МИНИСТЕРСТВО ВЫСШЕГО И СРЕДНЕГО СПЕЦИАЛЬНОГО ОБРАЗОВАНИЯ РЕСПУБЛИКИ УЗБЕКИСТАН
ФЕРГАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
СОЦИАЛЬНО – ЭКОНОМИЧЕСКИЙ ФАКУЛЬТЕТ
КАФЕДРА СОЦИОЛОГИИ

Б. НУМАНЖАНОВ

Методы выборки в социологических исследованиях

(Учебно - методическое указание)

Фергана 2009

В данном учебно – методическом указании были разработаны методологии, относящиеся к «методы выборки в социологических исследования» и, в соответствии с учебным планом, систематизированы методы исследования. Это учебно - методическое указание рассчитано для профессоров – преподавателей, студентов по направлению социологии.

Данное учебно - методическое указание было рассмотрено на на учёном совете ФерГУ от 24 апреля 2009 года.

Составитель: **старший преподаватель, кандидат философских наук. Нуманжанов. Б.**

Ответсвенный редактор: **доктор философских наук, профессор. Тургунбаев Ф.**

Рецензенты: **доктор философских наук, проф. Т. Абдуллаев**

доцент кафедры истории Узбекистана ФерПИ А. Набижанов

Оглавление

Введение……..3 -4 ст

Понятие «выборки» в социологии ……. 4 -12 ст

Использование других исследовательских методов в методе

выборки …… 12 -14 ст

Простая случайная выборка и её особенности …….14 -18 ст

Систематическая и продолжительная выборка ……18 -20 ст

Территориальная выборка и её особенности …… ……20 -22 ст

Общая характеристика квотной, многоступенчатой выборки …….23 -31 ст

Методы кластерного анализа ……32 -33 ст

Неконтролируемое обучение против контролируемого обучения
общая классификация кластерной выборки ……34 -40 ст
Выборка теоретическая ……..40 -50 ст

Выборка ошибка ……50 -53 ст

Литература ……..53 -54 ст

Темы для самостоятельной работы …… 54 ст

ВВЕДЕНИЕ

В структуре социологии выделяют три взаимосвязанных уровня: общесоциологическую теорию, специальные социологические теории и социологические исследования. Их называют еще частными, эмпирическими, прикладными или конкретными социологическими исследованиями. Все три уровня дополняют друг друга, что позволяет получить при изучении социальных явлений и процессов научно обоснованные результаты.

Социологическое исследование – это система логически последовательных методологических, методических и организационно-технических процедур, подчиненных единой цели: получить точные объективные данные об изучаемом социальном явлении.

Исследование начинается с его подготовки: обдумывания целей, программы, плана, определения средств, сроков, способов обработки и т. д.

Второй этап – сбор первичной социологической информации (записи исследователя, выписки из документов).

Третий этап – подготовка собранной в ходе социологического исследования информации к обработке, составление программы обработки и сама обработка.

Заключительный, четвертый этап – анализ обработанной информации, подготовка научного отчета по итогам исследования, формулирование выводов и рекомендации для заказчика, субъекта.

Понятие «выборки» в социологии

Вид социологического исследования определяется характером поставленных целей и задач, глубиной анализа социального процесса.

Различают три основных вида социологического исследования: разведывательное (пилотажное), описательное и аналитическое.

Разведывательное (или пилотажное, зондажное) исследование самый простой вид социологического анализа, позволяющий решать ограниченные задачи. Идет обработка методических документов: анкеты, бланк-интервью, опросного листа. Программа такого исследования упрощена. Обследуемые совокупности невелики: от 20 до 100 человек.

Разведывательное исследование обычно предваряет глубокое изучение проблемы. В ходе его уточняются цели, гипотезы, задачи, вопросы, их формулировка.

Описательное исследование – более сложный вид социологического анализа. С его помощью получают эмпирическую информацию, дающую относительно целостное представление об изучаемом социальном явлении. В описательном исследование возможно применение одного или нескольких методов сбора эмпирических данных. Сочетание методов повышает достоверность и полноту информации, позволяет сделать более глубокие выводы и обоснованные рекомендации.

Самый серьезный вид социологического исследования – аналитическое исследование. Оно не только описывает элементы изучаемого явления или процесса, но и позволяет выяснить причины лежащие в его основе. Главное назначение такого исследования – поиск причинно-следственных связей.

Аналитическое исследование завершает разведывательное и описательное исследования, в ходе которых собираются сведения, дающие предварительное представление об определенных элементах изучаемого социального явления или процесса.

Подготовка социологического исследования непосредственно начинается не с составления анкеты, а с разработки его программы, состоящей из дух разделов – методологического и методического.

В методологический раздел программы входят:

а) формулировка и обоснование объекта и предмета социальной проблемы;

б) определение объекта и предмета социологического исследования;

в) определение задач исследователя и формулировка гипотез.

Методический раздел программы предполагает определение изучаемой совокупности, характеристику методов сбора первичной социологической информации, последовательность применения инструментария для ее сбора, логическую схему обработки собранных данных.

Существенной частью программы любого исследования является прежде всего глубокое и всестороннее обоснование методологических подходов и методических приемов изучения социальной проблемы, под которой следует понимать «социальное противоречие», осознаваемое субъектами как значимое для них несоответствие между существующим и должностным, между целями и результатами деятельности, возникающее из-за отсутствия или недостаточности средств для достижения целей, препятствий на этом пути, борьбы вокруг целей между различными субъектами деятельности, что ведет к неудовлетворению социальных потребностей[1].

Важно различать объект и предмет исследования. Выбор объекта и предмета исследования в определенной мере уже заложен в самой социальной проблеме.

Объектом исследования могут любой социальный процесс, сфера социальной жизни, трудовой коллектив, какие-либо общественные отношения, документы. Главное, чтобы все они содержали социальное противоречие и порождали проблемную ситуацию.

Предмет исследования – те или иные идеи, свойства, характеристики, присущие данному коллективу, наиболее значимые с практической или теоретической точки зрения, т. е. то, что подлежит непосредственному изучению. Другие свойства, черты объекта остаются вне поля зрения социолога.

Анализ любой проблемы можно провести в теоретическом и прикладном направлениях в зависимости от цели исследования. Цель исследования может быть сформулирована как теоретическая. Тогда при подготовке программы основное внимание уделяется теоретическим и методологическим вопросам. Объект исследования определяется только после того, как выполнена предварительная теоретическая работа.

Объект исследования чаще всего насчитывает сотни, тысячи, десятки сотни тысяч людей. Если объект исследования состоит из 200-500 человек, они все могут быть опрошены. Такой опрос будет сплошным. Но если объект исследования насчитывает более 500 человек, то единственно верным способом будет применение выборочного метода.

Выборка – это совокупность элементов объекта социологического исследования, подлежащая непосредственному изучению[2].

Выборка должна учитывать взаимосвязи и взаимообусловленности качественных характеристик и признаков социальных объектов, говоря проще, единицы опроса выбираются на основании учета важнейших признаков социального объекта – образования, квалификации, пола. Второе условие: при подготовке выборки необходимо, чтобы отобранная часть являлась микромоделью
целого, или генеральной совокупности. В определенной степени генеральная совокупность есть объект исследования, на который распространяются выводы социологического анализа.

Выборочная совокупность – это определенное число элементов генеральной совокупности, отобранное по строго заданному правилу. Элементы выборочной совокупности, подлежащие изучению, есть единицы анализа. Ими могут выступать как отдельные люди, так и целые группы (студенческие), рабочие коллективы.

1.Понятие выборки. Выборочный метод.

Одной из задач, которые стоят перед социологом при проведении исследования, является сбор необходимых эмпирических данных об объекте исследования. Данные о массовых социальных явлениях и процессах социолог может получить из двух видов источников:

1. Объективных, к которым относятся официальная государственная статистика, статистика министерств и ведомств, служб социальной защиты, профессиональных союзов, общественных партий и движений и такое прочее. Они обычно представляют собой обобщённые количественные характеристики социальных общностей, явлений, процессов (например, уровень безработицы, численность и состав партий и общественных объединений, национальный валовой продукт, численность населения и другое). Но эти данные не всегда могут гарантируют точность и однозначность. Например, занижены данные о распространённости наркомании или пьянства, так как регистрируются далеко не все такие случаи.
2. Субъективных, к которым и относятся сами люди. Только от них мы можем узнать о настроениях населения или отдельных социальных групп, только с их помощью спрогнозировать результаты выборов и определить рейтинги телепередач. При работе с людьми возникают, как минимум, две методологические проблемы:

- все данные, которые мы получим от отдельных людей, должны быть обобщены, если мы хотим охарактеризовать изучаемое явление или процесс;

- наиболее точные данные мы сможем получить, если изучить всю совокупность объектов, которые имеют отношение к изучаемой проблеме (например, перепись населения). Но такие исследования (сплошные обследования) очень трудоёмки и дорогостоящи, а в информации от

субъективных источников общество нуждается постоянно. Поэтому большинство исследований бывают выборочными[1].

Как только нужно собрать информацию о некоторой группе или большой совокупности людей, возникает проблема построения выборки. Её как правило используют в опросах, ориентированных на статистические методы, в исследованиях политических и культурных элит, при отборе «случаев» для включённого наблюдения и качественного анализа.

Считается, что статистические обследования населения и ресурсов зародились одновременно с первыми формами централизованной социальной и политической организации: информацию такого рода использовали при решении различных управленческих задач – начиная с политики и заканчивая строительством общественных бань еще в развитых аграрных обществах и древних городах-государствах. Иногда эти обследования принимали форму сплошных переписей населения, но чаще всего довольствовались «сведениями о какой-то части совокупности: об урожайности судили по пробному обмолоту, о партии товара – по образцу, а о прихожанах – по их духовному наставнику.

Выборка – это подмножество заданной совокупности (популяции), позволяющее делать более или менее точные выводы относительно совокупности в целом»[2]. Но вообще-то термин "выборка" имеет двоякое значение. Это и процедура отбора элементов исследуемого объекта, и совокупность элементов объекта, выбранных для непосредственного обследования. Причины применения выборочного метода:

a) экономит силы и средства исследователей;
b) процедура представляет собой удобную и экономичную форму индуктивного вывода (рассуждение по схеме «от частных наблюдений – к общей эмпирической закономерности»);
c) реализует принцип рандомизации (случайного отбора).

[1] Бабосов Е.М. Прикладная социология: Учеб. пособие для студентов вузов. 2-е изд., стереотип. – Мн.: «ТетраСистемс», 2001. С. 330 – 331.
[2] Девятко И.Ф. Методы социологического исследования. – 3-е изд. – М.: КДУ, 2003. С. 200.

«Представление о том, что отбор наблюдений должен носить случайный, непредумышленный характер, в общем, соответствует нашему интуитивному знанию об условиях вынесения объективного и непредвзятого суждения»[3]. Но стоит заметить, что теория случайной выборки не часто использовалась вплоть до конца XIX – начала XX веков профессиональными статистиками, хотя теория вероятностей достигла высочайшего уровня развития уже в XVIII – первой половине XIX веков, так как сложилось убеждение о том, что в основе отбора должна лежать не «игра случая», а поиск типичных, характерных наблюдений. «Применимость выборочного метода для изучения случайно распределённых признаков, например дохода или размера семьи, была впервые обоснована в работах норвежца А. Киэра, англичан А. Боули и К. Пирсона, а также русского статистика А.И. Чупрова.

Следующий шаг в развитии выборочного метода можно связать с именем Р. Фишера, который разработал технику рандомизации в эксперименте и выборочном наблюдении. Выборочный метод часто используется как «замена» экспериментального метода. Например, мы не можем провести эксперимент, в котором людям в случайном порядке присваиваются определённые значения переменных «пол» или «цвет кожи», но выборочный метод помогает справиться с этими ограничениями и делать выводы о взаимосвязях между разными переменными, в том числе и вышеуказанными. На изучаемые переменные оказывают систематическое влияние посторонние факторы, которые в свою очередь мешают сделать обоснованные выводы. Единственное средство для достижения обоснованных выводов – это абсолютно случайный характер отбора наблюдений. Только равенство шансов для каждого наблюдения попасть в выборку (отбор «наугад»), гарантирует отсутствие намеренных или ненамеренных искажений.

Поэтому наилучшим способом отбора считается вероятностная или случайная выборка, в которой строго соблюдается принцип равенства

[3] Там же.

шансов попадания в выборку и для всех единиц изучаемой совокупности, и для любых последовательностей таких единиц[4].

Определившись с понятием выборки, необходимо определить еще несколько понятий, касающихся выборочного метода.

«Все множество социальных объектов, которые являются объектом изучения в пределах, очерченных программой социологического исследования и территориально-временными границами, образует генеральную совокупность»[5]. Любую генеральную совокупность характеризует какой-либо значимый признак или набор признаков, по которым мы можем отнести конкретный объект к данной совокупности. А «выборочная совокупность – это уменьшенная модель генеральной совокупности; те, кому социолог раздаёт анкеты, кого называют респондентами…Иначе говоря, это множество людей, которых социолог опрашивает»[6].

Корректное определение генеральной совокупности «включает ответы на следующие вопросы:

1) какие именно объекты (элементы) составляют генеральную совокупность – отдельные люди, семьи, академические группы, предприятия, населённые пункты или целые государства;
2) какими признаками обладают элементы генеральной совокупности, насколько они доступны для определения;
3) какова численность генеральной совокупности;
4) как генеральная совокупность размещена территориально;
5) как генеральная совокупность ограничена во времени»[7].

Следует различать единицы отбора и единицы наблюдения. Единицами отбора являются единицы или группы единиц генеральной совокупности,

---

[4] Девятко И.Ф. Методы социологического исследования. – 3-е изд. – М.: КДУ, 2003. С. 199 – 200.
[5] Агабекян Р.Л. Математические методы в социологии. Анализ данных и логика вывода в эмпирическом исследовании: Учебное пособие для вузов / Р.Л. Агабекян, М.М. Кириченко, С.В. Усатиков. – Ростов н/Д: Феникс, 2005. С.82.
[6] Кравченко А.И. Социология: Общий курс: Учебное пособие для вузов. – М.: ПЕРСЭ; Логос, 2000. С.115.
[7] Бабосов Е.М. Прикладная социология: Учеб. пособие для студентов вузов. 2-е изд., стереотип. – Мн.: «ТетраСистемс», 2001. С. 333.

которые отбираются на каждом этапе формирования выборочной совокупности и выступают единицами счёта. Единицы наблюдения – это отобранные единицы генеральной совокупности, характеристики которых непосредственно измеряются, то есть элементы сформированной выборочной совокупности. Если выборка проходит в несколько этапов (многоступенчатая выборка), то единицы отбора и единицы наблюдения могут не совпадать[8].

Выборочный метод позволяет не только сократить временные и материальные затраты на проведения исследования, но и повысить достоверность результатов исследования. Это утверждение может вызвать недоумение: как можно получить более достоверные данные, обследовав менее половины генеральной совокупности. Как когда-то сказал Джордж Гэллап: «Если хорошо помешать суп, повар возьмёт на пробу одну ложку и скажет, какой вкус у всего горшка!»[9]. То есть можно сказать, что достоверность полученной информации может быть не только не ниже, чем при сплошном обследовании, но и выше вследствие возможности привлечения персонала более высокого класса и применения различных процедур контроля качества получаемой информации.

Кроме того, выборочный метод имеет широкую область применения. Широта области применения выборочного метода объясняется тем, что небольшой (по сравнению с генеральной совокупностью) объем выборки позволяет использовать более сложные методы обследования, включая использование различных технических средств (например, видео- и аудиоаппаратуры).

ИСПОЛЬЗОВАНИЕ ДРУГИХ ИССЛЕДОВАТЕЛЬСКИХ МЕТОДОВ В МЕТОДЕ И ВЫБОРКИ

На первом этапе выбираются какие-либо трудовые коллективы, предприятия, учреждения. Среди них отбираются элементы, имеющие типичные для всей группы признаки. Эти отобранные элементы называются

[8] Кравченко А.И. Социология: Общий курс: Учебное пособие для вузов. – М.: ПЕРСЭ; Логос, 2000. С.116.
[9] Анурин В.Ф. Эмпирическая социология: Учебное пособие для вузов. – М.: Академический Проект, 2003. С. 68.

– единицами отбора, а среди них выбираются единицы анализа. Данный метод называют механической выборкой. При такой выборке отбор может быть произведен через 10, 20, 50 и т. д. человек. Промежуток между отбираемыми называется – шагом отбора.

Довольно популярен метод серийной выборки. В нем генеральная совокупность делится по заданному признаку (полу, возрасту) на однородные части. Затем отбор респондентов идет отдельно из каждой части. Число респондентов, отбираемых из серии, пропорционально общему числу элементов в ней.

Иногда социологи используют метод гнездовой выборки. В качестве единиц исследования отбираются не отдельные респонденты, а целые группы и коллективы. Гнездовая выборка дает научно обоснованную социологическую информацию, если группы максимально схожи по важнейшим признакам, например по полу, возрасту, видам обучения.

Также в исследованиях применяется целенаправленная выборка. В ней чаще всего используются методы стихийной выборки, основного массива и квотной выборки. Метод стихийной выборки – обычный почтовый опрос телезрителей, читателей газет, журналов. Здесь заранее невозможно определить структуру массива респондентов, которые заполнят и отправят анкеты по почте. Выводы такого исследования можно распространять лишь на опрошенную совокупность.

При проведении пилотажного, или разведывательного, исследования обычно применяют метод основного массива. Он практикуется при зондаже какого-либо контрольного вопроса. В подобных случаях опрашивается до 60-70% респондентов, попавших в отборочную совокупность.

Метод квотной выборки часто применяется при опросах общественного мнения. Им пользуются в случаях, когда до начала исследования имеются статистические данные о контрольных признаках элементов генеральной совокупности. Число признаков, данные о которых выбираются в качестве квот, обычно не превышает четырех, так как при

большем числе показателей отбор респондентов становиться практически невозможным.

ПРОСТАЯ СЛУЧАЙНАЯ ВЫБОРКА И ЕЁ ОСОБЕННОСТИ

Выборка случайная простая - метод извлечения случайной выборки из генеральной совокупности за один этап. Предполагается, что имеется репрезентативная выборки основа в виде более или менее полного списка элементов генеральной совокупности и что объекты из этого списка извлекаются с помощью случайных (вероятностных) или рандомизирующих (обеспечивающих квазислучайность) процедур.

Наиболее простой и известной процедурой простого случайного отбора является лотерея. Если генеральная совокупность имеет значительный объем, применяются компьютерные программы-датчики случайных чисел (до широкого распространения персональных компьютеров обычно использовались таблицы случайных чисел), которые позволяют получить необходимое количество равномерно распределенных номеров из списка.

Наиболее известными рандомизирующими процедурами являются систематическая выборка, маршрутная выборка, некоторые способы отбора респондента в семье (по дате дня рождения и т.п.).

Такая выборка является наиболее точной, репрезентативность (способность выборки «правильно отражать состояние дел в генеральной совокупности, из которой она извлечена и для изучения которой предназначена»[10]) её достигается при помощи математических методов. Особенность случайной выборки заключается в том, что все единицы генеральной совокупности имеют равную вероятность попасть в выборочную совокупность.[11] По определению, при случайной выборке выполняется принцип случайности. «Равенство шансов попасть в выборочную совокупность – насколько необходимое, настолько же и сложно

[10] Бабосов Е.М. Прикладная социология: Учеб. пособие для студентов вузов. 2-е изд., стереотип. – Мн.: «ТетраСистемс», 2001. С. 331.
[11] Зборовский Г.Е., Шуклина Е.А. Прикладная социология: Учебное пособие. – М.: Гардарики,2004. С. 95.

осуществимое требование. Для обеспечения этой «статистической демократии» равенства шансов социолог, как правило, формирует основу выборки»[12], то есть полный и точный перечень или пронумерованный список всех элементов генеральной совокупности. Например, основой выборки могут выступать списки работников предприятия, телефонные справочники, регистрационные списки владельцев автомобилей, списки избирателей на избирательных участках, домовые книги, а так же составленные самим социологом различные списки в зависимости от целей исследования (список улиц, на которых потом проводится отбор респондентов).

Случайная выборка обычно применяется при опросах общественного мнения перед выборами, референдумами и другими массовыми мероприятиями[13].

Плюсом данного метода является полное соблюдения принципа случайности и, как следствие – избежание систематических ошибок.

Случайная выборка обладает рядом недостатков, которые затрудняют ее применение на практике:

1. Необходимость наличия списка элементов генеральной совокупности. Трудность здесь заключается в том, что получить такой список далеко не всегда представляется возможным. Следовательно, в тех случаях, когда невозможно получить список элементов генеральной совокупности, невозможно проводить и случайный отбор.

2. Сложность проведения опроса. Процедура опроса при случайном отборе является очень громоздкой и требующей много времени. Ведь в результате случайного отбора исследователь получает на выходе список фамилий респондентов (телефонов, адресов и т.д.), которых необходимо опросить. То есть, интервьюерам приходится «бегать» за каждым респондентом и добиваться от него согласия ответить на «парочку вопросов».

---

[12] Агабекян Р.Л. Математические методы в социологии. Анализ данных и логика вывода в эмпирическом исследовании: Учебное пособие для вузов / Р.Л. Агабекян, М.М. Кириченко, С.В. Усатиков. – Ростов н/Д: Феникс, 2005. С.83.

[13] Зборовский Г.Е., Шуклина Е.А. Прикладная социология: Учебное пособие. – М.: Гардарики,2004. С. 95.

Усложняет эту задачу и то, что респондентов порой бывает не так просто найти; в случае отсутствия респондента его приходится посещать по нескольку раз (по крайней мере, не менее трех раз).

Все вышеперечисленное ведет к повышенным временным затратам на проведение опроса. Временные затраты можно уменьшить только благодаря привлечению дополнительных интервьюеров, т.е. только за счет дополнительных денежных расходов. Кроме этого возникает еще так называемая проблема неответивших.

3. Сравнительно большой объем выборки. Для получения результатов со сравнительно высокой степенью точности случайный отбор требует достаточно большого объема выборки по сравнению с другими видами отбора. Другими словами, случайный отбор обладает меньшей степенью точности, что, в конечном счете, является причиной его меньшей эффективности. А выборка считается более эффективной, если: при одинаковых расходах она более точна, а при одинаковой точности она более дешевая.

«Простой случайный отбор из генеральной совокупности предполагает что:

- генеральная совокупность однородна;
- все её элементы доступны для исследования в одинаковой степени;
- имеется полный список элементов, составляющих генеральную совокупность (или хотя бы репрезентативная основа выборки);
- к этому списку применяются процедуры случайного отбора, с использованием таблиц или компьютерных генераторов случайных чисел»[14].

а) Метод жребия (или лотерейный метод).

[14] Бабосов Е.М. Прикладная социология: Учеб. пособие для студентов вузов. 2-е изд., стереотип. – Мн.: «ТетраСистемс», 2001. С. 336.

Каждый элемент (респондент) генеральной совокупности заносится на карточку (это могут быть фамилии, адреса, просто номера (в этом случае номера ставят в соответствие с людьми в списках) и т.д.), затем бумажки помещаются в урну или барабан, перемешиваются и, не глядя, вынимаются. Номера на выбранных карточках указывают на элементы генеральной совокупности, которые попадают в выборочную совокупность. После доставания каждой карточки, оставшиеся снова перемешиваются.

- простой случайно-повторный отбор – отбор, при котором выбранная карточка возвращаются обратно в урну, и затем отбор продолжается;
- простой случайно -бесповторный отбор – отбор, при котором выбранная карточка откладываются в сторону и отбор продолжается.

Отбор заканчивается, когда будет выбрано заранее заданное количество элементов выборочной совокупности[15].

Осуществление этого метода довольно трудоёмкая и продолжительная операция (особенно при больших объемах выборки), а для обеспечения равного шанса выбора каждого элемента генеральной совокупности, требуется тщательное перемешивание карточек после каждой выемки очередного номера[16].

б) Метод таблиц случайных чисел.

Для осуществления этого метода используют таблицы случайных чисел, которые «можно найти в справочниках по математической статистике. Отбор номеров из таблицы случайных чисел формирует выборочную совокупность. Таблицы устроены таким образом, что отбор можно осуществлять с начала, с конца, из середины, по горизонтали, по вертикали, поскольку числа от 0 до 9 имеют равную вероятность появиться в любой позиции таблицы»[17]. Сначала мы присваиваем элементам (респондентам) генеральной совокупности номера. Например, номера от 01 до 70 (если число элементов генеральной

[15] Основы прикладной социологии. Учебник для вузов. Колл. авторов. Под ред. Ф.Э. Шереги и М.К. Горшкова. М.: Интерпракс, 1996. С. 33.
[16] Анурин В.Ф. Эмпирическая социология: Учебное пособие для вузов. – М.: Академический Проект, 2003. С. 69.
[17] Зборовский Г.Е., Шуклина Е.А. Прикладная социология: Учебное пособие. – М.: Гардарики,2004. С. 96.

совокупности равно 70), но если бы максимальный номер в списке (количество элементов генеральной совокупности) был трёхзначным (например, 456), мы бы присваивали им трёхзначные номера, используя нули в отсутствующих разрядах (например, 067 или 005). Затем задаёмся произвольными номерами строки и столбца, цифра, находящаяся на их пересечение и будет номером первого респондента, а далее отбор можно проводить по любому правилу: подряд, через строку через два столбца и такое прочее. Выбирается количество чисел равное количеству элементов выборочной совокупности.

Если в процессе отбора попадаются числа, превосходящие по величине самый большой номер в списке или повторяющиеся, то их положено пропускать.

Так же если нужны, например, трёхзначные числа, а таблица состоит из пятизначных чисел, то используют, как правило, только первые три цифры каждого пятизначного числа, а оставшиеся две игнорируют[18].

Кроме таблиц случайных чисел в этом методе нередко используется генератор случайных чисел. Это то же самое, что и таблицы случайных чисел, только числа вырабатываются компьютером (для этого существует специальная программа).

СИСТЕМАТИЧЕСКАЯ И ПРОДОЛЖИТЕЛЬНАЯ ВЫБОРКА

Выборка стратифицированная (расслоенная) - метод извлечения выборки, основанный на предварительном расслоении (стратификации, разукрупнении) генеральной совокупности на крупные подсовокупности, называемые слоями. Выборка извлекается из каждого слоя, причем в разных слоях отбор производится независимо, и могут применяться разные способы отбора как статистические, так и нестатистические. Общий объем выборки

[18] Девятко И.Ф. Методы социологического исследования. – 3-е изд. – М.: КДУ, 2003. С. 206 – 207.

распределяется между слоями пропорционально их численности. Если в каждом слое берут простую случайную выборку, то способ отбора в целом называется расслоенным случайным отбором. Примером В.С. является национальная выборка для опросов общественного мнения, когда территория страны делится на области (регионы и пр.), и для каждой области строится отдельная выборка.

Расслоенный отбор рекомендуется применять в следующих случаях:

1) если каждый слой внутренне однороден в том смысле, что результаты измерения внутри слоя изменяются от объекта к объекту значительно меньше, чем результаты измерения от слоя к слою; это позволяет получить выигрыш в точности результатов;

2) если желательно получить репрезентативные данные не только о генеральной совокупности в целом, но и об ее структурных частях; каждая из которых рассматривается в этом случае как слой;

3) если это продиктовано организационными соображениями (например, использование административного деления территорий; см. также:

4) если трудно (дорого) получить основу выборки для всей генеральной совокупности, но это можно сделать для каждого слоя;

5) если проблемы, связанные с отбором в различных частях генеральной совокупности, сильно разнятся (например, крупные предприятия могут быть выделены в отдельный слой и подвергнуты сплошному отбору, в то время как мелкие фирмы обследуются выборочно).

Выборка систематическая - процедура квази-случайного отбора респондентов из списка генеральной совокупности, аналог выборки случайной простой. Шаг отбора устанавливается в зависимости от необходимого объема выборки n и объема генеральной совокупности N: $l = [N/n]$. Первый элемент В.С. выбирается случайным образом из первых l номеров списка: пусть это будет элемент с номером k ($1 \leq k \leq l$). Затем в выборку последовательно включаются объекты с номерами $k + l$, $k + 2l$, $k + (n\text{-}1)l$. То обстоятельство, что В.С. распределена по генеральной

совокупности более равномерно, делает систематический отбор иногда более точным, чем простой случайный отбор, однако его эффективность существенно зависит от особенностей генеральной совокупности.

Если в списке генеральной совокупности единицы расположены случайно, в нем нет никаких статистических закономерностей, ни корреляции между соседними единицами, то можно ожидать, что систематический отбор будет, в сущности, равносилен простому случайному отбору. В этом случае к В.С. применим весь математический аппарат, разработанный для простого случайного отбора. Такими качествами обычно обладают списки и картотеки, составленные в алфавитном порядке.

Если элементы генеральной совокупности упорядочены по возрастанию или убыванию некоторого показателя, коррелирующего с изучаемым признаком, систематический отбор может оказаться более эффективным, чем простой случайный.

Наконец, если генеральная совокупность содержит периодический тренд, то эффективность В.С. зависит от шага отбора *l*. Он не должен быть кратным периоду изменения значений признака; иначе выборка почти наверняка будет иметь систематическую ошибку. Например, если в качестве единицы отбора выступает квартира ("домохозяйство"), то при организации систематического выборочного опроса в многоквартирном доме (или на улице, застроенной многоквартирными домами) шаг отбора не должен быть кратен числу квартир на лестничной клетке. Иначе интервьюер каждый раз будет попадать в однотипные квартиры, что, конечно, повлияет на состав выборки.

ТЕРРИТОРИАЛЬНАЯ ВЫБОРКА И ЕЁ ОСОБЕННОСТИ

Территориальная выборки; может быть организована по-разному, даже если она будет примерно одного и того же типа.

Первая переменная, определяющая организацию исследования, относится к степени централизации сбора информации.

Вариант организации исследования с максимальной централизацией может быть реализован в двух модификациях. В соответствии с первой вся и формация, необходимая для построения выборки, собирается центром еще до того, как построена выборка для промежуточных объектов. Это означает, что после отбора областей и краев сведения о всех районах и городах этих областей, а также другие данные, относящиеся к более мелким территориальным подразделениям (сельсоветам, жилищным организациям и т. д.), должны быть сосредоточены в центре исследования.

Вторая модификация опирается на «волновой» подход. Необходимая информация : спрашивается центром только после того, как построена выбор ;а для более крупных промежуточных объектов. Так, информация о сельсоветах запрашивается только по районам, включенным в выборку. Вторая модификация требует меньших затрат на сбор информации, но при этом удлиняется стадия простояния выборки, тогда как при использовании первой модификации сбор информации можно осуществлять задолго до начала реализации проекта.

При втором, децентрализованном, варианте сбор информации и выборка для каждого из промежуточных объектов осуществляются па местах соответствующим организуемым там подразделением (например, краевой или областной группой, уполномоченным по району или городу).

Оба варианта предполагают определенный характер контактов между центром исследования и исполнителями на местах. Максимально централизованный подход требует, чтобы практически все решения исполнителей апробировались центром. При ином подходе центр требует, чтобы его информировали только о наиболее важных решениях.

Второй вневременной, определяющей характер организации выборочного обследования, является ориентация на разовое или повторное посещение отобранных семей. Ясно, что в зависимости от принятого решения видоизменяются многие стороны проекта исследования.

Третьей переменной следует считать характер контингента.

В зависимости от того, являются ли они штатными работниками социологических подразделений или привлекаются и работу на общественных началах, многие организационные проблемы также решаются по-разному. В этой связи ведущее значение имеет и вопрос о том, направляются ли из центра или же рекрутируются на местах и могут, проводит, опрос лишь недалеко от своего места жительства.

Четвертая переменная выделяет выборки, опирающиеся на контакт разового или многократного использования. Как уже отмечалось, в последние годы возросла тенденция к многократному использованию выборки и созданию на последней ступени нескольких вариантов выборки.

Определенные проблемы далеко не исчерпывают всего круга организационных проблем построения территориальной выборки. Причём в каждой конкретной исследовательской ситуации существуют особые способы их решения. Направлять поиск этих решений могут стратегии, разработанные теорией выборочного метода, и практикованные проектирования выборок, и, прежде всего схемы конструирования таких основных частей проекта выборки как. 1) отдельные последовательности промежуточных объектов , 2) стратификация единиц отбора", 3) использование того или иного способа отбора; 5) определение объема выборки.

Многоступенчатая территориальная выборка

Одним из важнейших стратегических решений при проектировании всех видов территориальной выборки — выбор первичных единиц отбора. Значение этого решения связано, прежде всего, в том, что именно оно в наибольшей степени (а не различные вариации последующих промежуточных объектов) предопределяет.

ОБЩАЯ ХАРАКТЕРИСТИКА КВОТНОЙ, МНОГОСТУПЕНЧАТОЙ ВЫБОРКИ

Выборка квотная - метод нестатистического формирования выборки, в основе которого лежит статистическая информация о генеральной совокупности. В.К. является частным случаем выборки стратифицированной.Генеральная совокупность разделяется на части по некоторым "контролируемым" показателям, объем выборки делится между выделенными частями пропорционально их объему - образуются квоты. Специфика В.К. состоит в том что, во-первых, расслоение обычно проводится одновременно по нескольким критериям и, во-вторых, в пределах сформированных квот интервьюер может выбирать своих респондентов более или менее произвольно. Чаще всего квоты формируются на основе социально-демографических показателей, таких как размер населенного пункта, пол, возраст, образование и т.п. В советский период, когда по организационным причинам предпочтение отдавалось производственным выборкам , квоты нередко формировались на основе профессиональных групп.

В социологической литературе имеются ссылки (Э. Ноэль, У. Кокрен и др.) на специальные методологические исследования, согласно которым качество данных, полученных по В.К., не значительно уступает качеству данных, собранных с использованием случайных выборок . Отмечается, что В.К. позволяют получить приемлемые результаты в исследованиях общественного мнения , ценностей, мотивов и т.п., однако их не рекомендуется использовать в исследованиях социальной стратификации и социальной мобильности .

В.К. весьма популярна среди социологов благодаря своей простоте и низкой стоимости реализации по сравнению со случайными выборками. Основным недостатком квотного отбора является принципиальная невозможность использовать формальные статистические средства для оценивания ошибки выборки и подтверждения репрезентативности собранных данных.

Иногда В.К. используется совместно с выборкой маршрутной . Тем самым отбор респондентов рандомизируется и репрезентативность исследования повышается.

Квоты применяются также в исследованиях со смешанными целями, когда надо, например, репрезентативно представить как генеральную совокупность в целом, так и составляющие ее социальные группы , в том числе малочисленные. Малочисленным группам присваивают завышенные квоты, которые обеспечивают репрезентативность полученных для них результатов. В этом случае при получении обобщенных характеристик всей выборки и генеральной совокупности, выборка должна быть взвешена для восстановления нарушенных пропорций. Если в таком исследовании респондентов отбирают с использованием случайных процедур, полученная выборка является не квотной, а случайной стратифицированной.

В социологической литературе имеются ссылки (Э. Ноэль, У. Кокрен и др.) на специальные методологические исследования, согласно которым качество данных, полученных по В.К., не значительно уступает качеству данных, собранных с использованием случайных выборок (Выборка случайная) . Отмечается, что В.К. позволяют получить приемлемые результаты в исследованиях общественного мнения , ценностей, мотивов и т.п., однако их не рекомендуется использовать в исследованиях социальной стратификации и социальной мобильности .

В.К. весьма популярна среди социологов благодаря своей простоте и низкой стоимости реализации по сравнению со случайными выборками. Основным недостатком квотного отбора является принципиальная невозможность использовать формальные статистические средства для оценивания ошибки выборки и подтверждения репрезентативности собранных данных. .

ВЫБОРКА КВОТНАЯ - микромодель объекта социологического исследования, формируемая на основе статистических сведений (параметров квот) преимущественного о социально-демографических

характеристиках элементов генеральной совокупности. Принцип выборки квотной, или же принцип отбора единиц наблюдения по методу квот (англ. **quota**), восходит к представлению о подобии объектов в случае пропорциональности их структурных элементов. Идея о правомерности экстраполяции результатов модели на моделируемый объект в случае подобия их структур была общепринята в статистической практике задолго до построения теории вероятностной выборки. В частности, применение подобного метода выборки для прогноза урожайности сельскохозяйственных культур предписывалось еще Петром I в "Регламенте или Уставе конюшенном".

В социологических исследованиях метод выборки квотной впервые стал применяться институтами опроса общественного мнения в начале XX в. Выборки квотной - органичный элемент построения модели социального эксперимента. Что касается опросов общественного мнения, здесь выборка квотная применяется наряду с вероятностными выборками, порой для взаимного контроля представительности результатов опроса. Метод квот удобен также для построения выборки в случае небольшой генеральной совокупности, либо в случае сильной "скошенности" распределения в ней элементов наблюдения.

Квотный метод выборки отличается от вероятностного тем, что предполагает предварительное наличие статистических сведений по ряду существенных либо коррелирующих с ними характеристик генеральной совокупности. Однако эти сведения не используются для определения объема выборки, т.к. в последующем отбор респондентов осуществляется не случайно, а целенаправленно, при помощи интервьюеров. Поэтому в случае применения выборки квотной ее величина определяется на основании сложившегося десятилетиями опыта и составляет от 1000 до 2500 единиц наблюдения, в зависимости от сложности структуры исследуемого объекта.

Общей проблемой как вероятностной выборки, так и выборки квотной являются затруднения, возникающие при выделении существенных

характеристик объекта исследования. До начала исследования статистические данные о них, как правило, отсутствуют, поэтому в качестве параметров квот приходится выбирать числовые значения, тесно коррелирующие с существенными (исследуемыми) контрольными признаками.

Число характеристик, данные о которых выбираются в качестве квот, как правило, не превышает четырех. При большем числе фиксированных признаков отбор респондентов становится чрезмерно трудоемким.

Квоты могут быть заданы как по независимым, так и по взаимосвязанным параметрам. Квота с независимыми параметрами есть не что иное, как статистические данные о значениях контрольных признаков, взятых каждый в отдельности.

Квоты со взаимосвязанными параметрами являются статистическими данными, полученными в результате группировки первичной информации по двум или нескольким признакам. Параметры квот в процентном выражении в точности воспроизводят структуру генеральной совокупности по контрольным признакам.

Число подлежащих опросу респондентов в соответствии с заданными квотами вычисляется путем умножения параметров квот на коэффициент $k = n/100$, *где* n *- объем выборочной совокупности. Слишком большое число параметров квот затрудняет работу интервьюеров, ведет к увеличению систематической ошибки. Поэтому в модели квотной выборки, как правило, опускают признаки, которые тесно коррелируют с какой-либо другой характеристикой, параметры которой также используются в качестве квот.*

Степень репрезентативности квотной выборки повышается прямо пропорционально степени устойчивости значений характеристик, по которым задаются квоты, в связи с чем признаки, изменяющие свои значения слишком быстро, в модели выборки квотной применяются весьма редко.

Теоретические ошибки для выборки квотной не вычисляются, в связи с чем ряд социологов сомневается в эффективности использования выборки квотной для исследований, требующих высокой точности данных.

Проверка эффективности выборки квотной обычно осуществляется при помощи ее сравнения с вероятностной выборкой. Из одной и той же совокупности генеральной (см.) извлекают две совокупности выборочные (см.) с тождественными объемами. Одна из выборочных совокупностей формируется вероятностным, другая квотным методом. Опрос проводится по обеим выборочным совокупностям, а результаты сравниваются между собой.

Метод квот позволяет существенно сократить время, затрачиваемое на опросы: интервьюер в случае задания ему параметров квот может осуществить интервью вдвое быстрее, чем при вероятностной выборке.

ВЫБОРКА МНОГОСТУПЕНЧАТАЯ - частный случай выборки кластерной , метод отбора, при котором на каждой, кроме последней, ступени построения выборки объекты группируются в некоторые структурные единицы (кластеры), среди которых и производится отбор. На последней ступени выборки кластеры, прошедшие все этапы отбора, обследуются полностью или выборочно.

При построении В.М. применяются термины "единица отбора" и "единица наблюдения". Единицами наблюдения являются объекты, составляющие генеральную совокупность (являющиеся ее элементами), часть из которых и должна быть, в конце концов, обследована, например, люди. Непосредственный отбор единиц наблюдения производится только на последней ступени В.М. На всех предшествующих ступенях производится отбор кластеров, которые объединяют некоторое количество единиц наблюдения, но сами являются только единицами отбора (в англоязычной литературе по отношению к человеческим генеральным совокупностям иногда применяется термин "выборочная точка".

В.М. рекомендуется применять:

1) если генеральная совокупность велика и имеет сложную многоуровневую структуру;

2) если средства на исследование ограничены;

3) если в разных частях генеральной совокупности целесообразно применять различные методы отбора.

Полное описание В.М. включает количество ступеней, критерии расслоения/кластеризации и методы отбора, применяемые на каждой ступени, например, В.М. часто используются в исследованиях населения. В частности, при национальном опросе общественного мнения на первом этапе может производиться расслоенный случайный отбор населенных пунктов (расслоение по статусу - столица/другие города /село, а городских населенных пунктов также по размеру - большие/средние/малые).

Причем на отбор сельских населенных пунктов может быть наложено дополнительное ограничение, например, они должны выбираться в тех же административных районах, где расположены попавшие в выборку города (это делается для сокращения расходов на исследование). В столице, больших и средних городах на второй ступени выборки может быть произведен простой случайный отбор улиц (домов), на третьем - отбор квартир (по маршрутной выборке), на четвертом - отбор респондента в семье. В малых городах и сельских населенных пунктах может проводиться случайный отбор респондентам по домовым книгам, хранящимся в городских и сельских Советах.

При построении В.М. применяются термины "единица отбора" и "единица наблюдения". Единицами наблюдения являются объекты, составляющие генеральную совокупность (являющиеся ее элементами), часть из которых и должна быть, в конце концов, обследована, например, люди. Непосредственный отбор единиц наблюдения производится только на последней ступени В.М. На всех предшествующих ступенях производится отбор кластеров, которые объединяют некоторое количество единиц

наблюдения, но сами являются только единицами отбора (в англоязычной литературе по отношению к человеческим генеральным совокупностям иногда применяется термин "выборочная точка" - sampling point).

В.М. рекомендуется применять:

1) если генеральная совокупность велика и имеет сложную многоуровневую структуру;

2) если средства на исследование ограничены;

3) если в разных частях генеральной совокупности целесообразно применять различные методы отбора.

Полное описание В.М. включает количество ступеней, критерии расслоения/кластеризации и методы отбора, применяемые на каждой ступени, например, В.М. часто используются в исследованиях населения. В частности, при национальном опросе общественного мнения на первом этапе может производиться расслоенный случайный отбор населенных пунктов (расслоение по статусу - столица/другие города /село, а городских населенных пунктов также по размеру - большие/средние/малые). Причем на отбор сельских населенных пунктов может быть наложено дополнительное ограничение, например, они должны выбираться в тех же административных районах, где расположены попавшие в выборку города (это делается для сокращения расходов на исследование). В столице, больших и средних городах на второй ступени выборки может быть произведен простой случайный отбор улиц (домов), на третьем - отбор квартир (по маршрутной выборке), на четвертом - отбор респондента в семье. В малых городах и сельских населенных пунктах может проводиться случайный отбор респондентам по домовым книгам, хранящимся в городских и сельских Советах.

ВЫБОРКА КЛАСТЕРНАЯ (ГНЕЗДОВАЯ) - метод извлечения выборки , основанный на предварительном разделении генеральной совокупности , на относительно компактные структурные части (кластеры, гнезда). Главным требованием является более широкая вариация основных

изучаемых показателей внутри кластера по сравнению с их вариацией между кластерами [в отличие от выборки стратифицированной ,цель которой - выделение страт, в которых вариация основных показателей была бы минимальной].

В.К. может осуществляться в несколько этапов. На первом этапе основа выборки представляет собой полный список кластеров; из этого списка тем или иным способом извлекается выборка кластеров. Далее в исследовании участвуют только выбранные кластеры. Они, в свою очередь, могут обследоваться полностью или выборочно. Если отобранные кластеры обследуются полностью, мы получаем выборку серийную . Примером серийной выборки является опрос студентов целыми академическими группами, когда на первом этапе отбираются группы, а на втором - опрашиваются все студенты из отобранных групп. Если отобранные кластеры обследуются выборочно, выборка является двухступенчатой или многоступенчатой. Примером двухступенчатой выборки является исследование населения, на первом этапе которого в качестве кластеров отбираются отдельные населенные пункты; на втором - в каждом из отобранных населенных пунктов извлекается простая случайная выборка его жителей (в качестве основы выборки могут использоваться, например, данные адресного стола). Для многоступенчатой выборки процесс кластеризации может продолжаться "вглубь", посредством разделения отобранных кластеров на все более мелкие.

Если на всех этапах В.К. применяются процедуры случайного (простого или стратифицированного) отбора либо сплошное обследование всех объектов, составляющих кластер, выборка называется случайной кластерной. Существуют две основные стратегии, обеспечивающие при кластерном отборе всем элементам из генеральной совокупности одинаковую вероятность попадания в выборку. Согласно первой из них, кластеры отбирают из списка с равными вероятностями; затем предполагаемый объем выборки делят между выбранными кластерами пропорционально их размеру.

При использовании второй стратегии, кластеры отбирают с вероятностями, пропорциональными их размеру, предполагаемый объем выборки делится между отобранными кластерами в равных частях.

Главным преимуществом В.К. является минимизация трудовых и финансовых затрат; недостатком - более высокая ошибка выборки по сравнению с простым или стратифицированным случайным отбором. Ошибка выборки тем выше, чем больше средний размер кластера и, соответственно, меньше число обследуемых кластеров. Кластерный отбор рекомендуется применять:

1) если основа выборки не может быть получена для всей генеральной совокупности, но может быть получена для ее отдельных структурных частей;

2) если простой или расслоенный случайный отбор становится неприемлемо дорогостоящим, например, из-за того, что генеральная совокупность рассеяна по слишком большой территории.

В некоторых случаях отбор кластеров производится нестатистическими методами. Например, целевым образом выбираются населенные пункты, в которых проживают интервьюеры. В этом случае выборка не будет случайной (даже если в населенных пунктах производится случайный отбор), и статистическое оценивание ошибки выборки не будет корректным.

МЕТОДЫ КЛАСТЕРНОГО АНАЛИЗА

Методы, представленные в модуле *Обобщенные методы кластерного анализа* программы схожи с алгоритмом *k*-средних, включенным в стандартные настройки модуля *Кластерный анализ*, и вы можете просмотреть раздел Кластеризация *k*-средних для основного обзора этих методов и их приложений. Назначение этих методов в основном определять кластеры в наблюдениях (или переменных), и для назначения этих наблюдений кластерам.

Типичный пример приложения такого типа анализа - маркетинговые исследования, в которых число связанных переменных поведения потребителя измеряется для больших выборок респондентов; цель изучения - определить "сегмент рынка", т.е. групп респондентов, каким-нибудь образом схожих друг с другом (для всех членов одного кластера) в сравнении с респондентами, которые "принадлежат к" другим кластерам. Вместе с идентификацией таких кластеров, представляет интерес определение различий между этими кластерами, т.е. специфики переменных или измерений, которыми различаются члены кластеров, и как.

Кластеризация k-средними. Классический алгоритм кластеризации *k*-средними стал общеизвестным благодаря Hartigan`y (1975; см. также Hartigan и Wong, 1978). Основная операция этого алгоритма относительно проста: заданное фиксированное число (желательное или гипотетическое) k кластеров, наблюдения сопоставляются кластерам так, что средние в кластере (для всех переменных) максимально возможно отличаются друг от друга. (Смотрите ниже Различия между алгоритмами *k*-средних в обобщенных методах кластерного анализа и кластерным анализом за дополнительной информацией касательно специфики вычислений алгоритмов, используемых в этих двух модулях .)

Расширения и обобщения. Методы, представленные в модуле *Обобщенные методы кластерного анализа* , расширяют эти основные приближения кластеризации тремя важными методами:

1. Вместо того чтобы задавать соответствие наблюдений кластерам так, чтобы максимизировать разницу в средних для непрерывных переменных, алгоритм кластеризации EM (поиск максимума) вычисляет вероятности членства в кластере, основываясь на одном или более вероятностном распределении. Цель алгоритма кластеризации - максимизировать вероятность полного правдоподобия данных, задаваемых в (последних) кластерах.

2. В отличие от классической реализации алгоритма кластеризации *k*-средними в модуле *Кластерный анализ*, алгоритмы *k*-средних и EM в модуле *Обобщенные методы кластерного анализа* могут быть применены равно для непрерывных и категориальных переменных.
3. Основное отличие алгоритма кластеризации *k*-средними в том, что вы должны указать число кластеров перед началом анализа (то есть, число кластеров должно быть *априори* известно); модуль *Обобщенные методы кластерного анализа* использует измененную схему v-кратной кросс-проверки (схожую с реализованной в модулях Деревья классификации, Общие модели деревьев классификации и регрессии, и *Общие CHAID*) для определения наилучшего числа кластеров по данным. Это расширение делает модуль *Обобщенные методы кластерного анализа* весьма полезным инструментом добычи данных для неконтролируемого обучения и распознавания образов.

Полный обзор различных методов кластеризации, в контексте добычи данных представлен в Witten и Frank (2000). Имеются также разделы модуля: Нейронные сети, Самоорганизующаяся карта Кохонена (СОКК) или сети *Кохонена*; эти архитектуры нейронных сетей могут применяться для схожих типов проблем, таких как методы, описанные в этом разделе. Однако, методы *EM кластеризации и кластеризации k-средними* реализованные в этом модуле обычно быстрее и легче масштабируются на очень большие множества данных и аналитических проблем.

НЕКОНТРОЛИРУЕМОЕ ОБУЧЕНИЕ ПРОТИВ КОНТРОЛИРУЕМОГО ОБУЧЕНИЯ ОБЩАЯ КЛАССИФИКАЦИЯ КЛАСТЕРНОЙ ВЫБОРКИ

Важное различие в машинном обучении, также применимое к добыче данных, между контролируемым и неконтролируемым алгоритмами обучения. Раздел "контролируемое" обучение обычно применим в случаях,

где текущая классификация уже выяснена и сохранена в тестовой выборке, и вы хотите построить модель для прогнозирования этой классификации (в новой тестовой выборке). Например, у вас может быть множество данных, содержащее информацию о том, кто из списка клиентов, направленных на специальное поощрение примет или не примет это предложение; цель классификационного анализа - построить модель для прогнозирования того, кто (из различных списков потенциальных клиентов) вероятно, ответит на такое же (или схожее) предложение в будущем. Можно просмотреть также описание методов к разделам *Общие модели деревьев классификации и регрессии (GCRT)*, Общие CHAID модели (GCHAID), *Анализ дискриминантных функций* и *Общие модели дискриминантного анализа (GDA)*, и Нейронные сети (см. Справка к системе) для изучения различных методов, используемых для построения или подгонки моделей по данным, где наблюдаются итоговые переменные (например, клиент ответил или не ответил на предложение). Эти методы называются алгоритмы контролируемого обучения, так как обучение (подгонка моделей) "управляется" или "контролируется" наблюденными классификациями, записанными в файле данных.

При неконтролируемом обучении ситуация другая. Здесь итоговые переменные непосредственно не наблюдаются (не могут наблюдаться) Взамен мы хотим выявить некоторую "структуру" или кластеры данных, которые заведомо не могут наблюдаться. Например, у вас может иметься база данных на клиентов с различными демографическими индикаторами и переменными, потенциально важными для поведения потребителя. Ваша цель - найти сегменты рынка, то есть группы наблюдений, сравнительно похожих друг на друга по некоторым переменным; задав однажды, вы можете затем определять, насколько хорошо достигается один или несколько кластеров при предоставлении определенных товаров или услуг, которые, по вашему мнению, могут иметь особую полезность или индивидуальную привлекательность в сегменте (кластере). Такой тип заданий называется

алгоритмом неконтролируемого обучения, так как обучение (подгонка моделей) в этом случае не может управляться уже известной классификацией. Только после определения известных кластеров вы сможете начать задавать метки, например, основываясь на последующих наблюдениях (например, после определения одной группы клиентов как "молодые опасные воры"). Другие методы (отличные от кластеризации *k*-средними или EM), которые попадут в категорию алгоритмов неконтролируемого обучения: *Факторный анализ*, *Главные компоненты и результаты анализа классификаций*, *Многомерное шкалирование*, Анализ соответствий, *Нейронные сети - Самоорганизующаяся карта Кохонена (СОКК)* и т.д.

Алгоритм *k*-средних

Классический алгоритм *k*-средних описан в деталях в разделе *Кластерный анализ*; полный вводный курс и обзор можно найти в Hartigan и Wong (1978). Для повторения, базовый алгоритм кластеризации *k*-средними достаточно прост: Дано заданное пользователем фиксированное число кластеров k, перемещайте наблюдения в кластере для максимизации расстояния между центрами кластеров; центры кластеров обычно определяются вектором средних значений для всех (непрерывных) переменных в анализе.

Кластеризация категориальных переменных. Модуль *Кластерный анализ* включает реализацию классического алгоритма *k*-средних, который типично применим только к непрерывным переменным. В модуле *Обобщенные методы кластерного анализа* вы можете задать категориальные переменные для анализа. Вместо определения центра кластера для текущего кластера и переменной с помощью среднего соответствующей (непрерывной) переменной (для наблюдения в том же кластере), для категориальной переменной определяется одиночный класс (значение категориальной переменной), которому принадлежит большинство наблюдений этого кластера. Например, если текущий кластер анализа, включающий

переменную *Пол*, содержит больше (>50%) мужчин, тогда центральное значение для этого кластера будет установлено *Мужской*.

Мера расстояний. Реализация алгоритма *k*-средних в модуле *Кластерный анализ* всегда будет вычислять кластерное расстояние, базируясь на простом (квадратичном) евклидовом расстоянии между кластерными центроидами (вектор значений для непрерывных переменных в анализе). В модуле *Обобщенные методы кластерного анализа*, у вас есть выбор различных мер расстояний для использования в анализе: евклидово, Квадрат евклидова, Манхэттенское и Чебышева. Эти различные меры расстояний всегда вычисляются из нормализованных расстояний; смотрите также Различия между алгоритмами k-средних в обобщенных методах кластерного анализа и кластерным анализом (ниже). Отметьте, что для категориальных переменных, все расстояния могут быть только 0 (ноль) или 1 (один): 0, если класс, которому принадлежит соответствующее наблюдение, принадлежит в тоже время к классу, в котором встречается лучшая частота соответствующего кластера (см. предыдущий параграф), и 1, если он отличает от такого класса. Следовательно, за исключением расстояния Чебышева, различные меры расстояния для категориальных переменных, доступные в программе, приведут к идентичным результатам.

EM алгоритм

EM алгоритм кластеризации детально разобран в Witten и Frank (2001). Базовое приближение и логика этого кластерного метода в следующем: Пусть вы измеряете одиночную непрерывную переменную в большой выборке наблюдений. Дальше, предположите, что выборка состоит из двух кластеров наблюдений с различными средними (и возможно с различным стандартным отклонением); в рамках каждой выборки, распределение значений для непрерывной переменной соответствует нормальному распределению. Итоговое распределение значений (в совокупности) может выглядеть так:

Смешение распределений. Рисунок показывает два нормальных распределения с различными средними и различным стандартным отклонением и сумму двух распределений. Только смесь (сумма) двух нормальных распределений (различными средними и различным стандартным отклонением) должна быть выведена. Цель EM кластеризации - вычислить средние и стандартное отклонение для каждого кластера, так что правдоподобие наблюдаемых данных (распределения) максимально. С другой стороны, EM алгоритм пытается приблизить наблюдаемые распределения значений, основываясь на смеси различных распределений в различных кластерах.

Реализация EM алгоритма в модуле *Обобщенные методы кластерного анализа* позволяет вам выбирать (для непрерывных переменных) распределение: Нормальное, Логнормальное, и Пуассоновское. Вы можете выбрать различные распределения для различных переменных, и, таким образом, получить кластеры для смеси различных типов распределений.

Категориальные переменные. Реализация EM алгоритма в также может обрабатывать категориальные переменные. Программа сперва случайно задаст различные вероятности (точнее, веса) для каждого класса или категории каждого кластера; в последующих итерациях эти вероятности улучшаются (подгоняются) к максимальному правдоподобию данных, давая указанное число кластеров.

Классификационные вероятности вместо классификаций. Результаты EM кластеризации отличаются от таких же, вычисленных методом кластеризации *k*-средних: Позднее будут заданы наблюдения кластеров для максимизации расстояния между кластерами. EM алгоритм не вычисляет фактического назначения наблюдений кластерам, но вычисляет вероятности классификации. Другими словами, каждое наблюдение принадлежит каждому кластеру с определенной вероятностью.

Поиск верного числа кластеров: V-кратная кросс-проверка

Методы кластеризации, доступные в модуле *Обобщенные методы кластерного анализа*, специально оптимизированы и усовершенствованы для типичных приложений в добыче данных. Основное сравнение добычи данных - ситуация аналитического поиска полезных структур и "самородков" в данных, обычно без *априори* устойчивых ожиданий того, что можно найти (контраст с гипотетически-тестовым приближением научного исследования). На практике аналитик обычно не знает наперед, сколько кластеров может быть в выборке. По этой причине, программа включает реализацию алгоритма V-кратной кросс-проверки для автоматического определения числа кластеров по данным.

Этот уникальный алгоритм весьма полезен во всех основных задачах "обучения распознавания". Для определения числа сегментов рынка в маркетинговых исследованиях, числа моделей индивидуальных затрат на изучение потребительского поведения, числа кластеров различных медицинских симптомов, числа различных типов (кластеров) документов в текстовой добыче, числа погодных моделей в метеорологических исследованиях, числа моделей отбраковки кремниевых вафель и т.д.

Алгоритм v-кратной кросс-проверки в приложении к кластеризации. Алгоритм v-кратной кросс-проверки детально описывается в контексте модулей Деревья классификации, Общие модели деревьев классификации и регрессии (GCRT), и *Общие CHAID* (см. Справка к системе). Основная идея этого метода - разделить выборку на v частей или случайно вытащить (нарушив структуру) подвыборки. Затем несколько типов анализов последовательно наложатся на наблюдения, принадлежащие к v-1 частям (обучающая выборка) и результаты анализов наложатся на выборку v (выборка или часть, не использующаяся при вычислении параметров, построения дерева, определения кластеров, и т.д.; то есть это - тестовая выборка) для вычисления индексов предсказательной точности. Результаты для v ответов собраны (усреднены) для одиночной выборки стабильности

соответствующей модели, то есть обоснованности модели для прогнозирования нового наблюдения.

Как упомянуто ранее, кластерный анализ - метод неконтролируемого обучения, и мы не можем наблюдать (реальное) число кластеров по данным. Однако разумно заменить понятие (применимое к контролируемому обучению) "соответствие" на "расстояние": В общем, мы можем использовать метод V-кратной кросс-проверки для упорядочивания чисел кластеров, и наблюдать результаты среднего расстояния от наблюдений (в кросс-проверке тестовых выборок) до центров их кластеров (для кластеризации *k*-средними); для EM кластеризации, подходящим эквивалентом меры может стать среднее значение отрицательного (лог-) правдоподобия, вычисленного для наблюдений в тестовой выборке.

Просмотр результатов v-кратной кросс-проверки. Результаты v-кратной кросс-проверки лучше всего просматриваются на простой линии графика.

Здесь показаны результаты анализа множества широко известных данных, содержащих три кластера наблюдений (особенно, популярен файл данных *Iris (Ирис)* описанный Fisher, 1936, на который много ссылаются в литературе по дискриминантному анализу). Также показаны (в правом верхнем углу графика) результаты анализа нормальнораспределенных случайных чисел. "Реальные" данные (показанные слева) выражают характеристики шаблона график осыпи, где функция стоимости (в этом наблюдении, 2 раза лог-правдоподобие кросс-проверки данных дают вычисляемые параметры) быстро снижается в то время как число кластеров растет, но затем (после 3 кластеров) выравнивается, и даже растет, пока данные переподгоняются. С другой стороны, случайные числа показывают, что такой схемы не должно быть, на самом деле, существенного понижения функция стоимости вовсе нет, она быстро начинает расти вместе с ростом числа кластеров и процессом переподгонки.

По этому рисунку легко видеть, насколько полезна схема v-кратной кросс-проверки, применимо к кластеризации *k*-средними и EM кластеризации при определении "верного" числа кластеров данных.

Реализация методов кластеризации в модуле *Обобщенные методы кластерного анализа* сильно расширяема, и эти методы могут быть использована даже для очень большого множества данных.

ВЫБОРКА ТЕОРЕТИЧЕСКАЯ

Метод формирования выборочной совокупности для исследований случая, применяется также в формировании фокус-групп и планировании экспериментов с выделяемыми факторами. В противоположность выборке случайной , репрезентативность В.Т. обосновывается не равной вероятностью попадания в выборку для всех элементов генеральной совокупности , но тщательным отбором единичных случаев, соответствующих заданным критериям. Е.М. Ковалев и И.Е. Штейнберг ("Качественные методы в полевых социологических исследованиях", 1999) приводят следующие модели В.Т.:

1) Выборка экстремальных (девиантных) случаев: отбор необычных, в некотором смысле специфических, случаев. Предполагается, что такие случаи могут в сжатом виде содержать всю информацию о более "типичных" представителях генеральной совокупности.

2) Интенсивная выборка: отбор информативно значимых случаев, которые в значительной (но не экстремальной) степени представляют интересующее социолога явление. Предполагается, что собрана предварительная информация о генеральной совокупности и проведен предварительный анализ ее отличительных особенностей.

3) Выборка максимальной вариации: отбор случаев, представляющих все распространенные модели изучаемого явления. Для конструирования

выборки необходимо предварительно выделить соответствующие модели и оценить их распространенность.

4) Гомогенная выборка: отбор случаев, с максимальной полнотой характеризующих некоторую относительно гомогенную часть генеральной совокупности. В частности, интервью с фокус-группами обычно проводятся на гомогенных выборках, насчитывающих от 5 до 8 участников с одинаковыми демографическими и социальными характеристиками.

5) Выборка типичных случаев: типичные случаи выбираются в ходе бесед с экспертами, определяющими, что, на их взгляд, является типичным, а также на основе проведенных ранее опросов, демографическом анализе и т.п. Используется как индуктивное, так и дедуктивное понимание типичного. В индуктивном ("статистическом") понимании типичное представляется как наиболее часто встречающееся (модальное). Дедуктивное понимание типичного восходит к веберовскому пониманию типа как синтеза представлений об идеальном. Дедуктивное понимание типа более информативно, но намного труднее в реализации.

6) Стратифицированная выборка: отбор случаев из предварительно выделенных страт (слоев, частей) генеральной совокупности. Целью стратифицированной В.Т. является не описание генеральной совокупности в целом (как в случае стратифицированной случайной выборки), но фиксация основных различий между ее объектами. Статистической репрезентативностью стратифицированная В.Т. не обладает.

7) Выборка критических случаев: отбор случаев, критических важных для понимания происходящего ("Если так случилось здесь, то это произойдет везде" или "Если здесь этого не случилось, то и нигде не случится"). Эта модель особенно эффективна, если имеющиеся ресурсы ограничивают исследование одним случаем. Например, если правительство вводит новые правила налогообложения, то на первом этапе достаточно проверить, как их понимает наиболее образованная часть общества. Если даже для нее правила непонятны, они тем более будут непонятны всем остальным. Наоборот, если

правила понятны наименее образованным гражданам, люди с более высоким уровнем образования их тоже поймут.

8) Критериальная выборка состоит в том, что изучению подвергаются только объекты, удовлетворяющие заранее определенным критериям. Например, "пациенты психиатрической больницы, неоднократно пытавшиеся совершить самоубийство". Такая выборка эффективна при изучении проблемных случаев. Основой для критериальной выборки могут служить результаты количественного анализа или тестирования.

Объем В.Т. зависит от целей исследования, его глубины, имеющихся ресурсов, т.к. более глубокое исследование каждого случая требует больше средств и времени. Качественная выборка должна одновременно удовлетворять двум часто противоречащим друг другу критериям - она, с одной стороны, должна быть компактной, с другой - покрывать цели исследования. Критерий максимизации информации требует, чтобы формирование выборки прекращалось только в тот момент, когда от включения в выборку новых случаев уже не ожидается получение дополнительной информации (ср.: в "количественных" исследованиях объем выборки планируют заранее). Это, в частности, означает, что должна сохраняться возможность увеличить выборку в зависимости от первых результатов исследования. В условиях ограниченных ресурсов этот идеал может оказаться недостижимым.

Теоретическая подготовка исследовательской программы Каждый из нас в той или иной степени соприкасается с эмпирическими социологическими исследованиями в качестве слушателя радио, читателя газет, журналов, научной литературы и т. д. Возможно, и сам бывает вовлечен в эти исследования в качестве респондента, т. е. источника первичной информации об изучаемых процессах и явлениях. Не исключена вероятность того, дто кому-либо из нас придется организовывать такое исследование Задачей данной темы является дать общее представление о методологии и методике эмпирического социологического исследования,

познакомить с основными понятиями и процедурами. В эмпирическом социологическом исследовании можно выделить три основных этапа, каждый из которых включает в себя ряд важных процедур: 1) подготовительный (разработка программы исследования); 2) основной (проведение эмпирического исследования); 3) завершающий (обработка и анализ данных, формирование выводов и рекомендаций). Всякое исследование начинается с постановки какой-либо проблемы. Проблема исследования может быть задана извне каким- ' либо заказчиком или обусловлена познавательным интересом. В-раскрытии-темымы будем ориентироваться на прикладное социологическое исследование, которое может быть проведено в какой-либо производственной организации. Поэтому мы будем исходить из тех или иных реальных проблем, которые встают в настоящее время перед производственными организациями. Проблема — это всегда противоречие между знаниями о потребностях людей в каких-то результативных практических или теоретических действиях и незнанием путей и средств их реализации. Решить проблему — значит получить новое знание или создать теоретическую модель, объясняющую то или иное явление, выявить факторы, позволяющие воздействовать на развитие явления в желаемом направлении. Заказ социологу чаще всего формируется в виде обозначения некоторой проблемной ситуации, указания на какое-то социальное противоречие либо просто указания на неудовлетворительное состояние дел в той или иной сфере производства, управления и т. д. Социологу предстоит перевести проблемную ситуацию в формулировку проблемы, которую он будет исследовать. Для этого он должен проделать специальную теоретическую работу:
1) установить реальное наличие данной проблемы: а) есть ли показатель, количественно или качественно характеризующий данную проблему; б) есть ли учет и статистика по этому показателю, в) достоверны ли учет и статистика по этому показателю;

2) вычленить наиболее существенные элементы или факторы проблемы, решение которых принадлежит социологии, а не экономической теории, технологии производства и т. д. Например, поставлейа проблема разобраться в причинах низкой эффективности управления тем или иным подразделением предприятия. Социологу предстоит решить, какие социальные группы и личности участвуют в возникновении и решении этой проблемы, как влияют здесь их интересы, как стимулируется их участие в разрешении данной,проблемы и т. д.;

3) вычленить уже известные элементы проблемной ситуации, которые не требуют специального анализа и выступают как информационная база для рассмотрения неизвестных элементов (например, данные статистики и учета представляют собой готовый важный материал);

4) выделить в проблемной ситуации главные и второстепенные компоненты, чтобы определить основное направление исследовательского поиска;

5) проанализировать уже имеющиеся решения аналогичных проблем. С этой целью необходимо изучить всю литературу по данному вопросу. Провести беседы с компетентными людьми — экспертами. В роли экспертов обычно выступают специалисты — ученые или опытные практики. Производственная проблема может быть описана с помощью пяти основных характеристик:

1) Сущность или содержание. Например, низкая эффективность производства, высокая социально-психологоческая напряженность в трудовом коллективе и т. д. При определении проблемы следует устанЬвить, с чем все это сравнивается и на каком основании. В данных примерах следует ответить на вопрос: почему мы считаем, что эффективность производства низкая, а социальная напряженность высокая? Низкая и высокая по сравнению с какими стандартами?

2) Организационное и физическое нахождение. В каком организационном подразделении (участках, отделах, филиалах) и физических объектах (заводы, здания, склады, конторы) была выявлена проблема. Насколько

широко она распространена в организации. Какие подразделения она затронула.

3) Владение проблемой. Является ли проблема «открытой» (знакомойвсем) или «закрытой» (то есть известной группе лиц)? Какие люди (управленцы, специалисты, рабочие и т. д.) затронуты проблемой и более всего заинтересованы в ее решении?

4) Абсолютная и относительная величина. Насколько важна проблема в абсолютных величинах? Например, объем потерянного рабочего времени или денег; объем неиспользуемых производственных мощностей. Насколько она важна в относительном выражении? Как она влияет на подразделения, в которых она обнаружена, и на людей, которые владеют ею? Насколько она важна для организации в целом? Что может получить организация от ее решения?

5) Временные рамки. С какого времени существует данная проблема? Наблюдалась ли она один раз, несколько раз или возникает периодически? Какова тенденция: проблема стабилизировалась, усиливается или ослабляется?В результате предварительного анализа проблемная ситуация получает четкое выражение в виде формулировки проблемы. Причем эта формулировка может значительно отличаться от первоначальной, сформулированной заказчиком.

На основе предварительного анализа разрабатывается программа исследования данной проблемы. Программа является обязательным исходным документом любого социологического исследования, независимо от того, является ли это исследование теоретическим или прикладным. Программа социологического исследования обычно включает в себя следующие разделы: 1) теоретический (цели, задачи, предмет и объект исследования, определение понятий); 2) методический (обоснование выборки, обоснование методов сбора данных, методы обработки и анализа данных; 3) организационный (план исследования, порядок исследования подразделений, распределение людских и финансовых ресурсов и т. д.).

Основные элементы программы прикладного социологического исследования показаны на рисунке.

Цели и задачи исследования. Этот раздел программы регулирует отношения заказчика и социолога на стадии предварительного определения ожидаемого результата, а также определяет объем затрат, времени и финансовых ресурсов, необходимых для получения результата. Цели исследования могут быть различны. Например, если сформулирована проблема как недостаточно высокий уровень управления подразделениями организации, то цель будет состоять в анализе реальной ситуации причин низкой эффективности управления организацией, выявлении скрытых резервов и разработке практических рекомендаций по изменению этой ситуации. Задачи исследования представляют собой содержательную, методическую и организационную конкретизацию цели. Предмет и объект исследования. Предмет исследования — это центральный вопрос проблемы. В одной и той же проблемной ситуации, в одном и том же эмпирическом объекте могут выделяться различные его аспекты, которые являются предметом исследования. Иначе говоря, когда социолог выбирает предмет исследования, он в то же время формулирует и гипотезу о возможном пути решения проблемы, а также методы и формы проведения социологического исследования. Так, в обозначенном нами примере исследования социолог может предположить, что причиной является неэффективная система принятия решения, тогда предметом исследования может служить система принятия решений и это может стимулировать: 1) исследование путей принятия решений; 2) роль коллективных органов в подготовке и принятии решений; 3) роль штатных специалистов и линейных руководителей в принятии решений; 4) решающая и совещательйая роль лиц, обладающих неофициальным влиянием, ответственность за решения, их внедрение и контроль за внедрением. Но социолог может предположить, что основная причина низкой эффективности управления заключается в стиле руководства. Тогда исследование будет развиваться по другому сценарию.

Если в первом случае большое значение будет иметь анализ документов, то во втором случае — анкетный опрос и психологическое тестирование. Теоретическая и эмпирическая интерпретация понятия — необходимый этап в разработке методологии исследования. Он позволяет решить три основные задачи: 1) Выяснить те аспекты теоретических понятий, которые используются в данном исследовании. 2) Это дает возможность вести анализ практических проблем на уровне теоретического знания и тем самым обеспечивать научное обоснование его результатов, выводов и рекомендаций. 3) Обеспечить измерение и регистрацию изучаемых явлений с помощью количественных, статистических показателей. Теоретическая интерпретация понятий осуществляется через ряд последовательных этапов. На первом этапе осуществляется перевод проблемной ситуации в формулировку в строгих научных рамках и терминах. На следующем этапе каждое понятие этой формулировки раскладывается на такие операционные составляющие, которые затем могут быть исследованы количественным методом. Кроме структурной интерпретации понятий, описывающих предмет исследования, необходимо провести их факторную интерпретацию, то есть определить систему его связей с внешними объектами и внутренними субъективными факторами. Конечной целью всей этой работы является выработка таких понятий, которые доступны учету и регистрации. Понятия, обозначающие такие элементарные фрагменты социальной реальности, называются понятиями-индикаторами. При этом социолог должен стремиться обеспечить максимальное описание изучаемого предмета в понятиях-индикаторах. Формирование гипотезы — заключительная часть теоретической подготовки эмпирического социологического исследования. Гипотеза исследования — это научно обоснованное предположение о структуре изучаемого социального явления или о характере связей между его компонентами. Гипотезы вырабатываются на основе имеющихся фактов. В науке существуют определенные правила выдвижения и проверки гипотез: 1)

Гипотеза должна находиться в согласии или, по крайней мере, быть совместимой со всеми фактами, которых она касается. 2) Из многих противостоящих друг другу гипотез, выдвинутых для объяснения серии фактов, предпочтительнее та, которая единообразно объясняет большее их число. 3) Для объяснения связанной серии фактов нужно выдвигать возможно меньше гипотез, и их связь должна быть возможно более тесной. 4) При выдвижении гипотез необходимо сознавать вероятностный характер ее выводов. 5) Невозможно руководствоваться противоречащими друг другу гипотезами.

Гипотезы — это отправные точки для исследования, далы ей-пвде этапы эмпирического социологического исследования нахо. -ят-ся в прямой зависимости от выдвинутых гипотез. Для отработки гипотезы и процедур исследованиянередко проводят предварительное, пилотажное исследование. В зависимости от теоретического уровня интерпретируемых понятий гипотезы делятся на основные и выводные (гипотезы причины и гипотезы следствия). Таким образом, они образуют иерархические цепочки, дублирующие теоретическую интерпретацию понятий. Следует подчеркнуть, что формирование гипотез — это не праздные теоретические упражнения, а разработка логических опор для сбора и анализа эмпирических данных. Если исследователем были сформулированы гипотезы, то эмпирические данные служат для их проверки, подтверждения или опровержения. Если же гипотезы с самого начала не выдвигались, то резко падает научный уровень социологического исследования, а его результаты и обобщения сводятся к описаниям процентных выражений тех или иных индикаторов и к довольно тривиальным рекомендациям.

Методы сбора социальной информации (выборка, анализ документов, наблюдение, опрос: анкетирование, интервьюирование).

Наряду с теоретическим большое значение в исследовании имеет методический раздел программ, который включает в себя описание методики и организации исследования. Центральное значение в этом разделе занимает

обоснование выборки. Характер решаемой проблемы, цели и задачи исследования определяют, каким должен быть объект исследования. Иногда, когда объект исследования сравнительно невелик и социолог располагает достоверными силами и возможностями его изучить, он может исследовать его целиком. Тогда, говорят социологи, объект исследования тождествен генеральной совокупности. Но часто сложное исследование невозможно или в нем нет необходимости. Поэтому для решения задач исследования осуществляется выборка.

В программе должно быть четко указано: 1) Каков объект эмпирического исследования. 2) Является исследование сплошным или выборочным. 3) Если оно является выборочным, то претендует ли оно на репрезентативность. Репрезентативность — это свойство выборочной совокупности воспроизводить параметры и значительные элементы генеральной совокупности. Генеральная совокупность — это совокупность всех возможных социальных объектов, которая подлежит изучению в пределах программы социологического исследования. Вторичная совокупность (выборка) — это часть объектов генеральной совокупности, отобраннац с помощью специальных приемов для получения информации о всей совокупности в целом. Число единиц наблюдения, составляющих выборочную совокупность, называется ее объемом (объемом выборки). Существует ряд процедур осуществления выборки. 4) Исследователь обязан указать, сколько ступеней отбора применяется в выборке, какова единица отбора на каждой ступени и какой темп отбора применяется на каждой ступени. 5) Что является основой выборки (список, картотека, карта)? 6) Какова единица наблюдения на последней ступени выборки. Попробуем на характерном примере описать выборку. Возьмем исследование эффективности труда на малых предприятиях, существующих в системе крупных государственных предприятий.

ВЫБОРКИ ОШИБКА

Разность между средними значениями переменной по выборке (x) и по генеральной совокупности (μ): $\Delta = x - \mu$. Различают две составляющие В.О. - систематическую и случайную.

Систематическая ошибка порождается ошибками планирования выборочного исследования, такими как неправильное определение генеральной совокупности или основы выборки, неудачный выбор метода извлечения выборки, ошибки в реализации выборочных процедур. Например, опрос аудитории через СМИ неизбежно приводит к систематическим ошибкам, вызванным различиями между той частью аудитории, которая принимает участие в опросе, и той ее частью, которая уклоняется от участия. Систематическая ошибка не уменьшается с увеличением объема выборки, она не может быть оценена статистически на основании данных исследования. Систематическая ошибка может быть обнаружена, когда известно (или со временем становится известным) распределение признака по генеральной совокупности, либо когда данные исследования с очевидностью противоречат имеющимся фактам и социальной теории (что, конечно, не исключает возможности наличия ошибок как в "фактах", так и в теории, либо в определении области ее приложения).

Случайная В.О. неизбежно возникает в выборочном исследовании как следствие применения выборочных процедур. При применении процедур случайного отбора она уменьшается с увеличением объема выборки и может контролироваться средствами статистики. Контроль случайной ошибки означает, во-первых, возможность определения ее величины с заданной доверительной вероятностью и, во-вторых, возможность ее уменьшения до некоторого допустимого значения посредством увеличения объема выборки.

Ошибка выборки-(англ. sample error) - значение ошибки при проведении выборочной аудиторской проверки. **Ошибка выборки** возникает, если при проверке обнаруживается то, что в действительности отсутствует (например, в процессе наблюдения экономических явлений, при

решении задач поиска). При определении объема выборки устанавливается риск выборки, допустимая и ожидаемая ошибки. Риск выборки заключается в том, что мнения исследователя по определенному вопросу, составленные на основе выборочных данных и на основании изучения генеральной совокупности, могут различаться. Риск выборки имеет место как при тестировании средств системы контроля, так и при проведении детальной проверки. В контексте проверки на соответствие термин «ошибка» означает, что результаты проверки свидетельствуют либо о неправильном функционировании контроля внутреннего, либо о его неэффективности. В контексте проверки по существу - наличие ложных заявлений, возникающих как случайно, так и преднамеренно.

Ошибка выборки допустимая - максимальное значение ошибки, обнаруженной в ходе выборки, в пределах которой исследователь все еще может сделать вывод о достоверности в целом данных, подлежащих проверке. **Ошибка выборки** допустимая определяется на стадии планирования аудита. Чем меньше размер **ошибки выборки** допустимой, тем больше должен быть объем выборки. Результатом высокой степени допустимой ошибки при малом объеме выборки могут быть данные, слишком приблизительные для подкрепления вывода о том, что совокупность в целом не имеет существенных искажений. Большая выборка помогает также обнаружить как частые, так и редкие отклонения или искажения, содержащиеся в денежной сумме. **Ошибка выборки** ожидаемая - примерное, субъективно оцениваемое значение ошибки, которое до начала проведения выборки аудитор предполагает получить в ходе ее проведения. Если ошибка превышает **ошибку выборки** ожидаемую, то объем выборки, необходимой для получения определения риска выборочной проверки на заданной допустимой степени отклонения, следует увеличить. Для определения **ошибки выборки** ожидаемой обычно используют результаты проверок за предшествующие годы; если их невозможно получить, то

пользуются небольшой выборкой из генеральной совокупности за текущий год или прибегают к опыту и интуиции аудитора.

Значение **ошибки выборки** ожидаемой не должно быть точным, т.к. оно влияет только на определение объема выборки, а не на оценку ее результатов. Исследователь обязан анализировать каждую ошибку, попавшую в выборку, экстраполировать полученные при выборке результаты на всю проверяемую совокупность и оценить риски выборки. Многие социологические исследования связаны с использованием случайной выборки из населения. Отдельная выборка может быть недостаточно репрезентативной - в этом случае говорят об ошибке выборки. Повторные выборки в конце концов выравнивают колебания между отдельными выборками и таким образом обеспечивают точную репрезентацию. Однако величину ошибки выборки можно оценивать и не прибегая к помощи повторных выборок, таким образом конструируя интервалы доверия и используя статистические проверки значимости. Вероятно то, что ошибка выборки будет больше, если выборка слишком мала, или если высока степень изменчивости характеристик населения. См.', Выборка; Значимости проверка: Неучастие в опросе: Смещение. Чем больше ступеней в многоступенчатом отборе, тем больше ошибка выборки. В любом случае, при многоступенчатом отборе ошибка всегда больше, чем при простом случайном. И еще, нужно отметить, что на каждой ступени все равно применяется случайный отбор. То есть, всякий раз важно, каким способом отбираются нужные единицы на всех ступенях выборки. расхождение между характеристиками выборочной и генеральной совокупности. Различают два вида ошибок выборки: случайную ошибку и систематическую ошибку, возникающую вследствие нарушения правил отбора (или из-за смещений при отборе). При определении случайной ошибки предполагается, что ошибка регистрации равна нулю. Систематическую ошибку часто называют ошибкой, вызванной смещением. Общая ошибка выборки складывается из случайной ошибки (вследствие

случайных различий между элементами совокупности, включенными в выборку и не попавшими в нее) и из смещения (систематической ошибки), если оно существует.

ЛИТЕРАТУРА

1. Горяченко Е. Е «планирование выборки для комплексного соц.экономического изучения деревни» // соц. исследования.
1975.3. с 45-52.
2. Давидюк Г. «Прикладная социология»// Обоснование выборки. Минск: Выщ. школа. 1979. с 153-159.
3. Мозер К. «Методы социального исследования» // Информ. Бюл. ИСМ АИ ССР. 1969. 31 с 93-94.
4. Процесс социального исследования. М.: 1975. с 116.
5. Рабочая книга социолога. Простой случайный отбор. // Под ред. Г. В. Осипова. Изд. 2-е перераб. и доп. М.: Наука. 1983.
с 205-214
6. Размещение выборки. // Территориальная выборка в социологических исследованиях. М.: Наука. 1970. с 40-50.
7. Шварц Г. «Выборочние метод» М.: Статистика. 1978. с 199.
8.Ноэль Э.»Массовые опросы». М.: Прогресс.1978 . с. 381.
9.Толстова Ю.Н.» Логика математического анализа социологических данных «.Москва. : Наука. 1991.
10.Дмитриев А.В. Политическая социология США.Л.,1971.
11.Рабочая книга социолога. М., 1983.Как провести социологическое исследование .М., 1985.
12. Заславская Т. об одном методе классификации объектов в социологии.// Социологические исследования .1974.№1
13. Йейтс Ф. Выборочный метод в переписях и обследованиях. М., 1965;
14. Дружинин Н.К. Выборочный метод и его применение в социально-экономических исследованиях. М., 1970;

15. Позулов А.И. Очерки истории отечественной статистики. М., 1972;
16. Пэнто Р., Гравитц М. Методы социальных наук. М., 1972;
17. Шереги Ф.Э. Применение метода квот в выборочных социологических исследованиях // Социологические исследования. 1975. 3;
18. Территориальная выборка в социологических исследованиях. М.: Наука. 1980.

ТЕМЫ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ:

1. Смешанные методы выборки.
2. Организация простой случайной выборки.
3. Цель метода выборки.
4. Основные элементы выборочного метода.
5. Возможности систематизированной выборки.
6. Кластерная выборка.
7. Территориальная выборка.
8. Квотная выборка.
9. Централизация информаций.
10. Условия стратифицированной выборки.