нормального (рис. 4.8, таб. 4.1).

Таким образом, проведенные численные исследования не опровергают выдвигаемого в [58] предположения об устойчивости критериев проверки гипотез о равенстве нулю парного коэффициента корреляции по отношению к на-

Таблица 4.1

Значения достигнутых уровней значимости по критериям согласия для распределений статистик t, t^p и F, смоделированных при различных параметрах формы λ, приведенных на рисунках 4.6, 4.7 и 4.8

	t (рис. 4.6)			t^p (рис. 4.7)			F (рис. 4.8)		
	$\lambda=1$	$\lambda=5$	$\lambda=10$	$\lambda=1$	$\lambda=5$	$\lambda=10$	$\lambda=1$	$\lambda=5$	$\lambda=10$
$P_{\chi^2}=$	0.32	0.95	0.72	0.88	0.59	0.88	0.60	0.91	0.71
$P_{\text{Колм}}=$	0.23	0.99	0.97	0.76	0.47	0.82	0.32	0.77	0.25
$P_{\omega^2}=$	0.32	0.97	0.99	0.53	0.42	0.85	0.44	0.86	0.50
$P_{\Omega^2}=$	0.35	0.96	0.95	0.37	0.42	0.85	0.41	0.90	0.22

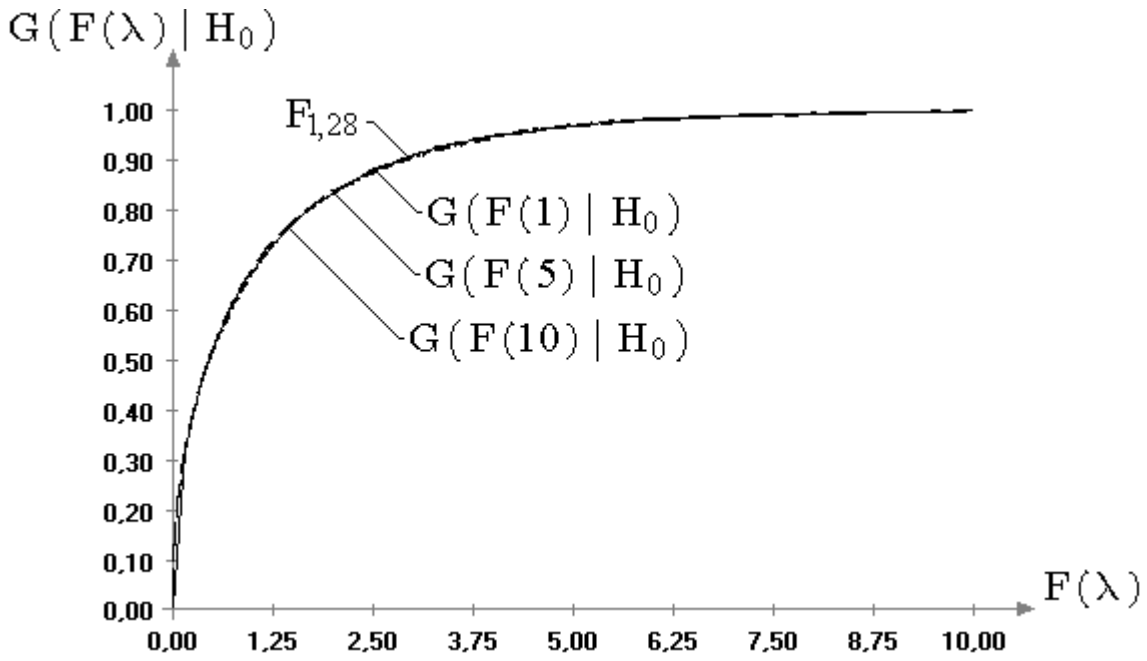


Рис. 4.8. Эмпирические распределения статистик $F(1)$, $F(3)$, $F(5)$ и классическое предельное $F_{1,28}$—распределение статистики (4.9) при проверке гипотезы $H_0 : r_{1\cdot3} = 0$, где $n = 30$, $m = 3$, $l = 2$ и (4.10)

рушению основного предположения корреляционного анализа о нормальности многомерного закона. Исследования проводились на различных размерностях псевдослучайных векторов и большом количестве повторных экспериментов с целью исключения ошибок возможных отдельных экспериментов. Поэтому можно выдвинуть более широкое предположение о том, что критерии проверки гипотез о нулевых значениях парного, частного и множественного коэффициентов корреляции являются устойчивыми к отклонениям от нормальности.

В критериях проверки гипотез о равенстве парного или частного коэффициента корреляции заданному значению распределения используемых статистик критериев очень чувствительны к виду наблюдаемого закона. Так, с ростом отклонения коэффициента корреляции от нулевого значения при прочих равных условиях происходит все более значимое отклонение распределения соответствующей статистики от классического предельного. Сказанное иллюстрирует рисунок 4.9, на котором показано, как с увеличением абсолютного значения коэффициента корреляции, изменяется распределение статистики данного кри-

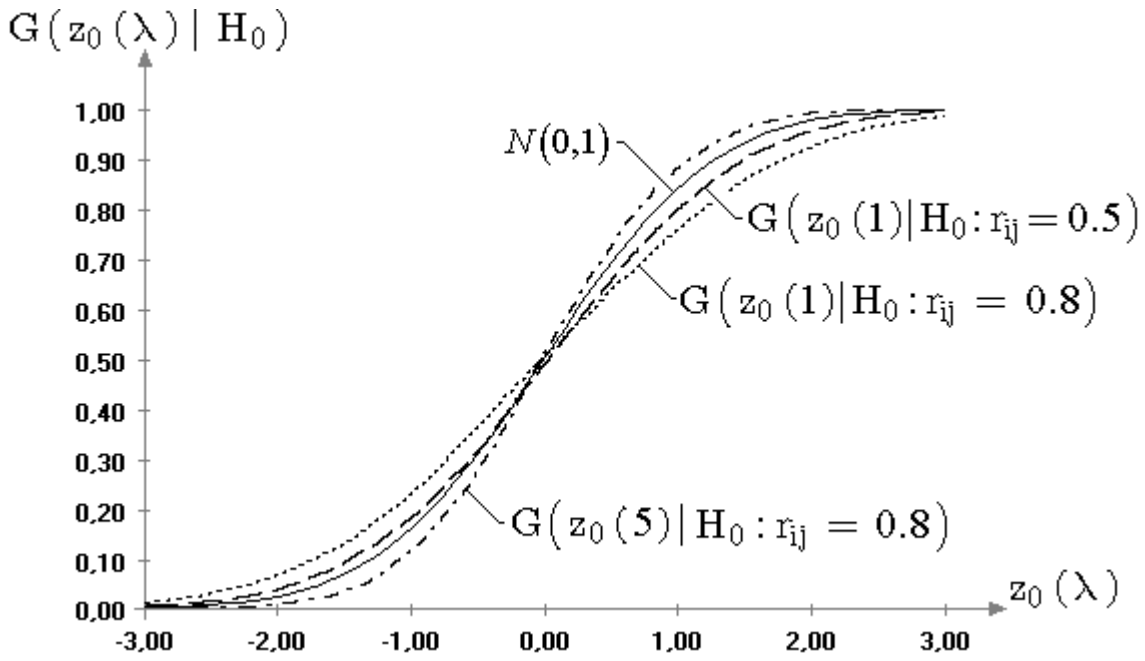


Рис. 4.9. Эмпирические распределения статистики $z_0(\lambda)$, построенные для проверки гипотез на равенство коэффициента парной корреляции различным значениям, и классическое предельное
$N(0,1)$—распределение статистики (4.3)

терия. В то время как в классическом случае распределение статистики в пределе стремится к стандартному нормальному распределению и не зависит от значения коэффициента корреляции.

На основании результатов исследований можно дать следующие рекомендации. При законах, отличных от нормального, и малых значениях парного (частного) коэффициента корреляции $0 < |r| \leq 0.15$ еще можно пользоваться стандартным нормальным распределением как предельным для статистики z_0 (z_0^p). Но при значениях коэффициента корреляции $|r| > 0.15$ требуется определение распределения статистики используемого критерия.

4.2.3. Случай принадлежности наблюдений многомерному закону Стьюдента

Исследования распределений статистик критериев проверки гипотез о коэффициентах корреляции на многомерном распределении Стьюдента показало

Таблица 4.2

Значения достигнутых уровней значимости по критериям согласия между t_{48}—распределением и распределением статистики t, смоделированной по многомерному закону Стьюдента при различных степенях свободы p, усредненных по 3 экспериментам ($m = 3$, $n = 50$)

	$p = 5$	$p = 15$	$p = 25$	$p = 35$	$p = 45$
$P_{\chi^2} =$	0.00	0.00	0.06	0.31	0.64
$P_{\text{Колм}} =$	0.00	0.02	0.19	0.15	0.76
$P_{\omega^2} =$	0.00	0.01	0.18	0.21	0.76
$P_{\Omega^2} =$	0.00	0.00	0.07	0.16	0.74
$P_{\text{сред}} =$	0.00	0.01	0.13	0.21	0.73

ограниченность применения классических результатов для выборок, не принадлежащих многомерному закону. Так, при наблюдении выборок, подчиняющихся закону Стьюдента с числом степеней свободы $p \leq 30$, распределения статистик t (4.2), t^p (4.6) и F (4.9) не сходятся к классическим предельным при объемах $n = 50 \div 100$, являющихся достаточными для нормального закона. Это отражено на рисунках 4.10 и 4.11, где видно, что эмпирические распределения данных статистик, полученные в результате моделирования многомерных величин по закону Стьюдента с числом степеней свободы $p = 5$ и $p = 15$, не подчиняются соответствующим предельным распределениям для нормального случая. Значительное увеличение объемов многомерных выборок $n > 250$ не улучшает сходимость распределений статистик t, t^p и F к классическим предельным.

При дальнейшем увеличении параметра $p > 30$ согласие между распределениями данных статистик и соответствующими предельными законами в нормальном случае заметно улучшается (см. рисунки 4.10, 4.11 и таблицу 4.2).

Исследования распределений статистик критериев по выборкам многомерного распределения, построенного по семейству распределений (6.4) с параметром формы $\lambda < 1$, демонстрируют аналогичные результаты, что и в

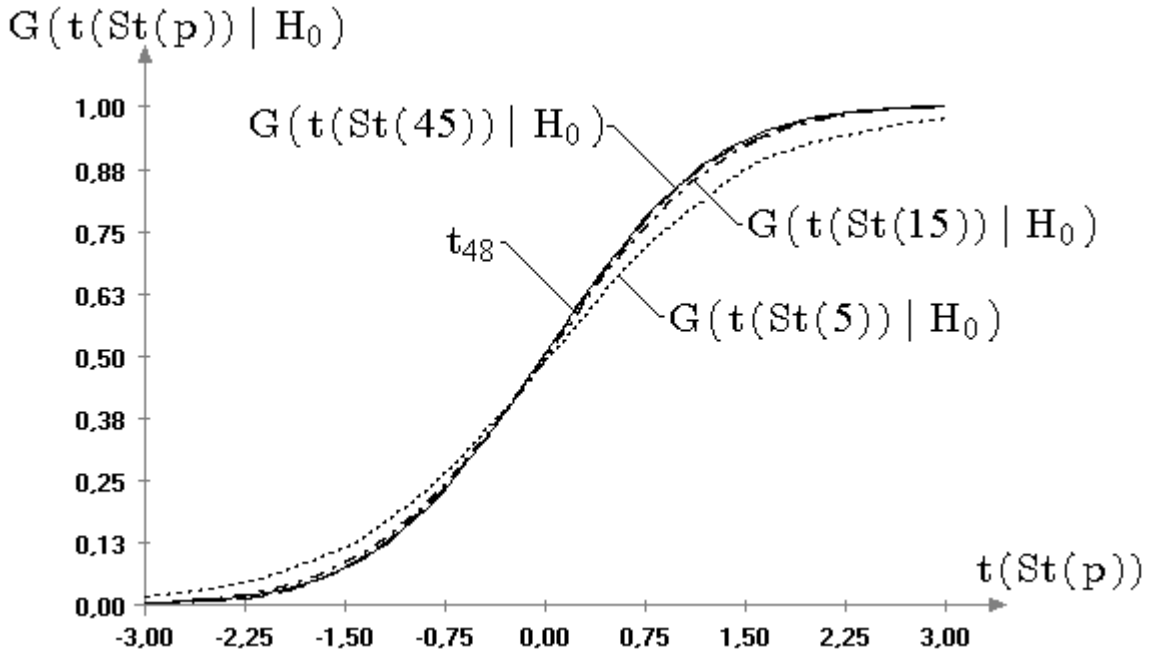
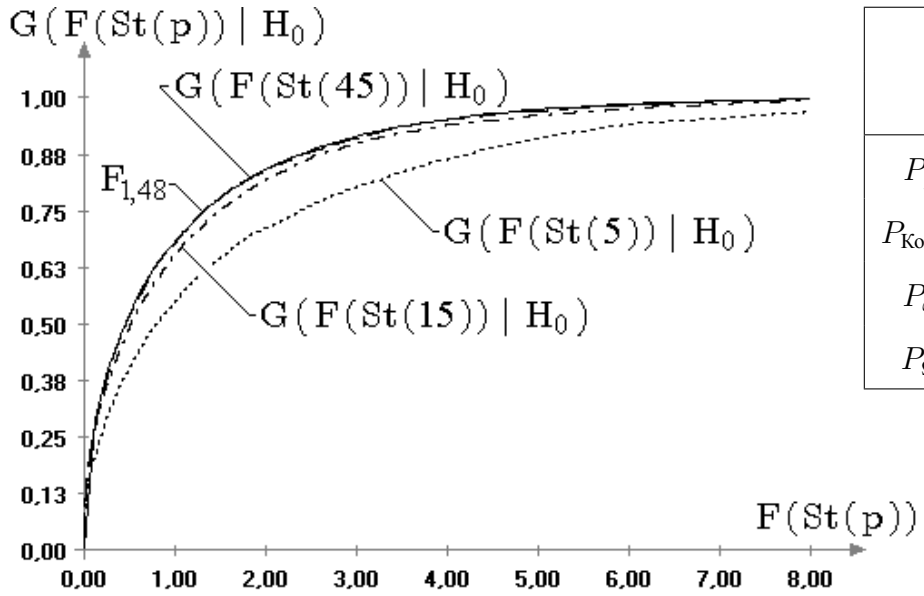


Рис. 4.10. Эмпирические распределения статистики t, построенные по многомерному распределению Стьюдента для $p = 5$, $p = 15$ и $p = 45$ числа степеней свободы, и классическое предельное t_{48}—распределение статистики (4.2) при проверке гипотезы $H_0 : r_{12} = 0$

случае многомерного распределения Стьюдента с числом степеней свободы $p \leq 30$: распределения статистик t, t^p и F претерпевают изменения и более не подчиняются соответствующим предельным распределениям, полученным в предположении о нормальности. Изменение предельного закона статистики t, моделирумой по семейству распределений (6.4) с параметрами формы $\lambda = 0.3$ и $\lambda = 0.5$, отражено на рисунке 4.12.

Многомерные распределения Стьюдента при $p < 30$ и многомерные распределения, моделируемые на основе семейства распределений (6.4) с параметром формы $\lambda < 1$, представляют собой законы с «тяжелыми хвостами». При $p = 1$ и $\lambda \to 0$ в том и другом случае мы приходим к многомерному распределению Коши.

Оценки максимального правдоподобия вектора математических ожиданий и, особенно, ковариационных матриц (а, следовательно, и ОМП коэффициентов корреляции) не являются робастными. Их асимптотические свойства



	p		
	5	15	45
P_{χ^2}	0.00	0.00	0.57
$P_{\text{Колм}}$	0.00	0.00	0.16
P_{ω^2}	0.00	0.00	0.45
P_{Ω^2}	0.00	0.00	0.57

Рис. 4.11. Эмпирические распределения статистики F, построенные по многомерному распределению Стьюдента для $p = 5$, $p = 15$ и $p = 45$ числа степеней свободы, и классическое предельное $F_{1,48}$—распределение статистики (4.9) при проверке гипотезы $H_0 : r_{1\cdot 3} = 0$

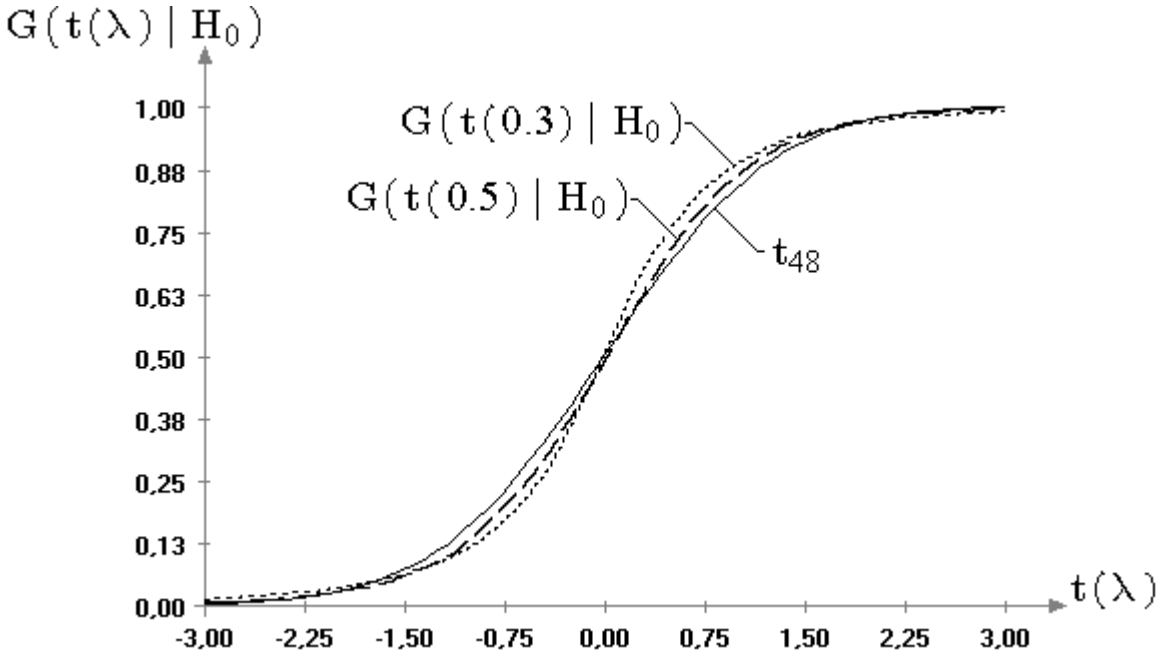


Рис. 4.12. Эмпирические распределения статистики $t(0.3)$, $t(0.5)$, и классическое предельное t_{48}—распределение статистики (4.2) при проверке гипотезы $H_0 : r_{12} = 0$, где $m = 3$, $n = 50$

резко ухудшаются (увеличивается рассеяние) в случае многомерных законов при $p < 30$ и $\lambda < 1$. Этим объясняется неустойчивость критериев проверки гипотез на нулевые значения коэффициентов корреляции для многомерных законов, построенных с помощью соответствующих процедур при $p < 30$ и $\lambda < 1$, и устойчивость этих же критериев для многомерных законов при $p > 30$ и $\lambda \geq 1$.

4.3. Выводы

Исследования эмпирических распределений статистик, используемых в критериях проверки гипотез о парных, частных и множественных коэффициентах корреляции, при псевдослучайных величинах, подчиняющихся многомерному нормальному закону, показали, что они хорошо согласуются с теоретическими предельными распределениями, полученными в классическом корреляционном анализе. Отмечено существенное влияние метода вычисления оценок коэффициентов корреляции на распределения статистик данных критериев.

Исследования распределений статистик t, t^p и F в случае многомерных законов, отличающихся от нормального в достаточно широких пределах, 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 раздвигает границы корректного применения методов классического корреляционного анализа при проверке гипотез о нулевых значениях парного, частного и множественного коэффициентов корреляции.

В случае многомерных законов с «тяжелыми хвостами» наблюдается значимое отличие распределений статистик t, t^p и F от предельных классических.

Используемые в критериях проверки гипотез о равенстве заданному значению парного или частного коэффициента корреляции статистики z_0 и z_0^p существенно зависят от наблюдаемого многомерного закона. Это подтвержда-

ет выдвинутое в [58] предположение о зависимости распределений данных статистик от вида многомерного закона. В то же время классическими результатами можно пользоваться при проверке гипотез вида $H_0 : r_{ij} = r_0$, при $|r_0| \leq 0.15$.

ГЛАВА 5
ИССЛЕДОВАНИЕ КРИТЕРИЕВ ПРОВЕРКИ ГИПОТЕЗ О КОРРЕЛЯЦИОННОМ ОТНОШЕНИИ

В классическом корреляционном анализе на основании соотношений между парным коэффициентом корреляции и корреляционным отношением можно судить о характере зависимости между компонентами случайного вектора.

5.1. Классические критерии проверки гипотез о корреляционном отношении

Корреляционное отношение случайной величины X_i по X_j определяется отношением дисперсии условного математического ожидания $E[X_i|X_j]$ к дисперсии X_i:

$$\rho_{ij}^2 = \frac{D\left\{E[X_i|X_j]\right\}}{D[X_i]}. \tag{5.1}$$

В отличие от коэффициента корреляции r_{ij} корреляционное отношение ρ_{ij} несимметрично относительно X_i и X_j. Соотношение между коэффициентом корреляции r_{ij} и корреляционным отношением ρ_{ij} в случае многомерного нормального закона позволяет утверждать следующее [58]:

1. $r_{ij}^2 = 0$, если X_i и X_j независимы;
2. $r_{ij}^2 = \rho_{ij}^2 = 1$, тогда и только тогда, когда имеется строгая линейная функциональная зависимость X_i от X_j;
3. $r_{ij}^2 < \rho_{ij}^2 = 1$, тогда и только тогда, когда имеется строгая нелинейная функциональная зависимость X_i от X_j;
4. $r_{ij}^2 = \rho_{ij}^2 < 1$, тогда и только тогда, когда регрессия X_i по X_j строго линейная, но нет функциональной зависимости;
5. $r_{ij}^2 < \rho_{ij}^2 < 1$, указывает на то, что не существует функциональной зависимости, и некоторая нелинейная кривая регрессии «подходит» лучше, чем «наилучшая» прямая линия.

Таким образом, равенство квадрата коэффициента корреляции корреляционному отношению указывает на то, что для регрессии нельзя найти лучшей кривой, чем прямая линия.

Оценка корреляционного отношения определяется выражением

$$\hat{\rho}_{ij}^2 = \frac{\sum\limits_{l=1}^{k} n_l \left(\bar{X}_l^i - \bar{X}^i\right)^2}{\sum\limits_{l=1}^{k}\sum\limits_{s=1}^{n_l} \left(X_{ls}^i - \bar{X}^i\right)^2}, \tag{5.2}$$

где k — количество интервалов сечений для компоненты X_j; $\bar{X}_l^i$ — среднее значение компоненты X_i в l—ом сечении; n_l — число наблюдений компоненты X_i в l—ом сечении; X_{ls}^i — значение компоненты X_i с номером s в l—ом сечении.

Относительно корреляционного отношения могут проверяться два вида гипотез: о равенстве корреляционного отношения нулю $H_0 : \rho_{ij}^2 = 0$ и о равенстве корреляционного отношения квадрату коэффициента корреляции $H_0 : \rho_{ij}^2 = r_{ij}^2$ (критерий линейности регрессии X_i по X_j).

1. В критерии проверки гипотезы $H_0 : \rho_{ij}^2 = 0$ используется статистика

$$F_1 = \frac{n-k}{k-1}\frac{\hat{\rho}_{ij}^2}{1-\hat{\rho}_{ij}^2}, \tag{5.3}$$

которая при справедливой гипотезе H_0 имеет F—распределение Фишера с числом степеней свободы $k-1$ и $n-k$: $G(F_1|H_0) = F_{k-1,n-k}$ [58].

2. При проверке гипотезы $H_0 : \rho_{ij}^2 = r_{ij}^2$ вычисляется статистика

$$F_2 = \frac{n-k}{k-2}\frac{\hat{\rho}_{ij}^2 - \hat{r}_{ij}^2}{1-\hat{\rho}_{ij}^2}, \tag{5.4}$$

которая при справедливой гипотезе H_0 имеет F—распределение Фишера с числом степеней свободы $k-2$ и $n-k$: $G(F_2|H_0) = F_{k-2,n-k}$ [58].

5.2. Влияние различных способов группирования и количества интервалов на оценку корреляционного отношения

Как ранее отмечалось, в данной работе использовались три способа группирования: равноинтервальное (РИГ), равночастотное (РЧГ) и асимптотически

оптимальное (АОГ). Отметим, что в случае равночастотного группирования, если количество случайных величин n не делится на число интервалов k нацело, то остаток распределяется равномерно от центральных до крайних интервалов группирования. Например, для $n = 10$ и $k = 4$ при РЧГ будем иметь следующие частоты попаданий в интервалы группирования $n_1 = 2$, $n_2 = 3$, $n_3 = 3$ и $n_4 = 2$.

Из выражения для оценки корреляционного отношения (5.2) можно увидеть, что увеличение числа интервалов группирования k приводит к росту самой оценки. Это подтверждают рисунки 5.1 и 5.2, где изображены полученные в результате моделирования плотности оценок корреляционного отношения, промоделированные при $\rho_{ij}^2 = 0$ и $\rho_{ij}^2 = 1$ соответственно и вычисленные с использованием РЧГ при различном количестве интервалов группирования k и объеме случайных наблюдений $n = 100$. На данных рисунках видно, что в общем случае с увеличением числа интервалов k растет параметр сдвига у функции плотности оценки. Аналогичная зависимость функций плотности распределения оценок корреляционного отношения от количества интервалов группирования наблюдается и в случае равноинтервального и асимптотически оптимального способов группирования.

Влияние способа группирования на оценку корреляционного отношения отражено на рисунке 5.3, где моделирование проводилось при $\rho_{ij}^2 = 0$ и объеме псевдослучайных величин $n = 100$. Для числа интервалов $k = 5$ функции плотности оценок $\hat{\rho}_{ij}^2$, вычисленных при различных способах группирования, совпадают. А с увеличением числа интервалов наблюдается расхождение функций плотности оценок для разных способов группирования. Например, при объеме $n = 100$, начиная с $k = 10$ плотность распределения оценки (5.2), вычисленная при асимптотически оптимальном группировании, смещается влево относительно функций плотности, вычисленных с использованием РИГ или РЧГ. Различие в распределениях оценок при РЧГ и РИГ наблюдается при $n = 100$, когда $k \geq 20$.

Так как моделирование оценок корреляционного отношения осуществлялось при $\rho_{ij}^2 = 0$, то казалось бы предпочтительней выбрать тот способ груп-

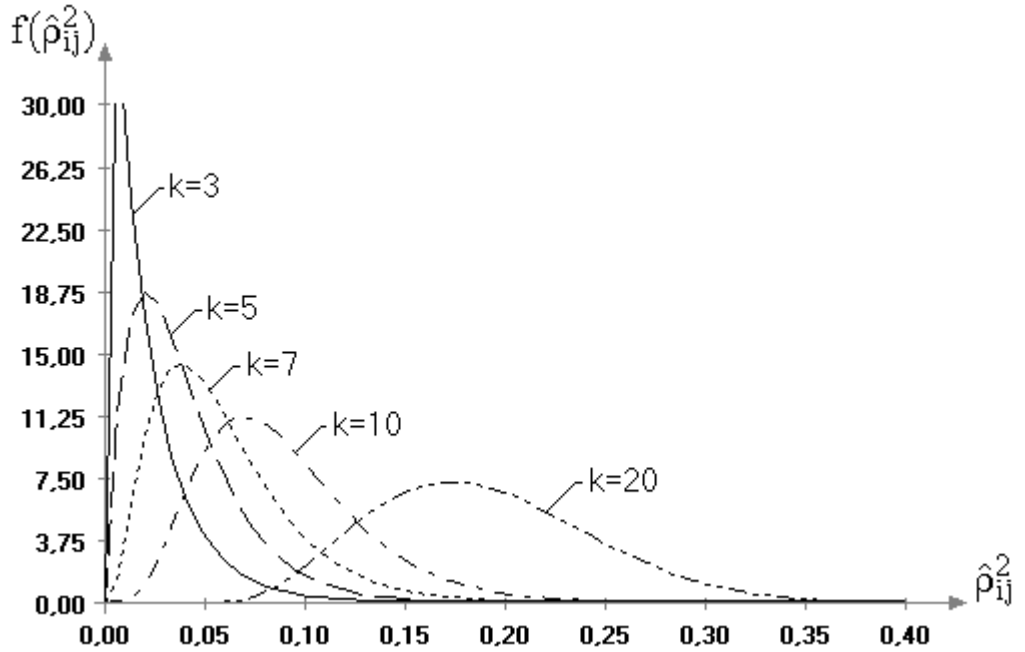


Рис. 5.1. Функции плотности распределения оценок корреляционного отношения, моделируемых при $\rho_{ij}^2 = 0$, в случае использования РЧГ для различного количества интервалов группирования k, $n = 100$

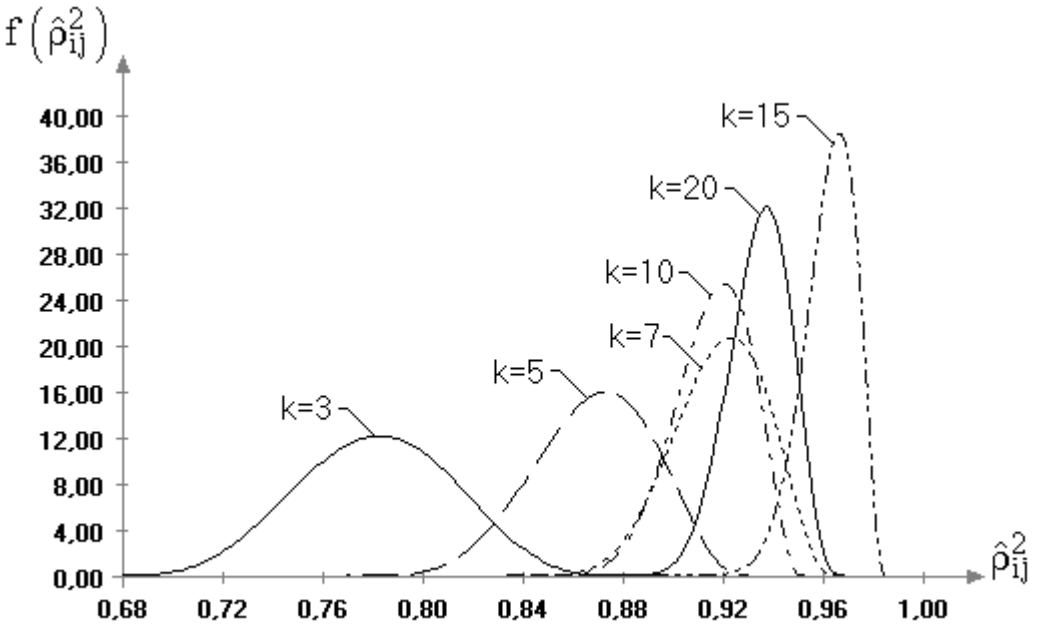


Рис. 5.2. Функции плотности распределения оценок корреляционного отношения, моделируемых при $\rho_{ij}^2 = 1$, в случае использования РЧГ для различного количества интервалов группирования k, $n = 100$

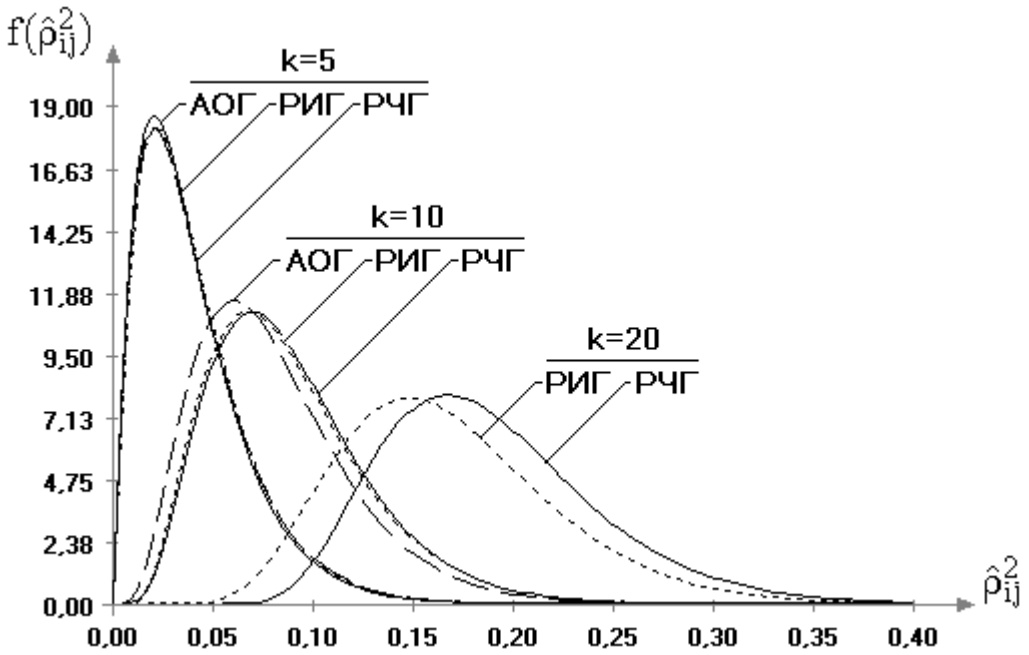


Рис. 5.3. Функции плотности распределения оценок корреляционного отношения, моделируемых при $\rho_{ij}^2 = 0$, где использовалось АОГ, РИГ, РЧГ для различного числа интервалов группирования k, $n = 100$

пирования, плотность оценок которого лежит левее. С другой стороны, рост числа интервалов группирования для АОГ и РИГ приводит к тому, что будут появляться интервалы, для которых число наблюдений n_l будет равно нулю. Для АОГ это крайние интервалы, а для РИГ — интервалы, находящиеся между «удаленными» наблюдениями и основной группой. Наличие интервалов с нулевыми частотами попадания приводит к искусственному занижению величины оценки корреляционного отношения $\hat{\rho}_{ij}^2$. Использование равночастотного группирования позволяет избежать таких ошибок.

Несколько сложнее выглядит ситуация когда рассматриваются оценки $\hat{\rho}_{ij}^2$, моделируемые для случая $\rho_{ij}^2 = 1$. На рисунке 5.4 изображены плотности оценок корреляционного отношения при различных способах группирования, моделируемых при $\rho_{ij}^2 = 1$ и объеме случайных величин $n = 100$. При малом числе интервалов группирования k относительно объема выборки n на данном рисунке РИГ выглядит предпочтительней, так как плотность оценок $\hat{\rho}_{ij}^2$, построенная с использованием РИГ, при равном числе интервалов располо-

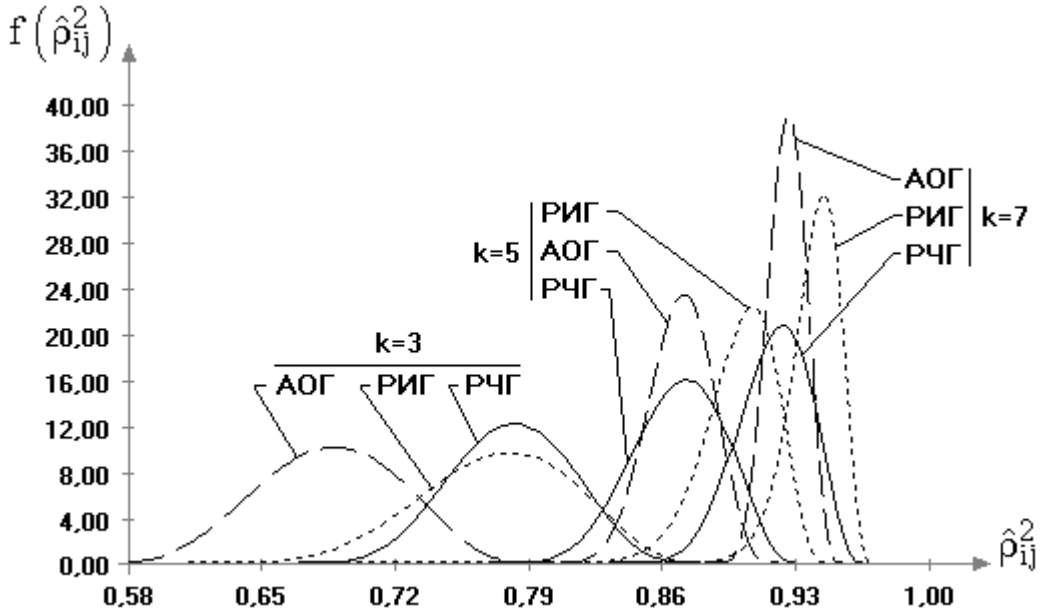


Рис. 5.4. Функции плотности распределения оценок корреляционного отношения, моделируемых при $\rho_{ij}^2 = 1$, где использовалось АОГ, РИГ, РЧГ для различного числа интервалов группирования k, $n = 100$

жена правее, чем плотности оценок для АОГ и РЧГ. Но, во—первых, оценки корреляционного отношения $\hat{\rho}_{ij}^2$, построенные с использованием РИГ, сильно зависят от крайних граничных точек интервалов группирования X_j^{min} и X_j^{max}, так как эти точки определяют длину интервалов. А во—вторых, как и в случае $\rho_{ij}^2 = 0$, неоправданное завышение числа интервалов при равноинтервальном группировании приводит к ухудшению свойств оценок корреляционного отношения. Существенное изменение функций плотности оценок $\hat{\rho}_{ij}^2$ показано на рисунке 5.5. Причина ухудшения свойств оценок $\hat{\rho}_{ij}^2$ есть описанное ранее обнуление частот попаданий n_l для нескольких интервалов группирования, которое вновь приводит к искусственному занижению величины оценки корреляционного отношения. Асимптотически оптимальное группирование при неправильном выборе количества интервалов, что и в случае РИГ, также приводит к искаженным функциям распределения.

Поэтому для асимптотически оптимального и равноинтервального группирования можно определить «критические» значения числа интервалов, начиная с которых появляются нулевые частоты попадания n_l, и, как следствие,

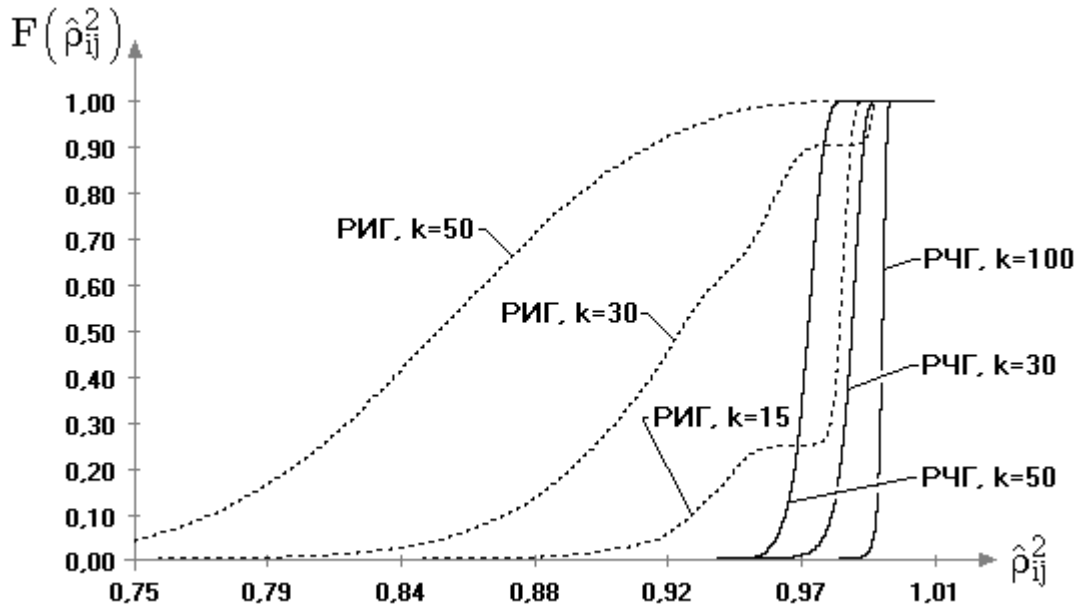


Рис. 5.5. Распределения оценок корреляционного отношения, моделируемых при $\rho_{ij}^2 = 1$, где использовалось РИГ, РЧГ для различного числа интервалов группирования k, $n = 250$

происходит ухудшение свойств оценок корреляционного отношения.

Увеличение объемов выборок случайных величин не изменяет выявленной закономерности по влиянию способов и количества интервалов группирования на распределения оценок корреляционного отношения. С ростом объемов происходит естественное увеличение значений для «критических» чисел интервалов группирования, начиная с которых наблюдается оговоренное ухудшение свойств оценок корреляционного отношения при использовании АОГ или РИГ.

Для вычислений оценок корреляционного отношения можно рекомендовать использовать равночастотное группирование, так как в данном случае свойства вычисляемых оценок меньше зависят от числа интервалов группирования. Если по каким—либо причинам было принято решение о применении АОГ или РИГ, тогда прежде всего требуется убедиться, что при разбиении на интервалы отсутствуют нулевые частоты попаданий n_l, в противном случае надо уменьшить число интервалов.

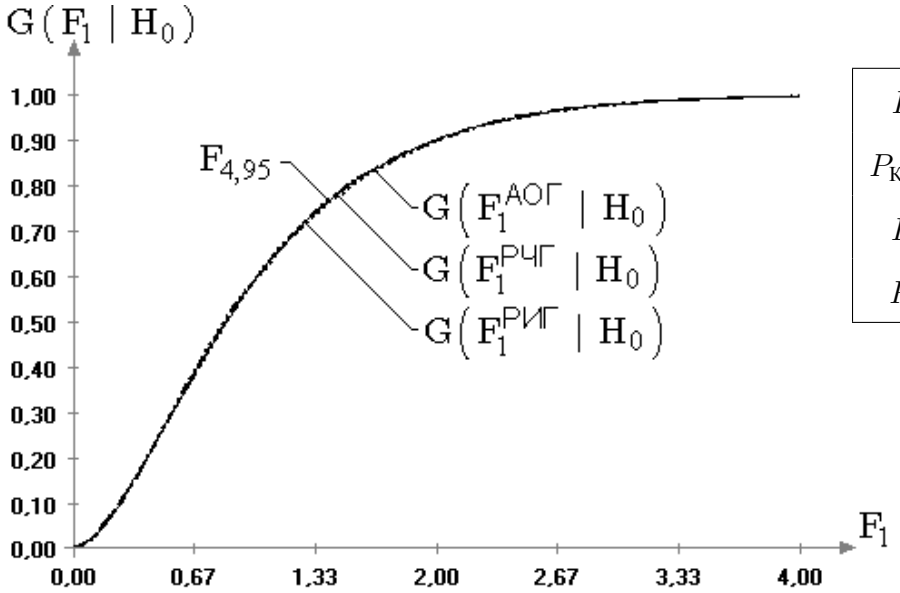
5.3. Исследование распределений статистики критерия проверки гипотезы о незначимости корреляционного отношения

В первую очередь с помощью методов статистического моделирования исследовались распределения статистик, используемых при проверке гипотез о корреляционном отношении, при условии, что наблюдения принадлежат многомерному нормальному закону.

Исследование распределения статистики критерия проверки гипотез о равенстве корреляционного отношения нулевому значению показало, что если осуществляется корректный выбор количества интервалов группирования k, то соответствующее теоретическое предельное F—распределение с $k-1$ и $n-k$ числом степеней свободы хорошо описывает эмпирическое распределение статистики F_1.

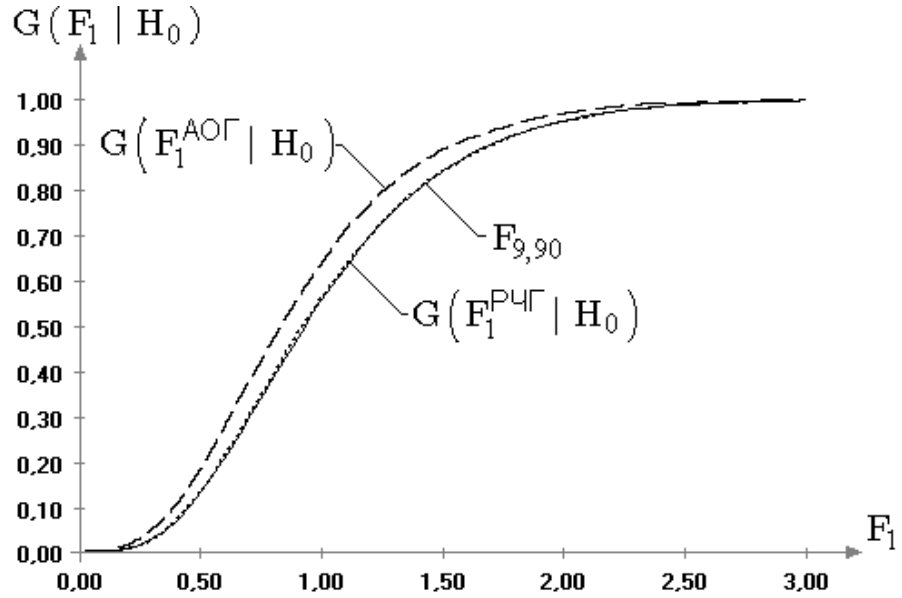
Например, на рисунке 5.6 представлены полученные в результате моделирования эмпирические распределения статистики F_1 (5.3), построенные с использованием АОГ, РЧГ и РИГ, а также соответствующее предельное $F_{k-1,n-k}$ — распределение при проверке гипотезы $H_0 : \rho_{ij}^2 = 0$ для числа интервалов $k = 5$ и объема выборки $n = 100$. Рисунок дополнен таблицей, где отражены результаты проверки согласия эмпирического распределения с теоретическим предельным по критериям согласия. Приведенные уровни значимости по критериям согласия свидетельствуют о том, что статистика F_1 действительно хорошо описывается соответствующим предельным распределением, и на данное согласие существенно не влияет выбор способа группирования при правильном выборе k.

Пример некорректного выбора числа интервалов для асимптотически оптимального группирования приведен на рисунке 5.7. Где явно видно изменение предельного закона распределения статистики F_1 при АОГ, в то время когда использование равночастотного группирования дает по—прежнему высокие значения для достигаемых уровней значимости. Превышение «критических» значений для числа интервалов k приводит к изменению предельного распределения статистики F_1 и в случае применения равноинтервального группиро-



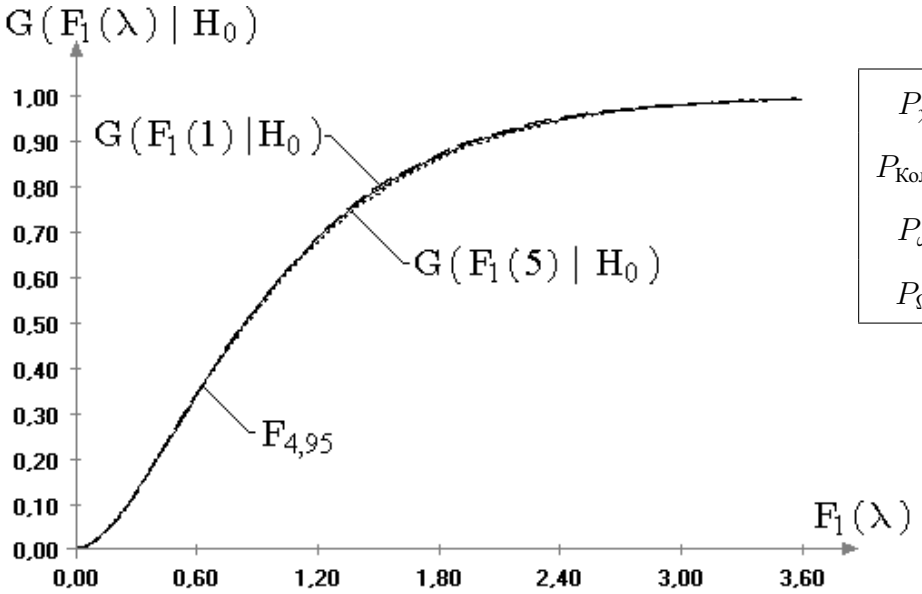
	АОГ	РЧГ	РИГ
$P_{\chi^2} =$	0.75	0.33	0.79
$P_{\text{Колм}} =$	0.73	0.74	0.94
$P_{\omega^2} =$	0.73	0.58	0.96
$P_{\Omega^2} =$	0.77	0.69	0.86

Рис. 5.6. Теоретическая и эмпирические функции распределения статистики F_1 (5.3) при проверке гипотезы $H_0 : \rho_{ij}^2 = 0$, построенные с использованием различных способов группирования: $k = 5$, $n = 100$



	АОГ	РЧГ
$P_{\chi^2} =$	0.00	0.88
$P_{\text{Колм}} =$	0.00	0.42
$P_{\omega^2} =$	0.00	0.33
$P_{\Omega^2} =$	0.00	0.43

Рис. 5.7. Теоретическая и эмпирические функции распределения статистики F_1 (5.3) при проверке гипотезы $H_0 : \rho_{ij}^2 = 0$, построенные с использованием АОГ и РЧГ: $k = 10$, $n = 100$



	$\lambda = 1$	$\lambda = 5$
$P_{\chi^2} =$	0.68	0.55
$P_{\text{Колм}} =$	0.64	0.92
$P_{\omega^2} =$	0.45	0.83
$P_{\Omega^2} =$	0.50	0.75

Рис. 5.8. Эмпирические функции распределения статистик $F_1(1)$, $F_1(5)$ и классическое предельное F—распределение при проверке гипотезы $H_0 : \rho_{ij}^2 = 0$: РЧГ, $k = 5$, $n = 100$

вания.

Исследование распределений статистики, используемой при проверке гипотезы вида $H_0 : \rho_{ij}^2 = 0$, проводилось для законов многомерных величин, моделируемых на основе предложенной в данной работе процедуры при различных способах группирования.

Из результатов, приведенных на рисунке 5.8, следует, что нет оснований для отклонения предположения о том, что предельным распределением статистики критерия проверки гипотезы о равенстве корреляционного отношения нулевому значению в случае многомерных законов, построенных по семейству распределений (6.4) с разными параметрами формы λ, является классическое предельное F—распределение Фишера с числом степеней свободы $k - 1$ и $n - k$.

Исследование влияния способа группирования на распределение статистики F_1 при многомерных законах, отличных от нормального, показало еще большую зависимость оценок корреляционного отношения от числа интервалов k при использовании асимптотически оптимального и равноинтервального группирования. В таблице 5.1 приведены значения достигаемых уровней значимо-

Таблица 5.1

Значения достигнутых уровней значимости по критериям согласия для распределений статистики F_1, смоделированных при различных параметрах формы λ: $k = 5$ и $n = 100$

	$\lambda = 1$			$\lambda = 5$		
	АОГ	РИГ	РЧГ	АОГ	РИГ	РЧГ
$P_{\chi^2} =$	0.02	0.00	0.50	0.28	0.52	0.44
$P_{\text{Колм}} =$	0.17	0.00	0.27	0.52	0.30	0.34
$P_{\omega^2} =$	0.10	0.00	0.26	0.57	0.18	0.21
$P_{\Omega^2} =$	0.04	0.00	0.28	0.49	0.19	0.26
$P_{\text{сред}} =$	0.08	0.00	0.33	0.46	0.30	0.31

сти при проверке согласия между эмпирическим распределением статистики F_1 и соответствующим классическим предельным $F_{k-1,n-k}$—распределением. Эксперименты показали, что в случае более островершинных многомерных законах для методов АОГ и РИГ желательно еще большее уменьшение числа интервалов группирования по сравнению с нормальным законом, а для более плосковершинных законов — наоборот, допустимо увеличение количества интервалов.

Для многомерных законов, моделируемых по семейству распределений (6.4), использование равночастотного группирования не ухудшает согласия между эмпирическим распределением статистики F_1 и соответствующим классическим предельным при любом выборе числа интервалов как при $\lambda < 2$, так и $\lambda > 2$. По–прежнему, разбиение допустимой области на интервалы с равными частотами попадания n_l видится более предпочтительным.

Таким образом, результаты исследования распределений статистики F_1 показали, что в случае многомерных законов, достаточно существенно отличающихся от нормального (более островершинных или более плосковершинных, и даже в случае многомерного закона, построенного по несимметричному одномерному распределению), значимого изменения предельного распределения

статистики F_1 не происходит.

Это позволяет утверждать, что статистические выводы, опирающиеся на классический аппарат, в задачах с применением критерия проверки гипотезы вида $H_0 : \rho_{ij}^2 = 0$ будут оставаться корректными и при нарушении предположений о нормальности наблюдаемого многомерного закона.

5.4. Исследование распределений статистики критерия линейности регрессии X_i по X_j

Указанные в начале данной главы соотношения $\rho_{ij}^2 \geq r_{ij}^2$ между теоретическими корреляционным отношением ρ_{ij}^2 и парным коэффициентом корреляции r_{ij} не всегда выполняются для их оценок, особенно, если связь (регрессионная или функциональная) линейная. Такое возможно, если ρ_{ij}^2 и r_{ij}^2 близки [103]. Нарушение условия происходит из–за вычислительных погрешностей, связанных с ограниченностью представления чисел в ЭВМ, случайностью оценок $\hat{\rho}_{ij}^2$ и $\hat{r}_{ij}^2$ и сильным влиянием на $\hat{\rho}_{ij}^2$ числа интервалов и способов группирования. Известно, что величина $\rho_{ij}^2 - r_{ij}^2 > 0$ является индикатором нелинейности [58]. Однако, как уже говорилось, величина $\hat{\rho}_{ij}^2 - \hat{r}_{ij}^2$ вследствие случайности оценок может оказаться отрицательной, хотя абсолютная величина разности, как правило, мала.

Возможность нарушения неравенства $\rho_{ij}^2 > r_{ij}^2$ для соответствующих оценок наглядно иллюстрирует рисунок 5.9, где представлены функции плотности квадрата оценки парного коэффициента корреляции $\hat{r}_{ij}^2$ и плотности оценок корреляционного отношения $\hat{\rho}_{ij}^2$, построенные для случая линейной зависимости X_i от X_j ($r_{ij}^2 = \rho_{ij}^2 = 1$). При вычислении оценок корреляционного отношения использовались интервалы равной частоты при объемах выборок случайных величин $n = 100$. На приведенном рисунке видно, что для объема $n = 100$ с ростом числа интервалов группирования вероятность появления значений $\hat{\rho}_{ij}^2 - \hat{r}_{ij}^2 < 0$ падает (плотности оценок «расходятся» дальше друг от друга), но остается положительной.

При увеличении объемов выборок n уменьшается дисперсия распределе-

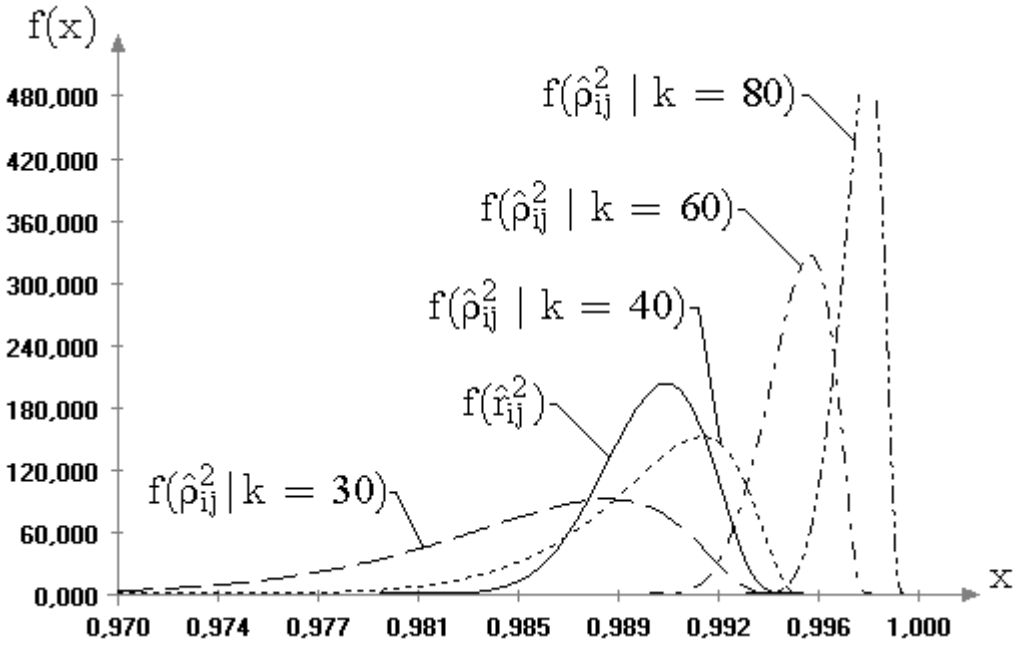


Рис. 5.9. Функции плотности распределения оценок корреляционного отношения $\hat{\rho}^2_{ij}$ и квадрата парного коэффициента корреляции $\hat{r}^2_{ij}$, моделируемых при линейной зависимости X_i от X_j: РЧГ, $n = 100$

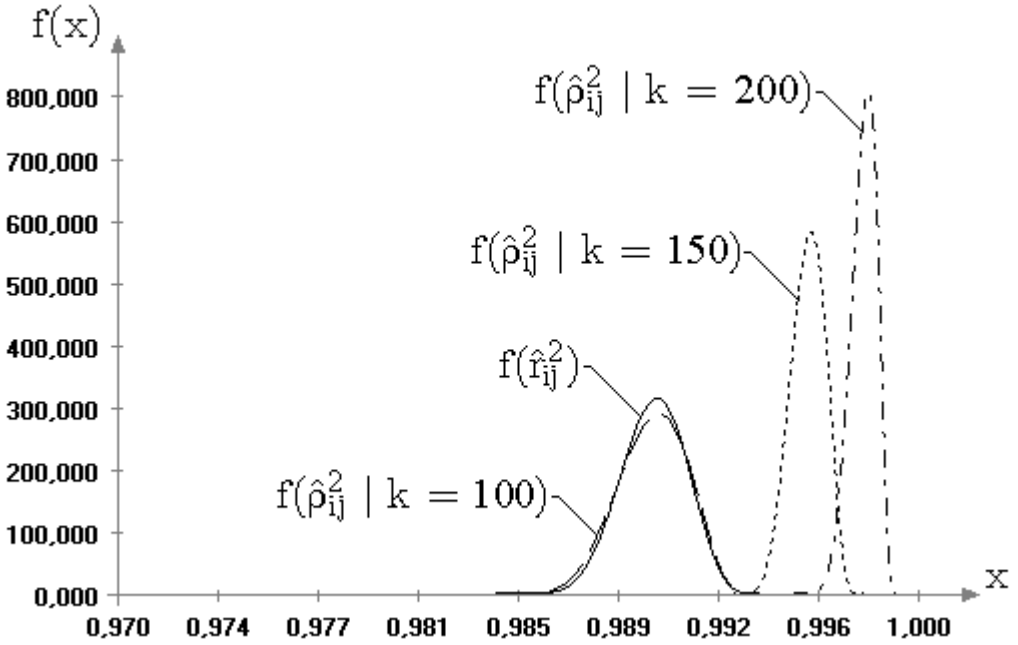


Рис. 5.10. Функции плотности распределения оценок корреляционного отношения $\hat{\rho}^2_{ij}$ и квадрата парного коэффициента корреляции $\hat{r}^2_{ij}$, моделируемых при линейной зависимости X_i от X_j: РЧГ, $n = 250$

ния оценки парного коэффициента корреляции. Поэтому для больших значений n и k вероятность появления значений $\hat{\rho}_{ij}^2 - \hat{r}_{ij}^2 < 0$ оказывается практически близкой к нулю. На рисунке 5.10 отображены плотности оценок $\hat{\rho}_{ij}^2$ и $\hat{r}_{ij}^2$, вид которых позволяет утверждать, что при объеме выборки $n = 250$ и числе интервалов группирования $k = 200$ при использовании РЧГ неравенство $\rho_{ij}^2 > r_{ij}^2$ с вероятностью 1 выполняется и для их оценок.

Однако и при значениях $n = 250$ и $k = 200$ распределение статистики F_2 даже в случае многомерного нормального закона не подчиняется F—распределению Фишера с числом степеней свободы $k - 2$ и $n - k$ (см. рис. 5.11). Дальнейшее увеличение объемов выборок и числа интервалов группирования существенно не улучшает согласия между распределением данной статистики и соответствующим предельным распределением.

С другой стороны, проведенные исследования не опровергают, что распределение статистики F_2 подчиняется $F_{k-2,n-k}$—распределению в пределе $n \to \infty$. При обработке реальных данных, когда вычисленное значение статистики оказывается $F_2 < 0$, можно рекомендовать рассмотреть значения оценок $\hat{\rho}_{ij}^2$ и $\hat{r}_{ij}^2$. И если они близки к единице можно выдвинуть предположение о линейной зависимости.

В случае многомерного закона, отличного от нормального, есть основания утверждать, что ни для конечных объемов выборок, ни при $n \to \infty$ распределение статистики F_2 не будет описываться $F_{k-2,n-k}$—распределением. Это следует, во–первых, из различия эмпирических распределений статистики F_2 для многомерного нормального закона и законов, моделируемых на основе семейства распределений (6.4) с параметрами формы $\lambda = 1$ и $\lambda = 5$ (см. рис. 5.12). Во-вторых, из показанной ранее неустойчивости критерия проверки гипотез о парном коэффициенте корреляции вида $H_0 : r_{ij} = r_0$, при $|r_0| > 0.15$ к отклонению от нормальности.

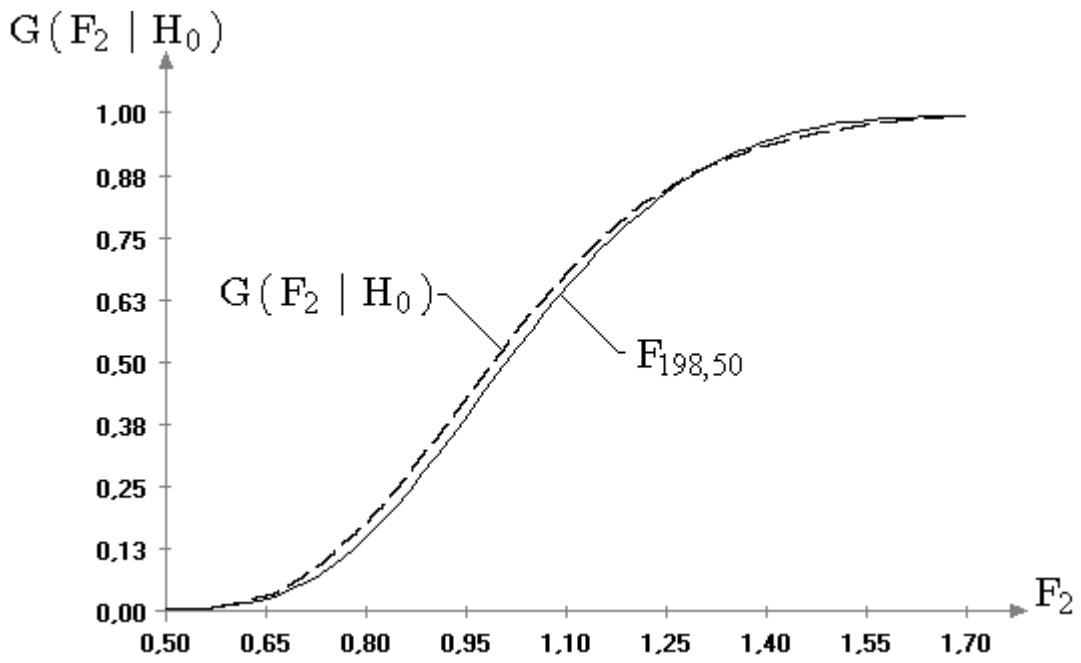


Рис. 5.11. Теоретическая и эмпирическая функции распределения статистики F_2 (5.4) при проверке гипотезы $H_0 : \rho_{ij}^2 = 1$: РЧГ, $k = 200$, $n = 250$

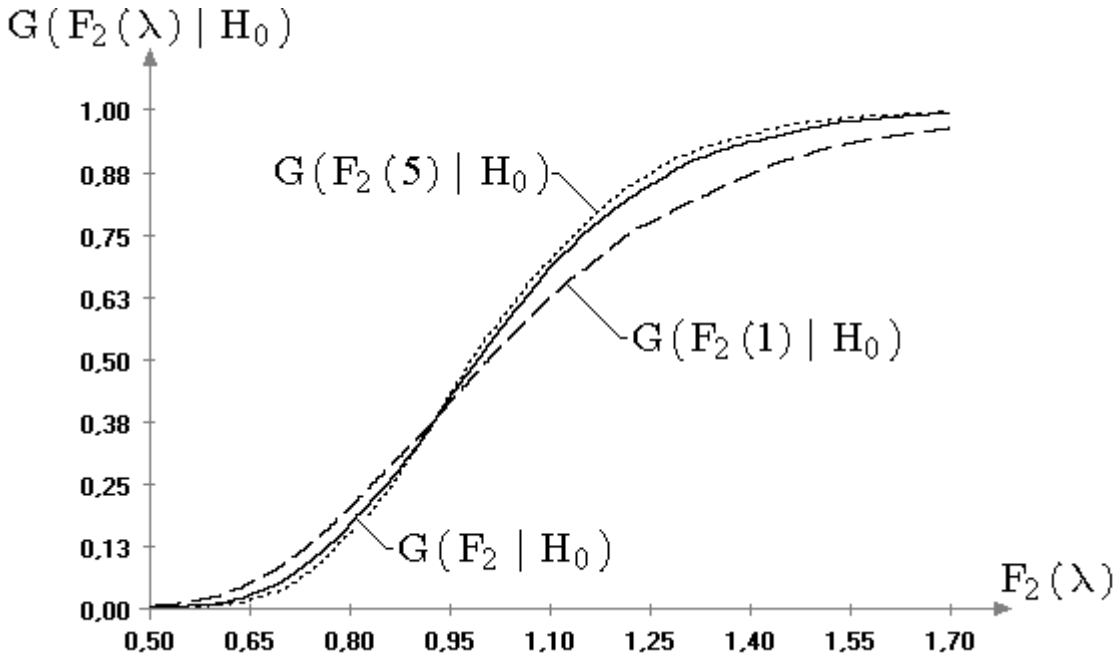


Рис. 5.12. Эмпирические функции распределения статистик F_2, $F_2(1)$ и $F_2(5)$ при проверке гипотезы $H_0 : \rho_{ij}^2 = 1$: РЧГ, $k = 200$, $n = 250$

5.5. Выводы

Исследование влияния способов группирования и количества интервалов на оценку корреляционного отношения показало, что оценка корреляционного отношения, прежде всего, сильно зависит от количества интервалов группирования. Как правило, уменьшение количества интервалов группирования приводит к уменьшению значений оценок корреляционного отношения, в то время как увеличение сопровождается ростом величины $\hat{\rho}_{ij}^2$. При использовании асимптотически оптимального и равноинтервального группирования необходимо корректно выбирать число интервалов, избегая нулевых частот попадания n_l в интервалы, приводящих к ухудшению свойств оценок корреляционного отношения. Разбиение области определения на интервалы равной частоты показало себя как наиболее предпочтительное для вычисления оценок $\hat{\rho}_{ij}^2$.

Исследования распределения статистики, используемой в критерии проверки гипотезы вида $H_0 : \rho_{ij}^2 = 0$, при псевдослучайных величинах, подчиняющихся многомерному нормальному закону, показали, что оно хорошо согласуется с теоретическим предельным распределением, полученными в классическом корреляционном анализе. В случае многомерных законов, отличающихся от нормального в достаточно широких пределах (более островершинных или более плосковершинных), изменения предельного распределения статистики F_1 не происходит. Эмпирическое распределение данной статистики по—прежнему хорошо описывается предельными законами, полученными в предположении о нормальности наблюдаемого вектора.

Полное исследование распределения статистики критерия, используемого при проверке гипотезы вида $H_0 : \rho_{ij}^2 = r_{ij}^2$, на данный момент затруднено вследствие указанных вычислительных проблем, суть которых заключается в том, что при линейной связи соотношение для теоретических величин $\rho_{ij}^2 > r_{ij}^2$ может не выполняться для их оценок.

ГЛАВА 6
ОПИСАНИЕ ПРОГРАММНОЙ СИСТЕМЫ

6.1. Общая характеристика программной системы

Методика компьютерного моделирования и анализа статистических закономерностей предполагает разработку соответствующего программного обеспечения для проведения исследований. Программная система предназначена для осуществления проверки рассматриваемых гипотез многомерного анализа, исследования распределений статистик критериев, вычисления оценок параметров многомерных законов, моделирования выборок различных одномерных и многомерных законов распределения. Разработанное программное обеспечение является продолжением и расширением основной идеи, заложенной еще в программной системе «Корреляционный анализ многомерных случайных величин» [65].

Изначально программная система разрабатывалась как функциональное расширение исследовательского программного пакета «Интервальная статистика (ISW)», разработанного Лемешко Б. Ю. и Постоваловым С. Н. Но в процессе реализации была оформлена как самостоятельная система. При этом использование совместимого формата данных позволило провести исследование распределений статистик, вычисляемых в критериях многомерного анализа, при помощи системы «Интервальной статистики (ISW)», хорошо зарекомендовавшей себя в задачах такого рода [111, 112].

Программная система позволяет решать следующие задачи:

- моделирование выборок псевдослучайных величин, подчиненных заданному закону распределения;
- моделирование выборок псевдослучайных векторов по методу, предложенному в диссертационной работе;
- моделирование распределений статистик, используемых при проверке гипотез о математическом ожидании и дисперсии;

– моделирование распределений статистик рассматриваемых критериев многомерного анализа;
– осуществлять проверку различных гипотез при помощи критериев многомерного анализа;
– строить оценки вектора математических ожиданий, ковариационной матрицы, парных, частных и множественных коэффициентов корреляции, корреляционного отношения.

Независимость ряда решаемых задач позволила спроектировать программную систему в виде совокупности самостоятельных блоков, что существенно упростило процесс разработки. Выбранный подход к реализации данных блоков позволяет легко использовать их функциональность в других программных системах. Например, блок моделирования псевдослучайных величин был реализован в виде подключаемой библиотеки.

При реализации были выделены следующие основные блоки.
– Блок моделирования одномерных и многомерных случайных величин, подчиняющихся различным законам распределения.
– Блок проверки гипотез.
– Процедуры вычисления оценок.
– Блок моделирования распределений статистик, используемых при проверке гипотез рассматриваемых критериев.

Код программный системы написан на языке C++ [105] в среде быстрой разработки приложений Borland C++ Builder 6.0 [34] с поддержкой объектно–ориентированного подхода и откомпилирован под 32–разрядные операционные системы семейства Microsoft Windows. Чтобы избежать возможных ошибок реализации математических соотношений, для нескольких алгоритмов были написаны дублирующие программы в среде математического программирования Maple [46, 96].

6.2. Краткое описание интерфейса программной системы

Программная система состоит из двух программ. Основная программа, которая носит название «Корреляционный анализ», позволяет решать и иссле-

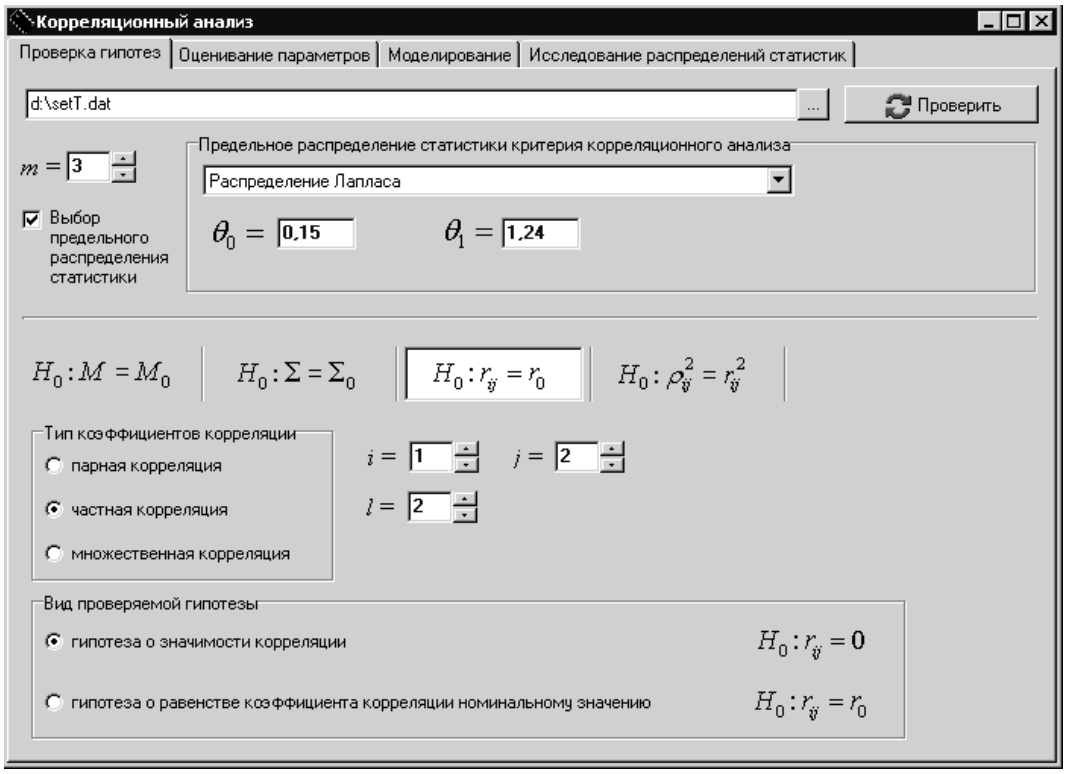


Рис. 6.1. Диалоговое окно «Проверка гипотез о коэффициентах корреляции»

довать задачи многомерного анализа. Вспомогательная программа позволяет моделировать распределения статистик, используемых при проверке гипотез о математическом ожидании и дисперсии в одномерном случае.

6.2.1. Основная программа

Первая закладка «Проверка гипотез» на главном диалоговом окне позволяет выбирать вид проверяемой гипотезы: гипотезу о равенстве вектора математических ожиданий заданному вектору $\left(H_0 : \bar{M} = \bar{M}_0\right)$; гипотезу о равенстве ковариационной матрицы заданной матрице $(H_0 : \Sigma = \Sigma_0)$; гипотезу о значении парного, частного и множественного коэффициентов корреляции $(H_0 : r_{ij} = r_0)$; гипотезу о корреляционном отношении $\left(H_0 : \rho_{ij}^2 = r_{ij}^2\right)$. Общими параметрами при проверке гипотез являются размерность, имя файла с выборкой случайных векторов и распределение статистики критерия проверяемой гипотезы. В зависимости от выбранного типа гипотезы может потребоваться задание дополнительных параметров. Например, для корреляционного отношения это способ группирования, количество интервалов группирования

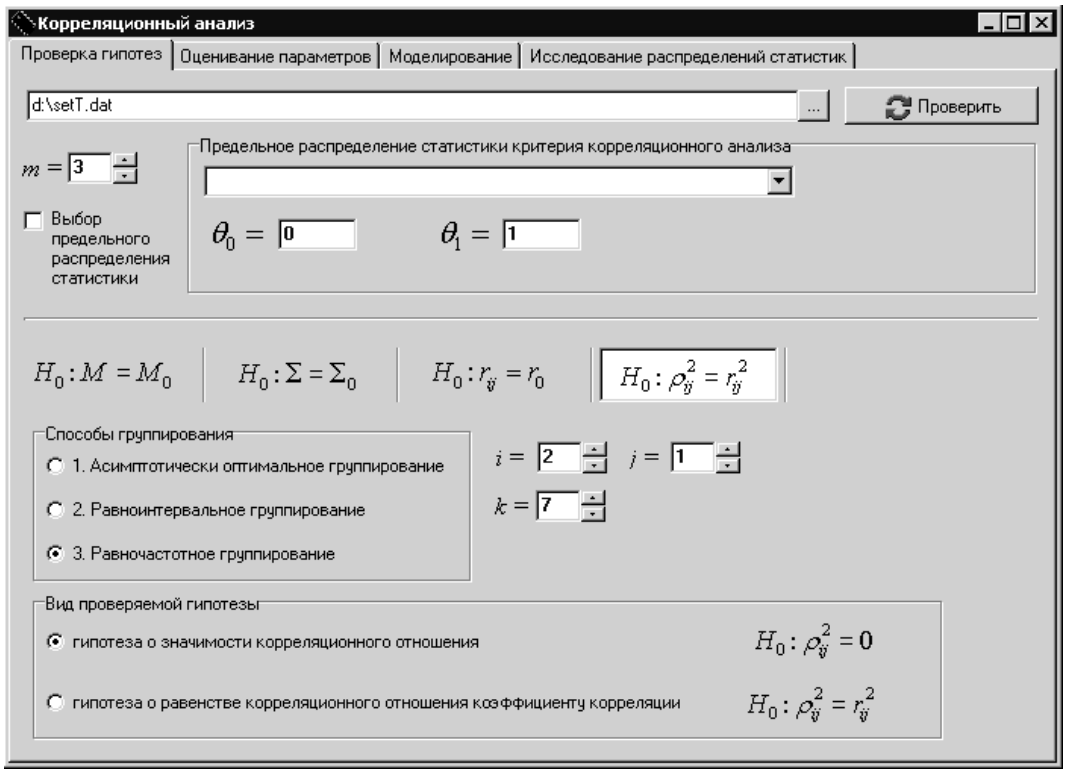


Рис. 6.2. Диалоговое окно «Проверка гипотез о корреляционном отношении»

и сам вид проверяемой гипотезы (см. рисунки 6.1 и 6.2).

Выбор и вычисление оценок рассматриваемых параметров по выборке случайных векторов можно осуществить через закладку «Оценивание параметров». Полученные результаты оформляются в виде HTML отчета средствами специально разработанной библиотеки. Изменение или доработка программного кода данной библиотеки позволяет легко добиться улучшения вида получаемого отчета без вмешательства в код основной программы.

Доступ к процедурам моделирования, описанным в разделе 6.3., осуществляется через одноименную закладку основного диалогового окна (рис. 6.3).

Закладка «Исследование распределений статистик» не содержит множества задаваемых параметров, кроме имени файла для выгрузки выборки значений статистики. При моделировании выборки используются установленные параметры на предыдущих закладках. В этом случае закладка «Моделирование» определяет закон распределения генерируемого псевдослучайного вектора, а «Проверка гипотез» — статистику критерия, используемую при проверке выбранной гипотезы.

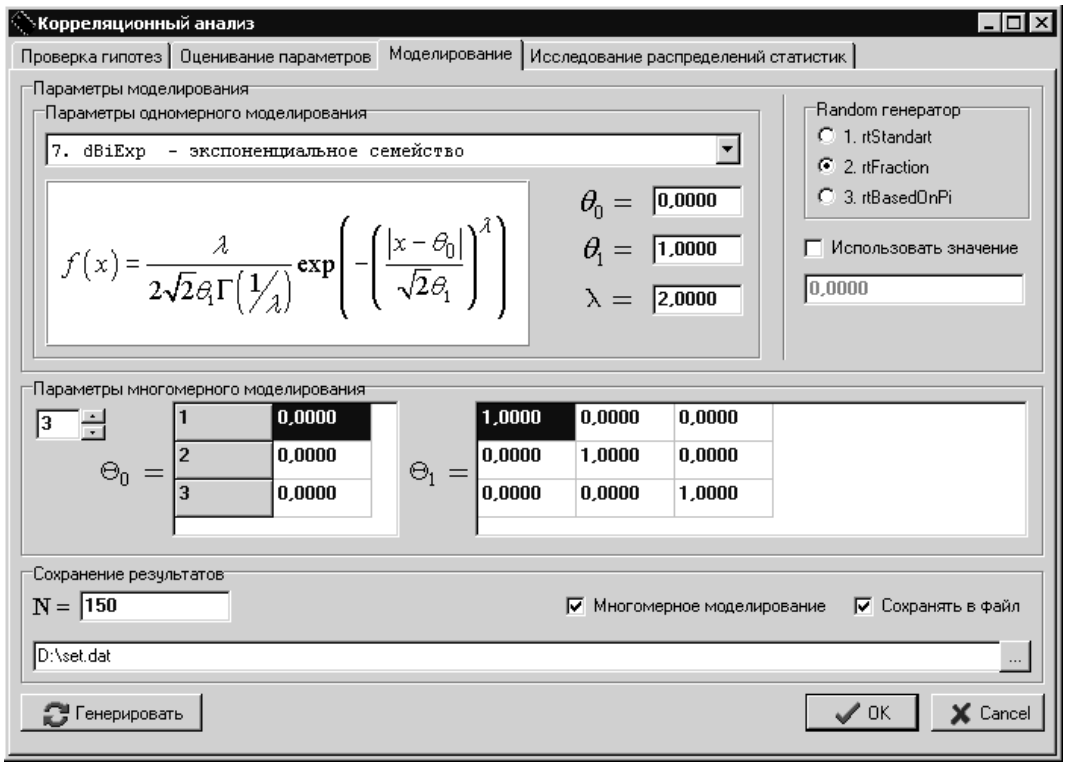


Рис. 6.3. Диалоговое окно «Моделирование»

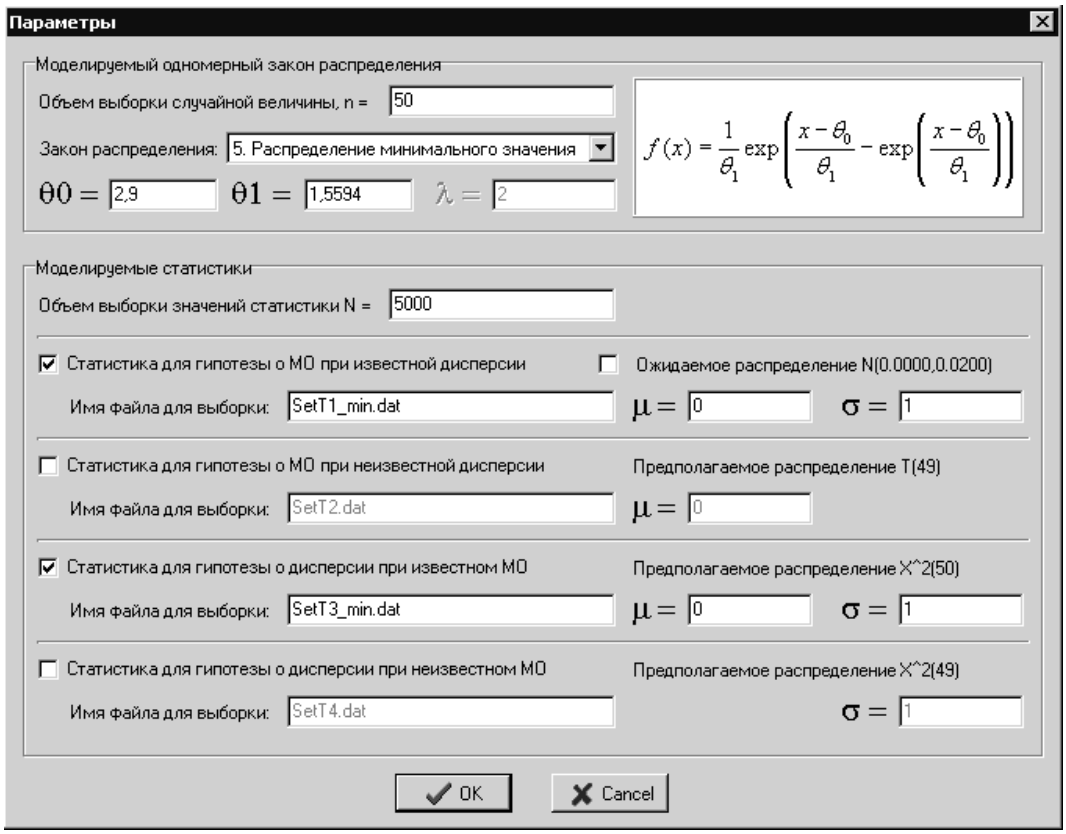


Рис. 6.4. Программа для исследования распределений статистик в одномерном случае

6.2.2. Вспомогательная программа

Для исследования одномерного случая, когда проверяются гипотезы о математическом ожидании и дисперсии, написана вспомогательная программа. Она позволяется моделировать распределения статистик, используемых при проверке данных гипотез (см. рисунок 6.4).

6.3. Моделирование псевдослучайных величин

Для проведения исследований по теме диссертационной работы ключевым блоком программной системы является блок моделирования. Средствами программной системы можно производить моделирование одномерных и многомерных псевдослучайных величин.

При построении любой системы статистического моделирования центральным элементом является датчик, генерирующий псевдослучайные числа по равномерному закону. Проверка качества такого датчика является непременным условием его использования. Важно, не только то, чтобы получаемые последовательности при любых объемах выборок хорошо соответствовали равномерному закону, но и то, чтобы они удовлетворяли целям исследований [50, 92]. Всегда хорошей дополнительной проверкой качества датчиков может являться построение в результате моделирования той статистической закономерности, которая является известным достоянием теории. Хорошее совпадение результатов моделирования с теоретическими является косвенным подтверждением качества используемого датчика.

В программную систему включены следующие алгоритмы имитации псевдослучайной величины, равномерно распределенной на отрезке (0, 1): встроенный датчик систем программирования C++ и мультипликативный датчик [6, 15, 51, 52, 110]. Оба датчика удовлетворяют требованиям, позволяющим использовать их в целях исследования статистических закономерностей.

Исследование датчиков проведено в работе [92], где было отмечено, что выбранные подходы к имитации псевдослучайной величины позволяют получать последовательности, достаточно хорошо подчиняющиеся равномерному

закону при различных объемах выборок. Они удовлетворяют требованиям, позволяющим использовать их в целях исследования статистических закономерностей. Датчик в системах программирования C++ обладает приемлемыми свойствами равномерности, но имеет один недостаток, который следует иметь в виду: в генерируемых выборках, начиная с объемов, примерно, в 1700–1800 наблюдений, начинают появляться повторные значения (этот недостаток исчезает при использовании вычислений с двойной точностью). Реализация мультипликативного датчика такого недостатка не имеет [52]. Поэтому в диссертационной работе при проведении исследований использовался мультипликативный алгоритм, так как для моделирования выборок значений статистик критериев требовались достаточно большие объемы выборок псевдослучайных величин, равномерно распределенных на отрезке (0, 1) .

В программной системе для реализации алгоритмов моделирования использовался объектно–ориентированный подход. Преимуществом такого построения программного кода является то, что при необходимости программная система может быть легко расширена любыми законами распределения. И тогда можно исследовать распределения статистик соответствующих критериев для добавленных одномерных и многомерных законов.

6.3.1. Моделирование одномерных распределений

Основные алгоритмы для имитации одномерных выборочных значений были взяты из [40,51,52,66], где наиболее часто используемым и общим методом формирования псевдослучайных величин является метод обратных функций. В этом методе случайная величина X, подчиняющаяся закону с функцией распределения $F(x)$, получается в соответствии с соотношением $X = F^{-1}(Y)$, где $F^{-1}(\cdot)$ — функция, обратная к $F(\cdot)$, а Y — случайная величина, равномерно распределенная на интервале (0, 1).

Введем обозначения аналогично [66]:

Y — случайные величины, равномерно распределенные на интервале (0, 1);

Z — случайные величины, распределенные по стандартному нормальному закону с параметрами (0, 1);

θ_0 — параметр сдвига;

θ_1 — параметр масштаба;

$E[x]$ — математическое ожидание случайной величины x;

$D[x]$ — дисперсия случайной величины x.

Тогда согласно [40, 41, 52]:

1. Пара псевдослучайных чисел, распределенных по стандартному нормальному закону с параметрами (0, 1), генерируется по формулам

$$\begin{aligned} Z_1 &= \sqrt{-2\ln Y_1}\sin(2\pi Y_2), \\ Z_2 &= \sqrt{-2\ln Y_1}\cos(2\pi Y_2), \end{aligned} \tag{6.1}$$

а нормальное распределение с математическим ожиданием $E[x] = \theta_0$ и дисперсией $D[x] = \theta_1^2$

$$f(x;\theta_0,\theta_1) = \frac{1}{\theta_1\sqrt{2\pi}}\exp\left(-\frac{(x-\theta_0)^2}{2\theta_1^2}\right), \tag{6.2}$$

получается преобразованием стандартной величины

$$X = \theta_0 + \theta_1 Z. \tag{6.3}$$

2. Псевдослучайная величина, принадлежащая семейству распределений с функцией плотности

$$f(x;\theta_0,\theta_1,\lambda) = \frac{\lambda}{2\sqrt{2}\theta_1\Gamma(1/\lambda)}\exp\left(-\left(\frac{|x-\theta_0|}{\sqrt{2}\theta_1}\right)^{\lambda}\right), \tag{6.4}$$

$$E[x] = \theta_0, \quad D[x] = 2\theta_1^2\frac{\Gamma(3/\lambda)}{\Gamma(1/\lambda)},$$

где λ — параметр формы, находится из численного решения уравнения $Y = F(X)$, так как в этом случае обратная функция $F^{-1}(Y)$ не выражается явно.

Дополнительно в программной системе реализовано моделирование псевдослучайных величин, подчиняющихся законам распределения, приведенным в таблице 6.1.

Таблица 6.1

Функции плотности моделируемых законов распределения

Распределение случайной величины	Функция плотности	Функциональное преобразование
Экспоненциальное	$f(x;\theta_0,\theta_1)=\frac{1}{\theta_1}\exp\left(-\frac{1}{\theta_1}(x-\theta_0)\right)$	$X=\theta_0-\theta_1\ln Y$
Логистическое	$f(x;\theta_0,\theta_1)=\frac{1}{\theta_1}\frac{e^{-z}}{\left[1+e^{-z}\right]^2},\quad z=\frac{x-\theta_0}{\theta_1}$	$X=\theta_0-\theta_1\ln\left[\frac{1-Y}{Y}\right]$
Лапласа	$f(x;\theta_0,\theta_1)=\frac{1}{\theta_1}\exp\left(-\frac{\lvert x-\theta_0\rvert}{\theta_1}\right)$	$\begin{cases} X=\theta_0+\theta_1\ln(2Y), & Y\le 0.5 \\ X=\theta_0-\theta_1\ln(2(1-Y)), & Y>0.5 \end{cases}$
Коши	$f(x;\theta_0,\theta_1)=\frac{\theta_1}{\pi[\theta_1^2+(x-\theta_0)^2]}$	$X=\theta_0+\theta_1\,\mathrm{tg}[\pi(Y-0.5)]$
Минимального значения	$f(x;\theta_0,\theta_1)=\frac{1}{\theta_1}\exp\left(z-\exp z\right),\quad z=\frac{x-\theta_0}{\theta_1}$	$X=\theta_0+\theta_1\ln[-\ln Y]$
Максимального значения	$f(x;\theta_0,\theta_1)=\frac{1}{\theta_1}\exp\left(-z-\exp\left(-z\right)\right),\quad z=\frac{x-\theta_0}{\theta_1}$	$X=\theta_0-\theta_1\ln[-\ln Y]$
Вейбулла	$f(x;\theta_0,\theta_1,\alpha)=\frac{\alpha}{\theta_1}z^{\alpha-1}\exp\left(-z^{\alpha}\right),\quad z=\frac{x-\theta_0}{\theta_1}$	$X=\theta_0+\theta_1(-\ln Y)^{1/\alpha}$

Продолжение табл. 6.1

Распределение случайной величины	Функция плотности	Функциональное преобразование
Рэлея	$f(x;\theta_0,\theta_1)=\frac{2z}{\theta_1}\exp\left(-z^2\right),\quad z=\frac{x-\theta_0}{\theta_1}$	$X=\theta_0+\theta_1\sqrt{-\ln Y}$
χ–распределение	$f(x;\theta_0,\theta_1,n)=\frac{2z^{n-1}}{\theta_1 2^{n/2}\Gamma(n/2)}\exp\left(-\frac{z^2}{2}\right),\quad z=\frac{x-\theta_0}{\theta_1}$	$X=\theta_0+\theta_1\sqrt{\sum\limits_{i=1}^{n}Z_i^2}$
Максвелла	$f(x;\theta_0,\theta_1)=\frac{4(x-\theta_0)^2}{\sqrt{\pi}\theta_1^3}\exp\left(-\frac{(x-\theta_0)^2}{\theta_1^2}\right)$	$X=\theta_1\sqrt{Z_1^2+Z_2^2+Z_3^2}$
Эрланга	$f(x;\theta_0,\theta_1,n)=\frac{z^{n-1}}{\theta_1\Gamma(n)}\exp\left(-z\right),\quad z=\frac{x-\theta_0}{\theta_1}$	$X=\theta_0-\theta_1\ln\left(\prod\limits_{i=1}^{n}Y_i\right)$
Гамма	$f(x;\theta_0,\theta_1,\alpha)=\frac{z^{\alpha-1}}{\theta_1\Gamma(\alpha)}\exp\left(-z\right),\quad z=\frac{x-\theta_0}{\theta_1}$	$X=F^{-1}(Y)$ решается численно
Бета I-го рода	$f(x;\theta_0,\theta_1,\alpha,\beta)=\frac{z^{\alpha-1}}{\theta_1 B(\alpha,\beta)}\left(1-z\right)^{\beta-1},\quad z=\frac{x-\theta_0}{\theta_1}$	$X=F^{-1}(Y)$ решается численно
Бета II-го рода	$f(x;\theta_0,\theta_1,\alpha,\beta)=\frac{z^{\alpha-1}}{\theta_1 B(\alpha,\beta)\left(1+z\right)^{\alpha+\beta}},\quad z=\frac{x-\theta_0}{\theta_1}$	$X=F^{-1}(Y)$ решается численно

Продолжение табл. 6.1

Распределение случайной величины	Функция плотности	Функциональное преобразование
Бета III-го рода	$f(x;\theta_0,\theta_1,\alpha,\beta,\delta)=\frac{\delta^\alpha}{\theta_1 B(\alpha,\beta)}*$ $*\frac{z^{\alpha-1}(1-z)^{\beta-1}}{[1+(\delta-1)z]^{\alpha+\beta}},\quad z=\frac{x-\theta_0}{\theta_1}$	$X=F^{-1}(Y)$ решается численно
Sb-Джонсона	$f(x;\theta_0,\theta_1,\alpha,\beta)=\frac{\beta}{\sqrt{2\pi}}\frac{\theta_1}{(x-\theta_0)(\theta_1-x+\theta_0)}*$ $*\exp\left\{-\frac{1}{2}\left[\alpha+\beta\ln\left(\frac{x-\theta_0}{\theta_1-x+\theta_0}\right)\right]^2\right\}$	$X=\theta_0+\frac{\theta_1}{2}\left[1+\text{th}\left(\frac{Z-\alpha}{\beta}\right)\right]$
Sl-Джонсона	$f(x;\theta_0,\theta_1,\alpha,\beta)=\frac{\beta}{z\theta_1\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left[\alpha+\beta\ln z\right]^2\right),$ $z=\frac{x-\theta_0}{\theta_1}$	$X=\theta_0+\theta_1\exp\left\{\frac{Z-\alpha}{\beta}\right\}$
Su-Джонсона	$f(x;\theta_0,\theta_1,\alpha,\beta)=\frac{\beta}{\sqrt{2\pi}}\frac{1}{\theta_1\sqrt{z^2+1}}*$ $*\exp\left\{-\frac{1}{2}\left[\alpha+\beta\ln\left\{z+\left[z^2+1\right]^{1/2}\right\}^2\right]\right\},$ $z=\frac{x-\theta_0}{\theta_1}$	$X=\theta_0+\theta_1\,\text{sh}\left[\frac{Z-\alpha}{\beta}\right]$

Продолжение табл. 6.1

Распределение случайной величины	Функция плотности	Функциональное преобразование
Накагами	$f(x;\theta_0,\theta_1,\alpha)=\dfrac{2\alpha^{\alpha}(x-\theta_0)^{2\alpha-1}}{\theta_1^{2\alpha}\Gamma(\alpha)}\exp\left\{-\alpha\dfrac{(x-\theta_0)^2}{\theta_1^2}\right\}$	$X=F^{-1}(Y)$ решается численно
H–распределение	$f(x;\theta_0,\theta_1,\alpha,\delta)=\dfrac{\alpha\lvert z\rvert^{\alpha\delta-1}}{2\theta_1\Gamma(\delta)}\exp\left\{-\lvert z\rvert^{\alpha}\right\},\quad z=\frac{x-\theta_0}{\theta_1}$	$X=F^{-1}(Y)$ решается численно
Γ–распределение	$f(x;\theta_0,\theta_1,\alpha,\beta)=\dfrac{\alpha z^{\alpha\beta-1}}{\theta_1\Gamma(\beta)}\exp\left\{-z^{\alpha}\right\},\quad z=\frac{x-\theta_0}{\theta_1}$	$X=F^{-1}(Y)$ решается численно
L–распределение	$f(x;\theta_0,\theta_1,\alpha,\beta)=\dfrac{e^{\alpha z}}{\theta_1 B(\alpha,\beta)(1+e^{z})^{\alpha+\beta}},\quad z=\frac{x-\theta_0}{\theta_1}$	$X=F^{-1}(Y)$ решается численно

6.3.2. 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 $\Sigma = \|\sigma_{ij}\|, \quad i, j = \overline{1, m}$.

Функция плотности многомерного нормального закона имеет вид

$$f(\bar{X}) = \frac{1}{\sqrt{(2\pi)^m |\Sigma|}} \exp\left(-\frac{1}{2}(\bar{X} - \bar{M})^T \Sigma^{-1} (\bar{X} - \bar{M})\right). \tag{6.5}$$

Хорошо зарекомендовавший себя алгоритм генерирования псевдослучайных нормальных векторов был подробно изложен в [52]. 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

$$\bar{X} = A\bar{Z} + \bar{M}. \tag{6.6}$$

В (6.6) обычно полагают, что A является нижней треугольной матрицей

$$A = \begin{bmatrix} a_{11} & 0 & 0 & \ldots & 0 \\ a_{21} & a_{22} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{m1} & a_{m2} & a_{m3} & \ldots & a_{mm} \end{bmatrix},$$

тогда коэффициенты a_{ij} легко определяются рекуррентной процедурой:

$$a_{ij} = \frac{\sigma_{ij} - \sum\limits_{k=1}^{j-1} a_{ik} a_{jk}}{\sqrt{\sigma_{jj} - \sum\limits_{k=1}^{j-1} a_{jk}^2}}, \quad 1 \le j \le i \le m, \tag{6.7}$$

через соотношение (6.6) и элементы ковариационной матрицы

$$\sigma_{ij} = E\left[(X_i - M_i)(X_j - M_j)\right].$$

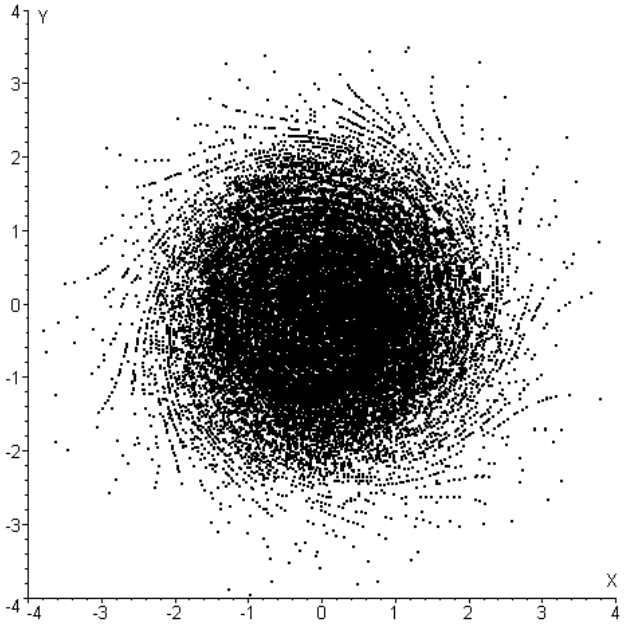


Рис. 6.5. Выборка двумерных случайных величин, смоделированная с использованием формул (6.1)

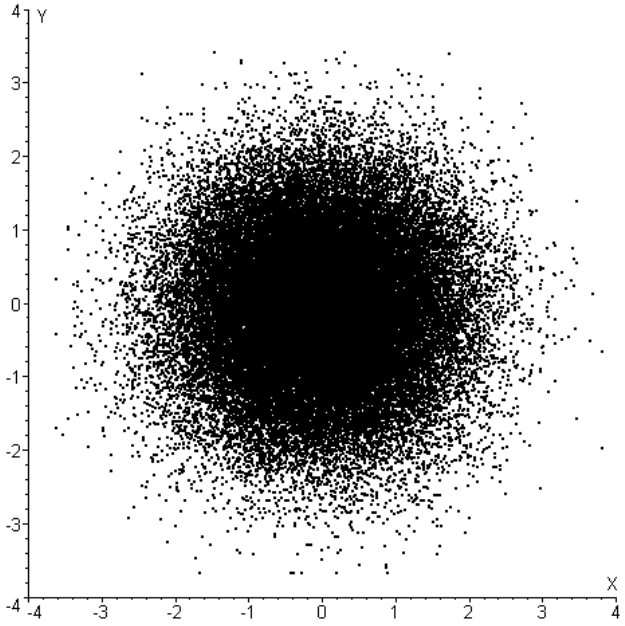


Рис. 6.6. Выборка двумерных случайных величин, смоделированная с использованием метода обратных функций

Исследование процедуры моделирования показало, что при моделировании больших объемов многомерных случайных векторов с использованием формул (6.1) для формирования совокупности $\{Z_i\}, i = \overline{1, m}$ уже в двумерном случае результаты оказываются неудовлетворительными. На рис. 6.5 явно видно появление регулярных структур, что может приводить к искажению результатов дальнейших исследований, опирающихся на процедуру моделирования. Если для моделирования одномерных стандартных нормальных величин использовать метод обратных функций, как и в случае семейства распределений (6.4), то появление регулярных структур не наблюдается (рис. 6.6). Поэтому целесообразней моделировать одномерные выборки нормальных случайных величин методом обратных функций.

6.3.3. Моделирование многомерных величин по законам, отличным от нормального

Процедуру моделирования многомерных величин, распределенных по законам, отличным от нормального, с некоторыми математическим ожиданием и ковариационной матрицей предложено [72] реализовать аналогично описанному выше алгоритму (6.6)–(6.7). Для этого определим в качестве параметров моделирования вектор $\bar{\Theta}_0$ и матрицу Θ_1, а совокупность величин $\{Z_i\}, i = \overline{1,m}$, будем формировать уже не по стандартному нормальному закону, а на основе некоторого одномерного распределения с нулевым математическим ожиданием ($E[Z_i] = 0$) и единичной дисперсией ($D[Z_i] = 1$). Элементы матрицы A' вычисляются по формуле (6.8), которая идентична соотношению (6.7). При этом вместо элементов ковариационной матрицы σ_{ij} используются элементы матрицы $\Theta_1 = \|\theta_{ij}^1\|$

$$a'_{ij} = \frac{\theta_{ij}^1 - \sum\limits_{k=1}^{j-1} a'_{ik} a'_{jk}}{\sqrt{\theta_{jj}^1 - \sum\limits_{k=1}^{j-1} a'^2_{jk}}}, \quad 1 \le j \le i \le m. \tag{6.8}$$

Псевдослучайный вектор $\bar{X}$ получается преобразованием вида

$$\bar{X} = A'\bar{Z} + \bar{\Theta}_0. \tag{6.9}$$

В результате на выходе процедуры мы имеем некоторый многомерный закон, отличный от нормального, но, вообще говоря, с неопределенными математическим ожиданием и ковариационной матрицей.

Определим математическое ожидание моделируемого случайного вектора $\bar{X}$. С использованием (6.9) вектор математического ожидания имеет вид

$$\bar{M} = E[\bar{X}] = E[A'\bar{Z} + \bar{\Theta}_0]. \tag{6.10}$$

Элементы вектора $\bar{M}$, если $\bar{\Theta}_0 = [\theta_1^0, \ldots, \theta_m^0]^T$, представимы в виде

$$M_i = E\left[\sum_{k=1}^{i} a'_{ik} Z_k + \theta_i^0\right] = \theta_i^0 + \sum_{k=1}^{i} a'_{ik} E[Z_k]. \tag{6.11}$$

А если учесть, что $E[Z_1] = \ldots = E[Z_m] = 0$, то получаем

$$\bar{M} = \bar{\Theta}_0. \tag{6.12}$$

Найдем ковариационную матрицу моделируемого многомерного закона. По определению ковариационная матрица находится как

$$\Sigma = E\left[\left(\bar{X} - \bar{M}\right)\left(\bar{X} - \bar{M}\right)^T\right]. \tag{6.13}$$

Если подставить в (6.13) представление (6.9) вектора $\bar{X}$ и учесть равенство (6.12), то получим

$$\Sigma = E\left[\left(A'\bar{Z}\right)\left(A'\bar{Z}\right)^T\right], \tag{6.14}$$

или для элементов матрицы

$$\sigma_{ij} = E\left[\sum_{k=1}^{m} a'_{ik} Z_k \sum_{k=1}^{m} a'_{jk} Z_k\right]. \tag{6.15}$$

Так как $\{Z_i\}, i = \overline{1,m}$, представляет собой совокупность моделируемых одинаково распределенных независимых случайных величин, то $cov(Z_i, Z_j) = E[Z_i Z_j] = 0, i \neq j$. И так как $D[Z_1] = \ldots = D[Z_m] = 1$, то (6.15) принимает вид

$$\sigma_{ij} = E\left[\sum_{k=1}^{m} a'_{ik} a'_{jk} Z_k^2\right] = \sum_{k=1}^{m} a'_{ik} a'_{jk} \tag{6.16}$$

А если учесть свойство $AA^T = \Theta_1$ разложения (6.8), получим окончательный результат

$$\Sigma = \Theta_1. \tag{6.17}$$

Таким образом соотношения (6.12) и (6.17) показывают, что у моделируемого случайного вектора математическое ожидание равно вектору параметров $\bar{\Theta}_0$, а ковариационная матрица — матрице параметров Θ_1.

Для моделирования различных совокупностей $\{Z_i\}, i = \overline{1,m}$, удобно использовать семейство распределений с плотностью (6.4) и параметром формы λ, так как оно охватывает целый класс симметричных распределений. Частными случаями данного закона являются распределение Лапласа (при $\lambda = 1$),

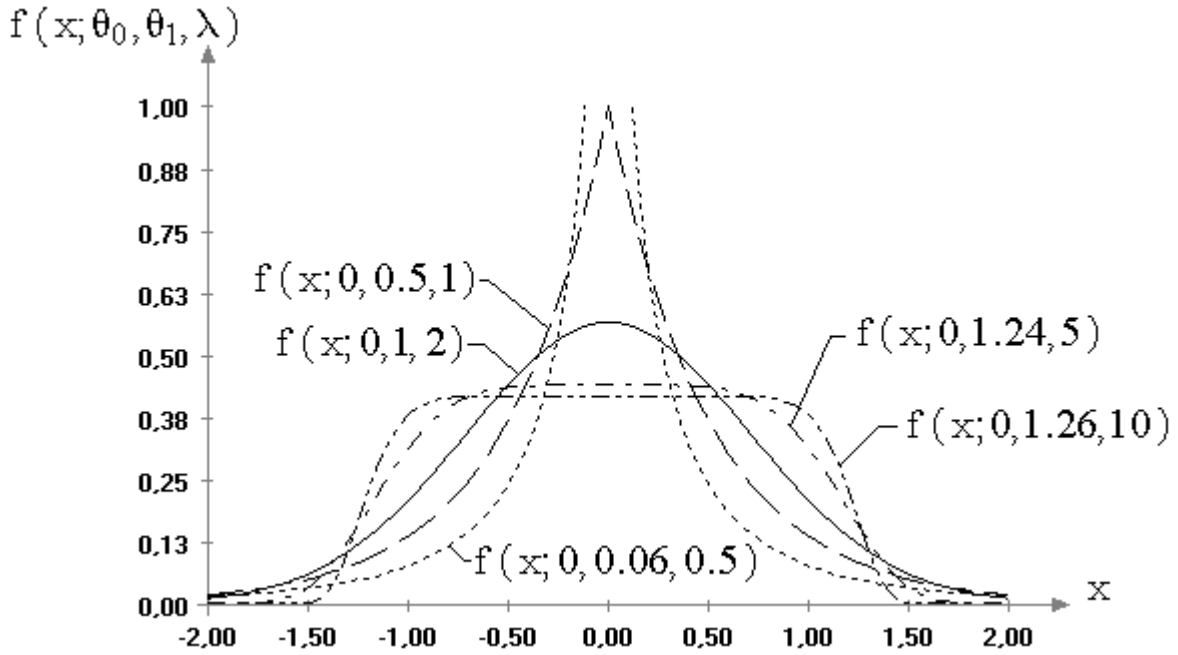


Рис. 6.7. Функции плотности семейства распределений (6.4) $f(x;\theta_0,\theta_1,\lambda)$ при различных параметрах формы ($E[x]=0$, $D[x]=1$)

нормальное ($\lambda=2$), а предельными — распределение Коши ($\lambda\to 0$) и равномерное ($\lambda\to+\infty$). Рис. 6.7 иллюстрирует изменение функции плотности данного семейства при изменении параметра формы от 0.5 до 10, где параметры сдвига и масштаба θ_0 и θ_1 выбраны из условия выполнения равенств $E[x]=0$, $D[x]=1$. 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.

Недостатком предложенной процедуры является то, что она не позволяет нам моделировать многомерный закон с некоторой произвольной функцией распределения, который находится на «заданном» расстоянии (определяемом в смысле некоторой меры) от многомерного нормального закона. Однако, при помощи этой процедуры мы можем построить датчик, генерирующий псевдослучайные векторы по закону, отличающемуся от нормального, с заданными

математическим ожиданием и ковариационной матрицей.

Если для моделирования $\{Z_i\}, i = \overline{1,m}$, использовать семейство распределений (6.4), то с учетом выражения для дисперсии можно получить выражение для параметра масштаба

$$\theta_1 = \sqrt{\frac{1}{2}\frac{\Gamma(1/\lambda)}{\Gamma(3/\lambda)}}, \tag{6.18}$$

при котором $D[Z_i] = 1$.

В качестве примера проверим полученные результаты и возможность моделирования многомерных величин с заданными вектором математических ожиданий $\bar{M}$ и ковариационной матрицей Σ, сравнивая оценки максимального правдоподобия $\hat{\bar{M}}$ и $\hat{\Sigma}$ по моделируемым выборкам многомерных величин достаточно большого объема $N = 100000$ для различных значений параметра формы λ. Выберем начальные параметры равными

$$\bar{\Theta}_0 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \quad \Theta_1 = \begin{bmatrix} 5 & 1 & 2.5 \\ 1 & 6 & 1 \\ 2.5 & 1 & 5 \end{bmatrix}.$$

Представленные ниже результаты приведены с округлением до 3-х десятичных знаков после запятой.

При $\lambda = 1$ величины Z_i моделировались с параметрами $\theta_0 = 0$ и $\theta_1 = 0.5$. Полученные оценки вектора математических ожиданий и ковариационной матрицы —

$$\hat{\bar{M}} = \begin{bmatrix} 0.999 \\ 2.008 \\ 2.997 \end{bmatrix}, \quad \hat{\Sigma} = \begin{bmatrix} 5.002 & 1.013 & 2.503 \\ 1.013 & 6.036 & 1.077 \\ 2.503 & 1.007 & 4.969 \end{bmatrix}.$$

При $\lambda = 2$ величины Z_i моделировались с параметрами $\theta_0 = 0$ и $\theta_1 = 1$. Соответствующие оценки оказались равными

$$\hat{\bar{M}} = \begin{bmatrix} 0.999 \\ 2.001 \\ 2.999 \end{bmatrix}, \quad \hat{\Sigma} = \begin{bmatrix} 4.998 & 1.002 & 2.499 \\ 1.002 & 5.998 & 1.004 \\ 2.499 & 1.004 & 4.999 \end{bmatrix}.$$

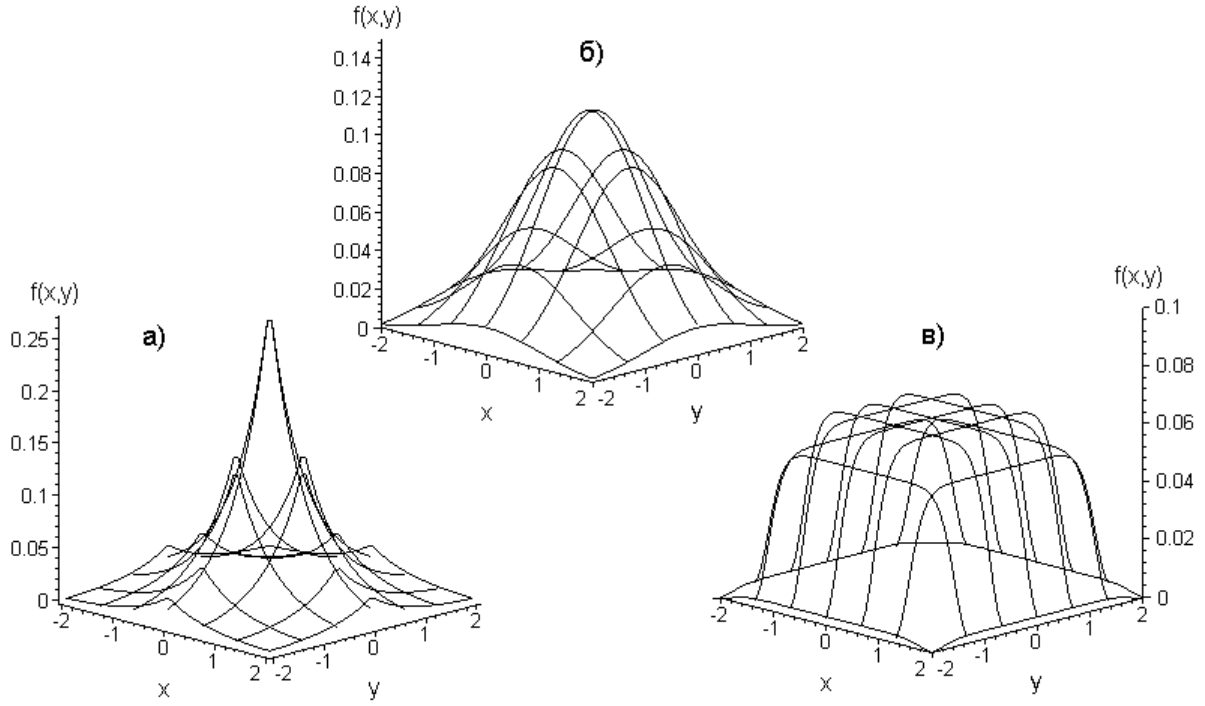


Рис. 6.8. Смоделированные плотности двумерных законов, построенных при различных значениях параметра формы: а) $\lambda = 1$, б) $\lambda = 2$ и в) $\lambda = 10$

При $\lambda = 5$ значения Z_i моделировались с $\theta_0 = 0$ и $\theta_1 = 1.2415$, соответствующие оценки —

$$\hat{\bar{M}} = \begin{bmatrix} 1.000 \\ 1.995 \\ 3.011 \end{bmatrix}, \quad \hat{\Sigma} = \begin{bmatrix} 5.024 & 0.993 & 2.511 \\ 0.993 & 5.991 & 0.994 \\ 2.511 & 0.994 & 4.999 \end{bmatrix}.$$

Во всех случаях оценки вектора математических ожиданий и ковариационных матриц дают основание говорить о выполнении равенств: $\bar{M} = \bar{\Theta}_0$ и $\Sigma = \Theta_1$. Таким образом, действительно каждый раз решалась задача по моделированию закона с заданными математическим ожиданием и ковариационной матрицей. Вообще говоря, реализации именно такой процедуры моделирования псевдослучайных векторов достаточно для целей настоящего исследования. На рис. 6.8 приведены полученные в результате моделирования функции плотностей двумерных законов с нулевым вектором математических ожиданий и единичной ковариационной матрицей: при $\lambda = 2$ (плотность нормального закона, в центре), при $\lambda = 1$ (слева) и $\lambda = 10$ (справа). Как видим, в первом случае наблюдается островершинное распределение, а во втором слу-

чае — плосковершинное. Полученное нормальное распределение существенно отличается от распределений, моделируемых с $\lambda \neq 2$.

В процессе исследования реализованной процедуры моделирования многомерных псевдослучайных величин исследовались и маргинальные распределения моделируемых многомерных векторов. Исследования показали, что маргинальные распределения многомерного закона, моделируемого с использованием выбранного семейства распределений (6.4) с параметром формы 2 (многомерный нормальный закон), хорошо согласуются с одномерным нормальным законом распределения. А маргинальные функции законов, получаемых при моделировании с параметром λ отличным от 2, существенно отличаются от нормального закона, но при этом хорошо согласуются с одномерным законом из семейства распределений (6.4).

6.3.4. Моделирование псевдослучайных векторов, подчиняющихся многомерному распределению Стьюдента

Случайный вектор $\bar{X}$ имеет m—мерное распределение Стьюдента с p степенями свободы, вектором сдвига $\bar{M}$ и матрицей точности T^{-1}, если функция плотности имеет вид

$$f(\bar{X}) = \frac{\Gamma\left(\frac{p+m}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\sqrt{(p\pi)^m|T|}}\left[1+\frac{1}{p}(\bar{X}-\bar{M})^T T^{-1}(\bar{X}-\bar{M})\right]^{-\frac{p+m}{2}}, \qquad (6.19)$$

где T — симметричная положительно определенная матрица.

Согласно [26] вектор математических ожиданий и ковариационная матрица многомерного распределения Стьюдента равны:

$$E[\bar{X}] = \bar{M}, \quad D[\bar{X}] = \Sigma = \frac{p}{p-2}T, \quad p > 2.$$

На рис. 6.9 приведены функции плотности двумерного распределения Стьюдента для степеней свободы $p = 3$, $p = 15$ и плотность двумерного нормального закона при равных значениях вектора математического ожидания и ковариационной матрицы. С ростом числа степеней свободы $p \to +\infty$ распределение Стьюдента стремится к нормальному распределению. Например,

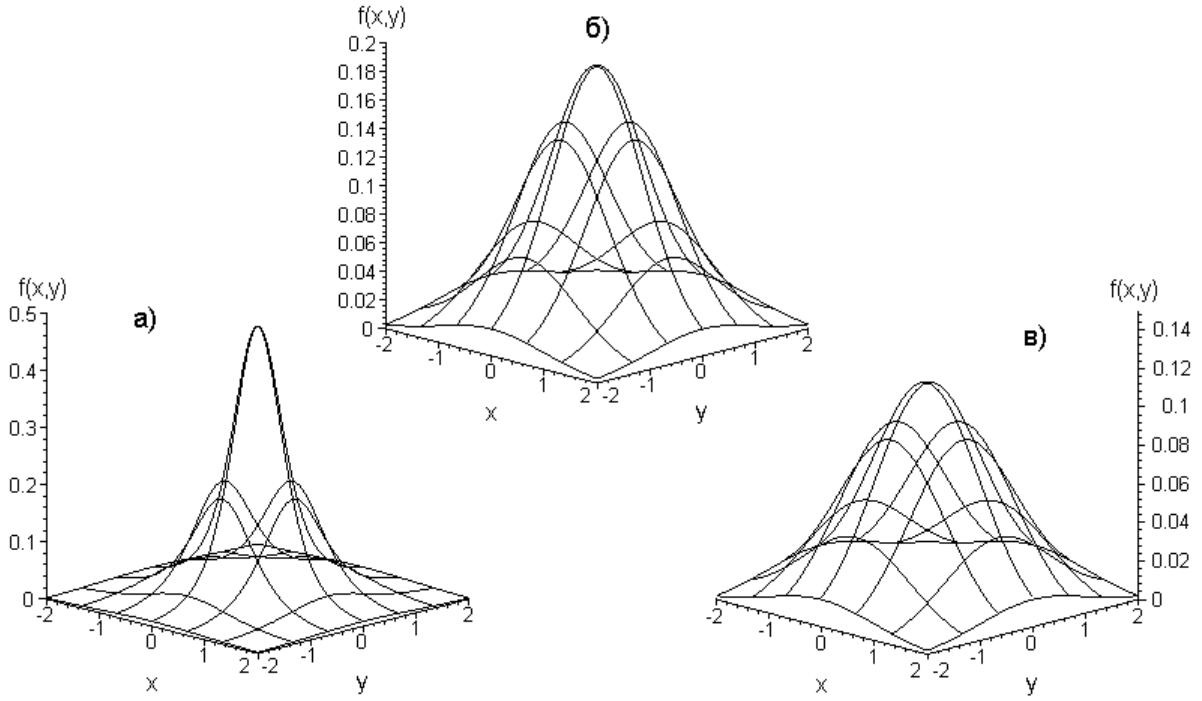


Рис. 6.9. Плотности двумерного закона Стьюдента, построенные при степенях свободы а) $p = 3$, б) $p = 15$, и в) нормальный закон

для значений $p \geq 200$ нормированная разность между двумерными функциями распределения Стьюдента и нормального не превышает по модулю 0.01.

В работе [26] приведен алгоритм моделирования псевдослучайных векторов, подчиняющихся многомерному распределению Стьюдента. Пусть вектор $\bar{Z}$ имеет многомерное нормальное распределение с нулевым вектором математических ожиданий и невырожденной ковариационной матрицей $\Sigma = T$, а ξ имеет χ^2—распределение с n степенями свободы, тогда вектор $\bar{X}$ определенный как

$$\bar{X} = \sqrt{\frac{n}{\xi}}\bar{Z} + \bar{M}, \tag{6.20}$$

имеет m—мерное распределение Стьюдента с p степенями свободы, вектором сдвига $\bar{M}$ и матрицей точности T^{-1}.

Используя формулу (6.20), мы можем генерировать псевдослучайные вектора, подчиняющиеся многомерному распределению Стьюдента с заданными параметрами: числом степеней свободы p, вектором математических ожиданий и ковариационной матрицей.

Описанные процедуры моделирования псевдослучайных векторов позволяют быстро получать выборки большого объема с любыми математическим ожиданием и ковариационной матрицей.

6.3.5. Моделирование функциональной линейной зависимости между X_i и X_j

Для исследования возможности выявления характера зависимости между компонентами случайного вектора необходимо моделировать псевдослучайные векторы с заданным видом зависимости, например, линейной. Рассмотрим двумерный случай, тогда вектор математических ожиданий $\bar{M}$ и ковариационная матрица Σ имеют вид

$$\bar{M} = \begin{bmatrix} M_1 \\ M_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}.$$

Вектор $\bar{X}$, распределенный по многомерному нормальному закону с параметрами $\bar{M}$ и Σ, получается через линейное преобразование вида (6.6).

Коэффициенты матрицы A вычисляются по формуле (6.7). В двумерном случае матрица A имеет вид

$$A = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 \\ \sigma_{12}/\sqrt{\sigma_{11}} & \sqrt{\sigma_{22} - \sigma_{12}^2/\sigma_{11}} \end{bmatrix}. \tag{6.21}$$

Подставляя матрицу A в выражение (6.6) получим два равенства:

$$\begin{aligned} X_1 &= a_{11}Z_1 + M_1 = \sqrt{\sigma_{11}}Z_1 + M_1, \\ X_2 &= a_{12}Z_1 + a_{22}Z_2 + M_2 = \sigma_{12}/\sqrt{\sigma_{11}}Z_1 + \sqrt{\sigma_{22} - \sigma_{12}^2/\sigma_{11}}Z_2 + M_2, \end{aligned} \tag{6.22}$$

где Z_i распределены по стандартному нормальному закону.

Если приравнять в (6.22) $M_2 = C_1M_1 + C_2$, $\sigma_{12} = C_1\sigma_{11}$ и $\sigma_{er} = \sqrt{\sigma_{22} - C_1^2\sigma_{11}}$, то получим линейную зависимость X_2 от X_1 вида

$$X_2 = C_1X_1 + C_2 + X_{er}, \tag{6.23}$$

где случайная величина X_1 имеет нормальное распределение $N(M_1, \sigma_{11})$, X_{er} распределена как $N(0, \sigma_{er})$, а C_1 и C_2 некоторые константы. Данная линейная

зависимость полностью определяется своими параметрами C_1, C_2, M_1, σ_{11} и σ_{er}.

Таким образом, если требуется смоделировать двумерную выборку с линейной зависимостью X_2 от X_1 вида (6.23), то потребуется задать следующие вектор математических ожиданий и ковариационную матрицу

$$\bar{M} = \begin{bmatrix} M_1 \\ C_1 M_1 + C_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_{11} & C_1\sigma_{11} \\ C_1\sigma_{11} & C_1^2\sigma_{11} + \sigma_{er} \end{bmatrix}. \tag{6.24}$$

6.4. Пример использования программной системы при обработке данных в медицине

Приведем пример использования программной системы для обработки данных лабораторных обследований при рассмотрении показателей липидного обмена у пациентов пожилого возраста с изолированной систолической артериальной гипертензией. Данные были получены после обследования 80 человек.

Из множества наблюдаемых показателей выберем два: общий холестерин (ОХС) и триглицериды (ТГ). Выборки значений по обоим показателям хорошо описываются семейством распределений (6.4): для ОХС при параметрах $\theta_0 = 5.34$, $\theta_1 = 0.97$ и $\lambda = 1.1$, а для ТГ при $\theta_0 = 1.61$, $\theta_1 = 0.63$ и $\lambda = 1.1$. Оценка коэффициента парной корреляции между ОХС и ТГ равна 0.88.

Пусть требуется проверить гипотезу вида $H_0 : r_{ij} = r_0$ для коэффициента парной корреляции. Из приведенных в главе 4 исследований следует, что распределение статистики z_0 (4.3) критерия проверки данной гипотезы существенно зависит от вида наблюдаемого закона. Поэтому из найденных моделей распределения показателей ОХС и ТГ вытекает, что использовать при проверке гипотезы $H_0 : r_{ij} = r_0$ классическое предельное распределение статистики z_0 (стандартный нормальный закон) некорректно.

Для определения распределения статистики z_0 воспользуемся разработанной программной системой. Смоделируем выборку значений статистики z_0 достаточно большого объема, например $N = 5000$, в случае наблюдения мно-

гомерного закона с параметром формы $\lambda = 1.1$. И идентифицируем закон распределения данной статистики по смоделированной выборке.

В программной системе для воспроизведения исходной модели обрабатываемых данных потребуется задать параметры моделирования многомерного закона. Здесь достаточно оценки коэффициента корреляции и найденных параметров законов распределения для показателей ОХС и ТГ. Так, с учетом нулевого математического ожидания и единичной дисперсии одномерное распределение компонент случайного вектора $\bar{Z}$ (6.9) есть распределение из семейства (6.4) с параметрами сдвига $\theta_0 = 0$, масштаба $\theta_1 = 0.58$ и формы $\lambda = 1.1$. А вектор $\bar{\Theta}_0$ и матрица Θ_1 из (6.8)–(6.9) имеют вид

$$\bar{\Theta}_0 = \begin{bmatrix} 5.34 \\ 1.61 \end{bmatrix}, \quad \Theta_1 = \begin{bmatrix} 2.80 & 1.60 \\ 0.87 & 1.18 \end{bmatrix}.$$

В результате моделирования при заданных параметрах будем наблюдать выборку псевдослучайного вектора $\bar{X}$, подчиняющегося, вообще говоря, неизвестному многомерному закону. Однако, маргинальные функции плотности моделируемого закона будут иметь вид для $X_1(\text{ОХС}) \in f(x; 5.34, 0.97, 1.1)$, а для $X_2(\text{ТГ}) \in f(x; 1.61, 0.63, 1.1)$, где $f(x; \theta_0, \theta_1, \lambda)$ — функция плотности (6.4). А коэффициент корреляции между X_1 и X_2 будет равен $r_{12} = 0.88$. Таким образом, воссоздан многомерный закон распределения для описания исследуемых показателей.

Смоделированные данным способом выборки значений статистики z_0 (4.3) хорошо описываются распределением семейства (6.4) при параметрах $\theta_0 = 0.01$, $\theta_1 = 1.36$ и $\lambda = 2.02$ (усредненный достигаемый уровень значимости по ряду критериев согласия $P_{\text{сред}} = 0.8$). Тогда как достигаемые уровени значимости при проверке согласия между смоделированными распределениями статистики z_0 и стандартным нормальным законом (классическое предельное распределение) меньше $1E - 5$.

Теперь, задавая в программной системе в качестве предельного закона статистики z_0 (4.3) найденное распределение, можно будет корректно проверить гипотезу вида $H_0 : r_{ij} = r_0$ для исходной выборки. При этом стоит отметить,

что при близких к нулю значениях статистики критерия z_0 различие между классическим предельным распределением и найденным несущественно. Но, например, если значение статистики $z_0 = -2.15$, то достигаемый уровень значимости при проверке гипотезы $H_0 : r_{ij} = 0.88$ с учетом найденного распределения статистики будет равен 0.12, тогда как использование стандартного нормального закона даст 0.03.

6.5. Выводы

1. В соответствии с целями диссертационной работы разработана программная система, которая реализует рассмотренные критерии классического корреляционного анализа, позволяет идентифицировать распределения статистик критериев через моделирование, осуществлять проверку гипотез для многомерных законов отличающихся от нормального по найденным распределениям статистик критериев.
2. В результате предложенного изменения метода моделирования псевдослучайных нормальных векторов реализована универсальная процедура, позволяющая на базе одномерного распределения моделировать многомерные псевдослучайные величины с заданными вектором математических ожиданий и ковариационной матрицей.

   Для исследований выбрано семейство распределений (6.4), позволяющее моделировать псевдослучайные величины, подчиненные как многомерному нормальному закону (параметр формы = 2), так и по закону отличному от нормального. Это было подтверждено численными исследованиями, в том числе маргинальных функций распределения моделируемых многомерных законов.
3. Реализована процедура моделирования псевдослучайных векторов, подчиняющихся m—мерному распределению Стьюдента с p степенями свободы, с заданными вектором математических ожиданий и ковариационной матрицей.

Разработанная программная система была использована Илюшенко А. Е.

[55] для расчета коэффициентов межвидовой сопряженности 4^x–польной матрицы для массива данных в диссертационной работе на соискание ученой степени кандидата биологических наук «Группировки почвенных водорослей сосновых фитоценозов в режиме рекреационной нагрузки».

В диссертационной работе на соискание ученой степени кандидата медицинских наук Вихман Е. А. «Некоторые особенности изолированной систолической артериальной гипертензии у мужчин пожилого возраста» программная система применялась для уточнения наличия связей и их характера при рассмотрении показателей периферической, центральной гемодинамики, данных метаболизма у пациентов с изолированной систолической артериальной гипертензией пожилого возраста.

Программное обеспечение используется на факультете прикладной математики и информатики НГТУ при проведении лабораторных работ по курсу «Компьютерные технологии анализа данных и исследования статистических закономерностей» по специальности 010200 — прикладная математика и информатика, результаты исследований закономерностей многомерного анализа при нарушении предположений включены в курс «Методы статистического анализа», читаемых по направлению магистерской подготовки 510200 — прикладная математика и информатика.

ЗАКЛЮЧЕНИЕ

В соответствии с целями исследований на базе разработанного программного обеспечения получены следующие основные результаты:

1. Показано, что получаемые методами компьютерного моделирования эмпирические распределения статистик корреляционного анализа в случае многомерного нормального закона хорошо согласуются с классическими предельными распределениями этих статистик. Для статистик различных критериев получены оценки объемов выборок n, начиная с которых распределения соответствующих статистик хорошо согласуются с предельными.
2. Реализована универсальная процедура, позволяющая на базе семейства распределений (6.4) моделировать псевдослучайные величины с заданными математическим ожиданием и ковариационной матрицей, распределенные как по многомерному нормальному закону, так и по законам отличным от нормального.
3. Показано, что распределения статистик, используемых при проверке гипотез о векторе математических ожиданий, устойчивы к отклонениям многомерного закона от нормального в достаточно широких пределах: значимого изменения распределений статистик не происходит. Как в случае более островершинных по сравнению с нормальным, так и в случае более плосковершинных многомерных законах распределения данных статистик по—прежнему хорошо описываются классическими результатами, полученными в предположении о нормальности наблюдаемого вектора. Аналогичная ситуация наблюдается и в одномерном случае при проверке гипотез вида $H_0 : \mu = \mu_0$.
4. Показано, что распределения статистик критериев, используемых при проверке гипотез о ковариационной матрице, существенно зависят от вида наблюдаемого многомерного закона. В случае принадлежности наблюдений m—мерным законам, хорошо описываемым моделями, получаемыми в соответствии с разработанной процедурой моделирования, для

распределений статистик L_1 и L_2 найдены аналитические модели законов, описывающие распределения этих статистик при определенных значениях размерности m и параметре формы λ.

Аналогичные результаты получены в одномерном случае для критериев проверки гипотез вида $H_0 : \sigma^2 = \sigma_0^2$ при известном и неизвестном математическом ожидании: построены модели распределений и таблицы процентных точек для соответствующих статистик в случае принадлежности наблюдений семейству распределений (6.4).

5. Показано, что распределения статистик критериев, используемых при проверке гипотез вида $H_0 : r_{ij} = 0$ для парных, частных и множественных коэффициентов корреляции, устойчивы к отклонениям наблюдаемого многомерного закона от нормального. Эмпирические распределения данных статистик по—прежнему хорошо описываются предельными законами, полученными в предположении о нормальности наблюдаемых величин.

   В то же время, в случае многомерных законов с «тяжелыми хвостами» наблюдается значимое отличие распределений статистик t, t^p и F соответствующих критериев от предельных классических.

6. Используемые в критериях проверки гипотез о равенстве заданному значению парного или частного коэффициента корреляции статистики z_0 и z_0^p существенно зависят от наблюдаемого многомерного закона. В то же время показано, что при $|r_0| \leq 0.15$ для проверки гипотез вида $H_0 : r_{ij} = r_0$ можно пользоваться классическими результатами.

7. Показано, что оценка корреляционного отношения сильно зависит от количества интервалов группирования. Показано, что разбиение области определения на интервалы равной частоты является наиболее предпочтительным для вычисления оценок $\hat{\rho}_{ij}^2$.

8. Показано, что распределение статистики критерия проверки гипотезы вида $H_0 : \rho_{ij}^2 = 0$ в случае многомерного нормального закона хорошо согласуется с теоретическим предельным распределением, полученным в классическом корреляционном анализе. В случае многомерных законов,

отличающихся от нормального в достаточно широких пределах (более островершинных или более плосковершинных), изменения предельного распределения статистики F_1 не происходит.

Показаны вычислительные проблемы, возникающие при проверке гипотез вида $H_0 : \rho_{ij}^2 = r_{ij}^2$, и плохая сходимость распределения статистики F_2 к предельному.

Полученные результаты расширяют сферу корректного применения методов классического многомерного анализа в приложениях. Разработанное программное обеспечение используется при проведении научных исследований и в учебном процессе.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. *Alqallaf F. A., Konis K. P., Martin R. D.* Scalable robust covariance and correlation estimates for data mining // Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. — 2002. — Pp. 14–23.
2. *Anderson T. W.* An Introduction to Multivariate Statistical Analysis. — Third edition. — Wiley-Interscience, 2003. — 752 pp.
3. *Bose R. C., Roy S. N.* The exact distribution of the studentized D^2–statistic // *Sankhya*. — 1938. — Vol. 4. — Pp. 19–38.
4. *Bose R. C., Roy S. N.* The use and distribution of the studentized D^2–statistic, when the variances and covariances are based on K samples // *Sankhya*. — 1938. — Vol. 4. — Pp. 535–542.
5. *Chandra M., Singpurwalla N. D., Stephens M. A.* Statistics for test of fit for the Extrem-Value and Weibull distribution // *J. Am. Statist. Assoc.* — 1981. — Vol. 76. — P. 375.
6. *Chen E. H.* A random normal number generator for 32-bit-word computers // *J. Am. Statist. Assoc.* — 1971. — Vol. 66. — Pp. 400–403.
7. *Devlin S. J., Gnanadesikan R., Kettenring J. R.* Robust estimation and outlier detection with correlation coefficient // *Biometrika*. — 1975. — Vol. 62. — Pp. 531–545.
8. *Fisher R. A.* The distribution of the partial correlation coefficient // *Metron*. — 1924. — Vol. 3. — Pp. 329–332.
9. *Fisher R. A.* The general sampling distribution of the multiple correlation coefficient // *Proc. Roy. Soc.* — 1928. — Vol. A121. — Pp. 654–673.
10. *Gayen A. K.* The frequency distribution of the Radial standard deviation // *Ann. Math. Soc.* — 1951. — Vol. 2. — Pp. 188–202.
11. *Hotelling H.* A generalized T–test and measure of multivariate dispersion // Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability. — University of California Press, 1951. — Pp. 23–42.

12. *Hotelling H.* New light on the correlation coefficient and its transforms // *J. Roy. Stat. Soc.* — 1953. — Vol. B 15. — Pp. 193–225.
13. *Huseby J. R., Schwertman N. C., Allen D. M.* Computation of the mean vector and dispersion matrix for incomplete multivariate data // *Communs Statist.* — 1980. — Vol. 9. — Pp. 301–309.
14. *Johnson M. E.* Multivariate Statistical Simulation: A Guide to Selecting and Generating Continuous Multivariate. — John Wiley & Sons, 1987. — 240 pp.
15. *L'Ecuyer P., Touzin R.* On the Deng-Lin random number generators and related methods // *Statistics and Computing*. — 2004. — Vol. 14. — Pp. 5–9.
16. *Lumley T., Diehr P., Emerson S.* The importance of the normality assumption in large public health data sets // *Annual Review of Public Health*. — 2002. — Vol. 23. — Pp. 151–169.
17. *Pearson E. S., Hartley H. O.* Biometrica tables for Statistics. — Cambridge: University Press, 1972. — Vol. 2. — 634 pp.
18. *Pearson K.* On the coefficients of Racial likeness // *Biometrika*. — 1926. — Vol. 18. — Pp. 105–117.
19. *Pearson K.* Note on standardization of method using the coefficients of Racial likeness // *Biometrika*. — 1928. — Vol. 20B. — Pp. 376–378.
20. *Shevlyakov G. L.* On robust estimation of a correlation coefficient // *Journal of Mathematical Sciences*. — 1997. — Vol. 83, no. 3. — Pp. 90–94.
21. *Shevlyakov G. L., Lee J. W.* Robust estimators of a correlation coefficient: Monte Carlo and asymptotics // *Korean Journal of Mathematical Sciences*. — 1997. — Vol. 4. — Pp. 205–212.
22. *Stein P. G., Matey J. R., Pitts K.* A review of statistical software for the Apple Macintosh // *The American Statistician*. — 1997. — Vol. 32, no. 1. — Pp. 67–82.
23. *Stephens M. A.* Use of Kolmogorov–Smirnov, Cramer–von Mises and related statistics – without extensive table // *J. R. Stat. Soc.* — 1970. — Vol. 32. — Pp. 115–122.

24. *Stephens M. A.* EDF statistics for goodness of fit and some comparisons // *J. Am. Statist. Assoc.* — 1974. — Vol. 69. — Pp. 730–737.

25. *Wilks S. S.* Moments and distribution of estimates of population parameters from fragmentary samples // *Ann. Math. Stat.* — 1932. — Vol. 3. — Pp. 163–195.

26. *Абусев Р. А., Колегова Н. В.* Байесовские оценки для некоторых характеристик многомерного t-распределения стьюдента // Мат. межд. научн.-практ. конференции «САКС-2001». — Т. 2. — Красноярск: САА, 2001. — С. 291–292.

27. *Айвазян С. А.* Программное обеспечение персональных ЭВМ по статистическому анализу данных // Компьютер и экономика: экономические проблемы компьютеризации общества. — М.: Наука, 1991. — С. 91–107.

28. *Айвазян С. А.* Программное обеспечение персональных ЭВМ по статистическому анализу данных (проблемы, тенденции, перспективы отечественных разработок) // *Заводская лаборатория. Диагностика материалов.* — 1991. — Т. 57, № 1. — С. 54–58.

29. *Айвазян С. А., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика: Основы моделирования и первичная обработка данных. — М.: Финансы и статистика, 1983. — 471 с.

30. *Айвазян С. А., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика: Исследование зависимостей. — М.: Финансы и статистика, 1985. — 487 с.

31. *Айвазян С. А., Мхитарян В.* Прикладная статистика и основы эконометрики. Учебник для вузов. — М.: ЮНИТИ, 1998. — 1022 с.

32. *Александров А. Д., Алексеев А. И., Горский Н. Д.* Анализ данных на ЭВМ (на примере системы СИТО). — М.: Финансы и статистика, 1990. — 192 с.

33. *Андерсон Т.* Введение в многомерный статистический анализ. — М.: Физматгиз, 1963. — 500 с.

34. *Архангельский А.* Программирование в C++ Builder 6. — М.: Бином, 2002. — 1152 с.

35. *Афифи А., Эйзен С.* Статистический анализ: Подход с использованием ЭВМ. — М.: Мир, 1982. — 488 с.

36. *Болч Б., Хуань К. Д.* Многомерные статистические методы для экономики. — М.: Статистика, 1979. — 317 с.

37. *Бусленко Н. П., Шрейдер Ю. А.* Метод статистических испытаний Монте-Карло и его реализация в цифровых машинах. — М.: Физматгиз, 1961. — 266 с.

38. *Векслер Л. С.* Статистический анализ на персональном компьютере // *Мир ПК*. — 1992. — № 2. — С. 89–97.

39. ГОСТ Р 50779.53-98. Приемочный контроль качества по количественному признаку для нормального распределения. Часть 1. Стандартное отклонение известно. — М.: Изд-во стандартов, 1998. — 23 с.

40. *Губарев В. В.* Вероятностные модели: Справочник. В 2-х ч. — Новосибирск: Изд-во НЭТИ, 1992. — Т. 2. — 188 с.

41. *Губарев В. В.* Вероятностные модели: Справочник. В 2-х ч. — Новосибирск: Изд-во НЭТИ, 1992. — Т. 1. — 198 с.

42. *Давидович М. И., Петрович М. Л.* Программное обеспечение ЭВМ: Библиотека прикладных программ БИМ. Вып. 20. (Прикладная статистика. Корреляционный анализ.). — Минск: Институт математики, АН БССР, 1989. — 187 с.

43. *Денисов В. И., Лемешко Б. Ю., Постовалов С. Н.* Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Методические рекомендации. Часть I. Критерии типа χ^2. — Новосибирск: Изд-во НГТУ, 1998. — 126 с.

44. *Джонстон Д.* Эконометрические методы. — М.: Статистика, 1980. — 446 с.

45. *Дубровский С. А.* Прикладной многомерный статистический анализ. — М.: Финансы и статистика, 1982. — 216 с.

46. *Дьяконов В.* Maple 6: учебный курс. — СПб.: Питер, 2001. — 608 с.

47. *Елисеева И. И., Семенова Е. В.* Основные процедуры многомерного статистического анализа. — Л.: УЭФ, 1993. — 78 с.

48. *Енюков И. С.* Методы, алгоритмы, программы многомерного статистического анализа: Пакет ППСА. — М.: Финансы и статистика, 1986. — 232 с.

49. *Ермаков С. М.* Метод Монте-Карло и смежные вопросы. — М.: Наука, 1975. — 471 с.

50. *Ермаков С. М.* О датчиках случайных чисел // *Заводская лаборатория. Диагностика материалов*. — 1993. — Т. 59, № 7. — С. 48–50.

51. *Ермаков С. М., Михайлов Г. А.* Курс статистического моделирования. — М.: Наука, 1976. — 320 с.

52. *Ермаков С. М., Михайлов Г. А.* Статистическое моделирование. — М.: Наука, 1982. — 296 с.

53. *Загоруйко Н. Г.* Анализ данных и анализ знаний // Анализ последовательностей и таблиц данных. Вып. 150: Вычислительные системы. — Новосибирск: 1994. — С. 3–17.

54. *Загоруйко Н. Г.* Прикладные методы анализа данных и знаний. — Новосибирск: Изд-во Ин-та математики, 1999. — 270 с.

55. *Илюшенко А. Е.* Группировки почвенных водорослей сосновых фитоценозов в режиме рекреационной нагрузки: Автореф. дисс. . . к–та биолог. наук. / ГУ. — Н., 2003. — 21 с.

56. *Кемени Д., Снелл Д.* Кибернетическое моделирование. — М.: Сов. радио, 1972. — 192 с.

57. *Кендалл М., Стьюарт А.* Теория распределений. — М.: Наука, 1966. — 588 с.

58. *Кендалл М., Стьюарт А.* Статистические выводы и связи. — М.: Наука, 1973. — 900 с.

59. *Кендалл М., Стьюарт А.* Многомерный статистический анализ и временные ряды. — М.: Наука, 1976. — 736 с.

60. *Кирьянов Б. Ф.* К проблеме моделирования случайных векторов // Вестник НовГУ. — № 3. — Новгород: 1996. — С. 87–89.

61. Компьютерные методы исследований статистических закономерностей / Б. Ю. Лемешко, С. Н. Постовалов, С. С. Помадин и др. // Тезисы докладов всероссийской НТК «Информационные системы и технологии ИСТ-2001». — Нижний Новгород: 2001. — С. 87–89.

62. *Куллдорф Г.* Введение в теорию оценивания по группированным и частично группированным выборкам. — М.: Наука, 1966. — 176 с.

63. *Лбов Г. С.* Методы обработки разнотипных экспериментальных данных. — Новосибирск: Наука, 1981. — 157 с.

64. *Лбов Г. С., Старцева Н. Г.* Логические решающие функции и вопросы статистической устойчивости решений. — Новосибирск: Изд-во Ин-та математики, 1999. — 212 с.

65. *Лемешко Б. Ю.* Корреляционный анализ многомерных наблюдений случайных величин: Программная система. — Новосибирск: Изд-во НГТУ, 1995. — 39 с.

66. *Лемешко Б. Ю.* Статистический анализ одномерных наблюдений случайных величин: Программная система. — Новосибирск: Изд-во НГТУ, 1995. — 125 с.

67. *Лемешко Б. Ю.* Компьютерные методы исследований статистических закономерностей // Сб. «Моделирование, автоматизация и оптимизация наукоемких технологий». — Новосибирск: Изд-во НГТУ, 2000. — С. 18–19.

68. *Лемешко Б. Ю., Ванюкевич О. Н.* Проверка гипотез о дисперсии при нарушении предположений о нормальности // Сб. научных трудов НГТУ. — Новосибирск: Изд-во НГТУ, 2002. — № 3(29). — С. 27–32.

69. *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.

70. *Лемешко Б. Ю., Помадин С. С.* Исследование распределений статистик в корреляционном анализе при отклонении многомерного закона от многомерного нормального // Сборник тезисов докладов Новосибирской меж-

вузовской научной студенческой конференции «Интеллектуальный потенциал Сибири» (Часть 1). — Новосибирск: 2000. — С. 15–16.

71. *Лемешко Б. Ю., Помадин С. С.* Исследование распределений статистик корреляционного анализа при отклонении многомерного закона от нормального // Материалы V международной конференции «Актуальные проблемы электронного приборостроения» АПЭП-2000. — Т. 7. — Новосибирск: 2000. — С. 184–187.
72. *Лемешко Б. Ю., Помадин С. С.* Один подход к моделированию псевдослучайных векторов с «заданными» числовыми характеристиками по законам, отличным от нормального // Российская научно-техническая конференция «Информатика и проблемы телекоммуникаций». Материалы конференции. — Новосибирск: 2001. — С. 121–122.
73. *Лемешко Б. Ю., Помадин С. С.* Статистическое моделирование распределений статистик корреляционного анализа при отклонении многомерного закона от нормального // Тезисы докладов региональной научной конференции студентов, аспирантов, молодых ученых «Наука. Техника. Инновации» (Часть 1). — Новосибирск: 2001. — С. 31–32.
74. *Лемешко Б. Ю., Помадин С. С.* Корреляционный анализ наблюдений многомерных случайных величин при нарушении предположений о нормальности // *Сибирский журнал индустриальной математики*. — 2002. — Т. 5, № 3(11). — С. 115–130.
75. *Лемешко Б. Ю., Помадин С. С.* Распределения статистик корреляционного анализа при отклонении многомерного закона от нормального // Материалы VI международной конференции «Актуальные проблемы электронного приборостроения» АПЭП-2002. — Т. 6. — Новосибирск: 2002. — С. 32–35.
76. *Лемешко Б. Ю., Помадин С. С.* Исследование распределений статистик, используемых при проверке гипотез о значениях математического ожидания и дисперсии, при наблюдаемых законах, отличных от нормального // Тезисы докладов МНТК «Информатика и проблемы телекоммуникаций». — Т. 2. — Новосибирск: 2003. — С. 142–143.

77. *Лемешко Б. Ю., Помадин С. С.* Исследование распределений статистик, используемых при проверке гипотез о математическом ожидании и дисперсии, в случае принадлежности наблюдаемых величин экспоненциальному семейству распределений // Материалы региональной конференции «Вероятностные идеи в науке и философии». — Новосибирск: 2003. — С. 102–105.

78. *Лемешко Б. Ю., Помадин С. С.* Исследование распределений статистик, используемых при проверке гипотез о ковариационных матрицах, при наблюдаемых законах, отличных от нормального // Тезисы докладов МНТК «Информатика и проблемы телекоммуникаций». — Т. 1. — Новосибирск: 2004. — С. 130–132.

79. *Лемешко Б. Ю., Помадин С. С.* Корреляционный анализ многомерных случайных величин при нарушении предположений о нормальности // Труды 10-го юбилейного симпозиума по непараметрическим и робастным статистическим методам в кибернетике. — Томск: 2004. — С. 114–128.

80. *Лемешко Б. Ю., Помадин С. С.* Проверка гипотез о математических ожиданиях и дисперсиях в задачах метрологии и контроля качества при вероятностных законах, отличающихся от нормального // *Метрология*. — 2004. — № 4. — С. 3–15.

81. *Лемешко Б. Ю., Помадин С. С., Кузьменко С. В.* Программное обеспечение компьютерного исследования статистических закономерностей в задачах корреляционного анализа // Российская научно-техническая конференция «Информатика и проблемы телекоммуникаций». Материалы конференции. — Новосибирск: 2001. — С. 79.

82. *Лемешко Б. Ю., Помадин С. С., Лемешко С. Б.* Численные исследования свойств критериев проверки статистических гипотез, используемых в задачах управления качеством // Тезисы докладов всероссийской НТК «Информационные системы и технологии ИСТ-2004». — Нижний Новгород: 2004. — С. 60–61.

83. *Лемешко Б. Ю., Помадин С. С., Французов А. В.* Статистическое моделирование распределений статистик, используемых в корреляционном анализе // Российская научно-техническая конференция «Информатика и проблемы телекоммуникаций». Материалы конференции. — Новосибирск: 2000. — С. 101–102.

84. *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.

85. *Лемешко Б. Ю., Постовалов С. Н.* Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Методические рекомендации. Часть II. Непараметрические критерии. — Новосибирск: Изд-во НГТУ, 1999. — 85 с.

86. *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.

87. *Лемешко Б. Ю., Постовалов С. Н.* Применение непараметрических критериев согласия при проверке сложных гипотез // *Автометрия*. — 2001. — № 2. — С. 88–102.

88. *Лемешко Б. Ю., Постовалов С. Н.* Непараметрические критерии при проверке сложных гипотез о согласии с распределениями Джонсона // *Доклады СО АН ВШ*. — 2002. — № 1(5). — С. 65–74.

89. *Лемешко Б. Ю., Постовалов С. Н.* Компьютерные технологии анализа данных и исследование статистических закономерностей: учеб. пособие. — Новосибирск: Изд-во НГТУ, 2004. — 120 с.

90. *Лемешко Б. Ю., Чимитова Е. В.* Методика компьютерного моделирования в исследовании статистических закономерностей // Тезисы докладов региональной НТК «Наука. Техника. Инновации». — Т. 2. — НТИ-2001, 2001. — С. 46–48.

91. *Лемешко Б. Ю., Чимитова Е. В.* Построение оптимальных L-оценок параметров сдвига и масштаба распределений по выборочным квантилям // *Сибирский журнал индустриальной математики*. — 2001. — Т. 4, № 2. — С. 166–183.

92. *Лемешко Б. Ю., Чимитова Е. В.* Оптимальные L–оценки параметров сдвига и масштаба распределений по выборочным квантилям // *Заводская лаборатория. Диагностика материалов*. — 2004. — Т. 70, № 1. — С. 54–66.

93. *Леонов В. П., Ижевский П. В.* Об использовании прикладной статистики при подготовке диссертационных работ по медицинским и биологическим специальностям // *Бюллетень ВАК РФ*. — 1997. — № 5. — С. 56–61.

94. *Леонов В. П., Ижевский П. В.* Применение статистики в медицине и биологии: анализ публикаций 1990-1997 гг. // *Сибирский медицинский журнал*. — 1997. — № 3-4. — С. 64–74.

95. *Маленво Э.* Статистические методы в эконометрии. — М.: Статистика, 1976. — 325 с.

96. *Манзон Б. М.* Maple V Power Edition. — М.: Информационно–издательский дом «Филинъ», 1998. — 240 с.

97. *Новицкий П. В., Зограф И. А.* Оценка погрешностей результатов измерений. — Л.: Энергоатомиздат, 1991. — 303 с.

98. *Орлов А. И.* Распространенная ошибка при использовании критериев Колмогорова и омега-квадрат // *Заводская лаборатория. Диагностика материалов*. — 1985. — Т. 51, № 1. — С. 60–62.

99. *Орлов А. И.* Часто ли распределение результатов наблюдений является нормальным? // *Заводская лаборатория. Диагностика материалов*. — 1991. — Т. 57, № 7. — С. 64–66.

100. *Орлов А. И.* О современных проблемах внедрения прикладной статистики и других статистичесих методов // *Заводская лаборатория. Диагностика материалов*. — 1992. — Т. 58, № 1. — С. 67–74.

101. *Орлов А. И.* Некоторые нерешенные вопросы в области математических методов исследования // *Заводская лаборатория. Диагностика материалов*. — 2002. — Т. 68, № 3. — С. 52–56.

102. *Пасман В. Р., Шевляков Г. Л.* Робастные методы оценивания коэффициента корреляции // *Автоматика и Телемеханика*. — 1987. — Т. 27, № 3. — С. 70–80.

103. *Петрович М. Л.* Численное исследование на ЭВМ некоторых алгоритмов прикладной статистики // *Заводская лаборатория. Диагностика материалов*. — 1991. — Т. 57, № 7. — С. 56–64.

104. *Петрович М. П., Давидович М. И.* Статистическое оценивание и проверка гипотез на ЭВМ. — М.: Финансы и статистика, 1989. — 192 с.

105. *Подбельский В. В.* Язык Си++: Учеб. пособие. — М.: Финансы и статистика, 1995. — 560 с.

106. *Полляк Ю. Г.* Вероятностное моделирование на электронных вычислительных машинах. — М.: Сов. радио, 1971. — 400 с.

107. *Помадин С. С.* К проверке гипотез о математических ожиданиях и дисперсиях при законах, отличающихся от нормального // *Сборник научных трудов НГТУ*. — 2003. — № 4(34). — С. 41–46.

108. *Райков Д. А.* Многомерный математический анализ. — М.: Высшая школа, 1989. — 271 с.

109. *Рао С. Р.* Линейные статистические методы и их применения. — М.: Наука, 1968. — 548 с.

110. *Рыданова Г. В.* Методика изучения временных зависимостей в последовательностях случайных чисел // *Заводская лаборатория. Диагностика материалов*. — 1986. — Т. 52, № 1. — С. 56–58.

111. Р 50.1.033-2001. Рекомендации по стандартизации. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть I. Критерии типа хи-квадрат. — М.: Изд-во стандартов, 2002. — 87 с.

112. Р 50.1.037-2002. Рекомендации по стандартизации. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть II. Непараметрические критерии. — М.: Изд-во стандартов, 2002. — 64 с.

113. *Соболь И. М.* Численные методы. — М.: Наука, 1973. — 312 с.

114. *Сошникова Л. А., Тамашевич В. Н., Уебе Г.* Многомерный статистический анализ в экономике: Учеб. пособие для вузов / Под ред. В. Н. Тамашевича. — М.: ЮНИТИ-ДАНА, 1999. — 598 с.

115. Статистические и математические системы // Каталог «Тысячи программных продуктов». — 1995. — № 2. — С. 88–92.

116. *Тьюки Д. У.* Анализ результатов наблюдений / Под ред. В. Э. Фигурнова. — М.: Мир, 1981. — 693 с.

117. *Тюрин Ю. Н.* О предельном распределении статистик Колмогорова–Смирнова для сложной гипотезы // *Изв. АН СССР. Сер. Матем.* — 1984. — Т. 48, № 6. — С. 1314–1343.

118. *Тюрин Ю. Н.* Исследования по непараметрической статистике (непараметрические методы и линейная модель): Автореф. дисс... д–ра физ.–мат. наук. / МГУ. — М., 1985. — 33 с.

119. *Тюрин Ю. Н., Макаров А. А.* Анализ данных на компьютере. — М.: Финансы и статистика, 1995. — 384 с.

120. *Тюрин Ю. Н., Макаров А. А.* Статистический анализ данных на компьютере / Под ред. В. Э. Фигурнова. — М.: ИНФРА, 1997. — 528 с.

121. *Тюрин Ю. Н., Саввушкина Н. Е.* Критерии согласия для распределения Вейбулла–Гнеденко // *Изв. АН СССР. Сер. Техн. Кибернетика*. — 1984. — № 3. — С. 109–112.

122. *Ферестер Э., Ренц Б.* Методы корреляционного и регрессионного анализа. — М.: Финансы и статистика, 1988. — 302 с.

123. *Шметтерер Л.* Введение в математическую статистику. — М.: Наука, 1976. — 520 с.

124. *Эфрон Б.* Нетрадиционные методы многомерного статистического анализа. — М.: Финансы и статистика, 1988. — 263 с.

Приложение

АКТЫ ВНЕДРЕНИЯ