МИНИСТЕРСТВО ЗДРАВООХРАНЕНИЯ И СОЦИАЛЬНОГО РАЗВИТИЯ РФ

ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ЗДРАВООХРАНЕНИЮ
И СОЦИАЛЬНОМУ РАЗВИТИЮ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО
ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ «САМАРСКИЙ
ГОСУДАРСТВЕННЫЙ МЕДИЦИНСКИЙ УНИВЕРСИТЕТ» РОСЗДРАВА

Г.В. НЕДУГОВ, В.В. НЕДУГОВА

СТАТИСТИЧЕСКИЙ АНАЛИЗ В СУДЕБНО-МЕДИЦИНСКОЙ АНТРОПОЛОГИИ

Самара 2007

УДК: 340.64:311
ББК 67.52
А 79

Недугов Г.В., Недугова В.В.
Статистический анализ в судебно-медицинской антропологии. – Самара, 2007. – 264 с.
Рисунков 41. Таблиц 42. Библиография: 159 наименований.

ISBN 5-86611-043-1 (978-5-86611-043-8)

Монография посвящена проблемам оптимизации статистического анализа биометрических данных при судебно-медицинской антропологической идентификации. В работе рассмотрены методы корреляционного, регрессионного, дискриминантного и кластерного видов анализа. Кроме традиционно используемых статистических процедур освещены перспективные, но пока еще не нашедшие должного применения методы одномерной биномиальной классификации при различных распределениях биометрических показателей. Особое внимание уделяется приемам статистического анализа при несоответствии эмпирических данных предпосылкам математических моделей статистических методов. Впервые в судебно-медицинской литературе предлагаются критерии и методы объективного сравнения точности регрессионных диагностических моделей и моделей классификации объектов экспертного познания.

Монография предназначена для исследователей, занимающихся проблемами судебно-медицинской антропологической идентификации, и судебно-медицинских экспертов медико-криминалистической специализации, преподавателей медицинских вузов, а также может быть полезной для любых исследователей, применяющих методы многомерного статистического анализа в биомедицине.

Научный редактор: заведующий кафедрой судебной медицины Самарского государственного медицинского университета, начальник ГУЗ «Самарское областное бюро судебно-медицинской экспертизы», заслуженный врач РФ, доктор медицинских наук *А.П. Ардашкин*

Рецензент: заведующий кафедрой естественнонаучных и технических дисциплин Кузнецкого института информационных и управленческих технологий, доктор технических наук *В.Г. Андреев*

ISBN 5-86611-043-1 (978-5-86611-043-8)

ПРЕДИСЛОВИЕ

> Способность применить статистический подход в медицине не сводится к заучиванию нескольких формул и умению отыскать табличное значение. Как и любая творческая деятельность, применение статистических методов и интерпретация полученных результатов требуют глубокого проникновения в суть дела.
>
> *С. Гланц*

> Если исследователи представляют цифры и притязают на то, что эти цифры что-то означают без использования статистических методов, они определенно ходят по тонкому льду.
>
> *Т. Гринхальх*

Методы статистического анализа в настоящее время находят все большее применение в научных биомедицинских исследованиях. При этом повышение интенсивности использования статистических методов сопровождается ростом абсолютного количества статистических ошибок в предлагаемых к публикации научных работах (относительная частота ошибок остается постоянной и примерно равна 50%). Аналогичная ситуация имеет место и в судебной медицине[1].

Отличительной чертой судебно-медицинской антропологии является относительная сформированность собственной, сугубо специфичной, методологии статистической обработки эмпирических данных. Сложность и специфичность применяемых в судебно-медицинской антропологии статистических процедур (в основном это – многомерные статистические методы) привела к появлению достаточно узкой группы исследователей, в совершенстве овладевших указанными методами статистической обработки данных и техническими приемами их сбора. Поэтому в работах, посвящен-

[1] Это утверждение основывается на данных сотрудников кафедры судебной медицины Самарского государственного медицинского университета, с 2001 г. изучающих эпидемиологию использования статистических методов и осуществляющих мониторинг статистических ошибок в отечественных научных исследованиях судебно-медицинской тематики.

ных судебно-медицинской антропологической идентификации, статистические ошибки практически отсутствуют.

Однако отсутствие ошибок не означает отсутствия проблем, требующих обсуждения. Объясняется это тем, что исследователи, в течение длительного времени занимавшиеся проблемами судебно-медицинской антропологической идентификации и практически монополизировавшие этот раздел судебной медицины и экспертной практики, давно уже сформировали собственные представления о приемах статистического анализа, предпочитая не выходить за их рамки. Между тем, математическая статистика развивается не менее бурно, чем судебная медицина, которая не успевает интегрировать не только все время появляющиеся новые, но и давно известные статистические методы. В этой связи одним из возможных путей повышения диагностической значимости результатов судебно-медицинских антропологических исследований является оптимизация методов статистического анализа эмпирических данных.

Изложенной проблеме и посвящена настоящая монография, при написании которой авторы поставили своей целью ознакомить судебных медиков с наиболее перспективными, но пока еще невостребованными методами статистического анализа. В работе рассмотрены методы корреляционно-регрессионного, дискриминантного и кластерного анализа, а также методы одномерной биномиальной классификации при различных непрерывных распределениях биометрических показателей. Особое внимание уделяется приемам статистического анализа при несоответствии эмпирических данных предпосылкам математических моделей традиционных статистических методов. В частности, подробно изложены приемы корреляционно-регрессионного анализа при неоднородности данных, нелинейности стохастических связей между изучаемыми показателями, наличии мультиколлинеарности, неоднородности дисперсии и серийной корреляции остатков, описаны процедуры сравнительного корреляционного и регрессионного анализа, автоматизированного подбора переменных в состав регрессионных и дискриминантных моделей, а также правила использования категориальных и ранговых показателей в качестве независимых переменных уравнений множественной регрессии.

Одним из основных положений, обосновываемых в данной работе, является утверждение о праве и обязанности судебно-медицинского эксперта самостоятельно осуществлять выбор опти-

мального способа антропологической идентификации, в связи с чем впервые предлагаются критерии объективного сравнения точности регрессионных диагностических моделей и моделей классификации объектов экспертного познания. Поэтому настоящая работа предназначается не только для исследователей, занимающихся проблемами медико-антропологической идентификации личности, но и для практических судебно-медицинских экспертов медико-криминалистической специализации.

Помимо теоретических выкладок авторы постоянно стремились обосновать необходимость использования обсуждаемых статистических методов с помощью подробно рассмотренных примеров, заимствованных из реальных исследований, посвященных судебно-медицинской антропологической идентификации. Значительная часть примеров приведена из собственной исследовательской практики авторов, преимущественно касающейся идентификации неопознанных трупов плодов и новорожденных. Однако подобное пристрастие к обсуждению собственных результатов объясняется только лишь отсутствием в научных публикациях подробных числовых данных, столь необходимых для демонстрации любой статистической процедуры[2]. Поэтому в случаях, когда изложение статистического метода позволяло оперировать усредненными данными, для примеров использовались результаты исследований других авторов, более актуальные для судебно-медицинской экспертной практики. Незначительная часть примеров сконструирована специально для демонстрации конкретного статистического приема и не отражает истинных результатов научных исследований. Подобные примеры в тексте книги сопровождены специальными сносками.

Определенные трудности составила проблема уровня освещения статистических методов. Дело в том, что в специальной литературе, посвященной популяризации статистических методов в различных практических приложениях, на этот счет существуют две полярные точки зрения. Первая из них предполагает строгое и подробное изложение основ математической статистики. Вторую точку зрения выразительно охарактеризовал американский исследователь-статистик Э. Сигел: «Жизнь слишком коротка, чтобы человек посвящал ее разбору таких сложных технических приемов, как деле-

[2] Определенную роль также сыграло банальное отсутствие в специальной литературе каких-либо альтернативных данных по некоторым обсуждаемым вопросам.

ние в столбик или обращение матрицы» [74, С. 29]. Учитывая преимущества и недостатки обеих указанных систем взглядов, в книге приведено систематическое изложение только лишь базовых математико-статистических методов, причем многие утверждения приводятся кратко, без доказательств. Идеи других, более специфичных разделов, которые, кстати, очень сложно или невозможно реализовать без использования вычислительной техники и специальных программных приложений, изложены с акцентом на существо описываемых методов без отвлечений на технические детали. И все же полноценное восприятие обсуждаемых вопросов невозможно без владения читателем основными понятиями дифференциального и матричного исчисления.

Авторы выражают глубокую благодарность своим родным Татьяне и Владимиру Недуговым, а также Валентине и Владимиру Рябовым, в очередной раз взявшим на себя значительную долю домашних забот в период подготовки данной книги, потребовавшей больших усилий. Авторы благодарны также своему четырехлетнему сыну Владимиру за его долгое терпение недостатка родительского внимания, на которое он имел и имеет полное право.

Авторы считают своим долгом выразить искреннюю признательность доктору медицинских наук А.П. Ардашкину и доктору технических наук В.Г. Андрееву, замечания и советы которых способствовали улучшению методики изложения и содержания книги.

Авторы выражают надежду, что предлагаемая работа поможет улучшить качество судебно-медицинской антропологической идентификации как на этапе научного, так и практического экспертного познания. Вместе с тем авторы в полной мере осознают, что некоторые положения могут быть предметом дискуссии, и потому с благодарностью примут все критические замечания и конструктивные предложения.

ГЛАВА 1. МЕТОДЫ СБОРА И СТАТИСТИЧЕСКОГО АНАЛИЗА ДАННЫХ ПРИ СУДЕБНО-МЕДИЦИНСКОЙ АНТРОПОЛОГИЧЕСКОЙ ИДЕНТИФИКАЦИИ

1.1. СУДЕБНО-МЕДИЦИНСКАЯ АНТРОПОЛОГИЧЕСКАЯ ИДЕНТИФИКАЦИЯ КАК КОМПЛЕКСНАЯ НАУЧНАЯ, ОБРАЗОВАТЕЛЬНАЯ И ЭКСПЕРТНАЯ ПРОБЛЕМА

Судебно-медицинская антропология представляет собой раздел судебной медицины, задачей которого является разработка специальных способов судебно-медицинской идентификации личности. Основными идентифицируемыми параметрами при судебно-медицинской антропологической идентификации служат унаследованные и появившиеся в процессе онтогенеза физические особенности человека: пол, возраст, раса, соматотип, длина тела и его сегментов, аномалии и пороки развития, травмы и заболевания, черты внешности, особенности дерматоглифики и т.д.

Современное состояние судебно-медицинской антропологической идентификации в силу настоятельных потребностей судебно-следственной практики характеризуется непрерывным увеличением количества и разнообразием разработанных приемов исследования идентифицируемых объектов и способов идентификации. При этом усложнение идентификации связано не только с увеличением числа идентифицируемых параметров человека и совершенствованием технических приемов изучения идентифицируемых объектов, но и с расширением спектра возможных объектов идентификации. В качестве последних в настоящее время могут быть представлены скелетированные и фрагментированные трупы неопознанных людей на любой стадии трупных изменений, объекты, похожие на какие-либо анатомические образования человека, особенности неопознанных трупов или их частей, документированные в виде их словесных, биометрических, графических, объемных, фотографических, рентгенологических и других моделей.

В указанных условиях достоверность судебно-медицинской антропологической идентификации во многом определяется правильным методологическим подходом: адекватностью выбора методик и конкретных способов и приемов исследования, правильностью их применения и корректностью оценивания полученных результатов. В этой связи важным положением является единство научного, об-

разовательного и практического элементов познания при судебно-медицинской идентификации личности. Так, точный результат идентификации невозможен или, по крайней мере, маловероятен без использования адекватных диагностических способов. В свою очередь, проведение научных исследований в области судебно-медицинской антропологии обусловлено соответствующими практическими экспертными потребностями.

Начальный этап практической судебно-антропологической идентификации характеризуется тем, что судебно-медицинский эксперт (субъект экспертного познания) после ознакомления с материалами о назначении экспертизы должен сформулировать экспертную задачу (определить идентифицируемые параметры) и выяснить характер представленных идентифицируемых объектов. Последний устанавливается путем изучения материалов о назначении экспертизы и последующего исследования идентифицируемых объектов общими методами, к которым принято относить методы, использующиеся для фиксации наиболее общих признаков [1].

Следующим этапом экспертного познания при судебно-медицинской антропологической идентификации, практически полностью определяющим успешность решения поставленной экспертной задачи, является выбор субъектом экспертного познания наиболее оптимального способа идентификации. Особенность медико-антропологической идентификации в этом отношении характеризуется тем, что фактически все известные способы установления общих идентификационных признаков индивида являются специальными диагностическими авторскими методиками, которые заключаются в использовании результатов применения того или иного комплекса общих методов для получения конкретных идентификационных данных. Каждая из указанных методик предусматривает индивидуальный комплекс и особую последовательность применения общих методов.

В этих условиях реализация волевого выбора наиболее адекватных способов и технических приемов идентификации вызывает у эксперта потребность в специальной научно-методической информации. При этом обнаружение сведений для решения конкретных идентификационных задач невозможно без владения субъектом экспертного познания навыками поиска и критической оценки научных данных на предмет их диагностической значимости и достоверности [47,66,147].

До сих пор владение навыками поиска, критической оценки и практического внедрения биомедицинской информации рекомендовалось лишь специалистам, занимающимся научно-исследовательской деятельностью. Между тем, экспертному судебно-медицинскому познанию присущи многие черты, свойственные научному познанию (системность, объективность, доказательность и др.). Невладение судебно-медицинским экспертом хотя бы одним из вышеперечисленных навыков отрицательно сказывается на его квалификации и ведет к снижению качества выполняемых экспертиз. Выходом из сложившейся ситуации является обучение каждого судебно-медицинского эксперта основным методам работы с информацией [47]. Важность владения экспертом такими навыками становится очевидной в условиях большого количества предложенных в судебной медицине способов идентификации, обладающих различной диагностической значимостью и достоверностью.

При отсутствии адекватных способов идентификации личности экспертом должен быть представлен обоснованный отказ от решения экспертной задачи. Накопление подобных отказов неизбежно приведет к обсуждению данной проблемы, к формулированию соответствующей научной задачи и инициированию научных исследований соответствующей тематики. Сущность указанных научных исследований сводится к набору объектов научного познания, тождественных таковым в экспертной практике, исследованию их характеристик (преимущественно метрических) с последующим отбором наиболее значимых в диагностическом отношении параметров и созданию на их основе с помощью математического аппарата собственно способа идентификации. Успешное решение уже субъектами научного познания поставленной научной задачи приведет к предложению ряда альтернативных способов судебно-медицинской антропологической идентификации, в пользу наилучшего из которых и должен сделать выбор практический эксперт при производстве судебно-медицинской экспертизы.

Выбор оптимального способа судебно-медицинской антропологической идентификации является одновременно и правом, и обязанностью эксперта. В плане обоснованности этого выбора любое научное исследование в области судебно-медицинской антропологии можно рассмотреть с позиций обобщаемости, достоверности и диагностической значимости.

Обобщаемость результатов исследования отражает обоснованность допущения того, что объекты исследования, включенные в него, сравнимы с другими, подобными им. Обобщаемость во многом зависит от степени соответствия изученной выборки основным характеристикам исследуемой популяции и степени сведения к минимуму систематических ошибок. Поэтому соответствие объектов научных исследований таковым в экспертной практике является непременным условием, определяющим выбор способа идентификации. Например, способ идентификации, разработанный для нативных костей, не может быть использован для идентификации их озоленных аналогов; методики краниометрической идентификации нельзя применять при наличии деформаций, аномалий и заболеваний черепа; определенную роль могут также играть возрастные и расовые ограничения, локализация парных костей [1,19].

Менее заметные погрешности, связанные с выбором способа идентификации с плохой обобщаемостью, характеризуются несоответствием научных данных остеометрическим характеристикам современной популяции вследствие процессов акселерации, произошедших со времени формирования объектов научного поиска. Например, давность коллекции черепов, использованных в работе В.И. Пашковой и Б.Д. Резникова для определения идентификационных параметров, на сегодняшний день превышает 100 лет [64]. Кроме того, основные руководства по остеологической идентификации изобилуют описаниями методик большой давности без подробного описания исследовавшихся костных массивов [24,64].

Достоверность научного исследования определяется тем, в какой степени структура исследования соответствует поставленным задачам, а полученные результаты корректны в отношении изучавшейся выборки. Достоверность зависит от правильности планирования, организации и структуры научного исследования. Общее суждение о достоверности суммируется из оценок методов фиксации свойств объектов, наличия рандомизации, мер, направленных на снижение влияния систематических ошибок, оценок использованных аналитических методов.

Заключительным этапом оценки способа судебно-медицинской антропологической идентификации является определение его диагностической значимости (точности идентификации). Критерии точности идентификации различны в зависимости от математической модели, на которой основан способ судебно-медицинской антропо-

логической идентификации. Для способов, созданных на основе регрессионного анализа, показателем точности идентификации является доверительная область для прогнозных оценок регрессионной модели. Для способов, основанных на применении различных статистических методов классификации, в качестве критерия точности идентификации пока рассматривается лишь один показатель - доля случаев ошибочной классификации объектов тестовой выборки.

В конечном счете, среди альтернативных способов судебно-медицинской антропологической идентификации приоритет должен быть отдан методике, характеризующейся хорошей обобщаемостью, наибольшими достоверностью и диагностической значимостью. Указанный выбор гарантирует реализацию способа судебно-медицинской антропологической идентификации и успешное решение экспертной задачи.

К сказанному следует добавить, что экспертизы отождествления личности являются одним из наиболее сложных видов судебно-медицинских экспертиз, отличаясь не только по характеру объектов и предмету исследования, но и по методам и техническим приемам последнего. Изложенное определяет необходимость дифференциации и официального закрепления самостоятельности судебно-медицинских антропологических экспертиз с утверждением соответствующих специальных программ подготовки кадров для учреждений судебно-медицинской экспертизы [48].

Таким образом, судебно-медицинскую антропологическую идентификацию в настоящее время необходимо рассматривать как специальный раздел судебной медицины (как научную и учебную дисциплины) и как отдельный вид судебно-медицинской экспертизы. Дальнейшее совершенствование судебно-медицинской антропологической идентификации в условиях постоянного увеличения объема специальных знаний и повышения требований к качеству экспертных заключений обусловлено не только разработкой новых диагностических способов, но и связано с выделением судебно-медицинской антропологической идентификации в качестве отдельного вида судебно-медицинской экспертизы. При этом одной из образовательных задач на современном этапе является обучение экспертов методам поиска необходимой медико-биологической информации и определения ее достоверности.

1.2. МЕТОДЫ СТАТИСТИЧЕСКОГО АНАЛИЗА В СУДЕБНО-МЕДИЦИНСКОЙ АНТРОПОЛОГИИ

Характерной чертой современной судебно-медицинской антропологии стало увеличение числа работ, посвященных поиску путей оптимизации статистических методов, используемых для анализа полученных эмпирических данных [28,51,52].

Все многообразие научных и практических экспертных проблем отождествления личности, решаемых с помощью статистических методов, можно условно разделить на две основные задачи. Одной из них является разработка и совершенствование способов прогнозирования каких-либо параметров идентифицируемых объектов, выраженных количественными признаками [34]. К числу основных идентифицируемых параметров, обладающих таким свойством, относятся биологический возраст, длина тела, различные соматические размеры человека. Средством реализации исследований названной группы являются поиск и изучение взаимосвязей между идентифицируемыми параметрами и идентифицирующими признаками и последующее построение аналитической модели, с определенной точностью позволяющей прогнозировать значение идентифицируемого параметра по значению идентифицирующего признака или группы признаков. Базовым методом создаваемых способов судебно-медицинского прогнозирования идентифицируемых параметров является корреляционно-регрессионный анализ.

Основным фактором, определяющим видовую структуру методов корреляционно-регрессионного анализа, следует назвать количество идентифицирующих признаков идентифицируемых объектов. В наиболее простом варианте прогнозирование идентифицируемого параметра может производиться всего лишь по одному биометрическому признаку. Как правило, более точное прогнозирование достигается по нескольким, наиболее информативным признакам, отобранным из множества характеристик, описывающих идентифицируемые объекты.

Другой задачей судебно-медицинской антропологии следует назвать разработку и совершенствование методов классификации объектов судебно-медицинского экспертного познания, целью которой является установление принадлежности идентифицируемого объекта к одной из нескольких, заранее известных, взаимоисключающих групп (кластеров) [34].

Основой видовой структуры методов судебно-медицинской классификации является число групп, на которые кластерообразующий параметр разделяет множество объектов экспертного познания. В зависимости от количества указанных групп целесообразно выделять биномиальную и полиномиальную схемы судебно-медицинской классификации. Вторым фактором, определяющим видовую структуру методов судебно-медицинской классификации, является количество идентифицирующих признаков классифицируемых объектов. В наиболее простом варианте одномерная и, как правило, биномиальная классификация объектов может производиться всего лишь по одному признаку. Математической моделью одномерной биномиальной классификации является одномерный дискриминантный анализ [57].

Как правило, более точная классификация достигается по группе наиболее информативных признаков, отобранных из множества характеристик, описывающих классифицируемые объекты. Математическими моделями, лежащими в основе многомерной биномиальной и полиномиальной классификации, являются метод главных компонент, дискриминантный, канонический корреляционный, факторный и кластерный методы анализа [57].

Помимо указанных основных при проведении судебно-медицинских антропологических исследований целесообразным является использование широкого ряда других статистических методов, имеющих вспомогательное значение и преимущественно направленных на выявление латентной неоднородности исследуемых данных. В качестве указанных статистических методов следует назвать статистические процедуры выявления кластеринга, обнаружения выбросов, тесты проверки согласия эмпирических типов распределения с теоретическими, тесты сравнительного анализа.

Таким образом, в настоящее время в судебно-медицинской антропологии в зависимости от задач исследования, свойств изучаемых объектов и других исходных условий используется широкий спектр методов статистического анализа, преимущественно многомерных, в том числе специфичных только для данного научно-практического приложения. Указанное обстоятельство предъявляет особые требования к подготовке как исследователей – авторов способов судебно-медицинской идентификации, так и практических судебно-медицинских экспертов – потенциальных пользователей этих диагностических методик.

1.3. УСЛОВИЯ ДОСТОВЕРНОСТИ РЕЗУЛЬТАТОВ СТАТИСТИЧЕСКОГО АНАЛИЗА ПРИ СУДЕБНО-МЕДИЦИНСКОЙ ИДЕНТИФИКАЦИИ ЛИЧНОСТИ

Одним из основных факторов, гарантирующих общую достоверность результатов научных исследований в области судебно-медицинской антропологической идентификации (как, впрочем, и в любой другой области научного познания), является достоверность результатов статистического анализа данных. В этой связи следует отметить, что проверка обоснованности и правильности применения статистических методов обработки эмпирических данных считается важнейшим компонентом оценки качества биомедицинских научных работ [16,20,149]. Широкий спектр проведенных в различное время исследований показал, что при отсутствии специального статистического рецензирования независимо от профиля и ранга издания около 50% опубликованных биомедицинских научных статей содержат статистические ошибки [87,98,101,137,144]. Схожие данные получены и относительно научных работ в области судебной медицины [10].

Однако в литературе, посвященной судебно-медицинской антропологической идентификации, проблеме оптимальности и правильности применения методов статистической обработки научных данных уделяется неоправданно мало внимания. Имеющиеся работы указанной тематики немногочисленны и преимущественно посвящены отдельным вопросам оптимизации регрессионного анализа [28,51-53,60]. В целом же априорно считается, что выбор статистического метода всегда является адекватным научной задаче, а субъект научного познания обладает должной подготовкой в данной области. Между тем, применение тех или иных методов математической статистики в биомедицинских исследованиях – процесс пока субъективный, базирующийся больше на интуиции исследователя, чем на строго формализованном подходе [18].

Оптимальность выбора статистических методов определяется степенью соответствия исходных данных тем условиям, наличие которых предполагает математическая модель конкретного алгоритма статистического анализа. Совокупность указанных условий можно разделить на 2 группы: общие и частные.

Общим условием применимости любого из известных методов статистического анализа является представительность (репрезента-

тивность) выборок, которая, в первую очередь, определяется случайностью формирования (рандомизацией) последних [18,150]. Важным также является требование обобщаемости результатов исследования, которое зависит от степени соответствия изучаемой выборки характеристикам генеральной совокупности [16,18,20].

Частные условия адекватности статистических методов определяются соответствием изучаемых данных специфическим предпосылкам математических моделей конкретных аналитических процедур. Наиболее распространенными предположениями этого типа являются показания к применению параметрических статистических методов. Например, стандартный регрессионный анализ предполагает соответствие данных допущениям нормальности значений идентифицируемого параметра для множества значений идентифицирующих показателей, постоянства дисперсии остатков и их взаимной независимости [23,72]. Имеются также и менее известные, но не менее жесткие ограничения применимости статистических процедур. Например, краниометрическая идентификация пола, основанная на одномерной классификации, предусматривает условие нормальности распределения данных [64]. Однако выполненная нами путем анализа отношения размаха к стандартному отклонению указанных в этой работе биометрических данных (иные виды проверки невозможны) показала отсутствие нормальности распределения самого информативного из предложенных краниометрических показателей – ширины большого затылочного отверстия ($p < 0{,}025$). Данное несоответствие прямо влияет на достоверность идентификации пола. Так, при условии соответствия краниометрических данных нормальному распределению половая принадлежность черепа с шириной большого затылочного отверстия, равной 35,1 мм, практически достоверно ($p \geq 99{,}9\%$) является мужской. Но на самом деле данная вероятность гораздо меньше ($p \geq 88{,}3\%$).

В этой связи анализ достоверности результатов статистических методов должен являться обязательным компонентом критического оценивания субъектом экспертного познания любых опубликованных диагностических методик. Для судебно-медицинских экспертов, занимающихся выполнением экспертиз отождествления личности, изложенный тезис имеет особое значение, поскольку применяющиеся в судебно-медицинской антропологии многомерные статистические методы отличаются многообразием и особой сложностью.

1.4. ОСНОВНЫЕ ИСТОЧНИКИ БИОМЕТРИЧЕСКИХ ДАННЫХ ПРИ ПРОВЕДЕНИИ СУДЕБНО-МЕДИЦИНСКИХ АНТРОПОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ

В зависимости от источников получения биометрических данных различают 2 основных типа организации научных исследований: первичные и вторичные [18,20]. Первичные исследования представляют результаты анализа данных, полученные его авторами, вследствие чего именуются оригинальными. Во вторичных исследованиях обобщаются данные и делаются выводы на основе проведенных ранее другими или этими же авторами первичных исследований.

В зависимости от поставленных задач вторичные исследования подразделяются на несистематические обзоры, систематические обзоры, характеризующиеся жесткой методологией обобщения результатов первичных исследований, а также мета-анализы, обобщающие количественные данные нескольких исследований.

Из всех видов исследований обобщающего характера применение статистических методов возможно только в мета-анализах, которые представляют собой количественный синтез первичных данных с целью получения суммарных статистических показателей [18]. Следует отметить, что важнейшей чертой мета-анализов является использование таких заранее определенных критериев включения первичных исследований в анализ, как полнота данных, отсутствие явных недостатков в организации исследования и т.д. [18,20]. Именно поэтому мета-анализы закономерно располагаются на вершине иерархии научных исследований, а их результаты характеризуются наибольшей достоверностью.

В настоящее время в судебно-медицинской антропологии основным источником получения биометрических данных являются первичные оригинальные исследования [17,41,65]. Вместе с тем часть способов судебно-медицинской идентификации основана на результатах статистического синтеза ранее проведенных первичных исследований. Большая часть таких вторичных исследований анализирует «сырые» данные (сводки) только одного первичного исследования [36,37], и лишь единичные – большой группы первичных исследований [33].

Положительным моментом следует считать тот факт, что авторы части вторичных исследований использовали определенные крите-

рии включения наблюдений в анализ. Так, авторы диагностических моделей определения массы зольных останков из индивидуальной сводки M. Warren и W. Maples, содержавшей результаты кремации 100 трупов американцев, в собственный анализ включили только лишь 85 наблюдений [36]. Основным критерием включения в данном случае являлся возраст кремированного субъекта.

Помимо вторичных исследований, использующих «сырые» числовые массивы, в судебно-медицинской антропологии стали появляться работы, основывающиеся на анализе усредненных данных (точечных оценок параметров распределений биометрических показателей), приводимых в первичных исследованиях [57]. При этом авторы названного вторичного исследования также использовали заранее определенный критерий включения данных в анализ – их соответствие нормальному распределению.

Учитывая присутствие в указанных вторичных исследованиях помимо количественного компонента (статистической обработки данных первичных исследований) также и качественного компонента, представленного заранее определенными критериями включения первичных исследований в анализ, можно сделать вывод о появлении мета-анализов в структуре научных публикаций в области судебной медицины. Это свидетельствует о необходимости пересмотра сделанных ранее выводов об отсутствии подобных работ в судебно-медицинской научной литературе [9]. Вместе с тем отнесение анализируемых исследований к мета-анализам в определенной степени пока является условным.

Таким образом, основным методом получения биометрических данных в судебно-медицинской антропологии является проведение оригинальных первичных исследований. В меньшей степени для научных работ названной тематики характерен статистический синтез ранее проведенных первичных исследований как на основе использования заранее определенных критериев включения, так и без таковых. Особенностью вторичных исследований в области судебно-медицинской антропологии на сегодняшний день является анализ «сырых» данных ограниченного количества первичных исследований. Учитывая имеющиеся тенденции в проведении и оформлении результатов судебно-медицинских антропологических исследований, можно прогнозировать появление в ближайшем будущем полноценных мета-анализов в данном разделе судебной медицины.

1.5. ВИДЫ БИОМЕТРИЧЕСКИХ ПОКАЗАТЕЛЕЙ

При проведении судебно-антропологического исследования следует четко разграничивать свойства изучаемых биометрических признаков. Анализ специальной литературы показывает, что унифицированный подход к классифицированию научных данных в биомедицине до сих пор отсутствует, вследствие чего представителями различных биомедицинских дисциплин предложено довольно большое количество классификационных схем [3,16,18].

В значительной степени терминологическая путаница связана с классифицированием биометрических показателей в зависимости от типа шкалы измерений, использованной для регистрации данного признака. При этом различают следующие шкалы: номинальную, порядковую (ординальную), интервальную и относительную (шкалу отношения) [13,18,29].

Номинальная шкала используется только для категориальной классификации. Порядковая шкала позволяет ранжировать (упорядочить) объекты, указав, какие из них в большей или меньшей степени обладают определенным качеством. Интервальная шкала позволяет не только упорядочивать объекты измерения, но и численно выражать и сравнивать различия между ними. Относительная шкала очень похожа на интервальную, отличаясь от последней наличием определенной точки абсолютного нуля, благодаря которой можно судить о том, во сколько раз измеряемое свойство у одного объекта больше такового у второго объекта.

В соответствии с этим классификационным критерием выделяют номинальные, порядковые, интервальные и относительные биометрические показатели. Вместе с тем с позиции математической статистики следует различать только три типа биометрических показателей: количественные, качественные и порядковые [16,26,54].

Количественные параметры являются основным видом биомедицинских показателей. Значения количественных признаков можно упорядочить, кроме того, над ними можно производить арифметические действия. Предложено разделять количественные показатели на счетные и мерные [3].

Мерные признаки получают путем измерения идентифицируемых объектов. В судебно-медицинской антропологии преимущественно определяются линейные характеристики идентифицируемых объектов, область значений которых представлена множеством

всех неотрицательных рациональных чисел. Однако в связи с возрастающей ролью методов количественной гистологии при идентификации личности, в судебно-медицинской антропологии стали применяться поверхностные и объемные гистостереометрические показатели [5,62,65,77,81]. При этом объемные, а зачастую и многие поверхностные величины гистоструктур вычисляются с помощью обычных математических формул, применение которых основывается на сходстве формы изучаемых микрообъектов с определенной геометрической фигурой или геометрическим телом [3]. Поэтому область значений мерных биомедицинских показателей правильнее определять множеством всех неотрицательных действительных чисел. В отличие от относительных (качественных) показателей мерные признаки называют также абсолютными.

Счетные признаки получают путем подсчета. Основной сферой приложения счетных показателей в судебно-медицинской антропологии также являются количественные гистологические методы. Примерами счетных признаков служат средняя плотность остеоцитов, сосудов и остеонов в единице площади костной ткани [65,81], кроветворная активность фетальной печени и плотность расположения лимфоидных узелков фетальной селезенки [11,53]. Область значений счетных биометрических показателей представлена множеством всех неотрицательных целых чисел. В связи с тем, что счетные количественные показатели по физической природе (множество всех натуральных чисел или число нуль) не могут иметь дробную часть, для обозначения указанного типа данных был предложен термин «дискретные» [18]. Однако указанный термин нельзя признать удачным, поскольку, несмотря на свою дискретность, счетные признаки, в отличие от качественных и порядковых показателей, в практических целях могут быть описаны только непрерывными типами распределений.

Качественные (альтернативные) признаки могут быть измерены только в терминах принадлежности к одному из двух или более взаимно исключающих классов (наличие или отсутствие). Наиболее частой разновидностью качественных показателей являются бинарные (дихотомные) переменные, кодирующие принадлежность к одному из двух взаимоисключающих классов. Качественные признаки не связаны между собой никакими арифметическими соотношениями, упорядочить их также нельзя. Единственный способ описания качественных признаков состоит в том, чтобы подсчитать

число объектов, имеющих одно и то же свойство, в этом случае областью определения качественных признаков является множество всех неотрицательных целых чисел. Кроме того, можно подсчитать, какая доля от общего числа объектов приходится на то или иное свойство. Тогда область значений относительных биометрических показателей будет представлена множеством рациональных чисел на замкнутом числовом промежутке от 0 до 1 или от 0 до 100 (в случае выражения величины признака в процентах). Некоторые авторы полагают, что исключением из описанного правила являются дихотомные показатели, два значения которых (как правило, обозначаемых числами 0 и 1) условно можно считать упорядоченными [26]. Часть исследователей из-за возможности выражения относительных переменных рациональными числами неправомерно относит указанные показатели к разряду количественных [18].

Примером качественного показателя в судебно-медицинской антропологии является относительная частота аномалий зубов [63]. В гистостереометрии широко используется такая разновидность качественных признаков, как относительные линейные, поверхностные и объемные показатели, например, относительные площади остеоида, жировой и миелоидной тканей в кости [65]. Существенным отличием от обычных качественных признаков является то, что в случае с относительными гистостереометрическими показателями определяется не доля микрообъектов, обладающих или не обладающих каким-либо свойством, а доля точек или делений системы анализа изображений, совпавших с профилями микрообъектов определенного типа [80].

Порядковые (ранговые) признаки позволяют распределить объекты в определенном порядке в зависимости от степени выраженности исследуемого свойства. Указанные признаки можно только упорядочить, производить арифметические действия над ними нельзя. Областью определения порядковых признаков является множество натуральных чисел, а область значений аналогична таковой для качественных переменных. Порядковые признаки лежат в основе полуколичественных оценочных шкал. Например, к порядковым признакам относятся показатели стертости зубов [64], показатель разрежения спонгиозы костных пластинок [86]. В настоящее время полуколичественные оценки находят все меньшее применение в судебно-медицинских антропологических исследова-

ниях, вытесняясь более точными количественными морфометрическими показателями.

В судебно-медицинской антропологии объекты научного познания чаще всего рассматриваются с точки зрения не одного, а нескольких признаков. Рассматриваемое множество признаков обозначается вектором x, имеющим k компонент, каждая из которых характеризует соответствующий признак x_j, $j = 1,2,\ldots,k$.

Таким образом, объектом судебно-антропологического исследования является система k случайных одномерных величин, называемая также k-мерной случайной величиной $(x_1, x_2, \ldots, x_k)$.

В зависимости от типа компонент различают непрерывные k-мерные случайные величины, все компоненты которых – непрерывные одномерные случайные величины (количественные биометрические показатели), дискретные k-мерные случайные величины, все компоненты которых дискретные (качественные и порядковые биометрические показатели) и смешанные k-мерные случайные величины, среди компонентов которых есть как дискретные, так и непрерывные случайные величины.

Разделение биометрических показателей на перечисленные классы необходимо в связи с их подчинением разным видам вероятностных распределений, обладающих различными математико-статистическими свойствами. Так, количественные признаки описываются различными видами непрерывных распределений. Распределение мерных показателей, как правило, подчиняется нормальному закону. Это является преимуществом мерных параметров, так как нормальное распределение обладает рядом благоприятных статистических свойств. Значения счетных показателей далеко не всегда подчиняются нормальному распределению. Качественные и порядковые признаки характеризуются каким-либо видом дискретных распределений, преимущественно биномиальным.

Изложенное означает, что принадлежность исследуемых биометрических данных определенному типу показателей, характеризующемуся специфическим законом распределения, полностью определяет весь спектр возможных методов описания данных, планирования оптимального объема наблюдений и дальнейшего статистического анализа. Наиболее предпочтительным является использование количественных данных, обладающих наибольшей информативностью и допускающих применение статистических методов анализа, характеризующихся наибольшей чувствительностью.

1.6. ОЦЕНИВАНИЕ ПАРАМЕТРОВ НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ

Исчерпывающей характеристикой генеральной совокупности признаков идентифицируемых объектов является функция плотности распределения. При рассмотрении большинства математических моделей статистических методов в судебно-медицинской антропологии предполагается нормальное распределение всех или некоторых признаков генеральной совокупности.

Нормальное распределение может рассматриваться как один из фундаментальных законов природы [13]. Нормальность распределения следует ожидать в тех случаях, когда исследуемые параметры подвержены влиянию многих независимых, примерно в равной степени влияющих факторов, при большом числе измерений и отсутствии их предварительного отбора.

Считается, что непрерывная k-мерная случайная величина распределена нормально, если плотность распределения имеет вид

$$p(x) = \left[(2\pi)^k |\Sigma|\right]^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}, \qquad (1)$$

где $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix}$ - k-мерный вектор математических ожиданий;

Σ - ковариационная матрица

$$\Sigma = M(x_{ij} - \mu_j)(x_{ij} - \mu_j)^T = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1j} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2j} & \cdots & \sigma_{2k} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \sigma_{i1} & \sigma_{i2} & \cdots & \sigma_{ij} & \cdots & \sigma_{ik} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_{kj} & \cdots & \sigma_{kk} \end{pmatrix},$$

Σ^{-1} - матрица, обратная ковариационной матрице Σ размерности $(k \times k)$; $|\Sigma|$ - определитель этой матрицы [26].

Матрица Σ является симметрической и положительно определенной.

Отсюда многомерный нормальный закон распределения определяется вектором математических ожиданий μ и ковариационной матрицей Σ, элементы главной диагонали которой $\sigma_{11}, \sigma_{22}, \ldots, \sigma_{kk} = \sigma_j^2$ представлены дисперсиями j-х компонент вектора $x = (x_1, x_2, \ldots, x_k)$, а остальные элементы – коэффициентами ковариации i-й и j-й компонент данного вектора. При этом коэффициентом ковариации нормированных случайных величин называется коэффициент парной корреляции

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}.$$

Таким образом, общее количество параметров многомерной нормально распределенной генеральной совокупности равняется $\left[k + \frac{k(k+1)}{2} \right]$.

При одномерном нормальном законе распределения $k = 1$, $\Sigma = \sigma_{11} = \sigma^2$. Тогда $|\Sigma| = \sigma^2$, а $\Sigma^{-1} = \frac{1}{\sigma^2}$. Отсюда из выражения (1) получаем плотность одномерного нормального распределения, зависящего от двух параметров: математического ожидания μ и стандартного отклонения σ:

$$P(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}.$$

Поскольку на практике анализ всей совокупности идентифицируемых объектов, как правило, невозможен, да и не нужен, оценка биометрических показателей обычно производится на основании изучения свойств ограниченных выборок. Полученное на основании изучения ограниченной выборки числовое значение интересующего биометрического параметра всегда отличается от его истинного значения, наиболее полно отражающего в количественном отношении свойства исследуемой генеральной совокупности идентифицируемых объектов. Поэтому главная задача субъекта научного судебно-медицинского антропологического исследования состоит в том, чтобы сделать максимально правдоподобные выводы о свойствах и характеристиках гипотетической генеральной совокупности идентифицируемых объектов на основе доступной части данных этой совокупности.

Оценками истинных количественных параметров генеральной совокупности изучаемых объектов могут служить различные статистики: арифметическая средняя, дисперсия, медиана и др. Для биометрического исследования целесообразно использовать точечную оценку, обладающую наилучшими качествами. В статистическом анализе качество статистик характеризуют четыре критерия: несмещенность, эффективность, состоятельность и достаточность.

Статистика считается несмещенной, если все выборочные значения располагаются симметрично относительно истинного значения оцениваемого параметра [73,151,153]. Критерий эффективности характеризует минимальность стандартной ошибки статистики, используемой в качестве точечной оценки параметра генеральной совокупности, то есть стандартная ошибка эффективной оценочной статистики должна быть меньше стандартной ошибки любой другой статистики, выбираемой в качестве точечной оценки [73,78]. Оценка истинного значения параметра является состоятельной, если по мере увеличения объема выборки ее значение приближается к истинному значению параметра [26,73,78]. Оценка является достаточной, если при ее вычислении используется вся содержащаяся в выборке информация [91,92,109].

Выборочная средняя чаще всего является наилучшей оценкой генеральной средней, удовлетворяющей всем четырем критериям. Поэтому в качестве точечной оценки исследуемого морфометрического параметра чаще всего используется именно выборочное среднее. Критерием эффективности указанной статистики, то есть погрешности исследования, служит стандартная ошибка среднего. Следует отметить, что лучшими оценками для параметров генеральной совокупности указанные статистики бывают не всегда, например, в случае распределения значения морфометрического параметра с выраженной асимметрией, когда в качестве точечной оценки приходится использовать медиану.

Выборку объема n из k-мерной генеральной совокупности X можно представить в виде матрицы данных

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \qquad (2)$$

строки которой рассматриваются как n независимых реализаций k-мерного случайного вектора. Тогда точечными оценками вектора математических ожиданий является k-мерный вектор выборочных средних $\bar{x}_l = \frac{1}{n}\sum_{i=1}^{n} x_{il}$, $l = 1,2,\ldots,k$. Несмещенной оценкой ковариационной матрицы Σ является матрица

$$S = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1k} \\ s_{21} & s_{22} & \cdots & s_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ s_{k1} & s_{k2} & \cdots & s_{kk} \end{pmatrix}, \qquad (3)$$

где $s_{lj} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{il} - \bar{x}_l)(x_{ij} - \bar{x}_j)$, $l, j = 1,2,\ldots,k$.

Поскольку полное совпадение точечной оценки с истинным значением параметра генеральной совокупности маловероятно, на практике необходимо рассчитывать интервальные оценки, определяющие интервал, внутри которого с известной вероятностью находится истинное значение параметра.

Следует отметить, что вычисления интервальных оценок генерального среднего различаются в зависимости от конечности и объема исследуемой генеральной совокупности [78]. Однако в судебно-медицинской антропологии любые объекты исследования гипотетически можно считать членами бесконечных генеральных совокупностей. Поэтому дальнейшее изложение приводится без учета поправки на конечность генеральной совокупности.

Для среднего одномерной нормально распределенной генеральной совокупности $100(1-\alpha)$- процентный доверительный интервал определяется из выражения

$$\mu \in \bar{x} \pm t_{\alpha;n-1} \cdot \frac{s}{\sqrt{n}},$$

где $t_{\alpha;n-1}$ - значение двустороннего критерия Стьюдента при уровне значимости α и $v = n-1$ количестве степеней свободы.

При построении доверительной области для вектора средних многомерной нормально распределенной генеральной совокупности используется статистика T^2 Хотеллинга:

$$T^2 = n(\bar{x} - \mu)S^{-1}(\bar{x} - \mu) = \frac{k(n-1)}{n-k} F_{\alpha;k;n-k},$$

где $F_{\alpha;k;n-k}$ - значение критерия Фишера при уровне значимости α, $v_1 = k$ и $v_2 = n - k$ количестве степеней свободы [26].

Доверительный интервал для стандартного отклонения может быть построен на основании χ^2 - распределения:

$$s \cdot \sqrt{\frac{n-1}{\chi^2_{\alpha/2;n-1}}} < \sigma < s \cdot \sqrt{\frac{n-1}{\chi^2_{1-\alpha/2;n-1}}},$$

где $\chi^2_{1-\alpha/2;n-1}$ - значение двустороннего критерия χ^2 - критерия при уровне значимости α и $v = n - 1$ количестве степеней свободы [30].

Нормальное распределение значений изучаемых количественных показателей при проведении судебно-медицинских антропологических исследований встречается часто, но далеко не всегда. Определение же параметров анормально распределенных совокупностей с помощью математических моделей, основанных на несоответствующей реальности гипотезе, приведет к неверным результатам.

Для проверки нормальности распределения разработано большое количество методов, которые можно разделить на две группы.

Первую группу оставляют методы визуализации данных и оценивание дескриптивных статистик. Несмотря на свою простоту эффективность данных методов весьма высока. Наименее трудоемкими являются визуальные методы, например, построение гистограмм, менее распространено использование нормальных вероятностных графиков [135]. Для проверки того, может ли быть полученная совокупность значений морфометрического параметра приближенно аппроксимирована нормальным распределением, на построенной гистограмме достаточно визуально оценить выполнение следующих условий:

- распределение должно быть унимодальным и симметричным;
- примерно 99% всех отклонений должны быть меньше 3*s*;
- примерно 95% всех отклонений должны быть меньше 2*s*;
- примерно 68% всех отклонений должны быть меньше *s*.

Если указанные условия выполняются, то распределение близко к нормальному и его можно описать при помощи среднего и стандартного отклонения.

Из дескриптивных статистик кроме показателей центральной тенденции и вариации признака для проверки нормальности распределения наиболее важными являются коэффициенты асимметрии и эксцесса и их стандартные ошибки [30,94].

Наиболее объективными являются численные методы проверки нормальности, основанные на оценивании степени согласия эмпирического распределения с теоретически нормальным. К основным из них следует отнести χ^2- критерий согласия, критерий согласия Колмогорова-Смирнова, критерий Колмогорова-Смирнова в модификации Лиллиефорса и тест Шапиро-Уилка [13,126,132]. Как правило, основная трудность состоит не в том, какой из перечисленных тестов выбрать, а в том, что объем выборки слишком мал, чтобы применить любой из них [16]. Например, в литературе приводятся данные, что основные критерии согласия имеют существенные ограничения по объему выборки: $n > 30$ и $n > 50$ для χ^2- критерия и критерия Колмогорова-Смирнова соответственно [29].

При проверке нормальности выборки часто руководствуются следующим принципом Р.А. Фишера: «Отклонения от нормального вида, если только они не слишком заметны, можно обнаружить лишь для больших выборок, однако сами по себе эти отклонения вносят малое отличие в статистические критерии и другие вопросы» (цит. по [13]). Вместе с тем в литературе имеются данные о том, что для анализа выборок объемом 3-50 наблюдений можно эффективно использовать критерий Шапиро-Уилка [29]. Наш собственный практический опыт показал эффективность использования χ^2- критерия для выборок объемом менее 30 наблюдений.

Относительным недостатком критериев согласия является зависимость их статистик от количества степеней свободы (числа категорий значений признака), выбираемого исследователем. Кроме того, для проверки правильности использования критериев согласия необходим полный набор эмпирических данных. Это затрудняет проверку результатов указанных тестов при анализе научных публикаций, в которых полный набор исходных данных из-за их громоздкости обычно не приводится.

Перечисленных недостатков лишен способ быстрой проверки выборки на нормальность, основанный на отношении размаха к стандартному отклонению [141]. Авторы указанной работы исследовали распределение отношения R/s для выборки объемом n из нормально распределенной генеральной совокупности и табулировали его критические границы.

Для описания данных, не подчиняющихся нормальному закону, лучше воспользоваться не средним, а медианой и процентилями.

1.7. ОЦЕНИВАНИЕ ПАРАМЕТРОВ БИНОМИАЛЬНОГО РАСПРЕДЕЛЕНИЯ

Наиболее частым видом дискретных распределений, используемым при проведении судебно-медицинских исследований, посвященных разработке способов идентификации личности, является биномиальное распределение. Считается, что случайная дискретная величина x подчиняется биномиальному распределению, если выполняются следующие условия, называемые свойствами независимых испытаний Бернулли [16]:

1. Каждое отдельное наблюдение имеет ровно два возможных взаимно исключающих исхода.

2. Вероятность данного исхода одна и та же для любого наблюдения.

3. Все наблюдения независимы друг от друга.

Точные доверительные границы биномиально распределенной генеральной совокупности определяются по формулам:

$$\pi_B = \frac{(x+1)\cdot F}{n-x+(x+1)\cdot F} \text{ при } F_{\{v_1=2\cdot(x+1),v_2=2\cdot(n-x)\}},$$

$$\pi_H = \frac{x}{x+(n-x+1)\cdot F} \text{ при } F_{\{v_1=2\cdot(n-x+1),v_2=2x\}},$$

где π_B и π_H - соответственно верхняя и нижняя односторонние доверительные границы; x – число положительных выборочных наблюдений ($x = pn$); F - значения критерия Фишера для выбранной доверительной вероятности в зависимости от обоих чисел степеней свободы [30].

Особый случай составляют нуль-событие и полное событие. Точную верхнюю интервальную оценку доли генеральной совокупности при $p = 0$ (нуль-событие) можно также вычислить с помощью выражения $\pi_B = 1-\sqrt[n]{\alpha}$. Для $p = 1$ (полное событие) нижняя граница определяется формулой $\pi_H = \sqrt[n]{\alpha}$ [30]. Важно, что точные доверительные интервалы асимметричны относительно выборочной оценки p. Симметричными доверительные границы являются только при $p = 0{,}5$.

Как свидетельствуют данные специальной литературы, методы вычисления точных доверительных границ для биномиальных величин мало известны в среде исследователей, занимающихся проблемами идентификации личности. Гораздо более широкое распро-

странение для решения указанной задачи получили методы, основанные на аппроксимации биномиального распределения нормальным с помощью выражения

$$p = \hat{p} \pm z_{0,95} \cdot \sqrt{\hat{p}(1-\hat{p})/n}, \qquad (4)$$

где p и $\hat{p}$ – неизвестная истинная относительная частота наблюдений случайного признака и ее точечная оценка; z – стандартная нормальная переменная при указанном уровне статистической надежности; n – количество наблюдений.

Из формулы (4) видно, что доверительные границы частотного показателя, основанные на данной аппроксимации, всегда являются симметричными относительно его выборочной оценки. Поэтому выражение (4) служит хорошим приближением биномиального распределения лишь при больших объемах выборок и условии отсутствия слишком больших и слишком малых относительных частот наблюдаемого признака, то есть при $n\hat{p}$ или $n(1-\hat{p}) > 5$ [16]. Кроме того, даже при использовании аппроксимации нормальным распределением для бесконечно больших генеральных совокупностей вводят поправку на непрерывность $\frac{1}{2\pi}$, а для конечных генеральных совокупностей – поправку на конечность $\sqrt{\frac{N-n}{N-1}}$ [30].

В этой связи нами было проведено исследование, целью которого явилась проверка правильности определения доверительных интервалов для относительных частот признаков в научных исследованиях в области судебной медицины.

Объектами анализа явились 314 оригинальных исследований, опубликованных отечественными авторами в журнале «Судебно-медицинская экспертиза» за период 2001-2005 гг. Протокол исследования включал выявление всех исследований, выводы которых опирались на результаты применения каких-либо методов аналитической статистики, и последующее выявление исследований, в которых осуществлялось определение доверительных интервалов для относительных частот признаков. В исследованиях последней группы отмечались указания авторов о методе расчета доверительных интервалов, объем выборочных данных и число итераций расчетов для каждой выборки. На заключительном этапе нами производилась проверка правильности определения доверительных интервалов путем расчета точных доверительных границ. Статистиче-

ская обработка полученных данных осуществлялась с использованием приложений Microsoft Excel пакета Microsoft Office 2003 и Statistica (StatSoft) версии 6.0.

Проведенный анализ показал, что определение доверительных интервалов для долей производилось всего лишь в двух оригинальных исследованиях, составивших 0,6% от всех изученных статей и 2,4% от всех работ, в которых для обработки данных использовалась аналитическая статистика. В обеих указанных статьях описывались популяционные исследования различных генетических маркеров. При этом авторы данных исследований рассчитывали частоты аллелей определенных локусов исходя из количества каждого генотипа в исследованной популяции, после чего определяли их 95% интервальные оценки. В одном исследовании производился расчет только верхних [69], а в другом – и верхних, и нижних доверительных границ [68]. Авторы указанных исследований оперировали достаточно большими объемами эмпирических данных и рассчитывали интервальные оценки для большого числа аллелей каждого локуса (табл. 1). Во всех итерациях использовалась одинаковая процедура определения доверительных интервалов относительных частот аллелей, основанная на аппроксимации биномиального распределения нормальным распределением по формуле (4). В обеих рассматриваемых статьях отсутствовало обоснование показаний к использованию названного метода определения доверительных интервалов.

Проведенная проверка показала, что в обоих исследованиях имелось достаточно большое (19% [69] и 43% [68]) количество аллелей, характеризовавшихся слишком малыми относительными частотами ($n\hat{p} < 5$). Вследствие этого приведенные в названных работах доверительные интервалы относительных частот таких аллелей являются ошибочными (табл. 2).

Таким образом, определение доверительных интервалов для относительных частот признаков в судебно-медицинских научных исследованиях, посвященных проблемам идентификации личности, применяется неоправданно редко, практически только в популяционных исследованиях генетических маркеров и по выборочным данным в 100% случаев является методически ошибочным. Ошибочность расчетов доверительных интервалов обусловлена необоснованным использованием аппроксимации нормальным распределением при слишком малых относительных частотах признаков.

Указанные ошибки определения доверительных интервалов для относительных частот признаков могут быть устранены только с помощью методов расчета точных доверительных границ.

Таблица 1

Характер судебно-медицинских данных, используемых при определении доверительных интервалов для частот признаков

Показатель	$\bar{x}$	s	$\tilde{x}$	$x_{\min}$	$x_{\max}$	R	n
Объем выборки	358,3	164,1	308	238	680	442	6
Число итераций*	9,3	7,7	6	5	25	20	6

Примечание. * - под числом итераций подразумевается сумма расчетов одно- или двусторонних доверительных интервалов для аллелей одного локуса (относительного частотного показателя одной выборки).

Таблица 2

D1S80 аллельные частоты в кавказоидной популяции Уральского региона России [68]

Аллель	$\hat{p}$	$n\hat{p}$	Нижние и верхние 95% интервальные оценки p					
			Аппроксимированные		Точные		Ошибка, %	
15	**0,001**	**0,68**	**0**	**0,003**	**0**	**0,0082**	**0**	**517**
17	**0,003**	**2,04**	**0**	**0,007**	**0,0004**	**0,0106**	**12**	**119**
18	0,282	191,76	0,248	0,316	0,2488	0,3178	0	1
20	0,024	16,32	0,012	0,036	0,0135	0,0379	6	8
21	0,018	12,24	0,008	0,028	0,0092	0,0306	6	15
22	0,037	25,16	0,023	0,051	0,0239	0,0538	3	8
23	**0,004**	**2,72**	**0**	**0,009**	**0,0009**	**0,0128**	**23**	**96**
24	0,331	225,08	0,295	0,366	0,2956	0,3677	0	1
25	0,075	51	0,055	0,095	0,0563	0,0974	2	3
26	0,019	12,92	0,009	0,029	0,0102	0,0325	6	18
28	0,053	36,04	0,036	0,070	0,0374	0,0725	3	5
29	0,034	23,12	0,020	0,048	0,0216	0,0503	5	7
30	0,022	14,96	0,011	0,033	0,0124	0,0361	6	14
31	0,069	46,92	0,050	0,088	0,0512	0,0909	2	4
37	0,009	6,12	0,002	0,016	0,0032	0,0191	14	35

Примечание. Жирным шрифтом выделены аллели со слишком малыми относительными частотами, не позволяющими использовать для определения доверительных границ аппроксимацию (4). Точные интервальные оценки частот аллелей 19,27,33,35,36,39,40 не рассчитывались ввиду их совпадения с частотами аллелей 15,17 и 23.

1.8. СТАТИСТИЧЕСКАЯ ОБРАБОТКА РЕЗУЛЬТАТОВ ИЗМЕРЕНИЙ ПРИ СУДЕБНО-МЕДИЦИНСКОЙ АНТРОПОЛОГИЧЕСКОЙ ИДЕНТИФИКАЦИИ

Для судебно-медицинской антропологической идентификации является характерным широкое использование измерительных методов, дающее возможность получения разносторонней количественной информации об исследуемых биологических структурах.

При проведении биометрического исследования результаты измерения всегда отличаются от истинных значений изучаемых морфологических параметров. Основным источником погрешностей, возникающих при фиксации свойств объектов как на этапе научного, так и в ходе экспертного исследования, являются погрешности измерения. Погрешности данного вида, называемые также инструментальными, связаны с точностными данными измерительных приборов, техническими особенностями конкретного измерительного метода, а также навыками исследователя. Являясь по своему характеру случайными, инструментальные погрешности достаточно легко контролируются исследователем, а существующие методы математической обработки результатов измерений позволяют сводить влияние этих ошибок к минимуму [45].

Однако в связи с созданием и внедрением в практику большого количества способов судебно-медицинской идентификации личности, основанных на методах количественной гистологии [5,28,62,77,81], при проведении метрологических процедур возник дополнительный, более существенный тип погрешностей, связанный с выбором оптимального количества единиц наблюдений.

В отличие от организменного и органного уровней изучения при гистологических исследованиях выбор оптимального количества единиц наблюдений значительно усложняется. Так, при антропометрии каждому объекту наблюдения соответствует только одно значение изучаемого признака. При микроскопическом исследовании только уже на светооптическом уровне для каждого исследуемого органа (единицы наблюдения 1 уровня) можно получить совокупность значений изучаемого признака, величина которой определяется количеством гистоструктур (единиц наблюдения 2 уровня) с анализируемым признаком в изучаемом органе или ткани [80].

В связи с наличием в пределах одного органа или ткани большого количества однотипных микрообъектов, количественная оценка

какого-либо гистометрического или гистостереометрического параметра обычно является значением, полученным на основании изучения массива числовых данных. Поскольку анализ всей совокупности микрообъектов, как правило, невозможен, оценка морфометрических показателей обычно производится на основании изучения свойств ограниченных выборок. Полученное на основании изучения ограниченной выборки микрообъектов числовое значение интересующего морфометрического параметра всегда отличается от его истинного значения, наиболее полно отражающего в количественном отношении свойства исследуемой генеральной совокупности гистоструктур данного органа или ткани. Поэтому главная задача исследователя-морфолога состоит в том, чтобы сделать максимально правдоподобные выводы о свойствах и характеристиках гипотетической генеральной совокупности гистоструктур на основе доступной части данных этой совокупности. При этом всегда существует риск, что эти выводы будут неправильными ввиду неполноты использованной информации. Отсюда возникает проблема количественных оценок степени этого риска. Наиболее адекватным научным подходом в данном случае является использование методов математической статистики.

При проведении количественного гистологического анализа исследователь всегда сталкивается с ситуацией, когда, с одной стороны, для уменьшения величины возникающих погрешностей требуется увеличение количества микрообъектов, подлежащих изучению, а с другой – необходимо уменьшение объема наблюдений для снижения трудоемкости морфометрии. В связи с этим планирование оптимального количества микрообъектов, анализ которого позволяет оценить нужный морфометрический параметр с минимальной, заранее известной, погрешностью, является важной самостоятельной проблемой любого научного и экспертного судебно-антропологического исследования, только успешное решение которой обеспечит его достоверность и репрезентативность [80].

В настоящее время имеется большое количество публикаций, подробно освещающих ошибки организации количественного гистологического исследования [2,3]. Вместе с тем в биомедицинской литературе имеется мало работ, посвященных детальному изучению возможных погрешностей измерения при проведении любых метрологических процедур, в том числе и гистометрических исследований. Кроме того, актуальной остается проблема планирования

объема наблюдений, поскольку применяющиеся математические методы не учитывают всего разнообразия ситуационных задач, возникающих при проведении гистометрического или гистостереометрического исследования. В связи с этим приводим попытку анализа вышеизложенных проблем и возможных путей их устранения.

При проведении морфометрического исследования следует четко разграничивать показатели дисперсии единиц наблюдений различных уровней. Подобное разграничение требует отдельного планирования подлежащего регистрации объема наблюдений для каждого уровня изучения. Например, нами было проведено изучение зависимости кроветворной активности фетальной печени от гестационного возраста путем определения численной плотности ядер гемопоэтических клеток в тестовом поле зрения микроскопа [11,53]. Задача исследования определила обязательность раздельного планирования количества полей зрения (единиц наблюдений 2 уровня), подлежащих изучению в конкретной печени для оценки средней величины ее кроветворной активности, и количества плодов разного гестационного возраста (единиц наблюдений 1 уровня), средние значения активности печеночного гемопоэза которых подверглись регрессионному анализу. Указанная необходимость была вызвана наличием выраженной вариабельности значений кроветворной активности в пределах одного органа, а также вариабельности средних значений кроветворной активности фетальной печени у плодов одного гестационного возраста. В данной ситуации планирование количества полей зрения печеночной паренхимы, подлежащих изучению, гарантировало заблаговременное определение величины возможных погрешностей при оценке среднего значения кроветворной активности отдельно взятой печени и выбор минимально допустимого их уровня.

Применительно к любым судебно-медицинским антропологическим исследованиям необходимый объем выборочной совокупности единиц наблюдений 1 уровня определяется исходя из задач конкретного исследования в соответствии с показателями чувствительности и специфичности используемых статистических критериев. Наоборот, описательные характеристики выборочных совокупностей единиц наблюдения 2 уровня являются одной из составляющей погрешности, возникающей при проведении измерительных гистологических процедур. Вследствие этого проведение любого количественного гистологического исследования требует ис-

пользования специальных методов планирования оптимального количества микрообъектов, подлежащих количественному оцениванию. Под оптимальностью в данном случае понимается обеспечение требуемого уровня точности гистометрии или гистостереометрии при их минимальной трудоемкости.

Отправной точкой любого количественного гистологического анализа является поисковое исследование, дающее представление о характере и выборочных оценках распределения изучаемого количественного параметра микрообъектов данного органа. Такое поисковое исследование необходимо даже при наличии литературных данных по соответствующей проблеме. Выбор числа микрообъектов, подлежащих морфометрическому оцениванию в ходе предварительного исследования, может производиться произвольно.

На следующем этапе нужно выяснить, к какому типу относится распределение полученных значений изучаемого морфометрического параметра. Если распределение близко к нормальному, то его следует описать при помощи среднего и стандартного отклонения.

Поскольку среднее генеральной совокупности находится между доверительными границами $\mu \in \bar{x} \pm t_{\alpha;n-1} s_{\bar{x}}$, где $s_{\bar{x}}$ – стандартная ошибка среднего, то необходимое минимальное количество наблюдений с максимально допустимой стандартной ошибкой (за величину которой лучше взять 1/4 допустимого двустороннего доверительного интервала) вычисляется из выражения $n_{\bar{x}} = s^2 / s_{\bar{x}}^2$.

Например, оценим минимальный объем выборки наблюдений, требующийся для обеспечения заданного уровня погрешности. Пусть в ходе предварительного исследования выборки объемом $n = 40$ получены следующие выборочные оценки нормально распределенной совокупности значений изучаемого морфометрического параметра: выборочное среднее $\bar{x} = 70$ и выборочное стандартное отклонение $s = 20$. Стандартная ошибка определяется как $s_{\bar{x}} = 20/\sqrt{40} \approx 3{,}16$. Тогда двусторонний 95% доверительный интервал для среднего находится в пределах $\mu \in 70 \pm 6{,}4$. Выберем максимально допустимую стандартную ошибку, равную 2. Тогда минимально необходимый объем выборки для достижения выбранного уровня стандартной ошибки будет равняться $n_{\bar{x}} = s^2 / s_{\bar{x}}^2 = 20^2 / 2^2 = 100$.

Если требуется оценить только максимальную величину возможной погрешности, вычисление ее производится по формуле:

$\varepsilon_{\max} = t_{\alpha/2;n-1} \cdot s_{\bar{x}}$, то есть определяется односторонний доверительный интервал, для расчета которого используется табличное значение двустороннего критерия Стьюдента при тех же степенях свободы, но с уровнем значимости, равным $\alpha/2$.

Определим 90% односторонний доверительный интервал для среднего значения морфометрического параметра из предыдущего примера. Поскольку $s_{\bar{x}} = 3{,}16$, то максимальная величина возможной погрешности с 90% статистической надежностью равняется $\varepsilon_{\max} = t_{0,05;39} \cdot 3{,}16 = 2{,}023 \cdot 3{,}16 = 6{,}4$. То есть полученная со статистической надежностью 90% односторонняя верхняя доверительная граница соответствует верхней границе 95% двустороннего доверительного интервала для среднего значения изученного морфометрического параметра.

Определение оптимального объема выборки возможно также на основании использования величины среднего абсолютного отклонения, которое определяется по формуле:

$$CAO = \frac{\sum |x_i - \bar{x}|}{n} \text{ [30].}$$

При этом доверительный интервал для среднего генеральной совокупности имеет вид: $\mu \in \bar{x} \pm коэффициент \cdot CAO$.

Поскольку используемый в формуле коэффициент зависит только от объема исследованной выборочной совокупности, то необходимый объем наблюдений можно определить, подбирая соответствующую величину коэффициента. Значения коэффициента для различных объемов выборок и доверительных границ табулированы и приводятся в специальной литературе [119,129]. Например, при гистометрии 9 микрообъектов получены следующие выборочные оценки: $\bar{x} = 120$ мкм, CAO = 20 мкм. Табулированное значение коэффициента для определения 95% доверительных границ для среднего значения по выборке объемом 9 наблюдений составляет 1,00. Отсюда среднее генеральной совокупности с 95% надежностью находится в интервале $120 \pm 1{,}00 \cdot 20$ мкм. Тогда для уменьшения 95% доверительного интервала вдвое необходима выборка из 30 наблюдений (соответствующий коэффициент равен 0,48).

Недостатком всех изложенных способов является то, что планирование объема выборки осуществляется, опираясь на выборочные оценки, которые могут значительно измениться после изучения запланированного количества микрообъектов. Поэтому целесообраз-

но при определении оптимального объема выборки использовать не точечную, а верхнюю интервальную оценку стандартного отклонения [50]. Затем определяется необходимый объем наблюдений, обеспечивающий любой требуемый уровень погрешности:

$$N = \frac{t^2_{\alpha;n-1} \cdot s^2_B}{\varepsilon^2},$$

где N – оптимальный объем наблюдений для обеспечения абсолютной погрешности морфометрии величиной ε. Следует отметить, данный способ является достаточно жестким, вследствие чего после исследования запланированного объема наблюдений дисперсия значений изучаемого морфометрического параметра, скорее всего, будет меньше уровня, выбранного для расчетов. Точнее, вероятность того, что истинная дисперсия превысит выбранный для расчетов уровень, равна α, соответственно, точность исследования в $(100 - 2\alpha)$% случаев окажется выше запланированного уровня.

Например, в указанном выше исследовании степень кроветворной активности фетальной печени определялась путем подсчета количества кроветворных клеток в стандартном поле зрения микроскопа, равном $1{,}39 \cdot 10^{-2}$ мм^2. Распределение совокупности полученных значений подчинялось нормальному распределению. Допустим, при изучении фрагмента фетальной печени после сканирования 31 поля зрения среднее значение кроветворной активности равнялось 60 со стандартным отклонением, равным 16. Максимально приемлемой абсолютной погрешностью морфометрии выбран интервал $\mu \in \bar{x} \pm 3$ клетки гемопоэза в тестовом поле зрения.

Определим верхнюю границу 95 % доверительного интервала для стандартного отклонения. Поскольку $\chi^2_{0,95;30} = 18{,}49$, то

$$s_B = \sqrt{\frac{s^2 \cdot (n-1)}{\chi^2_{1-\alpha;\,n-1}}} = \sqrt{\frac{16^2(31-1)}{18{,}49}} = 20{,}38.$$

Учитывая, что табличное значение двустороннего критерия Стьюдента при уровне значимости $\alpha = 0{,}05$ и количестве степеней свободы $\nu = 31 - 1 = 30$ равняется 2,042, определим необходимый объем наблюдений для обеспечения заданной $(\varepsilon = 3)$ абсолютной величины погрешности морфометрии:

$$N = \frac{t^2_{\alpha;\,n-1} \cdot s^2_B}{\varepsilon^2} = \frac{2{,}042^2 \cdot 20{,}38^2}{3^2} \approx 192.$$

Таким образом, для обеспечения заданной ($\varepsilon = 3$) абсолютной величины погрешности морфометрии необходимо произвести подсчет миелоидных элементов в 192 тестовых полях зрения. После изучения запланированного количества наблюдений вероятность превышения рассчитанного уровня погрешности составляет не менее 0,25% (вероятность одновременного превышения интервальных оценок среднего и стандартного отклонения).

Иногда средние значения исследуемого морфометрического параметра в разных анатомических отделах органа значительно отличаются. Указанное обстоятельство может привести к тому, что данные морфометрии в каждом отдельно взятом фрагменте подчиняются нормальному распределению, а в совокупности по всем отделам органа – нет. Распределение значений морфометрического параметра на графике в этом случае может выглядеть в виде нескольких (в зависимости от количества изученных фрагментов органа) наслоившихся за счет сдвига вдоль оси абсцисс нормальных распределений похожей формы (рис. 1). В таких случаях нужно проверить совокупное распределение данных на подчинение нормальному закону. Если аппроксимация нормальным распределением возможна, то генеральное среднее допустимо оценивать по всей совокупности морфометрических данных [80].

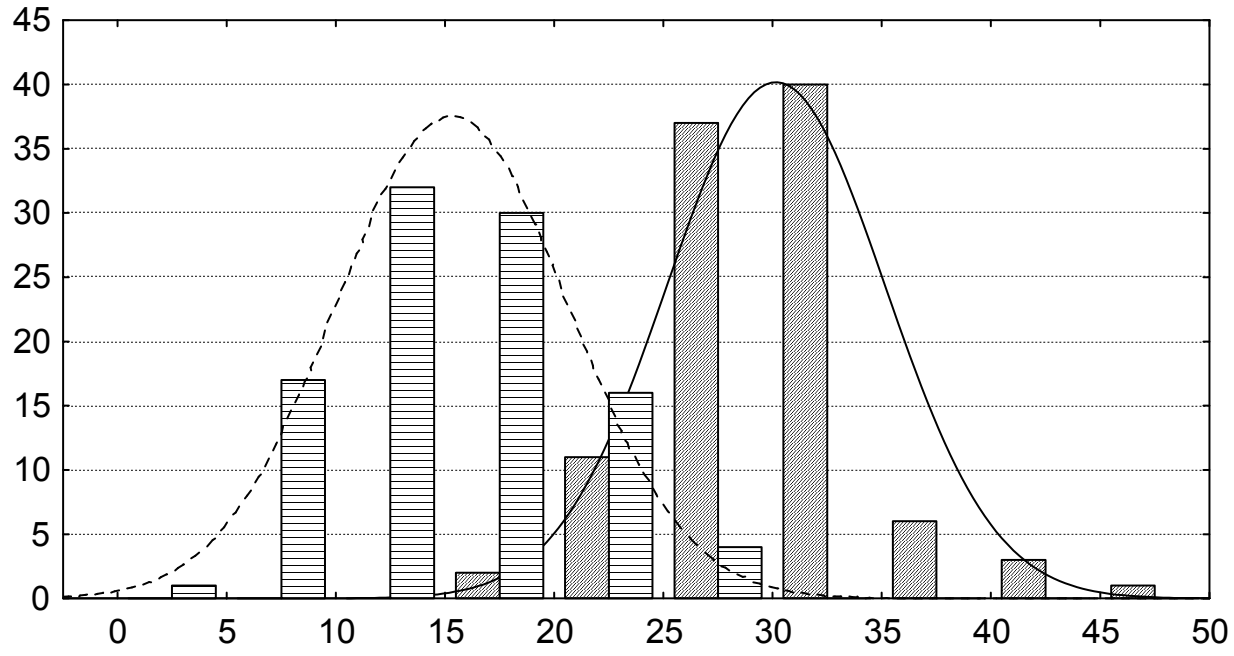

Рис. 1. Графическое изображение распределений значений кроветворной активности в разных отделах фетальной печени. По оси абсцисс - кроветворная активность, число ядер в тестовой площади; по оси ординат - их частоты.

В противном случае приходится планировать количество замеров для обеспечения необходимой погрешности исследования и определять выборочные средние отдельно для каждого фрагмента органа, причем величина $s_{\bar{x}}$ и, соответственно, погрешности исследования должна быть одинаковой для каждой серии измерений. После этого определяется среднее выборочных средних, которое служит оценкой среднего генеральной совокупности, а стандартное отклонение выборочных средних – стандартной ошибкой среднего, пригодной для определения погрешности морфометрии.

В случае неподчинения выборочной совокупности значений гистометрического показателя нормальному распределению, можно воспользоваться следующими тремя приемами [50].

Первый прием - стандартизация исследования с целью измерения не всех имеющихся в органе микрообъектов данного вида, а одного определенного микрообъекта. Выбор последнего определяется рядом дополнительно вводимых характеристик: конкретная локализация, максимальный или минимальный экстремум и т.д. Например, при измерении плотности сосудов в костной ткани можно не вычислять ее среднее значение на основании большой суммы замеров по нескольким фрагментам кости, а определить ее максимальную величину. В данном случае одному идентифицируемому объекту соответствует одно числовое значение, и проведение всех вышеизложенных этапов будет ненужным.

Второй прием - изменение иерархического уровня исследования для уменьшения разброса значений интересующего морфометрического параметра относительно выборочного среднего. Этим достигается уменьшение асимметрии и увеличение плотности расположения результатов измерений вокруг выборочного среднего, вследствие чего возможна аппроксимация полученных данных нормальным распределением. Например, при работе со счетными признаками для смены иерархического уровня исследования достаточно изменить единицу площади зрения микроскопа, в которой определялось количество элементов. Увеличить площадь зрения можно путем изменения увеличения микроскопа или смены окуляров.

Третий прием - определение средней величины показателя, его доверительной области и необходимого количества микроструктур на основании использования положений центральной предельной теоремы. Основываясь на принципах данной теоремы, можно показать, что если из совокупности значений какого-либо морфометри-

ческого показателя извлекать ограниченные выборки одинакового объема и определять их средние, то распределение выборочных средних будет подчиняться нормальному закону независимо от распределения исходной совокупности, причем, чем больше объем выборок, тем точнее приближение. Среднее выборочных средних будет являться характеристикой среднего исходной совокупности, а стандартное отклонение выборочных средних – стандартной ошибкой среднего, пригодной для расчета доверительных интервалов для среднего с помощью критерия Стьюдента и определения необходимого объема выборки по выше приведенному методу. При статистических заключениях считается, что использование центральной предельной теоремы дает приемлемые результаты, если объем выборки не меньше 30.

Например, известно, что в позднефетальном периоде выраженность миелопоэза в печени значительно уменьшается. Кроме того, по мере увеличения гестационного возраста отмечается переход от диффузного распределения кроветворных элементов к их группированию в очажки. При этом распределение значений кроветворной активности в позднефетальном периоде, полученных путем подсчета кроветворных элементов в $1{,}39 \cdot 10^{-2}$ мм^2 (площадь поля зрения микроскопа Биолам при увеличении 945^x) среза уже не подчинялось нормальному распределению. Например, можно было получить такую выборку объемом n = 9: 0;0;0;0;1;0;0;10;3 со средним $\bar{x} = 1{,}56$ и стандартным отклонением s = 3,32. Данную выборку нельзя считать извлеченной из нормально распределенной совокупности: стандартное отклонение более чем в 2 раза превышает среднее, при этом среди данных нет отрицательных значений (и не может быть по самой природе данных). Для приближения значений кроветворной активности нормальному распределению можно увеличить площадь среза, на единицу которой осуществлялся подсчет. Так, при увеличении площади поля зрения микроскопа в 5,05 раза (площадь поля зрения микроскопа Микмед-2 при увеличении 1000^x) среднее выборки из 9 полей зрения равнялось 7,78; стандартное отклонение – 3,99, среднее абсолютное отклонение – 3,41. Данные, полученные вторым способом, уже могут быть приближенно аппроксимированы нормальным распределением.

Доверительный интервал для медианы приходится определять, если не удалось достичь хотя бы приближенного соответствия совокупности значений исследуемого морфометрического параметра

нормальному распределению. При построении доверительного интервала для медианы сначала нужно упорядочить по величине полученные значения количественного параметра. Если упорядоченные значения обозначить $x_1, x_2, ..., x_{n,}$ то доверительные границы медианы, независимо от характера распределения исследуемого морфометрического параметра задаются формулой:

$$x_h \le \tilde{x} \le x_{n-h+1}$$ [30].

Для $n > 50$ значение h можно вычислить по формуле

$$h = \frac{n - z_\alpha \sqrt{n} - 1}{2}$$ [30].

Доверительные интервалы для относительных частотных показателей при $n\hat{p}$ или $n(1-\hat{p}) > 5$ удобно определять с помощью аппроксимации биномиального распределения нормальным распределением. Из выражения (4) легко получить формулу для определения минимального объема выборки, необходимого для обеспечения заданного уровня относительной погрешности Δ :

$$n_p = z^2 \cdot \hat{p}(1-\hat{p}) / \Delta^2.$$

Так как при увеличении числа степеней свободы t-распределение стремится к нормальному, критические значения z для любого уровня значимости можно найти в последней строке (при $v = \infty$) таблицы с критическими значениями t - критерия. Легко заметить, что n_p достигает максимума при максимуме $\hat{p}(1-\hat{p}),$ то есть при $\hat{p} = 0{,}5$. Поэтому в случаях, когда перед началом исследования отсутствуют данные о частоте обнаружения исследуемого признака, можно рассчитать необходимый объем выборки, принимая $\hat{p} = 0{,}5$.

Например, при разработке способа идентификации гестационного возраста нами рассматривалась возможность использования в указанных целях данные об изменении доли портальных трактов, содержащих желчные протоки. Требуемая относительная погрешность равнялась 1%. Тогда необходимое количество подлежащих изучению портальных трактов каждой печени с надежностью, равной 95%, составило $n_p = 1{,}96^2 \cdot 0{,}5 \cdot 0{,}5 / 0{,}01^2 = 9604$.

Для устранения погрешностей, связанных с количественной оценкой свойств исследуемой генеральной совокупности гистоструктур органа или ткани, целесообразно использование во всех случаях унифицированной тактической схемы, которая включает в себя следующие этапы:

- установление перечня морфометрических параметров, подлежащих количественной оценке;

- проведение поискового исследования для определения вида и свойств распределения изучаемого морфометрического параметра;

- определение необходимой степени точности исследования;

- определение минимального объема выборки, обеспечивающего получение количественной оценки исследуемого показателя генеральной совокупности микрообъектов (единиц наблюдения 2 уровня) данного органа или ткани (единицы наблюдения 1 уровня) с выбранной степенью точности;

- морфометрическое изучение определенного количества случайно выбранных микрообъектов в небольшой группе единиц наблюдения 1 уровня и фиксация результатов морфометрии;

- анализ полученных результатов и определение необходимого количества единиц наблюдения 1 уровня, исходя из задач исследования и вида последующих методов статистического анализа.

Для повышения воспроизводимости результатов количественных гистологических исследований следует учитывать также погрешности, обусловленные влиянием усадки тканей, неоднородности толщины срезов и аналитических тест-систем [54].

Таким образом, приведенные математические способы позволяют определять минимальное количество микрообъектов, необходимое для получения достоверной количественной оценки интересующего морфометрического параметра исследуемого органа с любой требуемой погрешностью. Возможности методов охватывают не только морфометрические показатели, подчиняющиеся нормальному закону, но и параметры, характеризующиеся другими видами непрерывных распределений, а также различными дискретными распределениями. Однако использование любых, даже самых сложных, математических приемов не освобождает от необходимости обеспечения рандомизации исследования и контроля погрешностей измерительной аппаратуры.

ГЛАВА 2. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

2.1. ЗАДАЧИ КОРРЕЛЯЦИОННОГО АНАЛИЗА ПРИ СУДЕБНО-МЕДИЦИНСКОЙ ИДЕНТИФИКАЦИИ ЛИЧНОСТИ

Одним из приоритетов судебно-медицинской антропологии является разработка и совершенствование способов прогнозирования каких-либо параметров идентифицируемых объектов, выраженных количественными признаками. К числу основных идентифицируемых параметров, обладающих таким свойством, относятся биологический возраст, длина тела, различные соматические размеры человека. Этим судебно-медицинское прогнозирование отличается от судебно-медицинской классификации, под которой понимается отнесение объекта экспертного познания к одному из нескольких взаимоисключающих классов, которые не могут быть описаны количественно (пол, соматотип, массивность скелета, порядковая локализация однотипных костей и др.).

Задачей научных исследований, посвященных созданию способов судебно-медицинского прогнозирования, являются поиск и изучение взаимосвязей между идентифицируемыми параметрами и идентифицирующими признаками и последующее построение математической модели, с определенной точностью позволяющей прогнозировать значение идентифицируемого параметра по значению идентифицирующего признака или группы признаков.

Искомые взаимосвязи между идентифицируемыми и идентифицирующими признаками, проявляющиеся в изменении статистических характеристик распределения одного признака с изменением другого, называются статистическими (стохастическими, случайными). Противоположность стохастической связи представлена функциональной зависимостью, когда факториальный признак полностью определяет прогнозируемый параметр (такие связи называются также детерминированными или жесткими).

В отличие от функциональной зависимости, дисперсия значений идентифицируемого параметра в статистических взаимосвязях не полностью определяется влиянием идентифицирующего признака (признаков). Более того, статистические взаимосвязи, как правило, не отражают истинный характер причинно-следственных связей.

Математической моделью, лежащей в основе создаваемых способов судебно-медицинского прогнозирования идентифицируемых

параметров, является корреляционно-регрессионный анализ, который включает в себя измерение тесноты, определение направления и установление аналитического выражения связи [79]. Корреляция и регрессия тесно связаны между собой: первая оценивает выраженность статистической связи, а вторая отражает ее форму.

В настоящее время корреляционный анализ (корреляционная модель) определяется как метод статистического анализа взаимозависимости нескольких признаков, применяемый тогда, когда данные наблюдений или эксперимента можно считать случайными и выбранными из генеральной совокупности, распределенной по многомерному нормальному закону [26].

Основным фактором, определяющим видовую структуру методов корреляционного анализа, следует назвать количество идентифицирующих признаков идентифицируемых объектов. В качестве идентифицирующих обычно выступают разнообразные биометрические показатели человека, характеризующиеся каким-либо непрерывным распределением, чаще всего подчиняющимся нормальному закону.

В наиболее простом варианте прогнозирование идентифицируемого параметра может производиться всего лишь по одному биометрическому признаку. Одним из основных показателей взаимозависимости двух случайных величин является парный коэффициент корреляции, служащий мерой линейной статистической зависимости между этими величинами. Этот показатель соответствует своему прямому назначению, когда статистическая связь между соответствующими признаками генеральной совокупности линейна. Как правило, более точное прогнозирование достигается по нескольким, наиболее информативным признакам, отобранным из множества характеристик, описывающих идентифицируемые объекты. Основным условием адекватности многомерного корреляционного анализа также является требование линейности статистических связей. Указанные требования выполняются, если генеральная совокупность распределена по многомерному нормальному закону.

Таким образом, диагностическая значимость результатов исследований, посвященных разработке способов судебно-медицинского прогнозирования идентифицируемых параметров, зависит от степени соответствия исходных данных тем условиям, наличие которых предполагает использованный метод корреляционного анализа.

2.2. ДВУМЕРНАЯ МОДЕЛЬ КОРРЕЛЯЦИОННОГО АНАЛИЗА

Двумерная модель корреляционного анализа лежит в основе статистической взаимосвязи пары признаков, один из которых является идентифицируемым, а второй - идентифицирующим. Как правило, двумерный корреляционный анализ используется на этапе поиска и отбора наиболее информативных признаков для последующего их включения в состав многофакторной регрессионной модели. Вместе с тем двумерный корреляционный анализ при идентификации личности по-прежнему имеет и самостоятельное значение.

Это нашло подтверждение проведенным нами специальным исследованием[3]. Целью его явилось изучение эпидемиологии и адекватности применения корреляционного анализа в научных работах, посвященных судебно-медицинской антропологической идентификации. Объектами анализа выступили 12 случайно отобранных статей, отображающих результаты оригинальных исследований, посвященных разработке способов прогнозирования идентифицируемых параметров, из числа работ, опубликованных в журналах «Судебно-медицинская экспертиза» и «Проблемы экспертизы в медицине» за период 2001-2006 гг. Протокол исследования включал также выявление всех опубликованных в журнале «Судебно-медицинская экспертиза» в течение 2001-2005 гг. оригинальных исследований, выводы которых базировались на результатах применения каких-либо методов аналитической статистики и последующий отбор исследований, в которых использовался корреляционный анализ.

Проведенное исследование показало, что определение силы связей с помощью парных коэффициентов корреляции, имевшее место в 29,8% (25) оригинальных исследований, явилось вторым по частоте применения методом статистической обработки данных, незначительно уступив лишь сравнительному анализу с использованием критерия Стьюдента (31,0%). Применительно к судебно-медицинским исследованиям, посвященным созданию способов идентификации личности на основе корреляционно-регрессионного анализа, двумерная модель последнего в качестве самостоятельного метода статистического анализа данных, использовалась в 5 (42%) статьях (95% доверительный интервал: 15-72%).

---

[3] На результаты этого исследования опираются все эпидемиологические данные, приводимые в главе 2.

Математической моделью однофакторного корреляционного анализа является генеральная совокупность двух признаков, совместное распределение которых задано плотностью двумерного нормального закона.

С помощью выражения (1) можно доказать, что плотность двумерного нормального распределения задается функцией:

$$P(x) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}Q(x_1, x_2)\right\},$$

где

$$Q(x_1, x_2) = \frac{1}{1-\rho^2}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)}{\sigma_1}\frac{(x_2-\mu_2)}{\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right].$$

Изложенное означает, что плотность двумерного нормального распределения определяется пятью параметрами: математическими ожиданиями μ_1 и μ_2 случайных величин x_1 и x_2, их стандартными отклонениями σ_1 и σ_2 и коэффициентом парной корреляции ρ, который в двумерной модели является единственным параметром тесноты связи.

Коэффициент корреляции ρ представляет собой меру линейной зависимости двух переменных. Для коэффициента корреляции ρ двумерной генеральной совокупности справедливо:

1) $-1 \le \rho \le +1$;

2) при $\rho = \pm 1$ имеется функциональная зависимость, совокупность значений пар признаков представляет собой прямую;

3) из равенства $\rho = 0$ следует стохастическая независимость признаков двумерной нормально распределенной генеральной совокупности;

4) чем ближе значение $|\rho|$ к единице, тем сильнее коррелированы две случайные переменные;

5) при $\rho < 0$ имеет место отрицательная корреляция, когда с увеличением одной переменной вторая в среднем уменьшается;

6) при $\rho > 0$ имеет место положительная корреляция, когда с увеличением одной переменной вторая в среднем увеличивается.

Двумерное нормальное распределение может быть графически представлено колоколообразной пространственной поверхностью. Сечение указанной поверхности горизонтальной плоскостью при $\rho = 0$ и $\sigma_1 = \sigma_2$, образует окружность, а при $\sigma_1 \neq \sigma_2$ - эллипс, сужающийся при $\rho \to 1$ (рис. 2).

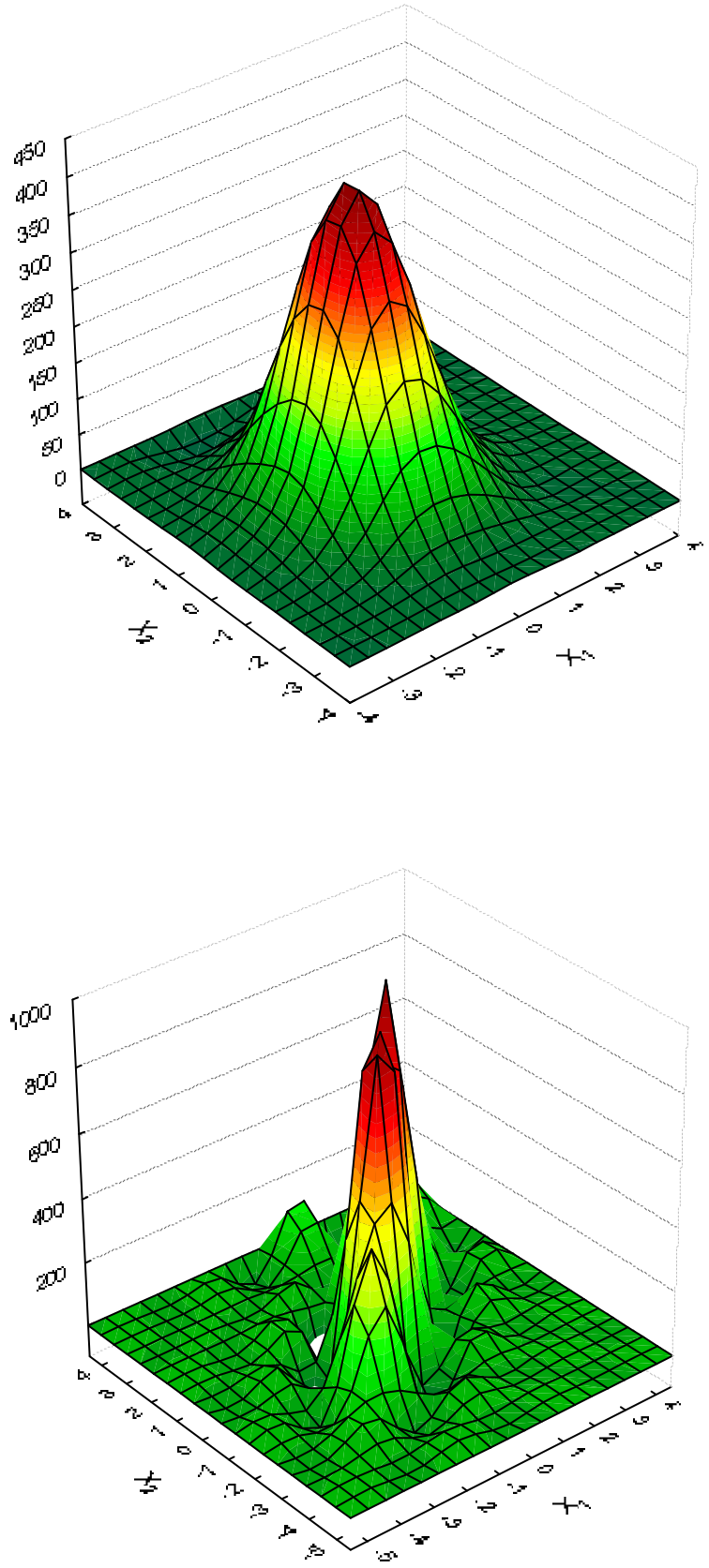

Рис. 2. Графики двумерного стандартного нормального распределения при $\rho = 0$ (верхний рисунок) и $\rho = 1$ (нижний рисунок).

Точечной оценкой параметра ρ тесноты связи между переменными x_1 и x_2 является выборочный парный коэффициент корреляции Пирсона r, который вычисляется как среднее произведение нормированных отклонений:

$$r = \sum \left(\frac{x_1 - \bar{x}_1}{s_1} \right) \left(\frac{x_2 - \bar{x}_2}{s_2} \right) / n = \frac{\sum (x_1 - \bar{x}_1)(x_2 - \bar{x}_2)}{n s_1 s_2}.$$

Статистическая значимость выборочного коэффициента корреляции, т.е. гипотеза о том, может ли выборочный коэффициент корреляции иметь случайные отклонения от нуля при генеральной совокупности с параметром $\rho = 0$, проверяется по Р.А. Фишеру на основании t-распределения с $v = n - 2$ степенями свободы

$$t_{\alpha;n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}. \qquad (5)$$

Существенность выборочного коэффициента корреляции также может быть проверена на основании F-распределения с помощью статистики

$$F = \frac{r^2 (n-2)}{1-r^2} \text{ при } v_1 = 1 \text{ и } v_2 = n - 2 \text{ [130].}$$

Для значимых параметров связи имеет смысл найти интервальные оценки. Для этого служит статистика, введенная Р.А. Фишером:

$$z_r = \frac{1}{2} \ln \frac{r+1}{1-r}, \qquad (6)$$

которая при $n > 10$ распределена приблизительно нормально со средним $\bar{z}_r \approx \frac{1}{2} \ln \frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}$ и стандартным отклонением $s_z = \frac{1}{\sqrt{n-3}}$.

Тогда доверительный интервал для ρ с надежностью 1-α задается выражением

$$z_\rho \in z_r \pm t_{\alpha;\infty} s_z = z_r \pm z_\alpha \sqrt{\frac{1}{n-3}}.$$

Для перехода от z_r к ρ используется специальная таблица [30]. Указанная таблица позволяет получить интервальные оценки ρ вида $r_{\min} \le \rho \le r_{\max}$, пренебрегая поправочным членом $\frac{\rho}{2(n-1)}$ у ρ.

Использование двумерной и других моделей корреляционного анализа можно показать на примере исследования гестационной динамики различных гистоструктур печени и селезенки человека на протяжении 21-40 недель антенатального развития, проведенного нами с целью создания способов судебно-медицинской идентификации гестационного возраста плодов и новорожденных [11,53].

Объектами данного исследования явились трупы 140 плодов и новорожденных. Морфометрическому исследованию были подвергнуты следующие показатели печени: кроветворная активность паренхимы, толщина соединительнотканной капсулы, относительные объемы стромы и долек. В селезенке оценивались толщина капсулы на диафрагмальной поверхности, диаметр и плотность расположения лимфоидных узелков, толщина стенок центральных артерий, относительные объемы белой пульпы, трабекулярного компонента и красной пульпы.

В ходе количественного исследования использовалась стандартная методика изготовления гистологических срезов с поддержанием величин коэффициентов усадки на неизменном уровне. Толщина капсулы печени и все относительные объемные показатели оценивались в 31 наблюдении, гистометрические показатели селезенки исследовались в 99, а степень кроветворной активности печени - в 140 наблюдениях. На всех этапах морфолого-статистического анализа использовались истинные количественные характеристики абсолютных морфометрических показателей.

В целях отбора морфометрических показателей, пригодных для использования в качестве факторных признаков для идентификации гестационного возраста, на первоначальном этапе исследования проводился корреляционный анализ для определения характера и степени тесноты связей между гестационным возрастом и исследовавшимися морфометрическими параметрами.

Одним из результатов исследования явилось основанное на изучении 99 наблюдений обнаружение умеренной положительной зависимости ($r = 0{,}633$; $r^2 = 0{,}400$) диаметра лимфоидных узелков от гестационного возраста (рис. 3).

Проверим значимость оцененного коэффициента корреляции с помощью t-распределения с $v = n - 2$ степенями свободы:

$$t_{97} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{,}633\sqrt{99-2}}{\sqrt{1-0{,}633^2}} = 8{,}050,\ p = 2{,}106 \cdot 10^{-12}.$$

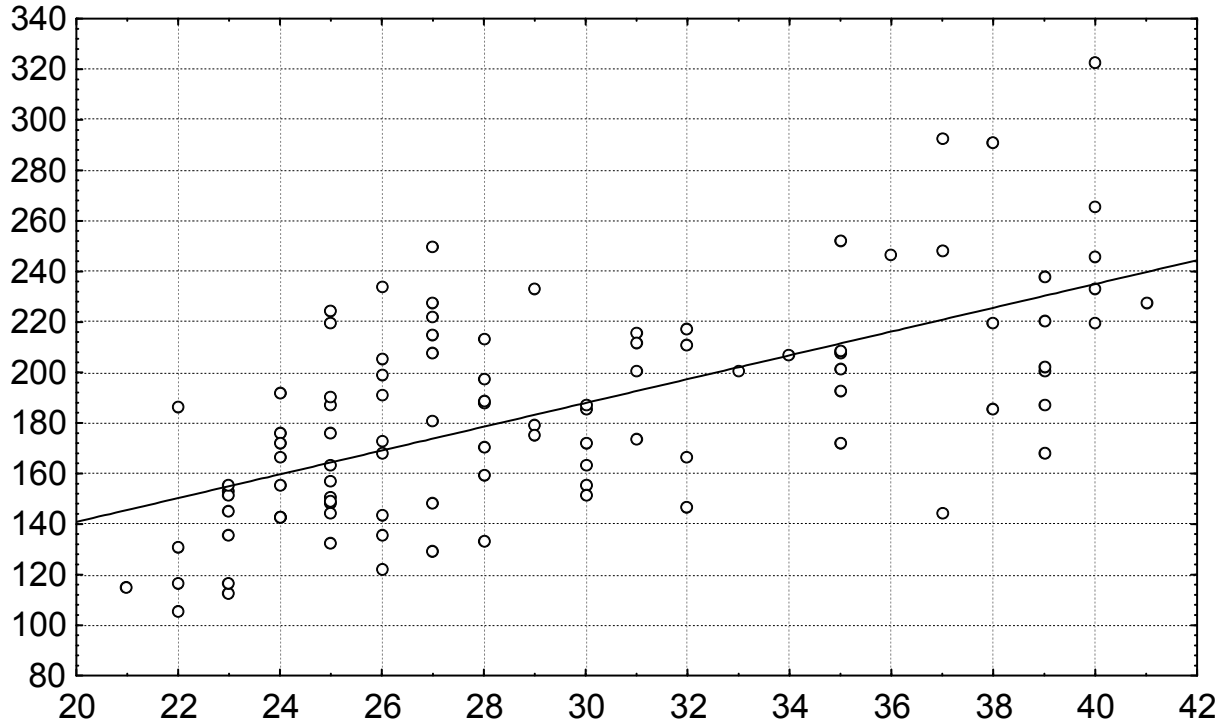

Рис. 3. Зависимость диаметра лимфоидных узелков фетальной селезенки от гестационного возраста. По оси абсцисс – гестационный возраст, недель; по оси ординат – диаметр лимфоидных узелков, мкм.

Аналогичный результат получаем с использованием *F*-критерия:

$$F_{1;97} = \frac{r^2(n-2)}{1-r^2} = \frac{0{,}633^2(99-2)}{1-0{,}633^2} = 64{,}795,\ p = 2{,}106\cdot 10^{-12}.$$

Для определения интервальных оценок рассчитаем статистику Р.А. Фишера

$$z_r = \frac{1}{2}\ln\frac{r+1}{1-r} = 0{,}5\ln\frac{0{,}633+1}{1-0{,}633} = 0{,}746$$

и ее стандартное отклонение

$$s_z = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{99-3}} = 0{,}102.$$

Тогда 95% доверительный интервал для ρ задается пределами

$$0{,}546 < z_r < 0{,}946.$$

Отсюда получаем интервальные оценки ρ и ρ^2:

$$0{,}496 < \rho < 0{,}736;$$

$$0{,}246 < \rho^2 < 0{,}542,$$

указывающие, что доля общей дисперсии показателя диаметра лимфоидных узелков селезенки, объясняемая гестационной динамикой, с 95% вероятностью находится в пределах 24,6—52,2%.

2.3. МНОГОМЕРНЫЙ КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

В современной судебно-медицинской антропологии чаще всего приходится оперировать с многофакторными статистическими связями, выражающими статистическую зависимость идентифицируемого параметра от группы идентифицирующих признаков. Среди судебно-медицинских исследований, основанных на корреляционно-регрессионном анализе, многомерная модель последнего использовалась в 7 (58%) статьях (95% доверительный интервал: 28-85%).

Математической моделью многомерного корреляционного анализа является генеральная совокупность k признаков, совместное распределение которых задано плотностью k-мерного нормального закона (1). С помощью преобразования $\rho_{ij} = \dfrac{\sigma_{ij}}{\sigma_i \sigma_j}$ из ковариационной матрицы Σ легко получить симметрическую и неотрицательно определенную корреляционную матрицу

$$R = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \cdots & \rho_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{k1} & \rho_{k2} & \cdots & 1 \end{pmatrix}.$$

Показателем меры линейной зависимости x_1 от $(x_2, \ldots, x_k)$ служит множественный коэффициент корреляции или его квадрат – множественный коэффициент детерминации, определяемый по формуле

$$\rho^2_{1/2,\ldots k} = 1 - \frac{|R|}{R_{11}},$$

где $|R|$ - определитель этой матрицы; R_{11} - алгебраическое дополнение элемента r_{11} корреляционной матрицы R[4].

---

[4] В матричном исчислении алгебраическим дополнением A_{ij} элемента a_{ij} определителя матрицы называется число, рассчитываемое по формуле $A_{ij} = (-1)^{i+j} M_{ij}$, где M_{ij} - минор данной матрицы. В свою очередь, минором M_{ij} элемента a_{ij} определителя матрицы n-го порядка называется определитель матрицы $(n - 1)$-го порядка, полученной из данной матрицы вычеркиванием ее i-й строки и j-го столбца.

Частным коэффициентом корреляции компонент x_1 и x_2, измеряющим тесноту связи между x_1 и x_2 после устранения влияния компонент $(x_3, x_4 \ldots, x_k)$, является величина

$$\rho_{12|34\ldots k} = -\frac{R_{12}}{\sqrt{R_{11}R_{22}}},$$

где R_{11}, R_{12} и R_{22} - алгебраические дополнения соответствующих элементов матрицы R.

Изложенное можно пояснить на трехмерной модели корреляционного анализа.

Для нормальной распределенной трехмерной генеральной совокупности с признаками (x_1, x_2, x_3) корреляционная матрица представлена квадратной матрицей третьего порядка

$$R_3 = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & 1 \end{pmatrix},$$

где ρ_{ij} - коэффициенты парной корреляции между соответствующими переменными.

Определитель матрицы равен

$$|R_3| = \begin{vmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & 1 \end{vmatrix} = 1 + \rho_{12}\rho_{23}\rho_{31} + \rho_{13}\rho_{21}\rho_{32} - 1\rho_{13}\rho_{31} -$$

$$-1\rho_{12}\rho_{21} - 1\rho_{23}\rho_{32} = 1 + 2\rho_{12}\rho_{23}\rho_{13} - \rho_{13}^2 - \rho_{12}^2 - \rho_{23}^2.$$

Вычислим алгебраические дополнения определителя матрицы $|R_3|$:

$$R_{11} = (-1)^{1+1} M_{11} = (-1)^2 \cdot \begin{vmatrix} 1 & \rho_{23} \\ \rho_{23} & 1 \end{vmatrix} = 1 - \rho_{23}^2;$$

$$R_{12} = (-1)^{1+2} M_{12} = (-1)^3 \cdot \begin{vmatrix} \rho_{21} & \rho_{23} \\ \rho_{31} & 1 \end{vmatrix} = \rho_{13}\rho_{23} - \rho_{12};$$

$$R_{13} = (-1)^{1+3} M_{13} = (-1)^4 \cdot \begin{vmatrix} \rho_{21} & 1 \\ \rho_{31} & \rho_{23} \end{vmatrix} = \rho_{12}\rho_{23} - \rho_{13};$$

$$R_{22} = (-1)^{2+2} M_{22} = (-1)^4 \cdot \begin{vmatrix} 1 & \rho_{13} \\ \rho_{31} & 1 \end{vmatrix} = 1 - \rho_{13}^2;$$

$$R_{23} = (-1)^{2+3} M_{23} = (-1)^5 \cdot \begin{vmatrix} 1 & \rho_{12} \\ \rho_{31} & \rho_{23} \end{vmatrix} = \rho_{12}\rho_{13} - \rho_{23};$$

$$R_{33} = (-1)^{3+3} M_{33} = (-1)^6 \cdot \begin{vmatrix} 1 & \rho_{12} \\ \rho_{21} & 1 \end{vmatrix} = 1 - \rho_{12}^2.$$

Отсюда находим, в частности, коэффициенты $\rho_{1/23}$ множественной и $\rho_{12/3}$ частной корреляции

$$\rho_{1/23} = \sqrt{1 - \frac{|R_3|}{R_{11}}} = \sqrt{\frac{\rho_{12}^2 + \rho_{13}^2 - 2\rho_{12}\rho_{13}\rho_{23}}{1 - \rho_{23}^2}} \text{ и}$$

$$\rho_{12/3} = -\frac{R_{12}}{(R_{11} \cdot R_{22})^{1/2}} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{(1 - \rho_{13}^2) \cdot (1 - \rho_{23}^2)}}.$$

Остальные коэффициенты $\rho_{2/13}$, $\rho_{3/12}$ множественной и $\rho_{13/2}$, $\rho_{23/1}$ частной корреляции вычисляются аналогично.

Выборку объема *n* из *k*-мерной генеральной совокупности *X* можно представить в виде матрицы *X* (2). В этом случае точечные оценки вектора математических ожиданий матрицы *X* представлены *k*-мерным вектором выборочных средних $\bar{x}_l = \frac{1}{n}\sum_{i=1}^{n} x_{il}$, $l = 1,2,\ldots,k$. Несмещенной оценкой ковариационной матрицы Σ является матрица *S* (3).

Оценкой корреляционной матрицы (ρ_{lj}) тогда служит матрица

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1k} \\ r & r_{22} & \cdots & r_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ r_{k1} & r_{k2} & \cdots & r_{kk} \end{pmatrix},$$

где $r_{lj} = \frac{s_{lj}}{s_l s_j}$ - точечная оценка парного коэффициента корреляции между *l*-й и *j*-й компонентами *x*, $l, j = 1,2,\ldots,k$.

Учитывая, что показателем тесноты связи между переменными является среднее произведение нормированных отклонений, матрицу *R* можно выразить также как $R = \frac{1}{n} Z^T Z$,

где $Z = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1k} \\ z_{21} & z_{22} & \cdots & z_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nk} \end{pmatrix}$, $z_{ij} = \dfrac{x_{ij} - \bar{x}_j}{s_j}$.

Используя матрицу выборочных коэффициентов корреляции, получаем формулы для расчета точечных оценок коэффициентов множественной и частной корреляции

$$r_{1/23} = \sqrt{1 - \frac{|R_3|}{R_{11}}} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \text{ и}$$

$$r_{12/3} = -\frac{R_{12}}{(R_{11} \cdot R_{22})^{1/2}} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2) \cdot (1 - r_{23}^2)}}.$$

Остальные выборочные коэффициенты $r_{2/13}$, $r_{3/12}$ множественной и $r_{13/2}$, $r_{23/1}$ частной корреляции вычисляются аналогично.

Проверка значимости множественного коэффициента детерминации ρ^2 осуществляется помощью F-распределения:

$$F_{v_1=k-1;v_2=n-k} = \frac{r^2/(k-1)}{(1-r^2)/(n-k)}.$$

При превышении F-статистики ее критического значения нулевая гипотеза $\rho^2 = 0$ отвергается с вероятностью ошибки, равной α.

Статистическая значимость частных коэффициентов корреляции проверяется на основании t-распределения с $v = n - k$ степенями свободы

$$t_{\alpha;n-k} = \frac{r\sqrt{n-k}}{\sqrt{1-r^2}}. \qquad (7)$$

Значимость парных, множественных и частных коэффициентов корреляции зависит от объема выборок [13,114]. Приведенные критерии значимости основываются на допущениях линейного регрессионного анализа относительно требований нормальности распределений и постоянства дисперсий остатков для множества значений каждой независимой переменной (подробнее см. раздел 3.2). Исследования методом Монте-Карло показали, что нарушение этих условий не является критичным, если размеры выборки не слишком малы, а отклонения от нормальности не очень большие [13].

Интервальные оценки для значимого множественного коэффициента корреляции находятся также с помощью преобразования Р.А. Фишера (6) с дисперсией, приблизительно равной $s_z^2 = \frac{1}{n}$ для достаточно больших значений n [26].

Для демонстрации трехмерной модели корреляционного анализа продолжим исследование возможности гистометрической идентификации гестационного возраста. Пусть целью исследования будет определение возможностей идентификации гестационного возраста одновременно по двум гистометрическим показателям селезенки: диаметру лимфоидных узелков и толщине стенок центральных артерий. В терминах математической статистики названная цель представляет собой задачу определения коэффициента множественной корреляции гестационного возраста с указанными гистометрическими структурами фетальной селезенки.

Введем условные обозначения: y – гестационный возраст, недель; x – диаметр лимфоидных узелков селезенки, мкм; z - толщина стенок центральных артерий селезенки, мкм.

Выборка в количестве 99 наблюдений из трехмерной генеральной совокупности X была представлена матрицей

$$X = \begin{pmatrix} y_1 & x_1 & z_1 \\ y_2 & x_2 & z_2 \\ \vdots & \vdots & \vdots \\ y_{99} & x_{99} & z_{99} \end{pmatrix} = \begin{pmatrix} 22 & 186 & 7{,}8 \\ 24 & 143 & 8{,}1 \\ \vdots & \vdots & \vdots \\ 40 & 246 & 9{,}7 \end{pmatrix} \text{ (рис. 4)}$$

с вектором точечных оценок математических ожиданий

$$\begin{pmatrix} \bar{y} \\ \bar{x} \\ \bar{z} \end{pmatrix} = \begin{pmatrix} 29{,}5 \\ 185{,}4 \\ 8{,}75 \end{pmatrix}.$$

Перейдем к центрированным величинам

$$U = \begin{pmatrix} y_1 - \bar{y} & x_1 - \bar{x} & z_1 - \bar{z} \\ y_2 - \bar{y} & x_2 - \bar{x} & z_2 - \bar{z} \\ \vdots & \vdots & \vdots \\ y_{99} - \bar{y} & x_{99} - \bar{x} & z_{99} - \bar{z} \end{pmatrix} = \begin{pmatrix} -7{,}5 & 0{,}6 & -0{,}95 \\ -5{,}5 & -42{,}4 & -0{,}65 \\ \vdots & \vdots & \vdots \\ 10{,}5 & 60{,}6 & 0{,}95 \end{pmatrix}.$$

Откуда

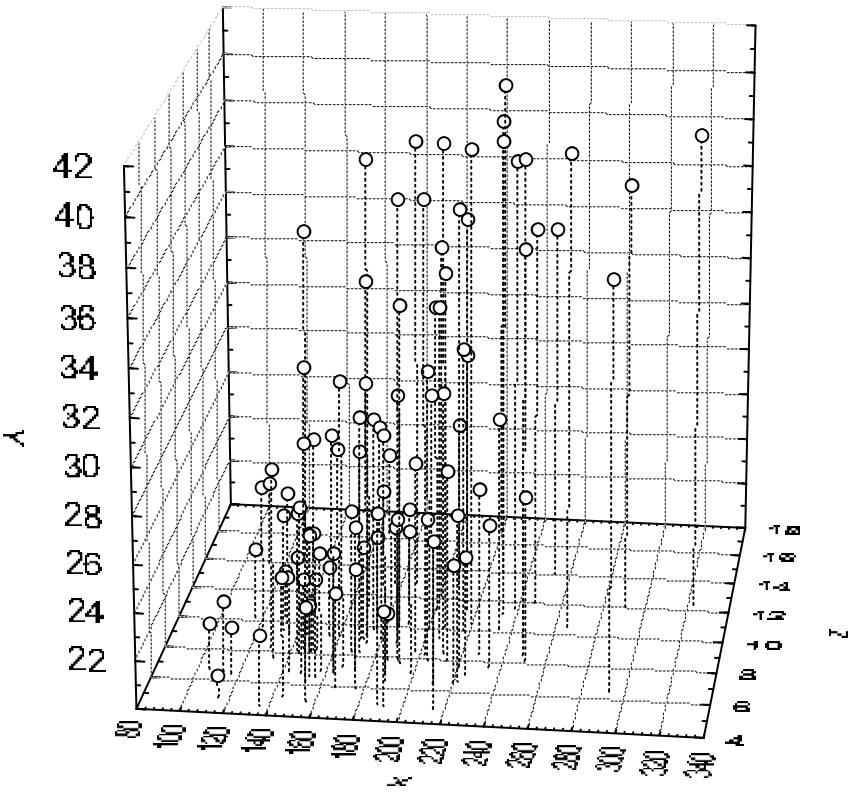

Рис. 4. Взаимозависимость гестационного возраста и гистометрических показателей фетальной селезенки (n = 99). По оси X – диаметр лимфоидных узелков, мкм; по оси Z – толщина стенок центральных артерий, мкм; по оси Y – гестационный возраст, недель.

$$U^T U = \begin{pmatrix} -7{,}5 & -5{,}5 & \cdots & 10{,}5 \\ 0{,}6 & -42{,}4 & \cdots & 60{,}6 \\ -0{,}95 & -0{,}65 & \cdots & 0{,}95 \end{pmatrix} \begin{pmatrix} -7{,}5 & 0{,}6 & -0{,}95 \\ -5{,}5 & -42{,}4 & -0{,}65 \\ \vdots & \vdots & \vdots \\ 10{,}5 & 60{,}6 & 0{,}95 \end{pmatrix} =$$

$$= \begin{pmatrix} 3126{,}545 & 14716{,}709 & 765{,}825 \\ 14716{,}709 & 172973{,}157 & 3861{,}502 \\ 765{,}825 & 3861{,}502 & 544{,}182 \end{pmatrix},$$

$$S = \frac{1}{n-1} U^T U = \frac{1}{98} \begin{pmatrix} 3126{,}545 & 14716{,}709 & 765{,}825 \\ 14716{,}709 & 172973{,}157 & 3861{,}502 \\ 765{,}825 & 3861{,}502 & 544{,}182 \end{pmatrix} =$$

$$= \begin{pmatrix} 31{,}904 & 150{,}171 & 7{,}815 \\ 150{,}171 & 1765{,}032 & 39{,}403 \\ 7{,}815 & 39{,}403 & 5{,}553 \end{pmatrix}.$$

Итак, точечные оценки дисперсий и стандартных отклонений следующие: $s_y^2 = 31{,}904$; $s_y = 5{,}648$; $s_x^2 = 1765{,}032$; $s_x = 42{,}012$; $s_z^2 = 5{,}553$; $s_z = 2{,}356$.

Получаем оценку корреляционной матрицы (ρ_{lj})

$$R = \begin{pmatrix} 1 & 0{,}633 & 0{,}587 \\ 0{,}633 & 1 & 0{,}398 \\ 0{,}587 & 0{,}398 & 1 \end{pmatrix}.$$

Используя матрицу выборочных коэффициентов корреляции, определим точечные оценки коэффициентов множественной и частной корреляции

$$r_{y/xz} = \sqrt{\frac{0{,}633^2 + 0{,}587^2 - 2 \cdot 0{,}633 \cdot 0{,}587 \cdot 0{,}398}{1 - 0{,}398^2}} = 0{,}731;$$

$$r_{yx/z} = \frac{0{,}633 - 0{,}587 \cdot 0{,}398}{\sqrt{(1 - 0{,}587^2) \cdot (1 - 0{,}398^2)}} = 0{,}537;$$

$$r_{yz/x} = \frac{0{,}587 - 0{,}633 \cdot 0{,}398}{\sqrt{(1 - 0{,}633^2) \cdot (1 - 0{,}398^2)}} = 0{,}472.$$

Проверим значимость множественного коэффициента детерминации:

$$F_{v_1=2;v_2=96} = \frac{0{,}731^2 /(3-1)}{(1 - 0{,}731^2)/(n-3)} = 55{,}009,\ p = 1{,}207 \cdot 10^{-16}.$$

Для определения интервальных оценок рассчитаем статистику Р.А. Фишера (6)

$$z_r = 0{,}5 \ln \frac{0{,}731 + 1}{1 - 0{,}731} = 0{,}930$$

и ее стандартное отклонение

$$s_{z_r} = \frac{1}{\sqrt{n}} = \frac{1}{\sqrt{98}} = 0{,}101.$$

Тогда 95% доверительный интервал для z_r задается выражением

$$0{,}729 < z_r < 1{,}131,$$

откуда получаем интервальные оценки ρ и ρ^2:

$$0{,}619 < \rho < 0{,}807;$$

$$0{,}384 < \rho^2 < 0{,}651,$$

которые позволяют утверждать, что доля общей дисперсии показателя гестационной динамики, которую можно прогнозировать по

показателям диаметра лимфоидных узелков и толщины стенок центральных артерий фетальной селезенки, с 95% вероятностью находится в пределах 38,6—65,5%.

Проверим значимость частных коэффициентов корреляции с уровнем значимости $\alpha = 0{,}05$.

$$t_{yx/z} = \frac{0{,}537\sqrt{99-3}}{\sqrt{1-0{,}537^2}} = 6{,}245;$$

$$t_{yz/x} = \frac{0{,}472\sqrt{99-3}}{\sqrt{1-0{,}472^2}} = 5{,}245.$$

При $v = n - k$ степенях свободы соответствующие вероятности ошибочного принятия альтернативной гипотезы $\rho \neq 0$ составляют $p = 1{,}153 \cdot 10^{-8}$ для $r_{yx/z}$ и $p = 9{,}298 \cdot 10^{-7}$ для $r_{yz/x}$.

Определим интервальные оценки частных коэффициентов корреляции

$$0{,}379 < \rho_{yx/z} < 0{,}662;$$

$$0{,}301 < \rho_{yz/x} < 0{,}610;$$

и детерминации

$$0{,}144 < \rho^2_{yx/z} < 0{,}439;$$

$$0{,}091 < \rho^2_{yz/x} < 0{,}372.$$

Частный коэффициент корреляции обладает всеми свойствами парного коэффициента корреляции и служит показателем линейной связи между двумя случайными переменными независимо от влияния остальных случайных переменных. Например, сравним модули коэффициентов парной корреляции диаметра лимфоидных узелков с гестационным возрастом ($r_{yx} = 0{,}633$) и аналогичной частной корреляции при фиксированном показателе толщины стенок центральных артерий ($r_{yx/z} = 0{,}537$). Неравенство $r_{yx} < r_{yx/z}$ означает, что взаимозависимость между диаметром лимфоидных узелков и гестационным возрастом частично обусловлена воздействием на эту пару показателя толщины стенок центральных артерий.

Кроме частного коэффициента корреляции существует также еще один показатель уникальности взаимозависимости между данной независимой переменной и результативным признаком, называемый получастным коэффициентом корреляции (semi-partial correlation) [13].

2.4. РАНГОВАЯ КОРРЕЛЯЦИЯ

Корреляционный анализ возможен при выполнении трех условий: линейность взаимосвязи, нормальность распределения и количественный характер признаков. В судебно-медицинской антропологии эти условия выполняются далеко не всегда. По выборочным данным доля судебно-медицинских антропологических исследований с использованием корреляционного анализа, в которых изучались взаимосвязи с ранговыми показателями, составила 8% (95% доверительный интервал: 0-39%). В этих случаях использование методов классического корреляционного анализ либо неэффективно, либо не применимо.

Поэтому для изучения тесноты связей между неколичественными признаками, а также количественными признаками с непрерывными неизвестными или не подчиняющимися нормальному закону распределениями применяются методы ранговой корреляции.

Под рангом наблюдаемого значения x_i признака x называется номер этого наблюдения в ранжированном ряду $x_1 \le x_2 \le \ldots \le x_n$ при условии, что неравенства - строгие. Если в ранжированном ряду встречаются одинаковые члены, то в качестве одинаковых рангов берется средняя арифметическая соответствующих номеров.

Пусть имеется выборка объема n из непрерывно распределенной двумерной генеральной совокупности (x,y): $(x_1, y_1)\ldots(x_n, y_n)$. Если выборка упорядочена по x, то ей соответствует следующая матрица

$$\begin{pmatrix} 1 & 2 & \cdots & n \\ R_1 & R_2 & \cdots & R_n \end{pmatrix},$$

в которой первая строка состоит из рангов наблюдений x, а вторая – из рангов y.

Очевидно, что детерминированной (функциональной) положительной связи между x и y соответствует подстановка

$$\begin{pmatrix} 1 & 2 & \cdots & n \\ 1 & 2 & \cdots & n \end{pmatrix};$$

детерминированной отрицательной связи - подстановка

$$\begin{pmatrix} 1 & 2 & \cdots & n \\ n & n-1 & \cdots & 1 \end{pmatrix}.$$

Остальные возможные подстановки получаются при той или иной степени связи, являющейся собственно стохастической. Количество таких подстановок равно $n-2$.

Для измерения степени связи между x и y используются понятия инверсии (беспорядка) и порядка. Если $R_i > R_j$, но в ранжированном ряду R_i стоит слева от R_j, то такая локализация называется инверсией между элементами перестановки R_i и R_j второй строки подстановки. Если при том же расположении $R_i < R_j$, то считают, что элементы R_i и R_j инверсии не образуют или образуют порядок [26].

В качестве меры связи принимают разность между суммами чисел порядков N и Q, образованных элементами второй строки подстановки. Учитывая, что общее количество различных подстановок составляет $n!$, можно определить вероятности получения перестановок с заданной мерой связи. Например, в таблице 3 приведены расчеты для подстановок из трех элементов.

Таблица 3

Расчетная таблица для подстановок из трех элементов

Число порядков N	Число инверсий Q	Мера сходства S_K	Подстановки	Вероятность
0	3	-3	321	1/6
1	2	-1	312, 231	1/3
2	1	1	213,132	1/3
3	0	3	123	1/6

В теории доказывается, что сумма числа порядков N и инверсий Q равна сумме номеров перестановки, т.е. $1+2+\ldots+n=n(n+1)/2$, а распределение вероятностей симметрично относительно центра S_K, равного нулю.

Коэффициент ранговой корреляции Кендалла определяется нормированием случайной величины S_K путем ее деления на $n(n-1)/2$:

$$r_K = \frac{2S_K}{n(n-1)} = 1 - \frac{4Q}{n(n-1)} = \frac{4N}{n(n-1)} - 1.$$

При больших объемах n можно использовать нормальный закон распределения r_K с математическим ожиданием, равным нулю и дисперсией $s^2_{r_K} = \frac{2(2n+5)}{9n(n-1)}$ [26].

Расчет другого рангового коэффициента корреляции Спирмена (r_S) аналогичен коэффициенту Пирсона, принимая вместо самих значений пар признаков их ранги:

$$r_S = 1 - \frac{6S_S}{n^3 - n},\ S_S = \sum_{i=1}^{n} (R_i - i)^2 .$$

При больших объемах n коэффициент r_S подчиняется нормальному распределению с математическим ожиданием, равным нулю и дисперсией $s_{r_S}^2 = \frac{1}{n-1}$ [26].

Коэффициенты корреляции Кендалла и Спирмена применяются только лишь как показатели тесноты связи между двумя признаками. Как и для обычного парного коэффициента корреляции Пирсона значения r_K и r_S изменяются в пределах от -1 до +1.

При изучении связей между числом порядковых признаков, большим двух, используют меру сходства соответствующего числа перестановок. В качестве показателя согласованности определяется коэффициент конкордации Кендалла, который, как и обычный множественный коэффициент корреляции, может изменяться в пределах от нуля (абсолютная несогласованность) до единицы (полное совпадение всех ранжировок). Непараметрическими выражениями зависимостей между категориальными переменными являются χ^2-критерий и точный критерий Фишера [13].

Поскольку мощность ранговой корреляции, как и любого другого непараметрического метода, когда его применяют на нормальном распределении, всегда слабее, чем соответствующий параметрический аналог, в статистике используется показатель E_n:

$$E_n = \frac{n \text{ для параметрического критерия}}{n \text{ для непараметрического критерия}},$$

который называют «эффективностью» непараметрического критерия [30]. При этом под n подразумевают объем выборки, необходимый для получения заданной мощности критерия. Понятие асимптотической эффективности применяется для случая бесконечно большой выборки нормально распределенной случайной величины.

Эффективность коэффициента ранговой корреляции Спирмена составляет 91% [30]. Ввиду небольших потерь в мощности использование коэффициента r_S рекомендуется при неуверенности соответствия исходных данных предпосылкам классического корреляционного анализа [16].

2.5. НЕЛИНЕЙНАЯ КОРРЕЛЯЦИЯ

Одним из возможных источников трудностей, возникающих при проведении корреляционного анализа данных в любых научно-практических приложениях, является форма исследуемой зависимости. Не являются исключением из этого правила и судебно-медицинские антропологические исследования.

Как уже упоминалось, классический корреляционный анализ хорошо подходит лишь для описания линейных зависимостей. Поэтому на первоначальном этапе корреляционного анализа биометрических данных необходимо определить форму (линейная или нелинейная) связи между исследуемыми параметрами. Существующие для проверки линейности регрессии статистические критерии являются очень трудоемкими и, самое главное, их применение возможно только тогда, когда каждому значению независимой переменной соответствует группа значений зависимой переменной. Визуальный же анализ диаграммы рассеяния (корреляционного поля) пригоден лишь при явно заметной форме связи между исследуемыми параметрами. Поэтому в судебно-медицинской антропологии и в других приложениях актуальной является проблема быстрого предварительного оценивания формы искомой зависимости.

Для решения указанной задачи нами был разработан метод, основанный на свойствах парных коэффициентов линейной и ранговой корреляции [8]. Суть его состоит в том, что корреляция Пирсона хорошо подходит лишь для описания линейной зависимости, поскольку любые отклонения от линейности, увеличивая общую сумму квадратов расстояний от регрессионной прямой, резко снижают величину коэффициента корреляции. Коэффициент корреляции Спирмена, использующий ранжированные данные, напротив, ослабляет влияние выбросов и эффективен при монотонной нелинейной зависимости, когда при увеличении одной переменной другая в среднем либо непрерывно возрастает, либо непрерывно уменьшается (отсутствуют точки минимума и максимума, называющиеся также точками экстремума функции).

Учитывая, что асимптотическая эффективность коэффициента ранговой корреляции составляет примерно 91% [30], то в случае нормальности распределения выборочной совокупности данных, наличия линейной связи между исследуемыми параметрами и при отсутствии выбросов, абсолютное значение коэффициента линей-

ной корреляции должно превышать абсолютное значение коэффициента ранговой корреляции. Исключением из указанного правила является лишь функциональная зависимость с $|r| = |r_S| = 1$.

Поэтому при $|r| < |r_S|$ и отсутствии выбросов форма исследуемой зависимости является нелинейной, и наилучшее сглаживание выборочной совокупности данных будет достигнуто только с помощью нелинейной регрессионной модели. При $|r| = |r_S| \neq 1$ форма исследуемой зависимости также не является линейной. При $|r| > |r_S|$ форма исследуемой зависимости является линейной или имеются небольшие отклонения от линейности. Если заранее известно, что искомая зависимость по форме является линейной, то превышение модуля коэффициента ранговой корреляции над абсолютным значением коэффициента линейной корреляции свидетельствует о наличии выбросов в исследуемой совокупности данных.

Указанный метод был многократно использован нами при разработке способа идентификации гестационного возраста плодов человека по ряду морфометрических показателей фетальных органов [53]. Особенностью выборочных данных в указанном исследовании явилось то, что каждому значению независимой переменной соответствовало лишь одно значение зависимой переменной. Данное обстоятельство исключало возможность проверки линейности регрессии с помощью *F* – статистики, являющейся отношением отклонений средних значений зависимой переменой от прямой регрессии к отклонению значений зависимой переменной от групповых средних, или основанной на расчете корреляционных отношений. Сравнение модулей коэффициентов линейной и ранговой корреляции позволило установить линейность большинства регрессий и избежать трудоемкой процедуры поиска нелинейных аппроксимаций.

Практическое использование вышеописанного метода показало, что при $|r| \leq |r_S|$ наилучшая аппроксимация выборочных данных может быть достигнута только с помощью нелинейных регрессионных моделей. При $|r| > |r_S|$ и небольшой величине разности $|r| - |r_S|$ в отдельных случаях сглаживание данных с наименьшей остаточной дисперсией также достигалось с помощью уравнений нелинейной регрессии. Однако данные регрессионные модели не соответствовали другим критериям качества. Указанное несоответствие проявлялось либо статистической незначимостью отдельных регрессионных коэффициентов, либо отсутствием монотонности регрессии,

что противоречило природе исследуемых данных, либо аппроксимация достигалась моделью, нелинейной по параметрам, то есть уравнением, не представимым в виде простой регрессионной модели с некоторыми преобразованиями независимых переменных. В любом случае при $|r| > |r_S|$ линейные модели, иногда обладая чуть большей остаточной дисперсией, были лишены вышеперечисленных недостатков альтернативных нелинейных регрессионных уравнений, что определило выбор первых в качестве наилучших аппроксимаций исследуемых данных.

Таким образом, метод опознавания нелинейности регрессии на основе сравнения модулей коэффициентов линейной и ранговой корреляции нетрудоемок в практическом применении и эффективен при исследовании выборок, в которых каждому значению независимой переменной соответствует только одно значение зависимой переменной.

Кроме использования непараметрической корреляции эффективный анализ монотонных зависимостей можно провести путем преобразования одной или обеих переменных, чтобы сделать зависимость линейной, а затем уже исследовать зависимость между преобразованными величинами. Для этого часто используется логарифмическое преобразование.

Для демонстрации практического использования изложенных методов продолжим исследование зависимости гистоструктур фетальной селезенки от гестационного возраста. Одним изучавшихся гистометрических показателей была плотность расположения лимфоидных узелков. Визуальный и численные методы анализа значений данного показателя не обнаружили неоднородности выборочных данных и их отклонений от нормальности [53].

При корреляционном анализе 99 наблюдений была выявлена сильная отрицательная зависимость плотности расположения лимфоидных узелков фетальной селезенки от гестационного возраста (рис. 5). При этом абсолютное значение парного коэффициента ранговой корреляции ($r_S = –0{,}769$; t = -11,863; $p = 1{,}372 \cdot 10^{-20}$) превысило соответствующую величину парного коэффициента линейной корреляции (r = -0,716; t = -10,111; $p = 7{,}740 \cdot 10^{-17}$), что при отсутствии выбросов в исследованной совокупности данных указывало на наличие нелинейной монотонной убывающей зависимости

плотности расположения лимфоидных узелков от гестационного возраста.

Поскольку зависимость плотности расположения лимфоидных узелков селезенки от гестационного возраста представляет собой монотонную убывающую кривую (см. рис 5), для расчетов долей дисперсии данного показателя, объясняемых гестационной динамикой и воздействием случайных факторов целесообразно использовать более чувствительный в данной ситуации коэффициент ранговой корреляции Спирмена (r_S = -0,769). Учитывая, что $s^2_{r_S} = \frac{1}{n-1}$, 95% доверительные интервалы для ρ_S и ρ_S^2 определяются как

$$-0{,}838 < \rho_S < -0{,}672;$$
$$0{,}452 < \rho_S^2 < 0{,}702.$$

Это позволяет утверждать, что доля общей дисперсии показателя плотности расположения лимфоидных узелков селезенки, объясняемая гестационной динамикой с 95% вероятностью находится в пределах 45,2-70,2%, а доля вариации, зависящая от воздействия случайных факторов, с 97,5% вероятностью не превысит 54,8%.

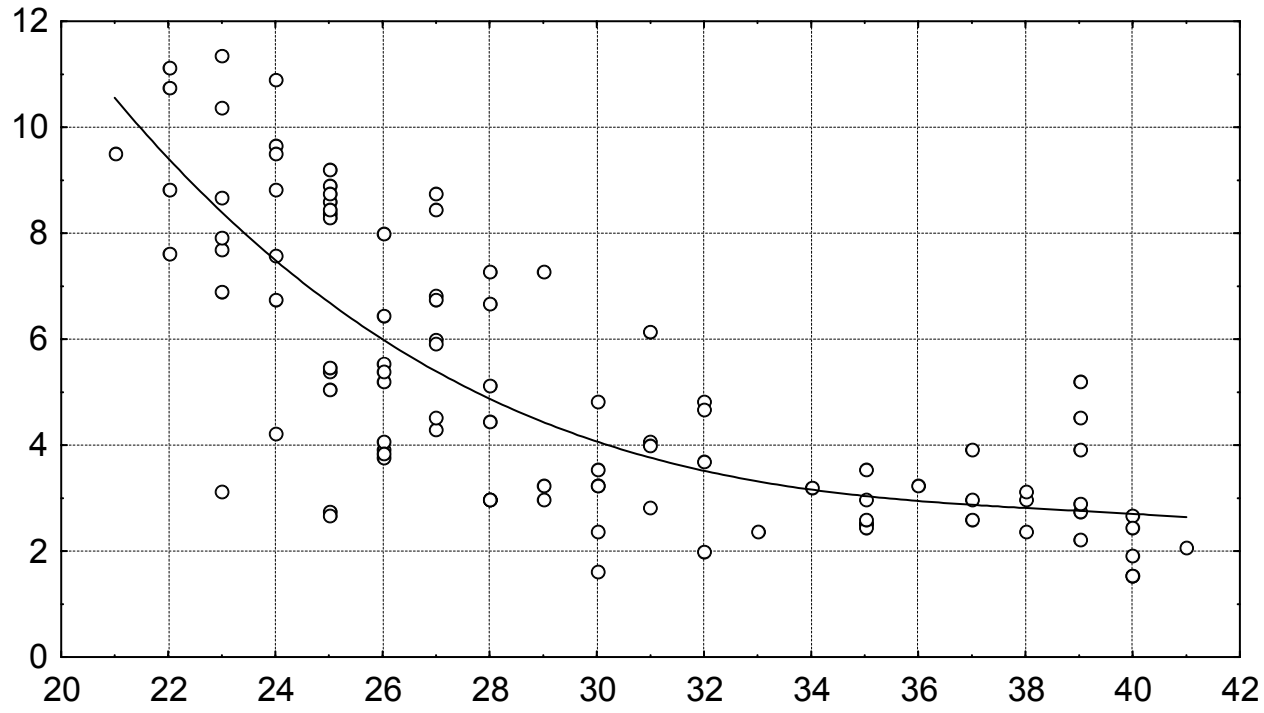

Рис. 5. Зависимость плотности расположения лимфоидных узелков фетальной селезенки от гестационного возраста. По оси абсцисс – гестационный возраст, недель; по оси ординат – плотность расположения лимфоидных узелков, число профилей.

Приведем также результаты описания исследуемой зависимости с помощью линеаризирующего преобразования одной из переменных путем ее логарифмирования. Полученный коэффициент парной корреляции между гестационным возрастом и логарифмом плотности лимфоидных узелков также превысил модуль линейной корреляции Пирсона (r = -0,741; t = -10,871; $p = 1{,}783 \cdot 10^{-18}$), что в данном случае подтверждает наличие монотонной нелинейной убывающей гестационной зависимости указанного гистометрического показателя.

В отличие от применения какого-либо линеаризирующего преобразования, выбранного наугад, самым точным методом исследования нелинейных зависимостей является поиск функции, которая наилучшим образом описывает данные. Указанный метод также является максимально эффективным при описании немонотонных зависимостей, имеющих точки экстремума. Однако данный поиск проводится уже в рамках регрессионного анализа.

Интересно отметить, что по выборочным данным нелинейный характер формы изучаемых зависимостей учитывался лишь в 3 (25%) судебно-медицинских антропологических исследованиях. При этом в двух из них факт нелинейности был известен по литературным данным еще до начала исследования. Проверка же нелинейности изучаемых зависимостей в условиях ее неочевидности была осуществлена лишь только в одном (95% доверительный интервал: 0-39%) судебно-медицинском антропологическом исследовании.

При обнаружении нелинейности зависимости нерешенным аспектом судебно-медицинских антропологических исследований является обоснование выбора метода ее описания. Анализ научной литературы указанной тематики показывает наличие четырех тактических вариантов описания нелинейных зависимостей.

Первый вариант описания заключается в игнорировании нелинейности и использовании в качестве исследовательского инструмента обычного корреляционного анализа с вычислением парных коэффициентов корреляции Пирсона. По выборочным данным указанный вариант изучения зависимостей встретился в 75% (9) судебно-антропологических исследований (95% доверительный интервал: 43-95%).

Второй вариант сводится к разбиению исследуемой зависимости на произвольно выбранные промежутки, которые исследуются как

отдельные зависимости с помощью линейного корреляционного анализа. Этот вариант изучения зависимостей встретился в 17% (2) судебно-антропологических исследований (95% доверительный интервал: 2-48%).

Третий вариант исследования зависимостей представлен ранговым корреляционным анализом. Данный подход был обнаружен нами в 8% (1) судебно-медицинских антропологических исследований (95% доверительный интервал: 0-39%). Но использовался он не как самостоятельный метод, а в качестве дополнения к четвертому варианту.

Четвертый вариант исследования зависимостей характеризовался поиском наилучшей аппроксимирующей функции путем использования различных линеаризирующих преобразований преимущественно независимой переменной (идентифицирующего признака). Сила искомой зависимости в этом случае выражалась коэффициентом парной или множественной корреляции между соответствующими преобразованиями переменных. Названный метод также был применен всего лишь в 8% (1) судебно-медицинских антропологических исследований.

Приведенные данные свидетельствуют об отсутствии обоснованного стандартизованного подхода к выбору метода изучения и описания зависимостей при проведении судебно-медицинских антропологических исследований. Указанное обстоятельство ввиду очевидной своей важности требует подробного обсуждения.

Как было показано выше, использование классического корреляционного анализа допустимо лишь при описании линейных зависимостей или при незначительных их отклонениях от линейности. В последнем случае ценой игнорирования нелинейности будет снижение эффективности корреляционного анализа, которое в итоге приведет к снижению точности разработанных способов судебно-медицинской идентификации.

Например, на рисунке 6 приведен график нелинейной монотонной убывающей детерминированной зависимости $(\rho = 1)$. Поскольку аппроксимирующая функция в данном случае представлена кубическим полиномом, то указанная зависимость характеризуется множественным коэффициентом корреляции, который может принимать значения на закрытом промежутке от 0 до 1, т.е. не может быть отрицательным, несмотря на убывающий характер функции.

При игнорировании нелинейности данной зависимости можно получить следующую точечную оценку ρ: $r = -0{,}932$, которая характеризуется потерей в точности идентификации на 13% ($r^2 = 0{,}869$). Показательно, что применение ранговой корреляции в данном случае позволило получить реальную оценку силы зависимости ($r_S = -1$).

В этой связи мы считаем возможным использование классического корреляционного анализа при изучении нелинейных зависимостей только на первоначальном этапе разведочного исследования данных. Окончательное же описание нелинейных зависимостей линейными функциями, на наш взгляд, следует считать неприемлемым.

Второй подход, связанный с разбиением исследуемой зависимости на произвольно выбранные промежутки, следует признать обоснованным и весьма полезным особенно в исследованиях теоретического плана. В частности, именно такой метод анализа применялся большинством судебных медиков при исследовании процессов старения [56,65,67,81]. К этому следует добавить, что с практической точки зрения (для уменьшения количества диагностических регрессионных моделей) в качестве концов выделяемых промежутков целесообразно выбирать точки экстремумов исследуемой функции [56].

В математическом анализе различают два вида точек экстремумов функции: точки максимума и минимума. При этом точка x_0 называется точкой минимума функции f, если найдется такая окрестность точки x_0, что для всех x из этой окрестности $f(x_0) \le f(x)$. Для точки максимума верно противоположное неравенство: $f(x_0) \ge f(x)$. Несколько более сложное определение точек экстремума связано с вычислением первых и вторых производных функции. Так, для максимума функции должно выполняться неравенство $f''(x_0) < 0$, для минимума - $f''(x_0) > 0$. Например, приведенная на рисунке 7 гипотетическая зависимость имеет одну точку максимума, которую следует выбрать в качестве точки разбиения исследуемой зависимости на два промежутка. На одном промежутке функция имеет строго монотонный возрастающий (т.е. из неравенства $x_1 < x_2$ следует неравенство $f(x_1) < f(x_2)$), а на другом – аналогичный убывающий характер (т.е. из неравенства $x_1 < x_2$ следует, что $f(x_1) > f(x_2)$).

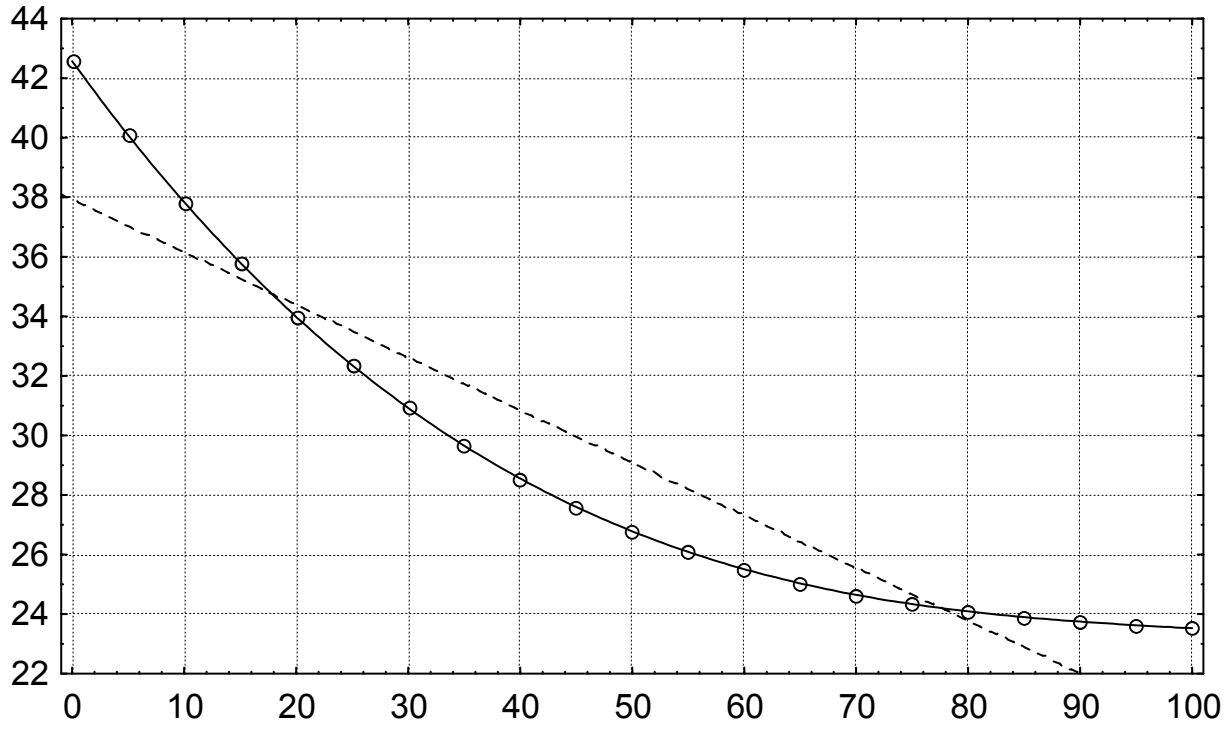

Рис. 6. График строго монотонной убывающей зависимости, $\rho = 1$; $r = -0{,}932$; $r_S = -1$. Здесь и на рис. 7: по оси абсцисс – гипотетический идентифицирующий признак, по оси ординат – идентифицируемый параметр; пунктирной линией показана аппроксимирующая линейная зависимость.

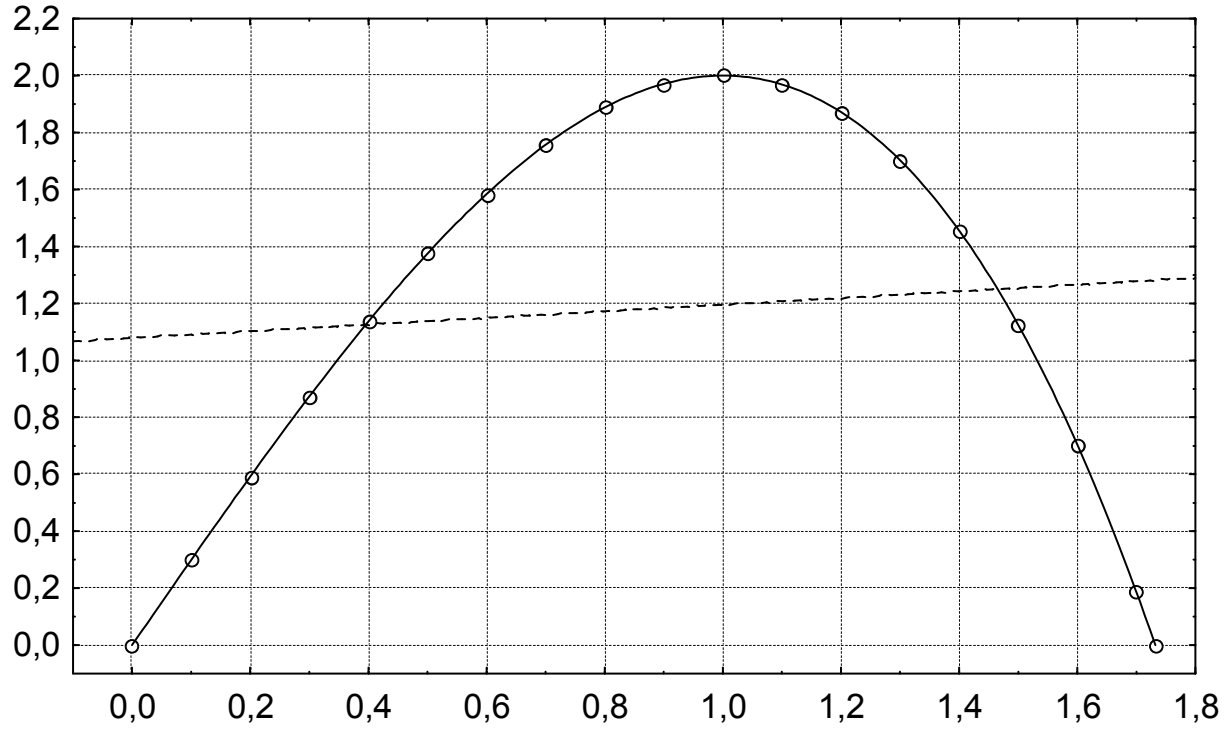

Рис. 7. График нелинейной немонотонной зависимости, имеющей экстремум в точке с абсциссой +1, $\rho = 1$; $r = 0{,}092$; $r_S = 0{,}079$.

Каждый из полученных промежутков уже можно приближенно охарактеризовать с помощью коэффициентов ранговой и даже линейной корреляции.

Необходимо заметить, что если для точек окрестности точки x_0 выполняется равенство $f(x_0) = f(x)$, то исследуемую зависимость целесообразно разделить на следующие промежутки: $f(x) < f(x_0)$, $f(x) = f(x_0)$ и $f(x) > f(x_0)$. Именно такое разделение характерно для изучения возрастных изменений, последовательно проходящих стадии развития, стабилизации и инволюции [21,84].

Ранговый корреляционный анализ эффективен лишь при исследовании монотонных нелинейных зависимостей. Например, ранговый корреляционный анализ изображенных на рисунке 7 данных показал отсутствие зависимости $r_S = 0{,}079$, так же как и свой линейный аналог $r = 0{,}092$. Кроме того, ранговая корреляция (кроме коэффициента конкордации) не применима при исследовании многофакторных зависимостей и неэффективна для последующего аналитического описания. Поэтому ранговая корреляция также может использоваться лишь на этапе разведочного анализа данных.

Последний вариант исследования зависимостей, связанный с поиском наилучшей аппроксимирующей функции, на наш взгляд, является «золотым стандартом» корреляционно-регрессионного анализа в судебно-медицинских антропологических исследованиях. Данный метод эффективен при исследовании любых зависимостей (монотонных и немонотонных) на всей области их определения. Это позволяет использовать метод поиска наилучшей аппроксимирующей функции при исследовании любых биометрических данных и сокращать количество итоговых диагностических регрессионных моделей до одной.

Еще одним ценным преимуществом излагаемого подхода является то, что поиск наилучшей аппроксимации одновременно представляет собой средство решения задачи аналитического описания изучаемой зависимости, т.е. построения регрессионной модели идентификации личности. Например, регрессионные модели нелинейных зависимостей, изображенных на рисунках 6 и 7, имеют соответственно вид

$$y = 42{,}574 + 0{,}521x + 0{,}0049x^2 - 1{,}597 \cdot 10^{-5} x^3;$$

$$y = 3x - x^3$$

и характеризуются отсутствием потерь в точности идентификации.

Приведенные данные позволили предложить нам алгоритм выбора методов описания изучаемых корреляционных зависимостей при проведении исследований, посвященных судебно-медицинской антропологической идентификации личности (рис. 8). Использование данного алгоритма позволяет унифицировать процедуру статистической обработки данных и обеспечивает выбор метода, математическая модель которого в наилучшей степени соответствует их характеру. Указанное обстоятельство, в конечном счете, приведет к созданию регрессионных моделей идентификации личности, обеспечивающих наибольшую точность результатов.

Рис. 8. Алгоритм выбора методов описания исследуемых корреляционных зависимостей при судебно-медицинской антропологической идентификации личности.

2.6. НЕОДНОРОДНАЯ КОРРЕЛЯЦИЯ

Еще одной важной проблемой, возникающей при проведении корреляционного анализа, является неоднородность изучаемых данных за счет наличия в них выбросов или кластеринга. Обычно неоднородность данных приводит к уменьшению реальных корреляционных связей. В отдельных случаях может наблюдаться искусственное увеличение значения коэффициента корреляции, иногда даже приводящее к обнаружению ложных корреляционных зависимостей.

Влияние выбросов на наклон линии регрессии объясняется тем, что при ее построении используется метод наименьших квадратов. Поэтому единичный выброс, значение которого возводится в квадрат, способен существенно изменить наклон прямой и, следовательно, модуль, а иногда и знак коэффициента корреляции. Особенно сильно влияние выбросов проявляется при небольших размерах исследуемых выборок.

Помимо выбросов, наличие кластеринга в выборке также является фактором, смещающим (в ту или иную сторону) выборочную корреляцию. В подобных ситуациях высокая корреляция может быть следствием разбиения данных на две группы, а вовсе не отражать реальную зависимость между двумя переменными, которая может практически отсутствовать (на практике чаще наблюдается обратное явление, когда кластеринг маскирует существующие корреляционные связи).

Изложенное доказывает необходимость проведения проверки любых биометрических данных на возможное наличие их неоднородности. При этом следует различать два подхода к обнаружению неоднородности.

Первый подход заключается в анализе данных литературы и умозрительном разделении совокупностей идентифицируемых объектов на кластеры. В качестве кластерообразующих параметров чаще всего выделяют следующие:

- половой диморфизм;
- возрастная и расовая принадлежность;
- приобретенная патология;
- врожденные аномалии развития;
- асимметрия парных органов или анатомических образований;
- анатомо-топографическая гетеротопия.

Половая принадлежность в большинстве случаев имеет место при изучении антропометрических и остеометрических данных и находит свое отражение при последующем создании регрессионных и дискриминантных моделей идентификации личности [36-38]. Однако выраженность половых различий сильно варьирует вплоть до полного своего отсутствия в зависимости от характера идентифицируемых объектов [33]. Особенно незначительным влияние половой принадлежности оказалось при исследовании гистометрических и гистостереометрических показателей различных органов и тканей [77,81]. Тем не менее, проверка наличия межполовых различий стала одним из стандартных методов выявления неоднородности идентифицируемых объектов.

Возрастная характеристика идентифицируемых объектов имеет значение в связи с нелинейностью протекающих в организме человека процессов старения, последовательно проходящих стадии развития, стабилизации и инволюции [56,67,84]. Наличие указанной стадийности часто побуждает авторов методик идентификации личности вводить соответствующие ограничения к их использованию [64]. Что касается расовой принадлежности, то последняя как фактор неоднородности имеет значение лишь для определенного круга идентифицируемых объектов [32,64].

Приобретенная патология и врожденные аномалии являются одним из наиболее существенных причин неоднородности идентифицируемых объектов. Как правило, авторы большинства способов идентификации личности указывают на недопустимость их использования при наличии указанных патологических состояний [1,64].

Рассмотрение асимметрии парных органов или анатомических образований в качестве фактора неоднородности было вызвано обнаружением асинхронности процессов развития и инволюции в парных анатомических структурах и даже левой и правой половинах одного органа [22,84]. Следствием этого явилось убеждение, что наличие или отсутствие симметрии заранее непредсказуемо и сравнение выраженности изучаемых процессов в парных органах должно стать необходимой частью любой программы судебно-антропологического исследования [6].

Возможность анатомо-топографической неоднородности идентифицируемых объектов также следует учитывать при исследовании их фрагментов. Например, при изучении возрастной динамики лимфоидной ткани червеобразного отростка была обнаружена не-

однородность мышечной оболочки его стенки за счет утолщения ее внутреннего слоя в дистальных отделах органа [56].

Второй подход заключается в обнаружении неоднородности исследуемых объектов статистическими методами. В отличие от предыдущей тактики статистические методы здесь имеют первостепенное значение и применяются именно для выявления выбросов или латентного кластеринга, а не только для объективного доказательства их наличия, предполагаемого теоретически. Большим преимуществом статистических методов выявления неоднородности исследуемых данных является их объективность и независимость от полноты и правильности теоретических рассуждений субъекта научного познания. Это позволяет выявлять скрытые причины неоднородности объектов исследования, которые не были известны до его начала. Иногда выявление факторов неоднородности представляет не меньший научный интерес, чем собственно разработанный способ идентификации.

В этой связи целесообразным является сочетание обоих вышеперечисленных способов выявления неоднородности исследуемых объектов. На этапе планирования исследования в его протокол следует включать проверку данных на возможное наличие выявленных с помощью анализа литературы потенциальных источников кластеринга. В ходе осуществления исследования необходимо использовать статистические методы обнаружения латентной неоднородности изучаемых данных.

Несмотря на перечисленные положительные качества, по выборочным данным статистические методы обнаружения выбросов и кластеринга были использованы лишь в 8% (1) судебно-антропологических исследований (95% доверительный интервал: 0-39%). Это свидетельствует об определенной неосведомленности части исследователей о возможностях названных методов. Данное обстоятельство делает актуальным поиск разработанных специалистами в области математической статистики методов выявления неоднородности данных и последующее их использование при проведении судебно-медицинских антропологических исследований.

К числу основных статистических методов обнаружения выбросов и кластеринга следует отнести следующие:

- проверка эмпирических распределений на согласие с нормальным законом;

- проверка подозрительных экстремальных значений на принадлежность к выбросам;

- кластерный анализ.

Для демонстрации эффективности статистических методов выявления скрытой неоднородности идентифицируемых объектов приведем результаты изучения гестационной динамики кроветворной активности паренхимы печени человека на протяжении 21-40 недель антенатального развития. Данное исследование проводилось нами с целью создания способов судебно-медицинской идентификации гестационного возраста плодов и новорожденных [11,53].

Объектами исследования явились трупы 140 плодов и новорожденных. Учитывая выраженное разнообразие патологических состояний в подлежавшей изучению выборке, зачастую наблюдавшихся у одного и того же плода, вполне обоснованным следовало считать предположение о неоднородности выборочной совокупности значений кроветворной активности. Поэтому первоочередной задачей корреляционно-регрессионного анализа являлось выявление возможной неоднородности выборочных данных. В связи с этим для выявления нетипичных наблюдений была произведена проверка упорядоченных рядов значений кроветворной активности печени у плодов и новорожденных каждой из недель гестации на выбросы всеми упомянутыми методами.

При проверке упорядоченных рядов показателя кроветворной активности у плодов и новорожденных 22-24, 26-32, 35, 37-40 недель гестации отсутствия согласия с нормальным распределением обнаружено не было (табл. 4). Вместе с тем обнаружилось выраженное несоответствие нормальному распределению значений кроветворной активности у плодов и новорожденных 25 недель гестации ($\chi^2 = 38{,}271$, $v = 16$, $p = 0{,}001$; $D = 0{,}183$, Lilliefors $p < 0{,}2$), что свидетельствовало о неоднородности данного показателя в указанной группе наблюдений.

При визуальном анализе диаграммы рассеяния среди значений кроветворной активности у плодов и новорожденных 21, 33, 34 и 36-38 недель гестации, значений, подозрительных на выбросы, не было выявлено (рис. 9). Среди значений кроветворной активности в остальных сроках гестации обнаружились наблюдения, подозрительные на принадлежность к выбросам, что и было проверено.

Выявление единичных выбросов в упорядоченных рядах данных осуществлялось по методу Диксона [30,104]. Если вычисленная M -

статистика превышала критическое значение соответствующего одностороннего критерия при $\alpha < 0{,}05$, то проверяемое значение морфометрического параметра рассматривалось как выброс. Обнаружение группы выбросов с одного из концов упорядоченных рядов выборочных значений проводилось по методу Титьена-Мура [26]. Вычисленные статистики сравнивались с критическими значениями критерия C_α. Группа из k наибольших или наименьших наблюдений признавалась выбросами при $L_k\left(\widetilde{L}_k\right) < C_{0,05}$. Выявление наибольших и наименьших экстремальных наблюдений одновременно в упорядоченных рядах выборочных значений проводилось с использованием специальной модификации метода Граббса [26]. Вычисленные значения статистик E_k сравнивались с критическими значениями C_α, а k рассматриваемых наблюдений признавались выбросами, если $E_k < C_{0,05}$.

Таблица 4

Результаты проверки распределений значений кроветворной активности печени на нормальность

Анализируемый ряд				χ^2 - критерий			$K-C$-критерий**	
y*	n	$\bar{x}$	s	χ^2	v	p	D	p
22	10	96,58	21,51	13,86	14	0,460	0,176	> 0,1
23	13	82,95	29,77	11,594	8	0,170	0,153	> 0,1
24	17	76,13	16,73	7,459	12	0,826	0,098	> 0,1
25	15	66,24	23,75	38,271	16	0,001	0,183	> 0,1
26	12	54,71	17,04	11,906	8	0,155	0,177	> 0,1
27	13	58,16	18,58	12,41	9	0,191	0,108	> 0,1
28	7	28,74	22,53	13,731	10	0,186	0,262	> 0,1
29	5	38,01	32,53	17,941	14	0,209	0,234	> 0,1
30	6	37,48	19,59	11,214	10	0,341	0,250	> 0,1
31	4	35,22	9,44	6,861	6	0,334	0,258	> 0,1
32	4	23,38	6,09	15,243	11	0,172	0,251	> 0,1
35	8	20,87	7,16	13,755	8	0,088	0,200	> 0,1
37	3	9,59	3,52	19,515	13	0,108	0,257	> 0,1
38	5	13,74	2,61	12,564	12	0,402	0,185	> 0,1
39	6	15,60	13,12	12,878	12	0,378	0,164	> 0,1
40	5	6,55	3,57	11,059	8	0,198	0,306	> 0,1

Примечание. * - здесь и в таблице 5: y - гестационный возраст, недель. ** - критерий согласия Колмогорова-Смирнова.

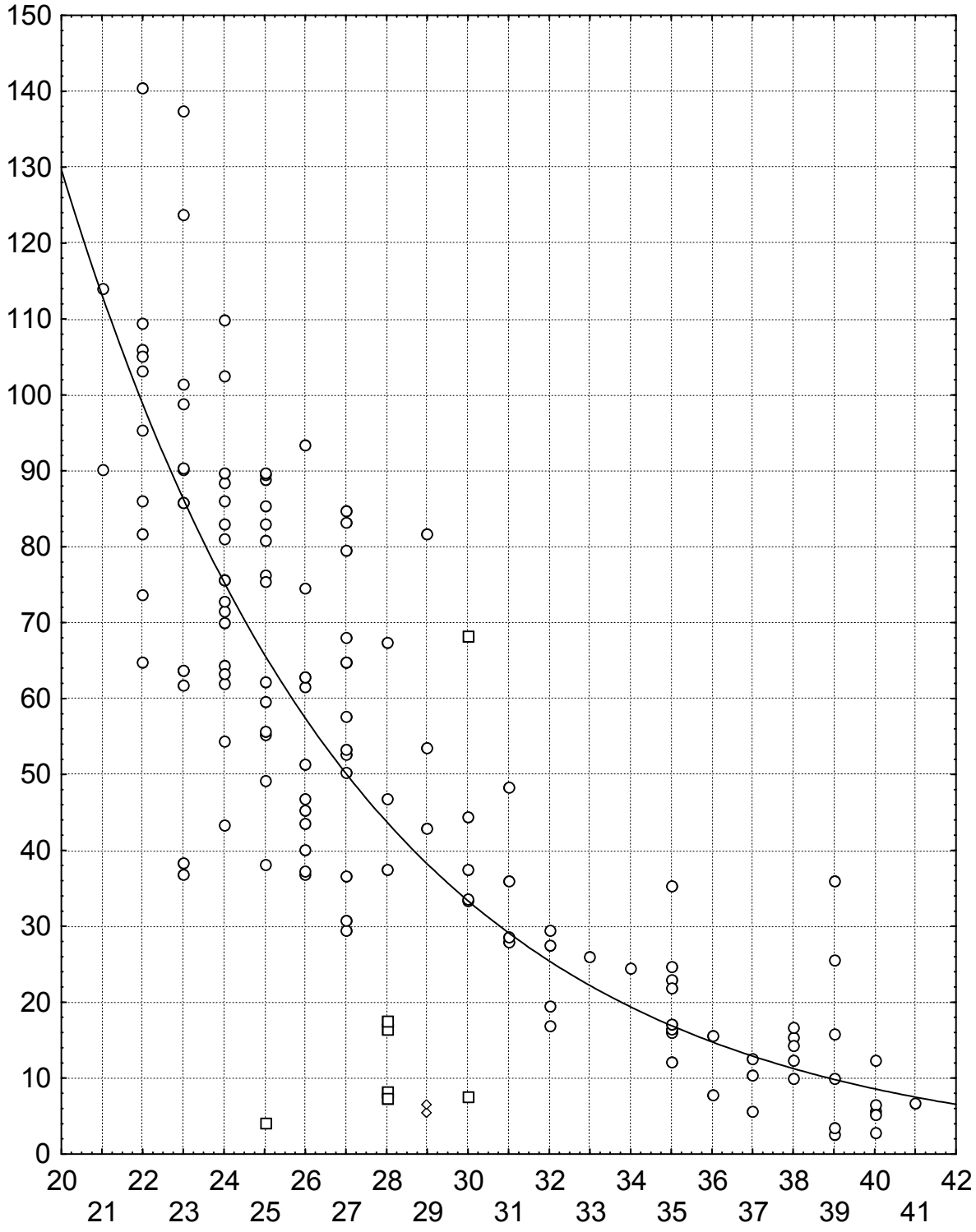

Рис. 9. Зависимость кроветворной активности паренхимы печени от гестационного возраста. По оси абсцисс – гестационный возраст, недель; по оси ординат – кроветворная активность, число профилей ядер. ○ – наблюдения, подвергнутые регрессионному анализу; □ – наблюдения, опознанные как выбросы; ◊ - наблюдения, не опознанные как выбросы, но также принадлежащие кластеру недоношенных новорожденных с постнатальной инволюцией кроветворной ткани печени.

При проверке упорядоченных рядов показателя кроветворной активности у плодов и новорожденных 22-24, 26, 27, 29, 31, 32, 35, 39 и 40 недель гестации выбросов не обнаружено (табл. 5).

Среди значений кроветворной активности в сроке 25 недель гестации опознано как выброс минимальное в данном упорядоченном ряду наблюдение (M = 0,530; $p < 0,05$). Распределение оставшихся в данном упорядоченном ряду после исключения указанного выброса значений кроветворной активности уже соответствовало нормальному закону (χ^2 = 13,800, v = 9, p = 0,130; D = 0,180, $p > 0,1$).

Таблица 5

Результаты проверки упорядоченных рядов показателя кроветворной активности печени на выбросы

Анализируемый ряд		Варианты проверок	Значение статистического критерия		
y	значения x		фактическое	критическое	p
22	$x_1 > x_2 > \ldots > x_{10}$	x_1	$M = 0,460$	$M = 0,477$	$> 0,05$
23	$x_1 > x_2 > \ldots > x_{13}$	x_1 x_1, x_2 x_{12}, x_{13} x_1, x_{12}, x_{13}	$M = 0,363$ $L_2 = 0,487$ $\widetilde{L}_2 = 0,544$ $E_3 = 0,314$	$M = 0,521$ $C = 0,337$ $C = 0,337$ $C = 0,165$	$> 0,05$ $> 0,05$ $> 0,05$ $> 0,05$
24	$x_1 > x_2 > \ldots > x_{17}$	x_1, x_2, x_{17} x_1, x_{17}	$E_3 = 0,272$ $E_2 = 0,442$	$C = 0,248$ $C = 0,362$	$> 0,05$ $> 0,05$
25	$x_1 > x_2 > \ldots > x_{15}$	x_{15}	$M = 0,530$	$M = 0,525$	$< 0,05$
26	$x_1 > x_2 > \ldots > x_{12}$	x_1	$M = 0,545$	$M = 0,546$	$> 0,05$
27	$x_1 > x_2 > \ldots > x_{13}$	x_1, x_2, x_3 x_{11}, x_{12}, x_{13} x_{12}, x_{13}	$L_3 = 0,440$ $\widetilde{L}_3 = 0,370$ $\widetilde{L}_2 = 0,551$	$C = 0,224$ $C = 0,224$ $C = 0,337$	$> 0,05$ $> 0,05$ $> 0,05$
28	$x_1 > x_2 > \ldots > x_7$	x_1, x_2, x_3	$L_3 = 0,029$	$C = 0,032$	$< 0,05$
29	$x_1 > x_2 > \ldots > x_5$	x_4, x_5	$\widetilde{L}_2 = 0,190$	$C = 0,018$	$> 0,05$
30	$x_1 > x_2 > \ldots > x_6$	x_1, x_6	$E_2 = 0,026$	$C = 0,034$	$< 0,05$
31	$x_1 > x_2 > x_3 > x_4$	x_1	$M = 0,627$	$M = 0,765$	$> 0,05$
32	$x_1 > x_2 > x_3 > x_4$	x_1, x_2	$L_2 = 0,033$	$C = 0,001$	$> 0,05$
35	$x_1 > x_2 > \ldots > x_8$	x_1	$M = 0,549$	$M = 0,554$	$> 0,05$
39	$x_1 > x_2 > \ldots > x_6$	x_1	$M = 0,306$	$M = 0,560$	$> 0,05$
40	$x_1 > x_2 > \ldots > x_5$	x_1	$M = 0,623$	$M = 0,642$	$> 0,05$

В сроке 30 недель гестации выбросами признаны максимальное и минимальное экстремальные значения кроветворной активности (E_2 = 0,026; $p < 0,05$). При визуальном анализе диаграммы рассеяния отмечалось деление данных кроветворной активности печени у плодов и новорожденных с гестационным возрастом, равным 28 неделям (n = 7), на две неоднородных группы, численностью из 3-х и 4-х наблюдений (см. рис. 9). Данная гипотеза также подтвердилась путем проверки группы из трех максимальных наблюдений на выбросы (L_3 = 0,029; $p < 0,05$). Однако рассмотрение всех изученных наблюдений показывает, что относительно всей совокупности данных (n = 140) нетипичными для срока в 28 недель гестации являются как раз не максимальные значения кроветворной активности, а группа из четырех минимальных ее значений.

Таким образом, проведенная проверка выявила 7 выбросов: один в сроке 25 недель гестации, четыре – в 28 недель и два - в 30 недель. В области максимальных значений упорядоченных рядов обнаружился только один выброс, который имел место в случае искусственного прерывания беременности из-за наличия множественных врожденных пороков развития (тяжелые аномалии мочеполовой системы с персистенцией протока аллантоиса и гипоплазией легких, дефект межжелудочковой перегородки, аплазия пупочной артерии, гипоплазия желудка). Из всех плодов и новорожденных с врожденными пороками развития данное наблюдение было единственным, в котором четко прослеживалась инфекционная этиология пороков в виде следов перенесенного в эмбриональном периоде экссудативного перитонита.

Остальные экстремальные значения были выявлены только в области минимальных концов упорядоченных рядов значений кроветворной активности. Каждому из них соответствовали преждевременные роды в сроке 25-30 недель гестации незрелым плодом с последующим развитием у последнего сразу после рождения тяжелой респираторной патологии (гиалиновые мембраны, первичный ателектаз, врожденная пневмония). Вследствие дыхательной недостаточности новорожденным проводилась аппаратная искусственная вентиляция легких. У 5 (83%) новорожденных постнатальная асфиксия осложнилась развитием внутрижелудочковых кровоизлияний. Наиболее характерным моментом для указанной группы недоношенных новорожденных явилась продолжительность внеутробной жизни, превысившая 48 ч.

На диаграмме рассеяния (см. рис. 9) видно, что помимо указанной группы выбросов в сроке 29 недель гестации имеются еще два низких значения кроветворной активности, соответствующие новорожденным с аналогичными клиническими проявлениями и продолжительностью неонатальной жизни. Указанные наблюдения, вероятно, не были опознаны как экстремальные из-за малого объема выборки (всего пять значений, два из которых предположительно являются выбросами).

Для проверки данной гипотезы использовалась процедура кластерного анализа. Объектами иерархической классификации являлись 33 плода 25, 28-30 недель гестации. Объекты исследования характеризовались двумя признаками: кроветворной активностью паренхимы печени и гестационным возрастом. Таким образом, все объекты кластерного анализа представляли собой одну двумерную совокупность признаков, состоявшую из 33 пар значений показателя кроветворной активности печени и гестационного возраста. Данная совокупность была упорядочена в порядке возрастания значений кроветворной активности. При этом каждому объекту присваивался номер, соответствующий положению показателя кроветворной активности в упорядоченном ряду. В результате проведенного ранжирования всем шести новорожденным с минимальными экстремальными значениями кроветворной активности, а также двум новорожденным с низкими значениями кроветворной активности в сроке 29 недель гестации были присвоены ранги с 1-го по 8-й. Плоду с максимальным экстремальным значением кроветворной активности в сроке 30 недель гестации был присвоен 24-й ранг.

Результаты иерархической классификации представлены в виде дендрограммы, которая показывает, что предпочтение следует отдать этапу классификации, на котором все наблюдения объединены в два кластера, состоящие из 8-ми и 25-ти объектов (рис. 10). Причем кластер, состоящий из 8-ми объектов, представлен шестью новорожденными с минимальными экстремальными значениями кроветворной активности и двумя новорожденными 29 недель гестации с минимальными значениями кроветворной активности среди плодов данного гестационного срока. Все остальные плоды и новорожденные были выделены во второй кластер, отделенный от первого значительным межкластерным расстоянием ($\rho_{(1-8),(9-33)} = 0{,}590$). Плод 30 недель гестации с максимальным экстремальным значением кроветворной активности в отдельный кластер выделен не был.

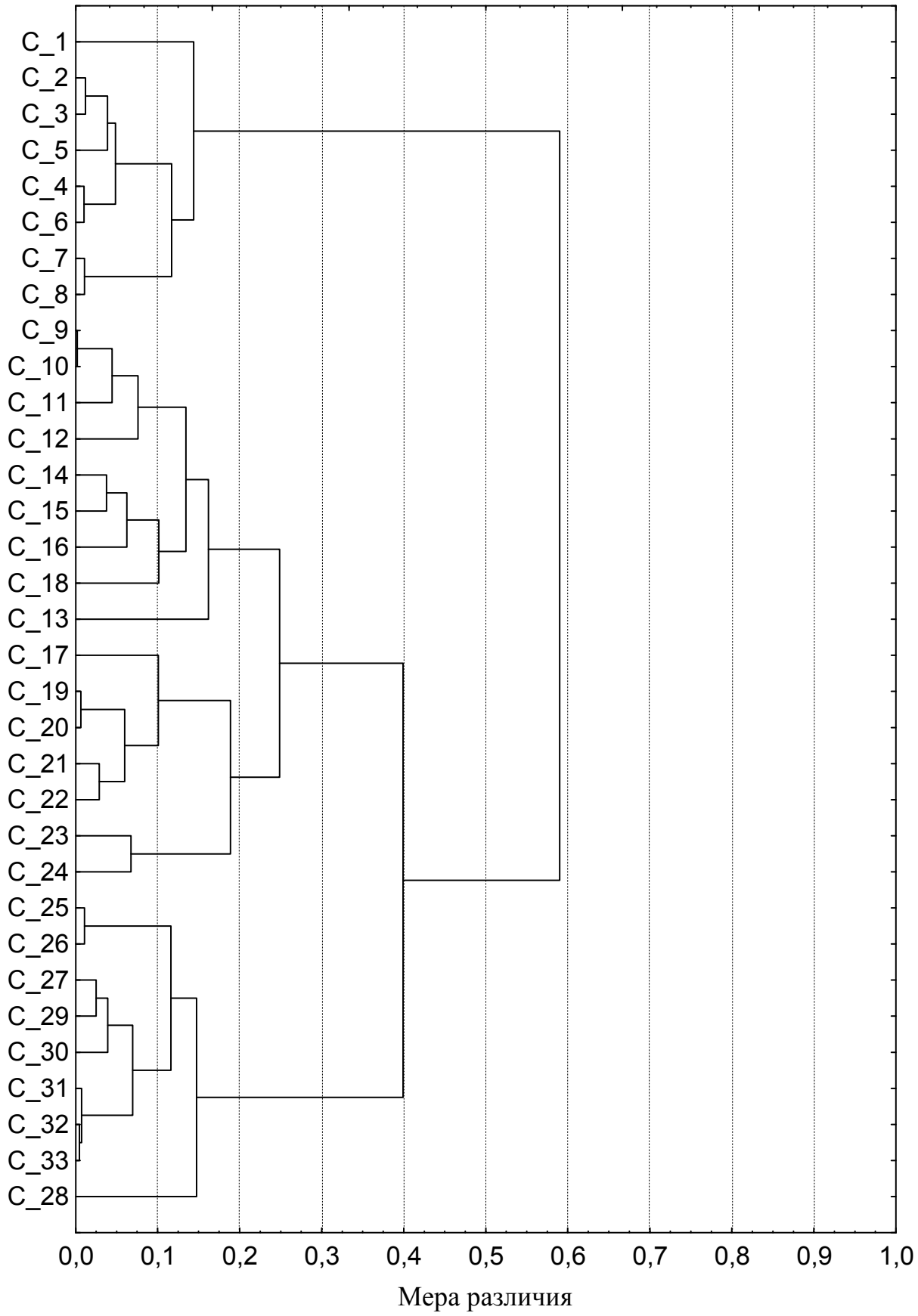

Рис. 10. Дендрограмма агломеративной иерархической классификации плодов и новорожденных 25, 28-30 недель гестации по степени кроветворной активности паренхимы печени и показателю гестационного возраста.

Полученные данные свидетельствовали о лавинообразном снижении миелоидной инфильтрации печени незрелых новорожденных на протяжении 3-5 суток раннего неонатального периода, что было подтверждено методами сравнительного анализа (рис. 11).

Таким образом, проведенная проверка исследуемых объектов на неоднородность выявила и доказала факт существования неоднородности кроветворной активности печени вследствие наличия отдельного кластера глубоко недоношенных новорожденных с выраженным постнатальным опустошением экстрамедуллярной кроветворной ткани. Это послужило причиной дальнейшего дифференцированного изучения гестационной динамики показателя кроветворной активности у мертворожденных плодов и постнатальной динамики данного показателя у недоношенных новорожденных.

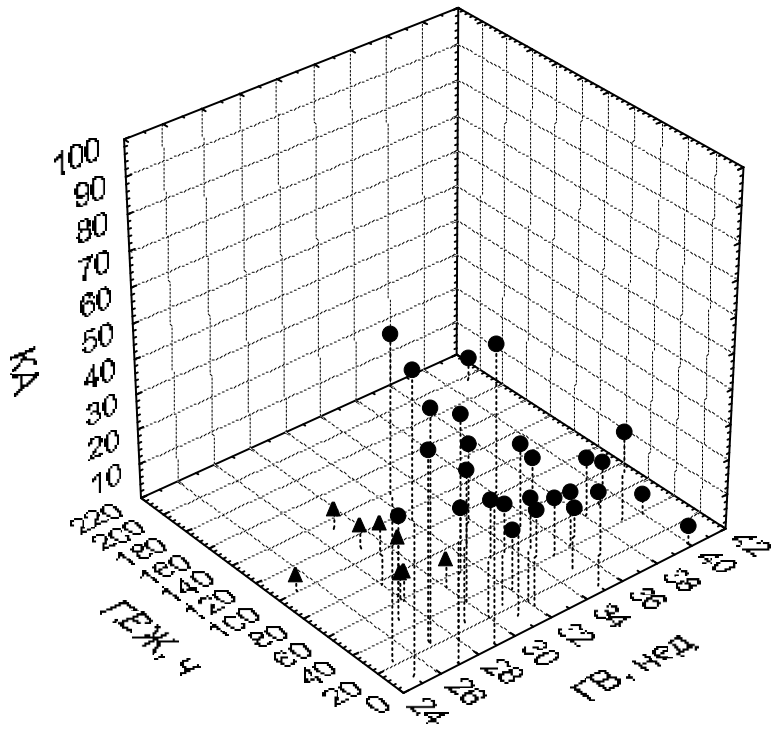

Рис. 11. Зависимость кроветворной активности (КА) паренхимы печени от гестационного возраста (ГВ) и продолжительности внеутробной жизни (ПВЖ) новорожденных. • - значения кроветворной активности у новорожденных 25-30 недель гестации с продолжительностью внеутробной жизни до 48 ч и у новорожденных 31-41 недель гестации с продолжительностью внеутробной жизни до 196 ч; ▲ – значения кроветворной активности у недоношенных новорожденных 25-30 недель гестации с продолжительностью внеутробной жизни более 48 ч ($U = 35$; $p < 0{,}005$).

После обнаружения неоднородности данных еще более важной проблемой корреляционно-регрессионного анализа, чем поиск выбросов и кластеринга, является тактика исследователя относительно их исключения из выборочной совокупности. Обычно считается, что выбросы представляют собой случайную ошибку, которую следует контролировать. Однако в настоящее время не только в судебно-медицинской антропологии, но и в других областях знания пока не существует определенного правила относительно необходимости удаления выбросов, и указанное решение рекомендуется принимать индивидуально в каждом исследовании с учетом его особенностей и сложившейся практики в данной области [13,74].

На наш взгляд, решение об исключении выбросов и кластеринга должно приниматься в зависимости от поставленной задачи исследования. Если таковой является преимущественно теоретическое изучение изменчивости каких-либо биометрических показателей и определение границ нормы, то любые проявления неоднородности исследуемых объектов должны быть выявлены, по возможности объяснены и в дальнейшем исключены из анализа. Например, при изучении возрастной динамики аппендикулярной лимфоидной ткани помимо топографической неоднородности толщины мышечной оболочки червеобразного отростка было обнаружено 3 выброса, при этом причина экстремальности была объяснена только у одного из них [56]. Несмотря на это, для устранения возможности искажения возрастной динамики аппендикулярной лимфоидной ткани и границ ее возрастной нормы данные наблюдения были исключены из корреляционно-регрессионного анализа.

Если задачей исследования является создание способа идентификации личности, то вопрос об исключении выбросов или кластеринга должен решаться в зависимости от возможности определения их принадлежности к отличающемуся кластеру и степени отличия от большинства объектов данного типа (рис. 12). Определенную роль может также играть степень вероятности поступления объектов из отличающихся кластеров на экспертизу идентификации личности.

Наибольшие трудности для проведения корреляционного анализа представляет сочетание научных задач обоих охарактеризованных типов. Такое сочетание, очевидно, требует дифференцированного подхода к проведению корреляционного анализа: с наличием неоднородности и без таковой.

Дифференцированный поход к решению диагностической задачи был применен нами при создании способов идентификации гестационного возраста плодов и новорожденных. Это позволило выявить и объяснить причины неоднородности исследовавшихся объектов, разработать комплекс регрессионных моделей идентификации гестационного возраста, удовлетворяющих всему многообразию исходных условий для идентификации (наличие или отсутствие достоверных данных о живорожденности и продолжительности внеутробной жизни плода, степень фрагментации трупа, выраженность гнилостных изменений и др.), а также создать альтернативные способы установления других идентифицируемых параметров (биологический возраст) [53].

Рис. 12. Алгоритм принятия решений относительно исключения выбросов или кластеринга из корреляционного анализа при проведении судебно-медицинских антропологических исследований, посвященных созданию способов идентификации личности.

2.7. СРАВНИТЕЛЬНЫЙ АНАЛИЗ ПАРАМЕТРОВ СВЯЗИ И АНАЛИЗ МОЩНОСТИ КОРРЕЛЯЦИОННОГО АНАЛИЗА

Кроме нахождения интервальной оценки для ρ с помощью преобразования Р.А. Фишера (6) можно также решать еще ряд задач, имеющих важное значение для судебно-медицинской антропологической идентификации: сравнение двух или нескольких выборочных коэффициентов корреляции и определение мощности (чувствительности) корреляционного анализа.

Уровень значимости корреляции представляет собой главный источник информации о ее надежности. Мощность (чувствительность) $1 - \beta$, необходимая для обнаружения линейной корреляции, не меньшей ρ, при уровне значимости α зависит от абсолютного значения ρ и объема выборки n:

$$z_{1-\beta} = z_\alpha - \frac{z_\rho}{s_z} = z_\alpha - 0{,}5 \ln\left(\frac{1+\rho}{1-\rho}\right) \cdot \sqrt{n-3}$$ [16,114].

Расчет чувствительности корреляционного анализа можно показать на примере исследования возможности диагностики роста человека с помощью остеометрии подъязычной кости с использованием данных В.Н. Звягина, Н.Л. Мальцевой, Л.А. Алексиной и О.И. Галицкой, полученных при исследовании 158 указанных костных объектов [41]. Программа остеометрии подъязычной кости, выполненная указанными авторами, включала определение 16 размеров, обозначенных ими как М1 – М16. Коэффициенты парной корреляции некоторых остеометрических показателей подъязычной кости с ростом приведены в таблице 6.

Таблица 6

Возрастная динамика некоторых остеометрических показателей подъязычной кости (по данным В.Н. Звягина, Н.Л. Мальцевой, Л.А. Алексиной и О.И. Галицкой [41] с дополнениями)

Показатель	n	r	t	p	$1-\beta$
М2	158	0,467	6,596	$6{,}198 \cdot 10^{-10}$	1,000
М12	158	0,407	5,565	$1{,}115 \cdot 10^{-7}$	1,000
М16	158	0,524	7,684	$1{,}602 \cdot 10^{-12}$	1,000
М3	158	0,336	4,456	$1{,}588 \cdot 10^{-5}$	0,992
М10	158	-0,107	-1,344	0,181	0,267

Проверка приведенных коэффициентов показала наличие статистической значимости корреляционных связей для признаков М2, М12, М16, М3 и отсутствие таковой для признака М10 (см. табл. 6). Допустим, что зависимость признака М10 с ростом существует. Определим вероятность ее выявления при указанном объеме выборки.

Для признака М10 данная вероятность равна

$$z_{1-\beta} = 1{,}960 - 0{,}5\ln\left(\frac{1+0{,}107}{1-0{,}107}\right)\cdot\sqrt{158-3} = 0{,}623,\ 1-\beta = 0{,}267.$$

Результаты определения чувствительности корреляционного анализа для остальных остеометрических показателей подъязычной кости выполняются аналогично (см. табл. 6).

Для планирования объема выборки, при котором чувствительность корреляционного анализа достигалась бы величины $1 - \beta$, рекомендуется использование следующей формулы:

$$n = \left(\frac{z_\alpha - z_{1-\beta}}{z_\rho}\right)^2 + 3 \text{ [16].}$$

Например, на рисунке 13 приведены графики чувствительности корреляционного анализа взаимосвязи признака М10 подъязычной кости с длиной тела в зависимости от объема наблюдений, абсолютного значения коэффициента корреляции и уровня значимости.

Важно, что дисперсия s_z^2 принимает разные значения при анализе парной линейной - $s_z^2 = \frac{1}{\sqrt{n-3}}$, множественной линейной - $s_z^2 = \frac{1}{n}$ и парной ранговой - $s_z^2 = \frac{1}{n-1}$ корреляций [26]. Это следует учитывать при определении чувствительности корреляционного анализа, планирования оптимального объема наблюдений и определении интервальных оценок указанных коэффициентов.

Напомним, что критерии значимости коэффициентов корреляции основываются на предположении, что распределение остатков является нормальным с постоянной дисперсией для всех значений независимой переменной. Нарушение этих условий не является слишком критичным, если размеры выборок достаточно велики, а отклонения от нормальности не слишком заметны [13]. Например, значимость ранговой корреляции Спирмена можно тестировать с помощью *t*-распределения уже при $n > 50$ [16].

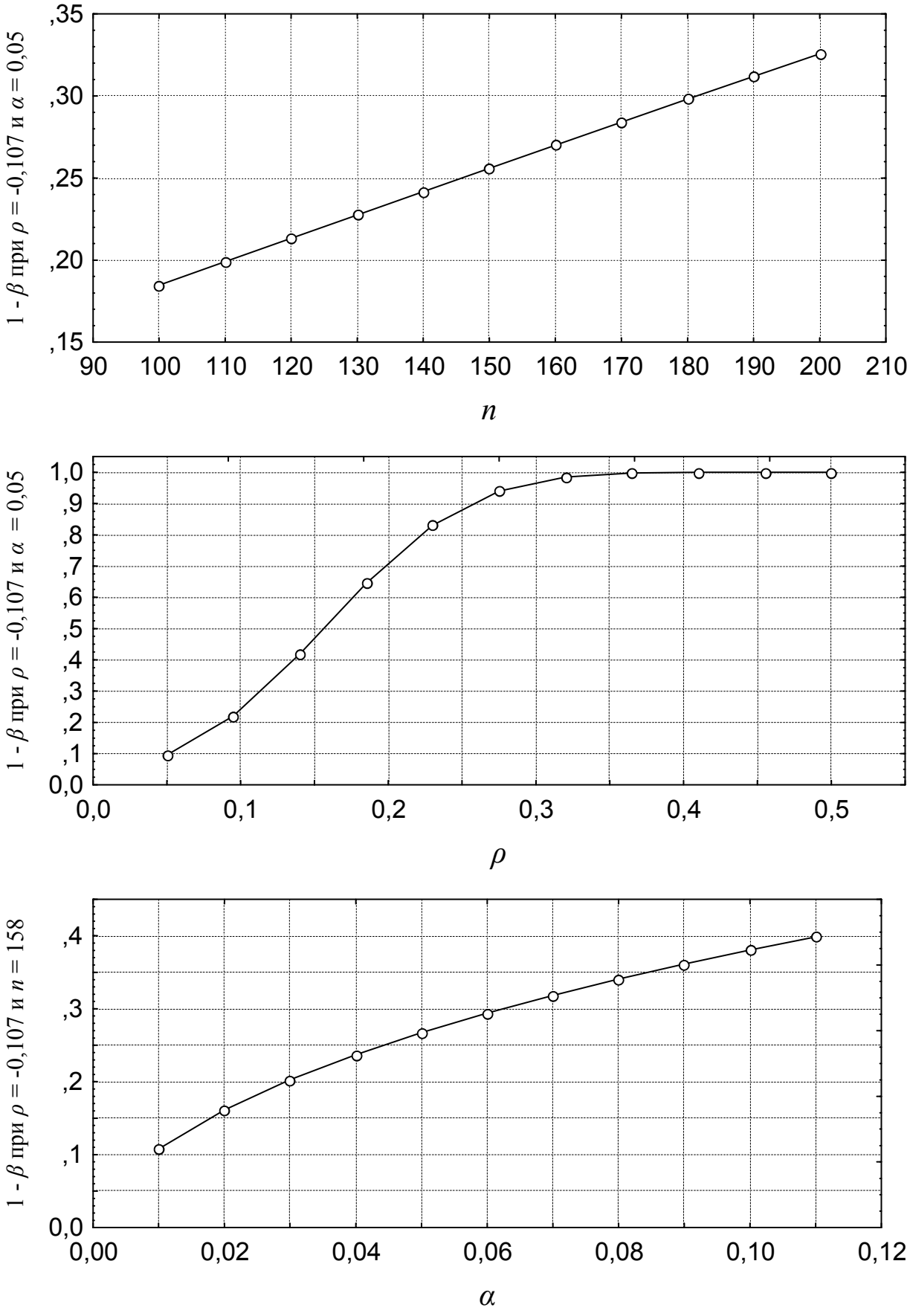

Рис. 13. Мощность корреляционного анализа при ρ = -0,107, n = 158, α = 0,05 в зависимости от объема выборки, величины корреляции и уровня значимости.

Выборочный анализ судебно-медицинских антропологических исследований показал, что объемы отдельных кластеров, использованные для проведения корреляционного анализа, варьировали от 85 до 564 членов (рис. 14). Такие объемы выборок следует охарактеризовать как достаточные для выявления даже слабых корреляционных зависимостей. Нетрудно показать, что даже наименьший из всех исследований объем выборки (n = 85) позволяет с достаточно большой надежностью выявлять все корреляционные зависимости с абсолютным значением коэффициента корреляции $\rho \geq 0{,}3$ (рис. 15). Возможно, именно по этой причине ни в одном из проанализированных исследований не производился расчет чувствительности корреляционного анализа, так как отсутствие значимости точечных оценок коэффициентов корреляции при такой степени чувствительности могло означать только отсутствие зависимости.

Помимо определения чувствительности корреляционного анализа иногда задачей судебно-антропологического исследования может являться сравнение двух или нескольких оцененных коэффициентов корреляции.

Сравнение нескольких оценок коэффициентов корреляции осуществляется по формуле

$$\chi^2 = \sum_{i=1}^{k} (n_i - 3)(z_{ri} - \bar{z})^2 , \qquad (8)$$

где k – количество оценок коэффициентов корреляции r_i; n_i – объем i-й выборки; $\bar{z} = \dfrac{\sum_{i=1}^{k} z_i (n_i - 3)}{\sum_{i=1}^{k} (n_i - 3)}$ [30].

Если полученная χ^2- статистика меньше границы значимости при $(k-1)$ степенях свободы, то нуль-гипотеза $\rho_1 = \rho_2 = \cdots = \rho_k$ не отклоняется.

Принципы сравнения нескольких точечных оценок коэффициентов корреляции можно показать на примере недавно проведенного исследования возрастной динамики коэффициента сократимости кожи различных областей тела [71]. Авторы указанной работы доказали наличие отрицательной корреляции с возрастом коэффициентов сократимости образцов кожи с 7 исследовавшихся областей тела человека.

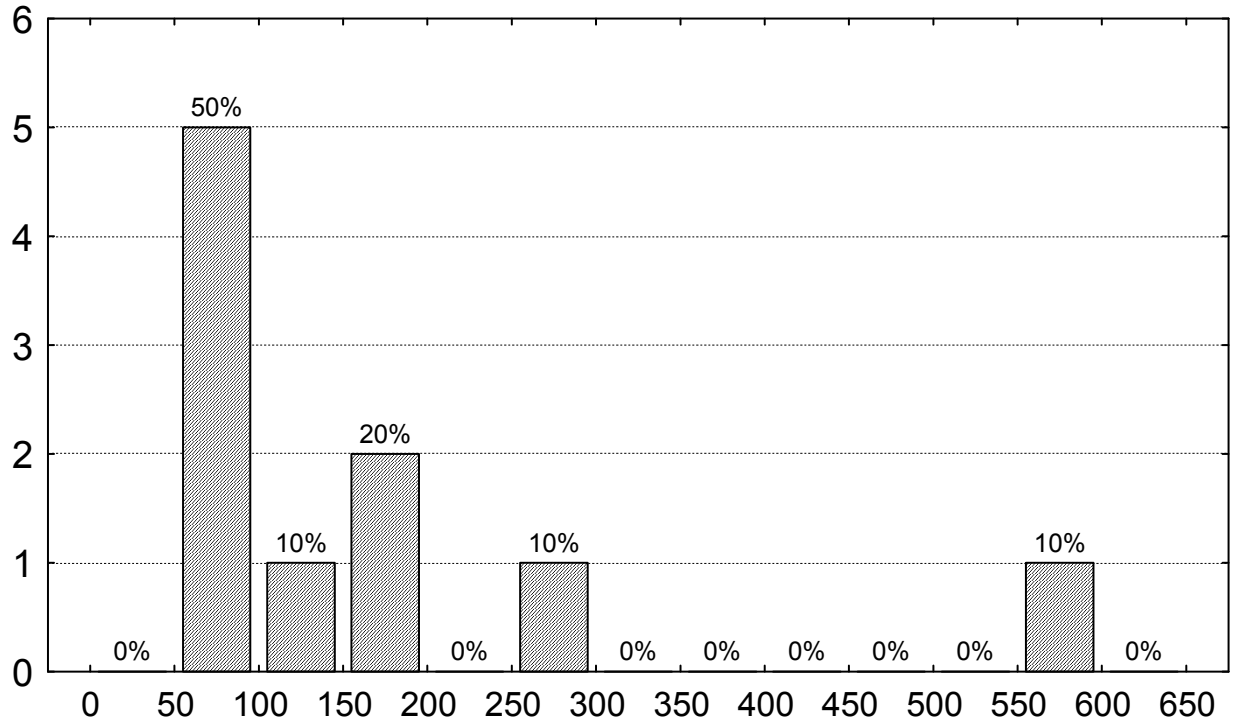

Рис. 14. Распределение значений минимальных объемов отдельных кластеров, использованных в судебно-медицинских антропологических исследованиях для проведения корреляционного анализа. По оси абсцисс – минимальный объем выборок, по оси ординат – абсолютная частота.

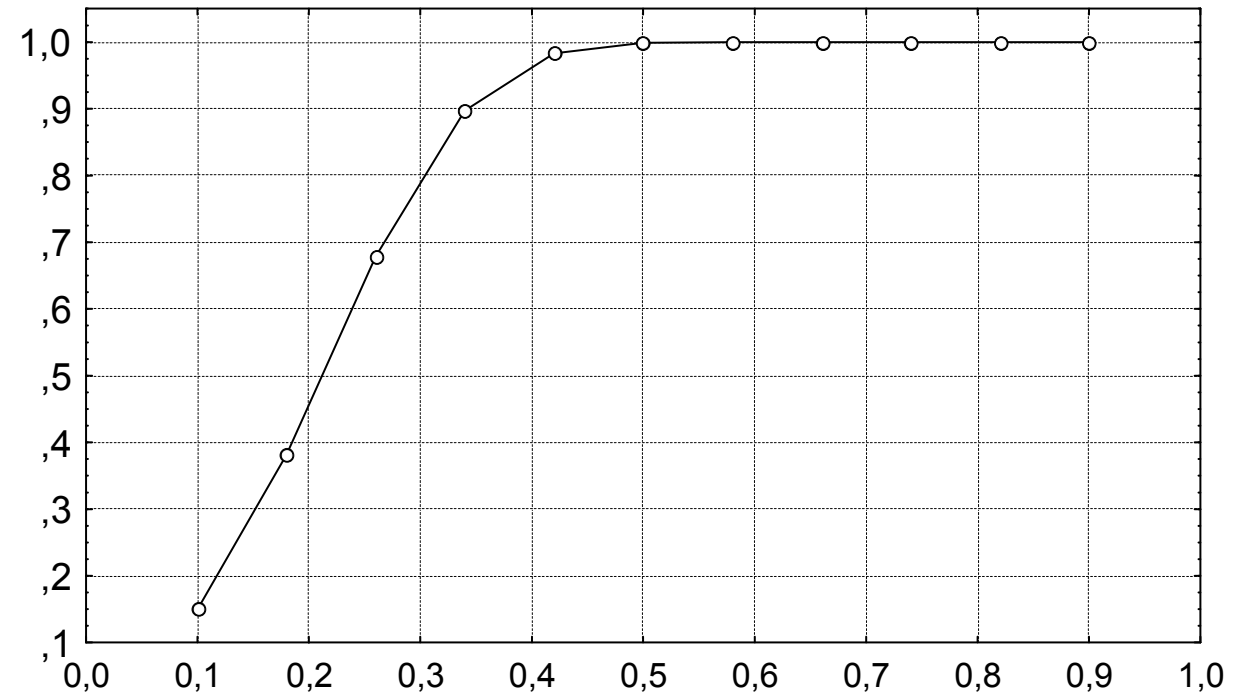

Рис. 15. Зависимость чувствительности корреляционного анализа от абсолютного значения коэффициента корреляции при n = 85 и α = 0,05. По оси абсцисс – модуль коэффициента корреляции, по оси ординат – мощность.

Вместе с тем остался неизученным вопрос о возможных различиях возрастной динамики сократимости кожи в различных областях тела. Восполним данный пробел. Показатели линейной корреляции коэффициентов сократимости кожи каждой из исследовавшихся областей тела с возрастом и промежуточные вычисления приведены в таблице 7.

Оценим

$$\bar{z} = \frac{\sum_{i=1}^{k} z_i (n_i - 3)}{\sum_{i=1}^{k} (n_i - 3)} = \frac{731{,}699}{679} = 1{,}078,$$

откуда $\chi^2 = \sum_{i=1}^{k} (n_i - 3)(z_{ri} - \bar{z})^2 = 44{,}791,\ p = 5{,}151 \cdot 10^{-8}$.

Таким образом, проведенный анализ доказал наличие регионарных различий в возрастной динамике коэффициентов сократимости кожи. Вместе с тем остается неясным, возрастная динамика каких именно регионов тела отличается от таковой других областей. Для этого необходимо осуществить попарные сравнения соответствующих коэффициентов корреляции.

Сравнение двух оцененных коэффициентов корреляции r_1 и r_2 производится по формуле

$$z = \frac{|z_{r1} - z_{r2}|}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}, \qquad (9)$$

которую для выборок равного объема можно представить как

Таблица 7

Возрастная динамика коэффициентов сократимости кожи различных областей тела человека (по данным Е.Н. Савенковой, Ю.А. Неклюдова и А.А. Ефимова [71] с дополнениями)

Область тела	n	r	$\lvert r \rvert$	z_r	$z_r(n-3)$	$z_r - \bar{z}$	$(n-3)(z_r - \bar{z})$
Бедро	100	-0,88	0,88	1,376	133,449	0,089	8,623
Предплечья	100	-0,86	0,86	1,293	125,454	0,047	4,514
Шея	100	-0,84	0,84	1,221	118,454	0,021	1,999
Спина	100	-0,80	0,8	1,099	106,565	0,000	0,043
Грудь	100	-0,78	0,78	1,045	101,401	0,001	0,101
Ягодицы	100	-0,75	0,75	0,973	94,377	0,011	1,062
Стопа	100	-0,49	0,49	0,536	51,998	0,293	28,448

$$z = \frac{\left| z_{r1} - z_{r2} \right|}{\sqrt{\frac{2}{n-3}}}. \qquad (10)$$

где z – стандартная нормальная переменная [30]. Если полученное отношение меньше, чем границы значимости, то предполагается равенство параметров ρ_1 и ρ_2 $(\rho_1 = \rho_2)$. В зависимости от варианта проверки для расчетов могут быть использованы одно- или двусторонние варианты стандартной нормальной переменной.

Для примера приведем результаты сравнения возрастной динамики объема головного мозга у мужчин и у женщин, выполненного нами в аспекте изучения влияния церебральной атрофии на клиническое течение и экспертную оценку травматического сдавления головного мозга. В рамках данного исследования помимо других биометрических показателей было произведено измерение объема головного мозга от трупов 61 мужчины и 32 женщин, умерших в возрасте 18-92 лет. В исследуемые группы не включались лица с наличием травматических или каких-либо других патологических изменений черепа, вне – и внутричерепных образований, кроме атеросклероза артерий головного мозга при отсутствии его инфарктов и внутримозговых кровоизлияний любого объема и сроков организации. Для исключения возможного влияния на объем головного мозга каких-либо медицинских вмешательств (инфузионная терапия, реанимационные мероприятия) в исследуемые группы также не включались лица, смерть которых наступила в стационаре. Относительная погрешность определения объема головного мозга в обеих группах не превышала 1%.

Проведенный корреляционный анализ показал наличие умеренно выраженной отрицательной возрастной динамики объема головного мозга как у мужчин (r = -0,341; t = -2,785; p = 0,007), так и у женщин (r = -0,417; t = -2,512; p = 0,018). Для проверки возможного наличия различий в возрастной динамике с помощью выражения (9) произведем сравнение коэффициентов корреляции, оцененных у мужчин (r = -0,341) и у женщин (r = -0,417):

$$z_{r1} = 0{,}5 \ln \frac{0{,}341 + 1}{1 - 0{,}341} = 0{,}355;$$

$$z_{r2} = 0{,}5 \ln \frac{0{,}417 + 1}{1 - 0{,}417} = 0{,}444;$$

$$z = \frac{|0{,}444 - 0{,}355|}{\sqrt{\frac{1}{61-3} + \frac{1}{32-3}}} = 0{,}391.$$

Откуда $p = 0{,}348$ для одностороннего и $p = 0{,}696$ для двустороннего вариантов проверки.

Таким образом, сравнительный анализ значимых межполовых различий в скорости развития церебральной атрофии не обнаружил. Отсутствие указанных различий вызвало необходимость охарактеризовать возрастную динамику объема головного мозга с помощью одного корреляционного коэффициента.

Однако коэффициенты корреляции не обладают свойством аддитивности, т.е. усредненный коэффициент корреляции, вычисленный по нескольким выборкам, не будет совпадать с корреляцией, вычисленной по объединенной выборке. Это объясняется тем, что коэффициент корреляции не является линейной функцией величины зависимости между переменными [13]. Кроме того, отсутствие значимости корреляций не означает схожесть регрессий. В данном случае регрессии, описывающие возрастную динамику объема головного мозга у мужчин и у женщин не отличаясь по коэффициентам наклона, отличаются по коэффициентам сдвига, т.е. не совпадают, но являются параллельными. Поэтому выборочные совокупности мужчин и женщин нельзя объединять в одну выборку для получения «среднего» коэффициента корреляции. Для этого следует преобразовать коэффициенты корреляции в какую-нибудь аддитивную меру зависимости. Таковыми являются коэффициенты детерминации, а также z_r-преобразование Р.А. Фишера. Это позволяет охарактеризовать возрастную динамику объема головного мозга одновременно у лиц обоих полов с помощью совместного коэффициента корреляции, вычисляемого по формуле:

$$\bar{z} = \frac{\sum_{i=1}^{k} z_i (n_i - 3)}{\sum_{i=1}^{k} (n_i - 3)}.$$

В данном случае $\bar{z} = 0{,}385$, $\bar{r} = -0{,}367$.

Нетрудно заметить, что полученный совместный коэффициент корреляции по модулю ближе к корреляции, оцененной по выборке

мужчин, характеризовавшейся большим объемом и имевший вследствие этого больший вес.

Выведем также формулу чувствительности сравнительного анализа двух коэффициентов корреляции. Учитывая, что величина $z = \Delta z_r / s_{\Delta z}$ подчинена стандартному нормальному распределению, то

$$z_{1-\beta} = \frac{z_\alpha s_{\Delta z} - \Delta z_r}{s_{\Delta z}}, \text{ где } s_{\Delta z} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}.$$

Следует отметить, что чувствительность сравнительного анализа одного или нескольких коэффициентов корреляции зависит не только от величины разности этих коэффициентов, но и от объема выборок. Поэтому, чем больше объем выборки, тем меньший эффект можно значимо обнаружить. Кроме того, на чувствительность корреляционного анализа влияют также и величины самих сравниваемых коэффициентов корреляции. Поскольку надежность коэффициента корреляции растет с увеличением его абсолютного значения, между большими коэффициентами корреляции могут быть значимыми даже относительно малые различия (рис. 16).

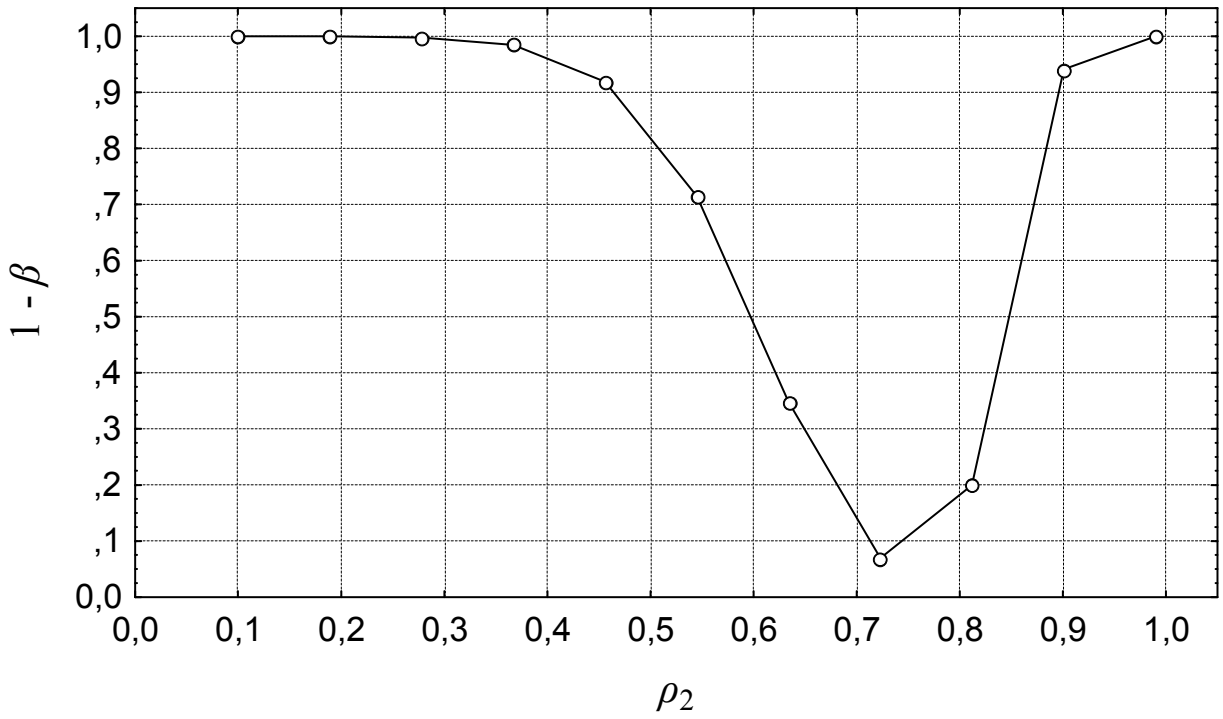

Рис. 16. Мощность корреляционного анализа при $\rho_1 = 0{,}75$, $n_1 = n_2 = 100$ и $\alpha = 0{,}05$ в зависимости от величины сравниваемого коэффициента корреляции ρ_2.

Изложенное делает необходимым завершение любого корреляционного анализа с незначимым результатом расчетом мощности.

Таким образом, основная задача корреляционного анализа при проведении судебно-медицинского антропологического исследования состоит в оценке $k(k+3)/2$ параметров, определяющих k-мерный нормальный закон распределения генеральной совокупности X по выборке. При этом один из компонентов k-мерной выборки представлен вектором выборочных значений идентифицируемого параметра, а остальные $k-1$ компонентов – векторами идентифицирующих признаков.

Для значимых парных коэффициентов корреляции идентифицируемого параметра и идентифицирующего показателя следует указать более предпочтительные точечные или интервальные оценки, а для незначимых – вычислить чувствительность корреляционного анализа.

Далее следует оценить и проверить значимость множественных коэффициентов корреляции или детерминации идентифицируемого параметра с системой идентифицирующих признаков. Для выяснения «чистых» взаимосвязей и истинных взаимозависимостей следует проанализировать выборочные частные коэффициенты корреляции.

Учитывая необходимость предварительного разведочного анализа эмпирических данных на предмет наличия их неоднородности, а также возможной нелинейности исследуемых взаимосвязей, при осуществлении судебно-медицинских антропологических исследований целесообразно использовать следующую оптимальную стратегию корреляционного анализа, приведенную на рисунке 17.

Необходимо подчеркнуть, что о статистической взаимозависимости говорят при обнаружении значимых корреляционных связей. Физический же смысл найденных статистических взаимозависимостей лежит за пределами статистического анализа. Объясняется это тем, что корреляция может быть обусловлена непосредственной причинной зависимостью между переменными, их общей зависимостью от третьей величины, неоднородностью материала или быть чисто формальной. Поэтому с помощью корреляционного анализа нельзя строго доказать наличие причинной зависимости между исследуемыми параметрами, можно лишь определить совместные корреляции, обусловленные влиянием других факторов, оставшихся вне рамок исследования.

Рис. 17. Оптимальная стратегия корреляционного анализа при судебно-медицинской антропологической идентификации личности.

Основная проблема совместной корреляции состоит в неопределенности вызвавшей ее причины. Например, установлено, что заболеваемость раком молочной железы связана с уровнем доходов, числом автомобилей и телевизоров в семье, схожие факторы риска установлены при раке толстой кишки [105]. Однако на основании таких данных правомочно лишь предположение, что какой-то фактор, связанный с уровнем жизни, влияет на риск рака молочной железы [16]. Значительное количество подобных корреляций было обнаружено в исследованиях, посвященных выявлению факторов риска синдрома внезапной смерти у грудных детей [15]. Известны такие эффектные примеры совместной корреляции как уменьшение числа гнезд аистов и числа новорожденных, увеличение числа радиослушателей и количества умственно отсталых людей [30,73].

Кроме совместных необходимо также различать ложные (ошибочные) и формальные (бессмысленные) корреляции.

Причинами ложных корреляций являются несоответствие исходных данных математической модели корреляционного анализа (неоднородность исходных данных, нелинейность зависимости, качественный или порядковый характер переменных), систематические ошибки, связанные с другими дефектами дизайна исследования, а также множественность оцениваний, в том числе и различных выборок.

В какой-то мере увеличение частоты обнаружения ложных корреляций в биомедицинских и других исследованиях связано с развитием компьютерной техники. В большинстве современных исследований первый шаг анализа состоит в вычислении корреляционной матрицы всех переменных и проверке значимых (ожидаемых и неожиданных) корреляций. Однако следует иметь в виду, что количество обнаруженных значимых оценок всегда прямо пропорционально количеству произведенных статистических оцениваний [13,16]. Например, в судебно-медицинской антропологии встречаются исследования, в которых было выявлено (и еще больше проверено на значимость) более двух с половиной сотен значимых корреляционных зависимостей, вероятность статистической ошибки для которых варьировала в диапазоне до $p = 0,05$. При таком количестве оцениваний можно ожидать обнаружение ложной значимости 2-3 коэффициентов корреляции. Такое положение является общим для всех методов анализа, применяющих «множественные сравнения и статистическую значимость».

Используя методы корреляционно-регрессионного анализа, иногда можно выявить взаимосвязи, которые противоречат здравому смыслу (формальные корреляции). Примером формальной корреляции является сравнение вероисповедания и роста людей, согласно которому при движении от Шотландии к Сицилии доля католиков в населении постепенно возрастает, а средний рост людей в то же время убывает [73]. Благоприятным фоном для обнаружения формальных корреляций является также исследование взаимосвязей между группами факторов. Вычисление совместных и формальных корреляций впервые было отмечено еще на рубеже XIX-XX веков, причем возникновение подобных абсурдных результатов чуть не дискредитировало всю математическую статистику [73].

В некотором роде противоположным заблуждением является расценивание некоррелированности факторов как их независимость. Это означает, что из независимости x и y следует утверждение $r(x, y) = 0$, но обратное утверждение неверно [73,154]. Например, при $y = |x|$ величины x и y сильно зависимы, но не коррелированы [73].

Таким образом, при проведении корреляционного анализа следует тщательно обдумывать физический смысл обнаруженных зависимостей, подходить с осторожностью ко всем неожиданным результатам и пытаться соотнести их с другими (надежными) результатами. Существуют правила, позволяющие лучше определить истинную корреляцию за счет исключения других возможных взаимозависимостей [127]. Опознание причинной корреляции, согласно этим правилам, осуществляется путем последовательного исключения формальной и совместной корреляций. Чаще всего причинная связь так и остается недоказанной в силу того, что не может быть отклонена возможность совместной корреляции.

ГЛАВА 3. РЕГРЕССИОННЫЙ АНАЛИЗ

3.1. РЕГРЕССИОННЫЙ АНАЛИЗ ПРИ СУДЕБНО-МЕДИЦИНСКОЙ ИДЕНТИФИКАЦИИ ЛИЧНОСТИ

После того как с помощью корреляционного анализа выявлено наличие статистически значимых связей между идентифицируемым параметром и идентифицирующими показателями (или их преобразованиями) и оценена степень их тесноты, переходят к математическому описанию конкретного вида зависимостей с использованием регрессионного анализа. Для точного прогнозирования идентифицируемого параметра необходимо знать вид функции, связывающей идентифицируемый параметр и идентифицирующие показатели. Поскольку такая информация обычно отсутствует, на практике ограничиваются поиском подходящих аппроксимаций искомой функции, основанных на исходных статистических данных.

С позиции статистического анализа идентифицирующие показатели рассматриваются как независимые переменные (аргументы) функции регрессии, а идентифицируемый параметр – как зависимая переменная или результативный показатель (оценка регрессии). Поскольку истинная функция регрессии не известна, в регрессионном анализе выделяют такие понятия оценок регрессии, как истинная, модельная и точечная [26].

В качестве истинной рассматривается оценка, полученная с помощью истинного уравнения регрессии, основанного на точном знании условного закона распределения результативного показателя. Аналитическим выражением истинного уравнения регрессии является функция $f(x) = M(y/x)$, где x – аргумент, y – истинная оценка результативного показателя.

Модельной называется оценка результативного показателя, полученная с помощью функции регрессии, класс которой выбран исследователем в качестве аппроксимации неизвестной истинной функции $f(x)$. Обобщенная форма модельных функций представлена выражением $\tilde{y} = \beta_0 + \sum_{i=1}^{i=k} \beta_i x_i$, где $\tilde{y}$ - модельная оценка результативного показателя; β – совокупность из $(k + 1)$ неизвестных истинных регрессионных коэффициентов модельной функции.

Точечной является оценка, полученная с помощью модельного уравнения регрессии, созданного на основе изучения выборочной

совокупности данных. В обобщенном виде такие регрессионные уравнения принято обозначать как $\hat{y} = b_0 + \sum_{i=1}^{i=k} b_i x_i$, где $\hat{y}$ - точечная оценка результативного показателя; b - совокупность из (k + 1) выборочных оценок параметров теоретической модельной функции регрессии.

Оценка $\hat{y}$ всегда сходится по вероятности к модельной оценке при неограниченном увеличении объема выборки ($n \to \infty$). В отличие от нее, модельная оценка $\tilde{y}$ сходится к истинной оценке y только при условии, что класс аппроксимирующей функции выбран правильно. В этом случае неточность в прогнозировании $\hat{y}$ объяснялась бы только ограниченностью исследованной выборки и могла бы быть сделана сколько угодно малой при $n \to \infty$. При неправильном выборе класса аппроксимирующей регрессионной функции точечные оценки $\hat{y}$ и параметры созданного регрессионного уравнения не будут обладать свойством состоятельности, т.е. при увеличении объема наблюдений оценка $\hat{y}$ не будет сходиться к истинной функции регрессии $f(x)$.

С целью наилучшего восстановления по исходным статистическим данным неизвестной функции регрессии $f(x) = M(y/x)$ наиболее часто используют следующие критерии адекватности (функции потерь) [26].

1. Метод наименьших квадратов (или его модификацию – метод взвешенных наименьших квадратов), согласно которому минимизируется квадрат отклонения наблюдаемых значений результативного показателя $y_i (i = 1,2,\ldots,n)$ от модельных значений $\tilde{y}_i = f(x_i, \beta_i)$:

$$\sum_{i=1}^{n} (y_i - f(x_i, \beta_i))^2 \to \min .$$

2. Метод наименьших модулей, согласно которому минимизируется сумма абсолютных отклонений наблюдаемых значений результативного показателя от модульных значений $\tilde{y}_i = f(x_i, \beta_i)$:

$$\sum_{i=1}^{n} |y_i - f(x_i, \beta_i)| \to \min .$$

3. Метод минимакса сводится к минимизации максимума модуля отклонения наблюдаемого значения результативного показателя y_i от модельного значения $f(x_i, \beta_i)$:

$$\max|y_i - f(x_i, \beta_i)| \to \min.$$

Получаемые при использовании указанных методов регрессии называются соответственно среднеквадратическими, среднеабсолютными (медианными) и минимаксными.

Таким образом, регрессионным анализом является метод статистического анализа зависимости случайной величины *y* от факторных переменных $x_i (i = 1,2,\ldots,k)$. Основной задачей регрессионного анализа является определение аналитического выражения связи, в котором изменение результативного параметра обусловлено влиянием одного или нескольких независимых факторов. При этом обычно предполагается, что *y* имеет нормальный закон распределения с условным математическим ожиданием $\tilde{y}$, являющимся функцией от аргументов $x_i (i = 1,2,\ldots,k)$ и постоянной, не зависящей от аргументов дисперсией σ^2. Однако предположение нормальности *y* необходимо лишь для проверки значимости уравнения регрессии и его параметров β_i, а также интервального оценивания β_i. Для получения точечных оценок $\beta_i (i = 0,1,2,\ldots,k)$ этого условия не требуется. Оценки неизвестных параметров уравнения регрессии находят обычно методом наименьших квадратов.

Основой видовой структуры регрессионных зависимостей являются такие параметры, как количество факторных показателей, направление регрессии и ее форма.

В зависимости от количества факторных показателей регрессия может быть однофакторной (парной) и многофакторной (множественной). По направлению связи различают:

- прямую регрессию (положительную), при которой с увеличением или уменьшением независимой переменной значения зависимой переменной также соответственно увеличиваются или уменьшаются;

- обратную (отрицательную) регрессию, при которой с увеличением или уменьшением независимой переменной зависимая переменная соответственно уменьшается или увеличивается.

По форме регрессия бывает линейной и нелинейной.

К настоящему времени наиболее полно разработана теория линейных регрессионных уравнений.

3.2. ОДНОФАКТОРНАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

В общем смысле линейной функцией с одной переменной называется функция вида $y = a_0 + a_1 x$, где a_0 и a_1 – некоторые константы. Областью определения линейной функции является множество всех действительных чисел (вся числовая прямая). Область значений линейной функции при $a_1 \neq 0$ также представлена числовой прямой. При $a_1 = 0$ область значений состоит из одной точки a_0.

Линейная функция дифференцируема на всей числовой прямой. Так как $f'(x) = a_1$, то при $a_1 > 0$ функция f возрастает на всей числовой прямой, при $a_1 < 0$ функция f на этом же промежутке убывает, а при $a_1 = 0$ функция постоянная. При $a_1 \neq 0$ линейная функция не имеет экстремумов. Коэффициент a_0 равен ординате точки пересечения прямой с осью ординат, коэффициент a_1 – тангенсу угла между прямой и осью абсцисс.

Применительно к судебно-медицинской антропологии независимая переменная x в составе однофакторной линейной функции представлена значениями идентифицирующего показателя, а зависимая переменная y – прогнозируемыми значениями идентифицируемого параметра. Области определения и значений линейной модели идентификации представляют собой множество неотрицательных действительных чисел. Как отмечалось, на практике неизвестными являются не только регрессионные коэффициенты, но и сам класс истинного аналитического выражения прогнозной зависимости идентифицируемого параметра от идентифицирующего показателя. В этой связи однофакторная линейная регрессия является всего лишь одной из множества альтернативных моделей, аппроксимирующих класс неизвестной истинной функции регрессии.

Моделью однофакторной линейной регрессии является выражение $\tilde{y} = \beta_0 + \beta_1 x$, где β_0 и β_1 – неизвестные параметры генеральной совокупности, которые необходимо оценить по результатам выборочных наблюдений.

Если для оценки параметров β_0 и β_1 из двумерной генеральной совокупности (x,y) взята выборка объемом n, где (x_i, y_i) результат i-го наблюдения $(i = 1,2,\ldots,n)$, то модель регрессионного анализа имеет вид $\tilde{y}_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, где ε_i – независимые нормально рас-

пределенные случайные величины с нулевым математическим ожиданием и дисперсией σ_ε^2.

Согласно методу наименьших квадратов для парной линейной регрессии задача заключается в отыскании неизвестных параметров b_0 и b_1, минимизирующих функцию Q:

$$Q = \sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n}\varepsilon_i^2 . \qquad (11)$$

Для этого продифференцируем функцию (11) отдельно по β_0 и β_1:

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = -2\sum\limits_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) \\ \dfrac{\partial Q}{\partial \beta_1} = -2\sum\limits_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i \end{cases} . \qquad (12)$$

Приравняв частные производные нулю и подставив в (12) вместо β_0 и β_1 их оценки b_0 и b_1, получим систему нормальных уравнений:

$$\begin{cases} b_0 n + b_1 \sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i \\ b_0 \sum\limits_{i=1}^{n} x_i + b_1 \sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} x_i y_i \end{cases} . \qquad (13)$$

Решая систему (13) относительно b_0 и b_1, получим

$$b_1 = \frac{\sum\limits_{i=1}^{n} x_i y_i - \frac{1}{n}\sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{\sum\limits_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum\limits_{i=1}^{n} x_i)^2} = \frac{\sum\limits_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum\limits_{i=1}^{n} x_i^2 - n\overline{x}^2}; \qquad (14)$$

$$b_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - b_1 \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{y} - b_1 \overline{x} . \qquad (15)$$

Величина b_1 является угловым коэффициентом линии регрессии и называется коэффициентом регрессии (наклона). Поскольку $f'(x) = b_1$, то зависимая переменная y при $b_1 > 0$ всегда возрастает в среднем на b_1 единиц, если независимая переменна x возрастает на единицу (или убывает в среднем на b_1 при $b_1 < 0$). Величина b_0 называется коэффициентом сдвига. В отличие от b_0 коэффициент регрессии более информативен и играет значительную роль в прикладном анализе.

Полученное на основе анализа выборочных данных уравнение однофакторной линейной регрессии $\hat{y} = b_0 + b_1 x_i + \varepsilon_i$, где $\varepsilon_i = y_i - \hat{y}_i$, можно использовать для прогнозирования значений y в зависимости от значений x. Ввиду наличия в составе регрессионного уравнения компонента ε_i, обозначающего величину отклонения точечной оценки $\hat{y}_i$ от истинного значения y_i, то прогнозирование зависимой переменной не будет точным. Поэтому для оценки надежности регрессионного уравнения используется такая характеристика, как остаточное стандартное отклонение (стандартная ошибка регрессии), вычисляемое по формуле:

$$s_\varepsilon = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}. \qquad (16)$$

Для определения интервальных оценок результативного показателя y с помощью регрессионной модели $\tilde{y}_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ необходимо сделать следующие допущения:

- каждому значению x соответствует распределение наблюдаемых значений y, которое является нормальным;
- дисперсия y остается постоянной при изменении x;
- ошибка ε является случайной величиной со средней, равной нулю, причем последовательные значения ε независимы друг от друга и имеют одинаковую дисперсию, т.е. $M\varepsilon_i = 0$; $D\varepsilon_i = \sigma^2$ для всех $i = 1,2,\ldots,n$ и $M\varepsilon_i\varepsilon_j = 0$ при $i \neq j$.

При линейной корреляционной связи между переменными x и y дисперсия распределения y вокруг прямой регрессии совпадает с дисперсией случайных ошибок ε.

Поскольку статистики b_0 и b_1 являются выборочными оценками неизвестных коэффициентов β_0 и β_1, то общая дисперсия прогноза, обеспечиваемого уравнением $\hat{y} = b_0 + b_1 x_i$, слагается из трех дисперсий: остаточной дисперсии и дисперсий статистик b_0 и b_1.

В теории доказано, что дисперсии статистик b_0 и b_1 определяются как

$$s_{b_0} = s_\varepsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{s_x^2(n-1)}} \text{ и } s_{b_1} = \frac{s_\varepsilon}{s_x\sqrt{n-1}} \text{ [73]}. \qquad (17)$$

Если регрессионное уравнение $\hat{y} = b_0 + b_1 x_i$ представить в виде $\hat{y} = \bar{y} + b_1(x_i - \bar{x})$, то, учитывая сумму компонентов (16) и (17), общая дисперсия прогноза $\hat{y}(x_i)$ будет иметь вид:

$$s_f^2 = s_\varepsilon^2 + \frac{s_\varepsilon^2}{n} + \frac{s_\varepsilon^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}(x_i - \bar{x})^2. \qquad (18)$$

Полученное стандартное отклонение s_f^2 называется стандартной ошибкой прогнозирования:

$$s_f = s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}. \qquad (19)$$

Выразив сумму в знаменателе (19) через стандартное отклонение x, получим

$$s_f = s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_X^2}} \qquad (20)$$

Ввиду (20) и допущений о нормальности распределения y вокруг прямой регрессии можно определить доверительный интервал для y_i при любом заданном x_i:

$$y_i \in \hat{y}_i \pm t_{\alpha;n-2} s_f = \hat{y}_i \pm t_{\alpha;n-2} s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_X^2}}. \qquad (21)$$

Из (21) следует, что по мере удаления x_i от среднего значения $\bar{x}$ ширина доверительного интервала для y_i снижается. Наименьшую величину доверительный интервал для y_i имеет при $x_i = \bar{x}$.

С помощью t-распределения можно также с доверительной вероятностью 1-α определить интервальные оценки параметров β_0 и β_1:

$$\beta_0 \in b_0 \pm t_{\alpha;n-2} s_{b_0} \text{ и } \beta_1 \in b_1 + t_{\alpha;n-2} s_{b_1}.$$

Установление значимости однофакторного линейного уравнения регрессии сводится к проверке при заданном α нулевой гипотезы о значимости коэффициента регрессии β_1, т.е. гипотезы $H_0 : \beta_1 = 0$ при альтернативной гипотезе: $H_0 : \beta_1 \neq 0$.

Для этого используется значение статистики

$$t_{b_1} = \frac{b_1}{s_{b_1}}, \qquad (22)$$

которое сравнивают с критическим значением t при уровне значимости α и $\nu = n-2$ степенях свободы. Нулевая гипотеза $H_0 : \beta_1 = 0$ отвергается при $\left|t_{b_1}\right| > t_{\alpha;n-2}$.В противном случае при $\left|t_{b_1}\right| < t_{\alpha;n-2}$ нулевая гипотеза принимается, и уравнение регрессии считают незначимым. В случае отсутствия значимости b_1 использование выборочного регрессионного уравнения для прогнозирования и анализа не имеет смысла, так как оно не отражает реальной связи между исследуемыми переменными.

Доказано, что результаты проверки значимости однофакторной регрессии и парного коэффициента линейной корреляции с помощью формул (22) и (5) тождественны [16].

Рассмотрим показатель, определяющий в какой степени изменением независимой переменной объясняется вариация зависимой переменной. Определение интенсивности связи основано на разложении на два слагаемых суммы квадратов отклонений зависимой переменной y_i от среднего $\bar{y}$:

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \qquad (23)$$

Величина r^2 называется коэффициентом детерминации. Данный коэффициент характеризует удельный вес общей дисперсии, который объясняется уравнением регрессии (или изменением факторного показателя x). Можно показать, что значение коэффициента детерминации равно квадрату коэффициента корреляции, поэтому $r^2 \in [0;1]$.

Из выражения (16) следует, что остаточная сумма квадратов связана с остаточной дисперсией соотношением

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = (n-2)s_\varepsilon^2.$$

В свою очередь, общая сумма квадратов связана с дисперсией s_y^2 соотношением

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = (n-1)s_y^2.$$

Отсюда $r^2 = 1 - \frac{n-2}{n-1}\frac{s_\varepsilon^2}{s_y^2}$.

В качестве примера однофакторной линейной регрессии продолжим построение регрессионной модели идентификации гестационного возраста плодов и новорожденных по диаметру лимфоидных узелков селезенки. В разделе 2.2 было доказано наличие и оценена степень тесноты связи между гестационным возрастом и указанным гистометрическим показателем фетальной селезенки. Однако для идентификации гестационного возраста необходимо получить аналитическое выражение зависимости данного идентифицируемого параметра от идентифицирующего показателя – диаметра лимфоидных узелков. Необходимые для этого результаты промежуточных вычислений приведены в таблице 8.

Таблица 8

Промежуточные результаты построения однофакторной линейной регрессионной модели идентификации гестационного возраста

$\bar{x}$	s_x^2	$\sum_{i=1}^{i=99} x_i$	$\sum_{i=1}^{i=99} (x_i - \bar{x})^2$	$\sum_{i=1}^{i=99} x_i^2$	$(\sum_{i=1}^{i=99} x_i)^2$
185,4	1765,0	18350	172973,2	3574365	336737830
$\bar{y}$	s_y^2	$\sum_{i=1}^{i=99} y_i$	$\sum_{i=1}^{i=99} (y_i - \bar{y})^2$	$\sum_{i=1}^{i=99} (y_i - \hat{y}_i)^2$	$\sum_{i=1}^{i=99} x_i y_i$
29,5	31,9	2916	3126,5	1874,4	555220

Исходя из (14), (15), (16) и (20) получаем оценки b_1, b_0, s_ε и s_f:

$$b_1 = \frac{555220 - 99 \cdot 185{,}4 \cdot 29{,}5}{3574365 - 99 \cdot 185{,}4^2} = 0{,}085;^5$$

$$b_0 = 29{,}5 - 0{,}085 \cdot 185{,}4 = 13{,}684;$$

$$s_\varepsilon = \sqrt{1874{,}4/(99-2)} = 4{,}396;$$

$$s_f = 4{,}396\sqrt{1{,}010 + (x_i - 185{,}4)^2 / 172973{,}2}.$$

Отсюда с учетом (21) получаем полную регрессионную модель идентификации гестационного возраста по диаметру лимфоидных узелков селезенки:

$$y_i = 13{,}684 + 0{,}085x_i \pm 4{,}396 t_{\alpha;97} \sqrt{1{,}010 + (x_i - 185{,}4)^2 / 172973{,}2},$$

где x_i – средний диаметр лимфоидных узелков с учетом усадки, мкм; y_i – гестационный возраст, недель (рис. 18).

[5] Здесь и в других примерах некоторые различия в итоговых результатах связаны с округлениями на стадии промежуточных вычислений.

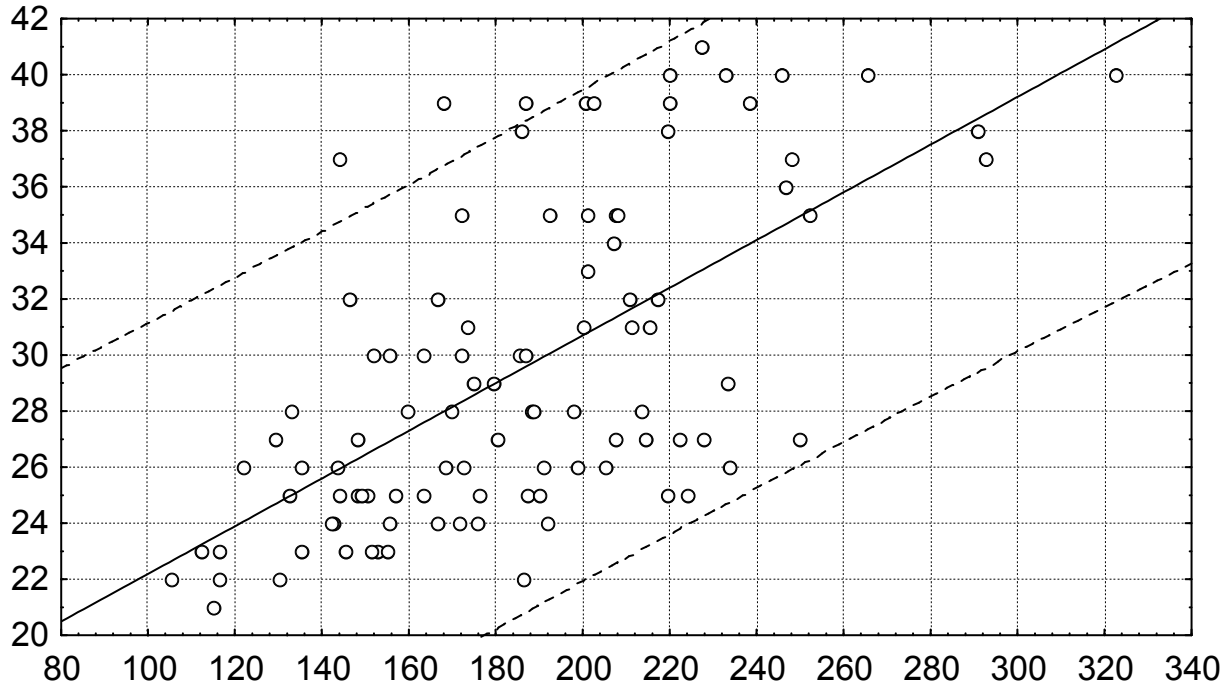

Рис. 18. Геометрическая интерпретация однофакторной линейной регрессионной модели идентификации гестационного возраста плодов и новорожденных по диаметру лимфоидных узелков селезенки. По оси абсцисс – средний диаметр лимфоидных узелков селезенки с учетом усадки, мкм; по оси ординат – гестационный возраст, недель. Непрерывной линией показана прямая регрессии, пунктирными - 95% доверительная область для значений гестационного возраста.

С помощью (17) вычислим стандартные ошибки коэффициентов регрессии:

$$s_{b_0} = \sqrt{\frac{4{,}396^2 \cdot 3574365}{99 \cdot 172973{,}2}} = 2{,}0084;$$

$$s_{b_1} = \sqrt{4{,}396^2 / 172973{,}2} = 0{,}0106.$$

Отсюда с помощью *t*-распределения определяем 95% интервальные оценки параметров β_0 и β_1:

$$9{,}698 < \beta_0 < 17{,}670;$$

$$0{,}064 < \beta_1 < 0{,}106.$$

Поскольку 95% доверительный интервал для β_1 не содержит нуль, можно со статистической надежностью, равной не менее 0,95, сделать вывод о значимости данного регрессионного уравнения. Точное значение вероятности ошибочного принятия альтернативной гипотезы о значимости регрессии в данном случае равняется $p = 2{,}106 \cdot 10^{-12}$.

3.3. МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

С помощью уравнений однофакторной регрессии можно прогнозировать значения идентифицируемого параметра всего лишь по одному идентифицирующему показателю. Полученные подобным образом прогнозные оценки редко бывают точными. Поэтому в судебно-медицинских антропологических исследованиях повышения точности обычно добиваются путем включения в диагностическую регрессионную модель нескольких признаков, обладающих наибольшей идентификационной значимостью. Подобные задачи в судебно-медицинской антропологии и других приложениях решаются методами множественной регрессии.

Важнейшей предпосылкой применения множественной регрессии является линейность связь между переменными. На практике это предположение, в сущности, никогда не может быть достоверно подтверждено. Однако процедуры множественного регрессионного анализа в незначительной степени подвержены воздействию малых отклонений от этого предположения [13].

В общем случае модель множественной линейной регрессии можно представить в виде уравнения

$$\tilde{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k .$$

Если для оценки неизвестных параметров $\beta_i (i = 1,2,\ldots,k)$ уравнения множественной регрессии взята случайная выборка из (k+1)-мерной нормально распределенной генеральной совокупности, где $y_i, x_{i1}, \ldots x_{ik}$ - результат i-го наблюдения, где $i = 1,2,\ldots,n$, то модель множественной линейной регрессии имеет вид:

$$\tilde{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i ,$$

где $\varepsilon_i = y_i - \tilde{y}_i$ - некоррелированные случайные ошибки с нулевым средним.

В матричной форме модель множественной линейной регрессии имеет вид:

$$Y = X\beta + \varepsilon ,$$

где $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ - вектор-столбец наблюдений размерности n;

$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$ - матрица размерности $n(k+1)$ известных величин ($n > k$);

$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}$ - вектор столбец размерности (k+1) неизвестных параметров, подлежащих оцениванию;

$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ - вектор-столбец (размерности n) случайных ошибок [26].

Причем ковариационная матрица $\sum(\varepsilon) = M\varepsilon\varepsilon^T$, где

$$M(\varepsilon\varepsilon^T) = \left[\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} (\varepsilon_1 \varepsilon_2 \ldots \varepsilon_n) \right] = \begin{pmatrix} \varepsilon_1^2 & \varepsilon_1\varepsilon_2 & \cdots & \varepsilon_1\varepsilon_n \\ \varepsilon_2\varepsilon_1 & \varepsilon_2^2 & \cdots & \varepsilon_2\varepsilon_n \\ \vdots & \vdots & \vdots & \vdots \\ \varepsilon_n\varepsilon_1 & \varepsilon_n\varepsilon_2 & \cdots & \varepsilon_n^2 \end{pmatrix}.$$

Для дальнейшего анализа повторим сделанные ранее допущения относительно ошибок: ε_i является случайной величиной со средней, равной нулю, причем последовательные значения ε_i независимы друг от друга и имеют одинаковую дисперсию, т.е. для всех $i = 1,2,\ldots,n$: $M\varepsilon_i = 0$; $M\varepsilon_i^2 = \sigma^2$ и $M\varepsilon_i\varepsilon_j = 0$ при $i \neq j$.

Тогда

$$\sum(\varepsilon) = M(\varepsilon\varepsilon^T) = \sigma^2 E_n,$$

где

$E_n = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$ - единичная матрица размерности $(n \times n)$.

Оценки неизвестных параметров β определяются из условия минимизации скалярной суммы квадратов Q по компонентам вектора β:

$$Q = (Y - X\beta)^T (Y - X\beta) = \varepsilon^T \varepsilon = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} x_{ij}\beta_j)^2 .$$

Условием обращения Q в минимум является система уравнений $\frac{\partial Q}{\partial \beta_j} = 0$, где $j = 1,2,\ldots,k$. Вычисляя частные производные, получим

$$-2X^T (Y - X\beta) = 0,$$

где X^T – транспонированная матрица X.

Заменяя вектор β на оценку метода наименьших квадратов, в конечном счете, можно получить

$$b = \widehat{\beta} = (X^T X)^{-1} X^T Y. \qquad (24)$$

В случае линейной модели b является несмещенной оценкой с минимальной дисперсией вектора β [26].

Ковариационная матрица вектора b равна

$$\sum(b) = \sigma^2 (X^T X)^{-1},$$

в которой элементы главной диагонали представлены дисперсиями вектора оценок b. Вне главных оценок ковариационной матрицы расположены значения коэффициентов ковариации:

$$\mathrm{cov}(b_i b_j) = M(b_i - \beta_i)(b_j - \beta_j),$$

где $i, j = 0,1,2,\ldots,k$.

Таким образом, оценка b_i коэффициента множественной линейной регрессии $\beta_i (i = 0,1,2,\ldots,k)$ имеет нормальный закон распределения с математическим ожиданием β_i и дисперсией $\sigma^2 \left|(X^T X)^{-1}\right|_{ii}$, где $\left|(X^T X)^{-1}\right|_{ii}$ - диагональный элемент обратной матрицы $(X^T X)^{-1}$, соответствующий i-й строке и i-му столбцу.

Поскольку несмещенная оценка остаточной дисперсии σ^2 равна

$$s_\varepsilon^2 = \frac{1}{n-k-1}(Y - Xb)^T (Y - Xb) = \frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{n-k-1},$$

то дисперсия оценки b_i характеризуется выражением

$$s_{b_i}^2 = s_\varepsilon^2 \left|(X^T X)^{-1}\right|_{ii}. \qquad (25)$$

Как уже отмечалось в разделе 3.2, общая вариация зависимой переменной является суммой объяснимой и необъяснимой вариаций:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2.$$

Тогда, с учетом (23), проверку статистической значимости уравнения множественной линейной регрессии можно осуществить через коэффициент множественной детерминации r^2:

$$F_{k;n-k-1}=\frac{r^2/k}{(1-r^2)/(n-k-1)}. \qquad (26)$$

Числитель F-отношения представляет собой долю объясненной вариации результативного показателя, деленную на число степеней свободы $v=k$. Знаменатель F-отношения характеризует долю необъясненной вариации, деленную на число степеней свободы $v=n-k-1$.

Для проверки значимости уравнения множественной регрессии служит также F-статистика:

$$F_{k;n-k-1}=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2/k}{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/(n-k-1)}. \qquad (27)$$

Результаты, получаемые с использованием (26) и (27) тождественны. Если полученные F-статистики незначимы, то принимается нулевая гипотеза $H_0:\beta=0$, т.е. все коэффициенты регрессии множественной регрессионной модели равны нулю, и на этом регрессионный анализ заканчивается. При статистической значимости регрессионного уравнения необходима проверка значимости отдельных коэффициентов регрессии и построение интервальных оценок для значимых из них.

С практической точки зрения большой интерес представляют данные о достаточно выраженной робастности (устойчивости) описанных F-тестов по отношению к отклонениям от предположения нормальности остатков [13].

Значимость коэффициентов регрессии можно проверить с помощью t-критерия:

$$t_i=\frac{b_i}{s_{b_i}}=\frac{b_i}{s\left[(X^TX)^{-1}\right]_{ii}^{1/2}},$$

которая при справедливости нулевой гипотезы $H_0 : \beta_i = 0$ имеет *t*-распределение с $v = n - k - 1$ числом степеней свободы.

Каждая из полученных *t*-статистик определяет, оказывает ли значимое влияние на результативный показатель данная независимая переменная, если все остальные факторные переменные остаются при этом неизменными. Иными словами, серия проведенных *t*-тестов помогает выяснить, какие именно из комплекса анализируемых факторных переменных действительно полезны в целях прогнозирования результативного показателя. В редких случаях *t*-тесты для отдельных коэффициентов регрессии могут быть значимыми даже тогда, когда *F*-тест значимым не является [74]. В этом случае результат *F*-теста считается более важным, свидетельствуя об отсутствии значимости всех коэффициентов регрессии.

В теории доказано, что ввиду свойства наименьших квадратов для множественной регрессии величина коэффициента множественной детерминации не уменьшается (а может только увеличиться) при добавлении в состав регрессионного уравнения дополнительной независимой переменной. Важно, что увеличение r^2 может произойти даже в случае, когда добавленная независимая переменная не связана с результативным показателем и характеризуется незначимым результатом *t*-теста. Высокий коэффициент детерминации также может быть получен при всех статистически незначимых коэффициентах множественной регрессии.

Для учета автоматического роста коэффициента r^2 вследствие увеличения количества независимых переменных рекомендуется использование скорректированного коэффициента множественной детерминации $\bar{r}^2$, вычисляемого по формуле

$$\bar{r}^2 = 1 - (1 - r^2)\frac{n-1}{n-k-1} = r^2 - \frac{k}{n-k-1}(1 - r^2),$$

где k – число независимых переменных.

Доказано, что введение новой независимой переменной в состав регрессионного уравнения приведет к увеличению $\bar{r}^2$ только в том случае, если *t*-статистика коэффициента регрессии, соответствующего этой переменной по абсолютной величине будет больше 1 [78].

С помощью *t*-распределения с надежностью 1-α возможно построение интервальных оценок для коэффициентов β множественной регрессионной модели:

$$\beta_i \in b_i \pm t_{\alpha;n-k-1} s_\varepsilon \left[(X^T X)^{-1} \right]_{ii}^{1/2}. \quad (28)$$

Определим интервальную оценку $\tilde{y}$ в точке, задаваемой вектором X^0 начальных условий, размерности (k+1):

$$x^0 = (1, x_1^0, x_2^0, \ldots, x_k^0)^T.$$

Тогда несмещенная оценка дисперсии значений признака $\tilde{y}$ при заданном векторе X^0 выражается формулой

$$s_{\tilde{y}}^2 = (X^0)^T \sum(b) X^0 = s_\varepsilon^2 (X^0)^T (X^T X)^{-1} X^0,$$

где $\sum(b)$ - ковариационная матрица вектора оценок b.

Тогда интервальная оценка для $\tilde{y}$ в точке, определяемой вектором X^0, с надежностью 1-α вычисляется как

$$\tilde{y} \in (X^0)^T b \pm t_{\alpha;n-k-1} s_\varepsilon \sqrt{(X_0)^T (X^T X)^{-1} X^0}. \quad (29)$$

Стандартная ошибка прогноза множественной регрессионной модели включает в себя дисперсию случайной ошибки s_ε^2 и дисперсию $s_{\tilde{y}}^2$, связанную с параметрами β модели:

$$s_f = \sqrt{s_\varepsilon^2 + s_{\tilde{y}}^2} = s_\varepsilon \sqrt{1 + (X^0)^T (X^T X)^{-1} X^0}.$$

Отсюда получаем доверительную оценку для интервала предсказания $\hat{y}$ с надежностью 1-α:

$$\hat{y} \in (X^0)^T b \pm t_{\alpha;n-k-1} s_\varepsilon \sqrt{1 + (X_0)^T (X^T X)^{-1} X^0}. \quad (30)$$

Продемонстрируем процедуру множественной линейной регрессии на примере построения регрессионной модели идентификации гестационного возраста плодов и новорожденных по двум гистометрическим показателям селезенки: диаметру лимфоидных узелков и толщине стенок центральных артерий. Необходимость введения дополнительной факторной переменной (толщина стенок центральных артерий) объясняется небольшой точностью однофакторного регрессионного уравнения, основанного на использовании одного только показателя диаметра лимфоидных узелков ($s_\varepsilon = 4{,}396$ недели). В разделе 2.3 было доказано наличие и оценена степень статистической зависимости гестационного возраста от диаметра лимфоидных узелков и толщины стенок центральных артерий фетальной селезенки. Однако для практического использования в целях идентификации гестационного возраста необходимо получить аналитическое выражение указанной статистической зависимости.

Введем условные обозначения: y – гестационный возраст, недель; x_1 – диаметр лимфоидных узелков селезенки, мкм; x_2 - толщина стенок центральных артерий селезенки, мкм. Тогда выборку объемом 99 наблюдений из трехмерной генеральной совокупности будет представлена матрицами Y и X:

$$Y = \begin{pmatrix} 22 \\ 24 \\ \vdots \\ 40 \end{pmatrix}$$ - вектор-столбец значений y_i размерности $n = 99$;

$$X = \begin{pmatrix} 1 & 186,2 & 7,83 \\ 1 & 142,6 & 8,14 \\ \vdots & \vdots & \vdots \\ 1 & 245,6 & 9,74 \end{pmatrix}$$ - матрица размерности $n = 297$ известных значений гистометрических показателей x_1 и x_2 селезенки;

Для определения вектора оценок b найдем предварительно симметричную матрицу X^TX, которая имеет вид:

$$X^T X = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} \\ \sum_{i=1}^{n} x_{i2} & \sum_{i=1}^{n} x_{i1}x_{i2} & \sum_{i=1}^{n} x_{i2}^2 \end{pmatrix}$$

и равна данном случае

$$X^T X = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 186,2 & 142,6 & \cdots & 245,6 \\ 7,83 & 8,14 & \cdots & 9,74 \end{pmatrix} \begin{pmatrix} 1 & 186,2 & 7,83 \\ 1 & 142,6 & 8,14 \\ \vdots & \vdots & \vdots \\ 1 & 245,6 & 9,74 \end{pmatrix} =$$

$$= \begin{pmatrix} 99 & 18350,4 & 866,01 \\ 18350,4 & 3574365,4 & 164383,75 \\ 866,01 & 164383,75 & 8119,725 \end{pmatrix}.$$

Вектор X^TY имеет вид:

$$X^T Y = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1} y_i \\ \sum_{i=1}^{n} x_{i2} y_i \end{pmatrix}$$

и в данном случае равен:

$$X^T Y = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 186{,}2 & 142{,}6 & \cdots & 245{,}6 \\ 7{,}83 & 8{,}14 & \cdots & 9{,}74 \end{pmatrix} \begin{pmatrix} 22 \\ 24 \\ \vdots \\ 40 \end{pmatrix} = \begin{pmatrix} 2916 \\ 555219{,}9 \\ 26273{,}85 \end{pmatrix}.$$

Получим обратную матрицу

$$(X^T X)^{-1} = \begin{pmatrix} 0{,}25512686 & -0{,}00084690 & -0{,}01006519 \\ -0{,}00084690 & 0{,}00000687 & -0{,}00004875 \\ -0{,}01006519 & -0{,}00004875 & 0{,}00218351 \end{pmatrix}.$$

Подставив найденные вектор $X^T Y$ и матрицу $(X^T X)^{-1}$ в (24), вычислим вектор оценок

$$b = (X^T X)^{-1} X^T Y = \begin{pmatrix} 0{,}25512686 & -0{,}00084690 & -0{,}01006519 \\ -0{,}00084690 & 0{,}00000687 & -0{,}00004875 \\ -0{,}01006519 & -0{,}00004875 & 0{,}00218351 \end{pmatrix} \begin{pmatrix} 2916 \\ 555219{,}9 \\ 26273{,}85 \end{pmatrix} = \begin{pmatrix} 9{,}2828 \\ 0{,}0638 \\ 0{,}9548 \end{pmatrix}.$$

Таким образом,

$$b = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 9{,}2828 \\ 0{,}0638 \\ 0{,}9548 \end{pmatrix}$$

и оценка уравнения регрессии имеет вид:

$$\hat{y} = 9{,}2828 + 0{,}0638 x_1 + 0{,}9548 x_2 .$$

Проверим значимость вычисленного в разделе 2.3 множественного коэффициента детерминации анализируемой регрессии:

$$F_{\nu_1=2;\nu_2=96} = \frac{0{,}731^2 / 2}{(1 - 0{,}731^2)/96} = 55{,}009, \; p = 1{,}207 \cdot 10^{-16}.$$

Найдем величину остаточной дисперсии:

$$s_\varepsilon^2 = \frac{1}{n-k-1} (Y - Xb)^T (Y - Xb) = 15{,}176.$$

Перед проверкой значимости отдельных коэффициентов регрессии определим оценку ковариационной матрицы $S(b)$ вектора b. Для этого нужно умножить элементы обратной матрицы $(X^T X)^{-1}$ на $s_\varepsilon^2 = 15{,}176$:

$$S(b) = s_\varepsilon^2 (X^T X)^{-1} = \begin{pmatrix} 3{,}8718 & -0{,}0129 & -0{,}1528 \\ -0{,}0129 & 1{,}043 \cdot 10^{-4} & -0{,}0007 \\ -0{,}1528 & -0{,}0007 & 0{,}0331 \end{pmatrix}.$$

Из статистического смысла полученной ковариационной матрицы следует, что оценки дисперсии коэффициентов регрессионного уравнения равны:

$$s_{b_0}^2 = 3{,}8718;\ s_{b_1}^2 = 0{,}0001043;\ s_{b_2}^2 = 0{,}0331.$$

Проверим значимость коэффициента β_1:

$$t_{b_1} = \frac{b_1}{s_{b_1}} = \frac{0{,}0638}{0{,}0102} = 6{,}245;\ p = 1{,}153 \cdot 10^{-8}.$$

Выполним аналогичную процедуру относительно β_2:

$$t_{b_2} = \frac{b_2}{s_{b_2}} = \frac{0{,}9548}{0{,}1820} = 5{,}245;\ p = 9{,}298 \cdot 10^{-7}.$$

Поскольку для обоих коэффициентов $p < 0{,}05$, то нулевые гипотезы H_0: $\beta_1 = 0$ и H_0: $\beta_2 = 0$ отвергаются, т.е. оба гистометрических параметра селезенки вносят значимый вклад в прогнозирование гестационного возраста.

Найдем с надежностью $1 - \alpha = 0{,}95$ интервальные оценки для β_1 и β_2. Согласно (25) и (28) и учитывая, что $t_{0{,}05;96} = 1{,}985$, имеем:

$$\beta_1 \in 0{,}0638 \pm 0{,}0203;$$
$$\beta_2 \in 0{,}9548 \pm 0{,}3613.$$

Определим с надежностью $1 - \alpha$ интервальную оценку для $\tilde{y}$ в точке, определяемой вектором X^0. С учетом полученных данных и согласно (29) находим:

$$\tilde{y} \in (X^0)^T b \pm 3{,}896 t_{\alpha;n-k-1} \sqrt{(X_0)^T (X^T X)^{-1} X^0},$$

где $X^0 = \begin{pmatrix} 1 & x_1^0 & x_2^0 \end{pmatrix}$ - вектор начальных условий; $(X^0)^T$ - транспонированный вектор

$$(X^0)^T = \begin{pmatrix} 1 \\ x_1^0 \\ x_2^0 \end{pmatrix};$$

$$b = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 9{,}2828 \\ 0{,}0638 \\ 0{,}9548 \end{pmatrix};$$

$$(X^T X)^{-1} = \begin{pmatrix} 0{,}25512686 & -0{,}00084690 & -0{,}01006519 \\ -0{,}00084690 & 0{,}00000687 & -0{,}00004875 \\ -0{,}01006519 & -0{,}00004875 & 0{,}00218351 \end{pmatrix}.$$

Согласно (30) получаем доверительную оценку для интервала предсказания $\hat{y}$:

$$\hat{y} \in (X^0)^T b \pm 3{,}896 t_{\alpha;n-k-1} \sqrt{1 + (X_0)^T (X^T X)^{-1} X^0}\,.$$

Например, вычислим с надежностью 1 - α = 0,95 доверительную оценку для значений гестационного возраста плодов и новорожденных со следующими гистометрическими показателями селезенки: средний диаметр лимфоидных узелков – 146,3 мкм, средняя толщина стенок центральных артерий – 8,13 мкм.

После формализации задания точка, определяемая вектором начальных условий, имеет вид:

$$X^0 = \begin{pmatrix} 1 \\ 146{,}3 \\ 8{,}13 \end{pmatrix},\ (X^0)^T = \begin{pmatrix} 1 & 146{,}3 & 8{,}13 \end{pmatrix}.$$

Предварительно найдем матричное произведение

$$(X^0)^T (X^T X)^{-1} X^0 = 0{,}01906158.$$

Тогда интервальная оценка для $\hat{y}$ равна:

$$\hat{y} = 26{,}374 \pm 3{,}896 \cdot 1{,}985 \sqrt{1 + 0{,}019} = 26{,}4 \pm 7{,}8 \text{ недель}.$$

Из приведенного выражения видно, что точность идентификации гестационного возраста плодов и новорожденных, обеспечиваемая использованием информации лишь о двух гистометрических показателях селезенки, очень низка (95% доверительный интервал охватывает почти 16 недель внутриутробного развития). Данное обстоятельство диктует необходимость создания более точной многофакторной регрессионной модели, включающей более информативные гистометрические показатели, в т.ч. и других фетальных органов.

3.4. НЕЛИНЕЙНАЯ РЕГРЕССИЯ

При построении регрессионных моделей идентификации личности возможны ситуации, когда связь между идентифицирующим признаком и идентифицируемым параметром является нелинейной. Выраженные отклонения парной связи от линейности могут быть легко установлены при визуальном анализе диаграммы рассеяния, когда точки наблюдений более точно сглаживаются не прямой, а кривой линией. В этом случае для построения криволинейной регрессии следует использовать подходы линейного регрессионного анализа. Для этого часто достаточно преобразовать переменные и соответствующие им значения наблюдений.

В зависимости от вида необходимых линеаризирующих преобразований различают следующие нелинейные регрессионные модели: нелинейные по переменным, нелинейные по параметрам и нелинейные и по переменным, и по параметрам [13].

Нелинейными по переменным являются модели, линеаризация которых достигается с помощью преобразований только лишь независимой переменной (переменных). Такие модели линейны по своей природе, отличаясь от обычных уравнений линейной регрессии только тем, что при оценивании их параметров приходится оперировать не с самой независимой переменной (переменными), а с ее преобразованием. Поэтому указанные нелинейные регрессии называются также регрессионными моделями с линейной структурой.

Основными типами регрессионных моделей, нелинейных по переменным, являются следующие функции:

- логарифмическая

$$\tilde{y} = \beta_0 + \beta_1 \log_a x;$$

- полиномиальная

$$\tilde{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k;$$

- гиперболическая

$$\tilde{y} = \beta_0 + \beta_1 \frac{1}{x};$$

- степенная

$$\tilde{y} = \beta_0 + \beta_1 x^n.$$

Необходимо пояснить, что выбор основания логарифмической модели влияет лишь на значения коэффициентов регрессии, пока-

затели же силы связи и точности прогнозирования остаются теми же. Поэтому на практике ограничиваются использованием лишь либо натурального, либо десятичного логарифмов. Относительно полиномиальной регрессии следует отметить, что, несмотря на принципиальную возможность использования для моделирования нелинейных взаимосвязей полиномов высоких степеней, на практике результаты прогнозирования с применением степеней, превышающих 3, зачастую оказываются нестабильными [74]. Кроме того, регрессионные коэффициенты полиномов высоких степеней вследствие мультиколлинеарности почти всегда становятся статистически незначимыми. Вышеперечисленное ограничивает практическое использование указанных аппроксимаций только лишь применением квадратного и кубического полиномов.

Доверительные интервалы для прогнозных оценок регрессионных уравнений нелинейных по переменным определяются по тем же стандартам, что и для прогнозных оценок одно- и многофакторных уравнений линейной регрессии. Отличием является лишь то, что при расчете доверительной области используется не сама независимая переменная (переменные), а ее преобразование. Для полиномиальных уравнений в указанных целях используются результаты возведения конкретного значения независимой переменной в соответствующие степени.

Например, 1-α - доверительный интервал для прогнозных оценок логарифмического уравнения $\hat{y} = b_0 + b_1 \lg x$, можно определить как

$$y \in \hat{y} \pm t_{\alpha;\,\text{n-2}} \cdot s_\varepsilon \cdot \left(1 + \frac{1}{n} + \frac{\left(\lg x_0 - \frac{\sum_{i=1}^{n} \lg x_i}{n} \right)^2}{\sum_{i=1}^{n} \left(\lg x_i - \frac{\sum_{i=1}^{n} \lg x_i}{n} \right)^2} \right)^{1/2}.$$

Подобным образом 1-α - доверительный интервал для прогнозных оценок гиперболического уравнения $\hat{y} = b_0 + \frac{b_1}{x}$, вычисляется по формуле

$$y \in \hat{y} \pm t_{\alpha;\,\text{n-2}} \cdot s_\varepsilon \cdot \left(1 + \frac{1}{n} + \frac{\left(\frac{1}{x_0} - \frac{\sum_{i=1}^{n} \frac{1}{x_i}}{n} \right)^2}{\sum_{i=1}^{n} \left(\frac{1}{x_i} - \frac{\sum_{i=1}^{n} \frac{1}{x_i}}{n} \right)^2} \right)^{1/2}.$$

Аналогично 1-α - доверительный интервал для прогнозных оценок полиномиального уравнения *k*-степени определяется из выражения:

$$y \in \hat{y} \pm t_{\alpha;\,\text{n-k-1}} \cdot s_\varepsilon \cdot \sqrt{1 + \overline{X}_0^T (X^T X)^{-1} \overline{X}_0},$$

где $\hat{y}(x_0, x_0^2, \ldots, x_0^{k-1}, x_0^k) = b_0 + b_1 x_0 + b_2 x_0^2 + \ldots + b_{k-1} x_0^{k-1} + b_k x_0^k$ - точечная оценка прогноза; X – матрица, состоящая из единичного первого столбца и k столбцов, содержащих n наблюдений, возведенных в каждую из k степеней:

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{k-1} & x_1^k \\ 1 & x_2 & x_2^2 & \cdots & x_2^{k-1} & x_2^k \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n-1} & x_{n-1}^2 & \cdots & x_{n-1}^{k-1} & x_{n-1}^k \\ 1 & x_n & x_n^2 & \cdots & x_n^{k-1} & x_n^k \end{pmatrix};$$

$\overline{X}_0$ - вектор-столбец, определяемый как

$$\overline{X}_0 = \begin{pmatrix} 1 \\ x_0 \\ x_0^2 \\ \vdots \\ x_0^{k-1} \\ x_0^k \end{pmatrix}.$$

Нелинейными по параметрам называются модели, линеаризация которых достигается с помощью преобразования только зависимой переменной. Такие модели по своей природе больше не являются линейными, т.е. их нельзя представить в виде простой регрессион-

ной модели с некоторыми преобразованиями независимых переменных. Поэтому указанные нелинейные регрессии называются также существенно нелинейными [13].

Основными типами нелинейных по параметрам регрессионных моделей являются следующие функции:

- показательная

$$\tilde{y} = \beta_0 + \beta_1^x;$$

- экспоненциальная

$$\tilde{y} = \beta_0 e^{\beta_1 x}.$$

Нелинейными по переменным и по параметрам называются модели, линеаризация которых достигается только лишь с помощью одновременного преобразования и независимой, и зависимой переменных. Примерами таких регрессионных уравнений являются функции:

$$\tilde{y} = \beta_0 x^{\beta_1};$$

$$\tilde{y} = \beta_0 e^{\frac{\beta_1}{x}};$$

$$\tilde{y} = \frac{\beta_0 x}{\beta_1 + x}.$$

Следующая таблица показывает наиболее часто используемые нелинейные зависимости y от x, которые легко могут быть линеаризованы (табл. 9).

Существенным недостатком регрессионных моделей последних двух видов является невозможность определения доверительных областей для прогнозных оценок стандартными методами линейной регрессии[6]. Поэтому с этой целью был предложен метод, основанный на использовании допустимых коэффициентов для нормального распределения [49].

В соответствии с указанным методом доверительные интервалы для прогнозных оценок гомоскедастичных нелинейных регрессионных уравнений определяются из выражения

$$y \in \hat{y} \pm k \cdot s_\varepsilon,$$

где $\hat{y}$ - точечная оценка прогноза; s_ε - остаточное стандартное отклонение; k - табличное значение допустимого коэффициента.

---

[6] Именно по этой причине статистические программные приложения определяют доверительные области для прогнозных оценок только лишь линейных регрессий и регрессий, нелинейных по переменным.

Таблица 9

Формулы перехода от параметров прямой линии $\hat{\dot{y}} = \dot{b}_0 + \dot{b}_1 \dot{x}$ к коэффициентам исходного соотношения [30]

Форма исходной зависимости	Преобразование переменных		Выражения для величин b_0 и b_1	
	$\dot{y} =$	$\dot{x} =$	$\dot{b}_0 =$	$\dot{b}_1 =$
$y = \frac{b_0}{b_1 + x}$	$\frac{1}{y}$	x	$\frac{b_1}{b_0}$	$\frac{1}{b_0}$
$y = \frac{x}{b_0 + b_1 x}$	$\frac{x}{y}$	x	b_0	b_1
$y = b_0 b_1^x$	$\lg y$	x	$\lg b_0$	$\lg b_1$
$y = b_0 x^{b_1}$	$\lg y$	$\lg x$	$\lg b_0$	b_1
$y = \beta_0 e^{\beta_1 x}$	$\ln y$	x	$\ln b_0$	b_1
$y = b_0 e^{\frac{b_1}{x}}$	$\ln y$	$\frac{1}{x}$	$\ln b_0$	b_1

Учитывая, что благодаря методу наименьших квадратов распределение остатков всегда является нормальным с генеральным средним, равным нулю, более точное определение доверительных областей для прогнозных оценок гомоскедастичных нелинейных регрессий может быть достигнуто с помощью использования односторонней верхней интервальной оценки остаточного стандартного отклонения:

$$y \in \hat{y} \pm z_\alpha \cdot s_{\varepsilon_B},$$

где $s_{\varepsilon_B} = \sqrt{\frac{s_\varepsilon^2 (n-1)}{\chi^2_{n-1;1-\alpha}}}$.

Для демонстрации практического использования изложенных принципов построения нелинейных регрессионных моделей приведем результаты поиска различных нелинейных аппроксимаций моделей идентификации гестационного возраста плодов и новорожденных по степени кроветворной активности паренхимы печени.

Объектами данного исследования явились трупы 131 плода и новорожденного, оставшиеся после исключения из выборки выбросов

и кластера недоношенных новорожденных с постнатальной инволюцией экстрамедуллярной миелоидной ткани (см. раздел 2.6).

Проведенный анализ показал, что статистическая зависимость гестационного возраста от указанного гистометрического показателя является нелинейной. Это определило необходимость поиска наиболее адекватной нелинейной регрессионной модели идентификации. В процессе данного поиска был построен комплекс охарактеризованных выше регрессионных моделей, нелинейных по переменным, по параметрам, а также по переменным и параметрам одновременно (рис. 19-22):

$$\hat{y} = 50{,}598 - 5{,}960 \ln x;$$

$$\hat{y} = 40{,}639 - 0{,}354x + 1{,}748 \cdot 10^{-3} x^2;$$

$$\hat{y} = 42{,}574 - 0{,}521x + 4{,}902 \cdot 10^{-3} x^2 - 1{,}597 \cdot 10^{-5} x^3;$$

$$\hat{y} = 25{,}715 + \frac{67{,}725}{x};$$

$$\hat{y} = 32{,}444 - 0{,}001x^2;$$

$$\hat{y} = 30{,}758 - 6{,}938 \cdot 10^{-6} x^3;$$

$$\hat{y} = 36{,}647 \cdot 0{,}995^x;$$

$$\hat{y} = 36{,}647 e^{-0{,}005x};$$

$$\hat{y} = 58{,}065 x^{-0{,}197};$$

$$\hat{y} = 25{,}558 e^{\frac{2{,}183}{x}};$$

$$\hat{y} = \frac{5739{,}432}{x + 155{,}179};$$

$$\hat{y} = \frac{x}{0{,}046x - 0{,}308}.$$

Несмотря на достаточно большое количество методов линеаризации, некоторые регрессионные модели нелинейные по параметрам не могут быть сведены к линейным простым преобразованием начальных данных. Задача построения подобных моделей поддается лишь численным методам решения с использованием программных средств [139]. Например, регрессионное уравнение идентификации гестационного возраста может быть аппроксимировано следующей моделью экспоненциального роста, полученной численным методом (рис. 23):

$$\hat{y} = 21{,}990 + e^{(3{,}047 - 0{,}029x)}$$ [113].

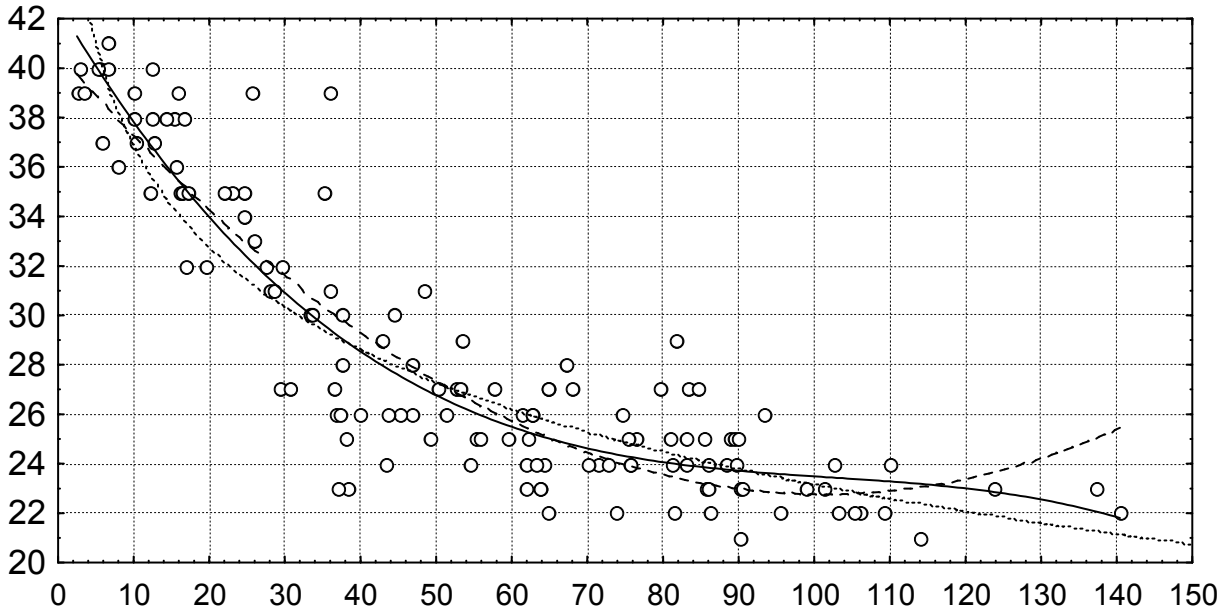

Рис. 19. Геометрическая интерпретация полиномиальной и логарифмической регрессионных моделей идентификации гестационного возраста по степени кроветворной активности печени. Здесь и на рис. 20-22: по оси абсцисс – кроветворная активность печени с учетом усадки, количество ядер миелоидных клеток в тестовой площади; по оси ординат – гестационный возраст, недель. Непрерывной линией показан кубический полином, пунктирной – квадратный полином, точечной – логарифмическая модель.

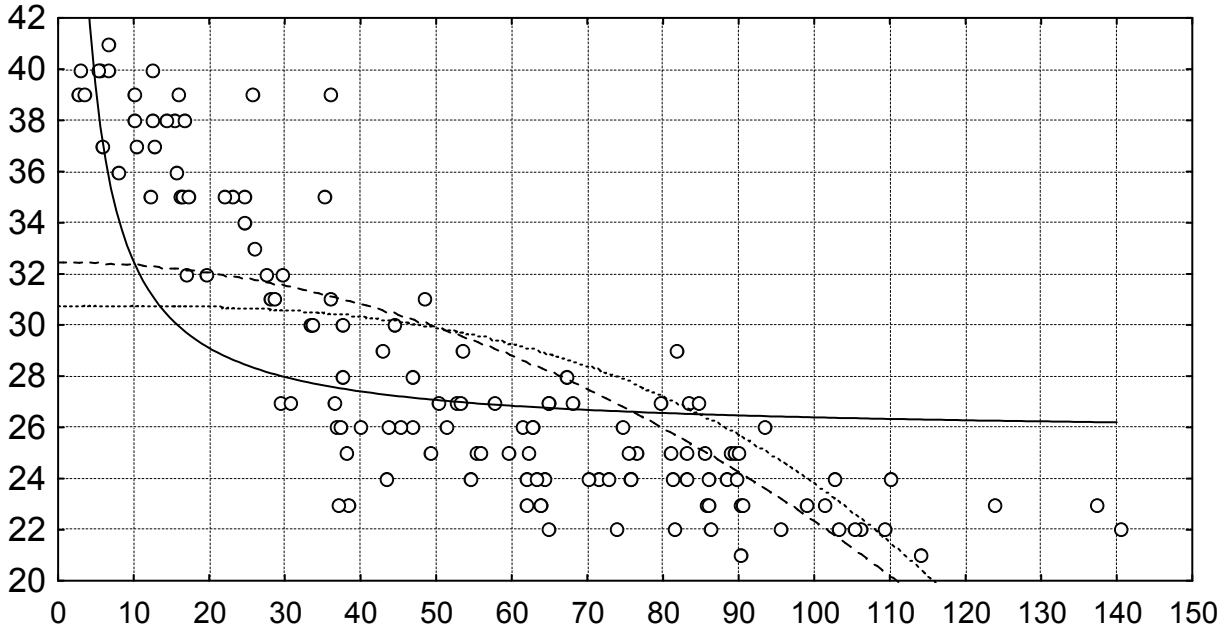

Рис. 20. Геометрическая интерпретация гиперболической и степенной регрессионных моделей идентификации гестационного возраста по степени кроветворной активности печени. Непрерывной линией показана гипербола, пунктирной и точечной – степенная модель с показателем степени, равным 2 и 3 соответственно.

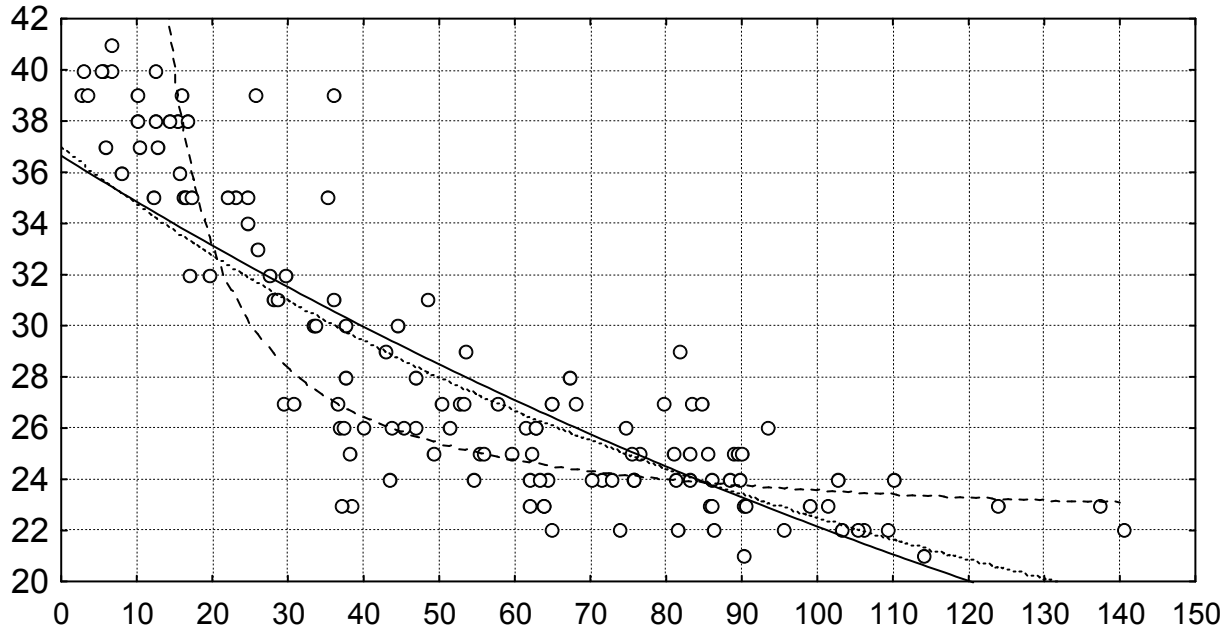

Рис. 21. Геометрическая интерпретация нелинейных по параметрам регрессионных моделей идентификации гестационного возраста по степени кроветворной активности печени. Непрерывной линией показаны тождественные графики функций: $\hat{y} = 36{,}647 \cdot 0{,}995^{x}$ и $\hat{y} = 36{,}647e^{-0{,}005x}$, пунктирной - $\hat{y} = x/(0{,}046x - 0{,}308)$, точечной - $\hat{y} = 5739{,}432/(x + 155{,}179)$.

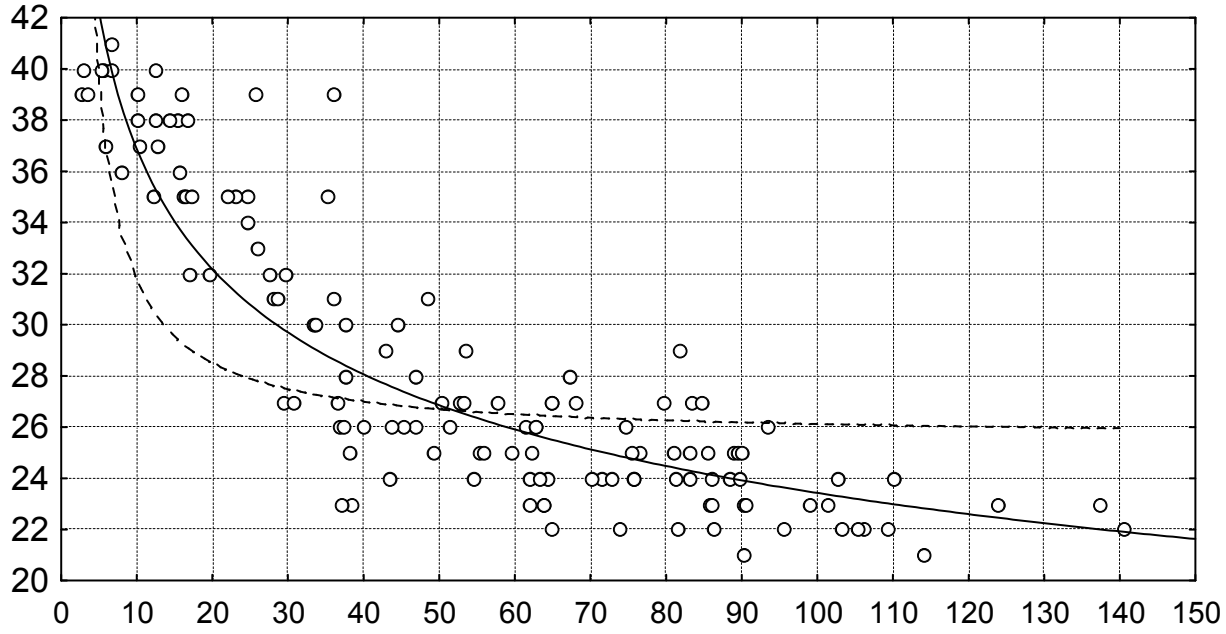

Рис. 22. Геометрическая интерпретация нелинейных по параметрам и переменным регрессионных моделей идентификации гестационного возраста по степени кроветворной активности печени. Непрерывной линией показана функция вида $\hat{y} = 58{,}065x^{-0{,}197}$, пунктирной – $\hat{y} = 25{,}558e^{\frac{2{,}183}{x}}$.

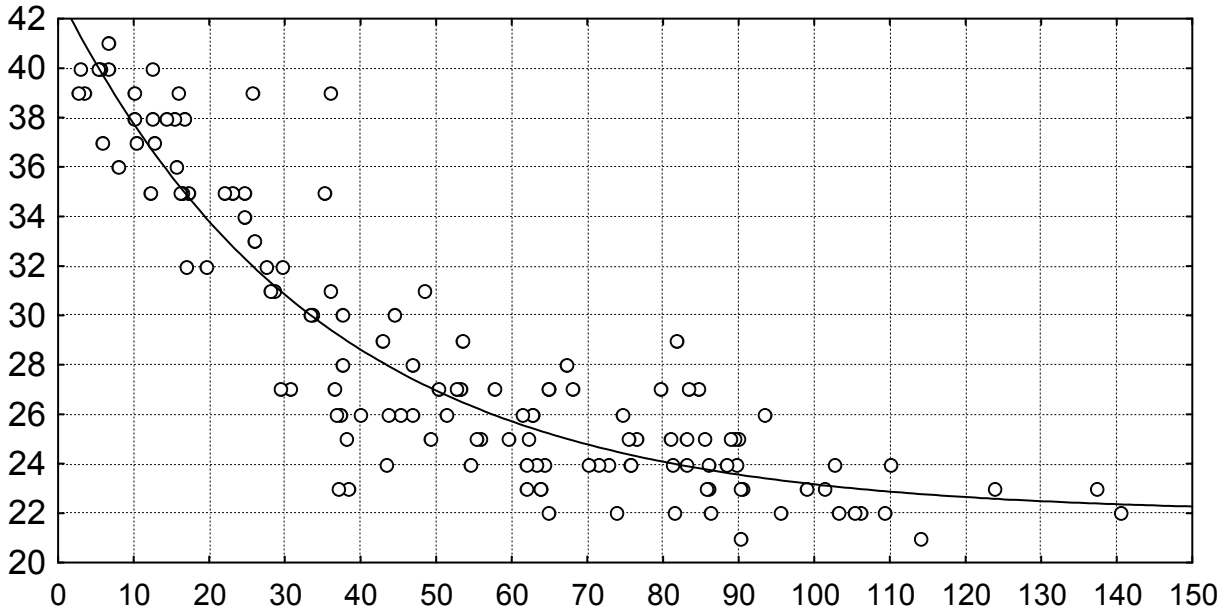

Рис. 23. Геометрическая интерпретация регрессионной модели идентификации гестационного возраста по степени кроветворной активности печени, аппроксимированной функцией экспоненциального роста. Функция потерь – метод наименьших квадратов. Алгоритм минимизации функции потерь – квазиньютоновский метод.

Существуют также способы создания нелинейных регрессий, основанные на использовании других, отличных от метода наименьших квадратов, функций потерь и различных алгоритмов их минимизации [13,120,124,138].

При подгонке нелинейных регрессионных моделей распространенной альтернативой минимизации функции потерь наименьших квадратов является максимизация функции правдоподобия или логарифма функции правдоподобия [89,108,152]. Применительно к регрессионному анализу функция правдоподобия определяется как

$$L = \prod_{i=1}^{n} f(y_i, x_i; \beta; \sigma_\varepsilon^2)$$

и позволяет вычислить вероятность (обозначенную как L, от слова likelihood - правдоподобие) появления конкретных значений зависимой переменной в выборке при заданной регрессионной модели [13,111].

Однако, поскольку линейная регрессия очень удобна для математико-статистического анализа и наиболее просто интерпретируема, то при описании нелинейных зависимостей предпочтение следует отдавать аппроксимациям, основанным на линеаризирующих преобразованиях независимых переменных.

3.5. НЕОДНОРОДНАЯ РЕГРЕССИЯ

Серьезной проблемой регрессионного анализа является неоднородность изучаемых данных за счет наличия в них выбросов или кластеринга, которые всегда приводят к искажению реальных взаимозависимостей между изучаемыми параметрами. В этой связи проверка наличия неоднородности данных и ее исключение должны быть осуществлены еще на этапе изучения корреляционных взаимосвязей (см. раздел 2.6). В данном разделе будут лишь изложены особенности обнаружения выбросов в совокупности данных, предназначенных для проведения регрессионного анализа.

Применительно к статистическому анализу в целом проблема наличия выбросов в исходных совокупностях данных давно известна, вследствие чего постоянно разрабатывались методы непосредственного выявления выбросов, а также методы обработки статистической информации, устойчивые к их наличию [76,83]. Существующие статистические критерии позволяют выявлять как одно, так и несколько максимальных или минимальных экстремальных наблюдений в упорядоченных рядах данных и даже несколько наибольших и наименьших экстремальных наблюдений одновременно[7]. Наиболее распространенными из указанных методов являются критерии Йейтса, Диксона, Смирнова-Граббса, Граббса и Титьена-Мура [26,102,103,155].

Единственным условием для применения указанных критериев является наличие в выборочной совокупности данных не менее чем трех наблюдений. Именно это обстоятельство и является характерным затруднением при осуществлении регрессионного анализа, так как, несмотря на значительное общее количество наблюдений, непосредственное выявление выбросов необходимо проводить отдельно в каждом из упорядоченных рядов, на которые разделена исходная совокупность данных. При этом каждый ряд образуют все

---

[7] Кроме упомянутых разработан также ряд статистических методов, предназначенных для выявления выбросов по остаткам регрессионных моделей (критерии Эконта, Титьена-Мура-Бекмана, Прескотта-Лунда и др.). Однако, на наш взгляд, выбросы целесообразнее искать (а в случае обнаружения решать вопрос об их исключении) до начала регрессионного анализа, а не после его проведения. Данный подход исключает влияние гетероскедастичности и качества регрессионных аппроксимаций на чувствительность методов обнаружения выбросов.

имеющиеся значения зависимой переменной, соответствующие конкретному значению независимой переменной. Чем меньше цена деления шкалы, по которой оценивается величина независимой переменной, тем больше количество рядов значений и, соответственно, меньше число наблюдений в них. После такой группировки могут образоваться ряды, содержащие только одно или два наблюдения. При подозрении на наличие выбросов в указанных рядах (относительно всей совокупности исследуемых данных) обнаружение и дальнейшее удаление первых с помощью имеющихся критериев становится невозможным. Между тем, единичный выброс может существенно повлиять на наклон прямой регрессии и, следовательно, на значение коэффициента корреляции. Исключение подозрительных наблюдений, которые, возможно, не являются выбросами, также способно оказать выраженное влияние на линию регрессии и сильно исказить истинное значение коэффициента корреляции, особенно при относительно небольшой общей совокупности исходных данных.

В связи этим нами был предложен метод выявления выбросов в упорядоченных рядах значений, предназначенных для проведения регрессионного анализа [52]. В соответствии с алгоритмом метода вначале производится группировка исходных данных с формированием упорядоченных рядов. После этого осуществляется объединение подозрительного на выброс значения из ряда небольшого объема с наблюдениями из соседнего ряда большего размера. Затем проводится проверка указанного значения на принадлежность к выбросам с помощью какого-либо из существующих критериев. В зависимости от знака коэффициента парной корреляции и расположения подозрительных на выбросы значений их объединение производится с противоположными соседними рядами. При положительной корреляционной связи между переменными, когда наклон линии регрессии направлен вверх, подозрительные на выбросы максимальные значения, объединяются с последующим, а минимальные – с предыдущим соседним рядом. В случае отрицательной корреляционной связи между переменными, когда наклон линии регрессии направлен вниз, подозрительные на выбросы максимальные значения объединяются с предыдущим, а минимальные – с последующим рядом.

Указанный метод является достаточно эффективным средством выявления выбросов при проведении регрессионного анализа в

случае недостаточного объема выборочных данных. Возможности метода не ограничиваются лишь однофакторной регрессией и могут быть использованы при проведении многомерного регрессионного анализа. Для этого предварительно необходимо проверить на выбросы все двумерные совокупности, образованные значениями зависимой и каждой из независимых переменных.

Поскольку объединение наблюдений осуществляется с соседними рядами с большим или меньшим ожидаемым значением зависимой переменной, чувствительность любых критериев обнаружения выбросов, будет снижена тем заметнее, чем больше наклон линии регрессии. Из-за высокой жесткости метода некоторые данные, в действительности являющиеся выбросами, могут быть не признаны таковыми, то есть возрастает риск ошибки второго рода. Кроме того, использование данного метода возможно только при условии монотонности регрессии и отсутствии гетероскедастичности.

Возможности изложенного метода можно показать на примере исследования гестационной динамики кроветворной активности печени плодов и новорожденных[8]. Кроветворная активность определялась путем подсчета среднего количества миелоидных клеток в поле зрения печеночной паренхимы без учета усадки тканей. Изучены фрагменты печени от 89 плодов. После построения диаграммы рассеяния подозрительными на выбросы признаны 3 значения, одно из которых соответствует гестационному сроку 25 недель, а два других – 30 неделям (рис. 24). Допустим, проверку на выбросы необходимо осуществить при уровне значимости, равном 0,10. Проверить ряд значений кроветворной активности, соответствующих гестационному сроку 25 недель, на наличие выбросов с помощью любого из существующих статистических методов не составляет труда. С помощью одностороннего варианта метода Диксона данное значение идентифицировано как выброс.

Сложнее обстоит дело с остальными значениями, поскольку упорядоченный ряд значений кроветворной активности, соответствующих гестационному возрасту 30 недель, содержит всего два наблюдения, оба из которых подозреваются на принадлежность к выбросам. Используем изложенный метод. Поскольку имеет место отрицательная корреляция, нижнее, подозрительное на выброс,

[8] Данный пример сконструирован исключительно в целях демонстрации описываемого метода идентификации выбросов и не отражает реальных результатов проведенного исследования.

значение, равное 13,4, из ряда с гестационным возрастом 30 недель добавляется в последующий соседний ряд, содержащий 5 наблюдений. После добавления указанного наблюдения получился упорядоченный ряд, содержащий следующие значения: 85,0; 84,1; 62,6; 49,8; 48,8; 13,4. При проверке по методу Диксона, минимальное значение в данной выборке опознано как выброс.

Верхнее, подозрительное на выброс, значение, равное 129,7, из ряда с гестационным возрастом 30 недель добавляется в предыдущий соседний ряд, содержащий 3 наблюдения. После добавления указанного наблюдения получился упорядоченный ряд, содержащий следующие значения: 129,7; 93,1; 83; 74,8. При проверке по методу Диксона максимальное значение в данной выборке на 10% уровне значимости нельзя признать как выброс. Не исключено, что выброс не опознан из-за большой жесткости использованного метода, так как подозрительное значение добавлялось в ряд с большим ожидаемым значением кроветворной активности.

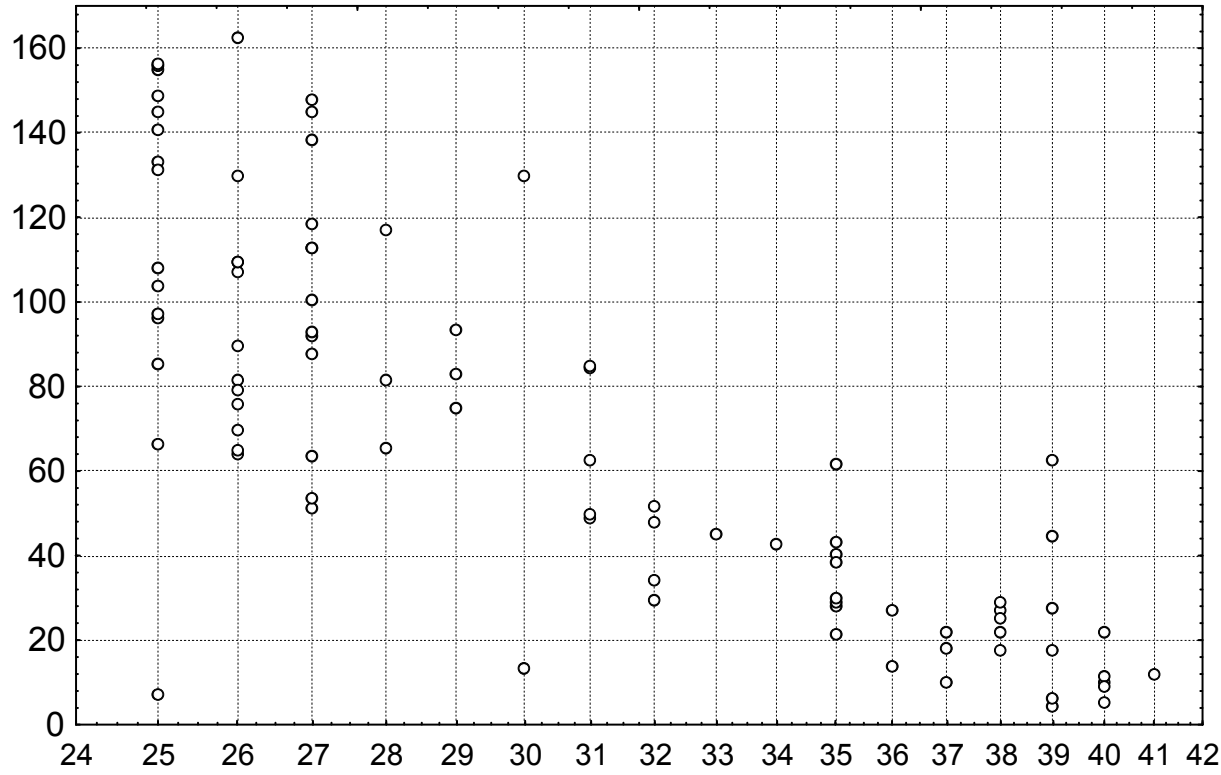

Рис. 24. Диаграмма рассеяния значений кроветворной активности паренхимы печени плодов и новорожденных. По оси абсцисс – гестационный возраст, недель; по оси ординат – кроветворная активность без учета усадки, количество фрагментов ядер миелоидных клеток в тестовой площади.

3.6. КРИТЕРИИ ТОЧНОСТИ РЕГРЕССИОННЫХ МОДЕЛЕЙ ИДЕНТИФИКАЦИИ ЛИЧНОСТИ И МЕТОДЫ ИХ СРАВНЕНИЯ

Характерной чертой судебно-медицинских научных исследований, посвященных идентификации личности, является широкое использование для обработки биометрических данных таких математико-статистических методов, как корреляционный и регрессионный анализ. Применение данных видов статистического анализа позволяет в рамках одного исследования создавать целый комплекс альтернативных регрессионных моделей идентификации личности, различающихся по точности получения прогнозных оценок идентифицируемых параметров и по числу входящих в уравнение биометрических показателей.

В настоящее время в качестве основных показателей точности альтернативных регрессионных моделей рассматриваются коэффициенты множественной (или парной) корреляции и остаточные дисперсии [см. напр. 42,60]. Однако выборочные оценки указанных параметров, как правило, не совпадают с их неизвестными истинными значениями [16]. Поэтому наименьшая дисперсия остатков (и, соответственно, наибольший модуль коэффициента парной или множественной корреляции) по данным изучения ограниченной выборки случайно может соответствовать регрессионному уравнению, в реальности не являющемуся самым точным из комплекса альтернативных моделей [73]. В этой связи важнейшей задачей, возникающей при проведении судебно-медицинского антропологических исследований, является объективный выбор из комплекса альтернативных моделей функции регрессии, характеризующейся наибольшей точностью определения идентифицируемого параметра [53,60].

Надежным решением проблемы объективного сравнения точности альтернативных регрессионных моделей идентификации личности, на наш взгляд, может быть использование разработанных в математической статистике критериев сравнения корреляционных коэффициентов и дисперсий [30]. Это позволило предложить два адаптированных метода объективного сравнения точности альтернативных регрессионных моделей идентификации личности, которые уже были успешно использованы нами при идентификации гестационного возраста плодов и новорожденных [53,59].

Первый метод сравнения точности альтернативных регрессионных моделей основан на сравнении их корреляционных коэффициентов. При реализации метода необходимо учитывать, что сравнению подлежат следующие показатели силы связи:

- коэффициенты парной корреляции между идентифицируемым параметром и идентифицирующим показателем (для моделей однофакторной линейной регрессии);
- коэффициенты парной корреляции между идентифицируемым параметром и линеаризирующим преобразованием идентифицирующего показателя (для однофакторных регрессионных моделей, нелинейных по переменным);
- коэффициенты парной корреляции между линеаризирующим преобразованием идентифицируемого параметра и идентифицирующим показателем (для однофакторных регрессионных моделей, нелинейных по параметрам);
- коэффициенты парной корреляции между линеаризирующими преобразованиями идентифицируемого параметра и идентифицирующего показателя (для однофакторных регрессионных моделей, нелинейных по переменным и параметрам);
- скорректированные коэффициенты множественной корреляции между идентифицируемым параметром и идентифицирующими показателями или их линеаризирующими преобразованиями (для многофакторных регрессионных моделей).

Алгоритм сравнения точности альтернативных регрессионных моделей по коэффициентам корреляции включает выполнение трех этапов.

На первом этапе производится сравнение оценок коэффициентов корреляции всех альтернативных регрессионных моделей, построенных при выполнении данного судебно-антропологического исследования, по формуле (8). Если полученная χ^2- статистика меньше границы значимости при $(k-1)$ степенях свободы, то нульгипотеза $\rho_1 = \rho_2 = \cdots = \rho_k$ не отклоняется и, соответственно, приоритет в точности любого из сравниваемых регрессионных уравнений не является доказанным. При превышении статистики критического значения χ^2 гетерогенность точности имеющихся регрессионных моделей является доказанной с надежностью, равной $1-\alpha$.

На втором этапе при наличии значимой неоднородности коэффициентов корреляции регрессионных моделей производят ранжи-

рование их модулей: $|r_1| > |r_2| > \ldots > |r_k|$ и последующие попарные сравнения. При этом сначала сравниваются модули $|r_1|$ и $|r_k|$, затем $|r_1|$ и $|r_{k-1}|$, потом, при наличии значимых различий, $|r_1|$ и $|r_{k-2}|$, и т.д. Сравнения осуществляются до получения первого статистически незначимого результата с помощью *z*-преобразования по формуле (10).

На третьем этапе применяют поправку Бонферрони для учета эффекта множественных сравнений: $p_{скор.} = p_\alpha n$, где $p_{скор.}$ - скорректированное значение вероятности ошибочного принятия альтернативной гипотезы о наличии различий; p_α - аналогичная вероятность без учета эффекта множественных сравнений; n – число сравнений [16].

Алгоритм второго метода сравнения точности альтернативных регрессионных моделей идентификации личности по остаточным дисперсиям включает выполнение аналогичных этапов.

На первом этапе производят сравнение остаточных дисперсий всех сравниваемых регрессионных уравнений, построенных при выполнении данного судебно-медицинского антропологического исследования, по Бартлету:

$$\chi^2 = \frac{1}{c}\left[2{,}3026k(n-1)\left\{\lg s_\varepsilon^2 - \frac{1}{k}\sum_{i=1}^{k}\lg s_{\varepsilon i}^2\right\}\right],$$

где $s_{\varepsilon i}^2$ - остаточная дисперсия *i*-й регрессионной модели;

$$c = \frac{k+1}{3k(n-1)} + 1, \text{ а } s_\varepsilon^2 = \frac{1}{k}\sum_{i=1}^{k} s_{\varepsilon i}^{\,2} \text{ [30,116].}$$

Если полученная статистика больше, чем критическое значение для заданной статистической надежности или равна ему, то нуль-гипотеза $\delta_{\varepsilon 1}^2 = \delta_{\varepsilon 2}^2 = \cdots = \delta_{\varepsilon k}^2$ отклоняется и, соответственно, является доказанной гетерогенность точности альтернативных регрессионных моделей.

На втором этапе при наличии значимой неоднородности остаточных дисперсий сравниваемых регрессий производят их ранжирование по величине: $s_{\varepsilon 1}^2 \rangle s_{\varepsilon 2}^2 \rangle \cdots s_{\varepsilon k}^2$ и последующие попарные сравнения. При этом сначала сравниваются модули $s_{\varepsilon k}^2$ и $s_{\varepsilon 1}^2$, затем, при наличии значимых различий, $s_{\varepsilon k}^2$ и $s_{\varepsilon 2}^2$ и т.д. Сравнения осуществляются до получения первого статистически незначимого результа-

та по формуле $F = s^2_{\varepsilon\max} / s^2_{\varepsilon\min}$ с числом степеней свободы $v_1 = v_2 = n - 1$ для одностороннего варианта критерия [30].

На третьем этапе применяют поправку Бонферрони для учета эффекта множественных сравнений.

Дальнейшее изложение методов сравнения точности альтернативных регрессий продолжим на примере идентификации гестационного возраста по степени кроветворной активности паренхимы фетальной печени. В ходе данного исследования был создан комплекс альтернативных регрессий (см. раздел 3.4), наибольшими критериями точности из которых обладала кубическая полиномиальная модель (табл. 10).

Комплексное сравнение основных альтернативных регрессий показало наличие их значимой неоднородности как по коэффициентам парной и множественной корреляции ($\chi^2 = 45{,}893; p = 3{,}110 \cdot 10^{-8}$), так и по остаточным дисперсиям ($\chi^2 = 69{,}998; p = 1{,}026 \cdot 10^{-13}$). Это вызвало необходимость последующего проведения попарных сравнений по указанным параметрам кубического полинома с остальными регрессиями (табл. 11).

Проведенные сравнения показали, что точность кубической полиномиальной модели идентификации гестационного возраста значимо превышает таковую гиперболической, линейной и экспоненциальной моделей. Вместе с тем она не отличается от точности логарифмической и квадратной полиномиальной моделей (см. табл. 11).

Таблица 10

Оценки качества некоторых регрессионных моделей идентификации гестационного возраста

Модель регрессии	F	p	r	r^2*	s_ε	s^2_ε
Кубический полином	232,240	$2{,}292 \cdot 10^{-51}$	0,920	0,842	2,279	5,195
Квадратный полином	311,291	$6{,}854 \cdot 10^{-50}$	0,911	0,827	2,388	5,701
Логарифмическая	547,161	$3{,}060 \cdot 10^{-48}$	0,900	0,809	2,516	6,329
Экспоненциальная	338,768	$6{,}772 \cdot 10^{-38}$	0,851	0,724	2,907	8,452
Линейная	308,004	$5{,}522 \cdot 10^{-36}$	0,840	0,705	3,129	9,792
Гипербола	119,584	$4{,}229 \cdot 10^{-20}$	0,694	0,481	4,149	17,215

Примечание. * - для полиномиальных регрессионных моделей приведены значения скорректированного r^2.

Таблица 11

Результаты сравнений точности кубического полинома с другими регрессионными моделями идентификации гестационного возраста

Модель сравнения	z	p_1^z*	$4p_1^z$**	F	p_1^F*	$4p_1^F$**
Гипербола	5,756	$4{,}3 \cdot 10^{-9}$	$1{,}7 \cdot 10^{-8}$	3,314	$1{,}7 \cdot 10^{-11}$	$6{,}8 \cdot 10^{-11}$
Линейная	2,838	0,002	0,009	1,885	$1{,}7 \cdot 10^{-4}$	$6{,}9 \cdot 10^{-4}$
Экспоненциальная	2,516	0,006	0,024	1,627	0,003	0,012
Логарифмическая	0,835	0,202	0,808	1,218	0,131	0,523
Квадратный полином	0,407	0,342	-	1,097	0,298	-

Примечание. * - вероятность ошибочного принятия нуль-гипотезы при одностороннем варианте критерия. ** - вероятность ошибочного принятия нуль-гипотезы с учетом эффекта множественных сравнений.

В аспекте практического использования описанного метода весьма важной является проблема выбора наиболее оптимальной тактики проведения попарных сравнений корреляционных коэффициентов или остаточных дисперсий. Под оптимальностью в данном случае подразумевается получение исчерпывающих результатов сравнительного анализа при наименьшем числе попарных сравнений. Объясняется это тем, что при большом количестве альтернативных регрессий коррекция эффекта множественных сравнений сопровождается выраженным снижением чувствительности используемых статистических критериев.

В этой связи целесообразно использовать следующую тактику, эффективность которой в общем виде доказана в теории информации [85].

В соответствии с указанным алгоритмом необходимо разбить ранжированное множество x всех подлежащих сравнению с кубическим полиномом регрессий на две равные по численности части и сравнить точность кубического полинома с точностью модели, занимающей срединное положение. Поскольку все регрессии были получены на основе анализа одной и той же совокупности данных, то чувствительность всех статистических критериев будет зависеть только от величины различий. Это позволяет при обнаружении различий точности кубического полинома с точностью регрессии, занимающей срединное положение, не проводить дальнейшие попарные сравнения со всеми регрессиями, точность которых ниже точности регрессионной модели, занимающей срединное положение. В

нашем примере срединное положение занимает экспоненциальная регрессионная модель. Поскольку точность экспоненциальной модели значимо уступает точности кубического полинома, то без проведения соответствующих сравнений можно утверждать, что точность гиперболы и линейной регрессии также значимо ниже точности кубического полинома.

Далее следует точно таким же образом разбить оставшееся множество альтернативных регрессий на две возможно более близкие по численности части и произвести сравнение с регрессионной моделью, занимающей срединное положение в данной группе регрессий. В нашем примере такое положение занимает логарифмическая модель.

Вообще, наименьшее число k сравнений, достаточное для полного выявления всех различий сравниваемой модели с n альтернативными регрессиями, определяется неравенствами

$$k - 1 \prec \log_2 n \le k \ \text{(или } 2^{k-1} \prec n \le 2^k\text{)}.$$

Независимо от значения n

$$k \ge \log_2 n;$$

при этом $k = \log_2 n$ только в том случае, когда число n является целой степенью числа 2 и, следовательно, $\log_2 n$ есть целое число [85].

При таком подходе для получения вышеописанного результата необходимо осуществление всего лишь двух сравнений точности кубического полинома: с точностью экспоненциальной модели, а затем – логарифмической.

Следует отметить, что метод сравнения точности альтернативных регрессий, основанный на сравнении коэффициентов корреляции, применим только при исследовании линейных и нелинейных монотонных зависимостей. В отличие от него метод, основанный на сравнении остаточных дисперсий, обладая большей чувствительностью, может быть использован при исследовании регрессионных зависимостей любой формы.

Необходимо также подчеркнуть, что спектр возможных методов сравнения дисперсий, разработанных математиками, не ограничивается лишь методом Бартлета. Кроме названных, известны также методы сравнения нескольких дисперсий как равных, так и неодинаковых объемов [90,97,100,118,131]. Последнюю группу методов или формулу (9) целесообразно использовать в экспертной практи-

ке для объективного выбора наиболее оптимального из группы альтернативных способов идентификации личности, разработанных разными авторами.

Таким образом, предложенные методы позволяют проводить объективное сравнение точности альтернативных регрессий и последующий объективный выбор наиболее точных регрессионных моделей для их использования в судебно-медицинской экспертной практике, повышая тем самым диагностическую значимость результатов судебно-антропологических исследований. Метод сравнения точности альтернативных регрессий, основанный на сравнении их остаточных дисперсий, является более предпочтительным, поскольку не имеет ограничений к применению и обладает большей чувствительностью по сравнению с аналогичным методом, основанным на сравнении коэффициентов корреляции.

Приведенные методы можно с успехом использовать не только в судебно-антропологических научных исследованиях, но и в экспертной судебно-медицинской практике в целях объективного выбора из комплекса разработанных разными авторами способов идентификации одной методики, характеризующейся наибольшей диагностической значимостью. Поскольку указанные способы идентификации, как правило, базируются на результатах исследования выборок разных объемов, сравнительный анализ остаточных дисперсий должен проводиться по несколько иной формуле:

$$\chi^2 = \frac{1}{c}\left[2{,}3026\left(v \lg s_\varepsilon^2 - \sum_{i=1}^{k} v_i \lg s_{\varepsilon i}^2\right)\right] \text{ с } (k-1) \text{ степенями свободы,}$$

где $v = n - k$ - общее число степеней свободы $= \sum_{i=1}^{k} v_i$, n – объем объединенной выборки; k – число групп, каждая из которых должна включать не менее 5 наблюдений; $v_i = n_i - 1$ - число степеней свободы в i-й группе;

$$c = \frac{\sum_{i=1}^{k} \frac{1}{v_i} - \frac{1}{v}}{3(k-1)} + 1;$$

$$s_\varepsilon^2 = \frac{\sum_{i=1}^{k} v_i s_{\varepsilon i}^2}{v} \text{ [30].}$$

3.7. ПРОБЛЕМА МУЛЬТИКОЛЛИНЕАРНОСТИ

Многофакторный регрессионный анализ в судебно-медицинских антропологических исследованиях практически всегда осложняется тем, что идентифицирующие признаки не являются взаимно независимыми. В теории корреляционно-регрессионного анализа явление связи между независимыми переменными, влияние которых на результативный показатель предполагается выразить посредством уравнения множественной регрессии, носит название мультиколлинеарности (избыточности информации). Поскольку каждая из независимых переменных включается в уравнение множественной регрессии без учета связи с другими переменными, то наличие мультиколлинеарности сопровождается ростом стандартных ошибок некоторых или всех коэффициентов регрессии. В результате доверительные интервалы для коэффициентов регрессии расширяются, а соответствующие *t*-тесты не будут значимыми. В случае сильной мультиколлинеарности может оказаться, что регрессия в целом очень высоко значима (исходя из результатов *F*-теста), однако ни один из *t*-тестов для отдельных факторных переменных значимым не является [74].

Обнаружение мультиколлинеарности при регрессионном анализе – самостоятельная важная задача. Существует довольно много статистических индикаторов избыточности информации (толерантность, получастные коэффициенты корреляции и др.). Основным методом обнаружения мультиколлинеарности является анализ коэффициентов детерминации.

Благодаря свойству аддитивности, при отсутствии корреляции между факторными переменными множественный коэффициент детерминации равен сумме коэффициентов парной детерминации, вычисленных для результативного показателя y и каждой из независимых переменных $x_1 x_2, \ldots, x_k$:

$$r^2_{y/1,2,\ldots,k} = \sum_{i=1}^{k} r^2_{yi} .$$

При наличии мультиколлинеарности указанное соотношение нарушается тем больше, чем сильнее она проявляется. Вследствие этого в качестве показателя интенсивности мультиколлинеарности M используют выражение:

$$M = 1 - \frac{\sum_{i=1}^{k} r_{yi}^2}{r_{y/1,2,\ldots,k}^2}$$ [75].

Мультиколлинеарность тем интенсивнее, чем ближе M к 1.

Факт наличия мультиколлинеарности можно проверить с помощью χ^2- критерия:

$$\chi_{\alpha;v}^2 = -\left[n - 1 - \frac{1}{6}(2k+5)\right]\ln|R|,$$

где $v = \frac{1}{2}k(k-1)$; $|R|$ – определитель матрицы

$$\begin{pmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ r_{k1} & r_{k2} & \cdots & 1 \end{pmatrix}$$ [75].

При $\chi^2 > \chi_{\alpha;v}^2$ наличие мультиколлинеарности между факторными переменными считается доказанным, в обратном случае сохраняется нулевая гипотеза об отсутствии мультиколлинеарности.

После установления наличия мультиколлинеарности, рекомендуется выяснить, какие именно факторные переменные играют в ней определяющую роль. Для этого предложено вычислять коэффициенты частной корреляции между каждыми двумя влияющими переменными при исключении влияния всех остальных переменных, а затем определять статистическую значимость каждого полученного частного критерия корреляции с помощью (7). При наличии значимой коллинеарности между двумя переменными одна из них может быть исключена из рассмотрения [75].

Изложенный метод характеризуется двумя недостатками: большой трудоемкостью и отсутствием критериев индикации переменной (из двух коррелированных между собой), подлежащей исключению. В этой связи нами предложен альтернативный метод, лишенный названных недостатков [49,53].

Суть данного метода заключается в том, что после построения регрессионного уравнения производится вычисление коэффициентов частной корреляции между результативным показателем и каждой из независимых переменных при исключении влияния остальных факторных переменных и определяется их значимость. Исклю-

чению подлежат независимые переменные, характеризующиеся статистически незначимыми частными корреляционными связями с результативным показателем.

Рассмотрим вновь пример с идентификацией гестационного возраста по гистометрическим показателям селезенки. Полный набор гистометрических показателей, рассматривавшихся нами в качестве потенциальных независимых переменных многофакторной регрессионной модели идентификации, включал следующие параметры: средние диаметр и плотность расположения лимфоидных узелков, среднюю толщину стенок центральных артерий, среднюю толщину капсулы на диафрагмальной поверхности селезенки. Исследованная совокупность состояла из 98 наблюдений (одно наблюдение было исключено как выброс по показателю толщины капсулы).

Введем условные обозначения: y – гестационный возраст, недель; x_1 – средний диаметр лимфоидных узелков селезенки, мкм; x_2 – средняя плотность расположения лимфоидных узелков; x_3 – средняя толщина стенок центральных артерий селезенки, мкм; x_4 – средняя толщина капсулы, мкм. Показатель интенсивности мультиколлинеарности многофакторного регрессионного уравнения $\tilde{y} = \beta_0 + \beta_1 x_1 + \beta_2 \lg x_2 + \beta_3 x_3 + \beta_4 x_4$ равнялся $M = 0{,}520$. Соответствующая проверка показала значимость мультиколлинеарности ($\chi^2 = 78{,}591; p = 6{,}981 \cdot 10^{-15}$)[9]. Результаты остальных вычислений приведены в таблице 12.

Таблица 12

Критерии мультиколлинеарности регрессионного уравнения $\tilde{y} = \beta_0 + \beta_1 x_1 + \beta_2 \lg x_2 + \beta_3 x_3 + \beta_4 x_4$

Показатель	Частный r	Толерантность	t	p
x_1	0,335	0,656	3,427	$9{,}111 \cdot 10^{-4}$
$\lg x_2$	-0,574	0,644	-6,766	$1{,}163 \cdot 10^{-9}$
x_3	0,362	0,696	3,739	$3{,}187 \cdot 10^{-4}$
x_4	0,114	0,802	1,106	0,272

Примечание. Толерантность вычисляется как разность 1 и коэффициента множественной детерминации между одной из факторных переменных и остальными независимыми переменными.

[9] Следует подчеркнуть, что алгоритм реализации излагаемого метода не требует ни обнаружения, ни проверки значимости мультиколлинеарности.

Из таблицы видно, что переменная x_4, соответствующая показателю толщины капсулы селезенки, практически не несет никакой дополнительной информации и может быть исключена из многофакторной регрессии. Исключение данной переменной сопровождалось уменьшением величины (M = 0,471) и снижением значимости ($\chi^2 = 57{,}886; p = 1{,}663 \cdot 10^{-12}$) показателя интенсивности мультиколлинеарности. Повторная проверка частных коэффициентов корреляции гестационного возраста с оставшимися после исключения x_4 гистометрическими показателями незначимых среди них не обнаружила ($p < 0{,}001$). Процедура исключения переменных из модели множественной регрессии на этом шаге останавливается.

Таким образом, наилучшая идентификация гестационного возраста может быть достигнута на основе использования трех гистометрических показателей селезенки: диаметра и плотности расположения лимфоидных узелков и толщины стенки центральных артерий. Исключение из регрессионной модели показателя толщины капсулы позволило уменьшить мультиколлинеарность, практически без потерь в точности прогнозирования результативного параметра ($\bar{r}^2 = 0{,}6799$ для четырехфакторной модели снизился лишь до $\bar{r}^2 = 0{,}6792$ - для трехфакторной). Необходимо заметить, что исключение показателя x_4 не привело к ликвидации мультиколлинеарности (M = 0,471). В этом смысле следует руководствоваться следующим принципом: слабая или умеренная мультиколлинеарность вообще не представляет собой ни аналитическую, ни прогностическую проблему; выраженная мультиколлинеарность представляет собой аналитическую проблему (т.е. искажает оценивание вклада каждой из независимых переменных в прогнозирование результативного показателя), но не влияет на точность прогноза; только лишь чрезвычайно сильная мультиколлинеарность всегда представляет собой проблему, приводя к неустойчивости и неправильности компьютерных вычислений [74].

В рассматриваемом примере имеет место умеренная мультиколлинеарность, которой можно пренебречь без риска уменьшения точности идентификации гестационного возраста и искажения истинных оценок диагностической значимости каждого из анализируемых гистометрических показателей.

Применительно к судебной антропологии чрезвычайно сильную мультиколлинеарность можно обнаружить лишь при подгонке по-

линомов высших порядков. Однако и в этом случае изложенный метод позволяет сделать объективный выбор полиномиальной модели с наиболее оптимальным набором преобразований независимой переменной. При анализе полиномиальных регрессий исключение переменных производится несколько иным способом: независимо от абсолютных значений и числа незначимых частных корреляций исключению подлежит только одна независимая переменная, соответствующая наивысшей степени идентифицируемого показателя, после чего анализ частных корреляций повторяется.

Например, предварительные этапы регрессионного анализа показали, что наилучшая идентификация гестационного возраста по степени кроветворной активности печени достигается с помощью полиномиальных регрессионных моделей (см. раздел 3.6). Однако недостатком данного вида линеаризирующих преобразований обычно является сильная мультиколлинеарность вследствие того, что все независимые переменные полиномиального уравнения фактически представлены одной переменной (в данном случае - показателем кроветворной активности). Закономерным следствием мультиколлинеарности является статистическая незначимость большинства коэффициентов регрессии и соответствующих коэффициентов частной корреляции. Так, модель идентификации, выраженная параболой 4-го порядка, характеризовалась незначимостью всех частных корреляций и, соответственно, коэффициентов регрессии всех степеней идентифицируемого показателя, больших 1 (табл. 13). Указанный результат вполне закономерен, поскольку каждая из этих переменных, в отдельности сильно коррелируя с гестационным возрастом, по сути, повторяет ту же информацию, которую объясняет базовая переменная – собственно показатель кроветворной активности печени.

Таблица 13

Критерии мультиколлинеарности регрессионной модели

$$\tilde{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$$

Переменная	Частный *r*	Толерантность	*t*	*p*
x	-0,386	0,003	-4,702	$6{,}659 \cdot 10^{-06}$
x^2	0,117	0,000	1,318	0,190
x^3	-0,020	0,000	-0,219	0,827
x^4	-0,021	0,001	-0,237	0,813

Это определило необходимость исключения из состава регрессионной модели независимой переменной, соответствующей наивысшей степени показателя кроветворной активности. Исключение указанной переменной привело к инверсии показателей статистической значимости оставшихся степеней базовой независимой переменной (табл. 14). В соответствии с алгоритмом метода процесс исключения переменных из состава полиномиальной регрессионной модели может быть остановлен, несмотря на то, что дальнейшее исключение привело бы к еще более заметному уменьшению мультиколлинеарности и росту частных корреляций (табл. 15).

Таблица 14

Критерии мультиколлинеарности регрессионной модели $\tilde{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

Переменная	Частный r	Толерантность	t	p
x	-0,676	0,015	-10,350	$1{,}399 \cdot 10^{-18}$
x^2	0,445	0,003	5,593	$1{,}309 \cdot 10^{-7}$
x^3	-0,310	0,009	-3,671	$3{,}550 \cdot 10^{-4}$

Таблица 15

Критерии мультиколлинеарности модели $\tilde{y} = \beta_0 + \beta_1 x + \beta_2 x^2$

Переменная	Частный r	Толерантность	t	p
x	-0,816	0,085	-15,961	$3{,}088 \cdot 10^{-32}$
x^2	0,650	0,085	9,673	$6{,}019 \cdot 10^{-17}$

Оценка критериев точности анализируемых полиномиальных регрессий показала адекватность изложенного метода сокращения избыточной информации и правильность выбора кубического полинома в качестве наилучшей регрессионной модели идентификации гестационного возраста (табл. 16).

Таблица 16

Показатели точности прогнозирования и мультиколлинеарности полиномиальных регрессионных моделей идентификации гестационного возраста

Модель	r	$\bar{r}^2$	s_ε	M	χ^2	p
Парабола 2-го порядка	0,911	0,827	2,388	0,306	316,668	$7{,}702 \cdot 10^{-71}$
Парабола 3-го порядка	0,920	0,842	2,279	0,444	912,988	$1{,}349 \cdot 10^{-197}$
Парабола 4-го порядка	0,920	0,841	2,288	0,513	1834,510	0

3.8. ГЕТЕРОСКЕДАСТИЧНОСТЬ

Наиболее распространенным в практике статистического оценивания параметров уравнений регрессии является метод наименьших квадратов. Этот метод основан на ряде предпосылок относительно результатов построения регрессионной модели. Одной из них является постоянство дисперсий остатков (гомоскедастичность). Однако в судебно-антропологических исследованиях зачастую приходится иметь дело с неоднородными данными, не обладающими таким свойством. Применение метода наименьших квадратов в этом случае может привести к такому нежелательному результату, как снижение эффективности оцениваемых параметров, что проявляется неадекватностью стандартных ошибок регрессионных коэффициентов. Это делает невозможным построение доверительных интервалов для прогнозных оценок регрессионных моделей.

В математической статистике существует несколько тестов, направленных на выявление неоднородности дисперсии остатков (гетероскедастичности). Во всех этих тестах проверяется нулевая гипотеза о равенстве остаточных дисперсий $H_0 = s_{\varepsilon 1}^2 = s_{\varepsilon 2}^2 = \ldots = s_{\varepsilon n}^2$, против альтернативной гипотезы $H_1 \neq s_{\varepsilon 1}^2 \neq s_{\varepsilon 2}^2 \neq \ldots \neq s_{\varepsilon n}^2$.

Одним из самых распространенных тестов обнаружения гетероскедастичности является тест Голдфелда-Квандта (цит. по [72]). Математическая модель данного метода предполагает, что остатки распределены нормально, автокорреляция отсутствует, а остаточное стандартное отклонение пропорционально значению независимой переменной.

В соответствии с алгоритмом метода Голдфелда-Квандта все n наблюдений в исследованной выборке упорядочиваются по величине x, после чего оцениваются отдельные регрессии для первых n_1 и последних n_2 наблюдений, средние $(n - n_1 - n_2)$ наблюдений отбрасываются. При наличии гетероскедастичности дисперсия в последних n_2 остатках будет больше, чем в первых n_1. На этом основана статистика

$$F_{\alpha;v_1=n_1-k-1;v_2=n_2-k-1} = \frac{\sum_{i=n-n_2}^{n} \varepsilon_i^2}{\sum_{i=1}^{n_1} \varepsilon_i^2},$$

где k – число независимых переменных в регрессионном уравнении. Мощность критерия зависит от выбора n_1 и n_2 по отношению к n. Рекомендуется следующее правило определения объемов выборок: $n_1 = n_2 = 0{,}37n$. При наличии в регрессионной модели нескольких независимых переменных упорядочивание должно производиться по той из них, которая, как предполагается, связана с s_ε.

Метод Голдфелда-Квандта может также использоваться для проверки на гетероскедастичность при предположении, что остаточная дисперсия обратно пропорциональна x. При этом используется статистика

$$F_{\alpha;v_1=n_1-k-1;v_2=n_2-k-1} = \frac{\sum_{i=1}^{n_1} \varepsilon_i^2}{\sum_{i=n-n_2}^{n} \varepsilon_i^2}.$$

Сравнение упорядоченных по величине x остатков можно также произвести по формуле $F_{\alpha;v_1=v_2=n-1} = s^2_{\varepsilon\max} / s^2_{\varepsilon\min}$ [30].

Для примера проверим на гетероскедастичность остатки регрессионной модели идентификации гестационного возраста по степени кроветворной активности печени, представленной кубическим полиномом $\hat{y} = 42{,}574 - 0{,}521x + 4{,}902 \cdot 10^{-3} x^2 - 1{,}597 \cdot 10^{-5} x^3$:

$$F_{\alpha=0{,}05;v_1=127;v_2=127} = \frac{373{,}654}{134{,}407} = 2{,}780;\ p = 8{,}917 \cdot 10^{-9}.$$

Таким образом, гетероскедастичность данной регрессионной модели доказана. Наличие гетероскедастичности ограничивает практическое использование указанной регрессионной модели только лишь определением точечных оценок гестационного возраста, прогнозирование же интервальных оценок данного идентифицируемого параметра является недоступным. Это объясняется тем, что доверительная область для прогнозных оценок регрессионной модели идентификации гестационного возраста будет излишне широкой на промежутке значений кроветворной активности печени 50-150 и чрезмерно узкой на промежутке 0-50, что соответствует прогнозным оценкам гестационного возраста 20-28 недель и 28-42 недели (рис. 25; 26). Наличие гетероскедастичности является скорее правилом, нежели исключением для регрессионного анализа любых биомедицинских данных и вызывает необходимость использования особых статистических методов.

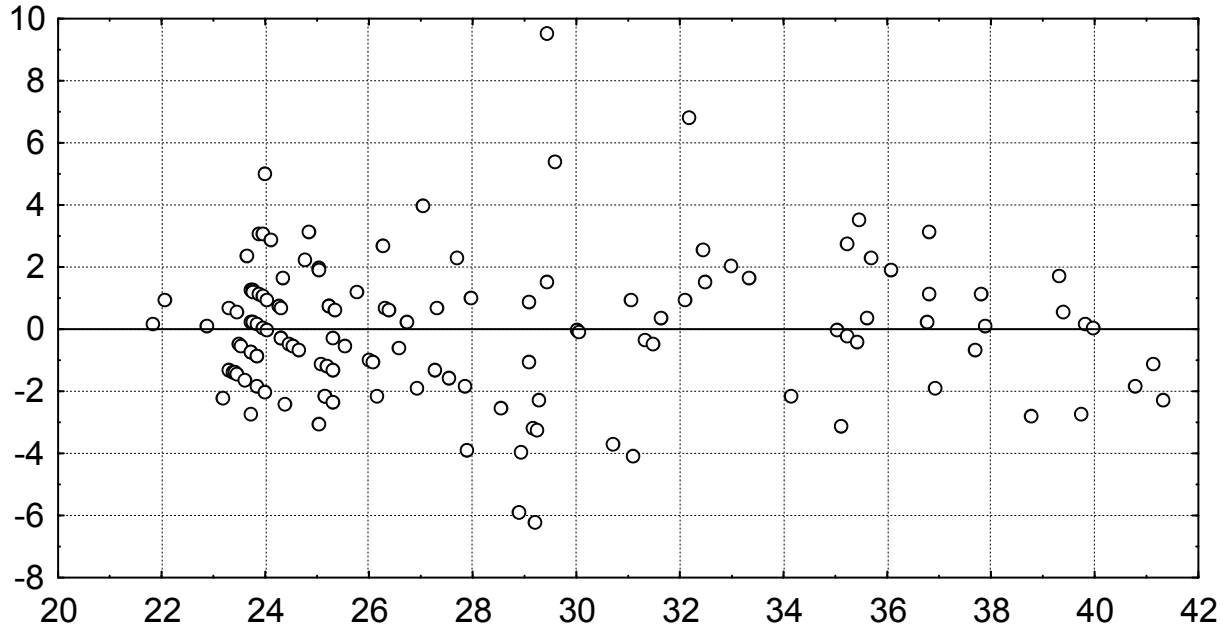

Рис. 25. Диаграмма рассеяния остатков регрессионной модели идентификации гестационного возраста по степени кроветворной активности фетальной печени $\hat{y} = 42{,}574 - 0{,}521x + 4{,}902 \cdot 10^{-3} x^2 - 1{,}597 \cdot 10^{-5} x^3$. По оси абсцисс – $\hat{y}$, недель; по оси ординат – остатки, недель.

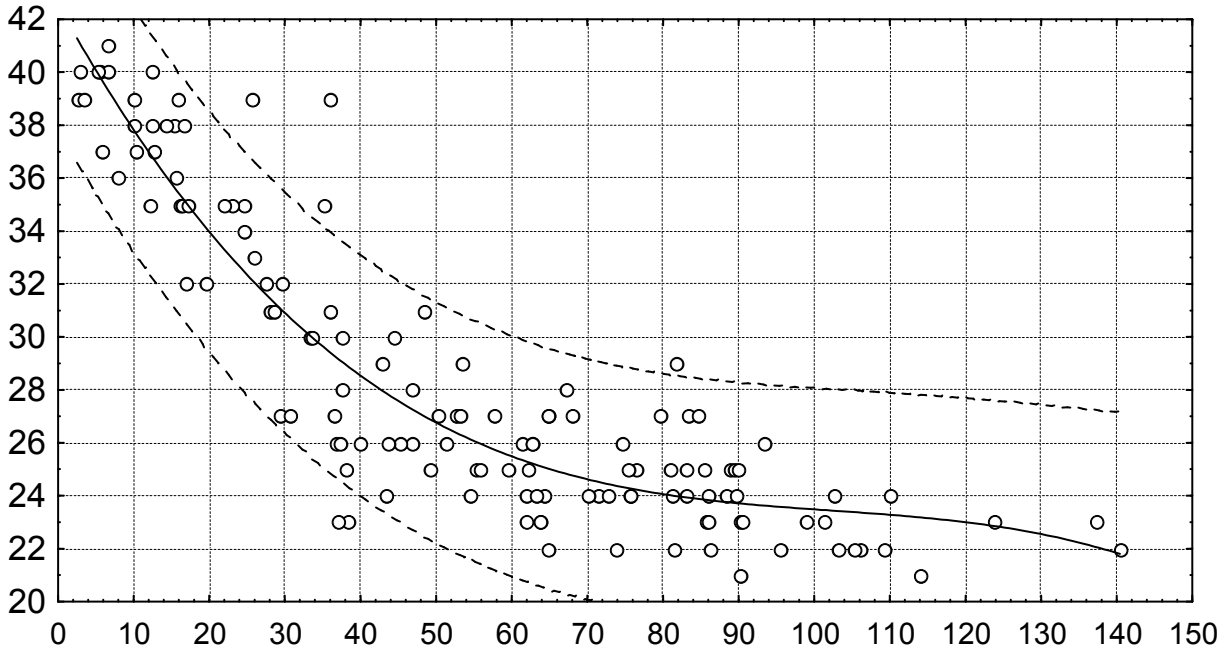

Рис. 26. Доверительная область для прогнозных оценок регрессионной модели $\hat{y} = 42{,}574 - 0{,}521x + 4{,}902 \cdot 10^{-3} x^2 - 1{,}597 \cdot 10^{-5} x^3$ идентификации гестационного возраста по степени кроветворной активности печени. По оси абсцисс – кроветворная активность с учетом усадки, число профилей ядер миелоидных клеток; по оси ординат – гестационный возраст, недель. На диаграмме маркирована выборочная совокупность наблюдений.

В теории регрессионного анализа любые отклонения наблюдаемых величин от прогнозных оценок рассматриваются как некоторые потери в точности прогнозирования. В традиционных методах множественной регрессии функция потерь определяется как сумма квадратов отклонений от предсказанных значений. Однако кроме этой можно рассмотреть и другие функции потерь, например, сумму абсолютных отклонений, которую, в частности, целесообразно использовать для уменьшения влияния выбросов [13]. В этой связи существует достаточно большое количество статистических методов, которые могут быть использованы для минимизации различных видов функций потерь.

Третьим по распространенности методом, в дополнение к методам наименьших квадратов и наименьших модулей отклонений, является метод взвешенных наименьших квадратов. Данный метод предполагает, что остатки независимы между собой, но имеют разные дисперсии. В этой связи при наличии гетероскедастичности отдается предпочтение именно методу взвешенных наименьших квадратов.

Основой метода взвешенных наименьших квадратов является регрессионная модель вида

$$Y = X\beta + \xi,$$

где $\sum(\xi) = \sigma^2\Omega$, а случайный вектор ξ имеет n-мерный закон распределения $\xi \in N_n(0;\sigma^2\Omega)$, Ω – известная положительно определенная матрица [26].

В теории доказано, что в этом случае несмещенные оценки вектора β и остаточной дисперсии σ^2 определяются как

$$b = (X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}Y;$$

$$s_\varepsilon^2 = \frac{1}{n-k-1}(Y-Xb)^T\Omega^{-1}(Y-Xb),$$

а остальные этапы анализа аналогичны таковым в классической линейной регрессии.

В методе взвешенных наименьших квадратов предполагают, что матрица Ω известна, так как на основании наблюдений Y ее определить невозможно. Выбор коэффициента взвешивания обычно основывается на предположении, что относительная ошибка σ_0 прогнозирования y_i постоянна, тогда стандартное отклонение пропорционально математическому ожиданию этой величины: $\sigma_i = \sigma_0\tilde{y}_i$. В

этом случае используют метод взвешенных наименьших квадратов с матрицей

$$\Omega^{\frac{1}{2}} = \sigma_0 \begin{pmatrix} \widehat{y}_1 & 0 & \cdots & 0 \\ & \widehat{y}_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \widehat{y}_n \end{pmatrix}, \qquad (31)$$

где $\sigma_0 = \frac{1}{n}\sum_{i=1}^{n}\frac{|\varepsilon_i|}{\widehat{y}_i}$.

Возможно также использование матрицы

$$\Omega = \begin{pmatrix} \varepsilon_1 & 0 & \cdots & 0 \\ & \varepsilon_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \varepsilon_n \end{pmatrix}. \qquad (32)$$

Иногда предполагают, что ошибка σ_i пропорциональна значению независимой переменной x. В этом случае применяют метод взвешенных наименьших квадратов с матрицей

$$\Omega = \begin{pmatrix} x_1 & 0 & \cdots & 0 \\ & x_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & x_n \end{pmatrix}. \qquad (33)$$

Применим метод взвешенных наименьших квадратов для построения кубической полиномиальной модели идентификации гестационного возраста по степени кроветворной активности печени (рис. 27). Использование матриц (31) и (32) привело к незначительному уменьшению гетероскедастичности (табл. 17).

Таблица 17

Оценки качества регрессионных моделей, полученных методом взвешенных наименьших квадратов с матрицей Ω

Ω	r^2	$\bar{r}^2$	s_ε	Гетероскедастичность	
				F	p
(31)	0,920	0,842	2,280	2,754	$1{,}190\cdot10^{-8}$
(32)	0,920	0,842	2,279	2,767	$1{,}031\cdot10^{-8}$
(33)	0,920	0,842	2,280	2,782	$8{,}716\cdot10^{-9}$
-*	0,920	0,842	2,279	2,780	$8{,}917\cdot10^{-9}$

Примечание. * - параметры стандартной регрессии.

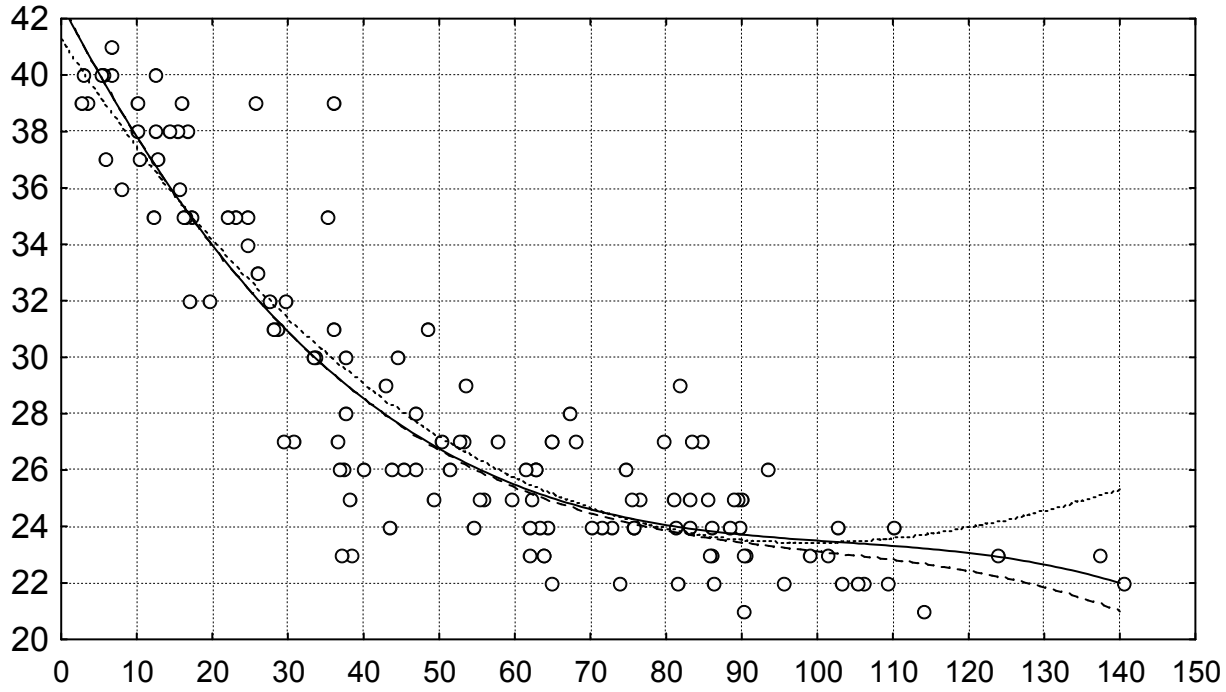

Рис. 27. Геометрическая интерпретация регрессионных моделей идентификации гестационного возраста по степени кроветворной активности печени, полученных методом взвешенных наименьших квадратов. По оси абсцисс – кроветворная активность с учетом усадки, число профилей ядер миелоидных клеток; по оси ординат – гестационный возраст, недель. Сплошной линией показана регрессионная модель, полученная с использованием матрицы (31), пунктиром – матрицы (32), точками – матрицы (33). На диаграмме маркирована выборочная совокупность наблюдений.

При этом уменьшение гетероскедастичности сопровождалось снижением точности прогнозирования. В этой связи необходимо отметить, что обычно уменьшение гетероскедастичности с помощью метода взвешенных наименьших квадратов достигается ценой увеличения дисперсии остатков на отрезке регрессионной линии с ранее небольшими ее значениями. Иногда уменьшение дисперсии остатков на отрезке регрессии с ранее большими ее значениями достигается за счет увеличения отклонений остатков со смещением их в одну сторону от линии регрессии.

Альтернативой аналитических методов регрессионного анализа являются численные методы минимизации функций потерь. Конкретные алгоритмы минимизации функций потерь могут использовать различные методы [72,115]. К настоящему времени разработано и исследовано большое число методов минимизации функций потерь. Наиболее известными из них, часто используемыми и реализованными в большинстве статистических и математических

программных продуктов являются семейства квазиньютоновских методов, симплекс-методов, методов Хука-Дживиса, Розенброка и Гессе [13,72].

Общим моментом всех методов оценивания является необходимость задания исследователем некоторых начальных значений, размера шагов и критерия сходимости алгоритма. Все методы начинают свою работу с особого набора предварительных оценок (начальных значений), которые в дальнейшем последовательно уточняются от итерации к итерации. При первой итерации размер шага определяет, как сильно будут меняться параметры. Наконец, критерий сходимости определяет, когда итерационный процесс можно прекратить. Например, процесс итераций можно остановить, когда изменение функции потерь на каждом шаге становится меньше определенной величины. Наиболее эффективным из приведенных алгоритмов является квазиньютоновский метод, основанный на вычислении первой и второй производной функции в различных точках и использовании эти данных для определения направления изменения параметров и минимизации функции потерь [115].

Применение некоторых из указанных алгоритмов (в частности, квазиньютоновского) для реализации методов наименьших квадратов или взвешенных наименьших квадратов приводит к результатам, аналогичным таковым для стандартного регрессионного анализа. Процедуры минимизации остальных функций потерь осуществимы только с помощью численных итерационных методов.

Учитывая, что рекомендуемый для коррекции неоднородности дисперсии остатков метод взвешенных наименьших квадратов очень трудоемок и часто не позволяет достичь желаемого результата, актуальной является проблема расчета доверительных интервалов для прогнозных оценок регрессионных моделей при наличии гетероскедастичности.

В связи с этим нами был разработан метод, позволяющий определять доверительные интервалы для прогнозных оценок регрессионных уравнений в указанной ситуации [51]. Математическое обоснование метода сводится к следующему.

Если регрессионное уравнение построено с помощью метода наименьших квадратов, то среднее остатков равно нулю. Поэтому при достаточно большом объеме выборки, выполнении условий о нормальности распределения остатков и однородности их диспер-

сии, доверительную область для прогнозных оценок регрессионных уравнений приближенно можно вычислить как

$$y_i \in \hat{y}_i \pm z_\alpha \cdot s_\varepsilon .$$

Погрешность в определении интервальных оценок $\hat{y}$ в данном случае может быть вызвана только неточностью выборочной оценки s_ε, которую можно ликвидировать, используя вместо s_ε ее одностороннюю верхнюю доверительную границу для любого требуемого уровня значимости, вычисляемую с помощью χ^2 - критерия.

Неадекватности интервальных оценок $\hat{y}$ при неоднородности дисперсии остатков можно избежать, если при их вычислении в каждой точке регрессионной линии использовать не общую для всех значений независимой переменной величину s_ε, а рассчитанную для каждого возможного значения x величину скользящего остаточного стандартного отклонения s_{ε_X} и его односторонней верхней доверительной границы. При таком подходе доверительные границы для прогнозных оценок регрессионных уравнений являлись бы «плавающими», расширяясь на отрезках регрессионной линии с большим s_ε и сужаясь на отрезках, где s_ε невелико.

В соответствии с алгоритмом метода остатки упорядочиваются по возрастанию соответствующих значений независимой переменной с формированием ряда $\varepsilon_1; \varepsilon_2; \ldots; \varepsilon_{\text{n-1}}; \varepsilon_{\text{n}}$. Затем, последовательно «скользя» вдоль ряда, определяется s_{ε_X} по k значениям остатков со сдвигом на одно значение. При этом для каждого значения независимой переменной вычисляется скользящее остаточное стандартное отклонение с формированием ряда

$$s_{\varepsilon_1} = \sqrt{\frac{\sum_{i=1}^{i=k} \varepsilon_i^2}{k-1}}; s_{\varepsilon_2} = \sqrt{\frac{\sum_{i=2}^{i=k+1} \varepsilon_i^2}{k-1}}; \ \ldots \ ; s_{\varepsilon_{n-k+1}} = \sqrt{\frac{\sum_{i=n-k}^{i=n} \varepsilon_i^2}{k-1}},$$

состоящего из $n-k+1$ значений s_{ε_X} (по $(k-1)/2$ значений s_{ε_X} теряется в начале и в конце упорядоченного ряда x_i). Далее для каждого значения s_{ε_X} вычисляется его односторонняя верхняя доверительная граница по формуле:

$$s_{\varepsilon_{XB}i} = s_{\varepsilon_X i} \cdot \sqrt{\frac{k-1}{\chi^2_{1-\alpha;k-1}}},$$

после чего формируются $n-k+1$ пар наблюдений x и соответствующих им значений $s_{\varepsilon_{XB}i}$.

В целях определения для любого возможного значения x_i величины $s_{\varepsilon_{XB}i}$ необходимо найти функцию зависимости $s_{\varepsilon_{XB}}$ от x. Для этого нужно добиться наиболее адекватного сглаживания значений $s_{\varepsilon_{XB}}$ любым из существующих методов. В случае успешной аппроксимации доверительную область для прогнозных оценок регрессионной модели можно определять по формуле:

$$y_i = \hat{y}_i \pm z_\alpha \cdot s_{\varepsilon_{XB}i}.$$

Из-за сложности формы искомой зависимости $s_{\varepsilon_{XB}}$ от x, зачастую наиболее адекватным методом подгонки полученных значений $s_{\varepsilon_{XB}i}$ является LOWESS – сглаживание, обеспечивающее расположение всех наблюдений $s_{\varepsilon_{XB}i}$ на аппроксимирующей кривой [95,96]. Недостатком LOWESS – сглаживания является необходимость графической визуализации зависимости $s_{\varepsilon_{XB}}$ от x вследствие невозможности ее представления в формульном виде. Немаловажным моментом для выполнения указанной процедуры является выбор величины k, который определяется правилом: чем меньше k, тем лучше $s_{\varepsilon_{XB}i}$ соответствует истинному значению s_ε в данной точке линии регрессии; чем больше k, тем меньше величина $s_{\varepsilon_{XB}i}$ и, соответственно, уже доверительный интервал для прогнозных оценок регрессионной модели. По мнению авторов метода при анализе выборок объемом более 100 наблюдений наиболее адекватным является значение k, соответствующее 21.

Изложенный метод расчета доверительных границ может быть использован и при наличии в составе регрессионной модели нескольких факторных переменных. В данном случае упорядочивание остатков производится по возрастанию прогнозных оценок зависимой переменной, после чего определяются ряды значений скользящего остаточного стандартного отклонения $s_{\varepsilon_{\hat{Y}}}$ и его односторонней верхней доверительной границы $s_{\varepsilon_{\hat{Y}B}}$. После успешной аппроксимации значений $s_{\varepsilon_{\hat{Y}B}}$ доверительная область для прогнозных оценок многофакторной регрессионной модели определяется по формуле:

$$y_i \in \hat{y} \pm z_\alpha \cdot s_{\varepsilon_{\hat{Y}B}}.$$

Метод скользящего остаточного стандартного отклонения был многократно использован нами при определении доверительных интервалов для прогнозных оценок разнообразных регрессионных моделей идентификации гестационного возраста плодов и новорожденных. Результатом явилось создание комплекса номограмм, позволяющих определять 50,60,70,80,90,95 и 99% интервальные оценки гестационного возраста [11,53]. Фрагмент одной из таких номограмм приведен на рисунке 28.

Таким образом, изложенный метод скользящего остаточного стандартного отклонения позволяет при наличии гетероскедастичности определять доверительные интервалы для прогнозных оценок любых (в том числе нелинейных и многофакторных) регрессионных уравнений. Данный метод более эффективен по сравнению с рекомендуемым при неоднородности дисперсии остатков методом взвешенных наименьших квадратов.

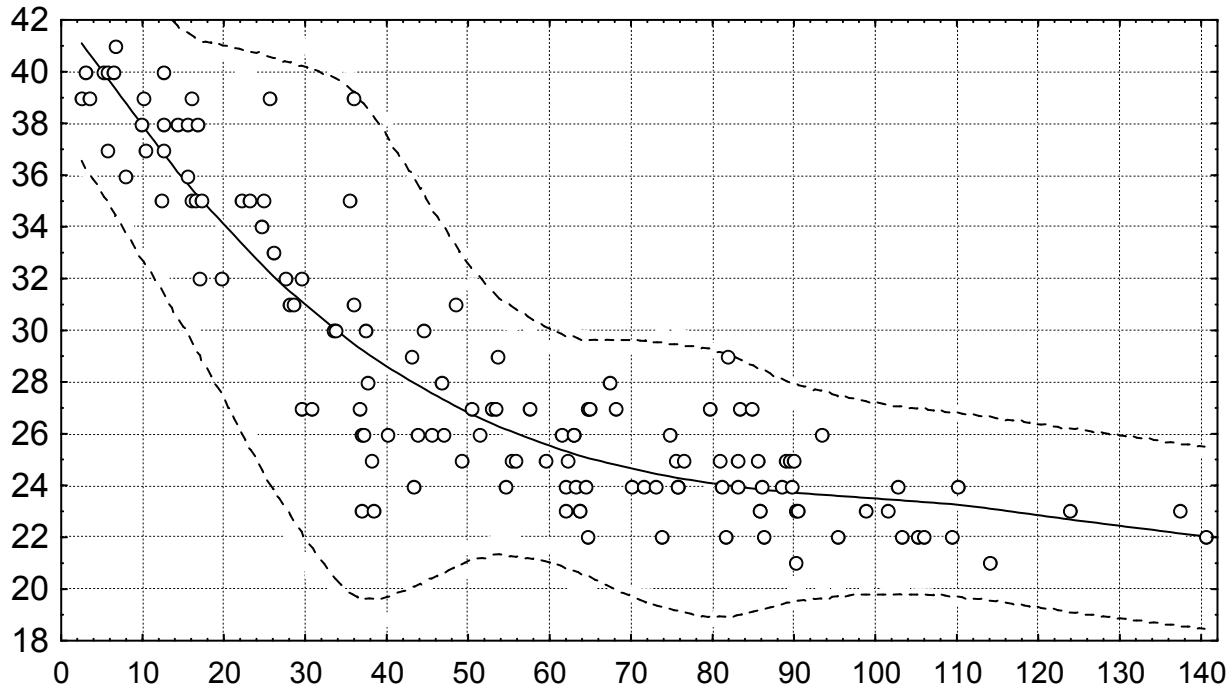

Рис. 28. Номограмма определения точечных и 95% интервальных оценок гестационного возраста плодов и новорожденных по степени кроветворной активности паренхимы печени. По оси абсцисс – кроветворная активность с учетом усадки, число профилей ядер миелоидных клеток; по оси ординат – гестационный возраст, недель. Сплошной линией показано множество точечных оценок, пунктирными линиями – множества 95% интервальных оценок гестационного возраста. На диаграмме маркирована выборочная совокупность наблюдений.

3.9. АВТОКОРРЕЛЯЦИЯ

Кроме гомоскедастичности второй предпосылкой адекватности метода наименьших квадратов относительно природы исследуемых данных и результатов регрессионной модели является отсутствие автокорреляции. Автокорреляцию (серийную корреляцию) можно определить как корреляционную зависимость между рядом $y_1, y_2, \ldots, y_n$ и этим же рядом, сдвинутым относительно первоначального положения на h моментов времени $y_{1+h}, y_{2+h}, \ldots, y_{h+n}$. С автокорреляцией обычно сталкиваются, когда данные, по которым строится регрессия, являются значениями временных рядов. Такой тип данных представлен в таблице 18.

Различают два вида автокорреляции:

1) автокорреляция в наблюдениях за переменными;

2) автокорреляция ошибок.

Автокорреляцию в наблюдениях за одной или более переменными измеряют при помощи нециклического коэффициента автокорреляции, который может рассчитываться между уровнями, сдвинутыми на любое число единиц времени. Это сдвиг, именуемый временным лагом, определяет и порядок коэффициентов автокорреляции: первого порядка (при $L = 1$), второго порядка (при $L = 2$) и т.д. Считается, что наиболее сильные искажения результатов анализа возникают при корреляции между исходными уровнями и теми же уровнями, сдвинутыми на одну единицу времени [72]. В этом случае коэффициент автокорреляции можно определить как корреляцию между рядами $y_1, y_2, \ldots, y_{n-1}$ и $y_2, y_3, \ldots, y_n$. Достоверность коэффициента автокорреляции проверяют с помощью тех же методов, что и для обычного коэффициента парной корреляции.

Таблица 18

Данные временных рядов для множественной регрессии [78]

Периоды времени (t)	y	x_1	x_2	…	x_k
1	y_1	x_{11}	x_{21}	⋯	x_{k1}
2	y_2	x_{12}	x_{22}	⋯	x_{k2}
⋮	⋮	⋮	⋮	⋮	⋮
n	y_n	x_{1n}	x_{2n}	⋯	x_{kn}

Однако несоответствие условиям применимости методов классического регрессионного анализа возникает при другом виде автокорреляции - серийной корреляции остатков. В этом случае нарушается одно из основных допущений о свойствах остатков ε_i регрессионной модели – свойство их статистической независимости между собой. При наличии серийной корреляции остатков искажаются стандартные ошибки коэффициентов регрессии, и все критерии существенности для уравнений регрессии не применимы даже для приблизительных расчетов.

Поскольку теоретические значения ошибок ε_i являются неизвестными, на практике исследуются их аналоги – остатки e_i. Коэффициент автокорреляции первого порядка, имеющий наибольшее практическое значение, рассчитывается по формуле:

$$r_e = \frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sqrt{\sum_{i=2}^{n} e_i^2 \cdot \sum_{i=1}^{n-1} e_i^2}} \approx \frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=1}^{n} e_i^2}$$

Для обнаружения автокорреляции остатков используется критерий Дарбина-Уотсона (DW):

$$DW = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \text{ [107].}$$

В теории доказано, что статистику Дарбина-Уотсона можно выразить через коэффициент автокорреляции:

$$DW \approx 2(1 - r_e).$$

При полной независимости остатков $r_e = 0$, а $DW = 2$. При положительной серийной корреляции $r_e \to +1$, а $0 \leftarrow DW$. При отрицательной серийной корреляции $r_e \to -1$, а DW возрастает до 4. Дарбин и Уотсон доказали, что существуют две границы, обозначаемые d_U и d_L, которые зависят лишь от объема выборки n, числа k независимых переменных и уровня значимости α [106,107]. Значения этих границ табулированы, поэтому результаты теста Дарбина-Уотсона можно представить в виде следующей таблицы (табл. 19).

Обычно считается, что автокорреляция имеет место практически только при изучении временных рядов и характерна преимущественно для анализа экономических процессов [72,74,78]. Примени-

тельно к судебно-медицинским антропологическим исследованиям автокорреляцию можно ожидать в случаях, когда изучаемые биометрические показатели регистрируются на объектах различной давности, когда стохастическое изменение исследуемых показателей обусловлено процессами акселерации, произошедшими со времени формирования объектов научного поиска. Указанное замечание может быть существенным, поскольку материалом многих современных способов идентификации служили разнородные костные коллекции: например, коллекция кафедры антропологии МГУ давностью 50 лет, дополненная скелетами из захоронений XVII – XVIII веков [42], биометрические данные R. Povilaitis, дополненные серией скелетов кафедры антропологии МГУ [38].

Изложенное делает целесообразными включение в исследование объектов схожей давности, контроль времени формирования объектов и итоговое тестирование наличия серийной корреляции остатков. Например, сбор исходных данных для разработки способов идентификации гестационного возраста был произведен на коротком отрезке времени длиной менее двух лет. Результатом явилось отсутствие серийной корреляции остатков. Так, для регрессионной модели идентификации гестационного возраста по кроветворной активности печени $r_e = 0{,}090;\ DW = 1{,}703;\ p > 0{,}05$. Резюмируя, следует отметить, что наличие автокорреляции в судебно-антропологических исследованиях можно вообще не контролировать, если объекты исследования характеризуются схожей давностью. Примером таких данных является сводка соматометрических характеристик шведов, умерших в 1971 г. [156].

Таблица 19

Интерпретация результатов DW –теста [72]

Значение DW	Вывод
$4 - d_L < DW < 4$	Есть отрицательная автокорреляция
$4 - d_U < DW < 4 - d_L$	Неопределенность
$2 < DW < 4 - d_U$	Автокорреляция отсутствует
$d_U < DW < 2$	Автокорреляция отсутствует
$d_L < DW < d_U$	Неопределенность
$0 < DW < d_L$	Есть положительная автокорреляция

3.10. СРАВНИТЕЛЬНЫЙ АНАЛИЗ РЕГРЕССИЙ

Необходимость сравнения двух или более линий регрессии в судебно-медицинских антропологических исследованиях встречается редко. Вместе с тем указанный аспект регрессионного анализа может быть источником важной информации о характере изучаемых зависимостей. Основным показанием к использованию сравнительного анализа регрессий в судебно-медицинской антропологии чаще всего является задача обнаружения кластеринга в исследуемых данных. Иногда указанная группа статистических методов может быть использована для решения обратной задачи – доказательства отсутствия выраженного кластеринга в исследуемой совокупности данных и возможности проведения регрессионного анализа для объединенной выборки. Например, при разработке регрессионных моделей идентификации гестационного возраста плодов и новорожденных помимо уже упомянутых причин кластеринга нами исследовалась также возможность влияния на степень кроветворной активности фетальной печени таких патологических процессов, как хроническая плацентарная недостаточность и внутриутробное инфицирование [55]. Проведенное исследование подтвердило большую выраженность экстрамедуллярного гемопоэза у плодов с названной патологией. Однако отсутствие выраженных различий между линиями регрессии в группах плодов и новорожденных от физиологически протекающей беременности и беременности с указанной патологией сделало возможным проведение регрессионного анализа для объединенной совокупности наблюдений [55].

Методы сравнения регрессий различаются в зависимости от количества сравниваемых линий, формы регрессионных зависимостей и количества переменных, входящих в состав регрессионных уравнений.

Рассмотрим наиболее простой случай – сравнение двух однофакторных линейных регрессий. Данную задачу можно выполнить тремя методами: сравнением коэффициентов сдвига b_0, сравнением коэффициентов наклона b_1 и сравнением линий регрессии в целом [16].

Сравнение коэффициентов двух однофакторных линейных регрессий производится однотипно по формуле:

$$t_{\alpha;\nu=n_1+n_2-4} = \frac{b_1 - b_2}{s_{b_1-b_2}},$$

где стандартная ошибка разности вычисляется как

$$s_{b_1-b_2} = \sqrt{s_{b_1}^2 + s_{b_2}^2},$$

если обе регрессии оценены по одинаковому числу наблюдений. В противном случае необходимо предварительно определить объединенную оценку остаточной дисперсии:

$$s_\varepsilon^2 = \frac{(n_1-2)s_{\varepsilon_1}^2 + (n_2-2)s_{\varepsilon_2}^2}{n_1+n_2-4}.$$

Тогда стандартная ошибка разности для коэффициента наклона b_1 вычисляется с помощью выражения

$$s_{b_1-b_2} = s_\varepsilon \sqrt{\frac{1}{(n_1-1)s_{x_1}^2} + \frac{1}{(n_2-1)s_{x_2}^2}}.$$

Аналогичная формула для стандартной ошибки коэффициента сдвига b_0 имеет вид:

$$s_{b_1-b_2} = s_\varepsilon \sqrt{\frac{1}{n_1} + \frac{\bar{x}_1^2}{(n_1-1)s_{x_1}^2} + \frac{1}{n_2} + \frac{\bar{x}_2^2}{(n_2-1)s_{x_2}^2}}.$$

Изложенные методы пригодны лишь для сравнений двух однофакторных линейных регрессий. Если вместо статистик $\bar{x}$ и s_x использовать соответствующие линеаризирующие преобразования, то методы сравнений регрессионных коэффициентов могут быть легко преобразованы для сравнений однофакторных нелинейных регрессий.

Сравнение линий регрессии в целом позволяет осуществлять сравнительный анализ многофакторных и нелинейных регрессий с различным количеством независимых переменных в их составе. Пусть сравнению подлежат m регрессий, каждая из которых содержит k_i независимых переменных. Тогда алгоритм метода включает выполнение следующих этапов.

1. Производится подгонка регрессии для каждой k-мерной выборки из общего их числа m.

2. По остаточным дисперсиям $s_{\varepsilon i}^2$ каждой из регрессий вычисляется объединенная оценка остаточной дисперсии

$$s_\varepsilon^2 = \frac{\sum_{i=1}^{m}(n_i-k_i-1)s_{\varepsilon i}^2}{\sum_{i=1}^{m}(n_i-k_i-1)}.$$

3. Осуществляются объединение выборок, подгонка регрессии, включающей K независимых переменных, для единой выборки и вычисление ее остаточной дисперсии s_{Σ}^2.

4. Вычисляется эффект использования раздельных регрессий, мерой которого служит величина:

$$s_{\Delta\varepsilon}^2 = \frac{s_{\Sigma}^2 \sum_{i=1}^{m}(n_i - K - 1) - s_{\varepsilon}^2 \sum_{i=1}^{m}(n_i - k_i - 1)}{m}.$$

5. Вычисляется статистика

$$F = \frac{s_{\Delta\varepsilon}^2}{s_{\varepsilon}^2}.$$

6. Полученная статистика сравнивается с критическим значением F-критерия при требуемом уровне значимости α и числе степеней свободы $v_1 = m$ и $v_2 = \sum_{i=1}^{m}(n_i - k_i - 1)$. Если полученное значение F-критерия больше критического, то гипотеза о совпадении линий регрессии должна быть отклонена.

Для примера продолжим сравнительный анализ возрастной динамики объема головного мозга у мужчин и у женщин (см. раздел 2.7). Напомним, что в рамках данного исследования было произведено измерение объема головного мозга от трупов 61 мужчины и 32 женщин, умерших в возрасте 18-92 лет. Сравнительный анализ корреляций значимых межполовых различий в скорости развития церебральной атрофии не обнаружил. Попытаемся выявить указанные различия методами сравнения регрессий.

Визуальное изучение диаграмм рассеяния и сравнение остаточных дисперсий альтернативных регрессионных моделей возрастной динамики объема головного мозга показало, что исследуемые зависимости по форме являются линейными и могут быть представлены следующими аналитическими выражениями:

$$V_1 = 1428{,}199 - 3{,}178x_1 \pm 134{,}410 \cdot t_{\alpha;59} \cdot \sqrt{1{,}016 + \frac{(x_1 - 49{,}867)^2}{13881{,}014}} \text{ и}$$

$$V_2 = 1264{,}460 - 2{,}747x_2 \pm 101{,}407 \cdot t_{\alpha;30} \cdot \sqrt{1{,}031 + \frac{(x_2 - 56{,}382)^2}{8598{,}286}},$$

где V_M – объем ГМ у мужчин, мл; V_W – объем ГМ у женщин, мл; x – возраст, лет; t – значение двустороннего варианта критерия Стью-

дента при требуемом уровне значимости α и указанном количестве степеней свободы.

Сравним коэффициенты наклона указанных регрессий. Результаты промежуточных вычислений и некоторые параметры регрессий приведены в таблице 20.

Таблица 20

Параметры регрессий и выборочных совокупностей

Группа	n	r	r^2	s_ε, мл	$\bar{x}$, лет	s_X, лет
Мужчины	61	0,341	0,116	134,410	56,4	16,7
Женщины	32	0,417	0,174	101,407	49,7	15,2

Вычислим объединенную оценку остаточной дисперсии:

$$s_\varepsilon^2 = \frac{(n_1-2)s_{\varepsilon 1}^2 + (n_2-2)s_{\varepsilon 2}^2}{n_1+n_2-4} = 15442{,}611.$$

Отсюда стандартные ошибки разности равны

$$s_{b_1-b_2} = 124{,}268\sqrt{\frac{1}{(61-1)\cdot 16{,}7^2}+\frac{1}{(32-1)\cdot 15{,}2}} = 1{,}705 \text{ для } b_1 \text{ и}$$

$$s_{b_1-b_2} = 124{,}268\sqrt{\frac{1}{61}+\frac{56{,}4^2}{60\cdot 16{,}7^2}+\frac{1}{32}+\frac{49{,}7^2}{31\cdot 15{,}2^2}} = 95{,}977 \text{ для } b_0.$$

В итоге получаем

$$t_{b_1} = \frac{3{,}178-2{,}747}{1{,}705} = 0{,}252;\ p = 0{,}801 \text{ и}$$

$$t_{b_0} = \frac{1428{,}2-1264{,}5}{95{,}977} = 1{,}706;\ p = 0{,}091.$$

Таким образом, сравнительный анализ не выявил различий между угловыми коэффициентами регрессий. Вместе с тем, при одностороннем варианте t-критерия обнаружены различия по коэффициентам сдвига регрессий $(p = 0{,}046)$. Полученные данные свидетельствуют об отсутствии межполовых различий в скорости церебральной инволюции, что подтверждается результатами сравнительного анализа коэффициентов корреляции (см. раздел 2.7), и наличии межполовых различий между средними значениями объема головного мозга. Геометрическая интерпретация полученных выводов приведена на рисунке 29.

Для сравнения регрессий в целом получим подгонку объединенной выборки, остаточная дисперсия которой равна $s^2_{\Sigma} = 19496{,}322$, и рассчитаем величину эффекта раздельных регрессий

$$s^2_{\Delta\varepsilon} = \frac{(61+32-2)\cdot 19496{,}322 - (61+32-4)\cdot 15442{,}611}{2} = 199886{,}487.$$

Отсюда

$$F = \frac{199886{,}487}{15442{,}611} = 12{,}944;\ p = 1{,}163\cdot 10^{-5}.$$

Таким образом, сравнение регрессий в целом доказало наличие межполовых различий в возрастной динамике объема головного мозга на протяжении 18-92 лет. Указанные различия обусловлены лишь разностью объемов головного мозга у мужчин и у женщин в начале исследуемого возрастного периода (иными словами, разностью ординат точек пересечения линий регрессии с осью ординат) (см. рис. 29). Скорость же возрастной инволюции головного мозга от пола индивида не зависит.

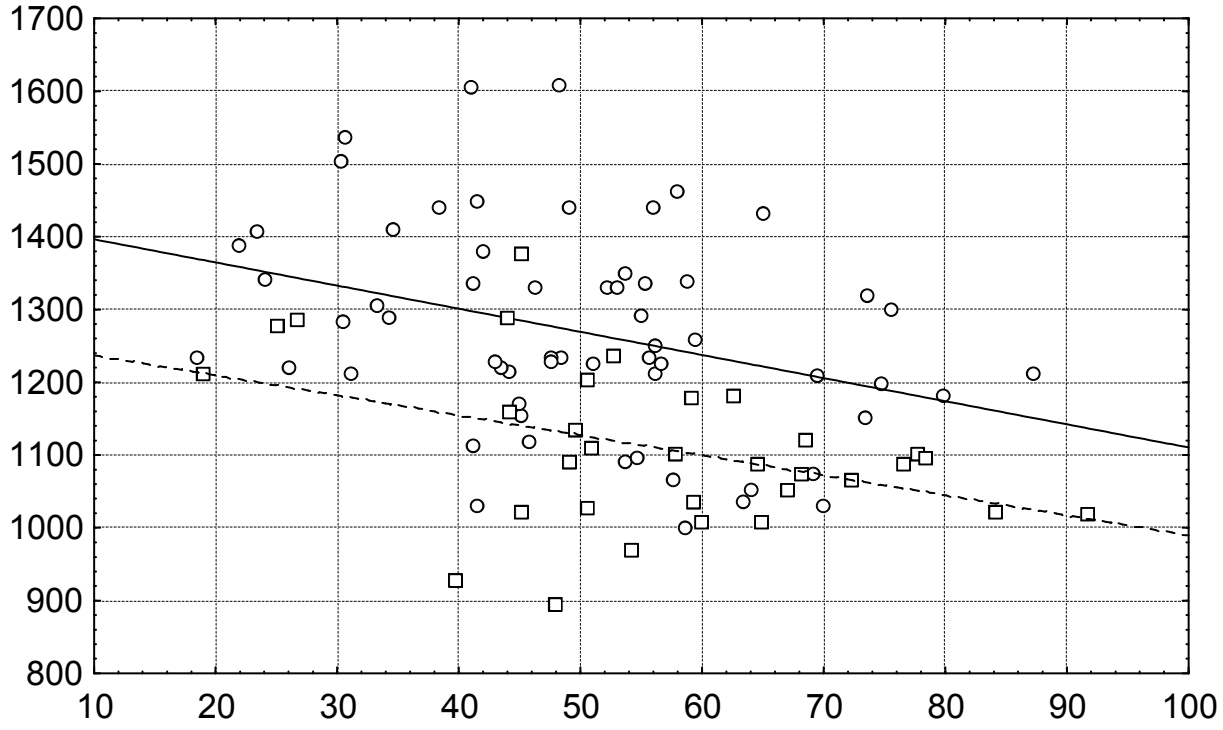

Рис. 29. Геометрическая интерпретация возрастной динамики объема головного мозга у мужчин и у женщин. По оси абсцисс – возраст, лет; по оси ординат – объем головного мозга, мл. ○ – выборка мужчин; □ – выборка женщин.

3.11. РЕГРЕССИИ С ИНДИКАТОРНЫМИ ПЕРЕМЕННЫМИ

Математическая модель регрессионного анализа предусматривает количественный тип данных для всех переменных, входящих в состав уравнения регрессии. Однако регрессионный анализ применим и в тех случаях, когда одна или несколько из независимых переменных являются качественными или порядковыми[10]. В судебно-медицинских антропологических исследованиях такая ситуация встречается, когда исходная совокупность данных обнаруживает выраженную неоднородность, чаще всего обусловленную половой принадлежностью идентифицируемых объектов.

В этих случаях в состав регрессионного уравнения вводят дополнительную индикаторную (фиктивную) переменную. Таким термином в регрессионном анализе называется переменная, предназначенная для представления качественных или порядковых данных. В судебной антропологии индикаторные переменные обычно используются для кодирования половой принадлежности идентифицируемых объектов [31,36-38]. В этом случае индикаторная переменная представляет качественный показатель и может принимать только одно из двух любых последовательных неотрицательных целых чисел. В математической статистике подобные бинарные (дихотомические) переменные рекомендуется кодировать числами 0 и 1 [74,85]. Между тем в отечественных судебно-медицинских антропологических исследованиях сложилась практика кодирования бинарных переменных числами 1 и 2 [31,36-38,41]. Хотя для точности идентификации выбор варианта кодирования значения не имеет, обозначение категорий индикаторной переменной числами, отличными от 0 и 1, лишает соответствующий коэффициент регрессии содержательной смысловой нагрузки.

Применим изложенный метод для аналитического описания возрастной динамики объема головного мозга человека, выраженного одной формулой. Для этого необходимо включить в состав уравнения регрессии кроме переменной, представляющей возраст, вторую переменную, представляющую качественный признак – пол, принимающую значения: 1 - для мужчин и 0 – для женщин.

---

[10] Ситуации, когда качественный или порядковый тип данных имеет зависимая переменная, рассматриваются при анализе методов классификации. Кроме них существуют, но пока не нашли своего применения в судебной антропологии разнообразные методы логистической регрессии.

Итоговое уравнение регрессии имеет вид:

$$V = 1279{,}445 - 3{,}013 \cdot Age + 140{,}544 \cdot Sex,$$

где V – объем головного мозга, мл; Age – возраст, лет; Sex – пол, $\in [0;1]$. Указанное регрессионное уравнение статистически значимо в целом (F = 24,296; p = $3{,}653 \cdot 10^{-9}$) и содержит значимые коэффициенты регрессии для переменной Age (t = -3,654; $p = 4{,}331 \cdot 10^{-4}$) и переменной Sex (t = 5,108; $p = 1{,}807 \cdot 10^{-6}$).

В данном примере значение коэффициента регрессии для индикаторной переменной означает, что средняя разность объема головного мозга у мужчин и женщин составляет 140,544 мл. Учитывая абсолютные значения t-статистик для коэффициентов регрессии, можно утверждать, что для прогнозирования объема головного мозга большее значение имеет половая принадлежность субъекта, чем его возраст. Об этом же свидетельствуют значения стандартизованных коэффициентов регрессии (0,442 – для переменной Sex и -0,317 - для переменной Age).

В регрессионном анализе можно использовать как одну, так и несколько индикаторных переменных. Более сложной процедурой отличается регрессионный анализ, если независимая переменная включает более двух категорий, т.е. является порядковым показателем. В судебно-медицинской антропологии подобные индикаторные переменные, как правило, используются для кодирования типа телосложения [41] и интервалов паспортного возраста [42]. Так же, как и в случае с качественными индикаторными переменными, способы кодирования порядковых переменных, принятые в отечественной судебно-медицинской антропологии, отличаются от методов, рекомендуемых в математической статистике.

Указанное отличие можно пояснить на примере регрессионной модели диагностики длины тела по подъязычной кости [41]. Данная модель помимо 9 остеометрических показателей включает также 2 индикаторных переменных, одна из которых кодирует половую принадлежность, а вторая – тип телосложения индивида. С позиции регрессионного анализа индикаторная переменная, кодирующая пол, является обычной бинарной переменной. В отличие от нее вторая индикаторная переменная кодирует одну из трех возможных категорий (типов телосложения): 1 – долихоморфный; 2 – мезоморфный; 3 – брахиморфный. В этом случае согласно модели множественной регрессии статистическая значимость коэффициента регрессии, соответствующего индикаторной переменной, коди-

рующей тип телосложения, означает наличие значимого вклада данной переменной (тип телосложения) для прогнозирования идентифицируемого параметра (длины тела). Вместе с тем само значение данного коэффициента регрессии не имеет содержательной смысловой нагрузки, и его интерпретация затруднительна.

В этой связи специалисты в области математической статистики рекомендуют другой метод кодирования порядковых биометрических показателей [74]. В соответствии с алгоритмом метода вначале нужно выбрать одну из категорий признака, которая будет служить в качестве базового значения, по отношению к которому будет измеряться влияние всех других категорий. Эта категория будет представлена в уравнении регрессии коэффициентом b_0. Индикаторные переменные создаются лишь для категорий переменной, отличных от базовой. Каждая созданная индикаторная переменная также кодируется лишь двумя числами: 0 и 1. В качестве базовой рекомендуется выбирать категорию признака, которая встречается чаще других. При таком подходе количество индикаторных переменных, используемых во множественной регрессии для замены переменной порядкового типа, будет на одну меньше количества категорий.

После замены качественных показателей индикаторными переменными множественная регрессия выполняется стандартным способом. В этом случае коэффициент регрессии, соответствующий индикаторной переменной, обозначает среднюю разницу значений прогнозируемого параметра в двух категориях – той, которую представляет данная индикаторная переменная и базовой категорией (остальные независимые переменные при этом остаются неизменными). Если коэффициент регрессии является положительным числом, то эта категория, характеризуется большим средним значением прогнозируемого параметра по сравнению с базовой категорией. Если коэффициент регрессии - отрицательное число, то среднее значение прогнозируемого параметра для этой категории оказывается меньше, чем для базовой. Значимость коэффициентов регрессии каждой индикаторной переменной проверяется обычным способом и означает наличие неслучайной разницы в значениях прогнозируемого параметра между категорией, которую представляет индикаторная переменная и базой.

Допустим, что в качестве базовой категории выбран мезоморфный тип телосложения. Тогда итоговая регрессионная модель диаг-

ностики длины тела по остеометрическим показателям подъязычной кости и показателям пода и типа телосложения должна включать не 11, а 12 переменных: 9 – остеометрические показатели; 1 – половая принадлежность; 1 – долихоморфный тип телосложения; 1 – брахиморфный тип телосложения (табл. 21).

Существенным недостатком регрессионного анализа является использование индикаторных переменных для представления количественных признаков. Так, кодирование интервалов паспортного возраста, примененное в работе [42], следует признать нецелесообразным, поскольку возраст является количественным признаком и должен быть представлен соответствующей переменной.

Резюмируя, следует заключить, что использование индикаторных переменных является простейшим методом исследования неоднородных регрессий, поскольку позволяет получить отдельный для каждой категории коэффициент регрессии, но одни и те же регрессионные коэффициенты для остальных независимых переменных. Более прогрессивный подход к регрессионному анализу многомерной совокупности данных, включающей качественные или порядковые переменные, заключается в выполнении отдельных регрессий для каждой категории этих переменных. В отличие от индикаторных переменных раздельный регрессионный анализ позволяет получить более гибкую модель с различными коэффициентами регрессии для каждой из независимых переменных и каждой категории качественной переменной [74]. Так, подобные анализы были выполнены при описании возрастной динамики объема головного мозга отдельно для мужчин и женщин (см. раздел 3.10).

Таким образом, поскольку математическая модель регрессионного анализа предполагает соответствие всех переменных количественному типу, то включение в регрессионное уравнение индикаторных переменных является нецелесообразным.

Таблица 21

Кодирование двух индикаторных переменных для представления трех категорий порядкового признака (тип телосложения), исключая мезоморфный тип как базовую категорию

Телосложение идентифицируемого субъекта	Индикаторная переменная (тип телосложения)	
	Долихоморфный	Брахиморфный
Долихоморфный	1	0
Мезоморфный	0	0
Брахиморфный	0	1

3.12. ОПТИМИЗАЦИЯ ПОДБОРА ПЕРЕМЕННЫХ В СОСТАВ МНОГОФАКТОРНОЙ РЕГРЕССИОННОЙ МОДЕЛИ

Важнейшей проблемой множественного регрессионного анализа во всех научно-практических приложениях, в том числе и в судебно-медицинской антропологии, является определение оптимального набора независимых переменных, входящих в состав регрессионной модели. В этой связи нами было проведено исследование, посвященное изучению эпидемиологии методов оптимизации состава множественных регрессионных моделей в научных работах в области судебно-медицинской антропологической идентификации. Объектами анализа явились 9 случайно отобранных статей, отображающих результаты оригинальных исследований, посвященных разработке способов прогнозирования идентифицируемых параметров, из числа работ, опубликованных в журнале «Судебно-медицинская экспертиза» за период 1997-2006 гг.

Проведенный анализ показал, что самыми распространенными в судебно-медицинских антропологических исследованиях критериями оптимальности состава множественных регрессионных моделей являются коэффициент множественной корреляции и/или детерминации, а также остаточное стандартное отклонение (табл. 22). В связи с этим уместно напомнить, что при добавлении в состав регрессионного уравнения дополнительной независимой переменной величина коэффициента множественной детерминации никогда не уменьшается, а может только увеличиться, в том числе и при отсутствии корреляции добавленной переменной с результативным показателем. Поэтому в судебно-медицинской литературе неоднократно подчеркивалось, что значения коэффициентов r и r^2 не могут быть критериями, полностью определяющими оптимальность набора идентифицирующих признаков, входящих в состав уравнения множественной регрессии [см. напр. 59].

Таблица 22

Критерии качества многофакторных регрессий, использованные в судебно-медицинских антропологических исследованиях

Количество исследований	r или r^2	$\bar{r}^2$	s_ε	F	t	M
Абсолютное число	9	1	9	1	1	3
95% верхняя оценка доли, %	66	0	66	0	0	7
Точечная оценка доли, %	100	11	100	11	11	33
95% нижняя оценка доли, %	100	48	100	48	48	70

Остаточное стандартное отклонение также не может являться достаточным критерием оптимальности состава уравнения множественной регрессии, поскольку самые низкие значения дисперсии остатков обычно соответствуют наиболее громоздкой регрессионной модели с большим количеством взаимно коррелирующих предикторов и высоким показателем интенсивности мультиколлинеарности. Между тем интенсивность мультиколлинеарности определялась лишь в 33% (7) судебно-антропологических исследований.

Изложенное доказывает необходимость дополнения упомянутых общепринятых критериев оптимальности множественных регрессий группой дополнительных критериев. На наш взгляд, такими критериями должны являться скорректированный коэффициент множественной детерминации, *F*-тест значимости регрессионного уравнения в целом и *t*-тесты значимости отдельных коэффициентов регрессии. Особенно важными следует считать *t*-статистики регрессионных коэффициентов, подтверждающие наличие значимого влияния на результативный показатель данной независимой переменной.

Еще более важной является проблема разработки стандартного алгоритма подбора независимых переменных в состав уравнения множественной регрессии, оптимальность которого во многом определяет диагностическую значимость создаваемых регрессионных моделей идентификации личности. Несмотря на это, теоретическим вопросам подбора переменных в состав многофакторных регрессионных моделей идентификации личности посвящены лишь единичные работы [38,60]. Благодаря работам указанных и ряда других авторов в судебной антропологии преимущественно используются следующие правила проведения регрессионного анализа и включения независимых переменных в состав уравнений множественной регрессии:

- планирование объема исследуемой совокупности не менее 50-100 наблюдений;
- приближенное соответствие распределений идентифицирующих признаков нормальному закону;
- наличие высоких парных корреляционных связей каждого идентифицирующего показателя с идентифицируемым параметром;
- отсутствие сильных ($r \leq 0{,}5$) взаимосвязей между самими идентифицирующими признаками.

Приведенный перечень правил вполне достаточен для анализа линейных взаимосвязей между идентифицируемым параметром и идентифицирующими признаками, однако является неадекватным при нелинейном характере указанных зависимостей и неоднородности исследуемых данных. Между тем, именно такие свойства характерны для большинства наборов данных, используемых при проведении судебно-медицинских антропологических исследований. Поэтому закономерным является появление интереса к обсуждению излагаемой проблемы в современной судебно-медицинской литературе [8,28,53]. Вместе с тем системного изучения теоретических основ оптимизации подбора идентифицирующих признаков в состав регрессионных моделей идентификации личности до настоящего времени не проводилось. Данную ситуацию усугубляет также тот факт, что проблема определения компактного перечня независимых переменных, обеспечивающих качественное прогнозирование результативного показателя, полностью пока не решена даже в теории математической статистики.

В настоящее время существует два альтернативных подхода к определению набора независимых переменных для регрессионных моделей [74]. Один из них заключается в классификации перечня независимых переменных по приоритетам, причем главным правилом классификации является волевое решение исследователя. Второй подход объединяет большую группу автоматизированных методов выбора переменных. Большинство из них основано на алгоритмах пошагового включения или исключения независимых переменных, которым также присуща доля субъективности. Наилучшим методом автоматического выбора переменных в настоящее время является анализ всех 2^k подмножеств из k независимых переменных. Однако серьезным недостатком данного метода является его значительная трудоемкость. Кроме того, все перечисленные методы полезны лишь при анализе линейных зависимостей между однородными группами данных.

Наличие указанных недостатков побудило нас предложить альтернативный автоматизированный алгоритм определения набора независимых переменных, входящих в состав многофакторной регрессионной модели и установления ее класса. Процедура указанного метода включает выполнение ряда этапов.

Инициализацией алгоритма является определение исходного множества независимых переменных $x_1, x_2, \ldots, x_k$. При этом мини-

мальное количество наблюдений, достаточное для проведения множественного регрессионного анализа, должно составлять $n \geq 10k$, в противном случае оценки регрессионной линии будут ненадежными и плохо воспроизводимыми [13].

На втором, самом трудоемком этапе производится раздельный корреляционно-регрессионный анализ каждой зависимости y от x_i, $i = 1,2,\ldots,k$, по стандартной схеме (рис. 30). По результатам комплекса однофакторных корреляционно-регрессионных анализов исключаются переменные, не коррелирующие с результативным показателем или характеризующиеся выраженным неконтролируемым кластерингом. Кроме этого, осуществляется подгонка наилучших аппроксимаций однофакторных зависимостей y от каждой неисключенной переменной x_i.

Далее производится построение многофакторной регрессионной модели $\tilde{y} = \beta_0 + \beta_i x_i$, включающей все лучшие однофакторные регрессии, и проверяется ее статистическая значимость в целом и значимость каждого отдельного коэффициента регрессии. При обнаружении незначимых регрессионных коэффициентов из состава многофакторной модели исключаются соответствующие независимые переменные. Независимые переменные или их преобразования, сохранившие статистическую значимость, и определяют класс итоговой многофакторной регрессионной модели.

На заключительном этапе итоговая регрессионная модель проверяется на неоднородность остатков. При отсутствии гетероскедастичности регрессионный анализ завершается. При наличии гетероскедастичности прибегают к одному из двух методов: либо используют регрессионный анализ с какой-либо другой функцией потерь, либо применяют метод скользящего остаточного стандартного отклонения.

В случае использования других функций потерь класс многофакторной регрессионной модели остается неизменным (в ходе анализа изменятся лишь значения коэффициентов регрессии), при ликвидации гетероскедастичности доверительная область для прогнозных оценок строится на основе односторонней верхней интервальной оценки остаточного стандартного отклонения. Изложенный алгоритм определения состава и установления класса многофакторной регрессионной модели приведен на рисунке 31.

Рис. 30. Оптимальная стратегия однофакторного регрессионного анализа при судебно-медицинской антропологической идентификации личности.

Рис. 31. Оптимальная стратегия многофакторного регрессионного анализа при судебно-медицинской антропологической идентификации личности.

Для демонстрации изложенного алгоритма изложим проведенную нами процедуру построения многофакторной регрессионной модели идентификации гестационного возраста по комплексу морфометрических показателей фетальных органов [53]. Исходное множество независимых переменных включало комплекс из 10 морфометрических показателей печени и селезенки. В печени определялись кроветворная активность паренхимы, толщина капсулы, относительный объем печеночной стромы, относительный объем печеночных долек. В селезенке оценивались толщина капсулы, диаметр и плотность расположения лимфоидных узелков, толщина стенок центральных артерий, относительные объемы белой пульпы, трабекулярного компонента и красной пульпы.

Корреляционный анализ статистических взаимосвязей каждого из исследовавшихся параметров с гестационным возрастом позволил исключить ряд морфометрических показателей, характеризовавшихся отсутствием или слабой выраженностью искомых зависимостей (табл. 23). Для каждого оставшегося показателя была определена форма его зависимости от гестационного возраста и оценена степень ее тесноты.

Таблица 23

Результаты анализа взаимозависимостей гестационного возраста с морфометрическими показателями фетальных органов

Морфометрические показатели	n	Форма зависимости	r (r_s)*
Печени			
Кроветворная активность**	131	Нелинейная	
Относительный объем стромы	31	-	0,047
Толщина капсулы	31	-	0,291
Селезенки			
Диаметр лимфоидных узелков	99	Линейная	0,633
Плотность лимфоидных узелков	99	Нелинейная	-0,769
Толщина стенок центральных артерий	99	Линейная	0,587
Толщина капсулы**	98	Линейная	0,379
Относительный объем трабекул	31	-	0,278
Относительный объем белой пульпы	31	-	-0,120
Относительный объем красной пульпы	31	-	-0,084

Примечание. * - сила линейных связей выражалась с помощью коэффициента корреляции Пирсона, нелинейных - коэффициента ранговой корреляции Спирмена; ** - коэффициенты корреляции рассчитаны для данных после исключения контролируемых видов кластеринга.

При анализе данных была обнаружена неоднородность наблюдений кроветворной активности печени за счет наличия кластера недоношенных новорожденных с острой неонатальной инволюцией экстрамедуллярной миелоидной ткани (см. раздел 2.6). Дальнейшая проверка доказала наличие относительной однородности исследованных совокупностей значений остальных гистометрических показателей, несмотря на наличие в них наблюдений с разнообразной перинатальной патологией. Возможность идентификации кластера недоношенных новорожденных с постнатальной инволюцией экстрамедуллярной кроветворной ткани позволила исключить его из регрессионного анализа.

Проведенный на основе различных линеаризирующих преобразований регрессионный анализ выделил комплекс наилучших подгонок однофакторных регрессионных моделей идентификации гестационного возраста (табл. 24, см. также разделы 3.4 и 3.6).

Таким образом, многофакторной моделью идентификации гестационного возраста, включающей все лучшие однофакторные регрессии, должна быть комплексная регрессионная модель, содержащая следующие факторные переменные: кубический полином кроветворной активности паренхимы печени, логарифм плотности лимфоидных узелков селезенки, диаметр лимфоидных узелков и толщина стенок центральных артерий селезенки.

Отсюда созданное на основе анализа 91 наблюдения многофакторное регрессионное уравнение

$$\hat{y} = 35{,}428 - 0{,}395x_1 + 0{,}004x_1^2 - 1{,}297 \cdot 10^{-5} x_1^3 + 0{,}023x_2 - 3{,}964 \lg x_3 + 0{,}240x_4$$

являлось статистически значимым ($F = 97{,}254$; $p = 1{,}132 \cdot 10^{-35}$), характеризовалось лучшими оценками коэффициентов множественной корреляции ($r = 0{,}935$) и детерминации ($\bar{r}^2 = 0{,}865$), минимальным остаточным стандартным отклонением ($s_\varepsilon = 2{,}154$) и содержало только значимые коэффициенты регрессии ($p < 0{,}05$).

Таблица 24

Комплекс наилучших подгонок однофакторных регрессионных моделей идентификации гестационного возраста

Показатель	Модель	r	r^2	$\bar{r}^2$	s_ε
x_1	$\tilde{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3$	0,920	0,846	0,842	2,279
x_2	$\tilde{y} = \beta_0 + \beta_1 x_2$	0,633	0,400	-	4,396
x_3	$\tilde{y} = \beta_0 + \beta_1 \lg x_3$	0,741	0,549	-	3,812
x_4	$\tilde{y} = \beta_0 + \beta_1 x_4$	0,587	0,345	-	4,596

Проведенное тестирование данной регрессионной модели выявило неоднородность дисперсии ее остатков (F = 2,916; p = 0,002). Учитывая, что гетероскедастичность оказывает влияние на адекватность t-статистик коэффициентов регрессии, справедливым являлось предположение о возможной излишней оптимистичности стандартных ошибок регрессионных коэффициентов. В данном случае наиболее вероятно превышение значения вероятности ошибки первого рода для t-статистики регрессионного коэффициента, соответствующего показателю толщины стенок центральных артерий (t = 2,066; p = 0,042).

Поскольку исключение показателя толщины стенок центральных артерий вследствие снижения мультиколлинеарности приведет к уменьшению вероятности ошибки первого рода для t-статистик остальных регрессионных коэффициентов, более адекватной задаче определения гестационного возраста является следующая комплексная регрессионная модель:

$$\hat{y} = 38{,}118 - 0{,}427x_1 + 4{,}374 \cdot 10^{-3} x_1^2 - 1{,}491 \cdot 10^{-5} x_1^3 + 0{,}024x_2 - 4{,}290 \lg x_3 .$$

Исключение показателя толщины стенок центральных артерий сопровождалось снижением интенсивности мультиколлинеарности. Это нашло выражение в повышении статистической значимости F – теста (F = 111,562; $p = 8{,}026 \cdot 10^{-36}$). По качеству итоговая модель лишь очень незначительно уступает предыдущей, характеризуясь приемлемыми точечными оценками коэффициентов множественной корреляции (r = 0,932) и детерминации ($\bar{r}^2 = 0{,}860$). Наиболее важным явилось повышение статистической значимости регрессионных коэффициентов (p < 0,02). Отрицательным эффектом явились незначительное увеличение остаточного стандартного отклонения ($s_\varepsilon = 2{,}195$) и умеренное увеличение гетероскедастичности (F = 3,045; $p = 9{,}730 \cdot 10^{-4}$). Однако с помощью метода остаточного скользящего стандартного отклонения удалось разработать номограмму определения интервальных оценок гестационного возраста по его точечным оценкам, рассчитанным на основе итоговой многофакторной регрессии [11,53].

В итоге проведенным исследованием была создана комплексная регрессионная модель, позволяющая по значениям трех морфометрических параметров печени и селезенки определять как точечные, так и интервальные оценки гестационного плодов и новорожденных.

Таким образом, изложенные методы корреляционно-регрессионного анализа обеспечивают построение регрессионных моделей, с определенной точностью позволяющих прогнозировать значение идентифицируемого параметра по значению идентифицирующего признака или группы признаков. В отличие от использующихся в настоящее время в судебно-медицинской антропологии алгоритмов, приведенные методы эффективны также при несоответствии изучаемых данных основным предпосылкам классического регрессионного анализа (неоднородность данных, нелинейность регрессий, качественный или порядковый тип независимых переменных, гетероскедастичность, серийная корреляция остатков). Приведенный алгоритм многофакторного регрессионного анализа является решением проблемы автоматизированного (лишенного субъективности) определения перечня независимых переменных или их преобразований, входящих в состав множественной регрессионной модели.

Наличие указанных преимуществ характеризует изложенные статистические процедуры как методы выбора при проведении судебно-медицинских антропологических исследований, связанных с созданием способов прогнозирования идентифицируемых параметров количественного типа. Кроме того, отдельные статистические алгоритмы, в частности, методы сравнения критериев точности регрессионных моделей, могут использоваться в судебно-медицинской экспертной практике для объективного выбора способов идентификации с наибольшей диагностической значимостью.

В заключение следует напомнить, что регрессионные уравнения, прежде всего, являются более или менее удачными подгонками эмпирических данных, удобными для практических целей, и могут не раскрывать глубинную сущность взаимосвязей между входящими в их состав величинами [82]. Данное обстоятельство, несмотря на наличие автоматизированных статистических алгоритмов, не освобождает исследователей - авторов способов идентификации и судебно-медицинских экспертов – пользователей этих способов от анализа сущности изучаемых биомедицинских данных (объектов научного и экспертного познания).

ГЛАВА 4. МЕТОДЫ ОДНОМЕРНОЙ КЛАССИФИКАЦИИ В СУДЕБНО-МЕДИЦИНСКОЙ АНТРОПОЛОГИИ

4.1. ОСНОВНЫЕ ПРИНЦИПЫ СУДЕБНО-МЕДИЦИНСКОГО КЛАССИФИЦИРОВАНИЯ

Одной из важнейших задач судебной медицины является разработка и совершенствование методов классификации объектов судебно-медицинского экспертного познания. Под термином «классификация» понимается отнесение объекта познания к одному из нескольких взаимоисключающих классов (групп), которые не могут быть упорядочены или описаны количественно. При этом следует различать две основные задачи классификации [26].

В большинстве случаев задачей классификации является определение принадлежности (идентификация) объекта к одной из нескольких, заранее известных, взаимоисключающих групп. Решение этой задачи достигается путем разработки методов судебно-медицинской классификации на основе априорной информации о распределении генеральных совокупностей значений классификационных признаков, которая представлена выборками из них.

Кроме этого, иногда задачей классификации является разбиение имеющейся конечной совокупности объектов, каждый из которых характеризуется одинаковым числом признаков, на однородные группы, количество которых заранее неизвестно. Указанные группы объектов обычно называются кластерами (от англ. cluster – группа элементов, характеризуемых каким-либо общим свойством), а также таксонами (от англ. taxon – систематизированная группа любой категории). Причем может оказаться, что исследуемое множество объектов не обнаруживает естественного расслоения на кластеры, т.е. образует один кластер. Подобные классификации основаны на использовании алгоритмов кластерного анализа. От других методов многомерной классификации кластерный анализ отличается отсутствием обучающих выборок, т.е. априорной информации о распределении классификационных признаков.

Следует отметить, что к определению понятий «классификация» и «идентификация» теоретическая криминалистика и математическая статистика подходят с несколько различных позиций. Математическая статистика указанные термины считает тождественными. С точки зрения криминалистики между классификацией и иденти-

фикацией существуют определенные, хотя и малосущественные различия. В частности, под идентификацией понимается решение классификационных задач первого типа, а классификация подразумевает решение классификационных задач второго типа [14].

Основой видовой структуры методов судебно-медицинской классификации является число групп, на которые кластерообразующий параметр разделяет потенциальное бесконечное множество объектов экспертного познания [57]. В зависимости от количества указанных групп целесообразно выделять биномиальную и полиномиальную схемы судебно-медицинской классификации. Биномиальная классификация разделяет классифицируемые объекты всего на два взаимоисключающих класса. Примером биномиальной классификации является идентификация пола. При полиномиальной схеме число групп, на которое производится разграничение объектов экспертного познания, более двух. Примерами полиномиальной классификации являются идентификация соматотипа и расы.

Вторым фактором, определяющим видовую структуру методов судебно-медицинской классификации, следует назвать количество признаков классифицируемых объектов. В качестве признаков обычно выступают какие-либо размерные показатели объектов классификации. В наиболее простом варианте одномерная классификация объектов может производиться всего лишь по одному признаку. Как правило, более точная классификация достигается по нескольким, наиболее информативным признакам, отобранным из множества характеристик, описывающих классифицируемые объекты. Математической моделью, лежащей в основе методов многомерной биномиальной и полиномиальной классификации, является дискриминантный анализ [121].

В случаях, когда ранжировать классификационные признаки по степени информативности не представляется возможным, применяется многомерная группировка путем создания интегрального показателя, функционально зависящего от исходных признаков с последующей классификацией по этому показателю. Развитием данного подхода можно назвать вариант классификации по нескольким обобщающим показателям (главным компонентам), полученным с помощью методов факторного анализа [4,25,43,157].

Таким образом, в настоящее время в судебной медицине в зависимости от задач классификации, свойств классифицируемых объектов и других исходных условий используется широкий спектр

методов статистического анализа, не включающий все же все известные статистические алгоритмы (рис. 32). В частности, пока не нашли должного применения такие эффективные методы как деревья классификации [133,134].

Основной, но далеко не единственной, областью практической реализации методов судебно-медицинской классификации являются экспертизы отождествления личности. Это объясняется тем, что главной целью данного вида судебно-медицинских экспертиз выступает научно обоснованное разграничение (идентификация) принадлежности объектов экспертного познания к одному из заранее известного конечного множества кластеров. Объектами идентификации могут быть фрагменты частей тела от неопознанных трупов людей, скелетированные трупы, части скелетов, отдельные кости и их фрагменты, зола из мест сожжения трупов. Основными кластерообразующими (идентифицируемыми) параметрами объектов, которые необходимо установить при производстве экспертиз отождествления личности, являются пол и соматотип человека, массивность скелета, порядковая локализация однотипных костей [34].

Наиболее простым из применяемых в судебно-медицинской антропологии методов биномиальной классификации является одномерная классификация, разграничивающая принадлежность идентифицируемых объектов по какому-либо одному их признаку к одной из двух взаимоисключающих групп. Фактически единственным кластерообразующим параметром при одномерном биномиальном классифицировании является половая принадлежность идентифицируемых объектов. Из-за этого данный метод в судебно-медицинской антропологии применяется не так часто, как дискриминантный анализ, который может использоваться при любых видах классификации. В то же время при идентификации пола обычно применяются оба указанных метода. В качестве классифицируемых признаков при этом обычно выступают размерные характеристики идентифицируемых объектов, характеризующиеся каким-либо типом непрерывных распределений. В зависимости от типа распределения исходных данных следует различать одномерную биномиальную классификацию нормально распределенных биометрических показателей и аналогичную процедуру при других типах непрерывных распределений. Наиболее хорошо изученной в судебной медицине является процедура одномерной биномиальной классификации при нормальном распределении исходных данных.

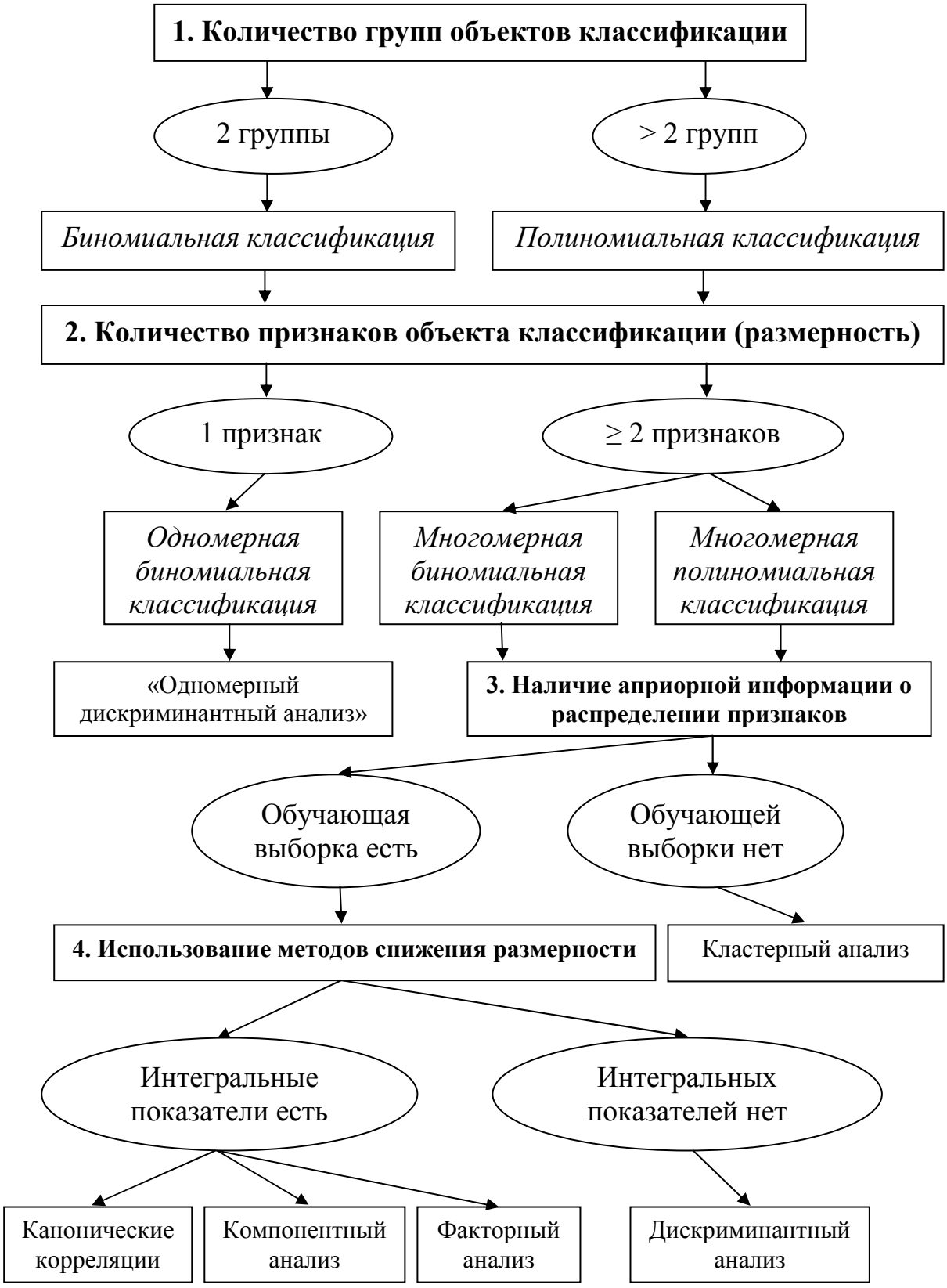

Рис. 32. Виды классификации объектов судебно-медицинского научного познания, структура статистических методов решения классификационных задач.

4.2. ОДНОМЕРНАЯ БИНОМИАЛЬНАЯ КЛАССИФИКАЦИЯ ПРИ НОРМАЛЬНОМ РАСПРЕДЕЛЕНИИ ПОКАЗАТЕЛЕЙ

Несмотря на разнообразие идентифицируемых объектов, математическая модель применяющихся в судебной медицине методов отождествления пола, основанных на одномерной биномиальной классификации, является единой и известна под термином «одномерный дискриминантный анализ» [42]. Алгоритм названной математической модели включает выполнение определенной последовательности действий [33,64].

Вначале фиксируются значения исследуемого показателя в каждой из выборочных совокупностей лиц мужского и женского пола, и проверяется соответствие полученных данных нормальному распределению. При отсутствии отклонений от нормальности вычисляются точечные оценки параметров распределения изучаемого показателя, по которым определяются его категории изменчивости в каждой из двух альтернативных групп объектов (табл. 25).

Полученные категории изменчивости служат основой интервальной шкалы принятия экспертных решений о половой принадлежности идентифицируемых объектов (табл. 26). Учитывая, что обычно размерные показатели у мужчин превышают таковые у женщин, то при соответствии исследуемого показателя категории «очень малый» для мужчин делается вероятный вывод о принадлежности идентифицируемого объекта женщине. При соответствии размера категории «очень большой» для женщин делается вероятный вывод о принадлежности идентифицируемого объекта мужчине.

Таблица 25

Категории изменчивости идентифицируемых объектов

Категория размера x	Концы соответствующего числового множества	Доля генеральной совокупности, %
Аномально малый	$x < \bar{x} - 3{,}30\sigma$	0,048
Очень малый	$\bar{x} - 3{,}30\sigma \le x < \bar{x} - 1{,}54\sigma$	6,130
Малый	$\bar{x} - 1{,}54\sigma \le x < \bar{x} - 0{,}56\sigma$	22,596
Средний	$\bar{x} \pm 0{,}56\sigma$	42,452
Большой	$\bar{x} + 0{,}56\sigma < x \le \bar{x} + 1{,}54\sigma$	22,596
Очень большой	$\bar{x} + 1{,}54\sigma < x \le \bar{x} + 3{,}30\sigma$	6,130
Аномально большой	$\bar{x} + 3{,}30\sigma < x$	0,048

Таблица 26

Интервальная шкала диагностики пола*

Значение биометрического показателя x	Пол
$x < \bar{x} - 3{,}30\sigma$ для мужчин	Достоверно женщина
$\bar{x} - 3{,}30\sigma \leq x < \bar{x} - 1{,}54\sigma$ для мужчин	Вероятно женщина
$x \geq \bar{x} - 1{,}54\sigma$ для мужчин $x \leq \bar{x} + 1{,}54\sigma$ для женщин	Не определен
$\bar{x} + 1{,}54\sigma < x \leq \bar{x} + 3{,}30\sigma$ для женщин	Вероятно мужчина
$x > \bar{x} + 3{,}30\sigma$ для женщин	Достоверно мужчина

Примечание. Шкала предназначена для классификации биометрических показателей, средние значения которых у мужчин превышают таковые для женщин.

Достоверный вывод о мужской или женской половой принадлежности формулируется только при превышении значения исследуемого показателя пределов категорий «очень большой» для женщин и соответственно «очень малый» для мужчин. При остальных значениях биометрического показателя половая принадлежность идентифицируемого объекта считается не установленной (см. табл. 26). Необходимо уточнить, что средние значения некоторых биометрических показателей у женщин превышают таковые у мужчин. В этом случае применяется аналогичная шкала с инверсией категорий изменчивости и экспертных выводов. Примером таких показателей являются толщина костей черепа в некоторых краниометрических ориентирах [33], размерные показатели таза [64].

Следует отметить, что выбор пределов категорий размеров и, соответственно, границ интервальной шкалы экспертных решений о половой принадлежности идентифицируемых объектов является субъективным. Например, в работе, посвященной разработке способов определения пола человека по рентгенограммам кисти, вместо интервалов 1,54s и 3,30s авторы использовали интервалы 2s и 3s [17]. Данное отличие является сугубо количественным и принципиального значения не имеет. Для сравнения результатов биометрических исследований, выполненных авторами с применением различных категорий изменчивости, целесообразно использовать следующие табулированные данные (табл. 27).

Таблица 27

Объем нормально распределенной генеральной совокупности, заключенный в пределах $\mu \pm z\sigma$, %

z	0	1	2	3	4	5	6	7	8	9
0	0,000	7,966	15,852	23,582	31,084	38,292	45,149	51,607	57,629	63,188
1	68,269	72,867	76,986	80,640	83,849	86,639	89,040	91,087	92,814	94,257
2	95,450	96,427	97,219	97,855	98,360	98,758	99,068	99,307	99,489	99,627
3	99,730	99,806	99,863	99,903	99,933	99,953	99,968	99,978	99,986	99,990

Недостатками одномерного дискриминантного анализа, рассматривающегося в настоящее время в качестве стандарта судебно-антропологического исследования, являются невозможность формулирования количественных вероятностных выводов о половой принадлежности идентифицируемых объектов и использование при определении категорий изменчивости биометрического показателя вместо истинных параметров его распределения их точечных оценок. При этом обязательны смещения рассчитанных категорий изменчивости от их истинных значений [16]. Изложенное приводит к тому, что при практическом использовании результатов одномерного дискриминантного анализа большая (15-20%) вероятность ошибочной классификации существует даже в случаях достоверного экспертного определения пола. Указанное обстоятельство обязывает делать завершающим этапом подобного судебно-антропологического исследования осуществление кросс-проверки.

В теории дискриминантного анализа кросс-проверка представляет собой процедуру оценки точности классификации объектов с помощью данных из специальной тестовой выборки (в математической статистике используется также термин кросс-проверочная выборка). При наличии выборок достаточно большого объема рекомендуется часть наблюдений (половину или две трети) использовать для обучающей выборки, а оставшиеся наблюдения - для тестовой [13]. Если на тестовой выборке модель классификации дает результаты того же качества, что и на обучающей выборке, то считается, что модель хорошо прошла кросс-проверку.

В связи с изложенным, тестирование точности классификации пола является неотъемлемой частью судебно-антропологического исследования. Поэтому процедура проведения большинства исследований соответствующей тематики предусматривает разделение

имеющихся биометрических данных на обучающую группу и группу верификации. При этом для разработки интервальной шкалы диагностики пола используются только данные обучающей группы, а группа верификации служит лишь для определения точности классификации. Данный подход дает ориентировочное (поскольку доверительные интервалы точности классификации не вычисляются) представление о доле случаев ошибочной классификации пола. Однако это достигается ценой значительного уменьшения объема потенциально пригодных для определения параметров распределения изучаемого показателя биометрических данных, что в итоге приводит к снижению точности идентификации.

Учитывая данное обстоятельство, отдельные авторы используют все имеющиеся в их распоряжении биометрические данные для разработки модели классификации пола и на них же проверяют ее точность [17]. Такой подход противоречит теории дискриминантного анализа, в соответствии с которой только классификация новых наблюдений позволяет определить качество дискриминантной модели [13].

Относительно необходимости кросс-проверки следует заметить, что так называемый в судебно-медицинской литературе «одномерный дискриминантный анализ», собственно говоря, таковым не является, представляя собой процедуру определения вероятности значений признака при заданных параметрах его распределения. Смысл сказанного заключается в том, что если истинные значения параметров распределения изучаемого биометрического показателя в совокупностях объектов мужской и женской половой принадлежности были известны, то проведение кросс-проверки стало бы ненужным. Необходимость кросс-проверки отпадает также, если вместо неизвестных истинных значений параметров распределения использовать их интервальные оценки.

Поэтому существенным недостатком одномерного дискриминантного анализа является использование при определении доверительных интервалов (каковыми являются категории изменчивости) для значений биометрического показателя вместо истинных параметров его распределения их точечных оценок, полученных при исследовании ограниченных выборок зачастую небольшого объема. При этом в зависимости от объема выборки точечные оценки генерального среднего и генерального стандартного отклонения могут значительно отклоняться от их истинных значений (рис. 33).

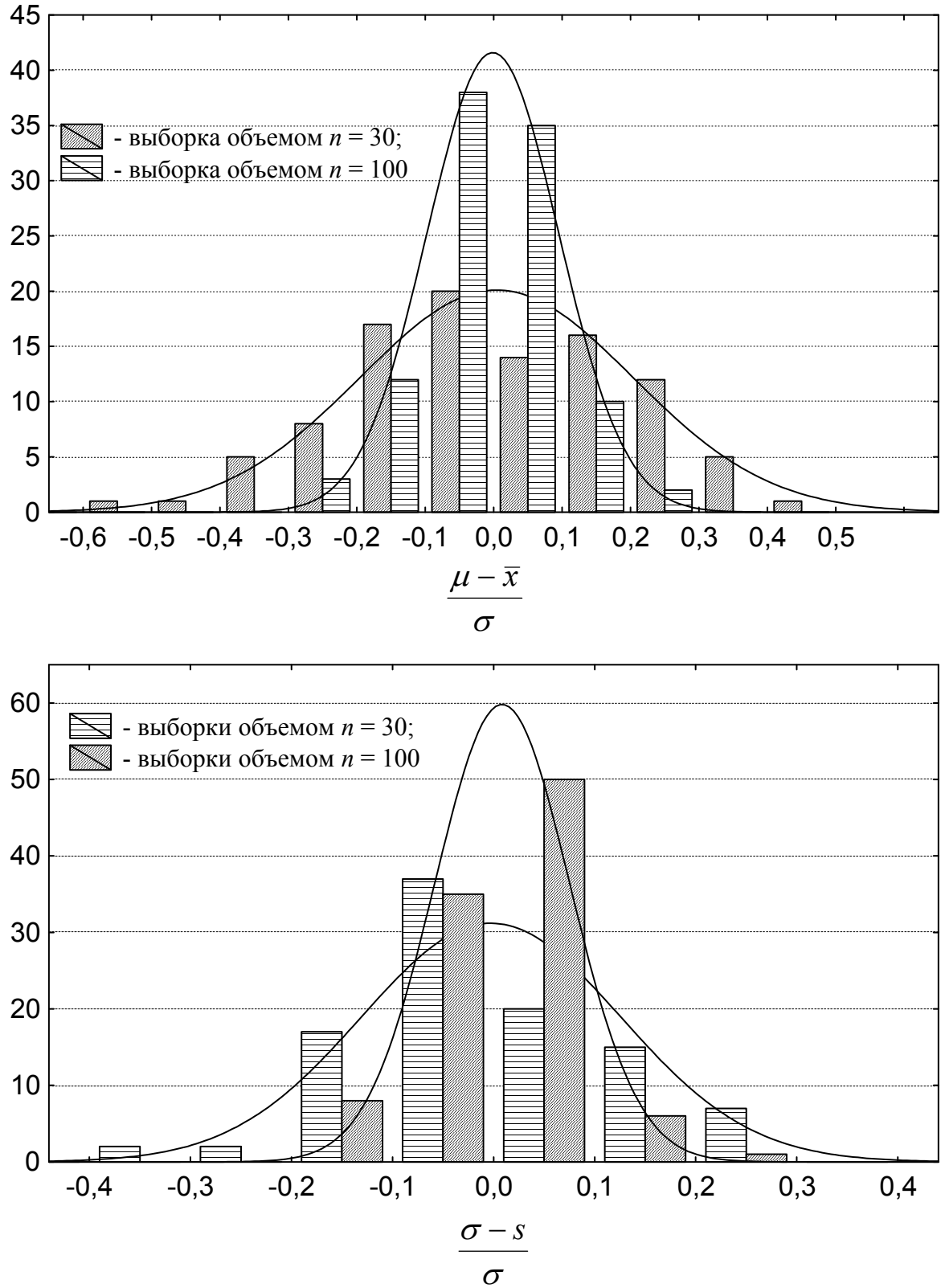

Рис. 33. Распределение отклонений точечных оценок генерального среднего (вверху) и генерального стандартного отклонения (внизу) от их истинных значений, полученные при исследовании 100 случайных выборок из 30 и 100 членов. По оси абсцисс – отклонение, по оси ординат – число наблюдений. Случайные выборки при заданных параметрах распределения генерированы с помощью приложения Microsoft Excel пакета Microsoft Office 2003.

Действительно, для нормально распределенной генеральной совокупности с известными значениями среднего и стандартного отклонения $(100-\alpha)\%$ допустимый интервал для значений определяется как

$$x_i \in \mu \pm z_\alpha \cdot \sigma .$$

Однако при неизвестных параметрах μ и δ определение доверительных интервалов для определенной части нормально распределенной генеральной совокупности значительно затрудняется, поскольку $\bar{x}$ и s являются не более чем точечными оценками истинных параметров μ и δ, как правило, не совпадающими с ними. При этом ошибка в оценке $\bar{x}$ накладывается на ошибку в оценке s – в результате шансы получить правильный результат становятся очень низкими [16]. Поэтому в качестве доверительного для определенной части генеральной совокупности берется более широкий интервал

$$x_i \in \bar{x} \pm k_{\alpha;\gamma;n} \cdot s ,$$

где $k_{\alpha;\gamma;n}$- допустимый коэффициент для нормального распределения. Величина $k_{\alpha;\gamma;n}$ зависит от доли γ членов генеральной совокупности, которые должны попасть в доверительный интервал, от вероятности 1 - α того, что они действительно туда попали и от объема выборки n [30,93]. Важно отметить, что допустимые коэффициенты недействительны при анормальных или неизвестных распределениях биометрических данных [110,140,159].

Однако в процессе создания критериев диагностики пола вместо истинных значений параметров распределения остеометрических показателей обычно используются их точечные оценки. Данное обстоятельство приводит к сужению и смещению границ категорий изменчивости каждого из исследовавшихся остеометрических показателей. Возникающие несоответствия между декларируемыми и реальными границами категорий размеров, хотя и не являются значительными, тем не менее, искажают интервальную шкалу диагностики пола и при практическом использовании могут привести к ошибочному экспертному заключению.

Наличие указанных недостатков одномерного дискриминантного анализа побудило нас предложить альтернативный метод одномерной биномиальной классификации нормально распределенных биометрических данных [57]. Данный метод уже был успешно использован нами при идентификации принадлежности фрагментов

печени и селезенки недоношенным новорожденным с постнатальной инволюцией экстрамедуллярной кроветворной ткани [11,53].

Алгоритм метода включает выполнение следующих этапов:

1. Фиксируют значения изучаемых количественных показателей у всех членов двух альтернативных совокупностей объектов классификации.

2. Проверяют соответствие полученных биометрических данных нормальному распределению с помощью любого статистического критерия или группы критериев (χ^2- критерий согласия, критерий согласия Колмогорова-Смирнова, критерий Колмогорова-Смирнова в модификации Лиллиефорса, тест Шапиро-Уилка, критерий отношения размаха к стандартному отклонению и др.). Предпочтительным является совместное использование χ^2- критерия и критерия Колмогорова-Смирнова в модификации Лиллиефорса.

3. При отсутствии отклонений от нормальности определяют точечные оценки параметров распределений биометрических показателей в каждой из двух альтернативных совокупностей объектов классификации.

4. Вычисляют объединенную оценку дисперсии каждого биометрического показателя:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

где s – объединенная оценка дисперсии; n_1 и n_2 - объемы выборок альтернативных групп идентифицируемых объектов; s_1 и s_2 - выборочные стандартные отклонения.

5. Для каждого биометрического показателя определяют нормированные межгрупповые различия по формуле:

$$\varphi = \frac{|\bar{x}_2 - \bar{x}_1|}{s},$$

где φ – параметр нецентральности; $\bar{x}_2$ и $\bar{x}_1$ – выборочные средние, и ранжируют потенциальные критерии классификации по абсолютным значениям φ. Наибольшая точность классификации пола будет соответствовать биометрическому показателю с максимальным, а наименьшая – с минимальным модулями параметра нецентральности.

6. Для каждого показателя определяют наихудшие с позиции точности классификации интервальные оценки генерального среднего и генерального стандартного отклонения. Для двух альтернативных групп объектов, состоящих из n_1 и n_2 членов с точечными оценками параметров $\bar{x}_1$, s_1 и $\bar{x}_2$, s_2, причем $\bar{x}_1 < \bar{x}_2$, наихудшими для точности классификации интервальными оценками неизвестных параметров распределения μ_1, σ_1 и μ_2, σ_2 являются:

$$\mu_1 = \bar{x}_1 + t_{\alpha;n_1-1} \cdot \frac{s_1}{\sqrt{n_1}} \text{ и } \sigma_1 = s_1 \cdot \sqrt{\frac{n_1 - 1}{\chi^2_{1-\alpha/2;n_1-1}}}; \quad (34)$$

$$\mu_2 = \bar{x}_2 - t_{\alpha;n_2-1} \cdot \frac{s_2}{\sqrt{n_2}} \text{ и } \sigma_2 = s_2 \cdot \sqrt{\frac{n_2 - 1}{\chi^2_{1-\alpha/2;n_2-1}}}, \quad (35)$$

где $t_{\alpha;n-1}$ и $\chi^2_{1-\alpha/2;n-1}$ - значения двустороннего критерия Стьюдента и χ^2- критерия при уровне значимости α и $v = n - 1$ количестве степеней свободы.

7. Строят упорядоченный ряд дискретных значений каждого исследуемого показателя: $x_1 > x_2 > \ldots x_{n-1} > x_n$, причем $x_1 \geq \bar{x}_2 + 3s_2$, $x_n \leq \bar{x}_1 - 3s_1$, а $x_i - x_{i+1} = \varepsilon$, где ε – максимальная погрешность измерительного инструмента.

8. Используя наихудшие для точности классификации интервальные оценки параметров распределения, дважды нормируют каждое значение x_i упорядоченного ряда $x_1 > x_2 > \ldots x_{n-1} > x_n$:

$$z_2^i = \frac{x_i - \mu_2}{s_2} \text{ и } z_1^i = \frac{x_i - \mu_1}{s_1},$$

где z – стандартная нормальная переменная.

9. Для каждого значения z определяют функцию плотности стандартного нормального распределения:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad (36)$$

после чего строят два ряда значений $f(z_2^i)$ и $1 - f(z_1^i)$.

10. Для каждого x_i определяют вероятность принадлежности к альтернативным классифицируемым группам по формулам:

$$p_1 = \frac{1 - f(z_1^i)}{(1 - f(z_1^i)) + f(z_2^i)} \cdot 100\% \text{ и } p_2 = \frac{f(z_2^i)}{(1 - f(z_1^i)) + f(z_2^i)} \cdot 100\%, \quad (37)$$

осуществляют проверку: $p_1 + p_2 = 1$. Полученные данные табулируют или представляют в форме набора номограмм.

Важнейшим отличием результатов приведенного алгоритма является формулирование экспертных выводов в количественной вероятностной форме, а не в терминах ранговой шкалы (достоверно, вероятно, пол не определен). Причем производится заблаговременный расчет вероятностей половой принадлежности для любого значения биометрического показателя идентифицируемого объекта, которое только возможно получить с помощью конкретного измерительного инструмента. Это позволяет в дальнейшем при производстве антропологических экспертиз по конкретному результату биометрии с помощью номограмм или табулированных данных сразу определять количественный вероятностный вывод о половой принадлежности идентифицируемого объекта.

Следующим принципиальным отличием приведенного метода одномерной биномиальной классификации от стандартного одномерного дискриминантного анализа является базирование всех расчетов на наихудших для точности классификации интервальных оценках параметров распределения исследуемого биометрического показателя. Следствием данного подхода является сведение риска ошибочности количественных вероятностных выводов о половой принадлежности идентифицируемых объектов к заранее определенному минимуму. Указанный минимум характеризуется вероятностью выхода истинного значения хотя бы одного какого-либо параметра распределения исследуемого показателя в одной из двух альтернативных групп объектов за пределы рассчитанных интервальных оценок. Данная вероятность полностью определяется уровнем значимости, равняясь 2α, и зависит от выбора исследователя. Изложенный подход позволяет утверждать, что вероятность половой принадлежности идентифицируемых объектов с надежностью $1 - 2\alpha$ не меньше значения, рассчитанного с помощью описанного метода одномерной биномиальной классификации. Это освобождает от необходимости последующего тестирования точности классификации пола.

Последнее отличие изложенного метода одномерной биномиальной классификации от одномерного дискриминантного анализа заключается в том, что ранжирование диагностических критериев позволяет использовать для идентификации лишь один показатель с максимальной точностью классификации пола, а при невозмож-

ности его определения в случаях исследования фрагментированных объектов – показатель, по точности следующий за первым.

Практическое использование метода одномерной биномиальной классификации можно показать на примере идентификации половой принадлежности подъязычной кости с использованием остеометрических данных В.Н. Звягина, Н.Л. Мальцевой, Л.А. Алексиной и О.И. Галицкой, полученных названными авторами при исследовании указанных костных объектов от 133 трупов лиц мужского и женского пола в возрасте 15-83 лет [41].

Программа остеометрии подъязычной кости, выполненная указанными авторами, включала определение 16 размеров, обозначенных ими как М1 – М16. Границы категорий отдельно взятого размера в данной работе устанавливали исходя из общепринятых величин (см. табл. 25). Отмеченная авторами встречаемость случаев в этих границах составила 33%, по 22%, 11,5% и 0% соответственно.

Осуществленная нами проверка показала соответствие распределений каждого из 16 количественных показателей подъязычной кости нормальному распределению ($\chi^2 > 10{,}359$; $p > 0{,}110$). Одномерная биномиальная классификация при уровне значимости $\alpha = 0{,}05$ выполнялась только для 10 остеометрических показателей, характеризовавшихся наличием половых различий.

Проведенное по точности классификации пола ранжирование остеометрических критериев показало, что в экспертной практике для определения половой принадлежности подъязычной кости необходимо использовать лишь показатель М16 (общая длина подъязычной кости), а при невозможности его определения в случаях исследования фрагментированных объектов – показатель М1 (усредненная длина больших рогов) и т.д. (табл. 28). Следует отметить, что показатели М5 (максимальный диаметр булавовидного утолщения больших рогов) и М12 (длина 3-го сегмента больших рогов) ввиду небольших межгрупповых различий оказались непригодными для проведения одномерной биномиальной классификации, поскольку интервальные оценки генерального среднего для женщин превысили таковые для мужчин.

Результаты выполнения остальных этапов одномерной биномиальной классификации приводим только для остеометрического показателя М16, характеризующегося наибольшей точностью диагностики пола (табл. 29).

Таблица 28

Исходные данные*, объединенная дисперсия, нормированные межполовые различия и интервальные оценки параметров распределения остеометрических показателей подъязычной кости

	M16**	M1	M4	M2	M3	M8	M13	M9	M5	M12
n_2	89	89	89	89	89	89	89	89	89	89
$\bar{x}_2$	41,1	31,7	30,8	24,1	46,2	8,1	15,0	10,2	4,7	8,1
s_2	3,5	2,9	3,3	2,9	6,9	1,5	2,1	2,1	1,0	3,1
n_1	44	44	44	44	44	44	44	44	44	44
$\bar{x}_1$	35,7	27,6	26,6	20,8	39,3	6,8	13,3	9,1	4,3	6,9
s_1	3,2	3,2	3,3	2,1	6,3	1,1	1,9	1,5	0,7	2,6
s	3,411	2,998	3,266	2,653	6,743	1,391	2,013	1,946	0,908	2,920
φ	1,583	1,367	1,286	1,244	1,023	0,935	0,844	0,565	0,441	0,411
μ_2	40,36	31,09	30,12	23,50	44,74	7,78	14,56	9,75	4,49	7,46
σ_2	4,12	3,40	3,81	3,37	8,14	1,78	2,44	2,49	1,16	3,59
μ_1	36,67	28,57	27,60	21,45	41,22	7,13	13,87	9,57	4,52	7,69
σ_1	4,05	4,04	4,18	2,71	8,01	1,37	2,37	1,94	0,90	3,31

Примечание. * - исходные данные приведены из работы В.Н. Звягина, Н.Л. Мальцевой, Л.А. Алексиной и О.И. Галицкой [41]; ** - показатели указаны в порядке убывания точности классификации пола. Поскольку средние значения остеометрических показателей у лиц мужского пола превышают таковые у лиц женского пола, то для всех статистик нижний индекс, равный 1, соответствует женщинам, равный 2 - мужчинам.

Учитывая, что для M16 $\bar{x}_2 + 3s_2 = 51{,}6$ мм, $\bar{x}_1 - 3s_1 = 26{,}1$ мм, а максимальная погрешность измерительного инструмента $\varepsilon = 0{,}1$ мм, упорядоченный ряд дискретных значений данного остеометрического критерия имеет вид: 51,6; 51,5; … 26,2; 26,1 мм. Графическое выражение полученных для показателя M16 результатов классификации пола представлено на следующей номограмме (рис. 34).

Таблица 29

Вероятности половой принадлежности для экстремумов M16

M16, мм	z_2^i	z_1^i	$f(z_2^i)$	$1 - f(z_1^i)$	p_2	p_1
51,6	2,730	3,682	0,997	$1{,}159 \cdot 10^{-4}$	99,988	0,012
26,1	-3,730	-2,879	$9{,}561 \cdot 10^{-5}$	0,998	0,010	99,990

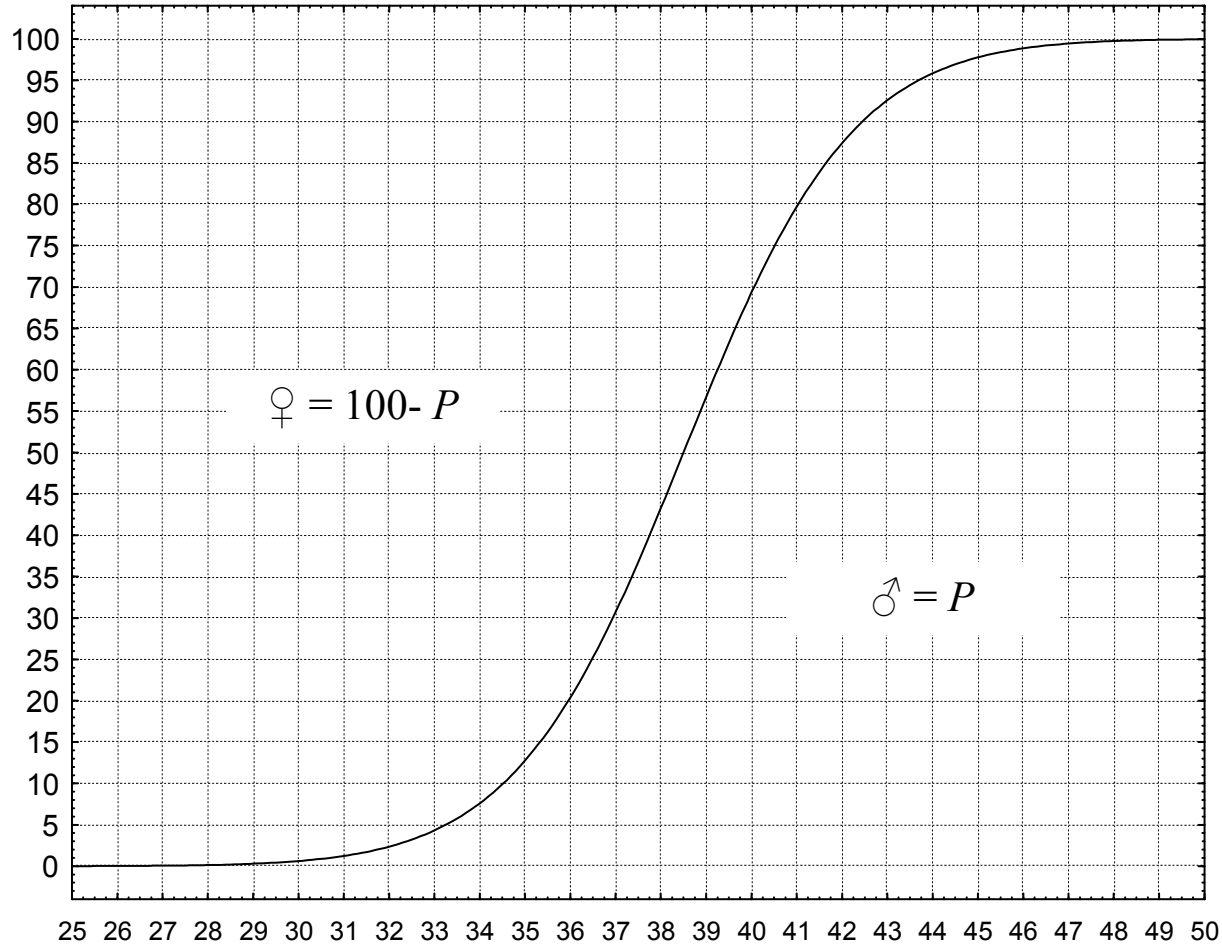

Рис. 34. Номограмма определения половой принадлежности подъязычной кости по признаку M16. По оси абсцисс – M16, мм, по оси ординат – вероятность принадлежности лицу мужского пола, %.

Аналогичные номограммы разработаны и для остальных остеометрических признаков подъязычной кости. Сравнительная характеристика результатов идентификации пола по показателю M16 подъязычной кости в соответствии с алгоритмами одномерного дискриминантного анализа и метода одномерной биномиальной классификации приведены в таблице 30.

Таким образом, предложенный нами метод одномерной биномиальной классификации нормально распределенных биометрических величин по сравнению с алгоритмом стандартного одномерного дискриминантного анализа обладает рядом следующих преимуществ:

1. Позволяет формулировать точные количественные вероятностные суждения о половой принадлежности классифицируемых объектов, повышая тем самым объективность и достоверность экспертных выводов.

Таблица 30

Результаты идентификации половой принадлежности подъязычной кости по признаку M16 с применением различных методов одномерной биномиальной классификации

M16, мм	Метод одномерной биномиальной классификации	
	Стандартный*	Альтернативный
29,1	Достоверно женщина	Женщина с вероятностью $p \geq 99{,}7\%$
31,3	Вероятно женщина	Женщина с вероятностью $p \geq 98{,}5\%$
36,1	Пол не определен	Женщина с вероятностью $p \geq 78{,}7\%$
40,4	Пол не определен	Мужчина с вероятностью $p \geq 73{,}8\%$
44,9	Вероятно мужчина	Мужчина с вероятностью $p \geq 97{,}6\%$
47,0	Достоверно мужчина	Мужчина с вероятностью $p \geq 99{,}4\%$

Примечание. * - идентификация проведена в соответствии с методикой В.Н. Звягина, Н.Л. Мальцевой, Л.А. Алексиной и О.И. Галицкой [41].

2. Реализация метода не требует последующего определения точности классификации пола с применением группы верификации. Это позволяет использовать для разработки методик судебно-медицинской идентификации весь имеющийся в распоряжении исследователя объем биометрических данных. В итоге метод идентификации пола, разработанный с помощью альтернативного алгоритма одномерной биномиальной классификации, при одинаковом объеме биометрических данных всегда будет более точным по сравнению с аналогичной методикой, выполненной с применением одномерного дискриминантного анализа.

3. Применение результатов метода одномерной биномиальной классификации в экспертной практике уменьшает трудоемкость классификации пола, поскольку для этого необходимо определение всего лишь одного, характеризующегося наибольшей диагностической точностью, количественного показателя идентифицируемого объекта.

Наличие указанных преимуществ делает приоритетным использование альтернативного метода одномерной биномиальной классификации в судебно-медицинских антропологических исследованиях, посвященных разработке критериев идентификации половой принадлежности останков человека.

4.3. ОДНОМЕРНАЯ БИНОМИАЛЬНАЯ КЛАССИФИКАЦИЯ НЕПРЕРЫВНЫХ БИОМЕТРИЧЕСКИХ ВЕЛИЧИН, НЕ ПОДЧИНЯЮЩИХСЯ НОРМАЛЬНОМУ РАСПРЕДЕЛЕНИЮ

Единственным предположением, которому должны соответствовать эмпирические данные при проведении одномерной биномиальной классификации, является хотя бы приближенное их подчинение нормальному закону [110,140,159]. При невыполнении данного предположения или при неизвестном типе распределения результаты любого из изложенных методов одномерной классификации будут некорректными. Несмотря на это авторы отдельных методик идентификации пола при создании последних, тем не менее, использовали одномерный дискриминантный анализ [17]. Для уменьшения числа случаев ошибочной идентификации пола по рентгенологическим показателям левой кисти, обусловленных анормальностью распределений биометрических данных, авторы указанной работы разработали дополнительные эмпирические критерии достоверности результатов классификации. В частности, по мнению данных исследователей, практически достоверная диагностика мужского пола имеет место, если выполнено одно из условий:

- в достоверно мужской интервал попали значения 4 или более биометрических показателей;
- разность количества значений биометрических показателей, попавших в вероятно мужской и вероятно женский интервалы, более или равна 14.

Схожие критерии были предложены и для остальных категорий интервальной шкалы диагностики пола [17].

В связи с этим следует отметить, что одновременное использование при классификации пола сразу всех биометрических показателей идентифицируемого объекта мало влияет на увеличение точности идентификации. Это связано с тем, что диагностические критерии по своей сути являются размерными показателями одного органа (идентифицируемого объекта) и, как правило, тесно коррелированы между собой. Это означает, например, что подъязычная кость, характеризующаяся большой общей длиной, вероятнее всего, также отличается большой величиной и других ее размерных характеристик. Именно поэтому для успешной одномерной биномиальной классификации необходим и достаточен лишь один биомет-

рический показатель, характеризующейся наибольшей точностью классификации пола. Одновременное использование нескольких количественных показателей идентифицируемого объекта возможно только в рамках многомерного дискриминантного анализа, алгоритм которого помимо прочего предусматривает исключение из дискриминирующих и классифицирующих функций тесно коррелирующих между собой переменных [13,121]. Введение же в рамках одномерной классификации пола дополнительных критериев достоверности, подобных использованным в анализируемой работе [17], приведет не столько к увеличению точности метода идентификации, сколько к снижению его чувствительности.

Таким образом, выраженное несоответствие биометрических данных нормальному распределению является противопоказанием к использованию методов одномерной биномиальной классификации нормально распределенных показателей.

К сожалению, при проведении судебно-медицинских антропологических исследований иногда приходится иметь дело с непрерывными биометрическими величинами, не подчиняющимися нормальному распределению. Выходом из этой ситуации является создание специфических математических моделей проведения одномерной биномиальной классификации, предусматривающих названную особенность распределения количественных данных.

Одним из возможных решений задач подобного рода является аппроксимация имеющихся биометрических данных каким-либо из существующих теоретических распределений, для которого можно рассчитать функцию плотности. Одномерный дискриминантный анализ применим при большинстве таких распределений. Метод одномерной биномиальной классификации в том виде, в каком он изложен в разделе 4.2, может быть использован только при подчинении биометрических данных логнормальному распределению. Данное распределение очень удобно тем, что путем преобразования каждого значения биометрического показателя по формуле $y = \ln x$ доступен переход к нормальному распределению с возможностью использования всех статистических свойств последнего, в том числе и в плане определения интервальных оценок его параметров [112]. Естественно, что перед применением такого подхода необходимо убедиться в хорошем соответствии количественных данных тому типу непрерывного распределения, статистические свойства которого будут использованы при решении задачи классификации.

Радикальным решением задачи одномерной биномиальной классификации любых непрерывных биометрических величин является использование неравенства Чебышева, доказанное в теории вероятностей [85]. В терминах статистической совокупности оно имеет следующую трактовку: для любой выборочной совокупности доля значений, попадающих в интервал $\bar{x} \pm ks$ будет равна, по крайней мере, $1-1/k^2$, где k – любое число, большее 1 [78]. Ценность неравенства Чебышева заключается в том, что оно будет верно для любого частотного распределения данных. Это позволило нам предложить универсальную интервальную шкалу диагностики пола, верную для любой совокупности биометрических данных независимо от формы кривой распределения (табл. 31).

Таблица 31

Универсальная шкала диагностики пола по показателю x

Значение x*	Пол	Максимальная ошибка, %
$x < \bar{x}_2 - 4{,}472 s_2$	Женский	5
$x < \bar{x}_2 - 3{,}162 s_2$	Женский	10
$x < \bar{x}_2 - 2{,}236 s_2$	Женский	20
$\begin{cases} x \geq \bar{x}_2 - 2{,}236 s_2 \\ x \leq \bar{x}_1 + 2{,}236 s_1 \end{cases}$	Не определен	0
$x > \bar{x}_1 + 2{,}236 s_1$	Мужской	20
$x > \bar{x}_1 + 3{,}162 s_1$	Мужской	10
$x > \bar{x}_1 + 4{,}472 s_1$	Мужской	5

Примечание. * - $\bar{x}_1$ и s_1 - оценки среднего и стандартного отклонения показателя в выборочной совокупности женщин, $\bar{x}_2$ и s_2 - в совокупности мужчин, причем $\bar{x}_1 < \bar{x}_2$.

Таким образом, предложенная универсальная шкала диагностики пола позволяет формулировать количественные вероятностные выводы о половой принадлежности классифицируемых объектов.

Еще одно преимущество универсальной шкалы по сравнению с одномерным дискриминантным анализом характеризуется тем, что при ее использовании максимальная ошибка идентификации пола известна заранее. Поэтому применение универсальной шкалы не требует последующего определения точности классификации с помощью кросс-проверки, что позволяет использовать для разработки

методик судебно-медицинской идентификации весь имеющийся в распоряжении исследователя объем биометрических данных.

Изложенное делает целесообразным применение универсальной шкалы диагностики пола в судебно-антропологических исследованиях, основанных на использовании биометрических показателей, не подчиняющихся нормальному распределению.

Важно подчеркнуть, что ценой универсальности предложенной диагностической шкалы явилось довольно значительное снижение точности классификации пола (табл. 32).

Таблица 32

Характеристика результатов различных методов одномерной идентификации пола по признаку M16 подъязычной кости

M16, мм	Метод одномерной биномиальной классификации	
	Универсальный	Для нормального распределения
29,1	Женщина с $p \geq 90\%$	Женщина с $p \geq 99{,}7\%$
31,3	Женщина с $p \geq 80\%$	Женщина с $p \geq 98{,}5\%$
36,1	Пол не определен	Женщина с $p \geq 78{,}7\%$
40,4	Пол не определен	Мужчина с $p \geq 73{,}8\%$
44,9	Мужчина с $p \geq 80\%$	Мужчина с $p \geq 97{,}6\%$
47,0	Мужчина с $p \geq 90\%$	Мужчина с $p \geq 99{,}4\%$

В этой связи единственным показанием к использованию универсальной шкалы при проведении антропологических исследований, посвященных разработке методов идентификации пола, следует считать отсутствие нормальности распределения исходных биометрических данных.

ГЛАВА 5. ДИСКРИМИНАНТНЫЙ АНАЛИЗ

5.1. ДИСКРИМИНАНТНЫЙ АНАЛИЗ ПРИ СУДЕБНО-МЕДИЦИНСКОЙ ИДЕНТИФИКАЦИИ ЛИЧНОСТИ

Дискриминантный анализ лежит в основе большинства методов многомерной классификации объектов судебно-медицинского экспертного познания. Так, из 13 случайно отобранных оригинальных исследований, посвященных судебно-медицинской классификации идентифицируемых объектов и опубликованных журналом «Судебно-медицинская экспертиза» в течение 2001-2006 гг., результаты 11 (85%) работ основывались на данных дискриминантного анализа. При этом в 2 (18%) из них использовалось несколько методов дискриминантного анализа.

В структуре многомерных статистических методов дискриминантный анализ является разделом, включающим в себя методы классификации многомерных наблюдений в ситуации, когда исследователь обладает априорной информацией о распределении классифицирующих (дискриминирующих) признаков. Априорная информация должна быть представлена выборками из альтернативных совокупностей (классифицируемых групп).

В отличие от одномерных методов дискриминантный анализ позволяет проводить как биномиальную, так и полиномиальную классификацию. При этом наиболее изучен случай, когда известно, что распределение признаков в каждой из классифицируемых групп наблюдений нормально, но отсутствует информация о параметрах этих распределений. Предположение о нормальности распределения признаков реализовано в параметрических методах дискриминации. Наряду с этим существуют менее известные непараметрические методы дискриминации, не требующие знаний о точном функциональном виде распределений и позволяющие решать задачи дискриминации на основе незначительной априорной информации о классифицируемых группах. Существующие методики судебно-медицинской идентификации личности основаны на наиболее изученных параметрических методах дискриминации.

В судебно-медицинской антропологии применяются все основные направления дискриминантного анализа: биномиальная дискриминация, линейный дискриминантный анализ Фишера и канонический дискриминантный анализ.

5.2. ДИСКРИМИНАНТНЫЙ АНАЛИЗ ПРИ БИНОМИАЛЬНОЙ КЛАССИФИКАЦИИ НА ОСНОВЕ ГРУППОВЫХ ЦЕНТРОИДОВ

Дискриминантный анализ на основе групповых центроидов, являясь инструментом реализации многомерной биномиальной классификации, позволяет разграничить принадлежность идентифицируемых объектов по множеству признаков к одной из двух взаимоисключающих групп и в судебно-медицинских антропологических исследованиях применяется практически только в целях разработки способов идентификации пола. Именно это, вероятно, является причиной того, что данный алгоритм дискриминантного анализа по частоте уступает линейному анализу Фишера (p = 0,046). Например, по выборочным данным, биномиальная дискриминация на основе групповых центроидов применялась лишь в 3 (27%) исследований, в то время как линейный дискриминантный анализ Фишера – в 8 (73%).

В терминах математической статистики задачу многомерной биномиальной классификации на основе дискриминантного анализа можно изложить следующим образом [26].

Пусть имеются две генеральные совокупности X и Y, имеющие n - мерный нормальный закон распределения с неизвестными, но равными ковариационными матрицами. Из них взяты обучающие выборки объемами n_1 у X и n_2 у Y:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \; Y = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1k} \\ y_{21} & y_{22} & \cdots & y_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nk} \end{pmatrix}.$$

Целью классификации является отнесение новых наблюдений, гипотетическая совокупность которых представлена матрицей Z, либо к X, либо к Y (единичное наблюдение представлено строкой матрицы Z).

$$Z = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1k} \\ z_{21} & z_{22} & \cdots & z_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nk} \end{pmatrix}.$$

Тогда алгоритм многомерной биномиальной дискриминации включает выполнение следующих этапов:

1. Определяют векторы выборочных средних

$$\overline{X} = \begin{pmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ x_k \end{pmatrix} \text{ и } \overline{Y} = \begin{pmatrix} \overline{y}_1 \\ \overline{y}_2 \\ \vdots \\ \overline{y}_k \end{pmatrix}.$$

2. От матриц X и Y переходят к матрицам центрированных величин X_C и Y_C путем определения $x_{ij} - \overline{x}_j$ и $y_{ij} - \overline{y}_j$.

3. Определяют оценки ковариационных матриц S_x и S_y:

$$S_x = \frac{1}{n_1} X_C^T X_C;\ S_y = \frac{1}{n_2} Y_C^T Y_C.$$

4. Получают несмещенную оценку суммарной ковариационной матрицы

$$S = \frac{1}{n_1 + n_2 - 2}(n_1 S_x + n_2 S_y).$$

5. Определяют матрицу S^{-1}, обратную к S.

6. Вычисляют вектор оценок коэффициентов дискриминантной функции $a = S^{-1}(\overline{X} - \overline{Y})$.

7. Рассчитывают оценки векторов значений дискриминантной функции для матриц исходных данных $U_x = Xa$, $U_y = Ya$.

8. Вычисляют средние значения оценок дискриминантной функции (групповые центроиды)

$$\overline{u}_x = \frac{1}{n_1}\sum_{i=1}^{n_1} u_{xi}\ ,\ \overline{u}_y = \frac{1}{n_2}\sum_{i=1}^{n_2} u_{yi}\ .$$

9. Определяют константу $C = \frac{1}{2}(\overline{u}_x + \overline{u}_y)$.

Дискриминантную функцию для нового наблюдения, подлежащего дискриминации, получают, решив уравнение

$$u_z = z_1 a_1 + z_2 a_2 + \ldots + z_k a_k.$$

При $u_z \geq C$ новое наблюдение, подлежащее дискриминации, относят к совокупности X, при $u_z < C$ - к совокупности Y.

В целях демонстрации практического использования изложенного алгоритма приводим следующий пример.

Известно, что для любого плода переход на внеутробное существование сопровождается изменениями со стороны всех систем органов. Особенно ярко указанные изменения проявляются у недо-

ношенных новорожденных, у которых развивается широкий спектр реакций адаптационного и патологического характера [11]. Одной из таких реакций у недошенных новорожденных является первоначальная гиперплазия экстрамедуллярной кроветворной ткани, происходящая в течение первых часов после рождения. Это делает потенциально возможной идентификацию живорожденности недоношенных плодов по степени кроветворной активности печени в зависимости от ее соответствия гестационной норме. Иными словами данная проблема формулируется как задача классификации недоношенных новорожденных и недоношенных мертворожденных плодов по группе биометрических показателей, включающей линейно-весовые показатели развития плода и показатель кроветворной активности печени.

Пусть априорная информация о распределении показателей дискриминируемых совокупностей представлена случайными выборками из них объемом $n_1 = n_2 = 5$ (табл. 33). Нужно идентифицировать принадлежность к одной из двух анализируемых групп трупа недоношенного ребенка массой 980 г, длиной 35 см и показателем кроветворной активности печени, равным 53,2[11].

Таблица 33

Исходные данные о распределении биометрических показателей

Классифицируемая группа	Масса, г	Длина тела, см	Кроветворная активность, число ядер
Недоношенные новорожденные, прожившие менее одних суток после родов (*X*)	995	34	68,1
	730	31	83,1
	1720	42	48,3
	2550	48	35,3
	1430	39	81,7
Недоношенные мертворожденные плоды (*Y*)	980	35	36,6
	720	26	43,7
	770	34	45,4
	775	34	46,9
	750	30	49,1

Решение задачи дискриминации:

1. Запишем исходные данные в виде матриц *X* и *Y*:

[11] Все примеры главы 5 сконструированы для демонстрации описываемых методов дискриминантного анализа и не отражают реальных результатов соответствующих исследований (подробнее см. [11,53]).

$$X = \begin{pmatrix} 995 & 34 & 68{,}1 \\ 730 & 31 & 83{,}1 \\ 1720 & 42 & 48{,}3 \\ 2550 & 48 & 35{,}3 \\ 1430 & 39 & 81{,}7 \end{pmatrix};\ Y = \begin{pmatrix} 980 & 35 & 36{,}6 \\ 720 & 26 & 43{,}7 \\ 770 & 34 & 45{,}4 \\ 775 & 34 & 46{,}9 \\ 750 & 30 & 49{,}1 \end{pmatrix},$$

где $n_1 = n_x = 5$ и $n_2 = n_y = 5$, а строка матрицы Z: $Z^T = (980 \quad 35 \quad 53{,}2)$.

2. Получим векторы средних

$$\overline{X} = \begin{pmatrix} 1485 \\ 38{,}8 \\ 63{,}3 \end{pmatrix};\ \overline{Y} = \begin{pmatrix} 799 \\ 31{,}8 \\ 44{,}3 \end{pmatrix}.$$

3. Определим матрицы центрированных величин X_C и Y_C:

$$X_C = \begin{pmatrix} -490 & -4{,}8 & 4{,}80 \\ -755 & -7{,}8 & 19{,}80 \\ 235 & 3{,}2 & -15{,}00 \\ 1065 & 9{,}2 & -28{,}00 \\ -55 & 0{,}2 & 18{,}40 \end{pmatrix};\ Y_C = \begin{pmatrix} 181 & 3{,}2 & -7{,}74 \\ -79 & -5{,}8 & -0{,}64 \\ -29 & 2{,}2 & 1{,}06 \\ -24 & 2{,}2 & 2{,}56 \\ -49 & -1{,}8 & 4{,}76 \end{pmatrix}.$$

4. Определим оценку ковариационных матриц

$$S_x = \begin{pmatrix} 400520 & 3756{,}0 & 10331{,}60 \\ & 35{,}8 & -95{,}88 \\ & & 352{,}53 \end{pmatrix};\ S_y = \begin{pmatrix} 8564 & 201{,}8 & -335{,}16 \\ & 11{,}4 & -4{,}33 \\ & & 18{,}13 \end{pmatrix}.$$

5. Получим несмещенную оценку суммарной ковариационной матрицы $S = \dfrac{1}{5+5-2}(5S_x + 5S_y)$:

$$S = \begin{pmatrix} 255677{,}5 & 2473{,}625 & -6666{,}7300 \\ & 29{,}450 & -62{,}6325 \\ & & 231{,}6615 \end{pmatrix}.$$

6. Определим обратную матрицу к S:

$$S^{-1} = \begin{pmatrix} 3{,}593 \cdot 10^{-5} & -0{,}001930 & 0{,}000513 \\ & 0{,}183219 & -0{,}005910 \\ & & 0{,}017483 \end{pmatrix}.$$

7. Найдем вектор оценок коэффициентов дискриминации

$$a = S^{-1}(\overline{X} - \overline{Y}) = S^{-1}\begin{pmatrix} 686 \\ 7{,}0 \\ 18{,}96 \end{pmatrix} = \begin{pmatrix} 0{,}02088957 \\ -0{,}1513922 \\ 0{,}642070195 \end{pmatrix}.$$

8. Вычислим оценки дискриминантной функции

$$U_x = Xa = \begin{pmatrix} 59{,}36276719 \\ 63{,}91226075 \\ 60{,}58357775 \\ 68{,}66665485 \\ 76{,}42492364 \end{pmatrix}; \; U_y = Ya = \begin{pmatrix} 38{,}67282032 \\ 39{,}16276041 \\ 40{,}08762060 \\ 41{,}15517374 \\ 42{,}65105773 \end{pmatrix}.$$

9. Определим средние значения оценок дискриминантной функции (групповые центроиды)

$$\overline{u}_x = 65{,}79003684; \; \overline{u}_y = 40{,}34588656.$$

10. Получим константу

$$C = \frac{1}{2}(65{,}79003684 + 40{,}34588656) = 53{,}067962.$$

11. Определим возможность отнесения нового наблюдения к кластеру недоношенных живорожденных плодов:

$$u_z = Z^T a = 49{,}33119.$$

Поскольку $u_z < C$, то на основании данного метода недоношенный плод массой 980 г, длиной 35 см и показателем кроветворной активности печени 53,2 не может быть признан живорожденным.

На практике в связи с повышенными требованиями к достоверности экспертных выводов результаты многомерной биномиальной классификации идентифицируемых объектов, как правило, формулируются следующим образом: при $u_z > u_x$ - новое наблюдение практически достоверно относится к совокупности X; при $u_z < u_y$ - достоверно относится к совокупности Y; при $C < u_z < u_x$ - вероятно относится к совокупности X; при $u_y < u_z < C$ - вероятно относится к совокупности Y [см. напр. 17].

5.3. ЛИНЕЙНЫЙ ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Линейный дискриминантный анализ Фишера позволяет разграничить принадлежность идентифицируемых объектов по множеству признаков к одной из нескольких (двух и более) взаимоисключающих групп. В этой связи в судебно-антропологических исследованиях линейный дискриминантный анализ чаще всего используется в целях создания способов судебно-медицинской идентификации соматотипа человека, отличающихся лишь характером идентифицируемых объектов и, соответственно, набором идентифицирующих признаков. В последние годы различными авторами на основе линейного дискриминантного анализа создан широкий спектр способов идентификации соматотипа по различным объектам: черепу [35,46], костям стопы [38] и кисти [40], длинным костям конечностей [19]. Еще одним полиномиальным параметром, идентифицируемым с помощью линейного дискриминантного анализа, является порядковая локализация множественных однотипных костей кисти [39] и стопы [27].

В терминах математической статистики задачу многомерной полиномиальной классификации на основе линейного дискриминантного анализа Фишера можно изложить следующим образом.

Пусть имеются k генеральных совокупностей $X_1, X_2, \ldots, X_k$, имеющие n - мерный нормальный закон распределения с неизвестными, но равными ковариационными матрицами. Из них взяты обучающие выборки объемами $n_1, n_2, \ldots, n_k$. Целью классификации является отнесение новых наблюдений, гипотетическая совокупность которых представлена матрицей X, к какому либо из классов X_i. В этом случае алгоритм многомерной классификации включает выполнение следующих этапов:

1. Определяют векторы выборочных средних $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_k$.
2. Определяют оценки ковариационных матриц $S_{X_1}, S_{X_2}, \ldots, S_{Xk}$.
3. Получают несмещенную оценку суммарной ковариационной матрицы

$$S = \frac{1}{n_1 + n_2 + \cdots + n_k - k}(n_1 S_{X_1} + n_2 S_{X_2} + \cdots + n_k S_{Xk}).$$

4. Определяют матрицу S^{-1}, обратную к S.

5. Вычисляют векторы оценок коэффициентов функций классификации $a_i = S^{-1}\overline{X}_i$.

6. Рассчитывают константы функций классификации

$$C_i = \frac{1}{2}\overline{X}_i^T a_i .$$

7. Новое наблюдение, представленное вектором X, относится к тому классу i, для которого линейная функция

$$h_i(x) = (X_i^T a_i) - C_i = \max .$$

Функции классификации не следует путать с дискриминирующими функциями. Функции классификации предназначены для определения того, к какой группе наиболее вероятно может быть отнесен каждый объект идентификации. Имеется столько же функций классификации, сколько существует групп объектов.

Для демонстрации практического применения изложенного алгоритма дискриминантного анализа продолжим разбор примера из раздела 5.2.

После первоначальной гиперпластической реакции закономерной стадией адаптивного ответа недоношенного новорожденного на его переход на внеутробное существование является акциденталь ная инволюция экстрамедуллярной кроветворной ткани [11,53]. Опустошение миелоидной ткани развивается в среднем через сутки после рождения. Данное обстоятельство делает потенциально возможной идентификацию мертворожденных недоношенных плодов и недоношенных новорожденных, проживших менее и более одних суток. Эту проблему можно сформулировать как задачу трехмерной трехгрупповой классификации недоношенных плодов и новорожденных.

Пусть априорная информация о распределении биометрических показателей в совокупностях недоношенных мертворожденных плодов и недоношенных новорожденных, проживших менее одних суток после родов, представлена теми же случайными выборками объемом по 5 наблюдений (см. табл. 33). При этом матрицу X обозначим как X_1, а матрицу Y – как X_2. Априорная информация о распределении этих же показателей в совокупности недоношенных новорожденных, проживших более одних суток, представлена дополнительной выборкой аналогичного объема, в матричной форме имеющей вид

$$X_3 = \begin{pmatrix} 1640 & 39 & 28{,}6 \\ 1910 & 41 & 44{,}5 \\ 2205 & 47 & 16{,}5 \\ 2345 & 45 & 17{,}1 \\ 2650 & 50 & 15{,}6 \end{pmatrix}.$$

Задачей классификации является определение принадлежности недоношенного плода массой 980 г, длиной 35 см и показателем кроветворной активности 53,2 к одной из трех генеральных совокупностей: X_1, X_2 или X_3.

Решение задачи трехгрупповой дискриминации:

1. Определим векторы выборочных средних

$$\overline{X}_1 = \begin{pmatrix} 1485 \\ 38{,}8 \\ 63{,}3 \end{pmatrix};\ X_2 = \begin{pmatrix} 799 \\ 31{,}8 \\ 44{,}3 \end{pmatrix},\ X_3 = \begin{pmatrix} 2150 \\ 44{,}4 \\ 24{,}46 \end{pmatrix}.$$

2. Рассчитаем оценки ковариационных матриц

$$S_{X_1} = \begin{pmatrix} 400520 & 3756{,}0 & 10331{,}6 \\ & 35{,}8 & -95{,}88 \\ & & 352{,}53 \end{pmatrix};$$

$$S_{X_2} = \begin{pmatrix} 8564 & 201{,}8 & -335{,}16 \\ & 11{,}4 & -4{,}33 \\ & & 18{,}13 \end{pmatrix};$$

$$S_{X_3} = \begin{pmatrix} 121750 & 1326 & -2644{,}8 \\ & 15{,}84 & -33{,}044 \\ & & 122{,}9544 \end{pmatrix}.$$

3. Получим несмещенную оценку суммарной ковариационной матрицы $S = \dfrac{1}{5+5+5-3}(5S_{X_1} + 5S_{X_2} + 5S_{X_3})$:

$$S = \begin{pmatrix} 221180{,}8 & 2201{,}583 & -5546{,}48 \\ & 26{,}23333 & -55{,}5233 \\ & & 205{,}672 \end{pmatrix}.$$

4. Определим обратную матрицу к S:

$$S^{-1} = \begin{pmatrix} 3{,}637 \cdot 10^{-5} & -0{,}00228 & 0{,}000366 \\ & 0{,}2316 & 0{,}001095 \\ & & 0{,}015024 \end{pmatrix}.$$

5. Найдем векторы оценок коэффициентов функций классификации

$$a_1 = \begin{pmatrix} -0{,}01122 \\ 5{,}67285 \\ 1{,}53677 \end{pmatrix};\ a_2 = \begin{pmatrix} -0{,}02716 \\ 5{,}59347 \\ 0{,}99329 \end{pmatrix};\ a_3 = \begin{pmatrix} -0{,}01400 \\ 5{,}41251 \\ 1{,}20267 \end{pmatrix}.$$

6. Вычислим константы функций классификации

$$C_1 = 150{,}3649;\ C_2 = 100{,}1088 \text{ и } C_3 = 119{,}8214.$$

7. Определим групповую принадлежность нового наблюдения $X^T = (980 \quad 35 \quad 53{,}2)$:

$$h_1(x) = 119{,}1785;\ h_2(x) = 122{,}0138;\ h_3(x) = 119{,}9132.$$

Поскольку $h_2(x) = \max$, то на основании данного метода недоношенный плод массой 980 г, длиной 35 см и показателем кроветворной активности печени 53,2 должен быть отнесен к группе мертворожденных.

Для оценки надежности экспертных выводов рекомендуется дополнять результаты изложенного алгоритма дискриминантного анализа определением функции *PI* [35,38]. Указанная процедура предусматривает ранжирование полученных значений $h_i(x)$: $h_1(x) < h_2(x) < \ldots < h_{k-1}(x) < h_k(x)$, где $h_k(x) = \max$, и вычисление $I = h_k(x) - h_{k-1}(x)$. Значения функции *PI* затем определяются по специальной таблице (табл. 34).

Таблица 34

Значения *PI* в зависимости от величины *I* [35,38]

I	0	1	2	3	4	5	6	7	8	9
0	0,500	0,525	0,550	0,574	0,599	0,622	0,646	0,668	0,690	0,711
1	0,731	0,750	0,768	0,786	0,802	0,818	0,832	0,846	0,858	0,870
2	0,881	0,891	0,900	0,909	0,917	0,924	0,931	0,937	0,943	0,948
3	0,953	0,957	0,961	0,934	0,968	0,971	0,973	0,976	0,978	0,980
4	0,982	0,984	0,985	0,987	0,988	0,989	0,990	0,991	0,992	0,993
5	0,9933	0,9939	0,9945	0,9950	0,9955	0,9959	0,9963	0,9967	0,9970	0,9975
6	0,9975	0,9978	0,9980	0,9982	0,9984	0,9986	0,9987	0,9988	0,9989	0,9990

В зависимости от величины PI формулируются следующие экспертные выводы:

- при $1{,}0 > PI \geq 0{,}95$ – решение достоверное;
- при $0{,}95 > PI \geq 0{,}75$ – решение вероятное;
- при $PI < 0{,}75$ – отказ от решения.

В нашем примере $I = h_1(x) - h_3(x) = 2{,}1$. По таблице 34 находим, что $PI = 0{,}891$. Поскольку $0{,}95 > 0{,}891 > 0{,}75$, то формулируется вероятный вывод о принадлежности недоношенного плода с данными биометрическими показателями к группе мертворожденных.

5.4. КАНОНИЧЕСКИЙ ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Задача многомерной полиномиальной классификации может быть также решена с помощью канонического дискриминантного анализа. В этой связи в судебно-медицинских антропологических исследованиях канонический дискриминантный анализ применяется в тех же целях, что и линейный анализ [46]. Вместе с тем частота использования канонического дискриминантного анализа значительно меньше таковой линейного анализа Фишера ($p = 0{,}018$). Так, по выборочным данным, каноническая дискриминация применялась лишь в 2 (18%) исследований, в то время как линейный дискриминантный анализ Фишера – в 8 (73%). Кроме того, во всех случаях использования канонического дискриминантного анализа, параллельно также применялся и его линейный аналог.

С вычислительной точки зрения канонический дискриминантный анализ основывается на анализе канонических корреляций, которые определяют последовательные канонические корни и функции [13,26,100]. Максимальное число канонических функций будет на 1 меньше числа дискриминируемых совокупностей.

Не вдаваясь в технологию вычислений, приведем полученные с помощью специальных программных приложений канонические функции для примеров из предыдущих разделов.

Кроме изложенного в разделе 5.2 способа биномиальную дискриминацию недоношенных новорожденных и недоношенных мертворожденных плодов можно также провести на основе канонической функции

$$u_z = Z^T a_K + C,$$

где $C = 10{,}52055$;

$$a_K = \begin{pmatrix} -0{,}00414 \\ 0{,}03001 \\ -0{,}12729 \end{pmatrix};$$

с групповыми центроидами $\bar{u}_x = -2{,}52211$ и $\bar{u}_y = 2{,}52211$. Распределение канонических оценок функции приведено на рисунке 35.

Трехгрупповую дискриминацию недоношенных новорожденных и мертворожденных плодов помимо трех линейных (см. раздел 5.3) можно осуществить с помощью двух канонических функций

$$h_1 = 0{,}003484 \cdot m + 0{,}013483 \cdot l + 0{,}115038 \cdot a - 10{,}731931;$$

$$h_2 = 0{,}001922 \cdot m - 0{,}100346 \cdot l - 0{,}041169 \cdot a + 2{,}817979,$$

где m масса тела, г; l – длина тела, см; a – кроветворная активность печени, количество ядер в тестовой площади.

Групповые центроиды канонических функций для каждой дискриминируемой группы приведены в таблице 35. Как видно, функция h_1 преимущественно дискриминирует новорожденных, проживших менее 1 суток, а функция h_2 – разделяет оставшуюся совокупность объектов на два других кластера (рис. 36).

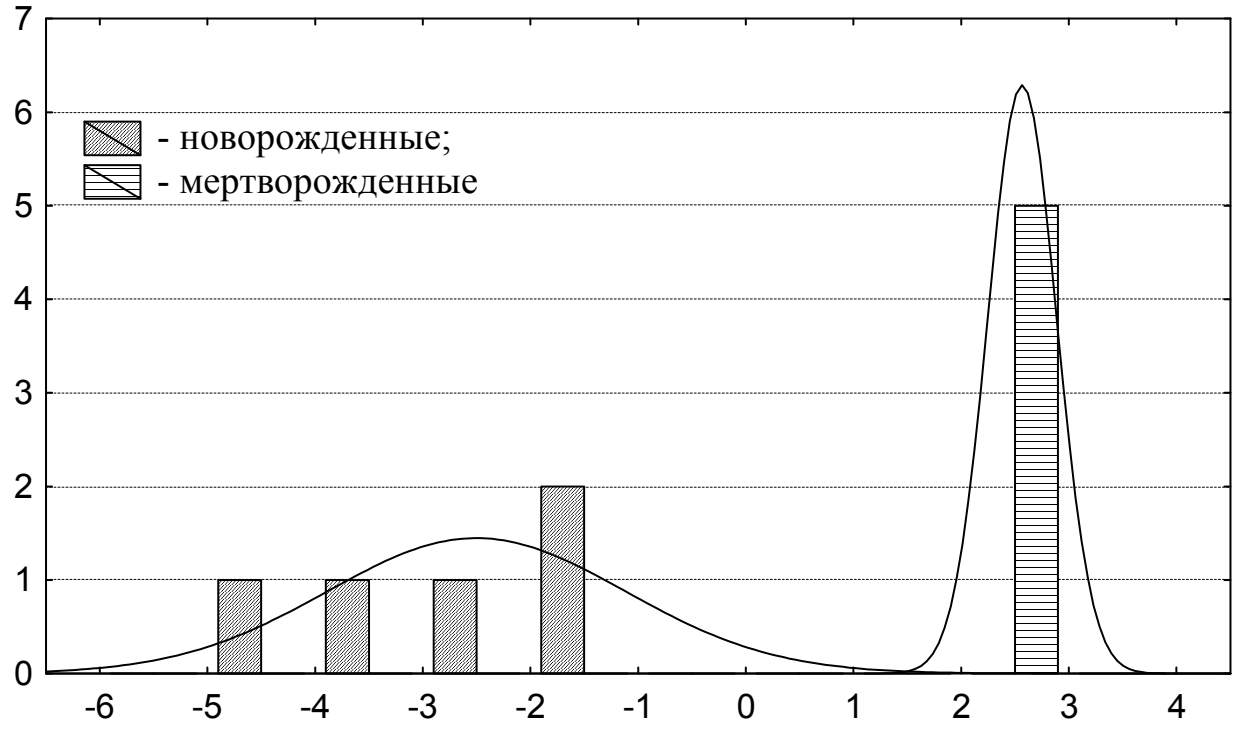

Рис. 35. Распределение канонических оценок дискриминируемых совокупностей плодов и новорожденных. По оси абсцисс – каноническая оценка, по оси ординат – частота.

Таблица 35

Средние значения канонических функций групп дискриминации

Дискриминируемые группы	h_1	h_2
Новорожденные, прожившие < 1 суток	2,24708	-0,826559
Мертворожденные плоды	-2,41855	-0,662394
Новорожденные, прожившие > 1 суток	0,17147	1,488953

Недостатками канонического дискриминантного анализа являются трудоемкость и сложность интерпретации результатов при полиномиальной дискриминации. В то же время, по результатам сравнительного анализа, проведенного рядом авторов, точность идентификации на основе канонического анализа примерно совпадает с таковой линейного дискриминантного анализа Фишера [46]. При этом последний, благодаря функции *PI*, обеспечивает более объективные идентификационные выводы. Данное обстоятельство делает линейный дискриминантный анализ методом выбора при проведении судебно-медицинских антропологических исследований, посвященных проблеме классификации идентифицируемых объектов.

Рис. 36. Канонические оценки дискриминируемых совокупностей новорожденных и мертворожденных плодов. По оси абсцисс – оценка функции h_1, по оси ординат – функции h_2.

5.5. ОПТИМАЛЬНАЯ СТРАТЕГИЯ ДИСКРИМИНАНТНОГО АНАЛИЗА

Исчерпывающий алгоритм дискриминантного анализа включает оценку критериев качества построенной дискриминантной модели [13,122]. Наиболее распространенными критериями являются:

- λ дискриминантной модели в целом;
- λ каждой из переменных дискриминантной модели;
- частная статистика λ каждой из переменных модели;
- F-статистика дискриминантной модели в целом;
- F-статистики каждой из переменных дискриминантной модели;
- толерантность каждой переменной дискриминантной модели.

Критерий λ определяется как 1 минус квадрат канонической корреляции и называется также *Wilks'*λ. Квадрат канонической корреляции является оценкой доли дисперсии, общей между двумя каноническими переменными, поэтому *Wilks'*λ равняется оценке необъясненной доли дисперсии. *Wilks'*λ используется в качестве статистики критерия значимости квадрата канонической корреляции и имеет χ^2 - распределение [26]. Для *Wilks'*λ дискриминантной модели в целом и каждой из входящих в ее состав переменных разработаны соответствующие F-аппроксимации [143].

Частная статистика λ каждой из переменных дискриминантной модели вычисляется как отношение λ модели после включения переменной в ее состав к значению λ модели без соответствующей переменной в ее составе. Для частных статистик λ также разработаны формулы вычисления F-аппроксимаций [71,143].

Еще одной важной мерой, влияющей на классификацию отдельных наблюдений, является расстояние Махаланобиса. В общем случае расстояние Махаланобиса расценивается как расстояние между двумя точками в пространстве, определяемом двумя или более коррелированными переменными. Например, при двух или трех некоррелированных переменных расстояние Махаланобиса между точками равно расстоянию Евклида, может быть представлено графически и непосредственно измерено (подробнее см. раздел 6.3). При коррелированности переменных оси на графике могут рассматриваться как неортогональные (т.е. направленными не под прямыми углами друг к другу). В этом случае определение дистанции Евклида является неадекватным и требуется использование других частных случаев расстояния Махаланобиса [13].

Для каждой совокупности в выборке можно определить положение точек, представляющих средние для всех переменных в многомерном пространстве, определенном переменными рассматриваемой дискриминантной модели. Эти точки называются групповыми центроидами. Полная процедура дискриминантного анализа предусматривает для каждого наблюдения тестовой выборки вычисление расстояния Махаланобиса до каждого центроида группы. Наблюдение признается принадлежащим к той группе, для которой минимально расстояние Махаланобиса до соответствующего группового центроида [13].

Некоторые критерии качества канонического дискриминантного анализа рассмотрены также в работах R. Barcikowski, J.P. Stevens [89] и C.J. Huberty [122].

Как и любой другой статистический метод дискриминантный анализ основан на определенных допущениях и может быть применен только при условии хотя бы приближенного соответствия исследуемых биометрических данных этим допущениям. Использование дискриминантного анализа при невыполнении указанного условия может привести к ошибочным результатам.

Главное предположение дискриминантного анализа касается того, что анализируемые данные должны представлять собой выборку из многомерного нормального распределения. Поэтому перед применением дискриминантного анализа следует проверить, являются ли данные нормально распределенными. Показано, однако, что отклонения от нормальности обычно не влияют сильно на достоверность *F*-критериев значимости дискриминантных функций [13].

Вторым важным предположением считается однородность ковариационных матриц исследуемых показателей. Различия между совокупностями должны касаться лишь векторов средних. Наличие этих различий и служит основой успешной классификации. Данное условие корректности применения дискриминантного анализа определяет необходимость доказательства равенства ковариационных матриц исследуемых совокупностей и различий их векторов средних. Считается, что небольшие отклонения от однородности ковариационных матриц не являются препятствием для проведения дискриминации [13].

Другое условие адекватного применения дискриминантного анализа заключается в том, что переменные, используемые для дис-

криминации между совокупностями, не должны быть полностью избыточными. Понятие избыточности в данном случае аналогично понятию мультиколлинеарности в многофакторном регрессионном анализе [149]. Как было показано ранее, при вычислении результатов дискриминантного анализа происходит обращение суммарной ковариационной матрицы для переменных в дискриминантной модели. Если одна из переменных полностью избыточна по отношению к другим переменным, то такая матрица называется плохо обусловленной и не может быть обращена.

Следует отметить, что большинство серьезных угроз корректности *F*-критериев значимости дискриминантных моделей возникает из-за возможной зависимости между средними по совокупностям и дисперсиями между собой, поскольку соответствующие критерии для относительно больших средних с большими дисперсиями будут ошибочно значимыми [13]. Наличие корреляции между средними и дисперсиями вообще характерно для биометрических данных [53]. Схожая проблема существует и в регрессионном анализе, в котором изначальная неоднородность дисперсий признаков в последующем приводит к неоднородности дисперсии остатков регрессионной модели (гетероскедастичности).

Таким образом, выполнимость данных предположений является условием корректности результатов дискриминантного анализа.

Так же как и для многофакторной регрессии, целью дискриминантного анализа является включение в модель тех переменных, которые наилучшим образом дискриминируют совокупности между собой. Для этого предусмотрены пошаговые процедуры с включением и исключением переменных.

В пошаговом анализе с включением на каждом этапе дискриминантного анализа просматриваются все переменные, и находится та из них, которая вносит наибольший вклад в различие между совокупностями. Эта переменная включается в модель на данном шаге, и происходит переход к следующему шагу. В пошаговом анализе с исключением все переменные сначала включаются в дискриминантную модель, а затем на каждом шаге устраняются те из них, которые вносят малый вклад в задачу классификации. Определяющим фактором для включения или исключения переменной из модели является значение соответствующей *F*-статистики. Это значение является мерой вклада переменной в предсказание членства в совокупности. На наш взгляд, более удобным является использова-

ние алгоритма пошагового дискриминантного анализа с исключением.

Для демонстрации пошагового дискриминантного анализа с исключением продолжим обсуждение идентификации живорожденности недоношенных плодов по степени кроветворной активности паренхимы печени в зависимости от ее соответствия гестационной норме (см. раздел 5.2). Исходное множество анализируемых биометрических параметров включает показатели массы и длины плода, а также показатель кроветворной активности печени.

С вычислительной точки зрения для двух классифицируемых групп дискриминантный анализ может рассматриваться как процедура, аналогичная множественной регрессии. Это объясняется тем, что одно из подмножеств случайных величин проводимого канонического анализа содержит только одну переменную. В этом случае каноническая корреляция является единственной и тождественна коэффициенту множественной корреляции дискриминируемого параметра с дискриминирующими признаками.

Например, если кодировать классифицируемые группы недоношенных новорожденных и мертворожденных плодов как 0 и 1, а затем использовать полученную дихотомическую переменную в качестве зависимой в модели множественной регрессии, то полученные результаты будут аналогичны результатам дискриминантного анализа. В частности, полученный коэффициент множественной детерминации будет равен квадрату канонической корреляции, *F*-статистика значимости регрессионной модели в целом будет аналогична таковой дискриминантной модели, а *t* - статистики регрессионных коэффициентов - *F*- статистикам переменных дискриминантной функции. Отсюда легко определить *Wilks'*λ дискриминантной модели в целом:

$$\lambda = 1 - r^2 .$$

*Wilks'*λ для каждой из исходного множества дискриминирующих переменных можно найти, вычисляя коэффициенты множественной детерминации дискриминируемого параметра с вектором дискриминирующих показателей за исключением анализируемой переменной. В этом случае *Wilks'*λ для каждой переменной, входящих на первом шаге в состав дискриминантной модели, будет также равняться разности 1 и значения соответствующего коэффициента множественной детерминации. Частные статистики λ каждой из переменных дискриминантной модели достаточно просто вычислить

как отношение λ исходной дискриминантной модели к значению λ соответствующей переменной.

Отсюда получаем оценки исходной трехмерной дискриминантной модели: $Wilks'\lambda = 0{,}11272$; $F_{3;6} = 15{,}903$; $p < 0{,}0029$. Остальные оценки, касающиеся каждой из входящих в состав дискриминантной модели переменных, приведены в таблице 36.

Таблица 36

Критерии качества трехмерной дискриминантной модели

Переменные	*Wilks'*λ	*Partial* λ	*F*	*p*
Масса	0,194	0,576	4,416	0,080
Длина	0,112	0,996	0,026	0,876
Кроветворная активность	0,6312	0,177	27,935	0,002

Полученные результаты свидетельствует об отсутствии дискриминирующего эффекта у показателя длины тела. Исключение данной переменной привело к улучшению оценок двумерной модели в целом: $Wilks'\lambda = 0{,}11221$; $F_{2;7} = 27{,}693$; $p < 0{,}0005$, а также оценок оставшихся в составе модели переменных, которая уже может использоваться в целях идентификации недоношенных новорожденных и недоношенных мертворожденных плодов (табл. 37).

Таблица 37

Критерии качества двумерной дискриминантной модели

Переменные	*Wilks'*λ	*Partial* λ	*F*	*p*
Масса	0,673	0,167	35,013	0,00059
Кроветворная активность	0,635	0,177	32,605	0,00073

Таким образом, дискриминантный анализ, при условии корректности его применения, является мощным средством для проведения многомерной биномиальной и полиномиальной классификации биометрических данных. Относительная простота и нетрудоемкость его алгоритмов характеризуют дискриминантный анализ как метод выбора при проведении многомерной классификации в судебно-медицинских антропологических исследованиях.

5.6. ТЕСТИРОВАНИЕ ТОЧНОСТИ СУДЕБНО-МЕДИЦИНСКОЙ АНТРОПОЛОГИЧЕСКОЙ ИДЕНТИФИКАЦИИ

Главной проблемой дискриминантного анализа после установления дискриминирующей функции, является вопрос о точности классификации. В математической статистике считается, что оценивание качества классификации нельзя проводить по той же самой выборке, по которой были оценены классифицирующие или дискриминирующая функции. Исключение составляют лишь так называемые *V*-кратная и глобальная кросс-проверки, применяющиеся при небольших объемах выборок при построении деревьев классификации [13]. Поэтому для установления точности классификации следует классифицировать только те наблюдения, которые не использовались при оценке функции дискриминации. Классификация же старых наблюдений может быть использована лишь для диагностики наличия выбросов или областей, где функция классификации кажется менее адекватной [13].

Изложенное требование вызывает необходимость деления имеющихся наблюдений на две выборки: обучающую и тестовую. При этом уменьшение объема данных, использованных при установлении дискриминирующей функции, в итоге приведет к снижению точности классификации по сравнению с той же функцией, рассчитанной по полному набору данных. Указанное обстоятельство принуждает многих судебных медиков, занимающихся проблемами идентификации личности, либо вообще отказаться от выделения тестовой выборки [17], либо, что более правильно, осуществлять тестирование точности классификации по объединенной совокупности данных обучающей и тестовой групп [41].

Между тем, оба названных метода тестирования точности классификации являются нарушением предпосылок адекватности результатов дискриминантного анализа. Использование этих и подобных подходов при проведении судебно-медицинских антропологических исследований чревато искажением итогов тестирования со смещением показателей точности классификации в сторону излишней оптимистичности.

В этой связи единственным допустимым методом тестирования точности идентификации, на наш взгляд, является рандомизированное разделение эмпирических данных на две выборки: обучающую и тестовую. Доли указанных выборок в общей совокупности

данных не обязательно должны равняться 2/3 и 1/3 и могут изменяться в зависимости от задач конкретного исследования. Наш собственный опыт применения дискриминантного анализа показывает, что уменьшение объема обучающей выборки незначительно меняя оценки коэффициентов дискриминации, не всегда сопровождается потерями в точности классификации. Вместе с тем синхронное увеличение объема тестовой выборки зачастую позволяет рассчитать достаточно узкие доверительные интервалы для оценок точности тестирования, что имеет большое практическое значение.

Следует отметить, что тестовая выборка должна быть не только случайной, но и стратифицированной. Иными словами, она должна содержать примерно в равных соотношениях объекты из каждой идентифицируемой группы. В противном случае, если одна из дискриминируемых групп (страта) будет избыточно или, наоборот, недостаточно представлена в тестовой выборке, подобное смещение может привнести дополнительные искажения в результаты тестирования точности классификации [20,74].

Следующей проблемой тестирования точности классификации в судебно-медицинских антропологических исследованиях является вопрос о критериях точности. В качестве критерия точности идентификации в настоящее время в судебной антропологии рассматривается лишь один показатель - доля случаев ошибочной классификации объектов тестовой выборки [17,46]. Однако указанный показатель отражает лишь точность идентификации в целом, не показывая, что означает результат классификации для отдельного идентифицируемого объекта.

Названный недостаток тестирования точности идентификации является потенциально устранимым при условии перехода на общепринятые критерии оценки результатов прикладных биомедицинских исследований. К основным характеристикам результатов этих исследований относятся: чувствительность (sensitivity/*Se*), специфичность (specificity/*Sp*), прогностическая ценность положительного (positive predictive value/*PPV*) и отрицательного (negative predictive value /*NPV*) результатов, индекс точности (accuracy/*Ac*) и отношение правдоподобия положительного результата (likelihood ratio of a positive test/*LRPT*) [123,124,137,147].

Изложенное позволило нам предложить группу критериев точности идентификации, адаптированных применительно к потребностям судебно-медицинской экспертной практики и лишенных не-

достатков, присущих общепринятым критериям [58]. Алгоритмы определения указанных критериев точности идентификации различаются в зависимости от уровней категоричности экспертных суждений о принадлежности идентифицируемых объектов к определенному кластеру, в соответствии с которыми экспертные суждения подразделяются на достоверные, вероятные и неопределенные.

Алгоритм тестирования точности идентификации без выделения степеней категоричности экспертных суждений осуществляется последовательно на пяти этапах:

1. Формируют рандомизированную стратифицированную тестовую выборку идентифицируемых объектов, не использовавшихся при разработке способа дискриминации, с известной принадлежностью каждого объекта к определенному идентифицируемому кластеру.

2. С помощью тестируемой дискриминантной модели осуществляют классификацию объектов тестовой выборки по их принадлежности к одному из кластеров.

3. По результатам классификации определяют следующие исходные параметры тестирования точности идентификации: k – количество кластеров, подлежащих дискриминации; x_i - доля истинных результатов идентификации принадлежности объектов к i-му кластеру; ε_{ij} - доля ошибочных результатов идентификации принадлежности объектов i-го кластера к j-му кластеру; ε_{ji} - доля ошибочных результатов идентификации принадлежности объектов j-го кластера к i-му кластеру. Тогда весь спектр возможных результатов тестирования будет представлен в виде матрицы

$$X = \begin{pmatrix} x_1 & \varepsilon_{12} & \cdots & \varepsilon_{1k} \\ \varepsilon_{21} & x_2 & \cdots & \varepsilon_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ \varepsilon_{k1} & \varepsilon_{k2} & \cdots & x_k \end{pmatrix}.$$

Каждая строка матрицы X содержит набор возможных результатов идентификации объектов i-го кластера, каждый столбец – набор возможных результатов идентификации принадлежности объектов к i-му кластеру. При этом должно выполняться условие

$$\begin{pmatrix} x_1 \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{k1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{12} \\ x_2 \\ \vdots \\ \varepsilon_{k2} \end{pmatrix} + \cdots + \begin{pmatrix} \varepsilon_{1k} \\ \varepsilon_{2k} \\ \vdots \\ x_k \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

4. По исходным параметрам тестирования определяют следующие критерии точности идентификации:

Чувствительность идентификации объектов i-го кластера – вероятность того, что объект *i*-го кластера будет правильно классифицирован:

$$Se_i = x_i .$$

Специфичность идентификации объектов i-го кластера – вероятность того, что объекты, не принадлежащие *i*-му кластеру, будут идентифицированы как таковые:

$$Sp_i = \frac{k - 1 - \sum_{i=1}^{k} \varepsilon_{ji}}{k - 1}.$$

Прогностическая ценность положительного результата идентификации объекта i-го кластера – вероятность правильности идентификации принадлежности объекта к *i*-му кластеру:

$$PPV_i = \frac{x_i}{x_i + \sum_{i=1}^{k} \varepsilon_{ji}}.$$

Средняя чувствительность идентификации – среднее значений чувствительности идентификации объектов каждого кластера:

$$Se = \sum_{i=1}^{k} x_i / k .$$

Отношение правдоподобия положительного результата идентификации объекта i-го кластера – насколько более вероятно то, что объект *i*-го кластера будет идентифицирован, как принадлежащий *i*-му кластеру, по сравнению с объектом, принадлежащим любому кластеру, отличному от *i*-го:

$$LRPT_i = Se_i /(1 - Sp_i).$$

Практическое использование приведенного алгоритма можно показать на примере тестирования точности краниометрической идентификации соматотипа мужчин по данным Н.В. Нариной и В.Н. Звягина (табл. 38).

Таблица 38

Итоги тестирования точности краниометрической идентификации соматотипа мужчин [46]

Действительная группа	Предсказанная группа, %			Всего
	1	2	3	
1. Грудной тип	62,5	12,5	25	100
2. Мускульный тип	19,3	64,5	16,1	100
3. Брюшной тип	0	25	75	100

Вычисленные выборочные оценки критериев точности краниометрической идентификации соматотипа мужчин приведены в таблице 39.

Таблица 39

Оценки точности краниометрической идентификации соматотипа

Классифицируемый соматотип	Se_i	Sp_i	PPV_i	$LRPT_i$
1. Грудной	0,625	0,903	0,764	6,5
2. Мускульный	0,646	0,813	0,632	3,4
3. Брюшной	0,750	0,794	0,646	3,6

Сравнение итогов двух процедур тестирования показывает, что приведенные указанными авторами критерии точности соответствуют лишь чувствительности идентификации определенного соматотипа и не отражают прогностическую ценность конкретного результата идентификации. Например, по данным таблицы 39 можно утверждать, что в случае идентификации грудного типа телосложения вероятность истинности классификации в среднем составляет 76,4%, при этом вероятность правильной идентификации указанного соматотипа в 6,5 раз больше вероятности его ложноположительной идентификации. Показательно, что результат идентификации грудного соматотипа, характеризующегося наименьшей чувствительностью, в то же время имеет наибольшую прогностическую ценность. Это объясняется тем, что идентификация данного соматотипа обладает наивысшей специфичностью (см. табл. 39).

Алгоритм тестирования точности идентификации с выделением степеней категоричности экспертных суждений включает выполнение двух первых аналогичных и следующих отличающихся этапов:

3. По результатам классификации определяют следующие исходные параметры тестирования точности идентификации: a_i - доля истинных результатов идентификации принадлежности объектов к i-му кластеру в достоверном диапазоне экспертных суждений; b_i – доля истинных результатов идентификации принадлежности объектов к i-му кластеру в вероятном диапазоне экспертных суждений; c_i – доля объектов i-го кластера, принадлежность которых не идентифицирована; ε_{ij}^{a} - доля ошибочных результатов идентификации принадлежности объектов i-го кластера к j-му кластеру в достоверном диапазоне экспертных суждений; ε_{ij}^{b} - доля ошибочных результатов идентификации принадлежности объектов i-го кластера к j-му кластеру в вероятном диапазоне экспертных суждений; ε_{ji}^{a} - доля ошибочных результатов идентификации принадлежности объектов j-го кластера к i-му кластеру в достоверном диапазоне экспертных суждений; ε_{ji}^{b} - доля ошибочных результатов идентификации принадлежности объектов j-го кластера к i-му кластеру в вероятном диапазоне экспертных суждений.

Тогда весь спектр возможных результатов тестирования будет представлен в виде суммы матриц

$$\begin{pmatrix} a_1 & \varepsilon_{12}^{a} & \cdots & \varepsilon_{1k}^{a} \\ \varepsilon_{21}^{a} & a_2 & \cdots & \varepsilon_{2k}^{a} \\ \vdots & \vdots & \vdots & \vdots \\ \varepsilon_{k1}^{a} & \varepsilon_{k2}^{a} & \cdots & a_k \end{pmatrix} + \begin{pmatrix} b_1 & \varepsilon_{12}^{b} & \cdots & \varepsilon_{1k}^{b} \\ \varepsilon_{21}^{b} & b_2 & \cdots & \varepsilon_{2k}^{b} \\ \vdots & \vdots & \vdots & \vdots \\ \varepsilon_{k1}^{b} & \varepsilon_{k2}^{b} & \cdots & b_k \end{pmatrix} + \begin{pmatrix} c_1 & 0 & \cdots & 0 \\ 0 & c_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & c_k \end{pmatrix}.$$

4. По исходным параметрам тестирования определяют следующие критерии точности идентификации:

Чувствительность достоверной идентификации объектов i-го кластера – вероятность правильной идентификации объектов i-го кластера в достоверном диапазоне экспертных суждений:

$$SeA_i = a_i.$$

Общая чувствительность идентификации объектов i-го кластера – вероятность правильной идентификации объектов i-го кластера в достоверном или вероятном диапазоне экспертных суждений:

$$Se_i = a_i + b_i .$$

Прогностическая ценность достоверной идентификации объекта i-го кластера – вероятность истинности идентификации объекта *i*-го кластера в достоверном диапазоне экспертных выводов:

$$PPVA_i = \frac{a_i}{a_i + \sum_{i=1}^{k} \varepsilon_{ji}^{a}} .$$

Прогностическая ценность вероятной идентификации объекта i-го кластера – вероятность истинности идентификации объекта *i*-го кластера в вероятностном диапазоне экспертных суждений:

$$PPVP_i = \frac{b_i}{b_i + \sum_{i=1}^{k} \varepsilon_{ji}^{b}} .$$

Средняя чувствительность достоверной идентификации – среднее значений чувствительности достоверной идентификации:

$$SeA = \sum_{i=1}^{k} a_i / k .$$

Средняя общая чувствительность идентификации – среднее значений общей чувствительности идентификации:

$$Se = \sum_{i=1}^{k} (a_i + b_i) / k .$$

Доля идентифицированных объектов i-го кластера – доля результатов идентификации объектов *i*-го кластера, попавших в достоверный или вероятный диапазоны экспертных суждений, по отношению ко всем результатам идентификации объектов данного кластера:

$$\Delta_i = 1 - c_i .$$

Общая доля идентифицированных объектов - доля результатов идентификации, попавших в достоверный или вероятный диапазоны экспертных суждений, по отношению ко всем результатам идентификации:

$$\Delta = k - \sum_{i=1}^{k} c_i .$$

Практическое использование данного алгоритма можно показать на примере тестирования точности идентификации пола по рентгенограммам кисти, проведенного Н.Н. Гончаровой, О.В. Самоходской, М.В. Федуловой и др. [17]. По данным указанного коллектива

авторов при тестировании 726 объектов мужской и 1003 объектов женской половой принадлежности достоверная диагностика мужского пола имела место в 369 случаях, вероятная – в 343 случаях, достоверная диагностика женского пола – в 501 случае, вероятная – в 516 случаях. При этом ошибки классификации наблюдались в 1 случае достоверной диагностики мужского и 11 случаях - женского пола, в 65 случаях вероятной диагностики мужского и 69 случаях - женского пола.

Вычисленные оценки критериев точности идентификации пола для данной дискриминантной модели приведены в таблице 40, из которой видно, что идентификация пола в зависимости от уровня категоричности экспертных суждений имеет различную прогностическую ценность, неодинаковую для каждого дискриминируемого кластера.

Таблица 40

Оценки точности идентификации пола по рентгенограммам кисти

Пол	SeA_i	Se_i	$PPVA_i$	$PPVP_i$	SeA	Se	Δ_i	Δ
Мужской	0,507	0,890	0,971	0,801	0,498	0,912	1	1
Женский	0,489	0,934	0,998	0,873			1	

Таким образом, предложенные критерии в полном объеме характеризуют точность любых способов идентификации объектов экспертного познания, позволяя ранжировать альтернативные дискриминантные модели по степени их диагностической значимости. Значительным преимуществом указанных критериев перед их аналогами, используемыми в клинических диагностических исследованиях, является независимость результатов теста от степени распространенности идентифицируемых признаков в тестовых выборках и человеческих популяциях. Дело в том, что в отличие от чувствительности и специфичности, положительная или отрицательная прогностическая ценность сильно зависят от распространенности идентифицируемых кластеров [20,144]. Однако данный недостаток легко устраним при использовании для расчетов не абсолютных, а относительных частотных показателей.

Еще одной важной проблемой тестирования точности идентификации является то, что простое определение доли правильно классифицированных объектов тестовой выборки дает лишь ориенти-

ровочное представление о точности классификации. В связи с этим для достоверного суждения о качестве классификации необходимо вычисление интервальных оценок долей правильной и ошибочной классификации наблюдений тестовой выборки. Данная рекомендация согласуется с общей тенденцией к расширению показаний к применению доверительных интервалов в биомедицинских исследованиях вплоть до замены ими обычных статистических критериев [146].

Например, в рассмотренной в разделе 4.2 работе, посвященной проблеме идентификации половой принадлежности подъязычной кости, ее авторы помимо метода определения пола, основанного на одномерной биномиальной классификации, применили также и дискриминантный анализ [41]. Проведенная данными авторами кросс-проверка, основанная на тестировании объединенных выборок 107 мужских и 51 женской подъязычных костей, показала, что точность классификации мужчин составляет 87,85%, женщин – 98,04%. Не вдаваясь в обсуждение обоснованности объединения обучающей и тестовой групп, заметим, что полученные числа являются лишь точечными оценками неизвестных истинных значений долей случаев правильной классификации пола при использовании полученной дискриминирующей функции в будущем. Так, вычисленные нами интервальные оценки показывают, например, что с 95% вероятностью истинные значения чувствительности идентификации мужчин находятся в пределах 81,38-92,66%, женщин – 90,27-99,99%.

Изложенное позволяет утверждать, что определение интервальных оценок точности классификации должно стать обязательным компонентом любой программы тестирования точности способов судебно-медицинской идентификации. К сожалению, определение интервальных оценок возможно не для всех предложенных нами критериев точности идентификации, в частности оно является затруднительным для наиболее ценного критерия точности – прогностической ценности положительного (отрицательного) результата идентификации. Однако названный недостаток для линейных дискриминантных моделей в определенной степени устраним применением функции *PI* [35,38], а для моделей биномиальной дискриминации – с помощью нижеизложенного метода.

Суть указанного метода сводится к тому, что, как и при одномерной биномиальной классификации, недостатком результатов

биномиальной многомерной классификации является формулирование экспертных выводов в терминах ранговой шкалы, а не в количественной вероятностной форме. В связи с этим судебные медики пытались модифицировать алгоритм дискриминантного анализа с целью устранения данного недостатка. В частности, такой модификацией можно назвать алгоритм дискриминантного анализа, использованный В.А. Клевно при идентификации половой принадлежности грудной клетки [44]. Большой заслугой указанного автора является предложение определять количественные вероятностные выводы о половой принадлежности идентифицируемых объектов на основе выборочных оценок параметров распределений значений дискриминантной функции для каждой из двух альтернативных совокупностей объектов. Однако нельзя не заметить, что недостатками данного подхода являются оценивание значений дискриминирующей функции по данным обучающей выборки и использование точечных оценок параметров распределений значений дискриминантной функции.

Для устранения указанных недостатков целесообразно использовать основные принципы изложенного в разделе 4.2 метода одномерной биномиальной классификации. С учетом данных принципов алгоритм многомерной биномиальной классификации на основе дискриминантного анализа следует дополнить следующими этапами.

1. Осуществляют кросс-проверку с использованием только лишь данных тестовой выборки и фиксируют полученные для каждой из дискриминируемых совокупностей значения дискриминантной функции u_z. Важно, что распределение значений u_Z для каждой дискриминируемой совокупности будет нормальным.

2. Используя (34) и (35), для каждой дискриминируемой совокупности объектов тестовой выборки определяют наихудшие с позиции точности классификации интервальные оценки параметров распределения значений u_z.

3. Строят упорядоченный ряд дискретных значений u_z: $u_z^1 > u_z^2 > \ldots u_z^{n-1} > u_z^n$, причем $u_z^1 \geq \bar{x}_2 + 3s_2$, $u_z^n \leq \bar{x}_1 - 3s_1$, а $u_z^i - u_z^{i+1} = \varepsilon$, где ε – последняя значащая цифра в результате вычисления u_z.

4. Используя наихудшие для точности классификации интервальные оценки параметров распределения, дважды нормируют каждое значение u_Z^i упорядоченного ряда $u_z^1 > u_z^2 > \ldots\ u_z^{n-1} > u_z^n$:

$$z_2^i = \frac{u_z^i - \mu_2}{s_2} \text{ и } z_1^i = \frac{u_z^i - \mu_1}{s_1}.$$

5. Используя (36), для каждого значения z определяют функцию плотности стандартного нормального распределения, после чего строят два ряда значений $f(z_2^i)$ и $1 - f(z_1^i)$.

6. С помощью (37) для каждого u_z^i определяют вероятность принадлежности к альтернативным дискриминируемым группам p_1 и p_2. Полученные данные табулируют или представляют в форме набора номограмм.

Таким образом, изложенная модель многомерной биномиальной классификации на основе дискриминантного анализа позволяет с надежностью не менее $1 - 2\alpha$ формулировать точные количественные вероятностные суждения о половой принадлежности классифицируемых объектов, повышая тем самым объективность и достоверность экспертных выводов. Указанные вероятностные суждения являются ни чем иным, как наиболее важным критерием точности судебно-медицинской классификации – показателем прогностической ценности положительного результата идентификации.

ГЛАВА 6. КЛАСТЕРНЫЙ АНАЛИЗ

6.1. ЗНАЧЕНИЕ КЛАСТЕРНОГО АНАЛИЗА В СУДЕБНО-МЕДИЦИНСКИХ АНТРОПОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

Кластерный анализ объединяет группу многомерных статистических методов, предназначенных для решения задач разбиения конечных совокупностей объектов, каждый из которых характеризуется одинаковым числом признаков, на однородные группы, количество которых заранее неизвестно. От других методов многомерной классификации кластерный анализ отличается отсутствием обучающих выборок, т.е. априорной информации о распределении классификационных признаков.

По сравнению с другими методами многомерной классификации кластерный анализ редко используется не только в судебной антропологии, но и в судебно-медицинских исследованиях любой другой тематики. Например, из 84 оригинальных исследований, опубликованных журналом «Судебно-медицинская экспертиза» в течение 2001-2005 гг., результаты которых основывались на использовании методов аналитической статистики, кластерный анализ был применен лишь в 2 (2,4%) статьях.

Редкость применения кластеризации объясняется тем, что в отличие от многих других статистических процедур, методы кластерного анализа используются в эксплораторной (разведочной) стадии исследования при отсутствии каких-либо априорных гипотез относительно кластерообразующих параметров и количества кластеров. Вследствие этого проверка статистической значимости к результатам кластерного анализа неприменима. Кроме того, несмотря на свою относительную элементарность, методы кластерного анализа, требующие большого числа арифметических и логических операций, стали доступными для практического использования только в последние годы в связи с бурным развитием вычислительной техники. Однако даже автоматизированная вычислительная реализация громоздких алгоритмов кластерного анализа при числе наблюдений, большем нескольких сотен, до сих пор нецелесообразна, а в ряде случаев и невозможна [26]. Вместе с тем, кластерный анализ представляет собой мощное средство выявления латентного кластеринга при использовании других статистических методов в судебно-медицинской антропологии.

6.2. ФОРМЫ ПРЕДСТАВЛЕНИЯ ИСХОДНЫХ ДАННЫХ В АЛГОРИТМАХ КЛАСТЕРНОГО АНАЛИЗА

В задачах кластерного анализа обычной формой представления исходных данных служит прямоугольная таблица, каждая строка которой представляет результат измерения k рассматриваемых признаков на одном из n обследованных объектов:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}.$$

В зависимости от конкретных задач исследования кластерный анализ может быть использован не только в целях группировки объектов, но и для группировки признаков. Существуют также методы, предназначенные для одновременной кластеризации объектов и признаков (так называемое двухвходовое объединение) [13]. В сравнении с другими методами кластерного анализа двухвходовое объединение является, вероятно, наименее часто используемым методом.

Числовые значения, входящие в матрицу X, могут соответствовать трем типам переменных: количественным, порядковым и качественным. В отличие от других статистических методов тип переменных влияет лишь на выбор конкретного алгоритма кластеризации (точнее, на выбор меры расстояний или близости между объектами), но не на ее допустимость. Поскольку теоретические основы кластерного анализа при использовании смешанных типов переменных исследованы недостаточно, при практической реализации кластерных алгоритмов необходимо, чтобы исходные данные соответствовали одному типу переменных. В противном случае разные типы переменных требуется свести к какому-либо одному типу с помощью ранжирования (для количественных переменных) или кодирования дихотомическими переменными (для качественных признаков) [26]. Альтернативными путями также являются выбор меры расстояния между объектами, предназначенной для смешанных типов переменных или отказ от использования переменных разных типов [29]. В большинстве случаев недопустимым также является использование для кластеризации переменных одного типа, но выраженных в разных единицах измерения.

6.3. РАССТОЯНИЕ МЕЖДУ ОБЪЕКТАМИ И МЕРА БЛИЗОСТИ

Наиболее трудным и наименее формализованным в задаче классификации является определение понятия однородности объектов. Чаще всего понятие однородности объектов задается введением правила вычислений расстояния $\rho(X_i, X_j)$ между любой парой исследуемых объектов [26]. Если задана функция $\rho(X_i, X_j)$, то близкие с точки зрения этой метрики объекты считаются однородными, принадлежащими одному классу. К часто используемым в задачах кластерного анализа мерам расстояний относятся следующие [13,26,29].

1. Обычное евклидово расстояние. Это наиболее часто используемая мера расстояния. Она является геометрическим расстоянием в многомерном пространстве и вычисляется следующим образом:

$$\rho_E(X_i, X_j) = \sqrt{\sum_{l=1}^{k} (x_{il} - x_{jl})^2},$$

где x_{il}, x_{jl} - величина l-й компоненты у i-го (j-го) объекта $(l = 1,2,\ldots,k; i, j = 1,2,\ldots,n)$.

Использование евклидова расстояния оправдано в случаях, если:

а) наблюдения берутся из совокупностей, имеющих многомерное нормальное распределение с одинаковыми ковариационными матрицами;

б) исследуемые признаки однородны по физическому смыслу и одинаково важны для классификации;

в) признаковое пространство совпадает с геометрическим пространством.

С геометрической точки зрения и содержательной интерпретации евклидово расстояние может оказаться бессмысленным, если его признаки имеют разные единицы измерения. Анализ биомедицинских исследований показывает, что зачастую евклидово расстояние применяется некорректно, в частности, для исследования неколичественных признаков либо количественных признаков, выраженных в разных единицах измерения [29]. Аналогичная ситуация имеет место и в судебно-медицинских исследованиях [12].

Для приведения количественных признаков к одинаковым единицам измерения необходимо производить нормирование каждого признака с последующим переходом от матрицы X к нормированной матрице с элементами

$$x_{il}^{H} = \frac{x_{il} - \bar{x}_l}{s_l},$$

где x_{il}^{H} - значение l-го признака у i-го объекта; $\bar{x}_l$ - среднее значение l-го признака; s_l - стандартное отклонение l-го признака.

Однако эта операция может привести к нежелательным последствиям. Если кластеры хорошо разделены по одному признаку и не разделены по другому, то после нормировки дискриминирующие возможности первого признака будут уменьшены в связи с увеличением «шумового» эффекта второго [26].

2. Взвешенное евклидово расстояние применяется в случаях, когда каждой компоненте x_l вектора наблюдений X удается приписать некоторый «вес» w_l, пропорциональный степени важности признака в задаче классификации:

$$\rho_{BE}(X_i, X_j) = \sqrt{\sum_{l=1}^{k} w_l (x_{il} - x_{jl})^2}.$$

Обычно принимают $0 \le w_i \le l$, где $l = 1,2,\ldots,k$.

3. Квадрат евклидова расстояния применяется с целью придать большие веса более отдаленным друг от друга объектам:

$$\rho_{E^2}(X_i, X_j) = \sum_{l=1}^{k} (x_{il} - x_{jl})^2.$$

4. Расстояние городских кварталов (манхэттенское расстояние) вычисляется по формуле:

$$\rho_M(X_i, X_j) = \sum_{l=1}^{k} \left| x_{il} - x_{jl} \right|.$$

Это расстояние является средним разностей по координатам. Данная метрика может использоваться для анализа порядковых и качественных переменных. В большинстве случаев эта мера расстояния приводит к таким же результатам, как и обычное евклидово расстояние. Следует отметить, что для этой меры влияние отдельных больших разностей (выбросов) уменьшается, так как они не возводятся в квадрат.

5. Расстояние Чебышева используется в том случае, когда нужно определить два объекта как "различные", если они различаются по какой-либо одной координате (каким-либо одним измерением):

$$\rho_Ч(X_i, X_j) = \max_l \left| x_{il} - x_{jl} \right|.$$

6. Степенное расстояние применяется, если необходимо прогрессивно увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются:

$$\rho_C(X_i, X_j) = \left(\sum_{l=1}^{k} \left| x_{il} - x_{jl} \right|^p \right)^{1/r}.$$

где r и p - параметры, определяемые исследователем. Параметр p ответственен за постепенное взвешивание разностей по отдельным координатам, параметр r ответственен за прогрессивное взвешивание больших расстояний между объектами. Если оба параметра - r и p, равны двум, то степенное расстояние совпадает с расстоянием Евклида.

7. Хеммингово расстояние используется как мера различия объектов, задаваемых переменными качественного типа. Хеммингово расстояние равно числу несовпадений значений соответствующих признаков в рассматриваемых i-м и j-м объектах:

$$\rho_H(X_i, X_j) = \sum_{l=1}^{k} \left| x_{il} - x_{jl} \right|.$$

Если кластерный анализ используется в целях группировки признаков, то исходные данные также могут быть представлены в виде матрицы расстояний X. В этом случае принципиальных различий между кластеризацией объектов и признаков нет. Однако при кластеризации признаков матрица X не является единственным способом представления исходных данных, последняя также может быть задана в виде квадратной матрицы

$$R = (r_{ij}),\ i, j = 1,2,\ldots,k,$$

элемент r_{ij} которой определяет степень близости i-го объекта к j-му (кластерный анализ «рассматривает» признак как объект). В этом случае различие между объектами и признаками является существенным, и понятие однородности объектов определяется заданием некоторой функции $r(X_i, X_j)$, характеризующей степень близости i-го и j-го объектов.

При задании меры близости $r(X_i, X_j)$ надо помнить о необходимости выполнения условий симметрии $r(X_i, X_j) = r(X_j, X_i)$; максимального сходства объекта с самим собой $r(X_i, X_i) = \max_j r(X_j, X_i)$ при $1 \le j \le n$; и монотонного убывания

$r(X_i, X_j)$ по $\rho(X_i, X_j)$, т.е. из $\rho(X_k, X_l) \geq \rho(X_i, X_j)$ должно следовать неравенство $r(X_k, X_l) \leq r(X_i, X_j)$ [26].

При кластеризации признаков мерами близости служат различные статистические коэффициенты связи. При количественном типе признаков можно использовать оценки парных коэффициентов корреляции Пирсона. При нелинейных зависимостях рекомендуется применение корреляционных отношений или преобразований шкал признаков. Существуют также различные коэффициенты связи, определенные для порядковых и качественных признаков.

Выбор метрики является узловым моментом исследования, от которого в основном зависит окончательный вариант разбиения объектов на классы при данном алгоритме кластеризации. В каждом конкретном случае этот выбор должен производиться по-своему в зависимости от типа и физической природы исследуемых данных, целей исследования, априорных сведений о характере вероятностного распределения *X*. Основными факторами, определяющими выбор метрики или меры близости, являются характер задачи кластеризации (объекты или признаки) и тип эмпирических данных (табл. 41).

Таблица 41

Зависимость мер расстояний или близости от типа данных

Тип признаков	Меры расстояний	Меры близостей
Количественный	ρ_E; ρ_{BE}; ρ_{E^2}; $\rho_Ч$; ρ_C	r; 1- r
Порядковый	ρ_M	r_S; r_K
Качественный и дихотомный	ρ_M ; ρ_H ; процент несогласия	Квадрантная корреляция; χ^2
Смешанный	-	Меры Журавлева, Гауэра, Воронина, Миркина

Изменение метрики, чаще всего (но не всегда) ведет к изменению классификации. Например, использование ненормированного евклидова расстояния в качестве метрики для кластеризации дихотомных или ранговых переменных, может привести к такому же результату, как и применение хеммингова расстояния [48]. Однако, поскольку результаты классификации считаются окончательными, при проведении кластерного анализа всегда следует тщательно обосновывать выбор меры расстояний или близости.

6.4. РАССТОЯНИЕ МЕЖДУ КЛАСТЕРАМИ

На первом шаге классификации, когда каждый объект представляет собой отдельный кластер, расстояния между этими объектами определяются выбранной мерой. Для связи двух кластеров, каждый из которых состоит из нескольких объектов, используют понятие расстояния между группами объектов и меры близости двух групп объектов. Наиболее употребительными расстояниями и мерами близости между группами объектов являются следующие [13,26,29,158].

1. Расстояние, измеряемое по методу «ближайшего соседа». В этом методе расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах. Данное правило строит "волокнистые" кластеры, т.е. кластеры, "сцепленные вместе" только отдельными элементами, случайно оказавшимися ближе остальных друг к другу. Метод также называется методом одиночной связи.

2. Расстояние, измеряемое по методу «дальнего соседа». В этом методе расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т.е. "наиболее удаленными соседями"). Использование этого правила определения расстояния между группами объектов является непригодным, если кластеры имеют удлиненную форму или их естественный тип является "цепочечным". Метод называется также методом полной связи.

3. Расстояние, измеряемое по методу невзвешенного попарного среднего. В этом методе расстояние между двумя различными кластерами вычисляется как среднее расстояние между всеми парами объектов в них. Метод одинаково эффективен при любой форме кластеров. В англоязычной научной литературе и статистических программных пакетах метод именуется аббревиатурой *UPGMA* (от англ. unweighted pair-group method using arithmetic averages - метод невзвешенного попарного арифметического среднего) [13].

4. Расстояние, измеряемое по методу взвешенного попарного среднего. Метод идентичен методу невзвешенного попарного среднего, за исключением того, что при вычислениях размер соответствующих кластеров (т.е. число объектов, содержащихся в них) используется в качестве весового коэффициента. Поэтому предлагаемый метод должен быть использован в случаях, когда предполага-

ются неравные размеры кластеров. Метод известен также под аббревиатурой *WPGMA* (от англ. weighted pair-group method using arithmetic averages - метод взвешенного попарного арифметического среднего) [13].

5. Расстояние, измеряемое с помощью невзвешенного центроидного метода. Здесь расстояние между двумя кластерами определяется как расстояние между их центрами тяжести. Для обозначения метода применяется также аббревиатура *UPGMC* (от англ. unweighted pair-group method using the centroid average - метод невзвешенного попарного центроидного усреднения) [13].

6. Расстояние, измеряемое с помощью взвешенного центроидного метода. Метод идентичен предыдущему, за исключением того, что при вычислениях используются веса для учёта разницы между размерами кластеров (т.е. числами объектов в них). Метод оказывается предпочтительнее предыдущего, если имеются (или подозреваются) значительные отличия в размерах кластеров. В статистических программных пакетах метод известен под аббревиатурой *WPGMC* (от англ. weighted pair-group method using the centroid average - метод взвешенного попарного центроидного усреднения) [13].

7. Расстояние, измеряемое по методу Варда. Этот метод отличается от всех других методов, поскольку он использует методы дисперсионного анализа для оценки расстояний между кластерами. Метод минимизирует сумму квадратов для любых двух (гипотетических) кластеров, которые могут быть сформированы на каждом шаге [158]. Метод считается очень эффективным при выявлении трудноуловимых различий, однако он стремится создавать кластеры малого размера [13,29].

В целом существует большое количество различных методов разбиения заданной совокупности объектов на кластеры. Поэтому большой интерес представляет задача сравнительного анализа качества этих способов разбиения. С этой целью используется понятие функционала качества разбиения, определенного на множестве всех разбиений. Наилучшим является такое разбиение, при котором достигается экстремум выбранного функционала качества.

В качестве основных характеристик функционала качества разбиения в математической статистике рассматриваются следующие: сумма внутриклассовых дисперсий, сумма попарных внутриклассовых расстояний между объектами, обобщенная внутриклассовая дисперсия и ее модификации [26].

6.5. ИЕРАРХИЧЕСКИЕ КЛАСТЕР-ПРОЦЕДУРЫ

Иерархические кластер-процедуры являются наиболее распространенными алгоритмами кластерного анализа. В иерархических методах выстраивается «дерево» кластеров, положение которых определяется матрицей попарных расстояний или попарных мер близости, а также выбором меры расстояний между кластерами.

Иерархические кластер-процедуры бывают двух типов: агломеративные и дивизимные [126]. В агломеративных процедурах начальным является разбиение, состоящее из n одноэлементных классов, а конечным – из одного класса; в дивизимных – наоборот. Принцип работы иерархических агломеративных (дивизимных) процедур состоит в последовательном объединении (разделении) групп элементов сначала самых близких (далеких), а затем все более отдаленных (близких) друг от друга.

Для демонстрации основных принципов кластеризации продолжим начатое в разделе 2.6 обсуждение латентной неоднородности кроветворной активности печени плодов и новорожденных 25, 28-30 недель гестации (см. рис. 9). Объектами проведенной иерархической классификации являлись 33 плода указанных сроков гестации, которые характеризовались двумя признаками: кроветворной активностью печени и гестационным возрастом (см. рис. 10). Напомним, что данная двумерная совокупность была упорядочена в порядке возрастания значений кроветворной активности. При этом каждому объекту присваивался номер, соответствующий положению показателя кроветворной активности в упорядоченном ряду. В результате проведенного ранжирования всем восьми новорожденным с постнатальной инволюцией экстрамедуллярной миелоидной ткани были присвоены ранги с 1-го по 8-й.

Проведенная иерархическая классификация показала, что предпочтение следует отдать этапу классификации, на котором все наблюдения объединены в два кластера, состоящие из 8-ми и 25-ти объектов (см. рис. 10). Причем кластер, состоящий из 8-ми объектов, был представлен восемью новорожденными с острой постнатальной инволюцией кроветворной ткани печени. Все остальные плоды и новорожденные были выделены во второй кластер, отделенный от первого значительным межкластерным расстоянием ($\rho_{(1-8),(9-33)} = 0{,}590$). Однако в разделе 2.6 не была охарактеризована использованная нами стратегия кластеризации.

Между тем, особенности последней в значительной степени влияют на окончательные результаты кластеризации. Более того, выбор другой метрики или меры расстояния между кластерами может привести к полярным результатам кластеризации одних и тех же эмпирических данных с соответствующими радикальными изменениями выводов экспериментального исследования. Поэтому обсудим подробнее характер использованного нами алгоритма кластерного анализа и возможных альтернативных стратегий кластеризации.

Задачей исследования является кластеризация 33 объектов. Каждый из объектов характеризуется двумя нормально распределенными количественными признаками, выраженными в разных единицах измерения. В этой связи предпочтительным является применение агломеративной иерархической кластер-процедуры с нормированным евклидовым расстоянием в качестве меры расстояния между объектами. В качестве меры расстояния между группами объектов лучше принять *UPGMA*-расстояние.

Результаты данной агломеративной классификации показывают, что предпочтение следует отдать этапу классификации, на котором все наблюдения объединены в четыре или пять кластеров, два из которых представлены новорожденными с постнатальной инволюцией экстрамедуллярной кроветворной ткани (рис. 37). Полученные результаты подтверждают выводы проведенной в разделе 2.6 кластеризации, но не являются тождественными им (ср. с рис. 10). Для наглядности выберем в качестве метрики ненормированное евклидово расстояние. На этот раз результаты кластеризации, отличаясь лишь конкретными числами шкалы различий, в принципиальном отношении практически полностью совпадают с выводами раздела 2.6 (рис. 38).

Таким образом, две использованные стратегии кластеризации подтверждают ранее сделанный вывод о латентной неоднородности кроветворной активности печени плодов и новорожденных за счет присутствия в выборке кластера новорожденных с постнатальной инволюцией экстрамедуллярного гемопоэза. Различия результатов двух указанных стратегий кластеризации связаны с подавлением дискриминирующего эффекта показателя кроветворной активности после нормировки обоих признаков в связи с увеличением «шумового» эффекта показателя гестационного возраста, не обладающего дискриминирующими возможностями.

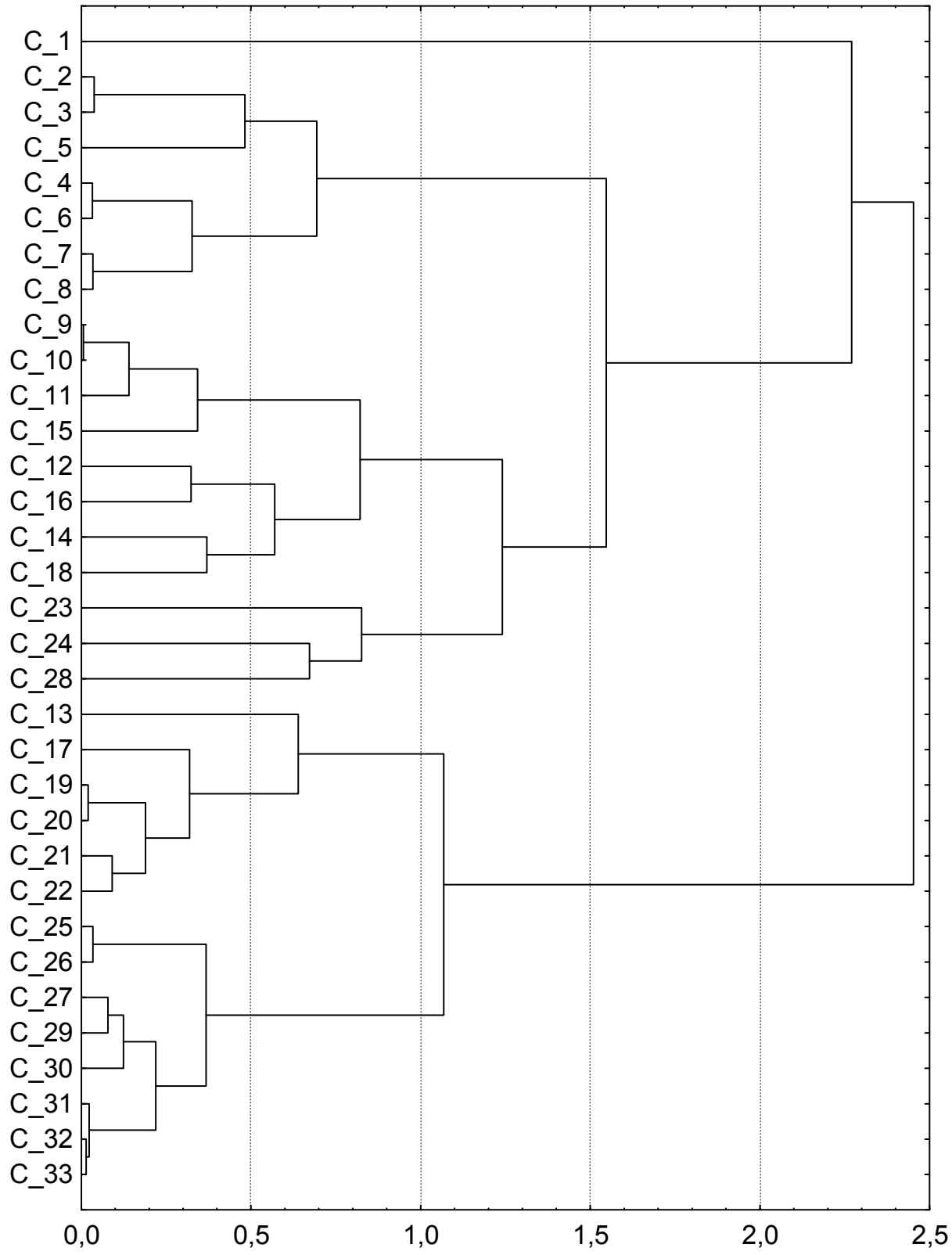

Рис. 37. Дендрограмма агломеративной иерархической классификации плодов и новорожденных 25, 28-30 недель гестации по степени кроветворной активности паренхимы печени и показателю гестационного возраста. Мера дистанции между объектами – нормированное евклидово расстояние, между кластерами – *UPGMA*-расстояние. Здесь и на рис. 38: по оси абсцисс – мера различия; по оси ординат – объекты.

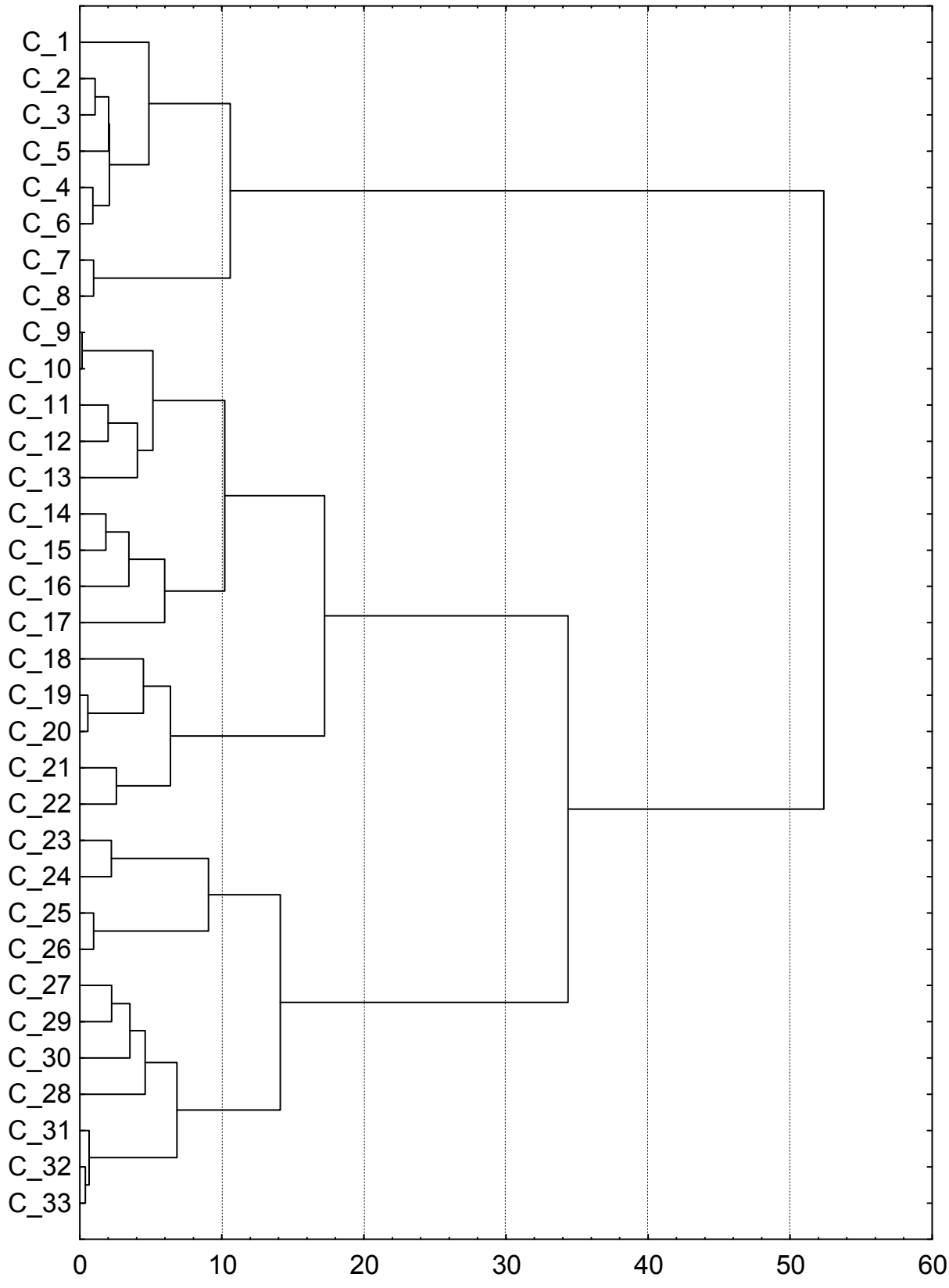

Рис. 38. Дендрограмма агломеративной иерархической классификации плодов и новорожденных 25, 28-30 недель гестации по степени кроветворной активности паренхимы печени и показателю гестационного возраста. Мера дистанции между объектами – ненормированное евклидово расстояние, между кластерами – *UPGMA*-расстояние.

В этом случае закономерен вопрос: с помощью какой же стратегии кластеризации получена дендрограмма, приведенная на рисунке 10? Дело в том, что существует альтернативный метод нормирования количественных признаков, лишенный основного недостатка обычной стандартизации [14]. Алгоритм указанного нормирования заключается в выполнении следующих двух приемов.

1. Производится ранжирование совокупности $x_1, x_2, \ldots, x_n$ выборочных значений каждого признака с построением упорядоченного ряда $x_i < x_{i+1} < \ldots < x_{\max}$.

2. Осуществляется нормировка значений каждого признака по формуле:

$$x_{z_i} = \frac{x_{\max} - x_i}{x_{\max}}.$$

Изложенный подход обеспечивает переход от размерных единиц к безразмерным без потерь в выраженности дискриминирующих эффектов признаков, связанных с их стандартизацией. Подобное нормирование и было использовано при построении дендрограммы, приведенной на рисунке 10. Остальные компоненты стратегии кластеризации те же: мера дистанции между объектами - нормированное евклидово расстояние, между кластерами - *UPGMA*-расстояние.

В качестве примера агломеративной иерархической классификации качественных признаков обсудим проблему обоснования оптимальной дифференциации медико-криминалистического экспертного познания [48].

Дифференциация судебно-медицинского экспертного познания является характерной закономерностью современного этапа развития судебно-медицинской экспертной деятельности. Однако, не будучи обоснованным теоретически, процесс дифференциации экспертного судебно-медицинского познания до сих пор имеет эмпирически-стихийный характер. Этим во многом обусловлено множество процессуальных, организационных, методических и кадровых проблем, существующих в настоящее время в практике назначения, производства и оценки судебно-медицинских экспертиз [7]. Одной из таких проблем является несоответствие между непрерывно возрастающим объемом судебно-медицинских знаний и существующими формально правом и обязанностью судебно-медицинских экспертов, работающих в разных структурных подразделениях уч-

реждений судебно-медицинской экспертизы, выполнять все виды экспертных судебно-медицинских исследований [7,48].

Наибольшей неоднородностью выполняемых видов экспертиз характеризуется деятельность медико-криминалистических отделений учреждений судебно-медицинской экспертизы. Не вызывает сомнения тот факт, что медико-криминалистические экспертизы являются совокупностью, по крайне мере, пяти различных видов судебно-медицинских экспертиз, отличающихся по объектам и предмету экспертных исследований. В качестве указанных основных видов медико-криминалистических экспертиз можно назвать следующие: судебно-медицинские трасологические, судебно-медицинские баллистические экспертизы, судебно-медицинские экспертизы отождествления личности, судебно-медицинские микрологические экспертизы и судебно-медицинские экспертизы реконструкции событий. Однако потенциальная обязанность исполнения всего перечня указанных экспертиз требует от экспертов медико-криминалистических отделений профессионального освоения и поддержания на необходимом уровне большого объема знаний и практических навыков. Причем в условиях постоянного увеличения объема специальных знаний, повышения требований к качеству экспертной деятельности и научной обоснованности экспертных заключений указанная потребность будет только возрастать. Поэтому научное обоснование необходимости и оптимальных пределов дифференциации, а также официального закрепления самостоятельных видов судебно-медицинских медико-криминалистических экспертиз является одной из первоочередных научно-практических задач, определившей цель соответствующего исследования.

В качестве средства достижения поставленной цели был использован кластерный анализ, а именно такой его алгоритм, как иерархическая агломеративная классификация основных видов медико-криминалистических экспертиз. Названная кластер-процедура предусматривала начальное разбиение совокупности видов медико-криминалистических экспертиз на 5 одноэлементных кластеров и последовательное объединение сначала самых близких видов судебно-медицинской экспертизы, а затем все более отдаленных друг от друга. Учитывая, что исследуемые признаки являлись качественными, в качестве меры расстояния между двумя различными видами медико-криминалистических экспертиз использовался процент несогласия. Мера расстояния между кластерами, состоящими

из объединенных видов медико-криминалистических экспертиз, определялась по *UPGMA*-алгоритму.

На первом этапе кластерного анализа исследовалась степень неоднородности медико-криминалистических экспертиз. Признаками, характеризовавшими виды этих экспертиз, являлось наличие или отсутствие использования при выполнении судебно-медицинских исследований данного вида каждого из 22 базовых методов наблюдения и фиксации свойств объектов экспертного познания, подготовительных методов и приемов, методов и приемов моделирования, а также аналитических методов (табл. 42).

Проведенный анализ показал, что по степени сходства наиболее оптимальным пределом дифференциации судебно-медицинских медико-криминалистических экспертиз является их разделение на 4 класса, один из которых представлен совокупностью двух видов экспертиз: трасологическими и баллистическими, а остальные – отдельными видами: отождествления личности, микрологическими и ситуационными (рис. 39).

Далее была определена степень сложности каждого вида медико-криминалистических экспертиз. Для этого каждый из 22 анализированных базовых методов медико-криминалистического познания был ранжирован по частоте использования при различных видах экспертиз. Ранжирование данных осуществлялось согласно общепринятой методике [16]. Наивысший ранг присваивался методам, специфичным лишь для одного вида экспертиз, наименьший – методам, применяющимся при выполнении любых медико-криминалистических экспертиз (см. табл. 42). Затем для каждого вида экспертиз вычислялась сумма рангов. При этом оказалось, что наибольшей сложностью характеризуются экспертизы отождествления личности (сумма рангов - 145), наименьшей – ситуационные (сумма рангов равна 42) и микрологические (сумма рангов равна 84) экспертизы. Сложность трасологических и баллистических экспертиз одинакова и приближается к таковой антропологических экспертиз (сумма рангов – по 143,5).

Полученные результаты позволяют утверждать, что, несмотря на различия по характеру объектов и предмету исследования, трасологические и баллистические экспертизы являются одинаковыми по своей сложности и очень схожими по методам и техническим приемам экспертных исследований, а также по методам анализа их результатов.

Таблица 42

Использование базовых методов исследования и технических приемов при различных видах экспертиз

№	Методы и технические приемы	ТЭ	БЭ	ЭОЛ	МЭ	ЭРС	Ранг
1	Визуальный	1	1	1	1	1	3,5
2	Измерительный	1	1	1	1	1	3,5
3	Фотографический	1	1	1	1	1	3,5
4	Графический	1	1	1	1	1	3,5
5	Рентгенологический	1	1	1	1	0	7
6	Визуальный в ИК и УФ спектрах	1	1	1	0	0	9,5
7	Стереомикроскопический	1	1	0	1	0	9,5
8	Микроскопический	0	0	1	1	0	15,5
9	Микрометрический и стереомикрометрический	0	0	1	1	0	15,5
10	Химический	1	1	0	0	0	15,5
11	Остеометрический	0	0	1	0	0	21
12	Изготовление макропрепаратов	1	1	1	0	0	9,5
13	Реставрация объектов исследования	1	1	1	0	0	9,5
14	Изготовление микропрепаратов и микрошлифов	0	0	1	1	0	15,5
15	Наливка рентгеноконтрастными и красящими веществами	1	1	0	0	0	15,5
16	Изготовление объемных слепков	1	1	0	0	0	15,5
17	Получение экспериментальных следов-повреждений	1	1	0	0	0	15,5
18	Моделирование процессов причинения повреждений при ЭРС	0	0	0	0	1	21
19	Статистический анализ	1	1	1	1	1	3,5
20	Сравнительный анализ	1	1	1	1	1	3,5
21	Векторно-графический анализ	1	1	0	0	0	15,5
22	Методы реконструкции признаков и динамических процессов	0	0	1	0	0	21

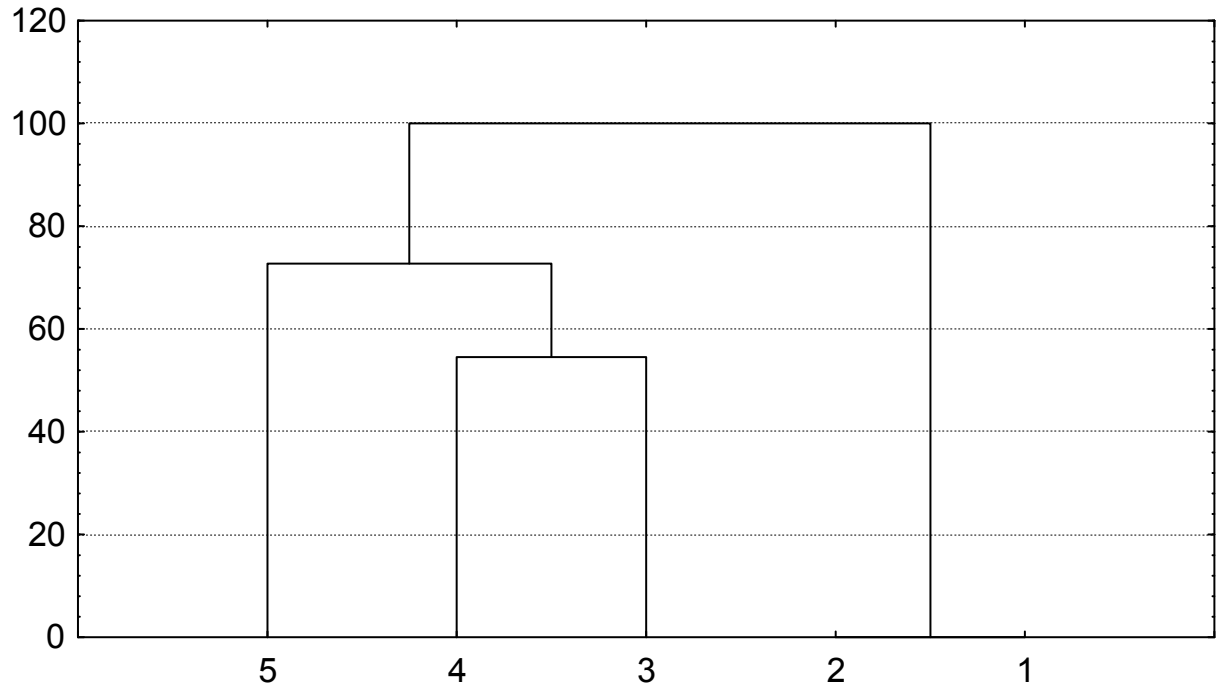

Рис. 39. Дендрограмма агломеративной иерархической классификации медико-криминалистических экспертиз по методам подготовки объектов к исследованию, наблюдения и фиксации свойств объектов, методам моделирования и аналитическим методам. Мера расстояния между объектами – процент несогласия, между кластерами – *UPGMA*-дистанция. По шкале абсцисс – виды судебно-медицинских экспертиз, по шкале ординат – мера различия, %. Здесь и на рис. 40: 1 - трасологические экспертизы; 2 – баллистические экспертизы; 3 – экспертизы отождествления личности; 4 – микрологические экспертизы; 5 – экспертизы реконструкции событий.

Напротив, экспертизы реконструкции событий, резко отличаясь по характеру применяемых методов исследования от остальных видов медико-криминалистических экспертиз, являются наименее сложными. Учитывая, что основными объектами экспертизы реконструкции событий являются материалы законченных трасологических и баллистических экспертиз, а также наибольшую частоту их назначения, целесообразным является выделение указанных трех видов медико-криминалистических экспертиз в группу базовых для подготовки судебно-медицинских экспертов медико-криминалистических отделений учреждений судебно-медицинской экспертизы.

Проведенное исследование также доказало, что экспертизы отождествления личности являются наиболее сложным видом медико-криминалистических экспертиз, отличаясь не только по характеру объектов и предмету исследования, но и по методам и техническим

приемам последнего. Причем основные для экспертиз отождествления личности методы исследования (например, остеометрический и стереомикрометрический) являются специфичными только для данного вида медико-криминалистических экспертиз.

Изложенное определяет необходимость дифференциации и официального закрепления самостоятельности экспертиз отождествления личности с утверждением соответствующих специальных программ подготовки кадров для медико-криминалистических отделений учреждений судебно-медицинской экспертизы.

Характерными особенностями микрологических экспертиз явилась их небольшая сложность на фоне выраженных различий по методам и техническим приемам исследования. В этой связи указанный вид экспертиз хотя и требует определенной дифференциации в подготовке специалистов, но может выполняться любыми судебно-медицинскими экспертами медико-криминалистической специализации независимо от их базовой подготовки, а также судебно-медицинскими экспертами других структурных подразделений учреждений судебно-медицинской экспертизы, например, экспертами судебно-гистологического или судебно-биологического отделений.

Таким образом, кластерный анализ показывает, что оптимальным пределом дифференциации экспертного познания при выполнении медико-криминалистических экспертиз является выделение трех классов. Основной класс, являющийся базовым при подготовке кадров для медико-криминалистических отделений учреждений судебно-медицинской экспертизы, должен быть сформирован из трасологических, баллистических и ситуационных экспертиз.

Второй, высокоспецифичный и сложный класс образуют экспертизы отождествления личности. Достижение высокой степени достоверности результатов и научной обоснованности выводов таких экспертиз на современном этапе невозможны без реализации официального закрепления самостоятельности данного вида экспертиз с утверждением соответствующих специальных программ подготовки медико-криминалистических кадров.

Третий, высокоспецифичный и относительно несложный класс медико-криминалистических экспертиз образуют микрологические экспертизы. Выполнение данного вида экспертиз возможно как любыми экспертами медико-криминалистической специализации

независимо от их базовой подготовки, так и экспертами других специализаций.

6.6. ДВУХВХОДОВОЕ ОБЪЕДИНЕНИЕ

В зависимости от конкретных задач исследования кластерный анализ может быть использован не только в целях группировки объектов или признаков, но также и для одновременной кластеризации объектов и признаков. Данный вид кластеризации является самым редким из существующих кластер-процедур и известен под терминами двухвходовое объединение, двувидовое соединение, двунаправленная ассоциация [13]. Указанную стратегию кластеризации целесообразно применять в обстоятельствах, когда ожидается, что и наблюдения, и переменные одновременно вносят вклад в обнаружение осмысленных кластеров.

Так, возвращаясь к примеру обоснования оптимальной дифференциации медико-криминалистического экспертного познания, можно показать избирательность использования конкретных методов и приемов экспертного познания по отношению к различным видам медико-криминалистических экспертиз (рис. 40). Некоторые исследователи полагают, что двухвходовое объединение представляет собой мощное средство разведочного анализа данных [13].

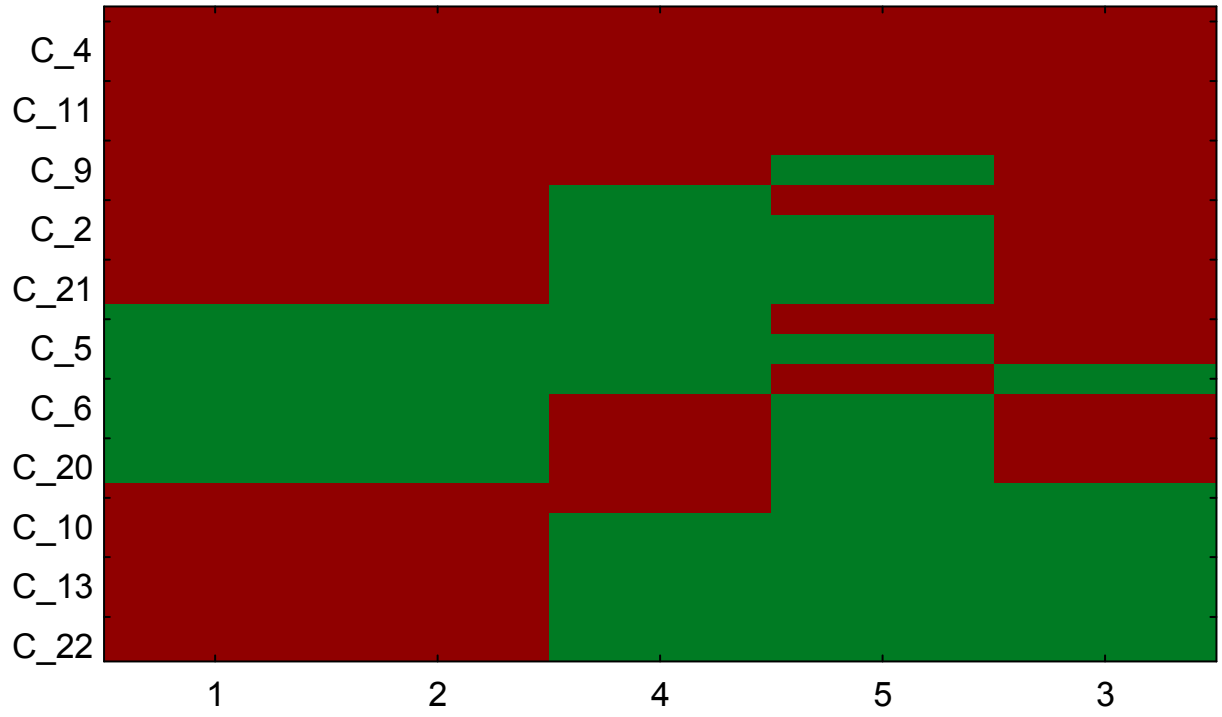

Рис. 40. Итог двухвходового объединения медико-криминалистических экспертиз и методов медико-криминалистического экспертного познания. По шкале абсцисс – виды судебно-медицинских экспертиз, по шкале ординат – нумерация методов.

6.7. МЕТОД *К*-СРЕДНИХ

В зависимости от особенностей технологии различают иерархические и итеративные приемы кластеризации [29]. В отличие от изложенных выше иерархических технологий в итеративных методах разбиение на кластеры ведет к последовательным перерасчетам приближений. Как и иерархические, итеративные методы подразделяются на агломеративные и дивизимные.

Одним из самостоятельных дивизимных алгоритмов является кластеризация методом *К*-средних. Этот метод кластеризации существенно отличается от таких методов, как древовидная кластеризация и двухвходовое объединение, поскольку строит ровно столько *k* различных кластеров, расположенных на возможно больших расстояниях друг от друга, сколько предполагает исследователь.

С вычислительной точки зрения метод *К*-средних можно рассматривать, как анализ, обратный дисперсионному [13,118]. Программа начинает с *k* случайно выбранных кластеров, а затем изменяет принадлежность объектов к ним, стремясь минимизировать изменчивость внутри кластеров и максимизировать изменчивость между кластерами.

Данный метод аналогичен дисперсионному анализу в том смысле, что *F*-критерий в дисперсионном анализе сравнивает межгрупповую изменчивость с внутригрупповой при проверке гипотезы о том, что средние в группах отличаются друг от друга. В кластеризации методом *К*-средних программа перемещает объекты из одних кластеров в другие для того, чтобы получить наиболее значимый результат при проведении дисперсионного анализа.

Вернемся к анализу латентной неоднородности кроветворной активности печени плодов и новорожденных 25, 28-30 недель гестации. На этот раз применим дивизимную кластер-процедуру, основанную на методе *К*-средних. Выберем количество кластеров разбиения 33 объектов, равное 3. Проведенный анализ разделил данную совокупность объектов на 3 кластера, состоящих из 14, 8 и 11 объектов (рис. 41). При этом кластер, состоящий из 8 объектов (№ 2) представлен группой глубоко недоношенных новорожденных с постнатальной инволюцией экстрамедуллярной кроветворной тка-

ни. Интересно, что кластер № 1 был представлен только мертворожденными поздними абортусами, а кластер № 3 – новорожденными и мертворожденными плодами с массой более 1000 г.

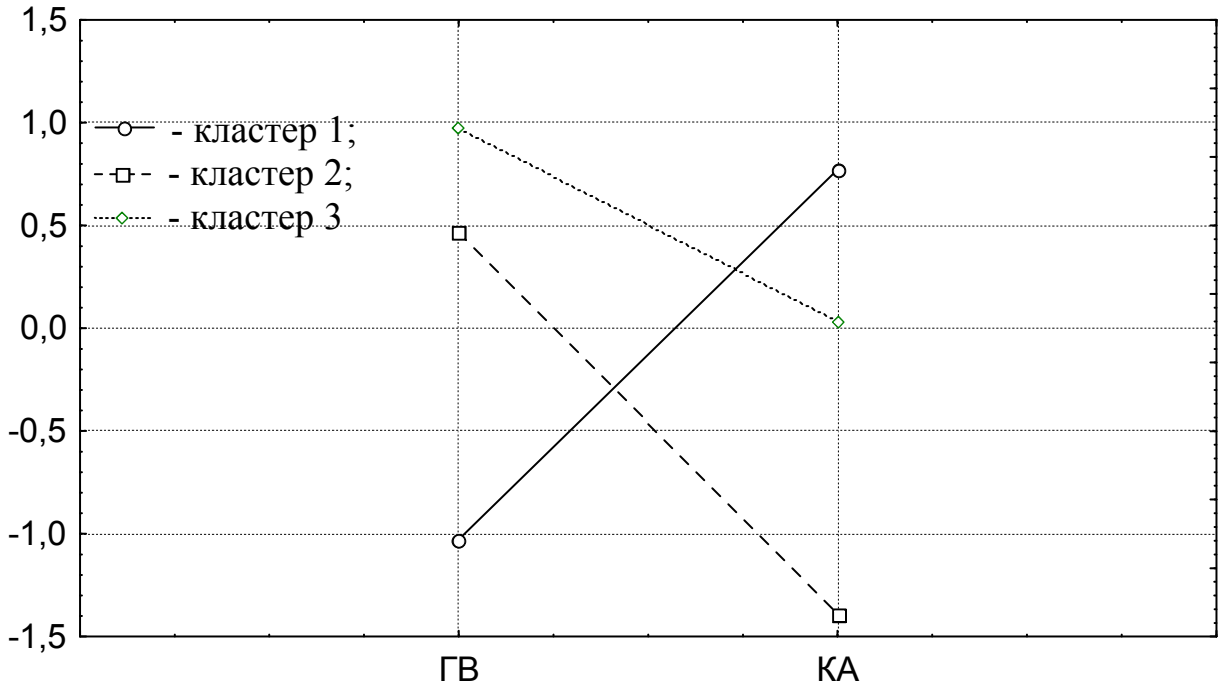

Рис. 41. График средних значений кластеров, полученных с помощью дивизимной классификации плодов и новорожденных 25, 28-30 недель гестации по степени кроветворной активности паренхимы печени и показателю гестационного возраста методом *K*-средних. Мера дистанции между объектами – нормированное евклидово расстояние. ГВ – гестационный возраст; КА - кроветворная активность.

Таким образом, метод *K*-средних также подтверждает доказанную ранее латентную неоднородность кроветворной ткани паренхимы печени за счет наличия в выборочной совокупности глубоко недоношенных новорожденных с постнатальным опустошением экстрамедуллярной миелоидной ткани.

В заключение следует отметить, что семейство процедур кластеризации представляет собой эффективное средство разведочного анализа данных. Основной целью использования кластерного анализа в судебно-медицинской антропологии является обнаружение латентного кластеринга перед применением других статистических методов, прежде всего, корреляционно-регрессионного анализа. Вместе с тем не стоит забывать, что кластерный анализ носит эвристический характер и в основном не имеет под собой достаточных статистических оснований. Поэтому результаты кластерного анализа нельзя выдавать за единственно возможные. В любой момент

может возникнуть потребность повторной кластеризации с использованием иных методов, которые могут привести к пересмотру ранее полученных результатов и выводов.

ЗАКЛЮЧЕНИЕ

В силу настоятельных потребностей судебно-следственной практики современное состояние судебно-медицинской антропологии характеризуется непрерывным увеличением количества известных способов идентификации личности. Основными факторами роста числа диагностических методик являются расширение спектра идентифицируемых объектов, совершенствование технических приемов их изучения, а также увеличение числа идентифицирующих показателей. Однако, несмотря на увеличение набора альтернативных диагностических методик и совершенствование инструментальной базы, точность идентификации пока остается недостаточно высокой. В этой связи одним из возможных путей повышения диагностической значимости результатов судебно-медицинских антропологических исследований является оптимизация методов статистического анализа эмпирических данных.

В настоящее время совокупность статистических методов, используемых при проведении судебно-медицинских антропологических исследований, можно разделить на три основные группы.

Первая группа включает в себя семейство методов корреляционно-регрессионного анализа, позволяющих изучать взаимосвязи между различными показателями и производить построение аналитических моделей, с определенной точностью прогнозирующих значение идентифицируемого параметра по значению идентифицирующего признака или группы признаков. Основные пути оптимизации корреляционно-регрессионного анализа связаны с предварительным выявлением латентной неоднородности данных, применением методов нелинейной и раздельной регрессии, альтернативных функций потерь и различных алгоритмов их минимизации, анализом мультиколлинеарности факторных показателей, неоднородности дисперсии и серийной корреляции ошибок прогнозирования.

Вторая группа статистических методов представлена совокупностью достаточно разнородных процедур классификации объектов судебно-медицинского экспертного познания. Основными представителями данной группы являются методы одномерного и многомерного дискриминантного анализа. При этом оптимизация бино-

миальной классификации может быть достигнута с помощью недавно разработанных методов, предназначенных как для нормально распределенных данных, так и для показателей с неизвестными или анормальными типами распределений.

В качестве общих путей оптимизации статистических методов прогнозирования и классификации следует назвать процедуры автоматизированного подбора независимых показателей в состав соответствующих регрессионных, дискриминирующих или классифицирующих функций.

Полноценное проведение судебно-медицинских антропологических исследований также невозможно без использования широкого ряда других статистических методов, имеющих вспомогательное значение и преимущественно направленных на выявление латентной неоднородности исследуемых данных. В качестве указанных статистических методов следует назвать тесты выявления кластеринга и выбросов, проверки согласия эмпирических распределений с теоретическими, а также процедуры сравнительного анализа. Названные методы помогают подтвердить или опровергнуть гипотезу о соответствии изучаемых данных специфическим предпосылкам математических моделей конкретных аналитических процедур, а потому являются обязательным компонентом программы любого судебно-медицинского антропологического исследования.

Итогом применения изложенных путей оптимизации должно являться повышение точности разрабатываемых способов идентификации и снижение трудоемкости научных исследований за счет минимизации объемов эмпирических данных, достаточных для достижения поставленных научных и практических судебно-медицинских задач.

Однако использование любых, даже самых сложных статистических процедур может оказаться малоэффективным в случае отсутствия единства научного, образовательного и практического элементов познания при судебно-медицинской идентификации личности. Объясняется это тем, что выбор наиболее оптимального способа идентификации из комплекса альтернативных диагностических методик является залогом успешного решения поставленной экспертной задачи. Объективность названного выбора базируется на данных критического анализа субъектом экспертного познания результатов альтернативных методик отождествления личности с позиций их обобщаемости, достоверности и диагностической значи-

мости. При этом оптимальность методов и достоверность результатов статистического анализа является основным фактором, гарантирующим общую достоверность анализируемых научных данных.

Оптимальность статистических методов и достоверность их результатов определяется степенью соответствия исходных данных тем условиям, наличие которых предполагает математическая модель конкретного алгоритма статистического анализа. Общим условием применимости любого из известных методов статистического анализа является репрезентативность выборок, которая, в первую очередь, определяется случайностью их формирования. Частные условия адекватности статистических методов определяются соответствием изучаемых данных специфическим предпосылкам математических моделей конкретных аналитических процедур.

Дополнительной возможностью повышения эффективности медико-антропологической идентификации с помощью аналитической статистики является использование критериев объективного выбора наиболее оптимального способа идентификации из комплекса альтернативных методик. В этой связи оценка способа судебно-медицинской антропологической идентификации должна завершаться определением его диагностической значимости (точности идентификации). Критерии точности идентификации различаются в зависимости от математической модели, на которой основан способ судебно-медицинской антропологической идентификации.

Для способов, созданных на основе регрессионного анализа, показателем точности идентификации служит доверительная область для прогнозных оценок регрессионной модели. Основными, относительно легко измеряемыми статистическими параметрами, косвенно характеризующими ширину доверительных интервалов, являются остаточная дисперсия и коэффициент детерминации (или его скорректированный аналог для моделей множественной регрессии). Существующие статистические методы сравнения дисперсий и коэффициентов корреляции для выборок одинакового объема позволяют проводить объективное сравнение точности альтернативных регрессий и последующий объективный выбор наиболее точных регрессионных моделей при проведении судебно-медицинских антропологических исследований. Аналогичные методы сравнения, предназначенные для выборок различных объемов, могут быть использованы в экспертной практике в целях объективного выбора из комплекса разработанных разными авторами способов идентифи-

кации одной методики, характеризующейся наибольшей диагностической значимостью.

Для способов, основанных на применении статистических методов классификации, в качестве показателя точности идентификации в настоящее время может быть использован комплекс критериев, адаптированных для способов идентификации с различными уровнями достоверности экспертных суждений и любым количеством идентифицируемых кластеров. Часть критериев характеризует способ идентификации в целом. Основными представителями данной группы являются чувствительность и специфичность идентификации какого-либо кластера. Использование этих критериев точности показано на этапе реализации выбора субъектом экспертного познания конкретной диагностической методики. Остальные критерии характеризуют точность конкретного диагностического результата на этапе практической реализации выбранного способа идентификации, вследствие чего имеют большее практическое значение. Важнейшим из критериев этой группы является прогностическая ценность положительного результата идентификации.

Необходимость анализа достоверности и диагностической значимости результатов статистических методов, использованных при создании любого способа судебно-медицинской идентификации личности, предъявляет определенные требования к подготовке не только субъектов судебно-медицинского научного, но и экспертного познания. В этой связи представляется, что достижение высокой степени достоверности результатов и научной обоснованности выводов экспертиз отождествления личности на современном этапе невозможны без реализации официального закрепления самостоятельности данного вида экспертизы с утверждением соответствующих специальных программ подготовки медико-криминалистических кадров.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Абрамов С.С.* Выбор методов исследования при судебно-медицинской остеологической идентификации // Суд. – мед. эксперт. – 1996. - № 4. – С. 13-20.

2. *Автандилов Г.Г.* Морфометрия в патологии. – М.: Медицина, 1973. – 248 с.

3. *Автандилов Г.Г.* Основы количественной патологической анатомии: Учебное пособие. – М.: Медицина, 2002. – 240 с.

4. *Айвазян С.А., Бежаева З.И., Староверов О.В.* Классификация многомерных наблюдений. – М.: Статистика, 1974. – 240 с.

5. *Алексеев Ю.Д.* Комплексная общепатологическая и судебно-медицинская оценка структурных изменений некоторых желез внутренней секреции в определении возраста человека: Автореф. дис. … докт. мед. наук. - Саратов, 1999. – 24 с.

6. *Аманмурадов А.Х., Пиголкин Ю.И., Богомолов Д.В. и др.* Значение общих и специфических признаков при судебно-медицинской идентификации личности морфологическими методами // Суд. – мед. эксперт. – 2003. - № 1. – С. 33-37.

7. *Ардашкин А.П.* Введение в теорию судебно-медицинской экспертизы: характеристика предмета. – Самара: ООО «Офорт», 2004. – 120 с.

8. *Ардашкин А.П., Недугов Г.В., Недугова В.В.* Способ опознавания нелинейности регрессии на основе сравнения модулей коэффициентов линейной и ранговой корреляции // Актуальные проблемы науки в России: Материалы всероссийской научно-практической конференции: В 3 т. – Кузнецк: КИИУТ, 2005. – Вып. 3. – Т. 3. - С. 10-12.

9. *Ардашкин А.П., Недугова В.В., Недугов Г.В.* Анализ качества научных работ обобщающего характера в области судебной медицины // Вопросы судебной медицины и медицинского права: Сборник научных трудов, посвященный 85-летию кафедры судебной медицины Самарского государственного университета / Под ред. А.П. Ардашкина, В.В. Сергеева. – Самара: ООО «Офорт»; ГОУ ВПО «СамГМУ», 2006. – С. 11-18.

10. *Ардашкин А.П., Недугова В.В., Недугов Г.В.* Ошибки сравнительного статистического анализа на основе критерия Стьюдента в судебно-медицинских научных исследованиях // Вопросы судебной медицины, медицинского права и биоэтики: Сборник научных трудов / Под ред. А.П. Ардашкина, В.В. Сергеева. – Самара: СамГМУ, 2007. – С. 12-18.

11. *Ардашкин А.П., Недугов Г.В.* Судебно-медицинская экспертиза трупов плодов и новорожденных (экспертно-правовая характеристика, гистологическая диагностика). – Самара: ООО «Офорт», 2006. – 145 с.

12. *Башкирева Е.А.* О морфологических особенностях строения волос млекопитающих вида кошка домашняя (felis catus) // Суд. – мед. эксперт. – 2002. - № 5. – С. 29-32.

13. *Боровиков В.П.* STATISTICA. Искусство анализа данных на компьютере: Для профессионалов. – 2-е изд. (+ CD). – СПб.: Питер, 2003. – 688 с.

14. *Ваганов П.А., Лукницкий В.А.* Нейтроны и криминалистика. – Л.: Изд-во Ленинградского ун-та, 1981. – 192 с.

15. *Воронцов И.М., Кельмансон И.А., Цинзерлинг А.В.* Синдром внезапной смерти грудных детей. – СПб.: Специальная литература, 1997. – 220 с.

16. *Гланц С.* Медико-биологическая статистика: Пер. с англ. – М.: Практика, 1998. – 459 с.

17. *Гончарова Н.Н., Самоходская О.В., Федулова М.В. и др.* Методы определения пола человека по рентгенограмме кисти // Суд. – мед. эксперт. – 2005. - № 5. – С. 21-26.

18. *Грачев С.В., Городнова Е.А., Олферьев А.М.* Научные исследования в биомедицине. – М.: ООО «Медицинское информационное агентство», 2005. – 272 с.

19. *Григорьева М.А.* Применение дискриминантного анализа в оценке соматотипа человека по длинным костям конечностей // Суд. – мед. эксперт. – 2004. - № 1. – С. 28-31.

20. *Гринхальх Т.* Основы доказательной медицины: Пер. с англ. – М.: ГЭОТАР - МЕД, 2004. – 240 с.

21. *Давыдовский И.В.* Геронтология. – М.: Медицина, 1966. – 300 с.

22. *Дгебуадзе М.А.* Морфологическое исследование клубочков правой и левой почек в возрастном аспекте // Морфология. – 2001. - № 1. – С. 59-62.

23. *Демиденко Е.З.* Линейная и нелинейная регрессия. – М.: Финансы и статистика, 1981. – 302 с.

24. *Добряк В.И.* Судебномедицинская экспертиза скелетированного трупа. – Киев: Государственное медицинское издательство УССР, 1960. – С. 100-111.

25. *Дубров А.М.* Обработка статистических данных методом главных компонент. – М.: Статистика, 1978. – 136 с.

26. *Дубров А.М., Мхитарян В.С., Трошин Л.И.* Многомерные статистические методы: Учебник. – М.: Финансы и статистика, 2000. – 352 с.

27. *Еременко Е.А., Звягин В.Н.* Установление порядковой локализации однотипных костей стопы // Суд. – мед. эксперт. – 2003. - № 5. – С. 32-36.

28. *Ефимов А.А., Луньков А.Е., Савенкова Е.Н.* Оптимизация регрессионных соотношений при определении возраста человека в судебно-медицинской практике // Пробл. эксперт. в мед. – 2007. – № 1. – С. 13-15.

29. *Жижин К.С.* Медицинская статистика: Учебное пособие. – Ростов-на-Дону: Феникс, 2007. – 160 с.

30. *Закс. Л.* Статистическое оценивание: Пер. с нем. – М.: Статистика, 1976. – 598 с.

31. *Звягин В.Н.* Определение длины окружности головы и размера головного убора при экспертизе черепа человека // Суд. – мед. эксперт. – 1999. - № 5. – С. 25-28.

32. *Звягин В.Н.* Реставрация фрагментированного черепа при экспертизе идентификации личности // Суд. – мед. эксперт. – 2001. - № 2. - С. 15-21.

33. *Звягин В.Н.* Критерии изменчивости толщины костей черепа человека // Суд. – мед. эксперт. – 2001. - № 5. - С. 24-26.

34. *Звягин В.Н.* Проблемный анализ медико-антропологической идентификации личности в судебной медицине // Суд. – мед. эксперт. – 2003. - № 5. - С. 6-14.

35. *Звягин В.Н., Галицкая О.И., Аунапу С.А.* Краниометрическая диагностика массивности скелета и соматотипа // Суд. – мед. эксперт. – 2002. - № 5. - С. 7-12.

36. *Звягин В.Н., Галицкая О.И.* Исследование зольной массы при экспертизе идентификации личности // Суд. – мед. эксперт. – 2002. - № 6. - С. 14-16.

37. *Звягин В.Н., Григорьева М.А.* Прогнозирование основных соматических характеристик человека при экспертизе отдельных расчлененных частей тела // Суд. – мед. эксперт. – 2006. - № 2. – С. 20-24.

38. *Звягин В.Н., Еременко Е.А.* Диагностика массивности скелета и соматотипа человека по костям стопы // Суд. – мед. эксперт. – 2003. - № 3. - С. 17-23.

39. *Звягин В.Н., Замятина А.О.* Установление порядковой локализации множественных однотипных костей кисти // Суд. – мед. эксперт. – 2003. - № 4. - С. 23-27.

40. *Звягин В.Н., Замятина А.О., Галицкая О.И.* Диагностика массивности скелета и соматотипа человека по костям кисти // Суд. – мед. эксперт. – 2003. - № 6. - С. 19-25.

41. *Звягин В.Н., Мальцева Н.Л., Алексина Л.А., Галицкая О.И.* Критерии идентификации личности по подъязычной кости // Суд. – мед. эксперт. – 2005. - № 6. - С. 27-34.

42. *Звягин В.Н., Самоходская О.В., Иванов Н.В., Григорьева М.А.* Диагностика пола и длины тела человека по фрагментированным костным останкам // Суд. – мед. эксперт. – 1997. - № 1. - С. 24-31.

43. *Иберла К.* Факторный анализ: Пер. с нем. – М.: Статистика, 1980. – 400 с.

44. *Клевно В.А.* Морфология и механика разрушения ребер (Судебно-медицинская диагностика механизмов, последовательности и прижизненности переломов). – Барнаул, 1993. – 300 с.

45. *Колосова В.М.* Математическая обработка результатов измерений при сравнительном исследовании // Лабораторные и специальные методы исследования в судебной медицине: Практическое руководство / Под ред. В.И. Пашковой, В.В. Томилина. – М.: Медицина, 1975. – С. 234-247.

46. *Нарина Н.В., Звягин В.Н.* Определение соматотипа мужчин при краниофациальной идентификации личности // Суд. – мед. эксперт. – 2004. - № 5. – С. 27-31.

47. *Недугова В.В.* Сравнительный анализ основных моделей принятия экспертных решений // Вопросы судебной медицины и медицинского права: Сборник научных трудов, посвященный 85-летию кафедры судебной медицины Самарского государственного университета / Под ред. А.П. Ардашкина, В.В. Сергеева. – Самара: ООО «Офорт»; ГОУ ВПО «СамГМУ», 2006. – С. 66-69.

48. *Недугова В.В.* Обоснование оптимальной дифференциации экспертного медико-криминалистического познания // Вопросы судебной медицины, медицинского права и биоэтики: Сборник научных трудов / Под ред. А.П. Ардашкина, В.В. Сергеева. – Самара: СамГМУ, 2007. – С. 26-32.

49. *Недугов Г.В.* Математическое моделирование пренатального морфогенеза гистоструктур фетальной селезенки в целях определения гестационного возраста // Пробл. эксперт. в мед. – 2003. – № 3. – С. 15-17.

50. *Недугов Г.В.* Метод определения оптимального объема наблюдений для морфолого-статистического анализа в гистологических исследованиях // Актуальные вопросы судебной и клинической медицины / Под ред. Н.В. Бастуева. – Ханты-Мансийск, 2004. – Вып. 7. – С. 116-121.

51. *Недугов Г.В.* Метод расчета доверительных границ для прогнозных оценок регрессионных уравнений при наличии гетероскедастичности // Актуальные проблемы науки в России: Материалы всероссийской научно-практической конференции: В 3 т. – Кузнецк: КИИУТ, 2005. – Вып. 3. – Т. 3. - С. 36-38.

52. *Недугов Г.В.* Метод выявления выбросов в выборках небольшого объема, предназначенных для проведения регрессионного анализа // Актуальные проблемы науки в России: Материалы всероссийской научно-практической конференции: В 3 т. – Кузнецк: КИИУТ, 2005. – Вып. 3. – Т. 3. - С. 38-40.

53. *Недугов Г.В.* Морфолого-математическая оценка развития фетальных органов в целях определения гестационного возраста: Дис. … канд. мед. наук. – Самара, 2005. – 179 с.

54. *Недугов Г.В.* Проблемы воспроизводимости результатов количественных судебно-гистологических исследований // Вопросы судебной медицины и медицинского права: Сборник научных трудов, посвященный 85-летию кафедры

судебной медицины Самарского государственного университета / Под ред. А.П. Ардашкина, В.В. Сергеева. – Самара: ООО «Офорт»; ГОУ ВПО «СамГМУ», 2006. – С. 74-79.

55. *Недугов Г.В., Ардашкин А.П., Недугова В.В.* Установление гестационного возраста на основании морфометрического исследования кроветворной активности фетальной печени // Пробл. эксперт. в мед. – 2002. - № 4. – С. 7-11.

56. *Недугов Г.В., Недугова В.В.* Медико-экспертная оценка патологии лимфоидной ткани червеобразного отростка // Пробл. эксперт. в мед. – 2006. – № 1. – С. 14-16.

57. *Недугов Г.В., Недугова В.В.* Идентификация пола человека методом одномерной биномиальной классификации // Пробл. эксперт. в мед. – 2007. – № 1. – С. 10-13.

58. *Недугов Г.В., Недугова В.В.* Критерии точности идентификации объектов судебно-медицинского экспертного познания // Вопросы судебной медицины, медицинского права и биоэтики: Сборник научных трудов / Под ред. А.П. Ардашкина, В.В. Сергеева. – Самара: СамГМУ, 2007. – С. 32-38.

59. *Недугов Г.В., Недугова В.В.* Методы сравнения точности альтернативных регрессионных моделей идентификации личности // Вопросы судебной медицины, медицинского права и биоэтики: Сборник научных трудов / Под ред. А.П. Ардашкина, В.В. Сергеева. – Самара: СамГМУ, 2007. – С. 39-44.

60. *Неклюдов Ю.А.* Биологический возраст: судебно-медицинские аспекты // Суд. – мед. эксперт. – 1997. - № 2. – С. 10-13.

61. *Неклюдов Ю.А., Алексеев Ю.Д., Спиридонов А.В. и др.* Возможности определения возраста по мягким тканям человека (морфометрическое исследование) // Суд. – мед. эксперт. – 2001. - № 2. – С. 41-43.

62. *Павлов А.В.* Возрастная динамика основных структурных компонентов семенников человека в оценке биологического возраста: Автореф. дис. … канд. мед. наук. - Саратов, 1997. – 29 с.

63. *Пашинян Г.А., Лебеденко И.Ю., Манин А.И.* Значение аномалий зубов при идентификации личности // Суд. – мед. эксперт. – 2004. - № 2. – С. 19-20.

64. *Пашкова В.И., Резников Б.Д.* Судебно-медицинское отождествление личности по костным останкам. – Саратов: Изд-во Саратовского ун-та, 1978. – 320 с.

65. *Пиголкин Ю.И., Богомолов Д.В., Федулова М.В и др.* Возрастные изменения микроструктуры костной ткани и возможности их использования для идентификации личности // Суд. – мед. эксперт. – 2002. - № 2. – С. 17-20.

66. *Пиголкин Ю.И., Богомолова И.Н.* Применение принципов доказательной медицины в качестве критериев полезности новых методов исследования в экспертной практике // Суд. – мед. эксперт. – 2004. - № 6. – С. 3-6.

67. *Пиголкин Ю.И., Гончарова Н.Н., Федулова М.В., Золотенкова Г.В.* Значение принципов возрастной морфологии для судебной антропологии // Суд. – мед. эксперт. – 2003. - № 4. – С. 47-49.

68. *Пушкарев В.П., Новиков П.И.* Популяционное исследование D1S80 локуса методом капиллярного электрофореза у представителей кавказоидов Уральского региона России // Суд. – мед. эксперт. – 2001. - № 2. – С. 21-26.

69. *Пушкарев В.П., Рахманина Л.В., Новиков П.И., Иванов П.Л.* Исследование с помощью капиллярного электрофореза аллельного разнообразия микросателлитных локусов D16S539, F13B, FESFPS, TH01 и TPOX у европеоидов Уральского региона России // Суд. – мед. эксперт. – 2004. - № 1. – С. 23-28.

70. *Рао С.Р.* Линейные статистические методы и их применение: Пер. с англ. – М.: Наука, 1968. – 548 с.

71. *Савенкова Е.Н., Неклюдов Ю.А., Ефимов А.А.* Возрастная динамика коэффициента сократимости кожи человека и возможность его использования при определении биологического возраста в судебной медицине // Пробл. эксперт. в мед. – 2006. - № 1. - С. 18-19.

72. *Салманов О. Н.* Математическая экономика с применением Mathcad и Excel. – СПб.: БХВ – Петербург, 2003. – 464 с.

73. *Секей Г.* Парадоксы в теории вероятностей и математической статистике: Пер. с англ. – М.-Ижевск: Институт компьютерных исследований, 2003. – 272 с.

74. *Сигел Э.* Практическая бизнес-статистика.: Пер. с англ. – М.: Издательский дом «Вильямс», 2002.- 1056 с.

75. *Славин М. Б.* Методы системного анализа в медицинских исследованиях. – М.: Медицина, 1989. – 304 с.

76. *Смоляк С.А., Титаренко Б.П.* Устойчивые методы оценивания. – М.: Статистика, 1980. – 208 с.

77. *Спиридонов А.В.* Возрастные изменения щитовидной железы и их судебно-медицинская оценка: Автореф. дис. … канд. мед. наук. – Саратов, 1997.

78. *Сулицкий В.Н.* Методы статистического анализа в управлении: Учеб. пособие. – М.: Дело, 2002. – 520 с.

79. *Урбах В.Ю.* Биометрические методы. Статистическая обработка опытных данных в биологии, сельском хозяйстве и медицине. – М.: Наука, 1964. – 416 с.

80. *Федорина Т.А., Недугов Г.В.* Комплексный подход к оценке погрешностей, возникающих при использовании методов количественного анализа в гистологических исследованиях: Учебное пособие. – Самара: ООО «Содружество Плюс», ГОУВПО «СамГМУ», 2004. – 48 с.

81. *Федулова М.В.* Зависимость параметров микроструктуры костной ткани, связанных с возрастом, от пола, роста и размеров ребра человека // Суд. – мед. эксперт. – 2004. - № 2. – С. 16-18.

82. *Фролов Ю.П.* Математические методы в биологии. ЭВМ и программирование: Теоретические основы и практикум. – Самара: Самарский университет, 1997. – 265 с.

83. *Хьюбер Д.* Робастность в статистике. – М.: Мир, 1984. – 304 с.

84. *Этинген Л. Е.* Нормальная морфология человека старческого возраста. – М., 2003. – 256 с.

85. *Яглом А.М., Яглом И.М.* Вероятность и информация. – М.: Главная редакция физико-математической литературы изд-ва «Наука», 1973. – 512 с.

86. *Янковский В.Э., Киселев В.Д., Пятчук С.В.* Исследование остеопоротических изменений длинных трубчатых костей нижних конечностей для определения биологического возраста человека // Суд. – мед. эксперт. – 2006. - № 3. – С. 9-12.

87. *Badgley R.F.* An assessment of research methods reported in 103 scientific articles from two Canadian medical journals // Can. M. A. J. – 1961. – Vol. 85. – P. 256-260.

88. *Barcikowski R., Stevens J.P.* A Monte Carlo study of the stability of canonical correlations, canonical weights, and canonical variate-variable correlations // Multivariate Behavioral Research. – 1975. - Vol. 10. – P. 353-364.

89. *Berkson J.* Minimum chi-square, not maximum likelihood! // Annals of Statist. – 1980. – Vol. 8. – P. 457-487.

90. *Bliss C.I., Cochran W.G., Tukey J.W.* A rejection criterion based upon the range // Biometrika. – 1956. – Vol. 43. – P. 418-422.

91. *Burkholder D.L.* Sufficiency in the undominated case // Annals of Math. Statist. – 1961. – Vol. 32. – P. 1191-1200.

92. *Burkholder D.L.* On the order structure of the set of sufficient σ-fields // Annals of Math. Statist. – 1962. – Vol. 33. – P. 596-599.

93. *Burrows G.L.* Statistical tolerance limits – what are they? // Applied Statistics. – 1963. – Vol. 12. – P. 133-144.

94. *Chissom B.S.* Interpretation of the kurtosis statistic // The American Statistician. – 1970. – Vol. 24. – P. 19-22.

95. *Cleveland W.S.* Robust locally weighted regression and smoothing scatterplots // J. Amer. Statist. Assoc. – 1979. - Vol. 74. – P. 829-836.

96. *Cleveland W.S.* Graphs in scientific publications // The American Statistician. – 1984. - Vol. 38. – P. 270-280.

97. *Cochran W.G.* The distribution of the largest of a set of estimated variances as a fraction of their total // Ann. Eugen. – 1941. Vol. 11. – P. 47-61.

98. *Cruess D.F.* Review of the use of statistics in the American Journal of Tropical Medicine and Hygiene for January – December 1988 // Am. J. Trop. Med. Hyg. – 1990. - Vol. 41. – P. 619-626.

99. *Darlington R.B., Weinberg S., Walberg H.* Canonical variate analysis and related techniques // Review of Educational Research. – 1973. – Vol. 43. – P. 433-454.

100. *David H.A.* The ranking of variance in normal populations // J. Amer. Statist. Assoc. – 1956. - Vol. 51. – P. 621-626.

101. *Davies J.* A critical survey of scientific methods in two psychiatry journals // Aust. N. Z. J. Psych. – 1987. - Vol. 21. – P. 367-373.

102. *Dixon W.J.* Analysis of extreme values // Annals of Math. Statist. – 1950. – Vol. 21. – P. 488-506.

103. *Dixon W.J.* Processing data for outliers // Biometrics. – 1953. – Vol. 9. – P. 74-89.

104. *Dixon W.J.* Rejection of Observations // Contributions to Order Statistics / Ed. by A.E. Sarhan, B.G. Greenberg. – New York, 1962. – P. 299-342.

105. *Drasar B.S., Irving D.* Environmental factors and cancer of the colon and breast // Br. J. Cancer. – 1973. – Vol. 27. – P. 167-172.

106. *Durbin J.* Testing for serial correlation in least-squares regression when some of the regressors are lagged dependent variables // Econometrica. – 1970. – Vol. 38. – P. 410-421.

107. *Durbin J., Watson G.S.* Testing for serial correlations in least squares regression. II // Biometrika. – 1951. – Vol. 38. – P. 159-178.

108. *Edwards A.V.T.* The history of likelihood estimate // Internal. Statist. Rev. – 1974. – Vol. 42. – P. 9-15.

109. *Efron B.* Controversies in the foundations of statistics // The American Math. Monthly. – 1978. – Vol. 85. – P. 231-246.

110. *Faulkenberry G.D., Daly J.C.* Sample size for tolerance limits on a normal distribution // Technometrics. – 1970. – Vol. 12. – P. 813-821.

111. *Ferguson T.*S. An inconsistent maximum likelihood estimate // J. Amer. Statist. Assoc. – 1982. – Vol. 77. – P. 831-834.

112. *Gaddum J.H.* Lognormal distributions // Nature. – 1945. – Vol. 156. – P. 463-466.

113. *Gardner E.S., Jr.* Exponential smoothing: The state of the art // Journal of Forecasting. - 1985. – Vol. 4. – P. 1-28.

114. *Gatsonis C., Sampson A.R.* Multiple correlation: exact power and sample size calculations // Psychological Bulletin. – 1989. Vol. 106. – P. 516-524.

115. *Gill P.E., Murray W.* Quasi-Newton methods for unconstrained optimization // Journal of the Institute of Mathematics and its Applications. – 1972. – Vol. 9. – P. 91-108.

116. *Harsaae E.* On the computation and use of a table of percentage points of Bartlett's M. // Biometrika. – 1969. – Vol. 56. – P. 273-281.

117. *Hartigan J.A., Wong M.A.* Algorithm 136. A *k*-means clustering algorithm // Applied Statistics. – 1978. – Vol. 28. – P. 100.

118. *Hartley H.O.* The maximum F-ratio as a short cut test for heterogeneity of variance // Biometrika. – 1950. – Vol. 37. – P. 308-312.

119. *Herrey E.M.J.* Confidence intervals based on the mean absolute deviation of a normal sample // J. Amer. Statist. Assoc. – 1965. – Vol. 60. – P. 257-269.

120. *Hoerl A.E., Kennard R.W.* Ridge regression: Applications to nonorthogonal problems // Technometrics. – 1970. – Vol. 12. P. 69-82.

121. *Huberty C.J.* Discriminant analysis // Review of Educational Research. – 1975. - Vol. 45. – P. 543-598.

122. *Jaeschke R., Guyatt G., Sackett D.L.* Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? // JAMA. – 1994. – Vol. 271. – P. 389-391.

123. *Jaenschke R., Guyatt G., Sackett D.L.* Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What were the results and will they help me in caring for my patients? // Ibid. – 1994. – P. 703-707.

124. *Jennrich R.I., Sampson P.F.* Application of stepwise regression to non-linear estimation // Technometrics. – 1968. - Vol. 10. – P. 63-72.

125. *Johnson S.C.* Hierarchical clustering schemes // Psychometrika. – 1967. – Vol. 32. – P. 241-254.

126. *Kac N., Kiefer J., Wofrowitz J.* On tests of normality and other tests of goodness of fit based on distance methods // Annals of Math. Statist. – 1955. – Vol. 26. – P. 189-211.

127. *Koller S.* Typlsierung korrelativer Zusammenhänge // Metrika. – 1963. – Vol. 6. – P. 65-75.

128. *Koller S.* Systematik der statistischen Schluβfehler // Method. Inform. Med. – 1964. – Vol. 3. – P. 113-117.

129. *Krutchkoff F.G.* The correct use of the sample mean absolute deviation in confidence intervals for a normal variate // Technometrics. – 1966. – Vol. 8. – P. 663-674.

130. *Kymn K.O.* The distribution of the sample correlation coefficient under the null hypothesis // Econometrica. – 1968. – Vol. 36. – P. 187-189.

131. *Leslie R.T., Brown B.M.* Use of range in testing heterogeneity of variance // Biometrika. – 1966. – Vol. 53. – P. 221-227.

132. *Lilliefors H.W.* On the Kolmogorov-Smirnov test for normality with mean and variance unknown // J. Amer. Statist. Assoc. – 1967. – Vol. 64. – P. 399-402.

133. *Loh W.Y., Shih Y.S.* Split selection methods for classification trees // Statistica Sinica. – 1997. – Vol. 7. – P. 815-840.

134. *Loh W.Y., Vanichestakul N.* Tree-structured classification via generalized discriminant analysis (with discussion) // J. Amer. Statist. Assoc. – 1988. – Vol. 83. – P. 715-728.

135. *Mahalanobis P.C.* A method of fractile graphical analysis // Econometrica. – 1960. – Vol. 28. – P. 325-351.

136. *Mant D.* Testing a test: three critical steps // Critical Reading for Primary Care / Ed by R. Jones, A.L. Kinmonth. – Oxford: Oxford University Press, 1995. – P. 183-190.

137. *McKinney W.P., Young M.J, Harta A., Lee M.B.* The inexact use of Fisher's exact test in six major medical journals // JAMA. – 1989. - Vol. 261. – P. 3430-3433.

138. *Neter J., Wasserman W., Kutner M.H.* Applied linear statistical models: Regression, analysis of variance, and experimental designs. - Homewood, IL: Irwin, 1985. – P. 168.

139. *Okunade A.A., Chang C.F., Evans R.D.* Comparative analysis of regression output summary statistics in common statistical packages // The American Statistician. – 1993. - Vol. 47. – P. 298-303.

140. *Owen D.B.* A survey of properties and applications of the noncentral *t*-distribution // Technometrics. – 1968. – Vol. 10. – P. 445-478.

141. *Pearson E.S., Stephens M.A.* The ratio of range to standard deviation in the same normal sample // Biometrika. – 1964. - Vol. 51. – P. 484-487.

142. *Rao C.R.* An asymptotic expansion of the distribution of Wilks' criterion // Bulletin of the International Statistical Institute. – 1951. - Vol. 33. – P. 177-181.

143. *Read M.C., Lachs M.S., Feinstein A.R.* Use of methodological standards in diagnostic test research: getting better but still not good // JAMA. – 1995. – Vol. 274. – P. 645-651.

144. *Ross O.B., Jr.* Use of controls in medical research // JAMA. – 1951. – Vol. 145. – P. 72-75.

145. *Rothman K.J.* A show of confidence // N. Engl. J. Med. – 1978. – Vol. 299. – P. 1362-1363.

146. *Sackett D.L., Haynes R.B., Guyatt G.H., Tugwell P.* Clinical Epidemiology – a Basis Science or Clinical Medicine. – London: Little, Brown, 1991. – P. 51-68.

147. *Sackett D.L., Haynes R.B.* On the need for evidence based medicine // Evidence based medicine. – 1995. – Vol. 1. - P. 4-5.

148. *Schmidt P., Muller E.N.* The problem of multicollinearity in a multistage causal alienation model: A comparison of ordinary least squares, maximum-likelihood and ridge estimators // Quality and Quantity. - 1978. – Vol. 12. – P. 267-297.

149. *Schor S., Karten I.* Statistical evaluation of medical journal manuscripts // JAMA. – 1966. – Vol. 195. – P. 1123-1128.

150. *Schulz K.F., Chalmers I., Hayes R.J., Altman D.J.* Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials // JAMA. – 1995. – Vol. 273. – P. 408-412.

151. *Sethuraman J.* Conflicting criteria of «goodness» of statistics // Sankhya. – 1961. – Vol. 23. – P. 187-190.

152. *Sheynin O.B.* C.F. Gauss and theory of errors // Archive for History of Exact Sciences // 1979. – Vol. 19. – P. 21-72.

153. *Stein C.* Inadmissibility of the usual estimate for the variance of a normal distribution with unknown mean // Annals Inst. Statist. Math. – 1964. – Vol. 16. – P. 1155-160.

154. *Stuart A.* A paradox in statistical estimation // Biometrika. – 1955. - Vol. 42. P. 527-529.

155. *Tukey J.W.* Exploratory Data Analysis. – Reading, Mass.: Addison-Wesley, 1977. – P. 44.

156. *Van Vark G.N.* The investigation of Human Cremated Skeletal Material by Multivariate Statistical Methods. II Measures // Ossa. – 1975. – Vol. 2, № 1. – P. 47-68.

157. *Velicer W.F., Jackson D.N.* Component analysis vs. factor analysis: some issues in selecting an appropriate procedure // Multivariate Behavioral Research. – 1990. – Vol. 25. – P. 1-28.

158. *Ward J.H.* Hierarchical grouping to optimize an objective function // J. Amer. Statist. Assoc. – 1963. – Vol. 58. – P. 236.

159. *Weissberg A., Betty G.H.* Tables of tolerance limit factors for normal distributions // Technometrics. – 1960. – Vol. 2. – P. 483-500.

СОДЕРЖАНИЕ

Недугов Герман Владимирович
Недугова Виолетта Владимировна

СТАТИСТИЧЕСКИЙ АНАЛИЗ В СУДЕБНО-МЕДИЦИНСКОЙ АНТРОПОЛОГИИ

Монография

Сдано в набор 10.08.2007 г. Подписано в печать 27.08.2007 г.
Формат 60x84 1/16
Бумага офсетная. Гарнитура Times New Roman.
Печать офсетная. Объем 16,5 печ. л. Тираж 200 экз.
Заказ № 3387.

443099 г. Самара, ул. Куйбышева, 42.
АНО «Типография ГУВД Самарской области».
Тел. (846)278-26-06, 232-77-85.