Министерство образования Республики Беларусь
Белорусский государственный университет
Факультет географии и геоинформатики
Кафедра почвоведения и геоинформационных систем

СОГЛАСОВАНО
Заведующий кафедрой
_________________Клебанович Н. В.
«29» марта 2021 г.

СОГЛАСОВАНО
Декан факультета
______________ Курлович Д. М.
«28» апреля 2021 г.

СОГЛАСОВАНО
Председатель
учебно-методической комиссии факультета
_________________ Кольмакова Е. Г.
«28» апреля 2021 г.

Математические методы в землеустройстве

Электронный учебно-методический комплекс для специальности:
1-56 02 02 «Геоинформационные системы (по направлениям)»
направления специальности:
1-56 02 02-01 «Геоинформационные системы (земельно-кадастровые)»,
1-56 02 02-02 «Геоинформационные системы (специальные)»

Регистрационный № 2.4.2-12/181

Автор:
Карпиченко А. А., кандидат географических наук, доцент

Рассмотрено и утверждено на заседании Научно-методического совета БГУ 27.08.2021 г., протокол № 10.

Минск 2021

УДК 332.3:330.43(075.8)+528.46(075.8)
К 264

Утверждено на заседании Научно-методического совета БГУ.
Протокол № 10 от 27.08.2021 г.

Решение о депонировании вынес:
Совет факультета географии и геоинформатики
Протокол № 8 от 28.04.2021 г.

Автор:

Карпиченко Александр Александрович, кандидат географических наук, доцент, доцент кафедры почвоведения и геоинформационных систем БГУ.

Рецензенты:

кафедра географии и методики преподавания географии УО «Белорусский государственный педагогический университет им. Максима Танка» (зав. кафедрой Таранчук А. В., кандидат географических наук, доцент);

Камышенко Г. А., ученый секретарь Института природопользования НАН Беларуси, кандидат технических наук.

Карпиченко, А. А. Математические методы в землеустройстве : электронный учебно-методический комплекс для специальности: 1-56 02 02 «Геоинформационные системы (по направлениям)», направления специальности: 1-56 02 02-01 «Геоинформационные системы (земельно-кадастровые)», 1-56 02 02-02 «Геоинформационные системы (специальные)» / А. А. Карпиченко ; БГУ, Фак. географии и геоинформатики, Каф. почвоведения и геоинформационных систем. – Минск : БГУ, 2021. – 50 с. : табл. – Библиогр.: с. 47–48.

Электронный учебно-методический комплекс (ЭУМК) предназначен для студентов, обучающихся по специальности 1-56 02 02 «Геоинформационные системы». Содержание ЭУМК посвящено методическим правилам применения математических методов для решения землеустроительных задач, способствует формированию у студентов навыков в области математического анализа и моделирования, обработки результатов полевых и экспериментальных исследований, оценки степени их достоверности.

ОГЛАВЛЕНИЕ

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Электронный учебно-методический комплекс по учебной дисциплине «Математические методы в землеустройстве» предназначен для реализации требований образовательных программ, образовательного стандарта и учебного плана специальности 1-56 02 02 «Геоинформационные системы (по направлениям)», направления специальности 1-56 02 02-01 «Геоинформационные системы (земельно-кадастровые)», 1-56 02 02-02 «Геоинформационные системы (специальные)». Его наличие обеспечивает стабильность качества образовательного процесса и является методической основой для обеспечения эффективной самостоятельной работы студентов при обучении и проведении научных исследований.

ЭУМК по учебной дисциплине «Математические методы в землеустройстве» создан на научно-методическом и программно-техническом уровнях, соответствующих современным информационно-коммуникационным технологиям и призван обеспечить реализацию учебных целей и задач на всех этапах образовательного процесса по данной дисциплине.

Назначение – реализация требований образовательного стандарта и учебной программы, обеспечение непрерывности и полноты процесса обучения, систематизации и контроля знаний по учебной дисциплине «Математические методы в землеустройстве».

Цель ЭУМК – повышение эффективности управления образовательным процессом и самостоятельной работой студентов по освоению учебной дисциплины «Математические методы в землеустройстве» с помощью внедрения в образовательный процесс инновационных образовательных технологий и современного программного обеспечения для подготовки высококвалифицированных специалистов.

Область применения – на лабораторных занятиях по курсу «Математические методы в землеустройстве», в ходе самостоятельной подготовки к аудиторным занятиям, текущему и итоговому контролю знаний по разделам дисциплины, ориентация в выполнении управляемой самостоятельной работы.

Функциональные возможности ЭУМК – средство ориентации в содержании дисциплины «Математические методы в землеустройстве» и порядке изучения учебного материала, освоение теоретического и практического материала, подготовка к контролю знаний. Материал ЭУМК структурирован по разделам таким образом, чтобы обеспечить возможность самостоятельного овладения знаниями по учебной дисциплине. ЭУМК по учебной дисциплине «Математические методы в землеустройстве» включает 4 основных раздела: теоретический, практический, контроля знаний и вспомогательный.

Теоретический раздел ЭУМК содержит конспект лекций для теоретического изучения учебной дисциплины на основе учебно-методического пособия

Чертко, Н.К. Математические методы в землеустройстве [Электронный ресурс]: учебно-методическое пособие / Н. К. Чертко. – Минск: БГУ, 2014. – Режим доступа: https://elib.bsu.by/handle/123456789/114777 – Дата доступа 15.03.2021.

Практический раздел ЭУМК включает методические рекомендации и варианты заданий для проведения лабораторно-практических занятий на основе учебно-методического пособия

Чертко, Н.К. Математические методы в землеустройстве [Электронный ресурс]: учебно-методическое пособие / Н. К. Чертко. – Минск: БГУ, 2014. – Режим доступа: https://elib.bsu.by/handle/123456789/114777 – Дата доступа 15.03.2021.

Раздел контроля знаний ЭУМК содержит материалы к контролю знаний и к аттестации, позволяющие определить соответствие результатов учебной деятельности обучающихся требованиям образовательного стандарта и учебно-программной документации по специальности. Данный раздел включает: варианты контрольных заданий, вопросы к зачету, перечень заданий и контрольных мероприятий управляемой самостоятельной работы.

Вспомогательный раздел ЭУМК содержит ссылки на учебную программу по учебной дисциплине «Математические методы в землеустройстве».

ЭУМК по учебной дисциплине «Математические методы в землеустройстве» предназначен для студентов, магистрантов и аспирантов, занимающихся исследованиями в области земельных ресурсов.

Дисциплина «Математические методы в землеустройстве» позволяет сформировать знания, умения и навыки в области решения землеустроительных задач, позволяет методически грамотно отобрать необходимую статистическую информацию и провести ее обработку с использованием соответствующих методов (дисперсионного, корреляционного, регрессионного, факторного, кластерного или линейного программирования).

1. ТЕОРЕТИЧЕСКИЙ РАЗДЕЛ

КОНСПЕКТ ЛЕКЦИЙ

1.1. Введение

Математические методы – специальная научная и учебная дисциплина, предметом изучения которой являются выборочные совокупности и их оценка. Варианты (результаты наблюдения или эксперимента) в выборочной совокупности не имеют постоянных, одних и тех же исходов. Однако многие хаотичные явления имеют упорядоченную структуру, поэтому могут иметь конкретные численные оценки. Главное условие для этого – статистическая устойчивость этих явлений, т. е. существование определенных закономерностей, которые можно описать математическими моделями статистически. Ведь большинству природных и экономических явлений свойственна вариабельность (изменение во времени в определенных пределах).

Большинство математических методов универсальны и применяются в различных отраслях деятельности человека. Поэтому многие компьютерные статистические программы не выступают чисто землеустроительными прикладными программами. Иногда выполнение отдельных функций в таких программах не является по сути статистическим. При этом следует помнить, что механический подход при использовании математических методов недопустим, поскольку каждый из методов анализа имеет свои возможности и ограниченную область применения.

Выбор конкретных методов статистического анализа и математических методов определяется целью и задачами исследования. Цель большинства математических методов и статистики – установление различия и на его основе проведение классификации, раскрытие взаимосвязи и оценка влияния факторов на состояние или развитие объекта или явления. Итогом таких исследований является изучение тенденций и закономерностей, которые проявляются в массе наблюдений и не могут быть достоверно проанализированы в отдельно взятых случаях.

Чем сложнее система, тем больше рассматриваемых взаимосвязанных переменных, тем труднее установить множество отношений. Количественные методы анализа помогают выбрать ведущие факторы, причины, признаки. В таких случаях важно понимание смысла математических методов, чтобы не допустить ошибочных выводов при интерпретации полученных результатов. Начинать изучение системы необходимо с усвоения методологических основ организации самих исследований и важнейших элементов системологии, которые определяют последовательность дальнейших действий.

Системный подход основан на исследовании объектов как систем, создает единую теоретическую модель.

Системный анализ представляет собой совокупность методологических средств, позволяющих обосновать проблемы научно-практического характера.

Успешное использование системного анализа возможно при реализации следующих важнейших принципов, опирающихся на математические методы:

выявляется и формулируется конечная цель исследования; система-объект рассматривается как единое целое, в ней выявляются все взаимосвязи и их результаты; строится обобщенная комбинированная модель (модели), где отображаются структура, иерархия и взаимосвязи.

Заметну часть статистических расчетов можно выполнять в программе Microsoft Excel, для более сложных видов анализа удобнее применять специализированное программное обеспечение – Statistica, SPSS, Mathematica и др.

1.2. Основы математической статистики

Источником данных для статистической обработки могут служить материалы, полученные при проведении эксперимента, а также официальная статистическая информация об объекте, фондовые материалы предприятий, землеустроительные и специальные карты, аэро- и космофотоснимки. При этом, чем больший регион занимает объект исследования, тем чаще используются карты, обобщающие материалы, литературные источники.

Понятия генеральной совокупности и выборки. Первичным элементом в статистике является единица наблюдения (варианта, дата). Ряд единиц наблюдения образует статистическую совокупность, которая характеризует определенный признак объекта исследования. По виду исследуемые признаки могут быть качественными и количественными. Количественные признаки имеют числовое выражение, качественные – словесное (плодородие: низкое, среднее, высокое). Для группировки качественных признаков используются альтернативная шкала и шкала рангов (шкала баллов, номиналов, рангов). При статистической сводке все качественные признаки кодируются (1, 2, 3) или им присваивается ранг (1, 2, 3 и т. д.), они приобретают количественное выражение. Ранги, баллы и другие показатели качественных признаков относятся к *непараметрическим*, необходимые величины рассчитываются по специальным формулам. Их нельзя включать в статистическую обработку вместе с количественными характеристиками изучаемого объекта, как и нельзя вычислять средние ранговые величины.

По времени наблюдение может быть текущим (непрерывным) и единовременным (в один и тот же момент времени в разных точках – хозяйства в различных областях). По охвату исследование может быть сплошное и не сплошное. Эта особенность определяет ход и методику статистического анализа.

Сплошное статистическое исследование (вся площадь земель Республики Беларусь) образует *генеральную совокупность*. Общее число членов генеральной совокупности называют *объемом генеральной совокупности*.

Из-за больших размеров генеральной совокупности (совокупность всех сельскохозяйственных предприятий страны) или из-за отсутствия определенных границ этой совокупности (Европа) сплошные исследования проводятся достаточно редко, поскольку требую много временных и финансовых затрат. В связи с этим в большинстве случаев ограничиваются методом *выборочного исследования* (не сплошного) из *генеральной совокупности.* Выборка образует совокупность наблюдений, полученных с целью объективной характеристики и получения информации о генеральной совокупности. Число ее членов называют *объемом выборочной совокупности*, который обозначается как *N*, *n*.

Отбор объектов для формирования выборочной совокупности производится следующими основными методами отбора: *случайным, направленным (типическим), смешанным*.

При случайном отборе все объекты должны иметь одинаковую возможность попасть в выборку. Для этого можно использовать таблицу случайных чисел (прил. таблица 2 в «Математические методы в землеустройстве»). По второму варианту из объектов в списке (алфавитном или ином) в выборку включают каждый третий или пятый, или десятый (механическая выборка) и т. д.

Случайная выборка может не отвечать условиям исследования из-за неоднородности. Тогда производят целенаправленный (когортный) отбор, выбирая для исследования типичные объекты. Правила отбора остаются те же, что и при случайном отборе.

Смешанный отбор производят в том случае, когда необходимо выбрать варианты для неоднородного объекта, например, хозяйства большие и малые, разной специализации и т. д.

Соблюдения правил составления выборки дают возможность наиболее полно и точно, т. е. репрезентативно, характеризовать генеральную совокупность. Репрезентативность выборочной совокупности бывает количественной и качественной (структурной). Количественная репрезентативность определяется числом наблюдений, гарантирующих получение статистически достоверных данных. Здесь действует основной постулат закона больших чисел: чем большее число наблюдений, тем больше значений характеристик выборочной совокупности приближаются к аналогичным значениям характеристик генеральной совокупности. Качественная репрезентативность обозначает структурное соответствие выборочной и генеральной совокупностей. Например, если в составе генеральной совокупности фермерские хозяйства составляют 50 %, то и в выборочной группе их должно быть 50 %.

Величина ошибки репрезентативности зависит от изменчивости изучаемого признака. Чем больше разброс значений изучаемого признака, тем больше статистическая ошибка. Отбор для выборки должен быть также научно обоснованным с учетом принятых методических правил, т. е. рандомизированным. В таких случаях при меньшем числе наблюдений уменьшается вероятность систематических ошибок наблюдений.

На втором этапе статистического исследования проводят сводку и группировку данных, применяются следующие варианты группировок: разделение анализируемой статистической совокупности на группы по тем или иным признакам; объединение мелких однородных групп в более крупные; комплексная группировка на основе многих учетных признаков, даже если они разнородные.

Типологическая группировка выделяет в совокупности качественно однородные в существенном отношении группы. Группировка по своей сути представляет собой процесс классификации. В государственной статистике используют классификаторы – специальные справочники, инструкции, указания.

Важным при исследованиях является определение минимально необходимого объема наблюдений в исследованиях, который обеспечивает достоверное представление о характере изменчивости признака в генеральной совокупности. Оптимальный объем выборки обычно прямо пропорционален степени изменчивости признака – чем она больше, тем большее количество наблюдений требуется. Если объект исследуется впервые, то определить объем выборки достаточно трудно, поэтому часто достаточно точные результаты получают при объеме выборки около 100.

Необходимая величина выборки может определяться по таблице достаточно больших чисел (прил. таблица 1 учебно-методического пособия), а также расчетным способом по формулам, исходя из величины допускаемой вероятности, с какой предполагается делать заключения, и величины точности опыта. Например, при допускаемом уровне вероятности $P = 0{,}95$ (95 %) и точности опыта $p = 5$ % число наблюдений по таблице достаточно больших чисел составит 384. Если точность опыта увеличить до 1 %, то число наблюдений следует увеличить до 9603.

Чаще всего ориентировочный объем (N) выборочной совокупности рассчитывают по формулам, в которых вероятность заменяют степенью варьирования:

$$N = \sigma^2 / m_M^2\,,$$

где σ – среднее квадратическое отклонение; m_M – ошибка среднего арифметического.

Допустим, варьирование признака (колебание температуры) составляет 7 °С, тогда число наблюдений выборочной совокупности с ошибкой среднего арифметического $m = \pm\,0{,}5$ °С составит: $N = \sigma^2 / m_M^2 = 7^2 / 0{,}5^2 = 196$.

Объем выборочной совокупности можно также определить по ожидаемому коэффициенту вариации (V) и точности опыта (p) с учетом поправочного коэффициента (1,96) для уровня вероятности 0,95 и 0,99:

$$N = (1{,}96 \cdot V)^2 / p^2.$$

Пример. Для расчета коэффициента увлажнения в зависимости от количества выпадающих осадков и испарения с ожидаемой точностью опыта 3 % и коэффициента вариации 30 % потребуется следующий объем выборочной совокупности $N = (1{,}96 \cdot 30)^2 / 3^2 = 384$.

Определение объема выборочной совокупности необходимо для получения достоверной информации о генеральной совокупности путем расчета минимального, но объективного количества наблюдений. Полученные параметры по выборке могут служить приблизительными оценками аналогичных параметров генеральной совокупности, т. е. указывать пределы, в которых они заключены ($M \pm m_М$; $\sigma \pm m_\sigma$).

Формирование и обработка вариационного ряда. Набор цифр в последовательности их увеличения или уменьшения представляет собой *вариационный ряд*. Перед началом вычислений статистических показателей рекомендуется выполнить поиск *артефактов* – вариант, которые резко отличаются от вариантов статистической совокупности и вызывают сомнение у исследователя. Они располагаются в начале или в конце вариационного ряда. Например, в приведенных вариационных рядах: 2, 9, 11, 12, 13, 15 и 25, 27, 29, 32, 55 почти все соседние показатели весьма близки по значению. Вызывают сомнение варианты 2 в первом ряду и 55 – во втором. Их можно принять за артефакт и исключить (выбраковать) из обработки. Выбраковка должна быть статистически доказана, в таком случае артефакт исключается из статистической совокупности и не подлежит обработке.

Существующие критерии выбраковки основываются, как правило, на допущении, что выборка распределяется по нормальному или близкому к нему закону. В качестве критерия выбраковки может быть использован критерий τ (прил. таблица 3 пособия). Если критерий τ вычисленный (фактический) больше или равен критерию τ табличному ($\tau_ф \geq \tau_т$) при объеме выборки N и уровне значимости α (0,05 или 0,01), то соответствующие значения вариантов выборки (x) допустимо отбросить как артефакт. Значения τ для вызывающей сомнение величины вычисляются по следующим формулам:

$$\tau_1 = (x_2 - x_1) / (x_{n-1} - x_1)$$

для наименьшего значения переменной величины в вариационном ряду (x_1);

$$\tau_n = (x_n - x_{n-1}) / (x_n - x_2)$$

для максимального значения переменной в вариационном ряду.

Пример. При составлении вариационного ряда по урожайности сельскохозяйственных культур в разрезе хозяйств одного из районов получен следующий ряд значений: 10,8; 12,5; 12,9; 13,2; 20,2 (ц/га). Вызывает сомнение максимальное значение в выборке варианты 20,2. Следует доказать, можно ли ее отнести к артефакту. Подставляем необходимые данные во вторую формулу:

$$\tau_5 = (x_5 - x_4) / (x_5 - x_2) = (20{,}2 - 13{,}2) / (20{,}2 - 12{,}5) = 0{,}958.$$

Вычисленное значение критерия (τ_5 = 0,958) сравнивают с табличным значением ($\tau_т$), учитывая объем выборки (N = 5). В прил. таблица 3 пособия критическое значение критерия артефакта для N = 5 и уровня значимости α 0,05 и 0,01 соответственно будут равны 0,807 и 0,916, что меньше расчетного значения (τ_5 = 0,958). Поэтому варианту 20,2 признают артефактом и исключают из статистической обработки как сомнительную.

После проверки на наличие артефактов приступают к вычислению показателей описательной статистики при условии, что *тип распределения* вариант соответствует *нормальному или логнормальному закону распределения.* В иных случаях с выборкой работают как с *непараметрической*, на которые теория вероятностей не распространяется и ряд параметрических статистических показателей и критериев вычислять не рекомендуется.

Вариационный ряд можно показать графически с помощью кривой распределения и гистограммы. Для построения гистограммы необходимо разбить вариационный ряд на интервалы. Для этого сначала определяется величина классового интервала i, которая зависит от принятого числа классов k и объема выборки N:

$$i = (x_{max} - x_{min}) / k.$$

Число классов в зависимости от объема выборки определяется по формуле:

$$k = 1 + 3{,}3 \lg N.$$

Можно рекомендовать следующее число классов в зависимости от объема выборки:

N	30–50	51–100	101–400	401–1000	1001–2000
k	4–5	6–7	8–9	9–10	11–12

Величина классового интервала должна быть одинаковой на протяжении всего вариационного ряда. Границы классов выбираются такими, чтобы каждая варианта могла быть отнесена только к одному классу. Примеры правильной границы классов: 5–9, 10–14, 15–19 или 5,1–9,1, 9,2–13,2, 13,3–17,3, первый и последний классы могут быть неполными. Границы классов желательно выбирать так, чтобы крайние варианты ряда по возможности оказались ближе к середине интервала своего класса.

На отклонение от нормального распределения могут указывать *показатели асимметрии и эксцесса*. Распределение частот в изучаемом объекте не всегда подчиняется закону нормального распределения. Это особенно четко проявляется при выражении вариационного ряда в виде графика. Распределение частот может быть представлено *асимметричной*, *островершинной* или *туповершинной* кривой.

Асимметрия кривой распределения обусловлена неравномерным размещением вариант по обе стороны от модального значения признака. Если число вариант больше справа от моды, распределение имеет положительную асимметрию, если слева – отрицательную.

При получении асимметричной кривой следует проверить асимметричность распределения. Если асимметричность не будет доказана по критерию Стьюдента, то рассматриваемое распределение относят к симметричному. Для проверки асимметричности распределения вычисляют коэффициент асимметрии, его ошибку, затем на основании показателя достоверности устанавливают вид кривой распределения. Коэффициент асимметрии находят:

$$K_{as} = (M - \text{Mo}) / \sigma, \text{ или } K_{as} = (M - \text{Me}) / \sigma.$$

Эксцесс кривой распределения (E) имеет место в тех случаях, когда большинство вариантов совокупности сосредоточено около среднего арифметического. Тогда эмпирическая кривая распределения отклоняется от нормальной теоретической кривой у ее вершины и количественно выражается показателем эксцесса.

Положительный эксцесс представлен островершинной кривой (эксцессивной, или лептокуртичной), отрицательный – плосковершинной (депрессивной, или платикуртичной). При сильном отрицательном эксцессе кривая может приобрести вид двухвершинной.

Показатель эксцесса определяется по формуле:

$$E = [\sum (x - M)^4 / N \cdot \sigma^4] - 3.$$

Вычисляют ошибку коэффициента эксцесса: $m_E = 2\sqrt{6/(N+5)}$.

Оценка достоверности показателя эксцесса производится аналогично оценке показателя асимметрии по критерию Стьюдента: $t = E / m_E$.

Оценить достоверность показателей эксцесса и асимметрии можно более простым способом. Отклонение эмпирического ряда по асимметрии и эксцессу от нормального распределения считают существенным, если K_{as} и E более, чем в 3 раза превышают свои ошибки (m_{as}, m_E). Если показатель эксцесса меньше –2, это указывает на возможное наличие в выборке вариант, относящихся к разным совокупностям. Эксцесс считается незначительным, если он по модулю меньше 0,4 (более точно это определяют с использованием критерия Стьюдента). Чем меньше показатель эксцесса, тем ближе распределение к нормальному.

Асимметрия и эксцесс эмпирических кривых указывают иногда на важные особенности объекта исследования, например, на изменение признака в ходе усовершенствования технологии на сельскохозяйственном предприятии при выращивании той же продукции. В таких случаях изучение степени и характера асимметрии и эксцесса вариационных кривых может быть самостоятельной задачей при проведении исследовательских работ, поскольку могут давать информацию о влиянии тех или иных факторов на форму распределения данных.

Показатели описательной статистики. Одна из основных задач статистической обработки – нахождение параметров, представляющих в обобщенном виде распределение данной статистической совокупности. Для решения этих задач используются методы описательной статистики. Для решения этих задач используются методы описательной статистики (таблица 1).

Таблица 1 – Статистические показатели распределения.

Показатели	Назначение показателей	Показатели распределения
Центра распределения (средние величины)	Описывают положение середины распределения	*Структурные (непараметрические) средние*: мода (Mo) и медиана (Me). *Степенные средние*: среднее арифметическое (M), среднее гармоническое ($M_{гар}$), среднее геометрическое ($M_г$), среднее квадратическое ($M_{кв}$), среднее кубическое ($M_{куб}$), среднее взвешенное ($M_{взв}$)
Рассеивания вариант	Описывают степень разброса (вариабельности, изменчивости) вариант	Лимит (lim) Размах варьирования (*ampl*) Среднеквадратическое отклонение (σ) Дисперсия (σ^2) Коэффициент вариации (V) Квантили ($V_{0,25;\ 0,5;\ 0,75}$)
Формы распределения	Описывают симметрию и островершинность распределения вариант около центра	Коэффициент асимметрии (As) Эксцесс (E) Гистограмма Полигон распределения

Показатели центра распределения. Для обоснования представления о генеральной совокупности на основании выборки необходимо использовать наиболее характерные параметры признаков. К ним относятся показатели центра распределения, или среднего положения: *мода, медиана, среднее арифметическое, гармоническое, геометрическое, квадратическое, кубическое, взвешенное*. Средняя величина выражает характерную, типичную для данного ряда величину признака и является равнодействующей всех факторов, влияющих на признак. В ней погашаются индивидуальные различия вариант в ряду, обусловленные случайными обстоятельствами.

Мода (Мо) представляет собой наиболее часто встречающуюся варианту в вариационном ряду. На графике она соответствует максимальной ординате и находится на вершине вариационной кривой. Если вариационный ряд разбит на классы, то мода соответствует максимальной частоте класса, который называется *модальным*. При полимодальном (многовершинном) распределении вариационный ряд имеет несколько значений моды.

Медиана (Ме) представляет собой среднюю варианту в ранжированном вариационном ряду, которая делит его на две равные части.

При наличии в вариационном ряду сильно отличающихся вариант медиана может характеризовать середину ряда более точно, чем среднее арифметическое. Мода и медиана относятся к *непараметрическим средним* и используются в тех случаях, когда о выборочных параметрах необходимо иметь ориентировочное представление.

Среднее арифметическое (M, $\bar{x}$) представляет собой величину, сумма положительных и отрицательных отклонений от которой равна нулю.

Среднее гармоническое ($M_{гар}$) может вычисляться при усреднении меняющихся скоростей процессов (скорость течения воды), показателей обратно пропорциональной зависимости между процессами и явлениями, сложных абсолютных величинах измерений (тонна/км^2).

Среднее геометрическое ($M_г$) вычисляется в тех случаях, когда в вариационном ряду отдельные значения распределяются в геометрической прогрессии.

Если варианты представлены логарифмами чисел (pH и др.), то вычисляют *среднее логарифмическое*.

Среднее квадратическое ($M_{кв}$) может использоваться для вычисления средних радиусов или диаметров объекта, площадей земельных участков, функциональных зон и т.д.

Среднее кубическое ($M_{куб}$) применяется при нахождении средней величины объема.

Средневзвешенная ($M_{взв}$). Сгруппированный вариационный ряд по классам иногда называют взвешенным из-за той роли, которую выполняют частоты. Чем больше частота вариант в классе, тем *больший вес* она имеет в характере распределения числового ряда. Среднее арифметическое, рассчитанное в этом ряду, называют *взвешенным средним*.

Показатели рассеивания вариант. Для полной характеристики распределения в вариационном ряду недостаточно лишь средней арифметической. В совершенно разных по величине вариантах двух выборок средняя может быть одной и той же. Для получения более полного представления о выборочных совокупностях используют показатели рассеяния вариант, или разнообразия (неоднородности) признаков: *лимит, размах варьирования (амплитуда), среднеквадратическое (стандартное) отклонение, средний квадрат отклонений (дисперсия), коэффициент вариации, квантили*. Эти показатели признаков характеризуют различную степень и особенности разброса.

Для быстрой оценки разнообразия можно сравнить минимальные и максимальные варианты со средним значением – чем они ближе и чем меньше амплитуда, тем меньше степень разнообразия между переменными в вариационном ряду, тем надежнее характеризуют статистические показатели искомую закономерность.

Более точно степень разнообразия признака следует характеризовать другими показателями, такими как среднеквадратическое отклонение и дисперсия, которые являются составляющими параметрами нормального распределения при вычислении ряда параметрических статистических критериев.

*Среднеквадратическое отклонение, или сигма (*σ*)* показывает степень рассеяния значений статистической совокупности около среднего значения, а точнее, интервал (M±σ), в который входит до 75 % вариант выборочной совокупности. Считается, если 75 % вариант выборки находится в пределах M±σ. то это вполне соответствует норме (стандартному отклонению); если в пределах M±2σ, то имеется незначительное отклонение от нормы; если выходит за пределы M±3σ, то можно утверждать о наличии анормального явления, процесса. Величина сигмы прямо пропорционально зависит от разброса вариант в вариационном ряду. Чем больший разброс, тем больше значение сигмы. Однако пределы колебаний не дают оценки разброса, как и дисперсия независимо от его величины.

Среднеквадратическое отклонение (σ) используется для:

- оценки данных одноименных вариационных рядов при близких средних: чем больше сигма, тем больший разброс вариант в совокупности, соответственно среднее арифметическое менее типично для данного ряда;
- для оценки типичности среднего арифметического в ряду, используя правило трех сигм (M±3σ);
- для определения доверительных интервалов статистических коэффициентов и репрезентативности выборочных исследований.

Средний квадрат отклонений, или дисперсия (σ^2*)*, указывает колебание значений признака внутри выборочной совокупности через отклонение всех вариант от среднего значения, т. е. показывает интервал, в который входят все варианты выборки (100 %). Однако сумма всех отрицательных и положительных отклонений от среднего равна нулю. Поэтому все отклонения от среднего возводятся в квадрат и суммируются: $\sum(x_i - M_x)^2$. При усреднении всех отклонений числового ряда путем деления на (N – 1) получаем средний квадрат отклонений, или дисперсию (D, σ^2).

Практическое применение дисперсии (σ^2) состоит в следующем:

- для оценки вариабельности рядов распределения;
- для факторного и дисперсионного анализа;
- для статистической оценки двух совокупностей по критерию Фишера.

Недостаток сигмы, как и дисперсии, заключается в том, что критерий представляет собой абсолютную именованную величину, поэтому его нельзя использовать при сравнении разнородных рядов, выраженных в различных единицах измерения. Для этой цели подходит коэффициент вариации.

Коэффициент вариации представляет собой относительный показатель разнообразия признаков, выражается в процентах. Он показывает отношение среднеквадратического отклонения к средней арифметической:

$$V = (\sigma / M) \cdot 100$$

Чем меньший по размаху варьирования будет признак, тем меньший будет коэффициент вариации для данной совокупности. Соответственно меньшими будут сигма и дисперсия.

Коэффициент вариации позволяет оценить вариабельность (разброс) признака в нормированных границах. Если его значение меньше 10 %, то разброс вариант относительно средней арифметической считается слабым, при 10–30 – средним, 30–60 – высоким, 60–100 – очень высоким, более 100 % – аномальным.

Квантили. В открытых вариационных рядах и рядах распределения качественных признаков для сжатого описания распределений используется другой параметр разброса – *квантиль*. Этот параметр может использоваться для перевода количественных признаков в качественные. В практике статистического анализа наиболее часто используются следующие квантили:

$V_{0,5}$ – медиана;

$V_{0,25}$, $V_{0,75}$ – квантили четверти, соответственно нижняя и верхняя квантиль;

$V_{0,1}$, $V_{0,2}$, … $V_{0,9}$ – децили (десятые);

$V_{0,01}$, $V_{0,02}$, … $V_{0,99}$ – процентили, или центили (сотые).

Квантили делят область возможных изменений вариант в выборке на определенные интервалы. Статистическая суть квантилей лучше раскрывается при построении графика.

Оценка статистических параметров по выборочным данным. Оценка в статистике – это правило вычисления оцениваемого параметра. Она указывает приближенное значение показателей выборки относительно этих же параметров для генеральной совокупности. По мере увеличения числа наблюдений выборочные средние и другие параметры все больше приближаются к этим значениям генеральной совокупности.

Степень соответствия показателей оценивается *ошибкой репрезентативности (m)*. Ее запись производится вместе с оцениваемым параметром, например, $M \pm m_M$, $\sigma \pm m_\sigma$, $V \pm m_V$.

Ошибка указывает интервал, в пределах которого находится этот показатель в генеральной совокупности. Чем меньше ошибка, тем ближе значение выборочного показателя к этому показателю генеральной совокупности. Чем больше число наблюдений и чем однороднее выборка, тем меньшая ошибка среднего и других показателей.

Представление средней арифметической выборки приводится обязательно с ее ошибкой. Стандартная ошибка средней рассчитывается:

$$m_M = \sqrt{\frac{\sigma^2}{N}}, \text{ или } m_M = \frac{\sigma}{\sqrt{N}}$$

Поскольку параметр m характеризует ошибку утверждения (прогноза) о том, что выборочное среднее равно генеральному среднему, то, чем выше требование к вероятности этого вывода, тем шире должен быть обеспечивающий точность такого прогноза интервал, называемый *доверительным интервалом*. Его величина задается вероятностью безошибочного прогноза, которую принято *называть доверительной вероятностью* (*уровень вероятности,* надежность опыта, вероятность безошибочного прогноза). В исследованиях допускается доверительная вероятность (*P*) не менее 95 % (0,95 частей от 1). В этих случаях *P* для средних арифметических при достаточно большом числе наблюдений ($N > 30$) равен $\pm 2\ m$. Предельная ошибка выборки $\Delta = M \pm 2\ m$. При доверительной вероятности 99 % (0,99) доверительный интервал составит $\pm 3\ m$, $\Delta = M \pm 3\ m$. По-иному, в отношении доверительного интервала можно сказать так: *он показывает какой процент вариант выборки (выборок) подтверждает искомую статистическую закономерность.*

Каждому значению доверительной вероятности соответствует свой *уровень значимости* (α). Он выражает вероятность нулевой гипотезы: вероятность того, что выборочная и генеральная средние не отличаются друг от друга. Иначе говоря, чем выше уровень значимости, тем меньше можно доверять утверждению, что различия существуют, т. е., *он показывает, какой процент вариант совокупности (выборок) отвергают искомую статистическую закономерность.* Уровень значимости 5 % (0,05) дополняет доверительную вероятность 95 % (0,95). В сумме они составляют 100 % (1). Если доказано подобие между выборками при α = 5 % (0,05), то из этого следует, что до 5 % вариант выборки подобие не подтверждают. В таблицах приложения приводятся численные значения для *P* или α соответственно 0,95 и 0,99; 0,05 и 0,01. В этих случаях при интерпретации мы можем утверждать нулевую гипотезу (H_0). При более высоких уровне вероятности 0,99 и уровне значимости 0,01 мы получаем сильный довод для утверждения нулевой гипотезы.

*Проверка статистических гипотез***.** Методологической основой любого исследования является формулировка рабочей гипотезы. В ходе исследования рабочая гипотеза либо принимается, либо отвергается. Статистической называют гипотезу о виде неизвестного распределения или о параметре распределения.

Для проверки нулевых гипотез используют статистические критерии. При сравнении дисперсий используют критерий Фишера. В большинстве исследований для статистической проверки гипотез существенности различий средних арифметических используют параметрический критерий Стьюдента. Если нулевая гипотеза принимается, это не означает ее доказательство. Доказать на основании однократной или косвенной проверки гипотезу нельзя, а опровергнуть можно. Для повышения точности статистических данных необходимо уменьшить вероятности ошибок первого и второго рода, увеличить объем выборок. Область применения того или иного критерия задается законом его распределения.

Оценка точности опыта. При исследованиях *методического* характера необходимо приводить их оценку по показателю *точность опыта* (*p*). Его смысл состоит в установлении величины ошибки среднего арифметического (m_M) в процентах от величины среднего арифметического (*M*). Показатель точности опыта можно определить по одной из двух формул:

$$p = (m_{\mathrm{M}} / \mathrm{M}) \cdot 100; \quad p = V / \sqrt{N},$$

где *V* – коэффициент вариации.

Опыт считается достаточно точным, если $p < 3$ %, удовлетворительным – при его величине 3–5 % . Если величина точности опыта более 5 %, к полученным выводам следует относиться осторожно и увеличить число повторностей в опыте. Эти градации обязательны для полевых опытов с растениями. Некоторые приборы для анализа или экспресс-тесты могут давать значительно большую погрешность (p до 15 %).

Теоретические функции распределения. В ходе работы с выборочной совокупностью иногда возникает необходимость описать вариационную кривую с помощью математической функции. Для характеристики вариационной кривой можно подобрать ряд математических зависимостей. Выбирают ту, которая наиболее реально отражает сущность объекта исследования. Выбор математической зависимости, описывающей распределение, проводится путем подбора подходящей математической модели, которая определяет вид функции распределения. Затем находят параметры функции и проверяют ее соответствие эмпирическому распределению.

В географии большинство закономерно повторяющихся явлений, процессов можно представить в виде *нормального и логнормального распределения*. Реже встречается *биномиальное распределение, распределение Пуассона и другие.*

Биномиальное распределение (распределение Бернулли) возникает, когда оценивается сколько раз происходит событие в серии определенного числа независимых, выполняемых в одинаковых условиях наблюдений. Разброс вариант – следствие влияния ряда независимых и случайно сочетающихся факторов (есть событие или его нет). Характерно для альтернативного типа изменчивости признака.

Распределение Пуассона рассматривается как предельный случай биномиального распределения и используется для характеристики редких событий. Отличительная особенность распределения Пуассона – величина дисперсии близка к величине среднего арифметического, например, длительное наводнение. Это проявляется в ситуациях, когда в определенный отрезок времени или на определенном пространстве происходит случайное число каких-либо событий, например, длительно повторяющиеся ураганы в течение одного летнего периода. На графике это распределение представляется в виде резко выраженной асимметрии.

Рассмотрим более детально наиболее характерные типы теоретических распределений в природе и обществе: нормальное и логнормальное распределение.

*Нормальное распределение***.** Нормальное (распределение Гаусса) используется для приближенного описания явлений, которые носят вероятностный, случайный характер. Распределение Гаусса имеет место среди природных и экономических явлений. В системе признак варьирует под влиянием большого количества взаимно независимых факторов, каждый из которых мало влияет на его общую вариабельность. Причем одни факторы приводят к возрастанию величины признака, другие – к уменьшению. Встречаемость вариантов, занимающих середину совокупности, максимальна. Такое распределение считается нормой для случайных величин, поэтому оно получило название нормального. Графически нормальное распределение выражается плавной симметричной куполообразной кривой с приближающимися к оси абсцисс ветвями (кривая плотности нормального распределения).

Кривая показывает, что большие отклонения от средней встречаются реже, чем малые. С уменьшением среднего квадратического отклонения (σ) кривая нормального распределения становится все более островершинной. Площадь, заключенная под кривой нормального, всегда принимается равной единице.

При нормальном распределении среднее, мода и медиана совпадают. Кривая плотности не пересекает оси абсцисс, что подтверждает вероятность существования неограниченно больших отклонений.

При нормальном распределении около 68,3 % всех вариант отклоняется от среднего значения не более, чем на величину среднего квадратического отклонения (±σ). Соответственно в пределах от –2σ до +2σ находится 95,5 % вариант, в пределах от –3σ до +3σ – 99,7 %.

Отклонение вариант от нормального закона распределения указывает на влияние какого-либо другого фактора на статистическую совокупность.

Логнормальное распределение. Некоторые распределения при изучении географических объектов имеют выраженную асимметрию, поэтому представляет практический интерес преобразование асимметричного распределения в симметричное (нормальное). Иногда это возможно, если каждую варианту выборки выразить в виде логарифма ($\lg x_i$). В тех случаях, когда логарифм случайной величины (x_i) подчиняется нормальному распределению, а сами значения случайных величин распределены асимметрично, распределение случайной величины принято называть логарифмически нормальным, или логнормальным. Например, к логнормальному распределению можно отнести распределение ряда микроэлементов в почвах, породах.

1.3. Методы установления сходства выборок

Проведение географических исследований предполагает не только изучение строения, развития, закономерностей распространения исследуемых объектов, явлений, но и установление сходства или различия между одноименными генеральными совокупностями изучаемых систем. Это зависит от условий, в которых протекает один и тот же процесс.

Сопряженный анализ одноименных признаков в выборках используется для классификации и районирования по одному или нескольким параметрам. При этом возникает необходимость применения объективного метода выделения классификационных групп или районов на основе методов математической статистики с использованием критериев достоверности.

Если достоверность различия между выборочными совокупностями доказана, то генеральные совокупности, сравниваемые по какому-либо признаку, выделяют как самостоятельные. В случае отсутствия достоверных различий их объединяют в одну группу.

Различие между двумя выборками устанавливается с помощью ряда критериев: t – распределение Стьюдента, наименьшего существенного различия (НСР), F – распределения Фишера, критерия соответствия (χ^2).

Каждый из критериев применяется при определенных условиях, которые задаются целью исследования. Несоблюдение указанных условий может привести к ошибочным выводам.

Прежде, чем приступать к статистической обработке и расчету критериев различия, следует убедиться в отсутствии артефакта в сравниваемых выборках.

Критерий Стьюдента. Используется для оценки сходства или различия между выборочными совокупностями по разности величин их средних арифметических ($d = M_{\text{большая}} - M_{\text{меньшая}}$) и ее отношения к ошибке этой разности (m_d) при условии распределения вариант в группах по закону нормального распределения и подтверждается равенство разброса вариант в выборке (близкие дисперсии сравниваемых выборок). Не допускается применения критерия в случае балльного характера сравниваемых числовых признаков.

Конкретная методика оценки для установления различий по критерию Стьюдента зависит от вида выборочных совокупностей: зависит от учета следующих особенностей выборочных совокупностей: сравниваются средние арифметические в *независимых* (*несвязанных*) выборках; различия устанавливаются в *сопряженных* (*парных*) *выборках*; устанавливается различие между выборочными и генеральными средними (теоретическими стандартами).

Независимые статистические совокупности могут быть получены на одном или нескольких объектах, но при одинаковых условиях проведения эксперимента: сравнение экономического показателя в хозяйстве или на предприятии по годам между собой; сравнение чистого дохода в хозяйствах с одинаковым экономическим развитием, но расположенных на значительном расстоянии. При сравнении независимых выборочных совокупностей объемы выборок могут быть одинаковы ($N_1 = N_2$) или разные ($N_1 \neq N_2$). В двух сравниваемых независимых выборках с одинаковым или разным объемом наблюдений *степень свободы* определяется так: $\nu = N_1 + N_2 - 2$.

Сопряженные статистические совокупности получают на одном или на разных объектах, но в разных условиях. Например, сравнение прибыли фермерских и подсобных хозяйств в любом районе или фермерских хозяйств Витебской и Гомельской области. Объем сравниваемых выборок должен быть одинаков ($N_1 = N_2$). Определение *степени свободы* для *сопряженных* выборок определяется как: $\nu = N_{\text{пар}} - 1$.

Ошибка разности между средними выборок (m_d) в зависимости от вида наблюдений (независимые, сопряженные) и объема наблюдений рассчитывается по разным формулам.

Сопоставляя критерий Стьюдента вычисленный с табличным (прил. 4 учебно-методического пособия), устанавливают или отвергают с некоторой долей уверенности различия между средними арифметическими выборок. В случае, если $t_ф > t_т$, то различие между выборками признается существенным. Если же $t_ф < t_т$, то различие между выборками является статистически не значимым и им можно пренебречь.

Расчет критерия Стьюдента в MS Excel приведен в прил. 10.

Наименьшая существенная разность (НСР). Используется в дисперсионном анализе. Она показывает то минимальное различие между средними, начиная с которого при выбранном уровне вероятности средние сравниваемые показатели существенно отличаются друг от друга. Величина критерия выражается в тех же единицах, что и сравниваемые средние выборочных совокупностей и определяется по формуле:

$$НСР = t_{табл} \cdot m_d,$$

где m_d – ошибка разницы средних; $t_{табл}$ – табличное значение критерия Стьюдента при уровне вероятности 0,95 или 0,99 и степени свободы, определяемой экспериментом.

Критерий Фишера. В выборочных совокупностях дисперсии могут существенно отличаться друг от друга. В таких случаях установление различий между выборочными совокупностями проводится по критерию Фишера (*F – положительное асимметричное распределение*). Расчет производится по формуле:

$$F = \sigma^2_{большая} / \sigma^2_{меньшая}$$

Если величина расчетного критерия Фишера ($F_ф$) не превышает величины приведенного в таблице ($F_т$) (приложение 5 учебно-методического пособия), то различие между сравниваемыми дисперсиями считается недостоверным. При $F_ф > F_т$ эти дисперсии достоверно различны, как и сравниваемые по ним генеральные совокупности. Степень свободы рассчитывается для сравниваемых выборок отдельно по формуле $\nu = N - 1$.

Критерий Пирсона (хи-квадрат, χ^2). Для оценки соответствия или расхождения полученных эмпирических данных и теоретических (расчетных, прогнозных) распределений применяются статистические *критерии согласия.* Среди них наибольшее распространение получил непараметрический критерий К. Пирсона – хи-квадрат. Его можно использовать с различными формами распределения совокупностей. Как и любой другой статистический критерий, он не доказывает справедливость нулевой гипотезы, а лишь устанавливает с определенной вероятностью ее согласие или несогласие с экспериментальными данными. Критерий применяется при условии наличия не менее 5 наблюдений или частот в каждой группе, классе или совокупности. Малые частоты объединяют. Вычисление проводят по формуле:

$$\chi^2 = \sum [(\varphi - \varphi')^2 / \sum \varphi'],$$

где φ, φ' – наблюдения или частоты в опыте соответственно эмпирически или теоретически ожидаемые.

Значения χ^2 могут быть только положительными и возрастать от нуля до бесконечности. Если вычисленный критерий хи-квадрат больше табличного (теоретического) значения, нулевая гипотеза, которая предполагает соответствие эмпирического и теоретического распределений, отвергается, при $\chi^2_{выч} < \chi^2_{табл}$ нулевая гипотеза принимается.

Критерий Пирсона тем меньше, чем меньше различаются эмпирические и теоретические частоты. Он не позволяет обнаружить различия, которые скрадывает группировка (объединение малых частот в одну группу). Его удобно использовать, так как не требуется вычислений средних дисперсий.

1.4. Дисперсионный анализ

При планировании эксперимента бывают ситуации, когда исследуемую систему необходимо разбить на группы, отличающиеся между собой в количественном отношении, и установить сходство или различие между ними по влиянию различных факторных величин на признак. Например, определить степень влияния географических условий на ход тех или иных процессов, явлений. Таким условиям лучше всего отвечает дисперсионный анализ, который нашел применение при проведении полевых и лабораторных экспериментов.

Дисперсионный анализ позволяет утверждать с определенной долей уверенности наличие влияния на изучаемый объект каждого из условий в отдельности или в их сочетаниях. Обязательным условием применения дисперсионного анализа является разбивка каждого учитываемого фактора не менее чем на две группы. Они могут быть представлены как качественными (в таком случае приводятся в виде баллов), так и количественными показателями. Анализу подвергаются лишь определяющие поведение объекта факторы, которые установлены исследователем. По количеству определяющих факторов дается название виду дисперсионного анализа (одно-, двух-, трехфакторный и т. д.).

Для получения достоверных результатов в ходе дисперсионного анализа необходима правильная организация опыта. Порядок расчета в различных видах дисперсионного анализа будет различным, но логическая схема остается единой. Факторы в дисперсионном анализе должны быть независимыми друг от друга; каждый фактор следует разделить на группы, количество которых зависит от поставленной задачи.

Дисперсионный анализ применяется в случаях нормального или близкого к нему распределения выборочных совокупностей. Выборки должны иметь близкие по значению показатели дисперсии σ^2. Количество повторностей в каждой выделенной группе принимается одинаковым.

Основная трудность при использовании дисперсионного анализа – составление комбинационной таблицы для обработки данных (дисперсионный комплекс). Если число наблюдений над результативным признаком по отдельным группам изучаемого фактора одинаково, то дисперсионный комплекс называет*ся равномерным, если разное, то неравномерным. Общее число наблюдений над результативным признаком принято называть объемом дисперсионного комплекса.*

Порядок действия по каждому виду дисперсионного анализа определяется его основной задачей, которая состоит в делении суммарного или общего варьирования изучаемого признака на доли: варьирование, вызываемое действием отдельных факторов; варьирование, вызываемое взаимодействием факторов между собой; остаточное варьирование объекта, которое определяется не учитываемыми факторами.

Среди различных видов дисперсионного анализа наиболее часто используется *однофакторный*. Для выполнения однофакторного анализа в опыте должно быть предусмотрено две повторности и более. Исследуемый фактор разбивается на группы с целью выявления его оптимальной величины, влияющей на результативный признак.

Дисперсионный комплекс составляется через нахождение сумм квадратов отклонений, т.е. расчленение общего варьирования признака на составные части исходя из равенства:

$$\Theta = \Theta_1 + \Theta_2 + \Theta_3,$$

где Θ – сумма квадратов отклонений по общему варьированию данных, Θ_1 – по группам фактора (варианты опыта), Θ_2 – по повторностям опыта, Θ_3 – по остаточному варьированию, вызванному неучтенными факторами.

Оценку сходства или различия между вариантами опыта можно проводить по критерию Фишера, критерию Стьюдента или НСР.

Если в дисперсионный анализ включают несколько факторов, влияющих на результативный признак, то они должны быть независимыми друг от друга.

Двухфакторный дисперсионный анализ можно представить в виде равенства:

$$\Theta = \Theta_1 + \Theta_2 + \Theta_3 + \Theta_4 + \Theta_5,$$

где Θ – общая сумма квадратов; Θ_1, Θ_2 – сумма квадратов отклонений для фактора I и II соответственно; Θ_3 – сумма квадратов отклонений, возникающих при взаимодействии факторов I и II; Θ_4 – сумма квадратов отклонений по повторностям; Θ_5 – остаточная сумма квадратов отклонений неучтенных факторов.

1.5. Кластерный анализ

При проведении географических исследований, как правило, возникает проблема *объединения по сходству (кластеризация)* объектов, которые характеризуются *множеством признаков, выраженных в разных единицах измерения*. Для этой цели используется *кластерный анализ*. Наилучшие результаты кластерный анализ дает в сочетании с факторным, поскольку первый удобен для проведения классификации объектов, а второй – для ее обоснования, поскольку исследует связи между объектами.

Методологические особенности кластерного анализа сводятся к выявлению единой меры, охватывающей ряд исследуемых признаков. Эти признаки объединяются с помощью метрики (расстояния) в один кластер сходства группируемых объектов.

Кластерный анализ удобно использовать для классификации или типологии, при проведении районирования и т.д.

Состояние любого объекта может быть описано с использованием *многомерного признака*, или *многомерной случайной величины* ($x_1, x_2, \ldots, x_n$). Примером количественных признаков при зонировании территории города может служить площадь строений (x_1), количество исторических памятников (x_2), количество промышленных предприятий (x_3) и т. д. Их можно объединить в один качественный признак – инфраструктурные условия города. Таким способом состояние любого объекта может быть описано с помощью многомерного признака.

Исследование нескольких аналогичных объектов (городов) обязывает проводить разбиение совокупности объектов на однородные группы, т. е. провести их классификацию по сходству признаков (x_1, x_2 ...). Содержательная постановка задачи при кластерном анализе заключается в следующем. Имеется некоторая совокупность объектов, которые характеризуются рядом признаков. Объекты необходимо разбить на несколько кластеров (классов) таким образом, чтобы объекты из одного класса были сходными по характеризующих их признакам, например, сравнение ландшафтов, выявление сходных тенденций в развитии экономических субъектов.

При относительной формализации методов кластерного анализа они носят эвристический (теоретический) характер, реализуют принцип здравого смысла. Для оценки сходства объектов по ряду признаков используют три типа мер:

- *коэффициент подобия* – для группировки объектов и признаков, если уровни показателей являются действительно целыми числами;
- *коэффициенты связи* – чаще применяются для группировки признаков с использованием коэффициента корреляции;
- *показатели расстояния* – характеризуют степень взаимной удаленности признаков и применяются в основном для кластеризации объектов; признаки объектов должны быть независимыми, что предварительно можно уточнить с помощью корреляционного анализа.

Многомерное наблюдение может быть интерпретировано геометрически в виде точки в многомерном пространстве. Геометрическая близость точек в пространстве означает близость физических состояний объектов, их однородность. Решающим в интерпретации остается выбор масштаба метрики, т. е. задание расстояния между объектами, которые объединяют или разъединяют объекты. В результате разбиения объектов на группы по сходству признаков образуются *кластеры (таксоны, образы)*. Необходимость разбиений совокупности объектов на однородные группы возникает при проведении социально-экономических, землеустроительных, географических исследований и т. д.

Выбор метрики (меры близости) является важнейшим моментом исследования, который определяет окончательный вариант разбиения объектов на группы. Это зависит от цели исследования, физической и статистической природы вектора наблюдений (x), полноты априорных сведений о характере вероятностного распределения x.

В задачах кластер-анализа широко используются следующие метрики: Эвклида, манхэттенское расстояние, Чебышева, Минковского, Хемминга, меры близости, задаваемые потенциальной функцией. Эвклидова метрика наиболее употребительна.

Эти метрики применяются в следующих случаях:

- наблюдения x извлекаются из генеральных совокупностей, описываемых многомерным нормальным законом с ковариационной матрицей (совместное изменение двух признаков), где компоненты x взаимно независимы и имеют одинаковую дисперсию;
- компоненты $x_1, x_2, \ldots, x_p$ вектора наблюдений x однородны по своему физическому смыслу и все важны;

- факторное пространство совпадает с геометрическим; понятие близости объектов соответственно совпадает с понятием геометрической близости в этом пространстве.

При решении задач классификации могут быть использованы разные меры сходства между объектами. Выбор метрики зависит от вида информации, характеризующей объекты в пространстве признаков, и требует тщательного критического анализа.

Метод дендритов. Исследуемые объекты, разделенные на кластеры, можно изобразить в виде дендрограммы, которая представляет собой графическое изображение матрицы расстояний или сходства. Такой анализ объектов исследования носит название метода дендритов. Имея n объектов, можно построить большое количество дендрограмм, которые соответствуют избранной процедуре кластеризации. Для конкретной матрицы расстояний или сходства существует только одна дендрограмма.

Вид дендрограммы зависит от выбора меры сходства или расстояния и метода кластеризации. Например, разработаны алгоритмы кластерного анализа, позволяющие проводить классификацию (группировку) многомерных наблюдений (строк и столбцов матрицы x) с помощью следующих мер сходства: выборочного коэффициента корреляции, модуля выборочного коэффициента корреляции, косинуса угла между векторами, модуля косинуса угла между векторами, эвклидова расстояния и т. д.

Решение задач классификации объектов с использованием кластерного анализа проводится в определенной последовательности. Многомерный анализ делится на три этапа:

- составляется таблица исходной информации с указанием объектов и их признаков;
- проводится нормализация исходной информации с использованием среднего квадратического отклонения;
- по нормализованным данным рассчитывается метрика, строится дендрограмма и проводится содержательная интерпретация полученных результатов.

На первом этапе при формировании таблицы выбор объекта зависит от места и масштаба исследования. Каждый объект должен быть пространственно локализован и одного ранга (уровня). Показатели должны отражать существенные черты или свойства исследуемых объектов и характеризовать их всесторонне.

На втором этапе нормализация значений исходных показателей по объектам проводится потому, что исходные данные выражены обычно в разных единицах измерения и проводить между ними арифметические действия невозможно без перевода их в безразмерные единицы.

На третьем этапе по нормализованным показателям рассчитывается метрика по одному из предложенных выше способов, учитывая условия задачи. Классификацию объектов производят приемами таксономического или факторного анализа.

При количестве координат (показателей) в многомерном пространстве более трех графически интерпретировать таксономические расстояния невозможно. Поэтому таксономические расстояния определяют на основе функции расстояний. Чаще всего используется эвклидова метрика.

На основе матрицы таксономических расстояний производится группировка объектов с использованием разных приемов, из них наиболее распространенные – вроцлавская таксономия, дендрограмма Берри, метод дендритов.

1.6. Корреляционный анализ

Термин «корреляция» означает соотношение, соответствие. Основоположниками теории корреляции считаются английские биометрики Френсис Гальтон (1822–1911) и Карл Пирсон (1857–1936). Представление о корреляции как о взаимозависимости случайных переменных величин лежит в основе статистической теории корреляции – изучение зависимости вариации признака от окружающих условий. Одни признаки выступают в роли влияющих (факторных), другие – на которые влияют, результативных. Зависимости между признаками могут быть функциональными и корреляционными. Функциональные связи характеризуются полным соответствием между изменением факторного признака и изменением результативной величины. Каждому значению признака-фактора соответствует определенное значение результативного признака.

В корреляционных связях между изменением факторного и результативного признака нет полного соответствия. В сложном взаимодействии находится сам результативный признак. Поэтому результаты корреляционного анализа имеют значение в данной связи, а интерпретация этих результатов в общем виде требует построения системы корреляционных связей. Они характеризуются множеством причин и следствий и с их помощью устанавливается тенденция изменения результативного признака при изменении величины факторного признака.

В природе и обществе явления и события протекают по характеру корреляционной связи, когда при изменении величины одного признака существует тенденция изменения другого признака. Корреляционная связь – это частный случай статистической связи. Корреляционный анализ используется при установлении тесноты возможной зависимости между явлениями, процессами, объектами.

Целью исследования часто бывает установление взаимосвязи (корреляции) между признаками. Знание зависимости дает возможность решать кардинальную задачу любого исследования – возможность предвидеть, прогнозировать развитие ситуации при изменении влияющего фактора. С помощью корреляции можно дать лишь формальную оценку взаимосвязей. Поэтому, прежде чем приступать к вычислению коэффициентов корреляции между любыми признаками, следует теоретически установить, имеется ли между этими признаками взаимосвязь. Ведь формально статистика может доказать несуществующие связи, например, между высотой здания в городе и урожайностью пшеницы в фермерских хозяйствах.

Связь между явлениями (корреляция) определяется путем постановки опытов, статистического анализа. *Корреляцию не следует отождествлять с причинностью*. Необходимо иметь в виду, что доказательство математической связи должно опираться на реальную зависимость между явлениями.

Любой показатель связи служит приближенной оценкой рассматриваемой зависимости и не является гарантией существования жесткой (функциональной) соподчиненности. Отсутствие жесткой зависимости в природе и обществе способствует саморегуляции процессов, явлений, систем.

По направлению связь может быть *прямой* и *обратной*; по характеру – *функциональной* или *статистической (корреляционной)*; по величине – *слабой, средней* или *сильной*; по форме – *линейной* и *нелинейной*; по количеству коррелируемых признаков – *парной и множественной*.

Функциональная зависимость характерна для геометрических форм, технических систем, когда каждому значению одного признака соответствует точное значение другого. Это пример взаимосвязи площади прямоугольника и длины его одной из сторон. Такая зависимость полная или исчерпывающая.

Выделяют несколько видов парной корреляционной связи:

- параллельно-соотносительную, или ассоциативную, когда оба признака изменяются сопряжено, частично под действием общих причин и следствий (приуроченность растительности и почв к определенным формам рельефа);
- субпричинную, когда один фактор выступает как отдельная причина сопряженного изменения признака (связь биомассы с количеством осадков; рост населения и рождаемости);
- взаимоупреждающую, когда причина и следствие, находясь в устойчивой взаимной связи, последовательно влияют друг на друга (влажность воздуха и осадки).

Если на признак влияет несколько факторов, то приходится оценивать множественную корреляцию. Множественная корреляция служит основой выявления связей между признаками, но требует строгой нормальности и прямолинейности распределения, поэтому использование ее может быть затруднено. С ростом числа переменных объем вычислительных работ увеличивается пропорционально квадрату числа переменных. В этом случае труднее оценивать значимость результатов, так как увеличиваются ошибки коэффициентов корреляции. Поэтому характер влияния главных факторов на признак более детально и точно исследуют путем факторного анализа.

В практической работе по установлению корреляции между признаками и явлениями необходимо придерживаться следующей последовательности:

- на основании проведенных исследований предварительно определяют, существует ли связь между рассматриваемыми признаками;
- если связь между ними существует, устанавливают ее форму, направление и тесноту, используя график.

Степень рассеяния частот или вариант относительно линии регрессии на графике указывает ориентировочно на тесноту связи: чем меньше рассеяние, тем сильнее связь.

Корреляционный анализ решает следующие задачи:

- установление направления и формы связи;
- оценка тесноты связи;
- оценка репрезентативности статистических оценок взаимосвязи;
- определение величины детерминации (доли взаимовлияния) коррелируемых факторов.

Для оценки связи используют следующие численные критерии (коэффициенты) корреляционной связи:

- коэффициент корреляции (r) при линейной зависимости,
- корреляционное отношение (η) при нелинейной зависимости,
- коэффициенты множественной регрессии,
- ранговые коэффициенты линейной корреляции Спирмена или Кендэла.

Линейная корреляция. Для установления формы зависимости по исходным (x, y) строится график. В случае линейной зависимости вычисляется коэффициент корреляции (r), при нелинейной – корреляционное отношение (η).

Принимается следующая характеристика тесноты корреляционной связи: если r (η) = 0 ± 0,4, то связь считается слабой; от ± 0,4 до ± 0,7 – средняя; от ± 0,7 до ± 1 – сильная.

Достоверность вычисленного коэффициента корреляции может быть установлена двумя путями: путем сравнения с табличным значением r (прил. таблица 6 учебно-методического пособия); второй путь – через критерий Стьюдента.

Если $r_{выч} > r_{табл}$, то влияние фактора на признак достоверно; если меньше табличного – не достоверно.

Достоверность связи устанавливается путем сравнения r (η) расчетного (фактического) и r (η) теоретического (табличного). Если $r(\eta)_{выч} > r(\eta)_{табл}$ при учете степени свободы (ν) вариационных рядов и уровня вероятности P = 0,95 и 0,99, то связь между признаками доказана без учета величины $r(\eta)$.

Регрессионный анализ обычно является продолжением корреляционного, в случае если r (η) $\geq \pm$ 0,7.

Коэффициент детерминации (*причинности*) R^2 (D^2) – это коэффициент корреляции, возведенный в квадрат, например, $R^2 = r^2 = 0{,}4^2 = 0{,}16$. С помощью коэффициента детерминации можно установить долю влияния анализируемого факторного признака на результативный признак. В случае, когда $R^2 = 0{,}16$, можно утверждать, что доля влияющего фактора (x) на признак (y) составляет 16 %. Следовательно, на долю других факторов приходится 84 % влияния.

Нелинейная корреляция. Зависимость между признаками не всегда выражается в виде прямой линии. Если рассеяние точек на графике приближается к кривой линии, то зависимость устанавливается с использованием корреляционного отношения (η), величина которого изменяется только от 0 до 1. Для него теоретические значения приводятся отдельно в таблице или находятся при перерасчете его в критерий Стьюдента. При нелинейной корреляции вычисляется корреляционное отношение (η).

Для установления формы связи иногда используется критерий криволинейности в случаях, когда кривая линия мало отличается от прямой. Существует несколько способов оценки степени криволинейности.

Первый способ менее точный. Оценка степени криволинейности определяется по разности коэффициента корреляции и корреляционного отношения с использованием неравенства: $\eta^2 - r^2 \geq 0{,}1$. Корреляция считается криволинейной, если полученный результат соответствует этому неравенству. Предварительно следует рассчитать между сравниваемыми выборками r и η.

Второй способ оценки степени криволинейности связан с применением критерия Стъюдента:

$$t = 0{,}5\sqrt{\frac{N}{(\eta^2 - r^2)^{-1} - 2 + (\eta^2 + r^2)}} \geq 3.$$

Если $t_{выч} < 3$ или $t_{выч} < t_{табл}$, то рассматриваемая связь несущественно отклоняется от прямолинейной, поэтому относим ее к линейной. В других случаях связь между признаками относят к криволинейной и рассчитывается корреляционное отношение.

Корреляционное отношение, как и коэффициент корреляции, используется для оценки прямой и обратной зависимости между признаками.

Оценка прямой нелинейной зависимости между признаками. Нелинейная зависимость прямая определяется как параболическая. Расчет корреляционного отношения производится по формуле с использованием функции y:

$$\eta = \sqrt{\frac{n\sum(\bar{y} - M_y)^2}{\sum(y_i - M_y)^2}},$$

где $\bar{y}$ – среднее арифметическое частных групп по y_i; n – число вариант в частной группе; $\bar{y} - M_y$ – отклонение общего среднего (M_y) от средних арифметических частных групп ($\bar{y}$).

Ошибка корреляционного отношения независимо от способа расчета вычисляется следующим образом:

$$m_\eta = \sqrt{(1 - \eta^2)/(N_{пар} - 2)}$$

Критерий Стьюдента определяется с использованием η:

$$t_\eta = \eta / m_\eta.$$

Если $t_{выч} > t_{табл}$, то корреляционное отношение признается достоверным.

Оценка обратной нелинейной зависимости между признаками. Алгоритм вычисления и доказательств при расчете корреляционного отношения обратной нелинейной (гиперболической) зависимости аналогичен алгоритму прямой нелинейной зависимости. Различие состоит в том, что в качестве исходных вариант используется выборка со значениями x.

Для нелинейной обратной (гиперболической) зависимости корреляционное отношение определяется с использованием аргумента x по формуле, условные обозначения в которой аналогичны формуле для параболической связи:

$$\eta_x = \sqrt{\frac{n\sum(\bar{x}_{гр} - M_x)^2}{\sum(x_i - M_x)^2}}$$

Расчетные величины η по x сравнивают с табличными (теоретическими) для степени свободы ($\nu = N_{пар} - 1$) и $P = 0{,}95$ и 0,99. Если расчетная величина больше табличной, то можно утверждать с уверенностью о наличии достоверной зависимости между признаком и фактором.

Частная (парциальная) корреляция. В практических целях часто приходится выявлять взаимодействие нескольких факторов. Производится комбинационная группировка собранного материала, которая требует большого числа наблюдений. Можно использовать специальные статистические методы. С помощью этих методов производится последовательная элиминация влияния одних факторов и выделение результатов влияния других факторов. К таким методам относится метод частной корреляции. Элиминация – это исключение неизвестного из системы уравнений.

Понятие о множественной корреляции. Метод множественной корреляции применяется в случаях, когда необходимо установить совокупное влияние всего комплекса факторов на результативный признак. Величина коэффициента множественной корреляции изменяется от 0 до 1. Его можно вычислить с использованием коэффициентов частной линейной корреляции по формуле:

$$R_{1,23} = \sqrt{1-(1-r_{12}^2)(1-r_{13}^2)} = \sqrt{1-(0,4^2)(1-(-0{,}7)^2)} = 0{,}75.$$

По коэффициенту $R = 0{,}75$ определяется коэффициент детерминации R^2 (R_D) $= 0{,}75^2 = 0{,}56$. Он показывает, что доля совместного влияния второго и третьего признаков составляет 56 %.

Оценка различий коэффициентов корреляции. Решение задач по оценке различий между коэффициентами корреляции возникает иногда в случае, если обе выборки принадлежали к одной генеральной совокупности. Следует иметь в виду при анализе коэффициентов корреляции: чем больше r, тем меньшие различия между ними становятся значимыми. Если для $r = 0{,}14$ и 0,24 (разница между ними в 0,1) может быть статистически не значимой, то для $r = 0{,}80$ и 0,90 (разница 0,1) может оказаться значимой.

Ранговая корреляция. В географических исследованиях иногда приходится обрабатывать быстро и с наименьшими затратами фактический материал, даже если получаются менее точные результаты. В некоторых случаях работают с качественной информацией или с громоздкими вычислениями. В таких случаях для установления зависимости между признаками используется ранговая корреляция.

Процесс упорядочения вариант по какому-либо признаку (например, увеличение или уменьшение количества населения по районам) называют ранжированием. Каждому члену ранжированного ряда присваивается *ранг.* Для обозначения рангов, как правило, используются числа в пределах единиц и десятков, например: 1, 2, 3, ..., n. Первой варианте или группе вариант присваивается ранг 1, второй варианте или группе – 2 и т. д. Следует иметь в виду, что одни и те же варианты в зависимости от цели группировки могут иметь различные ранги. Величина ранга не позволяет нам судить о том, насколько близко друг к другу расположены на шкале измерения различные варианты совокупности или качественные признаки.

Ранговую корреляцию можно применять для всех упорядоченных признаков (например, экспертные оценки, баллы, бонитеты). Объем сопряженных выборок должен быть не менее пяти.

Для расчета зависимости (*x, y*) существуют следующие коэффициенты ранговой корреляции: коэффициент неупорядоченности $r_{н}$ и коэффициент Спирмена $r_{с}$. В большинстве случаев используется коэффициент ранговой корреляции Спирмена, который представляет собой следующее соотношение:

$$r_c = 1 - \frac{6\Sigma(x' - y')^2}{{N_n}^3 - N_n}, \text{ или } r_c = 1 - \frac{6\Sigma(d^2)}{{N_n}^3 - N_n},$$

где d – разность между сопряженными рангами; x' – величины рангов, заменяющие фактические варианты или качественные признаки по аргументу *x;* y' – величины рангов, заменяющие фактические варианты или качественные признаки по функции *y;* $N_{п}$ – количество сопряженных пар.

Достоверность полученного рангового коэффициента можно установить аналогично достоверности линейного коэффициента корреляции.

1.7. Регрессионный анализ

Логическим продолжением корреляционного анализа является регрессионный анализ, который развивает и углубляет представление о корреляционной связи. Если корреляционный анализ позволяет установить лишь форму и тесноту связи между случайными переменными, то регрессионный анализ математически описывает выявленную связь, т. е. дает возможность численно оценить одни параметры через другие. Составив и решив уравнения регрессии, можно произвести выравнивание эмпирических линий регрессии, т. е. моделировать наблюдаемую зависимость путем подбора функции, график которой представляет собой теоретическую линию регрессии. Если подобранная функция отражает сущность процесса или явления, то возможно прогнозирование значений признака за пределами сделанных наблюдений.

Подобно корреляции, регрессия может быть *парной* (простой) и *множественной,* по форме связи – *линейной* и *нелинейной*, по зависимости – *односторонней* (изменяется лишь один признак под влиянием другого) и *двусторонней* (изменяются оба признака под воздействием друг друга).

Регрессия выражается несколькими способами: построением эмпирических линий, составлением уравнения и затем – построением теоретических линий регрессии, а также с помощью коэффициента регрессии. Уравнение наиболее точно выражает зависимость между двумя переменными (*x,y*), если корреляция между ними близка к единице.

Регрессионный анализ возможен при наличии всего лишь нескольких пар сопряженных наблюдений, но при условии сильных связей между признаками ($r \geq 0{,}7$).

Для вывода уравнения линейной регрессии достаточно двух пар наблюдений. Обычно рядом с уравнением регрессии приводится коэффициент корреляции или корреляционного отношения: $y = 0{,}2x + 1{,}3$, $r_{0,95} = 0{,}75$ (это обусловлено практическим использованием уравнения регрессии). В MS Excel вместо r приводится R^2 (коэффициент детерминации, величина достоверности аппроксимации, показывающий долю взаимной связи между признаками).

При достаточно сильном влиянии аргумента (x) на функцию (y), имея данные по аргументу, можно по формуле уравнения регрессии вычислить значение функции, не прибегая к полевым наблюдениям.

Существует два способа составления уравнений регрессии: а) способ координат точек, с использованием двух-трех точек, расположенных на эмпирической линии (желательно в начале, середине и конце ее),– для тех случаев, когда расчет не требует большой точности; б) способ наименьших квадратов, более точный, так как для составления уравнения регрессии привлекаются все сопряженные наблюдения.

Линейная регрессия на графике изображается в виде прямой так, чтобы точки эмпирической линии располагались по обе стороны ее и по возможности ближе к ней.

Известно следующее уравнение линейной регрессии:

$$y = ax + b$$

где y – значение зависимой переменной (признак); x – значение независимой переменной (фактор, влияющий на признак); a – коэффициент регрессии, показывающий степень зависимости между переменными (может быть также выражен тангенсом угла наклона линии регрессии к оси абсцисс); b – ордината линии, показывающая смещение начала прямой относительно начала координат.

Степень совпадения теоретической и эмпирической линии регрессии можно проверить, используя критерий хи-квадрат. Если $\chi_ф^2 > \chi_т^2$, то можно указать на недостаточное соответствие теоретической линии регрессии эмпирическому ряду.

Составленные уравнения регрессии можно проверить на точность зависимости между переменными (x, y) не только по критерию хи-квадрат, но и по коэффициенту точности выравнивания линии r_1, отражающему степень приближения (соответствия) фактических данных наблюдения к вероятным. Этот коэффициент определяем по формуле, в которой $(y_ф - M_ф) = \alpha$ – отклонение индивидуальных вариант от общего среднего арифметического по y; $(y_ф - y_в) = \beta$ – отклонение индивидуальных экспериментальных вариант по y от расчетных по уравнению.

Принято считать: если $r_1 > 0{,}95$, то уравнение регрессии соответствует более точному положению линии на графике. При $r_1 < 0{,}95$ необходимо найти другую математическую зависимость. В приведенном примере $r_1 = 0{,}88 < 0{,}95$, поэтому следует подобрать другую математическую зависимость. Такие же выводы получены при проверке на точность зависимости между переменными по критерию хи-квадрат. Оба критерия оценки (χ^2, r_1) на точность выравнивания линии уравнения регрессии используются и для других форм регрессионной зависимости (гиперболической, параболической).

Гиперболическая зависимость. При нелинейной зависимости между аргументом и функцией, представляющая собой на графике кривую в виде гиперболы. Общее уравнение регрессии для гиперболической зависимости имеет вид:

$$y = a/x + b$$

где x – аргумент; y – функция; a и b – коэффициенты, величину которых следует установить. Правильный вид записи уравнения регрессии:

$$y = \mathrm{a}/x + \mathrm{b};\ \eta_{0,95} = 0{,}84$$

Параболическая зависимость. Общее уравнение параболы n-го порядка имеет вид

$$y = ax^n + bx^{n-1} + cx^{n-2} + \ldots + kx + l.$$

Если ограничиться второй ступенью независимой переменной величины x, будем иметь частный случай параболы второго порядка:

$$y = ax^2 + bx + c$$

Для вычисления коэффициентов $a,\ b,\ c$ по способу наименьших квадратов используется общее уравнение параболы второго порядка.

Множественная регрессия. Если при установлении зависимости между признаками используется больше одной независимой переменной, то применяют множественный регрессионный анализ. Проведение такого анализа возможно в следующих условиях: распределение зависимой переменной при различных значениях независимых должно быть близко к нормальному; дисперсия зависимой переменной при разных значениях признаков x должна считаться одинаковой.

С увеличением числа признаков и в случаях нелинейной множественной регрессии необходимо использовать специализированное программное обеспечение. В простейшем случае, когда один признак зависит от двух факторов, общее уравнение линейной множественной регрессии имеет вид

$$\mathrm{y} = \mathrm{a} + \mathrm{bx} + \mathrm{cz}$$

Соответствие между теоретическими (y') и эмпирическими (y) значениями признака устанавливают с помощью критериев хи-квадрат или Стьюдента.

1.8. Факторный анализ

При изучении взаимного влияния многих процессов и явлений в последнее время все чаще обращаются к методам многомерного статистического анализа, в частности факторного анализа. Методы многомерного статистического анализа практически всегда выполняются на специализированном ПО.

Факторный анализ основывается на использовании статистических знаний (вычислении стандартных отклонений, знании корреляционного и регрессионного анализов). В большинстве случаев исследуется система корреляций, отраженных в корреляционной матрице.

Факторный анализ представляет собой ветвь математической статистики, цель которого – разработка моделей, понятий и методов, позволяющих анализировать и интерпретировать массивы экспериментальных данных независимо от их физической природы. Анализ данных включает краткое описание распределения объектов, установление взаимоотношения процессов и явлений, отражающихся в виде *параметров.*

Используемый набор моделей и методов предназначен для «сжатия» информации, содержащейся в корреляционной матрице. В основе различных моделей факторного анализа лежит следующая гипотеза: параметры – это косвенные характеристики объекта или явления и представляют в совокупности тот или иной фактор. В связи с этим задача факторного анализа состоит в том, чтобы показать наблюдаемые параметры в виде линейных комбинаций факторов. Изменение фактора не всегда одинаково отражается на параметрах, поэтому среди последних могут быть выделены группы, реагирующие на каждый из факторов порознь. Параметры, входящие в одну и ту же группу, сильно коррелируют между собой; параметры, входящие в разные группы, слабо коррелируют между собой. Задача выявления факторов понимается как разбиение параметров на группы таким образом, чтобы можно было описать взаимоотношения между параметрами.

Разработано несколько вариантов факторного анализа с использованием коэффициентов только линейной корреляции (нелинейная корреляция вызывает затруднения при обработке материала). Наиболее употребительны при этом *метод главных компонент, метод главных факторов* и *центроидный метод.*

Метод факторного отображения используется при решении ряда географических задач: для целей инженерно-географического районирования и количественной оценки влияния природных условий на производство; для организации отдыха и т. д. По матрицам значений факторов можно составлять картосхемы, на основе которых осуществляется территориальный анализ выражения важнейших факторов.

Элементами исходной матрицы в факторном анализе являются коэффициенты корреляции. В ходе анализа вычисляется также общая дисперсия σ^2, указывающая, в каких границах находятся значения параметров, которые характеризуют фактор. Кроме общей дисперсии в анализе учитывается факторная дисперсия (общность) и специфическая дисперсия, связанная с некоторой переменной и характеризующая только ее. Дисперсию, обусловленную ошибкой, стремятся свести к минимуму.

Элементы столбцов факторной матрицы представляют собой факторные нагрузки, или коэффициенты факторного отображения, выраженные коэффициентами корреляции данной переменной с данным фактором. Таким образом, коэффициенты факторного отображения характеризуют фактор и его влияние на все параметры.

Часто целями факторного анализа являются: сокращение числа переменных и определение структуры взаимосвязей между переменными. Поэтому факторный анализ используется или как метод сокращения данных или как метод классификации.

Метод главных компонент – это метод, который переводит большое количество связанных между собой (зависимых, коррелирующих) переменных в меньшее количество независимых переменных, так как большое количество переменных часто затрудняет анализ и интерпретацию информации.

Факторный анализ отличается от метода главных компонент тем, что в его основе лежит предположение о некотором небольшом количестве фундаментальных переменных, которые не могут быть измерены прямо. Основное отличие между факторным анализом и методом главных компонент заключается в том, что главные компоненты являются линейными функциями от наблюдаемых переменных, в то время как общие факторы не выражаются через комбинацию наблюдаемых переменных.

1.9. Методы линейного программирования

Линейные модели активно используются в экономике и экономической географии как достаточно эффективные в ряде ситуаций. Линейная функция (тройное правило) самая удобная, простая, хорошо разработанная математическая модель.

Линейность – это свойство математических выражений и функций. Выражение типа $ax + by$, где x, y – переменные величины, a, b – постоянные числа, называется линейным относительно переменных x, y. Если переменных больше двух (x_1, x_2, ...x_n), линейное выражение относительно их имеет вид:

$$a_1x_1 + a_2x_2 + \ldots a_nx_n.$$

В линейное выражение все переменные входят в первой степени и никакие переменные не перемножаются.

Линейное программирование – это совокупность методов решения экстремальных задач, в которых цель (критерий оптимальности) и условия (ограничения) заданы уравнениями и неравенствами первой степени. Программирование используется в данной ситуации как планирование, линейное – означает, что ищется экстремум линейной целевой функции при линейных ограничениях (линейных уравнениях, линейных неравенствах). Однако вычислительные средства при решении задач этого класса играют существенную роль в повышении эффективности их приложений.

Линейное программирование эффективно при решении следующих задач:

- составление смеси продукции предполагает выбор наиболее экономичного топлива, пищевых продуктов и т. д.;
- задачи производства – подбор наиболее выгодной производственной программы выпуска одного или нескольких видов продукции при использовании некоторого числа ограниченных источников сырья;
- задачи распределения, или транспортные задачи;
- комбинированные задачи – производство товаров в разных местах, задачи производства и распределения объединяют в единую задачу.

Все модели линейного программирования состоят из стандартных составных частей: совокупность основных переменных, линейные ограничения (условия), целевая функция, определяющая критерий оптимальности задачи.

Совокупность основных переменных характеризует размеры землепользований, площади, объемы производства, затраты материальных, трудовых, финансовых ресурсов.

Система линейных ограничений (условий) определяет область допустимых значений основных переменных. Каждое отдельное условие отражает реальное ограничение (нормы внесения удобрений, выполнение контрольных цифр бизнес-плана и т. д.).

Целевая функция представляется показателем, который обобщенно характеризует один из аспектов деятельности хозяйства данной землеустроительной задачи, например, чистый доход, валовую продукцию и т. д.

Критерий оптимальности в зависимости от условий задачи требует максимизации или минимизации целевой функции при заданных ограничениях.

Распределительный метод среди задач линейного программирования получил распространение из-за упрощения расчетов, точности вычислений и снижения затрат времени на ввод исходной информации. Метод предложили А. Толстой и Л. В. Конторович в 1939–40 гг. Первоначально он применялся в задачах, связанных с транспортировкой грузов, их распределением между поставщиком и потребителем, поэтому получил название «транспортная задача». Применяется при решении ряда землеустроительных задач: распределение севооборотов и угодий по участкам, размещение культур на землях различных категорий, перераспределение участков между хозяйствами для экономии транспортных затрат и др.

Особенности распределительных транспортных задач следующие:

- условия задачи описываются уравнениями (в симплекс-методе описываются и неравенствами);
- все переменные выражаются в одних и тех же единицах измерения;
- во всех уравнениях коэффициенты при переменных равны единице;
- каждая переменная встречается только в двух уравнениях системы ограничений: в одном по строке (по запасам) и в одном по столбцу (по потребностям).

Целевая функция Z выражает суммарные расходы на транспортировку грузов. Ограничения по запасам и по потребностям означают, что сумма ресурса, забираемого из i-го источника, должна быть равна запасу ресурса в нем, как и сумма ресурса, доставляемого j-му потребителю, должна быть равна его потребности.

Величина C_{ij} может выражать транспортные расходы (минимизация) или прибыль от транспортных операций (максимизация) и другие показатели.

Для решения транспортных задач разработан ряд методов: функционала, потенциала, дельта-метод, лямбда-задача. Используются модифицированные модели: транспортно-производственная, многоэтапная, многопродуктовая.

Транспортно-производственная задача. В географических исследованиях, посвященных вопросам определения границ зон сбыта продукции или рациональных связей по прикреплению потребителей к поставщикам, должны учитываться не только транспортные, но и производственные затраты. Такие задачи получили название транспортно-производственных. В качестве $c_{ij} = S_i + t_{ij}$ выступают транспортно-производственные затраты, т. е. S_i – затраты на производство единицы продукции (себестоимость, цена единицы продукции или приведенные удельные затраты) i-м поставщиком; t_{ij} – затраты на перевозку продукции между i-м поставщиком и j-м потребителем. Если увеличить или уменьшить на одну и ту же величину все показатели c_{ij} в матрице или в строке, или в столбце, то свойства матрицы не изменятся. Суммарные мощности поставщиков равны суммарному спросу потребителей. Следовательно, какой бы ни была стоимость производства, потребители для удовлетворения своего спроса возьмут продукцию у всех поставщиков. От каких поставщиков получит каждый потребитель продукцию, зависит от транспортных затрат.

Решение открытой транспортно-производственной задачи должно учитывать показатель S_i, например, себестоимость продукции. При суммарной мощности поставщиков, предположим на 20 единиц, превышающих суммарный спрос потребителей, у последних появляется свобода выбора в получении продукции от более выгодных поставщиков, поэтому оптимальный план может быть экономически более эффективным.

Модель транспортно-производственной задачи при введении дополнительных условий можно использовать для оптимизации развития и размещения промышленного и сельскохозяйственного производства, получить ответ, где должны располагаться новые торговые объекты и т. д.

1.10. Тренд-анализ

Ряд расположенных в хронологической последовательности значений статистических показателей представляет собой *временной (динамический) ряд*.

Статистические показатели, характеризующие изучаемый объект, называют *уровнями ряда*. В динамическом ряду они могут быть *абсолютными, относительными или средними величинами.* Ряды динамики, представленные за определенный промежуток времени, называются *интервальными*. В результате суммирования уровней интервального динамического ряда получаем *накопленные итоги.*

Вследствие многих обстоятельств однородность величин, составляющих динамический ряд, может нарушаться, и таким образом изменяется сопоставимость уровней динамического ряда. Если каждый уровень динамического ряда сравнивается с одним и тем же предшествующим уровнем, как правило, первоначальным – это сравнение с *первоначальной базой*. Если сравнение проводится с предшествующим уровнем – это сравнение с *переменной базой*.

Для представления модели динамического ряда используется *аналитическое выравнивание ряда динамики*. Закономерно изменяющийся уровень изучаемого показателя оценивается как функция времени.

Выбор формы кривой определяет результаты экстраполяции тренда. Одним из наиболее распространенных приемов сглаживания уровней первоначального ряда динамики – это *метод скользящей средней*.

Выполнить прогноз по уравнению тренда можно путем экстраполяции тенденции, наблюдавшейся в прошлом. Уровень динамического ряда ($\hat{y}$), полученный в результате экстраполяции, используется для определения прогнозного значения на будущее.

Наличие зависимости между последующими и предшествующими уровнями динамического ряда называют *автокорреляцией*, а построение модели зависимости будущих значений рассматриваемого показателя от прошлых его значений называется *авторегрессией.*

Ряд исследований проводятся длительное время (мониторинг), чтобы выявить тенденцию или закономерность развития и прогнозирования какого-либо процесса или явления. Для оценки таких событий используют динамические ряды (тренд-анализ). Они представляют собой однородные статистические величины, показывающие изменение явления или процесса во времени.

С помощью тренд-анализа описываются характерные тенденции изменения явления во времени, подбираются статистические модели, описывающие эти изменения, производится поиск промежуточных значений путем интерполяции, предсказание результатов значений в перспективе (экстраполяция).

Динамические ряды бывают *простые* (описание одного явления), *сложные* (несколько явлений), *производные* (составленные из средних или относительных величин), *моментный* (оценка события за определенный момент времени), *интервальный* (анализ явления за год, полгода, месяц).

На первом этапе статистической обработки динамических рядов анализируются основные тенденции (*тренд*) изменения явления во времени. Используется графическое изображение, которое дает исчерпывающую информацию. Вычисляется комплекс специальных показателей, позволяющих дать количественную оценку динамики анализируемого явления.

Абсолютный прирост или *убыль* характеризует изменение явления в единицу или интервал времени. Вычитают из данных последующего периода данные предыдущего. Если ряд возрастает, то прирост считается положительным.

Темп роста или снижения – соотношение в процентах последующего уровня к предыдущему и умноженное на 100. Положительный прирост имеет показатель более 100%, отрицательный – менее 100%.

Темп прироста показывает, на сколько процентов увеличился или уменьшился уровень явления. Отражает относительную скорость изменения явления от одного отрезка времени к другому. Вычисляется путем деления абсолютного прироста на предыдущий уровень, либо вычитанием из показателя темпа роста 100. При положительном приросте показатель больше нуля, при отрицательном – меньше нуля.

Абсолютное значение 1% прироста характеризует значение или стоимость 1 % прироста изучаемого явления. Может вычисляться делением абсолютного прироста на темп прироста, или делением показателя предыдущего уровня на 100. «Стоимость» 1 % темпа роста и прироста в различных совокупностях разная.

Показатель наглядности характеризует динамику явления в процентах относительно исходного уровня, который принимается за 100. В отличие от других показателей стоимость одного процента здесь остается неизменной. Однако динамика изменения исходных данных от одного промежутка времени к другому становится менее выразительной.

Средний темп прироста (среднее хронологическое) вычисляется в виде среднего геометрического.

Динамический характер всех используемых показателей может принимать самые разнообразные формы. Например, абсолютные приросты могут быть стабильными, а темпы роста (прироста) при этом увеличиваться или уменьшаться.

Углубленный анализ временных рядов требует использования более сложных методик математической статистики. При наличии в динамических рядах значительной случайной ошибки (шума) применяют один из двух простых приемов – *сглаживание* или *выравнивание* путем укрупнения интервалов и вычисления групповых средних. Этот метод позволяет повысить наглядность ряда, если большинство «шумовых» составляющих находятся внутри интервалов. Однако, если «шум» не согласуется с периодичностью, распределение уровней показателей становится грубым, что ограничивает возможности детального анализа изменения явления во времени.

Более точные характеристики получаются, если используют *скользящие средние* – широко применяемый способ для сглаживания показателей среднего ряда. Он основан на переходе от начальных значений ряда к средним в определенном интервале времени. В этом случае интервал времени при вычислении каждого последующего показателя как бы скользит по временному ряду.

Применение скользящего среднего полезно при неопределенных тенденциях динамического ряда или при сильном воздействии на показатели циклически повторяющихся выбросов (резко выделяющиеся варианты или интервенция).

Чем больше интервал сглаживания, тем более плавный вид имеет диаграмма скользящих средних. При выборе величины интервала сглаживания необходимо исходить из величины динамического ряда и содержательного смысла отражаемой динамики. Большая величина динамического ряда с большим числом исходных точек позволяет использовать более крупные временные интервалы сглаживания.

Более сложным и результативным методом является сглаживание (выравнивание) рядов динамики с помощью различных *функций аппроксимации.* Они позволяют формировать плавный уровень общей тенденции и основную ось динамики.

Наиболее эффективным методом сглаживания с помощью математических функций является *простое экспоненциальное сглаживание.* Этим методом учитываются все предшествующие наблюдения ряда по формуле:

$$S_t = \alpha \cdot X_t + (1 - \alpha) \cdot S_{t-1}\,,$$

где S_t – каждое новое сглаживание в момент времени t; S_{t-1} – сглаженное значение в предыдущий момент времени $t-1$; X_t – фактическое значение ряда в момент времени t; α – параметр сглаживания.

Если $\alpha = 1$, то предыдущие наблюдения полностью игнорируются; при величине $\alpha = 0$ игнорируются текущие наблюдения; значения α между 0 и 1 дают промежуточные результаты. Изменяя значения этого параметра можно подобрать наиболее приемлемый вариант выравнивания. Выбор оптимального значения α осуществляется путем анализа полученных графических изображений исходной и выравненной кривых, либо на основе учета суммы квадратов ошибок (погрешностей) вычисленных точек. Математическое выражение закономерности динамики данных можно получить с помощью *функции экспоненциального сглаживания*.

Суть его следующая: из бесконечного числа линий, которые могли бы быть теоретически проведены между точками, изображающими исходный ряд, выбирается только одна прямая, которая имела бы наименьшую сумму квадратов отклонений исходных (эмпирических) точек от этой теоретической прямой. Выравнивание проводят по уравнению прямой $y = a + bt$, или по уравнению параболы второго порядка $y = a + bt + ct^2$. В основе выбора параболы для выравнивания лежит предположение о том, что не скорость динамики, а ускорение является постоянной величиной. В качестве постоянных величин выступают a, b, c порядкового номера какого-либо периода – t.

Показателем правильности выбора того или иного уравнения служит коэффициент детерминации R^2. Чем ближе его значение к единице, тем больше соответствие фактического и выровненного распределений.

Современные программы статистической обработки позволяют получать различные теоретические кривые в автоматическом режиме. По результатам можно проводить экстраполяцию или интерполяцию рядов.

Достоверность статистического прогноза зависит от степени интеграции взаимосвязи явлений, которая обеспечивает сохранение механизма формирования явления и инерционность характера динамики (темп, направление, устойчивость) на протяжении длительного времени. Экстраполяция на очень большой период времени вперед или назад резко снижает точность прогноза при R^2 меньше 0,6.

Полный курс лекций приведен в учебно-методическом пособии:

Чертко, Н.К. Математические методы в землеустройстве [Электронный ресурс]: учебно-методическое пособие / Н. К. Чертко. – Минск: БГУ, 2014. – Режим доступа: https://elib.bsu.by/handle/123456789/114777 – Дата доступа 15.03.2021.

Презентации лекций по дисциплине «Математические методы в землеустройстве» доступны на сайте факультета географии и геоинформатики БГУ https://geo.bsu.by/index.php/departments/soil-science/present.html.

2. ПРАКТИЧЕСКИЙ РАЗДЕЛ

Практические работы

Тема 1. Описательная статистика. Нахождение сходства или отличия между выборками

Форма выполнения работы: Расчет в Microsoft Excel с помощью формул и надстройки «Пакет анализа» основных статистических показателей, характеризующие данные выборки: среднее арифметическое (M), медиана (Me), наименьшее, наибольшее, коэффициент вариации (V), среднеквадратическое отклонение (σ), дисперсия (σ^2). Нахождения отличия между выборками по критерию Стьюдента.

При выполнении данной работы в качестве исходных данных используются статистические показатели в зависимости от специальности и специализации студентов (содержание тяжелых металлов в почве, площади сельскохозяйственных угодий, количество предприятий и т.д.)

Цель работы: Освоить методику первичной обработки статистической информации и нахождения основных статистических показателей; выявить сходство или отличие между выборками с применением критерия Стьюдента; научиться анализировать полученные данные и представлять их в надлежащем виде.

Методика выполнения работы приведена в учебно-методическом пособии:

Чертко, Н.К. Математические методы в землеустройстве [Электронный ресурс]: учебно-методическое пособие / Н. К. Чертко. – Минск: БГУ, 2014. – Режим доступа: https://elib.bsu.by/handle/123456789/114777 – Дата доступа 15.03.2021.

Тема 2. Однофакторный дисперсионный анализ

Форма выполнения работы: Расчет статистических показателей в Microsoft Excel с помощью надстройки «Пакет анализа». В задании с помощью дисперсионного анализа необходимо установить влияние внесения мелиорантов на величину урожая сельскохозяйственных культур по вариантам опыта.

Дисперсионный анализ позволяет утверждать с определенной долей уверенности наличие влияния на изучаемый объект каждого из условий (факторов) в отдельности или в их сочетаниях, поэтому получил широкое распространение при планировании и постановке полевых и лабораторных экспериментов по установлению влияния тех или иных факторов на объекты.

Цель работы: научиться обрабатывать и анализировать результаты лабораторных и полевых экспериментов с помощью дисперсионного анализа с применением критерия Фишера.

Методика выполнения работы приведена в учебно-методическом пособии:

Чертко, Н.К. Математические методы в землеустройстве [Электронный ресурс]: учебно-методическое пособие / Н. К. Чертко. – Минск: БГУ, 2014. – Режим доступа: https://elib.bsu.by/handle/123456789/114777 – Дата доступа 15.03.2021.

Тема 3. Корреляционный и регрессионный анализ

Форма выполнения работы: Построение графиков, выявление артефактов, формы, направления и тесноты корреляционной связи между явлениями в Microsoft Excel. Анализ корреляционной связи, оценка достоверности корреляционной связи с использованием критерия Стьюдента.

Исследование корреляционных связей между явлениями и объектами – наиболее часто используемый метод анализа географической и землеустроительной информации. Знание зависимости дает возможность решать кардинальную задачу любого исследования – возможность предвидеть, прогнозировать развитие ситуации при изменении влияющего фактора.

Цель работы: Проанализировать корреляционные связи между показателями, выявить форму и направление связи, найти возможные артефакты, научиться анализировать результаты корреляционного анализа, проверять значимость коэффициента корреляции.

Методика выполнения работы приведена в учебно-методическом пособии:

Чертко, Н.К. Математические методы в землеустройстве [Электронный ресурс]: учебно-методическое пособие / Н. К. Чертко. – Минск: БГУ, 2014. – Режим доступа: https://elib.bsu.by/handle/123456789/114777 – Дата доступа 15.03.2021.

Тема 4. Кластерный анализ

Форма выполнения работы: Проведение кластерного анализа областей Республики Беларусь по сельскохозяйственным показателям в среде Statistica.

Кластерный анализ позволяет решать проблему объединения по сходству (кластеризация) объектов, которые характеризуются множеством признаков, выраженных в разных единицах измерения, что применяется в различных видах районирования (физико-географическом, экономико-географическом, в картографии). Метод позволяет повысить объективность выделения провинций и районов, учесть множество параметров, выявить внутреннюю структуру.

Цель работы: Освоить методику выполнения кластерного анализа в программе Statistica, научиться правильно интерпретировать результаты кластерного анализа, использовать дендрограммы для целей многоступенчатой классификации.

Методика выполнения работы приведена в учебно-методическом пособии:

Чертко, Н.К. Математические методы в землеустройстве [Электронный ресурс]: учебно-методическое пособие / Н. К. Чертко. – Минск: БГУ, 2014. – Режим доступа: https://elib.bsu.by/handle/123456789/114777 – Дата доступа 15.03.2021.

Тема 5. Факторный анализ

Форма выполнения работы: Выявления влияния факторов на эволюцию агроландшафтов среде Statistica, изучение и содержательная интерпретация результатов факторного анализа.

Методы многомерного статического анализа и метод факторного анализа позволяет разработка моделей, понятий и методов, позволяющих анализировать и интерпретировать массивы экспериментальных данных независимо от их физической природы. Анализ данных включает краткое описание распределения объектов, установление взаимоотношения процессов и явлений, отражающихся в виде параметров. Гипотеза факторного анализа о существовании небольшого числа скрытых факторов, через которые линейно выражаются все анализируемые переменные и в которых содержится вся существенная информация, позволяющая проводить системное изучение географических объектов.

Цель работы: С помощью факторного анализа или метода главных компонент в среде Statistica оценить плодородие почв в Минском районе под влиянием природных и агротехногенных факторов. Научиться содержательно интерпретировать результаты факторного анализа и метода главных компонент.

Методика выполнения работы приведена в учебно-методическом пособии:

Чертко, Н.К. Математические методы в землеустройстве [Электронный ресурс]: учебно-методическое пособие / Н. К. Чертко. – Минск: БГУ, 2014. – Режим доступа: https://elib.bsu.by/handle/123456789/114777 – Дата доступа 15.03.2021.

Тема 6. Решение землеустроительных задач на оптимальность

Форма выполнения работы: Определение оптимальных размеров фермерского хозяйства с целью получения максимальной прибыли для данных ограничений в Microsoft Excel с помощью надстройки «Поиск решения».

Для решения данной работы используются методы линейного программирование, представляющих собой совокупность методов решения экстремальных задач, в которых цель (критерий оптимальности) и условия (ограничения) заданы уравнениями и неравенствами первой степени. Программирование используется в данной ситуации как планирование, линейное – означает, что ищется экстремум линейной целевой функции при линейных ограничениях.

Цель работы: С помощью надстройки «Поиск решения» в Microsoft Excel найти оптимальное решение задачи по определению оптимальных размеров фермерского хозяйства.

Методика выполнения работы приведена в учебно-методическом пособии:

Чертко, Н.К. Математические методы в землеустройстве [Электронный ресурс]: учебно-методическое пособие / Н. К. Чертко. – Минск: БГУ, 2014. – Режим доступа: https://elib.bsu.by/handle/123456789/114777 – Дата доступа 15.03.2021.

3. РАЗДЕЛ КОНТРОЛЯ ЗНАНИЙ

3.1. Перечень тестов и контрольных заданий

Тесты и контрольные задания проводятся в течение шестого семестра по следующим темам:

1. Основы математической статистики и статистические критерии различия
2. Дисперсионный анализ
3. Корреляционный и регрессионный анализы
4. Кластерный анализ
5. Факторный анализ
6. Методы линейного программирования
7. Тренд-анализ

Примеры тестовых заданий по указанным темам:

1. Метод, используемый для установления связи между явлениями:

линейное программирование *корреляционный анализ*
тренд-анализ *дисперсионный анализ*

2. Критерий Фишера используется в каком анализе?

корреляционном *кластерном*
дисперсионном *факторном*

3. Параметрические показатели среднего положения:

максимум *минимум* *размах*
мода *медиана*

4. В каком анализе используется критерий НСР?
5. По каким формулам вычисляют коэффициент асимметрии Пирсона:

$K_{as} = (M - Me) / \sigma$ $K_{as} = (M - Mo) / \sigma$
$K_{as} = (Mo - Me) / \sigma^2$ $K_{as} = (Mo - m) / \sigma^2$

6. Среднее кубическое рассчитывается для вариационного ряда:

1) со значениями площади 2) времени 3) объема
4) обратно пропорциональной зависимости.

7. Для каких целей используется линейное программирование?

для установления оптимальных условий *оценки связей*
оценки влияния фактора на объект *для моделирования*

8. Перечислите последовательность этапов в факторном анализе.
9. Артефакт рассчитывается по формулам:

$\tau_1 = (x_1 - x_2) / (x_n - x_2)$ $\tau_1 = (x_2 - x_1) / (x_{n-1} - x_1)$
$\tau_n = (x_n - x_{n-1}) / (x_n - x_2)$

10. От чего зависит минимальный необходимый объем в выборочной совокупности:

точности опыта *разнообразия признака*
уровня значимости *средней арифметической*

11. Какие функции используются при тренд-анализе?
12. Что показывает коэффициент детерминации?

13. В каких случаях вычисляют ŋ:

при нелинейной связи *при линейной связи*

при работе с качественными данными

14. При какой величине r проводится регрессионный анализ?

$r \le 0,4$ $r \le 0,7$ $r \ge 0,7$

15. Величина классового интервала в выборке определяется:

$i = (x_{max} - x_{min}) / k$ $i = M / (x_{max} - x_{min})$

16. Какие уровни значимости (α) чаще всего используются в научных исследованиях?

17. Назовите показатели, которые характеризуют разброс вариант в вариационном ряду.

σ^2 r σ M.

18. Назовите метод, который используется для выявления закономерностей развития объектов в пространстве

дисперсионный *корреляционный*

теории графов *теории инфляции*

19. Какое варьирование у выборки при величине V = 23 %

слабое *среднее* *высокое* *аномальное*

20. В каком анализе используется критерий Кайзера?

3.2. Вопросы к зачету по дисциплине

1. Цель и задачи математических методов в землеустройстве.
2. Ранговая корреляция.
3. Генеральная и выборочная совокупность.
4. Регрессионный анализ, виды регрессий.
5. Определение объема выборочной совокупности.
6. Способы составления уравнения регрессии.
7. Правила составления репрезентативных выборок.
8. Моделирование линейного уравнения регрессии.
9. Артефакт и его использование.
10. Моделирование нелинейного уравнения параболической зависимости.
11. Ошибки выборочных критериев.
12. Моделирование нелинейного уравнения гиперболической зависимости.
13. Степень свободы и ее использование.
14. Факторный анализ и его использование.
15. Определение точности опыта, использование показателя.
16. Этапы факторного анализа.
17. Графическое представление вариационного ряда. Деление выборки на классы.
18. Интерпретация результатов факторного анализа.
19. Показатели асимметрии и эксцесса.
20. Области применения линейного программирования.
21. Теоретические распределения.
22. Показатели среднего положения.

23. Показатели разнообразия выборки.
24. Независимые и сопряженные выборочные совокупности.
25. Использование и расчет критерия Стьюдента.
26. Использование и расчет наименьшей существенной разницы (НСР).
27. Использование и расчет критерия Фишера.
28. Использование и расчет критерия Пирсона.
29. Использование дисперсионного анализа.
30. Составление дисперсионного комплекса.
31. Информационный анализ и его применение.
32. Меры теории графов, их назначение.
33. Основные элементы теории графов.
34. Кластерный анализ и условия его применения.
35. Классификация на основе теории графов.
36. Этапы вычислений в кластерном анализе.
37. Правила построения дендрограммы в кластерном анализе.
38. Корреляционный анализ и его использование.
39. Виды связей между явлениями, объектами.
40. Условия для расчета коэффициента корреляции.
41. Моделирование в географии.
42. Условия для расчета корреляционного отношения.
43. Виды моделей и отбор информации.
44. Тренд-анализ.
45. Моделирование уравнения множественной регрессии.

3.3. Организация самостоятельной работы

Самостоятельная работа ведется на основании Положения о самостоятельной работе студентов (курсантов, слушателей), утвержденном Министром образования Республики Беларусь от 06 апреля 2015 г.

По изучаемой дисциплине планируется:

- работа с литературными и статистическими источниками;
- выполнение творческих исследовательских заданий к зачету;
- изучение тем и проблем, не рассмотренных в лекционном курсе;
- выполнение расчетов для курсовых работ;
- подготовка докладов к студенческим научным конференциям;
- написание материалов и тезисов.

Перечень рекомендуемых средств диагностики

- коллоквиумы;
- электронные тесты;
- проверка расчетно-графических работ;
- оценивание на основе модульно-рейтинговой системы;
- оценивание на основе проектного метода.

Методика формирования итоговой оценки

Итоговая оценка формируется на основе 3-х документов:

1. Правила проведения аттестации (Постановление № 53 от 29 мая 2012 г.);
2. Положение о рейтинговой системе БГУ;
3. Критерии оценки студентов (10 баллов).

3.4. Перечень заданий по управляемой самостоятельной работе студентов

УСР 1. Тренд-анализ официальной статической информации

Форма контроля – отчет.

УСР 2. Сопряженное использование кластерного и факторного анализа для изучения и визуализации структуры взаимосвязей в базах данных

Форма контроля – отчет.

УСР 3. Решение задач на оптимальность методами линейного программирования

Форма контроля – отчет.

4. ВСПОМОГАТЕЛЬНЫЙ РАЗДЕЛ

4.1. Рекомендуемая литература

Основная

1. Чертко, Н.К. Математические методы в землеустройстве: учебно-методическое пособие / Н. К. Чертко. – Минск: БГУ, 2014. – 156 с.

Дополнительная

2. Экономико-математические методы и модели в практике землеустройства: учеб. пособие / В.И. Колеснев [и др.]. – Горки: БГСХА, 2006. – 456 с.
3. Боровиков, В. П. Популярное введение в современный анализ данных в системе STATISTICA. Учебное пособие для вузов / В.П. Боровиков. – Москва: Горячая Линия–Телеком, 2018. – 288 с.
4. Методы оптимальных решений: учебник / К.В. Балдин и др.– 3-е изд., стереотип. – М, 2015.
5. Волков, С.Н. Экономико-математические методы и модели в землеустройстве / С.Н.Волков. – Москва: КолосС, 2007. – 696 с.
6. Карпиченко, А.А. ГИС-картографирование факторов накопления тяжелых металлов в почвах города Молодечно / А.А. Карпиченко, А.С. Семенюк // Географические аспекты устойчивого развития регионов: сб. материалов IV Междунар. науч.-практ. конф., Гомель, 27–29 мая 2021 г.; редкол.: А.И. Павловский (гл. ред.) [и др.]. – Гомель: ГГУ им. Ф. Скорины, 2021. – С. 65–69.
7. Карпиченко, А.А. Группировка элювиальных ландшафтов Беларуси с применением кластерного анализа / А.А. Карпиченко // Почвенно-земельные ресурсы: оценка, устойчивое использование, геоинформационное обеспечение: материалы междунар. науч.-практ. конф., Минск, 6–8 июня 2012 г. – Минск: Изд. центр БГУ, 2012. – С. 173–175.
8. Карпиченко, А.А. Использование статистических методов для выявления особенностей накопления тяжелых металлов в поверхностных горизонтах почв / А.А. Карпиченко // Современные направления развития физической географии: научные и образовательные аспекты в целях устойчивого развития: материалы междунар. науч.-практ. конф., Минск, 13–15 нояб. 2019 г. – Минск: БГУ, 2019. – С. 127–130. – Режим доступа: http://elib.bsu.by/handle/123456789/234139 – Дата доступа 15.03.2021.

4.2. Электронные ресурсы

1. Национальный статистический комитет Республики Беларусь [Электронный ресурс]. – Режим доступа: https://www.belstat.gov.by/. – Дата доступа 15.03.2021.
2. Образовательный портал БГУ [Электронный ресурс]. – Режим доступа: http://dl.bsu.by. – Дата доступа 15.03.2021.

3. Реестр земельных ресурсов Республики Беларусь [Электронный ресурс]. – Режим доступа: https://www.gki.gov.by/ru/activity_branches-land-reestr/. – Дата доступа 15.03.2021.
4. Результаты кадастровой оценки сельскохозяйственных земель Республики Беларусь [Электронный ресурс]. – Режим доступа: https://www.gki.gov.by/ru/rezultati_kadastrovoi_ocenki/. – Дата доступа 15.03.2021.
5. Сведения о данных дистанционного зондирования Земли на территорию Республики Беларусь [Электронный ресурс]. – Режим доступа: https://beldzz.by/distantsionnoe-zondirovanie-zemli/karta.php. – Дата доступа 15.03.2021.
6. Электронная библиотека БГУ [Электронный ресурс]. – Режим доступа: http://elib.bsu.by. – Дата доступа 15.03.2021.

4.3. Учебно-методическая карта по учебной дисциплине

Номер раздела, темы, занятия	Название раздела, темы	Количество аудиторных часов				Количество часов УСР	Формы контроля знаний
		лекции	практические (семинарские) занятия	лабораторные занятия	Иное		
1	2	3	4	5	6	7	8
1	**Основы математической статистики**	**8**	**2**			**2**	
1.1	Введение. Развитие и использование математических методов в землеустройстве. Количественные и качественные учетные признаки. Исследование материальных и абстрактных систем.	2					Тест № 1
1.2	Статистические показатели распределения.Генеральная совокупность и выборка. Правила отбора вариант для выборки. Определение объема выборки. Артефакт. Репрезентативность выборок. Показатели центра распределения (средние). Показатели рассеивания вариант.	2	2				Проверка расчетно-графических работ (РГР)
1.3	*Оценка статистических параметров*. Ошибки статистических параметров. Точность опыта.	2					Опрос
1.4	Теоретические функции распределения. Построение кривой нормального распределения.	2				2	Проверка РГР
2	**Методы установления сходства выборок**	**6**	**6**				
2.1	Статистические критерии различия. Критерий Стьюдента, критерий Фишера, критерий Пирсона (χ^2), наименьшая существенная разность (НСР).	2	2				Проверка РГР
2.2	*Дисперсионный анализ*. Составление дисперсионного комплекса. Однофакторный и двухфакторный дисперсионный анализ. Нахождение НСР и F.	2	2				Тест № 2
2.3	*Кластерный анализ.* Порядок работ в кластерном анализе. Составление таблицы исходных показателей. Трансформация исходных данных. Вычисление метрик. Построение или дендрограммы. Классификация.	2	2				Проверка РГР
3	**Методы установления связи между явлениями**	**6**	**4**				

3.1	*Корреляционный анализ.* Виды корреляционной связи. Порядок работ. Линейная корреляция. Вычисление коэффициента корреляции. Нелинейная корреляция. Оценка прямой и обратной нелинейной связи между признаками. Вычисление корреляционного отношения. Множественная корреляция. Ранговая корреляция.	2	2				Тест № 3, Проверка РГР
3.2	*Регрессионный анализ.* Условия и цель применения регрессионного анализа. Виды регрессии. Линейная зависимость. Порядок работ. Способы составления уравнения регрессии. Нелинейная зависимость. Множественная регрессия.	2					Опрос
3.3	*Факторный анализ*. Условия и цель применения факторного анализа. Этапы работ. Выводы.	2	2				Проверка РГР
4	**Организация оптимальной системы землеустройства**	**6**	**4**			**2**	
4.1	*Методы линейного программирования*. Основные теоретические положения. Виды задач. Методы решения. Способы составления базисного допустимого плана. Правила составления цепи. Закрытые и открытые задачи.	1				2	Проверка РГР
4.2	*Метод потенциала.* Составление базисного допустимого плана. Установление степени оптимальности плана по величине характеристики клеток при стремлении функционала к максимуму и минимуму.	1	2				Проверка РГР
4.3	*Дельта-метод*. Условия его применения. Составление базисного допустимого плана. Правила построения открытой цепи и перераспределения поставок с учетом баланса строк. Условия получения оптимального плана.	1					Тест № 4
4.4	*Модификация методов транспортных задач.* Открытая транспортная задача. Максимизация целевой функции. Учет ограничения времени в задачах. Производственная задача. Многоэтапная задача. Многопродуктовая задача. Лямбда-метод.	1					Опрос
4.5	Применение методов линейного программирования при трансформации земель, разработке схем севооборотов, установления границ сырьевых зон для предприятий перерабатывающей промышленности.	2	2				Проверка РГР
4.6	Динамические ряды**.** Условия и цель применения динамических рядов. Виды динамических рядов. Показатели динамического ряда. Сглаживание динамических рядов.	2					Опрос