МИНИСТЕРСТВО ОБРАЗОВАНИЯ
И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ
АРХИТЕКТУРНО-СТРОИТЕЛЬНЫЙ УНИВЕРСИТЕТ (СИБСТРИН)

**О.А. Добрина**

# АНАЛИЗ ДАННЫХ В СОЦИОЛОГИИ

Учебное пособие

НОВОСИБИРСК 2013

УДК 316
ББК 60.5
Д 559

Добрина О. А.
Анализ данных в социологии : учеб. пособие / О. А. Добрина ; Новосиб. гос. архитектур.-строит. ун-т (Сибстрин). – Новосибирск : НГАСУ (Сибстрин), 2013. – 100 с.

**ISBN 978-5-7795-0666-3**

Пособие раскрывает сущностные характеристики анализа данных как завершающего этапа социологического исследования, логическую модель анализа, этапы проведения. Для закрепления материала в конце каждой темы предлагаются вопросы для самопроверки.

Учебное пособие предназначено для студентов, обучающихся по направлению 040100.62 «Социология». Пособие может быть использовано преподавателями в образовательном процессе и теми, кто интересуется методикой социальных исследований.

Печатается по решению издательско-библиотечного совета НГАСУ (Сибстрин)

Рецензенты:

- П.А. Кулаков, канд. ист. наук, доцент кафедры социологии, педагогики и психологии НГАСУ (Сибстрин);
- С.А. Ильиных, д-р социол. наук, профессор кафедры социальных коммуникаций и социологии управления НГУЭУ

**ISBN 978-5-7795-0666-3**

## ОГЛАВЛЕНИЕ

## ВВЕДЕНИЕ

Курс является составной частью программы подготовки студентов по направлению 040100.62 «Социология» и предназначен для студентов III курса ФЭМГО НГАСУ (Сибстрин).

**Цели изучения дисциплины**

Формирование целостного представления о процессе получения эмпирического знания в социологии, последовательно и доказательно выстроенного на основе работы с эмпирическими данными.

В силу специфики своего предмета анализ данных в социологии относится к числу базовых дисциплин в социологическом образовании, владение материалом которых необходимо социологу вне зависимости от его специализации.

Данный курс читается на третьем году обучения студентов-социологов, поскольку предполагает знание методики конкретных социологических исследований и теории, а также методики измерения в социологии.

**Задачи изучения дисциплины**

1. Введение оснований логической и математической формализации.
2. Освоение отдельных приемов измерения в эмпирической социологии.
3. Освоение различных приемов построения логики анализа социологических данных.
4. Обучение технике работы с современными программными системами анализа социологических данных на персональных компьютерах.

**Место дисциплины в учебном процессе**

Учебный курс «Анализ данных в социологии» конкретизирует и закрепляет знания, полученные студентами ранее при изучении таких учебных дисциплин, как «Методы прикладной

статистики для социологов», «Методология и методы социологических исследований», «Методы измерений в социологии». Этот курс служит некоторым основанием для последующих дисциплин «Социальная статистика», «Основы социального прогнозирования и управления».

Учебный курс содержит 44 лекционных часа и 16 часов практических занятий, текущие контрольные мероприятия, самостоятельную работу и сдачу зачета.

**Процесс изучения дисциплины направлен на формирование следующих компетенций:**

- способность к восприятию, обобщению, анализу информации, постановке цели и выбору путей ее достижения (ОК-1);
- способность использовать основные законы естественнонаучных дисциплин в профессиональной деятельности, применять методы математического анализа и моделирования, теоретического и экспериментального исследования (ОК-11);
- владение основными методами, способами и средствами получения, хранения, переработки информации, навыки работы с компьютером как средством управления информацией (ОК-13);
- способность и готовность участвовать в составлении и оформлении научно-технической документации, научных отчетов, представлять результаты исследовательской работы с учетом особенностей потенциальной аудитории (ПК-3);
- умение обрабатывать и анализировать данные для подготовки аналитических решений, экспертных заключений и рекомендаций (ПК-8);
- способность использовать базовые теоретические знания, практические навыки и умения для участия в научных и научно-прикладных исследованиях, аналитической и консалтинговой деятельности (ПК-10);

– способность использовать методы сбора, обработки и интерпретации комплексной социальной информации для решения организационно-управленческих задач, в том числе находящихся за пределами непосредственной сферы деятельности (ПК-11).

**В результате изучения дисциплины обучающийся должен:**

– знать основные виды и принципы анализа данных в социологии;
– уметь выбирать вид анализа данных, релевантный поставленным исследовательским задачам;
– владеть компьютерными технологиями анализа данных в социологии.

В процессе освоения дисциплины «Анализ данных в социологии» используются следующие **образовательные технологии.**

*Стандартные методы обучения:*

– лекционные занятия;
– практические (семинарские) занятия;
– самостоятельная работа студентов.

В ходе лекционных занятий раскрываются основные вопросы в рамках рассматриваемой темы, делаются акценты на наиболее сложные и важные положения изучаемого материала, которые должны быть приняты студентами во внимание. Материалы лекций являются основой для подготовки студентов к практическим (семинарским) занятиям и выполнения заданий самостоятельной работы.

Целью практических (семинарских) занятий является контроль за степенью усвоения пройденного материала, ходом выполнения студентами самостоятельной работы и рассмотрение наиболее сложных и спорных вопросов в рамках темы занятия.

Самостоятельная работа студентов включает подготовку к практическим (семинарским) занятиям в соответствии с вопросами, представленными в рабочей программе дисциплины, выполнение заданий для самостоятельной работы студентов, ре-

шение тестов. Отдельные задания для самостоятельной работы предусматривают представление доклада и/или презентации и обсуждение полученных результатов на практических (семинарских) занятиях. Работа выполняется с использованием текстового редактора MS Word, MS Ecxel – для текстов, таблиц и диаграмм, MS Power Point – для подготовки слайдов и презентаций.

При необходимости в процессе работы над заданием студент может получить индивидуальную консультацию у преподавателя. Также предусмотрено проведение консультаций студентов в ходе изучения материала дисциплины в течение семестра.

*Методы обучения с применением интерактивных форм образовательных технологий:*

- лекции-консультации и интерактивные лекции с применением мультимедийного оборудования;
- эвристические беседы;
- учебные дискуссии в виде «займи позицию», «один – вдвоем – все вместе», дебатов и т.д.;
- брифинги;
- творческие задания в форме изложения проблемного материала;
- групповые оценки и взаимооценки, а именно: рецензирование студентами друг друга, оппонирование студентами докладов и аналитических и исследовательских работ;
- презентации отдельных тем в частичном разрезе их содержания с последующим обсуждением;
- конкурсы творческих заданий и исследований с их дальнейшим обсуждением, позволяющим оценить уровень приобретенных знаний, умений и сформированных компетенций обучающихся.

# Глава 1. ОДНОМЕРНЫЙ ОПИСАТЕЛЬНЫЙ АНАЛИЗ



## 1.1. Основные понятия в анализе данных

Анализ данных является ключевым этапом всего исследования, в ходе которого происходит непосредственная проверка (математическими методами) соответствия собранной информации тем моделям социальных явлений, которые, явно или скрыто, имеются у социологов [12].

В случае простой визуализации собранной информации имеется дело лишь с *обработкой* социологических данных.

Если же ставятся задачи построения определенной модели изучаемого социального явления и проверки соответствия этой модели имеющимся данным, можно говорить именно об *анализе данных*.

При работе с социологическими данными используются два основополагающих понятия: единица анализа (анкета, случай); переменная. *Единица анализа* – это элементарная, единичная часть объекта исследования. В большинстве случаев единица анализа совпадает с единицей наблюдения, т.е. с тем объектом, о котором непосредственно получают информацию в ходе сбора данных. *Переменная* – это элементарный показатель, признак, характеризующий одно из изучаемых свойств единицы анализа. Например, переменными являются пол или зарплата респондента.

Ключевыми характеристиками переменной является то, что, с одной стороны, для каждой единицы анализа она имеет одно, вполне определенное значение, а, с другой стороны, то,

что не все единицы анализа имеют одинаковое значение переменой.

Простые количественные методы анализа данных делятся на три основных типа:

1. *Одномерный описательный анализ.* Раскрывает некоторые характеристики частотных распределений.
2. *Двумерный описательный анализ.* Связан с описанием взаимосвязи между переменными, а также со сравнением значений некоторой переменной в разных социальных группах.
3. *Объяснительный анализ.* Направлен на выявление силы влияния переменных друг на друга.

Задача одномерного описательного анализа – сжать полученную информацию, компактно представить ее для дальнейшего осмысления.

### 1.2. Измерения для номинальных переменных

Для различных уровней измерений подходят различные способы исчислений средней тенденции и дисперсии. Например, «тип занятий» – номинальная переменная. Начнем изучение этих исчислений с рассмотрения статистических процедур, подходящих для номинального уровня измерения. На этом уровне, где цифры просто обозначают категории безотносительно к порядку их расположения, единственно возможный способ измерения средней тенденции – это исчисление моды.

Рассмотрим пример, приведенный Дж.Б. Мангеймом, Р.К. Ричем [7, с. 395–396]. Мы задали ста респондентам вопрос об их занятии в настоящее время и затем распределили их ответы по типам. Тогда частотное распределение для переменной «тип занятий» может выглядеть так, как это показано в табл. 1.1.

Таблица 1.1

Частотное распределение: типы занятий респондентов

| Код | Значение | Число случаев |
|---|---|---|
| 1 | «Синие воротнички» | 25 |
| 2 | «Белые воротнички» | 23 |
| 3 | Специалисты | 22 |
| 4 | Фермеры | 20 |
| 5 | Безработные | 10 |

В частотном распределении исследователь просто перечисляет все значения переменной и показывает, сколько имеется случаев каждого значения. Та же самая информация может быть представлена в виде гистограммы, как показано на рис. 1.1. Используя эту информацию, можно выделить наиболее типичный случай и определить его репрезентативность.

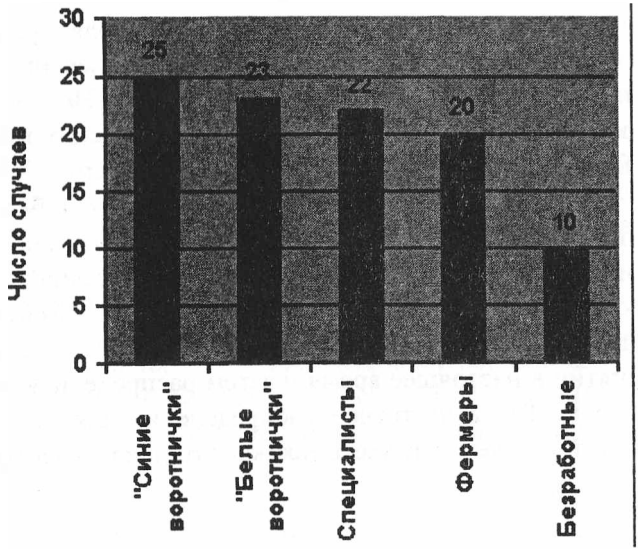



Рис. 1.1. Гистограмма: тип занятий респондентов

**Мода** – это просто наиболее часто встречающееся значение признака, т.е. то значение, которое наиболее часто может встречаться в серии зарегистрированных наблюдений. В нашем случае это первая категория, или градация «синие воротнички». Можно назвать их как модой, так и модальной категорией. (Распределенное, в котором две категории имеются с наибольшим количеством случаев, называется распределением с двумя модами, или *бимодальным,* возможно также распределение с большим количеством таких категорий.) Таким образом, занятие уровня «синих воротничков» являются наиболее типичными в нашем примере из ста человек.

Однако ясно, что большинство людей в этом примере (фактически 75 %) не являются рабочими – «синими воротничками», т.е. даже если мы выделим наиболее типичное значение в данном распределении, информация эта не обязательно полностью верно отражает картину. Более точно об этом можно судить, если подсчитать точное значение дисперсии для номинального уровня измерений, или **коэффициент вариации**, формула которого выглядит следующим образом:

$$v = \frac{f_{немодальное}}{N}$$

или

$$v = 1 - \frac{f_{модальное}}{N},$$

где $f_{немодальное}$ – сумма всех случаев, не входящих в модальную категорию;

$f_{модальное}$ – количество случаев в модальной категории;

$N$ – общее число случаев.

По сути дела, этот коэффициент дает *процентную долю всех признаков, которые не входят в модальную категорию.* В нашем примере

$$v = \frac{23 + 22 + 20 + 10}{100} = 0,75$$

или по упрощенной формуле

$$v = 1 - \frac{25}{100} = 0,75.$$

Значение коэффициента вариации колеблется между 0 (когда все случаи принимают одно и то же значение) и 1 – 1/*N* (когда каждый случай имеет свое значение). В общем, чем меньше коэффициент вариации, тем типичнее, или значимее (верно отражает картину), мода. В случае бимодального или многомодального распределения произвольно выбирается одно модальное значение в зависимости от целей подсчетов, и *v* определяется так, как указано выше.

### 1.3. Измерения для порядковых переменных

Когда мы имеем дело с данными порядкового уровня, у нас несколько больше информации, поскольку коды представляют не только категоризацию, но и относительные позиции, или ранжирование. Выбор способа измерения средней тенденции и дисперсии должен как отражать этот факт, так и использовать его возможности. Наиболее подходящий способ измерения средней тенденции для порядковых данных – медиана.

**Медиана** – это просто значение среднего признака в упорядоченном ряду, признака, до и после которого находится равное количество признаков. Вычисление медианы, таким образом, требует лишь того, чтобы отсчитать с обоих концов частотного распределения равное количество признаков, до тех пор пока не доберемся до срединного, и определить затем его значение. Там, где имеется нечетное количество признаков, можно определить единственный срединный признак (например, для 99 признаков 50-я от любого конца частотного распределения единица будет иметь 49 единиц как до, так и после себя). Значение этого признака и будет медианой. Если же *N* (количество единиц) – чет-

ное число, появятся две срединных единицы (например, для 100 единиц 50-я и 51-я вместе составят середину распределения). Если обе эти единицы имеют одно и то же значение, оно и будет медианой. Если у них разные значения, медианой будет среднее арифметическое между ними. Поясним на примере. Давайте рассмотрим распределение уровней образования по трем массивам данных (табл. 1.2).

Таблица 1.2

Уровни образования по трем массивам

| Код | Значение | Массив 1 (*N*) | Массив 2 (*N*) | Массив 3 (*N*) |
|---|---|---|---|---|
| 1 | Начальная школа | 25 | 25 | 10 |
| 2 | Незаконченное среднее | 23 | 23 | 40 |
| 3 | Законченное среднее | 22 | 22 | 35 |
| 4 | Высшее | 20 | 20 | 10 |
| 5 | Наличие ученой степени | 9 | 10 | 5 |
| Общее количество | | 99 | 100 | 100 |

В первом массиве выделяется один срединный случай (50-й с обоих концов), определяется его значение и выясняется, таким образом, что медианный уровень образования – 3, или «законченное среднее». Во втором массиве выделяется два срединных случая (50-й и 51-й с обоих концов), определяется, что каждый принимает одно и то же значение и выясняется, что медиана – опять 3. В третьем же массиве срединные случаи включают две категории – «незаконченное среднее» и «законченное среднее». Здесь медианой является среднее арифметическое между этими величинами, т.е. (2+3)/2 = 2,5. Поскольку дробные значения не

имеют смысла в порядковом измерении, эта цифра просто говорит нам, что середина распределения лежит *примерно* между 2 и 3.

Любой из нескольких способов измерения дисперсии для порядковых переменных, называемый **квантильным рангом**, показывает, насколько плотно различные значения группируются вокруг медианы, или опять насколько типична или репрезентативна медиана для распределения в целом. **Квантиль** – это мера положения внутри распределения. Например, персентиль (или процентиль) делит совокупность на 100 равных частей так, что первый персентиль – это такая точка или значение в этой совокупности (считая от меньшего значения вверх), ниже которой находится 1 % всех случаев, второй персентиль – такая точка или значение, ниже которой находятся 2 % всех признаков, и т.д. Или, используя более знакомый пример, будущий студент колледжа, достигший 85-го персентиля в тесте на эрудицию, дошел до уровня более высокого, чем уровни 85 % всех, кто проходил тест. Точно так же дециль делит совокупность на десятки (например, третий дециль – это точка, ниже которой находятся 30 % случаев), квантиль – на пятые доли, квартиль – на четвертые. Любой из них может быть использован для определения дисперсии вокруг медианы, хотя децильные и квантильные ранги наиболее часто встречаются в литературе.

Давайте проиллюстрируем эту процедуру на примере квантильных рангов. Квантильный ранг ($q$) определяется следующим образом:

$$q = q_4 - q_1,$$

где $q_4$ – четвертый квантиль (значение, ниже которого находится 4/5, или 80 % всех признаков);

$q_1$ – первый квантиль (значение, ниже которого находится 1/5 или 20 % всех признаков).

Чем меньше степень разброса величин между этими двумя точками совокупности, тем плотнее сгруппированы случаи вокруг медианы и тем точнее представляет медиана всю совокупность.

В массиве 2 табл. 1.2, например, где $N = 100$, можно подсчитать $q$, определив 81 признак (ниже которого расположено 80 % признаков) и 21 признак (ниже которого расположены 20 % признаков), начиная счет внутри частотного распределения с наименьших значений. Затем вычитаем значение 21-го признака из значения 81-го ($q = q_4 - q_1 = 4 - 1 = 3$) и получаем квантильный ранг.

Одна из трудностей интерпретации квантильных рангов состоит в том, что они чрезвычайно чувствительны к изменениям в количестве градаций самой переменной. Чем больше градаций, тем вероятнее большой разброс. Поэтому квантильные ранги не всегда поддаются интерпретации в случаях сравнений переменных с разным количеством градаций. Для переменных же с примерно равным количеством градаций для построчного или постолбцового сравнения значений одной переменной или для какого-либо абсолютного измерения разброса вокруг медианы они вполне подходят.

### 1.4. Измерения для интервальных переменных

Интервальные данные, безусловно, предоставляют наиболее полную информацию, включая категоризацию, ранжирование и установление интервалов. Интервальные значения могут быть подвержены любым арифметическим манипуляциям. Следовательно, приступая к исчислению средней тенденции и дисперсии для интервальных данных, мы можем и должны принять эту информацию о дополнительных возможностях во внимание.

Для количественных переменных самой важной и распространенной является мера центральной тенденции – *среднее*

*арифметическое,* которое чаще всего называют просто *средним* (и обозначают как $\overline{X}$). Процедура определения среднего общеизвестна: нужно просуммировать все значения наблюдений и разделить полученную сумму на число наблюдений. В общем случае

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n},$$

т.е.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

где $X_1 \ldots X_i$ – наблюдаемые значения;

$n$ – число наблюдений;

$\sum$ – знак арифметической суммы.

Очевидно, важно не только знать, что типично для выборки наблюдений, но и установить, насколько выражены отклонения от типичных значений. Чтобы определить, насколько хорошо та или иная мера центральной тенденции описывает распределение, нужно воспользоваться какой-либо мерой изменчивости, разброса.

Мерой разброса признаков для данного типа шкал является дисперсия, равная сумме квадратов остатков, деленной на количество наблюдений минус 1.

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - x_{cp})^2}{n-1}.$$

Дисперсия имеет недостаток – большое значение, не дающее представления – хорошо это или плохо. Это связано с тем, что остатки в дисперсии берутся в квадрате. Поэтому чаще ис-

пользуется в качестве меры разброса признаков среднее квадратичное, или стандартное, отклонение.

Введем понятие стандартного отклонения. Стандартное отклонение (*s*) является тем математическим инструментом, который может помочь выполнить вашу задачу. По сути дела, это процедура, которая сводит на нет свойства разнонаправленных интервалов уравновешивать друг друга путем простого возведения в квадрат утих интервалов (и избавляясь таким образом от отрицательных значений), измерения разброса *квадратов* интервалов вокруг среднего арифметического и затем извлечения из результата квадратного корня, с тем чтобы вернуться к начальным единицам интервалов. Формула, по которой все это вычисляется, напоминает прежнюю, кроме возведения в квадрат и извлечения квадратного корня. Формула эта такова:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - x_{cp})^2}{n-1}}.$$

где $x_i$ – значение каждого отдельного случая;

$x_{cp}$ – среднее арифметическое;

$n$ – количество случаев;

$\sum_{i=1}^{n}$ – знак суммы всех отдельных случаев от 1 до $n$.

Эта мера выражена в тех же единицах, что и исходные данные. Если переменные измеряются в одних и тех же единицах, то стандартное отклонение может быть основой для выяснения репрезентативности средних геометрических; чем больше стандартное отклонение, тем более репрезентативно ее среднее геометрическое. Но если единицы принципиально отличны или если анализируется одна переменная, интерпретация стандартного отклонения уже не столь проста.

Существует одно исключение из этого: переменные, чье распределение близко к *нормальному,* т.е. такие, у которых существует единственная мода в самом центре распределения, а частоты симметрично убывают по направлениям к предельным значениям (графическое изображение нормального распределения, с которым вы, наверно, хорошо знакомы, – это просто колоколообразная кривая). Известно (из рассуждений, которые не входят в рамки нашего разговора), что в таких случаях 68,3 % всех случаев лежат в пределах одного стандартного отклонения, от среднего геометрического ( $\overline{X} \pm s$), 95,5 % – в пределах двух стандартных отклонений от среднего геометрического ( $\overline{X} \pm 2s$) и 99,7 % – в пределах трех стандартных отклонений от среднего геометрического ( $\overline{X} \pm 3s$). Фактически в случае таких распределений для любой точки можно определить, на сколько стандартных отклонений ниже или выше среднего геометрического она находится, и затем использовать эту информацию для выяснения относительного положения двух признаков в одной переменной или, наоборот, относительного значения двух переменных для одного и того же признака. Позволяет это сделать *стандартная оценка* (*z*) которая вычисляется по следующей формуле:

$$z = \frac{(x_i - x_{cp})}{S}.$$

Представьте, что мы располагаем данными, например, по затратам на образование на душу населения в каждом регионе, количеству работающих преподавателей на 1000 студентов в каждом регионе и количеству награжденных выпускников средней школы на 100 тыс. населения в каждом регионе в определенном году и что значения этих переменных по регионам распределяются по кривой, близкой к нормальной. Представьте затем, что мы хотим использовать эти данные для изучения политики в области образования в двух регионах. Сначала подсчита-

ем среднее арифметическое ( $\overline{X}$ ) и стандартное отклонение ($s$) для каждой переменной по всем регионам, затем определим соответствующие стандартные оценки ($z$) для каждой переменной по двум нужным нам регионам. Результатом будут два набора значений в стандартных единицах (уже не в рублях или долларах, количестве учителей и документов, а в количестве стандартных отклонений от среднего геометрического), которые могут быть использованы для определения индексов политики в области образования, для выяснения относительной позиции двух регионов среди других или для стандартизации при необходимости сравнения принципиально отличных измерений. Таким образом, при использовании стандартного подсчета стандартное отклонение может оказаться очень полезным.

**Вопросы для самопроверки**

1. Чем отличается обработка данных от их анализа?
2. Каковы основные виды анализа данных в социологии?
3. Каковы основные измерения для номинальных шкал?
4. Каковы основные измерения для порядковых шкал?
5. Каковы основные измерения для интервальных шкал?

## Глава 2. ВЗАИМОСВЯЗЬ ПЕРЕМЕННЫХ



### 2.1. Основные понятия взаимосвязи переменных

Предыдущий вид обработки данных, как правило, является первым, исходным этапом анализа собранной информации – это анализ информации по каждой из переменных.

Вместе с тем, интересным представляется и одновременный анализ значений более одной переменной – это анализ, подразумевающий изучение взаимосвязи переменных.

Здесь моделью, которая предполагает свою проверку, является модель типа: «*социальные группы с разным уровнем образования (дохода, места жительства и т.п.) отличаются по характеру проведения досуга (политическим предпочтениям, степени удовлетворения жизнью и т.п.)*». То есть допускается, что существует переменная, которая объясняет поведение других переменных. Первая является основанием, вторая – следствием. Такие объясняющие переменные (причины) называются независимыми, а объясняемые переменные – зависимыми.

В том случае, если знание значений одной переменной по определенному случаю позволяет сделать некоторые предположения относительно соответствующих значений другой переменной, между этими переменными существует *связь*. Если, например, мы исследуем взаимосвязь между численностью населения какой-либо страны и долей взрослых, получивших высшее образование (принимая во внимание, что мы располагаем такими данными), то возможны три варианта:

1) более крупные страны обычно имеют большую долю взрослых, получивших высшее образование, чем менее крупные;
2) малые страны обычно имеют большую долю взрослых, получивших высшее образование, чем более крупные;
3) систематических различий нет; некоторые страны из обеих групп имеют относительно высокую долю таких людей, а другие – тоже из обеих групп – относительно низкую [7].

Если исследование покажет, что верен случай 1 или случай 2, то можно использовать знание значений независимой переменной – *количество населения,* – для того чтобы примерно представить или предсказать значения зависимой переменной – *доля взрослых, получивших высшее образование* – для любой из взятых стран. В первом случае для густонаселенных стран можно предсказать и относительно высокую долю взрослых с высшим образованием, а для малонаселенных стран – более низкую их долю. Во втором случае наши предположения будут прямо противоположны. В обоих случаях, хотя мы можем и не угадать каждый случай точно, мы будем чаще всего правы, поскольку между этими переменными существует связь. И конечно, чем теснее связь между двумя переменными, тем более вероятно, что наши догадки в каждом отдельном случае будут верны. Если существует полная зависимость значений одной переменной от значений другой, т.е. высокие значения одной переменной вызывают высокие значения другой или, наоборот, высокие значения одной вызывают низкие значения другой, можно вывести одну из другой с довольно большой степенью точности. Все это в корне отличается от третьего случая, который не позволяет с достаточной долей точности предугадать значения переменной *образование,* основываясь на знании количества населения. Если признаки по двум переменным распределяются, по сути дела, произвольно, то считается, что эти переменные не имеют связи.

Понятно, что между переменными может существовать более или менее сильная связь. Естественно, возникает вопрос, насколько сильна эта связь. На помощь приходит статистика. Из статистики возьмем показатель, который называется *коэффициентом связи.* Коэффициент связи – это показатель, который обозначает степень возможности определения значений одной переменной для любого случая, базируясь на значении другой. В нашем примере этот коэффициент может показать, *насколько* знание количества населения страны поможет в определении доли взрослых, получивших высшее образование. Чем больше коэффициент, тем сильнее связь и, следовательно, выше наши возможности прогноза. Вообще коэффициент колеблется в переделах от 0 до 1 или от –1 до 1, где значения, близкие к единице, обозначают относительно сильную связь, а значения, близкие к 0, – относительно слабую. Как было в случае с одномерной статистикой – и по тем же причинам, – каждый уровень измерения требует своего типа исчислений, и поэтому каждый из них требует своего способа измерения связи.

В дополнение к величине связи полезно также знать направление или форму взаимоотношений между двумя переменными. Еще раз обратите внимание на вышеприведенный пример, особенно на варианты 1 и 2. Можно предположить, что, чем теснее связаны признаки, тем больше будет коэффициент связи и тем выше шансы угадать долю взрослых с высшим образованием на основании знаний о количестве населения в данной стране. Очевидно, что наши прогнозы относительно каждого случая будут совершенно противоположны. В первом случае большие значения одной переменной вероятнее всего связаны с большими значениями другой, тогда как во втором случае большие значения одной переменной вероятнее всего связаны с меньшими значениями другой. Такие связи называются связями, имеющими разное *направление.* А такой тип связей, как в первом случае, когда обе переменные возрастают и убывают одновременно, называется *прямой,* или *положительной,* связью. Тип связей второго случая, когда значения постоянно изменя-

ются в разных направлениях, называется *обратной,* или *отрицательной,* связью. Эта добавочная информация – о знаке (плюс или минус) перед коэффициентом связи – способна сделать наши предположения более эффективными. Таким образом, коэффициент, равный –0,87 (отрицательный и близкий к единице), может описывать относительно сильную взаимосвязь, в которой значения двух данных переменных обратно связаны (изменяются в разных направлениях), коэффициент же, равный 0,10 (положительный – знак «плюс» обычно опускают – и близкий скорее к 0), может описывать слабую прямую связь.

Для всех случаев понятие направления или формы имеет разный смысл для разных уровней измерения. На номинальном уровне, где цифры играют роль просто обозначений, концепция направления вообще не имеет смысла и, соответственно, номинальные коэффициенты связи не изменяют знака. Все они положительны и просто показывают силу связи. На интервальном же уровне, наоборот, знаки могут не только изменяться, но и иметь достаточно сложную геометрическую интерпретацию. Проверка на связь на этом уровне измерений обладает очень высокими прогностическими способностями, причем знак коэффициента является в этом случае ключевым элементом.

Чтобы выбрать правильную статистику для этого случая, вам необходимо придерживаться простого правила: использовать статистику, разработанную для низшего уровня измерений, не игнорируя при этом данные для измерений высококачественного уровня. Вполне законно можно применять статистику для номинальных признаков с одноуровневыми данными, но совершенно невозможно использовать одноуровневую статистику для номинальных измерений. Это означает, что, когда проводится сравнение переменных, которые измеряются на разных уровнях, необходимо так выбирать статистический критерий, чтобы он соответствовал нижнему из двух уровней.

Функцию меры качества модели взаимосвязи переменных выполняют *коэффициенты связи.*

То есть в отличие от предыдущего опыта проведения перекрестного анализа, при более грамотном и продуктивном подходе к анализу необходимо учитывать эти коэффициенты связи.

Принцип един – чем выше коэффициент, тем больше взаимосвязь, тем выше качество модели.

## 2.2. Двумерные таблицы

К наиболее часто используемым инструментам изучения взаимосвязи двух переменных относятся методы анализа таблицы сопряженности.

Проанализируем пример, приведенный в работе И.Ф. Девятко [2]. Пусть мы располагаем совокупностью данных о занятиях физзарядкой и образовании для выборки горожан. Для простоты предположим, что обе переменные имеют лишь два уровня: высокий и низкий. Так как данные об образовании исходно разбиты на большее количество категорий, придется их перегруппировать, разбив весь диапазон значений на два класса. Предположим, мы выберем в качестве граничного значения 10 лет обучения, так что люди, получившие неполное среднее и среднее образование, попадут в «низкую» градацию, а остальные – в «высокую». Для занятий физическими упражнениями мы соответственно воспользуемся двумя категориями – «делают зарядку» и «не делают зарядку». В табл. 2.1 показано, как могло бы выглядеть совместное распределение этих двух переменных.

Таблица 2.1

Взаимосвязь между уровнем образования и занятиями физкультурой

| Занятия физкультурой | Уровень образования | | Всего |
|---|---|---|---|
| | Низкий | Высокий | |
| Делают зарядку | 50 | 200 | 250 |
| Не делают зарядку | 205 | 45 | 250 |
| Всего | 255 | 245 | 500 |

При анализе, например, табл. 2.1 можно руководствоваться простыми правилами. Во-первых, нужно определить независимую переменную и, в соответствии с принятым определением, пересчитать абсолютные частоты в проценты. Если независимая переменная расположена по горизонтали таблицы, считаем проценты по столбцу; если независимая переменная расположена по вертикали, проценты берутся от сумм по строке. Далее сравниваются процентные показатели, полученные для подгрупп с разным уровнем независимой переменной, каждый раз *внутри* одной категории зависимой переменной (например, внутри категории делающих зарядку). Обнаруженные различия свидетельствуют о существовании взаимосвязи между двумя переменными. (В качестве упражнения примените описанную процедуру к табл. 2.1, чтобы убедиться в наличии связи между уровнем образования и занятиями физкультурой.)

Отметим специально, что элементарная таблица сопряженности размерности 2×2 – это минимально необходимое условие для вывода о наличии взаимосвязи двух переменных. Знания о распределении зависимой переменной недостаточно. *Варьировать должна не только зависимая, но и независимая переменная.*

## 2.3. Коэффициенты связи для номинальных переменных

**Коэффициент Хи-квадрат** – наиболее известный из существующих показателей измерения степени зависимости двух переменных [6].

В теории вероятностей существует четкое определение независимости двух событий: два события считаются независимыми, если вероятность того, что они произойдут одновременно, равна произведению вероятностей того, что произойдет каждое из них. К примеру, мы бросаем две монеты – вероятность выпадения «орла» у каждой из них равна 1/2. То есть в случае

отсутствия зависимости между результатами подбрасывания двух монет вероятность одновременного выпадения «орлов» на обеих монетах равна произведению вероятностей выпадения «орла» на каждой из монет: $1/2 \cdot 1/2 = 1/4$.

В социологии, если в массиве данных оказалось 1/2 мужчин и 1/3 респондентов с высшим образованием, то при отсутствии зависимости между полом и образованием мужчин с высшим образованием в массиве должно быть $1/2 \cdot 1/3 = 1/6$. Количество респондентов, которые реально ответили на вопрос – это наблюдаемая частота. Количество респондентов, которое должно быть обозначено в клетке таблицы в случае независимости двух событий, называется ожидаемой частотой. Например, если мы опросили 1000 чел., среди них 1/2 – мужчины и 1/3 – с высшим образованием, то в случае независимости этих признаков ожидаемая частота составит $1/2 \cdot 1/3 \cdot 1000 = 166{,}7$. Но на вскидку делать сравнения между двумя числами – это не по-статистически.

Поэтому механизм проверки независимости переменных усложняется и проводится следующим образом:

1. Вычисляется степень суммарного расхождения реальных и ожидаемых частот. Здесь имеют место два обстоятельства:

- поскольку расхождение может принимать отрицательное значение, поэтому общая сумма может занижаться (поэтому суммируют квадраты разностей);
- квадрат разности может принимать различные по величине значения.

Например, для одной клетки реальное значение 12, а ожидаемое значение 2,2, квадрат разности будет 96,04, а в другой клетке – реальное значение 701, ожидаемое – 565,8, квадрат разности – 18279,04.

То есть первый квадрат разности даст меньший вклад в сумму, чем второй, при том, что разница между реальным и ожидаемым значением в первом случае составит более чем

5 раз. Чтобы преодолеть эту несообразность, складывают не абсолютные, а относительные расхождения частот.

Сумма таких относительных расхождений, или показатель, фиксирующий степень расхождения реальных и ожидаемых частот, называется коэффициентом хи-квадрат и определяется формулой:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i},$$

где $O_i$ – наблюдаемые частоты;

$E_i$ – ожидаемые частоты;

$n$ – число клеток в таблице.

Полученный результат, однако, не приближает нас к выяснению того, зависимы или нет между собой переменные, так как непонятно, значение хи-квадрат – большое или малое расхождение?

Если расхождение около 0, то степень независимости переменных очень высока, но хочется найти точное значение, относительно которого можно было бы сказать, что если хи-квадрат меньше $Z$ – значит модель независимости этих переменных подтверждается.

Для этого вводится понятие «степень свободы» – хи-квадрат напрямую зависит от количества этих степеней свободы.

Большую величину хи-квадрат можно получить в случае сильного расхождения между наблюдаемыми и ожидаемыми частотами, и во втором – когда расхождения маленькие, а количество самих слагаемых – большое. Поэтому когда говорят о величине хи-квадрат – говорят и о количестве клеток в таблице, т.е. о степенях свободы.

Число степеней свободы вычисляется по формуле:

$$df = (r - 1)(c - 1),$$

где $df$ (degreas freedom) – число степеней свободы;
$r$ – количество строк в таблице;
$c$ – количество столбцов.

Полученное значение хи-квадрат соотносят с количеством степеней свободы – для этого существует стандартизированная таблица распределений. По строкам идут значения степени свободы, а по столбцам – вероятность того, что значение сравниваемых переменных независимы (от 0 до 1) – эту вероятность еще называют уровень значимости. Поэтому, сначала находят степень свободы, а потом смотрят в каком диапазоне находится значение хи-квадрат – на основе этого делают вывод о том, что гипотеза о независимости данных переменных отвергается на уровне значимости 0,025.

К примеру, социолог хочет узнать, действительно ли то, что учителя более предвзято относятся к мальчикам, чем к девочкам. То есть более склонны хвалить девочек [18]. Для этого психологом были проанализированы характеристики учеников, написанные учителями, на предмет частоты встречаемости трех слов: «активный», «старательный», «дисциплинированный», синонимы слов подсчитывались. Данные о частоте встречаемости слов были занесены в табл. 2.2.

Таблица 2.2

Взаимосвязь между полом ученика и его оценкой педагогами

| | «Активный» | «Старательный» | «Дисциплинированный» |
|---|---|---|---|
| Мальчики | 10 | 5 | 6 |
| Девочки | 6 | 12 | 9 |

Для обработки полученных данных используем критерий **хи-квадрат**. Для этого построим таблицу распределения эмпирических частот, т.е. тех частот, которые наблюдаем (табл. 2.3).

Таблица 2.3

Взаимосвязь между полом ученика и его оценкой педагогами (эмпирические частоты)

| | «Активный» | «Старательный» | «Дисциплинированный» | Итого |
|---|---|---|---|---|
| Мальчики | 10 | 5 | 6 | 21 |
| Девочки | 6 | 12 | 9 | 27 |
| Итого | 16 | 17 | 15 | $n = 48$ |

Теоретически ожидается, что частоты распределятся равновероятно, т.е. частота распределится пропорционально между мальчиками и девочками. Построим таблицу теоретических частот. Для этого умножим сумму по строке на сумму по столбцу и разделим получившееся число на общую сумму $n$ (табл. 2.4).

Таблица 2.4

Взаимосвязь между полом ученика и его оценкой педагогами (теоретические частоты)

| | «Активный» | «Старательный» | «Дисциплинированный» | Итого |
|---|---|---|---|---|
| Мальчики | $(21 \cdot 16)/48 = 7$ | $(21 \cdot 17)/48 = 7{,}44$ | $(21 \cdot 15)/48 = 6{,}56$ | 21 |
| Девочки | $(27 \cdot 16)/48 = 9$ | $(27 \cdot 17)/48 = 9{,}56$ | $(27 \cdot 15)/48 = 8{,}44$ | 27 |
| Итого | 16 | 17 | 15 | $n = 48$ |

Итоговая таблица для вычислений будет выглядеть следующим образом (табл. 2.5):

Таблица 2.5

Взаимосвязь между полом ученика и его оценкой педагогами (итоговая таблица)*

| Категория 1 | Категория 2 | $O$ | $E$ | $(O-E)^2/E$ |
|---|---|---|---|---|
| Мальчики | «Активный» | 10 | 7 | 1,28 |
| | «Старательный» | 5 | 7,44 | 0,8 |
| | «Дисциплинированный» | 6 | 6,56 | 0,47 |
| Девочки | «Активный» | 6 | 9 | 1 |
| | «Старательный» | 12 | 9,56 | 0,62 |
| | «Дисциплинированный» | 9 | 8,44 | 0,04 |
| | | | | Сумма: 4,21 |

* $O$ – эмпирические частоты; $E$ – теоретические частоты.

$$\chi^2 = \sum(O - E)^2 / E,$$

$$df = (r - 1) \cdot (c - 1),$$

где $r$ – количество строк в таблице;

$c$ – количество столбцов.

В нашем случае $\chi^2 = 4{,}21$; $df = 2$.

По таблице критических значений критерия находим: при $df = 2$ и уровне ошибки 0,05 критическое значение $\chi^2 = 5{,}99$.

Полученное значение меньше критического, а значит принимается нулевая гипотеза. Вывод: учителя не придают значение полу ребенка при написании ему характеристики.

У статистики хи-квадрат есть одно ограничение – нельзя пользоваться данной формулой при значении ожидаемой частоты меньше 5, только 5 и более. Хотя, существует допущение, что клеток с такими частотами должно быть как можно меньше (на уровне 10–15 %).

*Коэффициенты связи, основанные на хи-квадрат*

В использовании показателя хи-квадрат кроется неудобство, поскольку само значение хи-квадрат ничего само по себе не значит. Для вывода о том, что значение хи-квадрат большое или маленькое, надо посчитать сначала степени свободы, затем заглянуть в таблицу распределений. Поэтому желателен такой коэффициент, который сразу будет показывать наличие взаимосвязи.

Один из таких коэффициентов – это коэффициент сопряженности Пирсона. Пирсон предложил коэффициент $C$, производный от хи-квадрат, само значение которого говорит о наличии, либо отсутствии связи между переменными.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}},$$

где $N$ – число опрошенных.

С ростом значения хи-квадрат значение показателя возрастает. При этом оно всегда больше нуля и меньше единицы. Но чем ближе значение к единице, тем выше уровень сопряженности или взаимосвязи между признаками.

Другой коэффициент, основанный на хи-квадрат, это коэффициент сопряженности Крамера, обозначаемый как $V$.

$$V = \sqrt{\frac{\chi^2}{N(K-1)}},$$

где $N$ – число опрошенных;

$K$ – наименьшее из чисел ($r$, $c$), где $r$ – количество строк; $c$ – количество столбцов.

Коэффициент Крамера меняется также от 0 до 1.

И коэффициент Пирсона, и коэффициент Крамера принимают значение 0, когда переменные независимы между собой, но коэффициент Крамера принимает значение 1 при жесткой связи (т.е. когда изменение одного значение всегда ведет к изменению другого значения), а коэффициент Пирсона – нет, оно всегда меньше 1.

Но, вспоминая о том, что зависимость – это отсутствие независимости, надо сказать, что значения хи-квадрат, коэффициентов Пирсона и Крамера не говорят о силе связи между переменными (например, если коэффициент Крамера для одной пары переменных равен 0,2, а для другой – 0,5) – данные значения говорят лишь об уменьшении вероятности, допускающей взаимосвязанность этих переменных.

*Коэффициенты связи, основанные на прогнозе*

Рассмотрим, что такое прогноз. У нас есть одномерное распределение какого-то признака – например, удовлетворенность учебой.

Интерпретировать его можно двумя способами:

- есть значения признака (различные степени удовлетворенности учебой);
- есть вероятности этих значений – относительные частоты – т.е. отношение количества выбравших данный вариант к общему числу опрошенных.

Например, опрошено 1000 чел., а удовлетворенность учебой «1» выбрало 200 чел. – это относительная частота, она же оценка вероятности – 0,2 – с этой вероятностью мы, выбрав случайного респондента в нашей выборке наткнемся на этого респондента. Эти вероятности в социологии по другому называются как вероятности статистического предсказания [9]. На основе этого строится прогноз – к примеру, 59,4 % опрошенных оценивает политическую обстановку в России как «критическую».

Соответственно, высказывание типа «Случайно взятый респондент охарактеризует политическую обстановку в стране как критическую» имеет вероятность подтверждения – 0,594, а прогноз ошибки этого высказывания – 1 – 0,594 = 0,406. То есть прогнозный вопрос будет звучать: какова вероятность предсказания оценки ситуации в стране как критической, взрывоопасной? Но это верно для одномерного распределения.

Если же мы вводим в анализ вторую переменную, то вопрос будет звучать уже следующим образом: как и насколько изменятся вероятности предсказания оценки ситуации, если мы учтем профессию (или уровень образования, материального благосостояния)?

Например, мы видим, что 80 % респондентов с высшим образованием оценивают ситуацию в стране как критическую. Соответственно, вероятность точности нашего попадания, если мы возьмем случайного респондента с высшим образованием, будет очень велика – с вероятностью 0,8 он нам ответит, что она критическая.

Для прогнозных подсчетов используется *коэффициент Гутмана,* который на практике обозначается буквой $\lambda$, а прогноз называется модальным. Назван он модальным потому, что считается по самым популярным (максимально вероятностным) значениям в таблице.

Считается по следующей формуле:

$$\lambda = \frac{\sum_{i=1}^{n} \max n_{ij} - \max n_j}{n - \max n_j},$$

где $\max n_{ij}$ обозначает наибольшую частоту в $i$-строке;
$\max n_j$ – наибольшую (итоговую) частоту по столбцам;
$n$ – общее количество ответов.

Коэффициент Гутмана позволяет определять, существуют ли в строках модальные группы, т.е. к примеру, в каждой группе *X* имеется ярко выраженная, часто встречаемая степень признака *Y*.

Рассмотрим пример, приведенный Дж.Б. Мангеймом, Р.К. Ричем [7, с. 417–419]. Представьте, например, что мы определяем партийную принадлежность 100 респондентов и выясняем, что частотное распределение выглядит следующим образом:

| | |
|---|---|
| Демократы | 50 |
| Республиканцы | 30 |
| Независимые | 20 |

Представьте также, что мы хотим установить партийную принадлежность каждого отдельного респондента и сделать подобные предположения для всех лиц и что мы хотим при этом совершить минимум ошибок. Наиболее очевидный путь – определить моду (самую распространенную категорию); мы предполагаем, что это будут демократы. Мы окажемся правы в 50 случаях (для 50 демократов) и не правы в 50 случаях (для 30 республиканцев и 1 независимых); это не просто стоящее внимания замечание, но самое лучшее, что мы можем сделать, поскольку, если мы выберем республиканцев, то окажемся не правы в 170 случаях, а если выберем независимых, то это приведет к 80 неверным предположениям. Таким образом, данная мода обеспечивает наилучший уровень предположений для имеющейся в распоряжении информации.

Но мы можем располагать еще одним набором данных, партийной принадлежности отца каждого респондента, представленным следующим распределением:

| | |
|---|---|
| Демократы | 60 |
| Республиканцы | 30 |
| Независимые | 10 |

Если эти две переменные связаны друг с другом, т.е., если каждый отдельный респондент, вероятнее всего, принадлежит к

той же партии, что и ее (или его) отец, то знание партийных предпочтений отца каждого респондента может помочь в определении партийных предпочтений самих респондентов. Это будет так в том случае, если, определяя для каждого респондента не моду всего распределения, как мы делали прежде, а просто партийную принадлежность его (или ее) отца, мы сможем снизить количество неверных предположений до уровня более низкого, чем 50 неверно определенных нами случаев.

Чтобы это проверить, нужно построить таблицу сопряженности, подытоживающую распределение признаков по этим двум переменным. В табл. 2.6 независимая, или определяющая, переменная (партийная принадлежность отца) дана по рядам, ее итоговое распределение находится в правой части таблицы. Зависимая переменная (партийная принадлежность респондента) расположена по колонкам, и ее итоговое распределение находится в низу таблицы. Значения в таблице даны произвольно, и в действительности они, конечно, должны пересчитываться самим исследователем.

Таблица 2.6

Определение партийности на основании партийной принадлежности отца

| Партийность отца | Партийность респондента | | | |
|---|---|---|---|---|
| | Демократ | Республиканец | Независимый | Всего |
| Демократ | 45 | 5 | 10 | 60 |
| Республиканец | 2 | 23 | 5 | 30 |
| Независимый | 3 | 2 | 5 | 30 |
| Всего | 50 | 30 | 20 | 100 |

По этой таблице можем партийные предпочтения родителей использовать для определения партийных предпочтений респондентов. Для этого, как и раньше, определим моду, но

только внутри *каждой категории независимой переменной,* а не по всему набору признаков. Таким образом, получится, что для тех респондентов, чьи отцы зафиксированы как демократы, прослеживаем предпочтение той же партии. Мы будем правы 45 раз и не правы 15 (для 5 республиканцев и 10 независимых). Для тех, чьи отцы зафиксированы республиканцами, предполагаем принадлежность к республиканской партии, при этом в 23 случаях мы окажемся правы и в 7 – не правы. Тех, чьи отцы зафиксированы независимыми, отнесем к независимым и будем правы в 5 из 10 случаев. Сравнив эти результаты, увидим, что теперь мы в состоянии верно предположить 73 раза и все еще ошибаемся 27 раз. Иными словами, наличие второй переменной существенно улучшило наши шансы. Для того чтобы точно определить процентную долю этого улучшения, используем общую формулу коэффициента связи.

В приведенном примере это выглядит следующим образом:

$$\lambda = \frac{5-(15+7+5)}{50} = \frac{23}{50} = 0,46.$$

Используя партийную принадлежность отца в качестве определителя партийной принадлежности респондента, можем улучшить (ограничить количество ошибок) наши предположения примерно на 46 %.

Свойства коэффициента Гутмана:

1. Изменяется от нуля до единицы.

2. Равен единице только в случае, когда в каждой группе $X$ все респонденты имеют одинаковую степень признака $Y$ и при этом в каждой отличную от другой.

3. Значение коэффициента равно нулю случае, когда имеет место отсутствие феномена модальности, т.е., условно говоря, полная «размытость» данных в таблице, когда в каждой из ячеек, присутствует понемногу.

## 2.4. Коэффициенты связи для порядковых переменных

Предыдущие рассуждения в основном касались связи номинальных переменных, или связи номинальных и порядковых переменных.

Когда же обе переменные измерены с помощью порядковой шкалы, то используются несколько коэффициентов. Рассмотрим один из самых распространенных – коэффициент «гамма» Гудмена-Краскела.

Коэффициент Гудмена-Краскела (гамма) – фиксирует степень соответствия двумерной модели:

1. « Если у респондента значение переменной *X* больше, чем у второго респондента, то у него будет больше и значение по переменной *Y*.

2. Или «чем больше значение *X*, тем меньше значение *Y*».

Рассмотрим пример И.Ф. Девятко [2, с. 59]. Для данных о внешней привлекательности (экспертные оценки) и популярности школьниц (данные опроса одноклассников).

Таблица 2.7

Ранги четырех школьниц по привлекательности (*X*) и популярности (*Y*)

| Случай | Переменная *X* (ранг по привлекательности) | Переменная *Y* (ранг по популярности) |
|---|---|---|
| Ольга | 1 | 1 |
| Светлана | 2 | 3 |
| Марьяна | 3 | 2 |
| Наташа | 4 | 4 |

Для того чтобы вручную рассчитать значение «гаммы» для небольшой выборки, нужно упорядочить наблюдения по независимой и зависимой переменным, как это показано в табл. 2.7.

Далее нужно сравнивать случаи (т.е. школьниц) попарно, определяя, сходится или расходится порядок расположения двух

этих случаев по двум переменным. Если упорядочения сходятся, пара называется *согласованной,* если они не сходятся, то пару нужно считать *несогласованной.* Результаты анализа для данных табл. 2.7 представлены в табл. 2.8.

Предполагается, что если согласованных (т.е. правильно предсказывающих порядок по зависимой переменной) пар больше, чем несогласованных, связь между переменными велика. Если несогласованных пар больше, то связь отрицательна (чем выше ранг по одной переменной, тем ниже ранг по другой). Если же различие между числом согласованных и несогласованных пар невелико, то связь между переменными просто отсутствует. Поэтому формула для «гаммы» такова:

$$\gamma = \frac{N_s - N_r}{N_s + N_r},$$

где $N_s$ – число согласованных пар;

$N_r$ – число несогласованных пар.

Таблица 2.8

Попарные сравнения рангов по переменным $X$ и $Y$

| Пара | Порядок по $X$* | Порядок по $Y$* | Знак пары («+» – согласованная, «–» – несогласованная) |
|---|---|---|---|
| Ольга – Светлана | О > С | О > С | + |
| Ольга – Марьяна | О > М | О > М | + |
| Ольга – Наташа | О > Н | О > Н | + |
| Светлана – Марьяна | С > М | М > С | – |
| Светлана – Наташа | С > Н | С > Н | + |
| Марьяна – Наташа | М > Н | М > Н | + |

*Здесь использованы лишь начальные буквы имен, т.е. «О > С» означает, что ранг Оли выше ранга Светы.

Для данных, используемых в нашем примере:

$$\gamma = \frac{5-1}{5+1} = 0{,}67.$$

Коэффициент гамма может принимать значения от –1 до +1. + 1 означает, что изменения по *X* влекут за собой такие же изменения и по *Y*, –1 – обратные изменения.

## 2.5. Коэффициент корреляции Пирсона (для интервальных, количественных переменных)

Линейный корреляционный анализ позволяет установить прямые связи между переменными величинами по их абсолютным значениям. Формула расчета коэффициента корреляции построена таким образом, что если связь между признаками имеет линейный характер, коэффициент Пирсона точно устанавливает тесноту этой связи. Поэтому он называется также коэффициентом линейной корреляции Пирсона.

В общем виде формула для подсчета коэффициента корреляции:

$$r_{xy} = \frac{\sum (x_i - x_{cp}) \cdot (y_i - y_{cp})}{\sqrt{\sum (x_i - x_{cp})^2 \cdot (y_i - y_{cp})^2}},$$

где $x_i$ – значения, принимаемые переменной *X*;
$y_i$ – значения, принимаемые переменой *Y*;
$x_{cp}$ – средняя по *X*;
$y_{cp}$ – средняя по *Y*.

Расчет коэффициента корреляции Пирсона предполагает, что переменные *X* и *Y* распределены нормально.

Даная формула предполагает, что из каждого значения $x_i$ переменной $X$, должно вычитаться ее среднее значение $x_{ср}$. Это неудобно, поэтому для расчета коэффициента корреляции используют не данную формулу, а ее аналог, получаемый с помощью преобразований:

$$r_{xy} = \frac{n \cdot \sum (x_i \cdot y_i) - (\sum x_i \cdot \sum y_i)}{\sqrt{[n \cdot \sum x_i^2 - (\sum x_i^2)] \cdot [n \cdot \sum y_i^2 - (\sum y_i^2)]}}.$$

Используя данную формулу, решим следующую задачу [19]:

*Пример*: 20 школьникам были даны тесты на наглядно-образное и вербальное мышление. Измерялось среднее время решения заданий теста в секундах. Психолога интересует вопрос: существует ли взаимосвязь между временем решения этих задач? Переменная $X$ – обозначает среднее время решения наглядно-образных, а переменная $Y$ – среднее время решения вербальных заданий тестов.

Для решения данной задачи представим исходные данные в виде табл. 2.9, в которой введены дополнительные столбцы, необходимые для расчета по формуле подсчета коэффициента корреляции (см. с. 39).

В табл. 2.9 даны индивидуальные значения переменных $X$ и $Y$, построчные произведения переменных $X$ и $Y$, квадраты переменных всех индивидуальных значений переменных $X$ и $Y$, а также суммы всех вышеперечисленных величин.

Рассчитываем эмпирическую величину коэффициента корреляции по формуле:

$$r_{xy} = \frac{20 \cdot 20089 \cdot 731 \cdot 518}{\sqrt{(20 \cdot 27873 - 731 \cdot 731) \cdot (20 \cdot 16000 - 518 \cdot 518)}}.$$

Таблица 2.9

Взаимосвязь выполнения заданий различного типа группой школьников

| Номер испытуемых | $X$ | $Y$ | $X \cdot Y$ | $X^2$ | $Y^2$ |
|---|---|---|---|---|---|
| 1 | 19 | 17 | 323 | 361 | 289 |
| 2 | 32 | 7 | 224 | 1024 | 49 |
| 3 | 33 | 17 | 561 | 1089 | 289 |
| 4 | 44 | 28 | 1232 | 1936 | 784 |
| 5 | 28 | 27 | 756 | 784 | 729 |
| 6 | 35 | 31 | 1085 | 1225 | 961 |
| 7 | 39 | 20 | 780 | 1521 | 400 |
| 8 | 39 | 17 | 663 | 1521 | 289 |
| 9 | 44 | 35 | 1540 | 1936 | 1225 |
| 10 | 44 | 43 | 1892 | 1936 | 1849 |
| 11 | 24 | 10 | 240 | 576 | 100 |
| 12 | 37 | 28 | 1036 | 1369 | 784 |
| 13 | 29 | 13 | 377 | 841 | 169 |
| 14 | 40 | 43 | 1720 | 1600 | 1849 |
| 15 | 42 | 45 | 1890 | 1764 | 2025 |
| 16 | 32 | 24 | 768 | 1024 | 5760 |
| 17 | 48 | 45 | 2160 | 2304 | 2025 |
| 18 | 42 | 26 | 1092 | 1764 | 676 |
| 19 | 33 | 16 | 528 | 1089 | 256 |
| 20 | 47 | 26 | 1222 | 2209 | 676 |
| Сумма | 731 | 518 | 20089 | 27873 | 16000 |

Для применения коэффициента корреляции Пирсона необходимо соблюдать следующие условия:

1. Сравниваемые переменные должны быть получены в интервальной шкале или шкале отношений.

2. Распределения переменных $X$ и $Y$ должны быть близки к нормальному.

3. Число варьирующих признаков в сравниваемых переменных *X* и *Y* должно быть одинаковым.

То есть коэффициент Пирсона показывает то, насколько переменные *X* и *Y* одновременно отклоняются от средних значений.

Нулевое или около нулевое значение может рассматриваться и как отсутствие зависимости между переменными вообще, и о том, что зависимость есть, но она носит нелинейный характер.

То есть нулевое значение коэффициента Пирсона говорит лишь об отсутствии линейной зависимости между переменными.

### Вопросы для самопроверки

1. Что называют коэффициентом корреляции и какова роль подобных коэффициентов в анализе социологических данных?
2. Какую роль в анализе данных играет вычисление статистической значимости?
3. Каковы основные правила построения двумерных таблиц?
4. Для чего предназначена и как рассчитывается статистика хи-квадрат?
5. Какие коэффициенты связи основаны на статистике хи-квадрат? В чем их особенности?
6. Для чего предназначены и как рассчитываются коэффициенты связи, основанные на прогнозе?
7. Какие коэффициенты предназначены для измерения связи порядковых переменных?
8. В чем особенности коэффициента Пирсона для интервальных данных?

# Глава 3. АНАЛИЗ ВЗАИМОСВЯЗЕЙ КАЧЕСТВЕННЫХ И КОЛИЧЕСТВЕННЫХ ПЕРЕМЕННЫХ



## 3.1. Визуализация различий средних значений

При анализе социологических данных и при построении социологических моделей, социолог чаще всего использует причинные модели, в которых некий показатель выступает как следствие каких-то причин. При таком анализе нас интересует то, насколько данные причины определяют данное следствие. Надо отметить, что ни один математико-статистических метод не в состоянии объяснить, почему получилось такое распределение (например, почему пол стал причиной увеличения доходов), но существует большое количество методов проверки таких моделей, которые конструирует социолог.

В этой главе рассмотрим методы анализа, позволяющие строить причинные модели в ситуации, когда переменная-следствие измерена по метрической шкале, а переменные-причины – по неметрическим шкалам (порядковым или номинальным).

Перед анализом необходимо продемонстрировать средние значения каких-то количественных показателей в социальных, демографических и др. группах [4].

Визуализация средних значений напоминает задачи описательного анализа с помощью одномерных частотных распределений, однако в этом случае средние значения необходимо получить не по выборке в целом, а по отдельным группам.

Осуществляется это в программе SPSS с помощью команды Means в меню Analyze – Compare Means (рис. 3.1).

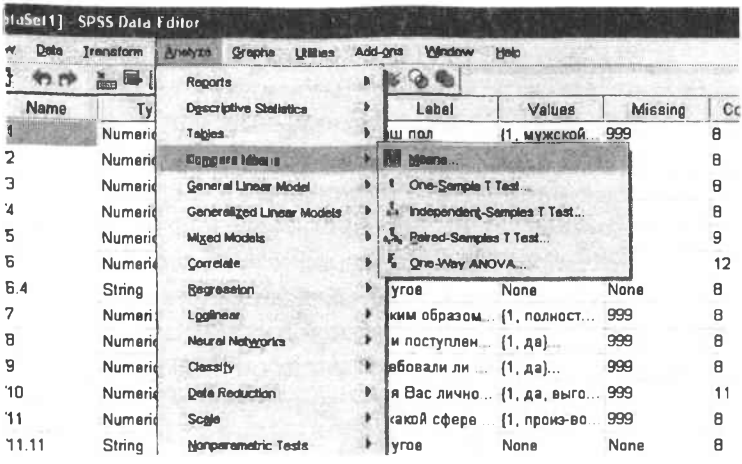



Рис. 3.1. Визуализация средних значений в программе SPSS

В главном меню *Means* необходимо задать два типа переменных: *Dependent List (зависимые переменные)* – это количественные переменные, средние значения которых необходимо вычислять. Второй тип – *Independent List (независимые)* – это переменные, которые определяют разделение всей совокупности опрошенных на определенные группы (рис. 3.2).

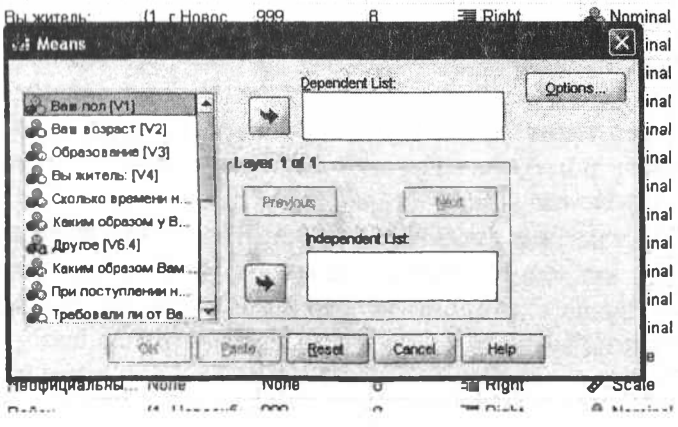



Рис. 3.2. Выбор зависимой и независимой переменных

Например, независимой переменной у нас стал *возраст*, а зависимой – *размер официальной зарплаты* (рис. 3.3).

**Report**

Официальный размер ЗП

| Ваш ... | Mean | N | Std. Deviation |
|---|---|---|---|
| до 20 лет | 8330,68 | 22 | 4179,712 |
| 20-24 | 7566,56 | 114 | 5599,209 |
| 25-29 | 7362,56 | 123 | 5797,386 |
| 30-34 | 7554,70 | 137 | 4800,204 |
| 35-39 | 8536,36 | 88 | 8007,298 |
| 40-44 | 7455,52 | 99 | 5166,388 |
| 45-49 | 8819,36 | 119 | 6474,605 |
| 50-55 | 7056,77 | 163 | 6385,968 |
| 56-60 | 9312,28 | 43 | 7197,868 |
| Total | 7792,88 | 908 | 6052,562 |

Рис. 3.3. Отчет программы

Как видно из рис. 3.3, команда *Means* по умолчанию вычисляет среднее значение для каждой из групп респондентов, количество самих респондентов в группе и стандартное отклонение анализируемого показателя в каждой из групп.

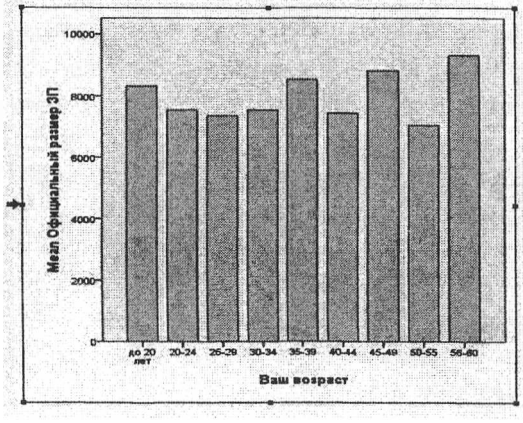



Рис. 3.4. Визуализация различий

С помощью визуализации (рис. 3.4) можем наглядно представить то, что усредненные данные в группах отличаются и даже сделать вывод о степени отличий между значениями в разных группах. Но этого явно недостаточно для полного доказательства действительных отличий. Поэтому для такого, более глубокого, анализа используется команда T-Test.

### 3.2. Команда T-Test (тест Стьюдента)

Команда T-Test (тест Стьюдента) решает задачу доказательства наличия различий средних значений количественной переменной в усеченном виде – а именно для случая, когда имеются только две сравниваемые группы (например пол, или две возрастных группы). То есть если мы хотим выяснить отличается ли заработок по всем возрастным группам – мы должны будем провести анализ несколько раз.

Есть три разновидности команды T-Test:

1. Команда T-Test для сравнения двух независимых выборок.
2. Команда T-Test для одной выборки.
3. Команда T-Test для парных данных.

***1. Команда T-Test для сравнения двух независимых выборок*** – т.е. для выборок, в которых сбор данных осуществлялся независимо, т.е. то, как отвечали в одной выборке (например, женщины) не влияло, как отвечают в другой выборке (мужчины). В социологических исследованиях типа анкетирования почти всегда выборки являются независимыми.

Вызов команды сравнения средних значений количественной переменной в двух независимых выборках осуществляется путем перехода к соответствующему меню – *Independent-Samples T-Test (Т-критерий для независимых выборок)*. Проиллюстрируем это (рис. 3.5).

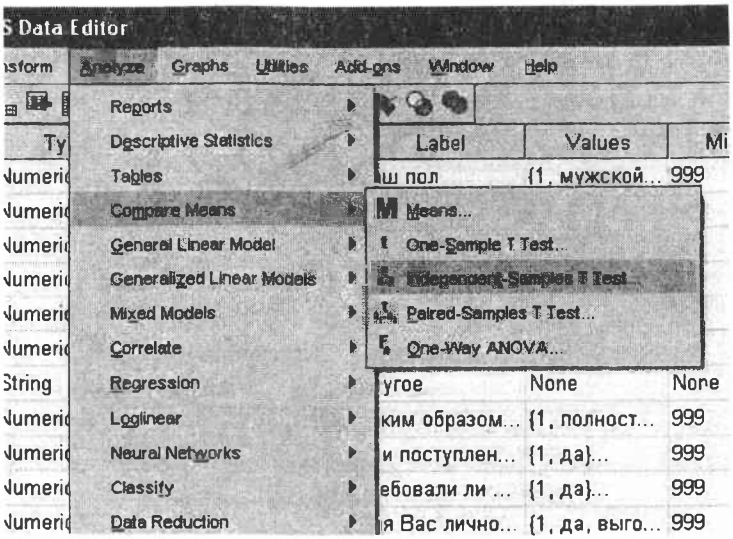



Рис. 3.5. T-Test для сравнения двух независимых выборок

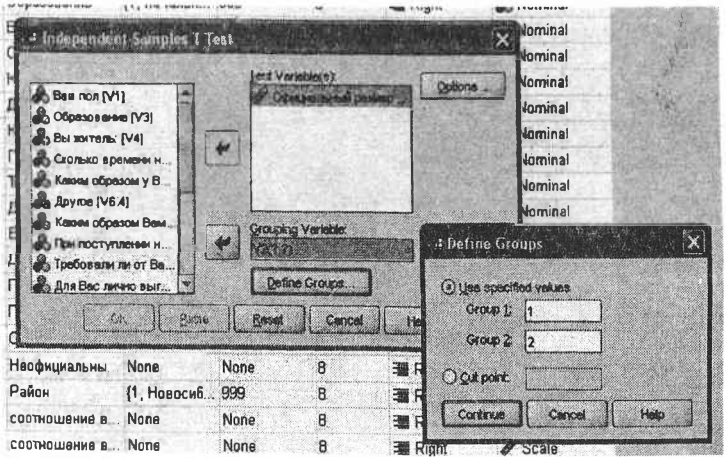



Рис. 3.6. Указание тестируемых групп

В строке *Grouping Variable (Группировать по:)* необходимо задать две группы сравниваемых между собой переменных, для этого нажимается кнопка *«Задать группы»*, где указать номера сравниваемых групп (рис. 3.6).

Выведенные результаты содержат в себе две таблицы – верхняя является повтором предыдущей функции «Среднее», где перечислено количество наблюдений, средние значения, стандартные отклонения и стандартные ошибки средних в обеих группах (рис. 3.7).

**Group Statistics**

| | Ваш возраст | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| Официальный размер ЗП | до 20 лет | 22 | 8330,68 | 4179,712 | 891,118 |
| | 20-24 | 114 | 7566,56 | 5599,209 | 524,414 |

Рис. 3.7. Результаты анализа сравнения средних значений

Вторая таблица (рис. 3.8) содержит в себе результаты *теста Левена*, который сравнивает дисперсии двух групп.

**Independent Samples Test**

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | 95% Confidence Interval of the Difference | |
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | Lower | Upper |
| Официальный размер ЗП | Equal variances assumed | 1,204 | ,274 | ,607 | 134 | ,545 | 764,120 | 1257,815 | -1723,618 | 3,252 E3 |
| | Equal variances not assumed | | | ,739 | 3,723 E1 | ,465 | 764,120 | 1033,973 | -1330,464 | 2,859 E3 |

Рис. 3.8. Результаты теста Левена

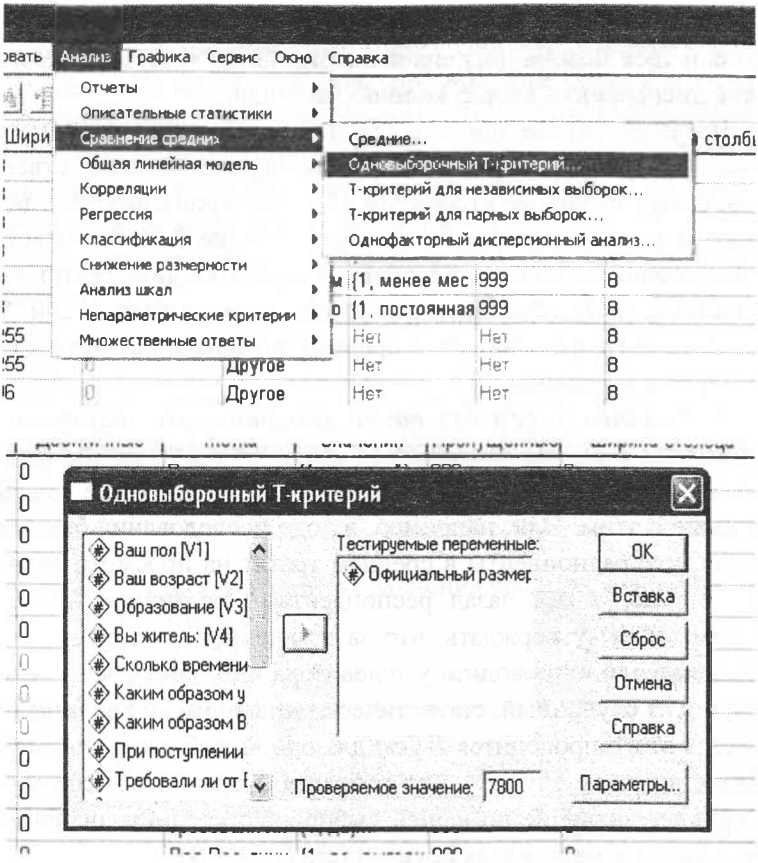
Как правило, гипотеза о равенстве дисперсий не принимается, если тест Левена дает значение *Sig. (Знч.)* < 0,05 (гетерогенность дисперсий) – вторая колонка таблицы.

В нашем случае *Знч.* = 0,274, т.е. это больше чем 0,05, поэтому предполагается равенство дисперсий. Соответственно проверяем с помощью второй *Sig.* (*Знч.* 2-сторон), поэтому далее смотрим по этому числу – оно также больше 0,05 – соответственно можно сказать что различия в заработке двух групп мало существенны. В обратном же случае (если значение 2-й *Sig.* меньше 0,05) – различие двух средних значений признается статистически значимым.

***2. Команда T-Test для одной выборки*** – это тестирование проводится в случае имеющегося среднего значения (например данные госстатистики) и необходимо сравнить наше среднее значение с этим. Или, например, в ходе исследования было выяснено, что респонденты в среднем тратят на просмотр телевизора 2 часа, а год назад респондентами тратилось 1,8 часа. Можем ли мы утверждать, что за прошедший год люди стали больше времени проводить у телевизора или обнаруженная разница носит случайный, статистически не значимый характер?

Для этого проводится *T-Test* для одной выборки *(one-sample T-Test),* в нашем случае – для того чтобы сравнить – совпадает ли среднее значение по нашей выборке со средним значением, имеющимся в материалах Госкомстата (рис. 3.9).

Снова обращаем внимание на показатель *Sig. (Знч. 2-сторон),* он равен 0,998 – т.е. можем сказать, что полученные данные близки официальной статистике.




**Одновыборочный критерий**

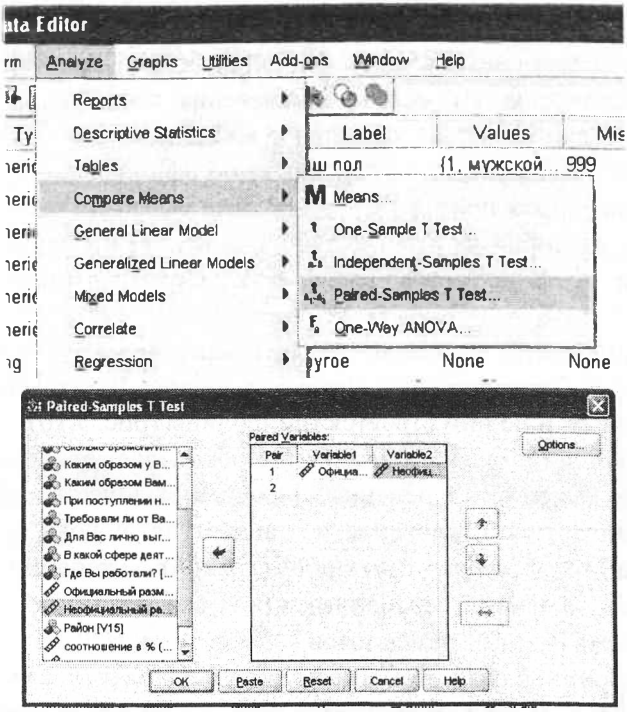
| | Тестовое значение = 7800 | | | | | |
|---|---|---|---|---|---|---|
| | | | | | 95% доверительный интервал разности средних | |
| | t | ст.св. | Знч. (2-сторон) | Разность Средних | Нижняя граница | Верхняя граница |
| Официальный размер ЗП | ,003 | 912 | ,998 | ,584 | -392,02 | 393,19 |

Рис. 3.9. Вывод данных анализа одновыборочного критерия

***3. Команда T-test для парных данных*** – когда сравниваем значения переменных в рамках одной выборки (например, употребление до начала и после начала лечения; потребление разных типов продуктов и т.д.) применяют команду T-Test для парных выборок. Например, имеются данные об официальной и неофициальной заработной плате. Необходимо выяснить связаны они или нет, т.е. имеются у нас достаточные основания утверждать о наличии различий между данными заработными платами или нет.

Выполняется данная команда из меню Анализ – Сравнение средних – Т-критерий для парных выборок. Выбирается пара переменных, и по ним строятся три таблицы (рис. 3.10).

Таблица *Paired Samples Statistics* показывает отдельные значения по каждой из переменных (среднее, количество наблюдений, стандартное отклонение и стандартная ошибка среднего) Таблица *Paired Samples Correlations* говорит о степени корреляции между данными переменными (а также об уровне значимости представленных результатов – *Sig*). Таблица *Paired Samples Test* говорит как раз о том, имеем ли мы статистические основания утверждать о наличии различий в значениях переменных – если меньше 0,05 – то имеем, если больше 0,05 – не имеем.




**Paired Samples Statistics**

| | | Mean | N | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| Pair 1 | Официальный размер ЗП | 6346,65 | 378 | 4760,899 | 244,874 |
| | Неофициальный размер ЗП | 9332,75 | 378 | 8976,320 | 461,692 |

**Paired Samples Correlations**

| | | N | Correlation | Sig |
|---|---|---|---|---|
| Pair 1 | Официальный размер ЗП & Неофициальный размер ЗП | 378 | ,131 | ,011 |

**Paired Samples Test**

| | | Paired Differences | | | | | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference Lower | Upper | | | |
| Pair 1 | Официальный размер ЗП - Неофициальный размер ЗП | -2,986 E3 | 9593,610 | 493,442 | -3956,341 | -2015,854 | -6,052 | 377 | ,000 |

Рис. 3.10. Вывод данных тестирования парных выборок

### 3.3. Однофакторный дисперсионный анализ

К сожалению, тест Стьюдента приспособлен только для изучения переменных, имеющих не более двух градаций – иначе нужно попарно их сравнивать – T-test дает возможность сопоставить только две градации – для большего количества градаций переменной используют метод дисперсионного анализа.

С точки зрения построения социологической модели? вопрос можно сформулировать следующим образом: оказывает ли значимое влияние на значение некоторой количественной переменной интересующая нас переменная, которая измерена на номинальном или порядковом уровне?

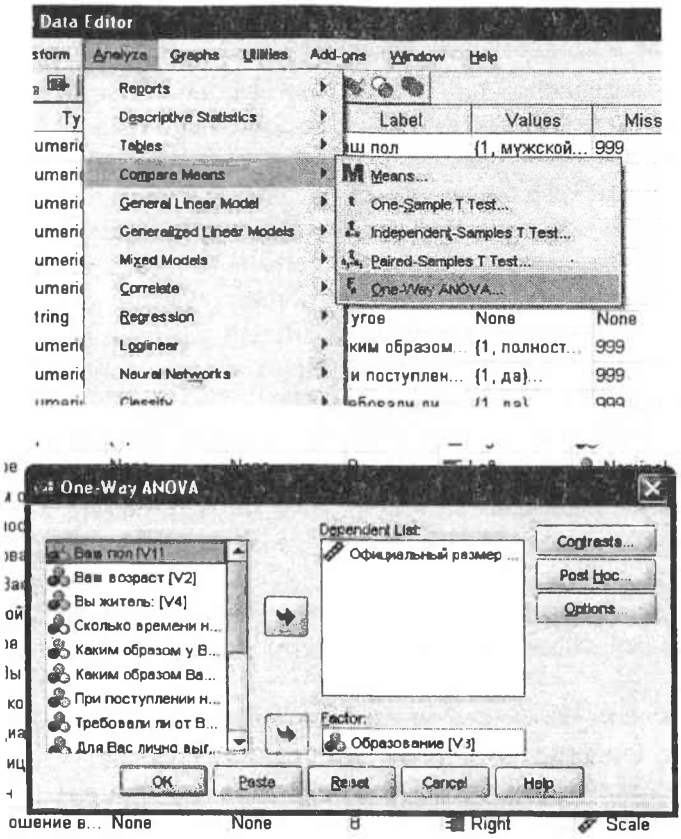
Та переменная, которая, как мы считаем, должна влиять на конечный результат – называется фактором. Например, если сравниваем зарплаты у респондентов, проживающих в разных населенных пунктах, то «Тип населенного пункта» – это *фактор.*

Конкретную реализацию, значение фактора (например, определенный тип населенного пункта) называют *уровнем фактора.*

Значение измеряемого признака (например, величина зарплаты) – называют *откликом.*

В рамках пакета SPSS команда, реализующая метод однофакторного дисперсионного анализа, называется *One-Way ANOVA* (*One-Way* – однофакторный; *Analisys Of VAriance* – дисперсионный анализ) [8].

Однофакторный дисперсионный анализ в SPSS осуществляется следующим образом (рис. 3.11):




**ANOVA**

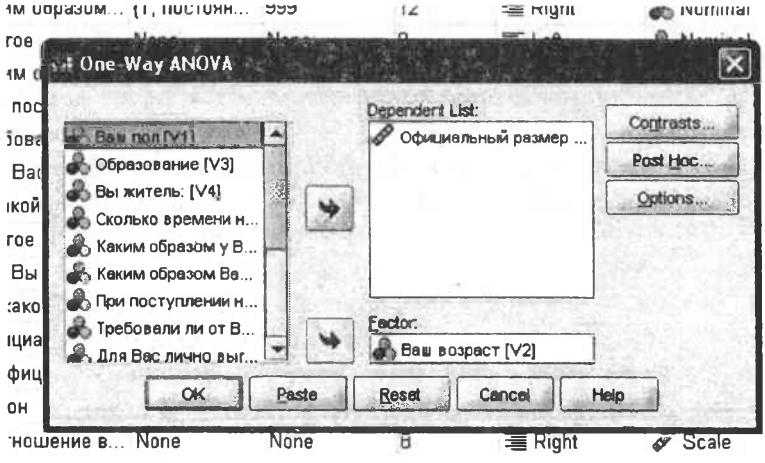
Официальный размер ЗП

| | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Between Groups | 1,681E9 | 6 | 2,801E8 | 8,514 | ,000 |
| Within Groups | 2,954E10 | 898 | 3,290E7 | | |
| Total | 3,122E10 | 904 | | | |

Рис. 3.11. Вывод данных однофакторного дисперсионного анализа

Значение Sig. стремится к нулю, поэтому можно сказать, что не все образовательные группы имеют одинаковую зарплату – другими словами зарплаты с разными уровнями образования – не имеют одинаковости, поэтому следует вывод о том, что разные уровни образования в данной выборке имеют разную зарплату.

В другом случае, когда мы в качестве фактора выбираем возраст, significance (степень значимости) равна 0,182, что больше 0,05, поэтому здесь можно сказать, что возрастные группы не имеют настолько четких отличий в зарплате, и зарплата распределена более равномерно (рис. 3.12).




**ANOVA**

Официальный размер ЗП

| | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Between Groups | 4,156E8 | 8 | 5,196E7 | 1,424 | ,182 |
| Within Groups | 3,281E10 | 899 | 3,650E7 | | |
| Total | 3,323E10 | 907 | | | |

Рис. 3.12. Результаты однофакторного дисперсионного анализа

Таким образом, можно сказать, что тест ANOVA необходимо сверять с обычной визуализацией различий средних значений, постоянно сопоставлять то, что считает машина, и тем, что мы видим.

### Вопросы для самопроверки

1. В каких случаях используется анализ средних значений?
2. Для чего предназначена процедура Т-тестирования и каковы ее разновидности?
3. В каких случаях рассчитывается Т-тест для независимых выборок?
4. Каковы возможности Т-тестирования для одной выборки?
5. Когда применяется и как рассчитывается Т-тест для парных выборок?
6. В чем заключаются особенности применения однофакторного дисперсионного анализа?

# Глава 4. МОДЕЛИ РЕГРЕССИОННОГО АНАЛИЗА



## 4.1. Общее описание регрессионной модели

Предположим, что необходимо выяснить причины хорошей и плохой успеваемости студентов. В ходе логического анализа понятий выводим факторы, которые влияют на успеваемость студентов – допустим, этими факторами являются:

- уровень подготовки студента;
- активность посещения занятий;
- активность самостоятельной работы;
- способности студента.

Изученные ранее коэффициенты фиксируют лишь то, насколько *тесно* связаны две переменные. А вот насколько *сильно* они связаны – данные коэффициенты не могут сказать.

Решение данной задачи начнем с упрощения данной модели. Смысл построения данной модели состоит в том, чтобы выяснить, каким образом на успеваемость влияет именно уровень предварительной подготовки, каково направление и сила этого влияния.

Предложенную модель зависимости можно представить в виде следующей математической зависимости:

$$y = f(x) + u,$$

где $y$ – показатель «Успеваемость студента»;

$x$ – показатель «Уровень предварительной подготовки»;

$f$ – функция, описывающая силу влияния $x$ на $y$;

$u$ – другие факторы, влияющие на $y$.

В качестве показателя «Уровень предварительной подготовки» выступает суммарный балл, полученный студентом на вступительных экзаменах в вуз, в качестве показателя «Успеваемость» – суммарный балл студента в 1 семестр обучения в вузе.

Коэффициент корреляции Пирсона для этих данных составит 0,43, его значимость будет на уровне значимости равный 0,06. То есть, можно сделать вывод о том, что между баллами имеется средняя связь, и модель отражает реально существующие закономерности.

Нас интересует значение $f$ (помним формулу), т.е. то, насколько влияет на успеваемость именно фактор «уровень предварительной подготовки». Данное значение характеризует использование уравнения простой (или парной) линейной регрессии:

$$y = b_0 + b_1 x + u,$$

где $b_0$ – это точка пересечения прямой с осью $y$.

График, демонстрирующий данное уравнение представлен на рис. 4.1.

Данный график наглядно демонстрирует, что если мы проведем прямую сквозь массу этих точек так, чтобы все точки максимально близко лежали к этой прямой, эта прямая называется *регрессионной прямой.*

Каждая точка находится на определенном расстоянии от этой прямой – вертикальное расстояние между точкой и прямой называется остатком. Поскольку остаток может принимать отрицательное значение, то остаток возводят в квадрат. Все квадраты остатков суммируют, и чем меньше сумма квадратов остатков, тем ближе данные находятся к прямой, тем теснее взаимосвязь между переменными.

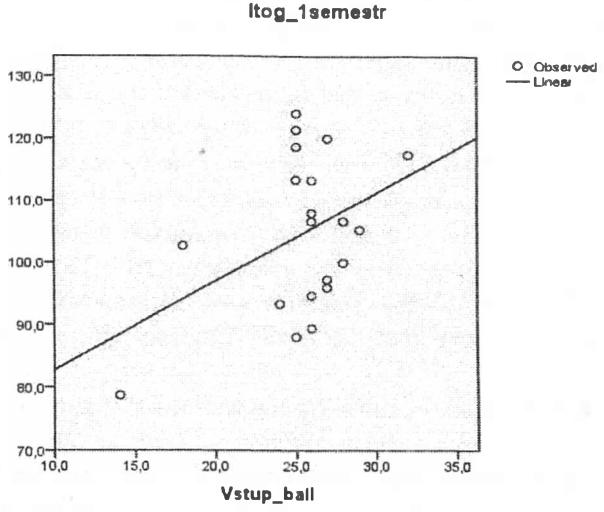



Рис. 4.1. Регрессионная прямая

В рассматриваемом случае это означает, что студенты, набравшие на вступительных экзаменах 0 баллов, по итогам сдачи 1-го семестра будут иметь успеваемость 68,4 балла.

Значение $b_1$ более интересно – это то, насколько баллов возрастает средняя успеваемость студента в 1-м семестре при увеличении на единицу балла на вступительных экзаменах.

То есть увеличение оценки на вступительных экзаменах на 1 балл дает улучшение успеваемости студента в 1-м семестре на 1,428 балла. Именно этот коэффициент демонстрирует силу связи между $y$ и $x$.

Еще один показатель, вычисляемый SPSS в меню Analyze – Regression (Регрессия), это показатель, который говорит о качестве влияния фактора (вступительный балл) на успеваемость в 1-м семестре – это коэффициент детерминации (R Square). Рег-

рессионная модель хороша, если большая часть значения изменений $y$ объясняется изменениями первой части формулы, соответственно чем меньше расхождения между собственно $y$ и $(b_0 + b_1x)$, то тем выше качество регрессионной модели.

R square – это отношение дисперсии значения $y$ по отношению к дисперсии суммы $(b_0 + b_1x)$. R square всегда положителен и равен 1, когда сумма $(b_0 + b_1x)$ полностью описывает $y$, или отсутствуют другие факторы $(u)$.

В нашем случае он равен 0,18, что можно объяснить как то, что успеваемость студентов в 1-м семестре на 0,18 (из единицы) или на 18 % объясняется уровнем предварительной подготовки студентов, соответственно, на 82 % – другими факторами.

### 4.2. Множественный регрессионный анализ

Для построения математической модели одновременного влияния на успеваемость нескольких факторов (независимых переменных, предикторов) на зависимую переменную используют усложненный вариант простой линейной регрессии – модель *множественной линейной регрессии.*

Общий вид модели множественной линейной регрессии – это естественное развитие уравнения для простой линейной регрессии:

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots b_nx_n + u.$$

В нашем случае всего два измеряемых по силе влияния фактора – балл по итогам 1-го семестра и новый показатель – процент занятий, пропущенных студентом.

Как и в случае с изучением силы влияния одного фактора, в случае влияния нескольких факторов, прежде всего нужно посмотреть – а есть ли связь вообще?

Для этого вычисляем коэффициент корреляции Пирсона.

В нашем случае – в отношении зависимости вступительного балла и экзаменационного балла – коэффициент корреляции со-

ставил 0,43, а в случае зависимости процента пропущенных занятий и экзаменационного балла – минус 0,62

Уравнение принимает вид:

$$y = 87{,}35 + 1{,}19x_1 - 1{,}32x_2,$$

где значение $b_0 = 87{,}35$ – это константа, характеризующая значение пересечения новой регрессионной прямой с осью $y$;
$b_1 = 1{,}19$, $b_2 = -1{,}32$

Положительный коэффициент при независимой переменной говорит о том, что с возрастанием последней значение зависимой переменной также возрастает. Верно и противоположное утверждение: при отрицательном коэффициенте с возрастанием значения независимой переменной значение зависимой переменной убывает.

*Значение Beta-коэффициента*

Переменные могут иметь разные единицы измерения (величина дохода ($x_1$) и возраст покупателя ($x_2$).

Коэффициент «*Beta*» приводит единицы измерения к единому соотношению – и можно увидеть разницу того, какая переменная оказывает большее влияние на зависимую.

В нашем случае, переменная пропуск оказывает большее влияние в 1,6 раза, чем переменная вступительный балл.

Ограничения множественного регрессионного анализа.

1. Ситуация, когда некоторые точки резко выпадают из общей тенденции и соответственно далеко стоят от регрессионной прямой – их называют выбросами. Выбросы ухудшают нормальность распределения наших значений, и это влечет неточность подсчетов квадратов расстояний до прямой. Причины как правило две: это или ошибка оператора при забивке данных, или это попадающиеся маргиналы. В любом случае от них необходимо избавляться.

Можно избавляться вручную, удаляя анкету из обработки, либо просто в меню линейной регрессии – в подменю «Статистики» указывать в ячейке – значение, которое превышает стандартное отклонение, например в 3 раза. Соответственно значения, которые выходят за эту границу, будут нивелированы.

2. Второй момент, на который необходимо обращать внимание – это высокий уровень корреляции между независимыми переменными (*значимая мультиколлинеарность*).

Например, когда мы хотим узнать, что влияет больше на количество покупок респондентами товаров за последнее время – личный доход респондента или среднедушевой доход его семьи?

Эти две переменные между собой коррелируют достаточно сильно, поэтому spss при вычислении значения $b_1$ для этих переменных будут вычислять как очень малые величины.

Требования к проведению множественного регрессионого анализа:

- исследование должно быть продумано по форме и исполнению;
- для того чтобы существующие корреляции были признаны значимыми, необходимо иметь достаточные размеры выборок ($N > 50$);
- данные должны быть корректными и не содержать ошибок;
- распределение значений предикторов должно быть близким к нормальному, без выбросов;
- наиболее жестким требованием является запрет па использование зависимых переменных, корреляции между которыми близки к 1 (–1). Не следует задействовать предикторы, схожие между собой но смыслу.

## 4.3. Логистическая регрессия

Ограничением регрессионного анализа является то, что переменные должны быть выражены количественно, т.е. иметь интервальный уровень измерения. Однако, если необходимо измерить действие разных факторов (предикторов) на то, покупает респондент товар или нет, курит или нет и т.д., т.е. на переменную, выраженную дихотомически – используют метод логистической регрессии.

Логистическая регрессия представляет собой расширение множественной регрессии и отличается от последней тем, что в качестве зависимой переменной используется не количественная, а дихотомическая переменная, имеющая лишь два возможных значения.

Как правило, эти два значения символизируют принадлежность или непринадлежность объекта какой-либо группе, ответ типа «да» или «нет» и т.п.

С логистической регрессией связаны математические понятия:

- вероятность;
- шанс;
- натуральный логарифм шанса.

Вероятность – это ожидаемая относительная частота некоторого события. Вероятность измеряется в диапазоне от 0 до 1. Мерой вероятности является дробь, числитель которой есть число всех замеренных случаев, а знаменатель – число всех возможных случаев.

Например, если необходимым случаем является дождь, то его вероятность пойти в данных условиях – это отношение наблюдаемых в этот день (или при таких условиях) по отношению ко всем наблюдениям. Например, это значение составит 0,2. В противоположность, вероятность того, что дождь не пойдет (событие не наступит) – составит 0,8.

Шанс представляет собой отношение вероятности того, что событие произойдет, к вероятности того, что событие не произойдет. Так, если вероятность дождя равна 0,2, следовательно,

вероятность отсутствия дождя равна 0,8 – шанс, что дождь все-таки прольется, равен 0,2/0,8 = 0,25. Шанс, в отличие от вероятности, не ограничен максимальным единичным значением (к примеру, если вероятность дождя составляет не 0,2, а 0,8, то получаем шанс 0,8/0,2 = 4).

*Натуральный логарифм шанса (логит)*

Логарифм – это степень, в которую нужно возвести другое число, называемое основанием логарифма, чтобы получить данное число. Например, логарифм числа 100 по основанию 10 равен 2. Иначе говоря, 10 нужно возвести в квадрат, чтобы получить число 100 (10^2 = 100). Если $n$ – заданное число; $b$ – основание; l – логарифм, то $b$^l = $n$.

Натуральный логарифм (ln) – это логарифм по основанию e (трансцендентному числу, приближенно равному 2,71828). Натуральные логарифмы записывают, не указывая основание, но используя специальное обозначение ln: например, ln2 = 0,6931, так как e^0,6931 = 2.

Натуральный логарифм шанса (логит) в регрессионном уравнении является значением $y$ (значением зависимой переменной), выраженной дихотомически. Например, анализируется влияние разных факторов на электоральное поведение. Для тех, кто голосовал за партию $A$ введем значение $y$ = 1, а для тех, кто голосовал за другие партии y = 0. Если по результатам исследования мы получили, что логит голосования за партию А для мужчин равен –0,847, а для женщин –1,386, это значит, что отношение предпочтения для мужчин равно e^ –0,847 = 0,43, а для женщин e^ –1,386 = 0,25. Другими словами – среди мужчин за партию $A$ проголосовали 43 %, а среди женщин – 25 %.

Формула логистической регрессии:

$$\ln\left[\frac{P}{1-P}\right] = b_0 + b_1x_1 + b_2x_2 + \ldots b_nx_n.$$

В классификационной таблице сравниваются прогнозируемые значения зависимой переменной, рассчитанные по уравне-

нию регрессии, и фактические наблюдаемые значения. Как показывают данные крайнего правого столбца таблицы, для 73,8 % объектов результаты прогноза оказались верными:

- B – коэффициенты регрессионного уравнения, отражающие влияние соответствующих предикторов на зависимую переменную;
- S.E. (Стандартная ошибка) – мера изменчивости коэффициентов *B*;
- Wald (Критерий Вальда) – показатель значимости коэффициента *B* для независимой переменной. Чем выше его значение, тем выше значимость;
- Sig. (Значимость) – значимость по критерию Вальда;
- Exp(B) – величина (Beta), которая может использоваться для интерпретации результатов анализа наравне с коэффициентом *B*.

## Вопросы для самопроверки

1. Для решения каких задач целесообразно построение регрессионной модели?
2. Как рассчитывается и интерпретируется коэффициент детерминации R-квадат?
3. Какова логика построения модели множественной регрессии?
4. Какую роль играют *Beta*-коэффициенты в процедуре множественного регрессионного анализа?
5. Как решается проблема выбросов в модели множественной регрессии?
6. Что такое значимая мультиколлинеарность и каковы пути ее снижения?
7. Каковы основные требования к множественному регрессионному анализу?
8. Каковы возможности логистического регрессионного анализа?

# Глава 5. ФАКТОРНЫЙ АНАЛИЗ



Факторный анализ – это процедура, с помощью которой большое число переменных, относящихся к имеющимся наблюдениям, сводят к меньшему количеству независимых влияющих величин, называемых факторами. При этом в один фактор объединяются переменные, сильно коррелирующие между собой. Переменные из разных факторов слабо коррелируют между собой. Таким образом, целью факторного анализа является нахождение таких комплексных факторов, которые как можно более полно объясняют наблюдаемые связи между переменными, имеющимися в наличии.

Факторный анализ может быть 1) *разведочным*. Он осуществляется при исследовании скрытой факторной структуры без предположения о числе факторов и их нагрузках; 2) *конфирматорным*, предназначенным для проверки гипотез о числе факторов и их нагрузках. Практическое выполнение факторного анализа начинается с проверки его условий. В обязательные условия факторного анализа входят [6]:

- все признаки должны быть количественными.
- число признаков должно быть в два раза больше числа переменных.
- выборка должна быть однородна.
- исходные переменные должны быть распределены симметрично.
- факторный анализ осуществляется по коррелирующим переменным.

## 5.1. Порядок выполнения факторного анализа

На первом шаге процедуры факторного анализа происходит стандартизация заданных значений переменных (*z*-преобразование); затем при помощи стандартизированных значений рассчитывают корреляционные коэффициенты Пирсона между рассматриваемыми переменными.

Исходным элементом для дальнейших расчетов является корреляционная матрица. Для понимания отдельных шагов этих расчетов потребуются хорошие знания, прежде всего, в области операций над матрицами; интересующимся подробностями советуем обратиться к специальной литературе. Для построенной корреляционной матрицы определяются, так называемые, собственные значения и соответствующие им собственные векторы, для определения которых используются оценочные значения диагональных элементов матрицы (так называемые относительные дисперсии простых факторов).

Собственные значения сортируются в порядке убывания, для чего обычно отбирается столько факторов, сколько имеется собственных значений, превосходящих по величине единицу. Собственные векторы, соответствующие этим собственным значениям, образуют факторы; элементы собственных векторов получили название факторной нагрузки. Их можно понимать как коэффициенты корреляции между соответствующими переменными и факторами. Для решения такой задачи определения факторов были разработаны многочисленные методы, наиболее часто употребляемым из которых является метод определения главных факторов (компонент).

Описанные выше шаги расчета еще не дают однозначного решения задачи определения факторов. Основываясь на геометрическом представлении рассматриваемой задачи, поиск однозначного решения называют задачей вращения факторов. Сущностью факторного анализа является процедура вращения факторов, т.е. перераспределения дисперсии по определенному ме-

тоду. Вращение бывает *ортогональным* и *косоугольным*. При первом виде вращения каждый последующий фактор определяется так, чтобы максимизировать изменчивость, оставшуюся от предыдущих, поэтому факторы оказываются независимыми, некоррелированными друг от друга (к этому типу относится МГК). Второй вид – это преобразование, при котором факторы коррелируют друг с другом. Преимущество косоугольного вращения состоит в следующем: когда в результате его выполнения получаются ортогональные факторы, можно быть уверенным, что эта ортогональность действительно им свойственна, а не привнесена искусственно. Однако если цель ортогональных вращений – определение простой структуры факторных нагрузок, то целью большинства косоугольных вращений является определение простой структуры вторичных факторов, т.е. косоугольное вращение следует использовать в частных случаях. Поэтому ортогональное вращение предпочтительнее. Существует около 13 методов вращения в обоих видах, в статистической программе SPSS 10 доступны пять: три ортогональных, один косоугольный и один комбинированный, однако из всех наиболее употребителен ортогональный метод «*варимакс*». Метод «варимакс» максимизирует разброс квадратов нагрузок для каждого фактора, что приводит к увеличению больших и уменьшению малых значений факторных нагрузок. В результате простая структура получается для каждого фактора в отдельности. Факторные нагрузки повернутой матрицы могут рассматриваться как результат выполнения процедуры факторного анализа. Кроме того на основании значений этих нагрузок необходимо попытаться дать толкование отдельным факторам.

Если факторы найдены и истолкованы, то на последнем шаге факторного анализа, отдельным наблюдениям можно присвоить значения этих факторов, так называемые факторные значения. Таким образом для каждого наблюдения значения большого количества переменных можно перевести в значения небольшого количества факторов.

Главной проблемой факторного анализа является выделение и интерпретация главных факторов. При отборе компонент исследователь обычно сталкивается с существенными трудностями, так как не существует однозначного критерия выделения факторов, и потому здесь неизбежен субъективизм интерпретаций результатов. Существует несколько часто употребляемых критериев определения числа факторов. Некоторые из них являются альтернативными по отношению к другим, а часть этих критериев можно использовать вместе, чтобы один дополнял другой.

*Критерий Кайзера* или *критерий собственных чисел*. Этот критерий предложен Кайзером, и является, вероятно, наиболее широко используемым. Отбираются только факторы с собственными значениями, равными или большими 1. Это означает, что если фактор не выделяет дисперсию, эквивалентную, по крайней мере, дисперсии одной переменной, то он опускается.

*Критерий каменистой осыпи* или *критерий отсеивания*. Он является графическим методом, впервые предложенным психологом Кэттелом. Собственные значения возможно изобразить в виде простого графика. Кэттел предложил найти такое место на графике, где убывание собственных значений слева направо максимально замедляется. Предполагается, что справа от этой точки находится только «факториальная осыпь» – «осыпь» является геологическим термином, обозначающим обломки горных пород, скапливающиеся в нижней части скалистого склона. Однако этот критерий отличается высокой субъективностью и, в отличие от предыдущего критерия, статистически необоснован. Недостатки обоих критериев заключаются в том, что первый иногда сохраняет слишком много факторов, в то время как второй, напротив, может сохранить слишком мало факторов; однако оба критерия вполне хороши при нормальных условиях, когда имеется относительно небольшое число факторов и много переменных. На практике возникает важный во-

прос: когда полученное решение может быть содержательно интерпретировано? В этой связи предлагается использовать еще несколько критериев.

*Критерий значимости.* Он особенно эффективен, когда модель генеральной совокупности известна и отсутствуют второстепенные факторы. Но критерий непригоден для поиска изменений в модели и реализуем только в факторном анализе по методу наименьших квадратов или максимального правдоподобия.

*Критерий доли воспроизводимой дисперсии.* Факторы ранжируются по доле детерминируемой дисперсии, когда процент дисперсии оказывается несущественным, выделение следует остановить. Желательно, чтобы выделенные факторы объясняли более 80 % разброса. Недостатки критерия: во-первых, субъективность выделения, во-вторых, специфика данных может быть такова, что все главные факторы не смогут совокупно объяснить желательного процента разброса. Поэтому главные факторы должны вместе объяснять не меньше 50,1 % дисперсии.

*Критерий интерпретируемости и инвариантности.* Данный критерий сочетает статистическую точность с субъективными интересами. Согласно ему, главные факторы можно выделять до тех пор, пока будет возможна их ясная интерпретация. Она, в свою очередь, зависит от величины факторных нагрузок, т.е. если в факторе есть хотя бы одна сильная нагрузка, он может быть интерпретирован. Возможен и обратный вариант – если сильные нагрузки имеются, однако интерпретация затруднительна, от этой компоненты предпочтительно отказаться.

Практика показывает, что если вращение не произвело существенных изменений в структуре факторного пространства, это свидетельствует о его устойчивости и стабильности данных. Возможны еще два варианта: 1) сильное перераспределение дисперсии – результат выявления латентного фактора; 2) очень незначительное изменение (десятые, сотые или тысячные доли

нагрузки) или его отсутствие вообще, при этом сильные корреляции может иметь только один фактор, – однофакторное распределение. Последнее возможно, например, когда на предмет наличия определенного свойства проверяются несколько социальных групп, однако искомое свойство есть только у одной из них.

Факторы имеют две характеристики: объем объясняемой дисперсии и нагрузки. Если рассматривать их с точки зрения геометрической аналогии, то касательно первой отметим, что фактор, лежащий вдоль оси OX, может максимально объяснять 70 % дисперсии (первый главный фактор), фактор, лежащий вдоль оси OY, способен детерминировать не более 30 % (второй главный фактор). То есть в идеальной ситуации вся дисперсия может быть объяснена двумя главными факторами с указанными долями. В обычной ситуации может наблюдаться два или более главных факторов, а также остается часть неинтерпретируемой дисперсии (геометрические искажения), исключаемая из анализа по причине незначимости. Нагрузки, опять же с точки зрения геометрии, есть проекции от точек на оси OX и OY (при трех- и более факторной структуре также на ось OZ). Проекции – это коэффициенты корреляции, точки – наблюдения, таким образом, факторные нагрузки являются мерами связи. Так как сильной считается корреляция с коэффициентом Пирсона $R \geq 0{,}7$, то в нагрузках нужно уделять внимание только сильным связям. Факторные нагрузки могут обладать свойством *биполярности* – наличием положительных и отрицательных показателей в одном факторе. Если биполярность присутствует, то показатели, входящие в состав фактора, дихотомичны и находятся в противоположных координатах.

## 5.2. Пример применения факторного анализа в области социологии

Изложенный метод будет проиллюстрирован на примере анкеты, составленной в Институте социологии Университета Марбурга. На основе этой анкеты на двух гессенских металлургических предприятиях было произведено исследование отношения к иностранцам. Опрашиваемым предложили высказать свое отношение к следующим пятнадцати положениям:

1. Необходимо улучшить интеграцию иностранцев.
2. Необходимо мягче относиться к беженцам.
3. Деньги Германии должны быть потрачены на нужды страны.
4. Германия – это не служба социальной помощи для всего мира.
5. Необходимо стараться налаживать хорошие отношения друг с другом.
6. Права беженцев следует ограничить.
7. Немцы станут меньшинством.
8. Право беженцев необходимо охранять во всей Европе.
9. Враждебность к иностранцам наносит вред экономике Германии.
10. Сначала необходимо создать нормальные жилищные условия для немцев.
11. Мы ведь тоже практически везде являемся иностранцами.
12. Мультикультура означает мультикриминал.
13. В лодке нет свободных мест.
14. Иностранцы вон.
15. Интеграция иностранцев – это убийство нации.

Оценки ставились по семибалльной шкале: от полного несогласия (1) до полного согласия (7).

Сначала приводятся первичные статистики (табл. 5.1)

Таблица 5.1

Объясненная суммарная дисперсия
(Total Variance Explained)

| Component (Компоненты) | Initial Eigenvalues (Первичные собственные значения) | | | Rotation Sums of Squared Loadings (Повернутые суммы квадратов нагрузок) | | |
|---|---|---|---|---|---|---|
| | Total (Сумма) | % of Variance (% дисперсии) | Cumulative % (Совокупный %) | Total (Сумма) | % of Variance (% дисперсии) | Cumu-lative % (Совокупный %) |
| 1 | 5,146 | 34,308 | 34,308 | 3,466 | 23,105 | 23,105 |
| 2 | 1,945 | 12,970 | 47,278 | 2,536 | 16,907 | 40,013 |
| 3 | 1,415 | 9,433 | 56,711 | 2,505 | 16,698 | 56,711 |
| 4 | 0,990 | 6,601 | 63,312 | | | |
| 5 | 0,936 | 6,238 | 69,550 | | | |
| 6 | 0,760 | 5,068 | 74,617 | | | |
| 7 | 0,693 | 4,622 | 79,240 | | | |
| 8 | 0,612 | 4,083 | 83,323 | | | |
| 9 | 0,529 | 3,529 | 86,852 | | | |
| 10 | 0,473 | 3,151 | 90,004 | | | |
| 11 | 0,433 | 2,889 | 92,893 | | | |
| 12 | 0,339 | 2,262 | 95,1555 | | | |
| 13 | 0,301 | 2,007 | 97,161 | | | |
| 14 | 0,245 | 1,635 | 98,797 | | | |
| 15 | 0,181 | 1,203 | 100,000 | | | |

Extraction Method: Principal Component Analysis (Метод отбора: Анализ главных компонент).

При анализе можно увидеть, что три собственных фактора имеют значения, превосходящие единицу. Следовательно, для

анализа отобрано только три фактора. Первый фактор объясняет 34,308 % суммарной дисперсии, второй фактор – 12,97 % и третий фактор 9,433 %. Так как мы запретили вывод неповернутой матрицы факторов, то далее приводится повернутая матрица.

Алгоритм факторного анализа в программе SPSS включает следующие команды:

*Extraction Method: Principal Component Analysis* (Метод отбора: Анализ главных компонентов).

*Rotation Method: Varimax with Kaiser Normalization* (Метод вращения: Варимакс с нормализацией Кайзера).

*Rotation converged in 8 iterations* (Вращение осуществлено за 8 итераций).

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.

Как уже было сказано, эти факторные нагрузки следует понимать как корреляционные коэффициенты между переменными и факторами. Так переменная $a_1$ сильнее всего коррелирует с фактором 2, а именно, величина корреляции составляет 0,628, переменная $a_2$ также сильнее всего коррелирует с фактором 2 (0,657), переменная же $a_3$ коррелирует сильнее всего с фактором 3 (0,711) и т.д. В большинстве случаев включение отдельной переменной в один фактор, осуществляемое на основе коэффициентов корреляции, является однозначным. В исключительных случаях, к примеру, как в ситуации с переменной $a_7$, переменная может относиться к двум факторам одновременно. Могут быть также и переменные, в нашем примере $a_{11}$, которыми нельзя нагрузить ни один из отобранных факторов.

Если поступить так, как изложено выше, то варианты мнений, указанные вначале рассмотрения примера, можно отнести в следующем порядке к трем факторам:

***Фактор 1***

Германия – это не служба социальной помощи для всего мира.

Немцы станут меньшинством.

Мультикультура означает мультикриминал.

В лодке нет свободных мест.

Иностранцы вон.

Интеграция иностранцев – это убийство нации.

***Фактор 2***

Необходимо улучшить интеграцию иностранцев.

Необходимо мягче относиться к беженцам.

Необходимо стараться налаживать хорошие отношения друг с другом.

Права беженцев необходимо охранять во всей Европе.

Враждебность к иностранцам наносит вред экономике Германии.

Мы ведь тоже практически везде являемся иностранцами.

***Фактор 3***

Деньги Германии должны быть потрачены на нужды страны.

Права беженцев следует ограничить.

Немцы станут меньшинством.

Сначала необходимо создать нормальные жилищные условия для немцев.

Из-за равных по величине нагрузок, как для фактора 3, так и для фактора 1, положение «Немцы станут меньшинством» включено в оба фактора. Теперь мы подошли к последнему и решающему шагу факторного анализа: необходимо обнаружить и описать смысловую связь факторов. В рассматриваемом примере это можно сделать без особых усилий.

Первый фактор собрал все положения, враждебно настроенные по отношению к иностранцам. На основании позитивных корреляционных коэффициентов участвующих переменных с фактором и принимая во внимание полярность значений пере-

менных (большое значение означает полное согласие) большое значение фактора означает высокую враждебность к иностранцам.

Во второй фактор входят те положения, которые указывают на дружелюбное отношение к иностранцам. Большое значение фактора означает здесь доброжелательное отношение к иностранцам.

Во второй фактор вошли точки зрения, соответствующие осторожному отношению к иностранцам; в противоположность к первому фактору это не враждебные точки зрения, а по большей части социальные страхи (деньги, жилье в первую очередь для немцев и т.д.). Большое значение фактора указывает здесь на высокую степень социального сомнения.

В соответствии с порядком изложения эти три фактора можно кратко охарактеризовать при помощи следующих выражений: Враждебная позиция, Доброжелательная позиция и Социальные страхи. Однако столь явно, как в приведенном примере факторы удается объяснить не всегда. Если нет возможности провести вербальное объяснение факторов, то факторный анализ можно считать неудавшимся.

### 5.3. Значения факторов

Поскольку мы пожелали произвести расчет значений факторов, то в соответствии с тремя отобранными факторам были сгенерированы три новые переменные, названные fac1_1, fac2_1 и fac3_1, которые содержат вычисленные значения факторов. Если Вы просмотрите текущий файл после поведения факторного анализа, то сможете увидеть имеющие нормализованные значения факторов. По каждому из отобранных фактору для каждого опрошенного было рассчитано специальное факторное значение. Факторное значение, как правило, лежит в пределах от –3 до +3.

Рассмотрим факторную переменную fac1_1. Она включает следующие элементарные переменные: $a_4$, $a_{12}$, $a_{13}$, $a_{14}$ и $a_{15}$. В качестве метки для этого фактора мы выбрали выражение: "Враждебная позиция". Большое положительное значение фактора означает одобрение элементарных переменных, т.е. положений, входящих в этот фактор. Одобрение элементарных переменных, относящихся к первому фактору, тождественно ярко выраженным расистским взглядам. Для подтверждения этого факта рассмотрим два примера. Наблюдение 4 характеризуется очень низким факторным значением в переменной fac1_1. Оно равно –2,00455. В данном случае можно сделать заключение о том, что здесь не наблюдается расистская направленность или она очень слаба. Соответственно этому ведут себя и отдельные значения элементарных переменных ($a_4 = 2$, $a_{13} = 1$, $a_{14} = 1$, $a_{15} = 1$). Наблюдение 17, в отличие от наблюдения 4, характеризуется очень высоким положительным значением фактора, который равен 3,14801. Основываясь на этом значении, мы можем исходить из того, что здесь явно заметна экстремально-расистская позиция. Соответственно этому ведут себя и отдельные значения элементарных переменных ($a_4 = 7$, $a_{13} = 7$, $a_{14} = 7$, $a_{15} = 7$).

Рассмотрим факторную переменную fac2_1. К ней относятся элементарные переменные: $a_1$, $a_2$, $a_5$, $a_8$, $a_9$ и $a_{11}$. В качестве метки для этого фактора мы выбрали выражение: «Доброжелательная позиция». Большое положительное значение фактора означает полное согласие. Полное согласие соответствует дружелюбному отношению к иностранцам. И здесь рассмотрим два выборочных примера. Наблюдение 17 характеризуется очень малым значением фактора, которое составляет –3,32632. Основываясь на значении этого фактора можно сделать вывод, что едва ли в этом случае присутствует доброжелательное отношение к иностранцам. Соответственным образом ведут себя и отдельные значения элементарных переменных ($a_1 = 1$, $a_2 = 1$, $a_5 = 1$, $a_8 = 2$, $a_9 = 4$, $a_{11} = 6$). В наблюдении 17 и следовало ожи-

дать низкого значения фактора, так как здесь наблюдается высокое положительное факторное значение для факторной переменной fac1_1. В таком случае говорят, что существует отчетливая консистенция. По сравнению с предыдущим наблюдением, наблюдение 6 характеризуется очень высоким положительным значением факторной переменной fac2_1. Оно равно 1,23438. Исходя из значения фактора, можно сделать вывод, что существует сильное дружелюбное отношение к иностранцам. Соответственным образом ведут себя и отдельные значения элементарных переменных ($a_1 = 7$, $a_2 = 7$, $a_5 = 7$, $a_8 = 7$, $a_9 = 7$, $a_{11} = 7$).

В заключение рассмотрим факторную переменную fac3_1. К ней относятся элементарные переменные $a_3$, $a_6$, $a_7$ и $a_{10}$. В качестве метки для этого фактора мы выбрали выражение: «Социальные страхи». Большое положительное значение фактора означает одобрение элементарных переменных. Одобрение элементарных переменных тождественно ярко выраженным социальным страхам. Рассмотрим для доказательства этого факта два примера. Наблюдение 5 характеризуется очень низким значением факторной переменной fac3_1. Оно равно –1,66369. В этом случае наблюдаются очень слабые социальные страхи и едва ли на основании социальных страхов можно наблюдать враждебное отношение к иностранцам. Соответственно этому ведут себя и отдельные значения элементарных переменных ($a_3 = 5$, $a_6 = 2$, $a_7 = 2$, $a_{10} = 1$). Наблюдение 43 в отличие от наблюдения 5 характеризуется очень высоким положительным факторным значением. Оно равно 1,93125. В этом случае наблюдаются очень сильные социальные страхи. Соответственным образом ведут себя и отдельные значения элементарных переменных ($a_3 = 7$, $a_6 = 7$, $a_7 = 7$, $a_{10} = 7$). В файле *ausland.sav* находятся еще несколько дополнительных переменных, а именно:

Таблица 5.2

Дополнительные переменные факторного анализа

| | |
|---|---|
| ewv | Удовлетворенность собственным местом в экономических отношениях (1 = да, 2 = нет) |
| gebjg | Год рождения (1 = 1935 – 1949, 2 = 1941 – 1950, 3 = 1951 – 1960, 4 = 1961 – 1970) |
| geschl | Пол (1 = мужской, 2 = женский) |
| sozeng | Социально-политическая активность (1 = да, 2 = нет) |
| s+ellung | Занимаемая должность (1 = рабочий, 2 = специалист, 3 = служащий) |

Эти переменные можно использовать для того, чтобы устанавливать связи для факторных значений.

## Вопросы для самопроверки

1. Каковы основные понятия факторного анализа?
2. В какой последовательности осуществляется процедура факторного анализа?
3. Каким образом осуществляется интерпретация вклада отдельных факторов в суммарную дисперсию?
4. Какую роль играет вращение факторов и каковы правила интерпретации повернутых факторов?
5. Каким правилам надо следовать при логическом объяснении результатов факторного анализа?

# Глава 6. КЛАСТЕРНЫЙ АНАЛИЗ



## 6.1. Возможности кластерного анализа

*Кластерный анализ* – это группа методов, используемых для классификации объектов или событий в относительно гомогенные (однородные) группы, которые называют *кластерами* (clusters). Синонимы: *классификационный анализ* (Classification Analysis), *численная таксономия* (Numerical Taxonomy). Если процедура факторного анализа сжимает матрицу признаков в матрицу с меньшим числом переменных, то кластерный анализ дает нам группы единиц анализа, т.е. выполняет классификацию объектов. Если в факторном анализе группируются столбцы *матрицы данных*, то в кластерном анализе группируются *строки*.

Если данные понимать как точки в признаковом пространстве, то задача кластерного анализа состоит в выделении «сгущений точек», в разбиении совокупности на однородные подмножества объектов.

*Статистики кластерного анализа:*

- протокол объединения (Agglomeration Scheduling);
- кластерный центроид (Cluster Centroid);
- кластерные центры, зерна (Clusters Centers);
- принадлежность кластеру (Cluster Membership);
- древовидная диаграмма (Dendrogram);
- расстояния между центрами (Distance between Centers);
- сосульчатая диаграмма (Icicle Diagram);
- матрица сходства (Distance Coefficient Matrix).

При проведении кластерного анализа обычно определяют расстояние на множестве объектов; алгоритмы кластерного анализа формулируют в терминах этих расстояний. Мер близости и расстояний между объектами существует великое множество. Их выбирают в зависимости от цели исследования. В частности, евклидово расстояние лучше использовать для количественных переменных, расстояние хи-квадрат – для исследования частотных таблиц, имеется множество мер для бинарных переменных.

Кластерный анализ является описательной процедурой, он не позволяет сделать никаких статистических выводов, но дает возможность провести своеобразную разведку – изучить структуру совокупности [6].

*Стадии кластерного анализа:*

- формулировка проблемы;
- выбор способа измерения расстояния;
- выбор метода кластеризации;
- принятие решения о количестве кластеров;
- интерпретация и профилирование кластеров;
- оценка достоверности кластеризации.

А. **Формулировка проблемы**. Имеются данные опроса 20 респондентов, отвечавших на 6 вопросов. Ответ на каждый вопрос по 7-балльной шкале, от не согласен, до полностью согласен. Требуется провести классификацию объектов, схожих по набору признаков.

Б. **Выбор меры расстояния или меры сходства**. Для каждого типа данных существует несколько способов измерения расстояния или определения меры сходства объектов. Наиболее часто используются для интервальных данных:

- евклидово расстояние (Euclidian Distance):

$$d(X,Y)=\sqrt{\sum_{i=1}^{m}(X_i-Y_i)^2};$$

– квадрат евклидова расстояния (Squared Euclidian distance)

$$d(X,Y)=\sum_{i=1}^{m}(X_i-Y_i)^2.$$

В. **Выбор метода кластеризации**. Мы можем осуществить выбор между следующими методами:

- среднее расстояние между кластерами (Between-groups linkage);
- среднее расстояние между всеми объектами пары кластеров с учетом расстояний внутри кластеров (Within-groups linkage);
- расстояние между ближайшими соседями (Nearest neighbor);
- расстояние между самыми далекими соседями (Furthest neighbor);
- расстояние между центрами кластеров (Centroid clustering);
- метод медиан (Median clustering);
- метод Варда (Ward's method).

Г. **Принятие решения о числе кластеров**. При принятии данного решения руководствуются следующими соображениями.

1. Руководствуются практическими и теоретическими соображениями. Исходя из цели исследования, например, может быть необходимо три кластера.

2. В иерархической кластеризации в качестве критерия можно использовать расстояния. При этом обращаем внимание на коэффициент в протоколе объединения.

3. В иерархической кластеризации можно воспользоваться графиком зависимости отношения суммарной внутригрупповой дисперсии к межгрупповой дисперсии от числа кластеров. Скачок указывает на число кластеров.

4. Размеры кластеров должны быть достаточно выразительными.

Д. **Интерпретация и профилирование кластеров**. Эта стадия включает проверку кластерных центроидов. *Центроиды* – это средние значения объектов, содержащихся в кластере, по каждой из переменных. Они позволяют описывать кластеры.

Е. **Оценка достоверности кластеризации**. Для выполнения этой задачи следуйте перечисленным ниже рекомендациям:

1. Выполняйте кластерный анализ на основании одних и тех же данных, но с использованием различных способов измерения расстояния. Сравните результаты, полученные на основе разных мер расстояния, чтобы определить, насколько совпадают полученные результаты.

2. Используйте разные методы кластерного анализа и сравните полученные результаты.

3. Разбейте данные на две равные части случайным образом. Выполните кластерный анализ отдельно для каждой половины. Сравните кластерные центроиды двух подвыборок.

4. Случайным образом удалите некоторые переменные. Выполните кластерный анализ по сокращенному набору переменных. Сравните результаты с полученными на основе полного набора переменных.

5. В неиерархической кластеризации решение может зависеть от порядка случаев в наборе данных. Выполните анализ несколько раз, меняя порядок случаев, до получения стабильного решения.

## 6.2. Иерархический кластерный анализ в SPSS

Процедура иерархического кластерного анализа в SPSS предусматривает группировку как объектов (строк матрицы данных), так и переменных (столбцов). Можно считать, что в последнем случае роль объектов играют переменные, а роль переменных – столбцы.

Этот метод реализует иерархический агломеративный алгоритм. Его смысл заключается в следующем. Перед началом кластеризации все $N$ объектов считаются отдельными кластерами, которые в ходе алгоритма объединяются. Вначале выбирается пара ближайших кластеров, которые объединяются в один кластер. В результате количество кластеров становится равным $N-1$. Процедура повторяется, пока все классы не объединятся. На любом этапе объединение можно прервать, получив нужное число кластеров. Таким образом, результат работы алгоритма агрегирования определяют способы вычисления расстояния между объектами и определения близости между кластерами.

Для определения расстояния между парой кластеров могут быть сформулированы различные подходы, для чего в SPSS предусмотрены методы, определяемые на основе расстояний между объектами:

- среднее расстояние между кластерами (Between-groups linkage);
- среднее расстояние между всеми объектами пары кластеров с учетом расстояний внутри кластеров (Within-groups linkage);
- расстояние между ближайшими соседями – ближайшими объектами кластеров (Nearest neighbour);
- расстояние между самыми далекими соседями (Furthest neighbour);
- расстояние между центрами кластеров (Centroid clustering), или «центроидный» метод. Недостатком этого метода является то, что центр объединенного кластера вычисляется как среднее центров объединяемых кластеров, без учета их объема;
- метод медиан – тот же «центроидный» метод, но центр объединенного кластера вычисляется как среднее всех объектов (Median clustering);
- метод Варда (Ward's method). В качестве расстояния между кластерами берется прирост суммы квадратов расстояний объектов до центров кластеров, получаемый в результате их объединения.

*Расстояния и меры близости между объектами.* У нас нет возможности сделать полный обзор всех коэффициентов, поэтому остановимся лишь на характерных расстояниях и мерах близости для определенных видов данных. Меры близости отличаются от расстояний тем, что они тем больше, чем более похожи объекты.

Пусть имеются два объекта $X = (X_1, \ldots, X_m)$ и $Y = (Y_1, \ldots, Y_m)$. Используя эту запись для объектов, определим основные виды расстояний, используемых в процедуре CLUSTER:

– евклидово расстояние (Euclidian distance)

$$d(X,Y) = \sqrt{\sum_{i=1}^{m}(X_i - Y_i)^2};$$

– квадрат евклидова расстояния (Squared Euclidian distance)

$$d(X,Y) = \sum_{i=1}^{m}(X_i - Y_i)^2$$

(евклидово расстояние и его квадрат целесообразно использовать для анализа количественных данных);

– мера близости – коэффициент корреляции

$$S(X,Y) = (\sum_{i=1}^{m} Z_{X_i} Z_{Y_i})/(m-1),$$

где $Z_{X_i}$ и $Z_{Y_i}$ компоненты стандартизованных векторов $X$ и $Y$ (эту меру целесообразно использовать для выявления кластеров переменных, а не объектов);

– расстояние хи-квадрат получается на основе таблицы сопряженности, составленной из объектов $X$ и $Y$, которые, предположительно, являются векторами частот.

| | | | | |
|---|---|---|---|---|
| $X$ | $X_1$ | ... | $X_m$ | $X.$ |
| $Y$ | $Y_1$ | ... | $Y_m$ | $Y.$ |
| $X+Y$ | $X_1+Y_1$ | ... | $X_m+Y_m$ | $X.+Y.$ |

– расстояние $\varphi$-квадрат является расстоянием хи-квадрат, нормированным на число объектов в таблице сопряженности, представляемой строками $X$ и $Y$, т.е. на корень квадратный из $N = X. + Y.$;

– в иерархическом кластерном анализе в SPSS также имеется несколько видов расстояний для бинарных данных (векторы $X$ и $Y$ состоят из нулей и единиц, обозначающих наличие или отсутствие определенных свойств объектов). Наиболее естественными из них, по-видимому, являются евклидово расстояние и его квадрат.

*Стандартизация*. Непосредственное использование переменных в анализе может привести к тому, что классификацию будут определять переменные, имеющие наибольший разброс значений. Поэтому применяются следующие виды стандартизации:

– $Z$-шкалы ($Z$-Scores). Из значений переменных вычитается их среднее и эти значения делятся на стандартное отклонение.

– Разброс от –1 до 1. Линейным преобразованием переменных добиваются разброса значений от –1 до 1.

– Разброс от 0 до 1. Линейным преобразованием переменных добиваются разброса значений от 0 до 1.

– Максимум 1. Значения переменных делятся на их максимум.

– Среднее 1. Значения переменных делятся на их среднее.

– Стандартное отклонение 1. Значения переменных делятся на стандартное отклонение.

– Кроме того, возможны преобразования самих расстояний, в частности, можно расстояния заменить их абсолютными значениями, это актуально для коэффициентов корреляции. Можно также все расстояния преобразовать так, чтобы они изменялись от 0 до 1.

Таким образом, работа с кластерным анализом может превратиться в увлекательную игру, связанную с подбором метода

агрегирования, расстояния и стандартизации переменных с целью получения наиболее интерпретируемого результата. Желательно только, чтобы это не стало самоцелью и исследователь получил действительно необходимые содержательные сведения о структуре данных.

Процесс агрегирования данных может быть представлен графически деревом объединения кластеров (Dendrogramm) либо «сосульковой» диаграммой (Icicle). Но подробнее о процессе кластеризации можно узнать по протоколу объединения кластеров (Schedule).

### 6.3. Быстрый кластерный анализ

Процедура иерархического кластерного анализа хороша для малого числа объектов. Ее преимущество в том, что каждый объект можно, образно говоря, пощупать руками. Но эта процедура не годится для огромных социологических данных из-за трудоемкости агломеративного алгоритма и слишком больших размеров дендрограмм.

Здесь наиболее приемлем быстрый алгоритм, носящий название метода *k-средних*. Он реализуется в пакете командой QUICK CLUSTER или командой меню *k*-means.

Алгоритм заключается в следующем: выбирается заданное число *k-точек* (объектов из данных), и на первом шаге эти точки рассматриваются как центры кластеров. Каждому кластеру соответствует один центр. Объекты распределяются в кластерах по такому принципу: каждый объект относится к кластеру с ближайшим к этому объекту центром. Таким образом, все объекты распределились по *k-кластерам*.

Затем заново вычисляются центры этих кластеров, которыми с этого момента считаются покоординатные средние кластеров. После этого опять перераспределяются объекты. Вычисле-

ние центров и перераспределение объектов происходит до тех пор, пока не стабилизируются центры.

Синтаксис команды:

```
QUICK CLUSTER W3d1 TO W3D6/CRITERIA
CLUSTERS(3) /MISSING = PAIRWISE /SAVE
CLUSTER(SAVCLU) /PRINT ANOVA.
```

За именем команды располагаются переменные, по которым происходит кластеризация. Параметр/CRITERIA CLUSTERS задает в скобках число кластеров. Подкомандой /SAVE CLUSTER можно сохранить полученную классификацию в виде переменной, имя которой дается в скобках. Подкоманда /PRINT ANOVA позволяет провести по каждой переменной одномерный дисперсионный анализ – сравнение средних в кластерах. Последний имеет лишь описательное значение и позволяет определить переменные, которые не оказывают никакого влияния на классификацию.

Команда использует только евклидово расстояние. При этом часть переменных может иметь неопределенные значения, расстояния до центров определяются по определенным значениям. Для использования такой возможности следует употребить подкоманду /MISSING = PAIRWISE.

Часто переменные имеют разный диапазон изменений, так как измерены они в различных шкалах или просто из-за того, что характеризуют разные свойства объектов (например, рост и вес, килограммы и граммы). В этих условиях основное влияние на кластеризацию окажут переменные, имеющие большую дисперсию. Поэтому перед кластеризацией полезно стандартизовать переменные. К сожалению, в «быстром» кластерном анализе средства стандартизации не предусмотрены непосредственно, как в процедуре иерархического кластерного анализа.

Для этого можно использовать команду DESCRIPTIVE. Напомним, что подкоманда /SAVE в ней позволяет автоматически сохранить стандартизованные переменные. Кроме того, хо-

рошие средства стандартизующих преобразований шкал дает команда RANK.

В выдаче распечатываются центры кластеров (средние значения переменных кластеризации для каждого кластера), получаемые на каждой итерации алгоритма. Однако для нас полезна лишь часть выдачи, помеченная текстом *Final centres*.

Интерпретация кластеров осуществляется на основе сравнения средних значений, выдаваемых процедурой, а также исследования сохраненной переменной средствами статистического пакета.

### Вопросы для самопроверки

1. Каковы возможности применения кластерного анализа?
2. Когда применяется иерархический кластерный анализ?
3. В чем причина ограниченности использования иерархического кластерного анализа?
4. Какие возможности демонстрирует быстрый кластерный анализ?

## ЗАКЛЮЧЕНИЕ

В первых конкретных социологических исследованиях при анализе социальных данных были взяты на вооружение простейшие математические и статистические методы – методы средних чисел, метод аналитических группировок, индексовый метод анализа, т.е. методы так называемой дескриптивной статистики. По мере развития конкретных социологических исследований применялись все более точные математические методы анализа социальных данных. Оперирование с большими массивами социальной информации привело к проблеме использования вычислительной техники. Социологи столкнулись с необходимостью измерения качественных социальных переменных и моделированием социальных процессов и явлений.

В настоящее время перед социологией стоит задача разработки методов измерения самых различных систем социальных, создания комплексных математических социально-экономических моделей и т.д.

## ГЛОССАРИЙ

*Временных рядов анализ (time-series analysis)* – один из методов прогнозирования спроса; основан на разбивке данных об объеме продаж в прошлом на компоненты, характеризующие тренды, циклические (например, циклы деловой активности) и сезонные колебания, а также случайные изменения. В.р.а. предполагает выявление причин измения спроса в прошлом для прогнозирования спроса в будущем.

*Дискриминантный анализ (discriminant analysis)* – статистический метод, используемый для прогнозирования вероятности какого-либо события. Относится к методам классификации с обучением. Используется для разделения респондентов в различающиеся между собой группы на основе некоторых характеристик. Обычно зависимая переменная номинальная или порядковая, а независимые переменные (предикторы) – метрические (интервальные).

*Дисперсионный анализ (analysis of variance)* – метод статистического анализа, позволяющий определить достоверность гипотезы о различиях в средних значениях на основании сравнения дисперсий распределений. Этот метод имеет смысл только лишь для интервальных переменных с наложенными дополнительными ограничениями.

*Дисперсия (variance)* – математическое ожидание (среднее) квадрата отклонения случайной величины от ее математического ожидания (среднего).

*Кластерный анализ (cluster analysis)* – совокупность методов, позволяющих классифицировать многомерные наблюдения, каждое из которых описывается неким набором переменных. Целью кластерного анализа является образование групп схожих между собой объектов, которые принято называть кластерами. Слова кластер английского происхождения (cluster), переводит-

ся как сгусток, пучок, группа. Родственные понятия, используемые в литературе, – класс, таксон, сгущение. В отличие от комбинационных группировок кластерный анализ приводит к разбиению на группы с учетом всех группировочных признаков одновременно.

*Корреляция (correlation)* – показатель наличия линейной зависимости двух переменных.

*Медиана (median)* – значение, которое делит ряд возрастающих (или убывающих) значений на две части, равные по числу значений. Если количество значений ряда нечетно ($2n$ + 1), то медиана равна $X_n$ + 1, при четном ($2n$) медиана равна $(X_n + X_n + 1)/2$.

*Переменная (variable)* – единица анализа данных в статистике.

*Процентиль (percentile)* – значение переменной, не более которого указала соответствующая доля респондентов. Так, 75-й процентиль соответствует такому значению, не более которого указали 75 % респондентов.

*Регрессионный анализ (regression analysis)* – статистический метод установления зависимости между независимыми и зависимыми переменными. Регрессионный анализ на основе построенного уравнения регрессии определяет вклад каждой независимой переменной в изменение изучаемой (прогнозируемой) зависимой переменной величины. Выделяют два вида регрессионного анализа – парный регрессионный анализ и анализ на основе множественной регрессии.

*Среднее квадратичное отклонение (mean square deviation, standard deviation)* – называют корень квадратный из дисперсии.

*Факторный анализ (factor analysis)* – совокупность методов, которые на основе реально существующих связей признаков (или объектов) позволяют выявлять латентные (или скры-

тые) обобщающие характеристики структуры и механизма развития изучаемых явлений и процессов. Понятие латентности в определении ключевое. Оно означает неявность характеристик, раскрываемых при помощи методов факторного анализа. Толчком для развития методов факторного анализа изначально послужили работы в области психологии и практически все первые работы по факторному анализу были напечатаны в журналах «Психология» и «Психометрика». Позже методы факторного анализа стали активно использоваться в социологических исследованиях, медицине, военной науке и экономике. Необходимым условием расширения сферы применения факторного анализа является математизация исследуемого процесса и применение компьютерных программ обработки данных.

*Частот таблицы (frequency tables)* – таблицы частот или одновходовые таблицы представляют собой простейший метод анализа категориальных (номинальных) переменных. Часто их используют как одну из процедур разведочного анализа, чтобы просмотреть, каким образом различные группы данных распределены в выборке.

*Шкала номинальная, шкала наименований* – используется для измерения объектов, обозначенных наименованием – пол, регион проживания, принадлежность к политической партии. Отношения между шкальными значениями – отношения неравенства, различия. Допустимые статистические расчеты – процент, доля, мода.

*Шкала порядковая* – измеряет уровень согласия с утверждением, степень удовлетворенности. Отношения между шкальными значениями – иерархия признаков, сравнение, отношение неравенства (больше, меньше, равно, не равно). Допустимые статистические расчеты – процент, доля, мода, медиана.

*Шкала интервальная* – измеряет в интервальных значениях возраст, доход. Отношения между шкальными значениями – равенство, неравенство, больше, меньше, больше на, меньше на, отношения между интервалами. Допустимые статистические расчеты – процент, доля, мода, медиана, среднее арифметическое, дисперсия, среднеквадратическое отклонение.

*Шкала отношений* – измеряет стаж работы, возраст, доход. Отношения между шкальными значениями – равенство, неравенство, больше, меньше, больше на, меньше на, больше в, меньше в. Допустимые статистические расчеты – процент, доля, мода, медиана, среднее арифметическое, дисперсия, среднеквадратическое отклонение.

## БИБЛИОГРАФИЧЕСКИЙ СПИСОК

*Основная литература*


1. Горшков М. К. Прикладная социология : методология и методы : учеб. пособие / М. К. Горшков, Ф. Э. Шереги. – М. : Альфа-М ; ИНФРА-М, 2011. – 414 с.
2. Девятко И. Ф. Методы социологического исследования: учеб. пособие / И. Ф. Девятко. – 6-е изд., испр. и доп. – М. : Книжный дом «Университет», 2010. – 208 с.
3. Ядов В. А. Стратегия социологического исследования : описание, объяснение, понимание социальной реальности : учеб. пособие / В. А. Ядов. – 3-е изд., испр. – М. : Омега-Л, 2008. – 567 с.


*Дополнительная литература*


4. Бююль А. SPSS: искусство обработки информации : пер. с нем. / А. Бююль, П. Цефель. – М. : DiaSoft, 2005. – 608 с.
5. Доугерти К. Введение в эконометрику : пер. с англ / К. Доугерти. – М. : ИНФРА-М, 1999. – 402 с.
6. Крыштановский А. О. Анализ социологических данных / А. О. Крыштановский. – М. : ГУ ВШЭ, 2006. – 283 с.
7. Мангейм Дж. Б. Политология. Методы исследования : пер. с англ. / Дж. Б. Мангейм, Р. К. Рич ; предисл. А. К. Соколова. – М. : Весь Мир, 1997. – 544 с.
8. Наследов А. Д. SPSS 15: профессиональный статистический анализ данных / А. Д. Наследов. – СПб. : Питер, 2008. – 416 с.
9. Наследов А. Д. Математические методы психологического исследования: анализ и интерпретация данных / А. Д. Наследов. – СПб. : Речь, 2008. – 392 с.

10. Пациорковский В. В. SPSS для социологов : учеб. пособие / В. В. Пациорковский, В. В. Пациорковская. – М. : ИСЭПН, 2005. – 433 с.
11. Татарова Г. Г. Типологический анализ в социологии / Г. Г. Татарова – М. : Наука, 1993. – 103 с.
12. Татарова Г. Г. Методология анализа данных в социологии (введение) : учеб. пособие для вузов / Г. Г. Татарова. – М. : Стратегия, 1998. – 224 с.
13. Толстова Ю. Н. Измерение в социологии : курс лекций / Ю. Н. Толстова. – М. : ИНФРА-М, 1998. – 224 с.
14. Толстова Ю. Н. Логика математического анализа социологических данных / Ю. Н. Толстова. – М. : Наука, 1991. – 285 с.
15. Толстова Ю. Н. Математико-статистические модели в социологии: математическая статистика для социологов : учеб. пособие / Ю. Н. Толстова. – М. : ГУ ВШЭ, 2007. – 244 с.
16. Тюрин Ю. Н. Анализ данных на компьютере : учеб. пособие / Ю. Н. Тюрин, А. А. Макаров. – М. : Форум, 2008. – 528 с.
17. Шляпентох В. Э. Проблемы качества социологической информации: достоверность, репрезентативность, прогностический потенциал / В. Э. Шляпентох. – М. : Центр соц. прогнозирования, 2006. – 644 с.
18. Электронный учебник по статистике. Москва, StatSoft. WEB: www.statsoft.ru/home/textbook/default.htm
19. http://psystat.at.ua/publ/1-1-0-29
20. http://cito-web.yspu.org/link1/metod/met125/node35.htm

## ПРИЛОЖЕНИЕ

### Тематика курсовых проектов

1. Страхи и опасения россиян на современном этапе (на основе вторичного анализа данных исследования «Чего опасаются россияне?», проведенного Центром социального прогнозирования и маркетинга, г. Москва, 2008 г.)

2. Социологический анализ процессов образования и распада браков и союзов россиян (на основе вторичного анализа данных исследования «Родители и дети, мужчины и женщины в семье и обществе», проведенного ЗАО «Демоскоп», г. Москва, 2004 г.)

3. Динамика доверия к правительству в разных странах в период с 1996 по 2006 год (на основе вторичного анализа социологических данных модуля «Роль правительства» программы Международного социального исследования (ISSP))

4. Отношение к религии в разных странах (на основе вторичного анализа социологических данных модуля «Религия» программы Международного социального исследования (ISSP))

5. Спорт и его роль в жизни граждан разных стран (на основе вторичного анализа социологических данных модуля «Досуг и спорт» программы Международного социального исследования (ISSP))

6. Труд как ценность в представлениях граждан разных стран (на основе вторичного анализа социологических данных модуля «Отношение к труду» программы Международного социального исследования (ISSP))

7. Социальный портрет потребителя в разных регионах страны (на основе вторичного анализа данных исследования «Типология потребителя», проведенного ГФК Русь, г. Москва, 2007 г.)

8. Проблемы преступности несовершеннолетних в России (на основе вторичного анализа данных исследования «Детская преступность», проведенного АНО «Левада-центр», г. Москва, 2004 г.)

9. Отношение россиян к благотворительности и благотворительным организациям (на основе вторичного анализа данных исследования «Благотворительность в России: осведомленность населения», проведенного АНО «Левада-центр», г. Москва, 2006 г.)

10. Плюсы и минусы бюрократизации властных органов глазами россиян (на основе вторичного анализа данных исследования «Бюрократия и власть в новой России», проведенного Центром социального прогнозирования и маркетинга, г. Москва, 2005 г.)

11. Отношение россиян к властным структурам различных уровней (на основе вторичного анализа данных исследования «Бюрократия и власть в новой России», проведенного Центром социального прогнозирования и маркетинга, г. Москва, 2005 г.)

12. Включенность женщин в общественно-политические процессы России (на основе вторичного анализа данных исследования «Женщина новой России», проведенного Центром социального прогнозирования и маркетинга, г. Москва, 2002 г.)

13. Проблемы и опасения российских женщин (на основе вторичного анализа данных исследования «Женщина новой России», проведенного Центром социального прогнозирования и маркетинга, г. Москва, 2002 г.)

14. Экономические и социальные стратегии среднего класса в России (на основе вторичного анализа данных исследования «Экономические и социальные стратегии среднего класса (базовая выборка)», проведенного Московским центром Карнеги, г. Москва, 2000 г.)

15. Ценностные приоритеты российской молодежи (на основе вторичного анализа данных исследования «Молодежь новой России: образ жизни и ценностные приоритеты (основная выборка)», проведенного Центром социального прогнозирования и маркетинга, г. Москва, 2007 г.)

16. Политические предпочтения россиян (на основе вторичного анализа данных экспресс-опросов ВЦИОМ, проведенных в 2007–2008 гг.)

17. Потребительское поведение россиян (на основе вторичного анализа данных четырех волн повторяющегося исследования «Российских индекс целевых групп (TGI)», проведенного COMCON, 2000 г.)

18. Организация труда на российских предприятиях глазами работников (на основе вторичного анализа данных опросов работников, проведенных ВЦИОМ с 1999 по 2002 гг.)

19. Институт семьи и его значимость в разных странах (на основе вторичного анализа социологических данных модуля «Семья и изменение гендерных ролей» программы Международного социального исследования (ISSP))

20. Гендерные нормы и стереотипы в разных странах: сходства и отличия (на основе вторичного анализа социологических данных модуля «Семья и изменение гендерных ролей» программы Международного социального исследования (ISSP))

21. Социально-экономическое положение на Южном Кавказе: оценка глазами жителей (на основе вторичного анализа социологических данных исследования «Социально-экономическая оценка положения домохозяйств на Южном Кавказе», проведенного Кавказским исследовательским ресурсным центром (CRRC), 2004 г.)

22. Отношение к экологии и важности ее охраны в разных странах (на основе вторичного анализа социологических данных модуля «Экология» программы Международного социального исследования (ISSP))

23. Перспективы развития России глазами россиян (на основе вторичного анализа социологических данных мониторинга социально-экономических перемен, проводимого АНО «Левада-Центр», 2008 г.)

Учебное издание

Добрина Ольга Александровна

# АНАЛИЗ ДАННЫХ В СОЦИОЛОГИИ

Учебное пособие

ISBN 978-5-7795-0666-3

Темплан 2013 г.

Редактор Г.К. Найденова

Санитарно-эпидемиологическое заключение
№ 54.НС.05.953.П.006252.06.06 от 26.06.2006 г.
Подписано к печати 17.12.2013. Формат 60×84 1/16 д.л.
Гарнитура Таймс. Бумага офсетная. Ризография.
Объем 5,9 уч.-изд.л.; 6,5 п.л. Тираж 85 экз. Заказ № 387

Новосибирский государственный
архитектурно-строительный университет (Сибстрин)
630008, Новосибирск, ул. Ленинградская, 113

Отпечатано мастерской оперативной полиграфии
НГАСУ (Сибстрин)