Ю.Н.Толстова
Математико-статистические модели в социологии
(математическая статистика для социологов)

Учебное пособие

ОГЛАВЛЕНИЕ

5.1. О задачах математической статистики
5.2. Точечные оценки параметров. Предъявляемые к ним требования

Тема 6. Интервальное оценивание параметров
6.1. Понятие доверительного интервала и принципы его построения (на примере математического ожидания)
6.2. Определение объема выборки
6.3. Доверительный интервал для медианы
6.4. Доверительный интервал для доли
6.5. Связь средних ошибок среднего арифметического и доли, обобщение этого факта на многомерный анализ
Примеры задач

Раздел III. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Тема 7. Понятие статистической гипотезы и принципы ее проверки. Проверка гипотезы об отсутствии связи между двумя признаками
7.1. Общее представление о статистической гипотезе
7.2. Логика проверки статистической гипотезы. Использование принципа невозможности реализации маловероятных событий
7.3. Проверка статистической гипотезы об отсутствии связи (критерий «Хи-квадрат»)
Примеры задач

Тема 8. Проверка гипотезы о равенстве средних
8.1. Понятие зависимых и независимых выборок
8.2. Проверка гипотезы для независимых выборок
8.3. Проверка гипотезы для зависимых выборок
Примеры задач

Тема 9. Направленные и ненаправленные альтернативные гипотезы. Односторонние и двусторонние критерии
9.1. Направленные и ненаправленные альтернативные гипотезы
9.2. Односторонние и двусторонние критерии

Тема 10. Проверка статистических гипотез: о равномерности генерального распределения, о равенстве дисперсий, о равенстве нулю коэффициента корреляции, о равенстве долей;
10.1. Проверка гипотезы о равномерности генерального распределения с помощью критерия «Хи-квадрат»
10.2 . Проверка гипотезы о равенстве дисперсий
10.3. Проверка гипотезы о равенстве нулю коэффициента корреляции
10.4. Проверка гипотезы о равенстве долей
Примеры задач

Тема 11. Методологические аспекты проверки математико-статистических гипотез
11.1. Ошибки первого и второго рода
11.2. Пример влияния содержательного характера задачи на выбор уровня значимости
11.3. Различие между статистической и содержательной гипотезой

Раздел IY. ПРОБЛЕМА ИЗУЧЕНИЯ ПРИЧИННО-СЛЕДСТВЕННЫХ ОТНОШЕНИЙ И ЭКСПЕРИМЕНТ В СОЦИОЛОГИИ; ОСНОВНЫЕ ИДЕИ ДИСПЕРСИОННОГО АНАЛИЗА

Тема 12. Методологические аспекты изучения причинно-следственных отношений с помощью математических методов. Эксперимент в социологии
12.1. Проблема изучения причинно-следственных отношений. Обзор некоторых подходов.
12.2. Невозможность полностью формализовать понятия причины и следствия. Выделение двух основных направлений изучения причинных отношений: построение структурных уравнений и проведение эксперимента
12.3. Роль математической статистики при проведении эксперимента. Нестатистический (индуктивный) подход: эксперимент по Миллю

ВВЕДЕНИЕ

В.1. Основная цель курса, адресат

Курс рассчитан на студентов-социологов и посвящен изложению основ математической статистики. По существу он является продолжением курса по теории вероятностей. Как известно, подобные курсы традиционно читаются студентам самых разных специальностей. Объясняется это тем, что изучение статистических закономерностей требуется практически в любой отрасли человеческого знания. Отечественная литература в соответствующем отношении очень богата, имеется множество учебников (в том числе переводных) и методических пособий самого разного плана: с разной широтой охвата проблематики, рассчитанных на читателей с различной подготовкой и т.д. Казалось бы, преподавание математической статистики для студентов - прикладников стало рутинным делом. Тем не менее, предлагаемое учебное пособие имеет ряд особенностей, позволяющих считать его в некоторых отношениях оригинальным (именно это обусловило изменение названия курса). Особенности эти вызваны желанием автора сделать курс хорошо воспринимаемым именно социологами. Ситуация, обусловившая потребность в соответствующих разработках, состоит в следующем.

В.2. Проблемы преподавания математических дисциплин студентам-социологам

Опыт показывает, что студенты-социологи часто бывают настроены на «гуманитарный» лад, и либо вообще отрицают необходимость серьезного рассмотрения каких бы то ни было математических

методов, либо делают это формально, в глубине души считая соответствующие знания для себя лишними. В результате - если не отсутствие знаний, то освоение материала на «абстрактном» уровне, без всякого сопряжения с практикой проведения социологических исследований. Во всяком случае, автору неоднократно приходилось наблюдать, что даже добросовестные студенты плохо представляют себе, как использовать знания, полученные им в курсе теории вероятностей и математической статистики, в практической работе социолога. Преодолеть соответствующую проблему, на наш взгляд, можно путем определенной «привязки» курса к социологическим проблемам.

В.3. Особенности курса

Основной чертой предлагаемого учебного пособия, отличающего его от других учебников соответствующей направленности, является прежде всего то, что все вводимые теоретические положения сопровождаются иллюстрациями их использования в социологических исследованиях. В качестве примеров случайных событий служат события, каждое из которых состоит в том, что какой-либо респондент обладает определенным сочетанием значений рассматриваемых признаков. Сами признаки служат примерами случайных величин (вместо вероятностей в примерах, естественно, фигурируют относительные частоты).

Еще одна особенность работы состоит в том, что в ней большое внимание уделяется проблеме измерения исходных данных. Дело в том, что в социологии проблемы выбора способа получения данных и метода их анализа (в том числе и с помощью алгоритмов математической статистики) не могут решаться отдельно друг от друга, поскольку отражают две стороны одного и того же процесса. В предлагаемом курсе это проявляется прежде всего в том, что, говоря о параметрах распределений, мы соотносим их с типами шкал, использованных при получении исходных данных.

Существенное внимание в курсе уделяется описанию роли статистического подхода в социологии; обсуждается возможность обеспечения того комплекса условий, реализация которого приводит к появлению интересующих социолога случайных событий; в частности, затрагивается проблема существования случайных величин. Рассматривается ряд часто встречающихся в социологии ситуаций, в которых не выполняются условия реализации известных математико-статистических методов. Показывается, как может действовать социолог в таких случаях.

В определенной мере затрагивается история применения статистического подхода к изучению социальных явлений. Потребность в этом объясняется следующими обстоятельствами. Статистический подход, зародившись в XVII веке именно при изучении общества, потом, на стыке XIX и XX веков, начал необоснованно отвергаться некоторыми обществоведами. В какой-то мере возникший кризис был преодолен. Но сейчас, через сто лет, история повторяется. И ретроспективный анализ работ наших предшественников оказывается весьма полезным для современной ситуации.

Одним из проявлений кризисности современной ситуации с использованием математического языка в социологии является присущая многим социологам механистичность использования методов, отсутствие потребности в анализе задействованных в методах моделей, в сопряжении их со смыслом решаемой задачи. Для исправления такого положения дел и представляется полезным обращение к «истокам», «корням», к рассмотрению тех обстоятельств, которые привели к рождению того или иного метода.

Разговор о роли математической статистики в социальных исследованиях в данной работе ведется на фоне обсуждения общих принципов использования математики в социологии. И в качестве основного пласта содержательных задач, выбранного для иллюстраций, используются задачи изучения причинно-следственных отношений. Это представляется естественным, поскольку в содержательном плане методы математической статистики в значительной мере направлены именно на решение соответствующих проблем. Для обеспечения возможности серьезного разговора по поводу связи содержания социологической задачи и математического формализма в работе коротко рассматривается развитие понятия причины и анализируется роль статистического (и не статистического) подхода к ее изучению. И здесь мы также пытаемся обратиться к истокам соответствующих теоретических положений, руководствуясь сформулированным выше принципом: в кризисной ситуации эффективным может быть обращение к «корням».

В.4. Общие организационные требования.

Курс продолжается в течение двух модулей. В конце первого модуля - контрольная работа, в конце второго – экзамен[1]. К середине второго модуля необходимо сдать реферат (эссе). Примерные

[1] Хочется коротко сказать о негативном отношении автора к современной тенденции безоговорочной замены устных экзаменов письменными. У устного и письменного экзаменов примерно те же

темы см. в конце книги. Они охватывают следующие направления: анализ исторических корней математической статистики, их общности с корнями эмпирической социологии; изучение методических достижений русской земской статистики; оценка гносеологических аспектов использования статистических закономерностей в социологии; рассмотрение проблемы построения выборки; осмысление методологических вопросов, возникающих при получении нового социологического знания с помощью математических методов.

Чтение лекций сопровождается проведением семинарских занятий. Примерная тематика последних отражается в списках ориентировочных задач, которые приведены после раскрытия большинства тем. Кроме того, на семинаре должна осуществляться та связь с курсом теории вероятностей, о которой идет речь ниже.

В.5. Связь с курсом теории вероятностей.

Успешное освоение предлагаемого курса возможно только после знакомства студента с элементами теории вероятностей в том объеме, который обычно предполагается рассчитанными на социологов учебными программами соответствующей дисциплины. Ниже указывается, какие именно знания по теории вероятностей требуются для освоения того или иного фрагмента настоящего курса. Такие указания оформлены в виде специальных рубрик (как было сказано, это – материал для семинарских занятий).

В.6. Специфика представления библиографии

В отечественной литературе имеется очень много работ (в том числе переводных), прекрасно описывающих основные положения математической статистики. Список этих работ приведен в конце книги. В них можно найти материал почти по все темам. К *обязательной* литературе мы отнесли работы, либо выпущенные в последние годы, либо ориентированные на социологов или по способу изложения, или по специфике рассматриваемых аспектов (к сожалению, эти работы зачастую опубликованы довольно давно).

После некоторых лекций приведены списки книг, содержание которых более узко - касается только рассматриваемой лекции. Эти списки называются *добавочными*.

Указываются отдельные работы и внутри текста (имеются в виду работы, содержание которых выходит за пределы стандартных курсов по теории вероятностей и математической статистике).

Раздел I. ОБЩЕЕ ПРЕДСТАВЛЕНИЕ О МАТЕМАТИЧЕСКОЙ СТАТИСТИКЕ. ПРЕДВАРИТЕЛЬНЫЕ СВЕДЕНИЯ ОБ ОСНОВНОМ ОБЪЕКТЕ ЕЕ ИЗУЧЕНИЯ – СЛУЧАЙНЫХ ВЕЛИЧИНАХ (измерение, стандартизация, виды распределений, предельные теоремы)

ТЕМА 1
Объект, предмет, цели и задачи математической статистики.

достоинства и недостатки, что и у мягких и жестких методов сбора данных. Устный экзамен позволяет преподавателю более глубоко проникнуть в сознание студента, более адекватно понять, что тот знает, а что не знает. Случайная ошибка студента здесь играет гораздо меньшую роль, чем в письменной работе. Списанная же или механически повторенная фраза быстро обретает свою истинную цену в глазах преподавателя. Экзаменатор и экзаменующийся в процессе устного экзамена взаимно воздействуют друг на друга, каждый из них имеет шанс получить в процессе этого общения нечто новое и полезное для себя. Но, с другой стороны, конечно, скорость проведения письменного экзамена намного выше скорости устного. Здесь необходимо упомянуть о тестовом опросе, который в особой мере повышает эту скорость. У нас имеется глубокое убеждение в том, что такой опрос в принципе не может позволить в должной мере оценить глубину понимания материала студентами. Напротив, он стимулирует студента осваивать материал весьма поверхностно. Использование тестов может быть полезно только на этапах промежуточного контроля. Для нас важнее адекватная оценка степени понимания студентом пройденного материала (вкупе с наличием контакта с экзаменующимися), а не возможность опросить как можно больше студентов за единицу времени. Поэтому мы стремимся по возможности проводить устные экзамены. О преимуществах и недостатках мягких и жестких методов сбора данных много сказано в социологической литературе, но почему-то до сих пор не проведено профессиональных исследований, направленных на оценку возможности переноса этих результатов на экзаменационную ситуацию.

1.1. Понятия выборки и генеральной совокупности.

Надеемся, читатель уже не в первый раз сталкивается с означенными в заголовке параграфа понятиями. Исследователь практически всегда хочет изучить генеральную совокупность, но практически всегда же имеет дело с выборкой. Генеральная совокупность обычно бывает «неуловима».

Все те положения математической статистики, которые мы будем изучать, справедливы лишь для случайной выборки. *Случайной выборкой* называется такая выборка, при построении которой обеспечена одинаковая вероятность попадания в неё любого объекта генеральной совокупности. Классический способ построения случайной выборки состоит в использовании датчика равномерно распределенных случайных чисел применительно к т.н. основе выборки, т.е. к перечню всех элементов генеральной совокупности. Используются также другие способы моделирования случайности, например, механическая выборка. К числу случайных иногда относят также стратифицированную (районированную) и гнездовую (кластерную) выборки. [2]

Социолог практически никогда не имеет основы выборки и поэтому не может обратиться ко всем тем объектам, номера которых выданы с помощью датчика случайных чисел. Следствием этого служит то, что используемая социологом выборка или является результатом лишь некоторого моделирования случайности, либо вообще не является случайной. Это надо иметь в виду, пользуясь результатами математической статистики.

1.2. Понятие случайной величины.

В курсе по теории вероятностей обычно вводится понятие случайной величины. Но, к сожалению, довольно типичным является следующий диалог студентов и преподавателя, начинающего читать курс математической статистики после того, как студенты освоили курс теории вероятностей.

- Что такое случайная величина? Приведите примеры случайных величин из социологической практики.

Молчание.

- Допустим, что мы провели анкетный опрос, собрали данные. Присутствуют ли в этих данных хотя бы в каком-то виде случайные величины?
- Ну, например, мы подсчитали, что у нас 20 процентов мужчин. Это и есть случайная величина.

Ответы студентов говорят о том, что у них зачастую складывается совершенно неверное представление о том, что такое случайная величина. Доля мужчин станет случайной величиной только в том случае, если единицей наблюдения у нас будет, скажем, вуз, и мы тем или иным способом будем подсчитывать, каков состав студентов каждого рассматриваемого вуза по полу. Скажем, в каком-то социологическом вузе – 20% юношей, в другом – 18%, в некоем техническом вузе – 83% и т.д. Здесь доля мужчин – случайная величина, 20% - одно из ее значений. А при анкетном опросе в качестве случайной величины может выступать, например, возраст респондентов. Единица наблюдения – человек. 20 лет, 35 лет, 16 лет – это значения нашей случайной величины. Далее мы увидим, что единицей наблюдения может быть даже выборка: для каждой такой единицы мы можем, например, вычислять среднее арифметическое значение возраста попавших в выборку респондентов. Средний возраст здесь – это случайная величина, среднее значение возраста для конкретной выборки – конкретное значение этой величины.

Подчеркнем, что случайная величина всегда задается некоторым распределением вероятностей встречаемости ее значений; далее вместо столь длинного оборота будем говорить либо просто о распределении, либо о распределении вероятностей, либо о распределении случайной величины. Задать случайную величину – значит задать отвечающее ей распределение; задать некоторое распределение – значит задать некоторую случайную величину.

Мы будет рассматривать в основном числовые случайные величины, т.е. такие, значениями которых служат числа (хотя для социолога огромное значение имеют нечисловые случайные величины, статистика которых разработана весьма слабо[3]). Случайная величина может быть

---

[2] Список литературы по выборке дан в Приложении 3.

[3] Мы получаем значения нечисловой случайной величины, когда, скажем, каждому респонденту приписываем данную им ранжировку каких-либо объектов, вершину графа (теория графов используется при изучении малых групп) и т.д. О статистическом анализе таких случайных величин можно прочитать, например, в работе: Орлов А.И. Эконометрика. М.: Экзамен, 2002. С.229-301.

одномерной, многомерной[4]; непрерывной (когда в принципе ее значением может быть любая точка числовой оси) и дискретной (когда она принимает счетное, чаще всего – конечное число значений).

Важно отметить, что понятия вероятности, распределения вероятностей и, соответственно, случайной величины, сопрягаются с генеральной совокупностью. Изучая выборку, мы имеем дело с выборочными оценками вероятностей и их распределений (в качестве таковых обычно фигурируют относительные частоты встречаемости соответствующих событий и частотные распределения), выборочными реализациями значений случайной величины (выступающих перед нами в виде значений некоторых признаков).

Подчеркнем также важность выделения двух видов случайных событий, задействованных при рассмотрении случайных величин в социологии.

Сама случайная величина определена на множестве случайных событий, имеющих определенные вероятности. В качестве такого события для социолога, как правило, выступает выбор того или иного респондента (конечно, вместо респондентов могут фигурировать и другие объекты – разного рода малые и большие социальные группы, регионы и т.д.). Ясно, что вероятность встречаемости подобных событий связана с тем, каков способ построения выборки.

Другой вид интересующих нас случайных событий – это события, состоящие в том, что те или иные случайные величины принимают те или иные значения. Другими словами, мы говорим о распределениях случайных величин. Выбрав того или иного респондента, мы можем определить соответствующее значение нашей случайной величины. Например, можем определить, что возраст выбранного респондента равен 23 годам. И мы относим возраст к категории случайных величин только в том случае, если можно говорить о распределении вероятностей встречаемости разных значений возраста (хотя это распределение может и не быть известным нам заранее).

Как известно, каждое распределение характеризуется определенным набором своих параметров. Наиболее популярные из них – меры отвечающих распределению средних тенденций (математическое ожидание, мода, медиана) и меры разброса значений случайной величины (например, дисперсия, среднее квадратическое отклонение, абсолютный размах). Изучением такого рода параметров мы и будем в основном заниматься.

1.3. Понятие статистической закономерности.

Статистической закономерностью обычно называют закономерность, характеризующая совокупность изучаемых объектов в целом, как систему. Чаще всего – это закономерность, говорящая об изучаемой совокупности «в среднем».

Для того, чтобы глубже понять, что именно здесь имеется в виду, совершим небольшой исторический экскурс.

Само представление о статистических закономерностях (и, соответственно, о статистических методах, о статистическом подходе) зародилось в XVII веке, когда родилось то направление в обществознании, которое впоследствии было названо политической арифметикой[5]. О статистических приемах изучения общества стали говорить в тех ситуациях, когда цель исследования заключалась «не в исследовании качественных признаков отдельного явления, а в определении количества явлений с известными качествами. … Каждый знает, что дети и старики подвергаются большей опасности умереть, чем люди в средних возрастах; мы получили это сведение из векового жизненного опыта; но лишь по переводе в числа оно приобретает в наших глазах полную убедительность, возвышается до степени общественного закона. Если нам покажут, что в Европейской России в среднем выводе за десятилетие 1874-1884, в течение первого года жизни из 1000 родившихся умирало 305 человек, в возрасте от 10 до 15 лет – только 6 человек, а в возрасте от 75 до 80 лет 130 человек на 1000, живущих этого возраста, то наше представление о распределении смертности по возрастам приобретет совершенно точный вид. Таким образом, систематическое изучение общества может состоять с одной стороны в качественном наблюдении отдельных явлений, с другой стороны в количественном

---

[4] Вообще говоря, многомерные случайные величины не являются числовыми, поскольку здесь речь идет о приписывании каждому респонденту (или любому другому измеряемому объекту) не числа, а набора чисел. Нечисловыми можно считать и такие случайные величины, значения которых получены, к примеру, по номинальной шкале, поскольку соответствующие числа не являются действительными числами в общепринятом смысле этого термина. Однако мы иногда будем рассматривать первые очень часто - вторые, наряду с «истинно» числовыми случайными величинами. Надеемся, что это не приведет к сумятице в сознании читателя.

[5] См., например: Птуха М.В. Очерки по истории статистики XYII - XYIII вв. М., 1945

наблюдении обширный масс явлений. Этот последний прием изучения и носит название *статистического*».[6]

Представляется небезынтересным отметить, что первыми стали пользоваться статистическими приемами именно обществоведы, а отнюдь не естествоиспытатели, как иногда пишется в ориентированной на социолога литературе

Приведем цитату из работы А.А.Чупрова:[7]. «В известных условиях массовый итог являет закономерность, постижимую для нас и без того, чтобы была необходимость знать в точности ход всех единичных процессов, которые к нему приводят. …Статистические формы знания … зародились в XVII столетии. Однако их применение долгое время ограничивалось исследованием явлений социальной жизни. … Потребовалось добрых два века, прежде чем они были осознаны во всей своей общепримени мости…. Статистическая точка зрения знаменует собой отказ от того прослеживания единичных событий, которое рисуется уму естествоиспытателя как идеал полноты и совершенства знания».[8]

Определение статистического подхода как подхода, позволяющего изучать рассматриваемую совокупность объектов «в среднем» господствовало в литературе примерно до второй половины XIX века. Не потеряло оно своего смысла и сейчас. Но мы не можем им ограничиться. В процессе институциализации математической статистики это определение претерпело изменение (уточнение).

Когда говорят об использовании математико-статистических приемов, представление о статистической закономерности обычно связывают с предположением о вероятностном порождении данных: предполагается, что все наши признаки – это выборочные представления случайных величин, каждое выборочное значение какого-либо признака – это реализация одного из значений случайной величины, и такая реализация имеет определенную вероятность. Поиск любой статистической закономерности сводится к поиску значений совокупности параметров распределений каких-либо случайных величин (одномерных, двумерных, многомерных). Подчеркнем, что сказанное означает, что само понятие закономерности мы в таком случае связываем не с выборкой, а с генеральной совокупностью.

Так, казалось бы, простейшими примерами статистических закономерностей, характеризующих студентов какого-либо вуза, мы можем считать утверждения вида: «20% студентов вуза – юноши»; «средняя успеваемость студентов – 6,7 баллов»; коэффициент корреляции между успеваемостью студента на первом и на пятом курсе равен 0,8 и т.д. Однако, в соответствии со сказанным, мы имеем право расценивать эти соотношения как статистические закономерности только в том случае, если «переведем» их на «язык» генеральной совокупности. К примеру, говоря о среднем арифметическом значении какого-либо признака для выборки, мы полагаем, что закономерность будет найдена только в том случае, если мы сумеем на базе выборочного среднего арифметического сделать какие-то выводы о генеральном среднем. Например, мы можем полагать, что найденное выборочное среднее само по себе является хорошей оценкой генерального. Но обычно такого рода утверждения являются не очень корректными. Оказывается, что можно на основе выборочного среднего по определенным правилам сформировать некое более адекватное (вероятностное) представление о генеральном. Собственно, такого рода формирование и является основной задачей математической статистики, о чем мы подробно будет говорить в следующих разделах.

Поскольку все интересующие нас статистические закономерности мы связали с поиском параметров распределений случайных величин в генеральной совокупности, то по сути дела само понятие генеральной совокупности мы связали с существованием, осмысленностью тех случайных величин, которые «стоят» за нашими наблюдаемыми признаками.

---

[6] Чупров А.И. Статистика. Лекции. СП-б: СПб политехнический институт, 1907. С. 6-7. Статистические данные взяты автором из работы: Янсон Ю.Э. Сравнительная статистика населения.

[7] А.А.Чупров. Вопросы статистики. М.: Госстатиздат, 1960, с. 143. Любой человек, хотя бы в какой-то мере изучавший основные приемы анализа данных, знает термин «коэффициент Чупрова». Это – один из самых используемых (мы говорим о современной *мировой* науке и практике) коэффициентов связи между двумя номинальными признаками. Но при этом совершенно забыто то, что А.А.Чупров был известным ученым, органично сочетавшим в своей работе знания социолога и математика (и имевшим два соответствующих высших образования), получившим математические результаты, позволяющие адаптировать понятие вероятности для социальных исследований, и написавшим огромное количество методологических работ, не потерявших своего значения и в наше время. К некоторым его идеям мы обратимся в главе 12.

[8] Отметим, что многие социологи зачастую из-за незнания истории используют несостоятельные аргументы, пытаясь доказать, что математика вообще и математико-статистические методы, в частности, - «инородное тело» для «истинного» социолога.

Для социолога очень важно то, что выполнение предположения о вероятностном порождении исходных данных при решении социологических задач далеко не всегда бывает очевидным. Здесь хотелось бы выделить две основные причины такой неочевидности (обе связаны с возможными сомнениями в существовании «генеральных» случайных величин).

Во-первых, нередко у исследователя имеются сомнения в том, что он имеет дело с выборкой из какой бы то ни было генеральной совокупности (и, соответственно, с выборочными реализациями значений какой-то случайной величины). Изучаем, скажем, 100 студентов, и у нас нет никаких оснований считать их частью какой-то генеральной совокупности, обобщать соответствующим образом результаты; все выводы считаем справедливыми только для этих 100 человек. В таком случае, естественно, сомнительным становится и использование положений математической статистики. Подобные ситуации были учтены при разработке ряда методов анализа данных. Существуют такие методы, которые заведомо не предполагают вероятностного порождения данных[9]. И мы не можем их сбрасывать со счета даже тогда, когда говорим о математической статистике. Дело в том, что одна и та же (с содержательной точки зрения) социологическая задача может решаться по-разному в зависимости от того, что думает исследователь по поводу модели порождения имеющихся в его распоряжении данных. Мы должны сознательно выбрать тот или иной подход (в данном случае речь идет о выборе математико-статистического подхода или отказа от него).[10] И не говорить об этом нельзя. К этому вопросу мы вернемся при изложении темы 12 (в п. 12.3, где идет речь об эксперименте по Миллю).

Во-вторых, мы можем, не сомневаясь в существовании генеральной совокупности, сомневаться в объективности нашего знания о том, как соотносятся наши наблюдаемые признаки и генеральные случайные величины. Так, к примеру, мы можем, опираясь на расчет средней выборочной зарплаты составляющих выборку респондентов, использовать мощный аппарат математической статистики и находить интервал, в который с определенной вероятностью попадает генеральное математическое ожидание рассматриваемого признака. А в действительности в генеральной совокупности существует, скажем, два распределения: одно для малооплачиваемых, нормальное со средним в 5000 рублей, а другое – для высоко оплачиваемых, тоже нормальное со средним 50000 рублей. Другими словами, в нашей генеральной совокупности существует не одна, а две случайные величины, и с каждой из них надо работать отдельно (отдельно осуществлять все требующиеся оценки). Математическая статистика может помочь «разделить» такую «смесь», но очень трудно заранее догадаться о том, что это *надо* делать.

Отметим, что здесь проблема существования случайной величины переплетается с проблемой однородности генеральной совокупности (о проблеме однородности мы будем также говорить в п. 1.7): под однородной совокупностью нередко понимают такую, на которой задана содержательно интерпретируемая нормально распределенная случайная переменная. [11]

Проблема однородности в социологии иногда бывает очень сложной.[12] Особенно тонкие и важные для нашей темы моменты возникают в связи с осмыслением понятия вероятности. Адекватный поиск статистических закономерностей (понимание которых не отделимо от понимания вероятности) предполагает умение исследователя различать две ситуации: (1) когда изменение относительной частоты изучаемого явления обусловлено действием случайных по отношению к этому явлению факторов и поэтому может быть нейтрализовано действием закона больших чисел (этот закон будет сформулирован в п. 4.2) и (2) когда то же изменение возникло из-за изменения того комплекса условий, который входит в само определение вероятности; в таком случае закон больших чисел не при

---

[9] О сходстве и различии подходов математической статистики и анализа данных к поиску статистических закономерностей см.: Толстова Ю.Н. Анализ социологических данных. Методология, дескриптивная статистика, анализ связей номинальных признаков. М.: Научный мир, 2000.

[10] Наверное, можно сказать, что рассмотрение подхода, не являющегося математико-статистическим, нужно хотя бы для того, чтобы более ярко оттенить возможности математической статистики.

[11] Так, предложенные Терстоуном измерительные модели предполагают, что мнение каждого респондента о некотором объекте представляет собой нормальное распределение (будучи спрошенным в разные моменты времени, респондент, вообще говоря, будет давать разные ответы, чаще всего – близкие к некоторому математическому ожиданию, и тем менее вероятные, чем далее они отстоят от последнего) и что усреднять оценки, данные респондентами некоторой совокупности этому объекту, можно только в том случае, если совокупность однородна, т.е. если такие нормальные распределения одинаковы для всех респондентов. Это может быть проинтерпретировано как существование единой случайной величины для всех наших респондентов (Толстова Ю.Н. Измерение в социологии. М.: Инфра-М, 1998). К проблеме однородности мы неоднократно будем возвращаться.

[12] Подробнее об этом см.: Толстова Ю.Н. Логика математического анализа социологических данных. М.: Наука, 1991.

чем, мы имеем дело с разными статистическими закономерностями[13]. К этому мы еще вернемся (п.12.3).

Иногда говорят о том, что статистическая закономерность как бы отвечает некой необходимости, «пробивающей себе дорогу» через массу случайностей (в том же смысле обычно говорят о наличии средней тенденции). Например, если коэффициент корреляции близок к единице, то можно говорить, что между признаками «в среднем» имеется линейная зависимость. В частности, с ростом значений одного признака «в среднем» растут значения другого. Но только «в среднем». В этом процессе могут быть «сбои»[14]. Ясно, что говорить о таком понимании статистической зависимости более целесообразно в том случае, когда речь идет о «средней» ситуации для разных выборок: взяли одну выборку – одни точки признакового пространства отклоняются от прямой линии, взяли другую выборку – другие, а «в среднем» все же большинство точек плотным облаком охватывают прямую (надеемся, читатель имеет представление о том, какова сущность коэффициента корреляции и какая прямая линия имеется в виду; мы еще вспомним об этом в конце данной темы и в п. 13.1).

Конечно, о какой-то средней «тенденции» можно говорить и в случае, когда нам кажется неадекватной реальности гипотеза о вероятностном порождении исходных данных. Однако обнаружение такой «тенденции» вряд ли можно считать нахождением научно осмысленной закономерности. Пусть, например, мы опросили какое-то количество (например, 100 человек) мужчин – студентов московских вузов, подсчитали их среднюю успеваемость (4,3 балла) и вычислили формально по известной формуле значение коэффициента корреляции между какими-то двумя переменными (0,9). Предположим также, что у нас нет оснований считать, что наши респонденты являются выборкой из некоторой генеральной совокупности и, соответственно, что значения любого из наших признаков – это реализации некоторой случайной величины). Тогда мы не имеем права хотя бы как-то обобщать эти результаты ни на московских студентов вообще, ни на студентов-мужчин, ни на какую-либо другую совокупность людей. Вполне может случиться так, что если мы добавим к этой совокупности еще 50 юношей-студентов московских вузов, то получим совсем другие цифры (скажем, среднюю успеваемость – 2.3 балла, коэффициент корреляции между теми же переменными – 0,1). И мы даже не можем сказать, какова вероятность такой метаморфозы. [15]

Мы вернемся к рассмотрению понятия статистической закономерности в п. 12.1, где нас будет интересовать его соотнесение с понятием причинно-следственной связи.

1.4. Объект изучения для математической статистики

Основным *объектом* изучения для математической статистики являются случайные величины. Эта наука изучает различные распределения (а, как мы уже отмечали, задать случайную величину и задать распределение вероятностей – это одно и то же), их выборочные представления, соотнесение одних с другими.

---

[13] Напомним, что вероятность определяется как «числовая характеристика степени возможности появления какого-либо определенного события в тех или иных определенных, могущих повторяться неограниченное число раз, условиях» (Колмогоров А.Н. Вероятность // Вероятность и математическая статистика. Энциклопедия. М.: БРЭ, 1999. С. 96). Вопрос о виде упомянутых условий и даже о самом их существовании для социологических задач нередко является проблематичным.

[14] О таком понимании статистических закономерностей и о том, как на базе соответствующих наблюдений рождались основные положения математической статистики (в частности, коэффициент корреляции), много говорится в работах А.А.Чупрова (например, из числа переизданных в книге: Чупров А.А. Основы статистики. М.: Госстатиздат ЦСУ СССР, 1960). О Чупрове мы еще будем говорить далее в основном тексте.

[15] Для дальнейшего нам важно отметить, что в описанных «некондиционных» условиях мы не только не можем считать, что за нашим вроде бы непрерывным признаком (например, за средней успеваемостью студента) стоит некоторая непрерывная случайная величина, но и вообще не имеем оснований полагать, что наш признак непрерывен. У нас имеется несколько его значений и остается неизвестным, имеют ли смысл остальные гипотетически мыслимые значения. Дело в том, что в социологии очень часто наблюдаемые признаки служат признаками-приборами, значения которых интересуют социолога только постольку, поскольку отражают какие-то латентные свойства изучаемых объектов (респондентов). Не все возможные значения признака могут отражать эти латентные свойства, некоторые свойства могут отражаться группами значений и т.д. К этому вопросу мы вернемся (см., например, конец п. 7.3).

1.5. Предмет изучения для математической статистики

Как было сказано, случайная величина отождествляется с определяющим ее значения распределением вероятностей, а поиск всех изучаемых математической статистикой закономерностей сводится к вычислению значений той или иной совокупности параметров распределений каких-либо случайных величин. Поэтому можно сказать, что *предметом* изучения для математической статистики являются параметры распределений случайных величин. Здесь, наверное, уместно сказать, что судить об этих параметрах исследователь может только на основе выборочных данных

1.6. *Основная задача математической статистики*

Основная задача – изучение проявления статистической закономерности на выборке и перенос результатов с выборки на генеральную совокупность. Перенос осуществляется на вероятностном языке. Существует два способа переноса (отвечающих двум мощным направлениям математической статистики) – статистическое оценивание параметров (этот подход, в свою очередь делится на точечное и интервальное оценивание) и проверка статистических гипотез. Все это подробно будет рассмотрено в следующих темах.

Ниже приводится таблица соотнесения основных понятий генеральной совокупности и выборки. В данном случае эта пара терминов в определенном плане синонимична паре «математическая статистика и эмпирическая социология». В соответствии с традицией, понятия, отвечающие генеральной совокупности, обозначаются преимущественно греческими буквами; выборочные их представления - созвучными латинскими буквами.

Генеральная совокупность	***Выборочная совокупность***
Случайная величина (ξ, η, ζ)	Признак (x, y, z)
Вероятность π, $p^{ген}$	Относительная частота p, $p^{выб}$
Распределение вероятностей	Частотное распределение
Параметр (характеристика вероятностного распределения)	Статистика (функция от выборочных значений признаков), служит для оценки того или иного параметра генерального вероятностного распределения
Примеры параметров и отвечающих им статистик	
Одномерные случайные величины (одномерные распределения)	
Математическое ожидание (μ, Mξ)	Среднее арифметическое (m, $\overline{X}$)
Мода (Mo)	Мода (Mo)
Медиана (Me)	Медиана (Me)
Среднее квадратическое отклонение (σ)	Среднее квадратическое отклонение (s)
Дисперсия (σ^2, Dξ)	Дисперсия (s^2, Dx)
Двумерные случайные величины (двумерные распределения)	
Коэффициент корреляции $\rho(\xi, \eta)$	Коэффициент корреляции r (x, y)
Многомерные случайные величины (многомерные распределения)	
Коэффициенты уравнения регрессии $\beta_1,\beta_2,\ldots,\beta_n$	Коэффициенты уравнения регрессии $b_1, b_2, \ldots, b_n$

Таблица 1.1. Соотнесение понятий, отвечающих генеральной и выборочной совокупностям.

Ниже мы (в соответствии со сложившимися обычаями) не будем очень строго выдерживать эти обозначения. Так, для обозначения случайных величин чаще будем иногда использовать латинские буквы, а не греческие. Очень строго будем придерживаться лишь обозначений для среднего квадратического отклонения (дисперсии)

1.7. Методологические принципы использования математики в социологии

Поскольку математическая статистика – ветвь математики, то, используя ее достижения в социологии, нельзя забывать об основных методологических принципах использования в социологии математических методов. Поясним подробнее, о чем идет речь.

Любой математический метод предполагает адекватной реальности определенную *модель того явления, которое с помощью этого метода изучается* (заметим, что это касается не только такой ситуации, когда мы пытаемся моделировать реальность с помощью математических методов, но и любого научного исследования вообще: любая наука имеет дело с моделью, и для успешности научных изысканий весьма часто требуется осознание того, какая модель используется исследователем). Конечно, об этом надо думать при использовании математики в любой отрасли знания. Но если в естественных и технических науках мы, применяя тот или иной математический метод, можем не задумываться о том, какая именно модель в нем заложена, то для социологии вопрос о выборе такой модели стоит довольно остро. Объясняется это в первую очередь тем, что наука пока не предложила методов, полностью адекватных большинству социологических ситуаций.

Поясним сказанное примерами. Применяя математику, скажем, в строительстве, мы рассчитываем нагрузку на некоторую балку, используя сложные формулы. При этом мы можем совершенно не помнить, как эта формула получена. Правильно рассчитаем – дом будет стоять. А в социологии не так. Например, одних только коэффициентов, измеряющих связь между двумя признаками – более сотни. Все они изменяются от 0 до 1 (или от -1 до +1). Значение «1» говорит о сильной связи, «0» – об отсутствии оной. Но каждый коэффициент по-своему «понимает» связь. Разные коэффициенты принимают значение «1» в разных ситуациях. Что же делать социологу, если один из коэффициентов равен 0,9, а другой – 0,2?

Описанный факт не уникален. Имеет место следующее обстоятельство: если для решения какой-то социологической задачи существует некий математический метод, то, как правило, он *не единствен*. Одной из основных трудностей использования математики в социологии является выбор метода, сравнение методов друг с другом и т.д. Однако на этом проблемы не кончаются. Назовем, по крайней мере, еще две.

Во-первых, каждый математический метод требует определенной *однородности* изучаемой с его помощью совокупности объектов. Один из самых очевидных примеров: среднее арифметическое значение какого-либо признака бессмысленно считать для такой совокупности, в которой разброс значений этого признака велик (скажем, для современной России бессмысленным является среднее значение зарплаты). Как мы уже упоминали в сноске 9, при изучении статистических закономерностей однородность совокупности нередко понимается как существование некоторой определенной для всех элементов совокупности случайной величины (или нескольких величин). А все интересующие нас закономерности, как было отмечено выше, – это параметры таких величин. Значит, однородность по существу должна отождествляться с осмысленностью для изучаемой совокупности выявляемой статистической закономерности.

Во-вторых, процесс применения математического аппарата в социологии, как правило, не может быть сведен непосредственно к выбору и реализации того или иного алгоритма анализа некой информации. В силу сложности проблемы концептуализации предмета исследования (в частности, вычленения и операционализации понятий, обеспечения процесса измерения, выбора модели изучаемой закономерности), неоднозначности толкования человеческих суждений, отсутствия строгой границы между объектом и субъектом исследования и т.д. использование любого математического аппарата «обрастает» огромным количеством проблем, решаемых на самых разных этапах социологического исследования.

Опираясь на сказанное, следующим образом сформулируем *основные методологические принципы* применения математики в социологии: (1) соотнесение модели, заложенной в методе, с содержанием решаемой с его помощью социологической задачи; (2) обеспечение однородности изучаемой совокупности объектов; связь представления об однородности с содержанием задачи; (3) обеспечение органической связи всех этапов исследования друг с другом, особенно этапа измерения и этапа анализа; (4) комплексное использование разных математических методов: последовательное (когда разные методы используются на разных этапах и, чаще всего «выход» одного метода служит «входом» для другого) и параллельное, когда разные методы используются для решения одной и той же задачи и исследователь должен сравнивать получающиеся результаты на базе сравнения заложенных в методах моделей. Подчеркнем, что мы сформулировали лишь самые основные принципы, которые в первую очередь будут учитываться нами при обсуждении проблем, связанных с обеспечением корректности применения в социологии теории вероятностей и математической статистики.

1.8. Некоторые замечания о терминах, использующихся в западной литературе.

Терминология, использующаяся в отечественной и западной литературе интересующего нас плана (мы имеем в виду в первую очередь американские учебники по т.н. статистике) в определенной мере различна[16]. И за этим различием зачастую стоят разные методологические позиции. Коснемся подобной ситуации, имеющей место для некоторых терминов, рассмотренных выше. Представляется, что анализ соответствующих методологических аспектов имеет непосредственное отношение к практике использования социологом достижений математической статистики.

Итак, мы разделил все интересующие нас статистические показатели на две большие группы – параметры распределений, т.е. как бы «истинные», генеральные характеристики случайных величин, и статистики – выборочные оценки «истинных» параметров. Процедура поиска параметров может иметь весьма разную степень сложности. Одни параметры просто говорят об описании совокупности. Это, скажем, математическое ожидание и дисперсия. Поиск их сравнительно прост. Другие – позволяют изучать причинно-следственные отношения. Это, например, коэффициент корреляции. Его найти уже сложнее, соответствующая процедура включает в себя, в частности, расчет математических ожиданий и средних квадратических отклонений. Третьи параметры дают возможность прогнозировать ситуацию. Это делает, например, регрессионный анализ (коэффициенты уравнения регрессии – это тоже набор параметров, характеризующих распределение случайной величины). Построение уравнения регрессии – еще более сложная процедура, включающая в себя, в частности, расчет разного рода коэффициентов корреляции. Выбор рассчитываемых параметров определяется решаемой содержательной задачей. И прежде всего здесь можно указать на три огромные класса задач, обычно выделяемые в методологии науки как основные, отвечающие целям любой наукой – описание, объяснение, предсказание. Первую задачу обычно связывают с термином «описательная (дескриптивная) статистика», называя таким образом и совокупность простейших параметров распределения, таких, как математическое ожидание и дисперсия (или их выборочные оценки), и ветвь статистики (как науки), позволяющую эти характеристики найти. Наверное, можно было бы вводить термины «объяснительная статистика», «предсказательная статистика» и т.д. Но этого не делается, поскольку слишком много методов могут «объяснять» или «предсказывать» изучаемые явления, слишком трудно провести границу между теми и другими задачами.

Подчеркнем, что всегда, какие бы параметры мы ни рассчитывали (и, соответственно, какую бы задачу ни решали), всегда мы сначала будем находить соответствующие выборочные статистики, а потом «соображать», с какой вероятностью в генеральной совокупности будет иметь место та или иная ситуация с параметрами изучаемых распределений. И всегда за нашими действиями будет стоять определенная модель, в наличии которой мы должны давать себе отчет. Модель стоит даже за самыми простейшими параметрами. Так, выбирая для оценки средней тенденции математическое ожидание, моду или медиану, мы по-разному понимаем смысл этой самой средней тенденции (а ведь бывают и другие меры средней тенденции – например, среднее геометрическое, среднее гармоническое, да и квантили тоже можно считать своеобразными средними). То же можно сказать о разных мерах разброса, коих тоже немало[17].

В западной же литературе вся статистика обычно делится на описательную (descriptive) и «выводимую» (inferential).

Descriptive statistics consists of the collections, organizations, summarizations, and presentations of data[18]. В дескриптивной статистике исследователь пытается описать ситуацию, определенным образом собирая данные, рассчитывая определенные средние и проценты и представляя собранную информацию в наглядном виде, используя диаграммы, графики, таблицы. Типичным примером использования описательной статистики является перепись населения.

Трудно возразить против описанного термина, если не попытаться уточнить его смысл путем сопоставления понятия «дескриптивная статистика» с понятием, ему противопоставляемым, - «выводимой статистикой».

«Inferential statistics consists of generalizing from samples to population, performing hypothesis tests, determining relationships among variables, and making predictions»[19]. Работая в рамках выводимой

---

[16] Под различием терминологии мы имеем в виду не то, что для обозначения одного и того же понятия можно использовать разные (с точностью до перевода) слова. Это непринципиально (как говорится, назови хоть горшком, да не сажай в печку). Речь идет о том, какие именно понятия «удостаиваются» введения отдельных терминов и о том, какую связь между этими понятиями терминология отражает. Различие таких аспектов может быть весьма принципиальным.

[17] О содержательном смысле некоторых мер средней тенденции и разброса говорится в: Толстова Ю.Н. Анализ социологических данных: методология, дескриптивная статистика, изучение связей номинальных признаков. М.: Научный мир, 2000.

[18] Bluman A.G. Elementary statistics. A step by step. McGraw-Hill Companies. 1992, 1995, 1998, 2001. C.5.

[19] Там же, с.7.

статистики, исследователь пытается перенести результаты с выборки (sample) на генеральную совокупность (population) путем проверки статистических гипотез. Но этим цели выводимой статистики (как науки) не ограничиваются. В нее еще входит изучение соотношений между переменными и осуществление предсказаний.

На наш взгляд, некорректно объединять в одну группу проверку статистических гипотез и, скажем, изучение связей между переменными. Остается абсолютно неясным основание выделения такой группы. Сомнительным выглядит и противопоставление этой группы методам дескриптивной статистики. Опишем причины наших сомнений.

Во-первых, проверка статистических гипотез требуется абсолютно для всех методов – и для описательной статистики, и для разного рода коэффициентов связи, и для прогнозных алгоритмов..

Во-вторых, четкой границы между методами описательной статистики и методами изучения отношений переменных нет. Так, коэффициенты корреляции нередко используются в чисто описательных целях.

В третьих, методы описательной статистики не менее «выводимы», чем методы изучения связей между переменными и прогнозирования. Так, собирая данные, мы используем иногда совсем не очевидные модели: именно так, а не иначе измеряем переменные (а в каждом методе измерения заложена своя модель); разбиваем диапазон изменения признака на интервалы (это – в значительной мере субъективная процедура, а от нее зависит и характер получаемого описания, и величина коэффициентов связи и т.д.); выбираем то или иное понимание средней тенденции или разброса и т.д.

Более адекватной представляется нам терминология, которую мы ввели выше и которая соответствует традициям отечественной школы.

Отметим, что описанная выше точка зрения отвечает взглядам и современных русских ученых, и российских исследователей начала 20-го века. Так, известный математик-социолог А.А.Чупров называл «вероятностным априори» то, для обозначения чего мы используем словосочетание «генеральные параметры», и говорил о необходимости отчетливого и выдержанного разграничения априорных искомых и эмпирических данных статистического исследования. В частности, он полагал, что именно смешение указанных объектов изучения приводят к сбивчивости в трудах представителей известной английской статистической школы, возглавляемой К.Пирсоном. Сбивчивость, по мнению Чупрова, затрудняла усвоение работ этой школы (которую Чупров ценил очень высоко; результаты, полученные Пирсоном и окружающими его исследователями, он активно пропагандировал в своих работах).

А.А.Чупров был воспитан на трудах крупнейших русских математиков – П.Л.Чебышева, А.М.Ляпунова, А.А.Маркова, занимающих лидирующие позиции в мировой математической статистике второй половины 19-го – начала 20-го века.

Повторение отдельных фрагментов курса по теории вероятностей

1. Функция плотности одномерного распределения и функция распределения для одномерных непрерывных и дискретных распределений; связь этих функций друг с другом; основные параметры любого[20] одномерного распределения - математическое ожидание, мода, медиана и другие квантили, дисперсия.
2. Площадь под кривой функции плотности как оценка вероятности попадания значения случайной величины в соответствующий отрезок.
3. Выборочное представление функции плотности распределения признака (непрерывного и дискретного): частотная таблица, полигон, гистограмма (в том числе с неравными интервалами). Различие стоящих за выбором полигона и гистограммы предположений о распределении признака внутри каждого интервала. Анализ моделей, заложенных в указанных способах выборочного представления случайной величины, их различие: при использовании полигона предполагаем, что все попавшие в интервал значения сосредоточены в одной точке (важно, что при построении графика на оси x может быть выбрана любая точка интервала, т.е. что выбор такой точки – тоже модельное предположение); при использовании гистограммы считаем, что распределение в каждом интервале равномерно; гистограмму имеет смысл рассчитывать только для непрерывного признака. Проблема построения выборочной функции плотности для непрерывного признака: разбиение диапазона изменения признака на интервалы, отнесение «стыка» соседних интервалов к одному из концов, пропущенные данные. Цели заполнения пропусков. Способы такого заполнения: средним арифметическим (может быть, с учетом значений других признаков) или

[20] На самом деле – не совсем любого, поскольку существуют распределения, не имеющие конечных моментов (напомним, что математическое ожидание – это первый момент распределения, дисперсия – второй и т.д.).

другими средними (с учетом шкал, о шкалах пойдет речь в лекции 2), равномерно по всем градациям, пропорционально получившимся частотам. Модели, стоящие за каждым названным подходам к заполнению пропущенных значений.

4. Выборочное представление функции распределения (кумулята): частотная таблица, полигон и гистограмма.
5. Статистики, отвечающие основным параметрам одномерного распределения: среднее арифметическое, дисперсия, мода, медиана и другие квантили. Медиану необходимо уметь считать двумя способами: как середину вариационного ряда и с помощью кумуляты. То же для других квантилей. Снова обратить внимание на модель, заложенную в методе.

Напомним основные формулы для расчета медианы и моды[21].

$$Me = x_0 + \delta \frac{\frac{1}{2}n - n_H}{n_{Me}},$$

где x_0 – начало (нижняя граница) медианного интервала; δ - величина медианного интервала; n - объем выборки (или 100%, либо 1); n_H – частота (или относительная частота в процентах, либо в долях), накопленная до медианного интервала; n_{Me} – частота (или относительная частота в процентах, либо в долях) медианного интервала.

$$Mo = x_0 + \delta \frac{n_{Mo} - n^-}{2n_{Mo} - n^- - n^+},$$

где x_0 – начало (нижняя граница) модального интервала; δ - величина модального интервала; n_{Mo} – частота модального интервала; n^- - частота интервала, предшествующего модальному; n^+ - частота интервала, следующего за модальным. Частоты, как и выше, везде могут быть заменены на относительные частоты, выраженные либо в процентах, либо в долях.

6. Функция плотности и функция распределения двумерных случайных величин. Основной параметр двумерного распределения – коэффициент корреляции.
7. Выборочное представление функции плотности двумерной случайной величины (частотная таблица, или таблица сопряженности). Маргинальные частоты, их связь с одномерными распределениями рассматриваемых признаков. Статистика, отвечающая генеральному коэффициенту корреляции.

Напомним формулу для вычисления последней названной статистики.

$$r = \sum_i \frac{(X_i - \overline{X})(Y_i - \overline{Y})}{n s_x s_y}$$

Кроме того, напомним важное свойство коэффициента корреляции: он измеряет только линейную связь. Это означает, что если он равен 1 или –1, то отвечающие нашим объектам точки рассматриваемого двумерного признакового пространства лежат на прямой линии, т.е. между признаками имеется точная линейная связь (прямая или обратная). А вот если r=0, то это означает не отсутствие связи вообще, а только отсутствие линейной связи. Нелинейная же связь при этом может быть и весьма сильной. Об этом мы будем говорить подробнее при обсуждении темы 13 (посвященной корреляционному отношению – коэффициенту, позволяющему измерить нелинейную связь).

8. Понятие случайной выборки. Ее построение с помощью таблицы случайных чисел.

Примеры задач.

1. Придумать пример, демонстрирующий, что при разных разбиениях диапазона изменения непрерывного признака на интервалы можно получить качественно разные полигоны

[21] Рабочая книга социолога, с.161-162

распределения – выборочные представлений функции плотности (разнокачественность распределений связать с пониманием описания данных как одной из задач науки). Примеры разнокачественных распределений: одновершинное и двухвершинное, одновершинное и равномерное, равномерное и с «ямой» и т.д.

2. Задана следующая частотная таблица:

Возраст (лет)	15-20	20-50	50-55
Относительная частота	1/3	1/3	1/3

Простроить соответствующую гистограмму (заметим, что представленное в таблице разбиение диапазона изменения возраста на интервалы не лишено смысла; например, такое разбиение может явиться следствием особого внимания исследователя к тем периодам жизни человека, когда он вступает в трудовую жизнь (15-20 лет) и постепенно выходит из нее, готовясь к пенсии (5—55 лет для женщи5н).

3. Описать, какие модели стоит за стандартными формулами расчета моды и медианы.

4. Вспомнить геометрические правила расчета медианы с помощью выборочной функции распределения – кумуляты (в виде полигона). Показать, что эти правила приводят к тому же результату, что и соответствующая формула из п. 5 выше (раздел «*Повторение отдельных фрагментов курса по теории вероятностей*»).

5. Разработать такой геометрический способ расчета моды с помощью выборочной функции плотности распределения (в виде гистограммы), который отвечал бы соответствующей формуле из п.5 выше.

6. Составить формулы (аналогичные формуле для расчета медианы), позволяющие рассчитывать квартили, децили, процентили и другие возможные квантили. Показать, как эти формулы могут быть заменены геометрическими построениями на основе кумуляты.

7. Предположим, что исследователя в первую очередь интересуют те возрастные категории, которые отвечают вхождению человека в работоспособный возраст (15-10 лет) и выходу из него (50-55 лет для женщин). Тогда естественным представляется разбиение диапазона изменения возраста на интервалы, представленные в следующей таблице:

Возрастной интервал	15-20	20-50	50-55
Доля лиц, попавших в интервал	1/3	1/3	1/3

Построить гистограмму, отвечающую отраженным в таблице данным. Обосновать теоретически выбранный способ построения.

8. Рассчитать средние и дисперсию для доли явившихся на голосование жителей некоторого региона, если известны аналогичные доли для каждого из находящихся на территории региона участков. Данные представлены следующей таблицей:

Доля явившихся на голосование	10 – 20	20 – 30	30 – 40
Количество избирательных участков	5	13	28

9. Рассчитать коэффициент корреляции между стажем работника и его зарплатой на основе следующей частотной таблицы

Зарплата (в т.р.)	Стаж (в годах)

	1-5	5-10	10-15	Нет ответа
0,5 – 1,5	40	30	30	0
1,5 – 2,5	2	2	6	10
2,5 – 3,5	10	10	20	0

10. У 12 школьников изучались две характеристики: оценки IQ, определенные с помощью шкалы интеллекта Стенфорда-Бине в шестом классе (X) и успеваемость по химии в средней школе, оцененная на основе теста, состоящего из 35 вопросов (Y). Полученные данные отражены в следующей таблице:

N	1	2	3	4	5	6	7	8	9	10	11	12
X	120	112	110	120	103	126	113	114	106	108	128	109
Y	31	25	19	24	17	28	18	20	16	15	27	19

Рассчитать коэффициент корреляции между X и Y.

---

11. Показать, каким образом связаны выборочные формулы для расчета статистик: среднего арифметического, дисперсии, коэффициента корреляции для непрерывного признака – и известные формулы для расчета (с помощью интегралов) отвечающих этим статистикам генеральных параметров: математического ожидания, дисперсии, коэффициента корреляции.

12. Показать, как выглядит функция плотности равномерного распределения и каким образом из нее с помощью интегрирования можно получить соответствующую функцию распределения. Как последняя выглядит?

13. Осуществить с помощью таблицы случайных чисел выбор 5-ти студентов из группы.

Добавочная литература к теме 1 .

Обязательная

(для повторения материала из курса по теории вероятностей: расчет выборочных статистик, отвечающих известным параметрам генеральных распределений)

Толстова Ю.Н. Анализ социологических данных: методология, дескриптивная статистика, изучение связей между номинальными признаками. М.: Научный мир, 2000

Ниворожкина Л.И., Морозова З.А. Основы статистики с элементами теории вероятностей. Для экономистов. Ростов-на-Дону: Феникс, 1999

Рабочая книга социолога. М.: Наука, 1983

Дополнительная

О методологических принципах использования математики в социологии

Толстова Ю.Н. Методология математического анализа данных // *Толстова Ю.Н.* Социология и математика. М.: Научный мир, 2003. С.80-94. А также: СОЦИС, 1990, №6, с. 77-87.

Проблемы пропущенных данных в массовых опросах

Алгоритмы и программы восстановления зависимостей. - М.: Наука, 1984.

Вапник В.Н. Восстановление зависимостей по эмпирическим данным. - М.: Наука, 1979.

Загоруйко Н.Г. Эмпирическое предсказание. - Новосибирск: Наука, 1979. С. 105-118.

Клюшина Н.А. Причины, вызывающие отказ от ответа // Социс (Социологические исследования). - 1990. - N1. С. 98-105.

Лакутин О.В. Учёт пропущенных данных / Применение математических методов и ЭВМ в социологических исследованиях. - М.: ИСИ АН СССР, 1982. С.86-90.

Лбов Г.С. Методы обработки разнотипных экспериментальных данных. - Новосибирск: Наука, 1981. С. 38-41, 52-55.

Литтл Р.Дж., Рубин Д.Б. Статистический анализ данных с пропусками. - М.: Финансы и статистика, 1991.

Фёдоров И.В. Причины пропуска ответа при анкетном опросе // Социс. - 1982. - N 2.

Проблемы разбиения диапазона изменения признака на интервалы

Орлов А.И. Асимптотика квантований и выбор числа градаций в социологических анкетах / Математические методы и модели в социологии. - М.: ИСИ АН СССР, 1977. С.42-55.

Пасхавер Б. Проблема интервалов в группировках // Вестник статистики. - 1972. - N 6.

Сиськов В.И. Об определении величины интервалов при группировках // Вестник статистики. - 1971. - N 12.

А.А.Чупров. О приемах группировки статистических наблюдений // Известия Санкт-Петербургского политехнического института. 1904. Т. 1. Вып. 1–2.

Doane D.P. Aesthetic frequency classification. American Statistician, 30, 1976. P. 181-183.

Freedman D., Diaconis P. On this histogram as a density estimator: L_2 theory. Zeit. Wahr. Ver. Geb.,57, 1981. P.453-476.

Scott D.W. On optimal and data-based histograms. Biometrika, 66, 1979. P. 605-610.

Scott D.W. Multivariate density estimation: theory, practice, and visualization. N.-Y.: John Wiley & Sons, 1992.

Sturges H. The choice of a class-interval. J.Amer. Statist. Assoc., 21, 1926. P.65-66.

Wand M.P. Data-based choice of histogram bin-width. Technical report, Australian Graduate Scool of management, university of NSW. 1995.

ТЕМА 2.
Общее представление о социологических шкалах.

1. Общие принципы понимания измерения в социологии

Как мы уже отмечали во Введении, социолог не может отвлечься от проблемы измерения используемых им признаков. Напомним основные принципы теории измерений. Используемые ниже сокращения: ЭС – эмпирическая система, МС – математическая система, ЧС – числовая система.

ЭС – это совокупность изучаемых объектов (например, в качестве таковой может служить множество студентов ГУ-ВШЭ), рассматриваемых как носители интересующих исследователя свойств (например, студенты могут интересовать исследователя как носители определенных политических взглядов). МС – это совокупность математических объектов, в которую ЭС отображается при измерении (в частности, ЭС может отображаться в ЧС). Итак, *измерение* – это отображение ЭС в МС:

$$\text{ЭС} \longrightarrow \text{МС}.$$

И отображение это должно быть таким, чтобы у нас была гарантия адекватности «перехода» интересующих нас отношений между эмпирическими объектами в соответствующие им математические отношения между объектами МС. Так, если студенты интересуют исследователя лишь с точки зрения того, в какой степени каждый из них склонен к демократии, а рассматриваемая МС – числовая, то от чисел требуется, чтобы студенту с большей склонностью отвечало бы большее число.

Чаще всего в качестве МС выступает ЧС. Тогда алгоритм, отображающий ЭС в МС, называется *шкалой,* соответствующий процесс измерения – *шкалированием,* а совокупность чисел, в которую мы отобразили элементы ЭС – *шкальными значениями.* Отметим, что нередко в социологии используются и нечисловые МС (например, граф, использующийся при изучении малой группы).

Подчеркнем, что и ЭС, и МС – модели. Элементы ЭС – это не полноценные люди, а определенные «срезы» с них. Скажем, если студенты интересуют социолога как носители политических взглядов, то социологу должно быть безразлично, каков у них цвет волос (но только не в том случае, если вдруг окажется что блондины, скажем, более склонны к демократии, чем брюнеты). Элементы ЧС – не полноценные числа, а тоже лишь некоторые «срезы» с тех чисел, к которым мы привыкли в школе. Например, если в описанной выше ситуации с изучением политических взглядов студентов нас интересует только то, о чем шла речь, т.е. только сравнение студентов друг с другом по степени их склонности к демократии, то для нас будет осмыслен порядок получающихся чисел, но, скажем, соотношение 5-4=3-2 в таком случае не будет иметь смысла.

2.2. Определение номинальной, порядковой, интервальной шкалы

Из-за того, что при шкалировании используются не все обычные свойства чисел, рассматриваемых в качестве шкальных значений, совокупность таких значений оказывается определенной не однозначно. Так, если мы отображаем лишь степень склонности студентов к демократии, то соответствующая эмпирическая упорядоченность с одинаковым успехом отобразится и в числах 2, 18, 19, и в числах 128, 154, 2037.

То преобразование, с точностью до которого определена совокупность шкальных значений каких-либо элементов рассматриваемой ЭС, называется *допустимым преобразованием шкалы.*

Тип шкалы определяется тем, какая совокупность допустимых преобразований этой шкале отвечает. Содержание таблицы 2 позволяет понять, каким образом определяются наиболее распространенные в социологии типы шкал - *номинальная, порядковая, интервальная).*

Тип шкалы	***Отношения между элементами ЭС, отражаемые в соответствующие отношения между элементами ЧС***	***Примеры возможных шкал (наборы "чисел", отражающих одну и ту же информацию)***	***Допустимые преобразования шкалы***
Номинальная шкала	a=b, a ≠b	5, 5, 10, 185, 15 25, 25, 3, 30, 1	Взаимно- однозначные преобразования
Порядковая шкала	a=b, a ≠b, a>b	5, 5, 10, 185, 15 25, 25, 35, 50, 40	Монотонно- возрастающие преобразования
Интервальная шкала	a=b, a ≠b, a>b a – b=c –d, a – b > c - d	5, 5, 10, 185, 15 25, 25, 35, 385, 45	Положительные линейные преобразования Y = αx + β, α > 0

Таблица 2. Определение основных типов шкал

Одна шкала называется *шкалой более высокого типа*, чем другая, если совокупность допустимых преобразований первой шкалы включается в совокупность допустимых преобразований второй. Среди шкал, из таблицы 2 наиболее высокий тип имеет интервальная шкала, наиболее низкий – номинальная. Результаты измерения по шкале более высокого типа больше похожи на числа.

Определение типа шкалы нужно нам не для «красоты». То, по какой шкале получены исходные данные, определяет, каким образом эти данные можно анализировать для получения нового знания. Интуитивно ясно, что далеко не все операции осмыслены для шкал любого типа. Скажем, если мы используем номинальную шкалу для измерения национальности и приписываем первому респонденту, русскому, 1, второму – татарину – 2, третьему, украинцу – 3, четвертому, чукче, - 4, то вряд ли будем считать имеющим смысл выражение: $\frac{1+3}{2}=2$ (среднее арифметическое между русским и украинцем равно татарину).

Как же определить, чтó именно мы имеем право делать с числами, полученными по той или иной шкале? В некоторых книгах ответ на этот вопрос дается путем перечня ряда методов, отвечающих определенным шкалам: для интервальной шкалы мы имеем право считать среднее арифметическое, моду и медиану; для порядковой – моду и медиану; для номинальной – только моду и т.д. Но такой подход вряд ли может считаться удовлетворительным.

Во-первых, для всех методов перечня не составишь. И даже если это удастся сделать сегодня, то что делать с теми методами, которые будут изобретены завтра?

Во-вторых, совершенно не ясно, что означает «разрешение» использовать метод. Ну, скажем, то, что среднее арифметическое нельзя использовать так, как мы это сделали в приведенном выше примере, ясно уже на уровне здравого смысла. А почему его нельзя использовать для порядковых шкал? А как определить, можно или нельзя использовать для номинальных шкал, скажем, регрессионный анализ? Ответ уже не столь очевиден.

В-третьих, приведенный выше перечень, вообще говоря, неверен. Так, бывают случаи, когда среднее арифметическое можно использовать и для номинальных данных. Пусть, скажем, мы измерили пол: мужчинам приписали 1, а женщинам - 0. Для 10 человек получили последовательность: 0, 0, 1, 0, 1, 0, 0, 1, 1, 0. Нетрудно видеть, что соответствующее среднее арифметическое будет равно 0,4. Если мы будем интерпретировать это обстоятельство так, как это обычно делается (наиболее типичный представитель рассматриваемой совокупности людей имеет пол 0,4), но, конечно, получим ерунду. Но попытаемся дать другую интерпретацию: доля единичных значений нашего признака в изучаемой совокупности составляет 40%. Вряд ли кто-нибудь будет возражать против того, что такая интерпретация вполне допустима. Так почему бы не считать среднее арифметическое для пола? Интерпретировать надо адекватно, и все будет в порядке.

Для решения поставленных вопросов в теории измерений был разработан специальный подход.

2.3. Проблема адекватности математического метода.

Интуитивно ясно, что, чем выше тип шкалы, тем более широкий круг методов применим для анализа соответствующих шкальных значений. Так, совершенно ясно, что к числам, полученным по номинальной шкале, многие традиционные математико-статистические методы не применимы. Это легко понять, рассмотрев две возможные номинальные шкалы из третьего столбца таблицы 2. Ясно, что, вообще говоря, очень многие методы будут давать разные результаты в зависимости от того, проанализируем ли мы с помощью выбранного метода числа (5, 5, 10, 185, 15) или же числа (25, 25, 3, 30, 1). Сравним, к примеру, средние арифметические значения первых трех объектов и последних двух объектов для каждой из номинальных шкал, приведенных в третьем столбце таблицы 2.

$$\frac{5+5+10}{3} = 6{,}7 < 100 = \frac{185+15}{2}.$$

Делаем содержательный вывод: первое среднее меньше второго. Проделаем то же для второй шкалы.

$$\frac{25+25+3}{3} = 17{,}7 > 15{,}5 = \frac{30+1}{2}$$

Делаем противоположный вывод: первое среднее больше второго.

А ведь если мы используем номинальную шкалу, то два рассмотренных набора шкальных значений содержат абсолютно одинаковую информацию! Значит, и результаты нашего анализа этих наборов должны быть одинаковыми. Естественно, у нас могут зародиться сомнения в целесообразности использования среднего арифметического для сравнения средних уровней двух рассматриваемых групп объектов.

Существуют ли методы, результаты применения которых не зависят от того, какую из двух рассмотренных шкал мы выберем? Конечно. Этому условию будут удовлетворять все методы, опирающиеся на подсчет частот. Скажем, модальным значением в первом случае будет то, которым обладают первый и второй объект (скажем, если цифрой 5 закодирована национальность «татарин», то в рассматриваемом случае татар у нас – больше всего). То же – и во втором случае (снова «татар» у нас больше всего). Выводы не изменились. И вроде бы нет причин считать моду неподходящей статистикой для номинальной шкалы.

Напротив, для интервальной шкалы большинство методов применимо. Попробуем, например, решить ту же задачу по сравнению средних для первых трех объектов и последних двух объектов для каждой из интервальных шкал, приведенных в третьем столбце таблицы 2.

Для первой шкалы имеет место:

$$\frac{5+5+10}{3} = 6{,}7 < 100 = \frac{185+15}{2}$$

Другими словами, наш содержательный вывод состоит в том, что первое среднее меньше второго. Теперь попробуем проделать то же для второй интервальной шкалы из третьего столбца таблицы.

$$\frac{25+25+35}{3} = 28{,}3 < 215 = \frac{385+45}{2}$$

Вывод – тот же. Рассмотренный пример не дает оснований для возникновения сомнений в допустимости сравнения средних арифметических для интервальной шкалы. Разумно полагать, что аналогичные соображения будут справедливы и насчет моды. Более того, интуитивно ясно, методы, подходящие для номинальной шкалы, будут подходить и для интервальной. Все сказанное не случайно.

Попытаемся выразить те же соображения в более строгом виде

Математический метод называется *формально адекватным,* если результаты его применения не зависят от применения к исходным данным допустимых преобразований тех шкал, по которым эти данные получены. Если использовать это определение, то становится ясным, почему нельзя использовать среднее арифметическое для национальностей. Мы показали, что сравнение двух средних арифметических формально не адекватно относительно номинальных шкал.

Покажем теперь, почему с формальной точки зрения нельзя рассчитывать средние для номинальной шкалы. Вернемся к обсужденному выше примеру. Напомним, что полученный содержательный вывод звучал так: «среднее между русским и украинцем равно татарину». Смысл этого утверждения изменится, если мы применим к набору исходных шкальных значений, скажем, следующее допустимое (для номинальной шкалы) преобразование: русским будем ставить в соответствие число 12, татарам – 5, украинцам – 10, чукчам – 11. Тогда среднее между русским и украинцем будет равно $\frac{12+10}{2} = 11$, т.е. чукче (а не татарину, как выше). Говоря формально, содержательный результат изменился в результате допустимого преобразования исходных (номинальных) шкал. Метод (подсчет среднего арифметического) формально не адекватен.

Напротив, наш результат, полученный с помощью расчета среднего арифметического для дихотомической шкалы, использованной нами при измерении пола, останется неизменным, если мы будем правильно его формулировать. А формулировку возьмем такую: среднее арифметическое делит интервал между числом, соответствующим женщине, и числом, соответствующим мужчине, в отношении 4:6 (что имело место выше и что по существу и говорило о доле единичных значений, равной 0,4). Покажем неизменность этого отношения на примере. Перекодируем произвольным образом значения пола: мужчинам припишем значение 48, а женщинам – 40. Нетрудно проверить, что тогда совокупность наших десяти шкальных значений превратиться в 40, 40, 48, 40, 48, 40, 40, 48, 48, 40, а среднее арифметическое будет равно 43,2, и будет верно соотношение: $\frac{43{,}2-40}{48-43{,}2} = \frac{3{,}2}{4{,}8} = \frac{4}{6}$.

Отношение, говорящее о доле значений «48», осталось тем же.

Поясним, почему выше мы дали определение не просто адекватности метода, а формальной адекватности. Дело в том, что формальная адекватность того или иного метода еще не делает его подходящим для решения социологической задачи соответствующего плана. Требуется еще другая адекватность, которую можно назвать содержательной – соответствие заложенной в методе модели сути решаемой задачи, априорным гипотезам исследователя (так, при осуществлении классификации объектов задействованная в алгоритме функция расстояния может быть формально адекватна типу используемых шкал, но содержательно совершенно не отвечать представлениям исследователя о похожести классифицируемых объектов).

Примеры задач.

1. Показать, что соотношение

$$\overline{X}_1 < \overline{X}_2$$

является инвариантным относительно допустимых преобразований интервальных шкал и не является таковым относительно допустимых преобразования порядковых шкал. На основе соответствующих рассуждений объяснить, почему нельзя усреднять результаты ранжировок респондентами каких-либо объектов с целью получения оценочных шкал (оценочная шкала – это

процесс приписывания рассматриваемым объектам чисел, отражающих усредненное отношение к этим объектам всей совокупности респондентов).
Рекомендация. Допустимые преобразования используемой шкалы должны быть одними и теми же для рассматриваемых шкальных значений. В данном случае – и тех, для которых рассчитывается $\overline{X}_1$, и тех для которых считается $\overline{X}_2$.

2. Доказать, что значение коэффициента корреляции не изменится, если к исходным данным применить допустимое преобразование интервальных шкал.

3. Предположим, что мы имеем совокупность значений номинального признака X с двумя значениями 0 и 1. Пусть p – доля "1", q – доля "0". Выразить $\overline{X}$ и S_x через p и q.

4. Описать, какова разница интерпретаций чисел 2, 3, 7 в ситуациях, когда эти числа получены по номинальной, порядковой или интервальной шкале.

5. Доказать формальную адекватность моды (в любом контексте ее использования) для номинальной шкалы

6. Доказать формальную адекватность рангового коэффициента корреляции (Спирмена или Кендалла) для порядковой шкалы.

Добавочная литература к теме 2

Обязательная

Толстова Ю.Н. Измерение в социологии. М.: Инфра-М, 1998

ТЕМА 3

Стандартизация значений случайных величин. Виды некоторых специфических распределений[22], использующихся при переносе результатов с выборки на генеральную совокупность

3.1. Стандартизация (нормировка) значений случайной величины: способы и цели

В эмпирических исследованиях зачастую бывают задействованы такие признаки, значения которых не сравнимы друг с другом по величине. И если мы не обратим на это внимания, то при анализе данных можем придти к нелепости. Например, предположим, что мы хотим построить типологию какой-то совокупности людей, описываемых, в частности, значениями их зарплаты и возраста. Включаем компьютер, «просим» его осуществить классификацию наших респондентов. Компьютер умеет работать с числами. В соответствии с большинством известных алгоритмов классификации, оценивая по определенным правилам степень близости между всевозможными парами объектов, программа будет близкие объекты относить к одному классу, далекие – к разным. Представим себе две пары людей: респонденты первой пары отличаются друг от друга только тем, что у одного – зарплата на 50 рублей больше, чем у другого; объекты же второй пары – только тем, что у них такая же разница в возрасте (50 лет). Вероятно, при любом разумном алгоритме, если уж первые два респондента окажутся включенными в один класс, то и вторые – тоже, и обратно. Вряд ли это можно считать разумным: какова бы ни была решающаяся задача, различие зарплаты в 50 рублей вряд ли стоит принимать во внимание, а различие в возрасте в 50 лет – напротив, по-видимому, надо будет учесть.

Могут возникнуть недоразумения и из-за того, что наблюдаемые значения рассматриваемых признаков будут «колебаться» вокруг сильно отличающихся друг от друга точек числовой прямой (если, например, среднее арифметическое значение одного признака равно 5000 (рублей), а другого – 50 (лет)).

Чтобы подобных недоразумений не происходило, признаки обычно определенным образом нормируют (хотя, вообще говоря, бывают задачи, когда этого делать не надо). Нормировка бывает

---

[22] Относительно каждого рассматриваемого распределения необходимо знать формулу, его определяющую, вид кривой плотности распределения, основные параметры распределения (математическое ожидание и дисперсию). Описание всех рассматриваемых распределений, в том числе отвечающие им графики функций плотности можно найти, например, в работе: Тюрин Ю.Н., , Макаров А.А. , Анализ данных на компьютере. М.: ИНФРА-М, 2003. С. 80 - 85.

разной. Чаще всего делают так, чтобы среднее значение признака стало равным нулю, а остальные значения измерялись в «сигмах». Нетрудно видеть, что к такой ситуации приводит следующая нормировка (стандартизация) всех значений признака x:

$$x \Rightarrow \frac{X-\mu}{\sigma}$$

3.2. Нормальное распределение (повторение).

Если некоторая случайная величина ξ имеет нормальное распределение, то будем использовать для обозначения этого обстоятельства выражение:

$$\xi \sim N(\mu, \sigma).$$

Напомним, что нормальное распределение задается двумя параметрами: μ - математическое ожидание, σ - стандартное (среднеквадратическое), отклонение (σ^2 – дисперсия).

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Из курса теории вероятностей читатель должен помнить, что кривая плотности нормального распределения представляет собой симметричный «холм», вершина которого находится над точкой X=μ; ширина же «холма» зависит от величины σ: чем больше эта величина, тем более пологим этот «холм» является.

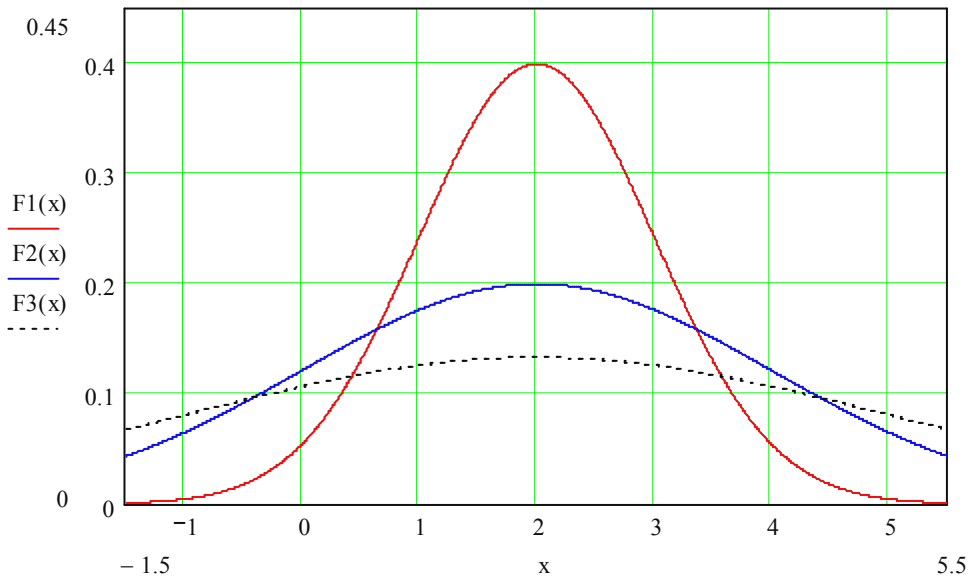


Рис. 3.1. Функции плотности нормального распределения с с математическим ожиданием μ = 2 и различными средними квадратическими отклонениями (σ = 1 для F1(x); σ = 2 для F2(x); σ = 3 для F3(x))

Произвольная нормально распределенная случайная величина превращается в *стандартизованную* $\xi_{станд} \sim N(0,1)$, если осуществить преобразование

$$x \Rightarrow \frac{X-\mu}{\sigma}$$

(мы получили тем самым стандартизованное распределение).

О роли такого представления нормально распределенной случайной величины вы слышали в курсе теории вероятностей.

Для хорошего восприятия дальнейшего материала важно вспомнить, что для нормального распределения хорошо изучено, какова вероятность попадания значения соответствующей случайной

величины в разные отрезки числовой оси (подробнее о том, что из курса теории вероятности следует повторить, см. в конце настоящей лекции)

3.3. Распределение Хи-квадрат

Пусть случайные величины ξ_1, ξ_2, ... , ξ_n - независимы и каждая имеет стандартное нормальное распределение N (0,1). Говорят, что случайная величина χ_n^2, определенная как

$$\chi_n^2 = \xi_1^2 + \xi_2^2, \dots , \xi_n^2$$

имеет *распределение хи-квадрат с n степенями свободы*[23]. Для этой случайной величины составлены разнообразные таблицы. Чаще всего они содержат значения p-квантилей.

Заметим, что плотность распределения этой функции можно выразить определенной формулой (как мы выражали, например, плотность функции нормального распределения). Но эта формула очень сложна, поэтому мы не выписываем ее в основной части текста[24].

Подчеркнем, что существует не одно распределение хи-квадрат, а целое *семейство* таких распределений, каждый член семейства задан отвечающим ему значением *n*. Другими словами, для каждого *n* существует *свое* распределение хи-квадрат. И если нам понадобится таблица, задающая это распределение, то мы должны будем выбрать нужную из целого семейства таблиц, каждая из которых задается своим числом степеней свободы.

Далее мы увидим, что понятие числа степеней свободы имеет смысл и для других распределений. Часто число степеней свободы рассматриваемого распределения обозначают сочетанием букв df (degree of freedom).

Известно, что $M\chi_n^2 = n$, $D\chi_n^2 = 2n$, $Mo\ \chi_n^2 = n - 2$ (для $n \geq 2$).

Для сравнения напомним, что нормальное распределение тоже задается парой двух параметров – математическим ожиданием и дисперсией; однако чтобы связать какое-либо значение нормально распределенной случайной величины $\xi \sim N(\mu,\sigma)$ с соответствующей вероятностью, нет необходимости выбирать нужную нам таблицу из некоторого семейства, поскольку в справочниках обычно фигурирует лишь одна таблица – для стандартизованной случайной величины $\xi_{станд} \sim N(0,1)$; все необходимые сведения для случайной величины с произвольным распределением $N(\mu,\sigma)$ из нее могут быть получены.

Приведем графики плотности распределения χ_2^2 и χ_3^2 (см. рис. 3.2).

---

[23] Это распределение было открыто астрономом Ф.Хельмертом в 1876 году, а в 1900 году получило название «Хи-квадрат» в работе Пирсона (ссылку см. в: [Айвазян, Мхитарян, 2001, с.142]).

[24] Эта формула выглядит так: $f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} \bullet x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, x \geq 0$, где используется большая греческая буква «гамма» для обозначения т.н. «гамма-функции» Эйлера (1707-1783)[24], определяемой следующим образом: $\Gamma(y) = \int_0^\infty e^{-t} t^{y-1} dt$

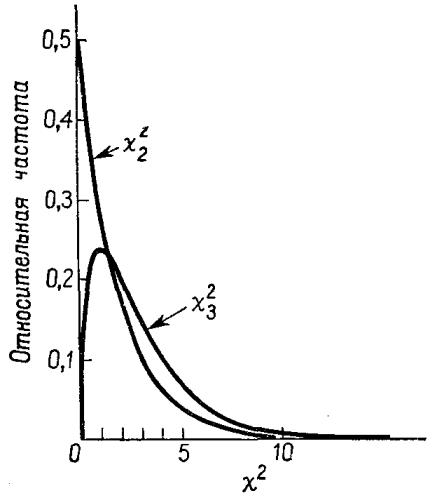


Рис. 3.2. Функции плотности распределения «Хи-квадрат» с числом степеней свободы 2 и 3.[25]

(вдоль вертикальной оси, строго говоря, следовало бы написать «плотность вероятности», а не «относительная частота»; второе, вообще говоря, годится лишь при работе с выборкой) !!!!!!!!!!!!!!!!!!!!

Приведем примеры «Хи-квадрат»-распределений с другими числами степеней свободы, указывая при этом, вероятности попадания значений рассматриваемых случайных величин в некоторые полуинтервалы.

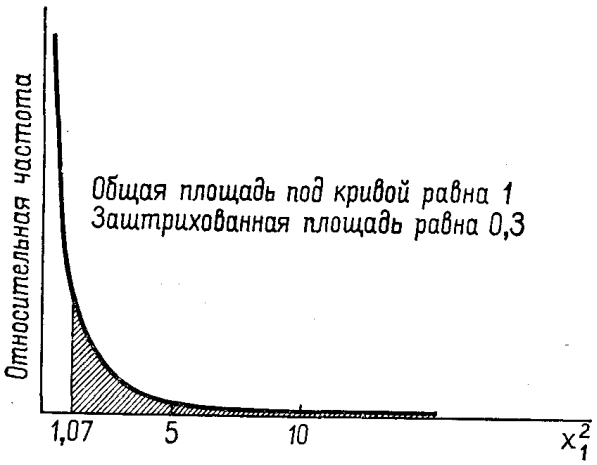


Рис. 3.3. Функция плотности распределения «Хи-квадрат» с числом степеней свободы 1. [26]

(вдоль вертикальной оси, строго говоря, следовало бы написать «плотность вероятности», а не «относительная частота»; второе, вообще говоря, годится лишь при работе с выборкой) !!!!!!!!!!!!!!!!!!!!!!!!!

[25] Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. М.: Прогресс,1976. С.209
[26] Там же, с. 208

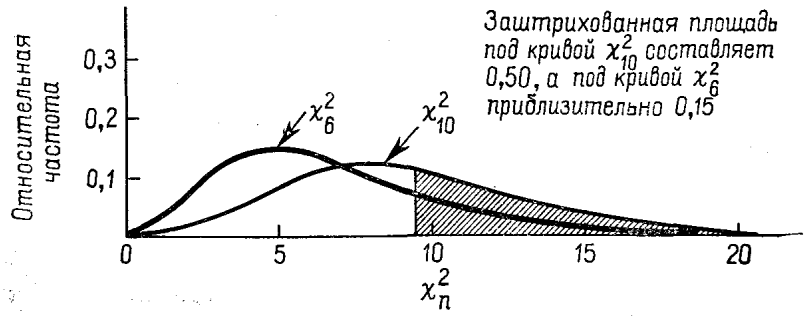


Рис. 3.4. Функции плотности распределения «Хи-квадрат» с числом степеней свободы 6 и 10.[27]

(вдоль вертикальной оси, строго говоря, следовало бы написать «плотность вероятности», а не «относительная частота»; второе, вообще говоря, годится лишь при работе с выборкой) !!!!!!!!!!!!!!!!!!!!

3.4.Распределение Стьюдента[28] *(t-распределение)*

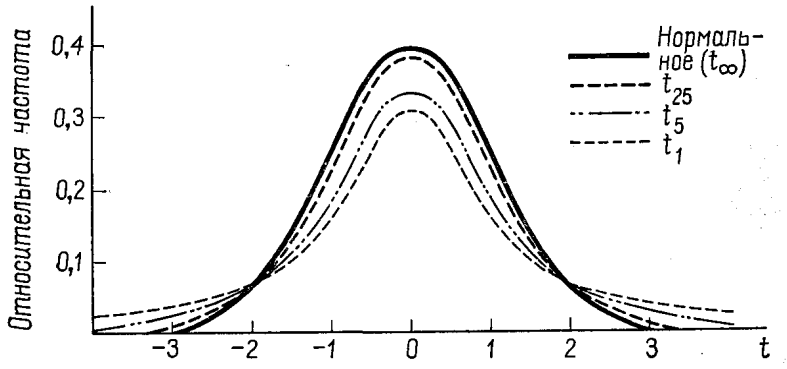
Пусть случайные величины ξ_0, ξ_1, ... , ξ_n – независимы и каждая имеет стандартное нормальное распределение $N(0,1)$[29]. Говорят, что случайная величина t_n, определенная как

$$t_n = \frac{\xi_0}{\sqrt{\frac{1}{n} \cdot \sum \xi_i^2}}$$

имеет *распределение Стьюдента с n степенями свободы.*

$M\, t_n = 0$; $D\, t_n = n / (n - 2)$ $(n > 2)$; $Mo = 0$.

При больших n вместо распределения Стьюдента обычно рекомендуется использовать стандартное нормальное распределение. Встает вопрос о том, какие выборки называть большими. В западной традиции большой называют выборку, если $n \geq 30$ (см., например, учебник Блумана). Отечественная наука более осторожна: большой называется выборка, если $n \geq 100$. Строгих правил здесь не может существовать в принципе. Все зависит от того, какая степень надежности получаемых выводов нам требуется. Ответ на вопрос об объеме выборки может дать только практика. При этом российские ученые опираются на практику использования математической статистики в естественных науках и технике, а западные учитывают и опыт гуманитарных наук, где погоня за большой точностью и надежностью зачастую теряет смысл, поскольку достаточно «грубыми» являются результаты измерений рассматриваемых признаков (социолога обычно волнует вопрос о том, то ли он меряет, что хочет, и вопрос о точности уходит на второй план). Так что мы далее будем считать, что выборки объема более 30 – достаточно большие для того, чтобы распределение Стьюдента подменять нормальным.



[27] Там же, с. 209

[28] Настоящая фамилия Стьюдента – У. Госсет (1876-1937). А.А.Чупров в своей лекции, прочитанной на шведском языке 19 сентября 1918 года в Обществе шведских актуариев, называет Стьюдента «талантливым учеником Пирсона» (см.: Чупров А.А. Вопросы статистики. М.: Госстатиздат ЦСУ СССР, 1960. С. 225).

[29] Как отмечается в работе [Айвазян, Мхитарян, 2001, с.144], со ссылкой на работу Стьюдента, здесь рассматриваемые случайные величины могут иметь произвольную дисперсию σ^2 , одинаковую для всех ξ_i. Итоговое распределение от этой дисперсии не зависит.

Рис. 3.5. Функции плотности распределения Стьюдента с различным числом степеней свободы. Для сравнение на том же графике приведена функция плотности нормального распределения. [30]

КРАСНОЕ – МЕЛКИМ ШРИФТОМ

И еще один вопрос должен был бы возникнуть у читателя (правда, за 10 лет чтения автором лекций по рассматриваемому предмету студентам-социологам нашелся лишь один слушатель, задавший этот вопрос): что делать, если n=1 или n=2? Дисперсия будет отрицательной? Нам представляется полезным дать ответ на этот вопрос, поскольку этот ответ позволит читателю-социологу вспомнить кое-что из интегрального исчисления и лишний раз убедиться в том, что полученные им на первом курсе вуза знания по высшей математике отнюдь не бесполезны.

Ответ на указанный вопрос таков: при n=1 или n=2 дисперсия соответствующего распределения Стьюдента не существует. ДАЛЕЕ, до п. 3.5, - МЕЛКИМ ШРИФТОМ.

Позволим себе привести здесь доказательство этого факта для n=1 (что, на наш взгляд, должно иметь определенное «воспитательное» значение).

Плотность распределения Стьюдента имеет вид: $f(x) = \frac{1}{\sqrt{\pi n}} \bullet \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \bullet (1 + \frac{x^2}{n})^{-\frac{n+1}{2}}$

(-∞ < x < +∞, напомним, что т.н. гамма-функция имеет вид: $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$). При n=1 f (x) принимает вид:

$$f(x) = \frac{1}{\sqrt{\pi}} \bullet \frac{\Gamma(1)}{\Gamma(\frac{1}{2})} \bullet (1 + x^2)^{-1} = \frac{const}{1 + x^2}$$

где *const* – это некоторая величина, не зависящая от нашей переменной x.

Напомним, что: (1) дисперсия случайной величины *x* равна $D(x) = M(x - M(x))^2$, (2) математическое ожидание случайной величины *x* с плотностью распределения *f(x)* вычисляется по формуле: $M(x) = \int_{-\infty}^{+\infty} x f(x) dx$, (3) математическое ожидание величины, распределенной по закону Стьюдента, равно нулю. Поэтому для рассматриваемого нами случая имеет место цепочка равенств:

$$D(x) = M(x^2) = \int_{-\infty}^{+\infty} x^2 f(x) dx = const \int_{-\infty}^{+\infty} \frac{x^2}{1 + x^2} dx = (x - arctg\, x) \Big|_{-\infty}^{+\infty} = -\infty - \infty = -\infty$$

Другими словами, интеграл расходится, т.е. интересующая нас дисперсия не существует

3.5. Распределение Фишера[31] (F – распределение, распределение дисперсионного отношения).

---

[30] Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. М.: Прогресс,1976. С. 213

[31] Р.А.Фишер (1890 - ??? в 1960 был еще жив) – известный английский статистик. Рассматриваемое распределение было открыто им в 1924 году (ссылку см. в книге [Айвазян, Мхитарян, 2001, с. 145]; там же указано, что все рассматриваемые случайные величины могут иметь одинаковую дисперсию, отличную от единицы, ср. сноску 28).

Пусть случайные величины $\eta_1, \eta_2, ..., \eta_m; \xi_1, \xi_2, ..., \xi_n$ (m и n – натуральные числа) – независимы и каждая имеет стандартное нормальное распределение $N(0,1)$. Говорят, что случайная величина $F_{m,n}$, определенная как

$$F_{m,n} = \frac{(\eta_1^2 + \eta_2^2 + ... + \eta_m^2)/m}{(\xi_1^2 + \xi_2^2 + ... + \xi_n^2)/n},$$

имеет *F-распределение* с параметрами m и n. Натуральные числа m и n называют числами степеней свободы.

$M F_{m,n} = n / (n - 2)$ для $n > 2$, $D F_{m,n} = (2 n^2(m + n - 2)) / (m (n - 2)^2(n - 4))$ для $n > 4$; $Mo = n (m - 2) / m (n + 2)$, $m > 1$.

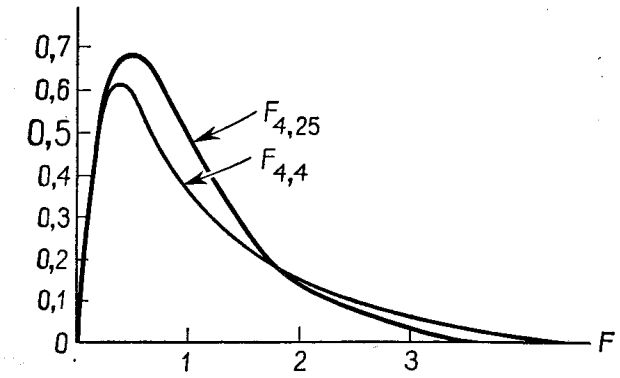


Рис. 3.6. Функции плотности F-распределения с разным числом степеней свободы[32]

Повторение отдельных фрагментов курса по теории вероятностей

1. Общее представление о вероятностных таблицах. Принципы их использования.
2. Определение нормального распределения. Вид нормальной кривой, «физический» смысл отвечающих ей математического ожидания и дисперсии. Целесообразность измерения значения случайной величины в единицах σ (среднего квадратического отклонения). Соотношение площадей под нормальной кривой и вероятностей попадания в отрезки ($\mu \pm \sigma$, $\mu \pm 2\sigma$, $\mu \pm 3\sigma$). Определение размера отрезка (в единицах сигма) при заданной доле попадания в него. Функция Лапласа.
   Следует запомнить, каковы доли площадей под разными частями нормальной кривой; учесть, что целому числу σ обычно отвечают «корявые» проценты: одна σ – 68,3 % площади, 2 σ – 95,4%; а «круглые» процентам отвечает «корявое» число «сигм»: 90% - 1,64 σ, 95% - 1,96 σ, 99% - 2, 57σ и т.д. Эти и другие значения представлены на рис. 3.7.
   Необходимо также понимать связь функции плотности распределения с функцией распределения и принципы нахождения квантилей распредлеления (например, процентилей). Этому также может способствовать рис. 3.7, на котором отражено, как из функции плотности получается функция распределения (отвечающая накопленным процентам) и как по функции распределения можно рассчитывать процентили.

[32] Там же, с. 211

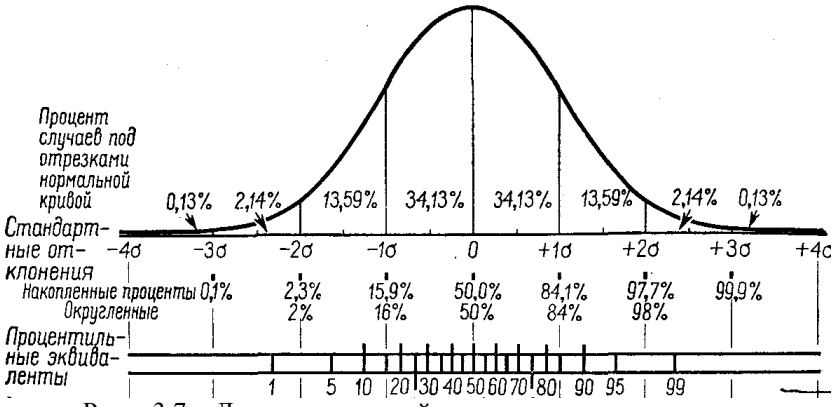


Рис. 3.7. Доли площадей под разными частями кривой плотности нормального распределения. Указание процентилей [33]
(подобные иллюстрации можно найти и в других работах[34]).

3. Правило трех сигм.
4. Роль стандартизации нормально распределенных случайных величин при использовании соответствующих вероятностных таблиц. Использование таблицы для стандартизованной величины при получении характеристик распределения произвольной нормально распределенной случайной величины.

   Пример того, как по величине отрезка под стандартизованной нормальной кривой можно определять вероятность попадания в этот отрезок соответствующей случайной величины, можно найти на рис. 3.8.

Значение z	Площадь слева от z	Ордината для z
0,00	0,5000	0,3989
0,50	0,6915	0,3521
−1,27	0,1020	0,1781
1,96	0,9750	0,0584

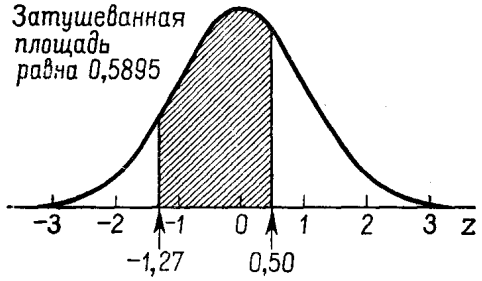


Рис. 3.8. Иллюстрация того, как отрезок диапазона изменения стандартизованной нормальной кривой связан с отвечающей этому отрезку площадью под этой кривой (т.е. с вероятностью попадания значения величины в рассматриваемый отрезок).[35]

5. Умение использовать вероятностные таблицы разного вида для нормального распределения. Желательно потренироваться на таблицах разных видов, например:

- таблицы значений функции Лапласа $\Phi(x)$[36];
- таблицы, в которой задана вероятность попадания в правый конец графика функции плотности (по заданному z определяется вероятность того, что $f(x) > z$)[37];
- таблицы значений удвоенной функции Лапласа[38];
- таблицы верхних процентных точек стандартного нормального распределения[39].

---

[33] Там же, с. 97

[34] См., например, Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. М.: ИНФРА-М, 2003.С. 79; Kachigan S.K. Statistical analysis. An interdisciplinar introduction to univariate and multivariate methods. - N.Y.: Radius Press,1986; С. 139.

[35] Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. М.: Прогресс,1976. С. 95

[36] См., например: Гмурман В.Е. Теория вероятностей и математическая статистика. М.: Высшая школа, 1998, С. 462-463; Колемаев В.А., Калинина В.Н. Теория вероятностей и математическая статистика, М.: Инфра-М,1997; Юнити, 2003. С. 293.

[37] Эддоус М., Стэнсфилд Р. Методы принятия решения. М.: Финансы и статистика, 1997. С. 575.

[38] Калинина В.Н., Панкин В.Ф. Математическая статистика. М.: Высшая школа, 1998. С. 331

[39] Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. М.: ИНФРА-М, 2003. С. 496

6. Таблицы, составленные для всех остальных рассмотренных выше распределений. Их разные виды. Квантили распределений. Понимание того, что таблицы, как правило, содержат значения *p*-квантилей. Принцип практической невозможности маловероятных событий

Примеры задач

1. Доказать, что среднее арифметическое значение стандартизованного признака равно нулю, а среднее квадратическое отклонение – единице.
2. Известно, что распределение оценок, полученных абитуриентами при ответе на некоторый 20-балльный тест, имеет вид N (10, 3). Какова вероятность того, что абитуриент получит балл от 7 до 11? Объяснить, как находится подобная вероятность при использовании четырех видов вероятностных таблиц.
3. С какой площадью под графиком функции плотности стандартизованного нормального распределения соотносится значение функции Лапласа?
4. Как следует понимать упомянутое выше выражение «верхние процентные точки» из книги Тюрина и Макарова? В некоторых работах эти точки называются процентилями. Что такое процентиль? Что такое квантиль (частным случаем которого является процентиль)? Какие еще виды квантилей вы знаете?

ТЕМА 4
Предельные теоремы.[40]

При изучении результатов наблюдений над реальными массовыми случайными явлениями (над выборочными значениями изучаемых случайных величин) часто наблюдаются определенные закономерности, обладающие свойством устойчивости при рассмотрении разных выборок. Суть устойчивости состоит в том, что конкретные свойства каждого явления почти не сказываются на среднем результате. Наблюдаемые на выборке характеристики случайных величин при неограниченном увеличении количества и объема выборок становятся практически не случайными. Предельные теоремы, о которых идет речь ниже, фактически устанавливают зависимость между случайностью и необходимостью. По смыслу эти теоремы можно разбить на две большие группы – центральную предельную теорему и закон больших чисел (хотя по сути они отражают одно и то же).

4.1. Центральная предельная теорема

Центральной предельной теоремой обычно называют группу утверждений (точнее, каждое утверждение из этой группы), которые устанавливают связь между законом распределения суммы случайных величин и его предельной формой - нормальным законом распределения. Различные формулировки отличаются условиями, которые накладываются на исходные случайные величины.

Прежде всего напомним формулировку (в упрощенном виде[41]) одной из известных теорем А.М.Ляпунова (1857–1918) (впервые, но в более простом виде, эта теорема была доказана П.Л. Чебышевым (1821–1894) в 1887 году).

Теорема Ляпунова (1901).

Распределение суммы независимых случайных величин X_1, X_2, X_3, ..., X_n приближается к нормальному закону распределения при неограниченном увеличении n, если выполняются следующие условия:

1) все величины имеют конечные математические ожидания и дисперсии;
2) ни одна из величин по своему значению резко не отличается от всех остальных, т.е. оказывает ничтожное влияние на их сумму.

---

[40] Огромную роль в становлении математической статистики, в частности, в доказательстве тех ее основных теорем, которые затрагиваются в настоящем параграфе, сыграли русские ученые. С середины 19-го и почти до конца 20-го века русская математико-статистическая школа играла лидирующую роль в мировой науке. О том, какое значение русские статистики придавали закону больших чисел, говорит то, что в декабре 1913 года было проведено специальное заседание Российской Академии наук (по инициативе А.А.Маркова (1856–1922)), посвященное 200-летию открытия этого закона (что связывалось с выходом в свет в 1713 году знаменитой работы Я.Бернулли (1654–1705), в которой была дана первая, самая простая его формулировка).

[41] См., например: Калинина В.Н., Панкин В.Ф. Математическая статистика. М.: Высшая школа, 1998. С. 130])

Таким образом, предельное распределение суммы случайных величин в условиях рассмотренной теоремы не зависит от вида распределений самих случайных величин.

На опыте установлено, что распределение суммы независимых случайных величин, у которых дисперсии не отличаются резко друг от друга, довольно быстро приближаются к нормальному. Уже при числе слагаемых, большем 10, распределение суммы можно заменить нормальным.

Теорема Ляпунова справедлива и для дискретных случайных величин.

Отметим, что центральная предельная теорема имеем большое значение для социолога, поскольку она объясняет причину большого распространения в природе (в том числе – в обществе) нормального распределения, оправдывает делаемые зачастую априори, без проверки, предположения о нормальности тех распределений, которые анализируются в социологических исследованиях. Приведем пример.

В ряде методов шкалирования предполагается, что мнение человека о любом объекте плюралистично. Это означает, что если бы у нас имелся инструмент измерения такого мнения, и мы имели бы возможность использовать его много раз, то, вообще говоря, каждый раз получали бы разные значения. Этим значениям отвечало бы некоторое распределение вероятностей. Важное для нас утверждение состоит в том, что при этом обычно полагают, что указанное распределение нормально (это делается, например, в методе парных сравнений – одном из известных способов получения экспертных оценок[42]). И такую посылку вполне можно принять, если опереться на центральную предельную теорему.

Другая формулировка теоремы Ляпунова.

Если случайная величина X имеет математическое ожидание MX и дисперсию DX, то распределение среднего арифметического $\overline{X} = (\Sigma\ X_i) / n$, вычисленного по наблюдавшимся значениям случайной величины в n независимых испытаниях, проведенных в одинаковых условиях, при $n \to \infty$ приближается к нормальному с математическим ожиданием MX и дисперсией DX / n. Другими словами,

$$\overline{X} \sim N\left(MX, \sqrt{\frac{DX}{n}}\right)$$

(в таких случаях говорят, что $\overline{X}$ имеет асимптотически нормальное распределение с указанными параметрами; «асимптотически» означает, что распределение тем ближе к нормальному, чем больше объем выборки).

Или, в других обозначениях, то же самое может быть записано следующим образом:

$$\overline{X} \sim \mathsf{N}\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right). \qquad (4.1)$$

Именно этот вариант формулировки теоремы Ляпунова будет активно использоваться нами далее.

Заметим, что, в соответствии с этой формулировкой, дисперсия средних, рассчитанных для случайных выборок объема n из совокупности с дисперсией σ^2, равна σ^2/n.

То, что, в соответствии с соотношением (4.1), средние арифметические значения рассматриваемого признака X, вычисленные для разных выборок, как бы сосредоточиваются вокруг математического ожидания μ_x, дает некоторые основания для того, чтобы использовать выборочное среднее арифметическое для оценки последнего. Еще одно основание для такого использования дает закон больших чисел.

4.2. Закон больших чисел.

Под законом больших чисел так же, как это имело место для центральной предельной теоремы, не следует понимать какой-то один общий закон, связанный с большими числами. Закон больших чисел – это обобщенное название нескольких теорем, из которых следует, что при неограниченном увеличении числа испытаний средние величины стремятся к некоторым постоянным. Мы рассмотрим лишь две формулировки соответствующего плана.

---

[42] Подробнее об этом см.: Толстова Ю.Н. Измерение в социологии. М.: Инфра-М, 1998. См. также сноску 11.

Во-первых, приведем формулировку Я.Бернулли (1654 – 1705) – ученого, открывшего закон больших чисел. Теорема Бернулли утверждает, что в последовательности независимых испытаний, в каждом из которых вероятность наступления некоторого события А имеет одно и то же значение p, 0<p<1, верно соотношение

$$P(\left|\frac{S}{n} - p\right| \geq \varepsilon) \to 0$$

при любом $\varepsilon > 0$ и $n \to \infty$. Здесь S – число появления события А в n испытаниях, S/n – частота появлений. Другими словами, при достаточно большом числе испытаний вероятность того, что частота появления события будет значимо отличаться от соответствующей вероятности, практически сводится к нулю.

Во-вторых, вспомним творчество Чебышева.

Частный случай теоремы Чебышева (1867)

Если $X_1, X_2, \ldots, X_n$ – выборочные значения случайной величины X, имеющей математическое ожидание MX и дисперсию DX, то для любого как угодно малого ε имеет место соотношение[43]:

$$\lim_{n\to\infty} P(\left|\overline{X} - MX\right| \leq \varepsilon) = 1$$

Итак, при увеличении объема выборки вероятность того, что вычисленное для выборки среднее арифметическое значение рассматриваемой случайной величины с вероятностью, практически равной единице, будет совпадать с соответствующим математическим ожиданием. Другими словами, среднее арифметическое случайной величины обладает определенной устойчивостью.

Этот факт дает нам еще одно основание для использования среднего арифметического как выборочной оценки генерального параметра – математического ожидания.

Основное значение теоремы Чебышева именно в том и состоит, что она позволяет, используя среднее арифметическое, получить представление о величине математического ожидания.[44]

Перейдем к более подробному рассмотрению того, каким образом генеральные параметры можно оценивать с помощью адекватным образом подобранных статистик.

Добавочная литература к теме 4.

Бернулли Я. О законе больших чисел. М., 1986

Пасхавер И.С. Закон больших чисел и статистические закономерности. М.: Статистика, 1974

Раздел II. ОЦЕНИВАНИЕ ПАРАМЕТРОВ

ТЕМА 5
Еще раз о задачах математической статистики. Точечное оценивание параметров

5.1. О задачах математической статистики

Главными задачами математической статистики, как мы уже отмечали, являются (а) изучение случайных величин путем осуществления выборочных оценок параметров их распределений и (б) перенос результатов с выборки на генеральную совокупность. Известны два подхода, позволяющие осуществлять упомянутый перенос: *статистическое оценивание параметров и проверка статистических гипотез*.

Прежде, чем перейти к их описанию, напомним, что сам термин «параметр распределения» имеет смысл лишь для генеральной совокупности. В выборке каждому параметру отвечает некоторая специальным образом подобранная статистика, т.е. функция от наблюдаемых, выборочных, значений рассматриваемой случайной величины (напомним, что для выборки «случайная величина»

---

[43] Строго говоря, здесь следовало бы величины $X_1, X_2, \ldots, X_n$ рассматривать как реализацию значений n независимых одинаково распределенных случайных величин.

[44] В середине и второй половине 19-го века в тех исследованиях, которые мы сейчас причисляем к области эмпирической социологии (тогда они отождествлялись с исследованиями в области политической арифметики, моральной статистики, а в России – еще и земской статистики), активно использовались идеи известного бельгийского ученого Л.А.Ж. Кетле (1796-1874), создателя т. н. теории «среднего человека». Однако подлинное теоретическое обоснование эта теория получила лишь в работах Чебышева.

превращается в «признак»). Удачность подбора статистики, отвечающей какому-либо параметру, по существу и означает то, что с помощью этой статистики можно судить о том, каков генеральный параметр. Ниже мы проанализируем, какими качествами для этого должна обладать статистика, и покажем, каким образом соответствующие оценки можно находить. Существует два способа, два вида оценивания параметров: *точечное и интервальное*. Точечное оценивание состоит в том, что мы вычисляем выборочное значение статистики, отвечающей нашему параметру, и именно это значение считаем хорошей выборочной оценкой генерального значения параметра. Интервальное оценивание происходит по иной схеме. Если τ -генеральное значение параметра, а t – выборочное значение соответствующей статистики, то интервальное оценивание означает указание того, что с некоторой вероятностью P значение τ будет заключено в интервале

$$t - \Delta \le \tau \le t + \Delta$$

(точнее: с вероятностью P указанный интервал «накроет» значение параметра τ).

При этом предполагается, что величины P и Δ очевидным образом детерминируют друг друга: чем больше одна, тем больше и другая (ясно, что исследователю всегда хочется, чтобы величина Δ была бы поменьше, а величина P – побольше, однако одно желание противоречит другому).

Правила построения и точечных, и интервальных оценок для большинства известных параметров вероятностных распределений даются математической статистикой. Заложены они и в большинстве известных пакетов прикладных программ для ЭВМ. Однако механическое использование вычислительной техники не может способствовать успешному решению социологических задач. Ряд величин, необходимых для осуществления оценки, исследователь должен задать сам (скажем, в рассмотренном выше соотношении между Δ и P исследователь должен сам выбрать, что для него важнее). А от выбора зависит интерпретация оценки. Эта интерпретация опирается на своеобразное статистическое видение реальности. Поэтому для эффективного использования положений математической статистики в эмпирическом социологическом исследовании необходимо понимать принципы осуществления статистических оценок. Мы опишем их на примерах рассмотрения наиболее популярных параметров.

5.2. Точечные оценки параметров. Предъявляемые к ним требования

В качестве статистики, отвечающей в вышеприведенном смысле математическому ожиданию, выступает среднее арифметическое. Другими словами, мы считаем, что, если $X_1, X_2, X_3, ..., X_n$ - выборочные значения некоторой случайной величины ξ (n – объем выборки), то точечной оценкой математического ожидания $M\xi$ (μ_x) этой случайной величины мы считаем число

$$\overline{X} = (X_1 + X_2 + X_3 + ... + X_n)/ n$$

Разумность выбора выборочного среднего арифметического в качестве точечной оценки генерального математического ожидания подтверждается центральной предельной теоремой и законом больших чисел, о чем пойдет речь ниже. Роль этих утверждений станет ясной, если мы обратимся к рассмотрению смысла тех свойств, которыми, в соответствии с положениями математической статистики, должна обладать «хорошая» точечная оценка того или иного параметра. Однако пока отвлечемся от качества точечных оценок и опишем некоторые модельные представления, типичные для названной науки.

Представим себе, что мы имеем некоторую генеральную совокупность и строим на ее основе бесконечное количество выборок одного и того же объема n, для каждой из которой вычисляем интересующую нас статистику – в данном случае среднее арифметическое значений нашей случайной величины. Схематически эту процедуру можно выразить следующим рисунком

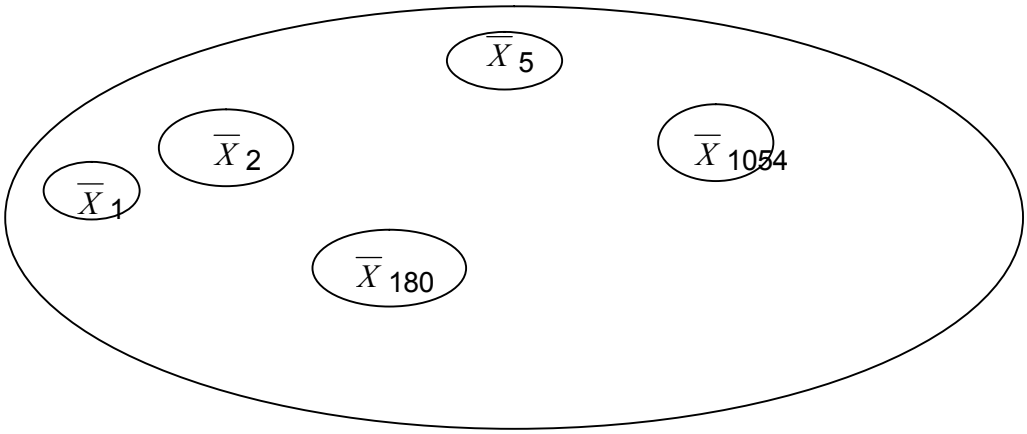


Рис. 5.1. Схематическое изображение процесса организации бесконечного количества выборок (одного и того же объема n) и получения соответствующей совокупности выборочных средних арифметических

Другими словами, мы имеем бесконечное количество выборочных средних

$$\overline{X}_1,\ \overline{X}_2,\ \overline{X}_3, ...,\ \overline{X}_n, \ldots .$$

Эти средние можно считать реализацией некоторой случайной величины. Распределение таких средних хорошо изучено. Оно является нормальным с параметрами $(\mu_x, \frac{\sigma_x}{\sqrt{n}})$. Это следует из рассмотренной в предыдущей лекции теоремы Ляпунова (второй ее формулировки).

Опр. Величина

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

называется *средней (стандартной) ошибкой среднего, или средней (стандартной) ошибкой выборки для признака X.*

Таким образом, стандартная ошибка среднего – это стандартное (среднее квадратическое) отклонение выборочного распределения средних значений X, отвечающих бесконечному числу разных мыслимых выборок объема n из изучаемой генеральной совокупности с дисперсией σ^2.

Подчеркнём, что средняя ошибка выборки говорит о порядке величины случайного отклонения выборочной оценки среднего от "истинного" значения параметра генеральной совокупности (в данном случае «истинное» значение – это μ_x). Ясно, что упомянутая ошибка уменьшается с увеличением объёма выборки и с уменьшением среднего квадратического отклонения самого признака, т.е. с увеличением однородности совокупности по этому признаку (можно показать, что та же ошибка увеличивается с увеличением объёма генеральной совокупности; однако генеральную совокупность в большинстве интересующих социолога случаев имеет смысл считать бесконечной, а в таком случае очевидно, что об увеличении ее объема говорить нет смысла).

Распределение, аналогичное описанному распределению выборочных средних, можно строить для значений любой статистики (т.е. для точечных оценок любого параметра заданного распределения). Далее мы этим будем активно пользоваться при обсуждении вопроса о том, что такое «хорошая» точечная оценка («хорошая» статистика). Все, что было сказано относительно математического ожидания и среднего арифметического, можно обобщить на любой параметр и отражающую его статистику.

Рассмотрим некоторый параметр τ (в качестве такового может выступать математическое ожидание, дисперсия, коэффициент корреляции и т.д.). Пусть имеется какая-то выборка, содержащая информацию о нашем параметре, и мы выбрали некую статистику t, значение которой для выборки служит точечной оценкой нашего параметра. Чтобы подобные точечные оценки были «хорошими», требуется, чтобы они удовлетворяли некоторым свойствам. Для понимания смысла этих свойств

представим себе картину, аналогичную изображенной на рис. 3.1, т.е. представим, что мы осуществляем огромное количество выборок, для каждой из которых рассчитываем значение рассматриваемой статистики. Этим значениям отвечает некоторое распределение.

Опр. Указанное распределение обычно называется *выборочным распределением рассматриваемой статистики t (*точнее, следовало бы говорить о распределении оценок, получаемых с помощью выбранной статистики).

Для большей ясности заметим, что распределение среднего арифметического (точнее, средних арифметических), представленное на рисунке 1, - частный случай такого выборочного распределения.

Каждое выборочное распределение любой статистики t (оценивающей любой генеральный параметр τ) имеет свои параметры – в частности, свое математическое ожидание и дисперсию (как выше это имело место для выборочного распределения среднего арифметического). Для многих параметров τ подобные распределения изучены, определен соответствующий закон, найдены основные его характеристики.

Ниже в соответствии со сложившейся в литературе традицией, термины статистика и оценка иногда будем использовать как синонимы (до сих пор оценками у нас служили конкретные выборочные значения статистики). А именно, введем следующее определение.

Опр. Иногда будем называть *оценкой* параметра τ самое статистику t (а не ее отдельное значение, как раньше). Соответственно, будем говорить о *выборочном распределении оценки* (вместо выборочного распределения статистики). Вместо t иногда будем использовать обозначение t_n в знак того, что при вычислении значений t используются выборки объема n.

Надеемся, что предлагаемое смешение понятий “оценка” и “статистика” не приведет к недоразумениям.

Итак, рассмотрим свойства «хороших» точечных оценок (Гласс и Стэнли, 1976, с. 228-232; Калинина, Панкин,1998,с.162-174).

Опр. Оценка t параметра τ называется *несмещенной,* если среднее выборочного распределения оценки t (при любом фиксированном объеме выборок n) равно величине оцениваемого параметра:

$$M\,t = \tau.$$

Несмещенность статистики требуется для повышения вероятности того, что наше единственное выборочное значение этой статистики будет достаточно близко к генеральному значению соответствующего параметра. Для смещенных оценок повышается вероятность большой ошибки.

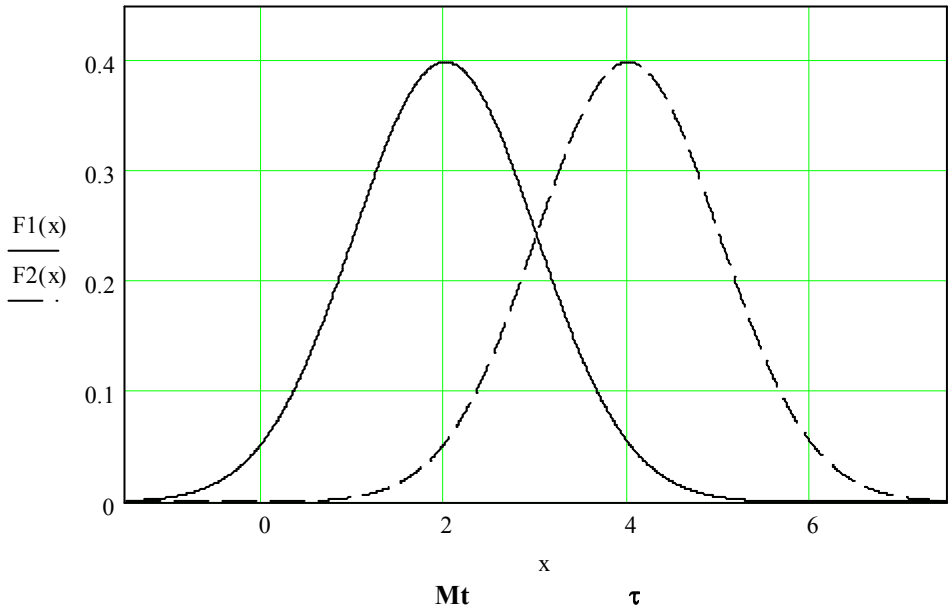


Рис. 5.2. Иллюстрация того, что смещенность оценки повышает вероятность того, что ее выборочное значение будет далеко отстоять от генерального (**Mt -** математическое ожидание выборочных оценок параметра **τ,** полученных с помощью смещенной статистики t ; **τ** - генеральное значение параметра; неравенство **Mt ≠ τ** означает смещенность статистики t; сплошной линией F1(x)

обозначено распределение упомянутых выборочных оценок; пунктирной F2(x)– то распределение гипотетических оценок, которые были бы получены с помощью несмещенной статистики).

Для пояснения обратимся к рис. 5.2. Предположим, что для оценки некоторого параметра τ используется значение нормально распределенной статистики t , распределение которой представлено на рисунке кривой F1 (x) (сплошная линия). В нашем распоряжении имеется только одно значение статистики t – то, которое мы вычислили для нашей единственной выборки. Очевидно, с относительно большой вероятностью это значение попадет в ближайшую окрестность точки x = 2 (поскольку Mt = 2). Вероятность же попасть в ближайшую окрестность точки x = 4 относительно мала. А ведь генеральное значение параметра τ равно именно 4. Это и означает смещенность статистики t: $Mt \neq \tau$. Ясно, что у нас резко возросла бы вероятность попадания выборочной оценки параметра в окрестность точки x=4, если бы мы пользовались другой статистикой, расределение которой представлено на рис. 5.2 кривой F2 (x) (пунктирная линия).

Выборочное среднее является несмещенной оценкой генерального математического ожидания (точнее, следовало бы говорить, что среднее арифметическое дает несмещенные оценки, если полагать, что оценка – это конкретное значение статистики для выборки). Это следует из центральной предельной теоремы.

Если исходная совокупность симметрична, то несмещенной оценкой того же математического ожидания является и выборочное значение медианы. Если совокупность будет не только симметричной, но и унимодальной, то несмещенной оценкой математического ожидания явится и мода[45].

Для несмещенных оценок имеют смысл следующие определения.

Опр. Среднее квадратическое отклонение выборочного распределения статистики, отвечающей некоторому рассматриваемому параметру, будем называть *средней ошибкой выборки для оцениваемого параметра.*

Таким образом, для каждого оцениваемого параметра существует своя средняя ошибка выборки. Если же говорят о средней ошибке выборки вообще, то имеют в виду среднюю ошибку выборки для математического ожидания.

Известно много представляющихся естественными, но смещенных оценок. Так, вообще говоря, смещенной является оценка генерального коэффициента корреляции ρ между двумя случайными величинами, когда в качестве оценивающей статистики фигурирует выборочный коэффициент корреляции r между соответствующими признаками, определяемый по знакомой нам формуле

$$r = \sum_i \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n s_x s_y}$$

Несмещенной эта оценка является только в том случае, когда $\rho = 0$.

Смещенной является и оценка генеральной дисперсии с помощью расчета известной формулы:

$$D^2 = \frac{\sum_i (X_i - \bar{X})^2}{n}$$

Именно для того, чтобы сделать эту оценку несмещенной, в знаменателе указанной формулы пишут не n, а (n – 1) (несмещенной такая оценка будет для любой исходной совокупности) .

Чтобы еще раз показать, зачем же нужно стремиться к тому, чтобы используемая статистика давала нам именно несмещенную оценку параметра, рассмотрим распределения только что упомянутых оценок дисперсии. Рассмотрим рис. 5.3.

---

[45] См., например: Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. М.: Прогресс, 1976. С. 230

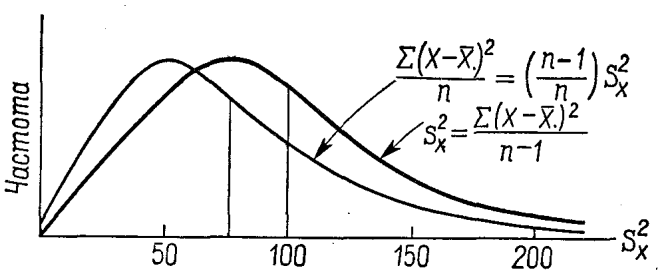


Рис. 5.3. Выборочные распределения величин ***Снова «частоту» надо заменить на «вероятность» !!!!!!!!!!!!!!!!! И убрать точку около икса с чертой***

$$(s'_x)^2 = \frac{\sum_i (X_i - \bar{X})^2}{n} \quad \text{и} \quad (s_x)^2 = \frac{\sum_i (X_i - \bar{X})^2}{n-1}$$

для случайных выборок объема 6 из нормального распределения с дисперсией $\sigma^2 = 100$[46]

Среднее, отвечающее распределению величины s_x^2 , равно 100, т.е. интересующему нас значению генерального параметра. Среднее, отвечающее величине $(s'_x)^2$, *смещено* относительно значения генерального параметра: оно равно 83,3. Величина подобного смещения может быть измерена с помощью коэффициента $\frac{n-1}{n}$, поскольку именно с помощью этого коэффициента $(s'_x)^2$ выражается через s_x^2:

$$(s'_x)^2 = \frac{n-1}{n} s_x^2$$

.В данном случае величина смещения довольно большая: она равна

$$\frac{n-1}{n} = 5/6 \ (= 83{,}3 : 100)$$

Опр. Оценка параметра называется *состоятельной*, если при увеличении объема выборки ее значение приближается к значению генерального параметра, который она оценивает:

$$\lim_{n\to\infty} P(|t_n - \tau| < \varepsilon) = 1.$$

Нетрудно понять смысл требования состоятельности. Если оценка не является состоятельной, то у нас не будет гарантии того, что увеличение объема выборочной совокупности приближает нашу оценку к «генеральному» значению изучаемого параметра (должно быть справедливым положение: чем больше объем выборки, тем ближе наша выборочная оценка генерального параметра к его истинному значению) .

Среднее арифметическое – состоятельная оценка математического ожидания. И следует это из закона больших чисел (см. приведенную выше формулировку частного случая теоремы Чебышева).

Если несмещенность и состоятельность – понятия абсолютные (относительно каждой статистики в принципе можно сказать, смещена она или не смещена, состоятельна или нет), то эффективность – понятие относительное: можно говорить только о том, что одна статистика более эффективна, чем другая. Более эффективной обычно считается та статистика, которая имеет меньшую дисперсию своего выборочного распределения.

Для примера упомянем, что выборочные мода и медиана являются несмещенными и состоятельными оценками математического ожидания (более точно, для медианы несмещенность имеет место только в случае симметричности генерального распределения, а для моды – для

[46] Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. М.: Прогресс, 1976. С.229

симметричного и унимодального)[47]. Но они менее эффективны, чем среднее арифметическое. Так, дисперсия ошибки выборочной медианы (т.е. дисперсия выборочного распределения медианы) равна $\frac{\pi}{2}\frac{\sigma^2}{n}$, т.е. примерно в 1,57 раз больше дисперсии среднего арифметического.

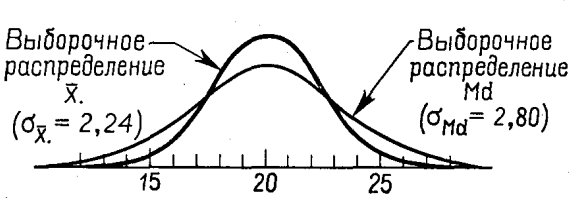


Рис. 5.4. Выборочное распределение среднего $\overline{X}$ и медианы Me для случайных выборок объема 10 из нормальной совокупности со средним μ=20 и дисперсией σ^2 = 50

(см. [Гласс и Стэнли, 1976, с. 232]) **Md заменить на Me !!!!!!!!!!**

Убрать точку около икса с чертой

Смысл требования эффективности тоже представляется очевидным. Если одна оценка (статистика) менее эффективна, чем другая, то, взяв значение первой (вычисленное для нашей одной - единственной выборки), мы имеем больший шанс «промахнуться», получить значение, сильно отличающееся от значения соответствующего генерального параметра.

Заметим, что, пользуясь вычисленной для выборки относительной частотой встречаемости того или иного интересующего нас события (скажем, тем, что в выборке мы имеем 40% женщин) мы фактически полагаем, что эта частота является хорошей точечной оценкой соответствующей генеральной вероятности (в нашем случае – полагаем, что эта вероятность близка к 0,4). О том, какова средняя ошибка выборки для доли (т.е. какова дисперсия выборочного распределения этой статистики) см. ниже.

ТЕМА 6
Интервальное оценивание параметров

6.1. Понятие доверительного интервала и принципы его построения (на примере математического ожидания)

Рассмотрим какой-нибудь параметр распределения изучаемой случайной величины, например, математическое ожидание, и попытаемся понять, каким образом можно судить о его значении на основе знания соответствующей выборочной точечной оценки, т.е. найденного для выборки среднего арифметического рассматриваемого признака.

Зададимся некоторой вероятностью α (обычно α = 0,05; подробнее об этой величине будет сказано ниже). Можно утверждать, что существует такое Δ, для которого имеет место соотношение:

$$P(\overline{X} - \Delta \le \mu_x \le \overline{X} + \Delta) = 1 - \alpha, \qquad (6.1)$$

Интервал вида (1) называется *доверительным*.

Чтобы понять, как находится Δ, напомним, что среднее арифметическое $\overline{X}$ для гипотетического бесконечного количества выборок имеет распределение $N(\mu_x, \frac{\sigma_x}{\sqrt{n}})$.

Вспомним теперь, что такое стандартизованное нормальное распределение, и попытаемся понять, как оно связано с нестандартизованным распределением случайной величины $\overline{X}$ (напомним, что значений средних мы рассматриваем как реализации некоторой случайной величины).

[47] Там же, с. 230

Нетрудно видеть , что величина

$$Z = \frac{\overline{X} - \mu_x}{\sigma_x / \sqrt{n}} \qquad (6.2)$$

имеет стандартизованное нормальное распределение.

Если мы зададимся целью найти тот интервал, в который попадает, скажем, 95% значений стандартизированной нормально распределенной величины, то, пользуясь известной таблицей, быстро установим, что этот интервал имеет вид

(-1,96; +1,96).

Использовав это обстоятельство применительно к величине (2), получим, что 95% значений этой величины удовлетворяет соотношению:

$$-1{,}96 \le \frac{\overline{X} - \mu_x}{\sigma_x / \sqrt{n}} \le 1{,}96$$

Значит, 95% значений случайной величины X удовлетворяет условию

$$\overline{X} - 1{,}96 \frac{\sigma_x}{\sqrt{n}} \le \mu_x \le \overline{X} + 1{,}96 \frac{\sigma_x}{\sqrt{n}} \qquad (6.3)$$

или, что то же самое,

$$\overline{X} - 1{,}96\, \sigma_{\overline{X}} \le \mu_x \le \overline{X} + 1{,}96\, \sigma_{\overline{X}}$$

Опр. Интервал (6.3) называется *95-%м доверительным интервалом для математического ожидания.*

Если мы захотим, чтобы аналогичному условию удовлетворяло 90% выборочных значений среднего арифметического, то должны число 1,96 заменить на 1,64; для 99% должны использовать множитель 2,57 и т.д.

К соотношению типа (6.3) можно придти и по-другому.

Рассмотрим рис. 1. Теоретически мы знаем, что Р % (выше – 95%) средних арифметических, рассчитанных для разных выборок, лежит вокруг μ_x в интервале, обозначенном овалом.

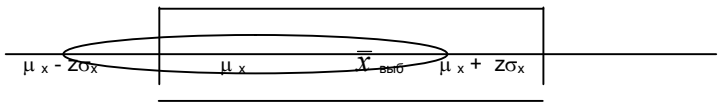
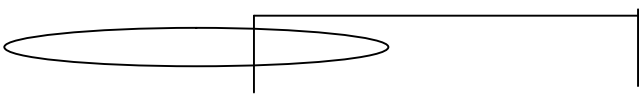


Рис.6.1. Ситуация, когда доверительный интервал (обозначен прямоугольником) «накрывает» математическое ожидание. Овал отвечает тому интервалу , в который попадают Р % выборочных средних арифметических

Теперь представим себе реальную ситуацию. У нас имеется единственная выборка и единственное значение среднего арифметического, вычисленное для нее. Обозначим его $\overline{x}_{выб}$. Нам надо выяснить,где находится μ_x. На помощь приходит соображение о том, что $\overline{x}_{выб.}$, очевидно, с вероятностью Р% попадает в «овальный» интервал. Поэтому, вероятно, логично было бы предположить, что μ_x с такой же вероятностью попадет в интервал такого же размера, но с центром не в μ_x, а в $\overline{x}_{выб}$. Этот интервал обозначен на рисунке прямоугольником. С помощью этого интервала мы можем с вероятностью Р% «поймать» математическое ожидание. Ясно, что это – интервал типа (3) (для последнего Р = 95%).

Конечно, не исключено, что $\overline{x}_{выб}$ не попадет в «овальный» интервал. Тогда мы мы будем иметь ситуацию, отраженную на рис.2. Ясно, что в таком случае реальное математическое ожидание не попадет в построенный для него интервал, не будем нами «поймано».



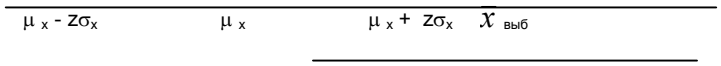


Рис. 6.2. Ситуация, когда математическое ожидание лежит вне доверительного интервала (последний обозначен прямоугольником. Овал отвечает тому интервалу , в который попадают Р % выборочных средних арифметических **СДЕЛАТЬ «ЗАРУБКИ» НА ОСИ**

Изобразим то же по-другому, прибегнув к изображению функции плотности распределения средних арифметических для выборок объема n, из генеральной совокупности с математическим ожиданием μ . Случаи, когда построенный по некоторому выборочному значению $\overline{X}$ доверительный интервал содержит , либо не содержит генеральное математическое ожидание, отражены, соответственно, на рисунка 6.3 и 6.4.

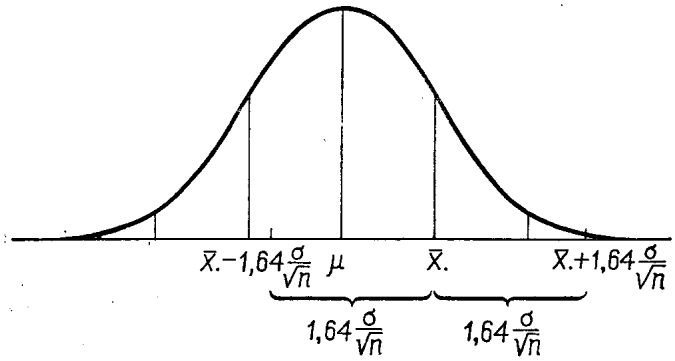


Рис. 6.3. Иллюстрация случая, когда интервал, установленный относительно $\overline{X}$, содержит μ в своих границах.[48] **НА РИСУНКЕ УБРАТЬ ТОЧКУ ПРИ Х с чертой. !!!!!!!!!!!!!!!!!!!**

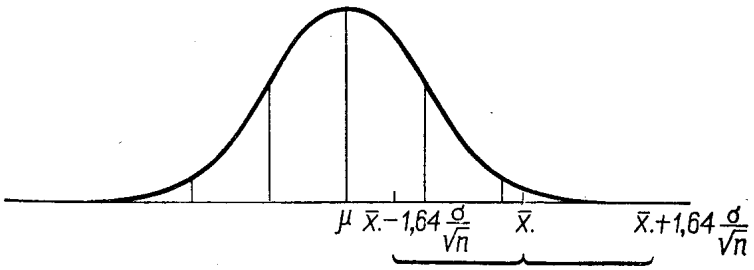


Рис. 6.4. Иллюстрация случая, когда интервал, установленный относительно $\overline{X}$, не содержит μ в своих границах.[49] **НА РИСУНКЕ УБРАТЬ ТОЧКУ ПРИ Х с чертой. !!!!!!!!!!!!!!!!!!!**

Возвращаясь к соотношению (6.1) и сравнивая его с (6.3), можно сказать, что для математического ожидания имеет место соотношение:

$$\Delta = z \frac{\sigma_x}{\sqrt{n}} \qquad (6.4)$$

Другими словами, соотношение (6.1) превращается в

[48] Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. М.: Прогресс, 1976. С. 235.
[49] Там же. С. 235.

$$P\left(\overline{X} - z\,\frac{\sigma_x}{\sqrt{n}} \le \mu_x \le \overline{X} + z\,\frac{\sigma_x}{\sqrt{n}}\right) = 1 - \alpha, \qquad (6.5)$$

где z определяется по таблице, исходя из выбранного α.

Опр. Интервал $(\overline{X} - \Delta, \overline{X} + \Delta)$, или, что то же самое, интервал

$$\left(\overline{X} - z\,\frac{\sigma_x}{\sqrt{n}}, \overline{X} + z\,\frac{\sigma_x}{\sqrt{n}}\right)$$

называется *доверительным интервалом* для μ_x.

Построение такого интервала - это и есть результат переноса сведений о выборочном среднем (коим является значение x) на генеральную совокупность.

Опр. α называется *уровнем значимости* доверительного интервала.

Как уже было сказано, он задается исследователем. Его выбор обуславливается содержательными соображениями. Чаще всего полагают, что $\alpha = 0{,}05$. Такой выбор означает, что 95-процентной уверенности в том, что генеральное ожидание принадлежит заданному интервалу, нам достаточно для того, чтобы считать это утверждение практически всегда верным. Другими словами, уровень значимости – это такая вероятность относительно которой мы предполагаем, что события, имеющие такую (или меньшую) вероятность, практически не происходят. Подчеркнем, что с оценкой подобной вероятности человек часто сталкивается в обыденной жизни. Именно на базе подобных оценок мы очень часто принимаем те или иные решения. К примеру, предположим, что по дороге на работу мы должны пройти мимо строящегося дома. Мы можем не давать себе в этом отчета, но где-то в подсознании у нас всегда будет происходить оценка вероятности того, что нам на голову свалится кирпич. Если нам случалось много раз проходить мимо этого дома без всяких неприятных последствий и мы никогда не слышали о том, что на кого-то что-то здесь свалилось, мы будем считать, что вероятность неприятности слишком мала для того, чтобы ее следовало принимать во внимание при принятии решения о нашем маршруте, и мы смело идем мимо стройки, не переходя на другую сторону улицы. В математической статистике обычно считается, что «слишком мала» означает «не более 5%». Напротив, если мы вчера прочитали в газете, что позавчера именно на этой стройке кирпич-таки свалился кому-то на голову[50], то мы, наверное, решим, что вероятность неприятности достаточно велика для того, чтобы ее надо было учитывать в своем поведении, и мы делаем крюк, чтобы обойти стройку, даже если опаздываем на работу. Опыт применения математической статистики говорит о том, что «достаточно велика» означает «превышает 5%».

Ясно, что, если суть задачи требует более надежной информации, то мы должны понизить уровень значимости, скажем, полагать, что он равен 0,01. Если, напротив, нас вполне устраивает меньшая уверенность, скажем, в 90%, то будем полагать, что $\alpha = 0{,}1$.

Мы вернемся к обсуждению смысла уровня значимости ниже, при рассмотрении способов проверки статистических гипотез (см. п. 7.2).

Значение z находится из таблицы нормального распределения. Величины z и α (а, стало быть, z и P) полностью определяют друг друга. Определение по таблице значения z для произвольного уровня доверительности интервала (величину разности (100 – уровень доверительности) надо поделить на 2, чтобы искать α).

Ясно, что исследователю всегда хочется, чтобы были поменьше и уровень значимости α (а P – побольше), и длина доверительного интервала (и, значит, z). Однако, к сожалению, законы природы так устроены, что уменьшение уровня значимости влечет за собой увеличение доверительного интервала. Поясним сказанное с помощью следующего рассуждения. Нетрудно понять, что, если X, к примеру, – возраст, выборочное среднее арифметическое значение которого оказалось равным 40 годам, то с вероятностью, практически равной 100% (т.е. $\alpha \approx 0$, математическое ожидание будет находиться в интервале (40 лет —100 лет, 40 лет + 100 лет). Однако от этой информации вряд ли может быть какая-либо практическая польза. Напротив, вероятность того, что генеральное математическое ожидание в той же ситуации в точности равно 40 годам (т.е. равен нулю доверительный интервал), практически нулевая (выборка всегда хотя бы в какой-то мере отличается

[50] Именно подобная ситуация, к великому нашему прискорбию, имела место прямо перед зданием ГУ-ВШЭ, где автор читала курс лекций по математической статистике. Поднимаемые кирпичи свалились на нескольких строителей высотного дома. Этот ужасный случай, будучи разобранным на лекции с точки зрения оценки описываемого уровня значимости, как нам кажется, способствовал лучшему усвоению студентами смысла этого показателя, а заодно и своеобразной математико-статистической логики рассуждений.

от генеральной совокупности, и поэтому выборочная статистика, как правило, будет отличаться от значения соответствующего генерального параметра).[51]

Отметим, что σ_x^2 социологу, как правило, неизвестно (хотя бывают ситуации, когда генеральную дисперсию признака удается хотя бы как-то оценить по каким-либо косвенным данным – скажем, воспользоваться результатами переписи, данными какого-то исследования, проведенного другим социологом и т.д.). Поэтому его вынуждены заменять выборочной дисперсией s_x^2 . Тогда, казалось бы, должно иметь место соотношение $s_{\bar{x}} = \frac{s_x}{\sqrt{n}}$ и, следовательно, равенство (6.4) заменяется на равенство $\Delta = z \frac{s_x}{\sqrt{n}}$. Однако это не так. Дело в том, что нормальное распределение при построении доверительного интервала для математического ожидания, вообще говоря, используется только при заданной генеральной дисперсии. В тех случаях, когда происходит замена σ_x^2 на s_x^2 нормальное распределение «превращается» в распределение Стьюдента. Коротко опишем, как в таких случаях надо действовать, не приводя строгих рассуждений, объясняющих описываемый алгоритм.

Если мы пользуемся выборочной оценкой s_x дисперсии признака , то доверительный интервал для μ_x приобретает вид:

$$(\bar{X} - t_{n-1} \frac{s_x}{\sqrt{n}}, \ \bar{X} + t_{n-1} \frac{s_x}{\sqrt{n}}),$$

где t – величина, найденная способом, аналогичным тому, с помощью которого мы искали z, но с использованием таблицы для распределения Стьюдента с числом степеней свободы, равным (n – 1). Другими словами, величина, полученная из (2) заменой σ_x на s_x, будет иметь не нормальное распределение, а распределение Стьюдента:

$$\frac{\bar{X} - \mu_x}{s_x / \sqrt{n}} \sim t_{n-1}$$

(Заметим, что нельзя пользоваться нормальным распределением и при малых объемах выборки, даже если σ_x известна; однако некоторая корректировка величины (2) все же приводит ее распределение к нормальному. А именно, нормально распределенной будет величина:

$$\frac{\bar{X} - \mu_x}{\sigma_x / \sqrt{n}} \sqrt{\frac{N-n}{N-1}}.$$

При бесконечной генеральной совокупности эта поправка не имеет смысла.)

Но, как мы отмечали при обсуждении темы 3 (п.3.4), при достаточно большом объеме выборки распределение Стьюдента можно считать приблизительно совпадающим с нормальным, поэтому при большой выборке можно пользоваться нормальным распределением (т.е. вместо t находить z) даже в том случае, когда генеральное значение σ_x мы вынуждены заменить на выборочную его оценку s_x.

Обычно применяют следующее правило[52].

[51] Смысл доверительных интервалов неплохо объясняется во многих книгах. См., например: Kachigan S.K. Statistical analysis. An interdisciplinar introduction to univariate and multivariate methods. - N.Y.: Radius Press,1986. С. 140-141.

[52] См., например: Bluman A.G. Elementary statistics. A step by step. McGraw-Hill Companies. 2th ed.C. 287.

Лишний раз повторим сказанное: относительно доверительного интервала для математического ожидания необходимо учитывать, что вид этого интервала зависит от того,

- известно ли σ;
- если σ неизвестно, то вместо него выступает s (которое рассчитывается либо по заданному набору значений признака, либо по частотной таблице), а величина z заменяется на t_{n-1};
- указанная в предыдущем пункте замена может не осуществляться, если объем выборки >30.

6.2. Определение объема выборки

Иногда ситуация складывается так, что мы из каких-либо содержательных (внешних по отношению к задаче построения доверительного интервала) соображений можем задать максимально возможное значение величины Δ, фигурирующую в определении доверительного интервала. В таком случае имеет смысл следующее определение.

Опр. Максимально возможное значение Δ называется *предельной ошибкой* выборки для признака x.

Нетрудно видеть, что предельная ошибка выборки - это та ошибка, которую мы допускаем, желая, чтобы генеральный показатель с заданным уровнем значимости находился в некотором заданном интервале.

Если и предельная ошибка выборки Δ, σ и уровень значимости будут заданы (а уровень значимости, как известно, определяет величину z), то приведенные выше соотношения позволят найти нижнюю границу объема выборки. Это следует из того, что о имеет место соотношение:

$$\frac{z^2\sigma_x^2}{n} \le \Delta^2,$$

и, следовательно, верно неравенство

$$n \ge \frac{z^2\sigma_x^2}{\Delta^2}.$$

(Отметим, что приведенные формулы лишь приблизительны. Ими можно пользоваться, если доля выборки в генеральной совокупности достаточно мала. Обычно считают, что должно выполняться соотношение

$$\mathrm{f} = \frac{n}{N} < 0{,}05.$$

Более точное соотношение для средней ошибки выборки имеет вид:

$$\mathrm{S_x} = \frac{s_x}{\sqrt{n}}\sqrt{1-f}\ .$$

При бесповторной выборке этого нельзя не учитывать.)

Отметим, что объем выборки может находиться только на основе известного генерального среднего квадратического отклонения σ (которую весьма нежелательно заменять на выборочную статистику S_x).

Важно подчеркнуть, что для социолога может быть очень мало пользы от описанного способа определения объема выборки. Дело в том, что объём, определенный таким образом, позволяет обеспечить представительность выборки в отношении только одного признака (и только в отношении одной его характеристики – среднего арифметического значения; как было отмечено выше своя ошибка выборки имеется и у других параметров рассматриваемых распределений). Поэтому такой признак выступает в роли "главного" для исследователя. Конечно, можно было бы таким же образом определить объём, принимая во внимание любое количество признаков: найти много значений n и

считать, что искомый объём должен быть равен максимальному из них. Но положение гораздо сложнее. Заранее, до сбора и анализа данных социолог обычно не имеет "в руках" того признака, который действительно можно считать "главным". Такого рода признаки обычно являются латентными, их значения находятся только в результате обработки некой первичной информации. И вопрос о том, какой должна быть выборка для сбора этой первичной информации, по существу остаётся открытым до проведения исследования.

6.3. Доверительный интервал для медианы.

Об ошибке выборки для медианы мы уже говорили при сравнении эффективности среднего арифметического и медианы как точечных оценок генерального математического ожидания. Распределение выборочных значений медиан является нормальным:

$$Me_{выб} \sim N(\mu_x, \sqrt{\frac{\pi}{2}}\frac{\sigma}{\sqrt{n}}).$$

Значит, в соответствии с описанными выше принципами, доверительный интервал для медианы выглядит следующим образом:

$$Me_{выб} - z\sqrt{\frac{\pi}{2}}\frac{\sigma}{\sqrt{n}} \le Me_{ген} \le Me_{выб} + z\sqrt{\frac{\pi}{2}}\frac{\sigma}{\sqrt{n}}.$$

Выражение $\Delta = z\sqrt{\frac{\pi}{2}}\frac{\sigma}{\sqrt{n}}$ называется предельной ошибкой выборки для медианы, если оно представляет собой определенное из содержательных соображений максимально возможное (естественно, это выражение носит вероятностный характер, задается определенный уровень значимости α, которому в формуле отвечает значение z) отклонение выборочных медиан от генеральной.

Отметим, что правило, в соответствии с которым при отсутствии сведений о генеральной дисперсии мы заменяем σ на s , а если к тому же и объем выборки мал (не больше 30 единиц), то вместо z используем t_{n-1}, действует при построении доверительного интервала для медианы так же, как и при построении аналогичного интервала для математического ожидания.

6.4. Доверительный интервал для доли

Доверительные интервалы могут быть построены отнюдь не только для генеральной средней и медианы, но и для многих других параметров распределений: дисперсии, коэффициента корреляции и т.д. Мы рассмотрим с соответствующей точки зрения ешё только одну статистику: долю встречаемости какого-либо значения одного из рассматриваемых признаков. Смысл соответствующей содержательной задачи представляется ясным. Представим, скажем, что доля мужчин в выборке оказалась равной 54%. Встает вопрос о какой-то оценке этой доля в генеральной совокупности. Как и выше, ответ на этот вопрос будет дан с помощью построения доверительного интервала для генеральной доли (т.е. – вероятности встречаемости свойства «быть мужчиной» среди объектов изучаемой генеральной совокупности).

Обозначения: p - упомянутая доля для выборки, π - для генеральной совокупности, $q = 1-p$, s_p - средняя ошибка выборки для доли p. Для любого уровня значимости α можно найти такое z, что будет справедливым соотношение:

$$P(p - \Delta \le \pi \le p + \Delta) = 1 - \alpha, \text{ где}$$

$$\Delta = z\sqrt{\frac{pq}{n}}, \text{ т.е.}$$

$$P\left(p - z\sqrt{\frac{pq}{n}} \le \pi \le p + z\sqrt{\frac{pq}{n}}\right) = 1 - \alpha \qquad (6)$$

Заметим, что величина $S_p = \sqrt{\frac{pq}{n}}$ - это средняя ошибка выборки для доли p (т.е. средний разброс таких долей, вычисленных для всех мыслимых выборок). Если как-то удастся определить максимально возможную величину Δ, то, по аналогии с соответствующими рассмотрениями выше, эта величина будет называться *предельной ошибкой выборки для доли.*

Подчеркнем, что значения p и q обычно рассчитываются для выборки, хотя в идеале здесь тоже (как и в случае расчета доверительного интервала для математического ожидания) должны быть генеральные показатели.

Формула для вычисления S_p фактически совпадает с формулой для вычисления S_x при определенном взгляде на принятое во внимание значение рассматриваемого признака. Рассмотрим этот аспект более подробно, поскольку этот факт имеет довольно принципиальное значение для выработки подходов к анализу социологических данных..

6.5. Связь средних ошибок среднего арифметического и доли, обобщение этого факта на многомерный анализ

Предположим теперь, что выделенному значению «а» рассматриваемого признака (тому, доля встречаемости которого равна p; в приведенном выше примере значение «а» означает «мужчина») поставлен в соответствие некоторый специальным образом построенный номинальный дихотомический признак Y:

$$Y=\begin{cases} 1, \text{ для респондентов, обладающих значением «а»,} \\ 0, \text{ для остальных респондентов.} \end{cases} \qquad (6.7)$$

Можно показать, что формула (6.5) для такого признака превращается в формулу (6.6)

Этот факт, несмотря на свою очевидность, очень важен для социолога. Его можно обобщить и на этой основе открыть путь широкому использованию традиционных количественных математико-статистических методов для изучения номинальной информации. С этой целью осуществляют так называемую дихотомизацию номинальных данных: каждому значению номинального признака только что описанным способом ставят в соответствие дихотомический номинальный признак, принимающий два значения: 0 и 1 (таким образом вместо одного номинального признака, имеющего k значений, появляется k новых дихотомических признаков). Можно показать, что в результате применения традиционных «числовых» методов к такого рода признакам, получается нечто осмысленное, если определенным образом разумно интерпретировать результаты (так, результатом решения задачи 3, приведенной после лекции 2, должны получиться соотношения, говорящие о том, как можно разумно интерпретировать среднее арифметическое значение и дисперсию такого дихотомического признака: $\bar{x}$ = p, Dx = pq, где p – доля единичных значений признака, а q – нулевых). Подобный подход широко используется в регрессионном анализе с фиктивными переменными (коими называются описанные выше дихотомические признаки), в логистической регрессии и т.д. Интерпретация коэффициентов уравнений регрессии при этом своеобразна, не похожа на то, что имеет место в классическом «числовом» случае.

Примеры задач.

Рекомендация. Если речь идет о выборе меры средней тенденции (для нас выбор ограничивается двумя средними - $\bar{x}$ и Me), то следует учитывать тип используемых шкал.

1. Доказать, что для признака вида (7) формула (5) превращается в формулу (6)

2. Министерство образования намеревается с помощью специального теста оценить средний уровень профпригодности молодых социологов - выпускников профильных вузов. Некоторые предварительные исследования позволяют считать, что среднее квадратическое отклонение, характеризующее разброс значений теста, равно 3,5. Какого объема выборку надо использовать, если исследователи хотят, чтобы с вероятностью 97% найденное среднее не отклонялось от генерального более, чем на 1,5?

3. Респонденты некоторой выборочной совокупности следующим образом распределились по возрасту

Возрастной интервал	Количество респондентов, попавших в

	интервал
15-20	20
20-25	40
25-30	40
30-35	60
35-40	40

Найти 92%-й доверительный интервал для математического ожидания и 97%-й доверительный интервал для доли людей, попавших по возрасту в интервал (25-30) лет.

4. На основе изучения выборки из 100 абитуриентов, прошедших тестирование, был подсчитан средний балл. Он оказался равным 7,8. Известно, что $\sigma = 2,0$. Найти 93%-й доверительный интервал для генерального среднего.

5. 10 случайно отобранных абитуриентов, поступающих в некоторый вуз, получили следующие баллы (использовалась интервальная шкала) на вступительных экзаменах:

5, 2, 1, 2, 3, 7, 5, 5, 6, 4.

Каков тот интервал, в котором с вероятностью 94% лежит генеральное математическое ожидание баллов, полученных всеми поступающими в вуз абитуриентами?

6. Предположим, что национальным меньшинством называется народность, составляющая менее 8% в общей совокупности жителей данной страны. В ходе выборочного опроса 2000 жителей страны 135 человек заявили, что являются бушменами. Можно ли с уверенностью 95% считать бушменов национальным меньшинством?

7. Респонденты некоторой выборочной совокупности были опрошены по шкале Лайкерта. Диапазон изменения установки был разбит на четыре интервала. Получилось следующее распределение.

Интервал изменения установки	Количество респондентов, попавших в интервал
10-15	50
15-20	20
20-25	40
25-30	40

Найти 96%-й доверительный интервал для медианы .

8. Результаты опроса некоторой совокупности респондентов по определенному тесту отражены в следующей таблице:

Значение теста	-2	1	2	3	4	5
Частота	2	1	2	2	2	1

Оценить с надежностью 0,95 математическое ожидание в соответствующей генеральной совокупности

9. Респонденты некоторой выборочной совокупности были опрошены по шкале Лайкерта. Диапазон изменения установки был разбит на четыре интервала. Получилось следующее распределение.

Интервал изменения установки	Количество респондентов, попавших в интервал
10-15	2
15-20	5
20-25	4
25-30	3

Найти 96%-й доверительный интервал для медианы .

10. Измеренная по шкале Лайкерта удовлетворенность 10-ти респондентов своей работой оказалась равной следующим величинам:

14, 10, 8, 21, 25, 21, 10, 16, 11, 10

Исследователь пока не определил, каков тип получившейся шкалы – порядковый или интервальный. Какие выводы мы можем сделать о средней удовлетворенности в генеральной совокупности в каждом из этих случаев ?

Раздел III. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

ТЕМА 7
Общее представление о статистической гипотезе. Проверка статистической гипотезы об отсутствии связи (критерий «Хи-квадрат»)

7.1. Общее представление о статистической гипотезе.

Начнем с примеров.

Первый пример.

Рассмотрим две дискретные переменные: X, принимающую значения из множества {1, …, r } и Y, принимающую значения из множества {1, …, c }. Мы будем их использовать как номинальные признаки, хотя вполне может быть, что их значения получены по шкалам более высоких типов.

Предположим, что нам задана частотная таблица вида $\| n_{ij} \|$, где i = 1, … , r (raw); j = 1,…,c (column), n_{ij} - количество объектов (например, респондентов), обладающих i –м значением признака X и j-м значением признака Y. Обозначим также через $n_{i\bullet}$ и $n_{\bullet j}$ маргинальные частоты (соответственно, по i – й строке и j- му столбцу), а через $n_{\bullet\bullet} = n$ - объем выборки.. Такую *таблицу называют частотной, или таблицей сопряженности.* Частоты, стоящие в клетках этой таблицы, назовем *эмпирическими, или наблюдаемыми.* Мы хотим на основе анализа эмпирических частот определить, имеется ли связь между рассматриваемыми переменными.

Здравый смысл подсказывает, что независимыми признаки можно считать в том случае, когда строки частотной таблицы пропорциональны[53]. Можно понятие независимости признаков отождествить и с другими свойствами частотной таблицы. Нетрудно проверить эквивалентность следующих утверждений:

1. Переменные X и Y являются независимыми.
2. Все частоты таблицы сопряженности являются теоретическими
3. Для всех i и j события (X = i) и (Y = j) являются независимыми.
4. Строки таблицы сопряженности пропорциональны.
5. Столбцы таблицы сопряженности пропорциональны.
6. Все частоты таблицы сопряженности вычисляются по формуле:

$$n_{ij} = \frac{n_{i\bullet} \bullet n_{\bullet j}}{n} \quad (7.2)$$

(7.1)

Предположим, что мы на основе собранной информации рассчитали частотную таблицу для некоторых двух переменных и хотим оценить, можно ли говорить о том, что связь между рассматриваемыми переменными отсутствует. Вопрос не так прост, как кажется на первый взгляд. Рассмотрим его подробнее.

Вспомним, что в действительности нас интересует генеральная совокупность, хотя имеющиеся в нашем распоряжении эмпирические данные, в том числе и таблица сопряженности, обычно отвечают выборке. Мы знаем, что выборочные данные никогда стопроцентно не отвечают генеральным. Любая, даже самая хорошая выборка будет отражать генеральную совокупность лишь с некоторым приближением, любая закономерность будет содержать так называемую выборочную ошибку. Случайную погрешность. Учитывая это, мы, вероятно, будем полагать, что если столбцы выборочной таблицы сопряженности мало отличаются от пропорциональных, то такое отличие, вероятнее всего, объясняется именно выборочной погрешностью и вряд ли говорит о том, что в генеральной совокупности наши признаки связаны. Сильное же отклонение от пропорциональности

---

[53] Подробнее об этом можно прочесть, например, в работе: Толстова Ю.Н. Анализа социологических данных. Методология, дескриптивная статистика, изучение связей номинальных признаков. М.: Научный мир, 2000].

заставит нас сомневаться в отсутствии связи в генеральной совокупности. Насколько же сильным должно быть такое отклонение для того, чтобы указанные сомнения у нас возникли? Наука не дает точного ответа. Она предлагает лишь такой вариант, который формулируется на вероятностном языке. Но прежде, чем ответить на поставленный вопрос, рассмотрим пример несколько иного плана.

Второй пример. Предположим, что в процессе решения социологической задачи мы хотим проверить гипотезу о том, что при оплате труда работников какого-то предприятия (отрасли и т.д.) нет дискриминации этих работников по полу. Это – *содержательная гипотеза*. Вероятно, наиболее естественными действиями исследователя, направленными на ее проверку, будет организация некоторой выборки из работников рассматриваемого предприятия и осуществление анкетного опроса с использованием, в частности, вопроса о поле респондента и его зарплате. Затем исследователь подсчитает среднюю зарплату мужчин и среднюю зарплату женщин. Обозначим зарплату буквой x и предположим, что получены соотношения (числа условны):

$$X_{жен} = 102,8; \; X_{муж} = 115,0.$$

Далее возможны разные рассуждения. Исследователь, пытающийся доказать отсутствие дискриминации, скажет: конечно, факт есть факт – средняя заплата женщин меньше средней зарплаты мужчин, но это различие очень мало. Наверное, его можно отнести за счет того, что мы взяли не всех работников, о только некоторую выборку из них. Другими словами, можно полагать, что наша статистика не дает оснований говорить о наличии дискриминации.

Другой исследователь, сторонник того, что дискриминация имеет место, выскажет совершенно твердую убежденность в своей правоте: статистические данные подтвердили его гипотезу, женщины в среднем получают меньше мужчин.

Кто прав? Где та граница, то значение разности зарплат, превышение которого говорит о том, что эти зарплаты действительно можно считать разными, что они отличны друг от друга не только в выборке, но и в генеральной совокупности?

Ответ получим, если воспользуемся логикой математической статистики, точнее, логикой проверки статистической гипотезы.. Ответ, конечно, будет носить вероятностный характер.

7.2. Логика проверки статистической гипотезы. Использование принципа невозможности реализации маловероятных событий

Приведенные в настоящем параграфе рассуждения, на наш взгляд, читателю было бы целесообразно прочитать дважды: сейчас и после прочтения дальнейших параграфов, посвященных описанию способов проверки конкретных гипотез.

Для примера заметим, что качестве проверяемой статистической гипотезы в описанной выше ситуации может фигурировать гипотеза о том, что в генеральной совокупности наши средние (т.е. математические ожидания зарплат для мужчин и для женщин) равны. Проверяемая гипотеза всегда обозначается H_0 и называется нуль-гипотезой. Заметим, что далеко не для каждой интересующей социолога гипотезы математическая статистика предоставляется возможность ее проверки, не для каждой гипотезы разработана соответствующая теория. Но если упомянутая возможность существует, то соответствующая логика рассуждений, коротко говоря, сводится к следующему.

Мы предполагаем, что для генеральной совокупности гипотеза верна. Изучаем выборку. Если выборочная ситуация резко отличается от того, что должно быть в генеральной совокупности при условии справедливости гипотезы, то гипотеза отвергается; если это отличие мало – гипотеза принимается (подчеркнем, что она не доказывается, а просто считается, что выборочные данные не дают оснований ее отвергнуть). Конечно, здесь возникают по крайней мере два вопроса: что значит «выборочная ситуация»? что значит «большое» или «малое» отличие выборочной ситуации от генеральной?

Прежде всего вспомним термины «параметр» и «статистика» и заметим, что выборочную ситуацию мы будем описывать с помощью некоторых статистик, в то время как проверяемая гипотеза будет касаться определенных предположений о характере параметров генеральных распределений изучаемых (одномерных и многомерных) случайных величин.

Предположим, что мы хотим проверить некую гипотезу H_0, для которой существует упомянутая выше теория. Последнее означает, что математическая статистика предлагает нам некоторый критерий, представляющий собой определенную статистику f - числовую функцию от наблюдаемых величин, например, рассчитанную на основе частот выборочной таблицы сопряженности :

$$f = f(n_{ij}).$$

Представим себе теперь, что у нас имеется много выборок (при доказательстве используемых нами положений математической статистики предполагается, что выборок - бесконечное количество), для каждой из которых вычисляется значение функции f. Распределение таких функций в предположении что H_0 верна, хорошо изучено, т.е. известно, какова вероятность попадания каждого значения в любой интервал. Грубо говоря, это означает, что для каждого полученного для конкретной выборки значения f , пользуясь соответствующей вероятностной таблицей, можно сказать, какова та вероятность, с которой мы могли на него «наткнуться».

Теперь необходимо пояснить, какого типа распределения будут нас интересовать. Мы уже говорили, что речь пойдет о нормальном распределении, о распределениях «χ^2», Стьюдента и F-распределении (распределении Фишера). «Маловероятными» для них всех являются области, лежащие в «хвостах» этих распределений. У первого распределения нас будет интересовать один «хвост» (правый), у трех других – два «хвоста» (и правый, и левый). Все названные распределения непрерывны. Значит, для них бессмысленно говорить о вероятности встречаемости точечного значения. И о вероятности «наткнуться» на конкретное значение $f_{выб}$ мы можем судить по одной из двух вероятностей: $P(f \geq f_{выб})$ (если $f_{выб} \geq 0$) или $P(f \leq f_{выб})$ (если $f_{выб} \leq 0$). Пока будем говорить о ситуации, когда $f_{выб} \geq 0$. Все, что будет по этому сказано, потом распространим на более общую ситуацию, специально посвятив этому параграф (см. ниже обсуждение вопроса о направленных и ненаправленных альтернативных гипотезах и об односторонних и двусторонних критериях в процессе рассмотрения темы 9).

Итак, вычисляем значение $f_{выб}$ критерия f для нашей единственной выборки. Находим по таблице вероятность $P(f \geq f_{выб})$.

Далее вступает в силу своеобразный принцип (уже затронутый нами в п.6.1) : *маловероятное событие практически не может произойти.* Другими словами, принимая практическое управленческое решение, мы, узнав, что некоторое событие имеет малую вероятность, будем вести себя так, как если бы это событие не могло произойти. Если такое маловероятное событие встречается в наших теоретических рассмотрениях, то мы делаем из этого вывод, что вероятность определялась нами неправильно, что в действительности рассматриваемое событие не маловероятно и что, следовательно, мы должны пересмотреть те положения, которые привели нас к выводу о незначительности величины вероятностиь его встречаемости мала.

Наше событие состоит в том, что критерий принял то или иное значение. Если вероятность этого события, (т.е. $P(f \geq f_{выб})$) очень мала, то, в соответствии с приведенными рассуждениями, мы полагаем, что неправильно ее определили. Встает вопрос о выяснении того, что именно привело нас к ошибке. Вспоминаем, что мы определяли упомянутую вероятность в предположении справедливости проверяемой гипотезы. Именно это предположение и заставило нас считать вероятность встреченного значения очень малой. Поскольку опыт дает основание полагать, что в действительности вероятность не столь мала, то остается опровергнуть нашу H_0.

Другими словами, если выборочная ситуация такова, что ее возникновение при справедливости в генеральной совокупности проверяемой гипотезы H_0 имеет очень малую вероятность, то гипотеза отвергается. Мы проанализировали выборку (вычислили статистику f) и увидели, что произошло событие, которое при условии справедливости H_0 можно считать маловероятным (статистика приняла маловероятное значение). Поскольку, в соответствии с обсуждаемым принципом, мы полагаем, что подобное событие произойти не может, то вынуждены допустить, что неверно то условие, при выполнении которого это событие маловероятно, т.е. неверна наша H_0. Другими словами, мы отвергаем нашу гипотезу.

Если же вероятность $P(f \geq f_{выб})$ достаточно велика для того, чтобы значение $f_{выб}$ могло встретиться практически, то мы полагаем, что у нас нет оснований сомневаться в справедливости проверяемой гипотезы. Мы принимаем последнюю, считаем, что она справедлива для генеральной совокупности.

Другими словами, если же анализируемая нами выборочная статистика приняла значение, вероятность появления которого при условии справедливости H_0 достаточно велика, то мы полагаем, что выборочная ситуация не противоречит проверяемой гипотезе. Мы эту гипотезу принимаем.

Таким образом, право именоваться критерием функция f обретает в силу того, что именно величина ее значения играет определяющую роль в выборе одной из двух альтернатив: принятия гипотезы H_0 или отвержения ее.

Правда, здесь снова возникает субъективный момент, связанный с неясностью того, какую вероятность мы назовем малой. Где граница между «малой» и «достаточно большой» вероятностью? Эта граница должна быть равна такому значению вероятности, относительно которого мы могли бы считать, что событие с такой (или с меньшей) вероятностью практически не может случиться – «не может быть, потому что не может быть никогда». Это значение называется *уровнем значимости принятия (отвержения) проверяемой гипотезы* и обозначается всегда греческой буквой α.

Итак, если вероятность $P(f \geq f_{выб}) > \alpha$, то мы гипотезу принимаем на уровне значимости α, если $P(f \geq f_{выб}) \leq \alpha$ - отвергаем на том же уровне значимости.

Иногда используется немного другая логика проверки. Мы задаемся уровнем значимости α и заранее ищем то значение критерия, обозначаемое обычно символом $f_{крит}$ (критическое) или $f_{табл}$ (табличное), для которого имеет место соотношение $P(f \geq f_{крит}) \leq \alpha$ и, вычислив $f_{выб}$, сравниваем его с $f_{крит}$: если $f_{выб} \geq f_{крит}$, то проверяемая гипотеза отвергается, если $f_{выб} \leq f_{крит}$, то принимается. Именно этой логики мы будем придерживаться ниже.

Ниже иногда будем использовать обозначение ${}_{\alpha}f$ (${}_{\alpha}f_{крит}$) в знак того, что речь идет о том табличном (критическом) значении, которое отвечает именно уровню значимости α.

Математическая статистика не дает нам правил определения α. Помочь установить уровень значимости может только практика. Обычно полагают, что

$$\alpha = 0{,}05.$$

В основе такого выбора не лежит никакая теория. Единственное его «оправдание» состоит в том, что, как показывает практика, если при проверке гипотез пользоваться таким уровнем значимости, то решения, принимаемые на основе проверки рассматриваемой гипотезы, как правило, оправдываются.

Однако, как мы увидим ниже при проверке конкретных гипотез, соответствующий уровень зачастую бывает целесообразно связывать с содержанием задачи. Он должен обусловливаться тем, насколько социально значимым может быть принятие ложной или отвержение истинной гипотезы (процесс проверки любой статистической гипотезы всегда сопряжен с тем, что мы рискуем совершить одну из упомянутых ошибок; ниже этот вопрос будет рассмотрен более подробно). Если большие затраты (материальные или духовные) связаны с отвержением гипотезы, то мы будем стремиться сделать α как можно меньше, чтобы максимально уменьшить вероятность отвержения нуль-гипотезы, являющейся в действительности верной.

О ситуации, когда затраты сопряжены с принятием гипотезы, и когда, следовательно, мы должны сделать так, чтобы минимизировать вероятность принятия неверной гипотезы, мы будем говорить ниже (мы имеем в виду обсуждение ошибок первого и второго рода; принятие неверной гипотезы – это ошибка второго рода; вероятность ее осуществления связана с понятием мощности критерия, о которой мы пока говорить не будем).

Перейдем к рассмотрению проверки конкретной нуль-гипотезы.

7.3. Проверка гипотезы об отсутствии связи между номинальными признаками на основе критерия «Хи-квадрат».

Вернемся к рассмотрению частотной эмпирической таблицы. Будем искать ответ на вопрос о существовании связи между признаками с помощью проверки статистической гипотезы об их независимости. Используя терминологию математической статистики, можно сказать, что речь пойдет о проверке нуль-гипотезы:

H_0: «связь между рассматриваемыми признаками отсутствует».

Функция, выступающая в качестве описанного выше статистического критерия, носит названиие «Хи-квадрат», обозначается как X^2 (X большое греческое «Хи»; подчеркнем, что дадлее будет фигурировать малая буква с тем же названием; и надо различать понятия, стоящие за этими обозначениями, что не всегда делается в ориентированной на социолога литературе). Определяется этот критерий следующим образом:

$$X^2 = \sum_{i,j} \frac{(n_{ij}^{теор} - n_{ij}^{эмп})^2}{n_{ij}^{теор}} \qquad (7.2)$$

где $n_{ij}^{эмп}$ - эмпирическая, наблюдаемая нами частота, стоящая на пересечении i –й строки и j –го столбца таблицы сопряженности; $n_{ij}^{теор}$ - та частота, которая стояла бы в той же клетке, если бы наши переменные были статистически независимы, т.е. та, которая отвечает пропорциональности столбцов (строк) таблицы сопряженности; она обычно называется теоретической, поскольку может быть найдена из теоретических соображений (см. формулу (7.2)); иногда ее называют также ожидаемой частотой, поскольку, действительно, ее появление ожидается при независимости переменных.

В соответствии со сказанным в предыдущем параграфе, представим себе, что мы организуем (теоретически) бесконечное количество выборок, для каждой из которых вычисляем величину X^2. Образуется последовательность таких величин:

$$X^2_{выб1}, X^2_{выб2}, X^2_{выб3,...} \quad (7.3)$$

Очевидно, имеет смысл говорить о соответствующем распределении, т.е. о вероятности попадания вычисленного для какой-либо выборки значения «Хи-квадрата» в тот или иной интервал. В математической статистике доказано следующее положение: если наши признаки в генеральной совокупности независимы, то величины (7.3) имеют хорошо изученное распределение, называемое «χ^2 – распределение». С ним мы уже знакомы (здесь используется малое греческое «хи»). Приблизительность можно игнорировать (т.е. считать, что величины (7.3) в точности распределены по закону χ^2), если ожидаемые (теоретические) частоты достаточно велики – обычно считают, что в каждой клетке таблицы, заполненной теоретическими частотами, должно быть по крайней мере 5 наблюдений. Будем считать, что это условие соблюдено (если это не так, то какие-то значения хотя бы одного из признаков следует объединить, чтобы соответствующие строки (столбцы) таблицы сопряженности сложились и частоты вследствие этого увеличились бы (отметим, что такое укрупнение должно быть осмысленным; скажем, если мы укрупняем градации возраста, то вполне допустимо объединить интервалы (15-20) и (20-25), но вряд ли при решении какой бы то ни было задачи будет разумно соединить интервалы (15-20) и (65-70)).

Вспомним, что «χ^2 – распределение» не одно. Чтобы выделить конкретный интересующий нас вариант из соответствующего семейства распределений, необходимо задать число степеней свободы. Оно равно

$$df = (r-1)(c-1).$$

Чтобы логика проверки нашей нуль гипотезы стала более ясной, отметим, что при отсутствии связи в генеральной совокупности среди выборочных значений (7.3) будут преобладать значения, близкие к нулю: отсутствие связи означает близость эмпирических и теоретических частот и, следовательно, близость к нулю всех слагаемых из определения критерия X^2 *(7.2).* Большие значения критерия будут встречаться относительно редко и поэтому они будут маловероятны. Мы имеем только одно значение – то, которое вычислено для нашей единственной выборки. Обозначим его через $X^2_{выб.}$ В силу сказанного, большое значение этой величины должно приводить нас к выводу о наличии связи, малое – об ее отсутствии. Описанная выше логика проверки статистической гипотезы превращается в следующее рассуждение.

Вычислим число степеней свободы df и зададимся уровнем значимости α. Найдем в таблице распределения χ^2 такое значение $\chi^2_{табл}$ (называемое иногда критическим значение критерия и обозначаемое через $\chi^2_{крит}$), для которого выполняется неравенство:

$$P(\xi \geq \chi^2_{табл}) = \alpha$$

(ξ - обозначение случайной величины, имеющей распределениеχ^2 с рассматриваемым числом степеней свободы).

Если $X^2_{выб} < \chi^2_{табл}$ (то есть вероятность появления $X^2_{выб}$ при справедливости нуль гипотезы о независимости достаточно велика), то полагаем, что наши выборочные наблюдения не дают оснований сомневаться в том, что в генеральной совокупности признаки действительно независимы – ведь, «ткнув» в одну выборку, мы встретили такое значение X^2, которое действительно вполне могло встретиться при независимости. В таком случае мы полагаем, что у нас нет оснований отвергать нашу нуль гипотезу, посольку эмпирия ей не противоречит. Мы ее принимаем – считаем, что признаки независимы. Если же $X^2_{выб} \geq \chi^2_{табл}$, (то есть вероятность появления $X^2_{выб}$ очень мала, меньше α), то мы вправе засомневаться в нашем предположении о независимости – ведь мы «наткнулись» на такое событие, которое вроде бы не должно было встретиться при таком предположении. В таком случае мы отвергаем нашу нуль-гипотезу, полагаем, что признаки зависимы.

Итак, рассматриваемый критерий не гарантирует наличие связи. Не измеряет ее величину. Он либо говорит о том, что эмпирия не дает оснований сомневаться в отсутствии связи, либо, напротив, дает повод для сомнений.

В заключение нельзя не сказать об очень важном (и с практической, и теоретической точки зрения) моменте: и величина критерия X^2, и его расположение по отношению к табличному значению (естественно, говоря об этом, мы предполагаем, что уровень значимости зафиксирован) может измениться. Другими словами, наш вывод о наличии или отсутствии связи между переменными

зависит от способа группировки значений рассматриваемых признаков. Представляется, что этот факт является интересной иллюстрацией к ведущейся в литературе дискуссии по вопросу объективности знания, получаемого социологом! [54]. Отметим, что при разной группировке значений какого-либо признака мы по существу переходим к разным признакам, отражающим разные стороны реальности. Так, сгруппировав значения возраста (не учитываем детство) так: (15-20), (20-50), (50-80), мы по существу отразим физическое состояние организма человека: растущий организм, стабильный, деградирующий. А сгруппировав по-другому: (15-20), (20-30), (30-80), получим признак, отражающий степень социальной зрелости человека (мы не претендуем на содержательную правильность предлагаемых разбиений). [55]

Примеры задач

1.Доказать эквивалентность соотношений (7.1).

Рекомендация к дальнейшему. Если имеются клетки, которым отвечают теоретические частоты, меньшие 5, то градации признаков надо укрупнять (разумным способом). При этом надо учесть, что чем меньше укрупнений – тем лучше, поскольку мы теряем меньше информации (естественно, какждое «слияние» градаций того или иного признака приводит к потере информации).

2. Проанализировав данные по абитуриентам, поступавшим в некоторый вуз, получили следующие частоты:

Занимался ли на подготовительных курсах	Набранный балл			
	До 10	11-15	16-20	21-25
Да	5	10	10	15
Нет	10	5	10	5

Можно ли считать, что обучение на подготовительных курсах способствует более эффективной подготовке к экзамену?

Рекомендация. Данная задача может решаться двумя способами. Применить оба и содержательно проинтерпретировать разницу в выводах

3. Данные опроса жителей некоторого промышленного региона об их электоральном поведении были сведены в следующую частотную таблицу (частоты в клетках выражены в десятках тысяч человек):

Место жительства	Голосование за представителя партии			
	ЕР	ЛДПР	КПРФ	Яблоко
Город	40	15	16	8
Село	15	10	10	5

Можно ли считать выбор партии респондентом статистически связанным с местом его проживания?

---

[54] Мы не сомневаемся в объективности знания, полученного с помощью методов математической статистики. Однако в само понятие «знание» должны включаться не только приобретенные сведения, но и то, каким образом эти сведения были получены. К вопросу об объективности знания, полученного с помощью критерия «Хи-квадрат» и других математико-статистических приемов мы вернемся в п. 11.3.

[55] Здесь мы сталкиваемся с одной из самых острых проблем эмпирической социологии: построением признаков, адекватных измеряемым с их помощью качествам людей. Легко наблюдаемые признаки обычно являются признаками-приборами (это понятие введено в книге: Клигер С.А., Косолапов М.С., Толстова Ю.Н. Шкалирование при сборе и анализе социологической информации. М.: Наука, 1978). Другими словами, их значения нас интересуют не сами по себе, а как индикаторы каких-то латентных свойств. При работе с такими признаками важную роль играет удачная группировка их значений. Существует много методов поиска таких группировок: методы разбиения на интервалы диапазона значений непрерывных признаков, объединения значения дискретных признаков, поиска т.н. взаимодействий, т.е. сочетаний значений разных исходных признаков (к обсуждению последнего аспекта мы вернемся при рассмотрении темы 15).

Рекомендация. Пояснить, почему один из способов, пригодных для решения предыдущей задачи, здесь неприменим.

4. При анкетном опросе жителей Татарстана (сотрудников госпредприятий) фрагмент одной из частотных таблиц имел вид:

Должность	Национальность	
	Татары	Русские
Руководитель предприятия	17	10
Рядовой квалифицированный сотрудник	123	124
Чернорабочий	4	3

Можно ли сказать, что в республике имеется определенная дискриминация по национальности при назначении человека на должность в государственной предприятии?

5. Придумайте пример задачи, в которой вывод о наличии (или отсутствии) связи между двумя непрерывными признаками (при использовании критерия «Хи-квадрат») зависел бы от группировки значений признаков (уровень значимости предполагается заданным). Попробуйте объяснить этот феномен.

ТЕМА 8.
Проверка гипотезы о равенстве средних

8.1. Понятие зависимых и независимых выборок.

Выбор критерия для проверки гипотезы

$$H_0: \mu_1 = \mu_2$$

в первую очередь определяется тем, являются ли рассматриваемые выборки зависимыми или независимыми. Введем соответствующие определения.

Опр. Выборки называются *независимыми*, если процедура отбора единиц в первую выборку никак не связана с процедурой отбора единиц во вторую выборку.

Примером двух независимых выборок могут служить обсуждавшиеся выше выборки мужчин и женщин, работающих на одном предприятии (в одной отрасли и т.д.).

Заметим, что независимость двух выборок отнюдь не означает отсутствие требования определенного рода сходства этих выборок (их однородности). Так, изучая уровень дохода мужчин и женщин, мы вряд ли допустим такую ситуацию, когда мужчины отбираются из среды московских бизнесменов, а женщины – из аборигенов Австралии. Женщины тоже должны быть москвичками и, более того – «бизнесвуменшами». Но здесь мы говорим не о зависимости выборок, а о требовании однородности изучаемой совокупности объектов, которое должно удовлетворяться и при сборе, и при анализе социологических данных[56].

Опр. Выборки называются *зависимыми, или парными,* если каждая единица одной выборки «привязывается» к определенной единице второй выборки.

Последнее определение, вероятно, станет более ясным, если мы приведем пример зависимых выборок.

Предположим, что мы хотим выяснить, является ли социальный статус отца в среднем ниже социального статуса сына (полагаем, что мы можем измерить эту сложную и неоднозначно понимаемую социальную характеристику человека). Представляется очевидным, что в такой ситуации целесообразно отбрать пары респондентов (отец, сын) и считать, что каждый элемент первой выборки (один из отцов) «привязан» к определенному элементу второй выборки (своему сыну). Эти две выборки и будут называться зависимыми.

[56] Снова возвращаемся к проблеме однородности, уже затрагивавшейся нами в сноске 11и в п. 1.7.

8.2. Проверка гипотезы для независимых выборок

Для *независимых* выборок выбор критерия зависит от того, знаем ли мы генеральные дисперсии σ_1^2 и σ_2^2 рассматриваемого признака для изучаемых выборок. Будем считать эту проблему решенной, полагая, что выборочные дисперсии совпадают с генеральными. В таком случае в качестве критерия выступает величина:

$$z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_2} + \frac{\sigma_2^2}{n_2}}} \qquad (8.1)$$

Прежде, чем переходить к обсуждению той ситуации, когда генеральные дисперсии (или хотя бы одна из них) нам неизвестны, заметим следующее.

Логика использования критерия (8.1) похожа на ту, которая была описана нами при рассмотрении критерия "Хи-квадрат" (7.2). Имеется лишь одно принципиальное отличие. Говоря о смысле критерия (7.2), мы рассматривали бесконечное количество выборок объема n, «черпающихся» из нашей генеральной совокупности. Здесь же, анализируя смысл критерия (8.1), мы переходим к рассмотрению бесконечного количества *пар* выборок объемом n_1 и n_2. Для каждой пары и рассчитывается статистика вида (8.1). Совокупности получаемых значений таких статистик, в соответствии с нашими обозначениями, отвечает нормальное распределение (как мы условились, буква z используется для обозначения такого критерия, которому отвечает именно нормальное распределение).

Итак, если генеральные дисперсии нам неизвестны, то мы вынуждены вместо них пользоваться их выборочными оценками s_1^2 и s_2^2. Однако при этом нормальное распределение должно замениться на распределение Стьюдента – z должно замениться на t (как это имело место в аналогичной ситуации при построения доверительного интервала для математического ожидания). Однако при достаточно больших объемах выборок ($n_1, n_2 \geq 30$) , как мы уже знаем, распределение Стьюдента практически совпадает с нормальным. Другими словами, при больших выборках мы можем продолжать пользоваться критерием:

$$z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_2} + \frac{s_2^2}{n_2}}} \qquad (8.2)$$

Сложнее обстоит дело с такой ситуацией, когда и дисперсии неизвестны, и объем хотя бы одной выборки мал. Тогда вступает в силу еще один фактор. Вид критерия зависит от того, можем ли мы считать неизвестные нам дисперсии рассматриваемого признака в двух анализируемых выборках равными. Для выяснения этого надо проверить гипотезу:

$$H_0 : \sigma_1^2 = \sigma_2^2. \qquad (8.3)$$

Для проверки этой гипотезы используется критерий

$$F_{n_1-1,n_2-1} = \frac{\sigma_1^2}{\sigma_2^2}. \qquad (8.4)$$

О специфике использования этого критерия пойдет речь ниже, а сейчас продолжим обсуждать алгоритм выбора критерия, использующего для проверки гипотез о равенстве математических ожиданий.

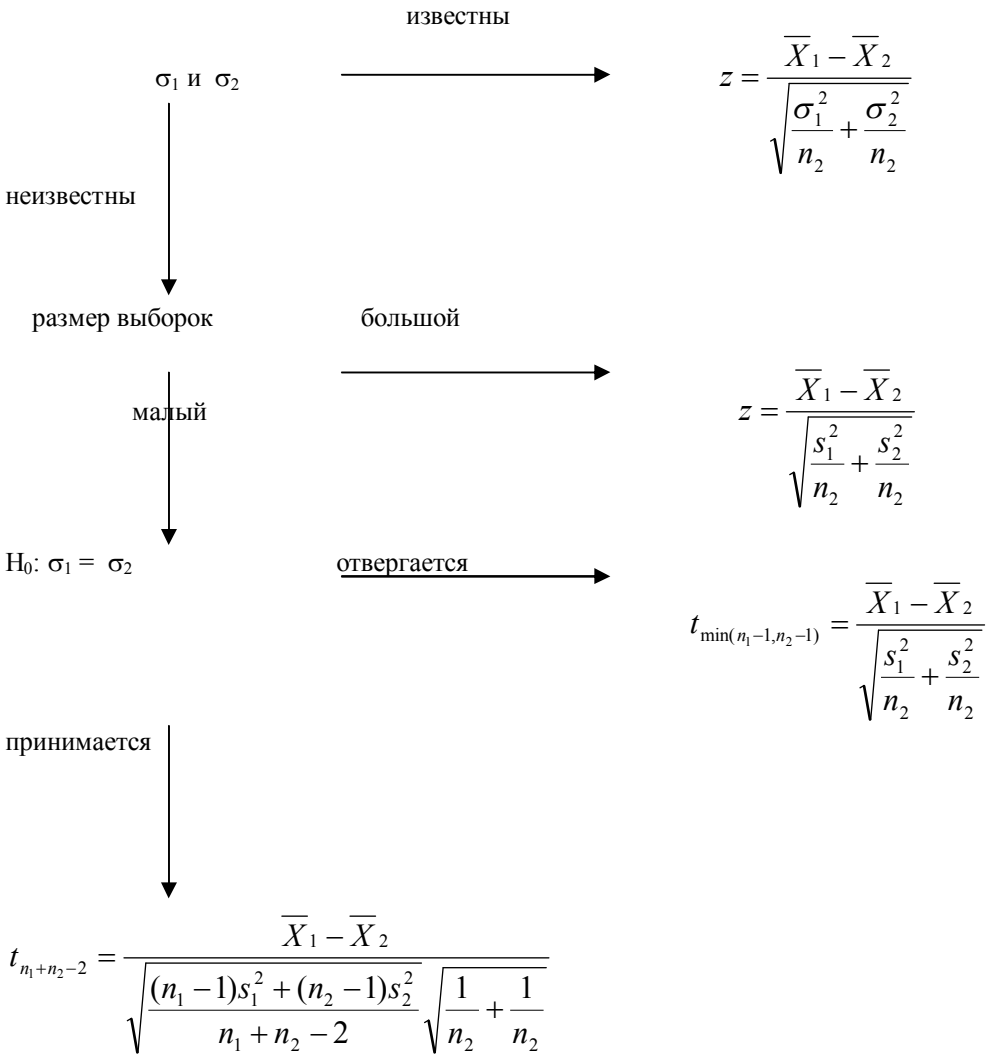
Если гипотеза (8.3) отвергается, то интересующий нас критерий приобретает вид:

$$t_{\min(n_1-1,n_2-1)} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad (8.5)$$

(т.е. отличается от критерия (8.2), использовавшегося при больших выборках, тем, что соответствующая статистика имеет не нормальное распределение, а распределение Стьюдента). Если гипотез (8.3) принимается, то вид используемого критерия меняется:

$$t_{n_1+n_2-2} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \qquad (8.6)$$

Подведем итог того, как выбирается критерий для проверки гипотезы о равенстве генеральных математических ожиданий на основе анализа двух независимых выборок.



8.3. Проверка гипотезы для зависимых выборок

Перейдем к рассмотрению зависимых выборок. Пусть последовательности чисел

$$X_1, X_2, \ldots, X_n;$$
$$Y_1, Y_2, \ldots, Y_n -$$

это значения рассматриваемой случайной для элементов двух зависимых выборок. Введем обозначение:

$$D_i = X_i - Y_i,\ i = 1, \ldots, n.$$

Для *зависимых* выборок критерий, позволяющий проверять гипотезу

$$H_0:\ \mu_1 = \mu_2$$

выглядит следующим образом:

$$t_{n-1} = \frac{\overline{D}}{s_D / \sqrt{n}}, \text{ где}$$

$$\overline{D} = \frac{\sum D_i}{n}$$

$$s_D = \sqrt{\frac{\sum D_i^2 - \frac{(\sum D_i)^2}{n}}{n-1}}$$

Заметим, что только что приведенное выражение для s_D есть не что иное, как новое выражение для известной формулы, выражающей среднее квадратическое отклонение. В данном случае речь идет о среднем квадратическом отклонении величин D_i. Подобная формула часто используется на практике как более простой (по сравнению с «лобовым» подсчетом суммы квадратов отклонений значений рассматриваемой величины от соответствующего среднего арифметического) способ расчета дисперсии.

Если сравнить приведенные формулы с теми, которые мы использовали при обсуждении принципов построения доверительного интервала, нетрудно заметить, что проверка гипотезы о равенстве средних для случая зависимых выборок по существу является проверкой равенства нулю математического ожидания величин D_i. Величина

$$s_D / \sqrt{n}$$

есть среднее квадратическое отклонение для D_i. Поэтому значение только что описанного критерия t_{n-1} по существу равно величине D_i, выраженной в долях среднего квадратического отклонения. Как мы говорили выше (при обсуждении способов построения доверительных интервалов), по такому показателю можно судить о вероятности рассматриваемого значения D_i. Отличие состоит в том, что выше шла речь о простом среднем арифметическом, распределенном нормально, а здесь – о средних разностей, такие средние имеют распределение Стьюдента. Но рассуждения о взаимосвязи вероятности отклонения выборочного среднего арифметического от нуля (при математическом ожидании, равном нулю) с тем, сколько единиц σ это отклонение составляет, остаются в силе.

Примеры задач

1. На основе обработки массива анкет была получена следующая частотная таблица:

Зарплата	Возраст	
	До 30 лет	Старше 30 лет
До 500	30	10
500-1000	20	10
1000-1500	20	15
1500-1200	10	25

Можно ли считать, что средняя зарплата молодых респондентов (моложе 30 лет) ниже средней зарплаты представителей более старшего возраста (старше 30 лет)? Пояснить статистический смысл ответа.

2. Имеем следующую статистику по регионам

№ региона	1	2	3	4	5
Уровень безработицы для мужчин	1,05	4,01	3,2	7,08	3,01
Уровень безработицы для женщин	1,02	5,005	3,05	2,03	3,1

Можно ли считать, что среди мужчин уровень безработицы в среднем выше?

3. В результате замеров верхнего давления респондентов были получены следующие данные:

№ респондента	Верхнее давление в спокойном состоянии	Верхнее давление при прослушивании концерта тяжелого рока
1	120	110
2	110	130
3	100	120
4	130	130
5	110	130

Можно ли считать, что прослушивание концерта тяжелого рока в среднем повышает верхнее давление? Пояснить статистический смысл ответа.

ТЕМА 9.
Направленные и ненаправленные альтернативные гипотезы. Односторонние и двусторонние критерии

Теперь, когда мы познакомились с техникой проверки некоторых математико-статистических гипотез, имеет смысл коснуться очень важного момента, обусловливающего способ поиска табличного значения используемого критерия. Сделаем это прежде, чем переходить к рассмотрению других гипотез, поскольку при таком рассмотрении соответствующие положения имеет смысл активно использовать.

9.1. Направленные и ненаправленные альтернативные гипотезы.

Наряду с каждой сформулированной выше нуль-гипотезой исследователь обычно формулирует и т.н. альтернативную гипотезу H_1 – то утверждение, которое он будет считать верным при отказе от H_0. При рассмотрении некоторых гипотез формулировка H_1 очевидна и говорить об альтернативной гипотезе не стоит. Это касается, например, той гипотезы об отсутствии связи, которую мы проверяли с помощью критерия "Хи-квадрат". Ей противостоит единственно возможная альтернативная гипотеза, утверждающая, что связь между переменными имеется. Подобная альтернативная гипотеза называется *ненаправленной.*

Однако не такова ситуация, возникающая при проверке статистической гипотезы о равенстве математических ожиданий. Гипотезе

$$H_0: \mu_1 = \mu_2$$

противостоят два альтернативных утверждения:

$$H_1: \mu_1 \neq \mu_2.$$

и

$$H_1: \mu_1 > \mu_2 \text{ (или } H_1: \mu_2 > \mu_1)$$

В первом случае альтернативная гипотеза называется *ненаправленной*, а во втором – *направленной.* В первом случае мы полагаем, что при неравенстве математических ожиданий одинаково возможны и такая ситуация, когда первое среднее больше второго, и такая, когда второе среднее больше первого. Во втором случае (когда гипотеза - направленная) считаем, что что только первое среднее может быть больше второго. Подобное предположение может объясняться либо чисто содержательными соображениями, когда вторая ситуация, не отраженная в нашей альтернативной гипотезе (второе среднее больше первого), просто не может возникнуть; либо тем, что вторая ситуация нас просто не интересует (например, нас может интересовать, имеется ли в какой-то отрасли промышленности дискриминация женщин по оплате; тогда мы будем проверять гипотезу о равенстве средних зарплат мужчин и женщин, ориентируясь на оценку вероятности того, что зарплата мужчин выше зарплаты женщин; ситуация же, когда средняя зарплата женщин выше средней зарплаты мужчин нас будет «волновать» ровно в той же мере, что и ситуация, когда указанные зарплаты равны).

За каждым видом выбранной альтернативы часто стоит свое понимание ситуации проверки гипотезы.

9.2. Односторонние и двусторонние критерии

Предположим, что мы проверяем гипотезу H_0: $\mu_1 = \mu_2$ и альтернативной гипотезой является гипотеза H_1: $\mu_1 \neq \mu_2$. Выбор альтернативной гипотезы означает следующее. При анализе описанной выше мысленной конструкции с бесконечным количеством пар выборок у нас могут встречаться и такие пары, для которых наш критерий (неважно, какой – (8.1), (8.2), (8.5) или (8.6)) положителен (т.е.первая средняя больше второй), и такие, для которых критерий отрицателен (вторая средняя больше первой). Соответственно, оценивая вероятность попадания конкретного выборочного значения критерия в тот или иной интервал, мы должны учитывать, что маловероятными областями являются не только правый конец оси, но и левый. Значит, маловероятная критериальная область, попадание в которую приведет нас к решению об отвержении гипотезы, распадется на две части. Уровень значимости α должен будет отвечать двум концам нормального распределения (этой величине должна быть равна сумма площадей и под правым, и под левым его концом). В таблицах же, как правило, указывается только такое табличное значение, которое отвечает правому концу. Чтобы это учесть, надо будет искать $z_{крит}$, отвечающее величине $\alpha/2$. И если наше $z_{выб}$ окажется отрицательным, то мы будем его сравнивать со значением ($- z_{крит}$). Другими словами, мы будем принимать проверяемую гипотезу, если

$$|z_{выб}| \leq z_{крит}$$

и отвергать ее, если $|z_{выб}| > z_{крит}$.

Совершенно аналогичные рассуждения справедливы и для того случая, когда используемый критерий имеет распределение Стьюдента. Соответствующие статистики тоже могут быть положительными и отрицательными, а распределение очень похоже на нормальное – тоже представляет собой симметричный «колокол».

При распределении F рассуждение несколько меняется. Соответствующие критерии всегда положительны (как мы увидим ниже, они «строятся» из дисперсий – статистик, не могущих принимать отрицательные значения). Тем не менее, они могут быть двусторонними. По таблице мы ищем $F_{крит}$ для правого конца распределения. Отвечать оно должно, как и выше, величине $\alpha/2$. А табличное значение для левого конца будет равно $(1/ F_{крит})$. И проверяемая гипотеза принимается, если

$$\frac{1}{F_{крит}} \leq F_{выб} \leq F_{крит},$$

и отвергается, если $F_{выб} > F_{крит}$ или $F_{выб} < \frac{1}{F_{крит}}$

Для критерия «Хи-квадрат» (и для других статистик, имеющих распределение χ^2) альтернативная гипотеза не направлена, но критерий всегда односторонен. Как мы увидим ниже, не так обстоит дело с другим показателем связи - коэффициентом корреляции, поскольку он, как известно, может быть и положительным, и отрицательным. Соответствующий критерий (а речь пойдет о распределении Стьюдента) может и односторонним, и двусторонним.

ТЕМА 10

Проверка статистических гипотез: о равномерности генерального распределения, о равенстве дисперсий, о равенстве нулю коэффициента корреляции, о равенстве двух долей; ошибки первого и второго рода. .

10.1. Проверка гипотезы о равномерности генерального распределения с помощью критерия Хи-квадрат

В социологических исследованиях нередко возникают ситуации, когда социологу надо знать, можно ли имеющееся у него частотное распределение, отвечающее какому-либо признаку, считать выборочным представлением распределения определенного характера: нормального, равномерного и

т.д. Так, многие математические методы требуют (если мы хотим, чтобы применение метода было корректным), чтобы распределение каждого из исходных признаков было нормальным. Выборочное частотное распределение может в какой-то степени походить на нормальное: скажем, быть одновершинным, в како-то степени симметричным и т.д.Можем ли мы считать, что в генеральной совокупности оно нормально? Что его отклонение от нормальности можно объяснить только случайными погрешностями выборки? Ясно, что подобная ситуация очень похожа на те описанные выше ситуации, когда мы отвевчали на вопросы, подобные только что сформулированным, с помощью проверки статистической гипотезы. Оказывается, что и здесь таким же образом можно найти ответ: математическая статистики предлагает нам соответствующий механизм.

Существует критерий, позволяющий проверять гипотезу о том, что распределение имеет определенный характер[57]. Это уже знакомый нам критерий «Хи-квадрат». В общем виде он имеет вид:

$$X^2 = \sum \frac{(E-O)^2}{E}$$

где O – эмпирические, наблюдаемые, фактические частоты, E – теоретические, ожидаемые частоты. Теоретические частоты – это те, которые мы видели бы (ожидали) в случае, если бы распределение имело характер, отвечающий проверяемой гипотезе.

Поясним сказанное на примере проверки гипотезы о равномерности генерального распределения. Пусть рассматриваемый признак принимает значения

$$1,2, \dots , k,$$

а фактически наблюдаемые (выборочные) частоты встречаемости этих значений равны, соответственно,

$$n_1^{эмп}, n_2^{эмп}, \dots , \quad n_k^{эмп}.$$

Отвечающие им теоретические частоты будут выглядеть так:

$$n_1^{теор} = n_2^{теор} = \dots = n_k^{теор} = \frac{n}{k}, n = \sum_{i=1}^{k} n_i^{эмп}$$

Интересующий нас критерий имеет вид:

$$X^2_{k-1} = \sum_i \frac{(n_i^{теор} - n_i^{эмп})^2}{n_i^{теор}}$$

Другими словами, отличие критерия, позволяющего нам проверить гипотезу о том. что генеральное распределение является равномерным, от того критерия Хи-квадрат, который выше мы использовали для проверки гипотезы об отсутствии связи, состоит в следующем:

- вместо $n_{ij}^{эмп}$ фигурирует $n_i^{эмп}$ – фактическая частота встречаемости i-го значения рассматриваемого признака;
- вместо $n_{ij}^{теор}$ выступает $n_i^{теор}$ – та частота, которая должна была бы соответствовать значению i рассматриваемого признака, если бы он имел проверяемое распределение;
- число степеней свободы равно не (c-1)(r-1), а (k – 1), где k – количество альтернатив в рассматриваемом признаке.

10.2. Проверка гипотезы о равенстве двух дисперсий

Вернемся к рассмотрению той статистической гипотезы, которую мы уже упомянули при обсуждением правил выбора критерия для проверки гипотезу о равенстве генеральных математических ожиданий - гипотезу о равенстве генеральных дисперсий:

$$H_0: \ \sigma_1^2 = \sigma_2^2.$$

Итак, предположим, что выборка объемом n_1 случайно извлекается из нормальной совокупности со средним μ_1 и дисперсией σ_1^2. Независимая случайная выборка объемом n_2 извлекается из второй нормальной совокупности со средним μ_2 и дисперсией σ_2^2. При проверке H_0 значения математических ожиданий несущественны.

Критериальная статистика $F = (s_1^2 / s_2^2)$ имеет распределение Фишера (в предположении справедливости нуль-гипотезы) со степенями свободы $(n_1 - 1)$ и $(n_2 - 1)$. В вероятностных таблицах для этого распределения, как правило, вверху указывается число степеней свободы для большей

[57] См., например: Рабочая книга социолога. М.: Наука, 1976 (переиздана в 2003 году). С. 170.

дисперсии, а слева – для меньшей (или вверху - число степеней свободы для числителя, а внизу – для знаменателя). По существу речь идет о проверке сходства степеней однородности двух выборок.

Критерий может быть двусторонним, поскольку интересующее нас отклонение F от 1 может быть и в правую, и в левую сторону. Значит, альтернативная гипотеза может быть ненаправленной и направленной.

Ненаправленная альтернативная гипотеза выглядит так:

$$H_1 : \sigma_1 \neq \sigma_2.$$

Критерий в таком случае – двусторонний. Находим $_{\alpha/2}F_{(n1-1),(n2-1)}$ по таблице (в таблице обычно дают только те значения критерия, которые больше 1) и

$$_{1-\alpha/2}F_{(n1-1),(n2-1)} = 1 / (_{\alpha/2}F_{(n1-1),(n2-1)}).$$

Нулевая гипотеза принимается, если F лежит между этими двумя табличными (критическими) значениями и отвергается в пользу ненаправленной H_1, если F выходит за пределы этого интервала, неважно в какую сторону- вправо, или влево.

Направленные альтернативные гипотезы имеют вид:

$$H_1 : \sigma_1 > \sigma_2 \text{ и } H_1 : \sigma_1 < \sigma_2 .$$

Логика их проверки – та же, что и логика проверки альтернативных гипотез, описанных выше.

10.3. Проверка гипотезы о равенстве нулю коэффициента корреляции

В социологических исследованиях довольно часто возникает ситуация, когда исследователь не знает, как следует интерпретировать коэффициент корреляции, рассчитанный для измерения связи между двумя интервальными. Скажем, равен этот коэффициент 0,3. Какой вывод должен сделать исследователь – имеется связь между переменными или нет? И снова, как и выше, на помощью приходит проверка статистической гипотезы. [58]

Рассмотрим способ проверки нуль-гипотезы - H_0: $\rho = 0$. Для такой проверки используется критерий

$$t_{n-2} = r\sqrt{\frac{n-2}{1-r^2}}$$

(напомним, что греческие буквы используются для обозначения генеральных параметров, а латинские – для отвечающих им выборочных статистик; в соответствии с этим r означает выборочный коэффициент корреляции). Эта статистика при справедливости нуль-гипотезы имеет распределение Стьюдента с df = (n – 2) степенями свободы.

Альтернативная гипотеза H_1 в данном случае может быть ненаправленной и направленной (это естественно, если вспомнить, что коэффициент корреляции может быть и положительным, и отрицательным), а рассмотренный критерий, соответственно, двусторонним и односторонним.

Направленный вариант альтернативной гипотезы - H_1: $\rho > 0$ (или – H_1: $\rho < 0$), ненаправленный - H_1: $\rho \neq 0$.

Подчеркнем, что при ненаправленной альтернативной гипотезе (и, соответственно, при использовании двустороннего критерия) гипотеза принимается, если - $t_{табл} \leq t_{набл.} \leq t_{табл.}$, и отвергается, если $t_{набл}$ выходит за пределы указанного интервала.

10.4. Проверка гипотезы о равенстве долей

Нередко в социологии возникает ситуация, когда требуется проверить гипотезу о равенстве генеральных долей:

$$H_{0:}\ p_1^{ген} = p_2^{ген} .$$

---

[58] Подчеркнем, что гипотезу о равенстве коэффициента корреляции нулю социологу надо проверять не только при непосредственном изучении причинно-следственных отношений. Скажем, при построении шкалы Лайкерта требуется проверка наличия корреляционной связи между каждым суждением и суммой всех остальных суждений (правда, и здесь расчет указанного коэффициента требуется для косвенной проверки наличия причинно-следственного отношения между наблюдаемыми переменными и латентной, измеряемой). И решить, оставлять ли проверяемое суждение в числе тех, которые будут участвовать в процессе измерения установки, можно только на базе проверки обсуждаемой статистической гипотезы.

Имеются в виду доли тех единиц рассматриваемых совокупностей (выборочной и генеральной), которые обладают каким-то заданным свойством. К примеру, исследователь выяснил, что среди опрошенных им студентов социологических вузов оказалось 13% юношей, а среди студентов экономических вузов- 21%. Встает вопрос, является ли это расхождение случайным (может быть объяснено тем, что выборки не достаточно адекватно отразили свойства генеральных совокупностей – всех студентов социологических и экономических вузов Москвы соответственно) или же оно принципиально – в экономические вузы действительно поступает относительно больше юношей, чем в социологические. И снова для ответа требуется проверка статистической гипотезы.[59]

$$z = \frac{p_1^{выб} - p_2^{выб}}{\sqrt{\overline{p}\overline{q}(\frac{1}{n_1} + \frac{1}{n_2})}}, \text{ где}$$

$\overline{p} = \frac{X_1 + X_2}{n_1 + n_2}$, $\overline{q} = 1 - \overline{p}$, X_1 – количество единиц первой выборки, обладающих рассматриваемым качеством, X_2 – то же для второй выборки.

Примеры задач.

1. Работая в одном из вузов, преподаватель убедился, что все его студенты являются достаточно однородными по уровню подготовки и разбиения отдельных учебных групп на подгруппы не требуется. Когда преподаватель решил начать работать в другом вузе, встал вопрос о том, следует ли группы этого вуза разбивать на однородные подгруппы. Чтобы быстро решить этот вопрос, преподаватель сформировал две выборки из студентов первого и второго вузов соответственно, составил специальный тест, с помощью которого опросил студентов обеих выборок. Было решено разбивать группы второго вуза, если разброс полученных по тесту оценок студентов второй выборки будет больше разброса аналогичных оценок студентов первой выборки и это различие будет статистически значимо. Какое решение примет преподаватель, если в выборке из первого вуза было 30 студентов, из второго – 25, а выборочные дисперсии равны, соответственно, 2,5 и 3,0 ?

2. Имеем следующую статистику по выборке из 5-ти регионов:

№ региона	1	2	3	4	5
Доля тяжких преступлений (на 1000 жителей)	0,05	0,01	0,2	0,08	0,9
Доля безработных (к численности работоспособного населения)	0,02	0,0005	0,05	0,03	0,1

Можно ли считать, что преступность обусловлена уровнем безработицы?

[59] Проверка гипотезы о равенстве долей часто требуется при проведении некоторых методических экспериментов, проведение которых требуется для грамотного построения анкеты. Так, желая проверить, влияет ли формулировка преамбулы перед анкетным вопросом на ответ респондента, можно дать респондентам сначала анкету с преамбулой одного вида, а потом, через некоторое время – другого и подсчитать соответствующие доли давших, скажем, перавый вариант ответа. Получим, к примеру, что при одной преамбуле указанным образом ответили 38% респондентов, а при другой – 43%. Решить, значимо ли такое различие в процентах, обусловливает ли вторая формулировка большее желание респондента дать именно первый вариант ответа, можно только на основе проверки статистической гипотезы. Жаль, что у нас подобные методические исследования редко проводятся.

3. При изучении электорального поведения населения часто возникает проблема определения того, за кандидата какой партии голосовали бы те избиратели, которые не пришли на голосование. Для доказательства того, что не пришли в основном сторонники партии А, было предложено подсчитать коэффициент корреляции между следующими наборами процентов:

№ избирательного участка	1	2	3	4	5	6	7
% явившихся на голосование	70	60	50	65	30	40	70
% проголосовавших за кандидата партии А	25	20	15	21	10	15	20

Логика рассуждений была такова: если существует связь между двумя выписанными рядами процентов (т.е. с ростом процента явившихся растет процент проголосовавших за кандидата А), то имеет смысл полагать, что отсутствующие на голосовании действительно являются сторонниками партии А. Можно ли считать, что в рассматриваемом случае связь есть?

4. При составлении анкеты проверялась гипотеза о том, что ответ респондента на некий закрытый вопрос анкеты зависит от порядка расположения вариантов предлагаемых ответов. В частности, предполагалось, что в случае, когда определенный вариант будет назван первым, то респондент отметит его с большей вероятностью, чем в случае, когда он назван последним. Для проверки гипотезы было опрошено две схожих группы респондентов. В анкете, использованной при опросе первой группы, состоявшей из 20 человек, рассматриваемый вариант ответа шел первым и его отметили 12 человек. В анкете, использованной при опросе второй группы, состоявшей из 25 человек, тот же вариант шел последним, и его отметили 10 человек. Можно ли считать, что гипотеза подтвердилась?

5. Предполагалось, что о стабильности экономической обстановки в стране (отсутствии войн, стихийных бедствий и т.д.) за последние 50 лет можно судить по характеру распределения населения по возрасту: при спокойной обстановке оно должно быть равномерным. В результате проведенного исследования, для одной из стран были получены следующие данные.

Возрастной интервал	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Доля населения	0,14	0,09	0,10	0,08	0,16	0,13	0,12	0,18

Имеются ли основания полагать, что в стране была нестабильная обстановка?

ТЕМА 11.
Методологические аспекты проверки математико-статистических гипотез

11.1. Ошибки первого и второго рода.

Надеемся, читателю ясно, что принятие решения на основе проверки той или иной математико-статистической гипотезы H_0 носит вероятностный характер. Мы можем совершить две ошибки: отвергнуть верную гипотезу и принять неверную. Какова вероятность каждой из этих ошибок? Как уменьшить эти вероятности? Попытаемся ответить на эти вопросы.

Прежде всего напомним, что если значение используемого критерия попало в «маловероятную» область (т.е. «зашкалило» за соответствующее критическое, табличное значение, вероятность чего мала при справедливости H_0, меньше выбранного нами уровня значимости α), гипотеза отвергается. Но ведь наш уровень значимости не равен нулю. Пусть очень редко, но значение критерия для конкретной выборки все же может «зашкалить» за найденное нами критическое значение и при справедливости H_0. Тогда отказ от H_0 будет ошибкой: гипотеза верна, а мы ее отвергли.

Эта ошибка называется *ошибкой первого рода*. Представляется очевидным, что вероятность совершения такой ошибки равна α. Ясно, что мы можем этой вероятностью управлять. Если для нас из содержательных соображений является важным недопущение отказа от справедливой гипотезы, будем уменьшать значение α и, соответственно, увеличивать критическое значение критерия.

А в каких случаях мы можем принять в действительности неверную гипотезу, т.е. совершить *ошибку второго рода?* Чтобы разобраться с этим, введем понятие *мощности критерия* и поясним, что это такое, на примере.

Рассмотрим один из возможных способов проверки затронутой нами в п. 10.3 гипотезы H_0: $\rho = 0$. В качестве альтернативной рассмотрим гипотезу H_1: $\rho \neq 0$. Ясно, что эта гипотеза – не направленная и поэтому критерий будет двусторонним. Поэтому при $\alpha = 0{,}05$ табличное значение будем считать для $\frac{\alpha}{2} = 0{,}025$. Воспользуемся рассуждениями из книги: Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. М.: Прогресс, 1976.

Обратимся к рисунку 11.1. Известно, что если в генеральной совокупности (для приводимых ниже рассуждений требуется, чтобы нашим признакам отвечало двумерное нормальное распределение) коэффициент корреляции равен нулю (т.е. если наша нуль-гипотеза верна), то всевозможные выборочные значения этого коэффициента для выборки объема n будут иметь приблизительно нормальное распределение с нулевым средним и стандартным отклонением, равным $1/\sqrt{n-1}$[60]. Так, если n = 200, то стандартное отклонение σ_r распределения выборочных коэффициентов корреляции будет равно примерно 0,071. Именно такое распределение изображено на рис. 11.1. слева. Нетрудно проверить также, что площадь под «хвостом»

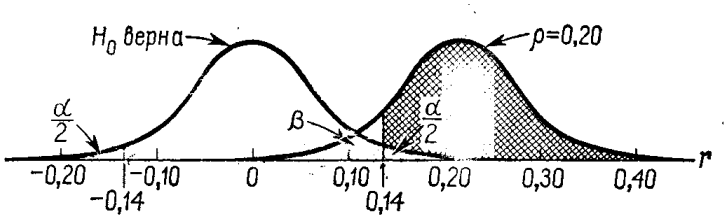


Рис. 11.1. Пример мощности критерия для H_0: $\rho = 0$ против H_1: $\rho \neq 0$ в случае, когда $\rho = 0{,}20$ (n =200, $\alpha = 0{,}05$)[61]; $r = r^{выб}$.

этого распределения, равная 0,025, будет находиться правее величины r = 0,14 (соответствующее табличное значение, равное, как известно 1, 96, будучи умноженным на среднее квадратическое отклонение, равное 0,071, даст как раз 0, 14). Значит, в соответствии с неоднократно описанной ранее логикой проверки статистических гипотез, мы будет отклонять нашу H_0, если $\left|r^{выб}\right| \geq 0{,}14$. Если наша гипотеза верна, и, следовательно, распределение выборочных коэффициентов корреляции совпадает с левым распределением рисунка 11.1, сделать такую ошибку – ошибку первого рода – мы рискуем с вероятностью 0,05.

А теперь представим себе, что в действительности в генеральной совокупности коэффициент корреляции равен не нулю, а, скажем, 0,2. Тогда распределение выборочных коэффициентов корреляции будет совпадать с правой кривой, представленной на рассматриваемом рисунке (соответствующее σ_r будет несколько меньше, чем у левой кривой, но на этом мы останавливаться не будем). Мы об этом не знаем и продолжаем использовать обычную логику принятия или отвержения

---

[60] Это утверждение носит лишь приблизительный характер. В других работах даются другие оценки характеристик распределения значений $r^{выб}$. Так, в учебнике [Прикладная статистика, т.1, с. 408] говорится о том, что в случае совместной нормальной распределенности исследуемых переменных и при достаточно большом объеме выборки n (а именно при n>200) распределение r можно считать приближенно нормальным со средним, равным своему генеральному значению, и дисперсией $\sigma_r^{\,2} = \frac{(1-r^2)^2}{n}$. При малых n и $|r|$, близкому к 1, это приближение становится очень грубым. В любом случае проверку гипотезы H_0: $\rho = 0$ лучше осуществлять с помощью критерия t_{n-2} из п. 10.3.

[61] См.: Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. М.: Прогресс, 1976. С. 259, рис. 13.6.

гипотезы. Естественно, если в такой ситуации мы примем нашу гипотезу, то тем самым совершим ошибку второго рода. В соответствии с нашими правилами, мы именно таким образом ошибемся, если наш единственный выборочный коэффициент корреляции попадет в интервал от –0,14 до + 0,14. Чтобы оценить вероятность этой ошибки, ошибки второго рода, вспомним, что в действительности распределение выборочных средних совпадает с правой кривой рисунка 11.1. И вероятность попадания нашего единственного выборочного значения коэффициента корреляции будет равна соответствующей площади под этой самой правой кривой. На рисунке она помечена буквой β. Это и есть вероятность ошибки второго рода. Мы видим, что она не очень велика. Это означает, что довольно большой является величина (1-β). Она называется мощностью критерия. На рис. 11.1 мощность критерия равна заштрихованной площади и составляет примерно 0,82.

Таким образом, для того, чтобы вероятность ошибки второго рода была мала, необходимо, чтобы мощность критерия была велика. Ясно, что мощность критерия равна вероятности отвергнуть неверную нуль-гипотезу.

Чтобы более ярко представить себе, что такое мощность критерия, предположим, что в реальности генеральное значение нашего коэффициента корреляции равно 0,1. Тогда ситуация, подобная только что рассмотренной, будет иметь вид, представленный на рис. 11.2.

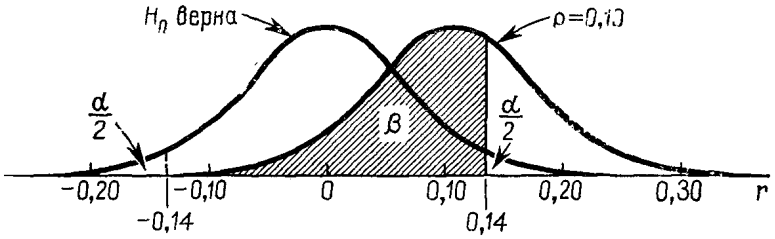


Рис. 11.2. Пример мощности критерия для H_0: $\rho = 0$ против H_1: $\rho \neq 0$ в случае, когда $\rho = 0{,}10$ ($n = 200$, $\alpha = 0{,}05$)[62].

Как и выше, мы принимаем нашу нуль гипотезу (и тем самым совершаем ошибку второго рода, если наш единственный выборочный коэффициент попадет в интервал от –0,14 до +0,14. Фактическая вероятность этого равна заштрихованной площади под правой кривой на рис. 11.2 (ведь в действительности именно эта кривая отвечает распределению выборочных значений коэффициентов корреляции). β, равная вероятности совершения ошибки второго рода, велика. Мощность критерия, равная (1-β), мала.

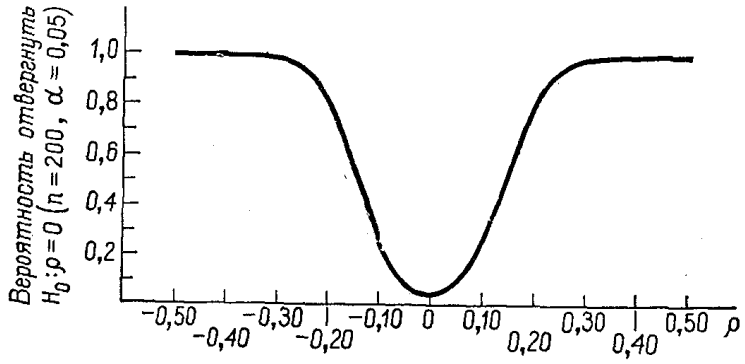


Рис. 11.3. Кривая мощности для критерия для H_0: $\rho = 0$ против H_1: $\rho \neq 0$ при $n = 200$ и $\alpha = 0{,}05$ [63].

Рассмотрим график на рис. 11.3. По горизонтали откладываются значений генеральных коэффициентов корреляции. По вертикали – вероятность отвергнуть неверную нуль-гипотезу, т.е. мощность критерия. Надеемся, что читателю понятен смысл этой кривой.

---

[62] Там же. С. 260, рис. 13.7.

[63] Там же. С. 261, рис. 13.8.

Если в генеральной совокупности коэффициент корреляции близок к нулю, то у нас будет очень мала вероятность отвергнуть нашу нулевую гипотезу (эта вероятность будет равна выбранному уровню значимости). Это означает, что мощность критерия мала: велика величина β, т.е. вероятность принятия H_0. Если же генеральный коэффициент корреляции достигает 0,3 и выше, то у нас вероятность отвержения нуль-гипотезы становится близкой к 1.

Все сказанное в этом параграфе можно свести в следующую таблицу.

Исследовательское решение	Состояние природы	
	H_0 верна	H_1 верна
Отвергнуть H_0 (принять H_1)	Ошибка I рода (Вероятность = α)	Правильное решение (Вероятность = 1- β)
Отвергнуть H_1 (принять H_0)	Правильное решение (Вероятность = 1- α)	Ошибка II рода (Вероятность = β)

α - уровень значимости, (1 - β) - мощность критерия.

11.2. Пример влияния содержательного характера задачи на выбор уровня значимости.

Выше мы говорили о том, что на выбор уровня значимости при проверке гипотез может влиять содержательный характер решаемой задачи. Приведем пример того, когда это имеет место применительно к проверке об отсутствии связи на основе критерия Хи-квадрат.

Приведем пример гипотетической задачи. Предположим, что руководитель некоторого региона должен принять решение по вопросу о том, стоит или не стоит осушать имеющиеся в этом регионе болота. Если между заболоченностью почвы в регионе и уровнем заболеваемости его жителей есть связь, то необходимо затрачивать огромные средства на осушение болот. Если связи нет – средства можно не тратить. А денег мало… В таком случае для руководителя естественным будет стремление уменьшить ту область, где он может совершить ошибку первого рода – отвергнуть в действительности верную гипотезу, т.е. уменьшить α - вероятность отвержения справедливой гипотезы.

Будем полагать, например, что свершение события, имевшего вероятность 0,05, еще не говорит о несправедливости наших посылок – вероятность, конечно, мала, но все же довольно существенна и можно полагать, что все же в генеральной совокупности связи нет (гипотеза верна). Может быть, и относительно вероятности 0,01 мы также будем рассуждать. А вот если мы встретим событие, вероятность реализации которого в предположении справедливости нуль-гипотезы была равна 0,001, тут уж мы засомневаемся. Наверное, такое событие не может реализоваться. Значит, действительная вероятность свершившегося события (отклонения выборочной частотной таблицы от ситуации пропорциональности) была не столь ничтожной. Другими словами, в генеральной совокупности столбцы (строки) таблицы, всего вероятнее, - не пропорциональны, т.е. между переменными имеется связь. Придется раскошеливаться на мелиорацию земель.

То, что обычно принимается соотношение $\alpha = 0{,}05$, опирается на большой практический опыт. В подавляющем числе реальных задач предположение о том, что события с такой вероятностью практически не случаются, вполне оправдано. Но еще раз подчеркнем, что и при проверке математико-статистических гипотез мы должны учитывать житейскую логику. Если речь будет идти, скажем, о том, что вероятность заболевания у нас горла после того, как мы съедим зимой на улице мороженое, равна 0,05, мы, наверное, предположим, что событие с такой вероятностью случиться не может, никаких неприятностей не произойдет, и смело будем потреблять мороженое. Однако в другой ситуации, скажем, если бы мы находились на месте архитектора, думающего о том, стоит ли при строительстве дома использовать сейсмостойкие конструкции, мы бы, вероятно, рассуждали по-другому. Если бы было известно, что вероятность землетрясения в рассматриваемом районе – 0,05, то мы бы, вероятно, поосторожничали, сочли бы землетрясение все же возможным и не стали бы рисковать человеческими жизнями, заложили бы в проект дорогостоящие сейсмостойкие элементы. Вот если бы наш архитектор работал в другом районе, где вероятность землетрясения - 0, 0001, то он, наверное, позволил бы себе считать, что такое событие практически произойти не может, и не стал бы делать дом сейсмостойким, сэкономив тем самым средства.

Если же тяжелые для нас затраты (любого плана – материальные, моральные и т.д.) связаны с ситуацией отсутствия связи, то мы будем «бояться» совершить ошибку второго рода – придти к

выводу об отсутствии связи в то время как она в действительности имеет место. В таком случае мы должны стремится увеличить мощность используемого критерия.

Можно также уменьшить ту область, где гипотеза принимается, т.е. увеличить уровень значимости (тем самым мы увеличим мощность критерия). Будем полагать, что не может произойти событие, имеющее вероятность не только 0,05, но и 0,1 ($\alpha = 0,1$) и будем отвергать гипотезу о независимости, если наш критерий достигнет соответствующей критической величины. А вот если встретим событие, вероятность реализации которого в предположении справедливости нуль-гипотезы была равна 0,2 - то ничего необычного в этом не найдем, т.е. будем считать, что у нас нет оснований отвергать нуль- гипотезу. И уж здесь придется идти на жертвы. Логика, прямо скажем, не очень надежная, поскольку ее реализация резко повышает вероятность ошибки первого рода – отказа от верной гипотезы.

11.3. Различие между статистической и содержательной гипотезой

Распространенным ответом студента на вопрос о том, какую математико-статистическую гипотезу надо проверить при решении той или иной конкретной задачи является фраза типа: «Надо проверить гипотезу о том, есть связь или нет». Такого рода предложение некорректно по крайней мере по двум причинам. Во-первых, сформулированная фраза вообще не содержит формулировку какой бы то ни было гипотезы. Гипотезой может быть или утверждение о том, что связь есть, или – о том, что связи нет. Во-вторых, даже гипотеза типа «здесь связь есть» может восприниматься как некоторое нечетко выраженное содержательное предположение (и как таковое имеющее право на существование), но никак не математико-статистическая гипотеза. Последняя должна быть сформулирована очень строго, например, так: «коэффициент корреляции в изучаемой генеральной совокупности равен 0,7». Именно такой строгий характер носили все рассмотренные нами выше гипотезы, обозначаемые знаком H_0. В п. 7.3 мы позволяли себе выражение «H_0: связи нет» только потому, что показали, что словосочетание «связи нет» может быть выражено формально, строго, на языке определенных формул (стоящих, например, за соотношениями (7.1)). . И мы не делали этого только из соображений краткости записи и стремления сделать текст более понятным для читателя-гуманитария.

Формулируя строгую математико-статистическую гипотезу, мы должны также понимать, что она должна опираться на определенные математические результаты, лежащие в основе разработки используемого для проверки гипотезы критерия. Другими словами, мы должны знать, что критерий для проверки нашей гипотезы существует. И если мы с помощью математической статистики хотим проверить какую-то интересующую нас из содержательных соображений гипотезу, то сначала мы должны погрузиться в математико-статистическую литературу, в справочники и убедиться, что критерий для проверки нашей гипотезы действительно существует. Особенно осторожно надо относиться к т.н. методам анализа данных. В методах анализа данных часто используются такие математические конструкты, для которых не разработаны способы переноса результатов с выборки на генеральную совокупность (подобные методам проверки статистических гипотез и построения доверительных интервалов). Таковы, например, многие методы кластерного анализа или многомерного шкалирования. Более того, как мы уже отмечали выше (п.1.3), при анализе данных зачастую возникают такие ситуации, когда мы не можем мыслить наблюдаемые объекты как выборку из некоторой генеральной совокупности. Это может иметь место, например, в силу того, что исследователю совершенно неясно, какой в его случае является генеральная совокупность. Возможны и другие причины, например, отсутствие оснований считать значения какого-то признака реализациями одной и той же случайной величины (скажем, у нас могут быть основания считать различными распределения зарплат мужчин и женщин; а, может быть, у русских и украинцев? А, может, у «либералов» и «коммунистов»?). Заранее сказать, так это или не так, невозможно. Какие категории объектов надо «испытывать» с целью выявления специфических распределений, неясно[64]. Об использовании математической статистики, основным объектом которой являются случайные величины, говорить в такой ситуации нельзя. Стало быть, нельзя говорить и о проверке математико-статистических гипотез. О содержательных же гипотезах в принципе может идти речь.

Различие между содержательными и математико-статистическими гипотезами имеет еще один очень важный для социолога аспект. Одна и та же содержательная гипотеза может быть проверена с помощью разных математико-статистических подходов. Приведем пример.

---

[64] По существу в такой ситуации речь идет о выделении в изучаемой совокупности объектов однородных подсовокупностей (о соответствующем методологическом принципе мы говорили в п.1.7). Однородность нередко отождествляется как раз с тем, что для рассматриваемых признаков осмыслены соответствующие распределения вероятностей (это мы уже отмечали в п. 1.3) .

Допустим, мы хотим выяснить, имеется ли статистическая связь между двумя признаками x и y, измеренными по интервальной шкале. Содержательная гипотеза может звучать, например, так: «связь между данными признаками существует»[65]. С помощью математической статистики проверку такой гипотезы можно осуществить несколькими способами, приводящими, вообще говоря, к разным результатам.

Во-первых, можно вычислить коэффициент корреляции и проверить гипотезу о его равенстве нулю.

Во-вторых, можно разбить диапазоны изменения значений обоих признаков на интервалы и применить критерий «Хи-квадрат». Здесь очень важно обратить внимание на то, что результат, вообще говоря, будет зависеть от способа упомянутого разбиения. При одном разбиении гипотеза об отсутствии связи может быть принята, при другом – нет. Здесь мы снова возвращаемся к одной из основных проблем математической социологии: построению признаков, отражающих то или иное социальное явление (ср. конец п. 7.3).

В-третьих, можно вычислить корреляционное отношение – коэффициент, отражающий криволинейную зависимость y от x, либо x от y и проверять значимость его отличия от нуля (см. тему 13).

В-четвертых, можно искать регрессионную зависимость среднего значения y от x и обосновывать статистическую значимость всех используемых при этом коэффициентов (в настоящем курсе изучение соответствующей темы не предусмотрено[66], однако регрессионный анализ описывается в очень большом количестве работ по математической статистике и анализу данных; и почти везде указываются способы статистической оценки всех получаемых при этом параметров).

В-пятых, можно разбить диапазон изменения x на группы (ячейки) и применить дисперсионный анализ для сравнения значений y для объектов, попавших в разные группы (см. тему 14).

Мы перечислили отнюдь не все известные способы изучения статистических связей между двумя переменными. Коснулись только самых популярных. И уже стало ясно, что множественность методов, решающих одну и ту же задачу – это проблема для социолога. Как же ее решать? Мы уже говорили в конце п.7.3, что наше знание будет объективным только в том случае, если мы наряду со сведениями, полученными в результате того или иного анализа статистических данных, в понятие «знание» будем включать также и тот способ, с помощью которого эти сведения получены. В данном случае речь идет по существу об учете той модели изучаемого явления (т.е. о том понимании статистической связи, или, может быть, причины), которая заложена в используемом нами методе (о важности учета такой модели как об одном из главных принципов использования математического аппарата в социологии шла речь в п. 1.7). Так, не имеет смысла говорить о том, что наша содержательная гипотеза о наличии связи между x и y подтвердилась. Полученный результат надо формулировать по-другому, например, так: на таком-то уровне значимости мы отвергли гипотезу о равенстве нулю коэффициента корреляции между рассматриваемыми признаками; на таком-то уровне значимости и при таком-то разбиении значений признаков на интервалы мы отвергли гипотезу об отсутствии связи с помощью критерия «Хи-квадрат», но приняли аналогичную гипотезу при некотором другом разбиении и т.д. Содержательно проинтерпретировать подобные факты можно только при том условии, что социолог достаточно хорошо представляет себе модель, заложенную в каждом методе. Надеемся, что сказанное говорит и о роли учета модели, заложенной в используемом математическом методе, и о различии между содержательной и математико-статистической гипотезами.

Мы показали, что одна и та же содержательная гипотеза может проверяться с помощью использования разных математико-статистических приемов (в том числе с помощью проверки

---

[65] Вообще говоря, содержательная постановка задачи обычно бывает связана с изучением не статистической связи, а причинно-следственных отношений между рассматриваемыми признаками; однако мы уже говорили о том, что никакие формальные методы не могут нам доказать, что один признак можно считать причиной, а другой следствием; подобные формулировки – дело социолога; в п. 12.1. мы более подробно коснемся проблемы соотнесения причинно-следственных отношений и статистической зависимости.

[66] Учебный план ГУ-ВШЭ, в соответствии с которым был написан настоящий учебник, предусматривает изучение регрессионного анализа в курсе анализа данных, слушать который студенты начинают несколько позже начала курса математической статистики (но частично параллельно последнему). Описание регрессионного анализа, рассчитанное на читателя-социолога, можно найти в работе автора [Толстова, 2000], более строгое изложение – в большинстве других работ, названных в Приложении 1 (один из наиболее фундаментальных учебников - [Прикладная статистика…, 2001].

разных математико-статистических гипотез, что в основном и интересует нас в настоящем параграфе). Можно показать и обратное утверждение: даже при рассмотрении одних и тех данных проверка одной и той же математико-статистической гипотезы может быть использована для решения разных содержательных задач, для проверки разных содержательных гипотез. Приведем пример.

Вспомним гипотезу H_0: $\mu_1 = \mu_2$. И вспомним уже упоминавшуюся задачу сравнения средних зарплат для мужчин и для женщин. Вопрос, на который мы намереваемся получить ответ с помощью проверки рассматриваемой гипотезы, может формулироваться по-разному, например: можно ли считать, что различие между средними зарплатами мужчин и женщин статистически значима? Можно ли полагать, что зарплата детерминируется полом? Первый вопрос может быть одним из многих других вопросов: о различии заплат у лиц разных национальностей, разных профессий и т.д. А второй может отражать работу специалиста по гендерной социологии о дискриминации женщин. Хотя задачи и схожи, но все же содержательные гипотезы здесь будут разными.

Раздел IY. ПРОБЛЕМА ИЗУЧЕНИЯ ПРИЧИННО-СЛЕДСТВЕННЫХ ОТНОШЕНИЙ И ЭКСПЕРИМЕНТ В СОЦИОЛОГИИ; ОСНОВНЫЕ ИДЕИ ДИСПЕРСИОННОГО АНАЛИЗА

Основной целью настоящего раздела является изложение идей одного из известнейших направлений математической статистики – дисперсионного анализа. Однако нам хотелось бы, чтобы читатель-социолог воспринял эти идеи не механически, а в свете понимания того, насколько он сможет использовать их для решения одной из самых главных задач любого научного исследования – изучения причинно-следственных отношений. Ведь дисперсионный анализ, по большому счету, предназначен именно для этого. И для того, чтобы его грамотно использовать, необходимо хотя бы кратко оговорить, в чем специфика того «понимания» этих отношений (в частности, понимания термина «причина»), которая заложена в соответствующем формализме. Этому, в частности, может способствовать также сравнение дисперсионного анализа с другими математико-статистическими методами изучения связи между переменными, с другими (не обязательно математико-статистическими) способами анализа каузальных отношений.

Поэтому прежде, чем перейти непосредственно к описанию основ дисперсионного анализа, мы очень кратко рассмотрим некоторые методологические аспекты изучения причинных отношений вообще, с помощью математико-статистических методов – в частности; коснемся сути понятия эксперимента - того подхода к изучению причин, формализацией которого по существу является дисперсионный анализ, коротко коснемся специфики эксперимента именно в социологии. Кроме того, мы проанализируем, как прообразы идей, лежащих в основе дисперсионного анализа, «работают» при использовании простейшего коэффициента парной связи – корреляционного отношения, которое тоже можно рассматривать как способ изучения причинных отношений.

Для более яркого преподнесения математико-статистического подхода мы коснемся также нестатистических методов изучения каузальных структур.

Т е м а 12

Методологические аспекты изучения причинно-следственных отношений с помощью математических методов. Эксперимент в социологии

12.1. Проблема изучения причинно-следственных отношений.

Подчеркнем специфику данного параграфа. Мы очень коротко охарактеризуем ряд известных, ставших классическими, подходов к пониманию причины. При этом сознательно обратимся к истокам интересующих нас положений. Наше изложение вынужденно является поверхностным: тема в какой-то мере уводит нас в сторону от основного содержания настоящего учебника. За кадром останутся многие интереснейшие методологические положения и споры, не потерявшие своего смысла и в наше время. Причина включения этого параграфа в книгу по сути оговорена выше. Мы хотим, чтобы математический аппарат воспринимался читателем не как нечто навязанное социологу «сверху», «со стороны», а как органическая часть социологического исследования. Хотелось бы, чтобы читатель понял, что математический язык – органическая часть социологического, что использование математического языка весьма удобно для адекватного изучения причинно-следственных отношений.

«Причина и следствие – философские категории, выражающие одну из форм всеобщей связи явлений. Причина обычно мыслится как явление, действие которого производит, определяет или вызывает другое явление; последнее называют следствием».[67] Для наших целей представляется существенным следующее добавление к сформулированному предложению, присутствующее в другом энциклопедическом издании: «Производимое причиной следствие зависит от условий. Одна и та же причина при разных условиях вызывает неодинаковые следствия. Различие между причиной и условием относительно. Каждое условие в определенном отношении является причиной, а каждая причина в соответствующем отношении есть следствие. Причина и следствие находятся в единстве: одинаковые причины в одних и тех же условиях вызывают одинаковые следствия. ... Совокупность всевозможных вещей и процессов природы составляют общее (универсальное) взаимодействие, исходя из которого мы приходим к действительному каузальному отношению ... Причина и следствие могут меняться местами: следствие может стать причиной другого следствия. Во многих областях объективной действительности само взаимодействие причины и следствия выступает как причина изменения явлений и процессов».[68] Для социолога, изучающего причинно-следственные отношения на основе анализа статистических связей, это весьма важно. Существуют математико-статистические методы измерения связей, которые позволяют в определенной степени учитывать указанные взаимодействия причин, следствий и условий протекания изучаемых процессов. Мы коснемся этого лишь в небольшой степени.

Несмотря на то, что изучение причинно-следственных отношений – главная задача любого научного исследования, единого определения понятия «причина» не существует. Приведем житейский пример, в какой-то мере говорящий о том, почему здесь возникают проблемы.

Предположим, что пожилой человек споткнулся о камень, лежащий на дорожке, и упал. Какова причина его падения? Естественно полагать таковой то, что человек споткнулся о камень. Падение – следствие этого. Одно явление (спотыкание о камень) повлекло за собой другое (падение). Вроде бы все ясно. Но представим себе, что по той же дорожке прошел молодой человек, тоже споткнулся, но не упал. Значит, камень – причина падения только при условии, что человек стар? Однако представим себе, что рядом с дорожкой был забор. И старый человек в той же ситуации не упал, поскольку схватился за забор. Значит, в число тех же условий попал и забор? Более того, то, что человек споткнулся, может быть не причиной падения, а следствием падения накануне, падения, приведшего к боли в ноге. И так – до бесконечности ... Естественно, глубокий анализ подобных ситуаций может привести к разным представлениям о том, что такое причина; в частности, о том, существуют ли вообще причины чего бы то ни было.

Анализу понятия причины уделялось внимание в трудах многих выдающихся ученых. И спектр мнений здесь очень широк. Мы не имеем возможности описать развитие этого понятия более или менее подробно (да это и не является нашей целью). Укажем лишь несколько ключевых фамилий с надеждой на то, что кто-то из студентов заинтересуется соответствующей проблемой и попытается найти ответ на многие стоящие перед исследователями и до сих пор не имеющие ответа вопросы, связанные с эффективным изучением каузальных структур, действующих в жизни общества.

О понятии причины говорил еще Аристотель (384-322 до н.э.). Этого мы касаться не будем, хотя использование взглядов выдающегося ученого Древней Греции отнюдь не бесполезно при анализе причинно-следственных отношений на основе современных математических методов. Не будем рассматривать и многие заслуживающие внимания положения других ученых, живших в следующие после Аристотеля два тысячелетия. «Перепрыгнем» в эпоху Просвещения.

Огромную роль в создании теории причинности, основанной на идее жестко детерминированного мира (все, что происходит, объясняется причинно-следственными связями), сыграл тот переворот в естествознании, который был осуществлен И.Ньютоном (1643-1727), создавшим классическую механику, что, помимо всего прочего, знаменовало собой очень важный методологический переход в научном познании мира от правдоподобных абстрактных рассуждений к точной количественной теории, к эксперименту[69]. Со временем, когда стало ясно, что огромное количество причин, порождающих очень многие интересующие ученых явления, в принципе не может быть учтено, понятие каузальных отношений стало связываться с понятием вероятности. Один из создателей современной теории вероятностей П.Лаплас (1749 – 1827), стоящий на ньютоновской

---

[67] Причина и следствие // Новая философская энциклопедия. Т.3. М.: Мысль, 2001. С. 353

[68] Причина и следствие // Философский энциклопедический словарь. М.: Советская энциклопедия, 1983. С. 531.

[69] О роли ньютоновских идей на умы исследователей XVII – XIX столетий говорит, например, то, что известный философ А.Сен-Симон (1760-1825) (у которого О.Конт (1798-1857) заимствовал многие идеи) опубликовал работу «Всеобщее тяготение», в которой призывал к построению истории по образцу классической механики.

позиции, в своей работе «Опыт философии теории вероятностей» (1795) писал, объясняя роль теории вероятностей в познании мира (и в «натуральной философии», и в «нравственных науках»): «Всякое имеющее место явление связано с предшествующим на основании того очевидного принципа, что какое-либо явление не может возникнуть без производящей его причины. ... Воля, самая свободная, не может породить ... действия без побуждающей причины ... Противоположное мнение есть иллюзия ума, который, теряя из виду мелкие причины того или другого выбора воли, убеждается, что она определяется самою собою и беспричинна».[70] Подчеркнем, что здесь речь идет о причинной обусловленности т.н. свободной воли отдельного человека. Следующую цитату (о том, как теория всеобщей причинной обусловленности совмещается с обоснованием необходимости прибегать к помощи математической статистики), заимствованную из той же работы Лапласа, мы дадим в переводе русского ученого А.А.Чупрова, ибо этот перевод представляется нам более адекватным: «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 движение мельчайших атомов; ничто не было бы для него недостоверным; будущее, как и прошедшее, были бы открыты его взору».[71] Далее Чупров пишет: «Величественное научное credo и в то же время грандиозная программа! Кто же из работников науки откажется исповедовать эту веру и в то же время не остановится в раздумьи перед непомерным размахом программы!». Каков же выход из положения? Ответ мы уже обсуждали в п. 1.3. По сути здесь речь идет о генезисе понятия статистической закономерности. «Статистическая точка зрения знаменует отказ от того прослеживания единичных событий, которое рисуется уму естествоиспытателя как идеал полноты и совершенства знания. Область, где статистический способ изучения утверждался впервые, - явления в совокупностях человеческого общежития, - вообще не поддается изучению путем «астрономическим». Необозримы в своем прихотливом многообразии действия отдельных людей с их трудно уловимыми мотивами поведения, события в их индивидуальной жизни. Вместе с тем рассматриваемые обособленно, они часто мало интересны и не привлекают к себе поэтому внимание исследователя. Нам важно знать именно массовые итоги и общие результаты, представляющие собой сгустки единичных явлений и событий».[72] Соответственно, с помощью статистических методов, надо изучать и причины разных явлений. При таком подходе к их изучению они тоже будут носить усредненный, статистический характер, о чем пойдет речь ниже.

Здесь же несколько скорректируем мнение Чупрова в соответствии с некоторыми современными представлениями. Во-первых, изучение причин отдельных событий зачастую бывает невозможно не только из-за «неподъемности» соответствующей работы, но и из-за принципиальных соображений: например, в силу невозможности объективного измерения каких-то параметров из-за влияния «прибора» на объект измерения. Это явление хорошо известно физикам, и они научились с ним «бороться» (принцип неопределенности Гейзенберга). Социологи тоже знают об указанном обстоятельстве (опросный инструментарий и обстановка опроса влияют на характер получаемых от респондентов ответов), но пока не разработано эффективных способов «борьбы» с этим явлением [73]. И, во-вторых, нельзя полностью отрицать необходимость разбираться в событиях индивидуальной жизни отдельных людей. Необходимость качественных методов исследования (например, использования биографического интервью) вряд ли можно отрицать. Но, наверное, все же говорить о научных выводах все же можно будет только в том случае, если мы, казалось бы, сугубо индивидуальные явления тоже встроим в статистическую канву (будем искать случаи повторения «экзотических» событий и анализировать возможность рассматривать их как причины или следствия других явлений).

Укажем еще некоторые соображения Чупрова. Их он высказал в статье, написанной для известного энциклопедического словаря Брокгауза и Эфрона. Чупров выделяет несколько научных

---

[70] Лаплас. Опыт философии теории вероятностей // Вероятность и математическая статистика. Энциклопедия. М.: Большая российская энциклопедия, 1999. С. 835. Небезынтересно отметить, что эта работа Лапласа в значительно мере опиралась как на его собственные социально-демографические исследования, так и аналогичные работы других авторов.

[71] Чупров А.А. Вопросы статистики. М.: Госстатиздат ЦСУ СССР, 1960. С. 142.

[72] Там же, с. 143.

[73] О принципе неопределенности для социологии см., например: Алексеев И.С., Бородкин Ф.М. Принцип дополнительности в социологии // Моделирование социальных процессов. М.: Наука, 1970. С. 37-48.

школ, связанных с изучением нравственной статистики[74]. Ту, к которой он причисляет самого себя, русский ученый называет *математической* (поскольку она ищет опору в математике и, в первую очередь, в теории вероятностей), или *логической* (поскольку она теснейшим образом примыкает к логике). Эта школа пришла «к более точному решению тех задач, над которыми бесплодно бились представители» других направлений в моральной статистике. Чупров подчеркивает, что «главную, хотя и не исключительную роль … при этом играла теория вероятности» (с.406). При этом он отмечает незавершённость работы и в то же время он говорит об «оживляющем веянии новых идей». О каких «новых идеях» идет речь?

В основании приложения теории вероятностей к статистике лежит представление об объективной возможности, введение которого было связано с углублением представлений о причинности. Если дана причина *A* некоторого явления, в полной её сложности и определённости, то явление необходимо. Но если имеет место не вся причина, а лишь некоторая часть ее, то явление становится только возможным. «Аналитическое исследование отношений *возможной* причинной связи является логической функцией статистического метода, тогда как аналитическое исследование отношений связи *необходимой* приходится на долю индукции» (с.407). Для количественного выражения объективной возможности «служит особая численная характеристика, получающая название математической вероятности. Значение этой характеристики раскрывается знаменитой теоремой Бернулли: в длинном ряду явлений, стоящих отчасти под действием одних и тех же общих причин, отчасти под действием причин, свойственных каждому из них в отдельности и с общими в связи не стоящих, числа повторений всех возможных событий всегда приблизительно пропорциональны их вероятностям при взятых общих условиях» [75](с. 407). Развивая эти положения в других своих работах, Чупров пришёл к некоторым соображениям, лежащим в основе современного причинного (путевого) анализа, предвосхитил отдельные идеи крупнейшего американского социолога-эмпирика и методолога эмпирической социологии П. Ф. Лазарсфельда (1901–1976)[76]. Отметим, что об этом говорил сам Лазарсфельд.[77]

Примерно тех же взглядов на понятие причинности, на связь причинности со статистическим подходом, разделял М.Вебер (1864-1920). Взгляды великого немецкого социолога по этому поводу (в рамках обсуждения категории «возможного» в структуре каузального культур-социологического анализа) в их сравнении со взглядами русский неокантианцев (прежде всего – в лице Б.(Ф.)А.Кистяковского (1868-1920)), немецких и русских статистиков (И.Криз (1853-1928), В.И.Борткевич (1868-1931), А.А.Чупров) подробно описаны в работе известного современного

---

[74] *Чупров А. А.* Нравственная статистика // Брокгауз Ф.А. (Лейпциг), Ефрон И.А. (СПб.). Энциклопедический словарь. Т. XXI. С.-Петербург: Типолитография И.А.Ефрона, 1897. С. 403–408. Отметим, что эту работу высоко оценил М.Вебер (1864-1920). Так, в процессе переписки с известным русским социологом-неокантианцем Б.А.Кистяковским, он, обсуждая актуальные для тогдашней социологии методологические проблемы, связанные с трактовкой понятий причинности, необходимости, возможности и т.д., апеллирует к мнению А.А.Чупрова, ссылаясь именно на эту статью [Давыдов Ю.Н. Макс Вебер и современная теоретическая социология. М.: Мартис, 1998, с. 164-165] .

[75] Ср. с законом больших чисел, одна из формулировок которого дана в п. 4.2.

[76] Так, логика, используемая А. А. Чупровым при рассуждениях о множественности причин, обусловливающих одно и то же следствие, и множественности следствий, возникающих под действием одной и той же причины, очень напоминает логику путевого анализа, дающую возможность построения адекватной реальности первичной причинной схемы. Высказанное им мнение о существенности того обстоятельства, что некая переменная может находиться в стохастической связи с рядом переменных, взятых в отдельности, и в функциональной зависимости от этих переменных, взятых вместе, заставляет вспомнить предлагаемую Лазарсфельдом логику операционализации понятий.

[77] В работе (Култыгин, с. 456) приводится следующая информация: «В начале 70-х годов автор настоящей работы, встретившись в Питсбургском университете с П.Лазарсфельдом, рассказал ему, что в Советском Союзе Лазарсфельда считают основателем применения математических методов в социологии. Однако сам Лазарсфельд заявил, что подобная оценка сильно преувеличена и что главное, что он сделал в этой области, - это удачное применение тех подходов и методик, которые были созданы А.А.Чупровым еще в 20-х годах нашего века».

российского социолога Ю.Н.Давыдова. [78] Не будем раскрывать эти очень интересные дискуссии, хотя они отнюдь не потеряли своей актуальности и для нашего времени.

Описанный взгляд на проблему причинности разделялся отнюдь не всеми исследователями. Упомянем только двух ученых, глубоко разрабатывавших проблему причинности, но отрицавшими объективное существование причин. Это Юм и Кант. Подчеркнем, что в целом мы не согласны с идеалистической направленностью их представлений, но, тем не менее, полагаем, что использование ряда разработанных ими идей полезно при обосновании необходимости изучения причинно-следственных отношений с помощью математико-статистических методов и при разработке адекватных способов такого изучения.

Для известного английского (шотландского) историка и мыслителя Д.Юма (1711-1776) понятие «Я» – это «пучок перцепций» (чувственных восприятий). Перцепции разрозненны и единичны, чисто индивидуальны, не связаны друг с другом; следовательно, в мире нет реальной причинности, мы просто привыкли к тому, что за одной перцепцией (причиной) часто следует другая (следствие), но логически показать, почему это так, мы не можем. Подчеркнем, что здесь речь идет отнюдь не о статистическом понимании причинности (с которым соглашались Лаплас, Вебер, Чупров). Здесь скорее имеется в виду индуктивный процесс (доказательство с помощью перехода от частного к общему). В связи с этим вглядимся внимательно в приведенную выше цитату из Чупрова: аналитическое исследование отношений *возможной* причинной связи русский ученый понимает как логическую функцию статистического метода и говорит о вероятности как численной характеристике объективной возможности; аналитическое же исследование *необходимой* связи он относит к индуктивному выводу. Таким образом, в изучении причинной структуры он как бы выделяет два разных направления: статистическое и индуктивное. Поскольку глубоко познать роль какого-то подхода в науке можно только в сопоставлении его с другими подходами, то ниже мы уделим индуктивному подходу особое внимание (это тем более важно, что, насколько нам известно, в литературе, ориентированной на социолога, эти *практически* важные моменты не рассматриваются).

Теперь вернемся к Юму и отметим, что и он не совсем все в человеческом познании сводил к перцепциям. По Юму, здесь имеются исключения: например, *математическое* познание не зависит от перцепций и, например, геометрия является результатом неких операций воображения, не связанных с опытом[79]. Именно такой взгляд на математическое познание в сочетании с идеей о возникновении понятия причины в результате привычки послужили одним из мощных толчков к тому обновлению европейской философии, которое было осуществлено великим немецким философом И.Кантом (1724-1804). Этот ученый рассматривал причинность как одну из тех категорий, без которых недостижимо упорядочение опыта (чувственных данных, результатов воздействия на нас вещей в себе). Кант полагал, что эту категорию нельзя извлечь из опыта: она является одной из априорных форм познания и как таковая требует предварительного критического анализа, изучения ее взаимодействия с опытом [80].

Нам важно отметить, что кантовская теория познания учитывает активность субъекта (она проявляется в представления о причинности как априорной формы познания). Это обстоятельство будет использоваться нами в следующих параграфах, при обсуждении вопросов, связанных с

---

[78] Давыдов Ю.Н. Макс Вебер и современная теоретическая социология. М.: Мартис, 1998. С. 150-190).

[79] Мы не согласны с тем, что геометрия не связана с опытом. С опытом связана любая ветвь математики, а геометрия – в особенности (как известно, она родилась в Древнем мире на основе измерения площадей земельных участков). Однако мы полностью согласны с тем, что, несмотря на связь с практикой, любая ветвь математики в значительной мере является плодом человеческого воображения. Математические теории – это результат особого видения мира их авторами. Нам близка идея видного русского политического деятеля конца XIX – начала XX С.Ю.Витте (1849-1915) (подхваченная современным российским математиком академиком В.И.Арнольдом) о разделении всех профессионалов-математиков на «философов» и «исчислителей» (об этом мы подробнее говорили в работе: *Ю.Н.Толстова.* «Дух» математики как основа научного социологического исследования // Математическое моделирование социальных процессов. Вып.7. М.: Макс Пресс, 2005. С.6-27). В данном случае мы говорим о «математиках-философах». Отметим еще то, что и математические объекты можно связывать с чувственным восприятием мира. Так, Аристотель писал, что «математические предметы, вопреки утверждениям некоторых, нельзя отделять от чувственно воспринимаемых вещей» (Аристотель. Метафизика // Аристотель. Собрание сочинений. Т.1. М.: Мысль, 1976. С. 367).

[80] И.Кант. Сочинения. Т.III. М: Мысль, 1964. С.174-175 (цит. по : Антология мировой философии Т.3. М.: Мысль, 1971. С. 122).

изучением каузальных структур на основе использования математического аппарата: как мы увидим, выбор того или иного метода и, соответственно, интерпретация понятия причины, зависит от самого исследователя. Тем самым мы позволим себе несколько перетолковать кантовское понятие причинности, связав представление об априорности соответствующей формы познания с заданием этой формы посредством выбора используемой математической модели причинности (т.е. выбора то ли коэффициента корреляции, то ли регрессионного анализа и т.д.). В оправдание такого «святотатства» приведем еще одну цитату из Канта (напомним, что, в соответствии с мнением этого ученого мы изучаем только явления не познаваемых в принципе предметов, вещей в себе): «Так как явления не вещи в себе, то в основе их должен лежать трансцендентальный предмет, определяющий их как одни лишь представления, и потому ничто не мешает нам приписывать этому трансцендентальному предмету кроме свойства, благодаря которому он является, также причинность, которая не есть явление, хотя результат ее находится тем не менее в явлении. Но всякая действующая причина должна иметь какой-то характер, т.е. закон своей каузальности, без которого она вообще не была бы причиной».[81] Будем считать, что характер причины, определяющий закон ее каузальности, - это и есть выбранная для изучения причины математическая модель. А то, что причина не есть явление, вполне согласуется с высказанным выше положением о том, что никакой математический метод не способен ответить на вопрос, что в рассматриваемых двух явлениях есть причина, а что – следствие. И то соединение чувственного созерцания с категориями рассудка, которое, по Канту, дает подлинное знание, при нашем подходе, вероятно, надо отождествить с органическим единством этапов применения математического метода (см. п. 1.7): формирования априорных соображений, дающих основу для выбора метода (условно это – чувственное созерцание), выбор алгоритма (использование категории рассудка), интерпретация результатов его применения, с возможной корректировкой результата (снова – чувственное созерцание)[82].

Итак, несмотря на обилие работ, посвященных осмыслению понятия причины, ответ на вопрос о том, что такое причина, до сих пор остается открытым, единого мнения по этому поводу не существует. Тем не менее, исследователи все же как-то выходят из положения, изучают системы причинно-следственных отношений, которые обычно считают объективно существующими. И делается это, как правило, с помощью тех или иных математических (иногда – логических) методов. Подробнее об этом – в следующем параграфе, где мы используем ряд описанных выше идей, связанных с пониманием причинности разными исследователями, идей, формализованных в той или иной степени в современном логико-математическом обеспечении социологических исследований.

12.2. Невозможность полностью формализовать понятия причины и следствия. Выделение двух основных направлений изучения причинно-следственных отношений: построение структурных уравнений и проведение эксперимента.

Ниже мы позволим себе упомянуть наименования нескольких методов анализа данных, пока вряд ли знакомых нашему читателю. Как и выше, мы делаем это сознательно, для того, чтобы студент получил хотя бы поверхностное представление о современном методном арсенале социолога.

Какое бы из известных определений причинности мы ни взяли, можно с уверенностью сказать, что не существует логико-математического метода, позволяющего полностью формализовать подход к выявлению причин, определяющих изучаемое исследователем явление (здесь, вероятно, требуется пояснить, почему мы говорим не только о математическом, но и о логическом анализе; однако это станет ясно из дальнейшего – логическими являются, например, правила причинного вывода на основе анализа результатов эксперимента, предложенные Миллем, о чем пойдет речь ниже).

Известно, что никакой формальный логико-математический анализ не может нам *доказать*, что такой-то признак (признаки) является причиной такого-то явления. Тем не менее, использование

---

[81] Антология мировой философии Т.3. М.: Мысль, 1971. С. 137.

[82] В «оправдание» осуществленного нами перетолковывания идей Канта отметим, что современные исследователи, говоря о том, какие идеи немецкого ученого имеют всеобщее, непреходящее значение, зачастую выделяют моменты, которыми мы по существу и воспользовались. Имеются в виду положения о том, что разум во всех своих начинаниях должен подвергать себя критике; что необходимо исследовать прежде всего внутренние основания знания; что догматизация любых положений науки должна быть исключена; что кантовская категориальная структура любого эмпирического знания есть теория опыта; что исследование субъективных форм познания (в том числе причинности как априорного понятия рассудка) необходимо даже в том случае, если исследователь – материалист и не ставит под сомнение объективную реальность причинности (100 этюдов о Канте. М.: КДУ, 2005. С.76-77).

логико-математического формализма – это единственный подход, позволяющий проверять гипотезы соответствующего плана, корректировать, принимать или отвергать их.

Логико-математических методов, позволяющих изучать причинно-следственные отношения, очень много. Сюда входят очень разнородные подходы, начиная с расчета парных коэффициентов связи (число которых, в свою очередь, измеряется сотнями) и кончая сложными комплексными методами изучения структур связей между переменными, опирающимися на разного рода причинные, регрессионные, логлинейные, факторные, дисперсионные и другие модели (отдельные примеры методов, позволяющих изучать причинные отношения между двумя переменными, мы приводили в п. 11.3).

Как мы уже неоднократно упоминали, в каждом математическом методе заложена определенная модель того явления, которое с помощью этого метода изучается. Встает вопрос о том, чем отличается один метод от другого, т.е. вопрос о сравнении моделей, заложенных в каждом из них. В рамках данной книги мы не будем подробно говорить обо всех этих методах. Выделим два главных направления.

Первое направление поиска причин связано с построением т.н. структурных уравнений. Речь идет о подходе, в соответствии с которым, на основе глубокого априорного анализа гипотетических причинных отношений между большим количеством переменных составляется система регрессионных моделей, анализ которых дает возможность четко понять, какие именно глубокие, опосредованные причины обусловливают ту или иную статистическую связь. Этот подход ранее назывался причинным, путевым анализом. Затем в соответствующие модели начали включать т.н. латентные (т.е. скрытые, не поддающиеся измерению с помощью непосредственного опроса респондентов). В результате подход объединил в себе, помимо своеобразных приемов использования регрессионного анализа, еще и факторный и латентно-структурный анализ. И подход стал коротко обозначаться аббревиатурой SEM (моделирование структурными уравнениями).

Представляется, что заложенные в указанном подходе модели близки к тому пониманию причинности, которое отвечает описанному выше статистическому подходу.

Второе направление отвечает поиску причин с помощью проведения эксперимента. Здесь выделим два поднаправления. Одно – статистическое. Современная наука включает в себя такой раздел, который называется «Планирование эксперимента». И опирается соответствующий подход на идеи дисперсионного анализа. Как уже было сказано, дисперсионный анализ отличается от ряда других подходов к выявлению каузальных отношений именно тем, что здесь речь идет о выявлении причинно-следственных отношений через эксперимент. При этом проверяются определенного рода статистические гипотезы. Конкретнее о «понимании» причины именно дисперсионным анализом пойдет ниже, в главах 14 и 15.

Второе поднаправление – индуктивное. Оно, как правило, не использует статистических предпосылок. Но мы его рассмотрим, во-первых, из-за его активного использования в социологии и, во-вторых, из-за того, его анализ даст возможность больше оттенить специфику (достоинства и недостатки) подхода математической статистики. Перейдем к более подробному описанию указанных поднаправлений.

12.3. Роль математической статистики в проведении эксперимента. Нестатистический (индуктивный) подход: эксперимент по Миллю

Надеемся, читатель-социолог помнит, что еще со времен Конта эксперимент в социологии считается одним из основных способов получения социологического знания. Но тот же Конт говорил и о сложности соответствующего процесса, его специфическом характере по сравнению с экспериментом в естественных науках.

Наверное, нетрудно понять, что эксперимент в социологии – дело не простое. Этой теме посвящено довольно много литературы. Ниже мы лишь очень коротко упомянем о наиболее острых проблемах, стоящих перед социологом, желающим получить новое знание с помощью эксперимента. Желающие познакомиться с этой проблематикой глубже могут воспользоваться библиографией, приведенной в списке добавочной литературы к настоящей теме.

Об основных моментах, обусловливающих специфику эксперимента в социологии, можно прочесть в отечественной литературе. [83] Очень коротко напомним, что такое эксперимент, зачем он нужен и в чем состоит специфика его использования именно в социологии.

---

[83] См.: *Методы сбора* информации в социологических исследованиях. Кн.2. М.: Наука, 1990. С.190-221, а также список работ по эксперименту в социологии в добавочной литературе к настоящей главе.

Воспользуемся несколькими цитатами из работы «Методы сбора информации в социологических исследованиях», кн.2.[84].

«Целью всякого эксперимента является проверка гипотез о причинной связи между явлениями: исследователь создает или изыскивает определенную ситуацию, приводит в действие гипотетическую причину и наблюдает за изменениями в естественном ходе событий, фиксирует их соответствие или несоответствие предположениям, гипотезам» (с.190). Удовлетворение всех предъявляемых к экспериментам требований (а эти требования родились в естественных науках и до сих пор никем не оспаривались) в социологии очень трудно. И основная трудность обычно заключается в обеспечении т.н. внешней и внутренней валидности эксперимента.

Определения упомянутых видов валидности даны в названной выше книге. Однако там нет четких указаний на способы их обеспечения. А такие способы опираются на математические методы. Остановимся на некоторых таких сторонах понятий внутренней и внешней валидности, которые нельзя оценивать без использования математического аппарата.

Внутренняя валидность. Проблема состоит в грамотном выделении таких факторов, которые действительно определяют вариацию интересующего нас признака. Мы должны быть уверены «в том, что именно изучаемый в данном эксперименте фактор, а не какой-либо иной, является причиной зарегистрированного изменения» (с.190). Как мы увидим в 14 и 15 главах, обеспечение такой уверенности дает дисперсионный анализ. Мы рассматриваем интересующие нас вариации значений некоторого признака Y в зависимости именно от изменения значения конкретного рассматриваемого фактора X (в однофакторном дисперсионном анализе, глава 14) или сочетаний значений двух рассматриваемых факторов X_1 и X_2 (в двухфакторном дисперсионном анализе, глава 15).

Внешняя валидность. Проблема состоит в обеспечении возможности переноса результатов с конкретной изученной исследователем экспериментальной ситуации в “живую” жизнь. Мы должны быть уверены «в том, что выявленная зависимость закономерна в определенных условиях, что полученные выводы можно распространять на внеэкспериментальные ситуации» (с.190). Ясно, что основным аппаратом, обеспечивающим подобную уверенность, может (и должна! Другого пути наука пока не создала) служить математическая статистика – построение доверительных интервалов, проверка статистических гипотез и т.д. Именно это делается в дисперсионном анализе, являющемся одним из разделов математической статистики.

Хотелось бы, чтобы особенное внимание читатель уделил следующему обстоятельству, слабо отраженному в социологических публикациях. Одним из первых социологов, обративших внимание на логику доказательства научных гипотез был английский ученый Дж.С. Милль (1806-1873), считающийся основателем позитивизма в Англии, создателем индуктивной логики, последователем Юма. Задача науки, с его точки зрения, - индуктивное упорядочение единичных явлений. Индукцию же «можно коротко определить как «обобщение из опыта». Она состоит в том, что на основании нескольких отдельных случаев, в которых известное явление наблюдалось, мы заключаем, что это явление имеет место и во всех случаях известного класса, т.е. во всех случаях, сходных с наблюдавшимися в некоторых обстоятельствах, признаваемых существенными»[85]. Милль не говорит о том, в чем именно должны быть сходны эти обстоятельства, но использует предположение о том, что «в природе существуют сходные, параллельные случаи, что то, что произошло один раз, будет иметь место при сходных обстоятельствах и вторично, и не только вторично, а всякий раз, как снова встретятся те же самые обстоятельства» [86]. Заметим, что в подобных предположениях прослеживается определенная аналогия с определением вероятности, а именно, с тем, что мы, как правило, не знаем, каков должен быть тот комплекс условий, который входит в это определение.[87]

Опишем предложенные Миллем логические схемы поиска причинных отношений, заимствуя их описание из цитированной выше фундаментальной работы Милля[88].

«Простейших и наиболее очевидных способов выделять из числа предшествующих явлению или следующих за ним обстоятельств те, с которыми это явление действительно связано при помощи неизменного закона, - таких способов два. Один состоит в сопоставлении тех отличных один от другого случаев, в которых данное явление имеет место; другой – в сравнении таких случаев, где это явление присутствует, со сходными в других отношениях случаями, где этого явления, тем не менее,

---

[84] Методы сбора информации в социологических исследованиях. Т.2. М.: Наука, 1990.

[85] Д.С.Милль. Система логики символической и индуктивной, изд.2. М., 1914 (цит. по: Антология мировой философии. Т.3. М.: Мысль, 1971. С. 602)

[86] Там же. С. 602.

[87] Напомним определение вероятности данное в сноске 13 (п. 1.3).

[88] Д.С.Милль. Система логики … С. 604-605. О тех же схемах кратко говорится также в книге «Методы сбора…», названной выше. Кн. 2, с. 191-192.

нет. Первый из этих способов можно назвать «методом сходства» (the Method of Agreement), второй – «методом разницы» или методом различия (the Method of Difference).

ПЕРВОЕ ПРАВИЛО.

Если два или более случаев подлежащего исследованию явления имеют общим лишь одно обстоятельство, то это обстоятельство, - в котором только и согласуются все эти случаи, - есть причина (или следствие) данного явления...

ВТОРОЕ ПРАВИЛО

Если случай, в котором исследуемое явление наступает, и случай, в котром оно не наступает, сходны во всех обстоятельствах, кроме одного, встречающегося лишь в первом случае, то это обстоятельство, в котором одном только и разнятся эти два случая, есть следствие, или причина, или необходимая часть причины явления...

Из этих двух методов «метод различия» есть по преимуществу метод искусственного опыта или эксперимента, тогда как к «методу сходства» прибегают преимущественно тогда, когда эксперимент невозможен.

Третий метод можно назвать «косвенным методом различия», или «соединенным методом сходства и различия» (the indirect Method of Difference, or the joint Method of Agreement and Difference). Он состоит в двойном приложении метода сходства, причем оба доказательства независимы одно от другого и друг друга подкрепляют … Правило для него можно выразить следующим образом:

ТРЕТЬЕ ПРАВИЛО

Если два или более случаев возникновения явления имеют общим лишь одно обстоятельство, и два или более случаев невозникновения того же явления имеют общим только отсутствие того же самого обстоятельства, то это обстоятельство, в котором только и разнятся оба ряда случаев, есть или следствие, или причина, или необходимая часть причины изучаемого явления.»

Кроме того, Милль вводит «метод остатков» и «метод сопутствующих изменений» (the Method of concomitant Variations) с помощью следующих правил:

«ЧЕТВЕРТОЕ ПРАВИЛО

Если из явления вычесть ту его часть, которая, как известно из прежних индукций, есть следствие некоторых определенных предыдущих, то остаток данного явления должен быть следствием остальных предыдущих.

ПЯТОЕ ПРАВИЛО

Всякое явление, изменяющееся определенным образом всякий раз, когда некоторым особенным образом изменяется другое явление, есть либо причина, либо следствие этого явления, либо соединено с ним какой-либо причинной связью.»

Вряд ли требует особого доказательства утверждение о том, что Милль, во-первых, говорит об индуктивном выводе и, во-вторых явно следует тому положению Юма, в соответствие с которым понятие причинности возникает просто в силу возникновения у нас привычки за одной перцепцией наблюдать другую.

Миллевским рассуждениям уделяли огромное внимание русские социологи 19-го века. А.А.Чупров (о котором шла речь выше), изучал Милля, еще будучи гимназистом. Большое внимание творчеству Милля уделил Н.Г.Чернышевский[89]. При этом логика Милля рассматривалась русскими учеными в основном именно как логика выявления причинно-следственных отношений (через эксперимент).

Слабо известным отечественным социологам является тот факт, что в наше время эти схемы были рассмотрены группой ученых под руководством В.К.Финна как основа того направления науки, которое связано с созданием искусственного интеллекта. Идеи Милля получили дальнейшее развитие. Был разработан пакет компьютерных программ, названный авторами “ДСМ-система” (по инициалам английского ученого)[90] Здесь по существу речь идет о порождении гипотез о причинных

---

[89] Популярность Милля в среде русской интеллигенции косвенно подтверждается тем, что его имя очень часто употребляется в русской художественной литературе второй половины XIX века в самых разных контекстах. Так, в числе мелких журнальных произведений А.П.Чехова имеется такая острота: «Превышение власти и административный произвол дантиста заключаются в вырывании здорового зуба рядом с больным. Это сказал околоточный, читая Логику Милля» (Антон Чехов. Собрание сочинений. М.-Л.: Гос. изд-во, 1929. Т.IV. С.342).

[90] См., например: *Финн В.К.* Интеллектуальные системы и общество: идеи и понятия // Научно-техническая информация, сер.2, 1999, №10, с. 6-20.

отношениях на основе анализа данных. Тем не менее, и эти рассуждения можно считать касающимися проведения своеобразных экспериментов.[91]

Подчеркнем, что в работах коллектива, руководимого Финном, не предполагается вероятностное порождение исходных данных. Исходные данные не считаются выборкой из какой бы то ни было генеральной совокупности. Рассматриваемые значения признаков не считаются реализациями значений каких-то известных случайных величин. Поэтому традиционные для математической статистики вопросы о переносе результатов с выборки на генеральную совокупность даже не ставятся. То же можно сказать и о логических схемах Милля. Тем не менее, мы говорим об этом подходе именно в данном учебнике. Причин, по крайней мере, три. Во-первых, с педагогической точки зрения неправильно было бы отделять друг от друга описания методов, решающих сходные в содержательном плане задачи. Во-вторых, предположение о вероятностном происхождении данных в социологических задачах довольно часто должно подвергаться сомнению (об этом мы уже упоминали в п. 1.3). И если это сомнение исследователь сочтет оправданным, он должен тут же попытаться решить свою задачу (в данном случае речь идет об анализе причинно-следственных отношений) каким-то другим способом, не использующим классические математико-статистические приемы. Сделать это и позволяет, скажем, та же система Финна. В-третьих, понять, что такое статистический подход, можно только в сравнении его с нестатистическим. В этом параграфе мы говорим о нестатистических способах проведения эксперимента с целью изучения причинно-следственных отношений, в темах 14 и 15 – о статистических.

Сказанное подтверждает, что понятие эксперимента имеет много аспектов. Так, проблемы, рассматриваемые Миллем, казалось бы, совсем не похожи на те, которые решает дисперсионный анализ. В действительности же речь идет о разных возможностях изучения одного и того же процесса – причинно-следственного воздействия одного явления на другое. Каждый подход имеет свои достоинства и недостатки. Каждый отвечает своему пониманию того, что такое причина. Как мы уже отмечали, в литературе имеется огромное количество определений этого понятия. И имеется острая нехватка работ, в которых серьезно бы анализировались те понимания причинности, которые стоят за тем или иным математическим аппаратом. По нашему мнению, без особой натяжки можно сказать, что речь идет о необходимости анализа той существующей в сознании человека априорной формы познания, которую Кант связывал с законом причинности. Мы сейчас делаем упор не на слове «априорный», полагая, что положение об априорности спорно, а на том, что прежде, чем исследователь приступит к изучению причинности с помощью логико-математических методов, у него в сознании должно сформироваться представление о том, что такое причинность. И это представление должно лечь в основу выбора способа изучения причинности. Если даже исследователь не задумывается о таких вопросах, все равно, выбирая тот или иной математико-логический способ изучения причинности, он волей-неволей пользуется некой априорной ее моделью – той, которая была придумана автором используемого метода.

На повестке дня науки (речь идет о методологии социологического исследования) явно стоит вопрос о разработке методики комплексного использования разных способов выявления и оценивания причинно-следственных отношений при решении одной и той же социологической задачи.

Примеры задач

1. Приведите примеры модельных ситуаций, когда причинно-следственные отношения могут быть установлены с помощью приведенных в главе 12 правил Милля.
2. Приведите примеры нескольких методических задач, встающих при построении социологического инструментария (см. сноску 55 и задачу 4 из темы 10), которые можно было бы решить с помощью эксперимента и проверки статистических гипотез.

Добавочная литература к главе 12

О понятии причины

Аристотель. Метафизика // Аристотель. Сочинения. Т.1. М.: Мысль, 1976. С.63-367.

[91] Учебный план ГУ-ВШЭ, в соответствии с которым написан данный учебник, предполагает изучение студентами четвертого курса, специализирующимися по специальности «Прикладные методы социологических исследований», дисциплины «Логико-комбинаторные методы анализа данных», программа которой базируется как раз на методах ДСМ-системы.

Давыдов Ю.Н. Макс Вебер и современная теоретическая социология. М.: Мартис, 1998. С. 150-190.

И.Кант. Теория познания // Антология мировой философии Т.3. М.: Мысль, 1971. С. 100-155.

Лаплас. Опыт философии теории вероятностей // Вероятность и математическая статистика. Энциклопедия. М.: Большая российская энциклопедия, 1999.

Милль Д.С. Система логики символической и индуктивной, изд.2. М., 1914 // Антология мировой философии. Т.3. М.: Мысль, 1971. С. 594-605.

Финн В.К. Интеллектуальные системы и общество. М.: КомКнига, 2006

Чупров А. А. Нравственная статистика // Брокгауз Ф.А. (Лейпциг), Ефрон И.А. (СПб.). Энциклопедический словарь. Т. XXI. С.-Петербург: Типолитография И.А.Ефрона, 1897. С. 403–408.

Чупров А.А. Вопросы статистики. М.: Госстатиздат ЦСУ СССР, 1960.

Юм Д. Сокращенное изложение «Трактата о человеческой природе» // Антология мировой философии. Т.2. М.: Мысль, 1970. С. 575-593.

Об эксперименте в социологии

Адлер Ю.П., Ковалёв А.Н. Математическая статистика и планирование эксперимента в науке о человеке // Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. М.: Прогресс, 1976. С. 477-490

Андреева Г.М. Социальная психология. М.: Наука, 1994. С.174-175 (Хоторнский эксперимент)

Батыгин Г.С. Лекции по методологии социологических исследований. М.: Аспект Пресс, 1995. С. 190-227

Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. М.: Прогресс, 1976, с. 405-436 однофакторный и многофакторный дисперсионный анализ: случайные, смешанные и постоянные эффекты), с. 437-458 (Основы планирования эксперимента).

Джонсон Н., Лион Ф. Статистика и планирование эксперимента в технике и науке. М.: Мир,т.1 - 1980, т.2 - 1981

Дэниел К. Применение статистики в промышленном эксперименте. М.: Мир, 1979

Кравченко А.И. Хоторнский эксперимент //Справочное пособие по истории немарксистской западной социологии. М.: Наука, 1986. С.452-453

Куприян А.П. Методологические проблемы социального эксперимента. М.: Наука, 1971

Куприян А.П. Проблема эксперимента в системе общественной практики. М.:Наука, 1981

Кемпбелл Д. Модели экспериментов в социальной психологии и прикладных исследованиях. М.: Прогресс, 1980. Переиздано в СПб: Изд-во «Социально-психологический центр», 1996

Киш Л. Представительность, рандомизация и контроль // Математика в социологии. М.: Мир, 1977, с. 201-223.

Методы сбора информации в социологических исследованиях. М.: Наука, 1990.Т.2. С. 190-214

Методы сбора данных: анализ документов, наблюдение, эксперимент. М.: ИСИ АН СССР, 1985

Методы социальной психологии. Л.: изд-во ЛГУ, 1977. С.132-150

Милграм С. Эксперимент в социальной психологии. Спб: Питер, 2000

Монсон П. Современная западная социология. Теории, традиции, перспективы. С.Пб.: Нотабене, 1992. с.160-162 (эксперимент Цимбардо)

Налимов В.В. Теория эксперимента. М.: Наука, 1971

Основы прикладной социологии. М.: Интерпракс, 1996. С.68-72

Рабочая книга социолога. М.: Наука, 1983. С. 411-432

Руткевич М.Н. Макросоциология. Методологический очерк. М.: РАН, 1995. С.16-19

Статистические методы анализа информации в социологических исследований. М.: Наука, 1979. С.178-194

Хагуров А.А. Социальный эксперимент: логико-методологические и социальные проблемы. Ростов-на-Дону,1989

Хикс Ч. Основные принципы планирования эксперимента. М.: Мир, 1967

Ядов В.А. Социологическое исследование: методология, программа, методы. Самара: «Самарский ун-т», 1995. С.220-231

Blalock H.M. Causal inferences in nonexperimental research. Chapel Hill: University of North Carolina Press, 1964

О причинном анализе

Бестужев-Лада И. В. , Варыгин В. Н. , Малахов В. А. Моделирование в социальных исследованиях. М. : Наука, 1978.

Богомолова Е.А., Наумова Н.Ф. Структурные модели как инструмент обобщения и интерпретации социальной информации на выходе системы моделирования // Неформализованные элементы системы моделирования. М.: ВНИИСИ, 1980

Бородкин Ф.М. Об одной схеме причинного анализа // Математика и социология. Новосибирск: ИЭиОПП СО АН СССР, 1970

Бородкин Ф.М. Научный эксперимент в социально-экономических исследованиях. Дисс. На соискание ученой степени доктора экономических наук. Новосибирск, 1975

Бунге М. Причинность. М.: Изд. Иностр. лит., 1962

Волд Г. Путевые модели с латентными переменными: подход NIPALS // Математика в социологии: моделирование и обработка информации. М.: Мир, 1977. С. 241-281.

Гаврилец Ю. Н. Структура связей и причинные зависимости между переменными // Математика в социологии: моделирование и обработка информации. М. : Мир, 1977. С. 135-169.

Елисеева И. И. Статистические методы измерения связей. Л. : Изд-во ЛГУ, 1982. С. 97-108. (Структурные модели. Путевой анализ).

Елисеева И. И. , Рукавишников В. О. Логика прикладного статистического анализа. М. : Финансы и статистика, 1982. С. 72-149 (Структурные и причинные модели).

Левин К. Теория поля в социальных науках. СПб: Сенсор, 2000, с. 7-14, 178-192, 213-220

Математические методы анализа и интерпретация социологических данных. М. : Наука, 1989. С.61-94 (Выбор стратегии анализа взаимосвязи признаков)

Новак С. Причинные интерпретации статистических связей в социальном исследовании // Математика в социологии: моделирование и обработка информации. М. : Мир, 1977. С. 76-123.

Осипов Г. В. , Андреев Э. П. Методы измерения в социологии. М. : Наука, 1977.

Осипов Г.В., Андреенков В.Г. Эмпирическое обоснование гипотез в социологических исследованиях // Социс, 1974, № 1. С. 160-173.

Рукавишников В.О. Информационный подход к причинному анализу // Модели социально-экономических процессов и социальное планирование. М., 1979.

Социальные исследования: построение и сравнение показателей. М.: Наука, 1978. С.104-111 (Построение показателей в процессе применения метода причинных моделей).

Статистические методы анализа информации в социологических исследованиях. М. : Наука, 1979. С. 267-282 (Модели для анализа структуры причинных связей).

Суппес П. Вероятностный анализ причинности // Математика в социологии: моделирование и обработка информации. М. : Мир, 1977. С. 50-75.

Таганов И.Н. Информационные меры причинного влияния // Математика в социологии: моделирование и обработка информации. М. : Мир, 1977. С. 124-134.

Татарова Г. Г. Структура многомерной случайной величины и проблема взаимосвязи признаков // Социологические исследования, 1986. N3. С. 142-148.

Трофимов В.П. Измерение взаимосвязей социально-экономических явлений. М.: Статистика. 1975. С. 15-29 (Соотношение причинной и корреляционной связи).

Хейс Д. Причинный анализ в статистических исследованиях. М. : Финансы и статистика, 1981.

Blalock H. M. Causal Inferences in Nonexperimental Research, Chapel Hill: university of North Carolina Press, 1964, c. 3-60, 95-96, 172-188

Blalock H. M. Causal models in the Social Sciences, 1970

Blalock H. M. Theory construction. From verbal to mathematical formulation. Prentice hall, New Jersey, 1969, p. 1-30

Bollen, K. A. Structural Equations with Latent Variables. NY: John Wiley & Sons, 1989

Knoke D. A causal model for the political party preferences of american men // Amer. Soc. Review, 1972. P. 679-689

Spilerman S. Forecasting social events // Social indicator model, S.N.Y. 1975

Suppes P. Probabylistic Theory of Causality. Amsterdam: North-Holl P/ Co.,1970

Sage University Paper series on Quantitative Applications in the Social Sciences:

n. 07-003. Asher H. Causal modeling, 1976,1980

n. 07-034. Long. Covariance Structure Models, 1983

n. 07-037. Berry W.D. Nonrecursive Causal Models, 1984

n. 07-055. Davis. The Logic of causal Order, 1985

n. 07-074. Brown, Melamed. Experimental design and analysis, 1990

n. 07-105. Finkel S.E/ Causal analysis of panel data, 1995

n. 07-114. Jaccard J., Wan C.K. LISREL Approaches to Interaction Effects in Multiple regression

n. 07-135. Jaccard J. Interaction effects in logistic regression, 2001

O SEM

Golob Thomas F. 2003. Structural Equation Modeling for Travel Behavior Research. Center for Activity Systems Analysis. Published in: Transportation Research, Vol. 37B, 2003, pp. 1-15.

Hox J.J. Bechger T.M. 1998. An Introduction to Structural Equation Modeling. Family Science Review, 11,354-373.

McArdle John J. , Johnson Ronald C. 2001. Structural Equation Modeling of Group Differences in CES-D Ratings of Native Hawaiian and Non-Hawaiian High School Students Journal of Adolescent Research, Vol. 16 No. 2, March 2001 108-149 Sage Publications, Inc.

Mueller R.O. 1996. Basic Principles of Structural Equation Modeling. Springer Verlag, January.

Тема 13

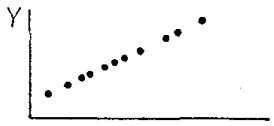
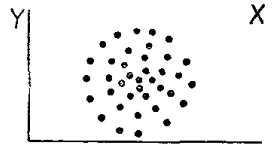
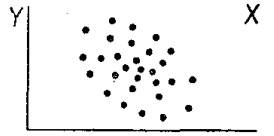
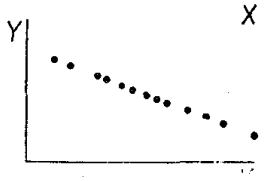
Корреляционное отношение

13.1. Линейная и нелинейная связи. Границы применимости коэффициента корреляции как показателя связи между изучаемыми переменными

Как известно, коэффициент корреляции Пирсона r измеряет лишь степень *линейной* связи между X и Y. Если он близок к 1 (или к –1), то мы можем быть уверены, что наши признаки связаны и что эта связь линейна. Близость же r к нулю может объясняться не только отсутствием связи, но и тем, что эта связь нелинейна (близость r к нулю говорит об отсутствии *линейной* связи). Рассмотрим таблицу 13.1.

Таблица13.1[92]

Интерпретация значений r_{xy}

Величина r_{xy}	Описание линейной связи	Диаграмма рассеивания
+1,00	Строгая прямая связь	
Около +0,50	Слабая прямая связь	
0,00	Нет связи (то есть ковариация X и $Y = 0$)	
Около —0,50	Слабая обратная связь	
—1,00	Строгая обратная связь	

Когда коэффициент корреляции равен нулю, возможны разные ситуации: статистическая связь может вообще отсутствовать (см. рис. 13.1), а может существовать, и даже быть довольно сильной, но криволинейной ((рис. 13.2).

[92] См.: Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. М.: Прогресс, 1976. С. 110.

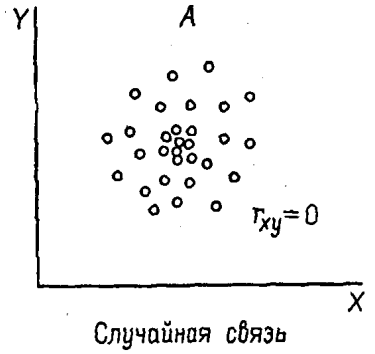


Рис. 13.1. Ситуация равенства нулю коэффициента корреляции: отсутствие статистической связи [Гласс, Стэнли, 1976, с.116, рис. 7.4]

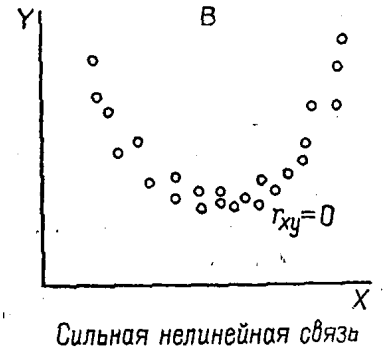


Рис. 13.2. Ситуация равенства нулю коэффициента корреляции: наличие сильной нелинейной статистической связи[93] .

Рассмотрим еще одну меру связи – меру, используемую для измерения связи в тех случаях, когда эта связь имеет произвольный вид, в частности, может быть нелинейной.

13.2. Корреляционное отношение. Общее представление о внутригрупповом и межгрупповом разбросе

Сущность новой меры связи - корреляционного отношения - продемонстрируем на примере, заимствованном из той же работы Гласса и Стэнли[94].
Изучается зависимость результатов ответа респондента на вопросы некоторого теста (Y) от его возраста (X). Опрашивалось 28 человек. По возрасту они были разделены на 8 групп. В представленных ниже данных каждая группа характеризуется средним возрастом попавших в нее респондентов (например, возраст 30 приписан всем людям, попавшим в возрастной интервал от 28 до 32 лет). Статистические данные представлены на рис. 13.3 и в таблице 13.2.

[93] Там же. С. 116, рис. 7.4.

[94] Там же. С. 138-139.

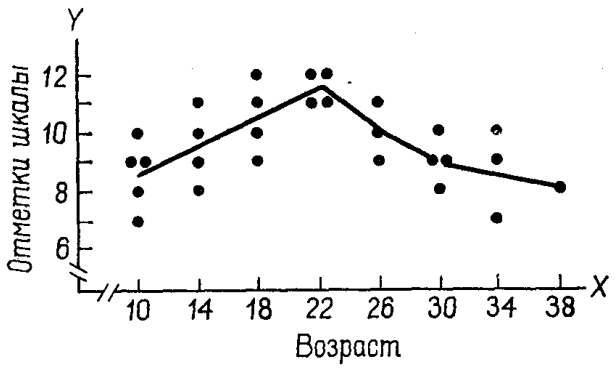


Рис. 13.3. Пример данных для расчета корреляционного отношения: связь между возрастом и характеристикой 28 людей по вспомогательному тесту цифра-знак шкалы интеллекта взрослых Векслера (WAIS)[95].

Таблица 13.2. Ответы респондентов на заданный тест в зависимости от их возраста (данные для расчета корреляционного отношения)[96]

Усредненные данные о группах	Возраст с точностью до ближайшего года							
	10	14	16	22	26	30	34	38
	$Y_{11}=7$ $Y_{21}=8$ $Y_{31}=9$ $Y_{41}=9$ $Y_{51}=10$	$Y_{12}=8$ $Y_{22}=9$ $Y_{32}=10$ $Y_{42}=11$	$Y_{13}=9$ $Y_{23}=10$ $Y_{33}=11$ $Y_{43}=12$	$Y_{14}=11$ $Y_{24}=11$ $Y_{34}=12$ $Y_{44}=12$	$Y_{15}=9$ $Y_{25}=10$ $Y_{35}=11$	$Y_{16}=8$ $Y_{26}=9$ $Y_{36}=9$ $Y_{46}=10$	$Y_{17}=7$ $Y_{27}=9$ $Y_{37}=10$	$Y_{18}=8$
Средние возрастных групп $\bar{Y}_{\bullet j}=\frac{\sum_{i=1}^{n_j} Y_{ij}}{n_j}$	8,60	9,50	10,50	11,50	10,00	9,00	8,67	8,0
Число членов группы (n_j)	$n_1=5$	$n_2=4$	$n_3=4$	$n_4=4$	$n_5=3$	$n_6=4$	$n_7=3$	$n_8=1$
Общее среднее всех значений	$\bar{Y}_{\bullet\bullet}=\frac{269}{28}=9{,}61$							

[95] Там же. С. 138.
[96] Там же. С.139.

Y_{ij} – значение зависимого признака для i–го респондента в j–й возрастной группе; i = 1,…,n_j , где n_j - число членов j–й группы; j = 1,…,J, где J – количество выделенных групп (в данном случае J = 8).

На рис. 13.3 видно, что наблюдаемые точки довольно плотно расположены вокруг нарисованной там ломаной линии. Это наводит на мысль о том, что указанная ломаная линия отражает определенную тенденцию: с изменением возраста от 10 до 22 лет показатели людей по рассматриваемому тесту растут, затем начинается спад. И говорить об этой тенденции мы можем только благодаря тому, что (1) если для каждого рассматриваемого значения возраста вычислить среднее арифметическое значение зависимой переменной, то наша ломаная линия пройдет через соответствующую точку; (2) для каждой возрастной группы разброс значений теста вокруг упомянутой средней является относительно небольшим. Это интуитивное соображение можно формализовать. Соответствующая формализация и лежит в основе рассматриваемого коэффициента связи.

Введем несколько новых понятий. Обозначим через $\overline{Y}_{\bullet j}$ среднее арифметическое значение зависимой переменной для j-й группы (j=1,…,J). Отметим, что точка перед индексом j в обозначении среднего арифметического означает, что по первому индексу величины Y_{ij} произошло суммирование.

ОПР. Назовем *внутригрупповой суммой квадратов* величину

$$SS_w = SS_{внутри} = \sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_{\bullet j})^2$$

Для нашего примера эта сумма будет равна:

$(7\text{-}8{,}60)^2 + (8\text{-}8{,}60)^2 + (9\text{-}8{,}60)^2 + (9\text{-}8{,}60)^2 + (10\text{-}8{,}60)^2 +$
(для первой группы)

$(8\text{-}9{,}50)^2 + (9\text{-}9{,}50)^2 + (10 - 9{,}50)^2 + (11\text{-}9{,}50)^2 + \ldots$

(для второй группы)

$+ \boxed{(8\text{-}8{,}00)^2} = 24,\ 87$

(для восьмой группы)

Обозначим через $\overline{Y}_{\bullet\bullet}$ среднее арифметическое всех значений независимого признака. Очевидно, что имеет место соотношение:

$$\overline{Y}_{\bullet\bullet} = \frac{\sum_{j=1}^{J}\sum_{i-1}^{n_j} Y_{ij}}{n},$$

где $n = \sum_{j=1}^{J} n_j = n_1 + n_2 + \ldots + n_J$ - объем выборки.

ОПР. Назовем *общей суммой квадратов* величину

$$SS_t = SS_{общ} = \sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_{\bullet\bullet})^2$$

ОПР. Назовем *корреляционным отношением* разность

$$\eta^2_{y/x} = 1 - (SS_{внутри} / SS_{общ})$$

Для рассмотренного примера, как нетрудно проверить, имеют место соотношения

$$SS_{общ} = 54, 68; \eta^2_{y/x} = 1 - (24,87/ 54,68) = 1 - 0,455 = 0,545.$$

Поясним смысл корреляционного отношения. Смысл любого коэффициента бывает легче понять, если рассмотреть, при каких условиях он принимает максимальное (в данном случае – 1) и минимальное (0) значение. Ясно, что $\eta^2_{y/x} = 1$, когда $SS_{внутри} = 0$, т.е. когда в каждой выделенной по признаку X группе (в нашем случае – в каждой возрастной группе) значения признака Y одинаковы. Для рассматриваемого примера это означает, что все точки лежат на выделенной ломаной линии. Ясно, что это действительно говорит о наличии криволинейной связи.

Что касается равенства $\eta^2_{y/x} = 0$, то оно имеет место в том случае, когда $SS_{внутри} = SS_{общ}$, т.е. когда фиксация признака X нисколько не уменьшает разброс признака Y. Ясно, что это говорит об отсутствии связи: получение информации об X не увеличивает информацию об Y. Здесь напрашивается аналогия с принципом построения коэффициентов связи, основанных на прогнозных моделях (Толстова, 2000).

Коэффициент $\eta^2_{y/x}$ является мерой степени предсказания Y по X с помощью "наилучшим образом подобранной" линии, либо прямой, либо кривой.

Заметим, что $\eta^2_{y/x} \neq \eta^2_{x/y}$ (заметим, что о подобной перестановке признаков можно говорить только в том случае, если оба признака измерены по интервальной шкале; хотя для измерения одного из означенных коэффициентов, скажем, $\eta^2_{Y/X}$, достаточно того, чтобы Y был интервальным, X может быть и номинальным).

Поясним это на нашем примере: если человеку 10 лет, то можно довольно уверенно предсказать, что результатом тестирования для него будет балл, равный примерно 8,60. Однако если некий человек получил балл 8,60, то его возраст может быть с одинаковой вероятностью как малым (10 лет), так и большим (38 лет). Значит, можно довольно хорошо предсказать Y по X, но нельзя хорошо прогнозировать X по Y. Это неизбежно отражается на величинах $\eta^2_{y/x}$ и $\eta^2_{x/y}$: $\eta^2_{y/x}$=0,545, а $\eta^2_{x/y}$ близка к нулю. Не будем ее вычислять. Такое вычисление потребовало бы перегруппировки данных. Ячейки должны были бы быть организованы по результатам тестирования (скажем, можно было бы сформировать три ячейки - в первую включить респондентов, получивших баллы 7-8, во вторую – баллы 9-10, в третью – баллы 11-12). А в качестве значений Y выступали бы возраста респондентов, вошедших в ту или иную ячейку.

Приведем еще один пример.

Пример. Дана частотная таблица

Возраст (X)	Зарплата (Y)		
	500-700	700-900	900-1100
18-22	10	5	5
22-26	10	10	20
26-30	5	20	20

Рассчитать корреляционное отношение $\eta^2_{y/x}$

Решение. Вспомним, что, разбив диапазон изменения признака на интервалы и составив частотную таблицу, мы потеряли исходную информацию и вынуждены считать, что респонденты, попавшие в один интервал, имеют одну и ту же зарплату, отвечающую середине этого интервала.
Расположим данные в более привычном (более часто использующемся при нахождении корреляционного отношения) виде. Правда, не будем выписывать конкретные зарплаты (Y) для людей, попавших в ту или иную возрастную группу (возраст – X), а укажем, сколько человек обладают тем или иным значением.

Интервал Y-ка	Середина интервала	I группа (возр. 18-22)	II группа (возраст 22-26)	III группа (возраст 26-30)
500-700	600	10 человек	10 человек	5 человек

700-900	800	5 человек	10 человек	20 человек
900-1100	1000	5 человек	20 человек	20 человек

Общее среднее по Y: $Y_{..} = (600*(10+10+5) + 800*(5+10+20) + 1000*(5+20+20))/100 = 880$

$SS_{общ} = (600-880)^2(10+10+5) + (800-880)^2(5+10+20) + (1000-880)^2(5+20+20) = 2\,832\,000$

$Y_{.1} = (600*10 + 800*5 + 1000*5)/20 = 750$

$Y_{.2} = (600*10 + 800*10 + 1000*20)/40 = 850$

$Y_{.3} = (600*5 + 800*20 + 1000*20)/45 = 866{,}7$

$SS_{внутри} = (600-Y_{.1})^2*10 + (800-Y_{.1})^2*5 + (1000-Y_{.1})^2*5 + (600-Y_{.2})^2*10 + (800-Y_{.2})^2*10 + (1000-Y_{.2})^2*20 + (600-Y_{.3})^2*5 + (800-Y_{.3})^2*20 + (1000-Y_{.3})^2*20 = 150^2*10 + 50^2*5 + 250^2*5 + 250^2*10 + 50^2*10 + 150^2*20 + 267^2*5 + 67^2*20 + 133^2*20 = 2\,449\,945$

$\eta^2_{y/x} = 1 - SS_{внутри}/SS_{общ} = 1 - 0{,}845;\ \ \eta_{y/x} = 0{,}367.$

13.3. Проблемы, не решаемые с помощью корреляционного отношения

Корреляционное отношение так же, как и коэффициент корреляции, не решает всех вопросов, возникающих у исследователя при решении задач, связанных с изучением влияния друг на друга каких-либо двух признаков. На некоторые вопросы анализ корреляционного отношения в принципе не дает ответа. Укажем два таких вопроса.

Во-первых, корреляционное отношение относится к числу т.н. интегральных коэффициентов связи. Информация по всем значения признака X здесь как бы усредняется. Не учитывается, что разные градации могут по-разному «влиять» на Y: при одном значении X признак Y может быть жестко детерминирован, при другом – соответствующие значения Y могут быть очень разбросаны. Вопрос о том, какие именно значения X действительно можно считать детерминирующими уровень Y, здесь не ставится.

Во-вторых, использование корреляционного отношения не предполагает постановку вопроса о возможности определения уровней Y не одним, а несколькими независимыми переменными.

Ответы на эти вопросы в достаточно полной мере дает дисперсионный анализ. Он будет рассмотрен ниже (Темы 14 и 15).

13.4. Соотношение между разными видами сумм квадратов

Для того, чтобы идеи дисперсионного анализа легче воспринимались читателем, уже здесь введем новое понятие и продемонстрируем его сущность на рассмотренном выше примере.

ОПР. Назовем *межгрупповой суммой квадратов* величину

$$SS_b = SS_{между} = \sum_{j=1}^{J} n_j (\overline{Y}_{\bullet j} - \overline{Y}_{\bullet\bullet})^2$$

Содержательный смысл этого понятия представляется очевидным: оно характеризует величину разброса групповых средних вокруг общего среднего с поправкой, состоящей в том, что вклад в этот разброс тех групповых средних, которым отвечают более многочисленные группы, считается более сильным, чем вклад средних малочисленных групп.

Нетрудно доказать соотношения:

$$SS_{общ} = SS_{внутри} + SS_{между}$$
$$\eta^2_{y/x} = 1 - (SS_{внутри}/SS_{общ}) = SS_{между}/SS_{общ}$$

Как следует из приведенных формул, для определенных выше сумм квадратов могут использоваться разные обозначения (могут задействоваться начальные буквы соответствующих и русских, и английских слов):

$SS_{общ} = SS_t$ (от англ "total"); $SS_{внутри} = SS_w$ (от "within"); $SS_{между} = SS_b$ (от "between") .

Само сочетание SS – начальные буквы слов "sum square".

Примеры задач

1. Доказать соотношение: $SS_b + SS_w = SS_t$ ($SS_{между} + SS_{внутри} = SS_{общ}$)[97]

2. Диапазон изменения некоторого признака X разбили на три интервала. Их середины равны, соответственно, 10, 20 и 30. Диапазон изменения признака Y разбили на два интервала с серединами 15 и 25. Считаем, что все значения признака, попавшие в один интервал, приравниваются к его середине. Для респондентов некоторой выборки из 50 человек двумерная частотная таблица имеет следующий вид:

Середина интервала	Середина интервала признака X		
признака Y	10	20	30
15	4	28	6
25	6	0	6

Рассчитать корреляционные отношения $\eta^2_{y/x}$ и $\eta^2_{x/y}$

Добавочная литература к теме 13.

Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. М.: Прогресс, 1976. С.138-141[98].

*Haggard E.A.*Intra-class Correlation and the Analysis of Variance. N.-Y., Dryden Press,1958

Rozeboom W.W. Foundations of the Theory of Prediction. Homewood, Illinois, Dorsey Press, 1966

ТЕМА 14
Однофакторный дисперсионный анализ.

14.1. Однофакторный дисперсионный анализ как метод анализа результатов эксперимента при изучении причинно-следственных отношений

Перейдем к рассмотрению того, что дает дисперсионный анализ исследователю, желающему изучать каузальные отношения на базе осуществления эксперимента.

Рассмотрим следующую задачу. Предположим, что некоторый вуз имеет возможность использовать разные формы обучения, и мы хотим изучить, зависит ли качество усвоения студентом некоторой учебной программы от того, по какой форме этот студент обучается. Другими словами, мы хотим понять, влияет ли форма обучения на качество последнего, обусловливает ли его причинно. Обозначим через X номинальный признак "форма обучения" (принимающий три значения). Заметим, что в действительности признак X может быть получен по шкале более высокого типа, чем номинальная, но в дисперсионном анализе мы его используем как номинальный. Таковым мы и будем его считать.

Опр. Признак X, обозначающий потенциальную причину, называется *фактором* (или *независимым* признаком). Если фактор один, то дисперсионный анализ назвается *однофакторным.*

Через Y , как и выше (и как это принято в литературе), обозначим признак, описывающий главное интересующее социолога явление. Для нас таким признаком является качество обучения. Будем полагать, что он измеряется с помощью ответов респондентов на вопросы некоторого теста

---

[97] Доказательство приведено в работе: Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. М.: Прогресс, 1976. С. 309-311.

[98] Там, помимо описания самого корреляционного отношения, осуществляется его сравнение с коэффициентом корреляции; анализируется возможность его использования для прогноза; предлагаются интересные задачи. На с. 140-141 называются два англоязычных учебника, где более глубоко анализируются вопросы, связанные со статистическим предсказанием и связи внутригрупповой корреляции с дисперсионным анализом (последнему посвящены темы 14 и 15):

(мы сейчас отвлекаемся от того, как именно “устроен” этот тест). В дисперсионном анализе требуется, чтобы уровень измерения Y был не ниже интервального (этого требует вычисление средних арифметических значений и дисперсий признака Y)..

Опр. Признак Y, обозначающий потенциальное следствие, называется *зависимым.*

Для решения поставленной задачи проводим эксперимент. Разделяем всех студентов на три группы и, скажем, в течение года используем свою форму обучения для каждой. Через год тестируем всех студентов. По результатам этого тестирования мы и должны понять, обусловливает ли форма обучения качество знания.

Организуем информацию по тому же принципу, какой был использован в параграфе, посвященном корреляционному отношению.

Группы (или, как говорят в дисперсионном анализе – *ячейки*) определяются значениями независимого номинального признака (здесь – формой обучения), или фактора X. В ячейку помещаются значения основного интересующего нас признака Y (в нашем случае – теста, используемого для оценки качества обучения), вычисленные для попавших в эту ячейку респондентов.

Пусть наша таблица примет вид 14.1.

Таблица 14.1. Гипотетический пример данных для однофакторного дисперсионного анализа.

Уровень фактора X		
1	2	3
$Y_{11}=2$, $Y_{21}=3$, $Y_{31}=4$	$Y_{12}= 1$, $Y_{22}=9$	$Y_{13}= 2$, $Y_{23}= 4$, $Y_{33}=4$, $Y_{43}= 3$
$n_1=3$	$n_2=2$	$n_3=4$
$\overline{Y}_{\bullet 1} = 3$	$\overline{Y}_{\bullet 2} = 5$	$\overline{Y}_{\bullet 3} = 3{,}3$

(X – форма обучения; Y – результаты тестирования)

Итак, наша основная задача состоит в том, чтобы на основе данных этой таблицы выявить, зависят ли значения признака Y от значений фактора X (т.е. зависит ли качество обучения от формы последнего). При решении подобных задач искомая зависимость часто ассоциируется с причинно-следственной. И мы будем иногда говорить о том, что фактор обусловливает Y, влияет на него, служит причиной того, что Y принял то или иное значение. Однако в этой связи следует напомнить, что дисперсионный анализ, как и любой другой математико-статистический метод, в принципе не может доказать наличие причинно-следственных отношений между рассматриваемыми признаками. И когда мы ниже будем говорить о том, что фактор детерминирует, причинно обусловливает рассматриваемый признак, следует иметь в виду, что подобные выражения могут быть приняты лишь условно.

Подчеркнем, что на вопрос о существовании зависимости уровня Y от значений фактора, невозможно ответить без использования разработанной в математической статистике логики переноса результатов с выборки на генеральную совокупность, точнее, без проверки статистической гипотезы. «На глаз» сравнить совокупности чисел, стоящих в разных ячейках, невозможно. Можно обратиться к средним, но то, что, скажем, среднее первой группы больше средней второй, будет говорить не вообще о влиянии фактора на уровень среднего (в нашем случае – о том, что первая форма обучения более эффективна, чем вторая), а о том, что указанное соотношение между средними верно только для данной конкретной выборки и, может быть, обусловлено лишь какими-то случайными обстоятельствами, а не влиянием фактора. Однако прежде, чем перейти к четкой формулировке проверяемой статистической гипотезы, вспомним еще один факт.

14.2. Модель однофакторного дисперсионного анализа

Прежде, чем говорить о строгом способе решения поставленной задачи, необходимо четко описать ту модель, которая лежит в основе всех формальных построений. Модель кажется очень простой. Но далее, перейдя к рассмотрению двухфакторного дисперсионного анализа, мы увидим, что подобные модели могут быть гораздо более сложными и неочевидными. Научиться же строить подобные модели надо. Причин тому, по крайней мере, две.

Во-первых, подобного рода модели используются очень часто: в регрессионном, логлинейном и других видах анализа данных. И их смысл надо хорошо понять, чтобы иметь возможность читать соответствующую литературу.

Во-вторых, хорошее понимание смысла подобных моделей, на наш взгляд, может способствовать усвоению очень актуального методологического принципа - прежде, чем собирать данные, необходимо сформировать систему «аксиом», четко обрисовывающих априорное представление социолога о том, что он изучает. При всей своей очевидности это положение на практике часто не выполняется. В результате используются анкеты, в которые включены вопросы, не имеющие отношения к делу, не включены необходимые вопросы и т.д.

Все методы анализа данных опираются на такие априорные аксиомы. Они и составляют суть модели, заложенной в том или ином методе. Привычка пользоваться методами, как нам представляется, должна способствовать выработке состоятельного с методологической точки зрения подхода к планированию социологического исследования. Или, коротко говоря, – математика может научить социолога методологической грамотности.

В данном случае имеется в виду одна «аксиома» – четкая формулировка того, отчего зависит основной интересующий исследователя признак Y. Предположение о том, из чего для каждого конкретного респондента формируется уровень признака Y, выглядит следующим образом:

$$Y_{ij} = \mu + \lambda_j^X + \varepsilon_{ij},$$

где μ - некий средний уровень, на фоне которого изучается действие фактора X на признак Y; λ_j^X - это вклад в формирование значения зависимого признака j - го уровня фактора X ; ε_{ij} - добавка, отражающая специфические характеристики именно того (неповторимого) респондента, который включен в j-ю ячейку и имеет в ней номер i.

Сделаем несколько замечаний по поводу отдельных членов модели.

Сущность μ поясним на примере: наверное, этот уровень будет один для выпускников московских школ, другой – для выпускников деревенских школ Иркутской области и третий – для молодежи Центрально-африканской республики. Модель действительна на генеральной совокупности. Выборочной оценкой среднего уровня μ служит $\overline{Y}_{\bullet\bullet}$.

По поводу λ_j^X можно сказать, что это – главный интересующий нас элемент, центральное звено модели. Размер именно этой величины и говорит о том, насколько j-я форма обучения детерминирует уровень полученных студентом знаний. Выборочной оценкой этого показателя служит разность $(\overline{Y}_{\bullet j} - \overline{Y}_{\bullet\bullet})$.

ε_{ij} - это некоторый поправочный элемент, корректирующий основную часть модели, выраженную суммой ($\mu + \lambda_j^X$). Коррекция требуется потому, что, как бы хорошо ни моделировалась величина Y_{ij} этой суммой (а мы ищем именно такой фактор, чтобы равенство Y_{ij} = ($\mu + \lambda_j^X$) по возможности отражало реальность), последняя никогда не сможет учесть все специфические особенности отдельных людей. Точным равенство Y_{ij} = ($\mu + \lambda_j^X$) не будет никогда. Но мы все же будем считать его приемлемым, если величины ε_{ij} в среднем равны нулю, независимы и как бы погашают друг друга. К примеру, если у i-го респондента из j-й группы ε_{ij} окажется большим по абсолютной величине и отрицательным (из-за того, что этот респондент, к примеру, много болел в течение того года, когда проводился эксперимент; это привело к тому, что успехи студента оказались ниже, чем следовало бы ожидать при рассматриваемом среднем уровне и воздействии соответствующей формы обучения), то в той же группе найдется такой k –й респондент, у которого ε_{kj} будет примерно таким же по модулю, но положительным (из-за того, скажем, что этот респондент весь год усиленно занимался дополнительно с преподавателем). Обычно предполагается, что величины ε_{ij} имеют нормальное распределение с нулевым математическим ожиданием.

Если же окажется, что ε_{ij} довольно велики и практически у всех респондентов положительны, - значит, мы не учли действие какого-то общего для всех респондентов фактора. И нам надо или заменить фактор X другим, или перейти к двухфакторному дисперсионному анализу. Величины ε_{ij} называются остатками, и математическая статистика обычно представляет исследователю средства проанализироваать эти остатки (это имеет место не только для дисперсионного анализа).

Выборочной оценкой величины ε_{ij} служит разность $Y_{ij} - \overline{Y}_{\bullet j}$.

Все три элемента модели могут быть расценены как вклады в вариацию признака Y, как источники такой вариации.

14.3. Однофакторный дисперсионный анализ как проверка статистической гипотезы

Взглянем на дисперсионный анализ с другой стороны.

Будем проверять гипотезу о равенстве средних для рассматриваемых групп. В данном случае:

$$H_0: \mu_1 = \mu_2 = \mu_3$$

(точнее, следовало бы писать: $H_0: \mu_{\bullet 1} = \mu_{\bullet 2} = \mu_{\bullet 3}$)

Заметим, что H_1 здесь формулируется достаточно неопределенно –

H_1: не все средние равны.

Для того, чтобы выяснить, какие именно из средних можно считать неравными, нало использовать специальную технику – методы множественного сравнения. Об этом мы пока не говорим. Вернемся к этому при обсуждении следующей темы.

Проверка гипотезы осуществляется знакомым нам образом. Чтобы реализовать соответствующую логику, надо знать критериальную статистику и закон ее распределения.

Используем введенные выше обозначения:

$SS_b = SS_{между}$ - межгрупповая сумма квадратов; $SS_w = SS_{внутри}$ – внутригрупповая сумма квадратов; $SS_t = SS_{общ}$ – общая сумма квадратов; n – объем выборки, J – число ячеек.

Каждой сумме квадратов отвечает свое число степеней свободы:

$$df_b = J-1;\ df_w = Jn - J;\ df_t = Jn - 1.$$

Заметим, что

$$df_b + df_w = df_t.$$

Введем еще два обозначения:

$MS_b = SS_b / df_b$; $MS_w = SS_w / df_w$. Это так называемые средние квадраты. Искомая статистика имеет вид:

$$F_{df_b, df_w} = \frac{MS_b}{MS_w}.$$

Чтобы пояснить тот содержательный смысл, который заложен в этом критерии, вспомним, как выглядит кривая F – распределения. Напомним, что эта кривая имеет два «хвоста» и, соответственно, могут быть найдены два табличных значения:

$$F_{табл} и \frac{1}{F_{табл}}.$$

Как правило, рассматриваемый критерий имеет смысл считать двусторонним (логика рассуждений, использующихся при выборе числа «сторон» критерия – та же, которая была использована нами при обсуждении аналогичного вопроса в случае проверки гипотезы о равенстве двух средних).

Для определенности (и в соответствии с традицией), положим, что первое значение ограничивает правый «хвост», а второе – левый. Гипотеза будет отвергнута в двух случаях.

Во-первых, если значение критерия достаточно велико. А это будет иметь место, когда числитель дроби MS_b / MS_w велик, а знаменатель мал. Другими словами, критерий «зашкалит» за правое табличное значение, если наши средние далеко отстоят друг от друга, а внутри каждой группы имеется однородность (т.е. каждое среднее действительно репрезентирует соответствующую группу). Хотелось бы, чтобы читатель понял, что это вполне отвечает здравому смыслу.

Заметим, что аналогичные критерии используются во многих алгоритмах классификации. Мы имеем в виду критерии, позволяющие судить о качестве разбиения. Эти критерии говорят о хорошем качестве, если внутри классов объектам «тесно», а сами классы расположены «просторно», между ними большие расстояния. Таким образом, можно сказать, что дисперсионный анализ не только выводит нас на причинно-следственные отношения, но и позволяет оценить качество классификации, состоящей в распределении объектов по ячейкам.

Во-вторых, гипотеза о равенстве средних будет отвергнута, если значение критерия достаточно мало. Смысл этого труднее понять. Однако и здесь обычные житейские рассуждения

приходят на помощь. Итак, пусть дробь MS_b / MS_w очень мала. Грубо говоря, это означает, что либо средние, вычисленные по отдельным классам, очень близки друг к другу, либо разброс значений внутри классов в среднем очень велик. Нетрудно увидеть, что и в том, и в другом случае нет абсолютно никаких оснований отвергать гипотезу, т.е. полагать, что у нас имеются различные средние, хорошо репрезентирующие свои группы. Этого нельзя сказать ни в том случае, если средние «слиплись» (раз они мало отличаются, то вряд ли можно говорить о том, что уровень Y определяется значением X), ни в том, если средние (даже если они разные) не надежны, не отражают ситуацию в группе. Подчеркнем, что мы *не доказываем,* что средние равны, мы просто полагаем, что выборка не дает нам оснований сомневаться в этом, не дает оснований отвергнуть нуль-гипотезу.

Можно показать, что гипотеза

$$H_0: \ \mu_1 = \mu_2 = \ldots = \mu_J \quad (14.1)$$

эквивалентна гипотезе

$$H_0: \lambda_1^X = \lambda_2^X = \ldots = \lambda_J^X . \quad (14.2)$$

Обычно считаются выполненными условия:

$$n_1\lambda_1^X + n_2\lambda_2^X + \ldots + n_J\lambda_J^X = 0^{99} \quad (14.3).$$

Если же это учесть, то гипотеза (14.1) оказывается эквивалентной гипотезе

$$H_0: \ \lambda_1^X = \lambda_2^X = \ldots = \lambda_J^X = 0$$

и, следовательно, гипотезе

$$H_0: \ \sum_{j=1}^{J} (\lambda_j^X)^2 = 0 .$$

14.4. О понимании термина «влияет» (или что значит доказать наличие причинно-следственного отношения с помощью дисперсионного анализа)

Итак, приняв гипотезу (14.1), мы полагаем, что фактор X не влияет на Y. Другими словами, форма обучения не обусловливает (в причинном смысле) уровень усвоения знаний студентами. Отвергнув же названную гипотезу, мы, напротив, полагаем, что уровень знаний студентов причинно обусловлен тем, по какой системе обучаются эти студенты. Хотелось бы, чтобы читатель понимал содержательную значимость подобных выводов. Термины «причинно обусловлен», «влияет» и т.д. здесь употребляются весьма условно. Точно так же мы говорили бы, если бы, скажем, составили частотную таблицу для наших двух признаков (предварительно, конечно, разбив диапазон изменения признака Y на интервалы и начав рассматривать этот признак как номинальный) и рассчитали, к примеру, критерий "Хи-квадрат". Ответ на вопрос о наличии (отсутствии) связи между признаками мы тоже могли бы интерпретировать как наличие, либо отсутствие соответствующих причинно-следственных отношений. И вывод этот совсем не обязательно совпал бы с выводом, сделанным на основе дисперсионного анализа.

Таким образом, исследователь должен четко понимать, что каждый математический метод дает нам лишь некоторый срез с того явления реальности, который мы назвали причинно-следственным отношением.

---

[99] Пусть условие (14.3) не будет выполнено. Например, пусть в этом равенстве справа вместо 0 стоит некоторая величина $\Delta \neq 0$. Тогда увеличим гипотетический средний уровень на величину $\frac{\Delta}{n}$ и начнем «отсчитывать» величины вкладов λ_j^X уровней фактора X как бы от нового среднего уровня: заменим каждый вклад на $(\lambda_j^X - \frac{\Delta}{n})$. Мы ничего не приобретем и не потеряем в смысле получения содержательного знания (допущение возможности сдвига всех вкладов говорит о том, что мы считаем, что вклады измерены по шкале разностей - это шкала более высокого типа, чем интервальная), но требование (101) окажется выполненным: $\sum_{j=1}^{J} n_j(\lambda_j^X - \frac{\Delta}{n}) = \sum n_j\lambda_j^X - \frac{\Delta\sum_{j=1}^{J} n_j}{n} = \Delta - \Delta = 0$).

14.5. Метод множественных сравнений для однофакторного дисперсионного анализа.

Метод множественных сравнений рассмотрим только для однофакторного дисперсионного анализа.

Предположим, что, применив однофакторный дисперсионный анализ, мы обнаружили, что проверяемая гипотеза (напомним, что это – гипотеза о равенстве средних всех рассматриваемых ячеек) должна быть отвергнута. Это означает отрицание выражения «средние всех ячеек равны», т.е. утверждение того, что среди средних имеются хотя бы два неравных. Естественно, что подобное утверждение малоинформативно для исследователя. Возникает ряд вопросов: какие именно средние не равны, в каком смысле не равны: может быть, первое – в пять раз больше второго? или же третье равно среднему арифметическому двух первых? и т.д. Найти некоторый (хотя снова не достаточно полный) ответ на эти вопросы и помогает найти метод множественных сравнений. Из двух известных подходов рассмотрим один – т.н. S-метод, связываемый обычно с именем Шеффе[100].

Метод позволяет проверить справедливость некоторой заранее заданной зависимости между генеральными математическими ожиданиями рассматриваемых ячеек. Зависимость эта выражается в определенном виде. Проверка ее справедливости снова происходит на статистическом языке: речь идет о проверке математико-статистической гипотезы о наличии определенного рода связей между изучаемыми средними. Чтобы описать, какого рода связь между средними мы имеем возможность проверить, введем новое определение.

Пусть μ_1 , μ_2, ... , μ_J – рассматриваемые групповые средние, т.е. те самые средние ячеек, гипотезу о равенстве которых мы отвергли в результате применения однофакторного дисперсионного анализа.

Опр. Назовем *контрастом* средних μ_1 , μ_2, ... , μ_J выражение вида

$$\psi = c_1\mu_1 + c_2\mu_2 + \ldots + c_J\mu_J,$$

где c_1, c_2, ... c_J - произвольные действительные числа, удовлетворяющие условию:

$$c_1 + c_2 + \ldots + c_J = 0.$$

Смысл введения понятия контраста станет ясным, если мы скажем, что, оказывается, математическая статистика представляет нам средства для проверки гипотез вида:

$$H_0 : \psi = 0$$

(для любого контраста ψ).

Чтобы привести примеры контрастов, предположим, что у нас имеется четыре средних μ_1, μ_2, μ_3, μ_J. Рассмотрим следующие гипотезы и коэффициенты отвечающих им контрастов:

Проверяемая гипотеза H_0	C_1	C_2	c_3	c_4
$\mu_1 - \mu_2 = 0$	1	- 1	0	0
$\mu_1 - (\mu_2 + \mu_3) / 2 = 0$	1	- (1/2)	- (1/2)	0
$5\mu_1 - 3\mu_2 - 2\mu_3 = 0$	5	-3	-2	0
$(\mu_1 + \mu_2) /2 - (\mu_1 + \mu_2) / 2 = 0$	½	½	- (1/2)	- (1/2)

Если будет принята первая гипотеза, это будет означать, что первое и второе средние равны. Значит, неравенство следует искать в отличии третьего или четвертого среднего от первых двух.

Принятие второй гипотезы означает, что первое среднее можно считать равным среднему арифметическому значению второго и третьего.

Принятие третьей гипотезы заставит нас полагать, что среднее первых двух средних равно среднему последних двух средних.

Примеры задач

1. Для изучения влияния семейного окружения на развитие ребенка были протестированы дети, растущие в разных условиях. Использовался специальный тест, позволяющий оценить уровень развития опрашиваемого (в качестве приписываемых каждому респонденту значений фигурировали целые числа от 0 до 10). Результаты опроса приведены в следующей таблице.

---

[100] См.: Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. М.: Прогресс, 1976. С. 350 – 354.

Дети из детского дома	Дети из неполных семей	Дети из полных семей
4,9,2, 3, 1,1	4, 5, 8, 3	5, 7, 3, 8

Можно ли сказать, что семейное окружение действительно влияет на развитие ребенка?

2. С помощью однофакторного дисперсионного анализа выявлялось, зависит ли успеваемость студента ГУ-ВШЭ от того, какую школу он окончил. Было изучено три группы студентов-первокурсников, окончивших школу, соответственно, при ГУ-ВШЭ (группа 1), другую среднюю школу в Москве (группа 2), среднюю школу на периферии (группа 3). Успеваемость измерялась с помощью некоторого теста. Гипотеза о равенстве средних была отвергнута. Можно ли каким-нибудь образом проверить гипотезу исследователей о том, что спецшкола при ГУ-ВШЭ дает лучшие знания, чем две другие школы, и что периферийная школа дает более слабые знания, чем каждая из московских. Если можно, то сделать это, воспользовавшись следующими статистическими данными:

Группа студентов	Выборочные средние	N	MS_w=8,35
1	12,86	20	
2	10,54	25	
3	7,17	15	

Добавочную литературу см. после главы 15.

ТЕМА 15

Двухфакторный дисперсионный анализ.

15.1. Двухфакторный дисперсионный анализ как метод анализа результатов эксперимента при изучении причинно-следственных отношений

Предположим, что решая рассмотренную выше задачу - выявление зависимости (независимости) успехов студентов от формы обучения, мы пришли к выводу о неполноте такой постановки задачи: к примеру, поняли, что эффективность формы обучения зависит от пола студента. Исходная матрица ячеек становится двумерной. Покажем, как она строится.

Пусть величина Y_{ijk} означает значение главного интересующего нас признака Y, вычисленное для i-го по счету (счет ведется внутри ячейки) объекта, помещенного в ячейку с номером (j, k), где j – отвечающее рассматриваемой ячейке значение первого фактора X^1 , а k – отвечающее той же ячейке значение второго фактора X^2 .

Будем рассматривать упрощенный случай, когда число объектов во всех ячейках одинаково и равно n (в приведенной ниже таблице n = 3). Фактор X^1 - значения 1, 2, …, J (в примере J=2), а фактор X^2 принимает значения 1, 2, …, K (в нашем примере K = 3) .

Уровень фактора X^1	Уровень фактора X^2			Числ-ть Групп, отв-х уровням X^1	Среднее арифметическое значение Y для фиксированного уровня фактора X^1
	1	2	3		
1	Y_{111}=2, Y_{211}=3, Y_{311}=4	Y_{112}= 1 Y_{212}=9 Y_{312}=4	Y_{113}=2 Y_{213}= 4 Y_{313} =4	$n_{1\bullet}$=nI=3 •3=9	$\overline{Y}_{\bullet 1\bullet} = \frac{2+3+4+1+9+4+2+4+4}{9} = 3,7$
2	Y_{121}=5, Y_{221}=6, Y_{321}=4	Y_{122}=3 Y_{222}=4 Y_{322}=4	Y_{123}=8, Y_{223}=9, Y_{323}=7	$n_{2\bullet}$=nI=3 •3=9	$\overline{Y}_{\bullet 2\bullet}$ =5,6
Числ-ть	$n_{\bullet 1}$ =nJ=3•2=6	$n_{\bullet 2}$=nJ=	$n_{\bullet 3}$=nJ=	$n_{\bullet\bullet}$=n=	

Групп, отв-х уровням X^2		=3•2=6	=3•2=6	*18*	
	$\overline{Y}_{\bullet 1} = \frac{2+3+4+5+6+4}{6} = 4$	$\overline{Y}_{\bullet 2} = 4{,}2$	$\overline{Y}_{\bullet 3}$= 5,6		$\overline{Y}_{\bullet\bullet}$= 4,6

6.1. Модель двухфакторного дисперсионного анализа

Общая модель двухфакторного дисперсионного анализа имеет вид:

$$Y_{ijk} = \mu + \lambda_j^1 + \lambda_k^2 + \lambda_{jk}^{12} + \varepsilon_{ijk} \ ,$$

где

- Y_{ijk} - значение зависимого признака для конкретного респондента, помещенного в ячейку, отвечающую j – му уровню первого фактора и k – му уровню второго фактора (скажем, обучающемуся по второй форме юноши, тогда j=2, k=1), имеющему номер i в соответствующей ячейке;

- λ_j^1 - вклад j-го уровня (нижний индекс) первого фактора (верхний индекс) X^1 в формирование значения Y (в нашем примере первый фактор – форма обучения; упомянутый вклад определяется тем, что студент обучался либо по первой форме обучения (j=1), либо по второй (j=2), либо по третьей (j=3));

- λ_k^2 - вклад k-го уровня второго фактора X^2 в формирование значения Y (в нашем примере второй фактор – пол студента; соответствующий вклад определяется тем, является ли студент юношей или девушкой);

- λ_{jk}^{12} - вклад, определяющийся взаимодействием i-го уровня первого фактора и j-го уровня второго;

- $\varepsilon_{i\,j\,k}$ – поправочный коэффициент (остаточный член), говорящий о различии между реальным значением Y-ка для рассматриваемого респондента и тем его значением, которое должно получиться для любого респондента, вошедшего в рассматриваемую ячейку (с номером (j,k)) в соответствии с нашей моделью.

Особое внимание стоит уделить понятию взаимодействия. Именно наличием этого элемента принципиально отличается двухфакторный дисперсионный анализ от однофакторного. Наличие взаимодействия двух признаков (в отношении некоторого зависимого признака Y) говорит о том, что вид связи первого признака с Y зависит от того, какое значение принимает второй признак. К примеру, по отдельности и форма обучения, и пол могут в среднем (статистически) не влиять на качество обучения, но, скажем, вторая форма обучения применительно к девушкам при этом может давать очень высокий положительный эффект. Другими словами, действие формы обучения зависит от того, для девушек или для юношей она используется. Когда в принципе уровень Y-ка зависит от сочетаний конкретных значений X^1 и X^2 , говорят о наличии взаимодействия между рассматриваемыми факторами. Иногда в том же смысле говорят о синергетическом эффекте, вызванном сочетанием значений факторов. В том же смысле ведут речь о нелинейности воздействия факторов на зависимую переменную. Иногда взаимодействием называют само сочетание значений факторов, обусловливающее некий конкретный уровень Y-ка.[101]

[101] Поиск взаимодействий – принципиальная задача для социолога. К такому выводу можно придти, если учесть, что в социологии в качестве результата измерения зачастую имеет смысл считать не найденные каким-то образом значения непрерывных латентных переменных, а сочетания значений номинальных признаков, детерминирующих то или иное поведение респондента. Такие сочетания и называются взаимодействиями. В этом – суть того, что иногда в литературе называется гуманитарным измерением. Подробнее об этом можно прочесть в работе: Толстова Ю.Н. Анализ социологических данных. Методология, дескриптивная статистика, изучение связей номинальных признаков. М.: Научный мир, 2000.

Выборочные оценки всех формирующих модель составляющих вычисляются аналогично тому, как это делалось для однофакторного дисперсионного анализа.

Название элемента модели	Обозначение	Выборочная оценка
Общий средний уровень	μ	$\bar{Y}_{\bullet\bullet\bullet}$
Вклад j-го уровня фактора X^1	λ^1_j	$\bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet\bullet}$
Вклад k-го уровня фактора X^2	λ^2_k	$\bar{Y}_{\bullet\bullet k} - \bar{Y}_{\bullet\bullet\bullet}$
Вклад, определяющийся взаимодействием i-го уровня фактора X^1 и j-го уровня фактора X^2;	λ^{12}_{jk}	$\bar{Y}_{\bullet jk} - \bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet k} + \bar{Y}_{\bullet\bullet\bullet}$
Остаточный член	$\varepsilon_{i\,j\,k}$	$Y_{ijk} - \bar{Y}_{\bullet jk}$

Определенную сложность представляет лишь оценка взаимодействия. Зададимся вопросом о том, каким образом можно оценить наличие (или отсутствие) взаимодействия в конкретной ячейке с номером (j,k). Наличие некоего среднего уровня, совокупное воздействие j-го уровня фактора X^1 и k-го уровня фактора X^2 в отдельности вкупе с взаимодействием указанных уровней приводят к тому, что средний уровень успеваемости для объектов рассматриваемой ячейки оказывается равным $\bar{Y}_{\bullet jk}$. Он состоит из общего среднего уровня $\bar{Y}_{\bullet\bullet\bullet}$, вклада j-го уровня первого фактора ($\bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet\bullet}$), вклада k-го уровня второго фактора ($\bar{Y}_{\bullet\bullet k} - \bar{Y}_{\bullet\bullet\bullet}$) и вклада упомянутого взаимодействия. Как же получить оценку последнего. Естественно, путем «вытаскивания» из среднего уровня, типичного для рассматриваемой ячейки, названных вкладов. Другими словами, выборочная оценка взаимодействия равна

$$\bar{Y}_{\bullet jk} - (\bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet\bullet}) - (\bar{Y}_{\bullet\bullet k} - \bar{Y}_{\bullet\bullet\bullet}) = \bar{Y}_{\bullet jk} - \bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet k} + \bar{Y}_{\bullet\bullet\bullet}. \quad (15.1)$$

Аналогичные рассуждения справедливы и для генеральной совокупности. Этим мы воспользуемся в следующем параграфе при обсуждении вида гипотез, проверяемых в двухфакторном дисперсионном анализе.

15.3. Двухфакторный дисперсионный анализ как проверка статистических гипотез

В двухфакторном дисперсионном анализе проверяются три статистические гипотезы. Первые две аналогичны гипотезе однофакторного анализа, это гипотезы:

$$H_0: \mu_{\bullet 1\bullet} = \mu_{\bullet 2\bullet} = \ldots = \mu_{\bullet J\bullet} \quad (15.2)$$

$$H_0: \mu_{\bullet\bullet 1} = \mu_{\bullet\bullet 2} = \ldots = \mu_{\bullet\bullet K}. \quad (15.3)$$

При проверке первой гипотезы мы как бы забываем о втором факторе и рассматриваем J ячеек, на которые наша совокупность делится в соответствии со значениями первого фактора. При проверке второй гипотезы аналогичные рассуждения используем применительно ко второму фактору. Использующиеся критерии очень похожи на те, которые фигурируют в однофакторном дисперсионном анализе (отличие состоит в виде внутригрупповой дисперсии, что станет ясно из приведенных ниже формул).

Третья гипотеза говорит об отсутствии взаимодействий, или, что то же самое, о равенстве всех взаимодействий нулю. В соответствии с формулой (15.1), речь идет о проверке равенства нулю выражений вида

$$\mu_{\bullet jk} - \mu_{\bullet j\bullet} - \mu_{\bullet\bullet k} + \mu_{\bullet\bullet\bullet}.$$

Известно, что можно говорить о равенстве нулю всех чисел из некоторой совокупности, если сумма квадратов этих чисел равна нулю. Поэтому третья проверяемая в двухфакторном дисперсионном анализе гипотеза имеет вид:

$$H_0: \sum_{j,k} (\mu_{\bullet jk} - \mu_{\bullet j\bullet} - \mu_{\bullet\bullet k} + \mu_{\bullet\bullet\bullet})^2 = 0. \quad (15.4)$$

Перейдем к обсуждению вида критериев, использующихся для проверки указанных трех гипотез.

Вместо межгрупповой суммы квадратов, обозначенной нами выше $SS_b = SS_{между}$, введем в рассмотрение три аналогичные суммы: SS_1, SS_2, SS_{12}, отвечающие, соответственно, гипотезам (15.2), (15.3), (15.4).

$$SS_1 = nK \sum_{j=1}^{J} (\bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet\bullet})^2, \quad SS_2 = nJ \sum_{k=1}^{K} (\bar{Y}_{\bullet\bullet k} - \bar{Y}_{\bullet\bullet\bullet})^2, \quad SS_{12} = n \sum_{j=1}^{J} \sum_{k=1}^{K} (\bar{Y}_{\bullet jk} - \bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet k} + \bar{Y}_{\bullet\bullet\bullet})^2$$

Внутригрупповая сумма квадратов $SS_w = SS_{внутри}$ вычисляется одинаковым образом для всех трех гипотез:

$$SS_w = \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{i-1}^{n} (\bar{Y}_{IJk} - \bar{Y}_{\bullet JK})^2$$

Каждой сумме квадратов отвечает свое число степеней свободы:

$$df_1 = J-1; \quad df_2 = K-1; \quad df_{12} = (J-1)(K-1); \quad df_w = JKn - JK;$$

Введем обозначения средних квадратов:

$MS_1 = SS_1 / df_1$; $MS_2 = SS_2 / df_2$; $MS_{12} = SS_{12} / df_{12}$; $MS_w = SS_w / df_w$. Искомые статистики имеют вид:

$$F_{df_1, df_w} = \frac{MS_1}{MS_w}; \quad F_{df_2, df_w} = \frac{MS_2}{MS_w}; \quad F_{df_{12}, df_w} = \frac{MS_{12}}{MS_w}.$$

Логика проверки соответствующих гипотез – та же, к которой читатель, как мы надеемся, уже привык.

В заключение обсуждения вопроса о дисперсионном анализа отметим, что существуют такие его варианты, которые рассчитаны на порядковые данные (непараметрический дисперсионный анализ).

Примеры задач

1. Исследователи решили выяснить, в какой мере крепость семьи связана с тем, одинаковы или нет национальности супругов, и каково количество детей в семье. Для ряда семей по определенной методике был рассчитан коэффициент прочности семьи (число от 1 до 10). Наблюдаемые данные были сведены в следующую таблицу (в клетках – коэффициенты прочности обследованных семей).

Сходство национальностей супругов	Количество детей в семье		
	0	1	Больше одного
Национальность мужа и жены одинаковы	3,4,4,1	5,2,6,3	10,8,4,7
Национальности мужа и жены разные	3,4,4,4	6,7,2,4	7,7,7,7

Что можно сказать о влиянии названных факторов на крепость семьи?

Добавочная литература к главам 14 и 15[102]

Основная

Гмурман В.Е. Теория вероятностей и математическая статистика. М.: Высшая школа, 1998. С.349-362 (однофакторный дисперсионный анализ)

Горелова Г.В., Кацко И.А. Теория вероятностей и математическая статистика в примерах и задачах с применением Excel. Ростов-на-Дону: Феникс, 2005. С.207-239 (однофакторный, двухфакторный, трехфакторный дисперсионный анализ)

Калинина В.Н., Панкин В.Ф. Математическая статистика. М.: Высшая школа, 1998. С. 244-266 (однофакторный и двухфакторный дисперсионный анализ)

Колемаев В.А., Калинина В.Н. Теория вероятностей и математическая статистика, М.: Инфра-М,1997.С. 184-191; Юнити, 2003 (однофакторный и двухфакторный дисперсионный анализ)

Кремер Н.Ш. Теория вероятностей и математическая статистика. М.: ЮНИТИ-ДАНА, 2001. С.375-391 (однофакторный и двухфакторный дисперсионный анализ)

Дополнительная

Гусев А.Н. Дисперсионный анализ в экспериментальной психологии: Учебное пособие для студентов факультетов психологии ВУЗов по направлению 512000 – «Психология», М.: Учебно-методический коллектор «Психология», 2000

Крыштановский А.О. Анализ социологических данных с помощью пакета SPSS.М.:Издательский дом ГУ-ВШЭ, 2005. С. 109-114 (непараметрический дисперсионный анализ Краскэла – Уоллиса)

Сидоренко Е. Методы математической обработки в психологии. СПб: Речь, 2000. С. 224-260 (однофакторный и двух факторный дисперсионный анализ)

Статистические методы анализа социологической информации. М.: Наука, 1979. Гл. 11.

Girden E.R. ANOVA: Repeated measures // Sage University Paper series on Quantitative applications in the social sciences; Beverly Hills: SAGE Publications. V. 84.

См

1. *Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. Учебник для вузов. – М.: ЮНИТИ, 1998. С. 282-319.*
2. *Крыштановский А.О. Анализ социологических данных с помощью пакета SPSS. Издательский дом ГУ-ВШЭ. Москва, 2006.*
3. *Эддоус М., Стэнсфилд Р. Методы принятия решений/ Пер. с англ. Под ред. Член-корр. РАН И.И.Елисеевой. – М.: Аудит, ЮНИТИ, 1997.*
4. *Руководство пользователя SPSS 11.0.*
5. *Бююль А., Цефель П. SPSS: искусство обработки информации, анализ статистических данных и восстановление скрытых закономерностей. DiaSoft, 2002.*

[102] Дисперсионный анализ описывается отнюдь не во всех книгах по теории вероятностей и математической статистике. Поэтому в данном пункте мы выписываем некоторые наименования из списка Приложения 1 с указанием страниц, посвященных дисперсионному анализу.

Приложение 1
Литература

Основная

Андропов А.М., Копытов Е.А., Гринглаз Л.Я. Теория вероятностей и математическая статистика. Учебник для вузов. С.-Пб: Питер, 2004

Бородин А.Н. Элементарный курс теории вероятностей и математической статистики. С.-Пб: Лань, 2004

Гмурман В.Е. Теория вероятностей и математическая статистика. М.: Высшая школа, 1998

Гмурман В.Е. Руководство к решению задач по теории вероятностей математической статистике. М.: Высшая школа, 1998

Калинина В.Н., Панкин В.Ф. Математическая статистика. М.: Высшая школа, 1998

Колемаев В.А., Калинина В.Н. Теория вероятностей и математическая статистика, М.: Инфра-М,1997; Юнити, 2003

Кочетков Е.С., Смерчинская С.О., Соколов В.В. Теория вероятностей и математическая статистика. М.: Изд. дом «Форум», 2003

Кремер Н.Ш. Теория вероятностей и математическая статистика. М.: ЮНИТИ-ДАНА, 2001

Прикладная статистика. Основы эконометрики. Т.1: Айвазян С.А., Мхитарян В.С. Теория вероятностей и прикладная статистика. М.: Юнити-Дана, 2001

Теория статистики с основами теории вероятностей / Под ред. И.И.Елисеевой. М.: Юнити, 2001

Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. М.: ИНФРА-М, 2003

Шведов А.С. Теория вероятностей и математическая статистика. М.: Изд. дом ГУ-ВШЭ, 2005

Эддоус М., Стэнсфилд Р. Методы принятия решения. М.: Финансы и статистика, 1997

Bluman A.G. Elementary statistics. A step by step. McGraw-Hill Companies. 1992, 1995, 1998, 2001

Дополнительная

Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. М.: Прогресс, 1976

Горелова Г.В., Кацко И.А. Теория вероятностей и математическая статистика в примерах и задачах с применением Excel. Ростов-на-Дону: Феникс, 2005

Интерпретация и анализ данных в социологических исследованиях. М.: Наука, 1987

Паниотто В.И. Количественные методы в социологических исследованиях. Киев: Наукова думка, 1982

Рабочая книга социолога. М.: Наука, 1983

Статистические методы анализа социологической информации. М.: Наука, 1989

Татарова Г.Г. Методология анализа данных в социологии. М., 1998

Толстова Ю.Н. Анализ социологических данных: методология, дескриптивная статистика, анализ связей номинальных признаков. М.: Научный мир, 2000

Шеффе Г. Дисперсионный анализ. М.: ГИФМЛ, 1963

Hinton P.R. Statistics Explained. A Guide for social Science Students. - N.-J.,L.: Routledge, 1995

Kachigan S.K. Statistical analysis. An interdisciplinar introduction to univariate and multivariate methods. - N.Y.: Radius Press,1986

Sirkin R.M. Statistical for the social sciences. Sage publ.,1995

Walsh A. Statistical for the social sciences: with computer - based applications. - N.Y.: Harper & Row, Publishers, 1990.

СПРАВОЧНИКИ, ЭНЦИКЛОПЕДИИ

Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. М., 1983

Вероятность и математическая статистика. Энциклопедия. М.: БРЭ, 1999

Корн Г., Корн Т. Справочник по математике для научных работников и инженеров. М., 1978

Математическая энциклопедия, в 5-ти томах. М.: БСЭ, 1977-1985

Поллард Дж. Справочник по вычислительным методам статистики. М.: Финансы и статистика, 1982

Прохоров Ю.В., Розанов Ю.А. Теория вероятностей. Основные понятия. Предельные теоремы. Случайные процессы (справочник). М., 1967

Рунион Р. Справочник по непараметрической статистике. Современный подход. М.: Финансы и статистика, 1982

Социологическая энциклопедия, в 2-х томах. М.: Мысль, 2003

Справочник по прикладной статистике (под ред. Э.Ллойда, У.Ледермана). В 2-х томах. М.: Финансы и статистика, 1989, 1990

Справочник по теории вероятностей и математической статистике. Киев, 1978

Хастингс Н., Пикок Дж. Справочник по статистическим распределениям. М.: Статистика, 1980

Энциклопедический социологический словарь. М., 1996

Handbook of survey research (Quantitative studies in social relations) (ed.by P.H.Rossi, J.D.Wright, A.B.Anderson). Academic Press, inc. LTD, 1983[103]

Приложение 2
ПРИМЕРНЫЕ ЭКЗАМЕНАЦИОННЫЕ ВОПРОСЫ

[103] В книге имеется следующее посвящение: “To the Memory of Paul.F.Lazarsfeld, Samuel A.Stouffer, and Angus Campbell, innovative pioneers in the development of social science applications of sample survey”.
Представляется, что наши читатели должны знать эти фамилии.

Вид случайных событий, изучаемых социологом. Понятие случайной величины.

Функция распределения и функция плотности распределения. Их связь друг с другом. Функция Лапласа

Понятия выборки и генеральной совокупности. Способы переноса результатов с первой на вторую

Представление о параметрах распределения и статистиках. Примеры

Основные параметры одномерных распределений – математическое ожидание, квантили, мода, медиана, среднеквадратическое отклонение, дисперсия

Основной параметр двумерного распределения – коэффициент корреляции. Его свойства, вид отражаемой им связи, его недостатки

Нормальное, равномерное распределения. Их основные параметры. Примеры важных для социолога случайных величин, имеющих названные распределения

Стандартизированные случайные величины. Использование вероятностных таблиц нормального распределения.

Распределения, основанные на нормальном: Хи-квадрат, Стьюдента, Фишера: аналитический вид случайной величины, графическое представление, расчет числа степеней свободы, формула для математического ожидания и дисперсии. Использование вероятностных таблиц для этих распределений

Центральная предельная теорема, ее значение для социолога

Закон больших чисел, его научное и практическое значение

Определение номинальной, порядковой, интервальной шкалы. Понятие допустимого преобразования шкалы. Общее представление об адекватности метода типу шкал

Обоснование адекватности (неадекватности) среднего арифметического для номинальной, порядковой, интервальной шкалы

Для какого типа шкал коэффициент корреляции является всегда адекватным и почему?

Для какого типа шкал медиана является всегда адекватной и почему?

Среднее арифметическое и дисперсия для дихотомической шкалы

Дискретные и непрерывные признаки. Выборочные представления генеральных распределений. Частотные таблицы, полигоны, гистограммы, кумулята. Проблемы, возникающие при их построении. Гистограммы с неравными интервалами

Способы нахождения моды и медианы для выборки и выборочного частотного распределения. Модели, используемые при расчете медианы

Формула для расчета коэффициента корреляции для выборки и выборочного частотного распределения.

Двумерные частотные распределения. Маргинальные частоты. Условные и безусловные распределения

Зависимые и независимые случайные величины. Теоретические частоты и формулы для них. Вид частотной таблицы для независимых случайных величин

Несколько определений независимости двух случайных величин. Доказательство их эквивалентности

Основные задачи математической статистики

Общее представление о точечном и интервальном оценивании параметров. Примеры

Свойства точечных оценок параметров (эффективность, несмещенность, состоятельность). Их определение и содержательный смысл

Доверительный интервал для математического ожидания. Принципы его построения. Связь с центральной предельной теоремой

Средняя ошибка выборки. Определение объема выборки на ее основе. Плюсы и минусы такого подхода

Логика проверки статистической гипотезы. Уровень значимости. Принципы его определения

Нулевая и альтернативная гипотезы. Примеры

Направленные и ненаправленные альтернативные гипотезы. Односторонний и двусторонний критерий проверки нуль-гипотезы. Примеры

Ошибки первого и второго рода. Мощность критерия

Проверка гипотезы об отсутствии связи между двумя номинальными переменными

Проверка гипотезы о равномерности распределения

Общая логика проверки гипотез о равенстве двух средних

Проверка гипотезы о равенстве двух средних для зависимых выборок

Проверка гипотезы о равенстве двух средних для независимых выборок

Проверка гипотезы о равенстве двух дисперсий для независимых выборок

Проверка гипотезы о равенстве двух долей для независимых выборок

Проверка гипотезы об отличии от нуля коэффициента корреляции

Корреляционное отношение. Его смысл, вид отражаемой им связи.

Межгрупповая, внутригрупповая, общая дисперсии при расчете корреляционного отношения . Соотношение между ними (с доказательством)

Модель, заложенная в однофакторном дисперсионном анализе. Выборочные оценки ее параметров. Смысл решаемых с помощью однофакторного дисперсионного анализа задач

Проверка гипотез в однофакторном дисперсионном анализе

Понятие взаимодействия и роль его изучения в социологии. Соотнесение логики изучения взаимодействий в двухфакторном дисперсионном анализе с логикой поиска взаимодействий в алгоритмах типа AID

Модель, заложенная в двухфакторном дисперсионном анализе. Выборочные оценки ее параметров

Проверка гипотез в двухфакторном дисперсионном анализе

Общее представление об эксперименте в социологии

Внутренняя и внешняя валидность эксперимента

Логические схемы экспериментальной проверки гипотез по Миллю

Методический эксперимент в социологии. Роль математической статистики при его проведении

Приложение 3

Ориентировочные темы эссе (рефератов)

Оглавление

Общие требования к написанию эссе (реферата)

Раздел 1. Изучение начала процесса взаимопроникновения математико-статистических и социологических идей.

Раздел 2. Роль статистических идей и границы их применимости в социологии

Раздел 3. Методы русской земской статистики

Раздел 4. Выборочный метод в социологии

Раздел 5. Социология и математика: проблема «взаимодействия»

Общие требования к написанию эссе (реферата)

1. Желательно написание эссе. Но положительную оценку можно получить и за реферат. Точную границу между этими двумя жанрами провести трудно. Эссе отличается от реферата большей выраженностью собственной позиции и оценивается более высокой оценкой.
2. Примерные темы эссе (реферата) даны ниже. Там же предлагается список ориентировочной литературы. Студент может выбрать по своему желанию (и согласованию с преподавателем) и другую тему, лежащую в русле рассматриваемой проблематики. В частности, может использовать тематика, отраженная в приведенной выше добавочной литературе к отдельным главам.
3. Содержание работы не может быть сведено к тому или иному фрагменту обязательной программы (по любому предмету).
4. Студент не должен ограничиваться использованием только тех работ (даже если это – работы из предлагаемого списка), которые являются учебниками, использующимися как при изучении рассматриваемой дисциплины, так и других дисциплин, включенных в учебную программу.
5. Как правило, требуется сравнение нескольких точек зрения и, соответственно, рассмотрение работ нескольких авторов (использование лишь одной работы должно оговариваться в общении с преподавателем заранее). Это требование должно быть выполнено и для эссе, и для реферата. Творческое осмысление, высказывание собственных мыслей в любом случае будет приветствоваться. Об отличии эссе от реферата говорилось в п.1.
6. Примерный объем – 10-15 страниц. Но надо иметь в виду, что главное – не объем, а содержание: «словам должно быть тесно, а мыслям – просторно».
7. В конце работы должен быть приведен список использованной литературы (в алфавитном порядке). В нем нежелательно присутствие учебников (хотя в прилагаемом ниже общем списке они на всякий случай приводятся). По тексту должны даваться ссылки (номер работы в прямых скобках, номер отвечает приведенному в конце работы библиографическому списку) с указанием той страницы, откуда берется цитата.
8. Использование интернета вполне возможно. Интернет желателен, если с его помощью студент обращается к первоисточникам. Использование же кем-то сделанных обзоров допустимо только в том случае, если первоисточник недоступен. Указание используемого сайта обязательно. Некритическое скачивание чужих текстов недопустимо.
9. Обязательно наличие оглавления.
10. Работа в обязательном порядке защищается автором в процессе непосредственного общения с преподавателем.

Раздел 1. Изучение начала процесса взаимопроникновения математико-статистических и социологических идей.

1. Русские ученые XIX - начала XX веков о роли статистических методов в социологии
2. Статистические методы в русской эмпирической социологии XIX - начала XX веков
3. Лаплас о роли статистики в изучении общества
4. Кондорсе о роли статистики в изучении общества
5. Кетле как родоначальник эмпирической социологии (базирующейся на идеях теории вероятностей)

6. Критика взглядов Кетле на законы развития общества и методы их изучения
7. Кетле как организатор науки (роль Кетле в институциализации статистики как науки)
8. Статистические методы в творчестве Кетле
9. Политическая арифметика и ее связь с социологией
10. Моральная статистика и ее место в социологии
11. Политическая арифметика и государствоведение: начало качественно-количественной дискуссии

Раздел 2. Роль статистических идей и границы их применимости в социологии

1. Роль статистических данных в творчестве великих социологов (Маркса, Вебера, Дюркгейма)
2. Статистическая связь и изучение причинно-следственных отношений
3. Детерминизм и статистичность в социологии
4. Границы применимости вероятностных критериев в социологии
5. Суть статистической парадигмы в социологии
6. Роль поиска статистических закономерностей при изучении социальных явлений
7. Роль закона больших чисел при изучении социальных явлений
8. Рождение русской эмпирической социологии. Роль идей теории вероятностей

Литература к разделам 1 и 2.

Алексеев В.Г. Математика как основание критики научно-философского мировоззрения. Юрьев: Тип. К.Маттисена. 50 с.

Алимов Ю.И. Альтернатива методу математической статистики. М., 1979

Андроновы А. и Е. Лаплас. М., 1930

Арсеньев К.И. Начертание статистики Российского государства, СПб, 1818-1819

Арсеньев К.И. Статистические очерки России. СПб, 1848

Архивная справка Российского государственного исторического архива, составленная заведующим отделом, к.и.н.Раскиным Д.И., подлинные архивные материалы **(студ.233 гр. Юдина Маргарита)**

Беляева Л.А. Эмпирическая социология в России и Восточной Европе. М.: Изд.дом ГУ-ВШЭ,2004

Бернулли Я. О законе больших чисел / Под общей ред Ю.В.Прохорова. М.: Наука, 1986

Бернштейн С.Н. Современное состояние теории вероятностей. М.-Л., 1933. Часть книги опубликована в: Вероятность и математическая статистика. Энциклопедия. М.: БРЭ, 1999. С. 870--871

Буняковский В.Я. Основания математической теории вероятностей. СПб, 1846. Предисловие и Подробное содержание книги опубликовано также в: Вероятность и математическая статистика. Энциклопедия. М.: БРЭ, 1999. С. 863—869

Воронцов-Вельяминов Б.А. Лаплас, 1985

Гернет М.Н. Моральная статистика. М.: ЦСУ, 1922

Гозулов А.И. Очерки истории отечественной статистики. М., 1972

Кондратьев Н. Д. Основные проблемы экономической статики и динамики. М. , 1991. С. 453-523.

Дубаева Е.Е. Взгляды А.Кетле на роль статистических методов в изучении общества // Социология: 4М, 2001, №13. С. 145-162

Елисеева И.И., Плошко Б.Г. История статистики. М.: Финансы и статистика, 1990

Елисеева И.И., Юзбашев М.М. Общая теория статистики. Учебник. М.: Финансы и статистика, 1995

История и теория статистики в монографиях Вагнера, Рюмелина. Пер. С нем. под ред Ю.Э.Янсона. СПб., 1879

История теоретической социологии. М.: Наука, т.1, 1995 (с. 215 –226 – «Программа статистически-вероятностно ориентированной науки об обществе», о творчестве Кондорсе)

Каблуков Н.А. Рец. на кн.: Майо-Смит Р. «Статистика и социология» // Вестник воспитания, №5, 1901, с. 27-30

Каблуков Н.А. Статистика. Теория и методы статистики. Основные моменты ее развития. Краткий очерк народонаселения. М., 1918

Каджаева С.Т. Лаплас о роли статистики в изучении общества // Социология: 4М, 2001, №13. С. 137-145

Карпенко Б.И. Развитие идей и категорий математической статистики. М.: Наука, 1980

Карпенко Б.И. Жизнь и деятельность А.А.Чупрова // Уч. Зап. по статистике, т.3. М., 1957

Карпухин Ю. Совершенствовать статистику преступлений // Вестник статистики, 1989, №8

Кауфман А.А. К вопросу о статистическом методе в историко-экономических исследованиях // Научно-исторический журнал. №1, 1913, с. 1 - 39

Кауфман А.А. Статистика. Ее приемы и значение для общественных наук. Лекции. М.: Тип. Т-ва И.Д.Сытина, 1911. 210 с.

Кауфман А.А. Теория и методы статистики. М.: Тип. И.Д.Сытина, 1912. 632 с.

Кауфман А.А. Вопросы второй всеобщей переписи // Статистический вестник, кн.1 и 2, 1914

Кауфман А.А. Статистическая наука в России. Теория и методология. 1806-1917 г.г. Историко-критический очерк. М.: 1922

Кетле А. Социальная система и законы, ею управляющие. СПб.,1899 (1866 ?)

Кетле А. Социальная физика или опыт исследования о развитии человеческих способностей. Т.1,2. К.: Киевский коммерческий институт , 1911-1913

Кириллов И. Потребление алкоголя и социальные последствия пьянства и алкоголизма // Вестник статистики, 1991, №9

Колмогоров А.Н. Теория вероятностей // Математика, ее содержание, методы и значение. М.,1956. Гл. X1. См. также: Вероятность и математическая статистика. Энциклопедия. М.: БРЭ, 1999. С. –871-881

Кондорсе Ж.А. Эскиз исторической картины прогресса человеческого разума. М., 1936

Краткие очерки по истории Российской государственной статистики, подготовлены коллективом авторов в следующесм составе: И.И.Елисеева, В.М.Проскуряков, В.И.Успенский, В.М.Ключников, В.Я.Роговая, Б.Г.Сивориновский, Д.Б.Сумкин

Лапин Н.И. Эмпирическая социология в Западной Европе. М.: Изд. Дом ГУ-ВШЭ, 2004

Лаплас П.С. Изложение системы мира. Т. 1,2. СПб., 1861

Лаплас П.С. Опыт философии теории вероятностей. М., 1908. Указанный перевод сочинения Лапласа опубликован в работе: Вероятность и математическая статистика. Энциклопедия. М.: БРЭ, 1999. С. 834-863.

Литвинова Е.Ф. Лаплас и Эйлер. СПб.,1982

Максименко В.С., Паниотто В.И. Зачем социологу математика. К.: Радянська школа, 1988.

Маркович, Хмельницкий. Организация моральной статистики во Франции // Вестник статистики, 1989, №7

Маслов П.П. Статистика в социологии. М.: Статистика, 1971. (Особенно с. 79-134: Границы вероятностных критериев в социологии. Применение математики в социально-экономическом исследовании)

Моральная статистика 20-х годов // История статистики. Вып. 1. М., 1991;Вып.2.М., 1990.

Некрасов П.А.Философия и логика науки о массовых проявлениях человеческой деятельности. Пересмотр оснований социальной физики Кетле.М., 1902

Петти В. Политическая арифметика // Экономические и статистические работы. М., 1940

Петти В. Пять ответов по политической арифметике // Там же

Плышевский Б.П. Проблемы развития территориальной статистики в Российской Федерации // Проблемы прогнозирования. !996, №2

Птуха М.В. Очерки по истории статистики XYII - XYIII вв. М., 1945

Райхесберг Н.М. А.Кетле. Его жизнь и научная деятельность. СПб, 1894

Ракитов А.И. Статистическая интерпретация факта и роль статистических методов в построении эмпирического знания. М., 1981

Саймон Г. Стратегия построения моделей в социальных науках // Математические методы в современной буржуазной социологии. М.: Прогресс, 1966. С.144-174

Сачков Ю.В. Вероятностная революция в науке (вероятность, случайность, независимость, иерархия). М.: Научный мир, 1999

Социальная статистика. Учебник. М.: Финансы и статистика, 1998. С. 144-157

Социология в России XIX - начала XX веков. Тексты (вып.1,2). М.: МУБиУ, 1997 (статьи ряда известных русских социологов, посвященные методам исследования)

Справочное пособие по истории немарксистской западной социологии. М.: Наука, 1986

Тахтарев К.М. Законы статистические и статистико-социологические ... // Социология в России XIX - начала XX веков. Тексты (вып. 2). М.: МУБиУ, 1997. С.223-244

Тихомиров П.В. Математический проект реформы социологии. Книга Некрасова П.А. Сергиев Посад, Свято-Сергиева Лавра, 1903

Толстова Ю.Н. Измерение в социологии. М.: Инфра-М, 1998

Толстова Ю.Н. Социология и математика. М.: Научный мир, 2003

Трофимов В.П. Измерение взаимосвязей социально-экономических явлений. М.: Статистика, 1975 (особенно с.5-29: Условия применимости математико-статистических методов для измерения взаимосвязей социально-экономических явлений)

Тутубалин В.Н. Границы применимости (вероятностно-статистические методы и их возможности). М.: Знание, 1977

Фортунатов А. Социология и статистика // Вестник воспитания, 1904, 5

Фортунатов А. Наука ли статистика? // Русское богатство, №2, 1896, с. 73-83

Фортунатов А. Количественный анализ в обществоведении // Народное хозяйство. №1, 1900, с.9-20

Хейс Д. Причинный анализ в статистических исследованиях. М.: Финансы и статистика, 1981

Чесноков С.В. Детерминационный анализ социально-экономических данных. М.: Наука, 1982

Чесноков С.В. Гуманитарные эмпирические исследования и обобщение силлогистики Аристотеля // Неклассические логики. М.: ИФАН СССР, 1985

Чесноков С.В. Основы гуманитарных измерений. М.,1986

Чупров А.А. О теории вероятностей и математической статистике. Переписка Маркова А.А. и Чупрова А.А. М., 1977

Чупров А.А. Нравственная статистика // Брокгауз Ф.А. (Лейпциг), Ефрон И.А. (СПб). Энциклопедический словарь. 1897

Чупров А.А. Очерки по теории статистики. СПб., 1909, 1910. М., 1959

Чупров А.А. Вопросы статистики. М.: Госстатиздат ЦСУ СССР, 1960

Чупров А.А. Закон больших чисел в современной науке (статистические основы научного мировоззрения) // Статистический вестник, №1, с. 1-21

Чупров А.И. Статистика. Киев: тип-литография «Прогресс», 1907. Перепечатано с издания кассы взаимопомощи С.-Петербургского Политехнического Института, исправленного А.А.Чупровым в 1907 году. Издание библиотеки студентов юристов

Чупров А.И. История статистики. М., 1892

Чупров А.И. О значении статистики для правоведения и ее успехи в России за последнее время // Юридический вестник, ч.4, с. 551-567

Чупров А.И. Статистика как связующее звено между естествознанием и обществоведением // Сборник правоведения и общественных знаний. СПб: тип. М.М.Стасюлевича. Т.3, с. 7-20

Шведовский В.А. Детерминизм и статистичность в динамических моделях // Социс,1985, 1, 128-134

Экспертные оценки в социологических исследованиях. - Киев: Наукова думка, 1990. С. 286-287 (принцип Кондорсе, парадокс Кондорсе).

Янсон Ю. Направления в научной обработке нравственной статистики. Введение в сравнительную нравственную статистику. Спб.: Тип. К.Вульфа, 1871

Янсон Ю.Э. Теория статистики. СПб, 1913, 1887

Янсон Ю.Э. Сравнительная статистика населения.

Янсон Ю. Сравнительно-статистические этюды // Знание. №2, с.133-151, №4, с. 59-87, №8, с. 223-254

Янсон Ю.Э. (ред.) Сборник монографий по теории и истории статистики.

Яхот О. Статистика в социологическом исследовании. М.: Знание, 1966

Раздел 3. Методы русской земской статистики

1. Анализ связи решаемых земской статистикой задач с соприкасающимися явлениями жизни (земские статистики ставили своей целью отражение местной общественной жизни как единого целого);
2. Опрос на деревенском сходе с точки зрения современной теории метода фокус-групп
3. Способы обеспечения достоверности собираемой информации
4. Способы построения “сети интервьюеров”. Требования, предъявляемые к последним
5. Методы сбора данных и их соотношение с современными подходами и друг с другом
6. Монографический метод в земской статистике и современная теория монографического подхода
7. Съезды и совещания земских статистиков о методах работы
8. Выборка в творчестве земских статистиков
9. Земская статистика о группировке материала (в частности, роль выбора основаниягрупировки)
10. Лонгитюд в творчестве земских статистиков
11. Земская статистика и традиции русской интеллигенции
12. Сравнительный анализ пообщинного и подворного способов исследования крестьянских хозяйств
13. Экспедиционный способ сбора данных; его соотношение с методами, традиционными для современной социологии
14. Сравнение экспедиционного метода с методом фокус-групп
15. Монографический метод в земской статистике, сравнение с подходами, развивавшимися в то же время в других странах (например, во Франции Ле Пле (1806-1882)) и с современной теорией монографического подхода
16. Съезды и совещания земских статистиков о методах работы
17. Выборка в земской статистике
18. Методы группировки в земской статистике (в частности, выбор основания группировки)
19. Лонгитюд в творчестве земских статистиков
20. Принципы земской статистики и традиции русской интеллигенции

Литература к разделу 3.

Абрамов В.Ф., Живоздрова С.А. Земская статистика - национальное достояние // Социс, 1996,2.С.89-99

Абрамов В.Ф. Земская корреспондентская сеть в Московской губернии // Социс, 1996, 5

Абрамов В.Ф. Земская статистика народного образования // СОЦИС, 1996, 9

Анисимов И.Я. Хозяйственно-экономические данные земской статистики. Вып.1. Полтава, 1885

Библиографический указатель земской оценочной литературы. Вып. 1-5; Московская, Черниговская, Тверская, Вятская, Тамбовская и Рязанская губернии. СПб, 1899 -1904

Богданов И.М. Обзор изданий губернских земств по статистике народного образования. Харьков, 1913

Велецкий С.Н. О приемах и способах исследования в 1898 г. В Уфимской губернии продовольственных, семенных и кормовых нужд населения в связи с некоторыми результатами этих исследований. Спб: тип. В.Димакова

Велецкий С. Земская статистика. В 3-х т. М., 1900 (с приложением земских статистических изданий по губерниям)

Веселовский Б.Б. История земства за 40 лет. В 4-х томах. СПб, 1909-1912.

Гозулов А.И. Очерки истории отечественной статистики. М., 1972

Григорьев В.Н. Предметный указатель в земско-статистических трудах с 1860-х годов по 1917 г. Вып. 1-2. М., 1926-27

Ермолаев А.В., Забаев И.В. К вопросу о методике земских статистических исследований // Социс, №11, 2001

Земское дело (журнал). 1910-1918

Каблуков Н.А. Статистика. Теория и методы статистики. Основные моменты ее развития. Краткий очерк народонаселения. М., 1918

Каблуков Н.А. Задачи и способы собирания статистических сведений (для чего и как собираются эти сведения). М.,1920

Караваев В.Ф. Издания земств 34-х губерний по общей экономической и оценочной статистике, вышедшей за время с 1864 по 1-е января 1911 г. СПб, 1911

Караваев В.Ф. Библиографический обзор земской статистической и оценочной литературы со времени учреждения земств, 1964 - 1903 гг. В 2-х выпусках. СПб, 1906-1913

Караваев В.Ф. Библиографический обзор земской статистической и оценочной литературы // Труды вольного экономического общества, 1909-1911

Карышев Н. Итоги экономического исследования России по данным земской статистики (вступительная статья: Фортунатов А. Общий обзор земской статистики крестьянского хозяйства). Т.1. М.: тип А.И.Мамонтова, 1892. Т.2. Дерпт: тип. Г.Лакмона, 1892.

Кауфман А.А. Сборник статей. Община. Переселение. Статистика. М.: изд-во Лемана и Плетнева, 1915

Кауфман А.А. Теории и методы статистики. М., 1912а

Кауфман А.А. Статистика, ее приемы и значение для общественных наук. М.: научное изд-е, 1912б

Кауфман А.А. Статистическая наука в России. Теория и методология. 1806-1917 г.г. Историко-критический очерк. М.: 1922

Куркин П.И. К вопросу о построении схемы земской санитарной статистики // Вопросы санитарной статистики. М., 1961

Куркин П. Земская санитарная статистика (опыт систематической библиографии). М.,1904

Пландовский В.А. Народная перепись. Спб, 1898

Сборник материалов для изучения сельской поземельной общины. Императорское Вольное экономическое Общество, 1880

Свавицкие З.Н. и Н.А. Земские подворные переписи 1880-1913 гг. Поуездные итоги. М.: ЦСУ СССР, 1926

Свавицкий Н.А. Земские подворные переписи (обзор методологии . М., 1961

Свод постановлений Казанского губернского земского собрания с 1865 по 1887 гг. Казань, 1887

Слуцкий Е.Е. Теория корреляции и элементы учения о кривых распределения. Киев, 1912

Статистические материалы по волостному и сельскому управлениям 34 губерний, в коих введены земские установления. СПб, 1886

Фортунатов А.Ф. Земско-статистическая литература в 1890 и 1891 гг // Юридический вестник, 1891. Кн. 5-6, 1892. Кн. 7-8

Фортунатов А.Ф. Общий обзор земской статистики крестьянского хозяйства. Т.1, М., 1892

Фортунатов А.Ф. Сельскохозяйственная статистика европейской России. М., 1893

Фортунатов А.Ф. О земско-статистических изданиях в 1910 г. СПб, б/г

Фортунатов А. Социология и статистика // Вестник воспитания, 1904, 5

Фортунатов А. Наука ли статистика? // Русское богатство, №2, 1896, с. 73-83

Фортунатов А. Количественный анализ в обществоведении // Народное хозяйство. №1, 1900, с.9-20

Черненков Н.Н. К характеристике крестьянского хозяйства. М.: изд. Лиги аграрных реформ, 1918

Чупров А.И. Статистика. Киев: тип-литография «Прогресс», 1907. Перепечатано с издания кассы взаимопомощи С.-Петербургского Политехнического Института, исправленного А.А.Чупровым в 1907 году. Издание библиотеки студентов юристов

Чупров А.И. История статистики. М., 1892

Щербина Ф.А. Крестьянские бюджеты

Юбилейный земский сборник \ под ред Б.Б.Веселовского, З.Г.Френкеля. СПб, 1914

Янсон Ю. Направления в научной обработке нравственной статистики. Введение в сравнительную нравственную статистику. Спб.: Тип. К.Вульфа, 1871

Янсон Ю.Э. Теория статистики. СПб, 1913, 1887

Янсон Ю. Сравнительно-статистические этюды // Знание. №2, с.133-151, №4, с. 59-87, №8, с. 223-254

Янсон Ю.Э. (ред.) Сборник монографий по теории и истории статистики.

Раздел 4. Выборочный метод в социологии

1. Виды неслучайных выборок, использующихся в социологии. Достоинства и недостатки каждого.
2. Способы построения случайной выборки. Моделирование случайности.
3. Стратификационная и гнездовая выборки. Их преимущества и недостатки.
4. Типологическая выборка. Использование методов многомерной классификации при ее построении.
5. Проблемы построения территориальной выборки.
6. Проблема недостижимых единиц при построении выборки
7. Малые выборки в социологии
8. Проблема олпределения генеральной совокупности
9. Роль математической статистики при построении выборки

10. Выборка в маркетинговых исследованиях

Литература к разделу 4.

Батыгин Г.С. Лекции по методологии социологических исследований. - М.: Аспект Пресс, 1995 С. 145-189.

Бауман М. Типические группы и способ их построения для проведения выборочного обследования. – Вопросы статистики. - 1996. – N 11.

Божков О.Б. Эта неуловимая генеральная совокупность // Социс. - 1987. - N 3.

Боярский А.Н. Выборочное наблюдение в статистике СССР.М.: Статистика, 1966.

Браверман Э.М. и др. Метод стратифицированной выборки в организации сбора эмпирических данных.- Автоматика и телемеханика. - 1995. - N 10.

Венецкий И.Г. Теоретические и практические основы применения выборочного метода. М.: НИИ по проектированию вычислительных центров и систем экономической информации ЦСУ СССР, 1971. –321 с.

Воронов Ю.П. Активный отбор объектов наблюдения при планировании выборочных социологических обследований. – Докл. Всес. Симп. По социологическим проблемам села. Наука, 1968. С. 170-180.

Воронов Ю.П. Применение методов таксономии в планировании выборочного исследования // Распознавание образов в социальных исследованиях. - Новосибирск: ИЭ и ОПП, 1968.

Воронов Ю.П. Районированные выборки в социологии. - Новосибирск, 1970.

Дружинин Н.К. Выборочный метод и его применение в социологических исследованиях. М.: Статистика, 1970.

Елисеева И.,Соколов Я. Выборочный метод в аудите // Вестник статистики. - 1993. - N9. С. 52-58.

Елисеева И.И., Юзбашев М.М.Общая теория статистики. - М.: Финансы и статистика, 1996, гл.7.

Йейтс Ф. Выборочный метод в переписях и обследованиях. - М.: Статистика, 1965.

Кокрен У. Методы выборочного исследования. - М.: Статистика, 1976.

Королев Ю.Г. Выборочный метод в социологии. М., 1976.

Косолапов М.С. Принципы построения многоступенчатой вероятностной выборки для субъектов Российской Федерации // Социс. – 1997. - № 10.

Малхотра, Нэреш К. Маркетинговые исследования. Практическое руководство, 3-е издание.: пер. с англ. – М.: Издательский дом «Вильямс», 2002. – С. 417-433.

Ноэль Э. Массовые опросы. Введение в методику демоскопии. - М.: Ава-Эстра, 1993. С. 85-134.

Основы прикладной социологии / Отв ред. Шереги Ф.Э.,Горшков М.К. М.: Интерпракс, 1996.С.31-38.

Паниотто В.И.Качество социологической информации. - Киев: Наукова думка, 1985.

Проектирование и организация выборочного социологического исследования. - М.: ИСИ АН СССР,1977.

Рабочая книга социолога / Отв. ред Осипов Г.В. М.: Наука, 1983. С. 200-236.

Римский В.Л. Описание процесса отбора административных районов России для проведения социологического исследования по проек-ту "Российский монитор" // Российский монитор. - 1993. - вып.3. С.197-206.

Рукавишников В.О., Паниотто В.И., Чурилов Н.Н. Опросы населения. - М.: Финансы и статистика, 1984. С. 7-81.

Сотникова Г.Н. Методы контроля и компенсации смещений от недоступных единиц в выборочном социологическом исследовании // Применение математических методов и ЭВМ в социологических исследованиях. М.: ИСИ АН СССР, 1982. С.102-117.

Статистические методы анализа информации в социологических исследованиях / Отв. ред Осипов Г.В. М.: Наука, 1979. С. 69-91.

Территориальная выборка в социологических исследованиях / Отв. ред Рябушкин Т.В. М.: Наука, 1980.

Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. М.:Инфра-М, 2003. С.445-462

Чурилов Н.И. Проектирование выборочного социологического исследования. - Киев: Наукова думка, 1986.

Шереги Ф.Э. , Гуцу В.Г., Пакоян Г.Б. Выборка в опросах общественного мнения. - Кишинев, 1989.

Шляпентох В.Э. Проблемы репрезентативности социологической информации (случайная и неслучайная выборки в социологии). - М.: Статистика, 1976.

Ядов В.А.Социологическое исследование: методология, программа, методы. - Самара: "Самарский ун-т",1995. С.69-76.

Kalton. Introduction to survey sampling. - Sage university papers series in quantitative applications in the social sciences. No 35, 1983.

Kish L. Survey sampling. - N.-Y.: Wiley and sons inc., 1965.

Раздел 5. Социология и математика: проблема «взаимодействия»

1. Роль математики в социологии
2. Методологические проблемы «стыковки» социологии и математики в творчестве советских ученых
3. Проблемы использования математических методов в социологии в творчестве русских ученых второй половины XIX, начала XX века
4. Проблемы использования математических методов в социологии в творчестве западных ученых
5. Взгляды Лазарсфельда на роль математического языка в социологии
6. Лазарсфельд об операционализации понятий
7. Блейлок о проблеме измерения в социологии. Использование им математико-статистических представлений о случайных величинах
8. Идеи причинного анализа в трудах Блейлока
9. Сорокин о «квантофрении»
10. Вероятностный анализ причинности в творчестве П.Суппеса
11. «Жесткость» и «мягкость» в процессе использования математического аппарата в социологии

Литература к разделу 6

Арнольд В.И. «Жесткие» и «мягкие» математические модели // Математическое моделирование социальных процессов. М.: МГУ, 1998. С. 29-51.

Батыгин Г.С. Ремесло Пауля Лазарсфельда: Введение в научную биографию // Вестник АН СССР. 1990. №8. С. 94-108

Блейлок Х. Косвенное измерение в социальных исследованиях: некоторые неаддитивные модели // Математика в социологии. М.: Мир, 1977, с.282-300.

Лазарсфельд П. Методологические проблемы социологии // Социология сегодня. Проблемы и перспективы. М.: Прогресс, 1965 .

Лазарсфельд П. Логические и математические основания латентно-структурного анализа // Математические методы в современной буржуазной социологии. М.: Прогресс, 1966, с. 344-401.

Лазарсфельд П. Измерение в социологии // Американская социология. М.: Прогресс, 1972.

Лазарсфельд П. Латентно-структурный анализ и теория тестов // Математические метолы в социальных науках. М.: Прогресс, 1973 С. 42-53 .

Максименко В. С. , Паниотто В. И. Зачем социологу математика. Киев: Радяньска школа, 1988.

Моисеев Н. Н. Математика в социальных науках // Математические методы в социологическом сследовании. М. : Наука, 1981.

Новак С. Причинные интерпретации статистических связей // Математика в социологии. М.: Мир, 1977, с. 76-123.

Поверить алгеброй гармонию (размышления о месте математики в социологии) // Социологические исследования, 1989. N6. С. 120-130.

Сатаров Г. А. Математика в социологии: стереотипы, предрассудки, заблуждения //Социологические исследования. 1986. N3. С. 137-141.

Суппес П. Вероятностный анализ причинности // Математика в социологии. М.: Мир, 1977, с.50-75.

Толстова Ю.Н. Анализ социологических данных: методология, дескриптивная статистика, изучение связей между номинальными признаками. Ч. 1. М.: Научный мир, 2000.

Толстова Ю.Н. Измерение в социологии. М.: Инфра-М, 1998

Толстова Ю.Н. Социология и математика. М.: Научный мир, 2003.

Швери Р. Теоретическая социология Джеймса Коулмена: аналитический обзор // Социологический журнал, 1996, № 1-2. С. 62 - 81.

Blalock H.M. The measurement problem: A gap between languages of theory and research // Methodology in social research / Ed. by H.M. Blalock. N.Y.: Mc Grow Hill, 1968, p. 5-27.

Blalock H. Theory construction: from verbal to mathematical formulations. Englewood Cliff, N.-Y.: Prentice-Hall, 1969 (метод построения особых причинных моделей, способствующих переводу вербальных теорий в математическую форму) .

Blalock H.M. Formalization of sociological theory // Theoretical sociology: Perspectives and developments Ed. By J.McKinney, E.A.Tiryakian. N.-Y.: Appleton-Century-Crofts, 1970, p. 271-300.

Blalock H. M. Power and conflict: Toward a general theory. Newburg Parc – L. – New Delhi : Sage publ., 1989 (Рецензия М.М.Назарова в: Социс, 1991, № 6. С. 148-150)

Blalock H.M. Basic dilemmas in the social sciences. London: Sage Publications, 1990.

Blalock H.M. Conceptualization and measurement in the social sciences. Beverly Hills, London: Sage Publications, 1990

Boudon R. L′analyse mathematic de faits social. Paris, 1967

The varied sociology of Paul F. Lazarsfeld / Wrighting collected and edited by P.L.Kendall. N.-Y.: Columbia university press, 1982. P. 239-285

Lazarsfeld P. The art of asking why: three principles underlying the formulation of questionnaires // National marketing reviews. 1935. No. 1. P.32-43

Lazarsfeld P. The controversy over detailed interviews: An offer for negotiation // Public opinion Quarterly. 1944. V.8. No.1. P. 38-60. Переиздано в: On social research and its language / Ed. by R. Boudon. Chicago: the university of Chicago Press. 1993. P. 109-129

Lazarsfeld P. Interpretation of statistical relations as a research operation // The language of social research. Glencoe, Ill: The Free Press, 1955

Lazarsfeld P. Evidence and inference in social research // Deadalus, 1958, #87/

Lazarsfeld P. Note on the history on quantification on sociology – trends, sources and problems // ISIS, 1961. V. 52, # 158, p. 277-333.

Lazarsfeld P. Concept formation and measurement in the behavioral sciences: some historical observations // Concept, theory and explanation in the behavioral sciences / Ed. By G.Di Renzo. N.-Y.: Random House, 1966, pp. 140-202.

Lazarsfeld P. The use of panels in social research // The varied sociology of Paul F. Lazarsfeld / Wrighting collected and edited by P.L.Kendall. N.-Y.: Columbia university press, 1982. P. 239-285

Lazarsfeld P. Analyzing the relations between variables // On social research and its language / Ed. by R. Boudon. Chicago: the university of Chicago Press. 1993

Lazarsfeld P. Some remarks on typological procedures in social science // On social research and its language / Ed. by R. Boudon. Chicago: the university of Chicago Press. 1993. P. 158-171

Lazarsfeld P., Rosenberg M. The language of social research // A reader in the methodology of social research, 1965

Nowak S. Some problems of causal interpretation of statistical relationships // Philosophy of science, XXVII, 1960, pp.23-38.

Nowak S. Studies in the methodology of the social sciences. Warsaw, 1965. Там, в частности: Causal interpretation of statistical relationships in social research, а также: General laws and historical generalizations in the social sciences.

Nowak S. Understanding and prediction: essays in the methodology of social and behavioral theories. Dordrecht, Boston: D. Reidal Publishing company, 1976

Nowak S. Methodology of social research: General problems. Dordrecht, Boston: D. Reidal Publishing company, 1977.

Sorokin P.A. Fads and Foibles in modern sociology and related sciences. Chicago: Henry Regnery Co., 1956 (с. 68-174 - о «квантофрении»).

Suppes P. A probabilistic theory of causality. Amsterdam: North-Holland Publishing Company, 1970

Zaller R.A., Carmines E.G. Measurement in the social sciences. The link between theory and data. Cambridge: Cambridge University Press, 1980 (выявление систематических ошибок измерения, оценок надежности и валидности с помощью факторного анализа).

Zaller R.A., Carmines E.G. Reliability and validity assessment. Sage University Paper series on quantitative applications in the social sciences, 07-017. Beverly Hills and London: Sage Publications, 1990

Zetterderg H. On theory and verifications in sociology. N.-Y.,1964 (метод дефиниционной редукции, т.е. постепенное понижение уровня обобщений исходных теоретических допущений до операционального уровня с помощью дедуктивного вывода)

Приложение 4
Статистические таблицы

Должны быть взяты из другой книги

……

Автор учебного пособия д.с.н., проф.

Ю.Н.Толстова