ДИСЦИПЛИНА Многомерные статистические методы (ч.1)

ИНСТИТУТ Технологий управления

КАФЕДРА Статистики и математических методов в управлении

ВИД УЧЕБНОГО МАТЕРИАЛА Лекции

ПРЕПОДАВАТЕЛЬ Есенин М.А.

СЕМЕСТР 5

План лекций МСМ 1 сем:

Раздел 1. Введение в статистический анализ данных

Лекция 1. Теоретические основы статистического анализа данных

1.Общие подходы к анализу данных

2.Классификация статистических данных

Лекция 2. Анализ одномерных статистических данных

1.Предварительный анализ категориальных данных

2.Предварительный анализ количественных данных

Лекция 3. Дискриптивные статистики и процедуры разведочного анализа данных

1.Основные описательные статистические характеристики

2.Диаграммы «ящик с усами» (box and whisker plots), нормирование и унификация данных

Раздел 2. Статистический анализ совокупностей

Лекция 4 Методы обнаружения засорения выборки

1.Визуальное обнаружение засорения выборки

2.Статистические критерии обнаружения засорения выборки

Лекция 5.Задачи статистического сравнения

1.Статистическое сравнение параметров со стандартом и проверка гипотез о законе распределения

2.Статистическое сравнение параметров нескольких совокупностей

Лекция 6. Статистический анализ многомерных совокупностей

1.Сравнение многомерных совокупностей

2.Основы дисперсионного анализа

Раздел 3. Статистический анализ взаимосвязи

Лекция 7.Многомерный линейный корреляционый анализ

1.Оценка взаимосвязи между парами признаков

2.Оценка множественной взаимосвязи между признаками

Лекция 8. Методы оценки зависимостей многомерных совокупностей

1.Многомерный линейный регрессионный анализ

Раздел 1. Введение в статистический анализ данных

Лекция 1. Теоретические основы статистического анализа данных

1.Общие подходы к анализу данных

Многомерный анализ данных, опирается на множество подходов и алгоритмов, используется практически во всех областях науки и деятельности общества. Он осуществляется исследователем с целью формирования определенных представлений о характере анализируемого явления.

В процессе анализа данных исследователь чаще всего пытается их сжать, стремясь потерять при этом как можно меньше заложенной в них полезной информации. Делается это обычно с помощью статистических методов. Сокращение объема данных достигается за счет применения двух взаимно дополняющих принципов: выборочного метода и свертки информации. Первый из них декларирует отказ от всей совокупности данных (генеральной совокупности) в пользу специально организованной ее части — выборки, а второй заменяет всю выборку ее несколькими характеристиками, например средней арифметической, дисперсией и т.д., а также результатами применения методов исследования зависимостей, снижения размерностей и классификации.

Развитие теории и практики статистических методов обработки данных идет в двух параллельных направлениях. Одно из них представлено методами, предусматривающими возможность вероятностной интерпретации данных, использования вероятностных моделей для построения и выбора наилучших методов статистической обработки.

Эти методы обычно называют вероятностно-статистическими. Они предполагают, что вероятностные модели адекватны явлениям, изучаемым с их помощью. В этом случае адекватность получаемых выводов основывается

на строго доказанных математических результатах, дающих возможность также устанавливать точность получаемых выводов.

Другое направление представлено логико-алгебраическими методами анализа данных, которые не предполагают вероятностных моделей изучаемых явлений. Эти исходные данные, подлежащие статистической обработке, не могут интерпретироваться как выборка из генеральной совокупности. Отсюда следует неправомерность использования вероятностных моделей при выборе методов статистической обработки данных и вероятностной интерпретации полученных результатов.

При этом процедуры свертки информации не всегда допускают формального алгоритмического подхода. Такое понимание термина многомерного анализ данных востребовано в социально-экономических приложениях и нашло отражение в работах многих современных статистиков и специалистов по обработке данных. Эти методы обработки статистических данных не основываются на строго доказанных математических результатах и, как следствие, не позволяют оценивать точность получаемых с их помощью выводов. При решении таких задач наилучший метод обработки данных обычно выбирается с помощью оптимизации некоторого функционала качества, задаваемого из содержательных соображений.

Естественно, что при этом проблема обоснованности выводов, получаемых с помощью методов многомерного анализа, требует дополнительного внимания, поэтому особую значимость приобретает вопрос согласования содержания задачи и используемых математических методов. Однако даже в случаях применения вероятностно-статистических методов анализа данных, когда исследователь имеет возможности опираться на формальные критерии, проверка адекватности вероятностной модели изучаемому явлению также должна опираться и на содержательные соображения. При этом методы анализа данных в обоих рассмотренных случаях могут служить средством получения фундаментальных знаний, выявления неизвестных ранее закономерностей.

2. Классификация статистических данных

В процессе управления экономическими и техническими системами статистические методы позволяют выработать обоснованные решения, сочетающие интуицию и опыт специалиста с тщательным анализом имеющейся информации. И с каждым годом интерес к статистической обработке данных неуклонно возрастает, так как объемы окружающей нас информации угрожающе увеличиваются и без грамотной их обработки и представления, исследования закономерностей невозможно правильно принимать решения на их основе.

При этом анализ данных может проводиться с целью:

• анализа и отображения конкретной собранной информации — в этом случае говорят о статистическом описании, *описательной* (*дескриптивной*)

статистике (*descriptive statistics*);

• описания всего класса явлений по имеющимся выборочным данным, характеризующим только часть этого класса. Эти задачи относятся к *аналитической статистике*.

Как правило, любое статистическое исследование начинается с дескриптивной статистики, а потом уже при необходимости углубляется аналитической.

Под *данными* (*data*) в статистике понимают совокупность сведений, зафиксированных на определенном носителе в форме, пригодной для их постоянного хранения, передачи и обработки. В статистике для характеристики изучаемых объектов используются различные типы данных, и к каждому типы применимы свои методы их обработки. Основные критерии классификации наборов статистических данных следующие:

1) по числу переменных, характеризующих объект исследования, различают *одномерные*, *двумерные и многомерные данные*;

2) по наличию или отсутствию упорядочения во времени различают *пространственные*, *временные и пространственно-временные данные*;

3) по типу шкалы измерения каждого признака различают *количественные* (числовые) признаки, которые делятся на *дискретные* и *непрерывные*,

и *категориальные* (качественные) признаки, которые делятся на *номинальные* и *порядковые*;

4) по способу получения данные делятся на *первичные* — если информация собиралась специально для данного анализа и *вторичные* — если используется информация из других источников, собранная для других целей.

Полная схема классификации данных по названным критериям представлена на рис. 1.1.

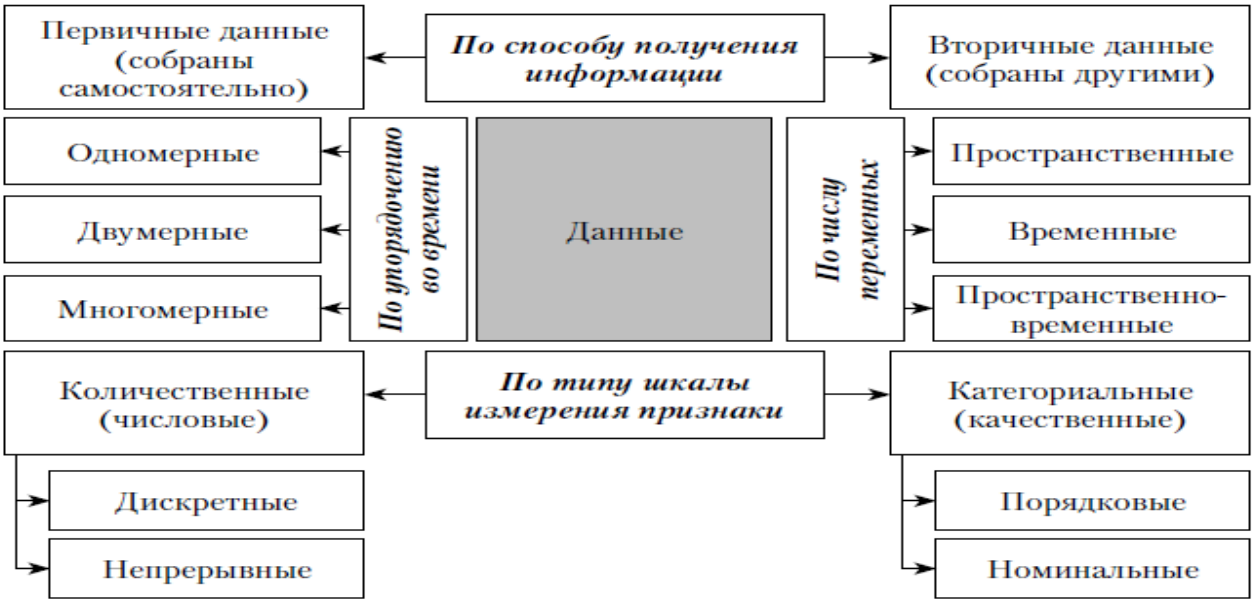

Рис. 1.1. Классификация статистических данных.

По числу переменных различают одномерный, двумерный и многомерный массивы данных.В *одномерных* наборах данных у каждого наблюдения регистрируется только один признак. В *многомерных* (двумерных, трехмерных и т.д.) наборах данных у каждого наблюдения регистрируется несколько признаков. Статистические методы в этом случае используются для решения задач определения основных характеристик по каждому одномерному признаку, анализа наличия и степени зависимости

между этими признаками, исследования вида зависимости одной переменной (результативной) от остальных (факторных), классификации наблюдений с целью получения однородных групп(кластеров) и выявления аномальных наблюдений, построения обобщающих, интегральных показателей с целью снижения размерности исходного признакового пространства, анализа рядов и прогнозирования (для временны данных).

Различают пространственные, временные и пространственно-временные данные.

Пространственные данные— значения переменных, относящихся к однотипным объектам в один и тот же фиксированный момент времени. Они позволяют сравнить значение признака, измеренное на разных объектах исследования, сравнить эти объекты по степени проявления в них того или иного свойства и разделить все объекты в соответствии с этим на группы и категории. *Временные данные* отражают динамику изменения переменных, характеризующих объект, на некотором промежутке времени. В зависимости от способа измерения значений признака временные данные делятся на *моментные* и *интервальные*.

Пространственно-временные данные — значения переменных, относящихся к сходным объектам за несколько моментов времени. Они могут быть также как моментными, так и интервальными.

В статистических исследованиях используют различные типы признаков, которые характеризуют состояние экономического объекта. Признаки могут иметь различный вид в зависимости от шкалы измерения, что в дальнейшем сказывается на выборе методов статистического анализа.

В зависимости от шкалы измерения различают *числовые* (*количественные*) и *категориальные* (*нечисловые, качественные*) данные. *Числовые* (*numerical*) *данные* — это показатели, принимающие числовые значения, которые получаются путем некоторых измерений или подсчетов. С точки зрения шкал измерений количественные данные считают измеренными

в интервальной шкале, которая применяется для отображения величины различия между характеристиками элементов.

Интервальная (*количественная*) *шкала* показывает, на сколько одно значение больше другого в принятых единицах измерения. Интервальная шкала может иметь произвольные начало отсчета и масштаб. Множество допустимых преобразований данной шкалы составляют все линейные преобразования. Основным свойством шкалы интервалов является сохранение отношения длин интервалов. Частными случаями шкалы интервалов являются *шкала отношений* нулевое начало отсчета) и *шкала разностей* (произвольное начало отсчета и единичный масштаб), а также *абсолютная шкала* (нулевое начало отсчета и единичный масштаб отсчета). Количественные шкалы допускают все арифметические действия над результатами измерения. В случае если данные получены путем измерений и они могут принимать абсолютно любые значения из некоторого промежутка или всей числовой оси, их называют *непрерывными* (*continuous*). Если данные образуют конечное или счетное множество и принимают только некоторые изолированные значения на числовой оси, между которыми значений быть не может, то такие признаки называют *дискретными* (*discrete*).

Другую группу, существенно отличающуюся от количественных данных, составляют нечисловые — *категориальные* (*categorical*) или *качественные* (*qualitative*) данные. В этом случае объект может принадлежать только к одной из множества категорий (классов). Особенно часто это имеет место при создании и обработке анкет, опросников, рейтингов и т.д.Даже если обозначить эти категории числами (например, перекодировать:0 — женский, 1 — мужской пол), то с такими данными все равно нельзя работать как с числовыми, а только как с категориальными.

В зависимости от того, можно ли эти категории упорядочивать, данные разделяют на *номинальные* (*nominal*) и *порядковые* (*ordinal*). Данные *шкалы наименований* (*номинальная, или классификационная, шкала*) определяются в

терминах категорий, которые нельзя содержательно упорядочить (профессия; регион страны; номер студенческой группы; и тд.). Номинальная шкала используется для описания принадлежности элементов к определенным классам. Нет оснований полагать, что одна категория лучше (или хуже), чем другая, поэтому при обработке таких данных применяются только операции сравнения: «равно» и «не равно».

Рис 1.2. Пример категориальных номинальных данных

Другой тип категориальных переменных — *порядковые* (*ординальные*). Порядковые шкалы используются для упорядочения элементов по одному или нескольким признакам. Они позволяют установить, что один элемент лучше, важнее, предпочтительнее другого или равноценен другому. Порядковая шкала отражает лишь порядок следования элементов и не дает возможности сказать, на сколько или во сколько раз один элемент предпочтительнее другого, в этой шкале нельзя определить меру степени предпочтительности. Для сравнения таких данных допускаются уже не только операции «равно» и «не равно», но и «больше-меньше» (без определения, на сколько).

По способу получения данные делятся на *первичные* — если информация собиралась специально для данного анализа и *вторичные* — если используется информация из других источников, собранная другими людьми и для других целей *Первичные* данные могут быть собраны самостоятельно или другими людьми по заказу, главное отличие — их сбор планируется и осуществляется в соответствии с задачами конкретного исследования. Поэтому они, конечно, обладают рядом преимуществ, прежде всего поскольку в этом случае достигается максимум соответствия

собранных данных требованиям и целям проводимого анализа. Но, к сожалению, часто такие исследования могут быть недоступны или слишком дороги. Но не все данные могут быть получены самостоятельно или заказаны профессионалам. Поэтому более распространенным вариантом является использование *вторичных* данных — уже собранных специализированными фирмами или просто другими исследователями.

Вторичные данные обладают своими преимуществами — они, как правило, гораздо дешевле или вообще бесплатны, могут быть велики по объему, зачастую собираются специальной труднодоступной аппаратурой и профессиональными в соответствующей области исследователями. Это могут быть данные, собранные научными институтами и подразделениями соответствующего направления, государственными учреждениями и маркетинговыми агентствами, учеными и исследователями. Поисковые системы Интернета значительно облегчают поиск вторичной информации.

Вопросы для самопроверки

1.Назовите основные критерии классификации данных.

2. В чем отличие пространственных, временных и пространственно-временных данных?

3. Назовите основные типы переменных в зависимости от шкалы измерения.

4. Чем отличаются номинальные категориальные переменные от порядковых?

Лекция 2. Анализ одномерных статистических данных

1.Предварительный анализ категориальных данных

Для того чтобы правильно применять те или иные статистические методы анализа данных, необходимо прежде всего определиться с типом данных.

К разным типам данных нужны разные подходы и методы. Начнем рассмотрение с наиболее простых с точки зрения анализа категориальных (качественных) данных. Категориальные данные — номинальные и порядковые — характеризуются тем, что все объекты исследования могут быть отнесены к разным категориям или классам, они позволяют произвести разделение объектов на подгруппы.

Измерение в номинальной (классификационной) шкале означает определение принадлежности объекта (наблюдения) к тому или иному классу. Например: пол, профессия, страна и т.д. В этой шкале, таким образом, можно лишь посчитать количество объектов в классах — *частоту* m_i и *относительную частоту* (*частость*) m_i/n.

Таблицы частот (*frequency tables*), или, как еще их называют, *одновходовые таблицы*, представляют собой простейший метод анализа категориальных переменных. В таких таблицах каждое значение переменной указывается вместе с частотой встречаемости такого значения в исследуемой совокупности объектов наблюдения (табл. 2.1).

Часто таблица частот дополняется накопленными (кумулятивными) частотами, показывающими суммарное число объектов во всех классах до рассматриваемого, включая его, и соответствующими относительными частотами (частостями) m_i/n или $m_i^{н}/n$.

Табл. 2.1.

Типичная таблица частот категориальной переменной.

Категория (класс) переменной	Частота m_i	Накопленная частота m_{in}	Относительная частота m_i/n	Относительная накопленная частота m_{in}/n
Класс 1	3	3	0,086	0,086
Класс 2	6	9	0,171	0,257
Класс 3	7	16	0,200	16/35 ≈ 0,457
Класс 4	11	27	0,314	0,771
Класс 5	5	32	0,143	0,914
Класс 6	2	34	0,057	0,971
Класс 7	1	35	0,029	1,000

Номинальные данные графически иллюстрируются при помощи гистограмм или круговых диаграмм. Для описания категориальных переменных не используются никакие числовые характеристики, так как они не принимают числовых значений, измеренных в интервальной шкале, и никакого смысла нет, например, в показателе «средний пол», посчитанном как среднее арифметическое между двумя категориями — мужской и женский пол, даже если они перекодированы в числа. Единственной полезной характеристикой является мода. Мода может быть не единственной, если два или несколько значений переменной обладают одинаковой максимальной частотой. В этом случае распределения называются бимодальными или полимодальными соответственно. Еще одним способом графического отображения и анализа категориальных номинальных данных является так называемая диаграмма Парето.

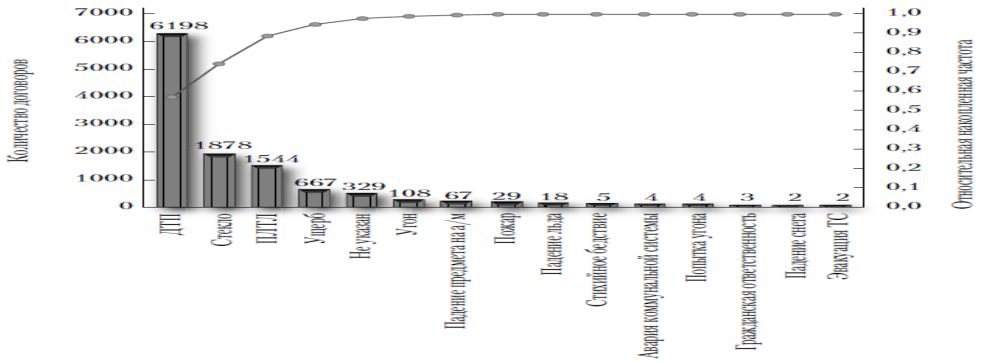

Рис. 2.1. Диаграмма Парето— распределение числа страховых случаев в портфеле автокаско по их причине.

Измерение в *порядковой* (*ординальной*) шкале, помимо определения класса принадлежности позволяет упорядочить наблюдения, сравнив их между собой в каком-то смысле — лучше/хуже, больше/меньше, *не определяя, насколько* лучше/хуже или больше/меньше. Поэтому порядковые экспериментальные данные, даже если они перекодированы в цифры, нельзя рассматривать как числа и выполнять над ними арифметические операции.

2. Предварительный анализ количественных данных

Проводя статистический анализ порядковых данных, кроме рассмотренных методов работы с номинальными, дополнительно к подсчету частот встречаемости категорий и формирования соответствующих таблиц частот, нахождения моды, можно вычислить ранг объекта. Порядковую шкалу также называют ***ранговой***, а место объекта в последовательности, которую она собой представляет, — ***рангом*** объекта.

Кроме того, так как шкала ранжированная, то графически для порядковых данных можно построить не только гистограмму и круговую диаграмму, но и изобразить исходные данные с их значениями в виде *столбиковой диаграммы*, если их число не очень велико. Высота столбца отражает значения переменной у объекта (для порядковой переменной определяет его ранг).

Рис. 2.2. Столбиковая диаграмма оценок за экзамен по 10-балльной шкале

Количественные переменные можно превратить в категориальные, т.е. их категоризовать (категорировать). Например, непрерывная переменная «вес человека в килограммах» может быть превращена в порядковую переменную «вес человека» с градациями: недостаточный, нормальный, избыточный. Переменные с двумя категориями (есть свойство — нет свойства; купил — не купил и т.д.) обычно кодируют цифрами 0 и 1 и называют *дихотомическими* или *бинарными*. Особенно широкое применение они имеют в экономических, медицинских и социологических исследованиях, в которых большинство переменных, интересующих специалистов, измеряется в качественных шкалах. При этом дихотомические данные зачастую являются более адекватными, чем результаты измерений по методикам, использующим большее число категорий (например, в психологических тестах).

Количественные данные — дискретные и особенно непрерывные — вследствие более совершенных шкал измерения и возможности количественной оценки различий между ними - допускают уже гораздо более широкий спектр возможностей в плане их статистического анализа. Так как практические исследования зачастую связаны с достаточно большими массивами данных, одной из первых у исследователя возникает задача их *сгруппировать*, изучить их внутреннюю структуру, представить в более компактном и удобном виде, дающем лучшую визуализацию таких данных.

Пусть имеются данные n наблюдений $x1, x2, \ldots, xn$, каждое из которых характеризует один количественный признак X. Решается задача обработки этих данных. Если число наблюдений (n) достаточно велико (по крайней мере $n \geq 50$), то их предварительно ранжируют и подвергают группировке.

Вариационный ряд — значения n независимых наблюдений $x1, x2, \ldots, xn$ количественного признака X, расположенные в порядке возрастания (неубывания) значений: $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(i)} \leq x_{(i+1)} \leq \ldots \leq x_{(n)}$.

Разность наибольшего и наименьшего значений признака называется *размахом* вариационного ряда: $R = x_{(n)} - x_{(1)} = x_{\max} - x_{\min.}$. (1)

С целью улучшения представления эмпирических данных при большом числе наблюдений их группируют, получая *сгруппированный вариационный ряд*. Для группировки *дискретных* количественных данных подсчитывают частоту встречаемости *mi* каждого признака *xi*. При достаточно большом т числе значений сгруппированный вариационный ряд может быть подвергнут дальнейшей группировке и преобразован в интервальный. Сгруппированный дискретный вариационный ряд графически представляют в виде *полигона* – ломаной линии, соединяющей точки, по оси абсцисс соответствующие всем возможным значениям признака, а по оси ординат — значениям частот *mi* или относительных частот *wi=mi/n*. Полигон позволяет оценить распределение частот значений дискретной переменной, выявить наиболее часто (мода) и редко встречающиеся значения признака.

Сгруппированный кумулятивный дискретный вариационный ряд представляет собой значения признака *xi*, указанные вместе с соответствующими накопленными частотами $mi^{н}$ или частостями $wi^{н}=mi^{н}/n$:

Значения признака x_i	$x_{(1)}$	$x_{(2)}$	...	$x_{(k)}$
Накопленная частота $m_{iн}$	$m_{1н} = m_1$	$m_{2н} = m_1 + m_2$	...	$m_{kн} = n = \sum_{i=1}^{k} m_i$

Сгруппированный кумулятивный дискретный вариационный ряд графически представляют в виде *кумуляты* (по оси абсцисс откладывают все возможные значения признака, а по оси ординат — накопленные частоты или накопленные относительные частоты). Кумулята показывает долю объектов совокупности значения признака которых не превышают заданного.

Рис. 2.3. Кумулята частот числа промахов спортсменов

При обработке больших массивов информации перед исследователем стоит серьезная задача правильной группировки исходных данных. Если исследуемый признак имеет *непрерывный* характер, то выбор оптимального числа интервалов группировки признака является отнюдь не тривиальной задачей. Для группировки непрерывных случайных величин весь вариационный т размах признака разбивают на некоторое количество интервалов *k*.

Сгруппированным интервальным (*непрерывным*) *вариационным рядом* называют ранжированные по значению признака интервалы ($ai \le x \le bi$), где $i = 1, 2, \ldots, k$, указанные вместе с соответствующими частотами (mi) числа наблюдений, попавших в *i*-й интервал, или относительными частотами (mi/n):

Интервалы значений признака $a_i \div b_i$	$a_1 \div b_1$	$a_2 \div b_2$	...	$a_i \div b_i$	...	$a_k \div b_k$
Частота m_i	m_1	m_2	...	m_i	...	m_k

Гистограмма и *кумулята* являются прекрасным средством визуализации данных, позволяющим получить первичное представление о структуре данных. Такие графики строятся для непрерывных данных так же, как и для дискретных, только с учетом того, что непрерывные данные сплошь заполняют область своих возможных значений, принимая любые значения.

Гистограмма и полигон являются аппроксимациями кривой плотности вероятности (дифференциальной функции) $f(x)$ теоретического распределения, поэтому их построение имеет такое важное значение при

первичной статистической обработке количественных непрерывных данных — по их виду можно судить о гипотетическом законе распределения. С кумулятой сопоставляется график интегральной функции распределения $F(x)$. В основном понятия гистограммы и кумуляты связывают именно с непрерывными данными и их интервальными вариационными рядами, так как их графики являются эмпирическими оценками функции плотности вероятности и функции распределения соответственно.

Построение интервального вариационного ряда начинают с определения числа интервалов k. Число интервалов не должно быть слишком малым, так как при этом гистограмма получается слишком сглаженной (*oversmoothed*), теряет все особенности изменчивости исходных данных. В то же время число интервалов не должно быть слишком велико —иначе мы не сможем оценить плотность распределения изучаемых данных по числовой оси: гистограмма получится недосглаженная (*undersmoothed*), с незаполненными интервалами, неравномерная.

Приблизительное число классов k, которое необходимо выбрать при группировке и построении гистограммы для n результатов измерений величины, полученных из нормально распределенной генеральной совокупности, определяется по правилу Стерджеса (*Sturges'rule*) следующим образом:

$$k = 1 + \log_2 n \approx 1 + 3{,}322 \cdot \lg n, \qquad (2)$$

где log2 n и lg n — логарифмы по основаниям 2 и 10 соответственно от числа наблюдений n.

Ширина интервалов h, на которые необходимо разбить всю область возможных значений исследуемого признака по имеющимся наблюдениям $\{x1, x2, \ldots, xn\}$, определяется тогда следующим образом:

$$h = \frac{R}{1 + \log_2 n} = \frac{x_{\max} - x_{\min}}{1 + \log_2 n} \approx \frac{x_{\max} - x_{\min}}{1 + 3{,}322 \lg n}, \qquad (3)$$

где R — размах значений признака. За нижнюю границу первого интервала обычно принимается величина $a1 = x\min - h/2$.

Верхняя граница первого интервала $b1=a1 + h$. При определении границ следующих интервалов исходят из условий: $ai+1=bi$; $bi=ai + h$. Построение интервалов продолжается до тех пор, пока начало следующего по порядку интервала не будет равным или больше x_{max}.

Число интервалов k для небольших объемов данных обычно берут следующим:

• 5—6 при $n \leq 50$;

• 6—8 при $50 < n \leq 100$;

• 8—10 при $n > 100$;

Считается, что формула Стерджеса позволяет строить удовлетворительные гистограммы при числе измерений менее 200. Для современных огромных массивов информации, например порядка 104—109 наблюдений, правило Стерджеса может приводить к слишком сглаженным гистограммам. Кроме того, если данные не являются выборкой из нормальной совокупности, распределение обладает существенной асимметрией, то формула Стерджеса также не подходит для группировки таких наблюдений — асимметричные распределения требуют большего числа интервалов группировки.

Таким образом, современному исследователю необходимо четко осознавать критерии и границы применимости используемых правил, и формула Стерджеса не является в этом смысле исключением. На ее замену в настоящее время существует достаточное количество альтернативных методов. Рассмотрим кратко основные из них. Предположим, что исследуемые данные имеют непрерывный закон распределения, функция плотности вероятности которого $g(x)$ дифференцируема, но неизвестна.

Одно из интуитивных предположений — взять ширину интервалов так, чтобы в каждый из них попадало равное количество значений выборочной совокупности. Однако Дэвид Скотт показал, что такой метод приводит к слишком узким интервалам в окрестности модальных значений. Поэтому рассмотрим более распространенный подход использования интервалов

равной ширины. Необходимо заметить, что предварительная обработка данных может привести к лучшим гистограммам. Например, распределения с существенной правосторонней асимметрией могут быть прологарифмированы. Гистограмма с равной шириной интервалов имеет два параметра: ширину интервала h (шаг вариационного ряда) и начало первого интервала $a0$. Определение первого параметра — шага h — является критическим, в то время как второй параметр не столь важен и во многих случаях принимается равным минимальному значению выборки $a0=x(1)$ (или отстоящим от него полшага влево).

Главный фактор, влияющий на качество гистограммы, — ширина интервала (шаг вариационного ряда). Рассмотрим основные проблемы, возникающие при выборе наилучшего значения h. Интуитивно понятно, что большие значения приводят к интервалам с большим количеством данных и меньшим их разбросом (дисперсией). Меньшие значения шага дают гистограммы с большей вариабельностью, но возможностью оценки истинной плотности $g(x)$ с учетом всех ее возможных изгибов.

Д. Скотт показал, что оптимальной шириной интервалов является

$$h^* = \sqrt[3]{\frac{6}{nR(g')}}, \tag{4}$$

где $R(g') = \int_{-\infty}^{\infty} g'(x)^2 dx$.

Формула практически очень значима, несмотря на то что функция плотности вероятности $g(x)$ неизвестна. Если можно сделать предположение о нормальном распределении изучаемой совокупности, то

$$R(g') = \int_{-\infty}^{\infty} g'(x)^2 dx = \frac{1}{4\sqrt{\pi}\sigma^3}$$ и $$h^* = \sqrt[3]{\frac{24\sqrt{\pi}\sigma^3}{n}} \approx \frac{3{,}5\sigma}{\sqrt[3]{n}}.$$

Д. Скотт предложил использовать в качестве оценки σ выборочное стандартное среднее квадратическое отклонение, тогда оптимальной шириной интервалов h является $$h^* \approx \frac{3{,}5S}{\sqrt[3]{n}}, \quad (5)$$

где S — среднее квадратическое отклонение значений переменной по всем наблюдениям.

Д. Фридман и П. Диаконис использовали вместо σ оценку 4/7**IQR*, основанную на выборочном интерквантильном размахе *IQR*, что дает для шага интервального ряда следующую формулу: $$h^* = \sqrt[3]{\frac{24\sqrt{\pi}\sigma^3}{n}} \approx \frac{2 \cdot IQR}{\sqrt[3]{n}}. \quad (6)$$

Это оценка более устойчива (робастна), но не так оптимальна для нормального распределения. Для этого случая авторы считают, что вместо коэффициента 2 в формуле лучше использовать 2,6.

В случае отличия изучаемого распределения от нормального необходимо, прежде всего, постараться оценить функцию $R(g'(x))$, а затем уже с помощью полученных результатов подобрать оптимальный шаг вариационного интервального ряда. В любом случае отмечают, что ширина интервала при построении вариационного ряда должна быть порядка $h^* \approx \frac{C}{\sqrt[3]{n}}$.

Метод квадратного корня (*square-root choice*) — число классов k выбирается равным квадратному корню из числа наблюдений n: $k \approx \sqrt{n}$.

Метод используется в *MS Excel* и некоторых статистических пакетах при построении гистограмм. Часто он приводит к слишком большому числу интервалов и недосглаженным гистограммам. Существуют и другие методы определения количества классов при группировке данных, но необходимо отметить, что в любом случае построение наилучшего вариационного ряда для изучаемого признака определяется в зависимости от конкретных задач и вида распределения данных, и его можно назвать искусством исследователя.

Вопросы для самопроверки

1. Какие методы анализа применимы к категориальным данным?

2. Как сгруппировать дискретные и непрерывные количественные переменные?

3. Какие методы определения ширины интервального ряда для группирования непрерывной переменной вы знаете?

Лекция 3. Дискриптивные статистики и процедуры разведочного анализа данных

1.Основные описательные статистические характеристики

Для того чтобы изучать какие-то количественные переменные и сравнивать их между собой, необходимо уметь рассчитывать основные числовые характеристики признаков. Для проведения дальнейших вычислений, как правило, интервальный вариационный ряд заменяется на дискретный. С этой целью все значения признака в пределах каждого интервала приравниваются к его срединному значению $x_i = \frac{a_i + b_i}{2}$. (1)

Как правило, рассчитывают следующие основные описательные (дискриптивные) статистики:

Средние представляют собой обобщающие показатели, характеризующие центр группирования данных. Обычно рассматривают средние: *арифметическую (arithmetic mean), гармоническую(garmonic mean) и геометрическую (geometric mean).*

Медиана (*median*) — значение признака, приходящего на середину вариационного (ранжированного) ряда наблюдений.

Для несгруппированных данных:

если число наблюдений нечетное, $n=2p+1$, то $Me=x_{(p+1)}$;

если число наблюдений четное, $n=2p$, то $Me = x_{me} = \frac{x_{(p)} + x_{(p+1)}}{2}$

Для интервального вариационного ряда *медианным* называют первый интервал [*ame*; *bme*), для которого накопленная частота впервые превышает половину объема наблюдений (или равна ей). Медиана для интервального вариационного ряда вычисляется по формуле

$$Me = x_{Me} = a_{Me} + h \cdot \frac{\frac{n}{2} - m^{(н)}_{Me-1}}{m_{Me}} \qquad (2)$$

где a_{Me}— нижняя граница медианного интервала;

$m^{(н)}_{Me\text{-}1}$ — накопленная частота интервала, предшествующего медианному;

m_{Me}— частота медианного интервала;

h — ширина интервала группирования.

Свойство медианы: сумма абсолютных отклонений признака от *Me* меньше, чем от любой другой величины: $\sum_{i=1}^{n}|x_i - x_{Me}| = \min$. (3)

Мода (*mode*) — наиболее часто встречаемое значение признака — для интервального сгруппированного вариационного одномодального ряда равна

$$Mo = x_{Mo} = a_{Mo} + h \cdot \frac{m_{Mo} - m_{Mo-1}}{2m_{Mo} - m_{Mo-1} - m_{Mo+1}}, \quad (4)$$

где a_{Mo}— нижняя граница модального интервала (интервала с наибольшей частотой);

m_{Mo}— частота модального интервала;

$m_{Mo\text{-}1}$ — частота интервала, предшествующего модальному;

m_{Mo+1} — частота интервала, следующего за модальным.

Показатели вариации характеризуют величину разброса наблюдаемых значений $x1, x2, \ldots, xn$ относительно среднего значения.

***Дисперсия* (*variation*)** — средняя арифметическая квадрата отклонения наблюдаемых значений *xi* от средней арифметической $\bar{x}$:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ или } S^2 = \frac{1}{n}\sum_{i=1}^{l}(x_i - \bar{x})^2 m_i, \quad (5)$$

если данные сгруппированы и *xi* — *i-е* значение признака — наблюдалось *mi* раз.

Иногда в качестве оценки дисперсии используют ее несмещенную оценку— так называемую исправленную выборочную дисперсию:

$$\hat{S}^2 = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \quad (6)$$

Дисперсия имеет размерность, равную квадрату размерности признака. Таким образом, если признак x измеряется в рублях, то размерность

дисперсии — [руб.^{2}], что представляется неудобным во многих задачах. Для этого была введена еще одна статистическая характеристика разброса, лишенная этого недостатка.

Среднее квадратическое отклонение (*standard deviation*) *S*, равное положительному корню из дисперсии S^2, имеет ту же размерность, что и *x*.

Среднее квадратическое отклонение равно $S = \sqrt{S^2}$. (7)

Размах вариации (*range*) равен разности между максимальным и минимальным значениями признака: $R = x_{\max} - x_{\min}$. (8)

Размах вариации и среднее квадратическое отклонение связаны примерным соотношением $R \approx (4+6)S$.

Квартили (от лат. *quarto* — четыре или *quarta* — четверть) **—** это такие значения изучаемого признака, левее (нижний квартиль) или правее (верхний квартиль) которых находится четверть всех наблюдений.

*Q*1 – это (n+1)/4 - ранжированное наблюдение — *первый* (*нижний*) *квартиль* — значение вариационного ряда данных, левее которого находится четверть (25%) всех наблюдений;

*Q*3 — это ¾*(n+1) -ранжированное наблюдение — *третий* (*верхний*) *квартиль* — значение вариационного ряда данных, правее которого находится четверть (25%) всех наблюдений.

Интерквартильный размах (*interquartile range*) отражает среднюю половину(50%) данных — различия между первым и третьим квартилем:

$$IQR = Q3 - Q1. \quad (9)$$

Коэффициент вариации (*coeffi cient of variation*) является процентным показателем вариации: $V_s = \frac{S}{\bar{x}} 100\%$. (10)

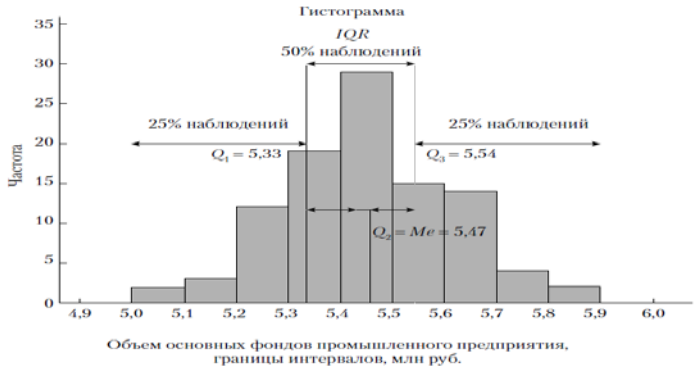

Рис. 3.1. Квартили и интерквартильный размах в ППП Statistica

Центральным моментом k-го порядка называют среднюю арифметическую *k*-й степениотклонения наблюдаемых значений *xi* (*i*=1, 2, …, *n*) от среднего значения, т.е. $\mu_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$. (11)

Начальным моментом k-го порядка называют среднюю арифметическую *k*-й степени значений *xi* (*i*=1, 2, …, *n*), т.е. $\hat{\vartheta}_k = \frac{1}{n}\sum_{i=1}^{n}x_i^k$. (12)

При *k*=1 имеем $\hat{\vartheta}_1 = \bar{x}$, т.е. средняя арифметическая есть начальный момент первого порядка. Центральный момент второго порядка *k*=2 есть дисперсия: $\mu_2 = S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Можно показать, что начальные и центральные моменты связаны соотношениями $\mu_2 = \hat{\vartheta}_2 - \hat{\vartheta}_1^2;\ \mu_3 = \hat{\vartheta}_3 - 3\hat{\vartheta}_2\hat{\vartheta}_1 + 2\hat{\vartheta}_1^3$.

Центральные моменты 3-го и 4-го порядков обычно используются не сами по себе, а для расчета коэффициентов асимметрии и эксцесса.

Коэффициент асимметрии (*skewness*) характеризует степень асимметричности, скошенности распределения данных и находится по формуле $Ac = \frac{\mu_3}{S^3}$. (13)

Коэффициент асимметрии был впервые введен Карлом Пирсоном в 1895 г.

При *Ac*>0 имеет место правосторонняя асимметрия (на графике более пологий спуск справа); при *Ac*<0 —левосторонняя асимметрия (на графике

более пологий спуск слева); при $Ac=0$ — идеально симметричное распределение. Если $|Ac|>0{,}5$, то асимметрия существенна.

Коэффициент эксцесса (*kurtosis*) — показатель, служащий мерой крутости (плосковершинности или островершинности) графика вариационного ряда в сравнении с кривой нормального распределения, определяемый по формуле $$Ek=\frac{\mu_4}{S^4}-3. \quad (14)$$

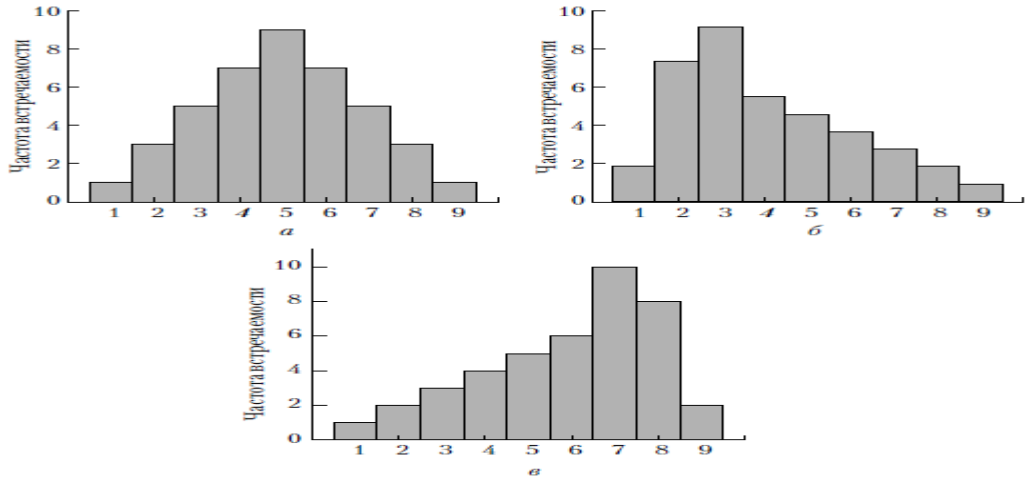

Рис. 3.2. Гистограммы симметричного распределения (*а*)и распределений с правосторонней (*б*) и левосторонней (*в*) асимметрией

Если $Ek>0$, то график ряда распределения является островершинным, если $Ek<0$ — плосковершинным по сравнению с нормальным(у нормального распределения коэффициент эксцесса равен 0).

Коэффициент эксцесса был впервые введен Карлом Пирсоном в 1905 г.1

Также он ввел термины островершинности (*leptokurtic*), плосковершинности (*platykurtic*) и *mesokurtic* (примерно такая же, как у нормальной кривой, вершина). Позже известный статистик Уильям Сили Госсет (William SealyGosset, работавший под псевдонимом Стьюдент) в своей работе привел забавные картинки, показывающие смысл и разницу “хвостов” у распределений с разными коэффициентами эксцесса (рис 3.3).

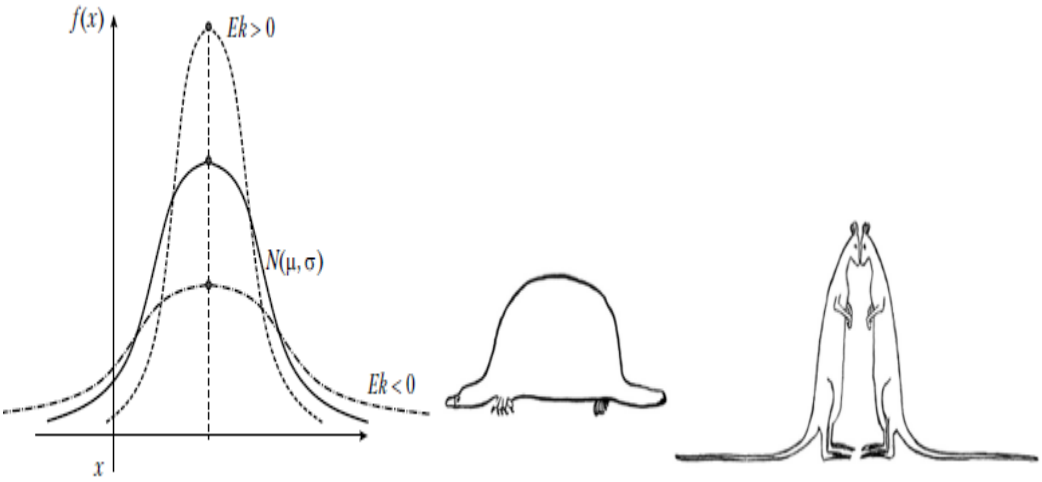

Рис. 3.3. Сравнение кривых трех распределений с различными коэффициентами эксцесса и картинки Стьюдента (У. Госсета), иллюстрирующие разницу плосковершинных и островершинных распределений

В некоторых источниках коэффициент эксцесса считают как отношение центрального момента 4-го порядка к квадрату дисперсии (не вычитая 3). В MS Excel коэффициенты асимметрии и эксцесса считаются по другим формулам несмещенных оценок, поэтому результаты расчетов будут различны.

2.Диаграммы «ящик с усами» (box and whisker plots), нормирование и унификация данных

Статистические пакеты позволяют строить еще один полезный и информативный вид графиков, объединяющий характеристики и центра группирования, и разброса, — так называемые ящичковые диаграммы (*Box&Whisker Plot).*Их отличительной особенностью является то, что они графически отображают три характеристики данных:

• центральная точка показывает центр группирования данных, центральную тенденцию;

• «ящик» (*box*) отражает разброс вокруг центра;

• «усы» (*whiskers*) характеризуют размах наблюдений.

При этом возможны различные варианты построения таких графиков по выбору исследователя. Например, статистический пакет *STATISTICA*

(*StatSoft*) строит для одномерных данных четыре вида таких диаграмм(аналогичные графики можно построить в *IBM SPSS Statistics* и других статистических пакетах):

• медиана — квартили — размах;

• среднее — стандартная ошибка среднего — стандартное отклонение;

• среднее — стандартное отклонение — 1,96*стандартное отклонение (95%-ный нормальный доверительный интервал для наблюдений вокруг среднего);

• среднее — стандартная ошибка среднего — 1,96*стандартная ошибка среднего (95%-ный нормальный доверительный интервал для генерального среднего).

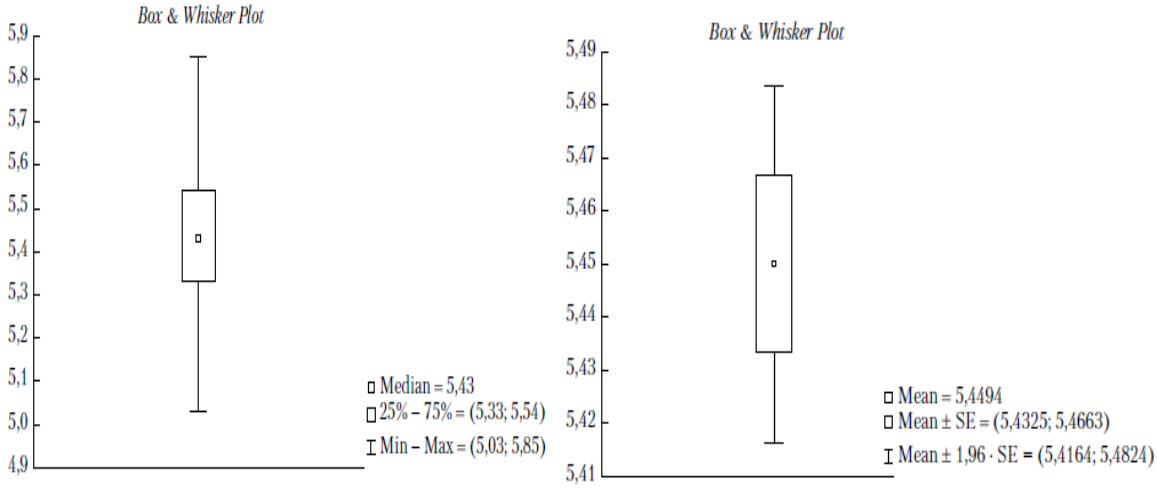

Рис. 3.4. Диаграмма «ящик с усами» вида «медиана — квартили — размах» и диаграмма «ящик с усами» вида «среднее — стандартная ошибка среднего — 95%-ный нормальный доверительный интервал для генерального среднего»

Такие графики позволяют быстро оценить основные характеристики дескриптивной статистики, проанализировать центральную тенденцию и степень разброса данных и даже получить доверительные интервалы. Представленные диаграммы являются хорошим способом анализа совокупности данных на аномальные наблюдения — выбросы (*outliers*) и экстремальные наблюдения (*extremes*).

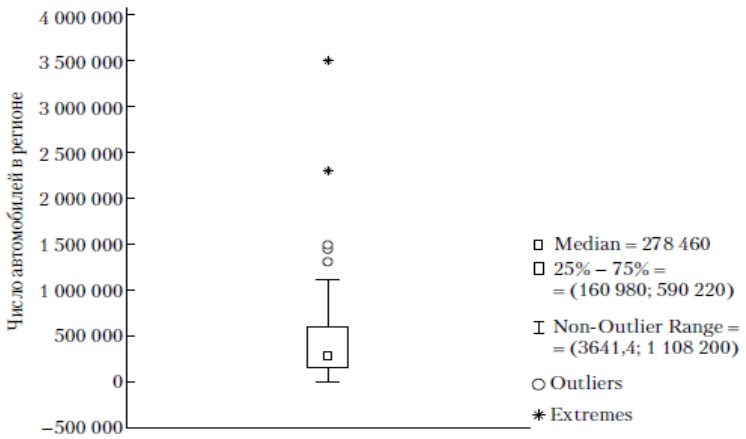

Рис. 3.5. Диаграмма «ящик с усами»вида «Медиана — интерквартильный размах — размах»для данных по количеству автомобилей в регионе по регионам России

Рисунок 3.5 позволяет получить множество характеристик представленной совокупности — медиану распределения (половина регионов имеют количество автомобилей, не превышающее 278 460 шт.), интерквартильный размах — при этом 25% регионов имеют менее 160 980 автомобилей, а 25% — более 590 220 автомобилей.

На графике сразу видны аномальные наблюдения (Свердловская обл. с 1 311 200 автомобилями)и экстремальные наблюдения — г. Санкт-Петербург, Краснодарский край, Московская обл. и г. Москва, число автомобилей в которых существенно превышает значения в остальных регионах. Кроме того, ящик и расположение в нем медианы показывают, что распределение обладает существенной правосторонней асимметрией (смещение вверх, в область высоких значений), в чем можно убедиться, построив гистограмму распределения.

В ряде задач бывает удобно или даже необходимо перейти от исходных наблюдений к нормированным (стандартизованным). Пусть имеются данные $x1, x2, \ldots, xi, \ldots, xn$, на основании которых получены $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i;\ S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Нормированными (*стандартизованными*) называют данные вида

$$x_i^* = \frac{x_i - \overline{x}}{S},\ i = 1, 2, ..., n, \tag{15}$$

—безразмерные величины, удовлетворяющие условию $\overline{x^*} = 0$ и $S^2_{x^*} = 1$.

Покажем, что средняя арифметическая нормированных данных равна нулю:

$$\overline{x^*} = \frac{1}{n}\sum_{i=1}^{n} x_i^* = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i - \overline{x}}{S} = \frac{1}{S}\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) = \frac{1}{S}\cdot\frac{1}{n}\left(\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n}\overline{x}\right) = 0,$$

а дисперсия равна единице:

$$S^2_{x^*} = \frac{1}{n}\sum_{i=1}^{n}(x_i^* - \overline{x^*})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{S}\right)^2 = \frac{1}{S^2}\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = 1.$$

При этом если нормированная величина больше нуля, то наблюдаемое значение больше среднего. Если же $x^*<0$, то $x_i < \overline{x}$.

Стандартизация (нормирование) данных является необходимым начальным этапом преобразования данных при использовании многих многомерных статистических методов — снижения размерности признакового пространства (факторный, компонентный анализ), классификации объектов и др., особенно если переменные измерены в шкалах, существенно отличающихся в величинах.

Вследствие распространенности и востребованности в статистических пакетах процедура нормирования (стандартизации) обычно вынесена в меню (рис. 3.6.).

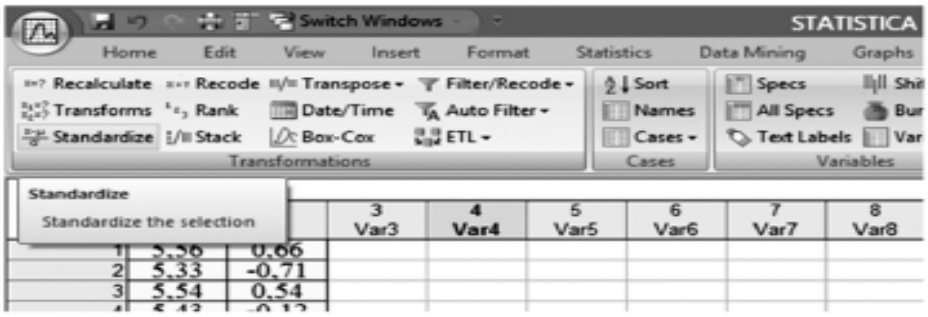

Рис. 3.6. Вызов процедуры нормирования (стандартизации) данных в меню пакетаSTATISTICA(StatSoft)

При построении интегральных обобщающих показателей часто возникает ситуация, когда нормирование данных не дает нужного результата. Например, нам необходимо построить интегральный показатель качества жизни в стране (регионе),включающий в себя три исходные переменные — продолжительность жизни, младенческую смертность и уровень безработицы. При этом, даже переведя эти три показателя в единую шкалу (например, со значениями от 0 до 1 или от 0 до N), мы будем иметь конфликт в интерпретации переменных, т.к. продолжительность жизни — характеризуется тем, что чем большие значения она принимает, тем выше качество жизни, а младенческая смертность — при повышении значений понижает качество жизни, в то время как безработица имеет свой некоторый оптимум (примерно 5%). И, соединив все три признака в один интегральный показатель, мы будем иметь отсутствие адекватной интерпретации полученного показателя. Для разрешения таких необходимо приведение всех переменных, участвующих в построении интегрального показателя, к единой *унифицированной шкале*.

Такая шкала используется при построении интегральных показателей из различных переменных. Она принимает значения от 0 до N и имеет единую систему интерпретации: чем выше значения переменной в унифицированной шкале, тем выше значение интегрального показателя. При N=1 получаем шкалу от 0 до 1.

Переменные первого типа — чем выше показатель, тем лучше (продолжительность жизни), — приводятся к унифицированной шкале следующим образом:

$$x_i^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \cdot N, \qquad (16)$$

где xi — значение переменной для i-го наблюдения;

$x$min и $x$max — соответственно наименьшее и наибольшее наблюдаемые значения переменной (чем больше значение переменной xi, тем выше (лучше) ее значение в унифицированной шкале).

Переменные второго типа — чем выше показатель, тем хуже (младенческая смертность) — приводятся к унифицированной шкале следующим образом:

$$x_i^* = \frac{x_{\max} - x_i}{x_{\max} - x_{\min}} \cdot N, \qquad (17)$$

т.е. чем больше значение переменной *xi*, тем ниже (хуже) ее значение в унифицированной шкале.

Переменные третьего типа — показатель имеет некий оптимум *х*опт, это значение наилучшее, чем больше отклонения от него, тем хуже (уровень безработицы) — приводятся к унифицированной шкале следующим образом:

$$x_i^* = \left(1 - \frac{|x_i - x_{\text{опт}}|}{\max\{(x_{\max} - x_{\text{опт}}); (x_{\text{опт}} - x_{\min})\}}\right) \cdot N. \qquad (18)$$

Если же *xi* имеет максимально возможное отклонение от *х*опт, то *xi**=0. Таким образом, чем больше значение переменной *xi* отклоняется от оптимального, тем ниже (хуже) значение *xi** в унифицированной шкале, а чем ближе значение *xi* к *х*опт, тем лучше.

Вопросы для самопроверки

1. Назовите основные характеристики центра группирования количественных данных.

2. Какие показатели вариации количественных данных вы можете назвать? В чем их отличия?

3. Что такое диаграмма «ящик с усами» (ящичковая диаграмма), какие характеристики показателя по ней можно определить?

4. Каким образом осуществляется нормирование (стандартизация) данных?

5. Для чего проводится унификация шкал различных данных?

Раздел 2. Статистический анализ совокупностей

Лекция 4. Методы обнаружения засорения выборки

1.Визуальное обнаружение засорения выборки

В процессе спецификации моделей и оценки их параметров на основе данных результатов измерений наличие резко выделяющихся наблюдений или значений может привести как к неправильному определению вида модели, так и к ошибкам в оценке ее основных характеристик.

Наличие аномальных наблюдений и грубых ошибок измерений часто порождает искаженное представление о структуре и взаимодействии исследуемых объектов. В свою очередь, неправильная спецификация модели обусловливает ее существенное несоответствие имеющимся данным, и некоторые результаты измерения могут быть восприняты как аномальные наблюдения. Именно поэтому моделирование требует постоянного внимания к изучению механизма, лежащего в основе моделируемого явления, его содержательного анализа.

При решении задач классификации даже одно аномальное значение может существенно исказить моделируемую структуру. Причинами «засорения» выборки могут служить ошибки в съеме или вводе данных, а также искажения при их передаче. Рассмотрим пример «засорения» нормально распределенных двумерных данных одним аномальным наблюдением.

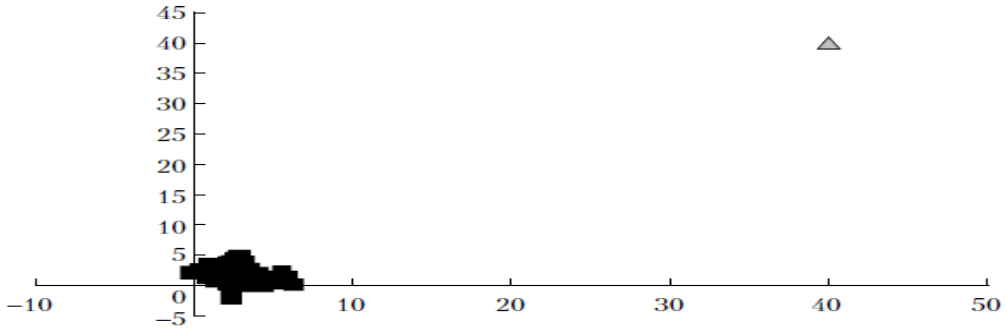

Рис.4.1. Диаграмма рассеяния при наличии аномального наблюдения

На рис. 4.1 представлено эмпирическое распределение нормально распределенных групп объектов и аномального наблюдения в двумерном признаковом пространстве. Присутствие отличающихся друг от друга однородных групп далеко неочевидно вследствие чрезмерно крупного масштаба, необходимого для охвата всех исходных объектов.

Решением задачи классификации для этого случая, очевидно, послужило бы разбиение объектов на две группы: группу, состоящую из аномального объекта, и группу из всех остальных объектов.

Удаление аномалии объекта приводит к иной картине распределения. На рис. 4.2 видна более детальная картина эмпирического распределения, в которой, в отличие от ситуации с аномальным наблюдением, расстояния между объектами при классификации будут играть существенную роль.

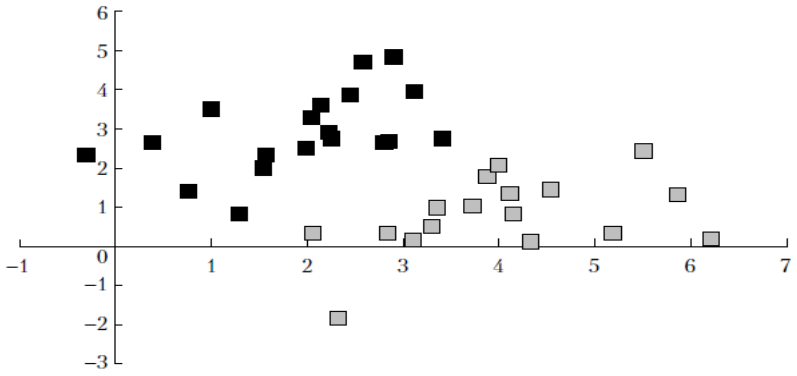

Рис. 4.2. Диаграмма рассеяния в отсутствие аномального наблюдения

Примером влияния ошибочной спецификации на решение задачи классификации может служить неправильный выбор закона распределения.

Чаще всего на практике предполагается нормальность распределения признаков. При отличии распределения признаков от нормального часть наблюдений образует группу явных «нарушителей» нормального закона.

Рассмотрим пример логарифмически нормального распределения. Таким примером может служить распределение однородной группы населения по уровню дохода. Предположение нормальности в данном случае теоретически неоправданно, но встречается в работах исследователей.

Ящичная диаграмма выборки из логарифмически нормальной совокупности представлена на рис. 4.3.

На ней заметны свойственная логарифмически нормальному распределению асимметрия и наличие аномальных (для гипотетического случая нормального распределения) значений в виде звездочек с подписанными возле них номерами наблюдений.

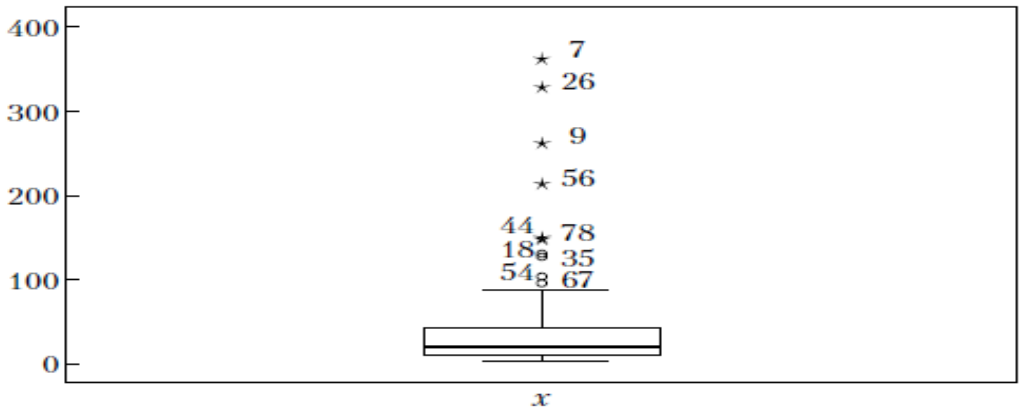

Рис. 4.3. Ящичная диаграмма логарифмически нормально распределенного признака *x*

Ящичная диаграмма для тех же данных после логарифмирования признака x (y=ln x) показывает отсутствие у логарифмированных данных аномальных значений (рис. 4.4).

Статистическое оценивание параметров генеральной совокупности, приводящее к проблеме появления аномальных значений вследствие неправильной спецификации модели распределения, может наблюдаться и при правильном определении вида закона распределения, но наличии в модели объектов из нескольких совокупностей с различными параметрами.

Эту проблему можно рассматривать как ошибку в описании модели, например, считать наблюдения нормально распределенного признака аномальными, если в совокупности содержатся объекты из различных однородных групп объектов, даже если в каждой из них аномальные наблюдения отсутствуют.

Распределение реальных совокупностей может представлять собой смесь нормального распределения с распределениями других видов, описывающих тяжелые «хвосты» эмпирического распределения. Примерами таких распределений могут служить распределение Парето, распределение Стьюдента с небольшим числом степеней свободы.

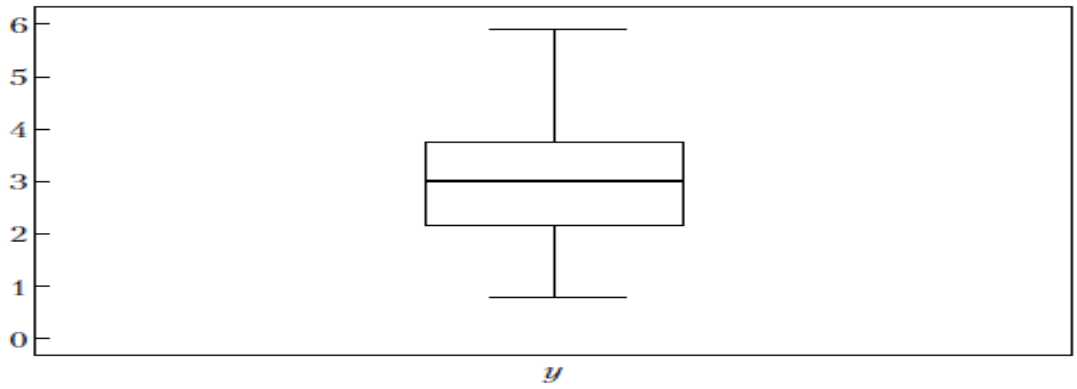

Рис. 4.4. Ящичная диаграмма логарифмически нормально распределенного признака после логарифмирования

Если содержательно недопустимо предположение о наличии нескольких групп объектов в генеральной совокупности, то чужеродные объекты могут рассматриваться как «засоряющие» однородную совокупность. Влияние эффекта «засорения» выборки на результат построения регрессионной модели может быть проиллюстрировано на примере его влияния на оценки коэффициентов.

На рис. 4.5 приведена диаграмма рассеяния однородной группы объектов и одного аномального наблюдения с наложенными на нее линиями регрессии, одна из которых построена по всем имеющимся данным, а вторая — только по данным об однородной группе, т.е. без аномального наблюдения.

Оценки получены обычным методом наименьших квадратов, обеспечивающим получение эффективных оценок при условии выполнения условий теоремы Гаусса — Маркова, но чувствительным к нарушению этих условий, в частности к наличию резко выделяющихся наблюдений. Как

очевидно из рис. 4.5, эффект от присутствия только одного аномального значения может проявляться в существенном изменении угла наклона линии регрессии и даже в смене знака коэффициента при регрессоре на противоположный.

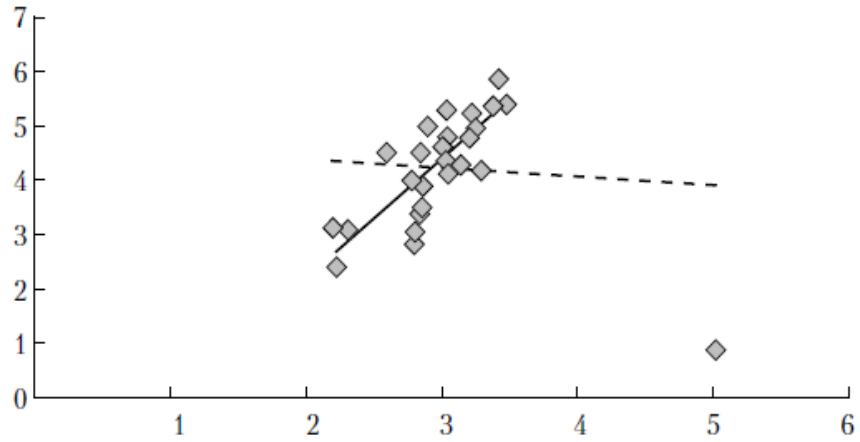

рис. 4.5. Линии регрессии в отсутствие (сплошная линия) и при наличии (пунктирная линия) аномального наблюдения

Можно провести определенную аналогию между формированием статистической оценки и поведением живого организма. Если оценивание производится в соответствии с алгоритмами, мало чувствительными к ошибкам спецификации модели (неблагоприятным внешним условиям)и присутствию мешающих объектов (некоему подобию инфекционного заражения), то получаемые оценки являются робастными (от английского слова *robust* — «здоровый», «крепкий», «стойкий»).

Альтернативный подход к оцениванию базируется на выявлении аномальных наблюдений, их удалении и построении оценки по неполной выборке. В силу возникающего при этом ухудшения репрезентативности такой подход пригоден только при небольшом числе аномальных значений. Особенно существенным ухудшение репрезентативности может быть в случае данных большой размерности, так как по каждой координате признакового пространства аномальными могут быть признаны значения

различных объектов, что обусловливает высокую засоренность всей выборки.

Распространенной причиной появления аномальных значений являются грубые ошибки. Они могут появиться при сборе исходной информации, а также в результате искажения информации в каналах ее передачи. Их причиной может также служить некорректный предварительный содержательный анализ исходных характеристик изучаемых объектов. Известны различные методы выявления аномальных наблюдений. Они, как правило, требуют предварительного определения структуры совокупности, в общем случае неоднородной, ее разбиения на однородные группы, каждая из которых характеризуется своим набором параметров.

При оценивании параметров исследуемой совокупности используют методы непосредственного выявления грубых ошибок и методы, сводящие к минимуму искажения, создаваемые этими ошибками. Кроме того, существуют и комбинированные методы, которые позволяют и выделять грубые ошибки, и давать наиболее правдоподобные оценки параметров распределения.

2.Статистические критерии обнаружения засорения выборки

В процессе реализации методов выявления грубых ошибок выборка подвергается своего рода цензуре, и ее называют *цензурированной*.

Примером цензурирования данных является удаление аномальных наблюдений в соответствии с *правилом трех сигм*. Согласно этому правилу практически все наблюдения нормально распределенного признака x отклоняются от своего математического ожидания μ менее чем на три средних квадратических отклонения σ: $P(\mu-3\sigma<x<\mu+3\sigma)=0{,}9973$.

Наблюдения, не попадающие в интервал$(\mu-3\sigma<x<\mu+3\sigma)$, признаются аномальными. Это правило является базовым для построения множества алгоритмов. Данные могут быть подвержены и более строгой «цензуре»,

когда задается более низкий порог ограничения, соответствующий большей вероятности отклонения наблюдения от математического ожидания.

Сложность непосредственного применения правила трех сигм и ему подобных состоит в том, что параметры μ и σ обычно неизвестны, а попытка получения их оценок непосредственно по имеющейся информации приведет к искажениям, обусловленным наличием аномальных значений. Учесть этот эффект можно путем предварительного удаления «подозрительных» наблюдений, оценивания параметров по «очищенной» выборке и применения базового правила (например, трех сигм) для выявления аномалий.

Можно также использовать статистику в виде нормированного значения экстремального отклонения от среднего, оцененного по всей выборке.

Примером алгоритма такого рода является метод Смирнова — Граббса для выявления грубых ошибок измерений.

При проверке на аномальность максимального значения в имеющейся совокупности метод Смирнова — Граббса предусматривает упорядочение результатов N наблюдений x_1, x_2, …, x_N, результат которого можно представить в виде вариационного ряда $x_{(1)} \le x_{(2)} \le \ldots \le x_{(N)}$, (1)

где $x(i)$, в отличие от xi, представляет собой наблюдение со значением признака, большим или равным значениям признака у других не менее чем (i -1) наблюдаемых объектов.

При неизвестных параметрах σ и μ необходимо определить нормированную величину модуля отклонения последнего члена вариационного ряда от среднего значения $T_{(N)} = \frac{|x_{(N)} - \bar{x}|}{\tilde{s}}$, (2)

где $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i;\ \tilde{s}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$ — оценки математического ожидания и дисперсии.

Рассчитанное согласно (2) значение $T(N)$ необходимо сравнить с критическим значением C_α из таблицы Граббса при односторонней альтернативе для соответствующей вероятности ошибки первого рода α.

Если $T_{(N)}>C_\alpha$, то гипотеза H_0 о том, что проверяемое наблюдение является типичным для данной совокупности, отвергается на выбранном уровне значимости α, и это значение признается грубой ошибкой. Иногда в таблицах вместо α используется доверительная вероятность γ, где γ=1-α. Для критического значения $C_\alpha=C_{1-\gamma}$ при справедливости гипотезы H_0 справедливо выражение $P(T_{ext}\leq C_\alpha)=\gamma$.

При проверке на аномальность наименьшей наблюдаемой величины целесообразно построение вариационного ряда в порядке уменьшения значений $x_{(1)} \geq x_{(2)} \geq \ldots \geq x_{(N)}$. При этом можно использовать статистику (2) и описанное выше правило проверки гипотезы.

Наряду с критерием $T(N)$ для проверки предположения, что наибольшее (наименьшее) из наблюдаемых значений является нетипичным, может быть использован эквивалентный ему *G*-критерий в виде отношения суммы квадратов отклонений от средних. Статистика критерия представляет собой частное от деления суммы квадратов отклонений от своего среднего значения $\overline{x}_{(N-1)}$ для выборки с исключенным проверяемым на аномальность значением $x_{(N)}$ на сумму квадратов отклонений от среднего по x, рассчитанную по всем имеющимся данным:

$$G = \frac{\sum_{i=1}^{N-1}(x_{(i)}-\overline{x}_{N-1})^2}{\sum_{i=1}^{N}(x_{(i)}-\overline{x})^2}, \qquad (3)$$

где $\overline{x}_{N-1}=\frac{1}{N-1}\sum_{i=1}^{N-1}x_i$.

Легко показать, что статистики $G_{(N)}$ и $T_{(N)}$ связаны между собой соотношением

$$G_{(N)}=1-\frac{1}{N-1}\left(\frac{x_{(N)}-\overline{x}}{\tilde{s}}\right)^2=1-\frac{1}{N-1}T^2_{(N)}.$$

Критические значения G_α для критерия G можно определить по таблице. Наблюдение относят к нетипичным, если рассчитанное для него значение статистики G окажется меньше критического G_α.

Альтернативным критерием, привлекательным с точки зрения меньших вычислительных затрат, является критерий, предложенный в середине XX в. В. Диксоном. Он основан на статистиках, рассчитываемых путем деления модуля разности экстремального и близкого к нему значения на размах, определяемый либо по всей совокупности, либо по совокупности, редуцированной путем удаления некоторых крайних значений.

Для проверки на аномальность наибольшего значения, т.е. величины $x_{(N)}$, в вариационном ряду (1) эти статистики имеют вид

$$R_{10}=\frac{x_{(N)}-x_{(N-1)}}{x_{(N)}-x_{(1)}};\ R_{11}=\frac{x_{(N)}-x_{(N-1)}}{x_{(N)}-x_{(2)}};\ R_{21}=\frac{x_{(N)}-x_{(N-2)}}{x_{(N)}-x_{(2)}};\ R_{22}=\frac{x_{(N)}-x_{(N-2)}}{x_{(N)}-x_{(3)}}. \quad (4)$$

Каждая из этих статистик рекомендуется к использованию при определенном числе наблюдений, и в соответствии с этими рекомендациями можно определить общую статистику

$$R^{(N)}=\begin{cases} R_{10}, 3\le n\le 7, \\ R_{11}, 8\le n\le 10, \\ R_{21}, 11\le n\le 13, \\ R_{22}, 14\le n\le 30. \end{cases}$$

Критические значения для этой статистики приведены в таблице.

Если рассчитанное значение статистики превышает соответствующее критическое на выбранном уровне значимости α, то наблюдение признается аномальным. Чтобы проверить на аномальность наименьшее наблюдение, необходимо вместо ряда (1) использовать вариационный ряд $x_{(1)} \ge x_{(2)} \ge \ldots \ge x_{(N)}$, построенный в порядке уменьшения значений.

Последовательное многократное применение критерия для проверки на аномальность отдельного значения с целью исключения нескольких выбросов производить не следует. При наличии более одного выброса возникает смещение параметров выборки, препятствующее обнаружению всех выбросов. Проверяемое на аномальность второе значение входит в

расчет средней величины и дисперсии и может таким образом себя маскировать, смещая в свою сторону среднее значение и увеличивая меру разброса значений признака. Это явление называют *маскирующим эффектом*, и для его предотвращения строят процедуры, изначально предназначенные для обнаружения нескольких выбросов. Обобщением критерия Граббса на случай проверки на аномальность нескольких экстремальных наблюдений является критерий Титьена — Мура.

В подходе, предложенном Г. Титьеном и Г. Муром, используется вариационный ряд (1) для проверки на типичность k наибольших наблюдений и вариационный ряд $x_{(1)} \geq x_{(2)} \geq \ldots \geq x_{(N)}$— для проверки k наименьших наблюдений. При этом формируется статистика

$$L_{(k)} = \frac{\sum_{i=1}^{N-k} (x_{(i)} - \bar{x}_k)^2}{\sum_{i=1}^{N} (x_{(i)} - \bar{x})^2}, \qquad (5)$$

где $\bar{x}_k = \frac{\sum_{i=1}^{N-k} x_{(i)}}{N-k}$ — средняя арифметическая первых $(N - k)$ наблюдений вариационного ряда.

Критические значения для $L_{(k)}$ приведены в таблице. Значение $L_{(k)}$ сравнивается с критическим значением. Если рассчитанное значение меньше критического, то k рассматриваемых наблюдений являются грубыми ошибками.

Иначе обстоит дело с проверкой на типичность одновременно наибольших и наименьших значений. Для этого требуется предварительное преобразование исходных данных. Необходим расчет абсолютных отклонений от среднего $z_1 = |x_1 - \bar{x}|, z_2 = |x_2 - \bar{x}|, \ldots, z_N = |x_N - \bar{x}|$ и построение по аналогии с рядом (1) вариационного ряда в порядке возрастания полученных значений с соответствующей переиндексацией величин z_i:

$z_{(1)} \leq z_{(2)} \leq \ldots \leq z_{(N)}$.

Для проверки на типичность k наблюдений, имеющих наибольшие по модулю отклонения от среднего значения, используется статистика

$$E_{(k)} = \frac{\sum_{i=1}^{N-k} (z_{(i)} - \overline{z}_k)^2}{\sum_{i=1}^{N} (z_{(i)} - \overline{z})^2}, \quad (6)$$

где $\overline{z}$ — средняя арифметическая всей выборки;

$\overline{z}_k = \frac{\sum_{i=1}^{N-k} z_i}{N-k}$ — средняя арифметическая из N — k наблюдений, оставшихся после исключения из выборки k элементов с наибольшими по модулю отклонениями от среднего $\overline{z}$.Критические значения для $E_{(k)}$ приведены в таблице.

К недостатку критерия Титьена — Мура можно отнести необходимость априорной информации о числе аномальных значений, при отсутствии которой определение этого числа производится по тем же данным, что и рассчитываемая статистика критерия. Это существенным образом сказывается на фактическом критическом значении. Кроме того, используемые статистики достаточно сильно искажаются при нарушении предположения о нормальности распределения исследуемого признака.

В статистической практике существуют и другие подходы обнаружения «выбросов»:

- устойчивые параметрические подходы, предложенные Тьюки, Хьюбертом, Л. Д. Мешалкиным, Д. Ф. Эндрюсом, Ф. Р. Хампелем и другими;

- оценки на основе порядковых статистик, предложенные Ж. А. Пуанкаре, Ч. П. Винзором и другими;

- непараметрические методы и оценки на основе бутстреп-анализа.

Вопросы для самопроверки:

1. В каких случаях требуется нахождение робастных оценок параметров?

2. В чем состоит основной недостаток непосредственного применения правила «трех сигм» для выявления аномальных наблюдений?

3. Укажите, как выявить аномальные наблюдения с помощью критерия Диксона.

4. В каком случае критерий Титьена — Мура эквивалентен критерию Граббса при выявлении аномальных наблюдений?

5. Для чего производится цензурирование выборки?

6. Укажите, как получить робастные оценки параметров линейной регрессионной модели.

Лекция 5. Задачи статистического сравнения

1. Статистическое сравнение параметров со стандартом и проверка гипотез о законе распределения

В процессе проведения процедур анализа данных часто возникает необходимость в сравнение нескольких одномерных или многомерных генеральных совокупностей на основе случайных выборок по основным статистическим характеристикам (прежде всего по параметрам центра распределения и параметрам рассеивания относительно центра).

В процессе сравнения параметров нескольких генеральных совокупностей может возникнуть вопрос о выборочной оценке (точечной или интервальной) основных параметров одномерной генеральной совокупности. В случае многомерной генеральной совокупности также может понадобиться нахождение основных оценок параметров, определение доверительной области для вектора математических ожиданий и генеральных дисперсий.

Также иногда необходимо сравнение со стандартом генерального среднего, генеральной дисперсии и генеральной доли. Такие задачи решают на основе соответствующих статистических критериев проверяющих гипотезы вида $H_0: p = p_0$, $\mu = \mu_0$, $\sigma^2 = \sigma_0^2$.

В случае многомерной генеральной совокупности речь будет идти, например, о сравнение вектора генеральных средних со стандартом. Для проверки в данном случае, как правило, опираются на статистику Хотеллинга при условии, что известна ковариационная матрица многомерной генеральной совокупности (или ее оценка). Данные статистические тесты подробно рассматриваются в классической математической статистике.

Другой важный вопрос - проверка гипотез о законе распределения значений признака X в генеральной совокупности. Осуществляется такая проверка с помощью критериев согласия, предназначенных для проверки гипотезы H_0 о том, что ряд наблюдений $x_1, x_2, \ldots, x_n$ образует случайную выборку, извлеченную из генеральной совокупности X с функцией

распределения $F(x) = F(x; \theta_1, \theta_2, ..., \theta_k)$, где общий вид функции $F(x)$ считается известным, а параметры $\theta_1, \theta_2, ..., \theta_k$ могут быть как известными, так и неизвестными.

Критерий согласия основан на использовании различных мер расстояний между анализируемой эмпирической функцией $F^*_n(x)$ распределения, определенной по выборке, и функцией распределения $F(x)$ генеральной совокупности X. Наиболее широко известны критерии согласия Пирсона и критерий Колмогорова.

2. Статистическое сравнение параметров нескольких совокупностей

Вначале рассмотрим гипотезы о сравнение основных параметров одномерных генеральных совокупностей:

Проверка гипотезы о равенстве генеральных средних двух нормальных совокупностей при неизвестных генеральных дисперсиях.

Пусть X и Y — нормальные совокупности с равными, но неизвестными дисперсиями $\sigma_x^2 = \sigma_y^2 = \sigma^2$ и математическими ожиданиями μ_xи μ_y. Из этих совокупностей взяты две случайные независимые выборки с параметрами $\overline{x}, S_x^2$ и $\overline{y}, S_y^2$.

На уровне значимости αтребуется проверить нулевую гипотезу H_0: $\mu_x=\mu_y$.В основу критерия для проверки нулевой гипотезы положена статистика

$$t_{\text{набл}} = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{n_x S_x^2 + n_y S_y^2}{n_x + n_y - 2}}} \sqrt{\frac{n_x n_y}{n_x + n_y}}, \quad (1)$$

которая при выполнении нулевой гипотезы H_0 имеет распределение Стьюдента с $\upsilon=n_x+n_y-2$ степенями свободы.

При заданном уровне значимости αвыбор критической области зависит от конкурирующей гипотезы: при H_1: $\mu_x>\mu_y$выбирают правостороннюю, при

H_1: $\mu_x < \mu_y$— левостороннюю, а при H_1: $\mu_x \neq \mu_y$ — двустороннюю критические области.

Критерий проверки гипотезы заключается в следующем: если | *t*набл | >$t_{кр}$, где $t_{кр} = St^{-1}(2\alpha; n_x + n_y - 2)$ (для правосторонней и левосторонней критических областей) или $t_{кр} = St^{-1}(\alpha; n_x + n_y - 2)$ (для двусторонней критической области), то гипотезу отвергают, если же | *t*набл | ≤*t*кр, то делают вывод, что гипотеза не противоречит опытным данным.

Проверка гипотезы о равенстве дисперсий двух генеральных совокупностей.

Пусть *X* и *Y* — генеральные совокупности, значения признаков которых распределены по нормальному закону с дисперсиями σ_x^2 и σ_y^2.

Из этих совокупностей взяты независимые случайные выборки объемом n_x и n_y, и пусть $\hat{S}_x^2$ и $\hat{S}_y^2$ — исправленные выборочные дисперсии, причем $\hat{S}_x^2 > \hat{S}_y^2$, где $\hat{S}_x^2 = \frac{1}{n_x - 1}\sum_{i=1}^{n_x}(x_i - \overline{x})^2$, $\hat{S}_y^2 = \frac{1}{n_y - 1}\sum_{i=1}^{n_y}(y_i - \overline{y})^2$.

Требуется проверить нулевую гипотезу $H_0: \sigma_x^2 = \sigma_y^2$ против альтернативной гипотезы $H_1: \sigma_x^2 > \sigma_y^2$. Основу критерия для проверки нулевой гипотезы составляет статистика $F_{набл} = \frac{\hat{S}_x^2}{\hat{S}_y^2}$, (2) которая при выполнении нулевой гипотезы имеет распределение Фишера — Снедекора (*F*-распределение) с n_x-1 и n_y-1 степенями свободы.

Для проверки гипотезы выбирают правостороннюю критическую область. Границу критической области $F_{кр}(\alpha; n_x - 1; n_y - 1)$ определяют по таблице *F*-распределения при заданном уровне значимости α и числе степеней свободы n_x-1 и n_y-1 из условия $P(F > F_{кр}(\alpha; n_x - 1; n_y - 1)) = \alpha$.

Критерий проверки гипотезы состоит в том, что при выполнении условия $F_{набл} \leq F_{кр}(\alpha; n_x - 1; n_y - 1)$ полагают, что гипотеза не

противоречит опытным данным; а если $F_{\text{набл}} \geq F_{\text{кр}}(\alpha;\, n_x - 1;\, n_y - 1)$, то гипотезу отвергают с вероятностью ошибки α.

Проверка гипотез об однородности ряда дисперсий.

а) Критерий Бартлетта.

Пусть X_1, X_2, …, X_l есть l нормальных генеральных совокупностей, из которых извлечены выборки объемом n_1, n_2, …, n_l соответственно, и пусть $\hat{S}_1^2, \hat{S}_2^2, \ldots, \hat{S}_l^2$ — исправленные выборочные дисперсии.

Требуется на уровне значимости αпроверить нулевую гипотезу о равенстве дисперсий l генеральных совокупностей, т.е. $H_0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_l^2$.

Введем обозначения:

- $\nu_i = n_i - 1$ — число степеней свободы i-й выборки;
- $\hat{S}_i^2 = \dfrac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2$ для i=1, …, l, где x_{ij}— результат j-го наблюдения i-й выборки;
- $\hat{S}_{\text{ср}}^2 = \dfrac{\sum_{i=1}^{l}\nu_i \hat{S}_i^2}{\sum_{i=1}^{l}\nu_i}$.

В качестве выборочной характеристики критерия Бартлетт предложил использовать статистику

$$\chi^2_{\text{набл}} = \frac{\nu \ln \hat{S}_{\text{ср}}^2 - \sum_{i=1}^{l}\nu_i \hat{S}_i^2}{1 + \dfrac{1}{3(l-1)}\left(\sum_{i=1}^{l}\dfrac{1}{\nu_i} - \dfrac{1}{\nu}\right)}, \qquad (3)$$

где $\nu = \sum_{i=1}^{l}\nu_i$.

При выполнении нулевой гипотезы H_0 и при $\nu_i > 3$ $\chi^2_{\text{набл}}$ приближенно имеет распределение χ^2 с l -1 степенями свободы.

Для проверки нулевой гипотезы строят правостороннюю критическую область, границу которой определяют по таблице распределения χ^2для уровня значимости αи числа степеней свободы l -1 из условия

$$P(\chi^2 > \chi^2_{кр}(\alpha; l-1)) = \alpha.$$

Критерий проверки гипотезы заключается в следующем: если выполняется условие $\chi^2_{набл} > \chi^2_{кр}(\alpha; l-1)$, то гипотезу отвергают, в противном случае считают, что гипотеза не противоречит опытным данным.

Критерий Бартлетта весьма чувствителен к отклонениям законов распределений X_i для i=1, …, l от нормального закона.

В случае, когда n_1=…=n_l, для проверки нулевой гипотезы H_0 используют критерий Кохрана.

б) Критерий Кохрана.

Пусть X_1, X_2, …, X_l — нормальные генеральные совокупности с неизвестными дисперсиями $\sigma_1^2, \sigma_2^2, ..., \sigma_l^2$, из которых взяты независимые случайные выборки одинакового объема n_1=n_2=…=n_l=n, и пусть $\hat{S}_1^2, \hat{S}_2^2, ..., \hat{S}_l^2$ — исправленные выборочные дисперсии соответствующих совокупностей. Требуется проверить нулевую гипотезу H_0: $\sigma_1^2 = \sigma_2^2 = ... = \sigma_l^2$.

С целью проверки нулевой гипотезы Кохран предложил критерий, основанный на статистике

$$G_{набл} = \frac{\hat{S}^2_{max}}{\hat{S}_1^2 + \hat{S}_2^2 + ... + \hat{S}_l^2} = \frac{S^2_{max}}{S_1^2 + S_2^2 + ... + S_l^2}, \quad (4)$$

которая при выполнении нулевой гипотезы имеет G-распределение с υ = n - 1 степенями свободы и k сравниваемыми совокупностями, где $\hat{S}^2_{max}$ — наибольшая из исправленных выборочных дисперсий.

Для проверки нулевой гипотезы H_0 на уровне значимости α строят правостороннюю критическую область.

Границу критической области $G_{кр}$ находят по таблице G-распределения из условия $P(G > G_{кр}(\alpha; n-1; l)) = \alpha.$

Критерий проверки гипотезы заключается в следующем: если выполняется условие $G_{набл} > G_{кр}(\alpha;\ n\ -1;\ l)$, то гипотезу отвергают, в противном случае считают, что гипотеза не противоречит опытным данным.

Проверка гипотез об однородности ряда вероятностей.

Пусть X_1, X_2, …,X_l-l генеральных совокупностей, каждая из которых характеризуется неизвестным параметром P_i, где P_i— вероятность появления события A в соответствующей выборке.

Требуется по результатам выборочных наблюдений проверить нулевую гипотезу о равенстве вероятностей появления события A в генеральных совокупностях, т.е. H_0: $P_1=P_2=\ldots=P_l=P$. Для проверки гипотезы можно использовать статистику

$$\chi^2_{набл} = \sum_{i=1}^{l} \frac{(m_i - n_i\hat{p})^2}{n_i\hat{p}(1-\hat{p})}, \qquad (5)$$

где $\hat{p} = \dfrac{\sum_{i=1}^{l} m_i}{\sum_{i=1}^{l} n_i}$ — средняя частость появления события A по всем выборкам.

Или, переходя к частостям по всем выборкам $\hat{p}_i = \dfrac{m_i}{n_i}$, получим статистику

$$\chi^2_{набл} = \frac{1}{\hat{p}(1-\hat{p})}\sum_{i=1}^{l}(\hat{p}_i - \hat{p})^2 n_i.$$

Статистика $\chi^2_{набл}$при выполнении нулевой гипотезы имеет асимптотическое распределение χ^2 с l -1 степенями свободы, где l — число генеральных совокупностей.

Для проверки нулевой гипотезы на уровне значимости α строят правостороннюю критическую область, границу $\chi^2_{кр}$ которой определяют из условия $P(\chi^2 > \chi^2_{кр}(\alpha; l-1)) = \alpha.$

Критерий проверки гипотезы заключается в следующем: если выполняется условие $\chi^2_{\text{набл}} > \chi^2_{\text{кр}}(\alpha; l-1)$, то гипотезу отвергают, в противном случае считают, что гипотеза не противоречит опытным данным.

Проверка однородности ряда вероятностей в случае полиномиального распределения.

Пусть X_1, X_2, …, X_l есть l генеральных совокупностей, из которых взяты случайные независимые выборки объемом n_1, n_2, … , n_l и пусть n_i элементов i-й выборки классифицируются по какому-либо признаку на h групп с числом элементов в каждой группе m_{i1}, m_{i2}, … , m_{ij}, … , m_{ih}, где j=1, 2, …, h.

Очевидно, что $\sum_{j=1}^{h} m_{ij} = n_i$ для всех i=1, 2, …, l. В результате классификации элементов по h группам мы получим lh чисел, которые представим в табличном виде:

Номер выборки	Число элементов в группах				Итого по строкам
1	m_{11}	m_{12}	…	m_{1h}	n_1
2	m_{21}	m_{22}	…	m_{2h}	n_2
…	…	…	…	…	…
…	…	…	…	…	…
l	m_{l1}	m_{l2}	…	m_{lh}	n_l
Итого по столбцам	$\sum_{i=1}^{l} m_{i1}$	$\sum_{i=1}^{l} m_{i2}$	…	$\sum_{i=1}^{l} m_{ih}$	N

Здесь $N = \sum_{i=1}^{l} n_i = \sum_{i=1}^{l}\sum_{j=1}^{h} m_{ij}$ — общее число всех наблюдений по всем выборкам.

Требуется проверить гипотезу о том, что вероятность P_j, j=1, 2, …, h попадания элемента в соответствующую группу равна для всех совокупностей, т.е. нулевую гипотезу H_0: P_{1j}=P_{2j}=…=P_{lj}=Pj для всех j=1, 2, …, h. Нулевую гипотезу проверяют против конкурирующей гипотезы H_1, состоящей в том, что вероятности не равны.

В качестве критерия можно использовать статистику

$$\chi^2_{\text{набл}} = \sum_{j=1}^{h}\sum_{i=1}^{l} \frac{(m_{ij} - n_i\hat{p}_j)^2}{n_i\hat{p}_j}, \qquad (6)$$

где $\hat{p}_j = \frac{\sum_{i=1}^{l} m_{ij}}{N}$ для всех j=1, 2, …, h.

При справедливости нулевой гипотезы H_0 статистика $\chi^2_{\text{набл}}$ имеет распределение χ^2 с $\nu = (l-1)(h-1)$ степенями свободы.

Для проверки нулевой гипотезы на уровне значимости α строят правостороннюю критическую область, границы которой определяются из условия $P(\chi^2 > \chi^2_{\text{кр}}(\alpha;\nu)) = \alpha.$

Критерий проверки гипотезы заключается в следующем: если $\chi^2_{\text{набл}} > \chi^2_{\text{кр}}$, то гипотезу отвергают, в противном случае считают, что гипотеза не противоречит опытным данным.

Вопросы для самопроверки:

1. Что понимают под статистической гипотезой и статистическим критерием, ошибками 1-го и 2-го рода?

2. Какие критерии используют для проверки гипотез относительно математических ожиданий одной и нескольких совокупностей?

3. Какие критерии используют для проверки гипотез относительно дисперсий одной и нескольких генеральных совокупностей?

4. Какие критерии используют для проверки гипотез относительно вероятностей p одной и нескольких генеральных совокупностей?

Лекция 6. Статистический анализ многомерных совокупностей

1. Сравнение многомерных совокупностей

Рассмотрим гипотезу о сравнение двух многомерных генеральных совокупностей:

Генеральные совокупности однородны, если кроме одних и тех же признаков они имеют одинаковые законы распределения вероятностей.

Рассмотрим две нормально распределенные совокупности X и Y. Их распределения полностью определяются заданием параметров μ_x, Σ_xи μ_y, Σ_y.

Следовательно, для проверки однородности этих совокупностей достаточно сравнить их ковариационные матрицы Σ_xи Σ_y. Затем в случае принятия гипотезы о равенстве этих ковариационных матриц надо сравнить генеральные средние μ_xи μ_yсовокупностей.

Для сравнения матриц генеральных коэффициентов ковариации проверяется гипотеза $H_0: \Sigma_x = \Sigma_y$ против $H_1: \Sigma_x \neq \Sigma_y$ с уровнем значимости α на основе выборок из совокупностей соответственно объемов n_x и n_y.

В качестве статистики критерия проверки берется случайная величина $W=ba$, где

$$b = 1 - \left(\frac{1}{n_x - 1} + \frac{1}{n_y - 1} - \frac{1}{n_x + n_y - 2} \right) \frac{2k^2 + 3k - 1}{6(k+1)};$$

$$a = (n_x + n_y - 2)\ln|\hat{S}_{xy}| - [(n_x - 1)\ln|\hat{S}_x| + (n_y - 1)\ln|\hat{S}_y|]; \quad (1)$$

$\hat{S}_x$ — несмещенная оценка ковариационной матрицы с элементами

$$\hat{s}_{xml} = \frac{1}{n_x - 1} \sum_{i=1}^{n_x} (x_{im} - \overline{x}_m)(x_{il} - \overline{x}_l);\ m, l = 1, \ldots, k;$$

$\hat{S}_y$ — несмещенная оценка ковариационной матрицы с элементами

$$\hat{s}_{yml} = \frac{1}{n_y - 1} \sum_{i=1}^{n_y} (y_{im} - \overline{y}_m)(y_{il} - \overline{y}_l);\ m, l = 1, \ldots, k;$$

$$\hat{S}_{xy} = \frac{1}{n_x + n_y - 2}[(n_x - 1)\hat{S}_x + (n_y - 1)\hat{S}_y]$$ — несмещенная оценка ковариационной матрицы, полученная по суммарной выборке, так как $\Sigma_x = \Sigma_y$.

При справедливости гипотезы H_0, достаточно больших n_x и n_y, а также при достаточно малой величине

$$C = \frac{k(k+1)}{48b^2}\left\{(k-1)(k+2)\left[\frac{1}{(n_x-1)^2} + \frac{1}{(n_y-1)^2} + \frac{1}{(n_x-1)^2 + (n_y-1)^2}\right] - 6(1-b)^2\right\}$$

статистика W аппроксимируется распределением χ^2 с числом степеней свободы $\nu = \frac{k(k+1)}{2}$.

Таким образом, критическая область имеет вид

$$W > W_{кр} = \chi^2\left(\alpha; \frac{k(k+1)}{2}\right).$$

Если $W_{набл} = ba$ попадает в критическую область (Wнабл $> W$кр), то гипотеза $H_0: \Sigma_x = \Sigma_y$ отвергается с вероятностью ошибки α. Тогда считается доказанным, что ковариационные матрицы Σ_x и Σ_y не равны и, следовательно, генеральные совокупности неоднородны.

Если $W_{набл}$ не попало в критическую область (Wнабл $\le W$кр), то гипотеза $H_0: \Sigma_x = \Sigma_y$ не противоречит наблюдениям и обычно принимается, т.е. считается, что ковариационные матрицы Σ_x и Σ_y равны. После этого переходят к сравнению генеральных средних, т.е. к проверке на уровне значимости α гипотезы $H_0: \mu_x = \mu_y$ при $H_1: \mu_x \neq \mu_y$.

Для проверки применяется критерий, основанный на статистике Хотеллинга вида

$$T^2 = \frac{n_x n_y}{n_x + n_y}(\bar{x} - \bar{y})^T \hat{S}_{xy}^{-1}(\bar{x} - \bar{y}). \quad (2)$$

Если гипотеза $H_0: \mu_x = \mu_y$ справедлива, то статистики T^2 и F связаны формулой

$$T_{кр}^2 = \frac{(n_x + n_y - 2)k}{n_x + n_y - k - 1} F_{\alpha;k;n_x+n_y-k-1}, \quad (3)$$

где $F_{\alpha;k;n_x+n_y-k-1}$ находится по таблицам F-распределения Фишера — Снедекора с числом степеней свободы в числителе $\upsilon_1=k$ и в знаменателе $\upsilon_2=n_x+n_y-k-1$. Критическая область имеет вид $T^2>T^2_{кр}$.

Если гипотеза $H_0: \mu_x = \mu_y$ отвергается с вероятностью ошибки α, то считается доказанной неоднородность генеральных совокупностей X и Y. Если же гипотеза $H_0: \mu_x = \mu_y$ не отвергается, то, принимая эту гипотезу, мы считаем, что генеральные совокупности однородны.

2. Основы дисперсионного анализа

Дисперсионный анализ — это статистический метод анализа результатов наблюдений, зависящих от различных, одновременно действующих факторов, выбор наиболее важных факторов и оценка их влияния. Дисперсионный анализ находит применение в различных областях науки и техники. Идея дисперсионного анализа, как и сам термин «дисперсия», принадлежат Р.Фишеру.

Суть анализа заключается в разложении общей вариации случайной величины на независимые слагаемые, каждое из которых характеризует влияние того или иного фактора или их взаимодействия.

Факторами обычно называют внешние условия, влияющие на эксперимент. Это, например, температура и атмосферное давление, сила тяготения, тип оборудования и т. п. Нас интересуют факторы, действие которых значительно и поддается проверке. В условиях эксперимента факторы могут варьировать, благодаря чему можно исследовать влияние контролируемого фактора на эксперимент. В этом случае говорят, что фактор варьирует на разных уровнях или имеет несколько уровней.

В зависимости от количества факторов, включенных в анализ, различают классификацию по одному признаку— *однофакторный анализ,* по двум признакам — *двухфакторный анализ* и многостороннюю —

перекрестную классификацию, изучением которой занимается многофакторный анализ.

Для проведения дисперсионного анализа необходимо соблюдать следующие условия: результаты наблюдений должны быть независимыми случайными величинами, имеющими нормальное распределение и одинаковую дисперсию. Только в этом случае можно оценить значимость полученных оценок дисперсий и математических ожиданий и построить доверительные интервалы.

На практике возможен случай, когда на автоматической линии несколько станков параллельно выполняют некоторую операцию. Для правильного планирования последующей обработки важно знать, насколько однотипны средние размеры деталей, получаемые на параллельно работающих станках. Здесь имеет место лишь один фактор, влияющий на размер деталей, — станки, на которых они изготовляются. Исследователя интересует, насколько существенно влияние этого фактора на размеры деталей?

Предположим, что совокупности размеров деталей, изготовленных на каждом станке, имеют нормальное распределение и равные дисперсии. Имеем *m* станков, следовательно, *m* совокупностей или уровней, на которых произведенон$_1$, n$_2$,..., n$_m$наблюдений. Для простоты рассуждений положим,

что n$_1$= n$_2$ = ... = n_m. Размеры деталей, составляющие n_i наблюдений на i-м уровне, обозначим x_{i1}, x_{i2},..., x_{in}. Тогда все наблюдения можно представить в виде таблицы, которую назовем *матрицей наблюдений*.

Табл.5.1

Матрица наблюдений

Уро-вень	Наблюдение					
	1	2	...	*j*	...	*n*
1	x_{11}	x_{12}	...	x_{1j}	...	x_{1n}
2	x_{21}	x_{22}	...	x_{2j}	...	x_{2n}
...	...	...	...	...	...	...
i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{in}
...	...	...	...	...	...	...
m	x_{m1}	x_{m2}	...	x_{mj}	...	x_{mn}

Будем полагать, что для 1-го уровня *n* наблюдений имеют среднюю β_i, равную сумме общей средней μ и вариации ее, обусловленной i-м уровнем фактора, т. е. $\beta_i = \mu + \gamma_i$. Тогда одно наблюдение можно представить в следующем виде:

$$x_{ij} = \mu + \gamma_i + \xi_{ij} = \beta_i + \xi_{ij}, \quad (4)$$

где μ — общая средняя; γ_i — эффект, обусловленный *i*-м уровнем фактора; ξ_{ij} — вариация результатов внутри отдельного уровня.

Член ξ_{ij}, характеризует влияние всех, не учтенных моделью факторов. Согласно общей задаче дисперсионного анализа нужно оценить существенность влияния фактора γ на размеры деталей. Общую вариацию переменной x_{ij} можно разложить на части, одна из которых характеризует влияние фактора γ, другая — влияние неучтенных факторов.

Для этого необходимо найти оценку общей средней μ и оценки средних по уровням β_i. Очевидно, что оценкой β является средняя арифметическая n наблюдений i-го уровня, т.е.

$$\bar{x}_{i*} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}. \quad (5)$$

Звездочка в индексе при *x* означает, что наблюдения фиксированы на i-м уровне. Средняя арифметическая всей совокупности наблюдений является оценкой общей средней *μ,* т.е.

$$\bar{x} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij} \text{ или } \bar{x} = \frac{1}{m}\sum_{i=1}^{m} \bar{x}_{i*}. \quad (6)$$

Найдем сумму квадратов отклонений x_{ij} от $\bar{x}$, т.е. $\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x})^2$.

Представим ее в виде

$$SS = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x})^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i*} + \bar{x}_{i*} - \bar{x})^2 =$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i*})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n}(\bar{x}_{i*} - \bar{x})^2 + 2\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i*})(\bar{x}_{i*} - \bar{x}). \quad (7)$$

Причем $\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i*})(\bar{x}_{i*} - \bar{x}) = \sum_{j=1}^{n}(x_{ij} - \bar{x}_{i*})\sum_{i=1}^{m}(\bar{x}_{i*} - x)$.

Но $\sum_{j=1}^{n}(x_{ij}-\bar{x}_{i*})=0,$ так как это есть сумма отклонений переменных одной совокупности от средней арифметической этой же совокупности, т.е. вся сумма равна 0. Второй член суммы запишем в виде $\sum_{i=1}^{m}\sum_{j=1}^{n}(\bar{x}_{i*}-\bar{x})^2=n\sum_{i=1}^{m}(\bar{x}_{i*}-\bar{x})^2.$

Тогда основное тождество можно представить следующим образом:

$$\underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\bar{x})^2}_{SS}=\underbrace{n\sum_{i=1}^{m}(\bar{x}_{i*}-\bar{x})^2}_{SS_1}+\underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\bar{x}_{i*})^2}_{SS_2};$$

или $SS=SS_1+SS_2.$ (8)

Слагаемое SS_1 является суммой квадратов разностей между средними уровней и средней всей совокупности наблюдений.

Эта сумма называется *суммой квадратов отклонений между группами* и характеризует расхождение между уровнями. Величину SS_1 называют также *рассеиванием по факторам***,** т.е. рассеиванием за счет исследуемого фактора.

Слагаемое SS_2является суммой квадратов разностей между отдельными наблюдениями и средней i-го уровня. Эта сумма называется *суммой квадратов отклонений внутри группы* и характеризует расхождение между наблюдениями i-го уровня. Величину SS_2называют также *остаточным рассеиванием***,** т.е. рассеиванием за счет неучтенных факторов. Наконец, *SS* называется *общей или полной суммой квадратов отклонений отдельных наблюдений от общей средней* $\bar{x}$.

Зная суммы квадратов *SS*, SS_1и SS_2, можно оценить соответствующие дисперсии: общую, межгрупповую и внутри групповую.

Оценим дисперсииs_1^2,s_2^2,s^2:

$$MS_1=s_1^2=\frac{1}{m-1}\sum_{i=1}^{m}(\bar{x}_{i*}-\bar{x})^2=\frac{SS_1}{m-1};$$

$$MS_2=s_2^2=\frac{1}{m(n-1)}\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\bar{x}_{i*})^2=\frac{SS_2}{m(n-1)};$$

$$s^2=\frac{1}{mn-1}\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\bar{x})^2=\frac{SS}{mn-1}.$$ (9)

Если влияние всех уровней фактора у одинаково, то s_1^2*и* s_2^2—оценки общей дисперсии. Тогда для оценки существенности влияния фактора у

достаточно проверить гипотезуH_0:$s_1^2=s_2^2$; для этого вычисляют статистику $F_B = s_1^2/s_2^2$с $\kappa_1 = \boldsymbol{m}$- 1 и $\kappa_2 = m(n - 1)$ степенями свободы.

Затем по таблице F-распределения для уровня значимости α находят критическое значение $F_{\alpha,k1,k2}$. Если$F_B > F_{\alpha,k1,k2}$, то нулевая гипотеза отвергается и делается заключение о существенном влиянии фактора γ. При $F_B < F_{\alpha,k1,k2}$ нет основания отвергать нулевую гипотезу и считать, что влияние фактора γ несущественно.

Сравнивая межгрупповую и остаточную дисперсии, по величине их отношения судят, насколько сильно проявляется влияние факторов. Однофакторный дисперсионный анализ удобно представить в виде таблицы (табл.5.2)

Табл.5.2

Однофакторный дисперсионный анализ

Компонента дисперсии	Сумма квадратов	Число степеней свободы k	Оценка дисперсий
Межгрупповая	$\sum_i (\bar{x}_{i\cdot} - \bar{x})^2$	$m-1$	$s_1^2 = \frac{1}{m-1}\sum_i (\bar{x}_{i\cdot} - \bar{x})^2$
Внутригрупповая	$\sum_{ij} (x_{ij} - \bar{x}_{i\cdot})^2$	$m(n-1)$	$s_2^2 = \frac{1}{m(n-1)}\sum_{ij} (x_{ij} - \bar{x}_{i\cdot})^2$
Полная (общая)	$\sum_{ij} (x_{ij} - \bar{x})^2$	$mn-1$	$s^2 = \frac{1}{mn-1}\sum_{ij} (x_{ij} - \bar{x})^2$

Если на результативный *признак влияют* несколько *факторов одновременно***,** то *имеет* место *многофакторный* анализ. Дисперсионный анализ в этом случае *имеет* свои особенности, так как *необходимо учитывать* взаимодействия *между факторами***.**

Частным случаем дисперсионного анализа при классификации по двум признакам является ситуация, когда в ячейке одно наблюдение и взаимодействие между факторами отсутствует. В общем случае в ячейке может быть несколько наблюдений (как равное количество, так и неравное) и между факторами может иметь место взаимодействие.

Когда в ячейке равное количество наблюдений, то при этом вычисления упрощаются.

Вопросы для самопроверки:

1. Каким количеством параметров может быть задана многомерная нормальная совокупность?

2. Что понимают под ковариационной матрицей? Какими свойствами она обладает?

3.На какую статистику опираются (в случае известной ковариационной матрицы) при проверке гипотезы о равенстве вектора генеральных средних заданному стандарту?

4. На какую статистику опираются (в случае неизвестной ковариационной матрицы) при проверке гипотезы о равенстве вектора генеральных средних заданному стандарту?

5. Какая формула связывает статистику Хотеллинга и Фишера-Снедекора при истинности гипотезы о равенстве вектора генеральных средних заданному стандарту?

6.Какая формула связывает статистику Хотеллинга и Фишера-Снедекора при истинности гипотезы о равенстве вектора генеральных средних двух нормальных совокупностей?

7. Для чего проводят дисперсионный анализ? Как классифицируют дисперсионный анализ в зависимости от количества факторов?

Раздел 3. Статистический анализ взаимосвязи

Лекция 7. Многомерный линейный корреляционый анализ

1.Оценка взаимосвязи между парами признаков

При статистическом анализе многомерной генеральной совокупности каждый из обследуемых объектов характеризируется набором признаков $X = (x_1, x_2, ..., x_p)^T$, а результат регистрации значений этих признаков для *i*-го объекта представляет собой *i*-е многомерное наблюдение и может быть представлен с помощью вектора $X_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$, где $i = 1,2,...,n$.

Различают два вида взаимосвязей между экономическими явлениями: функциональную и стохастическую (вероятностную). При функциональной взаимосвязи имеет место однозначность отображения множества значений изучаемых величин, то есть существует правило y = f(x) – соответствия независимой переменной x и зависимой переменной y. В экономике примером функциональной связи может служить зависимость производительности труда от объема произведенной продукции и затрат рабочего времени.

При изучении массовых явлений зависимость между наблюдаемыми величинами проявляется часто лишь в случае, когда число единиц изучаемой совокупности достаточно велико. При этом каждому фиксированному значению аргумента соответствует определенный закон распределения значений функции и, наоборот, заданному значению зависимой переменной соответствует закон распределения объясняющей переменной. Например, при изучении потребления электроэнергии **y** в зависимости от объема производства **x** каждому конкретному значению переменной **x** соответствует не одно, а множество значений переменной **y** и наоборот. В этом случае существует стохастическая (корреляционная) связь между переменными.

Множественность результатов при анализе связи **x** и **y** объясняется тем, что зависимая переменная **y** испытывает влияние не только фактора **x**, но и целого ряда других факторов, которые не учитываются. Кроме того, влияние

выделенного фактора может быть не прямым, а проявляется через цепочку других факторов.

Корреляционный анализ, разработанный К. Пирсоном(1857–1936) и Дж. Юлом (1871–1951), является одним из методов статистического анализа взаимозависимости нескольких признаков – компонент случайного вектора *x*. Использование методов корреляционного анализа позволяет ответить на такие интересующие исследователя вопросы как, например, «Связан ли уровень безработицы в стране с ВВП?», «Существует ли зависимость между доходом семьи и ее расходами на питание?» и т.д. Он применятся тогда, когда данные наблюдений можно считать случайными и выбранными из генеральной совокупности, распределенной по многомерному нормальному закону.

В корреляционном анализе многомерной генеральной совокупности рассматриваются следующие вопросы:

1. Как выбрать с учетом специфики и природы анализируемых переменных подходящий измеритель тесноты статистической связи (коэффициент корреляции, корреляционное отношение, ранговый коэффициент корреляции и т.д.).

2. Как оценить с помощью точечной и интервальной оценок его числовое значение по имеющимся выборочным данным.

3. Как проверить гипотезу о том, что полученное значение анализируемого измерителя связи действительно свидетельствует о наличие статистической связи, *т.е. проверить исследуемую корреляционную характеристику на статистическое значимое отличие от нуля*.

4. Как определить структуру связей между компонентами исследуемого многомерного признака, сопоставив каждой паре ответ: *связь есть или нет.*

Характеристики статистической связи, рассматриваемые в корреляционном анализе используются в качестве «входной» информации при определении вида зависимости; снижении размерности анализируемого признакового пространства; классификации объектов и признаков. Поэтому с

корреляционного анализа начинаются все многомерные статистические исследования.

Основная задача корреляционного анализа состоит в оценке корреляционной матрицы генеральной совокупности по выборке и определении на ее основе оценок частных и множественных коэффициентов корреляции и детерминации.

Исходной для анализа является матрица «объект-признак»

$$X = \begin{pmatrix} x_{11} \cdots x_{1j} \cdots x_{1p} \\ \cdots\cdots\cdots\cdots \\ x_{i1} \cdots x_{ij} \cdots x_{ip} \\ \cdots\cdots\cdots\cdots \\ x_{n1} \cdots x_{nj} \cdots x_{np} \end{pmatrix}$$

размерности (nxp), i-я строка которой характеризует i-е наблюдение(объект) по всем p-м показателям (j = 1, 2, ..., p).

В корреляционном анализе количественных признаков матрицу X рассматривают как выборку объема n из p-мерной генеральной совокупности, подчиняющейся p-мерному нормальному закону распределения.

Парный коэффициент корреляции характеризует тесноту линейной зависимости между двумя переменными (x1 и x2) на фоне действия всех остальных переменных, входящих в модель. Парный коэффициент корреляции ρ в силу своих свойств является одним из самых распространенных способов измерения связи между случайными величинами в генеральной совокупности; для выборочных данных используется эмпирическая мера связи r. Коэффициент корреляции не имеет размерности и, следовательно, его можно сопоставлять для разных статистических совокупностей. Величина его лежит в пределах (-1 до +1).

Значение ρ=±1 свидетельствует о наличии функциональной зависимости между рассматриваемыми признаками. Если ρ=0, можно сделать вывод, что линейная связь между x и y отсутствует, однако это не означает, что они статистически независимы. В этом случае не отрицается

возможность существования иной формы зависимости между переменными, которую можно оценить с помощью нелинейных измерителей взаимосвязи, таких как индекса корреляции или эмпирическое корреляционное отношение (рис. 7.1).

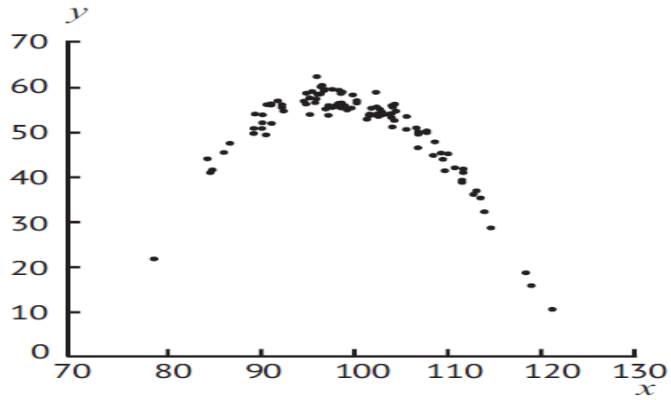

Рис 7.1. Нелинейная взаимосвязь между x и y

Положительный знак коэффициента корреляции указывает на положительную корреляцию. Отрицательный знак коэффициента свидетельствует об отрицательной корреляции. Чем ближе значение |r| к единице, тем связь теснее, приближение |r| к нулю означает ослабление линейной зависимости между переменными. При |r|=1 корреляционная связь перерождается в функциональную.

На практике при изучении зависимости между двумя случайными величинами используют поле корреляции, с помощью которого можно установить наличие корреляционной зависимости. Поле корреляции представляет собой диаграмму, на которой изображается совокупность значений двух признаков. Каждая точка этой диаграммы имеет координаты (xi, yi), соответствующие размерам признаков в i-м наблюдении.

Три варианта распределения точек на поле корреляции показаны на рисунке 7.2.

В первом случае основная масса точек укладывается в эллипсе, главная диагональ которого образует положительный угол с осью X. Это график положительной корреляции. Второй вариант распределения соответствует отрицательной корреляции. Равномерное распределение точек в пространстве (XУ) свидетельствует об отсутствии корреляционной зависимости.

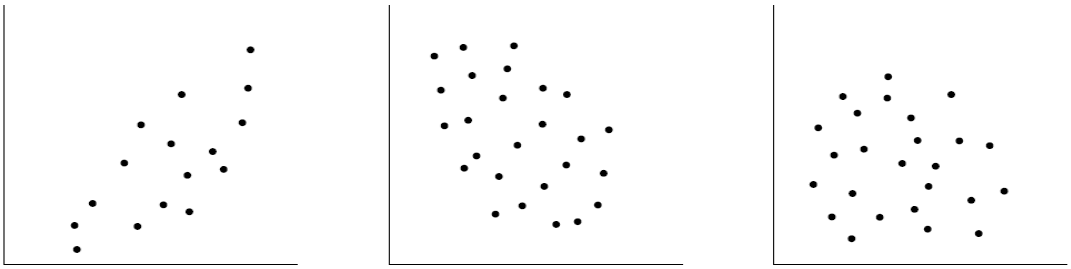

Рис 7.2.Вид поля корреляции в зависимости от характера связи

Отметим, что сила связи не зависит от ее направленности и определяется по абсолютному значению коэффициента корреляции. По имеющимся выборочным данным определяют оценки параметров генеральной совокупности, а именно: вектор средних $(\bar{x})$, вектор среднеквадратических отклонений s и корреляционная матрица (R) порядка *p*:

$$\bar{x} = \begin{pmatrix} \overline{x_1} \\ \overline{x_2} \\ . \\ . \\ . \\ \overline{x_p} \end{pmatrix}, \; s = \begin{pmatrix} s_1 \\ s_2 \\ . \\ . \\ . \\ s_p \end{pmatrix}.$$

Матрица парных коэффициентов корреляции имеет вид:

$$R = \begin{pmatrix} 1 & r_{12} & . & . & r_{1p} \\ r_{21} & 1 & . & . & r_p \\ \hline r_{p1} & r_{p2} & . & . & 1 \end{pmatrix}$$

Матрица R является симметричной ($r_{je} = r_{ej}$) и положительно определенной, где

$$\bar{x} = \frac{1}{n}\sum x_{ij}\,,\; s_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2}\,,$$

$$r_{jl} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \overline{x_j}\right)\left(x_{il} - \overline{x_l}\right)}{s_j s_l} \qquad (1)$$

а x_{ij}– значение *i*-го наблюдения *j*-го фактора;

r_{ie}– выборочный парный коэффициент корреляции, характеризует тесноту линейной связи между показателями x_j и x_e.

При этом r_{je} является оценкой генерального парного коэффициента корреляции ρ.

Частный коэффициент корреляции $\rho_{12/3,\ldots,p}$характеризует тесноту линейной зависимости между двумя переменными (x_1и x_2) при исключении влияния всех остальных переменных, входящих в модель.

Частный коэффициент корреляции обладает всеми свойствами парного коэффициента корреляции. Если парный коэффициент корреляции между двумя случайными величинами оказался больше соответствующего частного коэффициента, то можно сделать вывод о том, что фиксирование всех других переменных приводит к усилению взаимосвязи между изучаемыми величинами, т.е. более высокое значение парного коэффициента обусловлено присутствием «третьей величины». Более низкое значение парного коэффициента корреляции в сравнении с соответствующими частными свидетельствует об ослаблении связи между изучаемыми величинами действием фиксируемых величин.

В случае *p*-мерного нормального закона распределения вектора $X = (x_1, x_2, \ldots, x_p)^T$ частный коэффициент корреляции, например $\rho_{12/3,4,\ldots,p}$ порядка $l = p - 2$, определяется по формуле

$$\rho_{12/3,4,\ldots,p} = -\frac{R_{12}}{\sqrt{R_{11} \cdot R_{22}}}, \qquad (2)$$

где R_{jm} алгебраическое дополнение элемента ρ_{jm} корреляционной матрицы R , лежащего на пересечении j -й строки и m-го столбца (j,m=1,2,..., k) .

Значимость парных и частных коэффициентов корреляции, т.е. гипотеза H_0: ρ=0, проверяется по t-критерию Стьюдента. Наблюдаемое значение критерия находится по формуле: $t_{\text{набл}} = \frac{r}{\sqrt{1-r^2}}\sqrt{n-l-2}$, (3)

где r – соответственно оценка парного или частного коэффициента корреляции; l – порядок частного коэффициент корреляции, т.е. число фиксируемых факторов. Для парного коэффициента корреляции l=0.

Коэффициент корреляции считается значимым, т.е. гипотеза H_0: ρ=0 отвергается с вероятностью ошибки α, если tнабл по модулю будет больше, чем tкр, определяемое по таблицам t-распределения (распределение Стьюдента) для заданного уровня значимости α и числа степеней свободы ν = $n - l - 2$. Если же $|t_{набл}| < t_{кр}$, то гипотеза H_0 не отвергается, т.е. гипотеза об отсутствии зависимости между признаками не противоречит наблюдениям.

Значимость парных и частных коэффициентов корреляции также можно проверить с помощью таблиц Фишера-Иейтса. В этом случае гипотеза H_0 отвергается с вероятностью ошибки α, если полученное значение r коэффициента корреляции по модулю окажется больше табличного значения $r_{кр}$, найденного по табл. Фишера-Иейтса при заданном α и числе степеней свободы ν= n - l - 2, из этого следует, что зависимость между величинами имеет место. В противном случае $|r| < r_{кр}$, при этом гипотеза H_0: ρ=0 не отвергается. Для значимых параметров связи имеет смысл найти интервальные оценки. Фишер доказал, что статистика $Z' = \frac{1}{2}\ln\frac{1+r}{1-r}$ уже при n >10 имеет асимптотически нормальное распределение приемлемой точности с математическим ожиданием $MZ' = Z \approx \frac{1}{2}ln\frac{1+\rho}{1-\rho}$ и дисперсией $DZ' \approx \frac{1}{n-l-3}$, т.е. $Z' \in N\left(Z; \sqrt{\frac{1}{n-l-3}}\right)$.

В этой связи при определении с надежностью γ доверительного интервала для значимого парного или частного коэффициентов корреляции ρ используют Z –преобразование Фишера и предварительно устанавливают интервальную оценку для Z $Z' - t_\gamma \sqrt{\frac{1}{n-l-3}} \le Z \le Z' + t_\gamma \sqrt{\frac{1}{n-l-3}}$, (4)

где t_γ вычисляют по таблице интегральной функции Лапласа из условия $\Phi(t_\gamma)=\gamma$.

Значение Z' определяют по таблице Z – преобразования по найденному значению r. Функция нечетная, т.е. $Z'(-r) = -Z'(r)$. Обратный переход от Z к ρ осуществляют также по таблице Z – преобразования, после использования которой получают интервальную оценку для ρ с надежностью γ : $r_{min} \le \rho \le r_{max}$.

Таким образом, с вероятностью γ гарантируется, что генеральный коэффициент корреляции ρ будет находиться в интервале (r_{min}, r_{max}).

2. Оценка множественной взаимосвязи между признаками

Множественный коэффициент корреляции характеризует степень линейной связи между одной переменной (результативной) и остальным массивом $l = p - 1$ статистических данных, включенных в модель.

Множественный коэффициент корреляции изменяется в пределах от 0 до 1, то есть $0 \le \rho_{1\cdot(2,3,\ldots,p)} \le 1$ и определяется по формуле

$$\rho_{1\cdot(2,3,\ldots,p)} = \sqrt{1 - \frac{|R|}{R_{11}}}, \qquad (5)$$

где $|R|$ – определитель корреляционной матрицы R, а R_{11} – алгебраическое дополнение первого диагонального элемента матрицы R.

Если $\rho_{1\cdot(2,3,\ldots,p)} = 1$, то между x_1 и остальными признаками $x_2, x_3, \ldots, x_p$ вектора X имеет место линейная функциональная взаимосвязь, а если $\rho_{1\cdot(2,3,\ldots,p)} = 0$, то x_1 линейно не зависит от переменных $x_2, x_3, \ldots, x_p$.

Квадрат коэффициента корреляции называют *коэффициентом детерминации. Коэффициент детерминации*, например $\rho^2_{1\cdot(2,3,\ldots,p)}$, характеризует долю дисперсии (результативной) величины x_1, обусловленной влиянием остальных переменных $x_2, x_3, \ldots, x_p$, входящих в модель.

Когда данные наблюдений можно считать случайными и выбранными из генеральной совокупности, распределенной по многомерному нормальному закону, основная задача корреляционного анализа состоит в оценке корреляционной матрицы генеральной совокупности по выборке и определении на ее основе оценок частных и множественных коэффициентов корреляции и детерминации.

Значимость множественного коэффициента корреляции (или его квадрата – коэффициента детерминации) проверяется с помощью F – критерия. Например, для множественного коэффициента корреляции $r_{1\cdot(2,\ldots,p)}$ проверка значимости сводится к проверке гипотезы, что генеральный множественный коэффициент корреляции равен нулю, то есть: $H_0: \rho_{1\cdot(2,\ldots,p)} = 0$, а наблюдаемое значение статистики находится по формуле:

$$F_{набл} = \frac{\frac{1}{p-1} r^2_{1\cdot(2,\ldots,p)}}{\frac{1}{n-p}\left(1 - r^2_{1\cdot(2,\ldots,p)}\right)} \qquad (6)$$

Множественный коэффициент корреляции считается значимым, т.е. имеет место линейная статистическая зависимость, между x_1 и остальными факторами $x_2,\ldots x_p$, если $F_{набл} > F_{кр}(\alpha,\ p-1,\ n-p)$, где $F_{кр}$ определяется по таблице F – распределения для заданных $\alpha\ \nu_1 = p-1,\ \nu_2 = n-p$.

Вопросы для самопроверки

1. С какой целью рассчитываются частные коэффициенты корреляции? В чем отличие парных и частных коэффициентов корреляции?
2. Как проверить значимость парных и частных коэффициентов корреляции? Как проверить значимость множественного коэффициента корреляции?

3. Какая статистика используется для построения интервальных оценок коэффициентов корреляции?

4. Как называется квадрат множественного коэффициента корреляции? Как он интерпретируется?

Лекция 8. Методы оценки зависимостей многомерных совокупностей

1. Многомерный линейный регрессионный анализ

На основе результатов корреляционного анализа может быть проведен *регрессионный анализ* многомерной генеральной совокупности. Регрессионный анализ – это статистический метод исследования зависимости случайной величины Y от переменных X_j(j = *1, 2, ..., k*), рассматриваемых в регрессионном анализе как неслучайные величины независимо от истинного закона распределения X_j.

Пусть из (k +1) - мерной генеральной совокупности $(y, x_1, x_2, ..., x_k)$, взята случайная выборка объемом n и пусть i-е наблюдение имеет вид $(y_i, x_{i1}, x_{i2}, ..., x_{ik})$, где i = 1,2,..., n . Тогда классическая линейная модель множественной регрессии (КЛММР) имеет вид:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i, \quad i = 1,2,..., n, \tag{1}$$

где $\beta_0, \beta_1, ..., \beta_k$ – неизвестные параметры модели, которые подлежат оцениванию по выборке, есть неслучайные величины, как параметры генеральной совокупности.

Объясняющие переменные и регрессионные остатки модели удовлетворяют требованиям:

а) объясняющие переменные $x_1, x_2, ..., x_k$ рассматриваются как неслучайные величины, т.е. предполагается, что они измерены без ошибок;

б) величины $x_1, x_2, ..., x_k$ не связаны между собой линейной функциональной зависимостью;

в) регрессионные остатки ε_i есть взаимно независимые случайные величины с нулевым математическим ожиданием $M\varepsilon_i = 0$ и дисперсией

равной $D\varepsilon_i = \sigma^2$ для всех i = 1,2,..., n. Отсюда следует, что коэффициент ковариации:

$$\operatorname{cov}(\varepsilon_i, \varepsilon_l) = \mathrm{M}[(\varepsilon_i - \mathrm{M}\varepsilon_i)(\varepsilon_l - \mathrm{M}\varepsilon_l)] = \mathrm{M}\varepsilon_i\varepsilon_l = \begin{cases} \sigma^2 & при\ i = l \\ 0 & при\ i \neq l' \end{cases}$$

где i,l = 1,2,..., n .(2)

Если регрессионная модель удовлетворяет предпосылкам а, б, в, то МНК-оценки коэффициентов модели, в соответствии с теоремой Гаусса-Маркова, имеют наименьшую дисперсию в классе всех линейных несмещенных оценок.

г) при анализе свойств оценки уравнения регрессии обычно исходят из того, что вектор $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^T$ регрессионных остатков подчиняется n мерному нормальному закону распределения с вектором математических ожиданий $\mathrm{M}\varepsilon = 0$ и ковариационной матрицей $\Sigma_{(\varepsilon)} = \sigma^2 E_n$,т.е. $\varepsilon \in N_n(0; \sigma^2 E_n)$, где E_n– единичная матрица размерности n * n.

Найдем математическое ожидание y_i при заданном векторе значений объясняющих переменных $X_i = (x_{i1}, x_{i2}, ..., x_{ik})^T$.

Получим: $$\tilde{y}_i = \mathrm{M}(y_i / X_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} \qquad (3)$$

Мы получили *уравнение регрессии, характеризующее функциональную зависимость среднего значения y от объясняющих переменных* $x_1, x_2, ..., x_k$.

В этом уравнении β_0 называют свободным членом уравнения. Обычно он содержательно не интерпретируется, т.к. в экономике случай, когда все объясняющие переменные $x_1, x_2, ..., x_k$ равны нулю не имеет содержательного смысла. Например, о каком производстве в регрессионной модели производительности труда может идти речь, если равны нулю производственные площади, число работающих и т.д.

Параметры модели $\beta_1, \beta_2, ..., \beta_k$ называются коэффициентами регрессии. Обычно предполагается, что случайная величина Y имеет нормальный закон

распределения с условным математическим ожиданием $\tilde{Y} = \varphi(x_1,\ldots,x_k)$, являющимся функцией от аргументов x_j, и с постоянной, не зависящей от аргументов дисперсией σ^2.

Коэффициент регрессии β_j показывает, на какую величину в среднем изменится результативный признак Y, если переменную X_j увеличить на единицу его измерения при неизменных значениях других переменных, входящих в модель. Это можно проверить следующим образом: $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k (x_{ik} + 1) = \tilde{y}_i + \beta_k$

В матричной форме классическая линейная модель множественной регрессии(КЛММР) имеет вид: $Y = X\beta + \varepsilon$, (4)

где $Y = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix}$ – вектор-столбец размерности *n* значений результативного показателя,

$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix}$ – матрица объясняющих переменных размерности n*(*k*+1).

Единицы в первом столбце матрицы X призваны обеспечить наличие свободного члена в модели. Здесь можно предположить, что существует переменная x_0, которая во всех наблюдениях принимает значения, равные 1.

$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_k \end{pmatrix}$ – вектор-столбец размерности (*к*+1) неизвестных параметров, которые подлежат оцениванию по выборке;

$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{pmatrix}$ – вектор-столбец размерности n случайных «ошибок», регрессионных остатков.

Причем, $$M\varepsilon = 0, \quad (5)$$

а ковариационная матрица

$$\sum(\varepsilon) = M\varepsilon\varepsilon^T = \left[\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{pmatrix}(\varepsilon_1\varepsilon_2 \ldots \varepsilon_n)\right] = M\begin{pmatrix} \varepsilon_1^2 & \varepsilon_1\varepsilon_2 & \ldots & \varepsilon_1\varepsilon_n \\ \varepsilon_2\varepsilon_1 & \varepsilon_2^2 & \ldots & \varepsilon_2\varepsilon_n \\ . & . & \ldots & . \\ . & . & \ldots & . \\ . & . & \ldots & . \\ \varepsilon_n\varepsilon_1 & \varepsilon_n\varepsilon_2 & \ldots & \varepsilon_n^2 \end{pmatrix}.$$

Для i=1,2,...,n $M\varepsilon_i^2 = \sigma^2$ и $M\varepsilon_{i_1}\varepsilon_{i_2} = 0$ при $i_1 \neq i_2$, тогда

$$\sum(\varepsilon) = M(\varepsilon\varepsilon^T) = \sigma^2 \mathrm{E}_n, \quad (6)$$

где $\mathrm{E}_n = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix}$ – единичная матрица размерности ($n \times n$).

Таким образом Y – случайный вектор – столбец размерности (n × 1) наблюдаемых значений результативного признака $(y_1, y_2, \ldots, y_n)$;

X – матрица размерности [n * (k+1)] наблюдаемых значений аргументов элементы которой x_{ij} рассматриваются как неслучайные величины (i =1, 2, ... , n; j=0, 1, 2, ... k).

На практике рекомендуется, чтобы n превышало k не менее, чем в 4-5 раза.

Так как, в регрессионном анализе x_j рассматриваются как неслучайные величины, а $M\varepsilon_i = 0$, то уравнение регрессии имеет вид:

$$\widetilde{Y} = X\beta, \quad (7)$$

где $\tilde{Y}$ – вектор-столбец с элементами $\tilde{y}_1,\ldots,\tilde{y}_i,\ldots,\tilde{y}_n$ и для всех i= 1,2,...,n $\tilde{y}_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_j x_{ij} + \ldots + \beta_k x_{ik}$.

Основная задача регрессионного анализа заключается в нахождении по выборке объемом n оценок неизвестных коэффициентов регрессии $\beta_0, \beta_1, \ldots, \beta_k$.

Для оценки вектора β наиболее часто используют метод наименьших квадратов (МНК), согласно которому в качестве оценки принимают вектор b, который минимизирует сумму квадратов отклонения наблюдаемых значений y_i от модельных значений $\tilde{y}_i$, т.е. квадратичную форму:

$$Q=(Y - X\beta)^T (Y - X\beta) = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 \Rightarrow \min_{\beta_0,\beta_1,\ldots,\beta_k} \quad (8)$$

Оценка определяется из условия минимизации скалярной суммы квадратов Q по компонентам вектора β, где $Q = \sum_{j=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} x_{ij}\beta_j \right)^2 = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$.

Условием обращения Q в минимум является система уравнений $\frac{\partial Q}{\partial \beta_j} = 0$, где j = 0, 1, 2, ... , k.

Наблюдаемые и модельные значения показаны на рисунке 8.1.

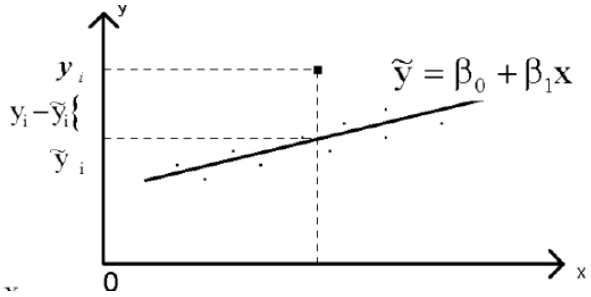

Рис 8.1.Наблюдаемые и модельные значения результативной величины y

Дифференцируя квадратичную форму Q по $\beta_0,\beta_1,\ldots,\beta_k$ и приравнивая производные нулю, получим систему нормальных уравнений:

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_i} = 0 \\ \text{для всех } j = 0, 1, 2, ..., k \end{cases}$$ решая которую и получаем вектор оценок b, где $b = (b_0\ b_1 ... b_k)^T$.

Согласно методу наименьших квадратов, вектор оценок коэффициентов регрессии определяется по формуле: $$b = (X^T X)^{-1} X^T Y \quad (9)$$

Зная вектор оценок коэффициентов регрессии *b*, найдем оценку $\hat{y}_i$ уравнения регрессии: $$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik}. \quad (10)$$

В матричном виде: $\hat{y} = Xb$, где $\hat{y} = (\hat{y}_1 \quad \hat{y}_2 ... \hat{y}_n)^T$.

Оценка ковариационной матрицы коэффициентов регрессии вектора b определяется из выражения: $$S(b) = \hat{S}^2 (X^T X)^{-1}, \quad (11)$$

где $$\hat{S}^2 = \frac{1}{n - k - 1} (Y - Xb)^T (Y - Xb). \quad (12)$$

На главной диагонали ковариационной матрицы находятся дисперсии коэффициентов регрессии: $$\hat{S}^2_{b_{(j-1)}} = \hat{S}^2 \left[(X^T X)^{-1} \right]_{jj} \text{ для } j=1, 2, ..., k, k+1. \quad (13)$$

Значимость уравнения регрессии, т.е. гипотеза H_0: $\beta=0$ $(\beta_0=\beta_1=...=\beta_k=0)$, проверяется по F-критерию, наблюдаемое значение которого определяется по формуле: $$F_{набл} = \frac{Q_R / (k+1)}{Q_{ост} / (n - k - 1)}, \quad (14)$$

где $$Q_R = (Xb)^T (Xb), \quad Q_{ост} = (Y - Xb)^T (Y - Xb) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

По таблице F-распределения Фишера-Снедекора для заданных α, $\nu_1=k+1, \nu_2=n-k-1$ находят Fкр. Гипотеза H_0 отклоняется с вероятностью α, если Fнабл>Fкр. Из этого следует, что уравнение является значимым, т.е. хотя бы один из коэффициентов регрессии отличен от нуля.

Для проверки значимости отдельных коэффициентов регрессии, т.е. гипотез $H_0 : \beta_j = 0$, где j = 1,2,...k, используют t-критерий и вычисляют: $$t_{набл}(b_j) = b_j / \hat{S}_{b_j}. \quad (15)$$

По таблице t-распределения Стьюдента для заданного α и ν= n – k – 1, находят $t_{кр}$. Гипотеза H_0 отвергается с вероятностью α, если |tнабл| >tкр.

Из этого следует, что соответствующий коэффициент регрессии β_j значим, т.е. $\beta_j \neq 0$. В противном случае коэффициент регрессии незначим и соответствующую переменную в модель не следует включать. Если часть коэффициентов уравнения регрессии значима, а часть нет, то реализуется алгоритм пошагового регрессионного анализа, состоящий в том, что исключается одна из незначимых переменных, которой соответствует минимальное по абсолютной величине значение $t_{набл}$. После этого вновь проводят регрессионный анализ с числом факторов, уменьшенным на единицу. Работа алгоритма завершается при получении уравнения регрессии со всеми значимыми коэффициентами.

Вопросы для самопроверки

1. Каким требованиям должны удовлетворять объясняющие переменные и регрессионные остатки?

2. Как содержательно интерпретируются коэффициенты регрессии?

3. Каковы свойства МНК-оценок в КЛММР?

4. В чем смысл проверки значимости уравнения регрессии? Какие задачи в регрессионном анализе решаются с помощью t- критерия Стьюдента?

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

Основная литература

1. Тарасов И. Е. Статистический анализ данных в информационных системах [Электронный ресурс]:учебно-методическое пособие. - Москва: РТУ МИРЭА, 2020. - 96 с. – Режим доступа: https://e.lanbook.com/book/163854
2. Каган Е. С. Прикладной статистический анализ данных [Электронный ресурс]:учебное пособие. - Кемерово: КемГУ, 2018. - 235 с. – Режим доступа: https://e.lanbook.com/book/134318
3. Мхитарян В. С., Архипова М. Ю., Дуброва Т. А., Миронкина Ю. Н., Сиротин В. П. Анализ данных [Электронный ресурс]:Учебник для вузов. - Москва: Юрайт, 2022. - 490 с – Режим доступа: URL: https://urait.ru/bcode/489100

Дополнительная литература

1. Халафян А. А. Statistica 6. Статистический анализ данных:Учеб. пособие для вузов. - М.: Бином, 2011. - 522 с.
2. Наследов А. SPSS 19: профессиональный статистический анализ данных:. - СПб.: Питер, 2011. - 399 с.
3. Берк К., Кэйри П. Анализ данных с помощью MicrosoftExcel:Адаптировано для OfficeXP. - М.: Изд. дом "Вильямс", 2005. - 555 с.
4. Тюрин Ю. Н., Макаров А. А. Статистический анализ данных на компьютере:. - М.: ИНФРА-М, 1998. - 528 с.
5. Миркин Б. Г. Введение в анализ данных:учебники практикум. - М.: Юрайт, 174 с. – Режим доступа: URL: https://urait.ru/bcode/469306
6. Боровиков В. П. Популярное введение в современный анализ данных в системе STATISTICA. Методология и технология современного анализа данных [Электронный ресурс]:. - Москва: Горячая линия-Телеком, 2018. - 288
7. Терехина А. Ю. Анализ данных методами многомерного шкалирования:. - М.: Наука, 1986. - 168 с.
8. Козлов А. Ю., Мхитарян В. С., Шишов В. Ф., Мхитарян В. С. Статистический анализ данных в MSEXCEL:Рек. УМО в кач. учеб. пособия для вузов. - М.: ИНФРА-М, 2014. - 320 с.

РЕКОМЕНДУЕМЫЙ ПЕРЕЧЕНЬ СОВРЕМЕННЫХ ПРОФЕССИОНАЛЬНЫХ БАЗ ДАННЫХ И ИНФОРМАЦИОННЫХ СПРАВОЧНЫХ СИСТЕМ

- Сайт Федеральной службы государственной статистики - https://rosstat.gov.ru/
- Аналитический центр при правительстве Российской Федерации -https://ac.gov.ru/
- Информационный портал Российского научного фонда - http://www.rscf.ru
- Научная электронная библиотека - http://www.elibrary.ru
- Министерство науки и высшего образования Российской Федерации -
- https://www.minobrnauki.gov.ru
  База данных Web of Science - http://www.webofknowledge.com
- Федеральное государственное бюджетное учреждение «Федеральный институт
- промышленной собственности» - http://www.fips.ru/
- Статистические сборники НИУ ВШЭ - https://www.hse.ru/primarydata/
- UNCTAD - https://unctad.org/
- Евростат - https://ec.europa.eu/eurostat
- World Trade Organization. International Trade Statistics - https://www.wto.org/

ДИСЦИПЛИНА Многомерные статистические методы (ч.2)

ИНСТИТУТ Технологий управления

КАФЕДРА Статистики и математических методов в управлении

ВИД УЧЕБНОГО МАТЕРИАЛА Лекции

ПРЕПОДАВАТЕЛЬ Есенин М.А.

СЕМЕСТР 6

План лекций МСМ 2 сем:

Раздел 1. Статистический анализ взаимосвязи (продолжение)

Лекция 1. Многомерный линейный регрессионный анализ: мультиколлинеарность и разновидности пошаговых процедур регрессионного анализа

1.Мультиколлинеарность и разновидности пошаговых процедур регрессионного анализа

Лекция 2. Методы оценки множественной взаимосвязи качественных признаков (часть 1)

1.Алгоритм выбора статистического критерия для анализа категориальных данных.

2.Анализ номинальных и дихотомических признаков

Лекция 3. Методы оценки множественной взаимосвязи качественных признаков (часть 2)

1.Критерии, оценивающие силу связи между номинальными переменными.

2.Анализ порядковых данных

3.Анализ парных выборок

Раздел 2. Методы редукции данных и многомерная классификация без обучения

Лекция 4. Методы снижения размерности - факторный анализ

1.Основные задачи методов снижения размерности

2.Факторный анализ

Лекция 5. Методы снижения размерности - компонентный анализ

1.Компонентный анализ

2.Построение модели регрессии по главным компонентам

Лекция 6. Кластерный анализ и его разновидности (часть 1)

1.Основные понятия кластерного анализа

2.Иерархические кластер-процедуры

Лекция 7. Кластерный анализ и его разновидности (часть 2)

1.Итерационные алгоритмы классификации, метод *k*-средних

2.Двухэтапный кластерный анализ

Раздел 3. Многомерная классификация с обучением и модели бинарного выбора

Лекция 8. Деревья решений, основы дискриминантного анализа и модели бинарного выбора

1.Основные понятия и критерии применения деревьев решений

2.Основы дискриминантного анализа

3.Модели бинарного выбора

Раздел 1. Статистический анализ взаимосвязи

Лекция 1. Многомерный линейный регрессионный анализ: мультиколлинеарность и разновидности пошаговых процедур регрессионного анализа.

1. Мультиколлинеарность и разновидности пошаговых процедур регрессионного анализа

При интерпретации уравнения линейной регрессии для каждого фактора может быть рассчитан средний линейный коэффициент эластичности (Эj). Особенно он актуален для многофакторных моделей, т.к. позволяет оценить вклад каждого регрессора в результативный признак у в %. Следует учитывать сложность оценивания вклада регрессора x_j в у на основе интерпретации коэффициентов модели из-за различия единиц измерения признаков и различия в средних.

Средний линейный коэффициент эластичности $Э_j = b_j \cdot \frac{\bar{x}_j}{\bar{y}}$ (1) показывает на сколько процентов в среднем изменится у при увеличении регрессора x_j на 1% (при фиксированном влиянии остальных факторов).

В статистических пакетах проверка значимости уравнения регрессии сводится к проверке гипотезы $H_0: \beta_1 = \ldots = \beta_k = 0$ и расчетное значение F-критерия сравнивают с табличным Fкр, которое определяют для заданных α, $\nu_1 = k$, $\nu_2 = n-k-1$. В этом случае реализация F-теста идентична проверке значимости множественного коэффициента детерминации, рассматриваемой регрессионной модели.

Модификацией t-теста Стьюдента на значимость коэффициента модели является t-тест, который проверяет гипотезу $H_0 : \beta_j = \beta^*$ о равенстве генерального коэффициента модели β_j при определенном регрессоре некоторому значению β* (например, если мы хотим округлить выборочный

коэффициент до целого значения). Альтернативной является гипотеза $H_1 : \beta_j \neq \beta^*$ для двухсторонней критической области.

Наблюдаемая статистика вычисляется по формуле $t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - \beta^*}{S_{\hat{\beta}_j}}$, (2)

где $\hat{\beta}_j = b_j$.

Критическое значение $t_{кр}$ берут также как в обычном варианте теста по таблице t-распределения Стьюдента для заданного α и ν= n – k – 1. Гипотеза H_0 не отвергается с вероятностью α, если |tнабл| < tкр. В этом случае мы можем сказать, что округление коэффициента перед j-м регрессором будет статистически корректно.

Наряду с точечными оценками b_j генеральных коэффициентов регрессии β_j, регрессионный анализ позволяет получать и интервальные оценки последних с доверительной вероятностью γ. Интервальная оценка с доверительной вероятностью γ для параметра β_j имеет вид:

$$b_j - t_\alpha \hat{S}_{b_j} \leq \beta_j \leq b_j + t_\alpha \hat{S}_{b_j}, \quad (3)$$

где t_α находят по таблице t-распределения при вероятности α=1-γ и числе степеней свободы ν=n-k-1.

Интервальная оценка для уравнения регрессий $\tilde{y}$ в точке, определяемой вектором начальных условий $X^0=(1, x_1^0, x_2^0,...,x_k^0)^T$, равна:

$$\tilde{y} \in \left[(X^0)^T b \pm t_\alpha \hat{S} \sqrt{(X^0)^T (X^T X)^{-1} X^0} \right], \quad (4)$$

где t_α определяется по таблице t-распределения при α=1-γ и числе степеней свободы ν=n-k-1.

По мере удаления вектора начальных условий x_0 от вектора средних $\bar{x} = (1, \bar{x}_1, \ldots, \bar{x}_k)$ ширина доверительного интервала при заданном γ будет увеличиваться (рис. 1).

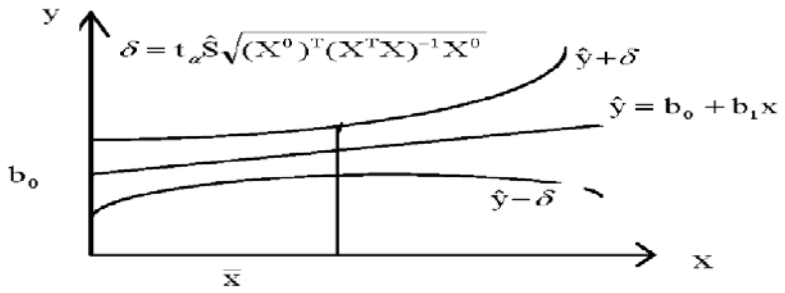

Рис 1.Точечная $\hat{y}$ и интервальная оценки $\left[\hat{y}-\delta < \tilde{y} < \hat{y}+\delta\right]$ уравнения регрессии $\tilde{y} = \beta_0 + \beta_1 x$

Существуют и другие алгоритмы пошагового регрессионного анализа, например, с последовательным включением факторов. Одним из основных препятствий эффективного применения множественного регрессионного анализа, является *мультиколлинеарность*. Она связана с линейной зависимостью между аргументами x_1, x_2, ..., x_k. В результате мультиколлинеарности, матрица парных коэффициентов корреляции и матрица (X^TX) становятся слабо обусловленными, то есть их определители близки к нулю. Это вызывает неустойчивость оценок коэффициентов регрессии, большие дисперсии $\hat{S}^2_{b_j}$ оценок этих коэффициентов, т. к. в их выражения входит обратная матрица $(X^TX)^{-1}$, получение которой связано с делением на близкий к нулю определитель $|X^TX|$ плохо обусловленной матрицы. Отсюда следуют содержательно не интерпретируемые знаки коэффициентов регрессии, заниженные значения t(bj). Кроме того, мультиколлинеарность приводит к завышению значения множественного коэффициента корреляции.

На практике о наличии мультиколлинеарности обычно судят по матрице парных коэффициентов корреляции и рассчитанным на ее основе множественным коэффициентам корреляции одного регрессора с другими регрессорами. Если хотя бы один из таких коэффициентов близок к единице, то считают, что имеет место мультиколлинеарность, и в уравнение регрессии следует включать только один из тесно связанных между собой показателей.

Кроме того, мультиколлинеарность можно выявить на основе критерия *VIF(Variance Inflation Factor*). Для этого нужно оценить регрессию любого из регрессоров на остальные регрессоры, где $x' = \left\{x^{(1)}, x^{(2)}, \ldots, x^{(j-1)}, x^{(j+1)}, \ldots, x^{(k)}\right\}$, и получить значение ρ^2_j. Если значение $VIF_j = 1/(1- \rho^2_j) > 10$, то возможна мультиколлинеарность. Параметр VIF_j для *j*-регрессора показывает, насколько увеличивается оценка стандартного отклонения для коэффициента при регрессоре по сравнению с ситуацией, если бы мультиколлинеарности не было.

С целью борьбы с мультиколлинеарностью часто осуществляют сбор дополнительных данных/изменение выборки, используют алгоритм пошагового регрессионного анализа с включением, применяют метод главных компонент (Principal Component Analysis, PCA) и строят уравнение регрессии на главных компонентах, используют «ридж-регрессию» (Ridge Regression).

Вопросы для самопроверки

1.Что такое мультиколлинеарность?

2. Как выявляют мультиколлинеарность?

3. Какие методы борьбы с мультиколлинеарностью используют на практике?

4. Что показывает коэффициент эластичности?

5. Для чего находят интервальные оценки параметров?

Лекция 2. Методы оценки множественной взаимосвязи качественных признаков (часть 1)

1. Алгоритм выбора статистического критерия для анализа категориальных данных

Перед тем как решить вопрос о выборе того или иного статистического критерия, следует ответить на вопрос о типе собранных данных. В принципе ответ на этот вопрос должен быть дан еще на стадии планирования исследования.

После решения вопроса о том, как представить данные, следует приступать к выбору статистического критерия для проверки той или иной гипотезы. Для этого необходимо знать: 1) данные какого типа будут использоваться для сравнений; 2) сколько групп планируется сравнивать; 3) являются ли сравниваемые группы независимыми (несвязанными).

Прежде чем выбрать статистический критерий для сравнения качественных, номинальных переменных (долей), необходимо определиться с количеством сравниваемых групп. При сравнении частоты встречаемости признака в одной выборке с заданным значением можно воспользоваться критерием Z, о котором подробно написано в большинстве пособий по статистике. Если предполагается сравнение двух независимых групп номинальных данных (например, сравнение удельного веса курящих среди мужчин и женщин), то можно воспользоваться критерием хи-квадрат Пирсона. Однако следует помнить, что в ситуациях, когда сравниваемых групп две и они малы, лучше пользоваться критерием хи-квадрат с поправкой Йейтса на непрерывность (Yates' continuity correction), если количество ожидаемых наблюдений хотя бы в одной из ячеек менее 10. Если ожидаемое число наблюдений в любой из ячеек четырехпольной таблицы 2×2 окажется меньше 5, то следует применять точный критерий Фишера (Fisher's exact test). Для сравнения долей в двух зависимых выборках используется критерий Мак-Нимара (MacNemar test). Если нужно сравнить

качественные данные в трех и более независимых группах, то также можно воспользоваться критерием $\chi 2$ Пирсона. При этом, если в многопольной таблице, например, 2×3, 3×3 и т.д., доля ячеек с ожидаемым числом меньше 5 окажется более 20 % или хотя бы одно ожидаемое значение будет менее единицы, то для расчета критерия $\chi 2$ рекомендуется объединить группы, чтобы увеличить количество наблюдений в клетках таблицы.

При сравнении качественных данных в трех и более зависимых группах используется Q-критерий Кокрена (Cochran's Q-test). Критерий Кокрена является глобальным критерием, то есть он проверяет, есть ли различия между тремя или более группами, но не сообщает, где эти различия. Поэтому при обнаружении статистически значимых различий с помощью критерия Кокрена можно провести попарные сравнения с помощью критерия Мак-Нимара с поправкой Бонферрони. Алгоритм выбора адекватного статистического критерия для номинальных данных представлен на рисунке 1.

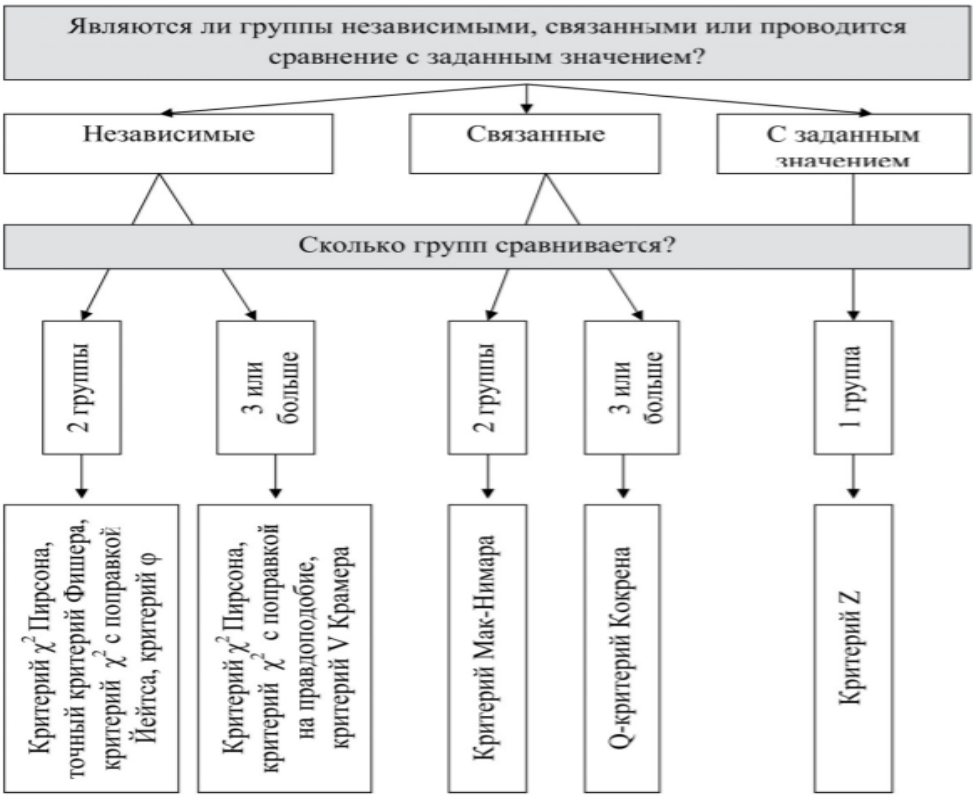

Рис. 1. Алгоритм выбора статистического критерия для анализа категориальных данных

Что касается порядковых (ранговых) признаков, то их можно анализировать как количественные переменные, которые не подчиняются закону нормального распределения, особенно если имеется много возможных категорий, как, например, при анализе оценки состояния здоровья новорожденных по шкале Апгар. Если рангов мало (как, например, в переменной образование), можно использовать критерии, предназначенные для сравнения номинальных переменных. Однако учитывая, что порядковые данные несут в себе больше информации, чем номинальные (известна направленность), применение критериев, предназначенных для сравнения номинальных данных, может привести к потере (точнее к неиспользованию) части информации.

Среди критериев, предназначенных для анализа порядковых данных, только критерий $\chi 2$ для линейного тренда используется для проверки гипотезы о наличии статистической взаимосвязи между переменными, а критерии γ-критерий Гудмана–Краскела, критерии τ-b и τ-c Кендалла, d-критерий Сомера применяются для оценки величины эффекта.

2.Анализ номинальных и дихотомических признаков Z-критерий

Если имеется всего одна выборка, размер которой составляет как минимум несколько десятков наблюдений, и планируется сравнить частоту встречаемости признака с заданным значением, то можно применить критерий Z, о котором подробно написано в большинстве пособий по статистике.

Расчет критерия несложен и может быть произведен «вручную» с помощью формулы:

$$Z = \frac{P - \pi}{\sqrt{\pi \cdot (1 - \pi)}} \cdot \sqrt{n} \quad (1)$$

где Z – абсолютное значение критерия,

P – выборочное значение доли, полученное в результате исследования,

π – заданная доля, с которой планируется провести сравнение,

n – объем выборки, на основании которой было получено значение P.

После подстановки всех значений в формулу необходимо сравнить Z с критическим значением, которое для традиционного уровня значимости 0,05 составляет 1,96, если $P > \pi$, и –1,96, если $P < \pi$ (используется аппроксимация нормального распределения). Если $Z > 1,96$ или $Z < –1,96$, то нулевую гипотезу об отсутствии различий между полученной в результате исследования частотой и заданным значением можно отвергнуть. Более точное значение уровня значимости для полученных значений Z можно найти в таблицах в любом учебнике по статистике.

Критерий хи-квадрат Пирсона

Критерий хи-квадрат Пирсона является, пожалуй, самым распространенным статистическим критерием для бивариантного анализа категориальных данных.

Интересно отметить, что исследователи на постсоветском пространстве часто сравнивают частоты и доли с помощью критерия Стьюдента, в то время как за рубежом для таких сравнений чаще всего используется критерий хи-квадрат ($\chi 2$) Пирсона. Причина, вероятно, кроется в простоте применения критерия Стьюдента, слабой информированности исследователей об ограничениях применения данного критерия. Следует отметить, что большинство авторов учебников и пособий указывают, что нормальная аппроксимация биномиального распределения актуальна лишь при наличии больших выборок и при частотах, близких к 0,5, однако исследователями это в большинстве случаев игнорируется. Игнорирование ограничений дает излишне приближенные результаты и может приводить к обнаружению различий там, где их нет, так как оценка ошибки частоты по формуле дает слишком «оптимистичные» результаты для ситуаций, когда частота события меньше 0,25 или больше 0,75.

Несмотря на то, что в некоторых пособиях сообщается, что свободным от подобного рода ограничений, а значит и более универсальным, является способ проверки равенства частот, основанный на угловом преобразовании

Фишера, он используется российскими исследователями редко. Кроме того, сравнение частот с помощью критерия Стьюдента возможно только для четырехпольных таблиц, то есть только в ситуациях, когда для каждой из изучаемых переменных может быть только два возможных значения. В исследованиях нередки ситуации, когда объемы выборок и/или частоты событий очень малы, а также когда качественные переменные могут принимать более двух значений.

Более универсальными способами сравнения частот и долей являются способы, основанные на идее сравнения фактических частот, полученных в результате исследования, с ожидаемыми частотами. К таким способам анализа качественных переменных относится критерий согласия $\chi 2$ Пирсона, который свободен от вышеперечисленных ограничений.

Критерий $\chi 2$ для таблиц сопряженности был предложен Карлом Пирсоном (1857–1936) еще в 1900 году. С помощью данного критерия оценивается значимость различий между фактическим (выявленным в результате исследования) количеством исходов или качественных характеристик выборки, попадающих в каждую категорию, и теоретическим количеством, которое можно ожидать в изучаемых группах при справедливости нулевой гипотезы. Для применения критерия $\chi 2$ Пирсона необходимо соблюдение следующих условий:

1. Номинальные или порядковые данные (возможно создание категорий из непрерывных данных).

2. Независимость наблюдений (отбор участников исследования из генеральной совокупности производится независимо друг от друга).

3. Независимость групп (метод нельзя применять для исследований типа «до – после»).

4. Ожидаемое (не фактическое) число наблюдений в любой из ячеек должно быть не менее 5 (или 10) для четырехпольных таблиц.

5. Доля ячеек с ожидаемым числом наблюдений менее 5 не должна превышать 20 % для многопольных таблиц.

6. Для расчета критерия χ2 используются только абсолютные фактические и ожидаемые числа (проценты и доли для расчетов не используются).

Рассмотрим принцип метода. Предположим, что проводится наблюдение за участниками исследования (A+B+C+D), причем у (A+B) из них имеется изучаемый фактор риска, а у остальных (C+D) этого фактора риска нет. После определенного времени изучаемый исход наблюдали у A человек из тех, у кого имелся фактор риска, и у C человек из тех, у кого изучаемого фактора риска не было.

Результаты исследования можно отобразить в виде четырехпольной таблицы (табл. 1).

Табл. 1

Пример таблицы сопряженности

	Исход есть (1)	Исхода нет (0)	Всего
Фактор риска есть (1)	(A)	(B)	23 (A+B)
Фактора риска нет (0)	(C)	(D)	25 (C+D)
Всего	(A+C)	(B+D)	(A+B+C+D)

Для ответа на вопрос о наличии статистической взаимосвязи между фактором риска и исходом с помощью критерия χ2 следует сначала рассчитать ожидаемое количество наблюдений в каждой из ячеек при условии справедливости нулевой гипотезы об отсутствии взаимосвязи.

Ожидаемое количество наблюдений для каждой ячейки рассчитывается путем перемножения сумм рядов и столбцов с последующим делением полученного произведения на общее число наблюдений. Так, например, для ячейки A ожидаемое число будет равно (A+B)*(A+C)/(A+B+C+D) и тд. Ожидаемые значения - необязательно целые числа, при этом до целого числа округлять ожидаемые значения нельзя.

Затем рассчитывается значение критерия χ2 по формуле:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}, \qquad (2)$$

где i – номер ряда (строки, от 1 до r),

j – номер столбца (от 1 до с O_{ij} – фактическое количество наблюдений в ячейке ij,

E_{ij} – ожидаемое число наблюдений в ячейке ij.

Затем значение критерия $\chi 2$ сравнивается с критическими значениями для $(r - 1) \cdot (c - 1)$ числа степеней свободы по таблицам, которые имеются в большинстве пособий по статистике. Для данного примера число степеней свободы равно $(2 - 1) \cdot (2 - 1)$, то есть 1. Для 1 степени свободы (а значит, для всех четырехпольных таблиц) критическое значение критерия равно 3,841 при уровне значимости 0,05.

Если фактическое значение превышает критическое, то на основании применения критерия $\chi 2$ Пирсона нулевая гипотеза об отсутствии статистической взаимосвязи между изучаемым фактором риска и исходом может быть отвергнута при критическом уровне значимости 5 %.

В целом, чем больше различия между фактическими и ожидаемыми числами в каждой из ячеек таблицы, тем больше будет значение критерия и тем меньше будет значение достигнутого уровня значимости (p). При равенстве ожидаемых и фактических чисел значение критерия будет равно 0, а p = 1. Хотелось бы подчеркнуть, что речь идет только о статистической взаимосвязи, поэтому выводы о наличии либо причинно-следственных, либо «достоверных» связей только на основании статистически значимых результатов были бы некорректны.

Критерий хи-квадрат с поправкой Йейтса

Вычисленное значение критерия $\chi 2$ изменяется скачкообразно, так как основывается на частотах, которые являются целыми числами. В то же время табличные значения для распределения $\chi 2$ составлены для непрерывной шкалы, поэтому в 1934 году английский статистик Фрэнк Йейтс (Frank Yates, 1902–1994) предложил поправку на непрерывность, которая сейчас известна под названием поправки Йейтса (Yates’s correction).

Поправка заключается в вычитании 0,5 из абсолютного значения разности между фактическим и ожидаемым количеством наблюдений в каждой ячейке, что ведет к уменьшению величины критерия:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(|O_{ij} - E_{ij}| - 0{,}5\right)^2}{E_{ij}} \quad (3)$$

Практически во всех российских пособиях отмечается, что применение поправки Йейтса целесообразно. В других оговаривается, что ее применение необходимо при небольших объемах выборки и/или при количестве ожидаемых наблюдений в любой из ячеек <5, либо < 10. Однако не все статистики согласны с необходимостью применять эту поправку, так как было показано, что она может приводить к получению заниженных значений критерия, а значит, увеличивать вероятность ошибки второго типа, то есть вероятность не найти различия там, где они есть.

Уже через несколько лет после опубликования Йейтсом работы о применении поправки на непрерывность целесообразность ее применения была оспорена другим известным английским статистиком Эгоном Пирсоном (Egon Pearson, 1895–1980, сын Карла Пирсона) и другими исследователями.

При наличии больших выборок различия в значениях критерия $\chi 2$, получаемых с использованием поправки Йейтса и без нее, незначительны, однако при малых выборках различия могут быть существенными. Следует помнить, что поправка Йейтса применяется только для четырехпольных таблиц, то есть при анализе двух дихотомических переменных. Если использование поправки приведет к получению противоположного результата от того, который был получен по исходному критерию, то нужно не зацикливаться на значении p=0,05 и не докладывать результаты применения только того критерия, который помогает «найти то, что очень хотелось». Далее будут рассмотрены методы, позволяющие смотреть более широко на результаты сравнения качественных переменных. Но прежде рассмотрим альтернативы критерию $\chi 2$.

Критерий хи-квадрат с поправкой на правдоподобие

Одной из таких альтернатив является расчет отношения правдоподобия (Λχ2), или критерия χ2 с поправкой на правдоподобие.

Расчет основан на методе максимального правдоподобия, при котором оценка неизвестного параметра производится путем максимизации функции правдоподобия. Расчет Λχ2 производится по формуле:

$$\Lambda\chi^2 = 2 \cdot \sum_{i=1}^{r} \sum_{j=1}^{c} O_{ij} \cdot \ln\left(\frac{O_{ij}}{E_{ij}}\right), \quad (4)$$

после чего полученные значения критерия χ2 сравниваются с табличными значениями, как описано выше. При больших выборках значения Λχ2 и χ2 приблизительно равны. При малых выборках значение Λχ2 обычно несколько меньше, а потому считается некоторыми авторами предпочтительнее.

Точный критерий Фишера

Все вышеперечисленные методы дают приблизительную (асимптотическую) оценку вероятности распределения чисел по ячейкам таблицы так, как было получено в результате исследования, если бы была верна нулевая гипотеза об отсутствии взаимосвязи между фактором риска и исходом.

Точную вероятность для всевозможных четырехпольных таблиц с совпадающими маргинальными итогами можно рассчитать с помощью точного критерия Фишера (Fisher's exact test) по формуле:

$$P = \frac{(A+B)!(C+D)!(A+C)!(B+D)!}{A!B!C!D!N!} \quad (5)$$

где ! – факториал.

Этот метод вызывает меньше споров, чем поправка Йейтса, хотя некоторыми исследователями также высказываются сомнения в целесообразности его применения для малых выборок ввиду его консервативности. Большинство статистиков, однако, по-прежнему

придерживаются мнения, что точный критерий Фишера следует применять при количестве ожидаемых наблюдений <5 (иногда <10) в любой из ячеек четырехпольной таблицы. Более того, некоторые исследователи рекомендуют применять этот критерий даже в ситуациях, когда объем выборки равен нескольким сотням. По мере увеличения числа наблюдений значение p, полученное с помощью точного критерия Фишера, будет приближаться к таковому, полученному с помощью критерия $\chi 2$.

Необходимые условия для применения точного критерия Фишера соответствуют условиям для применения критерия $\chi 2$ за исключением пунктов 4 и 5, подразумевается также гипергеометрическое распределение значения в левой верхней ячейке четырехпольной таблицы, чего мы проверить не можем.

Может получится ситуация, при которой два из четырех статистических критериев говорят о том, что нулевую гипотезу можно отвергнуть, а два других – наоборот. Значение уровня значимости (p) во многом зависит от объема выборки, поэтому даже сильную статистическую связь сложно выявить при малом числе наблюдений, в то время как при больших выборках даже слабая и маловажная связь становится статистически значимой. Поэтому ошибочно было бы делать вывод о силе взаимосвязи между переменными только на основании достигнутого уровня значимости, а также сравнивать по значениям p силу взаимосвязи между признаками в совокупностях с разным числом наблюдений.

Поэтому рекомендуется не только представлять достигнутые уровни значимости при проверке статистических гипотез, но и оценивать величину эффекта, то есть силу связи между признаками. Критерии, оценивающие силу связи между номинальными переменными, могут принимать значения от 0 до 1. Они не могут иметь отрицательных значений, так как данные, измеряемые на номинальной шкале, не имеют порядкового отношения, что не позволяет изучать направление зависимости.

Вопросы для самопроверки

1. В чем заключается алгоритм выбора статистического критерия для анализа категориальных данных?

2. Для чего используют z критерий?

3. В чем заключаются условия, необходимые для применения критерия $\chi 2$ Пирсона?

4. В чем состоят отличия критерия хи-квадрат с поправкой Йейтса и с поправкой на правдоподобие?

5. Для чего используют точный критерий Фишера?

Лекция 3. Методы оценки множественной взаимосвязи качественных признаков (часть 2).

1.Критерии, оценивающие силу связи между номинальными переменными.

Критерии φ и V Крамера

Критерий φ (фи, phi) предназначен для оценки силы взаимосвязи только для четырехпольных таблиц. Для многопольных таблиц целесообразнее применять критерий V Крамера (Cramer's V). Значения обоих критериев варьируют от 0 до 1 (за исключением критерия φ для многопольных таблиц, поэтому для них его применение и не рекомендуется).

Оба критерия основаны на критерии χ2 и могут быть рассчитаны вручную по формулам: $\varphi = \sqrt{\frac{\chi^2}{n}}$ и $V = \sqrt{\frac{\chi^2}{n \cdot (r-1) \cdot (c-1)}}$. (1,2)

Для четырехпольных таблиц значения обоих критериев будут совпадать. Интерпретировать полученные значения критериев φ и V Крамера можно согласно рекомендациям Rea & Parker (табл. 1).

Табл. 1

Интерпретация значений критериев φ и V Крамера согласно рекомендациям Rea & Parker

Значение критериев φ или V Крамера	Сила взаимосвязи
<0,1	Несущественная
0,1 – <0,2	Слабая
0,2 – <0,4	Средняя
0,4 – <0,6	Относительно сильная
0,6 – <0,8	Сильная
0,8 –1,0	Очень сильная

Критерий Чупрова

Коэффициент сопряженности представляет собой меру оценки силы взаимосвязи, основанной на критерии χ2. Зарубежные исследователи чаще применяют коэффициент сопряженности Пирсона (C), в то время как в

российских пособиях сообщается, что для малых таблиц (не более 5 х 5) более точную оценку дает критерий Чупрова (К), который в зарубежной литературе фигурирует как Tshuprow's T.

Расчет коэффициентов сопряженности может быть выполнен вручную по формулам: $$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \quad \text{и} \quad K = \sqrt{\frac{\chi^2}{n\sqrt{(r-1)\cdot(c-1)}}} \qquad (3,4)$$

где n – объем выборки,

r – количество рядов (строк),

c – количество столбцов,

$\chi 2$ –значение критерия хи-квадрат.

Коэффициенты сопряженности принимают значения от 0 (нет взаимосвязи) до значений, приближающихся к 1, но не достигающих ее (сильная взаимосвязь). Максимально возможное значение С зависит от размера таблицы, поэтому для симметричных таблиц можно вручную рассчитать нормированное или скорректированное значение С' по формуле Sakoda:

$$C' = \frac{C}{\sqrt{\frac{r-1}{r}}}, \qquad (5)$$

где r – количество рядов (или столбцов, так как формула предназначена только для симметричных таблиц).

В знаменателе рассчитывается максимально возможное значение С.

Критерий λ Гудмана – Краскела

Следующие два критерия отнести к мерам силы взаимосвязи признаков можно лишь условно. Критерий λ Гудмана – Краскела основан на принципе относительного уменьшения ошибки при прогнозировании значений зависимой переменной с помощью независимой переменной. Критерий λ принимает значения от 0 до 1, где 0 означает, что наличие информации о независимой переменной никак не улучшает возможности прогнозирования

значений зависимой переменной. Вручную λ-критерий Гудмана – Краскела рассчитывается по формуле:

$$\lambda = \frac{\sum f_i - f_t}{n - f_d}, \quad (6)$$

где fi – наибольшие числа в ячейках в каждом из классов независимой переменной;

fd – наибольший из маргинальных итогов (сумм) зависимой переменной,

n – объем выборки.

Для нашего примера (см. ЛК_2) зависимой переменной является исход, так как мы хотим прогнозировать его на основании наличия или отсутствия фактора риска. Таким образом, если λ=0, то это означает, что знания о наличии фактора риска не уменьшают ошибки предсказания исхода. В отличие от всех рассматриваемых выше, λ-критерий асимметричен, то есть его значение зависит от того, какая переменная является зависимой, а какая независимой.

Если бы мы хотели прогнозировать наличие фактора риска, зная исход, то значение критерия было бы отлично от нуля, например, равно 0,261, то есть знание исхода снизило бы количество неверно предсказанных значений для фактора риска на 26,1 %.

Данный критерий очень чувствителен к значениям маргинальных итогов для независимой переменной. Учитывая суть критерия, можно записать λ как $\lambda = \frac{Ошибка1 - Ошибка2}{Ошибка1}$, где ошибка 1 – доля неверно предсказанных значений зависимой переменной без учета значений независимой переменной; ошибка 2 – доля неверно предсказанных значений зависимой переменной с учетом значений независимой переменной.

Таким образом, наглядно видно, что λ показывает процент снижения ошибок прогнозирования при наличии информации о независимой переменной.

Коэффициент энтропии

Последним критерием для оценки силы взаимосвязи, является коэффициент неопределенности, в литературе он также встречается как коэффициент энтропии или энтропийный коэффициент Тейла (Theils U-coefficient). Коэффициент неопределенности имеет несколько отличные от λ-критерия Гудмена – Краскела теоретические обоснования, но также относится к коэффициентам, показывающим значение относительного уменьшения ошибки.

Обычно считают, что коэффициент неопределенности предпочтительнее λ-критерия Гудмена – Краскела. Значение коэффициента неопределенности может варьировать от 0 до 1 и интерпретируется так же, как и λ-критерий Гудмена – Краскела. Учитывая название коэффициента, говорят, что его значение отражает степень неточности прогноза. Значение 0 говорит о том, что зависимую переменную невозможно предсказать по значениям независимой переменной, а 1 – о том, что значения первой полностью предсказываются значениями второй. Расчет коэффициента Тейла более сложен, поэтому формула не приводится.

Относительный риск

Мы рассмотрели способы проверки гипотез о наличии статистической связи между номинальными переменными, а также способы оценки силы взаимосвязи между этими переменными. Тем не менее сообщение о том, что была обнаружена статистически значимая связь средней силы между фактором риска и исходом недостаточно информативно. Гораздо продуктивнее было бы говорить о количественной оценке вероятности исхода, связанной с наличием фактора риска. Однако не все исследования позволяют говорить о риске и оценивать вероятность возникновения исхода в зависимости от наличия или отсутствия фактора риска.

Мы можем рассчитать относительный риск (Relative Risk, RR). Поскольку в примере ничего не сообщается о времени наблюдения, но подразумевается, что оно было одинаковым для обеих групп (с наличием

фактора риска и без него), относительный риск будет равен отношению рисков. Отношение рисков отражает, во сколько раз риск исхода при наличии фактора риска выше риска исхода при отсутствии фактора риска, и рассчитывается применимо к таблице 1 (см. ЛК_2) следующим образом:

$$RR = \frac{\frac{A}{A+B}}{\frac{C}{C+D}} = \frac{A \cdot (C+D)}{C \cdot (A+B)} \quad . \qquad (7)$$

Например, пусть RR=2,7, тогда это говорит о том, что фактор риска может увеличивать вероятность возникновения исхода в 2,7 раза или что риск исхода у тех, у кого есть фактор риска, в 2,7 раза выше, чем у тех, у кого фактора риска нет. Такой результат гораздо более информативен. Однако различия в 2,7 раза будут справедливы только для нашей выборочной совокупности. Даже если допустить, что наша выборка репрезентативна, систематические ошибки отсутствуют, а влияние вмешивающихся факторов минимально, относительный риск для генеральной совокупности может отличаться, поэтому всегда рекомендуется представлять интервальную оценку относительного риска с помощью 95 % доверительного интервала. Этот интервал представляет собой область, в которую попадает истинное значение доли в 95 % случаев.

Для относительного риска 95 % доверительный интервал можно рассчитать по формуле:

Верхняя граница: e^x,

где $$x = \ln(RR) + 1{,}96 \cdot \sqrt{\frac{B}{A \cdot (A+B)} + \frac{D}{C \cdot (C+D)}}$$

Нижняя граница: e^x,

где, $$x = \ln(RR) - 1{,}96 \cdot \sqrt{\frac{B}{A \cdot (A+B)} + \frac{D}{C \cdot (C+D)}} \qquad (8,9)$$

а e – основание натурального логарифма (число Эйлера ~ 2,7).

Если доверительный интервал имеет значительную ширину, то это может быть вызвано малым объемом выборки.

Отношение шансов

Если бы наше исследование было типа «случай – контроль», было бы неверным рассчитывать относительный риск. В таких исследованиях в качестве меры эффекта выступает отношение шансов (Odds Ratio, OR). Представим, что наше исследование было исследованием типа «случай – контроль». Тогда,

$$OR = \frac{A \cdot D}{B \cdot C} \quad (10)$$

Пусть OR будет равно 4, тогда шансы на изучаемый исход были в 4 раза выше у тех участников исследования, у кого имелся фактор риска, чем у тех, у кого фактора риска не было.

При проецировании результатов на генеральную совокупность также необходимо рассчитать 95 % доверительный интервал, в который попадут значения от $e^{\ln(OR)-1,96\cdot\sqrt{\frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\frac{1}{D}}}$ до $e^{\ln(OR)+1,96\cdot\sqrt{\frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\frac{1}{D}}}$, где e – основание натурального логарифма.

Следует помнить, что вышеприведенные формулы для расчета доверительных интервалов предназначены только для независимых данных и неприменимы в исследованиях типа «до – после», а также в исследованиях типа «случай – контроль» по методу подобранных пар (Matched case-control study). Не стоит представлять относительный риск и отношение шансов в одном и том же исследовании. Для исследований типа «случай – контроль» описанные выше расчеты относительного риска и разницы рисков провести невозможно.

Предположим, что рассчитанные значения критериев φ, V Крамера, коэффициентов сопряженности и особенно относительного риска позволяют заподозрить, что изучаемый фактор риска может оказать значительное влияние на вероятность возникновения исхода даже при значении $p>0,05$ (точный критерий Фишера или поправка Йейтса).

Если бы мы ограничились проверкой статистической гипотезы при критическом уровне значимости 5 % и дихотомическом подходе к трактовке

результатов, то пришлось бы принять нулевую гипотезу об отсутствии связи между фактором риска и исходом и сделать вывод о безвредности изучаемого фактора. Еще интереснее, если бы исследование было повторено другими на большей выборке. Тогда (при прочих равных условиях) по причине большей статистической мощности достигнутый уровень значимости был бы <0,05, а значит, при аналогичном подходе к величине p вывод был бы противоположным, хоть величина эффекта была бы та же самая.

Рассуждения о том, что проверка статистических гипотез сообщает только часть информации, были опубликованы Пирсоном еще в 1901 году, продолжены Фишером и, наконец, нашли свое выражение в современных рекомендациях где четко говорится о том, что помимо результатов статистических тестов необходимо представлять меры силы взаимосвязи между изучаемыми факторами. Еще более грамотной стратегией является принятие решения еще на этапе планирования исследования, то есть задолго до начала сбора данных, о том, какие значения относительного риска или отношения шансов будут считаться важными, после чего рассчитывается необходимый объем выборки. Уже на этом этапе может стать очевидным, что некоторые исследования проводить нецелесообразно по причине невозможности набрать достаточное количество участников исследования для того, чтобы выявить статистически значимые различия на желаемом уровне.

2.Анализ порядковых данных

Учитывая, что данные, измеряемые на порядковой (ранговой, ординальной) шкале, также относятся к качественным данным, критерии, которые можно применять для анализа многопольных таблиц номинальных данных, применимы и для порядковых. Однако, учитывая, что порядковые данные несут в себе больше информации, чем номинальные (известна направленность), применение критериев, предназначенных для сравнения

номинальных данных, приведет к потере (точнее к неиспользованию) части информации. Все из рассматриваемых далее критериев предназначены для оценки величины эффекта, и только критерий χ2 для линейного тренда – для проверки гипотезы о наличии статистической взаимосвязи между переменными. Кроме того, γ -критерий Гудмана – Краскела, критерии τ-b и τ-c Кендалла являются симметричными мерами взаимосвязи между переменными, то есть при их расчете не имеет значения, какая из переменных является зависимой, а какая независимой. В то же время d-критерий Сомера является асимметричным и может принимать различные значения в зависимости от того, какая из переменных является зависимой.

Перед тем, как перейти к разбору критериев, предназначенных для ранговых переменных, остановимся на основных принципах, лежащих в основе расчетов. Поскольку для порядковых переменных характерна направленность, то в основе всех критериев лежит расчет количества нарушений порядка. Все пары данных можно классифицировать как конкордантные, дискордантные и связанные либо по зависимой, либо по независимой переменной. Конкордантные пары также называют проверсиями, а дискордантные – инверсиями. Помимо конкордантных и дискордантных пар существуют так называемые связанные или сцепленные пары, которые в англоязычной литературе обозначаются как ties или tied ranks. Расчеты количества конкордантных пар наблюдений, дискордантных пар наблюдений, и т.д. пригодятся позже при расчете статистических критериев.

Критерий хи-квадрат для линейного тренда

«Вручную» критерий χ2 для линейного тренда рассчитывается по формуле:

$$M^2 = r_\delta^{\,2} \cdot (N-1), \qquad (11)$$

где M2 – значение критерия χ2 для тренда (для 1 степени свободы),

N – объем выборки,

r_p – значение коэффициента корреляции Пирсона между изучаемыми переменными.

Интересно отметить, что критерий $\chi 2$ для линейного тренда менее чувствителен к ситуациям, когда ожидаемое количество наблюдений в некоторых ячейках мало, то есть при наличии малого числа наблюдений ценность этого критерия может быть выше, нежели классического критерия $\chi 2$ Пирсона.

Гамма-критерий Гудмена – Краскела

Гамма-критерий Гудмена – Краскела (Goodman – Kruskal's gamma) основан на сравнении количества конкордантных и дискордантных пар.

Расчет критерия может производиться вручную по формуле:

$$\gamma = \frac{C - D}{C + D}, \qquad (12)$$

где C – количество конкордантных,

D – количество дискордантных пар.

Критерий гамма может варьировать от -1 до 1, причем 1 означает полную прямо пропорциональную взаимосвязь между переменными, -1 – полную обратную взаимосвязь между переменными, а 0 – полное отсутствие какой-либо связи между изучаемыми признаками. Чем ближе значение критерия к 1 или -1, тем сильнее взаимосвязь. Гамма является симметричным критерием и не зависит от того, какая из переменных является зависимой. Технически гамма показывает насколько больше в исследуемой выборке конкордантных пар, чем дискордантных, относительно общего числа конкордантных и дискордантных пар. При этом полностью игнорируются связанные пары наблюдений (см. формулу 12). Можно также интерпретировать гамма-критерий как пропорциональное уменьшение ошибки прогнозирования одной переменной при наличии информации о другой.

Если полученное значение критерия будет равно 0,54, то можно интерпретировать этот результат следующим образом: наличие информации о факторе риска может уменьшить ошибку предсказания степени тяжести, например, заболевания на 54 %.

При расчете гамма-критерия совершенно не используется информация о связанных парах наблюдений.

Критерий тау-b Кендалла

Из критериев, предназначенных для сравнения порядковых данных с учетом связанных пар наблюдений можно рассчитывать критерии тау-b и тау-c Кендалла (Kendall's tau- b tau-c, соответственно). Оба критерия могут принимать значения в том же диапазоне, что и гамма-критерий Гудмена – Краскела. Оба тау-критерия Краскела также показывают силу взаимосвязи между переменными. Критерий тау-b чаще всего применяется для таблиц 2x2, однако возможно его применение и для многопольных таблиц. Так же как и критерий гамма, он показывает разность между количеством конкордантных и дискордантных пар, но с делением на геометрическое среднее количества пар, связанных по рядам, и количества пар, связанных по столбцам, что можно представить в виде:

$$\tau_b = \frac{C - D}{\sqrt{(C + D + X) \cdot (C + D + Y)}}, \qquad (13)$$

где X – количество пар наблюдений, связанных по рядам,

Y – количество пар наблюдений, связанных по столбцам,

C и D – количество конкордантных и дискордантных пар, соответственно.

Данный критерий лучше использовать только для квадратных таблиц, то есть для таблиц, в которых количество рядов равно количеству столбцов. Критерий тау-b является симметричным критерием, для которого неважно какая из переменных является зависимой.

Интерпретировать значение критерия достаточно сложно, но следует помнить, что он также, как и гамма-критерий может принимать значения от -1 до 1 и показывает силу взаимосвязи между переменными.

Критерий тау-с Кендалла

Критерий тау-с Кендалла, называемый иногда еще критерием тау-с Стюарта (Stuart›s tau-c) или критерием Кендалла – Стюарта, использует коррекцию на общее количество рядов и столбцов в таблице сопряженности и использует общее количество наблюдений, а не только конкордантные и дискордантные пары, как гамма-критерий. Кроме того, в отличие от критерия тау-b, он может применяться не только для четырехпольных или для квадратных таблиц, что обеспечивает его более широкое использование, чем критерия тау-b.

Расчет критерия тау-с Кендалла производится по формуле:

$$\tau_c = \frac{2m(C-D)}{N^2(m-1)}, \qquad (14)$$

где m – меньшее значение количества рядов или столбцов (в нашем примере таблица имеет 2 ряда и 3 столбца, значит m=2),

N – объем выборки, C и D – количество конкордантных и дискордантных пар, соответственно.

Из проблем, связанных с применением этого критерия, следует отметить то, что его значение в значительной степени зависит от размеров таблицы (количества рядов и столбцов), то есть от степени категоризации данных исследователем, что является поводом для критики данного критерия.

Критерий d Сомера

Критерий d Сомера учитывает только связанные пары данных по столбцам, если в столбцах записана зависимая переменная. Пусть, зависимой переменной является тяжесть заболевания, которая была занесена в столбцы,

а независимая переменная - фактор риска – в ряды. Критерий d Сомера рассчитывается по формуле:

$$d = \frac{C - D}{C + D + Y}, \quad (15)$$

где C и D – количество конкордантных и дискордантных пар, соответственно,

Y – количество пар, связанных по зависимой переменной.

Значение d Сомера показывает разность между вероятностью того, что случайно выбранная пара наблюдений конкордантна, и вероятностью того, что эта пара дискордантна при условии, что наблюдения не связаны по независимой переменной.

Возможные значения d Сомера варьируют от -1 до 1 (от полной прямо пропорциональной взаимосвязи до полной обратно пропорциональной взаимосвязи), а 0 обозначает полную независимость переменных друг от друга. В зависимости от количества связанных пар, d Сомера всегда будет несколько меньше, чем значение гамма-критерия Гудмена – Краскела.

3. Анализ парных выборок

Критерий Мак – Нимара

Критерий Мак– Нимара применяется для анализа связанных измерений в случае измерения реакции с помощью дихотомической переменной.

По результатам такого исследования строится результирующая таблица 2×2 в виде (табл. 2):

Табл. 3

Общий вид таблицы 2×2 для критерия Мак – Нимара

До/После	0	1	Всего
1	A	B	A + B
0	C	D	C + D
Всего	A + C	B + D	N

В клетках A и D представлены изменения от «до» к «после», причем в клетке A изменения благоприятных результатов на неблагоприятные, а в клетке D – наоборот.

Нулевая гипотеза состоит в том, что в генеральной совокупности доля тех, кто изменяет благоприятную реакцию на неблагоприятную в результате воздействия, равна доле тех, кто изменяет реакцию в обратном порядке. Объем выборки N определяется как сумма частот в диагональных клетках A и D.

Для проверки гипотезы в случае N > 50 рассчитывается статистика $\chi 2$ по упрощенной формуле (для данного критерия число степеней свободы всегда равно 1:

$$\chi^2 = \frac{(|A - D| - 1)^2}{A + D}, \quad (16)$$

где |A – D| – абсолютное значение разности значений соответствующих клеток, единица вычитается в качестве поправки на непрерывность.

Если рассчитанное значение статистики превосходит соответствующее табличное (рассчитанное исходя из объема выборки N и уровня значимости), то нулевая гипотеза отклоняется. В качестве табличных значений используются критические значения критерия $\chi 2$.

Практически все современные программные пакеты рассчитывают достигнутый уровень значимости, поэтому для сегодняшних исследователей статистические таблицы имеют скорее историческую, чем практическую ценность в повседневной научной практике.

Q-критерий Кокрена

Критерий Кокрена используется при сравнении влияний различных воздействий на одну группу или однородные группы. Исходной для проверки критерия является таблица результатов исследования в следующем виде: по столбцам – значения эффекта, например от соответствующей терапии: 0 - нет эффекта и 1 - есть эффект; по строкам – повторные измерения значения эффекта для каждого индивидуума (всего n объектов). Нулевая гипотеза

состоит в том, что в генеральной совокупности доли всех изучаемых воздействий одинаковы.

Статистика критерия Кокрена выражается формулой:

$$Q = \frac{(m-1)\left[m\sum(\sum X_k)^2 - (\sum X_T)^2\right]}{m\sum(\sum X_R) - \sum(\sum X_R^2)}, \quad (17)$$

где m – число изучаемых воздействий;

$(\sum X_T)^2 = (\sum X_1 + \sum X_2 + \ldots + \sum X_m)^2$, где $\sum X_k = \sum X_{ik}$ – сумма значений по k – му столбцу;

$\sum(\sum X_k)^2 = (\sum X_1)^2 + (\sum X_2)^2 + \ldots + (\sum X_m)^2$ $\sum(\sum X_R)^2$– сумма итоговых значений по строкам;

$\sum(\sum X^2_R)$ – сумма квадратов итоговых значений по строкам.

Полученное значение статистики Q проверяется по таблице $\chi 2$ для выбранного уровня значимости и числа степеней свободы, равного m – 1. Если рассчитанное значение превосходит табличное, нулевая гипотеза отклоняется на выбранном уровне значимости α.

Помимо воздействия различных видов лечения на одну и ту же группу критерий Кокрена может использоваться для повторных наблюдений в одной и той же выборке в разные моменты времени, например, сегодня, затем через неделю, через месяц и через год. Таблица результатов может оформляться так же, как было описано выше. Формула для расчета также может использоваться та же.

Вопросы для самопроверки

1.Какие основные критерии, оценивающие силу связи между номинальными переменными, вы знаете?

2.Что позволяют определить относительный риск и отношение шансов?

3.Как можно построить доверительный интервал для относительного риска и отношения шансов?

4.Какие основные критерии анализа порядковых данных, вы знаете?

5.Какие основные критерии применяются для анализа парных выборок?

Раздел 2. Методы редукции данных и многомерная классификация без обучения

Лекция 4. Методы снижения размерности – факторный анализ

1. Основные задачи методов снижения размерности

Во многих практических задачах исследователя интересуют признаки, которые обнаруживают наибольшую изменчивость (т. е. разброс, дисперсию) при переходе от одного объекта к другому, при этом такие признаки часто невозможно наблюдать непосредственно на объектах.

Приведем несколько примеров: склонность населения к миграции определяется по данным о достаточно большом числе социально-экономических, демографических, географических и др. показателей и результатам социологических опросов; только большая совокупность непосредственно измеряемых признаков позволяет сравнивать страны, регионы и города по уровню жизни, продукцию различных производителей — по качеству.

Приведенные примеры иллюстрируют сущность задач снижения размерности многомерного пространства, которая заключается в выражении большого числа исходных факторов, непосредственно измеренных на объектах, через меньшее (как правило, намного меньшее) число более емких, максимально информативных, но непосредственно не наблюдаемых внутренних характеристик объектов. При этом предполагается, что более емкие признаки будут отражать наиболее существенные свойства объектов.

Целью методов снижения размерности является исследование внутренней структуры изучаемой системы k случайных величин, «сжатие» этой системы без существенной потери содержащейся в ней информации путем выявления небольшого числа факторов, объясняющих изменчивость и взаимосвязи исходных случайных величин.

Метод главных компонент выявляет k компонент−факторов, объясняющих всю дисперсию и корреляции исходных k случайных величин;

при этом компоненты строятся в порядке убывания объясняемой ими доли суммарной дисперсии исходных величин, что позволяет зачастую ограничиться несколькими первыми компонентами.

Факторный анализ выявляет m ($m < k$) общих для всех исходных величин факторов, объясняя оставшуюся после этого дисперсию влиянием специфических факторов.

Среди прикладных задач, решаемых указанными методами: – поиск скрытых, но объективно существующих взаимосвязей между экономическими и социальными показателями, проверка гипотез о взаимосвязях этих показателей, выявление природы различий между объектами, описание изучаемой системы числом признаков, значительно меньшим числа исходных факторов, при этом выявленные факторы или главные компоненты содержат в среднем больше информации, чем непосредственно зафиксированные на объектах значения исходных факторов; построение обобщенных экономических и социальных показателей, таких как качество продукции, размер предприятия, интенсивность ведения хозяйства и т. п., визуализация исходных многомерных наблюдений путем их проецирования на специально подобранную прямую, плоскость или трехмерное пространство; построение регрессионных моделей по главным компонентам, – в социальных и экономических задачах исходные факторы часто обладают мультиколлинеарностью, что затрудняет построение и интерпретацию регрессионных моделей, не позволяя часто получать сколь-нибудь точные прогнозы, а главные компоненты, сохраняя всю информацию об изучаемых объектах, являются не коррелированными по построению; классификация по обобщенным экономическим показателям, практика показывает, что классификация объектов, проведенная по факторам или по главным компонентам, оказывается более объективной, чем классификация тех же объектов по исходным признакам; по одному —трем факторам или главным компонентам возможно проведение визуальной классификации, в случае

большей размерности пространства обобщенных показателей, полученного в результате компонентного или факторного анализа, необходимо привлечение методов многомерной классификации; сжатие исходной информации, значительное уменьшение объемов информации, хранимой в базах данных, без существенных потерь в информативности.

2. Факторный анализ

Факторный анализ со статистической точки зрения связан с поиском новых признаков, характеризующих объекты наблюдения на основе имеющейся информации, которая содержится в измеренных значениях *k* исходных признаков $x = (x_1, x_2, ..., x_k)^T$. Всю информацию об *n* объектах наблюдения можно представить в виде матрицы $\underset{n\times k}{X} = (x_{ij})$ «объект — признак». Для дальнейшего анализа удобнее использовать матрицу наблюдаемых стандартизованных признаков, которые тоже относятся к категории измеримых, как рассчитанных непосредственно по результатам произведенных наблюдений $\underset{[k\times n]}{Z} = \{z_{ji}\}.$.

Обычно неизвестные математические ожидания μ_j и дисперсии σ^2_j заменяются их выборочными аналогами: выборочной средней $\hat{\mu}_j = \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ и несмещенной оценкой дисперсии $\hat{s}_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$.

Средние значения стандартизованных переменных равны нулю (z_j=0),а дисперсии — единице (s^2_j=1). Связь новых переменных с наблюдаемыми признаками в факторном анализе аналогична регрессионной, но с тем существенным отличием, что эти новые объясняющие переменные, или факторы, неизвестны и нуждаются в идентификации.

В моделях факторного анализа используются общие и индивидуальные факторы. Общие факторы g_l связаны значимыми коэффициентами более чем с одной измеримой переменной. Каждый из индивидуальных факторов v_j

связан только с одной j-й измеримой переменной. При этом обычно предполагается, что индивидуальные факторы не коррелированы между собой и с общими факторами. Кроме того, для удобства факторы выбираются как стандартизованные:

$$z_{ji} = \sum_{l=1}^{m} a_{jl} g_{li} + b_j \nu_{ji}. \quad (1)$$

Второй индекс переменных z_{ji}, g_{li} и υ_{ji} обозначает номер объекта наблюдения i=1, 2, ..., n. Первый индекс j =1, 2, ..., k характеризует номер исходного признака z_{ji} и соответствующего ему индивидуального эффекта υ_{ji}, а для g_{li} первый индекс l =1, 2, ..., m обозначает номер общего фактора. Коэффициенты при общих факторах можно свести в матрицу $A = \begin{pmatrix} a_{11} & \dots & a_{1m} \\ \dots & \dots & \dots \\ a_{k1} & \dots & a_{km} \end{pmatrix}$, а коэффициенты при индивидуальных факторах для дальнейшего матричного представления модели будут диагональными элементами в диагональной матрице $B = \begin{pmatrix} b_1 & 0 & \cdots & 0 \\ 0 & b_2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & b_k \end{pmatrix}$.

Включающая нагрузки всех факторов общая матрица коэффициентов, или *матрица факторного отображения*, будет представлять собой результат объединения элементов обеих матриц: $K = (A\ B) = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} & b_1 & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & a_{2m} & 0 & b_2 & 0 & \vdots \\ \dots & \dots & \dots & \dots & \vdots & 0 & \ddots & 0 \\ a_{k1} & a_{k2} & \cdots & a_{km} & 0 & \cdots & 0 & b_k \end{pmatrix}$.

Матрица значений общих факторов представляет собой матрицу размерности m*n, где $m \le k$: $F = \begin{pmatrix} f_{11} & \dots & f_{1n} \\ \dots & \dots & \dots \\ f_{m1} & \dots & f_{mn} \end{pmatrix}$.

Матрица значений индивидуальных факторов имеет размерность k*n: $V = \begin{pmatrix} \nu_{11} & \dots & \nu_{1n} \\ \dots & \dots & \dots \\ \nu_{k1} & \dots & \nu_{kn} \end{pmatrix}$.

Общая матрица значений факторов может быть образована как результат объединения матриц общих и индивидуальных факторов:

$$\tilde{F} = \begin{pmatrix} F \\ V \end{pmatrix} = \begin{pmatrix} f_{11} & \cdots & f_{1n} \\ \cdots & \cdots & \cdots \\ f_{m1} & \cdots & f_{mn} \\ \nu_{11} & \cdots & \nu_{1n} \\ \cdots & \cdots & \cdots \\ \nu_{k1} & \cdots & \nu_{kn} \end{pmatrix}.$$

С учетом введенных обозначений модель факторного анализа в матричной форме может быть представлена в виде: $Z = K\tilde{F} = (A\ B)\begin{pmatrix} F \\ V \end{pmatrix}.$ (2)

Модель факторного анализа с учетом неполного содержания исходной информации об объектах исследования в новой системе координат меньшей размерности ($m < k$) неизбежно будет содержать помимо общности в виде информации об объектах в системе координат общих факторов и специфичность, представляемую в виде значений характерных факторов.

В то же время с учетом случайности выборки и погрешности измерения нормированное наблюдаемое значение z_{ji} содержит истинное значение, индивидуальную особенность Ind_{ji} каждого объекта и ошибку измерения ε_{ji}: $z_{ji} = M(z_{ji}) + Ind_{ji} + \varepsilon_{ji}$. В рамках статистического подхода под истинным значением понимается математическое ожидание признака $M(z_{ji})$, вторая и третья составляющие характеризуют отклонение отдельного показателя на данном объекте от среднего.

Если первая составляющая является общей статистической характеристикой совокупности объектов исследования, то вторая и третья компоненты являются носителями особенностей, присущих данному объекту и методу измерения. В процессе управления важнейшим моментом являются знание и умение учитывать индивидуальные черты отдельных объектов исследования.

Характеристика вариативности — дисперсия — для нормированного значения наблюдаемого признака может быть представлена в следующем виде:

$$\hat{s}_j^2 = 1 = \frac{1}{n}\sum_{i=1}^{n} z_{ji}^2 = \frac{1}{n}\Big[a_{j1}^2 \sum_{i=1}^{n} f_{1i}^2 + a_{j2}^2 \sum_{i=1}^{n} f_{2i}^2 + \ldots + a_{jm}^2 \sum_{i=1}^{n} f_{mi}^2 + b_j^2 \sum_{i=1}^{n} v_{ji}^2 + 2\Big(a_{j1}a_{j2}\sum_{i=1}^{n} f_{1i}f_{2i} + a_{j1}a_{j3}\sum_{i=1}^{n} f_{1i}f_{3i} + \ldots + a_{j(m-1)}a_{jm}\sum_{i=1}^{n} f_{(m-1)i}f_{mi} + a_{jm}b_j\sum_{i=1}^{n} f_{mi}v_{ji}\Big)\Big], \quad (3)$$

где $j = 1, 2, \ldots, k,\ i = 1, 2, \ldots, n.$

Ошибка измерения обычно оказывается значительно меньше вариативной компоненты, поэтому их часто объединяют. Однако поскольку вариативная составляющая и ошибки измерения возникают независимо друг от друга, то их рассматривают как некоррелированные. Рассмотрим слагаемые, содержащие сомножитель $\frac{1}{n}\sum_{i=1}^{n} f_{ri}^2$, величина которого является дисперсией произвольного общего фактора f'_{ri} после нормировки: $f_{ri} = \frac{f'_{ri} - \bar{f}'_r}{s_{f'_r}}$.

Величина дисперсии нормированного общего фактора равна единице: $\frac{1}{n}\sum_{i=1}^{n} f_{ri}^2 = \hat{s}_{f_r}^2 = 1.$

Рассмотрим в формуле слагаемые, содержащие сомножитель $\frac{1}{n}\sum_{i=1}^{n} f_{ri}f_{li}$.

Это коэффициент корреляции между двумя общими факторами, т.е.

$$\frac{1}{n}\sum_{i=1}^{n} f_{ri}f_{li} = r_{f_r f_l}, \quad (4)$$

где r=1, 2, ..., m;

l=1, 2, ..., m;

$r \neq l.$

После введения обозначения для коэффициента корреляции общих и индивидуальных эффектов $\frac{1}{n}\sum_{i=1}^{n} f_{mi} v_{ji} = r_{f_m v_j}$ выражение можно представить в виде $\hat{s}_j^2 = a_{j1}^2 + a_{j2}^2 + \ldots + a_{jm}^2 + b_j^2 + 2(a_{j1}a_{j2}r_{f_1 f_2} + a_{j1}a_{j3}r_{f_1 f_3} + \ldots + a_{jm}b_j r_{f_m v_j})$.

Из этого представления следует, что

$$\hat{s}_j^2 = 1 = \sum_{r=1}^{m} a_{jr}^2 + b_j^2 + 2\sum_{l>r=1}^{m} a_{jr}a_{jl}r_{f_r f_l} + 2b_j\sum_{r=1}^{m} a_{jr}r_{f_r v_j}. \quad (5)$$

Так как характерный фактор присущ только данной *j*-й переменной и не коррелирован с общими факторами, то $r_{f_r v_j} = 0$ и выражение можно упростить: $\hat{s}_j^2 = \sum_{r=1}^{m} a_{jr}^2 + b_j^2 + 2\sum_{l>r=1}^{m} a_{jr}a_{jl}r_{f_r f_l}$.

Дальнейшее упрощение может быть получено для некоррелированных общих факторов, когда и $r_{f_r f_l} = 0$, тогда $\hat{s}_j^2 = b_j^2 + \sum_{r=1}^{m} a_{jr}^2$. В этом случае дисперсия признака z_j равна сумме относительных вкладов в дисперсию этого признака каждого из *m* общих и одного характерного фактора.

Компонент общей дисперсии $\sum_{r=1}^{m} a_{jr}^2 = h_j^2$ называется общностью показателя z_j, т.е. суммой относительных вкладов всех *m* общих факторов в дисперсию признака z_j. Вклад в дисперсию признака z_j характерного фактора v_j, или характерность, определяется слагаемым b^2_j. В свою очередь дисперсия характерного фактора состоит из двух составляющих: связанной со спецификой параметра S_j и связанной с ошибками измерений E_j.

Если факторы специфичности S_j и ошибки E_j не коррелированы между собой, то модель факторного анализа примет вид

$$z_j = a_{j1}f_1 + a_{j2}f_2 + \ldots + a_{jm}f_m + k_j S_j + c_j E_j. \quad (6)$$

Вклад характерного фактора в дисперсию признака может быть представлен следующим образом: $b_j^2 = k_j^2 + c_j^2$.

Если выделить из дисперсии признака составляющую ошибки, то получим характеристику, называемую надежностью: $r_j^2 = h_j^2 + k_j^2.$. Вклад фактора *fr* в суммарную дисперсию всех признаков определяется соответствующей суммой квадратов коэффициентов при нормированных значениях: $\Delta_r = \sum_{j=1}^{n} a_{jr}^2.$. Вклад всех общих факторов в суммарную дисперсию признаков рассчитывается как сумма вкладов всех факторов: $\Delta_o = \sum_{r=1}^{m} \Delta_r.$.

Отношение этой суммы к размерности исходного признакового пространства $\xi = \Delta_o / k$ называют *полнотой факторизации*. Исходные данные матрицы *X* (или *Z*) позволяют получить матрицу парных коэффициентов корреляции *R*. Для воспроизведения всех связей переменных в корреляционной матрице может быть использована матрица

$$K = (A\ B): \quad R = KK^T = (A\ B)\begin{pmatrix} A^T \\ B^T \end{pmatrix} = AA^T + BB^T. \quad (7)$$

Введем обозначение для первого слагаемого — редуцированной корреляционной матрицы: $R_h = AA^T$.

Матрицу BB^T вследствие того, что *B* является диагональной матрицей, можно представить в виде $BB^T=B^2$. Таким образом, матрица парных коэффициентов корреляции исходных показателей может быть представлена в виде суммы:$R=R_h+B^2$. В то время как *R* является корреляционной матрицей с единицами на главной диагонали, матрица R_h представляет собой корреляционную матрицу с общностями на главной диагонали.

Для стандартизованных исходных признаков *Z* корреляционная матрица *R* тождественна ковариационной матрице Σ. Если рассматривать данные как выборку из генеральной совокупности, то определенная по выборочным данным матрица Σ (или *R*) является оценкой истинной ковариационной (корреляционной) матрицы.

Несмещенная оценка может быть получена в виде $R = \frac{1}{n-1} ZZ^T$. (8)

Рассчитаем редуцированную корреляционную матрицу с учетом равенства, используя для восстановления нормированных исходных признаков только общие факторы:

$$R = \frac{1}{n-1} AF(AF)^T = \frac{1}{n-1} AF(AF)^T = \frac{1}{n-1} AFF^T A^T = A\frac{1}{n-1} FF^T A^T.$$

Выражение, стоящее между A и A^T, является корреляционной матрицей стохастических связей между общими факторами $C = \frac{1}{n-1} FF^T$.

При этом общее выражение для редуцированной корреляционной матрицы примет вид $R_h = ACA^T$. (9)

Если общие факторы не коррелированы между собой, то матрица C будет единичной, и при этом $R_h = AA^T$. (10)

Два последних выражения представляют собой *фундаментальную теорему факторного анализа*.

При использовании факторного анализа исследователь сталкивается с различными проблемами. Наиболее часто они возникают в процессе содержательной интерпретации результатов анализа. Многие из проблем носят частный характер, не относящийся непосредственно к факторному анализу и присущий определенному классу задач, например, наличие плохо обусловленных матриц парных коэффициентов корреляций, присущее классу экономико-статистических задач.

Среди проблем проведения факторного анализа можно выделить проблемы робастности, общности, выбора факторов, вращения факторов и оценки их значений и содержательной интерпретации, а также проблему построения динамических моделей.

В классическом факторном анализе на основе исходной матрицы «объект — признак» формируется матрица нормированных значений исходных признаков. Опыт решения практических задач показывает, что наличие грубых ошибок данных при многомерном анализе может привести к дальнейшим трудностям. Малую чувствительность к наличию грубых

ошибок данных обеспечивают робастные оценки параметров: среднего значения и дисперсии или среднего квадратического отклонения. Рассчитываемая матрица парных коэффициентов корреляции является симметрической матрицей порядка k. Она является диагональной, и на ее главной диагонали стоят единицы, соответствующие дисперсиям исходных нормированных показателей. Данная матрица R является исходной для проведения компонентного анализа. Для факторного анализа необходимо получить редуцированную матрицу R_h. Редуцированная корреляционная матрица R_h служит основной для факторного анализа. Она также является симметрической порядка k, но на ее главной диагонали вместо единиц стоят общности h^2_j. На основе этой матрицы рассчитывается матрица весовых коэффициентов A. Ее элементы являются характеристиками стохастической связи между исходными признаками и общими факторами.

При переходе от редуцированной корреляционной матрицы к матрице весовых коэффициентов необходимо решить проблему нахождения факторов, включающую вопросы определения числа извлекаемых общих факторов и их вида.

Значения весовых коэффициентов являются координатами признаков на новых осях координат. Этими координатными осями являются общие факторы. Чаще всего для их нахождения используется метод главных компонент.

Задача воспроизведения матрицы R_h по матрице A не имеет однозначного решения. Выбор одной из возможных матриц является составной частью решения задачи вращения координатных осей.

После получения новой интегральной системы измерения — общих факторов — можно оценить значения индивидуальных факторов для каждого объекта исследования. Сопоставление факторных решений в течение длительного периода обеспечивается динамическим моделированием, позволяющим выявить те признаки, влияние которых в будущем будет снижаться или, наоборот, возрастать.

Вопросы для самопроверки

1. В чем заключаются основные задачи снижения размерности?

2. В чем отличие метода главных компонент от общей модели факторного анализа?

3. Как можно записать модель факторного анализа в матричной форме?

4. Как можно записать в матричном виде фундаментальную теорему факторного анализа?

5. С какими основными проблемами сталкивается исследователь при применении факторного анализа?

Лекция 5. Методы снижения размерности – компонентный анализ

1. Компонентный анализ

Компонентный анализ предназначен для преобразования системы k исходных признаков, в систему k новых показателей (главных компонент).

Главные компоненты не коррелированны между собой и упорядочены по величине их дисперсий, причем, первая главная компонента, имеет наибольшую дисперсию, а последняя, k-я, наименьшую. При этом выявляются неявные, непосредственно не измеряемые, но объективно существующие закономерности, обусловленные действием как внутренних, так и внешних причин. Компонентный анализ является одним из основных методов факторного анализа. В задачах снижения размерности и классификации обычно используются m первых компонент (m<< k).

При наличии результативного показателя Y может быть построено уравнение регрессии на главных компонентах.

На основании матрицы исходных данных:

$$X=\begin{pmatrix} x_{11} \dots x_{1j} \dots x_{1k} \\ x_{i1} \dots x_{ij} \dots x_{ik} \\ \dots\dots\dots\dots\dots \\ x_{n1} \dots x_{nj} \dots x_{nk} \end{pmatrix}$$

размерности (n×k) , где x_{ij} – значение j-го показателя у i-го наблюдения (i=1,2,...,n; j=1,2,...,k) вычисляют средние значения показателей $\overline{x}_1, \overline{x}_2, \dots \overline{x}_k$ а также s_1, ..., s_k и матрицу нормированных значений:

$$Z=\begin{pmatrix} z_{11} \cdots z_{1j} \cdots z_{1k} \\ \dots\dots\dots\dots\dots \\ z_{i1} \cdots z_{ij} \cdots z_{ik} \\ \dots\dots\dots\dots\dots \\ z_{n1} \cdots z_{nj} \cdots z_{nk} \end{pmatrix}$$

с элементами:

$$z_{ij}=\frac{x_{ij}-\overline{x_j}}{s_j} \quad . \qquad (1)$$

Если корреляция признаков $z1$ и $z2$ отсутствует (ρ=0), то эллипс вырождается в окружность. Наличие корреляции приводит к повороту осей эллипса соответственно на 45^0 или 45^0 относительно оси $z1$ в зависимости от знака коэффициента корреляции ρ. Размеры сечений по осям координат одинаковы (*bc*=*de*), так как они пропорциональны единичным средним квадратическим отклонениям нормированных признаков $z1$ и $z2$ с одним и тем же коэффициентом пропорциональности, определяемым параметром эллипса *c*.

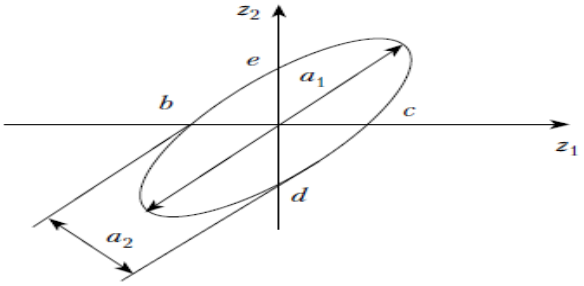

Рис 1. Сечение плотности вероятности двумерного нормального распределения стандартизованных коррелированных признаков

Рассчитывается матрица парных коэффициентов корреляции:

$$R=\frac{1}{n}Z^{T}Z \quad . \qquad (2)$$

с элементами:

$$r_{jl}=\frac{1}{n}\sum_{i=1}^{n}z_{ij}z_{il}=\frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij}-\overline{x_j}\right)\left(x_{il}-\overline{x_l}\right)}{Sj\times S_e} \quad \text{где, } j1 = 1,2, \ldots, k. \qquad (3)$$

На главной диагонали матрицы R, т.е. при j=l, $r_{jj}=\frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij}-\overline{x_j}\right)^2}{s_j^2}=1$

Модель компонентного анализа имеет вид:

$$Z_{ij}=\sum_{v=1}^{k}a_{jv}f_{iv} \quad , \qquad (4)$$

где: a_{iv} – “вес”, факторная нагрузка, ν-ой главной компоненты на j-ой переменной;

f_{iv} – значение ν-й главной компоненты для i-го наблюдения (объекта), где ν=1,2, ... ,k.

В матричной форме модель имеет вид: $Z = F A^{T}$ (5)

где: $F = \begin{pmatrix} f_{11} \cdots f_{1v} \cdots f_{1k} \\ \cdots\cdots\cdots\cdots \\ f_{i1} \cdots f_{iv} \cdots f_{ik} \\ \cdots\cdots\cdots\cdots \\ f_{n1} \cdots f_{nv} \cdots f_{nk} \end{pmatrix}$ – матрица значений главных компонент размерности (n×k)

$A = \begin{pmatrix} a_{11} \cdots a_{1v} \cdots a_{1k} \\ \cdots\cdots\cdots\cdots \\ a_{i1} \cdots a_{iv} \cdots a_{ik} \\ \cdots\cdots\cdots\cdots \\ a_{k1} \cdots a_{kv} \cdots a_{kk} \end{pmatrix}$ – матрица факторных нагрузок размерности (k×k).

A^{T} – транспонированная матрица A;

f_{iv} – значение ν-й главной компоненты у i-го наблюдения (объекта);

a_{jv} – значение факторной нагрузки ν-й главной компоненты на j-й переменной.

Матрица F описывает n наблюдений в пространстве k главных компонент. При этом элементы матрицы F нормированы, то есть: $\overline{f_v} = \frac{1}{n}\sum_{i=1}^{n} f_{iv} = 0, \qquad S_{f_v}^2 = \frac{1}{n}\sum_{i=1}^{n} f_{jv}^2 = 1$, а главные компоненты не коррелированы между собой. Из этого следует, что, $(1/n)\, F^{T} F = E$, (6)

где $E = \left(\begin{array}{cc:c} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \hdashline 0 & 0 & 1 \end{array}\right)$ – единичная матрица размерности (k×k).

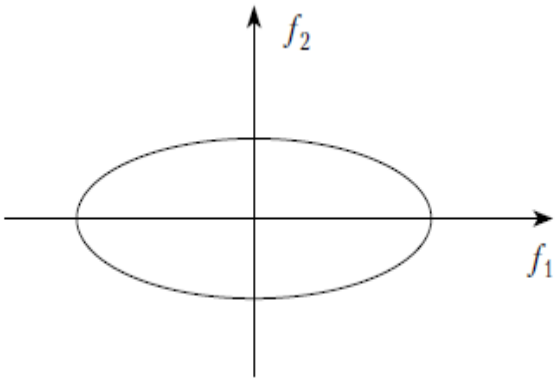

Рис. 2. Сечение плотности вероятности двумерного нормального распределения главных компонент

Выражение может быть также представлено в виде:

$$\frac{1}{n}\sum_{i=1}^{n} f_{iv} f_{iv'} = \begin{cases} 1 & \text{при} \quad v = v' \\ 0 & \text{при} \quad v \neq v' \end{cases}$$

$v, v' = 1,2,\ldots,k.$. (7)

С целью интерпретации элементов матрицы A, рассмотрим выражение для парного коэффициента корреляции, между Z_j-переменной и, например, f_1-й главной компонентой.

Так как, z_j и f_1 нормированы, будем иметь:

$$r_{z_j f_i} = \frac{1}{n}\sum_{i=1}^{n} z_{ij} f_{i1} = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{v=1}^{k} a_{jv} f_{iv}\right) f_{i1} = a_{j1}\frac{1}{n}\sum_{i=1}^{n} f_{i1}^2 + \sum_{v=2}^{k} a_{jv}\left(\frac{1}{n}\sum_{i=1}^{n} f_{i1} f_{iv}\right).$$

, окончательно получим: $r_{z_j f_v} = a_{j1}$. (8)

Рассуждая аналогично, можно записать в общем виде: $r_{z_j f_v} = a_{jv}$ для всех j=1,2, ... , k и v=1,2, ... , k.

Таким образом, элемент a_{jv} матрицы факторных нагрузок A, характеризует тесноту линейной связи между z_j-исходной переменной и f_v-й главной компонентой, то есть $-1 \le a_{jv} \le +1$.

Рассмотрим теперь выражение для дисперсии z_j-й нормированной переменной:

$$S_j^2 = \frac{1}{n}\sum_{i=1}^{n} z_{ij}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{v=1}^{k} a_{jv} f_{iv}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{v=1}^{k} a_{jv}^2 f_{iv}^2 + 2\sum_{v \neq v'} a_{jv} a_{jv'} \times f_{iv} f_{iv'}\right] =$$

$$= \sum_{v=1}^{k} a_{jv}^2 \left(\frac{1}{n}\sum_{i=1}^{n} f_{iv}^2\right) + 2\sum_{v \neq v'} a_{jv} a_{jv'} \left(\frac{1}{n}\sum_{i=1}^{n} f_{iv} f_{iv'}\right),$$

где ν, ν'=1,2, ... , k.

Окончательно получим: $S_j^2 = \sum_{v=1}^{k} a_{jv}^2 = 1.$. (9)

По условию переменные z_j нормированы и $s_j^2 = 1$. Таким образом, дисперсия z_j-й переменной представлена своими составляющими, определяющими долю вклада в нее всех k главных компонент.

Полный вклад ν-й главной компоненты в дисперсию всех k исходных признаков вычисляется по формуле: $\lambda_k = \sum_{j=1}^{k} a_{jv}^2$. . (10)

Одно из основополагающих условий метода главных компонент, связано с представлением корреляционной матрицы R, через матрицу факторных нагрузок A. $R=(1/n)Z^TZ=(1/n)(FA^T)^TFA^T=A((1/n)F^TF)A^T$.

Окончательно получим: $R = AA^T$. (11)

Перейдем теперь непосредственно к отысканию собственных значений и собственных векторов корреляционной матрицы R.

Из линейной алгебры известно, что для любой симметрической матрицы R, всегда существует такая ортогональная матрица U, что выполняется условие:

$$U^T R U = \Lambda \qquad (12)$$

где

$\Lambda = \begin{pmatrix} \lambda_1 & 0 \cdots 0 \\ 0 & \lambda_2 \cdots 0 \\ \cdots\cdots\cdots & \\ 0 & 0 \cdots \lambda_k \end{pmatrix}$ – диагональная матрица собственных значений размерности (k×k);

$U = \begin{pmatrix} u_{11} \cdots u_{1v} \cdots u_{1k} \\ \cdots\cdots\cdots\cdots\cdots \\ u_{j1} \cdots u_{jv} \cdots u_{jk} \\ \cdots\cdots\cdots\cdots\cdots \\ u_{k1} \cdots u_{kv} \cdots u_{kk} \end{pmatrix}$ – ортогональная матрица собственных векторов размерности (k×k).

Так как матрица R положительно определена, т.е. ее главные миноры положительны, то все собственные значения положительны – $\lambda_v > 0$ для всех v=1,2,.,к. В компонентном анализе элементы матрицы Λ ранжированы $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_v \geq \ldots \geq \lambda_к > 0$. Собственное значение λv характеризует вклад v-й главной компоненты в суммарную дисперсию исходного признакового пространства.

Таким образом, первая главная компонента вносит наибольший вклад в суммарную дисперсию, а последняя k-я – наименьший. В ортогональной матрице U собственных векторов, v-й столбец является собственным вектором, соответствующим λ_v-му значению.

Собственные значения $\lambda_1 \geq \ldots \geq \lambda_v \geq \ldots \geq \lambda_к$ находятся как корни характеристического уравнения: /ΛE-R/=0. Собственный вектор V_v, соответствующий собственному значению λ_v корреляционной матрицы R, определяется как отличное от нуля решение уравнения: $(\lambda_v E - R)V_v = 0$.

Нормированный собственный вектор Uv равен: $U_v = \frac{V_v}{\sqrt{V_v^T V_v}}$. (13)

Из условия ортогональности матрицы U следует, что $U^{-1}=U^{T}$, но тогда по определению матрицы R и Λ подобны, так как они удовлетворяют условию: $U^{-1}RU=\Lambda$.

Так как следы у подобных матриц равны, то $tr\Lambda=tr(U^{-1}RU)=tr[R(UU^{-1})]=trR$.

Умножение матрицы U на обратную матрицу U^{-1}, дает единичную матрицу E. Следы матричных произведений $(U^{-1})\times(RU)$ и $R\times(UU^{-1})$ также равны.

Учитывая, что сумма диагональных элементов матрицы R равна k, будем иметь: $tr\Lambda=trR=k$. Таким образом, $\sum_{\nu=1}^{k}\lambda_\nu = k$. (14)

Представим матрицу факторных нагрузок A в виде: $A=UA^{1/2}$, а ν-й столбец матрицы A: $A\nu=U\nu\cdot\lambda_\nu^{1/2}$, где U_ν – собственный вектор матрицы R, соответствующий собственному значению λ_ν. Найдем норму вектора A_ν: $/A_\nu/^2=A_\nu^T A_\nu=\lambda_\nu^{1/2} U_\nu^T U_\nu\lambda_\nu^{1/2}=\lambda_\nu$. Здесь учитывалось, что вектор U_ν нормированный и $U_\nu^T U_\nu=1$. Таким образом, $\lambda_\nu=\sum_{j=1}^{k}a_{j\nu}^2$. (15)

Можно сделать вывод, что собственное значение λ_ν характеризует вклад ν-й главной компоненты в суммарную дисперсию всех исходных признаков. Можно показать, что $A^T A=\Lambda$. Общий вклад всех главных компонент в суммарную дисперсию равен k. Тогда удельный вклад ν-й главной компоненты определяется по формуле: $\frac{\lambda_\nu}{k}100\%$.

Суммарный вклад m первых главных компонент определяется из выражения: $\frac{\sum_{\nu=1}^{m}\lambda_\nu}{k}100\%$. (16)

Обычно для анализа используют m первых главных компонент, суммарный вклад которых превышает 60–70%. Матрица факторных нагрузок A используется для экономической интерпретации главных компонент, которые представляют линейные функции исходных признаков. Для экономической интерпретации f_v используются лишь те x_j, для которых, $/a_{jv}/>0,5$.

Значения главных компонент для каждого i-го объекта (i=1,2,...,n) задаются матрицей F. Матрицу значений главных компонент можно получить из формулы: $Z=FA^{T}$, откуда, $F=Z(A^{T})^{-1}=ZV\Lambda^{-1/2}$, (17)

где $F=\begin{pmatrix} f_{11}\cdots f_{1v}\cdots f_{1k} \\ \cdots\cdots\cdots\cdots \\ f_{i1}\cdots f_{iv}\cdots f_{ik} \\ \cdots\cdots\cdots\cdots \\ f_{n1}\cdots f_{nv}\cdots f_{nk} \end{pmatrix}$;

Z – матрица нормированных значений исходных показателей.

2. Построение модели регрессии по главным компонентам

Уравнение регрессии на главных компонентах строится по данным вектора значений результативного показателя y и матрицы значений нормированных компонент F. Данный вид регрессии может быть радикальным способом борьбы с мультиколлинеарностью исходных признаков.

Значения ненормированных главных компонент f* jv можно получить из значений нормированных главных компонент f jv по формуле:

$$f^{*}_{jv}=\sqrt{\lambda_j}\,f_{jv}. \qquad (18)$$

Некоррелированность главных компонент между собой и тесноту их связи с результативным показателем y показывает матрица парных коэффициентов корреляции.

Например, если из матрицы парных коэффициентов корреляции следует, что зависимая переменная y наиболее тесно связана с первой, второй и третьей главными компонентами, то можно предположить, что только эти

главные компоненты войдут в регрессионную модель. Уравнение регрессии на главных компонентах строится по алгоритму пошагового регрессионного анализа, где в качестве аргументов используются не исходные показатели, а интегральные факторы.

Первоначально в модель, как правило, включают все главные компоненты. Если Fнабл > Fкр при заданном уровне значимости, то следовательно хотя бы один из коэффициентов регрессии β1 -βк не равен нулю.

К достоинству данной модели следует отнести тот факт, что главные компоненты не коррелированы между собой. Учитывая это, можно сразу исключить из уравнения все незначимые по t –тесту интегральные факторы, и уравнение примет скорректированный вид, без изменения значений коэффициентов при оставшихся главных компонентах. Далее можно сравнить уравнение регрессии по главным компонентам с регрессионными моделями по исходным показателям, полученными разными методами.

Может получиться, что уравнение регрессии по главным компонентам обладает несколько лучшими аппроксимирующими свойствами по сравнению с регрессионными моделями по исходным показателям. Кроме того, в уравнении регрессии по интегральным факторам главные компоненты являются линейными функциями всех исходных показателей, в то время как в конечное уравнение регрессии по исходным признакам, как правило, включает только некоторые из исходных переменных, оказавшихся статистически значимыми. Однако приходится учитывать тот факт, что модель регрессии на главные компоненты может быть трудно интерпретируема, так как в нее могут входить значимые главные компоненты, не имеющие однозначной интерпретации в рамках решаемой задачи, с существенным вкладом в суммарную дисперсию исходных показателей. Но исключение из уравнения таких не интерпретируемых, но значимых интегральных, факторов значительно ухудшает аппроксимирующие свойства модели из-за снижения коэффициента

детерминации и увеличения средней относительной ошибки аппроксимации и выборочной дисперсии. Поэтому в таком случае, в качестве регрессионной модели лучше выбрать уравнение по исходным регрессорам.

Вопросы для самопроверки

1. Какова цель проведения компонентного анализа?

2. Опишите модель метода главных компонент.

3. Что представляют собой собственные векторы и собственные значения корреляционной матрицы и как они могут быть использованы для получения матрицы весовых коэффициентов?

4. Дайте определение квадратичных форм и главных компонент. Укажите главные компоненты для двумерного, трехмерного и конечномерного пространств.

5. Как получают и для чего используют матрицы индивидуальных значений главных компонент?

6. Каковы свойства ортогональной матрицы собственных векторов в модели метода главных компонент?

7. В чем сущность регрессии на главные компоненты?

8. В чем заключается основная проблема применимости метода регрессии на главные компоненты?

9. Как определить относительный вклад m первых главных компонент в суммарную дисперсию?

10. Сколько главных компонент (факторов) следует выделять при снижении признакового пространства?

11. Как проинтерпретировать выделенные главные компоненты (факторы)?

Лекция 6. Кластерный анализ и его разновидности (часть 1)

1.Основные понятия кластерного анализа

В статистических исследованиях группировка первичных данных является основным приемом решения задачи классификации, а поэтому и основой всей дальнейшей работы с собранной информацией.

Традиционно эта задача решается следующим образом. Из множества признаков, описывающих объект, отбирается один, наиболее информативный с точки зрения исследователя, и производится группировка в соответствии со значениями данного признака. Если требуется провести классификацию по нескольким признакам, ранжированным между собой по степени важности, то сначала производится классификация по первому признаку, затем каждый из полученных классов разбивается на подклассы по второму признаку и т.д. Подобным образом строится большинство комбинационных статистических группировок. В тех случаях, когда не представляется возможным упорядочить классификационные признаки, применяется наиболее простой метод многомерной группировки – создание интегрального показателя (индекса), функционально зависящего от исходных признаков, с последующей классификацией по этому показателю. Развитием этого подхода является вариант классификации по нескольким обобщающим показателям (главным компонентам), полученным с помощью методов факторного или компонентного анализа. При наличии нескольких признаков (исходных или обобщенных), задача классификации может быть решена методами кластерного анализа, которые отличаются от других методов многомерной классификации отсутствием обучающих выборок, т.е. априорной информации о распределении генеральной совокупности, которая представляет собой вектор X.

Различия между схемами решения задачи по классификации во многом определяются тем, что понимают под понятием “сходство” и “степень сходства”. После того как сформулирована цель работы, естественно

попытаться определить критерии качества, целевую функцию, значения которой позволят сопоставить различные схемы классификации.

В экономических исследованиях целевая функция, как правило, должна минимизировать некоторый параметр, определенный на множестве объектов (например, цель классифицировать оборудование может явиться группировка, минимизирующая совокупность затрат времени и средств на ремонтные работы). В случаях, когда формализовать цель задачи не удается, критерием качества классификации может служить возможность содержательной интерпретации найденных групп.

Пусть исследуется совокупность n объектов, каждый из которых характеризуется по k замеренным на нем признакам X. Требуется разбить эту совокупность на однородные, в некотором смысле, группы (классы). При этом практически отсутствует априорная информация о характере распределения измерений X внутри классов. Полученные в результате разбиения группы обычно называются кластерами(таксонами, образами), методы их нахождения – кластер-анализом (соответственно численной таксономией или распознаванием образов с самообучением). При этом, необходимо с самого начала, четко представить, какая из двух задач классификации подлежит решению. Если решается обычная задача типизации, то совокупность наблюдений разбивают на сравнительно небольшое число областей группирования (например, интервальный вариационный ряд в случае одномерных наблюдений) так, чтобы элементы одной такой области находились друг от друга по возможности на небольшом расстоянии. Решение другой задачи, заключается в определении естественного расслоения исходных наблюдений на четко выраженные кластеры, лежащие друг от друга на некотором расстоянии. Если первая задача типизации всегда имеет решение, то при второй постановке, может оказаться, что множество исходных наблюдений не обнаруживает естественного расслоения на кластеры, т.е. образует один кластер.

Многие методы кластерного анализа довольно просты, однако эффективное решение задач поиска кластеров требует большого числа арифметических и логических операций, поэтому необходимо применение специального программного обеспечения.

2.Иерархические кластер-процедуры

Обычной формой представления исходных данных в задачах кластерного анализа служит матрица X: $X=\begin{pmatrix} x_{11}\cdots & x_{1j}\cdots & x_{1k} \\ \vdots & \vdots & \vdots \\ x_{i1}\cdots & x_{ij}\cdots & x_{ik} \\ \vdots & \vdots & \vdots \\ x_{n1}\cdots & x_{nj}\cdots & x_{nk} \end{pmatrix}$, каждая строка которой, представляет результат измерений k, рассматриваемых признаков на одном из обследованных объектов.

В конкретных ситуациях, может представлять интерес как группировка объектов, так и группировка признаков. В тех случаях, когда разница между двумя этими задачами не существенна, например, при описании некоторых алгоритмов, мы будем пользоваться только термином “объект”, включая в это понятие и “признак”.

Матрица X не является единственным способом представления данных в задачах кластерного анализа. Иногда, исходная информация задана в виде квадратной матрицы: $R=(r_{ij}),\ i,j=1,\ 2,\ \ldots,\ k,$ элемент r_{ij} , который определяет степень близости i-го объекта к j-му.

Большинство алгоритмов кластерного анализа полностью исходит из матрицы расстояний (или близостей), либо требует вычисления отдельных ее элементов, поэтому, если данные представлены в форме X, то первым этапом решения задачи поиска кластеров будет выбор способа вычисления расстояний, или близости, между объектами или признаками.

Относительно проще решается вопрос об определении близости между признаками. Как правило, кластерный анализ признаков преследует те же цели, что и факторный анализ – выделение групп связанных между собой

признаков, отражающих определенную сторону изучаемых объектов. Мерами близости в этом случае служат различные статистические коэффициенты связи.

Наиболее трудным и наименее формализованным в задаче классификации является определение понятия однородности объектов. В общем случае, понятие однородности объектов задается либо введение правила вычисления расстояний $\rho(x_i,x_j)$ между любой парой исследуемых объектов (x_1, x_2, ... , x_n), либо заданием некоторой функции $r(x_i,x_j)$, характеризующей степень близости i-го и j-го объектов. Если задана функция $\rho(x_i,x_j)$, то близкие с точки зрения этой метрики объекты считаются однородными, принадлежащими к одному классу. Очевидно, что необходимо при этом сопоставлять $\rho(x_i,x_j)$ с некоторыми пороговыми значениями, определяемыми в каждом конкретном случае по-своему.

Аналогично используется и мера близости $r(x_i,x_j)$, при задании которой мы должны помнить о необходимости выполнения следующих условий: симметрии $r(x_i,x_j)= r(x_j,x_i)$; максимального сходства объекта с самим собой $r(x_i,x_i)= r(xmaxi\ j_i,x_j)$, при $1\leq i,j\leq n$, и монотонного убывания $r(x_i,x_j)$ по мере увеличения $\rho(x_i,x_j)$, т.е. из $\rho(x_k,x_l)\geq\rho\ (x_i,x_j)$ должно следовать неравенство $r(x_k,x_l)\leq r(x_i,x_j)$.

Выбор метрики или меры близости является узловым моментом исследования, от которого в основном зависит окончательный вариант разбиения объектов на классы при данном алгоритме разбиения. В каждом, конкретном случае, этот выбор должен производиться по-своему, в зависимости от целей исследования, физической и статистической природы вектора наблюдений X, априорных сведений о характере вероятностного распределения X.

Рассмотрим наиболее широко используемые в задачах кластерного анализа расстояния и меры близости:

– расстояние Махаланобиса — способ (мера, метрика) нахождения расстояния между объектами в задачах кластерного анализа.

Данное расстояние было предложено в 1936 г. индийским статистиком П. Ч. Махаланобисом и рассчитывается по формуле

$$d_M(X_i, X_j) = \sqrt{(X_i - X_j)^T \Sigma^{-1} (X_i - X_j)}, \quad (1)$$

где $X_i = (x_{i1}, x_{i2}, ..., x_{il}, ..., x_{ik})^T$ – вектор-столбец, соответствующий *i*-му наблюдению;

x_{il} – значение *l*-го показателя для *i*-го наблюдения (объекта);

Σ– ковариационная матрица генеральной совокупности, соответствующая *k*-мерному вектору наблюдений *Xi*.

Расстояние Махаланобиса за счет ковариационной матрицы Σ учитывает как дисперсии *k* признаков, так и степень их взаимосвязи. Однако формула предполагает, что все *k* признаков одинаково значимы для результатов классификации. В противном случае используют *обобщенное* («взвешенное») расстояние Махаланобиса, которое задается формулой

$$d_{OM}(X_i, X_j) = \sqrt{(X_i - X_j)^T \Delta^T \Sigma^{-1} \Delta (X_i - X_j)}, \quad (2)$$

где Δ — симметричная неотрицательно определенная матрица весовых коэффициентов, которая обычно выбирается диагональной.

Отметим, что расстояние Махаланобиса учитывает корреляции между переменными и инвариантно к масштабу. Расстояние Махаланобиса тесно связано с распределением *T*-квадрат Хотеллинга (*Hotteling's T-squared distribution*) и используется в многомерном статистическом тестировании, а также в линейном дискриминантном анализе.

Следующие три вида расстояний являются частным случаем расстояния Махаланобиса.

- обычное Евклидово расстояние $\rho_E(x_i, x_j) = \sqrt{\sum_{c=1}^{k} (x_{ic} - x_{jc})^2}$, (3)

где $x_{ie,}$, x_{je} – величина e-ой компоненты у i-го (j-го) объекта (e=1,2,...,к, i,j=1,2,...,n).

Использование этого расстояния оправдано в следующих случаях:

а) наблюдения берутся из генеральной совокупности, имеющей многомерное нормальное распределение с ковариационной матрицей вида $\sigma^2 E_к$, т.е. компоненты X взаимно независимы и имеют одну и ту же дисперсию, где $E_к$ – единичная матрица;

б) компоненты вектора наблюдений X однородны по физическому смыслу и одинаково важны для классификации;

в) признаковое пространство совпадает с геометрическим пространством.

Естественное, с геометрической точки зрения, евклидово пространство может оказаться бессмысленным (с точки зрения содержательной интерпретации), если признаки измерены в разных единицах.

Чтобы исправить положение, прибегают к нормированию каждого признака путем деления центрированной величины на среднее квадратическое отклонение и переходят от матрицы X, к нормированной матрице. Однако, эта операция может привести к нежелательным последствиям. Если, кластеры хорошо разделены по одному признаку, и не разделены по другому, то после нормирования дискриминирующие возможности первого признака, будут уменьшены в связи с увеличением “шумового” эффекта второго.

– “взвешенное” Евклидово пространство $$\rho_{ВЕ}(x_i, x_e) = \sqrt{\sum_{e=1}^{k} \omega_e (x_{ie} - x_{je})^2} \quad (4)$$

применяется в тех случаях, когда каждой компоненте x_l вектора наблюдений X, удается приписать некоторый “вес” ω_l, пропорционально степени важности признака в задаче классификации.

Обычно принимают $0 \le \omega_e \le 1$, где e=1,2,...k. Определение “весов”, как правило, связано с дополнительными исследованиями, например,

организацией опроса экспертов и обработкой их мнений. Определение весов ω_l, только по данным выборки, может привести к ложным выводам.

– Хеммингово расстояние используется как мера различия объектов, задаваемых дихотомическими признаками.

Это расстояние определяется по формуле: $$\rho_H(x_i, x_j) = \sum_{e=1}^{k} |x_{ie} - x_{je}| \quad (5)$$

и равно числу несовпадений значений соответствующих признаков, в рассматриваемых i-м и j-м объектах.

– расстоянием Чебышева между *n*-мерными числовыми векторами называется максимум модуля разности компонент этих векторов:

$$d_{ij} = \max |x_{il} - x_{jl}|. \quad (6)$$

Это расстояние лучше использовать, когда необходимо определить два объекта как различные, если они отличаются по какому-то одному измерению.

Как правило, решение задач классификации многомерных данных, предусматривает в качестве предварительного этапа исследования реализацию методов, позволяющих выбрать из компонент x_1, x_2, ..., $x_к$, наблюдаемых векторов X, сравнительно небольшое число наиболее существенно информативных, т.е. уменьшить размерность наблюдаемого пространства.

В ряде процедур классификации (кластер-процедур) используют понятия расстояния между группами объектов и меры близости двух групп объектов.

Пусть, s_i – i-я группа (класс, кластер), состоящая из n_i объектов; $\bar{x}_i$ – среднее арифметическое векторных наблюдений s_i группы, т.е. "центр тяжести" i-й группы; $\rho(s_l, s_m)$ – расстояние между группами s_l и s_m.

Наиболее употребительными расстояниями и мерами близости между классами объектов являются:

–расстояние, измеряемое по принципу “ближайшего соседа” –

$$\rho_{min}(S_e, S_m) = \min_{x_i \in S_e, x_j \in S_m} \rho(x_i, x_j); \quad (7)$$

–расстояние, измеряемого по принципу “дальнего соседа” –

$$\rho_{max}(S_e, S_m) = \max_{x_i \in S_e, x_j \in S_m} \rho(x_i, x_j); \quad (8)$$

–расстояние, измеряемое по “центрам тяжести” групп –

$$\rho_{ц.т.}(S_e, S_m) = \rho(\overline{x}_e, \overline{x}_m); \quad (9)$$

–расстояние, измеряемое по принципу “средней связи”, определяется как среднее арифметическое всех попарных расстояний между представителями рассматриваемых групп –

$$\rho_{ср}(S_e, S_m) = \frac{1}{n_e n_m} \sum_{x_i \in S_e} \sum_{x_j \in S_m} \rho(x_i, x_j). \quad (10)$$

Академиком А.Н. Колмогоровым было предложено “обобщенное расстояние” между классами, которое включает в себя, в качестве частных случаев, все рассмотренные выше виды расстояний.

Расстояния между группами элементов особенно важно, в так называемых, агломеративных иерархических кластер-процедурах, так как принцип работы таких алгоритмов состоит в последовательном объединении элементов, а затем и целых групп, сначала самых близких, а затем все более и более отдаленных друг от друга. При этом расстояние между классами s_l и $s_{(m,q)}$, являющиеся объединением двух других классов s_m и s_q, можно определить по формуле: $\rho_{e,(m,q)} = \rho(S_e, S_{(m,q)}) = \alpha\rho_{em} + \beta\rho_{eq} + \gamma\rho_{mq} + \delta(\rho_{em} - \rho_{eq})$, (11)

где $\rho_{em} = \rho(S_e, S_m)$; $\rho_{eq} = \rho(S_e, S_q)$ $\rho_{mq} = \rho(S_m, S_q)$ – расстояния между классами s_l, s_m и s_q;

α, β, δ и γ – числовые коэффициенты, значения которых определяют специфику процедуры, ее алгоритм.

Например, при α= β=-δ=1/2и γ=0 приходим к расстоянию, построенному по принципу “ближайшего соседа”. При α= β=δ=1/2 и γ=0 – расстояние между классами определяется по принципу “дальнего соседа”, то

есть как расстояние между двумя самыми дальними элементами этих классов. А при $\alpha = \frac{n_m}{n_m + n_q}; \quad \beta = \frac{n_q}{n_m + n_q}, \quad \gamma=\delta=0$ приходим к расстоянию $\rho_{ср}$ между классами, вычисленному как среднее из расстояний между всеми парами элементов, один из которых берется из одного класса, а другой из другого.

Помимо рассмотренных метрик существуют и другие подходы, например, метод Уорда, приводящий, как правило, к образованию кластеров приблизительно равных размеров с минимальной внутригрупповой вариацией. Этот метод отличается от всех других методов, поскольку он использует дисперсионный анализ для оценки расстояний между кластерами. Метод минимизирует сумму квадратов (*SS*) расстояний для любых двух (гипотетических) кластеров, которые могут быть сформированы на каждом шаге.

Помимо перечисленных расстояний, широко используемых на практике в силу их легкой интерпретации и формализации в пакетах статистических программ (таких как, например, *SPSS*, *Statistica* и др.), разработаны и могут быть использованы и другие расстояния:

• метод групповых средних Ланса и Уильямса (G. N. Lance, W. T. Williams);

• двухгрупповой метод Миченера и Сокала (C. D. Michener, R. R. Sokal);

• метод Болла и Холла (G. H. Ball, D. J. Hall);

• методы Боннера и Хиверинена (R. E. Bonner, L. Hyvarinen);

• метод Мак-Куина Дж. (J. B. MacQueen);

• метод Себестьена (G. Sebestyen) и др.

Существует большое количество различных способов разбиения заданной совокупности элементов на классы. Поэтому представляет интерес, задача сравнительного анализа качества этих способов разбиения Q(S), определенного на множестве всех возможных разбиений. Под наилучшим

разбиением будем понимать такое разбиение, при котором достигается экстремум выбранного функционала качества.

Рис.1. Расстояние между точкой *x* и центром тяжести класса *l* в разбиении *S*

Следует отметить, что выбор того или иного функционала качества, как правило, опирается на эмпирические, профессиональные соображения исследователя, а не на строго формализованную схему соображения.

Наиболее распространенные функционалы качества разбиения, как правило, связаны с минимизацией внутригрупповых дисперсий. Пусть исследованием выбрана метрика ρ, в пространстве X и пусть $S=(s_1,s_2,...,s_p)$ – некоторое фиксированное разбиение наблюдений $x_1, ...,x_n$ на заданное число p классов $s_1,s_2,...,s_p$. За функционал качества берут сумму (“взвешенную”) внутриклассовых дисперсий:

$$Q_1(S) = \sum_{e=1}^{p} \sum_{x_i \in S_e} \rho^2(x_i, x_j). \qquad (12)$$

Существуют и другие функционалы качества разбиения.

Иерархические (древообразные) процедуры, являются наиболее распространенными, алгоритмами кластерного анализа. Они бывают двух типов: агломеративные и дивизимные.

В агломеративных процедурах начальным является разбиение, состоящее из n-одноэлементных классов, а конечным – из одного класса; в дивизимных – наоборот. Принцип работы иерархических агломеративных (дивизимных) процедур состоит в последовательном объединении (разделении) групп элементов, сначала самых близких (далеких), а затем все более отдаленных (близких) друг от друга. Большинство этих алгоритмов

исходит из матрицы расстояний (сходства). К недостаткам иерархических процедур следует отнести громоздкость их вычислительной реализации. Алгоритмы требуют вычисления на каждом шаге матрицы расстояний, а следовательно, емкой машинной памяти и большого количества времени. В этой связи, реализация таких алгоритмов при числе наблюдений, большем нескольких сотен, нецелесообразна, а в ряде случаев и невозможна.

Рассмотрим агломеративный иерархический алгоритм. На первом шаге алгоритма каждое наблюдение x_i (i=1,2,...n) рассматривается как отдельный кластер. В дальнейшем, на каждом шаге работы алгоритма, происходит объединение двух самых близких кластеров, и с учетом принятого расстояния по формуле пересчитывается матрица расстояний, размерность которой, очевидно, снижается на единицу.

Работа алгоритма заканчивается, когда все наблюдения объединены в один класс. Большинство программ, реализующих алгоритм иерархической классификации, предусматривает графическое представление результатов классификации в виде дендрограммы.

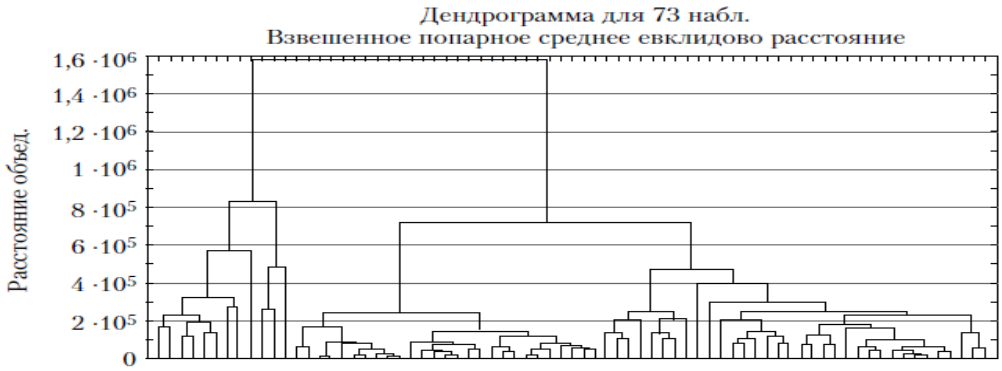

Рис 2. Пример построения дендрограммы в ППП STATISTICA

К недостаткам иерархических процедур следует отнести громоздкость их вычислительной реализации (алгоритм требует вычисления на каждом шаге матрицы расстояний). В связи с этим наглядность алгоритмов при числе наблюдений, большем нескольких сотен, теряется. Отметим, что дендрограмма позволяет исследователю наглядно увидеть последовательность объединения объектов в кластеры и расстояния, на

которых происходят объединения объектов, но не дает четкого ответа на вопрос «сколько кластеров необходимо выделить?».

Считается, что наибольший скачок в расстояниях при объединении объектов в кластеры сигнализирует о необходимости остановки процедуры объединения и изучения полученной кластерной структуры. Если мы выберем слишком большое число кластеров, то их наполняемость будет невысокой и мы можем упустить возможность изучения взаимосвязей внутри кластеров с помощью аппарата типологической регрессии. При малом числе кластеров характеристики объектов в них будут слишком размыты, что не позволит изучать присущие кластерам закономерности. Поэтому целесообразнее выбрать некую «золотую середину».

Отметим, что вывод о количестве кластеров решается в каждом конкретном случае по-своему и зависит от целей исследования и характера исходной информации. В некоторых случаях рекомендуется рассмотреть состав кластеров при разных решениях и выбрать тот вариант, который наиболее хорошо интерпретируется и понятен исследователю.

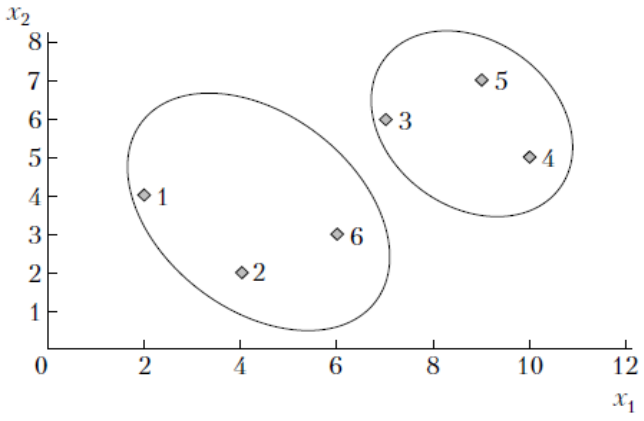

Рис. 3. Один из возможных вариантов классификации

Также при вынесении решения о целесообразности разбиения совокупности объектов на кластеры необходимо рассмотреть варианты разбиения при различных расстояниях между группами объектов и мерах близости групп объектов. Рекомендуется выбирать тот вариант разбиения, который получился наибольшим числом способов (т.е. является наиболее устойчивым), имеет минимальное значение функционала качества разбиения.

Помимо рассмотренных иерархических кластер – процедур существуют и другие методы классификации: метод «к – средних» и его модификации, двухэтапный (двухшаговый) кластерный анализ и другие подходы.

Вопросы для самопроверки

1. С какой целью проводится кластерный анализ?

2. Назовите способы представления информации для проведения кластерного

анализа.

3. Как можно рассчитать расстояние между объектами (признаками)?

4. В каких случаях может быть использовано обычное евклидово расстояние?

5. Как можно рассчитать расстояние между объектами, если они представлены дихотомическими признаками?

6. Что представляет собой функционал качества разбиения, с какой целью он используется?

7. Какие вы знаете иерархические кластер-процедуры?

8. В чем отличие агломеративных и дивизимных кластер-процедур?

9. Назовите наиболее часто употребляемые расстояния между классами объектов.

10. Что представляет собой дендрограмма, с какой целью она используется?

Лекция 7. Кластерный анализ и его разновидности (часть 2)

1. Итерационные алгоритмы классификации, метод *k*-средних

Если число кластеров заранее задано, то для классификации часто используют *параллельные кластер-процедуры — итерационные алгоритмы***.** Основной целью таких алгоритмов является нахождение способов сокращения числа перебора вариантов.

В параллельных кластер-процедурах реализуется идея оптимизации разбиения в соответствии с некоторым функционалом качества:

• для некоторого начального разбиения R_0 вычисляют значение $f(R_0)$;

• затем каждую из точек x_i поочередно перемещают во все кластеры и оставляют в том положении, которое соответствует наилучшему значению функционала качества;

• работа алгоритма заканчивается, когда перемещение точек не дает улучшения качества разбиения.

Обычно такой алгоритм применяют несколько раз, начиная с разных начальных разбиений R_0, и выбирают наилучший вариант разбиения. Если число объектов X_1, X_2, ..., X_n, подлежащих классификации, достаточно велико, то в этом случае целесообразно использовать итерационные алгоритмы, преимущество которых заключается в том, что на каждом шаге последовательно обсчитывается не вся, а лишь небольшая часть исходных данных.

Одним из наиболее распространенных среди таких неиерархических алгоритмов кластерного анализа является итерационный алгоритм *k*-средних (*k-mens*), также называемый *быстрым кластерным анализом*. В отличие от иерархических методов, которые не требуют предварительных предположений относительно числа кластеров, для использования этого метода необходимо иметь априорную информацию о наиболее вероятном количестве кластеров.

Идея метода *k*-средних заключается в разбиение множества объектов X_1, X_2, ..., X_n на заранее известное число p кластеров так, чтобы минимизировать функционал качества разбиения — сумму внутриклассовых дисперсий

$$Q_1(S) = \sum_{l=1}^{k} \sum_{X_i \in S_l} d^2(X_i, \bar{X}_l), \quad (13)$$

где $\bar{X}_l$ — вектор средних (центр тяжести) для группы S_l.

Пусть наблюдения X_1, X_2, ..., X_n требуется разбить на заданное число p однородных (в смысле некоторой метрики расстояний) классов. Алгоритм состоит в последовательном уточнении *эталонных точек* $E^{(\nu)} = \{e_1^{(\nu)}, e_2^{(\nu)}, ..., e_p^{(\nu)}\}$, где υ — номер итерации (0, 1, 2, …) с соответствующим пересчетом приписываемых им весов $\Omega^{(\nu)} = \{w_1^{(\nu)}, w_2^{(\nu)}, ..., w_p^{(\nu)}\}$.

Случайным образом выбираются k точек (например, первые) исследуемой совокупности, которые принимаются за центры классов. Таким образом, $e_i^0 = X_i; w_i^0 = 1; i = 1, 2, ..., k.$

На первом шаге алгоритма извлекается наблюдение X_{k+1} и выясняется, к какому из центров e_i^0 оно оказалось ближе всего. Именно этот самый близкий к X_{k+1} центр тяжести (эталон) заменяется новым эталоном, определяемым как центр тяжести старого эталона и присоединенной к нему точки X_{k+1} (с увеличением на единицу соответствующего ему веса). Все другие эталоны остаются неизменными.

Пересчет центров тяжести кластеров и их весов на δ-м шаге после извлечения наблюдения $X_{k+\delta}$ происходит для *i*-го кластера по следующим формулам:

$$e_i^{(\nu)} = \begin{cases} \dfrac{w_i^{(\nu-1)} e_i^{(\nu-1)} + X_{k+\nu}}{w_i^{(\nu-1)} + 1}, \text{ если } d(X_{k+\nu}, e_i^{(\nu-1)}) = \min d(X_{k+\nu}, e_j^{(\nu-1)}), \\ e_i^{(\nu-1)} \text{ в противном случае;} \end{cases}$$

$$w_i^{(\nu)} = \begin{cases} w_i^{(\nu-1)} + 1, \text{ если } d(X_{k+\nu}, e_i^{(\nu-1)}) = \min d(X_{k+\nu}, e_j^{(\nu-1)}), \\ w_i^{(\nu-1)} \text{ в противном случае.} \end{cases} \quad (14)$$

Если обнаружится несколько одинаковых минимальных значений $d(X_{l+v}, e_i^{(v-1)}) = \min d(X_{l+v}, e_i^{(v-1)})$, то можно условиться относить точку X_{l+v} к эталону с минимальным порядковым номером.

При достаточно большом числе итераций или при достижении большой совокупности (n — велико) дальнейший пересчет центров тяжести практически не приводит к изменению, т.е. имеет место сходимость к некоторому пределу. На этом работа итерационного алгоритма метода k-средних заканчивается.

Достоинства алгоритма k-средних – это простота и быстрота использования, понятность и прозрачность алгоритма, а также возможность наглядной интерпретации кластеров с использованием графика средних значений показателей в кластерах.

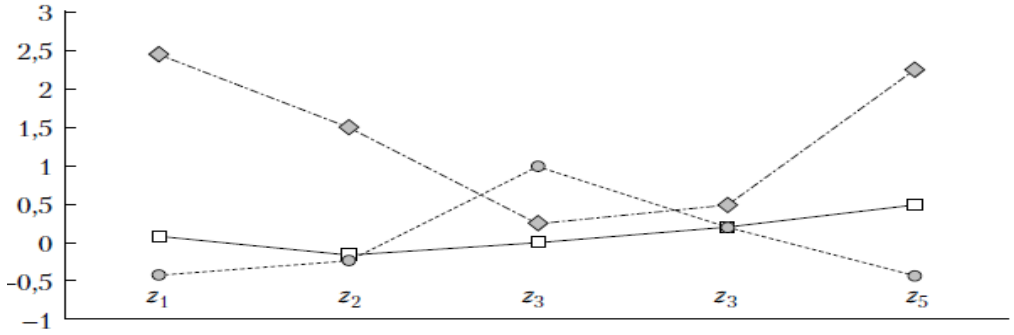

Рис. 1. График средних значений показателей в кластерах

Недостатки алгоритма k-средних состоят в том, что метод не гарантирует достижение глобального минимума суммарного квадратичного отклонения, а только одного из локальных минимумов, кроме этого алгоритм k-средних крайне чувствителен к выбору начальных приближений центров.

Случайный выбор центров может привести к неудовлетворительным результатам кластеризации (для формирования начального приближения лучше выделить k *наиболее удаленных точек выборки).* Для реализации алгоритма число кластеров необходимо знать заранее (т.е. требуется либо экспертная информация, либо проведение разведывательного анализа). Сам алгоритм чувствителен к выбросам, которые могут искажать среднее

значение (возможным решением этой проблемы является использование модификации алгоритма — алгоритм *k*-медианы).

Альтернативой кластерным методам являются иерархические алгоритмы, использующие понятие порога. Пороговые алгоритмы эффективны для исходных совокупностей, у которых слабо выражен цепочный эффект и они естественно распадаются на какое-то количество достаточно отдаленных скоплений точек (кластеров). Общая схема таких алгоритмов отличается от иерархических агломеративных алгоритмов наличием монотонной последовательности порогов C_1,C_2, ...,C_α, (порог C_α соответствует максимальному расстоянию между объектами), которые используются следующим образом:

• на первом шаге попарно объединяются элементы, расстояние между которыми не превышает C_1;

• на втором шаге объединяются элементы (группы элементов), расстояние между которыми не превышает C_2;

…

• на последнем шаге объединяются все *n* элементов в один класс.

Наиболее популярным среди таких алгоритмов является алгоритм типа *FOREL* («формальный элемент»). *FOREL* является примером эвристического дивизимного алгоритма классификации, основанный на идее объединения в один кластер объектов в областях их наибольшего сгущения.

Данный термин был предложен Н. Г. Загоруйко, В. Н. Ёлкиной и другими учеными в 1967 г. в Институте математики СО РАН при решении прикладной задачи в области палеонтологии. Таксоны, получаемые этим алгоритмом, имеют сферическую форму. Количество таксонов зависит от радиуса сфер: чем меньше радиус, тем больше получается таксонов.

Существуют и другие алгоритмы классификации. Так, Р. Боннер описывает метод похожий на алгоритм *FOREL*. В этом методе объект, служащий начальной точкой, выбирается случайно. Все объекты, лежащие на расстоянии от начальной точки не больше *r*, принимаются за первый кластер.

Из оставшихся точек снова случайным образом выбирается объект, и процесс повторяется, как и предшествовавший. В результате все точки будут разбиты на группы.

Л. Хиверинен рассматривает процедуру, аналогичную Боннеру, но в качестве начального объекта кластеризации выбирает не случайную, а так называемую типическую точку. Для определения типических точек он пользуется статистикой потери информации, при чем эти точки лежат на минимальном расстоянии от центра оставшегося множества объектов.

В процедуре, авторами которой являются Ж. Болл и Д. Холл, первоначальные l кластеров формируются случайным отбором l точек, к которым затем присоединяется каждая из оставшихся n — 1 точек — по минимальному расстоянию к той или иной из них. Затем находятся центры кластеров, и два кластера I и J объединяются, если Dij меньше некоторого порогового значения r. Наоборот, если внутригрупповая дисперсия кластера Sx^2 по некоторой переменной x превосходит пороговое значение S^2, то кластер разбивается. Таким образом, дисперсии кластеров S_I^2, получающихся в результате этой процедуры, ограничены. Вместо центра первоначального кластера рассматриваются центры новых образовавшихся кластеров, и процесс продолжается до тех пор, пока не сойдется.

Дж. Мак Куин предлагает метод, аналогичный методу Болла и Холла. Случайным образом отбирается l объектов, которые принимаются в качестве центров кластеризации. Для каждого объекта отыскивается ближайшая точка кластеризации, и если расстояние от выбранного объекта до этой т точки не больше заданного уровня r, то объект приписывают к кластеру найденной точки кластеризации. Если это расстояние больше r, то объект образует новый кластер. После этого вычисляются новые центры кластеров. Если расстояние между центрами двух кластеров меньше другого априорно заданного уровня, то соответствующие кластеры объединяются. Процесс продолжается до разбиения всех n объектов на классы.

Метод Г. Себестьена имеет много общего с предыдущим. Однако по Себестьену объект принадлежит кластеру, если расстояние d до центра кластера меньше r; если же это расстояние больше r ($R>r$), то этот объект образует новую точку кластеризации. Однако если $r<d<R$, то объект выбывает из рассмотрения до следующей итерации. Отметим, что зависимость результатов кластеризации от методов тем сильнее, чем менее явно изучаемая совокупность данных разделяется на группы объектов.

Существенное влияние на характеристики кластерной структуры оказывают, во-первых, набор признаков, по которым осуществляется классификация, во-вторых, тип выбранного алгоритма. Например, иерархические и итеративные методы приводят к образованию различного числа кластеров. При этом сами кластеры различаются и по составу, и по степени близости объектов. Выбор меры сходства также влияет на результат разбиения. Если используются методы с эталонными алгоритмами, например, метод k-средних, то задаваемые начальные условия разбиения в значительной степени определяют конечный результат разбиения.

2.Двухэтапный кластерный анализ

Процедура двухэтапного кластерного анализа представляет собой средство разведочного анализа для выявления естественного разбиения набора данных на группы (или кластеры), которое без ее применения трудно обнаружить.

Алгоритм, используемый этой процедурой, имеет несколько привлекательных особенностей, которые отличают его от традиционных методов кластерного анализа:

- работа с категориальными и непрерывными переменными: предполагая независимость переменных, можно считать, что категориальные и непрерывные переменные имеют совместное полиномиально-нормальное распределение.

- автоматический выбор числа кластеров: сравнивая значения критерия отбора модели для различных кластерных решений, процедура может автоматически определить оптимальное число кластеров.

- масштабируемость.

Формируя дерево свойств кластеров (СК), которое является компактным представлением информации о наблюдениях, двухэтапный алгоритм позволяет анализировать большие файлы данных.

Например, компании производства потребительских товаров и розничной торговли регулярно применяют методы кластерного анализа к данным, описывающим покупательские привычки их клиентов, а также их пол, возраст, уровень доходов и т.д. Эти компании настраивают стратегии маркетинга и развития производства на каждую из групп потребителей, чтобы увеличить продажи и повысить приверженность потребителей маркам товаров.

Рассмотрим основные этапы реализации процедуры двухэтапного кластерного анализа в ППП SPSS:

Этап 1. Задание меры расстояния - определяет, как вычисляется сходство между двумя кластерами:

- log-правдоподобия: мера правдоподобия приписывает переменным вероятностное распределение. Предполагается, что непрерывные переменные имеют нормальное распределение, а категориальные переменные - полиномиальное. Все переменные предполагаются независимыми.

- Евклидова: Евклидова мера является расстоянием "по прямой линии" между двумя кластерами. Она может быть использована, только когда все переменные являются непрерывными.

Этап 2. Выбор числа кластеров в рамках реализации данного алгоритма позволяет задать, как будет определяться число классов.

- определять автоматически: процедура автоматически определит "наилучшее" число классов, используя критерий, заданный в группе *критерий кластеризации*. Дополнительно можно ввести положительное

целое число, задающее максимальное число кластеров, которое должна рассмотреть процедура.

- задать: позволяет зафиксировать число кластеров в решении. Нужно задать положительное целое число.

Этап 3. Информация о количестве непрерывных переменных дает сводную информацию об установках, касающихся стандартизации непрерывных переменных, заданных в диалоговом окне программы.

Этап 4. Критерий кластеризации - задает способ, которым автоматический алгоритм кластеризации определяет число кластеров. Можно задать либо Байесовский информационный критерий (BIC), либо информационный критерий Акаике (AIC).

Этап 5. Данные для двухэтапного кластерного анализа. Данная процедура работает как с непрерывными, так и с категориальными переменными. Наблюдения представляют собой объекты кластеризации, а переменные являются атрибутами, на которых основывается кластеризация.

Этап 6. Порядок наблюдений. Дерево свойств кластеров и окончательное решение могут зависеть от порядка наблюдений. Чтобы минимизировать эффект порядка наблюдений, нужно расположить их в случайном порядке.

Возможно, что исследователь захочет получить несколько различных решений с наблюдениями, упорядоченными случайным образом, чтобы проверить стабильность данного решения. В ситуациях, когда это трудно сделать в силу чрезвычайно больших размеров файлов, можно в качестве альтернативы несколько раз выполнить процедуру с выборкой наблюдений, отсортировывая ее в случайном порядке.

Этап 7. Предположения. Мера расстояния, основанная на правдоподобии, предполагает, что переменные в кластерной модели являются независимыми. Кроме того, предполагается, что каждая непрерывная переменная имеет нормальное (гауссово) распределение, а каждая категориальная переменная - полиномиальное распределение.

Эмпирические исследования показывают, что эта процедура вполне устойчива к нарушениям предположений как о независимости, так и о распределениях, однако следует проверить, насколько эти предположения выполняются. Для проверки независимости двух непрерывных переменных используют процедуру – *парные корреляции*. Для проверки независимости двух категориальных переменных используют процедуру – *таблицы сопряженности*. Для проверки независимости между непрерывной переменной и категориальной переменной используют процедуру – *средние*. Для проверки нормальности непрерывной переменной используют процедуру – *исследовать*. Для проверки того, что категориальная переменная имеет заданное полиномиальное распределение, используют процедуру – *критерий хи-квадрат*.

Сама процедура двухэтапного кластерного анализа может быть запущена через меню: *Анализ* - *Классификация* - *Двухэтапный кластерный анализ*.

Далее необходимо выбрать одну или несколько категориальных или непрерывных переменных. Дополнительно возможно установить критерии, по которым формируются кластеры, выбрать установки для обработки шумов, выделения памяти, стандартизации переменных и ввода кластерной модели, задать запрос вывода средства просмотра моделей и сохранить результаты построения модели в рабочем файле или внешнем XML файле.

Вопросы для самопроверки

1. Расскажите об итерационных алгоритмах классификации и методе *k*-средних.

2. Какие модификации метода к-средних вы знаете?

3. В чем отличие метода к- средних от иерархических процедур кластеризации?

4. В каких случаях применяется двухэтапный кластерный анализ?

5. Перечислите основные этапы реализации алгоритма двухэтапного кластерного анализа в ППП SPSS.

Раздел 3. Многомерная классификация с обучением и модели бинарного выбора

Лекция 8. Деревья решений, основы дискриминантного анализа и модели бинарного выбора

1. Основные понятия и критерии применения деревьев решений

Дерево решений — эффективный инструмент интеллектуального анализа данных и предсказательной аналитики. Он помогает в решении задач по классификации и регрессии. Дерево решений представляет собой иерархическую древовидную структуру, состоящую из правила вида «Если …, то ...». За счет обучающего множества правила генерируются автоматически в процессе обучения.

В отличие от нейронных сетей, деревья как аналитические модели проще, потому что правила генерируются на естественном языке: например, «Если реклама привела 1000 клиентов, то она настроена хорошо». Правила генерируются за счет обобщения множества отдельных наблюдений (обучающих примеров), описывающих предметную область. Поэтому их называют индуктивными правилами, а сам процесс обучения — индукцией деревьев решений. В обучающем множестве для примеров должно быть задано целевое значение, так как деревья решений — модели, создаваемые на основе обучения с учителем.

По типу переменной выделяют два типа деревьев:

- дерево классификации — когда целевая переменная дискретная;
- дерево регрессии — когда целевая переменная непрерывная.

Развитие данного инструмента началось в 1950-х годах. Тогда были предложены основные идеи в области исследований моделирования человеческого поведения с помощью компьютерных систем.

Дальнейшее развитие деревьев решений как самообучающихся моделей для анализа данных связано с Джоном Р. Куинленом (автором алгоритма ID3 и последующих модификаций C4.5 и C5.0) и Лео Брейманом,

предложившим алгоритм CART и метод случайного леса. Модули для построения и исследования деревьев решений входят в состав множества аналитических платформ.

Успешнее всего деревья применяют в следующих областях: банковское дело, промышленность, медицина, молекулярная биология, торговля.

Круг использования деревьев решений постоянно расширяется, они становятся важным инструментом управления бизнес-процессами и поддержки принятия решений. Дерево решений применяют для поддержки процессов принятия управленческих решений, используемых в статистике, анализе данных и машинном обучении. Инструмент помогает решать задачи классификации**,** регрессии (численное предсказание), описания объектов.

Рассмотрим структуру дерева решений более подробно. Дерево решений — метод представления решающих правил в определенной иерархии, включающей в себя элементы двух типов — узлов (node) и листьев (leaf). Узлы включают в себя решающие правила и производят проверку примеров на соответствие выбранного атрибута обучающего множества.

Простой случай: примеры попадают в узел, проходят проверку и разбиваются на два подмножества: 1 — те, которые удовлетворяют установленное правило; 2 — те, которые не удовлетворяют установленное правило. Далее к каждому подмножеству снова применяется правило, процедура повторяется. Это продолжается, пока не будет достигнуто условие остановки алгоритма. Последний узел, когда не осуществляется проверка и разбиение, становится листом. Лист определяет решение для каждого попавшего в него примера. Для дерева классификации — это класс, ассоциируемый с узлом, а для дерева регрессии — соответствующий листу модальный интервал целевой переменной. В листе содержится не правило, а подмножество объектов, удовлетворяющих всем правилам ветви, которая заканчивается этим листом. Пример попадает в лист, если соответствует всем правилам на пути к нему. К каждому листу есть только один путь. Таким

образом, пример может попасть только в один лист, что обеспечивает единственность решения.

Основная задача при построении дерева решений — последовательно и рекурсивно разбить обучающее множество на подмножества с применением решающих правил в узлах. Этот процесс продолжают до того, пока все узлы в конце ветвей не станут листами.

Узел становится листом в двух случаях:

- естественным образом — когда он содержит единственный объект или объект только одного класса;
- после достижения заданного условия остановки алгоритм — например, минимально допустимое число примеров в узле или максимальная глубина дерева.

В основе построения лежат «жадные» алгоритмы, допускающие локально-оптимальные решения на каждом шаге (разбиения в узлах), которые приводят к оптимальному итоговому решению. То есть при выборе одного атрибута и произведении разбиения по нему на подмножества, алгоритм не может вернуться назад и выбрать другой атрибут, даже если это даст лучшее итоговое разбиение. Следовательно, на этапе построения дерева решений нельзя точно утверждать, что удастся добиться оптимального разбиения.

Популярные алгоритмы, используемых для обучения деревьев решений, строятся на базе принципа «разделяй и властвуй». Задают общее множество S, содержащее:

- n примеров, для каждого из которых задана метка класса Ci(i = 1..k);
- m атрибутов Aj(j = 1..m), которые определяют принадлежность объекта к тому или иному классу.

Тогда возможно три случая:

1. Примеры множества S имеют одинаковую метку Ci, следовательно, все обучающие примеры относятся к одному классу.

2. Множество S — пустое множество без примеров.
3. Множество S состоит из обучающих примеров всех классов Ck.

Третий случай применяется в большинстве алгоритмов, используемых для построения деревьев решений. Эта методика формирует дерево сверху вниз, то есть от корневого узла к листьям. Сегодня существует много алгоритмов обучения: ID3, CART, C4.5, C5.0, NewId, ITrule, CHAID, CN2 и другие.

Само построение дерева решений осуществляется в 4 этапа:

1.Выбор атрибута для осуществления разбиения в данном узле.

2.Определение критерия остановки обучения.

3.Выбр метода отсечения ветвей.

4.Оценка точности построенного дерева (извлечение правил).

К преимуществам дерева решений относят, то, что они формируют четкие и понятные правила классификации, способны генерировать правила в областях, где специалисту трудно формализовать свои знания, легко визуализируются, быстро обучаются и прогнозируют, не требуют большого числа параметров модели, работают как с числовыми, так и категориальные признаками. К недостаткам дерева решений относят, то, что они чувствительны к шумам во входных данных, уступает другим методам по качеству классификации, в них возможно переобучение, также поиск оптимального дерева решений может быть достаточно сложным процессом.

2. Основы дискриминантного анализа

Дискриминантный анализ — один из методов многомерного анализа, целью которого является классификация объектов, т.е. отнесение его к одной из известных групп некоторым оптимальным способом (например, разбиение совокупности предприятий на несколько однородных групп по значениям каких-либо показателей производственно-хозяйственной деятельности).

Методы дискриминантного анализа разрабатывались начиная с конца 1950-х гг. такими учеными, как Прасанта Чандра Махаланобис (индийский

экономист и статистик, 1893—1972), Гарольд Хотеллинг (американский экономист и статистик, 1895—1973), Рональд Фишер (английский статистик, биолог-эволюционист, генетик, 1890—1962), и другими учеными.

Отличительным свойством дискриминантного анализа как метода классификации является то, что исследователю заранее известно число групп, на которые нужно разбить рассматриваемую совокупность объектов и их свойства; известно также, что объект заведомо принадлежит к одной из определенных групп (но к какой именно — неизвестно). Например, некий исследователь в области образования исследует какие переменные относят выпускника школы к одной из трех категорий: (1) поступающий в вуз; (2) поступающий в профессиональную школу; (3) отказывающийся от дальнейшего образования или профессиональной подготовки.

С этой целью исследователь может собрать данные о различных переменных, характеризующих учащихся школы. Отметим, что нас здесь интересует только вероятность подачи документов в вуз, в профессиональную школу или же отказ от дальнейшего образования или профессиональной подготовки, т.е., другими словами, мы хотим моделировать вероятность выбора выпускниками школы своего дальнейшего путина основании отнесения их к одной из трех названных категорий.

Для решения данной задачи можно использовать дискриминантный анализ, который позволит выделить переменные, вносящий решающий вклад в выбор учащимися дальнейшего пути при условии наличия обучающих выборок. Медик может регистрировать значения различных переменных, описывающие состояние больного. Для того чтобы выяснить, какие переменные лучше предсказывают вероятность того, что пациент выздоровел полностью (группа 1), частично (группа2) или совсем не выздоровел (группа3) он может использовать дискриминантный анализ (ДА). Биолог может записать различные характеристики сходных типов (групп) цветов, чтобы затем провести анализ дискриминантной функции, наилучшим образом разделяющей типы или группы. Во всех приведенных примерах

исследователь обладает обучающими выборками, относительно которых он знает все необходимые характеристики, которые позволяют ему построить дискриминантные функции (ДФ). Подстановка значений для нового наблюдения (объекта) в ДФ позволяет предсказать вероятность наступления интересующего исследователя события, а также выделить переменные, вносящие наибольший вклад в процесс такого разделения.

В соответствии со свойствами ДА возникают задачи двух типов:

1) описания различий между классами;

2) классификации объектов, не входивших в первоначальную обучающую выборку.

Для решения первой задачи (описания различий между классами) строятся канонические дискриминантные функции, которые позволяют с максимальной эффективностью разделить классы.

Для того чтобы выделить p классов, требуется не более (p - 1) канонических дискриминантных функций. Например, для разделения двух классов достаточно одной функции, для разделения трех классов — двух функций и т.д.

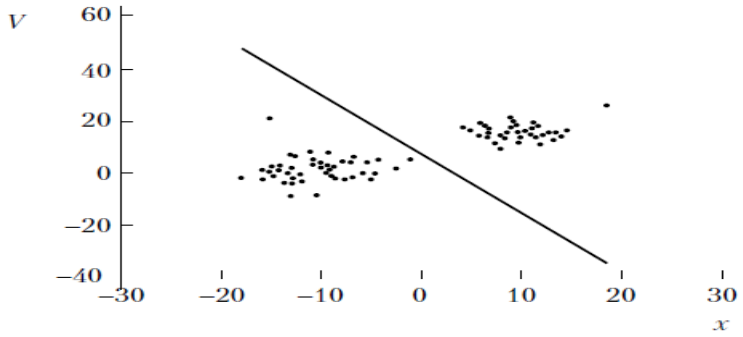

Рис. 1. Разделение совокупности на два класса с помощью одной дискриминантной функции

Канонические дискриминантные функции можно рассматривать как аналог регрессионной модели, построенной с целью классификации объектов. В ДА дискриминантные переменные являются независимыми переменными. Для измерения абсолютного и относительного вкладов дискриминантных переменных в разделение классов используются

нестандартизированные и стандартизированные коэффициенты канонических функций. Чем больше значение коэффициента, тем больший вклад в дискриминацию вносит переменная.

Одним из способов отбора информативных дискриминантных переменных является пошаговый ДА. Логика пошагового ДА такова: вначале определяется та переменная, для которой средние значения в априорно заданных группах наиболее различны. На каждом следующем шаге рассматриваются условные распределения оставшихся переменных и определяется та, для которой средние значение в группах наиболее различны, и т.д.

Процесс завершается, когда ни одна из оставшихся переменных не вносит значимого вклада в различие групп. От выбора критерия отбора дискриминантных переменных зависит результат проведения ДА.

В начале каждого шага ДА происходит проверка всех дискриминантных переменных на соответствие двум условиям: необходимой точности вычисления (толерантности) и превышения заданного уровня различия (на основе использования *F*-распределения и статистик *F*-ввода и *F*-исключения). Статистика *F*-ввода оценивает улучшение разбиения благодаря использованию данной переменной по сравнению с различием, достигнутым с помощью уже отобранных переменных. Статистика *F*-исключения определяет значимость ухудшения различия после удаления переменной из списка уже отобранных переменных. На заключительном шаге статистика *F*-исключения может быть использована для оценки дискриминантных возможностей отобранных переменных. *Переменная с наибольшим значением F-исключения дает наибольший вклад в различие, достигнутое посредством других переменных.* Переменная, имеющая вторую по величине статистику *F*-исключения, является второй по значимости и т.д.

На следующем этапе ДА отобранное подмножество наиболее информативных переменных используется для вычисления ДФ. ДФ является

линейной комбинацией дискриминантных переменных и выглядит как правая часть уравнения множественной регрессии. Таким образом, исследователь получает одну или две ДФ. Эти ДФ называются каноническими дискриминантными функциями.

Отметим, что после получения канонических дискриминантных функций необходимо определить, все ли из этих функций полезны для описания межгрупповых различий. С этой целью используются собственные значения, процент объясненной дисперсии каждой из вычисленных функций, коэффициенты канонической корреляции, тест равенства средних значений канонических дискриминантных функций в группах. Использование ДА для решения практических задач показало, что о полезности ДФ для выделения различий между объектами можно судить по величине коэффициента канонической корреляции. Если значение этого коэффициента невысоко, то данную каноническую функцию не используют.

Самым лучшим показателем информативности отобранных дискриминантных переменных и полезности применения дискриминантной функции для интерпретации межгрупповых различий является процент правильно распознанных объектов с использованием полученных ДФ. Число правильно распознанных новых объектов (как в целом, так и по отдельным группам) свидетельствует о соответствии дискриминантной модели эмпирическим данным.

Для решения второй задачи (классификации объектов, не входивших в первоначальную обучающую выборку) вычисляются расстояния от каждого нового объекта, подлежащего классификации, до геометрического центра (центра тяжести) каждого класса.

Дискриминантный анализ предъявляет строгие требования к исходным данным: в модели должно быть не менее двух классов, в каждом классе — не менее двух объектов из обучающей выборки, число дискриминантных переменных не должно превосходить объем обучающей выборки, дискриминантные переменные должны быть количественными и линейно

независимыми. Для каждого класса требуются приблизительное равенство ковариационных матриц, а также многомерная нормальность распределения.

Отметим, что на практике не всегда удается обеспечить выполнимость этих требований, что зачастую не позволяет корректно использовать аппарат дискриминантного анализа.

Итак, решается задача отнесения каждого из n наблюдений Xi, где i=1,2, …, n, к одному из p классов. В дискриминантном анализе под *классом* понимается генеральная совокупность, описываемая одномодальной функцией плотности распределения $f(x)$ или одномодальным полигоном вероятностей (в случае дискретных признаков X).

Идея вероятностных методов классификации состоит в следующем: наблюдение Xi будет относится к тому классу (той генеральной совокупности), в рамках которого оно выглядит наиболее правдоподобно. Этот принцип может корректироваться с учетом удельных весов классов (обозначим их через π_m) и особенности, так называемой функции потерь $C(l/m)$, которая определяет стоимость потерь от ошибочного отнесения объекта, принадлежащего классу m, к классу с номером l (l, m=1, 2, …, p).

Для реализации подхода необходимо знание функций $f1(x)$, ..., $fp(x)$, задающих закон распределения вероятностей соответствующих классов. В дальнейшем при описании аппарата ДА будем исходить из нормального закона распределения, т.е. $x \sim N_k(\mu_l, \Sigma_l)$, где l =1, 2, ..., p, $p \geq 2$. при p= 1 имеем

$$x \sim N(\mu_l, \sigma_l) \text{ и } f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma_l} \cdot e^{\frac{(x-\mu_l)^2}{2\sigma_l^2}}. \quad (1)$$

Методы классификации следует выбирать исходя из условия *минимизации потерь* или *вероятности неправильной классификации* объектов.

Для формализации данной задачи введем понятие «функция потерь». Обозначим $C(l/m)$ потери, связанные с ошибочным отнесением объекта m-го

класса к классу l (при $l=m$, очевидно, $C(l/m)=0$). Если потери $C(l/m)$ одинаковы для любой пары l и m, т.е. $C(l/m)=C_0=\text{const}>0, \text{при } l\neq m, C(l/m)=0, \text{при } l=m.$

Пусть в процессе классификации число ошибок составило $\upsilon(l/m)$.

Тогда потери, связанные с ошибочным отнесением объектов m-го класса к классу l, составляют $\upsilon(l/m)*C(l/m)$ по всем l, m=1, 2, …, p.

Общие потери Cn (n — число наблюдений) при такой процедуре классификации составят

$$C_n=\sum_{l=1}^{p}\sum_{m=1}^{p}C(l/m)\cdot\nu(l/m). \quad (2)$$

Удельная характеристика потерь C при $n\to\infty$ равна

$$C=\frac{1}{n}C_n=\sum_{l=1}^{p}\sum_{m=1}^{p}C(l/m)\cdot\frac{\nu(l/m)}{n_m}\cdot\frac{n_m}{n}\to\sum_{m=1}^{p}\pi_m\sum_{l=1}^{p}C(l/m)\cdot P(l/m), \quad (3)$$

где $P(l/m)$ — вероятность отнесения объект класса m к классу l;

π_m— вероятность извлечения объекта класса m из общей совокупности объектов, или априорная вероятность (удельный вес) класса m.

Здесь предел понимается в смысле сходимости по вероятности относительных частот к соответствующим вероятностям:

$$\frac{\nu(l/m)}{n_m}\approx P(l/m) \text{ и } \frac{n_m}{n}\approx\pi_m. \quad (4)$$

Средние потери от неправильной классификации объектов m-го класса равны

$$C^{(m)}=\sum_{l=1}^{p}C(l/m)\cdot P(l/m). \quad (5)$$

Средние удельные потери от неправильной классификации всех анализируемых объектов составят

$$C=\sum_{m=1}^{p}\pi_m\cdot C^{(m)}. \quad (6)$$

Минимизация средних удельных потерь C эквивалентна вероятности правильной классификации объектов, равной

$$\sum_{m=1}^{p}\pi_m\cdot P(m/m). \quad (7)$$

Можно показать, что

$$C = \sum_{m=1}^{p} \pi_m \cdot \sum_{l=1}^{p} C(l/m) \cdot P(l/m) = C_0 \sum_{m=1}^{p} \pi_m \sum_{\substack{l=1 \\ l \neq m}}^{p} P(l/m) =$$

$$= C_0 \sum_{m=1}^{p} \pi_m (1 - P(m/m)) = C_0 \left(1 - \sum_{m=1}^{p} \pi_m \cdot P(m/m) \right). \quad (8)$$

При этом учитывалось, что $C(m/m) = 0, \sum_{l=1}^{p} P(l/m) = 1$ для любого m. В этом случае при построении процедур классификации часто говорят не о потерях, а о вероятности неправильной классификации объектов $1 - \sum_{m=1}^{p} \pi_m P(m/m).$ (9)

В дискриминантном анализе используют Лямбду - статистику Уилкса(Wilks'Lambda), которая является мерой достоверности различения классов при помощи данного набора переменных. Это мера *остаточной дискриминативной способности* переменных при учете данного набора канонических функций. Чем меньше Лямбда - статистика Уилкса, тем лучше, так как она измеряет остаточную дискриминацию. Величины Лямбда - статистики Уилкса близкие к нулю говорят о высоком различении (центроиды классов хорошо разделены и сильно отличаются друг от друга по отношению к степени разброса внутри классов).

3. Модели бинарного выбора

До настоящего момента мы рассматривали модели, в которых зависимая переменная была количественной, бинарная переменная (или фиктивная переменная) выступала только в качестве регрессора.

Бывают ситуации, в которых необходимо отвечать не на «количественный» вопрос (например, как зависит объем спроса на товар от его цены), а на «качественный», например, вернет человек вовремя кредит или нет, поступит школьник в университет или нет, купит семья автомобиль или нет. Для ответов на подобные вопросы в качестве зависимой переменной

в регрессионном уравнении используется бинарная переменная и для оценивания таких уравнений применяются logit- и probit-модели.

В общем случае необходимо оценить вероятность наступления события. Так, например, если мы принимаем за единицу факт поступления школьника в университет (y = 1), а за ноль факт непоступления (y = 0), то в качестве зависимой переменной в модели будет вероятность поступления школьника в университет P(y=1). Основной идеей как logit-, так и probit-модели является оценивание вероятности наступления события через функцию, область значений которой лежит в диапазоне [0,1]. Наиболее часто в качестве такой функции используют функцию стандартного нормального распределения и функцию логистического распределения.

Рассмотрим зависимые переменные, принимающие конечное число значений. Как правило, без потери общности можно считать, что это значения 0, 1, …, m. Подобные ситуации возникают в тех случаях, когда значения объясняемой переменной соответствуют выбору решения из набора m + 1 возможных решений.

Наиболее простыми моделями оказываются *модели бинарного выбора*. В этом случае объясняемая переменная принимает всего два значения — 0 и 1. Можно считать, что 1 соответствует положительному решению, а 0 — отрицательному.

Пусть y_i— объясняемая величина, где y_i= 0 или y_i= 1.

Величина y принимает одно из своих возможных значений под воздействием факторов $x_1 \ldots x_k$, которые могут принимать непрерывные значения.

Рассматривается регрессионная модель $y_i = F(x_{1i}, \ldots, x_{ki};\ \beta_1, \ldots, \beta_k) + \varepsilon_i,$ (10)

в которой $\beta_1, \ldots, \beta_k$ - параметры модели, соответствующие регрессорам $x_1, \ldots, x_k$.

Наиболее простой является линейная модель регрессии

$$y_i = \sum_{j=1}^{k} \beta_j x_{ji} + \varepsilon_i . \qquad (11)$$

Так как полагается, что $M(\varepsilon_i)=0$, то$M(y_i)=F(x, \beta)$. В то же время $M(y_i) = 0 \cdot P(y_i = 0) + 1 \cdot P(y_i = 1) = P(y_i = 1)$, откуда следует, что

$$P(y_i=1)=F(x,\beta). \quad (12)$$

Такое уравнение называется уравнением модели бинарного выбора.

В частности, для линейной модели имеем:

$$P(y_i = 1) = \sum_{j=1}^{k} \beta_j x_{ji}. \quad (13)$$

Использование такого уравнения сопряжено со значительными трудностями. В первую очередь, это связано с тем, что величины $\sum_{j=1}^{n} \hat{\beta}_i x_{ji}$, получаемые оцениванием параметров β, могут не попадать в промежуток [0, 1].

Кроме этого, неправомочно предположение о том, что ошибки ε_i имеют нормальное распределение. Линейную модель можно использовать в ряде случаев при большом числе наблюдений и достаточно точной спецификации модели. В основном же она используется лишь как грубый инструмент первичной обработки данных.

Перечисленные трудности легко преодолимы, если в качестве функции F(x,β) выбирается функция распределения некоторой случайной величины.

Подобный выбор становится естественным, если предположить, что значение дискретной переменной *y* скачкообразно изменяется в зависимости от значений некоторой — как правило, ненаблюдаемой — количественной переменной y*. Наиболее простой случай соответствует наличию порогового значения, т.е. $y_i = \begin{cases} 1, & \text{если } y_i^* \geq 0; \\ 0, & \text{если } y_i^* < 0. \end{cases}$

Пусть количественная переменная y_i* удовлетворяет регрессионному уравнению $y_i^* = \sum_{j=1}^{k} \tilde{\beta}_j x_{ij} + \varepsilon_i,$ (14)

причем константа включена в число регрессоров, а ошибки ε_i независимы и одинаково распределены с нулевым средним и дисперсией σ^2.

Пусть функция F (t) — функция распределения случайной величины ε_i/σ.

$$P(y_i = 1) = P(y_i^* \geq 0) = P\left(\sum_{j=1}^{k} x_{ij}\tilde{\beta}_j + \varepsilon_i \geq 0\right) =$$

$$= P\left(\varepsilon_i \geq -\sum_{j=1}^{k} x_{ij}\tilde{\beta}_j\right) = P\left(\varepsilon_i \leq \sum_{j=1}^{k} x_{ij}\tilde{\beta}_j\right) = F\left(\sum_{j=1}^{k} x_{ij}\frac{\tilde{\beta}_j}{\sigma}\right).$$

Тогда

Обозначив $\beta = \frac{\tilde{\beta}}{\sigma}$, получим уравнение в такой форме:

$$P(y_i = 1) = F\left(\sum_{j=1}^{k} \beta_j x_{ij}\right). \qquad (15)$$

Уравнение (15) представляет собой стандартную форму модели бинарного выбора. Наиболее естественно в качестве функции F (t) выбрать функцию стандартного нормального распределения. Соответствующая дискретная модель (15) в этом случае называется *probit-моделью.*

Также используется функция логистического распределения

$$F(t) = \Lambda(t) = \frac{e^t}{1+e^t} \qquad (16)$$

и соответствующая модель при таком выборе называется *logit—моделью*.

Выбор функции (16) объясняется тем, что ее вид существенно проще, чем у функции стандартного нормального распределения. При этом для небольших (по модулю) t эти функции достаточно близки друг к другу, и качественные выводы, сделанные по probit- и logit-моделям, в основном совпадают. В то же время для больших значений регрессоров возможны и значимые различия.

Важно отметить то обстоятельство, что модель (15) нелинейная. Это приводит к тому, что предельный эффект каждого фактора x_j (производная $\frac{\partial P(y=1)}{\partial x}$) является переменным и зависящим от значений всех остальных факторов. Так что для получения представления о «среднем» предельном эффекте следует вычислить производные $\frac{\partial P(y=1)}{\partial x_i} = F'\left(\sum_{j=1}^{k} \beta_j x_{ij}\right)\beta_i$ для средних по выборке значений независимых переменных *x*.

Для оценивания модели используется метод максимального правдоподобия. Если наблюдения y_1, ..., y_n независимы, то функция правдоподобия имеет вид:

$$L = \prod_{y_i=0}\left(1 - F\left(\sum_{j=1}^{k}\beta_j x_{ij}\right)\right) \prod_{y_i=1} F\sum_{j=1}^{k}\beta_j x_{ij} =$$

$$= \prod_i F\left(\sum_{j=1}^{k}\beta_j x_{ij}\right)^{y_i}\left(1 - F\sum_{j=1}^{k}\beta_j x_{ij}\right)^{1-y_i}.$$

Тогда $$\ln L = \sum_{i=1}^{n}\left[y_i \ln F\left(\sum_{j=1}^{k}\beta_j x_{ij}\right) + (1-y_i)\ln\left(1 - F\sum_{j=1}^{k}\beta_j x_{ij}\right)\right].$$

Условия правдоподобия имеют вид: $\dfrac{\partial \ln L}{\partial \beta} = 0.$

Можно доказать, что для probit- и logit-моделей функции правдоподобия *lnL* вогнуты, и следовательно, уравнения дают оценки максимального правдоподобия.

Значимость отдельных коэффициентов при регрессорах в бинарных моделях (как logit, так и probit) можно проверять, ориентируясь либо на *p*-значения, либо сравнивая расчетные значения статистики с критическими значениями. Для проверки гипотез о коэффициентах модели используется тест отношения правдоподобия, в частности он позволяет проверять гипотезу о совместной незначимости коэффициентов модели.

Для оценки качества модели используется аналог ρ^2 – это $\rho^2_{McFadden}$. Коэффициент детерминации *McFadden* вычисляется как $1-\frac{\mathcal{L}_{est}}{\mathcal{L}_{const}}$, где – $\mathcal{L}_{est}$ это значение функции правдоподобия для оцениваемой модели, а $\mathcal{L}_{const}$ – значение функции правдоподобия для модели, которая включает в себя только константу. Если максимум правдоподобия для полной модели близок к значению максимума правдоподобия для модели только с константой (то есть их частное близко к единице), то объясняющая способность модели $\rho^2_{McFadden}$ будет близка к нулю.

Также достаточно часто для оценки качества модели бинарного выбора используют число корректно предсказанных значений. Этот показатель не

всегда является приемлемым, так как в случае, если в наблюдаемых значениях зависимой переменной имеется смещение в сторону одного из вариантов (допустим, единиц существенно больше в наблюдениях, чем нулей), то даже если модель не точна и в качестве расчетных значений зависимой переменной мы будем получать только единицы (как результат округления расчетной вероятности в ту или иную сторону по правилам округления), процент корректно предсказанных значений будет в любом случае велик.

Теперь рассмотрим, как интерпретировать оценки для коэффициентов регрессионного уравнения. Поскольку наблюдается нелинейная зависимость, то очевидно, что эффект от изменения одного регрессора при прочих равных условиях не будет постоянным, а будет изменяться. В общем виде эффект от изменения *i*-го регрессора можно выписать следующим образом:

$$\frac{\partial P(y=1)}{\partial x_i} = F'\left(\sum_{j=1}^{k}\beta_j x_{ij}\right)\beta_i ,$$

где $F'\left(\sum_{j=1}^{k}\beta_j x_{ij}\right)$ – первая производная функции,

β_i – значение коэффициента при *i*-м регрессоре.

Для того, чтобы оценить эффект от изменения одного регрессора, необходимо также знать значения остальных регрессоров. На базовом уровне считают среднее значение для каждого из регрессоров и при заданном среднем вычисляют предельный эффект от изменения отдельно взятого регрессора, но для более точного прогноза стоит считать предельный эффект для конкретных интересующих значений регрессоров (не средних значений).

Для logit-моделей помимо оценки предельного эффекта от изменения регрессоров также существует и другая интерпретация коэффициентов, а именно – число e, возведенное в степень, равную коэффициенту перед регрессором, показывает изменение шансов на реализацию события, при этом число $\exp(\beta_0 + x\beta)$ показывает исходные шансы на реализацию события при заданных условиях.

Вопросы для самопроверки

1. Для чего на практике применяют деревья решений?

2. В чем отличие деревьев классификации и регрессии?

3. Для чего используют дискриминантный анализ?

4. В чем отличие дискриминантного анализа от других методов многомерного статистического анализа?

5. Как можно судить о качестве полученной классификации?

6. Для чего используются модели бинарного выбора?

7. В чем состоит отличие logit-, probit- моделей?

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

Основная литература

1. Тарасов И. Е. Статистический анализ данных в информационных системах [Электронный ресурс]:учебно-методическое пособие. - Москва: РТУ МИРЭА, 2020. - 96 с. – Режим доступа: https://e.lanbook.com/book/163854
2. Каган Е. С. Прикладной статистический анализ данных [Электронный ресурс]:учебное пособие. - Кемерово: КемГУ, 2018. - 235 с. – Режим доступа: https://e.lanbook.com/book/134318
3. Мхитарян В. С., Архипова М. Ю., Дуброва Т. А., Миронкина Ю. Н., Сиротин В. П. Анализ данных [Электронный ресурс]:Учебник для вузов. - Москва: Юрайт, 2022. - 490 с – Режим доступа: URL: https://urait.ru/bcode/489100

Дополнительная литература

1. Халафян А. А. Statistica 6. Статистический анализ данных:Учеб. пособие для вузов. - М.: Бином, 2011. - 522 с.
2. Наследов А. SPSS 19: профессиональный статистический анализ данных:. - СПб.: Питер, 2011. - 399 с.
3. Берк К., Кэйри П. Анализ данных с помощью MicrosoftExcel:Адаптировано для OfficeXP. - М.: Изд. дом "Вильямс", 2005. - 555 с.
4. Тюрин Ю. Н., Макаров А. А. Статистический анализ данных на компьютере:. - М.: ИНФРА-М, 1998. - 528 с.
5. Миркин Б. Г. Введение в анализ данных:учебники практикум. - М.: Юрайт, 174 с. – Режим доступа: URL: https://urait.ru/bcode/469306
6. Боровиков В. П. Популярное введение в современный анализ данных в системе STATISTICA. Методология и технология современного анализа данных [Электронный ресурс]:. - Москва: Горячая линия-Телеком, 2018. - 288
7. Терехина А. Ю. Анализ данных методами многомерного шкалирования:. - М.: Наука, 1986. - 168 с.
8. Козлов А. Ю., Мхитарян В. С., Шишов В. Ф., Мхитарян В. С. Статистический анализ данных в MSEXCEL:Рек. УМО в кач. учеб. пособия для вузов. - М.: ИНФРА-М, 2014. - 320 с.

РЕКОМЕНДУЕМЫЙ ПЕРЕЧЕНЬ СОВРЕМЕННЫХ ПРОФЕССИОНАЛЬНЫХ БАЗ ДАННЫХ И ИНФОРМАЦИОННЫХ СПРАВОЧНЫХ СИСТЕМ

- Сайт Федеральной службы государственной статистики - https://rosstat.gov.ru/
- Аналитический центр при правительстве Российской Федерации -https://ac.gov.ru/
- Информационный портал Российского научного фонда - http://www.rscf.ru
- Научная электронная библиотека - http://www.elibrary.ru
- Министерство науки и высшего образования Российской Федерации -
- https://www.minobrnauki.gov.ru
  База данных Web of Science - http://www.webofknowledge.com
- Федеральное государственное бюджетное учреждение «Федеральный институт
- промышленной собственности» - http://www.fips.ru/
- Статистические сборники НИУ ВШЭ - https://www.hse.ru/primarydata/
- UNCTAD - https://unctad.org/
- Евростат - https://ec.europa.eu/eurostat
- World Trade Organization. International Trade Statistics - https://www.wto.org/