Министерство образования Российской Федерации
Ярославский государственный университет им. П.Г. Демидова

А.Ю. Левин, В.В. Майоров, М.Л. Мячин

О логике математической статистики

Текст лекций по курсу
«Дополнительные главы математической статистики»

Второе издание
переработанное и дополненное

Ярославль
2003

ББК В182я73
Л36
УДК 621

Левин А.Ю., Майоров В.В., Мячин М.Л.
О логике математической статистики: Текст лекций по курсу «Дополнительные главы математической статистики». / Яросл. гос. ун-т. 2-е изд., перераб. и доп. Ярославль, 2003. 44 с.

ISBN 5-8397-0256-О

На простейших примерах рассматривается логика применения статистических и вероятностных рассуждений. Этот опыт часто вызывает затруднения у студентов при первоначальном знакомстве с курсом математической статистики. Позднее непонимание у повзрослевших студентов (специалистов) возводит математическую статистику в ранг своего рода «шаманства», что совершенно не соответствует сути данной науки.

Рис. 1. Табл. 1. Библиогр.: 7 назв.

Рецензенты: кафедра математического анализа Ярославского государственного педагогического университета им. К.Д. Ушинского; д-р пед. наук, проф. В.В. Афанасьев.

ISBN 5-8397-0256-О

Аннотация к первому изданию

Опыт показывает, что специфическая природа статистических выводов вызывает трудности у студентов при первом знакомстве с предметом. Цель данного текста лекций — помочь преодолеть эти трудности. Основное внимание уделено проверке статистических гипотез и правильной интерпретации таких понятий, как практическая достоверность и т.п.

Предисловие ко второму изданию

Страницы книг первого издания [1] уже давно пожелтели, а само издание стало раритетным. В то же время, несмотря на малый объем, книга зарекомендовала себя как удачное введение в математическую статистику.

Нам хотелось бы выразить искреннюю признательность магистрантам факультета информатики и вычислительной техники ЯрГУ им. П.Г. Демидова, принявшим активное участие в подготовке оригинал-макета. Особую благодарность хотелось бы принести А.П. Янкевичу и А.А. Зинякову, которые постоянно указывали нам на ошибки, закравшиеся при верстке.

Второе издание дополнено критерием А.Ю. Левина сравнения векторных выборок на однородность (в изложении второго автора) и разделом «О датчиках случайных чисел». Техническая поддержка книги осуществляется в Интернете по адресу {`http://ltwood.hotbox.ru/statbook`}.

Лекция 1

Математическая статистика является обширным разделом математики, включающим в себя множество проблем, более или менее связанных между собой. Приблизительно говоря, эту совокупность проблем можно характеризовать как методику выявления закономерностей в экспериментальных данных. Некоторые новейшие разделы статистики (планирование эксперимента) включают в себя также и методику организации эксперимента. Таким образом, статистические идеи могут применяться не только после сбора экспериментальных данных, а еще на стадии организации исследований.

Излагаемый ниже материал относится, в основном, к разделу классической математической статистики — проверке статистических гипотез. Попутно мы коснемся также (в силу необходимости) вопроса об оценке неизвестных параметров. Настоящее пособие не претендует на сколько-нибудь полное изложение основных результатов теории проверки гипотез (такое изложение заняло бы несколько томов). Наша цель значительно скромнее — описать некоторые идеи и методы рассуждений, характерных для этой области. Для иллюстрации мы будем прибегать к примерам, носящим иногда не слишком серьезный характер, впрочем, сами по себе обсуждаемые идеи весьма серьезны и важны.

1.1 Что значит знать?

Достоверность и невозможность на различных уровнях

Теория вероятностей и математическая статистика используют некоторый особый тип рассуждений, практически не встречающийся в других областях математики. Ввиду важности этого вопроса остановимся на нем подробнее. Дело заключается в том, какие события считать достоверными или, наоборот, — невозможными.

Простейшим примером достоверного события является факт, содержащийся в геометрической теореме: сумма углов треугольника равна 180 градусам (геометрия, естественно, евклидова). Как мы знаем, какое бы множество треугольников ни взять, измеряя углы любого треугольника, мы всегда получим указанную сумму углов (пренебрегая погрешностями). Отличие суммы углов от 180 градусов, следовательно, является невозможным событием. Здесь мы имеем дело с абсолютной достоверностью и невозможностью на логическом уровне. Однако откуда берется наша уверенность в непререкаемой истинности данного утверждения? Этот факт имеет математическое доказательство. Именно: он с помощью вспомогательных построений и логических рассуждений выводится из некоторых первичных утверждений, называемых аксиомами, истинность которых постулируется априори (либо на основе «абсолютной очевидности», либо на основе «общей договоренности»). Кстати, подобная ситуация характерна для большинства математических утверждений.

Однако, в принципе, возможны совершенно другие подходы к обоснованию истинности утверждений. Представим себе, например, что некто доказывает теорему о сумме углов треугольника следующим образом. Он рассматривает достаточно большое количество треугольников (может, тысячи, а может, миллионы), взятые наудачу, экспериментально убеждается, что сумма углов каждого из них 180 градусов, и делает вывод, что так обстоит дело всегда. Для сомневающихся предлагается повторить эксперимент. Что же можно сказать о таком обосновании? Является ли оно правомерным? Само собой, способ действия выглядит нерациональным, даже глупым и математически несостоятельным, так как требует большой работы, с одной стороны, с другой же стороны, никогда не дает полной уверенности. То же самое следует из аксиом сразу же и бесспорно.

Казалось бы, «экспериментальный» способ нужно забраковать как неразумный и ненадежный. Внутри математики это действительно так. За пределами же чистой математики ситуация меняется. Здесь уместно напомнить, что великие геометры (Ф. Гаусс и Н.И. Лобачевский) экспериментально, очень тщательно измеряли суммы углов треугольников, образованных материальными объектами (ставилась под сомнение справедливость евклидовых аксиом для физического мира). Вообще, для естествознания характерен именно экспериментальный способ установления истины. Логический (дедуктивный) способ доказательства должен базироваться на аксиомах. Роль этих аксиом в естествознании играют законы природы. Основная задача — найти их. Естественно, когда законы уже известны, то следствия могут быть получены логическими рассуждениями и математическими выкладками, без дополнительных экспериментов.

Таким образом, в естествознании нельзя руководствоваться тем понятием достоверности, которое характерно для чистой математики. Как правило, закон считается установленным, если он проявляется в большом количестве экспериментов. Общеизвестным примером могут служить ньютоновские законы механики, которые ниоткуда не были «выведены», но, будучи взятыми в качестве аксиом, позволяли объяснить огромное количество экспериментальных фактов. В силу прекрасного совпадения с экспериментальными фактами, законы эти общепризнаны. Насколько надежна истина, установленная таким способом? Естественно, здесь нет и не может быть абсолютной логической убежденности. Более того, известно, что законы эти несправедливы в мире больших скоростей, где действуют другие (релятивистские) правила. Следовательно, факты, которые в течение долгого времени казались достоверными, таковыми, строго говоря, не были. Ясно, что такого рода ситуация может повториться и с любыми другими истинами в, казалось бы, «конечной инстанции». Таким образом, у нас нет абсолютно достоверного надежного формально-логического способа проверки предполагаемых закономерностей. Возникает необходимость широко использовать экспериментальный метод, который является, тем самым, основным.

Остановимся подробнее на способе экспериментального обоснования какой-либо закономерности. Что лежит в основе этой процедуры?

Мы говорим, что закономерность правильно выявлена, если она объясняет результаты всех экспериментов, относящихся к данной области (чем больше экспериментальных фактов объясняется, тем надежнее обоснование). Возникает вопрос, а нельзя ли согласование теории и экспериментов объяснить простым *совпадением*? Разумеется, логически такая возможность никогда не исключена (измерение суммы углов миллиарда треугольников не доказывает теорему о сумме углов). Суть дела, однако, в том, что, когда закономерность объясняет большое число различных экспериментов, *вероятность случайного совпадения ничтожно мала*.

Итак, вероятностные соображения неявно проходят через все естествознание. Понятие вероятности при этом, как правило, не упоминается, тем более не ставится вопрос о ее вычислении[1]. Однако, подчеркнем, что все развитие естествознания требует от нас считать маловероятные события практически невозможными. Впрочем, поскольку понятие вероятности здесь не является четким и значение вероятности не может быть оценено количественно, вряд ли здесь уместно говорить о теории вероятностей как таковой.

Теория вероятностей начинается тогда, когда имеется возможность хотя бы в принципе точно определить вероятность. Переходя к вопросам этого цикла, мы должны воспользоваться конкретной моделью. Традиционным является бросание монеты, не будем отступать от этой традиции и мы.

Итак, предположим, что подбрасывается симметричная монета, т.е. герб и цифра при каждом бросании выпадают с вероятностью $1/2$ (результаты отдельных бросаний, как обычно, предполагаются независимыми). Предположим, что монета бросается неограниченно много раз. Возможно ли, чтобы всегда выпадал

[1]Здесь не имеются в виду такие области знания, как статистическая физика и квантовая механика, где сами законы имеют четко выраженную вероятностно-статистическую природу.

герб (такие «мелочи», как ограниченность человеческой жизни при такого рода опытах, к делу не относятся)? Ясно, что логическая возможность интересующего нас случая налицо. Однако, элементарные рассуждения показывают, что вероятность данного события равна нулю (очевидно, данная вероятность меньше, чем $1/2^n$ для любого n). Итак, возможно или невозможно *всегда* видеть после бросания монеты герб? Как видим, само понятие «возможности» не является четким — его можно понимать по-разному.

На логическом уровне — это возможно. На логическом уровне невозможны лишь варианты, которых нет (например одновременное выпадение герба и цифры).

В вероятностном смысле невозможным (практически) является любое событие, имеющее нулевую вероятность. В этом смысле упомянутое событие невозможно.

Третий уровень, который для статистики наиболее значим: событие практически невозможно, если его вероятность достаточно мала. Здесь есть свободный параметр — малость вероятности, значение которого определяется смысловым содержанием каждой конкретной задачи и до некоторой степени нашим субъективным выбором. О третьем уровне мы будем говорить ниже весьма подробно.

Проанализируем понятие невозможности на втором уровне. Может показаться, что сам описанный эксперимент — бесконечное бросание монеты — является настолько искусственным, что не представляет никакого интереса. Тем более, что для его проведения требуется вечность. Поэтому можно предложить другие, гораздо «более быстрые» эксперименты той же природы, важность которых очевидна. Именно: пусть ξ — непрерывная случайная величина, распределенная, например, по нормальному закону, а C — некоторая постоянная. Возможно ли событие $\xi = C$? Ответ здесь, в принципе, ничем не отличается от рассмотренного выше случая: событие логически возможно (нормально распределенная величина, напомним, может принимать любые значения), но вероятность равна нулю. Стоит обратить внимание еще на один трудный для понимания парадокс. С одной стороны, мы знаем, что в результате эксперимента случайная величина примет какое-то конкретное значение, например C. С другой стороны, событие $\xi = C$ практически невозможно. Получается, что в результате эксперимента обязательно произойдет невозможное событие. Создается впечатление, что это противоречит здравому смыслу (с логически невозможными событиями, естественно, такого не бывает). Момент действительно непростой как для понимания, так и для объяснения. Когда мы говорим: событие $\xi = C$ невозможно, предполагается, что константа C задана *apriori*. Какое бы конкретное C ни было взято, нет шансов, что ξ примет значение C. Речь идет о полном совпадении всех десятичных цифр записи числа — первой, второй... миллионной и так далее.

Кстати, с физической точки зрения сама возможность измерения величины со всеми значащими цифрами (за исключением некоторых весьма частных случаев, например, целочисленных величин) совершенно исключена. Главное, правда, не в этом. Даже если мы договоримся считать числа одинаковыми, если совпадают, например, первые 10 цифр, то вероятность приближенного равенства $\xi = C$ уже равна не нулю, а некоторому малому положительному числу. Однако такие события с пренебрежимо малой вероятностью зачастую считаются тоже практически невозможными. Как часто говорят статистики: «Маловероятные события не происходят». Это и есть третий уровень невозможности.

Любопытно, что рассматриваемый парадокс почти никогда не затрагивается в учебниках по математической статистике, хотя очевидно, что для ясного понимания сути дела разобраться в вопросе необходимо. Возможно, ситуация прояснится, если не рассматривать ее в отвлеченных терминах, а обратиться к конкретному примеру. Мы заимствуем его из книги известного математика и педагога Д. Пойа «Математика и правдоподобные рассуждения». Д. Пойа приводит следующий эпизод, описанный в труде Бертрана «Исчисление вероятностей».

1.2 Случай в Неаполе

Нуль-гипотеза и альтернативные гипотезы

> *Однажды в Неаполе преподобный Галиани увидел человека из Базиликаты, который, встряхивая три игральные кости в чашке, держал пари, что выбросит три шестерки; и действительно, он немедленно получил три шестерки. Вы скажете: такая удача возможна. Однако человеку из Базиликаты это удалось во второй раз, и пари повторилось. Он клал кости назад в чашку три, четыре, пять раз и каждый раз выбрасывал три шестерки. «Sangue di Bacco (кровь Вакха), вскричал преподобный, — кости налиты свинцом». И так оно и было. Но почему преподобный воспользовался нечестивым выражением?*

Надо сказать, что в этой забавной истории как в капле воды отражается специфика вероятностных рассуждений.

Далее Д. Пойа приводит следующий комментарий:

> *Преподобный Галиани вывел правдоподобные заключения очень важного типа. Если он внезапно для себя открыл этот важный тип правдоподобного заключения, то его возбуждение вполне понятно, и я лично не упрекнул бы его за эти нечестивые выражения.*
>
> *Правильно будет к каждому относиться как к джентльмену, если нет доказательств противного... Я не сомневаюсь, что преподобный поступил правильно и сначала допустил, что человек из Базиликаты имел честные кости и пользовался ими честно. Такое допущение, правильно сформулированное в терминах вероятности, является статистической гипотезой. Статистическая гипотеза вообще вводит (допускаемые ею) значения некоторых вероятностей. Так, преподобный сначала в более или менее явной форме принял, что на любой из трех костей шесть очков будут выпадать с вероятностью* $1/6$... *Исчисление вероятностей позволяет нам вычислить искомые верояности по данным вероятностям на основании данной статистической гипотезы. Так, на основании статистической гипотезы, принятой преподобным сначала, мы можем*

вычислить вероятность выпадения на трех костях трех шестерок: она равна $1/216$ *— довольно малая вероятность. Вероятность повторения этого фокуса два раза, то есть выпадения трех шестерок и в первом и во втором испытании равна* $1/6^6 = 1/46656$*; действительно, это очень маленькая вероятность. Однако, человек из Базиликаты повторил эту необычную вещь пять раз. Вероятности повторения того же в трех, четырех и пяти испытаниях равны соответственно:* $1/6^9$*,* $1/6^{12}$*,* $1/6^{15} \approx 2 \cdot 10^{-12}$.

Возможно, преподобный принял свое первоначальное допущение из простой вежливости: глядя на человека из Базиликаты, он, быть может, имел сомнения относительно честности костей. Преподобный сохранял молчание, пока три шестерки выпали последовательно дважды, событие, которое при исходном допущении должно было бы произойти не намного чаще, чем раз за пятьдесят тысяч испытаний. Он сохранял молчание даже дольше. Однако когда события становились все более и более невероятными, достигли и, возможно, превзошли ту степень невероятия, которую люди рассматривают как сверхъестественную, преподобный потерял терпение, вывел свое заключение, отказался от первоначального допущения и энергично высказался.

Анекдот, который мы только что обсудили, интересен только в одном отношении: он типичен. Он ясно показывает обстоятельства, при которых мы можем разумно отказаться от статистической гипотезы.

Авторы разделяют, естественно, мнение крупного математика и педагога Д. Пойа, однако уместны некоторые уточнения и дополнения.

Во-первых, исходная гипотеза преподобного Галиани состояла не только в том, что вероятность выпадения шестерки на каждой кости составляла $1/6$, но и в независимости отдельных бросаний (по-видимому, Д. Пойа имел это в виду, иначе нельзя было бы пользоваться теоремой умножения).

Во-вторых, то обстоятельство, на котором мы остановимся более подробно, представляется весьма важным, хотя Д. Пойа его не затрагивает. Для прояснения сути ограничимся двумя испытаниями: так как $1/6^6$ — достаточно малое число, то были все основания отклонить первоначальную гипотезу после двух опытов.

Итак, вкратце схема рассуждений в данном случае выглядит следующим образом. Допустим, что исходная гипотеза (как ее называют, нуль-гипотеза) верна. Тогда мы вынуждены признать, что произошло событие, имеющее вероятность $1/6^6$, то есть «практически невозможное». Получено противоречие, которое позволяет исключить нуль-гипотезу. Надо сказать, что такой примитивный аналог доказательства от противного является слишком упрощенным и, по существу, несостоятельным. Легко понять, что таким способом можно опровергнуть чуть ли не все, что угодно. Поясним это.

Будем считать, что три игральные кости занумерованы и результаты двух бросаний, к примеру, таковы: $(6, 2, 5)$ и $(4, 1, 4)$. Легко видеть, что вероятность этого ничем не примечательного результата при условии, что нуль гипотеза о честности костей справедлива, равна *той же самой малой величине* $1/6^6$. Значит, и в этом случае, казалось бы, можно опровергнуть нуль-гипотезу по тем же самым соображениям. Другими словами, при любом результате отклоняется предположение о «честности» костей, что явно нелепо.

Что же характерно для данной ситуации? В наличии имеется огромное количество вариантов (в данном случае 6^6). При опыте один из них обязательно реализуется, хотя вероятность любого из них весьма мала (при нулевой гипотезе $1/6^6$)[2].

Характерно то, что мы конкретизируем исход эксперимента лишь после его окончания. Какой конкретный маловероятный вариант реализуется в опыте, заранее неизвестно, мы его определяем апостериорно. Все это, вместе взятое, делает приведенные нами выше рассуждения «от противного» несостоятельными. Наблюдаем пример абсолютно ошибочного применения правила — «маловероятные события не происходят», которое верно при разумном его использовании.

Покажем, как разумно действовать в нашем примере. Заметим сначала, что выдвижение нуль-гипотезы обычно означает утверждение об отсутствии какого-то определенного эффекта, наличие которого проверяется. В «Неапольском деле» эффект заключался в нечестности хозяина костей, стало быть, нуль-гипотеза преподобного предполагала честность хозяина и костей. Самое простое — предположить, что кости симметричны:

$$\mathrm{P}(1) = \mathrm{P}(2) = \ldots = \mathrm{P}(6) = 1/6,$$

где $\mathrm{P}(K)$ — вероятность выпадения на кости K очков. Для определенности мы принимам эту гипотезу, хотя возможны и другие варианты. Например: $\mathrm{P}(6) \neq 1/6$.

Если нулевая гипотеза неверна и в «Неапольском деле» имеет место жульничество, то естественно предположить, что вероятность выпадения шести очков на каждой из костей не есть $1/6$, а существенно ближе к единице. Тем самым, наряду с нуль-гипотезой появляется *альтернативная гипотеза* (о большой вероятности выпадения шестерок). Альтернативная гипотеза должна определять нашу стратегию при проверке нуль-гипотезы.

Такая стратегия (решающее правило, выбор критической статистики и т.п.) должна формулироваться до проведения эксперимента. Стратегия чаще всего сводится к тому, что выбирается некая случайная величина ξ (критическая статистика) и проверяется, превзошло ли ее значение в результате эксперимента определенный порог h_0. Если $\xi > h_0$, то нуль-гипотеза отклоняется, в противном случае принимается, что экспериментальные данные не позволяют отклонить нуль-гипотезу.

В «Неапольском деле» напрашивается в качестве ξ взять количество выпавших шестерок. Если при двух бросаниях их выпадет «слишком много», то можно сделать обоснованный вывод, что владелец костей — жулик. Как же конкретизировать понятие «слишком много», то есть определить пороговый уровень h_0?

[2] Мы здесь, очевидно, вернулись к упомянутому ранее парадоксу: одно из множества практически невозможных событий заведомо происходит.

Это зависит от выбираемого заранее уровня значимости ε; события с вероятностью меньше ε следует считать практически невозможными. Хотя в выборе ε, конечно, есть произвол, общепринятыми («каноническими») являются несколько значений, из которых упомянем $\varepsilon = 0.05$ и $\varepsilon = 0.01$. Ясно, что чем меньше ε, тем мы более осторожны при отбрасывании нуль-гипотезы. Преподобный Галиани был архиосторожным человеком, специалисты в области статистики отклоняют гипотезы обычно значительно решительнее. Пусть в рассматриваемом примере $\varepsilon = 0.01$. Найдем соответствующее пороговое значение h_0. Элементарные вычисления показывают, что

$$\mathrm{P}(\xi \geqslant 3) = \mathrm{P}(\xi = 3) + \mathrm{P}(\xi = 4) + \mathrm{P}(\xi = 5) + \mathrm{P}(\xi = 6) = \\ = \frac{C_6^3 \cdot 5^3 + C_6^4 \cdot 5^2 + C_6^5 \cdot 5 + 1}{6^6} = \frac{2500 + 375 + 30 + 1}{46656} = 6.23 \cdot 10^{-2},$$

$$\mathrm{P}(\xi \geqslant 4) = \mathrm{P}(\xi = 4) + \mathrm{P}(\xi = 5) + \mathrm{P}(\xi = 6) = \\ = \frac{C_6^4 \cdot 5^2 + C_6^5 \cdot 5 + 1}{6^6} = \frac{375 + 30 + 1}{46656} = 8.7 \cdot 10^{-3} < 0.01.$$

Таким образом, при выбранном выше уровне значимости $\varepsilon = 0.01$ пороговым значением ξ является $h_0 = 3$: если при двукратном бросании трех костей выпадет больше трех шестерок, то мы отклоняем нуль-гипотезу и соглашаемся с преподобным, в противном случае воздерживаемся от отклонения гипотезы (хотя это не означает, что мы нуль-гипотезу обосновали).

Итак, мы видим, что выбор случайной величины ξ (ее называют критической статистикой) происходит с учетом как нуль-гипотезы, так и альтернативной гипотезы. Формальных правил для выбора критической статистики, вообще говоря, нет, и во многих ситуациях существует несколько «достаточно разумных» претендентов на ее роль. К этому вопросу мы вернемся немного позднее.

Существенно отметить следующие обстоятельства. Стратегия принятия решения выбиралась таким образом, что эффект «множественности» маловероятных событий, о которых говорилось выше, исключен. В самом деле, мы отклоняем нуль-гипотезу при осуществлении одного-единственного события $\xi > 4$, причем данное событие фиксировано до проведения эксперимента. Второй важный момент: рассуждения, с помощью которых обосновывалась малость вероятности данного события, проведены в предположении, что нуль-гипотеза выполнена и в противном случае теряют силу. Учитывая эти факты, легко понять, что соображения, приведенные при рассмотрении шуточного примера, полностью корректны, а возражения против высказывания «маловероятные события не происходят» теперь явно оказываются неуместными.

Заметим, что нуль-гипотеза в данном примере относилась к числу так называемых *простых гипотез*: она полностью определяла (гипотетически) все исходные вероятности так, что в предположении нуль-гипотезы (по крайней мере, в принципе) можно вычислить вероятность любого события. В частности, мы вычислили вероятность события $\mathrm{P}(\xi > K)$. Простая гипотеза, таким образом, исключает всякую неопределенность. Но альтернативная гипотеза в нашем случае, сводящаяся к утверждению о том, что вероятность выпадения шестерки больше 1/6, очевидно, не относится к классу простых гипотез, так как содержит значительную неопределенность. Гипотеза не позволяет вычислять вероятность никаких событий, кроме, конечно, вероятности заведомо достоверных или невозможных. Существенно, однако, что в предположении альтернативной гипотезы расчет, доказывающий малость вероятности события выпадения четырех шестерок, теряет силу. Данная вероятность заведомо больше $C_6^4 \cdot 5^2/6^6$ (хотя неизвестно — насколько).

Примером простой альтернативной гипотезы могла бы быть, например, следующая: $\mathrm{P}(1) = \mathrm{P}(2) = \mathrm{P}(3) = \mathrm{P}(4) = \mathrm{P}(5) = 0.04$, $\mathrm{P}(6) = 0.8$.

На практике альтернативные гипотезы обычно бывают сложными, то есть не являющимися простыми, тогда как нуль-гипотезы бывают как сложными, так и простыми (чаще первое).

Рассмотренный пример показывает, что альтернативная гипотеза не всегда есть отрицание нуль-гипотезы (принцип «третьего не дано» не обязательно выполнен). В самом деле, отрицание нуль-гипотезы означает в данном случае несимметричность кости, то есть только тот факт, что среди чисел $\mathrm{P}(1), \ldots, \mathrm{P}(6)$ есть не равные 1/6. Это само по себе можно было бы выбрать в качестве альтернативной гипотезы. Однако мы с учетом поведения владельца костей выдвинули более конкретизированную альтернативную гипотезу: $\mathrm{P}(6) > 1/6$, а все остальные отклонения от симметричности игнорировали. Так, если в результате опыта выпали бы, скажем, шесть единиц, мы нуль-гипотезу не отклонили бы (в этом случае статистика не превосходила бы своего критического значения — четыре). Конечно, мы имеем веские основания считать кости несимметричными, однако в этом случае у нас нет оснований подозревать в нечестности владельца костей (учитывая предложенные им условия пари).

Альтернативная гипотеза тем самым ориентирована на содержание задачи. В «Неапольском деле» мы проверяли лишь порядочность владельца костей.

Что дает нам такое сужение отрицания нуль-гипотезы? Дело в том, что, фиксируя достаточно конкретную альтернативную гипотезу (по сравнению с отрицанием нуль-гипотезы), мы увеличиваем чувствительность критериев, так как проверка носит более направленный характер. В любой детективной истории следователь выдвигает версии; чем их меньше, тем легче их проверить.

1.3 Гипотеза о полиномиальном распределении и критерий χ^2

Далеко не всегда существуют достаточно «узкие» альтернативные гипотезы. Большой интерес поэтому представляет случай, когда в качестве альтернативной берется отрицание нуль-гипотезы.

Рассмотрим одну достаточно общую ситуацию, имеющую большое практическое значение. Предположим, что мы наблюдаем серию из n независимых испытаний. Каждое испытание может завершаться одним из m исходов $A_1, \ldots, A_m$. Вероятности $\mathrm{P}(A_1), \ldots, \mathrm{P}(A_m)$ не меняются от испытания к испытанию. Подлежащая проверке нуль-гипотеза H_0 состоит в том, что эти вероятности равны некоторым заданным заранее числам:

$$\mathrm{P}(A_1) = P_1 \quad \dots \quad \mathrm{P}(A_m) = P_m \tag{1}$$

(естественно, $\sum P_i = 1$). Относительно числа n будем предполагать, что оно велико (выборка достаточно обширна). Пусть ν_i — количество наступлений исхода A_i в данной серии из n испытаний. Априори $\nu_1, \dots, \nu_m$ — случайные величины, связанные соотношением $\nu_1 + \nu_2 + \dots + \nu_m = n$. Если нуль-гипотеза H_0 верна, то каждая из ν_i распределена по биномиальному закону с параметрами n, p_i. Отсюда, в частности, следует, что математические ожидания

$$\mathrm{M}\nu_i = np_i, \qquad i = 1, \dots, m.$$

Из каких соображений можно подобрать критическую статистику? Ясно, что при большом n частоты ν_i/n исходов A_i должны быть (при нуль-гипотезе) согласно закону больших чисел близки к p_i. Итак, естественно в качестве критической статистики выбрать некоторую величину типа расстояния между m-мерными векторами. Это можно сделать многими способами. Возможными вариантами могли бы быть, например, величины

$$\sum_{i=1}^{m} \left| p_i - \frac{\nu_i}{n} \right|$$

или

$$\sum_{i=1}^{m} \left(p_i - \frac{\nu_i}{n} \right)^2.$$

Однако на практике общеупотребительна величина

$$S = \sum_{i=1}^{m} \frac{(\nu_i - np_i)^2}{np_i}$$

или, что то же самое,

$$S = n \cdot \sum_{i=1}^{m} \frac{(\nu_i/n - p_i)^2}{p_i},$$

которая часто называется суммой Пирсона. Основное достоинство сумм Пирсона состоит в удобстве их применения: оказывается, что с ростом n распределение этой статистики стремится к некоторому предельному распределению (распределению χ^2 с $m-1$ степенью свободы), зависящему лишь от m, но не зависящему от чисел $p_1, \dots, p_m$. Сформулированное утверждение составляет предельную теорему Пирсона (для полиномиальных распределений). На доказательстве ее мы останавливаться не будем, тем более, что сама теорема является чисто вероятностной. (При $m = 2$ она вытекает из известной теоремы Муавра-Лапласа.) Ограничимся здесь лишь простой выкладкой, показывающей, что математическое ожидание суммы Пирсона (при нуль-гипотезе) всегда равно $m-1$. Действи-

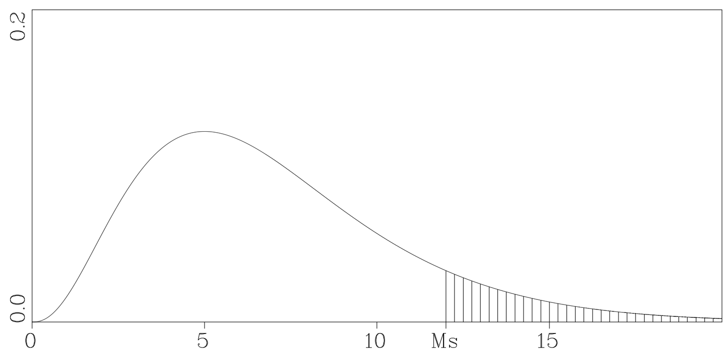

Рис. 1. Плотность распределения χ_s^2 при $s = 7$

тельно:

$$\mathrm{M}\left[\sum_{i=1}^{m}\frac{(\nu_i - np_i)^2}{np_i}\right] = \sum_{i=1}^{m}\mathrm{M}\left[\frac{(\nu_i - np_i)^2}{np_i}\right] = \sum_{i=1}^{m}\frac{\mathrm{D}\nu_i}{np_i} = \\ = \sum_{i=1}^{m}\frac{np_iq_i}{np_i} = \sum_{i=1}^{m}q_i = \sum_{i=1}^{m}(1 - p_i) = m - \sum_{i=1}^{m}p_i = m - 1.$$

(Здесь мы воспользовались тем фактом, что дисперсия биномиального распределения с параметрами n и p равна npq, где $q = 1 - p$.)

Хотя теорема и вероятностная, но основные свои приложения она находит в математической статистике.

В связи с этим стоит пояснить грань между теорией вероятностей и математической статистикой. В самой теореме не идет речь о проверке каких-либо гипотез. Равенства (1) предполагаются выполненными и затем доказывают сходимость распределений сумм Пирсона к упомянутому χ^2-распределению. Статистическая же ситуация возникает тогда, когда равенства (1) ставятся под сомнение и подлежат проверке.

Для пояснения схемы применения χ^2-критерия скажем несколько слов о распределении величины χ^2 с s степенями свободы, которую далее обозначим через χ_s^2. Это непрерывное распределение с математическим ожиданием s и дисперсией $2s$. Плотность $p(x)$ распределения равна нулю при $x \leqslant 0$ и положительна при $x > 0$ (явное выражение для плотности для нас сейчас не имеет значения). Из положительности плотности $p(x)$ на полуоси $x > 0$ следует, что случайная величина, распределенная по этому закону, не ограничена сверху, то есть не существует такого числа M, что всегда $\chi_s^2 < M$. Однако при любом $\varepsilon > 0$ можно указать практическую верхнюю границу $M_s(\varepsilon)$ такую, что

$$\mathrm{P}\left(\chi_s^2 \geqslant M_s(\varepsilon)\right) \leqslant \varepsilon,$$

другими словами, неравенство $\chi_s^2 \geqslant M_s(\varepsilon)$ практически невозможно. Принцип выбора числа M_s очевиден; именно в качестве $M_s(\varepsilon)$ нужно взять корень уравнения

$$F(x) = 1 - \varepsilon,$$

где $F(x)$ — функция распределения величины χ^2_s.

Если исходить не из функции распределения, а из плотности $p(x)$, то выбор числа M_s иллюстрируется рисунком, на котором площадь заштрихованной части равна ε. Величину $M_s(\varepsilon)$ часто называют квантилью порядка $1 - \varepsilon$. Естественно, работу по вычислению $M_s(\varepsilon)$ (квантили) нет нужды проводить каждый раз заново исходя из распределения, поскольку существуют соответствующие таблицы (для наиболее распространенных уровней значимости $\varepsilon = 0.01$; 0.05; 0.1). Отметим, что значения $M_s(\varepsilon)$ обычно не слишком отличаются от

$$\mathrm{M}\chi^2_s + 2\sqrt{\mathrm{D}\chi^2_s} = s + 2\sqrt{2s},$$

а таблица дает более точное значение с учетом и ε и s.

Из сказанного вытекает следующая схема проверки нуль-гипотезы. Задаемся уровнем значимости ε и вычисляем на серии из n наблюдений сумму Пирсона. Если она оказывается «слишком» велика, то есть больше $M_{m-1}(\varepsilon)$, то нуль-гипотеза отвергается. Итак, в этом случае критической статистикой является сама сумма Пирсона, а ее пороговое значение определяется из таблицы. Аргументы, по которым мы отклоняем гипотезу, того же рода, что и в «Неапольском деле». Если нуль-гипотеза верна, то большие значения суммы Пирсона (больше $M_{m-1}(\varepsilon)$) маловероятны. Таким образом, при больших суммах Пирсона предположение о нуль-гипотезе становится неправдоподобным. Отметим, что в силу предельного характера использованной теоремы Пирсона, число испытаний должно быть достаточно большим. Обычное требование состоит в том, что количество наступлений каждого из событий $A_1, \ldots, A_m$ в серии наблюдений должно составлять несколько десятков (или более).

Как интерпретировать случай, когда сумма Пирсона не превосходит порогового значения? Здесь мы, естественно, нуль-гипотезу не отклоняем. Наиболее точная интерпретация данной ситуации состоит в том, что у нас нет достаточных оснований для отклонения нуль-гипотезы. Встречаются и менее осторожные высказывания типа: «нуль-гипотеза согласуется с результатами эксперимента», «нуль-гипотеза подтверждается результатами эксперимента» и т.п.

Насколько допустимы такие высказывания? Конечно, такие термины, как «согласуется», являются довольно расплывчатыми и могут трактоваться по-разному. Здесь нужно иметь в виду несколько моментов. Во-первых, фиксируя определенный уровень значимости ε и жестко связывая с ним стратегию «отвергнуть — не отвергнуть» нуль-гипотезу, мы вносим в ситуацию элемент разрывности, не присущий самой задаче. Ясно, что случаи, когда критическая статистика приняла значение немного больше или меньше пороговых, по существу ничем не отличаются друг от друга. Однако выводы будут совершенно противоположными. Пусть, скажем, выбран уровень значимости $\varepsilon = 0.01$, а значение суммы Пирсона оказалось чуть меньше порогового значения $M_{m-1}(\varepsilon)$, но больше $M_{m-1}(0.02)$.

Нуль-гипотезу в этом случае мы не отклоним. Но можно ли ее считать хорошо согласованной с результатами эксперимента? Очевидно, что нет, ибо при нуль-гипотезе подобные суммы Пирсона могут встречаться реже, чем 1 раз на 50 случаев. Будь, например, у нас выбран уровень значимости $\varepsilon = 0.05$, мы бы при той же сумме Пирсона нуль-гипотезу отвергли (напомним, что выбор уровня значимости носит довольно произвольный характер).

Однако, если сумма Пирсона не превзошла своего математического ожидания $m - 1$, то здесь уместно сказать, что нуль-гипотеза согласуется с результатами эксперимента. Действительно, в этом случае мы далеко находимся от порогового значения (неважно, выбран ли уровень значимости $\varepsilon = 0.01$; 0.05 или 0.1). Согласие нуль-гипотезы с экспериментом в этом случае выглядит весьма убедительно.

Далее, пусть сумма Пирсона мала (скажем, меньше $m - 1$), можно ли утверждение усилить и говорить, что нуль-гипотеза «обоснована» или «доказана». Строго говоря, нет.

Само собой разумеется, что здесь о доказательстве на логическом уровне говорить вообще не приходится, но и на вероятностно-статистическом уровне доказать нуль-гипотезу в принципе невозможно. Действительно, если точное значение вероятностей исходов A_i мало отличается от чисел p_i (например в 20-м знаке), никакой реальный эксперимент выявить такое расхождение не сможет. Так, реальная игральная кость заведомо не является абсолютно симметричной. В этом случае нуль-гипотеза, что вероятность выпадения каждой грани в точности равна $1/6$, заведомо неверна, что ясно без всякой проверки. Фактически, выдвигая эту нуль-гипотезу, мы имеем в виду «почти симметричность» кости, то есть соотношения $\mathrm{P}(k) \approx 1/6$. Но такая формулировка выглядит расплывчатой и обычно не употребляется. Если понимать нуль-гипотезу в таком расширенном смысле, то малость суммы Пирсона при большом объеме выборки n может рассматриваться в некотором смысле как доказательство нуль-гипотезы. Как видим, здесь приходится употреблять столько оговорок и уточнений, что термина «доказательство» лучше избегать и говорить лишь о согласии с экспериментом. Заметим, что рассмотренный случай относится к наиболее благоприятным, так как здесь имеются некоторые основания говорить о доказательстве (в других ситуациях говорить о доказательстве совсем не приходится). Дело в том, что χ^2-критерий при проверке простой гипотезы является, как говорят в статистике, состоятельным против любых альтернатив. Это значит, что всякое отклонение от нуль-гипотезы наш тест «почувствует» и выявит, если объем n выборки достаточно велик. Подчеркнем своеобразную диалектику: с одной стороны, если отклонение от нуль-гипотезы есть, то анализ достаточно большой выборки его обнаружит; с другой стороны, при любом объеме выборки существуют столь малые отклонения от нуль-гипотезы, которые χ^2-критерий не почувствует. Здесь нет никакого противоречия, все зависит от того, что фиксируется — величина отклонения от нуль-гипотезы или объем выборки.

1.4 Пример из генетики

Приведем иллюстрацию применения χ^2-критерия для проверки простой гипотезы. Напомним первый закон Г. Менделя. Пусть есть два внешних проявления некоторого признака A и a, управляемых парой генов. За генами мы закрепим те же обозначения A и a. Здесь A — так называемый доминантный ген, в свою очередь a — рецессивный ген. Индивид будет обязательно обладать признаком A, если среди двух имеющихся у него генов присутствует ген A, признак же a, стало быть, наблюдается, лишь если пара генов выглядит как aa. Кстати сказать, пара генов называется генотипом. Согласно первому закону Менделя, механизм формирования генотипа с вероятностной точки зрения заключается в случайном выборе по одному гену у каждого из родителей (независимо и с равной вероятностью). Если, скажем, оба родителя имеют генотип Aa, то потомок может иметь генотипы AA, Aa, aa с вероятностями $1/4$, $1/2$, $1/4$; следовательно, внешнее проявление A признака будет иметь вероятность $3/4$, проявление a признака — вероятность $1/4$. (Биологи обычно говорят о расщеплении $3:1$). Отметим, что в ряде случаев наследование признака носит более сложный характер, так как признак может управляться и несколькими парами генов. Здесь первый закон Менделя в чистом виде уже не выполняется. Естественный способ проверки выполнения первого закона Менделя состоит в проведении экспериментов с последующей их статистической обработкой.

В частности, Мендель скрещивал между собой большое число растений гороха. Признак — цвет семян — имеет два проявления: A — желтые семена, a — зеленые семена. На подготовительном этапе Мендель скрещивал две чистопородные разновидности гороха с генотипами AA и aa, получив гибрид с генотипом Aa. Второй этап состоял в скрещивании гибридов между собой и получении второго поколения гибридов. Оказалось, что среди 8023 растений этого поколения 6022 обладали желтыми семенами, а 2001 — зелеными. Согласуются ли эти результаты с нуль-гипотезой, что цвет семян наследуется согласно первому закону Менделя? Здесь мы имеем $m = 2$ исхода, в соответствии с цветом семян A и a. Если нуль-гипотеза верна, то $\mathrm{P}(A) = 3/4$, а $\mathrm{P}(a) = 1/4$. Зададимся уровнем значимости $\varepsilon = 0.05$ (этот уровень наиболее популярен в биологии).

Распределение суммы Пирсона для столь большого числа, как $n = 8023$ практически не отличается от χ^2-распределения с $s = 1$ степенью свободы. По таблицам определяем $M_1(0.05) \approx 3.84$ — критическое значение (практическая верхняя граница). Вычислим сумму Пирсона для нашего случая:

$$S = 8023 \cdot \left[\frac{(6022/8023 - 3/4)^2}{3/4} + \frac{(2001/8023 - 1/4)^2}{1/4} \right] \approx 0.015.$$

Так как сумма Пирсона получилась значительно меньше $M_1(0.05)$ (и даже 1), то нуль-гипотеза хорошо согласуется с результатами эксперимента. Получаем веское основание считать, что первый закой Менделя справедлив.

Если же, например, оказалось бы, что число растений с желтыми семенами равно не 6022, а 6200, то подсчет показывает, что

$$S = 8023 \cdot \left[\frac{(6200/8023 - 3/4)^2}{3/4} + \frac{(1823/8023 - 1/4)^2}{1/4} \right] \approx 22.$$

В этом случае гипотезу о применимости в «чистом» виде закона Менделя пришлось бы отклонить. Отметим еще, что в рассмотренном случае $m = 2$ можно ссылаться не на предельную теорему Пирсона, а на интегральную предельную теорему Муавра-Лапласа.

Лекция 2

2.1 Как интерпретировать «слишком хорошее» согласие

Остановимся теперь на следующем любопытном вопросе. Мы использовали верхние оценки для суммы Пирсона и отклоняли нуль-гипотезу, когда сумма Пирсона оказывалась «слишком велика». Не следует ли отклонять нуль-гипотезу также и в случае «слишком малых» сумм Пирсона? Другими словами, не может ли слишком хорошее согласие послужить основанием для отклонения гипотезы? С чисто вероятностной точки зрения здесь не возникает никаких трудностей. Практическая нижняя граница для распределения χ^2_s (s — число степеней свободы) также может быть определена по таблицам. Если практической верхней границей служила квантиль порядка $1 - \varepsilon$, то практической нижней границей служит квантиль порядка ε, т.е. такое число Q, что

$$\mathrm{P}(\chi^2_s \leqslant Q) = \varepsilon.$$

Данный вопрос по существу относится не к теории вероятностей, а именно к статистике. Здесь все определяется, так сказать, контекстом, а именно наличием соответствующей альтернативной гипотезы.

Рассмотрим случай, когда есть основания предполагать фальсификацию результатов эксперимента с целью подгонки (таким образом, возникает ситуация, чем-то напоминающая «Неапольское дело»). Известен, в частности, реальный случай, когда при защите одной докторской диссертации по биологии (лысенковского направления) приводились экспериментальные данные, настолько «хорошо согласующиеся» с отстаиваемой гипотезой, что сразу вызвали недоверие. Подсчет сумм Пирсона показал, что столь малое значение этих сумм встречается (естественно, при выполнении нуль-гипотезы) реже, чем раз из 30 тысяч[3].

Если и раньше специалисты сомневались в правильности работы, то после такого подсчета все сомнения отпали. Диссертация была отклонена (диссертанта погубило, помимо прочего, полное незнание математической статистики).

Рассмотрим теперь противоположный случай, когда достоверность результатов не вызывает никакого сомнения. Скажем, мы сами добросовестно и качественно проводили эксперименты и не занимались никакой подгонкой результатов. Предположим, что в такой ситуации сумма Пирсона оказалась «слишком»

[3] Для сравнения можно отметить, что столь хорошее согласие, как в опыте Менделя, бывает примерно раз из десяти.

малой. Можно ли на этом основании отклонить нуль-гипотезу? Разумеется, нет, такое отклонение было бы попросту глупостью. В данном случае малость сумм Пирсона следует объяснить простым совпадением, так как других объяснений нет.

Сказанное еще раз иллюстрирует огромную роль тех или иных альтернативных гипотез. Естественно, что между рассмотренными двумя крайними ситуациями (сильное априорное недоверие и абсолютное доверие) есть и различные промежуточные варианты. Понятно, что здесь принятие решения может быть более трудным делом.

2.2 Малые выборки

Или как прожить без теоремы Пирсона при наличии компьютера

До сих пор мы предполагали, что выборки были достаточно обширными, и можно было пользоваться предельной теоремой Пирсона. Допустим теперь, что это не так. В случае малых выборок достаточно типична ситуация, когда одно или несколько чисел ν_i (напомним, что ν_i — количество наступлений исхода A_i) малы или даже равны нулю. В подобных ситуациях применять теорему Пирсона не рекомендуется, так как реальное распределение сумм Пирсона может быть весьма далеко от χ^2. Естественная рекомендация — увеличить объем выборки — по понятным причинам не всегда выполнима. Как быть в таких случаях? (Конечно, малая выборка дает меньше возможностей отклонить нуль-гипотезу, однако такое отклонение все же возможно, если величины ν_i/n резко отличаются от гипотетических вероятностей p_i). Поскольку пользоваться таблицами теперь нельзя, нужно изыскать другие способы определения практических верхних границ.

Опишем два способа действия, оба они будут предполагать использование компьютера. Первый способ состоит в точном нахождении распределения суммы Пирсона. Как уже говорилось, случайный вектор $\nu_1, \dots, \nu_m$ имеет полиномиальное распределение, то есть вероятности

$$p_{k_1,\dots,k_m} = p(\nu_1 = k_1, \nu_2 = k_2, \dots, \nu_m = k_m)$$

равны (при $\nu_i \geqslant 0, \sum_i \nu_i = n$) соответственно

$$p_{k_1,\dots,k_m} = \frac{n!}{k_1! \dots k_m!} p_1^{k_1} \dots p_m^{k_m}.$$

Практическая верхняя граница при заданном уровне значимости ε есть наибольшее из чисел M, для которых

$$\sum_{S(k_1,\dots,k_m) \geqslant M} p_{k_1,\dots,k_m} \geqslant \varepsilon,$$

где через $S(k_1, \dots, k_m)$ обозначена сумма Пирсона при $\nu_i = k_i$:

$$S(k_1, \dots, k_m) = \sum_{i=1}^{m} \frac{(k_i - np_i)^2}{np_i}.$$

Вычисления по этим формулам иногда могут осуществляться и вручную (особенно при $m = 2$, то есть для биномиального распределения). Значительно в большем диапазоне вычисления осуществимы с помощью компьютера. Однако если не мало m, например $m > 4$, то вычисления могут потребовать слишком больших затрат машинного времени.

В связи с этим изложим другой подход, который по временным затратам слабо зависит от m и величины выборки. В то время как все предыдущие способы действия использовали вероятностные факты, данный метод является «чисто статистическим». Здесь практические верхние границы определяются с помощью моделирования по методу Монте-Карло. Именно: исходя из нуль-гипотезы мы можем случайным образом генерировать любое число (например, 1000) выборок данного объема n. Для каждой выборки считается своя сумма Пирсона. Пусть задан уровень значимости $\varepsilon = 0.05$. Тогда мы отберем из полученных экспериментально выборок 5% с самыми большими суммами Пирсона. Наибольшая сумма Пирсона среди оставшихся выборок и дает нам приближенно практическую верхнюю границу. Впрочем, за практическую границу можно взять наименьшую сумму Пирсона среди выбранных.

Указанный прием используется не слишком часто, скорее всего из-за традиций математической статистики, которые закладывались во времена «бескомпьютерных» вычислений. Можно предполагать, что с появлением все более мощных компьютеров этот прием будет получать и более широкое распространение. Однако гораздо важнее методическое значение описанного приема, так как он ясно показывает истинный смысл наших действий, который зачастую бывает завуалирован различными аналитическими преобразованиями и ссылками на разного рода специальные (подчас довольно сложные) теоремы.

2.3 О сравнении статистических критериев

Ранее уже обращалось внимание на определенный произвол в выборе критической статистики. Возникает естественный вопрос, нельзя ли определить «наилучшую» критическую статистику для проверки данной нуль-гипотезы. Выше отмечалось, что сумма Пирсона является более удобной, чем другие разумные характеристики. Однако речь сейчас идет не о соображениях удобства, а о принципиальной стороне дела: как характеризовать качество критерия, основанного на той или иной критической статистике (трудоемкостью вычислений мы пренебрегаем). Сейчас мы увидим, что задать такой вопрос легче, чем на него ответить.

Некоторые соображения носят совсем простой характер. Очевидно, применяя какой-либо критерий, мы можем совершить в выводах ошибки двух видов: 1) отклонить нуль-гипотезу, когда она в действительности верна; 2) принять нуль-гипотезу, когда она в действительности неверна. Принято называть эти ошибки соответственно ошибками первого и второго рода.

Мы до сих пор имели дело только с ошибками первого рода, то есть строились такие области (промежутки) для критической статистики S, что попадание S в эту область влекло отклонение нуль-гипотезы (в статистике часто используется термин «критическая область»). При этом мы следили за тем, чтобы вероятность ошибки первого рода (то есть попадания S в критическую область при выполнении нуль-гипотезы) была равна заданному малому числу ε. Критические области могут быть определены, по существу, для любой статистики, при этом, вообще говоря, не единственным способом, что было видно на примере суммы Пирсона. Таким образом, статистический критерий определяется парой: статистика, критическая область. Итак, существует много критериев с одной и той же вероятностью ε ошибки первого рода (строго говоря, это так в непрерывном случае, поскольку в дискретном вероятности меняются скачками; данная сторона носит второстепенный характер). Естественно из двух таких критериев предпочесть тот, для которого меньше вероятность ошибки второго рода. Такой критерий называют более мощным. Мы рассматриваем, следовательно, сейчас вопрос о наиболее мощном критерии.

Почти в любом учебнике можно найти «решение» этого вопроса (так называемая теорема Неймана-Пирсона). Теорема представляет собой законченный строгий математический результат, но, к сожалению, имеющий весьма ограниченное значение для практики. Дело в том, что в условиях теоремы предполагается существование простой альтернативной гипотезы. Это требование исключает массу практически важных ситуаций. Возвращаясь к примеру с игральной костью, когда нуль-гипотеза состоит в ее симметричности, мы видим, насколько неестественно выглядит требование простой альтернативной гипотезы (например $\mathrm{P}(1) = \mathrm{P}(2) = 0.2$; $\mathrm{P}(3) = \mathrm{P}(4) = \mathrm{P}(5) = \mathrm{P}(6) = 0.15$). Естественные альтернативные гипотезы (кость несимметрична; вероятность выпадения шестерки больше $1/6$), очевидно, не являются простыми.

Что касается сложных альтернатив (либо сложных нуль-гипотез), то здесь известен ряд специальных случаев существования оптимальных критериев; подобные случаи — скорее исключение, чем правило.

Не следует, однако, надеяться, что в будущем появятся более общие результаты, ликвидирующие эту трудность. Она заключается в самом существе дела. Действительно, мы должны вычислять вероятность ошибок второго рода, то есть вероятность каких-то событий, при условии, что нуль-гипотеза неверна. Ясно, что такая негативная информация не позволяет вычислять никакие вероятности. Если, например, нуль-гипотеза состояла в симметричности костей, то утверждение «нуль-гипотеза неверна» эквивалентно утверждению, что не все $\mathrm{P}(C) = 1/6$ $(i = 1, \ldots, 6)$. Ни о каком вычислении вероятностей в данной ситуации не может быть и речи.

Итак, вообще говоря, понятие оптимального статистического критерия лишено смысла (если не говорить об упомянутых специальных случаях). Однако более или менее четкая информация об альтернативах позволяет формировать хоть и не оптимальные, но достаточно разумные статистические критерии. Под «достаточно разумными» мы понимаем критерии, которые, с одной стороны, были бы достаточно удобны в вычислительном отношении (при определении вероятностей ошибок первого рода) и, с другой стороны, были бы чувствительны к нарушениям нуль-гипотезы в направлении возможных альтернатив. В силу сказанного выше эту чувствительность обычно не удается точно характеризовать количественно и приходится довольствоваться соображениями здравого смысла (подобные неформальные моменты дают основание считать математическую статистику не только наукой, но и искусством).

2.4 Группировка исходов, непрерывные распределения, проверка датчиков случайных чисел

Остановимся теперь на таком моменте. Как уже говорилось выше, применение предельной теоремы Пирсона (удобной для приложений) допустимо в тех случаях, когда каждое из чисел ν_i (число наступлений исхода A_i) достаточно велико (5–10 или более). На практике, когда мы имеем дело со сравнительно небольшими выборками, это условие достаточно часто нарушается и некоторые из ν_i оказываются малыми и даже равными нулю. Это не значит, что в данном случае мы должны отказаться от применения теоремы Пирсона. Однако предпосылкой для такого применения должно быть предварительное «укрупнение» исходов, когда несколько различных исходов сливаются в один. Такого сорта укрупнение при любом объеме выборки заведомо необходимо в случае, когда число исходов бесконечно (например, счетно). Пусть, скажем, нуль-гипотеза состоит в том, что случайная величина γ (неотрицательная и целочисленная) распределена по закону Пуассона с заданным параметром a. Здесь возможно счетное число исходов $A_1, A_2, \ldots$, где A_k обозначает событие $\gamma = k-1$, вероятность которого при условии нуль-гипотезы есть

$$p_k = e^{-a}\frac{a^{k-1}}{(k-1)!}, \qquad (k = 1, 2, \ldots).$$

Ясно, что при любом объеме выборки только конечное число из величин $\nu_1, \nu_2, \ldots$ отлично от нуля. Наиболее естественное в рассматриваемом случае укрупнение состоит в том, что для некоторого m исходы $A_m, A_{m+1}, \ldots$ объединяются в один исход A'_m, который означает, таким образом, событие $\gamma \geqslant m$. При малых выборках зачастую возникает необходимость в дополнительном слияния других соседних событий A_k.

Рассмотрим случай нуль-гипотезы, состоящей в том, что γ — величина с наперед заданным непрерывным законом распределения. Другими словами, проверяется, что γ имеет заданную плотность $p_\gamma(x)$. В отличие от предыдущего примера в этом случае γ принимает несчетное множество значений. Для применимости теоремы Пирсона нужно провести еще более радикальное «укрупнение» множества исходов. Разобьем область значений на конечное число m непересекающихся областей $G_1, \ldots, G_m$. Связывая с каждой областью G исход $\gamma \in G$, мы приходим к случаю конечного числа исходов $A_1, \ldots, A_m$. При этом вероятность исхода A_k (в случае выполнения нуль-гипотезы) равна

$$P_k = \int\limits_{G_k} p_\gamma(x)dx.$$

Заметим, что γ может быть и многомерной случайной величиной. В этом случае x — вектор, то есть $x = (x_1, \ldots, x_n)$, $dx = dx_1 \ldots dx_n$ — элемент объема.

В качестве примера рассмотрим практически важную задачу о проверке качества датчика случайных чисел. Структуры таких датчиков могут быть различными, но в идеале они, как правило, должны давать последовательность независимых реализаций $\gamma_1, \gamma_2, \ldots$ случайной величины, равномерно распределенной на отрезке $[0,1]$. В силу того что реальный датчик может отклоняться от требований независимости и равномерности, обычно проводят экспериментальную проверку этих свойств. Пусть, например, нас интересует только равномерность. Разбивая отрезок $[0,1]$ на m непересекающихся интервалов равной длины, генерируя достаточно большую выборку $\gamma_1, \ldots, \gamma_n$, можно применить вышеуказанную схему. Ясно, что вероятность при нуль-гипотезе попасть в k-й интервал $P_k = 1/m$ для всех k. Если сумма Пирсона превзойдет критическое значение, мы либо отклоняем гипотезу о равномерности, то есть бракуем датчик, либо (более осторожный способ действия) повторяем испытание с большей выборкой.

Повторное превышение критического уровня служит веским основанием, чтобы забраковать датчик. Заметим, что χ^2-критерий в таких ситуациях содержит значительный элемент произвола, в частности, в выборе числа m. Практически можно рекомендовать, например, значение $m = 20$. Однако четких и однозначных рецептов в этом отношении не существует, так как выбор и малых, и больших m имеет свои достоинства и недостатки. Особенно очевидны эти недостатки при малых m, когда интервалы достаточно велики. Группируя вместе точки одного интервала, мы теряем более детальную информацию о расположении этих точек. Ясно, что χ^2-критерий следит лишь за количеством точек, попавших в данный интервал, и не чувствителен к «внутриинтервальным» нарушениям равномерности. С этой точки зрения целесообразно брать m как можно большим. Однако при этом следует учитывать такие обстоятельства. Во-первых, с ростом m растет и требуемый объем выборки и, следовательно, машинное время. Тоже существенным является другой момент. Если взять m очень большим и, соответственно, очень большую выборку, то критерий станет слишком чувствительным и будет браковать любой датчик, который дает распределение, даже слабо отличающееся от равномерного. Это обычно нежелательно, так как реальные (качественные) датчики дают не равномерное распределение, но очень близкое к нему, чего для практики обычно вполне достаточно. В силу сказанного вряд ли следует брать $m > 100$.

Вопрос о проверке независимости является более сложным, ибо свойство независимости относится ко всей совокупности $\gamma_1, \ldots, \gamma_n$. На практике обычно осуществляют лишь частичную проверку этого свойства, ограничиваясь, например, проверкой независимости пар γ_k, γ_{k+1}. Вопрос сводится к проверке равномерности распределения пар точек (γ_k, γ_{k+1}) в единичном квадрате $0 \leqslant x, y \leqslant 1$. Здесь применяется описанная выше схема: разбиваем квадрат на m_1 непересекающихся частей и так далее. В других случаях может проверяться независимость r-ок соседних значений $(\gamma_k, \gamma_{k+1}, \ldots, \gamma_{k+r})$. В этом случае проверяется равномерность распределения соответствующих точек в r-мерном кубе ($r > 2$). Конкретный выбор r зависит от специфики задачи.

Отметим, что критерий χ^2 — не единственный критерий при проверке гипотез о распределении непрерывной случайной величины. Сильным конкурентом его является критерий Колмогорова, основанный на совершенно иной критической статистике — максимуме модуля разности гипотетической функции распределения и так называемой выборочной функции распределения. Не останавливаясь подробно, отметим лишь следующее: большим достоинством критерия Колмогорова является отсутствие произвола, характерного для χ^2-критерия и связанного с числом интервалов разбиения и их длинами (для критерия Колмогорова группировка точек не требуется). Однако χ^2-критерий имеет более широкую область применения. В отличие от критерия Колмогорова он применим не только в скалярных, но и в многомерных случаях, и, что еще более важно, применим также для проверки сложных гипотез. Эти обстоятельства делают χ^2-критерий наиболее употребительным. Хотя в скалярном случае, когда нужно проверить простую гипотезу с непрерывным распределением, критерий Колмогорова [3, 4] представляется более предпочтительным.

Лекция 3

3.1 Сложные нуль-гипотезы

На практике сложные нуль-гипотезы встречаются значительно чаще, чем простые. Собственно, упомянутые выше вопросы, связанные с проверкой законов Менделя для наследуемых признаков и с проверкой качества равномерного датчика, — чуть ли не единственные практически важные примеры простых нуль-гипотез. Мир сложных нуль-гипотез значительно богаче и разнообразнее.

Рассмотрим некоторые примеры. Начнем с тех, для которых может быть применен критерий χ^2 (естественно, в модифицированной форме).

Выше было показано, как применяется χ^2-критерий, когда гипотетическое распределение определено однозначно. В приложениях, однако, гораздо чаще известен лишь тип гипотетического распределения, то есть оно определено с точностью до некоторых неизвестных параметров. Скажем, может проверяться гипотеза о том, что наблюдаемая величина распределена по закону Пуассона (один параметр), нормальному закону (два параметра) и так далее. Непосредственное применение старой схемы здесь, очевидно, невозможно, так как без знания параметров распределения мы не можем определить вероятности p_i, а следовательно, не можем построить суммы Пирсона. Возникает естественная мысль: вначале приближенно по выборке определить параметры $C_1, \ldots, C_k$, затем найти значения $p_i = p_i(C_1, \ldots, C_k)$ и, наконец, применить старую схему. Является ли законным такой способ действия? Ответ на этот вопрос положителен, но с определенными оговорками и уточнениями. Остановимся на основных из них.

Первое состоит, естественно, в том, что приближенное вычисление параметров по выборке (или, как говорят в статистике, оценка параметров) должно быть достаточно «разумным». Методами оценки параметров занимается большой раздел статистики. Поскольку этот материал можно найти в любом учебнике, мы не будем останавливаться на нем подробно. Упомянем лишь, что двумя основными методами получения оценок являются так называемые метод моментов и метод максимального правдоподобия. Первый из них основан на законе больших чисел и применяется для нахождения величин вида $\mathrm{M}f(\xi)$ (ξ — изучаемая случайная величина, $f(\cdot)$ — некоторая функция); достоинством его является то, что не требуется дополнительных знаний о типе распределения. Если же тип распределения заранее известен, и неизвестны лишь параметры, более употребительным является метод максимального правдоподобия. Суть его состоит в том, что по выборке $x = (x_1, \dots, x_n)$ строится так называемая функция правдоподобия $\mathcal{L}(x, C)$, где $C = (C_1, \dots, C_n)$, которая затем (при фиксированном x) максимизируется по C. Оптимальное значение C^* выбирается в качестве оценки неизвестного параметра C. Сама функция правдоподобия строится несколько по-разному в зависимости от того, дискретно или непрерывно распределение.

Если ξ распределена по дискретному закону, то

$$\mathcal{L}(x, C) = \prod_{i=1}^{n} \mathrm{P}(\xi = x_i)$$

(вероятности в правой части, естественно, зависят от C).

В случае непрерывного распределения величины ξ с плотностью $P_\xi(x, C)$

$$\mathcal{L}(x, C) = \prod_{i=1}^{n} P_\xi(x_i, C).$$

Иногда под функцией правдоподобия понимают $\ln \mathcal{L}(x, C)$, которую обычно удобнее максимизировать. На значение C^* это, очевидно, не влияет.

Приведем один пример. Пусть имеется нормальная выборка $x = (x_1, \dots, x_n)$; требуется оценить параметры распределения a (математическое ожидание) и σ (стандартное, или среднеквадратичное, уклонение). Здесь в роли вектора $C = (C_1, C_2)$ выступает вектор (a, σ). Как известно, плотность нормального распределения имеет вид

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-a)^2}{2\sigma^2}\right],$$

поэтому функция правдоподобия есть

$$\mathcal{L}(x, a, \sigma) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - a)^2\right].$$

Данную функцию нужно максимизировать в полуплоскости $\sigma > 0$. Нетрудно проверить, что эта гладкая положительная функция (при $\sigma > 0$) стремится к нулю как при $\sigma \to 0$, так и при $\sigma^2 + a^2 \to \infty$, и, кроме этого, в полуплоскости $\sigma > 0$ имеет единственную стационарную точку:

$$a^* = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \sigma^* = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - a^*)^2}.$$

Ясно, что данная точка является точкой глобального максимума.

Вернемся теперь к нашей основной задаче — проверке гипотез. Пусть, например, требуется проверить гипотезу о нормальности распределения. Это сложная гипотеза, поскольку параметры неизвестны. Если бы мы знали их, то можно было бы действовать, как и выше, и применить критерий χ^2. Напрашивается такой способ действия: 1. Оцениваем параметры по выборке, например, из принципа максимального правдоподобия. 2. Применяем χ^2 - критерий, обращаясь с оценками параметров, как если бы они были точными значениями. Является ли законной такая процедура (напомним, что в теореме Пирсона оценки параметров вообще не фигурируют)? Оказывается, эта процедура законна, но с некоторыми уточнениями, а именно: справедлива теорема (Фишер), обобщающая теорему Пирсона на интересующий нас случай. Суть ее, приблизительно говоря, сводится к тому, что при наличии r оцениваемых параметров предельное распределение для сумм Пирсона есть по-прежнему χ^2, но с $m - r - 1$ степенями свободы. Здесь число m имеет тот же смысл, что и в теореме Пирсона (число взаимоисключающих исходов A_i). В нашем примере с нормальным распределением m — количество интервалов, на которое разбивается числовая прямая. Итак, теорема Фишера показывает, что характер распределения не меняется, но число степеней свободы уменьшается на r. Именно благодаря этому факту критерий χ^2 стал одним из наиболее употребительных инструментов статистики.

Отмечая общность теоремы Фишера, следует все же сказать, что она применима не всегда и требует выполнения некоторых (на практике чаще всего выполненных) условий. Строгую математическую формулировку теоремы Фишера можно найти в [2]; нас сейчас, главным образом, интересует общий характер этого результата. Возвращаясь к примеру с нормальным распределением, предположим, что числовая прямая разбита на 10 интервалов, причем заданный уровень значимости есть 0.05. В силу сказанного выше сумма Пирсона (при достаточно большом n) имеет распределение, близкое к χ^2_7 (здесь $10 - 2 - 1 = 7$). В таблице для χ^2-распределения находим критическое значение $M_7(0.05) = 14.07$.

Итак, гипотеза о нормальности распределения отклоняется в том случае, если значения суммы Пирсона окажутся больше 14.07.

3.2 Пример: таблицы сопряженности признаков

Рассмотрим теперь пример применения критерия χ^2, который внешне будет существенно отличаться от предыдущего. Рассмотрим вопрос об обработке так называемых таблиц сопряженности.

Пусть рассматривается группа из n объектов, каждый из которых классифицируется по двум различным признакам. Далее, каждый признак имеет несколько уровней. Скажем, если речь идет о возрасте человека (признак человека), то он может характеризоваться уровнями: 0–10 лет, 10–20 лет, 20–30 лет, 30 лет и старше (это соответствует молодежной градации, согласно которой все люди старше 30 лет являются стариками). Для другого признака — температуры тела человека — можно выбрать, например, уровни: ниже 37 градусов, от 37 до 38, от 38 до 39, выше 39 градусов. При выборе уровней количественных признаков имеется определенный произвол. Однако такой качественный признак, как пол, имеет всего два «уровня».

Основной вопрос, который нас будет интересовать: наблюдается ли зависимость между двумя выбранными признаками, которые мы обозначим, скажем, через B и C.

Пусть B имеет m_1 уровней $B_1, \ldots, B_{m_1}$, C — m_2 уровней $C_1, \ldots, C_{m_2}$. Наблюдая группу из n объектов, мы относим каждый из них к одной из $m_1 \cdot m_2$ категорий, отвечающих различным комбинациям уровней B_iC_j. Обозначим через n_{ij} количество объектов, соответствующих категории B_iC_j. В результате получим целочисленную неотрицательную матрицу размером $m_1 \times m_2$.

Рассматриваемая задача разительно отличается от упомянутой выше проверки гипотезы о нормальности распределения; кажется, между ними нет ничего общего. Тем не менее обе эти задачи относятся к задачам о проверке сложных гипотез, и для решения применяется один и тот же критерий χ^2. Проверяемая нуль-гипотеза о независимости признаков B и C является сложной. Действительно, если предположить, что она верна, то это еще не позволяет нам полностью характеризовать вероятностную картину распределения по категориям B_iC_j. Данная картина была бы известна, если бы были известны вероятности (доли) каждого признака в отдельности $\tilde{p}_i = \mathrm{P}(B_i)$, $\tilde{\tilde{p}}_j = \mathrm{P}(C_j)$. В этом случае в силу независимости мы получили бы

$$p_{ij} = \mathrm{P}(B_iC_j) = \tilde{p}_i\tilde{\tilde{p}}_j.$$

Таким образом, в нашей задаче наблюдалась бы уже встречавшаяся ситуация с $m = m_1 \cdot m_2$ несовместными исходами

$$A_1 = B_1C_1, \quad A_2 = B_2C_1, \quad \ldots \quad A_m = B_{m_1}C_{m_2},$$

вероятности которых суть $\tilde{p}_1\tilde{\tilde{p}}_1, \tilde{p}_2\tilde{\tilde{p}}_1, \ldots, \tilde{p}_{m_1}\tilde{\tilde{p}}_{m_2}$.

Итак, если бы все $\tilde{p}_i$, $\tilde{\tilde{p}}_j$ были известны, то можно было бы воспользоваться теоремой Пирсона, согласно которой сумма Пирсона имеет при достаточно большой выборке распределение, близкое к $\chi^2_{m_1m_2-1}$.

Фактически, однако, параметры

$$\tilde{p}_1, \ldots, \tilde{p}_{m_1}; \tilde{\tilde{p}}_1, \ldots, \tilde{\tilde{p}}_{m_2}$$

неизвестны и мы их должны оценить по выборке. Число «существенно» неизвестных параметров есть $m_1 + m_2 - 2$ (так как $\sum \tilde{p}_i = \sum \tilde{\tilde{p}}_j = 1$, параметры $\tilde{p}_{m_1}$, $\tilde{\tilde{p}}_{m_2}$ можно не включать в число неизвестных).

Итак, число степеней свободы в предельном распределении для сумм Пирсона (согласно теореме Пирсона) будет равно

$$m_1m_2 - 1 - (m_1 + m_2 - 2) = m_1m_2 - m_1 - m_2 + 1 = (m_1 - 1)(m_2 - 1).$$

Оценим теперь неизвестные параметры по методу наибольшего правдоподобия. Введем обозначения

$$n_{i\cdot} = \sum_{j=1}^{m_2} n_{ij}, \qquad n_{\cdot j} = \sum_{i=1}^{m_1} n_{ij}$$

(эти числа описывают количество объектов выборки, отвечающих тому или иному уровню какого-либо из признаков B, C). Функция правдоподобия

$$\mathcal{L} = \prod_{i=1}^{m_1} \prod_{j=1}^{m_2} (\tilde{p}_i \tilde{\tilde{p}}_j)^{n_{ij}}.$$

Сделаем одно пояснение. Величина $\mathcal{L}$ есть вероятность того, что таблица будет заполняться в определенном порядке, то есть первые n_{11} объектов попадут в первую клетку B_1C_1, n_{21} объектов — в следующую клетку B_2C_1 и т.д. Можно было бы рассмотреть случай, когда таблица заполняется в произвольном порядке. Тогда у величины $\mathcal{L}$ появился бы дополнительный положительный множитель: число перестановок с повторениями. Постоянный множитель на точку максимума, очевидно, не влияет.

Максимизируя $\mathcal{L}$ (элементарные выкладки мы опускаем) на множестве

$$\sum \tilde{p}_i = 1, \qquad \sum \tilde{\tilde{p}}_j = 1, \qquad \tilde{p}_i \geqslant 0, \qquad \tilde{\tilde{p}}_j \geqslant 0,$$

получим

$$\tilde{p}_i^* = \frac{n_{i\cdot}}{n} \qquad (i = 1, \ldots, m_1),$$
$$\tilde{\tilde{p}}_j^* = \frac{n_{\cdot j}}{n} \qquad (j = 1, \ldots, m_2),$$

где n, как и раньше, — число объектов в данной совокупности. Другими словами, получили напрашивающуюся оценку: в качестве приближенного значения того или иного уровня следует брать соответствующую частоту в рассматриваемой совокупности.

Сумма Пирсона в нашем случае принимает вид

$$\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{\left(n_{ij} - n \cdot \dfrac{n_{i\cdot}}{n} \cdot \dfrac{n_{\cdot j}}{n}\right)^2}{n \cdot \dfrac{n_{i\cdot}}{n} \cdot \dfrac{n_{\cdot j}}{n}} =$$
$$= n \cdot \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \left(\frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} - 2 \cdot \frac{n_{ij}}{n} + \frac{n_{i\cdot} n_{\cdot j}}{n^2} \right) = n \cdot \left[\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} - 1 \right].$$

Здесь при выводе использовано, что

$$\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} n_{ij} = n, \qquad \sum_{i=1}^{m_1} n_{i\cdot} = n, \qquad \sum_{j=1}^{m_2} n_{\cdot j} = n.$$

Приведен конкретный пример таблицы сопряженности (см. табл. 1), заимствованный из книги Крамера «Математические методы статистики». В таблице отражено распределение супружеских пар по годовому доходу и количеству детей (выборка взята из шведской переписи населения 1936 года).

Таблица 1. Распределение супружеских пар по годовому доходу и количеству детей (за единицу дохода принята сумма в 1000 крон).

Количество детей	Доходы				**Сумма**
	0–1	1–2	2–3	3	
0	2161	3577	2184	1636	**9558**
1	2755	5081	2222	1052	**11110**
2	936	1753	640	306	**3635**
3	225	419	96	38	**778**
4 и более	39	98	31	14	**182**
Сумма	**6116**	**10928**	**5173**	**3046**	**25263**

Подсчет суммы Пирсона дает значение 568.5. Критическим значением для χ^2-распределения с $m = (5-1)(4-1) = 12$ степенями свободы и уровнем значимости 0.01 есть 26.2. Таким образом, значение суммы Пирсона намного превосходит критическое, и нуль-гипотезу о независимости факторов можно отбросить с полной уверенностью.

Таким образом, приведенный численный анализ подтверждает интуитивно ожидаемый факт связи дохода семьи с количеством детей.

3.3 Интерпретация результатов обработки. Сила и надежность связи

Сделаем ряд замечаний по поводу таблиц сопряженности вообще и рассматриваемого примера в частности.

1. Мы установили сам факт связи между рассмотренными факторами. Однако часто представляет интерес вопрос о характере этой связи, скажем, положительная или отрицательная корреляция между числом детей и доходом семьи (то есть верно ли, что в больших семьях в среднем больший или, наоборот, меньший доход).

Критерий χ^2 не дает ответа на подобные вопросы. Другое дело, что в случае заметной зависимости, обнаруженной с помощью критерия χ^2, этот вопрос становится естественным и может быть исследован дополнительно. Например, можно исследовать разности

$$n_{ij} - n \cdot \frac{n_{i\cdot}}{n} \cdot \frac{n_{j\cdot}}{n}$$

(при вычислении суммы Пирсона эти разности возводятся в квадрат и теряется информация об их знаках). Напомним, что таблицы сопряженности применяются также и для чисто качественных факторов (социальное положение, цвет глаз и так далее), когда понятия «больше» или «меньше» лишены смысла. Что же касается исследования корреляции между количественными факторами, то для этой цели обычно используется другая статистика — так называемый выборочный коэффициент корреляции. При его вычислении группировка данных по уровням (необходимая для критерия χ^2) нецелесообразна.

2. Отвлекаясь от направленности связи, можно поставить вопрос о «силе» связи. Скажем, является ли связь, выявленная в последнее примере, сильной или слабой. На первый взгляд связь выглядит как «сильная», поскольку критическое значение статистики было превышено во много раз. Однако это не так. Важно уяснить различие между надежностью проявления связи и ее силой. Упомянутое превышение означает лишь, что мы практически достоверно можем отвергнуть нуль-гипотезу, но отнюдь не означает, что эта гипотеза нарушается в «сильной» степени. Здесь существенную роль играет объем выборки: даже слабое отклонение от нуль-гипотезы приводит при достаточно большом объеме выборки к резкому превышению суммой Пирсона критического уровня. Это очевидно из следующих соображений: если нуль-гипотеза неверна, то есть для некоторых i, j $p_{ij} \neq \tilde{p}_i\tilde{\tilde{p}}_j$, то поскольку с ростом n величина n_{ij} ведет себя как np_{ij}, соответствующее слагаемое

$$\left(n_{ij} - n \cdot \frac{n_{i\cdot}}{n} \cdot \frac{n_{j\cdot}}{n}\right)^2 \cdot \left(n \cdot \frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n}\right)^{-1}$$

в сумме Пирсона растет линейно по n. Стало быть, данное слагаемое, а вместе с ним и вся сумма Пирсона неограниченно растет при увеличении объема выборки. Это резко отличается от поведения суммы Пирсона в случае справедливости нуль-гипотезы.

Если все же нас интересует, насколько интенсивна связь между факторами, то мы должны исключить влияние объема выборки. Существует несколько количественных мер силы связей на основе сумм Пирсона (коэффициенты Пирсона, Чупрова, Крамера и др.). Употребителен, в частности, коэффициент Крамера (точнее, Чупрова-Крамера), равный

$$C^2 = \frac{S}{n[\min(m_1, m_2) - 1]}.$$

Данный коэффициент при любых m_1, m_2 меняется в пределах от 0 до 1, при этом большее значение соответствует более «сильной» связи. В свете сказанного ясно, что при больших значениях C^2, но малых объемах выборки мы получаем данные о сильной связи, которые, однако, не обязательно являются достаточно надежными.

В примере со шведскими семьями получим

$$C^2 = \frac{568.5}{25263 \cdot [\min(5;4) - 1]} = \frac{568.5}{25263 \cdot 3} = 0.0075,$$

то есть в данном случае мы имеем дело с очень слабой связью между признаками (C^2 близко к нулю).

Итак, мы рассмотрели пример, когда связь надежно установлена, хотя и является весьма слабой.

С точки зрения здравого смысла факт наличия связи между доходами и численностью семьи можно было бы заранее предвидеть (пожалуй, несколько неожиданной может показаться слабость этой связи). Возможны, как легко видеть, четыре комбинации: 1. Сильная и надежная связь; 2. Слабая и надежная связь; 3. Сильная и ненадежная связь; 4. Слабая и ненадежная связь (здесь мы имеем в виду чисто качественную сторону дела, не уточняя количественных характеристик силы и надежности). Естественно, каждый из перечисленных случаев интерпретируется по-разному.

Если говорить только о справедливости нуль-гипотезы, то в вариантах 1 и 2 гипотеза отклоняется, а в вариантах 3 и 4 — нет. Существуют, однако, и другие аспекты интерпретации. Различие между случаями 1 и 2 состоит в том, что для практики основной интерес представляет сильная зависимость; углубленное изучение таких зависимостей может быть плодотворным во многих отношениях. Слабая же зависимость (пусть даже практически достоверно установленная) представляет гораздо меньший интерес. В рассмотренном примере зависимость между доходом и числом детей в семье настолько мала, что вряд ли заслуживает серьезного изучения. Что же касается вариантов 3 и 4, то здесь можно заметить следующее: если в случае 4 мы принимаем нуль-гипотезу «со спокойной душой», то в случае 3 имеем недостаточно надежное указание о наличии сильной связи (такие связи, как уже говорилось, особенно важны). Поэтому естественной реакцией в случае 3 является попытка провести повторные исследования с выборками больших объемов (естественно, такие исследования не всегда возможны).

Остановимся еще на одном существенном моменте. Допустим, что между исследуемыми факторами установлено наличие связи. Пусть для определенности эта связь будет типа сильной положительной корреляции (напомним, что знак корреляции устанавливается специальными методами). Возникает естественный вопрос, что является причиной, а что следствием (то ли многодетность вынуждает людей искать пусть трудную, но хорошо оплачиваемую работу, то ли, наоборот, хорошо обеспеченные родители заводят много детей). Важно подчеркнуть, что математическая статистика, вообще говоря, не дает ответа на подобные вопросы.

Вычисляя суммы Пирсона или выборочные коэффициенты корреляции, мы можем установить сам факт связи, ее силу, положительность или отрицательность корреляции, но не причинно-следственную структуру. Вопрос о причинах и следствиях вообще иногда не имеет смысла (скажем, блондины чаще имеют голубые глаза, однако, рассуждения на тему, являются ли светлые волосы «следствием» голубых глаз или наоборот, лишены всякого содержания). Имеет ли смысл вопрос о причинно-следственной связи, и каким может быть ответ, зависит от природы изучаемых факторов. В статистических выборках эта природа, очевидно, никак не отражается. Иногда вопрос о причинах и следствиях решается на основе здравого смысла с учетом характера факторов. Например, надежные статистические исследования показывают, что часто курение сочетается с болезнями легких. Отсюда однозначно делается правильный вывод, что курение ведет к ухудшению состояния здоровья, но не наоборот (однако такой информации в соответствующих статистических таблицах нет). Правда, в отдельных случаях, привлекая данные дополнительных экспериментов, статистическими методами можно пролить свет на причинно-следственную структуру связи, но более типично здесь применение аргументации, весьма далекой от математической статистики.

3.4 Критерий Вилкоксона для сравнения двух выборок

Выше было уделено большое внимание критерию согласия χ^2. Не должно, однако, создаваться впечатление, что этот критерий является единственным при проверке статистических гипотез. Мы остановимся сейчас еще на одном важном критерии, известном в литературе как критерий Вилкоксона или Манна-Уитни. Он применяется в тех случаях, когда нужно сравнить две выборки из непрерывных случайных величин и проверить гипотезу об однородности.

Подробнее, пусть имеются две выборки $X_1, \ldots, X_{n_1}$, $Y_1, \ldots, Y_{n_2}$. Эти выборки отражают обычно ситуацию такого типа: X_i — экспериментальные значения (независимые) некоторой случайной величины, а Y_i — значения той же величины, наблюдаемые, однако, при определенном изменении условий эксперимента — обычно при введении некоторого дополнительного фактора. Нас интересует вопрос: отражается ли этот фактор существенным образом на наблюдаемой случайной величине — в первую очередь, способствует ли он увеличению (уменьшению) этой величины. Как обычно, нуль-гипотеза будет отражать позицию «скептика»: никакого влияния этот фактор не оказывает. Другими словами, нуль-гипотеза состоит в следующем: обе выборки отвечают одному и тому же распределению (или, как часто говорят статистики, выбраны из одной и той же генеральной совокупности). Функцию распределения этой совокупности обозначим через $F_\xi(x)$. Единственное предположение относительно $F_\xi(x)$ состоит в том, что она непрерывна, поэтому без ограничения общности можно считать, что все $n_1 + n_2$ чисел X_i и Y_i различны (так как вероятность совпадения равна нулю). Поскольку гипотетическая функция распределения $F_\xi(x)$ нам неизвестна, данная нуль-гипотеза является сложной. В принципе для решения поставленной задачи можно применить χ^2-критерий, разбив численную прямую на m более или менее удачно выбранных интервалов, проводя группировку и рассматривая соответствующие таблицы сопряженности размером $2 \times m$. Критерий Вилкоксона, о котором будет идти речь, не требует группировки величин, а следовательно, не содержит произвола, характерного для критерия χ^2. Однако, главным преимуществом его является то, что он применим и при небольших объемах выборок. Наконец, как мы увидим, попутно будет получена информация, является ли данный фактор «увеличивающим» или «уменьшающим» (приблизительно говоря).

Будем действовать следующим образом. Объединим все величины X_i и Y_i, упорядочим их в порядке возрастания. Полученную выборку обозначим через

$$Z_1, \ldots, Z_{n_1+n_2} \qquad (Z_1 < Z_2 < \ldots < Z_{n_1+n_2}).$$

Каждая из величин Z_i совпадает либо с некоторым элементом множества $X = \{X_1, \ldots, X_{n_1}\}$, либо с некоторым элементом $Y = \{Y_1, \ldots, Y_{n_2}\}$. Решающей для критерия Вилкоксона является статистика

$$S = \sum_{i:\, Z_i \in X} i,$$

то есть сумма номеров (так называемых рангов) элементов в объединенной выборке, которые ранее принадлежали выборке $X_1, \ldots, X_{n_1}$.

Приблизительно говоря, слишком большие (малые) значения S указывают на то, что элементы первой выборки «велики» («малы») по сравнению с элементами второй выборки. Для того чтобы эти соображения приобрели точный смысл, надо, как и ранее, определить критические значения статистики S, отвечающие тому или иному уровню значимости. Важной особенностью ранговых критериев и критерия Вилкоксона, в частности, представляется следующее обстоятельство. Хотя нуль-гипотеза, строго говоря, здесь является сложной (предполагается лишь, что обе выборки X и Y отвечают одной и той же функции распределения, но какой — неизвестно), распределение решающей (критической) статистики S в случае нуль-гипотезы, в принципе, известно и зависит только от n_1 и n_2. Это распределение можно интерпретировать следующим образом. На $n_1 + n_2$ карточках (одинаковых) написаны числа $1, 2, \ldots$. Карточки тасуются, и наудачу выбирается n_1 из них (без возвращения). Сумма чисел на выбранных карточках распределена так же, как S (при нуль-гипотезе). Роль нуль-гипотезы заключается в том, что все X_i $(i = 1, \ldots, n_1)$, Y_i $(i = 1, \ldots, n_2)$ находятся в равном положении и с равной вероятностью $1/(n_1 + n_2)$ каждый из них может занять любое место в объединенной выборке.

При малых n_1, n_2 непосредственное нахождение критических значений осуществляется без труда. Пусть, например, $n_1 = 4$, $n_2 = 6$ и выбран уровень значимости $\varepsilon = 0.05$. Всего возможно $C_{10}^4 = 210$ взаимных расположений элементов выборки X в общей совокупности. В силу нуль-гипотезы все случаи равновероятны. Статистика S принимает возможные значения от минимального 10 (соответствующего $XXXXYYYYYY$) до максимального 34 (соответствующего $YYYYYYXXXX$). Критическая область значений S (где нуль-гипотеза отвергается) определена при фиксированном уровне значимости не вполне однозначно. Примем сейчас, что критическая область симметрична, то есть отклоняется одинаковое количество «слишком малых» и «слишком больших» значений S (в некоторых ситуациях такой подход нецелесообразен, о чем будет сказано ниже). Простой подсчет показывает, что ряд распределения для S имеет вид:

S	10	11	12	13	$\ldots$	30	31	32	33	34
P	$\frac{1}{210}$	$\frac{1}{210}$	$\frac{2}{210}$	$\frac{3}{210}$	$\ldots$	$\frac{5}{210}$	$\frac{3}{210}$	$\frac{2}{210}$	$\frac{1}{210}$	$\frac{1}{210}$

(вероятности других значений нам не понадобятся). Действительно, значение $S = 11$, например, принимается в единственном случае $(XXXYXYYYYY)$, а значение $S = 12$ в двух случаях $(XXXYYXYYYY)$ и $(XXYXXYYYYY)$. Таким образом, вероятность того, что

$$\mathrm{P}(S \leqslant 12 \vee S \geqslant 32) = \frac{8}{210} < 0.05,$$

но
$$\mathrm{P}(S \leqslant 13 \vee S \geqslant 31) = \frac{14}{210} > 0.05.$$
Итак, нуль-гипотеза отклоняется в случаях $S \leqslant 12$ или $S \geqslant 32$. При этом отклонение нуль-гипотезы сопровождается заключением, что значения выборки X «преобладают» над значениями выборки Y при $S \geqslant 32$ (и наоборот при $S \leqslant 12$).

Строго говоря, здесь преобладание относится лишь к средним рангам, а не к средним значениям самих выборок. Дело в том, что ранги учитывают только взаимное расположение элементов, но не расстояние между ними. Поэтому не исключено, что среднее значение может быть больше у одной выборки, а средний ранг — у другой. На практике, однако, такие ситуации встречаются довольно редко.

Мы проиллюстрировали, как находить простым перебором критические значения S при малых объемах выборок. С ростом n_1, n_2 такой подсчет становится затруднителен. В связи с этим заметим, что для не слишком больших значений n_1, n_2 (порядка нескольких десятков) и наиболее употребительных уровней значимости критические значения можно найти в таблицах. Если же n_1 и n_2 велики, то обычно применяется нормальная аппроксимация распределения S (известны соответствующие теоремы при $n_1, n_2 \to \infty$). Что касается математического ожидания и дисперсии S, то они могут быть найдены в самом общем виде. Пусть r_i — ранг величины X_i в объединенной выборке Z. При нуль-гипотезе величина r_i равновероятно принимает значения от 1 до $n_1 + n_2$. В силу этого
$$\mathrm{M}r_i = \frac{n_1 + n_2 + 1}{2}, \qquad (i = 1, \ldots, n_1).$$
Отсюда
$$\mathrm{M}S = \mathrm{M}\left[\sum_{i=1}^{n_1} r_i\right] = \sum_{i=1}^{n_1} \mathrm{M}r_i = n_1 \cdot \frac{n_1 + n_2 + 1}{2}.$$
Дисперия S вычисляется несколько сложнее, так как величины r_i зависимы; опуская выкладки, приведем окончательный результат:
$$\mathrm{D}S = \frac{1}{12}\, n_1 n_2 (n_1 + n_2 + 1).$$
Итак, при больших n_1, n_2 распределение величины близко к нормальному распределению
$$N\left(n_1 \cdot \frac{n_1 + n_2 + 1}{2}, \sqrt{\frac{1}{12}\, n_1 n_2 (n_1 + n_2 + 1)}\right).$$
В частности, при $\varepsilon = 0.05$ в силу известного правила 2σ для нормального закона мы будем отклонять нуль-гипотезу в случае, если
$$\left|S - n_1 \cdot \frac{n_1 + n_2 + 1}{2}\right| \geqslant \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{3}}.$$

Выше речь шла о так называемом двухстороннем критерии Вилкоксона в «симметричном» варианте. Не всегда, однако, такой подход уместен. Дело здесь снова заключается в характере альтернативной гипотезы. В некоторых случаях по самому смыслу задачи заранее известно, что речь может идти лишь об одной из альтернатив: значения выборки X «преобладают» над значениями выборки Y, значения Y «преобладают» над X. Для иллюстрации приведем пример, связанный с исследованиями на биологическом факультете Ярославского государственного университета. Изучался вопрос о влиянии промышленных отходов на жизнедеятельность микроорганизмов озера Байкал. Для проверки воздействия проводились оценки некоторых параметров, характер которых здесь не существен. Нуль-гипотеза, естественно, состояла в том, что существенное влияние отсутствует, альтернативная гипотеза — что влияние вредно (гипотетическую возможность, что это влияние полезно, всерьез принять трудно по очевидным причинам). Ясно, что здесь должен был применяться односторонний критерий.

Если вернуться к рассмотренному выше примеру ($n_1 = 4$, $n_2 = 6$, $\varepsilon = 0.05$) и принять, например, что по смыслу задачи реальной является лишь альтернатива: X «преобладает» над Y, то нуль-гипотезу следует отклонить при $S \geqslant 31$. Подчеркнем еще раз, что вопрос о выборе одно- или двухстороннего критерия, равно как и уровня значимости ε, лежит вне пределов математической статистики как таковой и определяется конкретным содержанием задачи.

Сделаем еще несколько замечаний по критерию Вилкоксона.

1. Этот критерий направлен, в первую очередь, на выявление альтернатив типа: X «преобладает» над Y, Y «преобладает» над X. Он может оказаться нечувствительным к неоднородностям и различиям (в выборках) другого характера, скажем, в дисперсиях. Если, например, одна из выборок распределена по закону $N(a, \sigma)$, а другая по закону $N(a, \sigma_1)$, то какое бы ни было различие между σ и σ_1, критерий Вилкоксона такой неоднородности «не почувствует». Как выражаются статистики, критерий Вилкоксона не является состоятельным против такой альтернативы.

Известны критерии, состоятельные против любой альтернативы, то есть позволяющие при достаточно больших объемах выборок обнаружить любое различие в распределениях. Наиболее известным критерием такого рода является критерий Смирнова. Он тоже носит ранговый характер (хотя и формулируется в других терминах). Для более подробного знакомства с относящимися сюда результатами Колмогорова и Смирнова мы отсылаем читателей к [3, 4].

2. До сих пор предполагалось, что в сравниваемых выборках нет совпадающих значений. Это естественное предположение для непрерывных распределений. Однако и в непрерывном случае равенство может возникнуть за счет округления. Вопрос о том, какие ранги приписывать X_i и Y_j, если $X_i = Y_j$, может решаться по-разному. Одним из вариантов является применение рандомизации, то есть случайное разыгрывание рангов между совпадающими значениями. Преимуществом такого подхода является то, что распределение статистики S не меняется. Другой вариант — использование «усредненных» рангов — связан с изменени-

ем распределения. Особую роль вопрос об обработке совпадений приобретает, конечно, при сравнении выборок, имеющих дискретное распределение.

3. Ранговые критерии могут использоваться и в тех случаях, когда объекты выборок не являются случайными величинами в обычном смысле слова, но допускают упорядочение по какому-либо признаку, возможно, с использованием экспертных оценок. Например, опытный врач может упорядочить группу больных по состоянию здоровья, не используя при этом числовой характеристики, описывающей состояние здоровья. Подобные ситуации характерны не только для медицинской практики.

3.5 Сравнение двух векторных выборок

При обработке экспериментальных данных часто приходится сравнивать на однородность две векторные выборки. Постановка задачи вполне аналогична скалярному случаю. Проводится серия из n_1 независимых экспериментов, в которых регистрируется некоторая векторная случайная величина. Пусть ее значение $X_1, \ldots, X_{n_1} \in R^n$. Затем условия эксперимента меняются. Проводится еще n_2 независимых опытов. Значения случайной величины суть $Y_1, \ldots, Y_{n_2} \in R^n$. Спрашивается: принадлежат ли выборки одной генеральной совокупности? Другими словами, изменилось ли распределение случайной величины во второй серии экспериментов. Отметим, что само распределение нам неизвестно. Для ответа на поставленный вопрос об однородности векторных выборок одним из авторов предложен [5] удобный непараметрический критерий (γ-критерий).

Как всегда, нуль-гипотеза отражает точку зрения «скептика»: изменение условий эксперимента не повлияло на распределение случайной величины. Более точно, пусть $p_1(x)$ и $p_2(x)$ — плотности распределения случайной величины в условиях первой и второй серий экспериментов (в [5] предполагается, что плотности существуют и интегрируемы по Лебегу). Нуль-гипотеза подразумевает, что почти всюду (за исключением, быть может, множества нулевой меры) $p_1(x) = p_2(x)$.

Для проверки нуль-гипотезы предлагается следующая конструкция. Выборки объединяются. Пусть Θ_1 — число элементов выборки $X_1, \ldots, X_{n_1}$, ближайшими для которых в объединенной выборке являются элементы той же выборки $X_1, \ldots, X_{n_1}$. Аналогично, Θ_2 — число элементов выборки $Y_1, \ldots, Y_{n_2}$, для которых ближайшими в объединенной выборке являются элементы выборки $Y_1, \ldots, Y_{n_2}$. Естественно, тривиальная близость — элемент является ближайшим к себе — не рассматривается.

Далее вводится случайная величина (статистика)

$$\gamma_n = \sqrt{n}\left(\frac{\Theta_1}{n_1} + \frac{\Theta_2}{n_2} - 1\right), \tag{2}$$

где $n = n_1 + n_2$ — длина объединенной выборки.

Если нуль-гипотеза верна, то вычисляется математическое ожидание случайной величины γ_n и оценивается ее дисперсия:

$$\mathrm{M}\gamma_n = -\frac{\sqrt{n}}{n-1},$$
$$\mathrm{D}\gamma_n \leqslant 2 \cdot \frac{n-2}{n-3}, \qquad n \geqslant 4.$$

Первая формула выводится весьма просто. Пусть ξ_i — случайная величина, связанная с X_i. Она принимает единичное значение, если ближайшим к X_i элементу в объединенной выборке является один из элементов $X_1, \ldots, X_{n_1}$. В противном случае ее значение равно нулю. Случайная величина ξ_i называется индикатором. Аналогично вводятся индикаторы η_i для выборки $Y_1, \ldots, Y_{n_2}$. Легко найти значения вероятностей:

$$\mathrm{P}(\xi_i = 1) = \frac{n_1 - 1}{n-1}, \qquad \mathrm{P}(\eta_i = 1) = \frac{n_2 - 1}{n-1}.$$

(Годятся рассуждения с урной, наполненной черными и белыми шарами.) Теперь

$$\Theta_1 = \sum_{i=1}^{n_1} \xi_i, \qquad \Theta_2 = \sum_{i=1}^{n_2} \eta_i.$$

Для математических ожиданий имеем:

$$\mathrm{M}\Theta_1 = \sum_{i=1}^{n_1} \mathrm{M}\xi_1 = n_1 \cdot \frac{n_1 - 1}{n-1},$$
$$\mathrm{M}\Theta_2 = \sum_{i=1}^{n_2} \mathrm{M}\eta_1 = n_2 \cdot \frac{n_2 - 1}{n-1}.$$

(Ясно, что $\mathrm{M}\xi_i = \mathrm{P}(\xi_i = 1)$ и $\mathrm{M}\eta_i = \mathrm{P}(\eta_i = 1)$.) Для математического ожидания самой величины γ_n получаем:

$$\mathrm{M}\gamma_n = \sqrt{n}\left(\frac{1}{n_1}\mathrm{M}\Theta_1 + \frac{1}{n_2}\mathrm{M}\Theta_2 - 1\right) =$$
$$= \sqrt{n}\left(\frac{n_1 - 1}{n-1} + \frac{n_2 - 1}{n-1} - 1\right) = \sqrt{n}\left(\frac{n-2}{n-1} - 1\right) = -\frac{\sqrt{n}}{n-1}.$$

К сожалению, дисперсия $\mathrm{D}\gamma_n$ оценивается сложнее в силу зависимости индикаторов.

Из конечности дисперсии и стремления к нулю (при $n \to \infty$) математического ожидания следует.

Утверждение 1. *Если нуль-гипотеза верна, то для любого $\varepsilon > 0$ существует константа $C(\varepsilon)$, зависящая только от ε, такая, что для всех n вероятность $\mathrm{P}(\gamma_n < C(\varepsilon)) > 1 - \varepsilon$.*

Другими словами, при справедливости нуль-гипотезы большое значение γ_n — редкое событие.

Альтернативой к нуль-гипотезе является предположение, что неравенство $p_1(x) \neq p_2(x)$ выполнено на некотором множестве ненулевой меры. Это — сложная гипотеза. В работе [6] доказано следующее

Утверждение 2. *Если нуль-гипотеза неверна, то для любого C выполнено* $\mathrm{P}(\gamma_n \leqslant C) \to 0$ *при* $n \to \infty$.

Таким образом, в условиях альтернативной гипотезы величина γ_n, как правило, неограниченно растет с увеличением n (в вероятностном смысле).

Практически нуль-гипотезу можно отклонить, если оказалось, что

$$\gamma_n > \mathrm{M}\gamma_n + 2\sqrt{\mathrm{D}\gamma_n}.$$

Данный критерий оказался действенным, в частности, при анализе узоров электрической активности коры головного мозга [7].

Добавление: О датчиках случайных чисел

Всякий, кто питает слабость к арифметическим методам
получения случайных чисел, грешен вне всяких сомнений.
Джон фон Нейман

Алгоритмические датчики случайных чисел появились почти одновременно с первыми вычислительными машинами. Первые датчики были капризными и все их пользователи внимательно относились к проверке качества получаемых случайных чисел. По мере совершенствования датчиков их качество повысилось, а интерес к методам проверки датчиков упал, поскольку в большинстве случаев простые тесты стали давать устойчивый положительный результат. Со временем датчики случайных чисел перешли в разряд стандартных алгоритмов и их стали встраивать в стандартные библиотеки, поставляемые с компиляторами. В результате в большинстве случаев встроенный датчик воспринимается как нечто само собой разумеющееся (пользователи обычно не отдают себе отчета даже в том, что в одних случаях датчик встроен в операционную систему, а в других — в стандартную библиотеку компилятора) и вопрос о качестве генерируемых случайных чисел не рассматривается вовсе. Тем не менее неудачные датчики получили широкое распространение именно в качестве встроенных датчиков. Истории о неверных результатах, полученных в результате использования плохих датчиков, прочно вошли в фольклор.

В большинстве случаев простые датчики используют линейный конгруэнтный алгоритм генерации случайных чисел. Этот алгоритм заключается в итеративном применении следующей формулы:

$$X_{k+1} = (aX_k + c) \bmod m,$$

где $a > 0$, $c \geqslant 0$, $m > 0$ — некоторые целочисленные константы. Получаемая последовательность зависит от выбора стартового числа X_0 и при разных его значениях получаются различные последовательности случайных чисел. В то же время многие свойства последовательности X_j определяются выбором коэффициентов в формуле и не зависят от выбора стартового числа. Ясно, что последовательность чисел, генерируемая таким алгоритмом, периодична с периодом, не большим чем m, поэтому число m обычно выбирают большим. При реализации выгодно выбирать $m = 2^e$, где e — число битов в двоичном представлении целого беззнакового числа, поскольку это позволяет избавиться от относительно медленной операции деления по модулю. Статистические свойства получаемой последовательности случайных чисел полностью определяются выбором констант a и c при заданной разрядности e. Для этих констант выписаны условия, гарантирующие удовлетворительное качество получаемых случайных чисел. В частности, известно, что младшие двоичные разряды сгенерированных таким образом случайных чисел демонстрируют поведение, далекое от случайного, поэтому рекомендуется использовать только старшие разряды.

Стандарт ANSI-C определяет (`stdlib.h`) функцию `rand()`, выдающую равномерно распределенное целое случайное число в диапазоне от `0` до `RAND_MAX`. Функция `srand()` устанавливает начальное значение датчика. Комиссия по стандарту ANSI-C опубликовала, хотя и не рекомендовала использовать[4], следующий пример реализации датчика:

```
#define RAND_MAX 32767

static unsigned long next_rand = 1;

int rand(void) {
  next_rand = next_rand * 1103515245 + 12345;
  return (unsigned int)(next_rand/65536)%32768;
}

void srand( unsigned int seed ){
  next_rand = seed;
}
```

В этой реализации выбраны значения констант $a = 1103515245$, $c = 12345$ и вычисления производятся в 32-битной арифметике, но сгенерированное число преобразуется в диапазон $[0, 32767]$ путем отбрасывания младших 16 разрядов и обнуления старшего разряда. Таким образом, в этой реализации обходится проблема с плохим поведением младших разрядов, но при этом происходит значительное уменьшение разрешения датчика. Действительно, мы получим всего $N = 32768$ различных случайных чисел, что оказывается неудовлетворительным для большинства статистических приложений. На самом деле выбор значений для констант a и c также не является оптимальным с точки зрения качества получаемых случайных чисел. Многие встроенные датчики используют приведенную выше реализацию (например Watcom C/C++), в некоторых случаях бывают

[4]Точную формулировку можно найти в следующем документе: *American National Standards Institute, International Standards Organisation, 1990, X3J11, ANSI X3.159, ISO/IEC 9899:1990.*

использованы другие значения констант (например, во всех C/C++ компиляторах фирмы Borland использованы значения $a = 22695477$, $c = 1$), но почти всегда у приведенной реализации заимствуется исключение потенциально опасных младших разрядов и ограничение на диапазон изменения получаемого случайного числа.

Внимательно изучив приведенный выше код, можно убедиться, что на самом деле приведение случайного числа к диапазону $[0, N - 1]$ производится только применительно к результату, возвращаемому функцией, а для получения следующего случайного числа используется непосредственно 32-битное число, полученное на предыдущем шаге. Из этого можно сделать вывод, что хотя генерируются случайные числа из небольшого диапазона, тем не менее период получаемой последовательности значительно превосходит N и может достигать величины 2^e. Многие алгоритмы, использующие случайные числа, оперируют d-мерными векторами со случайными компонентами. Казалось бы, можно надеяться, что алгоритм, оперирующий d-мерными векторами, составленными из последовательных случайных чисел, сможет сгенерировать N^d различных векторов, если $N^d \leqslant 2^e$. Тем не менее доказано[5] следующее важное свойство случайных чисел, сгенерированных с помощью линейного конгруэнтного алгоритма: *случайные точки d-мерного пространства лежат на конечном семействе $(d-1)$-мерных гиперплоскостей*. Значения коэффициентов a и c определяют количество гиперплоскостей и расстояние между ними и должны выбираться таким образом, чтобы получить по возможности равномерное заполнение d-мерного пространства при всех относительно небольших значениях d. Интересно, что если ограничиться 32-битной арифметикой и случаем $m = 2^e$ при $e = 32$, то оптимальными оказываются значения коэффициентов, принятые в компиляторах фирмы Borland.

Описанный выше результат делает невозможным использование линейных конгруэнтных датчиков в статистических алгоритмах, требующих высокого разрешения. В связи с этим линейный конгруэнтный алгоритм постепенно потерял свою популярность и его место заняло семейство фибоначчиевых алгоритмов[6], которые могут быть рекомендованы для использования в алгоритмах, критичных к качеству случайных чисел. Наибольшую популярность фибоначчиевы датчики получили в связи с тем, что скорость выполнения арифметических операций с вещественными числами сравнялась со скоростью целочисленной арифметики, а фибоначчиевы датчики естественно реализуются в вещественной арифметике[7]. При этом исчезает необходимость выполнять операцию перехода от целого

[5]См. *Marsaglia G. «Random numbers fall mainly in the planes», Proc. Nat. Acad. Sciences USA, 61 (1968), pp. 25–28.*

[6]См. *Marsaglia G. and Zaman A. «A New Class of Random Number Generators» Annals of Applied Probability, no. 3, vol. 3, pp. 462–480.* В англоязычной литературе фибоначчиевы датчики такого типа называют обычно «Subtract-with-borrow Generators» (SWBG).

[7]Первые реализации таких датчиков работали в целых числах. Реализацию в вещественной арифметике впервые предложили Каханер и Марсалья (D. Kahaner, G. Marsaglia), но по сути она ничем не отличается от целочисленной. Краткое описание «канонической» вещественной реализации на фортране приведено в книге *Каханер Д., Моулер К., Нэш С., Численные методы и программное обеспечение, М.: Мир, 1999.* Исходный код этой реализации можно найти по адресу {`ftp://ftp.mathworks.com/pub/books/kahaner/ch10/duni.f`}.

случайного числа к вещественному и появляется возможность получать вещественные случайные числа, все разряды которых случайны.

Рекомендуемый нами фибоначчиев датчик основан на следующей итеративной формуле:

$$X_k = \begin{cases} X_{k-a} - X_{k-b}, & X_{k-a} \geqslant X_{k-b}, \\ X_{k-a} - X_{k-b} + 1, & X_{k-a} < X_{k-b}, \end{cases} \tag{3}$$

где X_k — вещественные числа из диапазона $[0, 1)$, a, b — целые положительные числа, называемые *лагами*. Для работы фибоначчиеву датчику требуется знать $\max(a, b)$ предыдущих сгенерированных случайных чисел. При программной реализации для хранения сгенерированных случайных чисел используется конечная циклическая очередь на базе массива. Для старта фибоначчиевому датчику требуется $\max(a, b)$ случайных чисел, которые могут быть сгенерированы простым конгруэнтным датчиком. Ниже приведена реализация фибоначчиева алгоритма генерации случайных чисел на языке ANSI-C, построенная на базе стандартного датчика и позволяющая получать вещественные случайные числа, равномерно распределенные на интервале $[0, 1)$.

```
#define REAL_BITS 64
#define SH0 55
#define SH1 24

static int index = -1;
static double buf[SH0];

void f_srand( unsigned int seed ){
  int j, k;
  double val;

  srand(seed);
  for( j = 0; j < SH0; j++ ){
    buf[j] = 0.0;
    val = 0.5;
    for( k = 0; k < REAL_BITS; k++, val /= 2.0 )
      buf[j] += (rand() > RAND_MAX/2) ? 0.0 : val;
  }
  index = 0;
}

double f_rand() {
  double res;

  if( index == -1 ) f_srand(1);
  res = buf[index] - buf[(index + SH1) % SH0];
```

```
  if( res < 0.0 ) res += 1.0;
  buf[index] = res;
  index = (index - 1 + SH0) % SH0;
  return res;
}
```

Лагам a и b соответствуют константы SH0 и SH1. Значения этих констант — «магические» и их не следует выбирать произвольно. Кроме указанных допустимы также следующие значения для этих констант: SH0 = 17, SH1 = 5 и SH0 = 97, SH1 = 33. Качество получаемых случайных чисел зависит от значения константы SH0 — чем оно больше, тем выше размерность пространства, в котором сохраняется равномерность случайных векторов, образованных из полученных случайных чисел. В то же время с увеличением величины константы SH0 увеличивается объем используемой алгоритмом памяти. Значения SH0 = 17, SH1 = 5 можно рекомендовать для простых приложений, не использующих векторы высокой размерности со случайными компонентами. Значения SH0 = 55, SH1 = 24 позволяют получать числа, удовлетворительные для большинства алгоритмов, требовательных к качеству случайных чисел. Значения SH0 = 97, SH1 = 33 позволяют получать очень качественные случайные числа и используются в алгоритмах, работающих со случайными векторами высокой размерности. Описанный фибоначчиев датчик случайных чисел (с лагами 20 и 5) используется в широко известной системе Matlab[8] начиная с версии 5.0 (автором первой версии этой системы был Д. Каханер).

Отметим, что стандартный датчик не требует обязательной инициализации функцией srand(), что тривиально достигается путем статической инициализации переменной, хранящей состояние датчика. В нашей реализации аналогичное поведение достигается более сложно — проверка производится при каждом вызове функции f_rand() и при первом вызове для неинициализированного датчика вызывается функция инициализации. Это приводит к ненужным затратам времени при каждом вызове функции f_rand(). Если приложение, использующее датчик, может гарантировать вызов функции инициализации f_srand() перед первым вызовом функции f_rand(), то временные характеристики фибоначчиева датчика можно слегка улучшить, убрав первую строку в функции f_rand().

Важной особенностью приведенного исходного кода является переносимая реализация функции f_srand(), которая случайно инициализирует все биты вещественных чисел из массива buf, используя стандартный датчик. Для корректности этой функции величина константы REAL_BITS должна быть не меньше числа битов в мантиссе вещественного числа. Значение, использованное в приведенной реализации, достаточно для всех реализаций, в которых sizeof(double) <= 8 (для вещественной арифметики стандарта IEEE значение этой константы можно уменьшить до 53).

Получаемые случайные числа обладают хорошими статистическими свойствами, причем все биты случайного числа равнозначны по статистическим свой-

[8] Matlab – зарегистрированная торговая марка фирмы The Math Works, Inc.

ствам. Период фибоначчиева датчика может быть оценен по следующей формуле:

$$T = (2^{\max(a,b)} - 1)2^e,$$

где e — число битов в мантиссе вещественного числа. Для стандартной IEEE-арифметики двойной точности (тип `double`, $e = 53$) и приведенных в листинге значений лагов период составляет $p \approx 3 \cdot 10^{32}$.

Список литературы

1. Левин А.Ю., Майоров В.В. О логике математической статистики: Текст лекций по курсу «Дополнительные главы математической статистики». Ярославль, 1989. 44 с.

2. Крамер Г. Математические методы статистики. М., 1948.

3. Гнеденко В.В. Курс теории вероятностей. М.; Л.: Гостехтеоретиздат, 1954.

4. Боровков А.А. Математическая статистика. М.: Наука, 1984.

5. Левин А.Ю. О состоятельном многомерном непараметрическом критерии однородности // Усп. матем. наук. 1993. Т. 48, № 6. С. 155–156.

6. Левин А.Ю. О состоятельном многомерном непараметрическом критерии однородности // Моделирование и анализ информационных систем. 2002. Т 9, № 2.

7. Мышкин И.Ю., Майоров В.В. Оценка различий вызванных корковых потенциалов на зрительные стимулы у человека // ЖВНД. 1996. Т. 46, вып. 3. С. 488–495.

Содержание

Левин Анатолий Юрьевич,
Майоров Вячеслав Владимирович,
Мячин Михаил Леонидович

О логике математической статистики.
Текст лекций по курсу
«Дополнительные главы математической статистики»

Корректор А.А. Аладьева
Верстка М.Л. Мячин

Подписано в печать 02.11.2003. Формат 60x84/8.
Бумага тип. Печать офсетная.
Усл. печ. л. 5,11. Уч.-изд. л. 3,5.
Тираж 100 экз. Заказ

Оригинал-макет подготовлен
в редакционно-издательском отделе ЯрГУ
150000 г. Ярославль, ул. Советская, 14
Отпечатано на ризографе.