Министерство сельского хозяйства Российской Федерации

**Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«Саратовский государственный аграрный университет
имени Н. И. Вавилова»**

СТАТИСТИЧЕСКИЕ И БИОМЕТРИЧЕСКИЕ МЕТОДЫ В СЕЛЕКЦИИ И СЕМЕНОВОДСТВЕ РАСТЕНИЙ

краткий курс лекций

для аспирантов II курса

Направление подготовки
35.06.01 Сельское хозяйство

Профиль подготовки
Селекция и семеноводство сельскохозяйственных растений

Саратов 2014

УДК 63
ББК 41.3

Рецензенты:

Доктор биологических наук, профессор,
главный научный сотрудник ГНУ НИИСХ Юго-Востока
В.А. Крупнов

Кандидат биологических наук, доцент кафедры «Растениеводство, селекция и генетика»
ФГБОУ ВПО «Саратовский ГАУ»
Л.Г. Курасова

Статистические и биометрические методы в селекции и семеноводстве растений: краткий курс лекций для аспирантов 2 курса направления подготовки 35.06.01 Сельское хозяйство / Сост.: Ю.В. Лобачев // ФГБОУ ВПО «Саратовский ГАУ».- Саратов, 2014.- 48 с.

Краткий курс лекций по дисциплине «Статистические и биометрические методы в селекции и семеноводстве растений» составлен в соответствие с рабочей программой дисциплины и предназначен для аспирантов направления подготовки 35.06.01 Сельское хозяйство. Краткий курс лекций содержит теоретический материал по основным методам исследований в селекции и семеноводстве растений. Направлен на формирование у аспирантов навыков владения методами статистических и биометрических анализов, применяемых в селекции и семеноводстве сельскохозяйственных растений.

УДК 63
ББК 41.3

Введение

Краткий курс лекций по дисциплине «Статистические и биометрические методы в селекции и семеноводстве растений» составлен в соответствие с рабочей программой дисциплины и предназначен для аспирантов направления подготовки 35.06.01 Сельское хозяйство. Краткий курс лекций содержит теоретический материал по основным методам исследований в селекции и семеноводстве растений. Направлен на формирование у аспирантов навыков владения методами статистических и биометрических анализов, применяемых в селекции и семеноводстве сельскохозяйственных растений.

Краткий курс лекций по дисциплине «Статистические и биометрические методы в селекции и семеноводстве растений» предназначен для аспирантов по направлению подготовки 35.06.01 Сельское хозяйство. Курс нацелен на формирование у аспирантов универсальных компетенций: «способностью к критическому анализу и оценке современных научных достижений, генерированию новых идей при решении исследовательских и практических задач, в том числе в междисциплинарных областях» (УК-1); «способностью проектировать и осуществлять комплексные исследования, в том числе междисциплинарные, на основе целостного системного научного мировоззрения с использованием знаний в области истории и философии науки» (УК-2); «готовностью участвовать в работе российских и международных исследовательских коллективов по решению научных и научно-образовательных задач» (УК-3); общепрофессиональных компетенций: «владением методологией теоретических и экспериментальных исследований в области сельского хозяйства, агрономии, защиты растений, селекции и генетики сельскохозяйственных культур, почвоведения, агрохимии, ландшафтного обустройства территорий, технологий производства сельскохозяйственной продукции» (ОПК-1); «владением культурой научного исследования в области сельского хозяйства, агрономии, защиты растений, селекции и генетики сельскохозяйственных культур, почвоведения, агрохимии, ландшафтного обустройства территорий, технологий производства сельскохозяйственной продукции, в том числе с использованием новейших информационно-коммуникационных технологий» (ОПК-2); «способностью к разработке новых методов исследования и их применению в области сельского хозяйства, агрономии, защиты растений, селекции и генетики сельскохозяйственных культур, почвоведения, агрохимии, ландшафтного обустройства территорий, технологий производства сельскохозяйственной продукции с учетом соблюдения авторских прав» (ОПК-3); профессиональных компетенций: «готовностью использовать методы выведения сортов и гибридов культурных растений для получения их высококачественных семян и посадочного материала» (ПК-1); «способностью планировать эксперименты и анализировать результаты научно-исследовательской деятельности в области селекции и семеноводства сельскохозяйственных растений» (ПК-5).

Лекция 1

ПЛАНИРОВАНИЕ ПОЛЕВОГО ЭКСПЕРИМЕНТА

1.1 Основные приемы научного исследования

В селекции и семеноводстве используют два основных приема научного исследования – наблюдение и эксперимент.

Наблюдение – это количественная или качественная регистрация интересующих исследователя сторон развития явления, констатация наличия того или иного его состояния, признака или свойства.

Эксперимент (опыт) – это такое изучение, при котором исследователь искусственно вызывает явления или изменяет условия так, чтобы лучше выяснить сущность явления, происхождение, причинность и взаимосвязь предметов и явлений. Эксперимент – ведущий метод исследования. Важной чертой любого эксперимента является его воспроизводимость и количественная оценка эффектов изучаемых в опыте вариантов. Один или несколько вариантов, с которыми сравнивают опытные варианты называют контролем или стандартом. Совокупность опытных и контрольных вариантов составляет схему эксперимента. Варианты бывают качественные (сорта, культуры, способы посева или обработки почвы и т.д.) и количественные (нормы высева, дозы удобрений или пестицидов и т.д.).

В селекции и семеноводстве применяют три типа экспериментов: лабораторный, вегетационный и полевой.

Лабораторный эксперимент – исследование, осуществляемое в лабораторной обстановке с целью установления действия и взаимодействия факторов на изучаемые объекты.

Вегетационный эксперимент – исследование, осуществляемое в контролируемых условиях – вегетационных домиках, теплицах, оранжереях, климатических камерах, фитотронах и других сооружениях с целью установления различий между вариантами опыта и количественной оценкой действия и взаимодействия изучаемых факторов на урожай растений и его качество.

Полевой эксперимент – исследование, осуществляемое в полевой обстановке на специально выделенном участке. Основной задачей полевого эксперимента является установление различий между изучаемыми вариантами опыта, количественная оценка действия факторов жизни, условий или приемов возделывания на урожай растений и его качество. Полевой эксперимент является ведущим методом исследований в селекции и семеноводстве. Только по результатам полевого эксперимента можно сделать объективные выводы и рекомендации производству.

1.2 Основные требования к проведению полевого эксперимента

Ценность полевого эксперимента зависит от соблюдения следующих требований:

1). Типичность опыта;
2). Соблюдение принципа единственного различия;
3). Проведение опыта на специально выделенном участке;
4). Достоверность опыта по существу.

Под типичностью (репрезентативностью) полевого эксперимента понимают соответствие условий его проведения почвенно-климатическим (природным) и агротехническим (антропогенным) условиям данного региона (зоны).

Соблюдение принципа единственного различия предполагает соблюдение в эксперименте единства всех условий, кроме одного – изучаемого.

Полевой эксперимент проводят на специально выбранном участке с хорошо известной историей, однородным качественным составом почвы, опытопригодным рельефом местности.

Под достоверностью опыта по существу понимают логически правильно построенную схему и методику проведения опыта, соответствие их поставленным целям и задачам, правильный выбор объекта и условий проведения эксперимента.

При проведении эксперимента исследователь встречается с тремя видами ошибок – случайными, систематическими и грубыми. Ошибка – это расхождение между результатами выборочного наблюдения и истинным значением измеряемой величины. Случайные ошибки – это ошибки, возникающие под действием очень большого числа факторов, эффекты действия которых столь незначительны, что их нельзя выделить и учесть в отдельности. Характерной особенностью случайных ошибок является их тенденция взаимно погашаться в результате приблизительно одинаковой вероятности положительного или отрицательного влияния на изучаемый фактор. Систематические ошибки – искажают измеряемую величину изучаемого фактора в сторону или только увеличения или только уменьшения за счет вполне определенной причины. Систематические ошибки не имеют свойства взаимного погашения, они входят в показатели отдельных наблюдений и средних значений признака. Грубые ошибки связаны, в основном, с нарушением требований к полевому эксперименту.

Для анализа результатов эксперимента и статистической обработки полученных данных нужно использовать лишь те результаты, которые не содержат грубых и систематически односторонних ошибок.

Каждый полевой эксперимент характеризуется определенным числом вариантов, повторностью, формой, направлением и площадью делянок, системой размещения вариантов и повторений, методами учета и статистической обработки изучаемых факторов, организацией опыта во времени.

В зависимости от условий эксперимента делянки применяют удлиненной (1), прямоугольной (2) или квадратной (3) формы:

1 2 3

В селекционно-семеноводческих полевых экспериментах обычно используют заранее определенное число вариантов (обычно 12-16), 4-6-кратную повторность, площадь делянок от 1 до 50 м^2.

1.3 Методы расположения полевого опыта

Полевые опыты располагают на земельном участке методом организованных повторений, методом неорганизованных повторений (полной рендомизацией), методом латинского квадрата, методом латинского прямоугольника, методом расщепленных делянок.

Метод организованных повторений – делянки с полным набором всех изучаемых вариантов объединяют территориально в компактную группу, составляя организованное повторение.

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

Метод неорганизованных повторений (полная рендомизация) – изучаемые варианты размещены в пределах повторений случайно:

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

Наиболее распространенный в мировой практике метод размещения вариантов по делянкам – метод рендомизированных повторений. Варианты размещаются в каждом повторении в случайном порядке (жребий, таблица случайных чисел, генератор случайных чисел).

При методе латинского квадрата земельный участок квадратной или прямоугольной формы разбивают в горизонтальном и вертикальном направлении на столько рядов и столбцов, сколько вариантов в опыте. Любой ряд и столбец включают полный набор изучаемых вариантов, которые размещают на делянках квадратной или прямоугольной формы:

1	2	3	4
2	3	4	1
4	1	2	3
3	4	1	2

Ни один из вариантов не повторяется дважды ни в строке, ни в столбце. Общее число делянок в опыте всегда равно квадрату числа вариантов (4^2=16). Недостаток метода – требование равенства числа вариантов числу повторений.

При большом числе вариантов используют метод латинского прямоугольника. В этом случае число вариантов должно быть кратным числу повторений. Число вариантов должно без остатка делиться на число повторностей. Частное от деления дает число делянок, на которое необходимо разделить столбец соответствующего латинского квадрата. Варианты по делянкам рендомизируют так, чтобы ряд и столбец имели полный набор всех вариантов:

1	2	3	4	5	6
5	4	1	6	3	2
3	6	5	2	1	4

Иногда в селекции или семеноводстве используют размещение делянок по методу расщепленных делянок. Делянки первого порядка расщепляют в вертикальном или горизонтальном направлении на делянки второго порядка, которые при необходимости делят на делянки третьего порядка и т.д.:

A_1				A_2				A_3			
B_1		B_2		B_3		B_4		B_5		B_6	
C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}

А – способ обработки почвы, В – доза удобрений, С – сорт.

До закладки полевого опыта исследователь в зависимости от целей и задач должен решить следующие вопросы:

1. какие наблюдения, учеты и анализы включить в программу эксперимента;
2. в какие сроки проводить наблюдения и учеты;
3. определить оптимальный объем выборок (проб);
4. обеспечить представительность отбираемых выборок;
5. выбрать опытопригодный участок поля;
6. выбрать схему размещения изучаемых вариантов на делянках (число, форма, площадь, направления).

После проведения подготовительных работ необходимо нанести расположение изучаемых вариантов на схематический план, где указаны точные размеры поля, повторений, делянок, и номеров изучаемых вариантов по делянкам. По схематическому плану затем размещают опыт в натуре, используя при этом теодолит или экер, мерную ленту, шнур, колышки и т.д.

При учете урожая за несколько дней до уборки осматривают посевы и при необходимости делают выключки. Под выключкой понимают часть учетной делянки, исключенную из учета вследствие случайных повреждений или ошибок, допущенных во время работы:

1) повреждения, вызванные стихийными явлениями природы;
2) случайные повреждения (потрава скотом, птицей, грызунами и т.д.);
3) ошибки при закладке опыта или во время проведения опыта (травма растений при культивации и т.д.).

Уборку всех изучаемых вариантов проводят одним способом в один день. У зерновых культур бункерный урожай с каждой делянки пересчитывают на 14%-ю влажность и 100%-ю чистоту по формуле:

$$X = \frac{Y \times (100 - B) \times (100 - C)}{(100 - B_1) \times 100}$$ где

X – урожай при 14%-й влажности (т/га),
Y – урожай без поправки на влажность (т/га),
В – влажность зерна при взвешивании (%),
B_1 – стандартная влажность (%),
С – засоренность зерна (%).

Первичная обработка результатов наблюдений, учетов и анализов включает в себя:

1) агрономический анализ полученных данных;
2) первичную цифровую обработку данных;
3) статистическую обработку результатов эксперимента.

Агрономический анализ включает в себя анализ методики и техники проведения полевого эксперимента, проверки записей по первоисточникам (полевые журналы), устранение описок и неточностей. Бессмысленно обрабатывать методически неправильно проведенные опыты.

Цифровая обработка включает пересчет урожая с делянки на 1 га, приведение урожая к стандартной влажности и засоренности, составление матриц для статистической обработки.

Результаты полевого эксперимента должны быть обязательно обработаны тем или иным статистическим методом.

Вопросы для самоконтроля

1.Основные типы экспериментов, применяемые в селекции.
2. Основные требования, предъявляемые к проведению полевого эксперимента.
3. Методы расположения полевого опыта.
4. Что включает в себя агрономический анализ.

СПИСОК ЛИТЕРАТУРЫ

Основная

1. Основы научных исследований в растениеводстве и селекции: Учебное пособие / А.Ф. Дружкин, Ю.В. Лобачев, Л.П. Шевцова, З.Д. Ляшенко // ФГБОУ ВПО «Саратовский ГАУ». – Саратов, 2013. – 264 с.

Дополнительная

1. **Доспехов Б.А.** Методика полевого опыта (с основами статистической обработки результатов исследований) – 5-е издание, доп. и перераб. – М.: Агропромиздат, 1985. – 351 с.
2. **Смиряев А.В., Мартынов С.П., Кильчевский А.В.** Биометрия в генетике и селекции растений. – М.: издательство МСХА, 1992. – 269 с.
3**. Рокицкий П.Ф.** Введение в статистическую генетику. – Минск: Вышейшая школа, 1978. – 442 с.

Лекция 2

ЗАДАЧИ И МЕТОДЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

2.1 Задачи статистической обработки

Главная задача статистической обработки экспериментальных данных – найти такие показатели, которые характеризуют особенности изучаемых совокупностей, и сравнить их между собой.

Статистическая гипотеза – это научное предположение о статистических законах распределения случайных величин, которые можно проверить на основе выборки.

В большинстве случаев задача сводится к проверке гипотезы об отсутствии реального различия между фактическими и теоретически ожидаемыми наблюдениями. Такую гипотезу называют нулевой гипотезой и обозначают $H_0 : d = 0$.

Если в результате проверки H_0, различия между фактическими и теоретическими показателями близки к 0 или находятся в области допустимых значений, то H_0 не отвергается, т.е. она принимается. Принятие H_0 означает, что между фактическими и теоретическими распределениями предполагается совпадение. Отбрасывание (отвержение) гипотезы означает, что эмпирические данные несовместимы с H_0, а верна какая-то другая противоположная (альтернативная) гипотеза.

В сельскохозяйственных и биологических исследованиях объектом изучения является чаще всего совокупность предметов и явлений. Поэтому всякое множественное явление, например, растений, их частей, делянок, в полевом опыте называют статистической совокупностью.

Предметы и явления в каждой совокупности строго индивидуальны. Они отличаются друг от друга рядом признаков, например, массой, объемом и т.д. Каждый признак имеет разную степень выраженности, то есть этот признак изменяется, варьирует в определенных пределах изучаемого признака.

Отсюда свойство предметов отличаться друг от друга даже в однородных совокупностях называется изменчивостью или варьированием, случайностью.

Варьирование значения признака у отдельных представителей совокупности возникает главным образом под влиянием комплекса внешних условий, обуславливающих случайные приемы, а также в результате того, что растения даже одного и того же сорта отличаются своей наследственностью.

Известно, что не всегда возможно исследовать по тому или иному признаку все особи, всю совокупность по двум причинам:

1) совокупность очень велика по объёму;
2) при изучении того или иного признака генеральной совокупности приходится уничтожать всю эту группу, совокупность.

Поэтому прибегают к изучению части совокупности, по которой делают обобщение о всей генеральной совокупности.

Таким образом, *генеральная совокупность* – это вся совокупность особей, которая подвергается изучению, или вся группа объектов, подлежащих изучению.

Выборочная совокупность, выборка – часть объектов совокупности, или выборка из генеральной совокупности, или та часть объектов, которая поставлена под контроль, под проверку.

Сущность выборочного метода заключается в том, чтобы по малой пробе получить сведения и сделать умозаключение о всей генеральной совокупности. Здесь очень важно правильно выбрать объем выборки.

Объем выборки – это число элементов в генеральной совокупности и выборке, её обозначают буквой n. Правильный объем выборки – это применение метода повторности опыта. Повторность опыта в пространстве исключает ошибку. Таким образом, ошибки нивелируются в процессе повторения.

В результате наблюдений за счет увеличения повторений получаем сведения о числовой величине изучаемого признака данной совокупности.

Возможные значения варьирующего признака X называют в математической статистике *вариантами* и обозначают x_1, x_2, x_3,…..x_n.

Полученный ряд данных варьирующих признаков можно расположить в порядке возрастания или убывания, т.е. этот ряд данных можно подвергнуть ранжированию. *Ранжирование* – это расположение значений варьирующего признака в порядке возрастания или убывания. После ранжирования нетрудно заметить, что отдельные значения варьирующих признаков повторяются несколько раз.

Числа, которые характеризуют, сколько раз повторяется, каждое значение признака в данной совокупности, называют *частотами признака* и обозначают *Σf*:

$$\Sigma f = n$$

В результате такой обработки наблюдений получаем вариационный ряд.

Вариационным рядом называют такой ряд данных, в котором указаны возможные значения варьирующего признака, расположенные в порядке убывания или возрастания и соответствующие им частоты.

Вариационные ряды могут быть с двумя типами изменчивости: с количественной и качественной.

2.2 Статистические характеристики количественной изменчивости

Основными статистическими характеристиками *количественной изменчивости* являются:

1. $\bar{\tilde{O}}$ – средняя арифметическая;
2. s^2 – дисперсия;
3. s – стандартное отклонение (среднее квадратическое отклонение);
4. $s_{\bar{x}}$ – ошибка средней арифметической (ошибка выборочной средней);
5. $s_{x\%}$ - относительная ошибка выборочной средней;
6. V – коэффициент вариации;
7. B – коэффициент выравненности.

<u>Средняя арифметическая</u> ($\bar{\tilde{O}}$) представляет собой обобщенную характеристику всей совокупности в целом. Она находится по формуле:

$$\bar{x} = \frac{\sum \tilde{O}}{n},$$

где $\sum X$ – сумма всех вариант, n – число всех вариант.

Иногда вычисляют так называемую <u>взвешенную среднюю арифметическую</u> по формуле:

$$\bar{x} = \frac{\sum f\tilde{O}}{n},$$

где $\sum X$ – сумма всех вариант, f – частота встречаемости признака, n – число всех вариант.

Основное свойство средней арифметической заключается в равенстве суммы всех положительных и всех отрицательных отклонений от нее:

$$\sum(X - \bar{x}) = 0.$$

<u>Дисперсия</u> (**s²**) и <u>стандартное отклонение</u> (**s**) служат основными мерами вариации, рассеяния изучаемого признака.

Дисперсия определяется по формуле:

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

Стандартное отклонение определяется по формуле:

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

Если исходные наблюдения сгруппированы и частоты групп обозначены через $\boldsymbol{f}$, то дисперсию и стандартное отклонение вычисляют по формулам:

$$s^2 = \frac{\sum f(X - \bar{x})^2}{n-1} \text{ и } s = \sqrt{s^2} = \sqrt{\frac{\sum f(X - \bar{x})^2}{n-1}}.$$

Число свободно варьирующих величин называется числом степеней свободы (число степеней свободы вариации), оно обозначается через **ν**:

$$\nu = n - 1.$$

Стандартное отклонение служит показателем, который дает представление о наиболее вероятной средней ошибке отдельного, единичного наблюдения, взятого из анализируемой совокупности. Стандартное отклонение называют также основным отклонением вариационного ряда. В пределах одного значения ($\pm$1s) укладывается 68% всех вариант, в пределах $\pm$2s – 95%, в пределах $\pm$3s – 99%. Утроенное значение стандартного отклонения принято считать предельной ошибкой отдельного наблюдения.

Ошибка средней арифметической ($s_{\bar{x}}$) является мерой отклонения выборочной средней **x** от средней всей (генеральной) совокупности **M**. Ошибки выборки возникают вследствии неполной репрезентативности (представительности) выборочной совокупности и свойственны только выборочному методу исследований. Ошибка средней арифметической вычисляется по формуле:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \sqrt{\frac{s^2}{n}}.$$

Ошибки выборки выражаются в тех же единицах измерения, что и варьирующий признак, и приписывают к соответствующим средним со знаком $\pm$, т.е. x $\pm s_{\bar{x}}$.

Ошибка средней арифметической тем меньше, чем меньше варьирует опытный материал и чем из большего числа измерений вычислено среднее арифметическое.

Ошибка выборки, выраженная в процентах от соответствующей средней, называется относительной ошибкой выборочной средней ($s_{x\%}$) вычисляемой по формуле:

$$s_{x\%} = \frac{s_x}{x} \times 100, \%.$$

Относительную ошибку средней иногда обозначают буквой **P** и называют «точностью опыта» или «точностью анализа».

Коэффициент вариации (**V**) является относительным показателем изменчивости и вычисляется по формуле:

$$V = \frac{s}{\bar{x}} \times 100, \%.$$

Изменчивость принято считать:

а) незначительной, если V < 10%;

б) средней, если V = 10…20%;

в) значительной, если V > 20%.

Для характеристики степени выравненности материала иногда целесообразно использовать коэффициент выравненности (**В**), который определяют по формуле:

$$B = 100\text{-}V, \%.$$

2.3 Статистические характеристики качественной изменчивости

В селекционно-семеноводческих экспериментах часто приходится иметь дело с качественной изменчивостью признаков: разная форма и окраска семян и плодов, расщепление гибридов и т.п. Частным случаем качественной изменчивости является альтернативная, при которой варьирующие признаки представляют собой одну из двух возможностей (альтернатив): наличие или отсутствие признака (ости у пшеницы). Группировка результатов наблюдений при качественном варьировании сводится к распределению совокупности объектов на группы с разными качественными признаками.

Основными статистическими показателями *качественной изменчивости* являются:

1) P – доля признака;
2) S – показатель изменчивости качественного признака;
3) V_p – коэффициент вариации качественного признака;
4) S_p – ошибка выборочной доли.

Доля признака (Р) – это относительная численность (частота) отдельной варианты в данной совокупности:

$$P_1 = \frac{n_1}{N}; \quad P_2 = \frac{n_2}{N}; \ P_3 = \frac{n_3}{N} \text{ и т.д.,}$$

где n_1, n_2, n_3 и т.д. – варианты,
N – численность изучаемой совокупности.

Доля признака выражается или в долях единицы, или в процентах.

При альтернативной изменчивости доля одного признака обозначается через **p**, а второго через **q**, где **p** + **q**=1 (100%), **q**=1-**p**.

Показатель изменчивости качественного признака (**s**) характеризует варьирование величин ряда относительно друг друга и определяется по формуле:

$$s = \sqrt[k]{p_1 \times p_2 \times p_3 \times \ldots p_k} \text{, где}$$

p_1, p_2, p_3 и т.д. – доли признака в общей совокупности;
к – число градаций признака.

Если к > 2, то формулу удобнее пролагарифмировать:

$$\lg s = \frac{\lg p_1 + \lg p_2 + \ldots + \lg p_k}{k}.$$

Если изучаемая совокупность представлена объектами с двумя градациями признака (альтернативная изменчивость, к = 2), то :

$$s = \sqrt{pq},$$

где p и q – доли признака, выраженные в долях единицы или процентах.

В зависимости от соотношения **p** и **q** значение **s** изменяется от 0 до 0,5. Максимальная изменчивость качественного признака (s_{max}) будет наблюдаться когда p = q = 0,5 и $s_{\max} = \sqrt{0{,}5 \times 0{,}5}$ = 0,5 или 50%. Значения максимальной изменчивости для распределений с разным числом градаций качественных признаков следующие:

Число градаций признака	s_{max}
2	0,5 (50%)
3	0,333 (33,3%)
4	0,25 (25%)
5	0,2 (20%)
6	0,167 (16,7%)
7	0,143 (14,3%)

Коэффициент вариации качественных признаков (**V_p**) – фактический показатель изменчивости, выраженный в процентах к максимально возможной изменчивости:

$$V_p = \frac{s}{s_{\max}} \times 100,\% .$$

Ошибка выборочной доли (**s_p**) – мера отклонения доли признака выборочной совокупности **p** от доли его по всей генеральной совокупности **P** вследствии неполной представительности (репрезентативности) выборки, вычисляется по формуле:

$$s_p = \frac{s}{\sqrt{N}}, \text{ где}$$

s – показатель изменчивости качественного признака;
N – объем выборки.

Для альтернативного варьирования ($s = \sqrt{pq}$) формула вычисления будет:

$$s_p = \sqrt{\frac{pq}{N}} .$$

Вероятность встретить p (или q) в интервале p $\pm$ s_p составляет 68%, в интервале p $\pm$ 2s_p – 95%, в интервале p $\pm$ 3s_p – 99%. Следовательно, подобно количественной изменчивости, все значения **p** с вероятностью 99% укладываются в пределах тройной ошибки выборочной доли.

Вопросы для самоконтроля

1.Что такое статистическая гипотеза?
2. Что называют в математической статистике вариантами?
3. Что называют в математической статистике вариационным рядом?
4. Основные статистические характеристики количественной изменчивости.
5. Основные меры вариации, рассеяния изучаемого признака.
6. Статистические характеристики качественной изменчивости.
7. Показатель изменчивости качественного признака.

СПИСОК ЛИТЕРАТУРЫ

Основная

1. Основы научных исследований в растениеводстве и селекции: Учебное пособие / А.Ф. Дружкин, Ю.В. Лобачев, Л.П. Шевцова, З.Д. Ляшенко // ФГБОУ ВПО «Саратовский ГАУ». – Саратов, 2013. – 264 с.

Дополнительная

1. **Доспехов Б.А.** Методика полевого опыта (с основами статистической обработки результатов исследований) – 5-е издание, доп. и перераб. – М.: Агропромиздат, 1985. – 351 с.
2. **Смиряев А.В., Мартынов С.П., Кильчевский А.В.** Биометрия в генетике и селекции растений. – М.: издательство МСХА, 1992. – 269 с.
3**. Рокицкий П.Ф.** Введение в статистическую генетику. – Минск: Вышейшая школа, 1978. – 442 с.

Лекция 3
ДИСПЕРСИОННЫЙ АНАЛИЗ

3.1 Применение и использование дисперсионного анализа

Дисперсионный анализ был разработан английским ученым Р. Фишером, который открыл закон распределения отношений дисперсий:

$$\frac{s_1^2}{s_2^2} = F\ .$$

Дисперсионный анализ широко используют для планирования эксперимента и статистической обработки его данных. При дисперсионном анализе одновременно обрабатывают данные нескольких выборок (вариантов), составляющих единый статистический комплекс, оформленный в виде специальной рабочей таблицы. Структура статистического комплекса и его последующий анализ определяются схемой и методикой эксперимента.

Дисперсионный анализ – это расчленение общей суммы квадратов отклонений и общего числа степеней свободы на части – компоненты, соответствующие структуре эксперимента, и оценка значимости действия и взаимодействия изучаемых факторов по F-критерию.

В однофакторном статистическом комплексе общая изменчивость признака, измеряемая общей суммой квадратов, расчленяется на два компонента: варьирование между выборками (вариантами) и внутри выборок:

$C_y = C_v + C_z$, где

C_y – общая изменчивость признака,
C_v – варьирование между вариантами,
C_z – варьирование внутри вариантов.

Здесь вариация между вариантами представляет ту часть общей дисперсии, которая обусловлена действием изучаемых факторов, а дисперсия внутри вариантов характеризует случайное варьирование изучаемого признака, т.е. ошибку эксперимента.

Общее число степеней свободы (N-1) также расчленяется на две части – степени свободы для вариантов (l-1) и для случайного варьирования (N-l):

$$(N-1)=(l-1) + (N - l).$$

Если обрабатывают дисперсионным анализом однофакторные сопряженные статистические комплексы с наличием **n** организованных повторений в полевом эксперименте, то общая сумма квадратов разлагается на три части: варьирование вариантов C_v, варьирование повторений C_p, случайное варьирование C_z. Тогда общая изменчивость и общее число степеней свободы могут быть представлены выражениями:

$$C_y = C_v + C_p + C_z,$$
$$(N-1)=(l-1) + (n-1) + (n-1)(l-1).$$

Суммы квадратов отклонений по данным полевого эксперимента (статистического комплекса с l-вариантами и n-повторениями) находят обычно в следующей последовательности. В исходной таблице определяют суммы по повторениям P, по вариантам V и общую сумму всех наблюдений ∑X. Затем вычисляют:

1) общее число всех наблюдений $N=l \cdot n$;
2) корректирующий фактор (поправку) $C=(\sum X)^2/N$;
3) общую сумму квадратов $C_y=\sum X^2-C$;
4) сумму квадратов для повторений $C_p=\sum P^2/l-C$;
5) сумму квадратов для вариантов $C_v=\sum V^2/n-C$;
6) сумму квадратов для ошибки (остаток) $C_z=C_y-C_p-C_v$.

Две последние суммы квадратов C_v и C_z делят на соответствующие им степени свободы, т.е. приводят к сравниваемому виду – одной степени свободы вариации. В результате получают две дисперсии: для вариантов $s_v^2=C_v/(l-1)$ и ошибки $s^2=C_z/(n-1)(l-1)$.

Эти средние квадраты и используют в дисперсионном анализе для оценки значимости действия изучаемых факторов.

Оценка дисперсий

Оценка проводится путем сравнения дисперсии вариантов (s_v^2) с дисперсией ошибки (s^2) по критерию Фишера:

$$F= s_v^2/ s^2.$$

За единицу сравнения принимают средний квадрат случайной дисперсии, которая определяет случайную ошибку эксперимента. При этом проверяемой нулевой гипотезой служит предположение: все выборочные средние являются оценками одной генеральной средней, следовательно различия между ними не существенны. Если $\mathbf{F_{факт.} = s_v^2/ s^2 < F_{теор.}}$, то нулевая гипотеза не отвергается (между всеми выборочными средними нет существенных различий), на этом проверка заканчивается. Нулевая гипотеза отвергается, если $\mathbf{F_{факт.} = s_v^2/s^2 \geq F_{теор.}}$ В этом случае дополнительно оценивают существенность частных различий по НСР и определяют, между какими средними имеются значимые разности. Теоретические значения критерия F для принятого в исследовании уровня значимости находят по специальным таблицам с учетом числа степеней свободы для дисперсии вариантов (числитель) и случайной дисперсии (знаменатель).

3.2 Общая схема дисперсионного анализа однофакторных экспериментов

В селекционно-семеноводческих экспериментах используют разные модели дисперсионного анализа, отражающие условия и методику проведения эксперимента:

Вид эксперимента	Суммы квадратов (в числителе) и степени свободы (в знаменателе)				
	общая	повторений (рядов)	столбцов	вариантов	ошибка (остаток)
1. Полевой опыт, проведенный методом неорганизованных повторений (полная рендомизация)	$\mathbf{C_y/N-1}$	-	-	$\mathbf{C_v/\mathit{l}-1}$	$\mathbf{C_z/N-1}$
2. Полевой опыт, проведенный методом организованных повторений	$\mathbf{C_y/N-1}$	$\mathbf{C_p/n-1}$	-	$\mathbf{C_v/\mathit{l}-1}$	$\mathbf{C_z/(n-1)(\mathit{l}-1)}$
3. Латинский квадрат	$\mathbf{C_y/N-1}$	$\mathbf{C_p/n-1}$	$\mathbf{C_c/n-1}$	$\mathbf{C_v/n-1}$	$\mathbf{C_z/(n-1)(n-2)}$
4. Латинский прямоугольник.	$\mathbf{C_y/N-1}$	$\mathbf{C_p/n-1}$	$\mathbf{C_c/n-1}$	$\mathbf{C_v/\mathit{l}-1}$	$\mathbf{C_z/(n-1)(\mathit{l}-2)}$

где: N – общее число наблюдений,

l – число вариантов,

n – число повторений, рядов и столбцов.

Анализ влияния изучаемых факторов

Дисперсионный анализ дает возможность получить представление о степени, или доле влияния того или иного фактора в общей дисперсии признака, которую принимают за единицу или 100%:

$\eta_v^2 = C_v/C_y$ – влияние вариантов;

$\eta_p^2 = C_p/C_y$ – влияние повторений;

$\eta_z^2 = C_z/C_y$ влияние случайных факторов.

$\eta_y^2 = \eta_v^2 + \eta_p^2 + \eta_z^2 = 1$ (или 100%) – влияние всех факторов.

При этом:

$\eta_v = \sqrt{\eta_v^2} = \sqrt{C_v / C_y}$ – корреляционное отношение, характеризующее тесноту связи результативного признака с факториальным;
η_v^2 – индекс детерминации, показывающий долю его варьирования под действием изучаемых факторов.

Дисперсионный анализ имеет следующие основные преимущества перед методом попарных сравнений по критерию Стьюдента:

1) вместо индивидуальных ошибок средних по каждому варианту, в дисперсионном анализе используется обобщенная ошибка средних, которая опирается на большее число наблюдений и является более надежной базой для оценок;
2) методом дисперсионного анализа можно обрабатывать данные простых и сложных, однолетних и многолетних, однофакторных и многофакторных опытов;
3) дисперсионный анализ позволяет избежать громоздких вычислений при большом числе вариантов в опыте и позволяет компактно в виде существенности разностей представить итоги статистической обработки.

Современная теория планирования эксперимента и статистический анализ базируются на принципах рендомизации. Теория требует, чтобы все наблюдения были независимы. В этом случае дисперсионный анализ дает правильную, несмещенную оценку ошибки эксперимента. Если эксперимент не рендомизирован, то экспериментатор может получить смещенную оценку ошибки опыта, и обычно используемые в дисперсионном анализе критерии значимости теряют законную силу и не могут использоваться в качестве аргументов строго статистического доказательства эффектов вариантов.

3.3 Однофакторный дисперсионный анализ

Однофакторный дисперсионный анализ позволяет выявить влияние только одного фактора. При этом общее варьирование результативного признака разлагается на два компонента – варьирование вариантов и случайное варьирование : $C_y = C_v + C_z$.

Однофакторный дисперсионный анализ невозможен для полевых экспериментов, проведенных без повторностей. Минимум повторностей – две, но лучше – четыре-шесть.

Статистический анализ данных однофакторного полевого эксперимента проводят в три этапа:

1) Составляют расчетную таблицу, располагая в ней исходные данные по рядам и столбцам, определяют суммы и средние по вариантам, общую сумму и среднее значение результативного признака по эксперименту:

Варианты	Исходные данные **X**	Число наблюдений **n**	Суммы по вариантам **V**	Средние по вариантам $\bar{\tilde{o}}$
1 2 3 … l	$X_{11}, X_{12}, X_{13}, \dots X_{1n}$ $X_{21}, X_{22}, X_{23}, \dots X_{2n}$ $X_{31}, X_{32}, X_{33}, \dots X_{3n}$ … $X_{l1}, X_{l2}, X_{l3}, \dots X_{ln}$	n_1 n_2 n_3 … n_l	V_1 V_2 V_3 … V_l	$\bar{\tilde{o}}_1$ $\bar{\tilde{o}}_2$ $\bar{\tilde{o}}_3$ … $\bar{\tilde{o}}_l$
Общая сумма		**N=∑n**	**∑X=∑V**	**x=∑X/N**

2) Вычисляют суммы квадратов отклонений и F-критерий по следующим формулам и заносят в таблицу:

Дисперсия	Сумма квадратов	Степени свободы	Средний квадрат	$F_{факт.}$	$F_{теор.}$
Общая C_y **Вариантов C_v** **Остаток C_z**	$\sum X^2$ - C $\sum V^2/n$ – C C_y - C_v	**N-1** ***l*-1** **N-l**	- s_v^2 s^2	- s_v^2/s^2 -	- по стандартным таблицам -

Корректирующий фактор (C) вычисляют по формулам:

$$C=(\sum X)^2/N \text{ или } C=x\sum X.$$

3) Определяют ошибку опыта и существенность частных различий.

$$\text{ошибка средней: } s_x = \sqrt{\frac{s^2}{n}};$$

$$\text{ошибку разности средних: } s_d = \sqrt{\frac{2s^2}{n}};$$

$$НСР_{05}=t_{05} \times s_d$$

$$\text{или НСР в процентах: } НСР_{05}= \frac{t_{05} \times s_d}{\bar{x}} \times 100.$$

4) Итоги результатов эксперимента и статистической обработки данных записывают в итоговую таблицу.

Пример 1. «Испытание пяти сортов по урожайности зерна».

1) Составляют расчетную таблицу:

Варианты	Урожайность зерна, г/делянка				Число наблюдений **n**	Суммы по вариантам **V**	Средние по вариантам $\bar{x}$
1 (st)	454	470	430	500	4	1854	463,5
2	502	550	490	507	4	2049	512,2
3	601	670	550	607	4	2428	607,0
4	407	412	475	402	4	1696	424,0
5	418	470	460	412	4	1760	440,0
Общая сумма					$N=\sum n=20$	$\sum X=9787$	$\bar{x}=489,4$

2) Вычисляют суммы квадратов отклонений и F-критерий.

$C=(\sum X)^2/N=(9787)^2/20=4789268;$

$C_y=\sum X^2 – C=(454^2+470^2+…+412^2)-4789268=104941;$

$C_v=\sum V^2/n – C=(1854^2+2049^2+…+1760^2)/4 – 4789268=86961;$

$C_z=C_y-C_v=104941-86961=17980;$

$F_{факт.}=21740/1199=18,13.$

Полученные результаты заносят в таблицу дисперсионного анализа:

Дисперсия	Сумма квадратов	Степени свободы	Средний квадрат	$F_{факт.}$	$F_{теор.}$
Общая	104941	19	-	-	-
Вариантов	86961	4	21740	18,13	3,06
Остаток	17980	15	1199	-	-

Проводят оценку нулевой гипотезы. Поскольку $F_{факт.} > F_{теор.}$, то нулевая гипотеза отвергается, т.е. в опыте есть существенные различия по вариантам на 5%-м уровне значимости.

3) Определяют ошибку опыта и существенность частных различий:

$$\text{ошибка средней: } s_x = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{1199}{4}} = 17.3 \text{ г.;}$$

$$\text{ошибку разности средних: } s_d = \sqrt{\frac{2s^2}{n}} = \sqrt{\frac{2 \times 1199}{4}} = 24.5 \text{ г.;}$$

$$НСР_{05} = t_{05} \times s_d = 2.13 \times 24.5 = 52.2 \text{ г,}$$

$$\text{или НСР в процентах: } НСР_{05} = \frac{t_{05} \times s_d}{\bar{\tilde{o}}} \times 100 = \frac{52.2}{489.4} \times 100 = 10.7\%.$$

4) Итоги результатов эксперимента и статистической обработки данных записывают в итоговую таблицу:

Сорт	Урожайность зерна, г/делянка	Разность со стандартом		Группа по урожайности зерна
		г/делянка	%	
Саратовская 55 (st)	463,5	-	-	1 (st)
Саратовская 29	424,0	-39,5	-8,1	1
Саратовская 46	440,0	-23,5	-4,8	1
Саратовская 58	512,2	48,7	10,0	1
Л 503	607,0	143,5	29,3	2

Вывод: в сортоиспытании установлены достоверные различия по урожайности зерна между изучаемыми сортами (вариантами). Достоверно превысил по урожайности зерна сорт-стандарт только сорт Л 503.

3.4 Многофакторный дисперсионный анализ

Дисперсионный анализ данных многофакторного эксперимента проводят в два этапа. Первый этап – разложение общей вариации результативного признака на два компонента – варьирование вариантов и случайное варьирование: $C_y = C_v + C_z$. Второй этап – разложение суммы квадратов отклонений для вариантов на компоненты, соответствующие источникам варьирования – главным эффектам изучаемых факторов и их взаимодействий.

В двухфакторном эксперименте – $C_v = C_A + C_B + C_{AB}$.

В трехфакторном эксперименте – $C_v = C_A + C_B + C_C + C_{AB} + C_{AC} + C_{BC} + C_{ABC}$.

Пример 2. «Двухфакторный полевой эксперимент по изучению действия двух доз азота и трех доз фосфора на урожайность зерна пшеницы».

1) Составляют расчетную таблицу:

Азот A	Фосфор B	Урожайность зерна, г/делянка				Суммы **V**	Средние $\bar{\tilde{o}}$
a_0	$в_0$	24,1	25,8	23,0	27,0	99,9	25,0
	$в_1$	28,4	29,7	30,1	27,4	115,6	28,9
	$в_2$	28,7	30,4	32,0	27,0	118,1	29,5
a_1	$в_0$	30,7	34,4	34,0	31,0	130,1	32,5
	$в_1$	46,7	45,4	47,1	46,3	185,5	46,4
	$в_2$	59,4	50,7	64,5	60,1	234,7	58,7
Общая сумма						$\sum X = 883,9$	$\bar{\tilde{o}} = 36,8$

2) Вычисляют суммы квадратов отклонений по следующим формулам:
$N=l_A\times\ l_B\times n=2\times 3\times 4=24.$
$C=(\sum X)^2/N=(883,9)^2/24=32553,3;$
$C_y=\sum X^2 - C=(24,1^2+25,8^2+\ldots+60,1^2)-32553,3=3505,2;$
$C_v=\sum V^2/n - C=(99,9^2+115,6^2+\ldots+234,7^2)/4 - 32553,3=3374,5;$
$C_z=C_y-C_v=3505,2-3374,5=130,7;$
3) Вычисляют суммы квадратов по факторам А, В и взаимодействию АВ и составляют таблицу:
сумма квадратов для фактора А (азот):
$C_A=\sum A^2/l_B\times n-C=(333,6^2+550,3^2)/3\times 4-32553,3=1956,6$
при $(l_A-1)=2-1=1$ степени свободы;
сумма квадратов для фактора В (фосфор):
$C_B=\sum B^2/l_A\times n-C=(230,0^2+301,1^2+352,8^2)/2\times 4-32553,3=950,3$
при $(l_B-1)=3-1=2$ степени свободы;
сумма квадратов для взаимодействия АВ (азот-фосфор) находят по разности:
$C_{AB}=C_v-C_A-C_B=3374,5-1956,6-950,3=467,6$
при $(l_A-1)(l_B-1)=(2-1)(3-1)=2$ степенях свободы.
$F_{факт.A}=1956,60/7,26=269,50;$
$F_{факт.B}=475,15/7,26=65,45;$
$F_{факт.AB}=233,80/7,26=32,20.$

Составляют таблицу результатов дисперсионного анализа:

Дисперсия	Сумма квадратов	Степени свободы	Средний квадрат	$F_{факт.}$	$F_{теор.}$
Общая	3505,2	23	-	-	-
Азота (А)	1956,6	1	1956,60	269,50	4,41
Фосфора (В)	950,3	2	475,15	65,45	3,55
Взаимодействия АВ	467,6	2	233,80	32,20	3,55
Остаток	130,7	18	7,26	-	-

4) Для оценки существенности частных различий вычисляют:
$$s_x=\sqrt{\frac{s^2}{n}}=\sqrt{\frac{7,26}{4}}=1,35\ \text{г};$$
$$s_d=\sqrt{\frac{2s^2}{n}}=\sqrt{\frac{2\times 7,26}{4}}=1,90\ \text{г};$$
$НСР_{05}=t_{05}\times s_d=2,10\times 1,9=3,99$ г;

Результаты эксперимента и статистической обработки данных представляют в виде итоговой таблицы:

Действие азота и фосфора на урожайность пшеницы, г/делянка

Дозы азота	Дозы фосфора		
	0	1	2
0	25,0 (st)	28,9	29,5
1	32,5	46,4	58,7
$НСР_{05}$	3,99		

<u>Вывод:</u> в полевом эксперименте установлены различия по изучаемым факторам (азоту, фосфору, и взаимодействию азота с фосфором) на урожайность пшеницы. Применение азота, фосфора (2-я доза), а также сочетание азота и фосфора достоверно увеличивает урожайность зерна. Первая доза фосфора достоверно не влияет на урожайность зерна пшеницы.

Вопросы для самоконтроля

1.Применение и использование дисперсионного анализа.
2. Оценка дисперсий.
3. Анализ влияния изучаемых факторов.
4. Основные преимущества дисперсионного анализа перед методом попарных сравнений по критерию Стьюдента.
5. Этапы проведения многофакторного дисперсионного анализа.

СПИСОК ЛИТЕРАТУРЫ

Основная

1. Основы научных исследований в растениеводстве и селекции: Учебное пособие / А.Ф. Дружкин, Ю.В. Лобачев, Л.П. Шевцова, З.Д. Ляшенко // ФГБОУ ВПО «Саратовский ГАУ». – Саратов, 2013. – 264 с.

Дополнительная

1. **Доспехов Б.А.** Методика полевого опыта (с основами статистической обработки результатов исследований) – 5-е издание, доп. и перераб. – М.: Агропромиздат, 1985. – 351 с.
2. **Смиряев А.В., Мартынов С.П., Кильчевский А.В.** Биометрия в генетике и селекции растений. – М.: издательство МСХА, 1992. – 269 с.
3**. Рокицкий П.Ф.** Введение в статистическую генетику. – Минск: Вышейшая школа, 1978. – 442 с.

Лекция 4

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

4.1 Применение и использование корреляционного анализа

В селекционно-семеноводческих исследованиях для изучения связи между двумя параметрами используют корреляционный анализ.

Обычно исследователь встречается с тремя ситуациями:

1) между величинами X и Y связь отсутствует;

2) между величинами X и Y существует функциональная связь (когда каждому значению одной величины соответствует строго определенное значение другой величины);

3) между величинами X и Y существует стохастическая (вероятностная, корреляционная) связь (когда каждому значению одной величины соответствует множество возможных значений другой величины).

Корреляционный анализ применяют для изучения стохастических связей. Корреляция может быть линейной и нелинейной. Линейную корреляцию называют положительной (прямой), когда с увеличением признака X возрастает и признак Y, и отрицательной (обратной), когда с увеличением одного признака другой признак уменьшается. Два признака могут быть скоррелированы в большей или меньшей степени. Численной мерой корреляции при линейной связи является коэффициент корреляции (r), а при нелинейной связи использовать этот показатель запрещено. Вместо него используют корреляционное отношение (η), которое можно применять и для оценки линейной связи.

4.2 Парная линейная корреляция

Корреляция называется парной, если на величину результативного признака влияет один факториальный признак. Численной мерой парной корреляции при линейной связи является коэффициент корреляции.

Коэффициент корреляции (r) является безразмерной величиной, изменяющейся в области –1 < r < +1. Коэффициент корреляции рассчитывают по разным формулам, в зависимости от того, какими признаками являются X и Y: количественными или качественными.

Если оба признака являются количественными, то коэффициент корреляции рассчитывают по формуле:

$$r = \frac{\sum (\tilde{O} - \bar{x}) \times (Y - \bar{y})}{\sqrt{\sum (\tilde{O} - \bar{x})^2 \times (Y - \bar{y})^2}} \qquad \textbf{(1)}$$

или

$$r = \frac{\sum \tilde{O}Y - (\sum \tilde{O} \times \sum Y)/n}{\sqrt{(\sum \tilde{O}^2 - (\sum \tilde{O})^2/n) \times (\sum Y^2 - (\sum Y)^2/n)}} \qquad \textbf{(1)}$$

Если r=+1 или r=-1, то стохастическая связь переходит в функциональную, которую можно считать частным случаем корреляционной связи. Если r=0, то между X и Y нет линейной связи (но нелинейная связь может существовать). Корреляционная зависимость между признаками слабая, если r < 0,3, средняя при r = 0,3…0,7 и сильная при r > 0,7.

Степень сопряженности в вариации двух величин более точно измеряется квадратом коэффициента корреляции (r^2), который называется коэффициентом детерминации (d_{yx}). Он показывает долю (%) тех изменений, которые в данном явлении зависят от изучаемого фактора. Коэффициент детерминации является более непосредственным и прямым способом выражения зависимости одной величины от другой, в этом отношении он предпочтительней коэффициента корреляции.

Для оценки надежности выборочного коэффициента корреляции вычисляют его ошибку и критерий существенности.

Стандартную ошибку коэффициента корреляции определяют по формуле:

$$s_r = \sqrt{\frac{1-r^2}{n-2}},$$

где s_r – стандартная ошибка коэффициента корреляции;
r – коэффициент корреляции;
n – численность выборки (число пар значений X и Y).

Критерий существенности коэффициента корреляции рассчитывают по формуле:

$$t_r = r/s_r.$$

В корреляционном анализе обязательно проверяют нулевую гипотезу H_0 (r=0). Если $t_r \geq t_{теор.}$, то корреляционная связь существенна, а если $t_r < t_{теор.}$ – несущественна. Теоретическое значение критерия t находят по стандартной таблице Стьюдента, принимая 5%-й уровень значимости. Число степеней свободы равно n-2. Нулевую гипотезу можно проверить и без расчета критерия t_r. Для этого пользуются специальной таблицей, в которой даны критические значения коэффициента корреляции на 5%-м и 1%-м уровнях значимости. Между величинами X и У имеется существенная связь и нулевая гипотеза отвергается, если $r_{факт.} \geq r_{теор.}$. Нулевая гипотеза остается в силе, если $r_{факт.} < r_{теор.}$.

При постановке экспериментов исследователь должен знать величины выборок для проведения корреляционного анализа. Для доказательства значимости слабых связей необходим размер выборки в 40-100, средних связей – 12-40 и сильных связей – 6-12 пар наблюдений.

Исследователь должен знать, что при малых выборках и значениях r, близких к единице, распределение выборочных коэффициентов корреляции отличается от нормального. Поэтому для оценки существенности (значимости) коэффициента корреляции, построения доверительных интервалов относительно корреляции в генеральной совокупности и сравнения коэффициентов корреляции критерий t Стьюдента становится ненадежным. Для этих ситуаций Р. Фишер предложил преобразовать r в величину z, которая априори распределена нормально. Для перехода от r к z и обратно используют специальную таблицу. Стандартную ошибку величины z определяют по формуле:

$$s_z = \frac{1}{\sqrt{n-3}},$$ где n – объем выборки.

Критерий существенности величины z определяют по формуле:

$$t_z = \frac{z}{s_z}.$$

Критерий значимости для разности z_1-z_2 и доверительные интервалы определяют по формулам:

$$t_{z1-z2} = \frac{z_1 - z_2}{\sqrt{s_{z1}^2 + s_{z2}^2}}, \quad z \pm t\, s_z.$$

После определения доверительных границ обратным преобразованием по специальной таблице находят соответствующие z_{max} и z_{min} величины r_{max} и r_{min}.

Исследователь также должен знать, что фактический коэффициент корреляции используется только для описания линейной зависимости при условии двумерного нормального распределения изучаемых признаков и характеризует взаимосвязь признаков только в данной изучаемой совокупности.

В селекционно-семеноводческих исследованиях часто приходится определять зависимость между качественными (альтернативными) признаками. Для этой ситуации коэффициент корреляции рассчитывают по формуле Юла:

$$r = \frac{n_1 \times n_4 - n_2 \times n_3}{\sqrt{N_1 \times N_2 \times N_3 \times N_4}}, \quad \mathbf{(2)}$$

где n_1 – частота клеток со знаком «++»,
n_2 – частота клеток со знаком «+ –»,
n_3 – частота клеток со знаком «–+»,
n_4 – частота клеток со знаком «– –»,
N_1 и N_2 – суммы частот по строкам,
N_3 и N_4 – суммы частот по столбцам.

Пример составления корреляционной решетки:

Действие бора на поражаемость растений гнилью

Растения	Добавка бора (+)	Без бора (–)	Сумма по строкам
Непораженные (+)	n_1	n_2	N_1
Пораженные (–)	n_3	n_4	N_2
Сумма по столбцам	N_3	N_4	-

Для установления сопряженности между качественными признаками, имеющими несколько градаций (рангов, баллов), применяют коэффициент ранговой корреляции Спирмана, который определяют по формуле:

$$r_s = 1 - \frac{\sum d^2}{n \times (n^2 - 1)}, \quad \mathbf{(3)}$$

где d – разность между рангами сопряженных рядов X и Y (d=X–Y);
n – число парных наблюдений.

Коэффициент ранговой корреляции целесообразно вычислять в тех случаях, когда совокупность двух переменных не имеет нормального распределения. Для выборок, взятых из нормальных совокупностей следует рассчитывать обычный коэффициент корреляции по формуле **(1)**.

Иногда в селекционно-семеноводческих исследованиях возникает необходимость изучения сопряженности двух признаков, один из которых можно измерить (количественный признак), а в отношении другого только отметить его наличие или отсутствие (качественный признак). Коэффициент корреляции между качественными и количественными признаками вычисляют по формуле:

$$r = \frac{x_1 - \bar{x}}{s} \times \sqrt{\frac{n_1}{n - n_1}}, \quad \mathbf{(4)}$$

где $\bar{x}$ – общее среднее значение для количественного признака;
x_1 – среднее значение количественного признака с наличием качественного;
n – общее число всех наблюдений;
n_1 – число случаев с наличием качественного признака;

s – общее стандартное отклонение для количественного признака.

Для расчета исходные данные заносят в таблицу:

Количественный признак X	Качественный признак			$f_1 \times X$	$f_2 \times X$	$f \times X$	X^2	$f \times X^2$
	f_1 (+)	f_2 (–)	f (f_1+f_2)					

$$\tilde{o} = \frac{\sum f \times X}{n}, \quad x_1 = \frac{\sum f_1 \times X}{n_1}, \quad x_2 = \frac{\sum f_2 \times X}{n_2}, \quad s = \sqrt{\frac{\sum f \times X^2 - (\sum f \times X)^2 / n}{n-1}},$$

$$s_r = \sqrt{\frac{1-r^2}{n-2}}, \quad t_r = \frac{r}{s_r}, \quad \text{v=n-2.}$$

4.3 Множественная линейная корреляция

Корреляция называется множественной, если на величину результативного признака одновременно влияют несколько факториальных признаков.

В качестве меры тесноты линейной связи трех признаков используют частные коэффициенты корреляции ($r_{xy\bullet z}$, $r_{xz\bullet y}$, $r_{yz\bullet x}$) и множественные коэффициенты корреляции ($R_{x\bullet yz}$, $R_{y\bullet xz}$, $R_{z\bullet xy}$).

Частный коэффициент корреляции – это показатель, измеряющий степень сопряженности двух признаков при постоянном значении третьего. Частные коэффициенты корреляции рассчитывают по формулам:

$$r_{xy\bullet z} = \frac{r_{xy} - r_{xz} \times r_{yz}}{\sqrt{(1-r_{xz}{}^2) \times (1-r_{yz}{}^2)}};$$

$$r_{xz\bullet y} = \frac{r_{xz} - r_{xy} \times r_{yz}}{\sqrt{(1-r_{xy}{}^2) \times (1-r_{yz}{}^2)}};$$

$$r_{yz\bullet x} = \frac{r_{yz} - r_{xy} \times r_{xz}}{\sqrt{(1-r_{xy}{}^2) \times (1-r_{xz}{}^2)}}.$$

Ошибку и критерий значимости частной корреляции определяют по тем же формулам, что и парной корреляции (число степеней свободы v=n-3). Частные коэффициенты корреляции могут принимать значения между -1 и +1. Частные коэффициенты детерминации находят путем возведения в квадрат частных коэффициентов корреляции.

Множественные коэффициенты корреляции трех переменных – это показатель тесноты линейной связи между одним из признаков и совокупностью двух других признаков. Множественные коэффициенты корреляции можно вычислить по формулам, если известны коэффициенты парной корреляции:

$$R_{x\bullet yz} = \sqrt{\frac{r_{xy}{}^2 + r_{xz}{}^2 - 2 \times r_{xy} \times r_{xz} \times r_{yz}}{1-r_{yz}{}^2}};$$

$$R_{y\bullet xz} = \sqrt{\frac{r_{xy}{}^2 + r_{yz}{}^2 - 2\times r_{xy}\times r_{xz}\times r_{yz}}{1-r_{xz}{}^2}}\ ;$$

$$R_{z\bullet xy} = \sqrt{\frac{r_{xz}{}^2 + r_{yz}{}^2 - 2\times r_{xy}\times r_{xz}\times r_{yz}}{1-r_{xy}{}^2}}\ .$$

Множественный коэффициент корреляции может принимать значения от 0 до +1. Квадрат коэффициента множественной корреляции (R^2) называется <u>коэффициентом множественной детерминации</u>.

Значимость множественной корреляции оценивают по F-критерию:

$F_{\text{ôàêò}.} = (\frac{R^2}{1-R^2})\times(\frac{n-k}{k-1})$, где n – объем выборки, k – число признаков.

Теоретические значения F-критерия берут по стандартным таблицам для v_1=k-1 и v_2=n-k степеней свободы и принятого уровня значимости. Нулевая гипотеза отвергается, если $F_{факт.} \geq F_{теор.}$.

4.4 Тест на линейность

Для определения степени линейности зависимости между изучаемыми признаками используют F-критерий, вычисляемый по формуле:

$$F = \frac{(\eta^2 - r^2)\times(n-k)}{(1-\eta^2)\times(k_x - 2)}\ ,$$

где η^2 –квадрат корреляционного отношения У по X;

r^2 – квадрат коэффициента линейной корреляции;

n – объем выборки;

k_x – число групп по ряду X.

Связь можно практически принять за линейную (прямолинейную), если $F_{факт.} < F_{теор.}$. Корреляция считается нелинейной (криволинейной), если $F_{факт.} \geq F_{теор.}$. Теоретические значения F-критерия берут из стандартной таблицы при $v_1=k_x$-2 и v_2=n-2 степеней свободы.

4.5 Криволинейная корреляция

При криволинейной корреляции степень зависимости между двумя величинами измеряет <u>корреляционное отношение</u> (η), которое при <u>малом объеме наблюдений</u> вычисляют по формуле:

$$\eta_{\text{óõ}} = \sqrt{\frac{\sum(Y-\bar{y})^2 - \sum(Y-\bar{y}_{\text{õ}})^2}{\sum(Y-\bar{y})^2}}\ , \textbf{(1)}$$

где $\sum(Y-\bar{y})^2$ – сумма квадратов отклонений индивидуальных значений Y от средней арифметической $\bar{y}$;

$\sum(Y-\bar{y}_x)^2$ – сумма квадратов отклонений вариант от частных средних $\bar{y}_x$, соответствующих определенным, фиксированным значениям независимой переменной X.

Для вычисления корреляционного отношения значения независимого признака X располагают по ранжиру в возрастающем порядке и разбивают весь ряд наблюдений на 4-7 групп с таким расчетом, чтобы в каждой группе по ряду X было не менее двух наблюдений. Затем определяют общую среднюю y, групповые средние y_x и суммы квадратов отклонений для общего $\sum (Y-\bar{y})^2$ и группового $\sum (Y-\bar{y}_x)^2$ варьирования признака Y.

При <u>большом объеме наблюдений</u> (n>30) для вычисления корреляционного отношения строят корреляционную таблицу. После группировки и разноски дат определяют сумму квадратов отклонений общего варьирования $\sum f \times (Y-\bar{y})^2$, сумму квадратов отклонений группового варьирования $\sum f \times (\bar{y}_x - \bar{y})^2$ и вычисляют корреляционное отношение по формуле:

$$\eta_{ух} = \sqrt{\frac{\sum f_х \times (\bar{y}_х - \bar{y})^2}{\sum f \times (Y-\bar{y})^2}}. \quad \mathbf{(2)}$$

При обработке экспериментального материала методом дисперсионного анализа корреляционное отношение можно определить по формуле:

$$\eta_{yx} = \sqrt{\frac{C_v}{C_y}}. \quad (3)$$

Если η=0, зависимость У от X отсутсвует, если η=1, между Y и X имеется функциональная связь, если 0<η<1, между У и X имеется стохастическая (корреляционная) зависимость.

<u>Стандартную ошибку корреляционного отношения</u> определяют по формуле:

$$s_\eta = \sqrt{\frac{1-\eta^2}{n-2}}.$$

<u>Критерий существенности корреляционного отношения</u> определяют по формуле:

$$t_\eta = \frac{\eta}{s_\eta}.$$

В корреляционном анализе обязательно проверяют нулевую гипотезу H_0 (r=0). Если $t_\eta \geq t_{теор.}$, то корреляционная связь существенна, а если $t_\eta \leq t_{теор.}$ – несущественна. Теоретическое значение t-критерия для 5%-го и 1%-го уровня значимости находят по стандартным таблицам при числе степеней свободы v=n-2.

Значимость корреляционного отношения также оценивают по F-крите-рию:

$$F_{факт.} = \left(\frac{\eta^2}{1-\eta^2}\right) \times \left(\frac{n-k}{k-1}\right),$$ где n – объем выборки, k – число признаков.

Теоретические значения F-критерия берут по стандартным таблицам для v_1=k-1 и v_2=n-k степеней свободы для 5%-го и 1%-го уровня значимости. Нулевая гипотеза отвергается, если $F_{факт.} \geq F_{теор.}$.

Вопросы для самоконтроля

1. Применение и использование корреляционного анализа.
2. Линейная и нелинейная корреляция.
3. Какая корреляция называется парной?
4. Как рассчитывают коэффициент корреляции?
5. Какая корреляция называется множественной?

СПИСОК ЛИТЕРАТУРЫ

Основная

1. **Лобачев Ю.В.** Генетический анализ: Учеб. пособие / ФГОУ ВПО «Саратовский ГАУ». – Саратов, 2011. – 104 с.
2**. Смиряев А.В., Кильчевский А.В.** Генетика популяций и количественных признаков. – М.: «КолосС», 2007. – 272 с.
3. Основы научных исследований в растениеводстве и селекции: Учебное пособие / А.Ф. Дружкин, Ю.В. Лобачев, Л.П. Шевцова, З.Д. Ляшенко // ФГБОУ ВПО «Саратовский ГАУ». – Саратов, 2013. – 264 с.

Дополнительная

1. **Доспехов Б.А.** Методика полевого опыта (с основами статистической обработки результатов исследований) – 5-е издание, доп. и перераб. – М.: Агропромиздат, 1985. – 351 с.
2. **Смиряев А.В., Мартынов С.П., Кильчевский А.В.** Биометрия в генетике и селекции растений. – М.: издательство МСХА, 1992. – 269 с.
3**. Рокицкий П.Ф.** Введение в статистическую генетику. – Минск: Вышейшая школа, 1978. – 442 с.

Лекция 5

РЕГРЕССИОННЫЙ АНАЛИЗ

Коэффициент корреляции указывает на направление и степень сопряженности в изменчивости признаков, но не позволяет судить о том, как количественно меняется результативный признак при изменении факториального на единицу измерения. При постановке и решении подобных задач исследователь должен использовать регрессионный анализ, основная задача которого заключается в определении формулы корреляционной зависимости.

5.1 Парная линейная зависимость

Под парной линейной (прямолинейной) корреляционной зависимостью между двумя признаками X и Y понимают такую зависимость, которая носит линейный характер и в общем виде выражается уравнением прямой линии:
$Y = a + bX$. Это уравнение называется уравнением регрессии Y на X, а соответствующая ему прямая линия – выборочной линией регрессии У на X. Величина b – выборочный коэффициент регрессии.

В селекционно-семеноводческих экспериментах используют следующее уравнение линейной регрессии: $Y = \bar{y} + b_{yx}\times(X-\bar{x})$, где

$\bar{x}$ и $\bar{y}$ – средние арифметические для ряда X и Y;
b_{yx} – коэффициент регрессии Y по X.

Коэффициенты регрессии вычисляют по формулам:

$$b_{yx} = \frac{\sum (X-\bar{x})\times(У-\bar{y})}{\sum (X-\bar{x})^2}$$

$$b_{xy} = \frac{\sum (X-\bar{x})\times(У-\bar{y})}{\sum (У-\bar{y})^2}$$

Коэффициент регрессии b_{yx} показывает, как изменяется Y при изменении X на единицу измерения и выражается в единицах Y, а b_{xy} указывает регрессию X на Y и выражается в единицах X. Таким образом, коэффициентом линейной регрессии называется число, показывающее, в каком направлении и на какую величину изменяется в среднем признак Y (функция) при изменении признака X (аргумент) на единицу измерения. Коэффициенты регрессии имеют знак коэффициента корреляции. Произведение коэффициентов регрессии равно квадрату коэффициента корреляции: $b_{yx}\times b_{xy}=r^2$.

Ошибку коэффициента регрессии вычисляют по формулам:

$$s_{b_{yx}} = s_r \times \sqrt{\frac{\sum (У-\bar{y})^2}{\sum (X-\bar{x})^2}},$$

$$s_{b_{xy}} = s_r \times \sqrt{\frac{\sum (X-\bar{x})^2}{\sum (У-\bar{y})^2}}.$$

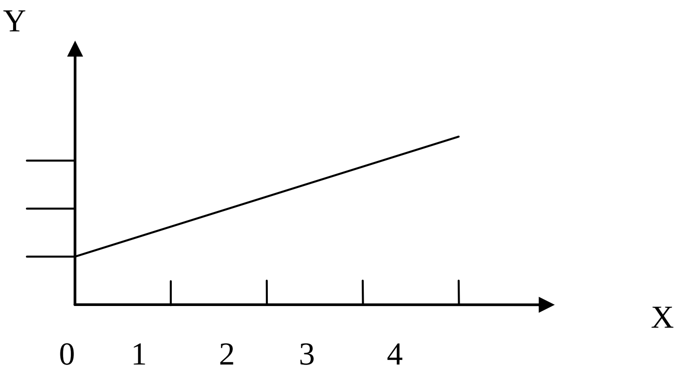
Критерий существенности коэффициента регрессии определяют по формуле:

$$t_b = \frac{b}{s_b}.$$

Существенность коэффициента регрессии оценивают по специальным таблицам, число степеней свободы принимают равным n-2. Если определен критерий существенности для коэффициента корреляции, он может быть использован и для оценки значимости коэффициента регрессии: $t_b = t_r$.

Теоретическую линию регрессии Y по X строят следующим образом: вычисляют $\bar{x}$, $\bar{y}$ и b_{yx}. Подставляя найденные значения в уравнение линейной регрессии $Y=\bar{y}+b_{yx}\times(X-\bar{x})$, определяют формулу уравнения прямой линии, которая примет общий вид Y=a+bX. По уравнению находят теоретически усредненные значения $\bar{y}_x$ для двух крайних (экстремальных) значений ряда X. Найденные точки (X_{min}, $\bar{y}_{min}$) и (X_{max}, $\bar{y}_{max}$) наносят на график и соединяют прямой линией. Это и будет теоретическая линия регрессии Y по X.

Например: Y=10+5X, $\bar{y}_{min}$=10+5×0=10, $\bar{y}_{max}$ =10+5×4=30.



В данном примере b_{yx}=5, это означает, что при возрастании значений X на одну единицу значение Y увеличивается в среднем на 5 единиц.

5.2 Множественная линейная зависимость

Математическое уравнение для прямолинейной зависимости между тремя переменными называется множественным линейным уравнением плоскости регрессии, которое имеет следующий общий вид:

$$Y=a+b_1X+b_2Z,$$

где Y – зависимая переменная,
X и Z – независимые переменные,
a – общее начало отсчета,
b_1 и b_2 – коэффициенты частной регрессии.

Коэффициент b_1 показывает на какую величину увеличивается Y при каждом увеличении на одну единицу X при постоянном значении Z; коэффициент b_2 показывает на какую величину увеличивается Y при каждом увеличении на одну единицу Z при постоянном значении X.

Параметры a, b_1, b_2 вычисляют методом наименьших квадратов. Параметры уравнения определяют по формулам:

$$b_1 = \frac{\sum (Z-\bar{z})^2 \times \sum (X-\bar{x}) \times (Y-\bar{y}) - \sum (X-\bar{x}) \times (Z-\bar{z}) \times \sum (Y-\bar{y}) \times (Z-\bar{z})}{\sum (X-\bar{x})^2 \times \sum (Z-\bar{z})^2 - \left[\sum (X-\bar{x}) \times (Z-\bar{z}) \right]^2};$$

$$b_2 = \frac{\sum (X-\bar{x})^2 \times \sum (Y-\bar{y}) \times (Z-\bar{z}) - \sum (X-\bar{x}) \times (Z-\bar{z}) \times \sum (X-\bar{x}) \times (Y-\bar{y})}{\sum (X-\bar{x})^2 \times \sum (Z-\bar{z})^2 - \left[\sum (X-\bar{x}) \times (Z-\bar{z}) \right]^2};$$

$$a = \bar{y} - b_1 \bar{x} - b_2 \bar{z}.$$

Установленное уравнением регрессии отношение зависимости коррелируемых признаков принято изображать графически в виде линий и поверхности регрессии:

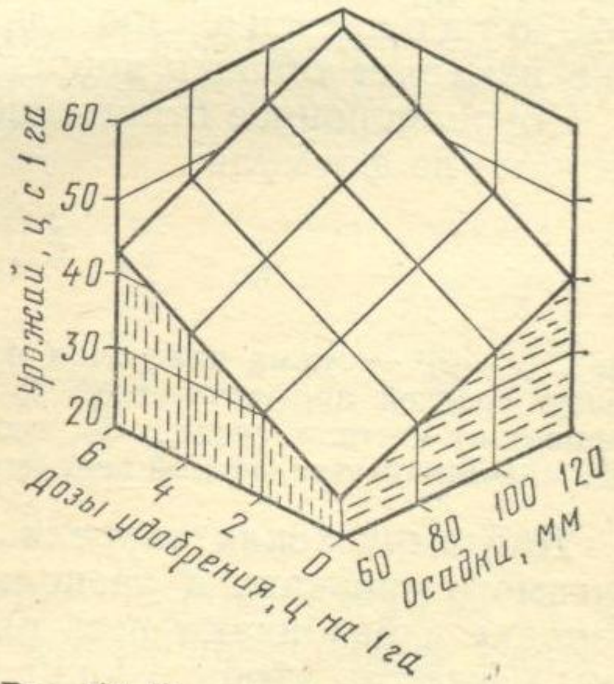
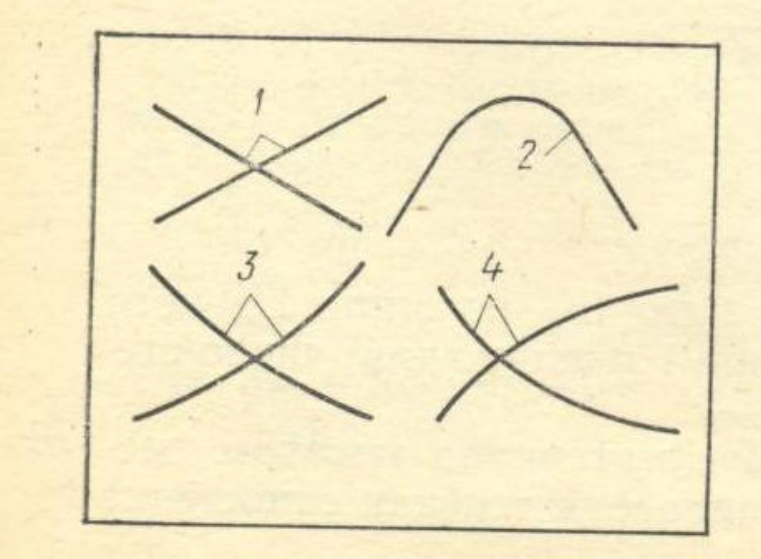


Рис. 54. Зависимость урожайности соломы льна от совместного действия осадков и доз удобрений (плоскость регрессии).

5.3 Криволинейная зависимость

Криволинейная зависимость между двумя переменными может быть выражена в виде кривой линии регрессии и соответствующего ей математического уравнения. Криволинейная регрессия – это такая зависимость, когда при одинаковых приращениях независимой переменной X зависимая переменная Y имеет неодинаковые приращения.

Эмпирические точки корреляционного поля при парной криволинейной корреляции располагаются около кривых различного типа – парабол, гипербол, логарифмических кривых и т.п. Основной метод построения математических уравнений заключается в подборе типа формулы и нахождения коэффициента к ней. Тип формулы подбирают по специальным компьютерным программам на основе типовых кривых:



При изучении множественной криволинейной зависимости, например для трех переменных, исходные данные табулируют и для нескольких фиксированных градаций аргумента X и Z определяют наиболее вероятное значение функции Y. Полученные результаты изображают графически в виде поверхности регрессии Y по X и Z, которая дает наглядное представление о форме зависимости результативного признака от совмещенного действия двух переменных:

Вопросы для самоконтроля

1. Что понимается под парной линейной зависимостью?
2. Уравнение линейной регрессии.
3. Множественное линейное уравнение плоскости регрессии.
4. Что такое криволинейная регрессия?

СПИСОК ЛИТЕРАТУРЫ

Основная

1. **Лобачев Ю.В.** Генетический анализ: Учеб. пособие / ФГОУ ВПО «Саратовский ГАУ». – Саратов, 2011. – 104 с.
2**. Смиряев А.В., Кильчевский А.В.** Генетика популяций и количественных признаков. – М.: «КолосС», 2007. – 272 с.
3. Основы научных исследований в растениеводстве и селекции: Учебное пособие / А.Ф. Дружкин, Ю.В. Лобачев, Л.П. Шевцова, З.Д. Ляшенко // ФГБОУ ВПО «Саратовский ГАУ». – Саратов, 2013. – 264 с.

Дополнительная

1. **Доспехов Б.А.** Методика полевого опыта (с основами статистической обработки результатов исследований) – 5-е издание, доп. и перераб. – М.: Агропромиздат, 1985. – 351 с.
2. **Смиряев А.В., Мартынов С.П., Кильчевский А.В.** Биометрия в генетике и селекции растений. – М.: издательство МСХА, 1992. – 269 с.
3**. Рокицкий П.Ф.** Введение в статистическую генетику. – Минск: Вышейшая школа, 1978. – 442 с.

Лекция 6

КОВАРИАЦИОННЫЙ АНАЛИЗ

6.1 Использование ковариационного анализа

Ковариационный анализ – это распространение методов дисперсионного анализа на случай нескольких переменных, а также корреляционного и регрессионного анализов на общие схемы полевых, вегетационных и лабораторных опытов. Ковариационный анализ – это одновременный анализ сумм квадратов и сумм произведений отклонений двух или более переменных от их средних. Ковариационный анализ используют для уменьшения ошибки эксперимента за счёт выравнивания сопутствующей эксперименту переменной X.

В узком смысле под ковариацией (cov) в математической статистике понимается среднее произведение отклонения двух переменных от их средних:

$$\mathrm{cov} = \frac{\sum\left(X - \bar{x}\right)\left(Y - \bar{y}\right)}{n-1}.$$

Ковариация может быть как положительной, так и отрицательной.

В широком смысле ковариацией называется совокупность трёх статистических показателей: средних арифметических $\bar{x}$ и $\bar{y}$, сумм квадратов отклонений $\sum\left(X - \bar{x}\right)^2$ и $\sum\left(Y - \bar{y}\right)^2$ и суммы произведений отклонений $\sum\left(X - \bar{x}\right)\left(Y - \bar{y}\right)$. Параллельное разложение этих величин по факторам варьирования и составляет суть ковариационного анализа.

Ковариационный анализ включает в себя три этапа:

1) дисперсионный анализ ряда *X*, *Y* и произведений *XY*;

2) разложение остаточной дисперсии C_z по ряду *Y* (остаток 1) на сумму квадратов отклонений, обусловленную регрессией *Y* по *X* (обозначаемую C_b) и сумму квадратов отклонений от регрессии $C_{d\,y \cdot x}$ (остаток 2);

3) приведение фактических средних по ряду *Y* к полной выравненности условий эксперимента по ряду сопутствующей переменной *X*.

Сумма квадратов отклонений (остаток 1), обусловленная регрессией *Y* по *X*, определяется по формуле:

$$C_b = \frac{\left[\sum\left(X - \bar{x}\right)\left(Y - \bar{y}\right)\right]^2}{\sum\left(X - \bar{x}\right)^2}.$$

Сумма квадратов отклонений от регрессии (остаток 2) находится по формуле:

$$C_{dy\cdot x} = C_z - C_b.$$

Коэффициент регрессии *Y* по *X* определяют по формуле:

$$b_{yx} = \frac{\sum\left(X - \bar{x}\right)\left(Y - \bar{y}\right)}{\sum\left(X - \bar{x}\right)^2}.$$

Выравнивание результативного признака *Y* проводится по соотношению:

$$Y_1 = Y + b_{yx}\left(\bar{x} - X\right),$$

где Y_1 – корректированное значение даты;

Y – фактическое значение даты;

b_{yx} – коэффициент регрессии *Y* по *X*;

$\bar{x} - X$ – разность между средним значением независимой переменной по опыту $\bar{x}$ и фактическим её значением X .

Ошибку отклонения от регрессии вычисляют по формуле:

$$s_{yx} = s_r \sqrt{\sum \left(Y - \bar{y}\right)^2} ,$$ где

s_r – ошибка коэффициента корреляции.

Выравнивают обычно только итоговые данные, поэтому в уравнении регрессии *Y* и *X* будут соответствовать средним по вариантам опыта.

В селекционных исследованиях ковариационный анализ целесообразно использовать для уточнения результатов опыта в двух случаях:

1) если на результативный признак может оказать заметное влияние разное исходное состояние условий эксперимента (почвенная пестрота, густота стояния растений и т.п.), которые могут быть измерены в начале опыта;

2) если на изучаемый признак в процессе эксперимента оказывают влияние не зависящие от вариантов опыта причины (повреждение растений вредителями, птицами и т.п.)

Правильное применение ковариационного анализа предполагает независимое от вариантов опыта распределение случайной величины *X*. Если сопутствующая *X* имеет отношение к изучаемым вариантам, то исключение её эффекта неправомерно, так как это ведёт к исключению части эффекта варианта.

Внесение поправок в результаты эксперимента проводят следующим образом. Вычисляют величину F-критерия путём деления среднего квадрата для регрессии на дисперсию остатка 2. Если фактическое значение F-критерия больше теоретического значения, то связь между *Y* и *X* не случайна и можно проводить корректирование результатов эксперимента. Если $F_{факт.} < F_{теор.}$, то введение поправок бесполезно, это не приведёт к уточнению эксперимента. Поправки вводят следующим образом (например, для урожая). К значениям урожая делянок, которые по данным предварительного учёта оказались ниже среднего уровня, прибавляют величину поправки, равную $b_{yx}\left(\bar{x} - X\right)$, а если их урожаи превышали средний уровень, то поправку вычитают. Таким образом, получают корректированные средние урожаи по вариантам опыта.

6.2 Принципиальная схема ковариационного анализа

Таблица 1

Исходные данные

Варианты		Повторения				Суммы V_x и V_y	Средние
		I	II	III	IV		
1	X	*	*	*	*	*	*
	Y	*	*	*	*	*	*
2	X	*	*	*	*	*	*
	Y	*	*	*	*	*	*
3	X	*	*	*	*	*	*
	Y	*	*	*	*	*	*
4	X	*	*	*	*	*	*
	Y	*	*	*	*	*	*
5	X	*	*	*	*	*	*
	Y	*	*	*	*	*	*
Суммы	P_x	*	*	*	*	$\sum X =*$	$\bar{x} =*$
	P_y	*	*	*	*	$\sum Y =+$	$\bar{y} =*$

* – значения показателя;

Таблица 2

Формулы для вычисления сумм квадратов отклонений и произведений

Дисперсия	Суммы квадратов и произведений		
	$\sum(X-\bar{x})^2$	$\sum(X-\bar{x})(Y-\bar{y})$	$\sum(Y-\bar{y})^2$
Общая (C_y)	$\sum X^2 - C$	$\sum XY - C$	$\sum Y^2 - C$
Повторений (C_p)	$\sum P_x^2 / l - C$	$\sum P_x \cdot P_y / l - C$	$\sum P_y^2 / l - C$
Вариантов (C_v)	$\sum V_x^2 / n - C$	$\sum V_x \cdot V_y / n - C$	$\sum V_y^2 / n - C$
Остаток (C_z)	$C_y - C_p - C_v$	$C_y - C_p - C_v$	$C_y - C_p - C_v$
	$C = \left(\sum X\right)^2 / N$	$C = \left(\sum X\right)\left(\sum Y\right) / N$	$C = \left(\sum Y\right)^2 / N$

где $N = l \cdot n$, l – число вариантов (=5), n – число повторений (=4)

Таблица 3

6.3 Результаты ковариационного анализа

Дисперсия	Суммы квадратов и произведений			Степени свободы	Коэффициент регрессии	Средний квадрат	$F_{факт.}$	F_{005}
	$\sum(X-\bar{x})^2$	$\sum(X-\bar{x})(Y-\bar{y})$	$\sum(Y-\bar{y})^2$		$b_{yx} = \frac{\sum(X-\bar{x})(Y-\bar{y})}{\sum(X-\bar{x})^2}$			
Общая	*	*	*	19	-	-	-	-
Повторений	*	*	*	3	-	-	-	-
Вариантов	*	*	*	4	-	*	*	3,36
Остаток 1	*	*	*	12	-	*	-	-
Регрессия $C_в$	-	-	*	1	*	*	*	4,84
Остаток 2	-	-	*	11	-	*	-	-

* – значения показателя;

l=5; n=4;

$F_{факт.\ вар.} = \frac{Средний\ квадрат\ для\ вариантов}{Средний\ квадрат\ остатка\ 2}$;

$F_{факт.\ рег.} = \frac{Средний\ квадрат\ для\ регрессии}{Средний\ квадрат\ остатка\ 2}$;

$C_в = \frac{\left[\sum(X-\bar{x})(Y-\bar{y})\right]^2}{\sum(X-\bar{x})^2}$;

Остаток 2 = остаток 1-$C_в$

Таблица 4

Коррекция урожая

Варианты	X	$\bar{x}-X$	$b_{yx}(\bar{x}-X)$	Урожай	
				Фактический Y	Корректированный $Y_1 = Y + b_{yx}(\bar{x}-X)$
1	*	*	*	*	*
2	*	*	*	*	*
3	*	*	*	*	*
4	*	*	*	*	*
5	*	*	*	*	*
Средние значения	$\bar{x}=*$	*	*	$\bar{y}=*$	$\bar{y}_1=*$

* – значения показателя.

Для оценки существенности частных различий вычисляют:

$$S_{\bar{y}} = \sqrt{\frac{S^2_{остатка2}}{n}},$$ где $S^2_{остатка2}$ – средний квадрат остатка 2;

$$S_d = \sqrt{\frac{2S^2_{остатка2}}{n}};$$

$$\text{ÍÑÐ}_{05} = t_{05} \times S_d.$$

Все разности между средними по вариантам, превышающие НСР, существенны на 5%-ом уровне значимости.

Вопросы для самоконтроля

1. Использование ковариационного анализа.
2. Положительная и отрицательная ковариация.
3.Этапы ковариационного анализа.

СПИСОК ЛИТЕРАТУРЫ

Основная

1. **Лобачев Ю.В.** Генетический анализ: Учеб. пособие / ФГОУ ВПО «Саратовский ГАУ». – Саратов, 2011. – 104 с.
2**. Смиряев А.В., Кильчевский А.В.** Генетика популяций и количественных признаков. – М.: «КолосС», 2007. – 272 с.
3. Основы научных исследований в растениеводстве и селекции: Учебное пособие / А.Ф. Дружкин, Ю.В. Лобачев, Л.П. Шевцова, З.Д. Ляшенко // ФГБОУ ВПО «Саратовский ГАУ». – Саратов, 2013. – 264 с.

Дополнительная

1. **Доспехов Б.А.** Методика полевого опыта (с основами статистической обработки результатов исследований) – 5-е издание, доп. и перераб. – М.: Агропромиздат, 1985. – 351 с.
2. **Смиряев А.В., Мартынов С.П., Кильчевский А.В.** Биометрия в генетике и селекции растений. – М.: издательство МСХА, 1992. – 269 с.

Лекция 7

ПРО́БИТ-АНАЛИЗ

7.1 Статистический метод – про́бит-анализ

В селекционно-генетических исследованиях при изучении влияния поражающих факторов (излучения, мутагены, пестициды и т.п.) на биологические организмы используется специальный статистический метод – про́бит-анализ.

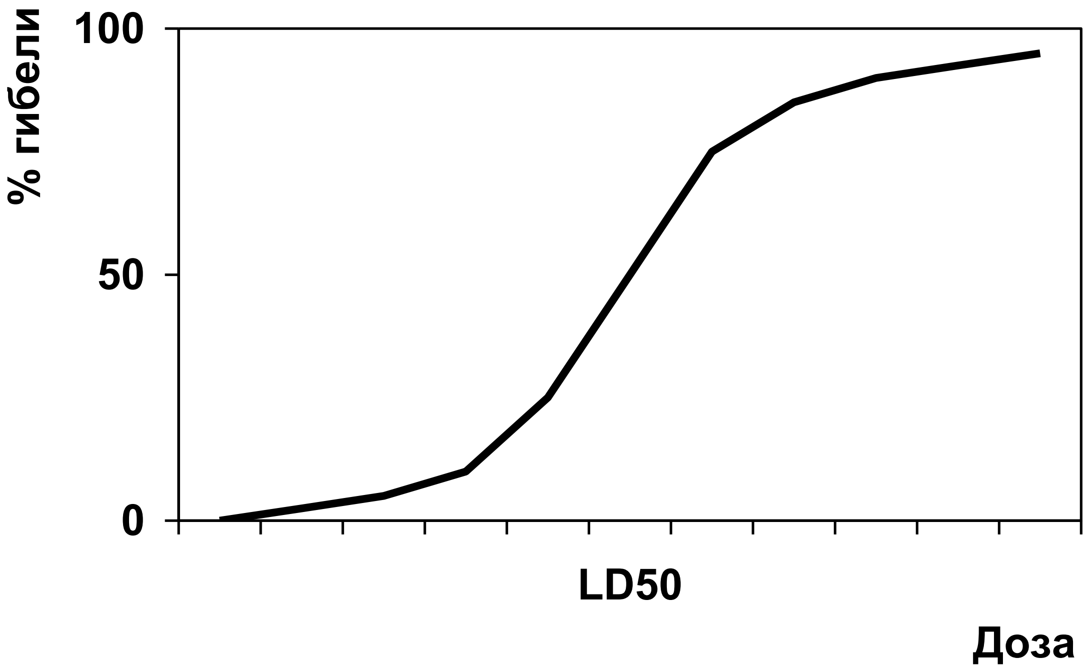
Часто в экспериментах требуется определить и измерить количественно летальную дозу (LD) вещества, приводящую к 100%-ной гибели биологического организма. Измерение такой дозы на одном единственном объекте невозможно по следующим причинам:

1) биологический объект состоит из многих особей, которые различаются генотипически друг от друга, в том числе и по устойчивости к изучаемому фактору;
2) заранее исследователь не знает: какую дозу вещества применить, чтобы получить 100%-ную гибель. Поэтому приходится изучать разные режимы (концентрации) поражающего фактора. Но сделать это на одном и том же организме нельзя, т. к. даже небольшие дозы могут вредно сказываться на жизнеспособности организма, а с другой стороны иногда отмечается реакция «привыкания» организма к поражающему фактору;
3) для подобных экспериментов требуется и большое количество объектов и большое количество испытуемого фактора (вещества), что не всегда бывает оправдано с точки зрения затрат на эксперименты.

Поэтому часто достаточно определить дозу, при которой погибает 50% особей, которую и принимают за усреднённую характеристику летального действия повреждающего фактора и обозначают LD_{50}.

Критерий LD_{50}, показывающий, какая доза препарата необходима для данной популяции, чтобы вероятность гибели особей составила 50%, определяют статистическим путём. Для этого всю изучаемую совокупность разбивают на несколько групп. На каждую группу особей, состоящую из большого числа (несколько десятков) воздействуют препаратом, причём на первую группу – в минимальной дозе, на вторую – более возрастающей дозой и т.д. На последнюю группу воздействуют препаратом в максимальной дозе. В итоге получают статистический ряд, в котором гибель биологических объектов увеличивается с повышением дозы препарата.

По результатам многочисленных экспериментов с химическими веществами и различными излучениями установили, что зависимость между дозой вещества и процентом погибших особей носит сложную S-образную форму:



Эта кривая несимметрична, крутизна изгиба нижней и верхней части кривой неодинакова. Математическая обработка данных, представленных графически в виде этой кривой очень сложна. Для того, чтобы сделать возможной линейную интерполяцию и графически определять любую дозу смертности (LD_{50}, LD_{90}, LD_{99}, кроме 0 и 100%), нужно подобрать такое преобразование осей координат, после которого график будет представлен прямой линией. Самое простое решение – трансформировать S-образную кривую в прямую линию. Это и составляет суть пробит-анализа. Для этого на оси абсцисс откладывают логарифмы доз изучаемого фактора, а по оси ординат вместо процента гибели подопытных особей откладывают условные вероятные величины, называемые пробитами (probability unit). Значение пробит, соответствующее конкретному проценту гибели особей, находят по специальной таблице.

Гибель, %	0	1	2	3	4	5	6	7	8	9
0	-	2,67	2,95	3,12	3,25	3,36	3,45	3,52	3,59	3,66
10	3,72	3,77	3,82	3,87	3,92	3,96	4,01	4,05	4,08	4,12
20	4,16	4,19	4,23	4,26	4,29	4,33	4,36	4,39	4,42	4,45
30	4,48	4,50	4,53	4,56	4,59	4,61	4,64	4,67	4,69	4,72
40	4,75	4,77	4,80	4,82	4,85	4,87	4,90	4,92	4,95	4,97
50	5,00	5,03	5,05	5,08	5,10	5,13	5,15	5,18	5,20	5,23
60	5,25	5,28	5,31	5,33	5,36	5,39	5,41	5,44	5,47	5,50
70	5,52	5,55	5,58	5,61	5,64	5,67	5,71	5,74	5,77	5,81
80	5,84	5,88	5,92	5,95	5,99	6,04	6,08	6,13	6,18	6,23
90	6,28	6,34	6,41	6,48	6,55	6,64	6,75	6,88	7,05	7,33

Описанный метод относится к категории простейших модификаций системы про́битов. Он позволяет лишь приблизительно оценить LD_{95} и LD_{99}, по нему нельзя рассчитать доверительные интервалы этих значений. Однако на точности определения LD_{50} это сказывается незначительно.

7.2 Алгоритм проведения про́бит-анализа

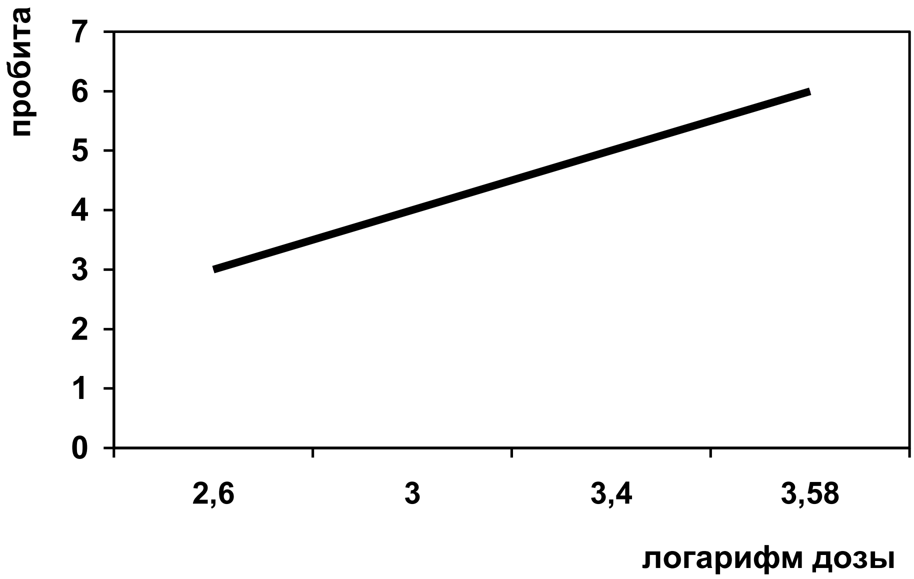
1) Результаты опыта по испытанию действия вещества на биологический объект заносят в таблицу 1.

Таблица 1

Доза (концентрация, мг/л) D	Средняя гибель (%)	Преобразованные значения доз и процента гибели	
		lg D (ось X)	пробиты (ось Y)
250	15	2,398	3,96
1000	39	3,000	4,72
…	…	…	…
8000	66	3,903	5,41

Проводят преобразование дозы вещества в их логарифмы (по специальным таблицам или на логарифмической линейке), а проценты гибели преобразуют по специальной таблице в пробиты.

2) Строят график зависимости «эффект – доза». Для этого на оси абсцисс откладывают логарифмы доз препарата, а по оси ординат – значения пробит. Через найденные точки проводят прямую линию, которая путём интерполяции (продолжения) позволяет определить LD_{50}, LD_{90} или LD_{99}.



3) После построения графика находят на нём значение 50% гибели (пробит равен 5,00) и определяют логарифм дозы (lg LD_{50}). Затем по антилогарифмам находят дозу LD_{50}, которая составляет… мг/л вещества. Аналогично определяют концентрацию для LD_{90} или

LD_{99}. Антилогарифмы вычисляют на логарифмической линейке. Результаты всех определений заносят в таблицу:

Летальные дозы	Концентрация вещества, мг/л
LD_{50}	…
LD_{90}	…
LD_{99}	…

На этом пр<u>о́</u>бит-анализ завершается.

Вопросы для самоконтроля

1. Использование про́бит-анализа.
2. Алгоритм проведения про́бит-анализа.

СПИСОК ЛИТЕРАТУРЫ

Основная

1. **Лобачев Ю.В.** Генетический анализ: Учеб. пособие / ФГОУ ВПО «Саратовский ГАУ». – Саратов, 2011. – 104 с.
2**. Смиряев А.В., Кильчевский А.В.** Генетика популяций и количественных признаков. – М.: «КолосС», 2007. – 272 с.
3. Основы научных исследований в растениеводстве и селекции: Учебное пособие / А.Ф. Дружкин, Ю.В. Лобачев, Л.П. Шевцова, З.Д. Ляшенко // ФГБОУ ВПО «Саратовский ГАУ». – Саратов, 2013. – 264 с.

Дополнительная

1. **Доспехов Б.А.** Методика полевого опыта (с основами статистической обработки результатов исследований) – 5-е издание, доп. и перераб. – М.: Агропромиздат, 1985. – 351 с.
2. **Смиряев А.В., Мартынов С.П., Кильчевский А.В.** Биометрия в генетике и селекции растений. – М.: издательство МСХА, 1992. – 269 с.
3**. Рокицкий П.Ф.** Введение в статистическую генетику. – Минск: Вышейшая школа, 1978. – 442 с.

Лекция 8

ОПРЕДЕЛЕНИЕ КОЭФФИЦИЕНТА НАСЛЕДУЕМОСТИ

8.1 Коэффициент наследуемости признака

В селекционно-генетических исследованиях экспериментатор имеет дело с признаками организма. Считается, что признак формируется под влиянием генотипа и среды. Поэтому важно знать, в какой мере проявление признака зависит от генотипа. Для этого вычисляют коэффициент наследуемости признака, который показывает долю генетической изменчивости в общей фенотипической изменчивости. Поскольку генотипическая дисперсия включает в себя аддитивный, доминантный и эпистатический компоненты, то Т. Лаш (1949) предложил различать наследуемость в широком (H^2) и узком (h^2) смысле:

$$H^2 = \frac{S_г^2}{S_ф^2};$$

$$h^2 = \frac{S_A^2}{S_ф^2}, \text{ где}$$

$S_г^2$ - генотипическая дисперсия;

$S_ф^2$ - фенотипическая дисперсия;

S_A^2 - аддитивная генотипическая дисперсия.

Коэффициент наследуемости в широком смысле есть доля генотипической, а в узком – доля аддитивной дисперсии в общей фенотипической.

Коэффициент наследуемости измеряется в процентах или долях единицы. Например, H^2=0,65, это означает, что 65% фенотипической изменчивости изучаемого признака обусловлено наследственной изменчивостью растений и отбор по этому признаку в популяции должен быть эффективным.

Коэффициент наследуемости в узком смысле можно использовать для прогноза эффективности массового отбора:

$R = h^2 \cdot S$, где

R – ответ на селекцию (генетический сдвиг);

S – селекционный дифференциал (разность между средним значением признака отобранной части популяции и популяционной средней).

Известно несколько методов вычисления коэффициента наследуемости в широком смысле. Наибольшую известность получили способы определения H^2 при помощи коэффициента корреляции или коэффициента регрессии между фенотипами родственных групп, а также с помощью дисперсионного анализа, позволяющего разложить фенотипическую изменчивость ($S_ф^2$) на составляющие её компоненты: дисперсию генотипическую ($S_г^2$) и паратипическую (S_h^2).

Оценка наследуемости с помощью анализа родственных связей имеет существенные ограничения:

1) в популяции должно иметь место аддитивное наследование признака (внутри- и межлокусные взаимодействия генов должны отсутствовать);

2) расщепляющаяся популяция должна находиться в состоянии равновесия (в популяции не должен происходить отбор).

В настоящее время отсутствуют тесты для проверки этих ограничений.

8.2 Алгоритмы вычисления коэффициента наследуемости в широком смысле (H^2) при помощи коэффициентов корреляции и регрессии.

1) Исходные данные:

Номер пары	X (материнская форма)	Y (гибрид)	X^2	Y^2	$X \times Y$
1	*	*	*	*	*
…	*	*	*	*	*
n	*	*	*	*	*
$\sum n=*$	$\sum X=*$	$\sum Y=*$	$\sum X^2=*$	$\sum Y^2=*$	$\sum X \times Y=*$

*– значения показателя.

2) Вычисляют $\sum X$, $\sum Y$, $\sum X^2$, $\sum Y^2$, $\sum X \times Y$, коэффициент корреляции (r), коэффициент регрессии (b_{xy}) и критерий $\mathbf{t_r}$.

$$r = \frac{\sum X \times Y - \left(\sum X \times \sum Y\right)/n}{\sqrt{\left(\sum X^2 - \left(\sum X\right)^2/n\right) \times \left(\sum Y^2 - \left(\sum Y\right)^2/n\right)}};$$

$$b_{YX} = \frac{\sum X \times Y - \left(\sum X \times \sum Y\right)/n}{\sum X^2 - \left(\sum X\right)^2/n};$$

$$S_r = \sqrt{\frac{1-r^2}{n-2}};$$

$$t_r = \frac{r}{S_r};\ V = n-2.$$

По t-критерию оценивают значимость корреляции и регрессии для 5%-го уровня. Если нулевая гипотеза отвергается, то вычисляют H^2.

3) Коэффициент наследуемости вычисляют исходя из теоретических допущений о том, что H^2 равен удвоенному коэффициенту корреляции или регрессии между фенотипами родителей и потомством:

$H^2 = 2r$ или $H^2 = 2byx$.

Считается, что H^2, вычисленный по коэффициенту регрессии Y по X, является более точной величиной.

Если оперируют со средними значениями признака у родителей, то $H^2 = byx$, если учитывается проявление признака только одного родителя, то $H^2 = 2byx$.

8.3 Алгоритм вычисления коэффициента наследуемости в широком смысле (H^2) при помощи дисперсионного анализа

1) Исходные данные:

Название гибрида		Повторения, X				Суммы V	Средние
материнская форма	отцовская форма	1	2	3	4		
1	1	*	*	*	*	*	*
1	2	*	*	*	*	*	*
1	3	*	*	*	*	*	*
Суммы P		*	*	*	*	$\sum X=*$	*

* – значение показателя.

2) Проводят дисперсионный анализ для однофакторного комплекса:

$N = l \times n = 3 \times 4 = 12$

$C = \left(\sum X\right)^2 / N$

$C_Y = \sum X^2 - C$

$C_P = \sum P^2 / l - C$

$C_V = \sum V^2 / n - C$

$C_Z = C_Y - C_P - C_V$

3) Полученные результаты записывают в таблицу дисперсионного анализа и определяют значимость действия генотипов на фенотипическую изменчивость признака по F-критерию:

Дисперсия	Сумма квадратов	Степени свободы	Средний квадрат	$F_{факт.}$	$F_{теор.}$
Общая	*	*N-1*	-	-	-
Повторений	*	*n-1*	-	-	-
Отцовских форм	*	*l-1*	*	*	*
Остаток	*	(*n-1*)×(*l-1*)	*	-	-

* – значение показателя.

Если $F_{факт.} > F_{теор.}$, то вычисляют коэффициент наследуемости, характеризующий степень генетической изменчивости.

Дисперсия групповых средних имеет сложную природу. Она определяется как генотипической изменчивостью признака ($S_г^2$) так и случайной, паратипической изменчивостью (S_h^2). Таким образом, средний квадрат для вариантов (отцовских форм) S_v^2 состоит из двух компонентов: $S_v^2 = S_h^2 + n \cdot S_{\tilde{a}}^2$. Здесь *«n»* поставлено перед генотипической изменчивостью групповых средних для того, чтобы привести её к уровню первичных наблюдений.

4) Проводят вычисления генотипической и фенотипической изменчивости признака:

$S_г^2 = \dfrac{S_v^2 - S_h^2}{n}$;

$S_ф^2 = S_г^2 + S_h^2$, где

S_v – средний квадрат для вариантов;
S_h – средний квадрат для остатка.

5) Проводят вычисления коэффициента наследуемости по результатам дисперсионного анализа:

$$H^2 = \frac{S_г^2}{S_ф^2}$$

(вычисленный коэффициент наследуемости характеризует степень передачи селекционного признака от отцовской формы к гибриду).

Для вычисления коэффициента наследуемости в узком смысле (h^2) необходимо дополнительно разложить генотипическую дисперсию на составляющие её компоненты и использовать аддитивную дисперсию:

$$h^2 = \frac{s_A^2}{s_\phi^2}.$$

СПИСОК ЛИТЕРАТУРЫ

Основная

1. **Лобачев Ю.В.** Генетический анализ: Учеб. пособие / ФГОУ ВПО «Саратовский ГАУ». – Саратов, 2011. – 104 с.
2**. Смиряев А.В., Кильчевский А.В.** Генетика популяций и количественных признаков. – М.: «КолосС», 2007. – 272 с.
3. Основы научных исследований в растениеводстве и селекции: Учебное пособие / А.Ф. Дружкин, Ю.В. Лобачев, Л.П. Шевцова, З.Д. Ляшенко // ФГБОУ ВПО «Саратовский ГАУ». – Саратов, 2013. – 264 с.

Дополнительная

1. **Доспехов Б.А.** Методика полевого опыта (с основами статистической обработки результатов исследований) – 5-е издание, доп. и перераб. – М.: Агропромиздат, 1985. – 351 с.
2. **Смиряев А.В., Мартынов С.П., Кильчевский А.В.** Биометрия в генетике и селекции растений. – М.: издательство МСХА, 1992. – 269 с.
3**. Рокицкий П.Ф.** Введение в статистическую генетику. – Минск: Вышейшая школа, 1978. – 442 с.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. **Лобачев Ю.В.** Генетический анализ: Учеб. пособие / ФГОУ ВПО «Саратовский ГАУ». – Саратов, 2011. – 104 с.

2**. Смиряев А.В., Кильчевский А.В.** Генетика популяций и количественных признаков. – М.: «КолосС», 2007. – 272 с.

3. Основы научных исследований в растениеводстве и селекции: Учебное пособие / А.Ф. Дружкин, Ю.В. Лобачев, Л.П. Шевцова, З.Д. Ляшенко // ФГБОУ ВПО «Саратовский ГАУ». – Саратов, 2013. – 264 с.

4. **Доспехов Б.А.** Методика полевого опыта (с основами статистической обработки результатов исследований) – 5-е издание, доп. и перераб. – М.: Агропромиздат, 1985. – 351 с.

5. **Смиряев А.В., Мартынов С.П., Кильчевский А.В.** Биометрия в генетике и селекции растений. – М.: издательство МСХА, 1992. – 269 с.

6**. Рокицкий П.Ф.** Введение в статистическую генетику. – Минск: Вышейшая школа, 1978. – 442

СОДЕРЖАНИЕ