МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»**

Кафедра социальной информатики

НОМИНАЛЬНЫЕ ШКАЛЫ: ОСОБЕННОСТИ АНАЛИЗА В ПРОГРАММЕ SPSS (НА ПРИМЕРЕ ИЗУЧЕНИЯ БЕЗРАБОТИЦЫ)

(автореферат бакалаврской работы)

Студента 5 курса 531 группы
направления 09.03.03 – Прикладная информатика,
профиль – Прикладная информатика в социологии
социологического факультета
Щекина Владимира Валерьевича

Научный руководитель
кандидат социологических наук, доцент С.В. Курганова

Заведующий кафедрой
кандидат социологических наук, доцент И.Г. Малинский

Саратов 2019 год

ВВЕДЕНИЕ

Актуальность проблемы. Измерение в социологии — один из ключевых моментов, правильное решение которого во многом определяет успех социологического исследования и качество получаемой информации. Необходимость измерения социальных характеристик объясняется как теоретико-методологическими, так и практическими соображениями. Главное практическое преимущество, которое достигается в результате измерения социальных характеристик, состоит в возможности использования математических методов анализа для дальнейшего изучения социальных явлений.

Вместе с тем, известно, что спецификой измерения социальных характеристик является их преимущественно качественный характер и, как результат, преобладание номинальных (или категориальных) шкал. «Качественность» исходных данных часто обусловливается природой соответствующей социальной характеристики. Например, такие признаки, как пол, национальная или конфессиональная принадлежность, профессия и мн. др., в принципе не могут быть измерены иначе как по номинальной шкале. Это накладывает серьезные ограничения на возможности применения традиционного математического аппарата.

В свою очередь, широкое использование номинальных шкал серьезно ограничивает исследователя в выборе методов анализа собранной информации. Решение этой задачи осуществляется с помощью теории измерений. Такой подход отвечает пониманию измерения как моделирования с помощью чисел. Построение шкалы с учетом этой теории позволяет конструктивно выделить моменты реальности, отражаемые в числовой модели, и отделить от них моменты, от которых исследователь в процессе моделирования абстрагируется. Это дает возможность, изучая интересующие социолога процессы, эффективно учитывать диалектическое единство их качественных и количественных

сторон, что мы и попытаемся продемонстрировать в данной работе на примере проблемы безработицы.

Степень научной разработанности. Социологические аспекты проведения эмпирического исследования изучаются социологами довольно пристально в течение длительного времени. Еще основатель социологии О. Конт заявил о необходимости эмпирической проверки теорий, возникающих в новой науке об обществе. Известнейшие социологи классического периода Э. Дюркгейм и М. Вебер следовали тем же принципам, в том числе и на практике, написав свои уже всемирно известные работы.

Ирландские статистики Дж. Граунт и У. Петтипо праву считаются основателями так называемой политической арифметики – направления социальных исследований официальных Данных для целей оптимизации государственного управления. Фундамент социальной статистики был заложен в XIX в. франко-бельгийским ученым Адольфом Кетле, открывшим ряд статистических закономерностей общественной жизни. Он одним из первых стал использовать в статистике достижения математики и ее методы, прежде всего теорию вероятности, которую применил в процессе анализа социальных явлений. Не отставали и отечественные исследователи. Так, в работе А.Г. Ковалевского были изложены основные идеи, положившие начало современной теории выборки. Мировую известность имел А.А. Чупров, фамилия которого дала название одному из наиболее часто использующихся в наше время коэффициентов парной связи между номинальными признаками.

С тех пор интерес исследователей к использованию математического аппарата в социологии не снижается. Только среди отечественных социологов разработкой общеметодологических вопросов проведения социологического исследования занимаются Ядов В.А., Здравомыслов А.Г., Батыгин Г.С., Шубкин В.Н., анализом конкретных методов сбора эмпирической информации – Белановский С.А., Исупова О., Козина И., проблемой сопряжения социологической информации и компьютерных технологий – Крыштановский А.О., Божков О.Б., проблемами измерения в социологии – Девятко И.Ф.,

Маслова О.М., Паниотто В.И., математическими методами обработки социологических данных – Андреенков В.Г., Толстова Ю.Н., Татарова Г.Г.

В то же время даже в специализированной литературе специфике обработки и анализа переменных с номинальными шкалами уделяется относительно малое внимание. Среди таких публикаций можно назвать работы Аптона Г., Аргуновой К.Д., Мирзоева А.А., Толстовой Ю.Н., что говорит об актуальности данного направления исследования.

Объектом исследования выступают номинальные шкалы; **предметом** – построение математической модели с помощью номинальных шкал (на примере изучения безработицы).

Цель исследования – раскрыть эвристический потенциал номинальных шкал, выявить возможности и ограничения их использования в процессе обработки социологической информации.

Для достижения цели были сформулированы следующие **задачи**:

1. определить специфику номинальных шкал как инструмента измерения социальных признаков;

2. раскрыть веер инструментов, предлагаемых в программе SPSS, для работы с переменными, имеющими номинальные шкалы;

3. разработать математическую модель безработицы на базе программы статистической обработки SPSS.

Эмпирическая база – результаты социологического исследования на тему «Безработица в условиях малого города», проведенного в 2016 г. и представленного в электронном виде на базе программы статистической обработки данных SPSS 17.0[1].

Научная новизна заключается в раскрытии принципа взаимодополнительности аналитических методов, работающих с номинальными шкалами.

[1] База данных содержит информацию по социологическому исследованию, проведенному в 2016 г. методом стандартизированного опроса, в ходе которого были проанкетированы безработные г. Новоузенска. Выборочная совокупность построена по принципу целевой, насчитывает всего 87 респондентов.

Структура работы. Данная работа состоит из введения, трех разделов (1 раздел «Номинальные шкалы как инструмент измерения социальных характеристик», 2 раздел «Возможности программы SPSS для анализа данных, представленных в формате номинальных шкал», 3 раздел «Практический пример анализа номинальных шкал с помощью программы SPSS на примере исследования безработицы»), заключения, списка использованных источников и приложений.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Номинальные шкалы как инструмент измерения социальных характеристик» посвящен раскрытию основных проблем конструирования шкал в социальных исследованиях и характеристике номинальной шкалы.

Шкала в социальном исследовании – это способ упорядочения эмпирической социальной информации, система чисел и отношений между ними, изоморфная ряду измеряемых социальных фактов.

Номинальная шкала является наиболее простой процедурой измерения, единственная разрешенная для нее процедура измерения – классификация по категориям.

Критерии адекватности измерения номинальных переменных: категории должны быть взаимоисключающими, исчерпывающими, однородными, числа, кодирующие категории, выступают лишь в качестве меток и не несут в себе численного значения.

Номинальные шкалы – это шкалы, допустимыми преобразованиями которых являются произвольные взаимно однозначные преобразования, т.е. преобразования, сохраняющие отношения равенства и неравенства между числами.

Возможные операции с числами для категориальной шкалы:

1. Нахождение частот распределения по пунктам шкалы с помощью процентирования или в натуральных единицах;

2. Поиск средней тенденции по модальной частоте (группе с наибольшей численностью).

3. Установление взаимосвязи между рядами свойств, расположенных неупорядоченно, для чего составляют перекрестные таблицы, критерий независимости Хи-квадрат Пирсона, логлинейный анализ, логистическую регрессию.

Во **втором разделе «Возможности программы SPSS для анализа данных, представленных в формате номинальных шкал»** описываются основные аналитические методы работы с номинальными шкалами, предоставляемые программой SPSS.

Таблица сопряженности представляет совместное распределение двух переменных, предназначенное для исследования связи между ними. Таблица сопряженности является универсальным средством изучения статистических связей, так как в ней могут быть представлены переменные с любым уровнем измерения.

Критерий независимости хи-квадрат часто применяется при проверке гипотез в социологии, поскольку для его использования требуется, чтобы переменные имели номинальные шкалы, также он является непараметрическим. Критерий независимости хи-квадрат отвечает лишь на вопрос о том, являются ли результаты статистически значимыми или нет, тогда как о силе связи между признаками ничего сказать не может. Для этого надо обратиться к особым статистическим методам - мерами связи - Ф Фишера и V Крамера. Данные показатели вычисляются на основе критерия хи-квадрат.

Логлинейный анализ применяется для анализа таблиц сопряженности нескольких категориальных признаков. Если анализируется сопряженность двух признаков, то вполне достаточно применения критерия Хи-квадрат. Однако зачастую данные содержат существенно большее число категориальных признаков, и тогда визуальный анализ таблиц сопряженности становится невозможным.

Логлинейная модель представляет собой множественную регрессию, в которой категориальные переменные и их взаимодействия выступают в качестве предикторов, а роль зависимой переменной играет натуральный логарифм частот категорий.

Логистическая регрессия представляет собой расширение множественной регрессии и отличается тем, что в качестве зависимой переменной используется дихотомическая переменная, имеющая два возможных значения, которые символизируют принадлежность или непринадлежность объекта какой-либо группе.

Смысл регрессионного анализа в нахождении математического выражения, наиболее адекватно отражающего связь между зависимой переменной и несколькими независимыми переменными. Также логистическая регрессия прогнозирует вероятность некоторого события, находящуюся в пределах от 0 до 1.

Третий раздел «Практический пример анализа номинальных шкал с помощью программы SPSS на примере исследования безработицы» описывается пример использования методов Хи-квадрат, логлинейный анализ и логистическая регрессия для описания модели безработицы в условиях малого города.

Построение таблиц сопряженности и анализ методом критерия независимости Хи-квадрат показал, что среди краткосрочно безработных доминируют женщины среднего возраста семейные и имеющие детей. Также доминирующими в данной группе являются такие признаки, как достаточный уровень семейного потребления, наличие профессионального образования и трудового стажа более 3 лет, а также личные причины потери работы. Среди долгосрочных безработных, напротив, преобладают мужчины молодого возраста, не женатые и без детей, с уровнем образования не выше школьного. Также для них характерны скудный уровень потребления в семье, отсутствие или незначительность трудового стажа и проблемы с образованием как причина безработицы.

Таким образом, было выявлено много зависимостей, причем во многих случаях были высказаны предположения о латентном влиянии других переменных. В данном случае эффективным аналитическим инструментом является метод логлинейного анализа, который позволяет исследовать связи более высоких уровней. В нашем случае мы подвергли анализу все 9 исследуемых нами переменных.

Полученная в ходе проведения логлинейного анализа модель содержит статистически значимые взаимодействия первого, второго и третьего порядка. В нашем случае являются статистически значимыми главные эффекты таких переменных, как возраст, наличие детей, уровень образования, уровень потребления семьи, длительность поиска работы, трудовой стаж и причины безработицы. Кроме отдельных переменных, к значимым можно отнести одно трехфакторное (семейное положение*уровень потребления*причины безработицы) и 7 двухфакторных взаимодействий (возраст*наличие детей, семейное положение*наличие детей, пол*уровень потребления, возраст*трудовой стаж, семейное положение*трудовой стаж, наличие детей*причины безработицы, уровень образования*причины безработицы).

Интерпретировать частичные ассоциации можно двумя способами. Первый предполагает построение таблиц сопряженности. Особенно удобно им пользоваться при интерпретации двухфакторных взаимодействий.

Второй способ интерпретации связан с анализом оценок параметров модели и является более сложным. Поскольку сумма всех параметров должна быть равна нулю, на основе полученных данных можно вычислить недостающую информацию. Например, в нашем случае видно, что среди семейных безработных с достаточным уровнем потребления в семье причиной безработицы чаще выступают проблемы с образованием, тогда как со скудным уровнем потребления – экономический кризис. По личным же причинам безработицы семейные респонденты с разным уровнем потребления практически не различаются. Холостые же безработные с достаточным уровнем потребления в семье чаще в качестве причины своей безработицы называют

экономический кризис, тогда как респонденты со скудным уровнем потребления, напротив, отмечают проблемы с образованием. По личным причинам данные группы респондентов также практически не различаются.

Результаты обратного пошагового исключения являются наиболее важными среди остальных частей логлинейного анализа, т.к. позволяют сформировать понимание наиболее состоятельной модели. В нашем случае была сформирована иерархическая модель, в которую вошли 36 взаимодействий, каждое из которых включает по 7 переменных, что невозможно интерпретировать простым способом при помощи таблиц сопряженности. В этом случае будет полезна множественная логистическая регрессия.

Основным результатом логистической регрессии являются бета-коэффициенты, нужные при построении уравнения регрессии и показывающие значимость каждой включенной в уравнение переменной. Из 6 независимых переменных рассматриваемого примера только три показали статистическую значимость: «Наличие детей», «Уровень потребления» и «Трудовой стаж», причем если первые две переменные показали значимое статистическое влияние.

Также бета-коэффициенты отражают влияние выбранных предикторов на зависимую переменную. Так, респонденты, находящиеся в поиске работы более 6 месяцев, как правило, не имеют детей, имеют трудовой стаж менее трех лет или совсем не имеют трудового стажа, при этом располагают скудным уровнем семейного потребления. Тогда как респонденты, ищущие работу менее 6 месяцев, напротив, как правило, детные, со значительным трудовым стажем и хорошим уровнем потребления в семье.

ЗАКЛЮЧЕНИЕ

Номинальные шкалы являются очень распространенным способом измерения признаков изучаемых социальных явлений. Например, такие признаки, как пол, национальность, профессия в принципе не могут быть измерены иначе как по номинальной шкале. Это существенно ограничивает

возможности применения традиционного математического аппарата, поскольку номинальная шкала является самым «примитивным» инструментом измерения. Тем не менее, современная социальная статистика располагает весьма эффективным и разнообразным инструментарием для обработки номинальных шкал.

По номинальной (или категориальной) шкалой понимается шкала, допустимыми преобразованиями которой являются произвольные взаимно однозначные преобразования, т.е. преобразования, сохраняющие отношения равенства и неравенства между числами.

При этом надо отметить, что существуют критерии, обеспечивающие адекватность измерения при помощи номинальных шкал. Это взаимоисключаемость, исчерпываемость и однородность категорий переменных, а также запрет на непосредственную математическую обработку данных, полученных при помощи номинальных шкал. Данные особенности номинальных шкал накладывают серьезные ограничения на возможности их математической обработки. Исследователь, работающий с номинальными шкалами, располагает следующим перечнем допустимых операций: подсчет частоты встречаемости каждой категории, поиск средней тенденции по модальной частоте, установление взаимосвязи между признаками с помощью перекрестных таблиц и статистических показателей (критерия независимости хи-квадрат, коэффициентов корреляции Ф Фишера, V Крамера), логистическая регрессия, логлинейный анализ таблиц сопряженности и некоторые другие.

Применение критерия независимости Хи-квадрат Пирсона при анализе данных социологического исследования безработных г. Новоузенска показал, что среди долгосрочных безработных две трети составляют мужчины, что является несколько неожиданным результатом. Молодежь также имеет более высокие шансы на долгосрочную безработицу. Среди краткосрочных безработных несколько преобладают семейные и детные респонденты, тогда как среди долгосрочных безработных – холостые и бездетные. У респондентов без профессионального образования гораздо больше шансов тратить на поиски

работы более 6 месяцев, чем у опрошенных с профессиональным образованием. При длительных сроках безработицы хотя бы одного члена семьи ее ресурсы оскудевают и уже не позволяют в полной мере удовлетворять основные потребности семьи. Наличие трудового стажа более 3 лет в значительной степени сокращает длительность поиска работы безработными. Наконец, для безработных, ищущих работу менее 6 месяцев, характерны личные причины потери работы, тогда как для безработных, находящихся в поиске работы более 6 месяцев, более свойственны проблемы с образованием.

Таким образом, было выявлено много зависимостей, причем во многих случаях были высказаны предположения о латентном влиянии других переменных. Если критерий независимости Хи-квадрат позволяет выявлять связи между исследуемыми переменными первого уровня (между двумя переменными), то для описания связей между тремя и более переменными более эффективным аналитическим инструментом является метод логлинейного анализа.

Проведение логлинейного анализа подтвердило статистическую значимость главных эффектов переменных возраст, наличие детей, уровень образования, уровень потребления семьи, длительность поиска работы, трудовой стаж и причины безработицы. Кроме отдельных переменных, значимым оказалось одно трехфакторное (семейное положение*уровень потребления*причины безработицы) и 7 двухфакторных взаимодействий (возраст*наличие детей, семейное положение*наличие детей и др.)

Возможности логлинейного анализа хорошо раскрываются при решении задачи интерпретации трехфакторной модели. Так, было выявлено, что среди семейных безработных с достаточным уровнем потребления в семье причиной безработицы чаще выступают проблемы с образованием, тогда как со скудным уровнем потребления – экономический кризис. По личным же причинам безработицы семейные респонденты с разным уровнем потребления практически не различаются. Холостые же безработные с достаточным уровнем потребления в семье чаще в качестве причины своей безработицы называют

экономический кризис, тогда как респонденты со скудным уровнем потребления, напротив, отмечают проблемы с образованием. По личным причинам данные группы респондентов также практически не различаются.

Еще одним способом интерпретации связи между тремя и более переменными с дихотомическими шкалами является логистическая регрессия. Главным результатом данного метода являются бета-коэффициенты, полезные при построении уравнения логистической регрессии, и значимости каждой включенной в уравнение переменной. В нашем случае из 6 независимых переменных только три показали статистическую значимость: «Наличие детей», «Уровень потребления» и «Трудовой стаж». Анализ бета-коэффициентов показал, что респонденты, находящиеся в поиске работы более 6 месяцев, как правило, не имеют детей, имеют трудовой стаж менее трех лет или совсем не имеют трудового стажа, при этом располагают скудным уровнем семейного потребления. Тогда как респонденты, ищущие работу менее 6 месяцев, напротив, как правило, детные, со значительным трудовым стажем и хорошим уровнем потребления в семье.

Подводя итоги проделанной работы, отметим, что даже при работе с самыми простыми номинальными шкалами исследователи могут использовать довольно широкий спектр методов, выявляющих взаимосвязи между переменными, причем с их помощью возможно построение сложных многомерных моделей. Одним из таких методов является логлинейный анализ, направленный на анализ таблиц сопряженности, включающих три и более переменных. Применение другого метода, логистической регрессии, позволяет проинтерпретировать сложные модели взаимодействия переменных, полученных при помощи логлинейного анализа.