И. Г. Венецкий

Вероятностные методы в демографии

ББК 60.7
В29

Венецкий И. Г.

В29 Вероятностные методы в демографии.— М.: Финансы и статистика, 1981.— 223 с., ил.

В пер.: 2 р. 50 к.

Книга посвящена использованию вероятностных методов для изучения различных демографических явлений: повозрастной смертности, средней продолжительности жизни, соотношения полов среди родившихся, влияния разных болезней на показатель смертности и др. Изучение демографических проблем и получение выводов осуществляется путем рассмотрения конкретных ситуаций и условий, возникающих перед исследователем, и подробного решения поставленных задач с привлечением соответствующих вероятностных методов.

Книга рассчитана на демографов, социологов, экономистов, медицинских работников.

В $\frac{10805-145}{010(01)-81}$ 47—81 (С) 0703000000

ББК 60.7
312

Введение

Первые опыты привлечения математики для изучения социальных, в том числе демографических, явлений относятся ко второй половине XVII в. С появлением теории вероятностей, основываясь на статистических данных, исследователи могли уже составить представление о таких явлениях, как повозрастная смертность, влияние различных болезней на смертность, продолжительность жизни и т. д. Современная демография широко использует достижения теории вероятностей, находящей новые способы отражения окружающей нас объективной социальной действительности, выражающей с помощью количественных характеристик конкретные взаимоотношения между изучаемыми явлениями и открывающей большие возможности для дальнейшего развития демографической теории и практики.

На нынешнем этапе строительства коммунистического общества в СССР задачи демографии состоят не только в объяснении демографических процессов, а главным образом в весьма трудном и ответственном изучении демографических явлений и множественных взаимосвязей, выявлении главных факторов и разработке системы практических мероприятий, позволяющих активно и планомерно воздействовать на эти процессы.

При описании демографических явлений с помощью вероятностных методов следует исходить из того, что объяснение выявленных путем массовых наблюдений соотношений и взаимозависимостей должно основываться в первую очередь на качественном анализе изучаемых явлений, на основных положениях демографической науки.

Главное в демографии *не вероятностная интерпретация событий и явлений, а принцип социально-экономической их обусловленности, рассмотрение демографических явлений в сочетании с условиями общественно-экономической жизни. Признание главенствующей роли общественно-экономических факторов является отличительной чертой марксистского подхода к вопросам демографии.*

Следовательно, вероятностные методы исследования количественных характеристик демографических явлений должны быть тесно связаны со знанием качественной стороны этих явлений. Без опоры на качественный анализ лежащие на поверхности факты могут быть приняты за конечное объяснение. Основываясь на одних вероятностных принципах, нельзя объяснить многие демографические явления: резкое падение рождаемости в нашей стране в 60-х годах, «демографический взрыв» в развивающихся странах, «старение» населения и др.

Чем отличается подход советских исследователей к демографическим явлениям от подхода буржуазных социологов и демографов? Прежде всего порочна методология многих буржуазных исследователей, которые, обращаясь к использованию вероятностных методов при изучении демографических явлений, главное и основное внимание уделяют биологической стороне, т. е. биологическим факторам, а социальной и экономической стороне человеческих отношений придается второстепенное, весьма малое значение.

Буржуазные социологи, статистики и демографы закрывают глаза на социальную действительность и делают необоснованные выводы из произвольно толкуемых и сомнительных фактов. Например, при оценке социальной мобильности делается вывод о «широких возможностях» продвижения представителей низших классов вверх по иерархической лестнице.

Как следует методически правильно изучать этот вопрос? Очевидно, надо прежде всего учесть совокупность родившихся за определенный период и достигших конкретного возраста и положения в обществе, опросить их о занятиях отцов и дедов и затем рассчитать вероятность перехода из одних социальных групп в другие. Тогда выяснилось бы, что в мире, основанном на частной собственности, вероятности перехода вертикально вверх от низов общества до его вершины ничтожно малы: сын фабриканта наверняка останется фабрикантом, сын рабочего — рабочим, а иногда и безработным.

Апологетический характер исследований буржуазных социологов и демографов ярко проявляется в бессмысленных упражнениях с математическими моделями, с помощью которых оцениваются вероятности того, что потомок рабочего через два-три десятка поколений окажется богачом-мультимиллионером.

В социалистическом обществе, где действительно имеются возможности расцвета личности, принцип социализма «от каждого по способностям, каждому по труду» находит свою полную реализацию.

В данной книге разъясняются основные идеи и возможные пути приложения теории вероятностей к описанию демографических явлений. Здесь используется главным образом вероятностный аппарат, известный статистикам, демографам и социологам из учебников и учебных пособий по математической статистике для вузов, одним из разделов которой является теория вероятностей. Таким образом, круг вопросов, включенных в работу, ограничивается в основном «классической» теорией вероятностей, без рассмотрения теоретико-вероятностных методов, которые хотя и используются в ряде случаев в демографическом анализе, но выходят за пределы программных положений указанных выше специальностей, так как они требуют гораздо бо́льших математических

знаний в области теории множеств, исследований операций, распознавания образов, случайных функций и т. д., чем те, которыми обладают статистики, демографы и социологи.

В настоящее время наблюдается разрыв между новыми подходами, применяемыми при статистической, математико-статистической и вероятностной обработке массовых демографических явлений, с одной стороны, и трудностями, испытываемыми практическими работниками в области статистики, демографии и социологии в понимании используемого при этом сложного математического аппарата, — с другой.

На наш взгляд, одной из наиболее важных причин этого разрыва является все еще недостаточный уровень статистической и математико-статистической подготовки практических работников (статистиков, демографов и социологов), не позволяющий развивать и совершенствовать в должной мере статистическое и математико-статистическое мышление.

Вместе с тем мы считаем, что экономисты-статистики, демографы и социологи не могут и не должны превращаться в математиков. «Именно экономисты, а не математики или механизаторы, обязаны строго следить за соблюдением принципа примата качества в статистике вообще, в применении математических методов в частности»[1]. Статистикам и социологам, занимающимся демографией, научиться прикладной математике, в том числе теории вероятностей, легче, чем специалисту-математику разобраться в специфике социальных, статистических и демографических проблем.

Опыт показывает, что привлечение статистиками, социологами и демографами вероятностных методов в качестве аппарата для практических исследований более плодотворно, чем привлечение математиками и специалистами в области теории вероятностей статистики, социологии и демографии в качестве объекта приложения своих знаний.

Именно поэтому об использовании методов математической статистики и теории вероятностей в демографии должны писать специалисты-статистики, социологи и демографы, достаточно осведомленные о возможностях математической статистики и теории вероятностей и имеющие опыт практического их использования при решении конкретных задач демографической науки.

Изучить основы теории вероятностей — особый язык, подчиненный жестким и строгим правилам логики, на котором можно описывать многие различные явления, происходящие в человеческом обществе, — довольно трудно; еще труднее описывать конкретные демографические явления с помощью вероятностных схем, производить расчет ошибок, находить показатели существенности различий демографических характеристик и т. д. Но пожалуй, самое ответственное и, следовательно, самое трудное — выявить возможность применения тех или иных вероятностных методов в демогра-

[1] *Боярский А. Я.* Корреляция или испорченная функция? — Вестник статистики, 1980, № 1, с. 63.

фическом анализе без нарушения ограничительных условий (т. е. правомерность использования теории вероятностей), интерпретировать со знанием специфики изучаемого явления полученные результаты. Это требует не формального скольжения по поверхности явления, а проникновения в его суть. Данным вопросам в работе уделяется большое внимание.

В ряде случаев вместо детального описания теоретических положений, определений и формул теории вероятностей (когда они уже известны читателю из вузовского курса) в книге дается либо весьма краткое их изложение, либо ссылки на источники, где они достаточно подробно изложены.

Чтобы сделать книгу доступной для демографов, статистиков и социологов, в нее не включены сложные выкладки, затрудняющие чтение, не применяется символика, малопонятная нематематикам. Автор стремился к тому, чтобы сложность математического аппарата не была чрезмерной, а проблемы вычислений не подменяли проблемы демографии. Цель книги — оказать помощь читателю при использовании им некоторых вероятностных методов решения демографических, медико-статистических, социальных и других задач.

Разумеется, что многие практические примеры упрощены, так как имеют характер числовой иллюстрации основных приемов вероятностных действий, привлекаемых методов, правил и теорем. Овладение этими методами и приемами поможет при самостоятельном изучении реальных демографических процессов.

Вероятностные методы все шире используются в журнальных статьях и книгах по проблемам демографии. Поэтому очень важно сориентировать читателя на самостоятельные изыскания в этой области: как в смысле выбора из большого арсенала методов теории вероятностей именно того, который соответствует решению конкретных задач, так и в смысле привлечения литературных источников.

Необходимо учитывать, что использование вероятностных показателей связано с точностью исходных данных. Разумеется, погоня за мнимой точностью никому не нужна. Вместе с тем следует иметь в виду, что информацию о демографических явлениях дают статистические наблюдения (сплошные и выборочные). При этом возможны различные неточности, в частности численностей отдельных групп при группировке по разным признакам, а также искажения самих возрастов живущих и умерших. Исследователь, привлекающий вероятностные показатели в своих расчетах, должен быть полностью уверен в том, что погрешности величины вероятностей никак не повлияют на окончательные выводы о закономерностях изучаемых явлений, тенденциях их изменения и т. д.

В ряде случаев простота статистического решения той или иной демографической задачи становится более важной, чем предполагаемая точность результата. Это означает, что привлечение более сложного математико-статистического аппарата не сможет привести к получению более точного результата.

Изложение демографических проблем и получение соответствующих выводов осуществляются в работе путем рассмотрения конкретных ситуаций и условий, возникающих перед исследователем, и достаточно подробного разбора и решения поставленных задач на основе привлечения и использования тех вероятностных приемов и методов, которые могут быть применены в подобных случаях.

Используя вероятностные методы, мы пытаемся добиться легкости восприятия всех построений, представления о правильности исходных предположений, простоты и наглядности. Однако не следует переоценивать значение простоты и наглядности. Из того факта, что не все эти построения просты и наглядны, не следует никаких выводов. Здесь уместно вспомнить слова К. Маркса о том, что, если бы все было наглядно, не нужна была бы и наука.

Данная монография является логическим продолжением ранее опубликованных работ, в которых частично рассматривались и вероятностные методы[1]. Учитывая необходимость последовательного изложения материала, в предлагаемой читателю книге допущено некоторое повторение в очень кратком виде наиболее существенного и важного в теории вероятностей из этих работ.

В ряде случаев вероятностные задачи настолько сложны для теоретического решения, что становится необходимым проведение статистического эксперимента с использованием электронно-вычислительных машин для обработки данных.

Статистики, демографы и социологи с удовлетворением могут отметить, что вероятностные расчеты брачности, рождаемости, заболеваний, миграций и смертности, выводы и предсказания о закономерностях массовых процессов стали в настоящее время не менее достоверными, чем если бы эти расчеты производились на основе точного знания законов единичных явлений.

---

[1] См.: *Венецкий И. Г.* Математические методы в демографии. М., 1971; *Он же.* Статистические методы в демографии. М., 1977.

Глава 1

Теория вероятностей в демографии

При статистической обработке демографических данных используются определения и правила, установленные теорией вероятностей. Теория вероятностей как раздел математики возникла в середине XVII столетия в результате поисков закономерности осуществления или неосуществления явлений. Основные понятия этой теории зарождались в работах Б. Паскаля, П. Ферма и Х. Гюйгенса. Дальнейшее развитие теории вероятностей связано с именем Я. Бернулли, доказавшего теорему, впоследствии названную законом больших чисел.

Суть этой теоремы состоит в том, что при изучении совокупностей, элементы которой носят вероятностный характер, при малом числе наблюдений изучаемое явление может показаться случайным, не подчиняющимся видимым закономерностям. Однако если такие явления подвергаются массовому наблюдению, то в них очень часто обнаруживаются определенные закономерности. В этом проявляется действие закона больших чисел — объективного закона, состоящего во взаимопогашении влияния второстепенных, не определяющих существо явлений факторов. Благодаря такому взаимопогашению влияния второстепенных факторов рельефно проявляются внутренние свойства данного явления, которые выступают затем в виде массовой закономерности.

До сих пор многие окутывают закон больших чисел дымкой мистицизма, говорят о статистическом фатализме и т. д. Вот пример упрощенного, а следовательно, неправильного, понимания закона больших чисел и закономерности массового процесса. В статье Я. Голованова «Посольство Нептуна», в целом хорошей и во многом правильной, автор отмечает, что в плавании наши спортсмены не могут считать себя признанными мировыми лидерами. Как же выправить положение? По мнению автора статьи, делать ничего не надо. В московском Дворце водного спорта «...появление чемпионов запрограммировано массовостью. Чемпионов не ищут, они сами непременно отыщутся, они просто не могут не отыскаться. Здесь, — пишет автор, — в силу вступает закон больших чисел, кривая вероятностей Гаусса... Появление чемпионов предопределено»[1]. Что же это за мудрое провидение, программирующее появление чемпионов и позволяющее ждать их сложа руки? На наш взгляд, нет ничего более нелепого, чем ориентировать на «предопределенность появления чемпионов». Будущих чемпионов, талантливых спортсменов надо терпеливо искать, а найдя, не менее терпеливо воспитывать.

Или вот представление одной студентки: «Я много читала о том, будто статистика доказывает наличие судьбы, богом установленного порядка. Число несчастных случаев из года в год постоянно, число смертей — тоже. Даже рождение определенного числа мальчиков и девочек как бы говорит о том, что свыше установлен какой-то порядок. Я не верю этому, но опровергнуть не могу. Что это за «мистика чисел»?[2] Студентка опровергнуть этот «божественный порядок» не может. Она не знает о существовании постоянных соотношений, определяемых в одних случаях естественными законами природы, а в других — законами развития социальных и экономических явлений. Между тем в этих соотношениях нет никакой «мистики чисел» или «божественного порядка».

Как раскрывали существо закона больших чисел классики марксизма-ленинизма?

К. Маркс считал, что закон больших чисел это такой закон, который при наблюдении значительного числа случайных явлений позволяет правилу прокладывать себе путь сквозь беспорядочный хаос, фиксирует погашение индивидуальных отклонений и проявляется в форме средних величин[3]. В. И. Ленин называл закон больших чисел массовой закономерностью и указывал, что изучение массового требует исследования каждого индивидуального и задача состоит в переходе от случайного и единичного к устойчивому массовому. При этом закон изучаемого явления, как правило, не совпадает с внешней формой его проявления. В. И. Ленин считал непременным условием успешного анализа фактов не арифметическое усердие при изучении столбцов цифр, а правильное понимание сущности изучаемых явлений, законов их развития. Нарушение этого условия, как отмечал В. И. Ленин, приводит к статистическому кретинизму.

Особенно важным для нас является понимание того, что в социалистическом обществе наряду со случайными процессами чаще имеют место процессы, регулируемые планово организованной, сознательной деятельностью человека, которая ограничивает действие закона больших чисел. В дальнейшем мы рассмотрим механизм действия закона больших чисел на демографических примерах.

Наиболее важный этап в развитии теории вероятностей связан с именами П. Л. Чебышева и его учеников — А. А. Маркова и

[1] Комсомольская правда, 1977, 17 июля.

[2] *Яхот О. О.* К вопросу о статистической закономерности. — В кн.: Ученые записки по статистике АН СССР. М., 1961, т. 6, с. 58.

[3] См.: *Маркс К., Энгельс Ф.* Соч., т. 25, ч. II. с. 396.

А. М. Ляпунова. Велика заслуга русских и советских ученых С. Н. Бернштейна, В. И. Романовского, А. Н. Колмогорова, А. Я. Хинчина, Е. Е. Слуцкого, Б. В. Гнеденко, Н. В. Смирнова, Б. С. Ястремского и др., которые своими исследованиями внесли крупный вклад в развитие теории вероятностей и сделали важные открытия, приведшие к созданию новых ее направлений.

1.1. О возможности привлечения теории вероятностей к изучению демографических явлений

Длительное время в нашей науке вероятностный подход к экономическим и социальным, в том числе демографическим, явлениям полностью отвергался.

Такое отрицательное отношение к теории вероятностей, сомнение в ее возможностях и приложимости к научным исследованиям, в частности к выборке, мешало ее развитию. Один из советских статистиков проф. П. П. Маслов в весьма категоричной форме выступил против использования вероятностного подхода при выборочном изучении социально-экономических и демографических явлений. Его критические замечания по поводу привлечения закона распределения Гаусса и вероятных ошибок средней были конкретно направлены против выборочного изучения грамотности населения Московской губернии и денежной оплаты трудодня, а в более широком смысле отрицали возможность использования теории вероятностей при выборочных исследованиях общественных и экономических явлений.

Действительно, применение теории вероятностей в любых случаях, без предварительного выяснения правомерности такого использования, возможного лишь при соблюдении некоторых ограничительных условий, в частности осуществления принципа случайности явлений, их равновозможности и независимости, лишено всякого познавательного смысла, хотя формально с точки зрения вычислительной техники (определение средней ошибки, прибавление и вычитание к полученным средним утроенной средней ошибки и т. д.) может оказаться правильным. В этом случае использование теории вероятностей будет вводить в заблуждение и приведет к неправильным выводам, отрыву от реальной действительности.

При отсутствии условий применимости вероятностных схем происходит скольжение по поверхности явлений, отрыв от главного, от реализации возможностей проникновений в сущность изучаемых явлений.

Коренной чертой социалистической экономики является общественный контроль производства и его планомерность, однако это не противоречит признанию вероятностного характера протекания некоторых экономических процессов. А так как экономические и демографические процессы взаимосвязаны, совершенно очевидно, что они во многом определяются «законами случая». Характеризуя главные черты социализма в «Критике Готской программы», К. Маркс указывал, что экономическая необходимость планомерного распределения основных частей совокупного общественного продукта при социализме будет определяться «...на основе наличных средств и сил, отчасти на основе теории вероятности...»[1].

Многие экономисты эти слова К. Маркса связывают только с действием стихийных элементов в экономике, с природными бедствиями и др. Однако эти слова можно понимать более широко. Дело в том, что плановость социалистической экономики не отрицает фактора неопределенности, связанного с рядом причин.

Вот почему мы считаем, что при изучении экономических и демографических явлений и процессов надо обязательно принимать во внимание случайные факторы, продолжающие играть некоторую, хотя и ограниченную роль. Возьмем такие демографические события, как рождаемость, заболеваемость, смертность и другие, представляющие собой ту часть социологических и общественных явлений, в которых медицинские и биологические факторы не потеряли своего значения. Ясно, что при их изучении следует принимать во внимание сопровождающие их случайные обстоятельства и не пренебрегать вероятностными методами их исследования, а, напротив, — брать на вооружение.

Советский статистик И. П. Суслов пишет: «Игнорирование наличия вероятностных свойств у общественных явлений, упрощенный и однозначный подход к ним приводят к просчетам в планировании и управлении народным хозяйством»[2]. В настоящее время вероятностный подход весьма успешно применяется при выборочном изучении демографических явлений, в частности при проведении Всесоюзных переписей населения. Привлекая теорию вероятностей к выборочному изучению демографических явлений, мы даем вероятностное толкование не материальной природе демографических фактов, а случайной форме проявления этих фактов.

Многие социально-экономические условия жизни общества имеют определяющее влияние на интенсивность указанных выше демографических процессов. Иными словами, характер, уровень и динамика демографических факторов не стихийно случайны, а регулируются социальными факторами и общими условиями развития общества. Вот почему демографические процессы неадекватны чисто случайным процессам, и, следовательно, применение вероятностных расчетов должно быть ограничено.

Однако к демографическим фактам нельзя подходить как к жестко детерминированным, им присущи в большой степени также неопределенность и случайность, которые можно учесть только путем привлечения теории вероятностей.

При этом, конечно, случайности в демографических явлениях отличаются от случайностей урновых схем. В демографических явлениях нельзя непосредственно рассчитывать шансы за и против

[1] *Маркс К., Энгельс Ф.* Соч., т. 19, с. 17.

[2] *Суслов И. П.* Вероятность в системе научных категорий.— В кн.: Динамическая и вероятностная оптимизация экономики/Отв. ред. К. К. Вальтух. Новосибирск, 1978, с. 57.

событий; можно лишь, наблюдая ход событий, подсчитывать частости и постулировать их в качестве апостериорных вероятностей. Вот почему при изучении массовых явлений, описывающих демографические процессы, используют не отношение шансов, благоприятствующих событию, к общему числу шансов, а отношение числа случаев, например смерти, вступления в брак, расторжения брака, рождения ребенка, миграции и т. д., в течение определенного времени к численности совокупности, в которой происходят эти события, в начале наблюдения.

Судьба человека — жить или не жить — это не простое извлечение белых и черных шаров из урны. Схемы теории вероятностей грубы, примитивны и вследствие этого неадекватны жизненным ситуациям. Смерть того или иного лица зависит от случая. Но среди большой совокупности людей число смертей уже не зависит всецело от случая. Вероятность смерти определяется в условиях действия закона больших чисел, весьма однообразного в своем проявлении и весьма достоверного в своих результатах и выводах.

Теория вероятностей исходит из принципа повторяемости явлений. Без устойчивой и прочной повторяемости нет объективного закона. Следовательно, для правильности выводов, которые должны иметь практическое значение, надо производить большое число «испытаний». Но, привлекая теорию вероятностей к построению таблиц смертности и фиксируя, например, рождение или смерть того или иного лица, мы не можем проводить повторные «испытания», ибо человек рождается и умирает только один раз. Каков же выход?

Демографы, социологи, статистики давно уже понимают, что значительный круг вопросов, которыми они занимаются, может быть представлен с математико-статистической точки зрения как определенная «статистическая совокупность», во многих случаях достаточно массовая и однородная, а следовательно, вполне пригодная для привлечения вероятностного аппарата при обработке данных. Таким образом, выход состоит в использовании могущественного орудия статистики, т. е. группировке изучаемого населения на однородные и достаточно многочисленные группы, чтобы могло проявиться действие закона больших чисел. Такая группировка позволит проводить «испытания» над множеством людей, незначительно отличающихся друг от друга по условиям и по образу жизни, следовательно, приблизительно одинаковых по доживаемости.

Если рассматривать население или отдельную достаточно многочисленную его группу как более или менее однородную совокупность, в которой происходят случайные, независимые события, то наиболее целесообразным подходом к оценке изучаемых процессов следует признать подход, ориентированный на использование методов математической статистики, позволяющий количественно анализировать изучаемые явления вероятностным путем.

В последние годы советские исследователи все шире привлекают вероятностные методы к изучению общественно-экономических

явлений[1]. И. П. Суслов, на наш взгляд, вполне правильно пишет: «... становится очевидным, что общественные явления вне вероятностной концепции понять невозможно. Общество является сложнейшей динамической системой, развивающейся под воздействием как необходимости, что определяет его детерминированность, так и многих случайных факторов, что придает ей объективные вероятностные свойства»[2].

Касаясь вопроса об использовании теории вероятностей в демографии, советский демограф Л. Е. Дарский пишет: «... вероятностная интерпретация демографических показателей достаточно сложна, но в то же время она абсолютно необходима ... Более того, сами демографические модели должны иметь вероятностное толкование, а вероятностная природа демографического прогноза бесспорна»[3]. П. Лаплас назвал теорию вероятностей здравым смыслом, переложенным на вычисления.

Привлечение теории вероятностей к изучению социальных явлений объясняется и тем, что эти явления чрезвычайно сложны. Для практических выводов на основе их изучения одних умозрительных теоретических заключений недостаточно, необходима обработка обширной информации, которая невозможна без использования вероятностного подхода.

Приведем пример. Во многих странах мира в последние десятилетия происходит процесс «старения» населения. Суть этого процесса состоит, в частности, в том, что происходит изменение долей различных возрастных групп: увеличивается доля населения 60 лет и старше и уменьшается доля молодежи до 20 лет. Главными причинами этого явления, действующими порознь и в различных комбинациях, различные исследователи считают динамику таких демографических показателей, как смертность, средняя продолжительность жизни и рождаемость.

Какой же из трех указанных факторов представляется наиболее существенным? На что в первую очередь следует обратить внимание, чтобы остановить или хотя бы временно задержать дальнейшее развитие процесса «старения» населения?

Для решения этой задачи французский демограф Ж. Буржуа-Пиша и американский Э. Коул в результате математико-статистического анализа и с привлечением теории стабильного населения доказали, а советский ученый-статистик А. Я. Боярский подтвердил, что «старение» населения является следствием не только и даже не столько снижения смертности и увеличения продолжительности жизни, сколько снижения рождаемости. Для приостановки «старения» нужно длительное, устойчивое и достаточно большое повышение рождаемости.

---

[1] Что касается демографических процессов, то уже в 20—30-х годах М. В. Птуха, В. В. Паевский, Е. Е. Слуцкий успешно применяли к их исследованию вероятностные методы.

[2] *Суслов И. П.* Указ. соч., с. 57.

[3] Место демографии в системе наук/Под ред. О. В. Лармина. М., 1975, с. 222.

Рассмотрим другой пример. Попытаемся ответить на вопрос: все ли признаки населения нужно изучать путем сплошного наблюдения? Нельзя ли хотя бы частично использовать выборочный метод? Ответить же на этот вопрос можно лишь на основе вероятностного расчета. В какой мере возможно использование выборочного метода для изучения таких признаков населения, как общественная группа, место работы, характер и продолжительность работы, длительность проживания в данном населенном пункте или прежнем месте постоянного жительства, мотивы миграции и число рожденных детей? Вероятностный расчет показывает, что если исходить из репрезентативности выборки в пределах областей СССР (а численность населения большинства областей составляет более 400 тыс. человек), доли генеральной совокупности не менее 1% всей совокупности и предельно допустимой ошибки 0,1, то с вероятностью 0,997 абсолютный объем проектируемой выборки составит около 100 тыс. человек:

$$n = \frac{t^2 w(1-w)}{\Delta^2} = \frac{9 \cdot 0,1 \cdot 0,99}{0,000001} \approx 100\,000 \text{ человек},$$

т. е. относительная численность выборки составит 25%. Эти расчеты и легли в основу выборки при переписях населения. Так, при переписи 1979 г. по ряду показателей опрос проводился в каждом четвертом помещении.

1.2. Случайность и необходимость в демографии

Демографические явления зависят от такого большого числа факторов, что все их невозможно зарегистрировать и учесть. Первостепенные, основные факторы переплетаются с второстепенными, случайными, и последние начинают играть значительную роль.

Такие основные понятия теории вероятностей, как *случайность* и *необходимость,* встречаются еще у древних философов. Однако трактовка этих понятий не предполагала перенесения их в область количественных отношений. Ранние решения простейших задач по теории вероятностей, содержавшиеся в работах математиков XVI в. Тартальи и Кардано, а в ряде случаев и в более поздних работах — в XVII и XVIII вв., — содержат грубые ошибки.

Существуют две метафизические концепции.

Одна утверждает, что все в мире или случайно, или необходимо. С точки зрения этой концепции случайность и необходимость диаметрально противоположные категории. Ф. Энгельс писал: «Что можно подвести под всеобщие законы, то считается необходимым, а чего нельзя подвести, то считается случайным. Легко видеть, что это такого сорта наука, которая выдает за естественное то, что она может объяснить, и приписывает сверхъестественным причинам то, что для нее необъяснимо. При этом для существа самого дела совершенно безразлично, назову ли я причину необъяснимых явлений случаем или богом ... Наука прекращается там, где теряет силу необходимая связь»[1].

Вторая метафизическая концепция — детерминизм — вообще отрицает случайность, утверждая, что никаких случайных, беспричинных явлений нет. С точки зрения этой концепции представление о случайности явления происходит из-за того, что мы недостаточно знали причины этого явления.

Представитель этой концепции французский математик П. Лаплас считал, что вообще случая не существует. Случай, по его мнению, есть не более как выражение нашего неведения. Миром явлений управляет не вмешательство духа, стремящегося достигнуть известных ему целей, а закон причинности. Любое явление может быть вполне, во всех подробностях определено совокупностью предшествующих ему явлений. Иначе говоря, каждой причине строго соответствует определенное следствие.

Детерминизм П. Лапласа, не оставляющий места для случайности, состоит, по его словам, в следующем: «Разум, который для некоторого данного мгновения знал бы все действующие в природе силы и взаимное расположение всех составляющих ее тел, если бы притом он был достаточно мощным, дабы подвергнуть эти данные вычислению, охватил бы в одной формуле движения величайших светил небесных и движения мельчайших атомов: ничто не было бы для него недостоверным: будущее, как и прошедшее, было бы открыто его взору»[2]. В этой образной мечте об «универсальном разуме» абсолютно все подчинено необходимости.

Таким образом, представители этой метафизической концепции объявляют случайными и вследствие этого безразличными для науки те явления, относительно которых неизвестны причины, их обусловливающие. Единственно достойным научного интереса ими признается лишь необходимое.

Диалектический материализм под необходимостью понимает объективную закономерность, т. е. такое развитие явления, которое с неизбежностью вытекает из его сущности, из всего внутреннего хода событий.

Вместе с тем диалектический материализм не отрицает и случайность, существующую тоже объективно, но не вытекающую с необходимостью из закономерного развития данного явления. Случайное существует как внешнее относительно необходимого. Одно и то же явление может одновременно быть и случайным, и необходимым. В процессе развития случайность может переходить в необходимость. С диалектической точки зрения случайное — необходимо, а необходимое — случайно.

При изучении демографических явлений надо иметь в виду слова В. И. Ленина о том, что наука «... во всех областях знания показывает нам проявление основных законов в кажущемся хаосе

---

[1] *Маркс К., Энгельс Ф.* Соч., т. 20, с. 533.

[2] *Лаплас П. С.* Опыт философии и теории вероятностей/Под ред. Власова. М., 1908, с. 9.

явлений»[1]. То обстоятельство, что любой закон, в том числе демографический, не осуществляется с полной строгостью, что всегда имеются случайные отклонения от него, нас не должно удивлять, ибо там, «... где на поверхности происходит игра случая, там сама эта случайность всегда оказывается подчиненной внутренним, скрытым законам. Все дело лишь в том, чтобы открыть эти законы»[2].

Таким образом, среди кажущегося нагромождения индивидуальных фактов, за видимым беспорядком скрывается порядок, который иначе чем через случайность и проявиться не может.

Имея в виду, что необходимость и случайность нельзя противопоставить и что они не только не исключают, а, напротив, дополняют друг друга, Ф. Энгельс писал: «...то, что утверждается как необходимое, слагается из чистых случайностей, а то, что считается случайным, представляет собой форму, за которой скрывается необходимость»[3].

Значит, демографические явления и факты складываются под влиянием необходимого и случайного. При этом каждое из случайных событий имеет свою вероятность, а сами эти вероятности определяются влиянием основных факторов на случайные. Все живые существа смертны, это закон биологии. Но индивидуальная продолжительность жизни разная, так как различно влияние на нее отдельных случайных обстоятельств.

Первым в истории статистики, понявшим, что явления, происходящие в человеческом обществе, кажущиеся случайными и несвязанными друг с другом, на самом деле обладают удивительной правильностью, постоянством соотношений и подчиняются определенным, медленно меняющимся законам, был Д. Граунт, изучавший соотношения между численностью населения, с одной стороны, и числом браков, рождений и смертей — с другой. Последователи Д. Граунта — У. Петти, Э. Галлей, Стрюйк, Керсебум, Депарсье — пошли дальше и возвели эти закономерности в ранг неизменных законов природы и общества.

Все многообразие демографических фактов можно разделить на три группы: случайные события, случайные явления и случайные функции, а если аргументом является время, то случайные процессы.

1. *Случайные события* представляют собой результат какого-либо наблюдения, проводимого при соблюдении некоторого комплекса условий.

В качестве примера случайных событий в демографии можно привести факты рождений, смерти, вступления в брак, переездов из одной местности в другую и т. д. Количественной оценкой случайного события является его *вероятность* (p — от слова probabilitas — вероятность).

---

[1] *Ленин В. И.* Полн. собр. соч., т. 25, с. 46.
[2] *Маркс К., Энгельс Ф.* Соч., т. 21, с. 306.
[3] Там же, с. 303.

Наиболее широко в демографии используются *случайные величины*, являющиеся обобщением случайных событий и отражающие их количественную характеристику. Примером случайных величин может служить удельный вес мальчиков среди родившихся и др. Различают дискретные и непрерывные случайные величины.

В ряде случаев результат наблюдения определяется несколькими случайными величинами, образующими систему. Особенностью такой системы является взаимосвязь между отдельными величинами.

2. *Случайные явления* заранее предвидеть нельзя. Они могут произойти или не произойти. Но они не беспричинны. В их основе всегда лежит цепь вполне определенных причин и следствий.

3. *Случайные функции времени* делятся на процессы с непрерывным и дискретным изменением аргумента. Конкретные значения непрерывного случайного процесса по состоянию на определенный момент (t) называют *реализациями процесса*. При этом для каждого фиксированного значения t случайный процесс превращается в обычную случайную величину, называемую в этих случаях *сечением* случайного процесса. Таким образом, случайный процесс образует систему, состоящую из бесконечного числа случайных величин. Описание случайных процессов производится с помощью различных статистических характеристик: распределения вероятностей, математических ожиданий (средних значений), дисперсий, среднеквадратических отклонений, вычисленных по совокупности реализаций и др.

Советский статистик Н. К. Дружинин пишет: «... в нашей общественной жизни имеются такие области, в которых получает известный простор проявление случайных процессов. Это, в первую очередь, область демографических процессов»[1].

1.3. Единичные явления и массовые

Один из родоначальников отечественной математической статистики А. А. Чупров, занимавшийся и демографией, указывал, что развитие социально-экономических и естественных наук начиная с XVII в. идет по пути увлечения статистикой (статистификации), т. е. отказа от прослеживания единичных событий и сосредоточения интереса на совокупности этих событий, на изучении массовых явлений (термин, введенный Фехнером), представляющих обобщенный результат их взаимодействия. Этот сгусток единичных событий «... являет закономерность, постижимую для нас и без того, чтобы была необходимость знать в точности ход всех единичных процессов»[2].

Математическая статистика установила, что при малом числе наблюдений изучаемое демографическое явление может пока-

---

[1] *Дружинин Н. К.* Статистика и эконометрия. — Вестник статистики, 1978, № 12, с. 58.
[2] *Чупров А. А.* Вопросы статистики. М., 1960, с. 143.

2 Заказ 2715

заться случайным. Однако если такие явления подвергнуть массовому наблюдению, то в них очень часто обнаруживают определенные закономерности.

Массовые явления в статистическом смысле при изучении населения характеризуются тем, что в каждом индивидуальном случае сочетается необходимость и случайность.

При научной трактовке такого понятия, как случайное событие, следует исходить из единства количественного и качественного в анализе. В тех случаях, когда заранее описать поведение отдельных людей невозможно, следует использовать вероятностный подход к действиям большой массы людей.

К. Маркс объяснил, почему кажущиеся случайности общественной жизни обладают внутренней необходимостью, и показал, что внутренний закон при массовом наблюдении однородной совокупности формирует не только средний уровень явления (главное, основное, существенное, необходимое), но и отклонения от него (неглавное, второстепенное, несущественное, случайное), теснейшим образом связанные с природой изучаемых явлений. Значит, в случае сложного переплетения основных причин, влияющих на формирование закономерности всей массы явлений, со случайными причинами, имеющими более или менее серьезное значение для отдельных единиц, составляющих совокупность, большая роль принадлежит вероятностным методам изучения, позволяющим на основе действия закона больших чисел выявлять существенное в явлении и отделять его от несущественного.

То, что кажется *случайным* (нетипичным) при изучении одного факта, перестает быть случайным и становится *закономерным в их совокупности*. Так, при изучении случаев смерти людей каждый такой случай представляется нам более или менее случайным. Проследить причины, вызывающие случайные явления, принципиально невозможно. Умирают мужчины и женщины всех возрастов от разных болезней, из-за несчастных случаев и т. д.; указать конкретно, кто, когда, от какой болезни или причины умрет, невозможно. Только при наблюдении массы смертных случаев обнаружатся определенные закономерности, проявится порядок вымирания населения.

Аналогично дело обстоит со вступлением в брак. Хотя здесь большую роль играют индивидуальные условия (личные вкусы, склонности и антипатии и множество других побуждений), но только массовое наблюдение проясняет картину. В этом проявляется действие закона больших чисел — закона, действующего объективно, состоящего во взаимопогашении влияния случайных, второстепенных факторов, не определяющих существа явления. Такое взаимопогашение случайных факторов рельефно проявляет внутренние свойства данного явления, которые выступают затем в виде массовой закономерности.

Следует иметь в виду, что выявленные закономерности, присущие массе явлений, не зависят от индивидуальных особенностей отдельных случайных событий, составляющих эту массу. Индивидуальные особенности взаимно погашаются, нивелируются, и средний результат перестает быть случайным. Так, если статистические данные позволяют утверждать, что уровень смертности по стране снизился, а уровень рождаемости повысился, то в этом проявилась закономерность этих явлений, погасившая случайности.

Таким образом, на базе массовых случайных процессов можно обнаружить закономерности развития демографических явлений, действующих в виде тенденции на фоне единичных случайных и преходящих фактов. Демографические законы, как и экономические, «... не имеют иной реальности, кроме как в приближении, в тенденции, в среднем, но не в *непосредственной* действительности»[1]. Методы установления этих законов путем привлечения большого числа наблюдений случайных явлений в демографии играют большую роль.

В последние годы в демографии находят применение вероятностные методы экспертных оценок. Речь идет о таких случаях, когда достоверность исходной информации невелика и неизвестную количественную характеристику явления принимают за случайную величину. Тогда возникает возможность, используя индивидуальные оценки значимости того или иного события, данные отдельными экспертами (специалистами-демографами), определить такое значение величины изучаемого признака, которое является обобщающим мнением всей группы экспертов. Такое обобщающее мнение, находящееся внутри диапазона оценок, данных отдельными экспертами, является более достоверным, чем индивидуальное.

Что касается перехода от случайности единичного явления к объективной закономерности массы таких явлений, то этот переход раскрывается специфическими вероятностными методами математического анализа, путем привлечения статистических данных.

Великий русский хирург Н. И. Пирогов был, по его собственному выражению, «ревностным сторонником рациональной статистики». Рассматривая достоверность выводов из научных исследований, Н. И. Пирогов подчеркивал, что «задачей статистики будет вычислить цифру самой случайности, т. е. показать, что она не так случайна»[2].

1.4. Определение вероятностей и способы их измерения

Явления общественной жизни глубоко отличны от явлений природы, но между обществом, являющимся частью природы, т. е. частью материального мира, и природой нет непроходимой пропасти. В основе законов природы и общества лежат всеобщие законы диалектики.

Если в различных точных технических науках: физике, механике и т. д. — математическая статистика и вероятностный подход

[1] *Маркс К., Энгельс Ф.* Соч., т. 39, с. 355.
[2] *Пирогов Н. И.* Собр. соч. М., 1961, т. 5, ч. 1, с. 19.

являются главным и, пожалуй, единственным инструментом изучения законов, управляющих движением различных видов материи, то в отношении общественных и социально-экономических явлений, в частности демографических, вероятностный подход является, хотя и весьма плодотворным, но вспомогательным средством анализа. Он помогает построить модель данного явления, установить закон распределения элементов, составляющих данное явление, и его устойчивость, проверить различные гипотезы о характере статистической закономерности и, наконец, оценить различные параметры установленных распределений.

Математическая вероятность того, что событие произойдет, измеряется отношением числа шансов (равновозможных, т. е. единственно возможных и несовместимых), благоприятствующих данному событию A (M), к общему числу шансов (N):

$$P_A = \frac{M}{N}.$$

Это определение называют классическим. При таком понимании вероятности мы ограничиваем сферу практического применения только такими событиями, которые допускают различие дискретных шансов.

Известно еще геометрическое определение вероятности, привлекаемое для случаев, когда общее число равновозможных исходов бесконечно. Тогда по классическому определению вероятность каждого исхода становится равной нулю. Это затруднение преодолевается путем представления меры равновозможности в виде геометрической площади. Если оценивается вероятность того, что какая-то точка окажется внутри области Q, расположенной в квадрате, то искомая вероятность может быть измерена отношением соответствующих геометрических площадей:

$$P_Q = \frac{S_Q}{S},$$

где P_Q — вероятность того, что случайная точка попадает в область Q;
S_Q — площадь области Q;
S — площадь всего квадрата, равная a^2.

При исчислении вероятностей можно использовать искусственный прием, представив расчет в виде дроби, в которой числитель и знаменатель могут возрастать до бесконечности, в то время как их отношение будет стремиться к конечной величине. Здесь аналогия с математическим анализом состоит в том, что производная функция представляет отношение двух бесконечно малых, а само отношение стремится к конечному и определенному пределу.

Имеется еще статистическое понимание вероятности. Если в результате достаточно большого числа испытаний установлено, что частость случайного события $A \left(\frac{m}{n}\right)$ приближается к некоторой величине, то эту величину в силу действия закона больших чисел

принимают за численное значение вероятности данного события (P_A). Ясно, что установление вероятности случайного события опытным путем носит несколько неопределенный характер.

Так как вопрос о соотношении частостей и вероятностей для демографии очень важен, остановимся на нем подробнее.

Частость есть свойство генеральной совокупности, характеризующее ее структуру, она выражается отношением единиц, обладающих данным признаком, к общему числу единиц. *Вероятность* — это не оторванная от реальной действительности, заранее определенная теоретическая величина, а показатель, возникающий только при наличии стохастического процесса — процесса, подчиненного игре случая. Количественно частость, или относительная частость, изменяющаяся в пределах от нуля до единицы, и вероятность события могут совпасть.

Вероятность — число, вблизи которого колеблется частость случайного события при массовых наблюдениях, представляет собой объективную величину, выражающую статистическую устойчивость частостей, возрастающую при увеличении числа элементов, составляющих данное явление. Статистики и демографы исходят из того, что на основе использования теории вероятностей по частостям можно достаточно приближенно судить о величине вероятностей, и наоборот, зная вероятность события, можно предсказать, с какой частостью (при большом числе испытаний) будет происходить это событие.

Как же связана частость с вероятностью? Вероятность, являющаяся мерой возможности реализации случайного события, проявляется при большом числе наблюдений. При увеличении числа наблюдений частость стремится к вероятности. Так, распределение новорожденных по полу при переходе к рассмотрению большой территории с большим населением показывает устойчивость частости рождения мальчиков и девочек. Наблюдаемая устойчивость дает основание считать, что рождение мальчика имеет определенную вероятность, вокруг которой колеблется частость. В этом случае отклонение частости от вероятности указывает на погрешность, которую мы допускаем, принимая частость за вероятность. На возможности определения величины допускаемой погрешности при принятии величины частости как значения вероятности и основано применение теории вероятностей к различным областям человеческого знания, в том числе и к демографии.

Практическое значение теории вероятностей именно в этом и состоит. В ряде случаев частость называют статистической вероятностью. А. А. Чупров писал: «Частости воспроизводят лежащую в их основе вероятность с той степенью точности, какую предуказывает закон больших чисел». И далее: «Такое согласие теории и опыта — верный залог того, что построения теории вероятностей покоятся на надежном фундаменте»[1].

В демографии приближенная с необходимой степенью точности величина вероятности может быть определена расчетным путем,

[1] *Чупров А. А.* Очерки по теории статистики. М., 1959, с. 199.

на основе частости. Изучая брачность и разводимость населения Украинской ССР, Л. В. Чуйко разработала методику вычисления суммарных (для первых и повторных браков) частостей вступления в брак[1].

Вычисленные по этой методике показатели аналогичны вероятностным показателям таблиц смертности. Вместе с тем вопрос о взаимосвязи частостей и вероятностей конкретно для изучения проблемы брачности, по нашему мнению, является дискуссионным и требует дальнейшего изучения. Тождественности между частостью и вероятностью вступления в брак нет, так как вычисление частостей вступления в брак производится и для повторных событий.

Помимо указанных определений понятия «вероятность» (классического, основанного на принципе равновозможности, геометрического и эмпирически-частотного) у буржуазных статистиков встречается еще одно — субъективистское — понимание вероятности как показателя степени уверенности в справедливости того или иного суждения, опирающегося на интуицию человека, его чувства и т. д. Впервые такое субъективистское представление о вероятности как мере незнания выдвинул П. Лаплас. А. Кетле ввел агностическое представление о вероятности как об «атрибуте» типа, существующего в идее природы и лишь приближенно улавливаемого исследователем при наблюдении массы случаев.

Очень интересны взгляды русского статистика А. А. Чупрова, считавшего, что вероятность не имеет отношения к единичному случаю и является числовой характеристикой связи между различными следствиями и одной и той же причиной. При этом стохастические схемы, например схема зависимости смертности лиц разных профессий от общих, бытовых, гигиенических и других условий, по его мнению, «не пустая игра творческого воображения забывших о мире житейском математиков», а очень важные и нужные понятия, находящие себе аналоги в действительности. Вообще А. А. Чупров считал математическую вероятность объективной величиной, но не в том смысле, что она объективно существует, а в том, что она представляет собой объективный продукт теоретической мысли. Н. К. Дружинин, характеризуя взгляды А. А. Чупрова на вероятность, справедливо отметил, что для А. А. Чупрова вероятность являлась не отражением объективно существующих в окружающей нас действительности закономерностей, а скорее приемом или способом упорядочения «хаоса» явлений человеческим разумом.

Английские эмпирики вообще отождествляли математическую вероятность с эмпирической частостью события. Разновидностью эмпиризма явились взгляды Мизеса, считавшего вероятность априорной величиной, недоступной непосредственному наблюдению. Лишь при изучении единиц, составляющих совокупность, и увеличении числа привлекаемых единиц эмпирическая частость асимптотически стремится к вероятности.

[1] См.: *Чуйко Л. В.* Браки и разводы. М., 1975, с. 50.

Советский ученый Б. С. Ястремский показал, что вероятность есть математическое выражение свойств реальной совокупности; при изучении общественных явлений оно проявляется в стохастической схеме проведения случайной выборки. Вместе с тем Б. С. Ястремский понимал, что общественные явления обусловлены взаимодействием множества факторов различного направления, характера и силы, и поэтому нельзя полностью свести весь стохастический процесс к одной лишь выборке.

Если вероятность случайного события определяется до испытания, приведшего к определенному исходу, то численное значение этой вероятности называют априорным (доопытным). Вероятность события после опыта, приведшего к определенному результату, в отличие от доопытной вероятности называют апостериорной вероятностью. Вероятность того, что событие A не произойдет (q_A), исчисляется в соответствии с классическим определением вероятности как отношение числа шансов, неблагоприятствующих событию A, к общему числу шансов. Если N — общее число шансов и событию A благоприятствуют M шансов, то не благоприятствуют ему $N—M$ шансов.

Тогда

$$q_A = \frac{N-M}{N} = 1 - \frac{M}{N} = 1 - p_A .$$

Полученную вероятность называют вероятностью противоположного события.

Установим пределы, в которых заключены вероятности.

Достоверные события. Если все N случаев благоприятствуют событию A, т. е. $M=N$, то вероятность события A равна единице. Такие события называются достоверными.

Невозможные события. Если число случаев, благоприятствующих событию A, равно нулю ($M=0$), то вероятность события A равна нулю. Такие события называются невозможными.

В чем же суть вероятностной логики? В чем отличие вероятностной логики от логики детерминизма, утверждающей строгую зависимость всего будущего от всего предшествующего?

Вероятностная логика предусматривает не два значения истины (да или нет), как классическая логика, а множество их с различными вероятностями.

Вероятность, равная нулю, означает, что событие наверняка не произойдет, оно невозможно.

Вероятность, равная единице, означает, что событие произойдет непременно, оно достоверно.

Между этими двумя мерами, т. е. 0 и 1, существует бесконечно большое число переходов: от состояния полной уверенности в том, что событие произойдет, неполной уверенности в этом, сомнений и колебаний и т. д. до полной неуверенности.

Помимо обычных вероятностей в демографический анализ были введены *«независимые» вероятности* (Каруп). Иногда их называют *чистыми* вероятностями[1]. Рассмотрим особенности построения «независимых», или чистых, вероятностей и их отличие от обычных на примере вероятностей смерти.

Обычная вероятность смерти — это отношение числа случаев смерти за определенный интервал (период наблюдения, чаще всего один год) к начальной численности населения. В действительности изменение численности населения происходит не только из-за смертности, но и вследствие рождаемости и миграции. Численность реального населения уменьшается вследствие вымирания, но увеличивается вследствие рождаемости. При этом родившиеся в период наблюдения, увеличивая численность населения, увеличивают также и численность умерших. Значит, в формуле вероятности смерти числитель (число умерших в реальном населении) возрастет, а знаменатель останется без изменения. Вот почему вероятность смерти для населения, численность которого изменяется в результате рождений и смертей, больше, чем в населении, численность которого изменяется только вследствие вымирания.

Следовательно, в условиях одновременного изменения численности населения под влиянием нескольких демографических процессов обычные вероятности недостаточно хорошо отражают фактическую интенсивность каждого из них. Теоретически правильными будут вероятности, измеряющие интенсивность процессов в искусственной совокупности с элиминированным воздействием других факторов, кроме изучаемого. Вот эти вероятности и назвали «независимыми», или «чистыми».

Итак, отличие «чистой» вероятности смерти от обычной состоит в том, что при одном и том же знаменателе в числителе формулы из числа смертей, происшедших за определенный интервал времени, вычтены умершие из родившихся в этом интервале.

«Чистые» вероятности можно вычислять приближенно. Для перехода от обычных вероятностей к «чистым» на практике вместо уменьшения числителя дроби увеличивают знаменатель. Исходя из изменения числа рождений (N) пропорционально времени к знаменателю формулы обычной вероятности смерти следует прибавить половину числа родившихся $\left(\frac{N}{2}\right)$ за изучаемый интервал. Тогда получаем «чистую» вероятность смерти:

$$q_{\text{чистая}} = \frac{M}{S_0 + \frac{N}{2}}.$$

Если численность населения в начале года $S_0 = 1000$ человек и за год родилось 30, а умерло 10 человек, то значения обычной и «чистой» вероятностей примут вид:

$$q_{\text{обычная}} = \frac{M}{S_0} = \frac{10}{1\,000} = 0{,}01;$$

[1] См.: *Корчак-Чепурковский Ю. А.* Избранные демографические исследования. М., 1970, с. 73.

$$q_{\text{чистая}} = \frac{M}{S_0 + \frac{N}{2}} = \frac{10}{1\,000 + 15} = 0{,}00985.$$

Логика состоит в признании того, что события, вероятность которых очень мала, совершаются крайне редко.

В переписке Ж. Л. Бюффона с Д. Бернулли рассматривался вопрос о численном значении вероятности, которой можно пренебречь из-за малости. Бюффон утверждал, что вероятность порядка 0,0001 нужно рассматривать как вероятность невозможного события, и приводил в качестве примера уверенность здорового человека в возрасте 56 лет, что он проживет еще 24 часа, хотя на основании статистических данных вероятность этому человеку умереть в течение суток составляет 0,0001. О такой опасности никто не думает. В. В. Паевский писал о том, что вероятность меньше 0,0001 считается пренебрежительно малой. В качестве примера В. В. Паевский приводил следующее: «... никто при устройстве своих дел не считается с вероятностью умереть в течение ближайшей недели: вероятность этого считается настолько малой, что никому обычно не приходит в голову считаться с возможностью предстоящей в течение недели смерти. Между тем вероятность смерти в течение ближайшей недели для взрослого человека 35—40 лет от роду есть величина порядка 0,0001—0,0002...»[1].

В таблицах смертности предельным возрастом жизни человека принимается 100 лет. А может ли человек жить дольше? Может, конечно, примеров этому немало. Но где граница? Оказывается, что в соответствии с формулами, на которых основаны современные таблицы смертности, вероятность дожить до 1000 лет имеет величину порядка единица, деленная на 10^{27}. Это означает столь малую вероятность, которая совместима с нашим понятием о невозможности.

Касаясь совершения маловероятных событий, А. А. Чупров писал: «Стоит, однако, терпеливо подождать, и мы увидим как осуществляются самые маловероятные комбинации. Только ждать-то придется, может быть, мириады тысячелетий ...»[2].

Всегда ли вероятности могут быть изучены и измерены математико-статистическими методами? Разумеется, что не всегда. Так, В. И. Ленин задавал вопрос: «... суждена ли нам революция типа 1789 или типа 1848 года? ... Спрашивается, какой тип вероятнее?»[3]. В данном случае вероятность, о которой пишет В. И. Ленин, хотя и объективно существовала, но не могла быть измерена в результате повторных экспериментов, ибо речь шла об уникальном событии.

[1] *Паевский В. В.* Вопросы демографической и медицинской статистики/Под ред. А. М. Меркова. М., 1970, с. 419.
[2] *Чупров А. А.* Вопросы статистики, с. 146.
[3] *Ленин В. И.* Полн. собр. соч., т. 9, с. 380.

1.5. Связь между коэффициентами демографических событий и вероятностями

Для измерения интенсивности демографических процессов привлекают относительные величины, характеризующие степень распространения явлений в определенной среде — среде, породившей эти явления. Используют четыре вида «относительных величин: коэффициенты, вероятности, «независимые» вероятности и силы событий.

В формулах расчета всех четырех видов интенсивных показателей числитель один и тот же — это число демографических фактов, происшедших в изучаемой совокупности за определенный интервал (период) времени. Если в качестве знаменателя используется средняя численность совокупности, то показатель интенсивности называют *коэффициентом* демографического события (m_x), а если начальная численность совокупности — *вероятностью* события (q_x). При изменении численности совокупности только под влиянием данного события вероятность называется *«независимой»*, а при стремлении интервала времени к бесконечно малой величине — *силой* демографического события.

Некоторые демографы и социологи не различают коэффициенты смертности и вероятности смерти. Между тем вероятность умереть в определенном возрасте отличается от коэффициента смертности в этом же возрасте тем, что при равенстве числителей в том и другом показателе (число умерших в течение года в этом возрасте) знаменатели их отличаются: при исчислении вероятностей в знаменателе — число лиц этого возраста, находящихся в живых к началу года, а при исчислении коэффициентов смертности — средняя численность населения этого возраста.

Н. А. Басалаева, описывая демографическую подсистему модели Месаровича-Пестела, называет коэффициенты фертильности и смертности, при вычислении которых в качестве знаменателя фигурирует средняя численность населения, возрастными вероятностями[1]. Другие вероятности, в частности переходные вероятности миграции, автор определяет правильно: как отношение числа выбывших в данном возрасте к численности населения на начало года.

Известно, что коэффициент смертности вычисляется по формуле

$$m_x = \frac{D_x}{S_x}.$$

Отсюда можно перейти к вероятности смерти:

$$q_x = 1 - e^{-m_x}.$$

Из этой формулы вытекает другая, являющаяся достаточно приближенной:

$$q_x = \frac{2m_x}{2 + m_x}.$$

[1] См.: *Басалаева Н. А.* Моделирование демографических процессов и трудовых ресурсов. М., 1978, с. 24.

Очевидно, что вероятность дожития может быть определена следующим образом:

$$p_x = 1 - \frac{2m_x}{2 + m_x} = \frac{2 - m_x}{2 + m_x}.$$

Использование формул перехода от годичных коэффициентов к годичным вероятностям может быть распространено, правда, с большой потерей точности, также на пятилетний и десятилетний интервалы. Так, при пятилетних интервалах формулы вероятностей смерти и дожития принимают вид:

$$q_{x/x+5} = \frac{2 \cdot 5m_{x/x+5}}{2+5m_{x/x+5}};$$

$$p_{x/x+5} = \frac{2-5m_{x/x+5}}{2+5m_{x/x+5}}.$$

Следует иметь в виду, что эти формулы действуют лишь при достаточно малых значениях коэффициентов смертности ($m_x < 0{,}06$). При значениях $m_{x/x+5} > 0{,}4$ формулы приводят к абсурду (величина, например, вероятности смерти превышает единицу).

Причина неточности указанных выше формул состоит в том, что в различных возрастных группах связи между коэффициентами и вероятностями смерти носит различный характер. При этом отклонения от указанных формул при небольших коэффициентах смертности (в младших возрастах) не очень значительны, но в старческих возрастах становятся весьма заметными.

Использование коэффициента смертности до перехода к вероятности в демографии продиктовано тем, что, как пишет В. В. Паевский, «... коэффициент смертности позволяет производить над собой операции сложения, вычитания, умножения на число, измеряющее собой величину периода наблюдения, деления на то же число — все это без нарушения логического смысла каждой такой операции, в то время как подобного же рода операции с величинами «вероятности смерти» невозможны без утери логического содержания такого арифметического действия»[1].

В тех случаях, когда переход от коэффициентов смертности к вероятностям приходится делать часто, для сокращения вычислительной работы можно пользоваться специальными таблицами соотношений[2].

При изучении миграции, так же как и при изучении смертности, более употребительными являются не коэффициенты, а вероятности. Дело в том, что при резких отличиях численности населения на начало и конец периода коэффициенты миграции искажают интенсивность процесса, в них растворяются характерные особенности таких, например, контингентов населения, как окончившие

[1] *Паевский В. В.* Вопросы демографической и медицинской статистики, с. 143.

[2] См.: *Пресса Р.* Народонаселение и его изучение. М., 1966, с. 136.

учебные заведения, призываемые в ряды Советской Армии и демобилизованные и т. д.

Переход от коэффициентов интенсивности миграции, например выбытия (V_t), к вероятности выбытия (q_t) совершается по формуле

$$q_t = \frac{2aV_t}{2 + aV_t},$$

где a — величина интервала[1].

1.6. Теория вероятностей в страховании жизни

Практический интерес к теории вероятностей возник еще в XIV в. в Италии вследствие создания обществ страхований от огня, несчастных случаев, потери посылок, пенсионных касс и т. д.

Известно, что еще Г. Галилей, имея в виду потребности страхового дела, начал интересоваться средней продолжительностью жизни и составил для этой цели нечто вроде таблиц смертности.

Лишь в XVII в. с возникновением учреждений по страхованию жизни людей выросло значение теории вероятностей, а смертность сделалась предметом изучения математической статистикой. Вот уже несколько столетий страховое дело строится на точном математико-статистическом расчете прав и обязанностей сторон в течение договора страхования. Страховые взносы рассчитываются по формулам теории вероятностей и основываются на таблицах смертности. Страховые компании не только не разоряются, но и получают большие прибыли.

В 1662 г. Д. Граунт на основе смертных списков по Лондону, издававшихся с 1603 г., построил таблицу смертности, определил уровень вероятностей повозрастной смертности, влияние различных болезней на смертность, периода удвоения населения и другие показатели, необходимые для расчетов страхования жизни и выплат пожизненной ренты. При построении таблицы смертности для стационарного населения, в которой были две основные графы: l_x — числа доживающих до точного возраста и d_x — числа умирающих, Д. Граунт использовал средние вероятности смерти для лиц разных возрастов и с их помощью получал числа доживающих. Хотя Д. Граунт не вычислял среднюю продолжительность жизни, но он имел о ней представление и понимал, что продолжительность жизни связана с вероятностями смерти, по которым определяется число смертных случаев и соответствующие им численности населения.

Первой работой, в которой теория вероятностей нашла свое практическое применение в расчетах страховых обществ, была работа ученика Р. Декарта, голландского ученого И. де Витта (1671 г.) по исчислению пожизненной ренты.

[1] См.: *Курман М. В.* О демографических методах исследования миграции кадров предприятия. — В кн.: Статистика миграции населения. М., 1973, с. 114—115.

Вначале возраст страхующихся во внимание не принимался. Считалось, что смерть выбирает свои жертвы случайно, без учета возраста и положения. Застраховавшееся лицо (любого возраста) делало единовременный взнос определенной суммы. Платежи годовой ренты, получаемой застрахованным, назывались аннуитетами.

Для вычисления стоимости пожизненных аннуитетов из 4% годового роста денег де Витт применил теорию вероятностей. Задача, поставленная де Виттом, состояла в попытке обеспечить участникам вместо возрастающих рент постоянную. Но для этого он должен был знать порядок вымирания населения. Де Витт предположил, что вероятность жизни между 3 и 53 годами — величина постоянная, равная единице; вероятность жизни в последующих возрастных интервалах принимает следующие значения: от 53 до 63 лет — $^2/_3$, от 63 до 73 лет — $^1/_2$, от 73 до 80 лет — $^1/_3$. Лица старше 80 лет, ввиду их малочисленности, в расчет не принимались.

Несмотря на то что используемые им формулы были верны, привлекаемые для расчета величины не согласовывались с имеющимися статистическими данными, что приводило к неточным результатам. Историки объясняют этот факт тем, что при выкладках де Витт сознательно стремился занизить стоимость ренты.

С 1653 г. во Франции начало осуществляться личное страхование в виде пожизненной ренты и так называемой «тантины».

Тантина (по имени итальянского врача Тонти) — такая форма одновременного страхования группы людей, при которой отдельные лица вносят государству определенные денежные суммы. Пожизненные ренты, получаемые участниками этих займов, оплачивались из процентов на внесенные суммы: эти проценты делились между дожившими пропорционально их вкладам. Выплачиваемые ренты увеличивались по мере вымирания членов группы.

Начиная с 1689 г. страхующиеся делились на 10 возрастных групп по 7-летним интервалам. Величина единовременного взноса зависела уже от возраста: с молодых групп страхующихся взнос предусматривался бо́льший. Последнее обстоятельство поставило вопрос об изучении порядка вымирания населения и сыграло значительную роль в определении вероятностных показателей продолжительности человеческой жизни и повозрастных чисел доживающих, т. е. основных показателей таблиц смертности.

Первой таблицей смертности, основанной на опыте тантинных займов, была таблица Депарсье, посвятившего свои работы изучению повозрастной силы смертности. Его таблица, опубликованная в 1746 г., помимо трех граф, имеющихся в ранее разработанных таблицах: возраст, число умерших и число оставшихся в живых, — содержала еще одну графу — среднюю продолжительность предстоящей жизни.

В 1693 г. английский астроном и математик Э. Галлей построил первую повозрастную таблицу смертности по данным Г. Бреславля и исходя из 6% нормы годового роста денег рассчитал стоимость пожизненной ренты.

Вычисляя вероятную продолжительность предстоящей жизни, Э. Галлей преследовал чисто практические цели: его интересовал размер аннуитета по отдельным возрастам. Положив в основу «норму» смертности, Э. Галлей первым разработал правильный метод для вычисления пожизненных аннуитетов. Построенная им таблица смертности позволяла на основе распределения умерших по возрасту, без привлечения данных о живущих и родившихся, изучать возрастное распределение всего населения. Исходя из этих расчетов в 1699 г. в Лондоне была создана вдовья и сиротская касса. Метод построения таблиц смертности Э. Галлея получил название *метод смертных списков*.

Э. Галлей впервые сформулировал в демографии ряд положений и применил на практике отдельные методы:

осуществил построение таблицы смертности для закрытого населения, т. е. при отсутствии миграции;

выравнивая первичные данные о повозрастных числах умерших, устранил случайные колебания;

различал совокупности умерших и исчислял вероятности дожития и смерти современников (а не ровесников);

ввел показатель вероятной продолжительности жизни, т. е. величину, показывающую, через сколько лет остается ровно половина из начальной совокупности, не дав, правда, этому показателю никакого названия;

привлек теорию вероятностей к исчислению вероятности дожития и смерти для двух и трех лиц (определив истинную цену пожизненной ренты);

для иллюстрации соотношений при исчислении вероятностей дожития и смерти использовал геометрические построения; параллелограмм — при выводе вероятности для двух лиц, а стереограмму — для трех лиц.

Таблица смертности Э. Галлея — наиболее простая форма такого рода таблиц. В ней приводится число остающихся в живых при переходе от одного возраста к другому. Ее недостаток присущ всем таблицам, построенным по методу смертных списков: она завышала смертность во всех возрастах.

Рассмотрим исчисление Э. Галлеем вероятностей дожития и смерти для двух и трех лиц. В частности, для расчета вероятности дожить и умереть в течение следующих 8 лет двум лицам, возраста которых равны 17 и 34 годам, он привлек данные из своей таблицы смертности: численность современников в возрасте 17 лет составила 610, а численность современников в возрасте 34 года — 490. За 8 лет из 17-летних умерло 50 человек, из 34-летних — 73 человека.

Числа равновозможных благоприятствующих исходов для каждого из следующих четырех возможных случаев составили:

выжить обоим: $560 \cdot 417 = 233\,520$;
умереть обоим: $50 \cdot 73 = 3\,650$;

младшему выжить, а старшему умереть: $560 \cdot 73 = 40\,880$;
младшему умереть, а старшему выжить: $417 \cdot 50 = 20\,850$.

Сумма всех равновозможных исходов указанных четырех случаев составляет:

$$233\,520 + 3\,650 + 40\,880 + 20\,850 = 298\,900.$$

Тогда соответствующие вероятности по Э. Галлею оказались равными:

$$p_1 = \frac{233\,520}{298\,900} = 0{,}78126;$$

$$p_2 = \frac{3\,650}{298\,900} = 0{,}01221;$$

$$p_3 = \frac{40\,880}{298\,900} = 0{,}13677;$$

$$p_4 = \frac{20\,850}{298\,900} = 0{,}06976.$$

Французский математик А. Муавр (A. Moivre) при расчете стоимости пожизненной ренты также исходил из признания того, что повозрастные показатели смертности для разных групп населения различны. Попытки найти связь между возрастом и смертностью в конкретных условиях места и времени в виде математической формулы предпринимались давно. Используя таблицы Э. Галлея, А. Муавр вывел закон смертности, упрощающий все расчеты по страхованию жизни, предположив, что в возрасте от 12 до 86 лет уменьшение числа лиц при переходе от одной возрастной группы к другой происходит по арифметической прогрессии (из каждых 86 новорожденных умирает за год один человек). В соответствии с этим, изучая стоимость совместных рент, он исходя из гипотетической средней продолжительности жизни вывел различные формулы и составил таблицы[1].

Линия смертности по гипотезе Муавра являлась прямой линией, выражаемой уравнением

$$l_x = 83 - x,$$

а кривая вероятностей смерти — равносторонней гиперболой

$$p = \frac{1}{86 - x}.$$

Русский математик и демограф акад. В. Я. Буняковский использовал в страховых расчетах другой показатель таблиц смертности — среднюю продолжительность предстоящей жизни (e_x). Он предложил свою эмпирическую формулу для выравнивания показателей средней продолжительности жизни для возрастов в интервале 10—68 лет. «Закон смертности», как он был назван В. Я. Буняковским, состоял в предположении об изменении сред-

[1] См.: *Птуха М.* Очерки по истории статистики XVII — XVIII вв. М., 1945, с. 223.

ней продолжительности жизни в возрастах 10—68 лет по прямой линии $\overline{e}_x = a_0 + a_1 x$ (где a_0 и a_1 — параметры, определяемые методом наименьших квадратов).

В. Я. Буняковский проверял свою формулу по таблице доживаемости Депарсье и получил, по его мнению, результаты, вполне пригодные для практических целей.

Русский демограф С. А. Новосельский проверил формулу В. Я. Буняковского для разных таблиц смертности и пришел к выводу, что результаты, получаемые по формуле В. Я. Буняковского, не могут быть признаны достаточно удовлетворительными и что отклонения табличных значений e_x от $\overline{e}_x$ по формуле Буняковского распределяются так, что для крайних возрастов $e_x > \overline{e}_x$, а для средних возрастов $e_x < \overline{e}_x$. Все это дало основания С. А. Новосельскому предложить в качестве выравнивающей функции не прямую линию, а параболу второго порядка

$$\overline{e}_x = a_0 + a_1 x + a_2 x^2 .$$

Вычисления, произведенные С. А. Новосельским, дали значительно лучшие результаты, чем результаты В. Я. Буняковского.

Назовем некоторые широко известные таблицы смертности, которыми пользовались страховые компании в отдельных странах: в Англии использовалась таблица, составленная Мильоном в 1815 г. на основании данных города Карлиеля. Эта же таблица использовалась в России иностранными обществами страхования жизни. В России известны таблицы, изданные В. Я. Буняковским, В. Борткевичем, Л. Бессером и К. Баллодом. В 1862 г. таблицы смертности были опубликованы Институтом актуариев Великобритании и Ирландии. В 1903 г. этот же институт составил таблицы смертности по данным 60 обществ страхования жизни за период с 1863 по 1893 г.

Разберем страхование жизни вида пожизненной ренты и страхование жизни на один год[1].

Страхование жизни вида пожизненной ренты состоит в том, что отдельные лица заключают договор со страховой организацией об ежегодной уплате с момента страхования до самой смерти определенной суммы a (ежегодный взнос). При страховании жизни на один год страхующийся производит только одни взнос b. Со своей стороны страховая организация обязуется в случае смерти застрахованных лиц выплатить их наследникам в первом случае сумму A, а во втором случае сумму B.

Отвлекаясь от расходов страховой организации на проведение всех работ по страхованию и величину процента на внесенные страхующимися суммы, рассмотрим факторы, определяющие величину ежегодного взноса, а затем схематично методы его расчета. Очевидно, что основными факторами, определяющими величину ежегодного взноса, является пол и возраст страхующегося, а также

[1] См.: *Малешевский Б. Ф.* Теория и практика пенсионных касс. Математическая статистика. Спб., 1890, т. II, ч. 1.

уровень смертности лиц этого пола и возраста и связанную со смертностью населения продолжительность жизни.

Рассмотрим сначала страхование жизни вида пожизненной ренты. Пусть S лиц одного пола в возрасте x лет заключили договор о страховании. Тогда сумма взносов страхующимися лицами составит Se_xa (где e_x — средняя продолжительность жизни для лиц указанного возраста x). Выплаты страховой организации составят SA.

Так как $Se_xa = SA$, ежегодный взнос определяется по формуле

$$a = \frac{A}{e_x} .$$

Если договор о страховании заключается с S лицами возраста x не пожизненно, а на один год, то каждый уплатит сумму b. Страховая организация выплатит наследникам тех из этих лиц, которые умрут в течение года, сумму B. Обозначив q_x — вероятность смерти в течение одного года, найдем, что страховая организация выплачивает BSq_x, а получает Sb. Значит, $Sb = BSq_x$. Откуда $b = Bq_x$. Вероятность q_x устанавливается путем наблюдения определенного контингента лиц этого возраста.

Разумеется, фактическая частость умерших (m_x) может отклоняться от вероятности q_x в обе стороны. Если $m_x < q_x$, то величина выплат страховой организации будет меньше полученных ею сумм, а если $m_x > q_x$, то страховая организация может оказаться не в состоянии выплатить положенные суммы. Мерой возможного отклонения является средняя квадратическая ошибка:

$$\mu = \sqrt{\frac{p_x q_x}{S}} \quad \text{при } p_x = 1 - q_x .$$

Отклонение m_x от q_x не может превзойти утроенную среднюю квадратическую ошибку (с вероятностью 0,997). Чтобы и в этом случае собранных взносов хватило для выплат, необходимо, чтобы

$$Sb = BS\left(q + 3\sqrt{\frac{pq}{S}}\right) = B\left(Sq + 3\sqrt{Spq}\right).$$

Величина $3\sqrt{Spq}$ является дополнительным резервом к страховому фонду. Она растет не пропорционально числу застрахованных, а пропорционально квадратному корню из него.

Для уменьшения риска, связанного с ограниченностью числа застрахованных, страховые компании перестраховывают свои контингенты друг у друга. Страховые организации, используя таблицы смертности, могут определить наиболее вероятное количество умирающих и доживающих лиц из числа застрахованных и, следовательно, вероятный размер ежегодных выплат.

Э. Качаловская приводит пример смешанного страхования жизни 1 000 000 лиц в возрасте 40 лет на 10 лет на страховую сумму 100 руб. каждому и демонстрирует расчет того, какова должна быть величина ежегодных страховых взносов[1]. Для этой цели в первую очередь автор привлекает таблицу смертности, разработанную Госстрахом в 1956 г.

Из таблиц смертности видно, что до 50 лет из числа 40-летних доживет 92 682 лица, каждому из которых нужно будет выплатить 100 руб. Следовательно, средства для выплат составляют: 92 682·100 руб.=9 268 200 руб. Если бы страхующиеся оплачивали положенные суммы в момент заключения договора, то Госстрах в течение 10 лет получал бы на эти средства 3% годового дохода. Используя соответствующие таблицы, можно найти величину V^{10}, называемую дисконтирующим множителем и позволяющую определять стоимость выплат Госстраха через 10 лет (фонд, необходимый в момент заключения договора). При трехпроцентной норме доходности $V^{10}=0{,}74409$. Умножая 9 268 200 на 0,74409, получаем 6 896 375 руб. — сумму, которой должен располагать Госстрах при заключении договора страхования. Следовательно, каждый из 1 000 000 страхующихся при заключении договора должен внести 68 р. 96 к.

Однако большинство страхующихся делают взносы не сразу, а постепенно, по годам. Используя специальные коэффициенты, учитывающие как норму роста денег, так и постепенное естественное уменьшение числа застрахованных вследствие смертности, получаем ежегодный взнос 8 р. 07 к. на 100 руб. страховой суммы.

Расчет фонда выплат по случаю смерти основывается на таблице смертности, позволяющей найти число лиц, не доживающих до 50 лет. Исходя из этого числа и 3%-ной нормы доходности можно определить, что ежегодная ставка при рассрочке на 10 лет равна 0 р. 72 к.

Что касается вероятности утраты нетрудоспособности в результате несчастного случая, почти не зависящей от возраста, то в действующих тарифах прибавляется 0 р. 15 к. на каждые 100 руб. страховой суммы. Нагрузка к нетто-ставке, предназначенная для покрытия расходов на ведение операции, составляет 1 р. 15 к. Учитывая все слагаемые, получается полная брутто-ставка: 8 р. 07 к.+0 р. 72 к.+0 р. 15 к.+1 р. 15 к.= 10 р. 09 к.

Значительный интерес при совместном страховании жизни супружеских пар имеет сочетание их возрастов. Изучению соответствия возрастов мужа и жены помогает построение пространства простых событий с помощью первого квадранта плоскости x, y. Обозначая возраст мужа x и возраст жены y, произведем построение пространства элементарных событий (рис. 1.1): событие A — мужу больше 48 лет (все точки лежат справа от прямой $x=48$); событие B — муж старше жены (все точки лежат между осью x и прямой $y=x$); событие C — жене больше 48 лет (все точки лежат над прямой $y=48$).

[1] См.: *Качаловская Э.* Повышение средней продолжительности жизни населения и его влияния на размеры страховых тарифов. — В кн.: Вопросы государственного страхования. М., 1964, с. 111.

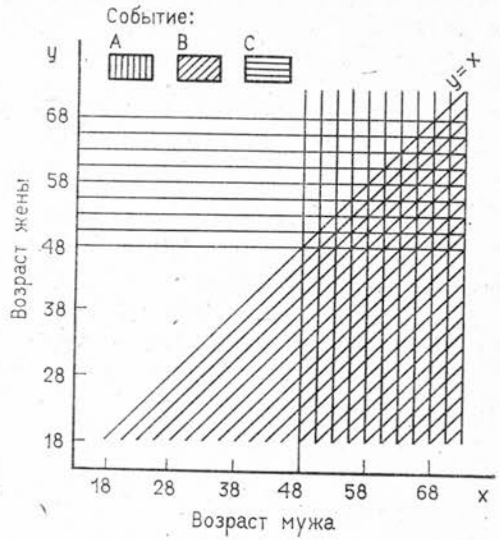

Рис. 1.1. Пространство элементарных событий

Пересечение плоскостей элементарных событий дает сложные события. Рассмотрим некоторые из них: событие AB означает, что мужу больше 48 лет и он старше жены (область, заключенная между осью x и прямыми $x=48$ и $y=x$); событие $\widehat{AB}$ означает, что мужу больше 48 лет и он не старше жены (область между прямыми $x=48$ и $y=x$). Возможны и другие сложные события.

Формулу Гомперца — Макегама используют в страховом деле для исчисления тарифных ставок при страховании двух лиц в возрастах x и y и определения одинакового промежуточного возраста этих лиц — v. Она имеет вид:

$$l_x = KS^x Z^{C^x},$$

где l_x — число лиц, доживающих до возраста x; K, S, Z, C — параметры, определяемые из статистических данных.

Зафиксируем вероятность того, что лицо в возрасте x лет проживет n лет ($_np_x$):

$$_np_x = \frac{l_{x+n}}{l_x}.$$

Заменяя значения l_{x+n} и l_x соответствующими выражениями, из формулы Гомперца — Макегама получим

$$_np_x = \frac{l_{x+n}}{l_x} = \frac{KS^{x+n} Z^{C^{x+n}}}{KS^x Z^{C^x}} = S^n Z^{C^x(C^n-1)}.$$

Аналогично находим вероятность того, что лицо в возрасте y проживет n лет:

$$_np_y = S^n Z^{C^y(C^n-1)}.$$

Вероятность того, что оба лица проживут n лет ($_np_{xy}$), получим по теореме умножения вероятностей:

$$_np_{xy} = {_np_x}\,{_np_y} = S^n Z^{C^x(C^n-1)} S^n Z^{C^y(C^n-1)} = S^{2n} Z^{C^x(C^n-1)+C^y(C^n-1)} = S^{2n} Z^{(C^n-1)(C^x+C^y)}.$$

Логарифмируя, получаем:

$$\lg {_np_{xy}} = 2n \lg S + (C^n-1)(C^x+C^y)\lg Z.$$

Пусть у обоих лиц одинаковый возраст v, заменяющий возраст x и y. Тогда

$$\lg {}_np_{vv}=2n\lg S+2(C^n-1)C^v\lg Z=$$
$$=2[n\lg S+C^v(C^n-1)\lg Z].$$

Учитывая, что выражение, заключенное в квадратные скобки, есть логарифм вероятности того, что лицо в возрасте v проживет n лет, т. е. $\lg {}_np_v$, получаем:

$$\lg {}_np_{xy}=2\lg {}_np_v\ .$$

Откуда ${}_np_{xy}={}_np^2{}_v={}_np_{vv}$.

Полученный результат означает, что вероятность того, что два лица в возрасте x и y проживут n лет, равна вероятности того, что этот срок проживут два лица, возраст каждого из которых v лет. Величину v можно рассчитать. Выше мы заменили C^x+C^y равным ему $2\,C^v$. Следовательно,

$$C^v=\frac{C^x+C^y}{2}\ .$$

Откуда

$$v=\frac{\lg(C^x+C^y)-\lg 2}{\lg C}\ .$$

В качестве примера укажем, что при страховании двух лиц в возрасте 40 и 30 лет промежуточный возраст оказывается равным 36 годам. С помощью формулы Гомперца — Макегама можно найти тарифы по страхованию двух лиц, заменив их тарифом страхования одного лица таким образом, чтобы вероятность наступления страхового случая двух лиц в возрастах x и y являлась равной вероятности наступления страхового случая одного лица в возрасте u лет.

Тогда

$$C^x+C^y=C^u;\quad {}_np_u={}_np_{xy}=S^nZ^{\,C^u(C^n-1)};$$

$$u=\frac{\lg(C^x+C^y)}{\lg C}\ .$$

При $x=40$ и $y=30$ общий возраст равен 43 годам, т. е. несколько больше возраста старшего из двух лиц.

Практика страхования жизни вызвала необходимость использования вероятностей смерти одного из двух лиц, пережившего другого, вероятности дожития до конца определенного срока по крайней мере одного из двух лиц, вероятности остаться в живых через n лет или умереть в течение этого же периода одному определенному (мужу или жене) лицу, вероятности лицу в возрасте x лет умереть раньше лица в возрасте y лет и т. д.

Рассмотрим некоторые вероятности. Кроме вероятностей умереть в течение ближайшего года жизни введем в анализ вероятности, относящиеся к более продолжительным периодам времени, которые могут заинтересовать исследователей в области демографии

и страхования. Учтем следующие соотношения показателей таблиц смертности:

$$d_x=l_{x+1}-l_x\ ;$$
$$p_x=\frac{l_{x+1}}{l_x};\quad q_x=\frac{d_x}{l_x};\quad p_x+q_x=1;$$
$$l_{\omega+1}=0;\quad \Sigma d_x=l_0;\quad p_\omega=0;\quad q_\omega=1,$$

где d_x — числа, характеризующие порядок вымирания по годам возраста, т. е. доля людей, умерших в возрасте x лет на $x+1$ году жизни из совокупности родившихся; ω — предельный возраст; l_x; l_{x+1} — число доживающих до возраста x; $x+1$ лет и т. д.; l_0 — начальное число родившихся; p_x — вероятность дожития, т. е. доля доживающих до возраста $x+1$ из доживших до возраста x лет; q_x — вероятность умереть в возрасте x лет, т. е. доля умерших в составе совокупности людей, достигших возраста x лет.

Вероятности, относящиеся к одному лицу. Определим вероятность для лица, дожившего до возраста x, *дожить до возраста $x+n$ лет.* Чтобы лицо, дожившее до возраста x лет, дожило до возраста $x+n$ лет, нужно, чтобы оно сначала дожило до $x+1$ года, затем до $x+2$ и т. д. А это есть сложное событие, состоящее из простых событий. Искомая вероятность определяется по теореме умножения вероятностей, так как доживший до возраста x лет не может дожить до $x+2$ года, если он не дожил до возраста $x+1$ год, и т. д. Поэтому

$${}_2p_x=p_x\,p_{x+1};$$
$${}_3p_x={}_2p_x\,p_{x+2}=p_x\,p_{x+1}\,p_{x+2};$$
$$\cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot$$
$${}_np_x=p_xp_{x+1}p_{x+2}\ldots p_{x+n-1}\ .$$

Заменяя значение вероятностей дожития отношением числа доживающих, получаем:

$$p_x=\frac{l_{x+1}}{l_x};\quad p_{x+1}=\frac{l_{x+2}}{l_{x+1}};\ldots;\ p_{x+n-1}=\frac{l_{x+n}}{l_{x+n-1}}\ .$$

Тогда

$${}_np_x=\frac{l_{x+1}}{l_x}\cdot\frac{l_{x+2}}{l_{x+1}}\cdots\frac{l_{x+n}}{l_{x+n-1}}=\frac{l_{x+n}}{l_x}\ . \qquad (1)$$

Найдем вероятность для дожившего до возраста x лет *умереть ранее достижения им возраста $x+n$ лет* — ${}_nq_x$. Совершенно ясно, что ${}_nq_x=1-{}_np_x$.

Отсюда

$${}_nq_x=1-\frac{l_{x+n}}{l_x}=\frac{l_x-l_{x+n}}{l_x}\ . \qquad (2)$$

Это можно записать и так:

$${}_nq_x=\frac{d_x+d_{x+1}+\ldots+d_{x+n-1}}{l_x}\ . \qquad (2a)$$

Зная вероятность дожить от возраста x до возраста $x+n$ лет, можно определить вероятность умереть в интервале возрастов от $x+n$ лет до $x+n+m$ лет. Символ ${}_{20/10}q_{50}$ означает вероятность человеку в возрасте 50 лет умереть в возрасте между 70 и 80 годами.

По формуле умножения вероятностей получим

$$ {}_{n/m}q_x = {}_np_{x\,m}q_{x+n} = \frac{l_{x+n} - l_{x+n+m}}{l_x} . \tag{3} $$

Окончательно получим

$$ {}_{n/m}q_x = {}_np_x - {}_{n+m}p_x . \tag{3a} $$

Вероятности, относящиеся к двум лицам[1]. Нас будут интересовать некоторые вероятности, относящиеся к совместной жизни двух лиц в возрасте x лет и y лет, а также «отсроченные вероятности».

Вероятность дожить обоим до конца периода n лет (${}_np_{xy}$). Если исходить из предположения о независимости дожития или смерти двух указанных лиц, то можно найти вероятность лицам в возрасте x и y лет прожить вместе еще n лет. При $n=1$ можно писать просто p_{xy}.

Рассматриваемое событие является сложным и определяется одновременно наступлением двух независимых событий:

1. Одно лицо, дожившее до возраста x лет, проживет еще n лет, т. е. доживет до возраста $x+n$ лет.

2. Другое лицо, дожившее до возраста y лет, проживет еще также n лет, т. е. доживет до возраста $y+n$ лет. Тогда

$$ {}_np_{xy} = {}_np_{x\,n}p_y = \frac{l_{x+n}\, l_{y+n}}{l_x\, l_y} . \tag{4} $$

Вероятность по крайней мере одному из двух лиц умереть в течение n лет с момента достижения возраста x и y лет (${}_nq_{xy}$) нетрудно получить из формулы (4):

$$ {}_nq_{xy} = 1 - {}_np_{xy} . $$

Откуда

$$ {}_nq_{xy} = \frac{l_x l_y - l_{x+n}\, l_{y+n}}{l_x\, l_y} . \tag{5} $$

Вероятность умереть обоим лицам в течение n лет является вероятностью совместного наступления двух событий и рассчитывается в соответствии с теоремой умножения вероятностей:

$$ {}_nq_{x\,n}q_y = (1 - {}_np_x)(1 - {}_np_y) = 1 + {}_np_{x\,n}p_y - {}_np_x - {}_np_y . \tag{6} $$

Вероятность умереть последним одному из двух лиц, достигших возраста x и y лет, в течение n лет представляет собой разность двух вероятностей ${}_nq_{xy} - {}_nq_{x\,n}q_y$ и составляет ${}_np_{x\,n}q_y + {}_np_{y\,n}q_x$.

Отсюда

$$ {}_nq_{xy} = {}_np_{x\,n}q_y + {}_np_{y\,n}q_x + {}_nq_{x\,n}q_y . \tag{7} $$

[1] Аналогичные построения вероятностей можно произвести и для трех лиц.

Полученное выражение можно трактовать следующим образом: вероятность умереть по крайней мере одному из двух лиц, достигших возраста x и y лет, в течение n лет равна сумме трех вероятностей:

вероятности лицу в возрасте x лет прожить n лет, а лицу в возрасте y лет умереть в течение этого же срока;

вероятности лицу в возрасте y лет прожить n лет, а лицу в возрасте x умереть в течение этого же срока;

вероятности обоим умереть в течение n лет.

Для обозначения того, что вероятность относится не к совокупности лиц, достигших возраста x и y лет, а к последнему пережившему из этих двоих лицу, условимся над xy проводить горизонтальную черту, т. е. писать $\overline{xy}$. Тогда вероятность умереть последним одному из двух лиц, достигших возраста x и y лет (т. е. умереть обоим), в течение n лет будет равна:

$$ {}_nq_{\overline{xy}} = {}_nq_{x\,n}q_y . \tag{8} $$

Теперь нетрудно найти вероятность противоположного события, т. е. *вероятность того, что по крайней мере хотя бы одно из двух лиц, доживших до возраста x и y лет, доживет до конца периода n лет:*

$$ {}_np_{\overline{xy}} = 1 - {}_nq_{\overline{xy}} = 1 - {}_nq_{x\,n}q_y . $$

Тогда

$$ {}_np_{\overline{xy}} = 1 - (1 - {}_np_x)(1 - {}_np_y) = {}_np_x + {}_np_y - {}_np_{xy} . \tag{9} $$

Найдем теперь *вероятность только одному лицу, например в возрасте x лет, умереть в течение периода n лет* (${}_nq_{x/y}$). Здесь налицо два независимых события: доживший до x лет умрет в течение n лет (${}_nq_x$), а доживший до y лет доживет до $y+n$ лет (${}_np_y$): ${}_nq_{x/y} = {}_nq_{x\,n}p_y$.

После преобразований получим

$$ {}_nq_{x/y} = {}_np_y - {}_np_{xy} . \tag{10} $$

Аналогично имеем вероятность для лица в возрасте y лет только одному умереть до конца периода n:

$$ {}_nq_{y/x} = {}_np_x - {}_np_{xy} . \tag{11} $$

Для получения вероятности любому из двух лиц прожить n лет следует использовать теорему сложения вероятностей и найти сумму вероятностей только одному из двух лиц прожить n лет. В результате получаем *вероятность прожить по крайней мере одному из двух лиц n лет:*

$$ {}_np_y - {}_np_{xy} + {}_np_x - {}_np_{xy} = {}_np_y + {}_np_x - 2\,{}_np_{xy} . \tag{12} $$

Можно также определять *отсроченные вероятности.* Так, вероятность того, что совместная жизнь двух лиц в возрасте x и y лет будет благополучно продолжаться еще n лет и нарушится в течение следующих m лет, можно вычислить по формуле

$$ {}_{n/m}q_{xy} = {}_np_{xy} - {}_{n+m}p_{xy} . \tag{13} $$

Вероятность пережившему из двух лиц период $x+n$ лет умереть после этого в течение следующих m лет равна:

$$_{n/m}q_{\overline{xy}} = {}_{n/m}q_x + {}_{n/m}q_y - {}_{n/m}q_{xy}\,. \qquad (14)$$

Для исследования вопроса о порядке, в котором следует смерть двух лиц, можно ввести понятие о средней продолжительности жизни двух лиц. Если учесть, что средняя продолжительность жизни одного лица, достигшего возраста x лет (e_x), равна $\Sigma_n p_x + \frac{1}{2}$, то по аналогии средняя продолжительность жизни двух лиц, достигших возраста x и y лет, составит: $e_{xy} = \Sigma_n p_{xy} + \frac{1}{2}$.

Тогда вероятность для лица, достигшего возраста x лет, умереть раньше лица, достигшего возраста y лет (Q_{xy}), будет равна сумме вероятностей этого события на 1, 2, 3, ..., $(n+1)$-м году и т. д. После различных преобразований получаем:

$$Q_{xy} = \frac{1}{2}\left(1 - \frac{e_{x/\,y-1}}{p_{y-1}} + \frac{e_{x-1/y}}{p_{x-1}}\right); \qquad (15)$$

$$Q_{yx} = \frac{1}{2}\left(1 - \frac{e_{y/x-1}}{p_{x-1}} + \frac{e_{y-1/x}}{p_{y-1}}\right). \qquad (16)$$

Учитывая, что кто-то из двух лиц непременно умрет раньше, $Q_{xy} + Q_{yx} = 1$.

Рассмотрим некоторые примеры и произведем соответствующие расчеты, используя данные, относящиеся к смертности городского и сельского населения, взятые из таблицы смертности и средней продолжительности жизни населения СССР 1958—1959 гг.[1]

По формуле (1) найдем вероятность лица, дожившего до 20 лет, дожить до 21, 22 и т. д. до 80 лет:

$$_1p_{20} = p_{20} = \frac{l_{21}}{l_{20}} = \frac{92\,767}{92\,917} = 0{,}99839;$$

$$_2p_{20} = \frac{l_{22}}{l_{20}} = \frac{92\,610}{92\,917} = 0{,}99670;$$

. . . .

$$_{10}p_{20} = \frac{l_{30}}{l_{20}} = \frac{91\,090}{92\,917} = 0{,}98034;$$

$$_{20}p_{20} = \frac{l_{40}}{l_{20}} = \frac{88\,565}{92\,917} = 0{,}95316;$$

$$_{40}p_{20} = \frac{l_{60}}{l_{20}} = \frac{76\,693}{92\,917} = 0{,}82539;$$

$$_{60}p_{20} = \frac{l_{80}}{l_{20}} = \frac{36\,481}{92\,917} = 0{,}39262.$$

[1] См.: Итоги Всесоюзной переписи населения 1959 года. СССР (Сводный том). М., 1962, с. 262—279.

По формуле (2) найдем вероятность лица, дожившего до 20 лет, умереть до достижения им 30; 40; 60 и 80 лет:

$$_{10}q_{20} = 1 - 0{,}98034 = 0{,}01966;$$

$$_{20}q_{20} = 1 - 0{,}95316 = 0{,}04684;$$

$$_{40}q_{20} = 1 - 0{,}82539 = 0{,}17461;$$

$$_{60}q_{20} = 1 - 0{,}39262 = 0{,}60738.$$

По формуле (3) найдем вероятность для человека в возрасте 50 лет дожить до 70 лет и умереть до истечения следующих 10 лет, т. е. в интервале от 70 до 80 лет:

$$_{20/10}q_{50} = \frac{l_{50+20} - l_{50+20+10}}{l_{50}} = \frac{l_{70} - l_{80}}{l_{50}} =$$

$$= \frac{61\,762 - 36\,481}{84\,502} = \frac{25\,281}{84\,502} = 0{,}29918.$$

По формуле (4) найдем вероятность двум лицам в возрасте 30 и 20 лет прожить вместе еще 10 лет:

$$_{10}p_{30\,:\,20} = \frac{l_{40}l_{30}}{l_{30}l_{20}} = \frac{88\,565}{92\,917} = 0{,}95316.$$

Можно и так:

$$_{10}p_{30\,:\,20} = {}_{10}p_{30}\,{}_{10}p_{20} = \frac{l_{40}}{l_{30}}\,{}_{10}p_{20} = \frac{88\,565}{91\,090}\cdot 0{,}98034 =$$

$$= 0{,}97228\cdot 0{,}98034 = 0{,}95316.$$

Отсюда находим вероятность того, что по крайней мере одно из двух указанных лиц умрет в течение 10 лет:

$$_{10}q_{30\,:\,20} = 1 - 0{,}95316 = 0{,}04684.$$

Вероятность смерти обоих лиц в течение 10 лет *по формуле (6)* равна:

$$_{10}q_{30}\cdot{}_{10}q_{20} = 1 + {}_{10}p_{30}\cdot{}_{10}p_{20} - {}_{10}p_{30} - {}_{10}p_{20} =$$

$$= 1 + \frac{88\,565}{91\,090}\cdot\frac{91\,090}{92\,917} - \frac{88\,565}{91\,090} - \frac{91\,090}{92\,917} = 1 + 0{,}95316 -$$

$$- 0{,}97228 - 0{,}98034 = 0{,}00054.$$

По формуле (7) можно получить ту же вероятность умереть по крайней мере одному из двух лиц, достигших 30 и 20 лет, в течение ближайших 10 лет:

$$_{10}q_{30\,:\,20} = {}_{10}p_{30}\,{}_{10}q_{20} + {}_{10}p_{20}\,{}_{10}q_{30} + {}_{10}q_{30}\,{}_{10}q_{20} =$$

$$= 0{,}97228\cdot 0{,}01966 + 0{,}98034\cdot 0{,}02772 + 0{,}00054 =$$

$$= 0{,}01912 + 0{,}02718 + 0{,}00054 = 0{,}04684.$$

По формуле (8) находим вероятность умереть последним одному из двух лиц, достигших возраста 30 и 20 лет, в течение ближайших 10 лет, которая равна вероятности умереть обоим:

$$_{10}q_{\overline{30:20}} = {}_{10}q_{30}\cdot{}_{10}q_{20} = 0{,}00054.$$

Рассчитаем вероятность того, что хотя бы одно из двух указанных выше лиц проживет еще 10 лет:

$$_{10}p_{\overline{30:20}} = 1 - 0{,}00054 = 0{,}99946,$$

или

$$_{10}p_{\overline{30:20}} = {_{10}p_{30}} + {_{10}p_{20}} - {_{10}p_{30:20}} =$$
$$= 0{,}97228 + 0{,}98034 - 0{,}95316 = 0{,}99946.$$

Вероятность только одному лицу (из двух доживших до 30 и 20 лет), а именно дожившему до 30 лет, умереть в течение ближайших 10 лет можно определить *по формуле (10):*

$$_{0}q_{30/20} = {_{10}p_{20}} - {_{10}p_{30:20}} = 0{,}98034 - 0{,}95316 = 0{,}02718.$$

Для дожившего до 20 лет эта же вероятность рассчитывается *по формуле (11):*

$$_{10}q_{20/30} = {_{10}p_{30}} - {_{10}p_{30:20}} = 0{,}97228 - 0{,}95316 = 0{,}01912.$$

Вероятность по крайней мере одному из двух указанных лиц прожить еще 10 лет определяется *по формуле (12):*

$$_{10}p_{20} + {_{10}p_{30}} - 2\,{_{10}p_{30:20}} = 0{,}98034 + 0{,}97228 - 2\cdot 0{,}95316 = 0{,}04630.$$

Для использования *формулы (13)* примем, что $m=n=10$. Тогда предварительно по формуле (4) определим:

$$_{20}p_{30:20} = \frac{l_{30+20}\ l_{20+20}}{l_{30}l_{20}} = \frac{l_{50}l_{40}}{l_{30}l_{20}} =$$
$$= \frac{0{,}84502\cdot 0{,}88565}{0{,}91090\cdot 0{,}92917} = \frac{0{,}74839}{0{,}84638} = 0{,}88422,$$

а затем получаем «отсроченную вероятность» нарушения совместной жизни двух лиц 30 и 20 лет после достижения ими возраста 40 и 30 лет:

$$_{10/10}q_{30:20} = {_{10}p_{30:20}} - {_{20}p_{30:20}} = 0{,}95316 - 0{,}88422 = 0{,}06894.$$

Приведем факт использования вероятностных расчетов из истории медицинской статистики России.

В 90-х годах прошлого столетия во врачебной среде обсуждался вопрос о создании такого страхового общества, которое позволяло бы при единовременном вступительном взносе 10 руб. с ежегодной уплатой 5 руб. застраховаться на случай смерти на сумму 2900 руб. Высказывались различные мнения.

Врач-статистик В. И. Гребенщиков, используя построенную им полную таблицу смертности русских врачей и произведя различные вероятностные расчеты, показал нереальность осуществления этого проекта: при приеме в страховое общество врачей всех возрастов оно станет несостоятельным всего лишь через 2—3 года. Для безубыточного функционирования такого общества ежегодные взносы страхующихся должны быть дифференцированными в зависимости от возраста. В. И. Гребенщиков рассчитал размеры этих взносов.

Б. С. Ястремский использовал формулу вероятности умереть (q_x) лиц, состоящих на страховании, вследствие выбытия из жизни или по случаю смерти, выведенную Спрейгом, $q = \frac{m}{a-Qb}$, где m — число смертных случаев в течение года; a — число застрахованных в начале года; b — число, выбывших со страхования при жизни в течение года; Q — некоторая правильная дробь.

Задача, которую поставил Б. С. Ястремский, состояла в определении Q. Величина q по методу Спрейга вычислялась на основе произвольного предположения, что лица, выбывшие со страхования, и лица, оставшиеся на страховании до конца года, подвержены одному и тому же закону вымирания. Это не совсем так, по крайней мере для первых лет страхования. Однако ввиду невозможности измерить расхождение между вероятностями умереть для указанных двух категорий лиц Б. С. Ястремский вынужден был принять их приблизительно равными. Но это обстоятельство помогло понять, что очень точного определения величины Q не требуется. Для выяснения влияния ошибки в определении Q на величину q Б. С. Ястремский предположил наличие двух значений (q_1 и q_2), при вычислении которых Q_1 взято точно, а Q_2 с грубым приближением:

$$q_1 = \frac{m}{a-Q_1b}; \qquad q_2 = \frac{m}{a-Q_2b},$$

тогда отношение

$$\frac{q_1}{q_2} = \frac{1-Q_2\frac{b}{a}}{1-Q_1\frac{b}{a}}.$$

Здесь величина $\frac{b}{a}$ — это отношение числа выбывших из страхования к числу состоящих на страховании к началу года.

Отклонение $\frac{q_1}{q_2}$ от единицы покажет величину погрешности из-за неточно взятого Q. Разбираются два предельных случая.

1. Все выбывшие в течение года пробудут под наблюдением немногим меньше полугода. В этом случае $Q_1=0{,}5$; $Q_2=1$. Тогда

$$\frac{q_1}{q_2} = \frac{1-\frac{b}{a}}{1-0{,}5\frac{b}{a}}.$$

2. Все выбывшие в течение года пробудут под наблюдением немногим больше полугода. В этом случае $Q_1=0{,}5$; $Q_2=0$. Тогда

$$\frac{q_1}{q_2} = \frac{1}{1-0{,}5\frac{b}{a}}.$$

При разных значениях $\frac{b}{a}$ отношения $\frac{q_1}{q_2}$ будут иметь следующие значения:

$b:a$	$q_1:q_2$ случай 1	$q_1:q_2$ случай 2
0,1	0,95	1,05
0,2	0,89	1,11
0,3	0,82	1,18

Даже при 10% выбытия ошибка в определении q не превышает 5%. При 20% выбытия ошибка составит 11%, при 30% — ошибка 18%, но фактически такой высокий процент выбытия маловероятен.

Используя статистические данные об отношениях $\frac{b}{a}$ и вычисляя $\frac{q_1}{q_2}$, Б. С. Ястремский получил величины, мало отличающиеся от единицы. Все это привело ученого к выводу, что особой точности в определении Q не требуется, потому простой, хотя и неточный, метод Спрейга в достаточной мере соответствует своему назначению.

1.7. Вероятностные методы выявления заболеваний

Один из основателей отечественной санитарной статистики проф. А. М. Мерков выделил группу проблем медицинской статистики, или статистики клинико-статистических исследований. «В ней, — писал А. М. Мерков, — особенно широко применяются приемы исследования, основанные на теории вероятностей и малой выборки»[1].

Установление правильного диагноза и указание конкретного заболевания у лица, обратившегося за медицинской помощью, представляется в большинстве случаев делом весьма трудным, потому что очень мало болезней, которым свойственны специфические симптомы, никогда не встречающиеся при других заболеваниях.

Используя вероятностные методы, можно прийти к выводу о наличии у обратившегося за врачебной помощью того или иного заболевания и установить диагноз или опровергнуть необоснованное предположение.

При установлении диагноза болезни вероятностным путем сначала устанавливают вероятность наличия симптомов A при заболевании D, т. е. $p(A/D)$. Дополнение этой вероятности до единицы, т. е. $1-p(A/D)$, означает вероятность отсутствия данного симптома. В качестве указанных вероятностей используют полученные статистическим путем данные о частостях симптомов A при предполагаемой болезни D.

[1] *Мерков А. М.* История санитарной и медицинской статистики в СССР. — В кн.: Советская статистика за полвека (1917—1967 гг.). М., 1970, с. 220. (Учен. зап. по статистике АН СССР, т. XVII).

Пусть вероятности наличия симптомов при предполагаемом заболевании D_1 составляют следующие величины[1]: $p(A_1/D_1)=0{,}90$; $p(A_2/D_1)=0{,}58$; $p(A_3/D_1)=0{,}38$; $p(A_4/D_1)=0{,}72$. Эти же симптомы могут иметь место у больных, страдающих другим заболеванием (D_2), с вероятностями: $p(A_1/D_2)=0{,}29$; $p(A_2/D_2)=0{,}30$; $p(A_3/D_2)=0{,}17$; $p(A_4/D_2)=0{,}51$. Сопоставляя эти величины и предположив, что у лиц с данным заболеванием вероятность равна: $p(D_1)=0{,}25$, а у прочих — $p(D_2)=0{,}75$, автор получил ответ на вопрос о вероятности заболевания D_1:

$$p(D_1/A)=\frac{p(D_1)[1-p(A_1/D_1)]p(A_3/D_1)}{p(D_2)[1-p(A_1/D_2)]p(A_3/D_2)}\cdot\frac{[1-p(A_4/D_1)]}{[1-p(A_4/D_2)]}=$$
$$=\frac{0{,}25\cdot0{,}10\cdot0{,}38\cdot0{,}28}{0{,}75\cdot0{,}71\cdot0{,}17\cdot0{,}49}=0{,}07.$$

На основании малой величины вероятности был сделан вывод о том, что, очевидно, больной не страдает тем заболеванием, которое первоначально у него предполагалось.

Приведем пример часто встречающейся ситуации. Известны частости заболевания больных, поступающих в специализированные больницы, и вероятности полного излечения от каждого заболевания (табл. 1.1). Больной, поступивший в больницу, был выписан здоровым. Какова вероятность того, что он страдал заболеванием k (первая строка табл. 1.1)? Воспользуемся для решения этой задачи теоремой Бейеса (теоремой о вероятности гипотез), позволяющей после опыта, приведшего к определенному результату (больной был выписан здоровым), переоценивать вероятности гипотез и получать апостериорные вероятности — $P_{\Gamma(i)}$. Примем частости поступления больных в больницу $(p_{\Gamma(i)})$ за вероятности i-х гипотез до испытания (априорные вероятности). Вероятности события A (т. е. полного излечения) при предположении, что данная гипотеза осуществлялась, приведены в графе $p_{A(i)}$.

Таблица 1.1. **Заболевания и вероятности** (априорные и апостериорные) **их излечения**

№ п/п	Заболевания (гипотезы)	Частости больных (априорные вероятности), $p_{\Gamma(i)}$	Вероятности полного излечения, $p_{A(i)}$	$p_{\Gamma(i)}\,p_{A(i)}$	Апостериорные вероятности, $P_{\Gamma(i)}$
1	k Г(1)	0,5	0,7	0,35	0,46
2	l Г(2)	0,3	0,8	0,24	0,31
3	m Г(3)	0,2	0,9	0,18	0,23
Итого	—	$\Sigma p_{\Gamma(i)}=1{,}0$	—	$\Sigma p_{\Gamma(i)}\cdot p_{A(i)}=0{,}77$	$\Sigma P_{\Gamma(i)}=1{,}00$

[1] См.: *Соколов Д. К.* Математическое моделирование в медицине. М., 1974, с. 97.

Находим полную вероятность события A как сумму произведений вероятностей гипотез на вероятности события A при условии осуществления каждой данной гипотезы:

$$P_A = \Sigma p_{\Gamma(i)} \cdot p_{A(i)} = 0{,}77.$$

Используя формулу Т. Бейеса, получаем, например, для первой строки:

$$P_{\Gamma(1)} = \frac{p_{\Gamma(1)} \cdot p_{A(1)}}{\Sigma p_{\Gamma(i)} p_{A(i)}} = \frac{0{,}35}{0{,}77} = 0{,}46.$$

Полученный результат дает ответ на интересующий нас вопрос. Таким образом, априорная вероятность заболевания k, равная 0,5, может быть заменена апостериорной вероятностью, равной 0,46.

Рассмотрим на примере переоценку априорных вероятностей, производимую с помощью теорем Д. Бернулли и Бейеса. При обследовании больного имеется подозрение на одно из трех заболеваний: A_1, A_2 и A_3. Априорные вероятности этих заболеваний (частости) соответственно равны: $p_{A_1} = 1/2$; $p_{A_2} = 1/6$; $p_{A_3} = 1/3$. Их сумма ($\Sigma p_{A(i)}$) и образует полную систему событий.

Для уточнения диагноза производится три анализа, давшие в двух случаях положительный результат (событие B), вероятность которого при указанных заболеваниях различна:

при заболевании $A_1 - p'_{A_1} = 0{,}1$;

при заболевании $A_2 - p'_{A_2} = 0{,}2$;

при заболевании $A_3 - p'_{A_3} = 0{,}9$.

Используя теорему Бернулли, найдем вероятность события B в случае заболевания A_1, A_2 и A_3:

$$P[B/A_1] = C_3^2 \cdot 0{,}1^2 \cdot 0{,}9 = 0{,}027;$$
$$P[B/A_2] = C_3^2 \cdot 0{,}2^2 \cdot 0{,}8 = 0{,}096;$$
$$P[B/A_3] = C_3^2 \cdot 0{,}9^2 \cdot 0{,}1 = 0{,}243.$$

Соответствующие произведения равны:

$$p_{A_1} P[B/A_1] = \frac{1}{2} \cdot 0{,}027 = 0{,}0135;$$

$$p_{A_2} P[B/A_2] = \frac{1}{6} \cdot 0{,}096 = 0{,}0160;$$

$$p_{A_3} P[B/A_3] = \frac{1}{3} \cdot 0{,}243 = 0{,}0810;$$

$$\Sigma p_{Ai} P[B/A_i] = 0{,}0135 + 0{,}0160 + 0{,}0810 = 0{,}1105.$$

Теперь, используя теорему Бейеса, найдем апостериорные вероятности каждого заболевания:

$$P[A_1] = \frac{p_{A_1} P[B/A_1]}{\Sigma p_{A_i} P[B/A_i]} = \frac{0{,}0135}{0{,}1105} \approx 0{,}12;$$

$$P[A_2] = \frac{p_{A_2} P[B/A_2]}{\Sigma p_{A_i} P[B/A_i]} = \frac{0{,}0160}{0{,}1105} \approx 0{,}15;$$

$$P[A_3] = \frac{p_{A_3} P[B/A_3]}{\Sigma p_{A_i} P[B/A_i]} = \frac{0{,}0810}{0{,}1105} \approx 0{,}73;$$

$$\Sigma P[A_i] = 0{,}12 + 0{,}15 + 0{,}73 = 1{,}00.$$

После анализов априорные вероятности заболеваний должны быть заменены на апостериорные. При этом, как мы видим, резко увеличилась априорная вероятность заболевания A_3 (с $\frac{1}{3}$ до 0,73) и снизилась априорная вероятность заболевания A_1 (с $\frac{1}{2}$ до 0,12).

При проведении медицинского обследования населения для выявления больных путем анализа крови вероятностные методы позволяют выбрать наиболее целесообразный метод. Разработаны два способа проведения анализа:

анализ крови каждого лица, составляющего изучаемую совокупность. Тогда число произведенных анализов будет равно численности совокупности (N);

предварительная группировка изучаемой совокупности на равные группы численностью n и проведение анализа *смешанной крови всех лиц, составляющих группы.* В этом случае отрицательный ответ означает, что все лица, входящие в состав группы, здоровы. Положительный ответ свидетельствует о наличии одного или нескольких больных в группе и необходимости последующего анализа каждого лица, входящего в данную группу.

Пусть вероятность того, что лицо, подвергнутое обследованию, является больным, равно p, а здоровым — $q = 1 - p$. Ответим на два вопроса:

1. При каком способе анализа крови число требуемых анализов будет меньшим? Каковы при этом ограничительные условия?

2. При какой численности групп потребуется меньшее число анализов, приходящееся в среднем на группу и на одно лицо в группе?

При первом способе, как мы уже знаем, число анализов будет равно числу обследуемых, т. е. N, а число анализов, приходящихся в среднем на группу, — соответственно n.

При втором способе возможны следующие исходы:

если анализ смешанной крови оказался отрицательным, т. е. больных в данной группе нет и все n человек здоровы, этот анализ остается единственным;

если первый анализ положительный, то для выявления конкретных больных необходимо произвести анализы крови всех лиц данной группы; в этом случае число анализов составит $n+1$.

Распределение случайной величины, т. е. числа анализов крови (x_i), будет иметь вид:

Значения случайной величины, x_i	1	$n+1$
Вероятности исходов, P_i	q^n	$1-q^n$

Найдем среднее число анализов на группу из n лиц, равное математическому ожиданию случайной величины x_i по формуле

$$\bar{x} = \mathrm{E}(x_i) = \Sigma x_i P_i = 1 \cdot q^n + (n+1)(1-q^n) = n - (nq^n - 1),$$

где $E(x_i)$ — математическое ожидание случайной величины x_i, равное сумме произведений всех значений случайной величины на их вероятность.

Таким образом, при первом способе среднее число анализов на группу равно n, при втором — $n-(nq^n-1)$. При $nq^n<1$ первый способ анализа крови целесообразнее, а при $nq^n>1$ второй способ приводит к меньшему числу анализов.

Из неравенства $nq^n>1$ установим, при каком значении вероятности лицу, подвергающемуся обследованию, быть здоровым (q), второй способ более целесообразен:

$$q>\frac{1}{\sqrt[n]{n}}\,.$$

Учитывая, что при целых значениях n минимальная величина $\frac{1}{\sqrt[n]{n}}$ достигается при $n=3$, получаем $q>0,694$.

При втором способе проведения анализа крови среднее число анализов, приходящееся на одно лицо (R_n), составит:

$$R_n=\frac{\overline{x}}{n}=1-q^n+\frac{1}{n}\,.$$

Решение этой задачи и определение n приводит к выводу о том, что при условии неравенства $0,694<q<0,876$ оптимальная численность групп, обусловливающая наименьшее число анализов, составляет 3 лица.

Во время второй мировой войны анализ смешанной крови применялся в армейских условиях и давал экономию в числе анализов до 80%[1].

1.8. Критика буржуазных концепций, неправомерно использующих вероятностные методы

Многие исследователи, привлекая теорию вероятностей для изучения демографических явлений, упускали из виду своеобразие этих явлений и приходили подчас к ложным выводам. В своем противоречивом и весьма слабом сочинении религиозного характера, написанном в 1741 г. по данным церковных книг и называвшемся «Божественный порядок в изменениях человеческого рода...», И. П. Зюссмильх пытался путем объединения различных таблиц смертности, построенных методом смертных списков для Парижа, Берлина, Вены и других городов и провинций, доказать, что, несмотря на различие условий жизни на разных территориях, в ходе смертности фиксируется какой-то общий, «высший» порядок. И. П. Зюссмильх писал, что его книга представляет собой опыт приложения теории вероятностей к явлениям человеческой жизни. На самом деле он и не пытался найти объяснение закономерностей и причиненных связей демографических явлений, а искал существование в человеческой жизни предопределенного, при этом *божественного порядка.*

[1] См.: *Савельев Л. Я.* Комбинаторика и вероятность. Новосибирск, 1975, с. 408.

По мысли И. П. Зюссмильха, творец при создании человека наделил его конкретными свойствами, а общественную жизнь подчинил определенным законам, которые человек изменить не может. Будучи не в состоянии правильно объяснить закономерность массовых явлений (наличие постоянных соотношений среди случайных явлений), он объявил ее выражением воли божества, которое при помощи закона больших чисел руководит человечеством. Таким образом, закон больших чисел приобрел черты таинственности и непознаваемости.

А. Кетле попытался создать науку о человеческом обществе, назвав ее социальной физикой. Он сформулировал положение о том, что закономерности общественных и демографических явлений можно выражать при помощи формул теории вероятностей (рост людей, размеры частей человеческого тела и их пропорции, закономерности их изменения с возрастом, рождаемость, смертность и т. д.), используя законы распределения вероятностей. А. Кетле показал, что статистические результаты измерения явлений, относящихся к антропологии, биологии, демографии, а в ряде случаев и к области социально-экономических явлений, имеют такой же характер, как и статистические результаты измерения в физике, астрономии и ряде других естественных наук. Именно поэтому к анализу этих результатов применима теория ошибок, выведенная К. Гауссом, а также возможно использование всего математико-статистического и вероятностного аппарата. Однако, по мнению А. Кетле, результаты измерений в области социально-экономических явлений более сложны, чем результаты простых измерений в физике, и требуют более тонких подходов и специальных формул.

Выяснив, что такие главные явления человеческой жизни, как рождение человека, переход из возраста в возраст и смерть, подчиняются определенным естественным законам, А. Кетле решил также, что и произвольные действия человека подчинены таким же законам. Таким образом, он встал на точку зрения детерминизма, приведшего его к фатализму — учению, отрицающему случайность и какую-либо свободу человеческих действий. Человек, по мнению А. Кетле, бессилен против железных законов окружающей природы и общества (капиталистического), и все его поступки независимо от его воли предопределены действием этих законов. Задача, следовательно, состоит в раскрытии премудрых законов, управляющих миром, против которых человек бессилен.

А. Кетле, пытаясь доказать, что присущие капитализму социальные факторы (безработица, нищета, преступность и др.) не зависят от социальных условий, являются «вечными» и «неизменными», выступил с «теорией устойчивости статистических рядов»». Математическое доказательство этой «теории» дал В. Лексис, предложивший критерий, с помощью которого можно якобы устанавливать, устойчив или неустойчив тот или иной динамический

4 Заказ 2715

ряд. При этом В. Лексис отвергал качественный анализ, т. е. игнорировал природу исследуемых явлений. Он считал, что достаточно одной статистики для доказательства полной устойчивости ряда.

Советские статистики Б. С. Ястремский, А. Я. Боярский и другие доказали, что «теория устойчивости» является удобным орудием апологетики капитализма, и вскрыли ее лженаучность. При различной группировке данных получаются разные результаты, т. е. один и тот же ряд может оказаться устойчивым и неустойчивым в зависимости от того, какими группами он представлен. Следовательно, подбор соответствующих группировок определяет устойчивость статистического ряда различных показателей, на основании чего делается вывод об устойчивости капитализма.

В противоположность буржуазной статистике, базирующейся на «теории устойчивости», советская статистика опирается на диалектический материализм и не ищет вечных и неизменных норм, а постигает сущность изменяющихся явлений на основе изучения их материальной природы.

А. Кетле является также создателем теории «среднего» человека, в котором воплощены все усредненные качества обычных людей и о сохранении которого как типа заботится природа. Этот воображаемый, логически немыслимый и фактически не существующий человек имеет все характеристики, складывающиеся из данных, характеризующих конкретных людей: вес, быстроту бега, вероятность умереть, наклонность к браку, к самоубийству, к добрым делам и т. д. Свойства этого «среднего» человека вечны и неизменны, а их изучение есть статистическая задача. Разумеется, что постичь физические, духовные, нравственные и интеллектуальные свойства этого человека можно лишь с использованием теории вероятностей.

Идею А. Кетле о «среднем» человеке поддержал В. Лексис. Он считал, что абстрактного «среднего» человека, демографическая биография которого должна быть описана с использованием теории вероятностей, следует представить вначале без всяких отличительных признаков: это не мужчина и не женщина, он не женат и не холост, не молод и не стар, но всякое из этих качеств присуще ему с *определенной вероятностью*. Например, прежде чем установлен пол новорожденного, существует вероятность — приблизительно 0,515, — что это мальчик, и 0,485, — что девочка.

Таким образом, сложная жизненная система человека, состоящая из различных взаимоувязанных органов, была заменена А. Кетле системой из усредненных значений, при этом оказалось, что далеко не всегда эти значения совместимы между собой, т. е. не всегда они обеспечивают жизнедеятельность человеческого организма.

Теорию «среднего» человека справедливо критиковали и статистики, и экономисты, и философы, хотя иногда эта критика подкреплялась не совсем удачными примерами.

Ж. Даламбер совершил ошибку, применяя теорию вероятностей к демографическим показателям. Его смутило различие между вероятной и средней продолжительностью жизни (в этом он увидел недостаток теории вероятностей).

Из того факта, что в качестве характеристики продолжительности жизни демография использует такие два показателя, как средняя продолжительность предстоящей жизни и вероятная продолжительности жизни[1], Ж. Даламбер делает вывод, что на один и тот же вопрос мы получаем два ответа. Эти два показателя действительно подводят итог условиям смертности, но с разных точек зрения.

Могут ли они быть равными для всех возрастов? Да, могут, но лишь тогда, когда распределение смертных случаев происходит равномерно в течение всей жизни, т. е. плотность смертей во всех возрастах одинакова, чего в действительности не наблюдается.

Известны ошибочные критические замечания Ж. Даламбера в адрес Д. Бернулли, доказавшего при помощи теории вероятностей, что оспопрививание увеличивает среднюю продолжительность жизни на 3—4 года и поэтому с точки зрения интересов государства является положительным фактором. Ж. Даламбер, используя геометрические построения, разработал собственный метод сравнения рисков смерти от оспы и от инокуляции и, используя свои расчеты, возражал против пользы оспопрививания. Он противопоставил интересам государства интересы отдельного человека, который подвергается опасности умереть через месяц непосредственно от прививки. При этом вероятность умереть от прививки принята им равной $^1/_{200}$.

Однако из литературных источников известно, что Ж. Даламбер к концу жизни отказался от своих ошибочных взглядов.

В своих работах П. Лаплас поддерживал выводы Д. Бернулли о пользе прививок.

Не избежали ошибок при использовании теории вероятностей в демографии и некоторые отечественные статистики. Так, Ю. Э. Янсон поставил перед собой задачу определить вероятность смерти от туберкулеза жителей Петербурга в возрасте 16—20 лет. Для этого он привлек статистические данные по Петербургу:

численность группы в возрасте 16—20 лет — 113 915 человек;
общее число умерших от разных причин в этом возрасте — 934 человека;
из них число умерших от туберкулеза — 366 человек.

Далее Ю. Э. Янсон вычислил следующие две вероятности:

$$\text{вероятность смерти от разных причин, равную } \frac{934}{113\,915} = 0{,}0082;$$

$$\text{вероятность смерти от туберкулеза, равную } \frac{366}{113\,915} = 0{,}0032.$$

[1] Напомним: средняя продолжительность жизни представляет собой математическое ожидание предстоящей жизни, равное отношению числа человеко-лет, которое проживут дожившие до x лет после этого возраста, на число доживших до x лет; вероятная продолжительность жизни — медиана в распределении по продолжительности предстоящей жизни, равная числу лет, которое проживет после возраста x лет ровно половина достигших этого возраста.

4*

Неправильно поняв теорему умножения вероятностей, Ю. Янсон продолжил вычисления и получил абсурдно малую величину $0{,}0082 \times 0{,}0032 = 0{,}00002627$, представляющую, по его мнению, вероятность смерти от туберкулеза.

В действительности 0,0032 — это и есть искомая вероятность, полученная непосредственно из исходных данных. Чтобы получить эту же вероятность, применяя теорему умножения вероятностей, надо было умножить вероятность смерти от разных причин (0,0082) на долю умерших от туберкулеза среди всех умерших, равную 366/934. Тогда и получилась бы правильная величина — 0,0032.

Внук Чарлза Дарвина — Чарлз Гальтон Дарвин — предложил термодинамическую теорию народонаселения, позволяющую якобы на основе теории вероятностей и распространенного на общественные явления закона Бойля — Мариотта о зависимости объема газов от давления и температуры предсказывать будущее человечества с такой же точностью, с какой астрономы вычисляют движение светил, солнечные и лунные затмения. Для этого только нужно очень немногое. Следует представить, что человеческое общество — это сосуд, наполненный газом, а люди — это молекулы.

Об ошибочном толковании вероятности, основанной на частоте событий, пишет Д. Пойя в анекдотическом примере. После осмотра больного врач заявил ему о крайней серьезности этого заболевания: из десяти заболевших ею остается в живых только один (т. е. вероятность умереть $p=0{,}9$). Продолжая разговор, врач сказал, обращаясь к больному, что тому повезло и что он останется в живых, потому что он 10-й: а из девяти, уже обратившихся к нему с этой болезнью, все девять умерли.

Пример того, как легко ошибиться, определяя вероятность случайного события, приводит В. В. Налимов[1]. Четыре короля из Гановерской династии — Эдуарды I, II, III и IV — умерли в один и тот же день недели — в субботу. Теоретически вероятность такого совпадения чрезвычайно мала:

$$p = \frac{1}{7^4} = \frac{1}{2\,401} \approx 0{,}0004.$$

Возникает вопрос: можно ли исходя из указанной малой вероятности сделать вывод о том, что суббота является роковым днем для данной династии? Одновременно возникает встречный вопрос: достаточно ли у нас данных и не нужно ли привлечь дополнительную информацию? Включаем в условие данные о том, что Гановерская династия насчитывала 53 короля. Тогда вероятность смерти четырех королей одного имени в какой-либо из дней недели составит уже $^1/_7$ и такое совпадение не покажется столь исключительным, как ранее.

Известно, что итальянский математик XVI столетия Д. Кардано, используя теорию вероятностей, рассчитал вероятную дату своей смерти. Чтобы действительность не опровергла его расчетов, ученый при приближении рокового дня подверг себя голоданию. Курьез этого случая состоял в том, что Д. Кардано закономерности массового явления — вероятность умереть в определенном возрасте, свойственную группе людей этого возраста, — перенес на конкретный случай, имея в виду самого себя.

[1] См.: *Налимов В. В.* Теория эксперимента, М., 1971, с. 35.

В ряде случаев для характеристики интенсивности рождений некоторые исследователи определяют долю новорожденных среди мужчин ($\Delta_м$) и долю новорожденных среди женщин ($\Delta_ж$), а затем находят среднюю арифметическую из указанных долей ($\bar{\Delta}$). Покажем, что в этом случае найденный показатель имеет систематическую ошибку: он завышает или занижает интенсивность рождений.

Имеем:

$$\Delta_м = \frac{N_м}{S_м} \text{ и } \Delta_ж = \frac{N_ж}{S_ж}; \ \bar{\Delta} = \frac{\dfrac{N_м}{S_м} + \dfrac{N_ж}{S_ж}}{2}.$$

Коэффициент рождаемости n равен:

$$n = \frac{N_м + N_ж}{S_м + S_ж}.$$

Тогда

$$\bar{\Delta} - n = \frac{\dfrac{N_м}{S_м} + \dfrac{N_ж}{S_ж}}{2} - \frac{N_м + N_ж}{S_м + S_ж}.$$

После преобразований получаем в правой части равенства:

$$\frac{0{,}5(S_ж - S_м)\cdot(S_ж N_м - S_м N_ж)}{S_м S_ж (S_м + S_ж)}.$$

В том случае, когда женщин больше, чем мужчин, имеем $S_ж - S_м > 0$ и $S_ж N_м - S_м N_ж > 0$, в числителе и в знаменателе все множители положительны. Следовательно, разность $\Delta - n$ есть величина положительная, т. е. $\bar{\Delta} > n$.

Приведем пример. Пусть $S_м = 2000$, $S_ж = 3000$, $S_м + S_ж = 5000$, $N_м = 60$, $N_ж = 50$, $N_м + N_ж = 110$.

Получаем:

$$n = \frac{110}{5\,000} = 0{,}022; \quad \Delta_м = \frac{60}{2\,000} = 0{,}03; \quad \Delta_ж = \frac{50}{3\,000} = 0{,}01667;$$

$$\bar{\Delta} = \frac{0{,}03 + 0{,}01667}{2} = \frac{0{,}04667}{2} = 0{,}02333.$$

Находим разность, т. е. величину завышения интенсивности рождаемости: $\bar{\Delta} - n = 0{,}02333 - 0{,}022 = 0{,}00133$. В других случаях получаем занижение интенсивности рождений. Ошибка данного метода состоит в осреднении долей мальчиков и девочек без взвешивания. Если же находить среднюю взвешенную долю и в качестве весов использовать численности мужчин и женщин, то она окажется равной коэффициенту рождаемости:

$$\overline{\Delta} = \frac{\Delta_м S_м + \Delta_ж S_ж}{S_м + S_ж} = \frac{N_м + N_ж}{S_м + S_ж} = n.$$

Пренебрежение законами случая, которым часто грешат социологические исследования, как считает А. Я. Боярский, «может за себя отомстить». Так, если показатель смертности равен 20‰ для возрастной группы численностью в 5 тыс. человек, то утроенная средняя квадратическая ошибка, указывающая на возможность колебания этого коэффициента смертности, равна 6‰. Значит, показатель смертности может колебаться в пределах от 14 до 26‰. При этом такие колебания могут носить случайный характер. Поэтому всякие выводы о существенности отклонений в этом интервале будут неправильными. Таким образом, мы видим, что многие исследователи из-за неверных исходных посылок и непонимания специфики науки, изучающей случайные явления, приходили к неправильным, иногда даже порочным выводам. Чтобы в какой-то степени снизить вероятность возникновения такого рода ошибок, исследователь должен глубже познать «законы случая».

Глава 2

Распределение признаков в демографии

2.1. Законы распределений

Расположенные в ранжированном порядке значения признаков образуют вариационный ряд (распределение признака), который можно изображать графически, например в виде полигона или гистограммы. Очень часто в демографии возникает необходимость определения степени приближения фактического распределения признака к теоретическому. При этом под теоретическим распределением признака понимают распределение вероятностей, лежащих в основе наблюдаемых частот вариантов.

Согласно махистской концепции, если фактическое распределение признака укладывается более или менее удовлетворительно в ту или иную математическую схему и аппроксимируется особой формой кривой, разработанной К. Пирсоном, то математическая формула, описывающая эту кривую, и является законом данного явления. Б. С. Ястремский, опираясь на марксистский диалектический метод, выявил эмпиризм построений К. Пирсона и опроверг эти построения.

Очень хорошим способом описания изменчивости демографических явлений служит привлечение определенных законов распределений, позволяющих найти вероятность того, что конкретные значения показателей изучаемых явлений находятся в определенном, заранее заданном интервале. Многие демографические явления аппроксимируются законом биномиального распределения или близкими к биномиальному — нормальным, логарифмически-нормальным (логнормальным), распределением Пуассона и др.

В демографии существуют и специфические построения распределений, например кривая распределения продолжительности жизни. Она строится таким образом, чтобы площадь, лежащая под каждым ее участком, представляла собой вероятность того, что новорожденный умрет в соответствующем возрастном интервале. Такая кривая называется кривой распределения вероятностей (рис. 2.1).

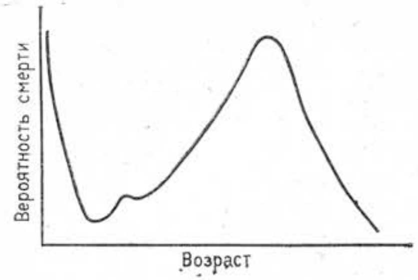

Рис. 2.1. Кривая распределения вероятностей

Для различных возрастов вероятности умереть неодинаковы и зависят от величины возрастного интервала, принятого за основу расчета, от времени, к которому приурочено построение кривой, от природно-климатических особенностей территории, на которой проживает изучаемое население, от социального состава изучаемой группы населения и т. д. Вот почему на основе одних лишь теоретических рассуждений построить кривую распределения вероятности жизни, которая выражала бы действительный порядок вымирания изучаемого населения, невозможно. Нужны статистические данные, позволяющие установить конкретную долю населения, умирающего в указанных возрастных границах. Рассматривая случаи смерти людей как взаимно независимые события и привлекая обширный статистический материал, исследователь приходит к выводу, что статистические частости (отношения числа смертей в данных возрастных интервалах к числу живущих в этих возрастных интервалах) хорошо воспроизводят действительный порядок вымирания населения. При изучении процесса или явления в динамике особое значение имеет определение тенденции изменения вероятностных показателей.

2.2. Биномиальное и полиномиальное распределения

Биномиальное распределение представляет распределение вероятностей возможных чисел появления события A при независимых испытаниях, в каждом из которых это событие A может осуществиться с одной и той же вероятностью p.

В теории вероятностей разработаны правила, выведены теоремы и формулы, позволяющие на основе формализации всех действий получать ответы на вопросы об исходах событий, о числе случаев при осуществлении каждого исхода и об общем числе исходов и случаев.

Вероятности любого числа событий соответствуют членам разложения бинома Ньютона в степени, равной числу испытаний:

$$(p+q)^n = p^n + np^{n-1}q + C_n^{n-2}p^{n-2}q^2 + C_n^{n-3}p^{n-3}q^3 + \ldots$$
$$\ldots + C_n^m p^m q^{n-m} + \ldots + q^n,$$

где p — вероятность события A; $q=1-p$ — вероятность того, что событие A не произойдет; n — число испытаний; m — число осуществлений события A, или частота события A; $C_n^{n-2}=C_n^2$; $C_n^{n-m}=C_n^m$ — числа сочетаний из n элементов по $n-2$, $n-m$ элементов; p^n — первый член биномиальной строки; его величина соответствует вероятности такого исхода (комбинации), при котором событие A осуществляется n раз; $np^{n-1}q$ — второй член биномиальной строки; его величина соответствует вероятности такого исхода, при котором событие A осуществляется $n-1$ раз, а не осуществляется один раз, и т. д. до q^n — последнего члена строки, дающего вероятность такого исхода, при котором событие A ни разу не осуществляется.

Таким образом, вероятность осуществления события A m раз в n независимых испытаниях с одинаковой вероятностью p можно рассчитать по формуле общего члена разложения бинома Ньютона:

$$P_m = C_n^m p^m q^{n-m} = \frac{n!}{m!\,(n-m)!}\, p^m q^{n-m},$$

где P_m — вероятность того, что при n испытаниях событие A осуществляется m раз; $n! = 1\cdot 2\cdot 3\ldots n$ — произведение натурального ряда чисел от 1 до n[1].

Покажем описание пространства событий при рождении в семье четверых детей. Получаем 16 равновозможных исходов, составляющих пространство событий (табл. 2.1).

Таблица 2.1. **Описание пространства событий для семей с четырьмя детьми**

1-й ребенок	2-й ребенок	3-й ребенок	4-й ребенок	Пространство событий
Мальчик	Мальчик	Мальчик	Мальчик	мммм
			Девочка	мммд
		Девочка	Мальчик	ммдм
			Девочка	ммдд
	Девочка	Мальчик	Мальчик	мдмм
			Девочка	мдмд
		Девочка	Мальчик	мддм
			Девочка	мддд
Девочка	Мальчик	Мальчик	Мальчик	дммм
			Девочка	дммд
		Девочка	Мальчик	дмдм
			Девочка	дмдд
	Девочка	Мальчик	Мальчик	ддмм
			Девочка	ддмд
		Девочка	Мальчик	дддм
			Девочка	дддд

[1] 0! считается равным единице.

Рассмотрим использование различных видов распределений на примере типа семей с шестью детьми. Для простоты рассуждений и последующих расчетов будем исходить из равновозможности всех комбинаций полов детей в семьях. Мы получим семь различных исходов:

1) все дети девочки (м=0; д=6);
2) один мальчик, пять девочек (м=1; д=5);
3) два мальчика, четыре девочки (м=2; д=4);
4) три мальчика, три девочки (м=3; д=3);
5) четыре мальчика, две девочки (м=4; д=2);
6) пять мальчиков, одна девочка (м=5; д=1);
7) все дети мальчики (м=6; д=0).

В разных исходах возможно различное число случаев. Найдем число случаев каждого исхода (C_n^m):

$$\text{при } n=6,\ m=0 \text{ получаем } C_6^0 = \frac{6!}{0!\,(6-1)!} = 1;$$

$$\text{при } n=6,\ m=1 \text{ имеем } C_6^1 = \frac{6!}{1!\,(6-1)} = 6;$$

далее соответственно $C_6^2=15$, $C_6^3=20$, $C_6^4=15$, $C_6^5=6$, $C_6^6=1$.

Найдем все вероятности исходов:

$$P\ (\text{м}=0,\ \text{д}=6) = C_6^0 \cdot 0{,}5^6 = \frac{1}{64} \approx 0{,}015625;$$

$$P\ (\text{м}=1,\ \text{д}=5) = C_6^1 \cdot 0{,}5^6 = \frac{6}{64} \approx 0{,}093750 \text{ и т. д.}$$

Результаты расчетов сводим в табл. 2.2.

Таблица 2.2. **Типы семей, число случаев в каждом типе и соответствующие вероятности**

Исходы	Типы семей (исходы)							
	м=0, д=6	м=1, д=5	м=2, д=4	м=3, д=3	м=4, д=2	м=5, д=1	м=6, д=0	итого
Число возможных случаев	1	6	15	20	15	6	1	64
Вероятность	0,015625	0,093750	0,234275	0,312500	0,234275	0,093750	0,015625	1

Исчисление вероятностей различных исходов по биномиальной формуле сопряжено с трудностями расчетов. П. Лаплас вывел формулы (локальную и интегральную), упрощающие все действия. Покажем расчет по локальной формуле Лапласа[1].

Найдем приближенное значение вероятности первого исхода, т. е. того, что в семье из шести детей все девочки. Запишем условие: $n=6$; $m=0$; $p=0{,}5$; $q=0{,}5$; наивероятнейшая частота: $m_0 \approx np = 6 \cdot 0{,}5 = 3$; $\sqrt{npq} = \sqrt{6 \cdot 0{,}5 \cdot 0{,}5} = \sqrt{1{,}5} = 1{,}22474$; отклонение искомой частоты от наивероятнейшей:

$$x = m - m_0 = 0 - 3 = -3;\ t = \frac{m - m_0}{\sqrt{npq}} = \frac{-3}{1{,}22474} \approx -2{,}45.$$

Тогда искомая вероятность равна: $P_t = \frac{\varphi t}{\sqrt{npq}}$, где $\varphi(t)$ определяется из соответствующих таблиц, при этом $\varphi(-t) = \varphi(t)$.

Получаем

$$P_{t=-2,45} = \frac{\varphi\ (-2{,}45)}{1{,}22474} = \frac{\varphi\ (2{,}45)}{1{,}22474} = \frac{0{,}0198}{1{,}22474} \approx 0{,}016.$$

Соответственно находим вероятность второго исхода по локальной формуле Лапласа: $n=6$; $m=1$; $x=1-3=-2$; $t = \frac{-2}{1{,}22474} \approx -1{,}63$; $\varphi(-1{,}63) = \varphi(1{,}63) = 0{,}1057$; после подстановки получаем:

$$P_{t=-1,63} = \frac{0{,}1057}{1{,}22474} \approx 0{,}086.$$

Далее рассчитываем вероятность третьего исхода: при $n=6$; $m=2$; $x=2-3=-1$; $t=-0{,}82$; $P_{t=-0,82}=0{,}233$.

Вероятность наивероятнейшего исхода, т. е. исхода, при котором число мальчиков и девочек равно трем, составляет $P_{t=0} = \frac{0{,}3989}{1{,}22474} \approx 0{,}330$. Сумма всех вероятностей равна единице.

Привлекая теорему сложения вероятностей, можно определить вероятность того, что число мальчиков в семьях с шестью детьми заключено в определенных границах, суммируя найденные по локальной формуле Лапласа вероятности различных исходов.

Однако вероятность того, что в семьях с шестью детьми число мальчиков заключается в пределах, например, от одного (a) до четверых (b), можно найти и минуя эти расчеты — по интегральной формуле Лапласа, имеющей вид:

$$P[a \leqslant m \leqslant b] = \frac{1}{\sqrt{2\pi}} \int_{\alpha}^{\beta} e^{-\frac{t^2}{2}}\,dt = \frac{1}{2}\left[F(t=\beta) - F(t=\alpha)\right],$$

$$\text{где } \alpha = \frac{a - m_0}{\sqrt{npq}};\quad \beta = \frac{b - m_0}{\sqrt{npq}}.$$

[1] См.: *Венецкий И. Г., Венецкая В. И.* Основные математико-статистические понятия и формулы в экономическом анализе. М., 1979, приложение.

Привлекая данные нашего примера ($a=1$ и $b=4$), получаем: $\alpha=$ $=\frac{1-3}{1,22474}\approx-1,63$; $\beta=\frac{4-3}{1,22474}=0,82$. Тогда

$$P[1\leqslant m\leqslant 4]=P[-1,63\leqslant m\leqslant 0,82]=\frac{1}{2}[F(0,82)-$$

$$-F(-1,63)]=\frac{1}{2}[F(0,82)+F(1,63)].$$

Используя таблицы, находим $F(0,82)=0,58778$; $F(1,63)=0,89690$; сумма равна 1,48468.

$$\frac{1}{2}\cdot 1,48468\approx 0,742.$$

Используя найденные вероятности исходов по формуле биномиального распределения из табл. 2.2, получаем примерно такую же величину: $0,093750+0,234375+0,312500+0,234375\approx 0,875$.

При оценке влияния различных причин на дожитие населения до старости Б. Ц. Урланис ввел в анализ 12 факторов (условия труда, условия отдыха, санитарно-гигиенический уровень и т. д.). Для рассуждения по самой элементарной схеме предполагалось, что влияние каждого из этих факторов отмечается не по степени их действия на человеческий организм, а фиксируется только как положительное или отрицательное (вероятность равна $\frac{1}{2}$). В результате получилось биномиальное распределение вероятностей $\left(\frac{1}{2}+\frac{1}{2}\right)^{12}$. У всех 13 слагаемых знаменатель равен: $2^{12}=4096$, а числители составляют: 1, 12, 66, 220, 495, 792, 924, 792, 495, 220, 66, 12, 1. Таким образом, если в совокупности 4096 человек, то только одному (первый член ряда) из них в этих условиях исключительно повезет: он будет иметь всю жизнь особенно благоприятные действия всех факторов, что приведет к долгой жизни. На организм двенадцати человек (второй член ряда) благоприятно действуют только одиннадцать факторов, а один — неблагоприятно и т. д.

Разумеется, что в действительности нужно учитывать влияние не 12 факторов, а значительно большего их числа. Вот почему при определении, например, уровней смертности от отдельных причин необходимо использовать не биномиальные прерывные распределения, а непрерывные кривые распределения вероятностей, более близкие к практическим нуждам демографии.

М. С. Бедный считает, что старение организма, как и смерть, носит вероятностный характер: по мере старения организма, т. е. по мере увеличения возраста человека, вероятность смерти повышается. Привлекая биномиальный закон распределения вероятностей и предположив, что вероятности p и q равны, он произвел отсчет от предельного, по его мнению, возраста 110 лет в сторону более ранних возрастов, действуя методом проб и ошибок.

При $n=10$ он получил: $\left(\frac{1}{2}+\frac{1}{2}\right)^{10}=0,001+0,010+0,044+0,117+$ $+0,205+0,246+0,205+0,117+0,044+0,010+0,001$. Этот результат автор отверг как неудовлетворяющий принятому условию: вероятность смерти к предельному возрасту 110 лет, равная в этом случае 0,001, не соответствует эмпирическим данным и недостаточно мала. При $n=20$ вероятность дожития до 110-летнего возраста практически равна нулю. М. С. Бедный счел такое распределение вероятностей $\left(\frac{1}{2}+\frac{1}{2}\right)^{20}$ соответствующим поставленной задаче. Тогда возраст начала старения устанавливается как разность 110—20 и составляет 90 лет. На основе полученных вероятностей была построена гипотетическая таблица смертности, устанавливающая средние сроки потенциальной продолжительности жизни. Таблица, по мнению М. С. Бедного, позволяет «... уяснить ряд закономерностей в доживаемости лиц престарелого возраста, что почти никогда не удавалось сделать на основании эмпирических данных»[1].

Анализ данных с помощью биномиального распределения позволил К. Пирсону разложить кривую смертности на пять одномодальных кривых, соответствующих пяти видам смертности: младенческой, в которой начало кривой лежит на $\frac{3}{4}$ года раньше рождения (мертворожденные), детской, юношеской, зрелого возраста и старческой, схожей с нормальной кривой (при небольшой асимметрии).

Рассмотрим использование теоремы о вероятности сложных событий, которая заключается в суммировании произведений вероятностей одного события на условные вероятности других (т. е. вероятности, определяемые при условии, что в каждом случае первое событие осуществилось). Примером такого использования может служить расчет вероятности заболевания при передаче наследственных признаков. Известно, что если ребенок получает от каждого из родителей рецессивный (вызывающий болезнь) ген, то он становится жертвой этой болезни; если ребенок получает такой ген только от одного из родителей, то он будет спасен от болезни, но передаст в свою очередь рецессивный ген примерно половине своих детей и т. д.

Допустим, что доля рецессивных генов в популяции генов составляет 0,02. Используем биномиальное распределение массы браков, рассматриваемых как случайный процесс: $p=0,98$; $q=$ $=0,02$; $n=2$; $(p+q)^2=p^2+2pq+q^2$. Делая подстановку, получаем:

0,9604	+	0,0392	+	0,0004	=1
вероятность того, что у ребенка рецессивных генов нет		вероятность того, что у ребенка один рецессивный ген		вероятность того, что у ребенка два рецессивных гена	

[1] *Бедный М. С.* Демографические процессы и прогнозы здоровья населения, М., 1972, с. 134.

Таким образом, при данных условиях из 10 000 рожденных только 4 ребенка (или 1 из 2500) могут получить два рецессивных гена и стать жертвой болезни. Если же вступление в брак не случайный процесс, а браки совершаются между родственниками (например, кузенами, т. е. лицами, имеющими общих дедушку и бабушку), то вероятность заболевания возрастает. Расчеты показывают, что в тех случаях, когда один участник брака между кузенами является носителем рецессивного гена, то имеется один шанс из восьми, что другой участник этого брака также является носителем такого гена. Следовательно, вероятность того, что оба члена брачной пары обладают рецессивным геном, равняется $\frac{1}{8}$. В итоге возможность заболевания увеличивается. Действительно, в тех случаях, когда браки совершаются между близкими родственниками (закрытые общины), наблюдается высокая заболеваемость.

Рассмотрим еще один пример использования закона биномиального распределения вероятностей (формула Бернулли) и локальной формулы Лапласа.

На дневном отделении факультета статистики в Московском экономико-статистическом институте обучается 730 студентов. Найдем наиболее вероятное число студентов, родившихся 1 января, а также вероятность того, что у трех студентов день рождения в один и тот же день. Имеем: $n=730$; $p=\frac{1}{365}$; $q=\frac{364}{365}$. Наиболее вероятное число студентов, родившихся в любой день года, в том числе и 1 января, составляет $m_0 \approx np = 730 \cdot \frac{1}{365} = 2$. Вероятность этого исхода равна 0,28.

Вероятность того, что у трех студентов один и тот же день рождения по формуле биномиального закона распределения равна:

$$P_{730(3)} = C_{730}^{3} \left(\frac{1}{365}\right)^{3} \cdot \left(\frac{364}{365}\right)^{727}.$$

Рассчитать эту величину довольно трудно. Используем локальную формулу Лапласа и получаем:

$$m=3;\ m_0=2;\ \sqrt{npq} = \sqrt{730 \cdot \frac{1}{365} \cdot \frac{364}{365}} \approx 1{,}41;$$

$$t = \frac{m-m_0}{\sqrt{npq}} = \frac{1}{1{,}41} \approx 0{,}71.$$

Тогда

$$P_{730\ (3)} = \frac{\varphi\ (t)}{\sqrt{npq}} = \frac{\varphi\ (0{,}71)}{1{,}41} = \frac{0{,}3101}{1{,}41} \approx 0{,}22.$$

Обобщением биномиального распределения для случая, когда в ходе испытаний может наступать одно из нескольких взаимо-

исключаемых событий $A_1, A_2, \ldots, A_s$ с соответствующими вероятностями $p_1, p_2, \ldots, p_s$ $(\Sigma p=1)$, является полиномиальное распределение. Вероятность того, что указанные события произойдут $m_1, m_2, \ldots, m_s$ раз $(\Sigma m=n)$, определяется формулой полиномиального распределения:

$$P(m_1, m_2, \ldots, m_s) = \frac{n!}{m_1! m_2! \ldots m_s!} \cdot p_1^{m_1} p_2^{m_2} \ldots p_s^{m_s}.$$

Приведем пример полиномиального распределения. Известно, что в изучаемой совокупности 25% семей имеет троих и более детей, 35% семей имеет двоих детей, 20% семей имеют одного ребенка, 20% семей не имеют детей. Отбирается 50 семей. Какова вероятность того, что все отобранные семьи будут иметь троих детей и более?

$$P(m_1=50;\ m_2=0;\ m_3=0;\ m_4=0) = \frac{50!}{50!0!0!} \cdot 0{,}25^{50} \times$$

$$\times 0{,}35^{0} \cdot 0{,}20^{0} \cdot 0{,}20^{0} = 0{,}25^{50} \approx 0.$$

Практически такой исход мало вероятен.

2.3. Нормальное распределение

Очень часто в качестве теоретической формулы распределения, описывающей демографические явления, используют формулу нормального распределения признаков[1]. Еще в 1773 г. А. Муавр открыл закон распределения вероятностей, названный законом нормального распределения и используемый вначале для анализа азартных игр. А. Кетле использовал нормальное распределение для изучения распределения людей по росту. Ф. Гальтон привлек нормальную кривую для статистического изучения законов, управляющих наследственностью, К. Пирсон рассматривал нормальную кривую как основу биометрических измерений и построений и т. д. Разработка относящихся к этому закону вопросов в XIX в. связана с именами К. Гаусса и П. Лапласа. Общие условия возникновения закона нормального распределения установил русский математик А. М. Ляпунов. В соответствии с принятыми обозначениями формула А. М. Ляпунова имеет следующий вид:

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{t^2}{2}}, \quad \text{где} \quad t = \frac{x-\bar{x}}{\sigma}.$$

Центр группирования частот и форма нормальной кривой определяются $\bar{x}$ и σ. Чем больше $\bar{x}$, т. е. средняя, тем правее по оси абсцисс находится центр нормального распределения. При малых среднеквадратических отклонениях σ кривая нормального

[1] Термины «нормальное распределение признака», «нормальная кривая», «кривая нормального распределения», «нормальный вариационный ряд» общеприняты в специальной литературе и используются в качестве синонимов.

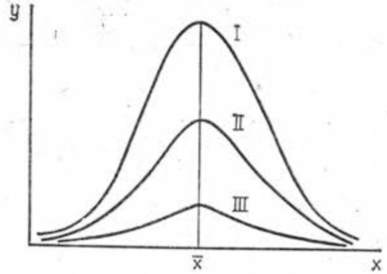

Рис. 2.2. Кривая нормального распределения при одинаковых $\bar{x}$ и разных σ

распределения вытянута и сжата с боков. При увеличении σ кривая растягивается вдоль оси абсцисс. Графическое изображение кривой нормального распределения приведено на рис. 2.2. У трех кривых, изображенных на этом рисунке, кривая *I* соответствует самому большому, а кривая *III* самому малому значению σ.

Закон нормального распределения, являющийся весьма распространенным, при большом числе наблюдений представляет собой предельный закон распределения. Нормальное распределение признака наблюдается в тех случаях, когда на величину вариантов, входящих в состав совокупности, действует множество случайных независимых или слабо зависимых факторов, каждый из которых играет в общей сумме примерно одинаковую роль (отсутствуют доминирующие факторы). Когда хотят сказать, что фактическое распределение признака более или менее близко к нормальному, говорят, что оно *подчинено закону нормального распределения.* Нормальное распределение может быть рассмотрено как конкретное выражение закономерности массового явления, которое возникает при взаимодействии множества причин. Нарушение нормального характера распределения признака чаще всего является свидетельством неоднородности совокупности. В математической статистике нормальное распределение играет роль некоторого стандарта, с которым сравнивают другие распределения.

Многие физические признаки человека, обладающие вариационной изменчивостью, с точки зрения теории вероятностей рассматриваются в качестве случайных величин, имеющих нормальные распределения. Советские исследователи М. В. Игнатьев и А. В. Пугачева на основании собственных экспериментов, а также статистических данных, полученных К. Пирсоном и П. Махалонобисом, утверждают, что распределение антропологических признаков в основной массе близко к нормальному распределению.

Существует несколько способов построения нормального распределения на основании фактических данных. Приведем, на наш взгляд, самый удобный. Все действия и расчеты продемонстрируем на примере данных, полученных при выборочном наблюдении роста 500 взрослых мужчин (табл. 2.3).

Задан фактический вариационный ряд с равными интервалами. Аппроксимация частот этого вариационного ряда частотами нормальной кривой производится по формуле

$$y=\frac{K\Sigma m}{\sigma}\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{{}^{2}}{2}}$$

Таблица 2.3. **Распределение 500 взрослых мужчин по росту**

Рост (центр интервала), см x	Число мужчин, m	xm	$x-\tilde{x}$	$t=\frac{x-\tilde{x}}{\sigma}$	$\varphi(t)$	$m'=\varphi(t)\cdot\frac{K\Sigma m}{\sigma}$	Частости, % $w=\frac{m_i}{\Sigma m_i'}\cdot 100$	Накопленные частости
150,5	2	301,0	—18	—3,08	0,004	1	0,2	0,2
153,5	3	460,5	—15	—2,56	0,015	4	0,8	1,0
156,5	10	1565,0	—12	—2,05	0,049	13	2,6	3,6
159,5	31	4944,5	—9	—1,54	0,122	31	6,2	9,8
162,5	72	11700,0	—6	—1,02	0,237	61	12,2	22,0
165,5	85	14067,5	—3	—0,51	0,350	90	18,0	40,0
168,5	94	15839,0	—0	0	0,399	100	20,0	60,0
171,5	88	15092,0	3	0,51	0,350	90	18,0	78,0
174,5	62	10819,0	6	1,02	0,237	61	12,2	90,2
177,5	37	6567,5	9	1,54	0,122	31	6,2	96,4
180,5	12	2166,0	12	2,05	0,049	13	2,6	99,0
183,5	3	550,5	15	2,56	0,015	4	0,8	99,8
186,5	1	186,5	18	3,08	0,004	1	0,2	100,0
Σ	500	84259,0	—	—	—	500	—	—

где K — величина интервала; Σm — сумма частот вариационного ряда, т. е. n; σ — среднее квадратическое отклонение; t — центрированное и нормированное отклонение, равное $\frac{x-\tilde{x}}{\sigma}$. Величина $\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}=\varphi(t)$ может быть найдена по соответствующим таблицам. Находим выборочную среднюю

$$\tilde{x}=\frac{\Sigma xm}{\Sigma m}=\frac{84259,0}{500}=168,5$$

и выборочную дисперсию $\sigma^2=34,218$; $\sigma=5,85$. Определяем для каждой строки в табл. 2.3 отношение отклонений центров интервалов от средней к среднему квадратическому отклонению $t=\frac{x-\tilde{x}}{\sigma}$. Используя соответственную таблицу, определяем плотность стандартизованного нормального распределения $\varphi(t)=\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$. Далее находим теоретические частоты нормального распределения $m'=\varphi(t)\cdot\frac{K\Sigma m}{\sigma}$. Для этого вычисляем $\frac{K\Sigma m}{\sigma}=\frac{3\cdot 500}{5,85}=\frac{1\,500}{5,85}\approx 256,4$ и получаем для первой строки $0,004\times\times 256,4=1,0256\approx 1$ и т. д. Установив, что рост мужчин хорошо аппроксимируется законом нормального распределения, можно

5 Заказ 2715

определить, например, удельный вес мужчин, рост которых находится в границах от 165,5 до 180,5 см:

$$\frac{90+100+90+61+31+13}{500}=0{,}77, \text{ или } 77\%.$$

Нормальное распределение признака дает возможность решения многих задач. Весьма близкое к нормальному распределению образует соотношение полов среди занятых в совокупности малых и средних городов РСФСР[1] (рис. 2.3). Так, доля женщин среди занятых, мода и медиана, соответственно равные 51,0; 51,2; 51,3, очень близки друг к другу.

Особенно важно, что при вариации признака, вызванной только случайными факторами, кривая должна соответствовать нормальному распределению. Если же в совокупности случайно отобранных единиц все же обнаруживается отклонение от нормального распределения, то это указывает на действие факторов не чисто случайных (по современной терминологии «шумы в системе»), а существенных. Возникает задача определить причины наблюдаемых отклонений от «нормальности» распределений.

В качестве примера того, что отклонение фактического распределения от нормального позволяет делать определенные выводы в отношении однородности изучаемой совокупности, приведем данные Ю. Э. Янсона. Используя статистические материалы Ж. Бертильона о росте французских новобранцев и о специфичности вида кривой для каждой национальности, Ю. Э. Янсон построил соответствующие кривые распределения для всей Франции и для одного департамента Дубс. Сравнив их с нормальной кривой, он писал: «Для целой Франции такие частные влияния теряются в общей массе, как причины случайные, и для нее получается кривая совершенно правильная»[2]. Что же касается кривой распределения роста новобранцев для департамента Дубс, то на ней явственно различаются две вершины (моды). Это свидетельствовало, по мнению Ю. Э. Янсона, о неоднородности совокупности призывников, т. е. о соединении новобранцев двух национальностей, существенно различающихся по росту (кельтов и бургунов), в одну совокупность. Такое смешение привело

Рис. 2.3. Распределение малых и средних городов РСФСР по соотношению полов среди занятых

[1] См.: Занятость в небольших городах/Под ред. А. Э. Котляра. М., 1978, с. 82.

[2] *Янсон Ю.* Теория статистики. Спб., 1891, с. 504.

к двумодальности. Надо полагать, что помимо двумодальности о разнородности совокупности населения могло бы также свидетельствовать расширение вершины кривой распределения. При этом при группировке населения по росту с бóльшими интервалами неоднородность как бы смазывалась (вершина кривой расширялась). Достаточно перегруппировать данные о росте населения с меньшими интервалами, как две кульминационные точки кривой (моды) обозначаются весьма ясно, свидетельствуя о неоднородности населения в отношении данного признака.

Интересный факт из работы А. Кетле по анализу отклонений фактических данных от теоретических привел советский статистик Б. И. Карпенко[1]. А. Кетле вычислил и сопоставил теоретическое и эмпирическое распределения 100 000 французских призывников по росту. Анализ отклонений показал, что не все они случайны: призывников с ростом ниже 1,570 м оказалось излишне много, тогда как группа имеющих рост в интервале 1,570 м — 1,624 м малочисленна. На основании этих данных ученый сделал вывод, что при измерении роста допускалась фальсификация с намерением побольше призывников освободить от военной службы по «недостаточности роста».

Известно, что статистические данные о числе долгожителей недостаточно достоверны. М. С. Бедный использовал теорию вероятностей для уточнения данных о совокупности умерших в предельно пожилых возрастах. Привлекая нормальное распределение, он показал, что интенсивность смертности после 90-летнего возраста возрастает более быстро, чем об этом свидетельствуют эмпирические данные.

2.4. Логарифмически-нормальное распределение

При практическом использовании нормального распределения для аппроксимации фактических данных мы часто встречаемся со случаями, когда эмпирическое распределение более или менее отличается от нормального и обнаруживает заметную асимметрию. Дело в том, что закон нормального распределения не обладает той универсальностью, которую ему раньше приписывали. Уже давно было замечено, что кривые, изображающие статистические распределения демографических явлений, не всегда симметричны относительно перпендикуляра из моды на основание. Первым обратил на это внимание Фехнер. Он предложил интересный метод, позволяющий аппроксимировать кривую фактического распределения, названную им двусторонней, путем подбора кривой нормального распределения отдельно для каждой части.

[1] См.: *Карпенко Б. И.* Случайность и необходимость и их взаимосвязи в статистических исследованиях. — В кн.: Проблемы теории статистики. М., 1978, с. 62.

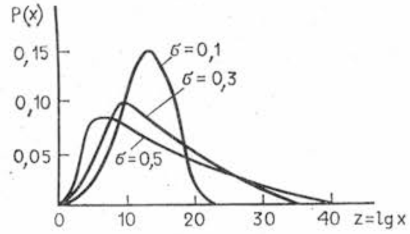

Рис. 2.4. Логарифмически-нормальные распределения

Логарифмически - нормальные распределения представляют собой распределение случайной величины *x*, логарифм которой ($t=\lg x$) подчинен закону нормального распределения.

Логнормальное распределение образуется в результате умножения большого числа независимых или слабо зависимых неотрицательных случайных величин, дисперсия каждой из которых мала по сравнению с дисперсией результата. Таким образом, в основе логарифмически-нормального распределения лежит мультипликативный (умножающий) процесс формирования случайных величин, т. е. такой, в котором действие каждого добавочного фактора на случайную величину пропорционально ее достигнутому уровню. Кривая, изображающая такие распределения, круто поднимается слева и полого спускается справа. При переходе к логарифмам исходных показателей подъем растягивается, а спуск сжимается, и все распределение переходит в нормальное (рис. 2.4).

Логнормальное распределение определяется двумя параметрами: средней M и среднеквадратическим отклонением логарифмов σ_M. Функция плотности логнормального распределения $f(t)$ имеет вид:

$$f(t)=\frac{M}{\sigma_M t\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\lg t-M}{\sigma_M}\right)^2}.$$

При малой дисперсии логнормальное распределение близко к нормальному.

Опыт показывает, что логарифмически-нормальному распределению подчиняются заработная плата рабочих, численность мальчиков и девочек по возрасту отцов и матерей и другие показатели. Приведем некоторые примеры.

При изучении интенсивности абортов В. К. Кузнецов установил два фактора, влияющих на их численность: материальные трудности и неудовлетворительные жилищные условия и как следствие этих факторов, уровень напряженности отношений в семье между супругами[1]. Автор обозначил F и f соответственно наличие и отсутствие жалоб на материальный недостаток, а D и d — на жилищные условия.

[1] См.: *Кузнецов В. К.* О соотношении субъективных и объективных оценок условий жизни как факторов, влияющих на численность абортов. — В кн.: Социально-гигиенические исследования. Труды второго медицинского института им. Н. И. Пирогова. М., 1973, т. XIX, вып. 3, с. 246.

На основе обработки 201 анонимной анкеты были построены упорядоченные вариационные ряды и различные распределения: распределение душевого дохода $F(F)$, распределение жилой площади в квадратных метрах на человека $F(D)$, распределение накопленных вероятностей того, что женщины не выразят жалобы на материальный недостаток при соответствующих величинах душевого дохода $P(f)$, и распределение того, что женщины не выразят жалобы на неудовлетворительные жилищные условия при определенной площади, приходящейся на одного человека $P(d)$. Автор показал, что все четыре распределения подчиняются закону логарифмически-нормального распределения, и проверил близость фактических распределений к теоретическим с помощью критерия Колмогорова. На основании этого можно считать доказанным, что при разработке средств демографической политики, в частности стимулирования рождаемости, *должна быть учтена инерциальность уменьшения жалоб женщин* на условия жизни: при улучшении условий жизни в геометрической прогрессии число жалоб женщин будет уменьшаться в арифметической прогрессии.

Г. Крамер оценивал гипотезы о «нормальности» различных распределений. Покажем его выводы в отношении распределения 447,6 тыс. мальчиков по возрасту их отцов[1] (табл. 2.4 и рис. 2.5).

Таблица 2.4. **Распределение мальчиков по возрасту отцов**

Возраст отцов, лет, x	Число мальчиков, тыс. m	xm	$x-\tilde{x}$	$(x-\tilde{x})$	$(x-\tilde{x})^2m$
17,5	2,2	38,50	—18,2	331,24	728,728
22,5	36,1	812,25	—13,2	174,24	6290,064
27,5	94,3	2593,25	—8,2	67,24	6340,732
32,5	112,7	3662,75	—3,2	10,24	1154,048
37,5	96,0	3600,00	1,8	3,24	311,040
42,5	69,7	2962,25	6,8	46,24	3222,928
47,5	38,9	1847,75	11,8	139,24	5416,436
52,5	17,2	903,00	16,8	282,24	4854,528
57,5	6,5	373,75	21,8	475,24	3089,060
62,5	2,6	162,50	26,8	718,24	1867,424
67,5	1,0	67,50	31,8	1011,24	1011,240
72,5	0,4	29,00	36,8	1354,24	541,696
Σ	477,6	17052,50	—	—	34827,924

Получены следующие характеристики:

$$\tilde{x}=\frac{\Sigma xm}{\Sigma m}=\frac{17052{,}50}{477{,}6}\approx 35{,}7 \text{ года};$$

[1] См.: *Крамер Г.* Математические методы статистики. М., 1975, с. 497.

$$\mu_2=\sigma^2=\frac{\Sigma(x-\tilde{x})^2 m}{\Sigma m}=\frac{34827,924}{477,6}\approx 72,9228\ ;$$

$$\sigma=\sqrt{\mu_2}=8,54.$$

Показатели асимметрии: 3-й нормированный момент (r_3), равный отношению третьего центрального момента (μ_3) к среднеквадратическому отклонению в кубе (σ^3), т. е. $r_3=\frac{\mu_3}{\sigma^3}$ значительно отличается от нуля, а 4-й нормированный момент (r_4), равный отношению четвертого центрального момента (μ_4) к среднеквадратическому отклонению в четвертой степени (σ^4), т. е. $r_4=\frac{\mu_4}{\sigma_4}$, значительно отличается от трех, что свидетельствует об отклонении фактических данных от нормального распределения. Дополнительные исследования показали, как пишет автор, близость этого распределения к логарифмически-нормальному.

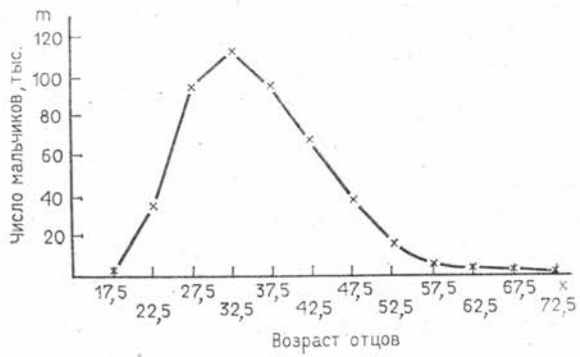

Рис. 2.5. Распределение мальчиков по возрасту их отцов

2.5. Распределение Пуассона и Полиа

Пуассон сделал сообщения о выведенной им формуле в 1835 г. Закон распределения Пуассона, по предложению В. Борткевича, начал использоваться при изучении распределений таких редко встречающихся явлений, как смерть от удара копытом лошади, рождения двух или трех близнецов при изучении смертности от крайне редких заболеваний и т. д. При этом имелось в виду, что даже очень редкие явления в силу своей случайности наступают в соответствии с расчетами, сделанными на основе теории вероятности. Если в биномиальном распределении вероятность событий p (или q) очень мала, но число испытаний n достаточно велико, так что произведение $np=a$ остается конечным, то биномиальное распределение приближается к распределению, называемому распределением Пуассона:

$$P_m=\lim_{n\to\infty} C_n^m\ p^m q^{n-m}=\frac{a^m e^{-a}}{m!},$$

где m — частота данного события.

При использовании фактических данных для определения теоретических ординат распределения Пуассона (m') нужно вместо наивероятнейшего исхода $np=a$ подставить значение математического ожидания, т. е. средней величины признака $E(x)=\overline{x}$, а вместо частоты события m — значение вариантов, т. е. x. Тогда получим формулу

$$m'=n\frac{\overline{x}^x e^{-\overline{x}}}{x!}.$$

Математическое ожидание случайной величины x, т. е. средняя, ее дисперсия и третий центральный момент распределения Пуассона равны np и x, т. е. ($\overline{x}=\sigma^2=\mu_3$). Это свойство распределения Пуассона используют на практике. Если значения средней величины признака в конкретном распределении и дисперсия близки друг к другу, то это является основанием гипотезы о том, что налицо распределение Пуассона.

Особенность распределения Пуассона заключается в том, что для его построения по фактическим данным требуется только знание средней величины. Значения $\frac{\overline{x}^x e^{-\overline{x}}}{x!}$ табулированы при различных значениях $\overline{x}$ и x.

Прежде чем применять различные законы распределения при изучении какого-либо конкретного процесса, например рождаемости, важно выяснить теоретическую безупречность приложения к этому процессу вероятностных методов. Чтобы демографический процесс можно было рассматривать как пуассоновский, необходимо выполнение трех условий: ординарности, стационарности, отсутствия последействия.

Условие ординарности заключается в практической невозможности появления двух или более событий за малый интервал времени (это означает, что вероятность появления двух и более событий должна быть пренебрежительно мала по сравнению с вероятностью появления одного события).

Условие стационарности означает, что вероятность наступления определенного числа событий на протяжении какого-либо интервала времени должна зависеть только от числа событий и длительности интервала, а не от сдвига времени на определенную величину.

Условие отсутствия последействия состоит в том, что вероятность событий не должна зависеть от того, сколько раз это событие происходило раньше.

Рассмотрим процесс деторождения с точки зрения соблюдения указанных условий. Условию ординарности процесс деторождения полностью не отвечает, так как имеются, правда, довольно редкие, случаи многоплодных родов. Если рассматривать не число рождений, а число родов, то условие ординарности будет соблюдено. Условию стационарности процесс деторождения также полностью не удовлетворяет, так как плодовитость женщин до некоторой степени зависит от ее возраста и продолжительности брака. Условию отсутствия последействия процесс деторождений также не отвечает полностью, так как вероятность рождения последующих детей во многом зависит от числа детей, рожденных женщиной ранее.

Следует иметь в виду, что, хотя условия теоретической безупречности приложения распределения Пуассона во многих практических вопросах демографии либо соблюдены нестрого, либо частично не соблюдены, практическое применение показывает хорошую аппроксимацию фактических данных.

Исследования в области демографии показывают, что многие явления, хотя они и не являются абсолютно случайными и независимыми, хорошо описываются такими вероятностными распределениями, как биномиальное, Пуассона, нормальное, логарифмически-нормальное и др. В качестве примера можно привести исследования В. А. Мозгляковой, показавшей достаточно хорошую аппроксимацию данных этими распределениями при изучении здоровья населения[1].

Приведем пример. По фактическим данным о числе детей, рожденных 1000 женщинами, построим распределение Пуассона (табл. 2.5).

Таблица 2.5. **Распределение 1000 женщин по числу рожденных детей**

Число рожденных детей, x	Число женщин, m	xm	$x-\overline{x}$	$(x-\overline{x})^2$	$(x-\overline{x})^2 m$	m'
0	232	0	—1,5	2,25	522,00	223
1	313	313	—0,5	0,25	78,25	335
2	260	520	0,5	0,25	65,00	251
3	130	390	1,5	2,25	292,50	125
4	52	208	2,5	6,25	325,00	47
5	10	50	3,5	12,25	122,50	14
6	2	12	4,5	20,25	40,50	4
7	1	7	5,5	30,25	30,25	1
Σ	1 000	1 500	—	—	1476,00	1 000

Находим, что $\overline{x}=\frac{\Sigma\, xm}{\Sigma\, m}=\frac{1\,500}{1\,000}=1{,}5$; $\sigma^2=\frac{\Sigma\,(x-\overline{x})}{\Sigma\, m}=1{,}476$. При $\overline{x}=1{,}5$ и $x=0$ получаем: $\frac{\overline{x}^x\, e^{-\overline{x}}}{0\,!}=e^{-1{,}5}=0{,}22313$. Умножая на

[1] См.: *Мозглякова В. А.* О применении теории вероятностей при изучении заболеваемости. — В кн.: Вопросы санитарной и медицинской статистики. М., 1971, с. 132.

объем совокупности $n=1000$, получаем частоту: $m'=1000\cdot 0{,}22313\approx\approx 223$. Аналогично, используя соответствующие математико-статистические таблицы[1], получаем все значения m'. Покажем оба распределения (фактическое и теоретическое) на рис. 2.6.

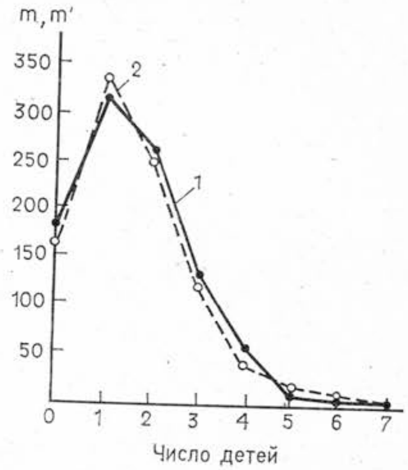

Рис. 2.6. Распределение женщин по числу рожденных детей:

1 — фактическое; *2* — теоретическое, построенное по формуле Пуассона

Д. Пойя при доказательстве возможности использования теории вероятностей для описания массовых явлений привел пример из демографии[2]. Конкретно речь идет о таких редких событиях, как рождения троен. В Швейцарии за 30-летний период с 1871 по 1900 г. было зарегистрировано 2 612 246 рождений, в том числе 300 случаев рождения тройни:

$$\text{Общее среднегодовое число рождений}=\frac{2\,612\,246}{30}=87\,075;$$

$$\text{Среднегодовое число рождений тройни}=\frac{300}{30}=10.$$

Погодовые числа рождений троен приведены в табл. 2.6.

Автор замечает, что фактические частоты кажутся разбросанными случайным образом, в то время как теоретические частоты, полученные, как пишет автор, по простому закону (очевидно, закону Пуассона), дают неплохое согласование. Чтобы сделать это согласование более очевидным, автор прибегает к приему накопления частот, позволяющему еще лучше судить о согласованности фактических частот с теоретическими. Полученные результаты Д. Пойя называет блестящими.

В демографии закону Пуассона подчиняются распределение городов СССР с населением свыше 50 тыс. (при исключении таких крупных городов, как Москва, Ленинград, Киев, Ташкент и Минск), распределение вступивших в брак мужчин по возрасту, распределение семей по числу членов, число рождений в год на 1000 женщин, число смертей на 1000 холостяков и т. д.

В ряде случаев, когда процессы развиваются по принципу цепной реакции (например, число заболеваний при эпидемии), используют распределение Полиа. До Полиа эту проблему изу-

[1] См.: *Большев Л. Н., Смирнов Н. В.* Таблицы математической статистики. М., 1968, с. 360.

[2] См.: *Пойя Д.* Математические и правдоподобные рассуждения. Пер. с англ. М., 1975, с. 297.

Таблица 2.6. **Тройни, родившиеся в Швейцарии в 1871—1900 гг.**

Число рождений тройни	В течение какого числа лет наблюдалось данное число случаев рождения тройни, m	Теоретические частоты, m'	Накопленные частоты	
			V_m	$V_{m'}$
0	0	0,00	0	0,00
1	0	0,00	0	0,00
2	0	0,09	0	0,09
3	1	0,21	1	0,30
4	0	0,57	1	0,87
5	1	1,14	2	2,01
6	1	1,89	3	3,90
7	5	2,70	8	6,60
8	1	3,39	9	9,99
9	4	3,75	13	13,74
10	4	3,75	17	17,49
11	4	3,42	21	20,91
12	3	2,85	24	23,76
13	2	2,16	26	25,92
14	1	1,59	27	27,51
15	2	1,02	29	28,53
16	0	0,66	29	29,19
17	1	0,39	30	29,58
18	0	0,21	30	29,79
19	0	0,22	30	29,91
Σ	30	29,91	—	—

чал А. А. Марков. В качестве примера рассмотрим сопоставление фактического распределения 256 месяцев по числу случаев смерти от оспы с распределением Полиа[1] (табл. 2.7).

Таблица 2.7. **Смертность от оспы в Швейцарии в 1877—1900 гг.**

Число случаев смерти в течение месяца	Фактическое число месяцев	Число месяцев по распределению Полиа	Число случаев смерти в течение месяца	Фактическое число месяцев	Число месяцев по распределению Полиа
0	100	100	8	5	7
1	39	36	9	6	6
2	28	24	10	1	5
3	16	17	11	6	5
4	13	14	12	2	4
5	16	11	13	2	4
6	11	9	14	3	3
7	5	8	15	3	3
			Σ	256	256

[1] См.: *Павловский З.* Введение в математическую статистику. Пер. с польск. М., 1967, с. 99.

2.6. Распределение полов среди родившихся и соответствующие вероятности

В демографии различают соотношения численности полов: при оплодотворении — первичное, при рождении — вторичное и в репродуктивном возрасте — третичное.

Соотношение полов среди родившихся, т. е. вторичное соотношение, основано на законах природы. Однако, хотя рождение мальчика или девочки биологически обусловлено, осуществление этой обусловленности происходит в рамках вероятностных законов, реализуемых при массовых наблюдениях.

Устойчивое соотношение полов среди родившихся фиксировалось еще в древнем Китае за 2238 лет до н. э. Соотношение мальчиков и девочек среди родившихся было признано равным.

Долгое время существовало мнение, что рождение мальчика или девочки в каждом отдельном случае есть события равновозможные, аналогичные событиям, происходящим в опыте с бросанием монеты, когда с равными вероятностями выпадает «орел» (рождается мальчик) и «решка» (рождается девочка). На самом деле, как показывают многолетние наблюдения во многих странах, вероятности рождения мальчика и девочки неодинаковы: вероятность рождения мальчика ($p_м$) составляет 0,516, а вероятность рождения девочки ($p_д$) равна 0,484. Так, еще член Лондонского королевского общества Д. Арбутнот (до И. П. Зюссмильха) установил, что с 1629 по 1710 г. (т. е. за 82 года) в Лондоне мальчиков в каждом году рождалось больше, чем девочек, и доказывал на основе теории вероятностей, что это не может быть простой случайностью, так как вероятность случайного превышения рождений мальчиков над девочками равна отношению 1 к 25-значному числу, т. е. чрезвычайно мала. В этом Д. Арбутнот усматривал божеское предвидение, т. е. здесь, по его мнению, видна забота «всевышнего» о продолжении человеческого рода.

Изучая статистические данные за 40 лет (1745—1784 гг.) по Парижу, П. Лаплас получил соотношение числа мальчиков и девочек среди родившихся не 22 : 21 = 1,048, как это наблюдалось для многих народов и Франции в целом, а меньшее, а именно 25 : 24 = 1,042. Для объяснения этого различия П. Лаплас привлек архивные материалы и выяснил, что в общее число рождений по Парижу включались подкидыши (окрестное население подкидывало в парижский приют главным образом девочек). Исключив из расчета подкидышей, П. Лаплас получил и для Парижа устойчивое соотношение мальчиков и девочек по всем годам, равное 22 : 21.

Классическая задача о соотношении рождений мальчиков и девочек привлекла внимание Н. Бернулли. Используя биномиальное распределение с параметром $\frac{m}{f}$ для вероятности новорожденному оказаться мальчиком, Н. Бернулли нашел вероятность того, что ежегодное число рождений мальчиков μ будет находиться в заданных границах:

$$P\{|\mu - rm| < l\} \approx \frac{t-1}{t},$$

где $r = \frac{n}{m+f}$; n — общее число рождений;

$$t = [1 - \frac{l(m+f)^2}{m+tn}] = e^{\frac{l^2(m+f)^2}{2\frac{m}{n}}}.$$

При решении этой задачи была проверена статистическая гипотеза о значении параметра биномиального распределения. Для этого было найдено математическое ожидание (M) числа мальчиков из общего числа рождений ($2M$) при относительной частоте рождений мальчиков и девочек, равной $a:b$. Путем вычисления максимального члена разложения $(a+b)^2$ было получено $M = \frac{2Ma}{a+b}$. Используя приращение вероятностей, автор вывел дифференциальное уравнение, решение которого позволило ему сформулировать интересные выводы.

Следует иметь в виду, что вероятности, определяющие пол новорожденного, не самоочевидны. Имеются фамильные предания, в которых фигурируют склонности в данной семье к преобладанию среди родившихся детей определенного пола. Это обстоятельство подтверждают статистические материалы, указывающие на некоторую связь индивидуальных особенностей родителей с полом потомства. Вот почему для получения однородных совокупностей должен быть использован основной статистический прием — группировка семей по признакам родителей. Это означает, что частоту рождения мальчиков и девочек нужно определять у родителей, имеющих как можно больше общих признаков: национальных, профессиональных, возрастных, условий питания, климата, имеющих одинаковое число детей, родителей, у которых прошло одинаковое время после рождения последнего ребенка, и т. д. Однако здесь нас подстерегает другая опасность — уменьшение числа наблюдаемых событий и усиление в связи с этим влияния случая. Если под наблюдение попала семья с семью сыновьями, то на основании того, что вероятность такого события представляется очень малой $\left(\frac{1}{2^7} = \frac{1}{128}\right)$, нельзя сделать вывод о тенденции у наблюдаемых родителей к преимущественному рождению мальчиков. Для правильного вывода о наличии такой тенденции необходимо привлечь большую совокупность

семей и установить, что в ней семьи с 7 мальчиками более многочисленны, чем это могло произойти под влиянием случая.

Разумеется, объяснение того, почему имеет место перевес мужских рождений, — не задача теории вероятностей. Объяснение этого факта должны дать генетики, врачи, гигиенисты. Современные исследования показали, что сразу после зачатия соотношение оплодотворенных яйцеклеток, несущих мужской и женский плоды (первичное соотношение полов), составляет 117 : 100. Вследствие меньшей жизнеспособности мужских зародышей, что проявляется в повышенной внутриутробной смертности мальчиков, а также в повышенной доле мальчиков среди мертворожденных, на 100 рожденных девочек приходится уже 105—106 мальчиков.

П. Лаплас занимался вопросом о влиянии климата на соотношение рождаемости мальчиков и девочек. Рассматривая это соотношение в Лондоне и Париже, он пришел к выводу о существенности расхождений.

А. А. Чупров предложил несколько гипотез о характере факторов, определяющих пол ребенка. В частности, рассматривая гипотезу о независимости пола ребенка от индивидуальных особенностей родителей, он привлек статистические материалы о многоплодных родах. Если гипотеза о независимости справедлива, то вероятности комбинации мальчиков и девочек в многоплодных родах должны находиться в соответствии с теоремой умножения вероятностей для независимых событий. Тогда вероятность рождения двух мальчиков должна была быть равной: $0{,}51 \cdot 0{,}51 = 0{,}26$; вероятность рождения двух девочек $0{,}49 \cdot 0{,}49 = 0{,}24$ и вероятность рождения одного мальчика и одной девочки $2 \cdot 0{,}51 \cdot 0{,}49 = 0{,}50$. Следовательно, вероятности рождения однополых и разнополых детей должны быть равными и в соответствии с законом больших чисел рождения однополых и разнополых детей должны встречаться одинаково часто.

Между тем статистические данные, на основе которых А. А. Чупров делает свои выводы, показывали, что фактическое распределение пола детей среди двоен и троен таково, что случаи однополых близнецов среди них более часты, чем дает вероятностный расчет. Именно поэтому А. А. Чупров отверг эту гипотезу. Дополнительные исследования потомства брачных пар с двумя детьми и более показали, что вероятности различных комбинаций пола детей в семьях не совпадают с фактическими частотами этих комбинаций: наблюдается перевес частот для однополых детей. При этом замечено, что в семьях, где первые дети — мальчики, удельный вес мальчиков среди последующих детей выше среднего, в тех семьях, где первые — девочки, удельный вес мальчиков среди последующих детей ниже среднего. Это подтверждает выводы о том, что среди факторов, действующих на пол ребенка, есть и такие, которые зависят от индивидуальных особенностей родителей.

Статистические данные позволяют предположить, что в годы, когда убыль мужского населения особенно значительна (например, во время войн), доля мальчиков среди родившихся увеличивается. Если принять удельный вес мужчин во всем населении за x, а долю мальчиков среди родившихся за y, а затем вычислить коэффициент корреляции между ними, то он укажет на связь между этими двумя факторами. Например, для населения Севера европейской части РСФСР по данным за период с 1897 по 1926 г. коэффициент корреляции составлял: $r_{x/y}=0{,}5\pm0{,}1$, т. е. превышал вероятную ошибку в 5 раз и заслуживал внимания.

В чем же суть процесса повышения доли мальчиков? Пример великолепного анализа факторов, влияющих на долю мальчиков среди рождающихся, мы находим у А. А. Чупрова. Сначала он на основании большого статистического материала устанавливает тенденции: с увеличением возраста зародыша внутриутробная смертность непрерывно убывает, сначала очень быстро, а затем медленнее; смертность тех зародышей, которые рождаются преждевременно живыми или умирают в момент рождения, тоже убывает до 7-го месяца, после чего очень сильно возрастет. В общем эти тенденции складываются таким образом, что частость выкидышей с увеличением возраста зародышей постепенно снижается и достигает минимума примерно на 8-м месяце беременности, после чего она снова поднимается. Это происходит в силу того, что быстрое увеличение смертности преждевременно родившихся после 7-го месяца перекрывает уменьшение собственно внутриутробной смертности.

Рассмотрим механизм этого процесса математически. Введем обозначения: V_0 — число зачатых мальчиков на 100 зачатых девочек; V_1 — число мальчиков на 100 девочек среди зародышей, достигших грани жизнеспособности; $\overline{V}$ — среднее число мальчиков на 100 девочек среди выкидышей; a — общая доля выкидышей из числа достигших грани жизнеспособности. Тогда

$$V_0=\frac{\dfrac{V_1}{100+V_1}+a\cdot\dfrac{\overline{V}}{100+V}}{\dfrac{1}{100+V_1}+a\cdot\dfrac{1}{100+V}}.$$

Используем в качестве границ величины a значения от 20 до 50% и для V_1 примем соотношение 106 : 100, это даст возможность подсчитать по указанной формуле значения V_0 при различных значениях a и $\overline{V}$ (табл. 2.8).

Интересно отметить, что А. А. Чупров считал соотношение 125—130 к 100 ниже истинного. Последующие наблюдения подтвердили это мнение. Таким образом, можно считать доказанным зависимость доли мальчиков среди родившихся от процента выкидышей.

Таблица 2.8. **Расчет числа зачатых мальчиков на 100 девочек**

Доля выкидышей, %, a	V_0 при	
	$\overline{V}=200$	$\overline{V}=400$
20	117	128
25	120	134
33	123	141
50	130	156

Венгерские демографы пришли к выводу о том, что пропорции полов среди новорожденных зависят от возраста родителей. Так, до начала войны 1914 г. на 100 девочек приходилось 105 мальчиков, после первой мировой войны эта пропорция резко увеличилась, затем снижалась, а после второй мировой войны медленно увеличивается. К 1968 г. соотношение составило 108 : 100. Объясняется это обстоятельство тем, что произошло снижение среднего возраста родителей, а в более молодом возрасте родителей частота рождения мальчиков выше. Так как число рождений на одну женщину снижается, то увеличился удельный вес рождений от матерей более молодого возраста.

К диаметрально противоположному выводу пришел Г. Крамер, выдвинувший тезис о независимости соотношения мальчиков и девочек среди родившихся от возраста родителей.

Основываясь на том, что мальчиков рождается больше, чем девочек, и что расчет ожидаемого соотношения численности полов в населении может быть произведен с помощью традиционной урны с черными и белыми шарами (с соотношением 51 к 49), а также исходя из того, что в самой урновой схеме не находит отражение другой факт — превышение смертности мальчиков по сравнению с девочками, П. П. Маслов сделал неожиданный вывод о невозможности применения вероятностных моделей к таким явлениям общественной жизни, как рождаемость и смертность[1]. П. П. Маслов видел противоречие между превышением доли мальчиков среди родившихся, с одной стороны, и тем, что в условиях длительного мирного времени соотношение полов уравновешивается — с другой.

На наш взгляд, никакого противоречия здесь нет. П. П. Маслов не учел возможности теоретико-вероятностного моделирования половозрастной смертности с помощью таблиц смертности, а также одновременного моделирования двух демографических факторов: рождаемости и смертности. При использовании «чистых» вероятностей одновременно учитывается и рождаемость, и смертность.

Рассмотрим некоторые вероятностные задачи на соотношение полов среди родившихся и примерный ход их решения. Начнем с простой задачи, показывающей, как трудно порой в теоретических рассуждениях, без практического подтверждения, установить истинность тех или иных утверждений. *Нужно оценить вероятность того, что в семье с двумя детьми оба ребенка маль-*

[1] См.: *Маслов П. П.* Измерение потребительского спроса. М., 1971, с. 101.

чики. Допустим для простоты последующих рассуждений, что мальчики и девочки рождаются поровну и вероятность рождения мальчика равна вероятности рождения девочки (0,5).

На первый взгляд возможны три исхода, т. е. три типа семей: 1) с двумя мальчиками; 2) с двумя девочками; 3) с одним мальчиком и одной девочкой. Учитывая, что из трех приведенных типов семей нас интересует один — оба ребенка мальчики, искомая вероятность его равна $\frac{1}{3}$.

Однако предположение, что все три типа семей равновозможны, ошибочно. Практические исследования показывают, что частость третьего типа семей (с одним мальчиком и одной девочкой) бо́льшая, чем у двух других типов. Дело в том, что с учетом последовательности рождений число равновозможных типов семей не три, а четыре (пол старшего ребенка обозначен первым): 1) мальчик, мальчик; 2) девочка, довочка; 3) мальчик, девочка; 4) девочка, мальчик. При этом последние два типа семей (3-й и 4-й) соответствуют предположению, что в семье один мальчик и одна девочка. Вот почему искомая вероятность того, что в семье с двумя детьми оба ребенка мальчики, равна не $\frac{1}{3}$, а $\frac{1}{4}$. Статистические наблюдения показывают правильность именно этой величины.

Однако можно сформулировать похожую задачу таким образом, что именно три типа семей будут удовлетворять условию: *из семей с двумя детьми, в которых есть мальчик (событие B), наудачу отбирается одна семья. Какова вероятность того, что в ней окажутся два мальчика (событие A и B)?*

Мы уже видели, что с учетом последовательности рождений число возможных распределений детей по полу в семье с двумя детьми равно четырем: 1) мальчик, мальчик; 2) девочка, девочка; 3) мальчик, девочка; 4) девочка, мальчик. Но из указанных четырех равновозможных типов семей имеются три типа, отвечающие условию задачи о том, что в семье из двух детей уже имеется один мальчик: это типы семей 1, 2 и 4-й. Из этих трех исходов только один (1-й) благоприятствует второму условию о том, что и второй ребенок — мальчик. Следовательно, искомая вероятность равна $\frac{1}{3}$.

Эту задачу можно решить и по формуле. В данном случае требуется определить вероятность события B при условии, что событие A уже состоялось. Формула имеет вид:

$$P\,(A/B) = \frac{p\,(AB)}{p\,(B)}\,.$$

Из условия нашей задачи имеем $p(B) = \frac{3}{4}$. Следовательно, искомая вероятность равна:

$$P\,(A/B) = \frac{1}{4} : \frac{3}{4} = \frac{1}{3}\,.$$

Если при соблюдении условия последней задачи поставить вопрос не об отборе семьи, а об отборе ребенка, то новая формулировка потребует, разумеется, другого решения. Пусть, например, *выбранный наугад из семей с двумя детьми ребенок оказался мальчиком (событие C). Оценим вероятность того, что оба ребенка в семье — мальчики?* Здесь возможны сочетания: 1) мальчик, мальчик; 2) мальчик, мальчик; 3) мальчик, девочка; 4) девочка, мальчик (сочетание «мальчик, мальчик» повторяется дважды, так как идет отбор мальчиков, а не семей). Следовательно, из четырех сочетаний два удовлетворяют условию; искомая вероятность равна $\frac{1}{2}$.

Решение этой задачи с помощью условной вероятности производится следующим образом. Отбирается первый ребенок в семье — мальчик; в этом случае наступлению события C благоприятствуют два исхода: 1) мальчик, мальчик; 2) мальчик, девочка из четырех (при отборе семей с мальчиком, когда не было известно, какой это был ребенок в семье — первый или второй, — событию C благоприятствовали три исхода: 1) мальчик, мальчик; 2) мальчик, девочка; 3) девочка, мальчик). Вероятность события C будет равна:

$$p\,(C) = \frac{2}{4} = \frac{1}{2}\,.$$

Вероятность пересечения событий остается прежней $P(AC) = p(A) = \frac{1}{4}$. Тогда условная вероятность того, что если первый отобранный ребенок из семей с двумя детьми оказался мальчиком, то и второй ребенок окажется тоже мальчиком, равна:

$$P\,(A/C) = \frac{P\,(AC)}{p\,(C)} = \frac{1}{4} : \frac{1}{2} = \frac{1}{2}\,.$$

Таким образом, при отборе семьи с мальчиком условная вероятность того, что второй тоже мальчик, оказывается равной $\frac{1}{3}$, а при отборе первого ребенка, оказавшегося мальчиком, условная вероятность того, что второй тоже мальчик, равна $\frac{1}{2}$.

В заключение укажем, что привлечение различных распределений в демографии с вероятностной точки зрения представляет не что иное, как попытку снижения «размерности» данных, т. е. сведения многообразия фактических результатов измерений к формуле с небольшим числом параметров-характеристик[1].

[1] Так, для построения нормального распределения необходимы две характеристики (средняя и дисперсия); для построения распределения Пуассона достаточно одной характеристики (средней) и т. д.

Глава 3

Изучение зависимостей между демографическими явлениями

3.1. Связи демографических явлений

Зависимости между массовыми демографическими явлениями проявляются чаще всего в виде корреляционных связей. На величину результативного признака оказывает влияние множество различных факторов, включающих в свой состав также и случайные факторы. Вот почему вывод о наличии корреляционной связи между изучаемыми признаками может носить только вероятностный характер. При этом надо также учитывать, что показатель тесноты корреляционной связи обладает определенными ошибками, расчет которых (в том числе и предельных значений этих ошибок) связан с теорией вероятностей.

Как и в других науках, в демографии связь между причиной (факторный признак) и следствием (результативный признак) не всегда имеет односторонний характер. Существует еще и обратная связь (взаимодействие), позволяющая утверждать, что следствие не всегда является пассивным по отношению к порождающим его причинам. На это обстоятельство указывал К. Маркс, имея в виду обратную связь в социальных процессах: «Но без некоторого смешения причины со следствием дело обойтись не может, так как причина и следствие во взаимодействии утрачивают свои отличительные признаки»[1]. Разумеется, хронологически причина предшествует следствию и порождает его.

Связи демографических явлений можно изучать четырьмя индуктивными способами:

1. *Способ совпадений,* основанный на сопоставлении случаев, в которых явления встречаются. Если за явлением A, сопровождающимся явлениями B и C (совокупное явление ABC), следует явление a, затем за соединением явлений A, D и E (совокупное явление ADE) снова следует явление a, то можно сделать вывод, что явление A, входящее в состав явлений ABC и ADE, предшествует явлению a, т. е. связано с ним. Этот способ используется в тех науках, где эксперимент невозможен. Так, этот способ применим в демографии, где действует множественность причин. Чем больше случаев совпадений A и a, тем связь между ними становится достовернее.

2. *Способ разницы,* основанный на сопоставлении различных случаев, в одних из которых явление встречается, в других — отсутствует. Если за совокупностью явлений ABC следует явление a, затем за совокупностью явления BC, в котором явление A отсутствует, не следует явление a, то делаем тот же вывод, что и в первом способе, — о связи явлений A с a. Этот способ в области демографии применяется крайне редко.

3. *Способ остатков,* который состоит в том, что если за совокупностью явлений ABC следует явление a, при этом из прошлых обследований известно, что ни явление B, ни явление C не может быть причиной явления a, то причиной явления a полагаем явление А. Этот способ при изучении многих демографических явлений совсем неприменим, так как требует экспериментирования.

4. *Способ сопутствующих изменений,* который состоит в допущении, что изменение одного явления, влекущее за собой изменение другого явления, должно быть связано с причиной его возникновения. Если за количественным изменением явления A происходит количественное изменение явления a и при этом известно числовое соотношение, в котором состоят изменения этих явлений, то можно утверждать, что явление A есть причина явления a. Этот способ чаще всего применяется при статистическом изучении демографических явлений.

Однако использование всех этих способов при изучении демографических явлений по ряду причин весьма условно. Во-первых, они могут указать лишь на сопутствие явлений, а не на причинную их связь. Во-вторых, в демографии имеет место сочетание множества факторов и смешение их действий; выявить изолированное действие какого-либо фактора этими методами крайне трудно, а порой просто невозможно. Делая вывод о связи явления A и a, можно ошибиться, так как эти явления могут зависеть не друг от друга, а от действия какого-нибудь иного, общего фактора. Именно поэтому заключение о закономерных связях между явлениями может носить только вероятностный характер, и задача при установлении связей демографических явлений состоит в оценке степени этой вероятности.

Очевидно, нужно выяснить следующие вопросы: какое число случаев совпадений указывает, что найденную причинную связь можно считать достоверной; насколько допустимы по величине отклонения явлений от найденной закономерности, чтобы их можно было считать случайными? На эти и множество других вопросов ответ дает теория вероятностей.

Привлечение и использование теории вероятностей в демографии возможно лишь на основе глубокого знания реальных свойств демографических явлений и понимания того, что вероятности этих

[1] *Маркс К., Энгельс Ф.* Соч., т. 8, с. 214.

демографических явлений вообще имеют смысл (в частности, и для последующих расчетов) при том непременном условии, что изученные реальные свойства и сочетания причин постоянны и в дальнейшем сколько-нибудь значительно изменяться не будут. В противном случае использование теории вероятностей бесцельно и не имеет практического значения.

Надо иметь в виду, что в вероятностном смысле зависимость событий не всегда самоочевидна. В ряде случаев необходима проверка путем следующего логического определения: два события A и B называются *независимыми*, если они удовлетворяют соотношению: $p(AB)=p(A)\cdot p(B)$, где $p(AB)$, $p(A)$ и $p(B)$ — соответствующие вероятности. Рассмотрим проверку зависимости событий на примере восьми сочетаний полов детей в трехдетных семьях: ммм, ммд, мдм, мдд, дмм, дмд, ддм, ддд (м — мальчик, д — девочка). Допустим, что каждое из этих восьми сочетаний равновозможно и имеет вероятность $^1/_8$. Возьмем два события: первое событие A — в семье есть дети обоего пола, второе событие B — в семье не более одной девочки. Надо ответить на вопрос зависимы или независимы эти события. Соответствующие вероятности будут равны: $p(A)=6/8$; $p(B)=4/8$. Используя теорему умножения вероятностей, получаем: $p(A)\cdot p(B)=\frac{6}{8}\cdot\frac{4}{8}=\frac{3}{8}$. Объединенным двум условиям A и B, т. е. AB, удовлетворяют три сочетания полов детей в семье: ммд, мдм, дмм. Следовательно, вероятность одновременного осуществления событий A и B, т. е. $p(AB)$, равна 3/8. Таким образом, $p(AB)=p(A)\cdot p(B)=3/8$. *Вывод: в семьях с тремя детьми события A и B независимы.*

3.2. Стохастические совокупности

Термин *стохастический* происходит от греческого слова *stochasis* — догадка. Мы встречаемся со стохастическими совокупностями, представляющими собой совокупности всевозможных случаев, реализующихся в вероятностном процессе. В качестве примера рассмотрим реальную совокупность, состоящую из 4 семей с числом членов семей 2, 3, 4 и 5 человек и средним размером семьи 3,5 человека. Возьмем под наблюдение две семьи, которые могут быть отобраны из указанных четырех в различных комбинациях. Получаемые вероятностным путем при реализации таких комбинаций совокупности называются *стохастическими*.

Возможны два способа извлечения семей, подвергаемых наблюдению: 1) когда извлеченная семья мыслится возвращенной в совокупность и продолжает участвовать в последующих извлечениях, тогда мы имеем дело с *повторным отбором;* 2) когда отобранная семья мыслится невозвращенной в совокупность и, следовательно, не может участвовать в последующих извлечениях — *бесповторный отбор.*

Теснота внутригрупповой связи между единицами стохастических совокупностей измеряется линейным коэффициентом корреляции ($r_{y}/_{x}$) по формуле

$$r_{y/x}=\frac{\overline{xy}-\bar{x}\cdot\bar{y}}{\sigma_x\cdot\sigma_y},$$

где $\overline{xy}=\frac{\Sigma xy}{n}$ — средняя из произведений вариантов x и y; $\bar{x}=\frac{\Sigma x}{n}$ — средняя из вариантов x; $\bar{y}=\frac{\Sigma y}{n}$ — средняя из вариантов y.

Теоретически доказано, что в стохастической совокупности, полученной повторным отбором единиц, линейный коэффициент корреляции $r_{y/x}$ равен нулю, тем самым подчеркивается независимость испытаний, а при бесповторном отборе $r_{y/x}=-\frac{1}{N-1}$. Рассмотрим стохастические совокупности семей при повторном (табл. 3.1) и бесповторном (табл. 3.2) отборах из указанных выше четырех семей.

Таблица 3.1. **Стохастические совокупности семей при повторном отборе**

Число членов семьи в первой отобранной семье, x	Число членов семьи во второй отобранной семье, y	xy	x^2	y^2
2	2	4	4	4
2	3	6	4	9
2	4	8	4	16
2	5	10	4	25
3	2	6	9	4
3	3	9	9	9
3	4	12	9	16
3	5	15	9	25
4	2	8	16	4
4	3	12	16	9
4	4	16	16	16
4	5	20	16	25
5	2	10	25	4
5	3	15	25	9
5	4	20	25	16
5	5	25	25	25
56	56	196	216	216
В среднем 3,5	3,5	12,25	13,5	13,5

Сделаем соответствующие расчеты:

$$\sigma_x^2=\overline{x^2}-(\bar{x})^2=\frac{\Sigma x^2}{n}-(3{,}5)^2=13{,}5-12{,}25=1{,}25;$$

$$\sigma_y^2=\overline{y^2}-(\bar{y})^2=\frac{\Sigma y^2}{n}-(3{,}5)^2=13{,}5-12{,}25=1{,}25.$$

Найдем линейный коэффициент корреляции:

$$r_{x/y}=\frac{12,5-12,5}{\sqrt{1,25\cdot 1,25}}=0.$$

Как и соответствует теории, он равен нулю.

Таблица 3.2. **Стохастические совокупности семей при бесповторном отборе**

Число членов семьи в первой отобранной семье, x	Число членов семьи во второй отобранной семье, y	xy	x^2	y^2	Среднее число членов семьи в группах, $\bar{x}$
2	2	6	4	9	2,5
2	4	8	4	16	3,0
2	5	10	4	25	3,5
3	2	6	9	4	2,5
3	4	12	9	16	3,5
3	5	15	9	25	4,0
4	2	8	16	4	3,0
4	3	12	16	9	3,5
4	5	20	16	25	4,5
5	2	10	25	4	3,5
5	3	15	25	9	4,0
5	4	20	25	16	4,5
42	42	142	162	162	42,0
В среднем 3,5	3,5	11,83333	13,5	13,5	3,5

В этом случае линейный коэффициент корреляции равен:

$$r_{y/x}=\frac{11,83333-12,25}{\sqrt{(13,5-12,25)\ (13,5-12,25)}}=-\frac{0,41667}{1,25}=-\frac{1}{3}.$$

В соответствии с теоретической формулой получаем:

$$r_{y/x}=-\frac{1}{N-1}=-\frac{1}{4-1}=-\frac{1}{3}.$$

Интересно отметить, что формулы средних ошибок типической выборки μ при повторном и бесповторном отборах вытекают из одной общей формулы

$$\mu=\sqrt{\frac{\sigma^2}{n}[1+(n-1)r_{y/x}]}.$$

Поэтому для повторного отбора при $r_{y/x}=0$ получаем:

$$\mu=\sqrt{\frac{\sigma^2}{n}},$$

для бесповторного отбора соответственно имеем:

$$\mu=\sqrt{\frac{\sigma^2}{n}\left[1+(n-1)\left(-\frac{1}{N-1}\right)\right]}=\sqrt{\frac{\sigma^2}{n}\cdot\frac{N-n}{N-1}}\approx$$

$$\approx\sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)}.$$

В качестве примера использования внутригрупповой связи в демографии укажем на контрольную перепись, учитывающую полноту охвата населения всеобщей переписью на 1 декабря 1965 г. в Народной Республике Болгарии. Использовалась следующая формула:

$$\Delta=t\sqrt{\frac{pq}{n}}\sqrt{1-\delta(\hat{n}-1)},$$

где t — коэффициент, связанный с вероятностью, что максимальный размер действительно не будет превышен (при $P=0,95$ $t=1,96$); p — удельный вес единиц в выборке, при которой установлен данный вид ошибки, например пропущенное лицо; q — удельный вес единиц в выборке, при которой не установлено данного вида ошибки; n — число единиц в выборке; δ — показатель степени внутригнездовой корреляции между единицами по данному признаку (наличие данного вида ошибок); $\hat{n}$ — среднее число наблюдаемых элементарных единиц, содержащихся в гнезде (в данном случае переписной участок). Формула, по которой исчислялась внутригнездовая корреляция, следующая:

$$\delta=\frac{S_c^2-\hat{n}S_2^2}{S_c^2+\hat{n}(\hat{n}-1)S_2^2},$$

где S_c^2 — междугнездовая дисперсия частоты данного вида ошибки в переписных участках; $S_2{}^2$ — внутригнездовая дисперсия единиц в переписных участках по данному виду ошибок. S_c^2 и $S_2{}^2$ определены по следующим формулам:

$$S_c^2=\frac{\Sigma x_i^2-\frac{(\Sigma x_i)^2}{m}}{m-1}+R^2\frac{\Sigma n_i^2-\frac{(\Sigma n_i)^2}{m}}{m-1}-$$

$$-2R\frac{\frac{\Sigma x_i n_i}{n}-\frac{\Sigma x_i m_i}{m}}{m-1};\quad S_2^2=\frac{1}{n}\Sigma\frac{n_i^2}{n_i-1}p_i q_i,$$

где n_i — размеры i-го гнезда (переписного участка); x_i — частота данного вида ошибки в i-м гнезде (переписном участке); m — число гнезд (переписных участков); R — относительная частота данного вида ошибки, определенная по всем m гнездам; n — общее число наблюдаемых лиц во всех m гнездах.

3.3. Вероятностное выявление тесноты корреляционных связей

Показатель тесноты связи между двумя факторами в случае, если известны соответствующие вероятности, определяется по формуле

$$r=\frac{p(A_1B_1)-p(A_1)p(B_1)}{\sqrt{p(A_1)[1-p(A_1)]\cdot p(B_1)[1-p(B_1)]}}$$

где p — соответствующие вероятности. Покажем пример использования этой формулы.

Покажем исследование вероятностным путем зависимости между признаками на примере наследования сыновьями изучаемого признака отцов. Результаты, полученные для совокупности, состоящей из 1000 пар отцов и сыновей, представлены в табл. 3.3.

Таблица 3.3. **Наследование сыновьями признака отцов**

Признак отцов	Признак сыновей: Наличие A_2	Признак сыновей: Отсутствие	Всего
Наличие, A_1	471 (A_1A_2)	151	622
Отсутствие	148	230	378
Всего	619	381	1 000

Введем соответствующие обозначения событий и их вероятностей.

Для отцов

Наличие признака — A_1; вероятность того, что случайно отобранное лицо будет обладать признаком A, — $p(A_1)=\frac{622}{1\,000}=0{,}622$, а вероятность того, что этот признак будет отсутствовать, — $q(A_1)=1-0{,}622=0{,}378$.

Для сыновей

Наличие признака — A_2; вероятность того, что случайно отобранное лицо обладает признаком, — $p(A_2)=\frac{619}{1\,000}=0{,}619$, а вероятность того, что признак будет отсутствовать, — $q(A_2)=1-0{,}619=0{,}381$.

Для сочетания признака у отцов и сыновей

Наличие признака у отцов и сыновей — A_1A_2; вероятность того, что отец и сын будут оба обладать признаком, — $p(A_1A_2)=\frac{471}{1\,000}=0{,}471$.

Если признаки у отцов и сыновей независимы, то вероятность того, что отец и сын будут обладать данным признаком, по теореме умножения вероятностей составит: $p(A_1)\,p(A_2)=0{,}622\cdot0{,}619=0{,}38502$. Таким образом, видим, что вероятность того, что у отца и сына признак присутствует, больше, чем если бы эти события были независимыми: $p(A_1A_2)>p(A_1)\,p(A_2)$. Можно зафиксировать в этом случае положительную связь. Коэффициент корреляции оказывается равным:

$$r=\frac{P(A_1A_2)-p(A_1)p(A_2)}{\sqrt{p(A_1)\cdot q(A_1)p(A_2)q(A_2)}}=\frac{0{,}471-0{,}38502}{\sqrt{0{,}622\cdot0{,}378\cdot0{,}619\cdot0{,}381}}=$$

$$=\frac{0{,}08598}{\sqrt{0{,}05545}}=\frac{0{,}08598}{0{,}23548}=0{,}36513.$$

Приведем еще примеры аналогичных расчетов. В совокупности, состоящей из 1000 человек, были сделаны прививки от некоторой болезни 300 человекам. Всего заболело 60 человек и из них 4 из числа тех, кому были сделаны прививки. Обозначим событие прививки — A_1, событие заболевания — A_2; отсюда вероятности того, что лицо, случайно отобранное из соответствующей совокупности:

а) подверглось прививке $p(A_1)=\frac{300}{1\,000}=0{,}3$;

б) не подверглось прививке $q(A_1)=\frac{700}{1\,000}=0{,}7$;

в) заболело $p(A_2)=\frac{60}{1\,000}=0{,}06$;

г) не заболело $q(A_2)=\frac{940}{1\,000}=0{,}94$;

д) заболело, будучи подвергнуто прививке, $p(A_1A_2)=\frac{4}{300}=0{,}013$.

Если вакцина не действует, то события A_1 и A_2 независимы; по теореме умножения вероятностей вероятность того, что человек, подвергшийся прививке, заболел, равна: $p(A_1)\,p(A_2)=0{,}3\times0{,}06=0{,}018$. Об эффективности прививок можно судить по коэффициенту корреляции r:

$$r=\frac{0{,}013-0{,}018}{\sqrt{0{,}3\cdot0{,}7\cdot0{,}06\cdot0{,}94}}=-\frac{0{,}005}{\sqrt{0{,}01184}}=-\frac{0{,}005}{0{,}01088}\approx-0{,}46.$$

Отрицательный знак коэффициента корреляции свидетельствует о том, что увеличение числа прививок снижает заболеваемость.

3.4. Использование критерия существенности

Для ответа на вопрос о влиянии возраста родителей на пол ребенка Г. Крамер[1] привлек выборочные данные о распределении мальчиков и девочек (всего 928,6 тыс. детей) по возрасту их от-

[1] См.: *Крамер Г.* Математические методы статистики. М., 1975, с. 496.

цов и матерей в Норвегии в 1871—1900 гг. (табл. 3.4). Производя соответствующие расчеты, получим следующие характеристики распределения (табл. 3.5).

Таблица 3.4. **Распределение родившихся по возрасту матери и отца**

Возраст отца, x	Возраст матери, y							
	до 20	20—25	25—30	30—35	35—40	40—45	45 и старше	всего
				Мальчики				
До 20	377	974	555	187	93	25	6	2 217
20—25	2 173	18 043	11 173	3 448	1 022	258	30	36 147
25—30	1 814	26 956	43 082	16 760	4 564	973	123	94 272
30—35	700	14 252	38 505	41 208	14 475	3 243	287	112 670
35—40	238	4 738	17 914	32 240	31 573	8 426	836	95 965
40—45	103	1 791	6 586	16 214	24 770	18 079	2 171	69 714
45—50	47	695	2 593	5 952	12 453	13 170	4 006	38 916
50—55	21	311	955	2 503	4 492	6 322	2 574	17 218
55—60	5	133	412	925	1 790	2 141	1 086	6 492
60—65	10	57	190	408	736	822	348	2 571
65—70	6	25	68	173	266	283	131	952
70 и старше	2	12	46	59	119	113	48	399
Всего	5 496	67 987	122 119	120 077	96 353	53 855	11 646	477 533
				Девочки				
До 20	319	861	504	206	91	22	3	2 006
20—25	2 133	16 990	10 643	3 193	979	242	45	34 225
25—30	1 793	25 147	40 817	15 637	4 305	943	96	88 738
30—35	707	13 254	36 745	38 619	13 699	3 018	292	106 304
35—40	236	4 676	17 165	30 453	29 858	7 883	772	91 043
40—45	101	1 670	6 278	15 323	23 803	16 983	1 941	66 099
45—50	38	640	2 384	5 603	11 764	12 336	3 823	36 588
50—55	16	284	964	2 469	4 221	5 815	2 480	16 249
55—60	12	120	406	874	1 726	2 000	1 079	6 217
60—65	6	54	171	381	591	750	325	2 278
65—70	3	29	87	154	277	247	114	911
70 и старше	1	18	30	67	108	115	40	379
Всего	5 365	63 743	116 194	112 979	91 392	50 354	11 010	451 037

Таблица 3.5. **Характеристики выборочных распределений мальчиков и девочек по возрасту их родителей**

Пол ребенка	Средний возраст		Среднеквадратическое отклонение в распределении по возрасту		Коэффициент корреляции между возрастами отцов и матерей
	отца, $\bar{x}$	матери, $\bar{y}$	отца, σ_1	матери, σ_2	
Мальчик	35,70	32,12	8,54	6,54	0,642
Девочка	35,70	32,12	8,40	6,54	0,641

Рассматривая возраста родителей x и y как фактические значения двумерной случайной величины, можно на основе совместного статистического распределения возрастов x и y в двух выборках (у мальчиков и девочек) проверить совместимость полученных данных с гипотезой о том, что обе выборки извлечены из одной и той же совокупности. Отрицательный ответ на вопрос о совместимости фактических данных с гипотезой будет свидетельствовать о том, что два полученных статистических распределения отцов и матерей по возрастам различны для мальчиков и девочек и, следовательно, соотношение полов зависит от возраста родителей. Положительный ответ на вопрос о совместимости фактических данных с проверяемой гипотезой не будет противоречить предположению, что оба распределения подчиняются одному и тому же закону распределения. Следовательно, можно сделать вывод о независимости пола ребенка от возраста родителей. Укрупняя возрастные интервалы для отцов и матерей и уменьшая число этих интервалов, можно сравнивать удельные веса одного из полов детей, например мальчиков, при переходе от младших возрастов родителей к старшим. При этом доведение агрегирования возрастных интервалов до получения четырехпольной таблицы (по два интервала у отцов и матерей) позволяет весьма наглядно проявить тенденцию изменения удельных весов мальчиков, если она имеется.

Выяснилось, что показатели асимметрии (3-й нормированный момент — r_3) и эксцесса (4-й нормированный момент — r_4) во всех полученных статистических распределениях значительно отличались от нуля, что свидетельствует об отклонении всех статистических распределений от нормального. Дополнительные исследования показали большую близость рассматриваемых распределений к логарифмически-нормальному распределению.

Для проверки того, извлечены ли обе выборки из одной и той же совокупности, был привлечен критерий хи-квадрат. Проверка различий основных выборочных характеристик для мальчиков и девочек ($\bar{x}$, σ, r и μ) показала, что они не существенны. Это позволило сделать вывод о том, что обе выборки извлечены из одной и той же совокупности, следовательно, двумерные распределения отцов и матерей для мальчиков и девочек близки друг к другу и соотношения полов среди родившихся не изменяются в зависимости от возрастов родителей.

На основе укрупнения возрастных интервалов можно получить для каждого из родителей две возрастные группы: для отцов — до 35 и старше 35 лет, для матерей — до 30 и старше 30 лет. Подсчитывая удельные веса рождения мальчиков и ошибки выборки в двух группах ук-

Таблица 3.6. **Доля мальчиков,** %

Возраст отца, x	Возраст матери, y	
	до 30	свыше 30
До 35	51,409±0,090	51,589±0,122
Старше 35	51,111±0,186	51,430±0,081

рупненных возрастных интервалов у родителей строим четырехпольную таблицу (табл. 3.6).

Сопоставляя полученные результаты, видим, что самое малое значение доли мальчиков равно 51,111, а самое большое — 51,589. Вычитая из этих величин и прибавляя к ним утроенные средние ошибки выборки, получаем с вероятностью 0,997 пересекающиеся границы долей мальчиков: $51{,}111 \pm 3 \cdot 0{,}186 = 51{,}111 \pm 0{,}558$ (т. е. в границах от 50,553 до 51,669); $51{,}589 \pm 3 \cdot 0{,}122 = 51{,}589 \pm 0{,}366$ (т. е. в границах от 51,223 до 51,955). Для наглядности изобразим полученные результаты на числовой оси (рис. 3.1).

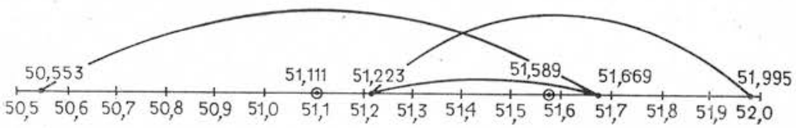

Рис. 3.1. Границы долей мальчиков у родителей разных возрастных групп

Таким образом, расчеты свидетельствуют о сравнительно большом участке пересечения долей мальчиков у родителей разных возрастных групп, следовательно, об отсутствии существенных различий в этих долях. Учитывая, что средние возраста отцов и матерей для совокупностей мальчиков совпадают с соответствующими средними для совокупности девочек (см. табл. 3.5), проверку существенности различия средних не приводим.

Все вышерассмотренное свидетельствует о том, что никакого существенного влияния возраста родителей на пол ребенка не обнаружено.

Для оценки влияния ликвидации смертности от различных причин на рост численности населения Н. Кейфитц привлек два типа возрастных распределений умирающих на Филиппинах: от болезней, свойственных младшим и свойственных старшим возрастным группам. В качестве примера болезней автором были взяты малярия (для младших возрастных групп) и болезни сердца (для старших возрастных групп), уровни смертности от которых в 1959 г. для Филиппин были равны друг другу. Привлекая вероятности смерти в разных возрастах для определения суммарной оценки последствий устранения смертности, автор пришел к выводу, что «...хотя абсолютные числа смертей от малярии и болезней сердца примерно равны, малярия влияет на рост населения примерно в четыре раза сильнее, чем болезни сердца»[1]. Таким образом, подчеркивается, что на росте численности населения ликвидация сердечных заболеваний отражается слабо, а ликвидация заболеваний малярией оказывает существенное влияние.

---

[1] *Кейфитц Н.* Возмещение долга: приложение к проблемам миграции и контроля над рождаемостью. — В кн.: Математика в социологии. Моделирование и обработка информации/Пер. с англ. М., 1977, с. 504—509.

Таблица 3.7. **Распределение браков по возрасту жениха и невесты, тыс.**

Возраст невесты, лет	Возраст жениха, лет										Всего
	моложе 20 лет (19)	20—24 (22,5)	25—29 (27,5)	30—34 (32,5)	35—39 (37,5)	40—44 (42,5)	45—49 (47,5)	50—54 (52,5)	55—59 (57,5)	60 лет и старше (62,5)	
Моложе 20 лет (18,0)	71	203	173	13	1	—	—	—	—	—	461
20—24 (22,5)	30	258	286	34	6	1	—	—	—	—	615
25—29 (27,5)	4	84	250	74	25	5	1	—	—	—	443
30—34 (32,5)	—	10	56	50	33	10	3	1	—	—	163
35—39 (37,5)	—	2	15	25	40	20	8	4	2	2	118
40—44 (42,5)	—	—	2	6	16	18	12	10	5	4	73
45—49 (47,5)	—	—	—	—	3	5	8	11	8	7	42
50 и старше (52,5)	—	—	—	—	—	1	4	19	41	102	167
Всего	105	557	782	202	124	60	36	45	56	115	2 082

При изучении связи возрастов женихов и невест привлечем данные о сопряженности этих двух признаков (табл. 3.7). Используем критерий существенности хи-квадрат, получаем:

$$\chi^2 = 2\,082 \cdot \left[\left(\frac{71^2}{105 \cdot 461} + \frac{203^2}{557 \cdot 461} + \ldots + \frac{41^2}{56 \cdot 167} + \right.\right.$$

$$\left.\left. + \frac{102^2}{115 \cdot 167}\right) - 1\right] = 2\,082 \cdot (0{,}10414 + 0{,}16049 + \ldots + 0{,}17975 +$$

$$+ 0{,}54173 - 1{,}00000) = 2\,082 \cdot (2{,}46853 - 1{,}00000) =$$

$$= 2\,082 \cdot 1{,}46853 \approx 3\,066.$$

Наименьшее число групп у невест равно 8, следовательно, $q = 8$. Тогда $f^2 = \frac{3\,066}{2\,082 \cdot 7} \approx 0{,}21$ и мера связанности $f \approx 0{,}46$. Полученный результат свидетельствует о том, что связь возрастов женихов и невест довольно сильная.

Можно устанавливать наличие связи не только у признаков, выражаемых количественно, как в предыдущем примере, но и качественно. Вот еще пример установления связи между различными комбинациями признаков у женихов и невест (данные условны — табл. 3.8).

Таблица 3.8. **Комбинации супружеских пар**

Женихи	Невесты			
	девицы	вдовы	разведенные	всего
Холостые	30	5	5	40
Вдовцы	20	10	5	35
Разведенные	10	5	10	25
Всего	60	20	20	100

Вычислим хи-квадрат и получаем:

$$\chi^2=100\left[\left(\frac{30^2}{60\cdot40}+\frac{5^2}{20\cdot40}+\frac{5^2}{20\cdot40}+\frac{20^2}{60\cdot35}+\frac{10^2}{20\cdot35}+\frac{5^2}{20\cdot35}+\right.\right.$$
$$\left.\left.+\frac{10^2}{60\cdot25}+\frac{5^2}{20\cdot25}+\frac{10^2}{20\cdot25}\right)-1\right]=12{,}322.$$

Число степеней свободы K равно: $(3-1)\cdot(3-1)=4$; $q=3$; $f^2=\frac{12{,}322}{100\cdot(3-1)}=\frac{12{,}322}{200}\approx0{,}06161$ $f=0{,}25$. Устанавливаем вывод: связь признаков не очень тесная.

3.5. Использование дисперсионного анализа для выявления влияния социального положения и пола на затраты времени на культурные мероприятия

Как мы уже видели, в демографии мы часто встречаемся с необходимостью изучать и оценивать степень влияния одного или нескольких факторных признаков на результативный признак. При этом нас интересует в первую очередь колеблемость признаков в отобранной совокупности, измеряемая с помощью дисперсий, приходящихся на одну степень свободы вариации. Отношение факторной дисперсии, полученной в результате обработки фактических данных, к случайной позволяет сделать определенный вывод о степени влияния факторов и оценить его с различной вероятностью. При изучении влияния одного факторного признака мы имеем дело с однофакторным комплексом, а при влиянии двух факторных признаков с двухфакторным комплексом и т. д.

Основными задачами дисперсионного анализа является обработка информации, полученной с помощью выборочного наблюдения, и выявление: 1) влияния факторных признаков на результативный; 2) роли каждого факторного признака и их сочетаний, а также определение достоверности выводов о влиянии факторных признаков и, следовательно, о возможности перенесения этих выводов на генеральную совокупность. Рассмотрим двухфакторный равномерный комплекс (с равным числом значений результативного признака): раздельное и совместное влияние факторных признаков — социального положения (рабочие — A_1, служащие — A_2) и пола (мужской — B_1, женский — B_2) на результативный признак (затраты времени на культурные мероприятия — y). Исходим из независимости факторных признаков A и B.

Таблица 3.9. **Двухфакторный статистический комплекс**

Группы факторного признака A	A_1		A_2		Всего
Группы факторного признака B	B_1	B_2	B_1	B_5	
Результативный признак, y	39 46 38	29 27 22	33 40 38	28 35 33	
Σy	123	78	111	96	$\Sigma(\Sigma y)=408$
n_x	3	3	3	3	$n=\Sigma n_x=12$
$h=\frac{(\Sigma y)^2}{n_x}$	5 043	2 028	4 107	3 072	$\Sigma\frac{(\Sigma y)^2}{n_x}=$ $=14\,250$
Σy^2	5 081	2 054	4 133	3 098	$\Sigma(\Sigma y^2)=$ $=14\,366$

Полученные в табл. 3.9 результаты и значения $H=\frac{[\Sigma(\Sigma y)]^2}{n}=\frac{408^2}{12}=13\,872$ позволяют вычислить основные три суммы квадратов отклонений результативного признака:

$$D_y^2=\Sigma(\Sigma y^2)-H=14\,366-13\,872=494;$$
$$D_z^2=\Sigma(\Sigma y)^2-\Sigma h=14\,366-14\,250=116;$$
$$D_x^2=\Sigma h-H=14\,250-13\,872=378.$$

Для вычисления сумм квадратов отклонений факторных признаков, используя данные табл. 3.9, строим таблицу (табл. 3.10).

Таблица 3.10. **Показатели по группам факторных признаков**

Факторный признак A					Факторный признак B				
Группа	n_A	Σy	$(\Sigma y)^2$	h_A	Группа	n_B	Σy	$(\Sigma y)^2$	h_B
A_1	6	201	40 401	6733,5	B_1	6	234	54 756	9 126
A_2	6	207	42 849	7141,5	B_2	6	174	30 276	5 046
Всего	12	408	—	13875,0	Всего	12	408	—	14 172

Вычислим суммы квадратов отклонений:

для факторного признака A: $D_A^2=\Sigma h_A-H=13\,875-13\,872=3$;

для факторного признака B: $D_B^2=\Sigma h_B-H=14\,172-13\,872=300$;

для сочетания факторных признаков A и B: $D_{AB}^2=D_x^2-D_A^2-D_B^2=378-3-300=75$.

Вычисление суммы квадратов отклонений позволяют оценить степень совокупного влияния факторных признаков и влияния каждого из них:

1) влияние совокупное

$$\eta_x^2=\frac{D_x^2}{D_y^2}=\frac{378}{494}=0{,}765\ (76{,}5\%);$$

2) влияние случайное

$$\eta_z^2 = \frac{D_z^2}{D_y^2} = \frac{116}{494} = 0{,}235 \ (23{,}5\%);$$

3) влияние факторного признака A (социального положения)

$$\eta_A^2 = \frac{D_A^2}{D_y^2} = \frac{3}{494} = 0{,}006 \ (0{,}6\%);$$

4) влияние факторного признака В (поля)

$$\eta_B^2 = \frac{D_B^2}{D_y^2} = \frac{300}{494} = 0{,}607 \ (60{,}7\%);$$

5) влияние сочетания признаков A и B

$$\eta_{AB}^2 = \frac{D_{AB}^2}{D_y^2} = \frac{75}{494} = 0{,}152 \ (15{,}2\%).$$

Для установления достоверности влияния факторов найдем число степеней свободы варьирования (n') и дисперсии на одну степень свободы варьирования (σ^2):

1) по всему комплексу: $n_y^1 = 12-1 = 11$; $\sigma_y^2 = \frac{494}{11} = 44{,}9$;

2) для дисперсии по признаку A: $n'_A = 2-1 = 1$; $\sigma_A^2 = \frac{3}{1} = 3$;

3) для дисперсии по признаку B: $n'_B = 2-1 = 1$; $\sigma_B^2 = \frac{300}{1} = 300$;

4) для дисперсии по сочетанию признаков A и B: $n'_{AB} = 1 \cdot 1 = 1$; $\sigma_{AB}^2 = \frac{75}{1} = 75$;

5) для дисперсии по случайному влиянию: $n'_z = 12 - 2 \cdot 2 = 8$; $\sigma_z^2 = \frac{116}{8} = 14{,}5$.

Теперь можно найти расчетное значение ($F_{расч}$), представляющее отношение всех вычисленных дисперсий к случайной дисперсии:

$$F_{расч.\ A} = \frac{\sigma_A^2}{\sigma_z^2} = \frac{3}{14{,}5} = 0{,}207;$$

$$F_{расч.\ B} = \frac{\sigma_B^2}{\sigma_z^2} = \frac{300}{14{,}5} = 20{,}690;$$

$$F_{расч.\ AB} = \frac{\sigma_{AB}^2}{\sigma_z^2} = \frac{75}{14{,}5} = 5{,}172;$$

$$F_{расч.\ x} = \frac{\sigma_x^2}{\sigma_z^2} = \frac{126}{14{,}5} = 8{,}690.$$

Находим по соответствующим таблицам[1] значения F по степеням свободы вариации сравниваемых дисперсий с вероятностями 0,95 и 0,99, т. е. $F_{0,95}$ и $F_{0,99}$, и строим табл. 3.11.

Таблица 3.11. **Выявление достоверности выводов**

Факторные признаки	$F_{расч}$	$F_{табл}$	
		$F_{0,95}$	$F_{0,99}$
A — социальное	0,207	5,32	11,26
B — пол	20,690	5,32	11,26
AB — сочетание	5,172	5,32	11,26
A, B, AB — сумма	8,690	4,11	7,60

Вывод. Из табл. 3.11 видно, что $F_{расч}$ больше $F_{табл}$ только для факторного признака B (пол) и суммы факторных признаков. Таким образом, с вероятностью, большей чем 0,99, делаем вывод о достоверном влиянии на результативный признак (т. е. на затраты времени на культурные мероприятия) признака B, т. е. пола, и суммы факторных признаков. Что касается влияния фактора A (социального положения) на результативный признак и сочетания факторов AB, то их действие недостоверно.

3.6. Уравнение регрессионной связи между демографическими признаками

Более 100 лет назад Мендель построил теорию наследственности, основанную на теоретико-вероятностных принципах. Он исходил из того, что в хромосомах родительских клеток заложены определенные наборы признаков. При оплодотворении эти признаки комбинируются между собой независимо и случайно. В дальнейшем основная идея Менделя подвергалась существенным изменениям.

В ходе изучения проблемы наследственности со статистической точки зрения было установлено, что наследование потомками признаков родителей происходит лишь частично, значение признака в той или иной мере возвращается к среднему значению для всей совокупности. Графическое изображение этой связи получило название линии регрессии.

Закон наследственной регрессии Гальтона состоит в том, что между величиной признаков, например ростом отца и сына, существует нормальная корреляция. По Гальтону наследственность у ребенка определяется наполовину двумя родителями, на одну четверть — дедом и бабушкой, на одну восьмую — прадедом и прабабушкой и т. д. Гальтон доказал, что максимально возможное зна-

[1] См.: *Венецкий И. Г., Кильдишев Г. С.* Теория вероятностей и математическая статистика. М., 1975, с. 262.

7 Заказ 2715

чение коэффициента корреляции между величиной признака у отцов и сыновей равно $^1/_2$. Рассматривая наследование признаков по Гальтону, акад. С. Н. Бернштейн возражал против следующей неправильной экстраполяции. «...Если средний рост мужского населения 1,65 м, то, выбирая из населения только мужчин роста 1,75, получим в следующем поколении средний рост 1,70 м. Однако и в этом случае допускать, что, продолжая таким образом, можно было бы создать расу великанов произвольно большого роста, является необоснованной экстраполяцией, и фактически указанное увеличение роста будет постепенно замедляться, так как нормальная корреляция и здесь представляет лишь приближение»[1].

Построение уравнения регрессии решает несколько задач. Во-первых, путем установления соответствия между осредненными особым образом значениями, например, двух признаков устраняется та вариабильность фактической связи между этими признаками (по состоянию на любой конкретный момент одинаковым значениям одного признака соответствуют всякий раз разные значения другого признака). Во-вторых, что особенно важно в перспективных расчетах, путем подстановки в найденное уравнение регрессии значения одного признака, например доходов семьи, можно определять (уже не для настоящего времени, а для будущего) значения другого признака — расходов на питание.

При рассмотрении зависимости между двумя демографическими признаками для построения уравнения регрессии будем рассматривать полученные результаты n пар измерений (x_i, y_i) как выборку из генеральной совокупности. Предполагая связь между факторным признаком x и результативным y линейной, можно построить эмпирическое (по выборочным данным) уравнение регрессии $\overline{y}_x = a'_0 + a'_1 x$. Здесь $\overline{y}_x$ — выборочное значение результативного признака при конкретном значении факторного признака; a_0' и a_1' — параметры, являющиеся оценками соответствующих параметров a_0 и a_1 в генеральной совокупности.

Рис. 3.2. Линейные уравнения регрессии (в генеральной и выборочный совокупностях)

На рис. 3.2 в качестве примера представлены две линии регрессии — в генеральной и выборочной совокупностях. Возьмем на рис. 3.2 точку $A(x_i, y_i)$, представляющую результат i-го измерения значения двух признаков x и y. Тогда $e_i = BA_i = y_i - \overline{y}_{xi}$ — это оценка по выборочным данным случайных значений $u_i = CA_i = y_i' - \overline{y}'_{x_i}$. Привлекая метод наименьших квадратов, мы используем следующее условие:

[1] *Бернштейн С. Н.* Теория вероятностей. 4-е изд., доп. М., 1946, с. 389.

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (a_0' + a_1' x)]^2 = \min.$$

Задача состоит в определении допустимого расхождения между этими линиями путем построения доверительной зоны, попадание в которую генеральной линии регрессии гарантировалось бы заданной доверительной вероятностью p. Две кривые $y_{\max}$ и $y_{\min}$ ограничивают линию регрессии выборочной совокупности и представляют собой доверительную зону, в которую должна попасть линия регрессии генеральной совокупности.

Для определения параметров a_0' и a_1' выборочного уравнения регрессии нужно решить систему нормальных уравнений (два линейных уравнения):

$$\begin{cases} na_0' + a_1' \Sigma x = \Sigma y; \\ a_0' \Sigma x + a_1' \Sigma x^2 = \Sigma xy \, . \end{cases}$$

Получаем:

$$a_0' = \frac{\Sigma x^2 \Sigma y - \Sigma x \Sigma xy}{n\Sigma x^2 - (\Sigma x)^2};$$

$$a_1' = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}.$$

Подставляя значения факторного признака x в уравнение регрессии, получаем выравненные значения результативного признака — $\overline{y}_x$. Доверительная зона, внутри которой с вероятностью $p = 1 - \alpha$ будет находиться линия регрессии генеральной совокупности $\overline{y}_x = a_0 + a_1 x$, определяется двумя кривыми линиями (гиперболами) по следующим формулам:

$$y_{\min} = a_0 + a_1 x - u_{n-2}(p, \lambda)\, \sigma_{\text{ост}} \sqrt{\frac{1}{n} - \frac{(x-\overline{x})^2}{\Sigma (x-\overline{x})^2}};$$

$$y_{\max} = a_0 + a_1 x + u_{n-2}(p, \lambda)\, \sigma_{\text{ост}} \sqrt{\frac{1}{n} - \frac{(x-\overline{x})^2}{\Sigma (x-\overline{x})^2}}.$$

Если $c = \dfrac{m-\overline{x}}{\sqrt{\Sigma (x-x)^2}}$, а $D = \dfrac{l-\overline{x}}{\Sigma (x-\overline{x})^2}$, то

$$\lambda = \sqrt{\frac{1}{2}\left[1 - \frac{1+nCD}{\sqrt{(1+nC^2)(1+nD^2)}}\right]},$$

где m и l — границы, в которых заключены значения факторного признака.

Значение $u_{n-2}(p, \lambda)$ определяется по таблицам с тремя входными параметрами: данной доверительной вероятностью $p = 1 - \alpha$, числом степеней свободы и λ[1]. Покажем все необходимые дей-

[1] См.: *Большев Л. Н., Смирнов Н. В.* Таблицы математической статистики. М., 1968, с. 80.

Таблица 3.12. **Рост и вес 10 человек**

Рост, см x	Вес, кг y	x^2	y^2	xy	Выровненные значения веса, $\overline{y}_x$	$y-\overline{y}_x$	$(y-\overline{y}_x)^2$
20,0	19,8	400,00	392,04	396,00	19,8	0,0	0,00
22,5	22,8	506,25	519,84	513,00	22,3	0,5	0,25
25,0	24,5	625,00	600,25	612,50	24,9	—0,4	0,16
28,5	27,3	812,25	745,29	778,05	28,4	—1,1	1,21
31,0	31,0	961,00	961,00	961,00	30,9	0,1	0,01
33,5	35,0	1122,25	1225,00	1172,50	33,4	1,6	2,56
35,5	35,1	1260,25	1232,01	1246,05	35,4	—0,3	0,09
37,0	37,1	1369,00	1376,41	1372,70	37,0	0,1	0,01
38,0	38,5	1444,00	1482,25	1463,00	38,1	0,5	0,25
40,0	39,0	1600,00	1521,00	1560,00	40,00	—1,0	1,00
311,0	310,1	10100,00	10055,09	10074,8	310,2	0	5,54

ствия на условном примере изучения связи между ростом и весом 10 человек, отобранных случайным образом из генеральной совокупности 100 человек (для упрощения расчетов из значений роста вычтена постоянная величина 140 см, а из значений веса — 45 кг — табл. 3.12).

Предположим, что между x и y существует линейная зависимость вида:

$$\overline{y}_x = a_0' + a_1'x.$$

Найдем параметры a_0' и a_1' методом наименьших квадратов. Соответствующие подготовительные вычисления произведены в табл. 3.12: $\Sigma x = 311$; $\Sigma y = 310{,}1$; $\Sigma x^2 = 10\,100$; $\Sigma y^2 = 10055{,}09$; $\Sigma xy = 10074{,}8$.

Находим:

$$\overline{x} = \frac{\Sigma x}{n} = \frac{311}{10} = 31{,}1; \quad (\overline{x})^2 = 31{,}1^2 = 967{,}21;$$

$$\overline{y} = \frac{\Sigma y}{n} = \frac{310{,}1}{10} = 31{,}01; \quad \overline{x^2} = \frac{\Sigma x^2}{n} = \frac{10\,100}{10} = 1\,010;$$

$$\overline{y^2} = \frac{\Sigma y^2}{n} = \frac{10055{,}09}{10} = 1005{,}509;$$

$$\Sigma(x-\overline{x})^2 = \Sigma x^2 - (\Sigma x)\overline{x} = 10100{,}0 - 311{,}0\cdot 31{,}1 = 10100{,}0 - 9672{,}1 = 427{,}9;$$

$$\Sigma(y-\overline{y})^2 = \Sigma y^2 - (\Sigma y)\overline{y} = 10055{,}09 - 310{,}1\cdot 31{,}01 = 10055{,}09 - 9616{,}201 = 438{,}889;$$

$$\overline{xy} = \frac{\Sigma xy}{n} = \frac{10074{,}8}{10} = 1007{,}48;$$

$$\Sigma(x-\overline{x})(y-\overline{y}) = \Sigma xy - (\Sigma x)\overline{y} = 10074{,}80 - 311{,}0\cdot 31{,}01 = 10074{,}80 - 9644{,}11 = 430{,}69;$$

$$[\Sigma(x-\overline{x})(y-\overline{y})]^2 = 430{,}69^2 = 185493{,}88;$$

$$\frac{[\Sigma(x-\overline{x})(y-\overline{y})]^2}{\Sigma(x-\overline{x})^2} = \frac{185493{,}88}{427{,}9} = 433{,}49820.$$

Найдем параметры линейного уравнения регрессии:

$$a_0' = \frac{10\,100\cdot 310{,}1 - 311\cdot 10074{,}8}{10\cdot 10\,100 - 311^2} = \frac{3\,132\,010 - 3\,133\,262{,}8}{101\,000 - 96\,721} = \frac{-1252{,}8}{4\,279} = -0{,}29278;$$

$$a_1' = \frac{10\cdot 10074{,}8 - 311\cdot 310{,}1}{10\cdot 10\,100 - 311^2} = \frac{100\,748 - 96441{,}1}{101\,000 - 96\,721} = \frac{4306{,}9}{4\,279} = 1{,}00652.$$

Таким образом, уравнение регрессии для выборочной совокупности имеет вид: $\overline{y}_x = -0{,}29278 + 1{,}00652\cdot x$.

Подставляя значения x из табл. 3.12, получаем для первой строки $\overline{y}_{x=20} = -0{,}29278 + 20{,}1304 \approx 19{,}8$ и т. д. Используя итог графы $(y-\overline{y}_x)^2$ из табл. 3.12, равный 5,54, вычислим остаточную дисперсию с учетом числа степеней свободы $(10-2=8)$.

$$\sigma^2_{\text{остат}} = \frac{\Sigma(y-\overline{y})^2}{n-2} = \frac{5{,}54}{8} = 0{,}6925; \quad \sigma_{\text{остат}} = 0{,}83.$$

Теперь построим в диапазоне значений факторного признака ($m=20$; $l=40$) доверительную зону для линии регрессии генеральной совокупности. Произведем подготовительные расчеты:

$$C = \frac{m-\overline{x}}{\sqrt{\Sigma(x-\overline{x})^2}} = \frac{20-31{,}1}{\sqrt{427{,}9}} = -\frac{11{,}1}{20{,}68} = -0{,}537;$$

$$D = \frac{l-\overline{x}}{\sqrt{\Sigma(x-\overline{x})^2}} = \frac{40-31{,}1}{\sqrt{427{,}9}} = \frac{8{,}9}{20{,}68} = 0{,}430;$$

$$\lambda = \sqrt{\frac{1}{2}\left[1 - \frac{1+nCD}{\sqrt{(1+nC^2)(1+nD^2)}}\right]} =$$

$$= \sqrt{\frac{1}{2}\left[1 - \frac{1+10\cdot(-0{,}537)\cdot 0{,}430}{\sqrt{[1+10\cdot(-0{,}537)^2]\,[1+10\cdot(0{,}430)^2]}}\right]} =$$

$$= \sqrt{\frac{1}{2}\left[1-\frac{1-2{,}3091}{\sqrt{(1+2{,}88)(1+1{,}85)}}\right]} \approx 0{,}83$$

Используя расчетные таблицы[1], находим $u_8\,(0{,}95;\ 0{,}83) = 2{,}946$.

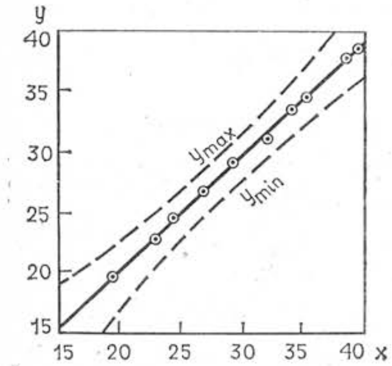

Рис. 3.3. Доверительная зона линий регрессии выборочной совокупности

Следовательно, неизвестная нам линия регрессии генеральной совокупности $\overline{y}_x = a_0 + a_1x$ заключена между гиперболами; $y_{\min\ и\ \max} = -\ 0{,}292 + 1{,}0065 \cdot x \pm \pm 2{,}946 \cdot 0{,}83 \sqrt{\frac{1}{10} - \frac{(x-31{,}1)^2}{427{,}9}}$. На рис. 3.3 представлена линия регрессии выборочной совокупности и доверительная зона, внутри которой заключена с вероятностью 0,95 линия регрессии генеральной совокупности. Произведенные в каждом конкретном случае расчеты и полученные при этом результаты дают демографу возможность судить о выходе или невыходе доверительной зоны за пределы, необходимые для прогнозирования.

3.7. Стохастические связи

При измерении связей между демографическими явлениями мы имеем дело со случайными переменными, т. е. величинами, принимающими те или иные значения с определенными вероятностями. Закон распределения случайных переменных — это совокупность всех значений случайной переменной с указанием вероятностей каждого из них. Между случайными переменными, например возрастами лиц, вступающих в брак, может существовать стохастическая связь, т. е. такая связь, при которой каждому значению одной случайной переменной величины, например возраста жениха, соответствуют различные значения другой переменной величины — возраста невесты — с определенными вероятностями.

Изучение стохастических связей покажем на следующем примере. Имеются данные о распределении браков по возрасту жениха и невесты (см. табл. 3.7). Обозначим возраст жениха — x и возраст невесты — y, найдем средние значения возраста невест для каждого значения возраста жениха и установим тенденцию изменения этих средних с увеличением возраста жениха. Получаем средний возраст невест для первой группы женихов:

[1] См.: *Большев Л. Н., Смирнов Н. В.* Таблицы математической статистики, с. 318.

$$\overline{y}_1 = \frac{\Sigma y_i m_i}{\Sigma m_i} = \frac{18{,}0 \cdot 71 + 22{,}5 \cdot 30 + 27{,}5 \cdot 4}{71+30+4} = 19{,}65.$$

Аналогично находим средние возраста невест для других групп женихов: $\overline{y}_2 = 21{,}8$; $\overline{y}_3 = 24{,}1$ и т. д. За средние возраста женихов берем центральные значения интервалов в каждой возрастной группе (см. табл. 3.7). Получаем соответствие возрастов женихов $(\overline{x})$ и невест $(\overline{y})$.

Таблица 3.13 **Средние возраста женихов $(\overline{x})$ и невест $(\overline{y})$**

$\overline{x}$	$\overline{y}$	$\overline{x}'$	$\overline{y}'$	$\overline{x}\overline{y}$	$(\overline{x})^2$
1	2	3	4	5	6
19,0	19,65	21,94	21,47	373,350	361,00
22,5	21,81			490,725	506,25
27,5	24,10	28,52	25,10	662,750	756,25
32,5	29,00			942,500	1056,25
37,5	34,16	39,13	35,47	1281,000	1406,25
42,5	38,17			1625,225	1806,25
47,5	43,57	50,31	45,66	2069,575	2256,25
52,5	47,28			2482,200	2756,25
57,5	49,47	60,84	50,89	2844,525	3306,25
62,5	51,59			3224,375	3906,25
401,5	358,80	—	—	15993,225	18117,25

Можно, укрупняя интервалы, соединить по две группы возрастов женихов с последующим расчетом для этих групп средних возрастов невест. Получаем для женихов новой первой укрупненной группы:

$$\overline{x}'_1 = \frac{19{,}0 \cdot 105 + 22{,}5 \cdot 558}{105+558} = \frac{1\,995 + 12\,555}{663} = \frac{14\,550}{663} = 21{,}94,$$

а для невест

$$\overline{y}'_1 = \frac{19{,}65 \cdot 105 + 21{,}81 \cdot 558}{105+558} = 21{,}47 \text{ и т. д.}$$

Запишем уравнение прямолинейной связи между найденными средними возрастами женихов и невест ($\overline{x}$ и $\overline{y}$, а также $\overline{x}'$ и $\overline{y}'$) и изобразим эмпирическую и теоретическую линии регрессии. Линейное уравнение связи имеет вид: $\overline{y}_x^{\,\prime} = a_0 + a_1\overline{x}$.

Для вычисления параметров a_0 и a_1 нужно решить следующую систему уравнений:

$$\begin{cases} a_0 n + a_1 \Sigma\overline{x} = \Sigma\overline{y}; \\ a_0 \Sigma\overline{x} + a_1 \Sigma\overline{x}^2 = \Sigma\overline{x}\overline{y}. \end{cases}$$

Производим расчет и получаем: $n = 10$; $\Sigma\overline{x} = 401{,}5$; $\Sigma\overline{y} = 358{,}8$; $\Sigma\overline{x}\overline{y} = 15993{,}225$; $\Sigma\overline{x}^2 = 18117{,}25$; $a_0 \approx 3{,}8$; $a_1 \approx 0{,}8$ (аналогичные значения параметров получаются и для укрупненных интервалов возрастов женихов и невест: $a_0=3{,}9$ и $a_1=0{,}8$). Тогда уравнение прямолинейной связи примет вид:

$$\overline{y_x}=3{,}8+0{,}8\overline{x}.$$

Подставляя в это уравнение значения $\overline{x}$, получаем выравненные значения средних возрастов невест $\overline{y_x}$. Так, для 19-летних женихов средний выравненный возраст невест оказывается равным: $3{,}8+0{,}8\cdot 19=3{,}8+15{,}2=19$ лет, а для 62,5-летних: $3{,}8+0{,}8\times\times 62{,}5=53{,}8$ года.

Строим эмпирическую и теоретическую линии регрессии возрастов женихов и невест (рис. 3.4), позволяющие выявить наличие стохастической связи между этими признаками.

Задача построения индексов брачности, измеряющих степень «притяжения» (предпочтения) или «отталкивания» (неодобрения) женихов и невест, решена с использованием теории вероятностей. Совершенным методом изучения склонностей брачных пар, основанным на теории вероятностей, является построение индексов брачности, разработанных Бенини и модифицированных М. В. Птухой.

В качестве примера исследуем вступление в брак лиц одной национальности. Число вступивших в брак мужчин данной национальности — B, женщин — C, число заключенных браков между женихами и невестами данной национальности — A, всего (в том числе и между лицами данной национальности) — D. Тогда

1) вероятность мужчине данной национальности найти себе невесту из совокупности невест C:

$$p_м=\frac{C}{D};$$

2) вероятность девушке данной национальности найти себе жениха из совокупности B:

$$p_ж=\frac{B}{D};$$

3) по теореме умножения вероятностей получаем вероятность вступления в брак жениха и невесты из указанных выше совокупностей B и C:

$$p_{м\cdot ж}=p_м p_ж=\frac{C}{D}\cdot\frac{B}{D}.$$

Теоретически число однородных в национальном отношении браков равно:

$$\frac{C}{D}\cdot\frac{B}{D}\cdot D=\frac{BC}{D}.$$

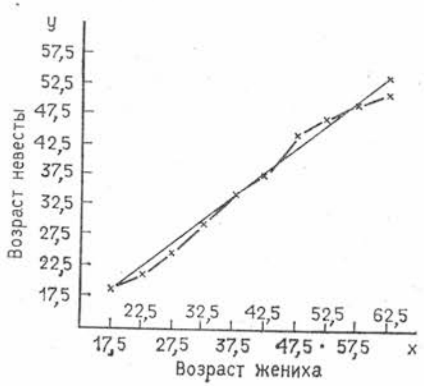

Рис. 3.4. Эмпирическая и теоретическая линия регрессии возрастов женихов и невест

Неоднородных браков должно быть:

$$B-\frac{BC}{D}=\frac{BD-BC}{D}.$$

При наличии «притяжения» однородных совокупностей первая группа будет увеличиваться за счет второй. При наличии «отталкивания» — наоборот. Найдем разность $A-\frac{BC}{D}$, показывающую размер увеличения первой группы за счет второй. Если теперь эту разность отнести к величине той совокупности, из которой она взята, то получим индекс «притяжения» $I_{пр}$:

для мужчин-женихов

$$I_{пр.м}=\frac{A-\frac{BC}{D}}{B-\frac{BC}{D}}=\frac{AD-BC}{B(D-C)};$$

для женщин-невест

$$I_{пр.ж}=\frac{A-\frac{BC}{D}}{C-\frac{BC}{D}}=\frac{AD-BC}{C(D-B)}.$$

Для обоих полов индекс «притяжения» (индекс гомогамии) можно получать как среднюю геометрическую между $I_{пр.м}$ и $I_{пр.ж}$:

$$I_{пр}=\sqrt{I_{пр.м}\,I_{пр.ж}}.$$

Аналогично рассчитывается индекс «отталкивания» (гетерогамии), имеющий тот же логический смысл и одинаковый для обоих полов, но с отрицательным знаком:

$$I_{отт}=\frac{A-\frac{BC}{D}}{\frac{BC}{D}}=\frac{AD}{BC}-1.$$

Значение получаемых индексов колеблется в пределах от -1 (полное «отталкивание») через 0 (безразличие) к $+1$ (полное «притяжение»).

Сравнивая индексы брачности в СССР с индексами в капиталистических странах, можно установить различия, обусловленные социально-экономическими факторами. Можно предположить, что индекс взаимного притяжения между группами однородных по ряду признаков молодых возрастов, вступающих в брак в СССР, будет выше, чем в капиталистических странах. Тем самым будет подтверждено уменьшение влияния косвенных мотивов при вступлении в брак (социально-экономических и имущественных).

Интересный вывод, на наш взгляд, может получиться при изучении индекса брачности по национальному признаку. Следует полагать, что индекс «притяжения» женихов отдельных национальностей по отношению к невестам той же национальности в нашей стране будет ниже. Это подтвердит, что национальные различия все меньше и меньше служат препятствием для вступления в брак.

Пользуясь этим методом анализа, следует помнить, что в основе системы индексов «притяжения» лежит гипотеза, соответствующая урновым схемам, но не подтвержденная действительностью о равновозможности «встреч» потенциальных женихов и невест. Так, вероятность «встречи» женихом-узбеком русской невесты меньше, чем невесты узбечки, хотя в СССР русских женщин в несколько десятков раз больше, чем узбечек.

Исследования Л. В. Чуйко по данным социально-демографических обследований в УССР показали, что «...объективный процесс — сближения национальностей — происходит в городских поселениях относительно быстрее, чем в сельских»[1].

Изложенный метод служит до некоторой степени тем же целям, что и показатель тесноты корреляционной связи, и может быть успешно применен для ответа на частный вопрос корреляционной теории — о тесноте «притяжения» или «отталкивания» каких-нибудь двух конкретных значений взаимосвязанных признаков у двух объектов, относительно которых известно фактическое число имевших место случаев комбинации.

[1] *Чуйко Л. В.* Браки и разводы. М., 1975, с. 79.

Глава 4

Вероятностные показатели таблиц смертности, брачности и плодовитости

По мысли некоторых демографов, математические модели в виде таблиц смертности, позволяющие изучать продолжительность жизни населения, представляют собой нечто вроде термометра или барометра, привлекаемых при исследовании физических явлений.

При использовании таблиц смертности следует иметь в виду, что абстрактные модели и формулы теории вероятностей отличаются от реальной демографической действительности. Нельзя механически переносить выводы, полученные по моделям и формулам, на реальность. Вот что писал А. А. Марков: «...признавая пользу таблиц смертности для практических целей, мы считаем невозможным доказывать законность их применения ссылками на формулы исчисления вероятностей»[1]. И. С. Пасхавер писал по этому поводу: «...когда речь идет о таблицах смертности, всецело базирующихся на принципе вероятностей и на вероятностных расчетах и полностью оправдавших себя в многолетней практике многих стран, нет оснований считать, что они имеют совершенно идентичную основу с абстрактными моделями теории вероятностей, к которым в полной мере применимы формулы исчисления вероятностей»[2]. Хотя абстрактные модели массовых процессов и процессов демографической действительности имеют много общего, на наш взгляд, последние значительно богаче, разнообразнее и сложнее.

4.1. Показатели таблиц смертности и их взаимосвязи

Демография разработала множество методов построения таблиц смертности. При этом специфика построения таблиц смертности во многом определяет специфику расчета вероятностных показателей. Вот почему рассмотрение вероятностных показате-

[1] *Марков А. А.* Исчисление вероятностей. 4-е изд. М., 1924, с. 311.
[2] *Пасхавер И. С.* Закон больших чисел и статистические закономерности. М., 1974, с. 104.

лей таблиц смертности должна предшествовать детализированная классификация методов построения таблиц смертности. Наиболее удобной, на наш взгляд, для этой цели служит классификация, разработанная А. Я. Боярским. Придерживаясь этой классификации, обратим главное внимание на один из показателей таблиц смертности, позволяющий развернуть все остальные показатели (графы таблиц), — вероятность умереть в течение года, построение которого основано на ряде допущений:

1. Принимается, что вероятность смерти в отрезок времени от t до $t+\Delta t$ выражается формулой $q(t, t+\Delta t)=f(t)\Delta t+Q(\Delta t)$, где $f(t)$ — неотрицательная непрерывная функция.

2. Предполагается, что смерть в промежуток времени от t_1 до t_2 не зависит от того, что было до момента t_1.

3. Допускается, что вероятность смерти в момент рождения равна нулю.

Учитывая указанные выше допущения, можно найти вероятность смерти лица до достижения им возраста t. Вероятность того, что лицо A доживает до возраста t, составит $p(t)$, а вероятность лица A дожить до возраста $t+\Delta t$, если оно уже дожило до возраста t, — $p(t+\Delta t, t)$, тогда

$$p(t+\Delta t)=p(t)\,p(t+\Delta t, t).$$

Вследствие первого и второго допущения получаем:

$$p(t+\Delta t, t)=1-f(t, \Delta t)-Q(\Delta t).$$

Отсюда находим, что $p(t)$ удовлетворяет следующему дифференциальному уравнению:

$$\frac{df(t)}{dt}=-f(t)\,p(t).$$

Учитывая третье допущение, находим решение дифференциального уравнения:

$$p(t)=e^{-\int_0^t f(z)dz}.$$

Следовательно, вероятность умереть прежде, чем будет достигнут возраст t лет, будет равна:

$$1-p(t)=1-e^{-\int_0^t f(z)dz}.$$

До перехода к классификации таблиц смертности и получению вероятности умереть напомним еще раз систему обозначений показателей таблиц смертности и формулы, устанавливающие взаимосвязи этих показателей.

Показатели таблиц смертности

l_x — числа доживающих до возраста x лет;
d_x — числа умирающих в возрасте от x до $x+1$ лет;
p_x — вероятность дожить от возраста x лет до $x+1$ лет;

q_x — вероятность умереть в возрасте от x до $x+1$ лет (доля умерших в составе совокупности людей, достигших возраста x лет);

L_x — числа живущих в данный момент в возрасте x лет (число человеко-лет, прожитых населением в один год) показывает число лиц в возрасте x лет в возрастном составе стационарного (постоянного) населения;

T_x — общее число человеко-лет, которое предстоит прожить совокупности l_x от возраста x лет до предельного возраста $\omega-1$;

e_x — средняя продолжительность предстоящей жизни, показатель, введенный Гюйгенсом, не зависящий от фактической возрастной структуры населения и вычисляемой только по таблицам смертности. С точки зрения теории вероятностей этот показатель представляет собой математическое ожидание длительности предстоящей жизни при достижении возраста x лет и вычисляется при допущении, что повозрастная смертность населения, положенная в основу построения таблицы смертности, для всего периода предстоящей жизни данного поколения остается неизменной. Нельзя путать этот показатель со средним возрастом умерших:

e_0 — средняя продолжительность жизни для новорожденного;

M_x — абсолютное число умерших в течение года в возрасте x лет;

V_x — вероятная продолжительность жизни (число лет, которое проживет после возраста x лет ровно половина достигших этого возраста);

μ_x — сила смертности (приведенная к одному году вероятность смерти в бесконечно малом возрастном интервале).

Взаимосвязь показателей таблиц смертности
(точные и приближенные)

$$d_x=l_x-l_{x+1};\quad \sum_{i=0}^{i=\omega-1} d_i=l_0=1$$

(исходное число родившихся);

$$l_x=1-d_0-d_1-\cdots-d_{x-1};\quad l_\omega=0;$$

$$l_x=d_x+d_{x+1}+\cdots+d_{\omega-1}=\sum_{i=x}^{i=\omega-1} d_i;$$

$$p_x=\frac{l_{x+1}}{l_x}=\frac{l_x-d_x}{l_x};$$

$$l_0p_0p_1\cdots p_x=l_xp_x=l_{x+1};$$

$$q_x=\frac{d_x}{l_x};\quad p_x+q_x=1;$$

$$e_x=\frac{L_x+L_{x+1}+\cdots+L_{\omega-1}}{l_x}=\frac{T_x}{l_x};$$

$$e_0=\frac{T_0}{l_0}=\frac{T_0}{1};\quad e_0=\frac{1}{2}d_0+\frac{3}{2}d_1+\frac{5}{2}d_2+\cdots;$$

$$T_x = \sum_{i=x}^{i=\omega-1} L_i\,;\quad L_x = T_x - T_{x+1}\,;\quad L_x \approx \frac{l_x + l_{x+1}}{2}\,;$$

$$\mu_x = -\frac{l'_x}{l_x}\,;\quad \mu_x = \frac{q_x}{1 - \frac{q_x}{2}}\,;$$

$$p_x \approx e^{-\mu_x}\,;\quad q_x = \frac{2\mu_x}{2 + \mu_x}\,;$$

$$\mu_x = \frac{2(l_x - l_{x+1})}{l_x + l_{x+1}}\,;$$

$$V_x = n + \frac{l_{x+n} - \frac{1}{2} l_x}{l_{x+1} - l_{x+n-1}},$$

где n — разность между возрастом x, для которого определяют величину вероятной продолжительности жизни, и тем возрастом $x+1$, в котором остается в живых несколько больше половины лиц возраста x лет; l_{x+n} и l_{x+n-1} — соседние табличные числа доживающих, из которых l_{x+n} несколько больше, а l_{x+n-1} несколько меньше $\frac{1}{2} l_x$.

4.2. Методы построения таблиц смертности и специфика вычисления вероятностей умереть

Классификация методов построения таблиц смертности дана А. Я. Боярским[1]:

I. Условный метод: 1) Дж. Граунта; 2) Л. Эйлера.

II. Прямой метод (П. Лапласа).

III. Косвенный метод: 1) В. Я. Буняковского; 2) чисел живущих; 3) индийских таблиц; 4) демографический.

IV. Метод определяющих функций (А. Я. Боярского).

V. Методы для детских возрастов.

Рассмотрим каждый из методов построения таблиц смертности с точки зрения тех его отличительных черт, которые наложили отпечаток и определили специфику вычисления вероятности умереть в том или ином возрасте, т. е. q_x.

I. **Условные методы.**

Метод Граунта (метод смертных списков) состоит в использовании возрастного состава умерших в течение определенного периода (чаще всего года). Основной предпосылкой этого метода является предположение о *стационарности* населения, т. е. сохранении постоянной плотности рождений и порядка вымирания l_x. В стационарном населении число ежегодно умирающих равно числу ежегодно рождающихся. Возрастная структура неизменна. Миграция отсутствует или взаимно уравновешивается.

[1] См.: *Боярский А. Я.* Курс демографической статистики. М., 1945, с. 167.

Численность каждого возраста в стационарном населении пропорциональна L_x. Численность всего населения $N\Sigma L_i = Ne_0$. Коэффициент рождаемости и коэффициент смертности равны между собой и равны $\frac{1}{e_0}$. Если M_x — число умерших в возрасте x лет, то вероятность умереть в этом возрасте может быть исчислена по формуле

$$q_x = \frac{M_x}{\sum_{i=x}^{i=\omega-1} M_i}\,.$$

Недостатком метода Граунта явилось преувеличение смертности младших возрастов и преуменьшение смертности старших возрастов. При росте рождаемости метод Граунта приводит к занижению средней продолжительности жизни.

Метод Эйлера, смягчающий недостаток метода Граунта, состоит в предположении *стабильности* населения (т. е. такого населения, изменение численности которого происходит в геометрической прогрессии с таким же знаменателем прогрессии, с каким изменяется и плотность рождений). При этом, следовательно, темпы роста (снижения) численности населения постоянны. Коэффициент естественного прироста неизменен.

В растущем стабильном населении удельный вес молодых возрастов более высок, чем в стационарном. Числа умерших M_x пропорциональны числам $e^{-kx}d_x$, которые рассчитываются на основе данных о численности населения в динамике. Определяя d_x, а затем l_x, находим

$$q_x = \frac{d_x}{l_x}\,.$$

II. **Прямой метод Лапласа** исходит из рассмотрения определенного поколения родившихся по мере его вымирания. Доли умирающих в общем числе родившихся составят d_x. Тогда

$$q_x = \frac{d_x}{l_x}\,.$$

Этот метод практического значения не имеет и используется главным образом для расчета когортных таблиц.

III. **Косвенные методы** характеризуют существующий в данное время повозрастной уровень смертности.

Метод В. Я. Буняковского основан на использовании чисел умирающих в течение года d_x и отличается от метода Эйлера тем, что вместо гипотетических чисел родившихся в течение года, меняющихся в геометрической прогрессии, используются фактические числа родившихся. Исходный показатель d_x определяется по формуле

$$q_x = \frac{M_x}{\frac{1}{2}(N_x + N_{x+1})}\,.$$

Откуда вероятность умереть

$$q_x = \frac{2M_x}{(N_x+N_{x+1})(1-q_0)(1-q_1)\dots(1-q_{x-1})}.$$

Этот метод часто используется для построения вероятностей умереть в детских возрастах. Недостатком этого метода является зависимость вероятности смерти в различных возрастах от предыдущей истории смертности.

Метод чисел живущих основан на использовании данных переписи. Исходным показателем является число живущих (L_x):

$$L_x = \frac{S_x}{N_x}$$

где S_x — абсолютное число живущих в возрасте x лет по переписи; N_x — абсолютное число родившихся в соответствующем году (x лет назад).

Исходя из $l_0=1$ и используя систему уравнений

$$L_0 = \frac{l_0+l_1}{2};\quad L_1 = \frac{l_1+l_2}{2},$$

и т. д., находим l_x и далее совершаем переход к вероятностям умереть. Для возрастов с высокой смертностью, например до одного года, расчет ведется более детализированно — по месяцам жизни.

Метод индийских таблиц основан на использовании числа умерших в период между двумя переписями. Исходным показателем являются коэффициенты дожития. При интервале между переписями, например, 5 лет получаем формулы коэффициентов дожития:

$$p_{x/x-5} = \frac{L_x}{L_{x-5}}.$$

Уравнения для определения числа живущих примут вид:

$$p_{5/0} = \frac{L_5}{L_0};\quad p_{6/1} = \frac{L_6}{L_1}\ \text{и т. д.}$$

Числа живущих для младших возрастов определяются с привлечением данных о родившихся за 5 лет между переписями. Переход к вероятностям смерти осуществляется так же, как и в методе живущих.

Демографический метод основан на использовании данных о рождаемости и смертности и отражает повозрастной уровень смертности в данное время. Для построения таблиц смертности используются числа умерших в их связи с числами живущих. Исходным показателем является q_x или μ_x. Впервые этот метод был применен В. Фарром и А. Кетле. Есть несколько модификаций этого метода, отличающихся исходной информацией.

А. Демографический метод с использованием данных о смертности и рождаемости. В этом случае для построения q_x используются четыре совокупности умерших первого рода, образующие прямоугольник $B_x\ A_{x+1}\ E_{x+1}\ F_x$ (см. рис. 4.1) и дожившие до возраста $x(B_x\ F_x)$:

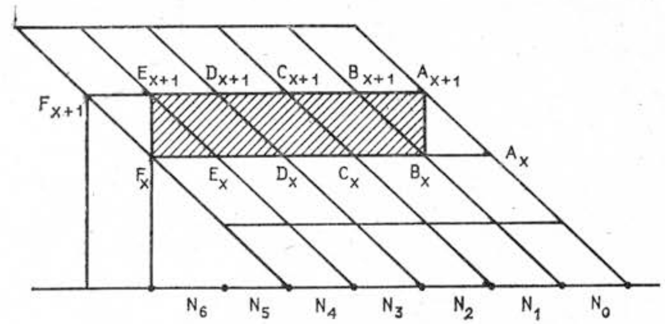

Рис. 4.1. Совокупности умерших первого рода и числа родившихся

$$q_x = \frac{M_x(B_xA_{x+1}E_{x+1}F_x)}{B_x F_x}.$$

Найденная вероятность умереть для достигших возраста x лет является средней для нескольких поколений.

Б. Демографический метод с использованием данных о смертности за некоторый период и переписи, проведенной в том же

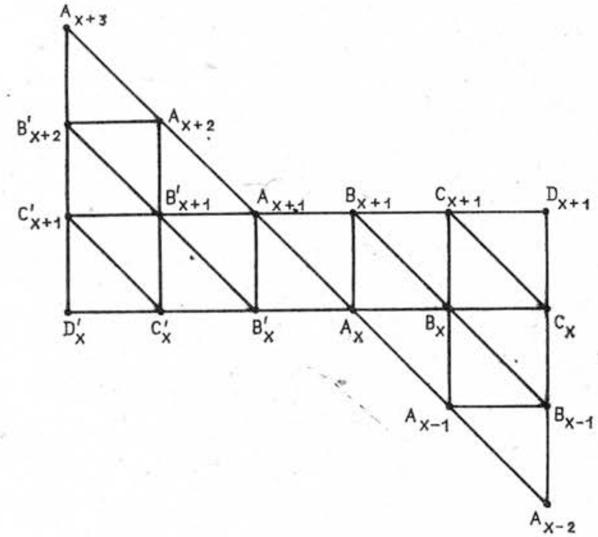

Рис. 4.2. Определение q_x с использованием переписных данных и элементарных совокупностей умерших

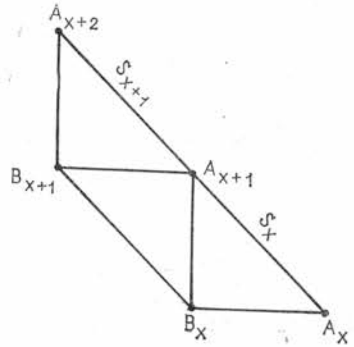

Рис. 4.3. Определение вероятности смерти методом Бёка

периоде (рис. 4.2). Имеются переписные данные и элементарные совокупности умерших. Исходный показатель — вероятность смерти q_x:

$$q_x=\frac{M_x(C_xD_{x+1}C'_{x+1}D'_x)}{C_x\,D'_x}.$$

Из рис. 4.2 видно, что $C_xD'_x=C_xA_x+A_xD'_x$, где

$$C_xA_x=A_{x-2}A_x-A_{x-2}A_xC_x,\text{ а}$$
$$A_xD'_x=A_xA_{x+3}+A_xA_{x+3}D'_x.$$

Тогда получаем:

$$C_xA_x=A_{x-2}A_{x+3}+A_xA_{x+3}D'_x-A_{x-2}A_xC_x.$$

Здесь $A_{x-2}A_{x+3}$ есть число живущих по переписи в возрасте от $x-2$ лет до $x+2$ лет (включительно).

Треугольники $A_xA_{x+3}D'_x$ и $A_{x-2}A_xC_x$ составляются суммированием образующих их элементарных совокупностей умерших.

Этот метод построения таблиц смертности использовался французскими демографами.

В. Метод Бёка — вариант демографического метода, основанный на привлечении данных переписи и умерших по элементарным совокупностям за год до переписи или за год после нее (рис. 4.3). Сначала в соответствии с рис. 4.3 определяют p_x:

$$p_x=\frac{S_x}{S_x+B_xA_xA_{x+1}}\cdot\frac{S_{x+1}+A_{x+1}A_{x+2}B_{x+1}}{S_{x+1}+B_xA_{x+1}A_{x+2}\,B_{x+1}}.$$

Затем можно вычислить вероятность смерти $q_x=1-p_x$.

Г. Демографический метод, основанный на данных двух переписей и смертности в промежуток между ними, имеет в свою очередь модификации.

Скандинавские таблицы смертности ориентированы на данные двух переписей населения, отделенных 10-летним периодом с использованием совокупностей умерших третьего рода за промежуток между переписями. Привлекая данные о числе умерших за 10 лет, а также значение среднего числа живущих в возрасте x лет в промежутке между переписями ($\overline{S}_x$), получаем:

$$q_x=\frac{A_xA_{x+1}H_{x+1}H_x}{10\overline{S}_x+(A_xA_{x+1}B_x+B_xB_{x+1}C_x+\ldots+N_xN_{x+1}H_x)}.$$

В знаменателе вероятности смерти — $10\overline{S}_x$ и сумма нижних элементарных совокупностей умерших (рис. 4.4).

Американские таблицы аналогичны скандинавским, но среднее число живущих в возрасте x лет рассчитывается как полусумма живущих в возрасте x лет по двум переписям:

$$\overline{S}_x=\frac{A_xA_{x+1}+H_xH_{x+2}}{2}=\frac{S_x+S'_x}{2}$$

или

$$\overline{S}_x=\frac{S'_x-S_x}{\ln S'_x-\ln S_x}$$

тогда

$$q_x=\frac{2M_x}{10(S_x+S'_x)+M_x}$$

или более точно

$$q_x=\frac{M_x}{10\left(\frac{S'_x-S_x}{\ln S'_x-\ln S_x}\right)+\frac{1}{2}M_x}.$$

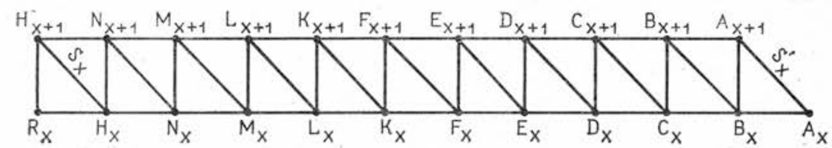

Рис. 4.4. Исходные данные для расчета показателей скандинавских и американских таблиц смертности

Д. Советские таблицы смертности ориентированы на данные о смертности за 1—2 года до переписи и за 1—2 после нее.

С. А. Новосельский при построении таблиц смертности использовал совокупности умерших третьего рода и данные переписи населения, приходящейся на середину периода, за который взята смертность. Исходным показателем является коэффициент смертности населения в возрасте x лет — m_x:

$$m_x=\frac{d_x}{L_x},\text{ откуда } q_x=\frac{2m_x}{2+m_x}.$$

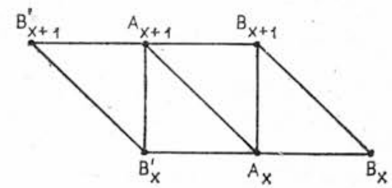

Рис. 4.5. Определение вероятности смерти методом Паевского

Представляя значение $m_x=\frac{M_x}{2S_x}$, получаем: $q_x=\frac{2M_x}{4S_x+M_x}$.

В. В. Паевский использовал совокупности умерших третьего рода за два календарных года и возрастные численности населения по переписи. Вероятность умереть при этом определяется по формуле

$$q_x=1-e^{-\mu_x},\text{ где } (\mu_x\approx m_x);$$

$$m_x=\frac{M_x}{2S_x}=\frac{B'_xB_xB_{x+1}B'_{x+1}}{2A_xA_{x+1}}.$$

IV. **Метод определяющих функций**, разработанный А. Я. Боярским для советских таблиц смертности 1938—1939 гг., представляет собой результат применения основных принципов теории средних величин и ориентирован, так же как и ряд вариантов демографического метода, на использование данных о смертности за двухлетний период и переписи, проведенной в середине того же периода.

Принимая, что

$$L_x = L_x' + \frac{1}{24}(d_{x+1} - d_{x-1}), \quad \text{где } L_x' = \frac{l_x + l_{x+1}}{2},$$

и, обозначив

$$\delta_x = \frac{1}{12}(d_{x+1} - d_{x-1}) \text{ и } \varepsilon_x = \frac{\delta_x}{q_x},$$

А. Я. Боярский получил:

$$M_x = S_{x+1}\frac{l_x - l_{x+1} + \delta_x}{l_{x+1} + l_{x+2} + \delta_{x+1}} + S_x\frac{2d_x}{l_x + l_{x+1} + \delta_x} + S_{x-1}\frac{l_x - l_{x+1} - \delta_x}{l_{x-1} + l_x + \delta_{x-1}}; \quad H_x = S_{x+1}\frac{l_x + \varepsilon_x}{l_{x+1} + l_{x+2} + \delta_{x+1}} + 2S_x\frac{l_x}{l_x + l_{x+1} + \delta_x} + S_{x-1}\frac{l_x - \varepsilon_x}{l_{x-1} + l_x + \delta_{x-1}}.$$

Величина H_x представляет собой функцию различных показателей таблиц смертности, полученных из исходных вероятностей умереть. Вероятность умереть равна: $q_x = \frac{M_x}{H_x}$. Определение q_x связано с привлечением q_0, q_1 и т. д. из прежних таблиц смертности и с использованием итерационного процесса вычисления, позволяющего постепенно уточнять значения q_x.

Практическое использование метода определяющих функций, требующего при своей реализации трудоемких расчетов, было ограничено технической возможностью — отсутствием расчетной базы. Сравнительно недавно (еще 30—40 лет назад) это обстоятельство было настолько серьезным, что предпочтение при выборе метода построения таблиц смертности отдавалось методам, которые хотя и приводили к приближенным результатам, но не требовали сложных вычислений. Приходилось мириться с неточностью вероятностных показателей.

В настоящее время электронно-вычислительная техника сняла это ограничение, увеличила свободу выбора методов построения таблиц смертности. Сейчас имеется возможность реализации метода А. Я. Боярского и получения вероятностных показателей таблиц смертности, отражающих реальную действительность с любой практически необходимой степенью точности. Вычислительные возможности гармонично дополняют теоретический подход.

При составлении таблиц смертности 1958—1959 гг. данные о возрастном составе живущих и умерших подвергались сглаживанию. «Числа живущих в тех возрастах, в которых были значительные колебания, не связанные с аккумуляцией, сглаживались методом десятилетней центрированной скользящей средней взвешенной с биномиальными коэффициентами»[1]. Для возрастов, имеющих значительную аккумуляцию, т. е. для возрастов 45—99 лет, применялся метод Б. С. Ястремского, предложенный им для обработки итогов переписи населения 1920 г. и названной им скользящими параболами. Исходным показателем является q_x, который для возрастов от 5 до 89 лет исчислялся по формуле А. Я. Боярского:

$$q_x = \frac{2(M_x + M_x')}{S_{x-1} + 2S_x + S_{x+1} + M_x + M_x' + \frac{1}{2}(M_x + M_{x+1}) - \frac{1}{2}(M_x' + M_{x-1}')},$$

где M_x и M'_x — число умерших в возрасте x лет соответственно в 1958 и 1959 гг.; S_x — число лиц в возрасте x лет по переписи 1959 г.

Вероятности q_1, q_2, q_3 и q_4 были получены отношением совокупностей умерших первого рода (умерших в определенном возрасте из числа родившихся в определенном году) к соответствующим совокупностям живущих первого рода (пережившие определенный возраст из числа родившихся в определенном году). Например,

$$q_1 = \frac{\text{число умерших в 1958 и 1959 гг. в возрасте до 1 года из поколения 1957 г.}}{\text{разность между числом родившихся в 1957 г. и числом умерших из них в 1957 и 1958 гг. в возрасте до 1 года}}.$$

Совокупности живущих первого рода для возрастов 1, 2 и 3 года были получены по данным о числе родившихся в 1955, 1956, 1957 гг., взятым из материалов разработки актов о рождении, а совокупности живущих первого рода для возраста 4 года — на основании данных переписи населения 1959 г. Показатели L_x для возрастов 1, 2, 3 и 4 года вычислялись по формуле $L_x = l_x - k_x d_x$, где k_x представляло собой отношение нижней элементарной совокупности умерших ко всей совокупности умерших первого рода. Начиная с возраста 90 лет экстраполяция проводилась по формуле Гомперца — Макегама: $\lg p_x = a + bc^x$. При этом в качестве непременного принимали условие, при котором экстраполированная кривая не уклонялась от вероятности дожития, а продолжала ее, и переход фактических данных в местах стыков в интер-

[1] *Родина Е., Дмитриева Р.* Построение таблиц смертности и средней продолжительности жизни населения СССР. — Вестник статистики, 1965, № 2, с. 24.

полированные был весьма плавным. Для определения параметров a, b и c брались значения p для трех предыдущих возрастов, образующих арифметическую прогрессию.

Интересен способ, предложенный кафедрой статистики населения Московского экономико-статистического института для вычисления вероятностей смерти по материалам переписи населения 1959 г. по РСФСР[1]. В основу способа был положен метод С. А. Новосельского, при котором привлекались данные текущего учета не за два смежных года, а лишь за один год.

Используя формулы

$$m_x = \frac{M_x}{\frac{1}{2}\,(S_x + S'_x)} \qquad q_x = \frac{2m_x}{2+m_x},$$

А. С. Семенова получила

$$q_x = \frac{4M_x}{2S_x + 2S_{x+1} + 3M_x + M_{x+1}}.$$

Для построения таблиц дожития населения Москвы расчеты производились по другой формуле:

$$q_x = \frac{2M_x}{S_x + S_{x-1} + \frac{1}{2}(M_x - M_{x-1})}.$$

IV. **Для детских возрастов** построение показателей смертности ориентировано на использование числа умерших детей, распределенных по элементарным совокупностям и совокупностям третьего рода, а также на привлечение числа родившихся за соответствующие годы (рис. 4.6). Рассмотрим формулы, по которым рассчитываются вероятности умереть для новорожденных.

Метод В. Я. Буняковского предполагает два варианта данных:

а) без элементарных совокупностей умерших:

$$q_0 = \frac{M_0}{\frac{1}{2}(N_0 + N_1)} = \frac{M_0\,(B_0A_0A_1B_1)}{\frac{1}{2}(A_0B_0 + C_0B_0)};$$

б) с элементарными совокупностями умерших:

$$q_0 = \frac{M_0(B_0A_0A_1)}{B_0A_0} + \frac{M_0(B_0A_1B_1)}{C_0B_0}.$$

Метод К. Ратса

$$q_0 = \frac{M_0}{\frac{2}{3}N_0 + \frac{1}{3}N_1} = \frac{M_0(B_0A_0A_1B_1)}{\frac{2}{3}(B_0A_0) + \frac{1}{3}(C_0B_0)}.$$

[1] См.: *Семенова А. С.* Из опыта построения советских таблиц дожития. — В кн.: Вопросы демографии. М., 1970, с. 90.

Метод Р. Бёка

$$q_0 = \frac{N_0(N_1 - C_0B_0B_1) - (N_0 - B_0A_0A_1)\cdot(N_1 - C_0B_0A_1B_1)N}{N_0(N_1 - C_0B_0B_1)}.$$

Метод Вестергарда для определения вероятностей умереть, т. е. q_x, для ранних детских возрастов (до 5 лет), использованный С. А. Новосельским при построении таблицы смертности населения Ленинграда 1938—1939 гг., состоял в замене знаменателя вероятности умереть — численности населения соответствующего возраста — на определенную, меняющуюся для каждого возраста совокупность (поколение) родившихся. Этот способ состоит в том, что умершие в данном промежутке возраста считаются принадлежащими к поколению, начало периода рождения которого отстоит от начала периода вымирания на среднюю арифметическую границу возрастного интервала.

На рис. 4.7 видно, что за поколение, к которому принадлежат умершие в возрасте 0—1 года в 1938—1939 гг., надо принять родившихся с 1/VII 1937 г. до 1/II 1939 г.; за поколение, к которому принадлежат умершие в возрасте 1—2 года, — родившихся с 1/VII 1936 г. до 1/VII 1938 г. и т. д. Вычитаем далее из числа родившихся соответствующие числа умерших до изучаемого периода 1938—1939 гг.: из числа родившихся в период с 1/VII 1936 г. до 1/VII 1938 г. — умерших в возрасте 0—1 года в 1937—1938 гг.; из числа родившихся с 1/VII 1935 г. до 1/VII 1937 г. — умерших в возрасте 1—2 года и т. д. Таким образом, получаем числитель и знаменатель вероятностей умереть.

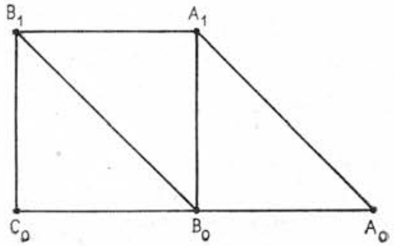

Рис. 4.6. Определение вероятности смерти для детских возрастов

Что касается показателя μ_x, то впервые его применил Ж. Б. Фурье в 1821 г., а затем Гомперц, но ни тот, ни другой не дали ему никакого названия. В 1832 г. Эдмонс назвал его на-

Рис. 4.7. Расчет показателей смертности в возрасте до 5 лет для таблиц 1938—1939 гг.:

1 — совокупности умерших в 1938—1939 гг.; *2* — совокупности (поколения) родившихся в соответствующие годы пряжением жизни. Воольгауз предложил очень удачный термин — сила смертности. Этот показатель, представляющий собой вероятность смерти в точном возрасте x лет, равный логарифмической производной функции дожития, взятой с отрицательным знаком. В. Борткевич назвал густотой вероятности смерти:

$$\mu_x = -\frac{d[l_x]}{l_x\, d_x}.$$

При гипотезе А. Муавра ($l_x = 86 - x$), получаем:

$$\frac{d[l_x]}{d_x} = -1,$$

а затем силу смертности

$$\mu_x = -\frac{1}{86-x} = \frac{1}{e_x} = \frac{d_x}{l_x} = q_x,$$

т. е. сила смертности равна вероятности смерти.

При определении силы смертности используются также и другие формулы.

Формула Воольгауза

$$\mu_x = \frac{8(l_{x-1} - l_{x+1}) - (l_{x-2} - l_{x+2})}{12 l_x}.$$

Формула Спрейга

$$\mu_x = \frac{9[5(l_{x-1} - l_{x+1}) - (l_{x-2} - l_{x+2})] + (l_{x-3} - l_{x+3})}{60 l_x}.$$

Формула Макегама

$$\mu_x = -\frac{1}{2l_x}\left(\Delta l_{x-1} - \frac{1}{24}\Delta^3 l_{x-3} + \frac{3}{640}\Delta^2 l_{x-5} - \cdots\right).$$

Используя приращения, можно также получить силу смертности

$$\mu_x = -\frac{1}{l_x}\left(\Delta l_x - \frac{1}{2}\Delta^2 l_x + \frac{1}{3}\Delta^3 l_x - \cdots\right).$$

Рассмотрим идею нахождения вероятности смерти, предложенную У. Фарром. Произведя группировку населения по возрасту на пятилетние группы (с погодовой дифференциацией первой группы) и зафиксировав фактические численности каждой возрастной группы по состоянию на критические моменты переписей 1841 и 1851 гг., У. Фарр предположил, что в межпереписной период численность населения возрастает в геометрической прогрессии со знаменателем, равным z, определяемым по численностям групп на указанные критические моменты. Далее У. Фарр предполагает, что изменение численностей возрастных групп в течение трех лет по обе стороны от дат переписей происходит по той же геометрической прогрессии с тем же знаменателем z. Используя полученные результаты, У. Фарр вычислял сначала средние вероятности смерти.

Предположим, что на начало изучаемого периода численность возрастной группы составляет L_0. Тогда через x лет ее численность составит $L_0 z$. В интервал времени dx число лет, прожитых лицами каждой возрастной группы, составит $L_0 z^x dx$. Интегрируя это выражение в пределах, например, от 0 до 17 лет, получим число лет, прожитое населением каждой возрастной группы за 17 лет (L):

$$L = \int_0^{17} L_0 z^x dx = \frac{L_0}{\ln z}(z^{17} - 1).$$

Привлекая к расчету число умерших за 17 лет (D) и находя отношение $\frac{D}{L}$, получаем *среднюю вероятность смерти*, относящуюся к среднему значению возраста в каждом возрастном интервале:

$$\bar{p} = \frac{D}{L}.$$

При вычислении вероятности дожития до следующего года для лиц, достигших средних значений возраста в возрастных интервалах, У. Фарр использовал формулу Эдмонса:

$$y = e^{\frac{m_a}{\ln z}(1 - z^x)},$$

где m_a — средняя вероятность смерти в течение одного года; a — возраст; y — число лиц, достигающих возраста $a+x$ лет; z — знаменатель прогрессии.

Примем число лиц в возрасте a лет равным единице. Вероятность смерти будет расти с увеличением возраста в геометрической прогрессии. Если $x=1$, то $y = e^{\frac{m_a}{\ln z}(1 - z^x)}$. Это число лиц, достигших возраста $a+1$ год, из числа лиц, состоявших в возрасте a, т. е. проживших один год.

Но так как число лиц в возрасте a лет принято за 1, то y будет означать вероятность для лиц в возрасте a лет прожить один год. Используя найденные по фактическим данным m_a и z, У. Фарр определял y и построил свою таблицу смертности.

Анализ методов построения таблиц смертности показывает, что изучение населения ведется как в «продольном», так и в «поперечном» разрезе.

При рассмотрении населения в *«продольном»*, или *вертикальном*, разрезе имеется в виду протяженность жизни поколения или жизнь некоторого среднего человека из одного поколения (сверстников). В этом случае регистрируются события в разные периоды жизни людей одного реального поколения. Когда наблюдению подвергается одно и то же поколение и фиксируются события на каждом этапе его жизни, то такое продольное наблюдение называют текущим. Если же изучаются события, имевшие место в прошлом этого поколения, наблюдение называют ретроспективным (анамнестическим).

При изучении интенсивности демографических явлений (брачности, рождаемости, смертности и т. д.) требуется знание последовательности их изменения на всех этапах жизни конкретного поколения. Такое изучение демографических показателей реальных поколений называют еще когортным методом. Когда говорят о демографической судьбе поколений, имеют в виду «продольный» разрез.

При изучении населения в *«поперечном» (по изохроне)* разрезе особое значение приобретает возрастная структура населения: анализу подвергается один год жизни всего населения. Иными словами, изучаются события, происходящие в один период у людей, живущих в данный момент и относящихся к разным поколениям. В данном случае мы имеем дело с фиктивным, или воображаемым, иначе говоря, условным поколением (современников). Такой метод изучения называют методом *гипотетического поколения,* основанным на реальных показателях разных поколений, но приуроченных к одному времени. В этом случае получаем моментный снимок населения.

При «продольном» долгосрочном изучении населения исследователь имеет дело со средней последовательностью. При «поперечном» сечении населения изучается последовательность средних. При демографическом анализе показатели гипотетического поколения переносят на реальные. Оценивая достоинства и недостатки двух методов построения таблиц смертности — когортного метода и метода гипотетического поколения, — следует, учитывая значительно большую зависимость смертности от современных факторов, чем от факторов, характеризующих поколение, отдать предпочтение методу, отражающему демографическую и социально-экономическую ситуацию данного конкретного периода времени, т. е. методу гипотетического поколения.

Что касается демографических таблиц, построенных когортным методом, то они могут найти применение в ряде случаев. Так, например, в послевоенные годы в некоторых странах с низкой смертностью наблюдается парадоксальное явление. Наряду с прогрессом здравоохранения и медицины и достигнутым в силу этого значительным общим снижением смертности в старших возрастах (более 45—50 лет) снижение смертности весьма малое, а иногда наблюдается даже и повышение смертности. Это дало основание некоторым исследователям предположить, что ввиду резкого снижения детской смертности и выживания слабых организмов происходит подавление очистительного действия естественного отбора. Иными словами, предполагается, что на смертность старших возрастов влияет история поколений. Вот почему был поставлен вопрос о необходимости наряду с гипотетическими таблицами смертности строить таблицы смертности на основе когорт, т. е. поколений.

Так, Ж. Легаре на основе построенных им таблиц смертности методом поколений пришел к выводу о том, что «...для достаточно старых поколений снижение смертности вызвано в большей части успехами медицины, а не улучшением биологических условий»[1]. Далее Ж. Легаре продолжает: «...прогресс становится все более и более трудным, он может только отсрочить смерть до какого-то предельного возраста, который еще не смогли отодвинуть»[2]. В заключение, однако, автор указывает, что ввиду малого времени, прошедшего после обнаружения этого парадокса, точного ответа на вопрос о том, ведет ли отсутствие естественного отбора к ослаблению населения, ни априори, ни апостериори получить не удалось.

4.3. Вероятностный метод взаимного контроля данных переписи и текущего учета

При построении таблиц смертности нужна уверенность в достаточной точности привлекаемых к расчету данных. Для проверки первичного материала можно использовать вероятностный метод взаимного контроля данных переписи и текущего учета, разработанный А. Я. Боярским и основанный на сопоставлении численности детей младшего возраста по переписи с соответствующими данными текущего учета.

Рассмотрим суть этой проверки на примере возраста 0 лет. Пусть численность детей возраста 0 лет на момент переписи составляет A_0A_1 (рис. 4.6). По данным текущего учета зарегистрировано рождений A_0B_0. По статистике смертности выделена элементарная совокупность умерших $A_0B_0A_1$, т. е. детей, родившихся и умерших в году, непосредственно предшествовавшем переписи. При абсолютной точности данных и отсутствии миграции должно соблюдаться равенство:

$$A_0A_1 = A_0B_0 - A_0B_0A_1.$$

Обстоятельством, затрудняющим проверку, А. Я. Боярский считает аккумуляцию на возрасте 1 год, из-за которой число детей в возрасте 0 лет занижается. Проверка первичного материала состоит в выявлении расхождения между величиной A_0A_1 и разностью $A_0B_0 - A_0B_0A_1$. Устанавливается коэффициент расхождения $K_р$:

$$K_р = \frac{A_0A_1}{A_0B_0 - A_0B_0A_1}.$$

---

[1] *Легаре Ж.* Некоторые соображения по поводу таблицы смертности поколения (на примере Англии и Уэлса). — В кн.: Демография поколений. М., 1972, с. 143.

[2] *Там же,* с. 148.

Для оценки отклонения K_p от единицы А. Я. Боярский предлагает проведение специальных выборочных обследований, целью которых является индивидуальная поименная проверка трех предварительно составленных списков:

зарегистрированных переписью детей в возрасте до 1 года;

родившихся в предшествующем году;

умерших в предшествующем году детей в возрасте 1 год из числа родившихся в том же году.

При сверке списков общая численность всех зарегистрированных переписью детей до 1 года (S_0) разделится на две части: имеющиеся и не имеющиеся в списке родившихся. Общее число всех родившихся (N) будет состоять из:

значащихся по переписи в группе 0 лет — $S_{0р}$;

не значащихся по переписи в группе 0 лет — $S_{0н}$;

значащихся по переписи в группе 1 год — $S_{1р}$;

значащихся в списке умерших в возрасте 0 лет — $N_у$;

недоучтенных переписью — $N_{нн}$;

недоучтенных при регистрации умерших — $N_{ну}$;

умерших, не учтенных в регистрации родившихся, — $M_{нр}$.

Тогда получаем:

$$S_0 = S_{0р} + S_{0н}\,; \quad N = S_{0р} + S_{1р} + N_у + N_{нн} + N_{ну}\,;$$
$$M = N_у + M_{нр}\,.$$

Вводятся вероятности:

вероятность регистрации ребенка при его рождении — p;

вероятность регистрации смерти для зарегистрированного ребенка — $q_р$;

вероятность смерти на первом году жизни до конца календарного года рождения — q_0;

вероятность регистрации смерти для ребенка, не зарегистрированного при рождении, — $q_{нр}$;

вероятность учета зарегистрированного при рождении ребенка переписью — $r_р$;

вероятность регистрации в переписи ребенка, рождение которого не было зарегистрировано, — $r_{нр}$.

С учетом введенных вероятностей получим:

$$N_у + N_{ну} = Nq_0; \quad S_{0р} + S_{1р} = N(1-q_0)\,r_р\,.$$

Откуда

$$r_р = \frac{S_{0р}+S_{1р}}{S_{0р}+S_{1р}+N_{нн}}, \text{ а}$$

$$q_0 = \frac{N_у}{N_у + N_{ну}}\,.$$

Используя теорему Бейеса, получаем:

$$\frac{S_{0р}}{S_0} = \frac{pr_р}{pr_р + (1-p)\,r_{нр}}\,.$$

$$\frac{N_у}{M} = \frac{pq_р}{pq_р + (1-p)\,q_{нр}}\,.$$

Итак, для определения пяти неизвестных p, $q_р$, $q_{нр}$, $r_р$, $r_{нр}$ имеем только 4 уравнения.

А. Я. Боярский делает предположение, по его мнению, достаточно близкое к истине, о равенстве $r_р$ и $r_{нр}$. Заменяя эти две вероятности одной — вероятностью вообще учета ребенка в переписи r — и производя различные преобразования, он получает:

$$p = \frac{S_{0р}}{S_0}; \quad q = p\,\frac{M}{N_у + N_{ну}}\,;$$

$$r = \frac{S_{0р}+S_{1р}}{S_{0р}+S_{1р}+N_{нн}}\,.$$

Используя эти три вероятности, можно найти искомые поправки. После исправления число родившихся примет вид:

$$\frac{N}{P} = \frac{NS_0}{S_{0р}}\,.$$

Число умерших

$$\frac{M}{q} = \frac{S_0}{S_{0р}}(N_у + N_{ну})\,.$$

Разность между ними составит:

$$\frac{N}{p} - \frac{M}{q} = \frac{S_0}{S_{0р}}(S_{0р} + S_{1р} + N_{нн})\,.$$

Число детей до 1 года по переписи составит:

$$\frac{S_0}{r} = \frac{S_0}{S_{0р}+S_{1р}}(S_{0р} + S_{1р} + N_{нн})\,.$$

Принимая, что дробь $\frac{S_0}{S_{0р}+S_{1р}}$ отражает вероятность аккумуляции на возрасте 1 год, находим, что действительное число детей до 1 года в переписи равно:

$$\frac{S_0}{r} : \frac{S_{0р}}{S_{0р}+S_{1р}} = \frac{S_0}{S_{0р}}\,(S_{0р} + S_{1р} + N_{нн})\,.$$

Теперь уже линия A_0A_1 дает в точности разность A_0B_0— —$A_0B_0A_1$.

4.4. Логиты и их использование

При построении модельных таблиц смертности и группировке стран на четыре группы по признаку средней продолжительности жизни (до 45 лет, 45—55, 55—65, свыше 65 лет) демографы ООН получали повозрастные вероятности смерти каждой группы путем осреднения (без взвешивания) повозрастных вероятностей смерти всех стран, попавших в состав группы. При этом оказалось, что отношение усредненных повозрастных вероятностей умереть (C) в группах стран меняется весьма сложно, приближаясь в старших возрастах к единице. Выяснилось также, что уменьшение различия между усредненными повозрастными вероятностями умереть в разных группах стран и приведение соотношений между ними к более простому виду может быть достигнуто привлечением весьма распространенных в биологии функций, представляющих натуральный логарифм отношения вероятности к ее дополнению до единицы, т. е. $\ln \frac{p}{1-p}$.

Эта функция, взятая с коэффициентом $^1/_2$, называемая логитом (logit) от p, использовалась ранее для измерения силы действия лекарств при лечении животных и выражалась формулой

$$v=\operatorname{logit} p=\frac{1}{2}\ln\frac{p}{1-p}=\frac{1}{2}\ln\frac{p}{q}.$$

При этом $\operatorname{logit} p=-\operatorname{logit} q$.

Тогда логит от $(1-l_x)$ можно записать так:

$$v_x=\operatorname{logit}(1-l_x)=\frac{1}{2}\ln\left(\frac{1-l_x}{l_x}\right).$$

При описании таблиц смертности была сделана попытка заменить числа доживающих разных таблиц смертности их логитами[1].

Для смягчения влияния перехода C от величины, значительно отличающейся от единицы, к единице в демографии был разработан метод оценки величины отношения C в младших, средних и старших возрастах. Это привело к использованию логитов

$$\ln\left(\frac{1-l_x}{l_x}\right).$$

Дифференцируя это выражение, получаем:

$$-\frac{l'_x}{l_x}\cdot\frac{1}{1-l_x}.$$

Первый множитель представляет собой относительное изменение числа доживающих, приведенное к одному году, и равен силе смертности μ_x. В результате получаем следующее соотношение показателей двух таблиц смертности:

$$\mu_{1(x)}\frac{1}{1-l_{1(x)}}=C\mu_{2(x)}\frac{1}{1-l_{2(x)}}.$$

Важным свойством полученного выражения является то, что при переходе l_x от 1 к 0 оно дает возможность приблизить C к постоянной величине. Так, при l_x, мало отличающемся от 1 (младшие возраста), множители $\frac{1}{1-l_{1(x)}}$ и $\frac{1}{1-l_{2(x)}}$ приближают значения $\mu_{1(x)}$ и $\mu_{2(x)}$ друг к другу. При значениях l_x, находящихся в интервале от 1 до 0 и более близких к 0,5 (средние возрасты), действие указанных множителей на $\mu_{1(x)}$ и $\mu_{2(x)}$ уменьшается. Наконец, при значениях l_x, мало отличающихся от 0 (старшие возраста), величина $1-l_x$ близка к единице, и мы приходим к следующему соотношению:

$$\mu_{1(x)}\approx C\mu_{2(x)}.$$

Использование логитов позволило привести довольно сложные соотношения между числами доживающих l_x к очень простому виду. Оказалось, что разности логитов усредненных таблиц смертности для последовательности возрастов меняются медленно и практически почти постоянны во всех возрастах. В качестве примера приведем табл. 4.1.

Таблица 4.1. **Логиты двух усредненных таблиц смертности (A и B) и их разности**

Возраст, лет	l_x		Логиты		Разность логитов
	A	B	A	B	$B-A$
1	0,9624	0,9625	—1,6216	—1,2670	0,3546
5	0,9535	0,9005	—1,5099	—1,1015	0,4084
20	0,9392	0,8710	—1,3684	—0,9551	0,4133
30	0,9208	0,8362	—1,2264	—0,8150	0,4114
40	0,8973	0,7963	—1,0836	—0,6816	0,4020
60	0,7666	0,6411	—0,5945	—0,2900	0,3045
65	0,6926	0,5657	—0,4062	—0,1316	0,2746
70	0,5897	0,4663	—0,1814	0,0676	0,2490
75	0,4537	0,3429	0,0928	0,3252	0,2324
80	0,2948	0,2094	0,4360	0,6642	0,2282

У. Брассом был замечен эмпирический факт, открывающий возможности прогнозирования и построения модельных таблиц смертности и состоящий в том, что связь логитов чисел доживающих, полученных из различных таблиц смертности, может быть аппроксимирована линейным уравнением, связывающим логиты вероятностей смерти:

$$v_{1(x)}=a_0+a_1v_{2(x)},$$

где a_0 и a_1 — параметры, определяемые методом наименьших квадратов.

4.5. Кривая вероятностей смерти

В. Лексис в 1875 г. обратил внимание на то, что конец кривой смертности d_x (рис. 4.8) по внешнему виду напоминает правую половину симметричной кривой вероятности. Сама кривая ведет себя следующим образом: она начинается очень высоко, быстро опускается до минимума, соответствующего 14 годам, затем медленно поднимается и достигает максимума к 71 году, после чего быстро опускается и сливается с осью возраста — к 106 годам.

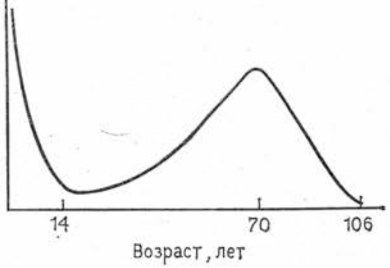

Рис. 4.8. Кривая смертности

Предположение К. Пирсона состояло в том, что кривая смерт-

[1] См.: *Брасс У.* Об одном способе выражения закономерностей смертности. — В кн.: Изучение продолжительности жизни. М., 1977, с. 39—93.

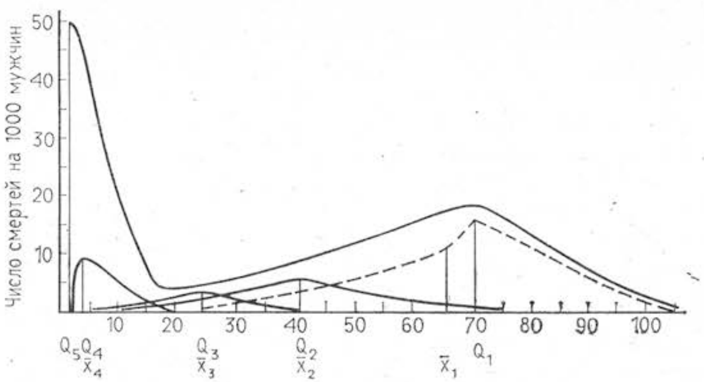

Рис. 4.9. Кривая смертности мужского населения (по данным английской статистики)

ности есть сложная линия, составленная складыванием ординат пяти более простых вероятностей. Для каждого из пяти возрастов: младенческого, детского, юношеского, зрелого (среднего), старческого — имеются свои причины смертности, и каждая из этих групп возрастов может быть выражена своей кривой. Сложением ординат всех пяти кривых вероятностей смерти получается первоначальная — общая кривая смерти.

На рис. 4.9 проведено 5 вертикальных линий — модальных значений вероятностей смерти каждой группы возрастов. При симметричности этих распределений модальные возраста совпадают со средними возрастами.

К. Пирсон очень образно описал путь толпы, состоящей из 1000 «человеческих существ мужского пола, зачатых одновременно и проходящих медленным шагом по жизненному пути». На небольшом расстоянии от этого пути в местах пяти наибольших вертикалей чертежа Q_5, Q_4, Q_3, Q_2 и Q_1 расположено пять стрелков, обладающих различной меткостью, стреляющих безостановочно в толпу проходящих (каждый в свою определенную цель) из оружия различной смертоносности. Следует ожидать, что больше всего будет убито в момент прохождения толпы напротив самого центра цели. Кроме того, учитывая разбросы около центра, некоторые выстрелы будут поражать толпу по сторонам цели. Толпа шествующих будет редеть в различных возрастах то медленными, то быстрыми темпами. На фоне этой общей картины К. Пирсон рассмотрел каждую из вероятностных кривых, начиная с группы старческого возраста.

Кривая *старческого возраста* в левой части полого склоняется к нижней горизонтальной линии и пересекает ее в очень молодом возрасте, а в правой части линия смертности опускается более круто. Стрелок, олицетворяющий все причины смерти этой группы (свойства организма, болезни, несчастные случаи и т. д.),

целит в проходящую перед ним толпу в возрасте, примерно равном 71 году (Q_1), но попадает из-за разбросанности результатов не только в более старых, но и более молодых — от 14-летних до 71-летних. Средний возраст старческой группы равен 66 годам. Преимущественное действие смертности в этой группе сосредоточено на возрастах 54—80 лет.

Кривые для *зрелого (среднего) возраста и юношей* очень симметричны. Максимум смертности для зрелого (среднего) возраста приходится на 42 года (Q_2) и охватывает возрастной интервал от 4 до 80 лет.

Наибольшая скученность смертности для *юношеского возраста* приходится на 23-й год жизни (Q_3). Разброс смертных случаев приходится на интервал от 4 до 45 лет.

Кривая смертности *детского возраста* резко асимметрична. Максимум смертности в интервале от 2 до 20 лет выпадает на 3-й год жизни (Q_4), а средний уровень смертности относится к 6 годам.

Из смертности детского возраста К. Пирсон выделил смертность в *младенческом* возрасте. Такое расчленение объяснялось тем, что суммирование всех уровней смертности, составляющих кривые вероятностей четырех возрастных распределений, при отсчете возраста наблюдаемых 1000 человек с начала года рождения не укладывались в общую линию смертности. Лишь в том случае, когда отсчет ведется не с года рождения, а девятью месяцами раньше, с момента зачатия, результаты суммирования стали приводить к ординатам общей кривой линии смертности.

Оценивая попытку К. Пирсона свести смертность всего населения к вероятностям смерти пяти возрастных групп, советский демограф Б. Ц. Урланис охарактеризовал ее как неудачную[1].

На наш взгляд, эта попытка К. Пирсона действительно является, хотя и весьма образной, не вполне удачной. В качестве недостатка построения К. Пирсона следует отметить тот факт, что вероятная смертность каждой из указанных выше пяти возрастных групп должна аппроксимироваться распределениями, отличающимися друг от друга различной степенью ассимметрии, не совпадающими для разных социальных групп, полов, регионов и т. д. и не остающимися одинаковыми во времени. Очевидно, при поиске таких форм распределений, которые улавливают особенности смертности каждой возрастной группы и поэтому пригодны для аппроксимации, надо «согласовывать» полученные по этим распределениям погодовые результаты со вторым условием — совпадением суммы этих погодовых вероятностей смерти возрастных распределений с общей смертностью. При несовпадении следует заменять предлагаемое для той или иной возрастной группы распределение другим. При такой замене возрастных распределений, представляющей по сути дела процесс их «подгонки», без больших натяжек не обойтись.

[1] См.: *Урланис Б. Ц.* Эволюция продолжительности жизни. М., 1978, с. 240.

9 Заказ 2715

4.6. Использование вероятностных показателей таблиц смертности для различных целей

Накладывая на данные переписи таблицу смертности, методом передвижки возрастов можно получить возрастной состав населения на перспективу. Расчетные данные расходятся с фактическим возрастным составом населения по последующей переписи. Эти расхождения исследователи прежде всего объясняют тем, что фактическая смертность в течение межпереписного периода не вполне соответствует показателям таблиц смертности, положенным в основу расчета, а также несовершенством регистрации рождений.

Наблюдая, с одной стороны, снижение смертности от более ранних периодов к поздним, а с другой — увеличение смертности от младших возрастных групп к старшим, можно сделать вывод о зависимости смертности от принадлежности к тому или иному поколению, связанному с периодом рождения. Выяснилось, что чем старше определенное поколение людей, тем в большей степени его фактическая смертность отличается от показателей таблицы смертности, относящейся к году рождения этого поколения. Для сопоставлений может быть использован особый показатель, называемый относительным значением вероятности умереть (${}^{b}_{a}u$), вычисляемый для каждой возрастной группы по формуле

$$ {}^{b}_{a}u = \frac{{}^{a}q - {}^{b}q}{{}^{a}q}, $$

где ${}^{a}q$ и ${}^{b}q$ — вероятности умереть для поколений или периодов a и b.

Это, в частности, означает необходимость внесения поправок в вероятностные показатели старших поколений таблиц смертности при их использовании в страховых расчетах. Следовательно, есть необходимость кроме обычных таблиц смертности строить также таблицы смертности по поколениям (когортам).

Из таблиц смертности (или аналогичных данных об уровне фактической смертности) видно, что вероятности смерти (или фактические ее уровни) у мужчин во всех возрастах выше, чем у женщин. Разности возрастов мужчин и женщин, имеющих одинаковые вероятности или фактические уровни смертности, будут указывать на величину опережения смертности у мужчин. При этом каждый возраст будет характеризоваться своей величиной такого рода лага. Напрашивается вопрос: как охарактеризовать опережение мужской смертности по сравнению с женской не для каждого возраста, а в целом одним числом? Очевидно, что средняя арифметическая из найденных лагов для каждого возраста сможет охарактеризовать величину опережения в целом.

Можно использовать для этой цели и другой прием, разработанный математической статистикой и состоящий в коррелировании повозрастных вероятностей смерти или фактических ее уровней у мужчин и женщин, постепенном сдвиге данных о мужчинах относительно женщин и совмещении возраста мужчин и женщин с различным сдвигом — на один, на два года и т. д. Наибольшая величина из полученных при таком сдвиге коэффициентов корреляции укажет на величину лага.

Об имеющихся резервах повышения средней продолжительности жизни населения может дать представление построение такой условной таблицы смертности, в которой каждая из фактических повозрастных вероятностей умереть (q_x) заменена минимальными повозрастными вероятностями смерти (q_x^{min}), взятыми из официальных таблиц смертности различных стран.

Производя аналогичный расчет с привлечением минимальных коэффициентов смертности (а не вероятностей умереть), М. С. Бедный получил в 1972 г. среднюю продолжительность жизни для мужчин 72, 61, а для женщин — 78,41 года. Наши расчеты с использованием минимальных повозрастных вероятностей умереть в каждом возрастном интервале, взятых из современных таблиц смертности, дали соответственно 74,3 и 79,6 года.

Вопросы долголетия населения уже давно привлекают внимание демографов и геронтологов. Обычно о долголетии населения судят по данным о возрастном распределении населения, полученным по переписи. В качестве измерителя долголетия долгое время использовался удельный вес лиц 80 лет и старше во всем населении. Недостатком этого до сих пор широко применяемого показателя является его зависимость от уровня рождаемости. Чем относительно меньше в составе населения численность молодых возрастов (при снижении рождаемости), тем выше, при прочих равных условиях, показатель долголетия.

Критикуя методическую неправильность этого показателя долголетия населения, Н. Н. Сачук предложила другой показатель — удельный вес лиц 80 лет и старше в населении 60 лет и старше, а не во всем населении:

$$ I' = \sum_{80}^{100} S_x : \sum_{60}^{100} S_x. $$

С мнением Н. Н. Сачук о фиксации внимания на возрастах 80 и 60 лет согласился А. М. Мерков, но предложил считать этот показатель не по фактическим переписным данным, а по числам доживающих (l_x) из таблиц смертности[1]. В соответствии с этим предложением вычисление вероятностных показателей долголетия, свободных от привходящего влияния различных факторов, производится по формуле

$$ q = l_{80} : l_{60}. $$

[1] См.: *Мерков А. М.* Методические вопросы санитарно-статистического изучения долголетия. — В кн.: Вопросы санитарной и медицинской статистики, М., 1971, с. 37.

Большое значение имеет сопоставление вероятностных показателей, относящихся к различным совокупностям, моментам и периодам времени. Этой проблемой занимались многие исследователи. В качестве примера укажем статью Р. М. Дмитриевой и Е. М. Андреева, сопоставляющих вероятности умереть мужчин, женщин и всего населения (q_x), взятые из таблиц смертности населения, приуроченных к годам всеобщих переписей населения России и СССР: 1896—1897 гг., 1926—1927 гг., 1938—1939 гг., 1958—1959 гг.

На основе использования повозрастных вероятностей смерти авторы пришли к выводу о том, что в СССР происходит снижение смертности, начавшееся еще в дореволюционный период и резко ускорившееся после Великой Октябрьской социалистической революции. Причины столь резкого снижения смертности авторы видят в социально-экономических мероприятиях Советской власти, направленных на повышение благосостояния населения и «...улучшение материальных и социальных условий его жизни, включая рост обеспеченности продуктами питания, повышение комфорта жилищ, развитие коммунального хозяйства и т. п.»[1].

При изучении эффективности применяемых методов лечения больных, находящихся под длительным наблюдением (рак, туберкулез, язвенные болезни, гипертония и т. д.), а также для анализа летальности (отношения числа умерших от определенной болезни к числу заболевших ею) Л. С. Каминский использовал вероятностный подход к построению обычных таблиц смертности и построил таблицу смертности больных. Обычно применяемый показатель летальности затушевывал различия возрастной структуры больных. Таблицы смертности больных позволяют устранить этот недостаток и построить рациональный и методологически обоснованный показатель, характеризующий эффективность лечения одним числом. Таким весьма характерным показателем, по мнению Л. С. Каминского, является величина продолжительности предстоящей жизни сравниваемых групп больных при различных методах лечения. Для вычисления этого показателя автор предложил заменить числа живущих обычных таблиц смертности числом лиц, поступивших под наблюдение, а совокупности умерших — числом умерших в том же году.

Идея Л. С. Каминского может быть реализована следующим образом:

1) лечебное учреждение фиксирует взятых под наблюдение лечащихся больных по годам (однородность в отношении пола, возраста, характера болезни достигается комбинированной группировкой);

2) по каждой группе больных определяют по годам числа умирающих и числа оставшихся в живых по истечении года лечения, на начало следующего года;

3) обозначая суммарное число больных, взятых под наблюдение за все годы, l_0 и приведя его к круглому числу, например 1000, можно определять число оставшихся в живых на конец каждого года наблюдения; найденные числа взятых под наблюдение и оставшихся в живых позволяют найти вероятность остаться в живых для лечившихся один год, два года и т. д., т. е. p_x (где x — число лет лечения);

4) из формулы $p_x+q_x=1$ находится вероятность умереть на том или ином году лечения — q_x;

5) все остальные показатели таблиц смертности больных получаются аналогично тому, как это делается в обычных таблицах смертности: $l_1=l_0p_0$; $l_2=l_0p_0p_1$; $l_3=l_0p_0p_1p_2$ и т. д.; $d_x=l_xq_x$; $l_x-l_{x+1}=d_x$ и т. д.;

6) при расчете средней продолжительности жизни лечащихся больных (e_x) исходят из равномерности распределения умирающих в течение года, т. е. в предположении, что каждый умерший в год своей смерти прожил в среднем полгода; тогда, вычисляя частное от деления числа прожитых всеми больными лет на число поступивших под наблюдение, мы определим среднее число лет, которое при данном уровне летальности предстоит прожить в среднем одному больному из первоначальной совокупности лиц, поступивших под наблюдение.

Многие демографы посвятили свои исследования изучению смертности по причинам. Принципиальная возможность построения таблиц смертности по причинам была доказана У. Фарром. В таблицах смертности показатель q_x — это вероятность умереть в промежутке возраста от x до $x+1$ лет. Исследователя может заинтересовать не просто вероятность смерти в возрастном интервале, а смертность в этом интервале от какой-то конкретной причины.

Д. Бернулли заменил комбинаторные методы решения задач, относящихся к смертности и продолжительности жизни населения, и попытался статистическим путем с привлечением вероятностных схем выявить значение прививок оспы для уменьшения смертности и увеличения средней продолжительности жизни населения. Особое внимание к заболеванию оспой объяснялось тем, что эта болезнь, по мнению Д. Бернулли, уносила $^1/_{64}$ часть населения ежегодно. Инокуляция (прививка оспы от больного человека здоровому) исключала оспу, но с малой вероятностью приводила к смертельному исходу. Д. Бернулли, рассчитав относительное число людей, не болевших оспой, построил таблицу смертности для инокулированного населения, продолжительность жизни которого оказалась увеличенной на 3 года 2 месяца.

Прямой подход к изучению и измерению эффекта от устранения различных причин смерти и определению увеличения продолжительности жизни при устранении тех или иных болезней впервые был сформулирован Макегамом. Одной из первых таб-

[1] *Дмитриева Р. М., Андреев Е. М.* Снижение смертности в СССР за годы Советской власти. — В кн.: Брачность, рождаемость, смертность в России и в СССР/ Под ред. А. Г. Вишневского. М., 1977, с. 48.

лиц смертности по причинам была таблица, построенная Л. Дублиным и А. Лоткой. В нашей стране этому вопросу посвятили работы В. В. Паевский, А. М. Мерков, А. Я. Боярский, М. С. Бедный и др.

Покажем вероятностный характер расчетов, связанных с выявлением резервов повышения средней продолжительности жизни.

Известна формула исчисления вероятности смерти при исключении какой-нибудь конкретной причины (q'_x), исходящая из того, что лица, болевшие, но не умершие от данной причины, выйдут из-под наблюдения и не будут влиять на уменьшение смертности от других причин:

$$q'_x = \frac{1-\sqrt{1-4x(1-w)\,q_x}}{2(1-w)},$$

где q_x — вероятность смерти от всех причин; w — доля смертности от данной причины.

Более простой является формула А. Я. Боярского, хотя и исходящая из тех же предположений, но выведенная из соотношения между силой смертности и вероятностью смерти; она имеет вид[1]:

$$q'_x = \frac{(1-w)q_x}{1-0{,}5q_x w}.$$

А. М. Мерковым и Р. Н. Бирюковой описываются и другие методы построения таблиц смертности, основывающиеся на использовании обычных таблиц смертности. А. М. Мерков предложил уменьшать d_x на долю смертей от определенной причины, Р. Н. Бирюкова уменьшала q_x на такой процент, на который уменьшается смертность в каждом возрасте при полной ликвидации смерти от этой причины. М. С. Бедный пошел другим путем и разработал специальную методику, позволяющую установить точную меру в изменении показателей таблиц смертности в зависимости от фактического снижения смертности от отдельных причин за определенный отрезок времени[2]. При построении таблиц смертности, дифференцированных по причинам, М. С. Бедный, предположив, что число доживающих до возраста x лет равно l_x, нашел вероятность умереть от всей совокупности причин смерти и от каждой причины в отдельности.

Если d_x^1, d_x^2, d_x^3 и т. д. — число умерших в возрасте x лет от каждой отдельной причины, то общее число умерших во всех возрастах от всех причин составит $d_x=\sum_{i=1}^{n} d_i$. Тогда вероятности смерти от отдельных причин будут равны:

$$\frac{d_x^i}{l_x}=q_x^i \quad \text{и} \quad \frac{\sum_{i=1}^{n} d_i}{l_x}=q_x .$$

Приведенные выше вероятности смерти от отдельных причин М. С. Бедный назвал *зависимыми парциальными вероятностями*. По его мнению, эти вероятности не выявляют истинных взаимосвязей между причинами смерти. Для устранения недостатка этих вероятностей М. С. Бедный предлагает число l_x уменьшать на число тех, которые умерли от прочих причин ($d_x - d^i_x$) в течение полупериода, т. е. на $\frac{d_x-d_x^i}{2}$. Тогда *независимые парциальные вероятности* смерти от отдельной причины будут равны:

$$\overline{q^i}_x = \frac{d^i_x}{l_x-\frac{d_x-d_x^i}{2}} .$$

По этим вероятностям М. С. Бедный построил таблицы смертности.

Некоторые исследователи считали, что существуют особенные возраста, опасные для здоровья человека. Это возраста, кратные 7 и 9. Особенно опасными, критическими считались такие возраста человека, как 49 и 63, якобы имеющие большое влияние на человеческую судьбу. Из таблиц смертности видно, что вероятность смерти в этих возрастах ничем не выделяется из остальных вероятностей.

Теория вероятностей с привлечением таблиц смертности может помочь нам развеять один предрассудок (поверье), состоящий в том, что если за стол сядут 13 человек, то в ближайший же год кто-то один из них обязательно умрет.

Чтобы определить вероятность смерти хотя бы одного из лиц, нужно было бы, привлекая таблицы смертности, найти вероятности дожития каждого из них до следующего года жизни, затем по теореме умножения вероятностей найти общую вероятность дожития всех 13 лиц до следующего года жизни. Вычитая из единицы найденную общую вероятность дожития всех 13 лиц до следующего года жизни, можно определить искомую вероятность смерти хотя бы одного из 13 конкретных лиц. Величина этой вероятности позволила бы сделать определенные выводы по рассматриваемому вопросу. Но этот путь закрыт, так как возраста каждого из 13 лиц, сидящих за столом, нам не известны. Попробуем найти обходный путь, позволяющий абстрагироваться от действительных возрастов указанных лиц и дающий возможность ответить на вопрос, каким должен быть *средний возраст всех 13 человек* ($\overline{x}=\frac{\Sigma x}{13}$), чтобы с достаточно большой ве-

[1] См.: *Боярский А. Я.* Население и методы его изучения. М., 1975, с. 290.
[2] См.: *Бедный М. С.* Продолжительность жизни. М., 1967, с. 127.

роятностью, т. е. почти наверняка, можно было утверждать, что кто-либо из них умрет в ближайшем году. В качестве такой вероятности примем величину 0,97. В этом случае вероятность дожития всем 13 лицам должна быть крайне незначительной величиной $P_x=1-0{,}97=0{,}03\approx\frac{1}{2^5}$. Если принять вероятность одному лицу в среднем возрасте x лет прожить еще один год за p_x, то вероятность всем 13 лицам прожить один год примет вид: $P_x=p_x^{13}$. Следовательно, вероятность того, что хотя бы одно лицо из 13 умрет в ближайшем году, определяется формулой $q_x=1-P_x$. Имеем:

$$p_x^{13}=\frac{1}{2^5};\quad p_x=\sqrt[13]{2^{-5}}=2^{-5,13}\approx0{,}766.$$

Тогда $q_x=1-0{,}766=0{,}234$.

Используя таблицу смертности и средней продолжительности жизни 1958—1959 гг.[1], видим, что такая величина вероятности смерти отвечает возрасту 100 лет. Следовательно, только в том случае, если за столом собирается общество из 13 человек, средний возраст которых более 100 лет, можно допустить с вероятностью 0,97, что один из них умрет в течение ближайшего года.

Может быть, мы слишком завысили требование к вероятности смерти? Произведем еще один расчет. Пусть вероятность того, что одно из лиц, обладающих средним возрастом x лет, умрет в течение ближайшего года, равна не 0,97, как раньше, а значительно меньше, скажем 0,75. Тогда вероятность дожития всем 13 лицам равна 0,25; значит, $P_x=0{,}25=(1/2)^2$. Отсюда

$$p_x^{13}=\frac{1}{2^2};\ p_x=\sqrt[13]{2^{-2}}=2^{-2/13}=2^{-0{,}1538}\approx0{,}90.$$

Тогда $q_x=1-0{,}90=0{,}10$.

В таблице смертности это отвечает возрасту 83 года. Вывод тот же.

В демографической литературе вопросы воздействия динамики смертности на экономику давно уже привлекли внимание исследователей[2].

Интересную методику для расчета вероятного резерва национального дохода в зависимости от изменения повозрастных показателей смертности разработал А. А. Ткаченко[3]. Имеется в виду снижение потери национального дохода, происходящее вследствие изменения смертности трудоспособного населения (мужчины в возрасте 15—59 лет, а женщины — 15—54 года).

[1] См.: Итоги Всесоюзной переписи населения 1959 года. СССР (Сводный том). М., 1962, с. 263.

[2] См., например: *Паевский В. В.* Вопросы демографической и медицинской статистики, с. 357.

[3] См.: *Ткаченко А. А.* Экономические последствия современных демографических процессов в СССР. М., 1978, с. 127.

Используются вероятностные показатели таблиц смертности. Автор предложил два варианта расчета. Укажем, например, на упрощенный вариант:

$$\Delta D=\frac{D}{W_i}\left(\sum_i W_m x_i q_{m_i}+\sum_i W_f x_i q_{f_i}\right),$$

где D — произведенный в определенном году национальный доход в текущих ценах; ΔD — вероятный резерв национального дохода, т. е. величина национального дохода, которая могла бы быть произведена той частью трудовых ресурсов, которая умирает при существующем порядке вымирания; W_i — численность занятых в сфере материального производства; x — возрастное распределение занятых; q — вероятность умереть в определенном возрастном интервале; m — мужское население; f — женское население; i — возраст.

Произведенный А. А. Ткаченко расчет вероятностного резерва национального дохода для 1973 г., носящий, правда, по его мнению, условный характер, составил 0,3% национального дохода. Автор считает, что вследствие увеличения национального дохода, приходящегося на одного работника, «ценность» каждого потерянного работника возрастает.

Интерес представляет возможность изучения совместного и раздельного воздействия различных факторов на величину вероятного резерва национального дохода. Расчеты А. А. Ткаченко с элиминированием влияния других факторов показывают, что старение населения и изменение возрастных вероятностей смерти приводят к различным потерям национального дохода.

Приведенные примеры далеко не исчерпывают всех возможностей, заложенных в анализе вероятностных показателей таблиц смертности. Такого рода анализ позволяет нам решать конкретные задачи как теоретического, так и практического характера, начиная от совершенствования самих показателей таблиц смертности и кончая выявлением сложнейших взаимосвязей демографических процессов и структур с экономикой.

4.7. Вероятная продолжительность жизни и порядок вымирания

Смерть каждого человека — явление биологическое, но уровень смертности всего населения определяется социально-экономическими условиями жизни общества.

В таблицах смертности в качестве инструмента углубленного и подробного изучения напряженности силы смертности используется совокупность показателей, связанных между собой определенными отношениями.

В обычном житейском представлении господствует правильное мнение о том, что возрастной состав оказывает влияние на уровень смертности. В стационарном, т. е. гипотетическом, населении наоборот: здесь смертность формирует возрастной состав населения.

Использование одной из основных теорем теории вероятностей — гипотезы Т. Бейеса — позволяет ввести в анализ такой показатель таблиц смертности, как вероятная продолжительность жизни[1]. После ряда преобразований формула, выражающая теорему Бейеса, принимает вид:

$$\Pi_E = \frac{\int_0^1 x^{n+1}\ (1-x)^m dx}{\int_0^1 x^n (1-x)^m dx},$$

где Π_E — вероятность сложного события E после наблюдения события E^1; E — сложное событие, состоящее в n-кратном осуществлении простого события A (вероятность которого x) и m-кратном осуществлении противоположного события $Б$ (вероятность которого $1-x$); E^1 — будущее событие, т. е. случай еще одного осуществления события A с вероятностью, равной x.

Применим эту формулу к таблицам смертности. Определим сначала вероятность прожить еще b лет человеку, достигшему возраста a лет, т. е. найдем $_bp_a$. Обозначим: число лиц, доживших до возраста a лет, — l_a; число лиц, доживших до возраста $a+b$ из лиц l_a, — l_{a+b}. Тогда в приведенной выше формуле $n=l_{a+b}$, $m=l_a-l_{a+b}$, при этом m — число умерших при переходе от возраста a к возрасту $a+b$. Получим:

$$_bp_a = \frac{\int_0^1 x^{l_{a+b}+1} (1-x)^{l_a-l_{a+b}} \cdot dx}{\int_0^1 x^{l_{a+b}} (1-x)^{l_a-l_{a+b}} \cdot dx} = \frac{l_{a+b}+1}{l_a+2}.$$

Если l_a и l_{a+b} достаточно больше числа, то

$$\lim {_bp_a} = \lim \frac{l_{a+b}+1}{l_a+2} \approx \frac{l_{a+b}}{l_a}.$$

Точная формула $_bp_a = \frac{l_{a+b}+1}{l_a+2}$ позволяет найти вероятную продолжительность предстоящей жизни, т. е. найти вероятное число лет b, которое проживает лицо, достигшее возраста a лет, при условии, что $_bp_a=0{,}5$ (это соответствует медианному возрасту). Получив по конкретной таблице смертности l_a, находим l_{a+b} из уравнения:

$$\frac{l_{a+b}+1}{l_a+2} = \frac{1}{2},$$

откуда $l_{a+b} = \frac{1}{2} l_a$. Определяя по этой же таблице половину l_a, находим тем самым l_{a+b} и, наконец, b.

Определим вероятную продолжительность жизни мужчин, проживающих в городе, достигших 50 лет, используя таблицу смертности СССР 1958—1959 гг. Имеем: $l_{a=50}=81\,056$, $l_{a+b} = \frac{1}{2} l_a = 40\,528$. Эта величина приближенно соответствует возрасту 73 года. Находим вероятную продолжительность жизни мужчин, достигших 50 лет: 73—50=23 года. Таким образом, вероятная продолжительность жизни представляет собой число лет, которое проживает после возраста a лет ровно половина достигших этого возраста, или, иными словами, число лет, через которое число доживающих уменьшится вдвое.

Более точная интерполяционная формула вероятной продолжительности жизни имеет вид:

$$V_a = K + \frac{l_{a+b} - \frac{1}{2} l_a}{l_{a+b} - l_{a+b+1}},$$

где l_{a+b} и l_{a+b+1} — соседние табличные числа доживающих, из которых первое несколько больше, а второе несколько меньше $\frac{1}{2} l_a$. Число K означает целую часть V_a или разность между возрастом, для которого определяют величину вероятной продолжительности жизни, и тем возрастом $a+b$, в котором остается в живых несколько больше половины лиц возраста a лет. Принимая a равным нулю, получаем вероятную продолжительность жизни для новорожденных, т. е. V_0. Например, из таблиц смертности 1958—1959 гг. для мужчин города и села имеем: $l_0 = 100\,000$, $K=70$, $\frac{l_0}{2} = 50\,000$, $l_{0+b} = 50\,920$, $l_{0+b+1} = 48\,677$.

Тогда

$$V_0 = 70 + \frac{50\,920 - 50\,000}{50\,920 - 48\,677} \approx 70{,}4 \text{ года}.$$

Вероятная продолжительность жизни подводит итог условиям смертности.

Полученная нами формула вероятной продолжительности жизни позволяет по числам доживающих в таблицах смертности, т. е. l_x, помимо медианного возраста определять и квартильные возраста. Так, приравнивая $\frac{l_{0+t}}{l_0}$ значению $^3/_4$ или $^1/_4$, найдем нижний и верхний квартили. Для нижнего квартиля

[1] В теории вероятностей принято называть события, имеющие вероятность 0,5, *вероятными событиями.*

$\frac{l_{0+c}}{l_0} = \frac{3}{4}$ и для верхнего квартиля $\frac{l_{0+d}}{l_0} = \frac{1}{4}$. Откуда $l_{0+c} = \frac{3}{4} l_0$ и $l_{0+d} = \frac{1}{4} l_0$.

Более точные интерполяционные формулы для вычислений квартилей $Q_1 = c$ и $Q_3 = d$ имеют вид:

$$Q_1 = z + \frac{l_{0+c} - \frac{3}{4} l_0}{l_{0+c} - l_{0+c+1}}; \quad Q_3 = t + \frac{l_{0+d} - \frac{1}{4} l_0}{l_{0+d} - l_{0+d+1}},$$

где все символы играют ту же роль, что и при нахождении вероятной продолжительности жизни по интерполяционной формуле.

Может возникнуть необходимость характеризовать длину возрастного интервала, в котором сосредоточена основная масса стационарного населения таблиц смертности. Тогда можно использовать квартильное расстояние возрастов, которое будет равно $Q_3 - Q_1$. Если же мы хотим характеризовать положение квартильных возрастов относительно медианного возраста, т. е. вероятной продолжительности жизни, и получить показатель, аналогичный коэффициенту вариации, то можно найти отношение половины квартильного расстояния возрастов к вероятной продолжительности жизни новорожденных и получить $r = \frac{Q_3 - Q_1}{2V_0}$. Этот показатель характеризует разброс квартильных возрастов относительно вероятной продолжительности жизни. Можно измерить величину разброса путем предварительного вычисления средней квадратической из отклонений квартильных возрастов от вероятной продолжительности жизни

$$\sqrt{\frac{(Q_3 - V_0)^2 + (Q_1 - V_0)^2}{2}}$$

и последующего деления этой величины на вероятную продолжительность жизни. Получаем коэффициент вариации $K_в$:

$$K_в = \frac{\sqrt{\frac{(Q_3 - V_0)^2 + (Q_1 - V_0)^2}{2}}}{V_0}.$$

Важным измерителем скорости вымирания населения помимо вероятной продолжительности жизни, на наш взгляд, является отношение квартильных возрастов, т. е. $R = \frac{Q_3}{Q_1}$. Дело в том, что если предположить почти идеальные условия вымирания населения, при которых в молодых возрастах нет никакой смертности, а вымирание происходит только в старческих возрастах, то вероятная продолжительность жизни будет близкой к предельному возрасту таблиц смертности, т. е. ω. В этом случае квартильные возраста примыкают с разных сторон к вероятной продолжи-

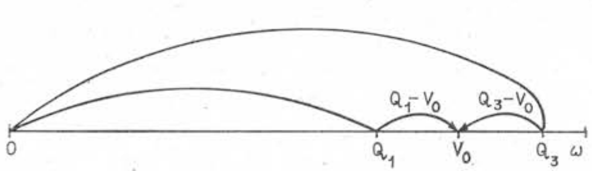

Рис. 4.10. Медианный и квартильные возраста при интенсивном вымирании в старческих возрастах

Рис. 4.11. Медианный и квартильные возраста при интенсивном вымирании в молодых возрастах

тельности жизни (графически это представлено на рис. 4.10). Тогда значение R будет довольно близким к единице. В случае наиболее интенсивного вымирания населения в молодых возрастах тоже может быть близким к единице (рис. 4.11), но это будет происходить на фоне значения V_0, более близкого к нулевому возрасту, чем к предельному.

Таким образом, для характеристики того, какие возраста — молодые или старческие — более интенсивно вымирают в период составления таблицы смертности, нужно привлечь два показателя — V_0 и R. Чем больше величина V_0, тем выше медианный возраст дожития населения, а чем больше величина R превышает единицу, тем интенсивнее вымирает население в молодые возраста. Следовательно, наилучшее положение с порядком вымирания имеем при

$$V_0 \to \omega \text{ и } R \to 1.$$

Рассмотренные нами ранее два показателя — r и $K_в$ — позволяют судить о плотности возрастов в таблицах смертности; третий показатель — R — дает в сочетании с V_0 характеристику порядка вымирания населения. Привлечение всех трех показателей — r, $K_в$ и R, — исчисленных по различным таблицам смертности, и их сопоставление по различным территориям и во времени позволит сделать определенные выводы о порядке вымирания в статике и в динамике, выявить соответствующие закономерности и тенденции.

Возьмем, например, данные таблиц смертности России и СССР 1896—1897 гг., 1926—1927 гг. и 1958—1959 гг. и сравним плотности возрастов, порядок вымирания у мужчин и женщин по каждой таблице, а также тенденции его изменения. Произведенные нами расчеты скомпонованы в табл. 4.2.

Из табл. 4.2 видно, что величины r и $K_в$ почти совпадают, поэтому для характеристики плотности возрастов в таблицах

Таблица 4.2. **Характер вымирания населения**

Год составления таблиц смертности	Пол	Нижний квартиль, Q_1	Верхний квартиль, Q_3	Вероятная продолжительность жизни, V_0	$r = \frac{Q_3-Q_1}{2V_0}$	$K_в$	$R = \frac{Q_3}{Q_1}$
1896—1897	Мужчины	0,8	63,3	17,9	1,7	1,8	79,1
	Женщины	0,9	65,0	26,8	1,2	1,2	72,3
	Оба пола	0,9	64,2	22,7	1,4	1,5	71,3
1926—1927	Мужчины	1,9	71,9	50,2	0,7	0,7	37,8
	Женщины	3,2	76,3	58,7	0,6	0,7	23,8
	Оба пола	2,4	74,0	54,1	0,7	0,7	30,8
1958—1959	Мужчины	55,8	80,7	70,4	0,2	0,2	1,4
	Женщины	66,4	85,7	77,9	0,1	0,1	1,3
	Оба пола	61,5	83,9	75,1	0,15	0,15	1,4

смертности следует привлечь тот из двух показателей, вычисление которого проще. Таким показателем является r. Сделаем некоторые выводы на основе данных табл. 4.2.

1. Во всех таблицах смертности у женщин по сравнению с мужчинами наблюдается: а) *более высокие значения вероятной продолжительности жизни, верхнего и нижнего квартилей;* при этом в более поздних таблицах смертности разности одноименных показателей Q_1 и Q_3 для женщин и мужчин возрастают, а разности V_0 уменьшаются; б) *меньшая плотность возрастов;* при этом значения показателя плотности (r) у женщин и мужчин в более поздних таблицах смертности приближаются: происходит нивелировка уровней плотности; в) *лучший порядок вымирания, т. е. менее интенсивное вымирание молодых возрастов;* при этом в более поздней таблице смертности разница в интенсивности этого вымирания становится у мужчин и женщин незначительной.

2. Динамические тенденции, вытекающие из сопоставления одноименных показателей более поздних таблиц смертности по сравнению с более ранними, могут быть сформулированы следующим образом: а) *возрастание итоговых (по двум полам) значений* Q_1, Q_3 и V_0. При этом Q_1 возрастает более быстро, чем Q_3, что свидетельствует о быстром приближении уровня Q_1 к V_0. Что касается V_0, то темп возрастания этого показателя замедляется (табл. 4.4);

Таблица 4.3. **Различия в показателях женщин и мужчин**

Год составления таблиц смертности	$Q_{1ж}-Q_{1м}$	$Q_{3ж}-Q_{3м}$	$V_ж-V_м$	$r_ж-r_м$	$R_ж-R_м$
1896—1897	0,1	1,7	8,9	—0,5	—6,8
1926—1927	1,3	4,4	8,5	—0,1	—14,0
1958—1959	10,6	5,0	7,5	—0,1	—0,1

б) *снижение показателя плотности возрастов до незначительной величины,* что свидетельствует о сужении межквартильной разницы возрастов, за

счет быстрого приближения Q_1 к V_0, а V_0 к ω; в) *улучшение порядка вымирания;* при этом в более поздних таблицах смертности вымирание молодых возрастов уступает место вымиранию в старческих возрастах, при высоком V_0 значение R достигло уже довольно низкого уровня (1.4).

Таблица 4.4. **Динамика показателей** Q_1, Q_3 **и** V_0

Сравниваемые таблицы смертности	Q_1	Q_3	V_0
1926—1927 к 1896—1897 гг.	2,7	1,1	2,4
1958—1959 к 1926—1927 гг.	25,6	1,1	1,4
1958—1959 к 1896—1897 гг.	68,3	1,3	3,3

Однако *возможности еще более значительного улучшения порядка вымирания налицо.* Несомненно, что дальнейшее снижение R и приближение его к единице будет происходить на фоне:

$$Q_1 \longrightarrow V_0 \longrightarrow Q_3 \longrightarrow \omega.$$

4.8. Точность вероятностных показателей таблиц смертности

Допустим, что мы хотим оценить достоверность исчисленных в таблице смертности вероятностей дожить или умереть, т. е. p_x или q_x. Из математической статистики известно, что средняя ошибка доли признака, т. е. относительной величины интенсивности, может быть вычислена по формуле

$$\mu_x = \sqrt{\frac{w(1-w)}{n}}$$

где μ_x — ошибка вероятности дожить или умереть в возрасте x лет; n — абсолютное число живущих в определенном возрасте; w и $1-w$ — соответственно p_x и q_x (т. е. вероятности). Предельная ошибка выборки равна утроенной средней ошибке. Используя соответствующие данные для городского населения, найдем, например, ошибки показателей p_0 и p_{80}. Численность населения в возрасте до 1 года примем в 1 млн. человек, а в возрасте 80 лет — 0,1 от численности населения до 1 года, т. е. 100 тыс. человек. Тогда

$$\mu_0 = \sqrt{\frac{0{,}95981 \cdot 0{,}04019}{1\,000\,000}} = 0{,}0002;$$

$$\Delta_0 = 3\mu_0 = 3 \cdot 0{,}0002 = 0{,}0006;$$

$$\mu_{80} = \sqrt{\frac{0{,}90983 \cdot 0{,}09017}{100\,000}} = 0{,}001;$$

$$\Delta_{80} = 3 \cdot 0{,}001 = 0{,}003.$$

Можно использовать теорию вероятностей при проверке точности таблиц смертности и другим способом. Известно, что интеграл вероятностей имеет вид:

$$P=\frac{2}{\sqrt{\pi}}\int\limits_0^v e^{-\frac{t^2}{2}}dt=\Phi(v).$$

Это соответствует вероятности того, что при n испытаниях число осуществлений события A будет находиться в пределах $np\pm a$. При этом

$$v=\frac{a}{\sqrt{2npq}}=\frac{a\sqrt{n}}{\sqrt{2m'n'}},$$

где m' и n' — наивероятнейшие исходы ($m'=np$; $n'=nq$). Примем отклонение вероятности $z=\frac{a}{n}$. Тогда

$$v=z\frac{\sqrt{n^3}}{\sqrt{2m'n'}}.$$

Подставим вместо n число доживающих до возраста x лет l_x, вместо m' — равное ему l_{x+1}, вместо n' — число умерших из числа l_x при переходе от возраста x к возрасту $x+1$, т. е. d_x. Получаем:

$$v=z\frac{\sqrt{l^3_x}}{\sqrt{2l_{x+1}\cdot d_x}}.$$

Зная наивероятнейший исход, определяем $p_v=\frac{1}{\sqrt{2\pi npq}}=\frac{1}{\sqrt{2\pi}}\times\times\frac{1}{\sqrt{npq}}=\frac{0,3989}{\sqrt{npq}}$ и, подставляя в формулу v, получаем:

$$p_v=0,3989\sqrt{\frac{l_{x+1}d_x}{l_x}}.$$

Допустим, что вероятность умереть мужчине в возрасте, например, 50 лет в течение следующего года жизни по таблицам смертности определена на основе 116 984 наблюдений, причем смертных случаев в течение года в этом возрасте было 998. Определим вероятность того, что в полученном значении q_{50}, равном $\frac{998}{116\,984}=0{,}00853$, верен: 1) второй знак после запятой; 2) третий знак после запятой; 3) четвертый знак после запятой.

Используем формулу

$$v=z\frac{\sqrt{l^3_x}}{\sqrt{2l_{x+1}d_x}}=z\frac{\sqrt{116\,984^3}}{\sqrt{2\cdot 115\,986\cdot 998}}=z\ 2629{,}703.$$

Придавая z значения $z_1=0{,}01$; $z_2=0{,}001$ и $z_3=0{,}0001$, получаем:

$$v_1=0{,}01\cdot 2629{,}703=26{,}297\approx 26;$$
$$v_2=0{,}001\cdot 2629{,}703=2{,}6297\approx 2{,}6;$$
$$v_3=0{,}0001\cdot 2629{,}703=0{,}26297\approx 0{,}26.$$

Используя таблицу интеграла вероятностей Маркова, находим значения вероятностей: $p_{v_1}\approx 1$; $p_{v_2}\approx 0{,}9998$; $p_{v_3}\approx 0{,}2900$. Полученные значения можно рассматривать как ответы на интересующий нас вопрос и трактовать их следующим образом: ошибки во втором и третьем знаках после запятой в вероятности умереть, т. е. q_x, мало вероятны. Что же касается ошибки в четвертом знаке (которая в конечном итоге может коснуться и третьего знака), то она очень вероятна, следовательно, на четвертый знак в вероятности умереть полагаться нельзя. Анализируя таблицу смертности, заключаем, что по мере увеличения возраста от 50 лет и старше необходимо располагать все большим и большим числом l_x, чтобы четвертый десятичный знак был точным.

Какова же минимальная численность возрастной группы, чтобы на основе теории вероятностей можно было быть уверенным в правильности четвертого или хотя бы третьего десятичного знака?

Возьмем, например, возраст $x=66$ годам. По таблице смертности 1958—1959 гг. для мужчин $q_{66}=0{,}03300$. Чтобы при вероятности 0,9998 был точен третий десятичный знак, нужно взять l_{66} больше 400 000 человек.

Все вышеизложенное позволяет сделать вывод о том, что вероятностные показатели в таблицах смертности верны только до третьего десятичного знака. Поэтому уже на четвертый десятичный знак следует смотреть как на величину, которая может отклоняться от действительной на несколько единиц.

В заключение разбора ошибок данного вида следует заметить, что как бы ни были велики эти ошибки, их можно рассчитать и учитывать в практической деятельности.

4.9. Вероятностные показатели таблиц брачности

Важность изучения брачной плодовитости и вероятности рождения детей у супругов диктует необходимость выявления влияния такого фактора, как уровень брачности, зависящий в свою очередь от возрастно-половой структуры бракоспособного контингента, сочетания возрастов вступающих в брак, длительности пребывания в браке, числа рожденных ранее детей, устойчивости браков и т. д.

По принципам построения таблицы брачности во многом сходны с таблицами смертности. Вступление в брак в них рассматривается как выбытие из совокупности несостоящих в браке, подобно тому как смерть есть выбытие из совокупности доживших до соответствующего возраста. Задача таблиц брачности состоит в выявлении повозрастного убывания исходной совокупности в результате действия либо одного фактора — вступления в брак, либо одновременно двух факторов — вступления в брак или смерти. При расчете вероятностей вступления в брак или смерти привлекаются данные о численности населения по состоянию на 1 января какого-то года (в качестве знаменателя дроби) и о числе браков и смертей за два года, примыкающие к этой дате (в качестве числителя).

Ю. А. Корчак-Чепурковский сделал попытку унифицировать и упорядочить систему обозначений таблиц брачности. Из основных показателей, предложенных им, укажем лишь те, которые строятся на основе вероятностных схем путем рассмотрения чисел холостых и девиц в возрасте от x до $x+1$ года и чисел доживающих и не вступивших в брак: 1) вероятность вступить в брак — δ_x; 2) вероятность умереть — q_x; 3) вероятность достигших возраста x вступить в брак когда-либо — δ'_∞; 4) вероятность навсегда остаться девицей — $1-\delta'_\infty$.

В обособленной таблице брачности Ю. А. Корчак-Чепурковский предложил вычислять «чистые» вероятности вступления в брак.

Вероятность вступления в брак в возрасте x лет (b_x) вычисляется по формуле

$$b_x = \frac{B_x}{S_x},$$

где B_x — число женщин, вступающих в брак в возрасте x лет; S_x — число незамужних женщин в возрасте x лет (общее число женщин, имеющих шанс вступить в брак).

Отсюда

$$b_x = \frac{S_x - S_{x+1}}{S_x} = 1 - \frac{S_{x+1}}{S_x}.$$

Принимая отношение $\frac{S_{x+1}}{S_x}$ за вероятность невступления в брак c_x, получаем: $b_x + c_x = 1$. S_{15} принимается равным 10 000.

Л. Е. Дарский при построении таблиц брачности находил также вероятность того, что женщина, вышедшая замуж в возрасте x лет и состоящая в этом браке y лет, останется в нем еще один год, а также вероятность сохранения брака на $y+1$ году его длительности с учетом вероятности дожития самой женщины[1]. Рассчитанные вероятности и соотношения между ними позволили Л. Е. Дарскому сформулировать некоторые выводы:

1) за начальный возраст женской таблицы брачности следует принять 15 лет, так как браки до этого возраста представляют исключение;

[1] См.: *Дарский Л. Е.* Формирование семьи. Демографо-статистическое исследование. М., 1972, с. 90.

2) не наблюдается существенной разницы в территориальных группах с низкой, средней и высокой рождаемостью между такими показателями, как «чистые» и комбинированные вероятности, возраста максимальной брачности, а также максимальные вероятности вступления в брак. Существенных различий нет также между кривыми распределения вероятностей вступления в брак;

3) наблюдаются существенные различия на территориях с разным уровнем рождаемости между вероятностями вступления в брак в раннем возрасте, а также средними возрастами вступления в брак;

4) следует считаться со стимулирующим влиянием добрачных зачатий на брачность и на возраст вступления в брак;

5) вероятность прекращения брака в плодовитом возрасте из-за смерти кого-либо из супругов даже при низком уровне смертности значительна;

6) из таблиц прочности браков следует, что ранние браки обладают сравнительно низкой прочностью; чем больше длительность брака, тем меньше вероятность его распада, наибольшей прочностью обладают браки, заключенные женщинами в возрасте 20—24 лет.

К изучению прочности браков можно привлечь показатель вероятности распада семьи на разных этапах ее существования. Обращая внимание на проблему прочности браков, некоторые демографы имеют в виду не только разводимость, но и смерть одного из супругов.

Для построения вероятности развода используются два коэффициента:

1) общий коэффициент разводимости строится по следующей схеме:

$$\frac{\text{число разводов за год}}{\text{численность всего населения}} \cdot 1000.$$

В этом показателе отсутствует соответствие между числителем и знаменателем (включающем в себя и детей, и тех взрослых, которые не состоят в браке);

2) специальный коэффициент разводимости получается по схеме:

$$\frac{\text{число разводов}}{\text{число браков}} \cdot 1000.$$

Этот показатель указывает на долю браков, кончающихся разводом. Недостатком этого второго коэффициента является несовпадение во времени фактов, входящих в состав числителя и знаменателя. Число разводов и браков берется за один год, но в число расторгнутых в данном году браков входят также браки, которые просуществовали свыше одного года.

А. Б. Синельников считает, что специальные коэффициенты разводимости не являются вероятностными показателями по ряду причин, в том числе и потому, что многие семьи перестают существовать не только из-за разводов, но и из-за смерти одного из супругов. Произведенные А. Б. Синельниковым расчеты для 1975 г. показали, что «...вероятность овдовения при любой длительности брака не столь мала, чтобы ею можно было пренебрегать»[1]. Чтобы получить общую вероятность прекращения брака, автор предлагает найти сумму вероятностей смерти для мужчин и женщин на первом пятилетии брака, на втором и т. д. и прибавить к ним долю разводов. Учитывая, что нас интересуют смертные случаи, которые ведут к распаду браков, надо исключить из расчетов умерших после развода или овдовения.

А. Б. Синельников пришел к выводу, что если существовавшие в 1975 г. уровни разводимости и смертности сохраняются в последующие 20 лет, то за это время из каждых 10 000 браков, заключенных в этом году, распадутся 4450 (2993 из-за развода и 1457 из-за овдовений).

Изучая рост разводимости, можно сделать вывод о значительном влиянии нестабильности браков на плодовитость женщин. В условиях половой диспропорции нестабильность браков, создавая неуверенность женщины в прочности брака, оказывает резко отрицательное влияние на рост рождаемости. Вот что пишет Л. Е. Дарский по этому поводу: «...повышенная вероятность распадения браков... способствует формированию более низкого уровня брачной плодовитости, так как женщина опасается остаться одна с детьми, и супруги не хотят «связывать» себя большим числом детей, учитывая потенциальную возможность развода»[2].

А. Г. Волков, учитывая наличие трех причин распада брака: смерть мужа, смерть жены и развода, — ввел в анализ независимые вероятности прекращения брака вследствие смерти мужа — q_y^m, смерти жены — q_y^f и развода — q_y^α. Тогда вероятность сохранения брака в течение года его существования при независимости указанных событий оказывается равной:

$$P_y = (1-\overset{m}{q_y})\cdot(1-\overset{f}{q_y})\cdot(1-\overset{\alpha}{q_y}),$$

а вероятность распада брака под влиянием указанных причин

$$Q_y = 1 - P_y \;.$$

Принимая численность браков, сохранившихся к началу очередного года брака, равной U_y, можно найти число распавшихся браков:

$$R_y = Q_y\, U_y \;.$$

Определяя числа браков, сохраняющихся к началу каждого года супружеской жизни,

$$U_{y+1} = U_y - R_y \;,$$

[1] *Синельников А.* Продолжительность существования современных браков.— В кн.: Возобновление поколений нашей страны (Народонаселение. Вып. 23). М., 1978, с. 110.

[2] *Дарский Л. Е.* Формирование семьи, с. 145.

числа браков, ежегодно распадающихся,

$$R_{y+1} = U_{y+1}\, Q_{y+1}$$

и используя указанные вероятности распада семьи и их доли в годичной вероятности прекращения брака:

$$\frac{q^f}{q^f+q^m+q^d}\,;\quad \frac{q^m}{q^f+q^m+q^d}\,;\quad \frac{q^d}{q^f+q^m+q^d}\,,$$

А. Г. Волков построил таблицу прекращения брака[1]. Ожидаемая продолжительность всех браков составила 32,3 года. В таблице, составленной А. Г. Волковым, предположение о том, что ведущей причиной распада браков являются разводы, подтвердилось не полностью. В конечном счете оказалось, что ведущей причиной распада семьи является смерть мужа. Что касается развода, то его роль велика на первых годах супружеской жизни.

В НИИ ЦСУ СССР использовали бюджетную сеть ЦСУ СССР для обследования рождаемости в семьях рабочих, служащих и колхозников[2]. Ответы 43 736 женщин, привлеченных к обследованию, на вопросы анкеты были использованы для построения таблиц брачности.

При построении «чистой» вероятности женщине в возрасте x лет вступить в брак, в которой влияние смертности было элиминировано, так как все обследованные женщины на момент обследования были живыми, использовались численности двух следующих совокупностей:

1) *знаменатель* (n) — число женщин, проживающих в периоде с 1/I 1949 г. по 1/I 1959 г. весь x-й год своей жизни, вступив в этот возраст никогда не состоявшими в браке;

2) *числитель* (m) — число женщин из совокупности n, вступивших в возрасте x лет в первый брак.

С точки зрения теории вероятностей в знаменателе искомой вероятности вступления в первый брак должно фигурировать общее число случаев, как благоприятствующих, так и не благоприятствующих вступлению в брак первый раз. Действительно, в знаменателе мы видим общее число женщин, имевших шанс вступить в первый брак в возрасте x лет в период наблюдения.

В числителе вероятности вступления в первый брак должно быть число случаев, благоприятствующих вступлению в первый брак в возрасте x лет. В нашем примере в числителе число женщин, действительно вступивших в первый брак в возрасте x лет.

На основе данных о вероятностях выжить в возрасте x лет из таблиц смертности женщин СССР за 1958—1959 гг. (p_x), а также полученных в результате соответствующих расчетов вероятности вступления в брак (b_x) была найдена комбинированная вероятность, например, выжить, не вступив в брак в интервале

[1] См.: *Волков А. Г.* Об ожидаемой продолжительности брака и ее демографических факторах. — В кн.: Демографическое развитие семьи/ Под ред. А. Г. Волкова. М., 1979, с. 73 и далее.

[2] В работе принимали участие А. Г. Волков, Л. Е. Дарский и др.

возраста от x до $x+1$. Полученные повозрастные вероятности вступления в первый брак, очищенные от влияния смертности, предварительно выравнивались, а далее служили основой построения чистых таблиц брачности раздельно для городского и сельского населения.

Л. В. Чуйко в отличие от рассматриваемого метода строит показатели вступления в брак без элиминирования влияния смертности. Основным ключевым признаком таблицы брачности является вероятность вступить в брак, вычисляемая по формулам, в которой в числителе — совокупность вступающих в брак и умерших (неженатыми, незамужними) первого рода, а в знаменателе — совокупность живущих первого рода (сверстников). В распоряжении статистики имеются совокупности вступающих в брак и умерших третьего рода и совокупности живущих второго рода (современники). Учитывая это, Л. В. Чуйко предлагает осуществлять переход от совокупностей вступающих в брак третьего рода к аналогичным совокупностям первого рода следующим путем:

«Если совокупность брачующихся третьего рода за два года, примыкающих к переписи, — $Б$, то совокупности брачующихся первого рода могут быть получены как $^1/_2 Б$ при условии равенства верхних и нижних элементарных совокупностей»[1].

С использованием вероятностей смерти из таблиц смертности 1958—1959 гг. были найдены абсолютные числа смертных случаев (совокупность умерших первого рода — M), половина которых учитывалась при исчислении «независимых» вероятностей вступления в брак, т. е. $^1/_2 M$ (при этом смертность не состоящих в браке приравнивается к смертности всего населения). Следовательно, числитель вероятности вступления в брак был равен $^1/_2 Б + ^1/_2 M$. Знаменатель представлял собой результат перехода от совокупности живущих, не состоящих в браке, второго рода к совокупности живущих, не состоящих в браке, первого рода.

При расчете частости (аналог вероятности) вступления в брак (δ) Л. В. Чуйко сочла необходимым устранить влияние смертности. Расчет независимой вероятности вступления в брак производится по формуле

$$b=\frac{\frac{1}{2}Б+\frac{1}{2}M}{L^2+\frac{1}{4}Б+\frac{1}{2}M};$$

где $Б$ — число первых и повторных браков за два года, примыкающие к переписи; M — число умерших, полученное на основании данных переписи населения и вероятностей смерти из таблиц смертности; L^2 — число холостяков или девушек, вдовых и разведенных по переписи населения.

[1] *Чуйко Л. В.* О вычислении суммарных таблиц брачности на основе современной статистической информации. — В кн.: Материалы Всесоюзной научной конференции по проблемам народонаселения Закавказья. Ереван, 1968, с. 58.

В ряде случаев данные о возрастном распределении вступающих в брак лиц различных брачных состояний отсутствуют. Вместе с тем возникает необходимость оценить возрастные распределения первых браков и найти число повторных браков. Вариант такого рода расчетов предложил М. С. Тольц[1]. Он использовал следующую систему обозначений: B^1_x — число вступивших в первый брак по «чистой» таблице брачности; β_x — коэффициенты брачности девиц; ω_x — числа девиц в возрасте x на момент переписи 1897 г.; W_x — общее число женщин в возрасте x лет по переписи; b^1_x — «чистая» вероятность вступления в первый брак (из модели Э. Коула); L_x — средние числа девиц в возрастном интервале $x/x+1$, полученные в таблицах брачности.

Определение числа первых браков М. С. Тольц производит путем использования «чистых» вероятностей вступления в первый брак по следующей формуле:

$$B^1_x=\beta_x\cdot\omega_x ,$$

где

$$\beta_x=\frac{2b^1_x}{2-b^1_x};\quad \omega_x=0{,}0001\cdot L_x\cdot W_x;$$

$$L_x=\frac{1}{2}(S^1_x+S^1_{x+1})+\frac{1}{24}(B^1_{x+1}-B^1_{x-1}).$$

Учитывая, что прекращение брака происходит помимо разводов еще из-за смерти одного из супругов, рассчитывается вероятность сохранения брака до конца плодовитого периода. Если женщина вступила в брак в возрасте x лет, то вероятность сохранения брака до конца плодовитого возраста (50 лет) равна:

$$p_x=\frac{\overset{ж}{l}_{50}\,\overset{м}{l}_{50+\tau}}{\overset{ж}{l}_x\,\overset{м}{l}_{x+\tau}},$$

где τ — средняя разница в возрасте мужа и жены, принимаемая равной 5 годам.

4.10. Вероятности увеличения семьи

При совершенствовании системы показателей плодовитости было обращено внимание на важность и необходимость изучения порядка рождений. Сначала для изучения порядка рождений реальных когорт были привлечены доли женщин, имевших к концу плодовитого возраста n рождений, а также вероятности

[1] См.: *Тольц М. С.* Брачность населения России в конце XIX — начале XX в. — В кн.: Брачность, рождаемость и смертность в России и в СССР/Под ред. А. Г. Вишневского. М., 1977, с. 147.

увеличения семьи. Затем вероятностный метод изучения семьи был распространен и на гипотетические поколения.

Вывод, к которому пришли исследователи, состоит в том, что «...для правильного применения метода гипотетического поколения к измерению плодовитости основным исходным показателем должна быть вероятность для женщин данного возраста родить следующего ребенка»[1].

Для одного календарного года табличный коэффициент плодовитости φ_x^n можно вычислить как отношение численности женщин в возрасте x лет, рожавших n раз (A_x^n), к численности женщин в этом возрасте, рожавших $n-1$ раз (A_x^{n-1}):

$$\varphi_x^n = \frac{A_x^n}{A_x^{n-1}}.$$

Учитывая формулу, связывающую коэффициенты с вероятностями, получаем вероятность родить n-й раз в возрасте x лет:

$$p_x^n = 1 - e^{-\varphi_x^n} = \frac{2\varphi_x^n}{2+\varphi_x^n}.$$

Используя указанные повозрастные вероятности, можно показать постепенный переход женщин от нерожавших к рожавшим один раз, а из этой совокупности — к рожавшим два раза и т. д. На основе этих данных получаются такие статистические характеристики, как вероятность увеличения семьи и средний возраст матерей при родах каждого порядка.

Аналогичные рассуждения позволили Л. Е. Дарскому построить таблицу плодовитости, уточняющую существующие методы, соединив аналитические возможности модели стабильного населения с вероятностями увеличения семьи.

Когортный метод анализа предполагает, что все исследованные рождения относятся к одной и той же исходной совокупности женщин. При обследованиях, касающихся, например, планируемого женщинами числа детей в брачных когортах, вероятность увеличения семьи или вероятность рождения следующего ребенка (a_n) вычисляется по формуле

$$a_n = \frac{\sum\limits_{n+1}^{\infty} w_n}{\sum\limits_{n}^{\infty} w_n},$$

где w_n — частость женщин, планирующих рождения n детей, в исходной совокупности женщин ($n=0, 1, 2, \ldots$).

В 1972 г. отдел демографии НИИ ЦСУ СССР провел выборочное изучение рождаемости путем опроса замужних женщин об ожидаемом числе детей[1].

[1] *Дарский Л. Е.* Формирование семьи, с. 90.

Покажем примерную схему, позволяющую осуществить расчет вероятности увеличения семьи по данным о планируемом числе рождений и частости женщин, планирующих то или иное число рождений.

Пусть в результате статистического наблюдения получено распределение женщин по планируемому числу детей. Расчет вероятностей увеличения семьи представлен в табл. 4.5.

Таблица 4.5. **Планируемое число рождений и соответствующие расчеты**

Планируемое число рождений, n	Частость женщин, w_n	Накопление частости с конца значений w_n, Σw_n	Вероятности увеличения семьи, a_n
0	0,12	1,00	0,88
1	0,18	0,88	0,80
2	0,23	0,70	0,67
3	0,14	0,47	0,70
4	0,13	0,33	0,61
5	0,10	0,20	0,50
6	0,06	0,10	0,40
7	0,04	0,04	—
Итого	1,00	—	—

Рассчитав накопленные частости снизу вверх, находим вероятности увеличения семьи путем деления последующих накопленных частостей на предыдущие. Так, вероятность рождения первого ребенка равна:

$$a_1 = \frac{0{,}88}{1{,}00} = 0{,}88;$$

вероятность рождения второго ребенка

$$a_2 = \frac{0{,}70}{0{,}88} = 0{,}80 \text{ и т. д.}$$

Для выявления той части ожидаемой рождаемости, которая уже реализована, были найдены доли реализованных рождений каждого порядка:

$$r_n = \frac{\sum\limits_{k}^{\infty} w_k}{\sum\limits_{n}^{\infty} w_n},$$

где w_k — частость женщин, родивших k детей; w_n — частость женщин, ожидающих n детей.

Величина r_2, равная, например, 0,5, означает, что из всех женщин данной совокупности, ожидающих двоих и более детей, половина уже родила второго ребенка.

Интересно, что различий между этнодемографическими группами, полученными в результате группировки женщин по их репродуктивным намерениям и по вероятности рождения первого

[1] См.: Сколько детей будет в советской семье (результат обследования). М., 1977, с. 9.

ребенка, почти не существует, но, начиная с вероятности рождения второго ребенка, появляются существенные различия, углубляющиеся с переходом к вероятностям рождения третьего и четвертого ребенка.

Чтобы изучить продуктивность браков, Л. Е. Дарский сопоставил вероятности увеличения семьи для браков, заключенных женщинами в 20—24 года, в городе и сельской местности с разными уровнями рождаемости в них. Полученные вероятностные показатели позволили ему сделать интересные заключения. Приведем некоторые из них[1]:

1. Существенные различия проявляются в вероятностях рождения третьего и четвертого ребенка.
2. Женщины, родившие троих детей, чаще рожают четвертого ребенка, чем родившие двоих детей рожают третьего.
3. Сельские жители территорий с высокой плодовитостью вообще не склонны ограничивать число детей в семье.
4. Предпочтительная величина семьи особенно ярко проявляется в показателях для городских жителей территорий с низкой рождаемостью, где свыше 50% женщин ограничиваются двумя детьми.

Изучение плодовитости женщин на основе выборочных обследований и вероятностных показателей производится и в других странах. Так, А. Никулин приводит ретроспективные данные о вероятностях очередного рождения у марокканских городских женщин в возрасте 40—49 лет[2]. Вероятность для бездетной женщины родить первенца $a_0=0{,}94$; вероятность родить двух детей $a_0 \cdot a_1=0{,}94 \cdot 0{,}97$ и т. д. Число рождений определяется суммой $a_0+a_0a_1+a_0a_1a_2+\ldots+a_0a_1\ldots a_{n-1}$.

В продольном разрезе вероятности увеличения семьи по Л. Анри (от брака до первого рождения, от первого рождения до второго и т. д.) вычисляются путем использования данных о рождениях по порядку рождения и возрасту матери, а также данных переписи о возрастном составе женщин, имеющих уже k детей. Формула вероятности увеличения семьи имеет вид:

$$a_k=\frac{\overline{N_{k+1}}}{\sum\limits_n a_{k,\,n} \cdot N_{k,\,n}},$$

где $\overline{N_{k+1}}$ — число рождений порядка $k+1$, происшедших в течение года; $N_{k,n}$ — число женщин, родивших k-го ребенка n лет назад; $a_{k,n}$ — доля рождений порядка $k+1$, происшедших через n лет (после k-го рождения).

---

[1] См.: *Дарский Л. Е.* Плодовитость браков в СССР, 1949—1959 гг., по выборочным данным. — В кн.: Факторы рождаемости/Под ред. А. Г. Волкова. М., 1971, с. 11.

[2] См.: *Никулин А.* Рождаемость в странах Магриба (Марокко, Алжир, Тунис). — В кн.: Демографический анализ рождаемости. М., 1974, с. 89.

При создании модели плодовитости супружеской пары Л. Анри использовал следующие вероятностные показатели: вероятность зачатия в интервале возраста от x до $x+dx$; вероятность для женщины, зачавшей в возрасте x лет, оставаться стерильной в возрасте $x+q$; вероятность того, что зачатие, происшедшее в возрасте x лет, закончится живорождением. Следует добавить, что модель Л. Анри предполагает отсутствие контрацепции. Этому условию не отвечает рождаемость среди большинства населения развитых в экономическом отношении стран.

Естественную рождаемость населения В. А. Борисов оценивает с помощью стандарта, разработанного на основе математической модели репродуктивного процесса и привлечения массовых фактических данных об уровнях рождаемости. Для этой цели автор вводит систему обозначений, рассматривая весь процесс поэтапно. Беременность женщины может иметь два исхода (живорождения и мертворождения), каждый из которых обладает определенной вероятностью. Затем идет период временной стерильности, после которого процесс может повториться. Таким образом, «...женщина в любой момент времени находится в одном из следующих состояний: S_0 — овуляторная экспозиция (время отсрочки зачатия); S_1 — беременность; S_2 — стерильность после живорождения; S_3 — стерильность после мертворождения или спонтанного аборта»[1]. Когда женщина становится беременной, она переходит из состояния S_0 в состояние S_1, а вероятность этого перехода, или коэффициент оплодотворяемости, — ρ. При живорождении женщина переходит в состояние S_2, а при мертворождении — в состояние S_3. Далее происходит возврат в состояние S_0. Переход из одного состояния в другое обладает определенной вероятностью — V_i; средняя продолжительность периода стерильности женщин, связанная с этим состоянием, — t_i. Обозначим вероятность рождения живого ребенка — V_2 и среднюю продолжительность беременности плюс среднюю продолжительность периода стерильности после живорождения — t_2, а V_3 и t_3 — соответствующие показатели, связанные с внутриутробной смертностью. Среднее ожидаемое время до зачатия определяется отношением $(1-\rho):\rho$. Таким образом, средний интервал между зачатиями (T_1) представляет собой сумму трех периодов.

$$T_1=\frac{1-\rho}{\rho}+V_2t_2+V_3t_3\ ,\ \text{где}\ V_2+V_3=1.$$

Средний интервал между живорождениями составляет:

$$T_2=\frac{T_1}{V_2}\ .$$

---

[1] *Борисов В. А.* Перспективы рождаемости. М., 1976, с. 51.

Средний коэффициент живорождаемости обратно пропорционален интервалу между рождениями, т. е.

$$F=\frac{V_2}{\frac{1-\rho}{\rho}+V_2t_2+V_3t_3}=\frac{\rho V_2}{1+\rho(V_2t_2+V_3t_3-1)}.$$

Учитывая, что некоторая часть замужних женщин бесплодна (α), а также и то, что часть женщин живет в длительной разлуке с мужьями (β), В. А. Борисов брачный годовой коэффициент рождаемости (на 1000 замужних женщин) окончательно представляет в следующем виде:

$$F=\frac{12\cdot 1\,000\rho V_2(1-\alpha-\beta)}{1+\rho(V_2t_2+V_3t_3-1)}.$$

На основе статистических данных и некоторых предположений относительно величины показателей, входящих в состав полученной формулы, В. А. Борисов получил для женщин в возрасте 20—24 года следующий брачный годовой коэффициент рождаемости:

$$F_{20-24}=\frac{12\cdot 1\,000\cdot 0{,}15\cdot 0{,}80(1-0{,}06-0{,}05)}{1+0{,}15\cdot(0{,}80\cdot 18+0{,}20\cdot 4{,}6-1)}=407{,}6‰.$$

Разумеется, что для женщин других возрастов, а также при иных предположениях о величине исходных показателей результаты получатся другие.

Вероятности рождения n-го ребенка могут рассчитываться не только в таблицах плодовитости, но и в таблицах продуктивности брака. В этом случае при их исчислении вместо возраста женщины используется длительность брака. Тогда получается вероятность родить n-го ребенка на $y+1$ году брака для женщины, вышедшей замуж в возрасте x лет и к y-му году брака родившей $n-1$ детей.

И. А. Герасимова, рассматривая распределение семей по числу рождений при любой фиксированной длительности брака, вывела систему формул, отвечающих на вопросы о формировании числа членов семьи и возрастной структуры поколения детей[1]. Эти формулы позволяют осуществлять переход от данных о рождаемости к составу семей, а также прогнозировать изменение режима брачной плодовитости. При ответе на вопрос о тенденции формирования семьи И. А. Герасимова рассматривает пять групп семей (с различными подразделениями, составляющими всего 15 типов) и оценивает вероятности перехода семьи из одного состояния в другое.

[1] См.: *Герасимова И. А.* Структура семьи. М., 1976, с. 131.

4.11. Осреднение демографических признаков, основанное на использовании принципов построения таблиц смертности

При изучении различных демографических событий большое значение имеет определение средних возрастов наступления этих событий. Специфика отдельных событий диктует необходимость особого подхода к этой проблеме. В частности, исследованием средних возрастов занимались В. Вейнберг при составлении таблиц болезненности населения, Г. А. Баткис при изучении показателей медицинской статистики и др. Не удовлетворившись использованием для этой цели обычной средней арифметической, исследователи разработали особую вероятностную методику расчета, опирающуюся на привлечение принципов построения таблиц смертности.

Рассмотрим в качестве примера осреднение признака *возраст вступления в брак* и определение *среднего возраста вступления в брак*.

При наличии данных о возрасте мужчин, вступающих в брак, и их числе статистика для определения среднего возраста рекомендует использовать среднюю арифметическую взвешенную ($\bar{x}$). В качестве варьирующего признака фигурирует возраст мужчин, вступающих в брак (x), а в качестве статистического веса — их численность в каждом возрасте (m). Формула средней арифметической взвешенной имеет вид:

$$\bar{x}=\frac{\Sigma xm}{\Sigma m}.$$

Рассмотрим конкретный пример. Допустим, что имеются две равные по общей численности совокупности мужчин (по 1305 человек). Часть из них вступили в брак. Численность и распределение по возрасту вступивших в брак в обеих совокупностях одинаково, а распределение по возрасту невступивших в брак (потенциальных женихов) различно (табл. 4.6).

Таблица 4.6. **Возраст вступления и невступления в брак мужчин**

Возраст, x	Совокупность I			Совокупность II		
	Вступившие в брак в данном возрасте, m	Невступившие в брак к данному возрасту, b	Всего мужчин, $m+b$	Вступившие в брак в данном возрасте, m	Невступившие в брак к данному возрасту, b^1	Всего мужчин, $m+b^1$
18	10	250	260	10	—	10
19	125	200	325	125	—	125
20	150	150	300	150	50	200
21	180	100	280	180	100	280
22	40	50	90	40	150	190
23	35	—	35	35	200	235
24	15	—	15	15	250	265
Всего	555	750	1 305	555	750	1 305

Рассчитаем средний возраст мужчин, вступивших в брак, по формуле средней арифметической взвешенной (в нашем случае он будет одинаков для обеих совокупностей):

$$\bar{x}=\frac{18\cdot10+19\cdot129+\dots+24\cdot15}{10+125+\dots+15}=\frac{11\,380}{555}\approx20{,}5 \text{ года.}$$

Однако такая средняя арифметическая с точки зрения демографии является ненадежной, потому что в ее расчете участвуют только те лица, которые в каждом возрасте обладают изучаемым признаком (в нашем примере женатые) и остаются вне наблюдения лица, не обладающие изучаемым признаком (неженатые). Более точно охарактеризовать брачность в каждой из этих совокупностей можно, используя принцип построения таблиц смертности, когда в расчет включаются все лица: как обладающие, так и не обладающие изучаемым признаком (в таблицах смертности — это умершие и живущие; в нашем случае — это вступившие и невступившие в брак). Мы вправе ожидать, что во второй совокупности, где в старших возрастах резерв потенциальных женихов значительно больше, чем в первой, браки должны заключаться позднее, а поэтому средний возраст мужчин, вступающих в брак, должен оказаться выше.

Такое рассмотрение привлекаемых к осреднению совокупностей мужчин видоизменяет наше представление о среднем возрасте мужчин, вступивших в брак, и приводит к представлению об этом среднем возрасте вступления в брак как о средней продолжительности жизни мужчин от возраста, принятого за начальный (в нашем примере 18 лет), до вступления в брак.

Методика, основанная на привлечении вероятностей дожития, состоит в исчислении средней продолжительности жизни людей, достигших определенного возраста, после этого возраста. Эта средняя представляет частное от деления числа лет, прожи-

Таблица 4.7. **Исчисление средней продолжительности жизни первой совокупности мужчин до вступления в брак**

Возраст, x	Вступившие в брак в данном возрасте, m	Невступившие в брак к данному возрасту, b	Вступившие в данный возраст неженатыми, $\Sigma(m+b)$	Дополнение к единице вероятности вступления в брак в предшествующем возрасте, q_x	Вероятность дожития до данного возраста без вступления в брак, v_x	Общая сумма лет, прожитых каждым членом возрастной группы, Σv_x
18	10	250	1 305	1,00	1,00	4,29
19	125	200	1 045	0,99	0,99	3,29
20	150	150	720	0,88	0,87	2,30
21	180	100	420	0,79	0,69	1,43
22	40	50	140	0,57	0,39	0,74
23	35	—	50	0,71	0,27	0,35
24	15	—	15	0,30	0,08	0,08
Всего	555	750	—	—	—	—

тых этими людьми, на их начальную численность. Покажем примерную схему расчета средней продолжительности жизни первой совокупности мужчин до вступления в брак по этой методике.

Расчет $\Sigma(m+b)$ производится путем последовательного суммирования снизу вверх данных двух граф — m и b — начиная с возраста 24 года (последняя строка). Получаем: $\Sigma(m_{24}+b_{24})=15$ (для предпоследней строки); $\Sigma(m_{23}+b_{23})=15+35=50$; $\Sigma(m_{22}+b_{22})=15+35+40+50=140$ и т. д.

Расчет q_x, т. е. вероятности невступления в брак в возрасте $x-1$, осуществляется как дополнение до единицы вероятности вступления в брак в предшествующем возрасте:

$$q_{18}=1;\ q_{19}=1-\frac{10}{1\,305}=0{,}99;\ q_{20}=1-\frac{125}{1\,045}=0{,}88 \text{ и т. д.;}$$

$$q_{24}=1-\frac{35}{50}=0{,}30.$$

Показатели v_x, представляющие собой вероятность дожития до данного возраста без вступления в брак, получаются последовательным умножением данных предыдущей графы сверху вниз:

$$v_{18}=1{,}00;\ v_{19}=1{,}00\cdot0{,}99=0{,}99;\ v_{20}=0{,}99\cdot0{,}88=0{,}87 \text{ и т. д.;}$$
$$v_{24}=0{,}08.$$

Значения Σv_x представляют собой общую сумму лет, прожитых каждым членом возрастной группы, вступившим под наблюдение до момента вступления в брак, и получаются последовательным суммированием показателей предыдущей графы снизу вверх. Получаем:

$$\sum v_{24}=0{,}08;\ \sum v_{23}=0{,}08+0{,}27=0{,}35;$$
$$\sum v_{22}=0{,}35+0{,}39=0{,}74 \text{ и т. д.; } \sum v_{18}=4{,}29.$$

Таким образом, конечная сумма указывает среднюю продолжительность жизни от начального возраста до момента вступления в брак.

Таблица 4.8. **Пересчет средней продолжительности жизни мужчин до вступления в брак**

Возраст, x	Число мужчин, вступивших в брак в данном возрасте, m	Число мужчин, вступивших в данный возраст неженатыми, Σm	q_x	v_x	Σv_x
18	10	555	0,99	1,00	3,50
19	125	545	0,77	0,99	2,50
20	150	420	0,64	0,76	1,53
21	180	270	0,33	0,49	0,77
22	40	90	0,55	0,16	0,28
23	35	50	0,30	0,09	0,12
24	15	15	0,03	0,03	0,03
Итого	555	—	—	—	—

Получаем: средний возраст вступления в брак равен: $18+3{,}29=21{,}29$ года. Аналогичный расчет для второй совокупности мужчин привел к результату: $18+4{,}76=22{,}76$ года.

Используя данные табл. 4.6, производим пересчет (см. табл. 4.8).

Следовательно, средняя продолжительность жизни мужчин, находящихся в возрасте 18 лет, до вступления в брак равна 2,5 года и составляет в итоге: $18+2{,}5=20{,}5$ года.

Таким образом, видим полное совпадение двух характеристик: среднего возраста мужчин, вступивших в брак, полученного нами по данным табл. 4.6, и общей средней продолжительности жизни мужчин до вступления в брак (по табл. 4.8).

Более совершенный подход к среднему возрасту вступления мужчин в брак как к средней продолжительности жизни мужчин от начального возраста вступления в брак (например, 18 лет) до фактического заключения брака приводит исследователя к демографическому пониманию этого показателя и требует привлечения для расчетов как численностей вступающих в брак, так и численностей невступающих в брак.

Глава 5

Вероятностное прогнозирование и моделирование демографических явлений

5.1. Прогнозирование и его задачи

Прогнозирование можно рассматривать как начальную стадию планирования, дающую необходимые материалы для выработки научно обоснованной политики. «Научным в современных условиях может быть лишь такое планирование, которое базируется на тщательно разработанных долгосрочных прогнозах — технических, экономических, демографических и других...»[1].

Следует отметить все возрастающую роль демографических факторов при определении перспектив экономического развития страны. В связи с этим усиливается значение демографических прогнозов, которые становятся одной из главных предпосылок проектов развития народного хозяйства.

Надо обязательно учитывать вероятностный характер демографических прогнозов, представляющих собой оценку возможного развития демографических процессов и требующих соответствующих интервальных показателей в виде областей допустимых значений.

Правильность демографического прогноза определяется соответствием принятой за основу прогнозирования математической модели характеру самого демографического процесса, закону его развития. Главная мысль, заложенная в прогнозировании, состоит в том, что современным состоянием явления в значительной степени определяется его будущее. При этом будущее не простой слепок с прошлого и настоящего, а такое проявление диалектической тенденции развития, которое имеет характер научного упреждения, носящего вероятностный характер. О тесной взаимосвязи экономических, социальных и демографических явлений говорил в своем выступлении на XXIV съезде КПСС Л. И. Брежнев: «...осуществление важнейших экономических и социально-политических задач требует не пяти лет, а гораздо большего сро-

[1] *Косыгин А. Н.* Социально-экономическое развитие Советского государства. — Коммунист, 1972, № 17, с. 39.

11 Заказ 2715

ка. В этой связи встает вопрос о перспективном долгосрочном планировании развития народного хозяйства, опирающемся на прогнозы роста населения страны, потребностей народного хозяйства, научно-технического прогресса»[1]. Большое значение при определении перспективных изменений демографических фактов имеет изучение влияния неопределенностей, сопровождающих такие изменения, т. е. непредсказуемостей каждого отдельного явления. В этих случаях наиболее важным методом определения наилучших оценок изучаемых фактов и процессов является вероятностный критерий наибольшего правдоподобия.

Если имеется динамический ряд, характеризующий изменения, например, численности населения $S_1, S_2, \ldots, S_m$, и требуется найти наилучшую оценку параметров модели, которую мы примем для описания динамики процесса, то, предполагая, что изменение этой численности населения происходит, например, по прямой линии $\overline{S_t}=a_0+a_1t$, по параболе второго порядка $\overline{S_t}=a_0+a_1t+a_2t^2$ и т. д., определяем такие значения параметров a_i, при которых значение наблюдаемых численностей населения было бы наиболее вероятным. Можно показать, что задача сводится к минимизации суммы вида:

$$\sum(S_t-\overline{S}_t)^2=\min,$$

т. е. к методу наименьших квадратов, являющемуся частным случаем метода наибольшего правдоподобия.

Большинство методов прогнозирования основано на предположении, что течение демографических процессов в будущем будет происходить аналогично тому, как оно происходило в прошлом. Вот почему в перспективных расчетах, связанных с определением будущей численности населения и его возрастно-половой структуры, все предположения носят вероятностный характер и прогноз дается, как правило, в нескольких вариантах.

Прогноз — некоторая гипотеза, вероятностная оценка протекания изучаемого процесса в будущем — может быть дан в виде одного числа (точечный) или в виде интервала, в котором находится прогнозируемый показатель (интервальный).

Прогнозирование демографических явлений (даже краткосрочное, а тем более долгосрочное) на основе использования статистических методов экстраполяции, выявленной тенденции — тренда (а в демографии во многих случаях до сих пор именно так и делается) — может быть успешным лишь в том случае, если оно учитывает вероятностный характер возможного изменения набора влияющих факторов, их направленности и силы, изменения структурных соотношений в составе населения и т. д.

Можно считать доказанным, что для прогнозов в границах 15—20 лет использование тончайших методов математики не только не противопоказано, а даже желательно, а вот для отдаленных прогнозов тонкие математические приемы и методы исчисления могут скорее затушевать неясность исходных данных и сомнительность гипотез, лежащих в их основе. Иногда предсказать возможность какого-то скачка в динамике явления помогает логическое осмысливание процесса. Так, рост научных кадров в США за период с 1940 по 1950 г. происходил по экспоненте со скоростью, превышающей рост всего населения страны. Это с большой вероятностью позволило предсказать будущее скачкообразное уменьшение коэффициента роста экспоненты, которое произошло в начале 70-х годов.

[1] Материалы XXIV съезда КПСС. М., 1973, с. 67.

5.2. Марковские процессы

Привлечение теоретико-вероятностных методов в демографии заслуживает большого внимания также в связи с моделированием демографических процессов, с широким применением регрессионных и корреляционных методов.

Наиболее затруднительным при построении вероятностных моделей в демографии является не создание математико-вероятностного аппарата, адекватного изучаемым процессам, хотя эта задача сама по себе ~~достаточно~~ трудна, а недостаточная изученность качественной природы демографических явлений и взаимосвязи факторов, формирующих их интенсивность.

Несомненно, что математико-статистические вероятностные модели не могут отразить всех аспектов каждого изучаемого демографического явления и математико-статистическое понимание явления не во всех деталях полностью совпадает с его объективной сущностью, но всегда надо стремиться к тому, чтобы такое моделирование было как можно теснее увязано с материальной природой явлений. Отличие вероятностной модели от детерминистской состоит в следующем: используя детерминистскую модель, можно по заданному на момент t, например, распределению населения по полу и возрасту рассчитать такую же структуру для каждого будущего момента времени. Но если учитывать влияние случая, т. е. если предположить, что процесс повторяется неоднократно, то от одного состояния (распределения населения) можно прийти к различным реализациям процесса. При этом каждая реализация обладает определенной вероятностью. Разумеется, что среди этих реализаций имеются наиболее вероятные — наивероятнейшие — и менее вероятные. Получаем вероятностную модель, достоинство которой заключается в том, что она дает возможность определить границы изменения параметров модели путем расчета их вероятных ошибок, а также сопоставления расчетных данных с фактическими.

Задача вероятностных построений в демографии состоит, в частности, в том, чтобы на основе, например, зафиксированной на некоторый исходный момент времени структуры населения, а также выявленной закономерности динамических изменений таких характеристик, как рождаемость и смертность, указать вероятности тех или иных состояний будущей структуры и уровня демографических характеристик.

Занимаясь практическими задачами страхового дела, русский математик А. А. Марков разработал методику особых моделей, названных впоследствии цепями (или процессами) Маркова. Вначале цепи Маркова использовались главным образом при решении некоторых статистических задач биологии. Сравнительно недавно выяснилась возможность изучения динамики численности населения, плодовитости, миграционных потоков и других демографических явлений путем привлечения коэффициентов рождаемости, смертности и т. д., на основе использования цепей Маркова. Идея при этом состоит не в выявлении величины значений тех или иных демографических показателей в момент t, а в нахождении их вероятностей. Тогда вместо определения состояния, в котором будет находиться население, можно, зная закономерности развития демографических процессов, происходящих в населении, определять вероятности различных переходов из одного состояния в другое.

При изучении многообразных признаков такой статистической совокупности, как население (пол, возраст, семейное положение, профессия, место жительства и т. д.), можно, при некоторых весьма общих допущениях, получить структуру этой совокупности на определенный момент как по различному набору признаков, так и по группировке по каждому признаку.

Статистическое наблюдение за данной совокупностью в течение какого-то сравнительно небольшого отрезка времени позволяет определить вероятности перехода единиц совокупности отдельных лиц из данной группы в другую, а также выход этих единиц из-под наблюдения (смерть, переход с одного места работы на другое, исключение из учебного заведения и т. д.).

Предполагая, что найденные вероятности перехода остаются неизменными в течение какого-то интервала времени, получаем марковский процесс. При таком переходе из одного состояния в другое через некоторое время в совокупности установится определенная структура, зависящая только от переходной матрицы, но не зависящая от начальной структуры. В этом и состоит принцип Маркова, иногда называемый законом или принципом эргодичности. Он заключается в том, что процесс перехода из одного состояния в другое за большой промежуток времени как бы отгораживается от далекого начального состояния. Это обстоятельство является выражением житейского наблюдения, состоящего в том, что хотя новое состояние явления непосредственно связано с предшествующим, но оно может оказаться и независимым от далекого начального состояния.

Что же касается вероятностей тех или иных значений в последующие моменты, то они не зависят от того, какие значения они принимают в предшествующие моменты. В демографии путем фиксации режима главных факторов, определяющих современную структуру населения, т. е. рождаемости и смертности, различают

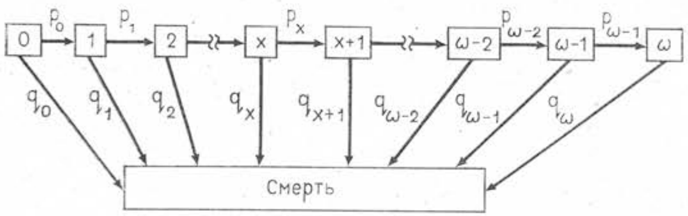

Рис. 5.1. Вероятности перехода человека из одной возрастной группы в другую $p(x)$ и вероятности выбытия (смерти) $q(x)$

слабую и сильную эргодичность. При изменяющемся режиме главных факторов свойство «забывать» прошлую структуру называют слабой эргодичностью.

Практические исследования по демографии показали, что влияние прошлой структуры населения по определенным признакам на современную структуру населения по этим же признакам по мере отдаления во времени уменьшается и сходит на нет.

Один из вариантов вероятностного изменения состояния человека заключается в том, что наряду с вероятностью перехода из одного состояния, например принадлежность к возрастной группе x лет, в другое — $p(x)$, существует вероятность выбытия из этой возрастной группы, т. е. вероятность умереть, не осуществив переход в возрастную группу $x+1$ год, — q_x. В этом случае процесс перехода из одного состояния в другое и процесс выбытия (смерти) может быть описан марковской цепью. Обозначив предельный возраст дожития через ω, покажем марковский граф такого перехода (рис. 5.1).

В результате получаем вероятностные матрицы, учитывающие вероятности переходов и смерти. Представляя эти матрицы в виде нескольких матриц, можно изучать процесс выбытия по причинам смерти. Аналогичные графы и матрицы можно строить и для изучения других демографических процессов.

Идею марковских цепей в приложении к демографии покажем на примере территориальной миграции.

Систему случайных событий A_i $(i=1, 2, \ldots, s)$ вместе с соответствующими им переходными вероятностями, т. е. вероятностями перехода из i-го состояния в j-е (p_{ij}), где i — номер исходного состояния, j — номер состояния, в которое переходит система, называют цепью Маркова. Для каждого k-го перехода (шага) совокупность переходных вероятностей образует матрицу:

$$p_k = \begin{vmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{vmatrix}.$$

В каждой строке $\sum_{j=1}^{n} p_{ij}=1$.

Если матрица переходов для системы событий в одном испытании известна, то можно определить вероятность появления интересующего нас события через k испытаний:

$$P_{ij}(k)=\Sigma p_{ij}(m)\,p_{ij}(k-m),$$

где $P_{ij}(m)$ — вероятность перехода из i-го состояния в j-е состояние через k шагов.

Марков доказал, что матрица переходов на k-м шаге равна матрице переходов на первом шаге в k-й степени, т. е.

$$p_k=p_1\cdot p_1\cdots p_1=p_1^k\ .$$

Рассматривая территориальную миграцию как марковский процесс, установим, например, наличие трех групп населения:

лица, не мигрировавшие и не собирающиеся мигрировать;

лица, не мигрировавшие, но собирающиеся мигрировать;

лица, мигрировавшие.

В результате массового статистического наблюдения можно оценить вероятность перехода отдельных лиц из одной группы в другую в течение одного года (табл. 5.1). В качестве оценок вероятностей используют частости наступления соответствующих событий.

Таблица 5.1 **Вероятности различных состояний**

Исходное состояние отдельных лиц	Вероятность состояния через один год, j			
	Не мигрировавшие и не собирающиеся мигрировать	Не мигрировавшие, но собирающиеся мигрировать	Мигрировавшие	Итого
Не мигрировавшие и не собирающиеся мигрировать	0,8	0,1	0,1	1,0
Не мигрировавшие, но собирающиеся мигрировать	0,1	0,6	0,3	1,0
Мигрировавшие	0,0	0,1	0,9	1,0

Тогда матрица вероятностей однолетнего интервала будет иметь вид:

$$p_1=\begin{vmatrix} 0{,}8 & 0{,}1 & 0{,}1 \\ 0{,}1 & 0{,}6 & 0{,}3 \\ 0{,}0 & 0{,}1 & 0{,}9 \end{vmatrix}\ .$$

Исходя из формулы полной вероятности можно определить вероятность перехода из i-го состояния в k-е через два однолетних перехода по формуле

$$p_{ik(2)}=p_{1k}\cdot p_{i1}+p_{2k}p_{i2}+\cdots+p_{nk}p_{in}\ .$$

Умножаем матрицу p_1 саму на себя и получаем матрицу переходных вероятностей для двухлетнего интервала, т. е.

$$p_2=p_1\cdot p_1=p_1^2\ .$$

Находим элементы первой строки:

$$p_{11(2)}=0{,}8\cdot0{,}8+0{,}1\cdot0{,}1+0{,}0\cdot0{,}1=0{,}64+0{,}01=0{,}65;$$

$$p_{12(2)}=0{,}8\cdot0{,}1+0{,}1\cdot0{,}6+0{,}1\cdot0{,}1=0{,}08+0{,}06+0{,}01=0{,}15;$$

$$p_{13(2)}=0{,}8\cdot01+0{,}1\cdot03+0{,}1\cdot0{,}9=0{,}08+0{,}03+0{,}09=0{,}20.$$

Затем находим элементы второй строки:

$$p_{21(2)}=0{,}1\cdot0{,}8+0{,}6\cdot0{,}1+0{,}3\cdot0{,}0=0{,}14;$$

$$p_{22(2)}=0{,}1\cdot0{,}1+0{,}6\cdot0{,}6+0{,}1\cdot0{,}3=0{,}40;$$

$$p_{23(2)}=0{,}1\cdot0{,}1+0{,}3\cdot0{,}6+0{,}9\cdot0{,}3=0{,}46$$

и т. д.

Получаем матрицу:

$$p_2=(p_1)^2=\begin{vmatrix} 0{,}65 & 0{,}15 & 0{,}20 \\ 0{,}14 & 0{,}40 & 0{,}46 \\ 0{,}01 & 0{,}15 & 0{,}84 \end{vmatrix}.$$

Интересно, что, производя дальнейшее умножение матриц переходных вероятностей для трехлетнего, четырехлетнего и т. д. интервалов, получим, что через 32 года матрица переходных вероятностей покажет стабильность процесса миграции:

$$p_{32}=(p_1)^{32}=\begin{vmatrix} 0{,}1 & 0{,}2 & 0{,}7 \\ 0{,}1 & 0{,}2 & 0{,}7 \\ 0{,}1 & 0{,}2 & 0{,}7 \end{vmatrix}.$$

Из приведенного примера видно, что полученные вероятности не зависят от начального состояния.

Распределение населения по территории зависит не только от интенсивности миграционных потоков, но и от естественного движения населения. Зная территориальное распределение населения на исходный момент и динамику рождаемости и смертности населения, проживающего на этой территории, можно найти законы развития территориальной структуры населения. Введем в анализ средние коэффициенты рождаемости и смертности n_i и m_i в период сравнения территориальных структур (от t до t_1), а также вероятности миграции из территории i в территорию $j\,(p_{ij})$. Вероятности, характеризующие мигрантов с точки зрения возраста, семейного положения, профессии, образования и т. д., а также характеризующие отдельные территории по степени их индустриализации, уровню заработной платы и т. д., могут быть определены на основе выборочных наблюдений о прошлой миграционной интенсивности. Пусть распределение населения по территории на начало и конец года представлено векторами $S^0(S_1^0, S_2^0, \ldots, S_\kappa^0)$ и $S^1(S_1, S_2, \ldots, S_\kappa^1)$.

Тогда связь между структурами территориального распределения может быть выражена следующим образом:

$$S'_i = S^0{}_i \left(1 - \sum_{\substack{i=1 \\ j \neq i}}^{k} p_{ij} + n_i - m_i\right) + \sum_{\substack{j=1 \\ j \neq i}}^{k} p_{ji} S_j{}^0 .$$

Так как $S_i^0\ n_i = N_i$ и $S_i^0\ m_i = M_i$ представляют абсолютные числа рождений и смертей на территории i, получаем:

$$S_i{}^1 = S_i{}^0 \left(1 - \sum_{\substack{i=1 \\ j \neq i}}^{k} p_{ij}\right) + \sum_{\substack{j=1 \\ j \neq i}}^{k} p_{ji} S_j{}^0 + N_i - M_i .$$

Правая часть равенства состоит из трех слагаемых:

$S_i{}^0 (1 - \sum_{\substack{i=1 \\ j \neq i}}^{k} p_{ij})$ — население i-й территории на начало года за вычетом мигрировавших в остальные территории;

$\sum_{\substack{i=1 \\ j \neq i}}^{k} p_{ji} S_j{}^0$ — население, мигрировавшее из всех остальных территорий на территорию i;

$N_i - M_i$ — естественный прирост населения территории i.

Учитывая необходимость многократного перехода от одной структуры населения к другой и связанное с этим усложнение расчетов, удобнее всего выразить указанную выше связь с помощью матричного равенства $S^1 = AS^0$, где A является матрицей, имеющей вид:

$$A = \begin{vmatrix} 1 - \sum_{j=2}^{k} p_{1j} + n_1 - m_1 & p_{21} & p_{31} \cdots & p_{k1} \\ p_{12} & 1 - \sum_{\substack{j=1 \\ j \neq 2}} p_{2j} + n_2 - m_2 p_{32} \cdots & & p_{k2} \\ \cdots & \cdots & \cdots\cdots & \cdots \\ p_{1k} & p_{2k} & p_{3k} \cdots 1 \sum_{j=1}^{k=1} p_{kj} + n_k - m_k & \end{vmatrix} .$$

Для получения распределения населения по территориям через несколько лет предполагают, что p_{ij}, n_i и m_i за эти годы не испытывают заметных изменений. Тогда территориальная структура через t лет, выраженная вектором S_t, может быть получена путем умножения вектора S^0 t раз на матрицу A:

$$S^1 = S^0 A;\quad S^2 = S^1 A = S^0 A^2;\ \cdots;\ S^t = S^{t-1} A = S^0 A^t .$$

Разумеется, что большая точность расчетов может быть достигнута путем ежегодных исправлений p_{ij}, n_i и m_i.

При расчете перспективной численности населения с учетом естественного и механического движения населения К. В. Папенов привлек вероятностные показатели[1].

Используя численность населения по административным единицам и численность миграционных перемещений, автор строит диагональную матрицу распределения по миграционным группам. Далее, на основе привлечения различных вероятностных показателей (переезда из одного пункта в другой, рождаемости у мигрантов, смерти после переезда, рождаемости или смерти на данном месте жительства) строится матрица переходных вероятностей. Полученная матрица позволяет рассматривать всю совокупность демографических факторов, характеризующих процесс воспроизводства и механического движения населения, во взаимосвязи и дает практическую возможность определения перспективной численности населения по административным единицам.

Аналогично тому, как мы рассматривали территориальную миграцию, можно рассматривать некоторую совокупность семей по признаку наличия детей и намерения в отношении их рождения. Пусть получено три группы:

семьи, имеющие детей;

семьи, не имеющие детей и не собирающиеся их иметь;

семьи, не имеющие детей, но собирающиеся их иметь.

Статистическое изучение позволило оценить вероятности перехода из одной группы в другую. Процессы изменений могут быть описаны с помощью матрицы вероятностей. Матрица переходных вероятностей имеет вид (пример условный):

$$A_1 = \begin{vmatrix} p_{21} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{vmatrix} = \begin{vmatrix} 0 & 0 & 1 \\ 0{,}7 & 0{,}2 & 0{,}1 \\ 0{,}1 & 0{,}6 & 0{,}3 \end{vmatrix}$$

Здесь p_{ij} — условная вероятность того, что после k-го измерения система окажется в j-м состоянии, если после $(k-1)$-го она находилась в i-м состоянии.

Случай $p_{ij} = 0$ означает, что случайное событие, состоящее в непосредственном переходе из i-го состояния в j-е состояние, невозможно; $p_{ij} = 1$ свидетельствует о достоверности такого перехода; $\sum_{j=1}^{s} p_{ij} = 1$, иными словами, сумма элементов каждой строки матрицы равна единице.

Рассмотрим на числовых значениях:

элемент первой строки матрицы $p_{13} = 1$ свидетельствует о том, что семьи, имеющие детей, и в последующем наверняка будут их иметь;

элемент $p_{33} = 0{,}3$ означает вероятность того, что семьи, не имеющие детей, но собирающиеся их иметь, не осуществят свое намерение;

---

[1] См.: *Папенов К. В.* Модели размещения населения с учетом естественного и механического движения населения. Автореф. канд. дис. М., 1969.

элемент $p_{31}=0{,}1$ означает малую вероятность того, что семьи, не имеющие детей, но собирающиеся их иметь, осуществят свое намерение.

При построении модели прогноза состава семей И. А. Герасимова, используя цепи Маркова, предположила, что после момента t изменения в структуре семьи не зависят от предшествующей эволюции, а зависят только от состава семьи на базисный момент времени t_0 (эргодичность)[1]. Известно распределение семей по демографическим типам и вероятности того, что на момент t_0 семья принадлежит к типу i $(i=1, 2, \ldots, m)$. В матрице b_i — это постоянные во времени вероятности того, что семья, принадлежавшая к моменту времени t к i-му типу, перейдет к моменту времени $t+\tau$ в j-й тип. По материалам обследования в Костроме 743 семей И. А. Герасимова построила два варианта одногодичных матриц переходных вероятностей, которые позволили ей определить стационарные распределения семей по демографическим типам и сделать ряд интересных выводов. При моделировании семейной структуры населения И. А. Герасимова выделила четыре этапа ее развития:

период плодовитости — увеличения численностей семей;

стабильный период существования семей;

период выделения из семей взрослых детей и образование новых семей;

период затухания семьи (смерть обоих супругов).

Границы этих периодов, длительности пребывания семей в том или ином состоянии изучались вероятностным путем. Были введены в анализ такие, например, вероятности: вероятность того, что в семье с фиксированным числом рождений первые j рождений произойдут до определенного момента, а рождения $j+1$, $j+2$ и т. д. — после него; условная вероятность того, что возраст ребенка в момент t меньше v и др.

Г. А. Павлов применил марковские процессы к изучению изменения брачной структуры женщин[2]. Общая численность женщин S представлена автором следующим образом:

$$S=C_0+C_1+B+D+W,$$

где C_0 — численность женщин, не достигших бракоспособного возраста; C_1 — численность женщин в бракоспособном возрасте, никогда не состоявших в браке; B — численность женщин, состоящих в браке; D — численность разведенных; W — численность вдов. Выход из-под наблюдения означает смерть женщины.

Связь между начальной брачной структурой женщины и конечной (через один год) выражается следующим матричным равенством:

$$S^1=A_0S^0,$$

[1] См.: *Герасимова И. А.* Структура семьи, с. 161 и далее.

[2] См.: *Павлов Г. А.* Система информации о населении и демографический анализ. Автореф. канд. дис. М., 1972, с. 9—11.

где A — матрица вида:

$$A=\begin{vmatrix} p_{00} & \Pi_1 L & \Pi_2 L & \Pi_3 L & \Pi_4 L \\ p_{01} & p_{11} & 0 & 0 & 0 \\ 0 & p_{12} & p_{22} & p_{32} & p_{42} \\ 0 & 0 & p_{23} & p_{33} & 0 \\ 0 & 0 & p_{24} & 0 & p_{44} \end{vmatrix}.$$

Здесь p_{ii} — это вероятности остаться женщине в том же брачном состоянии через год: $p_{ii}=1-\Sigma p_{ij}-M_i$;

p_{ij} — вероятности смены брачного состояния, т. е. перехода из i-го состояния в j-е;

Π_i — общий коэффициент рождаемости ($\Pi_0=0$);

M_i — общий коэффициент смертности;

L — вероятность новорожденной дожить до конца года.

Параметры Π_i и L включены в первую строку матрицы вероятностей для определения пополнения контингента женщин, не достигших бракоспособного возраста, за счет естественного прироста за год.

Г. А. Павлов рассматривает вероятности перехода как величины, полностью зависящие от исходной брачной структуры.

А. Я. Боярский применил цепи Маркова для прогноза распределения семей по размеру. Такое же применение цепей Маркова находим у В. В. Кашки и А. П. Великого, А. Х. Карапетяна и др.

На марковские процессы может также опираться построение межотраслевых моделей движения трудовых ресурсов. В соответствии с этим сначала следует зафиксировать численность и состав работников на конец тех или иных временных интервалов и определить вероятности перехода (перемещения) из одного состояния (категории) в другое. Если предположить, что формирование потоков перехода из одного состояния в другое происходит в соответствии, например, с биномиальным законом распределения и вероятности перехода представляют собой среднестатические величины из индивидуальных вероятностей в различии пола, возраста, образования, стажа работы и др., то в этом случае средние и дисперсии должны быть взвешенными. При минимизации средней ошибки следует учитывать, что главное условие, которому должны удовлетворять искомые веса m_i (сумма их равна единице), состоит в том, что они должны быть обратно пропорциональны дисперсиям:

$$m_1 : m_2 : \cdots : m_n=\frac{1}{\sigma_1^2}:\frac{1}{\sigma_2^2}:\cdots:\frac{1}{\sigma^2_n}.$$

В этом случае дисперсия равна:

$$\sigma_1^2=p_i(1-p_i)N_i.$$

Построенная модель даст возможность оценить влияние изменения каждого межотраслевого потока трудовых ресурсов на итоговую отраслевую структуру занятости.

В Норвежском центральном статистическом управлении совершенствовалась расчетная модель перспективной численности населения. Одной из решенных при этом задач явилось создание модели вероятности бракосочетаний и плодовитости, в том числе для изучения первых родов в течение первого брака.

Наблюдению были подвергнуты женщины, вышедшие замуж в возрасте y лет и живущих в браке u лет, и бездетные женщины в возрасте $y+u$ лет. Предполагалось наличие шести состояний: нулевое состояние; первые роды; смерть; овдовение; расторжение брака; выбытие (эмиграция).

Применялись цепи Маркова. Вероятность $p_k(y, u, t)$ указывала, что женщина в возрасте y лет, живущая u лет в браке, в момент t находится в состоянии k.

5.3. Переход от живущих в возрасте x лет к возрасту $x+1$ лет

Метод передвижки возрастов был разработан в советской демографической науке в 20-х годах. Примерно в это же время появилась работа Р. Кучинского. Авторы этого метода показали, что на основе его использования с привлечением показателей режима воспроизводства населения можно осуществлять перспективные расчеты с высокой степенью точности.

Передвижкой возрастов называют такой переход от живущих в возрасте x лет к возрасту $x+1$ год, при котором учитывается уменьшение их численности вследствие смертности и увеличение вследствие рождаемости. Коэффициенты дожития имеют вид:

$$P_{x+\frac{t}{x}}=\frac{L_{x+t}}{L_x},$$

где L_x — числа живущих. Приближенно

$$L_x=\frac{l_x+l_{x+1}}{2}.$$

Производя передвижку возрастов, мы получаем динамическую модель воспроизводства населения. При этом, конечно, правильность расчетов зависит от точности используемых коэффициентов, входящих в модель, от гипотез, допускаемых исследователем, и дальности перспективы.

При расчетах перспективной численности населения всего Советского Союза и многонаселенных республик метод передвижки возрастов дает хорошее приближение к действительности. При использовании же этого метода применительно к населению краев, областей и других более мелких территорий, где численность населения меньше 2,0—2,5 млн. человек, нужно иметь в виду недостаточную точность показателей вероятности дожития и рождения, привлекаемых из таблиц плодовитости и смертности. Именно поэтому для увеличения точности этих показателей нужно учитывать сходство населения различных территорий в экономическом развитии, в демографических особенностях и строить таблицы смертности и плодовитости по объединенным исходным территориям. Тогда точность перспективных расчетов увеличивается.

При прогнозировании численности населения Литовской ССР на 1976—2000 гг. использовалась передвижка возрастов с учетом миграции.

При перспективных расчетах численности пенсионеров по старости до настоящего времени используется обычная методика передвижки возрастов. Реализация этой методики требует привлечения распределения пенсионеров по возрастам в t-м году (l_x^t) и двух вероятностей: p_x — вероятность дожития (из таблиц смертности) и v_x — вероятность выхода на пенсию.

Расчет Σl_x^{t+1}, т. е. общей перспективной численности всех пенсионеров в $t+1$ году, производится по формуле

$$\Sigma l_x^{t+1}=\Sigma l_x^t p_x+\Sigma S_x^{t+1},$$

где S_x^{t+1} — возрастной состав населения, вступающего в пенсионный возраст в перспективном году.

С. А. Карапетян и Р. Ш. Назарей предложили новую методику расчета численности пенсионеров по старости, ориентированную на использование не возрастного распределения пенсионеров, а их распределения по стажу получения пенсии — R_y. При этом, как показали авторы, процедура расчета упрощается.

На передвижке возрастов основан разработанный А. С. Семеновой перспективно-балансовый метод, используемый для оценки демографических последствий войн[1]. Суть этого метода состоит в использовании вероятностных показателей таблиц смертности, отражающих довоенный режим воспроизводства населения, и передвижке возрастов от довоенной переписи населения до послевоенной.

Привлекая данные о возрастно-половом составе населения до войны и производя указанную выше передвижку возрастов, а затем определяя с помощью таблиц повозрастной рождаемости число предполагаемых рождений в военный период, можно сопоставить полученные расчетные данные, в которых элиминировано влияние войны (данные, отвечающие на вопрос, какова была бы численность населения, если бы не было войны) с фактическими данными послевоенной переписи населения, в которых влияние войны не устранено. Результат сопоставления укажет на демографические потери.

Этот способ, примененный А. С. Семеновой для расчета демографических потерь всех стран — участниц первой мировой войны, позволил ей оценить общие демографические потери примерно в 50 млн. человек, в числе которых прямые военные потери

[1] См.: *Семенова А. С.* Демографические потери в первой мировой войне. — В кн.: Ученые записки МЭСИ, т. 6. Труды кафедры демографии. М., 1965, с. 87.

составили 10 млн. мужчин (в возрасте 20—39 лет), потери в результате увеличения смертности гражданского населения (включая повышенную детскую смертность) — 19 млн. человек, «недородившиеся» (косвенные потери вследствие снижения рождаемости в годы войны) — 22 млн. детей. Такой «...счет погубленных и не начавшихся жизней... достаточно убедительно говорит о цене, уплачиваемой народами за кровавые затеи империалистов»[1].

5.4. Моделирование брачности

Американский демограф Э. Коул разработал методику моделирования брачности. При этом «чистая» вероятность вступления в брак в возрасте x лет вычислялась по формуле

$$b'_x = \frac{s'_x - s'_{x+1}}{s'_x},$$

где s'_x и s'_{x+1} — числа девушек соответственно в возрасте x и $x+1$ лет.

Стандартизованная вероятность вступления в первый брак, или вероятность вступить в брак в возрасте x лет для женщин, никогда не состоявших в браке, но исчисленная только для тех женщин, которые вообще вступят в брак, равна:

$$r_x = \frac{s'_x - s'_{x+1}}{s'_x - s'_\omega},$$

где ω — возраст, в котором вступление в первый брак полностью прекращается.

Производя преобразования, Э. Коул получил

$$b'_x = \frac{s'_x - s'_\omega}{s'_x} r_x.$$

Эмпирическая формула, связывающая стандартную вероятность с начальным возрастом вступления в первый брак — x_0 и параметром, характеризующим скорость протекания процесса вступления в первый брак соответствующего поколения женщин,

$$r_x = \frac{0{,}174}{k} \cdot e^{-4{,}411 \cdot e^{-\frac{0{,}309}{k}(x - x_0)}}.$$

Используя данную методику, М. С. Тольц, привлекая материалы переписи населения 1897 г. по Европейской России, определил значения параметров $x_0 = 15{,}0$ и $k = 0{,}58$. Окончательная доля невступивших в брак s'_ω оценивалась как средняя арифметическая для групп 40—49 и 50—59 лет[2].

Для построения комбинированной таблицы брачности, в которой наряду с брачностью учитывается также и смертность,

[1] *Семенова А. С.* Демографические потери в первой мировой войне. — В кн.: Ученые записки МЭСИ, т. 6, с. 120.

[2] См.: *Тольц М. С.* Брачность населения России в конце XIX — начале XX в. — Там же, с. 153.

М. С. Тольц принял смертность девушек по Европейской России равной смертности женщин всех брачных состояний. Следовательно, в распоряжении автора имелись b^1_x из чистой таблицы брачности и q_x из таблицы смертности женского населения Европейской России за 1896—1897 гг., построенной С. А. Новосельским.

Расчет зависимых вероятностей вступления в брак в интервале $x/x+1$ автор производил по формуле

$$b_x = b^1_x \left(1 - \frac{q^1_x}{2}\right),$$

а зависимой вероятности умереть, не вступив в брак, в интервале $x/x+1$ по формуле

$$q_x = q^1_x \left(1 - \frac{b^1_x}{2}\right).$$

Числа вступивших в брак с учетом смертности составили:

$$B_x = S_x b_x,$$

числа умерших девушек:

$$M_x = S_x q_x.$$

Число девушек в начале каждого возраста определялось из соотношения:

$$S_{x+1} = S_x - B_x - M_x.$$

И. П. Ильина, привлекая таблицы брачности для реальных поколений женщин 1913—1947 гг. рождения и вероятности вступления в первый брак для женщин СССР трех групп поколений (1913—1917, 1918—1922 и 1923—1927 гг.), проанализировала нарушение процесса заключения браков в результате Великой Отечественной войны[1].

Используя эмпирическую формулу Э. Коула для получения r_x женщин поколения 1913—1917 гг. рождения, автор нашла две оценки — x_0 и k. Для возрастов 19, 21 и 23 лет они составили: $x_0 = 15{,}46$ и $k = 0{,}5658$; для возрастов 18, 20 и 22 года — $x_0 = 14{,}58$ и $k = 0{,}7379$. Используя средние из этих величин, автор получила $\bar{x} = 15$ и $\bar{k} = 0{,}6519$.

Расчет r_x был осуществлен для значений x в границах возрастов $15 < x < 30$. Для $x > 30$ расчет стандартизованной вероятности производился по экспоненте

$$r_{x+1} = r_x e^x.$$

Полученные результаты позволили автору оценить влияние войн и сделать вывод, что в военные годы для женщин рождений 1913—1917 гг. и 1918—1921 гг. имел место «провал» возрастных вероятностей вступления в брак, выразившийся в снижении брачности. Для женщин рождения 1923—1927 гг. вероятности вступ-

[1] См.: *Ильина И. П.* Изучение брачности поколений женщин из семей рабочих и служащих в СССР. — В кн.: Рождаемость. М., 1976, с. 133.

ления в брак были очень низкими, свидетельствовавшими об откладывании браков. Окончательные выводы автора состоят в фиксации пагубного влияния, оказываемого войнами на процесс возникновения и распадания браков в нашей стране: нарушение брачности в каждом поколении в результате войн необратимо.

5.5. Модели экспоненциального, стационарного и стабильного населения

Укажем на некоторые демографические модели, при построении которых авторы привлекали вероятностные показатели.

Модель экспоненциального населения А. Лотки исходит из постоянства долей лиц (c_x) в возрасте между $x+dx$ и вероятностей дожития в возрасте x лет:

$$b=\frac{1}{\int_0^\infty e^{-rx}p_x dx}.$$

В этой формуле кроме показателей b (коэффициента рождаемости) и r (коэффициента прироста, равного разности между b и коэффициентом смертности d) фигурирует вероятностный показатель дожития в возрасте x, обозначаемый p_x.

Вариация значений c_x и p_x приводит к различным видам экспоненциальных моделей. Частным случаем экспоненциального населения является стационарное население. Формула этой модели вытекает из приведенной выше при $r=0$. Тогда $b=d$ и

$$b=d=\frac{1}{\int_0^\infty p_x d^x}=\frac{1}{e_0},$$

где e_0 — средняя продолжительность предстоящей жизни при рождении.

Рассмотрим построение простой модели стационарного населения и обобщенной модели стабильного населения, предложенное А. Я. Боярским[1]. Для стационарного населения характерна неизменность численности и половозрастного состава. Следовательно, в таком населении:

1) отсутствует внешняя миграция;

2) сохраняется неизменный порядок вымирания (изменение численности исходной совокупности для мужчин l_x^m и для женщин l_x^f под влиянием смертности в различных возрастах d_x);

3) неизменна плотность рождений (при неизменной доле девочек Q), т. е. число рождений в единицу времени, например один год.

В стационарном населении в возрасте x лет у мужчин и женщин в любой момент находится одно и то же число лиц.

[1] См.: Марксистско-ленинская теория народонаселения/Под ред. Д. И. Валентея. 2-е изд. М., 1974, с. 167.

Облекая условия, сформулированные выше, в математическую формулу, имеем: в возрасте от x до $x+dx$ лет находятся родившихся x лет назад за период времени dx:

$$Ndx=(1-Q)Ndx+QNdx,$$

где Q и $1-Q$ — соответственно доли девочек и мальчиков среди родившихся.

До данного момента t дожило мальчиков $(1-Q)Nl_x^m dx$ и девочек $QNl_x^f dx$.

При переходе к одногодичному интервалу возраста от x до $x+1$ получаем сумму чисел, доживших в стационарном населении:

$$\int_x^{x+1}(1-Q)Nl_x^m dx+\int_x^{x+1}QNl_x^f dx.$$

Учитывая, что

$$\int_x^{x+1} l_x^m dx=L_x^m \text{ и } \int_x^{x+1} l_x^f dx=L_x,$$

получаем численность населения в одногодичном интервале:

$$(1-Q)NL_x^m+QNL_x^f.$$

Для сохранения условия о постоянстве «порядка вымирания» для двух полов используем среднюю из чисел доживающих, взвешенную по величинам $1-Q$ и Q, и получаем:

$$L_x=L_x^m(1-Q)+L_x^f Q.$$

Тогда общая численность стационарного населения в возрасте от x до $x+1$ окажется равной:

$$P=NL_0+NL_1+\ldots+NL_\omega=N\Sigma L_x=N\int_0^\infty l_x dx.$$

Так как второй множитель есть средняя продолжительность жизни, т. е. e_0, то $P=Ne_0$. В этом случае коэффициент рождаемости (n) равен коэффициенту смертности (m) и равен обратному значению средней продолжительности жизни:

$$n=m=\frac{N}{Ne_0}=\frac{1}{e_0}.$$

Гипотеза стационарного населения может быть развита и превращена в гипотезу стабильного населения.

При решении задачи обобщенной теории стабильности населения, состоящей в определении показателей повозрастной рождаемости при данном порядке вымирания проф. А. Я. Боярский принял во внимание следующие показатели и взаимосвязи между ними[1]:

[1] См.: *Боярский А. Я.* Об обратной задаче общей теории стабильного населения. — В кн.: Модели демографических связей. М., 1972, с. 6—13.

p_x — вероятность дожития до конца предстоящего года для лиц возраста x лет;

Φ_x — годовое число рождений на одно лицо;

r_x — число детей, которые родятся в предстоящем году в расчете на одно лицо, родившееся x лет назад и дожившее до конца года рождения.

При этом $r_x = p_0 p_1 \dots p_{x-1} \cdot \Phi_x$.

Учитывая, что в стабильном населении ежегодные числа рождений изменяются в отношении h, общее число родившихся в предстоящем году, уменьшенное на число умерших до конца этого года, составит:

$$Nh^{-1}r_0 + Nh^{-2}r_1 + \dots + Nh^{-(\omega+1)}r_\omega .$$

В обзорном докладе на Всемирной конференции по народонаселению Л. Таба привел построение простейшей стохастической модели, основанной на гипотезе о соответствующем изменении коэффициентов рождаемости и смертности и предположении о независимости этих показателей от времени[1].

Пусть b и d — коэффициенты рождаемости и смертности замкнутого (без миграций) женского населения при соответствующей нормальной половой структуре населения; $n_t = n$ — численность населения в момент t; $p_n(t+dt)$ — вероятность того, что в момент $t+dt$ численность населения будет составлять $n = n_t$.

Учтем следующее приближенное равенство:

$$p_n(t+dt) \approx p_{n+1}(t)\, d(n+1)\, dt + p_{n-1}(t)\, b(n-1)\, dt + \\ + p_n(t)\,[1-(b+d)\,dt]^n ,$$

где $p_{n+1}(t)\,d(n+1)\,dt$ — вероятности того, что n лиц, живущих в момент времени $t+dt$, происходят из группы $n+1$ лиц, живших в момент времени t, в которой в интервале времени dt имел место один случай смерти; $p_{n-1}(t)\,b(n-1)\,dt$ — вероятность того, что n лиц, живущих в момент времени $t+dt$, происходят из группы $n-1$ лиц, живших в момент времени t, в которой в интервале времени dt произошло одно рождение; $p_n(t)\,[1-(b+d)\,dt]^n$ — вероятность того, что среди n лиц, живших в момент t, в интервале времени от t до $t+dt$ не имело места ни одного случая смерти или рождения.

Решая полученные дифференциальные уравнения и производя при этом различные преобразования, можно определить вероятность того, что население со временем вымрет (при $t \longrightarrow \infty$):

$$p_0(t) = \frac{de^{rt} - d}{be^{rt} - d} .$$

В зависимости от величины коэффициентов естественного прироста мы имеем три случая: 1) $r<0$ (отрицательный прирост); 2) $r=0$; 3) $r>0$.

[1] См.: Демографические модели. Сб. статей/ Под ред. Е. М. Андреева и А. Г. Волкова. М., 1977, с. 77—78.

В первом и втором случаях вероятность стремится к единице: $p_0(t) \longrightarrow 1$. Это означает вымирание населения; полученный при этом вывод для второго случая дает результат, ранее даже и не предполагаемый: оказывается, что «стационарное» население с равными b и d *почти достоверно вымирает.* Да и в третьем случае вероятность вымирания населения не равна нулю: $p_0(t)$ при $t \longrightarrow \infty$ равна $\left(\frac{d}{b}\right)^r$; чем больше r, тем вероятность меньше. Из этой модели вытекает, что численность населения при возрастании t не стремится асимптотически к какой-то величине.

Использование вероятностей смерти стационарного населения позволяет ответить на вопрос, очень интересующий демографов развивающихся стран: каково влияние контроля над рождаемостью на динамику численности населения? Будет ли, учитывая инерцию демографических процессов, в условиях контроля над рождаемостью численность населения продолжать расти теми же темпами, что и раньше, до введения контроля, или темп роста численности населения начнет сразу же резко снижаться? Что при этом будет происходить с численностью населения? Попытку ответить на эти вопросы предпринял Н. Кейфитц (Гарвардский университет). Сначала автор устанавливает некоторые факты:

1) возрастное распределение быстро растущего населения является благоприятным для ускорения роста;

2) чтобы коэффициент воспроизводства находился на уровне единицы ($R_0 \approx 1$), нужно, чтобы все супружеские пары, у которых могут родиться дети, имели в среднем на семью 2—3 ребенка.

Далее вводится условие, что ограничение рождаемости предпринято одновременно и число рождений снижается до уровня, необходимого для простого воспроизводства.

Тогда исходя из вероятностей смерти стационарного населения, а также предположения, что возрастное распределение населения более или менее стабильно, привлекая среднюю продолжительность жизни, рассчитываем предельную численность стационарного населения, а затем, что самое главное, отношение этой предельной численности населения к численности населения в момент введения ограничения рождаемости и снижения прироста. Используя в качестве примера данные для Эквадора за 1965 г., автор сформулировал следующие утверждения, отвечающие на поставленный вопрос: «...При резком снижении рождаемости до уровня простого воспроизводства население Эквадора и других демографически сходных с ним стран увеличится примерно на две трети, прежде чем будет достигнуто стационарное состояние»[1].

Широко, но большей частью неудачно применялись вероятностные показатели при попытке разрешить противоречие, возникшее из рассмотрения нетто-коэффициента воспроизводства.

[1] *Кейфитц Н.* Возмещение долга. Приложение к проблемам миграции и контроля над рождаемостью.— В кн.: Математика в социологии. Моделирование и обработка информации. Пер. с англ. М., 1977, с. 504.

Следует иметь в виду, что для стабильного населения коэффициенты естественного прироста женского и мужского населения, полученные из модели нетто-коэффициента воспроизводства, должны быть равны друг другу. Между тем демографо-статистический анализ фактических данных для мужчин и женщин приводил к различным результатам. При разрешении этого «конфликта» между показателями нетто-коэффициента воспроизводства для мужчин и женщин были предложены различные методы. Остановимся на тех из них, которые основаны на привлечении вероятностных показателей.

А. Поллард разработал модель, в основе которой лежали две вероятности: $\Psi(x)$ — вероятность того, что родившийся мальчик станет в x лет отцом девочки, и $\xi(y)$ — вероятность того, что родившаяся девочка в возрасте x лет родит мальчика. На основе этой формальной модели А. Поллард вычислил индекс воспроизводства, уже единый для обоих полов. Л. Е. Дарский по этому поводу пишет: «Это решение... ни на чем не основано, кроме бессмысленной гипотезы о женщинах, рождающих мальчиков, и мужчинах, рождающих девочек»[1].

так ли?

Вторая неудачная попытка разрешения «конфликта» была предпринята П. Кармеллом. Он ввел в анализ кроме показателей брачности такие вероятностные оценки, как $M_M(x, y)$ — вероятность того, что новорожденный мальчик к возрасту y будет жив и женат на женщине возраста x, и $M_F(x, y)$ — вероятность того, что новорожденная девочка к возрасту x будет жива и замужем за мужчиной возраста y.

Х. Хирениус, используя вероятностные показатели рождения женщинами в различных возрастах, предпринял попытку построения коэффициента плодовитости для когорт, закончивших репродуктивную деятельность. Привлекалась вероятность того, что женщина, вышедшая замуж в возрасте x, родит первого ребенка в возрасте y и др.

Таким образом, мы видим, что не все демографические модели, основанные на вероятностном подходе, имеют практическую ценность. Привлечение конкретных статистических данных и выработка на их основе с использованием соответствующих моделей выводов и рекомендаций не всегда полностью удачны. Дело в том, что допущения исследователей часто не соответствуют реальной действительности и противоречат друг другу, а отобранные для включения в модель факторы не являются существенными.

Работу в этом направлении, более отвечающую практическим потребностям, следует продолжать.

5.6. Перспективные расчеты в ЦСУ СССР

В ЦСУ СССР перспективные расчеты населения проводятся методом компонентов, предполагающим использование возрастно-

[1] *Дарский Л. Е.* Формирование семьи, с. 22.

половой структуры населения, вероятностных показателей смерти, коэффициентов дожития, показателей миграции и др.

Рассмотрим общую формулу для расчета в t-м календарном году перспективной численности населения на $t+1$ год. При этом используем обозначения: α — городское или сельское население, x — возраст, β — пол.

$$S_{x+1}^{\alpha\beta(t+1)} = S_x^{\alpha\beta t} p_x^{\alpha\beta t} \pm B^t v^{\beta t} b_x^{\beta t} \pm$$
$$\pm R^t \rho^{\beta t} r^{\beta t} \pm \frac{S_x^{(\text{гор.+село})\beta t}}{\Sigma S_x^{(\text{гор.+село})\beta t}} \cdot M^t .$$

Численность детей в возрасте до года рассчитывается по формуле

$$S_0^{\alpha\beta(t+1)} = \frac{1}{2} d \Sigma F_x^{\alpha t} (S_x^{\alpha t} + S_{x-1}^{\alpha t} \cdot p_{x-1}^{\alpha t}).$$

Численность населения в возрасте 100 лет и старше составит:

$$S_{100}^{\alpha\beta(t+1)} = S_{99}^{\alpha\beta t} p_{99}^{\alpha\beta t} + S_{100}^{\alpha\beta t} p_{100}^{\alpha\beta t},$$

где $S_x^{\alpha\beta t}$ — численности однолетних возрастных групп раздельно по городу и селу и по полу на начало календарного года;
$p_x^{\alpha\beta t}$ — повозрастные коэффициенты дожития;
$q_x^{\alpha\beta t}$ — вероятность умереть;
p_N — коэффициенты дожития новорожденных до конца календарного года;
$F_x^{\alpha t}$ — повозрастные коэффициенты плодовитости;
d — доля девочек среди родившихся;
B^t — сальдо миграции между селом и городом;
R^t — сальдо миграции между отдельными территориями;
M^t — численность жителей сельских населенных пунктов, преобразованных в городские;
$v^{\beta t}$ и $b_x^{\beta t}$ — распределение мигрантов по полу и возрасту во внутритерриториальной миграции;
$\rho^{\beta t}$ и $r_x^{\beta t}$ — распределение мигрантов по полу и возрасту в межтерриториальной миграции.

Если прогнозируемые явления инерционны и в какой-то мере сознательно управляемы (а это почти всегда наблюдается в демографии), то состояние явления в будущем (с малым временем упреждения) в большой степени предопределено его историей и режимом управления.

В НИИ ЦСУ СССР разработан новый подход к снижению систематической ошибки прогноза и выбору модели, а также к оценке ее параметров. Суть этого подхода состоит в том, что вид функции, описывающей тренд, не задается априорно, а определяется по статистическим данным и указаниям о последующих изменениях. Таким образом, прогнозирование не ограничивается простой экстраполяцией, основанной на предположении, что динамика изучаемого процесса не претерпит изменений. В условиях, когда заведомо известно, что динамика процесса будет так или иначе изменяться, надо сначала выявить возможные факторы, воздействующие на явление и управляющие его динамикой, не содержащиеся в предыстории изучаемого процесса, определить силу их воздействия и, что самое главное, выявить *наиболее вероятную тенденцию развития явления*.

В ряде случаев достоверная информация о тенденциях демографических явлений содержится в различных плановых документах, устанавливающих на будущее величину явлений, связанных с изучаемыми процессами. Кроме того, ведущаяся статистика систематических ошибок прогнозов позволяет корректировать эти прогнозы и повышать их точность. Статистический подход к показателям воспроизводства населения достаточно крупных территорий, например населения всей страны, позволяет представить эти показатели как взвешенные средние из соответствующих показателей более мелких территорий, например населения отдельных республик страны. В качестве статистических весов в данном случае можно использовать численность населения республик.

?С

Привлекая динамику показателей воспроизводства населения республик за определенный период времени, заменив абсолютные значения этих показателей темпами роста, а затем, осреднив найденные темпы, можно выйти на темп роста показателей воспроизводства населения всей страны. Это позволяет подойти к прогнозированию темпов роста показателей воспроизводства населения страны как множественной функции от темпов роста соответствующих показателей для населения отдельных союзных республик.

5.7. Построение различных моделей на основе вероятностных показателей

При расчете *перспективной численности пенсионеров* используются следующие показатели, установленные отдельно для мужчин и женщин: $A_{x,t}$ — число пенсионеров в возрасте x лет в t-м году; q_x — взятые из таблиц смертности вероятности смерти; z_x — рассчитанные на основе специальных обследований вероятности выбытия (выезда). Расчет последующих состояний при переходе из одного возраста в другой производится в конце года по формуле

$$A_{x+1,\ t+1} = A_{x,\ t}(1 - q_x - z_x).$$

Для получения общего числа пенсионеров используем формулу

$$A_{t+1} = \sum_{x}^{100} A_{x,\ t}(1 - q_x - z_x).$$

Для расчета перспективной численности инвалидов указанные выше показатели и вероятности модифицируются.

Обычно при построении модели прогноза численности пенсионеров и пенсионных выплат используют показатели таблиц смертности, вероятности инвалидности и потери кормильцев. Покажем пример построения модели для мужчин и женщин отдельно. Для этой цели используем следующую систему обозначений:

$A^a_{x,t}$ — число пенсионеров по старости в возрасте x лет в t-м году;
q_x — вероятность умереть на x-м году жизни;
z_x — вероятность выбытия из числа пенсионеров (вследствие миграции);
$l^a_{x,t}$ — численность инвалидов в возрасте x лет в t-м году;
α_x — вероятность выбытия из числа инвалидов по различным причинам.

Тогда расчетные формулы для различных категорий примут вид:

1) ожидаемая численность пенсионеров для n-го года:

$$A^a_{x+n,\ t+n} = A^a_{x+n-1,\ t+n-1}(1 - q_{x+n-1} - z_{x+n-1});$$

2) ожидаемая численность инвалидов для n-го года:

$$l^a_{x+n,\ t+n} = l^a_{x+n-1,\ t+n-1}(1 - \alpha_{x+n-1}).$$

Рассмортим некоторые *модели плодовитости*. В модели Ж. Буржуа-Пиша дан биометрический метод, позволяющий измерить плодовитость с помощью четырех показателей, из которых один носит вероятностный характер: вероятность зачатия у новобрачных, возраст окончания репродуктивного периода (мало изменяющиеся величины) и переменные факторы, характеризующие поведение в настоящий момент — рождения по возрасту матери и браки по возрасту невест.

В модели Л. Анри, привлекающего переписные данные о числе женщин по возрасту и числу детей, рожденных живыми к моменту переписи, фигурируют вероятности увеличения семьи (a_k):

$$a_k = \frac{\overline{N}_{k+1}}{\sum\limits_{n} a_{k,\ n} N_{k,\ n}},$$

где $\overline{N_{k+1}}$ — число женщин, родивших $k+1$ раз в течение того или иного года;
$N_{k,n}$ — число женщин, родивших k-го ребенка n лет назад;
$a_{k,n}$ — доля рождений порядка $k+1$, происшедших через n лет после k-го рождения.

В Московском экономико-статистическом институте В. В. Шураковым и Я. Л. Ципесом проводился опытный расчет перспективной численности населения и отдельных его категорий по республикам Советского Союза с использованием ЭВМ, давший хорошие результаты[1]. В качестве исходной информации для решения задачи принималась возрастно-половая структура населения

[1] См.: *Шураков В., Ципес Я.* Опыт расчета перспективной численности населения на ЭВМ «Минск». — Вестник статистики, 1964, № 2, с. 65—72.

по переписи на 15 января 1959 г. При этом в ЭВМ были введены следующие данные по каждой республике:

численность населения по возрасту и полу, по городу и селу;

коэффициенты дожития каждой возрастной группы;

коэффициенты рождаемости женщин в возрасте 15—49 лет по городу и селу;

коэффициенты дожития родившихся до конца календарного года;

коэффициенты дожития родившихся до одного года;

доля мальчиков и девочек в общем числе родившихся по городу и селу.

В ряде случаев помимо указанных факторов при перспективных расчетах принимаются во внимание также предполагаемые изменения уровней рождаемости и смертности, миграции и др.

Д. Г. Кендел построил вероятностную модель изменения численности и структуры населения, основанную на предположениях, не соответствующих реальной действительности (например, независимость порядка вымирания от времени и возраста). Аналогичные вероятностные модели раздельно для женского и мужского населения построили Гудмен и Джоши. На основе этих моделей авторы пришли к выводу, что отношение математических ожиданий чисел мужчин и женщин стремится к некоторой постоянной величине, т. е. в пределе устанавливается постоянная доля мужчин.

При построении модели, характеризующей *механизм формирования социально-профессиональных групп,* Э. К. Васильева использовала вероятности межгрупповых переходов, а также вероятности вовлечения молодежи в состав занятых в общественном производстве и выбытия по разным причинам[1]. Приняты следующие обозначения: S — общая численность занятых; S_i — численность i-й социально-профессиональной группы; p_i — вероятности выбытия из i-й группы или включения в нее и межгрупповых переходов населения.

Полученная функция имела вид:

$$S_i = f(p_1, p_2, \ldots, p_n); \quad S = \Sigma S_i \ .$$

Построение вероятностных моделей для воспроизведения основных черт трудовой биографии поколения производилось Э. К. Васильевой аналогично тому, как это делается в демографических моделях. Величины исходных вероятностей определялись путем использования анамнестических обследований населения, характеризующих движения населения за предшествующие годы. Например, привлекались данные о числе лиц, у которых к моменту достижения возраста x лет были изменения в социально-профессиональной принадлежности, и о числе лиц, у которых изменений не было, и определялись соответствующие этим событиям вероятности. И так по всем показателям.

[1] См.: *Васильева Э. К.* Социально-экономическая структура населения СССР. М., 1978, с. 176—177.

Расчет вероятностей перемещения производился по Ленинграду. За базу, т. е. знаменатель вероятностей, были приняты численности всех полученных автором 13 групп в 1960 г. с расчленением каждой группы по полу. В качестве числителя фигурировали численности лиц, доживших до 1975 г. и сохранивших принадлежность к той или иной группе, а также лиц, перешедших за период с 1960 по 1975 г. в другие группы. Найденные вероятности перемещений, а также привлеченные вероятности смерти и миграции для периода 1960—1975 гг. были распространены на последующий 15-летний период 1975—1990 гг. Таким образом, Э. А. Васильева произвела перспективный расчет ожидаемой численности всех выделенных групп на основе следующей системы формул:

$$S_1^2 = (S_0^1 q^{1\to 2} + S_0^2 q^2 + S_0^3 q^{3\to 2} + \ldots + S_0^{13} q^{13\to 2})\, p^2 d^2 \ ;$$

$$S_1^3 = (S_0^1 q^{1\to 3} + S_0^2 q^{2\to 3} + S_0^3 q^3 + \ldots + S_0^{13} q^{13\to 3})\, p^3 d^3 \ ; \text{ и т. д.,}$$

где $S_0{}^1$, $S_0{}^2$, $S_0{}^3, \ldots, S_0{}^{13}$ — численность населения в каждой из 13 групп на начало 1975 г.;

$p_1, p_2, p_3, \ldots$ — вероятности дожития до 1990 г.;

$d^1, d^2, d^3, \ldots$ — вероятности изменения численности группы за 15-летний период за счет миграции;

$q^{1\to 2}, q^2, q^{2\to 3}, \ldots$ — вероятности перемещений из одной группы в другую: первый индекс обозначает номер группы, из которой переходят, а второй — номер группы, в которую осуществляется переход.

На основе построенных моделей был сделан ряд интересных выводов.

При моделировании и определении *расчетной численности иногороднего населения, обслуживаемого культурно-бытовыми учреждениями города — центра системы расселения* (H_{ij}), используются статистические данные о частоте поездок. По результатам обследования учреждений и предприятий, обслуживающих население, и по данным о фактической численности пригородного населения изучается динамика доли пригородного населения, тяготеющего к городу, в его общей численности, принятой за величину вероятности. Распределение вероятностей поездок в город-центр изучается по формуле гиперболы

$$p(H_{ij}) = \frac{A}{R^n{}_{ij}} \ ,$$

где R_{ij} — расстояние от i-го поселения до j-го города-центра; A и n — постоянные параметры, определяемые методом наименьших квадратов.

Расчетная численность искомой категории населения устанавливается в зависимости от распределения вероятностей поездок по формуле

$$H_{ij} = \Sigma\, H_i \overline{p}_{ij} \ ,$$

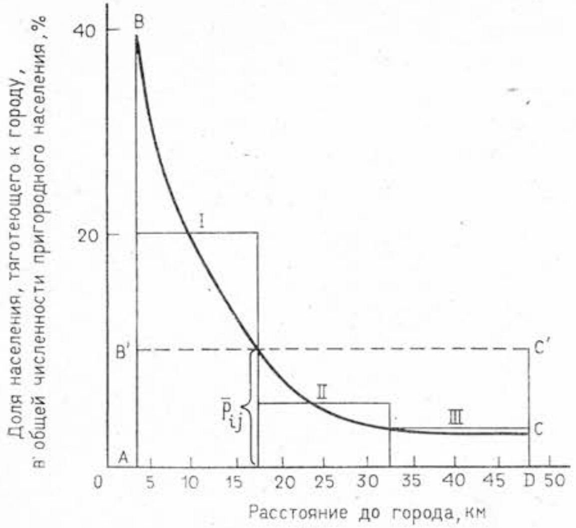

Рис. 5.2. Средняя вероятность поездки в город-центр (I, II, III — зоны притяжения города-центра)

где ΣH_i — общая численность населения, проживающего в районе влияния j-го города; $\overline{p}_{ij}$ — средняя вероятность поездки иногороднего населения в город-центр.

Величину $\overline{p}_{ij}$ можно представить графически (рис. 5.2). Из рис. 5.2 видно, что величина средней вероятности $\overline{p}_{ij}$ соответствует высоте прямоугольника $AB'C'D$, равновеликого по площади фигуре $ABCD$, ограниченной кривой гиперболической функцией, радиусом района обслуживания и соответствующими ординатами[1]. Значение p_{ij} определяется по формуле

$$\overline{p}_{ij}=\frac{S_{ABCD}}{R_{ab}},$$

где R_{ab} — среднее расстояние между границами обслуживания.

Построение вероятностной динамической модели расселения 32 000 жителей города, а также изучение подвижности населения на основе метода Монте-Карло произвела Г. В. Красоленко[2].

Пронумеровав промышленные предприятия ($i=1, 2, \ldots, m$) и жилые районы ($j=1, 2, \ldots, n$), автор вычислила вероятности нахождения жителей на данном предприятии и в данном жилом районе, т. е. p_{ij}. При этом разумеется, что $\sum\limits_{i}\sum\limits_{j} p_{ij}=1$.

Наблюдая и анализируя расселение населения по состоянию на фиксированные моменты времени, автор получила матрицу

[1] См.: *Тимчук Н. Ф.* Методы экономического обоснования развития городов и районов. Киев, 1974, с. 173 и далее.

[2] См.: *Красоленко Г. В.* Вероятностная динамическая модель расселения в городе. — В кн.: Статистика миграции населения. М., 1973, с. 74.

вероятностей состояний $|P_{ij}|$. Задача состояла в определении перспективного расселения через любой промежуток времени $[0, T]$. Кроме вероятностей для фиксированного начального момента $p_{ij}(0)$, были определены вероятности перехода из одного состояния в другое (E_y, b, E_{xb}), осуществляемого под воздействием различных градостроительных, социологических, демографических и экономических факторов, рассматриваемых автором как независимые, случайные[1].

При переходе из одного состояния в другое изменяется «сумма благ», которыми обладает житель. Производя соответствующие расчеты, автор полагала при этом, что такие изменения состояния, как повышение заработной платы, длительность проживания в данном районе и другие, увеличивают «сумму благ», а такие, как семейные обстоятельства (женитьба, рождение ребенка), длительность поездки, переезд в район со старым жилым фондом и другие, уменьшают «сумму благ».

Произведя расчеты по соответствующему вычислительному алгоритму, имитируя постепенный процесс расселения, автор после каждой реализации случайного процесса находила финальные вероятности и финальные переходные вероятности и определяла средние значения и дисперсии. Как утверждает Г. В. Красоленко, данная модель расселения позволила получить первое приближение научно обоснованного прогноза расселения на длительный период.

Моделирование, основанное на вероятностном принципе, используется во многих случаях расчета баланса трудовых ресурсов, для изучения распределения дошкольников, распределения оканчивающих школы и др.

Можно построить модель, характеризующую распределение *численности учащихся, окончивших 8-е классы общеобразовательной школы, по видам последующего обучения:* 9—10 классы общеобразовательной школы, профессионально-технические училища, средние специальные учебные заведения, направление на работу и продолжение обучения на производстве. Для нахождения распределения учащихся (y) по видам обучения (i) по состоянию на определенный момент t, т. е. $y^i(t)$, исходными показателями являются: $v^8(t)$ — численность учащихся, окончивших 8-е классы (может быть получена путем наблюдения, а также модельным путем); $p^i(t)$ — вероятность распределения учащихся, окончивших 8-е классы, по видам последующего обучения (практически эта вероятность может быть определена путем использования отчетных данных о величине приема в учебные заведения i-го вида). Получаем:

$$y^i(t)=v^8(t)\cdot p^i(t).$$

[1] Были привлечены следующие факторы: производственный стаж, возможность повышения заработной платы, возможность изменения численности семьи, затраты времени на передвижение от места работы до места жительства, переход в районы новостроек или районы со старым жилым фондом, длительность проживания в районе и т. д.

Разработана и экспериментально опробована матричная социально-демографическая модель для перспективных расчетов контингентов учащихся[1]. В модель включены следующие звенья системы образования и подготовки специалистов: 1) неполная средняя школа; 2) полная средняя школа; 3) профессионально-технические училища обычного типа; 4) средние профессионально-технические училища; 5) технические училища; 6) средние специальные учебные заведения; 7) высшие учебные заведения. Каждый человек в конкретный момент может находиться в состоянии обучения в одном из звеньев. Кроме того, в модель включены два состояния: отсев учащихся, занятые в народном хозяйстве в возрасте 15—30 лет.

За расчетную единицу времени принимается один год. В основе модели лежат вероятности перехода учащихся из одного состояния в другое (из i в j) в течение года t. Основное соотношение модели имеет вид:

$$S^t = S^{t-1}p^t + b^t,$$

где S^t — численность учащихся на конец года t; p^t — вероятность перехода из одного состояния в другое; b^t — численность детей, достигших семилетнего возраста на конец года t.

Матрица переходных вероятностей вида $P^t = p^t_{ij}$ позволяет рассчитывать численность приема учащихся для каждого года, выпуска из различных звеньев системы образования, потоки учащихся в самой системе образования на ее входе и выходе, соотношение, устанавливающее связь выпуска этого года с выпусками учащихся прошлых лет. Матричная модель предполагает возможность прогнозирования вероятностей перехода.

При необходимости можно построить модель *обучения в высшем учебном заведении* в целом и с подразделениями на факультеты. При этом исходными данными послужат численность набора, число лет обучения и вероятности перехода учащихся с курса на курс.

В. В. Никитенко, используя данные двух переписей населения СССР (1959 и 1970 гг.), придал численную определенность закономерности, замеченной социологами и состоящей в том, что чем выше на момент наблюдения образовательный уровень индивида, тем больше вероятности его перехода на следующие, более высокие ступени образования. Расчет вероятностей перехода производится для четырех поколений (1910—1919, 1920—1929, 1930—1939 и 1940—1949 гг.) как отношение числа лиц, выбывших из данной образовательной группы за межпереписной период, к суммарной численности лиц, находившихся в ней на начало периода и вошедших в нее на протяжении указанного периода из нижестоящих по образованию групп. Укажем на один важный вывод, полученный автором, который состоит в том, что «...современная молодежь не останавливается на достигнутом образовании, а стремится к дальнейшему пополнению знаний»[1].

Известны попытки построения моделей, позволяющих осуществлять *долгосрочное прогнозирование заболеваемости населения,* в частности гипертонической болезнью. Была построена модель, основанная на применении вероятностного метода Бейеса, для выявления зависимости распространения болезни от влияния комплекса таких демографических и социально-гигиенических факторов, как: пол, возраст, уровень образования, отношение к браку.

Модель имела вид:

$$\overline{y} = d_1p_1 + d_2p_2 + \ldots + d_np_n ,$$

где $\overline{y}$ — уровень распространенности гипертонической болезни;
$d_1, d_2, \ldots, d_n$ — доли соответствующих групп населения;
$p_1, p_2, \ldots, p_n$ — вероятности наличия больных среди указанных групп населения.

Американские демографы М. Шепс и Ж. Ридли при построении модели рождаемости привлекли данные о вступлении в первый и повторные браки, овдовении, а также такие элементы плодовитости, как зачатие, беременность, живорождение или смерть плода, выживание или смерть ребенка, стерильный период после родов и т. д. Авторы этой модели указывают, что если вероятности перехода женщин из одних состояний в другие определены правильно, то модель может довольно точно указать ожидаемый уровень рождаемости.

При решении задачи на *удовлетворение потребности в кадрах конкретной профессии* проф. А. Я. Боярский сделал несколько допущений[2]. Обозначив определенные возрастные группы 0 и 1 и приняв занятость населения в этих возрастах в данной профессии равной 10%, он получил

$$\Sigma s_x = 0{,}1\,(S_0 + S_1),$$

где s_x — повозрастная численность лиц данной профессии; S_0 и S_1 — общие численности выделенных возрастных групп.

Учитывая, что $\dfrac{s_x}{S_0} = p_0p_1 \ldots p_{x-1}h^{-x} = \alpha_x$; $\dfrac{s_x}{S_x} = \beta_x$ и h — коэффициент роста, автор получил после преобразований:

$$\alpha_{x-1}u_{x-1} + \alpha_{x-1}\cdot\beta_{x-1}(v_{x-1} - u_{x-1}) = h\alpha_x\beta_x ,$$

где v_x — переходная вероятность, т. е. вероятность для лица, принадлежащего выделенной группе, остаться в живых и через год оказаться в данной группе;
u_x — переходная вероятность, аналогичная v_x, но не для лица, принадлежащего выделенной группе, а для лица из остального населения.

[1] См.: *Ляляев В.* Модель демографических процессов развития образования. — В кн.: Образовательная и социально-профессиональная структура населения СССР. М., 1975, с. 73.

[1] *Никитенко В. В.* Демографический анализ поколений. М., 1979, с. 117.
[2] См.: *Боярский А. Я.* Население и методы его изучения. М., 1975, с. 110 и далее.

Устанавливается, что

$$a_x=\frac{\alpha_x u_x}{h}; \quad b_x=\frac{v_x-u_x}{h}; \quad \gamma_x=\alpha_x\cdot\beta_x.$$

Получается рекуррентная формула

$$\gamma_x=a_{x-1}+b_{x-1}\gamma_{x-1}.$$

Подставляя и преобразуя, устанавливаем, что

$$\Sigma\gamma_x=0{,}1\,(\alpha_0+\alpha_1).$$

На основании полученной формулы А. Я. Боярский решает конкретные задачи по определению долей выделенного субнаселения в каждой возрастной группе (прямая задача) и вероятностей пополнения подгрупп (обратная задача).

Рассмотрим его расчеты.

Таблица 5.2. **Расчет занятости населения**

x	p_x	u_x	v_x	α_x	a_x	b_x	γ_x	β_x
0	0,8	0,16	0,4	1	0,128	0	0	0
1	0,5	0,05	0,4	0,64	0,0256	0,28	0,128	0,20
2	0,25	0,025	0,1	0,256	0,00512	0,06	0,0614	0,24
3	0	0	0	0,0512	0	0	0,0088	0,172

Получаем $\Sigma\gamma_x=0{,}1\cdot(1+0{,}64)=0{,}164$ (приблизительно 16%).

Значения γ_x и зависящие от них искомые величины могут быть определены путем раскладки полученных 16% на три возрастные группы: 1, 2 и 3. А. Я. Боярский вводит 4 варианта раскладки, в сумме дающие 16%: 1) 4, 4, 8; 2) 4, 8, 4; 3) 8, 4, 4; 4) 2, 13, 1.

Делением γ_x на α_x находим β_x. Задача теперь состоит в том, чтобы определить вероятности пополнения состава групп, обеспечивающие достижение заданных величин.

Получаем

$$u_x=\frac{p_x\beta_{x+1}-\beta_x v_x}{1-\beta_x}.$$

Значения вычисленных по этой формуле u_x для четырех вариантов раскладки $\Sigma\gamma_x$ между тремя возрастными группами приведены в табл. 5.3.

Здесь при расчетах для каждого варианта добавлена графа значений w_x — фактических вероятностей умереть ($w_x=u_x:p_x$).

Анализ показывает, что не все указанные в таблице варианты распределения γ_x между возрастными группами могут быть реализованы. Так, первый вариант не реализуется из-за превышения возможностей при $x=3$, а последний — из-за переполнения при $x=2$.

Хотя современные электронно-вычислительные машины легко справляются с такими задачами, автор считает, что из-за несоот-

Таблица 5.3. **Варианты раскладки удовлетворения потребностей в кадрах**

Варианты	x	p_x	α_x	v_x	γ_x	β_x	u_x	w_x
Первый	0	0,80	1,0000	—	—	—	0,0500	0,0620
	1	0,50	0,6400	0,4	0,04	0,0625	0,0566	0,1130
	2	0,25	0,2560	0,1	0,04	0,1562	0,4444	1,7770
	3	0,00	0,0512	0,0	0,08	1,5625	—	—
Второй	0	0,80	1,0000	—	—	—	—	0,0620
	1	0,50	0,6400	0,4	0,04	0,0625	0,0500	0,2800
	2	0,25	0,2560	0,1	0,08	0,3125	0,1400	0,9545
	3	0,00	0,0512	0,0	0,04	0,7812	0,2386	—
Третий	0	0,80	1,000	—	—	—	0,1000	0,1250
	1	0,50	0,6400	0,4	0,08	0,1250	0,0321	0,0642
	2	0,25	0,2560	0,1	0,04	0,1562	0,2129	0,8516
	3	0,00	0,0512	0,0	0,04	0,7812	—	—
Четвертый	0	0,80	1,0000	—	—	—	0,2500	0,3100
	1	0,50	0,6400	0,4	0,02	0,0312	0,2423	0,4986
	2	0,25	0,2560	0,1	0,13	0,5080	0,0042	—
	3	0,00	0,0512	0,0	0,01	0,1950	—	—

ветствия фактического населения стабильному такие расчеты могут служить лишь ориентиром, который нуждается в дальнейших коррективах.

В вероятностном аспекте можно записать и модель воспроизводства населения. Рассмотрим такую модель, предложенную Г. А. Павловым[1]. Представим общую численность женщин плодовитого возраста S_{15-49} в закрытом населении (отсутствует миграция) в зависимости от возраста (x), семейного состояния (C — никогда не состоявшие в браке, B — состоящие в браке, W — вдовы, D — разведенные) и числа рожденных детей — n:

$$S_{15-49}=\sum_{x=15}^{49}\sum_{n=0}^{10+}C_{x,n}+\sum_{x=15}^{49}\sum_{n=0}^{10+}B_{x,n}+\sum_{x=15}^{49}\sum_{n=0}^{10+}W_{x,n}+$$

$$+\sum_{x=15}^{49}\sum_{n=0}^{10+}D_{x,n}.$$

Каждая женщина характеризуется некоторым набором вероятностей перехода в течение года из одного состояния в другое, а также вероятностью выбытия из-под наблюдения по достижении 50-летнего возраста или смерти.

Примем во внимание вероятность для женщины, состоящей в браке (B) в возрасте x лет и родившей к этому возрасту n детей, дожить до возраста $x+1$, оставаясь в браке, — P_x и вероятность выйти из этого состояния (овдоветь или развестись) — $Q_x=1-P_x$.

[1] *См.: Павлов Г. А.* Информация о естественном движении населения и перспективные расчеты. — В кн.: Вопросы демографии. М., 1970, с. 72—74.

Для получения искомой вероятности следует учесть, что изменение значения каждого признака при условии неизменности состояния по другим признакам является независимым событием. Тогда, обозначая вероятности умереть — q_x, развестись — d_x, овдоветь — w_x и родить $(n+1)$-го ребенка — $A_{x,n/n+1}$, применим теорему умножения вероятностей и получим

$$P_x=(1-q_x)(1-d_x)(1-W_x)(1-f_{x,\,n/n+1}).$$

Исходя из принципа ординарности, т. е. практической невозможности появления двух или нескольких событий за достаточно малый интервал времени, можно представить изменение общей численности женщин плодовитого возраста от момента t_1 к моменту t_2 в виде $t_2-t_1=\Delta t$, т. е. совершить переход от $S^{t_1}_{15-49}$ к $S^{t_2}_{15-49}$.

Для каждого возрастного интервала нужно сначала определить изменения по отдельным категориям семейного состояния, а затем, объединив все в одну формулу, судить об изменении общей численности каждого возрастного интервала. В результате Г. А. Павлов получил следующее выражение:

$$\begin{aligned}S_{x+\Delta t,n}=S_{x,n}&-(C_{x,n}q^{C}_{x,n}+B_{x,n}q^{B}_{x,n}+W_{x,n}q^{W}_{x,n}+\\&+D_{x,n}q^{D}_{x,n})+(C_{x,n-1}\cdot f^{C}_{x,n-1/n}+B_{x,n-1/n}\cdot f^{B}_{x,n-1/n}+\\&+W_{x,n-1}f^{W}_{x,n-1/n}+D_{x,n-1}f^{D}_{x,n-1/n})-(C_{x,n}f^{C}_{x,n/n+1}+\\&+B_{x,n}f^{B}_{x,n/n+1}+W_{x,n}f^{W}_{x,n/n+1}+D_{x,n}f^{D}_{x,n/n+1}\,.\end{aligned}$$

В демографической литературе при попытках прогнозирования распределения семей по их величине встречаются различные методы построения модели распределения семей по числу членов.

В тех случаях, когда распределение семей по величине почти однозначно определяется средней величиной семьи и имеет незначительную асимметрию, это распределение семей подчиняется закону Пуассона. При этом, разумеется, учитывается, что число членов семьи не может быть равно 0, а начинается с единицы. Распределение Пуассона принимает вид:

$$P_x=\frac{e^{-\lambda'}(\lambda')^{x-1}}{(x-1)!},\quad x=1,2,3,\ldots,$$

где λ' — средняя преобразованного распределения Пуассона на 1 меньше средней исходного распределения.

В этих случаях расчет производится на основе экстраполяции средней величины семьи на прогнозируемый период.

Теоретически же обоснованной моделью структуры семьи является модель А. Я. Боярского, предположившего при ее построении, что общее число членов семьи состоит из суммы четырех групп: $D+B+V+S$, где D — дети добрачного возраста; B — лица, состоящие в браке, где оба супруга в плодовитом возрасте; V — другие взрослые; S — старики.

Число основных демографических событий в каждой группе должно исчисляться по данным о повозрастной интенсивности этих событий и структуре стабильного населения. Для каждой из групп устанавливается показатель смертности, для D, B и V — вероятности перехода в следующую возрастную группу и для B — показатель рождаемости.

Частота образования новых семей определяется следующим образом: в 20% случаев в семью вступает новый член, а в 80% случаев образуется новая семья. Автор метода свел общее число переходов из состояния D, B, V, S в другие состояния к семи. Решение системы из 6 уравнений и использование матрицы переходных вероятностей для семей различной величины, а также предельное распределение семей по величине позволяют представить процесс развития семьи как случайной марковский процесс.

Э. К. Васильева использовала вероятностный подход для оценки особенностей демографического состава семей в различных сельских местностях Удмуртской АССР под влиянием двух факторов: принадлежности населенных пунктов к числу колхозных или неколхозных и национального состава жителей. Определялась вероятность существенности расхождений показателей, относящихся к различным типам семей, по формуле

$$t=\frac{x_1-x_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\,.$$

Один из выводов состоял, в частности, в том, что «...демографический состав удмурдских семей зависит в большей степени от принадлежности населенных пунктов к числу колхозных или неколхозных, чем от национального состава их жителей»[1].

В настоящее время назрела необходимость большего внимания к определению вероятностей перехода семей из одного типа в другие. Э. К. Васильева приводит модель развития семьи с указанием вероятности ее перехода из одной стадии в другую в связи с различными демографическими событиями (брак, рождение, смерть, прибытие или выбытие членов семьи, расторжение брака и т. д.). Такое построение позволило вывести закономерности, имеющие не только теоретическое, но и большое практическое значение.

Сейчас исследователи нашли способы перспективных расчетов, в которых наряду с использованием основных показателей режимов воспроизводства привлекаются также вероятности изменения численности различных групп семей. Это обстоятельство вселяет уверенность в повышении точности перспективных расчетов.

---

[1] *Васильева Э. К.* Социально-экономическая структура населения СССР, с. 61—63.

13 Заказ 2715

Вопрос об уровне, смертности и средней продолжительности жизни в ближайшем будущем с вероятностных позиций рассматривали французские демографы А. Низар и Ж. Валлен, которые определили биологический предел снижения смертности путем использования наименьших повозрастных вероятностей умереть, зафиксированных в развитых капиталистических странах. В соответствии с этими расчетами средняя продолжительность жизни оказалась равной для мужчин 72,2 года и для женщин 76,8 года. Аналогичные расчеты производились Ж. Буржуа-Пиша, который, пытаясь выделить эндогенную смертность, обусловленную старением организма, и отделить ее от экзогенной, вызванной внешними причинами, не связанными с развитием человеческого организма, — голод, инфекционные заболевания, насильственная смерть и т. д. — определил нижний биологический предел снижения смерти в каждом возрастном интервале и получил следующие оценки средней продолжительности жизни: для мужчин — 76,4 года, а для женщин — 78,7 года.

Интересно, что наименьшие фактические вероятности умереть во всех без исключения возрастных интервалах, которые получили А. Низар и Ж. Валлен, были выше предельных вероятностей Ж. Буржуа-Пиша. Отношение указанных вероятностей колебались в границах от 22 (для возрастов 1—4 года) до 1 (для возрастов 90—94 года).

В настоящее время разработана методика прогнозирования средней продолжительности жизни и уровня смертности на далекую перспективу. Она предполагает построение «идеальных» таблиц смертности, в которых используются самые низкие для каждого возраста коэффициенты смертности отдельных территорий. На основании ожидаемой динамики вероятностей дожития, использованных при передвижке возрастов, можно сделать различные предположения и выводы.

5.8. Систематические ошибки моделирования при использовании вероятностей дожития или смерти

В практических перспективных расчетах очень часто используется извлеченный из таблиц смертности показатель вероятности умереть — q_x или дожить — p_x. Покажем, что при этом допускается систематическая ошибка, ибо вероятность, например, q_x для этой цели не совсем пригодна и нужно вводить в расчет поправку на возрастную смертность.

В чем состоит дефект в построении q_x, приводящий к неточностям прогнозирования? Дело в том, что исчисление q_x производится по формуле $q_x=d_x:l_x$, т. е. ориентируется на совокупность людей, находящихся в точном возрасте x лет. Перепись населения этих чисел не дает, так как учитываемые ею лица имеют на дату переписи возраст от x до $x+1$ года. Если речь идет, например, о 30-летних, то получаем на дату переписи лиц в возрасте от 30 до 31 года, т. е. в возрасте 30 лет и несколько месяцев, не-

сколько недель или дней. Значит, перепись дает не совокупность доживающих l_x, а совокупность живущих L_x. При этом следует заметить, что $L_x<l_x$, ибо часть доживающих до возраста x лет к дате переписи умрет до перехода к возрасту $x+1$ год. Поэтому только в том случае, если мы умножим q_x на число доживающих до точного возраста x лет, мы получим число доживающих до точного возраста $x+1$ год.

Обозначим Q_{30} среднюю вероятность умереть для лиц, живущих в возрасте 30 лет (L_{30}) на протяжении года до перехода к 31-летнему возрасту (L_{31}). Q_x строится на основании таблиц смертности. Используя формулу вероятности остаться в живых для лиц, живущих в возрасте x лет, на протяжении года жизни, получим $P_x=L_{x+1}:L_x$. Тогда $Q_x=1-P_x=1-L_{x+1}:L_x$.

Конечно, Q_x не совпадает с q_x, потому что при исчислении, например, Q_{30} часть умирает в возрасте 30 лет, а часть — в возрасте 31 год, а при исчислении q_{30} учитываются все умирающие в возрасте 30 лет.

В тех возрастах, где смертность сравнительно невелика (возраст 5—35 лет), разность между Q_x и q_x мала. Используя для примера таблицу смертности населения СССР 1958—1959 гг., получаем для 30-летних:

$$P_{30}=\frac{L_{31}}{L_{30}}=\frac{90\,758}{90\,981}=0{,}99755;\ Q_{30}=1-P_{30}=0{,}00245,$$

$$\text{а } q_{30}=0{,}00240.$$

Разница, как мы видим, невелика, а вот для возрастов 0 лет, 1 года или 70 и 80 лет она уже значительна:

$$P_0=\frac{95\,390}{97\,272}=0{,}98065;\ Q_0=0{,}01935, \text{ а } q_0=0{,}04060;$$

$$P_1=\frac{94\,887}{95\,390}=0{,}99\,473;\ Q_1=0{,}00527, \text{ а } q_1=0{,}00840;$$

$$P_{70}=\frac{58\,611}{60\,738}=0{,}96498;\ Q_{70}=0{,}03502, \text{ а } q_{70}=0{,}03350;$$

$$P_{80}=\frac{32\,076}{35\,013}=0{,}91612;\ Q_{80}=0{,}08388, \text{ а } q_{80}=0{,}08501;$$

Систематизируем полученные результаты:

Возраст, лет	q_x	Q_x
0	0,04060	0,01935
1	0,00840	0,00527
30	0,00240	0,00245
70	0,03350	0,03502
80	0,08051	0,08388

Их анализ позволяет сделать вывод, что для молодых возрастов систематическая ошибка получается вследствие того, что $q_x>Q_x$, а для старых возрастов $q_x<Q_x$. Следовательно, если перспективные расчеты производятся на 15—20 лет вперед, то использование q_x вместо Q_x приведет к преуменьшению детских и

завышению старческих возрастов, т. е. исказит картину возрастного состава населения. Поэтому поправка на смертность должна производиться с использованием Q_x. Практически в расчетах удобнее пользоваться не Q_x, а P_x.

Что касается расчетов L_0, то для перехода от числа родившихся к числу учитываемых переписью в возрасте 0 лет (живущих в возрасте от 0 до одного года) нужно число родившихся умножить на отношение $\frac{L_0}{l_0}$, т. е. отношение числа живущих в возрасте 0 лет по таблице смертности к числу родившихся по этой таблице, т. е. 10 000 или 100 000. Для возрастов 75 лет и старше лучше всего поправку на смертность исчислять суммарно для всех последующих возрастов.

Несовпадение результатов моделирования с фактическими результатами можно объяснить недостаточной правильностью сделанных допущений. При попытках выявить границы несовпадения данных, т. е. ошибок, полученных в результате моделирования, в ряде случаев привлекают теорию вероятностей.

Очень резко выступил против этого советский статистик-демограф Б. Ц. Урланис, считая такие действия «лишенными оснований»[1]. Однако при такой критике не принималось во внимание, что в подавляющем большинстве случаев моделирование демографических процессов проводится на основе не сплошных, а выборочных данных. Вот почему при распространении полученных результатов на генеральную совокупность без теории вероятностей обойтись нельзя.

Вместе с тем опасения Б. Ц. Урланиса имели основание. Действительно, демографический расчет, основывающийся на привлечении вероятностных схем, может быть далек от совершенства, недостаточно достоверен, может обладать большой погрешностью и т. д. Но вина в этом не только вероятностного аппарата — схем теории вероятностей. По нашему мнению, такое расхождение расчетов с действительностью не должно служить причиной отказа от привлечения теории вероятностей в демографических расчетах, но оно требует принятия необходимых мер для повышения «степени правильности» расчетов, уменьшения ошибок. Для этого надо более аргументированно и научно отбирать факторы, воздействующие на величину изучаемых явлений, более тонко анализировать поведение и жизненный уровень людей, учитывать эффективность мероприятий, направленных на увеличение продолжительности жизни населения, и т. д. Надо, наконец, по ограниченному числу изучаемого населения делать научные выводы о всем населении. В конечном счете это означает необходимость привлечения к изучению демографических процессов не только теории вероятностей, но и мощного статистического, математико-статистического аппарата с факторным, корреляционным анализом, с проверкой статистических гипотез, ориентированных на всемерное расширение применения выборочного метода, и т. д.

[1] См.: Марксистско-ленинская теория народонаселения, с. 221.

Глава 6

Статистическая проверка демографических гипотез

6.1. Нулевая гипотеза и критерии согласия

Под гипотезой в демографии часто понимают предположение о виде распределения, однородности совокупности, существенности различия нескольких средних, наличии зависимости между явлениями, принадлежности выборочных данных к одной генеральной совокупности и т. д. Проверка гипотез производится на основе установленного уровня значимости — малой вероятности (0,05 и 0,01) и критической области, попадание в которую имеет вероятность, равную уровню значимости.

Нулевая гипотеза используется при статистической проверке гипотез. Суть ее состоит в признании того, что фактическое распределение укладывается в теоретическое, совокупность однородна, частость близка к вероятности, различия между средними несущественны, зависимость между признаками отсутствует, выборочные данные принадлежат к одной генеральной совокупности и т. д.

Проверка нулевой гипотезы производится путем использования различных критериев согласия, позволяющих с помощью доверительных вероятностей сделать вывод о ее опровержении. При этом следует иметь в виду, что неопровержение нулевой гипотезы не означает ее подтверждения, а свидетельствует о возможности существования других гипотез, которые также окажутся совместимыми с опытными данными, и о необходимости проведения дальнейшей проверки, в частности путем увеличения числа наблюдений. При проверке нулевой гипотезы наибольшее значение придается практической неосуществимости маловероятных событий. Так, если вероятность критерия согласия, выражающего вероятность случайного расхождения, очень мала ($<0{,}05$), то это свидетельствует о существенном различии, и нулевая гипотеза опровергается; если же она достаточно велика ($>0{,}05$), то вопрос о существенности расхождений остается без ответа.

Критерии согласия являются объективными оценками, например, близости фактических распределений к теоретическим. Они позволяют ответить на вопрос, вызвано ли расхождение фактического распределения с теоретическим случайными причинами, связанными с недостаточным числом наблюдений, или существенными, т. е. тем, что теоретическое распределение плохо воспроизводит фактическое.

Критерий согласия выступает обычно в виде некоторой величины, оцениваемой с определенной вероятностью. Применяются следующие критерии согласия: А. Н. Колмогорова (λ), К. Пирсона (χ^2), В. И. Романовского, Б. С. Ястремского (l).

Статистическая проверка гипотез состоит в следующем: устанавливается уровень значимости, затем определяется критическая область проверяемой гипотезы — такая область, попадание в которую имеет вероятность, равную значимости. Чаще всего эта область охватывает все значения критерия, превосходящие некоторые критические значения.

Если фактическая величина критерия оказывается в критической области, то гипотеза опровергается. В противном случае считается, что данные результаты наблюдений ей не противоречат, т. е. гипотеза приемлема, что ни в коем случае не следует смешивать с ее доказанностью или даже подтвержденностью. Смысл опровержения гипотезы в случае попадания критерия в критическую область состоит в том, что приходится либо отбросить гипотезу, либо признать, что произошло весьма маловероятное событие. Выбирая первое, мы, однако, рискуем отбросить как опровергнутую в действительности верную гипотезу: ведь маловероятные события все же возможны. Этот риск тем меньше, чем меньше уровень значимости, а последний тем меньше, чем уже критическая область, поэтому нередко вместо фиксации критического значения, отвечающего заранее назначенной малой вероятности, определяется вероятность того, что в условиях справедливости проверяемой гипотезы критерий достигнет своей фактической величины или превысит ее.

Критерий согласия К. Пирсона (хи-квадрат) чаще всего используется для проверки согласования фактических данных с гипотезой о наличии некоторого теоретического закона распределения, т. е. для оценки приближения фактического распределения к теоретическому:

$$\chi^2=\sum_{i=1}^{k}\frac{(m-m')^2}{m'},$$

где m — фактические частоты; m' — теоретические частоты, вычисленные на основании проверяемого закона распределения.

По соответствующим математико-статистическим таблицам находят при данном числе степеней свободы вероятность достижения χ^2 данного значения, т. е. $P(\chi^2)$. При вероятностях, значительно отличающихся от нуля, расхождение между теоретическими и эмпирическими частотами считают случайным, а гипотеза признается не противоречащей опыту.

При использовании критерия χ^2 необходимо, чтобы объем ряда распределения и частоты интервалов были достаточно велики ($n>50$ и $m_i>5$). В случае необходимости соединяют два или несколько соседних интервалов. Хи-квадрат может также служить для проверки принадлежности нескольких выборочных данных одной и той же генеральной совокупности.

В демографии критерий хи-квадрат может быть использован для изучения изменения структуры (возрастной, профессиональной, социальной и др.) населения, выявления равномерности распределения по территориям, изучения распределения населения по различным признакам, выявления наличия связей различных факторов, оценки существенности различных показателей и т. д.

Критерий согласия А. Н. Колмогорова (λ) оценивает близость фактического распределения к теоретическому путем нахождения величины D — максимальной разности накопленных частостей фактического и теоретического распределений.

В случае привлечения не частостей, а частот, еще не переведенных в частости, практическое использование данного критерия состоит в следующем.

1. Исчисляют величину

$$\lambda=\frac{D\sqrt{n}}{n}=\frac{D}{\sqrt{n}}.$$

2. Определяют вероятность того, что если конкретный демографический признак распределен по рассматриваемому теоретическому распределению, то из-за чисто случайных причин максимальное расхождение между фактическими и теоретическими накопленными частотами будет не меньше, чем фактически наблюденное.

3. Величина $P(\lambda)$ позволяет сделать определенные выводы:

при достаточно большой вероятности $P(\lambda)$ гипотезу о том, что фактическое распределение близко воспроизводит теоретическое, можно считать неопровергнутой, т. е. правдоподобной;

если же вероятность $P(\lambda)$ мала, то гипотеза считается неправдоподобной и отвергается; расхождения между фактическим и теоретическим расхождениями признают существенными.

Критерий согласия А. Н. Колмогорова можно применить и для решения вопроса о принадлежности двух выборок объемом n_1 и n_2 к одной генеральной совокупности, а также для проверки независимости двух или нескольких признаков совокупности.

Критерий согласия В. И. Романовского используется для оценки приближения фактического распределения к теоретическому. Определяется величина χ^2. Затем, привлекая число степеней свободы (K), равное числу групп фактического распределения без числа параметров, найденных по фактическому распределению, вместе с числом дополнительных соотношений, которым подчинены фактические данные, используют выражение

$$\frac{\chi^2-K}{\sqrt{2K}}.$$

Если при испытании близости фактического распределения к теоретическому величина этого выражения меньше трех, то это дает основание для утверждения о возможности принятия теоретического распределения за закон данного распределения. Можно для этой цели привлекать неравенство

$$\sqrt{2\chi^2}-\sqrt{2K-1}<2\,.$$

Критерий согласия Б. С. Ястремского (l). В 1948 г. Б. С. Ястремский показал, что недостатком всех указанных выше критериев согласия является то, что они не дают прямого ответа на вопрос о применимости к данному фактическому распределению предполагаемого распределения, а указывают лишь на вероятностную оценку расхождения между ними. Указанный недостаток устраняется критерием l, используемым для прямого ответа на вопрос о мере расхождения между эмпирическим и теоретическим распределением.

В критерии Б. С. Ястремского, так же как и при исчислении критерия В. И. Романовского, используется критерий К. Пирсона (хи-квадрат). В общем виде критерий Б. С. Ястремского может быть записан следующим образом:

$$l=\frac{\chi^2-n}{\sqrt{2n-4Q}}\,,$$

где n — число групп; Q — величина, зависящая от числа групп, но при числе групп, меньшем 20, принимается равной 0,6.

Так как величина l распределена нормально, то с вероятностью 0,997 по абсолютной величине l не должна превышать трех. Значит при $|l|\leqslant 3$ эмпирическое распределение не укладывается в теоретическое распределение.

В качестве элементарного приема определения близости фактического распределения к нормальному иногда прибегают к числам 0,3; 0,7; 1,1; 3,0, называемым числами Вестергарда. Данное фактическое распределение близко к нормальному, если соблюдены условия, изложенные в табл. 6.1.

Таблица 6.1. **Условия близости фактического распределения к нормальному**

Числа Вестергарда	Интервалы	Объем совокупности
0,3	от $\bar{x}-0{,}3\,\sigma$ до $\bar{x}+0{,}3\,\sigma$	0,25
0,7	от $\bar{x}-0{,}7\,\sigma$ до $\bar{x}+0{,}7\,\sigma$	0,50
1,1	от $\bar{x}-1{,}1\,\sigma$ до $\bar{x}+1{,}1\,\sigma$	0,75
3,0	от $\bar{x}-3{,}0\,\sigma$ до $\bar{x}+3{,}0\,\sigma$	0,99

Для проверки соответствия того или иного распределения закону нормального распределения может быть также использована «вероятностная бумага» (вариационная сетка, или сетка В. М. Турбина). Для этой цели строится диаграмма, у которой

по оси абсцисс нанесена равномерная шкала, на которую откладывают верхние границы фактического распределения, а по оси ординат — неравномерная, логарифмическая шкала (построенная применительно к закону нормального распределения), на которую наносятся кумулятивные частости. Так как кумулята частостей в процентах нормального распределения асимптотична 0 и 100, то вертикальная шкала не может достичь ни нуля, ни ста. Нижняя отметка начинается с величины 0,1 и заканчивается величиной, близкой к 100, т. е. 99,9. Используя данные о распределении 500 взрослых мужчин по росту (см. табл. 2.3) и результаты построения нормальной кривой, наносим на вероятностную бумагу фактические накопленные частости. В результате получаем рис. 6.1, из которого видно, что полученная ломаная линия может быть аппроксимирована прямой линией. Это является свидетельством правдоподобия принятой гипотезы о нормальности данного распределения. Степень приближения к прямой линии и характер отклонений дают возможность оценивать фактическое распределение.

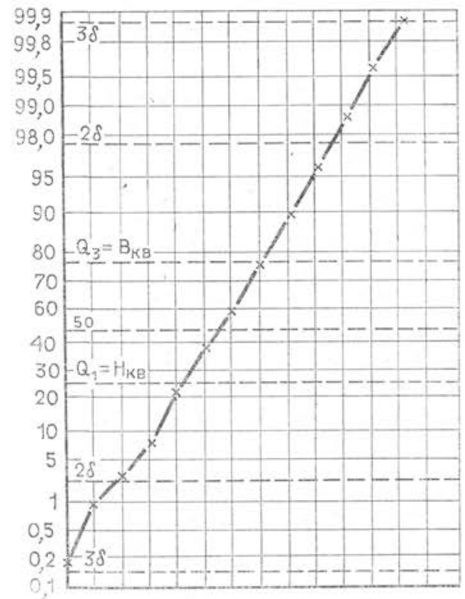

Рис. 6.1. Вариационная сетка В. М. Турбина — «вероятностная бумага»

6.2. Конкретное использование критериев согласия А. Н. Колмогорова, В. И. Романовского и Б. С. Ястремского

Имеются данные выборочного наблюдения бюджетов 400 семей рабочих, сгруппированных по уровню среднедушевого дохода в 8 групп (табл. 6.2).

Значения m — это фактические частоты распределения, а m' — частоты, рассчитанные в предположении о распределении по логарифмически-нормальному закону (рис. 6.2). Задача состоит в выяснении того, насколько логарифмически-нормальное распределение соответствует фактическому распределению.

Покажем использование для этой цели критерия λ А. Н. Колмогорова

$$\lambda=\frac{|S-S'_{\max}|}{\sqrt{n}}=\frac{D_{\max}}{\sqrt{n}}$$

Таблица 6.2. **Среднедушевой доход рабочих**

Группы семей по уровню среднедушевого дохода	Число семей (частоты), m	Частоты логарифмически-нормального распределения, m'	$S=\Sigma m'$	$S'=\Sigma m'$	$D=\lvert S-S'\rvert$	$(m-m')^2$	$\frac{(m-m')^2}{m'}$
1	5	4	5	4	1	1	0,25
2	5	5	10	9	1	0	0
3	37	31	47	40	7	36	1,16
4	62	62	109	102	7	0	0
5	116	118	225	220	5	4	0,03
6	119	131	344	351	7	144	1,10
7	44	40	388	391	3	16	0,40
8	12	9	400	400	0	9	1,00
Итого	400	400	—	—	—	—	3,94

где S и S' — накопленные частоты фактического и логарифмически-нормального распределения.

В нашем примере $D_{\max}=7$, $\sqrt{n}=20$.

Тогда $\lambda=\frac{7}{20}=0,35.$

Используя таблицу вероятностей того, что λ достигнет некоторой величины[1], и получив $P(\lambda)=0,9997$ (т. е. значительную величину $>0,05$), делаем вывод о том, что расхождение между частотами фактического и логарифмически-нормального распределения не является существенным, т. е. логарифмически-нормальное распределение удовлетворительно воспроизводит фактическое распределение.

Рис. 6.2. Распределение семей рабочих по среднедушевому доходу:

1 — фактическое распределение; *2* — логарифмически-нормальное распределение

[1] См.: *Венецкий И. Г., Венецкая В. И.* Основные математико-статистические понятия и формулы в экономическом анализе. 2-е изд. М., 1979, с. 374.

Можно также для этой цели использовать критерий В. И. Романовского. Расчет критерия К. Пирсона χ^2 (см. значения $(m-m')^2 : m'$ в табл. 6.2) дал величину, равную 3,94, $K=5$. Тогда критерий В. И. Романовского равен:

$$\frac{|3,94-5|}{\sqrt{10}}=\frac{1,06}{3,16}\approx 0,34.$$

Эта величина значительно меньше трех. Таким образом, делаем вывод, что логарифмически-нормальное распределение отражает существенные черты фактического распределения.

Критерий Б. С. Ястремского имеет такое же толкование, что и критерий В. И. Романовского, и равен:

$$l=\frac{|3,94-8|}{\sqrt{16+4\cdot 0,6}}=\frac{4,06}{\sqrt{18,4}}\approx 0,95.$$

Значение критерия меньше трех. Это свидетельствует о возможности принятия для фактических данных теоретической модели распределения по логарифмически-нормальному закону. Вывод тот же.

6.3. Использование критерия К. Пирсона хи-квадрат для установления равновероятности отбора различных групп населения и др.

Производится выборочное изучение мнения четырех групп населения об идеальном числе детей в семьях.

Отбор 1000 семей из различных групп населения

Группы населения	Первая	Вторая	Третья	Четвертая	Всего
Число отобранных семей	262	238	239	261	1000

Можно ли на основании указанного числа отобранных семей принять гипотезу о равновероятности отбора всех четырех групп населения? Для ответа на поставленный вопрос привлечем критерий хи-квадрат:

$$\chi^2_{\text{факт}}=\sum\frac{(m-m')^2}{m'},$$

где m — фактические числа отобранных семей из групп; m' — числа отобранных семей в предположении равновероятности попадания в выборку из каждой группы, т. е. одного и того же числа, равного $1000:4=250$ семей.

Используя соответствующие таблицы по фактической величине $\chi^2_{\text{факт}}$ и числу степеней свободы данного распределения, определим вероятность того, что $\chi^2_{\text{факт}}$ превзойдет теоретическое значение — $\chi^2_{\text{теор}}$, т. е. $P(\chi^2)$.

Находим $\chi^2_{факт}$ по нашим данным:

$$\chi^2_{факт} = \frac{(262-250)^2}{250} + \frac{(238-250)^2}{250} + \frac{(239-250)^2}{250} + \frac{(261-250)^2}{250} =$$

$$= \frac{144+144+121+121}{250} = \frac{530}{250} = 2,12 \approx 2.$$

Число степеней свободы равно $K=4-1=3$. По таблице $P(\chi^2)$ находим $P(2)$ при $K=3$ и получаем вероятность 0,57. Это не отвергает нашу гипотезу о равновероятности отбора семей из групп.

Критерий хи-квадрат можно привлечь для оценки существенности различия показателей смертности населения различных регионов, людей различных профессий и т. д. Рассмотрим последовательность действий.

1. Вычисляются групповые значения хи-квадрат (с одной степенью свободы):

$$\chi_i^2 = \frac{n_i n'_i (n_i + n'_i)}{(d_i + d'_i)(n_i + n'_i - d_i - d'_i)} \cdot \left(\frac{d_i}{n_i} - \frac{d'_i}{n'_i}\right)^2,$$

где n_i и n_i' — численности двух групп; d_i и d_i' — численности изучаемых явлений по группам.

2. Определяется сумма всех групповых значений критерия хи-квадрат:

$$\chi^2 = \Sigma \chi_i^2 .$$

3. По соответствующим таблицам находится вероятность полученного значения χ^2, которая служит основанием для вывода о существенности различия.

Таблица 6.3. Смертность среди больных, страдающих туберкулезом легких в открытой форме, в течение первого года после заболевания

Возрастные группы, лет	Мужчины			Женщины			χ_i^2
	заболели n_i	умерли d_i	смертность, %	заболели n'_i	умерли d'_i	смертность, %	
15—20	406	156	38,4	500	174	34,8	1,25
20—25	695	204	29,4	816	246	30,0	0,11
25—30	585	169	28,9	619	184	29,7	0,09
30—35	454	128	28,2	433	150	34,6	4,22
35—40	274	82	29,9	257	92	35,8	2,10
40—45	221	68	30,8	194	83	42,8	6,43
45—50	153	41	26,8	94	39	41,5	5,75
50—55	110	34	30,9	58	20	34,5	0,23
55—60	69	36	52,2	29	13	44,8	0,45
60 и старше	89	43	48,3	47	28	59,6	1,57
Всего	3 056	961	—	3 047	1 029	—	22,20

В качестве примера используем условные данные Г. Крамера о смертности от туберкулеза мужчин и женщин в течение первого года после установления заболевания (табл. 6.3)[1].

Получаем значения критерия хи-квадрат для первой группы ($i=1$):

$$\chi_1^2 = \frac{406 \cdot 500 \cdot (406+500)}{(156+174) \cdot (406+500-156-174)} \cdot \left(\frac{156}{406} - \frac{174}{500}\right)^2 = 1,25.$$

Для второй группы ($i=2$): $\chi_2^2=0,11$ и т. д. Сумма составит: $\Sigma\chi_i^2=22,2$.

По математико-статистической таблице находим, что значению критерия $\chi^2=22,2$ и числу степеней свободы $K=10$ соответствует вероятность P (22,2), равная 0,014.

Таким образом, делаем вывод о существенности отклонений смертности у мужчин и женщин: смертность среди женщин значительно выше, чем среди мужчин. При этом видно, что наибольшей смертности подвержены возраста 30—50 лет.

Укажем еще на одно применение критерия χ^2. Отделом демографии НИИ ЦСУ СССР было произведено графическое выравнивание вероятностей вступления в брак для возрастов старше 29 лет. Достаточность аппроксимации проверялась критерием χ^2.

6.4. Проверка гипотезы о равенстве частостей рождения мальчиков

Произведем проверку гипотезы о равенстве частостей рождения мальчиков в семьях с 8 детьми, сгруппированных по числу сыновей. Используем статистические данные из работы академика С. Н. Бернштейна[2].

Обследовано 53 680 семейств, имевших по 8 детей с общим числом детей 429 440, среди которых мальчиков было 219 823. Таким образом, частость рождения мальчика составляла

$$p = \frac{219\,823}{429\,440} \approx 0,512.$$

Распределение семей по числу сыновей дано в табл. 6.4.

Для проверки данной гипотезы автором был привлечен эмпирический коэффициент дисперсии Маркова:

$$D = \sum \frac{(m_i - np)^2 S_i}{Npq},$$

где S_i — число обследований в каждой группе: $\Sigma S_i = S$ — общее число обследованных; n — число опытов в каждой группе; m_i — число появлений изучаемого события в каждой группе; $N=Sn$.

Теоретическое значение этого коэффициента равно единице.

[1] См.: *Крамер Г.* Математические методы статистики. М., 1975, с. 486—487.
[2] *Бернштейн С. Н.* Теория вероятностей. 4-е изд. М., 1946, с. 230.

Таблица 6.4. **Распределение семей по числу сыновей**

Число сыновей в семьях с 8 детьми, m_i	Число семей, S_i	Расчетные графы		
		m_i-np	$(m_i-np)^2$	$(m_i-np)^2 S_i$
0	215	—4,096	16,777	3607,055
1	1 485	—3,096	9,585	14233,725
2	5 331	—2,096	4,393	23419,083
3	10 649	—1,096	1,201	12789,449
4	14 959	—0,096	0,009	134,631
5	11 929	0,904	0,817	9745,993
6	6 678	1,904	3,625	24207,750
7	2 092	2,904	8,433	17641,836
8	342	3,904	15,241	5212,422
Итого $S=53\,680$		—	—	110991,944

Исходные данные для расчета по формуле Маркова:

$$n=8;\ p=0{,}512;\ q=1-p=0{,}488;\ pq=0{,}249856;\ np=8\cdot 0{,}512=$$
$$=4{,}096;\ npq=8\cdot 0{,}512\cdot 0{,}488=1{,}99885;\ Npq=429440\times$$
$$\times\,0{,}249856=107298{,}16.$$

Тогда, производя расчеты, получаем соответствующие графы табл. 6.4:

$$m_1-np=0-4{,}096=-4{,}096;$$
$$m_2-np=1-4{,}096=-3{,}096 \text{ и т. д.};$$
$$(m_1-np)^2=(-4{,}096)^2=16{,}777;$$
$$(m_2-np)^2=(-3{,}096)^2=9{,}585 \text{ и т. д.};$$
$$(m_1-np)^2S_1=16{,}777\cdot 215=3607{,}55;$$
$$(m_2-np)^2S_2=9{,}585\cdot 1485=14233{,}725 \text{ и т. д.};$$
$$\Sigma(m_i-np)^2S_i=110991{,}944.$$

Вычисляем фактическое значение критерия Маркова:

$$D=\frac{110991{,}944}{107298{,}16}\approx 1{,}034.$$

Находим вероятность $D\geqslant 1+\varepsilon$, где $\varepsilon=0{,}034$. Доказано, что эта вероятность при достаточно большом значении S меньше $e^{-\frac{\varepsilon^2 S}{n}}$.

В нашем примере получаем $p<e^{-\frac{0{,}034^2\cdot 53680}{8}}=e^{-7{,}757}$.

Полученная вероятность очень мала. Это дает основание для вывода о чрезвычайно малой вероятности того, что частость рождения мальчика, равная во всей совокупности семей 0,512, одинакова во всех группах семей.

Эту же задачу можно решить совсем иным способом.

Из табл. 6.4 видно, что в 215 семьях с восьмью детьми были только одни девочки, а в 342 семьях — только мальчики. Число

семей с детьми одного пола составляло $215+342=557$. Вероятность рождения мальчика $p_0=0{,}512$. Вероятность того, что в 557 семьях все дети окажутся одного пола, равна:

$$p=p_0^8+q_0^8=0{,}512^8+0{,}488^8=0{,}00793.$$

Определяем вероятность неравенства

$$m-Sp\geqslant t\sqrt{Spq}+\frac{q-p}{6}t^2+\frac{p^3+q^3}{12\sqrt{Spq}}t^3,$$

которая окажется меньше $e^{-\frac{t^2}{2}}$ при условии, что $\frac{t}{\sqrt{Spq}}<\frac{5}{2}$.

Оценим вероятность того, что величина 0,512 является постоянной величиной для всех семей.

В нашем примере получаем: $S=53\,680$, математическое ожидание числа семей с детьми одного пола составляет $Sp=53\,680\times$ $\times 0{,}00793\approx 425{,}68$;

$$\sqrt{Spq}=\sqrt{53\,680\cdot 0{,}00793\cdot 0{,}99207}=\sqrt{422{,}31}\approx 20{,}55.$$

Тогда

$$m-425{,}68\geqslant 20{,}55t+\frac{0{,}98414}{6}t^2+\frac{0{,}09862}{12\cdot 20{,}55}t^3.$$

При $t=6$ получаем, что вероятность того, что $m-425{,}68\geqslant$ $\geqslant 123{,}30+5{,}90+0{,}89=130{,}09$, или что $m\geqslant 425{,}68+130{,}09\approx 556$, $e^{-18}<\frac{1}{65\,000\,000}$.

Так как число семей с детьми одинакового пола, равное 557, достаточно велико, то маловероятно, что вероятность рождения мальчика 0,512 является постоянной величиной для всех семей.

Рассуждая статистически, мы могли подойти к пониманию и решению этой задачи совершенно иначе. Надо выявить степень однородности изучаемой совокупности семей по числу сыновей. Как устанавливает теория проверки статистических гипотез, гипотеза об однородности совокупности не может быть подтверждена, а может быть с помощью статистических критериев лишь отвергнута. Отвергнув гипотезу об однородности совокупности

Таблица 6.5. **Расчеты для нахождения среднего числа сыновей на одну семью и дисперсии**

Число сыновей в семьях с 8 детьми, m_i	Число семей, S_i	$m_i\ S_i$	m_i^2	$m_i^2\ S_i$
0	215	0	0	0
1	1 485	1 485	1	1 485
2	5 331	10 662	4	21 324
3	10 649	31 947	9	95 841
4	14 959	59 836	16	239 344
5	11 929	59 645	25	298 225
6	6 678	40 068	36	240 408
7	2 092	14 644	49	102 508
8	342	2 736	64	21 888
Итого	53 680	221 023	—	1 021 023

и тем самым зафиксировав неправомерность использования полученной средней в этой совокупности в качестве типичной величины для всех групп этой совокупности, мы приходим к выводу о том, что средняя величина в неоднородной совокупности — это фикция, являющаяся огульной величиной, скрадывающей различия отдельных групп совокупности. Соответствующие расчеты для нашего примера произведены в табл. 6.5.

Для вычисления среднего числа сыновей на одну семью используем ΣS_i и $\Sigma m_i S_i$ и получаем

$$\overline{m}=\frac{\Sigma m_i S_i}{\Sigma S_i}=\frac{221\,023}{53\,680}\approx 4{,}12.$$

Для определения дисперсии привлекаем формулу $\sigma^2=\overline{m^2}-(\overline{m})^2$. Используя данные табл. 6.5, получаем

$$\overline{m^2}=\frac{\Sigma m_i^2 S_i}{\Sigma S_i}=\frac{1\,021\,023}{53\,680}\approx 19{,}02; \qquad (\overline{m})^2=4{,}12^2=16{,}97.$$

Тогда дисперсия равна: $\sigma^2=19{,}02-16{,}97=2{,}05$ и среднее квадратическое отклонение $\sigma\approx 1{,}43$. Коэффициент вариации составляет

$$V=\frac{\sigma}{\overline{m}}\cdot 100=\frac{1{,}43}{4{,}12}\cdot 100\%\approx 35\%.$$

Величина коэффициента вариации больше 30% позволяет опровергнуть гипотезу об однородности данной совокупности по изучаемому признаку. Следовательно, среднее число сыновей, приходящееся на одну семью, нехарактерно для всех групп семей.

Эмпирический коэффициент дисперсии Маркова можно использовать и для проверки гипотезы о том, что частость рождения мальчика для конкретного населения и заданного периода есть величина постоянная.

Пусть для десяти последовательных лет частости рождений имели следующие значения: 0,5160; 0,5157; 0,5161; 0,5158; 0,5156; 0,5150; 0,5166; 0,5152; 0,5160; 0,5161 (наибольшее 0,5166, наименьшее 0,5150, разность равна 0,0016). При этом допустим, что в течение изучаемых десяти лет колебания абсолютных чисел родившихся были небольшими и ежегодное число рождений (n) было равно 1 250 000. Тогда частость рождения мальчика за весь период будет представлять среднюю арифметическую невзвешенную из приведенных выше десяти частостей:

$$p=\frac{0{,}5160+\ldots+0{,}5161}{10}=0{,}5158.$$

Используем эмпирический коэффициент дисперсии и находим

$$D=\Sigma\,\frac{(p_i-0{,}518)^2}{10\cdot 0{,}5158\cdot 0{,}4842}\cdot 1\,250\,000=0{,}97.$$

Полученный результат не противоречит гипотезе о постоянстве частости рождения мальчиков в течение десяти лет.

Рассмотренные примеры проверки различных гипотез о постоянстве частостей имеют большое практическое значение.

6.5. Существенность отклонений частости рождения мальчиков от апостериорной вероятности

Рассмотрим методику выявления существенности отклонения фактической частости мальчиков от теоретической, принятой равной 0,516 (апостериорная вероятность), и причины такого отклонения. Пусть из 4040 зарегистрированных в течение года рождений в населенной местности N мальчиков оказалось 2055. Следовательно, $w=\frac{2055}{4040}=0{,}509$; $p=0{,}516$. Определим, вследствие влияния какой причины — систематической или случайной — имеет место отклонение $0{,}516-0{,}509=0{,}007$. Используем интегральную формулу Лапласа:

$$P(a\leqslant m\leqslant b)=\frac{1}{\sqrt{2\pi}}\int\limits_{\alpha}^{\beta} e^{-\frac{t^2}{2}}\,dt=\frac{1}{2}\,[F(t=\beta)-F(t=\alpha)],$$

где $P(a\leqslant m\leqslant b)$ — вероятность того, что при n числе независимых испытаний число появления события A будет заключено между числами a и b; $F(t)$ — интеграл вероятностей; α и β — пределы интегрирования, связанные с числами a и b следующими формулами:

$$\alpha=\frac{a-np}{\sqrt{npq}} \quad \text{и} \quad \beta=\frac{b-np}{\sqrt{npq}}.$$

Произведем замену

$$\alpha=-\varepsilon\sqrt{\frac{n}{pq}}; \qquad \beta=\varepsilon\sqrt{\frac{n}{pq}}$$

и после соответствующих преобразований получим равенство

$$P\{|m-np|<\varepsilon\}=F\left(\frac{\varepsilon}{\sqrt{npq}}\right),$$

где ε — абсолютное значение разности между фактическим числом рождений и наивероятнейшим.

Если вероятность противоположного события, т. е. $|m-np|\geqslant\varepsilon$, окажется достаточно большой, то это будет свидетельством того, что отклонение фактической частости рождения мальчиков от теоретической при данном числе зарегистрированных рождений можно считать обусловленным случайными причинами.

По условию задачи дано:

$$n=4040; \quad m=2055; \quad w=\frac{m}{n}\approx 0{,}509; \quad p=0{,}516; \quad q=0{,}484;$$

$$np=2085; \quad npq=1008{,}95; \quad \sqrt{npq}=31{,}76; \quad \varepsilon=|m-np|=30.$$

Тогда $P\{|m-np|<30\}=F(0{,}94)$.

По соответствующей таблице[1] находим $F\ (0{,}94)=0{,}653$. Значит вероятность противоположного события $P\{|m-np|\geqslant 30\}=$ $=1-0{,}653=0{,}347$.
По этой вероятности заключаем, что в данном случае отклонение фактической частости рождения мальчика, равной 0,509, от теоретической 0,516 можно считать обусловленным случайными причинами.

Можно подойти к решению этой задачи иначе. Полагая, что наблюдавшиеся 4040 рождений являются выборочной совокупностью, извлеченной из генеральной совокупности, найдем среднюю ошибку выборки для доли (μ):

$$\mu=\sqrt{\frac{pq}{n}}=\sqrt{\frac{0{,}516\cdot 0{,}484}{4040}}=0{,}00786.$$

Тогда, с вероятностью 0,997 ($t=3$) предельная ошибка выборки (Δ) составит

$$\Delta=t\mu=3\cdot 0{,}00786\approx 0{,}024.$$

Следовательно, с очень большой вероятностью можно утверждать, что значение фактической частости может колебаться в границах от 0,509—0,024 до 0,509+0,024, т. е. от 0,485 до 0,533. Этот результат также свидетельствует о том, что отклонение полученной в примере фактической частости 0,509 от теоретической 0,516 не выходит за границы случайных колебаний.

Иногда возникает необходимость сопоставления частостей, исчисленных для двух совокупностей, при ответе на вопрос о существенности или несущественности различий между лежащими в их основе апостериорными вероятностями. Имеем для сопоставления $w_1=\frac{m_1}{n_1}$ и $w_2=\frac{m_2}{n_2}$. Пусть $w_1<w_2$, требуется оценить существенность различия апостериорных вероятностей p_1 и p_2.

Первый путь. Находим с определенной достаточно большой (0,954 или 0,997) вероятностью границы каждого показателя. Если нижняя граница w_1, т. е. $w_1-\Delta_1$, больше верхней границы w_2, т. е. $w_2+\Delta_2$, то апостериорные вероятности p_1 и p_2, лежащие в основе частостей w_1 и w_2, различны. Если же интервалы $w_1-\Delta_1$ и $w_2+\Delta_2$ находят один на другой, то вопрос остается открытым.

Рассмотрим конкретный пример. Имеются две совокупности новорожденных. Первая совокупность состоит из 1 436 587 новорожденных, в числе которых 737 629 мальчиков. Тогда

$$m_1=737629;\ n=1\,436\,587;\ w_1=\frac{737\,629}{1\,436\,587}=0{,}51346;$$

$$1-w_1=0{,}48654;\ w_1(1-w_1)=0{,}24982;$$

$$\mu_1=\sqrt{\frac{0{,}24982}{1\,436\,587}}\approx 0{,}0003;$$

[1] См.: *Венецкий И. Г., Венецкая В. И.* Основные математико-статистические понятия и формулы в экономическом анализе, с. 402.

с вероятностью 0,997 имеем $\Delta_1=3\,\mu_1=0{,}0009$. Таким образом, апостериорная вероятность рождения мальчика в этом случае находится в границах

$$w_1-\Delta_1\leqslant p_1\leqslant w_1+\Delta_1;\ 0{,}51346-0{,}0009\leqslant p_1\leqslant$$
$$\leqslant 0{,}51346+0{,}0009;\ 0{,}51256\leqslant p_1\leqslant 0{,}51436.$$

Вторая совокупность состоит из 770 941 новорожденных, в числе которых 393 386 мальчиков. Тогда

$$m_2=393\,386;\ n_2=770\,941;\ w_2=\frac{393\,386}{770\,941}=0{,}51027;$$

$$1-w_2=0{,}48973;\ w_2(1-w_2)=0{,}24989;$$

$$\mu_2=\sqrt{\frac{0{,}24989}{770\,941}}\approx 0{,}0005;$$

с вероятностью 0,997 имеем $\Delta_2=3\,\mu_2=0{,}0015$. Таким образом, апостериорная вероятность рождения мальчика в этом случае находится в границах

$$w_2-\Delta_2\leqslant p_2\leqslant w_2+\Delta_2;\ 0{,}51027-0{,}00150\leqslant p_2\leqslant 0{,}51027+$$
$$+0{,}00150;\ 0{,}50877\leqslant p_2\leqslant 0{,}51177.$$

Видим, что нижняя граница p_1, равная 0,51256, больше верхней границы p_2, равной 0,51177. Это свидетельствует о существенности различий апостериорных вероятностей рождения мальчика в этих совокупностях. Имеются какие-то глубокие причины, лежащие в основе данного различия.

Второй путь. Используем критерий существенности различия

$$\frac{|w_1-w_2|}{\sqrt{\frac{w_1(1-w_1)}{n_1}+\frac{w_2(1-w_2)}{n_2}}}>3.$$

Получаем

$$\frac{0{,}51346-0{,}51027}{\sqrt{\frac{0{,}24982}{1\,436\,587}+\frac{0{,}24989}{770\,941}}}=\frac{0{,}00319}{\sqrt{0{,}0000001+0{,}0000003}}=\frac{0{,}00319}{0{,}00063}\approx 5;$$

$$5>3.$$

Значит, можно считать различие существенным.

По данным примера оценим влияние численности изучаемой совокупности на вывод о существенности различий частостей от апостериорных вероятностей.

Пример 1. Допустим, что в приведенном выше примере (см. с. 209) родившиеся составляют не 4040, а достаточно большое число, равное, например, 100 000. Тогда средняя ошибка будет

$$\mu=\sqrt{\frac{w(1-w)}{n}}=\sqrt{\frac{0{,}516\cdot 0{,}484}{100\,000}}\approx 0{,}00158.$$

Предельная ошибка, найденная с вероятностью 0,997 ($t=3$), будет равна: $t=t\mu=3\cdot0{,}00158\approx0{,}005$. Следовательно, границы найденной величины 0,509 составляют $0{,}509\pm0{,}005$; т. е. лежат в интервале от 0,504 до 0,514. Так как величина 0,516 не входит в этот интервал, то отклонение 0,509 от 0,516 значимо. Вероятность рождения мальчика, равная 0,516, не согласуется с утверждением о равенстве вероятностей рождения мальчиков и девочек друг другу. Только для упрощения расчетов и рассуждений их иногда принимают равными друг другу, т. е. 0,5.

Пример 2, приведенный С. Н. Бернштейном[1]. Из общего числа членов профсоюза, равного 100 000 человек, в два срока производят выборки численностью в 1000 и 1500 человек. Доля женщин среди попавших в выборку по срокам оказалась соответственно равной 0,2150 и 0,2007. Можно ли считать, что среди членов данного профсоюза удельный вес женщин уменьшился? Получаем средние ошибки выборки:

$$\mu_1=\sqrt{\frac{w_1(1-w_1)}{n_1}}=\sqrt{\frac{0{,}2150\cdot0{,}7850}{1000}}=\sqrt{0{,}00016775}=0{,}012988.$$

$$\mu_2=\sqrt{\frac{w_2(1-w_2)}{n_2}}=\sqrt{\frac{0{,}2007\cdot0{,}7993}{1500}}=\sqrt{0{,}000107}=0{,}01034.$$

Средняя ошибка разности долей составляет

$$\mu_{\text{разн}}=\sqrt{\mu_1^2+\mu_2^2}=\sqrt{0{,}000168+0{,}0001070}=\sqrt{0{,}000275}=0{,}01630.$$

Разность двух долей равна: $w_1-w_2=0{,}2150-0{,}2007=0{,}0143$. Тогда

$$\frac{|w_1-w_2|}{\mu_{\text{разн}}}=\frac{0{,}0143}{0{,}0166}=0{,}86.$$

Следовательно, нет серьезных оснований считать, что процент женщин среди членов профсоюза действительно уменьшился. В данном случае нулевая гипотеза не отвергается. Если бы объем выборки был больше и в обоих случаях равнялся бы 10 000, то

$$\mu_1=\sqrt{\frac{0{,}2150\cdot0{,}7850}{1000}}=\sqrt{0{,}000017};$$

$$\mu_2=\sqrt{\frac{0{,}2007\cdot0{,}7993}{10\,000}}=\sqrt{0{,}000016};$$

$$\mu_{\text{разн}}=\sqrt{0{,}000017+0{,}000016}=\sqrt{0{,}000033}\approx0{,}0057.$$

Тогда

$$\frac{0{,}0143}{0{,}0057}\approx2{,}5.$$

[1] *Бернштейн С. Н.* Теория вероятностей, с. 298.

В этом случае с большой вероятностью можно утверждать, что разность долей существенна, т. е. удельный вес женщин — членов профсоюза действительно уменьшился.

Иногда разность между показателями хотя и мала, но численность изучаемых совокупностей так велика, что такая малая разность может оказаться существенной. Например, в Латвийской ССР в 60-х годах среднее число детей у женщин, вступивших в брак до 25 лет (доля которых равна 54%), составило 1,42, а у женщин, вступивших в брак в 25 лет и старше (доля 46%), — 1,36. Как показало исследование, даже такое небольшое отличие при значительной численности обследованных явилось существенным.

Для оценки существенности расхождения между числами родившихся мальчиков и девочек В. Я. Буняковский вывел следующую формулу:

$$P=1-\frac{(p+q)^{p+q+\frac{3}{2}}}{(p-q)^{p+\frac{1}{2}}\,q^{q+\frac{1}{2}}\sqrt{\pi}}\left\{1-\frac{p+q}{(p-q)^2}+\frac{13pq-(p+q)^2}{12pq(p+q)}-\cdots\right\}$$

где p и q — абсолютные числа родившихся мальчиков и девочек; P — вероятность того, что расхождения между p и q существенны.

В качестве примера были привлечены статистические данные о рождаемости в Петербурге за 10 лет. Число зарегистрированных рождений за этот период составило 56 917 мальчиков и 54 636 девочек. В результате использования указанной выше формулы В. Я. Буняковский получил

$$P=1-\frac{0{,}375209}{10^{11}}.$$

Вывод был следующий: «Эта вероятность так близка к единице, что возможность рождения младенца мужского пола преимущественно перед женским, в Петербурге, должно считать в высшей степени вероятной»[1]. Используя критерий существенности, к этому же выводу можно подойти иным путем. Имеем

$$t=\frac{p_1-p_2}{\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}}.$$

Получаем

$$t=\frac{0{,}51-0{,}49}{\sqrt{\frac{0{,}51\cdot0{,}49}{56\,917}+\frac{0{,}49\cdot0{,}51}{54\,636}}}\approx6{,}8.$$

При достаточно больших n (>100) значение $t=2{,}6$ с вероятностью 0,99 можно считать достаточным для подтверждения существенности различий сравниваемых долей рождения мальчиков и девочек.

[1] *Буняковский В. Я.* Основания математической теории вероятностей. Спб., 1846, с. 206.

Глава 7

Вероятностный подход к изучению миграций

7.1. Вероятностные показатели миграции

При анализе перемещений людей, т. е. миграции населения, а также демографических процессов, связанных с этими перемещениями, есть все основания использовать вероятностные методы: налицо массовые явления, состоящие из элементов, носящих случайный характер, с взаимопогашением индивидуальных отклонений.

В самом деле, мигранты составляют большую совокупность людей, изъявивших по тем или иным, чаще всего самопроизвольным и случайным, причинам желание совершить переселение. При этом отклонения результативного признака — величины миграции — в обе стороны от средней нейтрализуются, т. е. взаимопогашаются. В ряде случаев высказывается мнение, что наряду с рождаемостью миграция населения является процессом, хотя до некоторой степени и управляемым, но преимущественно стихийным. Между тем вероятности наступления миграции можно исчислять довольно точно, ибо, как показали Б. С. Хорев и В. Н. Чапек, «...вероятность определенного контакта с населением (речь идет о мигрантах. — *И. Г. В.*) и адаптации в районах вселения оказывается пропорциональной численности населения в этих районах»[1].

При изучении причинно-следственных связей миграции следует иметь в виду, что миграционные процессы, в которых очень сложно переплетаются экономические, исторические, нравственные, психологические, демографические и этнические факторы внутрирайонного перераспределения, в большей степени являются вероятностными.

Именно поэтому, как мы уже видели, в последнее время широкое распространение получили модели вероятностного типа, основанные на использовании марковских цепей (см. разд. 5.2).

[1] *Хорев Б. С., Чапек В. Н.* Проблемы изучения миграции населения. М., 1978, с. 58.

Переходные вероятности рассчитываются путем фиксации распределения численностей миграционных потоков для двух моментов наблюдения. При этом в расчет включается вектор начального распределения и матрица переходных вероятностей. Определяется вектор конечного состояния. Учитывая реальность нескольких направлений будущей эволюции демографических процессов, вероятностный подход позволяет осуществлять многовариантный прогноз возрастно-половой структуры и численности населения.

Разработанная Т. И. Заславской схема миграции использует показатель вероятности переселения[1]:

$$P_{kj}=Q_{kj}-E=S_{kj}-S_{ki}-E=\Sigma\, \Delta_{kjl}V_{lk}J_{lkj}-$$
$$-\Sigma\, Z_{kjm}W_{km}J_{kmj}-E,$$

где P_{kj} — вероятность переселения k-го члена группы; i — индекс исходного населенного пункта; j — место вселения; l — характеристика условий жизни; m — затраты, необходимые для перемены места жительства; V — нормированные показатели сравнительной важности разных жизненных условий, принимающие значения от нуля до единицы; J — показатель наличия или отсутствия информации по определенному вопросу, принимающий для отдельных индивидов значения либо нуля, либо единицы; S — положительный стимул к миграции; по отношению к одному индивиду $S_{kj}=\Sigma\,\Delta_{kjl}V_{lk}J_{lkj}$; Z — нормализованный показатель затрат на миграцию, стимулирующий сохранение статус-кво; E — пороговая оценка эффективности переселения для индивида, превышение которой приводит к принятию соответствующего решения; W — нормированный показатель сравнительной важности для индивидов отдельных видов затрат, принимающий значение от нуля до единицы.

На основе полученного уравнения автор считает возможным определять вероятные размеры миграции из одних населенных пунктов в другие.

Вероятности используются также при когортном анализе географической и профессиональной мобильности. Изучаются вероятности перемещений, т. е. вероятности того, что лицо, совершившее перемещение порядка n, сменит место жительства еще раз. Расчет вероятностей перемещений производится аналогично вероятностям увеличения семьи путем суммирования частостей перемещений более высоких, чем данное, и нахождения отношений двух последовательных сумм. Возьмем, например, поколение 50-летних и рассмотрим частости перемещений людей, принадлежащих к этому поколению.

[1] См.: *Заславская Т. И.* Методологические проблемы изучения миграции сельского населения. — В кн.: Статистика миграций населения. М., 1972, с. 138—164.

Перемещения	Частости	Перемещения	Частости
0	0,266	4	0,072
1	0,302	5 и более	0,051
2	0,187	Всего	1,000
3	0,122		

Обозначая вероятности смены места жительства d, получим: d_0 — вероятность первого перемещения; d_1 — вероятность смены места жительства второй раз при условии, что одно переселение уже произведено, и т. д.

$$d_0=\frac{0{,}302+0{,}187+0{,}122+0{,}072+0{,}051}{0{,}266+0{,}302+0{,}187+0{,}122+0{,}072+0{,}051}=\frac{0{,}734}{1{,}000}=0{,}734\,;$$

$$d_1=\frac{0{,}187+0{,}122+0{,}072+0{,}051}{0{,}302+0{,}187+0{,}122+0{,}072+0{,}051}=\frac{0{,}432}{0{,}734}=0{,}589.$$

Найденные вероятности показывают интенсивность последовательных перемещений.

При изучении миграционных процессов используется вероятностный подход, состоящий, в частности, в утверждении, что поток мигрантов в район вселения составит не S человек, а $S\pm\Delta$ человек. При этом каждая из возможных величин потока имеет определенную вероятность. В качестве характеристик потоков используются максимальные и минимальные значения потока, среднее значение, дисперсия, среднее квадратическое отклонение, коэффициент вариации и темп его роста по годам, вероятности попадания возможных значений миграционных потоков в тот или иной интервал, вероятности отклонений расчетных миграционных потоков от их фактических значений за различные годы[1]. В ряде случаев при изучении миграции при отсутствии данных о численности населения на начало периода вычисляют поствероятности, в которых в знаменателе — численность населения не на начало, а на конец исследования.

Вероятностный подход позволяет изучать и маятниковые миграции[2]. Вероятность маятниковой миграции между пунктами i и k можно обозначить следующим образом:

$$p_{ik}=\frac{d_{ik}}{\sum\limits_{k=1}^{n} d_{ik}}\,,$$

где d_{ik} — функция оценки затрат времени равна $\frac{1}{t^2_{ik}}$; t_{ik} — затраты времени на передвижение между пунктами i и k. В этом случае вероятность миграции уменьшается при увеличении времени на ее реализацию. Учитывая численность населения жилого района поселения (A_j) и общее количество мигрантов, отправляющихся из пункта i (G_i), получаем поток мигрантов между i и j. Данную логическую схему маятниковой миграции называют гравитационной формулой маятниковой миграции. Очевидно, что $\sum\limits_j d_{ij}=G_i$ и $\sum\limits_i d_{ij}=A_j$, т. е. сумма вероятного числа жителей, желающих работать в районе, должна быть равна фактическому числу занятых в этом районе, а сумма вероятного числа трудящихся, желающих жить в районе j, должна быть равна фактической численности самодеятельного населения этого района.

[1] См.: *Матлин И. С.* Моделирование размещения населения. М., 1975, с. 157.

[2] См.: *Заблодский Г. А.* Общие предпосылки прогнозирования миграции населения.— В кн.: Демографические тетради. Киев, 1972, вып. 4—5, с. 96.

Обычно величина d_{ij} рассчитывается по формулам, получаемым эмпирическим путем:

$$d_{ij}=t_{ij}^{-\alpha} \quad \text{или} \quad d_{ij}=e^{-\beta t_{ij}}.$$

При изучении закономерности изменения плотности населения в городах относительно главных центров были выведены различные законы: закон убывания плотности населения по показательной кривой — закон Кларка; закон убывания плотности по положительному нормальному закону распределения — закон Г. Шеррата и др. В дальнейшем функции, по которым аппроксимируются изменения плотностей населения, усложнились.

Г. В. Шелниковский, рассматривая функцию распределения общей численности передвигающегося населения по дальности и используя дифференциальное уравнение психофизиологического закона Вебера-Фехнера

$$d\varphi(t)=-a\frac{dt}{t}\,,$$

получил плотность вероятности трудовых передвижений

$$\varphi(t)=\frac{1}{T}\ln\frac{T}{t}\,,$$

где t — время передвижения; T — предел расселения[1]. Без использования параметра T данная функция получает вид показательной функции.

При определении уровня урбанизации (V_j) Э. Арриага предложил индекс, исчисляемый по формуле

$$V_j=\frac{\sum\limits_{i=1}^{n} C_i^2}{P_j}\,,$$

где C_i — численность населения i-го города; P_j — численность населения j-го региона.

Учитывая тот недостаток этого индекса, что он плохо характеризует уровень урбанизации районов с одним доминирующим центром, было внесено предложение учитывать пространственные особенности изучаемых районов, а также отражать степень освоенности районов страны городами[2]. Для этой цели при измерении степени урбанизации используется формула

$$T_j=V_jK_j\,,$$

[1] См.: *Шелниковский Г. В.* Планировка, транспорт и расселение.

[2] См.: *Зорин И. В., Канцебовская И. В.* Некоторые методы измерения уровня урбанизации. — В кн.: Проблемы современной урбанизации/ Под ред. Ю. Л. Пивоварова. М., 1972, с. 195.

где K_j — коэффициент площади j-го района, вычисляемый путем привлечения вероятностного показателя (p_{sj})

$$K_j = \frac{1 - p_{sj}}{p_{sj}} \quad \text{при} \quad p_{sj} = \frac{S_j}{S},$$

где S_j и S — площади j-го региона и всей страны.

7.2. Вероятностные методы измерения смертности (летальности) в связи с миграцией

Миграция оказывает большое влияние на величину вероятности дожития и смерти. Установим погрешности вероятностных показателей, найденных без учета миграции.

Следуя системе обозначений, принятой Г. А. Баткисом[1], введем следующие показатели: L — показатель смертности (летальности); B — прибывшие в данное место за календарный период; A — находившиеся в данном месте к началу данного периода; E — находившиеся в данном месте в конце данного периода; D — умершие в данном месте в период наблюдения; C — выбывшие в данный период.

Путем рассуждения можно установить, что показатель летальности (L) находится в следующих границах:

$$\frac{D}{B} < L < \frac{D}{C+D} \text{ либо } \frac{D}{B} > L > \frac{D}{C+D}.$$

Предположим, что величина L есть средняя арифметическая своих крайних значений. Тогда окончательно формула примет вид

$$\overline{L} = \frac{1}{2}\left(\frac{D}{B} + \frac{D}{C+D}\right) + \left(\frac{D}{B} - \frac{D}{C+D}\right).$$

Указанные методы Г. А. Баткис рассматривает на примере движения больных. Предположим, что на начало отчетного периода больных было 70 (A), поступило 140 (B), выписано 74 (C), умерло 6 (D), осталось на конец отчетного периода 130 (E).

По этим данным вычислим показатель летальности:

$$L_1 = \frac{D}{B} = \frac{6}{140} = 0{,}043;$$

$$L_2 = \frac{D}{C+D} = \frac{6}{74+6} = \frac{6}{80} = 0{,}075;$$

$$L_3 = \frac{D}{\frac{B+E+D}{2}} = \frac{6}{\frac{140+74+6}{2}} = \frac{6}{110} = 0{,}055.$$

[1] См.: *Баткис Г. А.* Вопросы санитарной и демографической статистики. М., 1964.

Первые два показателя представляют рассмотренные ранее границы L (летальности): $0{,}043 < L < 0{,}075$. Третий показатель пригоден в условиях быстрой сменяемости больных. Величину этого показателя Г. А. Баткис рассчитывает следующим образом:

$$\overline{L} = \frac{0{,}043 + 0{,}075}{2} \pm \frac{0{,}043 - 0{,}075}{2} = 0{,}059 \pm 0{,}016.$$

Таким образом, величина $\overline{L}$ (равная 0,059), исчисленная методом Г. А. Баткиса, чуть выше величины $\overline{L}$ (0,055), рассчитанной другим методом. Очень интересно, что в данном случае метод Г. А. Баткиса можно дополнить и исчисление коэффициента смертности (летальности) поставить на вероятностную основу.

Найдем отношение числа умерших к средней величине из числа больных, состоящих на начало периода и оставшихся на конец его, т. е. вычислим долю умерших (w):

$$w = D : \frac{A+E}{2} = \frac{2D}{A+E}.$$

Тогда дисперсию альтернативного признака можно вычислить следующим путем:

$$\sigma^2 = w(1-w) = \frac{2D}{A+E}\left(1 - \frac{2D}{A+E}\right) = \frac{2D(A+E-2D)}{(A+E)^2}.$$

Найдем предельную ошибку выборки:

$$\Delta = t\sqrt{\frac{2D \cdot 2(A+E-2D)}{(A+E)^2(A+E)}} = \frac{2}{A+E}\, t\sqrt{\frac{D(A+E-2D)}{A+E}}.$$

Отсюда нетрудно определить t, а затем по таблице найти вероятность (p).

Принимая разность $L_2 - L_1$ за величину ошибки (без деления на два, как у Г. А. Баткиса), получим

$$\frac{A+E}{2} = \frac{70+130}{2} = 100; \quad w = \frac{6}{100} = 0{,}06;$$

$$1 - w = 0{,}94; \quad w(1-w) = 0{,}06 \cdot 0{,}94 = 0{,}0564;$$

$$\Delta = t\sqrt{\frac{0{,}0564}{100}} = t\sqrt{0{,}000564} \approx t0{,}024.$$

Итак, предельная ошибка выборки (Δ) равна 0,024 t. Зная, что без деления на два она равна 0,032, найдем t, т. е. $t = \frac{0{,}032}{0{,}024} = 1{,}34$. Значит, вероятность p равна 0,8.

Следовательно, вероятность того, что смертность (летальность) находится в границах $0{,}059 \pm 0{,}016$, равна 0,8, это не дает полной уверенности в том, что смертность не выйдет за указанные границы. Если бы в предыдущем примере число умерших составляло не 6 (как в примере Г. А. Баткиса), а, например, 20 человек, то расчет привел бы к другим показателям и вывод изменился бы. Определим величину оставшихся под наблюдением на конец периода: $E = A + B - C - D = 70 + 140 - 74 - 20 = 116$.

Тогда $$\frac{A+E}{2} = \frac{70+116}{2} = 93;$$

$$w = \frac{20}{93} = 0{,}215;\ 1-w = 0{,}785;\ w(1-w) = 0{,}1688.$$

Предельная ошибка выборки (Δ) равна $0{,}035\,t$.

Кроме того, следует учитывать изменение смертности (L) и ошибку смертности (Δ). Определим отношение умерших к поступившим

$$L_1 = \frac{D}{B} = \frac{20}{140} = 0{,}143$$

и отношение умерших к выбывшим

$$L_2 = \frac{D}{C+D} = \frac{20}{74+20} = \frac{20}{94} = 0{,}213.$$

Тогда смертность будет равна следующей величине:

$$\overline{L} = \frac{0{,}143+0{,}213}{2} = \frac{0{,}356}{2} = 0{,}178,$$

а ошибка составит $\Delta = 0{,}213 - 0{,}143 = 0{,}070$.

Следовательно, $t = \dfrac{0{,}070}{0{,}035} = 2$ и $p > 0{,}95$, т. е. вероятность существенно повысилась.

Таким образом, получив вероятность $p > 0{,}95$, можно с практической достоверностью утверждать, что смертность находится в границах $0{,}178 \pm 0{,}035$.

Попробуем установить меру неточности показателей таблиц смертности, если в расчет не принято механическое движение населения. В таблицах смертности населения Советского Союза ввиду крайне малой внешней миграции неточностью можно пренебречь. В таблицах же смертности населения какой-нибудь республики или меньшей территориальной единицы с группировкой населения на городское и сельское, где миграция может быть ощутимой, влияние ошибки вырастает.

Вероятность дожития до возраста $x+1$ год с учетом миграции можно рассчитать по формуле

$$p'_x = \frac{L_x - m'_в}{L_x + m'_н \pm A},$$

где $m'_в$ и $m'_н$ — верхняя и нижняя элементарные совокупности умерших; A — абсолютная величина механического прироста в возрасте x лет того же года рождения.

Расчет вероятности дожития с учетом миграции отличается от расчета вероятности дожития без учета миграции величиной A в знаменателе.

Определим влияние величин A на p_x. Для этого найдем разность между вероятностями дожития с учетом и без учета миграции:

$$\Delta = p_x - p_x' = \frac{L_x - m'_в}{L_x + m'_н} - \frac{L_x - m'_в}{L_x + m'_н \pm A}.$$

Обозначив $L_x - m'_в$ через L'_x и $L_x + m'_н$ через L''_x, получим

$$\Delta = \frac{L'_x}{L''_x} - \frac{L'_x}{L''_x \pm A} = \frac{\pm L'_x A}{(L''_x)^2 \pm A L''_x}.$$

Считая сальдо миграции A положительным, разделим на эту величину числитель и знаменатель:

$$\Delta = \frac{L'_x}{\dfrac{(L''_x)^2}{A} + L''_x}.$$

Рассмотрим случай, когда относительная миграция составляет 5% естественного прироста $\left(\dfrac{A}{L_x} = 0{,}05\ K_{е.пр}\right)$, а коэффициент естественного прироста — 15‰. Тогда $A = L_x\ 0{,}05 \cdot 0{,}015 = 0{,}00075\,L_x$.

Допустим, что $L''_x \approx L'_x = L_x$, получим

$$\Delta = \frac{L_x}{\dfrac{L_x^2}{0{,}00075} + L_x} \approx 0{,}0001.$$

Таким образом, при этих предположениях ошибка составляет 0,0001 и, следовательно, может влиять на четвертый знак после запятой в величине вероятности дожития или смерти.

Бо́льшая величина миграции может влиять еще больше. Так, если предположить, что относительная миграция составляет 10% естественного прироста, то в величине вероятности дожития или смерти уже третий знак после запятой будет неточным. Следовательно, при достаточно высокой миграции ошибка показателей таблиц смертности составляет значительную величину. Поэтому использование в планировании таблиц смертности, составленных без учета миграции, может привести к значительным просчетам.

Оглавление

Илья Григорьевич Венецкий

ВЕРОЯТНОСТНЫЕ МЕТОДЫ В ДЕМОГРАФИИ

Рецензенты **Е. М. Андреев, С. И. Пирожков**
Зав. редакцией **В. П. Томин**
Редактор **Г. И. Чертова**
Мл. редакторы **Е. М. Рудый, Л. А. Ципивко**
Корректоры **Г. А. Башарина, З. С. Кандыба**
Техн. редактор **И. В. Завгородняя**
Худож. редактор **Э. А. Смирнов**

ИБ № 1067

Сдано в набор 2.03.81 г. Подписано в печать 25.11.81. А 11979. Формат 60×90 1/16. Бум. тип. № 3. Гарнитура «Литературная». Печать высокая. П. л. 14,0. Усл. п. л. 14,0. Уч.-изд. л. 15,16. Тираж 3500 экз. Заказ 2715. Цена 2 р. 50 к.

Издательство «Финансы и статистика», Москва, ул. Чернышевского, 7.

Областная типография управления издательств, полиграфии и книжной торговли Ивановского облисполкома, г. Иваново-8, ул. Типографская, 6.