

Доказательная медицина

**ПРИМЕНЕНИЕ
СТАТИСТИЧЕСКИХ
МЕТОДОВ**

Основы теории вероятностей
и математической статистики

Диагностика онкологических
заболеваний на основании
результатов исследований
изменений в тканях и клетках

Влияние величины дозы
облучения на вероятность развития
злокачественных новообразований

Диагностика гастроудоденальных
заболеваний методами статистической
теории распознавания

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k$$

$$s^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2$$

$$r(x, y) = \frac{k(x, y)}{\sigma(x)\sigma(y)}$$

$$k(x, y) = m[x - m(x)][y - m(y)]$$

$$P(x'_k \in (x_{(i)}, x_{(j)})) = \frac{j-i}{n+1}$$

$$P(|x - m(x)| \geq \lambda \sigma(x)) \leq \frac{4}{9} \frac{1}{\lambda^2}$$

$$p(E, A) = \lim_{n \rightarrow \infty} h_n(T, A) = \lim_{n \rightarrow \infty} \frac{k}{n}$$

$$p((E_1, E_2), (A_1, A_2)) = p(E_1, A_1) p(E_2, A_2)$$

$$x_i = \sum_{j=1}^n \sum_{l=1}^m r_{ij}$$

$$x_8 = \frac{x_1}{\dots}$$

Д.А. Ключин Ю.И. Петунин

Доказательная медицина

**ПРИМЕНЕНИЕ
СТАТИСТИЧЕСКИХ
МЕТОДОВ**

Доказательная медицина

ПРИМЕНЕНИЕ СТАТИСТИЧЕСКИХ МЕТОДОВ

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k$$

$$s^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2$$

$$r(x, y) = \frac{k(x, y)}{\sigma(x)\sigma(y)}$$

$$k(x, y) = m[x - m(x)][y - m(y)]$$

$$P(x'_k \in (x_{(i)}, x_{(j)})) = \frac{j-i}{n+1}$$

$$p(|x - m(x)| \geq \lambda \sigma(x)) \leq \frac{4}{9} \frac{1}{\lambda^2}$$

$$p(E, A) = \lim_{n \rightarrow \infty} h_n(T, A) = \lim_{n \rightarrow \infty} \frac{k}{n}$$

$$p((E_1, E_2), (A_1, A_2)) = p(E_1, A_1) p(E_2, A_2)$$

$$x_7 = \sum_{i=1}^m \sum_{j=1}^n r_{ij} \quad x_8 = \frac{x_1}{m - p}$$

ББК (P)53.4
K52
УДК 616-079

Компьютерное издательство “Диалектика”

По общим вопросам обращайтесь в издательство “Диалектика” по адресу:
info@dialektika.com, <http://www.dialektika.com>

Клюшин, Д.А., Петунин, Ю.И.

K52 Доказательная медицина. Применение статистических методов. — М. : ООО
“И.Д. Вильямс”, 2017 — 316 с. : ил.

ISBN 978-5-8459-1321-0 (рус.)

ББК (P)53.4

Рекомендовано к изданию Советом факультета кибернетики Киевского национального университета имени Тараса Шевченко 23 апреля 2007 года (протокол № 8)

Рецензенты:

доктор физико-математических наук, профессор И.К. Мацак

доктор физико-математических наук, профессор Б.В. Рублев

доктор физико-математических наук Г.И. Соляник

Все названия программных продуктов являются зарегистрированными торговыми марками соответствующих фирм.

Научное издание

Дмитрий Анатольевич Клюшин, Юрий Иванович Петунин

Доказательная медицина

Применение статистических методов

Издательский дом “Вильямс”
127055, г. Москва, ул. Лесная, д. 43, стр. 1

ISBN 978-5-8459-1321-0 (рус.)

© Компьютерное изд-во “Диалектика”, 2017,
текст, оформление, макетирование

Содержание

Благодарности	10
Введение	11
От издательства “Диалектика”	14
Глава 1. Основания теории вероятностей	15
1.1. Детерминированные, стохастические и хаотические системы	16
1.1.1. Понятие системы	16
1.1.2. Отчего возникает неопределенность	17
1.1.3. Все ли недетерминированные системы обнаруживают одинаковое поведение? Стохастические и хаотические системы	19
Резюме	23
1.2. Случайные испытания и случайные события. Сравнение случайных событий. Операции над случайными событиями	23
1.2.1. Случайные испытания и случайные события	24
1.2.2. Сравнение случайных событий	24
1.2.3. Операции над случайными событиями	25
Резюме	28
1.3. Случайные эксперимент и вероятность случайного события. Поле случайных событий, происходящих при реализации одного случайного эксперимента. Независимые случайные события и эксперименты	29
1.3.1. Что такое вероятность случайного события	29
1.3.2. Поле случайных событий	31
1.3.3. Независимые случайные события и эксперименты	32
1.3.4. Существует ли математическое понятие случайного эксперимента?	39
1.3.5. Можно ли складывать случайные величины, возникающие при проведении различных случайных испытаний?	40
1.3.6. Является ли распределение вероятностей, порожденное случайной величиной, счетно-аддитивной функцией?	41
Резюме	44
1.4. Случайные величины и их вероятностные характеристики. Основные неравенства теории вероятностей. Правило 3 σ	45

1.4.1. Случайные величины и их характеристики	45
1.4.2. Числовые характеристики системы случайных величин	59
1.4.3. Основные неравенства теории вероятностей	62
Резюме	65
1.5. Случайные процессы и последовательности	66
1.5.1. Вероятностные характеристики случайных процессов и корреляционные функции	66
1.5.2. Стационарные случайные процессы и последовательности	70
1.5.3. Корреляционные функции и их свойства	75
Резюме	78
Глава 2. Основные понятия математической статистики	81
2.1. Понятие генеральной совокупности. Случайный выбор элементов генеральной совокупности. Простой случайный выбор	81
Резюме	82
2.2. Оценка вероятностных характеристик случайных событий и величин. Несмещенные оценки неизвестного математического ожидания	83
Резюме	86
2.3. Оценка неизвестной дисперсии случайной величины. Квадратичные оценки дисперсии	86
Резюме	90
2.4. Оценка коэффициентов ковариации и корреляции	91
Резюме	95
2.5. Оценка неизвестных параметров. Метод наименьших квадратов и метод максимального правдоподобия	95
Резюме	98
2.6. Схема линейной регрессии. Несмещенные линейные оценки неизвестных параметров	99
Резюме	103
2.7. Модель сплайновой регрессии	105
Резюме	108
2.8. Оценки неизвестных функций распределения и плотности вероятностей	109
Резюме	113
2.9. Доверительное оценивание. Доверительные интервалы для основной распределенной массы генеральной совокупности	114
2.9.1. Правило 3σ	115
2.9.2. Доверительные интервалы для параметров, построенные по правилу 3σ	115
2.9.3. Доверительные интервалы, порожденные порядковыми статистиками	123
Резюме	126

2.10. Статистические критерии для проверки гипотез	128
2.10.1. Основные принципы проверки гипотез. Теория Неймана–Пирсона	128
2.10.2. Оптимальные статистические критерии	130
2.10.3. Статистические критерии, использующие непринятие решений	132
2.10.4. Индивидуальные статистические критерии	136
2.10.5. Статистические критерии, построенные на основании обучающих выборок	144
Резюме	147
2.11. Меры близости между функциями распределения и выборками. Непараметрические критерии эквивалентности генеральных совокупностей	148
2.11.1. Меры близости и метрики	148
2.11.2. Введение	151
2.11.3. Доверительные интервалы и уровни значимости	152
2.11.4. Мера близости между выборками (p -статистика)	153
2.11.5. Оценка асимптотического доверительного уровня значимости	154
2.11.6. Сравнение p -статистики со статистиками Колмогорова–Смирнова и Вилкоксона	159
Резюме	167
2.12. Стратифицированный анализ генеральных совокупностей	167
Резюме	181
2.13. Статистический критерий для сравнения двух вероятностей	182
2.13.1. Введение	182
2.13.2. Оценки дисперсии частоты случайного события в схеме Бернулли	183
2.13.3. Доверительные границы для неизвестной вероятности	184
2.13.4. Статистические критерии для сравнения двух вероятностей, построенные на основании доверительных интервалов	186
2.13.5. Статистический критерий для проверки гипотезы о равенстве двух вероятностей, основанный на правиле 2σ	187
Резюме	187
2.14. “Точные” доверительные границы для неизвестной вероятности	188
Резюме	194
Глава 3. Применение статистических методов в медицине и радиобиологии	197
3.1. Компьютерный цитогенетический метод диагностики рака молочной и щитовидной железы	197
3.1.1. Введение	197
3.1.2. Меры близости между выборкой и множеством обучающих выборок, используемые для компьютерной диагностики рака груди	198

3.2. Цитоспектрофотометрические исследования изменений слизистой оболочки рта, ассоциированных с малигнизацией молочной железы	203
3.2.1. Биомедицинские основы	203
3.2.2. Материал и метод	205
3.2.3. Калибровка обучающих выборок	208
Резюме	220
3.3. Анализ текстурных показателей ядер буккального эпителия при доброкачественных и злокачественных процессах в щитовидной железе	220
3.3.1. Введение	220
3.3.2. Методика исследований	222
3.3.3. Результаты и их обсуждение	223
3.3.4. Оценка чувствительности и специфичности метода	225
Резюме	226
3.4. Оценка степени прогностической значимости цитогенетических, морфологических и клинических показателей у больных меланомой кожи	226
3.4.1. Введение	226
3.4.2. Материал и методы исследования	227
3.4.3. Результаты собственных исследований и их обсуждение	232
Резюме	235
3.5. Применение статистических методов в радиобиологии: оценка величины поглощенной дозы и степени острой лучевой болезни на основании хромосомных aberrаций	236
3.5.1. Материал и методы исследования	236
3.5.2. Результаты и обсуждение	239
Резюме	244
3.6. Влияние величины дозы облучения на вероятность развития злокачественных новообразований	244
3.6.1. Материалы и методы исследования	245
3.6.2. Математические основы обработки статистических данных. Модифицированный полигон распределения	246
3.6.3. Аппроксимация модифицированных полигонов линейными сплайнами с двумя узлами. Метод сплайновой регрессии	248
3.6.4. Определение количества точек перехода	249
3.6.5. Результаты вычислений	251
Резюме	254
3.7. Стратификационный анализ популяций раковых клеток, резистентных к цисплатину	255
3.7.1. Экспериментальные модели резистентности опухоли к действию цисплатина	255
3.7.2. Приготовление препаратов для проведения морфометрического исследования	256

3.7.3. Статистические методы исследования гетерогенных генеральных совокупностей	256
3.7.4. Результаты исследования	259
Резюме	266
3.8. Количественная оценка изменчивости и гетерогенности культивируемых опухолевых клеток при формировании их радиорезистентности	266
3.8.1. Материал и методы исследования	267
3.8.2. Результаты и их обсуждение	273
Резюме	278
3.9. Новые калибровочные кривые для дозовых зависимостей радиационно-индуцированных хромосомных aberrаций в культуре лимфоцитов человека	278
3.9.1. Введение	278
3.9.2. Материал и методы исследования	279
3.9.3. Результаты исследований и их обсуждение	280
Резюме	283
3.10. Диагностика гастродуоденальных заболеваний методами статистической теории распознавания	283
3.10.1. Метод дуоденального зондирования для исследования функционального состояния органов желудочного тракта	283
3.10.2. Комплексный метод фракционного гастродуоденального зондирования	284
3.10.3. Регистрация результатов гастродуоденального зондирования	284
3.10.4. Доверительные границы для основной распределенной массы значений генеральной совокупности. Правило 3σ и $3s$	288
3.10.5. Метод сплайновой регрессии для анализа временных рядов	291
3.10.6. Результаты диагностирования гастродуоденальных заболеваний	297
Резюме	301
Приложение. Необходимые сведения из математики	303
Литература	307

Благодарности

Авторы искренне признательны своим коллегам — биологам и медикам, работающим в Институте экспериментальной патологии, онкологии и радиобиологии им. Р.Е. Кавецкого Национальной Академии Наук Украины, Институте онкологии Академии медицинских наук Украины, Институте клинической и экспериментальной хирургии Академии медицинских наук Украины, Н.В. Бородай, В.А. Барабою, А.Н. Бурому, Э.А. Деминой, В.А. Зинченко, С.П. Кожара, С.И. Коровину, Г.И. Кулик, Л.А. Налескиной, В.Э. Орлу, А.В. Романову, И.Н. Тодору, В.Ф. Чехуну, О.В. Юрченко, а также сотруднику Института математики Национальной Академии Наук Украины М.Ю. Савкиной.

С большой теплотой и благодарностью мы вспоминаем совместную работу с нашим дорогим учителем, выдающимся ученым в области онкоморфологии и цитогенетики профессором Калерией Павловной Ганиной, безвременно ушедшей от нас, без которой эта книга никогда не появилась бы на свет.

Введение

Концепция доказательной медицины

Результатам каких методов диагностики следует доверять? Какие методы диагностики и лечения можно рекомендовать для клинической медицины и практического здравоохранения? Для ответа на эти вопросы и решения возникающих проблем специалистами клинической эпидемиологии Р. Флетчер, С. Флетчер и Э. Вагнер [123] была разработана концепция доказательной медицины, основанная на результатах правильно организованных клинических исследований, в которых решение врача должно базироваться на строго обоснованных и точно доказанных научных фактах, и “вес” каждого факта становится тем больше, чем строже методика исследования, в ходе которого этот факт был получен. Здесь у специалиста по точным наукам возникает вопрос: если факт естественным образом объективен и доказан, то при чем тут его “вес”? Однако факты, доказанные в биологии и медицине, совсем непохожи на факты, доказанные в математике. Математический факт представляет собой логически доказанное утверждение. Например, то, что площадь треугольника равна половине произведения его основания на высоту, доказывается с помощью аксиом и логических умозаключений. Это утверждение верно для любого треугольника. С другой стороны, в биологии и медицине утверждения, касающиеся исследуемых объектов, неизбежно носят вероятностный характер и являются следствием многочисленных и повторяющихся эмпирических наблюдений и экспериментов.

Объекты исследования в биологии и медицине имеют намного более сложное строение, чем объекты изучения математики. Основное отличие между ними заключается в том, что математики сами создают объект своего исследования и вправе поступать с ним по законам собственной логики. В то же время объекты изучения в биологии настолько сложны, что исследователь может лишь усреднить факты и сделать выводы о “типичных” объектах; чем “типичнее” объект, тем точнее вывод исследователя. В этой ситуации единственным инструментом, обеспечивающим непротиворечивость выводов, является статистический анализ.

Понятие о статистическом методе. Роль статистического метода анализа в доказательной медицине

Предположим, что мы проверяем эффективность двух фармакологических препаратов *A* и *B* при лечении некоторого заболевания *D*. Для этого выбираем две группы пациентов по 30 человек, страдающих заболеванием *D*, причем эти две группы постараемся по возможности сделать одинаковыми, так чтобы они

мало отличались друг от друга: возраст пациентов в первой и второй группе должен быть в одинаковых пределах (например от 20 до 40 лет), у пациентов первой и второй группы не должно быть никаких других заболеваний, кроме *Д* и т.д. После проведения курса лечения оказалось, что благодаря применению препарата *А* выздоровело 18 человек, а препарата *В* — 24 человека. У остальных пациентов положительного и отрицательного эффекта не наблюдалось. Означает ли это, что препарат *В* более эффективен, чем препарат *А*? На первый взгляд кажется, что да, поскольку количество выздоровевших пациентов при лечении препаратом *В* больше, чем *А*. Однако тут не все так просто, поскольку при втором клиническом испытании в группах по 30 пациентов, страдающих заболеванием *Д*, может случиться, что при лечении с использованием препарата *А* выздоровело 20 человек, а в группе, где применялся препарат *В* — 17 человек. При этом, если учитывать лишь интуитивные соображения, мы должны прийти к выводу, что препарат *А* более эффективный, чем *В*, и мы получим два противоречивых заключения об эффективности этих препаратов.

В более общем случае, когда группы испытуемых состоят из различного количества пациентов, сравнивают частоту выздоровления: если группа содержит n пациентов и после лечения выздоровело k больных, то частота $h = \frac{k}{n}$. Сравнение частот может свидетельствовать об эффективности препарата, да вот только частота очень изменчива. Если бы мы знали вероятности P_A и P_B выздоровления пациентов при лечении их препаратами *А* и *В* соответственно, то неравенство $P_A > P_B$ означало бы, что препарат *А* более эффективен по сравнению с *В*, поскольку вероятность есть объективная мера возможности события (в данном случае выздоровления пациента) (понятие вероятности будет подробно описано в разделе 1.4). Однако вероятность события, как правило, неизвестна, поэтому возникает проблема сравнения двух вероятностей по соответствующим частотам; (решению этих проблем посвящен раздел 2.12). Сейчас мы укажем лишь, что колебание частот может быть статистически незначимым (тогда соответствующие им вероятности могут быть равными) или статистически значимым (тогда соответствующие вероятности не равны между собой).

Из рассматриваемого выше примера можно убедиться в огромной силе и мощи теории вероятностей и математической статистики: не зная механизма действия фармакологических препаратов, можно сравнивать их эффективность на основании клинических испытаний, которые обязательно необходимы при исследовании новых препаратов, поскольку порождают исходную информацию для статистиков. Более того, достоверные выводы в медицинских исследованиях вообще нельзя сделать без корректно проведенного статистического анализа.

Рассмотрим теперь основные принципы и методы изложения материала в книге, а также проблемы его использования биологами, медиками и математиками. Сразу же отметим, что самая трудная и сложная проблема, которая возникла перед авторами, сделать так, чтобы книга была вполне доступна лицам, окончившим медицинский институт или биологический факультет университета, поэто-

му там нет сложных математических доказательств. Читателю придется просто поверить фактам теории вероятностей и математической статистике, которые там приводятся; однако в книге четко указано, где эти факты строго доказаны, поскольку есть ссылки на соответствующую литературу. Основное внимание уделено применению различных методов статистического анализа в доказательной медицине (диагностика онкологических и внутренних заболеваний, радиационная биология, оценка степени острой лучевой болезни, анализ популяции раковых клеток, резистентных к цисплатину, исследование радиорезистентности и т.д.). Все предлагаемые методы статистического анализа в основном созданы авторами на протяжении более чем 40 лет и являются строго обоснованными.

Что же касается использования книги биологами и медиками при обработке и статистическом анализе экспериментальных и клинических данных, то следует признать, что наиболее эффективный метод — это **совместная** работа биологов, медиков и статистиков от начала (планирования эксперимента — определения метода формирования выборок, вычисление требуемого объема выборок и т.д.) до конца исследования (анализа результатов статистической обработки и их медико-биологическая интерпретация). Мы не питаем иллюзий, что биологи и медики смогут овладеть всеми тонкостями математической статистики и самостоятельно, без помощи статистиков, станут правильно применять те или иные статистические методы для анализа необходимой информации. Цель нашей книги — доказать сомневающимся продуктивность тесного сотрудничества медиков и математиков на всей этапах исследований, подразумевающего четкое разграничением сфер компетенции и ясное понимание основных принципов научного анализа. Если, прочтя нашу книгу, медики станут искать контакты с математиками, а не слепо применять компьютерные программы, получая “на любые вопросы любые ответы”, мы будем считать нашу задачу выполненной.

От издательства “Диалектика”

Вы, читатель этой книги, и есть главный ее критик. Мы ценим ваше мнение и хотим знать, что было сделано нами правильно, что можно было сделать лучше и что еще вы хотели бы увидеть изданным нами. Нам интересны любые ваши замечания в наш адрес.

Мы ждем ваших комментариев и надеемся на них. Вы можете прислать нам бумажное или электронное письмо либо просто посетить наш Web-сервер и оставить свои замечания там. Одним словом, любым удобным для вас способом дайте нам знать, нравится ли вам эта книга, а также выскажите свое мнение о том, как сделать наши книги более интересными для вас.

Отправляя письмо или сообщение, не забудьте указать название книги и ее авторов, а также свой обратный адрес. Мы внимательно ознакомимся с вашим мнением и обязательно учтем его при отборе и подготовке к изданию новых книг.

Наши электронные адреса:

E-mail: info@dialektika.com

WWW: <http://www.dialektika.com>

Глава 1

Основания теории вероятностей

Теория вероятностей и математическая статистика возникли очень давно. Уже в XVII веке выдающийся французский ученый Б. Паскаль уделял много внимания этой науке и сделал много выдающихся открытий в этой области. В XVIII и особенно в XIX столетии, которое справедливо называют “золотым веком”, были созданы основы теории вероятностей в трудах братьев Бернулли, Пуассона, К.Ф. Гаусса, П.С. Лапласа, П.Л. Чебышева, К. Пирсона, А.А. Маркова, Р. Фишера и многих других выдающихся математиков.

Уже с самого начала своего развития теория вероятностей была очень не похожа на другие математические дисциплины. Такое отличие становится заметным даже при первоначальном, поверхностном знакомстве с ней: вместо строго определенных понятий в чистой и в значительной степени прикладной математики в теории вероятностей такие понятия отсутствуют и очень много места занимают метаматематические рассуждения, т.е. рассуждения, которые находятся за пределами математики, совершенно нехарактерные для нее. Эта особенность давно была замечена многими математиками, поэтому в 1900 году одним из самых великих математиков XX столетия Д. Гильбертом на международном конгрессе математиков в Париже была поставлена проблема аксиоматизации теории вероятностей, причем он хотел, чтобы это было сделано таким же способом, как и в геометрии. К сожалению, эта проблема до сих пор не решена, существуют лишь подходы к ее решению.

Однако для приложений теории вероятностей и математической статистики такой особой необходимости нет, поскольку многое из того, что нужно, можно описать на неформальном уровне. Следуя этой логике, ограничимся неформальной теорией вероятностей, которую часто называют описательной или дескриптивной теорией вероятностей. Тут все начинается со случайного испытания, однако описать сущность этого реально существующего понятия можно на основании рассмотрения одного из самых важных и основополагающих понятий современной науки — понятия системы.

1.1. Детерминированные, стохастические и хаотические системы

1.1.1. Понятие системы

Всякая наука занимается исследованием организованных объектов, состоящих из отдельных элементов, которые находятся в определенной взаимосвязи. Такие организованные объекты получили название систем; природа систем может быть самой различной, начиная от простейших механических устройств и кончая самыми сложными биологическими или социальными структурами, однако все системы, как бы сложны они ни были, представляют совокупность отдельных элементов и частей, связанных между собой некоторым образом.

Данное выше описание системы не претендует на логическую строгость и завершенность; по поводу точного определения системы ведутся многочисленные споры и написано много статей (см. например, [35, 36, 88]), но подробное обсуждение этих вопросов едва ли уместно в такой книге, как эта.

Система S может принимать различные состояния $S_1, S_2, \dots, S_n$; под состоянием системы, как правило, понимают всю информацию про ее элементы и части. Одна и та же система в разных ситуациях может вести себя и как детерминированная, и как недетерминированная система, в зависимости от поступающих на нее воздействий. Например, если подкожно ввести подопытному животному смертельную дозу яда, то оно, конечно, погибнет, т.е. система будет вполне детерминированной. С другой стороны, если на кожу этого животного систематически наносить определенную дозу канцерогена, то реакция этой системы будет недетерминированной — злокачественная опухоль может возникнуть, а может и не возникнуть.

Таким образом, понятие недетерминированной системы можно описать следующим образом: предположим, что мы имеем биологическую систему и знаем интересующее нас состояние этой системы до проведения эксперимента (например какого-либо известного фармакологического воздействия). Если после многократного повторения этого эксперимента (воздействия) при неизменных условиях получаются различные реакции исследуемой системы, которые мы не в состоянии предвидеть, то мы имеем дело с недетерминированной системой (рис. 1.1).

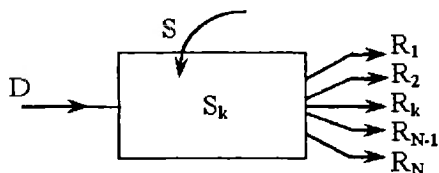


Рис. 1.1. Схема недетерминированной системы

1.1.2. Отчего возникает неопределенность

Попытаемся теперь выяснить причину неопределенности функционирования недетерминированных систем, или, проще говоря, почему они обнаруживают такое странное поведение, когда, несмотря на практически одинаковые воздействия и состояния, получается их совершенно различная реакция. Для того чтобы разобраться в этой, довольно странной ситуации, рассмотрим две достаточно очевидные задачи. Предположим, что на отрезке $[0, 2]$ заданы две простые функции $f_1(x)$ и $f_2(x)$:

$$f_1(x) = x^3,$$
$$f_2(x) = \begin{cases} x^3, & \text{если } x \neq \sqrt{2} \\ 10, & \text{если } x = \sqrt{2} \end{cases}$$

Требуется вычислить их значения в точке $\sqrt{2} = 1,412135\dots$. Точно найти число $\sqrt{2}$ невозможно, это число является иррациональным; оно представляет собой бесконечную непериодическую дробь, в которой десятичные знаки (цифры) изменяются по никому неизвестному закону, однако мы можем определить значение функции $f_1(x)$ в точке $\sqrt{2}$ с любой, практически необходимой для нас точностью. Действительно, рассмотрим приближение к числу $\sqrt{2}$ с точностью до целого, первого десятичного знака (десятых), второго десятичного знака (сотых) и так далее: $x_1 = 1$; $x_2 = 1,4$; $x_3 = 1,41$; $x_4 = 1,412$; $x_5 = 1,4121$; $x_6 = 1,41213$; $x_7 = 1,412135$; ...

Если нам нужно вычислить $f_1(\sqrt{2}) = (\sqrt{2})^3$ с точностью до сотого десятичного знака, то, выбирая приближенное значение с заведомо большей точностью (скажем, до миллионного знака, а существующие персональные компьютеры с помощью определенных программ позволяют это сделать), мы можем возвести это приближенное значение в куб и получить то, что нам нужно. Таким образом, получаем последовательность приближений:

$$x_1^3 = 1; x_2^3 = 3,8416; x_3^3 = 3,9525415; x_4^3 = 3,9750151; x_5^3 = 3,9761412;$$
$$x_6^3 = 3,976479; x_7^3 = 3,9765353; \dots$$

к числу $(\sqrt{2})^3$ и выбираем приближение к $(\sqrt{2})^3$ с необходимой для нас точностью, которая вполне устраивает ту или иную практически важную задачу. Только почему мы уверены, что последовательность приближений $x_1^3, x_2^3, x_3^3, x_4^3$ и так далее сходится или, как говорят, приближается к числу $f_1(\sqrt{2}) = (\sqrt{2})^3$, да при том еще с какой угодно степенью точности? Да потому, что функция $f_1(x) = x^3$ является непрерывной, т.е. ее график представляет собой сплошную линию.

График функции $f_1(x)$ обладает тем свойством, что значением аргументов x_n , близким к точке $x_0 = \sqrt{2} : x_n = x_0$, соответствует значение функции $f_1(x_n)$, которые также близки к значению функции в точке $x_0 : f_1(x_n) \approx f_1(x_0)$. Итак, если функция непрерывна, то мы можем сколько угодно точно вычислить ее значение в точке, скажем, $x_0 = \sqrt{2}$, хотя до вычислений нам это значение было неизвестно (рис. 1.2).

Совершенно иная картина возникает, если мы постараемся вычислить значение функции $f_2(x)$ в точке $\sqrt{2}$ при условии, конечно, когда нам это значение неизвестно (рис. 1.3). Какие бы близкие значения аргумента x_n к точке $\sqrt{2}$ мы ни выбирали, значение функции $f_2(x_n)$ в этих точках не дает никакой информации о значении функции $f_2(x_0) = f_2(\sqrt{2})$, поскольку функция $f_2(x)$ в точке $x_0 = \sqrt{2}$ не является непрерывной, так что значениям аргумента x_n , близким к точке $x_0 = \sqrt{2}$, отвечают значения функции $f_2(x_n)$, которые никоим образом не приближаются к значению функции $f_2(x)$ в точке $\sqrt{2}$ (напомним, $f_2(\sqrt{2}) = 10$).

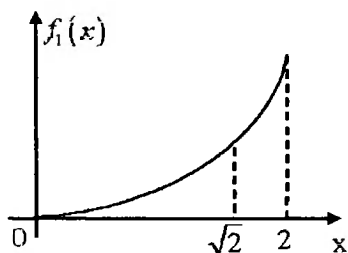


Рис. 1.2. График функции $f_1(x)$

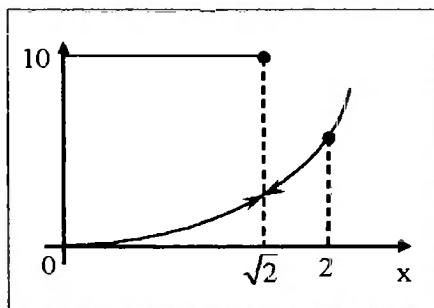


Рис. 1.3. График функции $f_2(x)$

Таким образом, в случае функции $f_2(x)$, которая разрывная в точке $x_0 = \sqrt{2}$, наши знания о ее значении в близких к x_0 точках x_n , т.е. $f_2(x_n)$, не дает нам никакой информации о значении этой функции в точке $x_0 = \sqrt{2}$ и все наши “априорные” знания $f_2(x_n)$ оказываются совершенно бесполезными.

Описанная выше ситуация с вычислением значения непрерывной и разрывной функции в точке весьма похожа на определение реакции R детерминированной и недетерминированной системы, когда известно ее состояние S и поступающее на ее воздействие D . Понятно, что реакция R зависит от S и D и полностью определяется этими факторами, поэтому R является функцией аргументов S и D . Эта функция f называется переходной функцией системы $R = f(D, S)$. Давайте вспомним пример детерминированной системы, описанный в начале этого раздела, — зная исходное состояние подопытного животного и воздействие D (дозу яда, которую ему вводят), мы можем точно предсказать ее

реакцию R (последующее состояние $S(D)$ этой системы). Нетрудно заметить, что для детерминированной системы ее переходная функция $R = f(D, S)$ является непрерывной: если доза яда изменяется незначительно, то и состояние организма изменится незначительно. Даже в случае, когда доза близка к смертельной, состояние организма не изменяется скачкообразно — подопытное животное проходит через клиническую смерть к физической. Итак, зная реакцию $\tilde{R} \approx f(\tilde{D}, \tilde{S})$, когда $\tilde{D} \approx D$, $\tilde{S} \approx S$, мы можем определить с практически нужной нам точностью ее реакцию $R = f(D, S)$, поскольку $\tilde{R} \approx R$.

Картина коренным образом изменяется, если переходная функция $R = f(D, S)$ является разрывной в каждой точке; такая система будет уже недетерминированной. Действительно, если мы знаем реакцию системы $\tilde{R}$, когда воздействие $\tilde{D}$ и состояние $\tilde{S}$ являются близкими к воздействию D и состоянию S , то реакция $R = f(D, S)$ может сколько угодно отличаться от $\tilde{R}$, поэтому реакция R непредсказуема на основании прошлого опыта $\tilde{R} \approx f(\tilde{D}, \tilde{S})$, а поскольку у нас нет никакой информации, кроме опыта, то реакция R вообще непредсказуема. Если вспомнить теперь пример недетерминированной системы, когда подопытному животному на кожу наносят канцероген, то даже при самом детальном исследовании организма (состояние S) и тщательной дозировке (воздействие D) нельзя точно предсказать, возникнет у животного рак или нет.

Таким образом, детерминированные системы — это системы, для которых переходная функция $R = f(D, S)$ непрерывна, а недетерминированная — для которых она разрывная.

1.1.3. Все ли недетерминированные системы обнаруживают одинаковое поведение? Стохастические и хаотические системы

Будем называть испытанием T всякое действие, состоящее в поступлении воздействия D на систему S , находящуюся в состоянии S_0 , и возникающей в ответ на воздействие D реакцию R_0 рассматриваемой системы. Мы назовем эту реакцию R результатом (или исходом) испытания T . Для детерминированной системы, зная ее переходную функцию $R = f(D, S)$, можно определить (предсказать) реакцию этой системы, а для недетерминированной системы в силу соображений, изложенных в предыдущем разделе, этого делать принципиально нельзя. Перечислим основные факторы, которые обуславливают невозможность предсказания реакции недетерминированной системы.

1. Отсутствие полной информации о внутреннем состоянии системы во время (или до) поступления воздействия D .
2. При наличии полной информации о первоначальном внутреннем состоянии системы в некоторых случаях бывает невозможным ее исполь-

зование для прогноза реакции R системы S из-за несовершенства наших знаний.

3. На реакцию R системы S после поступления воздействия D могут оказывать большое влияние неконтролируемые внешние факторы.
4. При повторении испытания T воздействие D не всегда можно сделать абсолютно точным предыдущему.

Если испытание T проводится над детерминированной системой, то оно называется детерминированным испытанием и его результат принципиально всегда можно предсказать или предвидеть. В случае, когда испытание T совершается над недетерминированной системой, оно называется недетерминированным и его результат (исход) предсказать невозможно.

Предположим, что исходом недетерминированного испытания может быть лишь две различные реакции системы S ; поскольку в теории вероятностей реакцию системы принято называть событием и обозначать ее символом A , то в нашем случае результатом или исходом испытания T могут быть лишь два события A и “не A ”, последнее событие обозначается символом $\bar{A}$ и считается как “отрицание события A ”. Проведем n повторений испытания T ; пусть событие A произошло k раз и $n - k$ раз появилось событие $\bar{A}$, тогда отношение

$$h_n(T, A) = \frac{k}{n}$$

называется частотой появления события A в n испытаниях. Рассмотрим поведение частоты $h_n(A)$, когда количество испытаний n неограниченно возрастает, т.е. когда $n \rightarrow \infty$; это поведение может быть двух видов:

- 1) частота $h_n(T, A)$ может совершать неконтролируемые колебания между 0 и 1, не обнаруживая какой-либо тенденции к стабилизации с ростом n (рис. 1.4);

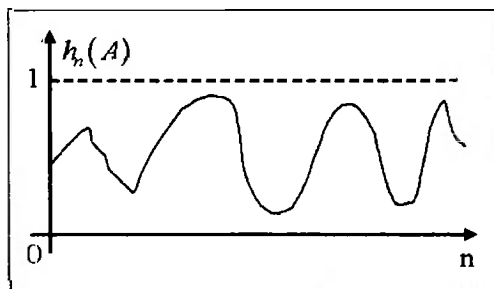


Рис. 1.4. Колебания частоты при хаотических испытаниях

- 2) частота $h_n(T, A)$ при малых n изменяется по неконтролируемому закону, совершая случайные колебания, однако при увеличении количества n

испытаний эти колебания становятся все менее и менее значительными; частота стремится сохранить почти постоянное значение (рис. 1.5).

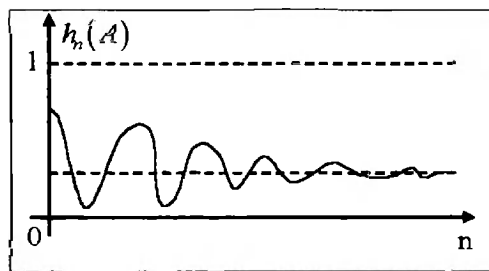


Рис. 1.5. Колебания частоты при случайных испытаниях

Недетерминированные системы, для которых справедливо условие 1, т.е. когда не наблюдается эффекта стабилизации частоты, носят название хаотических; при этом испытания, проводимые над такими системами, называются хаотическими испытаниями, а получающиеся в результате этих испытаний события A — хаотическими событиями. Забегая вперед, отметим, что непосредственное применение к хаотическим системам и процессам теории вероятностей и статистики, без их предварительной обработки, представляется незаконным и недопустимым, поскольку при этом могут возникать ошибочные выводы и противоречивые результаты.

Недетерминированные системы и связанные с ними испытания и события, удовлетворяющие условию 2, для которых выполняется закон стабилизации частоты, называются стохастическими, или случайными.

Перейдем теперь к исследованию и анализу стохастических систем. Прежде всего, определим понятие вероятности $p(T, A)$ случайного события A , которое может происходить при реализации случайного испытания T ; вероятность события представляет объективную меру его возможности; это число должно быть в пределах от 0 до 1, причем более возможные события должны иметь большую вероятность, невозможные события должны иметь нулевую вероятность, а достоверное (т.е. такое, которое всегда происходит) — единичную вероятность. Для вычисления вероятности $p(T, A)$ предположим, что мы можем бесконечное количество раз повторять случайное испытание T ; n -е повторение T обозначим через T_n , так что мы можем наблюдать последовательность $T_1, T_2, \dots, T_n, \dots$. Эта последовательность $\{T_n\}_{n=1}^{\infty}$ называется серией испытаний или случайным экспериментом и обозначается символом E : $E = (T_1, T_2, \dots, T_n, \dots)$. Поскольку по мере увеличения количества испытаний частота стабилизируется и все меньше уклоняется от некоторой постоянной p , то естественно считать, что частота $h_n(T, A)$ стремится к p при $n \rightarrow \infty$

$$\lim_{n \rightarrow \infty} h_n(T, A) = p = p(T, A).$$

Примером эксперимента, в котором происходит стабилизация частоты, является подбрасывание идеальной монеты (табл. 1.1).

Таблица 1.1. Стабилизация колебаний при увеличении количества испытаний

Количество испытаний	Частота выпадений герба	Количество испытаний	Частота выпадений герба	Количество испытаний	Частота выпадений герба	Количество испытаний	Частота выпадений герба
1	0,0000	20	0,6000	300	0,5006	2500	0,4972
2	0,0000	30	0,5666	400	0,5015	3000	0,4997
3	0,0000	40	0,5750	500	0,5220	3500	0,5071
4	0,2500	50	0,5800	600	0,5011	4000	0,5007
5	0,4000	60	0,6000	700	0,5157	5000	0,5005
6	0,5000	70	0,6000	800	0,5002	6000	0,5001
7	0,4285	80	0,5887	900	0,5000	7000	0,5000
8	0,5000	90	0,555	1000	0,5010	8000	0,4999
9	0,5555	100	0,5200	1500	0,4940	9000	0,4998
10	0,6000	200	0,5030	2000	0,4970	10000	0,4989

Согласно определению вероятности, за вероятность события $p(T, A)$ принимается тот предел, к которому стремится частота $h_n(T, A)$, когда количество испытаний неограниченно возрастает. Таким образом недетерминированные системы, как и недетерминированные испытания, распределены на два класса — стохастические системы (испытания) и хаотические системы (испытания); в свою очередь детерминированные системы состоят из двух классов — детерминированные системы с постоянным и переменным исходом. Указанная выше классификация систем представлена на рис. 1.6.



Рис. 1.6. Классификация систем

Резюме

Система — это организованный объект, состоящий из отдельных элементов, находящихся в определенной взаимосвязи. Система S может принимать различные состояния $S_1, S_2, \dots, S_n$. Под состоянием системы подразумевается вся информация о ее элементах. Системы разделяются на детерминированные и недетерминированные. Детерминированные системы характеризуются тем, что в ответ на одно и то же воздействие в ней возникает одна и та же реакция (ответ). У недетерминированной системы реакция на одинаковые воздействия может быть разной. Недетерминированные системы состоят из двух классов — вероятностные (стохастические) и хаотические системы. Для вероятностной системы в ответ на определенные воздействия могут возникать хотя и разные, но относительно устойчивые реакции, частота которых при увеличении количества испытаний (повторяющихся воздействий) стремится к определенному пределу, т.е. стабилизируется после определенного периода неконтролируемых колебаний. У хаотических систем этот эффект не наблюдается и частота совершает неконтролируемые колебания, амплитуда которых с увеличением количества испытаний не изменяется.

1.2. Случайные испытания и случайные события. Сравнение случайных событий. Операции над случайными событиями

В предыдущем разделе мы ввели понятие стохастической системы S ; было указано, что случайное испытание — это комплекс (D, S_0, R) , состоящий из воздействия D , поступающего на стохастическую систему, ее состояния S_0 и реакции этой системы R . В дальнейшем мы будем обозначать случайное испытание символом $T=(D, S_0, R)$ (T — начальная буква английского слова *trial* — испытание). Результат случайного испытания T , т.е. реакция R стохастической системы S в ответ на поступающее воздействие D , называется случайным событием, которое обычно обозначается прописными буквами латинского алфавита A, B, C и т.д., однако более правильно обозначать случайное событие символами (T, A) : на первом месте стоит случайное испытание T , при реализации которого может произойти событие A , а на втором — само событие A . Поясним, почему так следует поступать. Во-первых, случайное событие A может появиться лишь после случайного испытания T ; без случайного испытания нет и не может быть случайных событий, поэтому отрыв события от порождающего его случайного испытания T является некорректным действием. В самом деле, одно и то же случайное событие A может появиться после различных случайных испытаний T_1, T_2 и т.д.; например, саркома (событие A) у подопытных животных (белых мышей) может появиться при подкожном введении бензипирина (случайное испытание T_1) или метилхолантрена (случайное испытание T_2). Поскольку эти два канцерогена относятся к различным классам (первый — слабый канцеро-

ген, а второй — сильный), то эти два события (T_1, A) и (T_2, A) имеют различную меру объективной возможности (т.е. вероятности), поэтому обозначение этих двух событий (T_1, A) и (T_2, A) лишь одним символом A не дает возможности различить эти их важные свойства, при этом возникает путаница. К сожалению, во многих учебниках и монографиях по теории вероятностей изучаются различные модели случайных явлений, в которых указываются случайные события A без указания соответствующих случайных испытаний (при этом они подразумеваются). В таких моделях участвуют лишь результаты испытаний, а не сами испытания, свойства которых не описываются и полностью игнорируются. Как будет показано в дальнейшем, такой отрыв случайных событий A от испытаний T , в которых они осуществляются, может привести к ошибочным результатам или некорректным (бессмысленным) понятиям, поэтому лучше для обозначения случайных событий A использовать символ (T, A) , указывая также случайное испытание T , в результате которого может появиться событие A . Во-вторых, при вычислении вероятности события A обязательно необходимо указать случайное испытание T , при котором оно может произойти, поскольку при различных испытаниях T_1, T_2 событие A будет иметь разные вероятности, и без использования символа (T, A) тут никак не обойтись.

1.2.1. Случайные испытания и случайные события

Рассмотрим произвольное случайное испытание T и обозначим через $S(T)$ множество (пространство) всех случайных событий, которые могут произойти при осуществлении испытания T . Как правило, в каждом конкретном случае описание этого множества $S(T)$ и его структуры не представляет особого труда. Так при бросании игральной кости (случайное испытание T) множество $S(T)$ состоит из невозможного события (скажем, выпадение 7 очков), достоверного события (выпадение любого количества очков), а также всевозможные комбинации из “элементарных” событий — появление 1, 2, 3, 4, 5, 6 очков (например, появление четного числа очков, нечетного, появление комбинаций $\{1, 2, 3\}$, $\{2, 3\}$ и т.д.); при этом $S(T)$ содержит, как это следует из комбинаторики, $2^6 = 64$ события.

1.2.2. Сравнение случайных событий

В множестве $S(T)$, как и в множестве R' всех действительных чисел, можно ввести структуру упорядочения, полагая, что $(T, A) \leq (T, B)$, если появление события (T, A) при осуществлении испытания T непременно влечет появление события (T, B) . Правда, в отличие от числовой прямой R' , где для любых двух чисел a и b можно сказать, какое из них больше, а какое меньше, для любых двух различных случайных событий соотношение “меньше или равно” не всегда определено. Например, в выше указанном примере с игральной костью при ее бросании событие (T, A_2) — выпадение двух очков влечет событие (T, A_1) — выпадение четного числа очков $(T, A_2) \leq (T, A_1)$, однако для таких двух событий,

как выпадение четного числа очков (T, A_e) и нечетного числа очков, соотношение “меньше или равно” не определено и эти два события несравнимы. В связи с этим обстоятельством, когда соотношение “меньше или равно” определено не для любой пары элементов, множество называется полуупорядоченным пространством. Подчеркнем, что отношение полуупорядоченности существует лишь для событий (T, A) и (T, B) , которые принадлежат одному пространству событий $S(T) : (T, A) \in S(T), (T, B) \in S(T)$ (знак $\in$ означает “принадлежит”), если же события $(T, A), (T, A_1)$, взяты из разных пространств $S(T)$ и $S(T_1) : (T, A) \in S(T), (T_1, A_1) \in S(T_1)$, то никакого разумного отношения полуупорядоченности между ними невозможно определить, поскольку нельзя определить то, чего нет. Это обстоятельство будет играть в дальнейшем очень важную роль. Отметим, что структура полуупорядоченности является одной из трех основных, фундаментальных математических структур; ее роль в современной математике невозможно переоценить. Тот факт, что множество случайных событий $S(T)$ представляет полуупорядоченное пространство, имеет основополагающее значение для исследования строения $S(T)$.

1.2.3. Операции над случайными событиями

Для случайных событий (T, A) из пространства событий $S(T)$ можно ввести операции сложения и умножения (произведения) событий. Пусть (T, A) и (T, B) — два события, принадлежащие одному пространству событий $S(T)$, порожденному случайным испытанием T ; суммой $(T, A + B)$ этих событий называется событие (T, C) , которое происходит только тогда, когда появляется хотя бы одно из событий (T, A) или (T, B) в результате проведения случайного испытания T . Например, при бросании игральной кости (испытание T) событие (T, A) — появление двух очков и событие (T, B) — четырех или шести очков в сумме дают событие $(T, C = A + B)$ — появление четного числа очков. Произведением двух событий (T, A) и (T, B) называется событие (T, C) , которое состоит в совместном осуществлении событий (T, A) и (T, B) при проведении случайного испытания T ; таким образом, событие $(T, C = AB)$ появляется тогда и только тогда, когда при реализации T происходит как событие (T, A) , так и (T, B) . Если в рассмотренном выше примере с бросанием игральной кости (T, A) — появление четного числа очков, (T, B) — появление простого числа очков, то их произведение (T, AB) — появление двух очков при осуществлении T . Эти операции сложения и умножения непосредственно можно обобщить на случай произвольного множества событий (как конечного, так и бесконечного); при этом определения суммы и произведения переносятся на этот общий случай дословно. А именно, если $A_\alpha (\alpha \in I)$ — произвольное множество событий, которые могут произойти при осуществлении случайного испытания T , то сумма событий $(T, \sum_{\alpha \in I} A_\alpha)$ состоит в появлении хотя бы одного из событий (T, A_α) , а произведение $(T, \prod_{\alpha \in I} A_\alpha)$ в появлении всех событий (T, A_α) , $\alpha \in I$, при проведении случайного испытания T .

Перейдем теперь к рассмотрению довольно сложной проблемы: какие события можно складывать и умножать? Попытаемся доказать, что эти действия можно проводить лишь над событиями из одного пространства событий $S(T)$, порожденного случайным испытанием T ; выполнение этих операций над событиями, которые могут произойти при реализации случайных испытаний T_1, T_2 , является совершенно незаконным и недопустимым действием. Дело в том, что операции сложения и умножения событий можно однозначно определить на основании отношения упорядочения между событиями $(T, A) \leq (T, B)$, а в свою очередь операции суммы и произведения однозначно определяют отношения упорядочения; это довольно сложный математический факт, который основан, например, в работах [1, 6]. Таким образом, операции сложения и умножения событий можно ввести лишь для того пространства событий, в котором определено отношение упорядочения.

Это можно сделать, как было указано в начале этого раздела, лишь для пространства событий $S(T)$, порожденных одним случайным испытанием (т.е. для множества случайных событий, которые могут произойти при осуществлении испытания T). Действительно, рассмотрим очень простой пример, когда в роли случайного испытания T_1 выступает первое подбрасывание монеты, а в роли T_2 — второе. Однако, если мы попытаемся ввести соотношение упорядочения между событиями (T_1, A_1) и (T_2, A_2) из разных пространств событий $S(T_1)$ и $S(T_2)$, то мы должны позаботиться о том, чтобы для событий (T_1, A_1) и (T_2, A_2) , где A_1 скажем, выпадение герба при первом бросании, выполнялось неравенство $(T_1, A_1) \leq (T_2, A_2)$. Это абсурдно, так как выпадение герба при первом бросании не обязательно влечет его выпадение при втором бросании. Итак, никакого разумного отношения упорядочения между событиями из разных пространств событий $S(T_1)$ и $S(T_2)$ ввести нельзя, а потому невозможно ввести операции сложения и умножения для событий из разных пространств событий $S(T_1)$ и $S(T_2)$. Отметим, что приведенные выше рассуждения о несравнимости случайных событий (T_1, A_1) и (T_2, A_2) можно перенести на случай произвольных событий $(T_1, A) \in S(T_1)$ и $(T_2, B) \in S(T_2)$, если $A, B \neq I$ или I , где I — достоверное событие. Таким образом для любых событий из разных пространств $S(T_1)$ и $S(T_2)$, кроме невозможных и достоверных, вообще нельзя установить отношение упорядочения; они всегда просто несравнимы между собой.

Кроме операций сложения и умножения событий в пространстве событий $S(T)$, можно ввести операцию отрицания (дополнение) (T, A') события (T, A) : появление события (T, A') — это отсутствие события (T, A) после проведения случайного испытания T . Операция дополнения события связана с операцией сложения и умножения следующим образом: дополнения суммы событий равно произведению дополнений, а дополнения произведения равно сумме дополнений. Последнее утверждение носит название “принцип двойственности”, и оно справедливо не только для суммы и произведения двух событий, но и для произвольного множества событий. Принцип двойственности можно записать в виде формулы:

$$(T, (A+B)') = (T, A'B'),$$

$$(T, (AB)') = (T, A' + B');$$

в общем случае

$$\left(T, \left(\sum_{\alpha \in I} A_{\alpha} \right)' \right) = \left(T, \prod_{\alpha \in I} A'_{\alpha} \right),$$

$$\left(T, \left(\prod_{\alpha \in I} A_{\alpha} \right)' \right) = \left(T, \sum_{\alpha \in I} A'_{\alpha} \right).$$

Существует еще ряд известных соотношений для операции сложения, умножения, отрицания событий, доказательство справедливости которых можно найти во многих учебниках по теории вероятностей [54, 122]; приведем наиболее важные из них:

$$(T, A+B) = (T, B+A), (T, AB) = (T, BA),$$

$$(T, A+(B+C)) = (T, (A+B)+C), (T, A(BC)) = (T, (AB)C),$$

$$(T, (A+B)C) = (T, AC+BC), (T, A(B+C)) = (T, AB+AC),$$

$$\left(T, \left(\sum_{i \in I} A_i \right) C \right) = \left(T, \sum_{i \in I} A_i C \right), \left(T, B \left(\sum_{i \in I} A_i \right) \right) = \left(T, \sum_{i \in I} BA_i \right),$$

$$(T, AA') = (T, \emptyset), (T, A+A') = (T, I),$$

$$(T, (A')') = (T, A),$$

где $\emptyset$ невозможное, а I — достоверное событие. Как известно, если в множестве введены операции суммы и умножения элементов, удовлетворяющие указанным естественным условиям, то в таком множестве можно определить отношение упорядоченности, превращающие его в полуупорядоченное пространство.

Перейдем теперь к изучению строения пространства событий $S(T)$, порожденного случайным испытанием T . Наша цель — установить структуру любого случайного события из пространства $S(T)$. Два события (T, A) и (T, B) из пространства $S(T)$ называются несовместимыми, если они не могут появиться при одной реализации случайного испытания T . Легко видеть, что два события (T, A) и (T, B) будут несовместимыми, если их произведение является невозможным событием $(T, AB) = (T, \emptyset)$ (например, любое событие (T, A) и его отрицание (T, A') являются несовместимыми). Рассмотрим произвольное случайное событие $(T, A) \in S(T)$; событие (T, A) называется неатомным, если его можно представить в виде суммы двух несовместимых событий (T, A_1) и (T, A_2) , каждое из которых не является как невозможным событием, так и событием (T, A) :

$$(T, A) = (T, A_1) + (T, A_2) = (T, A_1 + A_2), \\ (T, A_i) \neq (T, A), (T, A_i) \neq (T, \emptyset), i = 1, 2.$$

Представление события (T, A) в виде суммы двух указанных выше событий (T, A_1) и (T, A_2) означает, что мы событие (T, A) как бы расщепили на две части — (T, A_1) и (T, A_2) . Все остальные события $(T, A) \in S(T)$, которые не являются неатомными, называются атомами (точками) пространства событий $S(T)$. Атом (T, A) нельзя расщепить, т.е. представить его в виде суммы двух несовместимых событий, каждое из которых есть его “правильная часть”. Заметим, что слово “атом” переводится с греческого языка как “неделимый”. В теории вероятностей атомы называют словами “элементарный исход” или “элементарное событие”. С помощью понятия атома или элементарного события можно установить структуру любого события из $S(T)$, а именно используя классическую теорему А. Тарского о строении вполне дистрибутивных полных булевых алгебр [96], можно показать [37], что всякое случайное событие (T, A) из пространства $S(T)$ есть сумма “атомов” (T, A_i) : $(T, A) = \left(T, \sum_{i \in I} A_i \right)$, поэтому полуупорядочен-

ное пространство $S(T)$ является атомно порожденной полной булевой алгеброй. Таким образом, атомы, или элементарные исходы, являются теми “кирпичиками”, из которых можно “сложить” любое случайное событие A из пространства $S(T)$. С другой стороны, любая сумма атомов $\left(T, \sum_{i \in I} A_i \right)$ является, как показано в работе [96], случайным событием из пространства $S(T)$, так что множество случайных событий — это всевозможные объединения из атомов (элементарных событий).

В связи с этим в современной теории вероятности фундаментальную роль играет понятие базового множества $B = \{B_i\}_{i \in J}$ случайных событий пространства $S(T)$. Этим термином называют систему событий $B = \{B_i\}_{i \in J}$, которые удовлетворяют следующим условиям:

- 1) все события B_i из B попарно несовместны, $B_i B_j = \emptyset$, если $i \neq j$;
- 2) всякое событие A из $S(T)$ можно представить в виде суммы событий B_i из B :

$$A = \sum_{i \in \bar{j}} B_i \quad (\bar{j} \subset J).$$

Можно доказать, что множество всех атомов (элементарных событий) образует базовое множество в пространстве $S(T)$ всех событий, порожденных случайным испытанием $S(T)$.

Резюме

Случайное испытание T — это комплекс, состоящий из воздействия, поступающего на стохастическую систему, ее состояния и реакции системы A . Случайное событие представляет собой реакцию системы или результат случайного испытания. Случайное испытание и случайное событие образуют недели-

мое целое и обозначаются символами (T, A) . Случайное событие без указания случайного испытания не имеет смысла. Множество всех случайных событий, которые могут произойти при испытании T , обозначается символами $S(T)$. Говорят, что событие (T, A) “меньше” события (T, B) (т.е. $(T, A) \leq (T, B)$), если появление события (T, A) при проведении испытания T непременно влечет событие (T, B) . Сравнение имеет смысл лишь для событий из одного множества $S(T)$. Суммой двух событий (T, A) и (T, B) называется событие $(T, A + B)$, которое происходит тогда и только тогда, когда в результате случайного испытания T появляется хотя бы одно из событий (T, A) или (T, B) . Произведением событий (T, A) и (T, B) называется событие (T, AB) , которое происходит тогда и только тогда, когда происходит как событие (T, A) , так и событие (T, B) . Подчеркнем, что операции сложения и умножения событий можно выполнять только над событиями из одного пространства событий $S(T)$.

1.3. Случайный эксперимент и вероятность случайного события. Поле случайных событий, происходящих при реализации одного случайного эксперимента. Независимые случайные события и эксперименты

1.3.1. Что такое вероятность случайного события

Понятие вероятности случайного события представляется очень сложным и мало изученным; в публикациях, посвященных этой проблеме, есть очень много путаницы и ошибочных утверждений. Это понятие было затронуто в подразделе 1.2.3 предыдущего раздела, теперь настала очередь рассмотреть его более подробно. Заметим, что прежде всего в определении стохастической системы было указано лишь на одно характеристическое свойство — эффект стабилизации частоты при неограниченном увеличении количества повторений случайного испытания T , однако для точного определения стохастической системы и случайного испытания необходимо еще указать второе характеристическое свойство — эффект плотности на отрезке $[0, 1]$ чисел, образованных строками и столбцами так называемой “характеристической матрицы случайного эксперимента” E (серии случайных испытаний), однако подробное описание этого эффекта связано с исследованием чисто математических проблем, рассмотрение которых едва ли уместно в такой книге, как эта [96].

Итак, пусть T — случайное испытание и $S(T)$ — полуупорядоченное пространство случайных событий, порожденное случайным испытанием T . Можно ли определить вероятность случайного события (T, A) , если мы можем наблюдать лишь одну реализацию случайного испытания T ? Ответ на этот вопрос отрицателен: те немногочисленные примеры “априорного” определения вероят-

ности события, которые приводятся в некоторых учебниках по теории вероятностей, основаны на гипотезе о равновероятности элементарных исходов, которая справедлива лишь для идеальных объектов (монеты и игральные кости должны быть сделаны из абсолютно однородного материала; они должны иметь абсолютно правильную геометрическую форму и т.д.).

Для реальных случайных испытаний вероятность события (T, A) можно определить, если мы имеем серию испытаний E , т.е. последовательность повторений $(T_1, T_2, \dots, T_n, \dots)$ одного и того же случайного испытания T . Мы будем называть эту последовательность повторений T случайным экспериментом, так что этот термин является синонимом понятия “серия испытаний”, и использовать символ E для обозначения случайного эксперимента $E = (T_1, T_2, \dots, T_n, \dots)$. Разумеется, осуществить бесконечное повторение случайного испытания T и получить тем самым случайный эксперимент E в реальной действительности нельзя (мы можем осуществить лишь так называемый “усеченный случайный эксперимент $E_N = (T_1, T_2, \dots, T_N)$ ”), однако в математических моделях такая возможность допускается; например, мы можем рассматривать и исследовать множество $N = \{1, 2, \dots, n, \dots\}$ всех натуральных чисел, всех простых чисел и т.д. Предположим, что при n повторениях случайного испытания T интересующее нас событие $A \in S(T)$ произошло k раз, тогда отношение

$$h_n(T, A) = \frac{k}{n}$$

называется частотой появления события (T, A) при n повторениях T (подробности — в подразделе 1.3.3). Если испытание T является случайным, то предел, к которому стремится частота $h_n(T, A)$, когда $n \rightarrow \infty$ существует, и называется вероятностью события A

$$p(E, A) = \lim_{n \rightarrow \infty} h_n(T, A) = \lim_{n \rightarrow \infty} \frac{k}{n}.$$

Поскольку для вычисления вероятности события A следует осуществить весь случайный эксперимент E , то мы будем в дальнейшем для ее обозначения использовать символ $p(E, A)$; таким образом вероятность $p(E, A)$ является функцией двух аргументов — случайного эксперимента E и события A . Необходимость такого обозначения вероятности A вызвана тем, что одно и то же событие A может иметь в различных случайных экспериментах разные вероятности. Например, в классических опытах Брайнера и Шимкина саркома (событие A) у подопытных животных могла появляться при подкожном введении бензипирита (случайное испытание T) и метилхолантрена (случайное испытание $\bar{T}$) в равных дозах. Если рассмотреть соответствующие T и $\bar{T}$ случайные эксперименты $E = (T_1, T_2, \dots, T_n, \dots)$ и $\bar{E} = (\bar{T}_1, \bar{T}_2, \dots, \bar{T}_n, \dots)$, то $p(E, A) \approx 0,5$, а $p(\bar{E}, A) \approx 1$, так что $p(E, A) \neq p(\bar{E}, A)$.

1.3.2. Поле случайных событий

После определения вероятности случайного события пространство случайных событий $S(T)$ превращается в множество, в котором для каждого его элемента (точки) A определено некоторое число — вероятность события $p(E, A)$. В математике такие множества называются полями (вспомним поле тяготения — пространство, в котором каждой точке на материальную массу, сосредоточенную в этой точке, действует сила тяжести, т.е. определен вес этой материальной массы). Таким образом $S(T)$ становится полем событий $S(E)$, где для каждого случайного события, которое мы теперь будем обозначать символом (E, A) , определена его вероятность. Как множества $S(T)$ и поля событий $S(E)$ совпадают; они состоят из одних и тех же событий, но в $S(E)$ для каждого события задана его вероятность $p(E, A)$; в дальнейшем события из поля событий $S(E)$ мы будем обозначать символом $(E, A) : (E, A) \in S(E)$.

Рассмотрим основные свойства вероятности $p(E, A)$; будем предполагать, что случайный эксперимент E является фиксированным, а переменным является второй аргумент — событие A . В случае двух произвольных событий $A, B \in S(E)$ справедлива формула

$$p(E, A + B) = p(E, A) + p(E, B) - p(E, AB), \quad (1.1)$$

если же события A и B несовместны, то событие AB является невозможным, т.е. $p(AB) = 0$, поэтому

$$p(E, A + B) = p(E, A) + p(E, B). \quad (1.2)$$

Эту формулу можно обобщить на случай конечного числа попарно несовместных событий $A_1, A_2, \dots, A_n$ из $S(E)$ ($A_i A_k = \emptyset$, если $i \neq k$):

$$p\left(E, \sum_{k=1}^n A_k\right) = \sum_{k=1}^n p(E, A_k). \quad (1.3)$$

Доказательство справедливости формул (1.1) – (1.3) можно легко получить на основании соответствующих очевидных свойств (1.1) – (1.3) для частоты случайных событий и операции предельного перехода. Свойство (1.2) вероятности $p(E, A)$ называется свойством аддитивности, а (1.3) — свойством конечной аддитивности $p(E, A)$. Интересной является проблема, можно ли в равенстве (1.3) положить $n = \infty$, т.е. будет ли вероятность $p(E, A)$ обладать свойством счетной аддитивности? К сожалению, это не так, несмотря на многочисленные уверения дилетантов, и вероятность не обладает, как правило, свойством счетной аддитивности. Обоснование этого утверждения будет дано в следующем подразделе.

Если событие (E, A') является отрицанием события (E, A) , то $p(E, A') = 1 - p(E, A)$.

1.3.3. Независимые случайные события и эксперименты

Перейдем теперь к описанию одного из самых сложных понятий классической теории вероятностей — понятию независимых событий и независимых случайных испытаний (экспериментов). При исследовании этого понятия возникают два принципиально различных случая:

- 1) события A и B могут осуществляться при проведении одного случайного эксперимента E ;
- 2) события A и B могут осуществляться при проведении двух, вообще говоря, различных экспериментов E_1 (когда может произойти событие A) и E_2 (при котором может реализоваться событие B).

Словосочетание “вообще говоря” означает, что эти эксперименты E_1 и E_2 могут быть одинаковыми $E_1 = E_2 = E$, т.е. проводится повторение одного и того же эксперимента E .

Рассмотрим вначале более простой случай 1. Если события A и B могут произойти при осуществлении одного случайного эксперимента E , то это в некоторых случаях означает, что они могут появиться или не появиться сразу после эксперимента E одновременно; в этой ситуации вообще не возникает проблема, какое из событий произошло первым, а какое — вторым. Например, при бросании игральной кости (случайный эксперимент E) могут появиться события A — выпадение двух очков и событие B — выпадение четного числа очков; очевидно, что тут бессмысленно обсуждать, какое событие произошло раньше. Другой, более интересный для естествоиспытателя пример: при попадании вирусов гриппа в организм человека (случайный эксперимент E) у пациента может повыситься температура (событие A) и он может заболеть гриппом (событие B); при этом события A и B также могут происходить одновременно.

Предположим, что события A и B могут произойти при осуществлении одного и того же случайного эксперимента E ; говорят, что событие A независимо от события B , если появление или отсутствие события B в данном испытании E не влияет на вероятность события A . Событие A считается зависимым от события B , если вероятность события A меняется в зависимости от того, произошло событие B или нет; вероятность события A , вычисленная при условии, что событие B имело место, называется условной вероятностью события A и обозначается символом $p(E, A/B)$.

Например, при бросании игральной кости (случайный эксперимент E) событие A — появление четного числа очков зависит от события B — выпадения простого числа очков (2, 3, 5). Действительно, если событие B произошло, то вероятность события A равна $1/3$, поскольку количество случаев, благоприятных A , при этом равно 1, а количество всех случаев равно 3:

$$p(E, A/B) = 1/3;$$

если же B не произошло, то тогда вероятность события A равна $2/3$ ($\bar{B} = \{1, 4, 6\}$, так как количество всех возможных случаев равно 3, из них благоприятных $A - 2$):

$$p(E, A/\bar{B}) = 2/3.$$

Перейдем теперь к математическому определению понятия зависимости событий A и B . Более точное определение независимых событий можно сформулировать следующим образом: говорят, что событие A независимо от события B , если вероятность события A , вычисленная в предположении, что в данном испытании E появилось событие B , равна вероятности события A , вычисленной в предположении, что событие B в этом испытании не произошло; в противном случае, когда эти вероятности не совпадают, событие A считается зависимым от B . Очевидно, что вероятности, о которых идет речь выше, являются условными вероятностями $p(E, A/B)$ и $p(E, A/\bar{B})$, где через $\bar{B}$ обозначается отсутствие события B в данном испытании. Таким образом, в случае, когда событие A не зависит от B , то

$$p(E, A/B) = p(E, A/\bar{B}), \quad (1.4)$$

так что появление или отсутствие события B в данном испытании E не влияет на вероятность события A , а если событие A зависит от B , то

$$p(E, A/B) \neq p(E, A/\bar{B}) \quad (1.5)$$

и вероятность события A меняется в зависимости от того, произошло событие B или нет.

Нетрудно показать (см., например, [54, с. 14]), что для независимых событий A и B справедлива формула

$$p(E, AB) = p(E, A)p(E, B); \quad (1.6)$$

в случае произвольных событий

$$p(E, AB) = p(E, B)p(E, A/B), \quad (1.7)$$

причем равенство (1.7) является необходимым и достаточным условием (критерием) статистической независимости событий A и B .

Докажем, что событие A будет независимым от B тогда и только тогда, когда $p(E, A/B) = p(E, A)$. Действительно, если выполняется равенство (1.4), то

$$p(E, A/B) = p(E, A/\bar{B}) = \frac{p(E, A/\bar{B})}{p(E, \bar{B})} = \frac{p(E, A)p(E, \bar{B}/A)}{1 - p(E, B)}.$$

Далее

$$1 = p(E, B + \bar{B}/A) = p(E, B/A) + p(E, \bar{B}/A);$$

$$p(E, \bar{B}/A) = 1 - p(E, B/A),$$

поэтому

$$p(E, A/B) - p(E, A/B)p(E, B) = p(E, A) - p(E, A)p(E, B/A);$$

$$p(E, A/B) - p(E, AB) = p(E, A) - p(E, AB);$$

$$p(E, A/B) = p(E, A).$$

Поскольку все проведенные выше рассуждения обратимы, из равенства $p(E, A/B) = p(E, A)$ вытекает соотношение (1.6). Итак, в случае независимости событий A и B ,

$$p(E, A/B) = p(E, A),$$

а для зависимых событий

$$p(E, A/B) \neq p(E, A).$$

Очевидно, что зависимость события A от B определяется случайным экспериментом E ; при одном эксперименте E событие A может зависеть от B , а в другом — нет. Если в данном случайном эксперименте событие A зависит от B (т.е. справедливо неравенство (1.5)), то в свою очередь событие B будет зависеть от события A : $p(E, B/A) \neq p(E, B)$, так что свойство независимости событий является взаимным.

Это свойство взаимной зависимости на первый взгляд кажется парадоксальным. Действительно, реально существующие события происходят в различные моменты времени, например событие B может предшествовать событию A , и тогда выражение “событие B зависит от A ” не согласуется со здравым смыслом, поскольку в философии естествознания существует принцип причинно-следственных отношений: если какое либо явление A происходит раньше явления B , то лишь явление A может оказывать влияние на B , а не наоборот. Короче говоря, причина A должна существовать до появления следствия B . Согласно этой концепции, событие B может зависеть от события A , если, конечно, событие A происходит раньше события B , а обратное влияние исключается. Однако на самом деле ничего парадоксального и сверхъестественного в свойстве взаимных зависимостей случайных событий нет. Рассмотрим эту проблему более подробно. Дело в том, что понятие статистической зависимости (или независимости) не равнозначно общежитейскому или философскому пониманию этого понятия. Подчеркнем, что статистическая зависимость события B от A означает, что появление или отсутствие события A изменяет вероятность появления события B .

Для выяснения смысла понятия статистической зависимости рассмотрим вероятностную систему S , а также ее первоначальное состояние $\hat{S}$ и воздействие J , поступающее на систему. Будем понимать под состоянием $\hat{S}$ внутреннее состояние системы S , а воздействие J будем считать внешним фактором. Фактически случайный эксперимент E представляет собой комплекс “воздействия J и внутреннее состояние $\hat{S}$ ”. Символически обозначим его через “ $J + \hat{S}$ ”: $E = “J + \hat{S}”$. Если бы мы знали этот комплекс “ $J + \hat{S}$ ” абсолютно точно, то, наверное, могли абсолютно точно определить последующее состояние системы, по крайней мере, принципиально. На самом деле, мы этот комплекс “ $J + \hat{S}$ ” не знаем достаточно

точно и можно лишь предположить, что комплекс " $E = J + \hat{S}$ " может принимать много различных состояний $E_1 = "j_1 + \hat{S}_1"$, $E_2 = "j_2 + \hat{S}_2"$ и т.д., а вероятность случайного события A , т.е. число $p(E, A)$, представляет усредненное значение вероятностей $p(E_1, A)$, $p(E_2, A)$ и т.д., причем эти вероятности $p(E_i, A)$ ($i = 1, 2, \dots$) и отличаются друг от друга, хотя для нас эти состояния неразличимы, например из-за несовершенства наших измерительных приборов. Предположим, что вероятность появления события B , которое во времени следует за событием A в состоянии E_k комплекса " $J + \hat{S}$ " ($E_k = "J + \hat{S}_k"$), велика, а для остальных состояний мала. Что можно сказать о вероятности события A , если после этого события появляется B ? Если произошло событие B , пусть вслед за A , то наиболее вероятно то, что комплекс " $J + \hat{S}$ " находится в состоянии $E_k = "J_k + \hat{S}_k"$, но тогда вероятность события A уже не равна $p(E, A)$, а будет определяться числом $p(E_k, A)$, которое может отличаться от $p(E, A)$. Таким образом, появление события B нам как бы раскрывает комплекс " $J_k + \hat{S}_k"$ и изменяет вероятность события A , хотя событие B происходит после появления события A .

Итак, для статистической зависимости случайных событий, которые принадлежат одному полю событий $S(E)$, порожденному случайным экспериментом E , не имеет значения время их появления, эти события происходят как бы "вне времени".

Понятие статистической независимости очевидным образом можно обобщить на случай конечного числа событий $A_1, A_2, \dots, A_k$ из поля событий $S(E)$:

$$p(E, A_1, A_2, \dots, A_k) = p(E, A_1)p(E, A_2) \dots p(E, A_k).$$

Перейдем теперь к определению статистической зависимости событий A и B , которые могут происходить при различных случайных экспериментах E_1 и E_2 , а также статистической независимости случайных экспериментов. Тут ситуация коренным образом меняется и большое влияние оказывает время проведения случайного эксперимента. Предположим, что случайный эксперимент E_1 осуществлялся раньше E_2 ; из этих экспериментов и можно образовать составной эксперимент $E = (E_1, E_2)$, если условиться на первом месте этой пары написать эксперимент E_1 , который предшествовал E_2 . При этом возникает две различные ситуации:

- 1) эксперимент проводится вслед за экспериментом E_1 , не дожидаясь результата эксперимента E_1 (последовательный составной эксперимент E_s);
- 2) эксперимент E_2 проводится вслед за экспериментом E_1 , после появления результата эксперимента E_1 (комбинированный составной эксперимент E_c).

При реализации последовательного составного эксперимента появляется лишь одно случайное событие, а комбинированного случайного эксперимента — два. Рассмотрим сперва последовательный случайный эксперимент E_s . Это новый случайный эксперимент, его поле событий $S(E_s)$ может отличаться

от $S(E_1)$ и $S(E_2)$. Кроме того, поле событий $S(E_s)$ и распределение вероятностей в этом поле $p(E_s, A) (A \in S(E_s))$ существенно зависят от порядка проведения экспериментов E_1 и E_2 , так что $E_s = (E_1, E_2) \neq \tilde{E}_s = (E_2, E_1)$ и $p(E_s, A)$ может не совпадать с $p(\tilde{E}_s, A)$, даже если $S(E_s) = S(E_1) = S(E_2)$. Приведем соответствующий пример. В классических опытах Баутвелла мышам наносят на кожу (выбрита спинка мышей) или скармливают канцерогенный углеводород ДМБА (диметилбензантрацен) (25 мг на спинку или 1 мг при скармливании) в дозе, недостаточной для индуцирования опухолей на протяжении нормального жизненного цикла подопытных животных. После этого мышам ежедневно наносят на спинку 0,5%-ный раствор кротонowego масла (масло семян тропического растения *Croton tiglium*) (эксперимент E_1) или раствор скипидара (эксперимент E_2). Если вначале наносят кротоновое масло, а затем скипидар, т.е. осуществляется последовательный составной эксперимент $E_s = (E_1, E_2)$, то через восемь недель почти у всех мышей (больше 95%) начинают визуально обнаруживать первые папилломы (случайное событие A), а в случае, когда вначале наносят скипидар, а затем кротоновое масло (последовательный составной эксперимент $\tilde{E}_s = (E_2, E_1)$), то у мышей папилломы практически не возникают (отсутствие опухоли наблюдается в более чем 95%). Таким образом, $p(E_s, A) > 0,95$, $p(\tilde{E}_s, A) < 0,05$ и $p(E_s, A) \neq p(\tilde{E}_s, A)$.

Кстати сказать, это обстоятельство подтверждает упомянутое выше правило о том, что при обозначении случайного события A лучше использовать символ (E, A) комплекса “эксперимент и случайное событие”, а также целесообразность обозначения вероятности $p(E_s, A)$.

Аналогично и для комбинированного составного эксперимента $E_c = (E_1, E_2)$ очень важен порядок проведения экспериментов E_1 и E_2 , так как поле событий $S(E_c) \neq S(\tilde{E}_c)$, где $\tilde{E}_c = (E_2, E_1)$, однако и в случае совпадения этих полей событий, так же как и для последовательных составных экспериментов, распределения вероятностей $p(E_c, A)$ и $p(\tilde{E}_c, A) (A \in S(E_c) = S(\tilde{E}_c))$ могут быть различными. В качестве примера такого различия $p(E_c, A)$ и $p(\tilde{E}_c, A)$ можно привести классический эксперимент Баутвелла по индуцированию опухолей кожи с помощью канцерогенов. Мышам наносят на кожу или скармливают канцерогенный углеводород (25 мг на спинку или 1 мг при скармливании) в дозе, недостаточной для образования опухолей на протяжении нормального жизненного цикла подопытных животных (случайный эксперимент E_1). После этого мышам ежедневно наносят на спинку 0,5%-ный раствор кротонowego масла (см. предыдущий пример) в течение восьми недель (случайный эксперимент E_2). После этого примерно через восемь недель начинают обнаруживаться первые папилломы. Если кротоновое масло наносят без предварительной обработки животного канцерогеном, то опухоли не возникают. Следовательно, кротоно-

вое масло в этих условиях не является канцерогеном. Обозначим через P случайное событие, состоящее в образовании папиллом у животного после этого эксперимента, а через $\bar{P}$ — его отрицание. Положим $A = (\bar{P}, P)$, тогда $S(E_c) = S(\bar{E}_c)$ и $p(E_c, A) = p((E_1, E_2), A) = 1$, а $p(\bar{E}_c, A) = p((E_2, E_1), A) = 0$, поэтому $p(E_c, A) \neq p(\bar{E}_c, A)$.

Мы будем в дальнейшем изучать лишь комбинированные составные эксперименты, опуская при их обозначении нижний индекс c : $E_c = E = (E_1, E_2)$ и слово “комбинированный”. Описать поле событий $S(E) = S(E_1, E_2)$, возникающих после составного случайного эксперимента $E = (E_1, E_2)$, в общем случае достаточно сложно, поскольку после проведения эксперимента E_2 вслед за экспериментом E_1 (даже когда результат эксперимента E_1 появился) могут возникать случайные события, которые не принадлежат полю событий $S(E_1)$, однако мы такие составные эксперименты не будем рассматривать. Всюду в дальнейшем предполагается, что при реализации составного эксперимента $E = (E_1, E_2)$ появляющаяся пара случайных событий $(A_1, A_2) = A$ имеет компоненты A_1 и A_2 , которые принадлежат соответствующим полям событий $S(E_1)$ и $S(E_2)$: $A_1 \in S(E_1)$ и $A_2 \in S(E_2)$. Итак, пусть E_1, E_2 — случайные эксперименты, обладающие вышеописанными свойствами, $E = (E_1, E_2)$ — составной эксперимент; будем называть случайное событие $A_2 \in S(E_2)$ статистически независимым событием от $A_1 \in S(E_1)$, если

$$p(E, A) = p((E_1, E_2), (A_1, A_2)) = p(E_1, A_1)p(E_2, A_2). \quad (1.8)$$

Приведем сравнение этой формулы с соотношением (1.6), являющимся фактически определением независимых событий A и B , которые могут появиться при реализации одного случайного эксперимента. Хотя чисто внешне эти формулы мало отличаются, они имеют совершенно различный смысл. Заметим, прежде всего, что в левой части (1.6) участвует произведение событий A, B , а в формуле (1.8) нет никакого произведения событий A_1, A_2 , поскольку для событий, не принадлежащих одному полю событий, понятие “произведение” бессмысленно. Даже если составной эксперимент E представляет повторение одного и того же эксперимента E_1 : $E = (E_1, E_1)$, то понятие независимости в смысле (1.8) не совпадает с понятием независимости, определяемым формулой (1.6). Действительно, если $A_1 = A_2 = A^*$, то при осуществлении одного случайного эксперимента E_1 случайные события $A_1 = A^*, A_2 = A^*$ будут всегда зависимыми, а при повторении эксперимента E_1 (т.е. для составного эксперимента $E = (E_1, E_1)$) эти события могут быть независимыми.

Аналогично определяется статистическая независимость A_1 от A_2 , если случайный эксперимент E_2 проводится раньше E_1 :

$$p(\tilde{E}, \tilde{A}) = p(E_2, E_1), (A_2, A_1) = p(E_2, A_2)p(E_1, A_1) \quad (\tilde{A} = (A_1, A_2)). \quad (1.9)$$

Отметим, однако, что для событий A_1 и A_2 , возникающих в результате проведения различных случайных экспериментов, понятие независимости не явля-

ется взаимным. Может случиться так, что событие A_2 зависит от A_1 при осуществлении составного эксперимента (E_1, E_2) , а A_1 — не зависит от A_2 при эксперименте (E_2, E_1) . Действительно, в рассмотренном выше предыдущем примере экспериментов по индуцированию папиллом у мышей

$$\begin{aligned} 1 &= p((E_1, E_2), (\bar{P}, P)) \neq p(E_1, \bar{P}) p(E_2, P) = 0 \quad (\text{ибо } p(E_2, P) = 0); \\ 0 &= p((E_2, E_1)(P, \bar{P})) = p(E_2, P) p(E_1, \bar{P}) = 0 \quad (A_1 = \bar{P}, A_2 = P). \end{aligned}$$

Если событие $A_1 \in S(E_1)$ и $A_2 \in S(E_2)$ статистически не зависят, т.е. выполняются равенства (1.5) и (1.6), то говорят, что эти события статистически независимы. Таким образом, для статистически независимых событий A_1 и A_2 справедливо равенство

$$p((E_1, E_2), (A_1, A_2)) = p((E_2, E_1), (A_2, A_1)). \quad (1.10)$$

Понятие статистической независимости можно обобщить на случай более чем двух событий $A_1 \in S(E_1), \dots, A_k \in S(E_k)$.

После введения понятия статистически независимых событий можно определить статистически независимые случайные эксперименты E_1 и E_2 ; под этим термином мы будем понимать такие эксперименты E_1 и E_2 , для которых произвольные случайные события $A_1 \in S(E_1)$ и $A_2 \in S(E_2)$ из поля событий $S(E_1)$ и $S(E_2)$ соответственно будут статистически независимыми. Понятие независимости двух случайных экспериментов E_1 и E_2 также можно обобщить на случай произвольного конечного числа n экспериментов $E_1, E_2, \dots, E_n$; если при этом случайные эксперименты $E_1, E_2, \dots, E_n$ и независимы и $E_k = E^*$, то такие эксперименты называются схемой независимых испытаний или серией независимых испытаний (схемой Бернулли).

Укажем в заключение этого подраздела две важные формулы теории вероятностей, которые часто используются в приложениях, однако для этого нужно предварительно ввести понятие гипотез. Говорят, что в данном случайном эксперименте E события $H_1, H_2, \dots, H_n$ образуют полную группу, если их сумма $H_1 + H_2 + \dots + H_n$ является достоверным событием (т.е. событием, которое в результате проведения экспериментов E обязательно происходит). Если к тому же события $H_1, H_2, \dots, H_n$ попарно несовместны, но они называются гипотезами. Имеет место следующая формула, которая называется формулой полной вероятности:

$$p(E, A) = \sum_{i=1}^n p(E, H_i) p(E, A / H_i). \quad (1.11)$$

Кроме того, большую роль в приложениях теории вероятностей играет теорема Байеса, которая утверждает справедливость равенства

$$p(E, H_k / A) = \frac{p(E, H_k) p(E, A / H_k)}{\sum_{i=1}^n p(E, H_i) p(E, A / H_i)}. \quad (1.12)$$

1.3.4. Существует ли математическое понятие случайного эксперимента?

В предыдущих подразделах этого раздела было описано неформальное определение случайного эксперимента E и основанное на нем понятие вероятности $p(E, A)$ случайного события (E, A) . В связи с этим естественно попытаться решить проблему математически точного определения этих понятий. К счастью, это можно сделать в некоторых практически важных частных случаях. Уточним постановку нашей задачи: необходимо установить, существуют ли такие модели серии испытаний $E = (T_1, T_2, \dots, T_n, \dots)$, для которых при неограниченном повторении испытания T частота $h_n(E, A)$ любого случайного события (E, A) сходится в обычном смысле к вероятности события $p(E, A)$, причем эта сходимости должна быть точно доказана, а не быть результатом экспериментальных исследований. В предыдущем подразделе шла речь об испытаниях (модели) Бернулли, состоящей из последовательности одинаковых и независимых в совокупности испытаний $T_1, T_2, \dots, T_n, \dots$, в результате каждого из которых может появиться событие (E, A) с постоянной вероятностью $p(T_n, A) = p(E, A) = p$ и противоположное событие $(T_n, \bar{A})$ с вероятностью $q = 1 - p = p(T_n, \bar{A})$. Согласно закону больших чисел, в этом случае частота $h_n(E, A)$ события (E, A) сходится по вероятности к величине p . Это означает, что для всяких сколь угодно малых $\epsilon > 0$ и $\delta > 0$ существует такой номер N , что при всех $n > N$ $p(|h_n(E, A) - p| > \epsilon) < \delta$. Можно ли отсюда доказать сходимости $h_n(E, A)$ к вероятности p в обычном смысле и тем самым дать логически безупречное определение вероятности $p(E, A)$? К сожалению, нельзя. Во-первых, из сходимости по вероятности последовательности $h_n(E, A)$ к числу p не вытекает обычная сходимости последовательности $\{h_n(E, A)\}$ к числу p (в математике на этот счет есть много контрпримеров), а во-вторых, определение вероятности p с помощью неравенства $p(|h_n(E, A) - p| > \epsilon) < \delta \forall n > N$ логически некорректно, поскольку при определении вероятности нельзя пользоваться понятием “вероятности”, так как при этом мы совершаем ошибку *idem per idem*, т.е. определение понятия через самого себя. Найти выход из этой неприятной ситуации помогает “усиленный закон больших чисел”, установленный французским математиком Э. Борелем. В силу этого закона каждая последовательность $\{h_n(E, A)\}$ сходится к вероятности $p(E, A)$ в обычном смысле с вероятностью единица, т.е. имеет место так называемая “строгая состоятельность”. Это означает, что последовательности $\{h_n(E, A)\}$, которые не сходятся к p , имеют нулевую вероятность, а события, имеющие нулевую вероятность в результате проведения серии испытаний, никогда не появляются, т.е. они не могут произойти и всегда будут наблюдаться лишь те последовательности частоты $\{h_n(E, A)\}$, которые обязательно сходятся к числу p . Таким образом, на основании усиленного закона больших чисел можно утверждать, что последовательность испытаний Бернулли является случайным экспериментом. Интересно отметить, что некоторые специалисты по ма-

тематической статистике ([42], с. 123) считают, что усиленный закон больших чисел “в математической статистике едва ли играет заметную роль”, а выходит, что без этого закона нельзя дать точного определения случайного эксперимента и вероятности случайного события.

Очень сложной является проблема сходимости частоты $h_n(E, A)$ к вероятности $p(E, A)$, когда испытания $(T_1, T_2, \dots, T_n, \dots)$ не являются независимыми. Можно показать, используя теорему С.Н. Берштейна (см. [54], с. 117–118), что в случае, когда зависимость между испытаниями T_n и T_{n+N} бесконечно ослабевает, то частота $h_n(E, A)$ сходится по вероятности к $p(E, A)$ при $n \rightarrow \infty$, однако будет ли сходимость $h_n(E, A)$ к $p(E, A)$ с вероятностью единица — эта проблема до сих пор не решена.

Отметим в заключение, что во многих случаях на вопрос, образует ли последовательность повторений $(T_1, T_2, \dots, T_n, \dots)$ случайного испытания T серию испытаний, когда частота $h_n(E, A)$ сходится по вероятности к $p(E, A)$, нельзя ответить, не выходя за рамки математической теории и не прибегая к рассуждениям, носящим гипотетический характер.

1.3.5. Можно ли складывать случайные величины, возникающие при проведении различных случайных испытаний?

В начале раздела мы уже дали отрицательный ответ на этот вопрос. В этом подразделе мы будем исследовать эту проблему более подробно и покажем, что для независимых случайных экспериментов E_1 и E_2 операции сложения (как и других арифметических действий) можно дать разумную интерпретацию.

Рассмотрим для простоты сложение для двух дискретных случайных величин x и y ; предположим, что x принимает свои значения после проведения случайного эксперимента E_1 , а y — E_2 . Для выполнения операции сложения нужно образовать составной случайный эксперимент из E_1 и E_2 . Как мы знаем, таких экспериментов может быть два — $E = (E_1, E_2)$, $\bar{E} = (E_2, E_1)$. Поскольку эти два эксперимента E и $\bar{E}$, вообще говоря, различны (они будут одинаковыми, если E_1 и E_2 совпадают), случайное событие $A_k = \{x + y = x_i + y_k\}$, которое может произойти при экспериментах E и $\bar{E}$, может иметь различную вероятность: $p(E, A_{ik}) \neq p(\bar{E}, A_{ik})$, поэтому в силу этой неопределенности понятие суммы двух случайных величин $x + y$, вообще говоря, является некорректным.

В процессе реализации составного случайного эксперимента $E = (E_1, E_2)$ могут произойти два случайных события: $A_1 = \{x = x_i\}$, состоящее в том, что x принимает значение x_i , и $A_2 = \{y = y_k\}$ — y принимает значение y_k . Однако, в отличие от произвольных случайных событий A_1 и A_2 , эти значения x_i и y_k можно сложить и получить одну случайную величину $z = x + y$, у которой возможные значения есть: $x_i + y_k = z_{ik}$. Таким образом, при осуществлении случайного эксперимента $E = (E_1, E_2)$ возникает случайная величина z с возможными значе-

ниями $z_{ik} = x_i + y_k$ и распределением вероятностей $p_{ik} = p(E, z = z_{ik}) = p((E_1, E_2), (A_2, A_1))$. Одним из основных свойств операции суммы (сложения) есть ее коммутативность (перестановочность) $x + y = y + x$, и мы, естественно, хотели бы, чтобы это свойство было присуще операции сложения случайных величин. Строго говоря, если случайные эксперименты E_1 и E_2 различны, то операция сложения случайных величин x и y никогда не может быть коммутативной, поскольку случайные величины $z = x + y$ принимают свои значения после проведения случайного эксперимента (E_1, E_2) , а $\bar{z} = y + x$ — случайного эксперимента (E_2, E_1) , которые не совпадают, поэтому говорить о равенстве z и $\bar{z}$ некорректно. С точки зрения здравого смысла, случайные величины z и $\bar{z}$ считаются одинаковыми, если они принимают свои значения в одном случайном эксперименте, причем при каждой реализации этого эксперимента эти значения совпадают. Однако в теории вероятностей две дискретные случайные величины z и $\bar{z}$ отождествляются, если их множества значений и распределение вероятностей совпадают, т.е. они имеют одинаковый закон распределения (такие случайные величины считаются эквивалентными). К сожалению, случайная величина $z = x + y$ принимает те же значения, что и $\bar{z}$, но в общем случае вероятность $\tilde{p}_{ki} = p((E_2, E_1) : z = y_k + x_i)$ не совпадает с вероятностью p_{ik} , так что операция сложения оказывается некоммутативной (даже если условиться считать одинаковыми эквивалентные случайные величины). К счастью, когда случайные эксперименты E_1 и E_2 будут независимы, эти вероятности равны $p_{ik} = \tilde{p}_{ik}$ (подробности — в подразделе 1.4.3), поэтому, если отождествить эквивалентные случайные величины, то можно считать, что операция сложения случайных величин x и y определена корректно. Проводя аналогичные рассуждения, можно показать, что для независимых случайных экспериментов операция сложения будет корректной и для произвольных случайных величин (не обязательно дискретных).

Если случайные эксперименты E_1 и E_2 являются зависимыми, то проблема определения двух случайных величин, возникающих после осуществления E_1 и E_2 , становится достаточно сложной [96].

1.3.6. Является ли распределение вероятностей, порожденное случайной величиной, счетно-аддитивной функцией?

Это очень важный для математиков вопрос, от ответа на который зависят многие свойства случайной величины. К сожалению, многие дилетанты, считающие себя “специалистами” в теории вероятностей, дают на него неправильный ответ и приписывают случайным величинам те свойства, которыми они не обладают, однако начнем все по порядку. Как мы знаем (см. подраздел 1.3.1), одна из интерпретаций случайной величины x утверждает, что x является случайным экспериментом E с числовым базовым множеством B_T —

множеством возможных значений x . Случайная величина x определяет на всевозможных подмножествах $A \in B_T$ распределение вероятностей $p(E, A)$, представляющее событие как функцию двух переменных — случайного эксперимента E и случайного события A (напомним, что случайное событие A наступает тогда и только тогда, когда возможное значение x , наблюдаемое после случайного испытания T , попадает в множество $A \in B_T$, так что событие A — это попадание значений x в множество A). Для экспериментатора изменение функции $p(E, A)$ при изменении первого аргумента E , когда второй аргумент — событие A — не меняется (является фиксированным) представляет наибольший интерес по сравнению с исследованием свойств $p(E, A)$ при фиксированном первом аргументе E . Действительно, в примере из раздела 1.4.1, где описано изменение вероятности инициации саркомы (событие A) у подопытных животных при введении двух различных канцерогенов (случайные эксперименты E и $\tilde{E}$) для экспериментатора очень важно было оценить силу этих двух канцерогенов, т.е. определить, для какого канцерогена вероятность возникновения саркомы будет большей. Следовательно, необходимо было установить, как изменяется вероятность $p(E, A)$, когда меняется первый аргумент $p(\tilde{E}, A)$, а второй аргумент A будет фиксированным. Для специалиста в теории вероятностей ситуация противоположная: считается, что первый аргумент функции $p(E, A)$, т.е. случайный эксперимент E , фиксированный, а переменным является второй аргумент — событие A , и необходимо изучить свойства функции $f(A) = p(E, A)$ относительно аргумента A . Как известно, если события A и B из поля событий $S(E)$ несовместны (т.е. их произведение AB — невозможное событие), то

$$p(E, A + B) = p(E, A) + p(E, B). \quad (1.13)$$

Такие функции, определенные на множестве $S(E)$, называются аддитивными. Функция, заданная на $S(E)$, называется конечно аддитивной, если для любой конечной группы попарно несовместных событий $A_1, A_2, \dots, A_n$ из множества $S(E)$ ($A_i A_k = \emptyset, i \neq k$):

$$p(E, A_1 + A_2 + \dots + A_n) = p(E, A_1) + \dots + p(E, A_n). \quad (1.14)$$

Для вероятности $p(E, A)$ формулу (1.14) можно легко получить из формулы (1.13), применяя метод математической индукции. Значительно более сложной является проблема исследования свойства счетной аддитивности функции $f(A) = p(E, A)$:

$$p(E, A_1 + A_2 + \dots + A_n + \dots) = p(E, A_1) + \dots + p(E, A_n) + \dots,$$

где $A_1 + A_2 + \dots + A_n + \dots$ — последовательность попарно несовместных случайных событий ($A_i A_k = \emptyset$, если $i \neq k$) из поля событий $S(E)$. Прежде чем рассматривать эту проблему, попытаемся выяснить, какие случайные величины x возникают на практике. Такие случайные величины, как правило, возникают за счет измерения или вычисления определенных показателей и их возможными значениями могут быть лишь рациональные числа. Даже если x принимает иррациональное значение, скажем, $x = \sqrt{2}$, все равно при введении таких чисел

в периферийные устройства персональных компьютеров мы вынуждены округлять их точные значения и использовать рациональные числа. Иррациональные числа никому еще не удалось точно вычислить и ввести в компьютер, они представляют лишь математическую абстракцию.

Будем называть случайную величину рациональной, если она может принимать лишь одни рациональные значения. Пусть теперь x — произвольная случайная величина с непрерывной функцией распределения $F(u)$; такие случайные величины называются непрерывными. Покажем, что непрерывная случайная величина x может принимать какое-либо одно фиксированное значение x_0 с нулевой вероятностью:

$$p(E, x = x_0) = 0.$$

В самом деле, для произвольного положительного числа ϵ случайное событие $(E, x = x_0) \leq (E, x_0 - \epsilon < x < x_0 + \epsilon)$, поэтому

$$\begin{aligned} 0 &\leq p(E, x = x_0) \leq p(E, x_0 - \epsilon, x_0 + \epsilon) = \\ &= F(x_0 + \epsilon) - F(x_0 - \epsilon) \rightarrow 0 \text{ при } \epsilon \rightarrow 0, \end{aligned}$$

поскольку функция $F(x_0)$ непрерывна в точке x_0 . Итак, неотрицательное число $p(E, x = x_0)$ может быть меньше сколь угодно малого числа $F(x_0 + \epsilon) - F(x_0 - \epsilon) > 0$, следовательно, $p(E, x = x_0) = 0$. В [96] показано, что для любой непрерывной функции распределения $F(u)$ существует такая рациональная случайная величина x , у которой функцией распределения $F_x(u)$ является данная функция $F(u)$: $F_x(u) = F(u)$ так, что x будет непрерывной случайной величиной. Рассмотрим распределение вероятностей $p(E, A)$, порождаемое на B_T непрерывной рациональной случайной величиной x , и покажем, что это распределение вероятностей не обладает свойством счетной аддитивности. Действительно, если предположить противное, то $B_T \subset Q$, а множество рациональных чисел счетно, поэтому будет не более чем счетным и множество возможных значений x : $B_T = \{x_1, x_2, \dots, x_n, \dots\}$, но тогда

$$\begin{aligned} 1 &= p(E, B_T) = p(E, \{x = x_1\} + \{x = x_2\} + \dots + \{x = x_n\} + \dots) = \\ &= p(E, \{x = x_1\}) + p(E, \{x = x_2\}) + \dots + p(E, \{x = x_n\}) + \dots = \\ &= 0 + 0 + \dots + 0 + \dots = 0, \end{aligned}$$

где через $(E, \{x = x_n\})$ обозначено случайное событие, состоящее в том, что после случайного испытания T случайная величина x принимает значение x_n . Следовательно, $1 = p(E, B_T) = 0$; полученное противоречие показывает, что предположение о счетной аддитивности распределения вероятностей $p(E, A)$ является ложным.

Может показаться, что отсутствие счетной аддитивности у распределения вероятностей (E, A) связано со специфичностью случайной величины x , а именно с тем, что x рациональная случайная величина, однако в [96] доказано, что и для обычных непрерывных случайных величин, принимающих любые действительные значения, нет свойства счетной аддитивности у распределения вероятностей $p(E, A)$, порожденного такими случайными величинами. Таким образом, как правило, несмотря на многочисленные уверения дилетантов, рас-

пределение вероятностей не будет счетно-аддитивной функцией, а обладает лишь свойством конечной аддитивности. Эта точка зрения авторов не является уникальной. Один из самых выдающихся специалистов в теории вероятностей, американский математик Г. Биркгофф писал в своей классической монографии [37, с. 358]: “к сожалению, мера $p_{\infty}[S]$ (в наших обозначениях $p(E, A)$. — *Примеч. авт.*) даже когда она определена, ни для какой заданной последовательности испытаний не обладает свойством счетной аддитивности (имеет место лишь конечная аддитивность)”. Аналогичной точки зрения придерживался известный итальянский ученый Де Финетти, а также ряд других специалистов по теории вероятностей.

Резюме

Случайный эксперимент E — это последовательность (серия) повторений одного испытания T : $E = (T_1, T_2, \dots, T_n)$, $T_k = T$. Вероятность случайного события (T, A) представляет собой предел, к которому стремится частота $h_n(T, A)$ (т.е. отношение $\frac{k}{n}$, где k — количество появления события, а n — количество испытаний T), когда количество испытаний n неограниченно возрастает:

$$p(E, A) = \lim_{n \rightarrow \infty} h_n(T, A) = \lim_{n \rightarrow \infty} \frac{k}{n}.$$

Множество (пространство) случайных событий $S(T)$, в котором для каждого события $A \in S(T)$ указана его вероятность $p(E, A)$, называется полем событий. Говорят, что событие $A \in S(T)$ не зависит от события $B \in S(T)$, если вероятность события A , вычисленная в предположении, что в испытании T появилось событие B , равна вероятности события A , вычисленного в предположении, что в испытании T событие B не произошло:

$$p(E, A|B) = p(E, A|\bar{B}).$$

Здесь символ $\bar{B}$ обозначает отсутствие события B в данном испытании. Для независимых событий A и B выполняется равенство

$$p(E, AB) = p(E, A)p(E, B).$$

Справедлива формула

$$p(E, A + B) = p(E, A) + p(E, B) - p(E, AB).$$

Для несовместных событий имеет место равенство

$$p(E, A + B) = p(E, A) + p(E, B).$$

Кроме того, для любых событий A и B

$$p(E, AB) = p(E, A)p(E, B|A).$$

1.4. Случайные величины и их вероятностные характеристики. Основные неравенства теории вероятностей. Правило 3σ

1.4.1. Случайные величины и их характеристики

В подавляющем большинстве случаев исследование систем сопряжено с измерениями каких-либо показателей или характеристик этих систем. Результатами этих измерений могут быть числа или функции (иногда их графики). Рассмотрим сейчас самую простую ситуацию, когда измеряется значение какого-либо одного показателя биологической системы и получается только одно число; если мы при этом имеем дело с вероятностной системой, то при повторении одного и того же случайного испытания T величина x этого показателя будет, вообще говоря, принимать различные числовые значения, поэтому мы будем получать ряд чисел $x_1, x_2, \dots, x_n$, отличающихся друг от друга (x_k — значение x , полученное после k -го повторения испытания T). Поскольку мы не можем предсказать последующую реакцию вероятностной системы до проведения эксперимента, то нельзя до опыта точно узнать, какое числовое значение будет иметь данный показатель. Случайной величиной называют такую величину, которая после проведения случайного испытания T может принять определенное значение, но неизвестно заранее, какое именно.

Случайная величина x представляет одно из важнейших понятий теории вероятностей, однако она не является ее основным понятием в силу следующих соображений. Рассмотрим случайные испытания T , порождающие пространство событий $S(T)$, в котором базовое множество (множество элементарных исходов) $B_T = \{B_i\}_{i \in J}$, является подмножеством числовой прямой R_1 : $B_T \subset R_1$. Поскольку после случайного испытания T обязательно появляется одно из событий базового множества $B_i (i \in J)$, при реализации испытания T непременно появляется некоторое действительное число. Следовательно, при проведении такого случайного испытания T наблюдается некоторое числовое значение B_i , т.е. значение случайной величины x . Таким образом случайную величину x можно отождествлять со случайным испытанием T с числовым базовым множеством B_T , а это означает, что мы определили случайную величину через основные понятия теории вероятностей; в роли базового множества (множества элементарных исходов) выступает множество B_T всех возможных значений случайной величины x , а пространство событий $S(T)$ состоит из всевозможных подмножеств B_T .

В приложениях теории вероятностей случайные величины могут возникать при проведении случайных испытаний T , у которых элементарные исходы не обязательно являются числами, а представляют собой некоторые объекты (события), обладающие числовыми показателями или характеристиками. Например, при диагностике онкологических заболеваний (рак желудка) с помощью компьютерного морфометрического метода берется соскоб (мазок) клеток

буккального эпителия из ротовой полости, затем из этой популяции эпителиоцитов выделяется одна клетка с интерфазным ядром (случайное испытание T) и измеряются следующие показатели этого эпителиоцита (окраска по Романовскому): площадь клетки (случайная величина x_1), ее периметр (случайная величина x_2), площадь ядра (случайная величина x_3), периметр ядра (x_4), фактор-форма клетки и ядра (x_5, x_6), яркость окраски клетки и ядра (x_7, x_8) и т.д. В этих случаях удобно интерпретировать случайную величину x как некоторую функцию, заданную на базовом множестве элементарных событий $B_T = \{B_i\}_{i \in J}$, $x = x(B_i)$, поскольку при этом каждому элементарному исходу (событию) B_i ставится в соответствие его числовой показатель $x(B_i)$, представляющий значение случайной величины x . Можно показать, что эти две интерпретации понятия случайной величины эквивалентны.

Интерпретация случайной величины как функции на базовом множестве случайного испытания T очень полезна при построении различных моделей теории вероятностей и проведении математических исследований случайных величин. Эта интерпретация позволяет корректно определить основные арифметические действия над случайными величинами. Действительно, пусть T — произвольное случайное испытание, $B_T = \{B_i\}_{i \in J}$ — его базовое множество элементарных событий и $\mathcal{F}(T)$ — совокупность всех числовых функций $x(B_i)$ ($i \in J$), заданных на множестве B_T , тогда для любых двух функций $x(B_i), y(B_i) \in \mathcal{F}(T)$, определены и имеют смысл основные арифметические действия над ними — сложение, умножение, вычитание и деление $z = x + y, v = x - y, w = xy, u = x:y$:

$$\begin{aligned} z(B_i) &= (x + y)(B_i) = x(B_i) + y(B_i), \quad v(B_i) = (x - y)(B_i) = x(B_i) - y(B_i) \\ w(B_i) &= (xy)(B_i) = x(B_i)y(B_i), \quad u(B_i) = (x:y)(B_i) = x(B_i):y(B_i) \\ &\quad (y(B_i) \neq 0, (i \in J)) \end{aligned}$$

Будем называть совокупность $\mathcal{F}(T)$ пространством случайных величин, порожденным случайным испытанием T . Таким образом, если две случайные величины x и y принадлежат одному пространству случайных величин $\mathcal{F}(T)$, порожденному случайным испытанием T , то для них определены все основные арифметические действия. Совершенно иная ситуация возникает, когда случайные величины x и y принадлежат пространствам случайных величин $\mathcal{F}(T_1)$ и $\mathcal{F}(T_2)$, порожденных разными случайными испытаниями T_1 и T_2 ; тогда в общем случае арифметические действия становятся некорректными и бессмысленными операциями. Поясним сказанное на примере операции сложения; предположим, что две функции $x(t) = t$ и $y(t) = t^2$ заданы на отрезке $[0, 1]$, при этом очевидно, что $z(t) = (x + y)(t) = x(t) + y(t) = t + t^2$; однако если функция $y(t) = t^2$ определена на отрезке $[2, 3]$, то операция $x(t) + y(\tilde{t})$ теряет всякий смысл, поскольку аргумент t у функции $x(t)$ принадлежит отрезку $[0, 1]$, а у функции $y(\tilde{t})$ — отрезку $[2, 3]$, и непонятно, в какой точке $\tilde{t}$ определена функция $z = x + y$. Таким образом, две функции можно складывать, т.е. для них опреде-

лена сумма, если они определены на одном множестве. В случае, когда две случайные величины x и y взяты из разных пространств $\mathcal{F}(T_1)$ и $\mathcal{F}(T_2)$: $x \in \mathcal{F}(T_1)$, $y \in \mathcal{F}(T_2)$, функции $x(B_i)$ ($B_i \in B_{T_1}$), $y(\tilde{B}_i)$ ($\tilde{B}_i \in B_{T_2}$), заданы, вообще говоря, на различных базовых множествах B_{T_1} и B_{T_2} элементарных исходов (событий), и операция суммы $x + y$, как и в рассмотренном выше примере, не определена. Аналогичная ситуация возникает, когда рассматриваются остальные операции — вычитание, умножение, деление.

Поскольку определение основных арифметических действий над случайными величинами вызывает такие трудности, естественно поставить вопрос: возникает ли необходимость сложения (вычитания и т.д.) случайных величин из различных пространств $\mathcal{F}(T_1)$ и $\mathcal{F}(T_2)$ в теории вероятностей? Ответ на этот вопрос положителен, так как в предельных теоремах используется операция сложения случайных величин из различных пространств $\mathcal{F}(T_1)$, $\mathcal{F}(T_2)$, ..., например, когда случайные испытания $T_1, T_2, \dots$, образуют схему (модель) Бернулли. В дальнейшем мы покажем, каким разумным образом эту операцию сложения можно определить так, что останутся справедливыми основные свойства арифметических операций, однако для этого необходимо ввести распределение вероятностей для случайных событий из множества (пространства) $S(T)$, порожденного случайной величиной x .

Как указано в предыдущем подразделе, пространство событий $S(T)$ превращается в поле событий $S(E)$, если вместо одного случайного испытания T провести серию испытаний T , т.е. последовательность повторений $T_1, T_2, \dots, T_n, \dots$ испытания T , которая называется серией испытаний или случайным экспериментом, и обозначается символом E . После случайного эксперимента E для каждого случайного события $A \in S(E)$ определена его вероятность $p(E, A)$, так что на $S(T)$ задано некоторое распределение вероятностей. Всякая случайная величина $x \in \mathcal{F}(T)$ имеет свое множество значений $R(x) = \{x(B_i), i \in J\} \subset R_1$. С помощью случайного испытания T и случайной величины x можно построить дополнительное случайное испытание T_x следующим образом: вначале осуществляется случайное испытание T , после этого происходит лишь одно элементарное событие B_i из множества B_T , а затем определяется числовое значение $x(B_i)$, которое является результатом вышеуказанного случайного испытания T_x . Таким образом, элементарными исходами испытания T_x являются точки подмножества $R(x)$, а произвольное случайное событие $\tilde{A}$ из пространства событий $S(T_x)$, порожденное случайным испытанием T_x , можно отождествить с некоторым подмножеством из $R(x)$ (точнее говоря, случайное событие $\tilde{A}$ из пространства $S(T_x)$ происходит тогда и только тогда, когда $x(B_i)$ попадает в множество $\tilde{A}$). Если теперь случайное испытание T заменить на соответствующий случайный эксперимент E , то T_x порождает соответствующий случайный эксперимент E_x , а пространство событий $S(T_x)$ превращается в поле событий $S(E_x)$,

которое как множество совпадает с $S(T_x)$, однако для каждого события $\tilde{A} \in S(E_x)$ теперь уже определена вероятность $p(E_x, \tilde{A})$. Эту вероятность можно вычислить непосредственно на основании следующей формулы:

$$p(E_x, \tilde{A}) = p\left(E, \sum_{i \in k} B_i\right),$$

где суммирование берется по всем индексам i , для которых $x(B_i) \in \tilde{A}$ (это множество индексов мы обозначим через k).

Использование распределения вероятностей $p(E_x, \tilde{A})$, которое порождает случайная величина x в поле событий $S(E_x)$, в самой теории вероятностей и ее приложениях не очень удобно и невыгодно, поскольку событий $\tilde{A}$ достаточно много. В связи с этим возникает мысль сузить класс событий $\tilde{A}$ и записать распределение вероятностей с помощью какой-либо аналитической формулы. К счастью, это действительно удастся сделать с помощью так называемой “функции распределения” $F(u)$ случайной величины x ; при этом вместо всех случайных событий $\tilde{A}$ можно использовать лишь случайные события из $S(E_x)$, порожденные множеством $M_u = R(x) \cap (-\infty, u)$ ($u \in R_x$). Прежде чем формулировать точное понятие функции распределения случайной величины, которая определяется на основании вероятностей событий (E, M_u) , мы опишем один простой класс случайных величин; для этого класса основные числовые характеристики можно вычислить, не прибегая к понятию функции распределения, а используя более простое понятие закона распределения.

Обычно принято различать так называемые “дискретные” и “непрерывные” случайные величины. Возможные значения дискретной случайной величины составляет конечное (или счетное) множество; случайная величина, значения которой сплошь заполняют некоторый отрезок луч или всю прямую, считается непрерывной. Примером дискретной случайной величины может быть количество опухолей после введения в организм подопытного животного (например, мыши) какого-либо канцерогена (например, метилхолантрена, бензипирина или ДМБА) в определенной дозе, достаточной для индуцирования опухолей после окончания инкубационного периода (т.е. в первые моменты, когда опухоли можно визуально обнаружить), а примером непрерывной случайной величины, которая возникает при этом эксперименте, может служить продолжительность инкубационного периода.

Опишем вначале основные вероятностные характеристики и их свойства для дискретных случайных величин. Пусть x — дискретная случайная величина и $x_1, x_2, \dots, x_k$ — ее возможные значения. Тогда тот факт, что x принимает какое-либо одно конкретное значение x_k будет случайным событием (E, A_k) , которое имеет вероятность p_k

$$p_k = p(E, A_k) = p(E, x = x_k).$$

Совокупность чисел $x_1, x_2, \dots, x_n$ и $p_1, p_2, \dots, p_n$ называется законом распределения случайной величины x ; обычно закон распределения записывается в виде таблицы, в верхней строке которой перечислены возможные значения случайной величины, а в нижней — соответствующие им вероятности:

x	x_1	x_2	$\dots$	x_n
p	p_1	p_2	$\dots$	p_n

События $(E, A_1), (E, A_2), \dots, (E, A_n)$ являются несовместными; кроме того, $(E, B) = (E, A_1 + A_2 + \dots + A_n)$ будет достоверным событием, поэтому

$$1 = p(E, B) = p(E, A_1 + A_2 + \dots + A_n),$$

в силу формулы (1.3)

$$p(E, A_1 + A_2 + \dots + A_n) = p(E, A_1) + p(E, A_2) + \dots + p(E, A_n),$$

следовательно,

$$p_1 + p_2 + \dots + p_n = 1 \text{ и } p_k \geq 0 \quad (1.15)$$

Наряду с дискретными случайными величинами, принимающими конечное число значений, в теоретических исследованиях очень часто встречаются случайные величины, множество возможных значений которых образует бесконечную последовательность $x_1, x_2, \dots, x_n, \dots$. Такие случайные величины также считаются дискретными; при этом вместо конечного распределения вероятностей $p_1, p_2, \dots, p_n$ приходится иметь дело с последовательностью $p_1, p_2, \dots, p_n, \dots$, где $p_k = p(x = x_k)$. Условие (1.15) для таких случайных величин принимает вид

$$\sum_{k=1}^{\infty} p_k = 1, \quad p_k \geq 0.$$

Наиболее важным для приложений теории вероятностей к биологии являются следующие распределения.

1. Равномерное распределение; случайная величина x принимает свои значения $x_1, x_2, \dots, x_n$ с одинаковой вероятностью $p_k = \frac{1}{n}$;

2. Биномиальное распределение; случайная величина x , принимающая значения $0, 1, \dots, n$, называется распределенной по биномиальному закону, если

$$p_k = C_n^k p^k q^{n-k},$$

где p — параметр, заключенный между 0 и 1, $q = 1 - p$, $C_n^k = \frac{n!}{k!(n-k)!}$ —

количество сочетаний из n элементов по k (считается, что $0! = 1$);

3. Пуассоновское распределение; случайная величина, распределенная по пуассоновскому закону, принимает значения $0, 1, \dots, n, \dots$ с вероятностью

$$p_k = \frac{\lambda^k}{k!} e^{-\lambda},$$

здесь λ — параметр, $0 < \lambda < \infty$;

4. Логарифмическое распределение; целочисленная случайная величина x (т.е. величина, принимающая целые неотрицательные значения) распределена по логарифмическому закону, если

$$p_k = p(E|x = k) = -\frac{1}{\log(1-q)} \frac{q^k}{k},$$

где $0 < q < 1$;

5. Гипергеометрическое распределение; случайная величина x со значениями $0, 1, \dots, n$ считается распределенной по гипергеометрическому закону, когда

$$p_k = \frac{C_{Np}^k C_{Nq}^{n-k}}{C_N^n},$$

где N, p, q — параметры, такие что $0 < p < 1, p + q = 1, N_p \geq n \geq N_q$.

Знание закона распределения дает полную информацию о случайной величине и поэтому представляет самую информативную характеристику случайной величины. Тем не менее, очень часто оказываются полезными другие, менее информативные, но зато более простые характеристики. Необходимость введения их основана главным образом на трех следующих причинах: во-первых, вероятности $p_k = (k = 1, 2, \dots, n, \dots)$ бывают довольно редко известны исследователю *a priori*, а их приближенное определение, как правило, связано с большими затратами и значительным объемом вычислительной работы. Во-вторых, при больших значениях n получается большой объем числовых данных, так что исследователь не в состоянии усвоить содержащуюся в них информацию; ему трудно работать с этими данными и использовать их при решении конкретных задач. Наконец, в-третьих, довольно часто для получения определенных фактов и результатов не требуется знание закона распределения случайной величины и можно ограничиться более простыми вероятностными характеристиками.

Наиболее простыми (и важными) вероятностными характеристиками дискретной случайной величины x являются математическое ожидание $m(x)$ и дисперсия $D(x)$ (рис. 1.7):

$$\hat{x} = m(x) = \sum_k x_k p_k; D(x) = m[x - m(x)]^2 = \sum_k (x_k - \hat{x})^2 p_k.$$

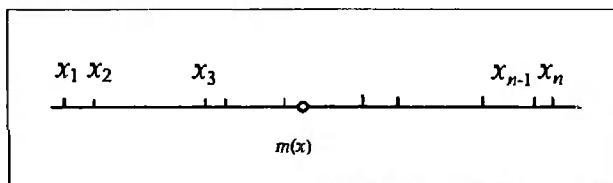


Рис. 1.7. Расположение возможных значений случайной величины относительно ее математического ожидания

Математическое ожидание и дисперсия классических дискретных распределений указаны в табл.; через α обозначена величина $\frac{1}{\log(1-q)}$.

Таблица 1.2. Математическое ожидание и дисперсия некоторых дискретных распределений

Распределение	Равномерное	Биномиальное	Пуассона	Логарифмическое	Гипергеометрическое
Математическое ожидание	$\hat{x} = \frac{1}{n} \sum_{k=1}^n x_k$	np	λ	$-\frac{\alpha q}{p}$	np
Дисперсия	$\frac{1}{n} \sum_{k=1}^n (x_k - \hat{x})^2$	npq	λ	$-\frac{\alpha q(1 + \alpha q)}{p^2}$	$\frac{npq(N-n)}{N-1}$

Перейдем теперь к изучению непрерывных случайных величин. Прежде всего заметим, что для непрерывной случайной величины x , как правило, вероятность события “ x принимает конкретное значение α ” равна нулю. Действительно, рассмотрим случайную величину $\tilde{x}$, равномерно распределенную на отрезке $[0, 1]$; точный смысл понятия “равномерно распределена” мы укажем позже, однако его интуитивный смысл совершенно ясен: это означает, что $\tilde{x}$ принимает значения из отрезка $[0, 1]$ ($0 \leq \tilde{x} \leq 1$), причем ни одно из событий $\tilde{x} = \alpha$ ($\alpha \in [0, 1]$) не является объективно более возможным, чем другое (все значения $\tilde{x}$ “равновозможны”). Тогда вероятность события $a \leq \tilde{x} \leq b$ ($0 \leq a, b \leq 1$), очевидно, равна $b - a$. Положим $a = \alpha - \frac{\epsilon}{2}$, $b = \alpha + \frac{\epsilon}{2}$, где ϵ — достаточно малое положительное число, такое что $0 \leq \alpha + \frac{\epsilon}{2} \leq 1$. Событие $(E, \hat{x}) = \alpha$ влечет за собой появление события $\left(E, \alpha - \frac{\epsilon}{2} \leq \hat{x} \leq \alpha + \frac{\epsilon}{2}\right)$, поэтому

$$0 \leq p(E, \hat{x} = \alpha) \leq p\left(E, \alpha - \frac{\epsilon}{2} \leq \hat{x} \leq \alpha + \frac{\epsilon}{2}\right) = \epsilon.$$

Итак

$$0 \leq p(E, \hat{x} = \alpha) \leq \epsilon,$$

где ϵ — произвольное, сколь угодно малое число, следовательно,
 $p(E, \hat{x} = \alpha) = 0$.

Приведенные рассуждения показывают, что при исследовании непрерывных случайных величин необходимо ввести новую вероятностную характеристику, отличную от набора вероятностей p_k .

Такой характеристикой x является функция $F(u)$, определяемая формулой

$$F(u) = p(x < u).$$

Нетрудно показать, что $F(u)$ — неубывающая функция. Действительно, пусть $a < b$, рассмотрим события $A = \{E, x < a\}$, $B = \{E, x < b\}$ и $C = \{E, a \leq x \leq b\}$; тогда $B = A + C$, причем A и C несовместны. На основании формул (1.1)

$$p(E, B) = p(E, A) + p(E, C),$$

но $p(E, A) = F(a)$, $p(E, B) = F(b)$, $p(E, C) = p(a \leq x < b)$, следовательно,

$$p(E, a \leq x < b) = F(b) - F(a) \geq 0 \text{ и } F(b) \geq F(a). \quad (1.16)$$

Кроме того, $F(u) \rightarrow 0$ при $u \rightarrow -\infty$ и $F(u) \rightarrow 1$ при $u \rightarrow \infty$ (рис. 1.8).

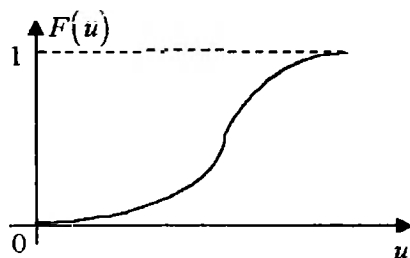


Рис. 1.8. График функции распределения случайной величины

Если $F(u)$ дифференцируемая, то ее производная $F'(u)$ называется плотностью вероятности и обозначается $f(u)$. Название “плотность вероятности” дано в силу следующих соображений. Рассмотрим отрезок $[l, l + \Delta l]$, где Δl — некоторое малое положительное число; из формулы (1.16) вытекает

$$p(E, l \leq x \leq l + \Delta l) = F(l + \Delta l) - F(l),$$

откуда

$$\frac{F(l + \Delta l) - F(l)}{\Delta l} = \frac{p(E, l \leq x \leq l + \Delta l)}{\Delta l}. \quad (1.17)$$

Мы видим, что отношение вероятности $p(E, l \leq x \leq l + \Delta l)$ к длине участка Δl , т.е. средняя вероятность, приходящаяся на единицу длины, равна отношению приращения функции распределения к приращению ее аргумента. В пределе при $\Delta l \rightarrow 0$ средняя вероятность станет плотностью вероятности, а левая часть равенства (1.17) будет равна производной $F'(l) = f(l)$.

Важнейшими как с теоретической, так и с практической точки зрения являются следующие распределения.

1. Равномерное распределение на отрезке $[a, b]$ (рис. 1.9, 1.10):

$$f(t) = \begin{cases} \frac{1}{b-a}, & \text{если } a \leq t \leq b, \\ 0, & \text{если } t < a, t > b; \end{cases}$$

$$F(u) = \begin{cases} 0, & \text{если } u \leq a; \\ \frac{u-a}{b-a}, & \text{если } a < u \leq b, \\ 1, & \text{если } b < u. \end{cases}$$

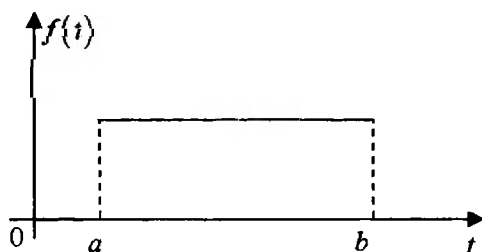


Рис. 1.9. График плотности вероятности равномерного распределения

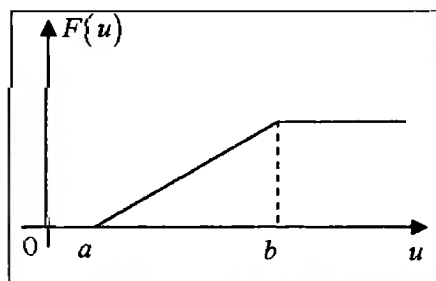


Рис. 1.10. График функции равномерного распределения

2. Нормальное распределение с параметрами $a, \sigma > 0$ (рис. 1.11):

$$f(t) = \left(\frac{1}{\sigma\sqrt{2\pi}} \right) e^{-\frac{1}{2} \frac{(t-a)^2}{\sigma^2}}; \quad F(u) = \Phi(u) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^u e^{-\frac{1}{2} \frac{(t-a)^2}{\sigma^2}} dt.$$

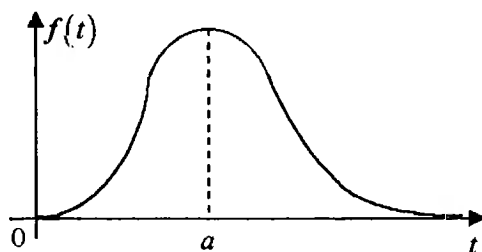


Рис. 1.11. График плотности вероятности нормального распределения

3. Показательное распределение с параметром λ (рис. 1.12):

$$f(t) = \begin{cases} 0, & \text{если } t < 0, \\ \lambda e^{-\lambda t}, & \text{если } t \geq 0; \end{cases}$$

$$F(u) = \begin{cases} 0, & \text{если } u < 0, \\ 1 - e^{-\lambda u}, & \text{если } u \geq 0. \end{cases}$$

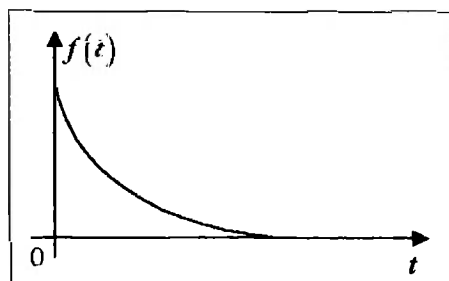


Рис. 1.12. График плотности вероятности показательного распределения

4. Гамма-распределение с параметрами χ и μ , $\chi > 0$:

$$f(t) = \left(\frac{\chi}{\mu}\right)^{\chi} \frac{t^{\chi-1} e^{-\chi t/\mu}}{\Gamma(\chi)}, \quad \text{если } t \geq 0,$$

$$f(t) = 0, \quad \text{при } t < 0;$$

$$F(u) = \begin{cases} 0, & \\ \frac{1}{\Gamma(\chi)} \int_0^{\frac{\chi u}{\mu}} t^{\chi-1} e^{-t} dt, & u \geq 0, \end{cases}$$

где $\Gamma(\chi)$ — гамма-функция [71]. При $\chi = 1$ гамма-распределение превращается в показательное; для $\chi > 1$ плотность вероятности $f(t)$ равна нулю в начале координат и имеет единственный максимум в точке $t = (\chi - 1)/\chi\mu$; если $0 < \chi < 1$, то $f(t)$ равна бесконечности в начале координат.

5. Логнормальное распределение: случайная величина x имеет логнормальное распределение, если $y = \log x$ будет нормально распределенной (рис. 1.13). Плотность вероятности такой случайной величины равна

$$f(t) = \begin{cases} 0, & \text{если } t < 0; \\ \frac{\delta}{\sqrt{2\pi t}} \exp\left\{-\frac{1}{2}(\gamma + \delta \log t)^2\right\}, & \text{если } t \geq 0, \end{cases}$$

где δ, γ — параметры, $\delta > 0$.

Функция $f(t)$ имеет лишь один максимум t_0 , который определяется из уравнения

$$1 + \delta(t_0 + \delta \log t_0) = 0.$$

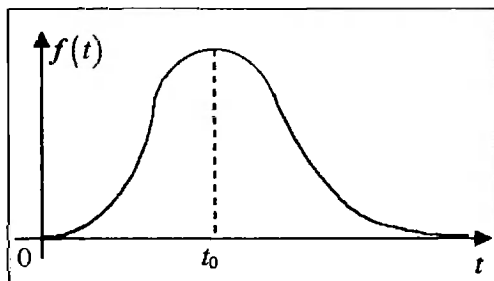


Рис. 1.13. График плотности вероятности логнормального распределения

6. Распределение χ^2 : говорят, что случайная величина x имеет χ^2 -распределение, если ее функция распределения $G(u)$ имеет вид:

$$G(u) = \begin{cases} 0, & \text{если } u < 0, \\ \alpha \int_0^u y^{\lambda-1} e^{-\frac{1}{2}y} dy & \text{при } u \geq 0. \end{cases}$$

где λ — параметр такой, что $f = 2\lambda$ — целое число.

Число $f = 2\lambda$ называется числом степеней свободы распределения $G(u)$; множитель α определяется формулой

$$\alpha = \frac{1}{2^\lambda \Gamma(\lambda)},$$

где

$$\Gamma(\lambda) = \int_0^\infty x^{\lambda-1} e^{-x} dx$$

гамма-функция [71]).

7. Распределение Стьюдента, или t -распределение: говорят, что случайная величина x подчиняется t -распределению с f степенями свободы, если ее плотность вероятности имеет вид

$$f(u) = \left\{ f^{\frac{1}{2}} B\left(\frac{1}{2}, \frac{1}{2}f\right) \left(1 + \frac{u^2}{f}\right)^{\frac{1}{2}(f+1)} \right\}^{-1}, \quad f > 0, \quad -\infty \leq u \leq \infty,$$

где

$$B\left(\frac{1}{2}, \frac{1}{2}f\right) = \frac{\Gamma\left\{\frac{1}{2}(f+1)\right\}}{\sqrt{\pi}\Gamma\left(\frac{1}{2}f\right)}.$$

Математическое ожидание и дисперсия для непрерывных случайных величин определяются с помощью формул

$$m(x) = \int_{-\infty}^{\infty} tf(t)dt,$$

$$D(x) = m[x - m(x)]^2 = \int_{-\infty}^{\infty} [t - m(x)]^2 f(t)dt,$$

где $f(t)$ — плотность вероятности x ; если $\varphi(t)$ — произвольная функция, то

$$m[\varphi(x)] = \int_{-\infty}^{\infty} \varphi(t) f(t)dt \quad (1.18)$$

Данные о математическом ожидании и дисперсии распределений 1-7 приведены в табл. 1.3; для логнормального распределения введены обозначения $\omega = \exp\left(\frac{1}{2}\delta^2\right)$, $p = \exp\left(-\frac{\gamma}{\sigma}\right)$.

Таблица 1.3. Математическое ожидание и дисперсия некоторых конкретных непрерывных распределений

Распределение	Равномерное на отрезке	Нормальное	Показательное	Гамма-распределение	Логнормальное	χ^2 -распределение с f степенями свободы	t -распределение с f степенями свободы
Математическое ожидание $m(x)$	$\frac{a+b}{2}$	a	$\frac{1}{\lambda}$	μ	ωp	f	0
Дисперсия $D(x)$	$\frac{(a-b)^2}{12}$	σ^2	$\frac{1}{\lambda^2}$	$\frac{\mu^2}{\chi}$	$\omega^2 p^2 (\omega^2 - 1)$	$2f$	$f \frac{\Gamma\left(\frac{3}{2}\right)\Gamma\left(\frac{1}{2}f-1\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{1}{2}f\right)}$ $2 < f$

Рассмотрим основные свойства математического ожидания и дисперсии. Оказывается, что математическое ожидание суммы двух произвольных случайных величин равно сумме их математических ожиданий:

$$m(x+y) = m(x) + m(y)$$

Аналогичный результат для произведений не имеет места, однако в некоторых случаях $m(xy) = m(x)m(y)$. Две случайные величины x и y называются независимыми, если для любых действительных чисел a и b события $A = \{E, x < a\}$, $B = \{E, y < b\}$ независимы, т.е.

$$p(E, x < a, y < b) = p(E, x < a)p(E, y < b).$$

Другими словами, независимость x и y означает, что значение x , которое она принимает в результате эксперимента E , не изменяет вероятности события B .

Пусть x и y — независимые случайные величины, тогда

$$m(xy) = m(x)m(y), \quad (1.19)$$

$$D(x \pm y) = D(x) + D(y); \quad (1.20)$$

произвольную константу C можно выносить из под знака дисперсии, возводя ее в при этом в квадрат:

$$D(Cx) = C^2 D(x)$$

Доказательство всех этих утверждений можно найти в учебниках по теории вероятностей (например, [54]).

Реальный смысл понятия математического ожидания заключается в следующем. Можно считать, что все значения случайной величины x образуют некоторое множество (называемое генеральной совокупностью), а эксперимент заключается в том, что мы выбираем наугад какой-либо элемент этого множества, регистрируем его значение и изображаем в виде точки на числовой оси, а затем возвращаем элемент обратно в генеральную совокупность. Если мы проведем указанный эксперимент n раз, то получим n чисел, которые называются выборочными значениями, а все множество этих чисел — выборкой. Можно сказать, что выборочные значения будут располагаться на числовой прямой около точки, являющейся математическим ожиданием $m(x)$.

Таким образом, математическое ожидание представляет собой наиболее “характерное” (или типичное) значение, которое принимает случайная величина. Не следует думать, что на роль этого наиболее “типичного” представителя претендует лишь $m(x)$. В классической математической статистике были и другие показатели меры расположения — это мода и медиана. Мода μ определяется, как значение x , при котором плотность вероятности $f(t)$ достигает своего максимального значения:

$$f(\mu) \geq f(t) \text{ при всех } t.$$

Если f непрерывна и дифференцируема, то мода удовлетворяет условиям

$$\frac{df}{dt}(\mu) = f' \frac{df}{dt} = 0, \quad f''(\mu) = \frac{d^2 f}{dt^2}(\mu) < 0.$$

Медиана μ , представляет такое значение x , которое разделяет множество возможных значений x на две равновероятностные части

$$p(E, x < \mu) = p(E, x \geq \mu).$$

Поскольку события $(E, x < \mu_e)$ и $(E, x \geq \mu_e)$ в сумме дают достоверное событие и несовместны,

$$p(E, x < \mu_e) + (E, x \geq \mu_e) = 1, \quad p(E, x < \mu_e) = \frac{1}{2},$$

следовательно,

$$F(\mu_e) = p(E, x < \mu_e) = \frac{1}{2}.$$

Дисперсия $D(x)$ служит мерой рассеивания возможных значений около ее математического ожидания: чем больше дисперсия, тем больше разбросаны значения случайных величин x вокруг $m(x)$ и наоборот, при малом $D(x)$ с вероятностью, близкой к единице, возможные значения x расположены вблизи $m(x)$. Дисперсия обладает превосходными математическими свойствами, выраженными формулами (1.19) и (1.20), поэтому этим показателем очень удобно пользоваться при теоретических исследованиях. Однако в процессе обработки статистических наблюдений становится заметным, что “эмпирические” свойства дисперсии не настолько хороши, как ее математические свойства.

Поэтому имеет смысл ознакомиться с другими показателями меры рассеяния, которыми пользовались в старой математической статистике, они незаслуженно забыты теперь. Такими показателями являются срединное отклонение $\delta(x)$, определяемое для непрерывной случайной величины x формулой

$$\delta(x) = m|x - m(x)| = \int_{-\infty}^{\infty} |t - m(x)| f(t) dt$$

и для дискретной случайной величины

$$\delta(x) = m|x - m(x)| = \sum_k |x_k - m(x)| p_k,$$

а также абсолютные центральные моменты p -го порядка $\delta_p(x)$ ($1 \leq p < \infty$):

$$\delta_p(x) = m|x - m(x)|^p = \int_{-\infty}^{\infty} |t - m(x)|^p f(t) dt,$$

$$\delta_p(x) = m|x - m(x)|^p = \sum_k |x_k - m(x)|^p p_k$$

и срединное отклонение от медианы $\delta_e(x)$:

$$\delta_p(x) = m|x - \mu_e| = \int_{-\infty}^{\infty} |t - \mu_e| f(t) dt,$$

$$\delta_e(x) = \sum_k |x_k - \mu_e| p_k.$$

Числовые характеристики $\delta_p(x)$ удовлетворяют так называемому “неравенству моментов” [83]

$$[\delta_p(x)]^{\frac{1}{p}} \leq [\delta_q(x)]^{\frac{1}{q}}, \quad 1 \leq p < q < \infty, \quad (1.21)$$

которое является прямым следствием хорошо известного неравенства Гельдера.

Конечно, для показателей $\delta_p(x)$, ($p \neq 2$) и $\delta_c(x)$ свойство (1.20), вообще говоря, не выполняется, но это не мешает им быть хорошими характеристиками меры рассеяния.

1.4.2. Числовые характеристики системы случайных величин

В предыдущем подразделе мы подробно изучали числовые характеристики одной случайной величины. Как было указано, одна случайная величина возникает тогда, когда после эксперимента мы интересуемся числовым значением какого-либо одного показателя; если же для описания реакции биологической системы требуется измерение нескольких показателей, то приходится исследовать несколько случайных величин и изучать их взаимосвязь. Такого рода ситуации бывают довольно часто: например, при изучении реакции нейронов зрительного центра на действие светового раздражителя записывается одновременно импульсная активность двух и нескольких нейронов; ясно, что при этом вместо одной случайной величины x , y или z мы будем получать несколько случайных величин этих типов.

Пусть x и y — две случайные величины, значение которых мы наблюдаем в результате одного эксперимента. Функция двух переменных $F(t, s)$, определяемая формулой

$$F(t, s) = p(E, x < t, y < s),$$

называется совместной функцией распределения пары случайных величин x и y . Функция $F(t, s)$ содержит всю информацию не только о каждой величине x и y , но и об их взаимосвязи. Например, $F(t, s)$ зная, легко получить функцию распределения $F_x(t)$, $F_y(s)$ случайных величин x и y . Действительно

$$F(t, \infty) = p(E, x < t, y < \infty),$$

так как событие $y < \infty$ всегда осуществляется, то произведение событий $x < t, y < \infty$ сводится к событию $x < t$, следовательно,

$$p(E, x < t, y < \infty) = p(E, x < t) = F_x(t),$$

так что

$$F(t, \infty) = F_x(t).$$

Аналогично

$$F(\infty, s) = F_y(s).$$

Если случайные величины x и y независимы, то совместная функция распределения $F(t, s)$ равна произведению функций распределения $F_x(t)$ и $F_y(s)$:

$$F(t, s) = F_x(t)F_y(s). \quad (1.22)$$

В самом деле

$$F(t, s) = p(E, x < t, y < s) = p(E, x < t)p(E, y < s) = F_x(t)F_y(s).$$

Справедливо обратное утверждение: равенство $F(t, s) = F_x(t)F_y(s)$ означает, что случайные величины x и y независимы.

Совместной плотностью вероятности $f(t, s)$ случайных величин x и y называется смешанная частная производная от $F(t, s)$:

$$f(t, s) = \frac{\partial^2 F_x(t)}{\partial t \partial s}.$$

Если $a < b$ и $c < d$, то вероятность события

$$p(E, a \leq x < b, c \leq y < d) = \int_a^b \int_c^d f(t, s) dt ds.$$

Как и в случае обычных функций распределения, совместная функция распределения представляет очень громоздкую и неудобную характеристику пары случайных величин x и y . Поэтому вводят другие, менее информативные, но зато более удобные и компактные характеристики системы x и y . Наряду с математическим ожиданием $m(x)$, $m(y)$ и дисперсиями $D(x)$, $D(y)$, вводят численную характеристику

$$k(x, y) = m[x - m(x)][y - m(y)] \quad (1.23)$$

которая называется коэффициентом ковариации x и y . Формулы для вычисления коэффициента ковариации имеют следующий вид:

$$k(x, y) = \sum_{j=1}^m \sum_{i=1}^n [x_i - m(x)][y_j - m(y)] p_{ij},$$

где $p_{ij} = p(E, x = x_i, y = y_j)$, когда x и y — дискретные случайные величины и

$$k(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [u - m(x)][v - m(y)] f(u, v) du dv,$$

когда x и y — непрерывные случайные величины.

Нетрудно заметить, что в случае независимых x и y

$$k(x, y) = 0;$$

действительно, пусть x и y — независимы, тогда в силу формулы (1.23)

$$\begin{aligned} k(x, y) &= m[x - m(x)][y - m(y)] = m[x - m(x)]m[y - m(y)] = \\ &= [m(x) - m(x)][m(y) - m(y)] = 0 \end{aligned}$$

поэтому коэффициент ковариации $k(x, y)$ в известной мере характеризует зависимость между x и y . Однако при более внимательном анализе формулы (1.23) становится заметным, что $k(x, y)$ характеризует не только зависимость между x и y , но и рассеяние x и y около их математического ожидания. В самом деле, предположим, например, что x принимает значения, которые мало отклоняются от $m(x)$; тогда $x - m(x)$ почти всегда будет близким к нулю: следовательно, коэффициент ковариации $k(x, y)$ будет мал, какой бы тесной зависимостью ни были

связаны между собой x и y . Для того чтобы устранить это отрицательное качество коэффициента ковариации, вводят другой показатель зависимости между x и y , называемый коэффициентом корреляции $r(x, y)$

$$r(x, y) = \frac{k(x, y)}{\sigma(x)\sigma(y)},$$

где $\sigma(x) = \sqrt{D(x)}$, $\sigma(y) = \sqrt{D(y)}$ — средние квадратические отклонения x и y . Этот коэффициент $r(x, y)$ при незначительном рассеянии x и y около своего математического ожидания не будет принимать такое малое значение, как $k(x, y)$. Действительно, при вычислении $r(x, y)$ мы должны поделить $k(x, y)$ на произведение $\sigma(x)\sigma(y)$, которое, в случае малого разброса x (или y) около $m(x)$ ($m(y)$), будет принимать значения, близкие к нулю. Таким образом, $r(x, y)$ является более информативной характеристикой степени зависимости между x и y , чем коэффициент ковариации $k(x, y)$.

Как коэффициент ковариации $k(x, y)$, так и коэффициент корреляции $r(x, y)$ равны нулю для независимых случайных величин. Тем не менее обращение в нуль коэффициента корреляции $r(x, y)$ еще не означает, что x и y независимы; вообще говоря, $r(x, y)$ может быть равен нулю и для зависимых случайных величин x, y . В связи с этим вводят понятие некоррелированных случайных величин: случайные величины x и y называются некоррелированными, если $r(x, y) = 0$.

Нетрудно доказать, что коэффициент корреляции $r(x, y)$ всегда имеет по абсолютной величине значения, не превосходящие единицу, поэтому значения $r(x, y)$ могут меняться от -1 до $+1$.

Как мы уже отметили, из некоррелированных x, y не следует независимость x и y , однако в одном частном случае такое заключение будет правильным. Говорят, что пара случайных величин x, y обладает совместным нормальным распределением, если их совместная плотность вероятности $f(u, v)$ имеет следующий вид:

$$f(u, v) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \exp \left\{ -\frac{1}{2(1-r^2)} \left[\frac{(u-a_1)^2}{\sigma_1^2} - \frac{2r(u-a_1)(v-a_2)}{\sigma_1\sigma_2} + \frac{(v-a_2)^2}{\sigma_2^2} \right] \right\},$$

где $a_1 = m(x)$, $a_2 = m(y)$, $\sigma_1 = \sigma(x)$, $\sigma_2 = \sigma(y)$, $r = r(x, y)$.

Если $r = 0$, то

$$\begin{aligned} f(u, v) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \exp \left\{ -\frac{1}{2} \left[\frac{(u-a_1)^2}{\sigma_1^2} + \frac{2r(u-a_1)(v-a_2)}{\sigma_1\sigma_2} + \frac{(v-a_2)^2}{\sigma_2^2} \right] \right\} = \\ &= f_1(u) f_2(v), \end{aligned}$$

причем $f_1(u)$, $f_2(v)$ представляют плотности случайных величин x и y . Мы видим, что в этом случае совместная плотность вероятности $f(u, v)$ равна произведению плотностей вероятности случайных величин x и y , а это означает независимость x, y .

При более пристальном изучении свойств коэффициента корреляции становится ясным, что $r(x, y)$ характеризует не всякую зависимость между x и y , а только линейную зависимость

$$y = ax + b,$$

где a, b — действительные числа. Можно сказать, что $r(x, y) = 1$ тогда и только тогда, когда

$$y = ax + b, \text{ где } a > 0$$

аналогично равенство $r(x, y) = -1$ эквивалентно условию

$$y = ax + b, \text{ где } a < 0.$$

Поэтому $r(x, y)$ является мерой линейной зависимости между x и y . Другие меры зависимости между x и y мы рассмотрим позже.

Перейдем теперь к краткому описанию вероятностных характеристик системы n случайных величин $x_1, x_2, \dots, x_n$. Исчерпывающую информацию об этой системе представляет совместная функция распределения $F(u_1, u_2, \dots, u_n)$, которая определяется с помощью формулы

$$F(u_1, u_2, \dots, u_n) = p(x_1 < u_1, x_2 < u_2, \dots, x_n < u_n);$$

n -я смешанная производная $F(u_1, u_2, \dots, u_n)$ случайных величин $x_1, x_2, \dots, x_n$:

$$f(u_1, u_2, \dots, u_n) = \frac{\partial^n F(u_1, u_2, \dots, u_n)}{\partial u_1 \partial u_2 \dots \partial u_n}.$$

Минимальное число характеристик, с помощью которых можно описать простейшие вероятностные свойства системы $x_1, x_2, \dots, x_n$, сводится к следующему:

- n математических ожиданий $m(x_1), \dots, m(x_n)$;
- n дисперсий $D(x_1), \dots, D(x_n)$;
- $n(n-1)$ корреляционных коэффициентов $r_{ij} = r(x_i, x_j)$.

Все коэффициенты корреляции удобно расположить в виде таблицы (корреляционной матрицы):

$$R = \begin{vmatrix} 1 & r(x_1, x_2) & \dots & r(x_1, x_n) \\ r(x_2, x_1) & 1 & \dots & r(x_2, x_n) \\ \dots & \dots & \dots & \dots \\ r(x_n, x_1) & r(x_n, x_2) & \dots & 1 \end{vmatrix}$$

1.4.3. Основные неравенства теории вероятностей

Всюду в дальнейшем, когда идет речь лишь об одном и том же случайном эксперименте E , мы будем использовать обозначение $p(A)$ для вероятности события A вместо $p(E, A)$ для сокращения записи.

Пусть x — произвольная случайная величина с конечным математическим ожиданием $m(x)$ и дисперсией $D(x)$. Классическое неравенство Чебышева–Бьенаме, полученное в 1853 году, утверждает, что для любого числа $\lambda > 0$

$$p(|x - m(x)| \geq \lambda) \leq \frac{D(x)}{\lambda^2} \quad (1.24)$$

Это неравенство можно переписать в следующей эквивалентной форме:

$$p(|x - m(x)| \geq \lambda \delta(x)) \leq \frac{1}{\lambda^2}, \quad (1.25)$$

где $\delta(x) = \sqrt{D(x)}$, или

$$p(m(x) - \lambda \delta(x) < x < m(x) + \lambda \delta(x)) > 1 - \frac{1}{\lambda^2};$$

однако при этом вероятность попадания в интервал $J = (m(x) - \lambda \delta(x), m(x) + \lambda \delta(x))$ оценивается очень грубо.

Например, если x — нормально распределенная случайная величина и $\lambda = 3$, то по неравенству Чебышева вероятность α выхода возможных значений x за интервал J (т.е. левая часть неравенства (1.25)) не превышает $\frac{1}{9} = 0,111\dots$, тогда

как на самом деле $\alpha \approx 0,03$, т.е. верхняя граница $0,111$ превышает истинное значение α в 37 раз! В связи с этим после опубликования неравенства Чебышева появилось большое количество неравенств чебышевского типа, в которых была предпринята попытка уточнить и обобщить неравенство Чебышева. Как указывает М. Кендалл и А. Стьюарт: “Вопросы, связанные с подобными неравенствами, привлекательны с математической точки зрения и сами по себе могли бы составить содержание отдельной монографии” ([71], с. 130). Однако все эти модификации неравенства Чебышева не позволили объяснить классическое эмпирическое правило 3σ , которое утверждает, что для распределений, встречающихся на практике, доверительный интервал $(m(x) - 3\sigma(x), m(x) + 3\sigma(x))$ содержит не менее чем 95% значений генеральной совокупности G , т.е.

$$p(|x - m(x)| \geq 3\sigma(x)) \leq 0,05,$$

где $m(x)$, $\sigma^2(x)$ — математическое ожидание и дисперсия случайной величины x соответственно. Не оправдалось также широко распространенное среди специалистов по математической статистике убеждение, что в силу центральной предельной теоремы правило 3σ всегда выполняется для случайных величин x , которые представляют сумму независимых одинаково распределенных случайных величин y_i ($i = \overline{1, n}$), если число слагаемых n достаточно велико. Действительно, как было показано в [44], для произвольной случайной величины x из класса

$$S_n = \left\{ x : x = \sum_{i=1}^n y_i, y_0 \text{ — независимы в совокупности и } Fy_i(u) \equiv F(u) \right\}$$

для всех $x \geq 1$ и каждого натурального числа n справедливо неравенство

$$\sup_{x \in S_n} p(|x - m(x)| \geq \lambda \sigma(x)) \geq \frac{1}{\lambda^2} \left(1 - \frac{1}{\lambda^2}\right);$$

следовательно, для $\lambda = 3$ и произвольного натурального числа n

$$\sup_{x \in S_n} p(|x - m(x)| \geq 3\delta(x)) \geq \frac{8}{81} \approx 0,0987,$$

так что для таких случайных величин x правило 3σ не выполняется ни при каком натуральном n .

Проблема обоснования правила 3σ была успешно решена для всех одновершинных распределений на пути, предложенным К.Ф. Гауссом. Напомним, что случайная величина x называется одномодальной (одновершинной), если ее плотность вероятности $f(u)$ имеет только один локальный максимум $f(u_0)$ (рис. 1.14).

Интересно отметить, что понятие одномодальности можно обобщить на случай, когда у случайной величины не существует плотности вероятности. С этой целью заметим, что для одномодального распределения $F(u)$ плотность вероятности $f(u)$ на луче $(-\infty, u_0)$, т.е. до точки u_0 , которая называется модой, возрастает, а на луче (u_0, ∞) , после точки u_0 , будет убывать. Поскольку $f(u) = F'(u)$ является производной функции распределения $F(u)$, в силу известных теорем математического анализа на луче $(-\infty, u_0)$ $F(u)$ будет выпуклой, а на луче (u_0, ∞) — вогнутой. В связи с этим А.Я. Хинчиным было дано следующее обобщение одномодальности: случайная величина x называется одномодальной, если существует такая точка u_0 (мода x), что на луче $(-\infty, u_0)$ функция распределения $F(u)$ случайной величины x будет выпуклой, а на луче (u_0, ∞) — вогнутой (рис. 1.15).

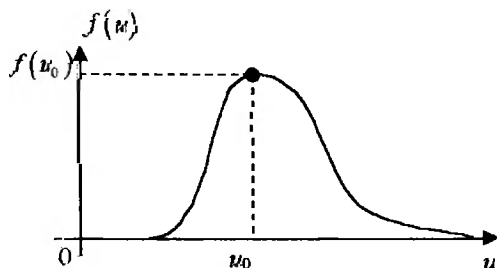


Рис. 1.14. График плотности вероятности одномодального распределения

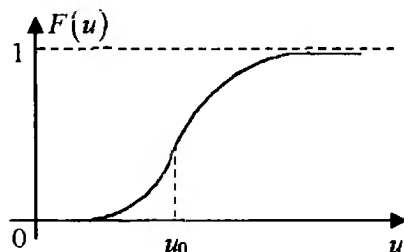


Рис. 1.15. График функции одномодального распределения

Будем говорить, что случайная величина x , обладающая плотностью вероятностей $f(u)$, является симметричной, если существует точка a , для которой для ее плотности вероятности $f(u)$ выполняется равенство $f(a - u) = f(a + u) \forall u \in \mathbb{R}^1$. Примеры симметричных случайных величин представляют нормальные, равномерно распределенные случайные величины, однако большинство случайных величин, встречающихся в теории вероятностей и математической статистике, не являются симметричными.

В 1821 году К.Ф. Гауссом для симметричных одномодальных случайных величин было доказано следующее неравенство, которое носит название неравенства Гаусса:

$$p(|x - m(x)| \geq \lambda \sigma(x)) \leq \frac{4}{9} \frac{1}{\lambda^2}, \quad (\lambda > 0),$$

А в 1979 году в работе [44] было показано, что при $\lambda > \sqrt{\frac{8}{3}}$ неравенство Гаусса будет справедливым для произвольной одномодальной (не обязательно симметричной) случайной величины (в случае $\lambda < \sqrt{\frac{8}{3}}$ неравенство Гаусса, вообще говоря, не имеет места). В связи с этим неравенство

$$p(|x - m(x)| \geq \lambda \sigma(x)) \leq \frac{4}{9} \frac{1}{\lambda^2}, \quad \left(\lambda > \sqrt{\frac{8}{3}} \right) \quad (1.26)$$

которое справедливо для всех одномодальных случайных величин, носит название неравенство Гаусса–Высочанского–Петунина; впоследствии, после выхода работы [44], в работах [22, 24] и монографии [6] были даны другие доказательства неравенства (1.25). Неравенство (1.25) дает возможность обосновать правило 3σ для класса всех одномодальных распределений: поскольку

$$\lambda = 3 > \sqrt{\frac{8}{3}},$$

$$p(|x - m(x)| \geq 3\sigma(x)) \leq \frac{4}{9} \cdot \frac{1}{9} < \frac{4}{80} = 0,05,$$

так что правило 3σ (1.25) для одномодальных случайных величин полностью обосновано.

Резюме

Случайную величину x можно интерпретировать как случайное испытание, после которого появляются числа. Другая интерпретация случайной величины — это числовой показатель так называемых “элементарных событий”, т.е. числовая функция, аргументом которой являются элементарные события, которые невозможно представить в виде суммы двух других случайных событий. Случайные величины, принимающие конечное или счетное множество значений, называются дискретными. Если же значения случайной величины полностью заполняют отрезок, луч или прямую, то они называются непрерывными. Вероятности всех возможных значений дискретной случайной величины $p_k = p(x = x_k)$, $k = 1, 2, \dots, n$, называется ее законом распределения. Непрерывная случайная величина x характеризуется функцией распределения $F_x(u) = p(x < u)$. Одними из важнейших показателей (числовых характеристик) случайной величины x является ее математическое ожидание $m(x)$ (среднее значение) и диспер-

сия $D(x)$ (мера разброса ее возможных значений вокруг математического ожидания): $D(x) = m[x - m(x)]^2$. Если x и y — произвольные случайные величины, то

$$m(x + y) = m(x) + m(y).$$

Если x и y — независимые случайные величины, то

$$m(xy) = m(x)m(y),$$

$$D(x \pm y) = D(x) + D(y).$$

Для двух случайных величины x и y основными характеристиками являются их совместная функция распределения

$$F(u, v) = p(x < u, y < v),$$

коэффициент ковариации $K(x, y) = m\{[x - m(x)][y - m(y)]\}$ и коэффициент корреляции

$$r(x, y) = \frac{K(x, y)}{\sigma(x)\sigma(y)},$$

где $\sigma(x) = \sqrt{D(x)}$, $\sigma(y) = \sqrt{D(y)}$. Если случайные величины x и y независимы, то $K(x, y) = r(x, y) = 0$.

Среди многочисленных неравенство теории вероятностей для математической статистики наибольший интерес представляет неравенство Гаусса–Высочанского–Петунина, которое утверждает, что для любых одномодальных случайных величин x

$$p(|x - m(x)| \geq \lambda \sigma(x)) \leq \frac{4}{9} \frac{1}{\lambda^2}, \quad \text{если } \lambda > \sqrt{\frac{8}{3}}.$$

Одномодальное распределение $F(u)$ — это такое распределение, у которого существует точка a , такая что при $u < a$ функция $F(u)$ является выпуклой, а при $u > a$ — вогнутой. Это неравенство дает возможность обосновать классическое правило 3σ , утверждающее, что для одномодальных распределений $p(|x - m(x)| \geq 3\sigma(x)) = 0,05$.

1.5. Случайные процессы и последовательности

1.5.1. Вероятностные характеристики случайных процессов и корреляционные функции

До сих пор мы рассматривали случайные величины, вероятностные свойства которых не изменяются с течением времени. Такой “статичный” подход к исследованию вероятностных систем был характерен для раннего периода развития теории вероятностей. В последнее время за счет приложений математики к естествознанию, медицине и технике появилась необходимость исследовать случайные величины, изменяющиеся во времени; такие случайные величины получили название случайных функций, или случайных процессов. Поэтому

теорию случайных процессов следует рассматривать как “динамическую” часть теории вероятностей.

Случайный процесс $x(t)$, $t \in [0, T]$, можно определить как случайное испытание T , у которого базовое множество событий B_T (элементарные события) состоит из функций $x^{(i)}(t)$, $t \in [0, T]$, определенных на отрезке $t \in [0, T]$ и принимающих числовые значения. Отрезок $t \in [0, T]$ может быть бесконечным $t \in [0, \infty]$ или даже совпадать со всей числовой прямой. Таким образом, после проведения случайного испытания T мы получим одну функцию $x^{(i)}(t)$ ($t \in [0, T]$), а при повторении T мы будем иметь новые функции $x^{(2)}(t)$, $x^{(3)}(t)$ и так далее, которые называются реализациями, или траекториями случайного процесса. Если проведем серию испытаний, $E = (T^1, T^2, \dots, T^n)$, т.е. осуществим случайный эксперимент E , то мы получим последовательность реализаций $x^{(1)}(t)$, $x^{(2)}(t)$, ..., $x^{(n)}(t)$, ..., где верхний индекс n показывает номер реализации (рис. 1.16); случайный эксперимент позволяет определить вероятностные характеристики случайного процесса, которые мы опишем ниже.

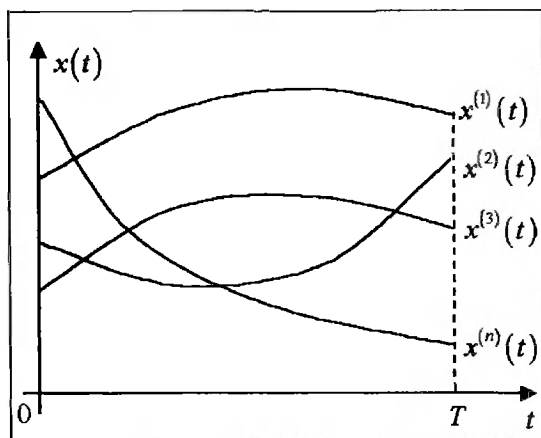


Рис. 1.16. Реализация случайного процесса

Случайная функция обычно обозначается символом $x(t)$; переменная t означает время, а $x(t)$ при фиксированном значении момента времени t является случайной величиной. Случайную функцию $x(t)$ принято называть случайным процессом, если время t меняется непрерывно и пробегает некоторый отрезок, луч или всю числовую прямую. В том случае, когда t пробегает некоторый дискретный ряд чисел (т.е. $t = t_1, t_2, \dots, t_n, \dots$), $x(t)$ называют случайным процессом с дискретным временем или случайной последовательностью и обозначают $x(t_1), x(t_2), \dots, x(t_n), \dots$

Случайные процессы и последовательности встречаются тогда, когда нам приходится измерять какие-либо числовые характеристики вероятностных систем в течение более или менее продолжительного интервала времени. Результат одного наблюдения этой случайной числовой характеристики, представляющей собой некоторую функцию или числовую последовательность, называется

реализацией (траекторией) случайного процесса или последовательности. Таким образом, реализация случайной функции $x(t)$ есть результат наблюдения значений $x(t)$ при одном эксперименте (рис. 1.17).

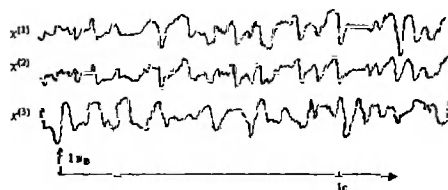


Рис. 1.17. Электрограмма фибриллирующего сердца собаки: три кривые, записанные в течение одной секунды каждая

Обозначим реализации случайной функции $x(t)$ через $x^{(1)}(t), x^{(2)}(t), \dots, x^{(n)}(t)$, где верхний индекс n указывает номер эксперимента, при котором наблюдается данная реализация. Мы имеем дело с наблюдениями случайной функции, поэтому очевидно функции $x^{(1)}(t), x^{(2)}(t), \dots, x^{(n)}(t)$ отличаются, и мы получаем n различных реализаций.

Типичным примером случайной последовательности является спонтанная (самопроизвольная) импульсная активность нейронов мозга лягушки (табл. 1.4).

Таблица 1.4. Пять реализаций случайного процесса, описывающего импульсную активность нейрона среднего мозга лягушки

N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	2,0	2,0	2,3	3,2	2,3	2,1	2,2	2,1	1,7	1,9	2,1	2,6	2,3	2,2	3,2	2,5	3,0	2,1
2	2,2	1,6	2,3	2,3	2,0	2,3	2,6	1,8	3,0	3,0	1,9	2,6	3,0	2,1	2,2	2,2	2,8	2,3
3	2,6	3,0	2,5	2,8	2,6	2,5	1,6	1,5	1,3	0,7	0,6	0,8	0,4	0,8	0,4	0,8	1,3	0,4
4	2,2	0,9	4,9	2,2	2,6	2,2	2,6	2,3	2,9	2,8	2,5	1,6	3,1	2,2	2,8	2,2	2,1	2,8
5	2,4	2,2	2,7	2,5	2,6	2,1	3,0	1,8	2,5	2,2	2,4	2,6	2,4	2,9	1,8	2,5	2,2	2,2

Перейдем теперь к проблеме вероятностных характеристик случайного процесса. Сразу же следует отметить, что это очень сложная и трудная проблема, которая еще далека от своего окончательного решения. Дело в том, что правильный выбор исходных данных во многом предопределяет тот класс задач, который в состоянии решить данная теория, а потому и ее дальнейшее успешное развитие.

Принципиально вопрос об описании случайной функции $x(t)$ решается довольно просто. Рассмотрим произвольный момент времени t_0 ; $x(t)$ при фиксированном значении t_0 будет случайной величиной, поэтому для описания вероятностных свойств $x(t)$ необходимо, прежде всего, задать функции распределения $x(t)$:

$$F_{t_0}(u) = p(x(t_0) < u).$$

Ясно, что, зная $F_t(u)$, при любом значении t мы будем иметь полную информацию о каждой отдельной случайной величине $x(t)$, но все семейство одномерных функций распределения $F_t(u)$ не даст никаких сведений о зависимости между случайными величинами $x(t)$ при различных значениях моментов времени. Так как в общем случае эта зависимость существует и играет очень важную роль, необходимо еще при любых фиксированных значениях $t_1, t_2, \dots, t_n$ знать совместную функцию распределения случайных величин $x(t_1), \dots, x(t_n)$:

$$F_{t_1, \dots, t_n}(u_1, \dots, u_n) = p(x(t_1) < u_1, \dots, x(t_n) < u_n) \quad (1.27)$$

Будем считать, что случайная функция $x(t)$ полностью определена, если для любого конечного набора $t_1, \dots, t_n$ задана совместная функция распределения (1.27).

Изложенный подход к заданию случайной функции принадлежит русскому математику Е.Е. Слуцкому. Естественно, что принципиальное решение проблемы описания случайной функции ни в коей мере не означает ее полного решения, которое было бы пригодно для исследования всех практически важных задач. Мы уже говорили о трудностях, связанных с использованием функции распределения в качестве вероятностной характеристики случайной величины. Заметим, что при использовании многомерных функций совместно-го распределения $F_{t_1, \dots, t_n}(u_1, \dots, u_n)$ эти трудности резко возрастают. Во-первых, в реальной ситуации мы не можем получить все многомерные функции совместного распределения (или оценки этих функций), так как их бесконечно много. Во-вторых, табулирование одной функции $F_{t_1, \dots, t_n}(u_1, \dots, u_n)$ связано с заданием огромного количества чисел: если считать, что одномерная функция распределения $F_t(u)$ удовлетворительно определяется своими значениями в 10 точках, то для двумерной функции распределения $F_{t_1, t_2}(u_1, u_2)$ необходимо знать ее значение в узлах сетки, состоящей из 100 точек, для трехмерной функции $F_{t_1, t_2, t_3}(u_1, u_2, u_3)$ необходимо уже 1000 значений и т.д. Ясно, что исследователь не в состоянии запомнить и осмыслить такую огромную по своему объему информацию. В связи с этим возникает довольно сложная проблема отыскания других, более простых и удобных характеристик случайной функции, которые были бы, тем не менее, достаточно информативными. Пользуясь этими характеристиками в качестве исходных данных, мы должны уметь удовлетворительно решить определенный класс теоретически или практически важных задач. Кроме того, нужно позаботиться о том, чтобы эти характеристики легко определялись из эксперимента, на основании реализаций случайной функции.

В настоящее время в этом направлении сделано немало: мы имеем интенсивно развитые теории стационарных случайных процессов (за исходные вероятностные характеристики здесь принимается математическое ожидание и корреляционная функция) и теорию марковских процессов (исходными вероятностными характеристиками здесь является первоначальное распределение, мат-

рица перехода или функция, управляющая марковским процессом). Краткое изложение этих теорий будет дано в следующих подразделах.

1.5.2. Стационарные случайные процессы и последовательности

1.5.2.1. Понятие о стационарном случайном процессе

Стационарные случайные процессы описывают изменение во времени установившегося, регулярного процесса, происходящего в реальной действительности. При этом все внешние условия, от которых зависит динамика этого процесса, остаются неизменными. Рассмотрим спонтанную импульсную активность нейрона крышки среднего мозга лягушки, когда животное находится в темноте. Так как этот отдел головного мозга осуществляет высший зрительный анализ у лягушки, при отсутствии зрительных раздражителей естественно считать внешние условия, вызывающие импульсную активность нейрона зрительного анализатора, неизменными в течение эксперимента. В связи с этим можно надеяться на то, что эта импульсная активность окажется стационарным процессом.

Перейдем теперь к математическому определению понятия стационарности.

Если мы говорим, что стационарная случайная функция описывает изменение во времени некоторого показателя $x(t)$ вероятностной системы, которое происходит при неизменных внешних условиях, то следует считать неизменными с течением времени все вероятностные свойства случайной функции $x(t)$. Поскольку все вероятностные свойства $x(t)$ определяются совместными функциями распределения $F_{t_1, t_2, \dots, t_n}(u_1, u_2, \dots, u_n)$, то мы должны потребовать, чтобы функции $F_{t_1, \dots, t_n}(u_1, \dots, u_n)$ не изменялись со временем t

$$F_{t_1, \dots, t_n}(u_1, \dots, u_n) = F_{t_1+t, \dots, t_n+t}(u_1, \dots, u_n). \quad (1.28)$$

Равенство (1.28) является математическим определением стационарной функции; при этом требуется, чтобы это равенство было справедливо для любого конечного набора моментов времени $t_1, t_2, \dots, t_n$ и при любом t из области определения $x(t)$. При внимательном анализе формулы (1.28) становится заметной следующая особенность области определения D стационарной случайной функции: наряду с точкой t_1 , в область определения входит и точка $t_1 + t$. Таким образом, сумма двух элементов из области определения D принадлежит D . Множество D , состоящее из элементов t , в котором введена операция сложения элементов (т.е. для любых двух элементов t_1, t_2 существует их сумма $t_1 + t_2$, входящая в данное множество D), обладающая свойством ассоциативности

$$t_1 + (t_2 + t_3) = (t_1 + t_2) + t_3,$$

называется полугруппой; из изложенного выше вытекает, что область определения D стационарной случайной функции $x(t)$ является полугруппой.

Примерами полугруппы на числовой прямой является множество натуральных чисел (при этом случайная функция превращается в случайную последовательность), луч $(0, \infty)$ или вся числовая прямая (случайная функция представляет тогда случайный процесс). В приложениях приходится иметь дело со случайными процессами, наблюдаемыми за некоторый конечный интервал времени $[0, T]$; хотя отрезок $[0, T]$ не является полугруппой, понятию стационарного процесса $x(t)$, наблюдаемому лишь при $t \in [0, T]$, можно придать определенный смысл. Будем говорить, что процесс $x(t)$ ($t \in [0, T]$) стационарен в узком смысле, если равенство (1.28) выполнено при всех $t_1, t_2, \dots, t_n; t_1 + t, t_2 + t, \dots, t_n + t$, входящих в отрезок $[0, T]$.

Непосредственно из определения стационарной случайной функции следует, что одномерная функция распределения $F_t(u) = P(x(t) < u)$ не зависит от t :

$$F_t(u) = F_{0+}(u) = F_0(u). \quad (1.29)$$

Двумерная функция распределения $F_{t_1, t_2}(u_1, u_2)$, описывающая зависимость между $x(t_1)$ и $x(t_2)$, полностью определяется значением $t_2 - t_1$. В самом деле, в силу равенства (1.27)

$$F_{t_1+t, t_2+t}(u_1, u_2) = F_{t_1, t_2}(u_1, u_2),$$

поэтому при сдвиге пары точек t_1, t_2 без изменения их расстояния (т.е. при переходе от t_1, t_2 к $t_1 + t, t_2 + t$) функция $F_{t_1, t_2}(u_1, u_2)$ не изменяется, а это означает, что вид функции $F_{t_1, t_2}(u_1, u_2)$ зависит лишь от расстояния между t_1 и t_2 (т.е. от разности $t_2 - t_1$), а не от их расположения на числовой оси. Итак $F_{t_1, t_2}(u_1, u_2)$ зависит от одного параметра $\tau = t_2 - t_1$:

$$F_{t_1, t_2}(u_1, u_2) = F_{t_2 - t_1}(u_1, u_2) = F_\tau(u_1, u_2). \quad (1.30)$$

Полученные соотношения (1.29) и (1.30) играют исключительно важную роль для определения вероятностных характеристик стационарной случайной функции. Из равенства (1.29) вытекает, что математическое ожидание $m(t) = m[x(t)]$ является постоянной величиной:

$$m(t) = m[x(t)] = \int_{-\infty}^{\infty} u f_t(u) du = \int_{-\infty}^{\infty} u f_0(u) du = m = \text{const},$$

где $f_t(u) = F_t'(u)$.

Коэффициент ковариации $x(t_1)$ и $x(t_2)$:

$$K(t_1, t_2) = K[x(t_1), x(t_2)] = m\{[m_1 - x(t_1)][m_2 - x(t_2)]\}, \quad m_i = m[x(t_i)], \quad i = \overline{1, 2},$$

который носит название корреляционной функции, в общем случае является функцией двух аргументов t_1 и t_2 . Для стационарных случайных функций $k(t_1, t_2)$ будет функцией одного аргумента $\tau = t_2 - t_1$; действительно, пусть $f_{11}, f_{22}(u_1, u_2)$ — совместная плотность вероятности $x(t_1)$ и $x(t_2)$; тогда в силу (1.30)

$$\begin{aligned} f_{i_1, i_2}(u_1, u_2) &= \frac{\partial^2 F_{i_1, i_2}(u_1, u_2)}{\partial u_1 \partial u_2} = \frac{\partial^2 F_{i_2 - i_1}(u_1, u_2)}{\partial u_1 \partial u_2} = \\ &= f_{i_2 - i_1}(u_1, u_2) = f_{\tau}(u_1, u_2), \end{aligned}$$

поэтому

$$K(t_1, t_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (u_1 - m)(u_2 - m) f_{i_1, i_2}(u_1, u_2) du_1 du_2 = K(t_2 - t_1) = K(\tau).$$

Таким образом для стационарной случайной величины функции математическое ожидание и корреляционная функция, будучи константой и функцией одной переменной, представляют очень удобные и компактные показатели. В связи с этим основными вероятностными характеристиками стационарных случайных функций считают их математическое ожидание m и корреляционную функцию $K(\tau)$, а теория, основанная на этих характеристиках, называется корреляционной теорией.

Случайный процесс $x(t)$ называется стационарным в широком смысле, если его математическое ожидание постоянно, а корреляционная функция $K(t_1, t_2)$ зависит лишь от разности аргументов:

$$K(t_1, t_2) = K(t_2 - t_1) = K(\tau), \quad \tau = t_2 - t_1.$$

Ясно, что стационарный в узком смысле процесс будет стационарен и в широком смысле; обратное утверждение, вообще говоря, неверно: можно привести пример стационарного в широком смысле процесса, который не является стационарным в узком смысле. Однако такие процессы будут очень экзотическими и они едва ли встречаются в приложениях. Поэтому мы всегда в дальнейшем будем говорить о стационарных процессах вообще, не уточняя, в каком смысле, и тем самым будем предполагать совпадение этих двух классов стационарных процессов в реальной действительности.

Рассмотрим вопрос об информативности этих показателей m и $K(\tau)$. Совершенно очевидно, что в общем случае математическое ожидание и корреляционная функция описывают лишь самые простые свойства случайной функции, поэтому m и $K(\tau)$ содержит лишь небольшой объем информации обо всей случайной функции $x(t)$. Тем не менее для некоторых классов случайных функций характеристики m и $K(\tau)$ могут быть очень информативны. Здесь в первую очередь следует отметить так называемые “нормальные” или “гауссовские” случайные функции; так называемые “случайные” функции, все конечномерные распределения которых оказываются многомерными нормальными распределениями:

$$f(u_1, u_2, \dots, u_n) = \sqrt{\frac{|K|}{\pi^n}} e^{-\sum_{i,j=1}^n K_{ij}^{-1}(u_i - m_i)(u_j - m_j)},$$

где $K = \|K_{ij}\|$ — корреляционная матрица

$$K_{ij} = m\{[x(t_i) - m_i][x(t_j) - m_j]\} = K(t_i, t_j) = K(t_j - t_i)$$

$K^{-1} = \|K_y^*\|$ — матрица, обратная к K , $|K|$ — определитель этой матрицы и $m_i = m[x(t_i)]$, $(i = \overline{1, 2, \dots, n})$. Очевидно, для таких нормальных случайных функций $m[x(t)]$ и $K(\tau)$ полностью определяют все конечномерные распределения, поэтому эти показатели содержат всю информацию о нормальной случайной функции. Поскольку гауссовские процессы получаются при взаимодействии (суперпозиции) большого количества независимых или слабо зависимых случайных функций, они могут встречаться в приложениях. Значит, для широкого класса случайных процессов (гауссовских или близких к гауссовским) корреляционная теория базируется на информационных показателях.

Корреляционная теория, как мы увидим в дальнейшем, может решить большой класс задач прикладной теории случайных процессов и в случае произвольных стационарных процессов.

Замечание. Если процесс $x(t)$ имеет нулевое математическое ожидание $m[x(t)] = 0$ при всех t , то

$$K(\tau) = m\{[x(t + \tau) - m][x(t) - m]\} = m[x(t + \tau)x(t)].$$

1.5.2.2. Эргодическая теорема

При исследовании случайных функций одной из самых первоочередных задач является определение их математических ожиданий и корреляционных функций. Эти показатели бывают известны до опыта в исключительных случаях. Так уж получилось, что никто не позаботился сообщить нам их заранее, поэтому остается только по имеющимся реализациям строить приближенные значения (оценки) указанных характеристик. К счастью, для стационарных случайных функций при выполнении довольно общих условий математическое ожидание m и корреляционную функцию приближенно можно определить по одной реализации, причем точность этих оценок можно сделать сколь угодно высокой, если увеличить продолжительность наблюдения реализации. Это утверждение носит название эргодической теоремы. Рассмотрим ее точную математическую формулировку. Будем предполагать, что корреляционная функция $K(\tau)$ удовлетворяет следующему соотношению:

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T K(\tau) d\tau = 0 \quad (1.31)$$

или

$$\lim_{N \rightarrow \infty} \frac{1}{N+1} \sum_{\tau=0}^N K(\tau) = 0. \quad (1.32)$$

Выполнение условия (1.31), (1.32) гарантирует справедливость формул

$$m = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x(t) dt, \quad m = \lim_{N \rightarrow \infty} \frac{1}{N+1} \sum_{\tau=0}^N x(\tau), \quad (1.33)$$

а при выполнении условий, аналогичных (1.31) и (1.32), наложенных на моменты более высокого порядка $\lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T}^T K(\tau) \left[1 - \frac{|\tau|}{\theta} \right] d\tau = 0$, $\lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T}^T K^2(\tau) d\tau = 0$, имеют место формулы

$$K(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T [x(t) - \bar{m}] [x(t+\tau) - \bar{m}] dt, \quad (1.34)$$

$$K(\tau) = \lim_{N \rightarrow \infty} \sum_{t=0}^N [x(t) - \bar{m}] [x(t+\tau) - \bar{m}];$$

если же условия (1.32), (1.33) не выполняются, то формулы (1.34) становятся неправильными.

Из формул (1.31) и (1.34) вытекает, что при больших значениях T в качестве хороших оценок для $\bar{m}$ и $K(\tau)$ можно принять

$$\bar{m} = \frac{1}{T} \int_0^T x^{(1)}(t) dt, \quad \bar{m} = \frac{1}{N+1} \sum_{\tau=0}^N x^{(1)}(\tau);$$

$$\bar{K}[\tau] = \frac{1}{T} \int_0^T [x^{(1)}(t) - \bar{m}] [x^{(1)}(t+\tau) - \bar{m}];$$

$$\bar{K}(\tau) = \frac{1}{N+1} \sum_{\tau=0}^N [x^{(1)}(\tau) - \bar{m}] [x^{(1)}(t+\tau) - \bar{m}].$$

При малых значениях T (т.е. когда время наблюдения процесса невелико) эти оценки для $\bar{m}$ уже не являются “самыми лучшими”; в следующей главе мы построим наилучшие оценки математического ожидания $x(t)$ и увидим, что они могут значительно отличаться от $\bar{m}$; кроме того, мы обсудим вопрос о том, какое время наблюдения T можно считать “большим”.

Условие (1.31) или (1.32), необходимое и достаточное для справедливости соотношений (1.33) и (1.34), практически всегда выполняется, так что в приложениях можно даже не заботиться о его соблюдении. Действительно, как правило, корреляционная функция $K(\tau)$ стремится к нулю при $\tau \rightarrow \infty$, ибо с течением времени τ зависимость между $x(t)$ и $x(t+\tau)$ бесконечно ослабевает. Выберем число T_0 настолько большим, чтобы при $\tau > T_0$

$$|K(\tau)| < \varepsilon,$$

где ε — малое, наперед заданное положительное число. Тогда

$$\left| \frac{1}{T} \int_0^T K(\tau) d\tau \right| = \frac{1}{T} \left| \int_0^{T_0} K(\tau) d\tau + \int_{T_0}^T K(\tau) d\tau \right| \leq \frac{1}{T} \left[\int_0^{T_0} K(\tau) d\tau + \int_{T_0}^T |K(\tau)| d\tau \right].$$

В следующем подразделе мы выясним, что $|K(\tau)| \leq K(0)$, поэтому

$$\frac{1}{T} \int_0^{T_0} |K(\tau)| d\tau \leq K(0) \frac{T_0}{T},$$

следовательно,

$$\frac{1}{T} \left| \int_0^T K(\tau) d\tau \right| \leq K(0) \frac{T}{T_0} + \varepsilon \frac{T - T_0}{T} \rightarrow \varepsilon \text{ при } \varepsilon \rightarrow \infty.$$

Таким образом, при больших T выражение

$$\frac{1}{T} \int_0^T K(\tau) d\tau$$

может быть сделано (и останется в дальнейшем) меньше ε , а это и означает, что

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T K(\tau) d\tau = 0.$$

Итак, для класса корреляционных функций $K(\tau) \rightarrow 0$ при $\tau \rightarrow 0$, который представляет наиболее часто встречающийся тип корреляционных функций, условие (1.31) всегда выполняется. Отсюда не следует, конечно, что свойство (1.31) присуще любой корреляционной функции; в следующем подразделе мы узнаем, что функция $K^*(\tau)$, являющаяся периодическим продолжением функции $\tilde{K}(\tau)$

$$\tilde{K}(\tau) = \begin{cases} 1 - \tau, & \text{если } 0 \leq \tau \leq 1; \\ 1 + \tau, & \text{если } -1 \leq \tau \leq 0, \end{cases}$$

будет корреляционной.

Однако при $T = 2N$

$$\frac{1}{T} \int_0^T K^*(\tau) d\tau = \frac{1}{2},$$

так что

$$\frac{1}{T} \int_0^T K^*(\tau) d\tau \not\rightarrow 0 \text{ при } T \rightarrow \infty.$$

1.5.3. Корреляционные функции и их свойства

Непосредственно из определения корреляционной функции

$$K(\tau) = m\{[x(t) - m][x(t + \tau) - m]\} \quad (1.35)$$

вытекает, что она обладает следующими свойствами.

1. $K(0) > 0$; в самом деле, обозначим через $\bar{x}(t) = x(t) - m$, тогда

$$K(0) = m[\bar{x}(t)\bar{x}(t)] - m[\bar{x}(t)]^2 > 0;$$

2. $K(\tau)$ четна, т.е. $K(\tau) = K(-\tau)$. Действительно, положим $t_1 = t - \tau$; так как равенство (1.35) справедливо при любых значениях t , при $t = t_1 + \tau$

$$K(-\tau) = m[\bar{x}(t)\bar{x}(t - \tau)] = m[\bar{x}(t + \tau)\bar{x}(t_1)] = K(\tau);$$

3. При любых τ справедливо равенство

$$|K(\tau)| \leq K(0).$$

Для доказательства этого соотношения рассмотрим выражение

$$\begin{aligned} m[\bar{x}(t+\tau) \pm \bar{x}(t)]^2 &= m[\bar{x}^2(t+\tau) \pm 2\bar{x}(t+\tau)\bar{x}(t) + \bar{x}^2(t)] = \\ &= m[\bar{x}(t+\tau)]^2 \pm 2m[\bar{x}(t)\bar{x}(t+\tau)] + m[\bar{x}(t)]^2 = \\ &= K(0) \pm 2K(\tau) + K(0) = 2[K(0) \pm K(\tau)]. \end{aligned}$$

Поскольку $m[\bar{x}(t+\tau) \pm \bar{x}(t)]^2 \geq 0$, $K(0) \pm K(\tau) \geq 0$, следовательно, $K(0) \geq \pm K(\tau)$, а это означает, что $K(0) \geq |K(\tau)|$.

Для того чтобы сформулировать четвертое свойство корреляционной функции, необходимо ввести одно из основных понятий математического анализа — понятие положительно определенной функции. Функция $K(\tau)$ называется положительно определенной, если для любых вещественных чисел $a_1, a_2, \dots, a_n$ и любых моментов времени $\tau_1, \tau_2, \dots, \tau_n$

$$\sum_{j=1}^n \sum_{i=1}^n K(\tau_i - \tau_j) a_i a_j \geq 0.$$

Нетрудно заметить, что корреляционная функция будет положительно определенной:

$$\begin{aligned} \sum_{i,j=1}^n K(\tau_i - \tau_j) a_i a_j &= \sum_{i,j=1}^n m[\bar{x}(t) \bar{x}(t + \tau_i - \tau_j)] a_i a_j = \sum_{i,j=1}^n m[\bar{x}(\tau_i) \bar{x}(\tau_j)] a_i a_j = \\ &= m \left\{ \sum_{i,j=1}^n \bar{x}(\tau_i) \bar{x}(\tau_j) a_i a_j \right\} = m \left\{ \sum_{i=1}^n \bar{x}(\tau_i) a_i \right\}^2 \geq 0 \end{aligned}$$

(переход от второго выражения к третьему осуществляется, если мы полагаем, что $t = \tau_j$).

Русским математиком А.Я. Хинчиным было доказано обратное допущение, всякая положительно определенная функция является корреляционной функцией некоторого стационарного случайного процесса (при этом можно считать процесс гауссовским) [98].

Таким образом, класс всех корреляционных функций совпадает с классом всех положительно определенных функций.

Описание класса P положительно определенных функций было дано американским математиком С. Бохнером. Он показал, что функция $K(\tau)$ будет положительно определенной тогда и только тогда, когда она может быть представлена в виде преобразования Фурье–Стилтьеса от некоторой неубывающей функции $F(\lambda)$ [83]:

$$K(\tau) = \int_{-\infty}^{\infty} e^{i\lambda\tau} dF(\lambda), \quad (i^2 = -1).$$

У нас нет сейчас возможности популярно объяснить точный смысл этого утверждения (это сделано во многих учебниках по теории случайных процессов ([80, 83]), поэтому мы постараемся дать представление о некоторой части класса P положительно определенных функций, которая незначительно отличается от всего класса P .

Напомним, что преобразованием Фурье называется отображение, которое каждой функции $f(\lambda)$ ставит в соответствие функцию

$$K(\tau) = \int_{-\infty}^{\infty} e^{i\lambda\tau} f(\lambda) d\lambda,$$

где $e^{i\lambda\tau} = \cos \lambda\tau + i \sin \lambda\tau$.

Множество функций $K(\tau)$, являющееся преобразованиями Фурье неотрицательных функций $f(\lambda)$, образует подмножество класса положительно определенных функций, почти полностью совпадающих с ним. Этим самым дана исчерпывающая характеристика класса положительно определенных функций (а поэтому и корреляционных функций): положительно определенные функции в своем подавляющем большинстве представляют образы $f(\lambda) \geq 0$ при действии на $f(\lambda)$ преобразование Фурье.

Изложенное выше описание класса корреляционных функций дает возможность принципиально ответить на вопрос: будет ли данная функция $K(\tau)$ корреляционной или нет? Для этого необходимо рассмотреть обратное преобразование Фурье:

$$f(\lambda) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\lambda\tau} K(\tau) d\tau;$$

если окажется, что $f(\lambda) \geq 0$, то $K(\tau)$ — корреляционная функция; в противном случае $K(\tau)$ не будет корреляционной функцией (при изложении этого алгоритма мы сознательно обошли многие тонкие теоретические проблемы, которые очень интересны с точки зрения чистой математики, но едва ли существенны для естествоиспытателя; разумеется, при таком популярном описании неизбежны небольшие математические неточности).

Нахождение обратного преобразования Фурье от функции $K(\tau)$ далеко не всегда представляет легкую задачу; поэтому для многих функций $K(\tau)$, заданных с помощью аналитического выражения, составлены таблицы обратных преобразований Фурье [63]. Иногда в приложениях и теоретических исследованиях появляются такие случаи, когда нельзя получить явное аналитическое выражение для обратного преобразования Фурье $f(\lambda)$ от корреляционной функции $K(\tau)$; при этом, как правило, очень трудно установить, будет ли $f(\lambda)$ при всех λ неотрицательным числом. Поэтому представляют интерес достаточные условия положительной функции $K(\tau)$, основанные на простых геометрических свойствах кривой $y = K(\tau)$ и не использующие понятие обратного преобразования Фурье. Одно из таких условий было найдено Пойа [122]: оказывается, что всякая четная функция $K(\tau)$, которая выпукла на луче $(0, \infty)$ и не возрастает, является

корреляционной функцией; в дальнейшем такие корреляционные функции мы будем ради краткости называть выпуклыми.

Другое достаточное условие положительной определенности установлено Феллером [122]: если корреляционная функция $K(\tau)$ равна нулю при $|\tau| > a$, то все ее периодические продолжения с периодом $\lambda \geq a$ снова будут корреляционной функцией.

Вернемся теперь к вопросу о том, почему функция $\bar{K}(\tau)$, описанная в предыдущем подразделе, является корреляционной функцией? Функция

$$K_1(\tau) = \begin{cases} \bar{K}(\tau), & \text{если } -1 \leq \tau \leq 1; \\ 0, & \text{если } |\tau| > 1 \end{cases}$$

будет корреляционной в силу условия Поля, а $K^*(\tau)$ представляет периодическое продолжение $K(\tau)$ с периодом $\lambda = 2$, поэтому на основании теоремы Феллера $K^*(\tau)$ — корреляционная функция.

В заключение этого подраздела приведем несколько наиболее часто встречающихся корреляционных функций:

- 1) $K(\tau) = ae^{-\alpha|\tau|}$, где $a > 0, \alpha > 0$;
- 2) $K(\tau) = ae^{-\alpha|\tau|}(1 + \alpha|\tau|)$, $a > 0, \alpha > 0$;
- 3) $K(\tau) = \sigma^2 e^{-\alpha|\tau|} \cos \beta \tau$, $\alpha, \beta > 0$;
- 4) $K(\tau) = \sigma^2 e^{-\alpha|\tau|} \left(\cos \beta \tau + \frac{\alpha}{\beta} \sin \beta |\tau| \right)$, $\alpha, \beta > 0$;
- 5) $K(\tau) = \sigma^2 e^{-\alpha|\tau|} \left(1 + \alpha|\tau| + \frac{1}{3} \alpha^2 \tau^2 \right)$, $\alpha > 0$.

Для всех этих корреляционных функций можно определить функцию $f(\lambda)$ и убедиться в ее неотрицательности.

Резюме

Случайным процессом $x(t)$, $t \in [0, T]$ называется случайная величина, изменяющаяся с течением времени (т.е. зависящая от времени). Важнейшими показателями случайного процесса есть математическое ожидание $m(t) = m[x(t)]$ и корреляционная функция $K(t, s) = m\{[x(t) - m(t)][x(s) - m(s)]\}$. Случайный процесс $x(t)$ называется стационарным (в широком смысле), если его математическое ожидание — постоянная величина $m[x(t)] = m(t) = m = \text{const}$, а корреляционная функция зависит от разности аргументов $K(t, s) = K(t - s)$. Корреляционная функция $K(\tau) = K(t - s)$, $\tau = t - s$, стационарного случайного процесса является положительно определенной: при любых a_i , $i = 1, 2, \dots, n$, справедливо неравенство

$$\sum_{j=1}^n \sum_{i=1}^n K(\tau_i - \tau_j) a_i a_j \geq 0.$$

Справедливо и обратное утверждение: всякая положительно определенная функция $K(\tau)$ является корреляционной функцией некоторого стационарного случайного процесса.

Основные понятия математической статистики

2.1. Понятие генеральной совокупности. Случайный выбор элементов генеральной совокупности. Простой случайный выбор

Одним из основных понятий общей биологии и доказательной медицины является понятие популяции, которое в статистике принято называть *генеральной совокупностью*. Под этим термином понимают совокупность живых индивидуумов одного вида, занимающих определенный ареал, т.е. проживающих в данной местности. Можно говорить о популяции сосен в отдельном лесу, о популяции эпителиоцитов в буккальном эпителии человека, о популяции онкологических больных, которая распадается на субпопуляции больных конкретными видами рака. Мы будем называть отдельные элементы генеральной совокупности (популяции) его *членами*, или *индивидуумами*.

Случайное испытание, о котором шла речь в главе 1, заключается в том, что из генеральной совокупности наугад выбирается какой-либо индивидуум, у которого регистрируется значение некоторой определенной характеристики ξ , после чего этот индивидуум возвращается в генеральную совокупность (выбор с возвращением). В некоторых случаях этот выбор, например, может проводить экспериментатор, когда он берет мазок у пациента при диагностике того или иного заболевания. Однако во многих ситуациях случайный выбор члена генеральной совокупности не зависит от исследователя: при диагностике заболевания у пациента клиницисты его сами не выбирают: больной поступает в клинику из-за развития патологического процесса. Кроме того, показатель ξ , характеризующий индивидуума, не всегда является отдельным числом. Например, при дифференциальной цитогенетической компьютерной диагностике рака молочной железы по данным сканограмм интерфейсных ядер буккального эпителия у каждого пациента берется соскоб, состоящий из 5070 клеток внутренней поверхности щеки, причем одна сканограмма состоит из $160 \times 160 = 25\,600$ чисел (в диапазоне от 0 до 255). Таким образом, показатель ξ может представлять собой случайную матрицу.

Исследование генеральной совокупности G путем случайного выбора ее индивидуумов для измерения показателя ξ представляет собой многократное повторение случайного испытания T . Группу индивидуумов, наблюдаемых при n повторениях случайного испытания T , будем называть *случайной выборкой* из генеральной совокупности G . Если при этом случайное испытание T организовано так, что вероятность быть выбранным одинакова для всех членов генеральной совокупности G и результат одного повторения испытания T не влияет на результаты других повторений, то описанный процесс выбора называется простым случайным выбором.

Значение характеристики ξ выбранного индивидуума из генеральной совокупности G является случайной величиной (одномерной, если показатель ξ является отдельным числом, или многомерным, если показатель ξ будет вектором или матрицей). Действительно, после проведения случайного испытания (выбора элемента из генеральной совокупности G) показатель ξ принимает конкретные значения, но до проведения этого испытания мы не можем точно предсказать, какое значение примет показатель ξ . Вектор, состоящий из n наблюдаемых значений $\xi_1, \xi_2, \dots, \xi_n$, полученных при n повторениях случайного испытания T , будем называть случайной выборкой объема n из генеральной совокупности G , а сами значения $\xi_k, k = 1, \dots, n$ — выборочными значениями.

Используя эти понятия можно дать более формальное определение простого случайного выбора: это выбор индивидуумов из генеральной совокупности G , при котором все выборочные значения имеют одинаковую функцию распределения $F_{\xi_k}(u) = p(\xi_k < u) \equiv F(u) \forall k = 1, 2, \dots, n$, которая называется функцией распределения генеральной совокупности G , и являются независимыми в совокупности случайными величинами.

Резюме

Генеральная совокупность — это множество индивидуумов, обладающих одним или несколькими числовыми характеристиками, которые представляют интерес для исследователя в связи с решением биологических или медицинских проблем. Отдельные элементы генеральной совокупности (популяции) называются его членами, или индивидуумами. Группа индивидуумов из генеральной совокупности, наблюдаемые при n повторениях случайного испытания T , состоящего в случайном выборе одного члена генеральной совокупности, называется случайной выборкой из генеральной совокупности G . Если при этом случайное испытание T организовано так, что вероятность быть выбранным одинакова для всех членов генеральной совокупности G и результат одного повторения испытания T не влияет на результат других его повторений, то описанный процесс называется простым случайным выбором.

2.2. Оценка вероятностных характеристик случайных событий и величин. Несмещенные оценки неизвестного математического ожидания

Как было показано в главе 1, случайным событием A принято называть реакцию вероятностной системы на проведение случайного испытания T . При этом предсказать результат такого случайного испытания, т.е. определить реакцию системы, до проведения испытания T невозможно. Однако можно определить вероятность этого события $p(A)$ с произвольной степенью точности, если мы сможем провести большое количество повторений случайного испытания T следующим образом: пусть при n повторениях испытания T событие A произошло k раз, тогда частота $h_k(A)$ события A , равная k/n , стремится к неизвестной вероятности $p(A)$, когда $n \rightarrow \infty$.

Таким образом, частота $h_k(A)$ является оценкой неизвестной вероятности случайного события A . В дальнейшем мы покажем, каким образом можно определить точность этой оценки в зависимости от количества повторения случайного испытания T . Отметим, что в некоторых ситуациях, которые исключительно редко возникают в реальной жизни, вероятность можно определить точно, не прибегая к подсчету частоты события. Однако это возможно лишь тогда, когда событие A распадается на схему случаев, как правило, это возможно лишь в теоретических задачах, предназначенных для учебников по теории вероятностей.

Перейдем к оценке вероятностных характеристик одномерной случайной величины x , среди которых самым важным является его *математическое ожидание*, или *среднее значение* $m(x)$. Этот показатель соответствует точке на прямой, вокруг которой расположены возможные значения случайной величины x , поэтому число $m(x)$ является характеристикой их пространственного распределения. Другим важным показателем, характеризующим свойства случайной величины x , является ее *дисперсия* $\sigma^2(x)$, т.е. мера ее разброса вокруг математического ожидания.

Предположим вначале, что мы имеем выборку $x_1, x_2, \dots, x_n$, полученную путем простого случайного выбора и состоящую из n значений случайной величины x . Отметим, что выборочные значения x_k также являются случайными величинами, причем $m(x_k) = m(x)$, $D(x_k) = \sigma^2(x)$. Это можно объяснить так: если мы будем бесконечно много раз извлекать из генеральной совокупности G случайной величины x выборки $x_1, x_2, \dots, x_n$, то каждый раз на k -м месте будут появляться новые числа, но вместе с тем эти числа будут по-прежнему являться значениями случайной величины x . Поэтому их математическое ожидание и дисперсия совпадают с математическим ожиданием и дисперсией случайной величины x .

Для того чтобы отыскать наилучшую оценку математического ожидания случайной величины x , рассмотрим класс всех несмещенных линейных оценок M математического ожидания $m(x)$. Такой выбор класса объясняется тем, что вычисления в нем наиболее простые. Оценка u математического ожидания $m(x)$ называется *несмещенной*, если ее математическое ожидание совпадает с величиной $m(x)$, т.е. $m(u) = m(x)$, а линейной оценкой называется величина вида $u = \sum_{k=1}^n c_k x_k$, где c_k — фиксированные числа. Если оценка u является одновременно несмещенной и линейной, то

$$m(u) = m\left(\sum_{k=1}^n c_k x_k\right) = \sum_{k=1}^n c_k m(x_k) = m(x) \sum_{k=1}^n c_k.$$

Таким образом, ограничения на коэффициенты c_k имеют очень простой вид:

$$\sum_{k=1}^n c_k = 1. \quad (2.1)$$

Отметим, что наилучшую оценку неизвестного параметра всегда ищут в конкретном классе, так как иначе нельзя указать, в каком смысле эта оценка является оптимальной.

Попробуем найти в классе M несмещенных линейных оценок наилучшую. Точность несмещенной оценки характеризуется ее дисперсией — чем меньше дисперсия, тем точнее оценка. Следовательно, поиск наилучшей оценки сводится к определению линейной несмещенной оценки с минимальной дисперсией. Сначала вычислим дисперсию оценки:

$$D(u) = D\left(\sum_{k=1}^n c_k x_k\right) = \sum_{k=1}^n c_k^2 D(x_k) = \sigma^2 \sum_{k=1}^n c_k^2.$$

Учитывая равенство (2.1), представим сумму $\sum_{k=1}^n c_k^2$ в следующем виде:

$$\sum_{k=1}^{n-1} c_k^2 + c_n^2 = \sum_{k=1}^{n-1} c_k^2 + \left(1 - \sum_{k=1}^{n-1} c_k\right)^2.$$

Тогда дисперсию оценки можно записать так.

$$D(u) = f(c_1, c_2, \dots, c_{n-1}) = \sigma^2 \left(\sum_{k=1}^{n-1} c_k^2 + \left(1 - \sum_{k=1}^{n-1} c_k\right)^2 \right).$$

Следовательно, дисперсия оценки является квадратичным положительным многочленом. Применяя стандартные методы поиска минимума многочлена, можно показать, что минимум функции $f(c_1, c_2, \dots, c_{n-1})$ достигается в точке, где $c_1 = c_2 = \dots = c_n = \frac{1}{n}$. Следовательно, наилучшая несмещенная линейная оценка $m(x)$ имеет вид

$$m(x) \approx \bar{x} = \frac{1}{n} \sum_{k=1}^n x_k.$$

Оценка $\bar{x}$ называется *выборочным средним*, ее дисперсия

$$D(\bar{x}) = D\left(\frac{1}{n} \sum_{k=1}^n x_k\right) = \frac{1}{n^2} \sum_{k=1}^n D(x_k) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

при увеличении объема выборки n быстро стремится к нулю, а выборочное среднее $\bar{x}$ приближается к истинному математическому ожиданию $m(x)$.

Если выборочные значения $x_1, x_2, \dots, x_n$ являются зависимыми, то для поиска наилучшей несмещенной линейной оценки u^* (т.е. оценки из класса M с минимальной дисперсией) необходимо знать коэффициенты ковариации случайных величин x_i и x_j , где $i, j = 1, 2, \dots, n$, которые удобно записать в виде корреляционной матрицы (см. главу 1, подраздел 1.4.2):

$$K = \begin{bmatrix} K(x_1, x_1) & K(x_1, x_2) & \dots & K(x_1, x_n) \\ K(x_2, x_1) & K(x_2, x_2) & \dots & K(x_2, x_n) \\ \dots & \dots & \dots & \dots \\ K(x_n, x_1) & K(x_n, x_2) & \dots & K(x_n, x_n) \end{bmatrix}$$

Здесь $K(x_i, x_j) = m[(x_i - m)(x_j - m)]$, $m = m(x)$, $i = 1, 2, \dots, n$. Наилучшая линейная оценка u^* математического ожидания случайной величины x по зависимым выборочным значениям имеет вид

$$u^* = \sum_{k=1}^n \dot{c}_k x_k,$$

где

$$(\dot{c}_1, \dot{c}_2, \dots, \dot{c}_n) = \frac{K^{-1} \epsilon^T}{\epsilon K^{-1} \epsilon^T}, \epsilon = (1, 1, \dots, 1).$$

Здесь знак T означает транспонирование вектора (это означает, что из строки получается столбец, а из столбца — строка), а K^{-1} — матрица, обратная к матрице K . Более того, можно показать, что

$$D(u^*) = (\epsilon K^{-1} \epsilon^T)^{-1}.$$

Если выборочные значения x_k являются независимыми, но имеют разную дисперсию $\sigma_k^2 = D(x_k)$, то наилучшую оценку u^* из класса M можно записать как

$$u^* = \sum_{k=1}^n \dot{c}_k x_k,$$

$$\text{где } \dot{c}_k = \frac{g_k}{\sum_{k=1}^n g_k}, \quad g_k = \frac{\sigma^2}{\sigma_k^2}, \quad \sigma^2 = \sum_{k=1}^n \sigma_k^2, \quad D(u^*) = \frac{\sigma^2}{\sum_{k=1}^n g_k}.$$

Оказывается, что для выборочных значений $x_1, x_2, \dots, x_n$, полученных в результате простого случайного выбора, выборочное значение $\bar{x}$ является оптимальной оценкой: все показатели ее разброса вокруг $m(x)$ минимальны. Говоря более формально, любой абсолютный центральный момент p -го порядка $\alpha_p = m(|\bar{x} - m(x)|^p)$, $p \geq 1$, величины $\bar{x}$ является минимальным.

Резюме

По определению вероятность $p(A)$ случайного события A является пределом его частоты: $p(A) = \lim_{n \rightarrow \infty} h_n(A)$. Следовательно, при достаточно большом количестве испытаний n (больше 30) частоту можно считать почти точной оценкой вероятности $p(A)$.

Если выборка $x_1, x_2, \dots, x_n$ получена путем простого случайного выбора из генеральной совокупности G (т.е. все выборочные значения имеют одинаковые математическое ожидание и дисперсию), то наилучшей линейной оценкой неизвестного математического ожидания является выборочное среднее. Кроме того, такая оценка является оптимальной, т.е. все показатели ее разброса вокруг математического ожидания являются минимальными.

2.3. Оценка неизвестной дисперсии случайной величины. Квадратичные оценки дисперсии

Как показано в подразделе 1.4.1, дисперсия $D(x)$ случайной величины x определяется формулой

$$D(x) = m[x - m(x)]^2$$

и характеризует меру разброса возможных значений x вокруг ее математического ожидания $m(x)$. Очевидно, что

$$\begin{aligned} D(x) &= m[x - m(x)]^2 = m[x^2 - 2xm(x) + m^2(x)] = \\ &= m(x^2) - 2[m(x)]^2 + [m(x)]^2 = m(x^2) - [m(x)]^2. \end{aligned}$$

Отсюда следует, что для оценки дисперсии $D(x)$ необходимо оценить момент второго порядка $m(x^2)$ случайной величины x^2 и ее математическое ожидание $m(x)$ на основании выборочных значений $x_1, x_2, \dots, x_n$. В дальнейшем будем считать, что эта выборка получена путем простого случайного выбора. В таком случае наилучшей оценкой $m(x)$ является выборочное среднее $\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k$. Со-

гласно практическим рекомендациям, выборочное среднее $\bar{x}$, как правило, практически совпадает с $m(x)$, если объем выборки превышает 30. В связи с этим основное внимание следует сосредоточить на оценке момента второго порядка $m(x^2)$. Кроме того, из сказанного следует, что оценку дисперсии $D(x)$ можно найти как при известном, так и при неизвестном $m(x)$.

Рассмотрим вариант, когда математическое ожидание известно. Введем величину $z = x^2$. Тогда $z_k = x_k^2, k = 1, 2, \dots, n$ и $m(x^2) = m(z)$. Наилучшей оценкой $m(z)$ в классе несмещенных линейных оценок является выборочное среднее $\bar{z}$:

$$\bar{z} = \frac{1}{n} \sum_{k=1}^n z_k = \frac{1}{n} \sum_{k=1}^n x_k^2.$$

Однако в более широком классе несмещенных оценок — в классе квадратичных оценок — можно получить более точные результаты. Обозначим через B класс несмещенных квадратичных оценок момента второго порядка $m(x^2)$:

$$B = \left\{ u = \sum_{i,k=1}^n a_{ik} x_i x_k, m(u) = m(x^2) \right\}.$$

Без ограничения общности можно считать, что матрица $A = \{a_{ik}\}_{i,k=1,2,\dots,n}$ квадратичной формы

$$u = \sum_{i,k=1}^n a_{ik} x_i x_k = (Ax, x), x = (x_1, x_2, \dots, x_n)$$

является симметричной, т.е. $a_{ik} = a_{ki}$. Поскольку $m(x^2) > 0$, естественно считать, что $(Ax, x) \geq 0$ для всех $x = (x_1, x_2, \dots, x_n)$. Найдем условия, которым должны удовлетворять коэффициенты a_{ik} , чтобы оценки u из класса B были несмещенными. Для этого сначала выделим в оценке “диагональную” часть (где $i = k$), а затем вычислим ее математическое ожидание:

$$\begin{aligned} m(u) &= m\left(\sum_{i,k=1}^n a_{ik} x_i x_k\right) = m\left(\sum_{\substack{i,k=1 \\ i \neq k}}^n a_{ik} x_i x_k + \sum_{i=1}^n a_{ii} x_i^2\right) = \\ &= \sum_{i,k=1}^n a_{ik} [m(x)]^2 + \sum_{i=1}^n a_{ii} m(x_i^2) = m(x^2). \end{aligned}$$

Поскольку это равенство справедливо для всех положительных значений $m(x^2)$, коэффициенты a_{ik} должны удовлетворять условиям

$$\sum_{\substack{i,k=1 \\ i \neq k}}^n a_{ik} = 0, \sum_{i=1}^n a_{ii} = 1.$$

Таким образом, класс всех несмещенных квадратичных оценок момента второго порядка имеет следующий вид:

$$B = \left\{ u = \sum_{i,k=1}^n a_{ik} x_i x_k = (Ax, x), \sum_{i,k=1}^n a_{ik} = 0, \sum_{i=1}^n a_{ii} = 1, (Ax, x) \geq 0 \text{ для всех } x = (x_1, \dots, x_n) \right\}.$$

Наилучшей несмещенной квадратичной оценкой момента второго порядка является величина

$$\bar{z} = \frac{1}{n} \sum_{k=1}^n x_k^2.$$

Теперь, зная наилучшую квадратичную оценку момента второго порядка, можно перейти к исследованию квадратичных оценок дисперсии на основании выборочных значений при условии, что математическое ожидание $m(x) = m$ случайной величины x известно. В этом случае

$$D(x) = m(x^2) - [m(x)]^2 = m(x^2) - m^2.$$

Класс B_{σ^2} несмещенных квадратичных оценок u для $D(u) = \sigma^2$ имеет следующий вид:

$$B_{\sigma^2} = \left\{ u = \sum_{i,k=1}^n a_{ik} x_i x_k - m^2, \sum_{i,k=1}^n a_{ik} = 0, \sum_{i=1}^n a_{ii} = 1, (Ax, x) \geq m^2 \text{ для всех } x = (x_1, \dots, x_n) \right\}.$$

Оценка u^* является наилучшей, если $u^* = \bar{z} - m$, т.е.

$$u^* = s_0^2 = \frac{1}{n} \sum_{k=1}^n x_k^2 - m^2. \quad (2.2)$$

К сожалению, в класс B_{σ^2} не попадают классические оценки неизвестной дисперсии в случае известного математического ожидания

$$s^2 = \frac{1}{n} \sum_{k=1}^n (x_k - m)^2, \quad (2.3)$$

так как в выражение s^2 входят линейные члены относительно выборочных значений x_k . В связи с этим необходимо рассмотреть еще более широкий класс $B_{\sigma^2}^*$ квадратичных несмещенных оценок

$$B_{\sigma^2}^* = \left\{ u = \sum_{i,k=1}^n a_{ik} x_i x_k + 2 \sum_{i=1}^n b_i x_i + c, m(u) = \sigma^2 \right\},$$

в который входят значения квадратичного многочлена относительно переменных $x_1, \dots, x_n$, так что $s^2 \in B_{\sigma^2}^*$. Таким образом, существуют две несмещенные оценки дисперсии — s_0 и s_0^2 . Можно показать, что для нормального распределения выполняется неравенство:

$$D(s_0^2) < D(s^2).$$

Следовательно, оценка s_0^2 является более точной, чем s^2 . Однако, если случайная величина x имеет равномерное или экспоненциальное распределение, то справедливо противоположное неравенство:

$$D(s_0^2) > D(s^2).$$

Итак, выбор наилучшей квадратичной несмещенной оценки дисперсии зависит от распределения случайной величины x .

Рассмотрим теперь случай неизвестного математического ожидания m . Если в формулах (2.2) и (2.3) заменить величину m ее оценкой $\bar{x}$, то полученные оценки совпадут:

$$\begin{aligned}\tilde{s}^2 &= \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2 = \frac{1}{n} \sum_{k=1}^n (x_k^2 - 2x_k \bar{x} + \bar{x}^2) = \\ &= \frac{1}{n} \sum_{k=1}^n x_k^2 - \frac{2}{n} \sum_{k=1}^n x_k \bar{x} + \bar{x}^2 = \frac{1}{n} \sum_{k=1}^n x_k^2 - 2\bar{x}^2 + \bar{x}^2 = \frac{1}{n} \sum_{k=1}^n x_k^2 - \bar{x}^2 = \tilde{s}_0^2.\end{aligned}$$

Однако оценка $\tilde{s}_0^2$ является смещенной, так как

$$m(\tilde{s}_0^2) = \frac{n-1}{n} \sigma^2.$$

Умножив оценку $\tilde{s}_0^2$ на поправочный коэффициент, можно получить несмещенную оценку:

$$\hat{s}^2 = \frac{n}{n-1} \tilde{s}_0^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2,$$

Для исследования точности оценок $\hat{s}^2$ и $\tilde{s}_0^2$ необходимо оценить средние квадратичные ошибки этих оценок:

$$\delta^2(\hat{s}) = m(\hat{s}^2 - \sigma^2)^2$$

и

$$\delta^2(\tilde{s}_0^2) = m(\tilde{s}_0^2 - \sigma^2)^2 = D(\tilde{s}_0^2).$$

Средняя квадратичная ошибка $\delta(\hat{\alpha})$ оценки $\hat{\alpha}$ параметра α определяется формулой

$$\delta(\hat{\alpha}) = m(\hat{\alpha} - \alpha)^2.$$

Для несмещенной оценки $\hat{\alpha}$ величина $\delta(\hat{\alpha})$ превращается в дисперсию $D(\hat{\alpha})$.

Можно показать, что в случае нормального распределения смещенная оценка $\tilde{s}_0^2$ более точная, чем несмещенная оценка $\hat{s}^2$, а для экспоненциального распределения справедливо противоположное утверждение.

Перейдем к исследованию класса C квадратичных несмещенных оценок дисперсии при неизвестном математическом ожидании. Класс C состоит из квадратичных оценок $u = \sum_{i,j=1}^n a_{ij} x_i x_j$, которые являются несмещенными, т.е. $m(u) = \sigma^2$, а матрица коэффициентов $A = \{a_{ik}\}_{i,k=1,2,\dots,n}$ положительно определена. Это требование является вполне естественным, поскольку дисперсия всегда является положительной.

Из условия несмещенности следует, что

$$\begin{aligned}m(u) &= m\left(\sum_{i,k=1}^n a_{ik} x_i x_k\right) = m\left(\sum_{\substack{i,k=1 \\ i \neq k}}^n a_{ik} x_i x_k + \sum_{i=1}^n a_{ii} x_i^2\right) = \\ &= m(x^2) \sum_{\substack{i,k=1 \\ i \neq k}}^n a_{ik} + m \sum_{i=1}^n a_{ii} = m(x^2) - m^2 = \sigma^2.\end{aligned}$$

Поскольку это равенство должно выполняться при всех положительных $m(x^2)$ и m^2 ,

$$\sum_{\substack{i,k=1 \\ i \neq k}}^n a_{ik} = -1, \sum_{i=1}^n a_{ii} = 1.$$

Таким образом,

$$C = \left\{ u = \sum_{i,k=1}^n a_{ik} x_i x_k - m^2, \sum_{\substack{i,k=1 \\ i \neq k}}^n a_{ik} = -1, \sum_{i=1}^n a_{ii} = 1, (Ax, x) \geq 0 \text{ для } x = (x_1, \dots, x_n) \right\}.$$

Легко видеть, что классическая оценка дисперсии

$$\bar{s}^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2 \quad (2.4)$$

принадлежит классу C . При этом матрица A имеет следующий вид:

$$A = \begin{pmatrix} \frac{1}{n} & -\frac{1}{n(n-1)} & -\frac{1}{n(n-1)} & \dots & -\frac{1}{n(n-1)} \\ -\frac{1}{n(n-1)} & \frac{1}{n} & -\frac{1}{n(n-1)} & \dots & -\frac{1}{n(n-1)} \\ \dots & \dots & \dots & \dots & \dots \\ -\frac{1}{n(n-1)} & -\frac{1}{n(n-1)} & -\frac{1}{n(n-1)} & \dots & \frac{1}{n} \end{pmatrix}.$$

Как известно, если выборка $x_1, x_2, \dots, x_n$ получена в результате простого случайного выбора, оценка $\bar{s}^2$ является наилучшей в классе C всех несмещенных квадратичных оценок.

Резюме

Оценку неизвестной дисперсии $\sigma^2(x)$, когда известно математическое ожидание $m = m(x)$, можно получить на основании двух следующих формул:

$$s_0^2 = \frac{1}{n} \sum_{k=1}^n x_k^2 - m^2 \text{ и } s^2 = \frac{1}{n} \sum_{k=1}^n (x_k - m)^2.$$

При этом в случае нормального распределения выборочных значений оценка s_0^2 является более точной, чем s^2 , а для равномерного и экспоненциального распределения выборочных значений оценка s^2 более точная, чем s_0^2 . Если математическое ожидание $m = m(x)$ неизвестно, то наилучшей оценкой для $\sigma^2(x)$ в классе всех несмещенных квадратичных оценок является оценка

$$\hat{s}^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2, \bar{x} = \frac{1}{n} \sum_{k=1}^n x_k.$$

2.4. Оценка коэффициентов ковариации и корреляции

Как указано в разделе 1.4, коэффициенты ковариации $K(x, y) = m[x - m(x)][y - m(y)]$ случайных величин x и y являются одними из самых важных характеристик меры зависимости случайных величин. Коэффициент ковариации между независимыми случайными величинами равен нулю. Обратное утверждение, вообще говоря, неверно, поскольку существуют примеры зависимых случайных величин, у которых коэффициент ковариации равен нулю. Однако для некоторых случайных величин (например, имеющих совместное нормальное распределение) их независимость эквивалентна равенству $K(x, y) = 0$. По этой причине в статистической практике малое значение коэффициента ковариации $K(x, y)$ считается признаком слабой зависимости между x и y .

Пусть x и y — пара случайных величин и $K(x, y) = m[x - m(x)][y - m(y)]$ — коэффициент ковариации. Выполним следующее преобразование:

$$\begin{aligned} K(x, y) &= m[x - m(x)][y - m(y)] = m[xy - xm(y) - m(x)y + m(x)m(y)] = \\ &= m[xy] - m(x)m(y) - m(x)m(y) + m(x)m(y) = m[xy] - m(x)m(y). \end{aligned}$$

Отсюда следует, что для оценки коэффициента ковариации $K(x, y)$ необходимо оценить смешанный момент второго порядка $\gamma = m(xy)$ и математические ожидания $\alpha = m(x)$ и $\beta = m(y)$. Прежде всего заметим, что для оценки смешанного момента γ необходимо иметь совокупность пар выборочных значений $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ из генеральных совокупностей G_x и G_y : $x_k \in G_x$ и $y_k \in G_y$. Будем предполагать, что эти пары выборочных значений получены с помощью простого случайного выбора. Это значит, что все выборочные значения (x_k, y_k) имеют одно и то же двумерное распределение $F(u, v)$ и взаимно не зависят, т.е. для двух множеств M_1 и M_2 выполняется следующее равенство:

$$p((x_i, y_i) \in M_1, (x_j, y_j) \in M_2) = p((x_i, y_i) \in M_1)p((x_j, y_j) \in M_2) \text{ при } i \neq j.$$

Как указано в разделе 2.3, проблему оценки неизвестного коэффициента ковариации можно исследовать как при известном, так и при неизвестном математическом ожидании.

Рассмотрим вначале проблему оценки коэффициента ковариации при известных математических ожиданиях. Начнем с оценки смешанного момента второго порядка $\gamma = m(xy)$. Введем класс билинейных несмещенных оценок $\gamma = m(xy)$:

$$T = \left\{ u = \sum_{i,k=1}^n a_{ik} x_i y_k = (Ax, y), x = (x_1, x_2, \dots, x_n), y = (y_1, y_2, \dots, y_n), m(u) = \gamma \right\}.$$

Из условия несмещенности оценки $m(u) = m(xy)$ следует, что

$$\begin{aligned}
m(u) &= m\left(\sum_{j,k=1}^n a_{jk} x_j y_k\right) = m\left(\sum_{\substack{j,k=1 \\ i \neq k}}^n a_{jk} x_j y_k + \sum_{i=1}^n a_{ii} x_i y_i\right) = \\
&= m(x)m(y) \sum_{j,k=1}^n a_{jk} + m(xy) \sum_{i=1}^n a_{ii} = \alpha\beta \sum_{j,k=1}^n a_{jk} + m(xy) \sum_{i=1}^n a_{ii} = m(xy).
\end{aligned}$$

Поскольку это равенство должно выполняться для произвольных значений $m(xy)$,

$$\sum_{\substack{i,j,k=1 \\ i \neq k}}^n a_{jk} = 0, \sum_{i=1}^n a_{ii} = 1.$$

Следовательно, класс билинейных несмещенных оценок смешанного момента второго порядка имеет следующий вид:

$$T = \left\{ u = \sum_{i,k=1}^n a_{ik} x_i y_k = (Ax, y), x = (x_1, x_2, \dots, x_n), y = (y_1, y_2, \dots, y_n), \sum_{\substack{i,k=1 \\ i \neq k}}^n a_{ik} = 0, \sum_{i=1}^n a_{ii} = 1 \right\}.$$

Как известно, если $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ образуют совокупность пар выборочных значений, полученных с помощью простого случайного выбора, то оценка

$$\bar{y} = \sum_{i=1}^n x_i y_i$$

является наилучшей оценкой (т.е. имеет минимальную дисперсию) в классе T .

Полученные результаты дают возможность найти наилучшие оценки для коэффициента ковариации в случае известных математических ожиданий $m(x)$ и $m(y)$. С этой целью рассмотрим класс оценок коэффициента ковариации, когда нам известны величины $\alpha = m(x)$ и $\beta = m(y)$:

$$E = \left\{ u = \sum_{i,k=1}^n a_{ik} x_i x_k + C\alpha\beta, m(u) = K(x, y) \right\}.$$

Здесь C — некоторая произвольная, но фиксированная константа, а равенство $m(u) = K(x, y)$ является условием несмещенности:

$$\begin{aligned}
m(u) &= m\left(\sum_{i,k=1}^n a_{ik} x_i y_k + C\alpha\beta\right) = m\left(\sum_{\substack{i,k=1 \\ i \neq k}}^n a_{ik} x_i y_k + \sum_{i=1}^n a_{ii} x_i y_i + C\alpha\beta\right) = \\
&= \alpha\beta \sum_{\substack{i,k=1 \\ i \neq k}}^n a_{ik} + m(xy) \sum_{i=1}^n a_{ii} + C\alpha\beta = \\
&= m(xy) \sum_{i=1}^n a_{ii} + \alpha\beta \left(\sum_{\substack{i,k=1 \\ i \neq k}}^n a_{ik} + C \right) = K(x, y) = m(xy) - \alpha\beta.
\end{aligned}$$

Поскольку это равенство должно выполняться при произвольных значениях $m(xy)$,

$$\sum_{\substack{i,k=1 \\ i \neq k}}^n a_{ik} + c = -1, \quad \sum_{i=1}^n a_{ii} = 1.$$

Таким образом, класс E состоит из следующих оценок:

$$E = \left\{ u = \sum_{i,k=1}^n a_{ik} x_i x_k + C\alpha\beta, \sum_{\substack{i,k=1 \\ i \neq k}}^n a_{ik} + c = -1, \sum_{i=1}^n a_{ii} = 1 \right\}.$$

Можно показать, что в классе E несмещенных билинейных оценок коэффициента ковариации наилучшей является оценка

$$\bar{K}(x, y) = \frac{1}{n} \sum_{k=1}^n x_k y_k$$

при условии, что пары (x_k, y_k) имеют нормальное совместное распределение с нулевым математическим ожиданием, а выборка $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ получена в результате простого случайного выбора.

Перейдем к исследованию оценок коэффициента ковариации при условии, что математические ожидания $m(x)$ и $m(y)$ неизвестны. Рассмотрим следующий класс несмещенных оценок коэффициента ковариации $K(x, y)$.

$$E_1 = \left\{ u = \sum_{i,k=1}^n a_{ik} x_i x_k, m(u) = K(x, y) \right\}.$$

Из условия несмещенности $m(u) = K(x, y)$ следует, что

$$\begin{aligned} m(u) &= m \left(\sum_{i,k=1}^n a_{ik} x_i x_k \right) = m \left(\sum_{\substack{i,k=1 \\ i \neq k}}^n a_{ik} x_i y_k + \sum_{i=1}^n a_{ii} x_i y_i \right) = \\ &= \alpha\beta \sum_{\substack{i,k=1 \\ i \neq k}}^n a_{ik} x_i y_k + m(xy) \sum_{i=1}^n a_{ii} x_i y_i + C\alpha\beta = \\ &= m(x)m(y) \sum_{\substack{i,k=1 \\ i \neq k}}^n a_{ik} + m(xy) \sum_{i=1}^n a_{ii} = m(xy) - m(x)m(y). \end{aligned}$$

Поскольку это равенство должно выполняться для произвольных значений $m(xy)$, $m(x)$ и $m(y)$,

$$\sum_{\substack{i,k=1 \\ i \neq k}}^n a_{ik} = -1, \quad \sum_{i=1}^n a_{ii} = 1.$$

Таким образом, класс билинейных несмещенных коэффициентов ковариации имеет следующий вид:

$$E_1 = \left\{ \begin{aligned} u &= \sum_{i,k=1}^n a_{ik} x_i x_k = (Ax, y) \\ \text{для всех } x &= (x_1, x_2, \dots, x_n), y = (y_1, y_2, \dots, y_n), \sum_{i,k=1}^n a_{ik} = -1, \sum_{i=1}^n a_{ii} = 1. \end{aligned} \right\}$$

Оценка

$$K^*(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}),$$

где $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$, является наилучшей в классе E_1 , если выборка $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ получена за счет простого случайного выбора.

Как указано в разделе 1.4, один из существенных недостатков коэффициента ковариации заключается в том, что он характеризует не только степень зависимости между случайными величинами x и y , но и их разброс. В связи с этим в статистике был введен другой показатель меры зависимости между случайными величинами x и y — коэффициент корреляции $r(x, y)$, имеющий вид

$$r(x, y) = \frac{K(x, y)}{\sigma(x)\sigma(y)}, \quad (2.5)$$

где $\sigma(x) = \sqrt{D(x)}$ и $\sigma(y) = \sqrt{D(y)}$ — стандартные отклонения случайных величин x и y соответственно. Коэффициент корреляции принимает значения от -1 до $+1$. К сожалению, построить хорошую теорию оценок коэффициента корреляции, т.е. указать достаточно простой класс его несмещенных оценок, представляет собой более сложную задачу, чем оценка неизвестного математического ожидания, дисперсии или коэффициента ковариации. Эта задача не решена до сих пор. Тем не менее полученные оценки позволяют удовлетворительно решить эту проблему. Действительно, если в формуле (2.5) заменить

показатели $K(x, y)$, $\sigma(x)$ и $\sigma(y)$ их оценками $K^*(x, y)$, $\hat{\sigma}(x) = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2}$,

$\hat{\sigma}(y) = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (y_k - \bar{y})^2}$, то полученная при этом оценка $r^*(x, y) = \frac{K^*(x, y)}{\hat{\sigma}(x)\hat{\sigma}(y)}$ яв-

ляется достаточно хорошим приближением для $r(x, y)$, поскольку $K^*(x, y) \approx K(x, y)$, $\hat{\sigma}(x) \approx \sigma(x)$, $\hat{\sigma}(y) \approx \sigma(y)$. Во всяком случае, при больших значениях объема выборки n на основании известных математических результатов можно гарантировать, что оценка $r^*(x, y)$ будет принимать достаточно близкие к $r(x, y)$ значения.

Резюме

Наилучшей оценкой в классе билинейных несмещенных оценок коэффициентов ковариации $K(x, y)$ является статистика

$$K^*(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}),$$

где $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$, а выборка $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ получена за счет простого случайного выбора. Для оценки коэффициента корреляции $r(x, y) = \frac{K(x, y)}{\sigma(x)\sigma(y)}$ используется статистика

$$r^*(x, y) = \frac{K^*(x, y)}{\hat{s}(x)\hat{s}(y)},$$

которая, хотя и не является наилучшей несмещенной оценкой, тем не менее, является ее приближением для коэффициента $r(x, y)$.

2.5. Оценка неизвестных параметров. Метод наименьших квадратов и метод максимального правдоподобия

До сих пор мы изучали статистические показатели случайных величин x , которые не изменяются со временем или рассматриваются в фиксированный момент времени $x(t_0)$. Такой статистический подход к исследованию случайных явлений был характерен для раннего периода развития теории вероятностей. Однако в дальнейшем появилась необходимость изучать случайные величины, которые изменяются во времени. Такие случайные величины $x(t)$ стали называться *случайными процессами* (если аргумент t пробегает некоторый отрезок, луч или числовую прямую) или *случайными последовательностями* (если аргумент t принимает дискретные значения $t = t_1, t_2, \dots, t_n, \dots$). Эти понятия описаны в разделе 1.5. В данном разделе рассматриваются некоторые довольно простые схемы случайных процессов и их статистические показатели.

Самая простая модель случайного процесса $y(t)$ представляет собой сумму линейной детерминированной функции $m(t) = at + b$ ($t = t_1, t_2, \dots, t_n, \dots$) и последовательности независимых одинаково распределенных случайных величин x_n ($n = 1, 2, \dots$) с нулевым математическим ожиданием $m(x_n) = 0$:

$$y(t_k) = x(t_k) = m(t_k) + x_k = a_k + b + x_k, \quad k = 1, 2, \dots$$

Таким образом, процесс $y(t)$ представляет собой сумму двух функций

$$y(t) = m(t) + x(t), \quad t = t_1, t_2, \dots, t_n, \dots$$

Функция $m(t)$ является детерминированной и называется *трендом* случайного процесса $y(t)$. Функция $x(t)$ представляет собой случайный процесс с нулевым математическим ожиданием $m[x(t)] = 0$ и называется *случайным шумом*, или *простым шумом*. Тренд $m(t)$ рассматривается как полезный сигнал, несущий информацию о состоянии системы, а случайный шум $x(t)$ — как помеха, искажающая полезную информацию.

Если бы не было случайного шума $x(t)$ в процессе $y(t)$, то случайный процесс $y(t)$ стал бы обычной линейной функцией:

$$m(t) = y(t) = at + b = f(t, a, b).$$

Эта функция зависит от двух параметров a и b . В простейшем варианте, когда $b = 0$, функция $y(t)$ зависит от одного параметра a :

$$m(t) = y(t) = at = f(t, a).$$

В этом случае для оценки параметров a и b достаточно вычислить или измерить значения функции $y_k = f(t_k, a, b) = m(t_k; a, b) = at_k + b$, $k = 1, 2$. Тогда определить параметры a и b , а вместе с ними и тренд $m(t; a, b)$ можно, решив систему, состоящую из двух линейных уравнений с двумя неизвестными:

$$\begin{aligned} y_1 &= at_1 + b, \\ y_2 &= at_2 + b. \end{aligned}$$

Совсем иная ситуация возникает, когда на полезный сигнал (тренд) накладывается шум $x(t)$: наличие возмущающего воздействия $x(t)$ приводит к тому, что точки с координатами $(t, m(t))$ “выбиваются” с прямой l , так что выборочные точки $(t_k, y(t_k))$, $k = 1, 2, \dots, n$, разбросаны вокруг этой прямой случайным образом.

Теперь задача оценки параметров a и b усложняется. Метод, позволяющий решить эту трудную задачу, был изобретен Лежандром и Гауссом. Он получил название “метод наименьших квадратов” и является одним из самых замечательных открытий в математике. Сущность этого метода заключается в следующем: после проведения случайного испытания возникает n точек наблюдения с координатами $(t_1, y_1), (t_2, y_2), \dots, (t_n, y_n)$, где $y_k = y(t_k)$, $k = 1, 2, \dots, n$. Для того чтобы прямая l наилучшим образом приближала неизвестную функциональную зависимость $m(t) = at + b$, параметры a и b следует выбрать так, чтобы прямая l прошла как можно ближе к точкам наблюдения $(t_1, y_1), (t_2, y_2), \dots, (t_n, y_n)$. Мерой отклонения прямой l от точек наблюдения естественно выбрать следующую квадратичную функцию:

$$\varphi(a, b) = \sum_{k=1}^n [y_k - m(t_k, a, b)]^2 = \sum_{k=1}^n [y_k - at_k - b]^2.$$

Наилучшими оценками параметров являются те оценки a и b , при которых мера отклонения принимает наименьшее значение. Поскольку $\varphi(a, b)$ представляет собой сумму квадратов, этот метод вычисления оценок стали называть методом наименьших квадратов (МНК). Нетрудно заметить, что приведенные выше рассуждения носят общий характер и не зависят от конкретного вида

тренда $m(t; a, b)$, поэтому их можно без существенных изменений применить к произвольному тренду $m(t; a_0, a_1, \dots, a_r)$. Точнее говоря, квадратичная функция $\varphi(a_0, a_1, \dots, a_r)$, определяющая меру отклонения точек наблюдения от кривой $(t_1, y_1), (t_2, y_2), \dots, (t_n, y_n)$ от кривой l , представляющей собой график тренда $m(t; a_0, a_1, \dots, a_r)$, имеет вид

$$\varphi(a_0, a_1, \dots, a_r) = \sum_{k=1}^n (y_k - m(t_k; a_0, a_1, \dots, a_r))^2.$$

Точка $a_0^*, a_1^*, \dots, a_r^*$, в которой функция $\varphi(a_0, a_1, \dots, a_r)$ достигает минимума, называется оценкой параметров $a_0, a_1, \dots, a_r$, найденной с помощью метода наименьших квадратов. Описанная выше модель называется *регрессией*.

Изложенная выше геометрическая интерпретация метода наименьших квадратов принадлежит французскому математику Лежандру. Немецкий математик К.Ф. Гаусс предложил вероятностную интерпретацию метода наименьших квадратов, сущность которой заключается в следующем. Предположим, что случайный шум $x(t)$ в модели

$$y(t) = m(t; a_0, a_1, \dots, a_r) + x(t), \quad t = t_1, t_2, \dots, t_n, \dots$$

состоит из нормально распределенных случайных величин $x(t_k)$, $k = 1, 2, \dots, n$ с нулевым математическим ожиданием и одинаковой дисперсией σ^2 , независимыми в совокупности. После первого наблюдения реализации случайного процесса $y(t)$ мы получим n пар чисел $(t_1, y(t_1)), (t_2, y(t_2)), \dots, (t_n, y(t_n))$. Однако при втором наблюдении эта выборка изменится и примет вид $(t, \tilde{y}(t_1)), (t, \tilde{y}(t_2)), \dots, (t, \tilde{y}(t_n))$. Можно показать, что для любого малого числа $\varepsilon > 0$ вероятность δP того, что значение $\tilde{y}(t_k)$ принадлежит интервалу $(y(t_k) - \varepsilon/2, y(t_k) + \varepsilon/2)$ при всех $k = 1, 2, \dots, n$, приближенно равна

$$\begin{aligned} \delta P &= P(\tilde{y}(t_1) \in (y(t_1) - \varepsilon/2, y(t_1) + \varepsilon/2), \dots, \tilde{y}(t_n) \in (y(t_n) - \varepsilon/2, y(t_n) + \varepsilon/2)) \approx \\ &\approx \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2} [y(t_1) - m(t_1; a_0, a_1, \dots, a_r)]^2} \dots \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2} [y(t_n) - m(t_n; a_0, a_1, \dots, a_r)]^2} \varepsilon^n = \\ &= \varepsilon^n (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2} \{ [y(t_1) - m(t_1; a_0, a_1, \dots, a_r)]^2 + \dots + [y(t_n) - m(t_n; a_0, a_1, \dots, a_r)]^2 \}}. \end{aligned}$$

Эта вероятность δP является произведением двух сомножителей: первый из них равен $\varepsilon^n (2\pi)^{-\frac{n}{2}}$ и не зависит от параметров $a_0, a_1, \dots, a_r$, а второй зависит от них. Для того чтобы величина δP стала максимальной, параметры $a_0, a_1, \dots, a_r$ следует подобрать так, чтобы сумма

$$[y(t_1) - m(t_1; a_0, a_1, \dots, a_r)]^2 + \dots + [y(t_n) - m(t_n; a_0, a_1, \dots, a_r)]^2$$

была минимальной, так как именно в этом случае величина

$$\varepsilon^n (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2} \{ [y(t_1) - m(t_1; a_0, a_1, \dots, a_r)]^2 + \dots + [y(t_n) - m(t_n; a_0, a_1, \dots, a_r)]^2 \}}$$

достигает максимума. Таким образом, как с геометрической, так и вероятностной точек зрения наилучшими оценками параметров являются оценки, полученные с помощью минимизации квадратичной функции, т.е. оценки МНК.

Перейдем к описанию одного из наиболее глубоких обобщений метода наименьших квадратов — *методу максимального правдоподобия*, предложенному английским статистиком Р. Фишером.

Рассмотрим параметрическую модель зависимости случайного процесса $y(t)$ от неизвестных параметров $a_0, a_1, \dots, a_r$:

$$y(t) = y(t; a_0, a_1, \dots, a_r).$$

Будем считать, что аргумент t принимает дискретные значения $t_1, t_2, \dots, t_n, \dots$. Предположим, что нам известна совместная плотность вероятности $f(u_1, \dots, u_n)$ случайных величин $y(t_1), \dots, y(t_n)$. Эта плотность вероятности зависит от параметров $a_0, a_1, \dots, a_r$. Обозначим через $f(u_1, \dots, u_n | a_0, a_1, \dots, a_r)$ совместную плотность вероятности величин $y(t_1), \dots, y(t_n)$ при условии, что параметры приняли данные значения $a_0, a_1, \dots, a_r$. Пусть в результате случайного испытания получена реализация $y_1 = y(t_1), \dots, y_n = y(t_n)$ процесса $y(t)$. Тогда можно показать, что вероятность δP того, что значение $y(t_i)$ принадлежит интервалу $(y_i - \delta/2, y_i + \delta/2)$ при малых значениях $\delta, i = 1, 2, \dots, n$, приближенно равна

$$\delta P \approx f(y_1, \dots, y_n | a_0, a_1, \dots, a_r) \delta_1 \delta_2 \dots \delta_n.$$

Мы не располагаем никакой предварительной информацией о значениях неизвестных параметров $a_0, a_1, \dots, a_r$. Единственное, что можно использовать при их оценке, — это наблюдаемые значения $y_1, \dots, y_n$. Хорошие оценки параметров должны наилучшим образом согласовываться с наблюдениями $y_1, \dots, y_n$. Очевидно, что о степени такой согласованности можно судить по величине вероятности δP . Чем больше δP , тем лучше оценки согласуются с выборочными значениями $y_1, \dots, y_n$. Степень согласования является наибольшей при тех значениях $a_0^*, a_1^*, \dots, a_r^*$, при которых функция $\delta P(a_0, a_1, \dots, a_r)$ достигает максимума. Точка максимума $(a_0^*, a_1^*, \dots, a_r^*)$ называется оценкой, полученной по методу максимального правдоподобия:

$$\delta P(a_0^*, a_1^*, \dots, a_r^*) = \max_{(a_0, a_1, \dots, a_r)} \delta P(a_0, a_1, \dots, a_r).$$

Легко видеть, что функция $\delta P(a_0, a_1, \dots, a_r)$ достигает максимального значения в той точке, где функция $f(y_1, \dots, y_n | a_0, a_1, \dots, a_r) = \varphi(a_0, a_1, \dots, a_r)$ принимает наибольшее значение, поэтому задача поиска правдоподобных оценок параметров сводится к поиску глобального максимума функции правдоподобия $\varphi(a_0, a_1, \dots, a_r)$.

Резюме

Схема регрессии представляет собой модель оценки неизвестной функциональной зависимости $\hat{y}(t) = m(t; a_0, a_1, \dots, a_r)$, в которой вид этой зависимости известен с точностью до параметров $(a_0, a_1, \dots, a_r)$, т.е. функция $m(t; a_0, a_1, \dots, a_r)$

выражается формулой, содержащей неизвестные параметры $a_0, a_1, \dots, a_r$. Характерная особенность этой модели состоит в том, что на наблюдаемые значения функции $m(t; a_0, a_1, \dots, a_r)$ накладывается случайный шум $x(t)$, поэтому мы наблюдаем не истинные значения $\tilde{y}(t) = m(t; a_0, a_1, \dots, a_r)$, а искаженные случайным шумом величины:

$$y(t) = \tilde{y}(t) + x(t) = m(t; a_0, a_1, \dots, a_r) + x(t).$$

Функция $m(t; a_0, a_1, \dots, a_r)$ называется трендом. Классическим методом оценки тренда является метод наименьших квадратов, который заключается в поиске минимума суммы квадратов

$$\Phi(a_0, a_1, \dots, a_r) = \sum_{k=1}^n (y_k - m(t_k; a_0, a_1, \dots, a_r))^2,$$

когда параметры $a_0, a_1, \dots, a_r$ принадлежат некоторой допустимой области, а $y_k = y(t_k)$. Выбирая значения параметров, при которых функция $\Phi(a_0, a_1, \dots, a_r)$ достигает минимума, мы можем приблизить неизвестный тренд оценкой $m(t; a_0^*, a_1^*, \dots, a_r^*)$.

Метод наибольшего правдоподобия позволяет получить более точные оценки параметров, однако для этого необходимо знать совместную плотность вероятностей случайных величин $y(t_1), y(t_2), \dots, y(t_n)$, что на практике бывает не всегда.

2.6. Схема линейной регрессии. Несмещенные линейные оценки неизвестных параметров

В предыдущем разделе мы описали модель случайного процесса $y(t)$, представляющего собой сумму тренда $m(t)$ и случайного шума $x(t)$. Рассмотрим теперь конкретные модели, в которых тренд $m(t)$ линейно зависит от неизвестных параметров $y(t_1), y(t_2), \dots, y(t_n)$. Как указано в разделе 2.2, для того чтобы получить хорошие оценки неизвестных параметров, необходимо указать класс таких оценок, среди которых существует наилучшая. По классической теореме К. Вейерштрасса всякую непрерывную функцию $m(t)$, определенную на отрезке $[0, T]$, можно с любой степенью точности ε приблизить алгебраическим многочленом $P(t) = a_0 + a_1 t + \dots + a_f t^f$, т.е.

$$\max_{t \in [0, T]} |m(t) - P(t)| < \varepsilon.$$

Естественно предположить, что тренд $m(t)$ представляет собой алгебраический многочлен r -й степени:

$$m(t) = a_0 + a_1 t + \dots + a_f t^f, t \in [0, T].$$

Коэффициенты $a_0, a_1, \dots, a_f$ называются *коэффициентами регрессии*, считаются неизвестными и подлежат оценке.

Итак, случайный процесс $y(t)$ с трендом $m(t)$ можно приблизить следующей моделью:

$$y(t) = a_0 + a_1 t + \dots + a_r t^r + x(t), \quad m[x(t)] = 0, \quad t \in [0, T].$$

Эта модель называется *схемой линейной регрессии*.

Для построения наилучших линейных оценок неизвестных коэффициентов регрессии рассмотрим случайный процесс $y(t)$ с дискретным временем $t = t_1, t_2, \dots, t_n, \dots$. Введем следующие обозначения: $\varphi_0(t) \equiv 1$, $\varphi_1(t) \equiv t$, ..., $\varphi_r(t) \equiv t^r$. Следовательно,

$$y(t) = \sum_{j=0}^r a_j \varphi_j(t) + x(t), \quad t = t_1, t_2, \dots, t_n, \dots$$

Более общую модель линейной регрессии можно построить, если функции $\varphi_k(t)$, $k = 1, 2, \dots, r$, являются произвольными *функциями регрессии*, известными заранее. Эта модель, предложенная выдающимся русским математиком А.А. Марковым, называется *общей моделью линейной регрессии*.

В качестве функций регрессии не обязательно выбирать степенные функции $\varphi_0(t) \equiv 1$, $\varphi_1(t) \equiv t$, ..., $\varphi_r(t) \equiv t^r$. Во многих случаях в качестве функций $\varphi_k(t)$ используются функции $\cos kt$ и $\sin kt$, так что тренд представляет собой не алгебраический, а тригонометрический многочлен

$$m(t; a_0, a_1, \dots, a_r, b_0, b_1, \dots, b_r) = \sum_{k=0}^r (a_k \cos kt + b_k \sin kt).$$

Кроме того, в качестве функций регрессии можно взять функции $\varphi_k(t) = e^{kt}$ и другие.

Далеко не всякий тренд $m(t; a_0, a_1, \dots, a_r)$ можно представить в виде $\sum_{k=0}^r a_k \varphi_k(t)$.

Например, тренд $m(t; a_0, a_1) = e^{a_0 t + a_1 t^2}$ нельзя представить в виде $a_0 \varphi_0(t) + a_1 \varphi_1(t)$, где $\varphi_0(t)$ и $\varphi_1(t)$ — известные функции. По этой причине многие параметрические модели трендов не являются моделями линейной регрессии.

Наилучшие оценки коэффициентов линейной регрессии a_i , $i = 0, 1, \dots, r$, естественно искать в классе несмещенных оценок L_r :

$$L_r = \left\{ u_i = \sum_{k=1}^n c_{ik} y(t_k), \quad m(u_i) = a_i \right\}.$$

Поскольку мы требуем, чтобы оценки были несмещенными, должно выполняться следующее равенство:

$$a_i = m(u_i) = \sum_{k=1}^n c_{ik} m[y(t_k)] = \sum_{k=1}^n c_{ik} \sum_{j=0}^r a_j \varphi_j(t_k) = \sum_{j=0}^r a_j c_{ik} \varphi_j(t_k).$$

Так как мы не знаем конкретных значений a_j , необходимо, чтобы равенство

$$a_i = \sum_{k=1}^n \sum_{j=0}^r a_j c_{ik} \varphi_j(t_k) = \sum_{j=0}^r a_j \sum_{k=1}^n c_{ik} \varphi_j(t_k)$$

выполнялось при любом значении a_i . Таким образом, должны выполняться следующие равенства:

$$\sum_{k=0}^n c_{ik} \varphi_j(t_k) = 0, \text{ если } i \neq j, \quad (2.6)$$

$$\sum_{k=0}^n c_{ik} \varphi_i(t_k) = 1. \quad (2.7)$$

Предположим, что случайные величины $x(t_1), x(t_2), \dots, x(t_n)$ являются независимыми в совокупности и имеют одинаковые дисперсии σ^2 . Тогда

$$D(u_i) = \sum_{k=1}^n c_{ik}^2 \sigma^2.$$

Минимизируя этот квадратичный многочлен относительно коэффициентов c_{ik} при условиях (2.6), пользуясь методами поиска условного экстремума, можно получить формулы для коэффициентов c_{ik}^* , позволяющие вычислить наилучшую несмещенную оценку u_i^* параметра a_i . Эти формулы довольно громоздки, поэтому мы их не приводим. Вместо этого немного ниже мы получим оценки u_i^* , используя метод наименьших квадратов.

В общем случае, когда наблюдения являются зависимыми, наилучшие линейные оценки u_i^* были найдены А. Айткеном.

Пусть K — корреляционная матрица случайных величин $y(t_1), \dots, y(t_n)$:

$$K = \begin{Bmatrix} K(t_1, t_1) & K(t_1, t_2) & \dots & K(t_1, t_n) \\ K(t_2, t_1) & K(t_2, t_2) & \dots & K(t_2, t_n) \\ \dots & \dots & \dots & \dots \\ K(t_n, t_1) & K(t_n, t_2) & \dots & K(t_n, t_n) \end{Bmatrix}.$$

Обозначим через Φ матрицу, составленную из элементов $\varphi_i(t_j)$, $i = 0, \dots, r$, $j = 1, \dots, n$,

$$\Phi = \begin{Bmatrix} \varphi_0(t_1) & \varphi_1(t_1) & \dots & \varphi_r(t_1) \\ \varphi_0(t_2) & \varphi_1(t_2) & \dots & \varphi_r(t_2) \\ \dots & \dots & \dots & \dots \\ \varphi_0(t_n) & \varphi_1(t_n) & \dots & \varphi_r(t_n) \end{Bmatrix},$$

а через Y и U^* — векторы-столбцы, состоящие из реализации последовательности $y(t)$ и наилучших оценок соответственно:

$$Y = \begin{Bmatrix} y(t_1) \\ y(t_2) \\ \dots \\ y(t_n) \end{Bmatrix}, \quad U^* = \begin{Bmatrix} u_0^* \\ u_1^* \\ \dots \\ u_n^* \end{Bmatrix}.$$

Тогда вектор наилучших оценок параметров $a_i, i = 0, 1, \dots, r$ можно записать следующим образом:

$$U^* = (\Phi^T K^{-1} \Phi)^{-1} (\Phi^T K Y).$$

Здесь K^{-1} — матрица, обратная к матрице K . Эта запись означает, что сначала вычисляется обратная матрица K^{-1} , а затем произведение матриц $\Phi^T K^{-1} \Phi$. Подчеркнем, что умножение матриц является некоммутативной операцией, поэтому вычисления следует проводить в строгом порядке: матрица K^{-1} умножается на Φ , а затем матрица Φ^T умножается на $K^{-1} \Phi$. По результату $\Phi^T K^{-1} \Phi$ вычисляется обратная матрица $(\Phi^T K^{-1} \Phi)^{-1}$. Следующим этапом вычислений является умножение матрицы Φ^T на матрицу K в указанном порядке. После этого матрица $\Phi^T K$ умножается на вектор Y . На последнем этапе матрица $(\Phi^T K^{-1} \Phi)^{-1}$ умножается на вектор $\Phi^T K Y$. Полученный в итоге вектор состоит из наилучших линейных оценок u_i^* .

Если известны оценки сверху для модуля коэффициентов регрессии, т.е. $|a_j| \leq \alpha_j$, то можно получить еще более точные линейные оценки u_j для коэффициентов a_j . Эти оценки являются смещенными, но имеют меньшую ошибку по сравнению с оценками u_i^* .

Формулы для вычисления оценок u_j имеют следующий вид. Обозначим через $c(j)$ вектор-столбец, состоящий из весовых коэффициентов оценки u_j , т.е. скалярных произведений

$$u_j = (Y, c(j)).$$

Тогда

$$c(j) = \alpha_j^2 \left[K + (\alpha_j)^2 \Phi(j) \right]^{-1} \left[f(j) - F(j) \right] \left\{ F'(j) \left[K + \alpha_j^2 \Phi(j) \right]^{-1} F(j) \right\}^{-1} \times \\ \times F'(j) \left[K + \alpha_j^2 \Phi(j) \right]^{-1} f(j).$$

Здесь

$$f(j) = (\varphi(t_1), \dots, \varphi(t_n)), \Phi(j) = \{\varphi(t_i)\varphi(t_k)\}_{i,k=1,\dots,n}, F(j) = \{\varphi_k(t_i)\}_{i=1,\dots,n, k=0,\dots,j-1,j+1,\dots,r}.$$

Оценки u_j оказываются особенно полезными при выделении слабого сигнала

$$m(t) = \sum_{j=0}^r a_j \varphi_j(t).$$

на фоне случайного шума. Следует отметить, что для последовательности независимых случайных величин $y(t_1), \dots, y(t_n)$ с одинаковыми дисперсиями σ^2 оценка имеет очень простой вид

$$u = \frac{1}{\gamma^2 + n} \sum_{i=1}^n y(t_i),$$

где $\gamma = \sigma/\alpha$, α — оценка сверху для $|m|$. При малых значениях α эта оценка сильно отличается от выборочной средней.

Рассмотрим линейную регрессию первого порядка:

Используя стандартные методы поиска минимума функции

нетрудно показать, что оценки МНК имеют вид

где $\bar{t} = \frac{1}{n} \sum_{i=1}^n t_i$, $\bar{y} = \frac{1}{n} \sum_{i=1}^n y(t_i)$.

$$\begin{aligned} na_0^* + [t]a_1^* + \dots + [t^r]a_r^* &= [y], \\ [t]a_0^* + [t^2]a_1^* + \dots + [t^{r+1}]a_r^* &= [ty], \end{aligned}$$

где $\sum_{i=1}^n t^k = [t^k]$, $k=1, 2, \dots$, $\sum_{i=1}^n t_i^k = [t^k y]$.

Резюме

$$y(t) = \sum_{j=0}^r a_j \varphi_j(t) + x(t), \quad t = t_1, t_2, \dots, t_n,$$

где $\varphi(t)$ — известная функция, a_j — неизвестные коэффициенты линейной регрессии, $x(t)$ — случайный шум. Первое слагаемое, т.е. тренд, является функцией, линейной относительно параметров $a_0, a_1, \dots, a_r$, поэтому эта модель называется схемой линейной регрессии. Для оценки параметров $a_0, a_1, \dots, a_r$ можно воспользоваться методом наименьших квадратов, если величины $x(t_1), \dots, x(t_n)$ независимы и одинаково распределенные величины. В общем случае, когда эти величины являются зависимыми, для оценки неизвестных параметров применяется формула Айткена:

$$U^* = (\Phi^T K^{-1} \Phi)^{-1} (\Phi^T K Y).$$

Здесь K — корреляционная матрица случайных величин $y(t_1), \dots, y(t_n)$:

$$K = \begin{Bmatrix} K(t_1, t_1) & K(t_1, t_2) & \dots & K(t_1, t_n) \\ K(t_2, t_1) & K(t_2, t_2) & \dots & K(t_2, t_n) \\ \dots & \dots & \dots & \dots \\ K(t_n, t_1) & K(t_n, t_2) & \dots & K(t_n, t_n) \end{Bmatrix},$$

Φ — матрица, составленная из элементов $\varphi(t_j)$, $i = 0, \dots, r, j = 1, \dots, n$,

$$\Phi = \begin{Bmatrix} \varphi_0(t_1) & \varphi_1(t_1) & \dots & \varphi_r(t_1) \\ \varphi_0(t_2) & \varphi_1(t_2) & \dots & \varphi_r(t_2) \\ \dots & \dots & \dots & \dots \\ \varphi_0(t_n) & \varphi_1(t_n) & \dots & \varphi_r(t_n) \end{Bmatrix},$$

а Y и U^* — векторы-столбцы, состоящие из реализации последовательности $y(t)$ и наилучших оценок соответственно:

$$Y = \begin{Bmatrix} y(t_1) \\ y(t_2) \\ \dots \\ y(t_n) \end{Bmatrix}, \quad U^* = \begin{Bmatrix} u_0^* \\ u_1^* \\ \dots \\ u_n^* \end{Bmatrix}.$$

В простейшем случае, когда $m(t; a_0, a_1) = a_0 + a_1 t$, оценки МНК (метода наименьших квадратов) имеют вид

$$a_1^* = \frac{\sum_{i=1}^n (t_i - \bar{t})(y_i - \bar{y})}{\sum_{i=1}^n (t_i - \bar{t})^2},$$

$$a_0^* = \bar{y} = a_1^* \bar{t},$$

где $\bar{t} = \frac{1}{n} \sum_{i=1}^n t_i$, $\bar{y} = \frac{1}{n} \sum_{i=1}^n y(t_i)$.

2.7. Модель сплайновой регрессии

Рассмотрим произвольный отрезок $[a, b]$, на котором выбраны фиксированные точки $\tau_1, \tau_2, \dots, \tau_r$, называемые *узлами*. Сплайном $S(t)$ называется функция, определенная на отрезке $[a, b]$, обладающая следующими свойствами:

- 1) на любом отрезке $[a, \tau_1], [\tau_1, \tau_2], \dots, [\tau_r, b]$ с концами в соседних узлах или в точках a и b функция $S(t)$ является алгебраическим многочленом;
- 2) на всем отрезке $[a, b]$ функция $S(t)$ непрерывна и имеет производные $S^{(k)}(t)$ до k -го порядка включительно, где число k может принимать натуральные значения и нуль; в случае, когда $k=0$, предполагается, что функция $S(t)$ в некоторых точках может не иметь производных, но всегда считается непрерывной.

Первое условие означает, что внутри отрезков $[a, \tau_1], [\tau_1, \tau_2], \dots, [\tau_r, b]$ у функции $S(t)$ существуют производные любого порядка, так что производные могут отсутствовать только в узлах сплайна.

Наиболее простым примером является линейный сплайн. У таких сплайнов на отрезках $[a, \tau_1], [\tau_1, \tau_2], \dots, [\tau_r, b]$ функция $S(t)$ является линейной. Таким образом, линейный сплайн представляет собой кусочно-линейную непрерывную функцию с вершинами в узлах сплайна $\tau_1, \tau_2, \dots, \tau_r$. Схематическое изображение линейного сплайна представлено на рис. 2.1.

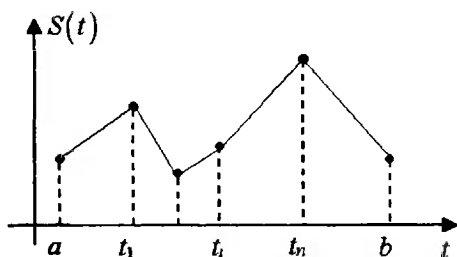


Рис. 2.1. Реализация случайного процесса

Интуитивное описание линейного сплайна можно формализовать следующим образом:

$$S(t) = \begin{cases} a_0 t + b_0, & \text{если } t \in [a, \tau_1], a_0 \tau_1 + b_0 = a_1 \tau_1 + b_1, \\ \dots \\ a_i t + b_i, & \text{если } t \in [\tau_i, \tau_{i+1}], a_i \tau_i + b_i = a_{i-1} \tau_i + b_{i-1} \\ \dots \\ a_r t + b_r, & \text{если } t \in [\tau_r, b], a_r \tau_r + b_r = a_{r-1} \tau_r + b_{r-1}. \end{cases}$$

Если у линейного сплайна $S(t)$ есть только один узел t_1 , то функция принимает следующий вид:

$$S(t) = \begin{cases} a_0 t + b_0, & \text{если } t \in [a, \tau_1], a_0 \tau_1 + b_0 = a_1 \tau_1 + b_1, \\ a_1 t + b_1, & \text{если } t \in [\tau_1, b]. \end{cases}$$

Нетрудно заметить, что любой сплайн $S(t)$ полностью определяется конечным набором параметров $a_0, \dots, a_r$, т.е. представляет собой функцию $S(t; a_0, \dots, a_r)$. Следовательно, модель сплайновой регрессии

$$y(t) = S(t) + x(t) = S(t; a_0, \dots, a_r) + x(t),$$

где $x(t)$ — случайный шум с нулевым математическим ожиданием $m[x(t)] = 0$, представляет собой параметрическую модель с неизвестными параметрами $a_0, \dots, a_r$. Естественным образом возникает вопрос: является ли сплайновая регрессия схемой линейной регрессии? Ответ на этот вопрос зависит от того, известны ли узлы сплайна $\tau_1, \tau_2, \dots, \tau_n$. Можно показать, что в том случае, когда все узлы сплайна известны, а неизвестны лишь коэффициенты многочленов, заданных на отрезках $[a, \tau_1], [\tau_1, \tau_2], \dots, [\tau_r, b]$, модель сплайновой регрессии является схемой линейной регрессии. Если же неизвестен хотя бы один узел τ_k , $k = 1, \dots, l$, то сплайновая модель является схемой нелинейной регрессии.

Как показано в разделе 2.6, при статистической обработке экспериментальных и клинических данных естественно считать, что время t изменяется дискретно, так что $t = t_1, t_2, \dots, t_n, \dots$. Будем предполагать, что узлы τ_k сплайна $S(t)$ расположены лишь в точках наблюдения значений сплайна $t_1, t_2, \dots, t_m$: $\tau_k = t_p$, $k = 1, 2, \dots, l$; $i = 1, 2, \dots, m$. Рассмотрим вначале самый простой (и наиболее распространенный на практике) линейный сплайн $S_{\tau_k}(t) = S_i(t)$ с узлом в точке $\tau_k = t_i$. В физике узел τ_k интерпретируется как точка перехода процесса $y(t)$, поскольку на отрезке $[a, \tau_k]$ процесс $y(t)$ описывается уравнением $y(t) = a_0 t + b_0 + x(t)$, а на отрезке $[\tau_k, b]$ динамика процесса $y(t)$ меняется и подчиняется закону $y(t) = a_1 t + b_1 + x(t)$ (рис. 2.2).

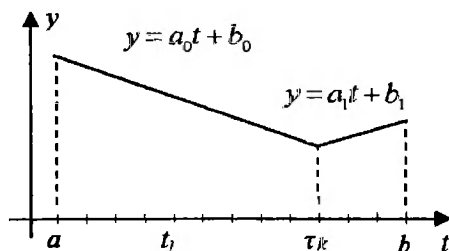


Рис. 2.2. Модель сплайновой регрессии

Нетрудно показать, что сплайн $S_{\tau_k}(t)$ можно представить в виде

$$S_{\tau_k}(t) = a_0 t + b_0 + c_1 (t - \tau_k),$$

$$\text{где } (t - \tau_k)_+ = \begin{cases} 0, & \text{если } t - \tau_k \leq 0, \\ t - \tau_k, & \text{если } t - \tau_k > 0 \end{cases}$$

является так называемой усеченной степенной функцией. Таким образом, в случае известной точки перехода τ_k (узла сплайна) мы приходим к уже рассмотренной нами модели линейной регрессии (см. раздел 2.6):

$$y(t) = \sum_{j=0}^r a_j \varphi_j(t) + x(t), \quad t = t_1, t_2, \dots, t_n, \dots,$$

в которой $r = 3$, $\varphi_0(t) \equiv t$, $\varphi_1(t) \equiv 1$, $\varphi_2(t) \equiv (t - \tau_k)_+$ — известные функции регрессии, $a_0, a_1 = b_0, a_2 = c_1$ — неизвестные коэффициенты линейной регрессии, подлежащие оценке. Эти неизвестные коэффициенты a_0, b_0, c_1 можно оценить на основании формул Айткена или с помощью метода наименьших квадратов (согласно теореме Гаусса–Маркова эти оценки совпадают). Если точка t_i в которой расположен узел сплайна $S_k(t)$, неизвестна, то вычисление наилучших оценок коэффициентов a_0, b_0, c_1 проводится следующим образом.

1. Вначале предполагается, что точкой перехода τ_k являются поочередно точки наблюдения значения процесса $y(t)$: $\tau_k = t_1, t_2, \dots, t_l$. Тогда вычисленные оценки a_0, b_0, c_1 будут зависеть от точки t_i , в которой мы помещаем узел τ_k : $\tau_k = t_i$. Эти оценки естественно обозначить символами $a_0^{(i)}, b_0^{(i)}, c_1^{(i)}$, поскольку они определяются точкой $\tau_k = t_i, i = 1, 2, \dots, l$.
2. После вычисления оценок $a_0^{(i)}, b_0^{(i)}, c_1^{(i)}$ определяется так называемая “остаточная сумма” квадратов s_i^2 :

$$s_i^2 = \sum_{j=1}^l \left(y(t_j) - a_0^{(i)} t_j - b_0^{(i)} - c_1^{(i)} (t_j - \tau_k)_+ \right)^2, \quad i = 1, 2, \dots, l.$$

3. Среди чисел s_i^2 выбирается минимальное число $\hat{s}_i^2$:

$$\hat{s}_i^2 = \min(s_1^2, s_2^2, \dots, s_l^2)$$

и соответствующая этому значению остаточная сумма квадратов оценки коэффициентов регрессии $a_0^{(i)}, b_0^{(i)}, c_1^{(i)}$ при условии, что $\tau_k = t_i$. Эти оценки называются оценками сплайновой регрессии.

Если сплайн $S(t)$ имеет больше одного узла в точках $\tau_1, \tau_2, \dots, \tau_k$, причем эти точки известны и расположены в точках наблюдения процесса $y(t)$: $t_{k_1} = \tau_1, t_{k_2} = \tau_2, \dots, t_{k_k} = \tau_k$, то модель сплайновой регрессии $y(t)$ можно записать в виде

$$y(t) = a_0 t + b_0 + c_1(t - \tau_1)_+ + c_2(t - \tau_2)_+ + \dots + c_k(t - \tau_k)_+,$$

где $(t - \tau_k)_+$ — усеченная степенная функция, числа $a_0, b_0, c_1, \dots, c_k$ — неизвестные коэффициенты линейной регрессии, а $\varphi_0(t) \equiv t$, $\varphi_1(t) \equiv 1$, $\varphi_2(t) \equiv (t - \tau_k)_+$, $\dots, \varphi_{k+2}(t) \equiv (t - \tau_k)_+$ — известные функции регрессии. Коэффициенты линейной

регрессии можно определить по формулам Айткена, так что мы имеем дело с обычной схемой линейной регрессии. В случае неизвестных узлов сплайна проблема сплайновой регрессии резко усложняется, поскольку остаточная сумма квадратов σ^2 становится суммой, зависящей от нескольких дискретных аргументов, а поиск экстремума такой функции является нелегкой задачей.

Можно рассмотреть еще более сложную задачу, в которой узлы сплайна $S(t)$ могут быть левыми концами отрезка $[a, b]$ и не совпадать с точками наблюдения. Такая модель является моделью нелинейной регрессии. Доказать существование оценок МНК в этой модели удалось лишь для линейных и кубических сплайнов.

Модель сплайновой регрессии нашла широкое применение во многих разделах доказательной медицины: при диагностике внутренних заболеваний, в радиобиологии, при описании изменения уровня антигенов у онкологических больных и т.д. В последующих разделах книги мы опишем некоторые приложения этой модели в клинической медицине.

Резюме

Сплайном $S(t)$ с узлами в точках $a < \tau_1 < \tau_2 < \dots < \tau_r < b$ называется функция, которая на любом отрезке $[a, \tau_1]$, $[\tau_1, \tau_2]$, ..., $[\tau_r, b]$ является алгебраическим многочленом, а на всем отрезке $[a, b]$ является дифференцируемой (гладкой) и имеет производные до k -го порядка включительно. В простейшем случае так называемый “линейный” сплайн $S(t)$ с одним узлом в точке τ_1 имеет вид

$$S(t) = \begin{cases} a_0 t + b_0, & \text{если } t \in [a, \tau_1], a_0 \tau_1 + b_0 = a_1 \tau_1 + b_1, \\ a_1 t + b_1, & \text{если } t \in [\tau_1, b]. \end{cases}$$

Если узлы сплайна $\tau_1, \tau_2, \dots, \tau_r$ известны, то вычисление оценок параметров в модели сплайновой регрессии

$$y(t) = S(t) + x(t), \quad t = t_1, t_2, \dots, t_l$$

сводится с помощью так называемых “усеченных степенных функций” к классической модели линейной регрессии. Если узел τ_k (точка перехода) линейного сплайна $S(t)$ неизвестен, то оценка коэффициентов a_0, b_0, c_1 в

$$S(t) = a_0 t + b_0 + c_1 (t - \tau_k)_+,$$

где

$$(t - \tau_k)_+ = \begin{cases} 0, & \text{если } t - \tau_k \leq 0, \\ t - \tau_k, & \text{если } t - \tau_k > 0, \end{cases}$$

является так называемой “усеченной степенной функцией”, проводится следующим образом.

1. Поочередно предполагается, что точкой перехода τ_k являются точки наблюдения значения процесса $y(t)$: $\tau_k = t_1, t_2, \dots, t_l$. Вычисленные оценки $a_0^{(i)}, b_0^{(i)}, c_1^{(i)}$ зависят от точки t_i , в которой мы помещаем узел τ_k : $\tau_k = t_i$.

2. После вычисления оценок $a_0^{(i)}, b_0^{(i)}, c_1^{(i)}$ определяется так называемая “остаточная сумма” квадратов s_i^2 :

$$s_i^2 = \sum_{j=1}^l \left(y(t_j) - a_0^{(i)} t_j - b_0^{(i)} - c_1^{(i)} (t_j - \tau_k)_+ \right)^2, i = 1, 2, \dots, l.$$

3. Среди чисел s_i^2 выбирается минимальное число $\hat{s}_i^2$:

$$\hat{s}_i^2 = \min(s_1^2, s_2^2, \dots, s_l^2)$$

и соответствующая этому значению остаточная сумма квадратов оценки коэффициентов регрессии $a_0^{(i)}, b_0^{(i)}, c_1^{(i)}$ при условии, что $\tau_k = t_i$. Эти оценки называются оценками сплайновой регрессии.

2.8. Оценки неизвестных функций распределения и плотности вероятностей

Как указано в разделе 2.1, у генеральной совокупности G существует функция распределения $F(u)$, поскольку элементы генеральной совокупности можно рассматривать как возможные значения некоторой случайной величины ξ . Функцию распределения $F(u) = P(\xi < u)$ будем называть базовой функцией распределения генеральной совокупности G . Значение ξ_1 случайной величины ξ можно получить после проведения случайного испытания T . Если это испытание T повторить n раз, то мы получим выборку $\xi_1, \xi_2, \dots, \xi_n$, на основании которой необходимо оценить базовую функцию распределения $F(u)$. Подчеркнем, что во всех наших дальнейших рассуждениях случайное испытание T является простым случайным выбором, так что все выборочные значения $\xi_1, \xi_2, \dots, \xi_n$ получены путем простого случайного выбора. Основная цель этого подраздела — указать оценки для базовой функции распределения и ее плотности вероятности по выборочным значениям из генеральной совокупности G .

В дальнейшем будем обозначать выборочные значения символами $x_1, x_2, \dots, x_n$. Построим по выборочным значениям $x_1, x_2, \dots, x_n$ вариационный ряд

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)},$$

а также ступенчатую функцию $F_n^*(u)$, которая в каждой точке $x_{(i)}$ совершает скачок величиной $1/n$, а в интервале $(x_{(i)}, x_{(i+1)})$, $i = 1, 2, \dots, n-1$, является постоянной (рис. 2.3).

Для определенности будем считать, что в точке $x_{(i)}$ имеет место равенство $F_n^*(x_{(i)}) = \frac{i-1}{n}$. Следовательно,

$$F_n^*(u) = \begin{cases} 0, & \text{если } u \leq x_{(1)}, \\ \frac{i}{n}, & \text{если } x_{(i)} < u \leq x_{(i+1)}, \\ 1, & \text{если } u > x_{(n)}. \end{cases}$$

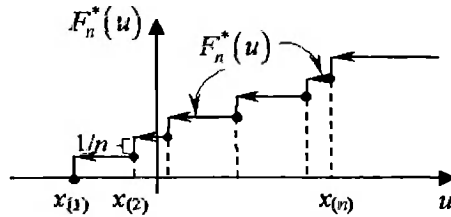


Рис. 2.3. Эмпирическая функция распределения

Функция $F_n^*(u)$ называется *выборочной*, или *эмпирической* функцией распределения, поскольку она построена на основе выборки $x_1, x_2, \dots, x_n$ (эмпирических данных). Используя выборочные значения, можно указать оценку для плотности вероятности $f(u)$, если она существует. Для этого следует рассмотреть ступенчатую функцию $f^*(u)$, которая принимает постоянные значения

$\gamma_i = \frac{1}{(n-1)(x_{(i+1)} - x_{(i)})}$ на каждом интервале $(x_{(i)}, x_{(i+1)})$ (рис. 2.4). Для определен-

ности можно считать, что в точке $x_{(i)}$ выполняется условие $f_n^*(x_{(i)}) = \gamma_i$. Если $u \notin (x_{(i)}, x_{(i+1)})$, то $f_n^*(u) = 0$.

Мы будем называть функцию $f_n^*(u)$ *унифицированной гистограммой* (рис. 2.4). Соответствующая ей функция распределения

$$\tilde{F}_n(u) = \int_{-\infty}^u f_n^*(t) dt, \quad (\tilde{F}_n'(u) = f_n^*(t)),$$

называется *модифицированной выборочной функцией распределения*. Эта функция $\tilde{F}_n(u)$ имеет следующий вид (рис. 2.5):

$$\tilde{F}_n(u) = \begin{cases} 0, & \text{если } u \leq x_{(1)}, \\ \frac{u + i(x_{(i+1)} - x_{(i)}) - x_{(i+1)}}{(n-1)(x_{(i+1)} - x_{(i)})}, & \text{если } x_{(i)} < u \leq x_{(i+1)}, \\ 1, & \text{если } u > x_{(n)}. \end{cases}$$

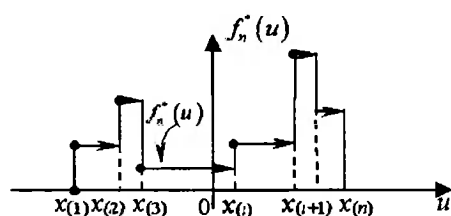


Рис. 2.4. Унифицированная гистограмма

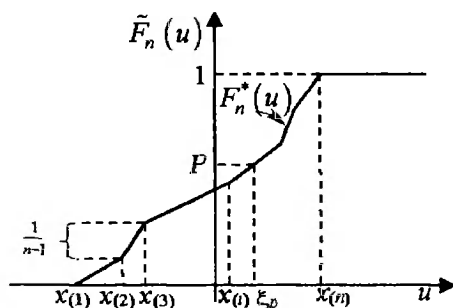


Рис. 2.5. Модифицированная выборочная функция распределения

Модифицированная выборочная функция распределения $\tilde{F}_n(u)$ является непрерывной кусочно-линейной функцией (сплайном). Она равна нулю при $u \leq x_{(1)}$, в точках $x_{(i)}$ принимает значения $\frac{i-1}{n-1}$, на интервале $(x_{(i)}, x_{(i+1)})$ линейна, а при $u > x_{(n)}$ тождественно равна единице. Эта оценка базовой функции распределения $F(u)$ генеральной совокупности G является более предпочтительной по сравнению с эмпирической функцией распределения $F_n^*(u)$. Это объясняется следующими соображениями. Во-первых, в отличие от разрывной кусочно-постоянной эмпирической функции распределения $F_n^*(u)$ функция $\tilde{F}_n(u)$ является непрерывной. В случае непрерывной базовой функции генеральной совокупности $F(u)$ модифицированная выборочная функция распределения $\tilde{F}_n(u)$ на интервале $(x_{(i)}, x_{(i+1)})$ осуществляет стохастическую линейную интерполяцию значений функции $F(u)$, которая, как правило, является более точной по сравнению с аппроксимацией на основе функции $F_n^*(u)$. Во-вторых, во многих задачах обработки эмпирических данных необходимо оценить квантиль распределения. Напомним, что величина ξ_p называется квантилем распределения $F(u)$ порядка p , если $F(\xi_p) = p$, где $p \in (0, 1)$. В частности, при $p = 1/2$ квантиль $\xi_{1/2}$ представляет собой медиану, которая является очень важной характеристикой меры расположения множества значений генеральной совокупности. Если пользоваться модифицированной выборочной функцией распределения $\tilde{F}_n(u)$, то у этого распределения квантиль ξ_p порядка p всегда существует и его нетрудно определить (см. рис. 2.5), в то время как квантиль порядка p у выборочной функции распределения $F_n^*(u)$ почти никогда не существует, поскольку эта функция принимает лишь конечное количество значений. Поскольку функция $\tilde{F}_n(u)$ аппроксимирует базовую функцию распределения $F(u)$, “выборочный” квантиль ξ_p , построенный по функции $\tilde{F}_n(u)$, является оценкой неизвестного квантиля ξ_p базовой функции распределения генеральной совокупности.

Оценку плотности вероятностей $f(u)$ распределения $F(u)$ можно значительно улучшить, если вместо унифицированной гистограммы $f_n^*(u)$ использовать модифицированную гистограмму $\tilde{f}_n(u)$. Для этого построим по выборочным значениям $x_1, x_2, \dots, x_n$ вариационный ряд

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}.$$

Выберем $l < n$ порядковых статистик $x_{(k_1)} \leq x_{(k_2)} \leq \dots \leq x_{(k_l)}$ (не обязательно последовательных!), которые будем называть *базовыми порядковыми статистиками*, и построим ступенчатую функцию $\tilde{f}_k(u)$, которая равна нулю вне интервала

$$(x_{(k_1)}, x_{(k_l)}), \text{ а в интервалах } (x_{(k_i)}, x_{(k_{i+1})}) \text{ принимает значения } \frac{1}{(l-1)(x_{(k_{i+1})} - x_{(k_i)})}$$

(рис. 2.6), как правило, $x_{(k_1)} = x_{(1)}$, $x_{(k_l)} = x_{(n)}$. Таким образом, модифицированная гистограмма $\tilde{f}_k(u)$ строится так же, как и унифицированная гистограмма, если вместо вариационного ряда взять группу всех базовых порядковых статистик $x_{(k_1)} \leq x_{(k_2)} \leq \dots \leq x_{(k_l)}$.

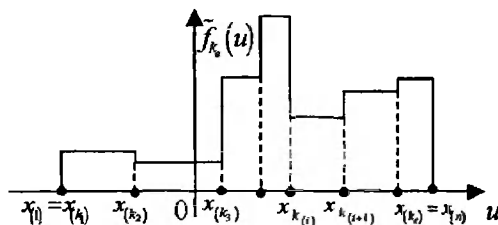


Рис. 2.6. Модифицированная гистограмма

Можно показать, что модифицированная гистограмма $\tilde{f}_k(u)$ является более предпочтительной оценкой плотности вероятностей, чем унифицированная гистограмма $f_n^*(u)$, поскольку она лучше оценивает функцию $f(u)$. Функция распределения

$$\hat{F}_k(u) = \int_{-\infty}^u \tilde{f}_k(t) dt,$$

соответствующая плотности вероятностей $\tilde{f}_k(u)$, также является более точной оценкой функции $F(u)$. Отметим, что количество l базовых статистик при использовании модифицированной гистограммы $\tilde{f}_k(u)$ следует выбирать так, чтобы оно возрастало при увеличении объема выборки n : $l = l_n = \varphi(n)$, причем $l_n \rightarrow \infty$ при $n \rightarrow \infty$. Однако величина l_n должна стремиться к бесконечности медленнее, чем число n , т.е. должно выполняться условие $l_n/n \rightarrow 0$ при $n \rightarrow \infty$.

При каком объеме n выборки $x_1, x_2, \dots, x_n$ выборочные функции распределения $F_n^*(u)$ и плотности вероятностей $f_n^*(u)$, а также их модифицированные аналоги будут удовлетворительно аппроксимировать базовые функцию распределения и плотность вероятности соответственно? Начнем обсуждение этой проблемы с модифицированной гистограммы $\tilde{f}_k(u)$. Если предположить, что плотность вероятности $f(u)$ базового распределения генеральной совокупности удовлетворительно определяется своими значениями в 10 точках, то мы должны выбрать 11 базовых порядковых статистик $x_{(k_1)} \leq x_{(k_2)} \leq \dots \leq x_{(k_l)}$, $l = 11$. Согласно практическим рекомендациям, для надежной оценки $f(u)$ в интервале $(x_{(k_l)}, x_{(k_{l+1})})$ должно находиться не менее 30 порядковых статистик. Следовательно, для надежной оценки функции $f(u)$ необходимо не менее чем 300 выборочных значений. Для надежной оценки функции распределения $F(u)$ на основании выборочной или модифицированной функций распределения $F_n^*(u)$ и $\tilde{F}_n(u)$ объем выборки $x_1, x_2, \dots, x_n$ может быть несколько меньшим.

В заключение отметим, что при всех $t \in R^1$ справедливо неравенство

$$|F_n^*(u) - \tilde{F}_n(u)| < \frac{1}{n},$$

так что при больших значениях n эмпирическая функция распределения и модифицированная функция распределения мало отличаются.

Резюме

Пусть $x_1, x_2, \dots, x_n$ — выборка, полученная путем простого случайного выбора, и $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ — вариационный ряд, построенный по этой выборке. Для оценки функции распределения $F(u)$ используются две статистики:

- эмпирическая функция распределения

$$F_n^*(u) = \begin{cases} 0, & \text{если } u \leq x_{(1)}, \\ \frac{i}{n}, & \text{если } x_{(i)} < u \leq x_{(i+1)}, \\ 1, & \text{если } u > x_{(n)}; \end{cases}$$

- модифицированная эмпирическая функция распределения

$$\tilde{F}_n(u) = \begin{cases} 0, & \text{если } u \leq x_{(1)}, \\ \frac{u + i(x_{(i+1)} - x_{(i)}) - x_{(i+1)}}{(n-1)(x_{(i+1)} - x_{(i)})}, & \text{если } x_{(i)} < u \leq x_{(i+1)}, \\ 1, & \text{если } u > x_{(n)}. \end{cases}$$

Оценка плотности вероятности $f(u)$ на основании выборочных значений $x_1, x_2, \dots, x_n$ проводится с помощью так называемой “унифицированной гистограммы”

$$\tilde{f}_n(u) = \begin{cases} 0, & \text{если } u \leq x_{(1)}, \\ \frac{1}{n-1} \frac{1}{x_{(i+1)} - x_{(i)}}, & \text{если } x_{(i)} < u \leq x_{(i+1)}, \\ 0, & \text{если } u > x_{(n)}. \end{cases}$$

Эта функция представляет собой производную от модифицированной эмпирической функции распределения $\tilde{F}_n(u)$.

2.9. Доверительное оценивание. Доверительные интервалы для основной распределенной массы генеральной совокупности

Во многих задачах биологии и медицины возникает необходимость определить интервал $J = (a, b)$, содержащий значения x генеральной совокупности G с заданной вероятностью β (например, $\beta = 0,95$):

$$p(x \in (a, b)) = \beta.$$

Такой интервал $J = (a, b)$ называется *доверительным интервалом* для значений генеральной совокупности G с *доверительным уровнем* β . Числа a и b называются *нижней и верхней доверительными границами* соответственно, значение $\alpha = 1 - \beta$ — *уровнем значимости* интервала J , а интервал (a, b) — *доверительным интервалом для основной массы значений генеральной совокупности G с уровнем значимости α* . В биологии и медицине наиболее распространенным является 5% уровень значимости, т.е. $\alpha = 0,05$.

Концы интервала $J = (a, b)$, т.е. доверительные границы a и b , могут быть как детерминированными числами, так и случайными величинами. Рассмотрим сначала относительно простой случай, когда границы a и b являются детерминированными, и опишем способ их вычисления. Естественно, что для определения чисел a и b необходима дополнительная информация о генеральной совокупности G (т.е. о случайной величине x), так как без этого ничего определенного о границах a и b сказать нельзя (кроме того, что $a = -\infty$, $b = \infty$, т.е. $\alpha = 0$). Очевидно, что такое тривиальное решение никого не устраивает, поэтому предположим, что нам известны наиболее важные показатели случайной величины x — математическое ожидание $m(x)$ и дисперсия $\sigma^2(x)$. Доверительные границы a и b можно получить с помощью неравенства Чебышева (см. подраздел 1.4.3), но на практике они практически не используются, так как полученный доверительный интервал слишком широк.

2.9.1. Правило 3σ

В течение долгого времени (свыше 150 лет) в математической статистике не было теоретически обоснованного и доказанного метода построения доверительных границ a и b для основной распределенной массы случайной величины, хотя в практических статистических вычислениях широко использовалось правило 3σ , утверждающее, что более чем 95% основной распределенной массы любой случайной величины содержится в интервале $(m(x) - 3\sigma, m(x) + 3\sigma)$. В настоящее время правило 3σ полностью обосновано лишь для одномодальной случайной величины (см. подраздел 1.4.3).

Заметим, что одномодальные распределения характерны для однородных (гомогенных) генеральных совокупностей, состоящих из индивидуумов одного вида, в то время как полимодальные распределения (т.е. распределения, у которых плотность вероятности имеет несколько точек локального максимума) характерны для неоднородных (гетерогенных) генеральных совокупностей, возникающих при смешивании однородных генеральных совокупностей.

На практике точные значения математического ожидания $m(x)$ и дисперсии $\sigma^2(x)$ известны лишь в редких случаях, поэтому эти показатели заменяют их оценками $m(x) \approx \bar{x}$, $\sigma^2(x) \approx s^2(x)$, описанными в разделах 2.2 и 2.3. Замена точных показателей $m(x)$ и $\sigma^2(x)$ их приближенными значениями приводит к правилу $3s$, утверждающему, что для генеральных совокупностей G , встречающихся на практике, доверительный интервал

$$\hat{J} = (\bar{x} - 3s, \bar{x} + 3s), \quad (2.28)$$

где $\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k$, $s^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2$, при больших значениях n содержит не менее чем 95% значений генеральной совокупности G .

Возникает вопрос: при каких значениях n выполняется правило $3s$? Согласно практическим рекомендациям, оценка $\bar{x}$ практически совпадает с $m(x)$, если объем выборки $n \geq 30$, а оценка s^2 близка к $D(x) = \sigma^2$ при $n \geq 150$. Однако расчеты, проведенные с помощью машинного моделирования, показывают, что интервал $\hat{J}$ содержит не менее чем 95% значений генеральной совокупности G , если объем выборки $n \geq 11$.

2.9.2. Доверительные интервалы для параметров, построенные по правилу $3s$

Во многих задачах биологии и медицины доверительные интервалы $J = (a, b)$ необходимо вычислять не только для основной распределенной массы значений генеральной совокупности G , но и для ее основных параметров — математического ожидания $m(x)$, дисперсии $\sigma^2(x)$ и т.д. Нетрудно заметить, что такие доверительные интервалы не могут быть детерминированными, т.е. их концы должны быть случайными величинами. В самом деле, если предположить, что

доверительный интервал $J = (a, b)$ для какого-то параметра θ (например, для математического ожидания $\theta = m(x)$) является детерминированным, то вероятность события $p(m(x) \in J)$ равна или нулю, если $m(x) \notin J$, или единице, если $m(x) \in J$. Следовательно, понятие уровня значимости для такого интервала совершенно лишено смысла, поскольку величины $m(x)$, a и b не связаны со случайными испытаниями. Если же величины a и b являются случайными величинами, то после одного случайного испытания интервал $J = (a, b)$ может содержать параметр $m(x)$, а при проведении другого испытания $m(x)$ может не принадлежать интервалу J . Таким образом, событие $m(x) \in J$ имеет определенную вероятность, а интервал J имеет доверительный уровень β .

С правилом 3 σ тесно связано правило 3 s , которое по выборке $x_1, x_2, \dots, x_n$, полученной путем простого случайного выбора, позволяет построить доверительные границы для неизвестного математического ожидания $m(x)$ с уровнем значимости, не превышающим 0,05. Как было указано, неравенство (2.8) справедливо для подавляющего большинства случайных величин x , встречающихся на практике, поэтому можно положить $x = \bar{x}$. В этом случае справедливы следующие формулы:

$$m(\bar{x}) = m\left(\frac{1}{n} \sum_{k=1}^n x_k\right) = \frac{1}{n} m\left(\sum_{k=1}^n x_k\right) = \frac{1}{n} \sum_{k=1}^n m(x_k) = m(x),$$

$$\sigma^2(x) = D(\bar{x}) = D\left(\frac{1}{n} \sum_{k=1}^n x_k\right) = \frac{1}{n^2} \sum_{k=1}^n D(x_k) = \frac{\sigma^2}{n}.$$

Следовательно,

$$\begin{aligned} \sigma^2(\bar{x}) &= \frac{\sigma^2}{n} \approx \frac{s^2}{n} \text{ и } \sigma(\bar{x}) \approx \frac{s}{\sqrt{n}}, \\ p(|\bar{x} - m| \leq 3s(\bar{x})) &\approx p(|\bar{x} - m| \leq 3\sigma(\bar{x})) = \\ &= p(-3\sigma(\bar{x}) \leq m - \bar{x} \leq 3\sigma(\bar{x})) = \\ &= p(\bar{x} - 3\sigma(\bar{x}) \leq m \leq \bar{x} + 3\sigma(\bar{x})) \approx \\ &\approx p\left(\bar{x} - 3\frac{s}{\sqrt{n}} \leq m \leq \bar{x} + 3\frac{s}{\sqrt{n}}\right) \geq 0,95. \end{aligned}$$

Поэтому интервал $J = \left(\bar{x} - 3\frac{s}{\sqrt{n}}, \bar{x} + 3\frac{s}{\sqrt{n}}\right)$ является случайным доверительным интервалом для неизвестного математического ожидания $m(x)$ генеральной совокупности G , доверительный уровень которого превосходит 0,95, а уровень значимости — не превосходит 0,05. Этот факт называется правилом 3 s_1 (от $s_1 = \frac{s}{\sqrt{n}}$).

Правило 3 s_1 напоминает хорошо известный критерий Стьюдента, который также позволяет вычислить доверительные границы для неизвестного матема-

тического ожидания нормальной генеральной совокупности. Напомним, суть этого критерия. Пусть $x_1, x_2, \dots, x_n$ — выборка, полученная путем простого случайного выбора из нормально распределенной генеральной совокупности G . Можно показать, что при этих предположениях случайная величина

$$t_1 = \frac{m(x) - \bar{x}}{s_1}$$

имеет так называемое распределение Стьюдента, поэтому для любого доверительного уровня β (например, $\beta = 0,95$) можно определить доверительную границу t_β , для которой выполняется равенство

$$P(|t| \leq t_\beta) = P\left(\frac{|m(x) - \bar{x}|}{s_1} \leq t_\beta\right) = \beta.$$

Доверительные границы t_β зависят от объема n выборки $x_1, x_2, \dots, x_n$. Однако при $\beta = 0,95$ для $n = 5, 6, 7, 8, 9, 10$ величина t_β изменяется в пределах от 2,571 до 2,228, а для $n \geq 11$ — от 1,96 до 2,2. Следовательно, доверительный интервал $(\bar{x} - t_\beta s_1, \bar{x} + t_\beta s_1)$ является доверительным интервалом для неизвестного математического ожидания с уровнем значимости $\alpha = 1 - \beta$. При $5 < n < 10$ интервал $(\bar{x} - t_\beta s_1, \bar{x} + t_\beta s_1)$ незначительно отличается от интервала $(\bar{x} - 3s_1, \bar{x} + 3s_1)$, построенного по правилу $3s_1$, и лишь при больших значениях n он имеет вид $(\bar{x} - 2s_1, \bar{x} + 2s_1)$, а его длина становится примерно в полтора раза меньшей, чем длина интервала, вычисленного по правилу $3s_1$.

Доверительный интервал Стьюдента был изобретен английским статистиком У. Госсетом, опубликовавшим свою работу в 1908 году под скромным псевдонимом “студент”. По мнению ведущих специалистов, эта работа произвела переворот в математической статистике. Не преуменьшая заслуги У. Госсета, отметим, что нормальные генеральные совокупности практически никогда не встречаются на практике. Известный статистик Ж.-Р. Барра писал: “Нормальность распределения — это миф. Его никогда не было и никогда не будет в реальной действительности”. В связи с этим критерием Стьюдента следует пользоваться осторожно, когда есть веские аргументы в пользу наличия нормального распределения у исследуемой генеральной совокупности. В то же время правило $3s_1$ не требует выполнения гипотезы о нормальности распределения генеральной совокупности G .

Другой аргумент в пользу правила $3s_1$ следует из центральной предельной теоремы, впервые сформулированной и доказанной выдающимся русским математиком А.А. Марковым. Согласно этой теореме, функция распределения случайной величины $\bar{x}$ при увеличении объема выборки n приближается к функции нормального распределения, поскольку для такой случайной величины справедливо правило $2s_1$:

$$P(|x - m(x)| \leq 2\sigma(x)) \geq 0,95,$$

а для случайной величины x выполняется правило $3s_1$:

$$p(|\bar{x} - m(x)| \leq 3\sigma(x)) \geq p(|\bar{x} - m(x)| \geq 2\sigma(x)) \geq 0,95.$$

Следовательно, при больших значениях n выполняется неравенство

$$\begin{aligned} p(|\bar{x} - m(x)| \leq 3\sigma(x)) &\approx p(|\bar{x} - m(x)| \leq 3s_1(x)) = \\ &= p\left(|\bar{x} - m(x)| \leq 3 \frac{s(x)}{\sqrt{n}}\right) = p\left(\bar{x} - \frac{s(x)}{\sqrt{n}} \leq m(x) \leq \bar{x} + \frac{s(x)}{\sqrt{n}}\right). \end{aligned}$$

Перейдем к проблеме определения доверительных границ неизвестных коэффициентов линейной регрессии:

$$y(t) = a_0 + a_1 t + \dots + a_r t^r + x(t), \quad t = t_1, t_2, \dots, t_n, \dots$$

Опишем вначале решение этой задачи при условии, что $x(t)$ является последовательностью независимых нормально распределенных случайных величин с одинаковыми дисперсиями $D[\bar{x}(t_i)] = \sigma^2$ ($i = 1, 2, \dots, n, \dots$).

Рассмотрим матрицу, составленную из коэффициентов нормальных уравнений

$$A = \begin{pmatrix} n & [t] & [t^2] & \dots & [t^r] \\ [t] & [t^2] & [t^3] & \dots & [t^{r+1}] \\ [t^2] & [t^3] & [t^4] & \dots & [t^{r+2}] \\ \dots & \dots & \dots & \dots & \dots \\ [t^r] & [t^{r+1}] & [t^{r+2}] & \dots & [t^{2r}] \end{pmatrix},$$

где

$$[t^k] = \sum_{i=1}^n t_i^k, \quad k = 1, 2, \dots, 2r.$$

Обозначим через A^{-1} матрицу, обратную матрице A . Нас интересуют лишь диагональные элементы этой матрицы, которые мы обозначим символами $h_{11}, h_{22}, \dots, h_{rr}$ (порядок матрицы A^{-1} совпадает с порядком матрицы A и равен $(r+1) \times (r+1)$):

$$A^{-1} = \begin{pmatrix} h_{11} & & & & \\ & h_{22} & & & \\ & & h_{33} & & \\ & & & \dots & \\ & & & & h_{rr} \\ & & & & & h_{r+1, r+1} \end{pmatrix}$$

Напомним, что оценки $a_0^*, a_1^*, \dots, a_r^*$ коэффициентов линейной регрессии получаются в результате минимизации суммы квадратов

$$Q(a_0, a_1, \dots, a_r) = \sum_{i=1}^n [y(t_i) - m(t_i; a_0, a_1, \dots, a_r)],$$

где $m(t; a_0, a_1, \dots, a_r) = a_0 + a_1 t + \dots + a_r t^r$.

Введем следующие обозначения:

$$\bar{Q} = Q(a_0^*, a_1^*, \dots, a_r^*) = \sum_{i=1}^n [y(t_i) - a_0^* - a_1^* t_i - \dots - a_r^* t_i^r]^2,$$

$$s^2 = \frac{\bar{Q}}{n - (r + 1)}, s_0^2 = h_{11} s^2, s_1^2 = h_{22} s^2, \dots, s_r^2 = h_{r+1, r+1} s^2.$$

Можно показать, что отношение

$$\tau_i = \frac{a_i^* - a_i}{s_i}, \quad i = 0, 1, \dots, r$$

подчиняется распределению Стьюдента с $n - (r + 1)$ степенями свободы. Это позволяет определить доверительные границы неизвестного параметра a_i с заданным доверительным уровнем. По таблице распределения Стьюдента можно найти значение t_β , для которого выполняется равенство

$$P(|\tau_i| \leq t_\beta) = \beta.$$

Значение t_β называется двусторонней доверительной границей, соответствующей уровню значимости $\alpha = 1 - \beta$ (именно такие границы приведены в таблице распределения Стьюдента). Отсюда следует, что

$$P(|\tau_i| \leq t_\beta) = \beta = 1 - \alpha$$

и

$$\begin{aligned} P(|\tau_i| \leq t_\beta) &= P\left(\left|\frac{a_i^* - a_i}{s_i}\right| \leq t_\beta\right) = P(|a_i^* - a_i| \leq t_\beta s_i) = \\ &= P(-t_\beta s_i \leq a_i^* - a_i \leq t_\beta s_i) = P(a_i^* - t_\beta s_i \leq a_i \leq a_i^* + t_\beta s_i) = 1 - \alpha = \beta. \end{aligned}$$

Таким образом, числа $a_i^* - t_\beta s_i$ и $a_i^* + t_\beta s_i$ являются нижней и верхней границами параметра a_i , соответствующими доверительному уровню β .

Если гипотеза о нормальности распределения выборочных значений $y(t_i)$, $i = 1, 2, \dots, n$, не выполняется (как это бывает в большинстве случаев в реальности, а не в теоретических моделях), то в качестве величины t_β в формуле для доверительных границ следует положить число 3. Эти доверительные границы, построенные по правилу $3s_i$, являются более надежными, а обоснование правила $3s_i$ совершенно аналогично обоснованию правила $3s_1$.

Для линейной регрессии первого порядка

$$m(t; a_0, a_1) = a_0 + a_1 t$$

описанная процедура вычисления доверительных границ представляет собой несложный алгоритм. Здесь матрица A имеет вид

$$A = \begin{pmatrix} n & \sum_{i=1}^n t_i \\ \sum_{i=1}^n t_i & \sum_{i=1}^n t_i^2 \end{pmatrix},$$

а диагональные элементы h_{11} и h_{22} соответственно равны

$$h_{11} = \frac{\sum_{i=1}^n t_i^2}{n \sum_{i=1}^n t_i^2 - \left(\sum_{i=1}^n t_i \right)^2},$$

$$h_{22} = \frac{n}{n \sum_{i=1}^n t_i^2 - \left(\sum_{i=1}^n t_i \right)^2}.$$

Кроме того,

$$\bar{Q} = \sum_{i=1}^n [y(t_i) - a_0^* - a_1^* t_i]^2, \quad s^2 = \frac{\bar{Q}}{n-2}, \quad s_0^2 = h_{11} s^2, \quad s_1^2 = h_{22} s^2.$$

Для зависимых наблюдений $y(t_1), y(t_2), \dots, y(t_n), \dots$ 95%-ные доверительные границы коэффициентов линейной регрессии можно определить приближенно (см. раздел 2.6). Применяя формулу Айткена (см. раздел 2.6), можно вычислить вектор $U^* = (u_0^*, u_1^*, \dots, u_r^*)$, составленный из наилучших оценок u_i^* , $i = 0, 1, \dots, r$, неизвестных коэффициентов линейной регрессии a_i , $i = 0, 1, \dots, r$. Как известно, матрица $\bar{K}$ ковариаций координат вектора u^* имеет вид

$$\bar{K} = [\Phi^T K^{-1} \Phi]^{-1}.$$

На главной диагонали матрицы $\bar{K}$ расположены значения дисперсии $D(u_i) = \sigma_{ii}^2$ случайных величин u_i^* . Затем, применяя правило 3σ , мы получим приближенные значения 95%-ных доверительных границ неизвестных параметров a_i :

$$u_i^* - 3\sigma_{ii}, u_i^* + 3\sigma_{ii}.$$

Рассмотрим теперь проблему построения доверительных границ для неизвестной дисперсии, коэффициентов ковариации и корреляции. Как показано в разделе 2.3, если математическое ожидание m известно, то существуют две оценки неизвестной дисперсии σ^2 :

$$s_0^2 = \frac{1}{n} \sum_{i=1}^n x_i - m^2$$

и

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - m)^2.$$

Как было показано, эти две оценки являются несмещенными: $m(s_0^2) = m(s^2) = \sigma^2$. Нетрудно вычислить значения дисперсии $\sigma^2(s_0^2)$ и $\sigma^2(s^2)$ этих оценок:

$$\sigma^2(s_0^2) = \frac{n-1}{n}L_2^2 + \frac{L_4}{n} - 2m^2L_2 + (L_2 - m^2),$$

$$\sigma^2(s^2) = \frac{n-1}{n}L_2^2 + \frac{L_4}{n} - 2m^2L_2 + m^4 + \frac{4}{n}(2mL_2 + L_2 - m)(L_2 - m^2),$$

где $L_k = m(x^k)$ — момент k -го порядка случайной величины x ($k = 1, 2, 3, 4$). Замена неизвестные моменты в формулах для дисперсии их оценками $L_k = \frac{1}{n} \sum_{i=1}^n x_i^k$ и обозначая оценки дисперсии символами $s^2(s_0^2)$ и $s^2(s^2)$ соответственно, получим два доверительных интервала для неизвестной дисперсии σ^2 , построенных по правилу $3s$, с уровнем значимости, не превосходящим 0,05:

$$J_1 = (-3s(s_0^2) + s_0^2, 3s(s_0^2) + s_0^2),$$

$$J_2 = (-3s(s^2) + s^2, 3s(s^2) + s^2).$$

Заметим, что величины s_0^2 и s^2 имеют одномодальное распределение, например, если величина x имеет нормальное распределение.

Поскольку доверительные интервалы J_1 и J_2 являются приближенными, их уровень значимости можно подтвердить лишь с помощью компьютерного моделирования. Если математическое ожидание $m(x)$ неизвестно, то его можно заменить оценкой $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$. Тогда приближенные формулы для доверительных границ неизвестной дисперсии имеют вид

$$(\bar{s} - 3\bar{\sigma}(\bar{s}^2), \bar{s} + 3\bar{\sigma}(\bar{s}^2)),$$

где

$$\bar{s}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2,$$

$$\bar{\sigma}(\bar{s}^2) = \sqrt{\frac{n-1}{n} \left(\sum_{i=1}^n x_i^2 \right)^2 + \frac{1}{n} \sum_{i=1}^n x_i^4 - 2\bar{x}^2 \sum_{i=1}^n x_i^2 + \bar{x}^4 - \left(\sum_{i=1}^n x_i^2 - \bar{x}^2 \right)}.$$

Как было указано в разделе 2.4, для оценки коэффициента ковариации $K(x, y) = m\{[x - m(x)][y - m(y)]\}$ случайных величин x и y необходимо провести случайные испытания (простой случайный выбор) и зарегистрировать пары случайных величин (x_k, y_k) , $k = 1, 2, \dots, n$. В этом случае несмещенная оценка $K^*(x, y)$ коэффициента ковариации имеет вид

$$K^*(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}),$$

где $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$. Эта оценка является наилучшей в классе билинейных несмещенных оценок.

Полагая $\gamma_i = (x_i - \bar{x})(y_i - \bar{y})$, получим, что

$$K^*(x, y) = \bar{\gamma} = \frac{1}{n-1} \sum_{i=1}^n \gamma_i.$$

Оценка $K^*(x, y)$ представляет собой величину, которая мало отличается от выборочного среднего, вычисленного по выборке $\gamma_1, \gamma_2, \dots, \gamma_n$. Однако случайные величины $\gamma_1, \gamma_2, \dots, \gamma_n$ будут зависимыми, поскольку в каждое слагаемое $\gamma_i = (x_i - \bar{x})(y_i - \bar{y})$ входят одни и те же случайные величины $\bar{x}$ и $\bar{y}$. Доверительный интервал для выборочного среднего, построенный по правилу $3s_1$, имеет вид

$$J = \left(\bar{x} - 3 \frac{s}{\sqrt{n}}, \bar{x} + 3 \frac{s}{\sqrt{n}} \right).$$

Если положить $x = \gamma$, то формула для доверительных границ коэффициентов ковариации, определенных по правилу $3s_1$, примет вид

$$J_K = \left(\bar{\gamma} - 3 \frac{s_\gamma}{\sqrt{n}}, \bar{\gamma} + 3 \frac{s_\gamma}{\sqrt{n}} \right),$$

где $s_\gamma^2 = \frac{1}{n-1} \sum_{i=1}^n (\gamma_i - \bar{\gamma})^2$, $\bar{\gamma} = \frac{1}{n-1} \sum_{i=1}^n \gamma_i$, а $\gamma_i = (x_i - \bar{x})(y_i - \bar{y})$, $i = 1, 2, \dots, n$.

Отметим, что указанные выше особенности случайных величин γ_i не могут оказать заметного влияния на уровень значимости доверительного интервала. Во-первых, отличие $\bar{\gamma}$ от выборочного среднего крайне незначительно, и, во-вторых, наличие зависимости между случайными величинами γ_i , $i = 1, 2, \dots, n$, не может существенно повлиять на величину s_γ^2 , так как можно показать, что при выполнении достаточно общих условий она является хорошей оценкой дисперсии $\sigma^2(\gamma)$.

Коэффициент корреляции

$$r(x, y) = \frac{K(x, y)}{\sigma(x)\sigma(y)}$$

отличается от коэффициента ковариации $K(x, y)$ лишь множителем $\lambda = \frac{1}{\sigma(x)\sigma(y)} > 0$. В связи с этим доверительные границы для коэффициента

корреляции $r(x, y)$, полученные на основании правила $3s_1$, можно вычислить с помощью следующих простых аналитических преобразований:

$$\begin{aligned} p \left(K(x, y) \in \left(\bar{\gamma} - \frac{3s_\gamma}{\sqrt{n}}, \bar{\gamma} + \frac{3s_\gamma}{\sqrt{n}} \right) \right) &= p \left(\lambda K(x, y) \in \left(\lambda \bar{\gamma} - \frac{3\lambda s_\gamma}{\sqrt{n}}, \lambda \bar{\gamma} + \frac{3\lambda s_\gamma}{\sqrt{n}} \right) \right) = \\ &= p \left(r(x, y) \in \left(\frac{\bar{\gamma}}{\sigma(x)\sigma(y)} - \frac{3}{\sqrt{n}} \frac{s_\gamma}{\sigma(x)\sigma(y)}, \frac{\bar{\gamma}}{\sigma(x)\sigma(y)} + \frac{3}{\sqrt{n}} \frac{s_\gamma}{\sigma(x)\sigma(y)} \right) \right) \geq 0,95. \end{aligned}$$

Если заменить $\sigma(x)$ и $\sigma(y)$ их оценками $s(x)$ и $s(y)$, то мы получим доверительный интервал для неизвестного коэффициента корреляции $r(x, y)$:

$$J_r = \left(\frac{\bar{y}}{s(x)s(y)} - \frac{3}{\sqrt{n}} \frac{s_y}{s(x)s(y)}, \frac{\bar{y}}{s(x)s(y)} + \frac{3}{\sqrt{n}} \frac{s_y}{s(x)s(y)} \right).$$

Для доверительных границ коэффициентов ковариации и корреляции можно получить более точные оценки [42], если предположить, что пара случайных величин имеет совместное нормальное распределение. Однако на практике эти границы применять вряд ли целесообразно, поскольку нормальное распределение редко встречается на практике.

2.9.3. Доверительные интервалы, порожденные порядковыми статистиками

Перейдем теперь к описанию другого метода построения доверительных интервалов для основной распределенной массы значений генеральной совокупности, основанного на понятии порядковой статистики.

Пусть — выборка, полученная путем простого случайного выбора T из генеральной совокупности G . Расставим выборочные значения $x_1, x_2, \dots, x_n$ в порядке их возрастания:

$$x_{k_1} \leq x_{k_2} \leq \dots \leq x_{k_n},$$

где $k_i = 1, 2, \dots, n$; $i = 1, 2, \dots, n$. Введем обозначения $x_{(1)} = x_{k_1}, x_{(2)} = x_{k_2}, \dots, x_{(n)} = x_{k_n}$ и перепишем цепочку неравенств в следующем виде:

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}.$$

Случайная величина $x_{(i)}$ называется i -й порядковой статистикой, а ряд чисел $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ — вариационным рядом. Число $x_{(1)}$ является минимальной, а число $x_{(n)}$ — максимальной порядковой статистиками. Обе эти статистики называются *крайними*.

В отличие от выборочных значений порядковые статистики $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ являются зависимыми случайными величинами и имеют разные функции распределения. Нетрудно показать [42], что функция распределения $F_k(u)$ k -й порядковой статистики $x_{(k)}$ имеет вид

$$F_k(u) = \sum_{i=k}^n C_n^i [F(u)]^i [1 - F(u)]^{n-i},$$

где $F(u) = p(x \leq u)$.

Будем называть случайную величину x непрерывной, если ее функция распределения непрерывна на числовой прямой. Используя эту формулу, можно показать, что, если выборочные значения $x_1, x_2, \dots, x_n$ непрерывной случайной величины x из генеральной совокупности G являются независимыми, то вероятность попадания следующего выборочного значения x_{n+1} в доверительный ин-

тервал $J = (x_{(i)}, x_{(j)})$, порожденный i -й и j -й порядковыми статистиками ($i < j$), не зависит от распределения $F(u)$ и равна

$$p\left(x_{n+1} \in (x_{(i)}, x_{(j)})\right) = \frac{j-i}{n+1}. \quad (2.9)$$

Эту формулу можно обобщить на случай симметрично зависимых случайных величин $x_1, x_2, \dots, x_n$, у которых совместная функция распределения $F(u_1, u_2, \dots, u_n) = p(x_1 < u_1, x_2 < u_2, \dots, x_n < u_n)$ не изменяет своего значения при любой перестановке ее аргументов

$$F(u_1, u_2, \dots, u_n) = F(u_{i_1}, u_{i_2}, \dots, u_{i_n})$$

и является абсолютно непрерывной, так что $p(x_i = x_k) = 0$, когда $i \neq k$.

Таким образом, если значение x не зависит от выборки $x_1, x_2, \dots, x_n$, полученной в результате простого случайного выбора T из генеральной совокупности G с непрерывным распределением $F(u)$, то

$$p\left(x \in (x_{(i)}, x_{(j)})\right) = \frac{j-i}{n+1}.$$

В частности,

$$p\left(x \in (x_{(1)}, x_{(n)})\right) = \frac{n-1}{n+1}.$$

Эти формулы дают возможность построить доверительный интервал для множества значений генеральной совокупности G с точно установленным уровнем значимости. Действительно, в силу формулы (2.9) доверительный уровень доверительного интервала $J_y = (x_{(i)}, x_{(j)})$, $i < j$, построенного с помощью порядковых статистик, равен $(j-i)/(n+1)$. Если $n \geq 39$, то уровень значимости доверительного интервала, порожденного крайними порядковыми статистиками $x_{(1)}$ и $x_{(n)}$, не превосходит 0,05, так как в этом случае $(n-1)/(n+1) \geq 0,95$. Таким образом, для многих средних и больших выборок случайный интервал $(x_{(1)}, x_{(n)})$ является доверительным интервалом для основной распределенной массы значений генеральной совокупности G с уровнем значимости $(n-1)/(n+1)$.

При исследовании свойств доверительного интервала J_p , порожденного порядковыми статистиками $x_{(i)}$ и $x_{(j)}$, следует оценить его надежность и точность. Пусть $a(u_1, u_2, \dots, u_n)$ и $b(u_1, u_2, \dots, u_n)$ — две произвольные функции, причем удовлетворяющие при всех $u_1, u_2, \dots, u_n$ неравенству $a(u_1, u_2, \dots, u_n) \leq b(u_1, u_2, \dots, u_n)$. Используя выборку $x_1, x_2, \dots, x_n$, можно построить случайный доверительный интервал $J = (a(x_1, x_2, \dots, x_n), b(x_1, x_2, \dots, x_n))$ для множества значений G . Пусть $m(a)$ и $m(b)$ — математические ожидания случайных величин $a(x_1, x_2, \dots, x_n)$ и $b(x_1, x_2, \dots, x_n)$ соответственно. Надежностью $\alpha(a, b)$ случайного доверительного интервала J называется его доверительный уровень

$$\alpha(a, b) = p(x \in (a, b)),$$

а его точностью $\beta(a, b)$ — величина

$$\beta(a, b) = \frac{1}{m(a) - m(b)}.$$

Точность $\beta(a, b)$ доверительного интервала характеризует его среднюю длину: чем более узким (в среднем) является доверительный интервал $J = (a, b)$, тем больше его точность. В теории проверки гипотез с помощью статистических критериев особый интерес представляют наиболее точные интервалы, выбранные из класса интервалов с одинаковой надежностью. Легко видеть, что два разных доверительных интервала $J_j = (x_{(j)}, x_{(0)})$ и $J_k = (x_{(k)}, x_{(0)})$, построенные с помощью порядковых статистик, имеют одинаковую надежность тогда и только тогда, когда $j - i = l - k$. Обозначим через $\mathfrak{A}_k$ множество всех доверительных интервалов J_j , для которых выполняется условие $j - i = n - 2k + 1 = \text{const}$, где $k \leq [n/2]$.

Возникает вопрос, какой интервал J_j из класса $\mathfrak{A}_k$ является наиболее точным? В работе [84] показано, что для симметричных одномодальных случайных величин в классе $\mathfrak{A}_k$ наиболее точным является доверительный интервал $J_{k, n-k+1} = (x_{(k)}, x_{(n-k+1)})$.

Рассмотрим две произвольные функции $a(u_1, u_2, \dots, u_n)$ и $b(u_1, u_2, \dots, u_n)$, удовлетворяющие неравенству $a(u_1, u_2, \dots, u_n) \leq b(u_1, u_2, \dots, u_n)$ при всех $u_1, u_2, \dots, u_n$. Как показано выше, используя выборку $x_1, x_2, \dots, x_n$, построим случайный доверительный интервал $J = (a(x_1, x_2, \dots, x_n), b(x_1, x_2, \dots, x_n))$ для множества значений генеральной совокупности G . Пусть $\mathfrak{F}$ — некоторый класс базовых распределений (например, класс всех непрерывных распределений $\mathfrak{F}_C$). Доверительный интервал $J = (a(x_1, x_2, \dots, x_n), b(x_1, x_2, \dots, x_n))$ из класса $\mathfrak{F}$ называется *инвариантным*, если для любой генеральной совокупности G с базовым распределением $F_G(u) \in \mathfrak{F}$ выполняется условие

$$p_G(x_{n+1} \in J) = p_G((a(x_1, x_2, \dots, x_n), b(x_1, x_2, \dots, x_n))) = \text{const},$$

где p_G — вероятность, определяемая базовым распределением $F_G(u)$.

Используя это определение, на основании формулы (2.9) можно утверждать, что доверительный интервал $J_j = (x_{(j)}, x_{(0)})$, порожденный порядковыми статистиками, является инвариантным доверительным интервалом в классе $\mathfrak{F}_C$, состоящим из всех непрерывных функций распределения. В связи с этим возникает естественный вопрос: существуют ли в классе $\mathfrak{F}_C$ еще какие-нибудь инвариантные доверительные интервалы, кроме J_j ? Ответ на этот вопрос отрицателен. Действительно, в работе [30] доказана теорема о единственности инвариантных доверительных интервалов, порожденных порядковыми статистиками. Кроме того, эта теорема позволяет описать множество всех доверительных уровней инвариантных доверительных интервалов: при фиксированном объеме выборки n множество B_n всех доверительных уровней таких интервалов имеет вид

$$B_n = \left\{ 0, \frac{1}{n+1}, \frac{2}{n+1}, \dots, 1 \right\}.$$

Объединение множеств B_n при всевозможных значениях n представляет собой все рациональные числа из отрезка $[0, 1]$.

Резюме

Доверительным интервалом, содержащим основную распределенную массу значений генеральной совокупности G , с доверительным уровнем β , называется интервал (a, b) , для которого выполняется условие

$$p(x \in (a, b)) = \beta, \quad x \in G.$$

Существуют два способа построения таких интервалов: 1) на основе математического ожидания $m(x)$ и дисперсии $\sigma^2(x)$ генеральной совокупности G и 2) с помощью выборочных значений $x_1, x_2, \dots, x_n$.

Первый метод построения доверительных интервалов основан на правиле 3σ . В силу неравенства Гаусса–Высочанского–Петунина уровень значимости интервала $(m(x) - \lambda\sigma(x), m(x) + \lambda\sigma(x))$ при $\lambda > \sqrt{\frac{8}{3}}$ (в частности, при $\lambda = 3$) не пре-

восходит $\frac{4}{9\lambda^2}$. Таким образом, интервал $(m(x) - 3\sigma(x), m(x) + 3\sigma(x))$, построен-

ный по правилу 3σ , имеет уровень значимости, не превосходящий 0,05. Для одномодальных распределений это утверждение является строго доказанным. Если математическое ожидание $m(x)$ и дисперсия $\sigma^2(x)$ генеральной совокупности G не известны, то их можно заменить оценками $\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k$ и

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

В этом случае интервал $(\bar{x} - 3s, \bar{x} + 3s)$, построенный по

правилу $3s$, тоже можно считать доверительным интервалом, содержащим основную распределенную массу значений генеральной совокупности G с уровнем значимости, не превосходящим 0,05.

Второй метод построения доверительного интервала для основной распределенной массе значений генеральной совокупности G основан на порядковых статистиках. Пусть $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ — вариационный ряд, порожденный выборкой $x_1, x_2, \dots, x_n$, полученной путем простого случайного выбора из генеральной совокупности G с непрерывной функцией распределения $F(u)$, а $x_{(i)}, x_{(j)}$, $i < j$ — порядковые статистики. Тогда

$$p\left(x \in (x_{(i)}, x_{(j)})\right) = \frac{j-i}{n+1},$$

где x — выборочное значение, не зависящее от $x_1, x_2, \dots, x_n$.

Следовательно, точное значение уровня значимости доверительного интервала $(x_{(i)}, x_{(j)})$ равно $1 - \frac{j-i}{n+1}$, и при $i = 1$ и $j = n$ эта величина не превосходит 0,05, если $n \geq 39$. Таким образом, интервал $(x_{(1)}, x_{(n)})$ можно считать доверительным

интервалом для основной распределенной массы значений генеральной совокупности G с уровнем значимости $(n-1)/(n+1)$.

Указанные результаты позволяют определить приближенные доверительные интервалы для неизвестного математического ожидания $m(x)$ и дисперсии $\sigma^2(x)$, уровни значимости которых не превосходят 0,05.

Приближенный доверительный интервал для неизвестного математического ожидания $m(x)$, построенный по правилу $3s_1$, имеет вид $J = \left(\bar{x} - 3 \frac{s}{\sqrt{n}}, \bar{x} + 3 \frac{s}{\sqrt{n}} \right)$.

Приближенные доверительные интервалы для неизвестной дисперсии $\sigma^2(x)$, построенные по правилу $3s_1$, имеют вид $J_1 = (-3s(s_0^2) + s_0^2, 3s(s_0^2) + s_0^2)$ и $J_2 = (-3s(s^2) + s^2, 3s(s^2) + s^2)$, где

$$s_0^2 = \frac{1}{n} \sum_{i=1}^n x_i - m^2,$$

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - m)^2,$$

$$s^2(s_0^2) = \frac{n-1}{n} L_2^2 + \frac{L_4}{n} - 2\bar{x}^2 L_2 + (L_2 - \bar{x}^2),$$

$$s^2(s^2) = \frac{n-1}{n} L_2^2 + \frac{L_4}{n} - 2\bar{x}^2 L_2 + \bar{x}^4 + \frac{4}{n} (2\bar{x} L_2 + L_2 - \bar{x}) - (L_2 - \bar{x}^2),$$

$$L_k = \frac{1}{n} \sum_{i=1}^n x_i^k.$$

Приближенный доверительный интервал для коэффициента ковариации $K(x, y)$ с уровнем значимости, не превышающим 0,05, построенный по правилу $3s_1$, имеет вид

$$J_K = \left(\bar{\gamma} - 3 \frac{s_\gamma}{\sqrt{n}}, \bar{\gamma} + 3 \frac{s_\gamma}{\sqrt{n}} \right),$$

где $\bar{\gamma} = \frac{1}{n-1} \sum_{i=1}^n \gamma_i$, $\gamma_i = (x_i - \bar{x})(y_i - \bar{y})$, $s_\gamma^2 = \frac{1}{n-1} \sum_{i=1}^n (\gamma_i - \bar{\gamma})^2$.

Приближенный доверительный интервал для коэффициента корреляции $r(x, y)$ с уровнем значимости, не превышающим 0,05, построенный по правилу $3s_1$, имеет вид

$$J_r = \left(\frac{\bar{\gamma}}{s(x)s(y)} - \frac{3s_\gamma}{\sqrt{n}} \frac{s_\gamma}{s(x)s(y)}, \frac{\bar{\gamma}}{s(x)s(y)} + \frac{3}{\sqrt{n}} \frac{s_\gamma}{s(x)s(y)} \right),$$

где $s^2(x) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$, $s^2(y) = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$.

2.10. Статистические критерии для проверки гипотез

2.10.1. Основные принципы проверки гипотез. Теория Неймана—Пирсона

Основоположниками классической теории проверки гипотез являются американский математик Ежи Нейман и английский математик Эгон Пирсон¹. Их идеи описаны во всех учебниках по математической статистике. Суть построенной ими теории заключается в следующем. Предположим, что необходимо проверить две конкурирующие альтернативные гипотезы H и H' (т.е. гипотезы взаимоисключающие, причем одна из них обязательно верна). Гипотеза H утверждает, что наблюдаемая в процессе эксперимента выборка $x = (x_1, x_2, \dots, x_n)$ принадлежит генеральной совокупности с совместным распределением $F(u_1, u_2, \dots, u_n)$, а гипотеза H' утверждает, что она извлечена из генеральной совокупности с совместным распределением $G(u_1, u_2, \dots, u_n)$. Гипотеза H называется основной, а H' — конкурирующей. Построение статистического критерия для проверки истинности гипотез H и H' осуществляется следующим образом. Рассмотрим n -мерное евклидово пространство R^n , состоящее из всевозможных выборок $x = (x_1, x_2, \dots, x_n)$. Оно называется *выборочным пространством*. Укажем в этом пространстве некоторую область V , обладающую следующим свойством: если выборка (точка пространства) x попадает в область V , то гипотеза H отклоняется и принимается гипотеза H' . В противном случае, когда выборка x не принадлежит области V , гипотеза H принимается, а гипотеза H' отклоняется. Область V , обладающая таким свойством, называется *критической*.

Не следует думать, что принятие гипотезы служит ее доказательством. Поскольку речь идет о случайных событиях, принятие гипотезы означает лишь то, что анализируемые данные (выборка) не противоречат данной гипотезе. В то же время отклонение гипотезы является доказательством, поскольку в этом случае данные опровергают гипотезу. Эти различия напоминают разницу между примером и контрпримером: если некий пример не противоречит высказанному утверждению, это еще не значит, что утверждение верно (иначе говоря, пример — это не доказательство!). Однако если найден контрпример, противоречащий высказанному утверждению, значит доказано, что оно неверно (т.е. контрпример — доказательство!).

Разумеется, выбрать область V в пространстве R^n можно по-разному. Следовательно, можно сконструировать много разных критериев. Возникает вопрос, не существует ли такой области V , которая была в некотором смысле наилучшей?

¹ Не путайте их с великими однофамильцами Джоном фон Нейманом (John von Neumann) и Карлом Пирсоном (Karl Pearson). Особенно легко перепутать Карла и Эгона Пирсона, так как они оба были статистиками. Кривая Пирсона и распределение Пирсона — изобретение Карла Пирсона, а теория проверки гипотез — Эгона Пирсона.

В таком случае критерий, основанный на проверке принадлежности выборки области V , был бы наилучшим. Для того чтобы ответить на этот вопрос, необходимо ввести понятие ошибки статистического критерия.

Во многих ситуациях основная (нулевая) и конкурирующая гипотезы H и H' не равнозначны. Например, предположим, что у некоего пациента наблюдается постоянное повышение температуры до $37,5^\circ\text{C}$. Возможно, у него туберкулез (гипотеза H). Однако такое постоянное повышение температуры может наблюдаться и при термоневрозе (гипотеза H'). Очевидно, что гипотеза H , предполагающая существование серьезной опасности для больного, намного значимее гипотезы H' . Следовательно, последствия ошибки, совершенной при постановке диагноза, имеют разный вес. Если врач ошибочно поставит диагноз “туберкулез легких”, приняв гипотезу H , то больной будет вынужден принимать лекарства. Для здорового человека это, разумеется, вредно. Однако если врач примет гипотезу H' , не заметив туберкулеза, то болезнь будет запущена, и опасность возрастет. Несмотря на очевидную наивность этого примера, совершенно очевидно, что в качестве основной гипотезы H всегда следует выбирать более важное утверждение.

Ошибка первого рода возникает тогда, когда основная гипотеза H отклоняется, несмотря на то, что она является истинной. Ошибка первого рода обозначается буквой α . Поскольку гипотеза H отклоняется, когда выборка $x = (x_1, x_2, \dots, x_n)$ попадает в критическую область V , величину α можно вычислить так:

$$\begin{aligned}\alpha &= p(\text{отклонить } H | H - \text{истинная}) = p(x \text{ попадает в } V | H - \text{истинная}) = \\ &= \int \dots \int_V f(u_1, u_2, \dots, u_n) du_1 du_2 \dots du_n.\end{aligned}$$

Здесь $f(u_1, u_2, \dots, u_n) = \frac{\partial^n F(u_1, u_2, \dots, u_n)}{\partial u_1 \partial u_2 \dots \partial u_n}$ — совместная плотность вероятности генеральной совокупности с функцией распределения $F(u_1, u_2, \dots, u_n)$.

Ошибка второго рода возникает, когда ложная гипотеза H не отвергается. Ошибка второго рода обозначается буквой β . Поскольку гипотеза H принимается, когда выборка $x = (x_1, x_2, \dots, x_n)$ не попадает в критическую область V , величину β можно вычислить так.

$$\begin{aligned}\alpha &= p(\text{принять } H | H - \text{ложная}) = p(x \text{ не попадает в } V | H - \text{ложная}) = \\ &= \int \dots \int_{R^n \setminus V} g(u_1, u_2, \dots, u_n) du_1 du_2 \dots du_n.\end{aligned}$$

Здесь $g(u_1, u_2, \dots, u_n) = \frac{\partial^n G(u_1, u_2, \dots, u_n)}{\partial u_1 \partial u_2 \dots \partial u_n}$ — совместная плотность вероятности генеральной совокупности с функцией распределения $G(u_1, u_2, \dots, u_n)$, а символами $R^n \setminus V$ называется множество точек, не принадлежащих области V (ее дополнение).

Число $\gamma = 1 - \beta$ называется *мощностью критерия*. Вероятность ошибки первого рода задается заранее. Как правило, в биологии и медицине принято выбирать $\alpha = 0,05$. Итак, будем считать, что величина α известна. Теперь возникает

следующая проблема. Критических областей V (соответственно критериев), для которых вероятность ошибки первого рода не превышает заданный уровень α , существует очень много. Следовательно, необходимо найти такую область V^* (и соответствующий ей критерий), у которой ошибка второго рода β является минимальной. Критерий с критической областью V^* является наилучшим среди всех критериев, у которых вероятность ошибки первого рода не превосходит α . Следуя подходу, предложенному Нейманом и Пирсоном, введем *отношение правдоподобия*

$$U(u_1, u_2, \dots, u_n) = \frac{g(u_1, u_2, \dots, u_n)}{f(u_1, u_2, \dots, u_n)}.$$

Выберем число v так, чтобы выполнялось равенство

$$P(U(x_1, x_2, \dots, x_n) \geq v | H) = \int_{\{u_1, u_2, \dots, u_n\} \in V} \dots \int g(u_1, u_2, \dots, u_n) du_1 du_2 \dots du_n.$$

Критическая область V^* , соответствующая наиболее мощному критерию, представляет множество всех точек, в которых отношение правдоподобия $U(u_1, u_2, \dots, u_n)$ принимает значения, превосходящие v :

$$V^* = \{u = (u_1, u_2, \dots, u_n): U(u_1, u_2, \dots, u_n) \geq v\}.$$

2.10.2. Оптимальные статистические критерии

Изложенная в предыдущем разделе теория Неймана–Пирсона, наряду с большими преимуществами, имеет ряд недостатков. Прежде всего, эту теорию нельзя перенести на случай N произвольных альтернативных гипотез $H_1, H_2, \dots, H_N$, хотя такая необходимость возникает довольно часто. Например, при диагностике опухолевых процессов слизистой оболочки желудка возникает не две, а четыре гипотезы: H_1 — нормальная слизистая, H_2 — одиночные полипы, H_3 — полипозно измененная слизистая, H_4 — рак желудка. Кроме того, в теории Неймана–Пирсона основная гипотеза H доминирует над гипотезой H' . Иногда такое доминирование является неестественным. Например, если гипотеза H заключается в том, что больной страдает сахарным диабетом, а H' — в том, что больной страдает несахарным диабетом, то ошибки как первого, так и второго рода являются крайне нежелательными. Следовательно, естественно попытаться построить теорию проверки нескольких равноправных альтернативных гипотез.

Пусть $x = (x_1, x_2, \dots, x_n)$ — выборка, совместное распределение которой принадлежит семейству функций $P = \{F_j(u_1, u_2, \dots, u_n), j = 1, 2, \dots, N\}$. Предположим, что гипотеза H_j соответствует распределению F_j . Во многих случаях при проверке гипотез объем n выборки x можно резко уменьшить, используя так называемые *достаточные статистики*. Этим термином называются функции, зависящие от выборочных значений $T_k = T_k(x_1, x_2, \dots, x_n)$, $k = 1, 2, \dots, m$, где m намного меньше n , в которых содержится столько же информации, как и в самой выборке $x_1, x_2, \dots, x_n$. Это позволяет использовать достаточные статистики $T_1, T_2, \dots, T_m$ вместо выборочных значений $x_1, x_2, \dots, x_n$. Предположим, гипотеза H утверждает,

что выборка $x_1, x_2, \dots, x_n$ принадлежит генеральной совокупности с нормальным распределением, имеющей нулевое математическое ожидание и дисперсию σ^2 , а гипотеза H' отличается от гипотезы H только тем, что математическое ожидание этой совокупности не равно нулю. В качестве достаточной статистики в этом случае можно выбрать одномерную случайную величину $t = \bar{x} = \frac{1}{n} \sum_{k=1}^n x_k$ (выборочное среднее). Таким образом, $m = 1$.

Обозначим через $T \in R^n$, $m \leq n$, достаточную статистику для семейства P . Без ограничения общности можно считать, что статистика T имеет плотность распределения $f(t)$, $t \in R^n$. *Нерандомизированная процедура* проверки гипотезы заключается в разбиении пространства R^n на N непересекающихся множеств $V_1, V_2, \dots, V_N$. Если после эксперимента статистика T попадает в область V_j , то принимается гипотеза H_j , а остальные гипотезы отклоняются.

Рандомизированная процедура имеет более общий характер. Она предусматривает задание в пространстве R^n критической векторной функции $\varphi(t) = (\varphi_1(t), \varphi_2(t), \dots, \varphi_N(t))$, $\varphi_j(t) \geq 0$, $\sum_{j=1}^N \varphi_j(t) = 1$. Для проверки гипотезы с помощью рандомизированной процедуры, кроме вычисления достаточной статистики t , необходимо провести дополнительный случайный эксперимент, результатом которого является одно из чисел $1, 2, \dots, N$, выпадающих с вероятностями $\varphi_1(t), \varphi_2(t), \dots, \varphi_N(t)$. Если в результате этого эксперимента появляется число j , то принимается гипотеза H_j , а остальные гипотезы отвергаются. Легко видеть, что рандомизированная процедура является обобщением нерандомизированной: для этого достаточно положить $\varphi_j(t) \equiv 1$, если $t \in V_j$, и $\varphi(t) \equiv 0$, если $t \notin V_j$. В этом случае дополнительный случайный эксперимент не дает никакой новой информации: при его проведении значение j появляется тогда и только тогда, когда $t \in V_j$.

Обозначим через $L_i(j)$, $i, j = 1, 2, \dots, N$, $i \neq j$, положительные числа, определяющие потери, связанные с принятием гипотезы H_j , когда на самом деле справедлива гипотеза H_i . Пусть $p(H_i|H_j)$, $i \neq j$ — вероятность принятия гипотезы H_i при условии, что верна гипотеза H_j . Рассмотрим *вектор риска* $r_\varphi = (r_\varphi(1), r_\varphi(2), \dots, r_\varphi(N))$, где

$$r_\varphi(j) = m_j[L(j)] = \sum_{i=1, i \neq j}^N L_i(j) p(H_i|H_j).$$

Введем в пространстве R^N две нормы векторов:

$$\|u\|_1 = \sum_{i=1}^N \pi_i |u_i|, \quad u = (u_1, u_2, \dots, u_N),$$

где $\pi_i > 0$ — некоторые заданные положительные числа, и

$$\|u\|_2 = \max \{|u_1|, |u_2|, \dots, |u_N|\}.$$

Обозначим через Φ множество всех критических функций. Будем говорить, что статистический критерий проверки конкурирующих альтернативных гипо-

тез $H_j, j = 1, 2, \dots, N$, соответствующий критической функции ϕ^* , является *оптимальным по норме* $\|u\|$, если норма вектора риска r_{ϕ^*} является минимальной в классе всех критериев, соответствующих критическим функциям $\phi \in \Phi$, т.е.

$$\|r_{\phi^*}\| = \min_{\phi \in \Phi} \|(r_{\phi}(1), r_{\phi}(2), \dots, r_{\phi}(N))\|.$$

Оказывается, что в случае двух конкурирующих альтернативных гипотез H и H' оптимальный статистический критерий определяется отношением правдоподобия $U(u_1, u_2, \dots, u_n) = \frac{g(u_1, u_2, \dots, u_n)}{f(u_1, u_2, \dots, u_n)}$ следующим образом: для любой нормы, определенной на плоскости R^2 , оптимальный статистический критерий задается критической функцией $\phi^* = (\phi_1^*, \phi_2^*)$, определяемой формулой

$$\phi_1^*(t_1, t_2, \dots, t_m) = \begin{cases} 1, & \text{если } U(t_1, t_2, \dots, t_m) < \lambda L, \\ \alpha, & \text{если } U(t_1, t_2, \dots, t_m) = \lambda L, \\ 0, & \text{если } U(t_1, t_2, \dots, t_m) > \lambda L, \end{cases}$$

$$\phi_2^*(t_1, t_2, \dots, t_m) = 1 - \phi_1^*(t_1, t_2, \dots, t_m),$$

где $L = L_2(1)/L_1(1)$, а α и λ — некоторые положительные числа, зависящие от вида нормы в пространстве R^2 . Аналогичное утверждение имеет место и для конкурирующих альтернативных гипотез.

Байесов критерий проверки гипотез заключается в следующем. Введем функцию

$$h_j(t) = \sum_{i=1, i \neq j}^N L_j(i) f_i(t) \pi_i, \quad j=1, 2, \dots, N.$$

Если статистика T принимает значение t из множества

$$V_j = \{t = (t_1, t_2, \dots, t_m): h_j(t) < h_i(t), i = 1, 2, \dots, N, i \neq j\},$$

то истинной считается гипотеза H_j ; если же значение t попадает в множество

$$V_{\mu_1, \mu_2, \dots, \mu_k} = \{t: h_{\mu_1}(t) = h_{\mu_2}(t) = \dots = h_{\mu_k}(t) < h_i(t), i = 1, 2, \dots, N, i \neq \mu_s, s = 1, 2, \dots, k\},$$

то произвольным образом принимается одна из гипотез $H_{\mu_1}, H_{\mu_2}, \dots, H_{\mu_k}$.

2.10.3. Статистические критерии, использующие неприятие решений

Как показано в 2.10.1, в классической теории Неймана–Пирсона проверки гипотез с помощью статистических критериев для двух конкурирующих альтернативных статистических гипотез H и H' оптимальный (наиболее мощный) критерий строится так, что вероятность ошибки первого рода $p(H|H)$ не превосходит некоторый критический уровень α^* , а вероятность ошибки второго рода $p(H|H')$ достигает минимального значения. При этом, однако, о вероятно-

сти ошибки второго рода нельзя сказать ничего определенного, кроме того, что она принимает значения от 0 до 1. Такой подход к проверке гипотез является неестественным для многих задач, возникающих в биологии и медицине. Например, при дифференциальной диагностике злокачественных опухолей (гипотеза H) от гиперпластических процессов (гипотеза H') по данным морфометрических и денситометрических показателей необходимо построить критерий так, чтобы вероятность ошибок первого и второго рода не превосходила некоторых достаточно малых критических уровней α^* и β^* :

$$p(H'|H) \leq \alpha^*, p(H|H') \leq \beta^*. \quad (2.10)$$

Если применять подход Неймана–Пирсона, то можно получить наиболее мощный статистический критерий, у которого вероятность ошибки второго рода, т.е. вероятность поставить диагноз злокачественного новообразования при наличии гиперплазии ткани, будет высокой, а применение такого критерия в клинической практике приведет к тяжелым последствиям для пациентов. К сожалению, если следует непременно принять или отвергнуть основную гипотезу H , то при фиксированном объеме выборки может не существовать ни одного статистического критерия, удовлетворяющего условию (2.10). Оказывается, что построить оптимальный критерий, удовлетворяющий условию (2.10), можно, введя процедуру непринятия решения RD .

Пусть две конкурирующие альтернативные гипотезы H и H' означают, что выборка $x = (x_1, x_2, \dots, x_n)$ принадлежит генеральным совокупностям с функциями распределения $F(u) = F(u_1, u_2, \dots, u_n)$ и $G(u) = G(u_1, u_2, \dots, u_n)$ с плотностями вероятности $f(u) = f(u_1, u_2, \dots, u_n)$ и $g(u) = g(u_1, u_2, \dots, u_n)$ соответственно. Разобьем выборочное пространство R^n на три непересекающиеся области V , W и T : $R^n = V \cup W \cup T$. Это разбиение определяет критерий $t(V, W, T)$ для проверки гипотез H и H' следующим образом: если $x \in V$, то принимается гипотеза H , если $x \in W$, то принимается гипотеза H' , а если $x \in T$, то не принимается ни одна из гипотез H и H' , т.е. решение не принимается (это событие обозначается символом RD).

$$\begin{aligned} p(H'|H) &= p(x \in W|H) = \int_{\dots} \int_W f(u_1, u_2, \dots, u_n) du_1 du_2 \dots du_n = \int_W f(u) du, \\ p(H|H') &= p(x \in V|H') = \int_{\dots} \int_V g(u_1, u_2, \dots, u_n) du_1 du_2 \dots du_n = \int_V g(u) du, \\ p(RD|H) &= p(x \in T|H) = \int_{\dots} \int_T f(u_1, u_2, \dots, u_n) du_1 du_2 \dots du_n = \int_T f(u) du, \\ p(RD|H') &= p(x \in T|H') = \int_{\dots} \int_T g(u_1, u_2, \dots, u_n) du_1 du_2 \dots du_n = \int_T g(u) du. \end{aligned}$$

Рассмотрим проблему построения оптимального критерия $t(V^*, W^*, T^*)$: найти разбиение выборочного пространства $R^n = V^* \cup W^* \cup T^*$, удовлетворяющее условию (2.10), для которого сумма вероятностей непринятия решения

$$p(RD|H) + p(RD|H') \quad (2.11)$$

достигает минимального значения. Покажем, что решить эту проблему с помощью отношения правдоподобия

$$L(u) = L(u_1, u_2, \dots, u_n) = \frac{g(u_1, u_2, \dots, u_n)}{f(u_1, u_2, \dots, u_n)}. \quad (2.12)$$

Как и в теории Неймана–Пирсона, если в некоторой части S пространства R^n функция f равна нулю, то мы можем, не нарушая условия (2.10), объединить области S и W . От такого увеличения вероятность ошибки первого рода не изменится, а вероятность ошибки второго рода и значение функции (2.11) может лишь уменьшиться. Следовательно, без ограничения общности можно считать, что плотность $f(u)$ всюду является положительной, поэтому отношение правдоподобия $L(u)$ определено во всех точках $u \in R^n$. Введем случайную величину (статистику) $\chi = L(x)$. В зависимости от распределения генеральной совокупности, из которой извлечена выборка x , статистика χ может иметь два разных распределения. Обозначим через χ_f статистику χ , если выборка x извлечена из генеральной совокупности с функцией распределения $F(u_1, u_2, \dots, u_n)$ (гипотеза H), а через χ_g — статистику χ , если выборка x извлечена из генеральной совокупности с функцией распределения $G(u_1, u_2, \dots, u_n)$ (гипотеза H'). Предположим, что для заданных критических уровней α^* и β^* существуют такие доверительные границы t_{α^*} и $\hat{t}_{\beta^*}$, что $p(\chi_f \geq t_{\alpha^*}) = p(L(x) \geq t_{\alpha^*} | H) = \alpha^*$, $p(\chi_g \leq \hat{t}_{\beta^*}) = p(L(x) \leq \hat{t}_{\beta^*} | H') = \beta^*$. Если одномерные случайные величины χ_f и χ_g имеют непрерывные функции распределения, то можно построить нерандомизированный критерий. Это условие всегда выполняется, когда плотности вероятности $f(u)$ и $g(u)$ являются непрерывными функциями. Если функции распределения χ_f и χ_g не принимают значений $1 - \alpha^*$ и β^* , то можно ввести рандомизированные процедуры. Однако эти процедуры редко применяются на практике, так что мы не будем их рассматривать.

Нерандомизированный оптимальный статистический критерий $\kappa(V^*, W^*, T^*)$ для проверки конкурирующих альтернативных гипотез H и H' , использующий процедуру непринятия решений RD , порождается разбиением $R^n = V^* \cup W^* \cup T^*$, полученным с помощью отношения правдоподобия (2.12) следующим образом:

- 1) если $p(\chi_f \geq \hat{t}_{\beta^*}) = p(L(x) \geq \hat{t}_{\beta^*} | H) \leq \alpha^*$, то $V^* = \{u : L(u) < \hat{t}_{\beta^*}\}$,
 $W^* = \{u : L(u) \geq \hat{t}_{\beta^*}\}$, $T^* = \emptyset$;
- 2) если $p(\chi_f \geq \hat{t}_{\beta^*}) = p(L(x) \geq \hat{t}_{\beta^*} | H) > \alpha^*$, то $V^* = \{u : L(u) \leq \hat{t}_{\beta^*}\}$,
 $W^* = \{u : L(u) \geq t_{\alpha^*}\}$, $T^* = \{u : \hat{t}_{\beta^*} < L(u) < t_{\alpha^*}\}$.

Итак, построение оптимального критерия сводится к разбиению числовой прямой на два луча и расположенного между ними интервала, который может вырождаться в пустое множество. Действительно, если вместо выборки

$x = (x_1, x_2, \dots, x_n)$ рассмотреть значение статистики $\chi = L(x)$, которая является одномерной случайной величиной, то проблема построения оптимального статистического критерия сводится к задаче разбиения числовой прямой на два луча $(-\infty, \hat{t}_\beta]$, $(\hat{t}_\beta, \infty)$ (первый вариант) или на два луча $(-\infty, \hat{t}_\beta]$, $[\hat{t}_\alpha, \infty)$ и интервал $(\hat{t}_\beta, \hat{t}_\alpha)$. Если $\chi \in (-\infty, \hat{t}_\beta]$, то принимается гипотеза H ; если $\chi \in [\hat{t}_\alpha, \infty)$ или $\chi \in (\hat{t}_\beta, \infty)$, то принимается гипотеза H' , а если $\chi \in (\hat{t}_\beta, \hat{t}_\alpha)$, то решение не принимается. В связи с этим случайная величина $\chi = L(x)$ называется статистикой, порождающей оптимальный критерий, или *порождающей статистикой*.

Для построения оптимального статистического критерия, кроме отношения правдоподобия $\chi = L(x)$, необходимо знать доверительные границы $\hat{t}_\alpha$ и $\hat{t}_\beta$, соответствующие критическим уровням α и β , а для этого, в свою очередь, нужно определить функции распределения статистик χ_β и χ_α . К сожалению, эта задача очень сложная, но во многих практически важных случаях порождающую статистику χ можно представить в виде монотонной функции φ от другой статистики $\mathcal{Y} = h(x_1, x_2, \dots, x_n)$, определяемой по выборочным значениям, распределение которой $F_{\mathcal{Y}}(t)$ вычисляется относительно просто: $\chi = \varphi(\mathcal{Y})$. Статистика $\mathcal{Y}$ называется *эквивалентной порождающей статистикой*. В самом деле, если функция φ монотонно возрастает, то

$$p(\chi = L(x) \leq \hat{t}_\beta | H) = p(\mathcal{Y} \leq \hat{\theta}_\beta | H) = \beta, \quad p(\chi = L(x) \geq \hat{t}_\alpha | H) = p(\mathcal{Y} \geq \theta_\alpha | H) = \alpha.$$

Следовательно, в первом варианте прямая разбивается на два луча $(-\infty, \hat{\theta}_\beta)$, $(\hat{\theta}_\beta, \infty)$ или $(-\infty, \hat{\theta}_\beta]$, $[\hat{\theta}_\alpha, \infty)$ и интервал $(\hat{\theta}_\beta, \theta_\alpha)$ во втором варианте, а порождающая статистика χ заменяется эквивалентной порождающей статистикой $\mathcal{Y}$. Аналогичные рассуждения можно провести, если φ является монотонно убывающей функцией. Приведем примеры эквивалентных порождающих статистик, предполагая, что выборка $x = (x_1, x_2, \dots, x_n)$ получена с помощью простого случайного выбора, т.е. случайные величины $x_1, x_2, \dots, x_n$ независимы в совокупности и имеют одинаковые функции распределения.

- 1. Нормальное распределение.** Пусть гипотеза H означает, что случайные величины x_k , $k = 1, 2, \dots, n$, имеют нормальное распределение $N(a, \sigma^2)$, а гипотеза H' утверждает, что они имеют нормальное распределение $N(a', \sigma'^2)$. Можно показать, что при $\sigma = \sigma'$ эквивалентной порождающей статисти-

кой является случайная величина $\sum_{k=1}^n x_k$, имеющая нормальное распре-

деление с параметрами na и $n\sigma^2$, если верна гипотеза H , и na и $n\sigma'^2$, если верна гипотеза H' . Если $\sigma \neq \sigma'$, то в качестве эквивалентной порождаю-

щей статистики можно взять случайную величину $\sum_{k=1}^n (x_k + b)^2$, где b — не-
которая известная постоянная.

2. **Гамма-распределение.** Пусть гипотеза H означает, что случайные величины x_k , $k = 1, 2, \dots, n$, имеют гамма-распределение с параметрами α и β , а гипотеза H' означает, что они имеют гамма-распределение с параметрами α' и β' . В этом случае эквивалентной порождающей статистикой является случайная величина $\sum_{k=1}^n x_k$.
3. **Показательное распределение.** Пусть гипотеза H означает, что случайные величины x_k , $k = 1, 2, \dots, n$, имеют показательное распределение с параметром λ , а гипотеза H' означает, что они имеют показательное распределение с параметром λ' . Эквивалентной порождающей статистикой является случайная величина $\sum_{k=1}^n \log x_k$.
4. **Логнормальное распределение.** Пусть гипотеза H означает, что случайные величины x_k , $k = 1, 2, \dots, n$, имеют логнормальное распределение с параметрами $\delta > 0$ и γ , а гипотеза H' означает, что они имеют логнормальное распределение с параметрами $\delta' > 0$ и γ' . Эквивалентной порождающей статистикой является случайная величина $x_1 x_2 \dots x_n$.

Отметим, что сведение многомерной задачи разбиения выборочного пространства R^n к числовой прямой с помощью порождающей статистики можно осуществить только в условиях классической постановки задачи, когда известны распределения, соответствующие гипотезам. Если же исследователь располагает только обучающими выборками, приходится проводить разбиение всего пространства R^n .

2.10.4. Индивидуальные статистические критерии

Теория Неймана–Пирсона имеет один существенный недостаток. Для того чтобы продемонстрировать его, рассмотрим вариант, в котором выборка $x = (x_1, x_2, \dots, x_n)$ состоит из одного показателя $x = x_1$, так что $n = 1$, случайный вектор x представляет собой одномерную случайную величину, а пространство R^n является числовой прямой. Можно показать, что при выполнении ряда естественных условий, наложенных на функции $f(u)$ и $g(u)$, критическая область V — это луч $(-\infty, v)$, определяемый критическим значением v , а ее дополнение — это луч (v, ∞) . Следовательно, при использовании наиболее мощного критерия гипотеза H принимается, если $x > v$, а гипотеза H' принимается, если $x \leq v$. Предположим, что гипотеза H означает сахарный диабет, а гипотеза H' — почечный диабет. Как известно, наиболее надежным показателем, характеризующим нарушение углеводного обмена, является изменение количества сахара в крови, которое у здорового человека колеблется в пределах 80–120 мг%. Повышение уровня сахара в крови называется гипергликемией. Основными признаками

сахарного диабета являются гипергликемия и глюкозурия (выделение сахара с мочой), интенсивность которых бывает разной. В тяжелых случаях сахарного диабета уровень сахара в крови доходит до 600 мг%. Почечный диабет характеризуется появлением глюкозурии при нормальном содержании сахара в крови. Отсюда следует, что в качестве статистики x при дифференциальной диагностике сахарного и почечного диабета естественно принять содержание сахара в крови. При этом критическое значение $v = 120$ мг%. Таким образом, наилучший критерий принимает следующий вид: если $x > v$, то принимается гипотеза H (сахарный диабет), а если $x \leq v$ — гипотеза H' (почечный диабет). Предположим, что вероятность ошибок первого и второго рода для этого критерия составляют 0,05 и 0,08 соответственно (их конкретные значения не играют существенной роли). Допустим также, что обследоване проходят два пациента с уровнями сахара в крови $x_1 = 130$ мг% и $x_2 = 600$ мг%. Руководствуясь выбранным критерием, мы должны обоих пациентов отнести к группе больных сахарным диабетом, т.е. принять гипотезу H . Подчеркнем, что вероятности ошибок первого и второго рода в обоих случаях одинаковы! Это противоречит здравому смыслу, поскольку у второго пациента уровень сахара в крови составляет 600 мг%, т.е. у него достоверно диагностируется сахарный диабет (при этом вероятность ошибки практически равна нулю). В то же время диагноз, поставленный первому пациенту, является не настолько определенным, и вероятность ошибки может быть большой. Таким образом, существенный недостаток классической теории Неймана–Пирсона заключается в том, что вероятность ошибки не зависит от конкретного значения случайной величины x , измеренной в результате эксперимента. Как будет показано в дальнейшем, в теории Неймана–Пирсона вероятность ошибки, соответствующей конкретному значению x , определить нельзя.

Для исправления этого недостатка введем следующее определение.

Определение 2.1.

Будем называть статистический критерий T , предназначенный для проверки двух конкурирующих альтернативных гипотез H и H' , индивидуальным, если вероятность ошибки, возникающей при принятии конкретного решения зависит от индивидуального вектора признаков $x_0 = (x_1^{(0)}, x_2^{(0)}, \dots, x_n^{(0)})$, измеренного в ходе случайного испытания.

Пусть H и H' — две конкурирующие альтернативные гипотезы, утверждающие, что n -мерный случайный вектор $x = (x_1, x_2, \dots, x_n)$ имеет совместную плотность распределения координат $f(u_1, u_2, \dots, u_n)$ или $g(u_1, u_2, \dots, u_n)$ соответственно. Предположим также, что, кроме совместных плотностей вероятности, нам известны безусловные вероятности гипотез $p(H)$ и $p(H')$. Эти показатели можно интерпретировать как вероятности случайных событий $\mathcal{H}$ и $\mathcal{H}'$, которые заключаются в том, что случайный вектор x , выбранный наугад из объединенной генеральной совокупности, имеет плотность распределения координат $f(u_1, u_2, \dots, u_n)$ (событие $\mathcal{H}$) или $g(u_1, u_2, \dots, u_n)$ (событие $\mathcal{H}'$). На практике эти величины заменяются частотами.

Рассмотрим подробнее математическую модель объединенной генеральной совокупности G , о которой идет речь при определении вероятностей $p(H)$ и $p(H')$. С точки зрения математической статистики, гипотеза H означает, что наблюдаемый в ходе эксперимента случайный вектор x принадлежит генеральной совокупности G_H с плотностью вероятности $f(u_1, u_2, \dots, u_n)$, а гипотеза H' означает, что случайный вектор x принадлежит генеральной совокупности $G_{H'}$ с плотностью вероятности $g(u_1, u_2, \dots, u_n)$. Обозначим через G генеральную совокупность, представляющую собой смесь генеральных совокупностей G_H и $G_{H'}$ с весами $p(H)$ и $p(H')$, соответственно: $G = p(H) G_H + p(H') G_{H'}$, где $p(H) + p(H') = 1$. Найдем совместную плотность вероятностей $h(u_1, u_2, \dots, u_n)$ координат вектора x из генеральной совокупности G . Используя формулу полной вероятности, нетрудно показать, что

$$h(u_1, u_2, \dots, u_n) = p(H)f(u_1, u_2, \dots, u_n) + p(H')g(u_1, u_2, \dots, u_n). \quad (2.13)$$

Таким образом, если $x \in G_H$, то ее распределение

$$p(x_1 < u_1, \dots, x_n < u_n | H) = F(u_1, u_2, \dots, u_n), \quad (2.14)$$

а если $x \in G_{H'}$, то

$$p(x_1 < u_1, \dots, x_n < u_n | H') = G(u_1, u_2, \dots, u_n). \quad (2.15)$$

Если выборка x принадлежит объединенной (составной) генеральной совокупности G , то ее распределение $H(u_1, u_2, \dots, u_n)$ определяется совместной плотностью вероятности $h(u_1, u_2, \dots, u_n)$, полученной по формуле (2.13):

$$p(x_1 < u_1, \dots, x_n < u_n) = H(u_1, u_2, \dots, u_n) \quad (2.16)$$

Вероятность (2.15) можно интерпретировать как безусловную вероятность, тогда как вероятности (2.14) и (2.15) можно считать условными и зависящими от наступления событий $\mathcal{H}$ и $\mathcal{H}'$.

Допустим, что в ходе случайного испытания статистика x приняла конкретное значение $x = x_0$. Рассмотрим условную вероятность $p(x_0) = p(H|x = x_0)$ того, что гипотеза H верна при условии, что $x = x_0$. Назовем это число апостериорной вероятностью гипотезы H , соответствующей значению x_0 . Апостериорная вероятность $p(x_0)$ является функцией, зависящей от n переменных. Апостериорная вероятность гипотезы H' определяется аналогично: $p'(x_0) = p(H'|x = x_0)$. Поскольку гипотезы H и H' являются конкурирующими и альтернативными, выполняется следующее равенство:

$$p(x_0) + p'(x_0) = p(H|x = x_0) + p(H'|x = x_0) = 1. \quad (2.17)$$

Для построения статистического критерия для проверки гипотез H и H' рассмотрим отношение апостериорных вероятностей:

$$\gamma(x_0) = \frac{p'(x_0)}{p(x_0)} = \frac{p'(H'|x = x_0)}{p(H|x = x_0)}. \quad (2.18)$$

Это отношение определено во всех точках $x \in R^n$, поскольку из формулы (2.18) следует, что числитель и знаменатель этой дроби не могут одновременно обращаться в нуль. Если вероятность $p(x)$ равна нулю, то $\gamma(x) = \infty$.

Пусть λ — некоторая константа, которая выбирается из практических соображений. Введем следующее обозначение:

$$W_\lambda = \left\{ u = (u_1, u_2, \dots, u_n) : \gamma(u) = \frac{p'(u)}{p(u)} < \frac{1}{\lambda} \right\} = \left\{ u = (u_1, u_2, \dots, u_n) : \lambda p'(u) < p(u) \right\}.$$

Критерий W_λ для проверки гипотез H и H' имеет следующий вид: если статистика x принимает значение $x_0 \in W_\lambda$, то принимается гипотеза H . В противном случае, когда $x_0 \notin W_\lambda$, принимается гипотеза H' . Если в результате испытания принимается гипотеза H , которая, как правило, является более важной по сравнению с конкурирующей альтернативной гипотезой H' , то апостериорная вероятность гипотезы H в λ раз превышает апостериорную вероятность гипотезы H' . В практике статистических вычислений во многих случаях число λ естественно считать равным единице. В этом случае гипотеза H принимается, когда ее апостериорная вероятность $p(H|x=x_0)$ превышает соответствующую вероятность $p(H'|x=x_0)$ гипотезы H' : $p(H|x=x_0) > p(H'|x=x_0)$. В отличие от теории Неймана-Пирсона, ошибка критерия W_λ характеризуется лишь одной величиной $\delta(x_0)$, которая принимает значение $1 - p(H|x=x_0)$, если гипотеза H принимается: $\delta(x_0) = 1 - p(H|x=x_0) = p(H'|x=x_0)$, и $1 - p(H'|x=x_0)$, если гипотеза H отклоняется, т.е. принимается гипотеза H' : $\delta(x_0) = 1 - p(H'|x=x_0) = p(H|x=x_0)$. Таким образом, в противоположность классической теории Неймана-Пирсона, ошибки $\delta(x_0)$ критерия W_λ зависят от того значения x_0 , которое принимает статистика x в результате случайного испытания, а также от гипотезы (H или H'), которая считается истинной. Отметим, что, если известно, какая из гипотез H или H' принимается, то ошибку критерия $\delta(x_0)$ можно определить с помощью отношения $\gamma(x_0)$ апостериорных вероятностей. Действительно, пусть в результате случайного испытания истинной считается гипотеза H и $\gamma(x_0) = \gamma$. В этом случае можно провести следующие вычисления:

$$\begin{aligned} p'(x_0) &= \gamma p(x_0), \\ \gamma p(x_0) + p(x_0) &= 1, \\ (\gamma + 1)p(x_0) &= 1. \end{aligned}$$

Следовательно,

$$\delta(x) = \delta_H(x_0) = p'(x_0) = \frac{\gamma}{1 + \gamma}.$$

В противном случае, когда гипотеза H отклоняется (т.е. принимается гипотеза H'), имеет место следующее равенство:

$$\delta(x_0) = \delta_{H'}(x_0) = p(x_0) = \frac{1}{1 + \gamma}.$$

Таким образом, если принимается гипотеза H , то

$$\delta(x_0) = \delta_H(x_0) = \frac{\gamma}{1 + \gamma}, \quad (2.19)$$

а если принимается гипотеза H' , то

$$\delta(x_0) = \delta_{H'}(x_0) = \frac{1}{1 + \gamma}. \quad (2.20)$$

Как видим, для того чтобы определить величину ошибки $\delta(x_0)$, достаточно знать отношение апостериорных вероятностей, а не сами апостериорные вероятности. Кроме того, на основании формул (2.19) и (2.20) можно определить верхнюю границу для ошибок критерия $\mathcal{W}_\lambda$. В самом деле, гипотеза считается истинной, если отношение апостериорных вероятностей не превышает $1/\lambda$. Из формулы (2.19) следует, что

$$\delta(x_0) = \delta_H(x_0) = \frac{\gamma}{1 + \gamma} < \frac{1}{1 + \lambda}.$$

Значит, величина $\frac{1}{1 + \lambda} = \delta_H^*$ представляет собой верхнюю границу для $\delta_H(x_0)$.

Это позволяет сформулировать практические рекомендации для выбора величины λ . Например, если необходимо обеспечить 5%-ный уровень значимости (т.е. $\delta_H(x_0) < 0,05$), то следует положить $\lambda = 19$. Аналогичные рассуждения можно провести и для верхней границы ошибки $\delta_{H'}(x_0)$.

Перейдем к проблеме вычисления отношения апостериорных вероятностей $\gamma(u)$, позволяющего построить критерий $\mathcal{W}_\lambda$. Можно доказать, что

$$\gamma(x_0) = \frac{p(H' | x = x_0)}{p(H | x = x_0)} = \frac{p(H') g(x_1^{(0)}, x_2^{(0)}, \dots, x_n^{(0)})}{p(H) f(x_1^{(0)}, x_2^{(0)}, \dots, x_n^{(0)})}, \quad (2.21)$$

где $x = (x_1, x_2, \dots, x_n)$, $x_0 = (x_1^{(0)}, x_2^{(0)}, \dots, x_n^{(0)})$. Из этой формулы следует, что $\gamma(x_0)$ представляет собой произведение двух сомножителей: первый сомножитель $\alpha = p(H')/p(H)$ — постоянная величина, не зависящая от значения x_0 (отношение вероятностей гипотез), а второй — хорошо известное отношение правдоподобия

$$\beta(x_0) = \mathcal{U}(x_0) = \frac{g(x_1^{(0)}, x_2^{(0)}, \dots, x_n^{(0)})}{f(x_1^{(0)}, x_2^{(0)}, \dots, x_n^{(0)})}. \quad (2.22)$$

Отношение правдоподобия $\mathcal{U}(x_0)$ является функцией, зависящей от аргумента x_0 , принадлежащего множеству M_G значений объединенной генеральной совокупности G , которое без ограничения общности можно отождествить со всем пространством R^n .

Предложенный статистический критерий $\mathcal{W}_\lambda$ для проверки гипотез H и H' можно сформулировать следующим образом:

- 1) если отношение правдоподобия меньше величины $\frac{1}{\lambda\alpha}$, т.е. $\mathcal{U}(x_0) < \frac{1}{\lambda\alpha}$, то гипотезу H следует принять; при этом критическое значение λ определяется, исходя из конкретных условий задачи, а вероятность ошибки $\delta_H(x_0)$ вычисляется по формуле (2.19);
- 2) если $\mathcal{U}(x_0) \geq \frac{1}{\lambda\alpha}$, то гипотезу H следует отвергнуть и принять гипотезу H' ; в этом случае вероятность ошибки определяется формулой (2.20).

Как было указано, в некоторых прикладных задачах основная гипотеза H доминирует над альтернативной гипотезой H' , поэтому вероятность ошибки $\delta_H(x_0)$ должна быть как можно более малой величиной, в то время как величина $\delta_{H'}(x_0)$ может быть достаточно большой. В то же время на практике часто возникают задачи, в которых гипотезы H и H' являются равноправными, а вероятности ошибок $\delta_H(x_0)$ и $\delta_{H'}(x_0)$ не должны превышать установленные границы:

$$\delta_H(x_0) \leq \delta_H^*, \delta_{H'}(x_0) \leq \delta_{H'}^*. \quad (2.23)$$

Описанный выше критерий $\mathcal{W}_\lambda$ не позволяет решать такие задачи, поскольку сумма ошибок этого критерия равна единице: $\delta_H(x_0) + \delta_{H'}(x_0) = 1$, и неравенства (2.23) могут не выполняться. Тем не менее, если использовать процедуру принятия решения RD , то такие задачи можно решить.

Предположим, что сумма критических уровней δ_H^* и $\delta_{H'}^*$ не превосходит единицу:

$$\delta_H^* + \delta_{H'}^* < 1. \quad (2.24)$$

(Это предположение является вполне естественным, поскольку критические уровни δ_H^* и $\delta_{H'}^*$, как правило, не велики; например, в медицине и биологии используются 5%-ные уровни значимости, т.е. $\delta_H^* = 0,05$ и $\delta_{H'}^* = 0,05$.)

Положим

$$\lambda_1 = \frac{1 - \delta_H^*}{\delta_H^*}, \lambda_2 = \frac{1 - \delta_{H'}^*}{\delta_{H'}^*}. \quad (2.25)$$

В силу неравенства (2.24)

$$(1 - \delta_H^*)(1 - \delta_{H'}^*) > \delta_H^* \delta_{H'}^*.$$

Следовательно,

$$\frac{1 - \delta_H^*}{\delta_H^*} > \frac{\delta_{H'}^*}{1 - \delta_{H'}^*}$$

и

$$\frac{1}{\lambda_1} < \frac{1}{\lambda_2}.$$

Статистический критерий для проверки гипотез H и H' , использующий процедуру непринятия решения, можно сформулировать следующим образом:

- 1) если отношение правдоподобия (2.22) меньше величины $1/\lambda_1\alpha$: $U(x_0) < 1/\lambda_1\alpha$, то гипотезу H следует принять;
- 2) если отношение правдоподобия (2.22) больше величины $1/\lambda_2\alpha$: $U(x_0) > 1/\lambda_2\alpha$, то гипотезу H следует отвергнуть;
- 3) если отношение правдоподобия (2.22) удовлетворяет неравенствам $1/\lambda_1\alpha \leq U(x_0) \leq 1/\lambda_2\alpha$, то никакое решение не принимается.

Вероятности ошибок $\delta_H(x_0)$ и $\delta_{H'}(x_0)$, вычисленные по формулам (2.19) и (2.20), не превосходят критические уровни δ_H^* и $\delta_{H'}^*$ соответственно, а числа λ_1 и λ_2 вычисляются по формулам (2.25). Например, для критических уровней $\delta_H^* = 0,05$ и $\delta_{H'}^* = 0,05$ числа λ_1 и λ_2 следует выбрать равными 19 и $1/19 = 0,0621$ соответственно.

В заключение рассмотрим иллюстративный пример. Предположим, что в ходе случайного испытания получена выборка $x = (x_1, x_2, \dots, x_n)$. Нулевая гипотеза H утверждает, что величины $x_1, x_2, \dots, x_n$ являются независимыми в совокупности и имеют нормальное распределение с нулевым математическим ожиданием и единичной дисперсией. В этом случае плотность вероятностей имеет вид

$$f(u) = f(u_1, u_2, \dots, u_n) = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}(u_1^2 + \dots + u_n^2)}.$$

Конкурирующая альтернативная гипотеза H' утверждает, что величины $x_1, x_2, \dots, x_n$ независимы в совокупности и имеют нормальное распределение с ненулевым математическим ожиданием и единичной дисперсией, т.е. плотность вероятностей имеет вид

$$g(u) = g(u_1, u_2, \dots, u_n) = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}((u_1-a)^2 + \dots + (u_n-a)^2)}.$$

Отношение правдоподобия

$$U(u) = \frac{g(u)}{f(u)} = e^{a \sum_{k=1}^n u_k - \frac{1}{2}na^2}$$

является монотонно возрастающей функцией от аргумента $\bar{u} = \frac{1}{n} \sum_{k=1}^n a_k$.

Таким образом, величина $\bar{u}$ является эквивалентной порождающей статистикой. Предположим, что безусловные вероятности гипотез $p(H) = p(H') = 1/2$. Тогда область W_λ имеет вид

$$W_\lambda = \{u = (u_1, u_2, \dots, u_n) : \frac{1}{n} \sum_{k=1}^n a_k < \frac{a}{2} - \frac{1}{na} \ln \lambda\}, \text{ если } a > 0,$$

$$W_\lambda = \{u = (u_1, u_2, \dots, u_n) : \frac{1}{n} \sum_{k=1}^n a_k > \frac{a}{2} - \frac{1}{na} \ln \lambda\}, \text{ если } a < 0.$$

С помощью эквивалентной порождающей статистики $\bar{u}$ область W_λ можно сделать одномерной (лучом):

$$W_\lambda \sim \tilde{W}_\lambda = \left\{ \bar{u} : \bar{u} < \frac{a}{2} - \frac{1}{na} \ln \lambda, a > 0 \right\},$$

$$W_\lambda \sim \tilde{W}_\lambda = \left\{ \bar{u} : \bar{u} > \frac{a}{2} - \frac{1}{na} \ln \lambda, a < 0 \right\}.$$

Ошибки критерия имеют следующий вид.

$$\delta_H(x_0) = \frac{e^{an\bar{x}_0}}{e^{\frac{1}{2}na^2} + e^{na\bar{x}_0}}, \text{ когда принимается гипотеза } H,$$

$$\delta_{H'}(x_0) = \frac{1 - e^{an\bar{x}_0}}{e^{\frac{1}{2}na^2} + e^{na\bar{x}_0}}, \text{ когда принимается гипотеза } H'.$$

$$\text{Здесь } \bar{x}_0 = \frac{1}{n} \sum_{k=1}^n x_k^{(0)}.$$

Использование процедуры непринятия решения позволяет разбить числовую прямую на два луча

$$W_{\lambda_1} = \left\{ u : \mathcal{U}(u) < \frac{1}{\lambda_1} \right\}; \tilde{W}_{\lambda_1} = \left\{ \bar{u} : \bar{u} \leq \frac{a}{2} - \frac{1}{na} \ln \lambda_1, a > 0 \right\},$$

$$\tilde{W}_\lambda = \left\{ \bar{u} : \bar{u} > \frac{a}{2} - \frac{1}{na} \ln \lambda_1, a < 0 \right\},$$

$$W_{\lambda_2}^* = \left\{ u : \mathcal{U}(u) > \frac{1}{\lambda_2} \right\}; \tilde{W}_{\lambda_2} = \left\{ \bar{u} : \bar{u} > \frac{a}{2} - \frac{1}{na} \ln \lambda_1, a > 0 \right\},$$

$$\tilde{W}_\lambda^* = \left\{ \bar{u} : \bar{u} < \frac{a}{2} - \frac{1}{na} \ln \lambda_2, a < 0 \right\},$$

и два отрезка “непринятия решения”

$$\tilde{V}_{\lambda_1\lambda_2} = \left\{ \bar{u} : \frac{a}{2} - \frac{1}{na} \ln \lambda_1 \leq \bar{u} \leq \frac{a}{2} - \frac{1}{na} \ln \lambda_2, a > 0 \right\},$$

$$\tilde{U}_{\lambda_1\lambda_2} = \left\{ \bar{u} : \frac{a}{2} - \frac{1}{na} \ln \lambda_2 \leq \bar{u} \leq \frac{a}{2} - \frac{1}{na} \ln \lambda_1, a < 0 \right\}.$$

Ошибки критерия вычисляются по формулам (2.19), (2.20).

2.10.5. Статистические критерии, построенные на основании обучающих выборок

Классическую теорию Неймана–Пирсона можно применять лишь тогда, когда известны функции распределения $F(u_1, u_2, \dots, u_n)$ и $G(u_1, u_2, \dots, u_n)$, соответствующие конкурирующим и альтернативным гипотезам H и H' . К сожалению, на практике эти функции, как правило, неизвестны, а вместо них есть лишь обучающие выборки из генеральных совокупностей G и G' , соответствующих гипотезам H и H' . Следовательно, необходимо построить статистические критерии для проверки гипотез, основанные на обучающих выборках.

Предположим, что G и G' — две генеральные совокупности, а $x = (x_1, x_2, \dots, x_n)$ и $x' = (x'_1, x'_2, \dots, x'_m)$ — выборки из G и G' соответственно, полученные с помощью простого случайного выбора. Пусть $z = (z_1, z_2, \dots, z_k)$ — выборка, о которой известно лишь, что она извлечена либо из G , либо из G' , но неизвестно, из какой именно генеральной совокупности. Обозначим через $H = \{z \in G\}$ и $H' = \{z \in G'\}$ гипотезы, утверждающие, что выборка z извлечена из генеральной совокупности G и G' соответственно. Предположим, что (a, b) ((a', b')) — доверительный интервал для основной распределенной массы значений генеральной совокупности G (G'), построенной по выборке x (x') с помощью правила $3s$ (см. подраздел 2.3.4):

$$(a, b) = (\bar{x} - 3s, \bar{x} + 3s), (a', b') = (\bar{x}' - 3s', \bar{x}' + 3s'),$$

где

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k, s^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2,$$

$$\bar{x}' = \frac{1}{m} \sum_{k=1}^m x'_k, s'^2 = \frac{1}{m-1} \sum_{k=1}^m (x'_k - \bar{x}')^2.$$

Кроме того, доверительный интервал можно построить с помощью порядковых статистик $a = x_{(j)}, b = x_{(j)}, i < j, a' = x'_{(r)}, b' = x'_{(r)}, r < t$ (см. подраздел 2.3.5).

Предположим для простоты, что $z = z_1$ (т.е. $k = 1$) и $a < a' < b < b'$. Тогда статистический критерий, построенный на основе обучающих выборок, принимает следующий вид:

- 1) если $z \in (a, a')$, то принимается гипотеза H ;
- 2) если $z \in (b, b')$, то принимается гипотеза H' ;
- 3) если $z \in (a', b)$, то решение не принимается.

Обозначим через α уровень значимости доверительного интервала (a, b) , а через α' — уровень значимости доверительного интервала (a', b') . В этом случае вероятность ошибки первого рода

$$\alpha^* = p(H'|H) = p(z \in (b, b')|H) \leq p(z \notin (a, b)|H) = \alpha,$$

а вероятность ошибки второго рода

$$\beta^* = p(H|H') = p(z \in (a, a')|H') \leq p(z \notin (a', b')|H') = \alpha'.$$

Рассмотрим теперь общий случай, когда k — произвольное натуральное число и $z = (z_1, z_2, \dots, z_k)$. Если порядковые статистики $x_{(1)}, x_{(n)}, x'_{(1)}, x'_{(m)}$ удовлетворяют неравенствам

$$x_{(1)} \leq x_{(n)} \leq x'_{(1)} \leq x'_{(m)},$$

то $a = x_{(1)}, b = x_{(n)}, a' = x'_{(1)}, b' = x'_{(m)}$.

Пусть θ — произвольное число из сегмента $[0, 1)$, т.е. $0 \leq \theta < 1$ и $k \geq 2$. Определим число $r = r(\theta, k) = [k/(1 + \theta)]$, где $[\alpha]$ — целая часть числа α . Статистический критерий T_θ для проверки гипотез H и H' относительно выборки $z = (z_1, z_2, \dots, z_k)$ формулируется следующим образом:

- 1) если по крайней мере r значений исследуемой выборки z принадлежат интервалу $(-\infty, x'_{(1)})$, то принимается гипотеза H ;
- 2) если по крайней мере r значений исследуемой выборки z принадлежат интервалу $(x'_{(m)}, \infty)$, то принимается гипотеза H' ;
- 3) во всех остальных случаях решение не принимается.

Вероятность ошибки первого рода критерия T_θ равна

$$\alpha^* = p(H|H) = \frac{(k-r+1)(k-r+2)\dots k}{(n+k-r+1)(n+k-r+2)\dots(n+k)},$$

а вероятность ошибки второго рода —

$$\beta^* = p(H|H') = \frac{(k-r+1)(k-r+2)\dots k}{(m+k-r+1)(m+k-r+2)\dots(m+k)}.$$

В дальнейшем мы будем пользоваться несколько иным вариантом порядкового критерия. Предположим, что выборка $z = (z_1, z_2, \dots, z_k)$ состоит из различных по своему смыслу показателей, т.е. представляет собой *вектор признаков*. С помощью обучающих выборок из генеральных совокупностей G и G' построим для каждого показателя $z_1, z_2, \dots, z_k$ доверительные интервалы (a_i, b_i) ((a'_i, b'_i) соответственно), $i = 1, 2, \dots, k$, для основной распределенной массы их значений. Обозначим через r (r' соответственно) количество выходов показателя z_i за границы интервала (a_i, b_i) ((a'_i, b'_i) соответственно) $i = 1, 2, \dots, k$. Порядковый критерий T для проверки гипотез H и H' формулируется следующим образом:

- 1) если $r < r'$, то принимается гипотеза H ;
- 2) если $r \geq r'$, то принимается гипотеза H' .

Соответствующий критерий T_1 , допускающий непринятие решения, выглядит так:

- 1) если $r < r'$, то принимается гипотеза H ;
- 2) если $r > r'$, то принимается гипотеза H' ;
- 3) если $r = r'$, то решение не принимается.

Формула (2.9) дает возможность построить статистические критерии, основанные на обучающих выборках, для которых можно точно определить вероятность ошибки первого рода, а также достаточно точно определить вероятность ошибки второго рода. Кроме того, среди этих критериев можно выбрать критерий, у которого вероятность ошибки первого рода является фиксированной, а оценка вероятности ошибки второго рода минимальна. Этот критерий представляет собой аналог наиболее мощного критерия в теории Неймана–Пирсона. Действительно, пусть H — основная, H' — конкурирующая альтернативная гипотеза, а G и G' — соответствующие им генеральные совокупности с непрерывными и неизвестными функциями распределения. Пусть $x = (x_1, x_2, \dots, x_n)$ и $x' = (x'_1, x'_2, \dots, x'_m)$ — обучающие выборки из генеральных совокупностей G и G' соответственно. По выборке x построим вариационный ряд

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)},$$

где $x_{(i)}$ — порядковые статистики.

С помощью инвариантных доверительных интервалов $J_i = [x_{(i)}, x_{(i+1)}]$, $i = 0, 1, \dots, n$, ($x_{(0)} = -\infty$, $x_{(n+1)} = +\infty$) определим инвариантные доверительные области $D_{i \dots i_k} = \bigcup_{j=1}^k J_{i_j}$, где $0 \leq i_j \leq n+1$, $j = 1, 2, \dots, k$, доверительные уровни которых вычисляются по формуле

$$p(z \in D_{i \dots i_k} | H) = \frac{k}{n+1}, \quad z \in G.$$

Для выборки z , состоящей из одного выборочного значения $z = z_1$, статистический критерий $T_{i \dots i_k}$ для проверки гипотез H и H' , не использующий процедуру непринятия решения, формулируется так: если $z = z_1 = D_{i \dots i_k}$, то принимается гипотеза H , в противном случае, когда $z_1 \in D_{i \dots i_k}$, принимается гипотеза H' . Вычислим вероятность ошибки первого рода $\alpha(T_{i \dots i_k})$ критерия $T_{i \dots i_k}$:

$$\alpha(T_{i \dots i_k}) = p(z \notin D_{i \dots i_k} | H) = p(z \notin D_{i \dots i_k} | z \in D) = 1 - p(z \in D_{i \dots i_k} | z \in D) = \frac{n+1-k}{n+1}.$$

В частности, если $D_{i \dots i_k} = (x_{(1)}, x_{(n)})$, то

$$\alpha(T_{1 \dots n-1}) = \frac{2}{n+1}$$

и при $n \geq 39$ эта ошибка не превышает 0,05.

Вероятность ошибки второго рода $\beta(T_{i \dots i_k})$ статистического критерия $T_{i \dots i_k}$ можно оценить по выборке $x' = (x'_1, x'_2, \dots, x'_m)$, заменяя ее соответствующей частотой:

$$\beta(T_{i \dots i_k}) = p(H | H') = p(z \in D_{i \dots i_k} | H') = p(z \in D_{i \dots i_k} | z \in G') = \frac{l}{m} = h(x').$$

Здесь l — количество элементов выборки $x' = (x'_1, x'_2, \dots, x'_m)$, попавших в область $D_{i_1 \dots i_k}$. Отметим, что при $m \geq 30$ частота $h(x')$ мало отличается от соответствующей вероятности $\beta(T_{i_1 \dots i_k})$.

Сформулируем оптимальный статистический критерий. Пусть k — фиксированное натуральное число. Обозначим через S_k класс статистических критериев $T_{i_1 \dots i_k}$, порожденных k доверительными интервалами $J_{i_1} = [x_{(i_1)}, x_{(i_1+1)}], \dots, J_{i_k} = [x_{(i_k)}, x_{(i_k+1)}]$. Поскольку число k фиксировано, все статистические критерии $T_{i_1 \dots i_k} \in S_k$ имеют одну и ту же вероятность ошибки первого рода. Естественно выбрать среди этого множества критериев такой критерий $T_{i_1 \dots i_k}^* = T^*$, имеющий минимальную вероятность ошибки второго рода. Этот критерий T^* называется *оптимальным*. Он аналогичен наиболее мощному критерию в теории Неймана–Пирсона с той лишь разницей, что наиболее мощный критерий имеет наименьшую вероятность ошибки второго рода в классе всех статистических критериев с фиксированной вероятностью ошибки первого рода, а оптимальный критерий T^* — в классе S_k . Однако не следует забывать, что наиболее мощный критерий можно построить, точно зная распределения $F_H(u)$ и $F_{H'}(u)$, соответствующие гипотезам H и H' , а в рассматриваемой ситуации мы располагаем лишь обучающими выборками $x = (x_1, x_2, \dots, x_n) \in G$ и $x' = (x'_1, x'_2, \dots, x'_m) \in G'$.

Расширение критерия $T_{i_1 \dots i_k}$ на случай выборок произвольного объема $z = (z_1, z_2, \dots, z_r)$ можно осуществить по схеме, описанной выше. Однако в этом нет необходимости, так как критерий $T_{i_1 \dots i_k}$ применяется к выборкам z , полученным после первичной обработки информации. Это значит, что число $z = z_i$ является маркером, с помощью которого можно достаточно точно провести проверку гипотез H и H' .

Резюме

Проверка гипотез с помощью статистических критериев является одним из важнейших разделов современной статистической теории. Она находит широкое применение в разных областях доказательной медицины, в частности в дифференциальной диагностике заболеваний. Основное внимание в главе уделено нерандомизированным критериям, поскольку рандомизированные критерии используются лишь в теоретических исследованиях. В общем случае задача проверки гипотез ставится следующим образом: по имеющейся выборке $x = (x_1, x_2, \dots, x_n)$ принять одну из двух гипотез: H — выборка принадлежит генеральной совокупности G с функцией распределения $F(u_1, u_2, \dots, u_n)$ (основная гипотеза) и H' — выборка принадлежит генеральной совокупности G' с функцией распределения $G(u_1, u_2, \dots, u_n)$ (альтернативная конкурирующая гипотеза). Эта задача решается так: в выборочном пространстве выбирается критическая область V ; если x попадает в область V , то принимается гипотеза H' , а гипотеза

H отклоняется; в противном случае принимается гипотеза H , а гипотеза H' отклоняется. Выбор области V осуществляется так, чтобы вероятность ошибки первого рода $p(H'|H)$ не превышала критического уровня α , а вероятность ошибки второго рода достигала минимального значения. Если гипотез не две, а несколько $H_1, H_2, \dots, H_k$, то выборочное пространство R^n разбивается на k частей так, чтобы сумма вероятностей ошибок, возникающих при проверке гипотез, была минимальной, и гипотеза H_i принимается, если $x \in V_i$.

В том случае, когда вероятность ошибок первого или второго рода превышает заданный критический уровень γ , вводится процедура непринятия решения. Эту процедуру можно выбрать таким образом, что $p(H'|H) \leq \gamma$ и $p(H|H') \leq \gamma$.

Индивидуальные статистические критерии применяются тогда, когда ошибка при проверке гипотезы полностью определяется значениями диагностического показателя.

2.11. Меры близости между функциями распределения и выборками. Непараметрические критерии эквивалентности генеральных совокупностей

2.11.1. Меры близости и метрики

Рассмотрим две пары функций распределения $F_1(u)$ и $F_2(u)$, а также $\tilde{F}_1(u)$ и $\tilde{F}_2(u)$, представленные на рис. 2.7 и 2.8. Интуитивно ясно, что функции распределения $F_1(u)$ и $F_2(u)$ значительно отличаются, в то время как функции распределения $\tilde{F}_1(u)$ и $\tilde{F}_2(u)$ являются близкими. Для того чтобы объективно оценить близость функций распределения $F_1(u)$ и $F_2(u)$, необходимо ввести меру близости $\delta(F_1, F_2)$. Этот показатель характеризует расстояние между функциями распределения. Для любых функций $F_1(u)$ и $F_2(u)$ мера близости $\delta(F_1, F_2)$ является неотрицательной величиной. Иногда меру близости выбирают так, что она может принимать любые неотрицательные значения, а иногда ее конструируют так, чтобы она изменялась в диапазоне от нуля до единицы. В последнем случае мера близости в окрестности единицы означает, что функции распределения почти совпадают, а в окрестности нуля — совершенно отличаются.

Меру близости между функциями распределения можно построить с помощью метрики, т.е. функции $\rho(F_1, F_2)$, удовлетворяющей четырем условиям:

- 1) для любых $F_1(u)$ и $F_2(u)$ выполняется неравенство $\rho(F_1, F_2) \geq 0$;
- 2) $\rho(F_1, F_2) = 0$ тогда и только тогда, когда $F_1(u) = F_2(u)$ при всех значениях u ;
- 3) $\rho(F_1, F_2) = \rho(F_2, F_1)$ (условие симметрии);
- 4) $\rho(F_1, F_2) \leq \rho(F_1, F_3) + \rho(F_3, F_2)$ для любых функций F_1, F_2, F_3 .

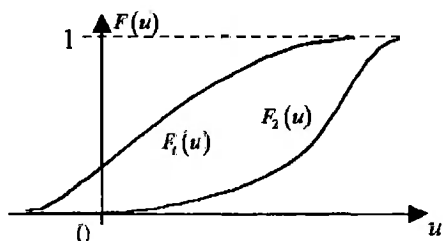


Рис. 2.7. Функции распределения двух "далеких" генеральных совокупностей

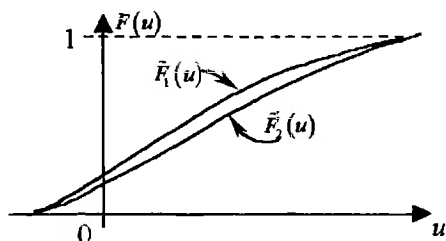


Рис. 2.8. Функции распределения "близких" генеральных совокупностей

Значение функции $\rho(F_1, F_2)$ для конкретной пары функций F_1 и F_2 называется *расстоянием* между этими функциями. Произвольное множество $\mathcal{F}$, в котором для любых двух ее элементов F_1 и F_2 определена метрика $\rho(F_1, F_2)$, называется *метрическим пространством*.

Пусть множество $\mathcal{F}$ — совокупность всех функций распределения $F(u)$, заданных на числовой прямой R^1 . Предположим, что в множестве $\mathcal{F}$ определена метрика $\rho(F_1, F_2)$. В таком случае ее можно считать мерой близости между функциями F_1 и F_2 . Интуитивно ясно, что для мало отличающихся функций F_1 и F_2 расстояние между ними невелико. В качестве примеров метрики можно указать, прежде всего, на чебышевскую метрику, представляющую собой максимальную разность между двумя функциями:

$$\rho_c(F_1, F_2) = \max_{-\infty < u < +\infty} |F_1(u) - F_2(u)|,$$

и лебегову метрику, характеризующую среднеквадратичное отклонение одной функции от другой:

$$\rho_L(F_1, F_2) = \sqrt{\int_{-\infty}^{+\infty} (F_1(u) - F_2(u))^2 du}.$$

Пусть F_1 и F_2 — функции распределения, а f_1 и f_2 — соответствующие плотности вероятности. Мера близости $\delta(F_1, F_2)$ между функциями распределения F_1 и F_2 называется *тонкой*, если из того, что величина $\delta(F_1, F_2)$ является малой, следует, что $\rho_L(f_1, f_2)$, т.е. из близости между функциями распределения по мере δ следует близость между плотностями вероятностей по чебышевской метрике (рис. 2.9 и 2.10). В противном случае мера близости называется *грубой*.

Нетрудно заметить, что чебышевская и лебегова метрики являются грубыми.

Различие между тонкой и грубой мерами близости можно продемонстрировать следующим примером. Предположим, что мы исследуем количественный показатель γ , характеризующий клетки (или ядра клеток) ткани, в которой по некоторым данным может существовать очаг злокачественных клеток. Допустим, что нам известно распределение показателя γ в здоровых тканях. Если распределение показателя γ сильно отличается от его эталонного распределения в здоровых тканях, то никаких проблем не возникнет. Однако, если отличие наблюдается в узком диапазоне значений, то грубые метрики его не выявят.

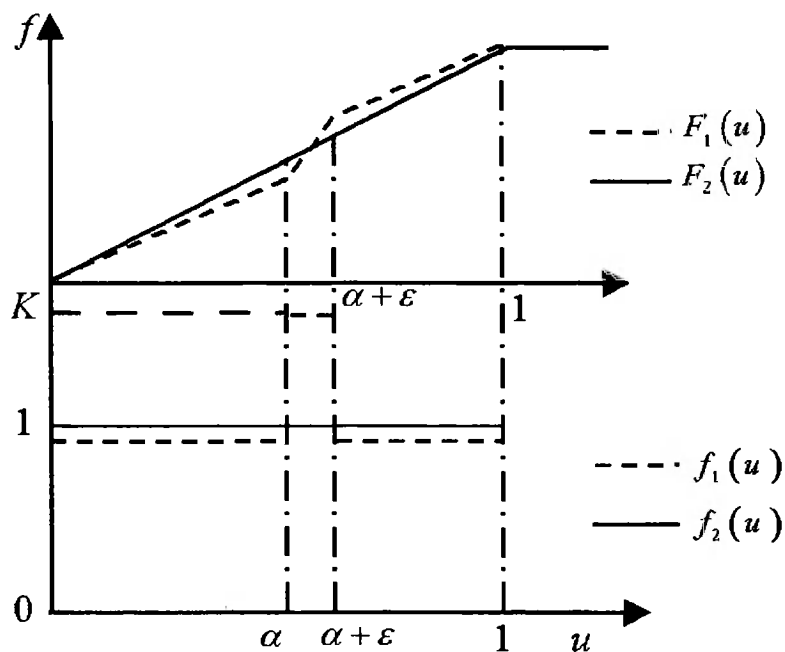


Рис. 2.9. Пример "тонкого" отличия между функциями распределения (непрерывное изменение)

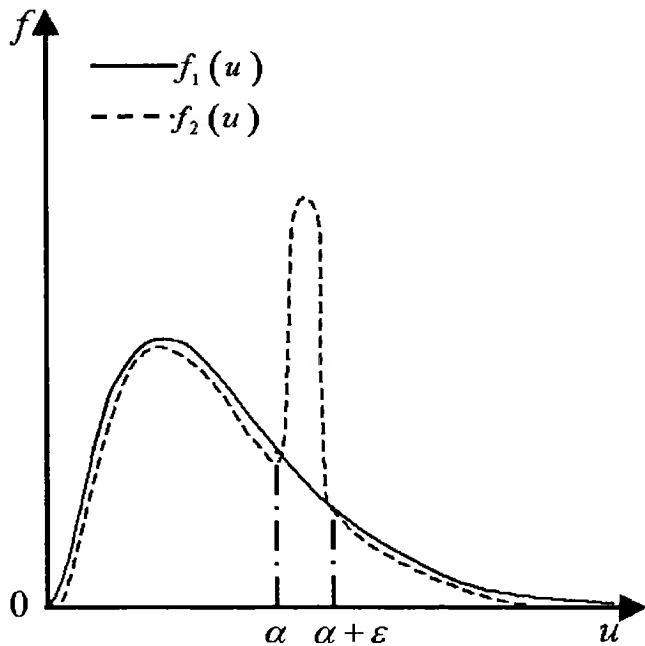


Рис. 2.10. Пример "тонкого" отличия между функциями распределения (скачок)

Типичным примером тонкой метрики на множестве функций распределения с ограниченной производной является C_1 -метрика:

$$\rho_{C_1}(F_1, F_2) = \max_{-\infty < u < +\infty} |F_1(u) - F_2(u)| + \max_{-\infty < u < +\infty} |f_1(u) - f_2(u)|.$$

Как видим, использование метрики сопряжено с большими трудностями, поскольку для их применения необходимо знать достаточно точные оценки функций распределения и плотностей вероятности. В данном разделе мы предлагаем новую меру близости между эмпирическими (выборочными) функциями распределения, основанную на доверительных границах для основной распределенной массы значений генеральной совокупности, вычисленных с помощью порядковых статистик. Поскольку эта мера близости вычисляется на основе выборок, мы называем ее также мерой близости между выборками. Так как выборки случайны, то и мера близости между ними является случайной величиной. Для нее можно найти доверительные границы, соответствующие заданному уровню значимости в случаях, когда основная (нулевая) гипотеза о равенстве гипотетических функций распределения может быть как правильной, так и ошибочной. На основании этой меры близости построен статистический критерий для проверки гипотезы о равенстве гипотетических функций распределения. Как показывают компьютерные эксперименты, эта мера близости лучше оценивает различия между функциями распределения, чем известные статистики Колмогорова–Смирнова и Вилкоксона.

2.11.2. Введение

Пусть $x = (x_1, x_2, \dots, x_n)$ и $x' = (x'_1, x'_2, \dots, x'_m)$ — две выборки из генеральных совокупностей G и G' соответственно и $z = (z_1, z_2, \dots, z_k)$ — выборка, принадлежащая одной из этих двух генеральных совокупностей. Полагаем, что все выборки получены путем простого случайного выбора. Необходимо идентифицировать генеральную совокупность, из которой взята выборка z .

Эту проблему можно решать с помощью классических непараметрических критериев Колмогорова–Смирнова, Вилкоксона или любого другого непараметрического двухвыборочного критерия [42]. Однако при использовании этих критериев применяются односторонние доверительные границы, соответствующие заданному уровню значимости (например, 5%-ному), что может привести к неопределенности и непринятию решения. Действительно, предположим, что мы используем критерий Колмогорова–Смирнова при сравнении выборки z с x и x' . Для этого вычисляется статистики Колмогорова–Смирнова $\rho(z, x)$ и $\rho(z, x')$, а затем находятся доверительные границы $t_\beta(k, n)$ и $t_\beta(k, m)$. Если $\rho(z, x) \geq t_\beta(k, n)$ и $\rho(z, y) < t_\beta(k, m)$, то считается, что $z \in G'$; если же $\rho(z, x) < t_\beta(k, n)$ и $\rho(z, y) \geq t_\beta(k, m)$, то $z \in G$; в случае остальных возможных неравенств возникает неопределенность и никакого решения не принимается. Другим существенным недостатком односторонних непараметрических критериев является тот факт, что вероятность попадания статистики $\rho(z, x)$ в доверительный интервал $[0, t_\beta(k, n)]$ (т.е. его доверительный уровень) можно точно определить лишь при

условии, когда $z \in G$ (гипотеза H); в противном случае эта вероятность может принимать любые значения от 0 до 1 и будет неизвестной. В связи с этим даже в случае принятия решения невозможно оценить вероятности ошибок 2-го рода, возникающие при использовании этого критерия. Эта ситуация характерна для любых односторонних непараметрических критериев.

Основная цель этой работы состоит в определении меры близости (p -статистики) $\rho_p(z, x)$ между выборками z и x , для которой можно построить двусторонний доверительный интервал $[\rho_p^{(l)}, \rho_p^{(w)}]$, соответствующий заданному уровню значимости, причем этот уровень значимости не зависит от истинности или ложности гипотезы H . Кроме того, на основании этой меры близости строится непараметрический критерий эквивалентности генеральных совокупностей, причем в некоторых практически важных случаях можно оценить его вероятность ошибки 1-го и 2-го рода.

2.11.3. Доверительные интервалы и уровни значимости

Пусть $x = (x_1, \dots, x_n)$ — выборка из генеральной совокупности G и p — некоторый известный или неизвестный показатель, значения которого могут зависеть от выборки x . Рассмотрим две непрерывные функции $a(u_1, \dots, u_n)$ и $b(u_1, \dots, u_n)$ от n переменных $u_1, \dots, u_n$, удовлетворяющие неравенству

$$a(u_1, \dots, u_n) \leq b(u_1, \dots, u_n) \quad \forall (u_1, \dots, u_n) \in R^n.$$

Случайный интервал $(a(u_1, \dots, u_n), b(u_1, \dots, u_n)) = (a, b)$ называется доверительным интервалом для p , соответствующим уровню значимости β , если

$$P(p \in (a, b)) = 1 - \beta, \quad (0 \leq \beta \leq 1);$$

при этом числа $a = a(x_1, \dots, x_n)$, $b = b(x_1, \dots, x_n)$ называются доверительными границами для p , отвечающие уровню значимости β . Это понятие можно обобщить на случай последовательности показателей $p_1, p_2, \dots, p_k, \dots$

Определение 2.2.

Интервалы $(a_k, b_k) = (a_k(x_1, \dots, x_k), b_k(x_1, \dots, x_k))$, $k = 1, 2, \dots$, называются асимптотическими интервалами для показателей p_k , $i = 1, 2, \dots, k, \dots$, отвечающими уровню значимости β , если

$$\lim_{k \rightarrow \infty} P(p_k \in (a_k(x_1, \dots, x_k), b_k(x_1, \dots, x_k))) = 1 - \beta, \quad (2.26)$$

а концы этих интервалов $a_k(x_1, \dots, x_k)$ и $b_k(x_1, \dots, x_k)$ называются асимптотическими доверительными границами.

Замечание. Для величины β мы будем также использовать термин “асимптотический уровень значимости последовательности (a_k, b_k) ”.

В частности, когда все показатели $p_1, p_2, \dots, p_k, \dots$ одинаковы, т.е. $p_k = p \quad \forall k = 1, 2, \dots$, то интервал (a_k, b_k) называется асимптотическим доверительным интервалом показателя p , а величина β в формуле (2.26) — асимптотическим уровнем значимости интервала (a_k, b_k) .

Во многих практически важных частных случаях вычислить точное значение уровня значимости β_k доверительного интервала (a_k, b_k) довольно сложно, а определить асимптотический уровень значимости β , используя соответствующие предельные теоремы, достаточно просто. Если $\beta = \lim_{k \rightarrow \infty} \beta_k$, то β можно принять за приближенное значение истинного уровня значимости β_k доверительных интервалов (a_k, b_k) : $\beta_k = \beta$. В практике статистических вычислений при $n \geq 30$ величина β является хорошим приближением для β_k (см., например, [42], с. 410, табл. 7). В связи с этим мы будем также называть число β приближенным уровнем значимости интервалов (a_k, b_k) .

2.11.4. Мера близости между выборками (p -статистика)

Обозначим через H гипотезу о равенстве непрерывных функций распределения $F_G(u)$ и $F_{G'}(u)$ генеральных совокупностей G и G' соответственно. Пусть $x = (x_1, \dots, x_n) \in G$ и $x' = (x'_1, \dots, x'_m) \in G'$, $x_{(1)} \leq \dots \leq x_{(n)}$, $x'_{(1)} \leq \dots \leq x'_{(m)}$ — порядковые статистики. Предположим, что $F_G(u) = F_{G'}(u)$. Обозначим через $A_j^{(k)}$, $k = 1, 2, \dots, m$, случайное событие, состоящее в том, что x'_k попадает в интервал $(x_{(j)}, x_{(j+1)})$, т.е. $A_j^{(k)} = \{x'_k \in (x_{(j)}, x_{(j+1)})\}$. Как известно [84], когда $F_G(u) = F_{G'}(u)$ (т.е. $G = G'$), вероятность этого события вычисляется по формуле:

$$p(A_j) = p(x'_k \in (x_{(j)}, x_{(j+1)})) = p_{ij} = \frac{j-i}{n+1}. \quad (2.27)$$

Положим

$$\begin{aligned} p_{ij}^{(1)} &= \frac{h_{ij}^{(n)}m + g^2/2 - g\sqrt{h_{ij}^{(n)}(1-h_{ij}^{(n)})m + g^2/4}}{m + g^2} \\ p_{ij}^{(2)} &= \frac{h_{ij}^{(n)}m + g^2/2 - g\sqrt{h_{ij}^{(n)}(1-h_{ij}^{(n)})m + g^2/4}}{m + g^2} \end{aligned} \quad (2.28)$$

где $h_{ij}^{(n)}$ — частота появления события $A_j^{(n)}$ в m испытаниях, величина $g = 3$.

Обозначим через N все доверительные интервалы $I_{ij}^{(n,m)} = (p_{ij}^{(1)}, p_{ij}^{(2)})$ ($N = n(n-1)/2$) и L — количество тех интервалов $I_{ij}^{(n,m)}$, которые содержат вероятности $p_{ij}^{(n)}$. Положим $h^{(n,m)} = \rho(F^*, F'') = \rho(x, x') = L/N$. Поскольку $h^{(n,m)}$ — частота случайного события $B = \{p_{ij}^{(n)} \in I_{ij}^{(n,m)}\}$, имеющего вероятность $p(B) = 1 - \beta$, то, полагая $h_{ij}^{(n)} = h^{(n,m)}$, $m = N$ и $g = 3$ в формулах (2.28), мы получим доверительный интервал $I^{(n,m)} = (p^{(1)}, p^{(2)})$ для вероятности $p(B)$. Статистика $h^{(n)}$ называется p -статистикой. Она является мерой близости $\rho(x, x')$ между выборками x и x' . Доверительные интервалы $I_{ij}^{(n,m)} = (p_{ij}^{(1)}, p_{ij}^{(2)})$ и $I^{(n,m)} = (p^{(1)}, p^{(2)})$ мы будем называть интервалами, построенными по правилу 3σ .

2.11.5. Оценка асимптотического доверительного уровня значимости

Всюду в дальнейшем будем предполагать, что выборки x и x' имеют одинаковый объем, равный n . Обоснование корректности предложенной меры близости связано с проблемой вычисления уровня значимости соответствующих доверительных интервалов $I_{ij}^{(n)}$ и $I'^{(n)}$. Для схемы испытаний Бернулли асимптотические уровни значимости этих интервалов при $n \rightarrow \infty$ можно вычислить на основании соответствующих предельных теорем, однако в нашем случае эти теоремы неприменимы, поскольку исследуемая схема испытаний не является схемой Бернулли, так как случайные события $A_{ij}^{(k)}, k=1, 2, \dots, n$, не являются независимыми. Схема испытаний, при которой могут появляться события $A_{ij}^{(k)}, k=1, 2, \dots, n$, в случае истинности гипотезы H носит название обобщенной схемы Бернулли [86, 87]. Если же гипотеза H неверна, указанная схема испытаний называется модифицированной схемой Бернулли [10, 86, 87]. В общем случае, когда может быть истинной любая гипотеза, как $F_{\sigma}(u) = F_{\sigma}(u)$, так и $F_{\sigma}(u) \neq F_{\sigma}(u)$, эта схема испытаний, согласно терминологии С. Котца и Н. Джонсона [10], называется МП-схемой (MP-scheme).

Найдем оценку для асимптотического уровня значимости доверительных интервалов $I_{ij}^{(n)}$.

Теорема 2.1.

Если в обобщенной схеме испытаний Бернулли, во-первых, $n = m$; во-вторых, $0 < \lim_{n \rightarrow \infty} p_q^{(n)} = p_0 < 1$ и в-третьих, $0 < \lim_{n \rightarrow \infty} \frac{i}{n+1} = p^* < 1$, то асимптотический уровень значимости β последовательности доверительных интервалов $I_{ij}^{(n)}$ для вероятностей $p_q^{(n)} = p_{ij}^{(n)} = p(A_{ij}^{(n)})$, построенных по правилу 3σ , не превышает 0,05.

Доказательство. Анализ вывода формул (2.28) показывает, что уровень значимости β_n интервала $I_{ij}^{(n)}$ равен уровню значимости $\tilde{\beta}_n$ доверительного интервала

$$\tilde{I}_n = \left(h_{ij}^{(n)} - 3\sqrt{\frac{p_q^{(n)}(1-p_q^{(n)})}{n}}, h_{ij}^{(n)} + 3\sqrt{\frac{p_q^{(n)}(1-p_q^{(n)})}{n}} \right) \quad (2.29)$$

для вероятности $p_q^{(n)} = p_{ij}^{(n)}$ при условии, что $E(h_{ij}^{(n)}) = p_{ij}^{(n)}$, поэтому

$$\beta = \lim_{n \rightarrow \infty} P(p_q^{(n)} \notin \tilde{I}_n).$$

В работе [86] показано, что в этом случае математическое ожидание частоты $h_{ij}^{(n)}$ равно $p_{ij}^{(n)}$:

$$E(h_{ij}^{(n)}) = p_{ij}^{(n)} = p_q^{(n)},$$

а ее дисперсия определяется формулой

$$D(h_{ij}^{(n)}) = \frac{p_{ij}^{(n)}(1-p_{ij}^{(n)})}{n} \frac{2n+1}{n+2} = \alpha_n \sigma_{ij,B}^{(n)},$$

$$\text{где } \alpha_n = \frac{2n+1}{n+2} < 2, \quad \sigma_{ij,B}^{(n)} = \frac{p_{ij}^{(n)}(1-p_{ij}^{(n)})}{n}.$$

Легко видеть, что доверительный уровень интервала $\bar{I}_n$ совпадает с доверительным уровнем интервала $\hat{I}_n$:

$$\hat{I}_n = (p_q^{(n)} - 3\sigma_{ij,B}^{(n)}, h_{ij} + 3\sigma_{ij,B}^{(n)}).$$

Поскольку $\alpha_n < 2$,

$$\begin{aligned} \hat{I}_n &= \left(p_q^{(n)} - \frac{3}{\sqrt{2}} \sqrt{2} \sigma_{ij,B}^{(n)}, p_q^{(n)} + \frac{3}{\sqrt{2}} \sqrt{2} \sigma_{ij,B}^{(n)} \right) = \\ &= (p_q^{(n)} - 2,1213 \sqrt{2} \sigma_{ij,B}^{(n)}, p_q^{(n)} + 2,1213 \sqrt{2} \sigma_{ij,B}^{(n)}) \Rightarrow \\ &\Rightarrow (h_{ij} - 2,1213 \sigma(h_{ij}^{(n)}), h_{ij} + 2,1213 \sigma(h_{ij}^{(n)})) \Rightarrow \\ &\Rightarrow (h_{ij} - 1,96 \sigma(h_{ij}^{(n)}), h_{ij} + 1,96 \sigma(h_{ij}^{(n)})) = \bar{I}_n. \end{aligned} \quad (2.30)$$

В силу условий теоремы, а также предельной теоремы, доказанной в [87], в обобщенной схеме Бернулли случайная величина $h_{ij}^{(n)}$ является асимптотически нормальной с параметрами $E(h_{ij}^{(n)}) = p_q^{(n)}$ и $D(h_{ij}^{(n)}) = \sigma^2(h_{ij}^{(n)})$, поэтому асимптотический уровень значимости интервала $\bar{I}_n$ равен 0,05. Поскольку $\bar{I}_n \subset \hat{I}_n$, асимптотический уровень значимости интервала $\bar{I}_n$, а значит, и интервалов $\bar{I}_n$ и $I_{ij}^{(n)}$ не превышает 0,05.

Теорема доказана.

Рассмотрим теперь проблему вычисления уровня значимости доверительного интервала $I^{(n)}$, построенного по правилу 3 σ . Подробное описание и обоснование правила 3 σ можно найти в книге [98]. Рассмотрим вначале случай, когда справедлива гипотеза H , т.е. $F_G(u) = F_G(u)$.

Покажем, что эту проблему можно свести к исследованию случайных величин ξ и ξ' , равномерно распределенных на отрезке $[0, 1]$. Действительно, пусть ξ — произвольная случайная величина из генеральной совокупности G со строго возрастающей непрерывной функцией распределения $F_G(u)$, $(\xi_1, \dots, \xi_n)$ — выборка из G , а $\xi_{(1)} \leq \dots \leq \xi_{(n)}$ — ее вариационный ряд. Рассмотрим случайную величину $\eta = F_G(\xi)$. Нетрудно заметить, что

$$F_\eta(u) = P(F_G(\xi)) = P(\xi < F_G^{-1}(u)) = F_G[F_G^{-1}(u)] = u.$$

Следовательно, η имеет равномерное распределение на отрезке $[0, 1]$:

$$F_{\eta}(u) = \begin{cases} u, & \text{если } u \in [0, 1], \\ 0, & \text{если } u = 0, \\ 1, & \text{если } u \geq 1 \end{cases}.$$

Кроме того, в силу монотонности функции распределения $F_G(u)$

$$\eta_{(i)} = F_G(\xi_{(i)}),$$

где $\eta_{(i)}$ — порядковые статистики, построенные по выборке $\eta_i = F_G(\xi_i)$. Далее, случайная величина $\eta' = F_G(\eta) = F_G(\eta')$ также имеет равномерное распределение на отрезке $[0, 1]$, причем случайные события $\xi'_k \in (\xi_{(i)}, \xi_{(j)})$ и $\eta'_k \in (\eta_{(i)}, \eta_{(j)})$ эквивалентны, поскольку функция $F_G(u)$ является монотонно возрастающей. Отсюда следует, что частота $h_{ij}^{(\xi)}$ случайного события, состоящего в том, что $\{\xi'_k \in (\xi_{(i)}, \xi_{(j)})\}$, совпадает с частотой $h_{ij}^{(\eta)}$ случайного события $\{\eta'_k \in (\eta_{(i)}, \eta_{(j)})\}$, поэтому эквивалентны между собой события $\{p_{ij} \in (p_{ij}^{(1)}(\xi), p_{ij}^{(2)}(\xi))\}$ и $\{p_{ij} \in (p_{ij}^{(1)}(\eta), p_{ij}^{(2)}(\eta))\}$, где $(p_{ij}^{(1)}(\xi), p_{ij}^{(2)}(\xi))$ и $(p_{ij}^{(1)}(\eta), p_{ij}^{(2)}(\eta))$ — доверительные границы для p_{ij} , построенные с помощью формул (2.28) при $g=3$ на основании частот $h_{ij}^{(\xi)}$ и $h_{ij}^{(\eta)}$. Таким образом, мера близости между выборками $(\xi_1, \dots, \xi_n)$ и $(\xi'_1, \dots, \xi'_n)$ совпадает с мерой близости между выборками $(\eta_1, \dots, \eta_n)$ и $(\eta'_1, \dots, \eta'_n)$. Следовательно, уровень значимости доверительного интервала $I^{(n)} = (p^{(1)}, p^{(2)})$ для меры близости можно вычислить с помощью машинного моделирования, используя для этого выборки случайных величин, равномерно распределенных на отрезке $[0, 1]$. Результаты машинного моделирования показывают, что для выборок объема $n = 100$ среднее значение доверительного уровня интервала $I^{(100)}$, вычисленное для 30 выборок, равно 0,997, что превышает, как и следовало ожидать на основании теоремы 2.1, величину 0,95. Кроме того, не было отмечено ни одного случая выхода числа 0,95 за пределы интервала $I^{(100)}$.

Перейдем теперь к оценке оценки уровня значимости доверительного интервала $I^{(n)} = (p^{(1)}, p^{(2)})$, когда гипотеза H неверна (т.е. $F_G(u) \neq F_G(u)$).

Теорема 2.2.

Если выборки $x = (x_1, \dots, x_n) \in G$ и $x' = (x'_1, \dots, x'_n) \in G'$ имеют одинаковый объем, то асимптотический уровень значимости интервала $I^{(n)} = (p^{(1)}, p^{(2)})$, построенный по правилу 3 σ при $g=3$ с помощью формул (2.28), не превосходит 0,05.

Доказательство. Рассмотрим доверительные интервалы $I_{ij}^{(n)} = (p_{ij}^{(1)}, p_{ij}^{(2)})$, $i < j$, построенные по частоте $h_{ij}^{(n)}$ случайного события $\{x'_k \in (x_{(i)}, x_{(j)})\}$ с помощью правила 3 σ . Поскольку $F_G(u) \neq F_G(u)$, вероятность $P(B_{ij}^{(n)})$ случайного события

$B_{ij}^{(n)} = \left\{ p_{ij}^{(n)} = \frac{j-i}{n+1} \in I_{ij}^{(n)} \right\}$, вообще говоря, не равна 0,95 и может зависеть от индексов i и j . Введем случайные величины

$$\delta_{ij}^{(n)} = \begin{cases} 1, & \text{если } p_{ij}^{(n)} \in I_{ij}, \\ 0, & \text{если } p_{ij}^{(n)} \notin I_{ij}, \end{cases}$$

$$z^{(n)} = \frac{1}{N} \sum_{i < j} \delta_{ij}^{(n)} = h^{(n)}, \quad N = \frac{n(n-1)}{2}.$$

Тогда

$$p^{(n)} = E(z^{(n)}) = \frac{1}{N} \sum_{i < j} p(B_{ij}^{(n)}). \quad (2.31)$$

Пронумеруем случайные величины $\delta_{ij}^{(n)}$ произвольным образом и обозначим полученное множество $X = (X_1, X_2, \dots, X_N)$, а совместную функцию распределения координат случайного вектора X — через $F(u_1, u_2, \dots, u_N)$. В работе [103] был предложен следующий метод получения выборки из множества X : выберем наугад один из элементов множества X и обозначим его через γ_1 ; затем из множества $X \setminus \{\gamma_1\}$ оставшихся элементов выбираем наугад элемент γ_2 и т.д. Таким образом, мы имеем серию испытаний, состоящую из N случайных выборок без возвращения элементов множества X , содержащихся в урне. Получившиеся при этом многомерную случайную величину $\gamma^{(N)} = (\gamma_1, \gamma_2, \dots, \gamma_N)$ будем называть индуцированной выборкой, полученной с помощью урновой модели. Как известно [103], функция совместного распределения $F_{\gamma^{(N)}}(u_1, u_2, \dots, u_N)$ индуцированной выборки $\gamma^{(N)}$ имеет вид

$$F_{\gamma^{(N)}}(u_1, u_2, \dots, u_N) = \frac{1}{N!} \sum_{(i_1, i_2, \dots, i_N) \in G_N} F(u_{i_1}, u_{i_2}, \dots, u_{i_N}), \quad (2.32)$$

где G_N — группа перестановок чисел $(1, 2, \dots, N)$. Из формулы (2.31) вытекает, что $F_{\gamma^{(N)}}(u_1, u_2, \dots, u_N)$ является симметрической функцией, так что элементы γ_k , $k = 1, 2, \dots, N$, индуцированной выборки являются симметрично зависимыми случайными величинами с маргинальной функцией распределения

$$F_{\gamma_k}(u) = \frac{1}{N!} [F(u, \infty, \dots, \infty)(n-1)! + F(\infty, u, \infty, \dots, \infty)(n-1)! + \dots + F(\infty, \dots, \infty, u)(n-1)!] =$$

$$= \frac{1}{N} [F_1(u) + \dots + F_N(u)].$$

Отсюда следует, что все маргинальные функции распределения $F_{\gamma_k}(u)$ одинаковы и имеют вид

$$F_{\gamma_k}(u) = \begin{cases} 0, & \text{если } u \leq 0, \\ q^{(n)} = 1 - p^{(n)}, & \text{если } 0 < u < 1, \\ p^{(n)}, & \text{если } \geq 1, \end{cases} \quad (2.33)$$

где $p^{(n)} = E(z)$ определяется по формуле (2.31). Обозначим через B случайное событие, состоящее в том, что вероятность $p_{ij}^{(n)} = \frac{j-1}{n+1}$ принадлежит доверительному интервалу $I_{ij}^{(n)}$ при случайном выборе этого интервала (или вероятности $p_{ij}^{(n)}$). Очевидно, что событие B происходит тогда и только тогда, когда $\delta_{ij}^{(n)} = \gamma_i = 1, k = 1, 2, \dots, N$, поэтому индуцированную выборку $\gamma^{(n)}$ можно интерпретировать как схему испытаний, в результате которых событие B может появиться с вероятностью $p(B) = p^{(n)}$. Легко видеть, что частота h_B этого события совпадает с $h = \rho(x, x')$, так что

$$h = \frac{1}{N} \sum_{k=1}^N \gamma_k$$

и

$$E(h) = \frac{1}{N} \sum_{k=1}^N E(\gamma_k) = p^{(n)},$$

$$D(h) = \frac{1}{N^2} \left(\sum_{k=1}^N D(\gamma_k) + \sum_{k \neq s} K(\gamma_k, \gamma_s) \right). \quad (2.34)$$

Легко видеть, что $D(\gamma_k) = p^{(n)} q^{(n)}$, а коэффициент ковариации $K(\gamma_k, \gamma_s)$ не зависит от k и s : $K(\gamma_k, \gamma_s) \equiv C$. Действительно, при любых k и s двумерная функция распределения γ_k и γ_s не зависит от k и s , так как в силу симметричности функций $F_{\gamma^{(n)}}(u_1, \dots, u_N)$ имеем

$$F_{\gamma_k \gamma_s}(u_k, u_s) = F_{\gamma^{(n)}}(\infty, \dots, \infty, u_k, \infty, \dots, \infty, u_s, \infty, \dots, \infty) = F_{\gamma^{(n)}}(u_k, u_s, \infty, \dots, \infty) = F_0(u_k, u_s).$$

Следовательно, коэффициент ковариации $K(\gamma_k, \gamma_s)$, который определяется функцией $F_{\gamma_k \gamma_s}(u_k, u_s)$, принимает постоянное значение и не зависит от γ_k и γ_s . Таким образом,

$$\frac{1}{N^2} \sum_{k \neq s} K(\gamma_k, \gamma_s) = \frac{N-1}{N} C,$$

где $C = K(\gamma_k, \gamma_s)$.

Далее,

$$C = K(\gamma_k, \gamma_s) = E(\gamma_k \gamma_s) - (p^{(n)})^2,$$

так что смешанный момент второго порядка $E(\gamma_k \gamma_s)$ тоже является постоянной величиной. Поскольку величины C и $p^{(n)}$, входящие в формулу (2.33), неизвестны, заменим их оценками

$$p^{(n)} = h,$$

$$C = \frac{1}{N(N-1)} \sum_{k \neq s} \gamma_k \gamma_s - h^2 = \frac{L(L-1)}{N(N-1)} - \frac{L^2}{N^2} = \frac{L(L-N)}{N^2(N-1)} = Q,$$

причем

$$|Q| = h \frac{N-L}{N(N-1)}.$$

При такой замене оценка практически совпадает с математическим ожиданием, если $N \geq 30$, т.е. $n \geq 9$. Кроме того, выборочное среднее является оптимальной оценкой в классе линейных оценок для симметрично зависимых случайных величин [122].

Следовательно,

$$\begin{aligned} D(h) &= \frac{p^{(n)}(1-p^{(n)})}{N} + \frac{N-1}{N} C \leq \frac{p^{(n)}(1-p^{(n)})}{N} + \frac{N-1}{N} |C| = \frac{h(1-h)}{N} + \frac{N-1}{N} |Q| = \\ &= \frac{h(1-h)}{N} + \frac{h}{N} \frac{N-L}{N} = \frac{h(1-h)}{N} + \frac{h(1-h)}{N} = 2 \frac{h(1-h)}{N}, \end{aligned}$$

так что

$$D(h) \leq 2 \frac{h(1-h)}{N}.$$

Используя вложения интервалов (2.30) и центральную предельную теорему для суммы симметрично зависимых случайных величин [122], получаем, что асимптотический уровень значимости доверительного интервала $I^{(n)}$ не превосходит 0,05.

Теорема доказана.

2.11.6. Сравнение p -статистики со статистиками Колмогорова–Смирнова и Вилкоксона

Для исследования эффективности p -статистики рассмотрим ее сравнение с наиболее распространенными статистиками, которые используются при построении порядковых критериев — статистикой Колмогорова–Смирнова и статистикой Вилкоксона — при вычислении меры близости между выборками из некоторых параметрических семейств распределений. Прежде всего, необходимо провести унификацию этих статистик, так чтобы они принимали свои значения из отрезка $[0, 1]$, причем их малым значениям должны соответствовать близкие выборки.

Для статистики Колмогорова–Смирнова (K-S)

$$d_{KS}(x, x') = \max_{t \in (-\infty, +\infty)} |F_x^*(t) - F_{x'}^*(t)|,$$

где $F_x^*(t)$ и $F_{x'}^*(t)$ — эмпирические функции распределения, построенные по выборкам x и x' соответственно, в такой унификации нет никакой необходимо-

сти, поскольку статистика Колмогорова–Смирнова принимает значения из отрезка $[0, 1]$ и ее малым значениям соответствуют близкие выборки.

Для U -статистики Вилкоксона такая унификация необходима, поскольку она принимает целые неотрицательные значения. Напомним определение статистики Вилкоксона: пусть $x = (x_1, x_2, \dots, x_g)$ и $x' = (x'_1, x'_2, \dots, x'_h)$ — выборки, полученные простым случайным выбором из генеральных совокупностей G и G' соответственно. Определим случайные величины

$$z_{ik} = \begin{cases} 1, & \text{если } x_i > x'_k, \\ 0, & \text{если } x_i \leq x'_k, \end{cases}$$

и

$$U = \sum_{i=1}^g \sum_{k=1}^h z_{ik}.$$

Положим $\hat{U} = 1/2 gh$ и $J = U - \hat{U}$. Для унификации статистики Вилкоксона введем случайную величину

$$d_U(x, x') = \frac{2}{gh} |J|,$$

которую мы будем называть унифицированной статистикой Вилкоксона и использовать в качестве меры близости между выборками x и x' . Легко видеть, что $0 \leq d_U(x, x') \leq 1$ и для выборок из одной генеральной совокупности $G = G'$ статистика $d_U(x, x')$ принимает значения, близкие к нулю.

Унификация p -статистики (P–K) проводится по формуле

$$d_p(x, x') = 1 - \rho(x, x').$$

Для сравнения эффективности указанных статистик мы рассмотрели выборки из генеральных совокупностей, которые имеют одинаковую дисперсию при различных математических ожиданиях, одинаковое математическое ожидание при различных значениях дисперсии, а также выборки из генеральных совокупностей, отличающихся как математическими ожиданиями, так и значениями дисперсии, и выборки из генеральных совокупностей, обладающих одинаковыми математическими ожиданиями и значениями дисперсии. Сравнивались выборки из параметрического семейства генеральных совокупностей G_α , обладающих гипотетическими функциями распределения $F_\alpha(u)$, $0 \leq \alpha \leq 1$, с выборкой из генеральной совокупности G_1 с распределением $F_1(u)$. Показателем эффективности статистики (меры близости) является значение α_0 параметра α , начиная с которого критерий, основанный на этой статистике, перестает различать выборки (генеральные совокупности). Введем следующие обозначения: $N(a, b)$ — нормальное распределение с математическим ожиданием a и дисперсией $b > 0$, $U(a, b)$ — равномерное распределение на отрезке $[a, b]$.

Рассмотрены следующие модели.

1. Одинаковые дисперсии при различных математических ожиданиях (рис. 2.11):

$$N(\alpha, 1) \text{ и } N(1, 1).$$

2. Одинаковые математические ожидания при различных значениях дисперсии (рис. 2.12):

$$N(0, \alpha) \text{ и } N(0, 1).$$

3. Различные математические ожидания и значения дисперсии:

- а) смеси распределений:

$$\alpha U(0,1) + (1-\alpha)U\left(\frac{1}{2}, \frac{3}{2}\right) \text{ и } U(0, 1) \text{ (рис. 2.13),}$$

$$\alpha U(0,1) + (1-\alpha)U\left(\frac{1}{6}, \frac{1}{2}\right) \text{ и } U(0, 1) \text{ (рис. 2.14),}$$

$$\alpha N(0,1) + (1-\alpha)N(1,1) \text{ и } N(0, 1) \text{ (рис. 2.15),}$$

- б) распределения одинакового типа:

$$N(\alpha, 1 + \alpha) \text{ и } N(1, 2) \text{ (рис. 2.16).}$$

4. Альтернативы Лемана на основе распределения $U(0, 1)$:

$$f_{\alpha}(u) = (1 + \alpha)F^{\alpha}(u)f(u), \quad -\frac{1}{2} \leq \alpha \leq 0 \text{ и } U(0, 1) \text{ (рис. 2.17),}$$

$$f_{\alpha}(u) = (1 + \alpha)F^{\alpha}(u)f(u), \quad -1 < \alpha \leq 0 \text{ и } U(0, 1) \text{ (рис. 2.18),}$$

где $f(u)$ — плотность вероятностей равномерного распределения на отрезке $[0, 1]$, $f_{\alpha}(u)$ — плотность вероятностей альтернативы Лемана.

5. Одинаковые математические ожидания и дисперсии:

$$\text{смесь распределений } \alpha N(0,1) + (1-\alpha)U\left(-\sqrt[3]{\frac{3}{2}}, \sqrt[3]{\frac{3}{2}}\right) \text{ и } N(0, 1) \text{ (рис. 2.19).}$$

В табл. 2.1 приведены значения порога чувствительности для рассмотренных выше критериев и генеральных совокупностей при 5%-ном уровне значимости. Результаты, приведенные в табл. 2.1, свидетельствуют о преимуществе Р-К критерия по сравнению с критериями Колмогорова-Смирнова и Вилкоксона. Однако для Р-К критерия наблюдается слабая чувствительность при значениях параметра α , близких к единице (т.е. когда выборки взяты из очень близких генеральных совокупностей), в то время как у критериев Колмогорова-Смирнова и Вилкоксона этот недостаток не так сильно выражен.

Таблица 2.1. Величины порога чувствительности для рассмотренных критериев и генеральных совокупностей при 5%-ном уровне значимости

Генеральные совокупности		Порог чувствительности критерия		
G_x	G_y	U	K-S	P-K
1 $N(\alpha, 1)$	$N(1, 1)$	0,9	0,7	0,8
2 $N(0, \alpha)$	$N(0, 1)$	—	0,2	0,7
3 $\alpha U(0,1) + (1-\alpha)U\left(\frac{1}{2}, \frac{3}{2}\right)$	$U(0, 1)$	0,9	0,7	0,9
4 $\alpha U(0,1) + (1-\alpha)U\left(\frac{1}{6}, \frac{1}{2}\right)$	$U(0, 1)$	—	0,2	0,8
5 $\alpha N(0,1) + (1-\alpha)N(1,1)$	$N(0, 1)$	0,9	0,5	0,8
6 $N(\alpha, 1 + \alpha)$	$N(1, 2)$	0,8	0,5	0,8
7 $f_\alpha(u) = (1 + \alpha)F^\alpha(u)f(u), -\frac{1}{2} \leq \alpha \leq 0$	$U(0, 1)$	0,8	0,4	0,7
8 $f_\alpha(u) = (1 + \alpha)F^\alpha(u)f(u), -1 < \alpha \leq 0$	$U(0, 1)$	0,9	0,7	0,9
9 $\alpha N(0,1) + (1-\alpha)U\left(-\sqrt[3]{\frac{3}{2}}, \sqrt[3]{\frac{3}{2}}\right)$	$N(0, 1)$	—	—	0,7

Примечание. Знак “—” означает отсутствие значимых отклонений при всех значениях параметра $0 \leq \alpha < 1$, так что в этом случае критерий не различает выборки.

На рис. 2.11–2.19 по оси абсцисс отложены значения параметра α , а по оси ординат — значения статистик, где РК-расстояние — унифицированная p -статистика, KS-расстояние — статистика Колмогорова–Смирнова, U-расстояние — унифицированная статистика Вилкоксона. Анализ рис. 2.11–2.19 показывает, что в области значимых отклонений p -статистика обладает более высокой эффективностью по сравнению с расстоянием Колмогорова–Смирнова и U-расстоянием. В связи с этим представляет интерес вычисление значений параметра α , начиная с которого наблюдаются значимые различия между выборками из генеральной совокупности с предельным распределением (при $\alpha = 1$) и выборками из генеральных совокупностей с распределениями, соответствующими параметрам α из интервала $[0, 1)$. Эти значения α мы будем называть *порогом чувствительности критерия*.

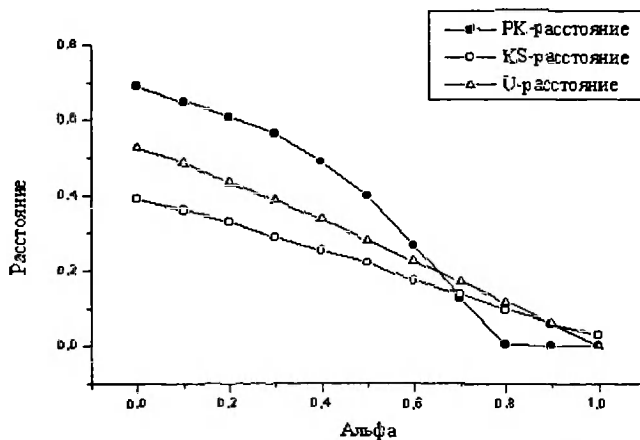


Рис. 2.11. Расстояния между выборкой из $N(1,1)$ и выборками из $N(\alpha,1)$

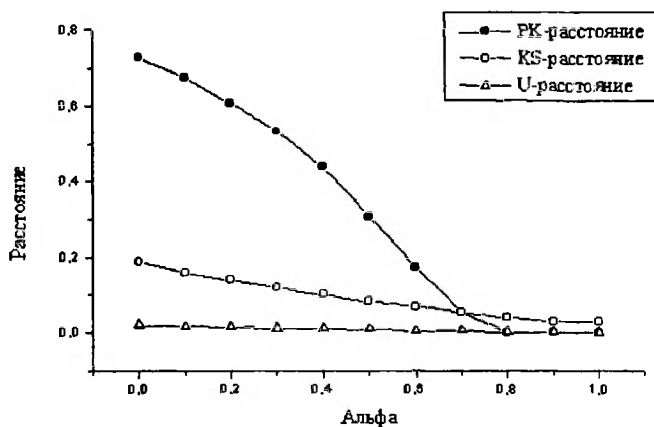


Рис. 2.12. Расстояния между выборкой из $N(0,1)$ и выборками из $N(0,\alpha)$

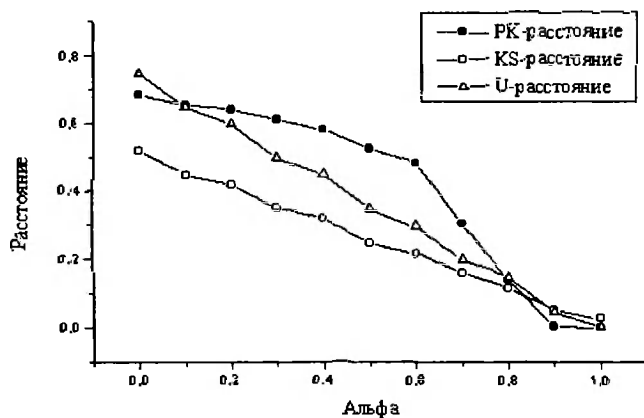


Рис. 2.13. Расстояния между выборкой из $U(0,1)$ и выборкой из смеси равномерных распределений $\alpha U(0,1) + (1-\alpha)U(0,5,1,5)$

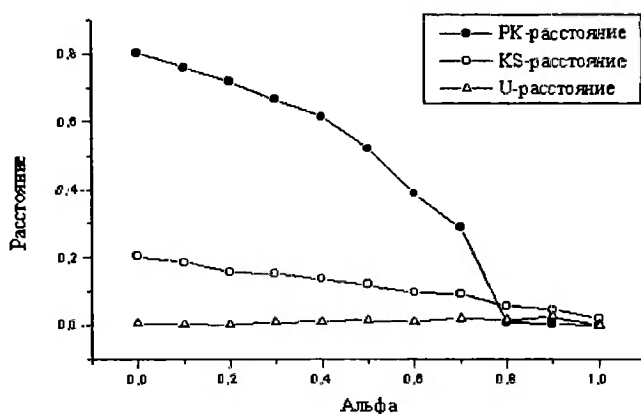


Рис. 2.14. Расстояния между выборкой из $U(0,1)$ и выборками из смеси равномерного и нормального распределений $\alpha U(0,1) + (1-\alpha)N(1/2, 1/6)$

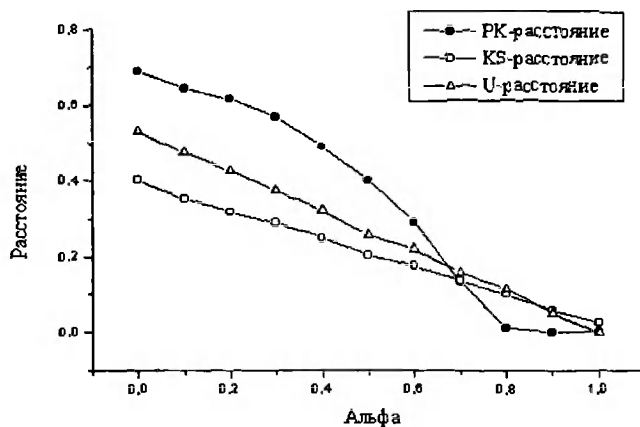


Рис. 2.15. Расстояния между выборкой из $N(0,1)$ и выборками из смеси $\alpha N(0,1) + (1-\alpha)N(1,1)$

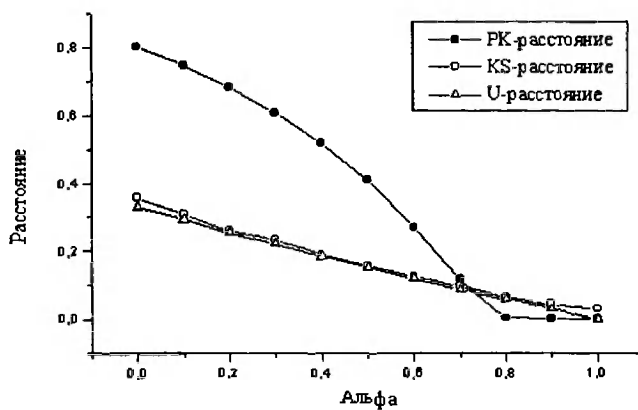


Рис. 2.16. Расстояния между выборкой из $N(1,2)$ и выборками из $N(\alpha, 1+\alpha)$

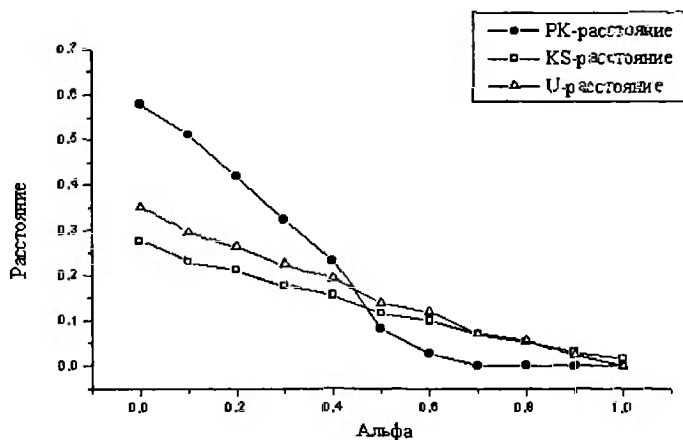


Рис. 2.17. Расстояния между выборкой из распределения $U(0,1)$ и выборками из семейства альтернатив Лемана на основе распределения $U(0,1)$ ($-0,5 \leq u \leq 0$)

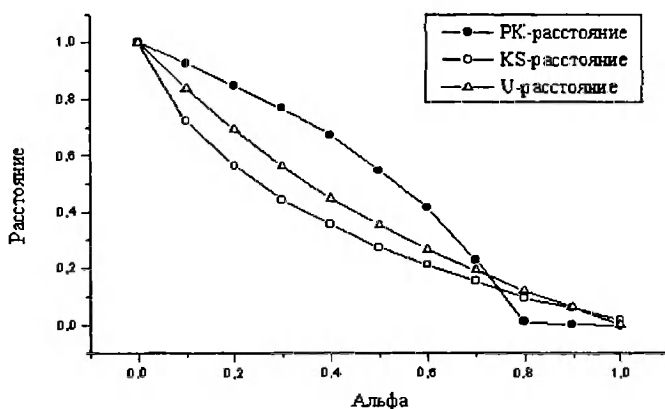


Рис. 2.18. Расстояния между выборкой из распределения $U(0,1)$ и выборками из семейства альтернатив Лемана на основе распределения $U(0,1)$ ($-1 < u \leq 0$)

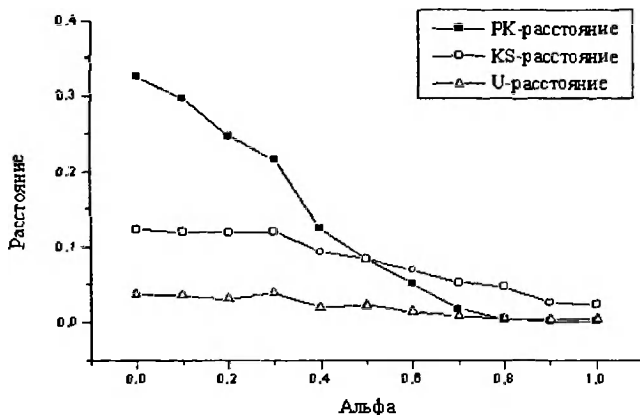


Рис. 2.19. Расстояния между выборкой из $N(0,1)$ и выборками из смеси равномерного и нормального распределений $\alpha N(0, 1) + (1 - \alpha)U\left(-\sqrt[3]{3/2}, \sqrt[3]{3/2}\right)$ при $g = 3$

Резюме

Меры близости между двумя выборками $x = (x_1, x_2, \dots, x_n)$ и $y = (y_1, y_2, \dots, y_n)$ вводятся для того, чтобы установить, принадлежат ли эти выборки одной и той же или разным генеральным совокупностям. От свойств меры близости зависит качество принятого решения. Предложенная новая мера близости обладает следующими преимуществами по сравнению со статистиками Колмогорова–Смирнова и Вилкоксона:

- 1) для p -статистики можно построить приближенные доверительные границы, соответствующие заданному уровню значимости в случае, когда нулевая гипотеза H не верна;
- 2) p -статистики более эффективны при проверке гипотезы об эквивалентности генеральных совокупностей с одинаковыми или близкими математическими ожиданиями;
- 3) p -статистика обладает высоким порогом чувствительности во всех проанализированных примерах выборок из разных генеральных совокупностей.

2.12. Стратифицированный анализ генеральных совокупностей

Рассмотрим новую концепцию генеральной совокупности, которая отличается от введенной в подразделе 2.2.1. Пусть $G = \{X_i, i \in I\}$ — произвольное множество элементов (индивидуумов); мы будем интересоваться не самими инди-

видуумами X_i как таковыми, а лишь значениями x_i их некоторой определенной числовой характеристики. Предположим, что в множестве G нет никакой структуры упорядочения, поэтому мы не можем выбрать какую-либо естественную определенную индексацию элементов множества G (введенная нами индексация X_i определена лишь для различения элементов, а не указывает порядок их следования). Во многих прикладных науках естественно принять следующую гипотезу: числовые характеристики x_i можно считать значениями некоторой случайной величины x , однако само множество G не является выборкой (поскольку оно не упорядочено), а представляет лишь множество выборочных значений некоторой неизвестной нам выборки G^* , которую генерирует природа. В выборке G^* есть структура упорядочения, однако она неизвестна экспериментаторам. Предположим, что множество G^* , а потому и G является конечным, причем $\text{card } G^* = n$. Пусть $G^* = \{x_1^*, x_2^*, \dots, x_n^*\}$, обозначим через $F^*(u_1, u_2, \dots, u_n)$ совместное распределение случайных величин $x_1^*, x_2^*, \dots, x_n^*$, которые могут быть зависимыми между собой, а через $F_1^*(u), F_2^*(u), \dots, F_n^*(u)$ — маргинальные распределения $F_k^*(u) = p(x_k^* < u)$ этих случайных величин. Положим

$$F(u) = F^*(u) = \frac{1}{n} \sum_{k=1}^n F_k^*(u).$$

Будем называть распределение $F(u)$ *базисным распределением* выборки G^* . Совместное распределение $F^*(u_1, u_2, \dots, u_n)$, а потому и маргинальные распределения $F_k^*(u)$, как и базисное распределение $F(u)$, нам неизвестны, однако $F(u)$ можно приближенно оценить на основании множества G выборочных значений выборки G^* . Действительно, пусть $\tilde{F}_n(u)$ — эмпирическая функция распределения, построенная на основании вариационного ряда выборки G^* или, что то же самое, множества выборочных значений G . На основании известных результатов теории вероятностей, выборочное распределение $\tilde{F}_n(u) \approx F(u)$, так что $\tilde{F}_n(u)$ можно считать приближением распределения $F(u)$. Заметим также, что вероятность $F(u)$ представляет собой вероятность события: выбранное наугад значение из множества G не превосходит u . Поскольку в множестве G нет упорядочения элементов, единственной вероятностной характеристикой этого множества является одномерное распределение $F(u) = p(x < u)$. Мы будем называть множество G с базисным распределением $F(u)$ *генеральной совокупностью*; таким образом, в нашем понимании генеральная совокупность представляет неупорядоченное множество зависимых или независимых выборочных значений.

Пусть G — генеральная совокупность и S — некоторый способ выбора элементов из G . В результате n -кратного применения данного способа выбора S мы получим упорядоченное множество чисел (вектор) $(x_1, \dots, x_n)$, состоящий из значений показателя x у элементов $X_1, \dots, X_n$. Мы будем называть вектор $(x_1, \dots, x_n)$ *выборкой* x , полученной с помощью способа выбора S из множества G . Для ге-

неральной совокупности G и данного способа выбора S при каждом натуральном n определена совместная функция распределения $F(u_1, \dots, u_n)$ случайных величин $x_1, \dots, x_n$

$$F(u_1, \dots, u_n) = p(x_1 < u_1, \dots, x_n < u_n).$$

Мы не располагаем самой выборкой G^* , имеющей совместное распределение $F^*(u_1, \dots, u_n)$, поэтому возникает следующая проблема о “восстановлении” выборки G^* на основании множества выборочных значений G : используя множество G , построить n случайных величин, совместное распределение которых совпадает с $F^*(u_1, \dots, u_n)$. Для решения этой проблемы было предложено два метода [79]. Первый метод основан на урновой модели. Будем предполагать, что в урне находится все множество выборочных значений G ; выберем наугад один из элементов γ_1 множества G , затем из множества $G \setminus \{\gamma_1\}$ оставшихся элементов выбираем наугад следующий элемент γ_2 и т.д. Мы имеем, таким образом, серию испытаний ω , состоящую из n случайных выборок без возвращения элементов множества G , содержащихся в урне. Получившуюся при этом многомерную случайную величину $\gamma^{(n)} = (\gamma_1, \gamma_2, \dots, \gamma_n)$ называем *индуцированной выборкой*, полученной с помощью урновой модели. Обозначим через Ω_n множество результатов ω таких серий испытаний; легко видеть, что число элементов множества Ω_n равно $n!$: $\Omega_n = \{\omega: r = 1, 2, \dots, n!\}$, причем $p\{\omega = \omega_r\} = 1/n!$. Справедливы следующие утверждения [79].

Теорема 2.3.

Пусть $F^*(u_1, \dots, u_n)$ — функция совместного распределения выборки G^* ; тогда функция совместного распределения $F_{\gamma^{(n)}}^*(u_1, \dots, u_n)$ индуцированной выборки $\gamma^{(n)}$, полученной с помощью урновой модели, имеет вид

$$F_{\gamma^{(n)}}^*(u_1, \dots, u_n) = \frac{1}{n!} \sum_{(i_1, \dots, i_n) \in I_n} F^*(u_{i_1}, \dots, u_{i_n}),$$

где I_n — группа перестановок чисел $(1, 2, \dots, n)$.

Следствие 2.1. Если G^* — выборка из генеральной совокупности G с распределением $F(u)$, то каждая компонента γ_i , $i = 1, 2, \dots, n$ случайного вектора $\gamma^{(n)}$ имеет распределение $F(u)$.

Следствие 2.2. Если элементы выборки G^* из генеральной совокупности G являются независимыми случайными величинами, то функция распределения $F_{\gamma^{(n)}}^*(u_1, \dots, u_n)$ индуцированной выборки $\gamma^{(n)}$, полученной с помощью урновой модели, совпадает с функцией распределения $F^*(u_1, \dots, u_n)$:

$$F_{\gamma^{(n)}}^*(u_1, \dots, u_n) = F^*(u_1, \dots, u_n) = F(u_1) \dots F(u_n).$$

Теорема 2.4.

Пусть G^* — выборка из генеральной совокупности G с распределением $F(u)$, причем выборочные значения из G^* имеют совместное нормальное распределение. Выборочные значения из G^* являются независимыми в совокупности тогда и только тогда, когда элементы $\gamma_1, \gamma_2, \dots, \gamma_n$ индуцированной выборки $\gamma^{(n)}$, полу-

ченной с помощью урновой модели, будут независимыми в совокупности случайными величинами.

Второй метод основан на понятии вариационного ряда и рандомизированной процедуры. Рассмотрим вариационный ряд $x_{(1)}^* \leq x_{(2)}^* \leq \dots \leq x_{(n)}^*$, построенный по неупорядоченному множеству выборочных значений G^* , полученному с помощью простого случайного выбора (это означает, что случайные величины $x_1^*, x_2^*, \dots, x_n^*$ являются независимыми в совокупности и имеют одинаковое маргинальное распределение). Определим m -мерную, $2 \leq m \leq n$, случайную величину $\xi^{(n)} = (\xi_1, \xi_2, \dots, \xi_m)$ следующим образом: проведем дополнительный случайный эксперимент (рандомизированную процедуру), состоящий в том, что мы извлекаем с помощью простого случайного выбора (без возвращения) $g_r = (i_1, i_2, \dots, i_n)$ объема m из множества чисел $(1, 2, \dots, n)$; при этом $r = 1, 2, \dots, A_n^m$, где $A_n^m = \frac{n!}{(n-m)!}$. Положим

$$\xi_1 = x_{(1)}^*, \xi_2 = x_{(2)}^*, \dots, \xi_m = x_{(m)}^*, \xi^{(m)} = (\xi_1, \xi_2, \dots, \xi_m).$$

Теорема 2.5.

Пусть функция распределения $F(u)$ генеральной совокупности G обладает плотностью вероятностей $f(u)$; тогда компоненты $\xi_1, \xi_2, \dots, \xi_m$ многомерной случайной величины $\xi^{(m)}$ являются независимыми одинаково распределенными случайными величинами, распределение которых $F_{\xi_j}(u)$ совпадает с $F(u)$, $j = 1, 2, \dots, m$.

Определение 2.3. Пусть G^* — неупорядоченное множество выборочных значений, полученных из генеральной совокупности G с помощью простого случайного выбора. Многомерные случайные величины $\gamma^{(n)}$ и $\xi^{(m)}$, о которых идет речь в теоремах 2.3 и 2.5, называются восстановленными выборками объема n и m .

Изложенная выше концепция генеральной совокупности показывает, что выборка $x = (x_1, \dots, x_k)$, $k < n$, полученная из генеральной совокупности G с помощью способа выбора S , далеко не всегда состоит из независимых выборочных значений, даже если пользоваться простым случайным выбором. Это противоречит известным представлениям классической математической статистики о том, что при использовании простого случайного выбора из генеральной совокупности G можно получить выборочные значения $x = (x_1, \dots, x_k)$, $k < n$, где " x_i — независимые величины, имеющие одну и ту же функцию распределения $F(x)$ " [42]. Таким образом, имеющуюся зависимость в выборке G^* нельзя исключить с помощью различных методов выбора S .

При определении генеральной совокупности G и метода выбора S мы не накладываем никаких условий на многомерные функции распределения (например, условие согласованности распределений и т.д.). Однако в дальнейшем мы будем изучать два способа выбора:

- простой случайный выбор, при котором случайные величины $x_1, \dots, x_n$ являются независимыми в совокупности и одинаково распределенными, так что $F(u_1, \dots, u_n) = F(u_1) \dots F(u_n)$, где $F(u)$ — маргинальное распределение;
- случайный выбор, при котором случайные величины независимы в совокупности, но могут иметь различные маргинальные распределения $F_k(u_k) = 1, \dots, n$, так что $F(u_1, \dots, u_n) = F_1(u_1) \dots F_n(u_n)$.

На основании изложенных выше результатов о восстановленной выборке мы можем гарантировать существование простого случайного выбора или случайного выбора в случае, когда первоначальная выборка G^* состояла из независимых в совокупности выборочных значений (гипотеза H). Всюду в дальнейшем мы будем предполагать, что гипотеза H верна.

В приложениях математической статистики большую роль играют понятия гомогенной и гетерогенной генеральной совокупности. *Гомогенная генеральная совокупность* состоит из однородных объектов, таких что значения $x_1, \dots, x_n$ количественного показателя x имеют одинаковые маргинальные распределения $F(u)$: $F_k(u) = F(u)$, $k = 1, \dots, n$. Однако это условие не достаточное для гомогенности распределения. На основании эмпирических данных принято считать, что *однородная (гомогенная) генеральная совокупность* характеризуется *одномодальным* распределением, поэтому мы будем считать генеральную совокупность *гомогенной*, если маргинальное распределение $F(u)$ является *одномодальным*. Это определение не будет логически безупречным, поскольку существуют контр-примеры смеси двух различных нормальных распределений, у которой плотность вероятности является *одномодальной*. Тем не менее такие нормальные распределения являются достаточно близкими и поэтому их считают практически одинаковыми. Таким образом, в нашем понимании термины *одномодальная* и *гомогенная генеральные совокупности* являются синонимами.

Клеточная популяция является частным случаем абстрактной генеральной совокупности, которая изучается в современной математической статистике и представляет собой произвольное множество значений G количественных характеристик некоторых объектов с заданным в множестве G распределением вероятностей $F(u)$, возможно, неизвестное экспериментатору. Как правило, распределение гетерогенной генеральной совокупности G является *полимодальным*, однако выборка $x = (x_1, \dots, x_n) \in G$ из гетерогенной генеральной совокупности может представлять в простейшем случае объединение двух выборок $x^{(1)} = (x_1, \dots, x_k) \in G_1$ и $x^{(2)} = (x_{k+1}, \dots, x_n) \in G_2$ из двух различных генеральных совокупностей G_1 и G_2 (такие выборки будем называть *составными*) или смесь двух выборок из этих совокупностей G_1 и G_2 (такие выборки мы будем называть *смешанными*). Составные выборки возникают тогда, когда индивидуумы, составляющие выборку, имеют различную локализацию (т.е. они расположены в разных частях пространства), в то время как индивидуумы, образующие смешанную выборку, такой компактной локализации не имеют и могут быть расположены в случайном порядке.

Рассмотрим проблему обнаружения гетерогенности генеральной совокупности. Очевидно, что распознать составную выборку значительно проще, чем смешанную. Действительно, если предположить, что составные части $x^{(1)} = (x_1, \dots, x_k) \in G_1$ и $x^{(2)} = (x_{k+1}, \dots, x_n) \in G_2$ имеют примерно одинаковый объем, то, вычисляя меру близости выборок $x^{(1)}$ и $x^{(2)}$, мы можем установить различие распределений этих выборок. Если это различие значимо, то выборка $x = x^{(1)} \cup x^{(2)}$ будет составной, в противном случае следует проверить, является ли она смешанной.

Приведем описание вычисления меры близости между выборками $x = (x_1, \dots, x_n) \in G$ и $x' = (x'_1, \dots, x'_m) \in G'$, полученными простым случайным выбором из генеральных совокупностей G и G' соответственно с неизвестными функциями распределения $F_G(u)$ и $F_{G'}(u)$. Рассмотрим самую общую ситуацию, когда функции распределения $F_G(u)$ и $F_{G'}(u)$ могут иметь атомы. Поясним это понятие более подробно. Как известно, произвольная функция распределения $F(u)$ является монотонно возрастающей функцией, определенной на всей числовой прямой, значения которой заключены в пределах от 0 до 1. Среди точек области определения $F(u)$, т.е. числовой прямой, есть точки непрерывности u_0 , когда

$$F(u_0 - 0) = \lim_{u \rightarrow u_0, u < u_0} F(u) = F(u_0 + 0) = \lim_{u \rightarrow u_0, u > u_0} F(u).$$

Однако на числовой прямой есть точки разрыва u_1 , когда

$$F(u_1 - 0) < F(u_1 + 0).$$

Точки разрыва функции распределения называются ее *атомами*. Для этих точек вероятность события “выборочное значение x в точности равно u_1 ” отлично от нуля, тогда как для точек непрерывности эта вероятность равна нулю. На основании выборки такие атомы можно установить, если, конечно, выборка имеет достаточно большой объем: при этом атомами являются те выборочные значения, которые встречаются больше одного раза, т.е. повторяются.

Пусть $x_{(1)} \leq \dots \leq x_{(n)}$, $x'_{(1)} \leq \dots \leq x'_{(m)}$ — вариационные ряды, построенные по выборкам $x = (x_1, \dots, x_n) \in G$ и $x' = (x'_1, \dots, x'_m) \in G'$. Как известно, элементы вариационного ряда $x_{(k)}$, $k = \overline{1, n}$ ($x'_{(i)}$, $i = \overline{1, m}$) называются *порядковыми статистиками*.

Если порядковая статистика $x_{(k)}$ встречается в вариационном ряду более чем один раз, то $x_{(k)}$ является атомом распределения $F_G(u)$. Предположим, что в полуинтервале $[x_{(j)}, x_{(j+1)})$ нет атомов, тогда на основании результатов работы [84]

$$p(A_j) = p(x'_k \in (x_{(j)}, x_{(j+1)})) = p_j = \frac{j-i}{n+1}, \quad (2.35)$$

где $\tilde{x}$ — последующее выборочное значение из генеральной совокупности G , не зависящее от выборки $x = (x_1, \dots, x_n) \in G$.

В случае, когда полуинтервал $[x_{(j)}, x_{(j+1)})$ содержит атомы, мы можем представить его в виде разбиения на смежные составляющие полуинтервалы:

$$[x_{(j)}, x_{(j+1)}) = [x_{(j)}, x_{(i+1)}) \cup [x_{(i+1)}, x_{(i+2)}) \cup \dots \cup [x_{(j-1)}, x_{(j)}).$$

Предположим, что левый конец какого-либо из составляющих полуинтервалов $[x_{(k)}, x_{(k+1)})$ является атомом. Этот факт можно установить на основании исследования выборки $x = (x_1, \dots, x_n)$: если элемент x_k встречается в выборке более чем один раз, то x_k будет атомом. При достаточно репрезентативной выборке $x = (x_1, \dots, x_n)$ все атомы распределений $F(u)$ можно определить на основании указанного выше правила. Обозначим через n_k количество повторений элемента x_k в выборке $x = (x_1, \dots, x_n)$; положим $\gamma_k = \frac{n_k}{n}$. Нетрудно заметить, что на основании закона больших чисел при достаточно больших n

$$\gamma_k \approx F(x_k + 0) - F(x_k - 0).$$

При этом в формулу (2.28) нужно внести следующие поправки:

$$\begin{aligned} p(A_{kk+1}) &= p(\tilde{x} \in [x_{(k)}, x_{(k+1)})) = p_{kk+1} = p(\tilde{x} \in \{x_k\} \cup (x_k, x_{k+1})) = \\ &= p(\tilde{x} = x_k) + p(\tilde{x} \in (x_k, x_{(k+1)})) \approx \gamma_k + \frac{1}{n+1} \end{aligned}$$

Учитывая это, имеем:

$$\begin{aligned} p(A_{ij}) &= p(\tilde{x} \in (x_{(i)}, x_{(j)})) = p_{ij} = \\ &= p(\tilde{x} \in [x_{(i)}, x_{(i+1)})) \cup [x_{(i+1)}, x_{(i+2)})) \cup \dots \cup [x_{(j-1)}, x_{(j)})) = \\ &= p(\tilde{x} \in [x_{(i)}, x_{(i+1)})) + p(\tilde{x} \in [x_{(i+1)}, x_{(i+2)})) + \dots + p(\tilde{x} \in [x_{(j-1)}, x_{(j)})) \end{aligned}$$

так, что

$$p(A_{ij}) = p_{ij} = \gamma_i + \gamma_{i+1} + \dots + \gamma_{j-1} + \frac{j-i}{n+1}. \quad (2.36)$$

Заметим, что формула (2.36) справедлива, независимо от того, содержит ли полуинтервал $[x_{(i)}, x_{(j)})$ атомы или нет. В случае, когда в полуинтервале $[x_{(i)}, x_{(j)})$ нет атомов, то $\gamma_i + \gamma_{i+1} + \dots + \gamma_{j-1} = 0$, и формула (2.36) переходит в формулу (2.35).

Если мы имеем выборку $x' = (x'_{(1)}, \dots, x'_{(m)})$, мы можем найти частоту h_y случайного события A_y и доверительные границы $p_{ij}^{(1)}$, $p_{ij}^{(2)}$ для вероятности p_y , соответствующих заданному уровню значимости β , такому что $B = \{p_{ij} \in (p_{ij}^{(1)}, p_{ij}^{(2)})\}$, $p(B) = 1 - \beta$. Эти границы были вычислены по формулам [42]:

$$\begin{aligned} p_{ij}^{(1)} &= \frac{h_y m + g^2/2 - g\sqrt{h_y(1-h_y)m + g^2/4}}{m + g^2}, \\ p_{ij}^{(2)} &= \frac{h_y m + g^2/2 + g\sqrt{h_y(1-h_y)m + g^2/4}}{m + g^2}, \end{aligned} \quad (2.37)$$

где g удовлетворяет условию $\Phi(g) = 1 - \beta/2$, $\Phi(u)$ — функция нормированного нормального распределения (если m мало, то в соответствии с правилом 3σ $g = 3$).

Обозначим через N все доверительные интервалы $I_y = (p_y^{(1)}, p_y^{(2)})$ ($N = n(n-1)/2$) и L — число тех интервалов I_y , которые содержат вероятности p_y . Положим $h = \rho(F^*, F^{\circ}) = \rho(x, x') = L/N$. Поскольку h — частота случайного события $B = \{p_y \in I_y\}$, имеющего вероятность $p(B) = 1 - \beta$, то, полагая $h_y = h$, $m = N$ и $g = 3$ в формулах (2.37), мы получим доверительный интервал $I = (p^{(1)}, p^{(2)})$ для вероятности $p(B)$, который имеет доверительный уровень, равный 0,95. Критерий для проверки гипотезы H с уровнем значимости, приближенно равным 0,05, может быть сформулирован следующим образом: если доверительный интервал $I = (p^{(1)}, p^{(2)})$ содержит вероятность $p(B) = 1 - \beta$, то гипотеза H принимается, в противном случае она отвергается. Статистика h называется p -статистикой. Она является мерой близости $\rho(x, x')$ между выборками x и x' .

Проблема распознавания смешанной выборки является более сложной, поскольку смесь двух одномодальных распределений при некоторых соотношениях их математических ожиданий и дисперсий может быть снова одномодальным распределением. Однако такая ситуация не характерна и, как правило, при смешивании двух одномодальных распределений получается бимодальная.

Нами предложен новый метод решения проблемы распознавания смешанной выборки. Этот метод основан на исследовании графика вариационного ряда выборки. Пусть $x = (x_1, \dots, x_n) \in G$ — выборка из генеральной совокупности G и $x_{(1)} \leq \dots \leq x_{(n)}$ — ее вариационный ряд. Рассмотрим на плоскости декартову прямоугольную систему координат и точки $(i, x_{(i)})$, $i = 1, 2, \dots, n$, проведем через эти точки плавную кривую (для определенности соединим соседние точки прямолинейными отрезками) и будем называть эту совокупность точек графиком вариационного ряда. Эта кривая в классической математической статистике называется кривой Кетле. Хорошо известно, что кривая Кетле состоит из трех частей. Ее начальная часть получена из первых минимальных порядковых статистик, затем идет ее центральная часть, которая образована точками $x_{(i)}$ из основной распределенной массы значений генеральной совокупности G , и ее конечной части, полученной из последних максимальных порядковых статистик. Во многих случаях начальная и конечная части кривой Кетле статистиками рассматривается как кривые, образованные выбросами, обусловленными артефактами. Поэтому наиболее надежной и заслуживающей доверия частью кривой Кетле является ее центральная часть. Если генеральная совокупность однородна, то центральная часть кривой Кетле хорошо аппроксимируется прямолинейным отрезком. Поэтому, когда центральная часть кривой Кетле аппроксимируется двумя отрезками, порождающими ломаную, состоящую из двух звеньев (первый случай), или двумя несвязными отрезками (второй случай), мы имеем дело с неоднородной (гетерогенной) генеральной совокупностью. В любом случае правый конец первого отрезка мы будем называть *осо-*

бой точкой кривой Кетле; в первом случае она является *точкой излома*, а во втором — *точкой разрыва* кривой Кетле.

Обоснование предложенного теста для обнаружения гетерогенности основано на следующих соображениях. Если генеральная совокупность является однородной, то ее плотность вероятности в начале возрастает, а затем, после моды, начинает убывать. Поэтому функция распределения такой генеральной совокупности до точки моды является выпуклой, а после вогнутой; в окрестности моды (точки перегиба) эта кривая хорошо аппроксимируется прямолинейным отрезком, так что основная распределенная масса значений генеральной совокупности расположена между абсциссами концов этого отрезка. Поскольку кривая Кетле аппроксимирует обратную функцию к функции распределения, график этой кривой представляет отражение графика функции распределения относительно биссектрисы первого и третьего координатных углов. Следовательно, в этом случае кривая Кетле в ее центральной части также хорошо аппроксимируется прямолинейным отрезком. Если же плотность вероятности генеральной совокупности G является двухмодальной кривой, то график ее функции распределения состоит из двух подобных частей: для одной части это будет кривая, меняющая свою выпуклость на вогнутость в окрестности первой моды, а для другой — аналогичная кривая в окрестности второй моды, поэтому график функции распределения будет хорошо аппроксимироваться двумя прямолинейными отрезками. Следовательно, график кривой Кетле также будет хорошо аппроксимироваться двумя прямолинейными отрезками. Отметим, однако что при вычислениях лучше иметь дело с кривой Кетле, поскольку ее график является более пологой кривой по сравнению с графиком функции распределения.

Алгоритм обнаружения и вычисления координат особой точки кривой Кетле состоит в следующем. Сначала по выборке $x = (x_1, \dots, x_n) \in G$ построим вариационный ряд $x_{(1)} \leq \dots \leq x_{(n)}$ и отбросим 10% его крайних порядковых статистик (5% первых минимальных порядковых статистик и 5% последних максимальных). Затем по оставшимся порядковым статистикам мы строим кривую, образованную приращениями ординат вершин смежных звеньев кривой Кетле (i, y_i) , где $y_i = x_{(i+1)} - x_{(i)}$, $i = 1, 2, \dots, n-1$, соединяя смежные вершины (i, y_i) , $(i+1, y_{i+1})$ прямолинейными отрезками.

Величины y_i носят название спейсингов (spacing), поэтому мы будем называть построенную таким образом кривую спейсинг-кривой. Разделим множество абсцисс и ординат вершин спейсинг-кривой на три равные (или почти равные) части U_1, U_2, U_3 и Y_1, Y_2, Y_3 :

$$U_1 = \{1, 2, \dots, k\}, U_2 = \{k+1, k+2, \dots, m\}, U_3 = \{m+1, m+2, \dots, n-l-1\},$$

$$Y_1 = \{y_1, y_2, \dots, y_k\}, Y_2 = \{y_{k+1}, y_{k+2}, \dots, y_m\}, Y_3 = \{y_{m+1}, y_{m+2}, \dots, y_{n-l-1}\},$$

$$\text{где } k = \left\lfloor \frac{n-2l-1}{3} \right\rfloor, m = n-l-k-2.$$

Положим $y_{\min}^{(i)} = \min Y_i$, $y_{\max}^{(i)} = \max Y_i$, $i = 1, 2, 3$. Тогда критерий обнаружения гетерогенности генеральной совокупности можно сформулировать следующим образом.

Если $y_{\min}^{(1)} \leq y_{\min}^{(3)}$ и $y_{\max}^{(1)} \geq y_{\max}^{(3)}$, то начиная с правого конца графика спейсингов проверяем неравенство $y_{\min}^{(3)} \leq y_{n-l-1-j} \leq y_{\max}^{(3)}$, $j = 1, 2, \dots, n-l-2$; первая точка с номером n_0 , выходящая за пределы интервала, является особой точкой и соответственно границей между двумя популяциями в генеральной совокупности. При этом множество $W_1 = \{y_i\}$, $i = 1, 2, \dots, n_0$, принадлежит к первой популяции, а множество $W_2 = \{y_i\}$, $i = n_0, n_0 + 1, \dots, n_0 - l - 1$ — ко второй.

В противном случае перебор точек графика спейсингов начинаем с левого конца, проверяя неравенство $y_{\min}^{(1)} \leq y_j \leq y_{\max}^{(1)}$, $j = 1, 2, \dots, n$; аналогично предыдущему случаю первая точка с номером n_0 , выходящая за пределы интервала, является особой точкой, и множество $W_1 = \{y_i\}$, $i = 1, 2, \dots, n_0$, принадлежит к первой популяции, а множество $W_2 = \{y_i\}$, $i = n_0, n_0 + 1, \dots, n_0 - l - 1$ — ко второй.

Возможен также вариант, когда особой точки не существует; при этом $y_{\min}^{(1)} = y_{\min}^{(3)}$ и $y_{\max}^{(1)} = y_{\max}^{(3)}$. В таком случае выборка является однородной.

Отметим, что применение изложенной выше методики исследования спейсинг-кривых следует проводить и в случае объединенных (составных) выборок, поскольку эта методика дает возможность установить одномодальность составляющих выборок и степень гетерогенности исследуемой выборки.

Кроме того, описанная методика может использоваться для более детального анализа гетерогенных выборок для выделения более чем двух компонент (модальных классов), в этом случае особые точки, которых может быть более чем одна, являются концами интервалов, образованных модальными классами.

На рис. 2.20–2.31 приведены кривые Кетле и спейсинг-кривые, построенные по вариационным рядам выборок, составленных из площадей ядер клеток карциномы Герена при исследовании поколений радиорезистентных клеток. Изложенные выше методы исследования гетерогенности генеральной совокупности позволили установить, что контрольная (необлученная) генеральная совокупность является однородной, а генеральные совокупности последующих поколений клеток являются гетерогенными, некоторые из которых являются составными, а другие — смешанными.

Результаты исследования гетерогенности с помощью метода, основанного на вычислении меры близости между двумя половинами $X_1 = \{x_1, x_2, \dots, x_k\}$ и $X_2 = \{x_{k+1}, x_{k+2}, \dots, x_n\}$ выборки $X = \{x_1, x_2, \dots, x_n\}$, где $k = \left[\frac{n}{2} \right]$ — целая часть числа $\frac{n}{2}$, приведены в табл. 2.2. Эту меру близости обозначим через $\rho(X_1, X_2)$ и введем обозначение $\mu(X_1, X_2) = 1 - \rho(X_1, X_2)$. Малые значения величины $\mu(X_1, X_2)$ свидетельствуют о том, что выборка является объединенной.

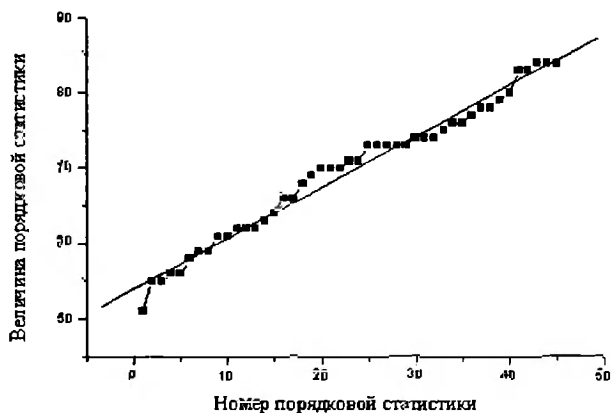


Рис. 2.20. Кривая Кетле для контрольной генерации

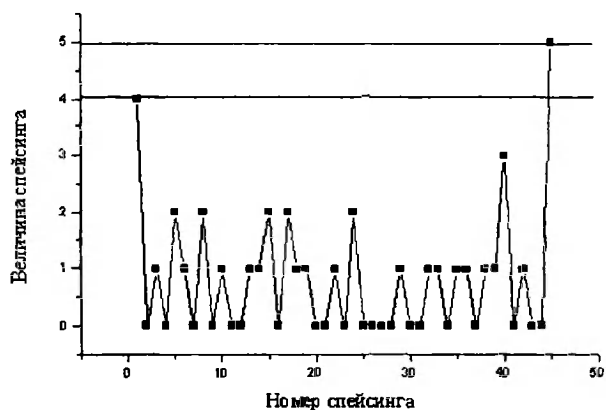


Рис. 2.21. Спейсинг-кривая для контрольной генерации

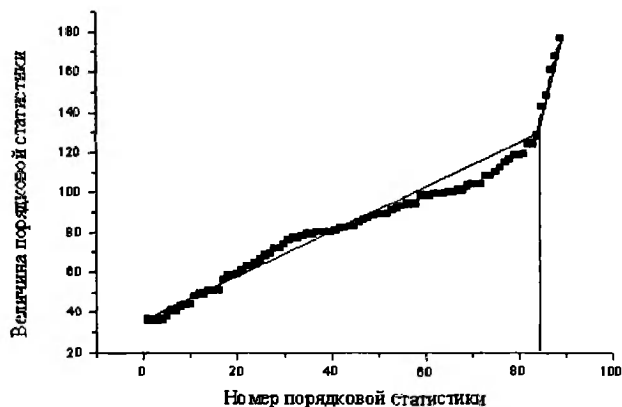


Рис. 2.22. Кривая Кетле для 3-й генерации

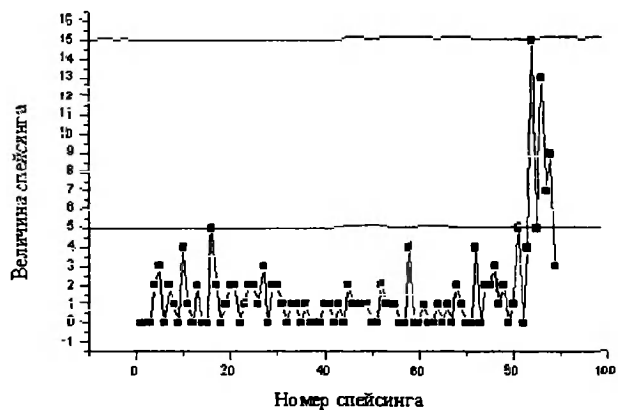


Рис. 2.23. Спейсинг-кривая для 3-й генерации

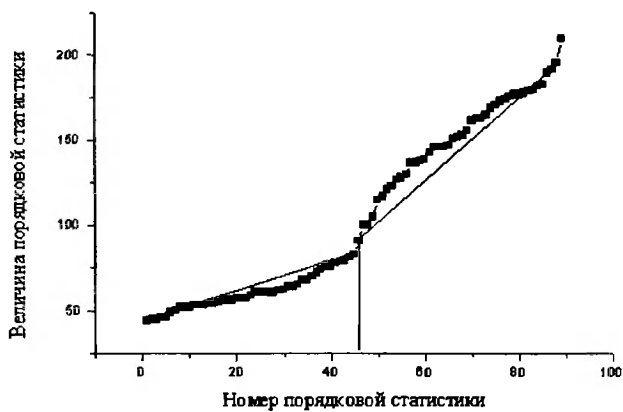


Рис. 2.24. Кривая Кетле для 5-й генерации

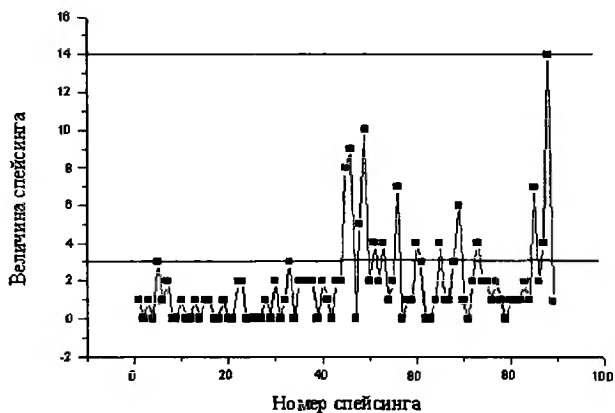


Рис. 2.25. Спейсинг-кривая для 5-й генерации

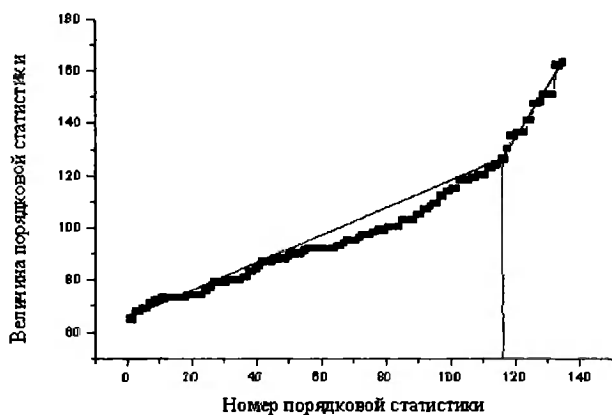


Рис. 2.26. Кривая Кетле для 6-й генерации

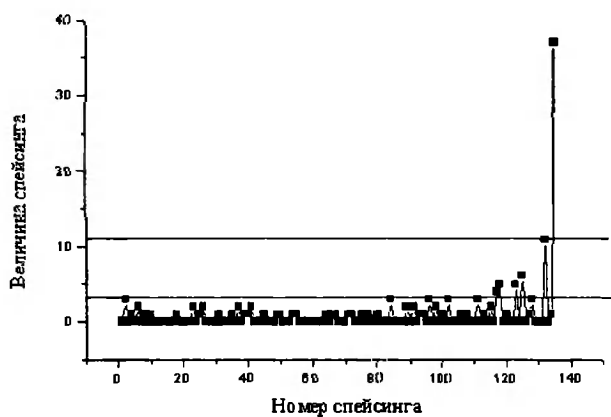


Рис. 2.27. Спейсинг-кривая для 6-й генерации

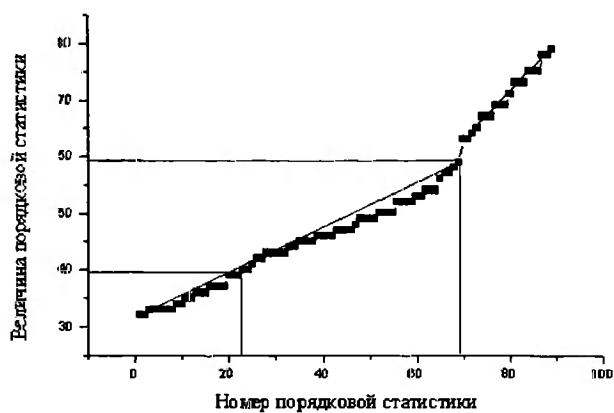


Рис. 2.28. Кривая Кетле для 7-й генерации

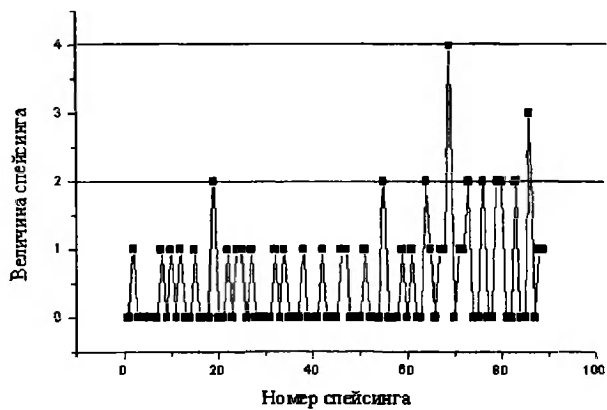


Рис. 2.29. Спейсинг-кривая для 7-й генерации

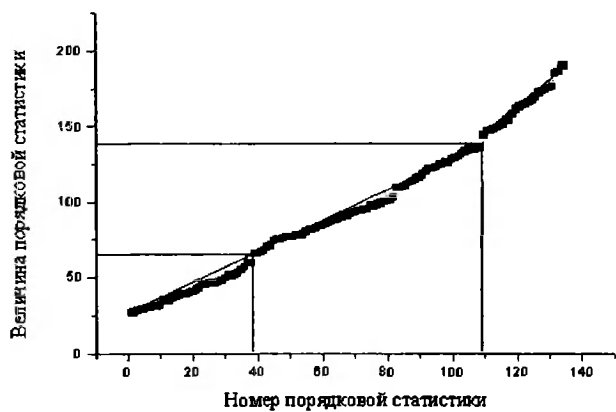


Рис. 2.30. Кривая Кетле для 12-й генерации

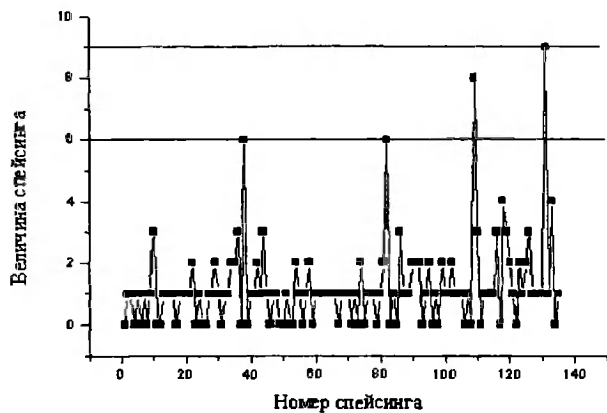


Рис. 2.31. Спейсинг-кривая для 12-й генерации

Таблица 2.2. Показатели гетерогенности генераций

Генерация	$\rho(X_1, X_2)$	НГ	ВГ	$\mu(X_1, X_2)$	НГ	ВГ
Контроль	1,000	0,944	1,000	0,000	0,000	0,056
3	0,974	0,950	0,987	0,026	0,013	0,050
5	0,249	0,190	0,318	0,751	0,682	0,810
6	0,998	0,991	0,999	0,002	0,001	0,009
7	0,493	0,419	0,566	0,507	0,434	0,581
12	0,441	0,409	0,475	0,559	0,525	0,591

Из табл. 2.2 следует, что выборки из 5-й, 7-й и 12-й генераций являются составными.

Далее было проведено исследование типа выборок с помощью метода, основанного на анализе спейсинг-кривых. При этом на основе анализа графиков кривых Кетле и спейсинг-кривых (см. рис. 2.20–2.31) были вычислены особые точки, доля популяций (модальных классов), составляющих генерацию, и пределы изменения площади ядер клеток из этих популяций (табл. 2.3).

Таблица 2.3. Оценка гетерогенности популяций с помощью спейсинг-кривой

№№	Первая популяция						Вторая популяция			
	Количество клеток	Особая точка	Количество клеток	Доля	Min	Max	Количество клеток	Доля	Min	Max
Контроль	45	Нет	45	1,00	55	85	0	0,00	0,00	0,00
3	89	83	83	0,93	30	125	6	0,07	125	225
5	89	44	44	0,49	30	80	45	0,51	100	225
6	135	118	118	0,87	65	125	17	0,13	125	160
7	89	69	69	0,78	30	60	20	0,22	60	90
12	135	110	110	0,82	25	125	25	0,18	125	200

Примечание. При исследовании гетерогенности мы отбрасывали 5% первых минимальных и 5% последних максимальных порядковых статистик для определения основной распределенной массы значений генеральной совокупности. Кроме того, выборка, составляющая менее чем 5% от общего числа клеток, не рассматривалась как отдельная популяция.

Резюме

Во многих задачах доказательной медицины очень важно установить, является ли выборка $x_1, x_2, \dots, x_n$ однородной (гомогенной) или она состоит из разных модальных классов, т.е. является составной или смешанной (гетерогенной). Для решения этой проблемы мы предложили специальный аппарат (стратифи-

кационный анализ), позволяющий установить гомогенность и гетерогенность выборок с помощью двух методов — кривых Кетле и спейсингов. Кривую Кетле можно определить как функцию, обратную к модифицированной эмпирической распределения, а спейсингами γ_i называется разность между порядковыми статистиками $x_{(i+1)} - x_{(i)}$. Если средняя часть кривой Кетле хорошо аппроксимируется линейной функцией, то выборка является однородной, в противном случае она считается неоднородной. Аналогично, если средняя часть кривой спейсингов является почти постоянной, то выборка однородна, а в противном случае — смешанной.

2.13. Статистический критерий для сравнения двух вероятностей

2.13.1. Введение

Напомним, что схемой Бернулли B называется последовательность независимых испытаний (случайных экспериментов $E_1, E_2, \dots, E_n, \dots$), в результате каждого из которых может произойти случайное событие A с одной и той же вероятностью p : $p(E_i, A) = p, i = 1, 2, \dots, n, \dots$. Предположим, что у нас две схемы Бернулли B_1 и B_2 со случайными экспериментами E_n^1 и $E_n^2, n = 1, 2, \dots$, причем $p(E_n^1, A_1) = p_1, p(E_n^2, A_2) = p_2$. Пусть проводится n_1 испытаний в первой схеме Бернулли B_1 и n_2 испытаний — в схеме B_2 , тогда можно найти частоты

$$h_1 = \frac{k_1}{n_1} \text{ и } h_2 = \frac{k_2}{n_2},$$

соответствующие вероятностям p_1 и p_2 . Проблема сравнения двух вероятностей состоит в том, чтобы построить статистический критерий, который позволяет по имеющимся частотам h_1 и h_2 сделать вывод о различии (значимом отклонении) этих вероятностей p_1 и p_2 . При условии независимости частот h_1 и h_2 эту проблему можно решить с помощью критерия χ^2 [42]. К сожалению, это предположение далеко не всегда выполняется при исследовании различного рода задач в биологии и медицине. Например, при исследовании распределений частот некоторых заболеваний по диапазонам поглощенных доз радиации в когорте ликвидаторов возникает следующая задача: в первом диапазоне доз частота заболевания D равна h_1 , а во втором диапазоне — h_2 , так что $h_1 + h_2 = 1$. Необходимо выяснить, значимо ли отличаются соответствующие вероятности p_1 и p_2 ? Очевидно, что в этом случае частоты h_1 и h_2 являются зависимыми, поэтому отсутствуют условия применимости критерия χ^2 . Другой пример, в котором частоты h_1 и h_2 зависимы между собой, представляет проблема сравнения вероятностей случайных событий, состоящих в том, что поглощенная доза радиации у ликвидатора из когорты находится в заданном интервале I (событие A_1), а у ликвидатора из некоторой части этой когорты (например, ликвидаторов,

страдающих некоторым заболеванием) поглощенная доза радиации принадлежит тому же интервалу I (событие A_2). В этом случае схема Бернулли B_2 , соответствующая событию A_2 , является частью схемы Бернулли B_1 , соответствующей событию A_1 , и частоты этих событий h_1 и h_2 являются зависимыми случайными величинами.

2.13.2. Оценки дисперсии частоты случайного события в схеме Бернулли

Пусть B — схема Бернулли, в результате каждого из случайных экспериментов E , случайное событие A может произойти с вероятностью p . Очевидно, что частота h случайного события A имеет математическое ожидание $m(h) = p$ и дисперсию

$$D(h) = \frac{pq}{n}, \quad (2.38)$$

где $q = 1 - p$.

Для построения статистических критериев сравнения двух вероятностей во многих случаях необходимо найти доверительный интервал для неизвестной вероятности p , а эта проблема в свою очередь требует вычисления оценки для дисперсии $D(h)$, поскольку точное значение h зависит от неизвестной вероятности p .

Самую грубую оценку для $D(h)$ можно получить на основании следующих простых соображений: легко видеть, что произведение $pq = p(1 - p) \leq \frac{1}{4}$, если $0 \leq p \leq 1$, поэтому $D(h) \leq \frac{1}{4n}$.

Для нахождения более тонких оценок в формуле (2.38) следует заменить неизвестную вероятность p на ее несмещенную оценку — частоту h : мы получим тогда оценку для неизвестной дисперсии $D(h)$:

$$s_n^2 = \frac{h(1-h)}{n}. \quad (2.39)$$

Однако эта оценка является смещенной. Для доказательства этого утверждения введем случайные величины

$$x_k = \begin{cases} 1, & \text{если в эксперименте событие } A \text{ происходит,} \\ 0, & \text{если в эксперименте событие } A \text{ не происходит.} \end{cases} \quad (2.40)$$

Тогда имеем

$$\begin{aligned} m(s_n^2) &= \frac{1}{n} (m(h) - m(h^2)) = \\ &= \frac{p}{n} - m\left(\frac{1}{n^2} \left(\sum_{k=1}^n x_k^2 + \sum_{i \neq k} x_i x_k \right)\right) = \\ &= \frac{p}{n} - \frac{1}{n^2} \left(\sum_{k=1}^n m(x_k^2) + \sum_{i \neq k} m(x_i) m(x_k) \right) = \\ &= \frac{pq}{n} - \frac{1}{n^2} pq = \frac{(n-1)pq}{n^2} = \frac{n-1}{n} D(h). \end{aligned}$$

Поэтому, если положить

$$\hat{s}_n^2 = \frac{n}{n-1} s_n^2 = \frac{h(1-h)}{n-1}, \quad (2.41)$$

то получим несмещенную оценку для $D(h)$.

Заметим, однако что оценка s_n^2 является асимптотически несмещенной, поскольку $\lim_{n \rightarrow \infty} s_n^2 = p$.

Для сравнения оценок (2.39) и (2.41) следует рассмотреть дисперсию несмещенной оценки $D(\hat{s}_n^2)$ и среднюю квадратичную ошибку $\Delta(s_n^2)$ оценки s_n^2 :

$$\Delta(s_n^2) = m(s_n^2 - D(h))^2. \quad (2.43)$$

Очевидно, что для несмещенной оценки ее средняя квадратичная ошибка превращается в дисперсию.

Теорема 2.6.

При любом натуральном $n \geq 2$

$$\Delta(s_n^2) < D(\hat{s}_n^2). \quad (2.44)$$

Теорема 2.6 показывает, что смещенная оценка s_n^2 для дисперсии частоты h является более точной, чем несмещенная оценка $\hat{s}_n^2$. Следовательно, оценка s_n^2 более предпочтительна, чем оценка $\hat{s}_n^2$, а при построении доверительных границ для неизвестной вероятности p по правилу $3s$ [98] следует пользоваться оценкой s_n^2 .

2.13.3. Доверительные границы для неизвестной вероятности

Одной из самых важных задач прикладной математической статистики является проблема построения доверительных границ для неизвестной вероятности p на основании частоты случайного события A в схеме Бернулли. Приближенное решение этой задачи, полученное на основании предельной теоремы Муавра–Лапласа, хорошо известно и описано в книге [42]. Такими границами являются числа

$$p_1 = \frac{hn + \frac{1}{2}g^2 - g\sqrt{h(1-h)n + \frac{1}{4}g^2}}{n + g^2}, \quad (2.45)$$

$$p_2 = \frac{hn + \frac{1}{2}g^2 + g\sqrt{h(1-h)n + \frac{1}{4}g^2}}{n + g^2}, \quad (2.46)$$

причем уровень значимости этого интервала приблизительно равен $2 - 2\Phi(g)$ (при 5%-ном уровне значимости $g = 1,96$). Однако точное значение уровня значимо-

сти этого доверительного интервала при различных значениях n колеблется в достаточно широких границах. Действительно, точное значение доверительного уровня этого интервала

$$1 - 2\beta = \sum_{k: p_1 < p < p_2} p_n(k),$$

где $p_n(k) = C_n^k p^k q^{n-k}$.

Другой метод построения доверительных границ для неизвестной вероятности можно предложить на основании правила $2s$, который утверждает, что доверительный интервал $(h - 2s_n, h + 2s_n)$ содержит неизвестную вероятность p с вероятностью, приближенно равной 0,95. Точные значения границ, в которых варьируются уровни значимости доверительных интервалов I_w , построенных по методу Вильсона при $g = 2$, а также доверительных интервалов I_{2s} , вычисленных на основании правила $2s$, приведены в табл. 2.4.

Таблица 2.4. Минимальные, максимальные и средние уровни значимости доверительного интервала Вильсона (при $g = 2$) и по правилу $2s$ (см. подраздел 2.9.1)

n	По Ван дер Вардену				По правилу $2s$			
	min	max	max-min	average	min	max	max-min	average
10	0,012	0,096	0,083	0,037	0,000	0,173	0,173	0,048
20	0,017	0,075	0,059	0,039	0,000	0,143	0,143	0,057
30	0,018	0,065	0,046	0,041	0,000	0,149	0,149	0,060
40	0,020	0,061	0,041	0,041	0,000	0,128	0,128	0,058
50	0,022	0,089	0,067	0,044	0,000	0,147	0,147	0,058
60	0,020	0,059	0,039	0,040	0,000	0,148	0,148	0,057
70	0,022	0,059	0,037	0,043	0,000	0,127	0,127	0,053
80	0,021	0,057	0,035	0,042	0,000	0,136	0,135	0,059
90	0,028	0,062	0,034	0,044	0,000	0,098	0,098	0,054
100	0,028	0,079	0,052	0,046	0,001	0,150	0,150	0,058

Примечание. N — количество элементов, min — минимальный уровень значимости, max — максимальный уровень значимости, max-min — разность между максимальным и минимальным уровнем значимости, average — средний уровень значимости.

Из этой таблицы следует, что доверительные интервалы I_w имеют меньшие уровни значимости, чем интервалы I_{2s} .

2.13.4. Статистические критерии для сравнения двух вероятностей, построенные на основании доверительных интервалов

Пусть B_1 и B_2 — схемы испытаний Бернулли, в которых событие A_1 происходит с вероятностью p_1 , а A_2 — с вероятностью p_2 соответственно. На основании доверительных границ для вероятностей p_1 и p_2 можно предложить различные статистические критерии для сравнения этих двух вероятностей, причем гипотеза о независимости частот h_1 и h_2 событий A_1 и A_2 в этих критериях не будет использоваться.

Действительно, пусть $I_1 = (p_1^{(1)}, p_2^{(1)})$ и $I_2 = (p_1^{(2)}, p_2^{(2)})$ — доверительные интервалы для неизвестных вероятностей p_1 и p_2 соответственно, уровень значимости которых приближенно равен 5%. Статистический критерий для сравнения вероятностей p_1 и p_2 имеет следующий вид: основная гипотеза H о равенстве вероятностей p_1 и p_2 ($p_1 = p_2 = p$) отклоняется, если доверительные интервалы I_1 и I_2 не пересекаются.

Оценим уровень значимости этого критерия. Для этого рассмотрим следующие отрезки: $I_1^- = [0, p_1^{(1)}]$, $I_1^+ = [p_2^{(1)}, 1]$, $I_2^- = [0, p_1^{(2)}]$, $I_2^+ = [p_2^{(2)}, 1]$. Предположим для определенности, что $p_2^{(1)} < p_1^{(2)}$. При любом расположении неизвестной вероятности p в интервале $[0, 1]$ возможны следующие случаи: 1) $p \in I_1^-$; 2) $p \in I_2$; 3) $p \in I_1^+$.

Если гипотеза H является истинной, то вероятность первого случая не превосходит 0,05, поскольку тогда $p \notin I_1$; если же $p \in I_1$, то $p \in I_2^-$, так как $I_1 \cap I_2 = \emptyset$ и $p \notin I_2^+$, поэтому вероятность второго случая также не превосходит 0,05. Наконец, если $p \in I_1^+$, то $p \notin I_1$, поэтому вероятность этого события не превышает 0,05.

Таким образом, во всех трех случаях уровень значимости предложенного критерия не превышает $2\beta = 0,05$. Заметим, однако, что полученная оценка завышена при больших значениях объемов обучающих выборок n_1 и n_2 . Действительно, в этом случае распределение частот h_1 и h_2 приближается к симметричному распределению, поэтому вероятность попадания p в интервалы I_1^- , I_1^+ , I_2^- не превышает половины уровня значимости, т.е. 0,025. Следовательно, при больших значениях n_1 и n_2 уровень значимости критерия, построенного на основании непересекающихся доверительных интервалов I_1 и I_2 , не превосходит 0,025.

Другой критерий для сравнения двух вероятностей в случае, когда выборка в схеме Бернулли B_2 является большой ($n_2 \geq 150 + 200$), а схеме B_1 — малой (т.е. $n_1 \leq 30$). Согласно этому критерию гипотеза H отклоняется, если середина интервала I_2 не принадлежит интервалу I_1 . Для оценки уровня значимости этого критерия заметим, что середина интервала I_2 практически совпадает с вероятностью p_2 , поэтому в случае истинности основной гипотезы H вероятность события $p \notin I_1$ не превосходит 0,05, так что ошибка критерия не превышает 0,05.

2.13.5. Статистический критерий для проверки гипотезы о равенстве двух вероятностей, основанный на правиле $2s$

Пусть x и y — две произвольные случайные величины, имеющие одинаковые математические ожидания $m(x) = m(y) = m$, которые принимают свои значения после проведения одного случайного эксперимента E . Покажем, что для среднеквадратического отклонения $\sigma(z)$ их разности $z = x - y$ справедливо неравенство

$$\sigma(z) = \sigma(x - y) \leq \sigma(x) + \sigma(y).$$

Действительно,

$$\sigma(z) = \sqrt{m(x - y)^2} = \sqrt{m[x - y + m(x) - m(y)]^2} \leq \sqrt{m[x - m(x)]^2} + \sqrt{m[y - m(y)]^2} = \sigma(x) + \sigma(y).$$

Пусть $x = h_1$, $y = h_2$, тогда $\sigma(h_1 - h_2) \leq \sigma(h_1) + \sigma(h_2) = \delta$, причем в силу результатов, изложенных в подразделе 2.13.1, имеем

$$\delta = \sqrt{\frac{p_1 q_1}{n_1}} + \sqrt{\frac{p_2 q_2}{n_2}} \approx \tilde{\delta} = \sqrt{\frac{h_1(1-h_1)}{n_1}} + \sqrt{\frac{h_2(1-h_2)}{n_2}}.$$

Рассмотрим следующий статистический критерий для проверки гипотезы H о равенстве неизвестных вероятностей p_1 и p_2 : если не выполняется хотя бы одно из двух следующих неравенств

$$h_1 - h_2 - 2\tilde{\delta} < 0;$$

$$h_1 - h_2 + 2\tilde{\delta} > 0,$$

то гипотеза H отклоняется; в противном случае считается, что экспериментальные данные не противоречат гипотезе H . Будем называть предложенный тест $2s$ -критерием.

Теорема 2.7.

В случае реализации схем Бернулли B_1 и B_2 уровень значимости $2s$ -критерия приближенно равен 0,05.

Подчеркнем, что в теореме 2.7 предположение о независимости моделей Бернулли B_1 и B_2 , а потому и частот h_1 и h_2 , снято, так что $2s$ -критерий можно применять к любым испытаниям Бернулли.

Резюме

Чаще всего в приложениях в качестве доверительных границ для неизвестной вероятности в схеме Бернулли используется границы интервала Вильсона

$$p_1 = \frac{hn + \frac{1}{2}g^2 - g\sqrt{h(1-h)n + \frac{1}{4}g^2}}{n + g^2},$$

$$p_2 = \frac{hn + \frac{1}{2}g^2 + g\sqrt{h(1-h)n + \frac{1}{4}g^2}}{n + g^2},$$

где n — количество испытаний, h — частота события, $g = 1,96$ для 5%-ного уровня значимости доверительного интервала (p_1, p_2). Другой метод построения доверительных границ основан на правиле $2s$ и утверждает, что доверительный интервал $(h - 2s_n, h + 2s_n)$, где $s_n^2 = \frac{h(1-h)}{n}$, содержит неизвестную вероятность p с вероятностью, приблизительно равной 0,95.

Статистический критерий для сравнения двух неизвестных вероятностей p_1 и p_2 в ситуации, когда частоты h_1 и h_2 могут быть зависимыми, можно построить следующим образом. Пусть $I_1 = (p_1^{(1)}, p_1^{(2)})$ и $I_2 = (p_1^{(2)}, p_2^{(2)})$ — доверительные интервалы для неизвестных вероятностей p_1 и p_2 соответственно, уровень значимости которых приближенно равен 5%. Основная (нулевая) гипотеза H отклоняется, если эти доверительные интервалы не пересекаются. Если же они пересекаются, то считается, что экспериментальные данные не противоречат гипотезе H . В качестве альтернативы можно предложить также критерий, в котором гипотеза H отклоняется, если середина интервала I_2 не принадлежит интервалу I_1 . Поскольку середина интервала I_2 практически совпадает с вероятностью p_2 , в случае истинности основной гипотезы H вероятность события $p \notin I_1$ не превосходит 0,05, так что ошибка критерия не превышает 0,05.

2.14. “Точные” доверительные границы для неизвестной вероятности

Во многих применениях медицинской статистики необходимо строить точный доверительный интервал, который содержит неизвестную вероятность p успеха в моделях Бернулли (так называемую биномиальную долю). Эта задача была исследована во многих работах, например [5, 28]. Цель этого раздела — построить “точный” доверительный интервал, который содержит неизвестную вероятность p успеха в классической и обобщенной моделях Бернулли. Термин “точный” взят в кавычки по следующим причинам. Как уже указывалось, уровень значимости α одного из самых популярных доверительных интервалов для неизвестной вероятности p в схеме Бернулли — интервала Вильсона — зависит от p и может варьироваться в широких пределах (например, если объем выборки равен 10, то минимальное значение $\alpha(p)$ равно 0,012, а максимальное — 0,096). В связи с этим представляет интерес построение такого доверительного интервала (p_0, p_1) для неизвестной вероятности p , уровень значимости которого при любых значениях p не превышал бы определенное число α_0 (например, 0,05). Именно такие доверительные интервалы называются “точными”. Это название оправдано тем, что надежность интервала, т.е. вероятность $p(x \in (p_1, p_2))$ при

всех p превышает $1 - \alpha_0$. Таким образом, для “точных” доверительных интервалов можно точно оценить минимальное значение его уровня значимости.

Рассмотрим следующий критерий для проверки неоднородности двух генеральных совокупностей. Пусть G_x и G_y — генеральные совокупности с неизвестными непрерывными функциями распределения $F_x(u)$ и $F_y(u)$ соответственно, $x = (x_1, x_2, \dots, x_n)$ — выборка из G_x и $y = (y_1, y_2, \dots, y_n)$ — выборка из G_y . Необходимо проверить, совпадают неизвестные функции распределения $F_x(u)$ и $F_y(u)$ (гипотеза H_0) или нет (гипотеза H_1). Если справедлива гипотеза H_0 , то составленная выборка $x_1, x_2, \dots, x_n, y_1, \dots, y_n$ является однородной, в противном случае составленная выборка является неоднородной.

Чтобы решить поставленную задачу, построим вариационный ряд $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$, считая, что $x_{(0)} = -\infty$ и $x_{(n+1)} = \infty$. Рассмотрим случайный интервал $I_{i,q} = (x_{(i)}, x_{(i+q)})$, где i и q — фиксированные числа, $0 \leq i \leq n$, $1 \leq q \leq n - i + 1$. Схема испытаний формулируется так: на k -м шаге ($k = 1, 2, \dots, m$) мы проверяем, принадлежит ли выборочное значение y_k интервалу $I_{i,q}$. Получаем множество событий $A_k = \{y_k \in I_{i,q}\}$, $k = 1, 2, \dots, m$. Каждое событие может происходить с определенной вероятностью $p_k = P(A_k)$, $k = 1, \dots, m$. Введем в рассмотрение случайную переменную κ , которое равняется количеству событий A_k , возникающих в m испытаниях. Если справедлива гипотеза H_0 , то все вероятности p_k одинаковы и равняются

$$p_q = P(A_k | H_0) = \frac{q}{n+1}. \quad (2.47)$$

Эта схема называется *обобщенной схемой Бернулли*. В работах [86, 87] было вычислено распределения вероятностей случайной величины κ :

$$P(\kappa = l | H_0) = \frac{C_{l+q-1}^{l-1} C_{m+n-l-q}^{m-l}}{C_{m+n}^m}, \quad (2.48)$$

$$l = 1, 2, \dots, m,$$

где n , m — объемы выборок x и y соответственно, q — фиксированное число, которое равняется количеству порядковых статистик в интервале $I_{i,q}$, C_r^s — количество сочетаний из r элементов по s .

Построим точный доверительный интервал $I(\kappa)$, который содержит вероятность (2.48), используя значения случайной переменной κ . Напомним, что слово “точный” означает, уровень значимости этого доверительного интервала не превышает заданного числа β (как правило, $\beta = 0,05$). С помощью этого интервала можно построить такой критерий для проверки гипотез H_0 и H_1 .

Строим на основе выборки x вариационный ряд и случайный интервал $I_{i,q}$, где i и q — фиксированные числа.

Вычисляем статистику κ , которая равна количеству элементов из выборки y , которые попадают в интервал $I_{i,q}$.

С помощью статистики κ строим доверительный интервал $I(\kappa)$, уровень значимости которого равен 2β .

Если интервал $I(k)$ не содержит вероятности p_k , гипотеза H_0 отвергается, в противном случае она не отвергается.

Построим доверительный интервал $I(k)$. Пусть x — произвольная целочисленная случайная величина с распределением $p_n(k) = p(x = k)$, $k = 0, 1, \dots, n$. Эта случайная величина порождает функцию $\varphi_x(k)$, определенную на множестве $M_x = \{0, 1, \dots, n\}$ по формуле $\varphi_x(k) = p_k(x)$. Рассмотрим произвольный отрезок $I = \{k: k_0 \leq k \leq k_1, 0 \leq k_0, k_1 \leq n\} \subset M$. Назовем этот отрезок I *областью монотонности* функции $\varphi_x(k)$, если из условия $k \geq i$, ($k, i \in I$) вытекает, что $\varphi(k) \geq \varphi(i)$. Целочисленная случайная величина x называется *одномодальной*, если множество ее значений M_x можно представить как объединение областей монотонности функции $\varphi_x(k)$. Например, случайная величина с биномиальным распределением и случайная величина с распределением (2.48) являются одномодальными.

Замечание 1. Концепция одномодальной целочисленной случайной величины, если $n > 2$, отличается от понятия, предложенного А. Хинчиным. Действительно, по Хинчину функция распределения $F_x(u)$ одномодальной случайной величины x является выпуклой на луче $(-\infty, x_0)$, где x_0 — мода величины x , и вогнутой на луче (x_0, ∞) . Итак, функция $F_x(u)$ может иметь не больше одной точки разрыва. Вместе с тем функция распределения дискретной случайной величины является ступенчатой. Следовательно, если $n > 2$, количество точек разрыва может быть больше двух.

Кроме целочисленной случайной величины x с распределением вероятностей $p_n(k) = p(x = k)$, $k = 0, 1, \dots, n$, рассмотрим непрерывную случайную величину y со следующей плотностью вероятности:

$$f(u) = \begin{cases} 0, & \text{если } u \leq 1, \\ p_n(k), & \text{если } k \leq u \leq k+1, k=0, 1, \dots, n, \\ 0, & \text{если } u \geq n+1. \end{cases}$$

Будем называть ее функцией, порождающей непрерывную случайную величину. Случайная величина y порождает случайную величину x с помощью функции $x = \lfloor y \rfloor = \text{Int}(y)$, отображающей каждое число $y \in R^1$ в ее целую часть $\lfloor y \rfloor = \text{Int}(y)$. Знак равенства в предшествующей формуле означает следующее. Обозначим через E_x случайное испытание, порождающее случайную величину x с заданным распределением вероятности $p_n(k)$, а через E_z — независимое случайное испытание, порождающее случайную величину z , равномерно распределенную на отрезке $[0, 1]$. В составном случайном испытании $E_y = (E_x, E_z)$ случайная величина $y = x + z$ имеет плотность $f(u)$, а случайная величина $\text{Int}(y)$ принимает значение k тогда и лишь тогда, когда $x = k$ в результате испытания E_x . Итак, мы можем считать, что составное случайное испытание E_y порождает случайные величины x и y , так что совпадают не только распределения величин x и $\text{Int}(y)$, но и значение x и $\text{Int}(y)$.

Пусть x — случайная величина с биномиальным распределением, а y — непрерывная случайная величина, индуцирующая переменную x . В модели Бернулли математическое ожидание $m(y)$ и дисперсия $\sigma^2(y)$ равны

$$\begin{aligned} m(y) &= \int_0^{n+1} uf(u) du = \sum_{k=0}^n \int_k^{k+1} uf(u) du = \sum_{k=0}^n p_n(k) \int_k^{k+1} u du = \\ &= \frac{1}{2} \sum_{k=0}^n (2k+1) p_n(k) = \sum_{k=0}^n k p_n(k) + \frac{1}{2} \sum_{k=0}^n p_n(k) = np + \frac{1}{2}; \\ m(y^2) &= \sum_{k=0}^n p_k \int_k^{k+1} u^2 du = \frac{1}{3} \sum_{k=0}^n p_k (3k^2 + 3k + 1) = \sum_{k=0}^n k^2 p_k + \sum_{k=0}^n k p_k + \frac{1}{3} \sum_{k=0}^n p_k = \\ &= m(x^2) + m(x) + \frac{1}{3} = (\sigma^2(x) + (m(x))^2) + np + \frac{1}{3} = npq + (np)^2 + np + \frac{1}{3}. \\ \sigma^2(y) &= npq + \frac{1}{12}, \\ \sigma(y) &= \sqrt{npq + \frac{1}{12}}. \end{aligned}$$

В обобщенной модели Бернулли

$$\begin{aligned} m(x) &= mp_q, \\ \sigma^2(x) &= \frac{(m+n+1)m}{n+2} p_q (1-p_q). \end{aligned}$$

Следовательно, в обобщенной модели Бернулли

$$\begin{aligned} m(y) &= mp_q + \frac{1}{2}, \\ \sigma^2(y) &= \frac{(m+n+1)m}{n+2} p_q (1-p_q) + \frac{1}{12}. \end{aligned}$$

Рассмотрим произвольный фиксированный доверительный интервал (a, b) , содержащий основную массу генеральной совокупности G , с уровнем значимости α . Поскольку $\text{Int}(y)$ является неубывающей функцией, следствием случайного события $A_y = \{y \in [a, b]\} = \{a \leq y \leq b\}$ является событие $A_x = \{x = \text{Int}(y) \in [\text{Int}(a), \text{Int}(b)]\}$. Таким образом, уровень значимости замкнутого доверительного интервала $[\text{Int}(a), \text{Int}(b)]$ для основной массы генеральной совокупности G_x не превышает α .

Далее

$$[\text{Int}(a), \text{Int}(b)] \subset [a-1, b],$$

следовательно,

$$p(x \in [\text{Int}(a), \text{Int}(b)]) \leq p(x \in [a-1, b]).$$

Таким образом, уровень значимости доверительного интервала $[a-1, b]$ для основной массы генеральной совокупности G_x также не превышает α :

$$p\{x \notin [a-1, b]\} \leq \alpha.$$

Легко видеть, что целочисленная случайная величина x является одномодальной тогда и только тогда, когда непрерывная случайная величина y является одномодальной по Хинчину. Для таких случайных величин y выполняется неравенство Гаусса–Височанского–Петунина [44]

$$p(|y - m(y)| \geq \lambda \sigma(y)) \leq \frac{4}{9} \frac{1}{\lambda^2},$$

где $\lambda > \sqrt{8/3}$.

Итак, уровень значимости доверительного интервала

$$[m(y) - \lambda \sigma(y), m(y) + \lambda \sigma(y)],$$

содержащего основную массу генеральной совокупности G , не превышает $\alpha = \frac{4}{9} \frac{1}{\lambda^2}$. В частности, если $\lambda = 3$, мы получаем неравенство $\alpha = \frac{4}{81} < 0.05$.

Положим в классической модели Бернулли

$$a = m(y) - \lambda \sigma(y) = np + \frac{1}{2} - \lambda \sqrt{npq + \frac{1}{12}},$$

$$b = m(y) + \lambda \sigma(y) = np + \frac{1}{2} + \lambda \sqrt{npq + \frac{1}{12}}.$$

На основе предшествующих соображений приходим к выводу, что доверительный интервал $[a-1, b]$ содержит основную массу значений случайной величины x с биномиальным распределением, т.е. доверительный интервал

$$I = \left[np - \frac{1}{2} - \lambda \sqrt{npq + \frac{1}{12}}, np - \frac{1}{2} + \lambda \sqrt{npq + \frac{1}{12}} \right]$$

имеет уровень значимости, не превышающий $\alpha = \frac{4}{9} \frac{1}{\lambda^2}$, где $\lambda > \sqrt{8/3}$.

Случайное событие $\{x \in I\}$ можно переписать в такой форме:

$$|x - np| \leq \frac{1}{2} + \lambda \sqrt{npq + \frac{1}{12}}.$$

Таким образом, в модели Бернулли

$$p \left(|h - p| \leq \frac{1}{2n} + \frac{\lambda}{n} \sqrt{npq + \frac{1}{12}} \right) \geq 1 - \alpha.$$

Для того чтобы построить доверительный интервал для неизвестной вероятности, содержащий неизвестную вероятность p , на основе частоты h в модели Бернулли, состоящей из n испытаний, рассмотрим две функции, зависящие от $p \in [0, 1]$:

$$\varphi(p) = |h - p|$$

$$\psi(p) = \frac{1}{2n} + \frac{\lambda}{n} \sqrt{np(1-p) + \frac{1}{12}}.$$

Обозначим

$$\tilde{\psi}(p) = \sqrt{np(1-p) + \frac{1}{12}}, \quad p \in R^1.$$

Легко проверить, что график функции $\tilde{\psi}(p)$, $p \in R^1$, представляет собой верхнюю часть эллипса E , который проходит через точки $A = \left(\frac{1}{2n} \left(n + \sqrt{\frac{n}{3} + n^2} \right), 0 \right)$, $B = \left(\frac{1}{2}, \sqrt{\frac{1}{12n} + \frac{1}{4}} \right)$, $C = \left(\frac{1}{2n} \left(n - \sqrt{\frac{n}{3} + n^2} \right), 0 \right)$, $D = \left(\frac{1}{2}, -\sqrt{\frac{1}{12n} + \frac{1}{4}} \right)$ с центром в точке $\left(\frac{1}{2}, 0 \right)$. График функции $\psi(p)$ строится на основе графика $\tilde{\psi}(p)$ на сегменте $[0, 1]$ его расширением или сжатием с коэффициентом $\frac{\lambda}{n}$ и переносом на величину $\frac{1}{2n}$.

Таким образом, график функции $\psi(p)$, не зависящей от h , представляет собой дугу эллипса Γ , проходящую через точки $(0, \psi(0))$, $\left(\frac{1}{2}, \psi\left(\frac{1}{2}\right) \right)$, $(1, \psi(1))$, так что функция $\psi(p)$ достигает своего минимума в точке $p = \frac{1}{2}$ и является симметричной относительно этой точки.

Нижняя доверительная граница p_1 является корнем квадратного уравнения

$$\left(1 + \frac{\lambda^2}{n} \right) p^2 - \left(\frac{\lambda^2}{n} - \frac{1}{n} + 2h \right) p + h^2 - \frac{h}{n} + \frac{1}{4n^2} \left(1 - \frac{\lambda^2}{3} \right) = 0. \quad (2.49)$$

Если $h > \psi(0) = \frac{1}{2n} + \frac{\lambda}{n\sqrt{12}}$, то нижняя доверительная граница p_1 является наименьшим корнем уравнения (2.49). Если $h \leq \psi(0)$, то $p_1 = 0$.

Аналогично, верхняя доверительная граница p_2 является корнем квадратного уравнения

$$\left(1 + \frac{\lambda^2}{n} \right) p^2 - \left(\frac{\lambda^2}{n} + \frac{1}{n} + 2h \right) p + h^2 + \frac{h}{n} + \frac{1}{4n^2} \left(1 - \frac{\lambda^2}{3} \right) = 0. \quad (2.50)$$

Если $1 - h > \psi(0)$, то верхняя доверительная граница p_2 является наибольшим корнем уравнения (2.50). Если $1 - h \leq \psi(1)$, то $p_2 = 1$.

Замечание 2. Поскольку $p_1 \leq h \leq p_2$, доля успехов всегда лежит в доверительном интервале $[p_1, p_2]$.

Для обобщенной схемы Бернулли аналогичные соображения приводят к следующему квадратному уравнению:

$$\left(1 + \frac{(m+n+1)\lambda^2}{(n+2)m}\right)p^2 + \left(\frac{1}{m} - \frac{(m+n+1)\lambda^2}{(n+2)m} - 2h\right)p + h^2 - \frac{h}{m} + \frac{1}{4m^2}\left(1 - \frac{\lambda^2}{3}\right) = 0. \quad (2.51)$$

Если $h > \frac{1}{2m} + \frac{\lambda}{m\sqrt{12}} = \gamma$, то нижняя доверительная граница p_1 в обобщенной схеме Бернулли является наименьшим корнем уравнения (2.51). Если $h \leq \gamma$, то $p_1 = 0$.

Аналогично, верхняя доверительная граница p_2 в обобщенной схеме Бернулли является корнем уравнения:

$$\left(1 + \frac{(m+n+1)\lambda^2}{(n+2)m}\right)p^2 - \left(\frac{1}{m} + \frac{(m+n+1)\lambda^2}{(n+2)m} + 2h\right)p + h^2 + \frac{h}{m} + \frac{1}{4m^2}\left(1 - \frac{\lambda^2}{3}\right) = 0. \quad (2.52)$$

Если $1 - h > \gamma$, то верхняя доверительная граница p_2 является наибольшим корнем уравнения (2.52). Если $1 - h \leq \gamma$, то $p_2 = 1$. Вследствие предыдущих результатов приходим к выводу, что уровень значимости построенного доверительного интервала не превышает $\frac{4}{9} \frac{1}{\lambda^2}$ (в частности, 0,05 для $\lambda = 3$).

Резюме

Доверительные границы p_1 и p_2 неизвестной вероятности p называются “точными”, если уровень значимости доверительного интервала (p_1, p_2) при всех значениях p не превосходит заданную границу β (например, 0,05).

В классической схеме Бернулли нижняя доверительная граница p_1 является корнем квадратного уравнения

$$\left(1 + \frac{\lambda^2}{n}\right)p^2 - \left(\frac{\lambda^2}{n} - \frac{1}{n} + 2h\right)p + h^2 - \frac{h}{n} + \frac{1}{4n^2}\left(1 - \frac{\lambda^2}{3}\right) = 0.$$

Если $h > \psi(0) = \frac{1}{2n} + \frac{\lambda}{n\sqrt{12}}$, то нижняя доверительная граница p_1 является наименьшим корнем этого уравнения. Если $h \leq \psi(0)$, то $p_1 = 0$. Аналогично, верхняя доверительная граница p_2 является корнем квадратного уравнения

$$\left(1 + \frac{\lambda^2}{n}\right)p^2 - \left(\frac{\lambda^2}{n} + \frac{1}{n} + 2h\right)p + h^2 + \frac{h}{n} + \frac{1}{4n^2}\left(1 - \frac{\lambda^2}{3}\right) = 0.$$

Если $1 - h > \psi(1)$, то верхняя доверительная граница p_2 является наибольшим корнем этого уравнения. Если $1 - h \leq \psi(1)$, то $p_2 = 1$.

В обобщенной схеме Бернулли нижняя доверительная граница p_1 является корнем квадратного уравнения:

$$\left(1 + \frac{(m+n+1)\lambda^2}{(n+2)m}\right)p^2 + \left(\frac{1}{m} - \frac{(m+n+1)\lambda^2}{(n+2)m} - 2h\right)p + h^2 - \frac{h}{m} + \frac{1}{4m^2}\left(1 - \frac{\lambda^2}{3}\right) = 0.$$

Если $h > \frac{1}{2m} + \frac{\lambda}{m\sqrt{12}} = \gamma$, то нижняя доверительная граница p_1 является наименьшим корнем этого уравнения. Если $h \leq \gamma$, то $p_1 = 0$. Верхняя доверительная граница p_2 является корнем уравнения

$$\left(1 + \frac{(m+n+1)\lambda^2}{(n+2)m}\right)p^2 - \left(\frac{1}{m} + \frac{(m+n+1)\lambda^2}{(n+2)m} + 2h\right)p + h^2 + \frac{h}{m} + \frac{1}{4m^2}\left(1 - \frac{\lambda^2}{3}\right) = 0.$$

Если $1 - h > \gamma$, то верхняя доверительная граница p_2 является наибольшим корнем этого уравнения. Если $1 - h \leq \gamma$, то $p_2 = 1$.

Глава 3

Применение статистических методов в медицине и радиобиологии

3.1. Компьютерный цитогенетический метод диагностики рака молочной и щитовидной железы

3.1.1. Введение

Рак молочной железы является одним из самых опасных и распространенных заболеваний у женщин; согласно данным медицинской статистики это заболевание занимает первое место по смертности — из 10 умерших женщин две умирают от рака молочной железы. Несмотря на то, что молочная железа не является жизненно важным органом, во второй и третьей стадии раковая опухоль в молочной железе интенсивно метастазирует, вызывая появление новых очагов во многих органах, что приводит к летальному исходу. В связи с этим исключительно важна ранняя диагностика этого заболевания. К сожалению, в настоящее время такая диагностика несовершенна, поскольку наиболее распространенный среди существующих методов диагностики — маммография — имеет вероятность ошибки (точнее, ошибки первого рода) в среднем 20%, так что каждая пятая пациентка, болеющая раком молочной железы, диагностируется как пациентка с доброкачественной опухолью. В этом разделе будет описан неинвазивный компьютерный цитогенетический метод диагностики рака молочной железы, основанный на исследовании содержания и распределения ДНК в интерфазных ядрах базального эпителия. Этот метод имеет значительно меньшую диагностическую ошибку и не представляет никакой опасности для здоровой пациентки. Необходимый для изложения этого метода математический аппарат описан в главах 1 и 2.

В настоящее время хорошо исследована цитологическая реактивность онкологических пациентов [51], но анализ разнообразных эффектов и показателей изменения соматических немалигнизованных клеток при предопухолевых и

опухолевых процессах у человека показывает, что эти отклонения по сравнению с нормой настолько малы и слабо выражены, что вряд ли могут использоваться для диагностики. Очевидно, некоторые определенные взаимоотношения между отдельными показателями имеют важное значение для диагностики, но представляется невозможным найти их без современных компьютеров, учитывая множественность и неограниченность вариантов.

В связи с этим мы предлагаем использовать современные методы статистической и геометрической теории распознавания образов, разработанные авторами на протяжении последних 20 лет. Мы предлагаем компьютерную диагностику рака молочной железы на основе соскоба со слизистой оболочки ротовой полости. Для диагностики мы получаем множество данных, изучая около 20–30 эпителиальных клеток от каждой пациентки. Этот метод основывается на хорошо известном эффекте отдаленной экстраоральной малигнизации при количественных цитологических оценках слизистой оболочки полости рта [17, 18].

3.1.2. Меры близости между выборкой и множеством обучающих выборок, используемые для компьютерной диагностики рака груди

Процесс диагностики раковой опухоли в молочной железе состоит из трех этапов.

На первом этапе на основе соскоба со слизистой оболочки полости рта мы можем получить множество данных, изучая 10–30 клеток.

На втором этапе, используя реакцию Фельгена, мы получаем ДНК (дезоксирибонуклеиновая кислота) сканограмму интерфазных ядер этих клеток. Затем мы записываем основные показатели (всего 15) сканограмм интерфазных ядер этих клеток: $x_1, x_2, \dots, x_{15}$. Мы назовем вектор $x_c = (x_1, x_2, \dots, x_{15})$ вектором показателей клетки C . Не обязательно использовать только указанные показатели. Можно предложить другие показатели, более информативные и дающие более точную оценку вероятности малигнизации молочной железы. Следовательно, мы получаем для каждой клетки (например, клетки под номером i) 15 показателей $x_1^{(i)}, x_2^{(i)}, \dots, x_{15}^{(i)}$, где $i = 1, 2, \dots, n$ ($10 \leq n \leq 30$).

На третьем этапе мы применяем специальные алгоритмы статистической и геометрической теории распознавания образов для определения изменений в соматических немалигнизованных клетках при раке молочной железы и фиброаденоматозе.

Эти тесты и алгоритмы будут рассмотрены более детально в следующем разделе. Здесь мы описываем три стадии упомянутых выше тестов и главные меры близости, на которых они основаны.

На первой стадии мы формируем две группы пациентов G_1 и G_2 ; первая группа включает пациенток, страдающих раком молочной железы (РМЖ), а вторая состоит из пациенток с фиброаденоматозом (ФАМ) (диагнозы пациенток в каждой группе должны быть установлены точно!).

Это — так называемые обучающие выборки. На основе этих групп мы будем диагностировать заболевания.

На второй стадии квадратичного теста с помощью p -статистик мы вычисляем расстояния (меры близости) между показателями сканограмм пациента и соответствующими показателями пациентов групп G_1 и G_2 . Это делается следующим образом. Предположим, что пациент Q принадлежит первой группе G_1 : $Q \in G_1$, $G_1 = (Q_1, \dots, Q_n)$, $Q = Q_i$ ($i = 1, 2, \dots, n$). Мы исключаем пациента $Q = Q_i$ из группы G_1 и получаем группу $G_1^{(i)} = \{Q_1, \dots, Q_{i-1}, Q_{i+1}, \dots, Q_n\}$.

Пусть

$$\begin{aligned} X_{C_1}^{(k)} &= (x_{1k}^{(1)}, x_{2k}^{(1)}, \dots, x_{15k}^{(1)}) \\ X_{C_2}^{(k)} &= (x_{1k}^{(2)}, x_{2k}^{(2)}, \dots, x_{15k}^{(2)}) \\ &\dots\dots\dots \\ X_{C_h}^{(k)} &= (x_{1k}^{(j_k)}, x_{2k}^{(j_k)}, \dots, x_{15k}^{(j_k)}), \end{aligned}$$

где ($k = 1, 2, \dots, n$; $10 \leq j_k \leq 30$) — вектор показателей клеток пациента Q_k . Здесь $x_{C_i}^{(k)} = (x_{1k}^{(i)}, x_{2k}^{(i)}, \dots, x_{15k}^{(i)})$ — вектор показателей клетки C_i пациента Q_k , $k = 1, 2, \dots, n$. Затем мы формируем обучающие выборки для каждого показателя x_i , $i = 1, 2, \dots, 15$; первая обучающая выборка для показателя x_1 :

$$\begin{aligned} X_1^{(1)} &= (x_{11}^{(1)}, x_{11}^{(2)}, \dots, x_{11}^{(j_1)}) \text{ (от первого пациента)} \\ X_2^{(1)} &= (x_{12}^{(1)}, x_{12}^{(2)}, \dots, x_{12}^{(j_2)}) \text{ (от второго пациента)} \\ &\dots\dots\dots \\ X_n^{(1)} &= (x_{1n}^{(1)}, x_{1n}^{(2)}, \dots, x_{1n}^{(j_n)}) \text{ (от } n\text{-го пациента);} \end{aligned}$$

вторая обучающая выборка (для показателя x_2):

$$\begin{aligned} X_1^{(2)} &= (x_{21}^{(1)}, x_{21}^{(2)}, \dots, x_{21}^{(j_1)}) \text{ (от первого пациента)} \\ X_2^{(2)} &= (x_{22}^{(1)}, x_{22}^{(2)}, \dots, x_{22}^{(j_2)}) \text{ (от второго пациента)} \\ &\dots\dots\dots \\ X_n^{(2)} &= (x_{2n}^{(1)}, x_{2n}^{(2)}, \dots, x_{2n}^{(j_n)}) \text{ (от } n\text{-го пациента)} \end{aligned}$$

и т.д.

После этого мы получим последнюю обучающую выборку, т.е. выборку для 15-го показателя $X_1^{(15)}, X_2^{(15)}, \dots, X_n^{(15)}$ (n — количество пациентов в группе G_1). Теперь мы можем вычислить значения p -статистик (см. раздел 2.11) для выборок $X_i^{(15)}, X_i^{(15)}, \dots, X_i^{(15)}$ i -го пациента и соответствующих выборок для других пациентов с номером k ($k \neq i$) (i фиксировано!):

$$\rho_{ik}^{(1)} = \rho(X_i^{(1)}, X_k^{(1)}), \rho_{ik}^{(2)} = \rho(X_i^{(2)}, X_k^{(2)}), \dots, \rho_{ik}^{(15)} = \rho(X_i^{(15)}, X_k^{(15)})$$

и найти значения усредненных p -статистик

$$\rho_i^{(1)} = \frac{1}{n-1} \sum_{k=1, k \neq i}^n \rho(X_i^{(1)}, X_k^{(1)}),$$

$$\rho_i^{(2)} = \frac{1}{n-1} \sum_{k=1, k \neq i}^n \rho(X_i^{(2)}, X_k^{(2)}),$$

.....

$$\rho_i^{(15)} = \frac{1}{n-1} \sum_{k=1, k \neq i}^n \rho(X_i^{(15)}, X_k^{(15)})$$

(i фиксировано!), которые представляют меру близости между Q_i (точнее, его показателями) и группой $G_i^{(1)} (i = 1, 2, \dots, n)$.

Заменяя пациента Q_i пациентом $\bar{Q}_i$ из группы G_2 (вспомним, что G_2 состоит из пациентов с фиброаденоматозом), мы получим аналогичные усредненные p -статистики для группы G_2 : $\bar{\rho}_i^{(1)}, \bar{\rho}_i^{(2)}, \dots, \bar{\rho}_i^{(15)}$ ($i = 1, 2, \dots, m$; $m = \text{card } G_2$):

$$\bar{\rho}_i^{(t)} = \frac{1}{n} \sum_{k=1}^n \rho(\bar{X}_i^{(t)}, X_k^{(t)}), (i = 1, 2, \dots, m; t = 1, 2, \dots, 15),$$

где $\bar{X}_i^{(t)}$ — соответствующий показатель пациента $\bar{Q}_i$.

Если мы заменим группу G_1 на G_2 и проведем аналогичные вычисления, мы получим усредненные p -статистики $d_i^{(1)}, d_i^{(2)}, \dots, d_i^{(15)}$ ($i = 1, 2, \dots, n$) и $\bar{d}_j^{(1)}, \bar{d}_j^{(2)}, \dots, \bar{d}_j^{(15)}$ ($j = 1, 2, \dots, m$):

$$d_i^{(t)} = \frac{1}{m-1} \sum_{k=1, k \neq i}^m \rho(\bar{X}_i^{(t)}, \bar{X}_k^{(t)}), (i = 1, 2, \dots, m; t = 1, 2, \dots, 15),$$

$$\bar{d}_j^{(t)} = \frac{1}{m} \sum_{k=1}^m \rho(X_j^{(t)}, \bar{X}_k^{(t)}), (j = 1, 2, \dots, n; t = 1, 2, \dots, 15).$$

На третьей стадии квадратичного теста мы проводим объединение этих усредненных p -статистик $(\rho_i^{(t)}, \rho_i^{(s)}), (\bar{\rho}_j^{(t)}, \bar{\rho}_j^{(s)})$ ($i = 1, 2, \dots, n; j = 1, 2, \dots, m; t, s = 1, 2, \dots, 15$) и аналогично получаем точки $(d_i^{(t)}, d_i^{(s)}), (\bar{d}_j^{(t)}, \bar{d}_j^{(s)})$ ($i = 1, 2, \dots, n; j = 1, 2, \dots, m; t, s = 1, 2, \dots, 15$). Затем мы строим так называемые “доверительные эллипсы” E_{α} , содержащие усредненные p -статистики $(\rho_i^{(t)}, \rho_i^{(s)})$ ($i = 1, 2, \dots, n$) для группы G_1 , т.е. эллипсы минимальной площади, содержащие точки $(\rho_i^{(t)}, \rho_i^{(s)})$ ($i = 1, 2, \dots, n; t, s = 1, 2, \dots, 15$). Точнее, мы строим эллипсы E_{α} с помощью алгоритма, который дает приближенное решение этой проблемы [102].

Затем мы определяем доверительные эллипсы $\bar{E}_{\alpha}$ для усредненных p -статистик $(\bar{\rho}_j^{(t)}, \bar{\rho}_j^{(s)})$ ($j = 1, 2, \dots, m; t, s = 1, 2, \dots, 15$) и аналогичные эллипсы E_{α}^* и $\bar{E}_{\alpha}^*$, используя точки $(d_i^{(t)}, d_i^{(s)}), (\bar{d}_j^{(t)}, \bar{d}_j^{(s)})$ ($i = 1, 2, \dots, n; j = 1, 2, \dots, m; t, s = 1, 2, \dots, 15$) соответственно.

Кроме того, для описания так называемого “линейного теста” мы рассматриваем дискриминантные функции Фишера $f_u(u, v)$ ($f_v^*(u, v)$), отделяющие множество $M_u^p = \{(\rho_i^{(t)}, \rho_i^{(s)}), i = 1, 2, \dots, n\}$ от множества $\bar{M}_u^p = \{(\bar{\rho}_j^{(t)}, \bar{\rho}_j^{(s)}), j = 1, 2, \dots, m\}$ и множество $M_u^d = \{(d_i^{(t)}, d_i^{(s)}), i = 1, 2, \dots, n\}$ от множества $\bar{M}_u^d = \{(\bar{d}_j^{(t)}, \bar{d}_j^{(s)}), j = 1, 2, \dots, m\}$. Функция $f_u(u, v)$ построена таким образом, что прямая линия $l_u^p = \{(u, v) : f_u(u, v) = 0\}$ перпендикулярна отрезку, соединяющему центры множеств M_u^p и $\bar{M}_u^p$ и проходит через середину этого отрезка; аналогично определяется функция $f_v^*(u, v)$. Кроме того, будем предполагать, что центр множества M_u^p принадлежит нижней полуплоскости π_u , а центр множества $\bar{M}_u^p$ принадлежит верхней полуплоскости λ_u (аналогично π_u^*, λ_u^*). Следовательно, мы имеем 15 показателей и 210 пар эллипсов $(E_u, \bar{E}_u)$ и $(E_u^*, \bar{E}_u^*)$ ($t < s; t, s = 1, 2, \dots, 15$) и, кроме того, 210 пар полуплоскостей (π_u, λ_u) , (π_u^*, λ_u^*) ($t < s; t, s = 1, 2, \dots, 15$). Пусть Q — пациент, страдающий раком молочной железы (гипотеза H_1) или фиброаденоматозом (гипотеза H_2). Используя алгоритмы, упомянутые выше, мы можем вычислить для этого пациента усредненные p -статистики $\rho_Q^{(t)}, d_Q^{(t)}$ ($t = 1, 2, \dots, 15$):

$$\rho_Q^{(t)} = \frac{1}{n} \sum_{k=1}^n \rho(X_Q^{(t)}, X_k^{(t)}), d_Q^{(t)} = \frac{1}{m} \sum_{k=1}^m \rho(X_Q^{(t)}, \bar{X}_k^{(t)}),$$

где $X_Q^{(t)}$ — соответствующий показатель пациента Q , и сформировать точки $(\rho_Q^{(t)}, \rho_Q^{(s)}), (d_Q^{(t)}, d_Q^{(s)})$ ($t < s; t, s = 1, 2, \dots, 15$).

Рассмотрим следующие случайные события:

$$\begin{aligned} A_1 &= \{(\rho_Q^{(t)}, \rho_Q^{(s)}) \in E_u\}, A_2 = \{(\rho_Q^{(t)}, \rho_Q^{(s)}) \in \bar{E}_u\}, A_3 = \{(\rho_Q^{(t)}, \rho_Q^{(s)}) \in E_u - \bar{E}_u\}, \\ A_4 &= \{(\rho_Q^{(t)}, \rho_Q^{(s)}) \in \bar{E}_u - E_u\}, A_1^* = \{(d_Q^{(t)}, d_Q^{(s)}) \in E_u^*\}, A_2^* = \{(d_Q^{(t)}, d_Q^{(s)}) \in \bar{E}_u^*\}, \\ A_3^* &= \{(d_Q^{(t)}, d_Q^{(s)}) \in E_u^* - \bar{E}_u^*\}, A_4^* = \{(d_Q^{(t)}, d_Q^{(s)}) \in \bar{E}_u^* - E_u^*\}, B_1 = \{(\rho_Q^{(t)}, \rho_Q^{(s)}) \in \pi_u\}, \\ B_2 &= \{(\rho_Q^{(t)}, \rho_Q^{(s)}) \in \lambda_u\}, B_1^* = \{(d_Q^{(t)}, d_Q^{(s)}) \in \pi_u^*\}, B_2^* = \{(d_Q^{(t)}, d_Q^{(s)}) \in \lambda_u^*\}, t < s, C_1 = A_3 \cup A_4^*, \\ C_2 &= A_4 \cup A_3^*, C_3 = A_1 \cup A_2^*, C_4 = A_2 \cup A_1^*, C_5 = B_1 \cup B_2^*, C_6 = B_2 \cup B_1^* \end{aligned}$$

и обозначим через $h_i = h(C_i)$, $i = 1, 2, \dots, 6$, частоту события C_i при 210 тестах (экспериментах), когда $t, s = 1, 2, \dots, 15; t < s$. Используя формулы (2.28), мы можем получить асимптотические доверительные границы, соответствующие заданному уровню значимости для вероятности $p_i = p(C_i)$ на основе частоты $h_i = h(C_i)$ (в связи с этим мы должны взять $h_j = h_j, m = 210$). Как показано в разделе 2.13, эти границы, очевидно, слишком широки. Мы назовем частоту h_i показателем РМЖ, а h_3 — тотальным показателем РМЖ (сокращенно ТРМЖ),

поскольку эти показатели являются мерой близости между сканограммами интерфазных ядер клеток исследуемого пациента Q и соответствующими сканограммами пациентов, страдающих карциномой молочной железы. На основании подобных аргументов мы обозначим частоты h_2 и h_4 как показатель ФАМ и тотальный показатель ФАМ (т.е. ТФАМ) соответственно. Аналогично мы обозначим частоты h_5 и h_6 как линейный показатель РМЖ (т.е. ЛРМЖ) и линейный показатель ФАМ (т.е. ЛФАМ).

Рассмотрим теперь так называемый “порядковый тест”. Первый этап этого теста аналогичен предыдущим тестам. Пусть $x_{C_i}^{(k)} = (x_{1k}^{(i)}, x_{2k}^{(i)}, \dots, x_{15k}^{(i)})$ ($i = 1, 2, \dots, j_k; k = 1, 2, \dots, n$) — вектор показателей клетки C_i пациента Q_k из группы G_1 , а $Y_{D_i}^{(k)} = (y_{1k}^{(i)}, y_{2k}^{(i)}, \dots, y_{15k}^{(i)})$, ($i = 1, 2, \dots, l_k; k = 1, 2, \dots, m$) — соответствующий вектор показателей пациента $\bar{Q}_k \in G_2$. На втором этапе вычисляется усредненный вектор показателей

$$\hat{X}^{(k)} = \frac{1}{j_k} \sum_{i=1}^{j_k} X_{C_i}^{(k)} = (\hat{x}_{1k}, \hat{x}_{2k}, \dots, \hat{x}_{15k})$$

для каждого пациента $Q_k \in G_1$; аналогично, для каждого пациента $\bar{Q}_k \in G_2$ усредненный вектор показателей $\hat{Y}^{(k)}$ имеет вид

$$\hat{Y}^{(k)} = \frac{1}{l_k} \sum_{i=1}^{l_k} Y_{D_i}^{(k)} = (\hat{y}_{1k}, \hat{y}_{2k}, \dots, \hat{y}_{15k}).$$

Положим

$$\begin{aligned} X_{ik}^{\min} &= \min(x_{ik}^{(1)}, x_{ik}^{(2)}, \dots, x_{ik}^{(j_k)}), \quad k=1, 2, \dots, n; \quad t=1, 2, \dots, 15; \\ X_{ik}^{\max} &= \max(x_{ik}^{(1)}, x_{ik}^{(2)}, \dots, x_{ik}^{(j_k)}), \quad k=1, 2, \dots, n; \quad t=1, 2, \dots, 15; \\ Y_{ik}^{\min} &= \min(y_{ik}^{(1)}, y_{ik}^{(2)}, \dots, y_{ik}^{(l_k)}), \quad k=1, 2, \dots, m; \quad t=1, 2, \dots, 15; \\ Y_{ik}^{\max} &= \max(y_{ik}^{(1)}, y_{ik}^{(2)}, \dots, y_{ik}^{(l_k)}), \quad k=1, 2, \dots, m; \quad t=1, 2, \dots, 15; \\ a_t^{\min} &= \min(x_{t1}^{\min}, x_{t2}^{\min}, \dots, x_{tn}^{\min}), \quad a_t^{\max} = \max(x_{t1}^{\min}, x_{t2}^{\min}, \dots, x_{tn}^{\min}), \\ b_t^{\min} &= \min(x_{t1}^{\max}, x_{t2}^{\max}, \dots, x_{tm}^{\max}), \quad b_t^{\max} = \max(x_{t1}^{\max}, x_{t2}^{\max}, \dots, x_{tm}^{\max}), \\ \bar{a}_t^{\min} &= \min(y_{ik}^{\min}, k=1, 2, \dots, m); \quad \bar{a}_t^{\max} = \max(y_{ik}^{\min}, k=1, 2, \dots, m); \\ \bar{b}_t^{\min} &= \min(y_{ik}^{\max}, k=1, 2, \dots, m); \quad \bar{b}_t^{\max} = \max(y_{ik}^{\max}, k=1, 2, \dots, m); \\ c_t^{\min} &= \min(\hat{x}_{ik}, k=1, 2, \dots, n), \quad c_t^{\max} = \max(\hat{x}_{ik}, k=1, 2, \dots, n), \\ \bar{c}_t^{\min} &= \min(\hat{y}_{ik}, k=1, 2, \dots, m), \quad \bar{c}_t^{\max} = \max(\hat{y}_{ik}, k=1, 2, \dots, m), \end{aligned}$$

$t = 1, 2, \dots, 15$; тогда $a_t^{\min}, a_t^{\max}$ будут минимальной и максимальной порядковыми статистиками соответственно, как и $b_t^{\min}, b_t^{\max}$, $\bar{a}_t^{\min}, \bar{a}_t^{\max}$, $\bar{b}_t^{\min}, \bar{b}_t^{\max}$, $c_t^{\min}, c_t^{\max}$, $\bar{c}_t^{\min}, \bar{c}_t^{\max}$. С помощью этих порядковых статистик определяются доверительные

интервалы $\alpha_i = (a_i^{\min}, a_i^{\max})$, $\beta_i = (b_i^{\min}, b_i^{\max})$, $\bar{\alpha}_i = (\bar{a}_i^{\min}, \bar{a}_i^{\max})$, $\bar{\beta}_i = (\bar{b}_i^{\min}, \bar{b}_i^{\max})$, $\gamma_i = (c_i^{\min}, c_i^{\max})$, $\bar{\gamma}_i = (\bar{c}_i^{\min}, \bar{c}_i^{\max})$.

Пусть Q — обследуемый пациент и $X_C = (x_1^{(i)}, x_2^{(i)}, \dots, x_{15}^{(i)})$, $i = 1, 2, \dots, j$ — вектор показателей этого пациента. На третьем этапе порядкового теста мы вычисляем усредненные векторы пациентов Q_j :

$$\hat{X} = \frac{1}{j} \sum_{i=1}^j X_{C_i} = (\hat{x}_1, \hat{x}_2, \dots, \hat{x}_{15})$$

и показатели $x_t^{\min} = \min(x_t^{(i)}, i = 1, 2, \dots, j)$, $x_t^{\max} = \max(x_t^{(i)}, i = 1, 2, \dots, j)$, $t = 1, 2, \dots, 15$ а затем определяем индикаторы выхода за границы $I_t^{\min}, I_t, I_t^{\max}$:

$$I_t^{\min} = \begin{cases} 1, & \text{если } x_t^{\min} \notin \alpha_t, \\ 0, & \text{если } x_t^{\min} \in \alpha_t, \end{cases} \quad I_t = \begin{cases} 1, & \text{если } \hat{x}_t \notin \gamma_t, \\ 0, & \text{если } \hat{x}_t \in \gamma_t, \end{cases} \quad I_t^{\max} = \begin{cases} 1, & \text{если } x_t^{\max} \notin \beta_t, \\ 0, & \text{если } x_t^{\max} \in \beta_t, \end{cases}$$

$t = 1, 2, \dots, 15$.

Аналогично определяются индикаторы $\bar{I}_t^{\min}, \bar{I}_t, \bar{I}_t^{\max}$. Далее, мы оцениваем показатели

$$\alpha_1 = \sum_{t=1}^{15} (I_t^{\min} + I_t + I_t^{\max}), \quad \alpha_2 = \sum_{t=1}^{15} (\bar{I}_t^{\min} + \bar{I}_t + \bar{I}_t^{\max}).$$

Эти показатели также являются мерами близости между сканограммами интерфазных ядер клеток обследуемого пациента Q и соответствующими сканограммами пациентов, болеющих раком молочной железы и фиброаденоматозом соответственно. Необходимость и приемлемость введения и использования этих мер близости $h_i = h(C)$, $i = 1, 2, \dots, 6$ и α_1, α_2 основывается на математических результатах, доказанных раньше (см. разделы 2.11, 2.13).

3.2. Цитоспектрофотометрические исследования изменений слизистой оболочки рта, ассоциированных с малигнизацией молочной железы

3.2.1. Биомедицинские основы

В настоящее время хорошо известно, что организм обладает потенциальной способностью адекватно отвечать на изменения внешней и внутренней среды в целом или некоторыми своими внутренними системами (нервной, эндокринной, иммунной системой и т.п.). Эта способность называется реактивностью организма [48]. Существует множество реакций, поскольку они проявляются в различных системах, органах и клетках организма, более того, мы можем говорить о цитологической реактивности организма, подразумевая под этим

изменения, возникающие в клетках в ответ на разнообразные факторы. Проблема цитологической реактивности во время образования злокачественной или доброкачественной опухоли в организме является довольно сложной и мало изученной. В 1962 году Г. Нейбург и соавторы [17] сообщили о характерных изменениях в клетках буккального эпителия пациента с опухолью, локализованной вне полости рта, и назвали это изменениями, ассоциированными с малигнизацией (MAC — Malignancy Assotiated Changes). Эти изменения характеризуются возрастанием размеров эпителиоцитов, разрывами ядерной мембраны, возрастанием размеров зон, связанных с хроматином, окруженным светлыми областями. Эти авторы регистрировали похожие изменения при наличии предопухолевых процессов в организме. В большом количестве работ, которые появились в это время, исследователи изучали качественные показатели, характеризующие влияние опухоли на различные структурные элементы организма, удаленные от опухоли, но лишь некоторые авторы сообщили об использовании количественных методов оценок состояния клеток, посвященных определению оптической плотности и размеров ядер.

К. Обрапалска с соавторами [18] обнаружили свойства MAC в буккальном эпителии пациентов со злокачественными опухолями в 74% наблюдений и у 31% пациентов без неоплазии. В 1990 году Г.Р. Огден с соавторами [19] вновь попытались охарактеризовать и обосновать представления о наличии влияния опухоли на функциональное состояние буккального эпителия, для того чтобы получить данные, характеризующие закономерности динамики процессов, возникающих в органах и системах, удаленных от опухоли. Обнаруженные изменения авторы назвали опухоль-ассоциированными изменениями (MAC), включающими следующие параметры: разрушение ядерной мембраны, неоднородность ядра и наличие многочисленных Фельген-отрицательных зон. Эти изменения были обнаружены у 77% больных с опухолями различной локализации (карциномы, лимфомы, семиномы). В исследованиях были использованы количественные методы оценки полученных результатов: цитоспектрофотометрические исследования ДНК, размеров ядер и цитоплазмы. Однако эти авторы не смогли выявить какой-либо закономерности, которая характеризовала бы опухолевый процесс, за исключением увеличения размеров ядер. Несмотря на это, они не отрицали существования влияния опухоли на функциональные отклонения, возникающие в буккальном эпителии, что согласуется с результатами, полученными в наших исследованиях [49].

Проблема взаимоотношений между опухолями и организмом изучалась на протяжении многих лет в Институте экспериментальной патологии, онкологии и радиобиологии Национальной Академии наук Украины. Многочисленные экспериментальные и клинические исследования установили, что существуют сложные взаимоотношения между опухолью и организмом-носителем, которые начинаются с момента трансформации нормальной соматической клетки и продолжаются на протяжении всей опухолевой болезни. Эти исследования, посвященные изучению изменений в буккальном эпителии, показали, что они представляют собой неспецифическую реакцию на наличие в нем опухолевого процесса.

3.2.2. Материал и метод

Для исследования были взяты группы женщин, страдающих раком молочной железы (на 2-й и 3-й стадиях) и фиброаденоматозом в возрасте от 25 до 53 лет (всего 104 пациентки). После полоскания и снятия поверхностного слоя клеток слизистой оболочки полости рта были получены соскобы с различной глубины шиповатого слоя (условно они обозначены как срединный и глубокий). Соскобы были высушены при комнатной температуре и зафиксированы в смеси Никифорова, а затем проведена реакция Фельгена с холодным гидролизом в 5 н. HCl на протяжении 15 минут при температуре $t = 21-22^{\circ}\text{C}$. Оптическая плотность ядра регистрировалась цитоспектрофотометром МЦФУ2МТ (ЛМО) с помощью метода сканирования с длиной волны 575 нм и диаметром зонда 0,05 мкм. В каждом препарате исследовалось от 10 до 30 клеток. Содержание ДНК-фуксина в ядрах эпителиоцитов определялось как произведение оптической плотности на площадь. Таким образом, при исследовании интерфазных ядер мы получали сканограмму распределения ДНК, которая является прямоугольной таблицей (матрицей) $R = (r_{ij})_{i=1, m}^{j=1, n}$, где r_{ij} характеризует содержание ДНК в ячейке сетки с номером (i, j) , m — количество строк и n — количество столбцов в матрице R ($5 \leq m, n \leq 10$). Ниже приведена сканограмма пациентов, страдающих фиброаденоматозом и раком молочной железы соответственно.

Сканограмма пациентки ФАМ:

0,046	0,128	0,251	0,234	0,156	0,170	0,233	0,147	0,065
0,147	0,147	0,222	0,220	0,128	0,152	0,222	0,249	0,125
0,046	0,064	0,129	0,222	0,152	0,147	0,155	0,254	0,147
0,133	0,147	0,258	0,222	0,131	0,132	0,222	0,146	0,048
0,064	0,155	0,170	0,262	0,170	0,128	0,226	0,247	0,156
0,152	0,153	0,348	0,336	0,242	0,149	0,251	0,156	0,147
0,170	0,251	0,254	0,224	0,147	0,147	0,224	0,258	0,228
0,348	0,244	0,219	0,336	0,265	0,221	0,132	0,222	0,150
0,262	0,252	0,244	0,214	0,254	0,341	0,334	0,219	0,228

Сканограмма пациентки РМЖ:

0,129	0,147	0,131	0,147	0,133	0,147	0,132	0,133	0,067
0,050	0,070	0,125	0,133	0,132	0,069	0,131	0,149	0,147
0,126	0,156	0,067	0,125	0,147	0,153	0,128	0,252	0,222
0,254	0,222	0,252	0,258	0,258	0,132	0,147	0,152	0,258
0,259	0,250	0,250	0,256	0,150	0,262	0,228	0,279	0,257
0,228	0,154	0,221	0,170	0,150	0,228	0,152	0,147	0,254
0,221	0,257	0,222	0,147	0,147	0,149	0,250	0,250	0,226
0,222	0,232	0,228	0,156	0,251	0,262	0,152	0,149	0,153
0,251	0,229	0,247	0,252	0,147	0,224	0,149	0,149	0,153

Значения $m \times n$ называются размером регистрационного поля сканограммы. При формировании обучающих выборок для обеспечения их статистической однородности этот размер варьировался в довольно узких пределах от 56 до 81.

Морфо- и денситометрические показатели интерфазного ядра. Сканограмма, или цифровой портрет клетки, представляет собой матрицу $R = \{r_{ij} \}_{i=1, n}^{j=1, m}$, где r_{ij} — значение поточечной оптической плотности хроматина интерфазного ядра клетки, выраженное в условных единицах; n, m — количество точек сканограмм по вертикали и горизонтали соответственно.

На основании этой цитоспектрофотометрической информации вычисляются следующие морфо- и денситометрические показатели, характеризующие структурные и текстурные особенности хроматина¹:

1. Площадь ядра:

x_1 — количество элементов матрицы R , таких что $r_{ij} \geq 0,08$.

2. Площадь конденсированного хроматина:

x_2 — количество элементов матрицы R , таких что $r_{ij} \geq 0,35$.

3. Площадь деконденсированного хроматина:

x_3 — количество элементов матрицы R , таких что $0,08 \leq r_{ij} < 0,35$.

4. Площадь сильно деконденсированного хроматина:

x_4 — количество элементов матрицы R , таких что $0,08 \leq r_{ij} < 0,15$.

5. Удельная площадь конденсированного хроматина:

$$x_5 = \frac{x_2}{x_1}$$

6. Удельная площадь деконденсированного хроматина:

$$x_6 = \frac{x_3}{x_1}$$

7. Интегральная плотность:

$$x_7 = \sum_{i=1}^m \sum_{\substack{j=1 \\ r_{ij} \geq 0.08}}^n r_{ij}$$

¹ См. Капанян А.Л., Каралова Е.М., Магакян Ю.А. Коэффициент взаиморасположения частиц и его использование при количественном анализе вещества в клетках // Цитология. — 1988. — 30, № 3. — С. 361–366; Папаян Г.В., Магакян Ю.А., Агроскин Л.С., Каралова Е.М. Использование данных сканирующей цитометрии для анализа поглощающего вещества в клетках // Цитология. — 1982. — 24, № 11. — С. 1359–1366; Бойко Ю.В., Литвинко П.Г., Петунин Ю.И., Ганина К.П. Показатель кластерности структуры хроматина интерфазных ядер // Докл. АН СССР. — 1990. — № 5. — С. 75–78; Магакян Ю.А., Каралова Е.М. // Цитофотометрия ДНК. — Ереван: Изд-во АН Армении, 1989. — 203 с.

где неравенство $r_{ij} \geq 0,08$ означает, что суммирование проводится лишь по тем индексам i и j , для которых $r_{ij} \geq 0,08$.

8. Средняя плотность:

$$x_8 = \frac{x_1}{nm - p},$$

где p — количество точек (i, j) , в которых $r_{ij} < 0,08$.

9. Усредненная сумма перепадов:

$$x_9 = \frac{1}{q} \sum_{k=1}^q v_k,$$

где q — количество точек (i, j) , таких что $\min(r_{ij}, r_{i+1,j}, r_{i,j+1}, r_{i+1,j+1}) \geq 0,08$;

$$v_k = \max(r_{ij}, r_{i+1,j}, r_{i,j+1}, r_{i+1,j+1}) - \min(r_{ij}, r_{i+1,j}, r_{i,j+1}, r_{i+1,j+1}), \quad k = \overline{1, q}.$$

(Суммирование проводится по вышеуказанным элементам.)

10. Общий показатель кластерности:

$$x_{10} = \frac{1}{q} \sum_{k=1}^q v_k^2.$$

11. Дисперсионный коэффициент:

$$x_{11} = \left(\frac{\sum_{k=1}^q (v_k - x_9)^2}{q - 1} \right)^{1/2}$$

12. Показатель вариации перепадов:

$$x_{12} = x_9 + x_{11}.$$

13. Показатель рельефности:

$$x_{13} = \frac{\sum_{i=2}^m \sum_{j=1}^n |r_{ij} - r_{i-1,j}|}{(2mn - m + n - q)},$$

где q — количество точек (i, j) , таких что $\max(r_{ij}, r_{i-1,j}) < 0,08$.

14. Текстульный коэффициент:

$$x_{14} = \frac{x_{13}}{\varepsilon}, \quad \varepsilon = \frac{\sum_{i=1}^m \sum_{j=1}^n (r_{ij} - x_7)}{\overline{r_{ij} \geq 0.08} \cdot mn - p}.$$

15. Коэффициент взаиморасположения:

$$x_{15} = \frac{a}{x_8^2 b},$$

$$\text{где } a = \sum_{i=1}^m \sum_{j=1}^n \left(\sum_{k=1}^m \sum_{l=j+1}^n \frac{r_{ij} r_{kl}}{(k-i)^2 + (l-j)^2} + \sum_{k=i+1}^m \sum_{l=1}^n \frac{r_{ij} r_{kl}}{(k-i)^2 + (l-j)^2} \right),$$

$$b = \sum_{i=1}^m \sum_{j=1}^n \left(\sum_{k=1}^m \sum_{l=j+1}^n \frac{1}{(k-i)^2 + (l-j)^2} + \sum_{k=i+1}^m \sum_{l=1}^n \frac{1}{(k-i)^2 + (l-j)^2} \right),$$

суммирование проводится лишь по тем элементам, для которых

$$\min(r_{ij}, r_{kl}) > 0.875 \max_{i=1,2,\dots,n; j=1,2,\dots,m; r_{ij} \geq 0.08} r_{ij}.$$

3.2.3. Калибровка обучающих выборок

Диагностика онкологических заболеваний, предлагаемая авторами, основывается на применении тестов и алгоритмов статистической и геометрической теории распознавания образов.

На первой стадии сформируем две группы сканограмм пациентов $A = \{X_i\}_{i=1, \overline{N}}$ и $B = \{Y_j\}_{j=1, \overline{M}}$, диагнозы которых точно идентифицированы. В дальнейшем для определенности будем полагать, что группа A содержит сканограммы больных со злокачественной опухолью (рак молочной железы), а группа B — с доброкачественным новообразованием (фиброаденоматоз). После процедур регистрации и определения морфо- и денситометрических показателей мы получаем так называемые “обучающие” выборки для каждого показателя x_k ($k = 1, 2, \dots, 15$): $G_A^{(1)}, G_A^{(2)}, \dots, G_A^{(15)}$ для пациентов группы A и $G_B^{(1)}, G_B^{(2)}, \dots, G_B^{(15)}$ для пациентов группы B .

Рассмотрим проблему определения количества обучающих выборок в группах A и B , обеспечивающего достаточно высокий уровень достоверности диагностики.

Вначале естественно предположить, что количество выборок в группах A и B должно быть одинаковым. Для принятия или отклонения этой гипотезы мы предложили *процедуру калибровки обучающих выборок*, которая состоит из следующих этапов.

1. Исключим пациента $X_i, i = \overline{1, N}$ (или $Y_j, j = \overline{1, M}$) из множества $A \cup B$.
2. На основе множества выборок $\{A \cup B\} \setminus X_i$ (или $\{A \cup B\} \setminus Y_j$) построим тесты, использующие пары эллипсов $(E_u, \bar{E}_u)$, $(E_u^*, \bar{E}_u^*)$ и полуплоскостей (π_u, λ_u) , (π_u^*, λ_u^*) .
3. Вычислим значения статистик $h_k = h(C_k)$ ($k = \overline{1, 6}$) для пациента $X_i, i = \overline{1, N}$ (или $Y_j, j = \overline{1, M}$).

4. Возвратим пациента $X_i, i = \overline{1, N}$ (или $Y_j, j = \overline{1, M}$) в множество $A \cup B$ и повторим процедуру для следующего пациента.

В табл. 3.1 и 3.2 представлены результаты калибровок для случая, когда множество A состояло из сканограмм 25 больных раком молочной железы (РМЖ), а B из сканограмм 25 больных фиброаденоматозом (ФАМ).

Таблица 3.1. Значения статистик h_k ($k = 1, 6$) для выборок сканограмм пациентов с РМЖ при калибровке обучающих выборок (25 РМЖ и 25 ФАМ)

Номер пациента	h_1 РМЖ	h_2 ФАМ	h_3 ТРМЖ	h_4 ТФАМ	h_5 ЛРМЖ	h_6 ЛФАМ
1	0,00952	0,00476	0,98571	0,98095	0,38095	0,61905
2	0,05238	0,17143	0,75238	0,87143	0,22857	0,77143
3	0,02857	0,00000	0,99524	0,96667	0,74762	0,25238
4	0,00952	0,33810	0,65714	0,98571	0,44286	0,55714
5	0,03333	0,18095	0,81429	0,96190	0,23810	0,76190
6	0,00952	0,04762	0,93333	0,97143	0,42857	0,57143
7	0,02857	0,00476	0,99524	0,97143	0,70000	0,30000
8	0,04762	0,13810	0,82857	0,91905	0,60476	0,39524
9	0,01429	0,08095	0,91905	0,98571	0,69524	0,30476
10	0,04762	0,07619	0,90476	0,93333	0,78571	0,21429
11	0,06667	0,10476	0,87619	0,91429	0,64286	0,35714
12	0,08095	0,09524	0,86667	0,88095	0,70952	0,29048
13	0,02381	0,01905	0,96190	0,95714	0,34286	0,65714
14	0,00000	0,01905	0,98095	1,00000	0,86190	0,13810
15	0,08571	0,23333	0,70476	0,85238	0,72857	0,27143
16	0,07143	0,07143	0,81905	0,81905	0,66190	0,33810
17	0,00476	0,00000	1,00000	0,99524	0,49048	0,50952
18	0,00476	0,03333	0,96190	0,99048	0,34762	0,65238
19	0,02381	0,01905	0,98095	0,97619	0,83333	0,16667
20	0,00000	0,00000	1,00000	1,00000	0,42381	0,57619
21	0,00476	0,03810	0,95238	0,98571	0,54762	0,45238
22	0,10476	0,04286	0,93333	0,87143	0,71905	0,28095
23	0,04762	0,25714	0,61905	0,82857	0,40952	0,59048
24	0,01905	0,08571	0,88571	0,95238	0,58095	0,41905
25	0,00000	0,01905	0,98095	1,00000	0,40952	0,59048

Таблица 3.2. Значения статистик h_k ($k = 1, 6$) для выборок сканограмм пациентов с ФАМ при калибровке обучающих выборок (25 РМЖ и 25 ФАМ)

Номер пациента	h_1 РМЖ	h_2 ФАМ	h_3 ТРМЖ	h_4 ТФАМ	h_5 ЛРМЖ	h_6 ЛФАМ
1	0,00000	0,03810	0,96190	1,00000	0,25238	0,74762
2	0,03333	0,35238	0,56667	0,88571	0,33810	0,66190
3	0,00000	0,31429	0,68095	0,99524	0,21429	0,78571
4	0,03333	0,08095	0,89524	0,94266	0,83333	0,16667
5	0,00952	0,01905	0,98095	0,99048	0,50952	0,49048
6	0,00476	0,02381	0,95714	0,97619	0,70476	0,29524
7	0,00000	0,03810	0,96190	1,00000	0,25714	0,74286
8	0,03810	0,20952	0,75238	0,92381	0,22857	0,77143
9	0,04286	0,24762	0,56190	0,76667	0,30476	0,69524
10	0,05238	0,19048	0,43333	0,57143	0,20476	0,79524
11	0,04286	0,03810	0,90952	0,90476	0,59048	0,40952
12	0,03810	0,05714	0,91905	0,93810	0,80952	0,19048
13	0,00476	0,05238	0,94762	0,99524	0,86667	0,13333
14	0,10000	0,10952	0,87143	0,88095	0,58095	0,41905
15	0,01905	0,21429	0,78571	0,98095	0,42857	0,57143
16	0,04286	0,00952	0,98571	0,95238	0,68095	0,31905
17	0,06190	0,00952	0,97143	0,91905	0,79524	0,20476
18	0,01429	0,03333	0,96667	0,98571	0,91905	0,08095
19	0,01429	0,02381	0,96667	0,97619	0,53333	0,46667
20	0,07143	0,13333	0,70952	0,77143	0,45238	0,54762
21	0,03810	0,00952	0,96667	0,93810	0,63810	0,36190
22	0,04286	0,10000	0,88571	0,94286	0,20476	0,79524
23	0,16190	0,14762	0,60000	0,58571	0,92857	0,07143
24	0,05238	0,29048	0,42381	0,66190	0,27143	0,72857
25	0,00952	0,11429	0,88571	0,99048	0,24762	0,75238

Введем в рассмотрение следующие критерии диагностики:

- 1) квадратичный: $h_3 > h_4 \Rightarrow \text{РМЖ}$; $h_3 \leq h_4 \Rightarrow \text{ФАМ}$;
- 2) линейный: $h_5 > h_6 \Rightarrow \text{РМЖ}$; $h_5 \leq h_6 \Rightarrow \text{ФАМ}$.

Введем следующие обозначения: D_1 — диагноз “РМЖ”, D_2 — диагноз “ФАМ”, v_{11} — частота события D_1 среди выборок РМЖ, v_{21} — частота события D_2 среди выборок РМЖ, v_{12} — частота события D_1 среди выборок ФАМ, v_{22} — частота события D_2 среди выборок ФАМ.

Анализ результатов, полученных при калибровке выборок из групп *A* и *B* одинакового объема (табл. 3.2), позволяет сделать следующие выводы.

1. В подавляющем большинстве случаев для группы *B* (обучающая выборка пациенток с ФАМ) наблюдается доминирование статистики h_4 (тотальный ФАМ) над h_3 (тотальный РМЖ), а также статистики h_2 (ФАМ) над h_1 (РМЖ) (мы будем называть этот феномен *эффектом устойчивого доминирования*), а для группы *A* такого эффекта не наблюдается.
2. Для линейного критерия события D_1 и D_2 практически равновероятны как для группы *A* (обучающая выборка пациентов с РМЖ), так и для группы *B* (обучающая выборка пациентов с ФАМ), поэтому этот критерий непригоден для дифференциальной диагностики РМЖ и ФАМ.
3. Квадратичный критерий для группы *B* дает значительно лучшие результаты (в 80% случаев ставится правильный диагноз — событие D_2 , а в 20% неправильный диагноз — событие D_1), а для группы *A* наблюдается противоположный эффект — в 28% случаев ставится правильный диагноз, а в 72% — неправильный диагноз, поэтому этот критерий также непригоден для дифференциальной диагностики.

Таблица 3.3. Частота случайного события D_k ($k = 1, 2$) при калибровке обучающих выборок (25 РМЖ и 25 ФАМ)

Частоты Критерии	V_{11}	V_{21}	V_{22}	V_{12}
Квадратичный	0,28	0,72	0,80	0,20
Линейный	0,56	0,44	0,48	0,52
Комбинированный	0,72	0,28	0,80	0,20

Таким образом, использование обучающих выборок *A* и *B* одинакового объема и лишь одних квадратичных и линейных критериев не позволяет получить приемлемые для клинической медицины результаты. В связи с этим была осуществлена калибровка обучающих выборок для случая, когда группа *A* (25 сканограмм пациентов с РМЖ) почти вдвое превышала по объему группу *B* (12 сканограмм пациентов с ФАМ); такой выбор объемов должен был обеспечить превышение (доминирование) статистики h_3 (тотальный РМЖ) над h_4 (тотальный ФАМ), а также h_1 (РМЖ) над h_2 (ФАМ).

Результаты калибровки обучающих выборок (25 РМЖ и 12 ФАМ) приведены в табл. 3.4–3.6. Анализируя эти результаты, мы приходим к следующим выводам.

1. В подавляющем большинстве случаев для группы *A* наблюдается превышение (доминирование) статистики h_3 (тотальный РМЖ) над h_4 (тотальный ФАМ), а также h_1 (РМЖ) над h_2 (ФАМ) (эффект устойчивого доминирования); а для группы *B* такого эффекта не наблюдается.

2. Для линейного критерия события D_1 и D_2 практически равновероятны, поэтому он непригоден для дифференциальной диагностики РМЖ и ФАМ.
3. Квадратичный критерий для группы A дает хорошие результаты (в 92% случаев ставится правильный диагноз — событие D_1 , а в 8% — неправильный событие D_2), а для группы B правильный диагноз ставится в 56% случаев, а неправильный — в 44%; поэтому этот критерий непригоден для дифференциальной диагностики.

Таблица 3.4. Значения статистик h_k ($k = 1, 6$) для выборок сканограмм пациентов с РМЖ при калибровке обучающих выборок (25 РМЖ и 12 ФАМ)

Номер пациента	h_1 РМЖ	h_2 ФАМ	h_3 ТРМЖ	h_4 ТФАМ	h_5 ЛРМЖ	h_6 ЛФАМ
1	0,16667	0,01905	0,97143	0,82381	0,42381	0,57619
2	0,14286	0,08095	0,81429	0,75238	0,27619	0,72381
3	0,07143	0,00000	0,99524	0,92381	0,80952	0,19048
4	0,12857	0,25714	0,69524	0,82381	0,41905	0,58095
5	0,23810	0,11905	0,77143	0,65238	0,26190	0,73810
6	0,05714	0,01905	0,93810	0,90000	0,55238	0,44762
7	0,34286	0,00952	0,90095	0,64762	0,80952	0,19048
8	0,20476	0,04286	0,82381	0,66190	0,60952	0,39048
9	0,18095	0,04762	0,91429	0,78095	0,85238	0,14762
10	0,26190	0,06190	0,89048	0,69048	0,83333	0,16667
11	0,27143	0,02381	0,85238	0,60476	0,68095	0,31905
12	0,35714	0,04286	0,88571	0,57143	0,75714	0,24286
13	0,23333	0,01905	0,95714	0,74286	0,33810	0,66190
14	0,29524	0,00000	0,97619	0,68095	0,91429	0,08571
15	0,32381	0,07619	0,70952	0,46190	0,79524	0,20476
16	0,21905	0,00476	0,80952	0,59524	0,65238	0,34762
17	0,15238	0,00000	1,00000	0,84762	0,51905	0,48095
18	0,11905	0,00476	0,98095	0,86667	0,40000	0,60000
19	0,13810	0,00476	0,99048	0,85714	0,83810	0,16190
20	0,00952	0,00000	1,00000	0,99048	0,66667	0,33333
21	0,16667	0,01905	0,97143	0,82381	0,61905	0,38095
22	0,17143	0,03333	0,92381	0,78571	0,71905	0,28095
23	0,15714	0,22381	0,67619	0,74286	0,38095	0,61905
24	0,20000	0,04286	0,89048	0,73333	0,60000	0,40000
25	0,08571	0,00952	0,97143	0,89524	0,41905	0,58095

Таблица 3.5. Значения статистик h_k ($k = 1, 6$) для выборок сканограмм пациентов с РМЖ при калибровке обучающих выборок (25 РМЖ и 12 ФАМ)

Номер пациента	h_1 РМЖ	h_2 ФАМ	h_3 ТРМЖ	h_4 ТФАМ	h_5 ЛРМЖ	h_6 ЛФАМ
1	0,04762	0,06190	0,92857	0,94286	0,34286	0,65714
2	0,10000	0,25714	0,58571	0,74286	0,31429	0,68571
3	0,02857	0,29048	0,69048	0,95238	0,21429	0,78571
4	0,20476	0,02857	0,91429	0,73810	0,86667	0,13333
5	0,05238	0,01905	0,97143	0,93810	0,71429	0,28571
6	0,11905	0,00952	0,93333	0,82381	0,80476	0,13333
7	0,05238	0,07619	0,91905	0,94286	0,34762	0,65238
8	0,09524	0,20952	0,72857	0,84286	0,27143	0,72857
9	0,14286	0,04286	0,50000	0,40000	0,37143	0,62857
10	0,09524	0,04286	0,43810	0,38571	0,13810	0,86190
11	0,24286	0,01905	0,92857	0,70476	0,55238	0,44762
12	0,21905	0,01429	0,95714	0,75238	0,85714	0,14286

Таблица 3.6. Частота случайного события D_k ($k = 1, 2$) при калибровке обучающих выборок (25 РМЖ и 12 ФАМ)

Частоты Критерии	V_{11}	V_{21}	V_{22}	V_{12}
Квадратичный	0,92	0,08	0,42	0,56
Линейный	0,68	0,32	0,58	0,42
Комбинированный	0,92	0,08	0,58	0,42

Отметим, что описанный выше эффект доминирования статистики h_3 над h_4 для группы A обнаруживается лишь в случае, когда размеры поля регистрации сканограммы варьируют в довольно узких пределах (в данном случае от 56 до 81); если этот принцип нарушается, то получается статистически неоднородная выборка, и эффект доминирования становится слабо выраженным. Это видно из табл. , где приведены результаты калибровки выборок, состоящих из 17 сканограмм больных РМЖ и 8 сканограмм больных ФАМ. Площадь поля регистрации сканограмм этих больных изменялась от 25 до 70.

Подводя итоги проведенных исследований, мы должны констатировать, что для выборок A и B с соотношением объема 2:1 (точнее говоря, $\lceil \frac{1}{2} \text{card } A \rceil = \text{card } B$, где $\lceil x \rceil$ означает целую часть числа x) использование лишь одних квадратичных и линейных критериев не приводит к удовлетворительным результатам. Тем не

менее описанный выше эффект устойчивого доминирования частот, который наблюдается при обучающих выборках одинакового объема ($\text{card } A \approx \text{card } B$) и выборках с объемами $\left[\frac{1}{2} \text{card } A \right] \approx \text{card } B$ позволяет сформулировать так называемый “фильтрующий критерий”, основная идея которого заключается в следующем.

Рассмотрим вначале результаты калибровки для обучающих выборок с одинаковым объемом. Как было указано, в этом случае для группы B (пациентки с ФАМ) возникает эффект устойчивого доминирования статистики h_4 (тотальный ФАМ) над h_3 (тотальный РМЖ), а для группы A (пациентки с РМЖ) этот эффект отсутствует. Пусть Q — пациентка, диагноз которой мы пытаемся поставить. Вычислим для нее на основании этих групп (обучающих выборок) A и B значения статистик $h_3(Q)$ и $h_4(Q)$. Предположим, что $h_3(Q) \geq h_4(Q)$. С какой гипотезой, D_1 или D_2 , лучше всего согласуются эти экспериментальные данные? Поскольку для группы B наблюдается эффект устойчивого доминирования $h_3(Q) \leq h_4(Q)$, то вероятность события “пациентка P ” имеет то же заболевание, что и пациентки группы B (т.е. ФАМ), будет малой, следовательно, у этой пациентки более вероятно заболевание РМЖ, значит, гипотеза D_2 лучше согласуется с экспериментальными данными. Если же для данной пациентки $h_3(Q) < h_4(Q)$, то мы не можем принять на основании двух групп A и B одинакового объема какое-либо решение о заболевании пациентки (т.е. не можем поставить диагноз), поскольку с такими данными могут быть как пациентки с РМЖ, так и с ФАМ.

Мы построили так называемый A -фильтр, который позволяет в некоторых случаях диагностировать РМЖ у обследуемых пациенток. Будем говорить, что сканограммы обследуемой пациентки проходят A -фильтр, если $h_3(Q) \geq h_4(Q)$, и не проходят A -фильтр, если $h_3(Q) < h_4(Q)$.

Перейдем теперь к описанию B -фильтра. Для этой цели используется другая пара обучающих выборок A и B с объемами $\left[\frac{1}{2} \text{card } A \right] \approx \text{card } B$, т.е. когда объем A вдвое превышает объем B . Данные калибровки показывают, что для этого случая наблюдается обратный эффект устойчивого доминирования $h_3(Q) > h_4(Q)$ для группы A . Следовательно, значения статистик $h_3^*(Q)$ и $h_4^*(Q)$, полученные для обследуемого пациента с помощью этой пары обучающих выборок A и B , которые удовлетворяют неравенству $h_3^*(Q) \leq h_4^*(Q)$, свидетельствуют с большей вероятностью о наличии у пациентки ФАМ, чем РМЖ; если же $h_3^*(Q) > h_4^*(Q)$, то диагноз не ставится (процедура непринятия решения). Таким образом, мы описали вторую часть фильтрующего критерия, который называется B -фильтром, позволяющего в некоторых случаях диагностировать ФАМ у обследуемых пациенток. Будем говорить, что сканограммы обследуемой пациентки проходят B -фильтр, если $h_3^*(Q) \leq h_4^*(Q)$, и не проходят B -фильтр, если $h_3^*(Q) > h_4^*(Q)$.

Схему фильтрующего квадратичного критерия можно описать следующим образом.

1. Формируются две пары обучающих выборок A и B с объемами $\text{card } A \approx \text{card } B$ и $\left[\frac{1}{2} \text{card } A \right] \approx \text{card } B$. Первая пара используется для построения A -фильтра, а вторая — B -фильтра.
2. Проводится описанный выше процесс фильтрации сканограмм обследуемого пациента через A -фильтр и B -фильтр. Если эти сканограммы проходят A -фильтр, то ставится диагноз РМЖ, а если же они проходят B -фильтр — диагноз ФАМ; в противном случае, когда они не проходят ни через один фильтр, диагноз не ставится (процедура непринятия решения).

Отметим, что в том случае, когда группа B второй пары обучающих выборок получается из группы B первой пары за счет удаления половины пациентов, принципиально невозможно прохождение сканограмм обследуемого пациента как через A -фильтр, так и B -фильтр критерия.

Фильтрующий квадратичный критерий представляет определенный интерес для клинической медицины, однако удовлетворительным его нельзя признать, поскольку у этого критерия достаточно велика вероятность непринятия решения.

В связи с этим возникает идея использования фильтрующего комбинированного критерия (линейного, квадратичного и порядкового), к описанию которого мы сейчас переходим.

Фильтрующий комбинированный критерий также состоит из двух фильтров — A -фильтра и B -фильтра. Как и раньше, обозначим через $h_i, i = \overline{1, 8}$ определенные ранее статистики обследуемого пациента P , которые получены с помощью обучающих выборок A и B приблизительно одинакового объема ($\text{card } A \approx \text{card } B$), а через $h_i^*, i = \overline{1, 8}$, — соответствующие статистики, полученные с помощью обучающих выборок A : $\text{card } A = 25$, и B : $\text{card } B = 12$. Будем считать, что сканограммы пациента P проходят A -фильтр, если выполняется хотя бы одно из следующих неравенств:

$$h_3 > h_4; \alpha_1 \leq \alpha_2$$

(т.е. справедливо высказывание $h_3 > h_4 \vee \alpha_1 \leq \alpha_2$), и проходят B -фильтр, если имеет место одно из неравенств:

$$h_3^* \leq h_4^*; \alpha_1^* > \alpha_2^*$$

(т.е. справедливо высказывание $h_3^* \leq h_4^* \vee \alpha_1^* > \alpha_2^*$).

Фильтрующий комбинированный критерий формируется следующим образом: если сканограммы обследуемого пациента проходят A -фильтр, то ставится диагноз РМЖ, если B -фильтр — ФАМ, а во всех остальных случаях диагноз не ставится (процедура непринятия решения). Поскольку мы использовали B -фильтр, построенный на основании группы B второй пары обучающих выборок ($\left[\frac{1}{2} \text{card } A \right] \approx \text{card } B$), которая получена удалением 13 пациентов с ФАМ

группы *B* первой пары обучающих выборок ($\text{card } A \approx \text{card } B$), прохождение сканограмм пациента *P* через оба фильтра принципиально невозможно. Использование процедуры калибровки обучающих выборок для оценки качества этого критерия некорректно, поскольку при этом получаются те же результаты, что и для фильтрующего квадратичного критерия; кроме того, лишь один порядковый критерий в этом случае дает безошибочный диагноз.

В связи с этим для экспериментальной проверки качества предложенных критериев были выбраны 17 больных раком молочной железы (РМЖ) и 7 больных фибroadеноматозом. Все эти пациенты не входили в обучающие выборки *A* и *B*. *A*-фильтр был построен на основании сканограмм 24 больных РМЖ (группа *A*) и 25 больных ФАМ (группа *B*), а *B*-фильтр — на основании сканограмм 25 больных РМЖ (группа *A*) и 12 больных ФАМ (группа *B*).

Результаты тестирования по обоим фильтрам приведены в табл. 3.7–3.10.

Таблица 3.7. Значения статистик h_k ($k = 1, 6$) для выборок сканограмм пациентов с РМЖ и ФАМ при тестировании по А-фильтру (24 РМЖ и 25 ФАМ)

Номер пациента	h_1 РМЖ	h_2 ФАМ	h_3 ТРМЖ	h_4 ТФАМ	h_5 ЛРМЖ	h_6 ЛФАМ	α_1	α_2
Рак молочной железы								
1	0,01905	0,01429	0,81905	0,54286	0,54286	0,45714	5	9
2	0,00000	0,01429	0,98571	1,00000	0,64762	0,35238	1	5
3	0,02857	0,07143	0,67143	0,71429	0,40000	0,60000	6	7
4	0,02857	0,14762	0,60476	0,72381	0,27143	0,72857	6	8
5	0,06667	0,20000	0,39524	0,52857	0,23333	0,76667	13	3
6	0,01905	0,10952	0,82381	0,91429	0,32857	0,67143	7	5
7	0,02381	0,06190	0,64762	0,68571	0,66667	0,33333	7	8
8	0,02857	0,01905	0,98095	0,97143	0,77143	0,22857	1	0
9	0,06190	0,22381	0,31429	0,47619	0,24762	0,75238	14	4
10	0,02381	0,00476	0,99048	9,97143	0,44762	0,55238	0	0
11	0,07143	0,01905	0,88571	0,83333	0,31905	0,68095	2	3
12	0,00952	0,25238	0,37143	0,61429	0,34286	0,65714	13	11
13	0,03810	0,09048	0,72381	0,77619	0,80000	0,20000	5	9
14	0,01905	0,12857	0,54762	0,65714	0,51429	0,48571	9	9
15	0,07619	0,05238	0,63333	0,60952	0,41905	0,58095	10	8
16	0,00952	0,12381	0,68571	0,80000	0,25238	0,74762	8	8
17	0,02857	0,11905	0,60476	0,69524	0,24286	0,75714	11	11
Фибroadеноматоз								
18	0,00000	0,04762	0,95238	1,00000	0,59524	0,40476	1	0
19	0,03333	0,37619	0,28571	0,62857	0,17143	0,82857	19	18

Номер пациента	h_1 РМЖ	h_2 ФАМ	h_3 ТРМЖ	h_4 ТФАМ	h_5 ЛРМЖ	h_6 ЛФАМ	α_1	α_2
20	0,01905	0,21905	0,56667	0,76667	0,26667	0,73333	16	9
21	0,00476	0,08571	0,91429	0,99524	0,17143	0,82857	11	0
22	0,00000	0,39524	0,60476	1,00000	0,88571	0,11429	1	0
23	0,00000	0,53810	0,45714	0,99524	0,16667	0,83333	5	0
24	0,00000	0,10952	0,89048	1,00000	0,18571	0,81429	3	0

Таблица 3.8. Значения статистик h_k ($k = 1, 6$) для выборок сканограмм пациентов с РМЖ и ФАМ при тестировании по В-фильтру (25 РМЖ и 12 ФАМ)

Номер пациента	h_1 РМЖ	h_2 ФАМ	h_3 ТРМЖ	h_4 ТФАМ	h_5 ЛРМЖ	h_6 ЛФАМ	α_1	α_2
Рак молочной железы								
1	0,18095	0,00000	0,80952	0,62857	0,35238	0,44762	5	10
2	0,08095	0,01429	0,98571	0,91905	0,71429	0,28571	1	5
3	0,17143	0,04286	0,66190	0,53333	0,40476	0,59524	6	9
4	0,12381	0,10476	0,60000	0,58095	0,30476	0,69524	6	9
5	0,21905	0,12381	0,41905	0,32381	0,28095	0,71905	13	14
6	0,20952	0,10952	0,79524	0,69524	0,34286	0,65714	7	7
7	0,17143	0,04762	0,60952	0,48571	0,67619	0,32381	7	9
8	0,20476	0,00000	0,97143	0,76667	0,87143	0,12857	1	3
9	0,17143	0,10476	0,35714	0,29048	0,27619	0,72381	14	16
10	0,14286	0,00000	0,98571	0,84286	0,50000	0,50000	0	1
11	0,20952	0,00476	0,89048	0,68571	0,35238	0,64762	2	9
12	0,12857	0,16667	0,36190	0,40000	0,37143	0,62857	13	13
13	0,28095	0,02857	0,72381	0,47143	0,77143	0,22857	5	12
14	0,10476	0,05714	0,52381	0,47619	0,55238	0,44762	9	12
15	0,19048	0,01905	0,60476	0,43333	0,46667	0,53333	10	11
16	0,19524	0,07143	0,70476	0,58095	0,30000	0,70000	8	9
17	0,23333	0,07619	0,63810	0,48095	0,29048	0,70952	11	12
Фиброаденоматоз								
18	0,10476	0,02381	0,95714	0,87619	0,69048	0,30952	1	1
19	0,11429	0,29524	0,32381	0,50476	0,22381	0,77619	19	21
20	0,17143	0,18571	0,56667	0,58095	0,29524	0,70476	16	15
21	0,06190	0,07143	0,86190	0,87143	0,18095	0,81905	11	9

Номер пациента	h_1 РМЖ	h_2 ФАМ	h_3 ТРМЖ	h_4 ТФАМ	h_5 ЛРМЖ	h_6 ЛФАМ	α_1	α_2
22	0,35714	0,08571	0,59524	0,32381	0,93810	0,06190	1	8
23	0,09524	0,24286	0,40476	0,55238	0,20476	0,79524	5	12
24	0,20476	0,09048	0,86190	0,74762	0,29048	0,70952	3	9

Таблица 3.9. Частота случайного события D_k ($k = 1, 2$) при тестировании сканограмм пациентов по А- и В-фильмам

Частоты Критерии	v_{11}	v_{21}	v_{22}	v_{12}
Квадратичный	0,29	0,06	0,22	0,11
Линейный	0,35	0,06	0,67	0,11
Комбинированный	0,94	0,06	0,43	0,0

Таблица 3.10. Значения статистик h_k ($k = 1, 6$) для выборок сканограмм (25 РМЖ и 12 ФАМ)

Номер пациента	h_1 РМЖ	h_2 ФАМ	h_3 ТРМЖ	h_4 ТФАМ	h_5 ЛРМЖ	h_6 ЛФАМ
Рак молочной железы						
1	0,23333	0,07143	0,86190	0,70000	0,77619	0,22381
2	0,18571	0,09524	0,74762	0,65714	0,76190	0,23810
3	0,21429	0,00952	0,98571	0,78095	0,97143	0,02857
4	0,19524	0,00952	0,98571	0,80000	0,41429	0,58571
5	0,02381	0,05714	0,33333	0,96671	0,14286	0,87714
6	0,32381	0,00000	0,97619	0,65238	0,79048	0,20952
7	0,22857	0,03333	0,94286	0,74762	0,69048	0,30952
8	0,17619	0,05174	0,90000	0,78095	0,70952	0,29048
9	0,04762	0,20952	0,68571	0,84762	0,17143	0,82857
10	0,31905	0,10952	0,68571	0,47619	0,57619	0,42381
11	0,44762	0,02381	0,76667	0,34286	0,60476	0,39524
12	0,27619	0,20952	0,65714	0,59048	0,32857	0,67143
13	0,29524	0,06667	0,56667	0,33810	0,37141	0,62857
14	0,39524	0,07143	0,77143	0,44762	0,43333	0,56667
15	0,17143	0,00952	0,93810	0,77619	0,75714	0,24286
16	0,44762	0,02381	0,89048	0,46667	0,66667	0,33333
17	0,31429	0,02857	0,82857	0,54286	0,59524	0,40476

Номер пациента	h_1 РМЖ	h_2 ФАМ	h_3 ТРМЖ	h_4 ТФАМ	h_5 ЛРМЖ	h_6 ЛФАМ
Фиброаденоматоз						
1	0,10952	0,21905	0,57143	0,68095	0,18571	0,81429
2	0,14762	0,07143	0,91905	0,84286	0,19524	0,80476
3	0,10952	0,11429	0,76667	0,77143	0,14762	0,85238
4	0,61429	0,00000	0,92381	0,30952	0,99048	0,00952
5	0,30000	0,06190	0,65714	0,41905	0,20476	0,79524
6	0,31429	0,00000	1,00000	0,68571	0,95238	0,04762
7	0,24762	0,20952	0,61429	0,57619	0,25714	0,74286
8	0,38095	0,04762	0,82381	0,49048	0,45238	0,54762

Анализ экспериментальных результатов показывает, что в случае использования комбинированного фильтрующего критерия мы можем получить три случая:

- мы диагностируем ФАМ для обследуемого пациента;
- мы диагностируем РМЖ;
- мы не диагностируем никакого заболевания (непринятие решения).

Если мы диагностируем ФАМ при условии, что пациент болен РМЖ, то возникает так называемая “ошибка первого рода”. Если мы диагностируем РМЖ при условии, что пациент болен ФАМ, то возникает так называемая “ошибка второго рода”. На основе результатов эксперимента мы можем констатировать следующее (см. табл. 3.7, 3.8): вероятность ошибки первого рода равна 6%, а второго рода — 0%. Это означает, что вероятность (точнее, частота) ФАМ-диагноза для РМЖ-пациента приблизительно равна 0,06, а вероятность РМЖ-диагноза для ФАМ-пациента равна 0,00. Кроме того, вероятность диагностировать заболевание (принятия решения) равна 94% для РМЖ-пациентов и 43% — для ФАМ-пациентов. Таким образом, мы можем диагностировать рак молочной железы с высокой вероятностью после первого анализа, а для более точного определения фиброаденоматоза мы должны повторить анализ.

Замечание. Как было указано, с помощью частот $h_i, i = \overline{1,6}$, можно получить доверительные границы $(p_i^{(H)}, p_i^{(B)})$ для неизвестной вероятности $p = p(c)$, отвечающие 5%-ному уровню значимости по формулам (2.28) из раздела 2.11 (при этом $h_j = h_i, m = 210, g = 3$). Эти доверительные границы $(p_i^{(H)}, p_i^{(B)})$, $i = \overline{1,6}$, могут быть полезными при исследовании прохождения сканограмм через А- и В-фильтры: если $h_3(P) > h_4(P)$ и доверительные границы $(p_3^{(H)}, p_3^{(B)})$ и $(p_4^{(H)}, p_4^{(B)})$ не пересекаются, то это служит еще одним аргументом в пользу прохождения сканограмм через А-фильтр для фильтрующего квадратичного критерия; аналогичное утверждения справедливо для h_5 и h_6 , h_1 и h_2 , а также для $h_i^*, i = \overline{1,6}$.

Резюме

Подводя итог выполненных исследований, мы можем утверждать, что предложенный компьютерный метод диагностики рака молочной железы (РМЖ) и фиброаденоматоза (ФАМ) позволяет идентифицировать с высокой вероятностью рак молочной железы у пациенток, страдающих этим заболеванием, сразу после первого анализа буккального соскоба, поскольку вероятность как ошибочного диагноза, так и непринятия решения не превышает 6%.

Для пациенток, болеющих фиброаденоматозом, вероятность ошибочного диагноза практически равна нулю, однако после первого анализа вероятность непринятия решения равна 43%.

Если решение не принято, мы должны повторить анализ, поскольку мы не имеем абсолютной гарантии, что обследуемая пациентка болеет только РМЖ или ФАМ и что отсутствуют заболевания, искажающие реальную ситуацию. Если анализ повторяется n раз, вероятность непринятия решения приблизительно равна $(\frac{1}{2})^n = \frac{1}{2^n}$ при условии, что результаты анализов являются независимыми (так называемые независимые испытания). Если точно известно, что пациент может болеть только одним из двух заболеваний (РМЖ либо ФАМ), значение $\frac{1}{2^n}$ быстро стремится к нулю и после 5-6 попыток (буккальных соскобов) мы можем диагностировать ФАМ.

Компьютерный метод диагностики является вспомогательным. Его следует применять совместно с другими традиционными методами клинического обследования пациентов (маммография, УЗИ). Кроме того, предложенный метод компьютерной диагностики рака молочной железы полезен при массовых клинических обследованиях населения для определения групп риска.

3.3. Анализ текстурных показателей ядер буккального эпителия при доброкачественных и злокачественных процессах в щитовидной железе

3.3.1. Введение

Рак щитовидной железы (РЩЖ) встречается достаточно редко и составляет во всех странах 1% от общего числа всех злокачественных новообразований в популяции. Однако в последние годы отмечается тенденция к росту этого заболевания. С одной стороны, частота выявления РЩЖ связана с внедрением в практику ряда современных методов исследования больных и с повышенной онкологической настороженностью врачей. С другой стороны, имеются данные о возрастании заболеваемости РЩЖ, связанной с неблагоприятной экологической обстановкой, например, с Чернобыльской катастрофой. В частности,

в Украине заболеваемость злокачественными новообразованиями щитовидной железы в расчете на 100 тысяч населения выросла с 0,9% в 1991 году до 1,2% в 1996 году [109].

При патологии ЩЖ в настоящее время широко распространен метод тонкоигльной аспирационной пункционной биопсии с последующим цитологическим исследованием пунктата. Этот метод имеет ряд преимуществ, а именно: он может использоваться в амбулаторных условиях, не требует применения анестезии, технически прост в исполнении, занимает мало времени и имеет невысокую стоимость. Анализ возможностей цитологической диагностики заболеваний ЩЖ показывает, что с его помощью можно судить о доброкачественности или злокачественности процесса, диагностировать определенные нозологические формы патологии ЩЖ. Однако этот метод имеет и достаточно существенные недостатки: возникают определенные трудности при диагностике сочетанной патологии ЩЖ [85], одних морфологических критериев оказывается недостаточно для постановки диагноза [94], при РЩЖ опухолевые клетки могут распространяться по ходу пункционного отверстия, способствуя диссеминации. Кроме того, точность данного метода зависит от целого ряда факторов и колеблется в весьма широких пределах — от 50 до 98% [82], а по данным ряда авторов [127] — от 40%. Низкий уровень точности дооперационной диагностики патологий ЩЖ и возможные ее осложнения указывают на несовершенство используемого метода и необходимость разработки новых диагностических критериев. С этих позиций особый интерес вызывают исследования слизистой оболочки полости рта (СОПР), или буккального эпителия, и в частности содержание ДНК в ядрах эпителиоцитов для диагностики различных патологических процессов в ЩЖ.

Анатомо-физиологические особенности СОПР и ее взаимосвязь с внутренними органами и системами определяют клинические варианты ее изменений при заболеваниях внутренних органов [32]. В последнее время клетки буккального эпителия широко используют для исследований общего состояния организма [125]. Кроме того, еще в 1962 году Г. Нейбург и др. в работе [17] отметили в эпителиоцитах буккального эпителия больных злокачественными опухолями характерное перераспределение хроматиновых масс, которое они оценили как изменения, связанные со злокачественным ростом, и назвали опухоль-ассоциированными изменениями (МАС — Malignancy Associated Changes). Подобные изменения были описаны нами [49], а также другими исследователями ([19]).

Целью наших исследований было усовершенствование метода дифференциальной диагностики доброкачественных и злокачественных процессов щитовидной железы на основе анализа распределения оптической плотности ДНК в интерфазных ядрах эпителиоцитов буккального эпителия.

Для этого предложено два новых текстурных показателя, основанных на двух различных и взаимно дополняющих интерпретациях информации об оптической плотности ДНК.

3.3.2. Методика исследований

Для изучения были взяты результаты цитоспектрофотометрического исследования содержания ДНК эпителиоцитов СОПР у больных со злокачественными (15 случаев аденокарциномы щитовидной железы АДК) и опухолеподобными (27 случаев аутоиммунного тиреоидита, АТ, и 15 случаев узлового зоба, УЗ) процессами в щитовидной железе. Методика взятия мазков из СОПР и их окраски изложены раньше [49]. Использован цитоспектрофотометр МЦФУ-2МТ (ЛОМО, Россия). Диагноз заболевания ставился на основании клинического и параклинического обследований, а также морфологического исследования патологического очага.

Сканограмма S распределения ДНК интерфазных ядер буккального эпителия представляла собой квадратную матрицу (таблицу), состоящую из чисел s_{ij} , характеризующих содержание ДНК в точке с номером (i, j) , m — количество строк и n — количество столбцов в матрице S . Исследованные сканограммы содержали 8 или 9 строк, так что каждая сканограмма состояла из 64 или 81 числа. У каждого пациента было обследовано 1020 клеток шиповатого слоя буккального эпителия.

Первый показатель, названный *отношением объемов модальных классов*, основывается на объединении всех сканограмм пациента в неупорядоченный набор случайных величин из некоторой генеральной совокупности, выделения в нем трех модальных классов, содержащих числа из заданных диапазонов $M_1 = \{s_{ij} : 0 \leq s_{ij} \leq 0,15\}$, $M_2 = \{s_{ij} : 0,15 \leq s_{ij} \leq 0,30\}$, $M_3 = \{s_{ij} : s_{ij} > 0,30\}$ и вычислении отношения объемов классов M_1 и M_2 в k -й сканограмме:

$$V_k = \frac{\text{card } M_1^{(k)}}{\text{card } M_2^{(k)}},$$

где $\text{card } M_j^{(k)}$, $j=1,2$ — количество элементов модального класса $M_j^{(k)}$ (например, $\text{card } M_2^{(k)}$ — количество точек в k -й сканограмме пациента, в которых оптическая плотность ДНК лежит в пределах от 0,15 до 0,30). *Отношение объемов модальных классов*, характерное для исследуемого пациента, вычисляется как среднее по всем сканограммам:

$$V = \frac{1}{N} \sum_{k=1}^N V_k.$$

Этот показатель является чисто статистическим, поскольку содержит информацию о распределении оптической плотности ДНК в интерфазных ядрах эпителиоцитов СОПР.

Второй показатель, названный *показателем рельефности*, напротив, основан на геометрической интерпретации сканограмм пациента. При этом каждая сканограмма пациента рассматривается как график функции двух переменных $s_{ij} = s(i, j)$, где (i, j) — координаты точек сканограммы. Пример поверхности, образованной таким графиком, представлен на рис. 3.1. Для характеристики кри-

визны поверхности, соответствующей k -й сканограмме, вычисляется средний угловой коэффициент ее сечений по координатам i и j :

$$R_k = \frac{1}{n^2} \left(\sum_{i=1}^n \sum_{j=1}^{n-1} |s_{ij+1} - s_{ij}| + \sum_{i=1}^{n-1} \sum_{j=1}^n |s_{i+1j} - s_{ij}| \right).$$

Показатель рельефности, характерный для исследуемого пациента, определяется как среднее значение по всем сканограммам

$$R = \frac{1}{N} \sum_{k=1}^N R_k,$$

где N — количество сканограмм.

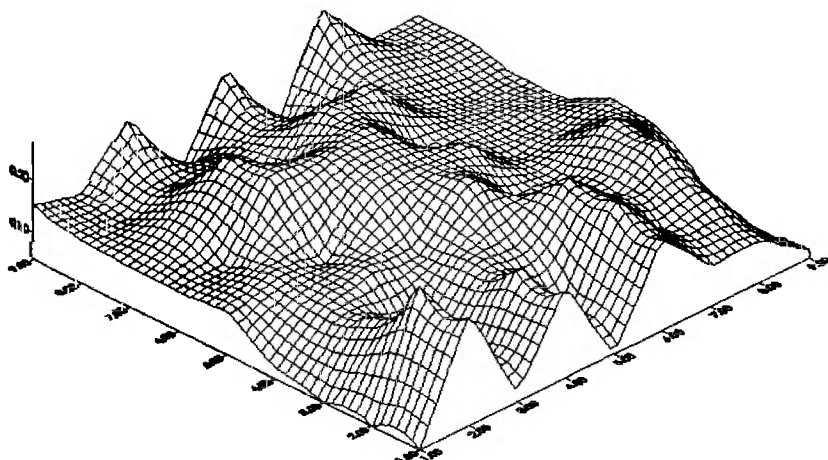


Рис. 3.1. Пример поверхности, образованной значениями оптической плотности ДНК в сканограмме

3.3.3. Результаты и их обсуждение

В предыдущем подразделе проведен анализ данных о содержании ДНК и 15 показателях, характеризующих морфометрические, денситометрические и текстурные параметры хроматина интерфазных ядер СОПР при различных патологических процессах в ШЖ. В данном подразделе поставлена задача усовершенствовать результаты диагностики с помощью двух новых описанных выше текстурных показателей.

Отношение объемов модальных классов у пациентов с УЗ и АТ изменяется в пределах от 0,220 до 2,053, а у пациентов с АДК оно изменяется в пределах от 0,148 до 1,098. Однако у основной части этих пациентов с АДК (13 из 15) этот показатель заключен между 0,148 и 0,439, и лишь два пациента имеют V , равный 1,075 и 1,098. При анализе цитологических мазков, окрашенных по Папаниколау, взятых у этих двух пациентов, было установлено, что клетки, попавшие в мазок, были не из шиповатого слоя эпителия СОПР, а из более

поверхностных слоев, где отсутствует синтез ДНК, что мы и подтвердили в наших исследованиях, проведенных раньше [48]. Этот факт еще раз свидетельствует о том, что наиболее информативным слоем СОПР при исследовании содержания ДНК является именно шиповатый слой.

Таким образом, если для выборки $V_1, V_2, \dots, V_{15}$ отношений объемов модальных классов у пациентов с АДК построить вариационный ряд

$$V_{(1)} \leq V_{(2)} \leq \dots \leq V_{(15)},$$

то $V_{(1)} = 0,148, \dots, V_{(13)} = 0,439, V_{(14)} = 1,075, V_{(15)} = 1,098$. Следовательно, основную распределенную массу показателя V в случае АДК можно представить в виде объединения двух доверительных интервалов

$$I = I_1 \cup I_2 = (V_{(1)}, V_{(13)}) \cup (V_{(14)}, V_{(15)}) = (0,148; 0,439) \cup (1,075; 1,098).$$

У исследованных пациентов с АДК щитовидной железы показатель рельефности изменялся в пределах от 0,0441 до 0,0960, а у пациентов с АТ и УЗ — от 0,0514 до 0,1267. Таким образом, доверительные интервалы этого показателя при различных патологиях имеют значительное пересечение.

Вычисленные доверительные интервалы для показателей R и V позволяют сформулировать следующий критерий для дифференциальной диагностики АДК, УЗ и АТ: если показатель рельефности R сканограмм пациента не превышает 0,0960, а его отношение объемов модальных классов принадлежит хотя бы одному из доверительных интервалов $I_1 = (0,148; 0,439)$ или $I_2 = (1,075; 1,098)$, то ставится диагноз АДК, в противном случае — УЗ или АТ (рис. 3.2). Чувствительность предложенного метода диагностики составляет 82, а специфичность 85,8%.

Заметим также, что при построении более осторожного метода диагностики доверительным интервалом $I_2 = (1,075; 1,098)$ следует пренебречь, что несколько повысит вероятность ложнонегативных диагнозов. В этом случае чувствительность составит 76,2%, а специфичность останется прежней — 85,8%.

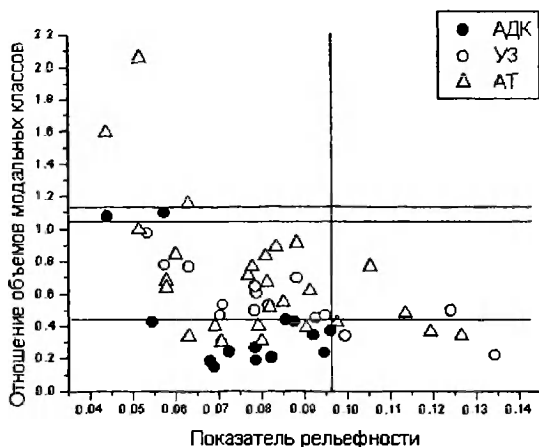


Рис. 3.2. Доверительные области показателей рельефности и отношения объемов модальных классов

3.3.4. Оценка чувствительности и специфичности метода

Вычислим вероятность ошибок 1-го и 2-го рода предложенных методов диагностики. Пусть основная гипотеза H состоит в том, что пациент болен аденокарциномой щитовидной железы, а альтернативная конкурирующая гипотеза H' означает, что у пациента узловой зоб или аутоиммунный тиреоидит. Как известно [42], вероятность ошибки 1-го рода определяется как вероятность отвергнуть гипотезу H , когда она истинна. В нашем случае она равна $p(H'|H)$. Для оценки вероятности $p(H'|H)$ воспользуемся следующей теоремой [84]: если $x_1, x_2, \dots, x_n$ — выборка, полученная с помощью простого случайного выбора, x_{n+1} — следующее выборочное значение, не зависящее от выборки, а $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ — вариационный ряд, построенный по выборке $x_1, x_2, \dots, x_n$, то

$$p(x \in (x_{(i)}, x_{(k)})) = \frac{k-i}{n+1}, \quad (i < k),$$

где $x_{(i)}, x_{(k)}$ порядковые статистики ($i < k$), $i, k = 0, 1, \dots, n, n+1$; $x_{(0)} = -\infty$; $x_{(n+1)} = +\infty$ (см. раздел 2.9).

При оценке чувствительности и специфичности методов мы будем предполагать, что показатели R и V являются независимыми случайными величинами. Данное предположение вполне допустимо, поскольку коэффициент корреляции между R и V равен 0,493, что свидетельствует об отсутствии сильной зависимости между этими показателями (хотя и не дает гарантий независимости этих показателей). В связи с этим определенная ниже вероятность ошибок является лишь приближением их точной теоретической вероятности.

Итак,

$$\begin{aligned} p(H'|P) &= p(R \leq 0,0960, 0 \leq V \leq 0,439 \vee 1,075 \leq V \leq 1,098) = \\ &= p(R \leq 0,0960)(p(0 \leq V \leq 0,439) + p(1,075 \leq V \leq 1,098)). \end{aligned}$$

Поскольку количество больных АДК $n = 15$, то

$$\begin{aligned} p(R \leq 0,0960) &= \frac{15}{16} = 0,9375; \\ p(0 \leq V \leq 0,439) &= p(V_{(0)} \leq V \leq V_{(13)}) = \frac{13}{16} = 0,8125; \\ p(1,075 \leq V \leq 1,098) &= p(V_{(14)} \leq V \leq V_{(15)}) = \frac{1}{16} = 0,0625. \end{aligned}$$

Следовательно, чувствительность предложенного критерия

$$p(H|H) \approx 0,82.$$

Соответственно, вероятность ошибки 1-го рода составляет 0,18 (или 18%).

Аналогичными рассуждениями можно доказать, что чувствительность “осторожного” критерия равна 76,2%, а вероятность ошибки 1-го рода — 0,238 (или 23,8%).

Вероятность ошибки 2-го рода определяется как вероятность принять гипотезу H , когда верна альтернативная конкурирующая гипотеза H' . Эту вероятность можно приближенно оценить как частоту случайного события, состоящего в том, что пациенты с УЗ и АТ диагностируются как АДК. В нашем исследовании количество пациентов с УЗ и АТ составляло 42, а ошибочно диагностировалось как АДК 6 пациентов с АТ. Заметим, кстати, что все пациенты с УЗ диагностировались верно. Таким образом,

$$p(H|H') = \frac{6}{42} = 0,1429.$$

Специфичность критерия составляет 85,8%. Отметим, что, согласно практическим рекомендациям [79], частота события практически совпадает с его вероятностью, когда количество испытаний превосходит 30. Следовательно, в нашем случае вероятность ошибки 2-го рода подсчитана практически точно для обоих предложенных критериев.

Резюме

Разработанные разработанные нами новые критерии в комплексе с имеющимися показателями позволяют рекомендовать их для дифференциальной диагностики между раком и доброкачественными пролиферативными процессами в ЩЖ.

3.4. Оценка степени прогностической значимости цитогенетических, морфологических и клинических показателей у больных меланомой кожи

3.4.1. Введение

За последние годы в большинстве стран мира отмечено значительное увеличение количества больных злокачественной меланомой кожи [2, 4, 20]. Наряду с этим имеются данные об улучшении ранней дифференциальной диагностики этого заболевания за счет внедрения новых неинвазивных методов исследования [23, 27]. Однако по-прежнему существуют определенные трудности при прогнозировании клинического течения опухолевого процесса и проспективном установлении продолжительности жизни этой категории больных, поскольку до настоящего времени не разработаны унифицированные подходы к систематизации, группировке и оценке факторов прогноза, хотя в литературе уже имеются отдельные сообщения о первых шагах к упорядочению вопросов прогностического моделирования. Мнения исследователей о прогностической важности разных показателей противоречивы. Одни из них [1, 7, 25] отдают предпочтение таким клиническим показателям, как пол, возраст больного,

клиническая стадия заболевания, локализация новообразования, наличие метастазов, не исключая при этом роли морфологических характеристик злокачественности новообразования (показателей инвазивности, митотической активности, некроза по поверхности и др.). Другие исследователи в качестве основных прогностических факторов признают глубину инвазии опухолевых клеток в дерму по Кларку и толщину меланомы по Бреслоу [9, 16]. Нет четко сформировавшегося представления о роли цитогенетических показателей среди признанных факторов прогноза, а также определенных суждений о степени значимости и месте каждого показателя в ряду используемых для прогноза признаков.

Аналогичная неоднозначность относительно информативности и значимости критериев для прогнозирования клинического течения опухолевого процесса имеет место при онкопатологии другого генеза [38]. Поэтому естественным является стремление онкологов к поиску математических подходов и моделей для объективизации представления об особенностях проявления опухолевого процесса на ранних этапах его развития, выборе оптимального объема индивидуализированного лечения онкологического больного и о прогнозе заболевания [69, 70, 120, 124].

Исходя из сказанного, целью настоящего исследования являлось определение степени значимости для прогноза ряда различных по биологической сущности цитогенетических, морфологических и клинических показателей у больных злокачественной меланомой с использованием модели линейной регрессии.

3.4.2. Материал и методы исследования

Исследование носит ретроспективный характер и основано на морфолого-цитогенетических данных, полученных в 1986–1988 годах, и показателях выживаемости 46 больных злокачественной меланомой кожи в 1993–1994 годах в Институте онкологии АМН Украины и Институте экспериментальной патологии, онкологии и радиобиологии им. Р.Е. Кавецкого НАН Украины. Основную группу составили 36 больных (20 женщин и 16 мужчин) в возрасте от 16 до 72 лет, умерших в сроки от 3 до 62 месяцев после операции. Десять больных в возрасте 21–75 лет живы (9 женщин и 1 мужчина). Продолжительность жизни в этой группе составляет 40–75 месяцев со дня удаления первичного новообразования.

Для оценки прогностической значимости нами использованы следующие клинические, цитогенетические и морфологические показатели лиц, умерших от злокачественной меланомы, в последующем обозначенные как $x_1, \dots, x_{16}$: x_1 — пол больного, x_2 — локализация меланомы, x_3 — возраст больного, x_4 — гистологическая форма роста меланомы, x_5 — наличие метастазов в регионарных лимфатических узлах, x_6 — уровень инвазии опухолевых клеток в дерму, x_7 — толщина опухоли, x_8 — митотический индекс опухоли, x_9 — количество патологических митозов, x_{10} — индекс метки, x_{11} — содержание ДНК в ядрах меланомных клеток, x_{12} — степень пигментации опухоли, x_{13} — степень лимфоцитарной инфильтрации в первичном очаге, x_{14} — некроз по поверхности новообразования, x_{15} — количество типов опухолевых клеток, x_{16} — типы клеток или их сочетания. Показатель выживаемости больных обозначен как y .

Для достижения поставленной цели использована модель линейной регрессии с применением классического метода наименьших квадратов (МНК), предложенного Лежандром и Гауссом [112]:

$$y = c_1 x_1 + c_2 x_2 + \dots + c_n x_n + c_0,$$

где $c_0, c_1, c_2, \dots, c_n$ — МНК-оценки коэффициентов линейной регрессии, $x_1, x_2, \dots, x_n$ — разнородные показатели и факторы, которые являются независимыми переменными, а y — зависимая переменная, значения которой определяются по $x_1, x_2, \dots, x_n$.

Точность модели и близость прогнозируемых значений зависимых переменных $\hat{y}^{(k)}$ к истинным $y^{(k)}$ оценены на основании определения ошибки модели:

$$s^2 = \sum_{k=1}^N \left(y^{(k)} - \sum_{i=1}^n c_i x_i^{(k)} - c_0 \right)^2,$$

где $N = 36$ (количество умерших пациентов), $n = 16$ (количество показателей).

Значимость каждого из анализируемых факторов оценивали путем исключения его из модели линейной регрессии и рассмотрением “усеченной” таким образом модели. Целесообразность такого подхода очевидна, поскольку потеря фактора приводит к возрастанию ошибки модели, и величина этой ошибки находится в прямой зависимости от важности фактора в модели, а отсутствие наиболее важного фактора влечет за собой самое значительное увеличение ошибки.

Построение вариационного ряда, составленного из ошибок модели в порядке убывания, позволяет составить представление о степени важности каждого из оцениваемых показателей.

1. Рассмотрим классическую модель многомерной линейной регрессии

$$y = c_1 x_1 + c_2 x_2 + \dots + c_n x_n + c_0 + \xi, \quad (3.1)$$

где $x_1, x_2, \dots, x_n$ — неслучайные величины, которые обычно интерпретируются как показатели или факторы; y — известный (или прогнозируемый) показатель; $c_1, c_2, \dots, c_n, c_0$ — неизвестные коэффициенты линейной регрессии; ξ — случайный шум (артефакт, искажающий истинное значение y).

Основная проблема, которую мы будем изучать в этом подразделе, заключается в следующем: определить важность (значимость) каждого из факторов $x_1, x_2, \dots, x_n$, участвующих в формировании значений известного показателя y , и установить наиболее важные факторы в модели (3.1). Предполагается, что при этом известны экспериментальные значения факторов $x_1^{(k)}, x_2^{(k)}, \dots, x_n^{(k)}$ и соответствующее им значение $y^{(k)}$, полученные после k -го испытания (эксперимента) ($k = 1, 2, \dots, N$); считается, что количество повторений эксперимента N значительно превышает количество факторов n .

Для решения этой проблемы мы будем использовать метод наименьших квадратов (МНК). Пусть $\hat{c}_1, \dots, \hat{c}_n, \hat{c}_0$ — оценки, полученные методом наименьших квадратов (МНК-оценки) [112] на основании экспериментальных данных $y^{(k)}, x_1^{(k)}, \dots, x_n^{(k)}$ ($k = \overline{1, N}$). Обозначим через s^2 ошибку модели (3.1)

$$s^2 = \sum_{k=1}^N \left\{ y^{(k)} - \sum_{i=1}^n c_i^* x_i^{(k)} - c_0^* \right\}^2,$$

которая характеризует точность этой модели. Для определения важности фактора x_j мы исключим его из модели (3.1) и рассмотрим “усеченную модель”

$$y = c_1^{(j)} x_1 + \dots + c_{j-1}^{(j)} x_{j-1} + c_{j+1}^{(j)} x_{j+1} + \dots + c_n^{(j)} x_n + c_0^{(j)} + \xi, \quad (3.2)$$

построенную по всем остальным факторам $x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_n$. При этом ошибка s_j^2 модели имеет вид

$$s_j^2 = \sum_{k=1}^N \left\{ y^{(k)} - \sum_{i=1, i \neq j}^n c_i^{(j)*} x_i^{(k)} - c_0^{(j)*} \right\}^2,$$

где $c_i^{(j)*}$, ($i \neq j$) — МНК-оценки в модели (3.2), вычисленные на основании экспериментальных значений $y^{(k)}, x_1^{(k)}, \dots, x_{j-1}^{(k)}, x_{j+1}^{(k)}, \dots, x_n^{(k)}$ ($k = \overline{1, N}$). Легко видеть, что эта ошибка s_j^2 возрастает: $s_j^2 \geq s^2$, причем для самого важного фактора x_j это возрастание будет наибольшим: $s_j^2 \geq s^2 \forall i = 1, 2, \dots, n$. Обоснование этого эффекта очевидно: если фактор x_j является наиболее важным, то его потеря (т.е. исключение из модели (3.1)) приведет к самому значительному увеличению ошибки модели. Таким образом, на основании ошибки s_j^2 , полученной за счет исключения фактора x_j из модели (3.1), мы можем судить о важности (значимости) этого фактора для показателя y . Если из ошибок s_j^2 ($j = 1, 2, \dots, n$) образовать вариационный ряд

$$s_{j_1}^2 \leq s_{j_2}^2 \leq \dots \leq s_{j_n}^2,$$

то наиболее важным фактором является показатель x_{j_n} , затем $x_{j_{n-1}}$ и т.д., а самым мало важным фактором будет показатель x_{j_1} с наименьшей ошибкой $s_{j_1}^2$.

2. Схема линейной регрессии (3.1) необходима для вычислений (прогнозирования) значений зависимой переменной на основании значений независимых переменных (случайных величин) $x_1, x_2, \dots, x_n$. Во многих практически важных задачах значения $x_1, x_2, \dots, x_n$ можно получить в ходе эксперимента (испытания), тогда как непосредственное вычисление значения y является либо трудновыполнимым, либо вообще невозможным. В связи с этим мы будем называть результаты эксперимента *полностью определенными*, если мы можем получить при этом значения как независимых переменных $x_1, x_2, \dots, x_n$, так и функции y ; если же в ходе эксперимента можно лишь определить значения $x_1, x_2, \dots, x_n$, а значения y получить нельзя, то такой эксперимент называется *частично определенным*. Используя полностью определенные результаты эксперимента, мы можем получить экспериментальные данные $y^{(k)}, x_1^{(k)}, \dots, x_n^{(k)}$ ($k = \overline{1, N}$) и с помощью этих данных и метода наименьших квадратов вычислить МНК-оценки коэффициентов линейной регрессии $c_1^*, \dots, c_n^*, c_0^*$. Зная значения этих коэффициентов, можно прогнозировать значение зависи-

мой переменной y на основании частично определенного результата эксперимента $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n$:

$$\bar{y} = c_1^* \tilde{x}_1 + c_2^* \tilde{x}_2 + \dots + c_n^* \tilde{x}_n + c_0^*,$$

где $\bar{y}$ — прогнозируемое значение для y .

Естественно, что прогнозируемое значение $\bar{y}$ не совпадает с истинным значением y , которое получено при полностью определенном результате эксперимента, даже если предположить, что модель линейной регрессии (3.1) абсолютно точно описывает реальную ситуацию, поскольку МНК-оценки c_k^* ($k = 0, 1, \dots, n$) отличаются от истинных значений c_k ($k = 0, 1, \dots, n$). В связи с этим возникает проблема отыскания доверительных границ (a, b) для истинного значения y на основании его прогнозируемого значения $\bar{y}$, соответствующих заданному уровню значимости 2β . Решение этой задачи в настоящее время получено в очень жестких предположениях (независимость случайных величин $x_1, x_2, \dots, x_n$, гипотеза об их нормальном распределении и т.д.) [71], которые никогда не выполняются на практике. Мы опишем значительно более простое решение задачи построения доверительных границ (a, b) в более общих предположениях, которые выполняются во многих практически важных случаях. Предположим, что $(n+1)$ -мерный случайный вектор $(y, x_1, x_2, \dots, x_n)$, представляющий результат полностью определенного эксперимента, получен при реализации простого случайного выбора, т.е. случайные векторы $(y^{(1)}, x_1^{(1)}, x_2^{(1)}, \dots, x_n^{(1)})$, ..., $(y^{(N)}, x_1^{(N)}, x_2^{(N)}, \dots, x_n^{(N)})$ независимы в совокупности и взяты из одной генеральной совокупности (и имеют поэтому одинаковое распределение).

Рассмотрим относительную ошибку δ прогнозируемого результата $\bar{y}$: $\delta = \frac{|y - \bar{y}|}{|\bar{y}|}$, где y — истинное значение зависимой переменной, соответствующее наблюдениям $x_1, x_2, \dots, x_n$, а $\bar{y} = c_1^* \tilde{x}_1 + c_2^* \tilde{x}_2 + \dots + c_n^* \tilde{x}_n + c_0^*$ — прогнозируемое значение.

Легко видеть, что вычислить относительную ошибку δ можно лишь при условии, что мы наблюдаем полностью определенный результат эксперимента. Поскольку у нас есть N таких результатов, мы можем вычислить N значений $\delta_1, \delta_2, \dots, \delta_N$, где $\delta_k = \frac{|y^{(k)} - \bar{y}^{(k)}|}{|\bar{y}^{(k)}|}$, ($k = 1, 2, \dots, N$).

В силу наших предположений при больших N можно считать, что эти значения δ_k представляют собой выборку, полученную после простого случайного выбора. Действительно, по условию, случайные векторы $(y^{(1)}, x_1^{(1)}, x_2^{(1)}, \dots, x_n^{(1)})$, ..., $(y^{(N)}, x_1^{(N)}, x_2^{(N)}, \dots, x_n^{(N)})$ независимы в совокупности и имеют одинаковые распределения; при больших значениях N МНК-оценки $c_1^*, \dots, c_n^*, c_0^*$ практически

совпадают с истинными значениями коэффициентов линейной регрессии $c_1, c_2, \dots, c_n, c_0 : \hat{c}_k \approx c_k, k = 0, 1, 2, \dots, n$, поэтому случайные величины $\tilde{y}^{(1)}, \tilde{y}^{(2)}, \dots, \tilde{y}^{(N)}$ также будут независимы в совокупности и имеют одинаковое распределение. Как известно [83], значение борелевской функции от независимых и одинаково распределенных случайных величин образует случайные величины, которые будут также независимы и имеют одинаковое распределение, следовательно, относительные ошибки $\delta_1, \delta_2, \dots, \delta_N$ можно рассматривать как случайную выборку, полученную за счет простого случайного выбора.

Пусть теперь $\delta^{\max} = \max(\delta_1, \delta_2, \dots, \delta_N)$. Как известно [98], для независимых и одинаково распределенных случайных величин справедливо утверждение: для произвольной случайной величины δ , которая имеет такое же распределение, как и δ_k , и является независимой от $\delta_1, \delta_2, \dots, \delta_N$, $p(\delta > \delta_k) = \frac{1}{2(N+1)}$. Следовательно, доверительный интервал

$$(a, b) = (\tilde{y} - |\tilde{y}| \delta^{\max}, \tilde{y} + |\tilde{y}| \delta^{\max}) \quad (3.3)$$

имеет уровень значимости $2\beta = \frac{1}{2(N+1)}$. На практике мы можем гарантировать, что при этом уровень значимости 2β приблизительно равен $\frac{1}{2(N+1)} : 2\beta \approx \frac{1}{2(N+1)}$, поскольку МНК-оценки $\hat{c}_k$ не равны в точности c_k ($k = 0, 1, 2, \dots, n$). Если из выборки $\delta_1, \delta_2, \dots, \delta_N$ образовать вариационный ряд

$$\delta_{(1)} \leq \delta_{(2)} \leq \dots \leq \delta_{(N-1)} \leq \delta_N \quad (3.4)$$

и провести его цензурирование, отбрасывая значение δ_N и принимая за $\delta^{\max}$ непосредственно предшествующее значение $\delta_{(N-1)} : \delta^{\max} = \delta_{(N-1)}$, то доверительный интервал

$$(\tilde{y} - |\tilde{y}| \delta_{(N-1)}, \tilde{y} + |\tilde{y}| \delta_{(N-1)}) \quad (3.5)$$

имеет уровень значимости $\frac{3}{2(N+1)}$. Например, при $N = 36$

$$2\beta = \frac{1}{2(N+1)} \approx 0,0135135 < 0,05, \quad 2\beta = \frac{3}{2(N+1)} \approx 0,0403405 < 0,05.$$

Таким образом, уровень значимости обоих доверительных интервалов (3.3) и (3.5) не превышает 5%-ный уровень, который принят за допустимый уровень в биологических и медицинских науках, и при $N = 36$ их можно использовать при проведении практических статистических вычислений.

Принимая за основу описанный выше математический принцип определения степени важности разных по биологической сущности показателей в прогнозе, мы использовали различные подходы к ретроспективной оценке морфолого-цитогенетических и клинических данных.

Прежде всего, рассмотрен весь ряд анализируемых показателей (в количестве 16) без распределения на группы с поочередным исключением одного из них

из модели, затем рассмотрены три группы показателей: цитогенетические, морфологические, клинические — с игнорированием одной из них в целом, в последнюю очередь — ряд показателей в каждой из трех групп с соблюдением очередности исключения одного из них. В итоге в этом разделе исследования было проведено три вида оценки степени прогностической значимости показателей: индивидуальная во всем вариационном ряду, групповая и индивидуальная в каждой группе.

В третьем подразделе сделана попытка проспективного прогнозирования продолжительности жизни больных злокачественной меланомой из группы “живых” с применением того же математического метода исследования — модели линейной регрессии.

Используя схему линейной регрессии для группы “умерших”, в которой все показатели, включая выживаемость (истинную и ожидаемую), известны или могут быть определены, и вычислив для нее МНК-оценки коэффициентов линейной регрессии ($c_0 = 70,076$, $c_1 = 5,999$, $c_2 = -3,236$, $c_3 = -0,006$, $c_4 = -1,325$, $c_5 = -19,788$, $c_6 = 4,957$, $c_7 = -1,725$, $c_8 = -1,456$, $c_9 = -0,252$, $c_{10} = -0,779$, $c_{11} = -1,492$, $c_{12} = 2,094$, $c_{13} = 4,400$, $c_{14} = -0,816$, $c_{15} = -2,689$, $c_{16} = 0,532$), мы рассчитали ожидаемые значения продолжительности жизни для каждого из 10 больных злокачественной меланомой по схеме линейной регрессии из группы “живых”. Для истинных значений продолжительности жизни больных на основании прогнозируемых значений вычислены доверительные интервалы, соответствующие 5%-ному уровню значимости.

3.4.3. Результаты собственных исследований и их обсуждение

При построении вариационного ряда из всех ошибок модели линейной регрессии в группе умерших от злокачественной меланомы для установления степени прогностической значимости каждого из использованных в исследовании клинических, цитогенетических и морфологических параметров (табл. 3.11) оказалось, что наиболее важным и неблагоприятным фактором является наличие метастазов в регионарных лимфатических узлах. Индивидуализированный анализ наблюдений показал, что у 22 больных этой группы из 36 прооперированных были гистологически подтвержденные метастазы в регионарных лимфатических узлах. Второе по степени значимости место занимает локализация опухоли. Действительно, преобладающими у лиц обоего пола являлись опухоли туловища, нижних конечностей, стопы и кисти, которые по анатомо-морфологическим характеристикам отличаются неблагоприятным течением и прогнозом. Как правило, это были нодулярные, акральные лентигинозные формы и меланомы, развившиеся на фоне невуса, которые, согласно данным литературы, имеют тяжелое клиническое течение опухолевого процесса ввиду их агрессивного инвазивного роста [16, 25, 46, 95, 108]. Третьим по степени прогностической значимости в ряду анализируемых показателей стоит пол больного. В наших исследованиях продолжительность жизни мужчин ока-

залась более низкой, чем женщин, что совпадает с данными работы [16], свидетельствующими о том, что выживаемость больных женского пола на 10% превышает этот показатель у мужчин. Меньшими числовыми значениями ошибки модели в порядке снижения прогностической значимости характеризуются митотический индекс, степень пигментации опухоли, уровень инвазии меланомных клеток в дерму. Все остальные показатели вариационного ряда мало отличались между собой, т.е. были равнозначными.

Таблица 3.11. Вариационный ряд исследованных показателей по степени их прогностической значимости у лиц, умерших от злокачественной меланомы

№ показателя	Исследованные показатели	Ошибка модели s^2
5	Метастазы в регионарные лимфатические узлы	3321,836
2	Локализация опухоли	2432,406
1	Пол больного	2320,438
8	Митотический индекс опухоли	2268,602
12	Степень пигментации опухоли	2242,315
6	Уровень инвазии меланомы в дерму	2216,926
4	Гистологическая форма роста опухоли	2209,708
13	Степень лимфоцитарной инфильтрации в первичном очаге	2186,212
7	Толщина опухоли	2184,832
15	Количество типов опухолевых клеток	2174,449
9	Количество патологических митозов	2170,380
16	Типы опухолевых клеток	2170,186
10	Индекс метки	2167,917
11	Содержание ДНК в ядрах меланомных клеток	2165,200
14	Некроз по поверхности опухоли	2161,007
3	Возраст больного	2160,007

При определении степени прогностической значимости целых групп признаков показано, что в данном исследовании по степени важности они распределяются следующим образом: клинические, цитогенетические, морфологические.

На основании внутригруппового анализа индивидуального вклада каждого из входящих в группу показателей (табл. 3.12) установлено, что в ряду клинических характеристик наиболее значимыми и неблагоприятными для прогноза являются наличие метастазов в регионарных лимфатических узлах, а также локализация меланомы, из цитогенетических параметров — митотический индекс опухоли и содержание ДНК в ядрах опухолевых клеток, среди морфологических — толщина опухоли, уровень ее инвазии в дерму и наличие некроза по поверхности новообразования.

При сопоставлении истинных значений продолжительности жизни лиц, умерших от злокачественной меланомы, с расчетно-ожидаемыми обнаружилось, что в большинстве случаев показатели совпадают или статистически мало отличаются между собой и находятся в границах вычисленных доверительных интервалов.

Результаты проспективного прогнозирования продолжительности жизни 10 больных злокачественной меланомой в группе “живых” показали (табл. 3.13), что у всех лиц истинные показатели выживаемости и ожидаемые не имеют статистически значимых отличий. Однако у 5-го и 6-го больного истинные сроки продолжительности жизни существенно превосходят соответствующие прогнозируемые значения, вычисленные на основании ретроспективного анализа показателей умерших пациентов. Следует отметить, что у этих лиц в отличие от других из группы “живых” имеются метастазы в регионарных лимфатических узлах, высокое содержание ДНК в ядрах меланомных клеток и некроз по поверхности новообразования. По-видимому, эти факторы могут в дальнейшем оказаться неблагоприятными для прогноза.

Таблица 3.12. Внутригрупповое распределение клинических, цитогенетических и морфологических показателей по степени прогностической значимости у лиц, умерших от злокачественной меланомы

Группы показателей	№ показателя	Исследованные показатели	Ошибка модели χ^2
Клинические	5	Метастазы в регионарные лимфатические узлы	11566,990
	2	Локализация опухоли	4512,733
	3	Возраст больного	4142,715
	1	Пол больного	4081,763
Цитогенетические	8	Митотический индекс опухоли	7719,246
	11	Содержание ДНК в ядрах меланомных клеток	7241,333
	10	Индекс метки опухоли	7194,529
	9	Количество патологических митозов в опухоли	7171,642
Морфологические	7	Толщина опухоли	6537,924
	6	Уровень инвазии меланомных клеток в дерму	5540,371
	14	Некроз по поверхности опухоли	5359,556
	12	Степень пигментации опухоли	5007,994
	13	Степень лимфоцитарной инфильтрации первичного очага	4851,468
	4	Гистологическая форма роста опухоли	4669,554
	15	Количество типов опухолевых клеток	4667,640
	16	Типы опухолевых клеток	4640,273

Таблица 3.13. Результаты проспективного прогнозирования продолжительности жизни больных злокачественной меланомой в группе “живых”

Номер пациента	Истинная продолжительность жизни γ (месяцев)	Прогнозируемая продолжительность жизни $\bar{\gamma}$ (месяцев)	Верхняя доверительная граница для γ (месяцев*)	Относительная ошибка прогноза δ
1	48	66	197	0,273
2	49	62	184	0,210
3	55	65	194	0,154
4	63	68	204	0,074
5	65	37	110	0,757
6	66	31	92	1,129
7	70	63	189	0,111
8	70	60	179	0,167
9	71	66	197	0,076
10	73	61	180	0,197

* Нижние доверительные границы для γ в таблице не приведены, поскольку их числовые значения отрицательны и их следует приравнять к нулю.

Резюме

Результаты ретроспективного анализа степени значимости для прогноза клинических, цитогенетических и морфологических показателей в группе умерших от злокачественной меланомы свидетельствуют о том, что все они имеют определенное влияние на продолжительность жизни больных, однако в каждой из оцениваемых групп на первый план выступают наиболее значимые факторы. В данном исследовании среди неблагоприятных клинических показателей первоочередное значение имеет наличие регионарных метастазов; в ряду цитогенетических параметров — митотический индекс опухоли и содержание ДНК в ядрах меланомных клеток; среди морфологических — толщина и уровень инвазии опухоли в дерму, а также некроз по поверхности новообразования. Не исключено, что именно совокупность отмеченных нами ведущих клинических, цитогенетических и морфологических показателей оказывает определяющее влияние на продолжительность жизни больных злокачественной меланомой.

Таким образом, использование математических подходов к прогностической оценке опухолевого процесса, с одной стороны, позволило ретроспективно определить степень значимости для прогноза каждого из наиболее важных цитогенетических, морфологических и клинических показателей в группе умерших от злокачественной меланомы кожи, с другой стороны — проспективно рассчитать прогнозируемую продолжительность жизни лиц из группы “живых” для сопоставления этих данных с реальными сроками их жизни.

3.5. Применение статистических методов в радиобиологии: оценка величины поглощенной дозы и степени острой лучевой болезни на основании хромосомных аберраций

Ряд оптимистических высказываний в постчернобыльском периоде о том, что радиационно-индуцированный генетический груз незначителен, оказались преждевременными. Цитогенетические последствия глобальной (трансграничной) аварии выявляются и в настоящее время и служат в качестве корректной основы для выполнения биологической индикации и/или дозиметрии эффекта облучения, для прогноза отдаленных постлучевых изменений на организменном уровне. Общеизвестно, что при отсутствии надежных данных физической дозиметрии используется анализ нестабильных хромосомных аберраций в лимфоцитах крови, позволяющий успешно реконструировать поглощенные дозы [34]. Существует ряд критериев, обуславливающих использование с этой целью лимфоцитов периферической крови как модель малых дозиметров со случайным распределением по всему телу облученного. Кратко эти критерии сводятся к следующему: высокая радиочувствительность лимфоцитов периферической крови; низкий уровень спонтанных аберраций хромосом; естественная синхронизация популяции лимфоцитов; равномерное распределение и циркуляция по всему организму; унификация цитогенетических протоколов; специфичность эффекта, индуцированного радиацией; дозозависимое повреждение хромосом и, наконец, близкое количественное соответствие хромосомных аберраций при идентичном облучении лимфоцитов крови *in vitro* и *in vivo* [57]. По сути, два последних достоинства культуры лимфоцитов человека явились основанием для использования этой тест-системы в биодозиметрических исследованиях при неконтролируемых облучениях человека.

В ряде цитогенетических лабораторий разработаны эталонные кривые выхода аберраций хромосом при облучении *in vitro* цельной крови, *in vivo* после терапевтического тотального облучения [89, 107]. Методы анализа аберраций хромосом имеют важное значение для диагностики степени острой лучевой болезни (ОЛБ), особенно первой степени (1-2 Гр), в связи с нечеткостью перехода субклинических признаков реакции организма в клинически значимые [110].

3.5.1. Материал и методы исследования

Исследование носит ретроспективный характер и основано на цитогенетических данных, полученных в 1986 году через 12, 24, 33 и 40 дней после момента облучения в Институте онкологии АМН Украины. Основную группу составили 30 ликвидаторов (все мужчины), страдающих от острой лучевой болезни 1-3 степени (ОЛБ). Исходные данные цитогенетического обследования представлены в табл. 3.14, а усредненные показатели — в табл. 3.15. Распределение ликвидаторов по степени ОЛБ и соответствующим оценкам дозы приведено на рис. 3.3.

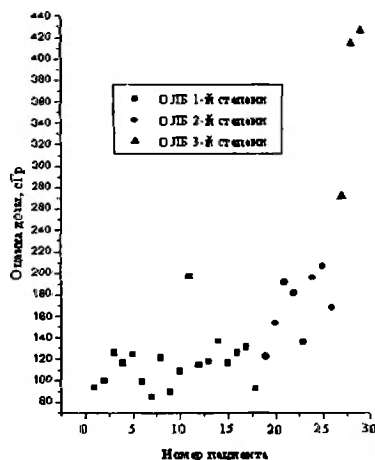


Рис. 3.3. Распределение ликвидаторов по степени ОЛБ и соответствующим оценкам дозы

Таблица 3.14. Данные цитогенетических обследований ликвидаторов с 1–3 степенью острой лучевой болезни

NN	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
1	2	12	200	4,0	4,5	0,0	0,0	0,0	0,0	0,0	4,5	0,0
2	1	12	200	2,5	2,5	1,0	0,5	0,0	0,0	0,0	1,0	0,0
3	1	12	50	4,0	4,0	2,0	0,0	0,0	2,0	0,0	0,0	0,0
4	1	12	200	6,0	6,0	1,5	0,5	0,0	1,0	0,0	3,0	0,0
5	1	12	200	12,5	13,5	6,5	2,5	0,0	3,0	0,0	1,5	0,0
7	2	12	300	6,3	7,0	0,3	0,6	0,0	0,3	0,3	5,2	0,3
8	2	12	250	15,2	16,0	2,4	4,4	0,8	7,2	0,4	0,4	0,4
11	1	24	145	4,0	4,0	0,0	0,6	0,0	1,3	0,0	2,1	0,0
12	3	24	300	21,6	25,0	6,6	3,0	0,5	10,0	1,0	3,9	0,0
13	2	24	100	9,0	9,0	2,0	1,0	0,0	2,0	0,0	0,0	0,0
14	1	24	200	4,0	4,0	0,0	0,0	1,0	0,0	0,0	1,0	2,0
15	1	24	100	2,0	2,0	1,0	1,0	0,0	0,0	0,0	0,0	0,0
16	3	24	114	51,7	80,5	22,8	14,0	1,7	30,7	5,2	6,1	0,0
17	1	24	200	11,0	13,0	1,0	1,0	2,0	9,0	0,0	0,0	0,0
18	1	24	200	4,0	4,0	0,0	2,0	1,0	1,0	0,0	0,0	0,0
19	1	24	200	6,0	6,0	3,0	0,0	0,0	2,0	0,0	1,0	0,0
24	1	24	200	2,0	2,0	1,0	0,0	0,0	0,0	0,0	1,0	0,0
40	1	40	200	9,5	9,5	1,0	0,5	0,0	0,0	0,5	6,5	1,0
41	2	40	200	5,0	5,0	0,0	0,0	0,0	0,0	0,0	5,0	0,0
42	1	40	200	3,0	3,0	0,0	0,0	0,0	0,5	0,0	2,5	0,0
43	1	40	60	3,3	3,3	0,0	0,0	0,0	1,7	0,0	0,0	1,6

NN	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
44	2	40	200	8,5	8,5	2,0	1,0	1,0	2,0	1,5	1,0	0,0
45	1	40	200	5,0	5,0	0,0	0,0	0,0	1,0	0,0	3,5	0,5
46	1	40	200	3,5	4,0	0,0	0,0	0,0	0,0	0,0	3,5	0,5
47	1	40	200	4,0	4,0	0,0	0,0	0,0	2,0	0,0	0,0	2,0
48	1	40	200	5,0	5,0	0,0	1,0	0,0	0,0	0,0	3,0	1,0
49	2	40	200	16,0	20,0	3,0	3,0	0,0	8,0	0,5	5,5	0,0
50	2	40	200	5,5	5,5	0,0	0,0	0,0	1,5	0,5	3,0	0,5
52	3	40	200	20,5	23,0	3,5	2,5	0,0	12,0	1,5	3,0	0,5
53	1	40	200	1,0	1,0	0,0	0,0	0,0	0,0	0,0	1,0	0,0

Таблица 3.15. Средние значения показателей цитогенетического обследования ликвидаторов

Показатели	Заключительный клинический диагноз (степень ОЛБ)		
	I	II	III
Количество обследованных больных	18	8	3
Количество проанализированных клеток	176,58±11,43	206,25±19,90	204,67±53,74
Процент клеток с абберациями	4,86±0,70	8,69±1,62	31,27±10,22
Суммарная частота аббераций	5,04±0,79	9,44±1,99	42,83±18,84
Процент аббераций хромосомного типа			
Всего:	2,98±0,83	5,71±2,19	38,33±18,04
фрагменты	0,95±0,36	1,21±0,44	10,97±5,98
ацентрические кольца	0,51±0,17	1,25±0,57	6,50±3,75
центрические кольца	0,21±0,12	0,23±0,15	0,73±0,50
дицентрики	1,29±0,48	2,63±1,13	17,57±6,59
транслокации	0,03±0,03	0,40±0,18	2,57±1,32
Процент аббераций хроматидного типа			
Всего:	2,06±0,44	3,23±0,81	4,50±0,81
делеции	1,61±0,40	3,08±0,81	4,33±0,92
обмены	0,45±0,16	0,15±0,08	0,17±0,17

При ретроспективной оценке полученной дозы нами использованы следующие клинические и цитогенетические показатели: x_1 — количество дней после момента облучения, x_2 — количество проанализированных клеток, x_3 — количество поврежденных клеток, x_4 — общее количество аббераций хромосом, x_5 — количество фрагментов хромосом, x_6 — количество точек, x_7 — количество колец, x_8 — количество дицентриков, x_9 — количество транслокаций,

x_{10} — количество фрагментов хроматидного типа, x_{11} — количество обменов хроматидного типа. Значения показателей x_3 – x_{11} вычисляются из расчета на каждые 100 клеток. Степень острой лучевой болезни обозначена как y .

Для достижения поставленной цели использована модель линейной регрессии с применением классического метода наименьших квадратов (МНК):

$$y = c_0x_1 + c_2x_2 + \dots + c_nx_n + c_0,$$

где $c_0, c_1, c_2, \dots, c_n$ — МНК-оценки коэффициентов линейной регрессии, $x_1, x_2, \dots, x_n$ — поясняющие переменные, а y — зависимая переменная, значения которой определяются по $x_1, x_2, \dots, x_n$.

Точность модели и близость прогнозируемых значений $\hat{y}^{(k)}$ к истинным $y^{(k)}$ оценены на основании определения ошибки модели:

$$s^2 = \sum_{k=1}^N \left(y^{(k)} - \sum_{i=1}^n c_i x_i^{(k)} - c_0 \right)^2,$$

где $N = 30$ (количество обследованных лиц), $n = 13$ (количество показателей).

Значимость каждой из анализируемых поясняющих переменных для определения степени ОЛБ мы оценивали путем исключения ее из модели линейной регрессии и применения “усеченной” таким образом модели. Поскольку потеря поясняющей переменной приводит к возрастанию ошибки модели и величина этой ошибки находится в прямой зависимости от важности (информативности) пропущенной поясняющей переменной в модели, отсутствие наиболее важной поясняющей переменной повлечет за собой наибольшую ошибку.

Построение вариационного ряда, составленного из ошибок модели в порядке убывания, позволяет составить представление о степени важности каждой из оцениваемых поясняющих переменных.

3.5.2. Результаты и обсуждение

В табл. 3.16 приведены результаты оценки информативности изученных нами цитогенетических показателей по методу, изложенному выше. При этом мы рассматривали лишь те показатели (поясняющие переменные), которые могут служить лучевыми маркерами, поэтому из модели линейной регрессии были исключены показатели x_1 и x_2 . Анализ полученных результатов показывает, что для ретроспективной оценки степени острой лучевой болезни и полученной дозы облучения, множественная регрессия, учитывающая комплекс цитогенетических показателей, предпочтительнее одномерной регрессии по отдельным поясняющим переменным. При этом выбор поясняющих переменных для множественной регрессии необходимо делать после детального анализа точности получаемых результатов для обучающей выборки.

В табл. 3.17 приведены результаты оценки точности множественной регрессии для обучающей выборки при различных комбинациях поясняющих переменных, которые включаются в рассмотрение по мере убывания их степени значимости. Как видим, наилучшая точность (в смысле минимального количества ошибок, но

не минимального среднеквадратичного отклонения) достигнута для набора показателей $x_3, x_5, x_7, x_9, x_{11}$. Заметим также, что приемлемая точность была получена уже при трех поясняющих переменных x_3, x_5, x_7 , и в дальнейшем, по мере добавления новых переменных, она улучшалась незначительно.

Таблица 3.16. Вариационный ряд исследованных цитогенетических показателей по степени их информативной значимости для ретроспективной оценки степени ОЛБ и коэффициентам корреляции

Номер показателя	Название показателя	Среднеквадратичное отклонение	Номер показателя	Название показателя	Коэффициент корреляции
3	Количество поврежденных клеток	4,34659	3	Количество поврежденных клеток	0,69216
9	Транслокации	3,71781	9	Транслокации	0,64728
5	Фрагменты хромосомного типа	3,52422	4	Общее количество аберраций	0,64087
11	Обмены	3,356637	8	Дицентрики	0,63964
7	Кольца	3,17102	6	Точки	0,55337
4	Общее количество аберраций	2,95759	5	Фрагменты хромосомного типа	0,52196
6	Точки	2,95435	10	Фрагменты хроматидного типа	0,44820
8	Дицентрики	2,89432	7	Кольца	0,22097
10	Фрагменты хроматидного типа	2,89398	11	Обмены	0,21422

Таблица 3.17. Результаты ретроспективной оценки степени ОЛБ с помощью одномерной и множественной линейной регрессии

Показатели	Кол-во ошибок	Процент ошибок	s^2	Показатели	Кол-во ошибок	Процент ошибок	s^2
x_3	12	40,000	6,56233	x_3, x_9	11	36,667	6,54992
x_4	11	36,667	7,54767	x_3, x_5, x_7	5	16,667	4,45414
x_5	10	33,333	9,54041	x_3, x_5, x_7, x_9	4	13,333	3,81872
x_8	10	33,333	9,05362	$x_3, x_5, x_7, x_9, x_{11}$	3	10,000	3,09612
x_7	12	40,000	12,76299	$x_3, x_4, x_5, x_7, x_8, x_{11}$	4	13,333	2,95482
x_6	9	30,000	7,57047	$x_3, x_4, x_5, x_6, x_7, x_8, x_{11}$	4	13,333	2,89499
x_9	9	30,000	7,42865	$x_3, x_4, x_5, x_6, x_8, x_7, x_9, x_{11}$	4	13,333	2,89398
x_{10}	12	40,000	10,57163	$x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}$	4	13,333	2,89012
x_{11}	18	60,000	12,80529				

Отметим, что до сих пор нами рассматривались лишь ретроспективные оценки степени острой лучевой болезни с помощью линейной регрессии, но не изучался вопрос об оценке полученной дозы облучения. Для этой цели мы предлагаем следующее: поскольку оценки степени острой лучевой болезни, полученные нами, выражаются дробными числами, то можно ввести пропорциональные коэффициенты для пересчета долей оценки степени ОЛБ в сантигрей. Очевидно, что для первой степени ОЛБ (диапазон 100–200 сГр) одна сотая оценки степени ОЛБ соответствует одному сантигрею, а для 2-й и 3-й степени ОЛБ (диапазоны 200–400 сГр и 400–600 сГр) одной сотой оценки степени ОЛБ выше 2-й соответствует 2 сГр.

В табл. 3.18 и 3.19 приведены результаты ретроспективной оценки полученной ликвидаторами индивидуальной дозы облучения по предложенному нами методу (на основе показателей $x_3, x_5, x_7, x_9, x_{11}$) и оценки этих доз по методикам, предложенным А.В. Севаньяевым и А.П. Насоновым [116] и Е.К. Пяткиным с соавторами [107]. В методических рекомендациях А.В. Севаньяева используется три вида оценок индивидуальных доз на основе линейно-квадратичной модели зависимости числа хромосомных aberrаций в культуре лимфоцитов человека *in vitro* от полученной дозы:

$$Y_1 = (1,9 \pm 0,9) + (0,274 \pm 0,012)D + (-0,000125 \pm 0,000034)D^2,$$

$$Y_2 = (1,0 \pm 0,6) + (0,143 \pm 0,019)D + (0,00108 \pm 0,000058)D^2,$$

$$Y_3 = (-0,9 \pm 0,5) + (0,099 \pm 0,017)D + (0,00049 \pm 0,000052)D^2,$$

где Y_1 — процент aberrантных клеток, Y_2 — общее количество aberrаций хромосом (на 100 клеток), Y_3 — сумма дицентриков и колец (на 100 клеток), D — доза, Р.

Методические рекомендации Е.К. Пяткина и соавторов [107] основаны на исследованиях *in vivo*, и для оценки индивидуальной дозы по результатам цитогенетического обследования культур лимфоцитов (при условии относительно равномерного гамма-облучения в дозах 1,5–5 Гр) авторами предлагается использовать зависимость

$$Y_4 = (9,55 \pm 1,74)D + (5,43 \pm 0,46)D^2,$$

где Y_4 — количество дицентриков в расчете на 100 клеток.

Таблица 3.18. Ретроспективные оценки степени ОЛБ и полученных доз облучения для ликвидаторов из обучающей выборки

Номер пациента	Степень ОЛБ	Оценка степени ОЛБ	Оценка дозы облучения, сГр	Оценки дозы облучения по Севаньяеву, сГр			Оценка дозы облучения по Пяткину, сГр
				1	2	3	
1	2	1,42	142	22	21	9	78
2	1	1,05	105	16	10	9	78
3	1	0,98	98	22	18	26	97
4	1	1,30	130	29	29	18	88

Номер пациента	Степень ОЛБ	Оценка сте- пени ОЛБ	Оценка дозы облучения, сГр	Оценки дозы облучения по Севаньяеву, сГр			Оценка дозы облуче- ния по Пяткину, сГр
				1	2	3	
5	1	0,87	87	54	60	34	106
7	2	1,65	165	30	33	11	81
8	2	2,13	225	64	69	67	135
11	1	1,43	143	22	18	20	91
12	3	2,22	244	89	97	82	152
13	2	1,94	194	41	42	26	97
14	1	0,57	57	22	18	18	78
15	1	0,99	99	14	7	9	78
16	3	3,21	442	217	213	17	244
17	1	1,21	121	48	58	85	146
18	1	1,06	106	22	18	26	88
19	1	0,97	97	29	29	26	97
24	1	0,99	99	14	7	9	78
40	1	1,74	174	42	44	9	78
41	2	1,53	153	25	24	9	78
42	1	1,32	132	18	13	13	83
43	1	0,96	96	19	14	24	95
44	2	1,73	173	39	40	34	97
45	1	1,40	140	25	24	18	88
46	1	1,25	125	20	18	9	78
47	1	0,93	93	22	18	26	97
48	1	1,28	128	25	24	9	78
49	2	2,22	244	67	82	67	140
50	2	1,67	167	27	26	22	93
52	3	2,88	376	85	91	90	163
53	1	1,11	111	11	0	9	78

Анализ данных, приведенных в табл. 3.18 и 3.19, показывает, что оценки индивидуальных доз облучения, полученные по предлагаемой нами методике, многократно завышены по сравнению с оценками по методике А.В. Севаньяева и лучше согласуются с оценками по методике Е.К. Пяткина и соавторов.

Таблица 3.19. Ретроспективные оценки степени ОЛБ и полученных доз облучения для ликвидаторов из тестовой выборки и интактных лиц

Номер пациента	Степень ОЛБ	Оценка сте- пени ОЛБ	Оценка дозы облучения, сГр	Оценки дозы облучения по Севанкяеву, сГр			Оценка дозы облуче- ния по Пяткину, сГр
				1	2	3	
6	1	1,35	135	19	14	15	85
9	1	1,56	156	25	29	13	78
10	1	0,86	86	14	7	18	78
20	1	1,39	139	37	38	26	97
21	1	0,74	74	25	24	26	83
22	1	1,42	142	22	18	9	78
23	1	1,15	115	33	33	26	88
25	1	1,10	110	18	13	9	78
26	1	1,12	112	39	42	18	88
27	1	1,65	165	42	49	18	88
28	1	1,97	197	32	34	12	82
29	1	1,18	118	25	26	13	83
30	1	1,71	171	29	29	23	84
31	1	1,32	132	18	13	22	93
32	1	1,21	121	14	7	9	78
33	1	1,55	155	35	38	26	97
34	1	1,11	111	11	0	9	78
35	1	2,07	214	46	60	18	88
36	1	1,28	128	25	24	18	88
37	1	1,01	101	7	0	9	78
38	1	1,29	129	25	24	9	78
39	1	1,32	132	18	13	9	78
51	1	2,04	208	228	206	221	279
Интактные лица							
54	0	0,95	95	8	0	9	79
55	0	0,89	89	10	0	11	79
56	0	0,96	96	11	0	11	79
57	0	0,94	94	12	0	11	79
58	0	0,96	96	12	0	9	79
59	0	1,06	106	13	0	9	79
60	0	0,94	94	13	1	9	79
61	0	1,04	104	13	1	11	79

Номер пациента	Степень ОЛБ	Оценка сте- пени ОЛБ	Оценка дозы облучения, сГр	Оценки дозы облучения по Севаньякаеву, сГр			Оценка дозы облуче- ния по Пяткину, сГр
				1	2	3	
62	0	0,86	86	14	1	11	79
63	0	1,12	112	15	3	9	79

По-видимому, это вызвано следующими причинами:

- исследования А.В. Севаньякаева основаны на облучении культур лимфоцитов человека *in vitro*, т.е. в идеальных условиях, когда отсутствует влияние элиминации поврежденных хромосом из организма облученного человека вследствие его индивидуальной лучевой реакции и проводимых терапевтических процедур;
- оценки в методике К.Е. Пяткина и соавторов получены при условии обследования пациента в течение либо суток, либо 4–6 недель с момента облучения, но не учитывают влияния терапевтических процедур, направленных на улучшение цитогенетических показателей крови облученного пациента.

Резюме

В нашем исследовании большинство облученных ликвидаторов в период с момента облучения до момента цитогенетического обследования (от 12 до 40 дней) подвергались интенсивным медицинским процедурам, направленным на лечение острой лучевой болезни (переливание крови, инъекции костного мозга и т.п.). Эти процедуры привели к нормализации цитогенетических показателей, но диагноз, поставленный по клиническим симптомам, не позволяет перевести таких больных в группу необлученных. Таким образом, мы получаем заведомо завышенные (по сравнению с методиками А.В. Севаньякаева и Е.К. Пяткина с соавторами) оценки индивидуальных доз в условиях интенсивного терапевтического вмешательства. Еще в большей степени это относится к группе интактных лиц, априори не получивших никакого облучения.

3.6. Влияние величины дозы облучения на вероятность развития злокачественных новообразований

Чернобыльская авария в связи с транскордонным характером влияния на население и природные объекты получила статус “Чернобыльской катастрофы” [92]. Только в Украине 8% населения оказалось вовлеченным в наибольшую радиоэкологическую катастрофу XX века [117]. Особые сложности возни-

заболеваний у людей, подвергшихся облучению [40]. Радиационный фактор Чернобыльской катастрофы в виде внешнего облучения, инкорпорации радионуклидов и радиационного загрязнения территории оказывает негативное влияние на здоровье людей [39]. Исследованиями зафиксировано повышение уровня злокачественных новообразований (ЗН) у участников ликвидации последствий аварии (УЛПА) на Чернобыльской АЭС, часть которых может быть радиационно-индуцированной [65, 105]. Несмотря на констатацию этого факта, диапазон представлений о значимости радиационного облучения, в частности в области действия низких доз, в развитии радиационного канцерогенеза широк — от гиперболизации роли радиации в возникновении ЗН [55] до недооценки ее влияния на клетки и в целом на организм человека [131]. Поэтому выяснение роли радиации, воздействию которой подверглись участники восстановительных работ в зоне аварии на ЧАЭС, в возникновении ЗН является актуальным. Эпидемиологические исследования среди пострадавших вследствие этой аварии сопряжены с рядом трудностей. Главная из них — несовершенство физической дозиметрии, хотя отчасти этот пробел восполняют данные биологической дозиметрии, выполненной в относительно ранние сроки после облучения [65].

Для оценки радиационных эффектов необходим выбор адекватной математической модели. Используемые линейно-квадратичные модели не всегда дают удовлетворительные приближения к истинным зависимостям [91, 129]. В связи с этим заслуживают внимания работы [12, 61, 91], в которых на различных объектах показано, что кусочно-линейные модели (так называемые модели сплайновой регрессии) являются более точными при аппроксимации зависимостей патологических эффектов от величины поглощенной дозы облучения в области малых доз. Цель этого раздела — изучить влияние радиации на вероятность возникновения ЗН в группе УЛПА на ЧАЭС в зависимости от величины дозы на основе использования сплайновой регрессии.

3.6.1. Материалы и методы исследования

Изучены спектр и частота распространенности заболеваний у 17 698 УЛПА, обратившихся в Киевскую региональную межведомственную экспертную комиссию для выявления причинной связи заболеваний с работами по ликвидации последствий аварии на ЧАЭС. Связь заболеваний с радиационным воздействием рассматривали индивидуально на основании характера клинических данных и документированных доз облучения. В комиссию обратились 1680 ликвидаторов со злокачественными новообразованиями и документированными дозами облучения, что составило 9,5% от общего количества обращений. Во всех случаях диагноз онкологического заболевания верифицирован патоморфологом.

В этом разделе использован метод внутреннего сравнения [53], который позволяет исследовать эпидемиологические показатели в зависимости от полученной дозы внешнего облучения, в данном случае в диапазоне доз 0,01–100 сГр. Для выявления дозовых зависимостей распределения заболеваний исследуемых

УЛПА разделили на девять групп согласно дозовым диапазонам. Под частотой класса заболеваний будем понимать отношение количества выявленных в этой группе заболеваний данного класса к общему количеству выявленных заболеваний в исследуемой группе, а под распространенностью — отношение количества выявленных заболеваний данного класса к количеству лиц в обследуемой группе [65].

Остановимся более подробно на математических основах обработки статистических данных.

3.6.2. Математические основы обработки статистических данных. Модифицированный полигон распределения

Введем понятие модифицированного полигона распределения, которое необходимо для статистического анализа зависимости вероятности злокачественного онкологического заболевания от величины поглощенной дозы радиации у ликвидатора аварии на ЧАЭС.

Пусть $x_1, x_2, \dots, x_n$ — выборка из генеральной совокупности, полученная простым случайным выбором. Введем следующие обозначения: $a = x_{\min} = \min(x_1, \dots, x_n)$, $b = x_{\max} = \max(x_1, \dots, x_n)$. Разобьем отрезок $[a, b]$ точками $y_1 < y_2 < \dots < y_l$ и обозначим через $x_1, \dots, x_{l_k}$ выборочные значения, попавшие в полуинтервал разбиения $[y_k, y_{k+1})$, $y_1 = a, y_l = b$ ($1 \leq k \leq l-1$), и отрезок $[y_{l-1}, y_l]$, тогда количество выборочных значений равно l_k . Пусть $h_k = \frac{l_k}{n}$ — частота таких выборочных значений, а $z_k = \frac{y_k + y_{k+1}}{2}$ — середина интервала (y_k, y_{k+1}) . Полигон распределения $P(u)$ представляет собой ломаную линию, состоящую из прямолинейных отрезков, соединяющих точки $M_0 = (a, 0)$, $M_1 = (z_1, h_1)$, $M_2 = (z_2, h_2)$, ..., $M_{l-1} = (z_{l-1}, h_{l-1})$, $M_l = (b, 0)$. За пределами отрезка (a, b) функция $P(u)$ считается равной нулю. Таким образом, полигон распределения $P(u)$ случайной величины x , построенный на основании выборки $x_1, x_2, \dots, x_n$, является кусочно-линейным сплайном с узлами в точках M_k ($k = 0, 1, \dots, l$). У модифицированного полигона распределения $P_k(u)$ в отличие от классического полигона $P(u)$ середина z_k отрезка разбиения (y_k, y_{k+1}) заменяется на выборочное среднее $\bar{z}_k$ тех выборочных значений x_p , которые принадлежат интервалу (y_k, y_{k+1}) . Если $x_1, \dots, x_m \in (y_k, y_{k+1})$, то $\bar{z}_k = \frac{1}{m} \sum_{j=1}^m x_{y_j}$.

Следовательно, модифицированный полигон $P_k(u)$ представляет собой линейный сплайн с узлами в точках $\bar{M}_0 = M_0 = (a, 0)$, $\bar{M}_1 = (z_1, h_1)$, $\bar{M}_2 = (z_2, h_2)$, ..., $\bar{M}_{l-1} = (z_{l-1}, h_{l-1})$, $\bar{M}_l = (b, 0)$; если $u \notin (a, b)$, то $P_k(u) = 0$ (рис. 3.4).

Гистограммы и полигоны распределения являются оценками для плотности вероятности случайной величины, однако последние представляют более точные оценки. Действительно, как известно из теории приближения, аппроксимация непрерывной функции линейными сплайнами имеет большую точность,

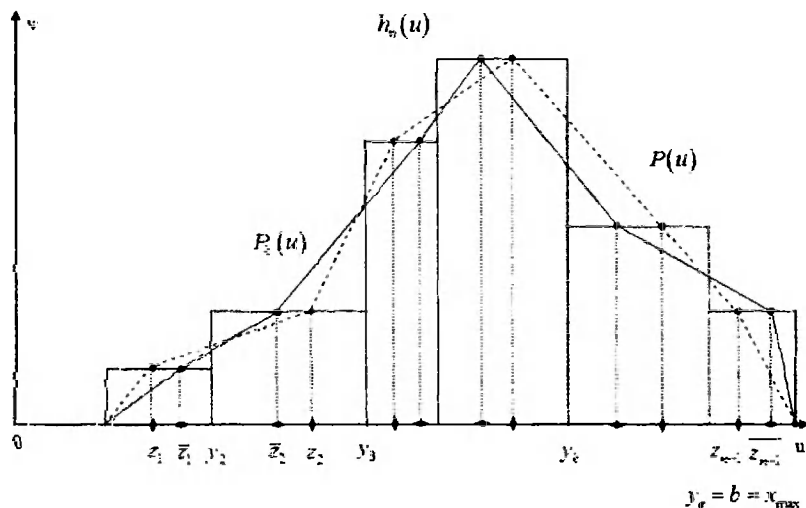


Рис. 3.4. Полигон распределения и модифицированный полигон распределения

чем ее приближение кусочно-постоянными функциями. Поскольку классические плотности вероятностей в своем большинстве являются непрерывными функциями, более предпочтительна их сплайновая аппроксимация. Кроме того, полигоны — это непрерывные функции, а гистограмма имеет много точек разрыва, что затрудняет ее использование в статистических вычислениях (например, для одномодальных распределений мода определяется по полигону распределения однозначно, а по гистограмме — неоднозначно и т.д.).

Если сравнивать между собой классические и модифицированные полигоны, то, на наш взгляд, последние имеют преимущества. Действительно, классические полигоны, у которых статистикой является лишь ордината узлов кусочно-линейного сплайна, в большой степени зависят от точек разбиения $y_1, y_2, \dots, y_i$ интервала (a, b) , которые в свою очередь определяются субъективно. У модифицированных полигонов и ордината и абсцисса узлов сплайна представляют собой функции выборочных значений, которые после выбора ячеек разбиения y_i не зависят от экспериментатора. В связи с этим в дальнейшем для оценки плотности вероятности случайной величины x (величина поглощенной дозы облучения у ликвидаторов со злокачественными новообразованиями) мы будем использовать модифицированный полигон.

Для исследования дозовой зависимости вероятности возникновения ЗН у УЛПА на ЧАЭС будем использовать линейные сплайны, состоящие из двух звеньев и имеющие соответственно два узла. Целесообразность такой аппроксимации основана на следующих соображениях. Основная цель этого раздела — показать, что низкие уровни радиации могут представлять серьезную опасность для здоровья облученных лиц, повышая вероятность развития ЗН. Если это предположение справедливо, то плотность вероятностей вышеуказанной случайной величины x , а потому и ее оценка — модифицированный полигон, будет значительно больше сосредоточена на интервале 3–20 сГр, чем на интервале

20–100 сГр, когда состояние УЛПА квалифицируется как лучевая реакция. В связи с этим при аппроксимации таких кривых двухзвенным сплайном первое звено (отрезок прямой) по сравнению со вторым должно иметь значительно больший по абсолютной величине угловой коэффициент, а второй узел можно интерпретировать как точку перехода процесса (функции), описывающего зависимость вероятности развития ЗН от величины поглощенной дозы радиации. Таким образом, для решения рассматриваемой задачи необходимо найти точку перехода и два указанных угловых коэффициента.

3.6.3. Аппроксимация модифицированных полигонов линейными сплайнами с двумя узлами. Метод сплайновой регрессии

Вычисление точки перехода (абсциссы второго узла двухзвенного сплайна) и угловых коэффициентов первого и второго звеньев осуществляется на основании метода сплайновой регрессии, описание которого изложим ниже.

Предположим, что плотность вероятностей $f(u)$ случайной величины x удовлетворительно можно аппроксимировать кусочно-линейный сплайном $g(u)$, состоящим из двух звеньев — $f(u) \approx g(u)$. Поскольку модифицированный полигон $P_k(u)$ является оценкой для плотности вероятностей $f(u)$, то $f(u) \approx P_k(u)$. Тогда полигон $P_k(u)$ можно приблизить сплайном $g(u)$: $P_k(u) \approx g(u)$ и определить точку перехода u^* полигона $P_k(u)$. Вычисление точки перехода u^* , которая по сути указывает на начало изменения процесса, описывающего зависимость вероятности ЗН у УЛПА от поглощенной дозы радиации, осуществляется с помощью метода сплайновой регрессии.

$$\text{Пусть } g(u, u^*, a_1, b_1, a_2, b_2) = \begin{cases} a_1 u + b_1, & \text{если } 0 \leq u \leq u^*, \\ a_2 u + b_2, & \text{если } u^* \leq u \leq T, \\ a_1 u^* + b_1 = a_2 u^* + b_2, \end{cases}$$

где u^* — точка перехода. Необходимо по значениям модифицированного полигона $P_k(u_i)$ ($i = 1, 2, \dots, n$) в точках u_i оценить точку перехода u^* и параметры $a_1, b_1; a_2, b_2$ сплайна $g(u)$ (рис. 3.5). Решение этой задачи осуществляется методом минимизации остаточной суммы квадратов. Предположим, что точка перехода u^* известна и совпадает с одной из точек наблюдения модифицированного полигона $P_k(u)$: $u^* = u_r$. Тогда задача вычисления параметров a_1, b_1, a_2, b_2 сводится к проблеме оценки параметров в схеме линейной регрессии. В этом случае математическое ожидание $m(g)$ можно представить в виде линейной комбинации фундаментальных сплайнов [66]. Используя классический метод наименьших квадратов (МНК) [112], можно получить так называемые оценки МНК $a_1^{(i)}, b_1^{(i)}, a_2^{(i)}, b_2^{(i)}$ для неизвестных параметров (эти оценки зависят от данной гипотезы $u^* = u_r$, а поэтому и от индекса i) и остаточную сумму квадратов

$$S_i^2 = \sum_{j=1}^n \left[P_k(u_j) - g(u_j, u^* = u_i, a_1, b_1, a_2, b_2) \right]^2.$$

Выбирая среди остаточных сумм $S_0^2, S_1^2, \dots, S_n^2$ наименьшую сумму $S_i^2 = \{S_0^2, S_1^2, \dots, S_n^2\}$ и определяя ее порядковый номер i^* , найдем оценку для точки перехода $u_i \approx u^*$ и оценки $a_1^{(i^*)}, b_1^{(i^*)}, a_2^{(i^*)}, b_2^{(i^*)}$ неизвестных параметров a_1, b_1, a_2, b_2 для сплайна $g(u)$, которые являются оценками, найденными по методу минимизации остаточных сумм квадратов. Эта оценка используется в модели сплайновой регрессии.

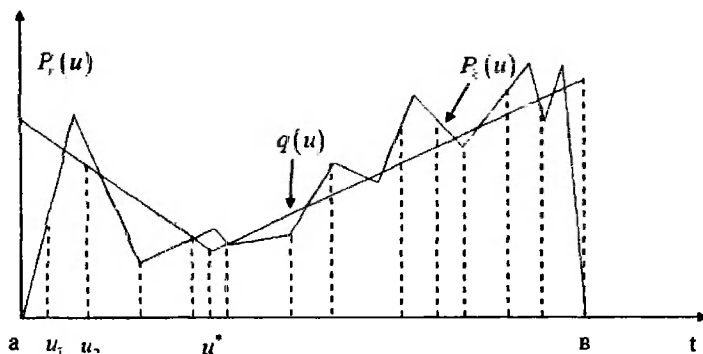


Рис. 3.5. Модель сплайновой регрессии

3.6.4. Определение количества точек перехода

Описанный выше алгоритм определения точки перехода имеет смысл лишь в том случае, когда функция $P_k(u) = f(u)$ имеет одну точку перехода. Очевидно, что у произвольной функции $P_k(u)$ точек перехода может быть несколько или они вообще могут отсутствовать. В связи с этим возникает сложная проблема определения точек перехода функции $P_k(u)$. Ее решение усложняется тем обстоятельством, что модифицированный полигон $P_k(u)$ представляет собой случайную функцию, поскольку ее значение зависит от выборки $x_1, x_2, \dots, x_n$ и точек разбиения (y_k, y_{k+1}) (если точки разбиения зафиксировать, то значения $P_k(u)$ будут зависеть лишь от выборки $x_1, x_2, \dots, x_n$).

Предлагаем следующее решение этой задачи, основанное на критерии Стьюдента. Пусть $(\bar{z}_1, P_k(\bar{z}))$, $\dots$, $(\bar{z}_{n-1}, P_k(\bar{z}_{n-1}))$ — вершины модифицированного полигона. Предположим, что этот полигон имеет лишь одну точку перехода u^* . Вычислим конечно-разностное отношение полигона $P_k(u)$:

$$\gamma_i = \frac{P_k(\bar{z}_{i+1}) - P_k(\bar{z}_i)}{\bar{z}_{i+1} - \bar{z}_i} \quad i = 1, 2, \dots, n-1.$$

Количество отношений γ_i на единицу меньше, чем количество вершин n модифицированного полигона $P_k(u)$. Предположим, что случайные величины $\gamma_1, \gamma_2, \dots, \gamma_{n-1}$ независимы в совокупности и имеют нормальное распределение с одинаковой дисперсией.

Рассмотрим первых два конечно-разностных отношения γ_1, γ_2 и построим по критерию Стьюдента доверительный интервал для основной распределенной массы с 5%-ным уровнем значимости [42]:

$$\gamma^{(1)} = \frac{1}{2}(\gamma_1 + \gamma_2) \quad s^2 = (\gamma^{(1)} - \gamma_1)^2 + (\gamma^{(1)} - \gamma_2)^2;$$

$$(a_1, b_1) = (\gamma^{(1)} - t_\beta s, \gamma^{(1)} + t_\beta s), t_\beta = 12,75; \quad \beta = 0,05,$$

где t_β — доверительная граница для статистики Стьюдента с одной степенью свободы, соответствующая уровню значимости β .

Если значение γ_3 не принадлежит интервалу (a_1, b_1) , то имеет место значимое различие математических ожиданий, так что (γ_1, γ_2) и γ_3 принадлежат разным генеральным совокупностям и тогда первая точка перехода находится между точками $\bar{z}_3$ и $\bar{z}_4$. Если $\gamma_3 \in (a_1, b_1)$, то рассматриваем следующее значение γ_4 и проверяем включения $\gamma_4 \in (a_1, b_1)$ и т.д. В случае, когда одно из γ_k (первое) отсутствует в интервале (a_1, b_1) , то существование точки перехода установлено — она расположена между точками $\bar{z}_k$ и $\bar{z}_{k+1}$. В противном случае, когда $\gamma_k \in (a_1, b_1)$ при всех k , то используемые методы не позволяют найти точку перехода. Предположим, что $\gamma_3 \notin (a_1, b_1)$, так что одну точку перехода мы отыскали. Рассмотрим далее γ_3 и γ_4 , построим аналогичный интервал (a_2, b_2) и будем проверять включение $\gamma_5, \gamma_6, \dots, \gamma_n$ в интервал (a_2, b_2) . Если при всех $k \geq 5$ $\gamma_k \in (a_2, b_2)$, то рассмотрим доверительный интервал (a_3, b_3) :

$$a_3 = \gamma^{(2)} - t_\beta s, \quad b_3 = \gamma^{(2)} + t_\beta s, \quad \text{где} \quad \gamma^{(2)} = \frac{1}{3}(\gamma_3 + \gamma_4 + \gamma_5),$$

$$s^2 = \frac{1}{2} \left[(\gamma^{(2)} - \gamma_3)^2 + (\gamma^{(2)} - \gamma_4)^2 + (\gamma^{(2)} - \gamma_5)^2 \right], \quad t_\beta = 4,303, \quad \beta = 0,05.$$

Если при этом $\gamma_k \in (a_3, b_3) \forall k \geq 6$, то строится интервал (a_4, b_4) ,

$$a_4 = \gamma^{(3)} - t_\beta s, \quad b_4 = \gamma^{(3)} + t_\beta s, \quad \text{где} \quad \gamma^{(3)} = \frac{1}{4}(\gamma_3 + \gamma_4 + \gamma_5 + \gamma_6),$$

$$s^2 = \frac{1}{3} \left[(\gamma^{(3)} - \gamma_3)^2 + (\gamma^{(3)} - \gamma_4)^2 + (\gamma^{(3)} - \gamma_5)^2 + (\gamma^{(3)} - \gamma_6)^2 \right], \quad t_\beta = 3,182,$$

$$\beta = 0,05.$$

Затем проверяем включение $\gamma_k \in (a_4, b_4) \forall k \geq 7$ и т.д. Отсутствие последующих значений γ_k во всех указанных интервалах (a_k, b_k) означает, что существует лишь одна точка перехода, расположенная между точками $\bar{z}_3$ и $\bar{z}_4$.

Если установлено, что существует лишь одна точка перехода u^* , то следует проверять статистическую значимость ее существования. Для этого необходимо

разбить множество конечно-разностных отношений γ_i на две группы — G_1 и G_2 . К первой группе G_1 относятся те величины γ_i

$$\gamma_i = \frac{P_k(\bar{z}_{i+1}) + P_k(\bar{z}_i)}{\bar{z}_{i+1} - \bar{z}_i},$$

для которых соответствующие значения $\bar{z}_i$ не превышают: $u^* : \bar{z}_i \leq u^*$. Вторая группа включает все остальные значения γ_i .

Пусть $G_1 = \{\gamma_1, \dots, \gamma_{k_0}\}$, $G_2 = \{\gamma_{k_0+1}, \dots, \gamma_{n-1}\}$,

$$\bar{\gamma}^{(2)} = \frac{1}{n-1-k_0} \sum_{i=k_0+1}^{n-1} \gamma_i, \quad \bar{s}^{(2)} = \sqrt{\frac{1}{n-2-k_0} \sum_{i=k_0+1}^{n-1} (\gamma_i - \bar{\gamma}^{(2)})^2},$$

а $I_2 = (\bar{\gamma}^{(2)} - 3\bar{s}^{(2)}, \bar{\gamma}^{(2)} + 3\bar{s}^{(2)})$ — доверительный интервал для величин γ_i , построенный по правилу $3s$ (рис. 3.6).

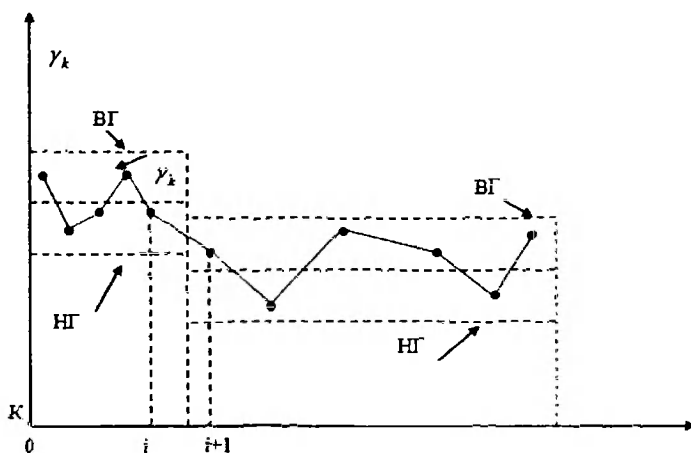


Рис. 3.6. Определение количества точек перехода

В силу правила $3s$ [98], которое является более общим критерием, чем критерий Стьюдента (при использовании правила $3s$ не предполагается, что выборочные значения имеют нормальное распределение), считается, что существование точки перехода u^* является статистически значимым, если хотя бы одно из конечно-разностных отношений группы G_1 не принадлежит интервалу I_2 .

3.6.5. Результаты вычислений

Для проверки истинности основной гипотезы H о статистически значимом увеличении вероятности развития раковых заболеваний среди УЛПА на Чернобыльской АЭС, подвергшихся воздействию радиационного фактора аварии, диапазон документированных доз (1–100 сГр) разбит на $i = 9$ интервалов: 1–3, 3–5, 5–10, 10–15, 15–20, 20–25, 25–50, 50–75, 75–100 сГр. В табл. 3.20 представлены

значения абсцисс и ординат вершин модифицированного полигона для этого разбиения. Абсцисса $\bar{z}_i$ — средняя доза поглощенной радиации в i -м диапазоне, а ордината $P(\bar{z}_i)$ — частота злокачественных новообразований в i -м диапазоне. Конечно-разностные отношения γ_i , построенные для этого модифицированного полигона, приведены в табл. 3.21.

Таблица 3.20. Координаты вершин модифицированного полигона

i	Дозы, сГр	$\bar{z}_i$	$P(\bar{z}_i)$
1	1–3	1,289	0,1185
2	3–5	3,812	0,0840
3	5–10	7,024	0,0364
4	10–15	12,455	0,0154
5	15–20	17,139	0,0168
6	20–25	21,513	0,0182
7	25–50	27,967	0,0056
8	50–75	53,333	0,00084
9	75–100	89,957	0,0000

Таблица 3.21. Конечно-разностные отношения модифицированного полигона

i	γ_i
1	–0,013674
2	–0,014819
3	–0,003866
4	0,0002988
5	0,000320
6	–0,001952
7	–0,0001876
8	–0,0000229

Доверительный интервал (a_1, b_1) , построенный по γ_1 и γ_2 на основании критерия Стьюдента, имеет вид $(-0,0240903; -0,0044025)$. Значение γ_3 не принадлежит этому интервалу, поэтому наблюдается статистически значимое различие и первая точка перехода u^* располагается на интервале $(\bar{z}_3, \bar{z}_4) = (7,024; 12,455)$. Следующие доверительные интервалы (a_k, b_k) , $k = \overline{1,4}$, имеют вид $(-0,028528; 0,0249608)$, $(-0,0113551; 0,0091903)$, $(-0,0076638; 0,0050642)$, $(-0,006043; 0,0038884)$. Анализ данных, представленных в табл. 3.21, показывает, что все последующие значения γ_i принадлежат указанным доверительным интервалам (например, $\gamma_5, \gamma_6, \gamma_7, \gamma_8 \in (-0,028528; 0,0249608)$). Поэтому модифицированный полигон $P_k(z)$, угловые коэффициенты которого представлены в табл. 3.20, не имеет других точек перехода, кроме одной точки u^* . Далее, доверительный интервал I_2 , построенный для основной распределенной массы значений для второй группы G_2 с помощью правила $3s$, имеет вид $I_2 = (-0,0071224; 0,0053192)$. При этом ни одно выборочное значение γ_i из первой группы G_1 (т.е. γ_1, γ_2) не принадлежит этому интервалу, что является еще одним аргументом в пользу статистически значимого существования точки перехода u^* в интервале $(\bar{z}_3, \bar{z}_4)$.

Аппроксимация модифицированного полигона $P_k(u)$ с помощью линейного сплайна с минимальной остаточной суммы квадратов (сплайновая регрессия) имеет следующий вид (рис. 3.7).

$$P_k(u) = S(u) = \begin{cases} -0,017287u + 0,144056, & \text{если } 0 \leq u \leq 7,024, \\ -0,0003473u + 0,02507, & \text{если } 7,024 \leq u \leq 100. \end{cases}$$

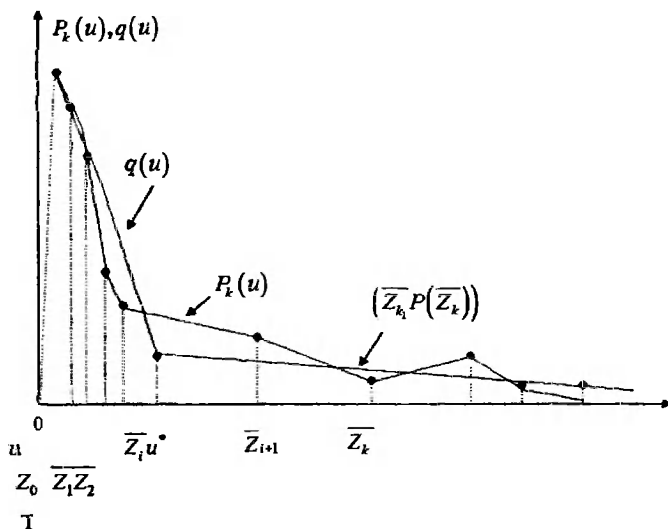


Рис. 3.7. Аппроксимация модифицированного полигона с помощью линейного сплайна

Таким образом, точка перехода равна 7,024, что хорошо согласуется с другими полученными результатами, которые описаны далее. При этом остаточная сумма квадратов равна $4,3461 \cdot 10^{-4}$, так что точность приближения сплайном $S(u)$ является достаточно высокой. Следует отметить, что наилучшим приближением модифицированного полигона в классе линейных функций является функция $l(u) = -0,00089634u + 0,05621$ ($0 \leq u \leq 100$), остаточная сумма квадратов которой равна $8,295 \cdot 10^{-3}$. Наилучшее приближение модифицированного полигона в классе линейно-квадратичных функций имеет вид $f(u) = 0,00003309u^2 - 0,003851u + 0,08651$ ($0 \leq u \leq 100$), причем ее остаточная сумма квадратов равна $3,806 \cdot 10^{-3}$. Таким образом, среди трех проанализированных аппроксимаций наилучшее приближение модифицированного полигона представляет сплайновая регрессия.

Дополнительные аргументы в пользу гипотезы H о статистически значимом увеличении вероятности раковых заболеваний при малых дозах облучения получены нами раньше при ретроспективном анализе данных о 2076 ликвидаторах, работавших в 1986–1987 годах в зоне аварии на ЧАЭС [60, 72]. Для выявления дозовых зависимостей исследованную когорту ликвидаторов разделили согласно дозовым диапазонам 1–10, 10–25 и 25–100 сГр и возрастным группам — младше и старше 40 лет. Для оценки общего характера дозовой зависимости была использована модель линейной регрессии, в которой независимой переменной x является величина дозы, а откликом y — частота заболеваний (в процентах). Установлено, что для ЗН наблюдается тенденция уменьшения частоты заболеваний при увеличении дозы в обеих возрастных группах. В модели линейной регрессии $y = A + Bx$ получены такие оценки углового коэффициента B прямой регрессии: для группы моложе 40 лет $B = -0,031$, а для группы

старше 40 лет $B = -0,089$. Кроме оценки общего характера дозовой зависимости, исследована согласованность полученных данных о частоте злокачественных новообразований с гипотезой H , утверждающей, что вероятность развития заболеваний в двух различных дозовых диапазонах одинакова (т.е. между соответствующими частотами нет значимых различий). Решение этой проблемы было получено с помощью критерия χ^2 с 5%-ным уровнем значимости. На основании данных и результатов статистического анализа можно сделать вывод: при сравнении вероятностей возникновения злокачественных новообразований в диапазонах доз 1–10 и 25–100 сГр в группе ликвидаторов младше 40 лет гипотезу H следует отклонить, поскольку между соответствующими частотами наблюдается статистически значимое различие. Отметим также, что среди лиц старше 40 лет этот эффект выражен более убедительно: вероятность возникновения ЗН в диапазоне доз 1–10 сГр статистически значимо отличается от вероятности возникновения ЗН в диапазонах 10–25 и 25–100 сГр. Таким образом, малые дозы облучения представляют не меньшую опасность для здоровья облученных лиц, чем средние и высокие дозы (от 25 до 100 сГр).

Другие аргументы в пользу гипотезы H приведены нами в работах [8, 73].

Подводя итог проведенным исследованиям, можно сделать следующее заключение. Поскольку модифицированный полигон характеризует зависимость вероятности возникновения ЗН у УЛПА на ЧАЭС от поглощенной дозы облучения, а его достаточно точная аппроксимация имеет точку перехода 7,024, после которой вероятность возникновения онкологических заболеваний резко снижается, то наиболее опасными для здоровья ликвидаторов следует считать малые поглощенные дозы радиации, заключенные в интервале от 3 до 15 сГр.

Полученные нами факты на основе использования модифицированного полигона распределения могут быть основой гипотезы, согласно которой низкие уровни радиации являются статистически значимыми факторами риска возникновения злокачественных новообразований. Это обусловлено недостаточной интенсивностью антиканцерогенной защиты организма при действии радиации в указанном диапазоне доз, включающей элиминацию аберрантно измененных клеток системой иммунного надзора за антигенным постоянством внутренней среды организма, процессы индуцибельной репарации, апоптоз, нормализацию функций и т.д.

Резюме

Итак, модифицированный полигон распределения правильно характеризует зависимость вероятности возникновения злокачественных новообразований у УЛПА на ЧАЭС от поглощенной дозы облучения, а его достаточно точная аппроксимация имеет точку перехода 7,024, после которой вероятность возникновения онкологических заболеваний резко снижается. Наиболее опасными для развития радиогенных злокачественных новообразований у облученных лиц следует считать малые поглощенные дозы радиации в интервале от 3 до 15 сГр.

3.7. Стратификационный анализ популяций раковых клеток, резистентных к цисплатину

Сложная проблема современной клинической онкологии — неудовлетворительное лечение больных с новообразованиями в связи с развитием резистентности клеток опухоли к применяемым противоопухолевым препаратам. В резистентных опухолях развиваются адаптационные клеточные реакции, что позволяет им не реагировать на цитотоксическое действие препарата, из-за чего клетки не погибают, что приводит к прогрессии заболевания.

В последние годы появилось большое количество работ по изучению молекулярных механизмов развития резистентности опухоли к противоопухолевым препаратам, за счет чего выявлены белки, экспрессия которых изменяется при развитии резистентности. Однако в резистентных клетках, помимо изменения экспрессии элементов протеома, происходит значительная перестройка морфологических структур, поэтому увеличивается их размер, в клетке появляется значительно большее количество клеточных органелл, изменяется ядерно-цитоплазматическое соотношение. Нами был применен стратификационный анализ популяций раковых клеток, резистентных к цитостатикам, что позволило показать гетерогенность интактной популяции и выделить основные морфометрические параметры, изменения которых определяются в клетках опухоли при развитии устойчивости к цитотоксическому действию противоопухолевого препарата — цисплатина.

3.7.1. Экспериментальные модели резистентности опухоли к действию цисплатина

Для исследования использована клеточная линия эпидермоидной карциномы ротовой полости человека KB (исследования *in vitro*). Линия получена из банка клеточных культур Института экспериментальной патологии, онкологии и радиобиологии им. Р.Е. Кавецкого НАН Украины. Резистентную к действию цисплатина линию KB/DDP5 получили после длительного (8 месяцев) культивирования клеток выходной линии KB в среде с добавлением препарата. Культивирование клеток с цисплатином начинали с концентрации цитостатика в среде 0,01 мкг/мл. Цисплатин добавляли через день после пересева клеток. Концентрацию цисплатина увеличивали в два раза за месяц. Каждые два месяца проводили тестирование клеток KB/DDP5 для установления степени резистентности с применением МТТ-теста. Через 8 месяцев культивирования клеток с цитостатиком, концентрация которого в среде составляла уже 0,5 мкг/мл, было установлено, что полученная нами линия KB/DDP5 была в 5,4 раза более стойкая к действию препарата, в отличие от выходной линии KB. Уровень резистентности определяли с помощью показателя торможения роста клеток IC_{50} (Inhibitable Concentration 50), который равняется такой дозе препарата, которая тормозит рост клеток на 50%. Такой показатель стабилен для определенного типа клеток и выбранного противоопухолевого препарата. Величина IC_{50} для

клеток линий KB и KB/DDP5 составляет 0,5 и 2,7 мкг/мл соответственно. Соотношение этих величин показывает, во сколько раз одна клеточная линия более стойкая к цитотоксическому действию препарата. *In vivo* выработка резистентности проводилась на лабораторных беспородных крысах-самцах массой от 150 до 200 г. разводки вивария Института экспериментальной патологии, онкологии и радиобиологии им Р.Е. Кавецкого. На первом этапе исследования животным переливали подкожно на бедро 20%-ную суспензию клеток карциномы Герена. В опытной группе животных проводили курс лечения цисплатином: 5 инъекций через 2 суток внутривенно в дозе 1,2 мг/кг. После окончания курса химиотерапии из опухоли готовили 20%-ную суспензию опухолевых клеток, которую на втором этапе переливали новым группам животных — контрольной и опытной. Лечение животных опытной группы на втором этапе проводили по схеме, описанной выше. В конечном итоге было проведено 27 курсов терапии цисплатином. Препарат тормозил рост исходной опухоли на 95–97%, а после 27 курсов на 0–4%.

3.7.2. Приготовление препаратов для проведения морфометрического исследования

Препараты для морфометрического исследования клеточной линии KB получены путем окрашивания клеток по Романовскому. Для морфометрического анализа опухолей Герена кусочки ткани опухоли фиксировали в 4%-ном забуференном растворе формалина. После обезвоживания в спиртах восходящей концентрации материал заливали в парафин по общепринятой методике. Парафиновые срезы окрашивали гематоксилин-эозином.

Изучение морфометрических показателей клеток проводили с помощью полуавтоматического цитоанализатора Интеграл 2-МТ в режиме “Маркер” на парафиновых срезах опухолей Герена, чувствительных и резистентных к цитотоксическому действию цисплатина.

3.7.3. Статистические методы исследования гетерогенных генеральных совокупностей

Основы стратификационного анализа генеральной совокупности G на основе выборки $x_1, x_2, \dots, x_n \in G$, полученной простым случайным выбором, были заложены в работах [13, 67]. Клеточная популяция является частным случаем абстрактной генеральной совокупности, которая представляет собой произвольное множество индивидуумов. В математической статистике изучаются лишь количественные характеристики индивидуумов (площади ядер, периметры клеток и т.п.), так что в дальнейшем под генеральной совокупностью мы будем понимать произвольное множество числовых значений G , на котором задано некоторое распределение вероятностей $F(u) = p(x < u)$, возможно, неизвестное экспериментатору.

Генеральные совокупности разделяются на два класса — гомогенные (однородные, или чистые) и гетерогенные (смешанные, или смеси). Несмотря на кажущуюся простоту, трудно дать правильное истолкование этих понятий, не прибегая к математической формализации. Следуя А.Я. Хинчину [122], будем называть функцию распределения $F(u)$ одномодальной, если существует такая точка M на прямой, называемая модой, что на луче $(-\infty, M)$ она выпукла, а на луче $(M, +\infty)$ — вогнута. Заметим, что мода M у одномодального по Хинчину распределения может быть неединственной. Название “одномодальное” возникло оттого, что плотность вероятностей $f(u)$, соответствующая функции распределения $F(u)$, имеет лишь одно максимальное значение, достигаемое в точке M .

Генеральная совокупность G с функцией распределения $F(u)$ называется гомогенной (однородной), если функция $F(u)$ является одномодальной, в противном случае генеральная совокупность G называется гетерогенной. Таким образом, основной отличительной чертой гетерогенной генеральной совокупности является полимодальность ее функции распределения.

Рассмотрим алгоритм стратификационного анализа генеральной совокупности. Предположим, что G — однородная генеральная совокупность, из которой простым случайным выбором извлечена выборка $x_1, x_2, \dots, x_n$. Построим по ней вариационный ряд $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$. Модифицированную эмпирическую функцию распределения, представляющую собой кусочно-линейный сплайн, заданный на отрезке $[a, b]$, можно определить следующим образом:

$$\tilde{F}_n^*(u) = \begin{cases} \frac{u}{n} + \frac{kx_{(k+1)} - (k+1)x_{(k)}}{n}, & \text{если } k=0, 1, 2, \dots, n; x_{(k)} \leq u \leq x_{(k+1)}, x_{(0)} = a; \\ 1, & \text{если } x_{(n)} \leq u \leq b. \end{cases}$$

Модифицированная эмпирическая функция распределения $\tilde{F}_n^*(u)$ является непрерывной монотонной возрастающей функцией с компактным носителем, поэтому существует обратная эмпирическая функция распределения $\tilde{F}_n^{*-1}(u)$. Обратная функция представляет собой кусочно-линейный сплайн с узлами в точках $(k/n, x_{(k)})$, $k=0, 1, 2, \dots, n$. В некоторых случаях эмпирическую обратную функцию распределения $\tilde{F}_n^{*-1}(u)$ удобно рассматривать как функцию натурального аргумента, полагая

$$\tilde{\Psi}_n^*(i) = \tilde{F}_n^*\left(\frac{i}{n}\right), i = 0, 1, 2, \dots, n.$$

Как показывает анализ обратной функции распределения, если G — гомогенная генеральная совокупность, то в окрестности точки перегиба u^* функция $\tilde{\Psi}_n^*(u)$ хорошо приближается линейной функцией $L(u) = au + b$. Если распределение $F_G(u)$ является многомодальным, т.е. G — гетерогенная генеральная совокупность, количество точек перегиба равно количеству мод, поэтому функция $\tilde{F}_n^*(u)$ хорошо аппроксимируется многозвенными кусочно-линейными

сплайнами. Это позволяет сформулировать следующие критерии гомогенности генеральной совокупности.

Критерий 1. Используя метод минимизации остаточной суммы квадратов, построим кусочно-линейный сплайн $S_n^*(u)$, аппроксимирующий функцию $\tilde{\Psi}_n^*(i)$:

$$S_n^*(u) = \begin{cases} a_1 u + b_1, & \text{если } 1 \leq u \leq i_1, \\ a_2 u + b_2, & \text{если } i_1 \leq u \leq i_2, \\ \dots, \\ a_m u + b_m, & \text{если } i_{m-1} \leq u \leq i_m, \\ a_{i-1} u + b_{i-1} = a_i u + b_i, & i = 1, 2, \dots, m. \end{cases}$$

Вычислим доверительные границы $(a_i^{(i)}, a_i^{(d)})$, $(b_i^{(i)}, b_i^{(d)})$, $i = 1, 2, \dots, m$ для неизвестных коэффициентов сплайновой регрессии по правилу $3s$ (эти доверительные интервалы имеют уровень значимости, приближенно равный 5%). Если между коэффициентами a_i или между коэффициентами b_i , $i = 1, 2, \dots, m$ существуют статистически значимые отличия, т.е. существует такой индекс k , что $a_i \notin (a_k^{(i)}, a_k^{(d)})$, или существует такой индекс l , что $b_i \notin (b_l^{(i)}, b_l^{(d)})$, то генеральная совокупность G считается гетерогенной и состоит как минимум из двух модальных классов. В противном случае генеральная совокупность G считается гомогенной. Как правило, количество m не превышает трех.

Предположим, что генеральная совокупность G состоит из двух модальных классов. Обозначим через $S_n^*(u)$ двухзвенный кусочно-линейный сплайн, наилучшее приближение функции $\tilde{F}_n^*(u)$ среди всех сплайнов этого типа. Пусть i_1 — точка перехода сплайна $S_n^*(u)$. Естественно считать, что порядковые статистики $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(i_1)}$ принадлежат к первому модальному классу, а $x_{(i_1+1)} \leq x_{(i_1+2)} \leq \dots \leq x_{(n)}$ — второму. Оценки математического ожидания и дисперсии модальных классов имеют вид

$$m_1 = \bar{x}^{(1)} = \frac{1}{i_1} \sum_{k=1}^{i_1} x_{(k)}, \quad \sigma_1^2 = s_1^2 = \frac{1}{i_1 - 1} \sum_{k=1}^{i_1} (x_{(k)} - \bar{x}^{(1)})^2, \\ m_2 = \bar{x}^{(2)} = \frac{1}{n - i_1 - 1} \sum_{k=i_1+1}^n x_{(k)}, \quad \sigma_2^2 = s_2^2 = \frac{1}{n - i_1 - 2} \sum_{k=i_1+1}^n (x_{(k)} - \bar{x}^{(2)})^2.$$

Доли первого и второго модальных классов равны $\frac{i_1}{n}$ и $\frac{n - i_1}{n}$ соответственно. Совершенно аналогично проводится анализ генеральной совокупности, состоящей из трех и больше модальных классов.

Критерий 2. Предположим, что генеральная совокупность G гомогенна. В этом случае функция $\tilde{F}_n^{-1}(u)$ хорошо аппроксимируется линейной функцией $L(u) = au + b$, а выборочное среднее величин $z_i = x_{(i)} - x_{(i-1)}$, называемых спейсингами, близко к угловому коэффициенту прямой $L(u) = au + b$, т.е. к числу a .

В самом деле, полагая $u_i = i/n$, имеем

$$a = \frac{L(u_{i+1}) - L(u_i)}{u_{i+1} - u_i} \approx \frac{\tilde{\Psi}_n^*(i+1) - \tilde{\Psi}_n^*(i)}{i+1-i} = x_{(i+1)} - x_{(i)} \approx \frac{1}{n} \sum_{i=1}^n z_i = \bar{z},$$

поскольку $L(u_i) \approx \tilde{\Psi}_n^*(i)$ и $x_{(i)} = \tilde{\Psi}_n^*(i)$.

Таким образом, для гомогенной генеральной совокупности G на отрезке $[x_{(1)}, x_{(n)}]$ спейсинги z_i имеют одинаковые математические ожидания, близкие к числу a . Если же генеральная совокупность G является смесью двух генеральных совокупностей, то обратная функция распределения также будет хорошо аппроксимироваться на некоторых отрезках $[x_{(1)}, x_{(n)}]$ и $[x_{(r+1)}, x_{(n)}]$ прямыми $L_1(u) = a_1 u + b_1$ и $L_2(u) = a_2 u + b_2$. Проводя аналогичные рассуждения, нетрудно убедиться, что спейсинги z_i на отрезке $[x_{(1)}, x_{(n)}]$ будут принимать значения, близкие к числу a_1 , а на отрезке $[x_{(r+1)}, x_{(n)}]$ — близкие к величине a_2 . Отсюда следует, что выборочное среднее спейсингов на отрезке $[x_{(1)}, x_{(n)}]$ при больших значениях n будет статистически значимо отличаться от выборочного среднего спейсингов на отрезке $[x_{(r+1)}, x_{(n)}]$. Если угловые коэффициенты a_1 и a_2 статистически значимо отличаются, то это различие можно обнаружить с помощью хорошо известных статистических критериев для сравнения двух средних, например, с помощью правила $3s_1$. Точка $x_{(r)}$ вычисляется методом скользящих средних: ею является очередная порядковая статистика, выходящая за пределы доверительного интервала для скользящих средних спейсингов $\bar{z}_1, \bar{z}_2, \dots, \bar{z}_{l-1}$. Легко видеть, что описанный метод позволяет обнаружить любое количество точек перехода.

Таким образом, можно сформулировать следующий критерий: генеральная совокупность G является гомогенной, если график ее вариационного ряда не имеет ни одной точки перехода. В противном случае генеральная совокупность G гетерогенна.

3.7.4. Результаты исследования

Анализ полученных результатов позволил установить различия между такими морфометрическими показателями, как площадь ядра, периметр ядра и фактором формы в чувствительных и резистентных клетках опухолей Герена. Показано, что площадь ядер чувствительных клеток ($46,14 \pm 1,50$) достоверно отличается от площади ядер резистентного штамма опухоли ($62,75 \pm 1,42$). Морфологически это выявляется в увеличении размеров ядра клеток резистентных опухолей и в значительной мере обусловлено изрезанностью ядерного контура, что и вызывает значительное увеличение их площадей.

Периметр ядер резистентных клеток ($30,16 \pm 0,34$) также был достоверно увеличен по сравнению с исходным подштаммом опухоли Герена ($26,42 \pm 0,44$).

Такой параметр, как фактор формы ядер, также имел различия в клетках карциномы Герена, имеющих разную чувствительность к цисплатину. В клетках резистентного подштамма было установлено снижение этого параметра по

сравнению с исходным ($1,18 \pm 0,0056$ против $1,27 \pm 0,0092$), что свидетельствует о более округлой форме ядра в чувствительных клетках, что также было показано при морфологическом и ультраструктурном исследовании [29].

Рассмотрим результаты стратификационного анализа исследования *in vitro* популяции клеток плоскоклеточного рака ротовой полости, резистентных к действию цисплатина.

Анализ данных, приведенных в табл. 3.22, показывает, что в резистентных клетках линии КВ наблюдается увеличение всех трех изученных показателей (форм-фактора, периметра и площади), но оно не является статистически значимым, поскольку доверительные интервалы для основной распределенной массы этих показателей в случае резистентных и интактных клеток пересекаются.

Таблица 3.22. Выборочные средние значения и 95%-ные доверительные интервалы для основной распределенной массы морфометрических показателей клеток, построенные по правилу 3s

Показатель	Среднее	НГ	ВГ	s	min	max
1	1,143	0,980	1,307	0,055	1,060	1,560
2	1,149	0,978	1,320	0,057	1,060	1,470
3	48,351	19,245	77,458	9,702	25,500	72,000
4	58,477	17,884	99,071	13,531	27,600	103,800
5	169,461	0	365,766	65,435	44,000	360,000
6	248,878	0	585,674	112,265	50,000	704,000

Расшифровка обозначений. 1 — форм-фактор интактной популяции; 2 — форм-фактор резистентной популяции; 3 — площадь клеток интактной популяции; 4 — площадь клеток резистентной популяции; 5 — периметр клеток интактной популяции; 6 — периметр клеток резистентной популяции; НГ — нижняя доверительная граница; ВГ — верхняя доверительная граница; s — стандартное отклонение.

Перейдем к анализу выборочных средних значений и доверительных интервалов для неизвестного математического ожидания указанных морфометрических показателей.

Как видим, средняя площадь и средний периметр клеток линии КВ, резистентных к действию цисплатина, значительно превышают среднюю площадь и средний периметр интактных клеток, поскольку соответствующие доверительные интервалы не пересекаются.

В табл. 3.23 приведены меры близости между соответствующими морфометрическими показателями резистентных и интактных клеток [74] и их 95%-ные доверительные границы. Во всех трех случаях различия между популяциями следует считать значимыми, поскольку доверительный уровень 0.95 не принадлежит указанным доверительным интервалам.

Таблица 3.23. Средние выборочные значения и доверительные интервалы для неизвестного математического ожидания морфометрических показателей клеток линии KB, построенные по правилу $3s$

Показатель	Среднее	НГ	ВГ	s ,
1	1,143	1,135	1,152	0,003
2	1,149	1,140	1,158	0,003
3	48,351	46,817	49,885	0,511
4	58,477	56,338	60,617	0,713
5	169,461	159,115	179,807	3,449
6	248,878	231,127	266,628	5,918

Расшифровка обозначений. 1 — форм-фактор интактной популяции; 2 — форм-фактор клеток резистентной популяции; 3 — площадь клеток интактной популяции; 4 — площадь клеток резистентной популяции; 5 — периметр клеток интактной популяции; 6 — периметр клеток резистентной популяции; $s_i = s/\sqrt{n}$.

Аналогичные исследования были проведены для морфометрических показателей ядер клеток карциномы Герена у крыс.

В табл. 3.24 приведены выборочные средние значения и 95%-ные доверительные интервалы для основной распределенной массы морфометрических показателей ядер карциномы Герена, построенные по правилу $3s$.

Таблица 3.24. Меры близости между морфометрическими показателями клеток линии KB

Показатель	Мера близости	Нижняя доверительная граница	Верхняя доверительная граница
Форм-фактор	0,693	0,687	0,698
Площадь	0,294	0,288	0,299
Периметр	0,302	0,297	0,307

Анализ данных, приведенных в табл. 3.25, показывает, что для клеток карциномы Герена, резистентных к действию цисплатина, наблюдается увеличение основных морфометрических показателей, однако значимого различия между этими показателями нет, поскольку соответствующие доверительные интервалы пересекаются.

Таблица 3.25. Выборочные средние значения и 95%-ные доверительные интервалы для основной распределенной массы морфометрических показателей ядер карциномы Герена, построенные по правилу $3s$

Показатель	Среднее	НГ	ВГ	s	min	max
1	1,183	0,950	1,416	0,078	1,070	1,450

Показатель	Среднее	НГ	ВГ	s	min	max
2	1,267	0,883	1,651	0,128	1,090	1,980
3	30,328	15,401	45,254	4,975	21,200	41,400
4	26,297	8,522	45,073	6,092	14,700	43,600
5	63,529	1,302	125,757	20,742	30,000	119,000
6	47,573	0	110,523	20,384	14,000	99,000

В связи с этим представляет интерес сравнение выборочных средних значений этих показателей, приведенных в табл. 3.26. Данные, приведенные в табл. 3.26, свидетельствуют о значимом увеличении средних значений всех основных морфометрических показателей ядер клеток карциномы Герена, поскольку соответствующие им доверительные интервалы не пересекаются.

Таблица 3.26. Выборочные средние значения и 95%-ные доверительные интервалы для основной распределенной массы морфометрических показателей ядер карциномы Герена, построенные по правилу $3s_i$

Показатель	Среднее	НГ	ВГ	s_i
1	1,183	1,155	1,211	0,009
2	1,267	1,241	1,293	0,009
3	30,328	28,518	32,138	0,603
4	26,297	25,565	28,029	0,411
5	63,529	55,983	71,076	2,515
6	47,573	43,329	51,817	1,415

Этот же вывод подтверждают данные из табл. 3.27, так как представленные там доверительные интервалы для меры близости между генеральными совокупностями основных морфологических показателей интактных и резистентных клеток не содержат доверительный уровень 0,95.

Таблица 3.27. Меры близости между морфометрическими показателями ядер карциномы Герена

Показатель	Мера близости	Нижняя доверительная граница	Верхняя доверительная граница
Форм-фактор	0,356	0,326	0,386
Площадь	0,478	0,446	0,509
Периметр	0,489	0,458	0,520

График вариационного ряда для площади клеток из интактной популяции линии КВ плоскоклеточного рака ротовой полости человека (рис. 3.8) аппроксимируется двухзвенным регрессионным сплайном. Это свидетельствует о том,

что в интактной популяции изначально содержатся клетки, резистентные к действию цисплатина. Анализ значений спейсингов (рис. 3.9), вычисленных для этой популяции, подтверждает сделанный вывод.



Рис. 3.8. График вариационного ряда для площади клеток из интактной популяции линии KB плоскоклеточного рака ротовой полости человека

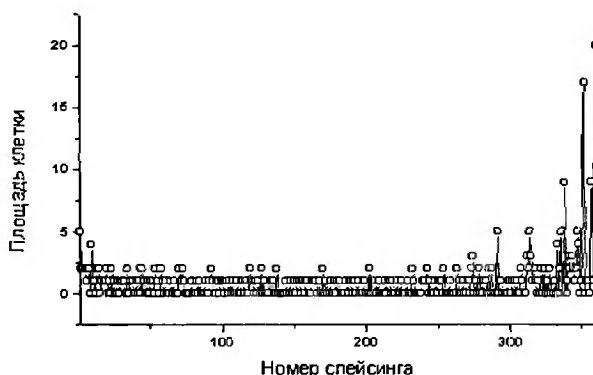


Рис. 3.9. График спейсингов для площади клеток из интактной популяции линии KB плоскоклеточного рака ротовой полости человека

График вариационного ряда, построенного для площади клеток линии KB, резистентных к цисплатину, также аппроксимируется двухзвенным сплайном (рис. 3.10). Его характер в целом совпадает с графиком, представленным на рис. 3.8, однако размах популяции увеличивается практически вдвое (с 300 до 700 единиц). Следует также подчеркнуть, что, несмотря на примерно одинаковый объем субпопуляций резистентных клеток в обеих популяциях, они сильно отличаются по размеру клеток. В то время как в интактной популяции к рези-

стентным клеткам были отнесены клетки, площадь которых превышала 225 единиц, в резистентной популяции этот порог увеличился до 350 единиц. Соответственно величина максимального спейсинга выросла с 20 до 60 (рис. 3.11).



Рис. 3.10. График вариационного ряда для площади клеток линии KB, резистентных к цисплатину

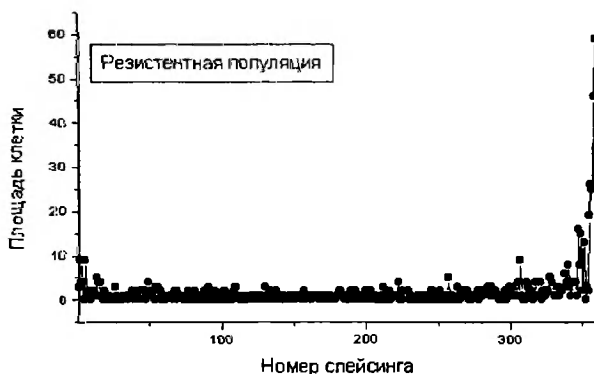


Рис. 3.11. График спейсингов для площади клеток линии KB, резистентных к цисплатину

График вариационного ряда для площади ядер из интактной популяции клеток карциномы Герена (рис. 3.12) также аппроксимируется двухзвенным сплайном, подтверждая гипотезу о том, что в интактной популяции резистентные клетки содержатся изначально. К резистентной субпопуляции относятся клетки, площадь которых превышает 60 единиц. В то же время величина максимального спейсинга не превышает 6 (рис. 3.13). График вариационного ряда для площади ядер клеток карциномы Герена, резистентных к цисплатину, аппроксимируется трехзвенным сплайном, что свидетельствует о сильной неоднородности этой популяции. Она распадается на три субпопуляции клеток, площадь которых не превышает 25, больше 25, но меньше 32, и больше 32 единиц. Размах площади этих клеток по сравнению с интактной популяцией увеличился



Рис. 3.12. График вариационного ряда для площади ядер из интактной популяции

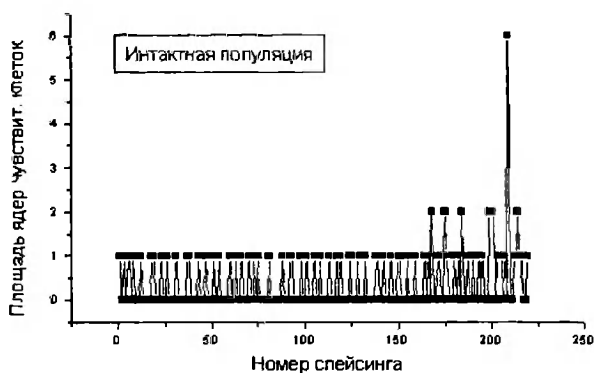


Рис. 3.13. График спейсингов для площади ядер из интактной популяции клеток карциномы Герена

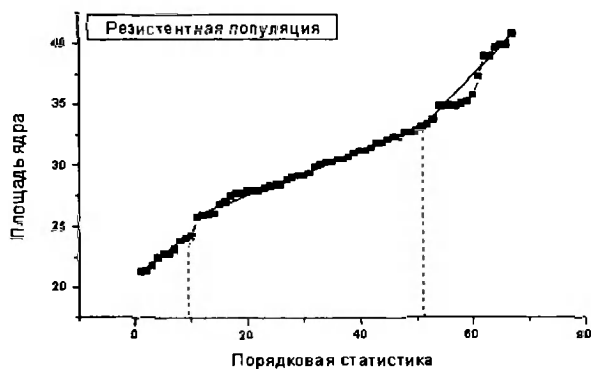


Рис. 3.14. График вариационного ряда для площади ядер клеток карциномы Герена, резистентных к цисплатину

вдвое (рис. 3.14), в то время как размах спейсингов уменьшился втрое (рис. 3.15). Это свидетельствует о внутренней однородности субпопуляций популяции резистентных клеток карциномы Герена.

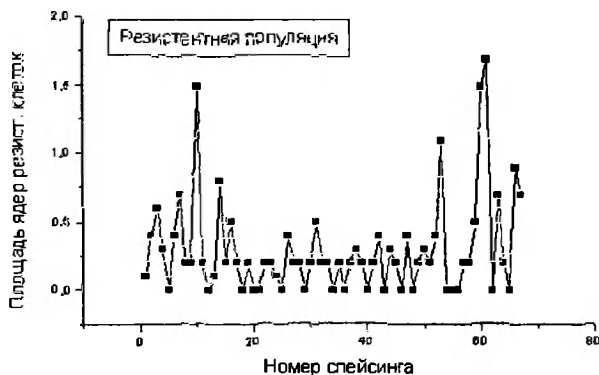


Рис. 3.15. График спейсингов для площади ядер клеток карциномы Герена, резистентных к цисплатину

Резюме

Исследование площади и периметра клеток из интактной популяции плоскоклеточного рака ротовой полости человека, а также интактной популяции клеток карциномы Герена у крыс являются гетерогенными генеральными совокупностями, состоящими из двух субпопуляций. Воздействие цисплатина на интактную популяцию раковых клеток в обоих случаях приводит к статистически значимому увеличению площади и периметра клеток и ядер, а также к увеличению размера их первого модального класса за счет смещения моды вправо. Кроме того, после воздействия цисплатина возрастает степень гетерогенности резистентной субпопуляции клеток карциномы Герена у крыс (количество модальных классов в распределении площади ядер увеличивается с двух до трех).

3.8. Количественная оценка изменчивости и гетерогенности культивируемых опухолевых клеток при формировании их радиорезистентности

Радиорезистентность (РР) опухолевых клеток — форма ответной приспособительной реакции их в виде совокупности специфических и неспецифических, обратимых и необратимых метаболических, структурных, функциональных, генетических и других изменений, возникших в определенные сроки по-

сле воздействия ионизирующей радиации. Понятие РР многогранно и его критерии многочисленны. Можно классифицировать РР по уровню биологической организации (субклеточная, клеточная и т.д.), по полноте восстановления нарушенного исходного состояния (компенсированная, субкомпенсированная, декомпенсированная), по эволюционной значимости (индивидуальная и популяционная), по пути приспособления (прогрессивная и регрессивная) и, что наиболее важно для онкологии, — по “происхождению”: генетически детерминированная или приобретенная [26, 90]. Вторичная (приобретенная) резистентность опухолевых клеток к ионизирующей радиации в результате фракционированной лучевой терапии является причиной снижения эффективности лечения онкологических больных [118].

В настоящее время сведений о механизмах возникновения РР опухолевых клеток при лучевой терапии настолько мало, что целью избранного направления исследований было не только получение новой информации, но и привлечение внимания к этому по-своему уникальному цитологическому феномену. В то же время следует отметить, что приобретенная химиорезистентность опухолей изучается весьма активно [130].

3.8.1. Материал и методы исследования

В Институте онкологии АМН Украины получен радиорезистентный штамм карциномы Герена крыс с помощью многократной ретрансплантации рецидивов опухолей после локального фракционированного облучения их в дозе 50 Гр (10 Гр × 5). Трансплантацию карциномы Герена осуществляли по общепринятой методике [68]. При достижении опухолью диаметра 1,6–2,2 см начинали курс лучевой терапии на рентгеновском аппарате РУМ-17 (180 кВ, 10 мА, фильтры — 0,5 мм Cu + 1,0 мм Al, КФР — 25 см, мощность дозы — 126 Гр/мин).

Исследования по формированию селективной РР опухолей проводили на крысах-самцах массой 110 ± 10 г разводки вивария УНИИОР МЗ Украины. На первом этапе исследований животным прививали под кожу бедра 20%-ную суспензию клеток карциномы Герена. Крыс делили на две группы (на всех этапах исследований в каждой из них было не меньше 10 животных). Опухоли крыс подопытной группы после достижения ими вышеуказанного размера подвергали локальному фракционированному рентгеновскому облучению в вышеуказанном режиме. Животных контрольной группы не подвергали никаким влияниям.

Животных с истинными рецидивами, которые возникли после окончания облучения, забивали декапитацией под легким эфирным наркозом. Опухоли после иссечения использовали для приготовления 20%-ной суспензии опухолевых клеток, которую на втором этапе ретрансплантировали двум новым группам животных — контрольной и подопытной, которую облучали по схеме, описанной выше. Третий и последующие этапы исследования полностью повторяли второй. Исследовали опухоли первых 12 генераций, применив метод культивирования в диффузионных камерах (ДК).

Для изготовления ДК применяли мембранные фильтры "Сынпор" (производство ЧССР) с диаметром пор 0,23 мкм. Фильтры монтировали на два кольца из тефлона шириной 2 мм с наружными диаметрами 18 и 16 мм (их внутренние диаметры — 16 и 14 мм соответственно). Камеры стерилизовали кипячением в дистиллированной воде в течение 40 минут с последующим высушиванием ультрафиолетовыми лучами. Опухолевую ткань измельчали в чашках Петри в стерильной среде 199 с добавлением 100 ед/мл пенициллина и 100 ед/мл стрептомицина. Забор материала производился иглой Кассирского И-11 в модификации, заключающейся в специальной заточке конца иглы, превращающей его в усеченный конус. Получаемая стандартизация количества опухолевого материала существенна для исследований, в которых становятся объективно сопоставимыми величины цитологических показателей зоны миграции клеток эксплантатов.

Полученный эксплантат помещали на увлажненный стерильной средой 199 фильтр большей половины камеры, закрывали его меньшей половиной, линию соприкосновения большого и малого колец герметизировали клеем БФ-6. Для инкубации эксплантатов по две ДК в асептических условиях имплантировали в брюшную полость крысы. На каждую точку исследования анализировали не меньше 12 ДК. Для определения оптимальных сроков пролиферации опухолевых клеток, удобных для регистрации зоны их миграции, был поставлен специальный эксперимент для изучения динамики роста карциномы Герена в ДК с первых по 11-е сутки. Каждый день извлекали из брюшной полости у нескольких крыс не меньше шести ДК, фиксировали, окрашивали гематоксилином Караччи, проводили радиоавтографическую обработку препаратов по общепринятой методике [81].

Кусочки ткани опухоли для культивирования брали на третьи сутки после окончания облучения. На седьмые сутки инкубации ДК извлекали из брюшной полости крыс, фиксировали, окрашивали, проводили качественное и количественное изучение морфологических изменений в монослое клеток.

При морфологическом анализе препаратов (для каждого случая не менее чем 2000 клеток) определяли: количество клеток, находящихся в зоне роста вокруг эксплантата; количество митотически делящихся клеток, митотический индекс (МИ), количество аберрантных митозов и их процент, количество меченых клеток, индекс метки (ИМ), интенсивность синтеза ДНК в каждом ядре. ДНК-синтезирующую функцию клеток оценивали с помощью метода радиоавтографии с использованием ^3H -тимидина. Для этого животным за 1 ч до забоя и извлечения ДК вводили внутрибрюшинно ^3H -тимидин в дозе 37 МБк/кг. Удельная активность ^3H колебалась в пределах 340–370 МБк в 1 мл раствора.

Наряду с вычислением процента патологических митозов, дифференцировали их типы [50]: первый тип — патологические митозы, связанные с повреждением хромосом (хромосомные и хроматидные мосты, отставание хромосом в анафазе, слипание хромосом); второй тип — патологические митозы, связанные с повреждением митотического аппарата (асимметрический митоз, полая метафаза, многополюсный митоз, рассеивание хромосом в метафазе, моноцен-

трический митоз, рассеивание ацентрических фрагментов вокруг митотического веретена).

Количественные морфометрические исследования проводили на цитоанализаторе "Интеграл-2 МТ" (Киев) в режиме "контур" с измерением на каждую точку 100 клеток. Из измеренных показателей (площадь ядра, периметр ядра, площадь и периметр клетки, коэффициент формы контура ядра и клетки, соотношение площадей и периметров ядер и клеток) использован один из наиболее информативных параметров — площадь ядра.

Для определения взаимосвязи изменчивости и гетерогенности популяции клеток со степенью РР выбраны значения площади ядер клеток в 3-й, 5-й, 6-й, 7-й и 12-й генерациях.

Оценка изменчивости и гетерогенности популяций клеток осуществлялась с помощью вычисления меры близости между выборками и анализа спейсинг-кривых, соответствующих различным генерациям [13]. Клеточная популяция является частным случаем абстрактной генеральной совокупности, которая изучается в современной математической статистике и представляет собой произвольное множество значений G количественных характеристик некоторых объектов с заданным в множестве G распределением вероятностей $F(u)$, возможно, неизвестным экспериментатору. Как правило, распределение гетерогенной генеральной совокупности G является полимодальным, однако выборка $x = (x_1, x_2, \dots, x_n) \in G$ из гетерогенной генеральной совокупности может представлять в простейшем случае объединение двух выборок $x^{(1)} = (x_1, \dots, x_k) \in G_1$ и $x^{(2)} = (x_{k+1}, \dots, x_n) \in G_2$ из двух различных генеральных совокупностей G_1 и G_2 (такие выборки будем называть *составными*) или смесь двух выборок из этих совокупностей G_1 и G_2 (такие выборки будем называть *смешанными*). Для распознавания типа гетерогенности выборки мы применили два метода: для составных выборок — метод разбиения выборки на две части и вычисления меры близости между ними, а для смешанных — метод анализа спейсинг-кривых, соответствующих кривым Кетле. При вычислении степени гетерогенности выборок использовалась гипотеза о двухмодальности популяции радиорезистентных клеток.

Идентификация составных выборок. Для идентификации и определения степени гетерогенности (СГ) составной выборки $X = \{x_1, x_2, \dots, x_n\}$ мы вычисляли меру близости $\rho(X_1, X_2)$ [13] между двумя ее половинами $X_1 = \{x_1, x_2, \dots, x_k\}$ и $X_2 = \{x_{k+1}, x_{k+2}, \dots, x_n\}$, где $k = \left\lfloor \frac{n}{2} \right\rfloor$ — целая часть числа $\frac{n}{2}$. Тогда степень гетерогенности выборки X определяется формулой $\mu(X_1, X_2) = 1 - \rho(X_1, X_2)$. Мера близости $\rho(X_1, X_2)$ принимает значения от 0 до 1. Если мера близости $\rho(X_1, X_2)$ мала, то степень гетерогенности $\mu(X_1, X_2)$ близка к единице, и выборка в этом случае идентифицируется как составная. В дальнейшем будем придерживаться следующей классификации составных популяций: $0 < \mu(X_1, X_2) \leq 0,3$ — слабо гетерогенная, $0,3 < \mu(X_1, X_2) \leq 0,7$ — умеренно гетерогенная, $0,7 < \mu(X_1, X_2) \leq 1$ — сильно гетерогенная составная популяция. Заметим, что если $\mu(X_1, X_2)$ близка к нулю, то выборка может быть как однородной, так и смешанной. В таком

случае необходимо применять метод анализа спейсинг-кривых для идентификации смешанных выборок.

Идентификация смешанных выборок. Пусть $x = (x_1, \dots, x_n) \in G$ — выборка из генеральной совокупности G и $x_{(1)} \leq \dots \leq x_{(n)}$ — ее вариационный ряд. Рассмотрим на плоскости декартову прямоугольную систему координат и точки $(i, x_{(i)})$, $i = 1, 2, \dots, n$, проведем через эти точки плавную кривую (для определенности соединим соседние точки прямолинейными отрезками) и будем называть эту совокупность точек графиком вариационного ряда. Эта кривая в классической математической статистике называется кривой Кетле (рис. 3.16, 3.18, 3.20). Если генеральная совокупность является однородной, то центральная часть кривой Кетле хорошо аппроксимируется прямолинейным отрезком. Поэтому, когда центральная часть кривой Кетле аппроксимируется двумя отрезками, порождающими ломаную, состоящую из двух звеньев (первый случай), или двумя несвязными отрезками (второй случай), мы имеем дело с неоднородной (гетерогенной) генеральной совокупностью. В любом случае правый конец первого отрезка мы будем называть *особой точкой* кривой Кетле; в первом случае она является *точкой излома*, а во втором — *точкой разрыва* кривой Кетле. Алгоритм обнаружения и вычисления координат особой точки кривой Кетле состоит в следующем. В начале по выборке $x = (x_1, \dots, x_n) \in G$ построим вариационный ряд $x_{(1)} \leq \dots \leq x_{(n)}$ и отбросим 10% его крайних порядковых статистик (5% первых минимальных порядковых статистик и 5% последних максимальных). Затем по оставшимся порядковым статистикам строим кривую, образованную приращением ординат вершин смежных звеньев кривой Кетле (i, y_i) , где $y_i = x_{(i+1)} - x_{(i)}$, $i = 1, 2, \dots, n - l - 1$, соединяя смежные вершины (i, y_i) , $(i + 1, y_{i+1})$ прямолинейными отрезками. Величины y_i носят название спейсингов (spacing), поэтому мы будем называть построенную таким образом кривую спейсинг-кривой (рис. 3.17, 3.19, 3.21). Для определения особой точки кривой Кетле определяется интервал, в котором расположены начальные точки спейсинг-кривой (рис. 3.16, 3.18, 3.20). Первая точка выхода за границы этого интервала и будет особой точкой, являясь в то же время границей между субпопуляциями. В предположении, что популяция клеток состоит из двух субпопуляций X_1 и X_2 , ее *степенью гетерогенности* $\lambda(X_1, X_2)$ назовем отношение объема меньшей субпопуляции к объему большей. Очевидно, что наиболее гетерогенной является популяция, образованная двумя субпопуляциями равного объема ($\lambda(X_1, X_2) = 1$), а абсолютно однородная популяция имеет нулевую степень гетерогенности ($\lambda(X_1, X_2) = 0$). Введем следующую классификацию гетерогенности смешанных популяций: $0 < \lambda(X_1, X_2) \leq 0,3$ — слабо гетерогенная, $0,3 < \lambda(X_1, X_2) \leq 0,7$ — умеренно гетерогенная, $0,7 < \lambda(X_1, X_2) \leq 1$ — сильно гетерогенная смешанная популяция.

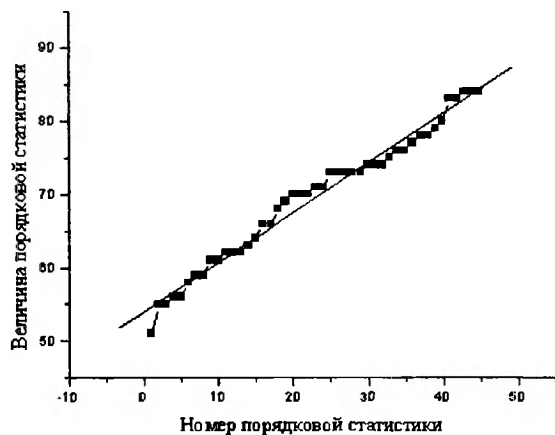


Рис. 3.16. Кривая Кетле абсолютно однородной контрольной популяции

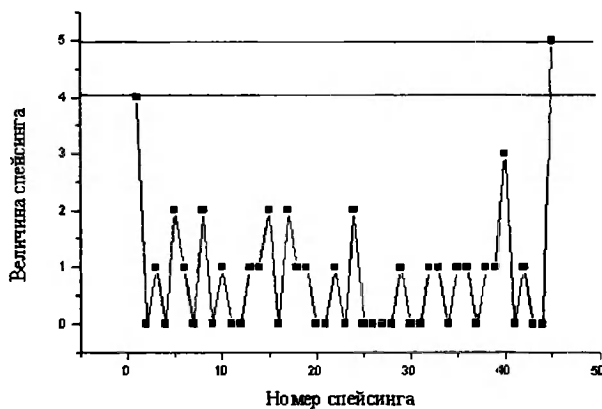


Рис. 3.17. Спейсинг-кривая абсолютно однородной контрольной популяции

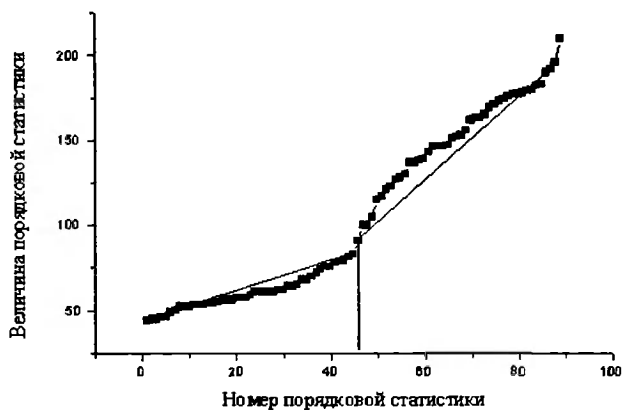


Рис. 3.18. Кривая Кетле составной сильно гетерогенной популяции (5-я генерация)

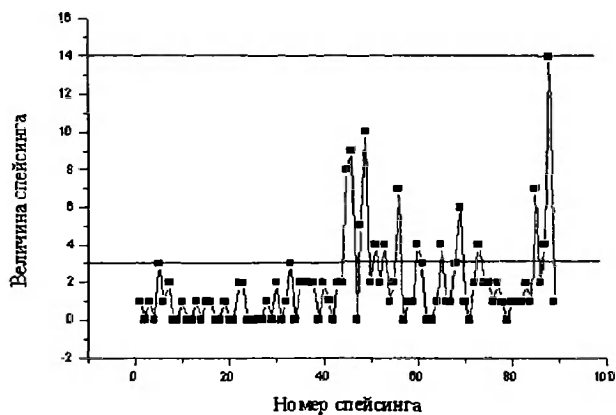


Рис. 3.19. Спейсинг-кривая составной сильно гетерогенной популяции (5-я генерация)



Рис. 3.20. Кривая Кетле смешанной слабо гетерогенной популяции (6-я генерация)

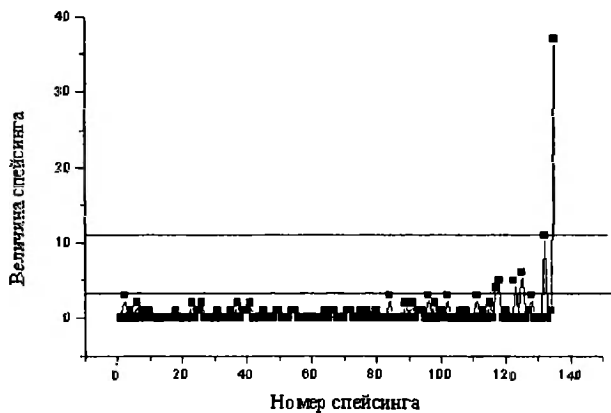


Рис. 3.21. Спейсинг-кривая смешанной слабо гетерогенной популяции (6-я генерация)

3.8.2. Результаты и их обсуждение

Морфометрические особенности радиорезистентных клеток. Одним из механизмов формирования приобретенной РР на тканевом уровне (по нашим предварительным гистологическим исследованиям) является перестройка клеточных взаимоотношений, главным образом, лимфоцитов и фибробластов с опухолевыми клетками как ответная реакция на облучение. В процессе выработки РР отличительной чертой патоморфоза была тенденция к дистрофическим изменениям в опухолевых клетках в отличие от деструктивных, ведущих к некрозу в классическом варианте.

Ранее нами были изучены особенности роста и изменения структуры карциномы Герена (*in vivo*), которые возникают в процессе выработки ее резистентности к действию ионизирующей радиации [68]. Установлено, что среди популяции клеток карциномы Герена существуют не меньше двух субпопуляций клеток, которые существенно отличаются как по характеру роста, так и по чувствительности к действию радиации. В процессе формирования РР соотношение этих субпопуляций изменяется в направлении уменьшения чувствительной к облучению субпопуляции и увеличения резистентной. Эти изменения согласуются с положениями клонально-селекционной теории резистентности и свидетельствуют об обратимом характере резистентности, приобретенной под влиянием радиации [68].

Усиление устойчивости к действию физического или химического фактора может быть достигнуто с помощью многократного ступенчатого отбора, при котором уровень селектирующего агента в каждом цикле отбора повышается. В результате этого нам удалось наблюдать приспособление клеточных популяций к рентгеновскому излучению. В каждой новой генерации адаптация популяций достигается за счет накопленного в ней резерва наследственной изменчивости. При адаптации на основе многократного ступенчатого отбора возникают новые наследственные изменения.

В клеточных популяциях важнейшие параметры структуры клеток обусловлены двумя основными процессами — возникновением наследственных изменений в соматических клетках и размножением клонов, наследственно различающихся между собой, т.е. взаимодействием процессов изменчивости и отбора.

Наследственная изменчивость в популяциях включает как генотипические (изменения количества, качества или порядка расположения авторепродуцирующих матриц клетки — вне зависимости от ядерной или цитоплазматической их локализации), так и эпигенетические изменения, обусловленные блокированием и деблокированием генетического материала (экспрессией, репрессией и дерепрессией генов). Генотипическая изменчивость связана с возникновением генных и хромосомных мутаций. Общей причиной мутаций любого типа в соматических клетках является нарушение внутриклеточного гомеостаза. Количественные признаки соматических клеток в своем фенотипическом проявлении никогда не бывают обусловлены лишь наследственными факторами, реализация признака всегда происходит за счет взаимодействия наследственной основы и среды [43].

При формировании устойчивости живых объектов как к любым воздействиям (адаптации), так и к лучевым увеличивается гетерогенность их клеточного

состава [33, 56]. Изучение особенностей реакции опухолевых клеток на облучение составило цель данного раздела.

Изучив кинетические параметры размножения клеток карциномы Герена в контроле в зоне роста вокруг эксплантата с 1-х по 11-е сутки, мы получили доказательство того, что оптимальным сроком для исследований являются 7-е сутки после эксплантации, так как именно на этот срок приходится максимальная пролиферация клеток.

Наиболее характерной особенностью изменений, связанных с формированием РР, явилась гетерогенность клеточного состава внутри каждой генерации, выраженная полиморфизмом, большим количеством многоядерных (зачастую с нечетным числом ядер) и особенно двуядерных клеток, в том числе неправильной формы, среди которых часто наблюдался анизоцитоз. Вероятно, они являются продуктом анеуплоидии, патологических эндомитозов или слияния трансформированных клеток.

Вариабельность и полиморфизм клеток и ядер выражались, главным образом, в изменении их размеров и формы, положением и структурой ядра, его гиперхроматозом, изменениями формы ядерной и клеточной поверхности и увеличенным количеством цитоплазматических включений. Параметры изменения размеров клеток явились наиболее выраженным морфологическим критерием при анализе особенностей формирования РР.

По данным морфометрических исследований площадей ядер модальный класс в контрольных опухолях составляет 76%. Два других варианта всего на 30 мкм² отклоняются от него вправо (4%) и влево (20%), что свидетельствует о мономорфности карциномы Герена и доминировании клеток с ядрами в узком диапазоне площадей — от 60 до 120 мкм².

Результаты анализа функциональной активности клеток карциномы Герена представлены в табл. 3.28. Отметим, что, несмотря на неоднократные последовательные облучения рецидивирующих опухолей, величина цитокинетических показателей остается приблизительно одинаковой, что свидетельствует о сохраняющейся жизнеспособности клеток этих опухолей.

Таблица 3.28 . Изменение параметров пролиферативной активности клеток карциномы Герена в динамике формирования РР

Номер генерации опухоли	Параметры пролиферации			
	МИ, %	Аберрантные митозы, %	ИМ, %	ИС ДНК ($M \pm m$)
Контроль	4,8	0,2	14,4	28,8 $\pm$ 2,1
3	3,6	1,0	14,4	24,8 $\pm$ 2,6
5	3,6	0,8	14,4	26,0 $\pm$ 3,2
6	3,8	1,0	13,8	24,6 $\pm$ 2,4
7	3,4	1,0	13,4	28,6 $\pm$ 3,8
12	3,4	1,4	13,4	29,0 $\pm$ 3,2

При формировании РР количество патологических митозов увеличивается в 4–7 раз, что свидетельствует о формировании определенных признаков в ходе адаптивно-приспособительной реакции опухолевых клеток на радиацию или о разделении клеточной популяции на потенциально нежизнеспособные клетки и фракции, сохраняющиеся после воздействия, для селективного отбора и создания фонда РР.

В 3-й генерации впервые наблюдался рост эксплантата, что было характерно и для других, особенно 7-й, генераций. В этих случаях размножение клеток осуществлялось двумя широкими ростками во взаимно перпендикулярных направлениях. Подобное явление наблюдали только в экспериментах, связанных с РР.

В препаратах опухолей 5-й генерации типичным было сближение ядер клеток до такой степени, что образовывался как бы плотный монослой ядер, находящихся в состоянии конфлюента. Однако ядерные мембраны оставались четкими, поэтому морфометрические исследования были адекватными.

Явление анизокариоза наблюдалось на всех этапах эксперимента, но наиболее выраженным оно было в опухолях 12-й генерации. Хотя число двуядерных клеток с ядрами разного размера в опухолях 12-й генерации наблюдалось до 4%, не отмечено ни одного ядра с неравномерным синтезом ДНК (в отличие от данных, полученных в гистологических исследованиях). ДНК-синтезирующая функция клеток при формировании РР находилась на уровне контроля (см. табл. 3.28).

Изменчивость генераций. Результаты исследования изменчивости популяций клеток с помощью метода, основанного на вычислении меры близости между выборками, образованными значениями площадей ядер клеток в контроле и различных генерациях, представлены в табл. 3.29.

Таблица 3.29. Величины меры близости между выборками, образованными значениями площадей ядер клеток в контроле, 3-й, 5-й, 6-й, 7-й и 12-й генерациях опухолей

	Номер генерации опухоли					
	Контроль	3	5	6	7	12
Контроль	1,0000	0,4484	0,4448	0,4202	0,4302	0,3774
3		1,0000	0,5164	0,8153	0,4059	0,6676
5			1,0000	0,3443	0,4809	0,7728
6				1,0000	0,2278	0,4038
7					1,0000	0,2825
12						1,0000

Анализ данных табл. 3.29 показывает, что для исследованных популяций клеток характерна тенденция убывания меры близости с контрольной выборкой при возрастании номера генерации опухоли, что свидетельствует об увеличении различия популяций клеток (т.е. их изменчивости) по сравнению с контролем. Однако стабилизации изменений не происходит, поскольку мера бли-

зости между 6-й, 7-й и 12-й генерациями изменяется нерегулярным образом и не является монотонной функцией.

Гетерогенность генераций. Оценка гетерогенности клеток (популяций) на основании исследования площадей их ядер в контроле и различных генерациях опухолей проводилась методами идентификации составных и смешанных выборок [13] (табл. 3.30).

Таблица 3.30. Результаты идентификации и оценки гетерогенности составных выборок разных генераций опухолей

Номер генерации опухоли	Показатели гетерогенности клеточных популяций					
	$\rho(X_1, X_2)$	НГ	ВГ	$\mu(X_1, X_2)$	НГ	ВГ
Контроль	1,000	0,944	1,000	0,000	0,000	0,056
3	0,974	0,950	0,987	0,026	0,013	0,050
5	0,249	0,190	0,318	0,751	0,682	0,810
6	0,998	0,991	0,999	0,002	0,001	0,009
7	0,493	0,419	0,566	0,507	0,434	0,581
12	0,441	0,409	0,475	0,559	0,525	0,591

Примечание. НГ и ВГ — нижняя и верхняя доверительная граница, соответствующая 5%-ному уровню значимости.

Из табл. 3.30 следует, что выборки из 5-й, 7-й и 12-й генерациях являются составными ($\mu(X_1, X_2) > 0,05$), причем опухоли 5-й генерации являются сильно гетерогенными, а 7-й и 12-й — умеренно гетерогенными.

Данные, относящиеся к контролю, опухолям 3-й и 6-й генераций составными не являются и в дальнейшем обрабатывались методом идентификации смешанных выборок. При этом на основе анализа графиков кривых Кетле и спейсинг-кривых (см. рис. 3.12–3.15) были вычислены особые точки, количество клеток в популяциях (модальных классах), составляющих генерацию, и пределы изменения площади ядер клеток из этих популяций (табл. 3.31).

Таблица 3.31. Оценка гетерогенности популяций клеток с помощью спейсинг-кривой

Номер генерации опухоли	Количество клеток	Особая точка	Первая популяция			Вторая популяция			$\lambda(X_1, X_2)$
			Количество клеток	$S_{\text{нп}}$	$S_{\text{вп}}$	Количество клеток	$S_{\text{нп}}$	$S_{\text{вп}}$	
Контроль	45	Нет	45	55	85	-	-	-	0,00
3	89	83	83	30	125	6	125	225	0,13
5	89	44	44	30	80	45	100	225	0,98
6	135	118	118	65	125	17	125	160	0,14
7	89	69	69	30	60	20	60	90	0,29
12	135	110	110	25	125	25	125	200	0,22

Примечание. При исследовании гетерогенности мы отбрасывали 5% первых минимальных и 5% последних максимальных порядковых статистик для определения основной распределенной массы значений генеральной совокупности. Кроме того, выборка, составляющая менее чем 5% от общего числа клеток, не рассматривалась как отдельная популяция; $S_{\min}$ и $S_{\max}$ — минимальная и максимальная площадь ядра в популяции.

Из табл. 3.31 следует, что контроль является абсолютно однородной популяцией ($\lambda(X, X) = 0$), а опухоли 3-й и 6-й генерации — слабо гетерогенные смешанные популяции. Остальные выборки были идентифицированы как составные.

Анализ спейсинг-кривых без ограничения модальности ядер клеток популяций указывает на то, что выборки из 7-й и 12-й генераций опухолей образованы из трех модальных классов, которые составлены из мелкоядерных, среднеядерных и крупноядерных клеток соответственно. При этом на кривых Кетле для 7-й генерации имеются две особые точки, соответствующие площадям ядер 39 и 59 мкм², а для 12-й генерации — 65 и 137 мкм². Популяции клеток, образованные первым и третьим модальными классами, имеют площади ядер, сильно отличающиеся от площадей ядер контрольной выборки. Таким образом, можно выдвинуть гипотезу, что, начиная с 7-й генерации опухолей, в популяции облученных клеток возникают две новые субпопуляции радиорезистентных клеток, образованные мелкоядерными и крупноядерными клетками.

Для оценки объема модальных классов, элементы которых лежат за пределами интервала, образованного минимальной и максимальной порядковыми статистиками контрольной выборки, в каждой генерации подсчитывалось количество клеток, площадь ядра которых меньше минимальной порядковой статистики (мелкоядерные клетки) или больше максимальной (крупноядерные клетки). Из результатов, представленных в табл. 3.32, следует, что процент среднеядерных клеток изменяется в достаточно узких пределах с тенденцией к убыванию, в то время как количество мелкоядерных и крупноядерных клеток меняется в значительно более широких пределах нерегулярным образом (в некоторых генерациях мелкоядерные клетки или крупноядерные клетки могут вообще отсутствовать).

Таблица 3.32. Размеры модальных классов по площади ядер в популяциях РР клеток

Номер генерации опухоли	Общее количество клеток	Количество мелкоядерных клеток	Количество среднеядерных клеток	Количество крупноядерных клеток	Доля мелкоядерных клеток	Доля среднеядерных клеток	Доля крупноядерных клеток
Контроль	45	0	45	0	0,00	1,00	0,00
3	89	13	32	44	0,15	0,36	0,49
5	89	7	38	44	0,08	0,43	0,49
6	135	0	40	95	0,00	0,30	0,70
7	89	55	34	0	0,62	0,38	0,00
12	135	30	30	75	0,22	0,22	0,56

Резюме

Наиболее характерной особенностью, связанной с формированием РР, тестируемой на цитологическом уровне при использовании метода ДК, явилась гетерогенность клеточного состава внутри каждой генерации опухолей. Гетерогенность и полиморфизм клеток и ядер выражались, главным образом, изменением их размеров и формы, положением и структурой ядра, его гиперхроматозом, изменениями формы ядерной и клеточной поверхности.

Параметры изменения размеров ядер клеток оказались наиболее выраженным морфологическим критерием при анализе особенностей формирования РР. С помощью статистического анализа удалось объективно оценить изменчивость между генерациями и гетерогенность клеточного состава в каждой генерации опухолей, в зависимости от варибельности площадей ядер в клетках.

Таким образом, на основе количественных данных мы установили, что все генерации опухолей, подвергшихся рентгеновскому облучению, являются гетерогенными по своему клеточному составу в отличие от гомогенной популяции клеток карциномы Герена в контроле. При этом наблюдается стабилизация показателей гетерогенности, начиная с 6-й генерации опухолей, а также соотношения количества мелко- и крупноядерных клеток с преобладанием относительно мелкоядерных.

3.9. Новые калибровочные кривые для дозовых зависимостей радиационно-индуцированных хромосомных aberrаций в культуре лимфоцитов человека

3.9.1. Введение

Анализ зависимостей “доза–эффект” является одним из основных средств исследования в количественной радиобиологии. Наиболее популярной является так называемая линейно-квадратичная модель, разработанная на основе микродозиметрических концепций, которые А. Келлерер и Г. Росси [11] сформулировали в их известной теории дуального действия. В рамках этой модели тот факт, что экспериментальные данные хорошо приближаются линейными или квадратичными функциями, объясняется тем, что энергия, поглощаемая мишенью, может быть получена в результате как одного, так и двух столкновений. В связи с этим дозовые кривые описываются уравнением $y = aD^2 + bD + c$, где y — количество хромосомных aberrаций, D — поглощенная доза, a , b , c — коэффициенты, отражающие относительную роль этих двух механизмов. Величина $\lambda = a/b$ соответствует той дозе, при которой роль одно- и двухударных компонентов одинакова. В отличие от редко ионизирующих излучений, для ко-

торых характерна линейно-квадратичная зависимость эффекта от дозы, при действии плотно ионизирующих излучений суммарный выход аберраций хромосом как и количество двухударных аберраций (хромосомные обмены), в некоторых случаях хорошо описываются линейной зависимостью от дозы излучения [59, 62, 93, 113]. Некоторые исследователи, исходя из принципа “попадания в мишень” (Н.В. Тимофеев-Ресовский, К.Г. Циммер), полагают, что с точки зрения механизма образования хромосомных аберраций более обоснованной является линейно-квадратичная модель [3, 14, 119]. Тем не менее в некоторых случаях как линейные, так и линейно-квадратичные зависимости “доза-эффект” не дают удовлетворительного приближения истинных функциональных зависимостей [115]. Одной из задач диагностики радиационных поражений является определение полученной организмом дозы облучения по изменениям биологических показателей — так называемая *биологическая дозиметрия*. Наиболее адекватно методам биологической дозиметрии соответствует подсчет аберраций хромосом в лимфоцитах периферической крови, причем требования, предъявляемые к нему, достаточно строги: погрешность в определении дозы не должна превышать 5% при дозах облучения, дающих острые эффекты, т.е. свыше 1 Гр [75].

В связи с этим целью этого раздела является определение новых калибровочных кривых, вычисленных после статистической обработки экспериментальных данных с помощью кусочно-линейных сплайнов.

3.9.2. Материал и методы исследования

Эксперименты выполнены в Институте онкологии АМН Украины по модифицированной методике культивирования лимфоцитов периферической крови человека, предложенной в [128]. Полученные препараты анализировали из расчета 200–300 клеток на экспериментальную точку. Анализировались метафазы, содержащие 45–47 хромосом.

Опыты с редко ионизирующим излучением выполнены на установке “Луч” с источником гамма-излучения ^{60}Co мощностью 0,5 Гр/мин. Исследуемый диапазон доз составляет 0,25–4,0 Гр. Измерение поглощенных доз на гамма-установке проводилось с помощью ионизационной камеры и ферросульфатного дозиметра. Облучение выполнено на G_0 -стадии митотического цикла лимфоцитов периферической крови человека в культуре.

В качестве плотно ионизирующего излучения использованы моноэнергетические нейтроны с энергией 6,0 МэВ мощностью 2,8 Гр/ч, получаемые на ускорителе КГ-2.5. Вклад гамма-компонента составлял не более чем 10%. Исследуемый диапазон доз составлял 0,25–2,0 Гр. Дозиметрия осуществлялась с помощью термолуминесцентного, ферросульфатного, трекового и расчетного методов. Облучение выполнено также в G_0 -стадии митотического цикла. Кроме того, культуру лимфоцитов облучали нейтронами со средней энергией 22 МэВ на циклотроне У-240. Вклад гамма-компонента не превышал 5%. Диапазон доз составлял 0,1–4,0 Гр.

Для аппроксимации дозовых зависимостей количества радиационно-индуцированных аберраций определенного вида (дицентрики, транслокации и т.д.) или общего количества всех хромосомных аберраций были использованы:

- линейные функции $y = aD + b$;
- линейно-квадратичные функции $y = aD^2 + bD + c$;
- кусочно-линейные сплайны с одной точкой перехода D_0 :

$$y = \begin{cases} a_1 D + b_1, & \text{если } D \leq D_0, \\ a_2 D + b_2, & \text{если } D > D_0. \end{cases}$$

Вычисление параметров линейной и линейно-квадратичной моделей проводилось стандартными методами расчета параметров в схеме линейной регрессии с помощью метода наименьших квадратов, а для вычисления параметров сплайнов и точки перехода был разработан специальный статистический метод нелинейной регрессии, основанный на методе наименьших квадратов и порядковых статистиках.

3.9.3. Результаты исследований и их обсуждение

Найденные значения параметров линейных, линейно-квадратичных и кусочно-линейных сплайновых дозовых зависимостей различных цитогенетических показателей для разных видов облучения приведены в табл. 3.33. Анализ результатов показывает, что при гамма-облучении кусочно-линейные сплайны имеют меньшую среднеквадратическую ошибку по сравнению с линейной и линейно-квадратичными моделями в четырех из шести случаев. При плотно ионизирующем облучении быстрыми нейтронами с энергией 6 МэВ кусочно-линейные сплайны имеют меньшую среднеквадратическую ошибку по сравнению с линейной и линейно-квадратичной моделями в пяти из шести случаев, а при облучении быстрыми нейтронами с энергией 22 МэВ — во всех случаях. Тем не менее следует заметить, что для облучения быстрыми нейтронами с энергией 6 МэВ различия между среднеквадратическими ошибками всех трех моделей в четырех случаях из шести крайне незначительны. Заметное преимущество модель, основанная на кусочно-линейных сплайнах, имела в пяти случаях из шести при облучении быстрыми нейтронами с энергией 22 МэВ.

Общепризнанно, что наиболее информативными цитогенетическими дозиметрами являются дицентрики [106]. Из табл. 3.33 следует, что по сравнению с линейной и кусочно-линейной моделями зависимость количества дицентриков от дозы наилучшим образом аппроксимировались кусочно-линейными сплайнами.

Таблица 3.33. Значения параметров линейных, линейно-квадратичных и кусочно-линейных сплайновых дозовых зависимостей различных цитогенетических показателей для разных видов облучения

ВО	ЦП	Среднеквадратичные отклонения			Линейная модель		Линейно-квадратичная модель			D0, Гр	Параметры до точки перехода		Параметры после точки перехода	
		Л	ЛК	С	а	б	а	б	с		а,	б,	а,	б,
I	1	90,87	81,98	55,94	23,2189	-2,2831	0,7952	20,0589	0,7624	0,50	8,7854	2,2000	24,3670	-5,5908
	2	2750,93	37,32	124,36	56,9200	-24,0267	13,9004	1,6840	2,5550	2,00	27,5882	-2,9843	88,1647	-124,1372
	3	8,58	5,12	3,24	1,2252	0,4891	12,4576	5,0658	2,7725	2,00	-9,7059	1,4377	2,6338	-4,0240
	4	2205,76	26,27	120,79	54,5684	-23,7947	12,4576	5,0658	2,7725	2,00	28,4353	-5,0407	82,4058	-112,9882
	5	139,85	19,11	1,81	12,7963	-8,9323	3,9626	-4,8906	4,2362	3,00	8,2576	-3,3186	27,8458	-62,0830
	6	331,43	84,45	16,73	32,9914	-22,4420	5,6675	7,6949	-3,6079	2,00	18,6987	-7,3891	43,6349	-57,2614
II	1	4,33	2,08	4,14	28,8873	4,5872	12,1157	49,1036	-1,9843	0,64	37,5648	0,5518	25,6355	8,1865
	2	3,65	3,49	3,264	45,2435	2,7582	-2,3599	49,1780	1,4792	0,64	47,8627	1,5401	44,2621	3,8445
	3	2,35	1,86	1,61	0,8691	4,6546	4,0719	-5,9196	6,8614	0,64	-2,7202	6,3238	2,2142	3,1638
	4	11,68	10,46	9,48	44,3744	-1,8965	-6,4321	55,0981	-5,3823	0,64	50,5829	-4,7836	42,0478	0,6787
	5	6,74	6,64	6,21	13,7345	-1,9081	-1,9090	16,9173	-2,9427	0,50	8,7038	0,2148	14,4564	-2,6615
	6	0,42	0,24	0,08	21,6810	0,1485	2,4965	17,5188	1,5015	0,50	17,6261	1,8596	22,2629	-0,4587
III	1	601,54	318,31	144,68	20,1568	13,1328	-4,1165	36,3805	4,8484	1,00	44,9504	0,2505	13,6978	31,5031
	2	752,12	751,37	710,67	55,9600	2,5741	0,2128	55,1213	3,0024	1,00	63,4284	-1,3063	54,0145	8,1076
	3	36,24	28,60	6,96	-0,9020	5,4651	-1,5667	3,9561	3,8930	0,50	16,5220	0,3945	-2,8682	10,0897
	4	435,93	317,61	222,53	57,2026	-7,4066	5,3215	41,4535	-1,9254	0,50	10,6640	6,2414	62,6981	-19,7756

Окончание табл. 3.33

ВО	ЦП	Среднеквадратичные отклонения			Линейная модель		Линейно-квадратичная модель			D0, Гр	Параметры до точки перехода		Параметры после точки перехода	
		Л	ЛК	С	а	б	а	б	с		а ₁	б ₁	а ₂	б ₂
5		658,53	614,59	315,83	11,3448	0,7363	-1,6214	17,7350	-2,5267	3,00	16,4886	-3,6685	-11,7972	81,1890
6		872,79	769,30	609,62	30,7208	-7,5002	2,4883	20,9142	-2,4926	3,00	26,2133	-3,6402	51,0001	-78,0006

Примечание. Здесь использованы следующие обозначения: ВО — вид облучения; I — гамма-облучения; II — быстрые нейтроны (6 МэВ); III — быстрые нейтроны (22 МэВ); ЦП — цитогенетические показатели; 1 — количество aberrантных метафаз (%); 2 — общее количество aberrаций хромосом (%); 3 — общее количество aberrаций хроматидного типа (%); 4 — общее количество хромосомных aberrаций (%); 5 — количество фрагментов (%); 6 — количество дисцентриков (%); Л — среднеквадратическая ошибка линейной регрессии; ЛК — среднеквадратическая ошибка линейно-квадратичной регрессии; С — среднеквадратическая ошибка кусочно-линейных сплайнов; а, б и с — параметры линейной и линейно-квадратичной регрессии; а₁, б₁, а₂, б₂ — параметры кусочно-линейных сплайнов.

Резюме

Обнаружены расхождения между полученными экспериментальными данными и общепринятыми линейной и линейно-квадратичной моделями зависимости “доза–эффект” для количества хромосомных aberrаций разного вида. Использование предложенного метода аппроксимации экспериментальных зависимостей “доза–эффект” на основе кусочно-линейных сплайнов приводит к более точным результатам в подавляющем большинстве случаев и с этой точки зрения предпочтительнее, чем линейная и линейно-квадратичная модели.

3.10. Диагностика гастродуоденальных заболеваний методами статистической теории распознавания

3.10.1. Метод дуоденального зондирования для исследования функционального состояния органов желудочного тракта

Дуоденальное зондирование существует уже почти 100 лет, но отношение клиницистов к этому методу до сих пор неоднозначное. Катетеризацию привратника и двенадцатиперстной кишки впервые осуществили Einhorn (1909) и Gross (1910); они же дали описание тонкого зонда, который получил название зонда Эйнгорна. Он представлял собой тонкую резиновую трубку с металлическим или деревянным наконечником с отверстиями на нижнем конце. Со временем этот зонд был усовершенствован и наиболее широкое применение в клинической практике получил зонд Огрена–Лагерлефа (1936), который представлял собой двойной плоский зонд с несколькими отверстиями для желудочного и дуоденального содержимого. Новым этапом в развития дуоденального зондирования стал метод многомоментного фракционного дуоденального зондирования (БФДЗ) (Lopez, Fuentes, Prado, 1950), с помощью которого можно сделать вывод о функциональном состоянии желчного пузыря и сфинктерного аппарата желчевыводных путей. Все это позволило значительно расширить диагностические возможности дуоденального зондирования.

Параллельно с исследованием функций желчного пузыря осуществлялось изучение функции поджелудочной железы. И.Л. Долинский (1894) и Л.Б. Попельский (1896) в лаборатории И.П. Павлова впервые показали, что введение соляной кислоты в двенадцатиперстную кишку собаки вызывает секрецию поджелудочного сока. После этого предлагались другие стимуляторы — разные гормональные препараты. Сейчас наиболее информативным считается секретин-панкреозиминотест [77, 126]. Несмотря на значительный прогресс в этой области, исследование функционального состояния желчного пузыря и поджелудочной железы с помощью дуоденального зонда сейчас не удовлетво-

ряет клиницистов вследствие сложности в интерпретации результатов. Причиной этого, с нашей точки зрения, состоит в особенностях самого объекта исследования, т.е. органов гепатогастропанкреатодуоденальной зоны, где наблюдаются очень сложные взаимные связи и влияния. Все это вызывает значительную вариабельность экспериментальных данных секреции, превращая их в реализации случайных процессов, поэтому нельзя получить достаточно емкую интерпретацию результатов анализа и диагностику гастродуоденальных заболеваний на основании этих данных без использования математических методов обработки экспериментальных данных.

3.10.2. Комплексный метод фракционного гастродуоденального зондирования

Мы опишем только суть этого метода, его детальное описание приведено в работе [76]. Для исследования функционального состояния органов желудочно-кишечного тракта в этом методе используется двойной гастродуоденальный зонд, который содержит два отверстия. Одно отверстие предназначено для изучения секреции желудка (так называемой *базальной секреции*), а другой — для поджелудочной железы и желчного пузыря. Содержимое двенадцатиперстной кишки и желудка вытекает спонтанно; считается, что его появление во время зондирования зависит от моторики у данного больного и является диагностическим признаком. Наличие желудочного зонда в желудке позволяет следить за количественными и качественными изменениями поджелудочной секреции на протяжении периода зондирования, а также регистрировать рефлюкс дуоденального содержимого в желудок. В качестве раздражителя для поджелудочной железы используется 30 мл 0,5%-ного раствора соляной кислоты, а для желчного пузыря — 20 мл масла. Сначала в дуоденальный зонд вводится 0,5%-ный раствор HCl, а после получения секреции поджелудочной железы вводится масло для получения рефлюкса желчного пузыря. Функция поджелудочной железы и желчного пузыря наблюдается на протяжении одного исследования. По мнению клиницистов, такой подход является единственно правильным способом получения достоверной информации о состоянии моторной и секреторной функции желудка, желчевыводящих путей, двенадцатиперстной кишки и поджелудочной железы.

Использованная нами методика разрешает проследить и оценить результат взаимного влияния желудка, поджелудочной железы, желчного пузыря и двенадцатиперстной кишки, а также дает много материала как для лабораторного, так и для клинического анализа.

3.10.3. Регистрация результатов гастродуоденального зондирования

Во время гастродуоденального зондирования у пациента регистрируется кислотность (p) желудочного содержимого и рефлюкс, пассаж и рефлюкс двенадцатиперстной кишки, базальная секреция, секреция поджелудочной железы

(после введения НСІ) и желчного пузыря (после введения масла), глюкоза, давление в двенадцатиперстной кишке (в мл водного столба). Все эти данные, а также фамилия, имя и отчество больного, его пол и диагноз записываются в виде таблицы функционального состояния органов желудочно-кишечного тракта пациента (табл. 3.34), которая потом используется для подтверждения или для корректирования (уточнения) диагноза. Надо заметить, что для диагностирования постхолецистоэктомического синдрома секрета желчного пузыря не регистрируется, поскольку этот синдром не может возникать у пациентов после резекции желчного пузыря. Поэтому в этом случае соответствующая таблица функционального состояния желудочно-кишечного тракта отличается от всех других таблиц, предназначенных для диагностирования других патологий (табл. 3.35).

Таблица 3.34. Функциональное состояние желудочно-кишечного тракта органов пациента

Фамилия, имя, отчество				А.Б.В.				12.04.01					
Возраст				43									
Пол				М									
Диагноз и дата обследования				Язвенная болезнь двенадцатиперстной кишки									
Желудок													
Кислотность (рН)				Рефлюкс				Цвет					
1,0				0									
Двенадцатиперстная кишка													
Пассаж				Цвет									
+				ж									
Базальная секреция													
№	Реф.1	Объем	Пас.1	Цвет	рН	№	Реф.1	Объем	Пас.1	Цвет	рН		
1		0,5	1	ж		6							
2		0,5	0	ж		7							
3		0,5	1	ж		8							
4						9							
5						10							
Поджелудочная железа после НСІ, время латентного периода 3–5 мин													
№	Реф.2	Объем	Цвет	рН	Пас. 2	Хлопья	№	Реф.2	Объем	Пас. 2	Цвет	рН	Хлопья
1		16	сж				11						
2		13	сж				12						
3		18	сж				13						
4		14	сж				14						
5		16	сж		1		15						

№	Реф.2	Объем	Цвет	pH	Пас. 2	Хлопья	№	Реф.2	Объем	Пас. 2	Цвет	pH	Хлопья
6		5	сж		1		16						
7		2	сж		1		17						
8					1		18						
9		2	сж		1		19						
10					1		20						

Желчный пузырь,
масло, продолжительность латентного периода 7–15 мин

№	Реф.2	Объем	Цвет	pH	Пас.2	№	Реф.2	Объем	Пас.2	Цвет	pH
1		4	тк	1,0		11					
2		5	тк	1,0		12					
3		13	тк	1,0		13					
4		18	тк	1,0	1	14					
5		15	тк	1,0		15					
6		16	тк	1,0	1	16					
7					1	17					
8						18					
9						19					
10						20					

глюкоза, продолжительность латентного периода 1–3 мин

№	Реф.2	Объем	Цвет	pH	Пас.2	№	Реф.2	Объем	Цвет	pH	Пас.2
1		14	жк	1,0	0	4		14	жк	1,0	1
2		7	тк	1,0	0	5		3	сж	1,0	1
3		13	тк	1,0	0	6					

Давление в двенадцатиперстной кишке (мл водного столба)

Пониженный (мин – 40)	Нормальный (80–120)	Повышенный (макс – 360)
		160

Таблица 3.35. Функциональное состояние желудочно-кишечного тракта органов пациента

Фамилия, имя, отчество	О.С.Т.	28.09.00
Возраст	40	
Пол	ж	
Диагноз и дата обследования	Постхолецисектомический синдром	

Желудок													
Кислотность (pH)				Рефлюкс				Цвет					
7,0				1									
Двенадцатиперстная кишка													
Пассаж								Цвет					
0								ж					
Базальная секреция													
№	Реф.1	Объем	Пас.1	Цвет	pH	№	Реф.1	Объем	Пас.1	Цвет	pH		
1		11		ск	7,0	6							
2		6		ск	7,0	7							
3		5		ск	7,0	8							
4		4		ск	7,0	9							
5						10							
Поджелудочная железа после НСИ, продолжительность латентного периода 3–5 мин													
№	Реф.2	Объем	Пас.2	Цвет	pH	Хлопья	№	Реф.2	Объем	Пас.2	Цвет	pH	Хлопья
1	+	27		ск	7,0	1	11	1	22		ск	7,0	1
2	+	25		ск	7,0	1	12	1	31		ск	7,0	1
3	+	20		ск	7,0	1	13	1	22		ск	7,0	0
4	+	11		ск	7,0	1	14	1	35		ск	7,0	1
5	+	24		ск	7,0	1	15	1	6		ск	7,0	0
6	+	22		ск	7,0	1	16						
7	+	21		ск	7,0	1	17						
8	+	32		ск	7,0	1	18						
9	+	22		ск	7,0	1	19						
10	+	31		ск	7,0	1	20						
глюкоза, продолжительность латентного периода 1–3 мин													
№	Реф.2	Объем	Пас.2	Цвет	pH	№	Реф.2	Объем	Пас.2	Цвет	pH		
1	+	12		ск	7,0	4	+	3		ск	7,0		
2	+	5		ск	7,0	5	+	4		ск	7,0		
3	+	4		ск	7,0	6							
Давление в двенадцатиперстной кишке (мм водного столба)													
Пониженный (мин – 40)				Нормальный (80–120)				Повышенный (макс. – 360)					
								130					

Примечание. В табл. 3.34 и 3.35 качественный показатель “цвет” базальной секреции, а также секреции двенадцатиперстной кишки, поджелудочной железы и желчного пузыря кодируется цифровыми данными таким образом: бесцветный — 1; светло-желтый — 2; желтый — 3; желто-коричневый — 3,5; светло-коричневый — 4; коричневый — 5; темно-коричневый — 6.

Анализ данных табл. 3.34 и 3.35 свидетельствует о том, что такие показатели, как кислотность (p), рефлюкс, цвет желудочного содержимого, рефлюкс и пассаж двенадцатиперстной кишки, давление двенадцатиперстной кишки, а также возраст и пол пациента являются случайными величинами, поскольку эти показатели варьируют в довольно широких границах у разных пациентов (точнее их следует рассматривать как реализацию возможных значений случайных величин). С другой стороны, показатели базальной секреции, секреции поджелудочной железы и желчного пузыря у одного пациента изменяются на протяжении времени проведения диагностирования, их динамика существенно зависит от индивидуальных особенностей больного, поэтому они являются случайными процессами. В процессе гастродуоденального зондирования пациента наблюдаются реализации этих случайных процессов. Эти реализации являются функциями непрерывного времени t . Тем не менее в таблицах указаны значения этих функций в моменты времени (после окончания латентного периода) $t_0 = 0$, $t_1 = 5$ мин, $t_2 = 10$ мин и т.д., $t_k = k \times 5$ мин, $k = 1, 2, \dots$, так что значение функций регистрируются каждые 5 мин. Заметим, что в таблицах функционального состояния органов желудочно-кишечного тракта указанные выше моменты времени t_k отождествляются с их порядковым номером k , поэтому считается, что $t_k = k$. Таким образом, случайные процессы $x(t)$ с непрерывным временем t аппроксимируются (приближаются) процессами с дискретным временем $x(t_k) = x_k$, которые в природоведении и технике называются временными рядами. Итак, в таблицах функционального состояния желудочно-кишечного тракта показатели базальной секреции, секреции поджелудочной железы и желчного пузыря для каждого пациента являются реализациями временных рядов.

3.10.4. Доверительные границы для основной распределенной массы значений генеральной совокупности. Правило 3σ и $3s$

Одной из основных задач современной теории вероятностей и математической статистики является проблема построения доверительного интервала (a, b) , содержащего значение x генеральной совокупности G с достаточно большой вероятностью, т.е. с вероятностью, которая превышает общепринятый в природоведении и технике доверительный уровень 0,95 (или 95%)

$$p(x \in (a, b)) \geq 0,95. \quad (3.6)$$

Такие доверительные интервалы получили название доверительных интервалов, содержащих основную распределенную массу значений x генеральной

совокупности G (сокращенно b -интервалы от английского названия *bulk interval*). Поскольку b -интервалы используются для построения разных статистических критериев для проверки гипотез (например, для медицинской диагностики заболеваний), которые должны иметь по возможности большую мощность (специфичность), эти доверительные интервалы необходимо избирать кратчайшими. Напомним также, что концы a, b доверительного интервала (a, b) называются соответственно нижней и верхней доверительной границей для основной распределенной массы значений генеральной совокупности G , а величина $p(x \notin (a, b))$ — уровнем значимости интервала (a, b) . Нетрудно заметить, что уровень значимости b -интервала (a, b) не превышает 0,05 (или 5%).

Для вычисления b -интервалов генеральной совокупности G необходима определенная информация. В роли такой информации можно использовать математическое ожидание (среднее значение) $m = m(x)$ и дисперсию $D(x)$ (или среднее квадратичное отклонение $\sigma = \sigma(x) = \sqrt{D(x)}$) генеральной совокупности G . Классическое правило 3σ утверждает, что для всех генеральных совокупностей G , встречающихся на практике, доверительный интервал $I = (m(x) - 3\sigma(x), m(x) + 3\sigma(x))$ содержит по крайней мере 95% значений G . Таким образом, считается, что доверительный интервал $I = (m - 3\sigma, m + 3\sigma)$ является b -интервалом и поэтому имеет уровень значимости, не превышающий 0,05.

Рассмотрим проблему обоснования правила 3σ . Хорошо известное неравенство Чебышева дает возможности оценить вероятность выхода x за доверительный интервал $(m(x) - \lambda\sigma(x), m(x) + \lambda\sigma(x))$ для произвольного положительного числа λ :

$$p(x \in (m(x) - \lambda\sigma(x), m(x) + \lambda\sigma(x))) = p(|x - m(x)| \geq \lambda\sigma(x)) \leq 1/\lambda^2. \quad (3.7)$$

Если положить $\lambda = 3$, то, согласно неравенству Чебышева, уровень значимости интервала I не превышает $1/9 = 0,11\dots$, что значительно превышает (по крайней мере в 2 раза) величину 0,05 уровня значимости для b -интервала. Поскольку неравенство Чебышева является точным (при любых $1/2\lambda \leq 1$ существуют генеральные совокупности G , для которых это неравенство превращается в равенство), обоснование правила 3σ для произвольной генеральной совокупности G невозможно. Среди разных генеральных совокупностей, встречающихся в приложениях статистики, наибольший интерес представляют так называемые *однородные генеральные совокупности*, отличающиеся от составных или смешанных генеральных совокупностей тем, что они характеризуют класс объектов, имеющих общие признаки, которые хотя и варьируют в определенных границах, но, тем не менее, похожи между собой. Так, можно говорить об однородной популяции (генеральной совокупности) больных, страдающих каким-нибудь одним заболеванием, в то время как популяции пациентов с разными (например, двумя) заболеваниями являются составными популяциями. Математическое определение однородной генеральной совокупности основано на понятии одномодального распределения $F(u)$ этой совокупности. Согласно Хинчину, назовем функцию распределения $F(u)$ одномодальной, если существ-

вует такая точка M на прямой (называемая модой), если на луче $(-\infty, M)$ функция $F(u)$ является выпуклой, а на луче $(M, +\infty)$ — вогнутой. Генеральная совокупность G с одномодальной функцией распределения $F(u)$ называется однородной. Характерным признаком однородных генеральных совокупностей является их одномодальность; нетрудно заметить, что двух- или многомодальные генеральные совокупности служат примером составных генеральных совокупностей или смесей. Обоснование правила 3σ для одномодальных генеральных совокупностей можно получить способом, предложенным К.Ф. Гауссом. В 1821 году для симметричных одномодальных генеральных совокупностей им было доказано неравенство, которое называется неравенством Гаусса:

$$p(|x - M(x)| \geq \lambda \sigma(x)) \leq \frac{4}{9\lambda^2}. \quad (3.8)$$

В своих мемуарах Гаусс поставил проблему доказательства этого неравенства при $\lambda = 3$ для произвольных (не обязательно симметричных) одномодальных распределений. Эта проблема была решена в 1979 году в работе [44], где было показано, что для любых одномодальных распределений при $\lambda > \sqrt{8/3}$ выполняется неравенство (3.8), называемое неравенством Гаусса–Высочанского–Петунина. Оно позволяет полностью обосновать справедливость эмпирического правила 3σ для одномодальных распределений. Действительно, при $\lambda = 3 > \sqrt{8/3}$ правая часть в неравенстве (3.8) равняется $\frac{4}{9\lambda^2} = \frac{4}{81} < \frac{4}{80} = 0,05$, так что уровень значимости доверительного интервала $I = (m - 3\sigma, m + 3\sigma)$ не превышает 0,05, ведь I является b -интервалом, если, конечно, G — одномодальная генеральная совокупность. Поскольку неравенство (3.8) решает фундаментальную проблему, то для подтверждения ее справедливости и уточнения в работах [6, 22, 24] были предложены другие доказательства этого неравенства.

В приложениях ни математическое ожидание $m(x)$, ни дисперсия $D(x)$ никогда не бывают известными абсолютно точно, однако на основании выборки $x_1, x_2, \dots, x_n$ из генеральной совокупности G , полученной простым случайным выбором, можно получить оптимальные оценки для $m(x)$ и $D(x)$:

$$m \approx \bar{x} = \frac{1}{n} \sum_{k=1}^n x_k, \quad (3.9)$$

$$D(x) \approx s^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2. \quad (3.10)$$

Поэтому доверительный интервал $\bar{I}$, построенный по правилу $3s$, имеет вид

$$\bar{I} = (\bar{x} - 3s, \bar{x} + 3s). \quad (3.11)$$

При больших значениях n объема обучающей выборки $x_1, x_2, \dots, x_n$ этот интервал будет близок к интервалу, полученному по правилу 3σ , поскольку при этих значениях n $m \approx \bar{x}$, а $\sigma(x) \approx s$. Проблема точной оценки уровня значимости

интервала $\bar{T}$ даже в случае одномодальных распределений при небольших значениях n представляет собой очень сложную математическую задачу, но, используя методы компьютерного моделирования, можно показать, что этот уровень значимости для распределений, встречающихся на практике, не превышает 0,05 при $n \geq 10$.

3.10.5. Метод сплайновой регрессии для анализа временных рядов

Рассмотрим случайный процесс $x(t)$, определенный на отрезке $[0, T]$. Будем считать, что тренд $m(t) = m(x(t))$ (переменная детерминированная компонента) является кусочно-линейным сплайном, имеющим такой вид:

$$m(t, t^*, a_1, b_1, a_2, b_2) = m(t) = \begin{cases} a_1 t + b_1, & \text{если } 0 \leq t < t^*, \\ a_2 t + b_2, & \text{если } t^* \leq t \leq T, \end{cases} \quad (3.12)$$

$$a_1 t^* + b_1 = a_2 t^* + b_2,$$

где t^* — точка перехода (рис. 3.22).

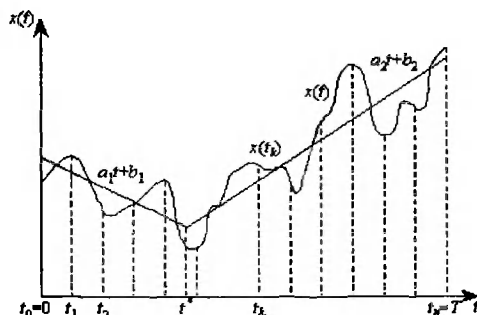


Рис. 3.22. Модель сплайновой регрессии

Необходимо по наблюдению реализации (траектории) $x(t_k)$ процесса $x(t)$ в точках t_k ($t_0 = 0, \dots, t_n = T$) оценить точку перехода t^* и параметры a_1, b_1, a_2, b_2 сплайна $m(t)$. Решение этой задачи будем осуществлять методом минимизации остаточной суммы квадратов. Опишем коротко суть этого метода. Предположим, что точка перехода t^* известна и совпадает с одной из точек наблюдения траектории $x(t)$: $t^* = t_p$, тогда задача оценки параметров a_1, b_1, a_2, b_2 сводится к проблеме оценки параметров в схеме линейной регрессии, поскольку в этом случае $m(t)$ можно представить в виде линейной комбинации фундаментальных сплайнов [66]. Используя классический метод наименьших квадратов (МНК) [112], можно получить так называемые оценки МНК $a_1^{(i)}, b_1^{(i)}, a_2^{(i)}, b_2^{(i)}$ для неизвестных параметров a_1, b_1, a_2, b_2 (эти оценки зависят от нашей гипотезы $t^* = t_p$, а потому и от индекса i) и остаточную сумму квадратов:

$$S_i^2 = \sum_{k=0}^{i-1} (x(t_k) - a_1 t_k - b_1)^2 + \sum_{k=i}^n (x(t_k) - a_2 t_k - b_2)^2. \quad (3.13)$$

Выбрав потом среди остаточных сумм $S_0^2, S_1^2, \dots, S_n^2$ наименьшую сумму $S_{i^*}^2 = \min\{S_0^2, S_1^2, \dots, S_n^2\}$ и находя ее порядковый номер i^* , мы найдем оценку для точки перехода $t_{i^*} \approx \hat{t}$ и оценки $a_1^{(i^*)}$, $b_1^{(i^*)}$, $a_2^{(i^*)}$, $b_2^{(i^*)}$ неизвестных параметров a_1 , b_1 , a_2 , b_2 для сплайна $m(t)$. Будем называть эти оценки t_{i^*} , $a_1^{(i^*)}$, $b_1^{(i^*)}$, $a_2^{(i^*)}$, $b_2^{(i^*)}$ оценками, полученными методом минимизации остаточных сумм квадратов (оценки метода МОСК). Покажем, что при выполнении некоторых условий эти оценки стремятся по вероятности к истинным значениям параметров $\hat{t}$, a_1 , b_1 , a_2 , b_2 .

Теорема 3.1.

Если в модели регрессии

$$x(t_k) = m(t_k, \hat{t}, a_1, b_1, a_2, b_2) + y_k, \quad (3.14)$$

где $k = 1, 2, \dots, n$, $t_1, t_2, \dots, t_n$ — узлы равномерного разбиения отрезка $[0, T]$, шум y_k ($k = 1, 2, \dots, n$) является последовательностью независимых в совокупности одинаково распределенных случайных величин и вектор параметров $\alpha = (\hat{t}, a_1, b_1, a_2, b_2)$ принадлежит компактному множеству Θ , то оценки параметров t_{i^*} , $a_1^{(i^*)}$, $b_1^{(i^*)}$, $a_2^{(i^*)}$, $b_2^{(i^*)}$, полученные методом МОСК, являются строго состоятельными.

Доказательство теоремы опирается на хорошо известную классической теореме Дженрича о строгой состоятельности параметров α в схеме нелинейной регрессии [10], если тренд $f(\alpha)$ является непрерывной функцией регрессии на компакте Θ . Если предположить, что для произвольных $\alpha, \beta \in \Theta$ существует предел

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (f_i(\alpha) - f_i(\beta))^2 = \varphi(\alpha, \beta), \quad (3.15)$$

причем сходимость в (3.15) равномерная и $\varphi(\alpha, \beta) = 0$ тогда и только тогда, когда $\alpha = \beta$, то, как показано в работе [58], оценка $\hat{\alpha}_n$ метода МОСК будет строго состоятельной, т.е. $p(\hat{\alpha}_n \rightarrow \alpha) = 1$, где α — истинное значение вектора параметров.

Проверим справедливость условий теоремы Дженрича в случае сплайновой регрессии. Для удобства введем такие обозначения:

$$f_i(\alpha) = f(t, \alpha) = \begin{cases} \alpha_2 t + \alpha_3, & \text{если } 0 \leq t < \alpha_1, \\ \alpha_4 t + \alpha_5, & \text{если } \alpha_1 \leq t \leq T, \end{cases}$$

$$f_i(\beta) = f(t, \beta) = \begin{cases} \beta_2 t + \beta_3, & \text{если } 0 \leq t < \beta_1, \\ \beta_4 t + \beta_5, & \text{если } \beta_1 \leq t \leq T, \end{cases}$$

$$\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5), \quad \beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5), \quad \alpha, \beta \in \Theta = \{(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) \mid |\theta_i| \leq C, i = \overline{1,5}\}$$

Без ограничения общности можно считать, что $\alpha_1 \leq \beta_1$. Положим

$$\Phi(t) = \Phi(t, \alpha, \beta) = f(t, \alpha) - f(t, \beta) = \begin{cases} c_1 + c_2 t, & \text{если } 0 \leq t < \alpha_1, \\ d_1 + d_2 t, & \text{если } \alpha_1 \leq t < \beta_1, \\ e_1 + e_2 t, & \text{если } \beta_1 \leq t \leq T, \end{cases}$$

где $c_1 = \alpha_3 - \beta_3$, $c_2 = \alpha_2 - \beta_2$, $d_1 = \alpha_3 - \beta_3$, $d_2 = \alpha_2 - \beta_4$, $e_1 = \alpha_3 - \beta_3$, $e_2 = \alpha_4 - \beta_4$.

Обозначим

$$\varphi_n(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^n \Phi^2(t_i) = \frac{1}{n} \sum_{i=1}^n (f(t_i, \alpha) - f(t_i, \beta))^2, \quad t_i = \frac{i}{n} T, \quad i = \overline{1, n},$$

$$\varphi(\alpha, \beta) = \lim_{n \rightarrow \infty} \varphi_n(\alpha, \beta) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \Phi^2(t_i) = \frac{1}{T} \lim_{n \rightarrow \infty} \sum_{i=1}^n \Phi^2(t_i) \Delta t = \frac{1}{T} \int_0^T \Phi^2(t) dt,$$

где $\Delta t = \frac{T}{n}$.

Покажем, что последовательность $\varphi_n(\alpha, \beta)$ равномерно сходится к функции $\varphi(\alpha, \beta)$ на декартовом произведении $\Theta \times \Theta$. Введем следующие обозначения:

$$r = \left[\frac{n}{T} \alpha_1 \right], \quad s = \left[\frac{n}{T} \beta_1 \right], \quad t_r = \frac{T}{n} r, \quad t_s = \frac{T}{n} s \quad (\text{рис. 3.23}).$$

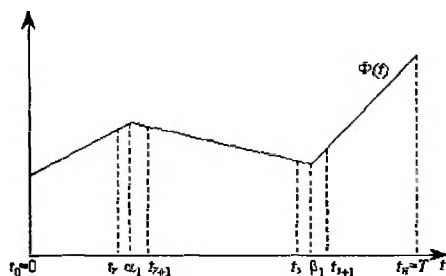


Рис. 3.23. Разность между двумя кусочно-линейными сплайнами

Легко видеть, что

$$\begin{aligned} \varphi(\alpha, \beta) - \varphi_n(\alpha, \beta) &= \frac{1}{T} \int_0^T \Phi^2(t) dt - \frac{1}{n} \sum_{i=1}^n \Phi^2(t_i) = \frac{1}{T} \int_0^{t_r} \Phi^2(t) dt - \frac{1}{n} \sum_{i=1}^r \Phi^2(t_i) + \frac{1}{T} \int_{t_r}^{t_s} \Phi^2(t) dt - \\ &- \frac{1}{n} \Phi^2(t_{r+1}) + \frac{1}{T} \int_{t_{r+1}}^{t_s} \Phi^2(t) dt - \frac{1}{n} \sum_{i=r+2}^s \Phi^2(t_i) + \frac{1}{T} \int_{t_s}^{t_{s+1}} \Phi^2(t) dt - \\ &- \frac{1}{n} \Phi^2(t_{s+1}) + \frac{1}{T} \int_{t_{s+1}}^T \Phi^2(t) dt - \\ &- \frac{1}{n} \sum_{i=s+2}^n \Phi^2(t_i) = I_1 + I_2 + I_3 + I_4 + I_5, \end{aligned}$$

где

$$I_1 = \frac{1}{T} \int_0^{t_r} \Phi^2(t) dt - \frac{1}{n} \sum_{i=1}^r \Phi^2(t_i), \quad I_2 = \frac{1}{T} \int_{t_{r+1}}^{t_s} \Phi^2(t) dt - \frac{1}{n} \sum_{i=r+2}^s \Phi^2(t_i),$$

$$I_3 = \frac{1}{T} \int_{t_{r+1}}^T \Phi^2(t) dt - \frac{1}{n} \sum_{i=s+2}^n \Phi^2(t_i), \quad I_4 = \frac{1}{T} \int_{t_r}^{t_{r+1}} \Phi^2(t) dt - \frac{1}{n} \Phi^2(t_{r+1}),$$

$$I_5 = \frac{1}{T} \int_{t_r}^{t_{r+1}} \Phi^2(t) dt - \frac{1}{n} \Phi^2(t_{s+1}).$$

Рассмотрим первое слагаемое I_1 :

$$I_1 = \frac{1}{n} \sum_{i=1}^r \left(\frac{n}{T} \int_{t_{i-1}}^{t_i} \Phi^2(t) dt - \Phi^2(t_i) \right).$$

Поскольку, $0 \leq t_{i-1} < t_i \leq \alpha_1$, $i = \overline{1, r}$, то

$$\int_{t_{i-1}}^{t_i} \Phi^2(t) dt = \int_{t_{i-1}}^{t_i} (c_1 + c_2 t)^2 dt = (t_i - t_{i-1}) \left(\frac{1}{2} (c_1 + c_2 t_i)^2 + \frac{1}{2} (c_1 + c_2 t_{i-1})^2 - \frac{1}{6} c_2^2 (t_i - t_{i-1})^2 \right) =$$

$$= \frac{T}{n} \left(\frac{1}{2} \Phi^2(t_i) + \frac{1}{2} \Phi^2(t_{i-1}) - \frac{1}{6} c_2^2 \frac{T^2}{n^2} \right),$$

поэтому

$$I_1 = \frac{1}{n} \sum_{i=1}^r \left(\frac{1}{2} \Phi^2(t_i) + \frac{1}{2} \Phi^2(t_{i-1}) - \frac{1}{6} c_2^2 \frac{T^2}{n^2} - \Phi^2(t_i) \right) = \frac{1}{n} \left(\frac{1}{2} \Phi^2(0) - \frac{1}{2} \Phi^2(t_r) - \frac{1}{12} c_2^2 T^2 \frac{r(r+1)}{n^2} \right).$$

Положим $\delta_1 = \left\{ \alpha_1 \frac{n}{T} \right\}$; учитывая, что $t_r = \alpha_1 - \delta_1 \frac{T}{n}$ и $t_{r+1} = \alpha_1 + (1 - \delta_1) \frac{T}{n}$, имеем

$$\Phi^2(t_r) = (c_1 + c_2 t_r)^2 = (c_1 + c_2 \alpha_1 - c_2 \delta_1 \frac{T}{n})^2 = (c_1 + c_2 \alpha_1)^2 + O\left(\frac{1}{n}\right) = \Phi^2(\alpha_1) + O\left(\frac{1}{n}\right),$$

аналогично

$$\Phi^2(t_{r+1}) = (d_1 + d_2 t_r)^2 = \Phi^2(\alpha_1) + O\left(\frac{1}{n}\right).$$

Итак,

$$I_1 = \frac{1}{2n} \left(\Phi^2(0) - \Phi^2(\alpha_1) - \frac{1}{6} c_2^2 T^2 \frac{r(r+1)}{n^2} \right) + o\left(\frac{1}{n}\right).$$

Проведя аналогичные вычисления, получим

$$I_2 = \frac{1}{2n} \left(\Phi^2(\alpha_1) - \Phi^2(\beta_1) - \frac{1}{6} d_2^2 T^2 \frac{(s-r-1)(s-r)}{n^2} \right) + o\left(\frac{1}{n}\right),$$

$$I_3 = \frac{1}{2n} \left(\Phi^2(\beta_1) - \Phi^2(T) - \frac{1}{6} e_2^2 T^2 \frac{(n-s-1)(n-s)}{n^2} \right) + o\left(\frac{1}{n}\right).$$

Оценим слагаемое I_4 :

$$I_4 = \frac{1}{T} \int_{t_r}^{\alpha_1} \Phi^2(t) dt + \frac{1}{T} \int_{\alpha_1}^{t_{r+1}} \Phi^2(t) dt - \frac{1}{n} \Phi^2(t_{r+1}).$$

Имеем

$$\begin{aligned} \frac{1}{T} \int_{t_r}^{\alpha_1} \Phi^2(t) dt &= \frac{1}{T} (\alpha_1 - t_r) \left(\frac{1}{2} \Phi^2(\alpha_1) + \frac{1}{2} \Phi^2(t_r) - \frac{1}{6} c_2^2 (\alpha_1 - t_r)^2 \right) = \\ &= \frac{1}{n} \delta_1 \left(\frac{1}{2} \Phi^2(\alpha_1) + \frac{1}{2} \Phi^2(t_r) - \frac{1}{6} c_2^2 \delta_1^2 \frac{T^2}{n^2} \right) = \frac{\delta_1}{2n} (\Phi^2(\alpha_1) + \Phi^2(t_r)) + o\left(\frac{1}{n}\right) = \frac{\delta_1}{n} \Phi^2(\alpha_1) + o\left(\frac{1}{n}\right), \\ \frac{1}{T} \int_{\alpha_1}^{t_{r+1}} \Phi^2(t) dt &= \frac{1}{T} (t_{r+1} - \alpha_1) \left(\frac{1}{2} \Phi^2(t_{r+1}) + \frac{1}{2} \Phi^2(\alpha_1) - \frac{1}{6} d_2^2 (t_{r+1} - \alpha_1)^2 \right) = \\ &= \frac{1}{n} (1 - \delta_1) \left(\frac{1}{2} \Phi^2(t_{r+1}) + \frac{1}{2} \Phi^2(\alpha_1) - \frac{1}{6} d_2^2 (1 - \delta_1)^2 \frac{T^2}{n^2} \right) = \\ &= \frac{1 - \delta_1}{n} \left(\Phi^2(\alpha_1) + O\left(\frac{1}{n}\right) + \Phi^2(\alpha_1) \right) + o\left(\frac{1}{n}\right) = \frac{1 - \delta_1}{n} \Phi^2(\alpha_1) + o\left(\frac{1}{n}\right), \end{aligned}$$

так что

$$I_4 = \frac{\delta_1}{n} \Phi^2(\alpha_1) + \frac{1 - \delta_1}{n} \Phi^2(\alpha_1) - \frac{1}{n} \Phi^2(t_{r+1}) + o\left(\frac{1}{n}\right) = \frac{1}{n} (\Phi^2(\alpha_1) - \Phi^2(t_{r+1})) + o\left(\frac{1}{n}\right) = o\left(\frac{1}{n}\right).$$

Положив $\delta_2 = \left\{ \beta_1 \frac{n}{T} \right\}$ и проводя аналогичные рассуждения, получим

$$I_5 = o\left(\frac{1}{n}\right).$$

Итак,

$$\begin{aligned} \varphi(\alpha, \beta) - \varphi_n(\alpha, \beta) &= \frac{1}{2n} \left(\Phi^2(0) - \Phi^2(\alpha_1) - \frac{1}{6} c_2^2 T^2 \frac{r(r+1)}{n^2} \right) + \\ &+ \frac{1}{2n} \left(\Phi^2(\alpha_1) - \Phi^2(\beta_1) - \frac{1}{6} d_2^2 T^2 \frac{(s-r-1)(s-r)}{n^2} \right) + \\ &+ \frac{1}{2n} \left(\Phi^2(\beta_1) - \Phi^2(T) - \frac{1}{6} e_2^2 T^2 \frac{(n-s-1)(n-s)}{n^2} \right) + o\left(\frac{1}{n}\right) = \\ &= \frac{1}{2n} (\Phi^2(0) - \Phi^2(T)) - \frac{T^2}{12n} \left(c_2^2 \frac{r(r+1)}{n^2} + d_2^2 \frac{(s-r-1)(s-r)}{n^2} + e_2^2 \frac{(n-s-1)(n-s)}{n^2} \right) + o\left(\frac{1}{n}\right), \end{aligned}$$

Значит,

$$\begin{aligned} |\varphi(\alpha, \beta) - \varphi_n(\alpha, \beta)| &\leq \frac{1}{2n} |\Phi^2(0) - \Phi^2(T)| + \\ &+ \frac{T^2}{12n} \left(c_2^2 \frac{r(r+1)}{n^2} + d_2^2 \frac{(s-r-1)(s-r)}{n^2} + e_2^2 \frac{(n-s-1)(n-s)}{n^2} \right) \leq \end{aligned}$$

$$\leq \frac{1}{2n}(\Phi^2(0) + \Phi^2(T)) + \frac{T^2}{12n}(c_2^2 + d_2^2 + e_2^2).$$

Затем положим, что

$$c_2^2 = (\alpha_2 - \beta_2)^2 \leq 4C^2,$$

аналогично $d_2^2 \leq 4C^2$ и $e_2^2 \leq 4C^2$, ведь

$$\begin{aligned}\Phi^2(0) &= c_1^2 = (\alpha_3 - \beta_3)^2 \leq 4C^2, \\ \Phi^2(T) &= (e_1 + e_2 T)^2 \leq e_1^2 + 2|e_1||e_2|T + e_2^2 T^2 = (\alpha_5 - \beta_5)^2 + 2|\alpha_5 - \beta_5||\alpha_4 - \beta_4|T + \\ &\quad + (\alpha_4 - \beta_4)^2 T^2 \leq 4C^2 + 8C^2 T + 4C^2 T^2.\end{aligned}$$

Таким образом

$$\begin{aligned}|\varphi(\alpha, \beta) - \varphi_n(\alpha, \beta)| &\leq \frac{1}{2n}(4C^2 + 4C^2 + 8C^2 T + 4C^2 T^2) + \frac{T^2}{12n}(4C^2 + 4C^2 + 4C^2) = \\ &= \frac{1}{n}(4C^2 + 4C^2 T + 3C^2 T^2).\end{aligned}$$

Полученное выражение меньше любого $\varepsilon > 0$, если

$$n \geq N = \left\lceil \frac{4C^2 + 4C^2 T + 3C^2 T^2}{\varepsilon} \right\rceil + 1.$$

Проверка других условий теоремы Дженрича очевидна. Теорема доказана.

Оценка $\hat{\alpha}$ по методу МОСК является смещенной. Для оценки точки перехода $\hat{t}^*$ в случае $n = 3$ можно получить смещение этой оценки и ее дисперсии. В работе [111] показано, что смещение $\Delta \hat{t}^*$ точки перехода зависит от двух переменных — дисперсии шума σ^2 и величины $(b_1 - b_2)^2$.

Теорема 3.2.

Если в модели регрессии (9) шум y_k является последовательностью независимых в совокупности нормально распределенных случайных величин с дисперсией σ^2 и нулевым математическим ожиданием, то смещение $\Delta \hat{t}^*$ оценки точки перехода имеет вид

$$\begin{aligned}\Delta \hat{t}^* &= F(\sigma^2, (b_1 - b_2)^2) = \frac{1}{6\pi} \left(\int_{\frac{\sigma}{2} - \varphi}^{\sigma + \varphi} \exp\left(-\frac{(b_1 - b_2)^2}{20\sigma^2(\cos\theta - \sin\theta)^2}\right) d\theta + \right. \\ &\quad \left. + \int_{\frac{\sigma}{2} + \varphi}^{\sigma + \varphi} \exp\left(-\frac{(b_1 - b_2)^2}{4\sigma^2(\cos\theta - \sin\theta)^2}\right) d\theta - \int_{-\varphi}^{\varphi} \exp\left(-\frac{(b_1 - b_2)^2(3 + \sqrt{5})}{40\sigma^2 \cos^2\theta}\right) d\theta \right),\end{aligned}\tag{3.16}$$

где $\varphi = \arccos \sqrt{\frac{3 + \sqrt{5}}{6}}$, а дисперсия оценки точки перехода определяется формулой

$$D(\hat{t}^*) = F(\sigma^2, (b_1 - b_2)^2) \left(\frac{1}{3} - F(\sigma^2, (b_1 - b_2)^2) \right). \quad (3.17)$$

Как известно, на основании смещения $\Delta \hat{t}^*$ и дисперсии $D(\hat{t}^*)$ можно получить среднеквадратичную погрешность оценки $\hat{t}^*$:

$$s^2(\hat{t}^*) = (\Delta \hat{t}^*)^2 + D(\hat{t}^*). \quad (3.18)$$

3.10.6. Результаты диагностирования гастродуоденальных заболеваний

Для диагностики гастродуоденального заболевания (например, язвы двенадцатиперстной кишки) сначала необходимо получить две обучающие выборки G_1 и G_2 , одна из которых G_1 состоит из практически здоровых пациентов, а другая G_2 — из пациентов, больных этим недугом. Подчеркнем, что диагнозы пациентов обеих групп должны быть строго верифицированы. У всех пациентов должно быть проведено дуоденальное зондирование и заполнены таблицы функционального состояния желудочно-кишечного тракта. С помощью этих таблиц можно определить числовые показатели (кислотность желудочного содержимого, рефлюкс, цвет и пассаж двенадцатиперстной кишки и т.д.), а также кривые секреции, которые являются временными рядами.

После вычисления параметров сплайновой регрессии для кривых, характеризующих динамику базальной секреции, секреции поджелудочной железы и желчного пузыря, для каждого пациента P получаем вместо этих кривых пятимерные векторы $(t_p^*, a_1^p, b_1^p, a_2^p, b_2^p)$, состоящие из вышеупомянутых параметров.

После этой процедуры все показатели пациента P представляют собою случайные величины, которые являются или указанными выше числовыми показателями, или оценкой параметров сплайновой регрессии; мы будем называть эти случайные величины x показателями функционального состояния желудочно-кишечного тракта пациента P . Потом, используя правило 3σ , описанное в 3.10.4, для любой из групп G_1 и G_2 (обучающих выборок) можно определить нижнюю $x_{G_i}^{(n)}$ и верхнюю $x_{G_i}^{(n)}$ доверительные границы для основной распределенной массы (b -интервалы). Таким образом, например, в случае диагностики язвы двенадцатиперстной кишки мы будем иметь b -интервалы для практически здоровых пациентов $(x_{G_1}^{(n)}, x_{G_1}^{(n)})$ и пациентов из группы G_2 , болеющих язвой двенадцатиперстной кишки $(x_{G_2}^{(n)}, x_{G_2}^{(n)})$.

Процедура компьютерной диагностики выполняется следующим образом. У пациента P с нераспознанным заболеванием с помощью описанного выше метода определяются показатели x функционального состояния желудочно-кишечного тракта. После этого проверяется, будут ли все эти показатели принадлежать соответствующим b -интервалам $(x_{G_1}^{(1)}, x_{G_1}^{(4)})$ и $(x_{G_2}^{(1)}, x_{G_2}^{(4)})$, построенным для группы G_1 практически здоровых пациентов и группы G_2 пациентов

с диагностированным гастродуоденальным заболеванием (например, язвой двенадцатиперстной кишки). Если все показатели x попадают в соответствующие b -интервалы для группы G_1 , а некоторая часть из них не принадлежит соответствующим b -интервалам из группы G_2 , то у обследуемого пациента данного гастродуоденального заболевания нет; если же все показатели x принадлежат соответствующим b -интервалам для группы G_2 , а часть из них не принадлежит соответствующим b -интервалам группы G_1 , то ставится диагноз о наличии данного гастродуоденального заболевания у обследуемого пациента. Теоретически может случиться так, что ни одна из этих возможностей не реализуется, тогда никакого решения не принимается, а анализ следует повторить. Однако на практике для тех обучающих выборок пациентов, которые являются в нашем распоряжении, такого никогда не происходило.

Конечно, описанная выше процедура постановки диагноза может привести к ошибочным результатам. Напомним, что точностью (или чувствительностью) метода диагностирования заболевания D называется процент верно поставленных диагнозов среди группы пациентов, страдающих заболеванием D , а специфичностью — процент верно поставленных диагнозов у практически здоровых пациентов. Среди большого количества указанных выше показателей x функционального состояния желудочно-кишечного тракта пациента P наиболее интересными являются так называемые *информативные показатели*, которые являются характеристическими для данного гастродуоденального заболевания и разрешают провести его диагностику с той самой точностью и специфичностью, как и весь комплекс показателей функционального состояния желудочно-кишечного тракта. Для каждого отдельного заболевания информативные показатели являются специфическими и могут значительно отличаться от информативных показателей другого заболевания. Путем тщательного анализа показателей функционального состояния желудочно-кишечного тракта нам удалось найти среди них информативные показатели, вычислить их доверительные границы (b -интервалы) и определить точность и специфичность описанного выше диагностического метода дуоденального зондирования. Приведем полный список исследованных гастродуоденальных заболеваний, а также их диагностические характеристики; каждая обучающая выборка G , состоящая из пациентов с точно верифицированным диагнозом конкретного гастродуоденального заболевания, содержит 30 пациентов. Проведена дифференциальная диагностика для группы G_1 практически здоровых пациентов и группы G_2 пациентов со следующим гастродуоденальным заболеванием.

1. Постхолецистектомический синдром. Для этого заболевания необходимы лишь два информационных параметра: секреции поджелудочной железы a_2 и s_2^2 — минимальная остаточная сумма квадратов, которая вычисляется методом МОСК для расчета параметров в схеме сплайновой регрессии (см. 3.10.5). Для диагностики этого синдрома используются их доверительные границы (b -интервалы): для a_2 они равняются $(-0,400; -0,006)$, а для s_2^2 — $(0,228; 3,588)$. Точность (чувствительность) метода составляет 100%, а специфичность — 96,7%.

2. Холециститопанкреатит. Укажем информационные параметры для этого заболевания; числа, написанные рядом с этими параметрами в скобках, обозначают соответствующие им доверительные границы (b -интервалы):

- поджелудочная железа (хлопья) $\hat{t}$ $(-1,765; 17,842)$, a_1 $(-0,256; 0,612)$, b_1 $(-1,510; 1,144)$, b_2 $(-6,055; 12,092)$, $s_{\hat{t}}$ $(-0,600; 3,277)$;
- желчный пузырь (масло), цвет $\hat{t}$ $(-5,609; 20,368)$, a_1 $(-2,445; 3,400)$, b_1 $(-3,625; 11,036)$, $s_{\hat{t}}$ $(-6,283; 11,628)$;
- желчный пузырь (глюкоза), цвет $\hat{t}$ $(-5,087; 10,754)$, a_2 $(-5,248; 4,049)$;
- желчный пузырь (масло), объем $\hat{t}$ $(-4,485; 18,002)$, a_1 $(-11,416; 12,712)$, a_2 $(-12,275; 8,989)$, b_2 $(-79,889; 130,568)$, $s_{\hat{t}}$ $(-247,675; 462,249)$;
- желчный пузырь (глюкоза), объем $s_{\hat{t}}$ $(-216,085; 277,414)$;
- поджелудочная железа, объем a_2 $(-7,936; 2,795)$, b_2 $(-28,653; 115,784)$, $s_{\hat{t}}$ $(-367,890; 1149,730)$;
- поджелудочная железа, цвет $\hat{t}$ $(-5,628; 13,828)$, b_1 $(-3,735; 5,071)$, a_2 $(-0,461; 0,210)$, b_2 $(-0,581; 7,576)$, $s_{\hat{t}}$ $(-4,000; 8,996)$;
- базальная секреция, объем b_1 $(-15,799; 21,260)$, a_2 $(-13,806; 11,228)$, b_2 $(-40,845; 54,710)$;
- базальная секреция, цвет a_1 $(-0,104; 0,092)$, a_2 $(-0,649; 0,687)$, b_2 $(-0,560; 4,624)$, $s_{\hat{t}}$ $(-0,076; 0,087)$.

Точность метода — 96,7%, специфичность — 87,0%.

3. Язвенная болезнь двенадцатиперстной кишки. Информационные параметры и их доверительные интервалы:

- поджелудочная железа (хлопья), a_2 $(-1,380; 2,101)$, b_2 $(-14,084; 8,959)$, $s_{\hat{t}}$ $(-0,594; 1,689)$;
- желчный пузырь (масло), цвет $\hat{t}$ $(-7,065; 24,199)$, a_1 $(-1,994; 2,938)$, a_2 $(-4,743; 2,355)$, b_2 $(-18,589; 53,878)$;
- желчный пузырь (глюкоза), цвет a_2 $(-4,305; 2,721)$, b_2 $(-13,377; 27,814)$;
- желчный пузырь (масло), объем $\hat{t}$ $(-6,391; 18,658)$, a_1 $(-11,597; 15,393)$, $s_{\hat{t}}$ $(-213,432; 394,954)$;
- желчный пузырь (масло), pH $\hat{t}$ $(-5,716; 10,716)$, a_2 $(-0,581; 0,514)$, b_2 $(-7,593; 10,993)$;
- желчный пузырь (глюкоза), pH $\hat{t}$ $(1,051; 2,793)$;
- базальная секреция, объем a_1 $(-10,118; 8,883)$, b_1 $(-17,076; 23,297)$;
- поджелудочная железа, объем a_1 $(-25,163; 16,627)$, $s_{\hat{t}}$ $(-297,663; 503,375)$;

- поджелудочная железа, цвет t ($-3,457; 9,390$), a_1 ($-0,797; 0,760$), b_1 ($-3,113; 6,657$), s_r ($-2,976; 4,452$).

Точность метода — 100%, чувствительность — 100%.

Для диагностики вышеупомянутых гастродуоденальных заболеваний необходимо вычислить доверительные границы по правилу 3σ для показателей функционального состояния желудочно-кишечного тракта для группы G_1 практически здоровых пациентов (так называемой группы контроля), значение этих границ приведено в табл. 3.36.

Таблица 3.36. Доверительные границы показателей функционального состояния желудочно-кишечного тракта для группы контроля

Вид секреции	Показатели					
	t	a_1	b_1	a_2	b_2	s_r
Поджелудочная железа, хлопья	1,011	-0,034	-0,648	-0,345	-0,351	0,442
	12,656	0,329	0,021	0,006	4,045	1,502
Базальная секреция, цвет	0,277	0,000	-1,446	0,000	-1,446	0,000
	3,102	0,000	4,204	0,000	4,204	0,000
Желчный пузырь (масло), цвет	-1,440	-2,078	-3,110	-3,406	-7,596	-4,095
	14,130	3,064	9,650	1,219	33,470	7,903
Желчный пузырь (глюкоза), цвет	-3,434	-2,390	-2,598	-2,656	-11,103	-1,269
	10,140	3,092	7,682	1,439	24,111	2,176
Желчный пузырь (масло), объем	-4,451	-11,417	-20,219	-12,168	-66,975	-105,127
	14,313	14,612	31,602	9,095	113,392	234,062
Желчный пузырь (глюкоза), объем	-3,548	-12,528	-7,216	-15,101	-66,849	-67,139
	11,077	9,260	30,407	11,636	102,832	115,959
Базальная секреция, pH	0,334	0,000	-2,084	0,000	-2,084	0,000
	3,094	0,000	10,655	0,000	10,665	0,000
Желчный пузырь (масло), pH	-4,794	0,000	-1,743	-0,592	-5,064	0,000
	9,690	0,000	7,674	0,523	12,029	0,000
Глюкоза, pH	2,000	0,000	-3,310	0,000	-3,310	0,000
	2,000	0,000	9,898	0,000	9,898	0,000
Базальная секреция, объем	-0,217	-11,178	-16,163	-14,540	-29,120	-10,037
	3,941	13,461	21,654	8,475	52,925	13,091
Поджелудочная железа, объем	-2,369	-21,930	-22,568	-8,289	-26,787	-171,126
	12,902	22,743	71,271	3,363	97,081	521,859
Поджелудочная железа, цвет	-1,185	-0,767	-1,314	-0,400	1,905	0,228
	7,719	2,680	3,233	-0,006	6,243	3,588

Заметим, что в медицине метод диагностики заболевания считается удовлетворительным, если его точность и специфичность превышает 80%. Подыто-

живая результаты проведенных исследований, следует отметить, что предложенные в этой работе методы компьютерной диагностики гастродуоденальных заболеваний, основанные на результатах статистической теории распознавания, целиком пригодны для использования их в клинической медицине.

Резюме

Предложен метод компьютерной диагностики гастродуоденальных заболеваний, который основан на дуоденальном зондировании и использовании статистической теории распознавания образов. Этот метод имеет высокую точность и специфичность.

Приложение

Необходимые сведения из математики

Для понимания содержания книги достаточно знания элементарной математики в объеме средней школы. Поскольку в книге встречаются ссылки на факты математического анализа и линейной алгебры, опишем основные операции и обозначения, которые встречаются в тексте.

Суммирование набора чисел

$$x_1 + x_2 + \dots + x_n$$

обозначается как

$$\sum_{i=1}^n x_i.$$

Вектор $x = (x_1, x_2, \dots, x_n)$ — это упорядоченный набор из n чисел. Над векторами выполняются три операции:

- 1) сложение: $x + y = (x_1, x_2, \dots, x_n) + (y_1, y_2, \dots, y_n) = (x_1 + y_1, x_2 + y_2, \dots, x_n + y_n)$,
- 2) вычитание: $x - y = (x_1, x_2, \dots, x_n) - (y_1, y_2, \dots, y_n) = (x_1 - y_1, x_2 - y_2, \dots, x_n - y_n)$,
- 3) умножение на число: $\lambda x = (\lambda x_1, \lambda x_2, \dots, \lambda x_n)$.

Матрицей A называется прямоугольная таблица, состоящая из чисел a_{ij} , где номер строки i изменяется от 1 до n , а номер столбца j — от 1 до m :

$$A = \begin{Bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{Bmatrix}$$

Таким образом, матрица A состоит из n строк и m столбцов, которые являются векторами. Если количество строк равно количеству столбцов, то матрица называется квадратной. Приведем пример прямоугольных и квадратных матриц:

$$A_1 = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 1 & 2 & 4 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}.$$

Элементы с одинаковыми индексами a_{ii} образуют *главную диагональ* матрицы.

Над матрицами выполняются шесть операций: умножение на число, сложение, вычитание, умножение, транспонирование и обращение.

1. Умножение на число — каждый элемент матрицы умножается на одно и то же число:

$$\lambda A = \begin{Bmatrix} \lambda a_{11} & \lambda a_{12} & \dots & \lambda a_{1n} \\ \lambda a_{21} & \lambda a_{22} & \dots & \lambda a_{2n} \\ \dots & \dots & \dots & \dots \\ \lambda a_{m1} & \lambda a_{m2} & \dots & \lambda a_{mn} \end{Bmatrix}.$$

2. Сложение матриц — элементы матриц складываются попарно:

$$A + B = \begin{Bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{Bmatrix} + \begin{Bmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \dots & \dots & \dots & \dots \\ b_{m1} & b_{m2} & \dots & b_{mn} \end{Bmatrix} = \begin{Bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2n} + b_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \dots & a_{mn} + b_{mn} \end{Bmatrix}$$

3. Вычитание матриц — элементы второй матрицы попарно вычитаются из элементов первой матрицы:

$$A - B = \begin{Bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{Bmatrix} - \begin{Bmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \dots & \dots & \dots & \dots \\ b_{m1} & b_{m2} & \dots & b_{mn} \end{Bmatrix} = \begin{Bmatrix} a_{11} - b_{11} & a_{12} - b_{12} & \dots & a_{1n} - b_{1n} \\ a_{21} - b_{21} & a_{22} - b_{22} & \dots & a_{2n} - b_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} - b_{m1} & a_{m2} - b_{m2} & \dots & a_{mn} - b_{mn} \end{Bmatrix}$$

4. Умножение матриц: элементы первой матрицы строка за строкой умножаются на соответствующие элементы второй матрицы столбец за столбцом, а полученные произведения складываются:

$$AB = \begin{Bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{Bmatrix} \times \begin{Bmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \dots & \dots & \dots & \dots \\ b_{m1} & b_{m2} & \dots & b_{mn} \end{Bmatrix} = \begin{Bmatrix} \sum_{i=1}^n a_{1i} b_{i1} & \sum_{i=1}^n a_{1i} b_{i2} & \dots & \sum_{i=1}^n a_{1i} b_{in} \\ \sum_{i=1}^n a_{2i} b_{i1} & \sum_{i=1}^n a_{2i} b_{i2} & \dots & \sum_{i=1}^n a_{2i} b_{in} \\ \dots & \dots & \dots & \dots \\ \sum_{i=1}^n a_{mi} b_{i1} & \sum_{i=1}^n a_{mi} b_{i2} & \dots & \sum_{i=1}^n a_{mi} b_{in} \end{Bmatrix}.$$

Операция умножения матриц не является коммутативной:

$$AB \neq BA.$$

5. Транспонирование квадратной матрицы — элементы матрицы меняются местами симметрично относительно главной диагонали: первая строка становится первым столбцом, вторая строка — вторым столбцом и т.д.:

$$A^T = \begin{Bmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & \dots & a_{n2} \\ \dots & \dots & & \dots \\ a_{1m} & a_{2m} & \dots & a_{nm} \end{Bmatrix}$$

6. **Обращение квадратной матрицы.** Для квадратной матрицы A вводится обратная матрица A^{-1} , такая что произведение матрицы A на ее обратную матрицу равно единичной матрице:

$$AA^{-1} = A^{-1}A = E = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}.$$

У единичной матрицы на диагонали стоят единицы, а все остальные элементы равны нулю. Обратная матрица существует не у всех квадратных матриц, а лишь у тех, у которых так называемый *определитель* не равен нулю.

Кроме того, в тексте используются некоторые понятия математического анализа. В частности, обозначение $x_n \rightarrow x$ при $n \rightarrow \infty$ означает, что число x_n приближается к числу x при неограниченном увеличении индекса n .

Литература

1. *Andersson A.P., Gottfried J., Sondergaard K., Hau-Jensen K., Drzewiecki K.T.* Prognostiske faktorer af betydning for den recidivfri overlevelse efter kirurgisk behandling af hoved-hals-melanom // *Ugeskr. Laeger.* — 1993. — 155, № 31. — P. 2397–2399.
2. *Baccard M.* Melanomes des tumeurs a depister tot // *Tempo med.* — 1993. — № 501. — P. 29–31.
3. *Barjaktarovich N., Savage J.R.K.* RBE for α (42 MeV) Be neutrons based on chromosome-type aberrations induced in human lymphocytes scored in cells at first division // *Int. J. Radiat. Iol.* — 1980. — 37. — P. 667–675.
4. *Barth A., Cerny Th.* Neuere Aspekte in der Behandlung des malignen Melanoms // *Schweiz. med. Wochenshr.* — 1994. — 124, № 36. — P. 1592–1603.
5. *Brown L.D., Cai T.T., DasGupta A.* Interval Estimation for a Binomial Proportion // *Stat. Sci.* — 2001. — 2, № 16. — P. 101–133.
6. *Dharmadhikari S., Joag-dev K.* Unimodality, Convexity, and Applications. — New York: Academic Press, 1988.
7. *Dikicioglu E., Kandologlu G., Yoleri L., Kandologlu A.R.* Review of factors for predicting prognosis in cases with malignant melanoma of the skin: Abstr. 13th Eur. Congr. Pathol. (Ljubljana Sept. 1–6, 1991) // *Pathol. Res. and Prait.* — 1991. — 187, № 6. — P. 678.
8. *Djomina E., Klyushin D., Petunin Yu.* Ocena wplywu promieniowania na czestosc zachorowan na nowotwory zlosliwe u likwidatorow skutkow katastrofy w Czarnobylu // *Polski Przegląd Radiologii.* — 2001. — 66, № 2. — P. 62–63.
9. *Furukawa M., Nakagawa K., Hamada T.* Malignant melanoma // *Asian Med. J.* — 1994. — 37, № 1. — P. 55–58.
10. *Johnson N.L., Kotz S.* Some generalizations of Bernoulli and Polya-Eggenberger contagion models // *Statist. Paper.* — 1991. — 32. — P. 117.
11. *Kellerer A.M., Rossi H.H.* The theory of dual radiation action // *Curr. Top. Radiat. Res. Q., North Holland.* — 1972. — 8. — P. 85–158.
12. *Klyushin D.A., Petunin Yu.I., Savkina M.Yu., Andrushkiw R.I., Djomina E.A.* Statistical tests for comparing two probabilities and their application to cancer risk analysis // *Proc. Int. conf. of Mathematics and Engineering Techniques in Medicine and Biological Sciences (Las Vegas, Nevada (USA)).* — Las Vegas: CSREA Press, 2002. — 1. — P. 165–169.
13. *Klyushin D.A., Petunin Yu.I., Bairamov I.* Test of heterogeneity of general population // *Istatistik.* — 1999. — 1, № 3. — P. 19–29.
14. *Lloyd D.C., Edwards A.A.* Chromosome aberrations in human lymphocytes: Effect of Radiation Quality, Dose and Dose Rate // *Radiation-Induced Chromosome Damage in Man*, Alan. R. Liss, Inc. — New York, 1983. — P. 23–49.

15. *Lloyd D.C., Prosser J.S., Moquet J.E., Malowany D.J.* Doses in radiation accidents investigated by chromosome aberration analysis. XIII. A review of cases investigated. — Oxfordshire: Harwell, 1982. — National Radiological Protection Board, Report NRP13-R148. — 14 p.
16. *Mansson-Brahme E., Carstensen J., Erhardt K. et al.* Prognostic factors in the cutaneous malignant melanoma // *Cancer*. — 1994. — 73, № 9. — P. 2324–2332.
17. *Nieburg H.E., Herman B.E., Reisman H.* Buccal cell changes in patients with malignant tumours // *Lab. Invest.* — 1962. — 11, № 1. — P. 80–88.
18. *Obrapalska E., Cadel Z., Kostirka J.* Ocena cytologiczna na nablonka jamy ustnej u chorych na nowotwory zlosliwe // *Nowotwory*. — 1973. — 23, № 12. — P. 25–29.
19. *Ogden G.R., Cowpe J.G., Green M.W.* The effect of distant malignancy upon quantitative cytologic assessment of normal oral mucosa // *Cancer*. — 1990. — 65, № 3. — P. 477–480.
20. *Ordóñez G. A., Felin B. J., González B. M.* Aspectos controvertidos en el melanoma maligno cutáneo // *Med. clin.* — 1993. — 101, № 3. — P. 110–114.
21. *Petunin Yu.I., Kljushin D.A., Andrushkiw R.I., Ganina K.P., Boroday N.V.* Computer diagnosis of the breast cancer // Abstracts of the 14th IMACS World Congress on Computational and Applied Mathematics (Atlanta, USA, July 11–15, 1995). — P. 888–892.
22. *Pukelsheim F.* The three sigma rule // *Amer. Statist.* — 1994. — 48. — P. 88–91.
23. *Schulz H.* Hohe Erkennungsstraten mit Auflichtmikroskopie // *TW Dermatol.* — 1994. — 24, № 4. — P. 219–225.
24. *Sellke Th.* Generalized Gauss-Chebyshev inequalities for unimodal random variables // Technical Report №94-17. Department of Statistics. Purdue University. — 1994. — P. 1–11.
25. *Thorn M., Ponten F., Bergström R., Sparen P., Adami H.-O.* Clinical and histopathologic predictors of survival in patients with malignant melanoma population-based study in Sweden // *J. Nat. Cancer Inst.* — 1994. — 86, № 10. — P. 761–769.
26. *Villalobos M., Becerra D., Nunez M.I. et al.* Radiosensitivity of human breast cancer cell lines of different hormonal responsiveness. Modulatory effects of oestradiol // *Int. J. Radiat. Biol.* — 1996. — 70. — P. 161–69.
27. *Urist M.M., Karnell L.H.* The National Cancer Data Base: Report on melanoma // *Cancer*. — 1994. — 74, № 2. — P. 782–788.
28. *Yoo S., David H.* Revisiting Clopper-Pearson. — Technical Report 2002-05. Department of Statistics and Statistical Laboratory. — Iowa University, 2002.
29. *Yurchenko O.V., Todor I.N., Tryndyak V.P., Tregubova N.A., Kovtonyuk O.V., Solyanik G.I., Kulik G.I., Chekhun V.F.* Resistance of Guerin's carcinoma cells to cisplatin: biochemical and morphological aspects. // *Experim. oncology*. — 2003. — 25. — P. 64–68.
30. *Байрамов И., Петунин Ю.И.* Структура инвариантных доверительных интервалов, содержащих основную распределенную массу // *Теор. вероятностей и матем. статистика*. — 1990. — 35, № 1. — С. 15–26.

31. Байрамов И., Петунин Ю.И. Статистические критерии, основанные на обучающих выборках // Кибернетика. — 1991. — № 3. — С. 74–77.
32. Банченко Г.В. Сочетание заболеваний слизистой оболочки полости рта и внутренних органов. — М.: Медицина, 1979.
33. Барабой В.А., Петунин Ю.И., Ганина К.П. и др. Взаимосвязь между уровнем формирования вторичной радиорезистентности опухолевых клеток и количественными критериями нарастания их гетерогенности // Тез. докл. Третьего съезда по радиационным исследованиям (14–17 окт. 1997, Пушкино). — 1997. — 2. — С. 272.
34. Барьяхтар В.Г. Чернобыльская катастрофа. — Киев: Наук. думка, 1995.
35. Берталанфи Л. фон. История и статус общей теории систем // Системные исследования. Ежегодн. — М.: Наука, 1973.
36. Берталанфи Л. фон., Общая теория систем: критический обзор // Исследования по общей теории систем. — М.: Прогресс, 1969.
37. Бирхгоф Г. Теория решеток. — М.: Наука, 1984.
38. Битюцкий П.Г., Великанов В.Д. Использование математического метода для прогнозирования выбора объема оперативного вмешательства у больных раком гортани // Вопр. онкологии. — 1996. — 42, № 1. — С. 53–55.
39. Бугайов В.М., Лагутин А.Ю., Рогожин О.Г., Козак В.С. Изменения здоровья населения Украины вследствие Чернобыльской катастрофы. — Киев, 1996.
40. Булдаков Л.А., Гуськова А.К. 15 лет после аварии на Чернобыльской АЭС // Радиационная биология. Радиозэкология. — 2002. — 42, № 2. — С. 228–233.
41. Бурлакова Е.Б. Последствия Чернобыльской катастрофы: здоровье человека // Науч. совет по радиобиол. РАН. — М., 1996. — С. 148–182.
42. Ван дер Вандер Б.Л. Математическая статистика. — М.: ИЛ, 1960. — 434 с.
43. Вахтин Ю.Б., Пинчук В.Г., Швембергер И.Н., Бутенко З.А. Клонально-селекционная концепция опухолевого роста. — Киев: Наук. думка, 1992. — 246 с.
44. Высочанский Д.Ф., Петунин Ю.И. Обоснование правила 3σ для одномодальных распределений // Теор. вероятностей и матем. статистика. — 1980. — 21. — С. 25–36.
45. Ганина К.П., Лысюк Л.П., Ищенко В.Е., Кавецкий Р.Е. Цитоспектрофотометрическое изучение содержания ДНК в интерфазных ядрах буккальных мазков женщин, больных раком молочной железы // Вопр. онкологии. — 1976. — 22, № 11. — С. 74–77.
46. Ганина К.П., Налескина Л.А. Злокачественная меланома и предшествующие изменения кожи. — Киев: Наук. думка, 1991. — 167 с.
47. Ганина К.П., Петунин Ю.И., Тимошенко Я.Г. Количественные характеристики ядерного полиморфизма эпителиальных клеток при фиброаденоматозе и раке молочной железы // Докл. АН УССР. — 1989. — 35, № 12. — С. 1444–1449.

48. Ганина К.П., Полищук Л.З., Бородай Н.В., Налескина Л.А., Исакова Л.М., Несина И.П., Бучинская Л.Г. Цитологическая реактивность онкологического больного. — Киев, Наук. думка, 1995.
49. Ганина К.П., Бородай Н.В., Петунин Ю.И., Ключин Д.А. Количественная оценка опухоль-ассоциированных изменений ДНК ядер букального эпителия при фиброаденоме, фиброаденоматозе и раке молочной железы // Эксперим. онкология. — 1998. — 2. — С. 1304.
50. Ганина К.П., Зиневич А.К., Жеро С.В. Специальные методы исследования при предопухолевых и опухолевых процессах желудка. — Киев: Наук думка, 1988.
51. Ганина К.П., Полищук Л.З., Бородай Н.В., Налескина Л.А., Исакова Л.М., Несина И.П., Бучинская Л.Г. Цитологическая реактивность онкологического больного. — Киев: Наук думка, 1995.
52. Ганина К.П., Бородай Н.В., Петунин Ю.И., Ключин Д.А., Пилипенко В.Н. Компьютерная цитологическая диагностика в онкологии // Материалы 6-го Конгресса патологов Украины (30.09–02.10.98, Винница), 1998. — С. 229–231.
53. Гастева Г.Н., Сорокин А.В., Фролов Г.П. Клинико-дозиметрическая характеристика отклонений в состоянии здоровья для персонала плутониевого производства по результатам выборочного эпидемиологического обследования // Тез. докл. III съезда по радиац. исследованиям (Пущино, 1997). — М., 1997. — 1. — С. 177–178.
54. Гливенко В.И. Теория вероятностей. — М.: Гос. учебно-педаг. изд-во, 1937.
55. Гофман Д. Рак, вызываемый облучением в малых дозах: неформальный анализ проблемы. /Пер. с англ. под ред. Е.Б. Бурлаковой, В.Н. Лысцова. — М.: Наука, 1994. — Т.1. — 320 с.; Т.2. — 250 с.
56. Гудков И.Н. Клеточные механизмы пострadiaционного восстановления растений. — Киев: Наук. думка, 1985.
57. Гуськова А.К., Баранов А.Е., Барабанова А.Е. и др. Острые эффекты облучения у пострадавших при аварии на ЧАЭС // Мед. радиология. — 1987. — № 12. — С. 3–18.
58. Демиденко Е.З. Линейная и нелинейная регрессия — М.: Финансы и статистика, 1981. — 304 с.
59. Демина Э.А. Кривые доза-эффект при действии нейтронов со средней энергией // Радиобиология. — 1987. — 27, № 3. — С. 357–361.
60. Демина Э.А., Ганина К.П., Ключин Д.А., Петунин Ю.И. Распространенность заболеваний у ликвидаторов последствий Чернобыльской катастрофы в зависимости от дозы // Довкілля та здоров'я. — 1999. — 9, № 2. — С. 43–47.
61. Демина Э.А., Ключин Д.А., Петунин Ю.И., Савкина М.Ю. Определение калибровочных кривых для цитогенетических показателей на основе модели сплайновой регрессии при нейтронном воздействии на лимфоциты человека // Променева діагностика, променева терапія: 36. наук. робіт Асоціації радіологів України. — Київ, 2001. — Вип. 11. — С. 23–30.

62. Демина Э.А., Черниченко В.А., Чеботарев Е.Е., Гулько Г.М. Цитогенетические эффекты 22 МэВ нейтронов // Цитология и генетика. — 1990. — № 4. — С. 43–46.
63. Диткин В.А., Прудников А.П. Интегральные преобразования и операционное исчисление. — М.: Физматгиз, 1961.
64. Дьоміна Е.А. Вплив опромінення на виникнення злоякісних новоутворень у ліквідаторів аварії на ЧАЕС // Пробл. екол. та мед. генетики і клін. імунології. — 2001. — Вип. 1(33). — С. 68–75.
65. Дьоміна Е.А. Радиогенні цитогенетичні ефекти у учасників ліквідації аварії на Чорнобильській АЕС: автореф. дис. ... д-ра біол. наук. — Київ, 2002. — 36 с.
66. Завьялов Ю.С., Квасов Б.И., Мирошниченко В.Л. Методы сплайн-функций. — М.: Наука, 1980. — 352 с.
67. Зинченко В.А., Ганина К.П., Петунин Ю.И., Ключин Д.А., Барабой В.А., ШUTOва Е.Б. Количественная оценка изменчивости и гетерогенности культивируемых опухолевых клеток при формировании их радиорезистентности // Эксперим. онкология. — 1999. — 21. — С. 30–37.
68. Зінченко В.А., Солянік Г.І., Барабой В.А., Кулик Г.І. Особливості росту та зміни структури карциноми Герена в процесі набування радіорезистентності // Укр. радіол. журн. — 1995. — 3. — С. 364–365.
69. Иванков А.А., Цодиков А.Д., Якимова Т.П., Яковлев А.Ю. Параметрический анализ данных о рецидивировании рака молочной железы после радикального лечения // Вопр. онкологии. — 1996. — 42, № 3. — С. 85–87.
70. Иванков А.А., Цодиков А.Д., Яковлев А.Ю. Оценка пролиферативного потенциала рецидивирующей опухоли. Верификация параметрических методов средствами имитационного моделирования // Вопр. онкологии. — 1996. — 42, № 1. — С. 34–39.
71. Кендалл М. Дж., Стьюарт А. Теория распределений. — М.: Наука, 1966.
72. Кіндзельський Л.П., Дьоміна Е.А., Ганіна К.П., Ключин Д.А., Петунін Ю.І. Вплив радіаційного фактора на частоту злоякісних новоутворень у ліквідаторів Чорнобильської катастрофи // Укр. радіол. журн. — 1999. — № 1. — С. 35–36.
73. Ключин Д.А., Демина Э.А., Петунин Ю.И., Ганина К.П. Малые дозы ионизирующей радиации как фактор риска возникновения злокачественных новообразований у ликвидаторов последствий аварии на ЧАЭС // Променева діагностика, променева терапія: 36. наук. робіт Асоціації радіологів України. — Київ, 2001. — Вип. 9. — С. 94–100.
74. Ключин Д.А., Петунин Ю.И. Непараметрический критерий эквивалентности генеральных совокупностей, основанный на мере близости между выборками // Укр. мат. журн. — 2003. — 55, № 2. — С. 147–163.
75. Комар В.Е. Современное состояние проблемы биологической индикации лучевых поражений // Радиобиология. — 1992. — 1. — С. 84–96.

76. Комаров Ф.И., Галкин В.А., Иванов А.И., Максимов В.А. Сочетанные заболевания органов дуоденохоледохопанкреатической зоны // Метод. рекомендации. — М.: Медицина, 1983. — 256 с.
77. Комаров Ф.И., Гребнев А.Л. (ред.) Руководство по гастроэнтерологии. — М.: Медицина, 1996. — Т.3. — 720 с.
78. Корнейчук Н.А. Сплайны в теории приближения. — М.: Наука, 1984. — 352 с.
79. Крамер Г. Математические методы статистики. — М.: Мир, 1975. — 648 с.
80. Крамер Г., Лидбеттер М. Стационарные случайные процессы. — М.: Мир, 1969.
81. Крутова Т.В., Корман Д.Б., Потапов Н.Н., Пашкова В.С. Усиление пролиферативной активности клеток опухолей легкого человека при культивировании в диффузионных камерах // Вопр. онкологии — 1984. — 30(8). — С. 91–95.
82. Кузина Г.И., Бирюкова Т.И., Милешко О.И и др. Интраоперационная цитологическая диагностика рака щитовидной железы // Тез. докл. VI научно-практ. конф. онкологов Республики Беларусь (16–18 ноября 1994). — Минск: 1994. — С. 657.
83. Лоев М. Теория вероятностей. — М.: Изд-во иностр. лит., 1964.
84. Мадреймов И., Петунин Ю.И. Характеризация равномерного распределения с помощью порядковых статистик // Теор. вероятностей и матем. статистика. — 1982. — 27. — С. 96–102.
85. Майер Н.Н., Цодикова Л.Б. Пункционная цитологическая диагностика заболеваний щитовидной железы: возможности и ограничения метода // Арх. патологии. — 1996. — 58(2). — С. 748.
86. Матвейчук С.А., Петунин Ю.И. Обобщение схемы Бернулли, возникающее в вариационной статистике. I // Укр. матем. журнал. — 1991. — 43(4). — С. 518–528.
87. Матвейчук С.А., Петунин Ю.И. Обобщение схемы Бернулли, возникающее в вариационной статистике. II // Укр. матем. журнал. — 1991. — 43(6). — С. 779–785.
88. Месарович М. Основания общей теории систем // Общая теория систем. — М.: Мир, 1966.
89. Моисеенко В.В. Применение методов биофизического моделирования для ретроспективной оценки доз по хромосомным aberrациям в различные сроки после облучения: автореф. дис. ... канд. биол. наук. — Обнинск, 1993.
90. Москалев Ю.И. Отдаленные последствия воздействия ионизирующих излучений. — М.: Медицина, 1991.
91. Нестеров Е.Б., Дикарев В.Г., Дикарева Н.С., Гераськин Р.А. Индукция цитогенетических эффектов в корневой меристеме облученных при разных мощностях дозы проростков ярового ячменя // Тез. докл. конф. “Проблемы радиационной генетики на рубеже веков” (М., 20–24 ноября 2000). — М.: Изд-во Рос. ун-та Дружбы народов, 2000. — С. 169.
92. Норми радіаційної безпеки України. — К.: М-во охорони здоров'я, 1997. — 122 с.

93. *Обажуров Г.М., Соколов В.А., Цыб Т.Г. и др.* Состояние и перспективы исследований в радиационной биофизике и нейтронной радиобиологии // Медрадиология. — 1992. — № 9-10. — С. 27-33.
94. *Олейник В.А., Безверхая Т.П., Эпштейн Е.В., Божок Ю.М.* Диагностика рака щитовидной железы // Пробл. эндокринологии. — 1995. — 5. — С. 37-41.
95. *Паршикова С.М., Рагимов Ч.Р.* Прогностическое значение морфологических признаков меланомы кожи головы и шеи // Арх. патологии. — 1985. — 47, № 9. — С. 74-80.
96. *Петунин Д.А., Ключин Д.А.* Структурний підхід до розв'язання шостої проблеми Гільберта I, II // Журнал обчисл. та прикл. матем. — № 89. — С. 80-90.
97. *Петунин Д.А., Ключин Д.А.* Структурний підхід до розв'язання шостої проблеми Гільберта II // Журнал обчисл. та прикл. матем. — № 90. — С. 83-88.
98. *Петунин Ю.И.* Приложение теории случайных процессов в биологии и медицине. — Киев: "Наук. думка", 1981.
99. *Петунин Ю.И., Ключин Д.А., Кулик Г.И., Юрченко О.В., Тодор И.Н., Чехун В.Ф.* Стратификационный анализ морфометрических показателей раковых клеток с фенотипом лекарственной резистентности // Кибернетика и системный анализ. — 2005. — № 6. — С. 158-167.
100. *Петунин Ю.И., Ключин Д.А., Демина Э.А., Савкина М.Ю.* Оценка влияния величины дозы облучения на вероятность развития злокачественных новообразований на основе сплайновой регрессии // Кибернетика и системный анализ. — 2006. — № 3. — С. 168-176.
101. *Петунин Ю.И., Матвейчук С.А.* Проверка гипотез с помощью статистических критериев, использующих процедуру принятия решений // Докл. АН России. — 1994. — 336, № 3. — С. 301-303.
102. *Петунин Ю.И., Рублев Б.В.* Распознавание образов с помощью квадратичных дискриминантных функций // Вычисл. и прикл. математика (в печати).
103. *Петунин Ю.И., Семейко Н.Г.* Случайные точечные процессы с независимым маркированием // Докл. АН СССР. — 1986. — 288, № 4. — С. 823-827.
104. *Петунин Ю.И., Тимошенко Я.Г., Петунина М.Ю.* Критерий для идентификации генеральной совокупности в случае конечного класса альтернативных гипотез // Докл. АН УССР. Сер. А. — 1984. — № 6. — С. 29-32.
105. *Присяжнюк А.Е., Грищенко В.Г., Федоренко З.П.* Эпидемиологическое изучение злокачественных новообразований у пострадавших вследствие аварии на Чернобыльской АЭС. Итоги, проблемы и перспективы // Междунар. журн. радиац. медицины. — 1999. — № 2. — С. 42-50.
106. *Пяткин Е.К., Баранов А.Е. и др.* Оценка дозы и равномерности облучения при острых радиационных поражениях человека с помощью анализа аберраций хромосом. Методические рекомендации. — М., 1988. — 26 с.
107. *Пяткин Е.К., Баранов А.Е., Филлюшкин И.В., Нугис В.Ю., Селидовкин Г.Д., Кончаловский М.В., Чирков А.А.* Оценка дозы и равномерности облучения при острых радиационных поражениях человека с помощью анализа аберраций хромосом. Методические рекомендации. — Москва, 1988. — 25 с.

108. *Рагимов Ч.Р.* Меланома кожи головы и шеи. Изучение факторов прогноза: автореф. дис. ... канд. мед. наук. — М., 1984. — 23 с.
109. Розповсюдженість злоякісних новоутворень в популяції України в 1991–1996 рр. (епідеміологічні та організаційні аспекти проблеми) / Під ред. С.О. Шалімова. — Київ, 1997.
110. Руководство по изучению генетических эффектов в популяциях человека. — ВОЗ: Женева, 1989. — 122 с.
111. *Савкина М.Ю.* Свойства оценки точки перехода случайного процесса метода минимизации остаточной суммы квадратов // Тез. докл. Шестой Крымской межд. матем. школы МФЛ. (Симферополь, 2002). — С. 125.
112. *Себер Дж.* Линейный регрессионный анализ. — М.: Мир, 1980. — 456 с.
113. *Севаньякаев А.В.* Закономерности возникновения аберраций хромосом в митотическом цикле клеток человека при гамма- и нейтрогенном облучении: автореферат дис. д-ра наук. — Обнинск, 1982. — 36 с.
114. *Севаньякаев А.В.* Радиочувствительность хромосом в митотическом цикле. — М.: Энергоатомиздат, 1987.
115. *Севаньякаев А.В., Деденков А.Н.* Актуальные проблемы современной радиобиологии в свете оценки и прогнозирования последствий аварии на Чернобыльской АЭС // Радиобиология. — 1990. — 30, № 5. — С. 579–584.
116. *Севаньякаев А.В., Насонов А.П.* Биологическая дозиметрия по хромосомным аберрациям в культуре лимфоцитов человека. Методические рекомендации. — Обнинск, 1979.
117. *Сердюк А.М., Бобылева О.А.* Чернобыль и здоровье населения Украины // Тр. 2-й Междунар. конф. “Отдаленные медицинские последствия Чернобыльской катастрофы” (Киев, 1–6 июня 1998 г.). — Киев: Чернобыльинтеринформ, 1998. — С. 132.
118. *Соляник Г.И., Коненко М.М., Чехун В.Ф., Кулик Г.И.* Изменение роста карциномы Герена при возникновении резистентности к действию цисплатина и тиофосамида // Эксперим. онкология. — 1992. — 14. — С. 68–72.
119. *Сусков И.И.* Мониторинг хромосомных аберраций как биоиндикаторов в оценке повреждения низкими уровнями радиации здоровья человека. // Пробл. радиационной генетики на рубеже веков. Тез. докл. Междунар. конф. (М., 2000). — М.: Изд-во Рос. Ун-та Дружбы народов. — С. 339.
120. *Теплоухова И.М., Введенская Н.А., Середа Г.Н., Арион А.И.* Подходы к прогнозированию при раке предстательной железы // Вопр. онкологии. — 1996. — 42, № 2. — С. 12–16.
121. *Тимошенко Я.Г., Петунин Ю.И., Ганина К.П., Суптело Л.П.* Анализ гетерогенности цитоспектрофотометрических и морфометрических показателей при гиперпластических процессах и раке желудка // Докл. АН УССР. Сер. В. — 1984. — № 6. — С. 75–78.
122. *Феллер В.* Введение в теорию вероятностей и ее приложения: В 2-х томах. — М.: Мир, 1964.

123. Флетчер С., Флетчер Р., Вагнер Э. Клиническая эпидемиология. Основы доказательной медицины. — М.: Медиа Сфера, 1998.
124. Харченко Е.П. Использование модели нейронной сети для прогнозирования рака на основе иммунологических показателей // Бюл. экспер. биол. и мед. — 1996. — 122, № 8. — С. 206–208.
125. Хусаинова И.С., Варзулева И.Ю., Копсина Н.А. Оценка цитологических показателей для диагностики функционального состояния человека. // Клини. лабор. диагностика. — 1997. — № 3. — С. 102.
126. Циммерман Я.С. Хронический панкреатит: диагностика и лечение. Метод. рекомендации. — Пермь, 1990. — 66 с.
127. Чазова Н.Л., Перчук Б.Д., Гольбурт Н.Н. и др. Некоторые спорные вопросы верификации и классификации опухолей щитовидной железы // Арх. патологии. — 1997. — 59, № 1. — С. 336.
128. Чеботарев Е.Е., Демина Э.А. Поражаемость хромосом лимфоцитов человека в различных стадиях митотического цикла при действии быстрых нейтронов со средней энергий 6 МэВ // Цитология и генетика. — 1985. — 19, № 4. — С. 260–263.
129. Шевченко В.А., Сьомов А.Б., Иофа Э.П., Акаева Э.А. Дозовая зависимость индукции хромосомных aberrаций у ликвидаторов Чернобыльской аварии // Радиационная биология. Радиоз экология. — 1994. — № 6. — С. 865–870.
130. Эйдус Л.Х. Общая схема модификации лучевого поражения клеток // Механизмы лучевой патологии / Под ред. Ю.Б. Кудряшова. — М.: Изд-во МГУ, 1984. — С. 128–135.
131. Яворовски З. Жертвы Чернобыля: реалистическая оценка медицинских последствий Чернобыльской аварии // Мед. радиология и радиационная безопасность. — 1999. — 44, № 1. — С. 19–30.

Д.А. Ключин Ю.И. Петунин

Доказательная медицина

ПРИМЕНЕНИЕ СТАТИСТИЧЕСКИХ МЕТОДОВ

В биологии и медицине утверждения, касающиеся исследуемых объектов, неизбежно носят вероятностный характер и являются следствием многочисленных и повторяющихся эмпирических наблюдений и экспериментов. В такой ситуации единственным научным инструментом, обеспечивающим непротиворечивость выводов, является статистический анализ.

Концепция доказательной медицины позволяет найти ответы на важнейшие вопросы: "Результатам каких методов диагностики следует доверять? Какие методы диагностики и лечения можно рекомендовать для клинической медицины и практического здравоохранения?" На ее основе можно сравнивать эффективность действия фармакологических препаратов без учета механизмов их воздействия, лишь на основании клинических испытаний, проводимых при исследовании новых препаратов и предоставляющих исходную информацию для статистических исследований.

Цель этой книги — доказать сомневающимся продуктивность тесного сотрудничества медиков и математиков на всех этапах исследований, осуществляемого при четком разграничении сфер компетенции.

