Министерство образования и науки Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра социальной информатики

ВЗАИМОДОПОЛНИТЕЛЬНОСТЬ АНАЛИТИЧЕСКИХ ОПЕРАЦИЙ ПРИ ИСПОЛЬЗОВАНИИ ПРОГРАММЫ SPSS

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 5 курса 531 группы

направления 09.03.03 - Прикладная информатика
профиль подготовки - Прикладная информатика в социологии
Социологического факультета
Утегалиевой Людмилы Ивановны

Научный руководитель
кандидат философских наук, доцент А.И. Завгородный

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

Саратов, 2016 год

Введение

Актуальность проблемы. Согласно истории развития эмпирической социологии, на середину XX столетия приходится расцвет массовых социологических исследований. Социолог Н.И. Лапин отмечает, что первые десятилетия после Второй мировой войны являются периодом развитых эмпирических социологических исследований. Конечно, достижение таких успехов эмпирической социологией было обусловлено всей предшествовавшей историей ее развития, разработкой теоретико-методологических оснований социологических исследований в целом, совершенствованием методов сбора информации, накоплением практического опыта организации исследований и т.д. Однако был еще один фактор, который внес свой вклад в расцвет эмпирических исследований, а именно появление и бурное развитие специализированных компьютерных программ, позволивших резко повысить качество сбора, хранения и обработки результатов проведенных исследований.

В 1965-м году американские студенты Норманн Най и Дейл Вент попытались найти компьютерную программу, с помощью которой можно было бы проанализировать статистическую информацию. Они перебрали все имевшиеся на тот момент программы, но ни одна из них не показалась им более или менее пригодной. Тогда студенты решили разработать собственную программу со своей единой концепцией и единым синтаксисом. Через год была готова первая версия программы, а еще через год – версия, которая смогла работать на IBM 360. ТакбыласозданапрограммаSPSS.За это время кроме SPSS на рынке появились и другие программы, также позволяющие решить проблему анализа статистических данных. Но SPSS до сих пор является одной из самых популярных программ, использующейся, в том числе, и на нашем социологическом факультете.

Краткая характеристика материалов исследования. Такая популярность программы не могла не вызвать публикацию многочисленных учебников, пособий и руководств по работе с SPSS. Конечно, первые работы, посвященные особенностям анализа социологических данных при помощи

компьютерных технологий, были переводными. Среди авторов таких работ можно назвать А. Бююля, П. Цефеля, Дж. Хили. Работы отечественных исследователей появились вскоре после публикации первых пособий зарубежных авторов по этой тематике и вскоре стали довольно разнообразными. Областью статистической обработки информации заинтересовались социологи, психологи, экономисты и мн.др.

Интерес исследователей был сосредоточен не только на SPSS как одной из наиболее популярных программ для статистической обработки данных, но и на других программных продуктах[1]. Появились попытки реализации комплексного соединения информационных технологий и социологии[2], а также теоретико-методологического осмысления проблем компьютерной поддержки социологического эмпирического исследования[3]. Тем не менее, подавляющее большинство работ содержат, в основном, лишь алгоритмы проведения основных статистических процедур. Попыток более или менее системного описания возможностей программных продуктов, в том числе и SPSS, практически не представлено, что делает нашу работу весьма актуальной

Цель исследования – раскрыть возможность взаимодополнения аналитических операций, имеющихся в программе SPSS, обусловленного, с одной стороны, разными возможностями отдельных методов, а с другой – особенностями исторического развития самой программы SPSS (на примере факторного и кластерного методов, критерия независимости Хи-квадрат Пирсона и метода сравнения средних). Постановка цели определила формулирование следующих **задач** исследования:

1. описать аналитические возможности программы SPSS, а также выявить особенности ее структуры, обусловленные, в том числе, историей развития данной программы;

---

[1] Макарова Н.В., Трофимец В. Я. Статистика в Excel: Учебное пособие. М.: Финансы и статистика, 2002. 368 с.

[2] Борисова С.Ф. Компьютер и интернет для социолога. Учебное пособие – справочник. Н. Новгород, 2002. 125 с.

[3]Божков О.Б. Компьютерные технологии в социологическом исследовании // Социологический журнал. 1998. № 1-2. С. 95-112.

2. сравнить некоторые методы анализа данных, представленных в SPSS, с целью выявления их потенциала взаимодополнительности (на примере факторного и кластерного анализа, критерия независимости Хи-квадрата Пирсона и метода сравнения средних).

Объектом данного исследования является программа обработки статистических данных SPSS; **предметом** выступает сопряженность методов анализа социологической информации, параметры которой соответствуют определенной математической модели.

В качестве **эмпирической базы** исследования были использованы данные массового социологического опроса, проведенного студентами социологического факультета СГУ по теме «Особенности профессиональной деятельности государственных служащих»[1].

Структура диплома. Данная работа состоит из введения, трех разделов, заключения, списка использованных источников и приложений.

Основное содержание работы. В первом разделе «Программа SPSS как инструмент статистического анализа социологической информации» приводитсяхарактеристика программы SPSS и ее аналитических возможностей.Как мы отметили ранее, SPSS является одной из старейших систем статистического анализа и управления данными.За полувековой период времени программа SPSS претерпела множество нововведений и изменений.

Первые версии программы содержали всего одиннадцать статистических процедур, но SPSS уже нашла отклик у потребителя. К моменту основания Норманом Наем фирмы было проведено уже шестьдесят инсталляций.

В течение следующих пяти лет велась активная работа по адаптации программы под разные операционные системы и расширению числа

[1] Данное социологическое исследование было проведено в 2012 году студентами социологического факультета СГУ. Метод сбора информации – стандартизированное анкетирование. Выборка строилась по принципу целевой (объектом исследования выступали государственные служащие, работающие в органах власти муниципального и регионального уровней). Всего было опрошено 100 респондентов.

статистических процедур. Уже в 1975 году была разработана 6 версия программы и проведено 600 инсталляций.

В 1983 году программа SPSS претерпела революционные изменения:был переработан командный язык; увеличилось число статистических процедур; была выпущена версия SPSS\PC+ для установки на персональные компьютеры; появилось европейское представительство компании.В результате SPSS стала самым популярным программным обеспечением для статистического анализа информации во всем мире.

Следующей важной вехой на пути развития SPSS стал выпуск новой версии SPSS для операционной системы Windows (SPSSforWindows). С одной стороны, в новой версии программы сохранились исходные возможности для больших ЭВМ, с другой – теперь программой могли пользоваться исследователи, не специализирующиеся в области прикладного программирования, причем делать это без особых затруднений. С тех пор программа продолжала совершенствоваться: была разработана концепция Viewer (Окно просмотра), технология мобильных таблиц, улучшился дизайн таблиц и графиков. Вместе с тем, с усложнением программы стали проявляться и ее слабые стороны: высокие требования к возможностям компьютера.

В настоящее время новейшей версией программы является IBMSPSSStatistics23.0, которая предоставляет еще больше возможностей для работы с информацией. Эргономичный интерфейс программы содержит все функции, необходимые для управления данными, процедуры для статистического анализа высокого уровня сложности, а также широкие возможности для написания отчетов по итогам проведенной работы. Таким образом, в настоящее времяIBMSPSSStatistics является полнофункциональной системой, направленной на решение задач в сфере бизнеса и науки с помощью анализа данных. Она позволяет глубоко и всесторонне исследовать данные, наглядно представлять итоговую информацию в виде таблиц и графиков.

SPSS выстроена как традиционная база данных. Сближает SPSS с рядовыми базами данных и особенности интерфейса, что также делает более

простым первое знакомство с программой. При этомSPSS специализируется на обработке результатов опросов, т.е. имеет структурные особенности, заключающиеся в принципах формализации накопляемого массива исходной информации, принципах статистической обработки и представления результатов информации.

Второй раздел «Факторный и кластерный методы исследования данных: сравнительная характеристика» приводятся результаты сравнительного анализа аналитических возможностей факторного и кластерногометодов.

Метод факторного анализа впервые был разработан психологом Ф. Гамильтоном, затем он стал успешно использоваться в других науках. Выделяют две основные цели факторного анализа: редукции переменных или сокращения их числа и выявления особенностей взаимосвязи между исследуемыми переменными или их классификации.Таким образом, факторный анализ позволяет дать всестороннее и одновременно компактное описание объекта изучения. Для этого в ходе факторного анализа выявляются латентные переменные или факторы, которые отвечают за наличие линейных корреляционных связей между наблюдаемыми переменными. Надежность результатов, полученных в ходе применения факторного анализа, тесно связана со степенью соответствия исследуемых эмпирических данных математической модели, лежащей в основе данного метода.

Процедура факторного анализа состоит из четырех стадий:

1. Вычисление корреляционной матрицы для всех переменных, участвующих в анализе;

2. Извлечение факторов;

3. Вращение факторов;

4. Интерпретация факторов.

В качестве эмпирических данных для проведения факторного анализа были выбраны переменные, содержащие оценку чиновниками тех или иных сторон свой работы по 5-балльной шкале.

В результате проведения факторного анализа при помощи Обобщенного метода наименьших квадратов с вращением факторов по методу Варимакса было получено решение, состоящее из 3 факторов. Первый фактор, получивший название *«Высокая социальная статусность»*, содержалпеременные, характеризующие работу как престижную (сила связи с фактором равна 0,83), хорошо оплачиваемую (0,77), ответственную (0,72), умственно тяжелую (0,65). Второй фактор *«Высокая социальная значимость»* включил переменные, описывающие работу как гарантированную (0,73), необходимую для общества (0,70), предоставляющую возможности для профессионального развития и самореализации (0,70), имеющую перспективы роста (0,64) и интересную по содержанию (0,52). К третьему фактору *«Благоприятный социально-психологический климат на работе»* были отнесены хорошие отношения с администрацией (0,95) и хорошие отношения с коллективом (0,83).

Кластерный анализ–это набор многомерных статистических методов, нацеленных на исследование структуры некоторой совокупности переменных или объектов. Соответственно, можно говорить о кластерном анализе переменных и кластерном анализе объектов.Главной задачей кластерного анализа переменных заключается в переходе от первоначальной совокупности множества переменных к значительно меньшему числу кластеров.

Кластерный анализ переменных реализуется с помощью следующих этапов:

1.Отбор выборки для кластеризации;

2. Выбор способа измерения расстояния между переменными;

3. Применение метода кластерного анализа для создания групп сходных переменных;

4. Проверка достоверности результатов кластерного решения.

Таким образом, кластерный анализ является эффективным и простым методом классификации. К его основным преимуществам можно отнести отсутствие ограничений на нормальное распределение переменных;

возможность классификации в случаях отсутствия априорной информации о классах; универсальность (применимость и к объектам, и к переменным).

С помощью иерархического кластерного анализа переменных мы получили несколько решений, содержащих 2, 3 и 4 кластера. В результате было выбрано 4-кластерное решение, хотя и отличающееся от факторного. В пользу данной модели говорит возможность ее более легкой и логически непротиворечивой интерпретации: первый кластер можно определить как *«Высокая социальная статусность»*, второй – *«Значительные возможности для профессионального развития»*, третий – *«Высокая социальная значимость»* и четвертый – *«Благоприятный социально-психологический климат на работе»*.

Третий раздел «Особенности анализа данных с использованием критерия независимости Хи-квадрат Пирсона и метода сравнения средних)» содержит сравнительную характеристику критерия независимости Хи-квадрат Пирсона и метода сравнения средних.Критерий независимости Хи-квадрат Пирсона является одним из самых популярных статистических критериев: для его применения достаточно, чтобы у исследуемых переменных были номинальные шкалы; его можно применять для анализа переменных с любым типом шкалы; его применение не ограничивается требованием нормальности распределения.

Концепция независимости, которая лежит в основании логики применения Хи-квадрата, предполагает, что две переменные будут считаться независимыми, если отнесение наблюдения к той или иной категории первой переменной не будет влиять на вероятность того, что данное наблюдение окажется в определенной категории второй переменной.Если различия между ожидаемыми и наблюдаемыми частотами оказываются минимальными, подтверждается нулевая гипотеза, утверждающая независимость исследуемых переменных друг от друга. Если же различия оказываются достаточно большими, нулевая гипотеза опровергается. Причем чем выявленные различия больше, тем вероятнее опровержение нулевой гипотезы и подтверждение альтернативной.

Хи-квадрат Пирсона имеет следующие ограничения:

1. у каждой переменной не должно быть более4-х категорий;
2. его надо осторожно применять при очень малой выборке;
3. а также при большой выборке из-за чувствительности к ее объему;
4. он может ответить лишь на вопрос о наличии или отсутствии связи между исследуемыми переменными. Для оценки силы связи между переменными надо обратиться к мерам связи–Ф Фишера илиVКрамера.

Применение критерия независимости Хи-квадрат Пирсона выявил влияние фактора «Социальное положение» на 7 переменных из 14-ти, характеризующих работу. Оказалось, что чиновники, занимающие руководящую должность, чаще характеризую свою работу как ненормированную, ответственную, перспективную, гарантированную, престижную, предоставляющую большие возможности для профессионального развития и самореализации, чем рядовые сотрудники. Последние же чаще отмечают физическую тяжесть своей работы, по сравнению с руководителями.

Сравнение средних является одним из наиболее часто используемых методов статистического анализа и включает в себя несколько вариантов обследования данных. Для сравнения оценок различных сторон работы, данных представителями двух независимых выборок: руководителями и рядовыми исполнителями, применяетсяt-тест для двух независимых выборок.

Логика применения теста следующая. Сначала для обследуемых выборок вычисляются средние значения и стандартные ошибки. Затем по t-критерию определяется статистическая значимость их различия. В случае попадания t-критерия в «зону частых значений» считается, что нулевая гипотеза об отсутствии различий подтверждается, и, следовательно, делается вывод об отсутствии различий между выборками. В случае же попадания t-критерия в «зону редких значений» нулевая гипотеза отклоняется, и делается вывод о наличии различий между выборками.

Ограничения применения t-тест для двух независимых выборок:

1. Зависимая переменная должна иметь количественную или условно количественную шкалу с нормальным распределением;

2. Независимая (группирующая) переменная должна быть дихотомической или делить выборочную совокупность на две выборки;

3. Дисперсии в двух сравниваемых выборках должны быть равны.

В результате применения t-теста были выявлены статистически значимые различия по 8 зависимым переменным из 14-ти:

руководители оценивают степень ненормированности своего рабочего дня значительно выше (в среднем на 2,64 балла), чем рядовые служащие (в среднем на 2,08 балла); руководители считают свою работу более интересной (4,17 баллов), чем рядовые сотрудники (3,75 баллов); в среднем руководители считают свою работу более значимой (4,61 балла), чем их подчиненные (4,33 балла); руководители оценивают свои перспективы роста в среднем на 4,28 баллов, тогда как рядовые сотрудники – лишь на 3,33 балла; средние оценки гарантированности работы руководителей выше (4,09 баллов), чем у рядовых сотрудников (3,68 баллов); для руководителей средний уровень престижности составил 4,47 баллов, для рядовых сотрудников – 4,08 баллов; руководители оценили уровень оплаты своего труда на 4,41 балла в среднем, исполнители – только на 3,95 баллов; наконец, руководители более высоко оценивают свои возможности по профессиональному развитию (в среднем на 4,39 баллов) по сравнению с рядовыми сотрудниками (в среднем на 3,65 баллов).

Заключение. Статистический пакет IBMSPSSStatistics является одним из старейших среди программных продуктов, ориентированных на анализ информации из разных областей знания, и доступным как для специалистов в программировании и математике, так и для тех, кто специализируется в других областях. Несомненными достоинствами программы являются навигация с помощью мыши и диалоговых окон, сопряженность с другими распространенными программами, модульная структура и мн. др.

Особая привлекательность этого пакета для социологов связана с тем, что SPSS предлагает широкий выбор методов анализа переменных с разными

шкалами. В том числе здесь имеются большие возможности анализа номинальных и порядковых шкал, которые оказались «крепкими орешками» для статистики, но являются очень распространенными в обществознании. Вместе с тем, длительная история развития SPSS, заключающаяся в расширении аналитических возможностей программы через добавление все новых и совершенствование уже существующих модулей, привела к тому, что в настоящее время программа предоставляет пользователю не просто широкие возможности, а чрезмерно широкие. С одной стороны, это делает более сложным процесс освоения программы для начинающих исследователей, с другой – уже продвинутый пользователь может выбрать для себя наиболее удобный метод и алгоритм действий.

В качестве одного из примеров анализа статистической информации при помощи разных методов можно привести поиск латентных переменных с помощью факторного анализа и иерархического кластерного анализа переменных. Оба указанных метода являются эффективными инструментами для решения данной задачи. При этом отметим, что действия, выполняемые в ходе статистических операций в каждом из методов, принципиально различаются. Если факторный анализ нацелен на вычленение дисперсии, объясняемой коррелирующими наблюдаемыми переменными, то кластерный анализ осуществляется за счет подсчета расстояния между наблюдаемыми случаями. Решение, предлагаемое каждым методом, зависит от того, какие меры связи между наблюдаемыми переменными были выбраны для расчетов. Тем не менее, итоговый набор латентных переменных (факторов и кластеров), как правило, совпадает. Поэтому с целью обеспечения более тщательного контроля над переменными исследователю целесообразно применять оба метода. В ходе авторского исследования использование факторного и иерархического кластерного методов анализа привело к выявлению одинаковой структуры латентных переменных, влияющих на оценку тех или иных сторон работы государственными чиновниками. Однако данный вариант решения оказался довольно сложно интерпретируемым. Поэтому в качестве итогового

решения была выбрана модель, построенная с помощью кластерного анализа и включающая4 кластера: «Высокая социальная статусность», «Значительные возможности для профессионального развития», «Высокая социальная значимость» и «Благоприятный социально-психологический климат на работе».

Второй пример, приведенный в нашем исследовании, демонстрирует возможности и ограничения применения критерия независимости Хи-квадрата Пирсона и метода сравнения средних на примере t-теста для двух независимых выборок. Хи-квадрат Пирсона и производные от него коэффициенты корреляции являются достаточно мощным инструментом анализа взаимосвязи между переменными с номинальными шкалами. Применение данного метода позволило нам прийти к выводу о наличии слабой, но статистически значимой связи между социальным положением респондента и такими оценками их работы, как степень физической тяжести, ненормированность, ответственность, наличие перспектив роста, гарантированность, престижность и наличие возможностей для профессионального роста и самореализации. Наиболее сильная корреляционная связь была отмечена между переменными «Социальный статус респондентов» и «Работа имеет перспективы роста»: руководители чаще указывали на наличие перспектив по сравнению со своими подчиненными. Применение же t-теста позволило выявить статистически значимые различия в выборках руководителей и рядовых работников по таким переменным, как ненормированность работы, ее интересное содержание, необходимость для общества, наличие перспектив роста, гарантированность, престижность, хорошая оплата труда и возможности для профессионального роста и самореализации. Было выявлено, что руководители в среднем дают более высокие оценки по всем указанным параметрам, по сравнению с рядовыми работниками, причем наиболее значимое статистическое различие характеризует переменные «Работа имеет перспективу роста» и «Работа предоставляет возможности для профессионального роста и самореализации». Сравнение результатов применения этих методов анализа позволило нам сделать вывод о том, что оба метода обладают серьезным эвристическим

потенциалом. Не смотря на то, что критерий Хи-квадрат применяется для анализа переменных с номинальными шкалами, а t-тест– для переменных с разными шкалами (категориальной и количественной), возможности трансформации переменных, встроенные в программу SPSS, позволяют использовать данные методы в паре. На эффективность их совместного использования для исследования данных указывают полученные результаты: пять переменных из десяти, показавших наличие статистически значимой связи или различия, были выделены обоими методами. Причем если t-тест дает средние значения по сравниваемым выборкам и оценку значимости их различия, то Хи-квадрат со своими производными указывает, в том числе, и на силу связи между переменными.