МИНИСТЕРСТВО СЕЛЬСКОГО ХОЗЯЙСТВА
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФГБОУ ВО «Кубанский государственный
аграрный университет имени И. Т. Трубилина»

ГОУ ЛНР «Луганский национальный
аграрный университет»

БИОМЕТРИЯ

Учебник

Под общей редакцией Л. П. Трошина

Краснодар
КубГАУ
2018

УДК 581.4:[519.2+004.9] (075.8)
ББК 28.56
Б63

Р е ц е н з е н т ы :

В. А. Драгавцев – академик РАН, доктор биол. наук, профессор
(Агрофизический научно-исследовательский институт);

Б. В. Бондарев – доктор физ.-мат. наук, профессор
(Донецкий национальный университет);

Е. В. Луценко – доктор экон. наук, профессор
(Кубанский государственный аграрный университет»)

К о л л е к т и в а в т о р о в :

И. Д. Соколов, Е. И. Соколова, Л. П. Трошин, О. М. Медведь,
О. М. Колтаков, С. Ю. Наумов

Б63 **Биометрия** : учебник / И. Д. Соколов [и др.]; под общ. ред. Л. П. Трошина. – Краснодар : КубГАУ, 2018. – 161 с.

ISBN 978-5-00097-616-6

В учебнике рассмотрены основные понятия биометрии, главные характеристики количественных признаков, законы распределения, построение статистических оценок, критерии значимости, дисперсионный, регрессионный, корреляционный, дискриминантный, факторный и кластерный анализ, вопросы планирования экспериментов.

Предназначен для обучающихся по агрономическим, биологическим специальностям, а также для сотрудников сельскохозяйственных комплексов.

УДК 581.4:[519.2+004.9] (075.8)
ББК 28.56

ISBN 978-5-00097-616-6

ВВЕДЕНИЕ

> Немного найдется областей, для которых высказывание «малое знание опасно» более истинно, чем для теории вероятностей и математической статистики
>
> *Дж. Теннант-Смит*

Если исходные данные подвержены изменениям, то для получения возможно более точного вывода необходимо использовать идеи и технику биометрии.

Точность вывода, полученного с помощью биометрических процедур, зависит от четырех важнейших факторов: 1) воображения и гибкости мышления исследователя; 2) подходящих методов получения выборки; 3) аккуратной регистрации измерений и учетов особенностей объектов выборки; 4) правильного выбора и использования биометрических методов. Ошибочно думать, что хороший биометрический метод может улучшить плохие данные. Биометрический анализ становится тем более эффективным, чем точнее соблюдаются при проведении эксперимента три первых условия.

В 1927 г. известный генетик С. С. Четвериков стал читать студентам Московского государственного университета отдельный курс «Введение в биометрию», но уже в начале 30-х гг. ученый был выслан из Москвы. С середины 30-х гг. в СССР стараниями Т. Д. Лысенко и его соратников при поддержке властей осуществлялся разгром не только генетики, но и биометрии, преследование специалистов этих направлений. Биометрия, способствующая установлению истины и подтверждающая, в частности, правильность выведенных Г. Менделем законов наследования, вообще была не нужна Т. Д. Лысенко. Возражая известному математику А. Н. Колмогорову, Т. Д. Лысенко писал: «…Нас, биологов, и не

интересуют математические выкладки, подтверждающие … статистические формулы менделистов». Следствием печально знаменитой августовской сессии ВАСХНИЛ 1948 г., завершившей разгром генетики в СССР, было уничтожение учебников по генетике и биометрии, исключение биологической статистики как учебной дисциплины из программ подготовки биологов и смежных специалистов. Лишь после осуждения лысенковщины в середине 60-х гг. по ряду специальностей курс «Биометрия» или аналогичные, но иные по названию дисциплины были снова включены в учебные планы.

Освоение биометрии в СССР в 50-х и первой половине 60-х гг. прошлого века было сильно затруднено, поскольку могло происходить лишь путем самообразования с использованием иностранных изданий. Первый автор предлагаемого учебника начинал изучение биометрии по книге Э. Вебер «Основы биологической статистики» на немецком языке (Weber E., 1957). Ситуация изменилась в 1964 г., когда стало возможным издать на русском языке университетский учебник «Биометрические методы», написанный доктором физико-математических наук В. Ю. Урбахом. Во многих отношениях удачный, этот учебник все же был сложноват для студентов-биологов. Еще в большей степени он был труден для восприятия студентов-аграриев, которые и тогда, и сейчас не изучают математическую дисциплину «Теория вероятностей», знание которой необходимо для осознанного применения биометрических методов.

В настоящее время в России биометрия считается частью научной специальности 03.01.09 «Математическая биология, биоинформатика», по которой защищаются диссертации на соискание ученых степеней кандидатов и докторов физико-математических, биологических и медицинских наук.

Использование математико-статистических методов в биологии, сельском и лесном хозяйстве, экологии – обязательный этап современных исследований. На русском языке имеется ряд удачных книг по биометрии (Урбах В. Ю., 1964; Плохинский Н. А., 1970; Доспехов Б. А., 1985; Лакин Г. Ф., 1990). Однако они ориентированы на применение при производстве расчетов устаревшей

вычислительной техники, простейших непрограммируемых калькуляторов. В этих книгах подробно излагается, как шаг за шагом производить вычисления при использовании тех или иных методов. Между тем сейчас, когда существует множество пакетов (комплексов) программ для математико-статистических вычислений на персональных компьютерах (ПК), главной задачей исследователей является выбор адекватного метода обработки данных и нужного пакета программ, умения работать с этим пакетом. При этом излишне знать в деталях ход вычислений, но необходимо четко представлять себе возможности и ограничения используемых методов статистики, суть вычисляемых параметров и критериев различий.

Отсутствие ко времени написания упомянутых руководств мощных ПК делало излишним подробное рассмотрение в них ряда эффективных, но практически не реализуемых без таких устройств математико-статистических методов (множественный корреляционный анализ, множественный регрессионный анализ и др.). Очевидно, что в новых руководствах по биометрии эти ранее мало использовавшиеся из-за громоздкости вычислений методы должны найти свое место.

В настоящее время существует много пакетов программ для математико-статистической обработки данных. Особое распространение в последние годы приобрел мощный и удобный пакет программ системы STATISTICA для Windows, разработанный компанией StatSoft. Он имеет свою систему управления базами данных, совместимую с другими системами, мощный пакет программ с графическим сопровождением, графический редактор и др. Тот, кто освоит систему STATISTICA в среде Windows, практически не будет иметь никаких проблем, связанных с обработкой данных. Вряд ли в ближайшее десятилетие кому-либо удастся потеснить эту систему на рынке программного обеспечения. С учетом этого в нашем учебнике сделан акцент именно на использовании этой системы.

Учебник разработан по дисциплине «Биометрия» для студентов специальностей «Садоводство», «Агрономия», «Лесное хозяй-

ство» и «Экология и природопользование». Поскольку методы математико-статистической обработки предметно не ориентированы, учебник представляет интерес также для студентов иных специальностей (биологов и др.), аспирантов и преподавателей. В качестве примеров были взяты материалы исследований авторов, использовались также работы Н. А. Плохинского (1970), Б. А. Доспехова (1985), Г. Ф. Лакина (1990), О. М. Царенко и др. (2000), Г. А. Стародворова (2006, 2007 а, 2007 б), Н. Н. Буреевой (2007), А. А, Халафян (2007), В. П. Боровикова (2008) и Ю. В. Тарасовой (2011).

Авторы выражают искреннюю благодарность рецензентам: академику РАН, доктору биол. наук, профессору В. А. Драгавцеву (Агрофизический научно-исследовательский институт), доктору физ.-мат. наук, профессору Б. В. Бондареву (ДонНУ), доктору техн. наук, профессору Е. В. Луценко (КубГАУ).

Особо благодарим ректоров обоих учреждений: академика УАЭН, доктора экон. наук, профессора В. Г. Ткаченко (Луганский НАУ) и доктора экон. наук, профессора А. И. Трубилина (Кубанский ГАУ) за поддержку нашей работы и помощь в издании учебника.

1 БИОМЕТРИЯ КАК НАУКА И ЕЕ СПЕЦИФИКА

Биометрия – это совокупность математико-статистических методов, применяемых в биологии, экологии, сельском и лесном хозяйстве и заимствованных, главным образом, из области теории вероятностей и математической статистики.

Термин «биометрия» ввел в науку Ф. Гальтон (1889). В аналогичном смысле используются также понятия «вариационная статистика», «биологическая статистика».

Биометрия – это, прежде всего, математическая обработка данных, полученных при измерениях и подсчетах. Обработка исходных данных позволяет вскрыть неочевидные при рассмотрении исходных данных закономерности, получить значимые выводы. В этом случае зачастую говорят, что обработка невзрачного алмаза превращает его в красивое украшение – бриллиант.

Задачи биометрии многочисленны, но чаще всего это либо установление значимости параметров (в том числе разности средних арифметических значений), либо установление значимости связей.

Значение биометрии

В биологии и смежных с ней науках с успехом применяют описательные методы, не требующие математико-статистической обработки исходных данных. Однако везде, где в результате измерений и подсчетов получают числовые данные, необходимо использование методов биометрии. Пренебрежение методами биометрии или неправильное их использование приводит к неоправданным затратам труда и времени, а главное – к малоубедительным, а нередко и ошибочным выводам.

Долгое время широкое применение математико-статистических методов сдерживалось трудоемкостью вычислений. Сейчас имеются хорошие системы, пакеты, комплексы программ для персональных компьютеров, позволяющие быстро производить нужные вычисления. Необходимо лишь, чтобы работающий с числами исследователь знал биометрию в такой мере, чтобы смог правильно выбрать адекватный стоящей перед ним задаче метод обработки результатов измерений и подсчетов.

Качественные и количественные признаки

Признаки довольно условно делятся на качественные (альтернативные, дискретные, атрибутивные) и количественные (мерные и счетные). Количественные признаки обычно модифицируются при изменении внешней среды. В подавляющем большинстве случаев можно без опасений исходить из предположения о непрерывном варьировании признака под влиянием условий среды.

Примеры:

– качественный признак (из ботаники): завязь – верхняя, полунижняя, нижняя;

– количественные признаки: мерные – высота деревьев – 10 м, 11 м; счетные – количество тычинок – 3, 4 шт.

Математико-статистическая обработка обязательна, если исходные данные числовые (в том числе и по качественным признакам – например, анализ долей).

Точность измерений и вычислений

Все измерения производятся с определенной точностью. Если для измерения длины используется линейка с расстоянием между черточками – 1 мм, то точность единичного измерения составляет 1 мм, с расстоянием между черточками 0,5 мм – 0,5 мм и т. д. В пределах одного опыта (эксперимента) измерения должен проводить один и тот же человек.

Исходные данные могут быть представлены как целыми числами (счетные всегда, мерные – иногда), так и дробными (мерные признаки). При вычислении на основе исходных данных математико-статистических характеристик или параметров обычно получают дробные числа.

ПК производит вычисления с очень большой точностью, но результат не бывает точнее тех данных, на основе которых он получен. Если измерения производились с точностью до единицы, вряд ли среднюю арифметическую следует приводить с точностью большей, чем одна десятая.

Округление чисел

Если за последней сохраняемой цифрой следуют цифры 0, 1, 2, 3, 4, они отбрасываются (округление с недостатком); если же за последней сохраняемой цифрой следуют цифры 6, 7, 8 и 9, то последняя сохраняемая цифра увеличивается на единицу (округление с избытком). Например, числа 45,346; 8,644 и 9,426 округляются до двух десятичных знаков следующим образом: 45,35; 8,64 и 9,43. Если за последней сохраняемой цифрой следует цифра 5 (с нулями или без нулей после нее), то округление осуществляется с недостатком при условии, что сохраняемая цифра четная. Если же сохраняемая цифра нечетная, то округление осуществляется с избытком. Например, числа 3,585 и 3,575 округляются до двух десятичных знаков таким образом: 3,58 и 3,58.

Во всех машинных языках есть функции округления чисел, так что знание правил округления необходимо лишь при работе с исходными данными.

Формы учета результатов наблюдений

Дневники, протоколы, журналы, бланки, формуляры или другие документы учета исходных данных (на бумажных и компьютерных носителях) должны сохраняться в установленном порядке.

Исходные данные могут представлять собой также данные метеостанций, статистических справочников и т. п.

Генеральная и выборочная совокупности

Совокупность, из которой отбирают определенную часть ее членов для совместного изучения, называют генеральной. Ее представляют как бесконечно большую ($n \to \infty$).

Отобранная для исследования часть генеральной совокупности называется выборочной совокупностью или просто выборкой ($n \geq 2$).

Репрезентативность (представительность) выборки

Чтобы выборка хорошо отображала генеральную совокупность, она должна быть репрезентативной. Это достигается способом рандомизации или случайным отбором объектов из генераль-

ной совокупности, обеспечивающим равную возможность для всех членов генеральной совокупности попасть в состав выборки (жеребьевка, таблицы случайных чисел, операторы RND).

Предложены разные способы обеспечения репрезентативности выборок в тех или иных конкретных ситуациях.

Ошибки типичности и систематические ошибки

Ошибки типичности возникают, если выборки нерепрезентативны. Они не могут быть устранены при последующей математико-статистической обработке. Эти ошибки могут возникнуть при несоблюдении условия объективности выборки, так и преднамеренным искажением (подтасовкой) исходных данных (например, селекционер характеристику своего сорта делает на основе выборки лучших растений, а контроля – худших).

Пример случая, когда неслучайный отбор объектов в выборку приводит к ошибкам типичности.

Исследователь посеял в 30 горшков по 4 семени нового сорта растений, а в другие 30 горшков – по 4 семени контрольных растений (стандарта, st). На стадии трех настоящих листьев оставил в каждом горшке по одному растению. При этом оставлял «типичные» для сорта растения, а надо было оставлять растения случайным способом.

В наших экспериментах с модельным растительным объектом арабидопсисом мы оставляем по одному растению в гнезде, удаляя лишние по закону случая.

Систематические ошибки обычно бывают следствием постоянных ошибок измерительных приборов (например, весы всегда показывают на 10 г больше, т. е. не установлены на нуль). Исследователи должны принимать меры к тому, чтобы все используемые ими измерительные приборы были проверены.

Поверка измерительных приборов

Существуют специальные лаборатории для поверки приборов, в том числе в системе Госстандарта.

Большие и малые выборки

Чем больше по объему выборка, тем точнее оцениваются параметры выборочной совокупности (тем меньше ошибки, связанные с переносом знаний, полученных на выборке, на генеральную совокупность). Однако работа с очень большими выборками обременительна, а при работе с крупными объектами (например, взрослые дубы) и невозможна.

Условно выборки подразделяют на малые ($n < 30$) и большие ($n \geq 30$). Исследователи обычно работают с большими выборками, хотя надежную информацию можно получить подчас и при работе с малыми выборками, когда $2 \leq n < 30$, (при рассмотрении критериев значимости).

Умножение и сложение вероятностей

Вероятность (*probability*) события A выражается отношением числа благоприятных (интересующих нас) исходов m к числу всех исходов n, т. е. $P(A) = m / n$. Вероятность – это число в интервале от 0 до 1.

На практике часто приходится умножать и складывать вероятности. Соответствующие теоремы здесь не рассматриваются; приведем лишь по одному примеру этих операций.

Пример 1

У гетерозиготного по одному гену растения генотипа *Aa* вероятность того, что гамета будет нести рецессивный аллель *a*, равна $\frac{1}{2}$ (0,5; 50 %). Чему равна вероятность того, что потомок от самооплодотворения будет гомозиготным по рецессивному аллелю *a*, т. е. будет иметь генотип *aa*?

Решение

Очевидно, что в этом случае необходимо, чтобы и участвующая в оплодотворении яйцеклетка, и сливающийся (объединяющийся) с ней спермий имели генотип *a*. Вероятность того, что яйцеклетка будет иметь генотип *a*, равна $\frac{1}{2}$; вероятность того, что

спермий будет такого же генотипа, тоже равна $\frac{1}{2}$. Рассматриваемые здесь события независимые. Из теории вероятностей известно, что одновременная реализация двух независимых событий равна произведению вероятностей частных (отдельных) событий. Она равна $\frac{1}{2}.\frac{1}{2} = \frac{1}{4}$ (0,25; 25 %). В общем, в потомстве ожидается 25 % особей, гомозиготных по рецессивному аллелю *a*.

Пример 2

При моногибридном скрещивании у гороха в F_2 происходит расщепление по генотипу в отношении 1 *AA*: 2 *Aa*: 1 *aa* (в понятиях вероятностей – $\frac{1}{4}AA : \frac{1}{2}Aa : \frac{1}{4}aa$). Из-за доминирования ($a < A$) оба генотипа (*AA* и *Aa)* детерминируют один и тот же фенотип – желтый цвет семян. Чему равна вероятность того, что горошина окажется желтой?

Решение

Вероятность того, что генотип горошины *AA,* равна $\frac{1}{4}$; вероятность того, что генотип горошины *Aa,* равна $\frac{1}{2}$. Общая вероятность, а это и есть вероятность того, что горошина будет желтой, равна сумме вероятностей, т. е. она равна $\frac{1}{4}+\frac{1}{2} = \frac{3}{4}$ (0,75; 75 %).

При решении вопроса о том, следует ли умножать или складывать вероятности, обычно достаточно понять, меньше или больше общая вероятность в сравнении с частными вероятностями (если меньше, то вероятности следует умножать; если больше, то складывать).

Считается, что события, имеющие очень малую вероятность, в единичных случаях (испытаниях) не произойдут, т. е. такие события рассматривают как практически невозможные. Если же вероятность события достаточно велика, его принято считать практически достоверным (принцип практической уверенности в прогнозировании исходов случайных событий).

Вопросы для обсуждения

1. Объясните понятие «биометрия», задачи и значение этой науки.

2. На основании чего признаки подразделяются на количественные и качественные?

3. Поясните точность измерений и расчетов, а также правила округления исходных данных.

4. Формы учета результатов наблюдений.

5. Рассмотрите понятия генеральной и выборочной совокупностей.

6. Укажите способы обеспечения репрезентативности выборки.

7. Ошибки типичности и систематические ошибки.

8. Большие и малые выборки.

9. Приведите пример, в котором необходимо перемножить вероятности.

10. Приведите пример, в котором вероятности следует складывать.

2 ОСНОВНЫЕ ХАРАКТЕРИСТИКИ КОЛИЧЕСТВЕННЫХ ПРИЗНАКОВ

Для описания количественных варьирующих признаков используются логически и теоретически обоснованные показатели, называемые «статистическими характеристиками», «характеристиками», «статистиками», «параметрами». При рассмотрении генеральных совокупностей обычно рекомендуется использовать слово «параметры», при изучении выборок – слово «статистики».

Среднее арифметическое значение признака и другие средние

Значение признака в выборке оценивается средней величиной.

Обычно используется среднее арифметическое значение признака, часто называемое сокращенно просто средним значением или средней (по-английски means).

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum x_i}{n},$$

где $\bar{x}$ – среднее арифметическое значение признака; x_1, x_2, …, x_n – значения отдельных (частных) значений измерений или подсчетов (дат); n – объем выборки; i – номера дат (от 1-й до n-й).

Пример:

Имеется выборка: 179, 180, 181 (рост людей в сантиметрах).

Тогда $x_1 = 179$, $x_2 = 180$, $x_3 = 181$; $n = 3$.

$$\bar{x} = \frac{179 + 180 + 181}{3} = 180 \text{ см}.$$

Ответ: Средняя высота людей в этой выборке равна 180 см.

Редко используются среднее геометрическое значение ($\bar{x}' = \sqrt[n]{x_1 \times x_2 \times \ldots \times x_n}$), мода (величина, наиболее часто встречающаяся в данной совокупности), медиана (средняя, относительно которой ряд распределения делится на две равные части) и другие средние.

Показатели изменчивости

Если средняя оценивает значение признака в выборке, то показатели изменчивости оценивают степень варьирования по этому признаку.

Эти показатели подразделяются на абсолютные и относительные.

Среди абсолютных показателей изменчивости особую роль играет выборочная ***дисперсия,*** или ***варианса*** (англ. *variance*).

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1},$$

где s^2 – дисперсия, $n - 1$ – число степеней свободы.

Как видим, в числителе стоит сумма квадратов отклонений (разностей) частных значений x_i и средней арифметической. Эту оценку дисперсии называют ***несмещенной*** в противовес ***смещенной*** (при вычислении последней в знаменателе стоит не $n - 1$ (число степеней свободы), а n – объем выборки. Заметим, что при увеличении n различия между смещенной и несмещенной оценками становятся все меньше. Разумеется, в общем случае лучше использовать несмещенную оценку дисперсии.

Сложное понятие «степень свободы» здесь не рассматривается.

При вычислении дисперсий разности $x_i - \bar{x}$ возводятся во вторую степень (в квадрат), поэтому дисперсию называют статистикой второго порядка. В отличие от дисперсии, средняя арифметическая – статистика первого порядка.

Рассмотрим вычисление s^2 на том же *примере:*

$x_1 = 179, x_2 = 180, x_3 = 181; \bar{x} = 180.$

Тогда $s^2 = \frac{(179-180)^2 + (180-180)^2 + (181-180)^2}{3-1} = \frac{1+1}{2} = 1.$

Ответ: дисперсия $s^2 = 1$.

Среднее квадратичное отклонение (стандартное отклонение, англ. *standart deviation*).

Наряду с дисперсией абсолютным показателем изменчивости служит и стандартное отклонение s, представляющее собой корень квадратный из дисперсии:

$$S = \sqrt{s^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}.$$

В нашем условном примере $s = \sqrt{1} = 1$, т. е. $s = s^2$, что в практической работе бывает крайне редко. Параметрический критерий

различия, *F*-критерий Фишера, применим лишь для оценки значимости различий дисперсий, но не стандартных отклонений.

Редко используется ***среднее линейное отклонение*** s_l.

Это усредненные отклонения данных от средней арифметической, взятые по модулю:

$$s_l = \frac{\sum\left|x_i - \bar{x}\right|}{n}.$$

С математико-статистической точки зрения этот параметр хуже s^2 и s.

Из относительных показателей изменчивости в странах СНГ (но не в дальнем зарубежье!) получил широкое распространение ***коэффициент вариации*** (коэффициент изменчивости):

$$C_v = \frac{s}{\bar{x}} \cdot 100(\%),$$

где C_v – коэффициент вариации.

Приведенные здесь формулы являются теоретическими, из них ясна суть используемых статистик. В практической работе с целью повышения точности вычислений нередко используют другие, рабочие, формулы.

Если $C_v < 10$, то изменчивость считается слабой, при $10 < C_v < 25$ – средней, при $C_v > 25$ – сильной (Лакин Г. Ф., 1990).

Тот же *пример*:

$$\bar{x} = 180,\ s = 1.$$

$$C_v = \frac{1}{180} \cdot 100 \approx 0{,}56\%.$$

Вывод: изменчивость в выборке людей по росту слабая.

Еще один условный *пример* (вторая выборка):

измерили высоту трех растений мелкого модельного растительного объекта арабидопсиса. Получили частные значения:

$$y_1 = 1,\ y_2 = 2,\ y_3 = 3.$$

Тогда $\bar{y} = 2$, $s_y = 1$, $C_v = \frac{1}{2} \cdot 100 = 50\%.$

Таким образом, здесь относительная изменчивость сильная.

Заметим, что дисперсии и стандартные отклонения в обеих выборках (люди и растения) одинаковы, а именно $s^2 = s = 1$. По абсолютным показателям изменчивости различий нет. Но разница в показателе C_v существенная, среди растений различия по высоте более значительные, «в разы».

Такие различия в размерах визуально воспринимаются людьми как очень сильные, а такие, менее чем в 1 %, как в примере с человеком, на глаз почти незаметны. Можно сказать, что визуально люди оценивают уровень относительной изменчивости.

К сожалению, оценить значимость различий C_v разных выборок объективно невозможно, поэтому многие специалисты по математической статистике предпочитают этот показатель не использовать.

Показатели изменчивости являются важной, иногда столь же экономически значимой, как и средние значения, характеристикой сортов сельскохозяйственных растений и пород животных. Допустим, при сортоиспытании в течение трех лет получены следующие результаты по двум сортам (в ц/га): 1) 1-й сорт 35, 40, 45; 2) 2-й сорт 30, 40, 50. Какой сорт предпочтительнее? Средние значения сравниваемых сортов одинаковы (40 ц/га). Однако и без вычисления показателей изменчивости очевидно, что второй сорт обнаруживает большую изменчивость урожайности по годам. По этой причине первый сорт является более пригодным для производства (проигрыш по второму сорту в ценах; потребности в дополнительной технике, складах и др.).

В ситуации, когда 1) 1-й сорт 35, 40, 45; 2) 2-й сорт 31, 41, 51 ($\overline{x_1} = 40,\ \overline{x_2} = 41$) без дополнительной информации трудно сказать, какой из них лучше. У второго несколько выше (на 1 ц/га) урожайность (это хороший признак), но явно больше ее изменчивость по годам (плохой).

В связи с этим отметим, что при сортоиспытании обычно ограничиваются сравнением средних, различия в изменчивости хозяйственно ценных признаков обычно объективно не оцениваются. Это неправильно.

Выборочная ошибка средней арифметической

Ошибка репрезентативности (как ее еще называют) всегда имеется. Ее причина – характеристика генеральной совокупности по выборочной.

Эта ошибка составляет: $s_{\bar{x}} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$.

Она тем выше, чем больше абсолютные показатели изменчивости s^2 (дисперсия) и s (среднее квадратичное отклонение). Она тем меньше, чем больше объем выборки n.

Большие выборки ($n \geq 30$) обеспечивают обычно достаточно низкое, приемлемое значение ошибки репрезентативности.

В таблицах и в тексте ошибку обычно приводят рядом со средней ($\bar{x} \pm s_{\bar{x}}$), например 50,2 ± 1,1.

Параметры $\bar{x}$, s^2, s, $s_{\bar{x}}$ нередко называют элементарными одномерными статистиками. Действительно, более простых, более элементарных математически приемлемых параметров совокупности не существует. Эти статистики одномерные в том смысле, что вычисляются по исходным значениям одного признака.

Вопросы для обсуждения

1. Перечислите основные характеристики совокупностей по количественным признакам.
2. Как вычисляется среднее арифметическое значение?
3. Какие другие средние (кроме средней арифметической) вы знаете?
4. Как вычисляется несмещенная оценка дисперсии?
5. Как вычисляется стандартное отклонение?
6. Как определяется коэффициент вариации?
7. Рассмотрите группирование изменчивости на слабую, среднюю и сильную в зависимости от значения коэффициента вариации.
8. Как вычисляется выборочная ошибка средней арифметической?
9. Объясните, почему рассмотренные в этом разделе статистики называются элементарными?
10. Объясните, почему рассмотренные в этом разделе статистики называются одномерными?

3 ГРУППИРОВКА ИСХОДНЫХ ДАННЫХ

Выборочная совокупность, представляющая собой исходные числовые значения $x_1, x_2, \ldots, x_n$, называется ***несгруппированной совокупностью***. Средние арифметические значения и другие элементарные статистики вычисляют по исходным данным, т. е. используя несгруппированную совокупность.

Выявление закономерностей варьирования признака предполагает использование сгруппированной совокупности. Часто данные группируют в таблицы. Среди группировок важное место занимают ***вариационные ряды***.

Вариационным рядом, или рядом распределения, называют двойной ряд чисел, показывающий, каким образом числовые значения признака связаны с его повторяемостью (частотой) в данной статистической совокупности. Например, фактическая урожайность озимой пшеницы в Луганской области (данные 1945–2013 гг.) была следующей (таблица 1).

Таблица 1 – Урожайность озимой пшеницы в Луганской области за период 1945–2013 гг., ц/га

Год	Урожайность	Год	Урожайность	Год	Урожайность
1	2	3	4	5	6
1945	8,1	1968	17,6	1991	32,3
1946	5,0	1969	18,1	1992	34,8
1947	10,3	1970	24,5	1993	31,1
1948	14,7	1971	18,9	1994	26,2
1949	8,0	1972	14,9	1995	21,0
1950	6,9	1973	29,2	1996	20,2
1951	10,0	1974	28,2	1997	23,9
1952	12,6	1975	20,1	1998	15,3
1953	9,6	1976	26,6	1999	12,6
1954	8,1	1977	24,8	2000	9,4
1955	18,3	1978	34,5	2001	33,8
1956	6,0	1979	16,8	2002	26,9
1957	18,9	1980	24,7	2003	16,3
1958	17,9	1981	21,5	2004	27,2
1959	19,2	1982	20,2	2005	31,8

Продолжение таблицы 1

1	2	3	4	5	6
1960	13,3	1983	26,8	2006	18,3
1961	24,9	1984	17,6	2007	22,5
1962	19,6	1985	27,2	2008	39,6
1963	6,9	1986	23,2	2009	24,3
1964	23,8	1987	29,9	2010	24,2
1965	12,6	1988	31,7	2011	25,9
1966	22,7	1989	39,6	2012	27,8
1967	18,0	1990	37,2	2013	24,1

Как видим, средняя урожайность озимой пшеницы в исследованный период времени составляет 21,14 ц/га и сильно варьирует по годам. Об этом свидетельствует значение коэффициента вариации *Cv*, превышающее 40,2 % (таблица 2).

В качестве примера количественного признака мы взяли урожайность озимой пшеницы, но это могла быть и урожайность желудей или еще чего-то. Методы биометрии предметно (объектно) не ориентированы; они пригодны для самых различных признаков.

Таблица 2 – Статистики признака «урожайность озимой пшеницы»

Переменная	Объем выборки	Средняя арифм. и ее ошибка	Станд. отклонение	Коэф. вариации, %	Коэф. эксцесса	Коэф. асимметрии
Урожайность	69	21,14 ± 1,02	8,50	40,22	–0,65	0,05

Приведенные в таблице 1 числа представляют собой несгруппированную совокупность. Наиболее важные статистики этого ряда представлены в таблице 2.

Лимиты (пределы) и размах изменчивости

Вначале следует найти лимиты изменчивости, т. е. x_{min} и x_{max}. В нашей совокупности x_{min} = 5 ц/га (в 1946 г.), x_{max} = 39,6 ц/га (в 1989 и 2008 гг.). Лимиты можно описать так: 5–39,6 ц/га.

Размах изменчивости R равен разности $x_{max} - x_{min}$ = 39,6–5 = = 34,6 ц/га, в данном случае он очень большой. За начало ряда x_0 принимается значение, меньшее или равное минимальному значению в выборке. Таким значением может быть и нуль, если минимальное значение равно или близко к нулю.

В рассматриваемом примере x_0 можно принять равным 0 или 5.

Число классов (групп)

Обычно данные группируют в равноинтервальный вариационный ряд. Число групп или классов в ситуации, когда изменчивость исключительно средовая (особи чистолинейных родителей, особи гибрида первого поколения от скрещивания чистых линий, вегетативные потомки или клоны), можно приблизительно установить, пользуясь таблицей 3.

Таблица 3 – Установление числа классов по числу наблюдений (Лакин Г. Ф., 1990)

Объем выборки *n* (диапазон)	Число классов К
25–40	5–6
40–60	6–8
60–100	7–10
100–200	8–12
Более 200	10–15

Иногда, например, при анализе гетерогенных природных или полученных в экспериментах популяций, число классов следует брать в 2–3 раза большим, чем указано в таблице 3. При этом объемы выборок должны составлять сотни и тысячи особей (результатов учетов).

В нашем примере n = 69, поэтому число классов вариационного ряда может быть от 7 до 10.

Классовый интервал

Классовый интервал I получают путем деления размаха изменчивости на выбранное число классов (R / K).

Если принять K = 7, то I = 34,6 / 7 ≈ 4,9 ц/га; при K = 10, I = 34,6 / 10 ≈ 3,5 ц/га, т. е. I может быть в интервале от 4,9 ц/га до 5,7 ц/га. Лучше принять для классового интервала «круглое» число. Этому условию в нашем примере удовлетворяет число 5, по-

этому принимаем $x_0 = 0$, $I = 5$. Вариационный ряд имеет следующий вид (таблица 4).

Таблица 4 – Вариационный ряд

Значения признака *x* (урожайность озимой пшеницы)	Класс								
	0	5	10	15	20	25	30	35	40
Среднее значение признака *x* в группе	2,5	7,5	12,5	17,5	22,5	27,5	32,5	37,5	42,5
Частота встречаемости в совокупности (f), шт.	–	9	8	14	17	11	7	3	–
Частота встречаемости, %	–	13,0	11,6	20,3	24,6	15,9	10,1	4,3	–

При этом $K = 7$, т. е. в пределах рекомендуемого. Данный ряд является равноинтервальным, поскольку классовый интервал один для всего ряда. Потребность в неравноинтервальных рядах возникает крайне редко.

Графическое представление вариационного ряда

Для того, чтобы более наглядно представить закономерность варьирования количественных признаков, вариационные ряды принято изображать в виде графиков: вариационных кривых (фактически ломаных линий) и гистограмм распределения частот.

Вариационная кривая по данным вариационного ряда, приведенного в таблице 4, изображена на рисунке 1.

Гистограмма (столбчатая диаграмма) распределения частот по данным этого же вариационного ряда изображена на рисунке 2.

Рассматривая эмпирический (фактический, наблюдаемый) вариационный ряд, вариационную кривую и гистограмму, можно утверждать, что чаще встречаются годы с урожайностью, близкой к средней (от 15 до 25 ц/га при средней урожайности около 21,1 ц/га). Наблюдаемое распределение одновершинное, вершина в области среднего значения признака. Годы с очень высокой (более 30 ц/га) и очень низкой (менее 10 ц/га) урожайностью встречались сравнительно редко (см. рисунки 1, 2).

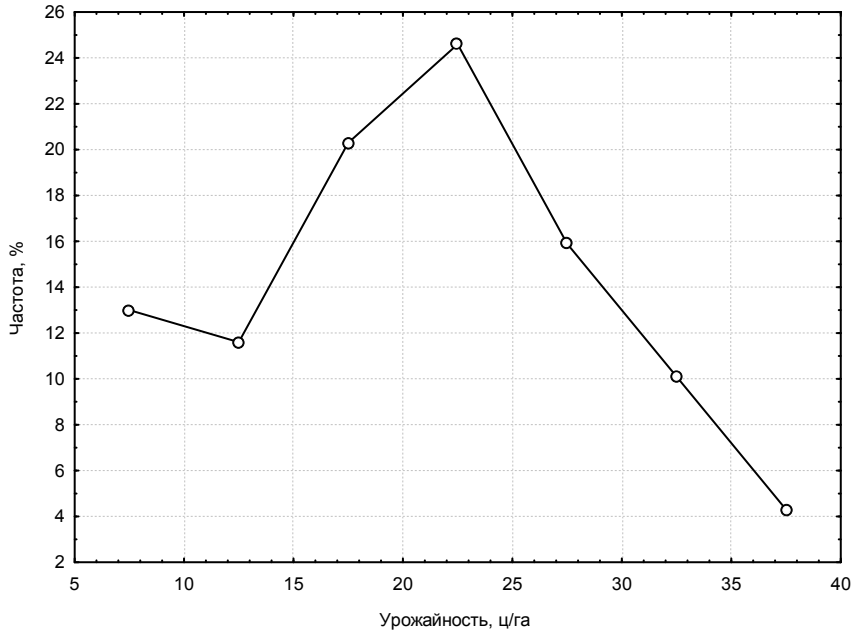


Рисунок 1 – Вариационная кривая

При группировке исходных данных нередко допускаются ошибки, которые исключают проведение стандартной математико-статистической обработки данных, а потому и формулировку надежных выводов.

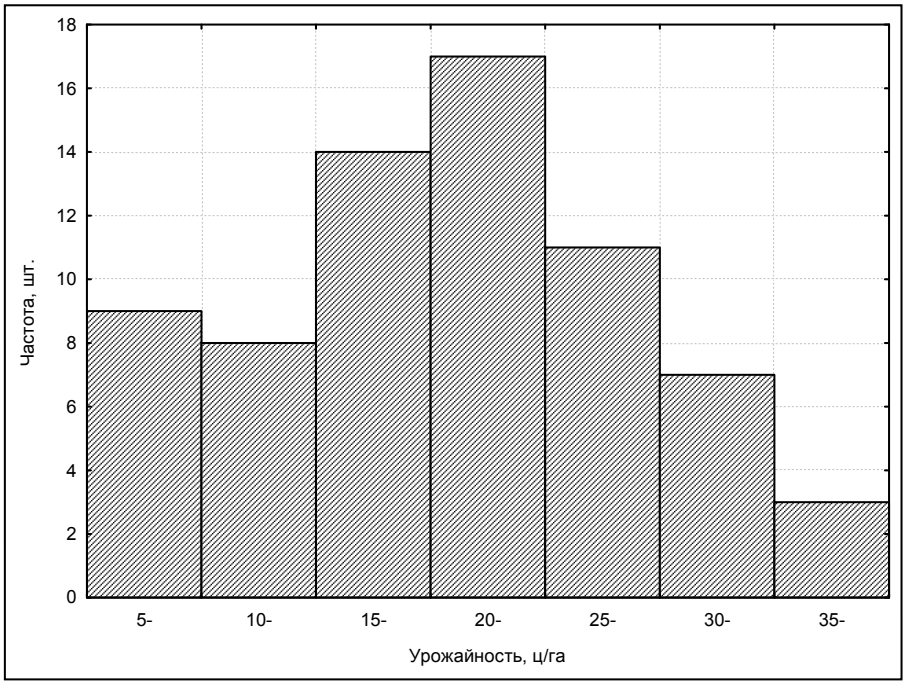


Рисунок 2 – Гистограмма распределения частот

Пример. В 2004 г. опубликована статья, в которой приводится следующая группировка показателей урожайности озимой пшеницы:

Границы классов, ц/га	Менее 10	10–19	20–24	25–29	30–39	Более 40
Частота встречаемости	3	2	5	5	5	3

Замечания. Ряд неравноинтервальный (используется интервал 5 ц/га и 10 ц/га). Обычная статистика в этом случае неприменима.

Число классов взято равным 6, но при объеме выборки $n = 23$ вариационный ряд составлять вообще не рекомендуется (n должно быть не меньше 25 – см. таблицу 2). При указании границ классов слова «менее» и «более» недопустимы.

Границы классов приведены так, что некоторые числа не попадают ни в какой класс. Скажем, предпоследний класс 30–39, а следующий «более 40». А к какому относятся значения 39,1–40?

Достоверность выводов, полученных при анализе подобных рядов, вызывает сомнение.

Существует немало в той или иной мере удобных машинных программ, позволяющих на ПК быстро вычислить элементарные статистики по исходным данным (без группировки совокупности) и составить вариационный ряд.

Вопросы для обсуждения

1. Рассмотрите понятие «несгруппированная совокупность».
2. Рассмотрите понятие «сгруппированная совокупность».
3. Вариационный ряд как один из способов группирования исходных данных. Равноинтервальные и неравноинтервальные ряды.
4. Нахождение лимитов (границ) изменчивости.
5. Объясните размах изменчивости.
6. Число классов (групп) при построении вариационного ряда.
7. Что такое «классовый интервал» и как его находят?
8. Графическое представление вариационного ряда в виде вариационной кривой.
9. Графическое представление вариационного ряда в виде гистограммы.
10. Ошибки, которые допускаются при построении вариационного ряда.

4 ЗАКОНЫ РАСПРЕДЕЛЕНИЯ

Распределения частот встречаемости значений исследуемых признаков в совокупностях могут быть одновершинными, двухвершинными и многовершинными. Одновершинные распределения бывают симметричными и асимметричными, обычными (без значимого эксцесса) и эксцессивными.

Вариационные ряды, вариационные кривые и гистограммы характеризуют эмпирические, или наблюдаемые, распределения.

Эмпирические распределения обычно можно достаточно близко описать одним из теоретических или ожидаемых распределений.

Среди теоретических распределений особое место занимает ***нормальное,*** или ***гауссово, распределение.***

$$f = \frac{n\lambda}{s} \times \frac{1}{\sqrt{2\pi}} \times e^{\frac{-(x_1 - \bar{x}/s)^2}{2}},$$

где f – теоретическая частота в классе;

$\bar{x}$ – средняя арифметическая;

s – среднеквадратичное отклонение;

λ – классовый интервал;

π и e – математические константы.

Теоретическая частота в классе f зависит от средней арифметической $\bar{x}$, среднего квадратичного отклонения s и классового интервала λ. Функция нормального распределения представляет собой колоколообразную кривую. Она симметрична и однозначно описывается параметрами $\bar{x}$ и s. При этом $\bar{x}$ является точкой максимума, через которую проходит ось симметрии.

На рисунке 3 изображены три кривые нормального распределения, причем $\bar{x}_1 = \bar{x}_2 = \bar{x}_3$ (средние равны), при разных значениях s, разных уровнях абсолютной изменчивости ($s_1 < s_2 < s_3$).

По закону, близкому к нормальному, распределены фенотипические отклонения многих количественных признаков.

Нормальное распределение изучаемых признаков позволяет использовать для математико-статистической обработки практиче-

ски любые известные методы, многие из них разработаны для нормального распределения. Можно использовать, в частности, корреляционно-регрессионный и дисперсионный анализы.

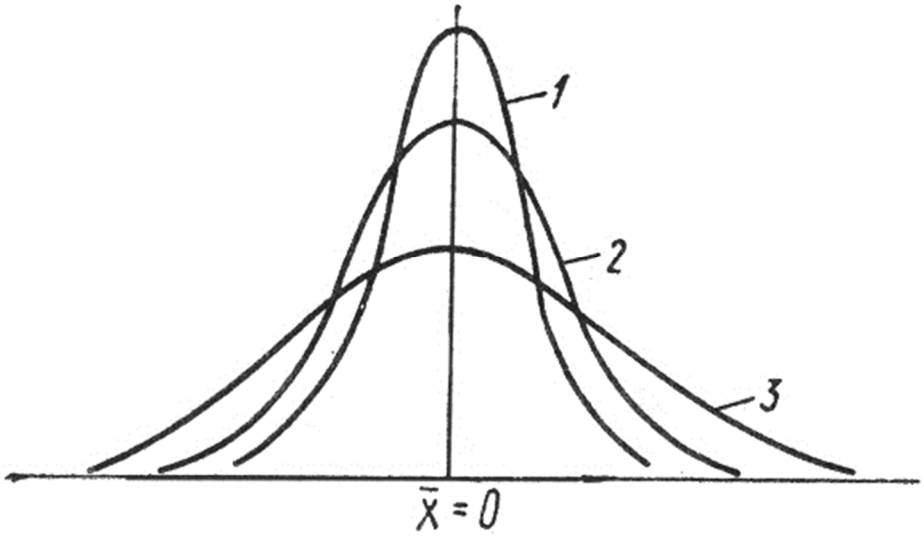


Рисунок 3 – Нормальные кривые (1, 2, 3) при разных значениях параметра *s* (Лакин Г. Ф., 1990)

Математическая обоснованность появления нормального распределения состоит в том, что это предельная форма распределения суммы влияний большого числа величин (переменных, факторов, причин), из которых ни одна сильно «не доминирует» над другими.

Наблюдаемое распределение урожайности озимой пшеницы в Луганской области можно аппроксимировать нормальным. Значит, изменчивость урожайности по годам определяется многими обстоятельствами, и попытки найти один или два решающих фактора изменчивости урожайности обречены на неудачу.

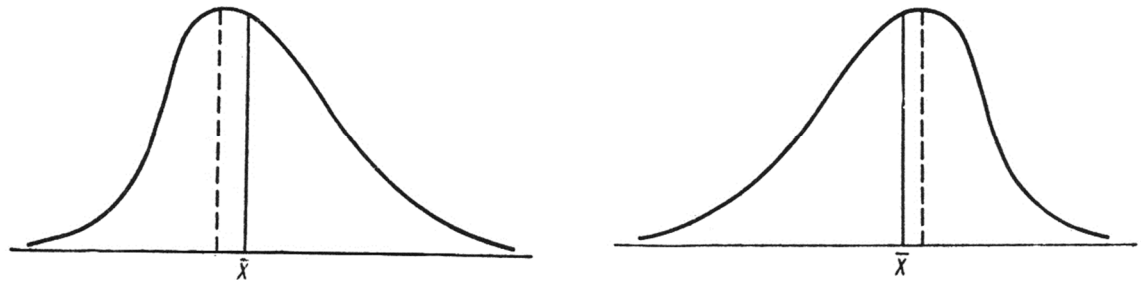
Измерение асимметрии и эксцесса распределений

Для нормального распределения не характерны ни асимметрия, ни эксцесс.

Графически асимметрия выражается в виде скошенной вариационной кривой (рисунок 4), вершина которой может находиться левее $\bar{x}$ (правосторонняя, или положительная, асимметрия) или правее (левосторонняя, или отрицательная, асимметрия).

Правостороннюю асимметрию обнаруживает, например, распределение исходных значений признака «количество дней от посева до цветения» у арабидопсиса.

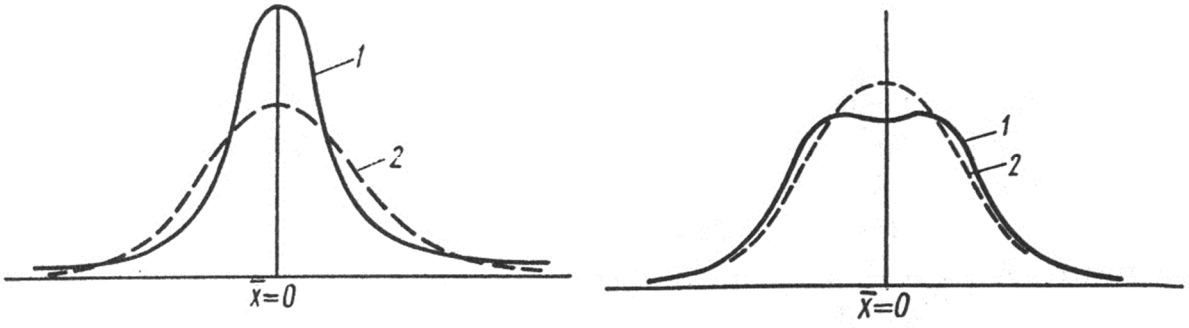
Величина асимметрии оценивается коэффициентом асимметрии As. Если он значим, распределение асимметричное и потому не нормальное.



а б

Рисунок 4 – Асимметричные кривые (Лакин Г. Ф., 1990):
а – положительная асимметрия, б – отрицательная асимметрия

Эксцесс тоже бывает положительный и отрицательный (рисунок 5). Излишне островершинные по сравнению с нормальной кривые имеют положительный эксцесс, излишне плосковершинные – отрицательный.



а б

Рисунок 5 – Эксцесс (1) в сравнении
с нормальной кривой (2) (Лакин Г. Ф., 1990):

а – положительный эксцесс (крутовершинная кривая); б – отрицательный эксцесс (плосковершинная кривая)

Величина эксцесса оценивается коэффициентом *Ex*. Если он значим, распределение эксцессивное и потому не нормальное.

В таблице 2 приведены небольшие по величине коэффициенты асимметрии и эксцесса, позволяющие принимать распределение значений урожайности озимой пшеницы в Луганской области за нормальное.

Ранее речь шла о распределении исходных единичных значений. Для обоснованного применения ряда статистических методов важен, однако, тип распределения не исходных данных, а производных от них параметров. В частности, при сравнении средних по *t*-критерию Стьюдента важно, как распределены выборочные средние, принимаемые за варьирующие величины.

В связи с этим следует отметить:

1. Если исходные значения распределены нормально, то и выборочные средние имеют такие же распределения.

2. Нормально распределены и параметры, полученные как алгебраические суммы нормально распределенных средних.

3. При сравнительно небольших (хотя и значимых) отклонениях распределений исходных величин от нормального закона средние арифметические распределены нормально.

Это значительно расширяет область применения таких критериев различий, как *t*-критерий Стьюдента.

В случаях значительной правосторонней асимметрии иногда используют преобразования шкалы для измерения признака, нормализующие преобразования. Наибольшее употребление нашла логнормальная трансформация $y_i = \lg (x_i + x_0)$, где x_i – исходные значения, y_i – преобразованные значения, а x_0 – неизвестный параметр, значение которого подбирают так, чтобы распределение y_i было близко к нормальному (Лакин Г. Ф., 1990).

Одним из распределений с правосторонней асимметрией является распределение, описываемое формулой **Максвелла**.

При таком распределении $s = 0{,}4224 \cdot \bar{x}$, а коэффициент вариации $C_v = 42{,}24$ %. Кстати, лишь при $C_v < 25$ % аппроксимация эмпирического распределения нормальным обычно бывает правомерной.

Если после вычисления $\bar{x}$ и s окажется, что между ними обнаруживается сходная связь ($s = 0{,}4224 \cdot \bar{x}$), можно производить выравнивание эмпирического вариационного ряда по Максвеллу с последующей оценкой степени соответствия эмпирического и теоретического распределений (обычно по критерию χ^2).

Распределение Пуассона (распределение редких событий)

Графики функции Пуассона приведены на рисунке 6.

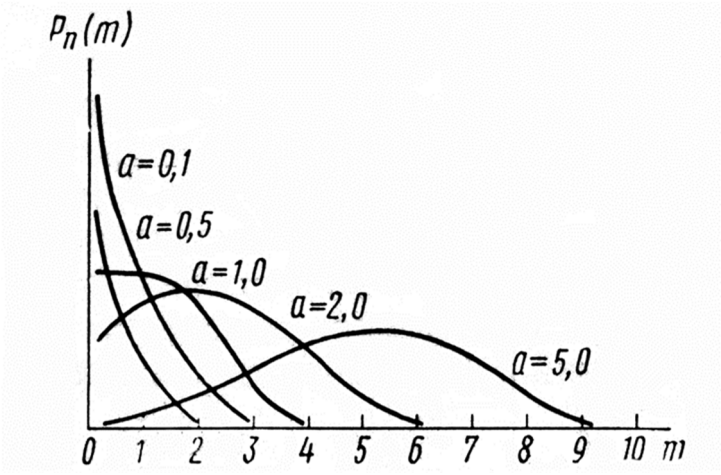


Рисунок 6 – График и функции Пуассона $P_n(m) = \frac{a^m}{m!}e^{-a}$ для разных значений a (Лакин Г. Ф., 1990)

Когда вероятность события очень мала (исчисляется сотыми и тысячными долями единицы), распределение частот таких событий становится крайне асимметричным и хорошо описывается формулой Пуассона.

По закону Пуассона распределяются редкие случайные события. Например, нами установлено, что число (количество) иногда формирующихся лепестков в цветках растений линии *ap1-1* (*apetala1-1*) арабидопсиса, обычно не имеющих лепестков, распределяется по этому закону.

Распределение Пуассона – частный случай биномиального распределения. Оно, как и любое такое распределение, приближается в известной ситуации (при возрастании числа $a \approx np$) к нормальной кривой (см. рисунок 6).

Эмпирические распределения никогда точно не совпадают с теоретическими, но в этом и нет необходимости. Для того, чтобы какое-то наблюдаемое распределение можно было аппроксимировать (считать) тем или иным теоретическим распределением, необходимо и достаточно, чтобы эмпирическое и теоретическое распределения частот ***значимо не различались по известным критериям различий*** (обычно используется для этих целей критерий χ^2).

Вопросы для обсуждения

1. Понятие эмпирического и теоретического распределения. Законы распределения.
2. Особенности нормального (Гауссового) распределения.
3. Математическое обоснование появления нормального распределения.
4. Поясните понятие «асимметрия распределения».
5. Поясните, что означает понятие «эксцесс распределений».
6. Распределение исходных данных и статистик.
7. Опишите распределение Максвелла.
8. Трансформация исходных данных: задачи и способы.
9. Рассмотрите распределение Пуассона.
10. Сравнение эмпирического и теоретического распределений.

5 ОЦЕНКА ЗНАЧИМОСТИ

Целями биометрических исследований обычно бывают либо установление (доказательство с определенной надежностью) различий, либо установление связей между переменными. Для этих целей применяются критерии непараметрические и параметрические.

Нулевая гипотеза (0-гипотеза)

Сущность ее в том, что обнаруженные различия выборочных параметров принимаются за исключительно случайные. Например, если средняя одной выборки $\bar{x}_1 = 15{,}0$; а средняя другой $\bar{x}_2 = 15{,}2$; то нулевая гипотеза исходит из того, что различием $d = 0{,}2$ можно пренебречь и считать его не значимым (не существенным, не достоверным).

Противоположная нулевой альтернативная гипотеза исходит из предположения о том, что генеральные средние значимо различаются.

Вероятность ошибки *p*

Выводы (суждения, заключения) о значимости – незначимости определенным образом связаны с вероятностью ошибки (таблица 5).

Таблица 5 – Вероятность ошибки (Бююль А., Цефель П., 2005)

Вероятность ошибки	Значимость	Обозначение
$p > 0{,}05$	незначимая	*ns*
$0{,}01 < p \le 0{,}05$	значимая	*
$0{,}001 < p \le 0{,}01$	очень значимая	**
$p \le 0{,}001$	максимально значимая	***

Если в тексте работы мы встречаем, например, значение параметра «коэффициент парной корреляции» $r = 0{,}92$***, то это значит, что корреляционная связь максимально значимая. Если $\bar{x}_1 - \bar{x}_2 = 2{,}5$*, то разность средних просто значимая. Иногда говорят о трех уровнях, или порогах, значимости (надежности) различий, трех

уровнях p (1-й – значимая, 2-й – очень значимая, 3-й – максимально значимая).

Вероятность ошибки дополняется до единицы «вероятностью достоверности» P ($P + p = 1$). Вероятность допустить ошибку первого рода (когда нулевая гипотеза отвергается, хотя она верна) равна вероятности ошибки p. Выбор уровня значимости определяется особенностями исследований и тем, насколько серьезными могут быть последствия при ошибочной браковке 0-гипотезы.

В таблице 6 приведены области применения трех порогов надежности (заимствовано из Н. А. Плохинского, 1970, с небольшими изменениями).

Таблица 6 – Три порога надежности (вероятности) безошибочных прогнозов

Порог	Применение	Уровень значимости	Объем выборок
1	В качестве обычных требований надежности в большинстве биологических, экологических, агрономических и лесоводческих исследований	$0{,}01 < p \leq 0{,}05$	$n \geq 30$
2	В качестве повышенных требований надежности при проверочных опытах и в экономических работах	$0{,}001 < p \leq 0{,}01$	$n \geq 100$
3	В качестве высоких требований надежности при разрешении спорных вопросов, при проверке гипотез и при исследованиях вредных и ядовитых веществ	$p \leq 0{,}001$	$n \geq 200$

В подавляющем большинстве случаев в биологии и других науках можно ориентироваться на первый уровень (если $0{,}05 < p$ различия незначимы, принимается 0-гипотеза; если $0{,}01 < p \leq 0{,}05$ различия значимы).

Некоторые авторы пишут о существовании различий при $0{,}05 < p$ (т. е. когда вероятность ошибки больше пороговой на 5 %). Примером могут служить отчеты межправительственной рабочей группы об изменении климата Земли.

Непараметрические критерии

Критерий знаков z

В тех случаях, когда результаты наблюдений можно выразить знаками плюс (+) и минус (–), различия между попарно связанными членами сравниваемых сопряженных выборок оценивают с помощью критерия знаков *z*. Этот критерий применим, если число пар значений $n \geq 6$.

Критерий знаков основывается на простых соображениях: если попарно сравниваемые значения двух зависимых выборок существенно не отличаются друг от друга, то число плюсовых и минусовых разностей окажется одинаковым (или почти одинаковым); если же заметно преобладают плюсы или минусы, это будет указывать на положительное или отрицательное действие изучаемого фактора на результативный признак.

Большее число однозначных разностей используется в качестве фактически найденной величины *z*-критерия знаков ($z_ф$). При этом нулевые разности, т. е. случаи, не давшие ни положительного, ни отрицательного результата, обозначаемые цифрой 0, в расчет не принимают, и число парных наблюдений соответственно уменьшается.

Если $z_ф \geq z_{st}$, то нулевую гипотезу отвергают, если $z_ф < z_{st}$, – ее принимают.

Критические значения *z*-критерия знаков при разных, а именно, первом и втором уровнях значимости и разных количествах пар значений признака приведены в приложении А.

Пример

В учхозе ЛНАУ в некоторые годы для получения зеленой массы использовали и кукурузу, и сорго. Сравнение урожайности зеленой массы сорго и кукурузы по критерию знаков *z* приведено в таблице 7.

Число пар значений в данном случае равно 12, поэтому критерий знаков применим. В шести случаях из двенадцати урожайность сорго превышала урожайность кукурузы, в остальных, наоборот, кукуруза была более урожайной. Общее количество положительных и отрицательных знаков одинаково ($\sum (+) = 6$, $\sum (–) = 6$).

Таблица 7 – Оценка урожайности сорго и кукурузы по критерию знаков *z*

Урожайность сравниваемых культур, ц/га												
Сорго	303,7	123,0	159,0	96,4	135,8	73,0	141,8	1169,0	270,4	133,4	134,0	145,6
Кукуруза	289,4	71,1	246,5	82,5	136,3	56,0	174,4	267,1	351,2	53,1	186,0	204,2
$z_ф$	+	+	–	+	–	+	–	+	–	+	–	–

Большая сумма (в нашем случае это любая из сумм, 6) – это фактический *z*-критерий ($z_ф = 6$). Стандартное значение определяем по таблице (приложение А) для данного числа $n = 12$ $z_{st} = 10$ (первый порог безошибочного прогноза). Поскольку фактическое значение меньше стандартного $z_ф < z_{st}$ (6 < 10), принимается 0-гипотеза. По критерию знаков различия в урожайности между сорго и кукурузой незначимы.

Критерий знаков имеет недостаток: он не учитывает значения разностей между величинами сопряженных признаков. Более надежным, но и более сложным является *T*-критерий Уилкоксона.

T-критерий Уилкоксона

Применим для зависимых выборок, т. е. в том случае, когда члены сравниваемых выборок связаны попарно некоторыми общими условиями.

T-критерий рассчитывают следующим образом.

1. Ранжируют попарные разности, как положительные, так и отрицательные, в один общий ряд. При этом нулевые разности в расчет не принимают, а все остальные независимо от знака ранжируют так, чтобы наименьшая абсолютная разность получила первый ранг, причем одинаковым по величине разностям присваивают один и тот же ранг.

2. Находят отдельно суммы рангов положительных и отрицательных разностей. Меньшую из двух сумм разностей, без учета ее знака, используют в качестве фактически установленной величины *T*-критерия.

3. Сравнивают эту величину $T_ф$ с критическим (стандартным) значением T_{st} для принятого уровня значимости и числа парных

наблюдений n, которое берут без учета нулевых разностей (приложение Б).

Если $T_ф \leq T_{st}$, то различия между средними считаются значимыми, нулевую гипотезу отвергают.

Если $T_ф > T_{st}$, различия между средними считаются незначимыми, нулевую гипотезу принимают.

Пример. Для демонстрации применения T-критерия Уилкоксона возьмем предыдущий пример. Сравнение урожайности зеленой массы сорго и кукурузы приведено в таблице 8.

Таблица 8 – Оценка урожайности сорго и кукурузы по критерию T-Уилкоксона

Урожайность		Разница выражена		Ранги разницы
сорго	кукуруза	знаками	числами	
303,7	289,4	+	15	3
123,0	71,1	+	52	6
159,0	246,5	–	88	10
96,4	82,5	+	13	2
135,8	136,3	–	0.5	1
73,0	56,0	+	17	4
141,8	174,4	–	32	5
1169,0	267,1	+	902	11
270,4	351,2	–	81	9
133,4	53,1	+	80	8
134,0	186,0	–	52	6
145,6	204,2	–	58	7

$T_{(+)} = 3 + 6 + 2 + 4 + 11 + 8 = 34$, $T(–) = 10 + 1 + 5 + 9 + 6 + 7 = 38$.

Меньшая разность дает $T_ф = 34$. Сравниваем эту величину со стандартным значением $T_{st} = 15$ для $n = 12$ и $p = 0{,}05$.

Так как $T_ф > T_{st}$, то принимаем нулевую гипотезу, различия между средними считаем незначимыми.

Таким образом, применяя *T*-критерий Уилкоксона, установили, что и по этому критерию 0-гипотеза состоятельна. Все правильно. Если вывод верный, он будет одинаковым при использовании всех адекватных поставленной задаче критериев различий.

Критерий χ^2 (хи-квадрат)

Критерий согласия, или соответствия, χ^2 применяется для проверки гипотез о законах распределения.

Он представляет собой сумму квадратов отклонений эмпирических частот f от вычисленных или ожидаемых частот f', отнесенную к теоретическим частотам:

$$\chi^2 = \sum \frac{(f_i - f_i')^2}{f_i'} = \sum \frac{d_i^2}{f_i'} \ ,$$

где d_i – разности между наблюдаемыми и ожидаемыми частотами.

Конструкция критерия χ^2 такова, что его значение тем меньше, чем больше совпадение наблюдаемых и ожидаемых частот. При полном совпадении $\chi^2 = 0$.

Пример

Получили 176 растений F_2, из которых 123 имели опушенные листья и стебли, а 53 были голыми. Проверяется гипотеза о расщеплении в отношении 3:1.

Используем критерий χ^2 (таблица 9).

Таблица 9 – Критерий χ^2

	Опушенные	Голые	Всего
f	123	53	176
f'	132	44	176
$d = f - f'$	–9	9	
d^2	81	81	
$\frac{d^2}{f'}$	0,61	1,84	
Примечание. f – фактическая частота, f' – ожидаемая частота.			

По таблице находим стандартные значения χ^2_{st} для трех уровней значимости и числа степеней свободы $\nu = k - 1 = 2 - 1 = 1$, где k – число классов расщепления (фенотипических классов);

$\chi^2_{st} = \{3,8 – 6,6 – 10,8\}$ (приложение В).

Полученное значение $\chi^2 = 2,45$ сравниваем со стандартными (табличными). В нашем случае $\chi^2 < \chi^2_{st}$, что позволяет принять нулевую гипотезу.

$$\chi^2 = \sum \frac{d_i^2}{f_i'} = 0,61 + 1,84 = 2,45.$$

Итак, гипотеза о расщеплении по моногибридной схеме в отношении 3:1 подтверждается.

Обычно принимаемым ограничением на использование критерия χ^2 является требование о том, чтобы f' не была меньше 5 ни в одном из классов (в данном случае минимальная $f' = 44$).

Число степеней свободы подсчитывается неединообразно. Рассмотрение этого вопроса требует немало времени, оно описано в учебных пособиях (например, Лакин Г. Ф., 1990).

В рассмотренном примере наблюдаемое распределение сравнивалось по χ^2 с ожидаемым. Подобным же образом можно проверять эмпирическую совокупность на нормальность, сравнивая эмпирический вариационный ряд с нормальным распределением. Существует много программ для нахождения на ПК теоретических частот нормального распределения, нужно знать только $\bar{x}$ и s эмпирической совокупности.

Используем данные о фактической урожайности озимой пшеницы в Луганской области, уже приводившиеся ранее (см. таблицу 1). На персональном компьютере вычисляем координаты точек, необходимых для построения кривой нормального распределения с $\bar{x} = 21,14058$ и $s = 8,503279$. Гистограмма эмпирического распределения и кривая нормального распределения, полученные в системе STATISTICA, приведены на рисунке 7.

Ход вычисления фактического значения χ^2 иллюстрирует таблица 10. При использовании χ^2 требуется, чтобы значения теорети-

ческих, полученных по нормальному закону частот f' во всех сравниваемых классах, не были меньше 5. Чтобы выполнялось это условие, крайние классы с обеих сторон распределения объединяли в укрупненные классы (группы). В результате вариационные ряды становятся неравноинтервальными, но зато пригодными для сравнения по критерию χ^2.

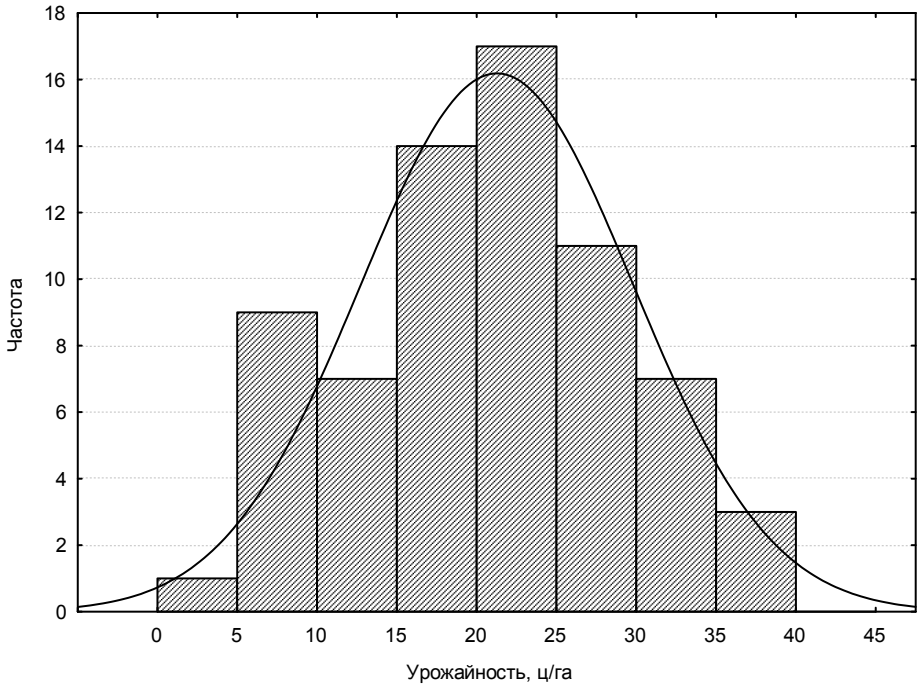


Рисунок 7 – Наблюдаемое (гистограмма) и ожидаемое (кривая линия) распределения признака «урожайность озимой пшеницы»

Таблица 10 – Проверка наблюдаемого распределения на нормальность по критерию χ^2

Границы классов	Менее 10	10-	15-	20-	25-	30 и более
Эмпирические частоты f	9	8	14	17	11	10
Теоретические частоты f'	6,6	9,7	14,6	15,8	12,2	10,1
Разность $d = f - f'$	2,4	–1,7	–0,6	1,2	–1,2	–0,1
d^2	5,76	2,89	0,36	1,44	1,44	0,01
$\frac{d^2}{f'}$	0,87	0,30	0,02	0,09	0,12	0,00

Фактическое значение χ^2 = 1,40. Число степеней свободы ν равно количеству классов минус единица ($\nu = k - 1$), т. е. $\nu = 6 - 1 = 5$. При пяти степенях свободы стандартное (табличное) значение «хи-квадрат» $\chi^2_{st} = \{11{,}1 - 15{,}1 - 20{,}5\}$. Три табличных значения соответствуют трем уровням значимости (p = 0,05; 0,01; 0,001). Поскольку фактическое значение χ^2 намного меньше стандартных, различие ожидаемого и наблюдаемого распределений можно считать случайным (принимается нулевая гипотеза). Распределение наблюдаемых значений урожайности можно аппроксимировать нормальным распределением.

Критерий χ^2 может использоваться и для оценки значимости различий двух и большего числа эмпирических совокупностей, но этот вопрос здесь не рассматривается.

Параметрические критерии

t-критерий Стьюдента

Этот критерий обычно применяется при сравнении средних арифметических двух выборок:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{s^2_{\overline{x_1}} + s^2_{\overline{x_2}}}} = \frac{d}{s_d},$$

где $\overline{x_1}$ и $\overline{x_2}$ – средние значения сравниваемых выборок;

$s_{\overline{x_1}}$ и $s_{\overline{x_2}}$ – выборочные ошибки соответствующих средних значений;

d – разность средних;

s_d – ошибка разности средних.

Если $n_1 \neq n_2$, используется другая формула для вычисления t. Разность средних d можно считать новым параметром, полученным на основе параметров $\overline{x_1}$ и $\overline{x_2}$. Фактическое t является отношением параметра d к его ошибке s_d.

При малых выборках t-критерий Стьюдента можно использовать тогда, когда дисперсии сравниваемых выборок по F-критерию значимо не отличаются.

Стандартные значения *t*-критерия Стьюдента для трех уровней надежности приведены в приложении Г. Число степеней свободы для $\overline{x_1} - \overline{x_2}$ $\nu = n_1 + n_2 - 2$. Если $t_{st} \leq t$, различия средних считаются значимыми, $t < t_{st}$ – незначимыми.

Пример. Сравниваются средние арифметические двух выборок:

1) Высота гречихи $n_1 = 3$, $\overline{x_1} = 180$, $s_1 = 1$, $s_{\overline{x_1}} = \frac{1}{\sqrt{3}} = 0{,}58.$

2) Высота арабидопсиса $n_2 = 3$, $\overline{x_2} = 2$, $s_2 = 1$, $s_{\overline{x2}} = \frac{1}{\sqrt{3}} = 0{,}58.$

Отсюда $d = 180 - 2 = 178$, $s_d = \sqrt{0{,}58^2 + 0{,}58^2} = \sqrt{0{,}34 + 0{,}34} = \sqrt{0{,}68} \approx 0{,}82.$

$t = 178 / 0{,}82 \approx 217$, $t_{st} = \{2{,}8 - 4{,}6 - 8{,}6\}$ (при $\nu = 4$).

$t_{st} < t$; параметр *d* максимально значим ($p < 0{,}001$), разность средних максимально значима.

Статистически доказано, что гречиха выше арабидопсиса.

Интересно, что в этом примере в высшей степени значимая разность получена в ситуации, когда $n_1 = n_2 = 3$ (очень малые выборки). Причина в очень большой разности средних *d*.

По таблице стандартных значений при $\nu \geq 27$ для первого уровня надежности ($p = 0{,}05$) $t_{st} = 2$ (приложение Г). Поскольку при больших выборках $\nu > 27$, есть возможность использовать так называемый упрощенный *t*-критерий.

Если $t \geq 2$, т. е. параметр в два или больше раз превышает ошибку, параметр считается значимым.

Генетики предложили много генетико-селекционных параметров, один из которых – эпистатическое отклонение *i*. Показано, что его можно вычислить по формуле $i = \overline{AABB} - \overline{aaBB} - \overline{AAbb} + \overline{aabb}$, где *AABB*, *aaBB*, *AAbb* и *aabb* – все четыре возможные при дигенном наследовании гомозиготы. Ошибка алгебраической суммы, каковой является *i*, $s_i = \sqrt{s^2_{AABB} + s^2_{aaBB} + s^2_{AAbb} + s^2_{aabb}}$. Упрощенный *t*-критерий $t = i / s_i$. Если $t \geq 2$, эпистатическое отклонение считается установленным, т. е. параметр *i* значимым (Соколова Е. И., 2003).

Напомним, что *t*-критерий можно использовать, если распределение параметра близко к нормальному. Кривая нормального распределения – симметричная кривая. Условием симметричности является равная возможность варьирования статистики в обе стороны – как в сторону уменьшения числовых значений, т. е. влево, так и в сторону их увеличения, т. е. вправо. Если распределение ненормальное, используются иные критерии. В частности, ненормально распределены выборочные доли в интервале до 0,25 и больше 0,75 (рисунок 8). Распределение малых долей всегда имеет положительную асимметрию вследствие того, что в левую сторону возможности варьирования таких долей меньше (до 0,0), чем в правую (до 1,0). Распределение больших долей, напротив, всегда с отрицательной асимметрией, поскольку возможности варьирования влево, в сторону нуля, больше, чем вправо.

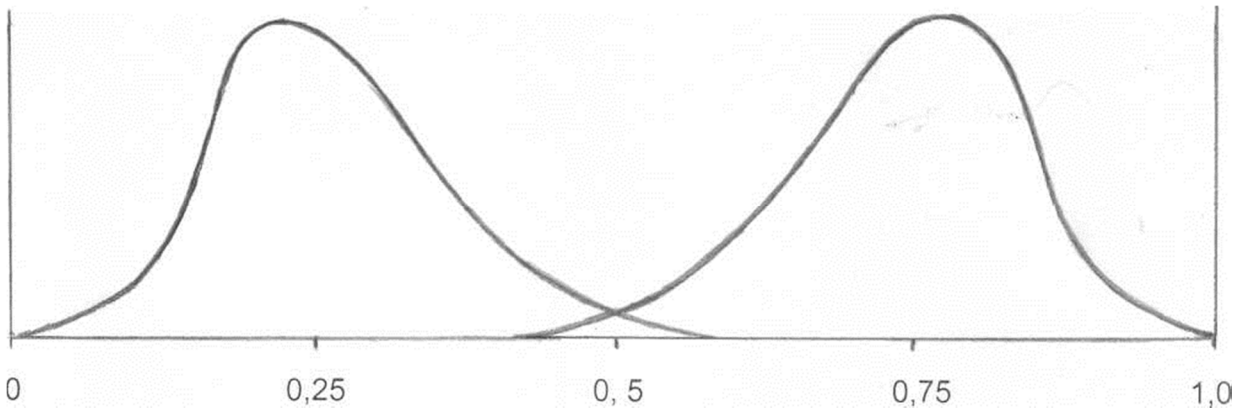


Рисунок 8 – Асимметричные распределения малых (< 0,25) и больших (> 0,75) значений эмпирических долей

При определении значимости разности малых ($f < 0{,}25$) или больших ($f > 0{,}75$) долей классическими способами могут получиться ошибочные результаты вследствие резкой асимметрии распределений с малыми или большими долями. Эти ошибки значительно снижаются, если вместо каждой доли взять угол φ, синус которого равен корню квадратному из этой доли (Плохинский Н. А., 1970). Важно, что ошибка репрезентативности угла φ не зависит от величины этого угла и определение значимости раз-

ности долей по соответствующим углам φ дает более правильные результаты. При сравнении долей в пределах $0{,}25 < f > 0{,}75$ метод φ Фишера дает практически такие же результаты, что и *t*-критерий. Использование метода можно пояснить следующим примером.

Пример. Исследовали частоту встречаемости хромосомных перестроек на разных этапах развития модельного растения арабидопсис. В раннем возрасте из 5000 исследованных клеток обнаружено 3 с перестройками хромосом. В более позднем возрасте из 500 клеток с перестройками тоже было 3. Значим ли прогноз об увеличении частоты перестроек с увеличением возраста?

Понятно, что $f_1 = 0{,}0006$, $f_2 = 0{,}006$. По методу φ Фишера $F_d = 5{,}1$ при $F_{st} = \{3{,}8 – 6{,}6 – 10{,}8\}$. Разность долей значима по первому порогу ($p < 0{,}05$). В данном случае $t = 1{,}5$ (< 2,0) показал незначимость разности долей, что следует считать ошибочным выводом для данного исследования.

F-критерий Фишера

Значение *F*-критерия Фишера при сравнении дисперсий определяют по формуле:

$$F = s_1^2 / s_2^2 \text{ (при } s_1^2 \geq s_2^2\text{)};\ \nu_1 = n_1 - 1,\ \nu_2 = n_2 - 1.$$

Если $F \geq F_{st}$, различие дисперсий считается значимым, т. е. абсолютная изменчивость в сравниваемых выборках принимается неодинаковой, если $F < F_{st}$ – незначимым. *F*-критерий используется как основной при дисперсионном анализе, множественном корреляционном анализе и др. Стандартные значения *F*-критерия Фишера приведены в приложении Д.

Предложены и многие другие критерии, описанные в специальной литературе. Однако они выходят за рамки настоящего издания.

Вопросы для обсуждения

1. Нулевая и альтернативная ей гипотезы.
2. Вероятность ошибки вывода (суждения, заключения) *p*.

3. Три уровня или порога значимости (по величине p).

4. Области применения разных порогов значимости при проведении исследований.

5. Объясните использование критерия знаков z.

6. Объясните использование T-критерия Уилкоксона.

7. Рассмотрите применение критерия χ^2.

8. Рассмотрите применение t-критерия Стьюдента при сравнении средних арифметических значений двух выборок.

9. Упрощенный t-критерий.

10. Рассмотрите применение F-критерия Фишера.

6 ПАРНЫЙ ЛИНЕЙНЫЙ И НЕЛИНЕЙНЫЙ КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Для описания связей между варьирующими признаками (переменными) применяют математическое понятие функции *f*, которая ставит в соответствие каждому определенному значению независимой переменной *X*, называемой аргументом, вполне определенное значение зависимой переменной *Y*: $y = f(x)$. Такого рода однозначные зависимости *Y* от *X* называются функциональными зависимостями или связями.

Однако чаще встречаются корреляционные связи или просто корреляции, когда определенному значению независимой переменной *X* соответствует не одно и то же числовое значение *Y*, а распределяющийся вариационный ряд числовых значений зависимой переменной, хотя связь и записывается в том же виде, что и функциональная $y = f(x)$.

Задачи корреляционного анализа сводятся к установлению направления (положительная или отрицательная), формы связи (линейная, точнее прямолинейная) или нелинейная (криволинейная), измерению ее тесноты (величины) и оценке значимости связи. Графически различие функциональной и корреляционной связи видно из рисунка 9.

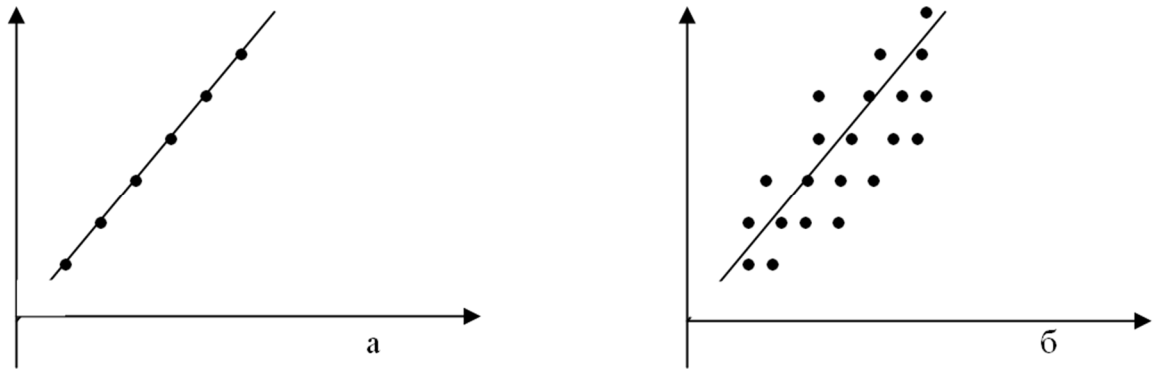


Рисунок 9 – Функциональная (а) и корреляционная (б) положительная линейная связь

Не все корреляционные связи можно называть корреляционными зависимостями. Например, если измерить на каком-то участке леса высоту и диаметр дубов, между этими переменными будет обнаружена корреляционная связь. Но диаметр не зависит от высоты, как и высота от диаметра. Эти признаки корреляционно связаны потому, что они оба увеличиваются при росте деревьев. Связь есть – зависимости нет.

Корреляционные связи не обязательно свидетельствуют о причинно-следственных связях.

Коэффициент корреляции r

Показателем парной линейной связи является двумерная статистика коэффициент корреляции *r:*

$$r_{xy} = \frac{\frac{1}{n}\left[\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})\right]}{s_x s_y}.$$

Коэффициент корреляции – отвлеченное число, лежащее в пределах от –1 до +1. При независимом варьировании признаков, когда связь между ними полностью отсутствует, $r = 0$; при функциональной связи $r = 1$ или $r = -1$.

Знак перед *r* указывает на направление корреляции. Например, мы получили $r = -0{,}9$. Это значит, что связь отрицательная.

Проще всего оценивать значимость коэффициента корреляции, пользуясь уже составленными таблицами критических значений коэффициента корреляции r_{xy}. Такая таблица имеется и у Г. Ф. Лакина (1990), правда, табулированная лишь для двух уровней значимости ($p = 0{,}05$ и $p = 0{,}01$). Допустим, для нахождения *r* использовали 7 пар значений признака. В этом случае число степеней свободы $v = n - 2 = = 7 - 2 = 5$; по таблице (приложение Е) значимым считается $r \geq 0{,}75$ ($p = 0{,}05$) и $r \geq 0{,}87$ ($p = 0{,}01$). При вычислении на ПК обычно получают оценки не только значений *r*, но и его значимости.

Коэффициент корреляции оценивает линейную связь, точнее линейный компонент связи. Довольно условно выделяют слабую

($r < 0{,}5$), среднюю ($0{,}5 \le r < 0{,}7$) и сильную $r \ge 0{,}7$ связь (Лакин Г. Ф., 1990).

Корреляционное отношение h

Для измерения нелинейной связи между X и Y используют корреляционное отношение. Оно всегда является величиной положительной, может иметь значения от 0 (отсутствие связи) до 1 (функциональная связь). Корреляционное отношение позволяет характеризовать любую форму корреляции – и линейную, и нелинейную. Если $h = r$ – связь линейная. Чем более различаются r и h, тем в большей степени нелинейной является связь. На сравнении r и h основывается вычисление критерия нелинейности.

Коэффициент детерминации

Коэффициент детерминации показывает, какая доля вариации одного признака зависит от варьирования другого.

При линейной связи коэффициент детерминации представляет собой r_{xy}^2, при нелинейной связи он равен h_{yx}^2.

В качестве примера парного корреляционного анализа приводим результаты изучения зависимости урожайности озимой пшеницы в Луганской области (y) от 48 переменных (x_1–x_{48}).

За независимые переменные принимали среднемесячные температуры года сбора урожая (x_{13}–x_{24}) и предшествующего ему года (x_1–x_{12}), а также месячные суммы осадков за эти же периоды (x_{25}–x_{48}). Заметим, что предшествующий году сбору урожая год для озимых культур является годом их посева.

В системе STATISTICA создаем базу данных, а потом находим всю совокупность парных коэффициентов корреляции. Особый интерес для нас имеют коэффициенты корреляции урожайности (y) с 48 независимыми переменными (x_1–x_{48}). Они приведены в таблице 11 с указанием уровней значимости.

Корреляция урожайности с независимыми переменными слабая. Положительные коэффициенты корреляции урожайности и температуры воздуха установлены по январю, февралю и марту года сбора урожая. Это значит, что чем теплее зима и начало весны (во время перезимовки озимых культур), тем выше урожайность

озимой пшеницы. Напротив, корреляции урожайности и температуры воздуха за май и летние месяцы года сбора урожая были отрицательными. Чем прохладнее май и последующее лето, тем выше урожайность озимой пшеницы, и наоборот, чем жарче май и лето, тем она ниже. Результат ожидаемый, поскольку зима в Луганской области для озимой пшеницы иногда слишком морозная, а лето – излишне жаркое.

В этой области, расположенной в засушливой зоне, озимая пшеница на втором году жизни обычно ощущает недостаток влаги, и потому почти во всех случаях положительно реагирует на увеличение осадков. Особенно большие положительные корреляции урожайности с осадками установлены по апрелю, маю, июню и июлю года сбора урожая (таблица 11).

Таблица 11 – Коэффициенты парной корреляции (r_1–r_{48}) урожайности озимой пшеницы (y) с температурой воздуха и осадками (x_1–x_{48})

Месяц	Незав. перем.	r_1–r_{48}	Месяц	Незав. переем.	r_1–r_{48}
1	2	*3*	4	5	6
Среднемесячная температура воздуха, град. С			Месячная сумма осадков, мм		
Год посева					
Январь	x_1	0,01	Январь	x_{25}	0,17
Февраль	x_2	0,02	Февраль	x_{26}	0,09
Март	x_3	0,21	Март	x_{27}	–0,11
Апрель	x_4	0,13	Апрель	x_{28}	0,18
Май	x_5	–0,03	Май	x_{29}	0,09
Июнь	x_6	–0,06	Июнь	x_{30}	0,28*
Июль	x_7	–0,04	Июль	x_{31}	0,22
Август	x_8	0,08	Август	x_{32}	–0,08
Сентябрь	x_9	–0,10	Сентябрь	x_{33}	0,31**
Октябрь	x_{10}	0,04	Октябрь	x_{34}	–0,15
Ноябрь	x_{11}	0,09	Ноябрь	x_{35}	0,13
Декабрь	x_{12}	0,18	Декабрь	x_{36}	0,07

Продолжение таблицы 11

1	2	3	4	5	6
Год сбора урожая					
Январь	x_{13}	0,25*	Январь	x_{37}	0,12
Февраль	x_{14}	0,29*	Февраль	x_{38}	0,12
Март	x_{15}	0,38**	Март	x_{39}	0,11
Апрель	x_{16}	0,13	Апрель	x_{40}	0,28*
Май	x_{17}	–0,21	Май	x_{41}	0,31**
Июнь	x_{18}	–0,30**	Июнь	x_{42}	0,28*
Июль	x_{19}	–0,10	Июль	x_{43}	0,25*
Август	x_{20}	–0,21	Август	x_{44}	–0,12
Сентябрь	x_{21}	–0,09	Сентябрь	x_{45}	0,29*
Октябрь	x_{22}	0,24*	Октябрь	x_{46}	–0,22
Ноябрь	x_{23}	0,17	Ноябрь	x_{47}	–0,06
Декабрь	x_{24}	–0,03	Декабрь	x_{48}	0,27*

* – Параметр значим при $0,05 < p < 0,01$.

** – При $0,01 < p < 0,001$.

Эти месяцы охватывают практически весь вегетационный период озимой пшеницы на втором году жизни (в год сбора урожая). При этом высокий коэффициент корреляции найден по паре признаков: урожайность – сумма осадков за июнь ($r = 0,28$). Почти такая же корреляция по абсолютному значению ($r = – 0,30$), но иная по знаку была обнаружена при рассмотрении пар признаков: урожайность – средняя температура июня (см. таблицу 11). При этом июньская температура и осадки связаны значимой отрицательной корреляцией ($r = –0,41$), т. е. чем больше осадков в июне, тем этот месяц прохладнее. Аналогичная связь, хотя и менее тесная, характерна также для июля, августа и сентября. Коэффициент детерминации июньской температурой изменчивости урожайности озимой пшеницы $r^2 = 0,09$ (9 %).

На основе этих исследований можно сделать заключение: для получения высоких урожаев озимой пшеницы в данном регионе определяющее значение имеет июнь года сбора урожая. Чем больше осадков в июне и чем он прохладнее, тем выше урожайность, и наоборот, чем меньше июньских осадков и чем жарче июнь, тем

урожайность ниже. В меньшей степени это касается и мая года сбора урожая.

На первый взгляд кажется странным, что урожайность озимой пшеницы значимо коррелирует с декабрьской суммой осадков года сбора урожая (см. таблицу 11): осадки выпадают уже после сбора урожая, и поэтому не могут прямо влиять на урожайность. Однако наличие такой связи можно понять, рассмотрев корреляцию декабрьских осадков с другими переменными. Сумма осадков декабря положительно связана с осадками апреля ($r = 0{,}14$) и июня ($r = 0{,}16$), оказывающими положительное влияние на урожайность (таблица 11). Таким образом, корреляционная связь не обязательно должна быть причинно-следственной.

В общем, на урожайность озимой пшеницы в целом влияет комплекс природных экологических факторов – среднемесячные температуры и месячные суммы осадков (48 переменных). Степень влияния на урожайность этих переменных неодинаковая, и они в той или иной степени коррелируют друг с другом.

Поскольку количество влияющих на урожайность факторов или независимых переменных велико, трудно точно оценить силу (степень) влияния отдельных факторов, индивидуализировать их влияние, нередко и доказать значимость отдельных влияний. Однако в подобной ситуации в этом и нет особой необходимости. Важнее оценить влияние всех факторов вместе, для чего и предназначен множественный корреляционно-регрессионный анализ, рассматриваемый в следующей лекции.

Корреляция между качественными и количественными признаками, а также между двумя качественными признаками

Предложено много разных методов и показателей, в том числе:

1) коэффициент ассоциации, или тетрахорический показатель связи (качественные признаки, группируемые в четырехпольную корреляционную таблицу).

Пример – оценка сцепления генов.

2) коэффициент взаимной сопряженности, или полихорический показатель связи (Пирсона – Чупрова) (качественные признаки, группируемые в многопольную корреляционную таблицу).

Имеется и возможность оценить корреляционную связь между качественным и количественным признаками.

Вопросы для обсуждения

1. Чем различаются понятия «функциональная связь (зависимость)» и «корреляционная связь». Связи линейные (прямолинейные) и нелинейные (криволинейные).

2. Понятия независимой переменной (аргумент) и зависимой переменной (функция).

3. Корреляционные связи и корреляционные зависимости. Причинно-следственные связи (зависимости).

4. Двумерная статистика коэффициент парной корреляции r.

5. Границы варьирования значений r. Установление направления и силы связи (слабая, средняя или тесная) по значениям r.

6. Определение значимости коэффициента парной корреляции r.

7. Корреляционное отношение h.

8. Коэффициенты детерминации (r^2, h^2).

9. Область применения тетрахорического показателя связи (коэффициента ассоциации).

10. Область применения полихорического показателя связи (коэффициента взаимного сопряжения).

7 ЧАСТНАЯ И МНОЖЕСТВЕННАЯ КОРРЕЛЯЦИЯ

Частная корреляция

Зная парные коэффициенты корреляции r_{xy}, r_{xz} и r_{yz}, можно определить так называемые частные, или парциальные, коэффициенты корреляции, показывающие корреляционную зависимость между двумя варьирующими признаками при постоянной величине третьего (т. е. при его исключенном влиянии).

Для определения частного коэффициента корреляции между признаками *X* и *Y* при постоянной величине признака Z применяют формулу:

$$r_{xy(z)} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}.$$

Заключение знака *z* в скобки означает, что влияние признака *Z* на корреляцию между *X* и *Y* исключено.

Аналогичные формулы получены для $r_{xz(y)}$ и $r_{yz(x)}$.

Пример

Установлены следующие коэффициенты парной корреляции:

1) урожайность (*Y*) – температура в июне (*X*) $r_{xy} = -0{,}30$;

2) урожайность (*Y*) – сумма осадков за июнь (*Z*) $r_{yz} = 0{,}28$;

3) температура в июне (*X*) – сумма осадков за июнь (*Z*) $r_{xz} = -0{,}41$.

Тогда

$$r_{xy(z)} = \frac{-0{,}30 - (-0{,}41 \cdot 0{,}28)}{\sqrt{(1 - (-0{,}41)^2)(1 - 0{,}28^2)}} = -0{,}24.$$

Сравнение частного коэффициента корреляции $r_{xy(z)} = -0{,}24$ с парным $r_{xy} = -0{,}30$ позволяет утверждать, что при одинаковом уровне осадков в июне связь урожайность – температура июня меньше, чем при изменчивом.

Частные корреляции используются редко.

Множественная корреляция

Простейшим случаем множественной корреляции является корреляция трех признаков: *X*, *Y* и *Z*. Тесноту связи одного из них Y (зависимая переменная) от двух других *X* и *Z* (независимых переменных) характеризует **коэффициент множественной корреляции**:

$$r_{y(xz)} = \frac{\sqrt{r_{xy}^2 + r_{yz}^2 - 2r_{xy} \cdot r_{xz} \cdot r_{yz}}}{1 - r_{xz}^2},$$

где r_{xy}, r_{xz} и r_{yz} – коэффициенты линейной корреляции между парами признаков *X* и *Y*, *X* и *Z*, *Y* и *Z*.

В последнее время чаще зависимую переменную обозначают буквой y, тогда как независимые – x_1, x_2 и т. д.

Уравнение для вычисления коэффициента множественной линейной корреляции (*R*) легко обобщается на любое число переменных.

Коэффициент множественной корреляции принимает значения от 0 до 1 ($0 \le R \le 1$).

Значимость этого совокупного коэффициента корреляции оценивают разными способами, в том числе и по *F*-критерию (Доспехов Б. А., 1985):

$$F = \frac{R^2}{1 - R^2}\left(\frac{n-k}{k-1}\right),$$

где *n* – объем выборки;

k – число признаков (число всех, зависимой и независимых переменных);

F_{st} определяют по таблице для $\nu_1 = k - 1$ и $\nu_2 = n - k$;

Нулевая гипотеза принимается, если $F < F_{st}$;

R – указывает, насколько сильно зависит одна (зависимая) переменная от всех других, учтенных при подсчете *R* независимых переменных.

Пример вычисления множественной корреляции тот же, что и при вычислении парных корреляций.

Анализ множественной корреляции позволяет оценить суммарное влияние 48 учитывавшихся независимых переменных (температуры воздуха и сумм осадков за каждый из 24 мес) на урожайность пшеницы.

Коэффициент множественной линейной корреляции вычисляли в пакете программ STATISTICA. Он очень высок (R = 0,920) и значим (p = 0,023). Можно говорить об очень тесной связи урожайности с 48 независимыми переменными.

Коэффициент множественной линейной детерминации, являющийся квадратом коэффициента множественной линейной корреляции R^2, тоже очень велик (R^2 = 0,8462).

Иначе говоря, почти 85 % (точнее 84,62 %) изменчивости по годам урожайности озимой пшеницы в Луганской области определяется погодными условиями, а именно температурой и осадками. Это главные лимитирующие факторы, детерминирующие урожайность озимой пшеницы.

Все другие, не учтенные в данном исследовании факторы, вместе детерминируют 100 – 84,62 = 15,38 % изменчивости урожайности озимой пшеницы по годам. Среди этих факторов такие природные экологические факторы, как скорость ветра, снеговой покров, минимальные температуры зимой и максимальные летом, ледяная корка, град, запасы влаги в почве и другие. Отсутствие снегового покрова и ледяная корка увеличивают вероятность повреждения, а то и гибели посевов озимой пшеницы при перезимовке, и поэтому могут снижать ее урожайность. Сильный ветер и град приводят подчас к полеганию пшеницы и недобору урожая. Однако добавление в математическую модель этих факторов не может значительно увеличить R и R^2, которые и так уже очень велики (R и R^2 в принципе не могут быть равны или больше 1). Неучтенные нами природные экологические факторы в той или иной мере коррелируют с организованными в настоящем исследовании независимыми переменными, и поэтому отчасти уже «учтены». Значи-

тельное увеличение R и R^2 при добавлении независимых переменных произойти не может, но небольшое увеличение ожидается. По изложенным соображениям наши оценки степени влияния независимых переменных (температуры и осадков) на урожайность озимой пшеницы можно принимать за минимальные оценки степени или силы влияния всех погодных факторов.

В эти же 15,38 % укладывается влияние ряда антропогенных факторов (фондо- и энергообеспеченность, объемы использования удобрений и пестицидов, сорта, изменения в организации производства, структуре посевных площадей, площадей под орошением и др.).

До этого исследования считалось, что все факторы внешней среды (погодные флуктуации, «капризы природы») обусловливают лишь 60–80 % изменчивости урожайности. В действительности для Луганской области это число больше.

Вопросы для обсуждения

1. Понятие частного (парциального) коэффициента корреляции.
2. Способ определения частного коэффициента корреляции между признаками X и Y ($r_{xy(z)}$) при постоянном значении признака Z.
3. Рассмотрите пример вычисления $r_{xy(z)}$.
4. Коэффициент множественной линейной корреляции R при трех переменных (Y – зависимая переменная, X и Z – независимые).
5. Возможно ли обобщение уравнения для вычисления коэффициента множественной линейной корреляции R для любого числа переменных?
6. Границы варьирования R.
7. Определение значимости R по F-критерию Фишера.
8. Коэффициент множественной линейной детерминации R^2.
9. Рассмотрите пример вычисления R, R^2, F-критерия и вероятности ошибки p.
10. Рассмотрите толкование смысла полученных значений R, R^2, F-критерия и p.

8 ПАРНЫЙ ЛИНЕЙНЫЙ И НЕЛИНЕЙНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

Любую форму связи можно выразить уравнением общего вида $y = f(x)$, где y – зависимая переменная (функция), а x – независимая переменная (аргумент).

Изменение зависимой переменной под действием одной или нескольких независимых называется регрессией. Известны эмпирические методы выравнивания (например, методом экспоненциального сглаживания, скользящей средней и др.) и аналитические (линейное сглаживание и др.).

Линейная регрессия

При одной независимой переменной и корреляции уравнение линейной регрессии имеет вид прямой линии:

$$y = a_0 + a_1 x \text{ (рисунок 10)},$$

где y – ожидаемое значение функции;

a_0 – свободный член в уравнении регрессии;

a_1 – угловой коэффициент, или коэффициент регрессии;

x – значение независимой переменной.

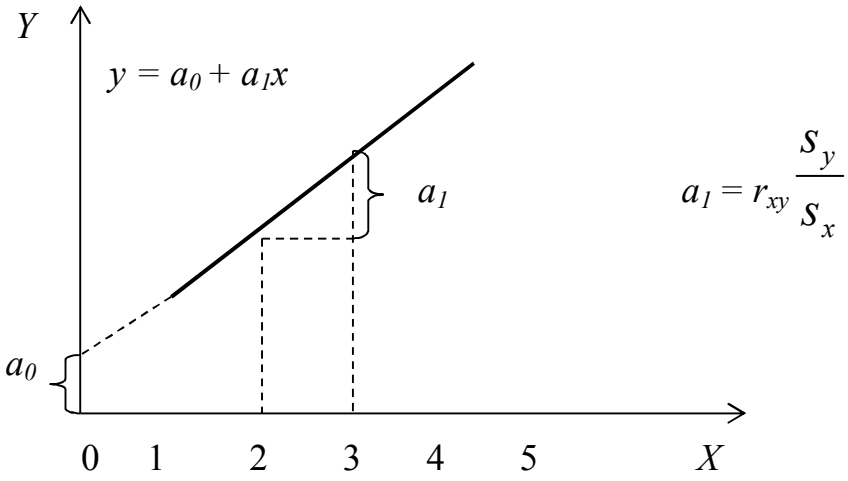


Рисунок 10 – График прямолинейной регрессии

Нахождение уравнения линейной регрессии – это вычисление a_0 и a_1 на ПК. Оба этих параметра при корреляционной связи оце-

ниваются с определенными статистическими ошибками, позволяющими установить значимость параметров a_0 и a_1.

Пример

Изучали количество сеянцев сосны обыкновенной, появившихся при самосеве в старом сосновом лесонасаждении. Исходные данные представлены в таблице 12. Необходимо оценить связь переменных.

Таблица 12 – Количество сеянцев сосны обыкновенной

Расстояние от соснового лесонасаждения, м	0-	20-	40-	60-	80-	100-	120-	140-	Всего сеянцев
Количество сеянцев, шт./10 м^2	42	42	38	21	13	14	1	0	171

Здесь расстояние – независимая переменная (аргумент) *X*, а количество сеянцев – зависимая (функция) *Y*.

При вычислении на ПК получили значение $r = –0,97***$ и $y = 48,95 – 0,3446x$.

Связь *Y* и *X* можно изобразить графически (рисунок 11).

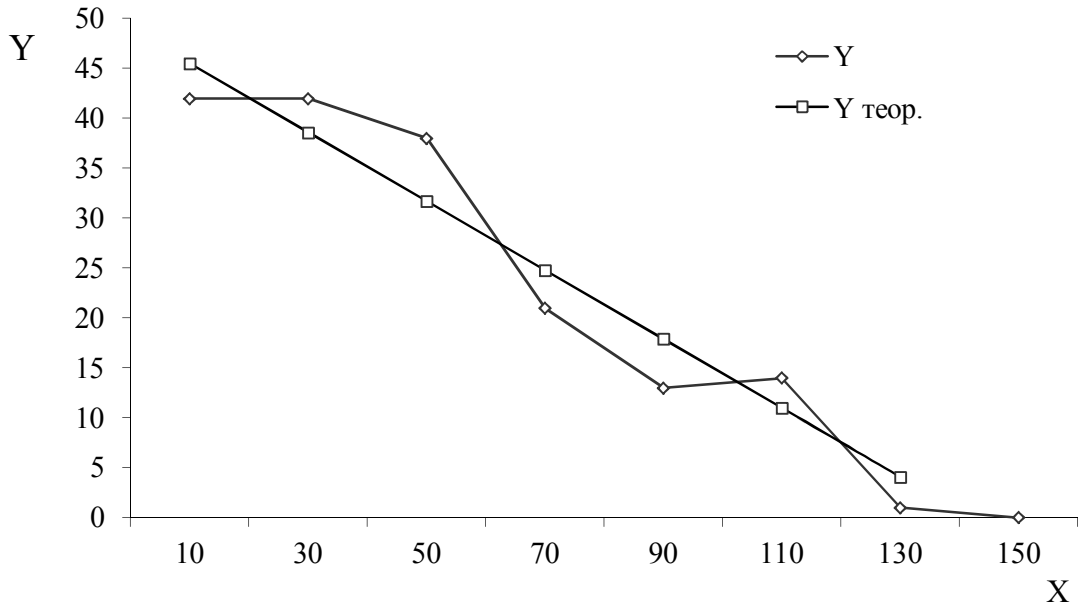


Рисунок 11 – Изменение густоты самосева сосны в зависимости от расстояния до соснового леса

Очевидно, что количество сеянцев зависит от расстояния до леса: чем дальше от него, тем сеянцев меньше. Зависимость их количества от расстояния до леса максимально значима ($p < 0{,}001$); ее можно считать прямолинейной.

Нелинейная регрессия

Полиномы второго и более высоких порядков

Уравнение прямой линии – частный случай полинома, это полином первой степени.

Полиномы больших степеней иногда лучше соответствуют исходным данным и потому тоже используются.

Полином второй степени – это квадратичная парабола

$$y = a_0 + a_1x + a_2x^2.$$

В этом случае необходимо найти a_0, a_1 и a_2 на ПК.

Прямая линия – частный случай квадратичной параболы, когда $a_2 = 0$.

В биологии, сельском и лесном хозяйствах квадратичная парабола обычно неплохо описывает связи: густота посева – урожайность, глубина заделки семян – урожайность и др. Во всех подобных случаях производственники должны найти оптимум фактора, обеспечивающего максимальную урожайность. Это так называемый поиск экстремума (рисунок 12).

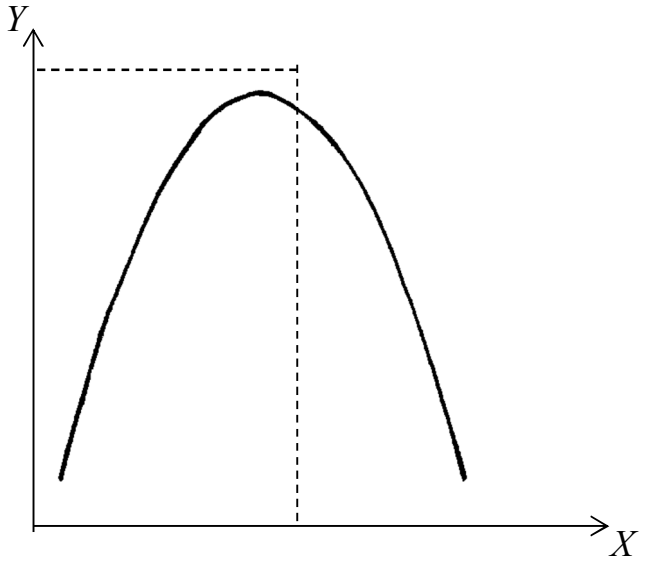


Рисунок 12 – Квадратичная парабола

Парабола (полином) третьего порядка описывается уравнением

$$y = a_0 + a_1x + a_2x^2 + a_3x^3.$$

Она используется реже квадратичной параболы, а полиномы еще больших степеней в наших областях знаний практически не применяются.

Гиперболы

Для аналитического сглаживания эмпирических рядов служат также гиперболы разных порядков с различным числом неизвестных.

Наиболее простая из них – гипербола первого порядка (рисунок 13):

$$y = a_0 + \frac{a_1}{x}$$

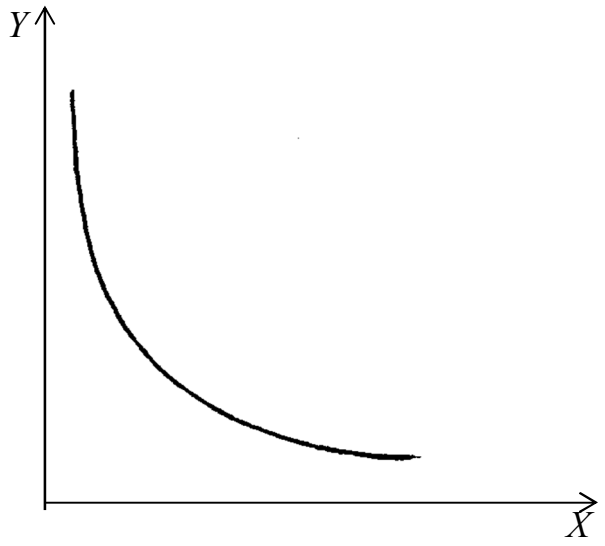


Рисунок 13 – Гипербола первого порядка

В ряде случаев именно так изменяется загрязнение почвы продуктами выхлопных газов автомобилей при уменьшении расстояния от автотрассы – источника загрязнения.

Регрессия, выражаемая уравнением логистической кривой

Логистическая кривая, называемая иногда *S*-образной, сигмоидной или кривой Сакса, изображена на рисунке 14.

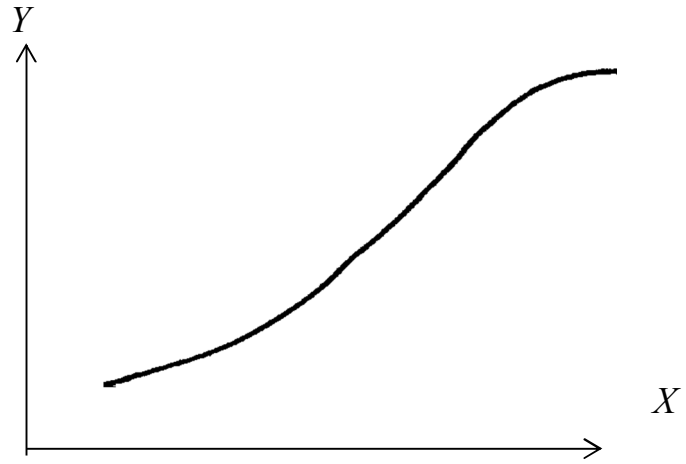


Рисунок 14 – Логистическая кривая

Кривая описывается уравнением Ферхлюста. Такие кривые хорошо аппроксимируют изменения во времени (X – время) высоты растений (в см, числе деревьев), их массы, числа особей популяции в замкнутом пространстве и т. п. Эти кривые называют поэтому также кривыми роста.

Периодическая регрессия

Ее описывают обычно тригонометрическим уравнением регрессии. Называют ее также циклической. Продолжительность периода (цикла) на рисунке 15 обозначена буквой t.

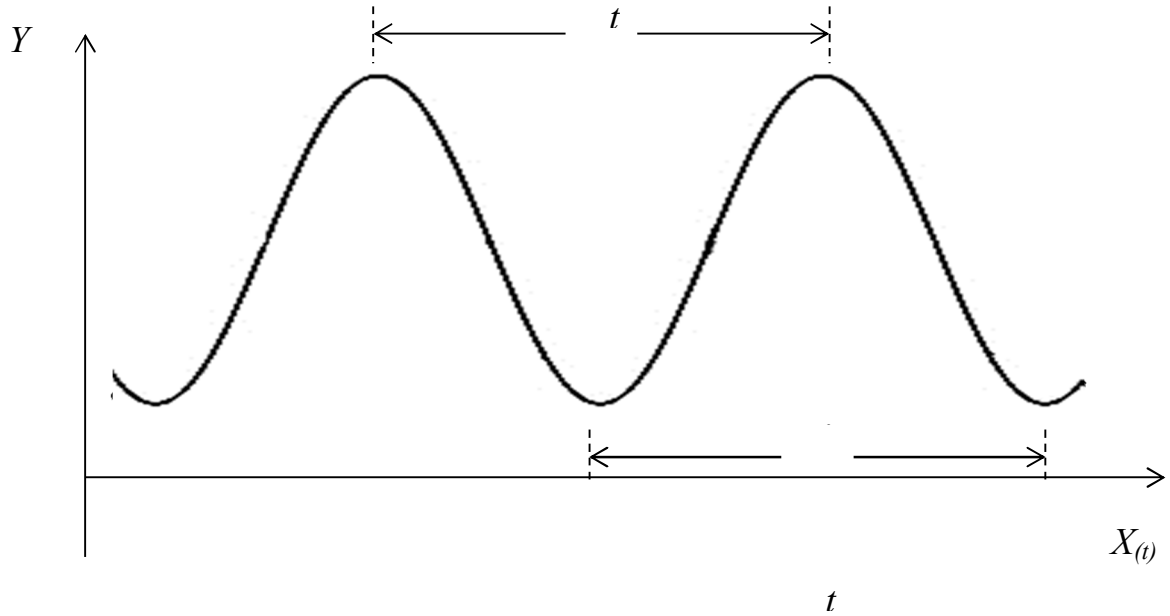


Рисунок 15 – Периодическая кривая

В естественных популяциях периодически изменяется численность зайцев, рысей и ряда других животных. Периодически изме-

няется урожайность озимой пшеницы и ряда других сельскохозяйственных культур. Обнаружен вековой цикл в динамике температуры и осадков в Луганской области.

В подобных случаях, когда независимой переменной является время (здесь – годы), а зависимой – какой-либо признак, двойной ряд чисел называется ***рядом динамики,*** или ***временным рядом***.

Анализ временных рядов производится с использованием парного корреляционного и регрессионного анализов так же, как это делают по любым парам других признаков.

В нашей программе REGAN предусмотрено аналитическое сглаживание следующими шестнадцатью функциями:

1) линейная;
2) периодическая;
3) гипербола вида

$$y = a_0 + \frac{a_1}{x}$$

4) степенная;
5) показательная;
6) экспоненциальная;
7) логарифмическая;
8) квадратичная парабола;
9) логистическая;
10) экспоненциальная вида $y = a_0 \cdot exp\ (a_1 \cdot x^2)$;
11) экспоненциальная вида $y = a_0 \cdot exp\ (a_1 \cdot x^4)$;
12) гипербола второго порядка;
13) гипербола третьего порядка;
14) гипербола первого порядка с тремя неизвестными;
15) гипербола второго порядка с тремя неизвестными;
16) гипербола третьего порядка с тремя неизвестными.

Выбор уравнений регрессии

Выбираться должно уравнение, которое лучше других описывает связь Y с X. Если преимущество более сложной функции над

более простой мало, лучше ограничиться более простой. Наконец, надо брать функции, которые не противоречат здравому смыслу.

Выбор линейного или нелинейного уравнения

При строго линейной зависимости коэффициенты корреляционного отношения h_{yx}^2 и h_{xy}^2 равны между собой и равны r_{xy}^2, т. е. $\nu = h^2 - r^2 = 0$. При нелинейной связи $\nu \neq 0$.

Если по F-критерию Фишера разность ν значима, лучше использовать нелинейное аналитическое сглаживание ($F = \frac{\nu}{1-h^2} \cdot \frac{N-\alpha}{a-2}$).

Выбор вида нелинейного сглаживания (уравнение)

Нередко возможные уравнения подбирают на основе сравнения эмпирического графика с известными образцами кривых.

Обычно проверяют на ПК несколько возможных функций, для которых оценивается и так называемая ***среднеквадратичная ошибка аппроксимации***. Предпочтение отдается той функции, которая обнаруживает явно меньшую ошибку аппроксимации.

Выбор функции – наука, но в некоторой степени и искусство. Возможно наложение одних функций на другие. Если коэффициент корреляции велик, имеет высокую значимость, и доказана периодическая функция, говорят о значимой линейной компоненте изменчивости, на которую наложена циклическая изменчивость.

Пример. Анализ временного ряда

Исследуется уже приводившийся ряд данных: урожайность озимой пшеницы в Луганской области по годам с 1945 г. по 2013 г., за 69 лет (рисунок 16).

Корреляция годы – урожайность максимально значима (r = = 0,602***, r^2 = 0,362). Линейная компонента изменчивости очевидна (см. рисунок 16). Уравнение линейной регрессии y = 12,21 + + 0,255*x. Угловой коэффициент 0,255, т. е. на данном 69-летнем временном интервале происходил рост урожайности в среднем на ~ 0,255 ц/га (на ~ 1 ц/га за 4 года). Среднеквадратичная погрешность аппроксимации прямой линией E = 45,43. Визуально кажется, что в последние 10–15 лет урожайность перестала расти и даже

несколько снизилась. Это требует проверки посредством сглаживания криволинейными функциями.

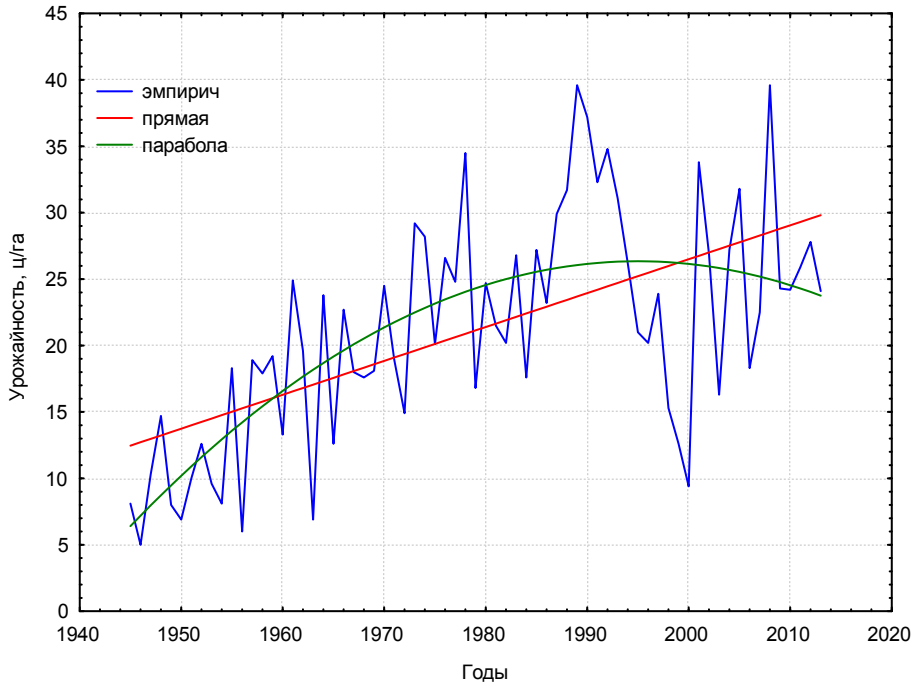


Рисунок 16 – Сглаживание временного ряда прямой линией и квадратичной параболой

Парабола

Уравнение квадратичной параболы $y = 5{,}59 + 0{,}814 \cdot x - 0{,}00798 \cdot x^2$. Видно, что аналитическое сглаживание параболой лучше, чем прямой линией (см. рисунок 16). Об этом же свидетельствует и ошибка аппроксимации параболой $E = 37{,}42$, в то время как такая ошибка при сглаживании прямой $E = 45{,}43$. Корреляция наблюдавшихся и ожидавшихся по параболе значений урожайности максимально значима ($r = 0{,}689$***). Кстати, в нашем примере парабола лучше, чем прямая линия, согласуется с изменением на исследованном временном интервале среднего уровня урожайности, почему ей и отдано предпочтение.

Проверяли также ***периодическую функцию***. Получено тригонометрическое уравнение регрессии, описывающее циклические колебания урожайности с периодом в 16 лет. Оно лучше, чем одна парабола, описывает динамику урожайности (рисунок 17). Амплитуда колебаний ≈ 6 ц/га. Ошибка аппроксимации в этом случае еще

меньше (E = 31,00). Корреляция наблюдавшихся и ожидавшихся по этой функции значений урожайности максимально значима (r = 0,7543***).

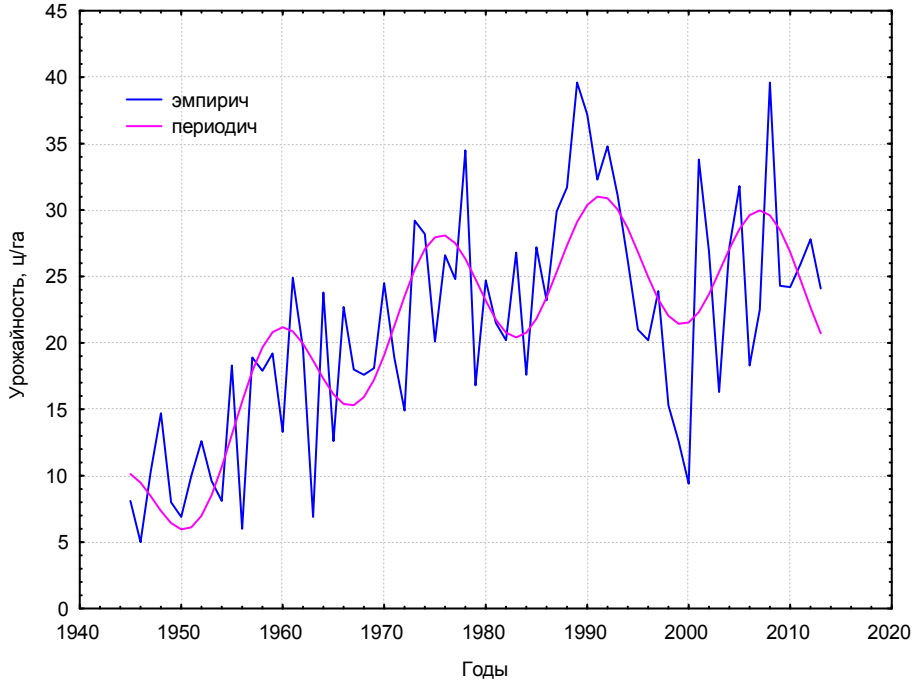


Рисунок 17 – Сглаживание временного ряда периодической функцией, наложенной на параболу

Можно использовать еще один способ оценки того, сколь хорошо согласуются фактические значения урожайности и полученные по уравнениям линий регрессии – вычисление коэффициента корреляции пар значений: наблюдавшиеся – ожидавшиеся.

Получены результаты: прямая линия r = 0,602***, парабола r = 0,689***; 16-летние циклы, наложенные на параболу r = 0,7543***. Чем выше r, тем лучше совпадение. Очевидно, что лучшие результаты дал третий способ аналитического сглаживания.

Итак, сглаживание прямой линией наиболее грубое (E = 45,43), сглаживание параболой значительно лучше (E = 37,42), еще лучшие результаты дает криволинейная функция, полученная при наложении на параболу периодической компоненты изменчивости (E = 31,00). Анализ временного ряда позволил создать более адекватную математическую модель динамики урожайности (тренд –

близкая к параболе кривая, на которую наложена циклическая изменчивость с периодом 16 лет), имеющую прогностическую ценность.

Выводы

1. В исследованный период времени средний ход урожайности такой: вначале рост – потом понижение (парабола).

2. На средний ход наложены периодические колебания урожайности (период 16 лет).

Вопросы для обсуждения

1. Что такое регрессия?

2. Эмпирические и аналитические методы выравнивания (сглаживания).

3. Линейная регрессия и ее графическое изображение.

4. Полиномы второго (квадратичная парабола) и более высоких степеней.

5. Гиперболы.

6. Логистическая кривая.

7. Периодическая регрессия.

8. Понятие рядов динамики, или временных рядов.

9. Определите, в каком случае следует использовать линейное аналитическое сглаживание и в каком – нелинейное.

10. Какому нелинейному сглаживанию (уравнению) следует отдать преимущество в том или другом случае?

9 МНОЖЕСТВЕННЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

Зависимость между несколькими переменными величинами принято выражать уравнением множественной регрессии, которая может быть линейной и нелинейной. Анализ множественных нелинейных связей сложен, он описан в специальной литературе. В простейшем виде множественная линейная регрессия (одна зависимая и две независимые переменные) выражается следующим уравнением:

$$y = a_0 + a_1x_1 + a_2x_2,$$

где a_0 – свободный член уравнения;

a_1 – коэффициент регрессии по первой независимой переменной;

a_2 – коэффициент регрессии по второй независимой переменной;

x_1, x_2 – значения первой и второй переменной соответственно.

По исходным данным на ПК находим значения a_0, a_1, a_2 и составляем уравнение множественной регрессии.

Если парная линейная регрессия графически изображается прямой линией, то такая регрессия – плоскостью (там сглаживание прямой – здесь сглаживание плоскостью).

Большие вычислительные трудности, возникавшие до создания PC (ПК) IBM, отталкивали исследователей от использования множественного регрессионного анализа, когда число независимых переменных больше двух. Тем более не использовался множественный нелинейный регрессионный анализ. Не случайно у Г. Ф. Лакина (1990) множественной регрессии посвящено всего лишь около двух страниц текста.

Современные ПК, оснащенные пакетами программ типа STATISTICA, позволяют вовлекать в анализ очень большое количество переменных и получать в результате анализа принципиально важные новые выводы по тем проблемам, которые давно волнуют ученых и общественность.

В общем виде уравнение множественной регрессии имеет следующий вид:

$$y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n,$$

где n – количество независимых переменных.

Очень высокую степень совпадения, или согласия, фактических (эмпирических, наблюдаемых) и теоретических (ожидаемых по уравнению множественной регрессии) значений урожайности иллюстрирует рисунок 18, полученный при работе с пакетом программ STATISTICA.

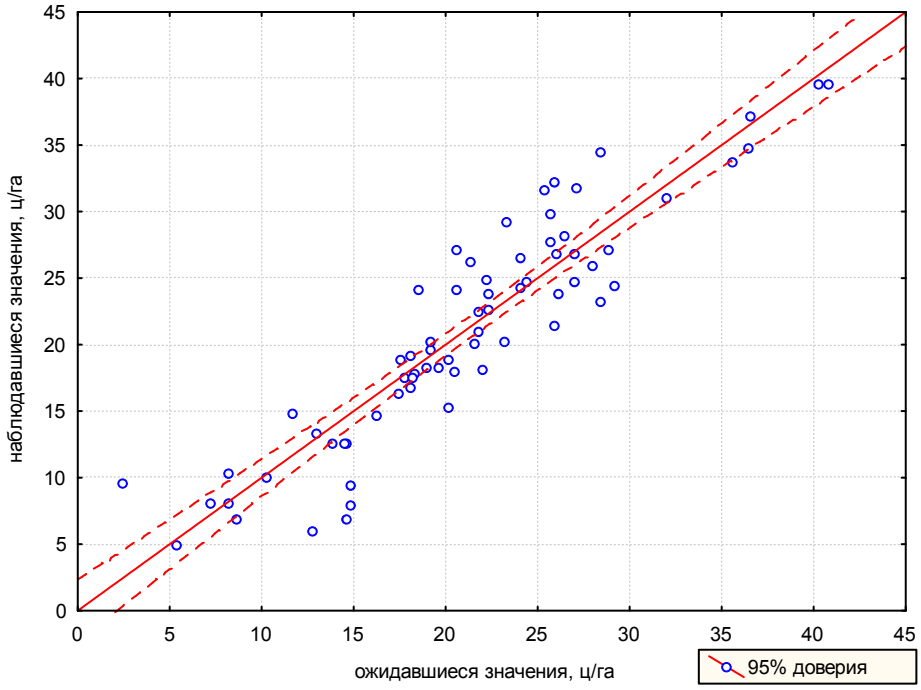


Рисунок 18 – График, иллюстрирующий степень совпадения ожидавшихся и наблюдавшихся величин урожайности озимой пшеницы

В пакете программ STATISTICA вычисляются также отклонения (разности) наблюдаемых значений урожайности от ожидаемых. Анализ этих отклонений позволяет утверждать, что они в среднем по модулю составляют ≈ 2,5 ц/га (около 55 % укладывается в 2 ц/га, 70 % – в 3 ц/га). Наибольшее отклонение фактического значения от теоретического в отрицательном направлении было в 1963 г. (–7,68 ц/га), в положительном – в 1953 г. (+7,20 ц/га).

Ряды динамики, или временные ряды ожидаемых и наблюдаемых значений урожайности, представленные на рисунке 19, очень сходны даже в деталях. Хорошее совпадение, или согласованность, наблюдается, в частности, как для особенно неурожайных, так и для рекордно урожайных лет. Особенно низкой в сравнении с ближайшими предшествующими и последующими годами фактическая урожайность была в 1946, 1950 и 1953 гг. (5,0; 6,9 и 9,6 ц/га

соответственно); по значениям изученных факторов (температура и осадки), являющихся независимыми переменными x_1-x_{48}, в эти годы ожидался неурожай (5,4; 8,5 и 2,4 ц/га соответственно). Наибольшей в сравнении с ближайшими предшествующими и последующими годами фактическая урожайность была в 1989, 1990 и 2001 гг. (39,6; 37,2 и 33,8 ц/га соответственно); в эти годы ожидалась необычайно высокая урожайность в связи с благоприятной температурой и осадками (40,2; 36,4 и 35,5 ц/га соответственно).

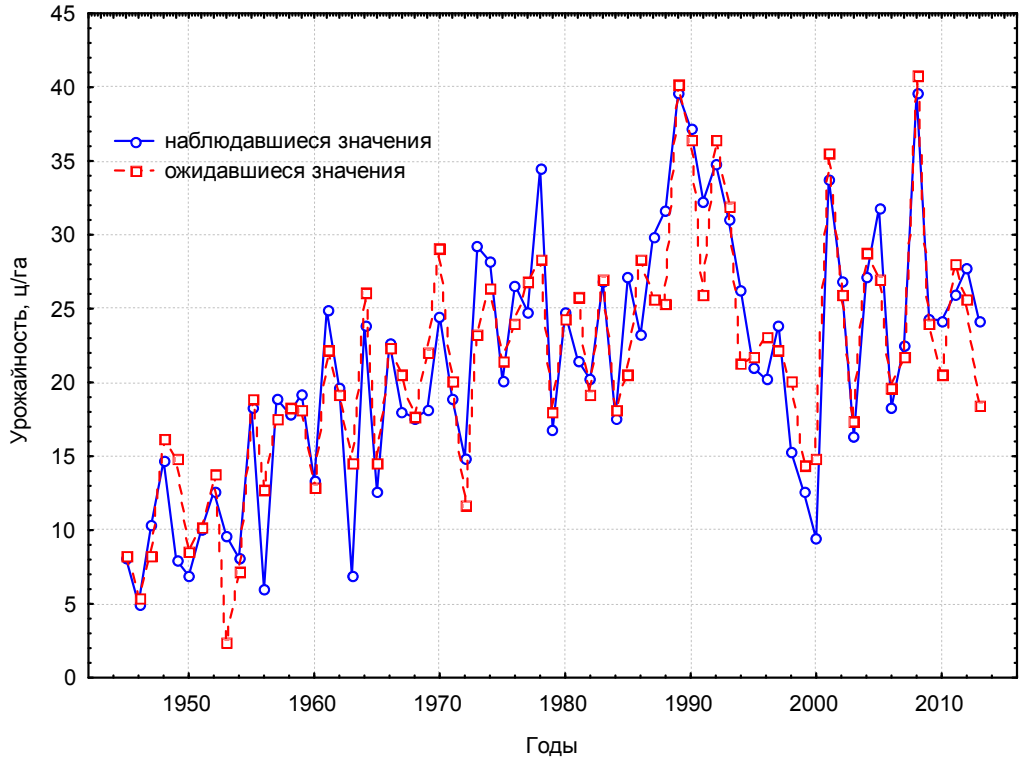


Рисунок 19 – Временные ряды ожидавшихся и наблюдавшихся величин урожайности озимой пшеницы

При таком совпадении фактической и теоретической урожайности коэффициент множественной линейной корреляции R, являющийся мерой зависимости урожайности от изученных экологических факторов, независимых переменных, должен быть высоким. И действительно, R = 0,920 (p = 0,023). Можно говорить об очень тесной связи урожайности с независимыми переменными.

Обычно считается, что уменьшение урожайности сельскохозяйственных культур в Луганской области в последнее десятилетие XX в. связано с экономическими и политическими причинами. Но

дело в том, что на временном интервале с 1991 по 2004 г. совпадение фактической урожайности озимой пшеницы с ожидавшейся по уравнению множественной регрессии, учитывающей лишь погодные условия (температуру и осадки), даже несколько лучше, чем в 1945–1990 гг. Степень согласия наблюдавшихся и ожидавшихся значений урожайности столь велика, что варьирование урожайности пшеницы по годам в 1991–2004 гг. следует связывать прежде всего с динамикой погодных условий.

При большом количестве независимых переменных (предикторов), в разной степени влияющих на зависимую переменную, для создания более удобной модели с меньшим количеством переменных рекомендуется использовать процедуру множественного регрессионного анализа с их последовательным пошаговым исключением. Целью пошагового анализа является построение модели с небольшим количеством предикторов, но, тем не менее, с высокими величинами R и R^2, значимыми по принятому p-уровню. Логика процедур пошагового анализа отражена в таблице 13.

Таблица 13 – Величины R, R^2, p и значимость коэффициентов регрессии в уравнениях множественной регрессии при разном наборе переменных

Шаг (step)	R	R^2	p	Предикторы			
				var1	var2	var3	var4
0	0,946	0,898	0,01	*ns*	*	*ns*	*
1	0,930	0,865	0,005	*	*	–	*
* – Значимый коэффициент регрессии; *ns* – незначимый коэффициент регрессии.							

Исходная модель (шаг 0) обеспечивает очень высокие и значимые показатели R и R^2. Однако коэффициенты регрессии уравнения множественной регрессии с переменными var1 и var3 незначимы. Если исключить из анализа var3, то все коэффициенты регрессии с оставшимися независимыми переменными оказываются значимыми (шаг 1). Однако величины R и R^2 меньше, чем в исходной модели. К сожалению, подобное уменьшение бывает всегда, когда строится модель с уменьшенным числом переменных. В данном примере уменьшение R и R^2 на шаге 1 в сравнении с шагом 0 не-

значительное, поэтому модель с тремя предикторами можно считать более адекватной, чем исходную.

Множественный корреляционно-регрессионный анализ на ПК позволяет перейти от догадок и неточных прогнозов о степени влияния на важные для агрономов и лесоводов признаки комплекса природных и антропогенных факторов к их числовой оценке с указанием степени значимости выводов. Такая возможность появилась лишь в последние годы (современные ПК и мощное программное обеспечение типа пакета программ STATICTICA).

Вопросы для обсуждения

1. Запишите уравнение множественной линейной регрессии в наиболее простом виде, при трех переменных (одна зависимая и две независимые).

2. Запишите уравнение множественной линейной регрессии в общем виде, при *n* независимых переменных.

3. Запишите уравнение множественной линейной регрессии в ситуации, когда количество независимых переменных равно 48.

4. Разберитесь с приведенным примером уравнения множественной регрессии.

5. Ознакомьтесь с графиком, который иллюстрирует связь ожидаемых и наблюдаемых по годам значений урожайности озимой пшеницы в Луганской области. Что имеют в виду в выражении «95 % достоверности»?

6. Ознакомьтесь с графиком, на котором представлены ожидаемые и наблюдаемые величины урожайности озимой пшеницы в Луганской области.

7. Как вы визуально оцениваете соответствие ожидаемых и наблюдаемых значений урожайности: плохое, удовлетворительное, хорошее или отличное?

8. Анализ отклонений (разница) наблюдаемых значений от ожидаемых.

9. Причины ограниченного использования множественного линейного регрессионного анализа и его перспективы.

10. Множественный нелинейный регрессионный анализ.

10 ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

Дисперсионный анализ (дословно: анализ дисперсий) основан на разложении общей изменчивости статистического комплекса на составляющие его компоненты, сравнивая которые, можно определить ***значимость влияния отдельных факторов***, а также ***долю общей вариации*** учитываемого признака, обусловленную действием на него организованных в опыте и неорганизованных факторов.

Чаще всего на практике встречаются однофакторные и двухфакторные комплексы, равномерные и неравномерные, сопряженные и несопряженные.

Дисперсионный анализ – наиболее часто применяемый в исследовательской работе в области сельского и лесного хозяйства метод, но возможности его используются обычно не в полной мере по той причине, что многие исследователи ограничиваются лишь сравнением средних по наименьшей существенной разности (НСР). В связи с этим подчеркнем, что главное в дисперсионном анализе – анализ дисперсий (о чем говорит его название).

Примером использования однофакторного дисперсионного анализа может служить сортоиспытание. Объективное сортоиспытание предполагает такую организацию опыта, при которой фенотипические различия между сортами можно связать с их генотипами. В частности, закладка опытного участка плодовых или лесных культур должна производиться в один и тот же год одновозрастными саженцами, полученными путем прививки на один и тот же стандартный для данной культуры и зоны подвой в одних и тех же условиях. Соблюдение принципа единственности различий – обязательное условие проведения сортоиспытания. По этой причине при расположении деревьев на опытном участке должен применяться принцип рендомизации.

На рисунке 20 представлено возможное размещение деревьев трех сравниваемых сортов при схеме посадки 9 × 6 м, когда места

посадки определялись по жребию (по закону случайных чисел), т. е. при полной рендомизации. Случайное расположение позволяет считать, что все сорта будут выращиваться при одинаковых внешних условиях (снимается эффект положения). Использование принципа рендомизации обеспечивает возможность применения дисперсионного метода анализа.

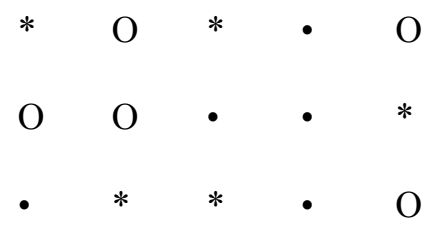


Рисунок 20 – Случайное размещение деревьев трех сравниваемых сортов (*, O , •) на опытном участке

В данном случае мы имеем дело с однофакторным несопряженным комплексом. Поскольку объемы градации фактора (число деревьев каждого сорта) равны ($n_1 = n_2 = n_3$), т. е. комплекс равномерный. Это частный случай неравномерного, когда отдельные или все градации фактора не равны по объему (например, $n_1 \neq n_2 = n_3$). Обычно равномерные комплексы обрабатывают на ПК по тем же программам, что и неравномерные. Приводимый ниже пример является примером анализа неравномерного комплекса: $n_1 = 4$, $n_2 = 6$, $n_3 = 3$, $n_4 = 2$).

Дисперсионный анализ однофакторных несопряженных неравномерных и равномерных комплексов

Форма представления исходных данных приведена в таблице 14.

Таблица 14 – Форма представления исходных данных дисперсионного анализа

Вариант	Исходные данные	Число наблюдений
A1	$x_{11}, x_{12}, x_{13}, \ldots, x_{1n}$	n_1
A2	$x_{21}, x_{22}, \ldots, x_{2n}$	n_2
.		.
.		.
.		.
A_a	$x_{a1}, x_{a2}, x_{a3}, \ldots, x_{an}$	n_a

Общая сумма квадратов отклонений значений признака (от x_n до x_{an}) от средней арифметической по комплексу $(\bar{x})$ D_y подразделяется лишь на две части – факториальную (межгрупповую) D_a и случайную (остаточную) D_z:

$$D_y = D_a + D_z.$$

Факториальная сумма квадратов обусловлена различиями средних по вариантам. Деля суммы квадратов на число степеней свободы, получают значения соответствующих дисперсий (варианс): s_y^2, s_A^2 и s_z^2 (таблица 15).

Таблица 15 – Форма представления результатов дисперсионного анализа

Источник изменчивости	Сумма квадратов	Степени свободы	Дисперсия (варианса)	Критерий F
Общая	D_y	$N - 1$	s_y^2	–
Факториальная	D_A	$a - 1$	s_A^2	s_A^2 / s_z^2
Остаточная	D_z	$N - a$	s_z^2	–

Достоверность (значимость) влияния фактора оценивается с использованием F-критерия (см. таблицу 2). Если $s_A^2 / s_z^2 \geq F_{st}$, влияние фактора считается значимым по тому или иному уровню значимости. Если $F < F_{st}$, влияние фактора считается незначимым, принимается нулевая гипотеза.

Если влияние фактора значимое, вычисляются НСР для трех уровней значимости – $НСР_{05}$, $НСР_{01}$ и $НСР_{001}$. Фактические разности между средними тех или иных вариантов сравнивают с НСР. Если $d \geq$ НСР, разность считается значимой.

Дискуссионным вопросом является метод оценки силы влияния факторов при дисперсионном анализе.

Наиболее простой и популярный метод оценки предложен Н. А. Плохинским: $h^2 = D_A / D_y$.

Используется и оценка по Снедекору: $h_A^2 = \frac{s_A^2 - s_z^2}{s_A^2 + (n-1)s_z^2}$.

Для неравномерных комплексов n вычисляется по формуле

$$n = \frac{1}{a-1}\left(N - \frac{\sum (n_i)^2}{N}\right).$$

Пример (Плохинский Н. А., 1970). Исходные значения приведены в таблице 16.

Таблица 16 – Исходные значения

Вариант	Исходные данные	Средние значения	Число наблюдений
А 1	8,0 8,4 9,0 8,6	8,5	4
А 2	8,2 9,0 10,0 10,0 9,2 10,0	9,4	6
А 3	11,0 13,0 12,0	12,0	3
А 4	7,5 8,5	8,0	2

Результаты дисперсионного анализа при $N = 15$ представлены в таблице 17.

Таблица 17 – Результаты дисперсионного анализа

Источник изменчивости	Сумма квадратов	Степени свободы	Дисперсия	Критерий F
Общая	33,05	14	2,36	–
Факториальная	27,31	3	9,10	17,45***
Остаточная	5,74	11	0,52	–

$F_{st} = \{3,6 – 6,2 – 11,6\}$, $F_{st} < F$ ($p < 0,001$).

Выводы

1. Различия в опыте между вариантами в целом максимально значимы.

2. Показатели силы влияния: по Снедекору – 0,82,
по Плохинскому – 0,83.

Более 80 % изменчивости в дисперсионном комплексе определяется различиями между вариантами. По НСР можно найти, какие же разности между средними значениями вариантов значимы.

Например, $\bar{x}_{A1} = 8,5$, $\bar{x}_{A4} = 8,0$; d = 0,5. При 5%-м уровне значимости $НСР_{05} = 1,38$. d < НСР. Средние значения А 1 и А 4 значимо не различаются, но в опыте имеются различающиеся варианты.

Так же поступают при сравнении других средних.

Предостережения

1. Не следует вообще сравнивать средние вариантов, находить НСР, если по *F*-критерию в опыте нет различающихся вариантов.

2. Не следует брать большого количества вариантов, когда их сравнивают по $НСР_{05}$. В этом случае из 20 сравнений в одном (в 5 % случаев) ожидается ошибка первого рода (будет отвергнута 0-гипотеза, хотя она верна).

3. Конкретный пример – конкурсное испытание. Если проверяемых образцов 40 и все они сравниваются со стандартом, то в ~ 2 случаях мы отвергнем 0-гипотезу, т. е. выделим 2 кандидатов на сорт, и это может оказаться ошибкой.

Дисперсионный анализ сопряженных равномерных комплексов

Если в полевом опыте использовалось случайное размещение опытных делянок в пределах повторности (блока), т. е. производилась рендомизация в блоках, то из общей изменчивости можно выделить изменчивость по повторностям, определяющую их различия.

Организация опытного участка при такой рендомизации представлена на рисунке 21.

A3	**A1**	**A2**	**A4**	**A2**	**A4**	**A1**	**A3**	**A1**	**A4**	**A2**	**A3**
I				II				III			

Рисунок 21 – Случайное размещение делянок в блоках (повторностях)

Здесь в каждой из трех повторностей имеется весь набор градаций изучаемого фактора (A1–A4), размещаемых в пределах повторности случайно, не систематически. Так располагаются делянки при изучении урожайности в опытах с полевыми и овощными культурами. Так же могут располагаться и «делянки», фактически представляющие собой ряды деревьев, в опытах с плодовыми и лесными культурами.

В этом случае общая сумма квадратов составит $D_y = D_A + D_p + D_z$, при этом D_p определяется различиями между повторностями.

При полной рендомизации в несопряженном комплексе $D_y = D_A + D_z$, в отличие от него, такой комплекс называют сопряженным.

Результаты дисперсионного анализа представляют по следующей форме (таблица 18).

Таблица 18 – Форма представления результатов дисперсионного анализа

Источник изменчивости	Сумма квадратов	Степени свободы	Дисперсия (варианса)	Критерий F
Общая	D_y	$N - 1$	s_y^2	–
Повторений	D_p	$n - 1$	s_p^2	–
Факториальная	D_A	$a - 1$	s_A^2	s_A^2 / s_z^2
Остаточная	D_z	$N - a$	s_z^2	–

Контрольный пример (Плохинский Н. А., 1970). Исходные значения приведены в таблице 19.

Таблица 19 – Исходные значения

Вариант	Значения по повторностям			Средние по вариантам
	I	II	III	
А 1	47,8	46,9	45,4	46,70
А 2	53,7	50,3	50,6	51,53
А 3	46,7	42,0	43,4	44,03
А 4	48,0	47,0	45,9	46,97
Средние по повторностям	49,05	46,55	46,32	47,31

Результаты анализа контрольного примера приведены в таблице 20.

Таблица 20 – Результаты дисперсионного анализа

Источник изменчивости	Сумма квадратов	Степени свободы	Дисперсия	Критерий F
Общая	111,06	11	10,10	–
Повторений	18,29	2	9,15	–
Факториальная	87,18	3	29,06	31,20***
Остаточная	5,59	6	0,93	–

$$F_{st} = \{4{,}8 - 9{,}8 - 23{,}7\},\ F_{st} < \mathrm{F}\ (p < 0{,}001).$$

Выводы

1. Различия в опыте максимально значимы, в опыте есть различающиеся варианты.

2. По $НСР_{05} = 1{,}89$ можно установить, какие именно варианты значимо различаются по средним значениям.

Вопросы для обсуждения

1. Задачи дисперсионного анализа статистических комплексов.

2. Объясните отличия одно-, двух- и многофакторных комплексов; равномерных и неравномерных; сопряженных и несопряженных.

3. Соблюдение принципа единственности различий между вариантами (градациями фактора).

4. Объясните форму представления исходных данных для дисперсионного анализа однофакторных несопряженных неравномерных и равномерных комплексов.

5. Рассмотрите форму представления результатов дисперсионного анализа.

6. Объясните оценку значимости влияния фактора с использованием F-критерия.

7. Показатели силы влияния фактора.

8. Сравнение отдельных средних (по градациям фактора) по НСР. Назовите ограничения использования НСР (предостережения).

9. Рассмотрите дисперсионный анализ сопряженных равномерных комплексов.

10. Почему в случае сопряженных комплексов общая сумма квадратов отклонений делится на компоненты по формуле $D_y = D_A + D_p + D_z$, а не по формуле $D_y = D_A + D_z$ (как в случае несопряженных комплексов)?

11 ДВУХФАКТОРНЫЙ И МНОГОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

Дисперсионный анализ двухфакторных несопряженных неравномерных и равномерных комплексов

Здесь $D_y = D_x + D_z$, где D_x – факториальная сумма квадратов. В свою очередь $D_x = D_A + D_B + D_{AB}$, где D_A – сумма квадратов по фактору А, D_B – сумма квадратов по фактору В, D_{AB} – сумма квадратов, обусловленная взаимодействием факторов А и В.

Схема представления исходных данных представлена в таблице 21.

Таблица 21 – Схема представления исходных данных дисперсионного анализа

Варианты		Исходные данные	Число наблюдений
A_1	B_1	$x_{11}, x_{11}, \ldots, x_{11}$	n_{11}
	.	...	.
	.		.
	.		.
	B_6	$x_{16}, x_{16}, \ldots, x_{16}$	n_{16}
.	.	...	.
	.		.
	.		.
A_a	B_1	$x_{a1}, x_{a1}, \ldots, x_{a1}$	n_{a1}
	.	...	.
	.		.
	.		.
	B_6	$x_{ab}, x_{ab}, \ldots, x_{ab}$	n_{ab}

Результаты дисперсионного анализа имеют следующий вид (таблица 22).

Таблица 22 – Результаты дисперсионного анализа

Источник изменчивости	Сумма квадратов	Степени свободы	Дисперсия (варианса)	Критерий F
1	2	3	4	5
Общая	D_y	$N - 1$	s_y^2	–
Факториальная	D_A	$ab - 1$	s_x^2	s_A^2 / s_z^2

Продолжение таблицы 22

1	2	3	4	5
Остаточная	D_z	$N - ab$	s_z^2	–
Фактор А	D_A	$a - 1$	s_A^2	s_A^2 / s_z^2
Фактор В	D_B	$b - 1$	s_B^2	s_B^2 / s_z^2
Взаимодействие факторов А и В	D_{AB}	$(a-1)(b-1)$	s_{AB}^2	s_{AB}^2 / s_z^2

Контрольный пример (Плохинский Н. А., 1970). Исходные значения приведены в таблице 23.

Таблица 23 – Исходные данные

Градации факторов		Исходные значения	Число наблюдений
А	В		
1	1	2 3 4	3
	2	3 3 3 4	4
	3	4 5 6 7 8	5
2	1	2 3 3 4	4
	2	3 7 8 8 9 9	6
	3	4 5 5 6	4

Результаты дисперсионного анализа контрольного примера (таблица 24).

Таблица 24 – Результаты дисперсионного анализа контрольного примера

Источник изменчивости	Сумма квадратов	Степени свободы	Дисперсия	Критерий F
Общая	116,00	25	4,64	–
Факториальная	96,45	5	19,29	19,73***
Остаточная	19,55	20	0,98	–
Фактор А	11,45	1	11,45	11,71***
Фактор В	40,30	2	20,15	20,61***
Взаимодействие факторов А и В	44,70	2	22,35	22,87***

Различия в опыте между вариантами максимально значимы как в целом, так и по отдельным факторам. Максимально значимо также взаимодействие факторов А и В.

Взаимодействие факторов связано с тем, что влияние градаций одного из них зависит от того, на фоне какой градации другого фактора оно исследуется. Взаимодействие – это отсутствие аддитивных влияний.

Приведем здесь лишь показатели силы влияния по Плохинскому: $h_x^2 = 0{,}83$, $h_A^2 = 0{,}10$, $h_B^2 = 0{,}35$, $h_{AB}^2 = 0{,}39$.

В целом факториальные влияния обеспечивают 0,83 изменчивости, по фактору А – 0,10, по фактору В – 0,35, взаимодействие факторов А и В – 0,39.

Очевидно, что сила влияния фактора В примерно в 3 раза выше, чем фактора А. Кроме того, имеет место сильное взаимодействие АВ.

Отдельные варианты сравниваются уже рассмотренным ранее способом (по НСР).

Дисперсионный анализ двухфакторных сопряженных равномерных комплексов

Используют такой анализ, если в полевом опыте производилась рендомизация в блоках. Разложение суммы квадратов ведется в соответствии с формулой $D_y = D_a + D_b + D_{ab} + D_p + D_z$.

Результаты дисперсионного анализа представляют в виде таблицы (таблица 25).

Таблица 25 – Результаты дисперсионного анализа

Источник изменчивости	Сумма квадратов	Степени свободы	Дисперсия (варианса)	Критерий F
Общая	D_y	$N - 1$	s_y^2	–
Повторений	D_p	$n - 1$	s_p^2	–
Факториальная	D_A	$ab - 1$	s_x^2	s_A^2 / s_z^2
Остаточная	D_z	$N - ab$	s_z^2	–
Фактор А	D_A	$a - 1$	s_A^2	s_A^2 / s_z^2
Фактор В	D_B	$b - 1$	s_B^2	s_B^2 / s_z^2
Взаимодействие факторов А и В	D_{AB}	$(a-1)(b-1)$	s_{AB}^2	s_{AB}^2 / s_z^2

Контрольный пример (Плохинский Н. А., 1970). Исходные данные представлены в таблице 26.

Таблица 26 – Исходные данные

Градации факторов		Значения по повторностям				Средние по вариантам
A	B	I	II	III	IV	
1	1	19	20	15	15	17,25
	2	20	20	20	18	19,50
	3	20	19	18	19	19,00
2	1	30	31	21	17	24,75
	2	42	35	28	33	34,50
	3	48	51	50	48	49,25
Средние по повторностям		29,83	29,33	25,33	25,00	27,38

Результаты дисперсионного анализа представлены в таблице 27.

Таблица 27 – Результаты дисперсионного анализа

Источник изменчивости	Сумма квадратов	Степени свободы	Дисперсия (варианса)	Критерий *F*
Общая	3357,62	23	145,98	–
Повторений	118,12	3	39,37	–
Факториальная	3083,38	5	616,68	59,25***
Остаточная	156,12	15	10,41	–
Фактор A	1855,04	1	1855,04	178,23***
Фактор B	690,75	2	345,38	33,18***
Взаимодействие факторов A и B	537,58	2	268,79	25,82***

Во всех случаях использования *F*-критерия $p < 0{,}001$, т. е. влияние изучавшихся факторов A и B, совместное и по отдельности максимально значимо. Столь же максимально значимо и взаимодействие AB.

Велика и сила влияния факторов A и B (по Плохинскому: фактора A – 55,2 %, фактора B – 20,6 %, взаимодействия факторов A и B – 16 %). Фактор A в данном случае влияет сильнее на изучавшийся признак, чем фактор B и их взаимодействие AB.

Оценка разностей частных средних (средних по вариантам) проводится обычным образом по НСР.

Трехфакторные и еще более сложные дисперсионные комплексы обычно не встречаются в практике биометрических исследований.

Если число факторов три (A, B, C) и более, то возрастает не только число взаимодействий первого порядка A – B, A – C, B – C, но и появляются взаимодействия более высоких порядков. Ничего такого в двухфакторном анализе не бывает, там есть лишь взаимодействие A – B. Из-за взаимодействий высоких порядков возникает плохо разработанная проблема интерпретации результатов многофакторного дисперсионного анализа.

Вопросы для обсуждения

1. Объясните форму представления исходных данных для дисперсионного анализа двухфакторных несопряженных неравномерных и равномерных комплексов.

2. Как при таком анализе разделяется сумма квадратов отклонений на компоненты?

3. Рассмотрите форму представления результатов дисперсионного анализа, обратив внимание на максимально значимые источники изменчивости (по *F*-критерию).

4. Показатели силы влияния факторов A и B, а также их взаимодействия AB.

5. Сравнение отдельных средних по НСР.

6. Разложение суммы квадратов отклонений на компоненты при дисперсионном анализе двухфакторных сопряженных равномерных комплексов.

7. Объясните форму представления результатов анализа двухфакторных сопряженных комплексов.

8. Рассмотрите контрольный пример анализа сопряженного комплекса, обратив внимание на интерпретацию результатов.

9. Причины ограниченного использования дисперсионного анализа трехфакторных и еще более громоздких комплексов.

10. Взаимодействие в трехфакторных и многофакторных экспериментах и их интерпретация.

12 ДИСКРИМИНАНТНЫЙ, КЛАСТЕРНЫЙ И ФАКТОРНЫЙ АНАЛИЗ

Рассматриваемые здесь методы относятся к многомерному статистическому анализу. Такие методы позволяют выбрать ту вероятностно-статистическую модель, которая соответствует исходным данным, характеризующим реальное поведение исследуемой совокупности объектов, сделать выводы в том числе и на основе ограниченного статистического материала.

Рассмотрим, для каких целей обычно используются эти типы анализа. Пусть есть исходно существующие группы (например, группы или классы расщепления в гибридном поколении), и нужны переменные, которые лучше всего их разделяют. Для этой цели применим дискриминантный анализ. Если есть несколько переменных, и на основе которых можно классифицировать выборку, проверить, не объединяются ли наблюдения в группы, потребуется кластерный анализ. При наличии нескольких переменных, в целях классификации или сокращения их числа следует использовать факторный анализ. В настоящее время в нашей предметной области исследований (биология, сельское хозяйство, лесное хозяйство, экология, медицина) чаще используется кластерный анализ.

Для облегчения выбора адекватного метода в краткой форме указаны основные области применения дискриминантного, кластерного и факторного анализа.

Дискриминантный анализ	*Кластерный анализ*
Есть исходно существующие группы. Поиск переменных, которые лучше всего их разделяют.	Есть несколько переменных. Цель – на их основе классифицировать выборку т. е. проверить, не объединяются ли наблюдения в группы.

Факторный анализ:
Многомерное шкалирование

Используется при нескольких переменных
с целью классификации или уменьшения их числа.

Дискриминантный анализ

Широкий круг задач, возникающих на практике и связанных с классификацией, можно решить методами дискриминантного анализа. Дискриминантный анализ используется: 1) для поиска переменных, позволяющих относить наблюдаемые объекты в одну или несколько реально наблюдаемых групп; 2) для классификации объектов в различные группы. С вычислительной точки зрения дискриминантный анализ похож на дисперсионный. В результате вычислений получают матрицу общих и внутригрупповых дисперсий. Их можно сравнить с помощью многомерного F-критерия для того, чтобы установить, имеются ли значимые различия между группами (с учетом всех переменных).

Дискриминантный анализ как раздел многомерного статистического анализа включает методы классификации многомерных наблюдений в ситуации, когда исследователь располагает так называемыми обучающими совокупностями (классификация с обучением).

Дискриминантный анализ для двух групп, называемый также линейным дискриминатным анализом Фишера, основывается на линейном уравнении следующего типа: Группа = $a + b_1x_1 + b_2x_2 + +\ldots b_mx_m$, где a является константой, а $b_1\ldots b_m$ – коэффициентами регрессии. Если имеется более двух групп, то оценивается более чем одна подобная дискриминантная функция.

Использование дискриминантного анализа возможно при выполнении следующих ограничений: 1) нормальные распределения переменных; 2) однородность дисперсий, проявляющаяся в отсутствии корреляции между групповыми средними и дисперсиями (или стандартными отклонениями); 3) используемые для дискриминации между совокупностями переменные не должны быть полностью избыточными. Если переменная почти полностью избыточна (и поэтому матрица задачи является плохо обусловленной), значение толерантности, равное $1 – R^2$, будет приближаться к нулю.

Частым применением дискриминантного анализа является включение в исследование многих переменных с целью определе-

ния тех из них, которые наилучшим образом разделяют совокупности между собой. Используется пошаговый анализ дискриминантных функций с последовательным включением или исключением переменных, обоснованием чего являются значения *F* для включения и *F* для исключения.

Можно сравнить, как две функции дискриминируют между группами, построив значения, которые принимают обе дискриминантные функции. В качестве примера приводим рисунок, взятый из электронного учебника STATISTICA StatSoft (рисунок 22). В этом примере функция Root 1 в основном дискриминирует между группой Setosa с объединением групп Virginic и Versicolor. По функции Root 2 (вертикаль) заметно некоторое смещение точек группы Versicolor вниз относительно центральной линии (0).

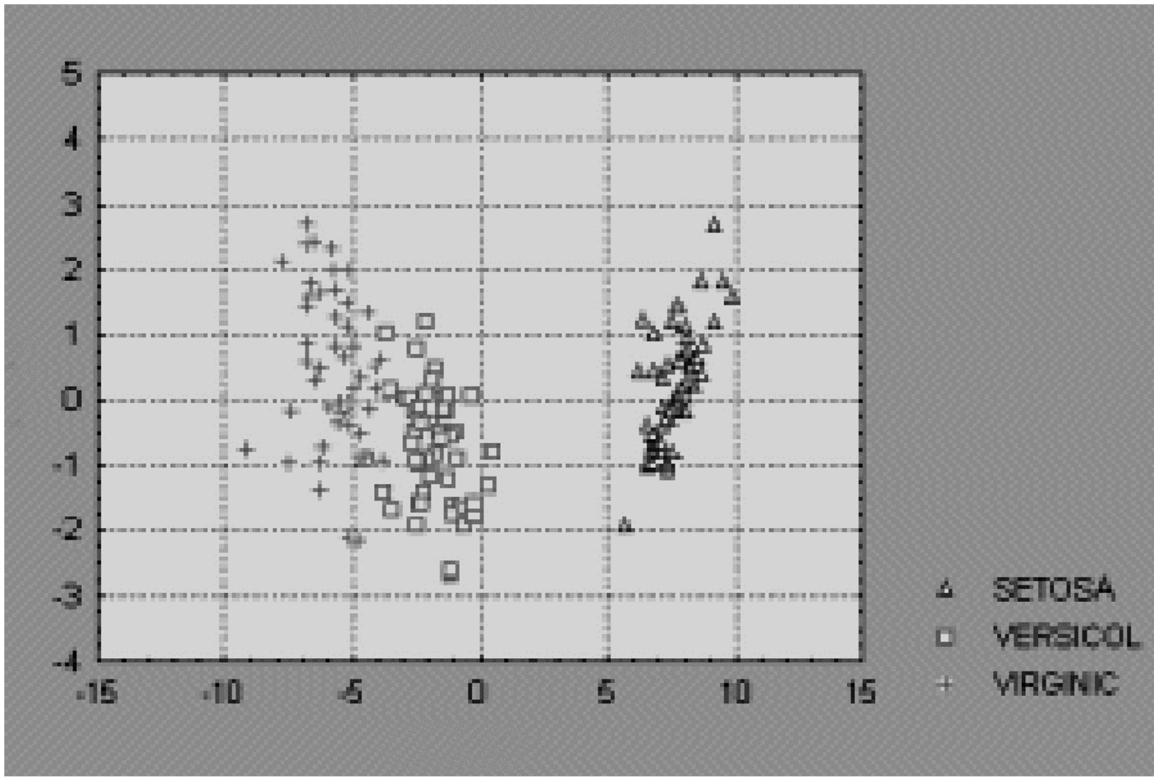


Рисунок 22 – Сравнение двух дискриминантных функций (Root 1, Root 2)

Классификация оказывается лучше для выборки, по которой была проведена оценка дискриминирующей функции (апостеорная классификация), чем для новой выборки (априорная классификация) Поэтому оценивание качества классификации не производят

по той же самой выборке, по которой была оценена дискриминантная функция. Производят кросс-проверку на новых объектах.

Расстояние Махалонобиса является мерой расстояния между изображаемыми точками-объектами (образцами). Объект признается принадлежащим к той группе, к которой он ближе, т. е. когда расстояние Махалонобиса до нее минимально.

Общим итогом классификации является матрица классификации, содержащая число образцов, корректно классифицированных и тех, которые попали не в свои группы. Только классификация новых объектов позволяет определить качество функции классификации. Модели, реализованные в STATISTICA, являются линейными, а функции классификации и дискриминантные функции – линейными комбинациями наблюдаемых величин.

Кластерный анализ

Английское слово «*cluster*» переводится на русский язык как группа, скопление. Главным назначением кластерного анализа является разбиение множества исследуемых объектов и признаков на однородные в определенном смысле группы, или кластеры. Этот анализ относят к классификационному анализу без обучения, противопоставляя его дискриминантному анализу, считающемуся классификационным анализом с обучением (Халафян А. А., 2007).

Задача кластерного анализа заключается в том, чтобы на основании данных, содержащихся во множестве X, разбить множество объектов G на m кластеров $Q_1, Q_2, \ldots, Q_m$ так, чтобы каждый объект Q_1 принадлежал одному и только одному подмножеству разбиения. При этом объекты, отнесенные к одному и тому же кластеру, должны быть сходными, а объекты, принадлежащие разным кластерам – разнородными. Подразделение на кластеры должно удовлетворять критерию оптимальности, в качестве которого обычно берется внутригрупповая сумма квадратов отклонений. Сходство между объектами определяется расстояниями между векторами измерений X_i, X_j.

Чаще всего используются следующие функции расстояний: *evclidian distances* (эвклидова метрика), *square evclidian distances*

(квадрат эвклидовой дистанции), *sity-block Manhatten distances* (манхэттеновское расстояние городских кварталов), *Chebyshev distances* (расстояние Чебышева), *power metric* (степенное расстояние Минковского), *percent disagreement* (процент несогласия). Все приведенные расстояния пригодны, если объекты кластеризации можно представить как точки в *k*-мерном пространстве. В противном случае в качестве дистанции используется $1 - r$, где r – коэффициент корреляции Пирсона. Понятием, противоположным расстоянию, является понятие сходства между объектами G_i и G_j.

Алгоритмы, или процедуры, кластерного анализа подразделяются на иерархические (древовидные) и неиерархические. Наиболее распространены иерархические. Большинство программ, реализующих алгоритм иерархической классификации, предусматривает графическое представление классификации в виде дендрограммы. В системе STATISTICA реализованы так называемые агломеративные методы минимальной дисперсии: *joining (tree clastering)* древовидная кластеризация, *two-way joining* двухходовая кластеризация, *k-means* дивизивный метод *k*-средних. В методе древовидной классификации предусмотрены правила иерархического объединения в кластеры.

Наиболее известный метод представления матрицы расстояний и сходства основан на идее дендрограммы. Это графическое изображение результатов последовательной кластеризации. На рисунке 23 показан один из примеров дендрограммы, заимствованной нами из учебника А. А. Халафян (2007).

В этом примере кластеризации подвергались 6 объектов (A, B, C, D, E, F) по комплексу признаков. Объекты A и C наиболее близки и поэтому объединяются в один кластер на уровне сходства, равном 0,85. Объекты D и E объединяются на уровне 0,8. Теперь имеем 4 кластера: (A, C), (F), (D, E), (B). Далее образуются кластеры (A, C, F) и (E, D, B), соответствующие уровням сходства, равным 0,7 и 0,6 соответственно. Наконец, все объекты группируются в один кластер при уровне сходства, равном 0,5 (см. рисунок 23).

Указывались различные причины принципиально неоднозначных решений задач кластеризации. Вероятно, главная причина в

том, что не существует однозначно наилучшего критерия качества кластеризации. Достаточно целесообразной считается кластеризация «по построению».

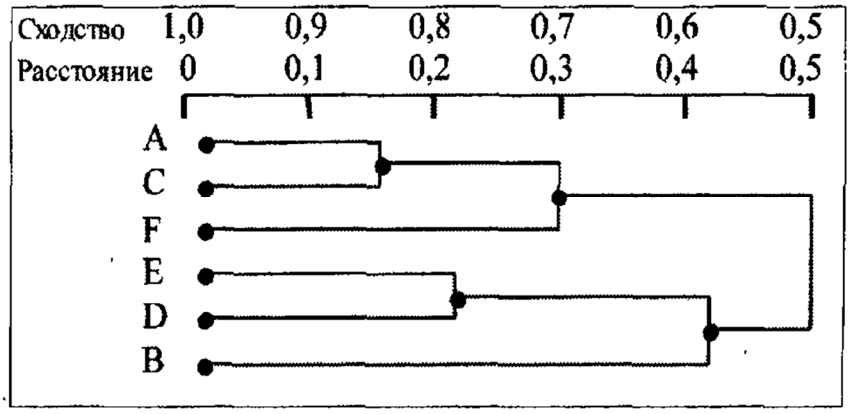


Рисунок 23 – Пример дендрограммы А. А. Халафян

Для окончательного определения качества кластеризации необходимо привлекать экспертов предметной науки, которые могли бы оценить рациональность выделения кластеров. Однако нередко оказывается, что мнения экспертов расходятся.

Факторный анализ

В этом анализе фактором называют гипотетические, непосредственно не измеряемые, скрытые (латентные) переменные, в той или иной мере связанные с измеряемыми характеристиками – проявлениями этих факторов (Буреева Н. Н., 2007). Идея факторного анализа основана на предположении, что имеется ряд величин, не известных исследователю, которые обуславливают проявление различных соотношений между переменными. Структура связей между вероятностью ошибки p и анализируемыми признаками $x^{(1)},\ldots, x^{(p)}$ может быть объяснена тем, что все эти переменные зависят от меньшего числа других, непосредственно не измеряемых факторов $f^{(1)},\ldots, f^{(m)}$ ($m < p$), которые принято называть общими. Таким образом, факторный анализ в широком смысле – совокупность моделей и методов, ориентированных на выявление, конструирование и анализ внутренних факторов по информации об их «внешних» проявлениях. В узком смысле под факторным анализом понимают методы выявления гипотетических (ненаблюдаемых) фак-

торов, призванных объяснить корреляционную матрицу наблюдаемых количественных переменных (Буреева Н. Н., 2007). Факторный анализ относят к методам редукции данных (Халафян А. А., 2007).

Главными целями факторного анализа являются сокращение числа переменных (редукция данных) и определение структуры взаимосвязей между переменными, т. е. классификация переменных. Факторный анализ используется или как метод сокращения данных, или как метод классификации переменных. Новые факторы (переменные) обычно являются линейной комбинацией исходных.

Обычно в моделях факторного анализа предполагаются выполненными следующие предположения (ограничения):

– x имеют многомерное нормальное распределение;

– общие факторы f являются либо некоррелированными случайными величинами с дисперсией 1, либо неизвестными случайными параметрами;

– остатки (остаточные факторы) имеют нормальные распределения, не коррелированны между собой и не зависят от общих факторов.

Кратко рассмотрим использование факторного анализа на примере из области медицины, взятом из работы Н. Н. Буреевой (2007). Исходные показатели представлены в таблице 28.

Таблица 28 – Медицинская характеристика (Буреева, 2007)

Страны	Численность населения	Количество человек на 1 врача	Расходы на здравоохранение	Уровень детской смертности	ВВП на душу населения	Смертность
Россия	145491	235	159	16,8	7700	13,9
Азербайджан	8041	256	99	29,3	3000	9,6
Армения	3787	198	152	15,4	3000	9,7
Белоруссия	101187	222	157	12,5	7500	14
Грузия	5262	182	152	17,6	4600	14,6
Казахстан	16172	265	154	42,1	5000	10,6
Киргизия	4921	301	118	37	2700	9,1
Молдовия	4295	251	143	20,5	2500	12,6
Таджикистан	6087	439	100	53,3	1140	8,6
Туркмения	4737	320	125	48,6	4300	9
Узбекистан	24881	299	116	36,7	2400	8
Украина	49568	224	131	15,3	3850	16,4

После выполнения ряда предусмотренных в системе STATISTICA процедур, приходим к решению, которое можно интерпретировать наглядным способом (рисунок 24).

Оказывается, первый фактор теснее всего связан с переменными *x2, x3, x4, x6*: второй фактор – с *x1* и *x5*. Таким образом, произвели классификацию переменных на две группы (см. рисунок 24).

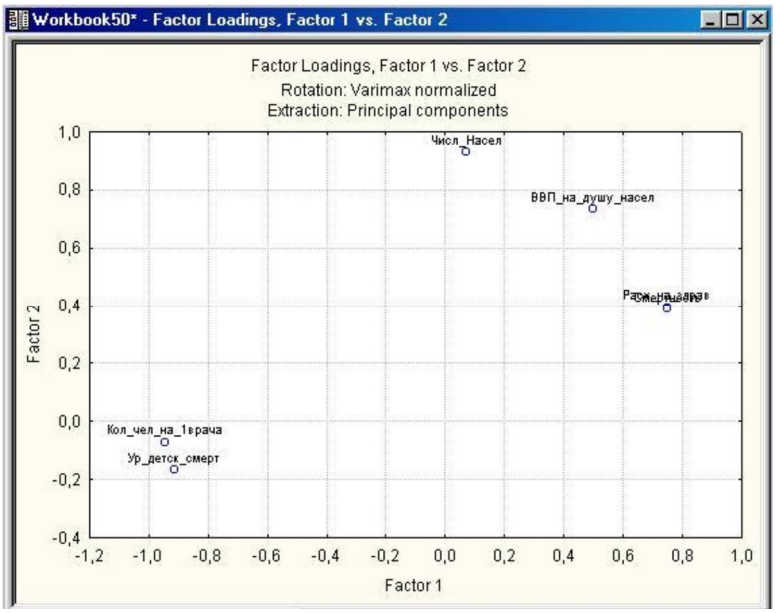


Рисунок 24 – Классификация переменных факторным анализом

Возникает вопрос: сколькими же факторами следует ограничиться на практике? Для этого в ППК STATISTICA реализован критерий Scree plot (Критерий каменистой осыпи) (рисунок 25).

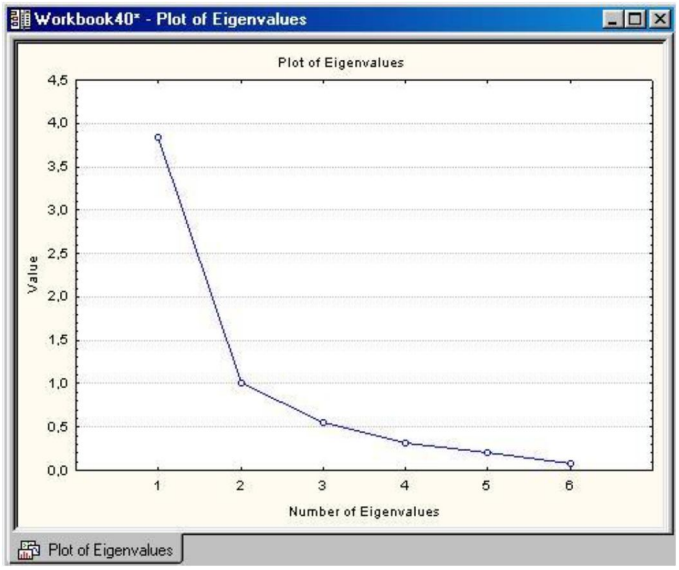


Рисунок 25 – Кривая Scree plot факторного анализа

Очевидно, что в точке с координатами 1, 2 осыпание замедляется наиболее существенно, следовательно, теоретически можно ограничиться двумя факторами (Буреева Н. Н., 2007).

Вопросы для обсуждения

1. Объясните, почему дискриминантный, кластерный и факторный анализы относят к многомерной статистике?
2. Для каких целей используется дискриминантный, кластерный и факторный анализ?
3. Ограничения дискриминантного анализа.
4. Матрица классификации в дискриминантном анализе.
5. Уравнение линейного дискриминантного анализа Фишера.
6. Функции расстояний в кластерном анализе.
7. Иерархические алгоритмы (процедуры) кластерного анализа. Интерпретация дендрограмм.
8. Основная причина принципиально неоднозначных решений задач классификации.
9. Ограничения факторного анализа.
10. Редукция данных как основное направление в использовании факторного анализа.

13 ПЛАНИРОВАНИЕ ИССЛЕДОВАНИЙ И ПРОБЛЕМА ПРОГНОЗИРОВАНИЯ

Вопросы планирования исследований

Планирование исследований и обработка их результатов – две тесно связанные между собой задачи статистического анализа.

Ошибки в планировании исследований приводят к необходимости их повторения, что отбрасывает ученых назад на месяцы, а в случае с лесными и плодовыми культурами – на многие годы.

Должны быть заранее определены и обеспечены:

– схема эксперимента;

– объемы выборок (обычно не менее 30);

– количество повторностей (обычно не менее 3–4);

– репрезентативность выборок (случайный отбор объектов в выборку, случайный выбор мест для растений и делянок);

– принцип единственности различий в однофакторном опыте и др.

Задача сводится к тому, чтобы при возможно минимальных объемах наблюдений получить значимую информацию (результаты).

Приближенные оценки основных статистических показателей и планирование объема выборки

По результатам поискового предварительного исследования приближенно (Плохинский Н. А., 1970):

$$\bar{x} = \frac{x_{\min} + x_{\max}}{2}, \quad s = \frac{x_{\max} - x_{\min}}{K},$$

где K определяется через объемы выборки n (если n = 2–5, то K = 2; n = 6–15, K = 3; n = 16–49, K = 4; n = 50–200, K = 5; n = 200–1000, K = 6; n > 1000, K = 7);

$$s_{\bar{x}} = \frac{x_{\max} - x_{\min}}{K \cdot \sqrt{n}}.$$

Определив так эти параметры, можно наметить объем выборки n для основного эксперимента, обеспечивающий необходимую $s_{\bar{x}}$ (ошибку средней).

Допустим, при приблизительных вычислениях $s_{\bar{x}} = \frac{14,7 - 9,0}{5\sqrt{100}} =$ = 0,114. Если ошибку средней нужно получить в опыте в два раза меньшую, объем выборки должен быть не 100, а 260 (n = 260):

$$s_{\bar{x}} = \frac{14,7 - 9,0}{5\sqrt{260}} = 0,06.$$

Проблема прогнозирования

Эмпирическое выравнивание (сглаживание)

Самое простое сглаживание – линейное сглаживание по трем точкам. При таком выравнивании $y_2' = \frac{y_1 + y_2 + y_3}{3}$, *где* y_1, y_2, y_3 – значения y для трех последовательных значений x_2, y_2' – выравненные значения для x = 2.

Пример 1. Зависимость урожайности семян донника от нормы высева (рисунок 26).

x	1	2	3	4	5	6
y	0,5	4,5	3,5	5,5	3	1
y'	–	2,8	4,5	4,0	3,2	–

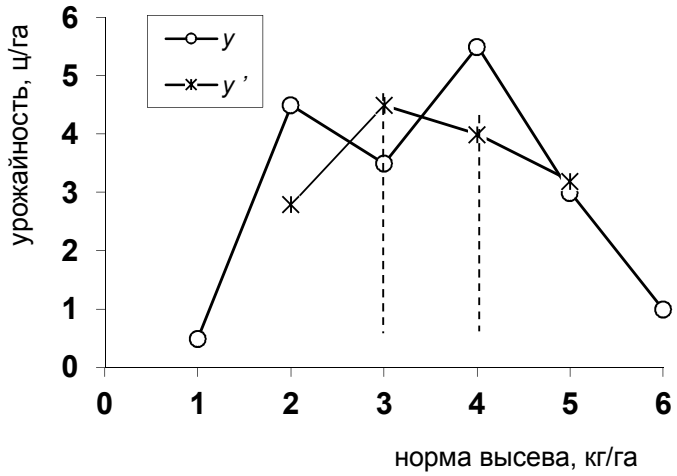


Рисунок 26 – Связь урожайности с нормой высева:

y – фактические значения (исходные данные);

y' – выравненные значения

Ломаная линия по выравненным значениям более плавная, поскольку уменьшена случайная компонента изменчивости. Эта линия свидетельствует о том, что наилучшая норма высева не 4 кг/га, как кажется по исходным данным. Более точную оценку оптимальной нормы высева можно получить при аналитическом сглаживании параболой (поиск экстремума).

При большем количестве пар значений x – y могут использоваться линейные сглаживания по большему числу точек (5, 7 и т. д.). В примере 1 при вычислении выравненных значений y' значения $y_1, y_2, y_3, y_4, y_5, y_6$ берутся с одинаковыми весами (вес каждого равен 1; $y_2' = \frac{y_1 + y_2 + y_3}{3}$, $y_3' = \frac{1y_2 + 1y_3 + 1y_4}{3}$ и т. д.). Так же поступают при линейном эмпирическом выравнивании по большему, чем 3, числу точек.

Линейное выравнивание применяют, если есть основания предполагать наличие линейной связи x – y. В иных случаях используют обычно более сложные нелинейные выравнивания. Широкое распространение получило выравнивание по методу взвешенной скользящей средней. При этом значения $y_1, y_2, \ldots y_n$ учитываются с разными весами. При вычислении y_2 максимальный вес придается y_2, y_3' – y_3 и т. д. Соответствующие формулы здесь представляются излишними. При прогнозировании в системе STATISTICA в качестве основного метода используется ARIMA (Autoregressive Integrated Moving Average) – модели авторегрессии и проинтегрированного скользящего среднего.

Экспоненциальное выравнивание, часто используемое для анализа и прогнозирования временных рядов, является более простым методом, чем ARIMA. В процессе используются все данные. В этом случае исходный ряд динамики $y(t)$ сглаживается с некоторыми экспоненциальными весами. Из самого названия следует, что экспоненциальное сглаживание придает большее значение показателям последних значений x (лет, кварталов, месяцев, дней и т. п.), чем более отдаленным (Боровиков В. П., Ивченко Г. И., 2006). Следует иметь в виду, что экспоненциальное сглаживание – наиболее простой способ прогнозов. Часто он дает более быстрые, эффективные результаты, чем другие методы.

Вообще существуют различные методы прогнозирования временных рядов. Пакет программ STATISTICA предлагает семь групп методов (опций), в том числе экспоненциальное сглаживание (Exponential smoothing & forecasting).

Суть метода в том, что исходный ряд сглаживается с некоторыми экспоненциальными весами, образуется новый временной ряд (с меньшим уровнем шума), поведение которого можно прогнозировать (Боровиков В. П., Ивченко Г. И., 2006). Имеется возможность использования как аддитивной (простое суммирование влияний), так и мультипликативной моделей шума.

На рисунке 27 в качестве примера представлен график, полученный в системе STATISTICA при использовании метода экспоненциального сглаживания и прогнозирования по показателю «годовая сумма осадков». По горизонтали указаны годы, причем 1838 г. присвоено число 1, 1839-му – 2 и т. д.

Выравненный временной ряд, изображенный на графике пунктирной линией, свидетельствует о наличии периодической компоненты изменчивости («вековой» цикл). В обозримом будущем вновь будет происходить рост количества осадков (рисунок 27).

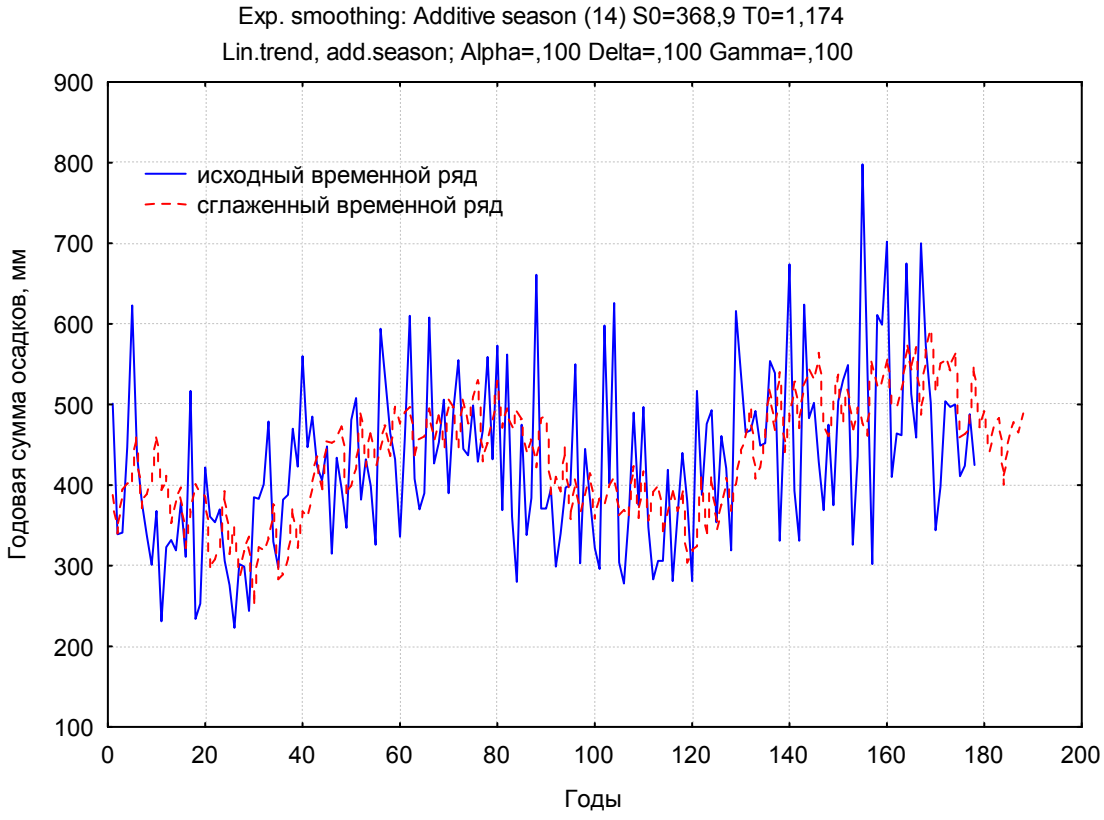


Рисунок 27 – Результаты обработки данных в системе STATISTICA (по данным Луганской метеостанции)

При работе системы STATISTICA вычисляются прогнозные значения зависимой переменной на несколько (по умолчанию на десять) будущих лет. На аддитивных моделях, как и в нашем примере, прогноз строится по формуле

$$Forecast(t) = S(t) + I\,(t - lag),$$

где t – сглаженный «сезонный фактор»:

$$i(t) = i(t - lag) + \lambda(1 - \alpha)e(t)$$

где $e(t)$ – разность между наблюдаемым рядом и прогнозом в момент времени t;

lag – «сезонный период».

Вековой цикл обнаруживается при экспоненциальном выравнивании также и во временных рядах температуры холодного (октябрь – март) и теплого (апрель – сентябрь) полугодий (рисунок 28). При этом периодические колебания температуры теплого и холодного сезонов происходят в противофазах, поэтому среднегодовая температура практически не обнаруживает вековых периодических колебаний.

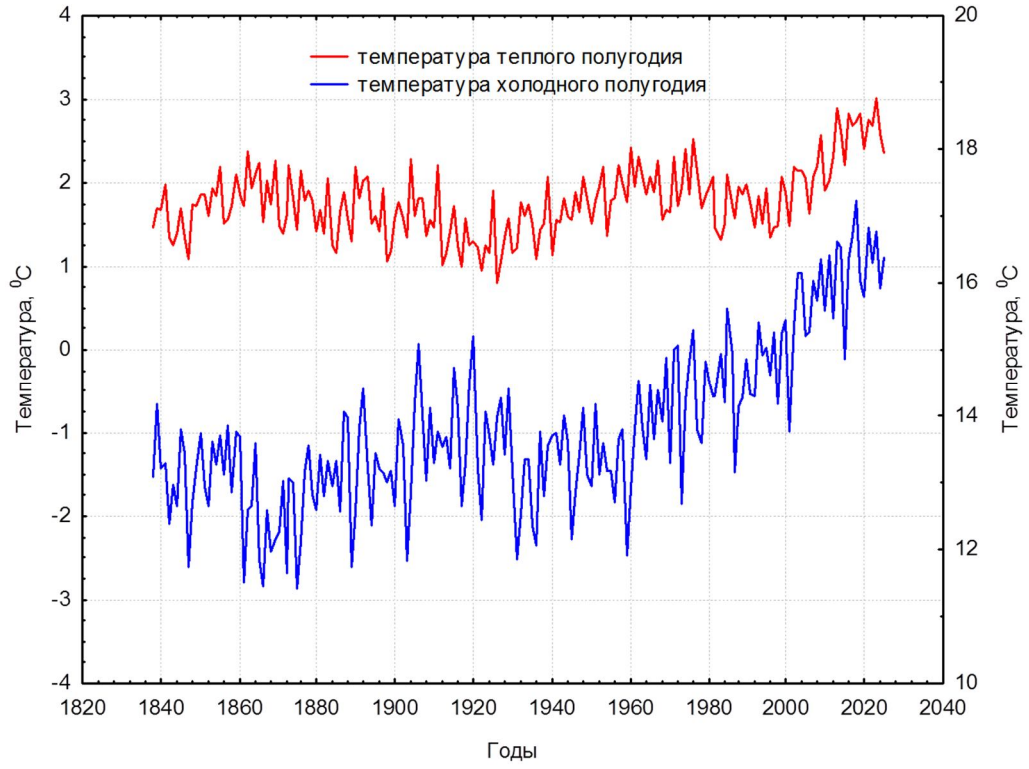


Рисунок 28 – Временные ряды, полученные при экспоненциальном сглаживании (по данным Луганской метеостанции)

Аналитическое сглаживание

Аналитическое сглаживание проводится по формулам (уравнениям регрессии), описывающим связь зависимой (Y) и независимой (X) переменных. Сглаженный ряд значений y в этом случае имеет вид плавной, а не ломаной, как при эмпирическом выравнивании, линии.

Уравнения регрессии позволяют вычислять величину y (зависимой переменной) при любых значениях независимой переменной (переменных) в области интерполяции, в пределах известных крайних значений независимых переменных. В определенной степени эти уравнения пригодны и для экстраполяции, нахождения ожидаемых значений независимой переменной за пределами крайних значений.

В качестве примера прогнозирования значений y в области интерполяции рассмотрим сглаживание квадратичной параболой.

Уравнение параболической регрессии имеет следующий вид:

$$Y = -3{,}00 + 4{,}5x - 0{,}64x^2,$$

где y – урожайность семян в ц/га,

x – нормы высева в килограммах на гектар.

На рисунке 29 изображена ломаная линия, построенная по исходным данным, и квадратичная парабола (ср. с рисунком 23).

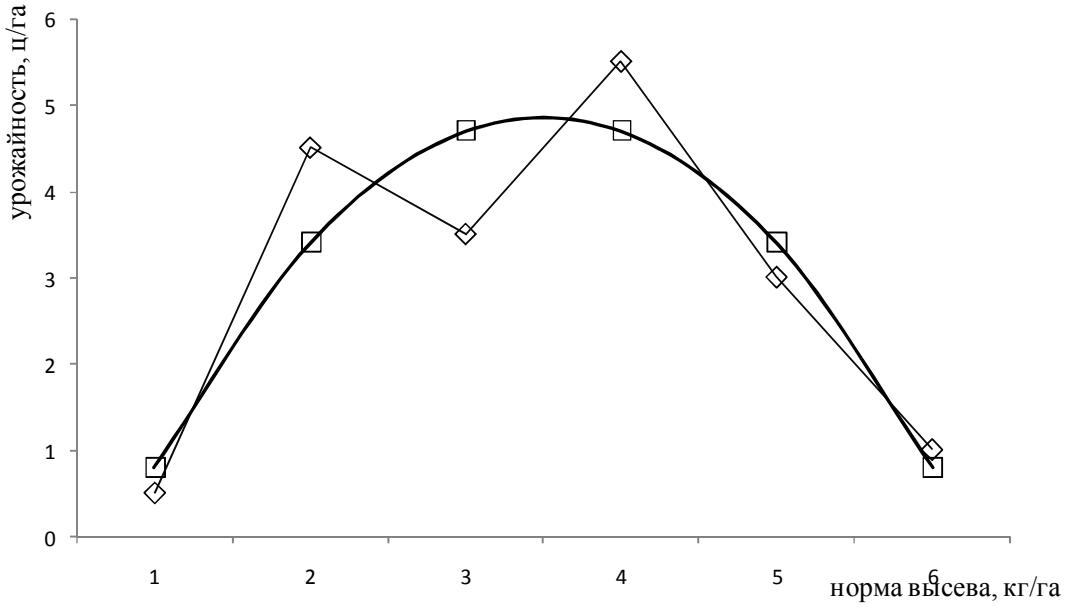


Рисунок 29 – Связь урожайности с нормой высева

При поисках экстремума по параболе получили следующие пары значений признаков *X* и *Y*: 3,4 – 4,869; 3,5 – 4,875; 3,6 – 4,869 (приложение Ж). Очевидно, что максимальную урожайность обеспечивает норма высева 3,5 кг/га. Немного меньшей, а именно 4,7, является ожидаемая урожайность для норм высева 3 кг/га и 4 кг/га.

Заметим, что интересующие агрономов связи нередко имеют такой, как в данном случае, вид (срок посева – урожайность, глубина заделки семян при посеве – урожайность, доза удобрений – урожайность, норма полива – урожайность и др.).

При вычислении ожидаемых значений по параболе или какой-либо другой функции в области интерполяции (в области определения функции) происходит прогнозирование в области интерполяции. Оно возможно и в области экстраполяции (вне области определения функции), хотя оно обычно оказывается несколько менее точным.

Прогнозировать можно только закономерные процессы, поэтому необходим глубокий анализ временных рядов. Должны быть выявлены тенденция или тренд зависимой переменной и периодические колебания (если они есть). Далее по установленной математической модели осуществляется прогнозирование.

Наиболее распространено прогнозирование при регрессионном анализе временных рядов.

Пример 1

При изучении динамики урожайности озимой пшеницы в Луганской области установлены три компоненты ее изменчивости:

1) изменение среднего уровня урожайности (тренд, или тенденция), неплохо описываемое на данном временном интервале параболой;

2) циклические изменения с периодом 16 лет;

3) случайные изменения, т. е. отклонения от наложенной на параболу периодической кривой (рисунок 30).

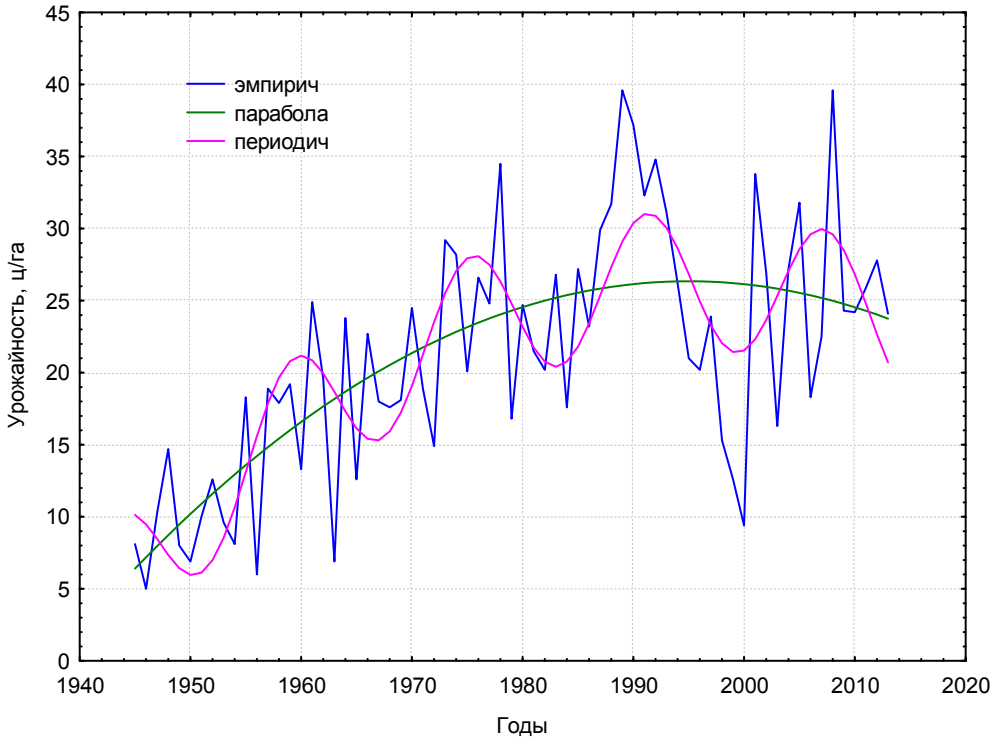


Рисунок 30 – Сглаживание временного ряда параболой и параболой с наложением на нее периодической компоненты

Случайная изменчивость потому и называется случайной, что не прогнозируется. Она велика – отклонение фактических значений от ожидаемых в ту или иную сторону составляет иногда 15 ц/га. Периодические изменения урожайности определяются, вероятно, космическими факторами, влияющими прямо или косвенно (через изменения климата и погодных условий).

Из-за случайной изменчивости прогноз урожайности по уравнению регрессии для какого-либо конкретного года будет ненадежным. Поэтому виды на урожай в этом случае приходится определять по долгосрочному прогнозу погоды. При этом каждый раз, когда ошибочным оказывается метеопрогноз, неверен и прогноз урожайности. А неверным долгосрочный метеопрогноз бывает, к сожалению, довольно часто.

Положение с прогнозом среднего уровня урожайности за несколько лет лучше. Случайные отклонения от тренда в обе стороны в значительной степени компенсируют друг друга, и прогноз оказывается довольно точным.

На основе учитывающих циклические колебания моделей динамики урожайности озимой пшеницы нами был сделан оптимистичный прогноз на конец 80-х – начало 90-х гг. прошлого века (верхний полупериод) на фоне весьма пессимистичных прогнозов других авторов. Подтвердился наш прогноз.

Все подобные прогнозы делаются при допущении статус-кво. Иначе говоря, прогноз подтвердится лишь в том случае, если в дальнейшем не будет происходить ничего такого, что может существенно изменить принятые в математической модели тренда закономерности (парабола, периодика).

Кстати, в нашем примере парабола лучше, чем прямая линия, согласуется с изменением среднего уровня урожайности на исследованном временном интервале, в связи с чем ей и отдано предпочтение. Однако парабола непригодна для прогнозирования на большие промежутки времени (как вперед, так и назад) в области экстраполяции. Дело в том, что ветви параболы пересекают ось *X*, т. е. со временем ожидаемые значения урожайности окажутся равными нулю (такую урожайность в принципе еще представить можно – урожая нет), а потом и меньше нуля (такая ситуация вообще невозможна).

Вероятно, средний уровень урожайности описывается не параболой, а периодической функцией с периодом около 80–100 лет, на которую «наложены» циклические колебания с периодом 16 лет. Однако анализ вековой цикличности пока невозможен – для этого нужны данные хотя бы об одном полном цикле (от самой верхней точки до следующей верхней или от самой нижней до другой такой же). Данные об урожайности сельскохозяйственных культур имеются менее чем за 80 лет (Луганская область образована в 1938 г.). Можно оценить, что же дает математико-статистическая обработка исходных данных.

Исходные данные (выборка) представляют собой 69 чисел, при рассмотрении которых без математико-статистической обработки какие-либо закономерности установить нельзя. Ясно лишь, что урожайность варьировала от 5 ц/га до почти 40 ц/га.

После вычисления элементарных статистик стало возможным утверждать, что средняя урожайность за исследованные 69 лет составляет около 21,1 ц/га. Изменчивость по годам очень сильная (по значению C_v), что ставит задачу изучения влияния факторов, обусловливающих эту изменчивость.

После составления вариационного ряда и анализа распределения установили, что фактическое распределение можно аппроксимировать нормальным распределением. Значит, на урожайность влияют многие факторы, ни один из которых сильно не доминирует над другими.

При использовании парного корреляционного анализа (у – урожайность, $x_1 – x_{48}$ – среднемесячные температуры и суммы осадков за два года) выяснилось, что урожайность действительно зависит от многих из этих факторов.

При использовании множественного корреляционного анализа установили, что почти 85 % изменчивости урожайности по годам определяется температурой и осадками (независимые переменные $x_1 – x_{48}$), 15 % – не организованными в нашем исследовании другими природными агроэкологическими, а также антропогенными факторами.

Анализ временного ряда позволил создать математическую модель динамики урожайности (тренд – близкая к параболе кривая, на которую наложена циклическая изменчивость с периодом 16 лет), имеющую прогностическую ценность биометрии.

После использования для анализа методов биометрии мы установили динамику урожайности озимой пшеницы в Луганской области. Следует иметь в виду, что методы прогнозирования требуют от пользователей ПК высокой профессиональной и математической грамотности. Составление удачного, т. е. оправдавшегося, прогноза – это и наука, и одновременно искусство, включающее элемент интуиции. Не случайно в цивилизованных странах труд специалистов по прогнозированию высоко оплачивается.

Наличие быстродействующих ПК с отличным программным обеспечением в области математико-статистической обработки

данных привело к тому, что сейчас от специалиста с высшим образованием требуется главным образом умение выбрать необходимый метод и подходящее программное обеспечение. Отпала необходимость проводить сами вычисления, обычно громоздкие. В результате интеллект исследователей используется с большей отдачей. Возможности биометрии продолжают расширяться (в частности, внедрение метода нейронных сетей и других методов так называемого искусственного интеллекта).

Вопросы для обсуждения

1. Вопросы планирования исследований.

2. Приближенные оценки $\overline{x}$, s^2 и $s_{\bar{x}}$ и планирование необходимого объема выборки.

3. Объясните суть проблемы прогнозирования.

4. Области интерполяции и экстраполяции.

5. Вычисление прогнозных значений зависимой переменной при различных значениях независимой переменной в области интерполяции.

6. Поиск экстремума зависимой переменной при использовании квадратичной параболы.

7. Временные ряды: нахождение тенденции или тренда, периодических колебаний (если они есть) и случайных изменений.

8. Регрессионный анализ временных рядов (независимая переменная X – время, Y – зависимая переменная).

9. Методы прогнозирования временных рядов, которые используются в системе STATISTICA и (или) других известных программных продуктах.

10. Что дает математико-статистическая обработка данных?

ТЕСТОВЫЕ ЗАДАНИЯ

МОДУЛЬ 1. Элементарная статистика

1. Кто из перечисленных ученых ввел в науку термин «биометрия»?

1) Г. Мендель

2) А. Колмогоров

3) Н. Бейли

4) П. Чебышев

5) Ф. Гальтон

2. Какая из математических наук является основой биометрии?

1) алгебра

2) теория вероятностей

3) геометрия

4) математический анализ

5) теория чисел

3. Какая из математических наук является основой биометрии?

1) алгебра

2) геометрия

3) математическая статистика

4) математический анализ

5) теория чисел

4. Выберите из приведенных выражений правильное определение биометрии:

1) совокупность математико-статистических методов

2) наука о закономерностях внешнего и внутреннего строения растений

3) наука о взаимоотношениях растений с окружающей средой

4) наука о количественных отношениях и пространственных форм действительного мира

5) наука о функционировании природных и техногенных систем

5. Какие из перечисленных задач стоят перед биометрией?

1) системный анализ

2) установление значимости параметров

3) изучение информационных процессов

4) разработка вычислительных систем

5) разработка программного обеспечения

6. Какие задачи из перечисленных стоят перед биометрией?

1) системный анализ

2) изучение информационных процессов

3) установление значимости связей

4) разработка вычислительных систем

5) разработка программного обеспечения

7. Какие из перечисленных признаков следует отнести к качественным?

1) длина листовой пластинки

2) высота растения

3) число листьев прикорневой розетки

4) венчик сростнолепестной

5) число тычинок в цветке

8. Какие из перечисленных признаков следует отнести к качественным?

1) длина листовой пластинки

2) высота растения

3) число листьев прикорневой розетки

4) число тычинок в цветке

5) окраска лепестков цветка

9. Какие из перечисленных признаков следует отнести к количественным?

1) длина листовой пластинки

2) венчик сростнолепестной

3) очередное расположение листьев

4) завязь нижняя

5) окраска лепестков цветка

10. Какие из перечисленных признаков следует отнести к количественным?

1) венчик сростнолепестной

2) число зерновок в колосе

3) очередное расположение листьев

4) завязь нижняя

5) окраска лепестков цветка

11. Укажите правильно округленные числа

1) 45,346 – 45,4

2) 8,644 – 8,65

3) 9,425 – 9,42

4) 3,585 – 3,5

5) 3,373 – 3,38

12. Что означает данное выражение: «Множество относительно однородных, но индивидуально различимых единиц, объединенных для совместного изучения»?

1) Определение вида

2) Определение популяции

3) Определение ареала

4) Определение статистической совокупности

5) Определение жизни

13. Что заключено в данном выражении: «все множество особей вида, произрастающих в пределах его ареала, отобранное с целью изучения изменчивости признаков»?

1) Определение вида

2) Определение популяции

3) Определение ареала

4) Определение статистической совокупности

5) Определение генеральной совокупности

14. Укажите правильное продолжение определения: «отобранная для исследования часть генеральной совокупности называется…»

1) выборкой

2) навеской

3) дозой

4) репрезентативностью

5) рендомизацией

15. Какой должна быть выборка обеспечивающая несмещенные оценки параметров генеральной совокупности?

1) большой

2) объемной

3) репрезентативной

4) малой

5) соответствующей

16. Какой должна быть выборка, которая должна обеспечить несмещенные оценки параметров генеральной совокупности?

1) большой

2) объемной

3) малой

4) рендомизированной

5) соответствующей

17. Что означает «… расхождение между результатами выборочного наблюдения и истинным значением наблюдаемой величины?»

1) отбор

2) ошибка

3) типичность

4) подтасовка

5) случай

18. К какому виду ошибок можно отнести методические огрехи в постановке опыта?

1) неустранимые

2) случайные

3) систематические

4) устранимые

5) грубые

19. Какая из ошибок характеризует невозможность обеспечения абсолютного совпадения выборочных и генеральных параметров?

1) ошибка репрезентативности

2) случайные ошибки

3) систематические ошибки

4) устранимые ошибки

5) грубые ошибки

20. Укажите пример систематической ошибки

1) нарушение правил рендомизации

2) нарушение требований к полевому опыту

3) использование неповеренных измерительных приборов

4) ошибки в записях наблюдений

5) субъективность

21. Какие из перечисленных ошибок отличаются однонаправленностью?

1) репрезентативности

2) случайные

3) систематические

4) устранимые

5) грубые

22. Выберите правильный пример с дискретным типом изменчивости признаков боба

1) число семян

2) длина

3) вес

4) цвет

5) возраст

23. Выберите правильный пример с непрерывным типом изменчивости

1) число семян в бобе

2) число цветков в соцветии

3) количество листьев в прикорневой розетке

4) длина листа

5) форма листа

24. Определение какого термина содержится в выражении: «свойство, проявлением которого один предмет отличается от другого»?

1) объект

2) форма

3) ошибка

4) признак

5) тип

25. Что является «абстрактной характеристикой ряда, заменяющей собой варьирующее значение признака»?

1) дисперсия

2) варианса

3) доверительный интервал

4) критерий Фишера

5) средняя арифметическая

26. Какая из приведенных формул представляет вычисление средней арифметической?

1) $\bar{x} = (x_1 + x_2 + x_3 + \ldots x_n) / n$

2) $c^2 = a^2 + b^2$

3) $m = (a+b) / 2$

4) $cv = s \; x \; 100 / \bar{x}$

5) $t = p_1 - p_2 / m$

27. Какой из приведенных показателей характеризует степень изменчивости признака?

1) среднее арифметическое

2) мода

3) медиана

4) среднее квадратичное отклонение

5) среднее геометрическое

28. Какой из приведенных показателей сохраняет ту же размерность, что и частные значения признака?

1) показатель асимметрии

2) показатель эксцесса

3) коэффициент вариации

4) среднее арифметическое значение

5) дисперсия

29. Что обозначает показатель, который определяет количество значений, необходимое для восстановления утерянного?

1) среднее арифметическое

2) коэффициент вариации

3) среднее геометрическое

4) объем выборки

5) число степеней свободы

30. Какая из приведенных формул представляет вычисление коэффициента вариации?

1) $\bar{x} = (x_1 + x_2 + x_3 + \ldots x_n) / n$

2) $c^2 = a^2 + b^2$

3) $m = (a + b) / 2$

4) $c_v = s \cdot 100 / \bar{x}$

5) $t = p_1 - p_2 / m$

31. Какой буквой обычно обозначается выборочное среднее квадратичное отклонение?

1) a

2) d

3) n

4) r

5) s

32. Какой критерий чаще используется на практике при сортоиспытании?

1) сравнение средних

2) выборочная дисперсия

3) среднее квадратичное отклонение

4) стандартное отклонение

5) коэффициент вариации

33. Какой из приведенных параметров выборки следует отнести к элементарным одномерным статистикам?

1) критерий Стьюдента

2) среднее арифметическое значение

3) критерий Фишера

4) хи-квадрат

5) критерий лямбда

34. Какой из приведенных параметров выборки следует отнести к элементарным одномерным статистикам?

1) критерий Стьюдента

2) критерий Фишера

3) дисперсия

4) хи-квадрат

5) критерий лямбда

35. Какой из приведенных параметров выборки следует отнести к элементарным одномерным статистикам?

1) критерий Стьюдента

2) критерий Фишера

3) хи-квадрат

4) коэффициент вариации

5) критерий лямбда

36. Какой из приведенных параметров выборки следует отнести к элементарным одномерным статистикам?

1) критерий Стьюдента

2) критерий Фишера

3) хи-квадрат

4) критерий лямбда

5) среднее квадратичное отклонение

37. Укажите правильное продолжение следующего определения: «Каждое частное значение, которое способен принимать данный признак, называется…»

1) вариантой

2) числом

3) порядком

4) последовательностью

5) изменчивостью

38. Укажите правильное продолжение следующего выражения: «Количество экземпляров, обладающих данным значением признака, называют…»

1) последовательностью

2) изменчивостью

3) группировкой

4) частотой встречаемости

5) несгруппированной совокупностью

39. Что образует совместный ряд вариант и соответствующих им частот?

1) несгруппированную совокупность

2) вариационный ряд

3) атрибутивный ряд

4) размах варьирования

5) дискретную изменчивость

40. Какую из приведенных целей преследует распределение исходных данных в вариационный ряд?

1) организацию отбора

2) нумерацию вариантов

3) выявление закономерностей варьирования

4) определение порядка признака

5) определение лимитов

41. При каком виде изменчивости признака при создании вариационного ряда следует объединять варианты в классы?

1) дискретной

2) непрерывной

3) мутационной

4) модификационной

5) адаптивной

42. По какой формуле рассчитывают величину классового интервала?

1) $\bar{x} = (x_1 + x_2 + x_3 + \ldots x_n) / n$

2) $c^2 = a^2 + b^2$

3) $m = (a + b) / 2$

4) $i = Xmax - Xmin / k$

5) $t = p_1 - p_2 / m$

43. Какой буквой греческого алфавита обозначается действие суммирования?

1) альфа (α)

2) бета (β)

3) эта (ή)

4) ню (ν)

5) сигма (σ)

44. Диаграмма, в которой величина показателя изображается графически в виде столбика, это?

1) вариационная кривая

2) парабола

3) гистограмма

4) гипербола

5) кумулятивная огива

45. Как на языке теории вероятностей называется всякий результат однократного испытания?

1) исход

2) случай

3) опыт

4) событие

5) акт

46. Назовите правильный исход события, которое является заранее предсказуемым

1) достоверный

2) невозможный

3) случайный

4) совместный

5) несовместный

47. По какой из приведенных формул можно рассчитать вероятность события?

1) $\bar{x} = (x_1 + x_2 + x_3 + \ldots x_n) / n$

2) $c^2 = a^2 + b^2$

3) $m = (a + b) / 2$

4) $P(A) = m / n$

5) $t = p_1 - p_2 / m$

48. Как называется вероятность, которую можно указать до опыта?

1) практически невозможной

2) практически достоверной

3) априорной

4) апостериорной

5) случайной

49. Найдите правильное продолжение выражения: «закон нормального распределения выражает …»

1) попытку найти один или два решающих фактора изменчивости

2) возможность проявления асимметрии и эксцесса

3) распределение исходных единичных значений

4) зависимость между вероятностью и нормированным отклонением

50. Какой кривой описывается функция нормального распределения?

1) колоколообразной кривой

2) параболой

3) синусоидой

4) гистограммой

5) кумулятивной огивой

51. Какими из перечисленных параметров определяется нормальное распределение?

1) числом степеней свободы

2) показателем линейной связи

3) коэффициентом ассоциации Юла

4) критерием Краскелла – Уоллиса

5) средней величиной и дисперсией

52. О чем свидетельствует скошенный график вариационной кривой?

1) о наличии асимметрии

2) о наличии эксцесса

3) о наличии нормального распределения

4) о наличии биномиального распределения

5) о наличии распределения Пуассона

53. О чем свидетельствует двухвершинная вариационная кривая?

1) о наличии асимметрии

2) о наличии эксцесса

3) о наличии нормального распределения

4) о наличии биномиального распределения

5) о наличии распределения Пуассона

54. В каких случаях используют формулу Пуассона?

1) при больших объемах выборки

2) при распределении редких событий

3) при нормальном распределении

4) при небольших отклонениях от нормального распределения

5) при строго симметричных распределениях

55. В каких случаях при сравнении признаков используют понятие нулевой гипотезы?

1) при нормальном распределении

2) при больших объемах выборки

3) при отсутствии различий между эмпирическими рядами

4) при распределении редких событий

5) если та или иная степень различия имеет место

56. Какой из перечисленных критериев представляет собой сумму квадратов отклонений эмпирических частот от вычисленных или ожидаемых частот, отнесенную к теоретическим частотам?

1) критерий хи-квадрат

2) критерий знаков

3) критерий Уилкоксона

4) Х-критерий Ван-дер-Вардена

5) критерий Фишера

57. Какой из перечисленных критериев обычно применяется при сравнении средних арифметических двух выборок?

1) критерий хи-квадрат

2) *t*-критерий Стьюдента

3) критерий Уилкоксона

4) *X*-критерий Ван-дер-Вардена

5) критерий знаков

58. Какой из перечисленных критериев применяется при оценке различий дисперсий выборок?

1) критерий хи-квадрат

2) *t*-критерий Стьюдента

3) критерий Уилкоксона

4) *X*-критерий Ван-дер-Вардена

5) критерий Фишера

59. Какой из перечисленных критериев применяется при оценке разности между эмпирическими долями из неравновеликих выборок?

1) критерий хи-квадрат

2) *t*-критерий Стьюдента

3) критерий Уилкоксона

4) *F*-критерий Фишера

5) *X*-критерий Ван-дер-Вардена

60. Если для нескольких параметров характерно гауссово распределение, то как распределена алгебраическая сумма этих параметров?

1) по Пуассону

2) по Максвеллу

3) по нормальному закону

4) экспоненциально

5) по закону случайных чисел

61. При установлении значимости различий в каком случае следует применять *t*-критерий Стьюдента?

1) разность выборочных средних арифметических значений

2) различия средних квадратичных отклонений

3) сравнение распределений

4) различие дисперсий

5) различие коэффициентов вариации

62. При установлении значимости различий в каком случае следует применять *F*-критерий Фишера?

1) разность выборочных средних арифметических значений

2) различия средних квадратичных отклонений

3) средние распределений

4) различие дисперсий

5) различие коэффициентов вариации

63. При установлении значимости различий в каком случае следует применять хи-квадрат?

1) разность выборочных средних арифметических значений

2) различия средних квадратичных отклонений

3) сравнение распределений

4) различие дисперсий

5) различие коэффициентов вариации

64. Какой из критериев различий следует использовать при оценке значимости разности средних арифметических значений?

1) *t*-критерий Стьюдента

2) критерий хи-квадрат

3) критерий знаков

4) *T*-критерий Уилкоксона

5) критерий лямбда Колмогорова – Смирнова

65. Какой из критериев следует использовать при оценке значимости различий дисперсий?

1) критерий хи-квадрат

2) критерий знаков

3) *T*-критерий Уилкоксона

4) критерий лямбда Колмогорова – Смирнова

5) *F*-критерий Фишера

66. При какой вероятности ошибки p различие статистик считается значимым?

1) $0{,}05 < p$

2) $0{,}01 < p < 0{,}05$

3) $0{,}001 < p < 0{,}01$

4) $p < 0{,}001$

67. При какой вероятности ошибки p различие статистик считается не значимым?

1) $0{,}05 < p$

2) $0{,}01 < p < 0{,}05$

3) $0{,}001 < p < 0{,}01$

4) $p < 0{,}001$

68. При какой вероятности ошибки p различие статистик считается очень значимым?

1) $0{,}05 < p$

2) $0{,}01 < p < 0{,}05$

3) $0{,}001 < p < 0{,}01$

4) $p < 0{,}001$

69. При какой вероятности ошибки p различие статистик считается максимально значимым?

1) $0,05 < p$

2) $0,01 < p < 0,05$

3) $0,001 < p < 0,01$

4) $p < 0,001$

70. Какой уровень значимости (p-уровень) принимается за достаточный в обычных биологических, сельскохозяйственных и лесоводческих исследованиях?

1) $0,05 < p$

2) $0,01 < p < 0,05$

3) $0,001 < p < 0,01$

4) $p < 0,001$

71. Какой уровень значимости (p-уровень) принимается за достаточный при проверочных опытах в биологических, сельскохозяйственных и лесоводческих исследованиях и в экономической работе?

1) $0,05 < p$

2) $0,01 < p < 0,05$

3) $0,001 < p < 0,01$

4) $p < 0,001$

72. Какой уровень значимости (p-уровень) принимается за достаточный при разрешении спорных вопросов и при исследовании вредных и ядовитых веществ?

1) $0,05 < p$

2) $0,01 < p < 0,05$

3) $0,001 < p < 0,01$

4) $p < 0,001$

73. Получен следующий вариационный ряд: 8, 9, 9, 10, 10, 10, 10, 11, 11, 12. Какое среднее арифметическое правильное?

1) 9,5

2) 10,0

3) 10,5

4) 11,0

5) 11,5

74. Сумма квадратов отклонений вариант от средней равна 90, а объем выборки 91. Чему равна несмещенная оценка дисперсии?

1) 1
2) 2
3) 3
4) 4
5) 5

75. Найдите значение коэффициента вариации, если среднее арифметическое значение равно 10, а дисперсия 4.

1) 5 %
2) 10 %
3) 15 %
4) 16 %
5) 20 %

76. Выборка содержит следующие варианты: 3, 4, 5. Укажите объем выборки.

1) 1
2) 2
3) 3
4) 4
5) 5

77. Выборка содержит следующие варианты: 3, 4, 5. Укажите нижний предел изменчивости (Lim_{min}).

1) 1
2) 2
3) 3
4) 4
5) 5

78. Выборка содержит следующие варианты: 3, 4, 5. Укажите верхний предел изменчивости (Lim_{max}).

1) 1
2) 2
3) 3
4) 4
5) 5

79. Выборка содержит следующие варианты: 3, 4, 5. Укажите размах изменчивости.

1) 1
2) 2
3) 3
4) 4
5) 5

80. Выборка содержит следующие варианты: 3, 4, 5. Укажите значение средней арифметической.

1) 1
2) 2
3) 3
4) 4
5) 5

81. Выборка содержит следующие варианты: 3, 4, 5. Определите значение дисперсии

1) 1
2) 2
3) 3
4) 4
5) 5

82. Получена выборочная совокупность: 1, 2, 2, 3, 3, 3, 3, 3, 4, 4, 5. Укажите нижний предел варьирования признака

1) 1
2) 2
3) 3
4) 4
5) 5

83. Получена выборочная совокупность: 1, 2, 2, 3, 3, 3, 3, 3, 4, 4, 5. Укажите верхний предел варьирования признака

1) 1
2) 2
3) 3
4) 4
5) 5

84. Получена выборочная совокупность: 1, 2, 2, 3, 3, 3, 3, 3, 4, 4, 5. Укажите верхний лимит варьирования в совокупности

1) 1
2) 2
3) 3
4) 4
5) 5

85. Получена выборочная совокупность: 1, 2, 2, 3, 3, 3, 3, 3, 4, 4, 5. Укажите нижний лимит варьирования в совокупности

1) 1
2) 2
3) 3
4) 4
5) 5

86. Получена выборочная совокупность: 1, 2, 2, 3, 3, 3, 3, 3, 4, 4, 5. Укажите размах изменчивости в совокупности

1) 1
2) 2
3) 3
4) 4
5) 5

87. Вероятность того, что ребенок будет мальчиком, равна $\frac{1}{2}$, а вероятность того, что он у данной супружеской пары будет светловолосым, равна $\frac{1}{4}$. Определите вероятность того, что ребенок будет голубоглазым мальчиком (расщепление по полу и цвету глаз – независимые события)

1) $\frac{9}{16}$
2) $\frac{3}{16}$
3) $\frac{1}{2}$
4) $\frac{1}{4}$
5) $\frac{1}{8}$

88. Признаки «цвет семян» и «форма поверхности семян» у гороха распределяется в F_2 независимо. Вероятность того, что семя будет желтым, равна $\frac{3}{4}$, зеленым – $\frac{1}{4}$; гладким – $\frac{3}{4}$, а морщинистым – $\frac{1}{4}$. Определите вероятность того, что горошина будет желтой и морщинистой

1) $\frac{9}{16}$

2) $\frac{3}{16}$

3) $\frac{1}{2}$

4) $\frac{1}{4}$

5) $\frac{1}{8}$

89. Предлагаемое студенту тестовое задание сопровождается пятью вариантами ответа. Определите среднюю вероятность чисто случайного угадывания (в процентах)

1) 10 %

2) 20 %

3) 30 %

4) 40 %

5) 50 %

90. Предлагаемое студенту тестовое задание сопровождается тремя вариантами ответа. Определите среднюю вероятность чисто случайного угадывания (в долях)

1) 1

2) $\frac{1}{2}$

3) $\frac{1}{3}$

4) $\frac{1}{4}$

5) $\frac{1}{5}$

91. Какое количество правильных ответов из 10 тестовых заданий в среднем можно случайно угадать, если число вариантов ответов равно 5?

1) 1
2) 2
3) 3
4) 4
5) 5

92. Какое количество правильных ответов из 12 тестовых заданий в среднем можно случайно угадать, если число вариантов ответов равно 3?

1) 1
2) 2
3) 3
4) 4
5) 5

93. Какое количество правильных ответов из 12 тестовых заданий в среднем можно случайно угадать, если число вариантов ответов равно 4?

1) 1
2) 2
3) 3
4) 4
5) 5

94. Чему равно среднее квадратичное отклонение, если дисперсия равна 25?

1) 1
2) 2
3) 3
4) 4
5) 5

95. Чему равно среднее квадратичное отклонение, если дисперсия равна 16?

1) 1

2) 2
3) 3
4) 4
5) 5

96. Чему равно среднее квадратичное отклонение, если дисперсия равно 0,1?

1) 0,01
2) 0,04
3) 0,09
4) 0,16
5) 0,25

97. Чему равна варианса, если квадратичное отклонение равно 0,2?

1) 0,01
2) 0,04
3) 0,09
4) 0,16
5) 0,25

98. Чему равна варианса, если квадратичное отклонение равно 0,3?

1) 0,01
2) 0,04
3) 0,09
4) 0,16
5) 0,25

99. Чему равна варианса (дисперсия), если квадратичное отклонение равно 0,4?

1) 0,25
2) 0,16
3) 0,09
4) 0,04
5) 0,01

100. Чему равна варианса (дисперсия), если квадратичное отклонение равно 0,5?

1) 0,25
2) 0,16
3) 0,09

4) 0,04

5) 0,01

МОДУЛЬ 2. Корреляционный, регрессионный и дисперсионный анализ

1. Определение какого термина заключено в следующем выражении: наличие взаимной согласованности в изменчивости двух или нескольких признаков?

1) корреляция

2) причинность

3) дисперсия

4) асимметрия

5) регрессия

2. Какой из перечисленных ученых ввел в науку термин «корреляция»?

1) Г. Мендель

2) А. Колмогоров

3) Н. Бейли

4) П. Чебышев

5) Ф. Гальтон

3. Какой из перечисленных видов анализа изучает сопряженную изменчивость двух или нескольких признаков?

1) однофакторный дисперсионный

2) двухфакторный дисперсионный

3) корреляционный

4) матричный

5) регрессионный

4. При какой корреляционной связи равномерные изменения одного признака соответствуют равномерным пропорциональным изменениям другого?

1) прямой (положительной)

2) обратной (отрицательной)

3) нулевой

4) полной

5) линейной

5. При какой корреляционной связи равномерным изменениям одного признака соответствуют неравномерные, но подчиняющиеся определенной закономерности изменения другого?

1) прямой (положительной)

2) обратной (отрицательной)

3) линейной

4) нелинейной

5) полной

6. В каких пределах колеблется величина коэффициента линейной корреляции?

1) около 0

2) от 0 до $\pm$ 1

3) в пределах 0,1

4) в пределах 0,01

5) в пределах 0,05

7. В рамки какого принципа укладывается следующее утверждение: «Все органы живого существа составляют систему, причем части ее зависят друг от друга таким образом, что изменение одной части влечет за собой изменение других»?

1) принцип координации

2) принцип причинности

3) принцип корреляции

4) принцип симметрии

5) закон равновесия органов

8. Какой показатель представляет собой разность между вариантой и средней арифметической, деленную на величину среднего квадратического отклонения?

1) критерий Стьюдента

2) нормированное отклонение

3) коэффициент вариации

4) коэффициент корреляции

5) критерий лямбда

9. Каким из приведенных символов обозначается нормированное отклонение?

1) *r*

2) *t*

3) *cv*

4) *n*

5) *s*

10. Какой из приведенных показателей является мерилом тесноты прямолинейной связи двух признаков?

1) критерий Стьюдента

2) коэффициент ассоциации Юла

3) коэффициент вариации

4) коэффициент корреляции

5) критерий лямбда

11. Каким из приведенных символов обозначается коэффициент парной корреляции?

1) *r*

2) *t*

3) *cv*

4) *n*

5) *s*

12. Коэффициент множественной корреляции $R = 0,7$. Чему равен коэффициент множественной детерминации?

1) 0,25

2) 0,36

3) 0,49

4) 0,64

5) 0,81

13. Какой показатель используется для измерения нелинейной зависимости сопряженных признаков и описывающий ее двусторонне?

1) критерий Стьюдента

2) нормированное отклонение

3) коэффициент вариации

4) коэффициент корреляции

5) корреляционное отношение

14. В каких пределах колеблется величина корреляционного отношения?

1) около 0

2) от 0 до –1

3) от 0 до +1

4) в пределах 0,01

5) в пределах 0,05

15. Какой буквой греческого алфавита обозначается корреляционное отношение?

1) альфа (α)

2) бета (β)

3) гамма (γ)

4) ню (v)

5) эта($\acute{\eta}$)

16. Какой из биометрических показателей определяет, какая доля вариации одного признака зависит от варьирования другого признака?

1) коэффициент детерминации

2) нормированное отклонение

3) коэффициент вариации

4) коэффициент корреляции

5) корреляционное отношение

17. Каким из приведенных символов обозначается коэффициент детерминации при наличии линейной связи?

1) r

2) r^2

3) cv

4) n

5) s

18. Каким из приведенных символов обозначается коэффициент детерминации при наличии нелинейной связи?

1) r

2) r^2

3) h^2

4) n

5) s

19. При каких показателях коэффициента детерминации можно судить о наличии сильной связи (когда около 50 % вариации одного признака зависит от вариации сопряженного признака)?

1) при $r < 0,5$

2) при $r = 0,5$

3) при $r = 0,6$

4) при $r > 0,7$

20. Корреляционная связь между двумя признаками при исключении влияния третьего, это…

1) линейная корреляция

2) частная корреляция

3) множественная корреляция

4) нелинейная корреляция

21. Установление степени связи одного признака с несколькими другими, вместе взятыми, это…

1) линейная корреляция

2) частная корреляция

3) множественная корреляция

4) нелинейная корреляция

22. Какой критерий обычно используют для измерения связи между качественными признаками?

1) коэффициент линейной корреляции

2) критерий согласия

3) критерий соответствия

4) критерий хи-квадрат

5) критерий Уилкоксона

23. Какой из видов анализа отвечает на вопрос о том, как именно изменяется один признак при определенном изменении другого?

1) математический

2) дисперсионный

3) анализ однофакторных комплексов

4) корреляционный

5) регрессионный

24. Какой буквой латинского алфавита в регрессионном анализе обычно обозначается зависимая переменная (функция)?

1) *a*

2) *d*

3) *f*

4) *p*

5) *y*

25. Какой буквой латинского алфавита в регрессионном анализе обычно обозначается независимая переменная (аргумент)?

1) *a*

2) *f*

3) *n*

4) *x*

5) *y*

26. С какого действия целесообразно начинать регрессионный анализ?

1) с построения эмпирических линий регрессий

2) с определения углового коэффициента

3) с расчета по методу наименьших квадратов

4) с решения системы нормальных уравнений

5) с расчета точек теоретической линии регрессии

27. О чем свидетельствует графическое изображение линии регрессии, которая пересекает под прямым углом ось *Y* и располагается параллельно оси *X*?

1) о наличии парной корреляции

2) о наличии сопряжения признаков

3) о наличии слабой связи

4) о полном отсутствии связи

5) о наличии полной связи

28. Коэффициент множественной корреляции $R = 0,7$. Чему равен коэффициент множественной детерминации?

1) 25 %

2) 36 %

3) 49 %

4) 64 %

5) 81 %

29. Какое уравнение служит уравнением линейной регрессии?

1) уравнение прямой линии

2) уравнение гиперболы

3) уравнение параболы 2-й степени

4) уравнение параболы 3-й степени

5) система нормальных уравнений

30. Какой буквой латинского алфавита обозначается в уравнении регрессии угловой коэффициент ($y = a + bx$)?

1) a

2) b

3) n

4) x

5) y

31. Как называется угловой коэффициент в уравнении линейной регрессии, отражающий пропорциональную зависимость между признаками?

1) детерминации

2) регрессии

3) вариации

4) корреляции

5) корреляционное отношение

32. В чем заключается биологический смысл коэффициентов регрессии?

1) дают оценку истинного значения признака

2) сравнивают средние двух независимых выборок

3) оценивают разность средних независимых выборок

4) оценивают меру изменения одного признака при определенном изменении другого

5) указывают границу предельным случайным отклонениям

33. Какие из приведенных коэффициентов характеризуют линейную зависимость каждого из признаков от другого?

1) детерминации

2) вариации

3) корреляционное отношение

4) корреляции

5) регрессии

34. Какой будет регрессия в случае, когда одинаковым приращениям одного признака сопутствуют неодинаковые приращения другого признака?

1) прямой (положительной)

2) обратной (отрицательной)

3) линейной

4) нелинейной

5) полной

35. В чем заключается сходство коэффициентов корреляции и регрессии?

1) оба изучают ошибку средней

2) оба дают оценку истинного значения признака

3) оба сравнивают средние двух независимых выборок

4) оба оценивают разность средних сопряженных выборок

5) оба свидетельствуют о наличии линейной связи

36. Какое из перечисленных уравнений является уравнением параболы второго порядка?

1) $x = (x_1 + x_2 + x_3 + \dots x_n) / n$

2) $c^2 = a^2 + b^2$

3) $y = a + bx + cx^2$

4) $m = (a + b) / 2$

5) $t = p_1 – p_2 / m$

37. Какое из перечисленных уравнений является уравнением гиперболы второго порядка?

1) $x = (x_1 + x_2 + x_3 + \ldots x_n) / n$

2) $c^2 = a^2 + b^2$

3) $y = a + bx + cx^2$

4) $y = a + b/x^2$

5) $t = p_1 – p_2 / m$

38. Определение какого понятия заключено в выражении: «построение по заданной функции другой, значения которой совпадают со значениями заданной функции в некотором числе точек»?

1) полином

2) интерполяция

3) парабола

4) гипербола

5) ряд динамики

39. С какой целью применяется интерполяция?

1) для описания случаев нелинейной регрессии

2) для моделирования биологического явления в целом

3) для биологического истолкования полученного результата

4) для построения кривых различной конфигурации

40. Какими методами изучается зависимость изменения одного признака от одновременного изменения нескольких?

1) методами корреляционного анализа

2) методами линейной регрессии

3) методами множественной регрессии

4) методами однофакторного дисперсионного анализа

5) методами двуфакторного дисперсионного анализа

41. С применением какого метода можно решить следующую задачу: зная высоту растения и длину листа, достаточно точно определить длину соцветия?

1) методами корреляционного анализа

2) методами линейной регрессии

3) методами однофакторного дисперсионного анализа

4) методами двуфакторного дисперсионного анализа

5) методами множественной регрессии

42. С какой целью осуществляется выравнивание ломаных линий эмпирических рядов?

1) для обеспечения наглядности изменчивости

2) для нахождения групповых средних

3) для демонстрации влияния второстепенных причин

4) для выявления основной тенденции вариации коррелирующих признаков

5) для визуализации срединных точек регрессии

43. К какому методу выравнивания следует отнести последовательное вычисление средних арифметических из трех соседних членов эмпирического ряда?

1) методу корреляционного анализа

2) методу линейной регрессии

3) методу дисперсионного анализа

4) методу скользящей средней

5) методу наименьших квадратов

44. Коэффициент множественной детерминации $R^2 = 0{,}25$. Чему равен коэффициент множественной корреляции?

1) 0,1

2) 0,2

3) 0,3

4) 0,4

5) 0,5

45. Дайте правильное продолжение выражения: «Изменение признаков во времени образует…»

1) вариационный ряд

2) эмпирический ряд

3) сглаженную кривую

4) ряд динамики

5) параболу

46. Какой из перечисленных факторов выступает в качестве независимой переменной при построении рядов динамики?

1) фактор времени

2) изменяющийся признак
3) угловой коэффициент
4) параметры уравнения
5) функция

47. С какого процесса следует начинать анализ рядов динамики?
1) с построения эмпирических линий регрессий
2) с выявления формы тренда
3) с расчетов по методу наименьших квадратов
4) с решения системы нормальных уравнений
5) с расчетов точек теоретической линии регрессии

48. Какой тип регрессии описывается уравнением Ферхюльста?
1) регрессия, выражаемая уравнением параболы
2) регрессия, выражаемая уравнением гиперболы 1-го порядка
3) регрессия, выражаемая уравнением гиперболы 2-го порядка
4) регрессия, выражаемая уравнением показательного типа
5) регрессия, выражаемая логистической кривой

49. Кем из перечисленных ученых были разработаны основы дисперсионного анализа?
1) Р. Фишер
2) А. Колмогоров
3) Н. Бейли
4) П. Чебышев
5) Ф. Гальтон

50. С помощью какого из перечисленных видов анализа можно оценить долю влияния нескольких факторов на общую изменчивость признака?
1) математический
2) дисперсионный
3) парный корреляционный
4) регрессионный

51. С помощью какого из перечисленных видов анализа можно выделить несколько ведущих факторов и исследовать именно их воздействие на изменчивость признака?
1) математический
2) корреляционный

3) регрессионный

4) дисперсионный

52. С помощью какого из перечисленных видов анализа можно установить степень влияния экологических факторов на выраженность тех или иных признаков в популяциях, т. е. проследить экологическую дифференциацию вида?

1) дисперсионный

2) корреляционный

3) регрессионный

4) математический

53. Задачей какого анализа является выявление той части общей изменчивости признака, которая обусловлена воздействием учитываемых факторов?

1) парный корреляционный

2) регрессионный

3) дисперсионный

4) математический

54. Коэффициент множественной детерминации R^2 = 25 %. Чему равен коэффициент множественной корреляции?

1) 0,1

2) 0,2

3) 0,3

4) 0,4

5) 0,5

55. Какому числу должна соответствовать сумма влияний, выраженных в долях, учтенных и неучтенных в однофакторном дисперсионном анализе факторов?

1) 0,001

2) 0,01

3) 0,05

4) 0,5

5) 1

56. Значение какого критерия в дисперсионном анализе рассчитывают как отношение факториальной дисперсии к остаточной?

1) критерий хи-квадрат
2) *F*-критерий Фишера
3) *T*-критерий Уилкоксона
4) критерий лямбда Колмогорова – Смирнова
5) критерий знаков

57. При дисперсионном анализе получили значение общей суммы квадратов отклонений вариант от средней арифметической, равное 1000, факториальной суммы квадратов 800 и остаточной (случайной) – 200. Укажите значение показателя силы влияния организованных в опыте факторов (по Н. А. Плохинскому).

1) 0,2
2) 0,4
3) 0,5
4) 0,6
5) 0,8

58. На какое количество компонент разлагается общая сумма квадратов в двухфакторном дисперсионном анализе несопряженных комплексов?

1) на 12
2) на 6
3) на 4
4) на 2

59. Определение какого термина содержит следующее выражение: «искусственно организуемый комплекс условий, в которых испытывают воздействие того или иного фактора на результативный признак»?

1) эксперимент
2) исследование
3) обучение
4) испытание
5) планирование

60. При дисперсионном анализе получили значение общей дисперсии равное 10, факториальной дисперсии – 6 и остаточной (случайной) – 2. Укажите правильное значение *F* – критерия Фишера.

1) 1
2) 2
3) 3
4) 4
5) 5

61. Временной ряд описывается уравнением регрессии $y = 5 + 0.01x$, где x – годы, y – температура в градусах по Цельсию. На сколько градусов изменяется температура за столетие?

1) на 0,5° С
2) на 0,8° С
3) на 1,0° С
4) на 1,2° С
5) на 2,0° С

62. При дисперсионном анализе получены следующие суммы квадратов отклонений значений вариантов от средней арифметической по комплексу: *Dy* (общая) = 100, *Dx* (факториальная) = 50, *Dz* (случайная) = 50. Определите показатель силы влияния организованных в опыте факторов по Плохинскому

1) 10 %
2) 20 %
3) 30 %
4) 40 %
5). 50 %

63. При дисперсионном анализе получены следующие суммы квадратов: *Dy* (общая) = 100, *Dx* (факториальная) = 50, *Dz* (случайная) = 50. Определите показатель силы влияния организованных в опыте факторов по Плохинскому

1) 0,1
2) 0,2
3) 0,3
4) 0,4
5) 0,5

64. При дисперсионном анализе получены следующие суммы квадратов: *Dy* (общая) = 100, *Dx* (факториальная) = 50, *Dz* (слу-

чайная) = 50, D_A (по фактору A) = 30, D_B (по фактору B) = 10, D_{AB} (взаимодействие факторов A и B) = 10.

Определите показатель силы влияния фактора A

1) 0,1
2) 0,2
3) 0,3
4) 0,4
5) 0,5

65. При дисперсионном анализе получены следующие суммы квадратов: D_y (общая) = 100, D_x (факториальная) = 50, D_z (случайная) = 50, D_A (по фактору A) = 30, D_B (по фактору B) = 10, D_{AB} (взаимодействие факторов A и B) = 10.

Определите показатель силы влияния фактора B

1) 0,1
2) 0,2
3) 0,3
4) 0,4
5) 0,5

66. При дисперсионном анализе получены следующие суммы квадратов: D_y (общая) = 100, D_x (факториальная) = 50, D_z (случайная) = 50, D_A (по фактору A) = 30, D_B (по фактору B) = 10, D_{AB} (взаимодействие факторов A и B) = 10. Определите показатель силы влияния взаимодействия факторов A и B

1) 0,1
2) 0,2
3) 0,3
4) 0,4
5) 0,5

67. При дисперсионном анализе получены следующие значения дисперсий (вариансы): общая $S^2_y = 2$, факториальная $S^2_x = 6$, остаточная $S^2_z = 1$. Найдите значение F-критерия

1) 1
2) 2
3) 3
4) 4

5) 5

6) 6

68. При дисперсионном анализе получены следующие значения дисперсий (вариансы): общая $S^2_y = 2$, факториальная $S^2_x = 6$, остаточная $S^2_z = 2$. Найдите значение F-критерия

1) 1

2) 2

3) 3

4) 4

5) 5

6) 6

69. При дисперсионном анализе получены следующие значения дисперсий (вариансы): общая $S^2_y = 2$, факториальная $S^2_x = 6$, остаточная $S^2_z = 3$. Найдите значение F-критерия

1) 1

2) 2

3) 3

4) 4

5) 5

6) 6

70. При дисперсионном анализе получены следующие значения дисперсий (вариансы): общая $S^2_y = 2$, факториальная $S^2_x = 6$, остаточная $S^2_z = 6$. Найдите значение F-критерия

1) 1

2) 2

3) 3

4) 4

5) 5

6) 6

71. При дисперсионном анализе получены следующие значения дисперсий (вариансы): общая $S^2_y = 5$, факториальная $S^2_x = 20$, остаточная $S^2_z = 5$, по фактору A $S^2_A = 10$, по фактору B $S^2_B = 20$, Взаимодействие факторов A и B $S^2_{AB} = 20$. Найдите значение F-критерия по фактору A

1) 1
2) 2
3) 3
4) 4
5) 5
6) 6

72. При дисперсионном анализе получены следующие значения дисперсий (вариансы): общая $S^2_y = 5$, факториальная $S^2_x = 20$, остаточная $S^2_z = 5$, по фактору A $S^2_A = 10$, по фактору B $S^2_B = 20$, Взаимодействие факторов A и B $S^2_{AB} = 20$. Найдите значение F-критерия по фактору B

1) 1
2) 2
3) 3
4) 4
5) 5
6) 6

73. При дисперсионном анализе получены следующие значения дисперсий (вариансы): общая $S^2_y = 5$, факториальная $S^2_x = 20$, остаточная $S^2_z = 5$, по фактору A $S^2_A = 10$, по фактору B $S^2_B = 20$, Взаимодействие факторов A и B $S^2_{AB} = 20$. Найдите значение F-критерия для оценки значимости взаимодействия факторов A и B

1) 1
2) 2
3) 3
4) 4
5) 5
6) 6

74. Коэффициент множественной детерминации $R^2 = 0{,}25$. Чему равен коэффициент множественной корреляции?

1) 0,1
2) 0,2
3) 0,3

4) 0,4

5) 0,5

75. Коэффициент множественной детерминации равен 25 %. Чему равен коэффициент множественной корреляции?

1) 0,1

2) 0,2

3) 0,3

4) 0,4

5) 0,5

76. При дисперсионном анализе получили значение общей дисперсии, равное 10, факториальной дисперсии – 6, остаточной (случайной) дисперсии – 2. Укажите правильное значение F-критерия Фишера.

1) 1

2) 2

3) 3

4) 4

5) 5

77. В какой компании была разработана система STATISTICA?

1) IBM

2) Intel

3) Microsoft

4) Windows

5) SPSS

78. При дисперсионном анализе получены следующие суммы квадратов: D_y (общая) = 100, D_x (факториальная) = 10, D_z (случайная) = 90. Определите показатель силы влияния организованных в опыте факторов по Плохинскому

1) 10 %

2) 20 %

3) 30 %

4) 40 %

5) 50 %

79. При дисперсионном анализе получены следующие суммы квадратов: *Dy* (общая) = 100, D_x (факториальная) = 10, D_z (случайная) = 90. Определите показатель силы влияния организованных в опыте факторов по Плохинскому

1) 0,1

2) 0,2

3) 0,3

4) 0,4

5) 0,5

80. При дисперсионном анализе получены следующие суммы квадратов: D_y (общая) = 100, D_x (факториальная) = 20, D_z (случайная) = 80. Определите показатель силы влияния организованных в опыте факторов по Плохинскому

1) 10 %

2) 20 %

3) 30 %

4) 40 %

5) 50 %

81. При дисперсионном анализе получены следующие суммы квадратов: D_y (общая) = 100, D_x (факториальная) = 20, D_z (случайная) = 80. Определите показатель силы влияния организованных в опыте факторов по Плохинскому

1) 0,1

2) 0,2

3) 0,3

4) 0,4

5) 0,5

82. При дисперсионном анализе получены следующие суммы квадратов: D_y (общая) = 100, D_x (факториальная) = 30, D_z (случайная) = 70. Определите показатель силы влияния организованных в опыте факторов по Плохинскому

1) 10 %
2) 20 %
3) 30 %
4) 40 %
5) 50 %

83. При дисперсионном анализе получены следующие суммы квадратов: D_y (общая) = 100, D_x (факториальная) = 30, D_z (случайная) = 70. Определите показатель силы влияния организованных в опыте факторов по Плохинскому

1) 0,1
2) 0,2
3) 0,3
4) 0,4
5) 0,5

84. При дисперсионном анализе получены следующие суммы квадратов: D_y (общая) = 100, D_x (факториальная) = 40, D_z (случайная) = 60. Определите показатель силы влияния организованных в опыте факторов по Плохинскому

1) 10 %
2) 20 %
3) 30 %
4) 40 %
5) 50 %

85. При дисперсионном анализе получены следующие суммы квадратов: D_y (общая) = 100, D_x (факториальная) = 40, D_z (случайная) = 60. Определите показатель силы влияния организованных в опыте факторов по Плохинскому

1) 0,1
2) 0,2
3) 0,3
4) 0,4
5) 0,5

86. При дисперсионном анализе получены следующие суммы квадратов: D_y (общая) = 100, D_x (факториальная) = 60, D_z (случайная) = 40, D_A (по фактору A) = 20, D_B (по фактору B) = 30, D_{AB} (взаимодействие факторов A и B) = 10. Определите показатель силы влияния фактора A

1) 0,1
2) 0,2
3) 0,3
4) 0,4
5) 0,5

87. При дисперсионном анализе получены следующие суммы квадратов: D_y (общая) = 100, D_x (факториальная) = 60, D_z (случайная) = 40, D_A (по фактору A) = 20, D_B (по фактору B) = 30, D_{AB} (взаимодействие факторов A и B) = 10. Определите показатель силы влияния фактора B

1) 0,1
2) 0,2
3) 0,3
4) 0,4
5) 0,5

88. При дисперсионном анализе получены следующие суммы квадратов: D_y (общая) = 100, D_x (факториальная) = 60, D_z (случайная) = 40, D_A (по фактору A) = 20, D_B (по фактору B) = 30, D_{AB} (взаимодействие факторов A и B) = 10. Определите показатель силы влияния взаимодействия факторов A и B.

1) 0,1
2) 0,2
3) 0,3
4) 0,4
5) 0,5
6) 0,6

89. При дисперсионном анализе получены следующие суммы квадратов: D_y (общая) = 100, D_x (факториальная) = 60, D_z (случайная) = 40, D_A (по фактору A) = 20, D_B (по фактору B) = 30, D_{AB} (вза-

имодействие факторов *A* и *B*) = 10. Определите показатель факторного влияния

1) 0,1
2) 0,2
3) 0,3
4) 0,4
5) 0,5
6) 0,6

90. Укажите, как в базах данных в системе STATISTICA называются столбцы (переменные)

1) variables
2) cases
3) characters
4) signs
5) files

91. Укажите, как в базах данных в системе STATISTICA называются строки

1) variables
2) cases
3) characters
4) signs
5) files

92. Какая команда при редактировании базы данных в системе STATISTICA означает удаление колонок (столбцов) или строк?

1) delete
2) copy
3) move

93. Какая команда при редактировании баз данных в системе STATISTICA означает копирование колонок (столбцов) или строк?

1) delete
2) copy
3) move

94. Какая команда при редактировании баз данных в системе STATISTICA означает перемещение колонок (столбцов) или строк?

1) delete
2) copy
3) move

95. Что нужно обозначить, если Вы хотите вычислить дисперсию совокупности?

1) Valid N
2) Mean
3) Standart Deviation
4) Varianse
5) Std. err. of mean

96. Что нужно обозначить, если Вы хотите вычислить среднее квадратичное отклонение?

1) Valid N
2) Mean
3) Standart Deviation
4) Varianse
5) Std. err. of mean

97. Что нужно обозначить, если Вы хотите вычислить среднее арифметическое значение?

1) Valid N
2) Mean
3) Standart Deviation
4) Varianse
5) Std. err. of mean

98. Что нужно обозначить, если Вы хотите вычислить статистическую ошибку среднего арифметического значения?

1) Valid N
2) Mean
3) Standart Deviation
4) Varianse
5) Std. err. of mean

99. Что нужно обозначить, если Вы хотите вычислить коэффициент асимметрии?

1) Skewness

2) Mean

3) Standart Deviation

4) Varianse

5) Kurtosis

100. Что нужно обозначить, если Вы хотите вычислить коэффициент эксцесса?

1) Skewness

2) Mean

3) Standart Deviation

4) Varianse

5) Kurtosis

101. При проведении множественного корреляционного анализа открывалось окно с приведенными ниже результатами. Укажите оценку уровня значимости (вероятности ошибки)

1) $R = 0{,}86997205$

2) $R = 0{,}75683137$

3) $F = 10{,}89449$

4) $df = 2{,}7$

5) $p = 0{,}007088$

102. При проведении множественного корреляционного анализа открывалось окно с приведенными ниже результатами. Укажите число степеней свободы

1) $R = 0{,}86997205$

2) $R = 0{,}75683137$

3) $F = 10{,}89449$

4) $df = 2{,}7$

5) $p = 0{,}007088$

103. При проведении множественного корреляционного анализа открывалось окно с приведенными ниже результатами. Укажите значение коэффициента множественной линейной корреляции

1) $R = 0{,}86997205$

2) $R = 0{,}75683137$

3) $F = 10{,}89449$

4) $df = 2{,}7$

5) $p = 0{,}007088$

104. При проведении множественного корреляционного анализа открывалось окно с приведенными ниже результатами. Укажите значение коэффициента множественной линейной детерминации

1) $R = 0{,}86997205$

2) $R? = 0{,}75683137$

3). $F = 10{,}89449$

4) $df = 2{,}7$

5) $p = 0{,}007088$

105. При проведении множественного корреляционного анализа открывалось окно с приведенными ниже результатами. Укажите значение параметрического критерия связи

1) $R = 0{,}86997205$

2) $R? = 0{,}75683137$

3) $F = 10{,}89449$

4) $df = 2{,}7$

5) $p = 0{,}007088$

106. При проведении линейного регрессионного анализа в окне результатов (p-level) появилось значение 0,000000. Какова значимость связи (регрессии)?

1) не значимая

2) значимая

3) сильно значимая

4) максимально значимая

107. Что означает в системе STATISTICA слово Intercept?

1) свободный член в уравнении

2) угловой коэффициент в уравнении регрессии

3) коэффициент регрессии

4) коэффициент корреляции

5) дисперсия

108. Сколько различных групп методов предполагается пользователем в модуле Time Series Analisis & Forecasting системы STATISTICA?

1) 1

2) 2

3) 3

4) 4

5) 5

6) 6

7) 7

109. Какая из перечисленных кнопок, стандартных для всех модулей системы STATISTICA, означает графики?

1) transformations

2) autocorrelations

3) plots

110. Какая из перечисленных кнопок, стандартных для всех модулей системы STATISTICA, означает преобразования?

1) transformations

2) autocorrelations

3) plots

111. Какая из перечисленных кнопок, стандартных для всех модулей системы STATISTICA, означает автокорреляции?

1) transformations

2) autocorrelations

3) plots

СПИСОК ЛИТЕРАТУРЫ

Основная литература

1. Боровиков В. П. STATISTICA. Искусство анализа данных на компьютере: Для профессионалов / В. П. Боровиков. – 2-е изд. – СПб. : Питер, 2003. – 688 с.

2. Боровиков В. П. Прогнозирование в системе STATISTICA в среде Windows: Основы теории и интенсивная практика на компьютере / В. П. Боровиков, Г. И. Ивченко. – М. : Финансы и статистика, 2006. – 368 с.

3. Буреева Н. Н. Многомерный статистический анализ с использованием ППП STATISTICA : учеб.-метод. материал по программе повышения квалификации / Н. Н. Буреева. – Нижний Новгород, 2007. – 112 с.

4. Бююль А. SPSS: искусство обработки информации. Platinum edition / А. Бююль, П. Цефель ; пер. с нем. – СПб. : Диа Софт ЮП, 2005. – 608 с.

5. Доспехов Б. П. Методика полевого опыта (с основами статистической обработки результатов исследований) / Б. П. Доспехов. – М. : Агропромиздат, 1985. – 351 с.

6. Категории земель в РФ (2015) [Электронный ресурс]. – Режим доступа : http://bigland.ru/o_kompanii/poleznye_stati/kategorii_-zemel_v_rf_2015.

7. Куприенко Н. В. Статистика. Анализ рядов динамики : учеб. пособие / Н. В. Куприенко, О. А. Пономарёва, Д. В. Тихонов. – СПб. : Изд-во Политехн. ун-та, 2009. – 204 с.

8. Лакин Г. Ф. Биометрия : учеб. пособие для биол. спец. вузов / Г. Ф. Лакин. – М. : Высш. шк., 1990. – 352 с.

9. Основы научных исследований : учеб. пособие / под ред. А. А. Лудченко. – Киев : Знання, 2000. – 114 с.

10. Нейронные сети. STATISTICA Neural Networks: Методология и технологии современного анализа данных / под ред. В. П. Бо-

ровикова. – 2-е изд., перераб. и доп. – М. : Горячая линия. – Телеком, 2008. – 392 с.

11. Олійник О. В. Тенденції та фактори економічної динаміки в аграрному секторі / О. В. Олійник // Вісник ХНАУ. Сер. Економіка і природокористування. – 2003. – № 4. – С. 21–32.

12. Олійник О. В. Циклічність у динаміці урожайності сільськогосподарських культур / О. В. Олійник // Економіка АПК. – 2003. – № 3. – С. 52–57.

13. П'ятницька-Позднякова. Основи наукових досліджень у вищій школі : навч. посіб. / П'ятницька-Позднякова. – Киев : 2003. – 116 с.

14. Плохинский Н. А. Биометрия / Н. А. Плохинский. – М. : Изд-во МГУ, 1970. – 363 с.

15. Компьютеризация агрономических и биологических расчетов / И. Д. Соколов, П. В. Шелихов, С. Ю. Наумов, Е. И. Сыч. – Луганск : Элтон-2, 2001. – 133 с.

16. Соколова Е. И. Новый метод оценки взаимодействия генов в количественной генетике растений / Е. И. Соколова // Збірн. наук. праць Луганського НАУ. Сер. Біологічні науки. – Луганськ : Елтон-2. – 2003. – № 22 (34). – С. 65–71.

17. Тарасова Ю. В. Молюски роду *Theodoxus* (Mollusca: *Gastropoda: Pectinibranchia: Neritidae*) України : дис. ... канд. біол. наук: 03.00.08 / Ю. В. Тарасова. – Житомир, 2011. – 155 с.

18. Урбах В. Ю. Биометрические методы / В. Ю. Урбах. – М. : Изд-во МГУ, 1964. – 415 с.

19. Федеральная служба государственной статистики [Электронный ресурс]. – Режим доступа : http://www.gks.ru/.

20. Халафян А. А. STATISTICA 6. Статистический анализ данных : учебник / А. А. Халафян. – 3-е изд. – М. : Бином-Пресс, 2007. – 512 с.

21. Комп'ютерні методи в сільському господарстві та біології / О. М. Царенко, Ю. А. Злобін, В. Г. Скляр [та інш.]. – Суми : Університетська книга, 2000. – 203 с.

22. Шмидт В. М. Математические методы в ботанике / В. М. Шмидт. – Л. : Изд-во ЛГУ, 1984. – 288 с.

Дополнительная литература

1. Биометрическая оценка высокопродуктивных с комплексной устойчивостью гибридных форм селекции ВНИИВиВ «Магарач» / П. Я. Голодрига, Л. П. Трошин, В. Т. Усатов [и др.] // Генетика и селекция винограда на иммунитет : АН УССР. УОГиС им. Н. И. Вавилова. ВНИИВиВ «Магарач». – Киев, 1978. – С. 186–193.

2. Голодрига П. Я. Аналіз фенотипічних кореляцій кількісних ознак винограду / П. Я. Голодрига, Л. П. Трошин // Вісн. с.-г. науки. – 1978. – № 9. – С. 49–54.

3. Луценко Е. В. Решение задач ампелографии с применением АСК-анализа изображений листьев по их внешним контурам (обобщение, абстрагирование, классификация и идентификация) / Е. В. Луценко, Д. К. Бандык, Л. П. Трошин // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета (Научный журнал КубГАУ) [Электронный ресурс]. – Краснодар : КубГАУ, 2015. – № 08(112). – С. 862 – 910. – IDA [article ID]: 1121508064. – Режим доступа : http://ej.kubagro.ru/2015/08/pdf/64.pdf.

4. Луценко Е. В. Количественное измерение сходства-различия клонов винограда по контурам листьев с применением АСК-анализа и системы «Эйдос» / Е. В. Луценко, Л. П. Трошин, Д. К. Бандык // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета (Научный журнал КубГАУ) [Электронный ресурс]. – Краснодар : КубГАУ, 2016. – № 02(116). – IDA [article ID]: 1161602077. – Режим доступа : http://ej.kubagro.ru/2016/02/pdf/77.pdf.

5. Островерхов В. О. Ковариационный анализ в селекционно-генетических исследованиях / В. О. Островерхов, Л. П. Трошин // Теория отбора в популяциях растений / АН СССР. Ин-т цитологии и генетики. – Новосибирск, 1976. – С. 94–103.

6. Островерхов В.О. Методические рекомендации по оценке стабильности количественных признаков у сортов винограда / В. О. Островерхов, Л. П. Трошин. – Ялта : ВНИИВиПП «Магарач», 1986. – 86 с.

7. Островерхов В.О. Статистический анализ в генетике и селекции винограда / В. О. Островерхов, Л. П. Трошин // Энциклопедия виноградарства. – Кишинев, 1987. – Т. 3. – С. 170.

8. Островерхов В. О. Методика вычисления несмещенных оценок стабильности признаков сортов сельскохозяйственных растений и прогнозирования эффективности сортосмены в изменяющихся условиях / В. О. Островерхов, Л. П. Трошин, Ю. В. Яловенко. – Ялта : ВНИИВиПП «Магарач», 1985. – Деп. в ВНИИТЭИСХ 27.05.1985, № 245 ВС-85. – 18 с.

9. Стародворов Г. А. Математическая модель зависимости урожайности озимой пшеницы от некоторых климатических факторов / Г. А. Стародворов // Серія: С.-г. науки. – Луганськ : Вид-во ЛНАУ, 2007а. – № 77 (100). – С. 100–104.

10. Стародворов Г. О. Залежність мінливості врожайності озимої пшениці від температури повітря та опадів / Г. О. Стародворов, І. Д. Соколов, О. А. Мостовий // Вісник Дніпропетровського університету. Серія: Біологія, екологія. – Дніпропетровськ : Вид-во ДНУ. – 2006. – № 3. – С. 164–169.

11. Стародворов Г. О. Зв'язок природних агроекологічних чинників з урожайністю основних сільскогосподарських культур / Г. О. Стародворов // Збірник наук. праць Уманського ДАУ. – 2007б. – № 64. – С. 44–48.

12. Трошин Л. П. Взаимодействие генотип-среда / Л. П. Трошин. – С. 52–57.

13. Трошин Л. П. Ковариационный анализ / Л. П. Трошин // Генет. анализ количеств. и качеств. признаков с помощью мат.-стат. методов. – М. : ВНИИТЭИСХ, 1973. – С. 34–40.

14. Трошин Л. П. Статистические методы в селекции винограда / Л. П. Трошин // Науч.-техн. прогресс в виноградарстве и виноделии : тез. докл. – Кишинев, 1980. – Ч. 1. – С. 87–88.

15. Трошин Л. П. Биометрический анализ взаимосвязи урожайности и сахаристости ягод реципрокных F_1-популяций винограда / Л. П. Трошин // Пробл. эволюц. и популяц. генетики. – Махачкала, 1978. – С. 42–48.

16. Трошин Л. П. Коррелятивная изменчивость количественных признаков винограда / Л. П. Трошин, П. Я. Голодрига // III съезд генетиков и селекционеров Украины. Частная генетика. – Киев, 1976. – Ч. 2. – С. 134.

17. Трошин Л. П. Методические рекомендации по клоновой селекции винограда на продуктивность / Л. П. Трошин, Л. А. Животовский. – Ялта : ВНИИВиПП «Магарач» ; ин-т общей генетики им. Н. И. Вавилова, 1987. – 36 с.

18. Трошин Л. П. Компьютеризация селекционного процесса винограда / Л. П. Трошин, А. В. Смиряев // Виноградарство и виноделие СССР. – 1990. – № 6 (9). – С. 28–35.

19. Трошин Л. П. Биометрический анализ генофонда винограда / Л. П. Трошин, Ю. К. Фёдоров. – Ялта : ВНИИВиПП «Магарач», 1988. – 90 с.

ПРИЛОЖЕНИЯ

Приложение А

Таблица А1 – Стандартные значения *z*-критерия знаков для двух уровней значимости (0,05; 0,01) при количестве сопряженных пар значений (объеме выборки) *n*

n	*p*		*n*	*p*		*n*	*p*		*n*	*p*	
	0,05	0,01		0,05	0,01		0,05	0,01		0,05	0,01
6	6	–	*30*	21	23	*54*	35	37	*78*	49	51
7	7	–	*31*	22	24	*55*	36	38	*79*	49	52
8	8	8	*32*	23	24	*56*	36	39	*80*	50	52
9	8	9	*33*	23	25	*57*	37	39	*81*	50	53
10	9	10	*34*	24	25	*58*	37	40	*82*	51	54
11	10	11	*35*	24	26	*59*	38	40	*83*	51	54
12	10	11	*36*	25	27	*60*	39	41	*84*	52	55
13	11	12	*37*	25	27	*61*	39	41	*85*	53	55
14	12	13	*38*	26	28	*62*	40	42	*86*	53	56
15	12	13	*39*	27	28	*63*	40	43	*87*	54	56
16	13	14	*40*	27	29	*64*	41	43	*88*	54	57
17	13	15	*41*	28	30	*65*	41	44	*89*	55	58
18	14	15	*42*	28	30	*66*	42	44	*90*	55	58
19	15	16	*43*	29	31	*67*	42	45	*91*	56	59
20	15	17	*44*	29	31	*68*	43	46	*92*	56	59
21	16	17	*45*	30	32	*69*	44	46	*93*	57	60
22	17	18	*46*	31	33	*70*	44	47	*94*	57	60
23	17	19	*47*	31	33	*71*	45	47	*95*	58	61
24	18	19	*48*	32	34	*72*	45	48	*96*	59	62
25	18	20	*49*	32	34	*73*	46	48	*97*	59	62
26	19	20	*50*	33	35	*74*	46	49	*98*	60	63
27	20	21	*51*	33	36	*75*	47	50	*99*	60	63
28	20	22	*52*	34	36	*76*	48	50	*100*	61	64
29	21	22	*53*	35	37	*77*	48	51	–	–	–

Приложение Б

Таблица Б1 – Стандартные значения парного *T*- критерия Уилкоксона для двух уровней значимости: 0,05 и 0,01 (двусторонний критерий)

Число парных наблюдений, *n*	Уровни значимости, *p*		Число парных наблюдений, *n*	Уровни значимости, *p*	
	0,05	0,01		0,05	0,01
6	1	–	16	31	21
7	3	–	17	36	24
8	5	1	18	41	29
9	7	3	19	47	33
10	9	4	20	53	39
11	12	6	21	60	44
12	15	8	22	67	50
13	18	11	23	74	56
14	22	14	24	82	62
15	26	17	25	90	69

Примечание. Для $n > 25$ критические значения *T*- критерия можно определить по формуле:

$$T_{st} = \frac{n(n+1)}{4} - t\sqrt{\frac{n(n+1)(2n+1)}{24}},$$

где n – число парных наблюдений; t зависит от принятого уровня значимости, т. е. $t_{0.05} = 1{,}96$ и $t_{0.01} = 2{,}58$.

Приложение В

Таблица В1 – Стандартные значения критерия χ^2 для трех уровней значимости (0,05; 0,01; 0,001) и чисел степеней свободы k

k	p			k	p		
	0,05	0,01	0,001		0,05	0,01	0,001
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1	3,8	6,6	10,8	51	68,7	77,4	88,0
2	6,0	9,2	13,8	52	69,8	78,6	89,3
3	7,8	11,3	16,3	53	71,0	79,8	90,6
4	9,5	13,3	18,5	54	72,2	81,1	91,9
5	11,1	15,1	20,5	55	73,3	82,3	93,2
6	12,6	16,8	22,5	56	74,5	83,5	94,5
7	14,1	18,5	24,3	57	75,6	84,7	95,8
8	15,5	20,1	26,1	58	76,8	86,0	97,0
9	16,9	21,7	27,9	59	77,9	87,2	98,3
10	18,3	23,2	29,6	60	79,1	88,4	99,6
11	19,7	24,7	31,3	61	80,2	89,6	100,9
12	21,0	26,2	32,9	62	81,4	90,8	102,2
13	22,4	27,7	34,5	63	82,5	92,0	103,4
14	23,7	29,1	36,1	64	83,7	93,2	104,7
15	25,0	30,6	37,7	65	84,8	94,4	106,0
16	26,3	32,0	39,2	66	86,0	95,6	107,3
17	27,6	33,4	40,8	67	87,1	96,8	108,5
18	28,9	34,8	42,3	68	88,2	98,0	109,8
19	30,1	36,2	43,8	69	89,4	99,2	111,1
20	31,4	37,6	45,3	70	90,5	100,4	112,3
21	32,7	38,9	46,8	71	91,7	101,6	113,6
22	33,9	40,3	48,3	72	92,8	102,8	114,8
23	35,2	41,6	49,7	73	93,9	104,0	116,1
24	36,4	43,0	51,2	74	95,1	105,2	117,4
25	37,6	44,3	52,6	75	96,2	106,4	118,6
26	38,9	45,6	54,0	76	97,4	107,6	119,8
27	40,1	47,0	55,5	77	98,5	108,8	121,1
28	41,3	48,3	56,9	78	99,6	110,0	122,4
29	42,6	49,6	58,3	79	100,8	111,1	123,6
30	43,8	50,9	59,7	80	101,9	112,3	124,8
31	44,9	52,2	61,1	81	103,0	113,5	126,1
32	46,2	53,5	62,5	82	104,1	114,7	127,3
33	47,4	54,8	63,9	83	105,3	115,9	128,6
34	48,6	56,1	65,2	84	106,4	117,1	129,8
35	49,8	57,3	66,6	85	107,5	118,2	131,0
36	51,0	58,6	68,0	86	108,6	119,4	132,3
37	52,2	59,9	69,4	87	109,8	120,6	133,5
38	53,4	61,2	70,7	88	110,9	121,8	134,7

Продолжение таблицы В1

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
39	54,6	62,4	72,1	89	112,0	122,9	136,0
40	55,8	63,7	73,4	90	113,1	124,1	137,2
41	56,9	65,0	74,7	91	114,3	125,3	138,4
42	58,1	66,2	76,1	92	115,4	126,5	139,7
43	59,3	67,5	77,4	93	116,5	127,6	140,9
44	60,5	68,7	78,8	94	117,6	128,8	142,1
45	61,7	70,0	80,1	95	118,8	130,0	143,3
46	62,8	71,2	81,4	96	119,9	131,1	144,6
47	64,0	72,4	82,7	97	121,0	132,3	145,8
48	65,2	73,7	84,0	98	122,1	133,5	147,0
49	66,3	74,9	85,4	99	123,2	134,6	148,2
50	67,5	76,2	86,7	100	124,3	135,8	149,4

Приложение Г

Таблица Г1 – Стандартные значения *t*-критерия Стьюдента (t_{st}) для трех уровней значимости (0,05; 0,01; 0,001)

Число степеней свободы *k*	*p*			Число степеней свободы *k*	*p*		
	0,05	0,01	0,001		0,05	0,01	0,001
1	12,7	63,7	64,6	18	2,1	2,9	3,9
2	4,3	9,9	31,6	19	2,1	2,9	3,9
3	3,2	5,8	12,9	20	2,1	2,8	3,8
4	2,8	4,6	8,6	21	2,1	2,8	3,8
5	2,6	4,0	6,9	22	2,1	2,8	3,8
6	2,4	3,7	6,0	23	2,1	2,8	3,8
7	2,4	3,5	5,4	24	2,1	2,8	3,8
8	2,3	3,4	5,0	25	2,1	2,8	3,7
9	2,3	3,2	4,8	26	2,1	2,8	3,7
10	2,2	3,2	4,6	27	2,0	2,8	3,7
11	2,2	3,1	4,4	28	2,0	2,8	3,7
12	2,2	3,0	4,3	29	2,0	2,8	3,7
13	2,2	3,0	4,2	30	2,0	2,8	3,6
14	2,1	3,0	4,1	40	2,0	2,7	3,6
15	2,1	3,0	4,1	60	2,0	2,7	3,5
16	2,1	2,9	4,0	120	2,0	2,6	3,4
17	2,1	2,9	4,0	∞	2,0	2,6	3,3

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

ОГЛАВЛЕНИЕ

У ч е б н о е и з д а н и е

Соколов Иван Дмитриевич, **Соколова** Елена Ивановна,
Трошин Леонид Петрович и др.

БИОМЕТРИЯ

Учебник

В авторской редакции

Подписано в печать 18.04.2018. Формат 60 × 84 $^{1}/_{16}$.
Усл. печ. л. – 9,4. Уч.-изд. л. – 7,3.
Тираж 100 экз. Заказ №

Типография Кубанского государственного аграрного университета.
350044, г. Краснодар, ул. Калинина, 13