Статистические методы анализа данных

Статистические методы анализа данных

Для решения задач, связанных с анализом данных (выявление скрытых взаимосвязей внутри массивов данных) при наличии случайных и непредсказуемых воздействий, математиками и другими исследователями за последние двести лет был выработан мощный и гибкий арсенал методов, называемых в совокупности статистическими методами анализа данных. За это время накоплен большой опыт успешного применения этих методов в разных сферах человеческой деятельности, от экономики до космических исследований. И при определенных условиях эти методы позволяют получать оптимальные решения.

География использования статистических методов анализа данных

Методы прикладной статистики используются в научных и технических исследованиях, работах по управлению (менеджменту), в медицине, биологии, социологии, психологии, истории, геологии, экологии и т.д. Как правило объекты человеческой деятельности характеризуются большим количеством различных свойств и связей между ними, что определяет многомерность данных. Многомерные данные можно представить в виде матрицы "объект-признак", строки которой соотнесены с анализируемыми объектами или номером опыта, а столбцы – со значениями изучаемых признаков.

Первоисточник многомерных данных определяется самой предметной областью или изучаемым объектом и целью проводимого исследования.

Примеры:

- Лабораторные журналы (записи экспериментов) – научные исследования, по установлению, например, зависимостей; Опросные листы – сегментация рынка, выявление предпочтений;
- Биометрические данные – машинное обучение для построения моделей идентификации личности;
- Медицинские изображения – поиск аномалий, классификация паталогий;
- Цифровые изображения – сегментация, поиск шаблонов и т.д.;
- Электрокардиограммы – обнаружение аномалий с классификацией;
- Звуковой ряд – распознование речи для идентификации и речевого управления, например, интерфейс с ПК;
- Финансово-экономические показатели – прогнозирование банкротств;
- Временные ряды показателей фондового рынка – выбор стратегии купли-продажи;
- Транзакции к БД – оптимизация запросов, построение распределенной БД;
- Текстовые сообщения – определение смыслового содержания, автоматическая рубрикация;
- Хранилища Данных (ХД) – поиск скрытых закономерностей методами Data Mining;
- Интернет – Web Mining (например, фильтация спама, определение закономерностей поведения посетителей, Semantic Web).

Примеры ХД.

База клиентов сотовой компании с полной историей вызовов и соединений. Задачи: разработка новых тарифных планов, выявление причин потери клиентов, выявление групп клиентов и их предпочтений;

Биоинформатика – выявление закономерностей в генетических текстах микроорганизмов, человека, например, для разработки лекарств нового поколения. Эффект – сокращение затрат вдвое.

Банковское дело – анализ кредитных историй с целью получения профилей надежных и ненадежных заемщиков, установления лимитов кредита, процентов и срока возврата для разных групп риска;

СТАТИСТИЧЕСКИЕ МЕТОДЫ (методы, основанные на использовании математической статистики) являются эффективным инструментом сбора и анализа информации. Применение этих методов не требует больших затрат и позволяет с заданной степенью точности и достоверностью судить о состоянии исследуемых явлений (объектов, процессов), прогнозировать и регулировать проблемы на всех этапах их жизненного цикла и на основе этого вырабатывать оптимальные управленческие решения.

Условно все методы можно классифицировать по признаку общности на три основные группы:

- графические методы,
- методы анализа статистических совокупностей
- экономико-математические методы.

Предложенная классификация не является ни универсальной, ни исчерпывающей, но она дает наглядное представление о разнообразии статистических методов и о тех потенциальных возможностях, которыми они располагают по части их использования при анализе данных.

Графические методы основаны на применении графических средств анализа статистических данных. В эту группу могут быть включены такие методы, как контрольный листок, диаграмма Парето, схема Исикавы, гистограмма, диаграмма разброса, расслоение, контрольная карта, график временного ряда и др. Данные методы не требуют сложных вычислений, могут использоваться как самостоятельно, так и в комплексе с другими методами. Овладение ими не представляет особого труда не только для инженерно-технических работников, но и для специалистов низшего звена. Вместе с тем это весьма эффективные методы. Недаром они находят самое широкое применение в промышленности, особенно в работе групп качества.

Методы анализа статистических совокупностей служат для исследования информации, когда изменение анализируемого параметра носит случайный характер. Основными методами, включаемыми в данную группу, являются: регрессивный, дисперсионный и факторный виды анализа, метод сравнения средних, метод сравнения дисперсий и др. Эти методы позволяют установить зависимость изучаемых явлений от случайных факторов как качественную (дисперсионный анализ), так и количественную (корреляционный анализ); исследовать связи между случайными и неслучайными величинами (регрессивный анализ); выявить роль отдельных факторов в изменении анализируемого параметра (факторный анализ) и т.д.

Экономико-математические методы представляют собой сочетание экономических, математических и кибернетических методов. Центральным понятием методов этой группы является оптимизация, т. е. процесс нахождения наилучшего варианта из множества возможных

с учетом принятого критерия (критерия оптимальности). Строго говоря, экономико-математические методы не являются чисто статистическими, но они широко используют аппарат математической статистики, что дает основание включить их в рассматриваемую классификацию статистических методов. Для целей, связанных с обеспечением качества, из достаточно обширной группы экономико-математических методов следует выделить в первую очередь следующие: математическое программирование (линейное, нелинейное, динамическое); планирование эксперимента; имитационное моделирование: теория игр; теория массового обслуживания; теория расписаний; функционально-стоимостной анализ и др.

АНАЛИЗ ДАННЫХ с помощью статистических методов может быть выполнен в несколько этапов:

	Этапы анализа данных	Статистические методы исследования
1.	Описание данных	Описательная статистика, определение необходимого объема выборки.
2.	Изучение сходств и различий	Статистические критерии: Крамера-Уэлча, Вилкоксона-Манна-Уитни, хи–квадрат, Фишера и др.
3.	Исследование зависимостей	Корреляционный анализ, дисперсионный анализ, регрессионный анализ.
4.	Снижение размерности	Факторный анализ, метод главных компонент.
5.	Классификация и прогноз	Дискриминантный анализ, кластерный анализ, группировка.

Рассмотрим более подробнее, каждый из этапов:

1. Описание данных. В практических задачах обычно имеется совокупность наблюдений (десятки, сотни, а то и тысячи результатов измерений индивидуальных характеристик), в связи с этим возникает задача компактного описания имеющихся данных.

Для этого используют методы описательной статистики - описания результатов с помощью различных агрегированных показателей графиков. Кроме того, некоторые показатели описательной статистики используются и в других статистических методах.

Для результатов измерений в шкале отношений показатели описательной статистики можно разбить на несколько групп.

Показатели положения – описывают положение экспериментальных данных на числовой оси. Примеры таких данных – максимальный и минимальный элементы выборки, среднее значение, медиана, мода и др.

Показатели разброса – описывают степень разброса данных относительно своего центра (среднего значения). К ним относятся: выборочная дисперсия, разность между минимальным и максимальными элементами (размах, интеграл) выборки и др.

Показатели асимметрии (положение медианы относительно среднего) и др.

Гистограмма и др.

Данные показатели используются для наглядного представления и первичного (визуального) анализа результатов измерений характеристик экспериментальной и контрольной групп.

2. Изучение сходств и различий (сравнение двух выборок) – задача заключается в установлении совпадений или различий характеристик двух имеющихся выборок.

Типовой задачей анализа данных является задача установления совпадений или различий характеристик экспериментальной и контрольной групп. Для этого формулируются статистические гипотезы: гипотеза об отсутствии различий (так называемая нулевая гипотеза) и гипотеза о значимости различий (так называемая альтернативная гипотеза).

Для принятия решения о том, какую из гипотез (нулевую или альтернативную) следует принять, используют решающие правила – статистические критерии. То есть на основании информации о результатах наблюдений (характеристиках членов экспериментальной и контрольной групп) вычисляется число, называемое эмпирическим значением критерия. Это число сравнивается с известным (например, заданным таблично) эталонным числом, называемым критическим значением критерия.

Критические значения приводятся, как правило, для нескольких уровней значимости. Уровнем значимости называется вероятность ошибки, заключающейся в отклонении (не принятии) нулевой гипотезы, когда она верна, то есть вероятность того, что различия сочтены существенными, а они на самом деле случайны. Обычно используют уровни значимости 0,05; 0,01; 0,001.

Если полученное исследователем эмпирическое значение критерия а оказывается меньше или равно критическому, то принимается нулевая гипотеза - считается, что на заданном уровне значимости (то есть при том значении критического показателя, для которого рассчитано критическое значение критерия) характеристики экспериментальных и контрольных групп совпадают. В противном случае, если эмпирическое значение критерия оказывается строго больше критического, то нулевая гипотеза отвергается и принимается альтернативная гипотеза – характеристики экспериментальной и контрольной групп считаются различными с достоверностью различий (1-а). Например, если а = 0,05 и принята альтернативная гипотеза, то достоверность различий равна 0,95 или 95%. То есть достоверность различия характеристик – это дополнение до единицы уровня значимости при проверке гипотезы о совпадении характеристик двух независимых выборок. Другими словами, чем меньше эмпирическое значение критерия (чем левее оно находится от критического значения), тем больше степень совпадения характеристик сравниваемых объектов. И наоборот, чем больше эмпирическое значение критерия (чем правее оно находится от критического значения), тем сильнее различаются характеристики сравниваемых объектов.

Статистические критерии различий. Все критерии различий условно подразделены на две группы: параметрические и непараметрические критерии.

Критерий различия называют параметрическим, если он основан на конкретном типе распределения генеральной совокупности (как правило, нормальном) или использует параметры этой совокупности (средние, дисперсии и т.д.).

Критерий различия называют непараметрическим, если он не базируется на предположении о типе распределения генеральной совокупности и не использует параметры этой совокупности. Поэтому для непараметрических критериев предлагается также использовать такой термин как «критерий, свободный от распределения».

При нормальном распределении генеральной совокупности параметрические критерии обладают большей мощностью по сравнению с непараметрическими. Иными словами, они спо-собны с большей достоверностью отвергать нулевую гипотезу, если последняя неверна. По этой причине в тех случаях, когда выборки взяты из нормально распределенных генеральных со-вокупностей, следует отдавать предпочтение параметрическим критериям.

Важно отметить, что:

- непараметрические критерии выявляют значимые различия и в том случае, если распределение близко к нормальному;
- при вычислениях вручную непараметрические критерии являются значительно менее трудоемкими, чем параметрические.

Для того, чтобы выяснить, являются ли совпадения или различия случайными, используются статистические методы, которые позволяют на основании данных, полученных в результате эксперимента, принять обоснованное решение о совпадениях или различиях.

Общий алгоритм использования статистических критериев прост: до начала и после окончания эксперимента на основании информации о результатах наблюдений (характеристиках членов экспериментальной и контрольной группы) вычисляется эмпирическое значение критерия (алгоритм выбора статистического критерия и формулы для вычислений приведены ниже). Это число сравнивается с известным (табличным) числом – критическим значением критерия (критические значения для всех рекомендуемых нами критериев приведены ниже). Если эмпирическое значение критерия попадает в зону незначимости, то можно утверждать, что «характеристики экспериментальной и контрольной групп совпадают с уровнем значимости 0,05 по статистическому критерию (Крамера-Уэлча, Вилкоксона-Манна-Уитни, хи-квадрат, Фишера).

В противном случае (если эмпирическое значение критерия оказывается вне зоны незначимости), можно утверждать, что "достоверность различий характеристик экспериментальной и контрольной групп по статистическому критерию … равна 95%".

Следовательно, если характеристики экспериментальной и контрольной групп до начала эксперимента совпадают с уровнем значимости 0,05, и, одновременно с этим, достоверность различий характеристик экспериментальной и контрольной групп после эксперимента равна 95%, то можно сделать вывод, что "применение предлагаемого педагогического воздействия(например, новой методики обучения) приводит к статистически значимым (на уровне 95% по критерию …) отличиям результатов".

Критерий Розенбаума

Назначение критерия. Критерий используется для оценки различий между двумя выборками по уровню какого-либо признака, количественно измеренного.

Описание критерия. Это очень простой непараметрический критерий, который позволяет быстро оценить различия между двумя выборками по какому-либо признаку. Однако если критерий *Q* не выявляет достоверных различий, это еще не означает, что их действительно нет.

В этом случае стоит применить критерий φ^* Фишера. Если же Q-критерий выявляет достоверные различия между выборками с уровнем значимости $p<0,01$, можно ограничиться только им и избежать трудностей применения других критериев.

Критерий применяется в тех случаях, когда данные представлены по крайней мере в порядковой шкале. Признак должен варьировать в каком-то диапазоне значений, иначе сопоставления с помощью Q -критерия просто невозможны. Например, если у нас только 3 значения признака, 1, 2 и 3, - нам очень трудно будет установить различия. Метод Розенбаума требует, следовательно, достаточно тонко измеренных признаков.

Применение критерия начинаем с того, что упорядочиваем значения признака в обеих выборках по нарастанию (или убыванию) признака. Лучше всего, если данные каждого испытуемого представлены на отдельной карточке. Тогда ничего не стоит упорядочить два ряда значений по интересующему нас признаку, раскладывая карточки на столе. Так мы сразу увидим, совпадают ли диапазоны значений, и если нет, то насколько один ряд значений "выше" *(S1)*, а второй - "ниже" *(S2)*. Для того, чтобы не запутаться, в этом и во многих других критериях рекомендуется первым рядом (выборкой, группой) считать тот ряд, где значения выше, а вторым рядом - тот, где значения ниже.

Мощность критерия не очень велика. В том случае, если он не выявляет различий, можно обратиться к другим статистическим критериям, например, к U-критерию Манна-Уитни или критерию φ* Фишера.

Данные для применения Q-критерия Розенбаума должны быть представлены хотя бы в порядковой шкале. Признак должен измеряться в значительном диапазоне значений (чем более значительном – тем лучше).

Ограничения применимости критерия

1. В каждой из выборок должно быть не менее 11 значений признака.
2. Объемы выборок должны примерно совпадать.
    1. Если объемы выборок меньше 50, то абсолютная величина разности n_1 (количество единиц в первой выборке) и n_2 (количество единиц во второй выборке) не должна быть больше 10.
    2. Если объемы выборок между 50 и 100, то абсолютная величина разности n_1 и n_2 не должна быть больше 20;
    3. Если объемы выборок больше 100, то допускается, чтобы одна из выборок превышала другую не более чем в 1,5 – 2 раза.
3. Диапазоны значений признака в двух выборках не должны совпадать между собой.

Использование критерия

Для применения Q-критерия Розенбаума нужно произвести следующие операции.

1. Упорядочить значения отдельно в каждой выборке по степени возрастания признака; принять за первую выборку ту, значения признака в которой предположительно выше, а за вторую – ту, где значения признака предположительно ниже.
2. Определить максимальное значение признака во второй выборке и подсчитать количество значений признака в первой выборке, которые больше его (S_1).
3. Определить минимальное значение признака в первой выборке и подсчитать количество значений признака во второй выборке, которые меньше его (S_2).
4. Рассчитать значение критерия $Q = S_1 + S_1$.
5. По таблице определить критические значения критерия для данных n_1 и n_2 . Если полученное значение Q превышает табличное или равно ему, то признается наличие существенного различия между уровнем признака в рассматриваемых выборках (принимается альтернативная гипотеза). Если же полученное значение Q меньше табличного, принимается нулевая гипотеза.

Таблица критических значений

Различия между двумя выборками достоверны с вероятностью 95% при p=0,05 и с вероятностью 99% при p=0,01. Для выборок, в которых больше чем 26 элементов, критические значения Q принимаются равными 8 (при p=0,05) и 10 (при p=0,01).

n	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
p=0,05																
11	6															
12	6	6														
13	6	6	6													
14	7	7	6	6												
15	7	7	6	6	6											
16	8	7	7	7	6	6										
17	7	7	7	7	7	7	7									
18	7	7	7	7	7	7	7	7								
19	7	7	7	7	7	7	7	7	7							
20	7	7	7	7	7	7	7	7	7	7						
21	8	7	7	7	7	7	7	7	7	7	7					
22	8	7	7	7	7	7	7	7	7	7	7	7				
23	8	8	7	7	7	7	7	7	7	7	7	7	7			
24	8	8	8	8	8	8	8	8	8	8	7	7	7	7		
25	8	8	8	8	8	8	8	8	8	7	7	7	7	7	7	
26	8	8	8	8	8	8	8	8	8	8	7	7	7	7	7	7

n	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
p=0,01																
11	9															
12	9	9														
13	9	9	9													
14	9	9	9	9												
15	9	9	9	9	9											
16	9	9	9	9	9	9										
17	10	9	9	9	9	9	9									
18	10	10	9	9	9	9	9	9								
19	10	10	10	9	9	9	9	9	9							
20	10	10	10	10	9	9	9	9	9	9						
21	11	10	10	10	9	9	9	9	9	9	9					
22	11	11	10	10	10	9	9	9	9	9	9	9				
23	11	11	10	10	10	10	9	9	9	9	9	9	9			
24	12	11	11	10	10	10	10	9	9	9	9	9	9	9		
25	12	11	11	10	10	10	10	10	9	9	9	9	9	9	9	
26	12	12	11	11	10	10	10	10	10	9	9	9	9	9	9	9

Критерий U Вилкоксона-Манна-Уитни

Назначение критерия. Статистический критерий, используемый для оценки различий между двумя независимыми выборками по уровню какого-либо признака, измеренного количественно. Позволяет выявлять различия в значении параметра между малыми выборками.

Описание критерия.

Простой непараметрический критерий. Мощность критерия выше, чем у *Q*-критерия Розенбаума.

Этот метод определяет, достаточно ли мала зона перекрещивающихся значений между двумя рядами (ранжированным рядом значений параметра в первой выборке и таким же во второй выборке). Чем меньше значение критерия, тем вероятнее, что различия между значениями параметра в выборках достоверны.

Ограничения применимости критерия

1. В каждой из выборок должно быть не менее 3 значений признака. Допускается, чтобы в одной выборке было два значения, но во второй тогда не менее пяти.
2. В выборочных данных не должно быть совпадающих значений (все числа — разные) или таких совпадений должно быть очень мало (до 10)

Использование критерия

Для применения U-критерия Манна — Уитни нужно произвести следующие операции.

1. Составить единый ранжированный ряд из обеих сопоставляемых выборок, расставив их элементы по степени нарастания признака и приписав меньшему значению меньший ранг (при наличии повторяющихся элементов в выборке использовать средний ранг). Общее количество рангов получится равным $N=n_1+n_2$, *где* n_1 — количество элементов в первой выборке, а — n_2 количество элементов во второй выборке.
2. Разделить единый ранжированный ряд на два, состоящих соответственно из единиц первой и второй выборок. Подсчитать отдельно сумму рангов, пришедшихся на долю элементов первой выборки R_1, и отдельно — на долю элементов второй выборки R_2, затем вычислить:

$$U_1 = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - R_1,$$
$$U_2 = n_1 \cdot n_2 + \frac{n_2 \cdot (n_2 + 1)}{2} - R_2,$$

если всё вычислено верно, то

$$U_1 + U_2 = n_1 \cdot n_2.,$$

3. Определить значение U-статистики Манна-Уитни по формуле $U=min\ \{U_1,U_2\}$.
4. По таблице для избранного **уровня статистической значимости** определить критическое значение критерия для данных n_1 и n_2. Если полученное значение *U* меньше табличного или равно ему, то признается наличие существенного различия между уровнем признака в рассматриваемых выборках (принимается альтернативная гипотеза). Если же полученное значение *U* больше табличного, принимается нулевая гипотеза. Достоверность различий тем выше, чем меньше значение *U*.

5. При справедливости нулевой гипотезы критерий имеет математическое ожидание $M(U)=n_1n_2/2$ и дисперсию $D(U)=n_1n_2(n_1+n_2+1)/12$ и при достаточно большом объёме выборочных данных ($n_1>19, n_2>19$) распределён практически нормально.

Критерий хи-квадрат

Назначение критерия.

Критерий хи-квадрат — любая статистическая проверка гипотезы, в которой выборочное распределение критерия имеет распределение хи-квадрат при условии верности нулевой гипотезы. Считается, что критерий хи-квадрат — это критерий, который асимптотически верен, то есть, выборочное распределение можно сделать как угодно близким к распределению хи-квадрат путём увеличения размера выборки.

В случае, когда распределение статистического критерия является в точности распределением хи-квадрат, критерий хи-квадрат является точным для конкретного значения дисперсии нормально распределённой совокупности на основе выборочной дисперсии. Такие критерии редко применяются на практике, поскольку величина дисперсии распределения обычно неизвестна.

Критерий χ^2 - статистический критерий для проверки гипотезы H_0, что наблюдаемая случайная величина подчиняется некому теоретическому закону распределения.

Определение

Пусть дана случайная величина X .

Гипотеза H_0: с. в. X подчиняется закону распределения $F(x)$.

Для проверки гипотезы рассмотрим выборку, состоящую из n независимых наблюдений над с.в. X: $X^n=(x_1,\cdots x_n),\ x_i\in[a,b],\ \forall i=1\cdots n$.

По выборке построим эмпирическое распределение $F^*(x)$ с.в X.

Сравнение эмпирического $F^*(x)$ и теоретического распределения $F(x)$ (предполагаемого в гипотезе) производится с помощью специально подобранной функции — критерия согласия. Рассмотрим критерий согласия Пирсона (критерий χ^2):

Гипотеза H_0^*: Xn порождается функцией $F^*(x)$.

Разделим [a,b] на k непересекающихся интервалов $(a_i,b_i],\ i=1\cdots k$;

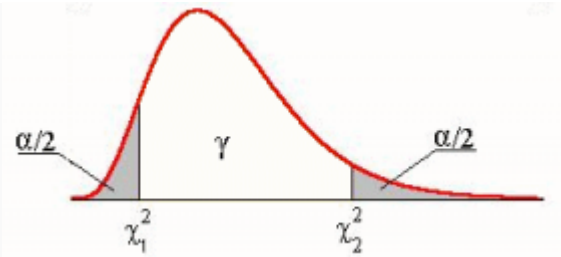
Пусть n_j - количество наблюдений в j-м интервале: $n_j=\sum_{i=1}^{n}\left[a_j<x_i\le b_j\right]$;

$p_j=F(b_j)-F(a_j)$ - вероятность попадания наблюдения в j-ый интервал при выполнении гипотезы H_0^*;

$E_j=np_j$ - ожидаемое число попаданий в j-ый интервал;

Статистика: $\chi^2 = \sum_{j=1}^{k} \frac{(n_j - E_j)^2}{E_j} \sim \chi^2_{k-1}$ - Распределение хи-квадрат с k-1 степенью свободы.

Проверка гипотезы H_0



Распределение хи-квадрат

В зависимости от значения критерия χ^2, гипотеза H_0 может приниматься, либо отвергаться:

- $\chi^2_1 < \chi^2 < \chi^2_2$, гипотеза H_0 выполняется.
- $\chi^2 \leq \chi^2_1$ (попадает в левый "хвост" распределения). Следовательно, теоретические и практические значения очень близки. Если, к примеру, происходит проверка генератора случайных чисел, который сгенерировал n чисел из отрезка [0,1] и гипотеза H_0: выборка X^n распределена равномерно на [0,1], тогда генератор нельзя называть случайным (гипотеза случайности не выполняется), т.к. выборка распределена слишком равномерно, но гипотеза H_0 выполняется.
- $\chi^2 \geq \chi^2_2$ (попадает в правый "хвост" распределения) гипотеза H_0 отвергается.

Пример 1

Проверим гипотезу H_0: если взять случайную выборку 100 человек из всего населения острова Кипр (генеральной совокупности), где количество мужчин и женщин примерно одинаково (встречаются с одинаковой частотой), то в наблюдаемой выборке отношение количества мужчин и женщин будет соотноситься с частотой как и во всей генеральной выборке(50/50). Пусть в наблюдаемой выборке 46 мужчин и 54 женщины, тогда число степеней свобод $k-1=2-1=1$ и

$$\chi^2 = \sum_{j=1}^{k} \frac{(n_j - E_j)^2}{E_j} = \frac{(46-50)^2}{50} + \frac{(54-50)^2}{50} = 0{,}64$$

Т.о. при уровне значимости $\alpha = 0.05$ о выполнении гипотезы H_0 ничего сказать нельзя т.к. значение $\chi^2 > \chi^2_{0.05,1}$ (см. Таблицу распределения χ^2_1).

Сложная гипотеза

Гипотеза H_0^*: Xn порождается функцией $F(x,\theta),\ \theta \in R^d,\ \theta$ - неизвестный параметр. Найдем приближенное значение параметра $\widehat{\theta}$ с помощью метода максимального правдоподобия, основанного на частотах (фиксируем интервалы $(a_j,b_j]$ для $j=1\cdots k$).

$n_j=\sum_{i=1}^{n}\left[a_j<x_i\le b_j\right]$ - число попаданий значений элементов выборки в j-ый интервал.

$p_j(\theta)=F(b_j,\theta)-F(a_j,\theta)$,

$\widehat{\theta}=\arg\max_{\theta}\sum n_j \ln p_j(\theta)$

Теорема Фишера Для проверки сложной гипотезы критерий χ^2 представляется в виде:

$\chi^2=\sum_{j=1}^{k}\frac{\left(n_j-E_j\right)^2}{E_j}\sim\chi^2_{k-d-1}$, где $E_j=np_j(\widehat{\theta})$

Пример 2

Задача о бомбардировках Лондона [Лагутин, Т2]. Задача возникла в связи с бомбардировками Лондона во время Второй мировой войны. Для улучшения организации оборонительных мероприятий, необходимо было понять цель противника. Для этого территорию города условно разделили сеткой из 24-ёх горизонтальных и 24-ёх вертикальных линий на 576 равных участков. В течении некторого времени в центре организации обороны города собиралась информация о количестве попаданий снарядов в каждый из участков. В итоге были получены следующие данные:

Число попаданий	0	1	2	3	4	5	6	7
Количество участков	229	211	93	35	7	0	0	1

Гипотеза H_0: стрельба случайна (нет "целевых" участков).

Закон редких событий (распределение Пуассона)

$P\{S=j\}=\frac{\lambda^j}{j!}e^{-\lambda}$, где S - число попаданий, $\widehat{\lambda}=0.924$.

$\chi^2=\sum_{j=1}^{k}\frac{\left(n_j-E_j\right)^2}{E_j}=32.6\sim\chi^2_{8-1-1}$

Тогда при уровне значимости $\alpha=0.05$ гипотеза H_0 не выполняется (см. таблицу значений ф-ии χ^2_6).

Объединим события (4,5,6,7) с малой частотой попаданий в одно, тогда имеем:

Число попаданий	0	1	2	3	4-7
Количество участков	229	211	93	35	8

$\chi^2 = 1.05 \sim \chi^2_{5-1-1}$, тогда при $\alpha = 0.05$ гипотеза H_0 верна.

Проблемы

Критерий χ^2 ошибается на выборках с низкочастотными (редкими) событиями. Решить эту проблему можно отбросив низкочастотные события, либо объединив их с другими событиями. Этот способ называется коррекцией Йетса (Yates' correction).

3. Исследование зависимостей.

Если рассмотренные в предыдущих разделах описательная статистика и статистические критерии позволяли, соответственно, компактно представлять полученные результаты и определять сходства и различия, то следующим этапом анализа данных обычно является исследование зависимостей. Для этих целей применяются корреляционный и дисперсионный анализ (для установления факта наличия или отсутствия зависимости между переменными), а также регрессионный анализ (для нахождения количественной зависимости между переменными).

Корреляционный анализ. Корреляция (Correlation) – связь между двумя или более переменными (в последнем случае корреляция называется множественной). Цель корреляционного анализа – установление наличия или отсутствия этой связи. В случае, когда имеются две переменные, значения которых измерены в шкале отношений, используется коэффициент линейной корреляции Пирсона r, который принимает значения от -1 до +1 (его нулевое значение свидетельствует об отсутствии корреляции). Термин «линейный» свидетельствует о том, что исследуется наличие линейной связи между переменными – если *r(x, y) = 1*, то одна переменная линейно зависит от другой (и наоборот), то есть существуют константы *a* и *b*, причем *a > 0*, такие что *y = a x + b*.

Для данных, измеренных в порядковой шкале, следует использовать коэффициент ранговой корреляции Спирмена (он может применяться и для данных, измеренных в интервальной шкале, так как является непараметрическим и улавливает тенденцию – изменения переменных в одном направлении), который обозначается *s* и определяется сравнением рангов – номеров значений сравниваемых переменных в их упорядочении. Коэффициент корреляции Спирмена является менее чувствительным, чем коэффициент корреляции Пирсона (так как первый в случае измерений в шкале отношений учитывает лишь упорядочение x элементов выборки). В то же время он позволяет выявлять корреляцию между монотонно нелинейно связанными переменными (для которых коэффициент Пирсона может показывать незначительную корреляцию.

Универсальных рецептов установления корреляции между немонотонно и нелинейно связанными переменными на сегодняшний день не существует. Отметим, что большое (близкое

к плюс единице или к минус единице) значение коэффициента корреляции говорит о связи переменных, но ничего не говорит о причинно-следственных отношениях между ними.

Дисперсионный анализ. Изучение наличия или отсутствия зависимости между переменными можно проводить и с помощью дисперсионного анализа (Analysis of Variance – ANOVA). Его суть заключается в следующем. Дисперсия характеризует «разброс» значений переменной. Переменные связаны, если для объектов, отличающихся значениями одной переменной, отличаются и значения другой переменной. Значит, нужно для всех объектов, имеющих одно и то же значение одной переменной (называемой независимой переменной), посмотреть, насколько различаются (насколько велика дисперсия) значения другой (или других) переменной, называемой зависимой переменной. Дисперсионный анализ как раз и дает возможность сравнить отношение дисперсии зависимой переменной (межгрупповой дисперсии) с дисперсией внутри групп объектов, характеризуемых одними и теми же значениями независимой переменной (внутригрупповой дисперсией). Другими словами, дисперсионный анализ «работает» следующим образом. Выдвигается гипотеза о наличии зависимости между переменными. Выделяются группы элементов выборки с одинаковыми значениями независимой переменной (число таких групп равно числу попарно различных значений независимой переменной). Если гипотеза о зависимости верна, то значения зависимой переменной внутри каждой группы должны не очень различаться (внутригрупповая дисперсия должна быть мала). Напротив, значения зависимой переменной для различных групп должны различаться сильно (межгрупповая дисперсия должна быть велика). То есть, переменные зависимы, если отношение межгрупповой дисперсии к внутригрупповой (обычно обозначаемое буквой F) велико. Если же гипотеза неверна, то это отношение должно быть мало.

Регрессионный анализ. Если корреляционный и дисперсионный анализ, качественно говоря, дают ответ на вопрос, существует ли взаимосвязь между переменными, то регрессионный анализ предназначен для того, чтобы найти «явный вид» этой зависимости. Цель регрессионного анализа – найти функциональную зависимость между переменными. Для этого предполагается, что зависимая переменная (иногда называемая откликом) определяется известной функцией (иногда говорят – моделью), зависящей от зависимой переменной или переменных (иногда называемых факторами) и некоторого параметра. Требуется найти такие значения этого параметра, чтобы полученная зависимость (модель) наилучшим образом описывала имеющиеся экспериментальные данные. Например, в простой линейной регрессии предполагается, что зависимая переменная y является линейной функцией $y = a\ x + b$ от независимой переменной x. Требуется найти значения параметров a и b, при которых прямая $ax + b$ будет наилучшим образом описывать (аппроксимировать) экспериментальные точки $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

4. Снижение размерности. Часто в результате экспериментальных исследований возникают большие массивы информации.

Например, каждый из исследуемых объектов описывается по нескольким критериям (измеряются значения нескольких переменных – признаков). Тогда результатом измерений будет таблица с числом ячеек, равным произведению числа объектов на число признаков. Возникает вопрос, а все ли переменные являются информативными, например, отражают изменения, произошедшие в результате изучаемого воздействия? Исследователю желательно было бы выявить эти существенные переменные (это важно с содержательной точки зрения) и сконцентрировать внимание на них. Кроме того, всегда желательно сокращать объемы

обрабатываемой информации (не теряя при этом сути). Статистические методы могут помочь и здесь. Существует целый класс задач статистического анализа – методы снижения размерности – цель которых как раз и заключается в уменьшении числа анализируемых переменных либо посредством выделения существенных переменных, либо построения новых показателей (на основании полученных в результате эксперимента). Но за все (в том числе за агрегирование информации) надо платить – такой платой в задачах снижения размерности является та часть вариации (изменений, дисперсии) исходных показателей, которая объясняется изменениями тех показателей, которые не «остаются» в результате снижения размерности (наименее изменчивые показатели или их комбинации).

Для снижения размерности используются **факторный анализ и метод главных компонент.**

Факторный анализ – один из наиболее популярных многомерных статистических методов. Если кластерный и дискриминантный методы классифицируют наблюдения, разделяя их на группы однородности, то факторный анализ классифицирует признаки (переменные), описывающие наблюдения. Поэтому главная цель факторного анализа – сокращение числа переменных на основе классификации переменных и определения структуры взаимосвязей между ними. Сокращение достигается путем выделения скрытых (латентных) общих факторов, объясняющих связи между наблюдаемыми признаками объекта, т.е. вместо исходного набора переменных появится возможность анализировать данные по выделенным факторам, число которых значительно меньше исходного числа взаимосвязанных переменных.

Метод главных компонент заключается в получении нескольких новых показателей – главных компонент, являющихся линейными комбинациями исходных показателей (напомним, что линейной комбинацией называется взвешенная сумма), полученных в результате эксперимента. Главные компоненты упорядочиваются в порядке убывания той дисперсии, которую они «объясняют». Первая главная компонента объясняет большую часть дисперсии, чем вторая, вторая – большую, чем третья и т.д. Понятно, что чем больше главных компонент будет учитываться, тем большую часть изменений можно будет объяснить.

Преимущество метода главных компонент заключается в том, что зачастую первые несколько главных компонент (одна-две-три) объясняют большую часть (например, 80-90%) изменений большого числа (десятков, а иногда и сотен) параметров. Кроме того, может оказаться, что в первые несколько главных компонент входят не все исходные параметры. Тогда можно сделать вывод о том, какие параметры являются существенными и на них следует обратить внимание в первую очередь.

5. Классификация. Обширную группу задач анализа данных, основывающихся на применении статистических методов, составляют так называемые задачи классификации. В близких смыслах (в зависимости от предметной области) используются также термины: «группировка», «систематизация», «таксономия», «диагностика», «прогноз», «принятие решений», «распознавание образов».

Выделяются три подобласти теории классификации: дискриминация (дискриминантный анализ), кластеризация (кластерный анализ) и группировка.

Дискриминантный анализ. Это собственно задача классификации – построение правила (классификатора) отнесения наблюдаемого объекта к одному из ранее описанных классов. Этот

тип Машинного обучения (Machine Learning) относится к классу алгоритмов обучения с учителем (supervised learning). Проблемы: на обучающей выборке можно получить нулевую ошибку классификации, но на тестовой она часто оказывается не нулевой.

Кластерный анализ. По статистическим данным необходимо разделить элементы выборки на группы (кластеры). Классы заранее не заданы – обучение без учителя (unsupervised learning). Проблемы: неоднозначность решений (от алгоритма, начальных условий), зависимость от признакового пространства, число кластеров не известно.

К задачам **прогноза** обычно относят и задачи анализа временных рядов.

Анализ временных рядов. Временным рядом называется последовательность чисел – значений некоторого показателя, измеренного в различные моменты времени. Временные ряды используются для описания динамики процессов, например, изменения температуры тела, концентрации определенного вещества в крови и т.д. На основании конечного отрезка временного ряда исследователь должен сделать выводы о свойствах рассматриваемого процесса и тех механизмах (в рамках статистики – вероятностных механизмах), которые порождают этот ряд. При изучении временных рядов ставятся следующие цели: агрегированное описание характерных особенностей ряда; подбор статистических моделей, описывающих временной ряд; предсказание будущих значений на основании прошлых наблюдений (прогноз динамики); выработка рекомендаций по управлению процессом, порождающим временной ряд. На сегодняшний день существует множество моделей и методов, позволяющих достигать перечисленных выше целей с учетом специфики исследуемого процесса.