БАЛТИЙСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ им. ИММАНУИЛА КАНТА

А. В. Щекотуров

СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ В ИССЛЕДОВАТЕЛЬСКОЙ ДЕЯТЕЛЬНОСТИ СОЦИОЛОГА

Учебное пособие

Издательство
Балтийского федерального университета им. Иммануила Канта
2023

УДК 316
ББК 60
Щ406

Рецензенты

Г. Л. Воронин, доктор социологических наук, профессор кафедры общей социологии и социальной работы, Нижегородский государственный университет им. Н.И. Лобачевского;
А. К. Леонов, кандидат социологических наук, доцент кафедры философии и социологии, Амурский государственный университет

Щекотуров, А. В.
Щ406 Статистический анализ данных в исследовательской деятельности социолога : учебное пособие / А. В. Щекотуров. — Калининград : Издательство БФУ им. И. Канта, 2023. — 113 с.
ISBN 978-5-9971-0799-4

Рассмотрены основные методы управления данными и статистические процедуры анализа количественной информации в программе IBM SPSS Statistics. В каждой главе представлены примеры использования программы, в которых продемонстрирован весь алгоритм исследовательской работы социолога от постановки проблемы до интерпретации обнаруженных связей.

Предназначено для студентов, обучающихся по направлению «Социология». Может быть полезно преподавателям, а также специалистам, занимающимся анализом количественных данных в социологии.

УДК 316
ББК 60

ISBN 978-5-9971-0799-4

ОГЛАВЛЕНИЕ

Посвящается моему главному учителю статистики профессору социологии Шаньянг Зао

Dedicated to my greatest teacher of statistics Professor of Sociology Shanyang Zao

ВВЕДЕНИЕ

Цель данного пособия — структурировать содержание учебного курса «Современные статистические пакеты (SPSS)», который читается студентам по направлению подготовки 39.03.01 «Социология» в Балтийском федеральном университете им. И. Канта. Однако одной структуры (какой бы логичной и систематизированной она ни была) не достаточно для полноценного освоения незаменимых в социологической деятельности компетенций по обработке и анализу количественных данных. В связи с этим в учебном пособии самым скрупулезным образом описаны не только процедуры использования статистических методов, но и продемонстрированы примеры выполнения исследовательских задач с применением каждого метода. Тем самым учебное пособие выполняет две функции: образовательную, интерпретируя логику работы в программе IBM SPSS Statistics, и научную, инкорпорируя теоретические знания в практику конкретных научных исследований. В качестве эмпирической базы были использованы следующие массивы, подготовленные сотрудниками социологической лаборатории Института геополитических и региональных исследований БФУ им. И. Канта:

1) опрос студентов БФУ им. И. Канта (тема: «Практики использования социальной сети “ВКонтакте”», n = 195, 2023 г.);

2) опрос жителей Калининградской области (тема: «Социально-политические настроения в Калининградской области», n = 915, 2022 г.);

3) опрос жителей Калининградской области (тема: «Историческая память населения Калининградской области», n = 997, 2022 г.);

4) опрос молодежи Калининградской области (тема: «Социально-политические настроения молодежи Калининградской области», n = 987, 2021 г.);

5) опрос женщин Калининградской области (тема: «Репродуктивные установки жительниц Калининградской области», n = 830, 2017 г.).

Такое разнообразие тем выбрано не случайно. Этим мы хотели продемонстрировать, как можно формулировать научную проблему и правильно выбирать статистические методы в различных областях социологического знания, что и определило название учебного пособия: «Статистический анализ данных в исследовательской деятельности социолога».

Еще одной отличительной особенностью данного пособия является обзор литературы, которым предваряется знакомство с основными методами статистического анализа данных в программе IBM SPSS Statistics. Автор искренне полагает, что это позволит удовлетворить запрос студентов на более широкое видение возможностей программы в исследовательской деятельности.

Несмотря на то что по задумке автора пособие должно служить мануалом по статистическому анализу, освоение материала будет наиболее эффективным, если читатель уже прослушал такие курсы, как «Высшая математика и математическая статистика» и «Методология и методы количественных и качественных исследований в социологии». Успешное прохождение дисциплины «Современные статистические пакеты (SPSS)», неотъемлемой частью которой является данное учебное пособие, несомненно, будет способствовать лучшему пониманию таких курсов, как «Прикладная статистика в социологии», «Население региона в социологическом измерении», «Социальная стратификация и социальное пространство», «Современные программы для обработки и базы данных в гуманитарных исследованиях», а также результативной подготовке и защите выпускной квалификационной работы.

По своей структуре учебное пособие состоит из 8 глав, которые гносеологически можно укрупнить до двух групп. К первой относятся главы 1 и 2, которые знакомят с базовыми статистическими категориями, функционалом самой программы и процедурами подготовки к анализу данных (их преобразованию). Без этого знания невозможно осмысленно работать дальше. Ко второй группе можно отнести главы 3—8: в них содержится разбор базовых и продвинутых методов анализа, описание областей их применения и примеры выполнения исследовательских задач. Каждая глава завершается вопросами для самопроверки, практическими заданиями и списком рекомендуемой литературы.

Глава 1
ОСНОВНЫЕ СТАТИСТИЧЕСКИЕ КАТЕГОРИИ

§ 1.1. Единица анализа. Переменная. Шкалы

Для социологического понимания статистических взаимосвязей базовыми терминами являются (1) единицы анализа, (2) их свойства, выраженные в (3) определенной шкале измерения, и (4) данные как результат обследований, замеров и разнообразных мониторингов.

Под единицей анализа понимают единичную часть объекта исследования [6, с. 11]. Например, если объектом исследования является городская молодежь, то каждый респондент в возрасте от 18 до 35 лет, проживающий в городе, будет квалифицирован как единица анализа. Однако, если объектом изучения являются, к примеру, молодежные субкультуры, то единицей анализа будет каждая отдельная субкультура молодежи. Здесь же нужно различать единицу анализа и единицу наблюдения. В последнем примере единицей наблюдения будет один представитель конкретной субкультуры, совокупность которых уже образует единицу анализа — молодежную субкультуру. Таким образом, единица анализа не всегда совпадает с единицей наблюдения. Статистический анализ позволяет преобразовывать информацию, собранную о единицах наблюдения, в информацию о единицах анализа. Об этом речь пойдет в следующей главе.

Под переменной понимают атрибут или признак, характеризующий одно из изучаемых свойств единицы анализа [6, с. 11]. Например, если единица анализа — респондент, то переменными могут выступать его пол, возраст и каждый вопрос в анкете, фиксирующий его мнение или оценку. Благодаря переменным исследователь может обобщать или различать единицы анализа. Различные проявления признака называют значениями. Например, переменная пол имеет два значения: мужской и женский. Каждое значение имеет определенное числовое выражение. Тем не менее в программе IBM SPSS Statistics представлены не только числовые переменные, но и текстовые, временные и ряд других. В аналитической статистике выделяют два типа переменных: зависимые и независимые. Объясняющие переменные называются независимыми, а объясняемые переменные — зависимыми [6, с. 39]. Например, в анализе таких переменных, как пол и уровень дохода, пол является независимой переменной, поскольку исследователь пытается объяснить различия в доходах в зависимости от гендерной принадлежности респондента.

Каждая переменная имеет свою шкалу измерения. В программе IBM SPSS Statistics реализованы три типа шкал: номинальная, порядковая и количественная. Номинальная шкала — способ классификации объектов или субъ-

ектов [4, с. 15], значения номинальной переменной представляют категории без естественного упорядочения [2, с. 14]. Например: пол, цвет кожи, национальность, семейное положение и прочее. Порядковая шкала — это шкала, классифицирующая по принципу «больше — меньше» [4, с. 15]. Примеры порядковых переменных включают баллы, представляющие степень удовлетворенности, уверенности оценивающие предпочтение [2, с. 14]. Например, оценка удовлетворенности жилищными условиями от 1 до 5, где 1 — максимальная неудовлетворенность, а 5 — абсолютная удовлетворенность. Количественная шкала всегда предполагает ответ на вопрос: «Сколько…» и потому позволяет сравнивать расстояния между значениями [2, с. 14], выполняя все арифметические операции. Примеры количественных переменных: возраст или образование в годах, количество членов семьи, стаж работы, доход и прочее. Подчеркнем, что количественную переменную всегда можно преобразовать в порядковую (разбив на интервалы), обратно — никогда. Каждая шкала предполагает свои статистические методы анализа и визуализации (табл. 1.1).

Таблица 1.1

Соотношение уровней измерения и статистических показателей

Характеристика шкалы	Шкала		
	Номинальная	Порядковая	Количественная
Определение	Неупорядоченные категории	Упорядоченные категории	Метрические / числовые значения
Примеры категорий	Пол, семейное положение	Рейтинг, уровень образования	Доходы, рост, вес
Меры центральной тенденции	Мода	Мода, медиана	Мода, медиана, среднее
Разброс	Нет анализа	Min, max, ранг, межквартильный размах	Min, max, ранг, межквартильный размах, стандартное отклонение, разброс
График	Круги, столбики	Круги, столбики	Гистограмма, ящичковые диаграммы
Анализ	Частоты	Частоты	Частоты, разведочный анализ

Составлено на основе [2, с. 35].

§ 1.2. Нормальное распределение. Статическая значимость

Перед применением любого статистического метода анализа в первую очередь следует выполнять оценку нормального распределения данных [1, с. 85], поскольку на этом этапе исследователь решает использовать ему пара-

метрические или непараметрические тесты. Нормальным распределением является такое распределение, при котором данные располагаются симметрично относительно своего среднего значения, образуя колоколообразную форму (рис. 1).

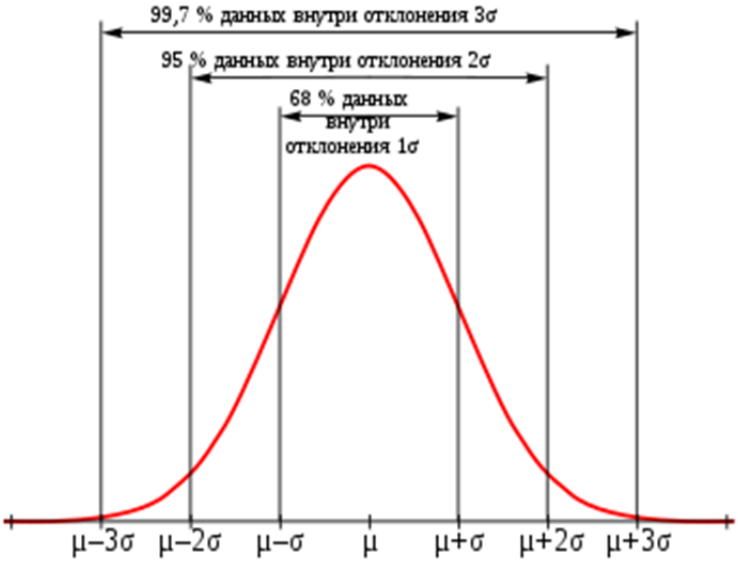


Рис. 1. График нормального распределения

Источник: [8].

В нормальном распределении 68,2 % данных находится в пределах одного среднеквадратического отклонения (σ) от среднего значения (μ), 95,4 % — в пределах двух среднеквадратических отклонений от среднего значения, 99,7 % — в пределах трех среднеквадратических отклонений от среднего значения и т. д.

О нормальном распределении свидетельствуют еще два коэффициента: асимметрия и эксцесс. Коэффициент асимметрии используется для измерения скошенности распределения в сторону больших или меньших значений признака. Нормальное распределение симметрично, если коэффициент асимметрии для него равен 0. Распределение со значимой положительной асимметрией имеет длинный хвост справа. Распределение со значимой отрицательной асимметрией имеет длинный хвост слева [3, с. 42]. Коэффициент эксцесса измеряет островершинность распределения. Считается, что распределение с эксцессом в диапазоне от –1 до +1 примерно соответствует нормальному виду [7].

Еще одним способом оценки нормальности распределения данных является тест Колмогорова — Смирнова. В этом случае исследователь формули-

рует нулевую гипотезу и проверяет ее, опираясь на полученный уровень значимости. Здесь следует сделать отступление и дать интерпретацию этим понятиям.

В статистическом анализе данных в большинстве случаев задача исследователя сводится к проверке нулевой гипотезы. Нулевая гипотеза — это предположение об отсутствии достоверных связей между переменными. Альтернативная гипотеза — это предположение о наличии статистически достоверных связей. Соответственно, принимая нулевую гипотезу, мы отвергаем альтернативную, и наоборот. Для проверки нулевой гипотезы проводится оценка уровня значимости. Уровень значимости (p-уровень или асимптотическая значимость в программе IBM SPSS Statistics) — это вероятность того, что различия между двумя группами или результаты исследования были получены случайно и не являются статистически значимыми. Обычно уровень значимости выбирается заранее (например, 5 %, или 0,05), чаще всего используется 0,05. Эта величина означает, что вероятность случайного получения статистически значимого результата составляет менее 5 %. Соответственно, если полученный p-уровень меньше или равен 0,05, мы отклоняем нулевую гипотезу и принимаем альтернативную, поскольку с вероятностью не более 5 % различия являются случайными. Другими словами, если полученное значение p-уровня значимости меньше выбранного уровня значимости, то различия между группами или результаты исследования считаются статистически значимыми.

В случае использования теста Колмогорова — Смирнова исследователь проверяет нулевую гипотезу о том, что распределение значений заданной переменной не отличается от нормального ($p > 0,05$). Для проверки этой гипотезы необходимо обратить внимание на уровень значимости статистики по критерию Колмогорова — Смирнова.

В таблице 1.2 показан уровень значимости для переменной с нормальным распределением (тест_1) и с распределением, отличающегося от нормального (тест_2). В последнем случае вовсе не обязательно отказываться от дальнейшего анализа данных, следует лишь выбрать подходящий непараметрический метод связи.

Таблица 1.2

Применение критерия Колмогорова — Смирнова для оценки нормального распределения

Вид значимости	Тест_1	Тест_2
Асимптотическая значимость (2-сторонняя)	0,231	0,001

Таким образом, нормальное распределение помогает оценить уровень значимости и определить вероятность полученных результатов в случайных выборках.

§ 1.3. Меры изменчивости и меры центральной тенденции

Меры изменчивости и меры центральной тенденции используются для описания распределения значений переменных. Они позволяют увидеть базовые характеристики как всей выборочной совокупности, так и отдельных переменных, поэтому служат первичным методом для разведывательного анализа данных.

Меры центральной тенденции указывают на расположение среднего, или типичного, значения признака, вокруг которого сгруппированы остальные наблюдения. К ним относятся среднее арифметическое, медиана и мода.

Мода — это такое значение в совокупности наблюдений, которое встречается чаще всего. Например, если в выборке содержится 60 % горожан, 30 % жителей районных центров и 10 % жителей деревень и сел, то модальным значением будут горожане. Мода значительно чаще используется при анализе номинальных переменных.

Медиана — значение, которое делит упорядоченное множество данных пополам, так что одна половина наблюдений оказывается меньше медианы, а другая — больше. Иными словами, медиана — это 50-й процентиль распределения [5]. Обычно используется для порядковых переменных, то есть таких переменных, значения которых могут быть упорядочены от меньших к большим.

Чтобы определить медианное значение, нужно упорядочить наблюдения по возрастанию переменной: значение, которое находится в середине списка, будет являться медианным. Например, в ряду значений 5—9 баллов, медианное значение будет равно 7 баллам. Если количество значений в группе четное, медиана будет равна среднему арифметическому двух центральных значений.

Медиану иногда называют «позиционным средним», так как она указывает именно на среднюю позицию в упорядоченном ряду наблюдений [5]. Медиана может совпадать или не совпадать с модой, но она наиболее точно определяет центральное значение в упорядоченном списке. Некоторые исследователи считают, что медиана более справедлива и точна при описании таких величин, как доход семьи, поскольку средняя арифметическая величина может искажаться очень большими или очень маленькими значениями. Например, средний доход может быть смещен в большую сторону из-за одного очень высокого дохода, но медиана не будет зависеть от таких «крайних» значений.

Среднее арифметическое — значение, получаемое путем сложения всех чисел в группе и последующего деления суммы на количество чисел в ней. Обычно используется с количественными переменными.

При выборе центральной меры тенденции нужно учитывать ее свойства, форму распределения и здравый смысл. Если распределение унимодально и симметрично (то есть половины гистограммы слева и справа от модального значения зеркально совпадают), то среднее, медиана и мода будут равны

между собой. В случае, если выборка взята из большой совокупности с нормальным распределением (когда большие и малые значения встречаются редко, а средние — часто), наиболее подходящей мерой тенденции будет среднее. Однако, при наличии крайних значений, которые могут значительно повлиять на среднее, следует использовать медиану [5].

Для достижения более полной и точной характеристики выборочных значений, необходимо учитывать не только стандартные или типичные значения, но также устанавливать степень отклонений от них. Для оценки эффективности центральных мер тенденции следует применять соответствующие меры изменчивости или разброса.

Наиболее простой и общепринятой мерой изменчивости является размах (диапазон) значений, который характеризует вариацию значений в выборке и определяется как разность между максимальным и минимальным значением. Следует отметить, что данная мера не учитывает индивидуальные отклонения от типичных значений выборки, а лишь информирует о ширине диапазона изменчивости выборочных данных.

Еще одна мера разброса значений, коэффициент вариации (CV) — это отношение стандартного отклонения выборки к ее среднему значению, обычно выражаемое в процентах. Он представляет меру относительной изменчивости выборочных данных и используется для сравнения изменчивости в разных выборках с разной средней величиной. Более высокое значение коэффициента вариации указывает на более высокую изменчивость в выборке. Представим, что мы исследуем доход людей в двух разных регионах. В первом регионе средний доход составляет 50 000 рублей, а стандартное отклонение — 5000 рублей. Во втором регионе средний доход равен 80 000 рублей, а стандартное отклонение — 12 000 рублей. Хотя стандартное отклонение дохода во втором регионе выше, мы не можем без дополнительной информации сравнивать изменчивость между выборками, так как они имеют разную среднюю величину. В этом случае мы можем использовать коэффициент вариации: для первого региона

$$CV = (5000 / 50\,000) \cdot 100\,\% = 10\,\%,$$

для второго региона

$$CV = (12\,000 / 80\,000) \cdot 100\,\% = 15\,\%.$$

Из этого следует, что доход второго региона относительно более изменчив, чем первого, с учетом разности их средних значений.

Стандартное отклонение — это мера разброса значений в выборке относительно среднего значения. Оно показывает, насколько значения выборки различаются от средней величины выборки. Чем выше стандартное отклонение, тем более разнородные значения. Формула стандартного отклонения включает вычисление среднего значения выборки и расстояния каждого значения выборки до него, затем квадратичного усреднения этих расстояний и извлечения квадратного корня из этого среднего значения. Стандартное от-

клонение — одна из самых часто используемых мер для описания изменчивости выборки в статистике и науке. Пример расчета и интерпретации стандартного отклонения выглядит так. Допустим, мы измерили вес в килограммах у 10 человек и получили следующие значения: 75, 68, 80, 72, 76, 82, 69, 73, 77, 70. Сначала необходимо вычислить среднее значение этой выборки, которое равно $(75+68+80+72+76+82+69+73+77+70)/10=74{,}2$. Далее для каждого значения в выборке необходимо вычислить расстояние до среднего значения:

- $(75-74{,}2)=0{,}8$,
- $(68-74{,}2)=-6{,}2$,
- $(80-74{,}2)=5{,}8$,
- $(72-74{,}2)=-2{,}2$,
- $(76-74{,}2)=1{,}8$,
- $(82-74{,}2)=7{,}8$,
- $(69-74{,}2)=-5{,}2$,
- $(73-74{,}2)=-1{,}2$,
- $(77-74{,}2)=2{,}8$,
- $(70-74{,}2)=-4{,}2$.

После чего каждое из расстояний необходимо возвести в квадрат и взять их среднее значение, которое затем нужно извлечь из под корня:

$$\text{sqrt}((0{,}8^2 + (-6{,}2)^2 + 5{,}8^2 + (-2{,}2)^2 + 1{,}8^2 + 7{,}8^2 + (-5{,}2)^2 + (-1{,}2)^2 + 2{,}8^2 + (-4{,}2)^2)/10) = \text{sqrt}(142{,}96/10) = 3{,}78.$$

Соответственно, стандартное отклонение выборки весов равно 3,78 кг. Это говорит о том, что значения выборки изменяются на промежуток, равный примерно 3,78 кг, относительно среднего значения веса.

Таким образом, если стандартное отклонение мало, значит, среднее значение репрезентативно для данной выборки наблюдений, и наоборот, если стандартное отклонение велико, среднее значение может быть менее представительным для данной выборки.

Одним из важных применений стандартного отклонения является возможность определения основных характеристик нормального распределения в сочетании со средним арифметическим. Графически нормальному распределению соответствует симметричная колоколообразная кривая, свойства нормального распределения хорошо изучены, что позволяет делать выводы о различных распределениях. Например, если среднее значение нормального распределения равно 100, а стандартное отклонение составляет 2, то можно предположить, что не менее 68 % наблюдений находятся в диапазоне от 98 до 102 (то есть среднее значение ± 1 стандартное отклонение). Из теории вероятности также известно, что около 99,73 % общего количества наблюдений будет находиться в пределах ± 3 стандартных отклонений. Это может быть использовано в прогнозировании различных значений всей генеральной совокупности при условии использования случайной выборки в сборе данных.

Если переменная измерена в порядковой шкале (например, доход или продолжительность образования), то можно легко получить выборочную оценку среднего. Единственный вопрос: насколько близка эта выборочная оценка дохода к истинному значению этого параметра, который мы бы получили, если бы имели возможность изучить всю совокупность. Если наша выборка была случайной, то мы способны ответить на этот вопрос. Можно определить числовой интервал, в который с заданной вероятностью попадает выбранный параметр. Большая часть этих оценок будет попадать в область, близкую к истинному значению среднего, и лишь некоторые выборки могут оказаться в «хвостах» распределения, значительно отклоняющиеся от этого значения. Для каждой выборки шансы оказаться близко к параметру совокупности выше, чем вероятность оказаться в «хвосте». Используя стандартную ошибку среднего, можно понять степень близости выборочной оценки к параметру во всей генеральной совокупности и с заданной вероятностью определить пределы, в которых находится среднее значение.

Стандартная ошибка — это оценка стандартного отклонения распределения выборочных средних. Она показывает, насколько сильно выборочное среднее может отличаться от истинного среднего по генеральной совокупности. Другими словами, стандартная ошибка среднего — мера того, насколько сильно среднее может различаться у разных выборок, взятых из одного распределения.

Для расчета стандартной ошибки часто используется формула:

стандартная ошибка = стандартное отклонение /
/ квадратный корень из размера выборки.

Поскольку 95 % значений генеральной совокупности находятся в пределах ±2 стандартных отклонений от среднего значения, мы можем использовать это правило для определения диапазона, в котором вероятнее всего находится истинное среднее значение генеральной совокупности. Правило двойного стандартного отклонения также применяется к стандартной ошибке выборки, что означает, что в 95 % случаев выборочное среднее будет лежать в пределах ±2 стандартных ошибок от реального среднего генеральной совокупности. Таким образом, любая конкретная выборка с 95%-ной вероятностью даст оценку, лежащую в интервале ±2 стандартных ошибок среднего совокупности. Заданный таким образом интервал для выборочных оценок называется доверительным интервалом, а та вероятность, с которой мы «попадаем» в этот интервал (например, 95 % или 99 %), называется доверительной вероятностью [5]. Если мы рассчитали, что средняя заработная плата для случайной выборки составляет 20 000 рублей, а стандартная ошибка равна 500 рублям, мы можем с уверенностью в 95 % утверждать, что средняя заработная плата для всех респондентов находится в диапазоне от 19 000 до 21 000 рублей. Если мы зададим интервал в 3 стандартные ошибки, то мы сможем достичь уровня доверительной вероятности, равной 99,73 %. Полезно

помнить о том, что стандартная ошибка обратно пропорциональна размеру выборки: чем больше выборка, тем меньше стандартная ошибка и тем точнее выборочное среднее представляет среднее по генеральной совокупности.

В данной главе были рассмотрены ключевые понятия статистического анализа данных, без понимания которых работа в IBM SPSS Statistics будет лишена смысла. Стоит добавить, что все упомянутые в параграфах 1.2 и 1.3 коэффициенты в IBM SPSS Statistics рассчитываются автоматически. О том, как это сделать и каковы ключевые опции интерфейса программы, пойдет речь в следующей главе.

§ 1.4. Контрольные вопросы

1. Дайте определения следующим понятиям: единица анализа, единица наблюдения, шкала, переменная, данные, нормальное распределение, статистическая значимость, среднее арифметическое, медиана, мода, размах, коэффициент вариации, стандартное отклонение и стандартная ошибка.

2. Какие меры центральной тенденции соответствуют каким типам шкал?

3. Как связаны меры центральной тенденции и меры изменчивости с нормальным распределением?

§ 1.5. Практические задания

1. Назовите единицы анализа, единицы наблюдения, зависимую и независимую переменные, а также их шкалы при анализе следующих исследовательских вопросов:

А) Как различаются медиаобразы региона у различных поколений его жителей?

Б) Влияет ли уровень дохода избирателя на его электоральную активность?

2. Рассчитайте медианное и среднее значение следующего набора значений:

А) 3, 5, 1, 7, 4, 2, 6;

Б) 5, 9, 4, 7, 1, 6, 2.

3. Рассчитайте среднее значение, стандартное отклонение и дайте интерпретацию полученным результатам при оценке следующих результатов тестирования:

А) 85, 72, 90, 68, 78;

Б) 74, 61, 79, 57, 67.

§ 1.6. Рекомендуемая литература

1. *Бююль А., Цёфель П.* SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей. СПб. : ДисСофтЮп, 2005.

2. *Воронин Г. Л.* IBM SPSS Statistics V21.0.0.0: Вводный курс. М. : Институт социологии РАН, 2014.

3. *Воронин Г. Л.* Статистический анализ данных в IBM SPSS Statistics V27.0.1.0. Н. Новгород : ННГУ им. Н. И. Лобачевского, 2022.

4. *Гаджигасанова Н. С.* Методы прикладной статистики для социологов. Ярославль : ЯрГУ, 2013.

5. *Девятко И. Ф.* Методы социологического исследования. Екатеринбург : Изд-во Урал. ун-та, 1998.

6. *Крыштановский А. О.* Анализ социологических данных с помощью пакета SPSS. М. : ГУ ВШЭ, 2006.

7. *Наследов А. Д.* SPSS 19: профессиональный статистический анализ данных. СПб. : Питер, 2011.

8. *Портал* знаний. Глобальный интеллектуальный ресурс. URL: http://statistica.ru/theory/normalnoe-raspredelenie/ (дата обращения: 05.10.2023).

Глава 2
УПРАВЛЕНИЕ ДАННЫМИ В SPSS

§2.1. Интерфейс SPSS и подготовка массива данных

Программа IBM SPSS Statistics представлена тремя окнами: «Редактор данных», «Редактор командного языка Syntax» и «Вывод».

Окно редактора данных по умолчанию открывается первым. Оно содержит две вкладки: «Данные» и «Переменные». Вкладка «Данные» представляет собой таблицу по сути идентичную таблице MS Excel, где в строках располагаются единицы анализа (переменные), а в столбцах — их значения. Вкладка «Переменные» (рис. 2.1) представляет собой свод переменных (строки) и их свойства (столбцы).

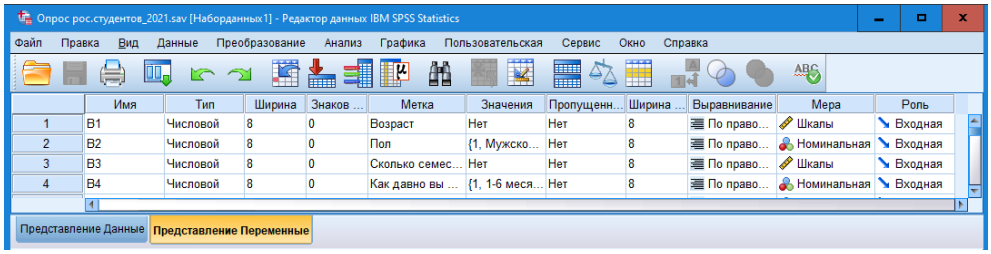


Рис. 2.1. Интерфейс редактора данных в программе IBM SPSS Statistics

Описание переменной включает имя, тип, количество цифр или символов в переменной, количество десятичных знаков, описательные метки переменных и значений, пользовательские пропущенные значения, ширина столбца, уровень измерения.

Имя переменной — краткое обозначение переменной в массиве. Тип переменной задается типом данных для каждой переменной. Например, числовая — для числовых данных, текстовая — для текстовых.

Столбец «Ширина» позволяет задать максимальное количество знаков, которое может иметь значение переменной, включая дробную часть. На практике заранее определить ширину переменной гораздо труднее, поскольку не всегда известно, какие данные нам понадобится вводить в будущем. Поэтому следует задавать ширину переменной с гарантированным запасом: ее можно ограничить потом, после ввода данных.

Столбец «Знаков…» предназначен для задания числа десятичных знаков после запятой в случае, если тип переменной допускает использование дробных чисел.

Метки переменных используются для развернутого обозначения переменной и могут включать до 256 символов. Метки переменных могут содер-

жать пробелы и символы, применение которых не допускается в именах переменных. Как правило, имя переменной — шифр вопроса анкеты, метка — сам вопрос анкеты целиком.

Столбец «Значения» предназначен для того, чтобы каждому значению переменной присвоить описательную метку, например, закодировать каждый вариант ответа в вопросе.

Столбец «Пропущенные значения» позволяет указать, какие значения данных будут считаться пропущенными. Например, мы можем исключить данные, содержащие вариант ответа «затрудняюсь ответить». Можно пропускать как отдельные значения, так и их диапазон.

Столбец «Ширина столбца» позволяет управлять шириной (в символах) столбцов вкладки «Данные».

Столбец «Выравнивание» позволяет управлять расположением данных внутри ячейки: по правому краю, по левому краю и по центру.

Столбец «Мера» позволяет выбрать одну из трех доступных шкал для переменной: количественную, порядковую и номинальную.

С помощью столбца «Роль» указываются роли, которые можно использовать для выбора переменных для анализа. Если переменные удовлетворяют этим ролям, они автоматически отображаются в списках выбора при открытии некоторых диалоговых окон анализа. На практике эта опция практически не используется.

Окно «Вывод» открывается автоматически при запуске программы (и сразу после выполнения какой-либо процедуры анализа) и содержит две панели. В левой панели содержимое вывода представлено в схематическом виде. Правая панель содержит сами результаты — статистические таблицы, диаграммы и текстовой вывод (рис. 2.2).

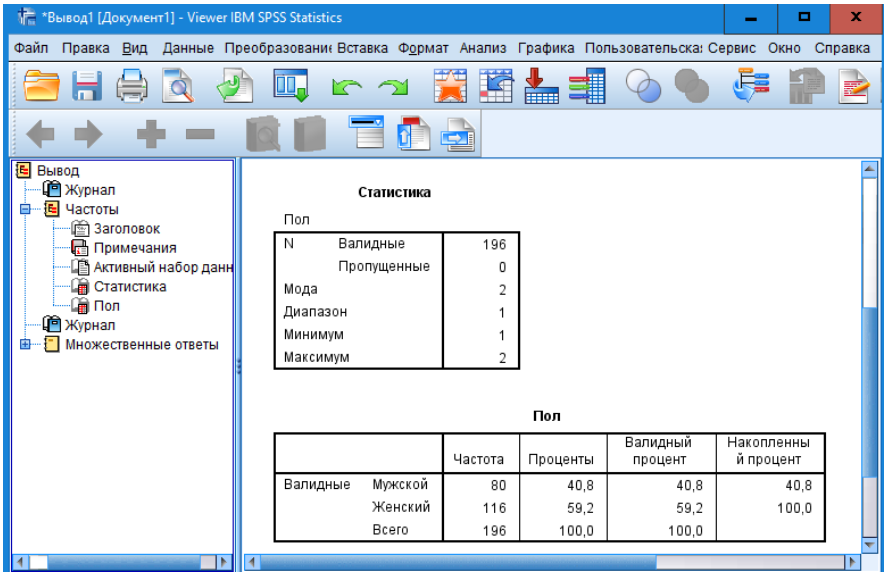


Рис. 2.2. Интерфейс окна «Вывод» в программе IBM SPSS Statistics

Результаты анализа, представленные в окне вывода, можно конвертировать и переносить в другие известные приложения: Microsoft Office (Word, Excel, Power Point), HTML, PDF.

Окно редактора командного языка синтаксиса (рис. 2.3) имеет текстовый формат записи команд и позволяет открывать и создавать командные файлы, которые можно сохранять и использовать в дальнейшем. Открыть его можно посредством нажатия кнопки «вставка» при выполнении любого вида анализа, либо пройдя путь: файл — создать — редактор синтаксиса.

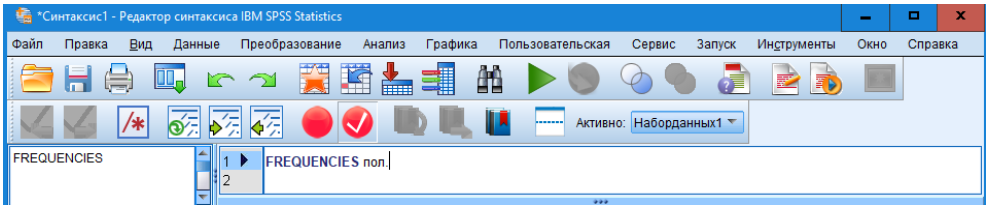


Рис. 2.3. Интерфейс редактора синтаксиса в программе IBM SPSS Statistics

Работа в окне синтаксиса выполняет как минимум две важные функции: 1) редактирование как переменных, их значений, так и выборочных совокупностей данных в процессе и впоследствии статистического анализа; 2) ведение конспекта всей работы в IBM SPSS Statistics, что позволяет сразу возвращаться к любому шагу на каждом этапе анализа, избегая повторного редактирования и ввода переменных. Однако мы солидарны с А. Д. Наследовым в том, что «необходимость обращения к командному языку SPSS обычно возникает лишь после достаточно уверенного освоения стандартного интерфейса программы — для облегчения управления данными, выполнения длинных последовательностей команд или реализации редких методов анализа данных» [3, с. 33]. Тем не менее основные команды синтаксиса будут представлены в данном пособии.

§ 2.2. Преобразование переменных и данных

Не секрет, что основная работа исследователя в программе IBM SPSS Statistics сводится к подбору подходящих переменных, их вычислению и преобразованию, равно как и управлению данными, включая определение границ выборочных совокупностей. Сам статистистический анализ занимает считанные минуты. В данном параграфе разберем наиболее популярные и востребованные техники по подготовке массива к анализу.

Отбор наблюдений. К примеру, нам требуется выполнить анализ только респондентов в возрасте от 18 до 35 лет в массиве, состоящем из респондентов в возрасте 18—80 лет. Для этого мы в любом окне программы переходим в раздел «Данные» (данные — отобрать наблюдения), выбираем переменную возраст и в новом окне указываем: «если больше или равно 18 и меньше или

равно 35» (рис. 2.4). Иногда может потребоваться задать несколько условий, например, возрастной диапазон и уровень образования (к примеру, только высший). В таком случае мы ставим знак «&» между указанными переменными. Когда нам снова потребуется весь массив без изменений, мы возвращаемся в это окно и выбираем «Все наблюдения».

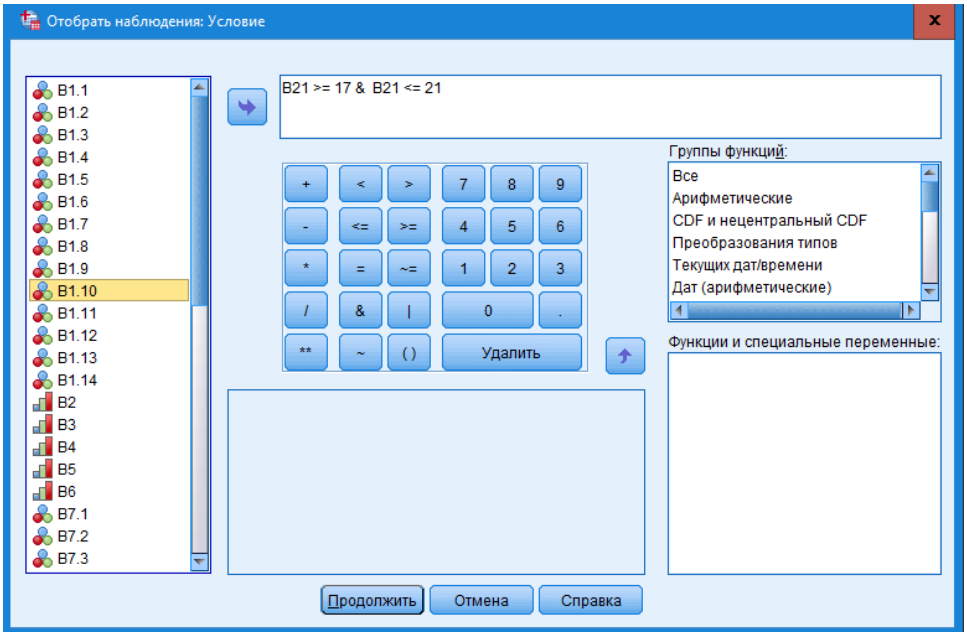


Рис. 2.4. Интерфейс функции «Отобрать наблюдения»

Исключение пропущенных значений. Например, нам может потребоваться исключить определенные значения переменной. В таком случае во вкладке «Переменные» в столбце «Пропущенные» в строке нашей переменной мы указываем значение, которое соответствует вариантам, которые требуется исключить (рис. 2.5). Программа исключит этот вариант при анализе данной переменной.

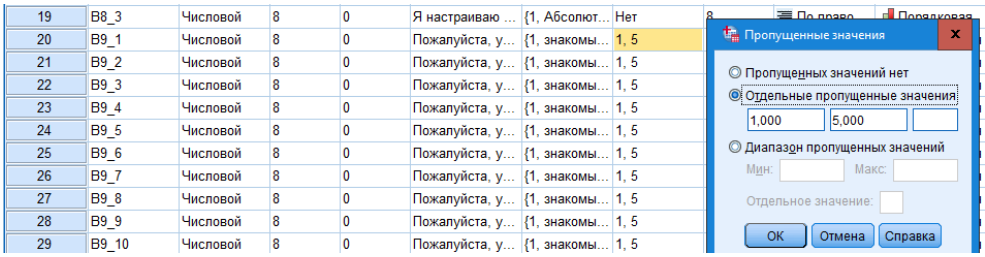


Рис. 2.5. Интерфейс функции «Пропущенные значения»

Преобразование переменной. Например, нам может потребоваться преобразовать количественную переменную в порядковую с разбивкой на определенные интервалы. Для этого нужно перейти во вкладку «Преобразование» (данные — преобразовать в новую переменную (если мы хотим сохранить исходную количественную переменную)). В новом окне следует выбрать значения будущих интервалов и дать им новое значение. Например, диапазон значений от 1 до 10 теперь будет обозначаться единицей, диапазон от 11 до 20 — двойкой и так далее (рис. 2.6). Дать этой порядковой переменной новое имя, кликнуть «Изменить», а потом ОК. Далее следует перейти в окно «Свойства переменных» (данные — свойства переменных), выбрать вновь созданную переменную и задать все необходимые свойства: количество десятичных знаков, метку, выбрать шкалу. Эту же процедуру следует проводить и в случае, когда требуется объединить некоторые значения, например, «согласен» и «скорее согласен» или, когда необходимо поменять порядок кодировки значений, например, от «абсолютно значимо» до «не имеет никакого значения», и наоборот.

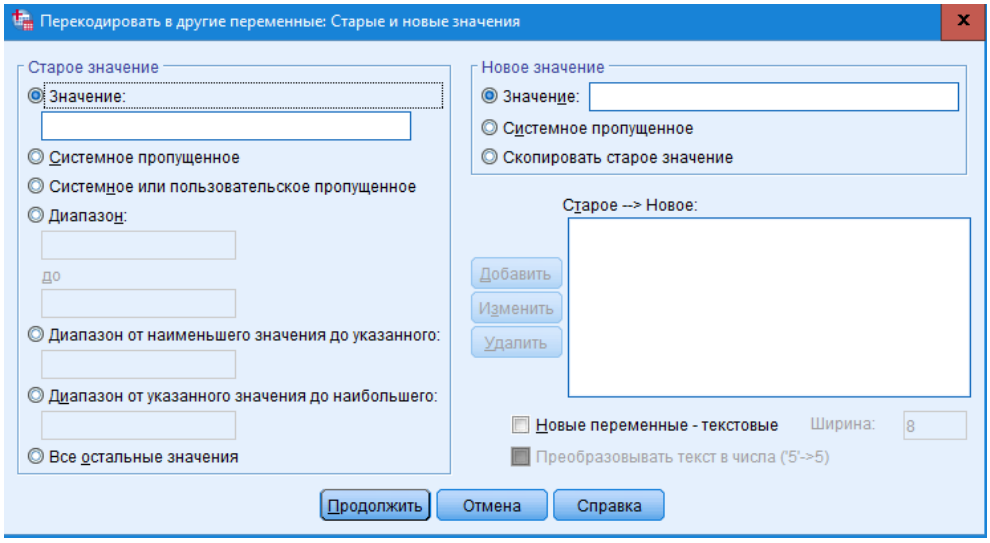


Рис. 2.6. Интерфейс функции «Перекодировать в другие переменные»

Если нам требуется выполнить аналогичное действие для нескольких переменных, уместно будет использовать окно редактора синтаксиса. В строке с переменными мы через пробел указываем переменные, с которыми хотим совершить идентичные преобразования. Выделяем команду целиком и кликаем на зеленый треугольник вверху (см. рис. 2.3).

Если нас интересует преобразование количественной переменной в порядковую с равным шагом или определенным процентилем, можно воспользоваться функцией «визульная категоризация» (преобразование — визуальная категоризация).

В окне «Визуальная категоризация» введем имя новой преобразованной переменной, например, V23 (возраст). Метка переменной преобразуется автоматически. Затем нажимаем кнопку «Границы интервалов», где зададим необходимый порядок группировки количественной переменной (рис. 2.7). В процессе группировки количественной переменной предусмотрены два алгоритма: равные интервалы и равные процентили. В первом случае необходимо заполнить минимум два поля: местоположение первой границы и ширину шага.

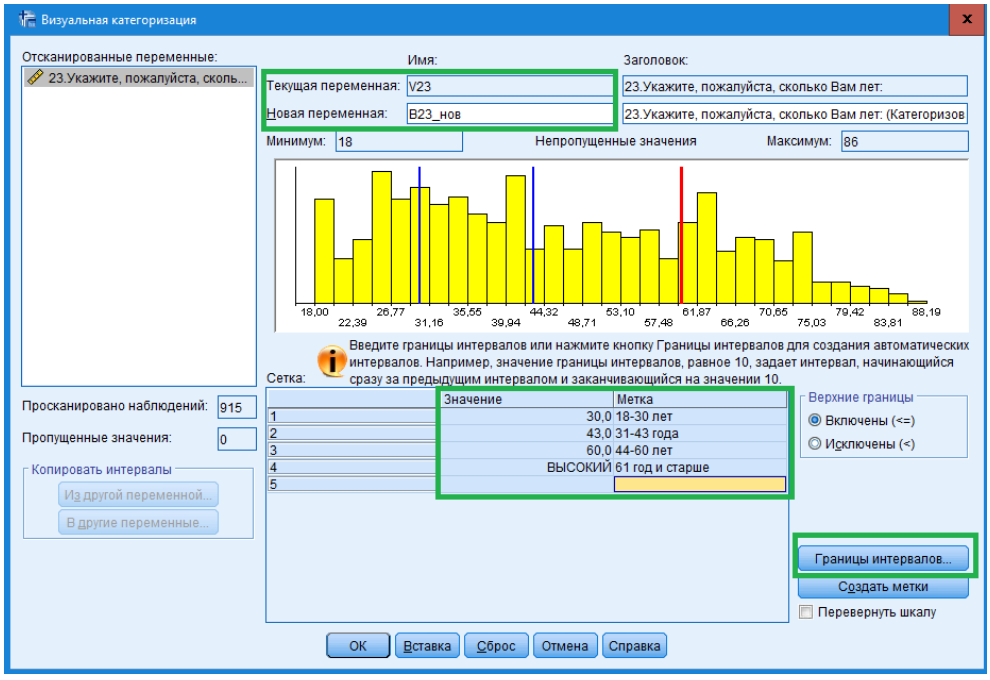


Рис. 2.7. Интерфейс функции «Визуальная категоризация»

Применительно к нашим данным целесообразно использовать метод разбиения на группы, называемый «Равные процентили». В этом случае необходимо заполнить только одно из двух полей. Мы планируем разбить выборочную совокупность на 4 группы, поэтому в поле «Количество границ» следует ввести цифру 3. В поле «Ширина (%)» появится значение 25 %, то есть в результате выполнения этой команды будет создана переменная с четырьмя категориями, при этом каждая из категорий будет включать примерно 25 % выборочной совокупности (табл. 2.1).

Таблица 2.1

Результат визуальной категоризации данных методом равных процентиль

Переменная		Частота	Проценты	Валидный процент	Накопленный процент
Валидные	18—30 лет	230	25,1	25,1	25,1
	31—43 года	236	25,8	25,8	50,9
	44—60 лет	228	24,9	24,9	75,8
	61 год и старше	221	24,2	24,2	100,0
	Всего	915	100,0	100,0	—

Отметим, что в результате проведенных преобразований мы получаем новую категоризованную переменную в редакторе данных, состоящую из четырех возрастных диапазонов.

Вычислить переменную. Эта команда используется, когда требуется рассчитать значения новой переменной, основываясь на значениях других переменных ее составляющих. В качестве иллюстрации воспользуемся примером Г. Л. Воронина [2, с. 57]. Допустим, нам требуется рассчитать среднедушевой доход индивида. У нас есть две переменные (RLMS-HSE за 2019 г.): XF14 — доход всех членов семьи, и x_nfm — количество членов семьи. Для того чтобы вычислить среднедушевой доход, следует открыть окно «Вычислить переменную» (преобразование — вычислить переменную). Там указать имя новой переменной (целевая переменная, например, NEW), а в поле «Числовое выражение» указать формулу среднего арифметического: XF14 / x_nfm (рис. 2.8). В результате проведенных преобразований мы получаем новую переменную.

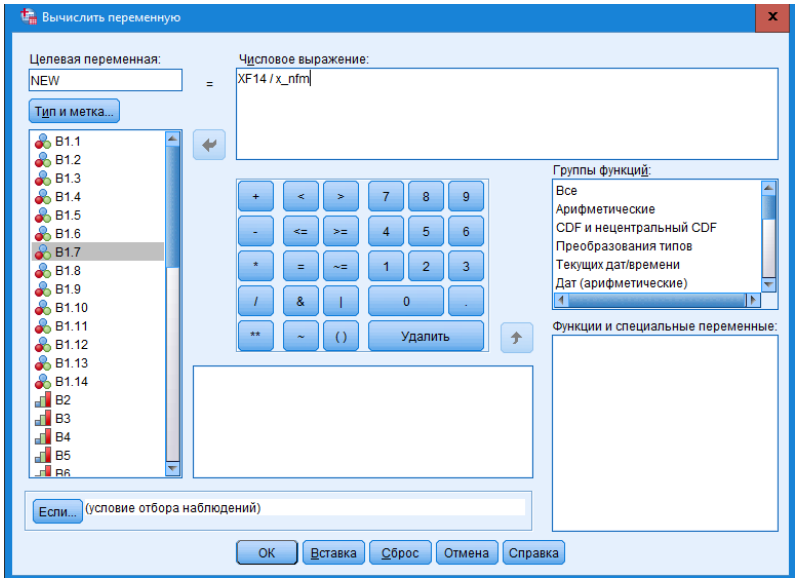


Рис. 2.8. Интерфейс функции «Вычислить переменную»

Подсчитать значения в наблюдениях. Эта команда предназначена для создания переменной, в которой для каждой единицы анализа задана информация о том, сколько раз содержится то или иное указанное значение или диапазон значений. В качестве иллюстрации снова воспользуемся примером Г. Л. Воронина [2, с. 57]. Допустим, мы решили создать переменную «физическая активность респондента», где хотим видеть информацию о включенности респондента в различные виды физической активности от бега трусцой до занятий борьбой, боксом, каратэ. В анкете имеется десять вопросов о физической активности респондента, предполагающие ответ «да» или «нет», где цифрой 1 помечен ответ респондента «Да, занимаюсь». Итоговая переменная будет предоставлять исследователю информацию о том, какое разнообразие типов физической активности характерно для респондентов, где 0 — отсутствие любых из перечисленных видов активности, 10 — респондент включен во все возможные в рамках анкеты виды физической активности.

Переходим в окно подсчета (преобразование — подсчитать значения в наблюдениях). Задаем имя и метку итоговой переменной, выбираем все переменные с указанием различных видов спорта (рис. 2.9). Нажимаем кнопку «Задать значения» и задаем значение «1», через клавишу «Добавить» переносим в окно «Подсчитываемые значения». В результате проведенных преобразований мы получаем новую переменную.

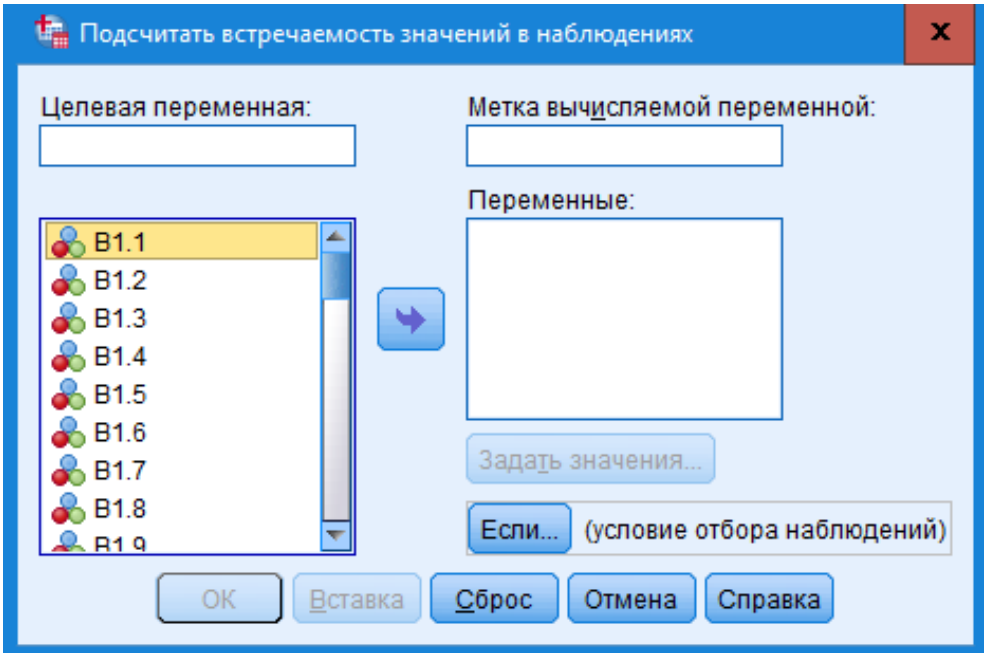


Рис. 2.9. Интерфейс функции «Подсчитать значения в наблюдениях»

Взвешивание данных. Данная процедура позволяет увеличить или уменьшить в выборке количество наблюдений с определенными характеристиками посредством назначения весовых коэффициентов, отображающих значимость наблюдений по сравнению с другими наблюдениями. Подобная корретировка выборки используется для: 1) более точной характеристики генеральной совокупности; 2) повышения значимости ответов респондентов с определенными признаками[1].

Взвешивание данных в программе IBM SPSS Statistics выполняется в несколько этапов:

1. Расчет весового коэффициента. К примеру, в полученной выборке 10,5 % респондентов являются служащими, хотя известно, что доля служащих в общем населении составляет только 8,4 % [1, с. 151—155]. Весовой коэффициент определяется как отношение значения генеральной совокупности к полученному значению. Таким образом, для служащих весовой коэффициент равен 8,4 / 10,5 = 0,8, а для остальных — 91,6 / 89,5 = 1,023.

2. Создание переменной взвешивания. В окне редактора Синтаксиса вводим следующую команду:

```
if (Статус= 1) weight1 = 0,8.
if (Статус= 2) weight1 = 1,023.
execute.
WEIGHT by weight1.
FREQUENCIES Статус.
```

Эта команда показывает, что если респондент является служащим (статус = 1), то его значения будут умножены на 0,8, в противном случае — они будут умножены на 1,023. Команда «WEIGHT by weight1» означает, что программа уже выполняет процедуру взвешивания данных по вновь созданной переменной weight1 (создается по умолчанию). Команда «FREQUENCIES Статус» показывает частотное распределение переменной «Статус» с учетом действия весовых коэффициентов, тем самым позволяет проверить правильность выполнения всех предыдущих действий.

Когда массив «взвешен», любые последующие статистические процедуры выполняются с учетом полученных коэффициентов. Чтобы отменить команду «взвешивание данных», следует перейти в окно «Взвесить наблюдения» (данные — взвесить наблюдения) и выбрать «Не взвешивать наблюдения».

§ 2.3. Анализ надежности (Альфа Кронбаха)

Анализ надежности является мерой точности, с которой проводится тестирование некоторого признака [1, с. 411]. Он применяется для отбора наиболее пригодных вопросов или заданий измерительной методики (вопросни-

[1] Например, если проводится опрос для определения, какие изменения стоит вносить в существующую продукцию, исследователь может принять решение присвоить больший весовой коэффициент ответам респондентов, которые пользуются данным товаром чаще других (https://nafi.ru/upload/spss/Lection_3.pdf).

ка, анкеты, теста) [3, с. 266]. Обычно при разработке методики сначала составляют ее предварительный вариант, включающий избыточное количество вопросов (переменных), который апробируется на достаточно представительной выборке респондентов. Затем проводят анализ надежности, который позволяет при помощи многочисленных критериев исключить неподходящие задания.

С одной стороны, это самостоятельная методика статистического анализа данных, с другой — она предполагает преобразование и создание новых переменных, значительно обогащая результаты, которые исследователь может получить из собранного массива данных. Именно поэтому мы рассматриваем эту процедуру в структуре других возможностей по управлению данными в программе IBM SPSS Statistics.

Альфа Кронбаха (α, коэффициент альфа Кронбаха) определяется как средний коэффициент корреляции между рядом всех вопросов в тесте или анкете. Другими словами, она позволяет проверить, насколько согласуется ряд вопросов между собой и насколько однородна и истинна данная шкала. Чем выше значение этого коэффициента, тем более надежным считается тест. Обычно, коэффициент альфа Кронбаха принимают значения от 0 до 1, где значение выше 0,7 считается приемлемым, а значение более 0,8 — хорошим качеством.

В социальных науках этот коэффициент обязательно используется, когда признак может быть определен только совокупностью ответов на различные вопросы и исследователю важно выяснить, какая часть этой совокупности наиболее согласована между собой. Примерами таких сложно концептуализированных признаков являются социальный капитал, социальная ответственность, доверие и прочее.

В качестве иллюстрации возьмем результаты опроса 195 студентов БФУ им. И. Канта относительно их практик использования социальной сети «ВКонтакте»[2]. На примере этого массива рассмотрим «процесс сборки» социального капитала, формируемого в социальной сети. Вслед за исследователями из Мичиганского университета [5], мы измерили социальный капитал социальной сети российских студентов, попросив респондентов оценить несколько утверждений по пятибальной шкале: «“ВКонтакте” стал частью моей повседневной жизни», «Я чувствую себя частью сообщества пользователей “ВКонтакте”», «Я буду огорчен, если “ВКонтакте” отключат», «Я использую “ВКонтакте”, чтобы узнать больше о том, кого я встретил лично оффлайн», «Я использую “ВКонтакте”, чтобы узнать больше о других людях, живущих рядом со мной», «Я использую “ВКонтакте”, чтобы быть на связи со своими давними друзьями», «Я использую “ВКонтакте”, чтобы познакомиться с новыми людьми». Для оценки связанности этих вопросов мы и будем использовать анализ надежности.

[2] Развитие этого исследования можно прочитать в источнике [4].

Основные этапы работы:

1. *Ввод переменных*. Для этого мы переходим в окно «Анализ надежности» (анализ — шкалы — анализ надежности). В поле «Пункты» переносим все указанные выше переменные. В окне «Статистики» выбираем все три параметра описательной статистики: пункты; шкалы; шкалы, если пункт удален. Остальные параметры не меняем[3]. Затем кликаем «Продолжить» и «ОК» (рис. 2.10).

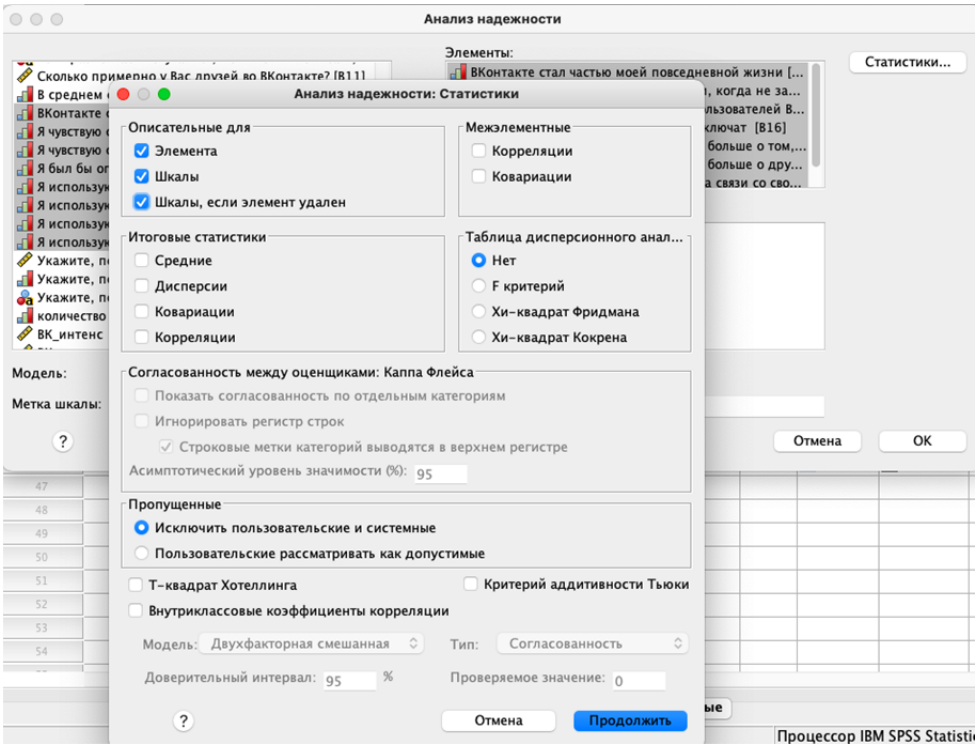


Рис. 2.10. Интерфейс функции «Анализ надежности»

2. *Анализ надежности всей совокупности переменных*. На данном этапе мы оцениваем полученный коэффициент Альфы Кронбаха и изучаем таблицу «Статистика пунктов по отношению к суммарному баллу» (рис. 2.11).

[3] Пункты — средние значения и стандартные отклонения указанных переменных.

Шкалы — средние значения, дисперсии, стандартные отклонения и размер выборки для суммы всех переменных шкалы. Под суммой переменных понимается сумма значений всех выбранных переменных (пунктов) для каждого объекта.

Шкалы, если пункт удален — значения коэффициента Альфы Кронбаха для каждой переменной, подсчитанного для всей шкалы в предположении, что данная переменная исключена.

Статистика надежности

Альфа Кронбаха	N элементов
,627	7

Статистика пунктов по отношению к суммарному баллу

	Шкалировать среднее при исключении пункта	Шкалировать дисперсию при исключении пункта	Исправленная корреляция между пунктом и итогом	Альфа Кронбаха при исключении пункта
ВКонтакте стал частью моей повседневной жизни	17,24	14,820	,513	,813
Я чувствую себя частью сообщества пользователей ВКонтакте	17,59	14,346	-,165	,895
Я был бы огорчен, если ВКонтакте отключат	17,00	13,906	,309	,899
Я использую ВКонтакте, чтобы узнать больше о том, кого я встретил лично оффлайн	17,38	14,527	,668	,805
Я использую ВКонтакте, чтобы узнать больше о других людях, живущих рядом со мной	17,64	14,159	,688	,802
Я использую ВКонтакте, чтобы быть на связи со своими давними друзьями	17,01	15,552	,696	,815
Я использую ВКонтакте, чтобы познакомиться с новыми людьми	17,95	15,580	,628	,823

Рис. 2.11. Исходная статистика пунктов по отношению к суммарному баллу

Значение Альфы 0,627 ниже допустимого 0,7. Более того, в таблице статистике пунктов мы видим, что первые три переменные имеют наименьшую корреляцию с суммой остальных переменных (столбец «Исправленная корреляция между пунктом и итогом», а переменная «Я чувствую себя частью сообщества пользователей Вконтакте» и вовсе имеет отрицательную корреля-

цию. Наконец столбец «Альфа Кронбаха при исключении пункта» демонстрирует более высокие значения коэффициента, если мы удалим эти переменные (пункты) из нашей модели.

3. *Анализ надежности выборочной совокупности переменных*. На данном этапе мы исключаем из анализа слабые переменные и повторяем процедуру. В результате получаем новые значения (рис. 2.12).

Статистика надежности

Альфа Кронбаха	N элементов
,819	4

Статистика пунктов по отношению к суммарному баллу

	Шкалировать среднее при исключении пункта	Шкалировать дисперсию при исключении пункта	Исправленная корреляция между пунктом и итогом	Альфа Кронбаха при исключении пункта
Я использую ВКонтакте, чтобы узнать больше о том, кого я встретил лично оффлайн	7,60	3,035	,817	,587
Я использую ВКонтакте, чтобы узнать больше о других людях, живущих рядом со мной	7,86	3,000	,873	,616
Я использую ВКонтакте, чтобы быть на связи со своими давними друзьями	7,22	3,874	,809	,711
Я использую ВКонтакте, чтобы познакомиться с новыми людьми	8,17	3,605	,841	,695

Рис. 2.12. Итоговая статистика пунктов по отношению к суммарному баллу

Здесь мы наблюдаем более высокое значение Альфы Кронбаха и меньшие значения коэффициента в случае удаления любого пункта. Это говорит о высокой согласованности переменных и пригодности данной методики. Следовательно, именно эту комбинацию пунктов мы оставляем для оценки социального капитала социальных сетей.

4. *Создание переменной*. Допустимое значение коэффициента не только свидетельствует о возможности использования выбранных переменных на репрезентативной выборке, но и позволяет создать новую переменную — социальный капитал. Данная процедура выполняется путем математического сложения указанных переменных (см. раздел «Вычисление переменных»). В результате программа каждому респонденту рассчитает свое значение социального капитала, а в самом массиве появится новая переменная (рис. 2.13).

Статистика

Социальный_капитал

N	Валидные	195
	Пропущенные	0
Мода		10,00
Диапазон		12,00
Минимум		4,00
Максимум		16,00

Социальный_капитал

		Частота	Проценты	Валидный процент	Накопленный процент
Валидные	4,00	7	3,6	3,6	3,6
	5,00	1	,5	,5	4,1
	6,00	8	4,1	4,1	8,2
	7,00	4	2,1	2,1	10,3
	8,00	14	7,2	7,2	17,4
	9,00	24	12,3	12,3	29,7
	10,00	44	22,6	22,6	52,3
	11,00	33	16,9	16,9	69,2
	12,00	35	17,9	17,9	87,2
	13,00	13	6,7	6,7	93,8
	14,00	7	3,6	3,6	97,4
	15,00	2	1,0	1,0	98,5
	16,00	3	1,5	1,5	100,0
	Всего	195	100,0	100,0	

Рис. 2.13. Статистика вновь созданной переменной «Социальный капитал»

На рисунке 2.13 видно, что значения по социальному капиталу социальных сетей варьируют от 4 до 16, а самым часто встречающимся значением является 10 (у 22,6% респондентов). Данная количественная переменная может быть преобразована в порядковую с выделением уровней социального капитала (например, низкий, средний и высокий) и может быть использована во всех других способах статистического анализа данных (которые будут рассмотрены нами далее). Таким образом, анализ надежности позволяет не только оценить согласованность методики, но создавать новые переменные, которые, несомненно, обогатят результаты любого исследования.

§ 2.4. Анализ множественных ответов

Создание набора множественных ответов также предполагает определенные манипуляции с данными и специфичность их интерпретации, именно поэтому этот метод анализа также рассмотрен в главе по преобразованию переменных.

Предваряя анализ множественных ответов, кратко охарактеризуем частотный анализ, поскольку по сути они идентичны. Важнейшее различие лишь в том, что первый используется для закрытых вопросов с несколькими вариантами ответов, а второй — с одним вариантом.

Примером частотного распределения данных является таблица 2.1. Каждый респондент был отнесен к одной из четырех возрастных групп. Столбец «Частоты» показывает количество единиц анализа, отнесенных к каждой группе, столбец «Проценты» рассчитывает проценты по каждой строке от всего массива, включая пропущенные здесь они отсутствуют, поэтому значения совпадают со столбцом «Валидный процент». Столбец «Валидный процент» рассчитывает процент каждой строки от заданного исследователем набора данных. Столбец «Накопленный процент» представляет собой сумму значений каждой следующей строки. В случае с частотным анализом в этой таблице следует включать данные из столбца «Валидный процент».

Анализ множественных ответов состоит из двух этапов: 1) создание набора; 2) анализ.

Процедура «Задать наборы множественных ответов» объединяет переменные в группы, состоящие из множественных категорий. Каждый набор должен иметь уникальное имя и может быть удален из списка нажатием на кнопку «Удалить». Также возможно изменить характеристики набора, выделив его в списке и нажав на кнопку «Изменить».

Чтобы задать набор множественных ответов, переходим в соответствующее окно: анализ — множественные ответы — задать наборы переменных. В поле «Параметры набора» выделяем все переменные, из которых будет состоять набор, и переносим в поле «Переменные в наборе». Ниже выбираем категории — числовые значения, которыми были закодированы варианты

ответа. В нашем случае — от 0 до 99. Далее присваиваем имя набору и кликаем «Добавить» (рис. 2.14). После этого закрываем окно, исключительно кликнув на «Закрыть» (закрытие окна любым другим образом не позволяет сохранить изменения).

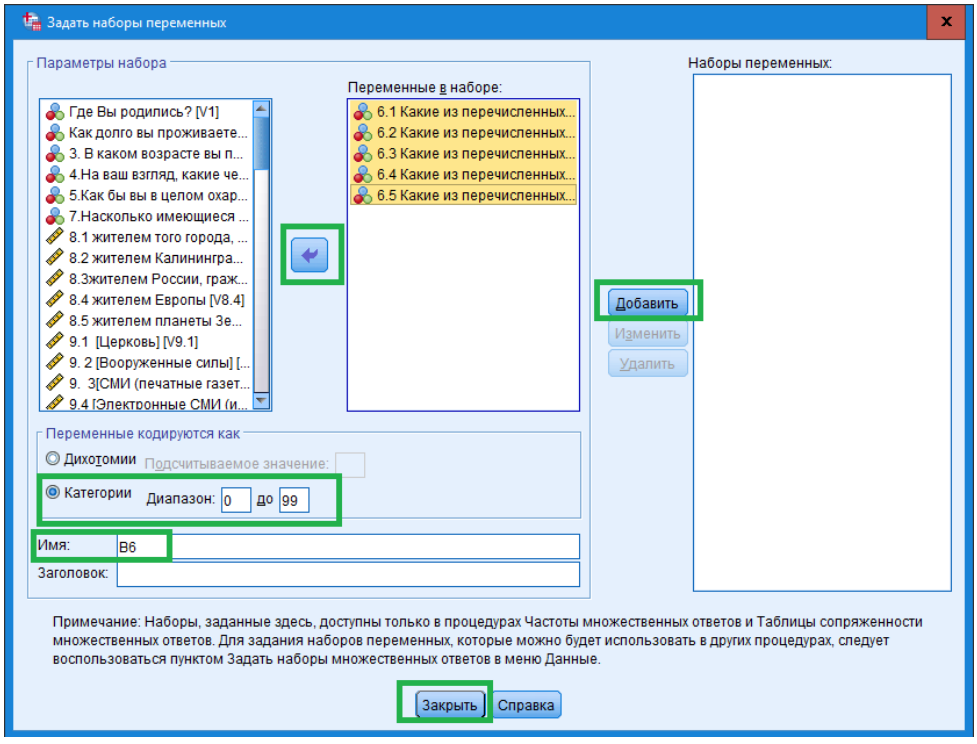


Рис. 2.14. Интерфейс функции «Множественные ответы»

После этого переходим по команде: анализ — множественные ответы — частоты. В открывшемся окне выбираем заданный набор и переносим в поле «Таблицы для», кликаем «ОК».

В результате в окне «Вывод» появится таблица, аналогичная той, что представлена ниже (рис. 2.15). Столбец «N» показывает количество респондентов, выбравших тот или иной вариант ответ. Столбец «Проценты» демонстрирует, какую долю каждый вариант занимает в общей суммы полученных ответов. Но поскольку в таких вопросах предполагается несколько вариантов ответа, то корректные данные будут представлены в столбце «Процент наблюдений», так как здесь значения считаются от числа респондентов. Другими словами, первую строчку таблицы можно интерпретировать следующим образом: 357 респондентов считают, что агрессивное поведение европейских стран вызывает лично у них наибольшее недовольство. Это 10,2 % данных ответов и 39,4 % респондентов.

Периодичность $B6

		Ответы		Процент наблюдений
		N	Проценты	
$B6[a]	Агрессивное поведение европейских стран	357	10,2%	39,4%
	Высокие цены на продукты и товары первой необходимости	449	12,9%	49,6%
	Неравенство в уровне жизни населения	170	4,9%	18,8%
	Высокие тарифы на услуги ЖКХ	282	8,1%	31,1%
	Низкий уровень доходов населения	475	13,6%	52,4%
	Приток мигрантов в регион	118	3,4%	13,0%
	Посещение других регионов России	125	3,6%	13,8%
	Транзит грузов в Калининградскую область	229	6,6%	25,3%
	Коррупция в органах власти	194	5,6%	21,4%
	Эпидемиологическая ситуация с COVID-19	90	2,6%	9,9%
	Наркомания и алкоголизм среди населения	152	4,4%	16,8%
	Угроза безработицы	159	4,6%	17,5%
	Ситуация в сфере образования	83	2,4%	9,2%
	Неблагоустроенные дворы и придомовые территории	112	3,2%	12,4%
	Недоступность, дороговизна жилья	234	6,7%	25,8%
	Террористическая угроза	48	1,4%	5,3%
	Положение дел в сельском хозяйстве	17	0,5%	1,9%
	Экологическая ситуация, состояние окружающей среды	62	1,8%	6,8%
	Состояние дорог	99	2,8%	10,9%
	Нет таких проблем	33	0,9%	3,6%
Всего		3488	100,0%	385,0%

а. Сгруппировать

Рис. 2.15. Статистика анализа множественных ответов

§ 2.5. Контрольные вопросы

1. Какой функционал выполняет каждое окно в программе IBM SPSS Statistics?
2. В чем практический смысл преобразования и вычисления переменных?
3. В чем практический смысл взвешивания данных? Как он выполняется?
4. Что такое таблица множественных ответов и как она создается?

§ 2.6. Практические задания

1. Преобразуйте в порядковую переменную «Возраст», используя массив «Практики использования социальной сети “ВКонтакте”», n = 195, 2023 г.
2. Выполните визуальную категоризацию исходной переменной «Возраст», используя массив «Практики использования социальной сети “ВКонтакте”», n = 195, 2023 г.
3. Вычислите среднедушевой доход, используя массив RLMS-HSE за 2019 г. (https://www.hse.ru/rlms/).
4. Выполните взвешивание данных по полу и возрасту, используя массив «Социально-политические настроения в Калининградской области», n = 915, 2022 г. (https://kantiana.ru/science/baza-dannykh/).
5. Создайте переменную «Социальный капитал», используя массив «Практики использования социальной сети “ВКонтакте”», n = 195, 2023 г.

§ 2.7. Рекомендуемая литература

1. *Бююль А., Цёфель П.* SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей. СПб. : ДисСофтЮп, 2005.
2. *Воронин Г. Л.* Статистический анализ данных в IBM SPSS Statistics V27.0.1.0. Н. Новгород : ННГУ им. Н. И. Лобачевского, 2022.
3. *Наследов А. Д.* SPSS 19: профессиональный статистический анализ данных. СПб. : Питер, 2011.
4. *Щекотуров А. В.* Влияние Facebook[4] на развитие социального капитала студентов // Коммуникативные практики современной молодежи: перспективы и вызовы : материалы междунар. науч.-практ. конф., Нижний Новгород, 15—16 сентября 2022 года. Н. Новгород : Национальный исследовательский Нижегородский государственный университет им. Н. И. Лобачевского, 2022. С. 618—622.
5. *Ellison N., Gray R., Lampe C. & Fiore, A.* Social capital and resource requests on Facebook. New Media & Society 16 (7). 2014. P. 1104—1121.

[4] Принадлежит корпорации Meta, признанной экстремистской и запрещенной на территории РФ.

Глава 3
АНАЛИЗ ТАБЛИЦ СОПРЯЖЕННОСТИ

§ 3.1. Цель и алгоритм выполнения анализа

Таблицы сопряженности служат для описания связи между номинальными и порядковыми переменными, когда значения одной переменной образуют строки, а значения другой — столбцы таблицы.

Этот метод является одним из самых популярных в социологических исследованиях, поскольку позволяет выявлять общие и специфичные характеристики различных подгрупп в структуре как выборки, так и всей генеральной совокупности. Построение таблиц сопряженности — достаточно простой метод, и потому его легко использовать как начинающим, так и опытным ученым во всех отраслях социологии.

Например, исследование поколенческих различий в восприятии различных аспектов региональной истории целиком основано на методе таблиц сопряженности [2]. Ниже представлен рисунок, демонстрирующий связь между отношением к полуразрушенным немецким зданиям в Калининградской области и четырьмя поколениями (рис. 3.1)[5].

Если бы у вас была возможность решить судьбу полуразрушенных зданий немецкого архитектурного наследия в Калининградской области, каким было бы ваше решение? (%)

Поколение	Вариант ответа				
	Здания были бы восстановлены в оригинальном виде	Здания были бы восстановлены, но их внешний вид был бы изменен	Здания были бы оставлены нетронутыми	Здания были бы снесены	Затрудняюсь ответить
Поколение Z	76,4	14,4	3,4	3,4	2,3
Миллениалы ЗС	66,2	19,1	5,5	3,3	5,9
Старшие миллениалы	60,2	19,7	5,9	4,8	9,3
Старшие поколения	50	11,9	1,2	15,6	21,3

Рис. 3.1. Пример таблицы сопряженности

Источник: [2, с. 84].

[5] Таблицы сопряженности часто используются и в анализе социально-политических настроений [3].

Выполнение данного метода статистического анализа предполагает прохождение семи этапов:

1) проверка переменных на нормальное распределение и выбор подходящего коэффициента связи;

2) формулирование исследовательского вопроса, то есть определение того, что именно необходимо узнать о явлениях, признаках, процессах или характере связей между ними [4, с. 83];

3) определение зависимых и независимых переменных;

4) указание нулевой и альтернативной гипотез;

5) проверка нулевой гипотезы посредством оценки асимптотической значимости хи-квадрата;

6) оценка силы связи переменных;

7) описание полученных значений в таблице сопряженности.

В данном параграфе мы рассмотрим анализ номинальных и порядковых переменных с применением параметрических коэффициентов связи (нормальное распределение).

Например, наш исследовательский вопрос звучит так: различается ли интерес к периодам в истории Калининградской области среди жителей региона с высшим образованием и без него?[6]

Проверка переменных по критерию Колмогорова — Смирнова, показала, что распределение значений не отличается от нормального, следовательно, мы можем выполнить анализ с использованием стандартных (параметрических) коэффициентов связи. Параметрический критерий — это метод статистического вывода, который применяется в отношении параметров генеральной совокупности. Самым главным условием для параметрических методов является нормальность распределения переменных и, как следствие, правомерность применения таких статистик, как среднее значение и стандартное отклонение [5, с. 164].

Исходя из исследовательского вопроса, независимой переменной выступает пол респондента, а зависимой — интерес к периодам в истории региона.

Нулевая гипотеза — предположение о том, что переменные не зависят друг от друга, альтернативная гипотеза — переменные статистически взаимосвязаны.

Далее мы строим таблицы сопряженности. В программе IBM SPSS Statistics выбираем: анализ — описательные статистики — таблицы сопряженности. Там в поле для строк переносим зависимую переменную, в поле для столбцов — независимую[7]. В окне «ячейки» выбираем процент там, где расположена независимая переменная (в нашем случае — столбцы). В окне «Статистики» выбираем «Значение хи-квадрата» и «Фи и V Крамера». Кликаем «ОК» и «Продолжить». Далее переходим в окно вывода и анализируем данные.

[6] Опрос жителей Калининградской области (тема: «Историческая память населения Калининградской области», n=997, 2022 г.).

[7] Переменные могут располагаться и иначе, все зависит от внешнего вида таблицы.

На следующем этапе нам требуется проверить нулевую гипотезу. Для этого необходимо оценить асимптотическую значимость хи-квадрата Пирсона (рис. 3.2). Если она выше 0,05 ($p > 0{,}05$), значит мы принимаем нулевую гипотезу и завершаем анализ. Если значимость ниже или равна 0,05 ($p \leqslant 0{,}05$), мы опровергаем нулевую гипотезу, принимаем альтернативную и продолжаем анализировать. Более того, при условии репрезентативности выборки критерий хи-квадрата Пирсона позволяет «определить наличие или отсутствие взаимосвязи между двумя переменными с экстраполяцией полученной информации на генеральную совокупность» [1, с. 80].

Критерии хи-квадрат

	Значение	ст.св.	Асимптотическая значимость (2-сторонняя)
Хи-квадрат Пирсона	15,572[a]	4	,004
Отношения правдоподобия	15,719	4	,003
Линейно-линейная связь	10,328	1	,001
Количество допустимых наблюдений	987		

a. Для числа ячеек 0 (0,0%) предполагается значение, меньше 5. Минимальное предполагаемое число равно 54,78.

Рис. 3.2. Значение хи-квадрата

В нашем случае $p = 0{,}004$, следовательно, мы устанавливаем факт влияния высшего образования на интерес к различным периодам в истории региона. И утверждаем с вероятностью 99 %, что в генеральной совокупности, то есть среди жителей Калининградской области от 18 лет и старше, интерес к региональной истории для респондентов с высшим образованием статистически значимо отличается от заинтересованности респондентов без него.

На следующем этапе мы должны оценить силу влияния независимой переменной на зависимую. Для этого используем коэффициент V Крамера[8].

V Крамера — это мера связи, основанная на статистике хи-квадрат и изменяется в интервале от 0 до 1, где 0 означает отсутствие взаимосвязи между переменными, значения ближе к 1 говорят о наличии такой взаимосвязи.

[8] Фи используется в случае с дихотомными переменными, когда каждая переменная состоит не более чем из двух градаций (например, пол).

Значение выше 0,7 свидетельствует о высокой силе связи, от 0,3 до 0,7 — об умеренной силе, ниже 0,3 — слабая сила связи. В нашем случае значение V Крамера составляет всего 0,126 (рис. 3.3), что говорит, о слабом влиянии уровня образования на интерес к определенным периодам в истории Калининградской области.

Симметричные меры

		Значение	Приблизительная значимость
Номинал/номинал	Фи	,126	,004
	V Крамера	,126	,004
Количество допустимых наблюдений		987	

Рис. 3.3. Значение симметричных мер связи

Наконец, мы должны выполнить анализ получившегося распределения двух переменных. На рисунке 3.4 мы видим, что вне зависимости от наличия высшего образования наиболее интересным периодом является прусский (34,1 %), а советский и российский периоды вызывают одинаковый интерес у каждой группы респондентов. Однако, выделяя фактор высшего образования, мы фиксируем, что статистически достоверно респонденты, окончившие высшие учебные заведения, больше интересуются прусско-немецким периодом, а респонденты без высшего образования — более чем в два раза склонны вообще не интересоваться историей региона.

Перекрестная таблица

			Образование		
			Есть высшее образование	Нет высшего образования	Всего
1. Какой период истории нашего региона вызывает у Вас наибольший интерес? (один вариант ответа)	Прусский (донемецкий)	Количество	171	169	340
		% в Образование	38,9%	30,3%	34,1%
	Немецкий	Количество	98	98	196
		% в Образование	22,3%	17,6%	19,7%
	Советский	Количество	78	102	180
		% в Образование	17,7%	18,3%	18,1%
	Российский (современный)	Количество	54	72	126
		% в Образование	12,3%	12,9%	12,6%
	Не интересуюсь историей региона	Количество	39	116	155
		% в Образование	8,9%	20,8%	15,5%
Всего		Количество	440	557	997
		% в Образование	100,0%	100,0%	100,0%

Рис. 3.4. Итоговая таблица сопряженности

В завершении анализа принято делать краткий вывод в виде ответа на исследовательский вопрос с указанием на значения всех используемых коэффициентов: интерес к различным периодам в истории Калининградской области достоверно различается среди жителей региона с высшим образованием и без него (p = 0,004). Прусско-немецкие годы региона вызывают наибольший интерес среди всех респондентов (с небольшим перевесом в пользу респондентов с высшим образованием), советско-российский период интересен меньшему количеству респондентов вне зависимости от образования. Опрошенные без высшего образования значительно более склонны не интересоваться историей региона вовсе. Однако фактор высшего образования не является определяющим в данном вопросе (V Крамера = 0,126).

Если распределение значений не отличается от нормального (тест Колмогорова — Смирнова), и одна переменная выражена в порядковой шкале, а вторая — в номинальной, то используется Лямбда[9], если обе переменных порядковые — Гамма[10] (табл. 3.1).

Таблица 3.1

Соотношение типа переменной и коэффициента связи

Тип переменной	Номинальная	Порядковая
Номинальная	Фи	Лямбда
Порядковая	Лямбда	Гамма

Гамма — симметричная мера связи между двумя порядковыми переменными (– 1 и 1). Значения, близкие по абсолютной величине к 1, указывают на сильную связь переменных. Значения, близкие к 0, говорят о слабой связи или ее отсутствии.

Если распределение значений отличается от нормального, алгоритм анализа остается прежним, но используются непараметрические коэффициенты связи: Тау-b и Тау-c Кендалла[11]. Непараметрические методы позволяют ис-

[9] Лямбда — мера связи, которая отражает относительное снижение ошибки, когда значения независимой переменной используются для предсказания значений зависимой переменной: значение 1 — независимая переменная точно предсказывает значения зависимой; значение 0 — независимая переменная абсолютно бесполезна для предсказания зависимой [6].

[10] Гамма — мера связи между двумя порядковыми переменными, значения которой меняются между – 1 и + 1. Значения по модулю, близкие по абсолютной величине к 1, указывают на сильную связь переменных. Значения, близкие к 0, говорят о слабой связи или ее отсутствии. Знак коэффициента указывает направление связи, а его модуль — ее силу, причем, чем он больше, тем связь сильнее [6].

[11] Тау-b Кендалла — непараметрическая мера корреляции для порядковых или ранговых переменных, которая учитывает возможные совпадения значений (связи). Знак

следовать данные без каких-либо допущений о характере распределения переменных, в том числе при нарушении требования нормальности распределения [5, с. 164].

§ 3.2. Многослойные таблицы сопряженности

В программе IBM SPSS Statistics есть возможность для конструирования многослойных таблиц сопряженности, когда в анализ включаются не две, а более переменных. Каждая дополнительная переменная становится внешним слоем для предыдущих. Это позволяет вычленять более узкие группы из всего массива данных и тем самым давать более точное описание социальных категорий.

В рисунке 3.5 представлен пример двухслойной таблицы сопряженности между переменными пол, наличие высшего образования и интерес к различным периодам в истории Калининградской области. Внешним слоем является пол, проценты выведены по строке. Читается эта таблица следующим образом: из 100 % мужчин с высшим образованием 40,7 % интересуются прусской (донемецкой) историей региона, в то время как среди мужчин без высшего образования таких оказалось 34,4 %. И так далее.

Комбинационная таблица Образование * 1. Какой период истории нашего региона вызывает у Вас наибольший интерес? (один вариант ответа) * 17. Укажите Ваш пол:

17. Укажите Ваш пол:				1. Какой период истории нашего региона вызывает у Вас наибольший интерес? (один вариант ответа)					
				Прусский (донемецкий)	Немецкий	Советский	Российский (современный)	Не интересуюсь историей региона	Всего
Мужской	Образование	Есть высшее образование	Количество	72	42	31	16	16	177
			% в Образование	40,7%	23,7%	17,5%	9,0%	9,0%	100,0%
		Нет высшего образования	Количество	89	59	42	25	44	259
			% в Образование	34,4%	22,8%	16,2%	9,7%	17,0%	100,0%
	Всего		Количество	161	101	73	41	60	436
			% в Образование	36,9%	23,2%	16,7%	9,4%	13,8%	100,0%
Женский	Образование	Есть высшее образование	Количество	99	56	47	38	23	263
			% в Образование	37,6%	21,3%	17,9%	14,4%	8,7%	100,0%
		Нет высшего образования	Количество	73	39	59	45	72	288
			% в Образование	25,3%	13,5%	20,5%	15,6%	25,0%	100,0%
	Всего		Количество	172	95	106	83	95	551
			% в Образование	31,2%	17,2%	19,2%	15,1%	17,2%	100,0%
Всего	Образование	Есть высшее образование	Количество	171	98	78	54	39	440
			% в Образование	38,9%	22,3%	17,7%	12,3%	8,9%	100,0%
		Нет высшего образования	Количество	162	98	101	70	116	547
			% в Образование	29,6%	17,9%	18,5%	12,8%	21,2%	100,0%
	Всего		Количество	333	196	179	124	155	987
			% в Образование	33,7%	19,9%	18,1%	12,6%	15,7%	100,0%

Рис. 3.5. Многослойная таблица сопряженности

коэффициента указывает направление связи, а его модуль — силу связи, причем, чем он больше, тем связь сильнее. Значения изменяются в диапазоне между –1 и +1, который можно получить только для квадратных таблиц.

Тау-с Кендалла — непараметрическая мера связи для порядковых переменных, игнорирующая возможные совпадения значений (связи). Знак коэффициента указывает направление связи, а его модуль — силу связи, причем, чем он больше, тем связь сильнее. Значения изменяются в диапазоне между –1 и +1, однако его можно получить только для квадратных таблиц [6].

Если запрошен вывод статистик и мер силы связи, то они вычисляются только для переменных внешнего слоя — в нашем случае по полу.

§ 3.3. Таблицы сопряженности для множественных ответов

Довольно распространенной практикой является построение таблиц сопряженности для переменных с множественными вариантами ответов. В таком случае задается набор переменных и в этом же окне строятся таблицы сопряженности (см. § 2.4). Там же можно добавлять новые слои, если это необходимо (см. § 3.2).

Пример такой таблицы сопряженности представлен на рисунке 3.6, поскольку вопрос про источники информации об истории Калининградской области предполагал любое количество вариантов ответа. Проценты в этой таблице выведены по столбцу, следовательно, правильно она читается так: респонденты, которые испытывают больший интерес к прусскому периоду в истории региона, получают информацию на интернет-сайтах (76,3 %), в музеях (47,9 %) и т. д.

Комбинационная таблица $B11 * @1

			1. Какой период истории нашего региона вызывает у Вас наибольший интерес? (один вариант ответа)					
			Прусский (донемецкий)	Немецкий	Советский	Российский (современный)	Не интересуюсь историей региона	Всего
$B11[a]	Научная и научно-популярная литература	Количество	127	71	53	20	13	284
		% в @1	38,0%	37,2%	30,1%	16,4%	9,0%	
	Документальные фильмы	Количество	134	73	67	32	33	339
		% в @1	40,1%	38,2%	38,1%	26,2%	22,9%	
	Школы, вузы	Количество	103	56	60	41	40	300
		% в @1	30,8%	29,3%	34,1%	33,6%	27,8%	
	Музеи	Количество	160	94	84	53	35	426
		% в @1	47,9%	49,2%	47,7%	43,4%	24,3%	
	Интернет-сайты	Количество	255	139	109	73	93	669
		% в @1	76,3%	72,8%	61,9%	59,8%	64,6%	
	Рассказы родственников, друзей, знакомых	Количество	138	88	63	35	63	387
		% в @1	41,3%	46,1%	35,8%	28,7%	43,8%	
	Социальные сети	Количество	141	83	58	51	48	381
		% в @1	42,2%	43,5%	33,0%	41,8%	33,3%	
Всего		Количество	334	191	176	122	144	967

Рис. 3.6. Таблица сопряженности для множественных ответов

Недостатком данного метода анализа данных является невозможность статистической проверки нулевой гипотезы и оценки силы связи переменных.

§ 3.4. Контрольные вопросы

1. В чем цель анализа таблиц сопряженности?
2. Каков алгоритм выполнения анализа таблиц сопряженности?
3. Каковы параметрические и непараметрические коэффициенты связи при анализе таблиц сопряженности?

4. В чем особенность многослойных таблиц сопряженности?
5. Каковы особенности таблиц сопряженности для множественных ответов?

§ 3.5. Практические задания

1. Выполните анализ таблиц сопряженности между полом и оценкой ситуации в регионе и стране, используя массив «Социально-политические настроения в Калининградской области», n=915, 2022 г. (https://kantiana.ru/science/baza-dannykh/).

2. Выполните анализ таблиц сопряженности между возрастом и оценкой ситуации в регионе и стране, используя массив «Социально-политические настроения в Калининградской области», n=915, 2022 г. (https://kantiana.ru/science/baza-dannykh/).

§ 3.6. Рекомендуемая литература

1. *Воронин Г.Л.* Статистический анализ данных в IBM SPSS Statistics V27.0.1.0. Н. Новгород : ННГУ им. Н.И. Лобачевского, 2022.

2. *Кришталь М.И.* Поколенческие различия жителей Калининградской области в восприятии региональной истории и историкокультурного наследия // Вестник Балтийского федерального университета им. И. Канта. Сер.: Гуманитарные и общественные науки. 2023. № 1. С. 77—89.

3. *Кришталь М.И., Щекотуров А.В.* Методология и методика анализа социально-политических настроений : учеб. пособие. Калининград : Страж Балтики, 2021.

4. *Методология* и методы социологического исследования : учебник / под ред. В.И. Дудиной, Е.Э. Смирновой; СПб. : Изд-во СПбГУ, 2014.

5. *Наследов А.Д.* SPSS 19: профессиональный статистический анализ данных. СПб. : Питер, 2011.

6. *Статистики*, рассчитываемые для таблиц сопряженности. URL: https://www.ibm.com/docs/ru/spss-statistics/26.0.0?topic=crosstabs-statistics (дата обращения: 05.10.2023).

Глава 4
СРАВНЕНИЕ СРЕДНИХ ЗНАЧЕНИЙ: Т-КРИТЕРИЙ

§ 4.1. Т-критерий для независимых выборок

Сравнение средних значений различных выборок — один из самых распространенных методов статистического анализа. Он позволяет определить, является ли различие между средними значениями статистически значимым или объясняется случайными колебаниями.

При сравнении средних значений выборок предполагается, что обе выборки имеют нормальное распределение данных. В случае, если это условие не выполняется, вычисляются медианы, а для сравнения выборок применяются непараметрические тесты.

В социологии сравнение средних чаще всего используется при выявлении различий по количественной переменной у представителей разных социодемографических групп. Например, t-тест позволяет проверить, существуют ли достоверные различия в оплате труда у мужчин и женщин или у жителей разных типов населенных пунктов.

Существует три варианта t-критерия:

1) для независимых выборок: предназначен для сравнения средних значений двух выборок одной переменной (например, пол: мужской и женский);

2) для парных выборок: используется для сравнений средних значений нескольких переменных одной группы между собой (например, значения какого-либо показателя до и после воздействия на группу);

3) одновыборочный t-критерий: разработан для сравнения средних одной переменной группы с «эталонным» значением (например, сравнение результатов тестирования с проходным, пороговым значением).

Примером научной публикации, в которой исследование основано на использовании t-критерия Стьюдента для независимых выборок, является работа А. С. Кубековой [4]. В ней рассмотрены механизмы адаптации иностранных студентов из двух групп: с низкой и высокой степенью адаптивности. Один из результатов представлен на рисунке ниже (рис. 4.1).

Защитный механизм	Группа	
	Низкая адаптивность (1-я группа)	Высокая адаптивность (2-я группа)
Отрицание	7,8 ± 35,9 *	3,9 ± 41,5
Подавление	8,2 ± 24,5 *	3,5 ± 18,1
Регрессия	7,6 ± 23,6 *	2,7 ± 11,2
Компенсация	4,9 ± 14,4	8,8 ± 13,6 *
Проекция	6,1 ± 26,9	9,3 ± 3,4
Замещение	5,3 ± 34,8	4,1 ± 3,6
Интеллектуализация	4,8 ± 13,4	5,3 ± 11,2
Регрессивное образование	3,6 ± 44,6	4,5 ± 6,4

* t-критерий Стьюдента ($p < 0,05$), достоверные различия между группами

Рис. 4.1. Пример использования t-критерия Стьюдента

Источник: [4].

Применение t-критерия предполагает выполнение следующих шагов:

1. Проверка переменных на нормальное распределение и выбор подходящего коэффициента связи.
2. Формулирование исследовательского вопроса.
3. Определение зависимых и независимых переменных.
4. Указание нулевой и альтернативной гипотез.
5. Проверка нулевой гипотезы.
6. Описание полученных значений в таблице средних значений.

Ниже представлены примеры анализа средних при условии, что данные в выборках распределены нормально.

Итак, сформулируем исследовательский вопрос следующим образом: в каком возрасте молодые люди больше интересуются переездом за границу[12]. Соответственно, зависимая переменная — возраст, независимая — интерес к переезду. Нулевая гипотеза — переменные независимы, альтернативная — переменные связаны, то есть различия в возрасте между теми, кто рассматривает и не рассматривает переезд за границу действительно существуют.

В таблице «Статистика группы» (рис. 4.2) показано количество респондентов, которые выбрали определенный вариант ответа («Да» или «Нет»), их средний возраст и его характеристики: стандартное отклонение и среднеквадратическая ошибка. Однако первостепенное значение имеет показатель p-уровня в таблице «Критерий для независимых выборок». Он равняется 0,011 в случае, если предполагаются равные дисперсии, и 0,009 в случае, если они не предполагаются. Оба значения менее 0,05, следовательно, мы отклоняем нулевую гипотезу и можем утверждать, что возрастные различия существуют. Далее мы возвращаемся в таблицу «Статистика группы» и делаем статистически достоверный вывод о том, что молодые люди, рассматривающие переезд за границу, моложе тех, кто его не рассматривает.

Статистика группы

	Скажите, пожалуйста, рассматриваете ли Вы в течение следующих 3-4 лет переезд за границу?	N	Среднее	Среднекв. отклонение	Среднекв. ошибка среднего
Сколько Вам лет?	Да	147	24,39	5,443	,449
	Нет	466	25,76	5,704	,264

Критерий для независимых выборок

		Критерий равенства дисперсий Ливиня		t-критерий для равенства средних						
									95% доверительный интервал для разности	
		F	Значимость	т	ст.св.	Знач. (двухсторонняя)	Средняя разность	Среднеквадратичная ошибка разности	Нижняя	Верхняя
Сколько Вам лет?	Предполагаются равные дисперсии	1,756	,186	-2,558	611	,011	-1,365	,534	-2,413	-,317
	Не предполагаются равные дисперсии			-2,621	255,061	,009	-1,365	,521	-2,391	-,339

Рис. 4.2. Таблицы анализа t-критерия для независимых выборок

[12] По результатам онлайн-опроса молодежи (18—35 лет) Калининградской области (n=987) в марте 2021 г. Вопрос был сформулирован так: «Скажите, пожалуйста, рассматриваете ли Вы в течение следующих 3—4 лет переезд за границу?». В анализ включены два варианта: «Да» и «Нет».

Если распределение отличается от нормального или зависимая переменная выражена в порядковой шкале для двух независимых выборок, используют непараметрический критерий Манна — Уитни[13].

Например, нас интересует, существуют ли достоверные различия в возрасте молодежи, которая очень интересуется деятельностью российских политиков, и той, что не интересуется ею вовсе[14]. Проверка по тесту Колмогорова — Смирнова показала, что распределения отличаются от нормального. Следовательно, мы используем критерий Манна — Уитни. При реализации метода программа сначала ранжирует все объекты без учета принадлежности к сравниваемым группам, а затем вычисляет средние ранги для каждой из двух групп. Чем выше средний ранг группы, тем выше сравниваемый показатель: в нашем случае — возраст. После нахождения средних рангов определяется p-уровень.

Для установки переменных переходим в соответствующее окно: анализ — непараметрические критерии — устаревшие диалоговые окна — для двух независимых выборок. Далее указываем группирующую и проверяемую переменную и выбираем U-критерий Манна — Уитни. Кликаем «ОК». В окне вывода программа продемонстрирует две таблицы: «Ранги» и «Статистические критерии» (рис. 4.3). Асимптотическая значимость менее 0,05, что позволяет нам отклонить нулевую гипотезу, а тот факт, что средний ранг значений респондентов, не интересующихся деятельностью российских политиков, выше, говорит, что и средний возраст этих респондентов выше.

Выяснить средние значения возраста можно так: анализ — сравнение средних — средние. В новом окне указываем зависимую и независимую переменную и кликаем «ОК». Результаты будут представлены в окне вывода (рис. 4.4). Теперь мы не только видим, что средний возраст тех, кто выбрал вариант «совсем не интересна» составляет 26,13 лет, а средний возраст, указавших «очень интересна» — 24,87 года, но и знаем, что эти различия не случайны, а статистически достоверны.

---

[13] Критерий Манна — Уитни проверяет гипотезу о том, что две генеральные совокупности, из которых были отобраны выборки, эквивалентны по расположению. Наблюдения из обеих групп объединяются и ранжируются, причем совпадающим значениям назначается средний ранг. Количество совпадающих значений должно быть мало по сравнению с общим количеством наблюдений. Если проверяемые совокупности эквивалентны по расположению, то ранги должны быть распределены между двумя выборками случайным образом. При расчете критерия подсчитываются число раз, когда значение из группы 1 предшествует значению из группы 2, и число раз, когда значение из группы 2 предшествует значению из группы 1. U-статистикой Манна — Уитни является меньшее из этих двух чисел [6].

[14] Опрос молодежи Калининградской области (тема: «Социально-политические настроения молодежи Калининградской области», n = 987, 2021 г.).

Критерий Манна-Уитни

Ранги

	Насколько в целом Вам интересна деятельность российских политиков	N	Средний ранг	Сумма рангов
Сколько Вам лет?	совсем не интересна	195	156,33	30485,00
	очень интересна	95	123,26	11710,00
	Всего	290		

Статистические критерии[a]

	Сколько Вам лет?
U Манна-Уитни	7150,000
W Вилкоксона	11710,000
Z	-3,161
Асимп. знач. (двухсторонняя)	,002

a. Группирующая переменная: Насколько в целом Вам интересна деятельность российских политиков

Рис. 4.3. Таблицы анализа t-критерия для независимых выборок методом Манна — Уитни

Отчет

Сколько Вам лет?

Насколько в целом Вам интересна деятельность российских политиков	Среднее	N	Стандартная отклонения
совсем не интересна	26,13	195	5,823
скорее не интересна	24,66	148	5,360
среднее значение	24,48	211	5,717
скорее интересна	24,62	164	5,285
очень интересна	23,92	95	5,563
Всего	24,87	813	5,612

Рис. 4.4. Средний возраст молодежи с разной степенью интереса к деятельности российских политиков

§ 4.2. Т-критерий для парных выборок

Т-критерий для парных выборок используется для сравнения средних значений двух связанных или парных наблюдений. Он применяется в следующих ситуациях:

1. Когда наблюдения собраны в парах или группах, и каждая пара является связанной, то есть каждое наблюдение в одной паре связано с определенным наблюдением в другой паре (например, до и после измерений у тех же самых индивидов).

2. Когда интересует разница между двумя связанными переменными, а не абсолютные значения.

Таким образом, t-критерий для парных выборок позволяет оценить, является ли средняя разница в парах статистически значимой.

Примером научной публикации, в которой исследование основано на использовании t-критерия Стьюдента для парных выборок, является работа А. В. Белинского [1]. В ней рассмотрены изменения реакции группы людей до и после воздействия негативного эмоционального стимула. Один из результатов представлен на рисунке ниже (рис. 4.5). О том, как интерпретировать подобные таблицы, пойдет речь в данном параграфе.

		Парные разности					Т	Ст. св.	Знач. (2-сторонняя)
		среднее	станд. отклонения	станд. средняя ошибка	95 % доверительный интервал для разности				
					нижняя	верхняя			
Пара 1	Контроль – увечье	,20580	,20587	,02911	,14729	,26430	7,069	49	,000
Пара 2	Контроль – угроза	–,37266	,27291	,03859	–,45022	–,29510	–9,656	49	,000

Рис. 4.5. Пример использования t-критерия Стьюдента для парных выборок

Источник: [1].

В качестве примера воспользуемся учебным массивом, подготовленным А. Д. Наследовым [5]. Допустим, у нас есть результаты успеваемости одних и тех же школьников в 10-х и 11-х классах. Соответственно, исследовательский вопрос здесь заключается в том, существуют ли достоверные различия между успеваемостью школьников в 11-м и 10-м классах.

В таблице «Статистика парных выборок» мы видим,что среднее значение отметок в 11-м классе выше (рис. 4.6), но без проверки на p-уровень мы не можем быть уверены в том, что эти различия получены не случайным образом.

Статистика парных выборок

		Среднее	N	Среднекв. отклонение	Среднекв. ошибка среднего
Пара 1	успеваемость в 10 классе	3,9630	100	,30597	,03060
	успеваемость в 11 классе	4,2205	100	,27589	,02759

Рис. 4.6. Статистика парных выборок

В таблице «Критерий парных выборок» (рис. 4.7) мы отмечаем, что уровень двухсторонней значимости меньше 0,05, следовательно, различия в успеваемости между 10-ми и 11-ми классами статистически достоверны. С исследовательской точки зрения, это говорит о существовании каких-либо обстоятельств или условий, которые способствовали повышению успеваемости в 11-м классе. Таким образом, обнаружение статистических связей может послужить основанием для проведения дополнительных научных работ.

Критерий парных выборок

		Парные разности					т	ст.св.	Знач. (двухсторонняя)
		Среднее	Среднекв. отклонение	Среднекв. ошибка среднего	95% доверительный интервал для разности				
					Нижняя	Верхняя			
Пара 1	успеваемость в 10 классе - успеваемость в 11 классе	-,25750	,31062	,03106	-,31913	-,19587	-8,290	99	,000

Рис. 4.7. Критерий парных выборок

Непараметрическим критерием для парных выборок является критерий Вилкоксона, который опирается на подсчет абсолютных различий между парами значений и их последующим ранжированием. Затем вычисляются средние ранги для положительных и отрицательных различий. Уровень значимости определяется на основе стандартизированного значения. Применение этого критерия может быть сомнительным в случае, если переменная имеет небольшое количество возможных значений, например, на 3-балльной шкале. В такой ситуации рекомендуется использовать точные критерии [5, с. 171—172].

Диалоговое окно для критерия Вилкоксона открывается так: анализ — непараметрические критерии — устаревшие диалоговые окна — для двух связанных выборок. Данный критерий будет выбран по умолчанию, добавление переменных для анализа аналогично при параметрическом t-критерии для парных выборок.

Разберем эту команду на примере исследования репродуктивных установок жительниц Калининградской области[15]. Допустим, нас интересует различается ли статистически достоверно количество детей, которых хотели бы иметь респондентки в идеале и исходя из обстоятельств. Выполнив описанные выше процедуры в IBM SPSS Statistics, и не забыв выбрать «Описательные статистики» на кнопке «Параметры», мы получим следующие таблицы в окне вывода (рис. 4.8, 4.9).

[15] Опрос женщин Калининградской области (тема: «Репродуктивные установки жительниц Калининградской области», n = 830, 2017 г.).

Критерий знаковых рангов Вилкоксона

Ранги

		N	Средний ранг	Сумма рангов
Сколько еще детей Вы реально планируете иметь, исходя из обстоятельств? - Сколько еще детей Вам хотелось бы иметь «в идеале»?	Отрицательные ранги	204[a]	107,12	21853,00
	Положительные ранги	8[b]	90,63	725,00
	Совпадающие наблюдения	251[c]		
	Всего	463		

a. Сколько еще детей Вы реально планируете иметь, исходя из обстоятельств? < Сколько еще детей Вам хотелось бы иметь «в идеале»?

b. Сколько еще детей Вы реально планируете иметь, исходя из обстоятельств? > Сколько еще детей Вам хотелось бы иметь «в идеале»?

c. Сколько еще детей Вы реально планируете иметь, исходя из обстоятельств? = Сколько еще детей Вам хотелось бы иметь «в идеале»?

Рис. 4.8. Критерий знаковых рангов Вилкоксона

Статистические критерии[a]

	Сколько еще детей Вы реально планируете иметь, исходя из обстоятельств? - Сколько еще детей Вам хотелось бы иметь «в идеале»?
Z	-12,065[b]
Асимп. знач. (двухсторонняя)	,000

a. Критерий знаковых рангов Вилкоксона

b. На основе положительных рангов.

Рис. 4.9. Асимптотическая значимость по критерию Вилкоксона

В первой таблице (рис. 4.8) мы видим, что средние значения и сумма отрицательных рангов выше для связанной пары исследуемых переменных. Здесь же находим информацию, что ранги отрицательные, когда количество детей, реально планируемых исходя из обстоятельств, меньше количества детей, которых хотелось бы иметь в идеале (a). Следовательно, в целом респондентки в реальности хотят иметь детей меньше, чем в идеале. Но мы не знаем, как именно различается количество детей и есть ли здесь статистически достоверные отличия.

Уровень значимости ($p < 0{,}001$) говорит о статистической достоверности различий.

Само количество детей мы можем наблюдать в таблице «Описательные статистики» (рис. 4.10). Таким образом, идеальное количество детей для респондентов женского пола близится к двум, но в реальности женщины рассчитывают всего на одного ребенка.

Описательные статистики

	N	Среднее	Среднекв. отклонения	Минимум	Максимум
Сколько еще детей Вам хотелось бы иметь «в идеале»?	616	1,87	1,004	0	6
Сколько еще детей Вы реально планируете иметь, исходя из обстоятельств?	475	,98	,943	0	6

Рис. 4.10. Статистика средних значений двух переменных

§ 4.3. Одновыборочный t-критерий

Этот метод анализа разработан для сравнения средних одной переменной группы с «эталонным» значением и находит свое применение в широком спектре научных отраслей: от иммунологии (например, при соотнесении уровня клеточного иммунитета с необходимым — эталонным — значением) до педагогики (например, при соотнесении среднего балла успеваемости группы учеников с проходным баллом по дисциплине).

А. О. Крыштановский так описывает актуальность использования одновыборочного t-критерия в социологии: «В ходе анализа социологических данных нередко возникает ситуация, когда необходимо сравнить среднее значение какой-то количественной переменной с некоторым фиксированным значением. Например, в ходе исследования образа жизни было выяснено, что в среднем респонденты тратят на просмотр телепередач около двух часов. Из материалов предыдущих исследований известно, что год назад респонденты тратили на этот вид деятельности приблизительно 1,8 часа. Можем ли мы, опираясь на эту информацию, утверждать, что за прошедший год люди стали больше времени проводить у телевизора или обнаруженная разница носит случайный, статистически не значимый характер? Другая исследовательская ситуация определяется необходимостью оценки репрезентативности проведенного опроса по количественным показателям. Если, скажем, мы провели всероссийский опрос, для оценки репрезентативности по параметру “возраст” требуется сопоставить данные опроса с материалами, представляемыми органами государственной статистики. Общим в двух рассмотренных примерах является то, что мы должны оценить значимость различий между данными опроса и некоторыми “внешними” цифрами» [2, с. 94].

Оригинальным представляется исследование, проведенное Т. Л. Крюковой и О. А. Екимчук. В своей работе они используют статистические методы анализа для оценки влияния фаббинга[16] на качество романтических отношений. В частности, используя одновыборочный t-критерий Стьюдента, они сравнивают уровень фаббинга партнера с эмпирической нормой, рассчитанной ими же по собственной методике. Авторы приходят к выводу «о наличии партнерского фаббинга в близких отношениях наших респондентов и о его яркой выраженности» [3, с. 69].

Рассмотрим возможности одновыборочного t-критерия на примере следующего исследовательского вопроса: отличается ли среднее количество детей, которых хотели бы иметь жительницы Калининградской области, от среднероссийских показателей?[17]

Для проверки нулевой гипотезы и ответа на исследовательский вопрос выполняем команду: анализ — сравнение средних — одновыборочный t-критерий. В появившемся окне указываем проверяемую переменную и выставляем контрольное значение: 2,2 (столько детей в среднем хотели бы иметь россиянки в 2015 году[18]). Описание полученных значений представлено в таблице «Одновыборочная статистика» (рис. 4.11).

Одновыборочная статистика

	N	Среднее	Станд. отклонения	Станд. средняя ошибка
Сколько еще детей Вам хотелось бы иметь «в идеале»?	616	1,87	1,004	,040

Рис. 4.11. Одновыборочная статистика

Опираясь на полученное значение двухсторонней значимости ($p < 0{,}001$), делаем вывод, что нулевую гипотезу следует отклонить (рис. 4.12). Таким образом, мы можем сказать, что полученные данные о среднем количестве детей, которых хотели бы иметь жительницы Калининградской области, значимо отличается от общероссийских показателей: россияне в целом хотят больше двух детей, калининградки — меньше двух.

[16] Фаббинг — злоупотреблением гаджетами в процессе коммуникации с другими людьми.

[17] Опрос женщин Калининградской области (тема: «Репродуктивные установки жительниц Калининградской области», $n = 830$, 2017 г.).

[18] По материалам публикации [7].

Одновыборочный критерий

Значение критерия = 2.2

	t	ст.св.	знач. (двухсторонняя)	Средняя разность	95% доверительный интервал для разности	
					Нижняя	Верхняя
Сколько еще детей Вам хотелось бы иметь «в идеале»?	-8,073	615	,000	-,327	-,41	-,25

Рис. 4.12. Оценка статистической значимости одновыборочного t-критерия

Если данные не имеют нормального распределения, но необходимо проверить значимость различий между средним значением в одной выборке и известным или предполагаемым значением среднего в генеральной совокупности, можно использовать непараметрический тест. Один из наиболее распространенных непараметрических тестов для такого сравнения — тест знаковых рангов Вилкоксона. Этот тест позволяет сравнить медиану выборки с заданным значением. Он основывается на ранжировании значений выборки, а не на распределении данных, и не требует нормальности данных.

В данном случае диалоговое окно в программе IBM SPSS Statistics открывается иначе. Следует перейти в следующее окно: анализ — непараметрические критерии — одновыборочные. В открывшемся окне перейти ко вкладке «Параметры», активировать кнопку «Выберите критерии» и в поле справа кликнуть на критерий Уилкоксона (рис. 4.13). Обратите внимание, что этот критерий дает результаты относительно медианы, а не среднего значения выборки.

Здесь в качестве исследуемой была взята переменная «Сколько еще детей Вы реально планируете иметь исходя из обстоятельств?». Поэтому в качестве контрольного значения было установлено 2. Соответственно, нулевая гипотеза формулируется так: медиана выборки равна заданному значению среднего в генеральной совокупности. Альтернативная гипотеза: медиана выборки отличается от заданного значения среднего в генеральной совокупности.

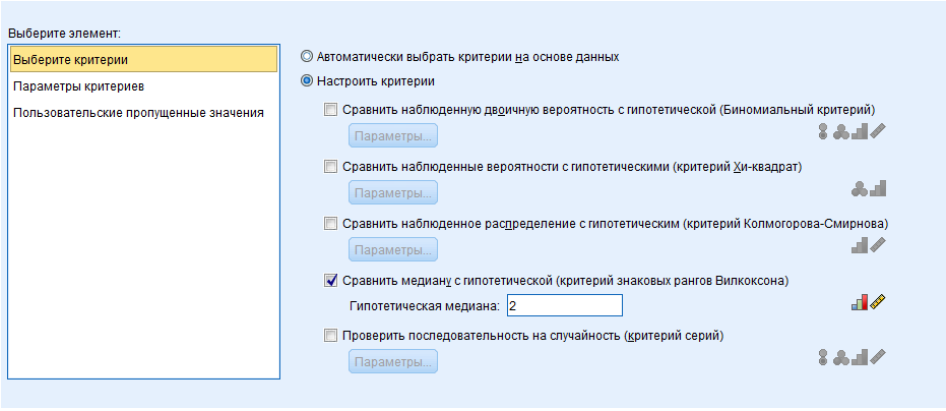


Рис. 4.13. Параметры критерия Вилкоксона
при использовании одновыборочного t-критерия

После установки контрольного значение необходимо кликнуть кнопку «Вставить». В открывшемся окне редактора синтаксиса появится команда, которую необходимо доработать. По умолчанию программа будет сравнивать указанное значение медианы по всем переменным в массиве. Для выбора только необходимой переменной мы указываем ее имя, поэтому команда в окне синтасиса будет выглядеть так:

```
NPTESTS
/ONESAMPLE TEST (B22) WILCOXON(TESTVALUE=2)
/MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
/CRITERIA ALPHA=0.05 CILEVEL=95.
```

Запустив эту команду, мы окажемся в окне вывода для анализа полученных таблиц (рис. 4.14). Здесь мы оцениваем значимость различий между средним значением в выборке и заданным значением среднего в генеральной совокупности на основе полученного p-значения (если p-значение меньше выбранного уровня значимости, отклоняем нулевую гипотезу и считаем различия статистически значимыми) — нулевая гипотеза отклоняется.

Итоги по проверке гипотезы

	Нулевая гипотеза	Критерий	Значимость	Решение
1	Медиана Сколько еще детей Вы реально планируете иметь, исходя из обстоятельств? равна 2.	Одновыборочный критерий знаковых рангов Вилкоксона	,000	Нулевая гипотеза отклоняется.

Выводятся асимптотические значимости. Уровень значимости равен ,050.

Рис. 4.14. Оценка статистической значимости критерия Вилкоксона

Рисунок ниже содержит сравнение значений проверяемой и наблюдаемой медиан (рис. 4.15).

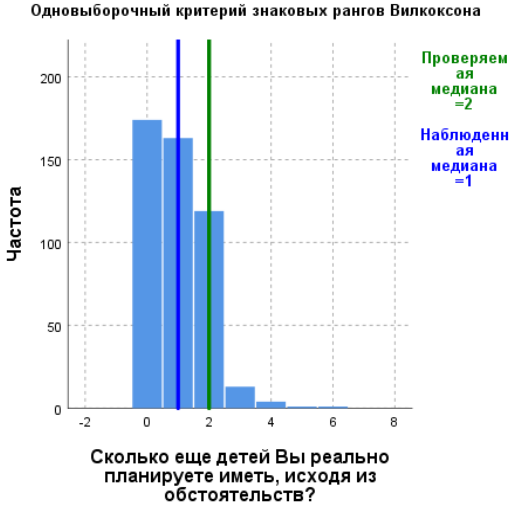


Рис. 4.15. Визуализация средних значений
с использованием одновыборочного критерия знаковых рангов Вилкоксона

Таким образом, мы можем сделать вывод о том, что количество детей, которое респондентки хотели бы иметь исходя из обстоятельств, статистически достоверно меньше двух.

§ 4.4. Контрольные вопросы

1. Дайте определение t-критерию для независимых выборок. Каков алгоритм его использования в программе IBM SPSS Statistics?
2. Дайте определение t-критерию для парных выборок. Каков алгоритм его использования в программе IBM SPSS Statistics?
3. Дайте определение одновыборочному t-критерию. Каков алгоритм его использования в программе IBM SPSS Statistics?
4. Назовите непараметрические критерии для каждого метода t-критерия Стьюдента.

§ 4.5. Практические задания

1. Определите, как связаны пол и доверие общественно-политическим институтам. Используйте массив «Социально-политические настроения в Калининградской области», n = 915, 2022 г. (https://kantiana.ru/science/baza-dannykh/).
2. Определите, как различается количество детей, которых хотели бы иметь респондентки в идеале и исходя из обстоятельств в группах с различным уровнем дохода. Используйте массив «Репродуктивные установки жительниц Калининградской области», n = 830, 2017 г.
3. Оцените, отличается ли средний доход жителей Калининградской области от среднероссийских показателей? Используйте массив «Социально-политические настроения в Калининградской области», n = 915, 2022 г. (https://kantiana.ru/science/baza-dannykh/).

§ 4.6. Рекомендуемая литература

1. *Белинский А. В.* Влияние наличия компонента угрозы как фактора деавтоматизации письма // Общество: социология, психология, педагогика. 2023. № 4. С. 105—112.
2. *Крыштановский А. О.* Анализ социологических данных с помощью пакета SPSS. М. : ГУ ВШЭ, 2006.
3. *Крюкова Т. Л., Екимчик О. А.* Фаббинг как угроза благополучию близких отношений // Консультативная психология и психотерапия. 2019. Т. 27, № 3. С. 61—76.
4. *Кубекова А. С.* Особенности защитных механизмов у иностранных студентов с разными уровнями адаптации // Общество: социология, психология, педагогика. 2020. № 3 (71). С. 89—92.

5. *Наследов А.Д.* SPSS 19: профессиональный статистический анализ данных. СПб. : Питер, 2011.

6. *Типы* непараметрических критериев для двух независимых выборок. URL: https://www.ibm.com/docs/ru/spss-statistics/25.0.0?topic=SSLVMB_25.0.0/spss/base/two_independent_samples_test_types.htm (дата обращения: 05.10.2023).

7. *Чурилова Е., Захаров С.* Репродуктивные установки населения России: есть ли повод для оптимизма? // Вестник общественного мнения. 2019. №2 (129). С. 69—89.

Глава 5
СРАВНЕНИЕ СРЕДНИХ ЗНАЧЕНИЙ: ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

§ 5.1. Цель и условия использования метода

Однофакторный дисперсионный анализ (ANOVA[19]) — это метод статистического анализа, который используется для определения статистически значимых различий между средними значениями трех или более групп данных.

Этот метод называется дисперсионным анализом, так как он исследует разброс выборочных значений и, в частности, такую меру разброса, как дисперсия. На основе этих двух оценок разброса можно будет сделать выводы о средних значениях в генеральных совокупностях. Если выборочные средние значения разбросаны больше, чем можно ожидать, исходя из разброса наблюдений внутри групп, можно сделать вывод, что в генеральных совокупностях не все средние значения равны между собой [2, с. 99].

Он используется для проверки гипотезы о том, что средние значения в трех или более группах статистически не отличаются друг от друга. ANOVA может быть полезен в случаях, когда нужно определить, есть ли статистически значимые различия между группами данных, например, при исследовании социальных классов. ANOVA может использоваться, чтобы сравнить средние значения доходов или образования в разных социальных классах. Это позволяет исследователям понять, как социальный класс влияет на статистически значимые различия между группами, если они есть.

В качестве примера социологического анализа данных, выполненных с использованием однофакторного дисперсионного анализа, отметим исследование региональной идентичности различных типов респондентов: социал-демократов, социал-этатистов, либерал-демократов и либерал-этатистов [5]. В частности, показано, что социал-демократы и социал-этатисты рассматривают себя преимущественно как жители того города или села, где сейчас проживают. Либерал-демократы и либерал-этатисты идентифицируют себя в большей степени с гражданином мира, жителем планеты Земля (рис. 5.1).

Для проведения ANOVA существует несколько условий, которые должны быть выполнены:

1. Нормальность распределения. Это означает, что основные статистические меры (среднее значение, медиана, стандартное отклонение) в каждой группе должны быть близки к нормальному распределению.

[19] Сокращение с английского языка «analysis of variance» — дисперсионный анализ.

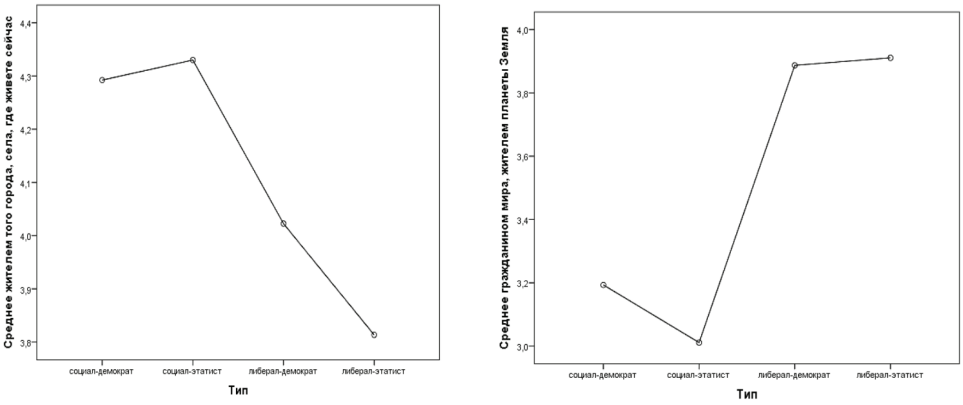


Рис. 5.1. Связь между идеологическим типом и региональной идентичностью

Источник: [5].

2. Гомогенность дисперсии. Дисперсия данных, или стандартное отклонение значений, в каждой группе должно быть одинаковым или близким к одинаковому. Предположение о равенстве дисперсий можно проверить при помощи критерия Ливиня, доступного в процедурах «Разведочный анализ» и «Однофакторный дисперсионный анализ».

3. Независимость наблюдений. Каждое наблюдение должно быть независимым от другого. Предположение о независимости означает отсутствие связей между наблюдениями в разных группах и между наблюдениями внутри каждой из них. Например, одна и та же семья относится только к одному из типов населенного пункта. Наблюдения, относящиеся к одной и той же семье, появляются только в одной из групп и не дублируются внутри нее [2, с. 98].

4. Масштаб переменной. Измерения в каждой группе должны быть в одном и том же масштабе. Например, если одна группа имеет значения от 1 до 10, а другая — от 100 до 1000, то данные не находятся в одном масштабе.

5. Отсутствие выбросов. В каждой группе не должно быть выбросов, так как они могут сильно исказить средние значения и вносить дополнительную переменность. При обнаружении таких значений анализ нужно выполнить еще раз, исключив необычные случаи, и таким образом проверить, совпадают ли выводы со сделанными ранее [2, 99].

Если при тестировании гипотезы о равенстве средних значений в группах какие-либо из предположений не выполняются, то необходимо использовать альтернативные методы, такие как непараметрические тесты (будут рассмотрены в параграфе 5.3).

§ 5.2. Алгоритм выполнения однофакторного дисперсионного анализа

Применение однофакторного дисперсионного анализа предполагает выполнение следующих шагов:

1) проверка на соответствие всем допущениям, указанным выше в параграфе 5.1;
2) формулирование исследовательского вопроса;
3) определение зависимых и независимых переменных;
4) указание нулевой и альтернативной гипотез;
5) проверка нулевой гипотезы;
6) определение силы влияния фактора;
7) апостериорный анализ;
8) описание полученных значений.

Ниже представлен пример ANOVA при условии, что данные в выборках соответствуют всем пяти допущениям, описанным выше. Разберем критерий однородности дисперсий Ливиня, поскольку мы с ним еще не работали. Команда «Исследовать» (в программе IBM SPSS Statistcs: анализ — описательные статистики — разведочный анализ) позволяет вывести его статистику (рис. 5.2).

Критерий однородности дисперсий

		Статистика Ливиня	ст.св.1	ст.св.2	Значимость
Сколько еще детей Вы реально планируете иметь, исходя из обстоятельств?	Основано на среднем	1,444	4	392	,219
	Основано на медиане	,909	4	392	,459
	Основано на медиане и с настроенными ст. св.	,909	4	287,970	,459
	Основано на усеченном среднем	1,347	4	392	,252

Рис. 5.2. Критерий однородности дисперсий

Таблица критериев однородности дисперсий Ливиня информирует исследователя о соблюдении требований равенства дисперсий при проведении однофакторного дисперсионного анализа [2, с. 103]. Нулевая гипотеза в отношении этого критерия формулируется следующим образом: дисперсии переменной «количество детей» равны. Это одно из требований проведения дисперсионного анализа. Из таблицы мы видим, что нулевая гипотеза о равенстве дисперсий Ливиня в анализируемых группах принимается ($p > 0{,}05$).

Для примера возьмем уже известный нам массив, отражающий репродуктивные установки женщин в возрасте 18—44 лет. Допустим, нас интересует, существуют ли статистически достоверные различия в количестве детей, которых хотели бы иметь респондентки исходя из обстоятельств, при различных формах подработки. Соответственно, независимая переменная (фактор) — возрастные группы, зависимая — количество детей, которых женщины хоте-

ли бы иметь. Нулевая гипотеза — переменные не связаны друг с другом, альтернативная — переменные связаны, форма подработки оказывает воздействие на количество детей.

В программе IBM SPSS Statistics открываем окно ANOVA (анализ — сравнение средних — однофакторный дисперсионный анализ). В соответствующих полях размещаем переменные (рис. 5.3). В опции «Апостериорные» выбираем критерий Шеффе, в опции «Параметры» отмечаем «Описательные» и «График средних». Кликаем «Продолжить» и «ОК».

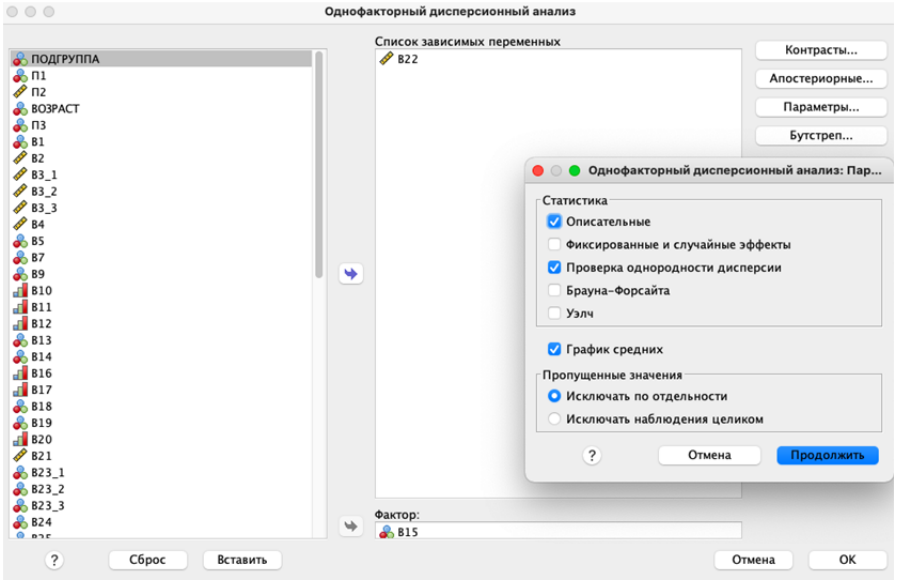


Рис. 5.3. Интерфейс окна «Однофакторный дисперсионный анализ»

Следующим шагом мы проверяем нулевую гипотезу. Для этого нам достаточно оценить уровень значимости для F-статистики (критерия Фишера) в таблице ANOVA (рис. 5.4). Поскольку его значения меньше 0,05, мы отклоняем нулевую гипотезу и утверждаем, что наличие подработок оказывает влияние на количество детей, которое респондентки хотели бы иметь.

ANOVA

Сколько еще детей Вы реально планируете иметь, исходя из обстоятельств?

	Сумма квадратов	ст.св.	Средний квадрат	F	Значимость
Между группами	24,383	4	6,096	7,180	,000
Внутри групп	332,801	392	,849		
Всего	357,184	396			

Рис. 5.4. Статистика F-критерия в ANOVA

Далее необходимо определить силу влияния фактора на зависимую переменную. Она определяется посредством коэффициента «эта в квадрате» (η^2)[20], которая рассчитывается как отношение суммы квадратов между группами к сумме квадратов всего. Значение эта-квадрата находится в диапазоне от 0 до 1, где значения, близкие к 1, указывают на более высокую долю дисперсии, которая может быть объяснена данной переменной в модели. Следующие эмпирические правила используются для интерпретации значений эта-квадрата:

0,01 — малый эффекта;

0,06 — средний эффекта;

0,14 или выше — большой эффект [6].

В нашем случае $\eta^2 = 0{,}067$ или 6,7%, то есть средний эффект. Это означает, что приблизительно 7% дисперсии зависимой переменной объясняется влиянием фактора. Другими словами, формы подработки оказывают незначительное влияние на количество планируемых детей.

Установив наличие влияния фактора и его силу, мы все еще не можем определить, какая именно форма подработки оказывает наибольшее и наименьшее влияние на количество детей, которых респондентки хотели бы иметь исходя из обстоятельств. Именно с этой целью мы проводим апостериорный анализ множественных сравнений. Суть метода состоит в определении различий посредством сравнения средних значений количественной переменной во всех возможных парах групп, определяемых градациями переменной — фактора [3, с. 104].

В программе IBM SPSS Statistics есть различные критерии для выполнения этого анализа. Вслед за известным специалистом в области статистического анализа социологических данных мы повторим, что «у нас нет серьезных аргументов для рекомендации выбора того или иного метода» и, как показывает практика, «все предлагаемые методы дают весьма близкие результаты» [3, с. 106]. Пожалуй, здесь важным фактором при выборе критерия является равенство дисперсий: если оно предполагается, прибегаем к одним критериям, если нет — к другим. Из числа тех критериев, где равенство дисперсий предполагается, наиболее «консервативным», то есть наименее подверженным случайным ошибкам, является критерий Шеффе [4, с. 188]. Он «производит одновременные сравнения совместных пар для всех возможных комбинаций пар средних. Использует выборочное F-распределение. Может применяться для проверки всех возможных линейных комбинаций групповых средних, а не только для парных сравнений» [1].

Чтобы выбрать соответствующий критерий, необходимо открыть окно апостериорных множественных сравнений (опция «Апостериорные» в окне ANOVA) (рис. 5.5).

[20] Эта-квадрат — это мера величины эффекта, которая измеряет долю дисперсии, связанную с каждым основным эффектом и эффектом взаимодействия в модели ANOVA.

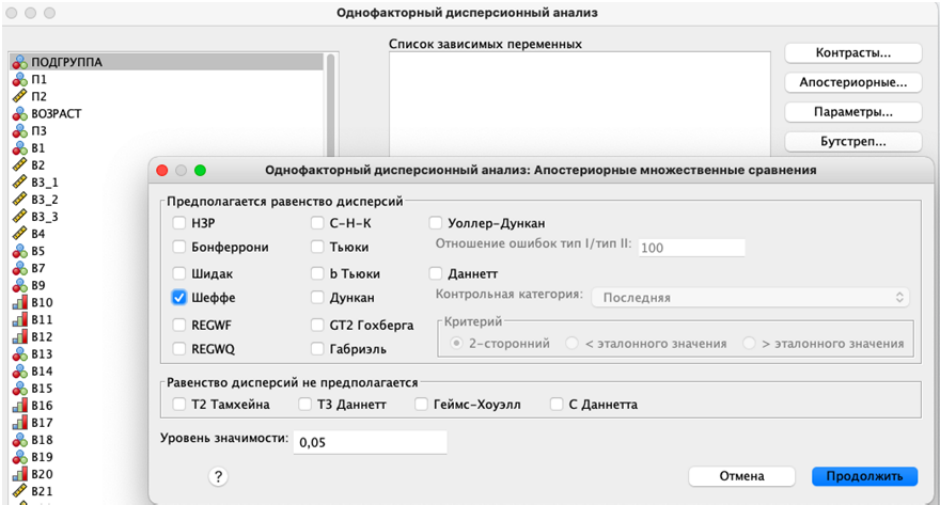


Рис. 5.5. Интерфейс окна «Апостериорные множественные сравнения»

В результате у нас получится таблица и рисунок, которые в разной форме сообщают одну и ту же информацию: какие существуют различия в средних значениях зависимой переменной между независимыми выборками (рис. 5.6). Рисунок наглядно демонстрирует, что наибольшее число детей приходится на тех респонденток, которые нигде, кроме основной работы, больше не трудятся. В то время как дополнительная занятость только уменьшает количество запланированных детей исходя из обстоятельств.

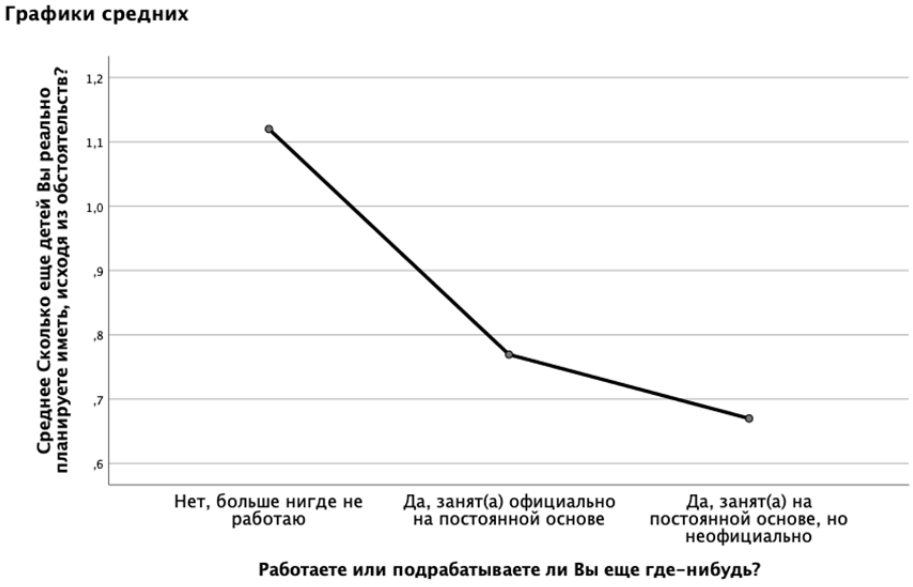


Рис. 5.6. График средних значений независимой переменной в ANOVA

Таблица «Множественные сравнения» позволяет убедиться в статистической достоверности обнаруженных различий между формами подработки (рис. 5.7). Читая данную таблицу, следует обращать внимание на: 1) те пары, где значимость менее 0,05; 2) положительную среднюю разность (I – J). Таким образом, мы фиксируем, что все пары с первой градацией («Нет, больше нигде не работаю») имеют не только удовлетворительный уровень значимости, но и положительную среднюю разность, что свидетельствует о том, что средние значения в этой группе статистически выше, чем в двух других.

Множественные сравнения

Зависимая переменная: Сколько еще детей Вы реально планируете иметь, исходя из обстоятельств?

Шеффе

(I) Работаете или подрабатываете ли Вы еще где-нибудь?	(J) Работаете или подрабатываете ли Вы еще где-нибудь?	Средняя разность (I-J)	Стандартная ошибка	Значимость	95% доверительный интервал	
					Нижняя граница	Верхняя граница
Нет, больше нигде не работаю	Да, занят(а) официально на постоянной основе	,351*	,126	,021	,04	,66
	Да, занят(а) на постоянной основе, но неофициально	,450*	,115	,001	,17	,73
Да, занят(а) официально на постоянной основе	Нет, больше нигде не работаю	-,351*	,126	,021	-,66	-,04
	Да, занят(а) на постоянной основе, но неофициально	,099	,139	,774	-,24	,44
Да, занят(а) на постоянной основе, но неофициально	Нет, больше нигде не работаю	-,450*	,115	,001	-,73	-,17
	Да, занят(а) официально на постоянной основе	-,099	,139	,774	-,44	,24

*. Средняя разность значима на уровне 0.05.

Рис. 5.7. Результат апостериорных множественных сравнений

Наконец, мы можем обратить внимание на таблицу с общей статистикой и констатировать, что респондентки, которые больше нигде не работают, кроме места основной занятости, в среднем хотели бы иметь более одного ребенка, в то время как остальные не готовы иметь больше детей исходя из обстоятельств (рис. 5.8).

Описательные статистики

Сколько еще детей Вы реально планируете иметь, исходя из обстоятельств?

	N	Среднее	Стандартная отклонение	Стандартная ошибка	95% доверительный интервал для среднего значения		Минимум	Максимум
					Нижняя граница	Верхняя граница		
Нет, больше нигде не работаю	175	1,12	,873	,066	,99	1,25	0	4
Да, занят(а) официально на постоянной основе	78	,77	,911	,103	,56	,97	0	4
Да, занят(а) на постоянной основе, но неофициально	103	,67	1,014	,100	,47	,87	0	6
Всего	356	,91	,944	,050	,81	1,01	0	6

Рис. 5.8. Статистика средних значений независимой переменной

Общий вывод по результатам однофакторного дисперсионного анализа, пример которого представлен выше, может выглядеть так: род занятости оказывает статистически достоверное влияние на количество детей, которое респондентки хотели бы иметь исходя из обстоятельств ($p < 0{,}001$; $F = 7{,}180$). Однако это влияние невелико ($\eta^2 = 0{,}067$), что говорит о существовании иных факторов, которые также могут определять отношение к количеству детей. Тем не менее в рамках поставленной задачи установлено, что женщины, имеющие дополнительные формы занятости, не готовы к рождению даже одного ребенка. Возможно, это связано с более высокой трудовой нагрузкой и отсутствием времени, которое потребовалось бы уделять детям.

§ 5.3. Непараметрические критерии

Для проведения классического (параметрического) однофакторного дисперсионного анализа (ANOVA) важно, чтобы данные были распределены нормально, а дисперсии выборок были примерно одинаковыми (гомогенными), так как этот анализ использует так называемую F-статистику, которая зависит от отношения межгрупповых и внутригрупповых дисперсий. Если дисперсии выборок значительно отличаются друг от друга (гетерогенность), то это может привести к неправильной интерпретации результатов теста.

Однако существует ряд критериев, которые могут быть использованы, когда одно или оба условия нарушаются. Если данные нормально распределены, но дисперсии в группах различны, то обычным выбором будет использование критерия Геймса — Хоуэлла. Он используется для сравнения всех пар групп между собой, даже если имеются значительные различия в дисперсии. Эти множественные сравнения выполняются для определения, какие пары групп значимо отличаются друг от друга.

В то же время если данные не являются нормально распределенными или имеют выбросы, то критерий Краскала — Уоллиса может быть предпочтительнее, поскольку он использует ранги данных, а не их исходные значения. Критерий Краскала — Уоллиса также может быть предпочтительнее в случае небольших выборок или, если данные содержат категориальные (порядковые) переменные, которые нельзя преобразовать в числовые значения для использования в ANOVA. Для этого непараметрического критерия равенство дисперсий не требуется, поскольку он использует ранговые данные вместо исходных значений переменных и не требует нормальности распределения. Этот критерий решает задачу гомогенности средних рангов в различных группах, но не учитывает величину различий.

Использование и интерпретация критерия Геймса — Хоуэлла полностью идентичны критерию Шеффе, описанному в параграфе 5.2. Использование критерия Краскала — Уоллиса (Kruskal-Wallis test) несколько отличается. Это непараметрический критерий однофакторного дисперсионного анализа для проверки гипотезы о равенстве медиан в двух или более группах.

Предположим, у нас есть несколько выборок, и мы хотим проверить, равны ли их медианы. Для этого мы объединяем их в одну большую выборку и заменяем значения на их ранги — порядковые номера в упорядоченных значениях. Затем мы сортируем данные по возрастанию и рассчитываем сумму рангов для каждой выборки. Критерий Краскала — Уоллиса использует эти суммы рангов, чтобы получить статистику критерия.

Гипотезы для критерия Краскала — Уоллиса (H-статистики) следующие:

— H0: медианы всех k групп равны;

— H1: не все медианы равны.

Если полученное значение H-статистики меньше критического значения, то мы не можем отклонить нулевую гипотезу и делаем вывод, что медианы в выборках равны. Если же H-статистика больше критического значения, то мы отклоняем нулевую гипотезу и сделаем вывод, что медианы в выборках различаются.

Важно отметить, что критерий Краскала — Уоллиса тестирует только гомогенность медиан, но не показывает, какие выборки завышают или занижают общую медиану.

Рассмотрим на примере. Диалоговое окно критерия Краскала — Уоллиса находится в непараметрических критериях: анализ — непараметрические критерии — устаревшие диалоговые окна — для К независимых выборок. Допустим, нас интересует, различается ли количество детей, которые респондентки хотели бы иметь в идеале, среди различных возрастных групп. Другими словами, мы проверяем эффект возрастных групп или влияние возраста как фактора. Проверка нулевой гипотезы об отсутствии связи между переменными также проверяется через оценку асимптотической значимости H-критерия Краскала — Уоллиса. В нашем случае она менее 0,05, следовательно, мы принимаем альтернативную гипотезу и констатируем факт взаимосвязи (рис. 5.9, с. 67). Более высокий средний ранг в группе 18—24 лет свидетельствует о больших значениях количественной переменной именно среди данных респондентов. Конкретные значения по каждой группе (описательную статистику) можно получить, выполнив процедуру анализа средних, описанную в пункте выше (анализ — сравнение средних — средние). Там же можно указать, что требуется вывести таблицу дисперсионного анализа и эта-коэффициент в квадрате.

§ 5.4. Контрольные вопросы

1. В чем суть однофакторного дисперсионного анализа? Каковы цели и условия его использования?

2. Каков алгоритм использования однофакторного дисперсионного анализа в IBM SPSS Statistics?

3. Какие непараметрические критерии используются в однофакторном дисперсионном анализе? Что общего и в чем разница между критериями Геймса — Хоуэлла и Краскала — Уоллиса?

Критерий Краскала-Уоллиса

Ранги

	ВОЗРАСТ	N	Средний ранг
Сколько еще детей Вам хотелось бы иметь «в идеале»?	18-24	266	338,59
	25-29	102	285,63
	30-34	99	293,85
	35-39	74	280,39
	40-44	71	261,11
	Всего	612	

Статистические критерии[a,b]

	Сколько еще детей Вам хотелось бы иметь «в идеале»?
H Крускала-Уоллеса	19,575
ст.св.	4
Асимп. знач.	,001

a. Критерий Краскела-Уоллиса

b. Группирующая переменная: ВОЗРАСТ

Рис. 5.9. Основные статистические показатели критерия Краскала — Уоллиса

§ 5.5. Практические задания

1. Определите, влияет ли возраст респондентов на то, к каким историческим эпохам они проявляют наибольший интерес. Используйте массив «Историческая память населения Калининградской области», n = 997, 2022 г.

2. Определите, как различается степень региональной идентичности среди жителей различных типов поселений (поселок, город, областная столица). Используйте массив «Социально-политические настроения молодежи Калининградской области», n = 987, 2021 г.

§ 5.6. Рекомендуемая литература

1. *Апостериорные* критерии для однофакторного дисперсионного анализа. URL: https://www.ibm.com/docs/ru/spss-statistics/25.0.0?topic=SSLVMB_25.0.0/spss/base/idh_onew_post.htm (дата обращения: 05.10.2023).

2. *Воронин Г. Л.* Статистический анализ данных в IBM SPSS Statistics V27.0.1.0. Н. Новгород : ННГУ им. Н. И. Лобачевского, 2022.

3. *Крыштановский А. О.* Анализ социологических данных с помощью пакета SPSS. М. : ГУ ВШЭ, 2006.

4. *Наследов А. Д.* SPSS 19: профессиональный статистический анализ данных. СПб. : Питер, 2011.

5. *Щекотуров А. В., Кришталь М. И.* Динамика территориальной идентичности и восприятия статуса региона жителями Калининградской области в 2016—2020 гг. // Вестник Московского университета. Сер. 18: Социология и политология. 2021. Т. 27, № 3. С. 43—62.

6. *Что* такое Эта в квадрате? URL: https://www.codecamp.ru/blog/eta-squared (дата обращения: 05.10.2023).

Глава 6
КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ

Корреляционный и регрессионный анализ — это два тесно связанных метода статистического анализа, используемых для изучения отношений между переменными.

Сходства между корреляционным и регрессионным анализом заключаются в следующем:

— основаны на предположении о линейной взаимосвязи переменных;

— чувствительны к экстремальным значениям (выбросам);

— используются для изучения отношений между количественными и порядковыми переменными;

— используются для проверки связи между переменными и определением силы этой связи.

Однако, также есть и различия между корреляционным и регрессионным анализом. Корреляционный анализ обычно используется для изучения силы отношения между двумя переменными, в то время как регрессионный позволяет выявить более точную природу этого отношения. Корреляции не предполагают причинно-следственных связей[21], в то время как регрессия способно предсказывать значения зависимой переменной на основе значений предиктора (предикторов).

В социологии каждый из этих методов довольно распространен. Например, выделим исследование доверия лояльной и оппозиционной молодежи общественно-политическим институтам и учреждениям [6]. В частности, анализ корреляций показал, что молодежь из разных групп с достоверно разной степенью доверяет тем или иным институтам (рис. 6.1).

В этой же публикации выполнен регрессионный анализ, демонстрирующий, как выраженность определенных ценностей в наибольшей и наименьшей степени определяет доверие государственным и оппозиционным лидерам (рис. 6.2).

---

[21] Довольно показательный пример приводится в книге Чарльза Уиллана «Голая статистика» [5]. В частности, автор предполагает существование положительной корреляции между результатами экзаменационного тестирования школьников и количеством телевизоров у них дома. Но, как справедливо замечает сам же Ч. Уиллан, «это не значит, что родители могут существенно повысить результаты тестов своих детей путем покупки еще пяти телевизоров… Как количество телевизоров, так и экзаменационные оценки обусловлены некой третьей переменной, коей является уровень образования родителей». Высокообразованные родители, как правило, имеют более высокий доход (могут позволить себе не один телевизор) и более высокие образовательные притязания (например, оплатить дополнительную подготовку к экзамену).

Институты	Лояльная молодежь	Оппозиционная молодежь
Церковь	,476**	,076*
Вооруженные силы	,525**	–,020
СМИ (печатные газеты, радио, телевидение и др.)	,349**	,057
Электронные СМИ (интернет-ресурсы)	,015	,249**
Частные ВУЗы	,116**	,207**
Государственные ВУЗы	,344**	,086**
Полиция	,534**	,027
Банки	,275**	,127**
Правозащитные организации	,207**	,192**
Благотворительные фонды	,104**	,284**
Международные организации	,078*	,264**
Частные лечебные учреждения	,128**	,134**
Государственные лечебные учреждения	,296**	,089**
Суды	,445**	,091**

* Корреляция значима на уровне 0,05 (двусторонняя).
** Корреляция значима на уровне 0,01 (двусторонняя).

Рис. 6.1. Пример корреляции

Источник: [6].

Предикторы	Лояльная молодежь			Оппозиционная молодежь		
	Коэффициент регрессии	Стандартная ошибка	Бета	Коэффициент регрессии	Стандартная ошибка	Бета
Константа	2,565	0,419	–	1,447	0,476	–
Свободное выражение личного мнения	–0,368	0,076	–0,172	0,264	0,090	0,113
Занятие высокого социального статуса	0,195	0,048	0,145	–	–	–
Культурное и интеллектуальное саморазвитие	–	–	–	–0,163	0,081	–0,075
Социально-экономическое равенство	–	–	–	0,191	0,057	0,124
R^2	0,091			0,052		

Рис. 6.2. Пример множественной регрессии

Источник: [6].

Схожее исследование доверия различным социальным институтам выполнено аспирантами из Томского политехнического университета [4]. В своей работе они опирались не только на корреляционный, но и на факторный анализ, что подчеркивает комплементарный характер простой линейной связи.

§ 6.1. Анализ корреляций

Корреляционный анализ — это метод статистического анализа, который используется для изучения отношений между двумя или более переменными и позволяет выявить наличие, направление и силу связи между ними.

Оценка корреляционной связи проводится на основе коэффициента корреляции, который измеряет степень линейной связи между переменными и может принимать значения от – 1 до 1. Значение 1 означает положительную связь между переменными (то есть, если одна переменная увеличивается, то и другая тоже увеличивается), – 1 означает отрицательную связь (если одна переменная увеличивается, то другая уменьшается) и 0 означает отсутствие связи между переменными.

Выделяют несколько коэффициентов корреляции: r-Пирсона, r-Спирмена и Tay-b Кендалла. Коэффициент корреляции Пирсона является параметрическим и используется только для анализа линейной взаимосвязи количественных переменных.

Ниже представлен рисунок, на котором визуализирована сильная положительная (B) и сильная отрицательная (A) корреляция, а также две ложные корреляции из-за неучтенных выбросов (C и D) (рис. 6.3).

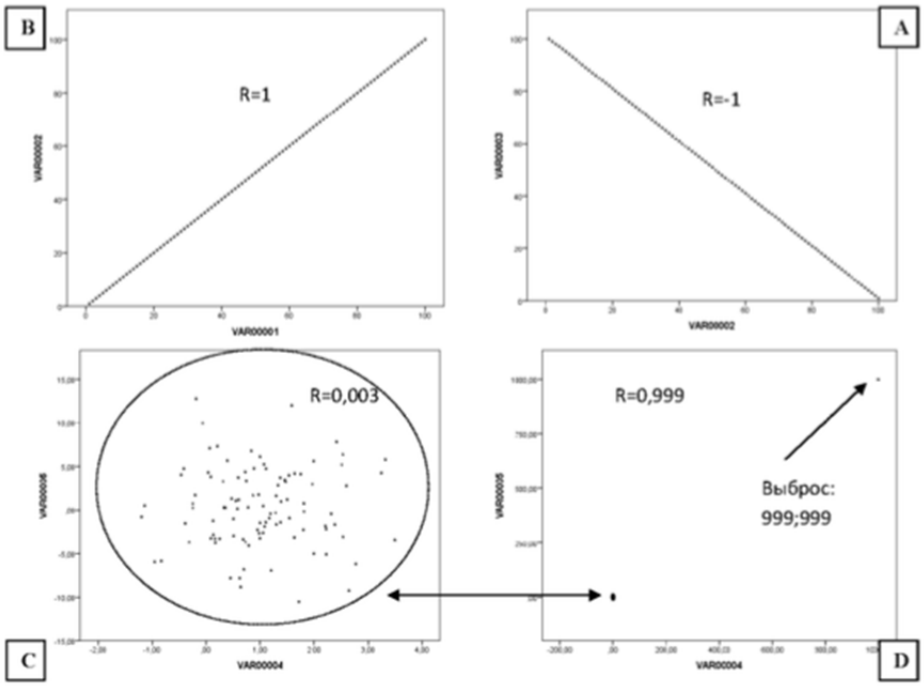


Рис. 6.3. Диаграммы рассеяния с различными показателями коэффициента корреляции Пирсона и влияния выбросов (ложная корреляция)

Источник: [2, с. 111].

Если направление корреляции определяется по знаку перед значением r, то силу связи мы оцениваем по величине самого значения. В социологии выделяют следующие уровни силы корреляции переменных [1 с. 257]:

до 0,2 — очень слабая корреляция;

до 0,5 — слабая корреляция;

до 0,7 — средняя корреляция;

до 0,9 — высокая корреляция;

свыше 0,9 — очень высокая корреляция.

Таким образом, проведение корреляционного анализа требует последовательного выполнения следующих шагов:

1) оценка переменных на наличие экстремальных значений;

2) оценка уровня значимости;

3) определение направления;

4) определение силы связи.

Рассмотрим анализ корреляции Пирсона на примере взаимосвязи между возрастом и количеством детей, которые респондентки хотели бы иметь в идеале и исходя из обстоятельств. Вывод окна корреляции осуществляется по следующей команде: анализ — корреляции — парные. После указания исследуемых переменных кликаем «ОК». Все нужные опции уже установлены по умолчанию[22].

В результате у нас появляется таблица корреляции (рис. 6.4).

Корреляции

		Количество полных лет	Сколько еще детей Вам хотелось бы иметь «в идеале»?	Сколько еще детей Вы реально планируете иметь, исходя из обстоятельств?
Количество полных лет	Корреляция Пирсона	1	-,192**	-,518**
	Знач. (двухсторонняя)		,000	,000
	N	826	611	472
Сколько еще детей Вам хотелось бы иметь «в идеале»?	Корреляция Пирсона	-,192**	1	,367**
	Знач. (двухсторонняя)	,000		,000
	N	611	616	463
Сколько еще детей Вы реально планируете иметь, исходя из обстоятельств?	Корреляция Пирсона	-,518**	,367**	1
	Знач. (двухсторонняя)	,000	,000	
	N	472	463	475

**. Корреляция значима на уровне 0,01 (двухсторонняя).

Рис. 6.4. Пример корреляции трех переменных

[22] При использовании любого коэффициента корреляции рекомендуется оставлять оценку двухсторонней значимости, заданную по умолчанию. Данный выбор обусловлен тем, что мы не имеем заранее определенных представлений о возможной направленности взаимосвязи между двумя переменными. В этом случае исследователь рассматривает обе возможности: положительное и отрицательное направление взаимосвязи.

Здесь мы наблюдаем сразу несколько статистически значимых корреляций ($p < 0,05$): между возрастом и количеством детей; между количеством детей, которые респондентки хотели бы иметь в идеале и исходя из обстоятельств. Однако в первых двух случаях корреляции отрицательные ($r = -0,192$ и $r = -0,518$), а в последнем связь положительная ($r = 0,367$). Интерпретировать это можно так: с возрастом количество детей, которые респондентки хотели бы иметь, снижается. Причем более сильная корреляция здесь зафиксирована с переменной, охватывающей реальную оценку ситуации. При этом, чем больше детей жительницы Калининградской области хотели бы иметь в идеале, тем больше они рассчитывают завести исходя из обстоятельств ($r = 0,367$, $p < 0,001$).

Для порядковых и ранговых переменных рассчитываются непараметрические коэффициенты корреляции r-Спирмена и Тау Кенделла. Выбор между ними зависит от конкретных условий и типа данных. Оба коэффициента корреляции являются непараметрическими мерами и используются для оценки силы связи между двумя переменными, когда данные не соответствуют условиям нормального распределения.

Коэффициент Спирмена предполагает, что данные имеют ранговую (порядковую) шкалу измерения. Коэффициент Тау Кенделла, в свою очередь, нечувствителен к выбросам и может хорошо работать с небольшими выборками.

Другими словами, если данные находятся в ранговой шкале и выбросов нет, то следует использовать коэффициент Спирмена. Однако, если данные не являются ранговыми и выборка небольшая, то следует использовать коэффициент Тау Кенделла.

Как и коэффициент Пирсона, коэффициенты Спирмена и Тау Кенделла принимают значения от -1 до 1, оцениваются и анализируются аналогичным образом.

§ 6.2. Простая линейная регрессия

Анализ корреляции отражает меру взаимосвязи между двумя переменными, тогда как регрессионный анализ позволяет определить тип этой связи и предсказать значения одной переменной (зависимой) на основе значений другой переменной (независимой) [1, с. 269].

Простой регрессионный анализ — это метод статистического анализа данных, который используется для изучения отношения между двумя переменными: зависимой и независимой. Он позволяет определить, насколько изменения в независимой влияют на изменения в зависимой переменной.

В простом регрессионном анализе учитывается только одна независимая и одна зависимая переменная. Обычно они представляют собой числовые данные и представлены в виде точек на графике. Регрессионный анализ строит линию или кривую тренда, которая проходит через эти точки и отображает отношение между переменными. Эту линию можно использовать для предсказания значений зависимой переменной на основе значений независимой.

Результатом регрессионного анализа является уравнение регрессии, которое представляет собой математическую формулу отношения между зависимой переменной и одной или несколькими независимыми.

Уравнение регрессии получается путем нахождения оптимальных коэффициентов, которые минимизируют разницу между фактическими значениями зависимой переменной и предсказанными значениями, полученными из уравнения регрессии. Коэффициенты определяют, насколько каждая независимая переменная влияет на зависимую и каким образом они взаимодействуют друг с другом.

Уравнение простой линейной регрессии представляет собой линейную функцию, которая описывает связь между зависимой и независимой переменными:

$$Y = a + bX + e,$$

где Y — зависимая переменная (или переменная отклика); X — независимая переменная (или переменная предиктора); a — константа или коэффициент сдвига (интерсепт), который представляет собой уровень изменений Y, когда X равен 0; b — регрессионный коэффициент или коэффициент наклона, который показывает, насколько изменится Y при изменении X на единицу; e — ошибка, случайная переменная, которая отражает любые другие факторы, которые могут влиять на Y.

Таким образом, задача социолога сводится к построению уравнения регрессии на основе соответствующих коэффициентов и его интерпретации. Стоить, однако, заметить, что, хотя соотношение между зависимой и независимой переменными в регрессионном анализе считается линейным, связь между ними может иметь иную форму. По этой причине рекомендуется дополнительно выполнять проверку на криволинейность.

Анализ регрессии предполагает выполнение следующих шагов:

1) формулирование исследовательского вопроса;
2) определение зависимой и независимой переменных;
3) формулирование нулевой и альтернативной гипотезы;
4) проверка нулевой гипотезы;
5) построение уравнения регрессии;
6) интерпретация коэффициентов регрессии.

Разберем все эти шаги на примере из массива о репродуктивном поведении жительниц Калининградской области.

Шаг 1. Допустим, нас интересует, существует ли связь между количеством детей, которые уже есть у респондентов, и количеством детей, которые они хотели бы иметь, с учетом текущих обстоятельств. И если эта связь существует, то как меняется количество планируемых детей в зависимости от количества имеющихся детей?

Шаг 2. Соответственно, независимая переменная — количество имеющихся детей, зависимая — количество детей, которых респондентки хотели бы иметь.

Шаг 3. Нулевая гипотеза — переменные не связаны ($p > 0{,}05$), альтернативная — количество планируемых детей значимо зависит от количества имеющихся.

Шаг 4. Переходим в диалоговое окно регрессии так, чтобы можно было проверить уравнение на криволинейность (анализ — регрессии — подгонка кривых). Указываем зависимую и независимую переменные в соответствующие поля. Выбираем линейную и квадратичную модели[23], кликаем «ОК» (рис. 6.5).

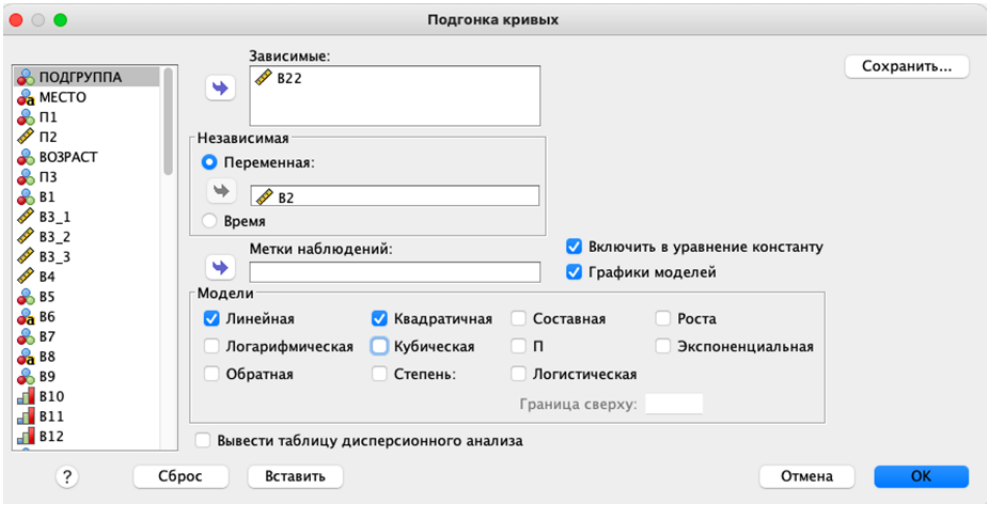


Рис. 6.5. Интерфейс окна «Подгонка кривых»

В результате у нас будет несколько таблиц, главная из которых — сводка для двух моделей (рис. 6.6). На данном шаге мы обращаем внимание на значимость и констатируем, что как для линейной, так и для квадратичной регрессии адекватна альтернативная гипотеза о наличии связей между переменными.

Сводка для модели и оценки параметров

Зависимая переменная: Сколько еще детей Вы реально планируете иметь, исходя из обстоятельств?

	Сводка для модели					Оценки параметров		
Уравнение	R-квадрат	F	ст.св.1	ст.св.2	Значимость	Константа	b1	b2
Линейная	,302	203,849	1	472	,000	1,466	-,558	
Квадратичная	,315	108,467	2	471	,000	1,528	-,873	,136

Независимая переменная – это Укажите, пожалуйста, общее количество детей, которые у Вас есть.

Рис. 6.6. Сводка для модели и оценка параметров простой регрессии

[23] Уравнение квадратичной регрессии имеет следующий вид: $y = a + b_1x + b_2x^2$.

Шаг 5. Для построения регрессии необходимо выбрать одну из предложенных моделей. Выбор между ними осуществляется через оценку значений R-квадрата. R-квадрат является статистической мерой, представляющей собой долю общей вариации зависимой переменной, которая может быть объяснена независимой переменной (предиктором) или набором предикторов в модели регрессии. R-квадрат принимает значения в диапазоне от 0 до 1, где 0 указывает на то, что модель не объясняет изменчивости данных, а 1 — на то, что модель идеально подходит к данным. В нашем примере обе модели примерно на 30% способны предсказать изменения значений зависимой переменной. В таком случае нет необходимости использовать более сложную модель и следует отдать предпочтение уравнению простой регрессии. Ошибок (или остатков) нет ни в одной модели, поэтому в уравнение мы их также не включаем.

Соответственно, уравнение регрессии будет иметь следующий вид:

$$\text{Количество запланированных детей исходя из обстоятельств} =$$

$$= 1{,}466 - 0{,}558 \times \text{Количество имеющихся детей.}$$

Шаг 6. Интерпретировать уравнение можно несколькими способами. Во-первых, если у респондента нет детей (значение $X = 0$), то он будет желать иметь от одного до двух детей ($У = 1{,}466$). Во-вторых, при увеличении имеющихся детей на одного ребенка количество запланированных будет уменьшаться на 0,558. Соответственно, те, у кого три и более детей, не хотят увеличивать семью ($1{,}466 - 0{,}558 \times 3 = -0{,}208$). Наконец, в-третьих, уравнение можно использовать для прогнозирования значений У на основе известных значений X. Например, если у респондента один ребенок, то с вероятностью в 95% он хотел бы завести еще одного ($1{,}466 - 0{,}558 \times 1 = 0{,}908$).

§ 6.3. Множественная линейная регрессия

Множественная регрессия расширяет идею простой линейной регрессии, которая используется для измерения влияния одной независимой переменной (предиктора) на зависимую (критерий). Однако в отличие от простой линейной регрессии, множественная исследует влияние двух или более независимых переменных на критерий. Другими словами, множественная регрессия может показать, как две или более независимые переменные взаимодействуют друг с другом, чтобы объяснить изменения в зависимой.

Еще одной особенностью множественной регрессии является способность контролировать влияние других переменных, которые не являются включенными в модель, при анализе влияния конкретных независимых переменных. Например, при анализе влияния дохода на расходы, множественная регрессия может учитывать влияние других факторов, таких как возраст, пол, образование, чтобы установить настоящую зависимость между доходом и расходами.

Также множественная регрессия может использоваться для прогнозирования будущих значений зависимой переменной на основе известных значе-

ний независимых переменных. Это позволяет принимать более обоснованные решения и планировать будущие действия, основанные на статистических данных и понимании связи между переменными.

Уравнение множественной линейной регрессии выглядит так:

$$Y = a + b_1X^1 + b_2X^2 + \dots b_nX_n,$$

где n — количество независимых переменных.

Алгоритм анализа множественной регрессии аналогичен работе с простой регрессией, однако предполагает знание еще нескольких коэффициентов. Разберем их на примере массива о репродуктивном поведении, но в этот раз добавим новые переменные: желаемый размер материнского капитала на первого ребенка, возраст, доход, количество имеющихся детей и количество желаемых детей в идеале.

Как было указано выше, одним из главных условием регрессионного анализа является отсутствие экстремальных значений. Поэтому в первую очередь проводим разведывательный анализ новых переменных. Он показал, что, в частности, средний размер желаемого материнского капитала на первого ребенка составляет 907 693 рублей (стандартное отклонение = 413 238). Огромные значения стандартного отклонения наводят на мысль о наличии выбросов, а статистика это подтверждает. В частности, ящичная диаграмма по материнскому капиталу на первого ребенка показывает (рис. 6.7), какие наблюдения являются максимальными, а какие принимают экстремальные значения и подлежат исключению из массива[24].

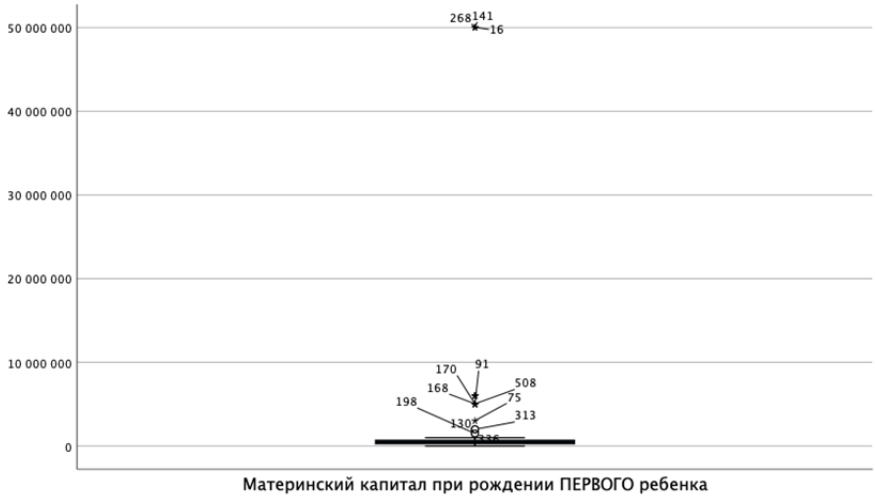


Рис. 6.7. Ящичная диаграмма в процедура разведывательного анализа

[24] Значения, удаленные от границ более, чем на три длины построенного прямоугольника (экстремальные значения), помечаются на диаграмме звездочками. Значения, удаленные более чем на полторы длины прямоугольника, помечаются кружками [2].

На диаграмме «ствол-лист» уже отмечено, какие значения относятся к выбросам (рис. 6.8). В нашем случае это 17 единиц анализа, чьи значения равны или превышают 150 000 рублей. Аналогичные данные получены и в отношении остальных переменных. В связи с этим требуется указать этот диапазон в качестве пропущенных значений.

Материнский капитал при рождении ПЕРВОГО ребенка

Материнский капитал при рождении ПЕРВО

Frequency	Stem &	Лист
23,00	0 .	0012334558
57,00	1 .	00000000000000000000000055&
46,00	2 .	00000000000000555555555
59,00	3 .	00000000000000000000000000005
17,00	4 .	00000555
103,00	5 .	000000000000000000000000000000000000000000000000000
25,00	6 .	0000000000005
21,00	7 .	0000000005
19,00	8 .	000000000
13,00	9 .	000000
43,00	10 .	00000000000000000000
17,00	Extremes	(>=1500000)

Ширина ствола: 100000
Каждый лист: 2 наблюдения

Рис. 6.8. Диаграмма «ствол-лист» в процедуре разведывательного анализа

В результате средние значения получились более реалистичными: желаемый материнский капитала на первого ребенка составил 393 859 рублей (стандартное отклонение = 229 986). А ящичная диаграмма для материнского капитала на первого ребенка стала более легкой для чтения (рис. 6.9).

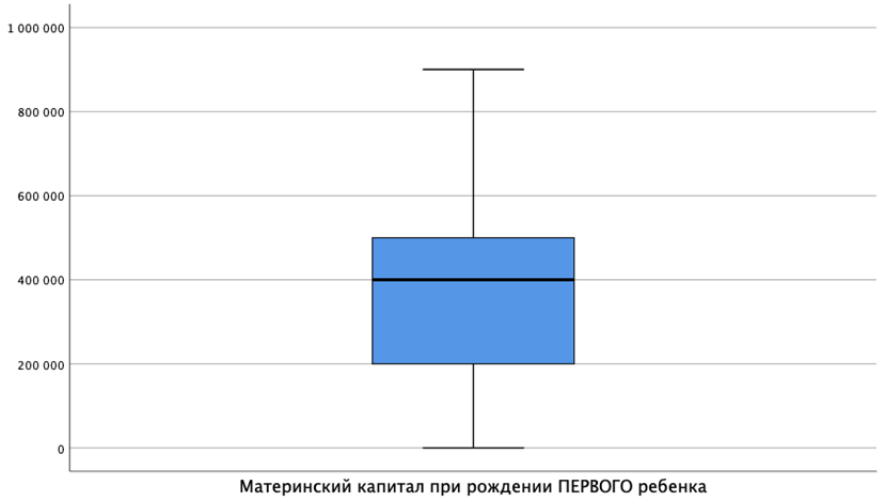


Рис. 6.9. Ящичная диаграмма после пропуска экстремальных значений переменной

Выполнив подготовительные процедуры, мы можем переходить к построению уравнения множественной регрессии.

Шаг 1. Исследовательский вопрос: как меняется ожидаемый размер материнского капитала на первого ребенка в зависимости от таких показателей, как возраст, доход, количество имеющихся детей и количество желаемых детей в идеале.

Шаг 2. Зависимая переменная — материнский капитал, остальные переменные — независимые (предикторы).

Шаг 3. Нулевая гипотеза — переменные не связаны, альтернативная — существует достоверное влияние всех или нескольких предикторов на размер материнского капитала.

Шаг 4. Диалоговое окно множественной регрессии находится так: анализ — регрессия — линейная. Заполняем все соответствующие поля и кликаем «ОК» (рис. 6.10).

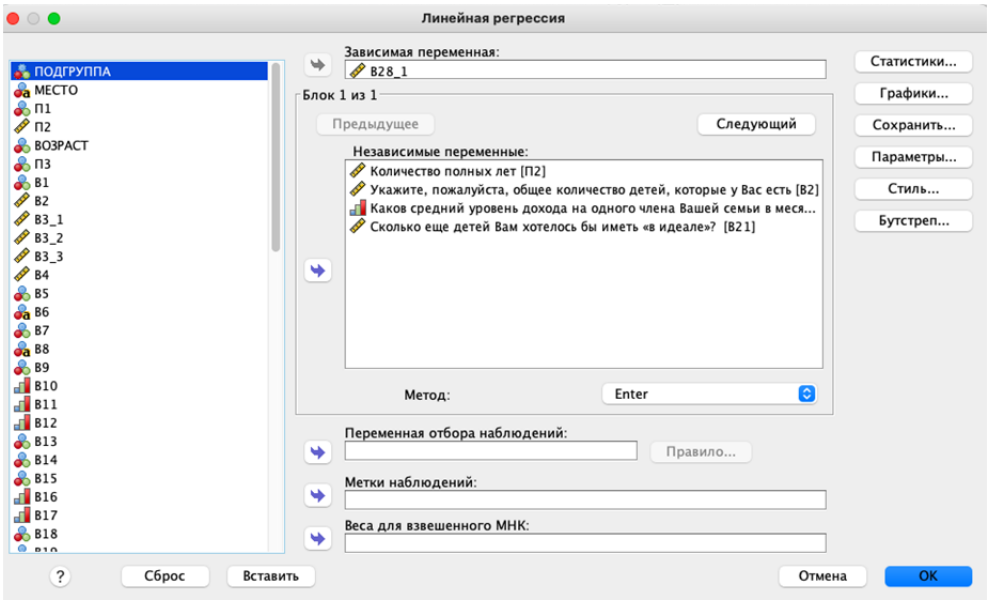


Рис. 6.10. Интерфейс окна линейной регрессии

Данный способ множественного регрессионного анализа выполнен методом ввода (Enter). В результате мы получили единственную модель, включающую все предикторы (рис. 6.11). Эта модель включает даже те переменные, в отношении которых не установлено значимого влияния на зависимую переменную (общее количество детей и средний доход). Здесь и далее следует обращать внимание еще на один важный коэффициент — бета (β), принимающий значения от –1 до 1. Он отражает *частную корреляцию* независимой и зависимой переменных. Если коэффициент бета положительный, то это означает, что при увеличении значений независимой переменной зависимая также

возрастает. Если коэффициент бета отрицательный, то при увеличении значений независимой переменной зависимая уменьшается[25]. Коэффициент бета является универсальной мерой влияния независимой переменной; его часто называют *стандартным коэффициентом регрессии*. И именно стандартные коэффициенты регрессии позволяют сравнивать независимые переменные по их влиянию на оценку зависимой переменной [3, с. 253]. Здесь мы можем принять нулевую гипотезу в отношении переменных, у которых самая большая значимость ($p > 0{,}05$) и наименьшие значения бета-коэффициента. Уравнение построить мы не можем, поскольку ряд предикторов не подходит по значимости, а их исключение даст иные значения в других коэффициентах. Поэтому социолог может выбрать следующие варианты представления данных: 1) представить эту модель в табличной форме; 2) выполнить множественный регрессионный анализ иным способом.

Коэффициенты[a]

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты		
		B	Стандартная ошибка	Бета	t	Значимость
1	(Константа)	230880,278	75498,688		3,058	,002
	Количество полных лет	7292,895	2266,696	,254	3,217	,001
	Укажите, пожалуйста, общее количество детей, которые у Вас есть	9247,845	20200,054	,038	,458	,648
	Каков средний уровень дохода на одного члена Вашей семьи в месяц?	14599,329	9705,775	,102	1,504	,134
	Сколько еще детей Вам хотелось бы иметь «в идеале»?	-56371,611	13830,199	-,251	-4,076	,000

а. Зависимая переменная: Материнский капитал при рождении ПЕРВОГО ребенка

Рис. 6.11. Модель множественной регрессии, выполненная методом ввода

Второй способ заключается в изменении способа ввода данных в модель и называется «пошагово». Этот метод позволяет сгенерировать данные, которые помогут определить, какая из независимых переменных оказывает наибольшее влияние на критерий. Для создания уравнения регрессии сначала

[25] Под частной корреляцией понимается воздействие, оказываемое на зависимую переменную со стороны одной из независимых переменных при фиксированных значениях других независимых переменных (с учетом влияния последних). Чем больше данная независимая переменная коррелирует с другими независимыми переменными, тем меньше абсолютная величина ее коэффициента бета. Для простой регрессии с одной независимой переменной коэффициент бета равен величине парной корреляции зависимой и независимой переменных.

включаются переменные, чья частная корреляция с зависимой переменной значима на уровне 0,05. Если какие-либо из включенных переменных показывают более низкий уровень значимости, они будут исключены из уравнения. Также будут созданы переменные для хранения прогнозируемых значений переменной помощи, рассчитанных по составленному уравнению регрессии.

В результате у нас получится таблица, содержащая значения коэффициентов для двух моделей (рис. 6.12). Для каждой из них уровень значимости менее 0,05, следовательно, нулевая гипотеза отклонена.

Коэффициенты[a]

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты		
		B	Стандартная ошибка	Бета	t	Значимость
1	(Константа)	497706,630	30699,287		16,212	,000
	Сколько еще детей Вам хотелось бы иметь «в идеале»?	-64690,867	13796,881	-,288	-4,689	,000
2	(Константа)	298049,208	55473,762		5,373	,000
	Сколько еще детей Вам хотелось бы иметь «в идеале»?	-61363,054	13357,372	-,273	-4,594	,000
	Количество полных лет	7288,113	1710,980	,253	4,260	,000

a. Зависимая переменная: Материнский капитал при рождении ПЕРВОГО ребенка

Рис. 6.12. Модели множественной регрессии выполненные методом «пошагово»

Также программа выводит исключенные из анализа переменные (рис. 6.13).

Исключенные переменные[a]

Модель		Бета-включения	t	Значимость	Частная корреляция	Статистика коллинеарности
						Допуск
1	Количество полных лет	,253[b]	4,260	,000	,264	,997
	Укажите, пожалуйста, общее количество детей, которые у Вас есть	,163[b]	2,681	,008	,170	,997
	Каков средний уровень дохода на одного члена Вашей семьи в месяц?	,018[b]	,279	,780	,018	,943
2	Укажите, пожалуйста, общее количество детей, которые у Вас есть	-,006[c]	-,073	,942	-,005	,568
	Каков средний уровень дохода на одного члена Вашей семьи в месяц?	,091[c]	1,437	,152	,092	,881

a. Зависимая переменная: Материнский капитал при рождении ПЕРВОГО ребенка

b. Предикторы в модели: (константа), Сколько еще детей Вам хотелось бы иметь «в идеале»?

c. Предикторы в модели: (константа), Сколько еще детей Вам хотелось бы иметь «в идеале»? , Количество полных лет

Рис. 6.13. Переменные, исключенные из моделей множественной регрессии

В окне вывода мы также обнаружим общую сводку для моделей (рис. 6.14). Во множественном регрессионном анализе R и R-квадрат — это две разные меры, используемые для оценки качества модели.

Сводка для модели

Модель	R	R-квадрат	Скорректированный R-квадрат	Стандартная ошибка оценки
1	,288[a]	,083	,079	189427,776
2	,383[b]	,147	,140	183079,464

a. Предикторы: (константа), Сколько еще детей Вам хотелось бы иметь «в идеале»?

b. Предикторы: (константа), Сколько еще детей Вам хотелось бы иметь «в идеале»? , Количество полных лет

Рис. 6.14. Сводка для модели множественной регрессии

R — это корреляционный коэффициент Пирсона между наблюдаемыми значениями зависимой переменной и прогнозируемыми значениями, полученными из множественной регрессии. R показывает, насколько точно модель прогнозирует значения зависимой переменной. Значение R может находиться в диапазоне от –1 до 1. Чем оно ближе к 1 или –1, тем сильнее положительная или отрицательная линейная связь между независимыми и зависимой переменными. Если значение R близко к 0, то связь между ними слабая или отсутствует.

R-квадрат (или коэффициент детерминации) — это процент вариации в зависимой переменной, который может быть объяснен множественным регрессионным уравнением. Он показывает, какую долю изменчивости зависимой переменной можно объяснить независимыми переменными в модели. Значение R-квадрат может находиться в диапазоне от 0 до 1. Чем оно ближе к 1, тем выше качество модели.

Таким образом, мы имеем две модели регрессии. При этом в каждой из них очень близкие значения бета-коэффициента (рис. 6.12), а значение R-квадрата во второй модели почти вдвое превышает значение в первой (рис. 6.14). Исходя из этого, мы делаем выбор в пользу второй модели.

Шаг 5. Выбрав модель, строим уравнение множественной регрессии по данным из таблицы «Коэффициенты».

Ожидаемый размер материнского капитала на первого ребенка =

$$= 497\,706 - 61\,363 \times \text{количество детей, которых хотелось бы иметь в идеале} +$$

$$+ 7288 \times \text{количество полных лет.}$$

Шаг 6. Интерпретируем аналогично уравнению простой регрессии. Коэффициент возраста равняется 7288. Это можно интерпретировать так: каждый дополнительный год к возрасту человека ассоциируется с 7288 дополнительными рублями к ожидаемому материнскому капиталу на первого ребенка *при неизменном количестве детей, которых хотелось бы иметь в идеале*. Для любой группы людей с одним и тем же количеством желаемых в идеале детей, кто на десять лет старше, ожидают размер материнского капитала в среднем на 72 000 рублей больше.

Обратим внимание, что в целом с возрастом ожидаемый размер материнского капитала увеличивается, но уменьшается с увеличением числа желаемых детей в идеале. С исследовательской точки зрения, это может говорить о двух типах респонденток: 1) репродуктивные намерения не определяются финансовой помощью со стороны государства; 2) финансовая поддержка государства требуется с возрастом.

§ 6.4. Контрольные вопросы

1. Что общего и различного в корреляционном и регрессионном анализе?
2. В чем суть анализа корреляций? Каков алгоритм выполнения корреляционного анализа в программе IBM SPSS Statistics?
3. Каковы параметрические и непараметрические критерии корреляции? Каковы условия их использования и особенности интерпретации?
4. В чем сходство и различие простой и множественной регрессии?
5. Каковы основные этапы выполнения простой регрессии в IBM SPSS Statistics? Как интерпретировать уравнение простой регрессии?
6. Каковы основные этапы выполнения множественной линейной регрессии в IBM SPSS Statistics? Как интерпретировать уравнение множественной регрессии?
7. В чем смысл основных коэффициентов множественной регрессии?

§ 6.5. Практические задания

1. Выполните анализ корреляций степени выраженности территориальной идентичности с возрастом респондентов, используя массив «Социально-политические настроения в Калининградской области», n = 915, 2022 г.
2. Выясните, существует ли статистически достоверная корреляция между таким переменными, как возраст, количество детей в идеале и размер материнского капитала. Используйте массив «Репродуктивные установки жительниц Калининградской области», n = 830, 2017 г.
3. Построив уравнение простой регрессии, определите, как меняется количество планируемых детей в идеале в зависимости от размера материнского капитала на первого ребенка. Используйте массив «Репродуктивные установки жительниц Калининградской области», n = 830, 2017 г.

4. Построив уравнение множественной регрессии, определите, как меняется количество планируемых детей в идеале в зависимости от материального объема различных форм государственной поддержки материнства и детства. Используйте массив «Репродуктивные установки жительниц Калининградской области», n = 830, 2017 г.

§ 6.6. Рекомендуемая литература

1. *Бююль А., Цёфель П.* SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей. СПб. : ДисСофтЮп, 2005.

2. *Воронин Г. Л.* Статистический анализ данных в IBM SPSS Statistics V27.0.1.0. Н. Новгород : ННГУ им. Н. И. Лобачевского, 2022.

3. *Наследов А. Д.* SPSS 19: профессиональный статистический анализ данных. СПб. : Питер, 2011.

4. *Романчуков С. В., Берестнева Е. В., Маклакова Т. Г.* Анализ социологических данных на основе корреляционного и факторного анализа // Information and mathematical technologies in science and management. 2017. № 2 (6). С. 72—77.

5. *Уиллан Ч.* Голая статистика. Самая интересная книга о самой скучной науке. М. : МИФ, 2016.

6. *Щекотуров А. В.* Политическое доверие и ценности лояльной и оппозиционной молодежи в эксклавном регионе России // Вестник Российского университета дружбы народов. Сер.: Политология. 2021. Т. 23, № 4. С. 570—583.

Глава 7
ФАКТОРНЫЙ АНАЛИЗ

§7.1. Суть и области применения

Факторный анализ — это метод статистического анализа, который позволяет исследовать структуру отношений между множеством переменных и помогает уменьшить количество измеряемых переменных, сводя их к меньшему количеству факторов. Факторы — это независимые переменные, которые имеют высокую корреляцию с группами измеряемых переменных, сведенных в один фактор. При этом в один фактор объединяются переменные, сильно коррелирующие между собой, в то время как переменные из разных факторов должны слабо коррелировать между собой [3, с. 368]. Главная задача факторного анализа — выявление скрытых (латентных) факторов, объясняющих вариацию в наборе измеряемых переменных.

Рассмотрим простой пример, чтобы проиллюстрировать суть факторного анализа.

Представьте, что у вас есть набор данных, который включает в себя информацию о доходах и образовании жителей города. При этом вы не знаете, какие конкретно признаки наиболее важны для определения общего уровня благосостояния. С помощью факторного анализа можно выделить скрытые факторы (например, социальный статус, образование, занятость и т. д.), которые наиболее сильно влияют на вариацию в доходах и образовании жителей города.

Другими словами, «факторный анализ позволяет установить для большого числа исходных признаков сравнительно узкий набор “свойств”, характеризующих связь между группами этих признаков и называемых факторами» [7, с. 278]. Концепция факторного анализа заключается в «сжатии» информации.

Таким образом, факторный анализ позволяет прояснить взаимосвязи между различными переменными, и выявить скрытые структуры, которые могут быть использованы для более точного моделирования, прогнозирования или принятия решений в различных областях. Социологический смысл анализа — измеряемые эмпирические показатели считаются следствием других, глубинных, скрытых характеристик (латентных переменных).

Метод факторного анализа является «продвинутой» статистической процедурой, поскольку предполагает не только знание программы IBM SPSS Statistics, но и понимание механизмов факторизации данных. Факторный анализ широко представлен в социологических публикациях, например, в работе [2], в которой они на основе факторного анализа выделяют три группы атти-

тюдов в позициях жителей Екатеринбурга по отношению к иноэтничным мигрантам из стран Центральной Азии. Результаты их исследования представлены в виде матрицы факторных нагрузок (рис. 7.1).

Индикатор		Фактор		
		1	2	3
1	Известность отрицательной или положительной информации о мигрантах	0,160	**0,577**	0,116
2	Отношение к росту численности мигрантов в городе	0,073	**0,743**	0,078
3	Готовность меняться под влиянием новых культурных норм мигрантов	0,031	**0,567**	0,047
4	Оценка степени привыкания к присутствию в городе большого числа представителей Центральной Азии	0,238	**0,600**	0,339
5	Оценка культуры поведения мигрантов	0,131	**0,690**	0,073
6	Приемлемость брачных отношений с мигрантами	**0,743**	0,051	–0,021
7	Приемлемость личной дружбы с мигрантами	**0,871**	0,116	0,208
8	Приемлемость соседства с мигрантами	**0,909**	0,141	0,182
9	Приемлемость быть коллегами по работе с мигрантами	**0,903**	0,166	0,200
10	Приемлемость быть жителями одного города с мигрантами	**0,821**	0,262	0,207
11	Приемлемость быть согражданами одной страны с мигрантами	**0,770**	0,220	0,170
12	Оценка числа факторов увеличения похожести мигрантов и местного населения	0,218	0,035	**0,688**
13	Число предъявляемых требований к мигрантам	0.012	–0,357	**0,433**
14	Знание культурных особенностей мигрантов	0,055	0,168	**0,479**
15	Готовность в деловых отношениях контактировать с мигрантами как с партнерами или руководителями	0,172	0,144	**0,694**
16	Готовность оказать мигрантам помощь (число видов помощи)	0,179	0,374	**0,577**
17	Отношение к получению мигрантами гражданства	0,240	**0,565**	0,154

Рис. 7.1. Пример результатов факторного анализа

Источник: [2].

Дополнительно о примерах использования факторного анализа в социологических публикациях можно прочитать в [1; 4].

§ 7.2. Алгоритм факторного анализа

Перед использованием факторного анализа необходимо убедиться в выполнении следующих условий:

1) нормальное распределение данных[26];

2) отсутствие выбросов;

[26] Поскольку в социологических исследованиях нормальное распределение переменных встречается редко, исследователь должен проанализировать средние значения и стандартное отклонение переменных, которые будут подвергнуты факторизации. Необходимо обратить внимание на те переменные, для которых среднее значение отличается от смыслового среднего значения на шкале [5, с. 128].

3) все признаки — количественные переменные (интервальные либо метрические).

Если данные соответствуют этим условиям, можно приступать к использованию факторного анализа.

Процедура факторного анализа состоит из пяти основных стадий:

1) постановка исследовательского вопроса и выбор релевантных ему переменных для анализа;

2) анализ корреляционной матрицы выбранных переменных, который позволяет выявить меру связи между всеми переменными;

3) извлечение факторов, которые объясняют вариацию в исходных переменных;

4) вращение факторов для создания упрощенной структуры, что позволяет получить более четкое представление о том, какие переменные наиболее связаны с каждым фактором;

5) интерпретация и дальнейшее использование полученных факторов.

Все пять операций мы рассмотрим на конкретном примере в следующем разделе, а стадии 2—4 рассмотрим в данном параграфе, поскольку их понимание принципиально важно для достижения корректных целей факторного анализа.

В программе IBM SPSS Statistics настройки факторного анализа устанавливаются так: анализ — снижение размерности — факторный анализ. Настройки корреляционной матрицы находятся во вкладке «Описательные». Здесь мы выбираем «Одномерные описательные», «Начальное решение», «Коэффициенты», «Уровни значимости» и «КМО и критерий сферичности Бартлетта» (рис. 7.2).

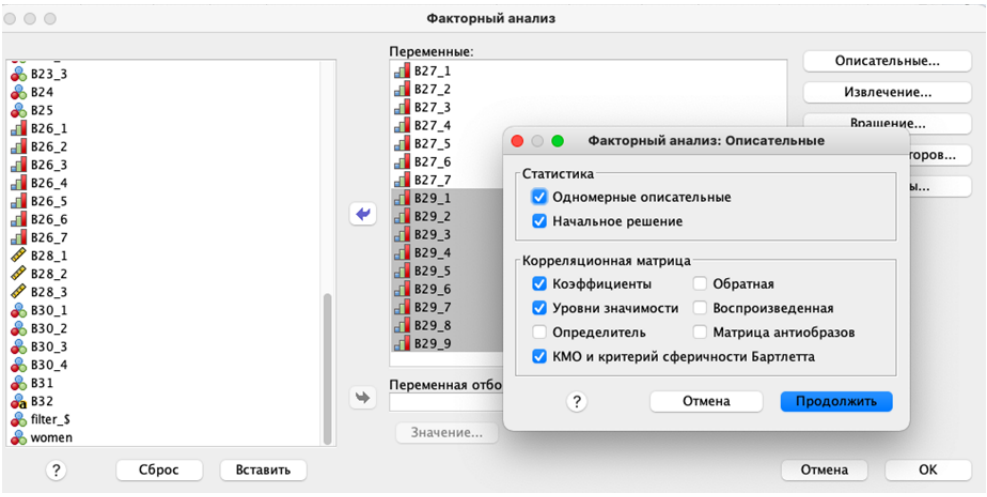


Рис. 7.2. Интерфейс команды «Факторный анализ»

Команда «Одномерные описательные» выведет таблицу со средними значениями и стандартными отклонениями для каждой переменной в анализе. Это позволит оценить наличие выбросов и внести коррективы (или вовсе отказаться от ряда переменных) в последующем. Флажок «Начальное решение» по умолчанию установлен и отвечает за включение в выводимые данные имен переменных, начальных общностей (по умолчанию равных 1,0), факторов, собственных значений, а также общего и кумулятивного процента общей дисперсии для каждого фактора.

Команды «Коэффициенты» и «Уровни значимости» построят корреляционную матрицу, в которой можно отследить наличие или отсутствие статистически достоверных корреляций. Для успешного факторного анализа важно, чтобы переменные были связаны друг с другом, но с различной силой.

KMO и критерий сферичности Барлетта — два критерия, описывающие многомерную нормальность (Бартлетта) и адекватность выборки. KMO — это мера Кайзера — Мейера — Олкина, которая оценивает соответствие между переменными и факторами, полученными из факторного анализа, определяет применимость факторного анализа к выбранным переменным. Она принимает значения от 0 до 1, где 0 означает, что выборка данных не подходит для факторного анализа, а 1 — что она идеально подходит. Обычно считается, что значение KMO больше 0,6 является достаточно хорошим для проведения факторного анализа. В целом исследователи придерживаются следующих степеней применимости факторного анализа [7, с. 294]:

— более 0,9 — безусловная адекватность;
— более 0,8 — высокая адекватность;
— более 0,7 — приемлемая адекватность;
— более 0,6 — удовлетворительная адекватность;
— более 0,5 — низкая адекватность;
— менее 0,5 — факторный анализ неприменим к выборке.

Критерий сферичности Бартлетта используется для проверки гипотезы о том, что данные коррелируют между собой и могут быть объединены в небольшое число факторов. Если корреляция между переменными в выборке данных значима, то можно провести факторный анализ ($p < 0{,}05$). Однако, если корреляция между переменными низкая, факторный анализ может быть неточен.

Извлечение факторов является следующим этапом факторного анализа. В программе IBM SPSS Statistics настройки извлечения факторов находятся на кнопке «Извлечение». Там следует выбрать «График собственных значений» и оставить выбранными остальные пункты по умолчанию. На данном этапе в окне вывода мы обратим внимание на таблицы «Общности», «Объясненная совокупная дисперсия» и «График собственных значений».

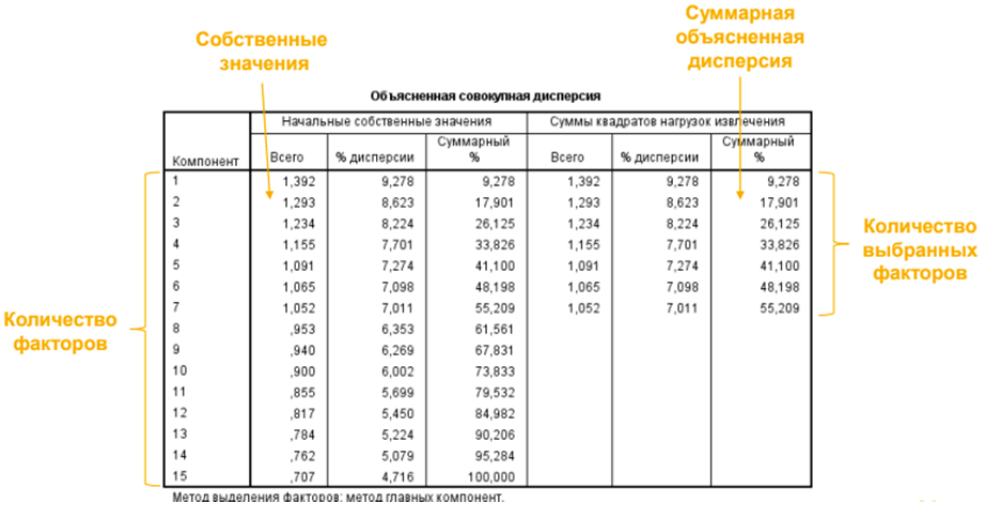
Извлечение факторного анализа начинается с вычисления суммарного разброса значений всех включенных в анализ переменных. Затем программа выбирает группу переменных, взаимная корреляция которых обусловливает наибольшую часть общей дисперсии. Эта группа станет первым фактором. После выделения он исключается, и процесс повторяется. Оставшиеся пере-

менные используются для выбора следующего фактора, который также определяется на основе максимальной взаимной корреляции для оставшейся общей дисперсии. Продолжение этой процедуры дает возможность выделять все остальные факторы до тех пор, пока не будет использована вся общая дисперсия переменных.

По умолчанию в процедуре факторного анализа каждая переменная имеет единичное значение *общности*. После извлечения каждая переменная получает свое собственное значение (см. таблицу «Общности» в программе). Этот показатель равен доле дисперсии переменной, обусловленной совокупным влиянием факторов. Общность можно сравнить с множественным коэффициентом корреляции R, принимающим значение 0 в случае, если факторы не влияют на переменную, и 1 в случае, если дисперсия переменной целиком определяется выделяемыми факторами [7, с. 280].

После того как программа извлекает первый фактор, напротив его номера появляется его *собственное значение* (рис. 7.2). Собственное значение фактора пропорционально доле общей дисперсии, определяемой данным фактором (обратите внимание, что здесь речь идет о факторе, а не о переменной, как было в случае общности). Затем вычисляется процент дисперсии, обусловливаемый данным фактором и равный отношению собственного значения фактора к числу переменных, а также соответствующий кумулятивный (накопленный) процент. С извлечением каждого нового фактора собственные значения уменьшаются, а кумулятивный процент приближается к 100.

Отбор факторов может выполняться по значениям в таблице «Объясненная совокупная дисперсия» (рис. 7.3). Согласно формальным критериям, отбираются те факторы, чьи собственные значения превышают единицу, а суммарный процент объясненной совокупной дисперсии достигает наибольших значений.



Объясненная совокупная дисперсия

Компонент	Начальные собственные значения			Суммы квадратов нагрузок извлечения		
	Всего	% дисперсии	Суммарный %	Всего	% дисперсии	Суммарный %
1	1,392	9,278	9,278	1,392	9,278	9,278
2	1,293	8,623	17,901	1,293	8,623	17,901
3	1,234	8,224	26,125	1,234	8,224	26,125
4	1,155	7,701	33,826	1,155	7,701	33,826
5	1,091	7,274	41,100	1,091	7,274	41,100
6	1,065	7,098	48,198	1,065	7,098	48,198
7	1,052	7,011	55,209	1,052	7,011	55,209
8	,953	6,353	61,561			
9	,940	6,269	67,831			
10	,900	6,002	73,833			
11	,855	5,699	79,532			
12	,817	5,450	84,982			
13	,784	5,224	90,206			
14	,762	5,079	95,284			
15	,707	4,716	100,000			

Метод выделения факторов: метод главных компонент.

Рис. 7.3. Объясненная совокупная дисперсия в ходе факторного анализа

Источник: [8].

Однако можно воспользоваться и критерием «каменистой осыпи» Р. Кеттелла, суть которого состоит в поиске точки, где убывание собственных значений замедляется наиболее сильно. Для выбора факторов методом Р. Кеттела следует обратить внимание на график собственных значений (рис. 7.4). Как правило, оба метода должны совпадать в своих результатах. Если факторов окажется столько же, сколько исходных переменных, факторный анализ теряет смысл, поскольку его целью является сокращение исходного набора переменных.

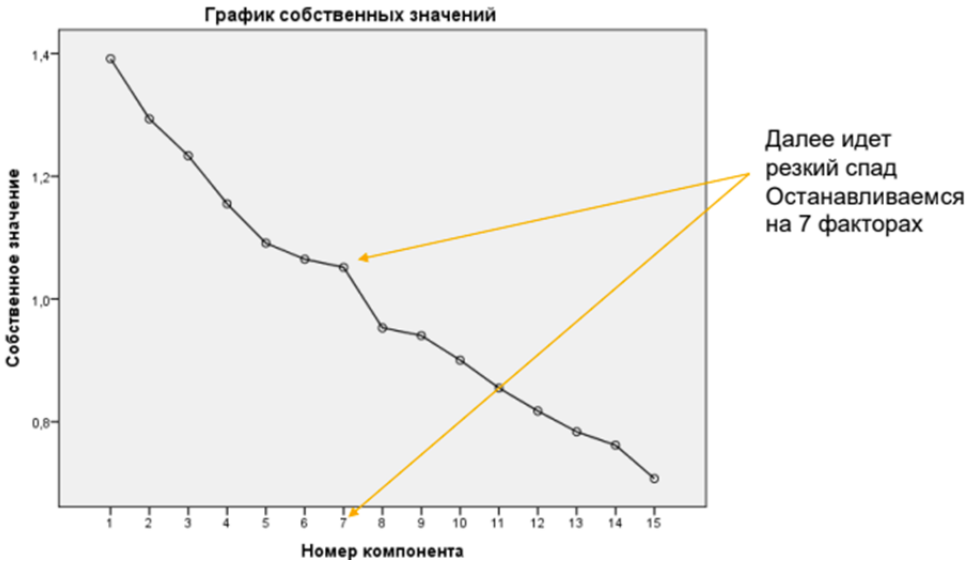


Рис. 7.4. График собственных значений факторов

Источник: [8].

Следующий этап — вращение и выбор факторов. В программе IBM SPSS Statistics настройки вращения факторов находятся на кнопке «Вращение». Там следует выбрать «Варимакс» и оставить выбранными остальные пункты по умолчанию. На данном этапе в окне вывода мы обратим внимание на таблицы «Объясненная совокупная дисперсия», «График собственных значений» и «Повернутая матрица компонентов».

Целью вращения при анализе факторов является упрощение структуры данных путём выявления переменных, которые сильно связаны только с одним фактором, и слабо связаны с остальными. Нагрузка, которая отражает степень связи между переменной и фактором, является аналогом коэффициента корреляции, и может принимать значения от – 1 до 1. Идеальной простой структурой является та, в которой каждая переменная имеет нулевые значения нагрузок для всех факторов, кроме одного, нагрузка которого для этой переменной близка к 1 (или – 1).

В учебнике А. Наследова хорошо проиллюстрирована процедура вращения факторов [7, с. 281].

До вращения (слева) точки, соответствующие переменным, расположены на удалении от осей факторов. После поворота осей (справа) они оказываются вблизи осей, что соответствует максимальной нагрузке каждой переменной только по одному фактору (рис. 7.5).

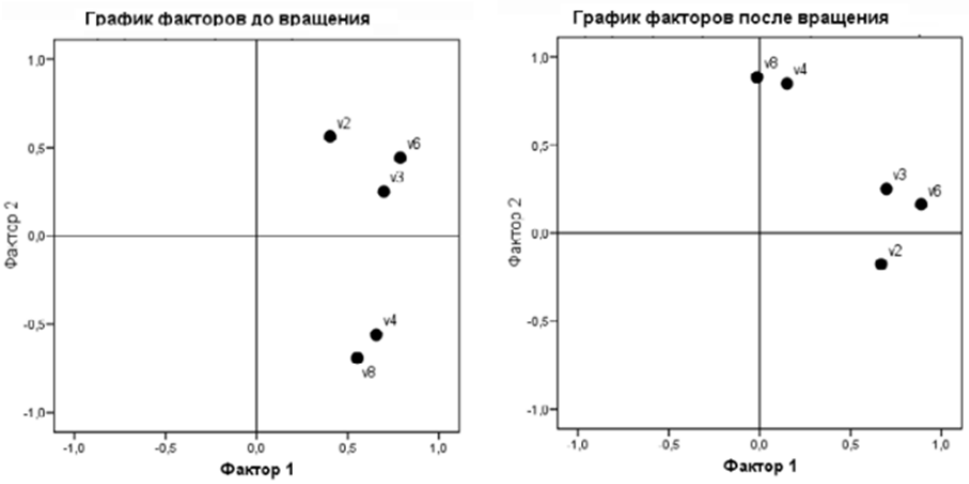


Рис. 7.5. Графики вращения факторов

Источник: [7, с. 281].

При этом вращение факторов не оказывает влияния на математическую точность анализа, поскольку взаимное расположение переменных не изменяется при повороте осей. Такой метод вращения факторов, как «Варимакс» сохраняет взаимное расположение осей, располагая их под прямым углом, что также не влияет на математическую точность анализа. Результатом применения метода «Варимакс» являются новые вращенные факторы, которые могут быть более простыми и интерпретируемыми, чем исходные невращенные.

В результате вращения «Варимакс» следует обратить внимание на таблицу «Повернутая матрица компонентов» (рис. 7.11). Именно эта матрица является главным итогом факторного анализа и подлежит содержательной интерпретации. Как пишет Бююль [3, с. 371], «здесь начинается самая интересная часть факторного анализа: вы должны попытаться объяснить отобранные факторы. Для этого возьмите в руки карандаш и в каждой строке повернутой факторной матрицы отметьте ту факторную нагрузку, которая имеет наибольшее абсолютное значение». Об особенностях проведения факторного анализа и интерпретации полученных факторов в реальной социологической практике мы расскажем в следующем параграфе.

§7.3. Пример факторного анализа

В качестве рабочего массива оставим результаты опроса женщин Калининградской области относительно их репродуктивных установок (2017 г.). Допустим, нас интересуют причины или условия, которые определяют отношение к различным формам государственной поддержки рождаемости. В ходе опроса респонденту предлагалось оценить степень эффективности шестнадцати вариантов форм государственной поддержки. Ответ было предложено выразить по пятибальной шкале, где 1 означает абсолютную неэффективность меры, 5 — высокую эффективность и 99 — вариант для тех, кто затрудняется ответить[27]. В результате выборка сократилась с 835 наблюдений до 271 респондента (рис. 7.6).

Описательные статистики

	Среднее	Станд. отклонения	Анализ N
Пособие по беременности и родам	2,02	1,202	271
Единовременное пособие при рождении ребенка	1,99	1,130	271
Материнский капитал	2,32	1,355	271
Частично оплачиваемый отпуск по уходу за ребенком 1,5 лет	1,93	1,230	271
Ежемесячное пособие на ребенка в семьях с низкими доходами	1,72	1,187	271
Жилищные кредиты на льготных условиях молодой семье	1,82	1,254	271
Компенсация затрат на оплату детских дошкольных учреждений	1,78	1,205	271
Региональный материнский капитал (в дополнение к федеральному)	3,00	1,402	271
Земельные участки под строительство (при рождении третьего и последующего ребенка)	3,07	1,438	271
Пособие на третьего ребенка до 3-х лет для семей, чей доход ниже среднего	2,79	1,441	271
Погашение кредита (для 2-х детей - частичное, 3 и более - полное)	2,96	1,527	271
Ясли или детский сад по выбору	2,84	1,464	271
Повышение компенсации на оплату детских дошкольных учреждений	2,70	1,511	271
Гибкий график работы для родителей	2,85	1,585	271
Оплата детского сада семьям, где работают оба родителя	2,76	1,612	271
Право на один дополнительный выходной в месяц	2,78	1,629	271

Рис. 7.6. Статистика переменных, участвующих в факторном анализе

[27] Этот вариант был исключен из анализа с целью получения адекватных средних значений.

Анализ корреляционной матрицы показал высокую статистически достоверную корреляцию между переменными ($p < 0{,}001$)[28]. Мера КМО и критерий Бартлетта свидетельствуют о безусловной адекватности выборки для факторного анализа (рис. 7.7).

КМО и критерий Бартлетта

Мера адекватности выборки Кайзера-Майера-Олкина (КМО).		,905
Критерий сферичности Бартлетта	Примерная Хи-квадрат	5617,767
	ст.св.	120
	Значимость	,000

Рис. 7.7. Критерии оценки адекватности выборки для факторного анализа

Таблица «Общности» содержит значения переменных после извлечения факторов. Как мы видим, значения всех переменных довольно высоки и подлежат факторизации[29] (рис. 7.8).

Общности

	Начальная	Извлечение
Пособие по беременности и родам	1,000	,769
Единовременное пособие при рождении ребенка	1,000	,745
Материнский капитал	1,000	,661
Частично оплачиваемый отпуск по уходу за ребенком 1,5 лет	1,000	,838
Ежемесячное пособие на ребенка в семьях с низкими доходами	1,000	,791
Жилищные кредиты на льготных условиях молодой семье	1,000	,744
Компенсация затрат на оплату детских дошкольных учреждений	1,000	,843
Региональный материнский капитал (в дополнение к федеральному)	1,000	,748
Земельные участки под строительство (при рождении третьего и последующего ребенка)	1,000	,731
Пособие на третьего ребенка до 3-х лет для семей, чей доход ниже среднего	1,000	,756
Погашение кредита (для 2-х детей – частичное, 3 и более – полное)	1,000	,849
Ясли или детский сад по выбору	1,000	,869
Повышение компенсации на оплату детских дошкольных учреждений	1,000	,870
Гибкий график работы для родителей	1,000	,800
Оплата детского сада семьям, где работают оба родителя	1,000	,844
Право на один дополнительный выходной в месяц	1,000	,815

Метод выделения факторов: метод главных компонент.

Рис. 7.8. Значения переменных после извлечения факторов

[28] Большой размер таблицы не позволяет разместить ее в учебном пособии.

[29] По рекомендации Г.Л. Воронина, первичные переменные с показателями менее 0,2 желательно в факторизацию не включать [5, с. 131].

Для выделения факторов проводим анализ таблицы «Объясненная совокупная дисперсия» (рис. 7.9) и графика собственных значений (рис. 7.10).

Объясненная совокупная дисперсия

Компонент	Начальные собственные значения			Извлечение суммы квадратов нагрузок			Ротация суммы квадратов нагрузок		
	Всего	% дисперсии	Суммарный %	Всего	% дисперсии	Суммарный %	Всего	% дисперсии	Суммарный %
1	8,547	53,417	53,417	8,547	53,417	53,417	7,241	45,258	45,258
2	4,123	25,771	79,188	4,123	25,771	79,188	5,429	33,930	79,188
3	,899	5,617	84,806						
4	,499	3,120	87,925						
5	,467	2,917	90,843						
6	,231	1,442	92,285						
7	,217	1,355	93,640						
8	,188	1,174	94,814						
9	,172	1,077	95,892						
10	,134	,838	96,730						
11	,123	,768	97,497						
12	,104	,648	98,145						
13	,099	,619	98,764						
14	,078	,485	99,249						
15	,061	,382	99,631						
16	,059	,369	100,000						

Метод выделения факторов: метод главных компонент.

Рис. 7.9. Объясненная совокупная дисперсия

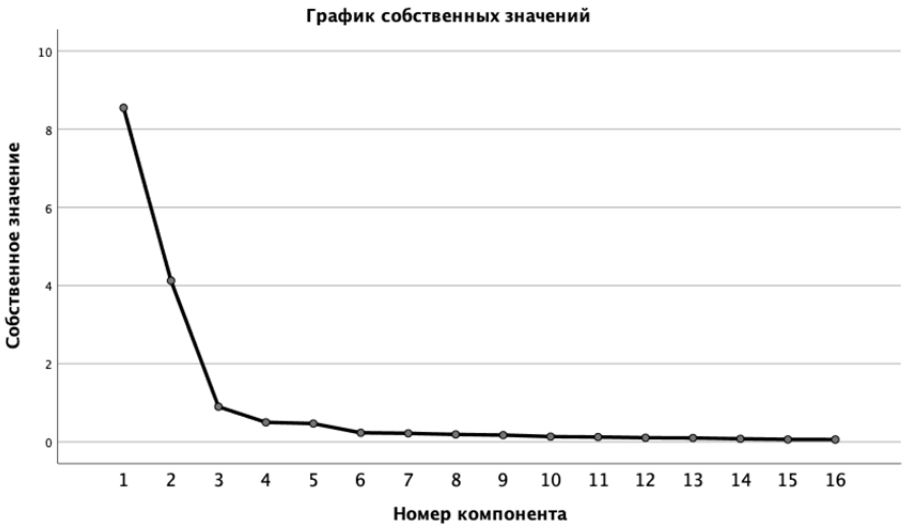


Рис. 7.10. График собственных значений факторов

Из этих рисунков мы видим, что существуют все основания для выбора двух факторов, чей кумулятивный процент объясняет 79 % общей дисперсии.

Наконец, мы подошли к финальной таблице — повернутой матрице компонентов (рис. 7.11). Для более наглядной демонстрации устанавливаем формат вывода коэффициентов на уровне не менее 0,3 (анализ — снижение размерности — факторный анализ — параметры — не выводить коэффициенты с низкими значениями) [6]. Первый фактор включает переменные, относящи-

еся к более зрелым (возрастным) и многодетным родителям, которым необходимы и дополнительные финансовые средства, землю под большой дом, гибкий график и так далее. Второй фактор группирует переменные, более значимые для более молодых респондентов, которые готовы пока завести первого ребенка. Таким образом, отвечая на исследовательский вопрос, мы констатируем, что за различным отношением к государственным мерам поддержки рождаемости стоят социально-экономические и возрастные факторы, а именно: более зрелые, работающие и многодетные родители с одной стороны, и низкодоходные, учащиеся или безработные респондентки — с другой.

Повернутая матрица компонентов[a]

	Компонент	
	1	2
Пособие по беременности и родам		,875
Единовременное пособие при рождении ребенка		,860
Материнский капитал		,799
Частично оплачиваемый отпуск по уходу за ребенком 1,5 лет		,900
Ежемесячное пособие на ребенка в семьях с низкими доходами		,875
Жилищные кредиты на льготных условиях молодой семье		,841
Компенсация затрат на оплату детских дошкольных учреждений		,895
Региональный материнский капитал (в дополнение к федеральному)	,848	
Земельные участки под строительство (при рождении третьего и последующего ребенка)	,845	
Пособие на третьего ребенка до 3-х лет для семей, чей доход ниже среднего	,850	
Погашение кредита (для 2-х детей - частичное, 3 и более - полное)	,898	
Ясли или детский сад по выбору	,921	
Повышение компенсации на оплату детских дошкольных учреждений	,920	
Гибкий график работы для родителей	,891	
Оплата детского сада семьям, где работают оба родителя	,912	
Право на один дополнительный выходной в месяц	,895	

Метод выделения факторов: метод главных компонент.
Метод вращения: варимакс с нормализацией Кайзера.

a. Вращение сошлось за 3 итераций.

Рис. 7.11. Повернутая матрица компонентов

После получения приемлемого решения можно вычислить факторные оценки для объектов, которые будут сохранены как новые переменные для дальнейшего анализа. Для этого в диалоговом окне «Факторный анализ» необходимо щелкнуть на кнопке «Значения факторов» и в открывшемся диалоговом окне установить флажок «Сохранить как переменные». В итоге будут созданы новые переменные (по количеству факторов), которые можно использовать в дальнейшем анализе в дополнении к исходным переменным.

§ 7.4. Контрольные вопросы

1. Каковы основные цели использования факторного анализа в социологических исследованиях?
2. Какова процедура факторного анализа в программе IBM SPSS Statistics?
3. Как интерпретировать результаты факторного анализа?

§ 7.5. Практические задания

1. Выполните процедуру факторного анализа на переменных о доверии взрослого населения Калининградской области (старше 35 лет) общественно-политическим институтам и учреждениям. Сформулируйте исследовательский вопрос и дайте интерпретацию полученным результатам. Используйте массив «Социально-политические настроения в Калининградской области», n = 915, 2022 г.
2. Выполните процедуру факторного анализа на переменных о доверии молодежи Калининградской области (18—35 лет) общественно-политическим институтам и учреждениям. Сформулируйте исследовательский вопрос и дайте интерпретацию полученным результатам. Используйте массив «Социально-политические настроения молодежи Калининградской области», n = 987, 2021 г.

§ 7.6. Рекомендуемая литература

1. *Бессокирная Г. Л.* Факторный анализ: традиции использования и новые возможности // Социология: 4М. 2000. № 12. С. 142—153.

2. *Бритвина И. Б., Могильчак Е. Л., Савчук Г. А.* Отношение жителей уральского мегаполиса к иноэтничным мигрантам из стран Центральной Азии: факторный анализ // Вестник Томского государственного университета. Философия. Социология. Политология. 2018. № 44. С. 137—146.

3. *Бююль А., Цёфель П.* SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей. СПб. : ДисСофтЮп, 2005.

4. *Воронин Г.Л.* К вопросу о конструировании социологического теста (измерение ментальности) // Социология: 4М. 2002. № 15. С. 93—109.

5. *Воронин Г.Л.* Статистический анализ данных в IBM SPSS Statistics V27.0.1.0. Н. Новгород : ННГУ им. Н.И. Лобачевского, 2022.

6. *Крыштановский А.О.* «Кластеры на факторах» — об одном распространенном заблуждении // Социология: 4М. 2005. № 21. С. 172—187.

7. *Наследов А.Д.* SPSS 19: профессиональный статистический анализ данных. СПб. : Питер, 2011.

8. *Преподавателям* SPSS. URL: https://nafi.ru/academy/prepodavatelyam-spss/ (дата обращения: 05.10.2023).

Глава 8
КЛАСТЕРНЫЙ АНАЛИЗ

§ 8.1. Сравнение факторного и кластерного анализа

Кластерный анализ — это метод статистического анализа, который используется для выявления групп (кластеров) сходных объектов на основе их признаков. Кластерный анализ предназначен для разбиения исходных данных на поддающиеся интерпретации группы, таким образом, чтобы элементы, входящие в одну группу, были максимально «схожи», а элементы из разных групп были максимально «отличными» друг от друга.

Этот метод широко используется в социологии для создания типологий или классификаций общественных явлений, например, групп людей с конкретными социально-демографическими и экономическими характеристиками или общественных организаций с определенными стратегиями деятельности и целями. Это позволяет получить более полное представление о структуре данных и выделить особенности каждой группы объектов.

Примером использования кластерного анализа в обработке социологической информации может послужить работа Н. А. Дерман [5], в которой исследуются социально-демографические характеристики кластеров, выделенных на основе степени пристрастия респондентов к потреблению алкогольных напитков и курению. Характеристики кластеров в разрезе исследуемых переменных представлены на рисунке 8.1.

Можно выделить и ряд других работ, в которых с разной степенью глубины и методологической новизны используется кластерный анализ [1; 4; 7; 10; 11].

Факторный анализ и кластерный анализ — это два различных метода многомерного анализа данных, которые используются для обработки и классификации множества переменных. Они имеют некоторые сходства, но основное различие заключается в том, что факторный анализ целевым объектом рассмотрения принимает переменные, а кластерный анализ — объекты.

Основные сходства между факторным и кластерным анализом:

1) оба метода используются для сжатия информации: факторный анализ позволяет свести кратное число переменных к меньшему числу факторов, а кластерный — сгруппировать объекты по сходству между ними;

2) оба метода используют матричный подход: факторный анализ основывается на матрице коэффициентов корреляции между переменными, а кластерный — на матрице расстояний между объектами.

ПЕРЕМЕННЫЕ	Кластеры				
	Воздержи-вающиеся	Умеренно выпиваю-щие	Умеренно выпиваю-щие и регулярно курящие	Пьющие	Злоупотребляю-щие
N кластера Уд. вес в выборке	485 33,6%	390 27,1%	143 9,9%	211 14,6%	213 14,8%
КУРЕНИЕ	*1,12* *(0,32)*	*1,17* *(0,41)*	*3,64* *(0,67)*	*1,26* *(0,44)*	*3,73* *(0,68)*
нет	88,5	83,5	–	74,3	–
редко	11,5	15,1	–	25,7	–
1–10 сигарет в день	–	1,0	46,8	–	39,5
11–20 сигарет в день	–	–	42,7	–	47,4
20 + сигарет в день	–	–	10,7	–	13,1
УПОТРЕБЛЕНИЕ АЛКОГОЛЯ	*1,61* *(0,49)*	*3,41* *(0,57)*	*2,12* *(0,81)*	*3,84* *(0,63)*	*4,02* *(0,62)*
нет	39,5	–	25,2	–	–
редко	60,5	–	39,2	–	–
1–2 раза в месяц	–	62,6	33,5	28,9	18,3
более 1 раза в неделю	–	33,6	2,1	58,3	62,0
почти каждый день	–	3,8	–	12,8	19,7
6 + РЮМОК ПОДРЯД	*1,30* *(0,46)*	*1,71* *(0,46)*	*1,67* *(0,60)*	*3,35* *(0,53)*	*3,26* *(0,93)*
нет	69,5	29,2	39,2	–	–
редко	30,5	70,8	53,8	–	23,6
1–2 раза в месяц	–	–	7,0	68,1	35,8
более 1 раза в неделю	–	–	–	29,5	30,7
почти каждый день	–	–	–	2,4	9,9

Рис. 8.1. Статистика кластеров в разрезе исследуемых переменных

Источник: [5].

Основные различия между факторным и кластерным анализами:

1) факторный анализ применяется для выделения скрытых факторов, которые описывают вариативность в исходном наборе переменных, а кластерный — для выделения групп сходных объектов;

2) факторный анализ использует методы уменьшения размерности, такие как метод главных компонент, а кластерный — различные методы группировки объектов, такие как иерархический кластерный анализ, метод k-средних и т. д.

§ 8.2. Особенности процедуры кластерного анализа

Программа SPSS реализует три метода кластерного анализа: двухэтапный кластерный анализ (TwoStep), кластеризация k-средними (k-means) и иерархическая кластеризация (Hierarchical).

Двухэтапный кластерный анализ позволяет выявить группы (кластеры) объектов по заданным переменным, если они действительно существуют. При этом программа автоматически определяет количество существующих кластеров (групп). Если невозможно однозначно определить количество кластеров, все объекты помещаются в один [9, с. 296].

Двухэтапный кластерный анализ. Для многих приложений процедура двухэтапного кластерного анализа окажется подходящим выбором. Она дает следующие уникальные возможности:

1) автоматический выбор наилучшего числа кластеров и мер для выбора кластерной модели;

2) возможность строить кластерные модели одновременно на основе и категориальных, и непрерывных переменных;

3) сохранение модели кластеров во внешнем XML файле для дальнейшего его считывания и обновления модели кластеров на основе новых данных [6].

Кроме того, процедура двухэтапного кластерного анализа может обрабатывать большие файлы данных.

Иерархическая кластеризация, как наиболее гибкий из рассматриваемых методов, позволяет детально исследовать структуру различий между объектами и выбрать наиболее оптимальное число кластеров. Используется при небольшом количестве наблюдений.

Применение процедуры иерархического кластерного анализа ограничивается небольшими файлами данных (сотни объектов для кластеризации), однако она обладает следующими уникальными возможностями:

1) способность разбивать на кластеры как наблюдения, так и переменные;

2) способность формировать диапазон возможных решений и сохранять принадлежность к кластерам для каждого из этих решений;

3) наличие нескольких методов формирования кластеров, преобразования переменных и измерения расстояний между кластерами [6].

Процедура иерархического кластерного анализа может обрабатывать интервальные (непрерывные), двоичные переменные или количества, если все переменные имеют один и тот же тип.

Кластеризация k-средними разбивает по заданным переменным все множество объектов на заданное пользователем число кластеров так, чтобы средние значения для кластеров по каждой из переменных максимально различались. Как правило, используется при большом количестве наблюдений. По мнению Г. Л. Воронина, является одним из популярных в социологии [3, с. 134].

Применение процедуры кластерного анализа методом k-средних ограничивается непрерывными данными и требует задания числа классов заранее, но она имеет следующие уникальные возможности:

1) способность сохранять расстояние от центра кластера до каждого объекта;

2) способность считывать начальные центры кластеров из внешнего файла IBM SPSS Statistics и сохранять в нем окончательные центры кластеров [6].

Иерархический кластерный анализ. В этом типе кластерного анализа каждое наблюдение образовывает сначала свой отдельный кластер. На первом шаге анализа два соседних кластера объединяются в один. Этот процесс продолжается до тех пор, пока их не останется только два. Расстояние между кластерами является средним значением всех расстояний между всеми возможными парами точек из обоих кластеров (between-groups linkage — связь между группами) [2, с. 387].

В целом иерархический кластерный анализ состоит из шести этапов (рис. 8.2).

Рис. 8.2. Основные этапы иерархического кластерного анализа

1. *Формулировка проблемы*. Как ранее уже было сказано, кластерный анализ используется для категоризации объектов на основе сходства измеренных признаков с целью выделения групп или типологий. В этом контексте научной проблемой могла бы стать потребность в классификации каких-либо объектов таким образом, чтобы внутри каждой группы они были бы более похо-

жи друг на друга, чем на объекты из других групп. Для анализа могут быть использованы как порядковые, так и количественные переменные. Допустимы случаи, когда переменные имеют различные шкалы измерения. В IBM SPSS Statistics используется стандартизация, в частности ее простой метод — нормализация переменных, приводящая все переменные к стандартной z-шкале [9, с. 299].

2. *Выбор способа измерения расстояния*. У нас нет возможности сделать полный обзор всех коэффициентов, поэтому остановимся лишь на некоторых. По умолчанию в программе IBM SPSS Statistics используется квадрат Евклидова расстояния, согласно которому расстояние между объектами равно сумме квадратов разностей между значениями одноименных переменных объектов. Евклидово расстояние и его квадрат разумно применять для анализа количественных данных [8, с. 208].

3. *Выбор метода кластеризации* — это способ вычисления расстояний между кластерами. В IBM SPSS Statistics существуют следующие основные методы кластеризации: межгрупповая связь; внутригрупповая связь; ближайший сосед; самый дальний сосед; центроидная кластеризация; медианная кластеризация; метод Варда. По умолчанию используется метод межгруппового связывания, который и будет рассмотрен в этом параграфе. Здесь программа вычисляет наименьшее среднее значение расстояния между всеми парами групп и объединяет две группы, оказавшиеся наиболее близкими. На первом шаге, когда все кластеры представляют собой одиночные объекты, данная операция сводится к обычному попарному сравнению расстояний между объектами. Термин «среднее значение» приобретает смысл лишь на втором этапе, когда сформированы кластеры, содержащие более одного объекта [9, с. 299].

4. Количество кластеров можно задать во вкладке «Статистики» или включить в вывод все кластеры, выбрав опцию «Нет» (рис. 8.3). Как и в случае факторного анализа, при выборе числа кластеров необходимо руководствоваться практическими и теоретическими соображениями. В иерархической кластеризации в качестве критерия используются расстояния. Необходимо смотреть на коэффициент в протоколе объединения (расстояние между двумя кластерами, определенное на основании выбранной дистанционной меры с учётом предусмотренного преобразования значений). Когда мера расстояния между двумя кластерами увеличивается скачкообразно, процесс объединения в новые кластеры необходимо остановить [2, с. 389]. Иначе будут объединены кластеры, находящиеся на большом расстоянии друг от друга. Оптимальным считается число кластеров равное разности количества наблюдений и шагов, после которого коэффициент увеличивается скачкообразно. Размеры кластеров должны быть значимыми.

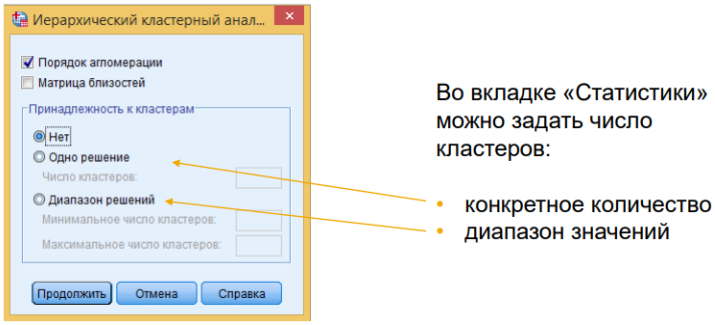


Рис. 8.3. Интерфейс команды иерархического кластерного анализа

Важным результатом иерархического кластерного анализа является порядок агломерации кластеров (рис. 8.4) и дендрограмма, демонстрирующая ход этого объединения (рис. 8.5). Пример таблицы и ее описания представлен у А. Д. Наследова: «В этой таблице вторая колонка “Кластер объединен с” содержит первый (Кластер 1) и второй (Кластер 2) столбцы, которые соответствуют номерам кластеров, объединяемых на данном шаге. После объединения кластеру присваивается номер, соответствующий номеру в колонке Кластер 1. Так, на первом шаге объединяются объекты 5 и 14, и кластеру присваивается номер 5, далее этот кластер на шаге 3 объединяется с элементом 4, и новому кластеру присваивается номер 4 и т. д. Следующая колонка “Коэффициент” содержит значение расстояния между кластерами, которые объединяются на данном шаге. Колонка “Этап” первого появления кластера показывает, на каком шаге до этого появлялся первый и второй из объединяемых кластеров. Последняя колонка “Следующий этап” показывает, на каком шаге снова появится кластер, образованный на этом шаге» [9, с. 310].

Шаги агломерации

Этап	Кластер объединен с		Коэффициенты	Этап первого появления кластера		Следующий этап
	Кластер 1	Кластер 2		Кластер 1	Кластер 2	
1	5	14	,439	0	0	3
2	2	12	,550	0	0	5
3	4	5	,671	0	1	6
4	7	13	,942	0	0	10
5	2	15	1,203	2	0	7
6	4	11	1,586	3	0	11
7	2	8	1,810	5	0	9
8	6	10	1,847	0	0	11
9	1	2	2,471	0	7	13
10	3	7	4,136	0	4	13
11	4	6	4,492	6	8	12
12	4	9	5,914	11	0	14
13	1	3	9,656	9	10	14
14	1	4	10,498	13	12	0

Рис. 8.4. Шаги агломерации в процедуре иерархического кластерного анализа

Источник: [9, с. 310].

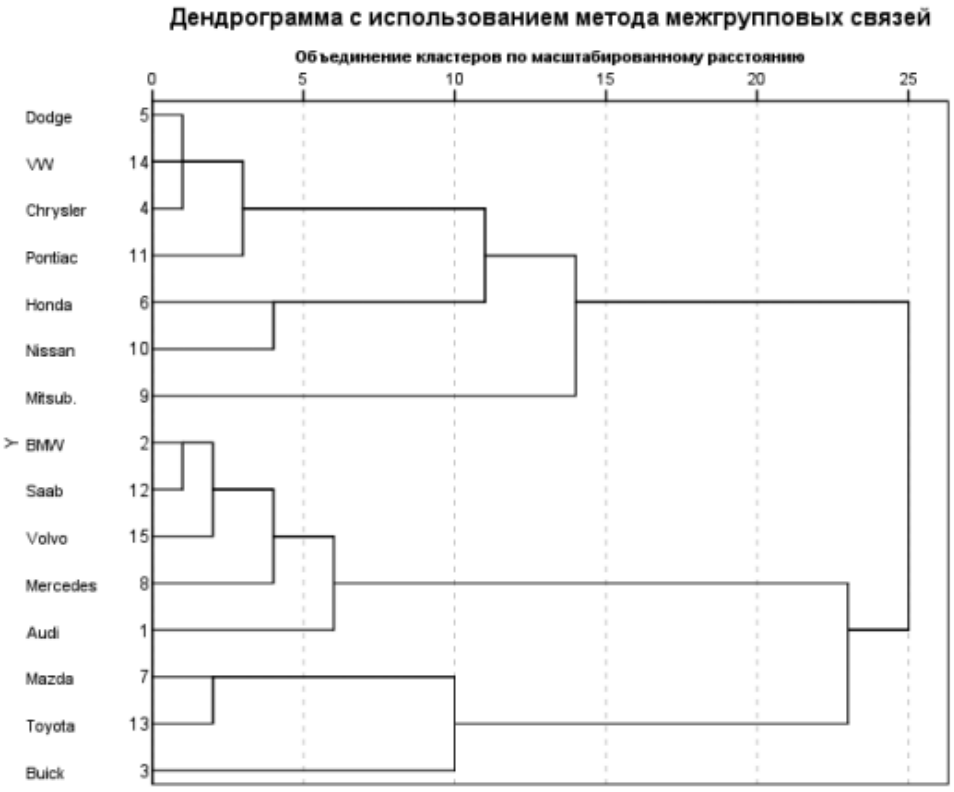


Рис. 8.5. Результат иерархического кластерного анализа, представленного в виде дендрограммы

Источник: [9, с. 310].

Дендрограмма представляет процесс кластеризации в форме древовидной структуры. Дендрограмма не только позволяет перейти к любому объекту на любом уровне кластеризации, но и дает возможность судить о том, каково расстояние между кластерами или объектами на каждом из уровней. Числа от 0 до 25 являются условной шкалой этих расстояний; 0 соответствует наименьшему расстоянию на первом этапе, а 25 — наибольшему на последнем.

5. *Интерпретация и профилирование кластеров* включает проверку кластерных центроидов[30]. Для этого, во-первых, информацию о принадлежности каждого наблюдения к определенному кластеру необходимо сохранить в но-

[30] Центроиды — средние значения объектов по каждой из переменных. Позволяют описывать кластеры.

вой переменной[31]. Во-вторых, выполнить анализ средних значений, описанный в главе 4. В результате будет выведена таблица кластерных центроидов, позволяющих содержательно описывать каждую группу. Например, в книге А. Бююля и П. Цёфеля приводится таблица, где независимыми переменными выступили вновь образованные кластеры, а зависимые переменные — тесты, по значениям которых и были сгруппированы наблюдения (рис. 8.6).

	Кластер 1	*Кластер 2*	*Кластер 3*	*Кластер 4*
Память на числа	10,00	10,00	4,20	4,80
Математические задачи	10,00	10,00	4,80	4,40
Находчиаость при прямом диалоге	9,00	4,25	10,00	4,00
Тест на составление алгоритмов	10,00	10,00	4,40	4,00
Уверенность во время выступления	10,00	4,75	10,00	4,20
Командный дух	9,50	4,50	4,40	10,00
Находчивость	9,25	3,75	10,00	4,40
Сотрудничество	9,75	4,25	4,00	10,00
Признание в коллективе	10,00	4,25	3,80	10,00
Сила убеждения	9,50	4,25	10,00	5,00

Рис. 8.6. Пример анализа средних значений
с использованием группировки объектов по кластерам

Источник: [2, с. 393].

Из этой таблицы мы видим, что люди, входящие в первый кластер, имеют высокие показатели во всех тестах. Во второй кластер включены те, кто имеет хорошие показатели по математическим тестам, но слабые оценки социальной компетентности и уверенности при выступлениях. В третий кластер вошли те, кто уверенно чувствует себя во время выступлений, но имеет слабые показатели в математических тестах и социальной компетентности. В четвертом кластере собраны люди с высоким уровнем социальной компетентности, но со слабыми результатами в тестах на решение математических задач и на силу убеждения.

6. *Оценка качества кластеризации* предполагает выполнение кластерного анализа одних и тех же данных с использованием различных способов изме-

[31] Опция «Сохранить» в окне «Иерархический кластерный анализ». Выбрать «Одно решение» или «Диапазон решений». Если поставить переключатель в «Одно решение» и указать в поле число 3, то получим новую переменную, значение которой будет равно 1, 2 или 3 в зависимости от того, какому кластеру будет принадлежать соответствующий объект в решении. Если же установить переключатель «Диапазон решений», в поле «Минимальное число кластеров» указать число 3, а в поле «Максимальное число кластеров» — число 5, это приведет к созданию трех новых переменных: первая будет принимать значения от 1 до 3, вторая — от 1 до 4, третья — от 1 до 5 [9, с. 307].

рения расстояния. Следует сравнить результаты, полученные на основе различных способов измерения, чтобы определить, насколько совпадают полученные результаты.

Кластерный анализ методом k-средних. Как уже было сказано выше, для больших выборок эффективнее будет использовать кластерный анализ методом k-средних, который позволяет разбить множество точек в многомерном пространстве на предварительно заданное число кластеров.

Процесс кластеризации основывается на минимизации расстояний между точками внутри всех кластеров и максимизации расстояний между кластерами. При этом каждый из них представляет собой группу точек, которые находятся как можно ближе к центру кластера, который называется центроидом. В процессе кластеризации методом k-средних сначала случайным образом выбираются центроиды кластеров, а затем каждая точка относится к ближайшему центроиду. Затем центроиды пересчитываются на основе точек, которые были отнесены к каждому кластеру, и процесс повторяется до тех пор, пока точки в кластерах не перестанут изменяться.

Следует также дополнить, что «метод кластерного анализа является разведочным, а это предполагает проведение нескольких процедур для поиска оптимальной модели. Верификация результатов исследования и определение оптимальности полученной модели осуществляются на уровне смысловых характеристик выделенных групп, а именно с точки зрения того, будут ли полезны эти кластеры в проводимом социологическом исследовании» [3, с. 135].

Алгоритм выполнения кластеризации k-средними несколько проще по сравнению с иерархическим кластерным анализом. Технически требуется указать количество заданных кластеров (анализ — классификация — кластеризация k-средними), увеличить число итераций до 100 (кнопка «Итерации»), сохранить принадлежность к кластеру и расстояние от центра кластера (кнопка «Сохранить») (рис. 8.7).

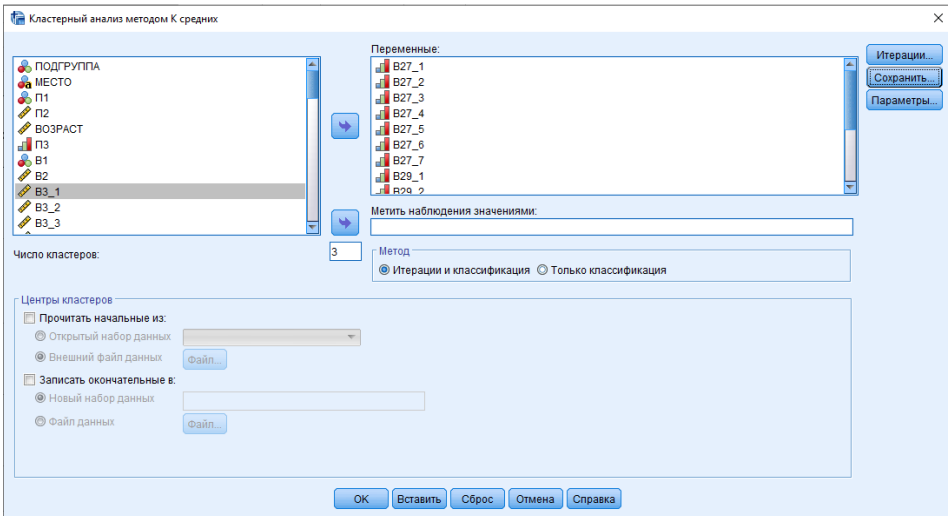


Рис. 8.7. Интерфейс команды кластеризации методом k-средних

В окне вывода будет информация о начальных и конечных центрах кластеров, количестве итераций, которые понадобились для выделения кластеров. Но главный результат — появление новых переменных в окне редактора данных, благодаря которым каждой единице анализа будет присвоен кластер с рассчитанным расстоянием до его центра. Именно с этими переменными в дальнейшем имеет дело исследователь. В частности, как и в иерархическом анализе эту переменную следует использовать в анализе средних в качестве фактора для построения типологии респондентов. И далее ее можно использовать во всем комплексе статистических процедур. Например, вычислить, какая доля респондентов проживает в городах, а какая — в селах, каков средний доход или уровень образования респондентов в каждом из кластеров.

§ 8.3. Пример кластерного анализа

В качестве примера воспользуемся тем же массивом о репродуктивных установках жительниц Калининградской области. Выполним как иерархический кластерный анализ, так и кластеризацию k-средними.

В каждом случае цель исследования — выявить типы респондентов на основе их отношения к государственным мерам поддержки рождаемости. Другими словами, с помощью кластерного анализа мы рассчитываем выявить типы репродуктивных установок респондентов.

Выполнив все процедуры иерархического кластерного анализа и создав новую переменную, содержащую значения о принадлежности к кластерам, мы получили три кластера (рис. 8.8).

Average Linkage (Between Groups)

		Частота	Проценты	Валидный процент	Накопленный процент
Валидные	1	100	12,0	36,9	36,9
	2	170	20,4	62,7	99,6
	3	1	,1	,4	100,0
	Всего	271	32,5	100,0	
Пропущенные	Системные	564	67,5		
Всего		835	100,0		

Рис. 8.8. Статистика трех исходных кластеров

Как мы видим, третий кластер состоит всего из одного наблюдения и не может быть использован в анализе. В качестве визуальной проверки кластеризации выведем дендрограмму, построенную уже не по наблюдениям, а по переменным (рис. 8.9).

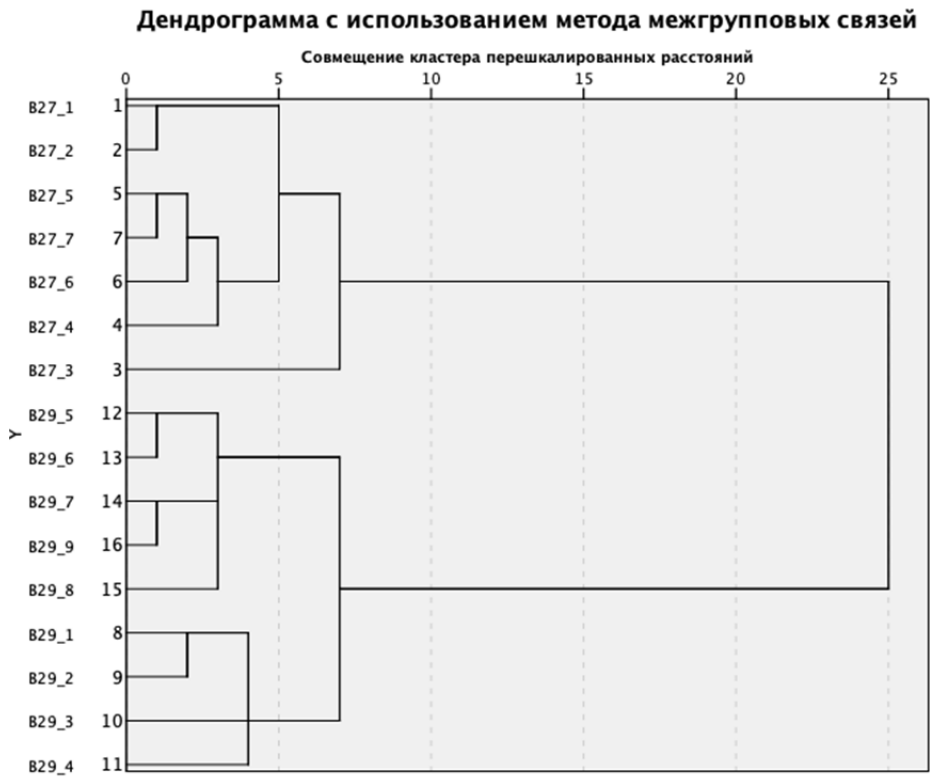


Рис. 8.9. Дендрограмма с использованием метода межгрупповых связей

Здесь мы видим довольно заметное выделение двух кластеров (по переменным B27 и B29). Это дает еще одно основание исключить из анализа третий кластер и работать только с первыми двумя.

На следующем этапе мы переходим к анализу средних, где в качестве независимых переменных указываем вновь созданную переменную по кластерам, а зависимых переменных — те, что отвечают целям исследования, в первую очередь отношение к мерам государственной поддержки (табл. 8.1).

Таблица 8.1

Анализ средних значений переменных в разрезе двух кластеров

Вид государственной поддержки	Кластер 1	Кластер 2
Пособие по беременности и родам	2,27	1,86
Единовременное пособие при рождении ребенка	2,22	1,86
Материнский капитал	2,79	2,06

Окончание табл. 8.1

Вид государственной поддержки	Кластер 1	Кластер 2
Частично оплачиваемый отпуск по уходу за ребенком длительностью 1,5 года	2,44	1,64
Ежемесячное пособие на ребенка в семьях с низкими доходами	2,17	1,44
Жилищные кредиты на льготных условиях молодым семьям	2,35	1,48
Компенсация затрат на оплату детских дошкольных учреждений	2,34	1,44
Региональный материнский капитал (в дополнение к федеральному)	4,31	2,21
Земельные участки под строительство (при рождении третьего и последующего ребенка)	4,34	2,32
Пособие на третьего ребенка до трех лет для семей, чей доход ниже среднего	4,11	2,01
Погашение кредита (для семей с двумя детьми — частичное, с тремя и более — полное)	4,53	2,04
Ясли или детский сад по выбору	4,34	1,94
Повышение компенсации на оплату детских дошкольных учреждений	4,35	1,72
Гибкий график работы для родителей	4,43	1,91
Оплата детского сада семьям, где работают оба родителя	4,56	1,71
Право на один дополнительный выходной в месяц	4,5	1,78

В таблице мы видим, что первый кластер имеет низкие оценки значимости первых семи параметров, направленных преимущественно на поддержку семей с одним ребенком, и высокую оценку остальных переменных. Второй кластер респондентов имеет низкие оценки всех предложенных мер государственной поддержки. Для уточнения типологии анализируем социально-демографические характеристики каждого показателя посредством анализа средних значений (рис. 8.10). Здесь мы наблюдаем, что а) респондентки из первого кластера моложе, б) как и у респонденток из второго кластера у них в среднем один ребенок, в) у респонденток из обоих классов сопоставимый уровень дохода. То, что значимо их отличает, так это намерение иметь больше детей как в идеале, так и исходя из обстоятельств. Такова типология респондентов согласно иерархическому кластерному анализу.

Average Linkage (Between Groups)		Количество полных лет	Укажите, пожалуйста, общее количество детей, которые у Вас есть	Каков средний уровень дохода на одного члена Вашей семьи в месяц?	Сколько еще детей Вам хотелось бы иметь «в идеале»?	Сколько еще детей Вы реально планируете иметь, исходя из обстоятельств?
1	Среднее	25,05	,82	3,48	2,32	1,08
	N	98	100	90	80	65
	Стандартная отклонения	7,085	,989	1,530	,612	1,020
2	Среднее	29,15	,98	3,66	1,81	,84
	N	170	170	154	73	83
	Стандартная отклонения	7,966	,765	1,145	,811	,757
Всего	Среднее	27,65	,92	3,59	2,08	,95
	N	268	270	244	153	148
	Стандартная отклонения	7,893	,857	1,300	,757	,887

Рис. 8.10. Статистика двух кластеров
по ряду социально-демографических характеристик

Теперь выполним кластеризацию методом k-средних. В данном случае зададим большее число кластеров (например, пять) на случай выделения более специфичных групп респондентов. В результате программа вычислила пять кластеров, но один из них снова состоит из одного наблюдения и подлежит исключению. Оставив заданными четыре кластера, мы получили примерно равную разбивку респондентов на все кластеры (рис. 8.11).

Номер кластера наблюдения

		Частота	Проценты	Валидный процент	Накопленный процент
Валидные	1	77	9,2	28,4	28,4
	2	83	9,9	30,6	59,0
	3	43	5,1	15,9	74,9
	4	68	8,1	25,1	100,0
	Всего	271	32,5	100,0	
Пропущенные	Системные	564	67,5		
Всего		835	100,0		

Рис. 8.11. Статистика кластеров, созданных методом k-средних

Далее также выполняем анализ средних по тем же переменным, что и в иерархическом кластерном анализе (таблица 8.2).

Таблица 8.2

Анализ средних значений переменных в разрезе четырех кластеров

Вид государственной поддержки	Кластер 1	Кластер 2	Кластер 3	Кластер 4
Пособие по беременности и родам	1,1	1,6	3,93	2,37
Единовременное пособие при рождении ребенка	1,19	1,54	3,56	2,44
Материнский капитал	1,64	1,73	4,14	2,68
Частично оплачиваемый отпуск по уходу за ребенком длительностью 1,5 года	1,32	1,39	3,93	2,03
Ежемесячное пособие на ребенка в семьях с низкими доходами	1,18	1,28	3,65	1,65
Жилищные кредиты на льготных условиях молодым семьям	1,42	1,31	3,63	1,74
Компенсация затрат на оплату детских дошкольных учреждений	1,26	1,29	3,93	1,62
Региональный материнский капитал (в дополнение к федеральному)	3,88	1,51	4,23	3,03
Земельные участки под строительство (при рождении третьего и последующего ребенка)	3,96	1,54	4,21	3,22
Пособие на третьего ребенка до трех лет для семей, чей доход ниже среднего	3,75	1,2	3,98	2,9
Погашение кредита (для семей с двумя детьми — частичное, с тремя и более — полное)	4,03	1,22	4,53	2,87
Ясли или детский сад по выбору	4,04	1,29	4,35	2,41
Повышение компенсации на оплату детских дошкольных учреждений	4,04	1,14	4,35	2,06
Гибкий график работы для родителей	4,31	1,22	4,35	2,25
Оплата детского сада семьям, где работают оба родителя	4,21	1,19	4,51	1,91
Право на один дополнительный выходной в месяц	4,36	1,12	4,35	2,01

Из таблицы 8.2 мы видим уже более детальные различия между респондентами. В частности, опрошенные из первого кластера крайне низко оценивают эффективность мер поддержки для одного ребенка и высоко все оставшиеся переменные. Респонденты из второго кластера продемонстрировали низкие оценки по всем предложенным параметрам. Респонденты из третьего кластера, напротив, выразили сравнительно высокие оценки по всем переменным. Наконец, опрошенные из четвертого кластера выразили умеренные оценки всем мерам поддержки семьи.

Рассмотрим их социально-демографические показатели и репродуктивные установки (рис. 8.12). Самым молодым и малочисленным кластером оказался третий: это девушки, как правило, еще не имеющие детей, при этом располагающие самым высоким уровнем ежемесячного дохода. Как и респонденты из первого кластера, они хотели бы иметь не менее двух детей и в целом ожидают завести именно столько, несмотря на все обстоятельства (самый высокий средний балл). У них самые высокие репродуктивные намерения. Наиболее возрастными респондентками стали опрошенные из второго кластера. У них уже есть один ребенок, они обладают более низким по сравнению с респондентками из предыдущего кластера уровнем дохода. Они также хотели бы иметь двоих детей, но демонстрируют самые низкие репродуктивные ожидания исходя из обстоятельств. Следовательно, это группа женщин, которые не могут реализовать свой репродуктивный потенциал. Наконец, опрошенные из первого и четвертого кластера близки по возрасту и количеству уже имеющихся детей, но отличаются по остальным параметрам. Так, девушки из первого имеют более низкий доход, но самое большое среди всех кластеров число детей в идеале. Однако исходя из обстоятельств больше готовы к рождению еще одного ребенка девушки из четвертого. Следовательно, респондентки из первого кластера более нуждаются в поддержке государства при рождении второго и последующего ребенка, чем и обусловлены их различия в оценках мер.

Номер кластера наблюдения		Количество полных лет	Укажите, пожалуйста, общее количество детей, которые у Вас есть	Каков средний уровень дохода на одного члена Вашей семьи в месяц?	Сколько еще детей Вам хотелось бы иметь «в идеале»?	Сколько еще детей Вы реально планируете иметь, исходя из обстоятельств?
1	Среднее	27,32	1,08	3,00	2,38	,84
	N	76	77	69	65	56
	Стандартная отклонения	7,443	,957	1,306	,630	,910
2	Среднее	31,01	1,10	3,62	1,82	,55
	N	83	83	74	33	42
	Стандартная отклонения	8,147	,775	1,167	,950	,670
3	Среднее	21,38	,30	4,13	2,13	1,62
	N	42	43	40	30	24
	Стандартная отклонения	4,690	,708	1,572	,507	,875
4	Среднее	27,72	,93	3,90	1,54	1,22
	N	68	68	62	26	27
	Стандартная отклонения	7,289	,739	,987	,647	,751
Всего	Среднее	27,63	,92	3,60	2,07	,95
	N	269	271	245	154	149
	Стандартная отклонения	7,884	,855	1,301	,759	,888

Рис. 8.12. Статистика четырех кластеров
по ряду социально-демографических показателей

Исходя из сравнительного анализа двух методов кластеризации, в данном случае предпочтительным является кластеризация k-средними, поскольку позволила выделить несколько групп респондентов. Методической рекомендацией здесь может послужить проведение кластеризации различными методами и выбор того, который более соответствует задачам исследования и здравому смыслу.

§ 8.4. Контрольные вопросы

1. Что общего и различного в факторном и кластерном анализе?
2. Какие виды кластерного анализа существуют? В чем их разница?
3. Каков алгоритм выполнения иерархического кластерного анализа в IBM SPSS Statistics?
4. Каков алгоритм выполнения кластерного анализа методом k-средних в IBM SPSS Statistics?

§ 8.5. Практические задания

1. Выполните процедуру кластерного анализа всеми тремя методами, используя переменные о доверии взрослого населения Калининградской области (старше 35 лет) общественно-политическим институтам и учреждениям. Сформулируйте исследовательский вопрос и дайте интерпретацию полученным результатам. Используйте массив «Социально-политические настроения в Калининградской области», n = 915, 2022 г.
2. Выполните процедуру кластерного анализа всеми тремя методами, используя переменные о доверии молодежи Калининградской области (18—35 лет) общественно-политическим институтам и учреждениям. Сформулируйте исследовательский вопрос и дайте интерпретацию полученным результатам. Используйте массив «Социально-политические настроения молодежи Калининградской области», n = 987, 2021 г.

§ 8.6. Рекомендуемая литература

1. *Беляева Л. А.* Социальные слои в россии: опыт кластерного анализа // Социологические исследования. 2005. № 12 (260). С. 57—64.
2. *Бююль А., Цёфель П.* SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей. СПб. : ДисСофтЮп, 2005.
3. *Воронин Г. Л.* Статистический анализ данных в IBM SPSS Statistics V27.0.1.0. Н. Новгород : ННГУ им. Н. И. Лобачевского, 2022.
4. *Воронин Г. Л.* Еще раз о «кластерах на факторах» // Социологический журнал. 2010. № 3. С. 21—34.
5. *Дерман Н. А.* Алкоголь и курение в Эстонии (кластерный анализ) // Социологические исследования. 2011. № 3 (323). С. 79—84.

6. *Выбор* процедуры кластеризации. URL: https://www.ibm.com/docs/ru/spss-statistics/saas?topic=features-choosing-procedure-clustering (дата обращения: 05.10.2023).

7. *Крыштановский А. О.* «Кластеры на факторах» — об одном распространенном заблуждении // Социология: 4М. 2005. № 21. С. 172—187.

8. *Крыштановский А. О.* Анализ социологических данных с помощью пакета SPSS. М. : ГУ ВШЭ, 2006.

9. *Наследов А.* SPSS 19: профессиональный статистический анализ данных. СПб. : Питер, 2011.

10. *Черныш М. Ф.* Опыт применения кластерного анализа // Социология: 4М. 2000. № 12. С. 129—141.

11. *Щемелева И. И.* Социальная активность студенческой молодежи: факторный и кластерный анализ // Социологические исследования. 2019. № 4. С. 133—141.

ЗАКЛЮЧЕНИЕ

В данном учебном пособии рассмотрены базовые и продвинутые методы статистического анализа данных в исследовательской деятельности социолога. Было показано, что каждый из них может быть использован для ответа на конкретные вопросы в области социологии.

Опыт преподавания учебного курса по работе в пакете IBM SPSS Statistics показывает, что у некоторых слушателей возникает потребность в более широком видении использования программы на практике. В связи с этим пособие содержит значительный материал, отсылающий к научным исследованиям, в которых были использованы те или иные методы статистического анализа данных. Более того, в каждой главе представлены примеры применения каждого метода на реальных данных, в которых продемонстрирован весь алгоритм работы исследователя от постановки проблемы до интерпретации обнаруженных связей. По сути, каждая глава представляет собой самостоятельное исследование, поскольку содержит не только описание метода в теории, но и анализ эмпирических данных на практике. Важной частью знакомства с методами был анализ непараметрических критериев, а также условий и ограничений, связанных с применением каждой статистической процедуры.

В заключение подчеркнем, что статистические методы в социологии всегда будут оставаться необходимым инструментом для получения объективной информации о количественных данных. Они позволяют выявлять закономерности и зависимости, которые не являются очевидными на первый взгляд. Следует также помнить, что статистический анализ является не целью сам по себе, а лишь инструментом для проверки исследовательских гипотез. В связи с этим описанные в пособии методы развивают компетенции, полученные после прохождения таких учебных курсов, как «Основы социологии» и «Методология социологического исследования». Соответственно, данное пособие встраивается в образовательный комплекс подготовки будущих социологов.

Надеемся, что учебное пособие «Статистический анализ данных в исследовательской деятельности социолога» будет полезным не только для студентов, преподавателей, исследователей, но и всех, кто интересуется статистическими методами в социологии.

Учебное издание

Щекотуров Александр Вячеславович

СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ
В ИССЛЕДОВАТЕЛЬСКОЙ ДЕЯТЕЛЬНОСТИ СОЦИОЛОГА

Учебное пособие

Корректор *О. И. Бессчастнова*
Компьютерная верстка *Г. И. Винокуровой*

Подписано в печать 27.11.2023 г.
Дата выхода в свет 14.12.2023 г.
Формат 70×100 $^{1}/_{16}$. Усл. печ. л. 9,3
Тираж 300 экз. (1-й завод 75 экз.). Заказ 126

Издательство Балтийского федерального университета им. И. Канта
236041, г. Калининград, ул. А. Невского, 14