**Содержание**

1. Проблема качества социологического измерения

2. Надежность и валидность измерения

##### 3. Конструирование индексов и шкал

Библиографический список

**1. Проблема качества социологического измерения**

Переменная-признак, включенная в модель измерения, может задаваться совокупностью эмпирических индикаторов (в предельном случае — одним индикатором), наблюдаемые значения которых содержат, помимо истинного, ошибочный компонент. Задача оценки качества измерения — это, говоря обобщенно, задача разделения истинного значения и ошибки. Еще одна проблема, стоящая перед исследователем, связана с необходимостью использовать несколько показателей (вопросов анкеты, данных наблюдения и т. п.) для построения суммарного индекса или шкалы, позволяющих получать более точные, т. е. менее подверженные смещению, оценки интересующей исследователя переменной. Индексы и шкалы позволяют перейти от многообразия конкретных, наблюдаемых переменных-индикаторов, отражающих лишь отдельные аспекты теоретического понятия, к более абстрактным переменным теоретической модели. Так, например, конкретный вопрос об участии в голосовании дозволяет сделать определенные выводы о политической активности людей, однако ограничившись только этим вопросом, вы ничего не узнаете о других аспектах и формах политической активности. Кроме того, использование нескольких показателей позволяет уменьшить влияние посторонних факторов на ваши оценки величины или разброса значений переменной и сделать получаемые результаты более устойчивыми и воспроизводимыми. Иными словами, использование индексов и шкал ведет к повышению надежности и валидности измерения. С рассмотрения этих понятий мы и начнем данную главу.

**2. Надежность и валидность измерения**

Качество отдельных индикаторов, суммарных показателей (индексов) и шкал определяется их надежностью и валидностью.

Надежность измерения связана с его устойчивостью и воспроизводимостью. Показатель может считаться надежным в той мере, в которой полученные оценки могут быть воспроизведены на данной совокупности объектов измерения. Основной вид надежности — это надежность-повторяемость (или диахронная, ретестовая надежность). Оценка ретестовой надежности отражает результаты повторного применения одного и того же показателя (вопроса, теста) для одной и той же выборки случаев (респондентов) в разные моменты времени. Если люди отвечают на вопрос одинаково и в первом, и во втором, и во всяком последующем случае, то этот вопрос надежен. Если тест умственных способностей дает те же результаты при повторной проверке на одной и той же группе старшеклассников, то это надежный тест.

Важно отметить, что надежность инструмента измерения не отражает его точность или правильность. Скажем, если на вопрос о доходе респонденты дважды дадут совершенно идентичные ответы, завышенные на какие-нибудь 100%, мы имеем дело с абсолютно надежным, но неточным показателем. Если, в другом случае, термометр с безукоризненной надежностью показывает электрическое сопротивление кожи, у нас нет оснований говорить о правильности, адекватности измерения. Оценка надежности-повторяемости — это корреляция между результатами повторных измерений. В случае единичного вопроса-индикатора достаточно просто сопоставить ответы одной и той же группы людей, полученные в первом опросе, с ответами, данными двумя неделями или месяцем позднее, и вычислить коэффициент корреляции. На практике хорошей можно считать корреляцию 0,8 и выше.

К недостаткам оценок надежности-повторяемости следует отнести, во-первых, сложность проведения повторных замеров на больших выборках, типичных для социологии. Во-вторых, истинное значение переменной также может изменяться с течением времени, например, политические симпатии могут измениться за неделю в результате каких-то скандальных разоблачений, а зарплата — существенно возрасти за месяц из-за введения обязательного индексирования в условиях инфляции. В последнем случае перед исследователем встанет трудноразрешимая задача отделить колебания, вызванные изменением истинного значения переменной, от колебаний, связанных с надежностью показателя. Поэтому так важен еще один вид надежности — надежность-согласованность. Оценить надежность-согласованность можно лишь в том случае, если для измерения одной и той же переменной используют множество индикаторов. Зато для такой оценки достаточно однократного измерения. В психологическом тестировании, например, используют батареи тестов, нацеленных на измерение одной способности или одного личностного качества. Социологи — если они располагают достаточными средствами и техническими возможностями — используют индексы и шкалы, состоящие из множества отдельных вопросов-индикаторов.

Самый простой способ оценки надежности-согласованности — это «расщепление пополам». Если, например, у нас есть 12 вопросов, предположительно измеряющих политическую активность, где каждому вопросу присвоен 1 балл, а максимальной политической активности соответствует суммарный индекс 12 баллов, то применяется следующая процедура:

1. Вопросы в случайном порядке разбиваются на два равных списка (по 6 вопросов в каждом) и предъявляются один раз группе испытуемых.

2. Далее подсчитывается коэффициент корреляции между результатами одних и тех же испытуемых по разным «половинам» теста: чем выше его значение, тем согласованнее оценки истинного значения политической активности, получаемые с помощью данного набора индикаторов. Не исключено, однако, что полученная таким способом оценка надежности окажется весьма чувствительной к способу «расщепления пополам»: коэффициент корреляции будет заметно меняться в зависимости от способа составления двух списков.

Еще одна элементарная процедура, позволяющая оценить надежность отдельного вопроса (высказывания, пункта шкалы), — это проверка его корреляции с суммарным баллом, т. е. с суммарным значением индекса. Если для данной группы опрошенных коэффициент корреляции между отдельным вопросом о частоте зарубежных поездок и суммарным «индексом ксенофобии» оказался равен 0,3, то можно предположить, что названный вопрос не отражает истинного значения переменной «уровень ксенофобии» и может быть исключен из опросника. Ведь строго определенная надежность — это та доля измеренного разброса оценок, которая относится к истинному разбросу значений измеряемой переменной (мы пользуемся здесь менее строгим и скорее содержательным определением, поскольку пока не обсуждали необходимые статистические понятия). Очевидно, что коррелирование с суммарным баллом — это процедура, применимая для имеющих довольно простую структуру суммарных индексов и шкал.

В любом случае важно располагать явной моделью измерения теоретической переменной, так как лишь она позволит предсказать, каковы ожидаемые отношения между отдельными индикаторами и насколько применимы описанные простые методы оценки надежности. Чтобы убедиться в этом, достаточно сравнить модель с множеством эффект-индикаторов латентной переменной с моделью, включающей только причинные индикаторы. Очевидно, что эффект-индикаторы должны быть высоко согласованны и с латентной, переменной-свойством, которую они призваны измерять, и друг с другом. Однако это не так уж очевидно для причинных индикаторов: скажем, и образование, и доход — важные компоненты понятия «социально-экономический статус». Однако даже если образование растет, доход имеет право вести себя как угодно, т. е. он вовсе не должен показывать непременно высокую корреляцию с образованием. Иными словами, если от надежных эффект-индикаторов следует ожидать высокой скоррелированности друг с другом (при использовании методов «расщепления пополам» или корреляции с суммарным баллом), то для причинных индикаторов столь простой подход к оценке надежности неприменим. Разработать подходящий метод оценки надежности здесь можно, лишь анализируя взаимосвязи разных индикаторов и разных теоретических переменных в модели измерения. Прогнозируя ожидаемую направленность и величину этих связей, исследователь может оценить степень соответствия своих предсказаний наблюдаемым данным и сделать вывод о качестве индикатора. Самыми универсальными методами оценки надежности эмпирических индикаторов являются факторный анализ и путевой анализ. В идеале для оценки надежности используют несколько индикаторов (два-три) и по крайней мере две волны панели.

Очень важно помнить, что понятие надежности связано со случайными ошибками измерения, т. е. с ошибками, которые никаким систематическим образом не связаны друг с другом или какими-то систематически действующими внешними переменными (скажем, полом или возрастом респондентов). Типичные **источники ненадежности** — это случайные несистематические факторы, связанные с колебаниями внимания респондентов, неоднозначностью формулировки вопроса, ведущей к различию в его восприятии в разных случаях; несистематическими различиями в проведении интервью; различиями в кодировании открытых вопросов, или с ошибками при вводе данных. Например, если предложить даже очень опытным специалистам классифицировать сотню населенных пунктов по заранее разработанной схеме кодирования типов поселений, то, вероятнее всего, можно будет найти по крайней мере несколько расхождений в получившихся классификациях. Некоторые расхождения будут связаны с наличием «предельных» случаев, не поддающихся однозначной классификации по предложенным правилам, некоторые — с механическими ошибками записи или невнимательностью. Предварительная оценка надежности вопросов социологической анкеты требует прежде всего «отбраковки» неясно сформулированных вопросов, на которые люди часто отвечают случайным образом. Столь же низка надежность вопросов, на которые респонденты попросту не способны ответить, так как не имеют никакого мнения по затронутой проблеме или ничего не знают о ней. Далеко не все опрошенные, столкнувшись с иррелевантным вопросом, честно ответят «не знаю» или «не помню». Многие дадут наугад выбранный ответ из вежливости или нежелания демонстрировать свою неосведомленность.

Методы увеличения надежности нами уже обсуждались. Во-первых, нужно стремиться к использованию множественных индикаторов. Когда же это невозможно, т. е. существуют теоретические или практические трудности в измерении одной переменной разными способами, то следует использовать самые устоявшиеся и общепринятые показатели (например, если можно лишь один раз спросить респондента о его возрасте, то лучше всего использовать в точности такую же формулировку вопроса и те же категории ответа, какие используются в общенациональных переписях, масштабных панельных исследованиях и т.п.).

К другим методам увеличения надежности можно отнести «отсев» иррелевантных вопросов, анализ словесной формулировки вопроса, обучение и контроль интервьюеров, совершенствование методов кодирования данных и процедур ввода.

Валидность измерения, в самом общем смысле, характеризует соответствие измерения его цели. Эмпирический показатель валиден (обоснован, правилен) в той мере, в какой он действительно отражает значение той теоретической переменной, которую предполагалось измерить. Очевидно, что нет смысла говорить о валидности какого-то индикатора самого по себе. Валидность инструмента измерения состоит в однозначностш и правильности получаемых результатов относительно измеряемого свойства объектов, т. е. относительно предмета измерения. Можно сказать, что валидность определяет «чистоту» измерения теоретического конструкта. Когда измерение является непосредственным, т.е. мы можем прямо подсчитать количество эталонных единиц измеряемого свойства, и на результаты измерения влияют только случайные ошибки, надежность и валидность неразличимы, валидность инструмента измерения равна его надежности. Если мы измеряем интересующее нас свойство лишь косвенно, используя какой-то индикатор, возникает различие между надежностью и валидностью. Индикатор может обладать высокой надежностью (воспроизводимостью), но при этом измерять интересующий нас социологический конструкт недостаточно «чисто». Уже на интуитивном уровне очевидно, что вполне надежный инструмент может измерять нечто другое, помимо интересующего исследователя качества (например, не столько политическую активность, сколько конформизм). Косвенное измерение обычно содержит и случайный, и неслучайный ошибочный компонент. Именно неслучайный компонент, включающий в себя систематическую (скоррелированную) ошибку и, реже, имеющую одну и ту же величину для каждого случая постоянную ошибку измерения, определяет валидность показателя. Характерными примерами систематической ошибки измерения в социологическом опросе или эксперименте являются уже упоминавшиеся эффекты «памяти», социальной желательности, установки за позитивный или негативный ответы. Они влияют не только на правильность, валидность индикатора интересующей исследователя переменной, но и на правильность и обоснованность результатов анализа данных: скоррелированная ошибка измерения может воздействовать на любые статистические показатели, в том числе на показатели взаимосвязи между переменными и на оценки значимости различий между подгруппами. Иными словами, конечным итогом «пользования невалидных индикаторов могут оказаться неверные содержательные выводы.

Проблема валидности измерения — сложнейшая проблема социологической методологии. Валидное измерение — это прежде всего результат валидной модели измерения, т. е. результат обоснованной и ясной концептуализации теоретических представлений. Здесь мы опишем лишь основные виды валидности и традиционные методы валидации, т. е. установления валидности измерений.

Валидностъ по содержанию показывает, в какой мере избранные исследователем индикаторы отражают различные аспекты теоретического понятия. Иными словами, речь идет о представительности данной совокупности измерений да отношению к концептуальной структуре переменной-признака, о полноте операционализации теоретических понятий. Например, экзамен по статистике может рассматриваться как валидный инструмент измерения статистических знаний студентов, так как экзаменационные вопросы отражают содержание лекций и учебников. Однако если все вопросы относятся лишь к одному разделу прочитанного курса — скажем, к нормальному распределению, — то результаты экзамена будут отражать, например, умение студентов переводить «сырые» баллы в стандартные оценки, но ничего не скажут о знании корреляции и регрессии.

Основная процедура оценки валидности по содержанию — это суждение эксперта. В некоторых случаях связь между теоретическими понятиями и измеряющими их индикаторами столь ясна, что никакие специальные обоснования попросту не требуются: понятно, что термометр измеряет температуру. Здесь можно говорить об очевидной (иногда—лицевой, от англ. face validity) валидности показателя. Очевидная валидность тем выше, чем тождественнее понимание цели вопроса, теста или иного показателя профессионалом-социологом и неискушенным респондентом. Вопрос о частоте покупки шампуня, по всей вероятности, не содержит в себе никаких подвохов и позволяет судить именно о том типе потребительского поведения, который описан в вопросе. Однако в более сложных случаях содержательная валидность отнюдь не сводится к очевидной. Набор простых вопросов о излюбленном способе проведения досуга, предпочитаемой марке автомобиля, частоте чтения престижного журнала и т. п. может быть нацелен на измерение «стиля жизни» респондента (в данном случае измерение позволяет отнести человека к одной из номинальных категорий стиля жизни — «выживающий», «достиженческий», «экзистенциальный», «социальный» и т. п.). Судить о полноте этого набора и относительной значимости вопросов для измерения понятия «стиль жизни» могут только специалисты. Основой такого экспертного суждения является теоретическое определение, концептуализация исследовательской переменной. Обычно экспертное суждение о валидности по содержанию выносится более или менее стихийно, после публикации результатов исследования. Иногда все же удается использовать более организованные процедуры — метод параллельных панелей или метод нескольких судей. В первом случае две или три последовательные панели специалистов проводят всю процедуру валидации по содержанию, т. е. сравнивают существующие дефиниции, составляют список возможных индикаторов и оценивают их репрезентативность по отношению к исследуемой концептуальной области. Если сравнение индикаторов, независимо отобранных в двух и более панелях, обнаруживает множество совпадений, можно говорить о высокой содержательной валидности.

Метод нескольких судей (экспертов) полезен в тех случаях, когда переменная-признак, которую предполагается измерить, имеет многомерную структуру. Если, например, социолог разрабатывает воображаемую шкалу социально-экономического благополучия регионов, то полезно обратиться к специалистам в таких областях, как демография, социальная политика, занятость, налоговая система и т. п. Опрос экспертов позволит выявить существенные факторы, входящие в шкалу, оценить их сравнительную значимость и найти релевантные эмпирические индикаторы.

Критериальная валидность (или валидность по критерию) показывает, насколько хорошо результаты по данному тесту или индикатору согласуются с результатами измерения другого показателя, называемого критерием. Чаще всего критерий — это та переменная, которая и представляет практический интерес для исследователя, но не может быть измерена в данный момент. Например, критериальная валидность вступительных экзаменов определяется той академической успеваемостью, которую в дальнейшем продемонстрируют студенты (т. е. академическая успеваемость является в данном случае критерием). Можно также предположить, что тест моральной оценки девиантного поведения для подростков обладает критериальной валидностью по отношению к реальному отклоняющемуся поведению. Индикатор, обладающий доказанной критериальной валидностью, может рассматриваться как переменная-предиктор, позволяющая предсказывать индивидуальные значения переменной-критерия. Конечно, нужда в предикторе, замещающем собственно критерий, возникает лишь в тех случаях, когда оценки по критерию получить трудно, т. е. речь идет о давно прошедших или еще не наступивших событиях, либо переменную-критерий трудно измерить из-за практических или этических соображений. Валидность предиктора обычно тем выше, чем ближе он к критерию. Скажем, идеальным методом отбора курсантов авиационного училища мог бы стать пробный краткосрочный курс обучения с проверкой практических навыков управления самолетом в финале: прошедшие проверку претенденты имели бы все шансы стать настоящими профессионалами. Однако такой метод слишком дорогостоящ и на практике используют тесты интеллекта, испытания визуально-моторной координации и другие показатели, установив предварительно их критериальную валидность. В другом случае проективный личностный тест (типа ТАТ — теста тематической апперцепции, подразумевающего составление рассказов по фотографиям с неопределенным сюжетом) позволит выявить признаки психоза либо травмирующего сексуального опыта в прошлом пациента. Полное психиатрическое обследование могло бы занять очень много времени, да и данные о плохом обращении в детстве получить довольно трудно.

К основным типам критериальной валидности относят прогностическую, конкурентную и постдиктивную («предсказывающую-назад») валидности.

Прогностическая критериальная валидность описывает точность, с которой значения данной переменной — обычно характеризующей отдельного индивида или группу — могут быть предсказаны на основании текущих значений какой-то другой переменной (предиктора). Очевидно, что наилучшим показателем такой прогностической точности будет корреляция между значениями переменной-предиктора и значениями переменной-критерия для одной и той же выборки. Тогда — в пределах ошибки выборки — коэффициент корреляции будет равен коэффициенту прогностической валидности.

Конкурентная валидность по критерию — это степень соответствия между текущими значениями переменной-критерия и переменной-предиктора. Попросту говоря, исследователь использует результаты измерения по одному признаку для того, чтобы оценить значение другой переменной. Причина может заключаться в том, что измерение непосредственно переменной-критерия трудноосуществимо, занимает слишком много времени и т. п. Проблема конкурентной валидности весьма значима для эмпирической социологии, где многие поведенческие или установочные переменные измеряются не непосредственно, а через самоописания, ответы на анкетные вопросы, иными словами, через вербальное поведение. Например, мы можем оценить конкурентную валидность анкетного вопроса о частоте посещения дискотеки студентами-отличниками с помощью серии включенных наблюдений за реальным поведением данной группы. Уже ранние исследования конкурентной валидности фактографических вопросов в социологии показали, что даже для относительно «безобидных», несензитивных индикаторов конкурентная валидность может изменяться в самых широких пределах.

В так называемом Денверском исследовании валидности, проведенном С. Стауффером и соавторами в 1947 г., сопоставлялись данные ответов респондентов на фактографические вопросы и данные официальной статистики местных организаций. Речь шла о регистрации избирателей, участии в голосовании, взносах в городскую казну, о наличии водительских прав и читательского билета и т. п. Сопоставляя данные официальной статистики (критерий) с ответами респондентов (предиктор), исследователи обнаружили, что величина расхождений составляла от нескольких процентов до почти 50%, в зависимости от содержания вопроса.

Следует, однако, помнить и об ограничениях, присущих объективным показателям-критериям: данные официальных документов также нередко основаны на самоотчетах и нередко подвержены ошибкам измерения. Все же в некоторых случаях процедура конкурентной валидации имеет преимущество перед предиктивной, так как первая не требует от исследователя длительного ожидания того момента, когда можно будет измерить значения критерия. Если, например, исследователь хочет оценить критериальную валидность теста профессиональных интересов, разработанного для студентов, как предиктора успешной профессиональной карьеры, то ему необязательно ждать десять лет, чтобы измерить значение переменной-критерия. Достаточно провести тестирование профессиональных интересов для двух «крайних» групп уже работающих специалистов — преуспевших и наименее преуспевших в профессии в данный момент времени. Высокая корреляция между тестовым баллом и успешностью работы (или статистически значимое различие тестовых баллов «крайних» групп) будет свидетельствовать о конкурентной валидности теста. Условиями, при которых выводы о конкурентной валидности индикатора могут все же оказаться ошибочными, являются избирательное выбывание из выборки (самоотбор) и реактивность переменной-предиктора. Самоотбор в нашем примере может иметь место в том случае, если среди выбывших из выборки (бывших студентов, отказавшихся от профессиональной карьеры в данной области и не охваченных по этой причине проводимым тестированием) будут сверхпредставлены высоко- либо низкомотивированные, т. е. естественное выбывание будет носить неслучайный характер. В случае реактивности индикатора наши испытуемые будут отвечать на вопросы теста мотивации не так, как они отвечали будучи студентами (из-за повлиявших на них профессионального опыта, изменения социального статуса и т. п.) И все же существуют нереактивные переменные-предикторы, конкурентная валидность которых вполне поддается обоснованию. Примером могут служить такие стабильные характеристики, как коэффициент интеллекта или «фоновые» переменные (социальное происхождение, национальная принадлежность и т. п.).

Наконец, в некоторых случаях мы заинтересованы в том, чтобы установить точность, с которой мы можем оценить наличие какого-то критерия-признака или черты, присущей индивиду (группе) в прошлом. Пример использования проективного психологического теста для постдиктивного «прогноза» детских травм приведен выше. Эта ситуация обозначается как оценка постдиктивной валидности.

Описанные нами виды валидности существенны в тех ситуациях, когда перед исследователем стоит задача сравнить некий показатель с уже существующими или с используемым в практике критерием. Иными словами, критериальная валидность показателя — это корреляция с другим, предположительно «чистым», эмпирическим показателем. Однако существует и другой подход к валидности, где оценка индикатора основана на том, насколько хорошо его «поведение» соответствует теоретическим ожиданиям. Такая оценка может быть проведена лишь в рамках целостной теоретической модели, описывающей отношения между теоретическими переменными, их индикаторами, случайными и неслучайными ошибками измерения. Предположим, мы хотим проверить валидность новой шкалы групповой сплоченности. Основываясь на существующих теоретических представлениях, мы можем предположить, что большей групповой сплоченности соответствует меньшая частота открытых конфликтов и большая интенсивность коммуникаций. Сравнив различные по уровню групповой сплоченности группы и определив для них значения других двух переменных (частота конфликтов и интенсивность коммуникаций), мы увидим, насколько хорошо «ведет» себя разработанный нами показатель групповой сплоченности. Если паттерн его отношений с двумя другими переменными соответствует предсказаниям теории, то мы можем заключить, что новая шкала валидна, т. е. измеряет именно ту теоретическую переменную, которая нас интересует. Этот вид валидности обычно обозначают термином «конструктная валидность». (Иногда используют также обозначение «концептуальная валидность».)

Со статистической точки зрения абсолютная конструктная валидность предполагает, что весь наблюдаемый разброс в значениях показателя связан исключительно с измеряемым теоретическим конструктом. Если же часть вариации индикатора связана с другой переменной — будь то другой теоретический конструкт или систематическая ошибка измерения,— конструктная валидность окажется меньше.

Легко заметить, что оценка конструктной валидности предполагает какую-то связь между проверкой содержательных теоретических гипотез и проверкой качества измерения. Действительно, оценка конструктной валидности посредством сопоставления теоретической модели «поведения» изучаемой переменной с реальными отношениями индикаторов требует включения модели измерения в более широкую теоретическую модель.

Предположим, мы используем некоторый суммарный показатель — индекс «демократизма политической системы», состоящий из ряда индикаторов (Х1, Х2, Х3), каждый из которых имеет свой собственный «вес» в индексе демократизма. К таким индикаторам могут относиться наличие парламента (номинальная дихотомическая переменная), количество независимых телерадиокомпаний и т.п. Исходя из теоретической модели, мы ожидаем, что степень «демократизма» находится в обратной связи с долей ВНП, расходуемой на модернизацию вооружений. Основываясь на этой модели, можно проверить конструктную валидность изобретенного нами индекса демократизма.

 **С**

Доля расходов на вооружение,

***Y***

*Х*2

*Х*3

*Х*1

Рис. 1. **Модель взаимосвязи для переменных «демократизм» и «доля расходов на вооружение»**

Собрав необходимые данные для 10—15-и национальных государств, мы можем обнаружить, что наш индекс «демократизма» невалиден, так как ожидаемое отношение между теоретическими переменными (с) не выполняется, их корреляция равна нулю. Просмотрев наши данные, мы, например, обнаружим, что в некоторых странах, почитаемых за образец демократического общественного устройства, изрядную часть бюджета составляют военные расходы, тогда как некоторые деспоты из «банановых республик» вполне обходятся кремневыми ружьями. Однако вывод о невалидности нашего измерения «демократизма» верен лишь в том случае, если верны наши теоретические представления о связи демократии и пацифизма. Если же демократия и пацифизм отнюдь не связаны друг с другом, наши результаты вовсе не доказывают низкую валидность индикатора: вполне возможно, что как раз «демократию» мы измеряли правильно, но неверна была наша теоретическая гипотеза. Существует своеобразное отношение дополнительности между собственно теоретическими моделями и моделями измерения. Оценить качество показателей в модели измерения можно, лишь приняв теоретическую модель как безусловно верную. Для оценки справедливости собственно теоретической модели, нужно принять предположение о конструктной валидности индикаторов и провести **новое** исследование с **новыми** данными.

Существуют сложные статистические методы, позволяющие одновременно оценивать модель измерения и теоретическую модель (часто их называют «LISREL-методы»). Они применимы лишь к моделям с несколькими индикаторами для каждой переменной. Однако некоторые методологи полагают — на наш взгляд, справедливо, — что попытки проверить модель измерения и совокупность теоретических гипотез **на одних и тех же данных** чреваты возможностью ошибочных выводов. Если теория, предсказания которой мы используем для проверки конструктной валидности, относительно нова и не стала еще общепринятой истиной, мы просто не сможем определить, связан ли отрицательный результат исследования с невалидностью показателя, или причиной всему — ложные теоретические представления. Крометого, может оказаться, что мы отберем худший из показателей, ибо именно онподдерживает неверную теорию. Поэтому проверка конструктной валидности индикаторов и проверка теорий требуют от нас разных исследований, множественных показателей и разных матриц данных.

**3. Конструирование индексов и шкал**

Использование нескольких индикаторов, как было показано выше, увеличивает валидность и надежность измерения переменных. Здесь, однако, возникает новая проблема: как использовать полученные значения индикаторов для того, чтобы охарактеризовать каждый «случай» (каждого респондента, группу, страну и т. п.) одним числовым значением, однозначно определяющим его положение на одномерном континууме переменной-признака, для измерения которой мы использовали данный набор индикаторов. Иными словами, нужно осуществить обратный переход от набора значений эмпирических индикаторов, описывающих каждую конкретную единицу анализа, к упорядочению всех единиц анализа по оси интересующей нас переменной. Такое упорядочение и называется собственно шкалой, мерой выраженности переменной-признака, а логика перехода от набора наблюдаемых значений к шкальным значениям называется моделью шкалирования. Заметим сразу, что некий набор индикаторов — например, набор оценочных шкал — может использоваться для измерения более чем одной переменной, и, следовательно, данные о наблюдаемых значениях этих индикаторов в принципе позволяют упорядочить «случай» по нескольким переменным, т.е. по нескольким шкалам. Однако это уже задача многомерного шкалирования, мы же пока ограничимся обсуждением одномерных шкал и индексов.

Если вернуться к структурированной матрице данных «переменная х случай», то можно увидеть, что процедура конструирования шкалы может быть описана и как процедура «сжатия» матрицы данных, уменьшения ее размерности. Предположим, три строки нашей матрицы соответствуют переменным-индикаторам «доход», «род занятий» и «образование». Мы включили эти индикаторы в наше исследование ради того, чтобы охарактеризовать социально-экономический статус каждого респондента, т.е. расположить их от низкого статуса к высокому. Если мы вместо трех строк, соответствующих доходу, образованию и профессии, введем в нашу матрицу данных одну строку, отражающую положение каждого респондента на сконструированной нами шкале СЭС, размерность матрицы уменьшится. Однако сначала нам нужно решить, как объединить три значения — три строки матрицы — в одно, т. е. нам нужно избрать модель шкалирования.

Пусть, скажем, три строки нашей матрицы данных — это полученные каким-то образом (тестирование, опрос экспертов и т. п.) оценки «жизнерадостности», «энергичности» и «независимости». Исследователь предполагает, что эти три индикатора могут быть использованы для измерения важной для его теории переменной «сила Я». Все, что ему нужно сделать — это решить, как перевести оценки в строках 1—3 в оценки «силы Я» (см. рис. 2).

Рис. 2. **Фрагмент матрицы данных «переменные х субъекты»**

|  |  |
| --- | --- |
| Переменные-индикаторы | Субъект («случай») |
| Л.М. | Ф.Ж. | К.Р. | … |
| 1. жизнерадостность | 2 | 0 | 2 | … |
| 2. энергичность | 2 | 2 | 2 | … |
| 3. независимость | 0 | 2 | 2 | … |

«Сила Я»?

Самый простой и очевидный способ — это суммировать для каждого индивида оценки по каждому индикатору. Получившийся суммарный балл будет отражать индивидуальные различия в «силе Я», так как позволит упорядочить всех респондентов от минимального к максимальному значению этой переменной (в нашем примере — от 0 до 6 баллов). Еще одно преимущество суммирования — увеличение разброса индивидуальных значений. Действительно, максимально возможное различие по первичным индикаторам составляло 2 балла (от 0 до 2). В суммарном показателе разница между индивидуальными значениями может составить 6 баллов. Следовательно, суммарный балл — это более «чуткий» и надежный инструмент для упорядочения и может быть назван шкалой в смысле определения, данного нами выше. Однако в социологии суммарные показатели чаще называют индексами, чтобы подчеркнуть их единственное важное отличие от «больших» шкал. Индекс позволяет эффективно «свернуть» информацию, содержавшуюся в исходных индикаторах (вопросах, пунктах, тестах), однако от суммарного балла нельзя вернуться к исходной матрице, точнее, к тому паттерну ответов, который стоит за данным значением индекса. Если сформулировать это корректнее, индекс не позволяет учитывать различия в структуре ответов респондентов. Если снова обратиться к рисунку 12, то можно заметить, что субъекты Л. М. и Ф.Ж. имеют одинаковый суммарный балл, равный 4 (достаточно высокое значение!). Но можно ли считать несущественным то обстоятельство, что у Л. М. нулевой уровень независимости, а Ф. Ж. получил тот же суммарный балл из-за недостатка оптимизма? Предположим, даже довольно мрачный человек может обладать значительной «силой Я», но следует ли считать столь же «сильным» того, кто легко поддается давлению окружения?

В принципе индексы безусловно применимы в тех случаях, когда модель измерения предполагает, что некая латентная, т. е. не измеряемая непосредственно переменная, может быть измерена с помощью совокупности качественно однородных показателей. Во многих случаях различия в значимости, важности отдельных индикаторов можно учесть с помощью «взвешивания», пересчета значений с учетом «веса» каждого индикатора в латентной переменной. Так в примере с «силой Я» можно домножить все индивидуальные значения в строке «независимость» на 2, если принять предположение о том, что независимость влияет на латентную переменную с двукратным эффектом.

Экономисты часто используют индексы розничных цен, отражающие динамику стоимости жизни. При этом разные товарные группы, например, имеющие неодинаковое значение в потребительском бюджете, — как, скажем, хлеб и деликатесы — учитываются с разными весовыми коэффициентами. Но и в этом случае индекс остается несовершенным типом шкалы: эмпирическая информация здесь используется лишь для шкалирования различий между субъектами (или другими единицами анализа), но не для шкалирования различий между пунктами-ответами {эмпирическими индикаторами). Используя «взвешивание», мы вводим априорные ограничения на упорядочение входящих в индекс индикаторов, не зависящие от данных наблюдения.

Своеобразным переходом между моделью суммарного балла (индекса) и основными моделями шкалирования является шкала Р. Ликерта (Лайкерта). Исходным материалом для ее построения служат оценочные шкалы согласия-несогласия с суждениями, которые выражают более или менее «благожелательную» установку. Количество категорий ответа — «согласен», «совершенно согласен» и т. п. — обычно варьирует от двух до семи. Респондент получает балл по каждому суждению в зависимости от избранного им ответа. Присуждаемый данному ответу балл в свою очередь определяется «благожелательностью» ответа по отношению к измеряемой установке (интенсивностью согласия с суждением), т. е. ответы также упорядочены на одномерном континууме (от крайне негативной установки к крайне позитивной). Баллы, полученные за каждый ответ, суммируются. Суммарный балл, полученный индивидуумом, характеризует уже его собственное положение на установочном континууме (например, «консерватор», «умеренный консерватор», «умеренный либерал», «либерал»). Отметим сразу, что эта же модель шкалирования может использоваться и для измерения мотивации или осведомленности (соответственно респондента просят оценить степень важности какого-то объекта или сказать, верно или неверно определенное утверждение). Для отбора списка суждений, составляющих шкалу Ликерта, исходный список высказываний предъявляют репрезентативной выборке респондентов (так называемой выборке стандартизации). В окончательный список попадают те высказывания, для которых были получены высокие оценки надежности — согласованности и валидности. Обычно используют описанные нами ранее методы оценки надежности и валидности (коррелирование с суммарным баллом, сравнение «крайних групп» и т. п.).

Приведем в качестве примера некоторые высказывания «Теста для измерения художественно-эстетической потребности молодежи» (в скобках дан ключ к каждому высказыванию, показывающий, за какой ответ присуждается балл):

1. Думаю, что вполне можно обойтись без общения с произведениями искусства (неверно).

2. Я не люблю стихов (неверно).

3. Я коллекционирую записи классической музыки (верно).

Шкалирование по описанной модели дает ординальный уровень измерения.

Шкалы социальной дистанции Э. Богардуса — старейшая модель социологического шкалирования, не утратившая, однако, своей популярности. Исследователь разрабатывает совокупность вопросов, отражающих различную степень близости отношений с определенной социальной или этнической группой, например:

1. Согласны ли Вы, чтобы хорваты жили с Вами в одном городе?

2. Согласны ли Вы жить по соседству с хорватами?

3. Согласны ли Вы работать в одном отделе (учреждении) с хорватом?

4. Позволите ли Вы своей дочери выйти замуж за хорвата?

Предполагается, что согласие с каждым последующим утверждением отражает переход к очередной градации ординальной шкалы установок — от меньшей близости к большей. Существенным требованием к избранной совокупности вопросов является их содержательная валидность, иными словами, здесь необходимы экспертные процедуры, описанные выше. Важно также убедиться в обоснованности предположения об одномерности шкалируемой переменной. Если в данных, полученных при использовании шкал социальной дистанции, встречаются «нелогичные» (так называемые нешкалируемые) индивидуальные паттерны ответов, причиной чаще всего бывает влияние другой переменной. Примером нешкалируемого паттерна ответов может служить ситуация, когда респондент, отрицательно ответивший на «слабые» вопросы, неожиданно соглашается с более «сильными», предполагающими высокую степень близости (среди специалистов по социологическим методам имеет хождение соответствующая шутка: если человек, не желающий жить в одном городе с черными, согласен выдать свою дочь замуж за черного, это не ошибка измерения: просто он одинаково ненавидит негров и собственную дочь).

Шкала равнокажущихся интервалов Л. Терстоуна позволяет достичь более высокого уровня измерения установок, чем ординальный. Она представляет собой целый класс методов интервального шкалирования и будет рассмотрена здесь в качестве наиболее простого примера.

Первая шкала равнокажущихся интервалов была описана в работе 1929 года и предназначалась для измерения остановок по отношению к церкви как социальному институту. Этой работой мы воспользуемся для того, чтобы проиллюстрировать основные этапы предложенной Терстоуном процедуры.

Шкала Терстоуна позволяет расположить и суждения, и индивидов вдоль одномерного континуума установки, полюсам которого соответствует крайне благожелательное и крайне негативное отношение к объекту установки (церкви, партии, прогрессивному налогообложению или чему-либо еще). Шкальный балл суждения или индивида отражает степень этой благожелательности или неблагожелательности.

На первом этапе исследователь составляет максимально широкий список суждений (высказываний), выражающих интересующую его установку. Так, Терстоун собирал мнения коллег, студентов, высказывания из публикаций, касающихся церкви. Здесь уместны также интервьюирование, использование открытых вопросов («Что Вы думаете о...?»), групповая дискуссия и т. п. Собранные суждения были подвергнуты первичному отбору. Исследователи отсеяли те высказывания, которые не удовлетворяли обычным требованиям к конструированию вопросов — двусмысленные, слишком длинные, содержащие специальные термины и т.п.. При первичном отборе суждений для шкалы Терстоуна используют и некоторые специальные критерии:

1. Исключаются суждения, относящиеся скорее к прошлому, чем к настоящему (например, «В средневековье церковь играла важную роль в общественной жизни»).

2. Исключаются суждения, описывающие факты, а не мнения и отношения. Конечно, далеко не всегда можно отделить высказывания, описывающие фактическое положение дел, от прочих. Скажем, слова «Бог любит нас всех» — факт для верующего, хотя другие люди могут усмотреть в них определенное отношение к религии. В практических целях вполне достаточно руководствоваться следующим критерием для выявления фактических суждений, подлежащих устранению из шкалы Терстоуна: фактом является любое высказывание, для установления истинности которого могут быть использованы какие-то «посюсторонние» процедуры верификации.

3. Исключаются также суждения, содержащие слова «все», «всегда», «никто», «никогда», так как этим словам люди обычно придают различный смысл, что затрудняет интерпретацию.

В результате исходный список из 350—400 суждений сокращается до 100—120. Следующим этапом является «судейская» процедура, позволяющая определить шкальное значение для каждого суждения и провести среди них окончательный отбор. Терстоун предложил разделить гипотетический континуум благожелательного-неблагожелательного отношения к церкви на 11 категорий (от «А» до «К»), разделенных субъективно равными интервалами. Требование субъективного равенства интервалов между градациями весьма существенно для построения шкалы Терстоуна и обычно его специально подчеркивают в инструкции для «судей» (например, «Представьте, что карточки с буквами от „А" до “К" представляют расположенные на равном расстоянии градации шкалы, так что градации „А" соответствует максимально благожелательное отношение к Х (объекту установки), а „К" — максимально неблагожелательное, негативное отношение»). Каждое из утверждений списка печатается на отдельной карточке, которые и раздаются «судьям» (в конструировании шкалы установок по отношению к церкви участвовало 300 таких экспертов). Задача «судей» заключается в том, чтобы разложить все 100—120 суждений по 11 рубрикам соответственно степени выраженного в них благожелательного или неблагожелательного отношения к объекту остановки.

Подчеркнем, что «судей» не просят высказать их собственное мнение, они должны лишь рассортировать высказывания.

Шкальное значение (балл) каждого из высказываний определяется распределением оценок «судей», поэтому началом следующего этапа (собственно построения шкалы) является подсчет процента экспертов, положивших высказывание в определенную стопку. Далее подсчитывается суммарный (кумулятивный) процент «судей», отнесших суждение к данной градации и предшествующим градациям. Терстоун присваивал использовавшимся градациям числовые значения от 1 (градация «А», максимально благожелательное отношение к церкви) до 11 (градация «К»). Проиллюстрируем дальнейшее на примере гипотетического суждения N, данные для которого представлены в таблице 1.

Таблица.1

**Распределение «судейских» оценок для суждения N**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Градация (числовое значение) | «А» (1) | «В» (2) | «С» (3) | «D» (4) | «Е» (5) | «F» (6) | «G» (7) | «Н» (8) | «I» (9) | «J» (10) | «K» (11) |
|  Процент судей, отнесших суждение к данной градации | 1 | 2 | 2 | 1 | 3 | 33 | 34 | 12 | 7 | 3 | 2 |
| Кумулятивный процент | 1 | 3 | 5 | 6 | 9 | 42 | 76 | 88 | 95 | 98 | 100 |

Распределение кумулятивных (накопленных) процентов позволяет вычислить значения медианы и междуквартильного размаха. Медиана, или процентиль 50 в распределении накопленных частот, — это такое значение на шкале «А» — «К», относительно которого половина судей дала большие, а другая половина — меньшие оценки данного утверждения. Медиана, таким образом, делит пополам упорядоченное множество значений признака. Вычислить медиану мы можем по следующей формуле:

В методе Терстоуна ширина интервала между соседними численными градациями по определению равна 1 (равнокажущиеся интервалы). В используемом нами примере границами интервала, где расположена медиана (процентиль 50), являются градации «F» и «G» (см. табл. 1). Фактической нижней границей интервала медианы будет значение 6,5, отсюда:

Значение медианы и принимается за шкальный балл («цену») суждения. Для гипотетического суждения N в нашем примере он оказался равен 6,7. (В принципе более простым является графический метод нахождения медианы. Для этогона миллиметровой бумаге строится кривая накопленных процентов — огива, позволяющая легко найти числовое значение, соответствующее процентилю 50.)

Ясно, однако, что не все суждения, получившие оценку «судей», в равной мере пригодны для шкалы: некоторые из суждений получат весьма согласованные и единодушные оценки экспертов, тогда как другие вызовут разнобой во мнениях. Для оценки внутренней согласованности отдельных высказываний шкалы Терстоун применил меру разброса судейских оценок — междуквартильный размах. (Здесь снова вместо распределения абсолютных частот экспертных оценок используется распределение процентилей, т. е. накопленные частоты выражают в кумулятивных процентах, что позволяет сравнивать выборки разного объема.) Междуквартильный размах — это расстояние между первым и третьим квартилем распределения. Первый квартиль (Q1) задается точкой на оси, до которой лежит 25% полученных оценок суждения, а третий ((Q3) — точкой, выше которой лежит 25% оценок. (Легко видеть, что второму квартилю соответствует медиана.) Для вычисления междуквартильного размаха (Q3 ⎯ Q1) сначала устанавливаются значения, соответствующие первому и третьему квартилям распределения. Для этого используются формулы, аналогичные формуле для медианы, с соответствующими поправками: берется фактическая нижняя граница интервала соответствующего квартиля, кумулятивный процент для нижней границы интервала данного квартиля и т. д. Так, для первого квартиля формула подсчета такова:

Для нашего примера с суждением N:

Читатель может самостоятельно выписать аналогичную формулу для третьего квартиля (75 процентиль), произвести необходимые подсчеты и убедиться, что для вымышленного суждения N междуквартильный размах (Q3 ⎯ Q1) составит 1,7. Те суждения, для которых разброс оценок, измеренный через междуквартильный размах, оказывается слишком велик, исключаются из шкалы Терстоуна. Предполагается, что высказывание, получившее столь разные оценки, воспринимается слишком неоднозначно. Так, Терстоун исключил из первоначально предъявленного «судьям» списка 90 высказываний из 130. В результирующей шкале оставляют одно-два высказывания для каждого деления шкалы, чтобы все градации предполагаемого установочного континуума оказались в равной мере представлены.

Если получившуюся шкалу предъявить теперь группе респондентов, то индивидуальным баллом каждого субъекта, выражающим меру «благожелательность» его установки, можно считать медиану (или средний балл) всех суждений, с которыми он согласился.

Многие критики шкалы Терстоуна указывали на возможность влияния на получаемые результаты характеристик «судейской» группы и широты диапазона предлагаемых суждений. Все же существуют веские основания считать, что такая шкала обладает вполне удовлетворительной воспроизводимостью и в среднем диапазоне дает уровень измерения, превосходящий ординальный (является так называемой шкалой разностей). Удаление или прибавление пункта шкалы не меняет шкальных значений других пунктов-высказываний. Приведем некоторые примеры высказываний, включенных Терстоуном в окончательный вариант шкалы установок по отношению к церкви (в скобках указан шкальный балл суждения):

• «Я думаю, что церковь – это наиважнейший социальный институт в современной Америке» (0,2);

• «Когда я нахожусь в храме, мне доставляет удовольствие наблюдать за службой, особенно если при этом звучит хорошая музыка» (4,0);

• «Я ощущаю потребность в какой-то религии, но не нахожу того, что мне нужно, ни в одной из существующих церквей» (6,1);

• «С моей точки зрения, церковь безнадежно устарела» (9,1).

В основе шкалы Гутмана также лежит идея кумулятивности: одни высказвания-пункты имеют большую «нагрузку» на одномерном континууме шкалируемой переменно-признака, другие – меньшую. Модель шкалирования, предложенная Гутманом, подразумевает, что в идеальном случае респондент, согласившийся с более «нагруженным» пунктом, согласится и со всеми менее «нагруженными». Таким образом, знание максимального шкального балла респондента позволяет полностью воссоздать паттерн его ответов. Шкалируемая переменная-признак не обязательно является установкой, она может характеризовать поведение: одна из первых гутмановских шкал, например, содержала пункты, описывающиен симптомы реактивного невроза, расстройства сна, тошнота, страх и т.п. В предложенной Гутманом процедуре обычно используется совокупность дихотомических вопросов, т.е. вопросов, на которые могут быть даны лишь два ответа: “да” или “нет”, “согласен” или “не согласен”. Совокупность вопросов или утверждений, составляющих гутмановскую шкалу, должна обладать одним существенным свойством: вероятность положительного ответа монотонно возрастает с ростом значения шкалируемой (латентной) переменной. Отсюда следует, что субъекты, имеющие больший шкальный балл, т.е. большее значение латентной переменной, с большей вероятностью дают положительный ответ на каждый отдельный вопрос. Это условие легко проиллюстрировать на примере следующих вопросов о росте (для простоты изложения предположим, что ошибка измерения отсутствует, т.е. все респонденты знают и точно сообщают свой истинный рост):

1. Верно ли, что Ваш рост превышает 1,50 м?
2. Верно ли, что Ваш рост превышает 1,60 м?
3. Верно ли, что Ваш рост превышает 1,70 м?

Эти вопросы образуют идеальную гутмановскую шкалу: если ошибка измерения отсутствует, любой респондент, ответивший положительно на вопрос 3, дает положительный ответ и на вопросы 2 и 1. Вообще, зная максимальный балл респондента, мы можем полностью воссоздать его паттерн ответов. Для вопросов о росте все возможные паттерны ответов (шкальные типы) приведены в табл. 2.

Таблица 2

**Ответы на вопросы о росте для четырех гипотетических респондентов**

|  |  |  |  |
| --- | --- | --- | --- |
| Вопрос, № п/пРеспондент | 1 | 2 | 3 |
| А | + | + | + |
| Б | + | + | ⎯ |
| В | + | ⎯ | ⎯ |
| Г | ⎯ | ⎯ | ⎯ |

Если респондентов и вопросы расположить на одной шкале латентной переменной (в данном случае, «роста»), то станет очевидным, что респонденты реагируют на вопросы в зависимости от своего ранга (положения) на этой же шкале: респондент данного роста, позитивно прореагировавший на некий вопрос-пункт, будет также позитивно реагировать на все пункты, имеющие более низкий ранг (в нашем примере, на все более «низкорослые» вопросы-пункты). Скажем, для примера с ростом совместное упорядочение вопросов и респондентов на гутмановской шкале могло бы выглядеть, как на рисунке 2.

Вопрос 3

Вопрос 2

Вопрос 1

Низкий рост Высокий рост

Респондент Г

Респондент В

Респондент Б

Респондент А

Рис. 2. **Шкала Гутмана для трех вопросов о росте и четырех респондентов**

Легко видеть, что в нашем примере самыми «популярными» (имеющими наибольшую частоту положительных ответов) будут менее «нагруженные» по шкалируемой переменной вопросы. С другой стороны, большее число положительных ответов будет у тех респондентов, для которых значение шкалируемой переменной выше.

Для того чтобы определить, является ли данная совокупность вопросов шкалируемой, т.е. дает ли она хорошее приближение к идеальной модели шкалирования по Гутману, нужно, во-первых, убедиться, что вопросы по своему смыслу могут соответствовать вышеописанному условию монотонного возрастания вероятности положительного ответа. Так, например, рассмотренные выше вопросы шкалы социальной дистанции Богардуса теоретически могут составить гутмановскую шкалу. Однако так называемые точечные вопросы ⎯ например: «Верно ли, что Ваш рост равен 1,65 м?» ⎯ наверняка не могут быть использованы в рамках рассматриваемой шкальной модели. Дальнейшая процедура построения гутмановской шкалы заключается в проверке соответствия реальных данных модели совершенной шкалы Гутмана и ее лучше рассмотреть на конкретном примере.

Допустим, некий социолог пытается измерить переменную «благопристойность», которую он почему-то связывает со стремлением следовать нормативным предписаниям, регулирующим публичное (т. е. безличное) социальное общение. Среди использованных им вопросов есть три следующих:

1. Вы обычно стучите в дверь кабинета или комнаты, прежде чем войти (верно, неверно).

2. В публичных обсуждениях или дискуссиях Вы высказываете свое мнение лишь тогда, когда Вас прямо об этом попросят (верно, неверно).

3. Вы предпочтете промолчать, если Вам покажется, что Ваше справедливое и существенное замечание может вызвать раздражение у окружающих (верно, неверно).

Конечно, можно предположить, что некоторые из изобретенных исследователем вопросов отражают скорее «социальную тревожность» или «конформизм», но в данном случае мы будем руководствоваться тем, что идеальных индикаторов не бывает: бывают индикаторы, дающие лучшее или худшее приближение к идеальной шкале для реальных данных.

Итак, социологу предстоит проверить, насколько полученное им эмпирическое распределение ответов соответствует тем теоретически возможным паттернам ответов на три вопроса, которые в таблице 6.2 образуют безупречный «параллелограмм», характерный для идеальной гутмановской шкалы. Для случая трех вопросов возможны четыре «правильных» паттерна ответов, обозначаемых обычно как школьные типы ответов:

социологический индекс шкала

1. + + +

2. + + ⎯

3. + ⎯ ⎯

4. ⎯ — —

Предположим, что наш социолог получил следующую картину распределения шкальных типов (см. табл. 3).

Таблица 3

##### Распределение ответов для шкальных типов

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Вопросы Ответы | Вопрос 1 («стук в дверь») | Вопрос 2 («публичная дискуссия») | Вопрос 3 («опасение вызвать раздражение») | Число случаев, N |
| Паттерн ответа («+» — «верно», «—» — «неверно»): | + | + | + | 30  |
| + | + | — | 50 |
| + | — | — | 45 |
| — | — | — | 10 |
|  | Всего 135 |

Судя по таблице 6.3, априорное упорядочение вопросов совпало с реальным: самый «легкий» первый вопрос оказался и самым популярным, тогда как на самый «тяжелый» вопрос шкалы положительно ответили лишь 30 опрошенных: нежелание высказывать свою точку зрения требует значительно большего количества «благопристойности», чем привычка стучать в дверь.

Если бы использованный нами исходный порядок вопросов не совпал бы с их реальным ранжированием по числу позитивных ответов, то это само по себе не доказывало бы «нешкалируемости» данной совокупности пунктов: для того, чтобы получить столь же красивую «гутмановскую» картину распределения ответов, как в предыдущей таблице 6.2, было бы достаточно просто переставить столбцы таблицы так, чтобы первым оказался самый популярный вопрос с наибольшим числом положительных ответов и т. д. (Упорядоченную таким образом таблицу обычно называют шкалограммной матрицей, или шкалограммой.)

Реальной проблемой в нашем примере, как и в большинстве случаев построения гутмановской шкалы, стало наличие так называемых нешкальных типов, т. е. таких паттернов ответа, которые попросту не укладываются в логику одномерной модели с монотонно возрастающей вероятностью ответа. Примером «нешкального» паттерна мог бы быть положительный ответ на третий вопрос при отрицательных ответах на первые два вопроса (— — +). То обстоятельство, что некий респондент, бесцеремонно входящий в чужую дверь без стука, боится открыто выразить свое мнение, может быть и случайной ошибкой, и результатом влияния какой-то посторонней переменной: возможно, отвечая на третий вопрос, этот человек думал не о хороших манерах, а о том, что высказывать свое мнение открыто в его привычной среде «невыгодно», недальновидно и т. п. Для того чтобы проверить шкальную гипотезу о том, что данная совокупность вопросов дает хорошее приближение к гутмановской шкале, нам следует трактовать «нешкальные» типы ответа как ошибки и оценить, насколько велико отклонение от идеальной модели. Пусть наш исследователь получил следующее распределение «нешкальных» типов (см. табл. 4).

Разумно предположить, что «нешкальный» тип — — + можно отнести к шкальному типу — — — с одной ошибкой. Второй «нешкальный» паттерн ответа — + + можно рассматривать как отклонение от школьного типа + + + также с одной ошибкой (если бы мы отнесли этот «нешкальный» паттерн к типу — — —, то ошибок было бы две, а не одна). Существуют разные способы оценки приемлемости наблюдаемых отклонений от совершенной шкалы, содержащей лишь шкальные паттерны ответа. Здесь мы воспользуемся самым простым и грубым, рассчитав коэффициент воспроизводимости шкалы Rep (от англ. reproducibility) по следующей формуле:

В нашем примере мы, основываясь на идеальной модели шкалы, можем воспроизвести (предсказать) по три ответа для 143 респондентов. Всего мы сделаем 429 предсказаний для отдельных ответов. Из них 8 ответов окажутся ошибочными (каждая ошибка будет отличаться от ожидаемого ответа только на 1 балл). Коэффициент воспроизводимости составит, таким образом, 0,98 (или 98%).

Таблица 4

##### Распределение ответов для «нешкальных» типов

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ВопросыОтветы | Вопрос 1(«стук вдверь») | Вопрос 2(«публичнаядискуссия») | Вопрос 3(«опасениевызватьраздражение») | Числослучаев |
| Паттерн ответа: | — | — | + | 3 |
| — | + | — | 5 |
|  | Всего 8 |

На практике принято считать приемлемым любое значение коэффициента воспроизводимости, которое превышает 0,90 (90%). Очевидно, что 100%-й воспроизводимостью может обладать лишь совершенная гутмановская шкала.

Если полученное значение коэффициента воспроизводимости превосходит заданный порог, данная совокупность вопросов может использоваться в качестве шкалы Гутмана. При этом вопросам присваиваются шкальные значения, отражающие их ранжирование по шкале (скажем, 1, 2 и 3), так что самый «легкий» вопрос получает самый низкий балл. Респонденты получают индивидуальный балл, соответствующий их шкальным типам (число положительных ответов либо суммарный балл).

Следует помнить о том, что полученная шкала отражает наличие определенной упорядоченности в той матрице реальных данных, для которых проверялась гутмановская модель. Иными словами, вывод о том, что данная совокупность вопросов составляет шкалу Гутмана, верен для данной выборки и для данной серии наблюдений. Перенос шкалы с одной популяции на другую требует новых данных и нового обоснования.

Мы рассмотрели лишь некоторые, относительно простые, методы конструирования индексов и шкал в социологии. Проанализированные нами примеры подтверждают полезность шкал для повышения качества социологического измерения (т. е. его надежности и валидности) и для экономного представления эмпирической информации, получаемой в ходе исследования. Наконец, анализ моделей измерения, лежащих в основании любой шкалы, часто помогает прояснить природу теоретических понятий и взаимосвязей между ними. Еще одним шагом к содержательным и основанным на реальных эмпирических наблюдениях выводам является анализ данных.

**Библиографический список**

1. Аванесов В.С. Тесты в социологическом исследовании. М.: Наука, 2010.

2. Грин Б.Ф. Измерение установки // Математические методы в современной буржуазной социологии. М.: Прогресс, 2009.

3. Девятко И.Ф. Диагностическая процедура в социологии: очерк истории и теории. М.: Наука, 2008.

4. Клигер С.А., Косолапов М.С., Толстова Ю.Н. Шкалирование при сборе и анализе социологической информации. М.: Наука, 2009.

5. Осипов Г.В., Андреев Э.П. Методы измерения в социологии. М.: Наука, 2010.

6. Толстова Ю.Н. Логика математического анализа социологических данных. М.: Наука, 2009.

7. Ядов В.А. Социологическое исследование: методология, программа, методы. 2-е изд. М.: Наука, 2010. Гл. 3.