Лекция 2. СТАТИСТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА ДАННЫХ

Основные понятия математической статистики. Математическая статистика – это наука, посвященная разработке методов описания и анализа статистических экспериментальных данных, полученных в результате наблюдений массовых случайных явлений. Математическая статистика изучает методы сбора, обработки и интерпретации результатов опытов (экспериментов).

Основные задачи МС:

1) первичная обработка результатов наблюдений (эксперимента);
2) проверка статистических гипотез;
3) статистические оценки параметров распределения;
4) исследование статистических зависимостей (связей).

Значительная часть математической статистики связана с описанием и анализом больших совокупностей объектов, объединенных по некоторому качественному или количественному признаку X. Такая группа объектов называется статистической совокупностью. Если исследуемая совокупность слишком многочисленна, либо ее элементы малодоступны, либо имеются другие причины, не позволяющие изучать сразу все ее элементы, прибегают к изучению какой-то части этой совокупности. Эта выбранная для полного исследования группа элементов называется выборочной совокупностью или выборкой, а все множество изучаемых элементов – генеральной совокупностью. Под выборкой понимается последовательность независимых, одинаково распределенных случайных величин, т. е. каждая выборка $(x_1, x_2, \ldots, x_n)$ значений случайной величины X рассматривается как результат n независимых повторных испытаний. Таки образом выборка должна быть организована так, чтобы каждый объект генеральной совокупности имел одинаковые шансы попасть в эту выборку Объемом совокупности называется число объектов, входящих в эту совокупность. Например, если из 10 000 микросхем для проверки качества отобрано 200 штук, то объем генеральной совокупности равен 10 000, а выборочной – 200.

Статистическая совокупность, расположенная в порядке возрастания или убывания значений изучаемого признака X, называется вариационным рядом, а ее объекты – вариантами. Наиболее распространенными характеристиками статистического

распределения являются средние величины: мода, медиана и средняя арифметическая, или выборочная средняя.

Вариационный ряд называется дискретным, если его члены принимают конкретные изолированные значения. Если элементы вариационного ряда заполняют некоторый интервал, то такой ряд называется непрерывным.

Методы анализа данных

Дескриптивная статистика. Описательная статистика — один из разделов статистической науки, в рамках которого изучаются методы описания и представления основных свойств данных. Цель описательной (дескриптивной) статистики — обработка эмпирических данных, их систематизация, наглядное представление в форме графиков и таблиц, а также их количественное описание посредством основных статистических показателей.

Позволяет обобщать первичные результаты, полученные при наблюдении или в эксперименте. Применение описательной статистики включает следующие этапы:

1. Сбор данных
2. Категоризация данных
3. Обобщение данных
4. Представление данных

Описательная статистика позволяет обобщать первичные результаты, полученные при наблюдении или в эксперименте. Все расчеты описательных статистик сводятся к группировке данных по их значениям, построению распределения их частот, выявлению центральных тенденций распределения и, наконец, к оценке разброса данных по отношению к найденной центральной тенденции.

Представление описательных статистик является, как правило, первым шагом любого анализа. Цель представления данных в виде описательных статистик – сделать выводы и принять стратегические (для анализа) решения, основанные на имеющихся данных.

Основные показатели описательной статистики:

- Среднее значение (среднее арифметическое, медиана, мода).
- Усредненное значение.
- Разброс (диапазон разброса данных).
- Дисперсия.
- Стандартное (среднеквадратическое) отклонение.
- Квартили.

- Доверительный интервал.

В рамках описательной статистики применяются следующие простейшие техники:

- Графическое представление данных.
- Табличное представление данных.
- Использование обобщающих статистик, таких, как математическое ожидание, медиана, дисперсия и т.д.

Параметрические методы. Все параметрические методы статистики работают с интервальной шкалой, в отличие от непараметрических методов, ориентированных прежде всего на первые две шкалы.

При рассмотрении большинства статистических методов предполагается, что наблюдения, о которых идет речь, выражены в интервальной шкале и являются реализациями случайной величины, распределение которой принадлежит некоторому параметрическому семейству распределений. Например, случайная величина имеет нормальное, или пуассоновское, или другое распределение. То есть, мы предполагаем, что известна форма распределения, например, мы можем предполагать нормальную $N(\mu, \delta)$ модель, но с неизвестными параметрами μ и δ. Методы оценивания и проверки гипотез позволяют делать выводы о неизвестных параметрах, при этом ценность любых заключений до некоторой степени должна зависеть от адекватности исходного предположения о параметрическом семействе, то есть о форме распределения. Однако существуют случайные величины, которые не подчиняются одной из распространенных форм распределения. Следовательно, к ним нельзя применить те математические методы, которые разработаны для параметрических распределений. Поэтому для таких признаков разработаны специальные математические модели, которые получили название непараметрических или свободных от распределения.

Преимущество параметрических методов состоит в том, что для них существует хорошо разработанный математический аппарат. Однако применение этих методов, кроме прочего, предполагает большой объем выборки. Параметрические методы используют для количественных признаков.

Непараметрические методы. Непараметрические методы позволяют обрабатывать данные "низкого качества" из выборок малого объёма с переменными, про распределение которых мало что или вообще ничего неизвестно.

Непараметрические методы не основываются на оценке параметров (таких как среднее или стандартное отклонение) при описании выборочного распределения интересующей величины. Поэтому эти методы иногда также называются свободными от параметров или свободно распределенными.

Для анализа номинальных и ранговых переменных используются только непараметрические методы, которые не требуют предварительных предположений относительно вида исходного распределения. В этом их достоинство. Но есть и недостаток – снижение т.н. мощности (чувствительности к различиям объектов). Поясним это.

В качестве примера непараметрического метода можно привести найденный А. Н. Колмогоровым способ проверки согласованности теоретических и эмпирических распределений (так называемый критерий Колмогорова). Пусть результаты n независимых наблюдений некоторой величины имеют функцию распределения F(x) и пусть $F_n(x)$ обозначает эмпирическую функцию распределения (см. Вариационный ряд), построенную по этим n наблюдениям, а D_n — наибольшее по абсолютной величине значение разности $F_n(x)$ — F(x). Случайная величина $\sqrt{n}\,D_n$ имеет в случае непрерывности F(x) функцию распределения K_n (λ), не зависящую от F(x) и стремящуюся при безграничном возрастании n к пределу

$$K(\lambda) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2 j \lambda^2}$$

Отсюда при достаточно больших n, для вероятности $p_{n,\lambda}$ для неравенства $\sqrt{n}\,D_n \geq \lambda$

получается приближённое выражение $p_{n,\lambda} \approx 1 - K(\lambda)()$

Функция К (λ) табулирована. Её значения для некоторых λ приведены в табл. 1.1.

Таблица 1.1. Значения функции К (λ)

λ	0,57	0,71	0,83	1,02	1,36	1,63
$K(\lambda)$	0,10	0,30	0,50	0,75	0,95	0,99

Равенство (*) следующим образом используется для проверки гипотезы о том, что наблюдаемая случайная величина имеет функцию распределения F(x): сначала по результатам наблюдений находят значение величины D_n, а затем по формуле (*) вычисляют вероятность получения отклонения F_n от F, большего или равного наблюдённому. Если указанная вероятность достаточно мала, то в соответствии с общими принципами проверки статистических гипотез проверяемую гипотезу отвергают. В противном случае считают, что результаты опыта не противоречат проверяемой гипотезе. Аналогично проверяется гипотеза о том, получены ли две независимые выборки, объёма n_1 и n_2 соответственно, из одной и той же генеральной совокупности с непрерывным законом распределения. При этом вместо формулы (*) пользуются тем, что вероятность неравенства

$$D_{n_1,n_2}\sqrt{\frac{n_1 n_2}{n_1+n_2}}<\lambda$$

как это было установлено Н. В. Смирновым, имеет пределом К (λ), здесь D_{n_1,n_2} есть наибольшее по абсолютной величине значение разности $F_{n_1}(x) — F_{n_2}(x)$.

Другим примером Н. м. могут служить методы проверки гипотезы о том, что теоретическое распределение принадлежит к семейству нормальных распределений. Отметим здесь лишь один из этих методов — так называемый метод выпрямленной диаграммы. Этот метод основывается на следующем замечании. Если случайная величина X имеет нормальное распределение с параметрами α и σ, то

$$\Phi^{-1}\left[F(x)\right]=\frac{x-\alpha}{\sigma}$$

где Φ^{-1} — функция, обратная нормальной:

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-u^2}du$$

Таким образом, график функции y = $\Phi^{-1}\left[F(x)\right]$ будет в этом случае прямой линией, а график функции y = $\Phi^{-1}\left[F_n(x)\right]$ — ломаной линией,

близкой к этой прямой (см. рис.2.1). Степень близости и служит критерием для проверки гипотезы нормальности распределения F (x).

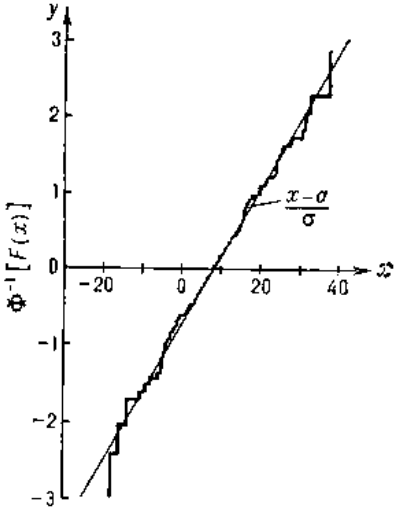

Рис.2.1 Графики функций

Корреляционный анализ. Корреляционный анализ — метод обработки статистических данных, заключающийся в изучении коэффициентов (корреляции). Его применение возможно в случае наличия достаточного количества (для конкретного вида коэффициента корреляции) наблюдений из более чем одной переменной. При этом сравниваются коэффициенты корреляции между одной парой или множеством пар признаков, для установления между ними статистических взаимосвязей.

Корреляционный анализ позволяет судить о том, насколько похоже ведут себя разные переменные. В самом общем виде принятие гипотезы о наличии корреляции означает, что изменение значения переменной А произойдет одновременно с пропорциональным изменением значения Б: если обе переменные растут, то корреляция положительная; если одна переменная растет, а вторая уменьшается – корреляция отрицательная.

При изучении корреляций стараются установить, существует ли какаято связь между двумя показателями в одной выборке (например, между ростом и весом детей или между уровнем IQ и школьной успеваемостью) либо между двумя различными выборками (например, при сравнении пар близнецов), и если эта связь

существует, то сопровождается ли увеличение одного показателя возрастанием (положительная корреляция) или уменьшением (отрицательная корреляция) другого.

Регрессионный анализ. Регрессионный анализ – методика анализа отношений между двумя и более переменными интервального уровня с целью предсказания значения одной по сравнению с другой или другими. Например, при уравнении регрессии, описывающим отношение между размером дохода и числом лет обучения, доход может быть предсказан, если известно число лет обучения.

Многократный линейный регрессионный анализ используется в тех случаях, когда имеются несколько независимых переменных интервального уровня. Например, можно вывести линейное уравнение, которое связывало бы доход с годами обучения, возрастом и годами опыта работы. Простая линейная регрессия: $y=b*x+a$, $b>0$ – связь прямопропорциональная, функция возрастает $b<0$ – связь обратнопропорциональная, функция убывает

Регрессионный анализ используют в тех случаях, когда: - необходимо установить, реально ли есть взаимосвязь между переменными; - необходимо установит тесноту связи зависимых и независимых переменных; - нужно определить форму связи; - нужно предсказать значение зависимой переменной; - необходимо осуществлять контроль над независимыми переменными при определении вкладов конкретной переменной.

Регрессионный анализ служит для выявления вида влияния одной переменной на другую. Корреляционный анализ устанавливает наличие зависимости, а регрессионный – вид зависимости: линейная, квадратичная, экспоненциальная и т.д. Предполагается, что связь между величинами линейная. Если мы знаем уравнение линейной регрессии, то по ответу человека на вопрос X мы можем спрогнозировать (с некоторой точностью) его ответ на вопрос Y.

Для проведения регрессионного анализа необходимо следующее:

-Выбор одного блока, из которого берется координатный интервал, чьи данные дают зависимую переменную регрессии.

-Выбор одного или нескольких блоков, из которых аналогично берутся факторы в качестве независимых переменных регрессии. При этом необходимо, чтобы блок, дающий зависимую переменную, и все блоки, дающие независимые переменные, имели какие-либо общие координаты (обычно пространство и время), которые служат переменными развертки и дают точки, по которым проводится регрессионная кривая или поверхность.

-Выбор типа и «степени» функций от независимых переменных, которые включаются в регрессию. -Задание координатных интервалов переменных сравнения, внутри которых регрессионная функция не должна значимо изменяться.

-Определяется точность предсказания. Для этого находится стандартная ошибка оценки регрессии. Регрессия проводится последовательно с увеличением числа независимых переменных и степени регрессионной функции. При этом общесистемным оптимизатором находится минимум среднеквадратичного отклонения точек данных от регрессионной кривой.

Для регрессионной кривой вычисляются характеристики неопределенности - показатели тесноты регрессии: кривые доверительного интервала и коэффициент детерминации. Последний может вычисляться сразу для всех комбинаций «зависимая переменная - независимая переменная». Как и корреляция, регрессия рассчитывается для фиксированных координатных интервалов каждой переменной сравнения. Проверяется устойчивость регрессии к смене координатного интервала на том же уровне иерархии.

Так же как и корреляционный анализ, регрессионный имеет свои особенности и направленности. Для установления математической зависимости между двумя метрическими переменными – зависимой и независимой используется парная регрессия. Множественная регрессия используется для определения математической зависимости между двумя или больше независимыми переменными и зависимой переменной, выраженной с помощью интервальной или относительной шкал. Силу тесноты связи в данном случае измеряют с помощью коэффициента множественной детерминации (аналогично, как и при корреляции). При пошаговой регрессии независимые переменные вводят и выводят из уравнения регрессии один за другим, чтобы выбрать меньшее их количество, которое объясняет большую часть вариации.

Пример регрессионного анализа: Ошеломительным примером такого анализа является пример компании Sun Microsystems, которая обошла по продажам компанию IBM. Взяв за основу регрессионный анализ конкурентных преимуществ, компания стала лидером на рынке технологий. Регрессионный анализ проводился следующим образом: было взято три набора независимых переменных: численность специалистов в компании конкурента, расходы на рекламу и расходы на разработки. И все они использовались только благодаря проведенному ранее бенмаркингу. Зависимой переменной являлся

объем сбыта. Проведение данного анализа показало, что именно из-за численности персонала страдала компания Sun Microsystems и была в лидерах IBM. Из-за большей численности персонала в компании Sun Microsystems возникала разобщенность на профессиональном уровне, и зачастую не было единого мнения по внедрению того или иного продукта, деньги на разработки выделялись, но большинство из разработок так и оставались разработками и не внедрялись. Напротив, в IBM менее крупной по численности компании разработки быстро уходили на рынок и скупались практически сразу. По итогам анализа, Sun Microsystems не решилась сокращать персонал, боясь утечки информации, а разделилась на филиалы и тем самым увеличила свои продажи, и 3 года находилась на пике в лидерах.

Дисперсионный анализ. С помощью дисперсионного анализа исследуют влияние одной или нескольких независимых переменных на одну зависимую переменную (одномерный анализ) или на несколько зависимых переменных (многомерный анализ). В обычном случае независимые переменные принимают только дискретные значения (и относятся к номинальной или порядковой шкале); в этой ситуации также говорят о факторном анализе. Если же независимые переменные принадлежат к интервальной шкале или к шкале отношений, то их называют ковариациями, а соответствующий анализ — ковариационным. Концепция дисперсионного анализа предложена Р. Фишером в 1920 г. и состоит в выделении и сравнении между собой различных компонент дисперсии признака Y (отсюда и название метода). Эти компоненты выделяются посредством разложения вариации (SS) признака Y на составные части. Сравнение компонент позволяет делать вывод о значимости или незначимости влияния отдельного фактора на изменчивость признака Y. Дисперсионный анализ, возникший как метод планирования эксперимента (Р. Фишер предложил его. для обработки результатов опытов по выявлению условий, при которых испытываемый сорт сельскохозяйственной культуры дает максимальный урожай), используется как метод анализа данных для выявления систематических различий между результатами непосредственных измерений, выполненных при тех или иных меняющихся условиях (что особенно важно для социологии).

Для применения дисперсионного анализа требуется определенная структура представления исходных данных. Задачу однофакторного дисперсионного анализа можно представить как проверку связи двух признаков, один из которых измеряется по интервальной шкале, а другой – по номинальной. Эта задача является логическим

обобщением сравнения средних значений в группах на ситуацию, когда групп не две, а больше. Действительно, признак, измеряемый по номинальной шкале, делит всех респондентов на группы (число групп равно числу вариантов ответа). В каждой группе можно вычислить среднее значение того признака, который измеряется по интервальной шкале. Если связи между факторами нет, то во всех группах средние значения будут равны. дисперсионный анализ проверяет более строгое условие, формулируемое как нулевая гипотеза: «Средние значения во всех группах равны».

Кластерный анализ. Термин «кластерный анализ» (впервые ввел Tryon в 1939 г.) в действительности включает в себя набор различных алгоритмов классификации. Общий вопрос, задаваемый исследователями во многих областях, состоит в том, как разбить данные на группы с близкими значениями параметров.

Например, при сегментации рынка можно кластеризовать потребителей по двум параметрам — цены и качества. Допустим, компания — производитель автомобилей провела опрос потребителей, в котором задавала два вопроса: «За какую цену Вы готовы купить автомобиль?» и «Оцените качество автомобиля X по 50-балльной шкале» (несколько странный вопрос, однако в качестве иллюстрации он вполне подходит). В результате опроса были получены следующие данные (таблица 1.2.)

Таблица 1.2. Данные опросов

№ участника опроса	Цена, тыс. $	Качество автомобиля X
1	27	19
2	11	46
3	25	15
4	36	27
5	35	25
6	10	43
7	11	44
8	36	24
9	26	14
10	26	14
11	9	45
12	33	23
13	27	16
14	10	47

Если посмотреть на диаграмму (так называемая диаграмма рассеяния) «цена — качество», представленную на рис. 2.2, то сразу будут видны группы потребителей:

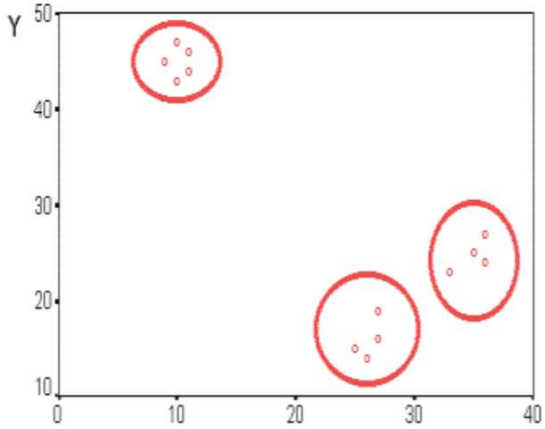

Рис. 2.2. Соотношение цены — качества

Владея этой информацией, каждой группе потребителей можно предложить именно то, что необходимо именно этой группе, и за счет этого увеличить уровень продаж компании.

Разумеется, в реальной жизни кластеры, различимые глазом, встречаются нечасто, гораздо чаще бывают ситуации, когда все результирующие параметры смешиваются в одну «кучу» (см. рис. 2.3)

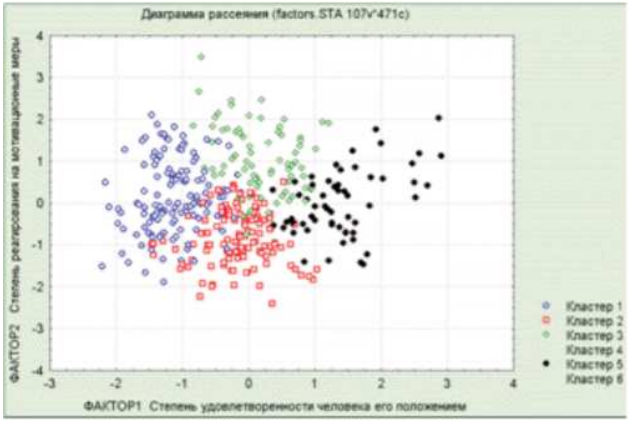

Рис. 2.3. Диаграмма рассеяния

Особенно часто это встречается, когда анализируемых параметров не два, а несколько десятков (кластерный анализ не ограничивает число анализируемых параметров, поэтому можно рассматривать всю проблему комплексно).

Для проведения кластерного анализа, кроме сбора данных, необходимо определить две вещи: на какое количество кластеров необходимо разделить данные и как определить меру сходства в данных. Например, все предприятия России можно кластеризовать по географическому признаку на 10 кластеров. Тогда мера сходства будет определяться коммуникационной близостью предприятий друг к другу.

Дискриминантный анализ. Дискриминантный анализ – это инструмент статистики, который используется для принятия решения о том, какие переменные разделяют возникающие наборы данных.

Например, некий исследователь в области образования решает исследовать, какие переменные относят выпускника средней школы к одной из трех категорий: (1) поступающий в колледж, (2) поступающий в профессиональную школу или (3) отказывающийся от дальнейшего образования или профессиональной подготовки. Для этой цели исследователь может собрать данные о различных переменных, связанных с учащимися школы: пол, возраст, успеваемость, материальное положение семьи и т. д. После выпуска большинство учащихся должно попасть в одну из названных категорий.

Затем можно использовать дискриминантный анализ для определения того, какие переменные дают наилучшее предсказание выбора учащимися дальнейшего пути. Например, можно математически определить, что учащиеся с низкой успеваемостью и низким достатком в семье скорее всех попадают в 3-ю категорию.

Еще пример: есть данные о клиентах / потребителях, которых можно разделить по группам (совершившие повторную покупку – не совершившие повторную покупку; покупатели марки А – покупатели марки В – покупатели марки С; высокие риски невозврата кредита – низкие риски невозврата кредита), также имеется дополнительная информация о клиентах / потребителях. Дискриминантный анализ позволяет выяснить, действительно ли группы различаются между собой, и если да, то каким образом (какие переменные вносят наибольший вклад в имеющиеся различия).

Факторный анализ. Факторный анализ — многомерный метод, применяемый для изучения взаимосвязей между значениями переменных. Предполагается, что известные переменные зависят от меньшего количества неизвестных переменных и случайной ошибки.

В случае наличия большого числа параметров (более 100) имеет смысл сгруппировать параметры и анализировать уже не каждый параметр в отдельности, а группы параметров как единый комплексный параметр (фактор).

В основе факторного анализа лежит идея о том, что за сложными взаимосвязями явно заданных признаков стоит относительно более простая структура, отражающая наиболее существенные черты изучаемого явления, а «внешние» признаки являются функциями скрытых общих факторов, определяющих эту структуру.

Факторный анализ позволяет решить две важные проблемы исследователя: описать объект измерения всесторонне и в то же время компактно. С помощью факторного анализа возможно выявление скрытых переменных факторов, отвечающих за наличие линейных статистических корреляций между наблюдаемыми переменными.

Две основных цели факторного анализа:

- определение взаимосвязей между переменными, (классификация переменных), то есть «объективная R-классификация»;
- сокращение числа переменных необходимых для описания данных.

Например, для анализа структуры экономического роста России можно проанализировать все макроэкономические параметры, предварительно объединив их в группы. Одним из таких факторов будет являться ВВП.

Объединение параметров можно делать вручную, эмпирически, как это сделано с ВВП, а можно с помощью метода факторного анализа. Применение факторного анализа позволяет, во-первых, уменьшать (редуцировать) число рассматриваемых параметров, во-вторых — находить осмысленные группы параметров, каждая из которых будет являться одним самостоятельным параметром.

Спецификой этого метода является то, что при объединении параметров в факторы каждый фактор аккумулирует в себе общие закономерности во всех параметрах, отбрасывая особенности каждого параметра в отдельности.