ФГБОУ ВО НОВОСИБИРСКИЙ ГАУ

Агрономический факультет

Кафедра селекции, генетики и лесоводства

СТАТИСТИЧЕСКИЙ АНАЛИЗ В АГРОНОМИИ

Методическое пособие
для практических занятий и самостоятельной работы

Новосибирск 2016

УДК 578. 087.1
ББК

Кафедра селекции, генетики и лесоводства

Составитель: канд. с.-х. наук, доц. *И.В. Кондратьева,*
д-р биол. наук, проф. *М.Л. Кочнева*
д-р биол. наук, проф. *Р.А Цильке.*
Рецензент: д-р с.- х. наук, проф. *Р.Р. Галеев*

Статистический анализ в агрономии: метод. пособие / Новосиб. гос. аграр. ун-т. Агроном. фак-т; сост. И.В. Кондратьева, М.Л. Кочнева, Р.А. Цильке. – Новосибирск, 2016. – 54 с.

Методическое пособие предназначено для практических занятий и самостоятельной работы по дисциплине «Статистический анализ в агрономии» для студентов очного и заочного обучения по направлению подготовки 35.03.04 Агрономия.

Утверждены и рекомендованы к изданию учебно-методическим советом агрономического факультета (протокол № 1 от 29.01. 2016 г.).

ВВЕДЕНИЕ

Роль статистических методов в агрономии

Основная цель курса «Статистический анализ в агрономии» - сформировать научное мировоззрение о математических основах описания биологических процессов, о применении методов математической статистики в агрономии. Полученные знания дают необходимую основу для проведения научных исследований.

Явления жизни, как и вообще все явления материального мира, имеют две неразрывно связанные стороны: качественную, воспринимаемую непосредственно органами чувств, и количественную, выражаемую числами при помощи счета и меры.

При исследовании различных явлений природы применяют одновременно и качественные, и количественные показатели. Количественные методы как более объективные и точные имеют преимущество перед качественной характеристикой предметов. Недаром еще в древности достоверность познания природы связывалась с математикой - наукой точной, изучающей количественные отношения и пространственные формы реальной действительности. Опираясь на количественные показатели, можно получить более достоверную информацию о предметах, что позволяет глубже постигнуть их качественное своеобразие.

Количественные методы не ограничиваются одними лишь змерениями или учетом живых существ и продуктов их жизнедеятельности. Сами по себе результаты измерений, хотя и имеют известное значение, еще недостаточны для того, чтобы сделать из них необходимые выводы.

Цифровые данные, собранные в процессе массовых испытаний, т.е. измерений или учета изучаемого объекта, - это всего лишь сырой фактический материал, который нуждается в соответствующей математической обработке.

Без обработки - упорядочения и систематизации цифровых данных - не удается извлечь заключенную в них информацию, оценить надежность отдельных суммарных показателей, убедиться в достоверности или недостоверности наблюдаемых между ними различий. Эта работа требует от специалистов определенных знаний, умения правильно обобщать и анализировать собранные в опыте данные. Система этих знаний и составляет содержание биометрии - науки, занимающейся главным образом вопросами статистического анализа результатов исследований как в области теоретической, так и прикладной биологии, в том числе и в агрономии.

Занятие 1

Тема: Совокупности. Группировка данных выборочной совокупности

Совокупностью называют множество сходных, в некотором смысле однородных объектов, на которых производят идентичные измерения (наблюдения). Например, совокупность растений определенного вида или совокупность животных, предназначенных для эксперимента, т.д. Совокупность состоит из единиц совокупности или членов. Совокупности делятся на генеральные и выборочные.

Генеральная совокупность т.е. бесконечное множество однородных, но индивидуально различимых объектов. Предположим, мы поставили цель изучить изменчивость массы тела конкретного вида животных, тогда генеральной совокупностью будут данные о массе всех без исключения особях этого вида. В теоретических рассуждениях n генеральной совокупности стремится к бесконечности.

В генеральных совокупностях, используемых в биологии количество единиц, как правило, настолько велико, что оценить состояние всех ее членов практически невозможно (или нецелесообразно), поэтому объектами изучения являются только части генеральной совокупности. Так при изучении природных популяций в подавляющем большинстве случаев мы не сможем иметь дело с генеральной совокупностью в её полном объёме, а будем располагать лишь какой-то, как

правило, очень небольшой, её частью. Отобранная для исследования часть генеральной совокупности получила название **выборочной совокупности** или просто выборки. Выборка должна быть как бы уменьшенной копией генеральной совокупности, т. е. правильно, без искажений представлять всю генеральную совокупность. Поэтому основное требование, предъявляемое к любой выборке, это её представительность или **репрезентативность** (от лат. represento - представляю). Выборка считается репрезентативной, если она получена путём случайного отбора, поскольку, если для анализа мы будем отбирать только самые крупные особи, то они будут характеризовать лишь собранный материал, но никак не генеральную совокупность. На основании репрезентативной выборки исследователь и формирует представление о свойствах генеральной совокупности.

Однако следует помнить, что и репрезентативность, и точность случайного выбора носят вероятностный характер.

Репрезентативная выборка:

а) случайный выбор

б) достаточная численность.

Традиционно в биометрии сумму членов генеральной совокупности обозначают буквой N, а число наблюдений, образующих выборку - буквой n.

При изучении тех или иных объектов исследователь имеет дело с признаками, проявлением которых один предмет отличается от другого. Примерами признаков могут служить - масса и длина тела, окраска и число яиц в гнезде и т.п

Переменные величины принято обозначать прописными латинскими буквами X, Y, Z, а их варианты – строчными буквами (x1, x2, x3 ... xn).

Важным свойством признаков является варьирование величины признака при переходе от одной единицы наблюдения к другой. Отдельные числовые значения варьирующего признака называются вариантами (лат. *variatio* - изменение) и обозначают, как правило, маленькими буквами латинского алфавита с цифровыми индексами: x1, x2, x3, …, xn.). В общем виде значения варианты отмечают символом xi. Если некоторое i-ое значение варианты при измерениях встретилось ni раз, то ni называют частотой.

Все признаки можно разделить на две группы – качественные и количественные. Типичные качественные признаки - окраска, пол, возраст, наличие заболевания и т.п. К примеру, в городской популяции сизых голубей можно выделить особей с типичной “сизой” окраской, черночеканных, альбиносов, меланистов, пегих, красных. В таком случае можно без специальных измерений достаточно определенно судить о наличии или отсутствии того или иного признака у конкретной особи.

Количественные признаки, в отличие от качественных, можно анализировать лишь на основании специальных измерений или подсчетов. Их подразделяют на **континуальные** (непрерывная изменчивость) и **дискретные** (прерывистая изменчивость).

Континуальные признаки теоретически могут принимать любые возможные значения в пределах между минимальным и максимальным показателем признака. К ним относятся масса, линейные размеры и температура организма, содержание биохимических и неорганических веществ в его тканях и т.п. В качестве примера дискретных признаков можно привести число глазков на клубне картофеля, пятен на надкрыльях жука, яиц в гнезде. Отсюда следует, что непосредственно наблюдаемые значения дискретного признака могут характеризоваться лишь целыми числами, тогда как при континуальном варьировании значения признака могут быть как целыми, так и дробными. Иногда вместо непосредственно измеренных значений количественного признака используют присвоенные им ранги или баллы. В таких случаях количественный признак называется **ранжированным**.

1. Группировка данных выборочной совокупности по признакам с дискретной изменчивостью.

Способ упорядочения данных, которые записаны в порядке нарастания или в порядке уменьшения величины называется **ранжированием.** При этом строго обозначены минимальное и максимальное значения вариант, которые носят название **лимитов изменчивости** – min - max. В центре такого упорядоченного (ранжированного) ряда цифр сосредоточено основное количество вариант со средним значением признака.

Ранжирование применяется при любом характере изменчивости при небольших размерах выборки (n < 30).

Задание. На каждой из 10 выбранных делянок регистрировалось число деревьев определенного вида: 13 15 13 15 14 16 14 12 14 14 15.

Составьте упорядоченный ряд. Определите тип совокупности признака.

Другой способ упорядоченья вариант совокупности состоит в записи значений вариант в порядке нарастания и определения количества раз встречаемости каждой цифры. Различные типы вариант распределения называются **классами** (x_i), а числа, соответствующие каждому классу **частотами** (f). Такой ряд чисел, состоящий из классов и частот, называется **вариационным рядом**. Таблицу, отображающую соответствие между значениями вариант (xi) и их частотами (ni) называют вариационным рядом.

Пример. Подсчитано число продуктивных стеблей на растение у сорта Кантегирская 89. Получены следующие данные: 6, 7, 8, 12, 5, 4, 9, 7, 15, 8, 8, 6, 6, 5, 13, 7, 8, 10, 8, 7, 14, 6, 5, 7, 5, 7, 8, 10, 8, 7, 6, 8, 7, 9, 7 , 7, 9, 7, 6, 11.

Записываем в табл. 1 значения класса и разносим все варианты по классам.

Таблица 1

Вариационный ряд

Классы, x_i	Частоты, f
4	1
5	4
6	6
7	11
8	8
9	3
10	2
11	1
12	1
13	1
14	1
15	1
	n = $\sum f$ = 40

Картина закономерностей распределения следующая:

1. В центре вариационного ряда сосредоточено основное число значений признака.
2. Чем больше значения вариант отклоняются от значений вариант, находящихся в середине ряда, тем меньше таких вариант встречается в совокупности.

Класс, в котором представлено больше всего вариант, называется **модальным** классом. Если распределение симметрично, то значение модального класса равно среднему значению анализируемого признака в данной совокупности.

Такая группировка признаков с дискретной изменчивостью возможна тогда, когда признак представлен небольшим число естественных классов.

Если дискретный количественный признак имеет большой размах изменчивости и, следовательно, большое число классов, то получится растянутое распределение и такой вариационный ряд сложно обсчитывать.

Когда естественных классов больше 6-7, то можно организовать искусственные классы.

Для составления вариационного ряда необходимо:

1. Найти в учетах данных максимальное (max) и минимальное (min) значения признака. Разница между максимальным и минимальным значениями признака (варианта) – это **размах изменчивости признака** (lim = max–min).

2. Исходя из объема выборки и размаха изменчивости, выбрать оптимальное число классов (k) для проведения группировки.

Таблица 2

Выбор числа интервалов (классов) согласно эмпирически выработанным рекомендациям:

Объем выборки, n	Число классов, k
25 - 40	5 – 6
40 – 60	6 – 8
60 – 100	7 – 10
100 - 200	8 – 12
> 200	10 – 15

3. На основании выбранного количества классов и размаха изменчивости признака установить величину классового промежутка (i), т.е.

величину, на которую один класс должен отличаться от другого:

$$i = \frac{\lim}{k} = \frac{\max - \min}{k}$$

4. Определение границ интервалов.

Началом первого класса обычно служит варианта с минимальным значением признака, концом первого класса – величина, равная началу первого класса, увеличенному на классовый промежуток (i). Конец последнего класса завершается максимальным значением варианты. Конец предыдущего и начало следующего классов не должны совпадать. Они должны отличаться или на целое число, или на десятые или сотые доли числа, в зависимости от величины изучаемого признака.

Пример.

Подсчитано количество зерен в колосе сорта мягкой яровой пшеницы Кантегирская 89:

54 45 44 44 49 47 48 53 55 47

61 59 52 52 44 47 39 50 34 57

47 37 46 31 45 59 43 44 43 42

48 37 46 32 59 43 37 44 43 42

Задание. Построить вариационный ряд по числу зерен в колосе.

Выполнение задания

1. Используя данные таблицы 2, определяем число искусственных классов – n = 40, количество классов k = 7.

2. Величина классового интервала:

$$i = \frac{\lim}{k} = \frac{\max - \min}{k}$$

$$\text{i} = \frac{61-31}{6} = 5$$

3. Определение границ интервалов и подсчет (разноска) частот наблюдений для каждого интервала. Установленные для нашего примера границы классов заносятся в табл.3.

Таблица 3

Группировка данных по числу зерен колоса

Классы, x_{min} - x_{max}	Среднее значение признака	Разноска вариант	Частоты, f
31- 35	33	▪ ▪ ▪	3
36 - 40	38		4
41 - 45	43		13
46 - 50	48		10
51 - 55	53		5
56 - 60	58		4
61 - 65	63		1
i = 5			n = 40

После распределения вариант по классам (таб. 3) получаем вариационный ряд, который показывает, как часто встречаются варианты каждого класса.

2. Группировка данных выборочной совокупности по признакам с непрерывной изменчивостью

Задание 1.

Получены следующие показатели длины зерна пшеницы (мм):

5, 39 5,42 5,38 5,47 5,51 5,30 5,40 5,40 5,40 5,28 5,43

Составьте упорядоченный ряд.

Задание 2.

Определена масса зерна колоса у сорта Кантегирская 89 (гр.). Построить вариационный ряд.

1,90; 1,65; 1,64; 1,55; 1,85; 1,66; 1,84; 2,01; 2,02; 1,72; 1,78; 2,13; 1,75; 1,85; 1,58; 1,77; 1,37; 1,43; 1,25; 2,02; 1,83; 1,41; 1,79; 1,83; 2,27; 1,39; 1,58; 2,26; 1,67; 1,65; 1,75; 1,84; 1,60; 1,75; 1,38; 1,44; 1,27; 2,03; 1,84; 1,45.

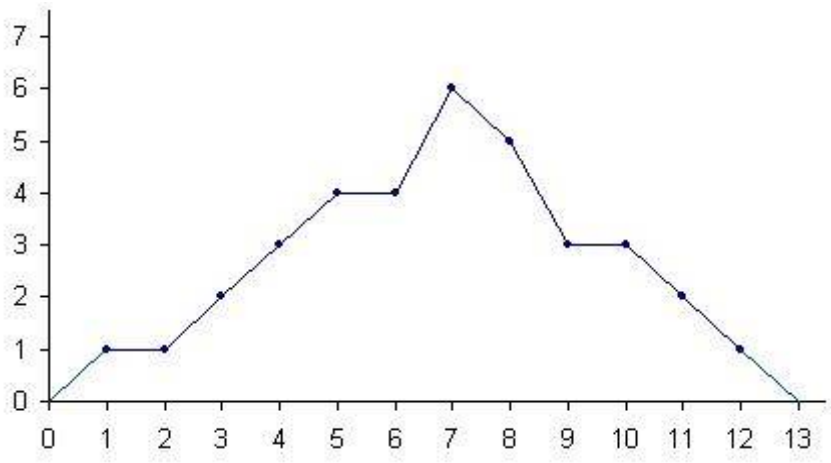
3. Графическое изображение вариационного ряда

Любой вариационный ряд можно изобразить графически в виде полигона распределения или гистограммы. Тогда он называется кривой распределения.

При построении и кривой распределения на оси абсцисс (x) наносятся естественные класса от 1 до 8, на оси ординат (y) – частоты. На месте пересечения класса и частоты ставиться точка, а затем точки соединяются кривой. Такая кривая носит название **полигона распределения**. Полигон чаще всего используют для изображения дискретных рядов.

Если полигон строят по данным интервального ряда, то в качестве абсцисс точек берут середины соответствующих интервалов. Конечно, в этом случае полигон лишь приближенно отображает зависимость частот от значений аргумента.

Пример построения полигона



Для признаков с дискретным характером изменчивости с искусственными классами на оси абсцисс откладываются классы, где конец предыдущего класса является началом следующего класса. На оси ординат откладываются частоты. Частота каждого класса фиксируется прямоугольником. Такой график называется **гистограммой распределения.**

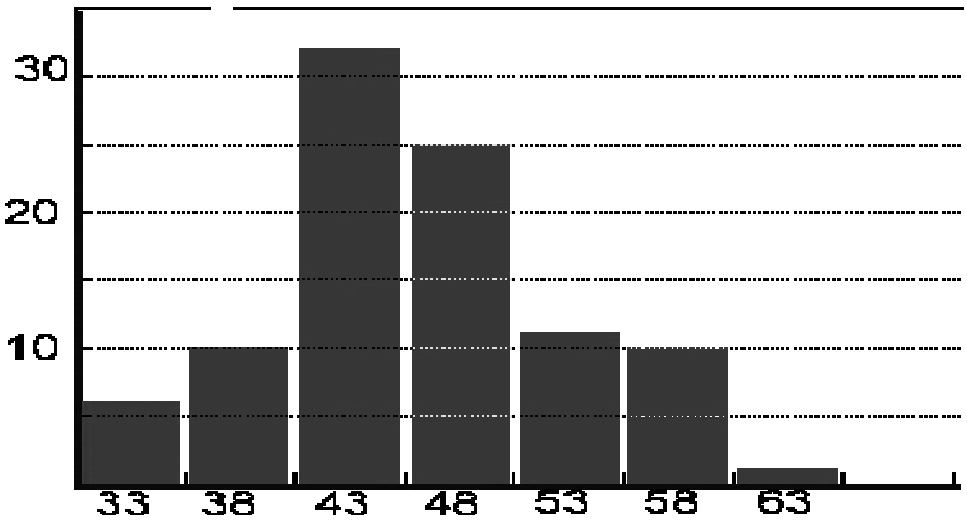


Рис. 1. Гистограмма распределения числа зерен в колосе сорта Кантегирская 89

Оцениваем соответствие полученного (эмпирического распределения теоретическому нормальному распределению.

Критерии оценки распределения как нормальное:

а) если $x = M_o = M_e$

б) min и max приблизительно равноудалены от x.

в) малые отклонения более вероятны, большие – менее.

Вычисление средней величины признака.

1. Средняя арифметическая (x)

Средняя взвешенная ($x_{взв}$) – используют в том случае, если строят вариационный ряд.

$X_i = \sum(x_i p_i) / n$

(определение % жира или белка в молоке)

2. Мода (M_o) – варианта, расположенная в середине (центре) ряда и делящая его на 2 равные части. Множество показателей (биологических и др.) следуют нормальному распределению.

Модальным классом или модой (M_o) называют класс или варианту, которым отвечает наибольшая частота. Медианой (M_e) называют варианту, расположенную по середине ранжированного ряда.

Наиболее часто используют для характеристики центральной тенденции ряда среднюю арифметическую, которую определяют по формуле:

Занятие 2

Тема: Оценка статистических показателей выборочных совокупностей

Описательная статистика занимается задачами – как наилучшим образом описать данные (описать данные сжато).

Полученные при проведении обследования данные характеризуют каждую особь совокупности в отдельности. Наиболее общие свойства этой совокупности можно установить после

статистической обработки данных. Основная задача статистической обработки наблюдений – нахождение ряда показателей, характеризующих в обобщенном виде свойства данной совокупности.

Статистические показатели выборочных совокупностей

1. Среднее значение признака (средняя арифметическая)

Одним из таких показателей является средняя арифметическая, характеризующая среднее значение признака.

Средняя арифметическая величина является обобщенной характеристикой выборочной совокупности. Средняя арифметическая показывает, какое значение признака наиболее характерно в целом для данной совокупности.

Средняя арифметическая представляет собой как бы точку равновесия вариационного ряда, отклонения от которой в сторону увеличения или уменьшения признака взаимно уравновешиваются. Она используется как для характеристики отдельных выборочных совокупностей (например, сорта) по какому-либо признаку, так и для сравнения их между собой.

Простейший метод вычисления средней арифметической величины для небольшой выборки (n < 30) – это нахождение суммы всех вариант выборки и деление ее на объем выборки. Среднюю арифметическую обозначают $\mathbf{X_{ср}}$ или **M**.

$$\overline{X} = \frac{X_1 + X_2 + X_3 + ... + X_n}{n} = \frac{\sum X_i}{n},$$

$$X_{ср} = \frac{X_1 + X_2 + X_3 + ... + X_n}{n} = \frac{\sum X}{n},$$

где Xi – варианта, величина варьирующего признака;

n – объем выборки;

Σ – знак суммирования.

$\overline{X}$ - среднее значение признака

Пример. Число зёрен в колосе у сорта Кантегирская 89 по пяти колосьям*:* 54, 45, 44, 49, 47.

$$\overline{X} = \frac{54 + 45 + 44 + 49 + 47}{5} = \frac{239}{5} = 47{,}8$$

Для больших выборок среднюю арифметическую удобнее вычислить косвенным методом по формуле:

$$X_{ср} = A + \frac{\sum pa}{n} \bullet i,$$

где A – условное среднее значение нулевого класса;
p – частоты;
a – условное отклонение;
n – объем выборки;
i – величина классового промежутка.

Основные свойства средней величины:

1. Имеется абстрактный характер, так как является обобщающей величиной, в ней стираются случайные колебания.

2. Занимает срединное положение в ряду (в строго симметричном ряду).

3. Если к каждой варианте ряда прибавить или отнять одну и ту же величину, или умножить и разделить на одну и ту же величину, средняя арифметическая увеличится или уменьшится на эту же величину.

4. Алгебраическая сумма отклонений отдельных вариант совокупности от средней арифметической этой совокупности равна нулю, то есть

$(x_1 - X) + (x_2 - X) + (x_3 - X) + \ldots + (x_i - X) = 0$

5. Сумма квадратов отклонений вариант совокупности от средней арифметической меньше суммы квадратов отклонений от любой другой величины (А), то есть $\Sigma (x_i - X)^2 < \Sigma (x_i - A)^2$.

Если значения интересующего нас признака у большинства объектов близки к их среднему и с равной вероятностью отклоняются от него в большую или меньшую сторону, лучшими характеристиками совокупности будут само *среднее* значение и *стандартное отклонение*. Напротив, когда значения признака распределены несимметрично относительно среднего, совокупность описать с помощью *медианы.*

В симметричном ряду вариант величина среднего арифметического, моды (Мо) и медианы (Ме) совпадает. Отличие Мо и Ме от $\bar{\delta}$ заключается в том, что эти параметры не зависят от крайних вариант совокупности (lim) и степени разброса отдельных вариант. Среднее арифметическое характеризует весь объем наблюдений, включая крайние варианты, имеющие нетипичный характер для той или иной совокупности.

Мо и Ме отражают основную массу наблюдений без учёта крайних вариант, величина которых иногда зависит от случайных причин.

Если выборочная совокупность имеет одну Мо, то распределение объектов (наблюдений) в такой выборке будет называться унимодальным.

В случае наличия двух и более Мо речь идет о би- или полимодальном распределении объектов (наблюдений) в совокупности. Полимодальный ряд распределения свидетельствует о неоднородности выборочной совокупности, т.е. наблюдается объединение качественно различных совокупностей.

2. Изменчивость признака, ее оценка

Средняя арифметическая даёт обобщённое представление о совокупности данных, характеризует лишь одно из свойств анализируемого явления, но она, в частности, не отражает такое важнейшее свойство, как изменчивость.

Вариационные ряды могут характеризоваться одинаковой средней арифметической, но сильно различаться по характеристике вариационного ряда, т.е. по характеру распределения вариант.

Поэтому если описывать данные совокупности только одним показателем – средней величиной, то разные выборочные совокупности, имеющие совершенно разные значения вариант будут считаться одинаковыми.

Наряду со средней арифметической важным параметром, является оценка вариации признака. Размах изменчивости или вариабельности признака можно охарактеризовать верхним и нижним лимитами.

Лимиты (lim) - это максимальное и минимальное значения признака в совокупности. Чем больше разность между максимальной (max) и минимальной (min) вариантой, тем выше изменчивость признака.

Однако при одинаковых лимитах изменчивость в сравниваемых группах может различаться, так как лимиты не учитывают распределения отдельных вариант в совокупности.

Рассмотрим два простейших вариационных ряда.

Ряд 1					**Ряд 2**				
x_1	1	2	3	4	5	x_2	1	2	3	4	5
f_x	1	2	8	2	1	f_x	1	4	4	4	1

Анализ показывает, что оба ряда полностью совпадают по средней величине, числу наблюдений и лимитам ($M = 3$; $n = 14$; $x_{min} = 1$; $x_{max} = 5$), и тем не менее первый вариационный ряд явно отличается меньшим разбросом.

Поэтому вариацию измеряют отклонением каждой варианты от средней арифметической.

Тогда $(x_1 - X) + (x_2 - X) + (x_3 - X) + \ldots + (x_i - X) = \Sigma (x_i - X)$

Величина $\Sigma (x_i - X)$ всегда равна нулю, так как сумма всех отклонений с положительным и отрицательным знаком равна нулю независимо от характера изменчивости совокупности.

$$\frac{2+4+6+7+9+2}{6} = \frac{30}{6} = 5 \qquad \overline{X} = 5$$

Сумма отклонений: $-3 - 1 + 1 + 2 + 4 - 3 = 0$, или $\sum(\overline{X} - X) = 0$

3. Дисперсия, варианса, среднее квадратическое отклонение

Чтобы отклонения от средней могли служить мерой варьирования, необходимо освободить их от знака путём возведения в квадрат:

$(x_i - X)^2$

Дисперсия (S) – сумма квадратов отклонений каждой варианты совокупности от среднего арифметического.

$\mathbf{S} = \Sigma (x_i - X)^2$

Величина дисперсии показывает рассеяние вариант вокруг средней величины.

Дисперсия является мерой изменчивости, вариации признака. В отличие от других показателей вариации дисперсия может быть разложена на составные части, что позволяет тем самым оценить влияние различных факторов на вариацию признака. Дисперсия – один из существеннейших показателей, характеризующих явление или процесс, один из основных критериев возможности создания достаточно точных моделей.

Важно отметить, что по мере увеличения численности совокупности дисперсия накапливается. В двух совокупностях разного объема при одинаковой вариации значение дисперсии будет выше в совокупности большей численности. Поэтому дисперсию необходимо усреднить. Усредненное значение дисперсии называется **вариансой.**

Вариансой, или средним квадратом, называют сумму квадратов центральных отклонений, деленную на число степеней свободы.

$$\sigma^2 = \frac{\sum (X_i - \overline{X})^2}{n}$$

Однако при делении суммы квадратов отклонений на n получаем величину, недостаточно характеризующую изменчивость, поэтому в знаменателе указанной формулы n заменяют на n-1:

$$\sigma^2 = \frac{\sum (X_i - \overline{X})^2}{n-1}; \qquad \sigma^2 = \frac{\sum (x_i - \bar{x})^2}{df}$$

где *df* (**degrees of freedom***)*- число степеней свободы, т. е. количество всех вариант совокупности, уменьшенных на единицу (*df* = n - 1). Для выборки из 100 особей (*n* = 100) число степеней свободы равно 99 **(***df* **=** *n* - 1 = 100 -1 = 99**).**

Варианса характеризует степень разнообразия величин, собранных в одну группу. Если выборка составлена из отдельных измерений признака, варианса характеризует разнообразие вариант этой группы по данному признаку.

Если группа составлена из средних величин для выборок, взятых из одной генеральной совокупности, то $\boldsymbol{\sigma}^2$ характеризует получившееся разнообразие этих выборок. В этом случае варианса средних величин $\boldsymbol{\sigma}^2_{\text{xi}}$ связана с вариансой индивидуальных значений $\boldsymbol{\sigma}^2$ равенствами $\sigma^2 = \frac{\sigma^2}{n}$, где n – одинаковая численность выборок.

Среднее квадратическое или стандартное отклонение используется как более точный показатель для характеристики изменчивости. Среднее квадратическое отклонение обозначается греческой буквой **σ** (сигма). Этот показатель указывает, насколько в среднем каждая варианта отклоняется от среднего арифметического. σ^2

Стандартное отклонение очень удобная и понятная характеристика, которая выражается в тех же единицах измерения, что и анализируемый признак (кг, см, % и т.д.). Чем больше величина **σ**, тем выше изменчивость признака.

Среднее квадратическое отклонение можно вычислить, исходя из следующей формулы:

$$\sigma = \sqrt{\sigma^2}\ .$$

$$\sigma = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n}}$$

Одно из важнейших его свойств заключается в том, что, зная среднюю величину и стандартное отклонение в отдельной выборке, можно с определённой уверенностью судить о генеральной совокупности, из которой взята эта выборка.

Из теории статистики и эмпирических исследований известно, что выборка, репрезентативно отражающая генеральную совокупность, как правило, обладает следующими свойствами:

в пределах $M \pm 1\sigma$ сконцентрировано 68, 3 % вариантов генеральной совокупности;

в пределах $M \pm 2\sigma$ сгруппировано 95, 5 % вариантов генеральной совокупности; в пределах $M \pm 3\sigma$ расположено 99, 7 % вариантов генеральной совокупности**.**

Особенно большое значение имеет при исследовании нормальных распределений. В нормальном распределении 68% всех случаев лежит в интервале + одного отклонения от среднего, 95% - + двух стандартных отклонений от среднего и 99,7% всех случаев - в интервале + трех стандартных отклонений от среднего.

Вся изменчивость признака лежит от среднего арифметического в пределах $\pm 3{,}3$ **σ**. Это называется *правилом «плюс-минус трех сигм».* Поэтому средняя арифметическая, увеличенная и уменьшенная на три сигмы, дает практически крайние значения признака при нормальном распределении объектов в совокупности.

4. Коэффициент вариации (CV)

Таким образом, стандартное отклонение представляет собой одну из наиболее обоснованных и эффективных описательных статистик.

Однако, если необходимо сопоставить изменчивость признаков, представленных в разных единицах измерения (например, число колосков в колосе, масса зерна, длина стебля и т. д.) этот показатель использовать нельзя, так как он измеряется в тех же величинах, что и средняя величина. Кроме того, одно и то же значение стандартного отклонения (например, $\sigma = 2$) может указывать как на очень малую ($M = 100$), так и на очень большую изменчивость ($M = 5$).

Для сравнения изменчивости признаков, выраженных в разных единицах измерения, рассчитывается коэффициент вариации (CV), равный процентному отношению стандартного отклонения к средней арифметической величине, то есть:

$$CV = \frac{\sigma}{X} \cdot 100\%$$

Коэффициент вариации находит применение и в селекционной работе [Федоров, 1957; Снедекор, 1961; Мацеевский, Земба, 1988 и др.]. Например, при сравнении двух сходных по продуктивности и качественным показателям сортов, предпочтение должно быть отдано тому из них, который при равных условиях обладает меньшей изменчивостью.

При характеристике совокупности коэффициент вариации является дополнительным показателем и должен применяться с основными параметрами σ и $\bar{x}$.

Коэффициент вариации, как дисперсия и стандартное отклонение, является показателем изменчивости признака. При величине коэффициента вариации до 10% изменчивость оценивается как слабая, 11-25% - средняя, более 25% - сильная (Лакин, 1990).

5. Ошибка средней арифметической

Средняя арифметическая любой выборочной совокупности характеризует среднюю генеральной совокупности не точно, а приближенно, отличаясь от неё на некоторую величину.

Средняя ошибка – это статистическая ошибка и не имеет ничего общего с ошибкой точности. Статистические показатели для выборочной совокупности всегда имеют так называемые ошибки выборочности, или ошибки репрезентативности, которые представляют собой среднюю величину расхождения между средними значениями признака в выборке и генеральной совокупности. Наиболее общую совокупность называют **генеральной**. Эта теоретически бесконечно большая или, во всяком случае, приближающаяся к бесконечности совокупность, как правило, не поддаётся исследованию, поэтому практически изучают ограниченную часть совокупности – **выборочную совокупность**.

Стандартная ошибка среднего это величина, на которую отличается среднее значение выборки от среднего значения генеральной совокупности при условии, что распределение близко к нормальному. С вероятностью 0,68 можно утверждать, что среднее значение генеральной совокупности лежит в интервале + одной стандартной ошибки от среднего, с вероятностью 0,95 - в интервале + двух стандартных ошибок от среднего и с вероятностью 0,99 - среднее значение генеральной совокупности лежит в интервале + трех стандартных ошибок от среднего.

Формула для определения средней ошибки:

$$S_{\overline{X}} = \frac{\sigma}{\sqrt{n}},$$

где S_x - ошибка средней арифметической, или средняя ошибка

σ- среднее квадратическое отклонение;

n – число наблюдений, вариант (xi).

2. Способы вычисления статистических параметров

2.1. Вычисление статистических параметров для данных, не сгруппированных в вариационный ряд

Таблица 1

Число колосков в колосе у сорта Кантегирская 89, 1996 г.

Номер растения	Число колосков в колосе	Отклонения от средней	Квадрат отклонений
1	14	-1	1
2	14	-1	1
3	14	-1	1
4	16	+1	1
5	14	-1	1
6	14	-1	1
7	16	+1	1
8	16	+1	1
9	16	+1	1
10	16	+1	1
N=10	=15	(-5)+(+5)=0	$\sum (X_i - \overline{X})^2 = 10$

$$\sigma^2 = \frac{\sum (X_i - \overline{X})^2}{n-1} = \frac{10}{10-1} = \frac{10}{9} = 1{,}11$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{1{,}11} = 1{,}05$$

$$CV = \frac{\sigma}{\overline{X}} \cdot 100\% = \frac{1{,}05 \cdot 100}{15} = 7\%$$

$$S_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1{,}05}{\sqrt{10}} = \frac{1{,}05}{3{,}16} = 0{,}33$$

2.2. Прямой способ вычисления статистических параметров для данных, сгруппированных в вариационный ряд

Если все варианты разнесены по классам, каждый из которых характеризуется определённым значением и частотой, то средняя арифметическая вычисляется по формуле:

$$\overline{X} = \frac{\sum fx}{n},$$

где f – частота класса;

x- значение класса;

n – число вариант.

Таблица 2

Число продуктивных стеблей на растение у сорта Кантегирская 89

Значение класса, x	Частота, f	fx	fx^2 (x fx)
4	1	4	16
5	4	20	100
6	6	36	216
7	11	77	539
8	8	64	512
9	3	27	243
10	2	20	200
11	1	11	121
12	1	12	144
13	1	13	169
14	1	14	196
15	1	15	225
	n = 40	$\sum$ fx = 313	$\sum fx^2$= **2681**

Варианса:

$$\sigma^2 = \frac{\Sigma fx^2 - (\Sigma fx)^2/n}{n-1}$$

2.3. Непрямой способ вычисления статистических параметров для группированных данных

Оценка параметров выборочной совокупности в вариационном ряду способом условных отклонений используется, когда варианты вариационный ряда имеют большие величины (трехзначные, четырехзначные значения). Метод условных отклонений позволяет упростить расчеты.

Таблица 3

Длина стебля у сорта Кантегирская 89, см

Границы классов	Среднее значение класса	Частота (f)	Условные отклонения			
			a	a^2	f a	$f a^2$
91 – 95	93	3	–3	9	–3	27
96 – 100	98	5	–2	4	–10	20
101 – 105	103	4	–1	1	– 4	4
106 – 110	108	10	0	0	0	0
111 – 115	113	8	1	1	8	8
116 – 120	118	8	2	4	16	32
121 – 125	123	2	3	9	6	18
i = 5		$\Sigma f = n = 40$			$\sum fa = (-23) + (+30) = + 7$	$\Sigma f^2 = 109$

Для вычисления средней арифметической необходимо:

1. Найти в построенном вариационном ряду условный средний класс. В качестве условного среднего класса рекомендуется брать класс, который занимает центральное место в данном вариационном ряду или имеет наибольшее значение частот (f). В нашем примере условным средним классом будет четвертый класс с наибольшей встречаемостью вариант (f = 10) и варьированием в пределах 106 –110 см.
2. Выбранный условный средний класс принимаем за нулевой.
3. Вычислить условное среднее значение нулевого класса. Его обозначают буквой **А**.

$$A = \frac{106+110}{2} = \frac{226}{2} = 108$$

Условное отклонение (**а**) каждого класса от нулевого определяем путем вычитания из значения конк класса значение нулевого класса и делим на классовый промежуток (i = 5), $a1 = \frac{52-55}{3} = -1$, следующи

$a2 = \frac{49-55}{3} = -2.$

Вверх от класса, принятого за условный нулевой, получим натуральный ряд отрицательных чисел (–1, –2, –3 и т.д.), вниз – натуральный ряд положительных чисел (+1, +2, +3 и т.д. в зависимости от класса). Дальнейшие расчеты ведутся с полученными условными значениями, как со значениями классов.

Среднее арифметическое при способе условных отклонений рассчитывается по формуле:

$$X_{ср} = A + \frac{\sum fa}{n} \bullet i,$$

где А – условное среднее значение нулевого класса;

Σ f a –сумма произведений положительных и отрицательных значений f a

i – величина классового промежутка. Для рассматриваемого примера.

$$X_{ср} = 108 + \frac{7}{40} \bullet 5 = 108{,}95$$

Формула для оценки дисперсии:

$$S = [\sum fa^2 - \frac{(\sum fa)^2}{n}] \cdot k^2$$

Занятие 3

Тема: Оценка достоверности различий между средними значениями двух выборочных совокупностей

Следующей задачей статистического анализа, решаемой после определения основных (выборочных) характеристик и анализа одной выборки, является совместный анализ нескольких выборок. Важнейшим вопросом, возникающем при анализе двух выборок, является вопрос о наличии различий между выборками. Обычно для этого проводят проверку статистических гипотез о принадлежности обеих выборок одной генеральной совокупности или о равенстве средних.

Если вид распределения или функция распределения выборки нам заданы, то в этом случае задача оценки различий двух групп независимых наблюдений может решаться с использованием **параметрических критериев** статистики: либо критерия Стьюдента (t), если сравнение выборок ведется по средним значениям (Х и У), либо с использованием критерия Фишера (F), если сравнение выборок ведется по их дисперсиям.

Использование параметрических критериев статистики без предварительной проверки вида распределения может привести к определенным ошибкам в ходе проверки рабочей гипотезы.

Непараметрические критерии статистики - свободны от допущения о законе распределения выборок и базируются на предположении о независимости наблюдений.

В группу **параметрических критериев** методов математической статистики входят методы для вычисления описательных статистик, построения графиков на нормальность распределения, проверка гипотез о принадлежности двух выборок одной совокупности. Эти методы основываются на предположении о том, что распределение выборок подчиняется нормальному (гауссовому) закону распределения. Среди параметрических критериев статистики будут рассмотрены критерий Стьюдента и Фишера.

1. Методы проверки выборки на нормальность

Чтобы определить, имеем ли мы дело с нормальным распределением, можно применять следующие методы:

1) в пределах осей можно нарисовать полигон частоты (эмпирическую функцию распределения) и кривую нормального распределения на основе данных исследования. Исследуя формы кривой нормального распределения и графика эмпирической функции распределения, можно выяснить те параметры, которыми последняя кривая отличается от первой;

2) вычисляется среднее, медиана и мода и на основе этого определяется отклонение от нормального распределения. Если мода, медиана и среднее арифметическое друг от друга значительно не отличаются, мы имеем дело с нормальным распределением. Если медиана значительно отличается от среднего, то мы имеем дело с асимметричной выборкой.

3) эксцесс кривой распределения должен быть равен 0. Кривые с положительным эксцессом значительно вертикальнее кривой нормального распределения. Кривые с отрицательным эксцессом являются более покатистыми по сравнению с кривой нормального распределения;

2. Критерий Стьюдента (t-критерий)

Критерий позволяет найти вероятность того, что оба средних значения в выборке относятся к одной и той же совокупности. Данный критерий наиболее часто используется для проверки гипотезы: «Средние двух выборок относятся к одной и той же совокупности».

При использовании критерия можно выделить два случая. В первом случае его применяют для проверки гипотезы о равенстве генеральных средних двух **независимых**, **несвязанных** выборок (так называемый **двухвыборочный t-критерий**). В этом случае есть контрольная группа и экспериментальная (опытная) группа, количество испытуемых в группах может быть различно.

Во втором случае, когда одна и та же группа объектов порождает числовой материал для проверки гипотез о средних, используется так называемый **парный t-критерий**. Выборки при этом называют **зависимыми**, **связанными**.

а) случай независимых выборок

Статистика критерия для случая несвязанных, независимых выборок равна:

$$t_{эмп} = \frac{\bar{x} - \bar{y}}{\sigma_{x-y}} \qquad t_{эмп} = \left|\frac{\bar{X} - \bar{Y}}{Sd}\right| \qquad Sd = \sqrt{S_x^2 + S_y^2} \qquad (1)$$

где $\bar{x}$, $\bar{y}$ — средние арифметические в экспериментальной и контрольной группах,

σ_{x-y} - стандартная ошибка разности средних арифметических. Находится из формулы:

$$\sigma_{x-y} = \sqrt{\frac{\sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2}{n_1 + n_2 - 2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \qquad (2)$$

где n_1 и n_2 соответственно величины первой и второй выборки.

Если $n_1 = n_2$, то стандартная ошибка разности средних арифметических будет считаться по формуле:

$$\sigma_{x-y} = \sqrt{\frac{\sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2}{(n-1) \cdot n}} \qquad (3)$$

где n - величина выборки.

$$Sd = \sqrt{S_x^2 + S_y^2} = \sqrt{\frac{\sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2}{(n-1) \times n}}$$

Подсчет **числа степеней свободы** осуществляется по формуле:

$k = n_1 + n_2 - 2.$ (4)

При численном равенстве выборок $k = 2n - 2$.

Далее необходимо сравнить полученное значение $t_{эмп}$ с теоретическим значением t—распределения Стьюдента (см. приложение к учебникам статистики). Если $t_{эмп} < t_{крит}$, то гипотеза H_0

принимается, в противном случае нулевая гипотеза отвергается и принимается альтернативная гипотеза.

Рассмотрим пример использования t-критерия Стьюдента для несвязных и неравных по численности выборок.

Пример 1. В двух группах учащихся — экспериментальной и контрольной — получены следующие результаты по учебному предмету (тестовые баллы; см. табл.).

Таблица Результаты эксперимента

Первая группа (экспериментальная) N_1=11 человек	Вторая группа (контрольная) N_2=9 человек
12 14 13 16 11 9 13 15 15 18 14	13 9 11 10 7 6 8 10 11

Общее количество членов выборки: n_1=11, n_2=9.

Расчет средних арифметических: $X_{ср}$=13,636; $Y_{ср}$=9,444

Стандартное отклонение: σ_x=2,460; σ_y=2,186

По формуле (2) рассчитываем стандартную ошибку разности арифметических средних:

$$\sigma_{x-y} = \sqrt{\frac{60{,}545 + 38{,}222}{11 + 9 - 2} \cdot \left(\frac{1}{11} + \frac{1}{9}\right)} = 1{,}053$$

Считаем статистику критерия:

$$t = \frac{13.636 - 9.444}{1.053} = 3.981$$

Сравниваем полученное в эксперименте значение t с табличным значением с учетом степеней свободы, равных по формуле (4) числу испытуемых минус два (18).

Табличное значение $t_{крит}$ равняется 2,1 при допущении возможности риска сделать ошибочное суждение в пяти случаях из ста (уровень значимости=5 % или 0,05).

Если полученное в эксперименте эмпирическое значение t превышает табличное, то есть основания принять альтернативную гипотезу (H_1) о том, что учащиеся экспериментальной группы показывают в среднем более высокий уровень знаний. В эксперименте t = 3,981, табличное t = 2,10, 3,981>2,10, откуда следует вывод о преимуществе экспериментального обучения.

Здесь могут возникнуть такие ***вопросы***:

1. Что если полученное в опыте значение t окажется меньше табличного? Тогда надо принять нулевую гипотезу.

2. Доказано ли преимущество экспериментального метода? Не столько доказано, сколько показано, потому что с самого начала допускается риск ошибиться в пяти случаях из ста (p=0,05). Наш эксперимент мог быть одним из этих пяти случаев. Но 95% возможных случаев говорит в пользу альтернативной гипотезы, а это достаточно убедительный аргумент в статистическом доказательстве.

3. Что если в контрольной группе результаты окажутся выше, чем в кспериментальной? Поменяем, например, местами, сделав $\bar{y}$ средней арифметической экспериментальной группы, а $\bar{x}$ — контрольной:

$$t = \frac{9.444 - 13.636}{1.053} = -3.981$$

Отсюда следует вывод, что новый метод пока не проявил себя с хорошей стороны по разным, возможно, причинам. Поскольку абсолютное значение 3,9811>2,1, принимается вторая альтернативная гипотеза (H_2) о преимуществе традиционного метода.

б) случай связанных (парных) выборок

В случае связанных выборок с равным числом измерений в каждой можно использовать более простую формулу t-критерия Стьюдента.

Вычисление значения t осуществляется по формуле:

$$t_{эмп} = \frac{\bar{d}}{Sd} \qquad (5)$$

где $d_i = x_i - y_i$ — разности между соответствующими значениями переменной X и переменной У, а d - среднее этих разностей;

Sd вычисляется по следующей формуле:

$$Sd = \sqrt{\frac{\sum d_i^2 - \frac{(\sum d_i)^2}{n}}{n \cdot (n-1)}} \qquad (6)$$

Число степеней свободы *k* определяется по формуле k=n-1. Рассмотрим пример использования t-критерия Стьюдента для связных и, очевидно, равных по численности выборок.

Если $t_{эмп} < t_{крит}$, то нулевая гипотеза принимается, в противном случае принимается альтернативная.

3. Критерий Фишера позволяет сравнивать величины выборочных дисперсий двух независимых выборок. Для вычисления $F_{эмп}$ нужно найти отношение дисперсий двух выборок, причем так, чтобы большая по величине дисперсия находилась бы в числителе, а меньшая – в знаменателе. Формула вычисления критерия Фишера такова:

$$F_{эмп} = \frac{\sigma_x^2}{\sigma_y^2}, \qquad (8)$$

где σ_x^2, σ_y^2 - дисперсии первой и второй выборки соответственно.

Так как, согласно условию критерия, величина числителя должна быть больше или равна величине знаменателя, то значение $F_{эмп}$ всегда будет больше или равно единице.

Число степеней свободы определяется также просто:

$k_1 = n_1 - 1$ для первой выборки (т.е. для той выборки, величина дисперсии которой больше) и $k_2 = n_2 - 1$ для второй выборки.

В Приложении 1 критические значения критерия Фишера находятся по величинам k_1 (верхняя строчка таблицы) и k_2 (левый столбец таблицы).

Если $t_{эмп} > t_{крит}$, то нулевая гипотеза принимается, в противном случае принимается альтернативная.

Задание.

В двух опытах изучено число зёрен в колосе пшеницы. Получены следующие данные. Оцените достоверность разности между средними значениями анализируемого признака двух опытов и напишите вывод.

Опыт	n	$\bar{x}$	S
1	61	33,9	22,8
2	75	30,5	18,4

2. Наименьшая существенная разность (НСР) — величина, указывающая границу возможных случайных отклонений в эксперименте; это та минимальная разность в урожаях между средними, которая в данном опыте признается существенной при 5%-ном ($НСР_{05}$) или 1%-ном ($НСР_{01}$) уровне значимости.

НСР=t*Sd.

$$S_d = \sqrt{S^2x + S^2x}$$

df=n_1+n_2-2

t – значение критерия Стьюдента, соответствующее числу степеней свободы.

Уровень значимости — риск сделать ошибочное заключение. В агрономических исследованиях допускается 5 и 1 %.

Занятие 4

Тема: Исследование зависимостей

Описательная статистика и статистические критерии позволяют, соответственно, компактно представлять полученные результаты и определять сходства и различия.

Следующим этапом анализа данных обычно является исследование зависимостей. Для этих целей применяются корреляционный анализ и дисперсионный анализ (для установления факта наличия/отсутствия зависимости между переменными), а также регрессионный анализ (для нахождения количественной зависимости между переменными).

1. Корреляционный анализ

Вообще, в природе, и в биологии в частности, существуют вполне определённые связи признаков. Например, между телосложением и темпераментом людей, между строением их тела и предрасположенностью к заболеваниям существует определенная связь (еще Гиппократ обратил внимание на существование данных связей).

У исследователя часто возникает вопрос о взаимосвязи отдельных признаков. Например, как связаны хозяйственно-ценные признаки полевых культур, определяющие продуктивность растения?

Для описания зависимости между различными величинами, описывающими интересующие его признаки,

служит математическое понятие функции, имеющее в виду случаи, когда определенному значению одной (независимой) переменной X, называемой ***аргументом***, соответствует определенное значение другой (зависимой) переменной Y, называемой ***функцией***. Однозначная зависимость между переменными величинами *Y* и *X* называется ***функциональной***, т.е. $Y = f(X)$ (“игрек есть функция от икс”).

Например, в функции $\boldsymbol{Y = 2X}$ каждому значению *X* соответствует в два раза большее значение *Y*. В функции $\boldsymbol{Y = 2X^2}$ каждому значению *Y* соответствует 2 определенных значения *X*.

Функциональная связь является наиболее простым видом связи между величинами, при которой каждому значению одной величины соответствуют строго определённые значения другой.

Например, к функциональной относится зависимость между высотой местности и насыщением гемоглобина кислородом.

Но такого рода однозначные или функциональные связи между переменными величинами встречаются не всегда. Нередко встречаются связи между величинами, которые нельзя отнести к функциональным зависимостям. К ним, например, относятся связи между урожаем и количеством осадков или между ростом отцов и сыновей.

Причина таких "исключений" в том, что каждый биологический признак, выражаясь математическим языком, является функцией многих переменных; на его величине сказывается влияние и генетических и средовых факторов, в том числе и случайных, что вызывает варьирование признаков. Отсюда зависимость между ними приобретает не функциональный, а ***статистический характер***, когда определенному значению одного признака, рассматриваемого в качестве независимой переменной, соответствует не одно и то же числовое значение, а целая гамма распределяемых в вариационный ряд числовых значений другого признака, рассматриваемого в качестве независимой переменной. Например, при росте человека 170 см масса тела может быть 70 кг, 65 кг, 72 кг и т.д. Случайный разброс этих возможных значений объясняется влияниембольшого числа дополнительных факторов, от которых отвлекаются, изучая связь между данными величинами.

Такого рода зависимость между переменными величинами называется ***корреляционной*** или ***корреляцией*** (термин "корреляция" происходит от лат. correlatio — соотношение, связь).

Задача корреляционного анализа сводится к установлению направления и формы связи между признаками, измерению ее тесноты и к оценке достоверности выборочных показателей корреляции.

Пусть сделаны измерения двух признаков X и У: X1, X2,...,Xn и Y1, Y2,...,Yn.

Необходимо установить, существует ли связь между изменениями признаков X и Y и, если эта связь существует, то определить её тип, глубину и достоверность.

Для качественной оценки связи между признаками строят график.

Экспериментальные графики для величин X и Y, находящихся в корреляционной зависимости, состоят из ряда точек, не укладывающихся на какую-либо определённую кривую. Каждая точка (x,y) на плоскости отображает результат одного измерения. Такой точечный график называют **корреляционным полем**. По корреляционному полю можно качественно оценить наличие или отсутствие зависимости и указать положительна она или отрицательна.

Коэффициент корреляции p для генеральной совокупности, как правило, неизвестен, поэтому он оценивается по экспериментальным данным, представляющим собой выборку объема n пар значений (x_i, y_i), полученную при совместном измерении двух признаков X и Y. Коэффициент корреляции, определяемый по выборочным данным, называется ***выборочным коэффициентом корреляции*** (или просто ***коэффициентом корреляции***). Его принято обозначать символом ***r***.

В случае, когда имеются две переменных, значения которых измерены в цифровой шкале отношений (единицы измерений при этом не важны – например, масса зерна может быть измерена в граммах, килограммах, тоннах – они не влияют на значение коэффициента корреляции), используется коэффициент линейной корреляции Пирсона r, который принимает значения от -1 до +1 (нулевое его значение свидетельствует об отсутствии корреляции.

Проанализировав знак коэффициента корреляции, определяют тип корреляционной связи:

если $r > 0$, то связь прямая (положительная), т.е. при возрастании одной величины другая в среднем тоже возрастает;

если $r < 0$, то связь обратная (отрицательная), т.е. при возрастании одной величины другая имеет тенденцию в среднем убывать.

Если статистическая связь между признаками отсутствует, то $r = 0$.

Величина коэффициента корреляции показывает глубину линейной связи между двумя выборками, т.е. характеризует степень близости зависимости величин X и Y к линейной функциональной зависимости. Графически это выражается теснотой или разбросанностью точек корреляционного поля.

В практической деятельности, когда число коррелируемых пар признаков *X* и *Y* не велико ($\mathbf{n \leq 30}$), то при оценке зависимости между показателями используется следующую градацию.

Глубина корреляционной связи определяется, исходя из следующих критериев:

если 0< |r| ≤0,3, то связь слабая;

если 0,3< |r| ≤0,5, то связь умеренная;

если 0,5< |r| ≤0,7, то связь значительная;

если 0,7< |r| ≤0,9, то связь сильная;

если 0,9< |r| <1, то связь очень сильная.

При |r| =1 связь между величинами функциональная.

Таким образом, чем ближе абсолютная величина r к единице, тем сильнее связь между признаками и теснее расположены точки на графике.

Однако, для обоснованного вывода о наличии связи не достаточно анализа величины коэффициента корреляции; необходимо проверить его достоверность.

Иными словами, требуется ответить на вопрос: является ли вычисленный поданным наблюдений коэффициент корреляции значимым, т.е. можно ли верить полученному значению коэффициента, учитывая случайный характер выборок значений исследуемых величин.

Значимость корреляционной связи при определённом уровне доверительной вероятности можно проверить с помощью критерия Стьюдента.

Из таблицы 1 для числа степеней свободы ν = n - 2 определяют стандартные значения критериев Стьюдента, соответствующие трем порогам достоверности: 0,95; 0,99; 0,999.

Сравнивают критерий достоверности tr со стандартными значениями критериев Стьюдента и делают вывод о достоверности коэффициента корреляции:

• если tr ≥ tst0,999 , то достоверность коэффициента корреляции 99,9%;

• если tr ≥ tst0,99 , то достоверность коэффициента корреляции 99%;

• если tr ≥ tst0,95 , то достоверность коэффициента корреляции 95%;

• если tr < tst0,95 , то коэффициент корреляции недостоверен, доверять ему нельзя.

1.1. Коэффициент линейной корреляции Пирсона

Наиболее распространенный коэффициент корреляции. Предназначен для расчета силы и направления линейной зависимости между переменными исследования.

Смысл коэффициента линейной корреляции.

Коэффициент линейной корреляции отражает меру линейной зависимости между двумя переменными. Предполагается, что переменные измерены в интервальной шкале либо в шкале отношений.

Общая формула:

$$r_{xy} = \frac{\sum_{i=1}^{n}\left[\left(x_i - \bar{x}\right)\cdot\left(y_i - \bar{y}\right)\right]}{\left(n-1\right)\cdot\sigma_x\cdot\sigma_y} \quad (1)$$

где x_i и y_i - сравниваемые количественные признаки, n – число сравниваемых наблюдений, σ_x и σ_y – стандартные отклонения в сопоставляемых рядах.

В формуле корреляции Пирсона используется среднее арифметическое и стандартное отклонение коррелируемых рядов, а в формуле Спирмена не используется. Таким образом, для получения адекватного результата по формуле Пирсона, необходимо, чтобы коррелируемые ряды были приближены к **нормальному распределению** (среднее и стандартное отклонение являются параметрами нормального распределения).

$$r_{xy} = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}} \quad (2)$$

где x_i — значения, принимаемые в выборке X,
y_i — значения, принимаемые в выборке Y;
$\bar{x}$ — средняя по X, $\bar{y}$ — средняя по Y.

В формуле (2) встречается величина $\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})$ при делении на n (число значений переменной X или Y) она называется **ковариацией**. Формула (2) предполагает также, что при расчете коэффициентов корреляции число значений переменной X равно числу значений переменной Y.

Для расчетов вручную используется преобразованная формула:

$$r_{xy} = \frac{n\sum (x_i \cdot y_i) - \sum x_i \cdot \sum y_i}{\sqrt{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right) \cdot \left(n\sum y_i^2 - \left(\sum y_i\right)^2\right)}}$$

Полученный коэффициент корреляции проверяется на значимость с помощью таблицы критических значений. Уровень значимости, вычисленный для каждой корреляции, представляет собой главный источник информации о надежности корреляции.

Коэффициент корреляции Пирсона (*r*) представляет собой меру линейной зависимости двух переменных. Если возвести его в квадрат, то полученное значение **коэффициента детерминации** (r^2) представляет долю вариации, общую для двух переменных (иными словами, "степень" зависимости или связанности двух переменных).

Задание. Провести корреляционный и регрессионный анализы линейной зависимости

Таблица 1

Корреляционная зависимость между произведением длинны листьев яблони на их ширину (X) и площадью листьев (У)

Номер листа	X, см	У,см	X - x	У- y	$(X - x)^2$	$(У - y)^2$	$\sum(X - x) * (У - y)$
1	15,8	7,2	-21,6	-16,8	466,56	282,24	362,88
2	18,8	11,8	-18,6	-12,2	345,96	148,84	226,92
3	27,0	18,6	-10,4	-5,4	108,16	29,16	56,16

Продожение

4	28,8	19,1	-8,6	-4,9	73,96	24,01	42,14
5	28,8	19,4	-8,6	-4,6	73,96	21,16	39,56
6	29,6	19,5	-7,8	-4,5	60,84	20,25	35,10
7	32,5	21,6	-4,9	-2,4	24,01	5,76	11,16
8	32,8	22,1	-4,6	-1,9	21,16	3,61	8,74
9	36,5	23,1	-0,9	-0,9	0,81	0,81	0,81
10	38,5	23,2	1,1	-0,8	1,21	0,64	0,88
11	39,6	23,6	2,2	-0,4	4,84	0,16	0,88
12	39,7	26,5	2,3	2,5	5,29	6,25	5,75
13	39,7	27,3	2,3	3,3	5,29	10,89	7,59
14	44,5	28,6	7,1	4,6	50,41	21,16	32,66
15	46,2	29,3	8,8	5,3	77,44	28,09	46,64
16	46,4	29,7	9,0	5,7	81,00	32,49	51,30
17	48,0	30,4	10,6	6,4	112,36	40,96	67,84
18	49,8	30,8	12,4	6,8	153,76	46,24	84,32
19	51,0	34,4	13,6	10,4	184,96	108,16	141,44
20	53,9	34,6	16,5	10,6	272,25	112,36	174,90
	x=37,4	y=24,0	∑(X – x) =-0,1	∑ (У – y) = 0,8	∑(X – x)^2 =2124,23	∑(У – y)^2 =943,24	∑(X – x) * (У – y) =1397,67

Число пар n = 20

1. Вычисление коэффициента корреляции (r).

$$r = \frac{\sum (X - \overline{x}) \cdot (Y - \overline{y})}{\sqrt{\sum (X - \overline{x})^2 \cdot \sum (Y - \overline{y})}} = \frac{1397,67}{\sqrt{2124,23 \cdot 943,24}} = +0,987$$

2 .Ошибка коэффициента корреляции (Sr).

$$Sr = \sqrt{\frac{1 - r^2}{n - 2}} = \sqrt{\frac{1 - 0,974}{20 - 2}} = 0,00144 .$$

3. Критерий достоверности коэффициента корреляции (tr).

$$t_r = \frac{r}{S_r} = \frac{0,987}{0,00144} = 685$$

Теоретическое значение критерия t находят по таблице Стьюдента (приложение1) при числе степеней свободы υr = n – 2 = 20-2 = 18; $t_{0,95} = 2,1$; $t_{0,99} = 2,88$.

Выводы

1. Так как коэффициент корреляции r = +0,987, то связь между изучаемыми показателями прямая и сильная, приближающаяся к полной.

2. Критерий достоверности tr (685) больше t0,95 и t0,99 следовательно, связь достоверна на самых высоких уровнях доверительной вероятности.

1.2. Регрессионный анализ

Регрессионный анализ проводится при сильной и достоверной связи и

любом направлении (прямом или обратном). В рассмотренном примере с это целесообразно сделать по произведению длины листа на его ширину, т.е. по значению X определить площадь листьев Y.

Уравнение линейной регрессии имеет вид Y = $\bar{y}$ Ryx (X - $\bar{x}$) , где y и x – средние арифметические анализируемых вариационных рядов; X – произведение длины на ширину для листьев, площадь которых надо определить.

$$Ryx = \frac{\sum(X-\bar{x})(Y-\bar{y})}{\sum(X-\bar{x})^2} = \frac{1397,67}{2124,23} = +0,658$$ см^2 на 1 см произведения длины на ширину. Тогда Y = 24,0 + 0,658 (X- 37,4).

Пусть средняя произведения длины на ширину для 30 листьев яблони равна 44,5 см. Подставив это значение в предыдущее уравнение, получим Y = 24,0 + 0,658 (44,5 – 37,4) = 24,0 + 4,64 = 28,67 см^2. Фактическая площадь 14-го листа равна 28,6 см^2. Разница между расчетным и фактическим значениями составляет 28,67 – 28,6 = 0,07 см^2, или (0,07 ⋅ 100) : 28,6 = 0,25 %. Ошибка 0,25 % свидетельствует о достаточно высокой точности определения площади листьев яблони по произведению длины на ширину.

1.3. Нелинейные зависимости между переменными

Другим возможным источником трудностей, связанным с линейной корреляцией Пирсона (r), является форма зависимости. Корреляция Пирсона r хорошо подходит для описания линейной зависимости. Отклонения от линейности увеличивают общую сумму квадратов расстояний от регрессионной прямой, даже если она представляет "истинные" и очень тесные связи между переменными.

Если корреляция сильная, однако зависимость явно нелинейная? К сожалению, не существует простого ответа на данный вопрос, так как не имеется естественного обобщения коэффициента корреляции Пирсона r на случай нелинейных зависимостей. Однако, если кривая монотонна (монотонно возрастает или, напротив, монотонно убывает), то можно преобразовать одну или обе переменные, чтобы сделать зависимость линейной, а затем уже вычислить корреляцию между преобразованными величинами. Для этого часто используется логарифмическое преобразование. Другой подход состоит в использовании непараметрической корреляции (например, корреляции Спирмена).

Ранговый коэффициент корреляции Спирмена

Коэффициентом ранговой корреляции Спирмена называют непараметрический метод, используемый при статистическом исследовании связи между различными явлениями.

Метод ранговой корреляции Спирмена позволяет определять тесноту (или силу) и направление корреляционной связи между двумя профилями признаков или признаками. Мощность параметрического коэффициента корреляции превосходит мощность коэффициента ранговой корреляции Спирмена.

Коэффициент ранговой корреляции Спирмена используется в случаях, когда:

- переменные имеют ранговую шкалу измерения (но могут быть измерены также в шкале

интервалов и отношений.);
- распределение данных слишком отличается от нормального или вообще неизвестно;
- выборки имеют небольшой объём (N < 30).

Число варьирующих признаков в сравниваемых переменных X и Y должно быть одинаковым.

Перед использованием коэффициента Спирмена для рядов данных с различным размахом, необходимо обязательно их ранжировать . Ранжирование приводит к тому, что значения этих рядов приобретают одинаковый минимум = 1 (минимальный ранг) и максимум, равный количеству значений (максимальный, последний ранг = N, т.е. максимальному количеству случаев в выборке).

Без ранжирования можно обойтись, когда данные имеют исходно ранговую шкалу.

Для расчета коэффициента ранговой корреляции Спирмена выделяют следующие действия:

1.Каждому из признаков присваивается порядковый номер (ранг). Ранг может присваиваться как по возрастанию, так и по убыванию.

2. Определяется разность рангов каждой пары сопоставляемых значений.

3. Каждая разность возводится в квадрат, а полученные результаты затем суммируются.

4. Коэффициент корреляции рангов высчитывается по формуле:

$$r = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

$\sum d^2$ - сумма квадратов разностей рангов (x – y)2

n – число парных наблюдений.

Слабой теснотой связи называют связь с коэффициентом равным или меньшим 0,3. Значения коэффициента от 0,4 до 0,7 считают показателями умеренной тесноты, а если полученное значение превышает 0,7, то говорят о высокой тесноте связи.

Таблицы для определения критических значений коэффициента корреляции Спирмена рассчитаны от числа признаков равных n = 5 до n = 40 и при большем числе сравниваемых переменных следует использовать таблицу для пирсоновского коэффициента корреляции. Нахождение критических значений осуществляется при k = n.

Задание. Проанализировать связь между баллом поражения и урожаем у гибридов яблони.

Таблица состоит из двух переменных:

Балл (x)	Урожай (y)
4	10
1	15
2	20
3	10
5	5
5	15
2	25
1	20
3	15
4	15

Выполнение задания

1. Вычислите среднее арифметическое по каждому признаку.
2. Проставьте ранги.
3. Проверьте H_0. Сравните результат с табличным значением t-критерия.
4. Охарактеризуйте связь между баллом поражения и урожаем и напишите вывод.

Рекомендуемая литература

1. *Лакин Г.Ф*. Биометрия / Г.Ф. Лакин - М.: Высш. шк., 2013.- 300 с.
2. *Рокицкий П.Ф*. Биологическая статистика / П.Ф. Рокицкий - Минск: Высшая школа, 1973.- 319 с.
3. *Мазер К., Джинкс Дж*. Биометрическая генетика /К. Мазер, Дж. Джинкс - Пер. с англ..- М.: Мир, 1985.- 463с.
4. *Плохинский Н.А*. Биометрия / Н.А. Плохинский - Новосибирск: Наука СО АН СССР, 1961.- 364 с.
5. *Снедекор Дж.У*. Статистические методы в приложении к исследованиям в сельском хозяйстве и биологии / Дж. У. Снедекор - М.: Сельхозиздат.- 1961.- 503 с.
6. *Урбах В.Ю*. Биометрические методы / В.Ю. Урбах - М.: Наука, 1964.- 415 с.
7. *Шеффе Г*. Дисперсионный анализ / Г. Шеффе - М.: Физикоматематическая лит-ра, 1963.- 625 с.
8. *Васильева Л.А*. Статистические методы в биологии: Учебное пособие к курсу лекций «Биометрия» / Л.А. Васильева – Новосибирск. 2004. – 127 с.

Содержание

ПРИЛОЖЕНИЯ

Список вопросов для подготовки к зачету по дисциплине «Статистический анализ в агрономии».

1. Классификация признаков биологических объектов.
2. Предмет, методы и задачи статистического анализа в агрономии.
3. Понятие о выборочных и генеральных совокупностях.
4. Статистические показатели, характеризующие количественную изменчивость. Среднее значение признака, мода, медиана.
5. Показатели изменчивости признака. Лимиты, дисперсия, варианса, среднее квадратическое отклонение, коэффициент вариации. Стандартные ошибки статистических параметров.
6. Виды группировки экспериментальных данных. Ранжирование данных.
7. Вариационный ряд и принципы его построения.
8. Графическое изображение вариационного ряда.
9. Стандартная выборочная ошибка. С какой целью её вычисляют?
10. Нормальное распределение случайной переменной.
11. Вероятность встречаемости различных вариант в нормальном распределении.
12. Распределение Пуассона.
13. Принцип построения треугольника Паскаля.
14. Уни- и полимодальное распределение.
15. Биномиальное распределение случайной переменной.
16. Асимметричное распределение случайной переменной.
17. Эксцессивное распределение случайной переменной.
18. Оценка параметров генеральной совокупности. Статистические гипотезы. Ошибки первого и второго рода, уровень значимости и мощность критерия.
19. Статистические критерии параметрической статистики.
20. Достоверность различий средних арифметических двух выборочных совокупностей. Критерий Стьюдента. Критерий хи-квадрат.
21. Наименьшая существенная разность (НСР).
22. Оценка связи между признаками. Коэффициент регрессии.
23. Коэффициент корреляции и его свойства.
24. Корреляционный анализ.
25. Регрессионный анализ.
26. Принципы дисперсионного анализа.
27. Однофакторный дисперсионный анализ
28. Двухфакторный дисперсионный анализ.
29. Модель I и II дисперсионного анализа. Коэффициент внутриклассовой корреляции.
30. Статистический анализ качественных признаков. Вероятность. Частоты. Среднее квадратическое отклонение, стандартная ошибка.

Таблица значений F-критерия Фишера при уровне значимости $\alpha = 0,05$

k_2 \ k_1	1	2	3	4	5	6	8	12	24	∞
1	161,5	199,5	215,7	224,6	230,2	233,9	238,9	243,9	249,0	254,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71

26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
35	4,12	3,26	2,87	2,64	2,48	2,37	2,22	2,04	1,83	1,57
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
45	4,06	3,21	2,81	2,58	2,42	2,31	2,15	1,97	1,76	1,48
50	4,03	3,18	2,79	2,56	2,40	2,29	2,13	1,95	1,74	1,44
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
70	3,98	3,13	2,74	2,50	2,35	2,23	2,07	1,89	1,67	1,35
80	3,96	3,11	2,72	2,49	2,33	2,21	2,06	1,88	1,65	1,31
90	3,95	3,10	2,71	2,47	2,32	2,20	2,04	1,86	1,64	1,28
100	3,94	3,09	2,70	2,46	2,30	2,19	2,03	1,85	1,63	1,26
125	3,92	3,07	2,68	2,44	2,29	2,17	2,01	1,83	1,60	1,21
150	3,90	3,06	2,66	2,43	2,27	2,16	2,00	1,82	1,59	1,18
200	3,89	3,04	2,65	2,42	2,26	2,14	1,98	1,80	1,57	1,14
300	3,87	3,03	2,64	2,41	2,25	2,13	1,97	1,79	1,55	1,10
400	3,86	3,02	2,63	2,40	2,24	2,12	1,96	1,78	1,54	1,07
500	3,86	3,01	2,62	2,39	2,23	2,11	1,96	1,77	1,54	1,06
1000	3,85	3,00	2,61	2,38	2,22	2,10	1,95	1,76	1,53	1,03
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1

Критические значения *t*-критерия Стьюдента при уровне значимости 0,05.

Число степеней свободы *k*	0,05
1	12,706
2	4,3027
3	3,1825
4	2,7764
5	2,5706
6	2,4469
7	2,3646
8	2,3060
9	2,2622
10	2,2281
11	2,2010
12	2,1788
13	2,1604
14	2,1448
15	2,1315
16	2,1199
17	2,1098
18	2,1009
19	2,0930
20	2,0860
21	2,0796
22	2,0739
23	2,0687
24	2,0639
25	2,0595
26	2,0555
27	2,0518
28	2,0484
29	2,0452
30	2,0423
40	2,0211
60	2,0003
120	1,9799
∞	1,9600

Составители

Кондратьева Инесса Витальевна

Кочнева Марина Львовна

Цильке Регинальд Александрович

СТАТИСТИЧЕСКИЙ АНАЛИЗ В АГРОНОМИИ

Методическое пособие

для практических занятий и самостоятельной работы

Редактор: Н.К. Крупина

Компьютерная верстка

Подписано к печати 2016 г.

Формат 60х84 1/16. объем уч.-изд. л.

Тираж экз. изд. № Заказ №

Отпечатано в издательстве НГАУ

630009, РФ, г. Новосибирск, ул. Добролюбова 160, офис 106. Тел.факс

(383) 267-09-10.