# Лекция 1. ЭКОЛОГИЧЕСКАЯ ИНФОРМАЦИЯ И ОСОБЕННОСТИ ЕЕ ОБРАБОТКИ. МЕТОДЫ СТАТИСТИЧЕСКОГО АНАЛИЗА ЭКОЛОГИЧЕСКИХ ДАННЫХ

1.1. Экологическая информация и особенности ее обработки

1.2. Статистические методы анализа результатов эксперимента

## 1.1. Экологическая информация и особенности ее обработки

В основе научных исследований и их практических приложений, как правило, лежит информация.

***Информация*** *– сведения об исследуемом объекте, зафиксированные каким-либо образом.*

***Геоэкологическая информация (ГЭИ)*** *– информация о геоэкологических объектах или объектах, изучаемых в геоэкологии.*

Информация может фиксироваться в памяти человека или на каких-либо технических носителях: бумага, магнитные диски, лазерные диски, USB-flash-drive (флеши) и т.д.

Информация сама может служить объектом хранения, переработки, передачи и изучения. Информацию можно создавать, передавать, запоминать, искать, принимать, копировать, обрабатывать, разрушать. Информация может быть в самых разнообразных формах: в форме световых, звуковых или радиоволн, электрического тока или напряжения, магнитных полей, знаков на бумажном носителе. В принципе, информацию может переносить любая материальная структура или поток энергии.

Масштабы использования информации являются одним из основных признаков, отличающих мыслящие особи от всех остальных существ (В. Иллингуорт, 1989).

**Источники геоэкологической информации**

Геоэкологическую информацию формируют многочисленные и разнообразные организации. Это законодательные и исполнительные органы власти РФ, аналогичные органы в субъектах федерации, которые издают различные нормативные акты, принимают планы и программы; службы соответствующих министерств и ведомств, которые ведут те или иные наблюдения за состоянием окр ужающей среды, здоровьем людей, а также за факторами воздействия на окружающую среду. Это многочисленные научно-иследовательские организации – институты, станции и пр. Кроме того, сбором и анализом экологической информации занимаются проектировщики, поскольку, например, без знания характеристик окружающей среды невозможно спроектировать прочное и полезное сооружение. Эта же информация оказывается и в распоряжении заказчиков проектной документации – государственных организаций или коммерческих компаний.

Также сбором экологической информации занимаются международные организации, координирующие деятельность по охране окружающей среды, и общественные экологические организации.

Одним из важнейших понятий, связанных с экологической информацией, ее получением, является экологический мониторинг. В функции различных российских ведомств, определенные нормативными документами, входит целый ряд видов мониторинга: от мониторинга состояния и загрязнения окружающей среды до социально-гигиенического мониторинга.

Такая информация может содержать данные:

– о состоянии воды, воздуха, фауны, флоры, земли, почвы, недр, природных ландшафтов и комплексов;

– об экологической угрозе или риске для здоровья и жизни людей;

– о химических, физических и биологических воздействиях на состояние окружающей среды и их источниках;

– о хозяйственной деятельности, отрицательно влияющей или могущей повлиять на природные объекты и человека;

– о мерах по охране окружающей среды, в том числе правовых, административных и иных мерах;

– о деятельности государственных органов, юридических лиц и граждан-предпринимателей в сфере распоряжения природными ресурсами, природопользования, охраны окружающей среды, обеспечения соблюдения и защиты экологических прав и законных интересов физических и юридических лиц, если необходимость осуществления такой деятельности установлена законодательством России.

**Методы изучения геоэкологических процессов**

Геоэкологические объекты могут оставаться неизменными или меняться во времени и/или в пространстве. В свою очередь смена состояний экологических объектов может быть естественной или вызванной человеческой деятельностью. Для исследования изменений состояний объекта во времени вводится понятие «процесс».

***Процесс*** *– последовательная смена состояний или стадий развития объекта во времени или пространстве.*

Геоэкологический процесс развертывается во времени в виде определенной последовательности событий. Представления об этом процессе получаются, как правило, по данным наблюдений или экспериментов.

До недавнего времени в естественных науках использовались почти исключительно два метода экспериментальных исследований: метод пассивного и метод активного эксперимента. В последние годы широкое распространение получил метод численных экспериментов.

***Метод пассивного эксперимента.*** Метод заключается в наблюдениях за природными явлениями в естественных условиях на обширной сети станций и постов. Результаты наблюдений хранятся в различных фондах и частично публикуются в виде ежегодников и в различного рода справочниках. Значение таких данных трудно переоценить, так как они являются фундаментом, на котором развиваются геоэкологические и все другие исследования

окружающей среды. Вместе с тем набор материалов, которые могут быть получены на основе пассивных экспериментов, имеет весьма существенные ограничения.

Во-первых, эти материалы являются выборочными во времени. Действительно, стационарные наблюдения охватывают период времени определенной продолжительности, не превышающий, как правило, 100–150 лет. В подавляющем числе случаев продолжительность этого периода меньше 40–50 лет. Кроме того, наблюдения в основном приурочены к нашему времени, т. е. охватывают только последний период развития рассматриваемых процессов. Вполне вероятно, что в него не попали годы экстремальной водности или годы с редким сочетанием тех или иных геофизических факторов. А именно эти данные часто интересуют потребителей геоэкологической информации.

Во-вторых, материалы наблюдений являются выборочными в пространстве. Как известно, преобладающее большинство пунктов наблюдений, особенно продолжительных во времени, приурочено к обжитым местам. В то же время в связи с развитием народного хозяйства происходит все большее смещение промышленности в северные и восточные, ранее малообжитые, районы. А именно в этих районах данные наблюдений отсутствуют или крайне немногочисленны. Кроме того, формирование климатических условий на отдельных территориях происходит, как правило, в верхней и средней части речных бассейнов, а наблюдения в основном приурочены к нижней, более обжитой части этих бассейнов.

В-третьих, многие природные явления, активно участвующие в формировании геоэкологичеких процессов, настолько завуалированы, что наблюдать их в природе практически невозможно.

***Метод активного эксперимента.*** Метод заключается в постановке полевых или лабораторных физических экспериментов.

Так как формирование природных процессов имеет сложный, многоступенчатый и многофакторный характер, то такие эксперименты во многих случаях основаны на более или менее грубом приближении физических моделей к природным объектам. Вследствие этого экспериментальные результаты, полученные в лаборатории и в полевых условиях, не всегда могут быть перенесены в естественные условия.

***Метод численных экспериментов.*** В настоящее время в связи с развитием вычислительной техники и повышением общего уровня исследований в практику расчетов и прогнозов широко внедряется третий метод экспериментальных исследований – метод численных экспериментов, осуществляемый на основе математического моделирования рассматриваемых процессов.

***Численным экспериментом*** *называется численное решение уравнений математической модели некоторой системы при тех или иных заданных условиях или значениях параметров.*

Особенно важную роль численные эксперименты играют в следующих областях наук о Земле.

1. Исследование состояний окружающей среды с помощью математических моделей для тех условий, которые не могут быть изучены путем непосредственных наблюдений или путем постановки лабораторных или полевых экспериментов.

Например, одной из важнейших задач при водохозяйственном проектировании является определение возможной продолжительности маловодных периодов. Ряды непосредственных наблюдений, всегда ограниченные по продолжительности, или результаты лабораторных и полевых экспериментов в этом случае помочь не могут. В то же время метод численных экспериментов, осуществляемый на основе статистической модели колебаний водности, – метод Монте-Карло – позволяет по статистическим характеристикам исследуемых рядов получить выборки практически неограниченной продолжительности, включающие в себя почти все возможные циклы колебаний водности.

2. Конструирование наиболее простых, физически обоснованных и достаточно точных моделей процессов в окружающей среде.

Действительно, численные эксперименты позволяют задавать любые сочетания гидрологических факторов в возможном диапазоне их изменения и предполагаемом соотношении и путем пересчета устанавливать рациональную структуру и параметры предп о ла гаемой математической модели.

3. Планирование и постановка экспериментов в полевых и лабораторных условиях на основе разработанных путем численных экспериментов математических моделей.

В настоящее время во многих отраслях науки и техники все чаще ставится требование, чтобы перед тем, как перейти к дорогостоящим лабораторным или полевым экспериментам, возможные ситуации, возникающие в ходе опытов, были проиграны на математических моделях путем численных экспериментов.

**Геоэкологические процессы и ряды наблюдений**

Из данных, полученных в результате экспериментов наблюдений, объединенных по какому-либо признаку, конструируются ряды наблюдений.

***Ряд наблюдений*** *– последовательность данных наблюдений или каких-либо других сведений, характеризующих данный объект во времени и/или в пространстве.*

Ряды наблюдений могут быть первичными и вторичными.

***Первичный ряд*** *– ряд данных непосредственных наблюдений за теми или иными процессами.*

***Вторичный ряд*** *– ряд данных, полученный на основе обработки первичного ряда данных наблюдений.*

Деление рядов на первичные и вторичные во многом условно, так как в зависимости от имеющихся приборов и принятой методики наблюдений или измерений те же самые данные могут составлять и первичный и вторичный ряд. Так влажность воздуха может измеряться непосредственно, например, с

помощью гигрографов. В этом случае ряд данных является первичным. Те же самые данные, полученные расчетом по данным наблюдений с помощью психрометров (сухой и смоченный термометр) являются вторичными.

Ряды наблюдений могут быть ***дискретными*** или ***непрерывными***. Например, запись уровней воды самописцем дает непрерывный ряд наблюдений. Аналогичные ряды наблюдений могут быть получены при непрерывной записи и по другим процессам. Однако чаще значения уровней наблюдаются на простейших уровнемерных устройствах в конкретные сроки, например, ежедневно в 8 и в 20 ч. Временной ряд, составленный из таких наблюдений, называется дискретным.

Для практических расчетов непрерывные ряды обычно дискретизируются (квантуются). Это делается для удобства обработки и анализа. При этом используются два метода дискретизации: по равномерным интервалам или по характерным точкам.

Наблюдения могут производиться через равные интервалы времени. Такие дискретные ряды называются ***эквидистентными***. Наблюдения могу производиться и не через равные интервалы. Такие дискретные ряды называются ***неэквидистентными***.

Ряды наблюдений могут быть временными, пространственными и пространственно-временными.

Ряды наблюдений обычно обозначают заглавной латинской буквой $X$, $Y$, или $Z$, и т.д. В этих обозначениях $X$ представляет всю последовательность значений ряда $X$, $Y$ – всю последовательность значений ряда $Y$. Отдельные значения каждого ряда обозначаются соответствующими строчными буквами с опущенным индексом. Например первое значение ряда $X$ – $x_1$, второе значение – $x_2$, и т.д.; любое $i$-ое значение – $x_i$, где $i$ может изменяться от 1 до $n$, $n$ – число членов ряда. В соответствии с этими обозначениями значения ряда $X$ можно представить в следующем виде: $x_1, x_2, \ldots, x_n$,

а отдельные значения ряда: $x_i$, ($i = 1, 2, \ldots, n$).

Данные в рядах наблюдений могут представляться в различном виде, чаще всего в цифровом и реже в символьном обозначении. Например, концентрация загрязняющих веществ в реке может задаваться в количестве миллиграмм на литр воды, а влажность почвы – в символьном обозначении.

Следует отметить, что в геоэкологии используются ряды наблюдений над разнообразными объектами окружающей среды. Эти ряды имеют различную продолжительность периода и частоту наблюдений. Так при анализе уровней загрязнения окружающей среды следует иметь в виду, что по большинству городов и промышленных зон материалы наблюдений, как правило, освещают только период установившейся урбанизации. Так, в пределах г. Санкт-Петербурга регулярные наблюдения за химическим составом речного стока начались в основном лишь в период с 1959 по 1969 г.

Кроме того, материалы наблюдений за факторами формирования окружающей среды во многих случаях являются неоднородными во времени. Так, если в первые годы наблюдений за качественными характеристиками окружающей среды отбор проб и определение концентраций загрязняющих

веществ или их характеристик проводились 4 раза в год, то затем 7 раз, а с 1989 г. – 12 раз в год. Естественно, что эти обстоятельства в значительной степени затрудняют ретроспективный анализ возможных средних годовых и экстремальных годовых значений, так и анализ тенденций изменяющегося качества воды во времени.

Следует также отметить, что определение химических и других веществ в пробах воды обычно определяется с достаточной для анализа точностью. Однако сами по себе пробы воды в некоторых случаях являются нерепрезентативными, то есть не отражают действительное содержание химических веществ в реке или водоеме.

**Особенности геоэкологической информации**

Предметом изучения большинства наук о земле являются естественные процессы, происходящие в природе. В геоэкологии особое внимание уделяется процессам, которые изначально формировались и поддерживались внешними естественными условиями и лишь в последние десятилетия или годы стали подвергаться антропогенному воздействию. В связи с этим геоэкологические наблюдения над многими процессами начались сравнительно недавно и часто носили отрывочный или нерегулярный характер.

В связи с этим информация, используемая в геоэкологии, зачастую во многом отличается от информации, используемой, например, в гидрологии и метеорологии.

Иногда первичная обработка данных гидрохимических и ряда других наблюдений, освещающих состояние природной среды, ведётся не совсем корректно, так как не учитывает особенности этих рядов наблюдений, что приводило и приводит к погрешностям в результатах расчетов и, конечно же, в какой-то степени сказывается на выводах на основе этих данных [труды РГГМУ]. Недостаточная корректность обработки данных в особой степени относится, например, к расчетам среднегодовых концентраций содержащихся в речном стоке веществ.

Так методика расчёта среднегодовых концентраций содержащихся в воде веществ, по гидрохимическим наблюдениям, принятая в настоящее время, основана на следующих теоретических предположениях:

1) временные ряды измеренных значений концентраций содержащихся в воде веществ в каждом пункте наблюдений описываются моделью в виде ряда значений случайной величины;

2) данные ряды наблюдений являются стационарными, регулярными и однородными.

Следует учитывать, что при анализе среднегодовых концентраций, рассчитанных на этой основе, могут быть получены противоречивые результаты, не имеющие какого-либо физического объяснения. В связи с этим возник вопрос о репрезентативности и надёжности расчётов среднегодовых концентраций содержащихся в воде веществ существующими методами. В выполненных исследованиях было показано, что гидрохимические наблюдения, имеют ряд особенностей, которые не укладываются в рамки рассмотренных выше теоретических положений. Во-первых, исходные ряды

неоднородны по числу измерений в год; во-вторых, концентрации рассматриваемых в анализе веществ зависят от расходов воды; в-третьих, исходные ряды являются неэквидистентными; в четвертых, ряды наблюдений являются неоднородными по генезису формирования.

## 1.2. Статистические методы анализа результатов эксперимента

Статистические методы основаны на использовании накопленной статистической информации об изменении показателей, характеризующих анализируемый объект или процесс. Для анализа с использованием статистических методов необходимо, чтобы число наблюдений было достаточно большим, не менее 20–30, иначе достоверность выводов существенно снижается. При исследовании взаимосвязей между признаками на основе статистического анализа обычно решают следующие задачи:

1. Существует ли связь между результатом и выбранными для анализа факторами;
2. Какова количественная мера связи;
3. Какова аналитическая форма выражения связи;
4. Какова надёжность найденной закономерности и возможности использования параметров уравнения для решения оптимизационных моделей.

Ответ на первый вопрос дают дисперсионный и корреляционный анализ. Количественную меру зависимости определяют с помощью регрессионного анализа.

С теорией статистического оценивания параметров тесно связана проверка статистических гипотез. Она используется всякий раз, когда необходим обоснованный вывод о преимуществах того или иного способа инвестиций, измерений, стрельбы, технологического прогресса, об эффективности нового метода обучения, управления, о пользе вносимого удобрения, лекарства, об уровне доходности ценных бумаг, о значимости математической модели и т. д.

***Задачи о выборках: анализ распределений, сравнение, поиск зависимостей***

Анализ каждой произвольной выборки, представляющей собой совокупность независимых, одинаково распределенных случайных измерений, начинается с расчета описательных статистик эмпирического ряда: средних, дисперсии, основных моментов высшего порядка, медианы, моды, стандартного отклонения, ошибки среднего и др. Расчету элементарных статистик посвящено огромное множество литературы (Урбах, 1963; Смирнов, Дунин-Барковский, 1965; Крамер, 1975; Гнеденко, 1988; Калинина, Панкин, 2001; Ю. Прохоров, 2002). Рядом авторов (Браунли, 1977; Айвазян с соавт., 1983; Зайцев, 1984) предлагаются также специальные критерии,

предназначенные для оценки показателей вариации, точности опыта, репрезентативности и случайности выборок и т. д. Можно привести также некоторые ссылки на источники, где статистические методы рассматриваются в контексте использования по-пулярных пакетов прикладных программ (Тюрин, Макаров, 1995; Боровиков, 2001; Алексахин с соавт., 2002) или в виде руководства к использованию офисного табличного процессора Excel (Лапач с соавт., 2000).

Особое место в анализе выборок занимает проверка соответствия характера эмпирического распределения какому-нибудь заданному закону распределения (Кендалл, Стьюарт, 1966; Гмурман, 1972; Джонсон, Лион, 1980, 1981). Это связано с тем, что вид функции распределения часто постулируется как одно из важнейших предположений применения большинства статистических методов.

Разработанную в первой трети XX в. теорию называют параметрической статистикой (Плошко, Елисеева, 1990), поскольку ее основной объект изучения – это выборки из распределений, описываемых одним или небольшим числом параметров. Наиболее общим является семейство кривых Пирсона, задаваемых четырьмя параметрами (Елисеева, Юзбашев, 1995; Вентцель, 1999). Как правило, нельзя указать каких-либо веских причин, по которым конкретное распределение результатов экологических наблюдений должно входить в то или иное параметрическое семейство. В подавляющем большинстве реальных ситуаций таких предположений сделать нельзя, но, тем не менее, приближение реального распределения с помощью кривых из семейства Пирсона или его подсемейств часто не является чисто формальной операцией. Закономерности расчета описательных статистик в зависимости от распределения эмпирического ряда хорошо известны: если вероятностная модель основана на нормальном распределении, то расчет математического ожидания предусматривает суммирование независимых случайных величин; если же модель приближается к логарифмически нормальному распределению, то итог естественно описывать как произведение таких величин и т. д.

В первой же трети XX в., одновременно с параметрической статистикой, в работах Ч. Спирмена и М. Кендалла появились первые непараметрические методы, основанные на коэффициентах ранговой корреляции, носящих ныне имена этих статистиков (Кендалл, 1975; Руниoн, 1982; Холлендер, Вулф, 1983). Но непараметрика, не делающая нереалистических предположений о том, что функции распределения результатов наблюдений принадлежат тем или иным параметрическим семействам распределений, стала заметной частью статистики лишь со второй трети XX в. В 30-е годы появились работы А. Н. Колмогорова и Н. В. Смирнова, предложивших и изучивших статистические критерии, носящие в настоящее время их имена и основанные на использовании так называемого эмпирического процесса – разности между эмпирической и теоретической функциями распределения (Большев, Смирнов, 1968; Гублер, Генкин, 1973).

Во второй половине XX в. развитие непараметрической статистики пошло быстрыми темпами, в чем большую роль сыграли работы Ф. Вилкоксона и его школы (Гаек, Шидак, 1971). К настоящему времени с помощью непараметрических методов можно решать практически тот же круг статистических задач, что и с помощью параметрических (Никитин, 1995). Все бóльшую роль играют непараметрические оценки плотности вероятности, непараметрические методы регрессии и распознавания образов (дискриминантного анализа).

Тем не менее, параметрические методы всё еще популярнее непараметрических, хотя неоднократно публиковались обзоры (Налимов, 1960; Максимов с соавт., 1999), свидетельствующие о том, что распределения реально наблюдаемых случайных величин (в частности, биологических данных) в подавляющем большинстве случаев отличны от нормальных (гауссовских). Теоретики продолжают строить и изучать статистические модели, основанные на гауссовости, а практики – применять подобные методы и модели («ищут под фонарем, а не там, где потеряли»). Однако полностью игнорировать классические методы не менее вредно, чем переоценивать их. Поэтому целесообразно использовать одновременно оба подхода – и параметрические методы, и непараметрическую статистику. Такая рекомендация находится в согласии с концепцией математической устойчивости (Орлов, 1979), рекомендующей использовать различные методы для обработки одних и тех же данных с целью выделить выводы, получаемые одновременно при всех методах.

Любая выборка экологических данных является принципиально неоднородной, поскольку измерения могут осуществляться в различные временные периоды, разных пространственных точках водоема, с использованием различных инструментальных методов и т. д. В связи с этим, важным этапом математической обработки является дисперсионный анализ, с помощью которого оценивается, имеют ли место статистические различия между отдельными подмножествами данных и можно ли считать их принадлежащими одной генеральной совокупности (Плохинский, 1970; Лисенков, 1979; Джонсон, Лион, 1980, 1981, Любищев, 1986). Если каждому измерению поставлен в соответствие один признак (фактор), определяющий условия его реализации, то говорят об однофакторном дисперсионном анализе. Если таких группообразующих факторов больше одного, то выполняется многофакторный дисперсионный анализ (Плохинский, 1982; Афифи, Эйзен, 1982).

Если выборка состоит из двух рядов сопряженных наблюдений, измеренных в идентичных условиях, то решается задача регрессионного анализа, т. е. один эмпирический ряд объявляется результативным показателем или «откликом» $Y$, а другой – независимой варьируемой переменной $X$ или «фактором». Теория и практика одномерного регрессионного анализа также представлена многочисленными

литературными источниками (Хальд, 1956; Андерсен, 1963; Себер, 1980; Дрейпер, Смит, 1986; Дюк, 1997).

***Корректность математической обработки результатов эксперимента – залог достоверности научных положений по диссертации***

Эксперимент является важнейшим средством получения новых знаний не только в области естественных и технических наук, но и в экономике, социологии, политике, психологии, литературоведении и в других отраслях. Экспериментальные исследования дают критерии оценки обоснованности и приемлемости на практике любых теорий и теоретических предположений. Одним из основных этапов любого эксперимента является статистическая обработка экспериментальных данных. Она направлена, как правило, на построение математической модели исследуемого объекта или явления, а также на получение ответа на вопрос: «Достоверны ли полученные опытные данные в пределах требуемой точности или допусков?».

Сама же математическая модель в зависимости от целей эксперимента (исследование, управление, контроль) может быть использована для разных целей: для предметно-смыслового анализа объекта или явления, прогнозирования их состояния в разных условиях функционирования, управления ими в конкретных ситуациях, оптимизации отдельных параметров, а также для решения каких-то других специфичных задач. Особенно важна тщательная математическая обработка результатов экспериментов, подтверждающая теоретические выводы и построения по диссертациям на соискание ученых степеней.

Анализ результатов работы советов по защите диссертаций, а также экспертных советов ВАК Беларуси за последнее время свидетельствует о том, что для обработки экспериментальных данных не всегда выбираются методически обоснованные приемы, да и степень владения соискателями методиками такой обработки результатов экспериментальных исследований оставляет желать лучшего. Эксперты и оппоненты по диссертациям отмечают негативную тенденцию к снижению уровня подготовленности соискателей в понимании того, что они делают с помощью современной компьютерной техники при обработке опытных данных.

Применение статистических методов обработки экспериментальных данных, критериев достоверности и соответствия моделей изучаемым процессам или явлениям, оценка точности и надежности результатов эксперимента требует знания основных положений теории вероятностей и математической статистики, умелого использования принципов и приемов программирования. Кроме того, в связи с усложнением алгоритмов обработки данных необходимы глубокие знания основных вычислительных методов. Статистические методы, методы вычислительной математики и программирование в вузах традиционно изучаются раздельно, однако только при комплексном использовании полученных из этих курсов знаний можно достигнуть успеха. Анализ учебных планов подготовки студентов по разным специальностям свидетельствует о постоянном увеличении числа изучаемых

предметов. Это ведет к уменьшению числа учебных часов, выделяемых на общеобразовательные дисциплины, в том числе и математическую подготовку. С сожалением приходится констатировать факт все более усиливающейся прагматичности обучения, т. е. все дальше мы удаляемся от основополагающего принципа советской высшей школы по фундаментальности образования. В этих условиях в учебных и научных учреждениях и организациях, имеющих аспирантуру и докторантуру, необходимо введение специальных курсов по обучению особенностям применения математических методов для планирования эксперимента, сбора информации в виде экспериментальных данных по исследуемому объекту или явлению, а также по их последующей обработке с обеспечением требований надежности и точности.

Конечной целью любой обработки экспериментальных данных является выдвижение гипотез о классе и структуре математической модели исследуемого явления, определение состава и объема дополнительных измерений, выбор возможных методов последующей статистической обработки и анализ выполнения основных предпосылок, лежащих в их основе. Для ее достижения необходимо решить некоторые частные задачи, среди которых можно выделить следующие:

1. Анализ, выбраковка и восстановление аномальных (сбитых) или пропущенных измерений. Эта задача связана с тем, что исходная экспериментальная информация обычно неоднородна по качеству. В основной массе результатов прямых измерений, получаемых с возможно малыми погрешностями, в экспериментальных данных часто имеются грубые ошибки, вызванные разными причинами. К ним могут быть отнесены просчеты экспериментатора, сбои вычислительной техники, аномалии в работе измерительных приборов и т. д. Без глубокого анализа качества данных, устранения или хотя бы существенного уменьшения влияния аномальных данных на результаты последующей обработки можно сделать ложные выводы об изучаемом объекте или явлении.

2. Экспериментальная проверка законов распределения экспериментальных данных, оценка параметров и числовых характеристик наблюдаемых случайных величин или процессов. Выбор методов последующей обработки, направленной на построение и проверку адекватности математической модели исследуемому явлению, существенно зависит от закона распределения наблюдаемых величин. Предварительная обработка требует также и содержательного анализа изучаемого процесса, схемы и методики проведения эксперимента.

3. Группировка исходной информации при большом объеме экспериментальных данных. При этом должны быть учтены особенности их законов распределения, которые выявлены на предыдущем этапе обработки.

4. Объединение нескольких групп измерений, полученных, возможно, в различное время или в различных условиях, для совместной обработки.

5. Выявление статистических связей и взаимовлияния различных измеряемых факторов и результирующих переменных, последовательных измерений одних и тех же величин. Решение этой задачи позволяет отобрать те переменные, которые оказывают наиболее сильное влияние на результирующий признак. Выделенные факторы используются для дальнейшей обработки, в частности, методами регрессионного анализа. Анализ корреляционных связей делает возможным выдвижение гипотез о структуре взаимосвязи переменных и, в конечном итоге, о структуре модели объекта исследований.

В ходе предварительной обработки, кроме указанных выше задач, часто решают и другие, имеющие частный характер: отображение, преобразование и унификацию типа наблюдений, визуализацию многомерных данных и др.

К вычисляемым в результате эксперимента оценкам случайных величин предъявляются три основных требования: состоятельности, несмещенности и эффективности. Полагают, что оценка состоятельна, если с ростом объема выборки она стремится по вероятности к истинному значению, несмещена, если ее математическое ожидание стремится к истинному значению, и эффективна, когда оценка обладает наименьшим рассеянием по сравнению с любыми другими оценками. Из двух оценок эффективнее та, которая обладает меньшей дисперсией, т. е. значения которой рассеиваются в более узком интервале.

На уровень рассеивания оценок значительное влияние оказывают ошибки, имеющие место при эксперименте.

Как известно, при выборочном наблюдении встречаются ошибки трех видов: грубые, систематические и случайные. Грубые ошибки, отличающиеся большим отклонением от центра группирования выборки, отсеиваются на этапе первичного анализа материалов.

Точность измерений любой физической величины характеризуется, как известно, абсолютной и относительной ошибками, которые, в свою очередь, состоят из суммы систематических и случайных ошибок.

Систематические ошибки постоянны при определении каждого члена выборки и зависят от технического уровня измерительной аппаратуры и техники эксперимента. Эти ошибки можно свести к минимуму периодической тарировкой приборов с помощью более совершенных и повышением точности метода определения исследуемых переменных.

Случайные ошибки обусловлены влиянием большого количества факторов. Их появление неодинаково и случайно от измерения к измерению и не может быть предварительно учтено из-за их зависимости от изменения условий измерений и изменчивости самих измеряемых величин. Однако при достаточно большом количестве экспериментов суммарное значение случайных ошибок, изменяющихся примерно одинаково в положительную и отрицательную сторону, приближается к нулю.

***Дисперсионный анализ***

Дисперсионный анализ (от латинского *dispersio* –рассеивание) – статистический метод, позволяющий анализировать влияние различных факторов на исследуемую переменную. Метод был разработан биологом Р. Фишером в 1925 году и применялся первоначально для оценки экспериментов в растениеводстве. В дальнейшем выяснилась общенаучная значимость дисперсионного анализа для экспериментов в психологии, педагогике, медицине и др. Целью дисперсионного анализа является проверка значимости различия между средними с помощью сравнения дисперсий. Дисперсию измеряемого признака разлагают на независимые слагаемые, каждое из которых характеризует влияние того или иного фактора или их взаимодействия. Последующее сравнение таких слагаемых позволяет оценить значимость каждого изучаемого фактора, а также их комбинации. При истинности нулевой гипотезы (о равенстве средних в нескольких группах наблюдений, выбранных из генеральной совокупности), оценка дисперсии, связанной с внутригрупповой изменчивостью, должна быть близкой к оценке межгрупповой дисперсии. На практике часто возникают задачи более общего характера – задачи проверки существенности различий средних выборочных нескольких совокупностей. Например, требуется оценить влияние различного сырья на качество производимой продукции, решить задачу о влиянии количества удобрений на урожайность сельскохозяйственной продукции. Иногда дисперсионный анализ применяется, чтобы установить однородность нескольких совокупностей (дисперсии этих совокупностей одинаковы по предположению; если дисперсионный анализ покажет, что и математические ожидания одинаковы, то в этом смысле совокупности однородны). Однородные же совокупности можно объединить в одну и тем самым получить о ней более полную информацию, следовательно, и более надежные выводы. Дисперсионный анализ бывает однофакторный (в эксперименте участвует один фактор) и многофакторный (более одного фактора).

Однофакторный дисперсионный анализ. На первом этапе выделяются входной и результативный факторы. Для входного используются одновременно несколько вариантов (уровней) и каждому варианту входного фактора соответствует определённый вариант проведения опыта. Кроме того на каждом уровне входного фактора проводят несколько наблюдений, т. е. повторяют эксперимент (4 или 5 повторностей). Методом дисперсионного анализа выявляется значимость изучаемого фактора и количественно оценивается степень его влияния. Суть метода заключается в разложении рассеивания случайной величины на независимые слагаемые, каждое из которых характеризует влияние того или иного фактора или их взаимодействия (разложение обычной дисперсии на составляющие). Степень влияния изучаемых входных факторов оценивается долей воздействия, т. е. величиной соответствующей дисперсии. Доля вклада или рассеивания рассчитывается как отношение суммы квадратов слагаемого (составляющей) к общей сумме квадратов и выражается в процентах. Общая вариация результативного признака разложена на три составляющих:

1. Вариация фактора – это и есть воздействие фактора на результативный фактор (результат).

2. Оценивает закономерное действие неконтролируемых факторов внутри каждого из уровней изучаемого фактора от повторения к повторению. Это так называемая вариация повторений.

3. Остаток – оценивает вклад случайных незакономерных факторов и может служить оценкой точности опыта. То есть чем меньше 2 и 3, тем больше основной фактор.

Для оценки значимости действия изучаемого фактора рассматривается показатель статистики – Критерий Фишера (*F*-критерий). Сравниваются фактическое и теоретическое значение Фишера. Фактическое значение – это отношение дисперсий варианта к дисперсии остатка (вычисляем в программе). *F* – теоретическое для принятого в исследовании уровня значимости находят по таблицам с учётом числа степеней свободы для дисперсии вариантов и случайной дисперсии. В большинстве случаев используют 5 % уровень значимости. Чем больше $F_{факт.} > F_{теор.}$, тем более закономерный характер имеет действие изучаемого фактора (между признаками существует взаимосвязь). Если это равенство выполняется, то рассчитывают НСР (наименьшую существенную разность) и d (разность двух выборочных средних).

НСР – теоретическое значение, означающее возможную предельную ошибку разности двух средних (как эталон, с которым сравнивается фактическое значение). Если НСР > или = *d*, то взаимосвязь существенна или значима, если НСР < *d*, то несущественна.

***Корреляционный анализ***

Между переменными (случайными величинами) может существовать функциональная связь, проявляющаяся в том, что одна из них определяется как функция от другой. Но между переменными может существовать и связь другого рода, проявляющаяся в том, что одна из них реагирует на изменение другой изменением своего закона распределения. Такую связь называют стохастической. Она появляется в том случае, когда имеются общие случайные факторы, влияющие на обе переменные. В качестве меры зависимости между переменными используется коэффициент корреляции (*r*), который изменяется в пределах от –1 до +1. Если коэффициент корреляции отрицательный, это означает, что с увеличением значений одной переменной значения другой убывают. Если переменные независимы, то коэффициент корреляции равен 0 (обратное утверждение верно только для переменных, имеющих нормальное распределение). Но если коэффициент корреляции не равен 0 (переменные называются некоррелированными), то это значит, что между переменными существует зависимость. Чем ближе значение *r* к 1, тем зависимость сильнее. Коэффициент корреляции достигает своих предельных значений +1 или –1, тогда и только тогда, когда зависимость между переменными линейная. Корреляционный анализ позволяет установить силу и направление стохастической взаимосвязи между переменными (случайными величинами). Если переменные измерены, как минимум, в интервальной

шкале и имеют нормальное распределение, то корреляционный анализ осуществляется посредством вычисления коэффициента корреляции Пирсона, в противном случае используются корреляции Спирмена, тау Кендала, или Гамма.

***Регрессионный анализ***

При изучении процессов функционирования сложных систем приходится иметь дело с целым рядом одновременно действующих случайных величин. Для уяснения механизма явлений, причинно-следственных связей между элементами системы и т. д., по полученным наблюдениям мы пытаемся установить взаимоотношения этих величин.

В математическом анализе зависимость, например, между двумя величинами выражается понятием функции $y = f(x)$, где каждому значению одной переменной соответствует только одно значение другой. Такая зависимость носит название функциональной. Гораздо сложнее обстоит дело с понятием зависимости случайных величин. Как правило, между случайными величинами (случайными факторами), определяющими процесс функционирования сложных систем, обычно существует такая связь, при которой с изменением одной величины меняется распределение другой. Такая связь называется стохастической, или вероятностной.

Если дано распределение двух случайных величин $X$ и $Y$, то регрессией $Y$ по $X$ называется любая функция $g(X)$, приближенно представляющая статистическую зависимость $Y$ от $X$.

Понятие регрессии введено в науку по предложению английского ученого Ф. Гальтона. Уравнение регрессии составляется исследователем на основе характера связи между функцией и аргументами. При установлении тесноты связи между $Y$ и $X$ решается задача установления строгости соблюдения функциональной зависимости между изменениями $Y$ и $X$.

***Канонический анализ***

Канонический анализ предназначен для анализа зависимостей между двумя списками признаков (независимых переменных), характеризующих объекты. Например, можно изучить зависимость между различными неблагоприятными факторами и появлением определенной группы симптомов заболевания, или взаимосвязь между двумя группами клинико-лабораторных показателей (синдромов) больного. Канонический анализ является обобщением множественной корреляции как меры связи между одной переменной и множеством других переменных. Множественная корреляция есть максимальная корреляция между одной переменной и линейной функцией других переменных. Эта концепция была обобщена на случай связи между множествами переменных – признаков, характеризующих объекты. При этом достаточно ограничиться рассмотрением небольшого числа наиболее коррелированных линейных комбинаций из каждого множества. Пусть, например, первое множество переменных состоит из признаков $y_1, \ldots, y_p$, второе множество состоит из $x_1, \ldots, x_q$, тогда взаимосвязь между данными множествами можно оценить как корреляцию между линейными

комбинациями $a_1y_1 + a_2y_2 + ... + a_py_p$, $b_1x_1 + b_2x_2 + ... + b_qx_q$, которая называется канонической корреляцией. Задача канонического анализа в нахождении весовых коэффициентов таким образом, чтобы каноническая корреляция была максимальной.

***Методы сравнения средних***

В прикладных исследованиях часто встречаются случаи, когда средний результат некоторого признака одной серии экспериментов отличается от среднего результата другой серии. Так как средние это результаты измерений, то, как правило, они всегда различаются, вопрос в том, можно ли объяснить обнаруженное расхождение средних неизбежными случайными ошибками эксперимента или оно вызвано определенными причинами. Если идет речь о сравнении двух средних, то можно применять критерий Стьюдента (*t*-критерий). Это параметрический критерий, так как предполагается, что признак имеет нормальное распределение в каждой серии экспериментов. В настоящее время модным стало применение непараметрических критериев сравнения средних.

Сравнение средних результата один из способов выявления зависимостей между переменными признаками, характеризующими исследуемую совокупность объектов (наблюдений). Если при разбиении объектов исследования на подгруппы при помощи категориальной независимой переменной (предиктора) верна гипотеза о неравенстве средних некоторой зависимой переменной в подгруппах, то это означает, что существует стохастическая взаимосвязь между этой зависимой переменной и категориальным предиктором. Так, например, если установлено, что неверна гипотеза о равенстве средних показателей физического и интеллектуального развития детей в группах матерей, куривших и не куривших в период беременности, то это означает, что существует зависимость между курением матери ребенка в период беременности и его интеллектуальным и физическим развитием.

Наиболее общий метод сравнения средних дисперсионный анализ. В терминологии дисперсионного анализа категориальный предиктор называется фактором.

Дисперсионный анализ можно определить как параметрический, статистический метод, предназначенный для оценки влияния различных факторов на результат эксперимента, а также для последующего планирования экспериментов. Поэтому в дисперсионном анализе можно исследовать зависимость количественного признака от одного или нескольких качественных признаков факторов. Если рассматривается один фактор, то применяют однофакторный дисперсионный анализ, в противном случае используют многофакторный дисперсионный анализ.

Частотный анализ. Таблицы частот, или как еще их называют одновходовые таблицы, представляют собой простейший метод анализа категориальных переменных. Таблицы частот могут быть с успехом использованы также для исследования количественных переменных, хотя при

этом могут возникнуть трудности с интерпретацией результатов. Данный вид статистического исследования часто используют как одну из процедур разведочного анализа, чтобы посмотреть, каким образом различные группы наблюдений распределены в выборке, или как распределено значение признака на интервале от минимального до максимального значения. Как правило, таблицы частот графически иллюстрируются при помощи гистограмм.

Кросстабуляция (сопряжение) – процесс объединения двух (или нескольких) таблиц частот так, что каждая ячейка в построенной таблице представляется единственной комбинацией значений или уровней табулированных переменных. Кросстабуляция позволяет совместить частоты появления наблюдений на разных уровнях рассматриваемых факторов. Исследуя эти частоты, можно выявить связи между табулированными переменными и исследовать структуру этой связи. Обычно табулируются категориальные или количественные переменные с относительно небольшим числом значений. Если надо табулировать непрерывную переменную (предположим, уровень сахара в крови), то вначале ее следует перекодировать, разбив диапазон изменения на небольшое число интервалов (например, уровень: низкий, средний, высокий).

Анализ соответствий. Анализ соответствий по сравнению с частотным анализом содержит более мощные описательные и разведочные методы анализа двухвходовых и многовходовых таблиц. Метод, так же, как и таблицы сопряженности, позволяет исследовать структуру и взаимосвязь группирующих переменных, включенных в таблицу. В классическом анализе соответствий частоты в таблице сопряженности стандартизуются (нормируются) таким образом, чтобы сумма элементов во всех ячейках была равна 1.

Одна из целей анализа соответствий – представление содержимого таблицы относительных частот в виде расстояний между отдельными строками и/или столбцами таблицы в пространстве более низкой размерности.

***Кластерный анализ***

Кластерный анализ – это метод классификационного анализа; его основное назначение – разбиение множества исследуемых объектов и признаков на однородные в некотором смысле группы, или кластеры. Это многомерный статистический метод, поэтому предполагается, что исходные данные могут быть значительного объема, т. е. существенно большим может быть как количество объектов исследования (наблюдений), так и признаков, характеризующих эти объекты. Большое достоинство кластерного анализа в том, что он дает возможность производить разбиение объектов не по одному признаку, а по ряду признаков. Кроме того, кластерный анализ в отличие от большинства математико-статистических методов не накладывает никаких ограничений на вид рассматриваемых объектов и позволяет исследовать множество исходных данных практически произвольной природы. Так как кластеры – это группы однородности, то задача кластерного анализа

заключается в том, чтобы на основании признаков объектов разбить их множество на $m$ ($m$ – целое) кластеров так, чтобы каждый объект принадлежал только одной группе разбиения. При этом объекты, принадлежащие одному кластеру, должны быть однородными (сходными), а объекты, принадлежащие разным кластерам, – разнородными. Если объекты кластеризации представить как точки в n-мерном пространстве признаков ($n$ – количество признаков, характеризующих объекты), то сходство между объектами определяется через понятие расстояния между точками, так как интуитивно понятно, что чем меньше расстояние между объектами, тем они более схожи.

***Дискриминантный анализ***

Дискриминантный анализ включает статистические методы классификации многомерных наблюдений в ситуации, когда исследователь обладает так называемыми обучающими выборками. Этот вид анализа является многомерным, так как использует несколько признаков объекта, число которых может быть сколь угодно большим. Цель дискриминантного анализ состоит в том, чтобы на основе измерения различных характеристик (признаков) объекта классифицировать его, т. е. отнести к одной из нескольких заданных групп (классов) некоторым оптимальным способом. При этом предполагается, что исходные данные наряду с признаками объектов содержат категориальную (группирующую) переменную, которая определяет принадлежность объекта к той или иной группе. Поэтому в дискриминантном анализе предусмотрена проверка непротиворечивости классификации, проведенной методом, с исходной эмпирической классификацией. Под оптимальным способом понимается либо минимум математического ожидания потерь, либо минимум вероятности ложной классификации. В общем случае задача различения (дискриминации) формулируется следующим образом. Пусть результатом наблюдения над объектом является построение $k$-мерного случайного вектора $X = (X_1, X_2, \ldots, X_k)$, где $X_1, X_2, \ldots, X_k$ – признаки объекта. Требуется установить правило, согласно которому по значениям координат вектора $X$ объект относят к одной из возможных совокупностей $i$, $i = 1, 2, \ldots, n$. Методы дискриминации можно условно разделить на параметрические и непараметрические. В параметрических известно, что распределение векторов признаков в каждой совокупности нормально, но нет информации о параметрах этих распределений. Непараметрические методы дискриминации не требуют знаний о точном функциональном виде распределений и позволяют решать задачи дискриминации на основе незначительной априорной информации о совокупностях, что особенно ценно для практических применений. Если выполняются условия применимости дискриминантного анализа – независимые переменные–признаки (их еще называют предикторами) должны быть измерены как минимум в интервальной шкале, их распределение должно соответствовать нормальному закону, необходимо воспользоваться классическим дискриминантным анализом, в противном случае – методом общие модели дискриминантного анализа.

***Факторный анализ***

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 классификация 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.

***Деревья классификации***

Деревья классификации – это метод классификационного анализа, позволяющий предсказывать принадлежность объектов к тому или иному классу в зависимости от соответствующих значений признаков, характеризующих объекты. Признаки называются независимыми переменными, а переменная, указывающая на принадлежность объектов к классам, называется зависимой. В отличие от классического дискриминантного анализа, деревья классификации способны выполнять одномерное ветвление по переменным различных типов категориальным, порядковым, интервальным. Не накладываются какие-либо ограничения на закон распределения количественных переменных. По аналогии с дискриминантным анализом метод дает возможность анализировать вклады отдельных переменных в процедуру классификации. Структура метода такова, что пользователь имеет возможность по управляемым параметрам строить деревья произвольной сложности, добиваясь минимальных ошибок классификации. Но по сложному дереву, из-за большой совокупности решающих правил, затруднительно классифицировать новый объект. Поэтому при построении дерева классификации пользователь должен найти разумный компромисс между сложностью дерева и трудоемкостью процедуры классификации.

***Анализ главных компонент и классификация***

На практике часто возникает задача анализа данных большой размерности. Метод анализ главных компонент и классификация позволяет решить эту задачу и служит для достижения двух целей:

– уменьшение общего числа переменных (редукция данных) для того, чтобы получить «главные» и «некоррелирующие» переменные;

– классификация переменных и наблюдений, при помощи строящегося факторного пространства.

Метод имеет сходство с факторным анализом в постановочной части решаемых задач, но имеет ряд существенных отличий:

– при анализе главных компонент не используются итеративные методы для извлечения факторов;

– наряду с активными переменными и наблюдениями, используемыми для извлечения главных компонент, можно задать вспомогательные переменные и/или наблюдения; затем вспомогательные переменные и наблюдения проектируются на факторное пространство, вычисленное на основе активных переменных и наблюдений;

– перечисленные возможности позволяют использовать метод как мощное средство для классификации одновременно переменных и наблюдений.

Решение основной задачи метода достигается созданием векторного пространства латентных (скрытых) переменных (факторов) с размерностью меньше исходной. Исходная размерность определяется числом переменных для анализа в исходных данных.

***Многомерное шкалирование***

Метод можно рассматривать как альтернативу факторному анализу, в котором достигается сокращение числа переменных, путем выделения латентных (непосредственно не наблюдаемых) факторов, объясняющих связи между наблюдаемыми переменными. Цель многомерного шкалирования – поиск и интерпретация латентных переменных, дающих возможность пользователю объяснить сходства между объектами, заданными точками в исходном пространстве признаков. Показателями сходства объектов на практике могут быть расстояния или степени связи между ними. Основное предположение многомерного шкалирования заключается в том, что существует некоторое метрическое пространство существенных базовых характеристик, которые неявно и послужили основой для полученных эмпирических данных о близости между парами объектов. Следовательно, объекты можно представить как точки в этом пространстве. Предполагают также, что более близким (по исходной матрице) объектам соответствуют меньшие расстояния в пространстве базовых характеристик. Поэтому, многомерное шкалирование – это совокупность методов анализа эмпирических данных о близости объектов, с помощью которых определяется размерность пространства существенных для данной содержательной задачи характеристик измеряемых объектов и конструируется конфигурация точек (объектов) в этом пространстве. Это пространство («многомерная шкала») аналогично обычно используемым шкалам в том смысле, что значениям существенных характеристик измеряемых объектов соответствуют определенные позиции на осях пространства.

***Методы анализа выживаемости***

Методы анализа выживаемости первоначально были развиты в медицинских, биологических исследованиях и страховании, но затем стали широко применяться в социальных и экономических науках, а также в промышленности в инженерных задачах (анализ надежности и времен отказов). Представьте, что изучается эффективность нового метода лечения

или лекарственного препарата. Очевидно, наиболее важной и объективной характеристикой является средняя продолжительность жизни пациентов с момента поступления в клинику или средняя продолжительность ремиссии заболевания. Для описания средних времен жизни или ремиссии можно было бы использовать стандартные параметрические и непараметрические методы. Однако в анализируемых данных есть существенная особенность – могут найтись пациенты, которые в течение всего периода наблюдения выжили, а у некоторых из них заболевание все еще находится в стадии ремиссии. Также может образоваться группа больных, контакт с которыми был потерян до завершения эксперимента (например, их перевели в другие клиники). При использовании стандартных методов оценки среднего эту группу пациентов пришлось бы исключить, тем самым, потеряв с трудом собранную важную информацию. К тому же большинство этих пациентов являются выжившими (выздоровевшими) в течение того времени, которое их наблюдали, что свидетельствует в пользу нового метода лечения (лекарственного препарата). Такого рода информация, когда нет данных о наступлении интересующего нас события, называется неполной. Если есть данные о наступлении интересующего нас события, то информация называется полной. Наблюдения, которые содержат неполную информацию, называются цензурированными наблюдениями. Цензурированные наблюдения типичны, когда наблюдаемая величина представляет время до наступления некоторого критического события, а продолжительность наблюдения ограничена по времени. Использование цензурированных наблюдений составляет специфику рассматриваемого метода – анализа выживаемости. В данном методе исследуются вероятностные характеристики интервалов времени между последовательным возникновением критических событий. Такого рода исследования называются анализом длительностей до момента прекращения, которые можно определить как интервалы времени между началом наблюдения за объектом и моментом прекращения, при котором объект перестает отвечать заданным для наблюдения свойствам. Цель исследований – определение условных вероятностей, связанных с длительностями до момента прекращения. Построение таблиц времен жизни, подгонка распределения выживаемости, оценивание функции выживания с помощью процедуры Каплана–Мейера относятся к описательным методам исследования цензурированных данных. Некоторые из предложенных методов позволяют сравнивать выживаемость в двух и более группах. Наконец, анализ выживаемости содержит регрессионные модели для оценивания зависимостей между многомерными непрерывными переменными со значениями, аналогичными временам жизни.