

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ  
АСТРАХАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

М.Ф. Козак, М.В. Козак

**БИОМЕТРИЧЕСКИЕ МЕТОДЫ  
В НАУЧНЫХ ИССЛЕДОВАНИЯХ**

*Монография*

Издательский дом «Астраханский университет»  
2019

УДК 578.8.58.59  
ББК 28.083.28.5.28.6  
К59

Рекомендовано к печати редакционно-издательским советом  
Астраханского государственного университета

*Рецензенты:*

заведующий кафедрой физиологии, морфологии, генетики и биомедицины  
Астраханского государственного университета,  
доктор биологических наук, профессор *Д.Л. Теплый*;  
доктор медицинских наук, профессор кафедры биохимии  
Астраханского государственного медицинского университета *О.В. Бойко*;  
доктор биологических наук, профессор кафедры гидробиологии и общей экологии  
Астраханского государственного технического университета, профессор *В.П. Иванов*

**Козак, М. Ф. Биометрические методы в научных исследованиях** : монография  
/ М. Ф. Козак, М. В. Козак. – Астрахань : Астраханский государственный университет,  
Издательский дом «Астраханский университет», 2019. – 168, [3] с.

Показана роль и место методов прикладной математики в исследованиях по биологии и смежных областях биомедицины, сельскохозяйственной биологии. Последовательно рассматриваются различные методы, необходимые для биологических исследований. В их числе: принципы составления случайной выборки, группировка первичного материала исследования на основе теории репрезентативности, точечные и интервальные оценки достоверности, статистический анализ вариации качественных и количественных признаков, параметрические и непараметрические методы, анализ закономерностей случайной вариации, соответствие эмпирических и теоретических распределений, расчет генеральных параметров, корреляционный и дисперсионный анализ. Рассматривается принцип каждого метода, схема его применения, ограничения и возможные ошибки использования. Для биологов-исследователей книга станет руководством по планированию, проведению и математической оценке достоверности результатов исследований, поможет начинающему исследователю достигнуть понимания связи статистических методов с задачами биологического и медицинского исследования.

Рекомендуется студентам биологических и экологических специальностей высших учебных заведений, а также аспирантам, магистрантам и исследователям в области биологии и смежных с ней областей биомедицины и сельскохозяйственной биологии, преподавателям научным работникам, изучающим и использующим математические методы самостоятельно.

ISBN 978-5-9926-1076-5

© Астраханский государственный университет,  
Издательский дом «Астраханский университет»,  
2019

© Козак М. Ф., Козак М. В., 2019

© Яценко Ю. А., оформление обложки, 2019

## ОГЛАВЛЕНИЕ

<b>СПИСОК УСЛОВНЫХ ОБОЗНАЧЕНИЙ .....</b>	<b>5</b>
<b>ВВЕДЕНИЕ .....</b>	<b>7</b>
<b>ГЛАВА 1. Основные понятия биометрии. Теория репрезентативности.</b>	
<b>Выборочный метод исследования .....</b>	<b>10</b>
1.1. Репрезентативность – степень соответствия выборочных показателей генеральным параметрам .....	11
1.2. Статистическое распределение выборки.....	12
1.3. Группировка дат при качественной и количественной вариации признака .....	14
1.4. Самостоятельная работа по разделу .....	18
<b>ГЛАВА 2. Основные статистические характеристики</b>	
<b>варьирующих признаков биологических объектов .....</b>	<b>20</b>
2.1. Степенные и структурные средние величины .....	20
2.2. Показатели варьирования.....	23
2.3. Самостоятельная работа по разделу .....	29
2.4. Программа «Методы описательной статистики в Excel» .....	31
<b>ГЛАВА 3. Закономерности случайной вариации. Основные типы</b>	
<b>распределений в биологии.....</b>	<b>33</b>
3.1. Вероятность. Основные свойства вероятности. Эмпирическая и теоретическая вероятность.....	33
3.2. Теоремы сложения и умножения вероятностей .....	38
3.3. Нормальное распределение в биологических совокупностях, его характеристика с помощью нормированного отклонения.....	41
3.4. Самостоятельная работа по разделу. Выравнивание эмпирических кривых распределения по нормальному закону .....	52
3.5. Распределение групп.....	55
<b>ГЛАВА 4. Оценка достоверности выборочных показателей .....</b>	<b>68</b>
4.1. Ошибки репрезентативности. Причины возникновения .....	68
4.2. Надежность. Пороги вероятности безошибочных прогнозов .....	70
4.3. Порядок оценки генеральных параметров .....	71
<b>ГЛАВА 5. Выборочный метод исследования</b>	
<b>и оценка генеральных параметров .....</b>	<b>77</b>
5.1. Принципы составления случайной выборки, точечные и интервальные оценки достоверности.....	77
5.2. Статистические гипотезы, их проверка .....	80

5.3. Параметрические и непараметрические критерии оценок достоверности .....	81
5.4. Причины асимметрии эмпирических распределений .....	90
<b>ГЛАВА 6. Определение достаточной численности выборки .....</b>	<b>93</b>
<b>ГЛАВА 7. Статистический анализ вариации</b>	
<b>качественных признаков.....</b>	<b>98</b>
7.1. Статистический анализ при альтернативной вариации .....	98
7.2. Дискретные переменные величины.....	102
7.3. Исследование вероятности распространения генов в популяции.....	107
<b>ГЛАВА 8. Корреляция-сопряженность признаков,</b>	
<b>их взаимосвязь .....</b>	<b>113</b>
<b>ГЛАВА 9. Дисперсионный анализ .....</b>	<b>122</b>
9.1. Сущность метода. Основные понятия и символы .....	122
9.2. Анализ однофакторных дисперсионных комплексов для количественных признаков .....	125
9.3. Анализ однофакторных дисперсионных комплексов для качественных признаков .....	131
9.4. Анализ двухфакторных дисперсионных комплексов для качественных признаков .....	135
9.5. Примеры заданий для самостоятельного изучения .....	140
9.6. Возможности многомерного анализа в биологии.....	143
<b>ЗАКЛЮЧЕНИЕ .....</b>	<b>145</b>
<b>ТЕРМИНОЛОГИЧЕСКИЙ СЛОВАРЬ .....</b>	<b>149</b>
<b>СПИСОК ЛИТЕРАТУРЫ .....</b>	<b>152</b>
<b>ПРИЛОЖЕНИЕ .....</b>	<b>157</b>

## СПИСОК УСЛОВНЫХ ОБОЗНАЧЕНИЙ

A, B, C, – постоянные величины; факторы и их градации в дисперсионном комплексе; случайные события;  
A – условная средняя, нулевой класс, где  $A = 0$ ;  
 $\alpha$  – уровень значимости (вероятность ошибки репрезентативности, ошибки I рода);  
a, b, c, ... – численность групп (градаций) факторов: A, B, C, в дисперсионных комплексах;  
a – отклонения вариант от условной средней, нулевого класса;  
CV – коэффициенты вариации;  
C – сумма квадратов отклонений вариант от их средних (девиата);  
d – разность между сравниваемыми величинами;  
F – критерий соответствия Р. Фишера;  
f – эмпирические частоты вариант в данной совокупности;  
f' – вычисленные или ожидаемые (теоретические) частоты вариант;  
f<sub>x,y</sub> – частоты вариант в клетках корреляционной таблицы;  
f<sub>x</sub> – первая функция нормированного отклонения;  
H – величина, используемая в дисперсионном анализе;  
H<sub>a</sub> – символ альтернативной гипотезы;  
H<sub>0</sub> – символ нулевой гипотезы;  
P – вероятность справедливости нулевой гипотезы  
 $\eta^2_x$  – показатель силы влияния фактора на результативный признак;  
i – порядковый номер варианты;  
K – биномиальный коэффициент;  
lim – [min ÷ max] – лимиты, символ, обозначающий границы вариации признака;  
Me – медиана;  
Mo – мода;  
N – объем генеральной совокупности; объем дисперсионного комплекса; число классов (групп) вариационного ряда; общая сумма сопоставляемых рядов;  
n – объем выборки; численность вариант в отдельных градациях;  
P – вероятность события;  
p – доля вариант, обладающих данным признаком (+);  
q – доля вариант, не обладающих данным признаком (–);  
R – ранг, или порядковый номер варианты в ранжированном ряду;  
r – коэффициент корреляции;  
 $s^x$  – среднее квадратическое отклонение;  
 $S^2_x$  – выборочная дисперсия (варианса);

$m, s_x$  – ошибка выборочной средней;  
 $m_p$  – ошибка доли;  
 $\bar{X}, x$  – выборочная средняя (средняя арифметическая);  
 $X_g$  – средняя геометрическая;  
 $X_h$  – средняя гармоническая;  
 $\mu$  (мю) – генеральная средняя; математическое ожидание;  
 $\nu$  (ню) – число степеней свободы;  
 $\Sigma$  (сигма прописная) – знак суммирования;  
 $\sigma$  – среднее квадратическое отклонение;  
 $\sigma^2$  – варианса;  
 $\chi^2$  (хи-квадрат) – критерий соответствия Пирсона;  
 $T$  – критерий Уилкоксона для независимых выборок;  
 $t$  – нормированное отклонение; критерий Стьюдента.

*Посвящается светлой памяти  
замечательного учителя и неутомимого  
исследователя – профессора МГУ Николая  
Александровича Плохинского.*

## **ВВЕДЕНИЕ**

Биология исследует проявления жизни в сложных совокупностях особей животных, растений, микроорганизмов, где роль математических методов анализа необычайно велика. Кроме элементов классической математики, широкое применение в современной биологии нашли методы прикладной математики, учитывающей специфику биологических явлений и особенности методов биологических исследований. Область биологии, анализирующая биологические явления с помощью методов прикладной математики, получила в России и за рубежом название биометрии. Основы пауки были заложены Ф. Гальтоном в 1899 г. Целью и содержанием биометрии являются планирование, проведение биологического эксперимента (или биологических наблюдений), статистический анализ результатов. Одна из задач данной работы – определение основных понятий, которые позволят их упорядочить и обозначить необходимость и направления применения при проведении биологических исследований. В зоологии, ботанике, генетике, цитологии, физиологии и других биологических науках математические методы используются для установления математических закономерностей, обнаруживаемых при анализе групповых биологических явлений в биологических совокупностях. Эти закономерности не применимы к отдельным единицам совокупностей. Значение биометрии в исследовательской работе биологов стало очевидным уже тогда, когда были открыты статистические законы, действующие в сфере массовых явлений. Биологи не сразу оценили важность этих открытий. Положение изменилось после того, как была обоснована теория малой выборки. Приоритет в этой области принадлежал В. Госсету (1876–1937), опубликовавшему в журнале «Биометрика» свой труд под псевдонимом «Стьюдент»<sup>1</sup>. Его теория малой выборки получила дальнейшее развитие в трудах К. Пирсона и Р. Фишера (1890–1962), Однако выдающийся датский ученый В. Иогансен<sup>2</sup> (1857–1927) в своих генетических опытах с фасолью пришел к выводу о том, что биологические проблемы должны решаться с помощью математики, но не только как математические задачи. «Статистике всегда должен предшествовать биологический анализ, иначе результаты могут быть

---

<sup>1</sup> Student. The probable Error of Mean // Biometrika. 1908. Vol. 6. P. 1–25.

<sup>2</sup> Иогансен В. Элементы точного учения об изменчивости и наследственности. М., 1933. С. 103.

«статистической ложью» (В. Иогансен). Это был новый подход к оценке роли математических методов в биологических исследованиях, показавший, что статистический анализ нельзя отрывать от биологического анализа. Роль математических методов в биологии в настоящее время необычайно велика и возрастает постоянно в связи с исследованием жизни на разных уровнях ее организации. Биометрия рассматривается как наука о математических методах анализа групповых биологических явлений. На протяжении последних лет активно развиваются исследования (и на их основе технологии) распознавания лица и голоса, анализ отпечатков пальцев, технологии распознавания лиц, обработка звуковых сигналов и видеоизображений. При этом на первый план выходит проблема оценки точности статистических показателей, критерии выбора биометрических методов, применяемых в научных исследованиях. В настоящее время биометрия это также комплекс постоянно развивающихся исследований, которые дали начало новой перспективной науке

По мнению Н.А. Плохинского (1970, 1978), эмпирический эксперимент - лишь начало исследования (10 %). Еще 90 % информации биолог получает в процессе математического анализа эмпирических данных исследования. Современные методы компьютерного анализа позволили глубоко проникнуть в математические закономерности биологических процессов. Однако, на первых этапах проникновения исследователя в *мир математических идей*, биолог-исследователь должен овладеть некоторыми специфическими понятиями, навыками, познакомиться с разнообразием методов прикладной математики, позволяющими успешно пользоваться языком математической биологии, чтобы стремительно и осмысленно погружаться в мир математической биологии. Математический анализ групповых биологических явлений не должен превращаться в оцифровку данных, что, к сожалению, происходит не редко. Учитывая непрерывное расширение сферы применения математических методов в биологии и развитие биометрии как самостоятельной научной дисциплины, настоящее издание посвящено погружению биолога-исследователя в мир математических методов анализа групповых биологических явлений на понятном материале из мира вероятностных биологических событий и явлений. В книге последовательно рассматриваются различные методы, необходимые для биологических исследований: группировка первичного материала исследования на основе теории репрезентативности, описательная статистика, параметрические и непараметрические методы, анализ закономерностей случайной вариации, соответствие эмпирических и теоретических распределений, оценка достоверности выборочных показателей, расчет генеральных параметров, корреляционный и дисперсионный анализ. Рассматривается принцип каждого метода, схема его применения, ограничения и возможные ошибки. Актуальность адекватного использования статисти-



ческих методов осознается все шире. И, хотя ошибки их применения не исчезли, все больше научных журналов прилагают усилия к их устранению. Во многих из них рецензирование включает отдельный этап проверки статистической обоснованности представляемых к публикации работ.

Данная работа написана на основе личного опыта преподавания математических методов биологии в Астраханском государственном университете (ранее Астраханском государственном педагогическом институте), где с 1989 г. началось преподавание этого курса многим поколениям студентов, аспирантов, магистрантов.

# ГЛАВА 1. ОСНОВНЫЕ ПОНЯТИЯ БИОМЕТРИИ. ТЕОРИЯ РЕПРЕЗЕНТАТИВНОСТИ. ВЫБОРОЧНЫЙ МЕТОД ИССЛЕДОВАНИЯ

В каждой естественной науке заключено  
столько истины, сколько в ней математики.

*И. Кант*

Предмет исследования биометрии – совокупность, группа биологических объектов, (*множество*), а также единичный биологический объект при многократном исследовании его с количественной стороны (например, артериальное давление пациента в течение определенного времени). Основная задача выборочного метода исследования: получить такую информацию, которая позволяет (более или менее точно) судить о состоянии *генеральной совокупности*. *Совокупность* – всякое множество отдельных, отличающихся друг от друга, но сходных в существенных отношениях объектов. Совокупностями являются: растения одного сорта на одном поле, одного вида на данной территории, популяция одного вида растений или животных на определенной территории (например, популяция большого баклана или узорчатого полоза в дельте реки Волги).

Объем совокупности – число единиц, ее составляющих. Каждый представитель совокупности может исследоваться по различным признакам. *Генеральная совокупность* (г.с.) – бесконечное множество относительно однородных величин, объектов: все представители данного вида, штамма микроорганизмов. Г.с. может мыслиться ограничено, например, животные одной породы конкретного района. Выборочная совокупность (выборка) – часть генеральной совокупности, отобранная рандомизированно, т.е. так, чтобы любой объект генеральной совокупности имел одинаковую вероятность попасть в выборку. Будем называть выборкой результаты конечного числа измерений, в некоторой идеальной совокупности, состоящей из бесконечного числа измерений. В то время как измерения, относящиеся к некоторой конечной выборке, можно охарактеризовать показателями вероятности, результаты идеальной, бесконечно большой совокупности измерений описываются параметрами.

### **1.1. Репрезентативность – степень соответствия выборочных показателей генеральным параметрам**

Основное требование к выборке – *репрезентативность* – представительность. *Репрезентативность* – это свойство выборочных групп характеризовать соответствующие генеральные совокупности с определенной точностью и надежностью. Репрезентативность – степень соответствия «выборочных» показателей генеральным параметрам. Каждый выборочный показатель имеет свою репрезентативность и может характеризовать соответствующий генеральный параметр с такой точностью и надежностью, с какой это позволяют основные детали организации «*выборочного исследования*». В биометрии существует понятие статистической совокупности. Статистическая совокупность – это множество относительно однородных, индивидуально различающихся единиц, объектов (особей), объединенных для совместного изучения. Недопустимо объединение в одну совокупность особей разного пола, возраста, изучение модифицирования признака на генетически неоднородном материале. Результат первичного измерения – *варианта (даты)*, в математике обозначается термином «величина» или случайная переменная величина, безотносительно к тем объектам, на которых проведено исследование. В биологии применяется более конкретный термин «значение признака». Это содержание вкладывается в термин «дата», «варианта», т. е. результат изучения отдельного признака у конкретного объекта. Биология, используя математические методы, исследует различные *категории признаков биологических объектов*:

- количественные признаки, счетные и мерные, поддающиеся точному измерению (масса, объем, длина);
- количественные признаки, не поддающиеся точному измерению, оцениваемые глазомерно в баллах (например, признаки экстерьера животных);
- качественные признаки (альтернативные), имеющие только две степени проявления – есть или нет: пол, женский или мужской; мутация, есть или нет;
- качественные признаки, имеющие много проявлений (например, масть лошадей: серые, буланные, вороные, каурые, рыжие, гнедые);
- порядковые, не измеряемые, но ранжируемые: цвет меха норок, соболей, лисиц; интенсивность окраски оперения птиц.

Точность выводов, сделанных с помощью биометрических методов, зависит от следующих факторов:

- 1) воображения, гибкости исследователя, понимания сущности и особенностей биологического процесса;
- 2) объективности методов составления выборки;
- 3) аккуратной регистрации первичных данных;

4) оптимальности выбора биометрических методов анализа и статистической оценки данных.

Биометрический анализ становится тем более эффективным, чем точнее соблюдаются при проведении эксперимента перечисленные условия. Однако ни один биометрический метод не может улучшить «плохие» первичные данные. *«Вычисления можно производить как угодно точно, но результат вычисления не может быть точнее тех данных, на которых оно основано»* – академик. А. Н. Крылов<sup>3</sup>.

Биологу, использующему математические методы, как инструмент исследования, необходимо овладеть определенными навыками и специфической терминологией прикладной математики с тем, чтобы, погружаясь в мир компьютерных программ, использовать их вполне эффективно и профессионально.

Каждая выборочная совокупность (множество) в своей основе структурно подразделена. Биологу-исследователю необходимо выявить эти элементы структуры, *составив статистическое распределение выборки* в зависимости от категории признаков биологических объектов.

## **1.2. Статистическое распределение выборки**

Допустим, что из генеральной совокупности рандомизированно извлечена выборка. Оказалось: варианта  $x_1$  встретилась  $n_1$  раз,  $x_2$  –  $n_2$  раза,  $x_k$  –  $n_k$  раз. Объем выборки:  $n_1 + n_2 + \dots + n_k = n$ . Числа:  $n_1, n_2, n_k$  называются частотами, а их отношения к общему объему выборки – относительными частотами:

$$\frac{n_1}{n} = p_1, \quad \frac{n_2}{n} = p_2, \quad \frac{n_x}{n} = p_x.$$

Сумма относительных частот равна единице:

$$p_1 + p_2 + p_x = \frac{n_1 + n_2 + \dots + n_x}{n} = 1.$$

Перечень вариантов и соответствующих им частот (или относительных частот) называется *статистическим распределением выборки*.

Статистическое распределение выборки можно представить в виде последовательности интервалов и соответствующих им частот (при непрерывном распределении): В качестве частот, соответствующих интервалу, принимают сумму вариантов, попавших в этот интервал. *Распределение* (с точки зрения теории вероятностей) – это соответствие между возможными значениями случайной переменной величины и их вероятностями. Распределение (в математической статистике) – это соответствие между наблюдаемыми вариантами и их частотами.

---

<sup>3</sup> Крылов А. Н. Лекции о приближенных вычислениях. М., 1933. С. 486.

**Пример.** Количество лопастей на листьях шелковицы черной варьирует от 3 до 7 (варианты: 1, 2, 3, 4, 5, 6, 7). При интервале 2, установлена следующая взаимосвязь между вариантами и их частотами:

*Распределение листьев шелковицы черной по числу лопастей листа*

Количество лопастей листа, $x_i$	3	5	7
Относительные частоты, $p_i$	0,24	0,36	0,40

Относительные частоты:

$$p_1 = \frac{12}{50} = 0,24; p_2 = \frac{18}{50} = 0,36; p_3 = \frac{20}{50} = 0,40.$$

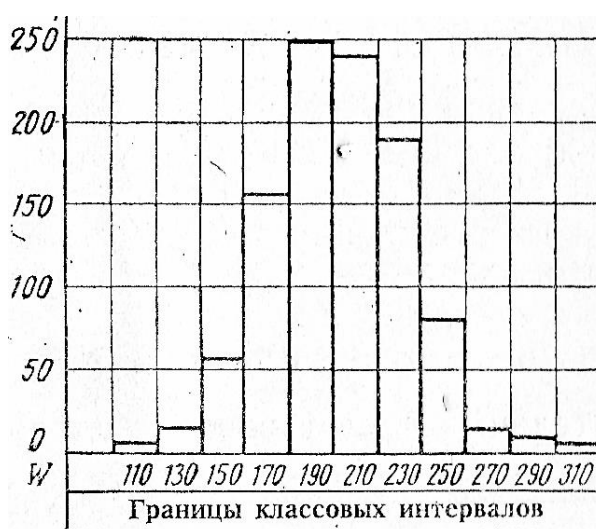


Рис. 1.1. Гистограмма

Графическим изображением статистического распределения являются полигоны распределения и гистограммы (рис. 1). Для построения полигона распределения на горизонтальной оси  $X$  откладывают значения вариантов ( $x_i$ ). На оси  $Y$  наносятся значения частот  $n_i$  или относительных частот  $p_i \cdot (f_i)$ . Полигоном чаще пользуются в случае небольшого числа вариантов, при построении графиков без интервального вариационного ряда.

Соединяя вершины перпендикуляров, высота которых соответствует частотам классов, прямыми линиями, получают геометрическую фигуру в виде многоугольника – полигон распределения частот. Линия, соединяющая вершины перпендикуляров, называется вариационной кривой (рис. 1.1) или кривой распределения частот вариационного ряда. При большом количестве дат и в случае непрерывного распределения вероятностей, чаще строят гистограммы распределения частот. Для этого на оси абсцисс откладывают границы классовых интервалов, а по оси ординат – частоты интервалов (рис. 1.2). Если из середины верхних сторон прямоугольников гистограммы опустить

перпендикуляры на ось абсцисс, а верхние точки соединить вариационной кривой, то гистограмма превращается в полигон распределения. Графические изображения распределений – удобный и наглядный способ их сравнения, особенно, когда на одном графике нанесено несколько распределений.

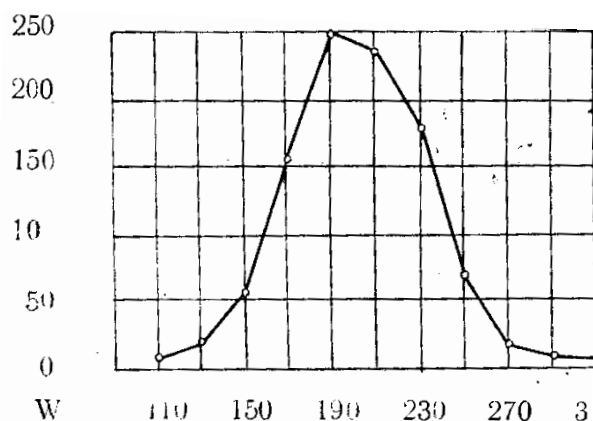


Рис. 1.2. Вариационная кривая

Вариационный ряд – упорядоченное расположение реально существующего распределения особей в группе по степени развития признака. Это двойной ряд чисел, состоящий из значений классов и соответствующих им частот. Методика составления вариационных рядов приводится в таблице 1.3 (алгоритм). Зафиксированные в документах учета сведения об изучаемом объекте (или объектах) представляют тот первичный фактический материал, который нуждается в соответствующей обработке и анализе, которые начинаются с упорядочения или систематизации собранных данных. Процесс систематизации результатов массовых наблюдений, объединения их в относительно однородные группы по некоторому признаку называется группировкой первичных данных.

### 1.3. Группировка дат при качественной и количественной вариации признака

**При качественной вариации** распределение вариантов (дат) совокупности выражается количеством особей определенного фенотипа, процентным соотношением их, или долями (пример, табл. 1.1).

**При количественной дискретной вариации признака.** Различия между вариантами (датами), отдельными значениями случайной переменной величины в этом случае выражаются целыми числами, между которыми не бывает переходов. Группировка дат в этом случае проводится по значениям отдельных вариантов (пример, табл. 1.1). Варианты (даты) в такой совокупности выражены целыми числами. Варьирующая величина, изменяющаяся под

влиянием многих случайных факторов и принимающая различные значения (количество щенков на одну самку), называется случайной переменной  $x_i$ .

Таблица 1.1

**Расщепление по фенотипу во втором поколении ( $F_2$ ) дигибридного скрещивания дрозофилы (пример группировки дат при качественной вариации признака)**

№ класса	Классы расщепления (фенотип)	Кол-во особей	% от общего числа особей	Доля, $p(f)$
1	Серое тело, нормальные крылья	280	56,0	9/16
2	Серое тело, редуцированные крылья	94	18,8	3/16
3	Черное тело, нормальные крылья	94	18,8	3/16
4	Черное тело, редуцированные крылья	32	6,4	1/16
Итого		500	100,0	16/16

Таблица 1.2

**Распределение 80 лисиц по количеству щенков в помете**

Классы (вариации): количество щенков на одну самку	1	2	3	4	5	6	7	8	9
Вероятность встречаемости (количество случаев)	1	4	10	39	13	7	3	2	1

При очень большой вариации признака классы не целесообразно намечать по значениям каждой вариации, так как вариационный ряд получится растянутым с перерывами в некоторых классах.

Целесообразно наметить классы, охватывающие несколько значений вариантов. Классовый промежуток ( $k$ ) – есть постоянное число. При количественной непрерывной вариации признака нет реально существующих классов. Эти классы необходимо наметить самому исследователю в соответствии с существующими правилами и рекомендациями, представленными в таблице 1.3. Разница между максимальным и минимальным значением вариантов – лимиты:

$$450 - 375 = 75.$$

Этот интервал надо разбить па определенное количество классов ( $g$ ). Если принять за оптимальное количество классов 7 или 8, классовый промежуток будет равен 10. Он должен быть целым числом или округленными дробями. Начало первого класса не обязательно должно совпадать со значением минимальной варианты. Величина классового промежутка должна быть одной и той же для всех классов. Одна и та же варианта не должна входить одновременно в два класса: например, ею должен заканчиваться предыдущий класс и не должен начинаться следующий. Число классов при статистическом анализе материала зависит от объема исследуемой совокупности ( $n$ ). На практике можно руководствоваться следующими рекомендациями статистики:

Таблица 1.3 (алгоритм)

### Составление вариационного ряда

Первичные данные (даты)																			
413	450	419	412	427	435	404	430	421	399	414	386	428	441	397	417	418	414	429	417
423	420	416	407	427	428	417	398	424	420	401	424	411	426	375	419	406	419	429	406
414	410	409	416	430	403	426	407	400	423	425	391	432	409	418	418	388	421	415	417
423	434	402	431	410	405	436	405	424	405	412	413	444	392	411	428	394	431	411	422
433	395	433	420	439	398	437	422	394	416	424	434	408	443	407	421	422	410	423	409
Число классов $g = 1 + 3,3lg\ n = 1 + 3,3lg\ 100 = 7,6$								Вариационный ряд											
Число дат		Число классов		Классы				Разноска		Частоты, $f$									
6–11		4		начало, $W_0$		середина, $W$													
12–22		5		455–		450		.		1									
23–46		6		435–		440		Π		7									
47–93		7		425–		430		☒ ☒		20									
94–187		8		415–		420		☒ ☒ ☒		30									
188–377		9		405–		410		☒ ☒ ☒		25									
378–755		10		395–		400		☒		10									
756–1515		11		385–		390		Γ.		6									
1516–3050		12		375–		380		.		1									
Размах $p = \max - \min = 450 - 375 = 75$ $k = \frac{p}{g} = \frac{75}{7,6} = 9,9 \approx 10$						–		–		–		$n = 100$							
<p>Начала классов <math>W_0</math> используются для разности дат по классам, середины классов <math>W</math> служат как представители классов при расчетах по способу взвешенных вариаций.</p> <p>Начало класса равно полусумме середин данного класса и следующего низшего:</p> $W_0 = \frac{W_i + W_{i-1}}{2} = \frac{450 + 440}{2} = 445.$ <p>Середина класса (вариация) равна полу сумме начала данного класса и следующего высшего:</p> $W_i = \frac{W_{\alpha(i)} + W_{\alpha(i+1)}}{2} = \frac{435 + 445}{2} = 440.$																			



Рекомендуемое количество классов в зависимости от объема выборки:

Количество вариант (дат) совокупности, n	Число классов, g
25–40	5–6
40–60	6–8
60–100	7–10
100–200	8–12
Более 200	10–15

#### Основные элементы вариационного ряда

- Классы, в которых собраны особи, сходные по фенотипу.
- Вариации (или середины классов).
- Классовые промежутки, равные разности вариаций соседних классов.
- Частоты – число особей каждого класса.
- Объем совокупности (распределения), общее число особей в группе.

В качестве примера рассмотрим распределение 100 початков кукурузы по их длине сорта ВПР-42 (табл. 1.4). Среди 100 початков, отобранных *рандомизированно*, большинство початков имеют длину приблизительно 16–18 см, хотя абсолютный размах (интервал) варьирования составляет 14,5–20,5 см. Средняя арифметическая рассчитывается путем деления суммы величин длины всех 100 початков на их число (100). Взвешенная средняя арифметическая рассчитывается с учетом математического веса вариации путем деления суммы величин  $\sum fx_i$  на число 100:

$$\bar{X}(M) = \frac{\sum fV}{n} = \frac{\sum fx_i}{n} = \frac{\sum 15 \cdot 4 + 16 \cdot 19 + 17 \cdot 34 + 18 \cdot 28 + 19 \cdot 11 + 20 \cdot 4}{100} = 17,35 \text{ см.}$$

Таблица 1.4

#### Распределение по длине (см) 100 початков кукурузы сорта ВПР-42 (статистическое распределение выборки)

№ класса	Классы (от... до...), min÷ max	Частота встречаемости в выборке, $f$	Среднее значение класса, $V(x_i)$
1	14,5–15,5	4	15
2	15,6–16,5	19	16
3	16,6–17,5	34	17
4	17,6–18,5	28	18
5	18,6–19,5	11	19
6	19,6–20,5	4	20
$\Sigma$		100	

Биолог-исследователь лишь на первых этапах использования биометрических методов испытывает определенные сложности, легко преодолеваемые с приобретением навыков применения математических методов в своей работе.

#### **1.4. Самостоятельная работа по разделу**

##### **СТАТИСТИЧЕСКОЕ РАСПРЕДЕЛЕНИЕ ВЫБОРКИ**

*Цель работы:* изучение правил составления статистических распределений выборочной совокупности.

*Задачи:*

- освоение методики первичной группировки выборочных данных при составлении вариационных рядов;
- изучение правил и методики составления статистического распределения выборки

*Материал для исследования:*

1. Материалом для исследования могут быть любые сравнимые объекты природы, представленные во множестве ( $n = 25-100$ ) для составления выборочных совокупностей. Они могут быть различными, но подлежащие сравнению по конкретным признакам методами биологической статистики.

*Содержание работы:*

1. Пользуясь литературой, познакомиться с понятиями: совокупность (выборочная и генеральная), случайная переменная величина, варианта (дата), объем совокупности, объем выборки, с основными правилами составления совокупности, особенностями группировки первичных данных исследования при дискретной и непрерывной вариациях признаков биологических объектов.

2. Объяснить сущность выборочного метода исследования, причины варьирования признаков биологических объектов.

3. Составить выборочную совокупность на основе экспериментальных данных, полученных в результате исследования конкретных количественных признаков биологических объектов.

4. Занести данные исследования в таблицу 1.3.1 (объем выборки),

5. Найти значения максимальной и минимальной даты, определив размах варьирования признака (лимиты).

Таблица 1.3.1

**Форма фиксации первичных данных исследования**

№ п/п	Значения исследуемого признака у отдельных особей совокупности $x_i$ (даты)		
	Длина початка кукурузы, см (1)	Количество зерен в початке, шт. (2)	Плотность початка (3)
1			
2			
3			
4			
...			
25			
n			
$\Sigma$ (сумма)			
M (среднее)			

6. Разбить выборочную совокупность (1) на классы (вариации) и определить количество дат каждого класса, (частота встречаемости их в совокупности). Все результаты занести в таблицу 1.3.2.

Таблица 1.3.2

**Статистическое распределение выборки  
(Возможная форма итоговой записи)**

**Исследуемый объект, признак**

Номер класса	Классы (min ÷ max)	Количество вариант класса, шт.	Среднее значение класса, вариации, $V_i$	Вероятность (относительная частота встречаемости), $f$	Модальный класс
1					
2					
3					
...					
6					

*Вопросы*

1. Что такое биологическая совокупность (множество)?
2. Привести примеры различных биологических совокупностей.
3. Чем отличается выборочная и генеральная совокупности?
4. Что такое варианта, дата, случайная, переменная величина?
5. Что такое вариационный ряд?
6. Каковы особенности распределения вариантов в вариационном ряду?
7. Каковы возможные причины многовершинности и асимметрии вариационных кривых?

## ГЛАВА 2. ОСНОВНЫЕ СТАТИСТИЧЕСКИЕ ХАРАКТЕРИСТИКИ ВАРЬИРУЮЩИХ ПРИЗНАКОВ БИОЛОГИЧЕСКИХ ОБЪЕКТОВ

Главная задача, которую решает исследователь с помощью методов биологической статистики, заключается в том, чтобы на основе изучения выборочной совокупности сделать выводы о свойствах генеральной совокупности. Выводы, которые делаются на основе изучения выборки, могут иметь лишь вероятностный характер. При решении статистической задачи исследуются массовые характеристики совокупности. Исследователь, собирая материал, идет от свойств выборки к свойствам генеральной совокупности. Построение статистической модели будет успешным, если первый этап, организации эксперимента был сделан правильно, если не была нарушена процедура составления репрезентативной выборки. Так, безосновательно судить о росте студентов института на основе исследования роста членов баскетбольной или гандбольной команды или описывать вкусовые свойства нового сорта арбуза по нескольким «типичным» экзemplярам. Вариационные ряды и их графики дают информацию о варьировании признаков, но они недостаточны для характеристики варьирующих объектов. Для этого служат статистические характеристики: средние величины и показатели вариации. Значение средних величин заключается в их способности уравнивать все индивидуальные отклонения, отражать качественное своеобразие варьирующего объекта, что позволяет отличить одну статистическую совокупность от другой.

В качестве статистических характеристик неравно интервальных вариационных рядов служит плотность распределения – отношение частот к ширине классовых интервалов. В качестве статистических характеристик равноинтервальных вариационных рядов используют степенные и структурные средние величины.

### 2.1. Степенные и структурные средние величины

Степенные средние вычисляются на основе общей формулы:

$$M = \left[ \frac{\sum x_i^k}{n} \right]^{1/k} \text{ или } M = \sqrt[k]{\frac{\sum x_i^k}{n}},$$

где  $M$  – средняя величина,  $x_i$  – варианта (дана),  $n$  – объем выборки,  $k$  – вид сравнения,  $k = 1$  – средняя арифметическая,  $k = 2$  – среднее квадратическое. Допустим, в выборке, состоящей из  $n$  особей надо измерить какой-нибудь признак: например, вес или рост. Обозначим величину признака у особи  $i$  как  $x_i$ . Тогда можно вычислить среднюю арифметическую – усредненное значение величин признака в выборке из  $n$  особей:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i,$$

где черта над  $x$  указывает на среднюю величину, а греческая буква  $\Sigma$  – на сумму величин признаков у всех особей. *Статистические показатели для характеристики совокупности* – это количественные показатели, которые логически и теоретически обоснованы и позволяют судить о качественном своеобразии варьирующих признаков биологических объектов и объективно сравнивать эти объекты между собой.

*Средние величины дают характеристику среднему уровню развития признака в исследуемой совокупности.*

**Степенные средние:**

а) средняя арифметическая:

$$M = \frac{\Sigma V}{n} \text{ или } \bar{X} = \frac{\Sigma X_i}{n} = \frac{1+2+3+4+5}{5} = 3;$$

б) взвешенная средняя арифметическая рассчитывается с учетом математического веса даты:

$$\bar{X}(M) = \frac{\Sigma fV}{n} = \frac{\Sigma fX_i}{n}; \quad \bar{X}(M) = \frac{v_1 p_1 + v_2 p_2 + \dots + v_n p_n}{p_1 + p_2 + \dots + p_n};$$

в) средняя геометрическая – равна корню  $n$  – степени из произведения всех дат. Основным критерием для применения средней геометрической является возрастание (увеличение) значений данного признака для усреднения приростов (при онтогенетической изменчивости):

$$G = \sqrt[n]{\pi v_i} = \sqrt[n]{v_1 \cdot v_2 \dots \cdot v_n},$$

$$\bar{x}_g = (x_1 x_2 x_3 \dots x_i \dots x_N)^{1/N} = \left( \prod_{i=1}^N x_i \right)^{1/N}.$$

Заглавная греческая буква  $\Pi$  (пи) указывает на умножение всех отдельно взятых величин.

г) средняя гармоническая  $\bar{X}_g$ . Совокупность: 1, 2, 3, 4, 5

$$\bar{X}_g = \frac{N}{\Sigma \frac{1}{V}} = \frac{5}{\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5}} = 2.2.$$

$$\bar{x}_h = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} \dots \frac{1}{x_i} \dots \frac{1}{x_N}} = \frac{N}{\sum_{i=1}^n \frac{1}{x_i}}.$$

При определении эффективного размера популяции, численность которой меняется в разных поколениях, более рационально вычислить гармоническую среднюю.

### **Структурные средние:**

а) мода ( $M_o$ ) – наиболее часто встречающееся значение признака в данном вариационном ряду.

1 – 2 – 3 – 4 – 5 – 5 – 6 – 7 – [7] – 7 – 8 – 9 – 9 – 10 – 11 – 11 – 12;

б) медиана ( $M_e$ ) – величина, по обе стороны которой встречается одинаковое количество дат:

1 – 2 – [3] – 4 – 5.

### **Математические свойства среднего арифметического (средней арифметической величины)**

На перечисленных ниже свойствах основаны различные методы вычисления среднего арифметического, а также возможности оперирования показателем: среднее арифметическое:

1. Если каждую из вариантов совокупности уменьшить или увеличить на одну и ту же величину, то средняя арифметическая величина изменится на такую же величину. В алгебраическом выражении: если совокупность:  $x_1, x_2, x_3, x_4 \dots x_n$ , имеющая среднюю арифметическую  $\bar{X}(M)$ , будет заменена совокупностью:  $(x_1 - a), (x_2 - a), (x_3 - a), (x_4 - a)$ , то средняя арифметическая второй совокупности будет:  $(\bar{X} - a)$ .

2. Алгебраическая сумма отклонений отдельных вариантов от средней арифметической равна нулю:  $\sum (x_i - \bar{X}) = 0$ .

3. Сумма квадратов центральных отклонений от средней арифметической меньше суммы квадратов отклонений от любой другой величины  $A$ , не равной  $\bar{X}(M)$ , т. е.  $\sum (x_i - \bar{X})^2 < \sum (x_i - A)^2$ , если  $A \neq \bar{X}$ . Эти свойства позволяют применить непрямой способ вычисления среднего арифметического и других биометрических показателей с помощью условной средней.

### **Сущность среднего арифметического:**

1. Среднее арифметическое является обобщающей величиной данной совокупности. Этот показатель отражает уровень всей совокупности, дает обобщающую характеристику исследуемого признака.

2. Цифровое значение средней арифметической величины может не встретиться среди вариантов совокупности. В этом смысле среднее арифметическое является абстрактной величиной: среднее количество детей в одной семье в данной местности – 1,5; среднее количество плодов на одно растение арбуза 8,25; среднее количество детенышей у жемчужной норки на одну самку (♀) 4,31.

3. Средняя арифметическая величина в тоже время конкретна: она выражается в тех же наименованиях, что и варианты ряда.

4. При определении средней арифметической погашаются случайные колебания вариантов.

5. Средняя арифметическая характеризует всю совокупность как целое, но не отдельные варианты. Так, число детенышей на одну самку у жемчужной норки, равное 4.31, относится ко всей группе, хотя каждая отдельная норка может дать от 1 до 7 детенышей.

6. Средняя арифметическая величина имеет смысловое значение по отношению к количественно однородной совокупности: например, для каждой возрастной группы животных, определенной фенологической фазы развития растений. Для каждой такой группы организмов характерна специфическая степень развития признака.

7. Механический перенос средней арифметической на явления, выходящие за рамки данной совокупности, неправилен без специального анализа этого вопроса. Особое место в биометрии имеет вопрос о том, на основе каких данных можно сделать выводы о других подобных совокупностях.

## 2.2. Показатели варьирования

### *Учет разнообразия (варьирования) признака*

Любая группа организмов состоит из объектов, отличающихся друг от друга по каждому из признаков. Эти различия могут быть почти незаметными или очень большими. Из всех групповых свойств наибольшее теоретическое и практическое значение имеет средний уровень развития признака. Вторым основным свойством любой группы (множества) является *разнообразие признака*. Часто этому групповому свойству даются другие названия: *изменчивость, варьирование, рассеяние, разброс* и другие. Термин «разнообразие» достаточно точно отражает *свойство группы состоять из неодинаковых объектов по любому признаку*. Причиной варьирования признака в популяциях микроорганизмов, животных, растений, человека являются: генетический полиморфизм, фенотипическая изменчивость, возникающая под влиянием генотипа и условий среды. При одинаковых средних значениях признаки могут отличаться по величине и характеру варьирования.

### *Лимиты и размах варьирования*

$\rho$  ( $\rho_0$ ) =  $[x_{min} \div x_{max}]$  – разность (интервал) между минимальной и максимальной вариантами совокупности (абсолютный размах варьирования признака).

**Пример.** Исследование диаметра пяти колоний у двух штаммов микроорганизмов (мм) дало следующие результаты:

Первый штамм: 2,0 – 2,2 – 2,4 – 2,6 – 2,8;  $M_1 = 2,4$ ;  
 $lim = 2,0 \div 2,8$ ;  $\rho_1 = 0,8$ .

Второй штамм: 1,6 – 2,0 – 2,4 – 2,8 – 3,2;  $M_2 = 2,4$ ;  
 $lim = 1,6 \div 3,2$ ;  $\rho_2 = 1,7$ .

Средний диаметр колоний оказался одинаковым, но изучение лимитов и размаха варьирования выявили большие различия штаммов по этому признаку. Эти характеристики совокупности не отражают характера варьирования признака. Кроме того, они могут изменяться для данной совокупности при повторных исследованиях. Поэтому, кроме лимитов и размаха варьирования, необходимо исследование других показателей разнообразия признака. Широкое применение в биометрии получили следующие показатели:

*Дисперсия* (от лат. *dispersio* – рассеяние) – сумма квадратов центральных отклонений

*Варианса (средний квадрат)*. Рабочие формулы, методика расчетов приводятся в алгоритмах (табл. 2.3.2 и 2.3.3).

В математике и биометрии существует несовпадение (разнообразие) в терминах «варианса» и «дисперсия». Математики назвали сумму квадратов центральных отклонений вариансой (от англ. *variance* – изменение), биометрики (Н.А. Плохинский, 1970, 1978; Дж Снедекор, 1964 и др.) – дисперсией, а варианса – средний квадрат. Кроме того, для обозначения суммы квадратов центральных отклонений используется предложенный Д. Юлом и М. Кендэлом (1960) термин «девиата» (от лат. *deviatio* – отклонение). Значение дисперсии и вариансы заключается в том, что они являются показателем варьирования числовых значений признака около средней арифметической и мерой изменчивости признака, зависящей от разности результатов наблюдений:

а) лимиты или размах варьирования ( $min \div max$ );

б) дисперсия (C) – сумма квадратов центральных отклонений частных средних от общей средней:

$$C = \sum (x_i - \bar{X})^2, \text{ или } C = \sum (V - M)^2,$$

или

$$C = \sum V^2 - \frac{(\sum V)^2}{n} = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 - \frac{15^2}{5} = 10.$$

При необходимости учета веса дат:

$$C = \sum f (V - M)^2 = \sum f \alpha^2,$$

где  $(V - M) = \alpha$ ;

в) средний квадрат (варианса):

$$\sigma^2 = S^2 = \frac{C}{v} = \frac{C}{n - 1},$$

где  $v = n - k$ ;  $v$  – число степеней свободы – количество свободно варьирующих элементов, равно числу всех имеющихся элементов без числа огра-



нений. Распределение варьирующих величин вокруг средней арифметической величины может быть различным. Например, если все особи одинакового роста или веса, то изменчивость почти отсутствует. Если в выборке равное количество высоких и низких особей, то получаются крайние варианты изменчивости. *Варианса* служит мерой отклонения дат от средней арифметической и вычисляется следующим образом:

$$V_x = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

*Варианса* – средняя сумма квадратов отклонений отдельных величин от средней арифметической. Выражение  $n - 1$  в знаменателе используется вместо  $n$ , поскольку вычисленная таким образом дисперсия является несмещенной оценкой, это будет более правильной оценкой при усреднении параметров выборки любого размера.

г) среднее квадратическое отклонение или стандартное отклонение ( $\sigma$ ), показывает средний размах варьирования признака:

$$\sigma = S = \pm \sqrt{\frac{C}{n-1}}$$

или

$$\sigma = \pm \sqrt{\frac{\sum f (Xi - \bar{X})^2}{n-1}} = \pm \sqrt{\frac{\sum f \alpha^2}{n-1}}.$$

Для группы особей: 1, 2, 3, 4, 5;  $n = 5$ ,  $M(\bar{X}) = 3$ ,

$$C = 10; \sigma = \pm \sqrt{\frac{10}{4}} = \pm 1,58,$$

где  $\sigma$  – показатель именованный, выражается в тех же единицах, что и среднее арифметическое.

Стандартное отклонение представляет собой квадратный корень из вариансы и измеряется в тех же единицах, что и среднее арифметическое. Математически, нормальное распределение является симметричным, со средней и вариансой, определяемыми стандартным отклонением. Для удобства в статистике иногда принимают величину стандартного отклонения, близкую к нормальному распределению. Выборки из природных популяций близки по распределению значений признаков к нормальному распределению;

д) коэффициент вариации (CV) является показателем степени изменчивости признака, используется для сравнения изменчивости признаков в различных вариационных рядах:

$$CV = \frac{100(V_x)^{1/2}}{\bar{x}},$$

или

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

или

$$CV = \frac{S_{\bar{X}}}{\bar{X}} \cdot 100 \%;$$

е) нормированное отклонение ( $t$ ).

Отклонение той или иной варианты от среднего арифметического, выраженное в количестве сигм ( $\sigma$ ):

$$t = \frac{(x_i - \bar{X})}{S_{\bar{X}}} = \frac{(V - M)}{\sigma}.$$

Наиболее общей формулой для вычисления дисперсии (среднего квадрата) является:

$$\sigma^2_x = S^2_x = \frac{\sum_i^k (x_i - \bar{X})^2}{n}$$

или

$$\sigma^2_x = \frac{\sum_i^k f_i (x_i - \bar{X})^2}{n},$$

где  $[\sum_{i=1}^k]$  – знак суммы всех произведений отклонений вариантов от среднего арифметического, умноженных на вес дат от первого до  $k$ -класса. Исследованиями установлено, что вычисленные таким способом сумма квадратов и средний квадрат смещены в сторону от своего генерального параметра на величину  $n(n-1)$ . Для получения несмещенной величины эта формула была преобразована:

$$\sigma^2_x = S^2_x = \frac{\sum_{i=1}^k (x_i - \bar{X})^2}{n} \cdot \frac{n}{n-1} = \frac{\sum_{i=1}^k f_i (x_i - \bar{X})^2}{n-1}$$

где  $n - 1 = k = \nu$  (ню) – число степеней свободы ( $k$ ).

При наличии не одного, а нескольких ограничений свободы варьирования число ее степеней будет:  $\nu = n - k$ .

### Свойства дисперсии

1. Если каждую варианту уменьшить (или увеличить) на одно и то же число ( $A$ ), то дисперсия не изменится. Следовательно, дисперсию можно вычислить не только по значениям варьирующего признака, но и по их отклонениям от какой-либо постоянной величины  $A$ .

2. Если умножить (или разделить) каждую варианту совокупности на одно и то же число ( $A$ ), то дисперсия увеличится (или уменьшится) в  $A^2$  раз. Таким образом, если совокупность состоит из многозначных дат, то каждую из них можно уменьшить на число  $A$  и на основе уменьшенных дат вычислить дисперсию.

Среднее квадратическое отклонение,  $\sigma$ , вместе с дисперсией лучше всего характеризует не только величину, но и характер варьирования признака. Кроме формул, приведенных выше, существует другие формулы расчета среднего квадратического отклонения. Все они дают одинаковый результат.

*Пример:* изучалось систолическое давление у двух групп спортсменов (мм рт. ст.) после нагрузки:  $n_1 = 10$ ;  $n_2 = 10$ .

I группа: 100–110–120–130–140–150–160–170–180–190.

II группа; 100–145–145–145–145–145–145–145–145–190.

$M_1 = 145$ ;  $lim_1 = 100 \div 190$ .  $M_2 = 145$ ;  $lim_2 = 100 \div 190$ .

Дисперсия:  $C = \sum(V-M)^2 = \sum(x_i - \bar{X})^2$ .

$$c_1 = (-45)^2 + (-35)^2 + (-25)^2 + (-15)^2 + (-5)^2 = \\ = +52 + 152 + 252 + 322 + 452 = 8250;$$

$$c_2 = (-45)^2 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 452 = 4050;$$

$$\sigma_1 = \pm \sqrt{\frac{8250}{9}} = \pm 30,3 \text{ мм}, \quad \sigma_2 = \pm \sqrt{\frac{4050}{9}} = \pm 21,2 \text{ мм}.$$

*Вывод:* несмотря на совпадение средних значений систолического давления и их лимитов, статистические показатели варьирования оказались разными.

*Расчет дисперсий* (табл. 2.1). Для отражения сущности дисперсии, как суммы взвешенных квадратов центральных отклонений частных средних от общей средней, рассмотрим предыдущий пример (см. табл. 1.2)

Таблица 2.1

**Распределение по длине 100 початков кукурузы сорта (гибрида) ВИР-42  
(расчет дисперсии и среднего квадратического отклонения)**

№ класса	Классы (от... до...) $min \div max$	Частота встречаемо- сти в выборке (f)	Среднее значение класса ( $x_i$ )	Центральные отклонения $\alpha = (x_i - \bar{X})$	$\alpha^2 =$ $(x_i - \bar{X})^2$	$f\alpha^2$
1	14,5–15,5	4	15	–2,35	5,5	22,1
2	15,6–16,5	19	16	–1,35	1,8	34,3
3	16,6–17,5	34	17	–0,35	0,1	4,2
4	17,6–18,5	28	18	+0,65	0,4	11,8
5	18,6–19,5	11	19	+1,65	2,6	28,8
6	19,6–20,5	4	20	+2,65	7,0	28,1
$\bar{X}$	17,35					
$C = \sum f\alpha^2$	Дисперсия					129,6

$C = \sum f\alpha^2$  – дисперсия – сумма взвешенных квадратов центральных отклонений частных средних от общей средней.

Среднее квадратическое отклонение ( $\sigma$ ):

$$\sigma = \pm \sqrt{\frac{\sum f (X_i - \bar{X})^2}{n - 1}} = \pm \sqrt{\frac{\sum f \alpha^2}{n - 1}} = \pm \sqrt{\frac{129,6}{99}} = \pm 1,14 \text{ см.}$$

Среднее квадратическое отклонение ( $\sigma$ ) служит основным показателем разнообразия признака в группе. Сигма используется как самостоятельный показатель и необходим для вычисления других статистических показателей: ошибка репрезентативности, коэффициент корреляции, коэффициент вариации, различных показателей распределения признака. Среднее квадратическое отклонение может непосредственно служить для сравнения разнообразия групп, однако только при соблюдении следующих условий:

а) при сравнении одних и тех же признаков у разных групп;

б) если сравниваемые группы не очень отличаются по величине  $\bar{X}$  (M).

Использование *коэффициента вариации* (CV %) снимает эти ограничения, так как в отличие от дисперсии и среднего квадратического отклонения является не абсолютным, а относительным показателем вариации, он показывает долю сигмы ( $\sigma$ ) от среднего арифметического в процентах:

$$CV = \frac{\sigma \cdot 100 \%}{M}.$$

При симметричных распределениях коэффициент вариации обычно не превышает 50 %. При асимметричных рядах распределения коэффициент вариации может достигать 100 % и даже более процентов.

*Нормированное отклонение* ( $t$ ) показывает, на сколько  $\sigma$  отклоняются значения крайних дат от среднего арифметического. В рассматриваемом нами примере лимиты:  $lim = 14,5 \div 20,5$ ;  $\bar{X} = 17,35$  см;  $\sigma = 1,14$  см;  $t_{min} = \frac{14,5-17,35}{1,14} = -2,6 \sigma$ ;  $t_{max} = \frac{20,5-17,35}{1,14} = +3,0 \sigma$ . Размах варьирования длины початка кукурузы не выходит за пределы:  $\bar{X} \pm 3\sigma$ .

*Коэффициент вариации* (CV) – это стандартное отклонение, деленное на среднее арифметическое, выраженное в процентах:

$$CV = \frac{100(V_x)^{1/2}}{\bar{x}},$$

Коэффициент вариации позволяет сравнивать изменчивость в различных выборках с разными значениями средней арифметической.

Стандартная ошибка средней (ошибка репрезентативности) равна:

$$se = \left( \frac{V_x}{n} \right)^{1/2},$$

или

$$se = m = \pm \frac{\sigma}{\sqrt{n}}.$$

Эту величину можно применять для оценки достоверности выборочной средней. Часто распределение величин не является нормальным или его нельзя рассматривать как нормальное. В таких случаях, полезно оценивать доверительный интервал вокруг средней с помощью других методов.

### 2.3. Самостоятельная работа по разделу

Конечным результатом изучения той или иной совокупности по определенным признакам является составление эмпирического вариационного ряда, его графического изображения в виде полигона или гистограммы и вычисление основных статистических показателей ( $\bar{X}$ ,  $S$ ,  $\sigma$ ,  $\sigma^2$ ), определение вероятности ошибки репрезентативности ( $p$ ).

*Цель работы:* изучение смысла и значения биометрических показателей, которые позволяют судить о данной совокупности в целом, о разнообразии данного признака внутри группы, а в дальнейшем, об отличиях ее от других совокупностей.

#### Содержание работы

1. Пользуясь первичными данными, полученными ранее на основе изучения статистического распределения выборки, вычислить и оценить две группы статистических показателей для характеристики совокупности с непрерывной вариацией признака.

Таблица 2.2

**Статистические показатели для характеристики совокупностей**

	Признак	Статистические показатели среднего уровня				Статистические показатели варьирования признака					
		M	$\bar{X}$	Me	Mo	lim	$\sigma$	$\sigma^2$	C	CV	t
1	Длина початка кукурузы										
2	Количество зерен в початке										

Таблица 2.3 (алгоритм)

**Вычисление M и  $\sigma$  без составления вариационных рядов для малых групп**

Даты малозначные	Даты многозначные
<p>Каждая дата возводится в квадрат; даты и их квадраты суммируются; на основе полученных сумм <math>\sum V</math> и <math>\sum V^2</math> рассчитываются:</p> <p>средняя арифметическая</p> $M = \frac{\sum V}{n};$ <p>сумма квадратов</p> $C = \sum V^2 - \frac{(\sum V)^2}{n};$	<p>Для каждой даты определяется условное отклонение:</p> $\Delta = V - A,$ <p>где <math>A</math> – любое число.</p> <p>Каждое отклонение возводится в квадрат; на основе двух сумм <math>\sum \Delta</math> и <math>\sum \Delta^2</math> рассчитываются:</p> <p>средняя арифметическая</p> $M = A + \frac{\sum \Delta}{n};$

сигма			сумма квадратов			
$\sigma = \sqrt{\frac{c}{n-1}}$			$C = \sum \Delta^2 - \frac{(\sum \Delta)^2}{n};$			
			сигма			
			$\sigma = \sqrt{\frac{c}{n-1}}$			
	V	V <sup>2</sup>		V	Δ(V – 2400)	Δ <sup>2</sup>
1	12	144	1	2536	136	18496
2	9	81	2	2703	303	91809
3	10	100	3	2815	415	172225
4	13	169	4	2487	87	7569
5	15	225	5	2644	244	59536
6	14	196	6	2521	121	14641
7	8	64	7	2452	52	2704
8	12	144	8	2463	63	3969
	∑ V = 93	∑ V <sup>2</sup> = 1123		–	∑ Δ = 1424	∑ Δ <sup>2</sup> = 370949
$M = \frac{93}{8} = 1,6;$			$M = 2400 + \frac{1421}{8} = 2577,6;$			
$C = 1123 - \frac{93^2}{8} = 41,88$			$C = 370949 - \frac{(1421)^2}{8} = 118544$			
$\sigma = \sqrt{\frac{41,88}{7}} = \pm 2,44$			$\sigma = \sqrt{\frac{118544}{7}} = \pm 130,13$			

Таблица 2.4 (алгоритм)

**Вычисление М и σ без составления вариационных рядов**

$\bar{X} = \frac{\sum x_1}{n} \left( M = \frac{\sum V}{n} \right)$										
для больших и малых групп										
	413	450	419	412	427	435	404	430	421	399
	414	386	428	441	397	417	418	414	429	417
	432	420	416	407	427	428	417	398	424	420
	401	424	411	426	375	419	406	419	429	406
	414	410	409	416	430	403	426	407	400	423
	425	391	432	409	418	418	388	421	415	417
	423	434	402	431	410	405	436	405	424	405
	412	413	444	392	411	428	394	431	411	422
	433	395	433	420	439	398	437	422	394	416
	424	434	408	443	407	421	422	410	423	409
$\sum V$	4191	4157	4202	4197	4146	4172	4148	4157	4170	4134
										$\sum V = 41674$

$$C = \sum \Delta^2 - \frac{(\sum \Delta)^2}{n};$$

$$\sigma = \sqrt{\frac{c}{n-1}};$$

$$n = 100$$

$\sum v^2$	1757349	1731959	1767320	1763761	1721682	1741846	1732070	1729121	1740186	1709590	$\sum v^2 = 17385884$
<p>средняя арифметическая</p> $M = \frac{41674}{100} = 416.7;$ <p>сумма квадратов</p> $C = 17385884 - \frac{41674^2}{100} = 18661;$ <p>сигма</p> $\sigma = \sqrt{\frac{18661}{99}} = 13.71$											

2. Вычисленные статистические показатели занести в таблицу 2.2. Нанести на построенные ранее графики медиану, моду, среднее взвешенное арифметическое, среднее квадратическое отклонение.

3. Дополнить выводы о размахе и характере варьирования признаков в двух исследуемых вариационных рядах с учетом вычисленных биометрических показателей. Пользуясь алгоритмом, освоить методику расчета статистических показателей для экспериментальных данных, не сгруппированных в вариационный ряд.

### Вопросы

1. Какие две группы показателей позволяют характеризовать вариационные ряды?
2. Совпадают или нет значения средней арифметической медианы и моды?
3. Сущность и свойства среднего арифметического.
4. Каковы способы вычисления статистических показателей, если данные не сгруппированы в вариационный ряд?
5. Среднее квадратическое отклонение как характеристика изменчивости признака в совокупности.
6. Как вычисляются взвешенные  $\bar{X}$  и  $\sigma$ ?

## 2.4. Программа «Методы описательной статистики в Excel»

Для ознакомления с принципами работы компьютерной программы «Методы описательной статистики» в «Excel» рекомендуется сравнить полученные статистические показатели с теми же показателями, но полученными с помощью программы «Методы описательной статистики в Excel», обращаясь

к источникам, рекомендованным в списке литературы, например, А.В. Фролов (2015) и др. С помощью этого инструмента, используя ресурсы программы, можно в короткие сроки обработать массив данных, а также получить о нем информацию по ряду статистических критериев, изучить, как работает данный инструмент для дальнейшего использования. Под описательной статистикой понимают программу систематизации эмпирических данных по ряду основных статистических критериев. На основе анализа результатов, полученных итоговых показателей, можно сформировать общие выводы об изучаемом массиве данных. В «Excel» существует отдельный инструмент, входящий в пакет анализа, с помощью которого можно провести данный вид обработки данных, который называется *описательной статистикой*. Среди критериев, которые рассчитывает данный инструмент, есть следующие уже знакомые показатели: медиана; мода; дисперсия; среднее взвешенное арифметическое; среднее квадратическое отклонение (стандартное отклонение); стандартная ошибка; коэффициент вариации, коэффициент асимметрии и др.

*Примечание.* Инструмент «Описательная статистика» входит в более широкий набор функций, который называется «Пакет анализа». По умолчанию данная надстройка в «Excel» отключена, для использования возможностей описательной статистики, ее следует включить.



### ГЛАВА 3. ЗАКОНОМЕРНОСТИ СЛУЧАЙНОЙ ВАРИАЦИИ. ОСНОВНЫЕ ТИПЫ РАСПРЕДЕЛЕНИЙ В БИОЛОГИИ

**Биологическая статистика** имеет дело не с единичными явлениями или объектами, а с их **совокупностями**. Отдельные члены совокупности, как правило, в той или другой степени отличаются друг от друга, *варьируют*. Каждый из них представляет собой *отдельный случай*, который осуществляется под влиянием многих определяющих причин. Однако этих причин может быть так много, что обнаружение их для каждого отдельного случая становится невозможным. Каждое *отдельное явление*, взятое само по себе, представляется случайным, но, взятые в массе, они обнаруживают определенные, **статистические закономерности**. Поэтому **можно предсказать результаты для массового явления в целом**. В отношении каждого единичного явления, каждого отдельного члена совокупности приходится говорить только об известной возможности, или вероятности, значения, которое они приобретают. Биологу нередко приходится встречаться с вероятностями, большими или малыми, *ответить на многие вопросы можно, с определенной долей вероятности*.

#### 3.1. Вероятность. Основные свойства вероятности.

##### Эмпирическая и теоретическая вероятность

*Вероятность – это возможность осуществления определенного события в некотором количестве случаев из общего числа возможных (степень уверенности в том, что событие произойдет).*

Опыт, эксперимент, наблюдение явления называется испытанием. Результат, исход испытания называется событием.

- Два события называются совместимыми, если появление одного из них не исключает появления другого в одном и том же испытании.
- Два события называются несовместимыми, если появление одного исключает появление другого в том же испытании.
- Два события (А и В), называются противоположными, если в данном испытании они несовместимы, а одно из них обязательно происходит.

Вероятностью Р события (А) называется отношение числа элементарных событий, благоприятствующих событию А к числу всех элементарных событий:  $P(A) = \frac{m}{n}$ . Иными словами, вероятность – числовая мера объективной возможности осуществления события А.

*Основные свойства вероятности*

- Вероятность достоверного события равна единице:

$$P(A) = \frac{m}{n} = \frac{n}{n} = 1.$$

- Вероятность невозможного события равна нулю:

$$P(A) = \frac{m}{n} = \frac{0}{n} = 0.$$

- Вероятность случайного события есть положительное число, заключенное между нулем и единицей:

$$0 < P(A) \leq 1.$$

- Число  $m$  – абсолютная частота события  $A$ .
- Отношение  $P(A) = \frac{m}{n}$  – относительная частота события  $A$ .

### *Статистическое определение вероятности*

Вероятностью случайного события называется число, около которого группируются частоты этого события по мере увеличения числа испытаний. Преимущество статистического способа определения вероятности состоит в том, что оно опирается на реальный эксперимент. Его недостаток заключается в том, что для определения вероятности необходимо выполнить большое число опытов. Вероятность события  $A$  есть число  $P(A)$ , около которого группируются значения относительной частоты при больших объемах испытаний. Так, относительная частота рождения девочек по месяцам по данным шведской статистики колеблется около числа 0,482 ( $0,462 \div 0,491$ ). Следовательно, относительная частота события приблизительно совпадает с его вероятностью, если число испытаний велико. В биологии для упрощения символики принято числовое значение вероятности первого события обозначать строчной латинской буквой  $p$ , а значение вероятности противоположного события – буквой  $q$ :  $P(A) = p$ ,  $P(\bar{A}) = q$ ;  $p + q = 1$ .

### *Эмпирическая и теоретическая вероятность*

В приведенных выше примерах биологических совокупностей рассматривались конкретные эмпирические и теоретические вероятности. Эмпирические вероятности применимы к конкретным совокупностям. Так, вероятность рождения детей с отрицательным резусом крови относится к определенной изученной группе людей. Для популяции, в которой много людей с отрицательным резусом вероятность рождения детей с резус-отрицательной кровью (фенотип  $Rh^-$ ) возрастает. В практике важно судить не только об отдельных случаях, но и всех возможных случаях этих явлений. Математическая теория, имея дело с отдельными, частными наблюдениями, выработала методы, позволяющие по отдельным, частным наблюдениям судить о тех случаях, которые имели бы место, если бы изучалась не только данная совокупность, но и теоретически мыслимая совокупность всех возможных случаев этого рода. По эмпирическим, опытным вероятностям, основанным на учете конкретных

частот, можно судить о теоретических (априорных) вероятностях, таких, о которых можно знать заранее, до проведения опыта. Теория вероятностей дает возможность построить абстрактные совокупности, представляющие собой отражение реальных генеральных совокупностей.

### *Распределение*

Вариация признаков биологических объектов есть результат совместного действия многих разнонаправленных и независимых друг от друга факторов на развитие особей (и их признаков), входящих в состав совокупности. Признаки особей варьируют в пределах границ, детерминируемых генотипом, и принимают в каждом конкретном случае одно из множества возможных значений. Таким образом, вариационный ряд с характерной для него концентрацией дат вблизи его центральной части и рассеиванием к краям ряда является в то же время и распределением вероятностей. Это значит, что в вариационном ряду случайная переменная величина « $x_i, v_i$ » принимает разные значения:  $x_1, x_2, x_3 \dots x_n$  под влиянием различных причин, как правило, не зависящих друг от друга. Поэтому вариацию величины  $x$  можно рассматривать как случайную. Отдельным значениям  $x_i$  соответствуют вероятности  $p_i$ :  $p_1, p_2, p_3 \dots p_n$ . Совокупность значений  $x_i$  и соответствующих им вероятностей  $p_i$  называется *распределением вероятностей*.

### *Понятие: случайные величины*

В математике случайной величиной называется переменная величина, которая в зависимости от исхода испытания, случайно принимает одно значение из множества возможных значений.

*Примеры случайных переменных величин в биологии:*

- 1) значения длины початков кукурузы одного сорта на одном поле;
- 2) количество гомозиготных особей из 10 лабораторных мышей второго поколения гибридов: 1, 2, 3, 4, 5...10 ( $p = 0,5$ );
- 3) прирост междоузлий стебля тыквы за каждый месяц вегетации.

Случайная величина, принимающая различные значения, которые можно записать в виде последовательности только определенных, фиксированных значений, которые выражаются целыми числами, называется *дискретной случайной переменной величиной*.

Случайная величина, которая может принимать все значения из некоторого числового промежутка, называется *непрерывной случайной величиной*.

## Необходимость и случайность

**Необходимая** связь явлений: если при осуществлении события *A* возможно только событие *B*. Но если же при осуществлении события *A* равно возможны *B* и *C*, мы имеем дело с **проявлением возможности в виде случайного**.

**Случайное** – это такое же **объективное явление**, как и необходимое, и оно так же обусловлено различными причинами, как и необходимое, только характер причинности здесь иной, а именно: возможен не один, а два результата или более. Эти возможности являются вероятностями.

Процесс осуществления явления на основе известной его возможности, или вероятности, **называется вероятностным или стохастическим**. Теория вероятностей изучает математические законы таких процессов.

**Вероятность** можно выразить математически по следующей формуле:  $p = \frac{m}{n}$ , где *m* – число благоприятных случаев, *n* – число всех возможных, или равновероятных, случаев.

*Пример:* Если на каждой из сторон кубика обозначены цифры 1, 2, 3, 4, 5, 6, то вероятность того, что наверху будет цифра 4, равна  $\frac{1}{6}$ , т.к. из всех возможных положений кубика *может быть шесть, и лишь один случай благоприятный*.

Какова вероятность того, что при одновременном выбрасывании двух кубиков сумма цифр наверху равна 6. Для этого следует рассчитать все возможные случаи сочетания цифр в двух кубиках:

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

Всего возможно 36 случаев ( $n = 36$ ). Благоприятных случаев, когда цифры двух кубиков дают в сумме число 6, будет только 5 (они подчеркнуты). Значит, вероятность выбрасывания двух кубиков с суммой цифр наверху, равной 6, может быть выражена формулой:

$$p = \frac{m}{n} : p = \frac{5}{36}.$$

На подобных примерах, **или моделях**, решаются многие задачи теории вероятностей.

*Пример.* Особи мужского и женского пола у многих видов животных рождаются примерно в равном количестве. Это значит, что на каждые 100 потомков в среднем может родиться примерно 50 самок (♀) и 50 самцов (♂), отсюда вероятность рождения телочки или бычка равна 0,5 (или 50 %).

*Пример.* Чтобы оценить вероятность рождения белого теленка, надо знать количество рождавшихся ранее в данном стаде или данной породе белых и рыжих животных. Так, если в данной популяции за несколько последних лет обнаружено 110 белых телят из общего количества 55000 родившихся, то вероятность рождения белого теленка равна  $p = 0,002$ . Это значит, что, в среднем, на каждые 1000 случаев приходится только 2 случая рождения белых телят.

*Противоположная вероятность*, обратная величине  $p$ , что родится не комолый, а рогатый теленок. Эта вероятность обозначается буквой  $q$ . Она выражается в данном случае величиной, равной 0,998. *Алгебраическая сумма величин  $p$  и  $q$  равна 1, т.е. сумма вероятностей противоположных событий равна единице.*

Из этих примеров вытекает теоретический вывод: количественной характеристикой вероятности явления *может быть относительная частота явления, установленная эмпирически на достаточно значительном фактическом материале.* Если некоторое явление имеет вероятность  $p$ , то относительная частота его, обнаруживаемая в опыте или при наблюдении, будет близка к  $p$ , при этом она будет тем ближе к  $p$ , чем больше было проведено опытов или наблюдений. Если заранее дается определенная вероятность, то по ней можно найти ожидаемую частоту явления, которая будет получена при проведении опытов. Зная относительную частоту, можно найти приближенное значение вероятности, которое может служить для характеристики частоты данного явления в последующих опытах *или наблюдениях. Вероятность проявляется только при большом числе наблюдений или опытов.* Приведенные примеры показывают, что вероятности  $p$  и соответствующие им  $q$  могут иметь самые разные значения – от величин, близких к нулю или равных ему, до величин, близких к единице или равных ей. Так, вероятность рождения белого теленка очень мала. Однако существует немало событий, обладающих еще меньшей вероятностью. Наконец, если  $p = 0$ , то на совершение данного события вообще нельзя рассчитывать. Но могут быть события, вероятность которых, хотя и очень близка к нулю, но все же нулю не равняется. Теоретически можно утверждать, что вероятность обнаружения в стаде породистых животных с жирностью молока 6,5 % равна нулю, но возможен все же какой-то исключительный случай, когда корова может дать молоко исключительно высокой жирности. Очевидно, вероятность подобного события будет выражена очень малой дробью.

### *Маловероятные события*

События, обладающие очень малой вероятностью, осуществляются вполне закономерно, хотя они могут казаться невозможными. Маловероятные события (например, мутации гена) при многократном повторении явления приобретают вполне устойчивую и определенную вероятность их осуществления, хотя бы такое событие происходило в одном случае из миллионов. Так, с точки зрения вероятности возникновение жизни на Земле является необычайно редким событием. Но каким бы невероятным показалось возникновение жизни на Земле, времени для этого было достаточно, поэтому оно могло произойти хотя бы один раз, что было уже достаточно для дальнейшего развития жизни.

### *Достоверные события*

По мере приближения величины  $p$  к единице событие становится все более достоверным. Если  $p = 1$ , то событие вполне достоверно.

#### *Явления, обладающие малой вероятностью*

Оценка того, насколько мала вероятность, чтобы с ней можно было не считаться, в значительной мере зависит от степени важности события. Так, если вероятность воздействия нового удобрения на понижение урожая равна 0.05, это не должно помешать его применению, так как в 0,95 случая оно окажется полезным. Однако если новый лекарственный препарат может с вероятностью, равной 0,05, принести не пользу, а вред организму пациента, его применение не может быть допущено. В опытах и наблюдениях заранее намечают приемлемую величину (или уровень) вероятности и считают ее достаточной для доказательства получения того или иного эффекта.

## **3.2. Теоремы сложения и умножения вероятностей**

Для понимания закономерностей случайной вариации важны две теоремы:

1. *Теорема сложения вероятностей* – относится к таким независимым друг от друга событиям, которые несовместимы.

2. *Теорема умножения вероятностей* – относится также к независимым событиям, но совместным друг с другом или следующим друг за другом.

*Пример 1.* На грядке растут 20 красных, 30 синих и 40 белых астр. Какова вероятность сорвать в темноте окрашенную астру?

*Ответ.* Вероятность сорвать в темноте окрашенную астру равна сумме вероятностей:  $P = \frac{20}{90} + \frac{30}{90} = \frac{50}{90}$ .

*Пример 2.* Какова вероятность того, что при выбрасывании двух кубиков, на гранях которых написаны цифры от 1 до 6, наверху будет сумма цифр не менее 10?

*Ответ.* Эта вероятность состоит из суммы трех вероятностей: получить сумму цифр 10, сумму 11 и сумму 12. Первая вероятность складывается из данных выше сочетаний трех цифр,  $p_{10} = \frac{3}{36}$ , вторая  $p_{11} = \frac{2}{36}$ , третья  $p_{12} = \frac{1}{36}$ . Сумма их составит  $\frac{6}{36}$ , или  $\frac{1}{6}$ .

Для умножения вероятностей необходимо, чтобы второе событие  $E_2$  осуществлялось только при осуществлении события  $E_1$  при этом осуществлении  $E_1$  не влияет на вероятность осуществления  $E_2$ , т.е. события  $E_1$  и  $E_2$  независимы.

*Пример 1.* Какова вероятность наличия цифры 4 наверху двух выброшенных одновременно кубиков? При выбрасывании одного кубика вероятность появления цифры 4 равна  $\frac{1}{6}$ . При выбрасывании второго кубика вероятность та же  $-\frac{1}{6}$ . Общая вероятность  $p = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$ .

*Пример 2.* Какова вероятность прохождения по лабиринту с шестью развилками и шестью тупиками? Очевидно, что на каждом развилке вероятности попасть или в тупик, или к следующей развилке одинаковы:  $\frac{1}{2}$ . Тогда при наличии шести развилки общая вероятность ( $p$ ) будет равна:

$$\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2^6} = \frac{1}{64}.$$

### *Теорема сложения вероятностей*

Если успех может осуществляться двумя и более различными, взаимоисключающими способами, полная вероятность успеха равна сумме индивидуальных вероятностей.

*Пример.* Какова вероятность того, что при двукратном выбрасывании кубика, на гранях которого написаны цифры от 1 до 6, наверху окажется цифра 5 или 6?

Эта вероятность складывается из суммы двух вероятностей:

- вероятность цифры 5 наверху  $-\frac{1}{6}$ .
- вероятность цифры 6 наверху  $-\frac{1}{6}$ .

Вероятность «цифры 5» или «цифры 6» наверху равна  $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$ . Таким образом, вероятность осуществления любого одного или нескольких взаимоисключающих «успехов» равна сумме индивидуальных вероятностей.

Если вероятность «успеха» равна  $p$ , а вероятность неудачи  $q$ , то вероятность либо успеха, либо неудачи равна  $(p + q)$ .

Если точно известно, что данное событие должно быть либо успехом, либо неудачей, то  $p + q = 1$ ,  $p = 1 - q$  и  $q = 1 - p$ .

### *Теорема умножения вероятностей*

Если второе событие наступает только при осуществлении первого и при этом наступление первого не влияет на вероятность второго, то вероятность осуществления нескольких независимых успехов равна произведению их индивидуальных вероятностей.

### *Эмпирические и теоретические вероятности*

В приведенных выше примерах вычислялись так называемые **эмпирические вероятности**. Они применимы только к тем конкретным совокупностям, для которых они вычислены. *Для практики же очень важно судить не только об отдельных конкретных случаях, но и всех возможных случаях этого рода.* Математическая теория, имея дело с отдельными, частными наблюдениями, выработала методы, позволяющие по результатам наблюдений судить о тех случаях, которые имели бы место, если бы изучалась не только данная совокупность осуществившихся случаев, но и теоретически мыслимая совокупность всех возможных случаев этого рода. По эмпирическим, опытным вероятностям, основанным на учете конкретных относительных частот тех или других явлений, можно судить о теоретических, или так называемых априорных, вероятностях, т.е. таких, которые можно брать заранее, до проведения опыта. В предыдущих разделах было приведено несколько вариационных рядов. Каждый из них являлся результатом изучения некоторого, сравнительно небольшого числа объектов. Так, суждение о плодовитости серебристо-черных лисиц, ее средней величине и изменчивости было сделано по 80 экземплярам. Но можно было бы изучить не эту маленькую группу лисиц, а всех лисиц, разводимых в РФ. Такая совокупность всех конкретных объектов, которую можно было бы изучить, называется, как уже сказано выше, генеральной (популяция). Изученная же небольшая группа представляет собой выборку из генеральной совокупности, поэтому ее называют выборочной. Можно себе представить и теоретически мыслимую совокупность, т.е. совокупность всех возможных наблюдений, в том числе и таких, которые практически не были осуществлены. Такую совокупность называют *стохастической*.

Теория вероятностей дает возможность построить абстрактные совокупности, представляющие собой отображение реальных совокупностей. В таких *абстрактных стохастических* совокупностях, доступных точному математи-



ческому анализу, вероятности становятся *теоретическими*. В биологии исследователь встречается, как правило, с выборочными совокупностями, но по ним стремится судить о генеральной или стохастической совокупности. Так, для изучения признаков окуня данного озера нет необходимости изучать всю его популяцию, т.е. генеральную совокупность. Достаточно взять выборочную совокупность в количестве 100, 200 или 1000 особей. По капле крови больного нередко делают выводы о состоянии всей крови, данные об изменчивости нескольких сотен обоей *Artemia salina* из озера *Тунаки* позволяли судить обо всей популяции *Artemia salina*.

*Распределение вероятностей – математическая основа вариации признаков*. Если бы все особи популяции были сходны, то уже по одной особи можно было бы получить полную информацию обо всей генеральной совокупности, всей популяции. Но в действительности существует очень большое разнообразие, как среди самих особей популяции, так и в отношении условий внешней среды, в которой они живут и развиваются. Поэтому и проводимые многократно выборки из генеральной совокупности, никогда не будут одинаковыми. Возникает вопрос: каковы закономерности вариации внутри каждой совокупности и каково взаимоотношение между разными типами совокупностей. Это дает возможность ответить на вопрос, можно ли по статистическим показателям, полученным на основе изучения одной совокупности, например, выборочной, судить о статистических показателях других видов совокупности, например генеральной. Это вопрос о том, насколько достоверны статистические показатели, полученные по выборочной совокупности, чтобы можно было судить по ним о генеральной совокупности.

### **3.3. Нормальное распределение в биологических совокупностях, его характеристика с помощью нормированного отклонения**

*Нормальное распределение* занимает важнейшее место в статистике вообще и в биологической статистике в частности, так **как очень многие эмпирические распределения биологических признаков, характеризующиеся непрерывной вариацией**, приближаются к нормальному, следуют ему.

#### *Теоретическая основа вариации*

**Фенотип** – результат реализации генотипа в конкретных условиях внешней среды в результате совместного действия многих разнонаправленных и независимых друг от друга факторов. Согласно теореме А.М. Ляпунова, если случайная величина является суммой большого числа независимых слагаемых, то она с достаточной степенью вероятности будет распределяться по нормальному закону. *Закон нормального распределения* – один из ос-

новых законов статистических явлений. Закономерности, которым подчиняются случайные события, изучаются в разделах математики, называемых теорией вероятностей и математической статистикой. Методы теории вероятностей и математической статистики широко используются в биологии. При исследовании биологических совокупностей, всюду в природе прослеживается закономерность: в однородных совокупностях большинство ее членов оказываются близкими к среднему уровню признака, чем они сильнее отклоняются от среднего уровня, тем реже встречаются. Следовательно, существует связь между числовыми значениями варьирующих признаков и частотой их встречаемости. Впервые эти закономерности распределения объектов в биологических совокупностях исследованы бельгийским математиком А. Кетле. *Закон распределения вероятностей случайной переменной величины ( $x$ ), называется нормальным*, если функция  $f(x)$  определяется формулой:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2} \frac{(x-\bar{X})^2}{\sigma^2}}.$$

Для нормального распределения характерно совпадение по абсолютным значениям средней арифметической, медианы и моды. Равенство этих показателей указывает на нормальность распределения.

Кривая нормального распределения одновременно с максимумом при  $\bar{X} = \mu$ , имеет две точки перегиба: при  $\bar{X} = (\mu - \sigma)$  и  $\bar{X} = (\mu + \sigma)$ , она симметрична относительно перпендикуляра, опущенного из вершины на ось абсцисс ( $\mu$  – математическое ожидание  $\bar{X}$ ). Согласно теореме выдающегося математика А.М. Ляпунова (1857–1918), если случайная переменная величина  $X_i$  представляет собой сумму большого числа взаимно независимых случайных величин, влияние каждой из которых на всю сумму ничтожно мало, то  $X$  имеет распределение, близкое к нормальному. Это утверждение именуется как центральная предельная теорема. В биологии очень распространены ситуации, близкие к условиям центральной предельной теоремы. Значения различных количественных признаков растений, животных, человека являются результатом длительной эволюции на разных уровнях организации жизни. Процесс формирования каждого признака включает бесконечное число разнонаправленных воздействий, поэтому большинство различных эмпирических распределений биологических признаков, характеризующихся непрерывной изменчивостью, следуют ему. Нормальным распределением является такое, которое с достаточной степенью приближения соответствует закону К. Гаусса, А. Муавра, П. Лапласа:

$$p' = f' = \frac{nk}{\sigma} \cdot \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}},$$

где

$$\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} = f(x),$$

где  $f(x)$  – первая функция нормированного отклонения;

$p'$  – теоретически ожидаемые частоты (ординаты нормальной кривой);

$n, k, \sigma$  – объем совокупности, классовый промежуток, среднее квадратическое отклонение.

$$\pi = 3,1416, \quad e = 2,71828,$$

$$x(t) = \frac{V - M}{\sigma} = \frac{x_i - \bar{X}}{\sigma} = f(x).$$

Функция  $f(x)$  связывает значения  $x_i$  (случайной переменной величины) с их вероятностями.

$$\text{Нормированное отклонение } f' = \frac{nk}{\sigma} f(x)$$

Для изучения закономерностей вариации при нормальном распределении вероятностей широко пользуются нормированным отклонением, которое обозначим буквой  $t^*$ .

*Нормированное отклонение* – отклонение той или другой варианты (или группы вариантов) от средней арифметической, выраженное в количестве сигм, отсюда:  $x_i - \bar{x} = t\sigma, t = \frac{x_i - \bar{x}}{\sigma}$ .

**Нормированное отклонение** ( $t^*$ ) имеет несколько более широкий смысл, и оно может выражаться не только в сигмах. Каждая варианта характеризуется определенным значением **нормированного отклонения**  $t$ , указывающим ее положение в вариационном ряду или на кривой распределения. Так, если варианта № 19 имеет значение  $t = +1,5$ , это значит, что она располагается в правой части кривой на расстоянии от  $M$  в  $1,5 \sigma$ . Если варианта № 28 имеет значение  $t = -2,6$ , она расположена в левой части кривой на расстоянии от  $\bar{X} (M)$  в  $2,6 \sigma$ .

В нормальном распределении отклонение вариантов от среднего арифметического захватывает приблизительно  $6\sigma$ , или  $\pm 3\sigma$ . Размещение вариантов в вариационном ряду имеет свои закономерности. Поэтому, если распределение является нормальным, можно заранее прогнозировать, сколько (%) особей находится в пределах: 1, 2, 3  $\sigma$ :

а)  $\pm 1\sigma$  – 68,3 % особей,

б)  $\pm 2\sigma$  – 95,5 % особей,

в)  $\pm 3\sigma$  – 99,7 % особей.

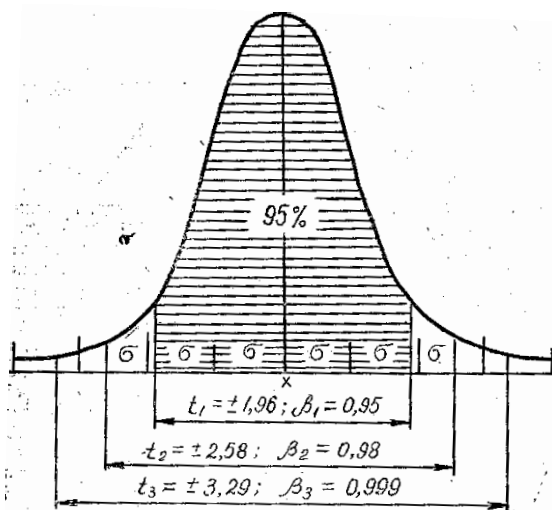


Рис. 3.1. Кривая нормального распределения:  $t$  – число сигм, на которое должны отстоять от средней в обе стороны границы:  $\bar{X} \pm t\sigma$ , чтобы вероятность появления дат, не выходящих за эти границы, равнялась заданным порогам – 0,95; 0,99; 0,999

### *Закономерности нормального распределения*

1. Значения нормированного отклонения колеблются приблизительно в пределах  $\pm 3 \sigma$ .
2. Чем дальше от среднего арифметического находится дата, тем больше соответствующее для нее значение нормированного отклонения.
3. Вероятность любого отклонения от  $\bar{X}$ , есть функция нормированного отклонения.

На этой основе рассчитана таблица нормального интеграла вероятностей для разных значений  $t$  (приложение, табл. 1), график порогов вероятностей безошибочного прогноза.

### *Закон распределения случайных величин при нормальном распределении*

Варьирующие признаки биологических объектов, принимающие в одних и тех же условиях испытания различные числовые значения, рассматриваются в математике как случайные переменные величины. Случайная переменная величина  $x$  или  $y$  в повторных испытаниях может принимать различные значения, но в каждом из них она имеет единственное значение:  $x_i, y_i$ , образуя дискретные (интервальные) или непрерывные (безинтервальные) вариационные ряды.

Функция  $f(x)$ , связывающая значение  $X_i$  с их вероятностями  $p_i$  называется законом распределения случайной переменной величины. Закон распределения любой случайной переменной величины можно выразить: 1) в таблице, 2) в виде кривой распределения вероятностей, 3) формулой. Непрерывная случайная величина может быть выражена лишь в значениях,

которые она может принять в определенном интервале с той или иной вероятностью. Выдающиеся математики К. Гаусс (1821), П. Лаплас (1795), А. Муавр (1821) и И.Г. Ламберт (1765) установили, что вероятность  $p$  любого значения  $x_i$  непрерывно распределяющейся случайной величины  $x$  находится в интервале от  $x$  до  $x + dx$  и выражается уравнением:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2} \frac{(xi-\mu)^2}{\sigma^2}} \cdot dx,$$

где  $dx$  – малая величина, определяющая ширину интервала;

$\sigma$  – стандартное отклонение, характеризующее степень рассеяния значения  $x_i$  случайной величины  $x$  вокруг средней  $\mu$ , (называемой математическим ожиданием).

В показатель степени входит **нормированное отклонение**:  $t = \frac{(xi-\mu)}{\sigma}$ .

Закон нормального распределения выражает функциональную зависимость между вероятностью  $p(x)$  и нормированным отклонением  $t$ . Этот закон утверждает: вероятность отклонения любой варианты (даты)  $x_i$  от центра распределения  $\mu$  определяется функцией нормированного отклонения ( $t$ ). Графически  $f(t)$  выражается нормальной кривой распределения вероятностей, форма которой определяется только двумя параметрами:  $\mu$  и  $\sigma$ . При изменении величины  $\mu$  форма нормальной кривой не изменится, лишь график ее смещается влево или вправо.

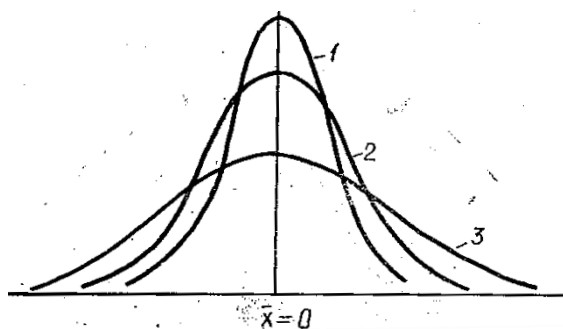


Рис. 3.2. Нормальные кривые (1, 2, 3) при различных значениях параметра  $\sigma$  ( $\sigma_1 < \sigma_2 < \sigma_3$ )

Уменьшение или увеличение величины  $\sigma$  приводит к изменению ширины кривой (рис. 3.2). Уменьшение величины сигмы ( $\sigma$ ) делает кривую более узкой за счет уменьшения варьирования признака около средней. При всех этих вариантах нормальная кривая распределения остается строго симметричной относительно центра распределения.

На рисунке 3.3 показана кривая распределения при  $\mu = 0$  и  $\sigma = 1$ , которая называется стандартизованной кривой. Любую нормальную кривую можно привести к стандартной путем вычитания  $\mu$  из  $x_i$  и делением на  $\sigma$ . Стандартная кривая имеет площадь, равную единице. Максимальная ордина-

та соответствует началу прямоугольных координат, перенесенному в центр распределения, где  $(x_i - \bar{\mu}) = 0$ . Влево и вправо от центра случайная величина  $x_i$  может принимать любые значения и величина  $(x_i - \bar{\mu})$  определяется функцией нормированного отклонения  $f(t)$ . Вероятности отклонений, соответствующие разным значениям  $t$ , приведены в таблице 1 (приложение).

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

первая функция нормированного отклонения.

Для того чтобы ордината выражала не вероятности, а абсолютные значения случайной величины, т.е. выравнивающие частоты вариант эмпирического распределения, в формулу частот вносятся дополнительные множители: число наблюдений  $n$ , умноженное на величину классового интервала  $k$ , в знаменатель – величину среднего квадратического отклонения эмпирического распределения. Теоретические частоты вариационного ряда  $f^1$  рассчитываются по формуле:  $f^1 = \frac{nk}{\sigma} f(t)$ . Значения функций нормированного отклонения приводятся в таблице 2 приложения. Таблица этих величин показывает ординаты нормальной кривой (первая функция нормированного отклонения).

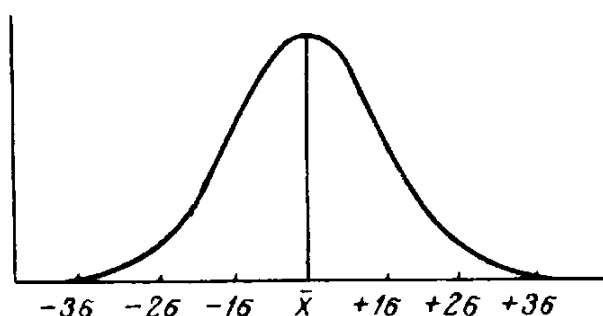


Рис. 3.3. Нормальная вариационная кривая.

Отклонения вариант вправо и влево от  $X$  охватывают несколько больше  $6\sigma$

Эмпирические распределения никогда не соответствуют в точности нормальному распределению. Математическими приемами можно выровнять эмпирические распределения и получить плавную вариационную кривую, в точности соответствующую закону нормального распределения. Выравнивание эмпирических кривых распределения вскрывает закономерности нормального распределения, скрытые под случайной формой проявления.

### *Возможности использования закономерностей нормального распределения*

Закономерности нормального распределения дают возможность по двум параметрам и  $\sigma$  построить весь вариационный ряд. Так, если известно, что  $x = 40$  см,  $\sigma = \pm 3$  см,  $n = 500$ , то размах вариации всего ряда должен быть от 31 до 49 см. 68 % особей, т.е. 340 из 500, будут иметь значение признака от 37 до 43 см.

Если, наоборот, знаем только *максимум и минимум* вариационного ряда, то можно приближенно, без вычислений, определить и среднее квадратическое отклонение, разделив вариационный размах на 6. Однако надо иметь в виду то обстоятельство, что изучаемая совокупность, являющаяся только выборкой из генеральной совокупности, обычно имеет ограниченный объем. Чем меньше объем ( $n$ ) изучаемой совокупности, тем менее точно фактический вариационный ряд с его минимумом и максимумом отображает теоретический вариационный ряд, который можно было бы построить на основе изучения генеральной совокупности и который должен охватывать  $6\sigma$ . Поэтому возможно определение  $\sigma$  по размаху варьирования:

- при  $n = 30$  величину размаха следует делить не на 6, а только на 4;

- при  $n = 50$  – на 4,5; при  $n = 100$  – на 5.

Однако определять среднее квадратическое отклонение следует более точными методами, пользуясь соответствующими формулами.

#### *А. Доверительные вероятности и уровни значимости*

**Понятие доверительной вероятности**, широко используемое в статистике, было введено английским биологом и статистиком **Р. Фишером**. Существенными являются три вероятности, которые постоянно используются исследователями в работах с использованием математических методов. Их обычно выражают величинами 0,95 и 0,99, 0,999. С вероятностью 0,95 любая случайно взятая особь будет отклоняться от  $x$  не более чем на  $1,96\sigma$ , или, иначе, с вероятностью 0,05 она будет за пределами  $1,96\sigma$ . С вероятностью же равной 0,99 она будет отклоняться от  $x$  не более чем на  $2,58\sigma$ . Вероятность выхода за пределы  $\pm 2,58\sigma$  равна 0,01. Если же взять в качестве границы  $3\sigma$ , то вероятность отклонения от  $x$  больше чем на  $3\sigma$  ( $t > \pm 3$ ) очень мала  $\sigma$  – всего 0,0027. Это очень важное правило часто называют *правилом трех сигм*. Три сигмы ограничивают пределы случайного рассеяния внутри вариационного ряда. Что находится в пределах  $3\sigma$ , относится к данному ряду; то, что за пределами  $3\sigma$ , вероятнее всего, к этому ряду уже не относится. Но для достижения вероятности 0,99 достаточно взять границы только

$\pm 2,58\sigma$ . Вероятность, выражающаяся величиной 0,99, достаточно велика, и в тех случаях, когда достигнута такая вероятность, можно с очень большой степенью уверенности делать вывод по поводу отнесения особи к той или иной группе относительно результатов опыта. Но нередко можно остановиться и на более низком уровне вероятности, например 0,95. В этом случае, отклонения от ожидаемого результата будут в 5 % случаев (вероятность 0,05). Вероятности 0,95 и 0,99, или 95 % и 99 %, получили название *доверительных вероятностей*, т.е. таких, значениям которых можно достаточно доверять или которыми можно уверенно пользоваться. Вероятности, принятые как доверительные, в свою очередь определяют **доверительные границы** и **доверительный интервал** между ними. На них можно основывать оценку той или иной величины и те границы, в которых она может находиться при разных вероятностях. Для различных вероятностей **доверительные интервалы** будут следующими:

Вероятность	Интервал
0,95	$-1,96 \sigma \dots +1,96 \sigma$
0,99	$-2,58 \sigma \dots +2,58 \sigma$
0,999	$-3,03 \sigma \dots + 3,03 \sigma$

Вероятности можно обозначать как в долях единицы, так и в процентах, поэтому в последующем будем употреблять параллельно оба обозначения.

#### *Уровни значимости*

Определенным значениям вероятностей соответствуют так называемые *уровни значимости*. Вероятности 0,95 (95 %) соответствует уровень значимости 0,05 (5 %). По отношению к закономерностям нормального распределения это означает, что выход за пределы принятых границ возможен в порядке случайности с вероятностью 0,05, т.е. в 5 % случаев есть вероятность (риск) ошибиться в своих выводах. При вероятности 0,99 уровень значимости 0,01 (1 %). Случайное отклонение возможно лишь с вероятностью 0,01.

**Таким образом, уровень значимости обозначает вероятность получения случайного отклонения** от установленных с определенной вероятностью результатов. С помощью уровня значимости можно установить, в каком проценте случаев (или с какой вероятностью) все же возможна ошибка в результатах, в тех выводах, которые делаются на основе опыта, в оценке достоверности показателей или различий между какими-то величинами, полученными в опытах или наблюдениях. Необходимо, чтобы выводы, вытекающие из научного исследования, имели достаточно высокую достоверность (употребляют также термин «значимость»). Например, 5 % уровень значимости ( $p = 0,05$ ) указывает, что ошибка репрезентативности возможна в 5 % случаев. В некоторых случаях можно удовлетвориться и таким результатом.



Но если нужна большая доказательность результатов, то уровень значимости должен быть повышен до 1 % ( $p = 0,01$ ). Чем эта цифра меньше, тем уровень значимости (и достоверность) результатов выше. При уровне значимости 0,01 в 1 % случаев вывод не обоснован из 100. Такую значимость считают уже высокой и широко ею пользуются. Но бывают случаи, когда уровень значимости может быть еще выше – 0,001. Тогда вывод не обоснован только в одном случае из 1000. В каждом конкретном случае, исходя из важности исследования, устанавливается граница той вероятности, с которой считаются, и той, с которой не считаются. Уровень значимости – вероятность ошибки, которой решено пренебрегать в данном исследовании или явлении.

*В. Эмпирические ряды распределения,  
их отклонение от теоретических. Артефакты*

Конечным результатом изучения той или иной совокупности по определенным признакам является составление эмпирического вариационного ряда, его графическое изображение в виде полигона или гистограммы и вычисление основных статистических показателей ( $x$  и  $\sigma$ ,  $\sigma^2$ ). Большое количество признаков биологических объектов варьирует в соответствии с закономерностями нормального распределения. *Однако возможны случаи, когда фактические распределения в той или иной степени отклоняются от теоретических распределений.* Это может проявиться как в форме кривой распределения, так и в особенностях полученных статистических показателей. Уже упоминалось о двух- или многовершинности кривых распределения, могущих быть результатом объединения в одну совокупность двух или нескольких групп, в действительности отличающихся друг от друга. Очевидно, что в таких структурно неоднородных рядах и нельзя ожидать проявления закономерностей нормального распределения. Перед исследователем будет стоять задача расчленения исходного материала на более однородные группы, чтобы каждую из них обработать самостоятельно, выразить ее в виде кривой распределения и вычислить статистические показатели.

*Одновершинная кривая распределения* может быть не вполне симметричной. Полезно вычислить показатель асимметрии, который дает объективную оценку степени асимметрии, едва уловимую при рассмотрении графиков. Неполная симметрия (*скошенность*) иногда есть результат неполноты материала, недостаточного количества изученных вариантов. Если асимметрия значительна, следует проверить, не является ли полученное распределение пуассоновским. Критерием его, служит примерное равенство  $\bar{X}$ ,  $Me$ ,  $Mo$  и  $\sigma^2$ . Однако асимметрия ряда может зависеть и от природы изучаемого признака, по каким-либо причинам легче варьирующего в одном направлении и труднее – в другом. *Соответствие фактического распределения нормаль-*

ному дает возможность судить и о том, в какой степени изучаемый эмпирический материал действительно однороден, нет ли в нем *отдельных вариантов*, которые по тем или иным причинам резко выделяются из изучаемой совокупности. При проведении опытов иногда получают так называемые «выскакивающие» результаты, которые явно не укладываются в общую картину вариации полученных данных. При изучении материала, взятого из природы или опыта, также бывают случаи, когда одна или несколько вариантов отклоняются от средней арифметической значительно больше, чем на  $3\sigma$ . Так, например, если при изучении веса (масса тела) при рождении большой группы телят получены  $\bar{X} = 32$  кг и  $\sigma = \pm 3$  кг, то теленок с весом 17 кг окажется далеко за пределами изменчивости этой труппы телят. Его отклонение от  $\bar{x}$  будет равно  $5\sigma$ .

В практике экспериментальной работы нередко такие «выскакивающие» значения – *артефакты* – исключают из анализируемого материала, считая их результатом незамеченной при проведении опыта или наблюдения неточности, ошибки или каких-либо частных «патологических» обстоятельств, нарушающих общую картину. Однако это можно делать лишь в тех случаях, когда весь остальной материал действительно укладывается в очень четкий и симметричный вариационный ряд. При асимметричном же распределении некоторые варианты могут отклоняться от  $\bar{x}$  значительно больше чем на  $3\sigma$  по объективной причине *самой закономерности вариации*. Поэтому исключение подобных вариантов из рассматриваемого материала будет неправильным. Хотя теоретический ряд распределения должен охватывать (если он нормальный) примерно 6 значений среднего квадратического отклонения, в конкретном эмпирическом вариационном ряду это будет наблюдаться довольно редко. При малых значениях  $n$  соответствующий нормальному эмпирический ряд может охватывать не 6, а 5 или  $4\sigma$ . Возможно также большее сгущение вариантов вблизи средней арифметической при недостатке их в боковых частях распределения («крутизна») и, наоборот, ненормально малая частота вариантов в классах, близких к средней арифметической («плосковершинность»). При «крутизне» значения  $\sigma$  малы по сравнению с теми же параметрами нормального ряда, при плосковершинности кривой, наоборот, велики. Это легко проверить, если, сохраняя  $\bar{X}$ , уменьшать или увеличивать значение  $\sigma$ . Вариационная кривая при этом будет делаться или более острой, или более плоской. Таким образом, исследователь должен очень внимательно анализировать эмпирические ряды распределения, оценивая их математически, не стремясь подогнать их к тому или иному виду теоретических кривых.

*Асимметрия и эксцесс* – явления, которые встречаются часто при исследовании эмпирических распределений. Асимметрия проявляется в виде скошенной вариационной кривой, вершина которой находится правее или ле-

вее центра распределения. На рисунке 3.4 (слева) имеет место отрицательная асимметрия, на рисунке 3.4 (справа) – положительная (в соответствии со знаком числовой характеристики). В первом случае пологая сторона кривой распределения находится левее и называется левосторонней, во втором – правосторонней асимметрией. В этих случаях распределения значительно отличаются от нормального. Такие распределения имеют многие признаки растений, животных и микроорганизмов.

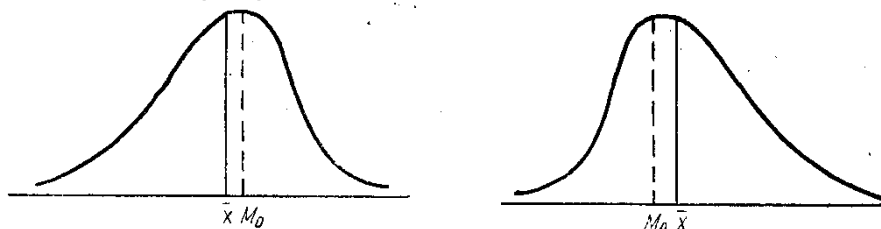


Рис. 3.4. Отрицательная и положительная асимметрия

В тех случаях, когда действуют факторы, благоприятствующие или препятствующие развитию признака, образуются асимметричные распределения. При асимметрии эмпирические распределения имеют увеличенные частоты (в сравнении с симметричным расположением) в левой или правой части. Кроме асимметричных, встречаются островершинные и плосковершинные распределения.

#### *Правило проверки артефактов*

Артефакты (выпады) – даты, выходящие за пределы ожидаемого.

Согласно теории вероятностей, действительные отклонения можно найти по следующему математическому правилу:

Если  $T_x = \frac{V-M}{\delta} \geq T_{st}$  (V) Отклоняющаяся дата исключается.

Таблица 3.1

**Нормированные отклонения выдающихся дат ( $T_{xst}$ )**

n	$T_{st}$	n	$T_{st}$	n	$T_{st}$	n	$T_{st}$
2	2,0	16–20	2,4	47–66	2,8	125–174	3,2
4	2,1	21–28	2,5	67–84	2,9	175–349	3,3
9	2,2	29–34	2,6	85–104	3,0	350–599	3,4
15	2,3	35–46	2,7	105–124	3,1	600–1500	3,5

**Пример 1.** V: 1, 2, 3, 10;  $n = 4$ ;  $M = 4,0$ ;  $\sigma = 4,1$ .

$$T_x = \frac{10,0 - 4,0}{4,1} = 1,5 < 2,1 < T_{st}.$$

**Вывод:** варианту 10 следует оставить.

**Пример 2.** V: 1, 2, 2, 3, 3, 4, 4, 5, 21 – артефакт ?

$$n = 9, \quad M = 5, \quad \sigma = 6,1, \quad T_{st} = 2,6 > T_x = 2,2.$$

**Вывод:** варианту 21 следует исключить и вычислить заново  $M$  ( $\bar{X}$ ),  $\sigma$ .

### 3.4. Самостоятельная работа по разделу

Цель и задачи:

1. На основе изучения закономерностей нормального распределения освоить методику определения артефактов.
2. Освоить методику выравнивания эмпирических кривых распределения.
3. Исследовать соответствие эмпирических кривых закону нормального распределения.

#### *Содержание работы*

1. Исследовать экспериментальную совокупность на наличие артефактов. В случае обнаружения артефактов исключить отклоняющиеся даты из исследуемых совокупностей и пересчитать заново статистические показатели.
2. Освоить алгоритм расчета теоретических частот в соответствии с уравнением нормального распределения (табл. 3.2, алгоритм).
3. На основе эмпирических вероятностей, основанных на учете конкретных частот, произвести расчет теоретических частот для исследуемой экспериментальной совокупности.
4. Построить теоретическую кривую распределения по исследуемым признакам.
5. Пользуясь алгоритмом расчета критерия  $\lambda$  (лямбда) А.Н. Колмогорова и Н.А. Смирнова, оценить вероятность соответствия теоретических и эмпирических кривых распределения вероятностей (табл. 3.3, алгоритм). При отсутствии соответствия теоретических и эмпирических кривых распределения, оценить достоверность различий по критерию  $\lambda$  (лямбда).

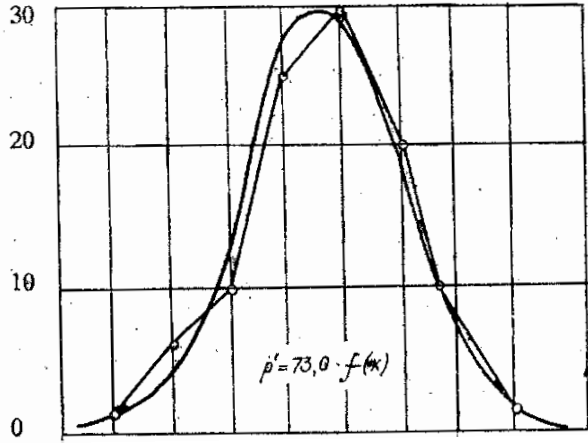
#### *Методика выравнивания эмпирических кривых распределения по нормальному закону*

Методика выравнивания эмпирических кривых распределения представлена в алгоритме (табл. 3.2) и состоит из следующих действий:

1. Весь объем выборочной совокупности разбивается на классы и составляется вариационный ряд.
2. Рассчитываются  $M(\bar{X})$ , определяется величина  $\frac{nk}{\sigma}$ , единая для данного вариационного ряда.
3. Рассчитываются нормированные отклонения  $x = \frac{W - M}{\sigma}$  по каждому классу.
4. Используя таблицу первой функции нормированного отклонения, (приложение, табл. 2) находят  $f(x)$  для каждого класса.

Таблица 3.2 (алгоритм)

**Выравнивание эмпирических вариационных кривых  
по нормальному закону  $p' = \frac{n \cdot k}{\sigma} \cdot f(x)$**

$p'$ – теоретическая частота; $n$ – объём ряда; $k$ – классовый промежуток; $\sigma$ – сигма;			$f(x)$ – первая функция нормированного отклонения (см. приложение) $x = \frac{W-M}{\sigma}$ – нормированное отклонение середин классов							
Вариации $W$	Эмпирические частоты $p$	$W - M$	$x = \frac{W - M}{\sigma}$	$f(x)$	Теоретические частоты					
					$\frac{n \cdot k}{\sigma} \cdot f(x)$	$p'$				
450	1	+33	+2,43	0,021	1,5	1				
440	7	+23	+1,69	0,096	7,0	7				
430	20	+13	+0,96	0,252	18,4	18				
420	30	+3	+0,22	0,389	28,5	29				
410	25	-7	-0,51	0,350	25,6	26				
400	10	-17	-1,25	0,183	13,5	14				
390	6	-27	-1,99	0,055	4,0	4				
380	1	-37	-2,72	0,010	0,7	1				
	100	–	–	–	99,2	100				
Пример: $n = 100$ ; $k = 10$ ; $M = 417,0$ ; $\sigma = 13,7$ ; $\frac{n \cdot k}{\sigma} = \frac{100 \cdot 10}{13,7} = 73,0$										
Вариации			380	390	400	410	420	430	440	450
Эмпирические частоты, $p$			1	6	10	25	30	20	7	1
Теоретические частоты, $p'$			1	4	14	26	29	18	7	1

5. Используя полученные данные и уравнение нормального распределения, производится расчет теоретических частот для каждого класса:

$$f^1 = \frac{nk}{\sigma} \cdot f(t).$$

6. На основе полученных ординат строится теоретическая кривая нормального распределения.

7. Сделать выводы о соответствии эмпирических кривых распределения нормальному закону.

8. Соответствие эмпирических и теоретических кривых распределения оценивается, пользуясь критериями  $\chi^2$  и  $\lambda$  (лямбда).

#### *Критерий $\lambda$ (лямбда) А.Н. Колмогорова и Н.А. Смирнова*

Критерий  $\lambda$  применяется для определения достоверности расхождений между эмпирическими и теоретическими распределениями, а также между любыми двумя распределениями частот одного и того же признака, даже в случае несовпадения числа классов и числа дат.

Для расчета критерия  $\lambda$  (алгоритм табл. 3.3) необходимо составить ряды накопленных (*кумулярованных*) частот для каждого из сравниваемых распределений:  $\sum f$  и  $\sum f^1$  и, определив их разницу по каждому классу, установить наибольшую, *максимальную* разницу:

$$[d] = [\sum f^1 - \sum f] \max.$$

Эмпирический критерий вычисляется по формуле:  $\lambda = \frac{[d]\max}{\sqrt{n}}$ , затем эмпирическое значение сравнивается с теоретическим (табл. 3.3, алгоритм) для установления порога вероятности различия распределений. Методика оценки расхождения между двумя любыми распределениями частот при разном числе дат показана во второй части алгоритма.

*Условием применения критерия лямбда* является достаточно большая численность сравниваемых распределений (несколько десятков дат). При определении достоверности различий между эмпирическим и теоретическим распределениями устанавливается обратный порядок планирования порога вероятности безошибочного прогноза. В этих исследованиях, чем выше ответственность, тем при меньшем расхождении распределений различие уже считается достоверным. Чем менее ответственно исследование, тем при большем расхождении распределений различие между ними все еще может считаться не достоверным. При оценке различий между эмпирическими и теоретическими распределениями в большинстве биологических работ устанавливают второй порог вероятности безошибочного прогноза  $\lambda = 1,63$ ,  $B_2 = 0,99$ . Для применения критерия лямбда не требуется определение числа степеней свободы.

Таблица 3.3 (алгоритм)

**Оценка расхождения любых распределений: критерий  $\lambda$  (лямбда)**

I. Оценка различий между теоретическим и эмпирическим распределениями:								
$\lambda = \frac{ d }{\sqrt{n}} = \frac{ \sum f_i - \sum f'_i _{max}}{\sqrt{n}} \geq \begin{cases} 1,95 & B_3 = 0,999 \text{ при малой (!)} \\ 1,63 & B_2 = 0,99 \text{ при обычной} \\ 1,36 & B_1 = 0,95 \text{ при большой (!)} \end{cases} \left. \begin{array}{l} \text{ответственности} \\ \text{результатов} \\ \text{исследований} \end{array} \right\}$								
W	f	f'	Накопляемые частоты		$n = 100$ – объём каждой группы; $\lambda = \frac{3}{\sqrt{100}} = 0,3 < 1,36$ ; Различие недостоверно, нет достаточных оснований считать, что выборки взяты из генеральных совокупностей, отличающихся своим распределением			
			$\sum f$	$\sum f'$				
450	1	1	100	100				
440	7	7	99	99				
430	20	18	92	92				
420	30	29	72	74				
410	25	26	42	45				
400	10	14	17	19				
390	6	4	7	5				
380	1	1	1	1				
n	100	100	–	–				
II. Оценка различий между двумя любыми распределениями:								
$\lambda = \frac{ d }{n_1} = \left  \frac{\sum f_1 - \sum f_2}{n_2} \right _{max} \geq \begin{cases} 1,95 & B_3 = 0,999 \text{ при малой (!)} \\ 1,63 & B_2 = 0,99 \text{ при обычной} \\ 1,36 & B_1 = 0,95 \text{ при большой (!)} \end{cases} \left. \begin{array}{l} \text{ответственности} \\ \text{результатов} \\ \text{исследований} \end{array} \right\}$								
Различия могут считаться случайными, если эмпирический критерий не достигает требуемого порога вероятности								
W	f <sub>1</sub>	f <sub>2</sub>	$\sum f_1$	$\sum f_2$	$\frac{\sum f_1}{n_1}$	$\frac{\sum f_2}{n_2}$	d	$\lambda = 0,30 \cdot \sqrt{\frac{100 \cdot 200}{100 + 200}} =$ $= \underline{\underline{2,45}} > 1,95$ ; Различия не могут считаться случайными, они в высшей степени достоверны. Выборки взяты из генеральных совокупностей, явно различающихся по своим распределениям
450	1	2	100	200	1,00	1,00	0	
440	7	4	99	198	0,99	0,99	0,00	
430	20	8	92	194	0,92	0,97	0,05	
420	30	42	72	186	0,72	0,93	0,21	
410	25	83	42	144	0,42	0,72	0,30	
400	10	37	17	61	0,17	0,31	0,14	
390	6	20	7	24	0,07	0,12	0,05	
380	1	4	1	4	0,01	0,04	0,03	
n	100	200	–	–	–	–	–	

**3.5. Распределение групп**

При исследовании качественных признаков составляются распределения не дат, а групп по числу объектов, обладающих (или не обладающих) изучаемым признаком, распределения разночленных групп по значению ( $n +$ ) и ( $n -$ ). Наибольшее значение из такого рода распределений в биоло-

гии имеют биномиальное распределение и распределение Пуассона, являющееся частным случаем биномиального распределения.

#### А. Биномиальное распределение

Закон биномиального распределения – закон распределения дискретной случайной величины  $x$ . Если вероятности появления отдельных значений  $p$  и  $q$  выражаются величинами, соответствующими коэффициентами бинома Ньютона, распределение называют биномиальным:

$$(p + q)^n = p^n + n \cdot p^{n-1} \cdot q + \frac{n(n-1)}{1 \cdot 2} p^{n-2} q^2 + q^n, \quad (1)$$

где  $p$  – вероятность проявления первого события (в генеральной совокупности);  $q$  – вероятность проявления второго события, т.е. отсутствия первого;  $n$  – одинаковая численность каждой изучаемой группы.

Таблица 3.4

**Распределение вероятностей появления разного количества курочек среди десяти цыплят**

Количество курочек, ♀	0	1	2	3	4	5	6	7	8	9	10
Количество петушков, ♂	1	10	45	120	210	252	210	120	45	10	1
Вероятность события, $p$	0,001	0,010	0,044	0,117	0,205	0,246	0,205	0,117	0,044	0,010	0,001

Эти 1024 случая (исхода) распределяются следующим образом:

$$p_{10}(m) = (0,5 + 0,5)^{10} = \frac{1}{1024} + \frac{10}{1024} + \frac{45}{1024} + \frac{120}{1024} + \frac{210}{1024} + \frac{252}{1024} + \frac{210}{1024} + \frac{120}{1024} + \frac{45}{1024} + \frac{10}{1024} + \frac{1}{1024} = 1.$$

При биномиальном распределении возможны различные значения  $p$  и  $q$ :  $p = 0,7$ ;  $q = 0,3$ , или  $p = 0,9$  и  $q = 0,1$ , или  $p = 0,5$  и  $q = 0,5$  и др. Форма полигона при этом меняется. По мере увеличения различий между  $p$  и  $q$  полигон распределения становится все более скошенным, асимметричным. Однако по мере увеличения  $n$  даже при значительном различии  $p$  и  $q$ , степень симметрии полигона вновь увеличивается.

**Резюме:** в отношении некоторого случайного события ( $A +$ ) производят  $n$  независимых испытаний (при этом вероятность  $p$  появления этого события постоянна). Учитываются только два исхода; появление события ( $A +$ ) или ему противоположного события ( $A -$ ), тоже имеющего постоянную вероят-



ность  $q$ , причем  $p + q = 1$ . Если событие ( $A+$ ) в  $n$  независимых испытаниях появится  $m$  раз, то событие ( $A-$ ) будет встречаться  $(n - m)$  раз. Вероятность любого исхода ( $P_m(n)$ ), независимо от того, в каком порядке эти события чередуются, выразится произведением  $p^m \cdot q^{n-m}$ , умноженным на биномиальный коэффициент:

$$C_n^m: P_n(m) = C_n^m \cdot p^m \cdot q^{n-m} = \frac{n!}{m! (n - m)!} p^m \cdot q^{n-m}.$$

Эта формула Бернулли позволяет находить вероятность того, что из  $n$ , взятых наугад элементов окажется  $m$  ожидаемых. Если графически изобразить это распределение (рис. 3.5), то получится полигон биномиального распределения. В конкретном примере число сочетаний различного количества курочек и петушков равно 10, т.е. бином представлен  $(p + q)^{10}$ , его разложение дано во второй строчке таблицы, вероятности отдельных случаев – в третьей строчке. Сумма всех вероятностей равна 1 (единице). График соответствует биномиальному распределению, где ординаты соответствуют членам разложения бинома  $(1/2 + 1/2)^{10}$ .

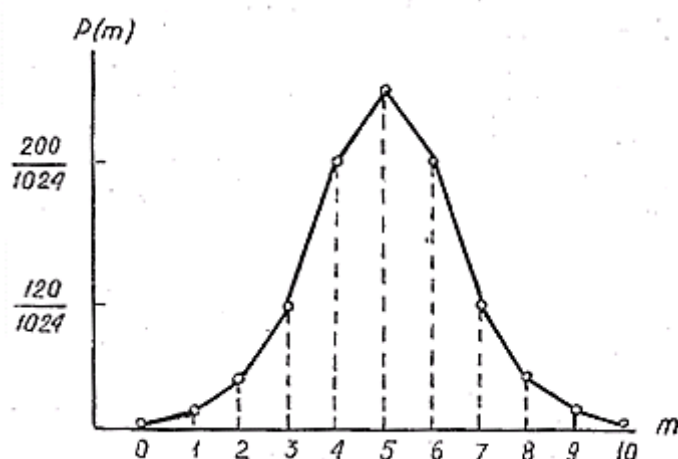


Рис. 3.5. Распределение вероятностей двучлена  $(1/2 + 1/2)^{10}$

Биномиальная кривая в этом случае строго симметрична относительно максимальной ординаты, являющейся центром биномиального распределения. При  $p = q$  биномиальная кривая приближается к своему пределу нормальной кривой. Если  $p \neq q$  – биномиальная кривая становится все более асимметричной и тем сильнее, чем больше разница между  $p$  и  $q$ . Например, при положительной асимметрии более пологой становится ее правая часть. Распределение частных равночисленных групп по значению  $(+n)$  и  $(-n)$  называется биномиальным. Такое название объясняется следующими причинами:

- признак может иметь только два варианта: он есть (+) или нет (-);

• закономерности таких распределений имеют количественное выражение, связанное с коэффициентами разложения бинома Ньютона, который, кроме уравнения, представленного выше (1), может быть представлен в следующем виде:

$$(p+q)^n = \frac{1}{1} p^n q^0 + \frac{n}{1} p^{n-1} q^1 + \frac{n(n-1)}{1 \cdot 2} p^{n-2} q^2 + \frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3} p^{n-3} q^3 + \dots + \frac{1}{1} p^0 q^n. \quad (2)$$

Вариациями этого распределения будут величины: « + » и « - », количество особей, имеющих изучаемый признак в отдельных разноточленных частных группах, а частотами – количество соответствующих групп. Каждый член бинома может быть представлен в виде произведения двух множителей, первый из которых зависит от величины  $n$ , второй от соотношения  $p$  и  $q$ .

n	Биномиальные коэффициенты															2n
0																1
1																2
2																4
3																8
4																16
5																32
6																64
7																128
8																256
9																512
10	1	10	45	120	210	252	210	120	45	10	1					1024

Рис. 3.6. Арифметический треугольник Паскаля

Первые множители каждого члена бинома – коэффициенты бинома, определяются в зависимости от величины  $n$  (число особей в каждой группе) по арифметическому треугольнику Паскаля, в котором каждое число равно сумме чисел, стоящих над ним.

*Биномиальные коэффициенты.* Подставляя в формулу бинома величины  $f(n)$  и  $f(p, q)$  и решая ее относительно величины  $p$ , можно получить следующие значения:

- $q^n$  – нулевой член бинома (содержит  $p$  в нулевой степени) дает ожидаемую долю таких разноточленных групп, в которых из  $n$  особей ни одна не имеет изучаемого признака;
- $n \cdot p^{n-1}$  – первый член бинома (с  $p^1$ ) дает долю групп, в которых только одна особь имеет ожидаемый признак;
- $f \cdot (n) \cdot k \cdot p^k \cdot q^{n-k}$ ,  $k$ -й член бинома, соответствует доле групп, в которых имеется  $k$  особей с изучаемым признаком;

•  $p^n$  – последний член бинома, дает долю групп, в которых все  $n$  особей имеют изучаемый признак. Рассмотренный пример показывает, что распределение вероятностей соответствует коэффициентам бинома Ньютона, отнесенным к одному и тому же знаменателю, равному  $2^n$  (в данном случае:  $2^{10} = 1024$ ).

Сумма биномиальных коэффициентов для бинома любой степени (рис. 3.6) равна  $2^n$ . Характер биномиального распределения не меняется от того, как выражены результаты (исходы испытания), в значениях вероятности или в абсолютных значениях частоты ожидаемого результата. В любом случае биномиальный закон выражает зависимость между частотой ожидаемого результата и числом независимых испытаний, проведенных в отношении случайного события: (А+). Частота ожидаемого события в  $n$  независимых испытаниях определяется его вероятностью  $p$ , которая остается постоянной в каждом отдельном испытании.

$$\sum f^1 = N(p + q)^n; \quad f^1 = \frac{NK}{\sum K},$$

где  $N$  – сумма всех частот эмпирического ряда ( $\sum f_i$ );  $K$  – биномиальные коэффициенты (по треугольнику Паскаля, рис. 3.6);  $p$  – вероятность ожидаемого события;  $q = 1 - p$ ,  $n$  – число членов ряда минус единица  $n = m - 1$ .

Таблица 3.5

**Эмпирическое распределение численности самок  
в 113 пометах лабораторных мышей:  $p = q = 0,5$**

Число самок в помете	0	1	2	3	4	5	6	7	8
Число пометов с указанным количеством самок	0	1	10	17	46	28	8	3	0

Число классов (без 0) – 7.

Сумма частот  $N = 113$ .

$n = m - 1$ :  $n = (7 - 1) = 6$ .

$K$  – биномиальные коэффициенты (по треугольнику Паскаля)

при  $n = 6$

$$K = 1 - 6 - 15 - 20 - 15 - 6 - 1.$$

Сумма биномиальных коэффициентов = 64:

$$\begin{aligned} \sum f^1 &= 113 \left(\frac{1}{2} + \frac{1}{2}\right)^6 = 113 \left(\frac{1}{64} + \frac{6}{64} + \frac{15}{64} + \frac{20}{64} + \frac{15}{64} + \frac{6}{64} + \frac{1}{64}\right) = \\ &= 1,77 + 10,59 + 26,48 + 35,32 + 26,48 + 10,59 + 1,77 = 113,00. \end{aligned}$$

Округляя числа, получаем теоретически ожидаемые частоты:

$$\begin{aligned} m &1 - 2 - 3 - 4 - 5 - 6 - 7, \\ f_i &2 - 11 - 26 - 35 - 26 - 11 - 2. \end{aligned}$$

В случаях с известной вероятностью ( $p = q = 1/2$ ) теоретически ожидаемые частоты рассчитываются также по формуле:  $f^1 = \frac{NK}{\sum K}$

Если значения  $p$  и  $q$  заранее неизвестны, их определяют по средней статистической величине (вероятности) полученных в опыте данных:

$$f^1 = N(p^m q^{n-m} K),$$

где  $p$  – статистическая вероятность первого события  $\left(\frac{m}{n}\right)$

$$\bar{m} = \frac{\sum mfi}{N} \cdot q = 1 - p.$$

Рассмотрим  $n$  независимых испытаний, в каждом из которых наступает событие  $A$  с вероятностью  $p$ . Обозначим через  $x$  случайную величину, равную числу появлений события  $A$  в  $n$  испытаниях. Событие  $A$  может вообще не наступить, наступить один раз, два раза, три раза, ...  $n$  раз. Таким образом, возможными значениями  $x$  будут числа  $1, 2, 3, \dots, n-1, n$ . По формуле Бернулли можно найти вероятности этих значений:

$x$	0	1	...	$m$	...	$n$
$p$	$q^n$	$Cn^1 p n^{-1}$	...	$Cn^m q^{n-m}$	...	$q^n$

Биномиальному распределению вероятностей соответствуют биологические совокупности, возникающие в результате *генетических расщеплений*. Расщепления при моногибридном, дигибридном, полигибридном скрещиваниях соответствуют формулам:  $(3 + 1)^1$ ;  $(3 + 1)^2$ ;  $(3 + 1)^3$ ;  $(3 + 1)^n$ . Примеры анализа генетических расщеплений приводятся в конце главы 3.

### В. Распределение Пуассона

Такой тип распределений также относится к типу дискретной (прерывистой) изменчивости. Оно имеет самостоятельное значение, хотя его рассматривают и как частный случай биномиального распределения.

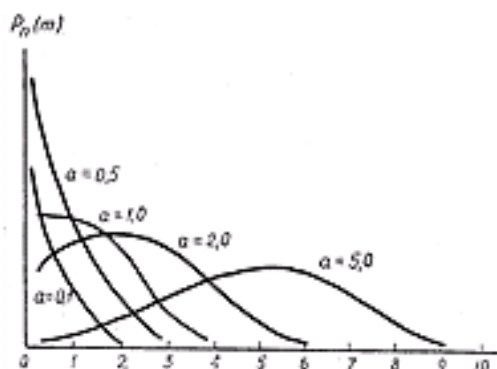


Рис. 3.7. График функции  $P_n(m) = \frac{a^m}{m!} e^{-a}$  различных значений  $a$

Когда вероятность ( $p$ ) события очень мала и выражается сотыми (тысячными) долями единицы, распределение частот таких редких событий

в  $n$  независимых испытаний становится крайне асимметричным. Закономерности распределения таких событий вскрыты Пуассоном и выражены уравнением:

$$P_n(m) = \frac{a^m}{m!} e^{-a} = \frac{a^m}{m! e^a},$$

$a \approx n \cdot p$  – наивероятнейшая частота ожидаемого события;  $m$  – частота ожидаемого события в  $n$  независимых испытаниях;  $e = 2,7183$  – основание натуральных логарифмов;  $m!$  – произведение натуральных чисел от 1 до  $m$  ( $1 \cdot 2 \cdot 3 \cdot \dots \cdot m$ ).

Уравнение Пуассона дает возможность определить вероятность для любых значений  $a$ : от 0 до  $n$ . В биологии закономерности распределения Пуассона применимы к таким явлениям, как частота рождения троен, четверок близнецов у человека, частота островков Лангерганса в тканях поджелудочной железы, частота спонтанных мутаций у кишечной палочки, число поврежденных хромосом в клетках меристемы корешков растений при облучении сухих семян.

Рождение двух близнецов у человека часто встречаемое явление. Это происходит в среднем один раз на 80 случаев рождения одного ребенка. Тройня рождается (по статистическим данным) один раз на 80 двоен. В дальнейшем эта пропорция сохраняется: на 80 рождений троих близнецов приходится одна четверка, на 80 рождений четверен – один случай рождения пятерых детей. Таким образом, пять близнецов появляется один раз на 40960000 обычных рождений, а шесть близнецов – на 3276800000 обычных рождений.

2 – 1/80 (от обычных рождений);

3 – 1/80 (от двоен);

4 – 1/80 (от троен);

5 близнецов – 1/80 (от четверок)

Вероятность рождения пяти близнецов:  $\frac{1}{80} \times \frac{1}{80} \times \frac{1}{80} \times \frac{1}{80} = \frac{1}{40960000}$ .

Кривая распределения таких величин является асимметричной. При малых значениях  $p$  и достаточно большом значении  $n$  распределение Пуассона приближается к биномиальному. Оно, как и биномиальное распределение, приближается к нормальной кривой (рис. 3.7) при возрастании числа случаев. Для того чтобы с помощью уравнения Пуассона произвести расчет теоретических ординат кривой распределения Пуассона, формуле придают следующий вид:

$$f^1 = n \frac{n^{-n}}{m!} \cdot e^{-x},$$

где  $f^1$  – ожидаемые частоты редкого события или ожидаемое число случаев редкого события в каждом отдельном классе испытаний: 0, 1, 2, 3, 4, 5, 6, ...;

$n$  – число испытаний;  $x$  – среднее число фактически наблюдаемых случаев (вместо  $a$ ).

В современной радиобиологии многие расчеты проводятся на основе уравнения Пуассона, так как в этих случаях исследователь имеет дело с редкими событиями. Так, при облучении штамма бактериальных клеток гамма-лучами число подвергнутых испытанию объектов очень велико ( $n = \kappa$ ). Возникновение биохимических мутантов, выявляемых методом селективных сред, является редкими событиями ( $m$ ), вероятность которых ( $p = \frac{m}{n}$ ) очень мала. Распределение Пуассона характеризуется только одним параметром  $\bar{X}$  ( $M$ ) (средней арифметической), так как  $\sigma^2$  в этом случае обычно равна  $\bar{X}$  (или близка к ней по значению). По равенству  $\bar{X} = \sigma^2$  можно определить, является ли распределение пуассоновским. Среднее арифметическое для распределения Пуассона обозначается  $m$  или  $\lambda$ :

$$\bar{X} = M = \lambda = n \cdot p = \sigma^2.$$

### **С. Параметры дискретных распределений**

- *Биномиальное распределение.*

Два параметра характеризуют биномиальное распределение:

1) среднее (наивероятнейшее) число ( $\mu$ ) ожидаемого результата:

$$\mu = n \cdot p,$$

где  $n$  – число испытаний;  $p$  – вероятность;

$$2) \sigma_m^2 = n \cdot p \cdot q, \quad \sigma = \sqrt{npq}.$$

• *Распределение Пуассона* характеризуется одним параметром: средней величиной:  $\bar{X} = M = n \cdot p$ , так как для этого распределения

$$\sigma_m^2 = m; \quad \bar{X} = \bar{m} = \frac{\sum fx}{n};$$

где

$$\sigma_m^2 = \frac{\sum fx^2 - \frac{(\sum fx)^2}{n}}{n - 1}.$$

Расчет теоретических частот можно произвести, применяя данные таблицы 3 приложения, в которой содержатся значения вероятности  $P(m)$  для каждого класса испытаний  $m$  и средней величины ( $X$ ).

Чтобы получить теоретические частоты  $f^1$ , нужно значения вероятностей (табл. 3.6) для них умножить на число наблюдений ( $n$ ). Полученные результаты округляются до целых чисел, сохраняя общее число испытаний. Биномиальное распределение, в котором величина  $p$  пренебрежимо мала, а величина  $N$  становится очень большой (произведение  $N \cdot p$  остается средней

величиной), приближается к *распределению Пуассона*. В этом случае вероятность числа событий составит:

$$Pr(i) = \frac{e^{-\mu} \mu^i}{i!},$$

где среднее число событий,  $N \cdot p$ , равно  $\mu$ . Если вероятность принадлежности особи к данной группе  $p = 0,01$ , а  $N = 100$  ( $\mu = 1$ ), то вероятность группы с 0 особями составит:

$$Pr(0) = \frac{e^{-1} \mu^0}{0!} = 0,368.$$

Таблица 3.6

**Расчет теоретических частот ( $f^1$ ) по уравнению Пуассона**

Число		$P(m)$ вероятность	Теоретические частоты	
клеток-мутантов	случаев, n		расчетные	с округлением
0	112	0,2231	115,34	115
1	168	0,3347	173,04	173
2	130	0,2510	129,77	130
3	68	0,1255	64,88	65
4	32	0,0471	24,35	24
5	5	0,0141	7,29	7
6	1	0,0035	1,81	2
7	1	0,0008	0,41	1
Сумма	517	–	516,90	517
$n = 517, X \approx 1,5; e^{-1,5} = 0,2231,$ классы испытаний, $m: 0, 1, 2, 3, 4, 5, 6, 7; \bar{m} = \lambda = \bar{X} = \sigma^2 = 1,5$				

Существенно то, что с помощью теории вероятностей можно оценивать экспериментальные результаты (Sokal, Rohlf, 1995; Zar, 1999). В эксперименте принимается нулевая гипотеза, которая подлежит проверке. При этом желательно свести к минимуму вероятности отклонения истинной нулевой гипотезы – ошибки репрезентативности, а также принятия ложной гипотезы. Перед началом эксперимента следует выбрать уровень величины ошибки репрезентативности  $\alpha$ , который означает вероятность такой ошибки. Например, если  $\alpha = 0,05$ , то при уровне значимости 5 %  $P < 0,05$ . Если нулевая гипотеза не отклонена, следует определить статистическую силу данного эксперимента (надежность), вероятность отказа от нулевой гипотезы, как от ложной.

*D. Использование критерия  $\chi^2$  для анализа генетических расщеплений  
(модельный объект – *Drosophila melanogaster*)*

Критерий  $\chi^2 = \sum \frac{d^2}{q}$  широко используется для анализа генетических расщеплений. Приведем примеры анализа соответствия эмпирических и теоретически ожидаемых расщеплений при моногибридном, дигибридном скрещивании в первом и втором поколении гибридов, а также результатов анализа сцепленного с полом наследования у дрозофилы (табл. 3.7–3.9).

Таблица 3.7

**Пример 1.** Схема анализа количественного расщепления по форме крыльев во втором поколении моногибридного скрещивания дрозофилы.

Комбинация скрещивания:  $\frac{vg}{vg} \times \frac{vg^+}{vg^+}$ .

№ опыта	Количество особей		Всего
	Нормальные крылья ( $p_1$ )	Редуцированные крылья ( $p_2$ )	
1	229	91	320
2	291	82	373
3	284	104	388
6	...	...	...
n	405	119	524
Суммарные эмпирические данные по классам расщепления ( $p$ )	1209	396	1605
Теоретически ожидаемые данные расщепления ( $q$ ), исходя из соотношения 3 : 1	1204	401	1605
Отклонение ( $d$ ) = $p - q$	+5	–5	
$d^2$	25	25	
$d^2/q$	0,02	0,08	
$\chi^2 = \sum \frac{d^2}{q}$	$\chi^2 = 0,02 + 0,06 = 0,08;$ $\chi^2_{st} = \{3,8 - 6,6 - 10,8\}$ (таблица стандартных значений $\chi^2$ , приложение). Число степеней свободы $\gamma = 1$ . (Количество классов расщепления: $n - 1 = 2 - 1 = 1$ .) $\chi^2 = 0,08 < \chi^2_{st}$		

**Выводы:**

- Отклонение фактического расщепления от теоретически ожидаемого имеет случайный характер. Значение эмпирического критерия  $\chi^2$  не достигает стандартных значений.
- Эмпирическое расщепление соответствует теоретически ожидаемому (3 : 1).



Таблица 3.8

**Пример 2.** Схема количественного анализа расщепления по двум парам альтернативных признаков: окраска тела (серая и чёрная) и форма крыльев (нормальные крылья и редуцированные) во втором поколении дигибридного скрещивания дрозофилы.

Комбинация скрещивания:  $\frac{vg}{vg} \frac{e}{e} \times \frac{vg^+}{vg^+} \frac{e^+}{e^+}$  (*vg*, vestigial × *e*, ebony).

Гены: *vg* – 2-я хромосома; *e* – 3-я хромосома

Формула расщепления в  $F_2$ :

$[9/16 e^+ - vg^+ -] : [3/16 e^+ - vg vg] : [3/16 e e vg^+ -] : [1/16 e e vg vg]$ :

№ пробирки или опыта	Фенотипы. Количество особей, шт.				
	Серое тело, нормальные крылья	Серое тело, редуцирован- ные крылья	Чёрное тело, нормальные крылья	Чёрное тело, редуцирован- ные крылья	Всего
1	44	20	21	5	
2	31	11	14	6	
3	34	10	18	3	
n	849	248	267	81	
Суммарные эмпириче- ские данные по классам расщепления ( <i>p</i> )	958	289	320	95	1662
Теоретически ожидае- мые классы расщепле- ния ( <i>q</i> ): 9 : 3 : 3 : 1	934	312	312	104	1662
Отклонение $d = (q - p)$	+24	-23	+8	-9	
$d^2$	576	529	64	81	
$d^2/q$	0,67	1,70	0,20	0,78	
$\chi^2 = \sum \frac{d^2}{q}$	$\chi^2 = 0,67 + 1,70 + 0,20 + 0,78 = 3,35$ ; число степеней свободы $\gamma = 3$ (количество классов расщепления $n - 1 = 4 - 1 = 3$ ); $\chi^2_{st}\{7,8 - 11,3 - 16,3\}$ (таблица стандартных значений $\chi^2$ , приложение); $\chi^2 = 3,35 < \chi^2_{st}$				

Выводы:

- Отклонение фактического расщепления от теоретически ожидаемого носит случайный характер. Значение эмпирического критерия  $\chi^2$  не достигает стандартных значений.

- Эмпирическое расщепление по фенотипу соответствует теоретически ожидаемому: 9 : 3 : 3 : 1.

Таблица 3.9

**Пример 3.** Схема количественного анализа наследования сцепленных полом признаков: красная и белая окраска глаз в реципрокных скрещиваниях дрозофилы  $F_1$  и  $F_2$ .

Комбинации скрещивания: 1) ♀  $\frac{w^+}{w^+} \times \text{♂} \frac{w}{Y}$ ; 2) ♀  $\frac{w}{w} \times \text{♂} \frac{w^+}{Y}$ .

Расщепление	♀ красноглазые × ♂ белоглазые					♀ белоглазые × ♂ красноглазые				
	Количество особей дрозофилы в F <sub>1</sub> и F <sub>2</sub>									
	Красногла- зые		Белоглазые		Все- го	Красногла- зые		Белоглазые		Все- го
	♀	♂	♀	♂		♀	♂	♀	♂	
Первое поколение F <sub>1</sub>										
Фактическое (p)	256	226	–	–	482	215	–	–	183	398
Формула рас- щепления	1	1	Нет	Нет	2	1	Нет	Нет	1	2
Теоретически ожидаемое (q)	241	241	–	–	482	199	–	–	199	398
Отклонение d d <sup>2</sup>	+15 225	–15 225				+16 256			–16 256	
$\chi^2 = \sum \frac{d^2}{q}$	$\chi^2 = 0,93 + 0,93 = 1,86;$ $\gamma = 1, \text{ p} < 0,05$					$\chi^2 = 1,29 + 1,29 = 2,58;$ $\gamma = 1, \text{ p} < 0,05$				
Второе поколение F <sub>2</sub>										
Фактическое (p)	312	163	Нет	129	604	130	151	160	127	568
Формула рас- щепления	2	1	Нет	1	4	1	1	1	1	4
Теоретически ожидаемое (q)	302	151	–	151	604	142	142	142	142	568
Отклонение d	10	13	–	22	0	12	9	18	15	0
d <sup>2</sup>	100	165	–	484	0	144	81	324	225	0
$\chi^2 = \sum \frac{d^2}{q}$	$\chi^2 = 0,33 + 0,95 = 4,49;$ $\gamma = 2, \text{ p} < 0,05$					$\chi^2 = 1,01 + 0,7 + 2,28 + 1,57 = 5,44;$ $\gamma = 3, \text{ p} < 0,05$				

Примечание: Y – игрек-хромосома;  $\gamma$  – число степеней свободы:

$$\chi^2 = \sum \frac{d^2}{q}.$$

Выводы:

1. Отклонения фактических данных от теоретически ожидаемого расщепления носят случайный характер, так как эмпирическое значение  $\chi^2 = \chi^2$  стандартных значений.

2. Эмпирические расщепления соответствуют теоретически ожидаемым результатам.

### *Вопросы*

1. Что такое вероятность?
2. Какие процессы называются вероятностными? Что такое «стохастическая совокупность»?
3. Какими параметрами характеризуется биномиальное распределение? Является ли оно дискретным или непрерывным?
4. Каким является нормальное распределение вероятностей, непрерывным или дискретным?
5. Чем отличаются эмпирические ряды распределений от теоретических?
6. Каковы причины асимметрии и эксцесса?
7. Обязательно ли эмпирический ряд нормального распределения должен охватывать  $\pm 3\sigma$ ?
8. Что такое артефакты? Как их определяют?

## ГЛАВА 4. ОЦЕНКА ДОСТОВЕРНОСТИ ВЫБОРОЧНЫХ ПОКАЗАТЕЛЕЙ

Все статистические величины основаны на изучении массовых явлений в различных совокупностях. При этом возникают следующие вопросы:

1. Насколько достоверно выборочные показатели отражают генеральную совокупность (часть – целое).
2. Все ли разнообразие величин представлено в множестве?
3. Репрезентативна ли выборка?

Исследованиями выдающихся математиков, В. Госсета (Стъудент) и Р. Фишера, доказано, что выборочная и генеральная совокупности характеризуются одними и теми же закономерностями случайной вариации. Следовательно, для генеральной совокупности могут быть вычислены те же показатели. Каково соотношение между выборочными показателями и генеральными параметрами (статистическими показателями, относящимися ко всей генеральной совокупности) – постоянная проблема исследователей.

Биолог почти всегда имеет дело с выборками: при проведении опытов, при изучении материала в природе. При этом генеральные совокупности остаются неизвестными. Поэтому всегда существует риск возникновения ошибки в выводах. Часто эти выводы основываются на изучении небольшого материала, поэтому полученные в опытах или наблюдениях статистические показатели могут иметь значительные *ошибки репрезентативности*.

### **4.1. Ошибки репрезентативности. Причины возникновения**

Варьирование выборочных средних вокруг средней генеральной совокупности приводит к тому, что некоторые опыты могут дать результаты, отклоняющиеся от генерального, параметра на две–три ошибки. Однако при большом количестве опытов их результаты будут группироваться близко к генеральному параметру. Исследование генерального параметра и его доверительных интервалов позволяет уверенно сделать объективные выводы. Связь между выборочными показателями и генеральными параметрами выражаются законом больших чисел. Закон больших чисел заключается в том, что частота  $\frac{m}{n}$  события  $A$  будет сколь угодно близкой к его вероятности, если число испытаний неограниченно возрастает. Разница между эмпирически исследованной частотой события и его вероятностью уменьшается при увеличении числа испытаний. Вероятность приближения  $\bar{X}(M)$  к  $\mu$  становится все большей, стремясь (при  $n = \infty$ ) к единице, т.е. к полной достоверности. В этом сущность теоремы П.Л. Чебышева.

Закон больших чисел, составляющий основу нормального распределения вариантов в вариационном ряду, является также основой распределения

выборочных средних вокруг генеральной средней. При возрастании объема выборочных совокупностей вариация значений их средних арифметических вокруг генеральной средней становится все меньше. Однако в биологических исследованиях часто приходится встречаться с выборочными совокупностями, состоящими из малого количества вариантов или наблюдений:  $n = 30$  и менее. Теоретическое обоснование закономерности распределения выборочных средних арифметических при малых выборках, открытого В. Госсетом (Стьюдент), обосновано Фишером. Это распределение вероятностей получило название  $t$ -распределения.

$$t = \frac{\bar{X} - \mu}{m_{\bar{X}}} n,$$

где

$$m_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

Оказалось, что распределение значений  $t$  отличается от нормального тем сильнее, чем меньше  $n$ . Поэтому, вероятности нахождения выборочных средних в пределах значений  $\pm t$  значительно снижаются по сравнению с нормальным распределением. Для достижения тех же вероятностей нужно взять значительно больший интервал  $\bar{X} \pm tm$ . В практической работе следует исходить из определения, необходимого при данном типе исследований порога вероятностей безошибочного прогноза. Если выборка включает только 10 наблюдений (число степеней свободы 9), а необходим по условиям опыта уровень значимости 0,01 (доверительная вероятность 0,99), то величина  $t$  должна быть не менее 3,25. Уровню значимости 0,05 ( $B = 0,95$ ) при  $\nu = 9$  соответствует величина  $t = 2,62$ . По мере увеличения  $n$  распределение приблизится к нормальному. Исследования Стьюдента (В. Госсет) позволили работать с малыми совокупностями.

Каждая биологическая совокупность может быть представлена в виде ряда распределения. Для распределения можно определить статистические показатели, указывающие:

- на средний уровень значений признака;
- на степень вариации дат, единиц совокупности вокруг этого уровня.

*Точность* – степень приближения выборочного показателя к генеральному параметру при определении надежности его оценки. Показатель точности (уровень ошибки репрезентативности) выборочного показателя определяется на основе выборочных показателей по специальным формулам (табл. 4.2).

*Ошибки репрезентативности* возникают только вследствие того, что целое характеризуется на основе исследования его частей. Ошибки репрезентативности могут быть вскрыты и учтены биометрическими методами. Они не могут быть устранены при самой правильной и безупречной организации

исследования (за исключением перехода на сплошное обследование генеральной совокупности).

В противоположность этому *организационные ошибки* (методические, ошибки точности, внимания, типичности) могут быть устранены безупречным исследованием (или сведены к минимуму). Организационные ошибки не могут быть учтены, устранены математическими приемами.

*Ошибки репрезентативности* можно свести к минимуму путем увеличения выборочной совокупности. Величину ошибок репрезентативности можно определить на основе анализа выборочных данных и учесть при оценке генеральных параметров с определенной точностью и надежностью. Учет ошибок репрезентативности дает возможность: а) определить доверительные границы генеральных параметров; б) определить достоверность выборочных показателей.

#### **4.2. Надежность. Пороги вероятности безошибочных прогнозов**

Надежность – это вероятность того, что генеральный параметр действительно окажется внутри доверительных границ.

Таблица 4.1

**Четыре порога вероятности безошибочных прогнозов**

Порог	Исследования	Надежность (вероятность безошибочного прогноза), $B$	Показатель надежности, $t$	Достаточный объем группы, $n$	Вероятность ошибочного про- гноза, $\alpha$
0	Исследование трудно измеряемых признаков; грубо ориентировочные характеристики	$B_0 \geq 0,90$	$t_0 = 1,645$	20	$\alpha_0 \leq 0,10$
1	Большинство биологических исследований; начальные стадии изучения; описание новых явлений	$B_1 \geq 0,95$	$t_1 = 1,960$	30	$\alpha_1 \leq 0,05$
2	Углубление результатов первых исследований; экономические рекомендации	$B_2 \geq 0,99$	$t_2 = 2,576$	100	$\alpha_2 \leq 0,01$
3	Разрешение спорных вопросов; изучение вредных и ядовитых веществ	$B_3 \geq 0,999$	$t_3 = 3,291$	200	$\alpha_3 \leq 0,001$

Критерий надежности определяется заранее при планировании исследования, исходя из представлений о большей или меньшей ответственности результатов работы. Критерий надежности ( $t$ ) – показатель вероятности безошибочных прогнозов. Практика биологических исследований с помощью теории вероятностей и биометрии выработала четыре порога вероятности безошибочных прогнозов (табл. 4.1). Для выборок, объем которых меньше указанных в таблице, значения показателя надежности ( $t$ ) определяется по таблице Стьюдента (приложение, табл. 4). Критерий Стьюдента используется при установлении доверительных границ генеральных параметров и для оценки достоверности разности. Многие авторы биологических работ вместо вероятности безошибочных прогнозов:  $B_0, B_1, B_2, B_3$  используют вероятность ошибочных прогнозов, каждый из которых равен:  $\alpha = 1 - B$  (или *уровень значимости*).

### **4.3. Порядок оценки генеральных параметров**

Генеральные параметры оцениваются в виде двух значений: минимального и максимального. Доверительные границы – пределы, в которых может находиться искомая величина генерального параметра. Доверительные границы любого генерального параметра определяются по следующему основному правилу: генеральный параметр может отличаться от выборочного показателя не более чем на величину, кратную ошибке репрезентативности выборочного показателя:

$$\begin{aligned}\bar{A} &= \tilde{A} \pm \Delta; \\ \bar{A} &= (\tilde{A} - \Delta) \div (\tilde{A} + \Delta); \\ \Delta &= tm,\end{aligned}$$

где  $\bar{A}$  – генеральный параметр;  $\tilde{A}$  – выборочный показатель;  $t$  – критерий надежности или показатель вероятности того, что величина генерального параметра будет внутри доверительных границ и не выйдет за эти границы. Величина  $t$  показателя надежности устанавливается при планировании исследования;  $m_{\tilde{A}}$  – показатель точности или ошибка репрезентативности выборочного показателя. Рассчитывается по выборочным данным;  $\Delta = tm$  – максимально возможная абсолютная погрешность оценки генерального параметра при данной точности и надежности, равная произведению показателя надежности на показатель точности.

Таблица 4.2

## Формулы для вычисления ошибок репрезентативности

Показатель (m)	Формула
1. Ошибка средней арифметической: — при бесконечной генеральной совокупности; — при конечной генеральной совокупности.	$m = \frac{\sigma}{\sqrt{n}}$ $m = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$
2. Ошибка доли: — если генеральные доли не известны; — если генеральные доли известны; — если в выборке $p=0$ или $p=1$ .	$m_p = \sqrt{\frac{pq}{n-1}}$ $m_p = \frac{0,5}{\sqrt{n}} = \frac{1}{2\sqrt{n}} = \frac{1}{\sqrt{4n}}$ $m_p = \sqrt{\frac{PQ}{n}}$ $m_p = 0 = m_p = 1 = \frac{1}{n+1}$
3. Ошибка разности средних арифметических, долей	$m_d = \sqrt{m_1^2 + m_2^2}$
4. Ошибка разности между выборочной и генеральной долями.	$m_{p-p} = m_p = \sqrt{\frac{PQ}{n}}$

В форме доверительных границ может оцениваться любой генеральный параметр:

- 1) генеральная средняя  $\bar{M}$ ;
- 2) генеральная доля ( $P$ ) – она равна числу плюсовых (+) особей к общему числу обследованных.

## А. Оценка генеральной средней

**Пример.** Урожайность нового сорта риса по 100 пробным участкам:  $M(\bar{X}) = 50$  ц/га,  $\sigma = 2$  ц/га. Дать прогноз среднего урожая при массовых посевах:

- Надежность прогноза ( $t$ ) по 1 порогу вероятности безошибочного прогноза:  $B = 0,95$ .
- Число степеней свободы  $\nu = 100 - 1 = 99$ .
- При  $B = 0,95$  и  $\nu = 99$  – показатель надежности по таблице стандартных значений критерия Стьюдента равен ( $t$ ) = 1,96 ( $\sim 2$ ).

- Точность прогноза:  $m = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{100}} = 0,2$  (ошибка репрезентативности).



- Возможная погрешность:  $\Delta = tm = 2 \cdot 0,2 = 0,4$  ц/га.
- Прогноз генерального параметра (среднего урожая):

$$M(\bar{X}) = M \pm \Delta = 50,0 \pm 0,4 \text{ ц/га.}$$

*Доверительные границы:*

гарантируемый минимум:  $M - \Delta = 50,0 - 0,4 = 49,6$  ц/га;

возможный максимум:  $M + \Delta = 50,0 + 0,4 = 50,4$  ц/га.

$$M \pm \Delta: 49,6 \text{ ц/га} \div 50,4 \text{ ц/га.}$$

### *Б. Оценка средней разности*

**Пример.** При испытании стимулятора сердечной деятельности на 100 особях мышей получено 100 разностей пульса до и после введения стимулятора. Получены следующие показатели:

$n = 100$ , средняя разность  $(\bar{X}) = +1,5$  (удара в минуту),  $\sigma$  разности  $\pm 2,0$  (удара в минуту). Принято:  $B_3 = 0,999$ ,  $\nu = 99$ ,  $t_{st} = 3,4$ .

$$m = \frac{\sigma}{\sqrt{n}} = \frac{2}{10} = 0,2, \Delta = t_m = 3,4 \cdot 0,2 = 0,68 (\sim 0,7).$$

Прогноз генерального параметра (изменения пульса при массовом применении стимулятора):

$$(\bar{X}) = X \pm \Delta = 1,5 \pm 0,7.$$

Гарантированный минимум:  $(\bar{X}) = X - \Delta = 1,5 - 0,7 = 0,8$  удара.

Возможный максимум:  $= 1,5 + 0,7 = +2,2$  удара.

**Вывод.** Получен прогноз действия испытуемого стимулятора: при его массовом применении гарантировано прибавление пульса в среднем на 0,8 ударов в минуту, причем возможно среднее увеличение числа ударов на +2,2 удара в минуту. Такая характеристика генеральной совокупности (случаев применения нового стимулятора) является достоверной, результаты, полученные в выборке, соответствуют тому, что можно наблюдать в генеральной совокупности.

Таблица 4.3 (алгоритм)

Определение доверительных границ генеральных параметров  $M$  и  $P$

$$M = \bar{M} \pm \Delta; \quad \Delta = t \cdot m; \quad m = \frac{\sigma}{\sqrt{n}};$$

$$n = 100, \quad \bar{M} = 200, \quad \bar{\sigma} = 20, \quad m = \frac{20}{\sqrt{100}} = 2,0$$

$$\beta = 0,95; \quad v = n - 1 = 99; \quad t_{st} = 2,0 - 2,6 - 3,4$$

$$\Delta = 2,0 \cdot 2,0 = 4$$

$$M = 20 \pm 4 \begin{cases} \text{не более } 204 \\ \text{не менее } 196 \end{cases}$$

204 — возможный максимум значения генеральной средней;

196 — гарантированный минимум значения генеральной средней

Генеральная доля

$$P = p \pm \Delta; \quad \Delta = t \cdot m; \quad m = \sqrt{\frac{p(1-p)}{n-1}}$$

$$n = 100; \quad p = 0,60; \quad m = \sqrt{\frac{0,6 \cdot 0,4}{99}} = 0,05$$

$$\beta = 0,95; \quad v = n - 1 = 99; \quad t_{st} = \{2,0 - 2,6 - 3,4\}$$

$$\Delta = 2,0 \cdot 0,05 = 0,10 \quad \text{Стьюдента}$$

$$P = 0,60 \pm 0,10 \begin{cases} \text{не более } 0,70 \\ \text{не менее } 0,50 \end{cases}$$

0,70 — возможный максимум генеральной доли;

0,50 — гарантированный минимум генеральной доли

Примечание:  $d$  — разность выборочных средних;  $m_d$  ( $m$ ) — ошибка разности;  $t_d$  — эмпирический критерий достоверности;  $t_{st}$  — стандартное значение критерия Стьюдента.

Оценка достоверности разности средних

**Пример 1.** Сравнивали вес (масса) двух пород кроликов после откорма.

$$n_1 = 20. \quad M_1 \pm m_1 = 4,0 \pm 0,3 \text{ кг};$$

$$n_2 = 25. \quad M_2 \pm m_2 = 4,6 \pm 0,4 \text{ кг};$$

$$d = +0,6; \quad m_d = \sqrt{0,32 + 0,42} = \pm 0,5;$$

$$t_d = 1,2; \quad v = 20 + 25 - 2 = 43;$$

$$t_{st} = \{2,0 - 2,7 - 3,5\}.$$

**Вывод:** Полученная разность не достоверна; осталось невыясненным, какая порода может иметь больший средний вес.

**Пример 2.** Предыдущее исследование было повторено на большей выборке.

$$n_1 = 100, M_1 = 4,1 \pm 0,1;$$

$$n_2 = 100, M_2 = 4,7 \pm 0,1;$$

$$d = +0,6; m_d = \sqrt{0,12 + 0,12} = \pm 0,14; td = 4,3;$$

$$\nu = 100 + 100 - 2 = 198; t_{st} = \{1,7 - 2,0 - 2,6 - 3,4\}.$$

*Выводы:*

1. Разность между средними достоверна по третьему порогу вероятности безошибочного прогноза ( $B \geq 0,999$ ).

2. Животные второй группы достоверно имеют в среднем большую массу тела, чем первой.

Вероятности 0,95; 0,99; 0,999 (95 %, 99 %, 99,9 %) получили название доверительных, т.е. таких значений вероятности безошибочного прогноза, которым можно достаточно доверять, или которыми можно уверенно пользоваться. Вероятности, принятые как доверительные, определяют доверительные границы и доверительный интервал между ними.

На рисунке 4.1 представлена нормальная вариационная кривая, на которой нанесен доверительный интервал при  $B \geq 0,95$ . Вероятность выхода за пределы этого интервала равна 0,05, которая распределяется на две стороны кривой по 0,025 с каждой стороны (2,5 %), – доли площади под нормальной кривой распределения. Уровень значимости – вероятность ошибки, которой решено пренебречь в данном исследовании. Вероятности 0,95 (95 %) соответствует уровень значимости 0,05 (5 %). Это означает, что выход за пределы принятых границ возможен в порядке случайности с вероятностью 0,05, т.е. в 5 % случаев вероятность. При вероятности 0,99 ( $B \geq 99,0$  %) риск получения случайного отклонения (уровень значимости) составляет 1 % (1 случай на 100).



Рис. 4.1. Нормальная вариационная кривая

При уровне значимости 0.01 (второй порог) вывод не обоснован только в 1 % случаев. Такой уровень значимости считают высоким и широко им пользуются при оценке данных опыта. При уровне значимости 0,001 вывод не обоснован только в одном случае из 1000.

### *Вопросы*

1. В какой степени средняя арифметическая выборочной совокупности характеризует среднюю арифметическую генеральной совокупности?
2. Что такое ошибка репрезентативности? Каковы причины ее возникновения?
3. Какая зависимость между величиной ошибки репрезентативности и объемом совокупности?
4. Как оценивается достоверность среднего арифметического?
5. Как оценивается достоверность разности между средними арифметическими?
6. Какие вероятности называются доверительными?
7. Что такое достоверная разность, генеральная разность?
8. Что значит «разность недостоверна»?
9. Означает ли это, что разности нет?
10. Перечислите факторы, определяющие достоверность разности.

## ГЛАВА 5. ВЫБОРОЧНЫЙ МЕТОД ИССЛЕДОВАНИЯ И ОЦЕНКА ГЕНЕРАЛЬНЫХ ПАРАМЕТРОВ

Совокупность, из которой рандомизированно отбирают определенную часть ее членов для совместного изучения особей по определенному признаку, называется генеральной. Теоретически объем генеральной совокупности ничем не ограничен ( $N \rightarrow \infty$ ). Генеральная совокупность (г.с.) мыслится как бесконечно большое множество однородных единиц. Однако на практике объем генеральной совокупности почти всегда ограничен и может быть различным в зависимости от объекта и задач исследования. Так, при анализе кариотипа растений нового вида генеральную совокупность мыслят как совокупность всех растений данного вида. Если же данный вопрос решают для определенного ареала, то генеральную совокупность составляют растения исследуемого вида, обитающие в данном регионе.

### **5.1. Принципы составления случайной выборки, точечные и интервальные оценки достоверности**

Объем выборки ( $n$ ) может быть большим и малым, но не менее двух единиц. Выборочный метод – основной путь изучения статистических биологических совокупностей. Рандомизированно отобранная выборка достаточно объективно отражает структуру генеральной совокупности, хотя полного совпадения выборочных показателей с генеральными параметрами нет. Чтобы это соответствие было максимальным, выборка должна быть репрезентативной (от лат. *represento* – представляю). *Репрезентативность* достигается способом *рандомизации* (от лат. *random* – случай), обеспечивающей равную возможность для всех членов г.с. попасть в выборочную совокупность (гл. 1).

В различных исследованиях применяют несколько разновидностей отбора вариант из генеральной совокупности: простой случайный отбор, типический, серийный, механический и другие разновидности отбора вариант для составления случайной выборки.

1. *Случайный повторный отбор*. Объекты изучения отбираются из генеральной совокупности в случайном порядке. После изучения по определенному признаку объект возвращается в г.с. и может повторно попасть в выборку.

2. *Случайный бесповторный отбор*. Объекты, отобранные случайно, не возвращаются в г.с.

3. *Типический пропорциональный отбор*. На основе предварительного изучения г.с. разбивается на части (например, по типу растительных сообществ, в которых обитают растения данного вида). Из каждой такой части для

исследования отбирается в случайном порядке число экземпляров, пропорциональное плотности (населенности) отдельных частей.

4. *Серийный (гнездовой) отбор*. Применяется тогда, когда объекты исследования равномерно распределены в определенном объеме, территории. Г.с. разбивается на части (серии). Отдельные (некоторые) серии исследуются целиком.

5. *Механический отбор*. Г.с. разбивается на несколько равных частей (квадратов, объемов). Из каждой части (без выбора) берется 1 объект. Или при механическом отборе в выборку должен попасть каждый пятый, десятый или сотый объект данной популяции определенного вида растений или животных.

Числовые показатели, характеризующие генеральную совокупность, называют параметрами (генеральные параметры), а числовые показатели, характеризующие выборочную совокупность, выборочными показателями или выборочными характеристиками. Выборочные показатели являются приблизительными оценками генеральных параметров, варьирующими вокруг них. Рассмотренные ранее оценки генеральных параметров по выборочным показателям подразделяются на *точечные* и *интервальные*.

Точечные оценки (т.о.) представляют собой числа, определяемые по случайной выборке, варьирующие вокруг генеральных параметров, как правило, не совпадающие с ними по абсолютному значению. Генеральные параметры принято обозначать буквами греческого алфавита, а выборочные характеристики – латинского.

Выборочная средняя  $\bar{X}(M)$  является приблизительной оценкой генеральной средней, а среднее квадратическое отклонение  $\sigma(S_x)$  – приблизительной оценкой стандартного отклонения, характеризующего генеральную совокупность. Точечные оценки должны удовлетворять следующим требованиям: быть состоятельными, несмещенными. Точечная оценка является состоятельной, если при увеличении объема выборки она стремится к величине генерального параметра. Для генеральной средней ( $\mu$ ) состоятельной оценкой является выборочная средняя  $\bar{X}$ . Для генеральной дисперсии состоятельной оценкой будет выборочная дисперсия. Точечная оценка эффективна, если она имеет наименьшую дисперсию выборочного распределения, т.е. обнаруживает наименьшую случайную вариацию. Точечная оценка является несмещенной, если математическое ожидание ее выборочного распределения совпадает со значением генерального параметра. Очевидно, что перечисленные свойства. **г.с.** выявляются при изучении большого числа выборок из бесконечно большой генеральной совокупности. При этом на основе каждой выборки проводится одно и то же количество наблюдений по данному признаку ( $n$ ). В силу случайных причин выбо-

рочные средние будут варьировать, образуя некоторое распределение (*выборочное распределение статистики*).

Выборочная средняя является несмещенной оценкой генеральной средней, а выборочная дисперсия представляет собою смещенную оценку относительно генерального параметра на величину  $n(n - 1)$ . Чтобы получить несмещенную оценку генеральной дисперсии, необходимо при вычислении выборочной дисперсии, а также стандартного отклонения сумму квадратов центральных отклонений делить не на число наблюдений, а на число степеней свободы:

$$v = n - 1.$$

Показателем точности определения выборочной средней является отношение ошибки репрезентативности к своей средней (в процентах); выборочные показатели подразделяются на точечные и интервальные.

### *Точечные оценки*

*Точечные оценки* представляют собой значения чисел, определяемые по случайной выборке, варьирующие вокруг генеральных параметров, как правило, не совпадающие с ними по абсолютному значению. Генеральные параметры принято обозначать буквами греческого алфавита, а выборочные характеристики – латинского. Так, выборочная средняя  $\bar{X}(M)$  является приблизительной оценкой генеральной средней, среднее квадратическое отклонение  $\sigma(S_x)$  – приблизительной оценкой стандартного отклонения генеральной совокупности.

*Точечные оценки* должны удовлетворять следующим требованиям: быть состоятельными, несмещенными. Точечная оценка является состоятельной, если при увеличении объема выборки она стремится к величине генерального параметра. Для генеральной средней состоятельной оценкой является выборочная средняя  $\bar{X}(\mu)$ , для генеральной дисперсии состоятельной оценкой является выборочная дисперсия. Точечная оценка будет эффективной, если она имеет наименьшую дисперсию выборочного распределения, т.е. обнаруживает наименьшую случайную вариацию. Точечная оценка является несмещенной, если математическое ожидание ее выборочного распределения совпадает со значением генерального параметра. Перечисленные свойства точечных оценок выявляются при изучении большого числа выборок из бесконечно большой генеральной совокупности. При этом на основе каждой выборки проводится одно и то же количество наблюдений по данному признаку. Под влиянием случайных причин выборочные средние будут варьировать, образуя распределение (*выборочное распределение статистики*).

Показателем точности выборочной средней является отношение ошибки репрезентативности (доля) к своей средней (в %):

$$p = \frac{m}{M} \cdot 100 \%,$$

или

$$p = \frac{m}{\bar{X}} \cdot 100 \%,$$

или

$$p = \frac{S_x}{\bar{X}} \cdot 100 \%.$$

### *Интервальные оценки*

Любой генеральный параметр оценивается в виде доверительных границ, пределов, в которых может находиться величина генерального параметра. Кроме рассмотренных выше способов оценки доверительных границ генеральной средней, генеральной доли и генеральной разности, существуют способы оценки генеральных параметров и других статистических показателей.

## **5.2. Статистические гипотезы, их проверка**

### *Нулевая гипотеза ( $H_0$ )*

Смысл ее заключается в предположении, что разность между генеральными параметрами двух сравниваемых совокупностей равна нулю, и что различия между ними по данному признаку носят случайный характер. Так, если параметры одной нормально распределяющейся совокупности  $\mu_x$  и  $\sigma_x$ , а другой  $\mu_y$  и  $\sigma_y$ , то нулевая гипотеза исходит из того, что  $\mu_x = \mu_y$ ;  $\sigma_x = \sigma_y$  ( $\mu_x - \mu_y = 0$  и  $\sigma_x - \sigma_y = 0$ ).

С помощью теории вероятностей можно оценивать экспериментальные результаты (Sokal, Rohlf, 1995; Zar, 1999). В эксперименте принимается нулевая гипотеза, которая подлежит проверке. При этом желательно свести к минимуму вероятности отклонения истинной нулевой гипотезы – ошибки первого типа, а также принятия ложной гипотезы – ошибки второго типа. Перед началом эксперимента следует выбрать уровень величины ошибки первого типа – ( $\alpha$ ) или  $P$ , который означает вероятность такой ошибки. Например, если  $\alpha \leq 0,05$ , то при уровне значимости 5 %  $P \leq 0,05$ . Если нулевая гипотеза не отброшена, следует определить статистическую силу данного эксперимента (надежность) – вероятность отказа от нулевой гипотезы как от ложной. Согласно этой гипотезе, первоначально принимается, что между данными показателями (или группами, на основе которых они получены) достоверных различий нет, т.е. обе группы вместе составляют один и тот же однородный материал, одну совокупность. Статистический анализ должен может привести к отклонению нулевой гипотезы, если достоверность полученных различий не доказана (различия признаны случайными). Поскольку все статистические



показатели и различия между ними характеризуются определенными уровнями значимости, отклонение нулевой гипотезы должно быть связано с принятием определенного уровня значимости. Так, если в эксперименте признан необходимым уровень значимости 0,01 и если вероятность достоверности данного статистического показателя или разница между показателями не удовлетворяет этому условию, т.е. она ниже 0,99 (например, 0,97, 0,91, 0,98), то нет основания для отклонения нулевой гипотезы. Ее надо считать правильной до тех пор, пока новые данные не дадут возможности ее опровергнуть, доказав, что существующие различия не являются случайными. В том случае, когда нулевая гипотеза считается опровергнутой, все же остается вероятность, что она в действительности верна. При уровне значимости 0,01 этот шанс составляет 1 из 100, т.е. в 1 % случаев отклонение нулевой гипотезы было бы ошибкой. Если уровень значимости достиг 0,001, то уверенность в том, что нулевая гипотеза отвергнута, резко возрастает (лишь 1 шанс на 1000 случаев, что она все же верна). При  $\alpha = 0,05$  уверенность в правильности вывода составляет лишь 95 случаев из 100, а в 5 % возможен неправильный вывод. Таким образом, если полученные данные характеризуются уровнем значимости  $\alpha > 0,05$ , то нет оснований отклонять нулевую гипотезу. Если уровень значимости ( $\alpha$ )  $< 0,01$ , то для отбрасывания нулевой гипотезы основания достаточны.

#### *Альтернативная гипотеза ( $H_a$ )*

Противоположна нулевой, *альтернативная гипотеза, которая* исходит из предположения, что  $\mu_x - \mu_y \neq 0$  и  $\sigma_x - \sigma_y \neq 0$ . В таком случае для проверки принятой гипотезы и достоверности оценки генеральных параметров по выборочным данным используют величины, функции распределения которых известны: *критерии достоверности оценок*.

### **5.3. Параметрические и непараметрические критерии оценок достоверности**

В биометрии используют два типа статистических критериев достоверности: *параметрические и непараметрические*.

*Параметрические критерии* построены на основе параметров данной совокупности ( $\bar{X}, \sigma^2$ ) и других, представляющих функции этих параметров. Они служат для проверки гипотез о параметрах совокупностей, распределяемых по нормальному закону. *Непараметрические критерии* представляют собой функции, зависящие непосредственно от вариантов данной совокупности с их частотами. Они служат для проверки рабочих гипотез независимо от

формы распределений совокупностей, из которых взяты сравниваемые выборки. Параметрические критерии способны более эффективно отвергнуть нулевую гипотезу. Поэтому при сравнении нормально распределяющихся совокупностей предпочтение отдается параметрическим критериям. Из параметрических критериев в биометрии широко применяются: уже обсуждавшийся критерий Стьюдента, используемый для сравнительной оценки средних величин,  $F$  критерий Фишера, используемый для оценки дисперсий.

### *Параметрические критерии.*

#### *t-Критерий Стьюдента (t-распределение)*

Разработанный Стьюдентом (В. Госсет) и обоснованный Р. Фишером, закон -распределения служит основой теории малой выборки. Закон характеризует распределением выборочных (*выборка*) средних в нормально распределяющейся совокупности в зависимости от объема выборки. В. Госсет нашел закон распределения величины  $t$ :

$$t = \frac{(\bar{X} - \bar{\mu})}{\sigma\sqrt{n}},$$

где генеральной параметр заменен на его выборочную характеристику:  $\sigma$ , т.е. нашел закон распределения значений  $t$ . Оказалось, что отношение разности выборочной и генеральной средних к ошибке выборочной средней непрерывно распределяется в соответствии с уравнением:

$$f(t) = C(1 + \frac{t^2}{n-1})^{-\frac{n-1}{2}} \quad (-\infty < t < +\infty),$$

где  $C$  – константа, зависящая от числа степеней свободы  $\nu = n - 1$ ;  $t$  – распределение зависит только от числа степеней свободы,  $\nu$ . С увеличением объема выборки ( $n$ ),  $t$  – распределение быстро приближается к нормальному, и уже при  $n = 30$  не отличается от него. Для выборок, объем которых превышает 30 единиц, величина  $t$  распределяется нормально и не зависит от числа наблюдений. Если  $n < 30$  характер распределения находится в зависимости от числа наблюдений ( $n$ ). Для практического использования - распределения в приложении приведена таблица 4 (Стьюдента), в которой содержатся критические точки ( $t_{st}$ ) для разных уровней значимости и числа степеней свободы.

*Оценка достоверности разности средних  
(достоверность разности)*

В биологии из всех математических функций наибольшее значение имеет разность двух величин. Критерий Стьюдента предназначен для сравнения двух групп (совокупностей). По разности средних сравнивают: разные группы, популяции, расы, породы, сорта, опытные и контрольные варианты, особей одной группы, но разного возраста, в разных сезонах года, в разных географических зонах, в разных условиях, выявляют результаты воздействия различных факторов на признаки особи.

*Критерий достоверности разности Стьюдента*

$$t_d = \frac{d}{m_d} \geq t_{st}, \quad \nu_d = n_1 + n_2 - 2;$$

$$t_d = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{m_1^2 + m_2^2}},$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}}}$$

$d - \bar{X}_1 - \bar{X}_2$  выборочная разность средних:  $M_1 - M_2$ ; а также  $r_1 - r_2$ ;  $p_1 - p_2$ ;  $m_d$  – ошибка разности.

При сравнении двух выборочных совокупностей возникает вопрос: насколько правильно выборочные данные отражают генеральные отношения? Достоверность разности – свойство выборочной разности с заданной надежностью отражать по знаку генеральную разность.

Разность достоверна – это значит, что разность выборочных показателей может быть перенесена на генеральную совокупность.

Разность недостоверна – не получено определенного ответа о разности между генеральными параметрами. Это не обозначает отсутствия разности генеральных средних.

Вывод: между генеральными средними могут быть любые отношения (из перечисленных выше), но какие – неизвестно.

$m_d$  – ошибка разности выборочных средних;  $t_d$  – эмпирический критерий достоверности разности;  $t_{st}$  – стандартные значения критерия Стьюдента (приложение, табл. 4) в зависимости от числа степеней свободы и принятого порога вероятности безошибочного прогноза.

Число степеней свободы:  $\nu = n_1 + n_2 - 2$ .

При  $t_d \geq t_{st}$  – разность достоверна; при  $t_d \leq t_{st}$  – разность недостоверна.

Таблица 5.1 (алгоритм)

**Критерий достоверности разности. Критерий Стьюдента**

$$t_d = \frac{|d|}{m_d} = \frac{M_2 - M_1}{\sqrt{m_1^2 + m_2^2}} = t_{st} \begin{cases} \beta_1 = 0,95 \\ \beta_2 = 0,99 \\ \beta_3 = 0,999 \end{cases} (v_d = n_1 + n_2 - 2),$$

$M_1, M_2$  – сравниваемые средние;

$m_1^2 = \frac{\sigma_1^2}{n_1}, m_2^2 = \frac{\sigma_2^2}{n_2}$  – квадраты ошибок репрезентативности;

$$\frac{\sigma^2}{n} = \frac{c}{n(n-1)}, \quad C = \sum (V - M)^2 = \sum V^2 - \frac{(\sum V)^2}{n},$$

$n_1, n_2$  – объемы выборок;

$t_{st}$  – стандартные значения критерия определяются по таблице критериев Стьюдента по числу степеней свободы ( $v = n_1 + n_2 - 2$ ) для одного из трех порогов вероятности; при  $t_d \geq t_{st}$  разность достоверна, подчеркивается одной, двумя или тремя чертами, при  $t_d < t_{st}$  разность недостоверна, подчеркивается волнистой чертой.

Если  $\bar{M}_2 > \bar{M}_1$  и  $t_d \geq t_{st}$ , то и  $\bar{M}_2 > \bar{M}_1$ .

Если  $\bar{M}_2 > \bar{M}_1$  и  $t_d < t_{st}$ , то и  $\bar{M}_2 \cong \bar{M}_1$ .

**Пример:**

$$n_1 = 25; M_1 = 232; \alpha_1 = 23; m_1^2 = \frac{23^2}{25} = 21,16;$$

$$n_2 = 36; M_2 = 210; \alpha_2 = 21; m_2^2 = \frac{21^2}{36} = 12,25;$$

$$t_d = \frac{22}{5,78} = \underline{\underline{3,8}}; |d| = 210 - 232 = |22|; m_d^2 = 33,41;$$

$$m_d = 5,78;$$

$$v_d = 25 + 36 - 2 = 59; t_{st} = \{2,0 - 2,7 - 3,5\}$$

**Достоверность разности (что это значит?)**

- возможность обобщения результатов выборочного исследования;
- достаточная вероятность: различия генеральных средних;
- вероятность такая, как получено в выборочном исследовании;
- достаточная вероятность действия изучаемого фактора при его массовом применении;
- достаточная вероятность такого действия, какое обнаружено в эксперименте.

Недостоверность разности: невозможность любого прогноза величины различия и знака разности между соответствующими генеральными совокупностями.

### Критерий Фишера (F-распределение)

Для сравнения вариации признаков биологических объектов и установления достоверности различий между группами более целесообразно использование F-критерия Фишера. Критерий представляет собой отношение дисперсий (средних квадратов):

$$F = \sigma_1^2 / \sigma_2^2.$$

Если обе дисперсии равны, тогда  $F = 1$ . Нулевой гипотезой является признание равенства дисперсий. Если они не равны, то нужно доказать, что это неравенство достоверно (не случайно).

Таблица 5.2 (алгоритм)

#### Критерий Фишера. F-критерий Фишера (F-распределение)

$$F_d = \frac{d^2}{\sigma_z^2} \cdot \frac{n_1 \cdot n_2}{n_1 + n_2} \geq F_{\alpha} \left\{ \begin{array}{l} \nu_1 = 1 \\ \nu_2 = n_1 + n_2 - 2 \end{array} \right\} \left\{ \begin{array}{l} \beta_1 = 0,95 \\ \beta_2 = 0,99 \\ \beta_3 = 0,999 \end{array} \right\}$$

$d^2$  — квадрат разности средних  $(M_2 - M_1)^2$

$$\sigma_z^2 = \frac{(n_1 - 1) \sigma_1^2 + (n_2 - 1) \sigma_2^2}{n_1 + n_2 - 2} = \frac{C_1 + C_2}{n_1 + n_2 - 2} = \text{дисперсия случайного разнообразия}$$

$n_1, n_2$  — объемы выборок

$F_{\alpha}$  — стандартные значения критерия Фишера находятся по специальной таблице на основе двух чисел свободы

$$(\nu_1 = 1; \nu_2 = n_1 + n_2 - 2)$$

для одного из трех порогов вероятности

При  $F_d \geq F_{\alpha}$  — разность достоверна;

при  $F_d < F_{\alpha}$  — разность недостоверна.

Употребляется критерий Фишера в тех случаях, когда нет противопоказаний к тому, чтобы считать разнообразие обеих выборок достаточно близким

Пример:

$$\begin{array}{l} n_1 = 25; \quad M_1 = 232; \quad \sigma_1 = 23 \\ n_2 = 36; \quad M_2 = 210; \quad \sigma_2 = 21 \end{array} \left\{ \begin{array}{l} d = 22; \quad d^2 = 484 \\ \sigma_z^2 = \frac{24 \cdot 23^2 + 35 \cdot 21^2}{25 + 36 - 2} = 476,8 \\ \nu_1 = 1; \quad \nu_2 = 25 + 36 - 2 = 59 \end{array} \right.$$

$$F_d = \frac{484}{476,8} \cdot \frac{25 \cdot 36}{25 + 36} = 14,9$$

$$F_{\alpha} = \{4,0 - 7,1 - 12,0\}$$

Критические значения  $F$ , являющиеся границами для признания достоверности разности между вариансами, приводятся в таблице 5 приложения. Обычно отношение вариансов берут так, чтобы в числителе была большая варианса. Если эмпирический критерий  $F > F_{st}$  при принятом уровне значимости, различие между вариансами является достоверным; если  $F \leq F_{st}$ , то различие между вариансами недостоверно, т.е. случайно, и нулевая гипотеза остается не опровергнутой. Значение критерия  $F$  особенно велико в дисперсионном анализе.

*Пример использования критерия Фишера.* Испытывалось влияние шести рационов кормления на яйценоскость кур.

Варианса между средними арифметическими групп с разными рационами:  $\sigma_1^2 = 1074,5$   $\nu_1 = 5$ .

Варианса внутригруппового разнообразия  $\sigma_2^2 = 312,4$  при  $\nu_2 = 114$ . Межгрупповое разнообразие оказалось больше внутригруппового:

$$F = \frac{1074,5}{312,4} = 3,4.$$

При  $\nu_1 = 5$   $\nu_2 = 114$  по таблице стандартных значений критерия Фишера:

$$F_{st} = \{2,3 - 3,2 - 4,5\}.$$

Эмпирическое значение критерия Фишера  $F=3,44$  превышает стандартное значение по второму порогу вероятности безошибочного прогноза. Следовательно, различия между группами кур с разными рационами кормления достоверны с вероятностью 0,99 (и только в одном случае из 100 эта разница может быть следствием случайных причин).

#### *Оценка достоверности разности между попарными данными*

В ряде случаев такой метод сравнения позволяет значительно упростить исследования по проверке достоверности разности.

Данные четвертого столбца (табл. 5.3, показатель  $d$ ) были обработаны, как вариационный ряд. В результате получено:

$$d(\text{средняя разность}) = 2,04 \text{ г}; \quad \sigma^2 = 13,18; \quad \sigma = 3,63 \text{ г},$$

$$m_d = 0,73; \quad t = \frac{d}{m_d} \cdot 2,81.$$

При  $\nu = 24$ ;  $B = 0,99$  (2-й порог вероятности). Разность достоверна при  $t = 2,81$ . Эмпирическое значение  $t = 2,81$  находится на границе требуемой достоверности. Таким образом, разница между средним весом самок и самцов достоверна, и нулевая гипотеза может быть отвергнута.

Таблица 5.3

**Парное сравнение массы тела (веса) самок и самцов мышей (г)  
в возрасте 12 дней в 25 пометах (по П.Ф. Рокицкому)**

Номер помета	Вес		$d$ раз- ность	Номер помета	Вес		$d$ раз- ность
	♀	♂			♀	♂	
1	2,6	16,5	9,5	14	22,5	20,5	2,0
2	20,0	17,0	3,0	15	23,5	19,5	4,0
3	18,0	16,0	2,0	16	23,5	22,5	1,0
4	28,5	21,0	7,5	17	25,0	20,0	5,0
5	23,5	23,0	0,5	18	24,5	20,5	4,0
6	20,0	19,5	0,5	19	23,5	18,0	5,5
7	22,5	18,0	4,5	20	20,5	24,5	-4,5
8	24,0	18,5	5,5	21	20,0	22,0	-2,0
9	24,0	20,0	4,0	22	20,5	20,0	0,5
10	25,0	28,0	-3,0	23	25,0	20,0	5,0
11	22,0	27,0	-5,5	24	23,5	23,0	0,5
12	24,0	20,5	3,5	25	22,0	24,0	-2,0
13	22,5	23,0	-0,5				

*Непараметрические критерии достоверности*

Использование параметрических критериев для проверки статистических гипотез основано на предположении о нормальном распределении совокупностей, из которых получены сравниваемые выборки.

Непараметрические критерии используются:

а) при исследовании биологических признаков, распределение которых в совокупностях не соответствует нормальному закону;

б) при исследовании качественных признаков, отмечаемых порядковыми номерами, индексами, долями.

Для проверки нулевой гипотезы при сравнении выборочных групп используются различные непараметрические критерии, к числу которых относятся:

$Z$  – критерий знаков;

$T$  – ранговый критерий Уилкоксона;

$U$  – критерий Уилкоксона (Манна – Уитни);

$X$  – критерий Ван-дер-Вардена и др.

**Пример** использования одного из них.

$Z$ – критерий знаков. Метод удобно использовать для первоначальной оценки результатов опытов.

Применение критерия знаков можно рассмотреть на основе приведенного выше сравнения массы тела самок и самцов мышей. Из 25 сравниваемых пар в 19 случаях вес самок был выше веса самцов: 19 (+) и в 6 случаях меньше: 6 (–). Из приводимой в приложении таблицы 8 сравнения пар видно, что при объеме выборочной совокупности  $n = 25$  достаточно 17–18 случаев со знаком (+), чтобы считать разницу достоверной с уровнем значимости  $p = 0,05$ , и 19–20 случаев (+), чтобы считать разницу достоверной с уровнем значимости  $p = 0,01$ . Следовательно, проверка с помощью критерия  $Z$  знаков дала приблизительно те же результаты, что и сравнение разности с ее ошибкой.

Этот метод недостаточно чувствителен и не всегда позволяет обнаружить реально существующие различия, как это можно сделать путем применения  $t$ -критерия достоверности разности Стьюдента. Наиболее адекватно  $t$ -критерий знаков используется в тех случаях, когда результаты наблюдений выражаются не числами, а знаками: (+) или (–).

$T$ -критерий Уилкоксона используется, когда члены сравниваемых выборок связаны попарно некоторыми общими условиями (зависимые выборки), например: животные (матери и дочери). При этом следует ранжировать разности.

$X$ -критерий Ван-дер-Вардена также относится к ранговым критериям. Его используют для проверки нулевой гипотезы при сравнении независимых выборок.

$U$ -ранговый критерий Уилкоксона (Манна – Уитни) используют для проверки гипотезы о принадлежности сравниваемых независимых выборок к одной и той же генеральной совокупности или к совокупностям с одинаковыми параметрами.

### Критерий $\chi^2$ (хи-квадрат) К. Пирсона

Критерий применяется во всех случаях, когда требуется определить соответствие эмпирического распределения вероятностей теоретическому:

$$\chi^2 = \sum \frac{(f-f')^2}{f'} = \sum \frac{d^2}{f'},$$

где  $d = (f - f')$  – разность между эмпирическими и теоретически ожидаемыми частотами;  $f$  – эмпирические частоты ординат;  $f'$  – теоретически ожидаемые частоты.

При определении различий между эмпирическими и теоретическими распределениями (так же, как и в случае применения критерия  $\lambda$ ) используется обратный порядок планирования порогов вероятности безошибочных прогнозов. Эмпирическое значение  $\chi^2$  необходимо сравнить со стандартными



значениями этого критерия (по таблице 6 приложения стандартных значений  $\chi^2$  с учетом числа степеней свободы).

При определении числа степеней свободы в нормальном распределении дат по классам существуют три ограничения:

- 1) определенный объем совокупности ( $n$ );
- 2) определенная средняя  $M(X)$ , от которой устанавливаются центральные отклонения;
- 3) определенная сигма ( $\sigma$ ), по которой нормируются центральные отклонения средних значений классов.

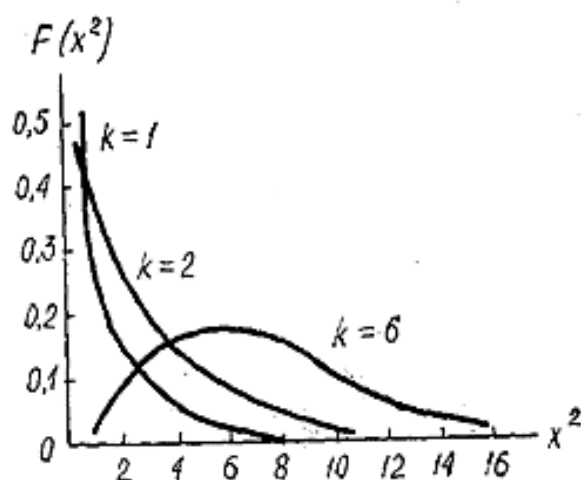


Рис. 5.1. Функция  $\chi^2$  распределения в зависимости от числа степени свободы ( $k$ )

Число степеней свободы устанавливается так:

1.  $\nu_1 = r - 3$  (имеющееся число классов без трех).
2. По первому числу степеней свободы устанавливается минимально допустимая теоретическая частота крайних классов.
3. Классы с малыми теоретическими частотами менее пяти дат объединяются в один общий класс, получается второе, уменьшенное число классов:

$$\nu_2 = r_2 - 3.$$

Распределение вероятностных значений случайной величины является непрерывным и асимметричным (рис. 5.1), оно зависит от числа степеней свободы и приближения к нормальной кривой по мере увеличения числа испытаний ( $n$ ). Поэтому применение критерия  $\chi^2$  к оценке дискретных распределений связано с некоторыми погрешностями, особенно при малых выборках. Чтобы оценки были точными, выборка, распределяемая в вариационный ряд, должна содержать не менее 50 вариантов.

*Применение критерия  $\chi^2$  для анализа генетических расщеплений рассмотрено в главе 3.*

Нулевая гипотеза сводится к тому, что различия между эмпирическими и теоретически ожидаемыми частотами носят случайный характер. Для проверки нулевой гипотезы следует эмпирическую величину критерия  $\chi^2$  сопоставить со стандартными значениями  $\chi^2_{st}$ .

#### **5.4. Причины асимметрии эмпирических распределений**

*Причина 1* – механическая, связана с неправильной группировкой выборочных данных. Такую асимметрию В. Иогансен назвал кажущейся или ложной, так как в результате незначительного изменения границ классов средняя арифметическая и сигма остаются неизменными, но асимметрия может уменьшиться в несколько раз.

*Причина 2* – модифицирующее влияние условий внешней среды для формирования признаков: «краевое» влияние, густота стояния растений, пестрота почвенного плодородия и другие. В таких случаях резко увеличиваются коэффициент вариации и асимметрия эмпирического распределения.

*Причина 3* – обусловлена различными типами взаимодействия генов: (эпистаз, полимерия, комплементарность). При полигенном наследовании количественных признаков одни гены ограничивают проявление в фенотипе других, поэтому кривая распределения таких признаков окажется асимметричной. Известно, что многие количественные признаки растений, животных обусловлены аддитивными генами: длина вегетационного периода у растений, удой и жирность молока у животных, длина початка у кукурузы и др.

Перечисленные причины могут действовать вместе, поэтому статистический анализ нельзя отрывать от биологического анализа причин асимметрии кривых эмпирических распределений. Возможно также большее сгущение вариантов вблизи средней арифметической при недостатке их в боковых частях распределения («крутизна») и, наоборот, ненормально малая частота вариантов в классах, близких к средней арифметической («плосковершинность»). Исследователь должен внимательно анализировать полученные эмпирические ряды распределения и, оценивая их математически, не забывать о биологической природе исследуемого признака, не стремясь подогнать кривые распределения к тому или иному типу теоретических кривых.

### Вопросы

1. Приведите примеры дискретной и непрерывной вариации признака.
2. Каковы особенности группировки дат при непрерывной и дискретной вариации?
3. Что такое статистическое распределение выборки?
4. Что значит, выборка репрезентативна?
5. Нужно ли считаться с возможностью событий, обладающих малой вероятностью?
6. Что такое вероятность? Как она определяется?
7. Какие процессы называются вероятностными или стохастическими?
8. Приведите примеры биологических явлений, осуществление которых может быть оценено известной вероятностью.
9. Какое значение имеет  $p$  для достоверных событий?
10. Какая взаимосвязь существует между частотой определенного явления и его вероятностью?
11. Чему равна сумма  $p + q$ ?
12. Какова разница между эмпирической и теоретической вероятностью?
13. Теоремы сложения и умножения вероятностей, примеры.
14. Какая связь существует между вариацией в пределах вариационного ряда и распределением вероятностей?
15. Что иллюстрирует аппарат Гальтона?
16. Что такое биномиальная кривая распределения?
17. Какая формула является основой для биномиального распределения?
18. Что такое  $k$  в бинOME  $(p + q)^k$ ?
19. Какими параметрами характеризуется биномиальное распределение? Является ли оно дискретным или непрерывным?
20. Чем отличается распределение Пуассона от биномиального?
21. Какими параметрами характеризуется распределение Пуассона?
22. Какой ряд выражает частоты распределения Пуассона?
23. Что такое нормальное распределение? Каковы его закономерности?
24. Что такое нормированное отклонение?
25. Сколько  $t$  охватывает вариационный ряд при нормальном распределении?
26. Что показывает таблица нормального интеграла вероятностей?
27. Какой процент особей укладывается в интервале  $\pm 1\sigma$ ,  $\pm 2\sigma$ ,  $\pm 3\sigma$ ?
28. Какие вероятности называются доверительными?

29. Что такое «доверительные границы» и «доверительные интервалы»?
30. Каков доверительный интервал при нормальном распределении с вероятностью 0,90; 0,95; 0,99; 0,999?
31. Что такое уровень значимости? Какая связь между уровнем значимости и вероятностью?
32. На что указывает уровень значимости 6 %?
33. Каков характер распределения при малых значениях  $n$ ?
34. Чем могут отличаться эмпирические ряды распределения от теоретических рядов?
35. Что такое распределение вероятностей, как оно может быть представлено?
36. Всегда ли кривые распределения симметричны?
37. Каковы причины асимметрии эмпирических кривых распределения?
38. Обязательно ли эмпирический ряд распределения соответствует интервалу  $\pm 3\sigma$ ?

## ГЛАВА 6. ОПРЕДЕЛЕНИЕ ДОСТАТОЧНОЙ ЧИСЛЕННОСТИ ВЫБОРКИ

При проведении биологических исследований одним из важнейших является вопрос о том, сколько объектов (растений, животных) данного вида необходимо включить в выборку, чтобы получить объективное представление о популяции вида (по исследуемому признаку). Практический опыт научных исследований, а также понимание закономерностей нормального распределения показывают, что нецелесообразно стремиться к неоправданно большому числу испытаний, если достоверный результат можно получить при минимально допустимом объеме выборки. Необходимая численность выборки, отвечающая точности ожидаемого результата, зависит от величины ошибки выборочной средней.

Чтобы оптимально прогнозировать объем выборки, необходимо:

- определить допустимую погрешность изучаемого показателя  $\Delta$ , вероятность безошибочного прогноза  $(B, t)$  или надежность доверительных границ;

- определить значение генеральных параметров:

$$\bar{X}(M) = \bar{X} \pm tm, \quad \bar{\sigma} = \bar{\sigma} \pm tm;$$

- освоить математические приемы, которые связывают все исходные данные в единую формулу, дающую в конечном итоге ответ о достаточном числе первичных наблюдений ( $n$ ),

$$n = \frac{t^2}{K^2},$$

где

$$K = \frac{\Delta}{\sigma}, \quad m = \frac{\sigma}{\sqrt{n}},$$

откуда

$$\sigma = m\sqrt{n}, \quad n = \frac{t^2 \sigma^2}{\Delta^2},$$

где  $K$  – нормированная погрешность;  $t$  – нормированное отклонение, с которым связан тот или иной уровень значимости,  $\sigma^2$  – дисперсия;  $\Delta = t \cdot m$  – абсолютная погрешность, величина, определяющая границы доверительного интервала.

Допустимая погрешность  $\Delta$  прогноза генерального параметра определяется по выборочному показателю, исходя из особенностей изучаемого признака и ответственности планируемой работы. При рекогносцировочных исследованиях допустима большая погрешность, чем в тех случаях, когда уже требуется получить более точное значение генеральной доли, генеральной средней.

Среднее квадратическое отклонение для количественных признаков можно прогнозировать одним из следующих способов:

а) если имеются данные об этом показателе на основе выборочных исследований, то берется средняя из частных сигм:

$$\sigma = \frac{\sum \sigma_i}{g},$$

где  $g$  – число исследований;

б) если подобных сведений нет, то можно наметить лимиты признака в генеральной совокупности и определить сигму;

в) при невозможности применения этих способов производится пробное исследование признака не менее чем у 30 случайно взятых объектов и на основе этих данных рассчитывается сигма.

Нормированная погрешность ( $K$ ) определения генеральной средней арифметической  $K = \frac{\Delta}{\sigma}$ .

Расчет достаточной численности выборки:

$$\hat{n} = \frac{t^2}{K^2}.$$

При

$$\begin{aligned} t &= 1,96 (2), & \Delta &= 10, \\ \sigma &= 50, & K &= \frac{10}{50} = 0,2, \\ \hat{n} &= \frac{2^2}{(0,2)^2} = 100. \end{aligned}$$

Можно воспользоваться таблицей 6.1, в которой объем выборки рассчитан в зависимости от  $t$  и  $K$  (приложение, табл. 7).

**Пример.** Необходимо определить возможную среднюю массу плода у томатов сорта Брекодей с допустимой погрешностью не более:  $\Delta = 1$  г и вероятностью безошибочного прогноза  $B_3 = 0,999$  (III порог)  $t_3 = 3,29$ . По массе уже взвешенных плодов при  $n = 100$  рассчитаны показатели:  $M = 70$  г,  $\sigma = 10$  г,  $K = \Delta/\sigma = 1/10 = 0,10$ .

Объем выборки, достаточной для более точного прогноза:

$$\hat{n} = \frac{3,29^2}{0,10^2} = 1082 \text{ плода.}$$

Таблица 6.1

**Достаточная численность выборки**

при прогнозе генеральной средней, при допустимой нормированной погрешности

 $K = \Delta/\sigma$  для четырех порогов вероятности показателей надежности $t$  (критерий Стьюдента)

K	Пороги			
	0	1	2	3
	$B = 0,90$ $t = 1,65$	$B = 0,95$ $t = 1,96$	$B = 0,99$ $t = 2,58$	$B = 0,999$ $t = 3,29$
0,01	27060	38420	66360	108310
0,02	6765	9605	16590	27078
0,03	3007	4269	7324	12035
0,04	1692	2402	4148	5194
0,05	1083	1537	2655	4343
0,06	752	1067	1844	3009
0,08	423	601	1037	1693
0,10	271	384	666	1082
0,12	188	260	461	752
0,14	138	196	339	553
0,16	105	150	260	423
0,18	90	119	204	334
0,20	68	96	166	271
0,22	56	80	137	224
0,24	47	67	115	188
0,26	40	57	98	160
0,28	35	49	85	138
0,30	30	43	74	121
0,35	22	32	54	89
0,40	17	24	40	68
0,50	11	16	27	44

*Качественные признаки.* Достаточный объем выборки определяется с учетом особенностей распределения этих признаков, исходя из общей формулы:

$$n = \frac{t^2}{K^2} = \frac{t^2 \sigma^2}{\Delta^2}.$$

Квадрат генеральной сигмы для долей равен произведению доли на ее дополнение до единицы:

$$\sigma^2 = p(1-p),$$

поэтому

$$K^2 = \frac{\Delta^2}{\sigma^2} = \frac{\Delta^2}{p(1-p)}; n = \frac{t^2 p(1-p)}{\Delta^2}.$$

$K$  – нормированная погрешность.

Ориентировочный способ определения достаточной численности выборки заключается в том, что  $\sigma^2$  приравнивается к наибольшей вероятности генеральных долей:  $\sigma^2 = 0,5 \cdot 0,5 = 0,25$ , при  $t = 1,96$  (при анализе количественных признаков).

Таблица 6.2 (алгоритм)

**Определение достаточной численности выборки  $\hat{n}$**

I. Если $K < 0,2$ , $\hat{n} = \frac{t^2}{K^2}$ .				
II. Если $K \geq 0,2$ , $\hat{n} = f$ – определяется по математической таблице 7.				
$t$ – показатель надежности для четырех порогов вероятности ( $B$ ) безошибочных прогнозов	Пороги	$B$	$t$	$t^2$
	нулевой (0)	0,90	1,645	2,706
	первый (1)	0,95	1,960	3,842
	второй (2)	0,99	2,576	6,636
	третий (3)	0,999	3,291	10,831
$K = \frac{\Delta}{\sigma}$ – показатель точности, нормированная погрешность. $\Delta$ – абсолютная максимально допустимая погрешность при оценке генерального параметра по выборочным данным. Определяется на основе имеющихся сведений с учетом изученности предмета исследования и ответственности возможных результатов биологического анализа. $\sigma$ – генеральная сигма, примерное значение которой намечается на основе имеющихся данных и предположений следующими способами: 1. При изучении средних величин: $\bar{\sigma} = \sum \sigma_i$ { на основе предыдущих исследований $r$ – число усредняемых сигм, $\bar{\sigma} = \frac{\max - \min}{5 \text{ или } 6}$ { по наблюдаемым или предполагаемым лимитам в генеральной совокупности, 2. При изучении долей: $\sigma = \sqrt{PQ}$ { где $P$ – предполагаемая доля в генеральной совокупности, $Q = 1 - P$ , $\sigma_p = 0,5$ { если нет никаких предположений о величине генеральной доли.				

**Пример.** При исследовании зараженности плодов арбузов антракнозом предполагается, что поражена половина урожая. Уточнить процент зараженности плодов с гарантией погрешности не более 5 %:  $\Delta = 0,05$ , вероятностью безошибочного прогноза:  $t = 1,96$  (I порог).

$$\hat{n} = \frac{(1,96)^2 \cdot 0,5 \cdot 0,5}{(0,05)^2} = \frac{3,84 \cdot 0,25}{0,0025} = 384.$$

*Ответ:* необходимо рандомизированно исследовать 384 плода со всей площади.



Таблица 6.3

**Достаточная численность выборки при изучении качественных признаков**

$\Delta$	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	0,10	0,20	0,3
$\hat{n}$	10000	2500	1111	625	400	278	204	157	124	100	25	12

Примечание:  $\Delta$  – допустимая погрешность в долях.

Для предварительного расчета  $n$  требуется установить:

- допустимую абсолютную погрешность  $\Delta$  (в долях);
- значение генеральной доли,  $p$ ;
- критерий надежности  $t$  (Стьюдента).

Невозможно устанавливать заранее объем выборки, при котором было бы гарантировано получение достоверной разности средних или долей. Достоверность разности обусловлена не только объемом выборки, но и степенью варьирования дат, величиной выборочной разности. Если генеральные совокупности не различаются по величине среднего показателя, то увеличение выборок из таких совокупностей приведет к уменьшению достоверности выборочной разности. Выборочная разность будет стремиться к генеральной разности, равной нулю. И, наоборот, при очень малом объеме выборок можно получить вполне достоверную разность, если варьирование дат будет незначительным, а сама разность достаточно велика.

## ГЛАВА 7. СТАТИСТИЧЕСКИЙ АНАЛИЗ ВАРИАЦИИ КАЧЕСТВЕННЫХ ПРИЗНАКОВ

Биологу-исследователю нередко приходится анализировать различия в совокупностях по качественным признакам, таким, как наличие и отсутствие хромосомных перестроек в клетках, наличие или отсутствие морфологических признаков, расщепление при различных типах скрещиваний по окраске цветков (или семян) у растений, по окраске меха у пушных животных. При анализе вариации по качественным признакам исходными данными являются следующие:

- а) абсолютная численность объектов каждой группы:  $p^0, p^1, p^2, p^3 \dots$ ;
- б) доля объектов каждой группы в общем объеме совокупности (в долях единицы или процентах).

### **7.1. Статистический анализ при альтернативной вариации**

Альтернативная вариация – частный случай качественной вариации, при которой совокупность состоит из двух групп: одна группа имеет признак, другая (другие) не имеет его.

Численность I группы –  $a$ .

Численность II группы –  $b$ .

$n$  – объем всей совокупности.

Доля особей, имеющих признак  $a$ :  $p = \frac{a}{n}$

Доля особей, имеющих признак  $b$ :  $q = \frac{b}{n}$

При  $p + q = 1$ ,  $q = 1 - p$ .

Совокупность из трех или четырех групп обычно рассматривают как две альтернативные группы.

*Среднее арифметическое и среднее квадратическое отклонение  
при альтернативной вариации*

А. Среднее арифметическое:  $M = \bar{X} = p = \frac{a}{n}$  – это доля особей определенного класса в общей совокупности.

*Вывод:* относительная доля в совокупности особей, имеющих данный признак, соответствует  $\bar{X}$  для количественной вариации.

Б. Среднее квадратическое отклонение. Этот показатель можно вычислить по обычным формулам количественной вариации:

$$\sigma_p = S_p = \pm \sqrt{\frac{p(1-p)}{n}} = \pm \sqrt{\frac{pq}{n}}.$$

В. Средняя ошибка среднего арифметического ( $m_p$ ). Ошибка репрезентативности возникает, поскольку уровень вариация признака изучается на основе выборочной совокупности. Значения полученных долей, определенные для ряда выборочных совокупностей ( $n_1, n_2, n_3 \dots$ ), будут колебаться вокруг долей генеральной средней, т.е. средней арифметической генеральной совокупности. Мерой этих колебаний является средняя (или статистическая) ошибка:  $m_p = \frac{\sigma_p}{\sqrt{n}}$ .

Ошибка выборочной доли:

$$m_p = \sqrt{\frac{p q}{n}},$$

или (в %):

$$m_p = \sqrt{\frac{p(100-p)}{n-1}}.$$

#### Оценка генеральной доли

**Пример.** На животных испытывался новый лечебный препарат, гарантирующий выздоровление в 90 % случаев. При проверке препарата из 40 испытуемых животных выздоровело 36.

Каков прогноз действия препарата?

Генеральная доля  $P = 0,90$ ,  $Q = 0,1$ .

Выборочная доля  $p = \frac{36}{40} = 0,90$ .

Ошибка выборочной доли:  $m_p = \pm \sqrt{\frac{pq}{n}} = \pm \sqrt{\frac{0,9 \cdot 0,1}{40}} = \pm 0,05$ .

Погрешность:  $\Delta = tm = 2,7 \cdot 0,05 = 0,14$ .

Доверительные границы генеральных параметров:  $= 0,90 \pm 0,14$ .

Гарантированный минимум:  $M - \Delta = 0,90 - 0,14 = 0,76$  (76 %).

Возможный максимум:  $0,90 + 0,14 = 1,04 = 104$  %. Оказалось, что гарантировать можно выздоровление не 90 %, а 76 % особей.

#### Оценка разности долей

$p = \frac{a}{n} = p_1$  – отношение числа объектов с признаком ( $a$ ) к объему группы ( $n$ ).  $q = 1 - p$  – доля объектов без указанного признака (альтернативная доля)  $= p_2$ ;  $m_p = \sqrt{\frac{p q}{n-1}}$  – ошибка генеральной доли;  $t$  – стандартное значение критерия Стьюдента.

## Оценка разности между выборочной и генеральной долями

1. Проверка гипотезы о принадлежности изучаемой выборки ( $p$ ) к определенной генеральной совокупности ( $P$ ).
2. Проверка гипотезы о величине генеральной доли ( $P$ ) по результатам выборочного исследования ( $p$ ).

$$t(p-P) = \frac{d}{md} \geq t_{\alpha} (\sqrt{V} = n-1)$$

$p, P$  — выборочная и генеральная доли

$d = p - P$  — разность между выборочной и генеральной долями

$$md = mp = \sqrt{\frac{PQ}{n}} \quad \left\{ \begin{array}{l} \text{ошибка разности между выборочной и генеральной} \\ \text{долями, равная ошибке выборочной доли,} \\ \text{определяемой на основе известных или предполагаемых} \\ \text{генеральных долей } P \text{ и } Q = 1 - P \end{array} \right.$$

$n$  — объем выборки

Пример 1. Впервые исследованная группа объемом  $n = 50$  содержала 35 плюсовых объектов (имеющих изучаемый признак),  $p = 0,70$ . Проверяется гипотеза о принадлежности этой группы к генеральной совокупности, в которой таких объектов обычно содержится 50 %,  $P = 0,50$

$$t_{p-P} = \frac{0,7 - 0,5}{\sqrt{\frac{0,5 \cdot 0,5}{50}}} = \frac{0,20}{0,07} = 2,9; \quad \sqrt{V} = 50 - 1 = 49; \quad t_{\alpha} = \{2,0 - 2,7 - 3,5\}$$

Вывод: разность достоверна с вероятностью  $\beta > 0,99$ ; ответ отрицателен: изученная группа не может принадлежать к этой генеральной совокупности.

Пример 2. Предложена гипотеза: в генеральной совокупности плюсовых объектов должно содержаться 75 %,  $P = 0,75$ . Проверка по выборке, в которой при  $n = 100$  плюсовых объектов оказалось 70 ( $p = 0,70$ ), показала:

$$t_{p-P} = \frac{0,75 - 0,70}{\sqrt{\frac{0,75 \cdot 0,25}{100}}} = 1,2; \quad \sqrt{V} = 99; \quad t_{\alpha} = \{2,0 - 2,6 - 3,4\}$$

Вывод: разность явно недостоверна. Ответ положителен: гипотеза не опровергнута и может считаться правильной до тех пор, пока не будет опровергнута или заменена более точной гипотезой.

Для  $t_{p-P}$   
пороги  
вероятности

$\left\{ \begin{array}{l} \beta_1 \geq 0,95 \text{ при большой (!)} \\ \beta_2 \geq 0,99 \text{ при обычной} \\ \beta_3 \geq 0,999 \text{ при малой (!)} \end{array} \right\}$	ответственности.
--	------------------

**Оценка разности выборочных долей  
критерий Стьюдента**

$$t_d = \frac{d}{m_d} = \frac{p_1 - p_2}{\sqrt{m_1^2 + m_2^2}} \geq t_{st} \{V_d = n_1 + n_2 - 2\}$$

$$\beta_1 = 0,95; \quad \beta_2 = 0,99; \quad \beta_3 = 0,999$$

$p_1, p_2$  — сравниваемые доли

$$p = \frac{A}{n} \begin{cases} n — \text{объем группы} \\ A — \text{число объектов с признаками} \end{cases}$$

$m_1^2, m_2^2$  — квадраты ошибок долей

$$m = \sqrt{\frac{p(1-p)}{n-1}}$$

$$m^2 = \frac{p(1-p)}{n-1}$$

$t_{st}$  — стандартное значение критерия Стьюдента

(табл. IV) для числа степеней свободы

$V_d = n_1 + n_2 - 2$  и трех порогов вероятности безошибочных прогнозов:

0,95; 0,99; 0,999.

Пример:

$$n_1 = 100; \quad A = 40; \quad p_1 = \frac{40}{100} = 0,4; \quad m_1^2 = \frac{0,4 \cdot 0,6}{99} = 0,0024$$

$$n_2 = 200; \quad A = 100; \quad p_2 = \frac{100}{200} = 0,5; \quad m_2^2 = \frac{0,5 \cdot 0,5}{199} = 0,0013$$

$$d = 0,1; \quad m_d = 0,061$$

$$m_d^2 = 0,0037$$

$$t_d = \frac{0,100}{0,061} = 1,6$$

$$m_d = 0,061$$

$$V_d = 100 + 200 - 2 = 298$$

$$t_{st} = \{2,0 - 2,6 - 3,3\}$$

**Пример.** Из тысячи цыплят, получавших кормовые дрожжи, заболело рахитом 10, а из 2000 цыплят, не получавших дрожжи, — 80. Какова эффективность добавки в корм кормовых дрожжей?

$$p_1 = \frac{10}{1000} = 0,01; \quad p_2 = \frac{80}{2000} = 0,04.$$

$$t_d = \pm \frac{p_1 - p_2}{\sqrt{m_1^2 + m_2^2}} \geq t_{st}; \quad v_d = \{n_1 + n_2 - 2\}.$$

$$m_1^2 = \frac{0,1 \cdot 0,99}{999} = \pm 0,0000099,$$

$$m_2^2 = \frac{0,4 \cdot 0,96}{1999} = \pm 0,0000192,$$

$$d = 0,1 - 0,4 = -0,03,$$

$$md = \sqrt{m_1^2 + m_2^2} = \pm 0,005.$$

$$v = \infty,$$

$$t_d = \frac{d}{md} = \frac{0,03}{0,005} = 6,0; \quad t_{st} = \{2,0 - 2,6 - 3,3\}.$$

**Вывод:** разность долей достоверна по III порогу вероятности безошибочного прогноза.

## 7.2. Дискретные переменные величины

При статистическом анализе вариации по качественным признакам исследователь имеет дело с дискретными переменными величинами. Эти переменные можно просто занумеровать, поскольку возможные результаты распадаются на отдельные, различимые классы (мальчик или девочка при рождении ребенка, голубая или красная пыльца гибрида при окрашивании йодом, желтые или зеленые семена у гороха при расщеплении гибридов). В случае дискретных переменных существует лишь конечное число альтернатив (небольшое число классов). С другой стороны, количественные признаки: вес (масса), высота, умственное развитие и другие описываются не дискретными, а непрерывными переменными. Измерения, относящиеся к некоторой конечной выборке, характеризуются частотой (вероятностью), а результаты идеальной, бесконечно большой совокупностью измерений, описываются параметрами.

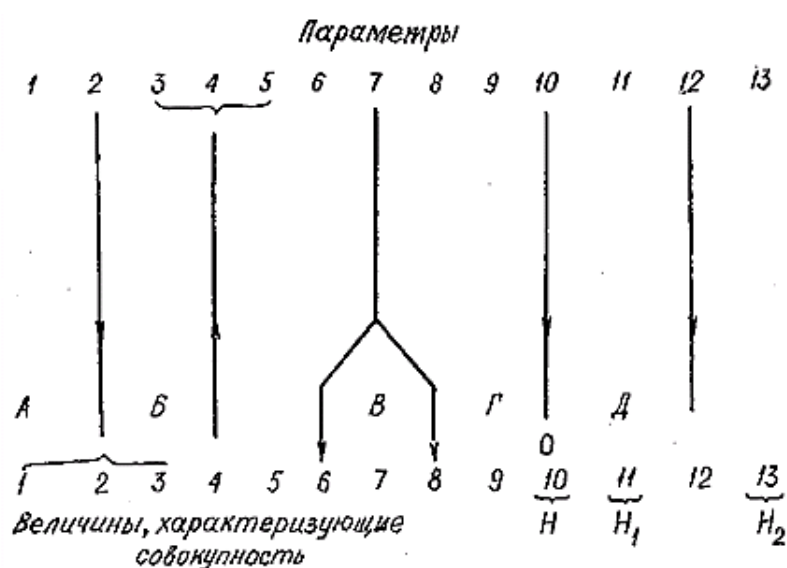


Рис. 7.1. Биометрические действия, обсуждаемые в разделе «Дискретные переменные»: Н – наблюдаемое; О – ожидаемое; стрелки – направление предсказаний (прогнозирования)

При анализе качественных признаков могут возникать различные задачи:

- по данному параметру предсказать интервал частот, ожидаемых для некоторой выборки;
- по заданной частоте определить интервал параметров, которые при выборке могли привести к наблюдаемой частоте,
- определить вероятности, т.е. параметры того, что в выборках, сделанных из некоторой идеальной совокупности, будут включаться различные альтернативы (рис. 7.1В).
- наблюдаемые (Н) в действительности (рис. 7.1Г, Д);
- сравнить две группы статистических совокупностей (рис. 7.1Н-1, Н-2).

### **7.2.1. Определение ожидаемого разнообразия частот по заданному параметру, зависящему от одной переменной**

Гипотеза формулируется в терминах вероятности возникновения исследуемого события. Необходимо знать, какие результаты получатся при проверке этой гипотезы (рис. 7.1А). Так, например, при анализе расщепления по полу у дрозофилы любая особь, взятая случайно из репрезентативной выборки с равной вероятностью, может оказаться самцом (♂) или самкой (♀). В биометрии и математике часто используется понятие «вероятность успеха». Будем считать «успехом» самца дрозофилы, альтернативное событие – самка. Гипотеза: вероятность успеха ( $p$ ) равна 50 %, или 0,5 от общего числа. В данном случае возможны только две альтернативы: успех или неудача. Вероятность неудачи ( $1 - p$ ). Для описания всех возможных результатов, достаточно одной переменной – вероятность успеха  $f$  – эмпирическая частота. Если число успехов  $a$ , число испытаний –  $n$ , то  $f = \frac{a}{n}$ . Если собрано много относительно больших выборок, то какие будут получены значения  $f$ ?

Значения  $f$  определяются величиной среднего квадратического отклонения:  $\sigma_f = S_f = \pm \sqrt{\frac{f(1-p)}{n}}$ . Если выборка  $\geq 25$ , то 95 % полученных значений  $f$  будут находиться между  $(p - 1,96\sigma)$  и  $(p + 1,96\sigma)$ . Если заявить, что значения  $f$  будут находиться в указанном доверительном интервале, то вероятность ошибки составит 5 %.

В примере с расщеплением по полу у дрозофилы ( $p = 0,5$ ) при  $n = 100$ ,  $\alpha = 0,05$  следует ожидать, что в 95 % случаев  $f$  будет лежать в интервале 0,4–0,6 ( $1,96\sigma = 0,1$ ). Если сделать много выборок при  $n = 100$ , то можно утверждать, что 95 % всех  $f$  попадут в интервал 0,4–0,6. Для одной выборки с  $n = 100$  утверждение, что  $f$  будет лежать в интервале 0,4–0,6, имело бы 95 шансов из 100 быть верным и 5 шансов – не верным. Если бы  $N$  было бесконечно велико,  $f$  равнялось бы  $p$ .

*Определение ожидаемого «разброса» параметров по заданной частоте, зависящей от одной переменной*

Если бы не была известна вероятность исследуемого события, то можно было определить возможные значения  $p$  из полученной в опыте выборки. Зная частоту  $f$ , можно оценить неизвестный параметр  $p$ . Допустим, что из 100 исследованных клеток 30 вступили в митоз. Величина  $p = 0,3$  представляет собою единичную частоту (наилучшая единичная оценка  $p$  есть  $f$ ). Оценка  $p = 0,3$  сделана по одному параметру, который может быть, равен, например, 0,31, 0,29... Разнообразие (разброс) значений  $p$  при  $f = 0,3$  и  $N = 100$  можно определить, вычислив величину  $\sigma_f$ :

$$\rightarrow \sigma_f = S_f = \pm \sqrt{\frac{f(1-f)}{n}}.$$

Значения  $p$ , находящиеся между  $(f - 1,96\sigma_f)$  и  $(1 + 1,96\sigma_f)$  образуют 95-процентный интервал  $p$ , поскольку в 95 % случаев следует ожидать, что данная выборка имеет  $p$  в этом интервале. Это утверждение может быть ошибочным лишь в 5 случаях из 100. Следовательно,  $\sigma_f = \pm 0,05(\sim)$ , и 95-процентный доверительный интервал  $p$  заключен примерно между 0,20 и 0,40. Следовательно, вывод о том, что  $p$  должно лежать между 0,20 и 0,40, будет неверно лишь в 5 % случаев.

**7.2.2. Определение ожидаемых относительных вероятностей по заданным параметрам, зависящим от одной переменной**

При анализе многих вероятностных (стохастических) процессов, не обращаясь к опыту, можно высказать априорно гипотезу о вероятности успеха. Однако иногда успех может осуществляться двумя и более различными способами. Чему равна в таких случаях полная вероятность успеха? Для ответа на этот вопрос следует обратиться к рассмотренным ранее теоремам сложения и умножения вероятностей (гл. 3).

*Теорема сложения.* Если успех может осуществляться двумя и более различными, взаимоисключающими способами, полная вероятность успеха равна сумме индивидуальных вероятностей. Вероятность осуществления любого одного или нескольких взаимоисключающих «успехов» равна сумме индивидуальных вероятностей. Если вероятность «успеха» равна  $p$ , а вероятность неудачи  $q$ , то вероятность либо успеха, либо неудачи равна  $(p + q)$ .



Если точно известно, что данное событие должно быть либо успехом, либо неудачей, то  $p + q = 1$ ,  $p = 1 - q$  и  $q = 1 - p$ .

*Теорема умножения.* Если второе событие наступает только при осуществлении первого, и при этом наступление первого не влияет на вероятность второго, то вероятность осуществления нескольких независимых успехов равна произведению их индивидуальных вероятностей. Вспомним пример с лабиринтом с шестью развилками и шестью тупиками (гл. 3): при наличии шести развилок общая вероятность прохождения по лабиринту ( $p$ ) будет равна:

$$\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2^6} = \frac{1}{64}.$$

### 7.2.3. Сравнение наблюдаемых вероятностей (частот) с ожидаемыми.

Биномиальный критерий для параметра, зависящего от одной переменной.

**Пример 1.** Допустим, для генетического расщепления в  $F_2$  ожидается расщепление по фенотипу 1:1 ( $p = 0,50$ ). Согласно биномиальному распределению, следует ожидать, что из 6 особей будут 3 особи одного и 3 другого фенотипа, в 5 случаях из 16. Такой результат будет осуществляться наиболее часто. Допустим, что в эксперименте все шесть особей оказались одного фенотипа. Следует ли считать это наблюдение статистически незначимым и обусловленным случайным результатом? Или этот результат статистически значимый и указывает на то, что ожидаемые результаты не всегда согласуются с результатами наблюдений?

Вероятность того, что каждая отдельная особь будет принадлежать к первому типу, равна  $1/2$ , вероятность того, что она будет принадлежать ко второму типу, также равна  $1/2$ . Вероятность того, что все шесть особей будут первого типа, равна  $(1/2)^6$ , вероятность того, что все шесть особей будут второго типа, тоже равна  $(1/2)^6$ . Вероятность того, что все шесть особей будут первого, или второго типа, будет  $(1/2)^6 + (1/2)^6 = 0,03$ .

Поскольку событие, вероятность которого 0,03, должно осуществляться только в трех случаях из 100, то следует, что исходная гипотеза или верна, но осуществилась крайне маловероятная ситуация, либо не согласуется с результатами наблюдений. Отсюда следует выбрать вторую альтернативу, что гипотеза, вероятно, неверна.

*Нулевая гипотеза* проверяется следующим образом. Исходя из заданного значения параметра, вычисляют такую вероятность получения частоты, которая является крайним случаем (или вообще лежит за пределами наблюдаемых частот). Если эта вероятность мала (0,05 и меньше), то можно сде-

лать вывод, что наблюдаемые результаты не согласуются с ожидаемыми. Исходная гипотеза отвергается с уверенностью 95 % и с уровнем значимости 5 % (следовательно, существует 5 % вероятность отвергнуть гипотезу, которая в действительности верна). Если вероятность превышает 5 %, то можно считать, что наблюдения не противоречат исходной гипотезе. Следовательно, эта гипотеза приемлема. Если вероятность меньше 0,05 (0,01 и меньше), то принято считать расхождение очень большим.

**Пример 2.** В группе из 8 особей обнаружено 6 особей одного типа и 2 особи другого типа. Допустим, что теоретическое отношение равно 1: 1. Согласно нулевой гипотезе, вероятность получить в группе из 8 особей только две особи одного типа или еще меньше, является суммой следующих членов, полученных при разложении бинома  $(1/2 + 1/2)^8$ .

Вероятность: 0 особей 1 типа =  $(1/2)^8$ .

Вероятность: 1 особь 1 типа =  $8 \cdot (1/2)^8$ .

Вероятность: 2 особи I типа =  $28 \cdot (1/2)^8$ .

Вероятность: 2 особи II типа =  $28 \cdot (1/2)^8$ .

Вероятность: 1 особи II типа =  $8 \cdot (1/2)^8$ .

Вероятность: 0 особей II типа =  $(1/2)^8$ .

Складывая отдельные вероятности, находим, что полная вероятность обнаружить только две (и меньше) особи одного типа =  $\frac{74}{256} = 0,29$ .

Поскольку полная вероятность больше 0,05, полученный результат согласуется с исходной гипотезой. Следовательно, гипотеза приемлема.

#### **7.2.4. Критерий доверительного интервала для параметра, зависящего от одной переменной**

В рассмотренных примерах биномиальный критерий принимался для случаев, когда  $n < 10$ . При  $n \geq 10$  это затруднительно. Значительно удобнее пользоваться ожидаемым интервалом  $f$ . Этот интервал ряд исследователей (Гершкович, 1968, с. 539; Гланц, 1999) предлагает определить как «диаграмму чувствительности», исходя из предполагаемого параметра, зависящего от одной переменной с помощью графика (рис. 7.2).

**Пример.** Пусть  $f = 0,3$ ,  $n = 100$ . Какой вывод можно сделать относительно нулевой гипотезы при  $p = 0,5$ ? Если  $p = 0,5$  и  $n = 100$ , то 95 %  $f$  лежало бы между 0,4 и 0,6, то можно было бы принять  $p = 0,5$ . Выводы относительно значения параметра, сделанные с помощью графика (рис. 7.2), имеют 5 % уровень значимости, т.е. дают основание лишь отвергать или принимать гипотетические параметры. Выборка же реально отражает наблюдаемые факты, ее нельзя отвергать. Критерий  $\chi^2$  для параметра, зависящего от одной переменной (применим при достаточно больших объемах выборки).

Ожидаемое отношение фенотипов 1: 1,  $n = 100$ . В идеальном случае выборка содержит 50 случаев первого типа и 50 случаев второго. Если реально наблюдается 55 случаев одного типа и 45 другого, нужно, исходя из нулевой гипотезы, определить вероятность получения такого результата в выборке ( $N = 100$ ), сделанной из идеальной совокупности. Эту вероятность можно найти, суммированием соответствующих членов в разложении  $(1/2 + 1/2)^{100}$  (с помощью компьютера). Из теории вероятностей известно, что эту вероятность можно выразить через величину  $\chi^2$  следующим образом:

$$\chi^2 = \sum \frac{[(\text{наблюдаемое} - \text{ожидаемое}) - \frac{1}{2}]^2}{\text{ожидаемое}}$$

(член  $\frac{1}{2}$  – поправка Йетса).

Чтобы прочесть диаграмму (рис. 7.2) для  $\chi^2$  равного 17, для случая 7 степеней свободы, нужно двигаться вверх по линии, соответствующей величине  $\chi^2 = 17$ , до пересечения с кривой, соответствующей  $N = 7$ . Найдя эту точку на шкале вероятностей, получаем: вероятность равна 0,017. Если диаграмму перевернуть, то таким же способом можно определить вероятности для t-распределения. Приведенная вероятность есть вероятность *наибольшего по величине отклонения*.

### **7.3. Исследование вероятности распространения генов в популяции**

Генетика популяций изучает генетическое разнообразие в популяциях и закономерности изменения этого разнообразия во времени (поколения) и пространстве. Популяционная генетика включает исследование природных, экспериментальных и теоретически мыслимых (стохастических) популяций. Цель генетики популяций: исследование генетической структуры популяций и факторов изменения этой структуры. В работе С.С. Четверикова (1926) «О некоторых моментах эволюционного процесса с точки зрения современной генетики» были обоснованы новые представления о роли и динамике мутаций в природных популяциях, о роли отбора и изоляции в эволюционном процессе. По С.С. Четверикову, в условиях свободного скрещивания (*панмиксия*), вид, «.....как губка, впитывает в себя гетерозиготные геновариации, оставаясь при этом внешне однородным».

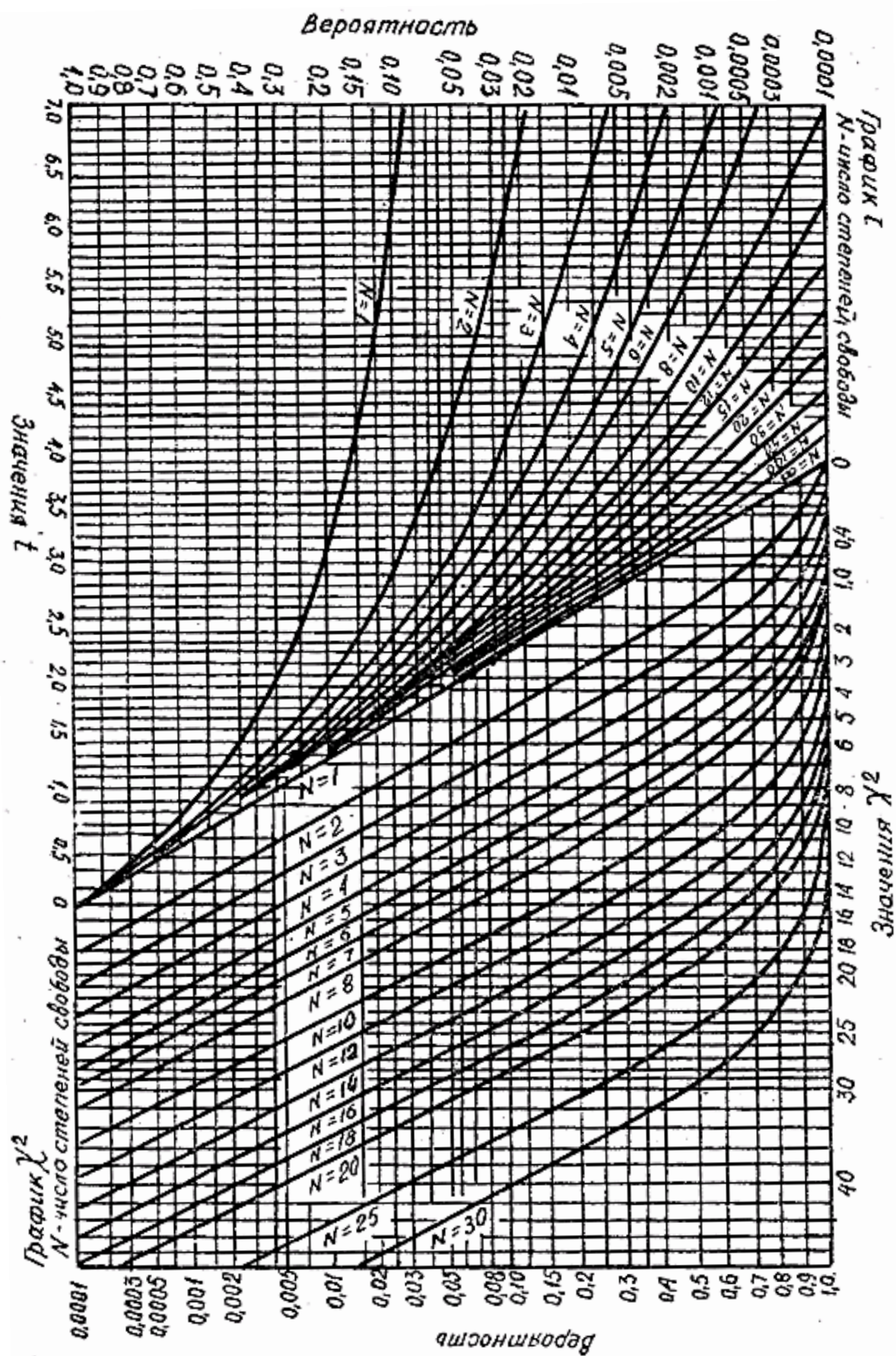


Рис. 7.2. Распределения  $\chi^2$  и  $t$

В начале 1930-х гг. ученики С.С. Четверикова (Н.В. Тимофеев-Ресовский, Б.Л. Астауров, С.М. Гершензон и др.) исследовали генетическую структуру природных популяций дрозофилы (*Drosophila*) и установили в них высокую концентрацию рецессивных мутаций. Большой вклад в изучение природных популяций и развитие генетики популяций внесли исследования А.С. Серебровского, Н.П. Дубинина, Ф. Добжанского, П.Ф. Рокицкого, Р. Фишера, С. Райта, Дж. Холдейна, Н.В. Тимофеева-Ресовского и др.

*Популяция* – это неограниченно большая группа (множество) особей вида (Н.В. Тимофеев-Ресовский), которая:

- в течение длительного (исторического) времени занимает конкретный ареал;
- находится в состоянии панмиксии, т.е. случайного скрещивания;
- не имеет внутри заметных изоляционных барьеров,
- популяция отделена от соседних групп определенными формами изоляции.

История генетики популяций и развитие математических идей и методов в этой области началась в 1908 г., когда был сформулирован закон Г. Харди – В. Вайнберга о распространении независимых генов и генотипов в панмиктической популяции. Однако еще в 1904 г. К. Пирсон пытался применить правила Менделя для анализа количественных признаков в биологических совокупностях. Расщепление любых гетерозигот по фенотипу (при полном доминировании) происходит в соответствии с коэффициентами разложения бинома Ньютона:

- Одна пара генов:  $Aa \times Aa \rightarrow (3 + 1)^1 \rightarrow AA + 2Aa + aa \rightarrow 3:1$ .
- Две пары генов:  $AaBb \times AaBb \rightarrow (3 + 1)^2 \Rightarrow 9:3:3:1$ .
- Три пары генов:

$$AaBbCc \times AaBbCc \rightarrow (3 + 1)^3 \rightarrow 27:9:9:9:3:3:3:1.$$

- При полигибридном скрещивании: степень гетерозиготности равна  $n$ , расщепление –  $(3 + 1)^n$ .

Согласно закону Харди – Вайнберга, в панмиктической популяции (популяция с преобладанием процесса свободного скрещивания) частоты двух генов одной аллельной пары ( $A$  и  $a$ ) равны соответственно  $p$  и  $q$  при ( $p + q = 1$ ).

Пусть скрещивание и репродукция по данному локусу происходят случайно. Тогда частоты аллелей будут оставаться постоянными, а относительные частоты генотипов ( $AA$ ,  $Aa$ ,  $aa$ ) будут соответственно:

$$p^2 + 2pq + q^2 = 1,$$

т.е. будут членами биномиального выражения:

$$(p + q)^2 \rightarrow p^2 AA + 2pq Aa + q^2 aa = 1.$$

Для аутосомных генов при отсутствии каких-либо факторов генетической динамики популяций это соотношение сохраняется во всех последующих поколениях. Распространение аллелей в неограниченно большой популяции при свободном скрещивании, отсутствии давления отбора и мутационного давления устанавливается на основе концентрации генов, имеющих в популяции. Концентрация генов – их относительная частота в популяции. Так, если частота гена  $A$  равна  $p$ , то частота гена  $a$  равна  $q$ , или  $(1 - p)$ .

Таблица 7.3.1

**Вероятность возникновения различных генотипов соответствует произведению частот генов в гаметах предшествующего поколения**

♂ гаметы	♀ гаметы	
	$pA$	$qa$
$pA$	$p^2AA$	$pqAa$
$qa$	$pqAa$	$q^2aa$

В связи с равномерным распределением генов между особями, у них будет образовываться  $p$  яйцеклеток с геном  $A$  и  $q$  яйцеклеток с геном  $a$ ,  $p$  спермиев с геном  $A$ ,  $q$  спермиев с геном  $a$ . Частота определенных генов в генотипах последующих поколений определяется частотой этих генов в гаметах предшествующих поколений. В итоге получаем суммарное уравнение Харди – Вайнберга:

$$p^2AA + 2pqAa + q^2aa = 1,$$

из которого следуют выводы:

- количество доминантных гомозигот ( $AA$ ) в популяции равно квадрату частоты доминантного гена ( $p^2$ );
- количество рецессивных гомозигот ( $aa$ ) в популяции равно квадрату частоты рецессивного гена ( $q^2$ );
- количество гетерозигот равно удвоенному произведению частот обоих этих генов ( $2pq$ ).

Согласно закону Харди – Вайнберга, исходная пропорция генов данной аллельной пары в панмиктической популяции может быть любой, но она остается постоянной в последующих поколениях. Если в популяции распределяется одна пара аллелей ( $A - a$ ), возможны три генотипа  $AA$ ,  $Aa$ ,  $aa$ . Допустим, что эти генотипы находятся в данном поколении в соотношении:

$$AA + 2Aa + aa \text{ (или } 1/4 AA + 1/2 Aa + 1/4 aa).$$

При свободном скрещивании особей в потомстве получается то же соотношение генотипов (табл. 7.3.2). Таким образом, распределение генотипов в следующем поколении будет таким же, как в предыдущем, и это будет справедливо для всех последующих поколений. Пользуясь уравнением Харди – Вайнберга, можно исследовать генетическую структуру популяции, т.е. часто-



ту генов и генотипов данной аллельной пары, если, например, с помощью демографической статистики определить частоту рецессивных гомозигот в популяции. Пользуясь формулой Харди – Вайнберга, можно рассчитать относительную частоту генов и генотипов в популяции, т.е. генетическую структуру популяции по данной аллельной паре.

Начиная с частоты распространения в популяции рецессивных гомозигот ( $aa$ ), можно определить частоту (вероятность) рецессивного аллеля.

Таблица 7.3.2

Частота генов в панмиктической популяции  $\frac{1}{4}AA + \frac{1}{2}Aa + \frac{1}{4}aa$

	Возможное скрещивание		Частота исходных генотипов	Частота генотипов в потомстве		
	♀	♂				
1.	AA	AA	$1/4 \times 1/4$	$1/16AA$		
2.	AA	Aa	$1/4 \times 1/2$	$1/16AA +$	$1/16Aa$	
3.	AA	aa	$1/4 \times 1/4$		$1/16Aa$	
4.	Aa	AA	$1/2 \times 1/4$	$1/16AA +$	$1/16Aa$	
5.	Aa	Aa	$1/2 \times 1/2$	$1/16AA +$	$1/8Aa$	$1/16aa$
6.	Aa	aa	$1/2 \times 1/4$		$1/16Aa +$	$1/16aa$
7.	aa	AA	$1/4 \times 1/4$		$1/16Aa$	
8.	aa	Aa	$1/4 \times 1/2$		$1/16Aa +$	$1/16aa$
9.	aa	aa	$1/4 \times 1/4$			$1/16aa$
Итого:				$4/16AA +$	$8/16Aa +$	$4/16aa$

Частоты аллелей в следующем поколении оказались равными частотам в исходном поколении, значит, частоты генотипов во втором поколении окажутся такими же, как в предыдущем.

**Пример.** Частота рецессивных гомозигот ( $aa$ ) по гену альбинизма ( $a$ ) в популяциях Европы =  $1/20000$ .

Какова генетическая структура популяции по данному гену? Исходя из общей формулы закона Харди – Вайнберга  $p^2AA + 2pqAa + q^2aa = 1$ , можно определить частоту рецессивного аллеля ( $a$ ):

Частота рецессивного аллеля  $a = \sqrt{q^2} = \sqrt{(1/20000)} = 1/140 = 0,007 = 0,7 \%$ . Концентрация гена альбинизма  $a$ :  $q = 0,7 \%$ .

Частота доминантного аллеля  $A = p$ .

При  $p + q = 1$ ,  $p = 1 - q = 1 - 0,007 = 0,993 = 99,3 \%$ .

Частота нормального аллеля ( $A$ ) =  $p = 99,3 \%$ .

*Расчет генетической структуры популяции*

Частота доминантных гомозигот:

$(AA) = p^2 = 0,9932 = 0,981 = 98,1 \%$ .

Частота рецессивных гомозигот:

$$(aa) = q^2 = \frac{1}{20000} \cdot 100 \% = 0,5 \%$$

Частота (вероятность) гетерозигот ( $Aa$ ) в популяции:

$$Aa = 2pq = 2 \cdot 99,3 \% \cdot 0,7 \% = 1,4 \%$$

Частота гетерозиготного носительства:

$$(Aa) = 2pq = 1,4 \%$$

Поскольку большинство наследственных болезней у человека имеет рецессивный характер, то, отталкиваясь от эмпирической частоты рецессивных гомозигот, можно рассчитать частоту гетерозиготного носительства ( $Aa$ ) по любому аутосомному гену описанным способом. Это важно для медико-генетического консультирования. Частота гетерозиготного носительства в популяциях оказывается достаточно высокой. Примеры частоты некоторых гомозигот ( $aa$ ) =  $q^2$  и гетерозигот  $Aa = 2pq$  по некоторым наследственным заболеваниям (по Lenz):

Наследственная патология	Частота рецессивных гомозигот $q^2(aa)$	Вероятность гетерозигот $2pq (Aa)$
Псевдохолинэстераза	1: 2500	1/25
Адреногенитальный синдром	1: 4900	1/35
Фенилкетонурия	1: 10000	1/50
Цистинурия	1: 40000	1/100

Из закона Харда – Вайнберга следует следующий вывод: если частоты аллелей у особей мужского и женского пола одинаковы, то при любом исходном соотношении частот генотипов равновесные частоты генотипов в каждом локусе достигаются за одно поколение. Если частоты аллелей у представителей разного пола различны, то для аутосомных локусов они становятся одинаковыми в следующем поколении, поскольку и самцы, и самки получают половину своих генов от отца и половину от матери. Следовательно, в этом случае равновесные частоты генотипов достигаются за два поколения. Закон Харди – Вайнберга описывает поведение системы во времени при указанных выше условиях, т.е. при соблюдении всех перечисленных условий будут наблюдаться соответствующие соотношения частот аллелей и генотипов. Однако отклонение от соотношения Харди – Вайнберга всегда свидетельствует о процессах, происходящих в популяции. Закон Харди – Вайнберга – элементарная математическая модель генетической структуры популяции. В реально существующих популяциях всегда действуют факторы, изменяющие в чреде поколений частоты аллелей и генотипов. К ним относятся отсутствие панмиксии (случайного скрещивания), ограниченная численность популяции, мутации, миграция, дрейф генов, генетическая изоляция, отбор.



## ГЛАВА 8. КОРРЕЛЯЦИЯ – СОПРЯЖЕННОСТЬ ПРИЗНАКОВ, ИХ ВЗАИМОСВЯЗЬ

При изучении изменчивости нередко наблюдается, что изменения в одном признаке связаны с изменением в другом. Корреляционные зависимости наблюдаются между многими признаками организмов, морфологическими, физиологическими, между различными биологическими процессами. Так, с увеличением дозы ионизирующего облучения увеличивается количество мутаций. Чем больше детенышей в помете многоплодных животных, тем меньше вес (масса) каждого из них. Так, у растений гороха с серыми семенами всегда пурпурные цветки, у гороха с прозрачной (белой) семенной кожурой венчик цветка всегда белый. Между перечисленными признаками наблюдается полная положительная корреляция. При изучении большинства количественных признаков полное соответствие в их развитии встречается редко. Корреляция большинства количественных признаков может быть разной. Исследование корреляционных зависимостей имеет большое практическое значение, поэтому существует необходимость в количественном измерении корреляции. Сущность корреляции заключается в сопряженности вариаций признаков. Для того чтобы установить наличие корреляционной зависимости и ее степень, необходимо установить, насколько параллельно идет вариация по этим признакам. Предварительным способом сравнения вариации признаков является построение графиков, на которых выражена кривыми вариация признаков у особей данной совокупности. Однако этот способ не дает величины корреляционной зависимости. Непосредственное сравнение вариации признаков затрудняется тем, что они, как правило, выражены в разных измерениях. Поэтому при изучении корреляции используется нормированное отклонение  $t$ . Нормированное отклонение  $t$  – отклонение тех или иных вариантов от их средней арифметической, выраженное в количестве сигм ( $\delta$ ), долях среднего квадратического отклонения. Выражая отклонения отдельных дат от средних арифметических по обоим признакам одновременно, можно объективно сопоставить вариацию по обоим признакам. Биологические объекты (растения, животные, микроорганизмы) развиваются под контролем генотипа и действием бесконечно большого числа факторов, которые по-разному модифицируют развитие признаков. Взаимосвязи между признаками также сильно варьирует, поэтому каждому значению одного признака соответствует не одно значение второго признака, а целое распределение значений второго признака. Учитывая исключительную значимость этого свойства, оно по своему описывается исследователями различных биологических наук. Так, Р. Ригер и А. Михаэлис (1967) определяют это свойство живых организмов, как фенотипическое изменение (англ. *correlated phenotypic variation*) двух или

нескольких различных или же обозначенных разными символами фенотипических признаков. Это не означает, что оба признака всегда изменяются в одинаковой мере или в том же направлении, а лишь значительную вероятность того, что изменение у особи одного из признаков будет сопровождаться определенными изменениями коррелированного с ним признака. Сама по себе эта корреляция представляет материал для эволюции и селекции лишь в том случае, если в основе ее лежит наследственность, т.е. если корреляция сама связана с механизмами наследственности. *Коэффициент корреляции* (нем. *Korrelations koefizient*; англ. *correlation coefficient*) – мера зависимости двух варьирующих свойств. Например, с помощью коэффициента корреляции связь между размерами и массой тела может быть вычислена по формуле Бравэ:

$$r = \frac{S(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sigma_{x_i}\sigma_{y_i}},$$

где  $(x_i - \bar{x})$  – отклонение каждого варианта от средней по одному свойству;  $(y_i - \bar{y})$  – отклонение каждого варианта от средней по другому свойству;  $S(x_i - \bar{x}) \times (y_i - \bar{y})$  – сумма произведений отдельных отклонений;  $\sigma_{x_i}\sigma_{y_i}$  – произведение средних квадратических отклонений коррелирующих признаков;  $(n - 1)$  – число степеней свободы.

Величина  $r$  может колебаться от  $r = -1$  через  $r = 0$  до  $r = +1$ . В данном случае при  $r = 0$  будет означать отсутствие какой-либо связи, а  $r = +1$  указывать на тесную корреляцию между исследуемыми величинами. Точно так же может быть исследована связь между признаками родителей и потомства. При тесной корреляции (высокий коэффициент корреляции) можно сделать вывод о степени наследуемости соответствующих признаков. Вычисление коэффициента корреляции по формуле Бравэ обычно применяется при большом количестве вариантов, но оно менее точно, чем вычисление прямым способом по формуле:

$$r = \frac{S(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}.$$

Необходимо установить достоверность коэффициента корреляции с помощью таблиц (приложение) или путем вычисления ошибки коэффициента корреляции  $S_r$  (или  $m_r$ ) и определения  $t$ :

$$t = \left( S_r = \frac{\sqrt{1 - r^2}}{\sqrt{n - 2}} \right) \text{ и определения } t \left( t = \frac{r}{S_r} \right).$$

Значение  $t = 2$  обеспечивает достоверность коэффициента корреляции с вероятностью  $p = 0,95$ .

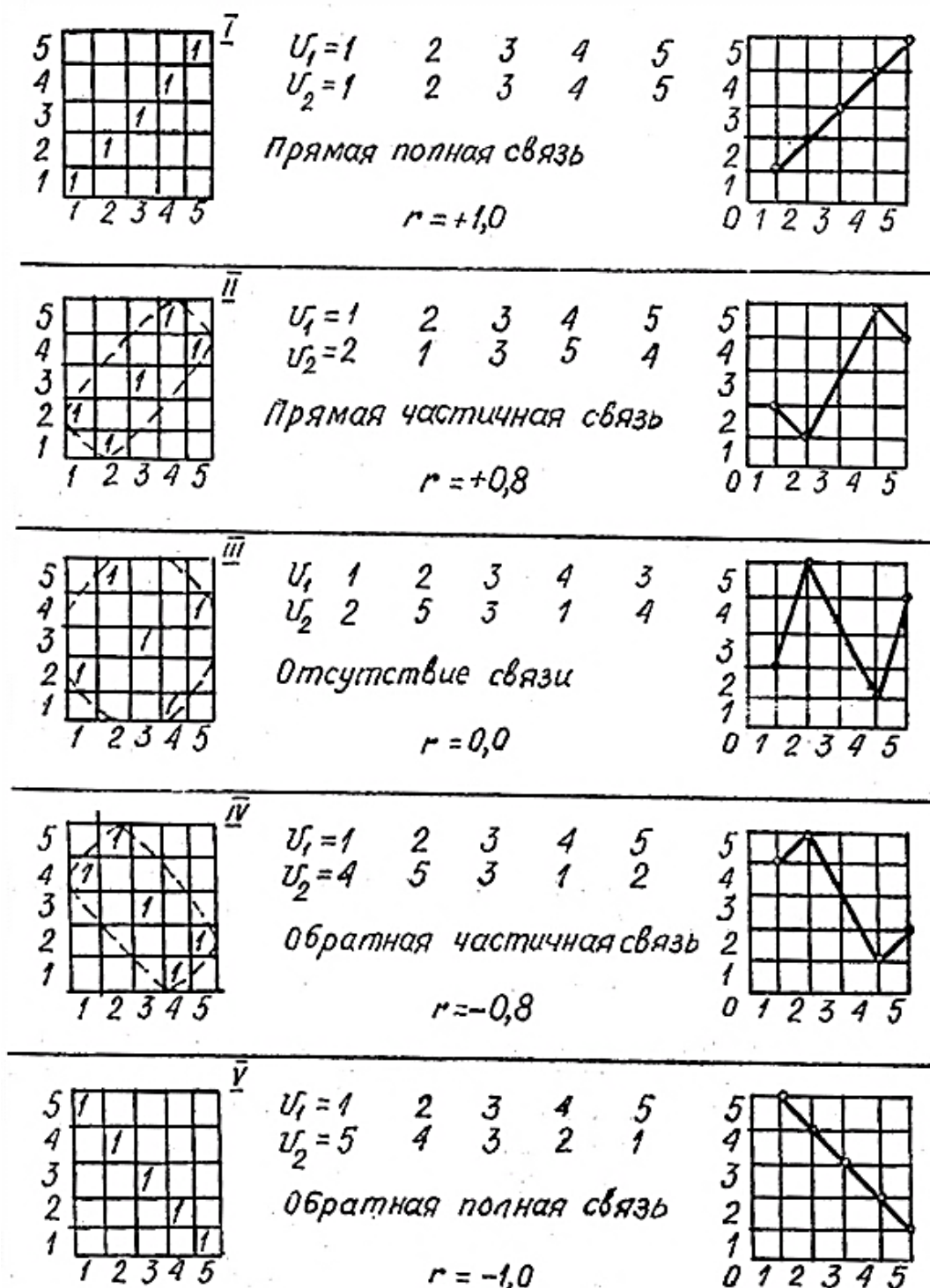


Рис. 8.1. Схемы прямолинейных корреляционных связей

Изобразить корреляционную взаимосвязь двух признаков можно различными способами (рис. 8.1):

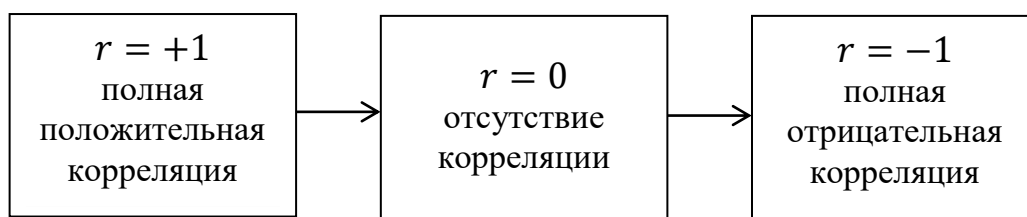
- с помощью корреляционного ряда, включающего пары соответствующих значений двух взаимосвязанных признаков;
- с помощью корреляционной решетки (на рис. 8.1 слева);

- с помощью линий регрессии (на рис. 8.1 справа показаны схемы линий регрессии).

Существуют следующие типы корреляции:

1. Прямая (положительная) взаимосвязь, при которой с увеличением значения одного признака, значение другого также увеличивается.

2. Обратная (отрицательная) взаимосвязь, при которой с увеличением значений одного признака значения другого систематически уменьшается. Значение коэффициента корреляции возможно в пределах от  $r = +1$  до  $r = -1$ .



Коэффициент корреляции дает значение только степени взаимосвязи признаков, но и направление прямолинейной зависимости. Основная формула (коэффициент корреляции К. Пирсона), вскрывающая сущность коэффициента корреляции:

$$r = \frac{\sum x_1 \cdot x_2}{\gamma},$$

$r$  – коэффициент корреляции;  $\sum x_1 \cdot x_2$  – сумма произведений нормированных отклонения дат по первому и второму признаку;  $\gamma$  – число степеней свободы, равное числу сравниваемых пар без одной.

Нормированное отклонение служит универсальной и неименованной мерой развития признака:

$$x(t) = \frac{V - M}{\delta}.$$

На основе исходной формулы коэффициента корреляции для удобства разработаны различные рабочие формулы, дающие одинаковые результаты. Наиболее часто в биологических работах используются две формулы:

$$r = \frac{\sum V_1 \cdot V_2 - \frac{\sum V_1 \sum V_2}{n}}{\sqrt{C_1 C_2}}, r = \frac{C_1 + C_2 - C_d}{2\sqrt{C_1 C_2}}.$$

По своему содержанию они не отличаются от формулы, приведенной выше. Применение этих формул показано в таблице 8.3:  $V_1, V_2$  – даты первого и второго признаков;  $n$  – число сравниваемых пар;  $S$  – дисперсии по первому и второму признакам и разности между датами сравниваемых признаков.

Рассмотрим на конкретном примере методику изучения корреляционной взаимосвязи признаков **с использованием корреляционной решетки и метода условной средней**:

$$r = \frac{\sum p\alpha_x \cdot p\alpha_y - nb_x b_y}{n\delta_x \delta_y},$$

где  $\alpha_x, \alpha_y$  – отклонения дат от условной средней по ряду  $x$  и  $y$ ;  $b_x, b_y$  – поправки, вводимые вследствие того, что отклонения находятся не от среднего арифметического, а от условной средней.

При исследовании корреляции между двумя признаками группа особей исследуется параллельно по обоим признакам и их пары заносятся в журнал. На этой основе составляется *корреляционная решетка*, которая наглядно демонстрирует расположение и разброс парных значений исследуемых признаков.

Рассмотрим правила и последовательность действий составления корреляционной решетки с *использованием метода «условной средней»*.

На двух сторонах квадрата (табл. 8.1, сверху и сбоку слева) наносим значения классов обоих исследуемых рядов, предварительно разбив оба ряда на удобное количество классов. Результаты парных значений исследуемых взаимосвязанных признаков разносятся в отдельные клетки. Каждая дата заносится в квадрат на месте пересечения значений двух признаков. Сумма всех частот в горизонтальных строчках пишется справа (ряд  $x$ ). Сумма частот ( $p$ ) в вертикальных строчках пишется внизу (ряд  $y$ ). Справа внизу в угловой клетке пишется сумма всех значений дат. По формуле требуется найти сумму взвешенных отклонений одновременно по обоим признакам.

В каждом ряду выбираем класс, величину которого принимаем за условную среднюю, обозначив их  $A$  и  $B$ .

Находим отклонение каждого класса от значения  $A$  по ряду  $x$  ( $\alpha_x$ ) и от значения  $B$  по признаку  $y$  ( $\alpha_y$ ), записываем их соответственно внизу и справа.

Находим  $p \cdot \alpha_x \cdot \alpha_y$  для каждого квадрата, учитывая знак произведения. Находим произведение  $p \cdot \alpha_x \cdot \alpha_y$ , перемножая частоты встречаемости дат в общем квадрате на отклонения по ряду  $x$  и ряду  $y$ .

Все значения  $p \cdot \alpha_x \cdot \alpha_y$  по вертикали суммируются:  $\sum p \cdot \alpha_x \cdot \alpha_y$ .

$$\sigma = \sqrt{\frac{\sum p\alpha^2}{n} - b^2}, \quad b = \frac{\sum p\alpha}{n},$$

$$b_x = \frac{\sum p\alpha_x}{n}, \quad b_y = \frac{\sum p\alpha_y}{n};$$

$$b_x = \frac{6}{50} = 0,12; \quad b_y = \frac{0}{50} = 0;$$

$$\sigma_x = \sqrt{\frac{\sum p\alpha_x^2}{n} - b_x^2} = \sqrt{\frac{148}{50} - 0,12^2} = \pm 1,72;$$

$$\sigma_y = \sqrt{\frac{\sum p\alpha_y^2}{n} - b_y^2} = \sqrt{\frac{1850}{50} - 0} = \pm 1,72;$$

$$r = \frac{\sum p\alpha_x \cdot p\alpha_y - nb_x b_y}{n\sigma_x \sigma_y} = \frac{290 - 50 \cdot 0,12 \cdot 0}{50 \cdot 1,7 \cdot 6,1} = +0,56.$$

Ошибка репрезентативности выборочного коэффициента корреляции:

$$m_r = \sqrt{\frac{1 - r^2}{n - 2}} = \sqrt{\frac{1 - 0,31}{48}} = \pm 0,12. \quad r = 0,56 \pm 0,12.$$

*Вывод:* исследуемые признаки находятся в прямой корреляционной зависимости. Достоверность выборочного коэффициента корреляции:

$$t_r = \frac{r}{m_r} = \frac{0,56}{0,12} = 4,67 > t_{st} \text{ при } t_{st} \{1,7 - 2,0 - 2,7 - 3,5\}$$

$$\gamma = n - 2 = 48,$$

где  $n$  – количество сравниваемых пар.

Выборочный коэффициент корреляции является достоверным. На основе проведенного исследования можно утверждать достоверность наличия взаимосвязи исследуемых признаков.

Таблица 8.1

## Корреляционная решетка

$\begin{matrix} y \\ x \end{matrix}$	15	20	25	30	35 A	40	45	50	$\rho$	$\alpha_x$	$\rho \cdot \alpha_x$	$\alpha_x^2$	$\rho \alpha^2 x$	$\rho \cdot \alpha_x \cdot \alpha_y$
$\alpha_y$	-20	-15	-10	-5	0	+5	+10	+15						
18 -6				1					1	-6	-6	36	36	30
20 -4		1		2					3	-4	-12	16	48	-60 -40
22 -2				2					2	-2	-4	4	8	+20
24 B			4	6	12	8			30	0	0	0	0	0
26 +2					6	4	2	2	14	+2	+28	4	56	+40 +60 +40
28 +4									0	+4	0	16	0	0
30 +6									0	+6	0	36	0	0
$\rho$	0	1	4	11	18	12	2	2	50				$\Sigma=148$	$\Sigma \rho \cdot \alpha_x \cdot \alpha_y = 290$
$\alpha_y$	-20	-15	-10	-5	0	+5	+10	+15				$\Sigma \rho \alpha^2 x = 148$ $\Sigma \rho \alpha^2 y = 1850$ $\Sigma \rho \cdot \alpha_x \cdot \alpha_y = 290$		
$\rho \alpha_y$	0	-15	-40	-55	0	60	20	-30						
$\alpha^2 y$	400	225	100	25	0	25	100	225						
$\alpha_y^2$	0	225	400	275	0	300	200	450	$\Sigma \rho \alpha_y^2 = 1850$					

Примечание: значения А и В – условные средние

Таблица 8.2 (алгоритм)

Вычисление коэффициента корреляции для малочисленных групп										
первый способ					второй способ					
$r = \frac{\sum V_1 V_2 - \frac{\sum V_1 \sum V_2}{n}}{\sqrt{C_1 C_2}}; \quad (n \geq n_{st})$					$r = \frac{C_1 + C_2 - C_d}{2\sqrt{C_1 C_2}}; \quad (n \geq n_{st})$					
$V_1 V_2$ — даты признаков $C_1 C_2$ — сумма квадратов $C = \sum V^2 - \frac{(\sum V)^2}{n}$ $n$ — число сравниваемых пар					$C_1 C_2 C_d$ — сумма квадратов по первому и второму признакам и по ряду разностей $d = V_1 - V_2$ $C = \sum V^2 - \frac{(\sum V)^2}{n}$ и $C_d = \sum d^2 - \frac{(\sum d)^2}{n}$ $n$ — число сравниваемых пар					
$V_1$	$V_2$	$V_1^2$	$V_2^2$	$V_1 V_2$	$V_1$	$V_2$	$V_1^2$	$V_2^2$	$d = V_1 - V_2$	$d^2$
3	11	9	121	33	31	27	961	729	+4	16
7	10	49	100	70	22	24	484	576	-2	4
1	7	1	49	7	27	32	728	1024	-5	25
11	4	121	16	44	29	29	841	841	0	0
9	3	81	9	27	21	24	441	576	-3	9
5	9	25	81	45	30	27	900	729	+3	9
2	7	4	49	14	23	23	529	529	0	0
10	4	100	16	40	28	31	784	961	-3	9
4	12	16	144	48	25	30	625	900	-5	25
8	3	64	9	24	24	23	576	529	+1	1
60	70	470	594	352	260	270	6870	7394	-10	98
$C_1 = 470 - \frac{60^2}{10} = 110$ $C_2 = 594 - \frac{70^2}{10} = 104$ $r = \frac{352 - \frac{60 \cdot 70}{10}}{\sqrt{110 \cdot 104}} = \frac{-68}{107} = -0,64$ $n = 10; \quad n_{st} = \{10 - 15 - 23\}$					$C_1 = 6870 - \frac{260^2}{10} = 110$ $C_2 = 7394 - \frac{270^2}{10} = 104$ $C_d = 98 - \frac{10^2}{10} = 88$ $r = \frac{110 + 104 - 88}{2\sqrt{110 \cdot 104}} = \frac{+126}{214} = +0,59$ $n = 10; \quad n_{st} = \{11 - 18 - 27\}$					
Вывод: отрицательная корреляция в генеральной совокупности достоверна с вероятностью первого порога $\beta > 0,95$					Вывод: положительная корреляция в генеральной совокупности на грани достоверности первого порога. В исследованиях пониженной ответственности такую корреляцию можно считать достоверной. В ответственных работах следует повторить оценку корреляции на более обширном материале.					



Таблица 8.3 (алгоритм)

**Составление корреляционной решетки для последующего измерения  
корреляционных связей первого признака (1) со вторым (2)**

Первичные измерения										
$V_1$	107	169	121	168	167	124	138	145	130	98
$V_2$	60	93	54	90	86	57	64	71	47	43
1	133	50	163	87	135	111	188	72	140	132
2	57	37	81	50	61	37	101	44	67	55
1	117	165	147	153	149	179	172	142	151	113
2	50	84	73	70	74	104	87	69	65	42
1	134	155	93	161	159	80	139	173	137	177
2	59	73	37	80	77	35	66	90	63	95
1	102	136	157	185	127	131	152	115	175	104
2	48	62	75	97	63	53	67	48	93	53
$n = 50, g - \text{число классов} = 1 + 3,3 \cdot \log 50 = 7,$ $\lim_1 = 50 - 188(138), \quad k_1 = 138/7 = 20,$ $\lim_2 = 35 - 140(69), \quad k_2 = 69/7 = 10$										
$W_{\alpha}$	1	50–	70–	90–	110–	130–	150–	170–	$n_2$	
2	95–							4	4	
	85–						3	3	6	
	75–						5		5	
	65–					6	4		10	
	55–			1	2	7			10	
	45–		1	2	3	2			8	
	35–	1	2	2	2				7	
$n_1$		1	3	5	7	15	12	7	$N = 50$	

*Ложная корреляция.* Несмотря на получение достоверного коэффициента корреляции, иногда в действительности корреляции нет. Причинами ложной корреляции, кроме субъективных причин, связанных с ошибками сбора исходного материала, могут быть объективные, обусловленные свойствами изучаемой группы растений, животных, микроорганизмов. Одна из причин ложной корреляции связана с неоднородностью изучаемой совокупности, наличием в ее составе двух или нескольких биотипов. При однородном исследуемом материале эти варианты располагаются в овале, вытянутом по одной из диагоналей решетки. При неоднородном материале варианты располагаются на графике корреляционной решетки удаленно и дают две (или более) группы, соответствующие нескольким совокупностям. По расположению вариантов на корреляционной решетке можно заранее судить о том, какова зависимость между признаками: прямолинейная или криволинейная. В случае криволинейной зависимости признаков коэффициент корреляции не подходит для характеристики взаимосвязи. В этом случае следует использовать метод регрессии или другие критерии.

## ГЛАВА 9. ДИСПЕРСИОННЫЙ АНАЛИЗ

Дисперсионный анализ возник в процессе совершенствования методики сельскохозяйственного опытного дела. Вскоре его стали применять в биологии, смежных с нею областях, затем в технике, психологии и педагогике. Ценность метода заключается в том, что он позволяет выявить суммарное действие факторов и действие каждого регулируемого в опыте фактора в отдельности, действие различных сочетаний факторов друг с другом на резуль- тативный признак.

### **9.1. Сущность метода. Основные понятия и символы**

Наряду с относительно простыми способами сравнения одной выборки с другой в исследовательской работе встречаются и более сложные задачи, когда приходится сравнивать одновременно несколько выборок, объединяе- мых в единый статистический комплекс. В таких случаях метод парных срав- нений выборочных характеристик оказывается нерациональным. Учитывая это, английский математик и биолог Р. Фишер (1925) предложил метод ком- плексной оценки сравниваемых средних, получивший название дисперсион- ного анализа. Методы дисперсионного анализа, разработанные Р. Фишером, применялись первоначально лишь для анализа результатов опытов в расте- ниеводстве и животноводстве. В дальнейшем была доказана возможность использования дисперсионного анализа при изучении биологического мате- риала из природы, любых экспериментальных данных. Затем этот метод ста- ли применять в медицинских исследованиях, технике, педагогике, психологии и других отраслях. Медицинские данные чаще требуют именно дисперсион- ного анализа, и именно он служит основой для всех параметрических крите- риев. Изложение метода дисперсионного анализа Р. Фишера на русском язы- ке было впервые дано применительно к полевым опытам Н.Ф. Деревицким (1933). Для измерения силы влияния факторов было предложено несколько методов. Наибольшее признание получили методы профессора МГУ Н.А. Плохинского (1960, 1966, 1970 и др.) и Д.У. Снедекора (1961). В 1960 г. в Новосибирске вышла книга Н.А. Плохинского «Дисперсионный анализ».

Дисперсионный анализ основан на разложении общей дисперсии стати- стического комплекса на составляющие ее компоненты, сравнивая которые с использованием критерия Фишера, можно определить, какая доля общей ва- риации резуль- тативного признака обусловлена действием на него регулиру- емых и нерегулируемых в опыте факторов. Так, если регулируемый фактор (например, доза удобрений) оказывает существенное влияние на результа- тивный признак (урожай культуры), оно отразится на величине групповых

средних, которые будут отличаться друг от друга. Таким образом, происходит варьирование групповых средних, причиной которого является влияние регулируемого фактора. Внутри каждой группы, входящей в статистический (дисперсионный) комплекс, также обнаружится варьирование, вызванное влиянием на признак нерегулируемых в опыте факторов. Зависимость между этими источниками варьирования выразится равенством:

$$D_y = D_x + D_e,$$

где  $D_x$  – межгрупповая вариация (девиата), представляющая собой сумму квадратов центральных отклонений групповых средних  $x_i$  (их общее число  $a$ ) от общей средней комплекса, взвешенную на численность вариантов в группах  $n$ , т.е. при  $N = \sum n$ ;  $D_e$  (внутригрупповая девиата, представляющая сумму из сумм квадратов отклонений отдельных вариантов  $x_i$  от их групповых средних, т.е.  $D_y$  – общая девиата или сумма квадратов центральных отклонений вариант (дат) от общей средней комплекса.

Выборочные дисперсии  $C_y = C_x + C_z$  служат основой оценками соответствующих генеральных параметров;  $C_y$  является общей дисперсией всего комплекса. Отношение межгрупповой дисперсией  $C_x$ , называемой также факториальной (так как она зависит от действия регулируемых факторов), к внутригрупповой ( $C_z$ ) дисперсии, служит критерием оценки влияния регулируемых в опыте факторов на результативный признак.

*Нулевая гипотеза* сводится к предположению о том, что генеральные межгрупповые средние и дисперсии равны между собой и различия, наблюдаемые между выборочными показателями, вызваны случайными причинами, а не влиянием на признак регулируемых факторов. Нулевую гипотезу отвергают, если  $F_{\text{факт}} \geq F_{st}$  принятого уровня значимости ( $\alpha$ ) и числа степеней свободы, и принимают, если  $F_{\text{факт}} < F_{st}$ ; при этом различия, наблюдаемые между групповыми средними комплекса, признают статистически недостоверными. После того как действие регулируемого фактора, нескольких факторов или их совместного влияния на признак будет доказано, т.е. окажется статистически достоверным, переходят к сравнительной оценке групповых средних. Заключительным этапом дисперсионного анализа является оценка силы влияния отдельных факторов или их совместного действия на признак. Являясь методом одновременных сравнений выборочных средних, дисперсионный анализ предъявляет высокие требования к группировке выборочных данных, к планированию наблюдений, подлежащих дисперсионному анализу. Выборочные данные группируют с учетом градаций каждого регулируемого фактора, действующего на признак, например по дозам удобрений, срокам или способам внесения их в почву, породной и линейной принадлежности экспериментальных животных, их возрасту. Применение дисперси-

онного анализа предполагает нормальное (или близкое к нормальному) распределение совокупности, из которой взяты выборки, объединяемые в дисперсионный комплекс. При этом важно, чтобы дисперсии выборочных групп были одинаковыми или не очень сильно отличались друг от друга. Весьма желательно, чтобы при планировании наблюдений и, особенно, при обработке их результатов в группах дисперсионного комплекса содержались одинаковое или пропорциональное число вариантов (дат).

*Основные понятия и символы.* Признаки, изменяющиеся под воздействием тех или иных причин, называют результативными. Причины, вызвавшие изменение величины результативного признака или признаков, принято называть факторами. Например, масса тела, его линейные размеры, урожай культуры; успеваемость учащихся и другие – все это признаки, на которые оказывают влияние самые различные факторы: элементы или режим питания, физические или умственные упражнения, дозы лекарственных препаратов, удобрений, микроэлементов, БАДов и др. Факторы обозначают прописными буквами латинского алфавита ( $A, B, C, \dots$ ), учитываемые признаки – конечными буквами ( $X, Y, Z$ ). Этот метод основывается на разложении общей дисперсии на составляющие ее компоненты, сравнивая которые между собою с помощью критерия  $F$  (Фишера), можно определить, какая доля от общего разнообразия исследуемого (результативного) признака обусловлена действием регулируемых и нерегулируемых (в данном опыте) факторов. Сущность дисперсионного анализа заключается в изучении статистического влияния одного или нескольких факторов на результативный признак. Под статистическим влиянием понимают отражение организованного разнообразия влияния градаций фактора (например, дозы) на разнообразие результативного признака.

Возможности дисперсионного анализа:

1. Оценка силы и достоверности влияний.
2. Оценка разности частных средних и частных долей.
3. Оценка наследуемости признаков в определенных группах особей при передаче генетической информации из поколения в поколение.
4. Анализ комбинационной способности линий.

Дисперсионный (вариансный) анализ необходим, так как влияние факторов на признак никогда не может быть выделено в чистом виде. Хотя условия опыта должны быть однородными, различные опыты дают несколько неодинаковые результаты. Причины:

- 1) влияние многочисленных случайных обстоятельств в природе и опыте;
- 2) влияние других факторов, меняющихся от опыта к опыту, не поддающихся контролю;
- 3) велика роль неконтролируемых факторов в природе.

Задача дисперсионного анализа: разложение общей изменчивости признака на составные части:

$$C_y = C_x + C_z,$$

$C_x$  – влияние изучаемого фактора;  $C_z$  – влияние случайных факторов;  $C_y$  – общее, влияние.

Дисперсионный анализ позволяет оценить значимость влияния отдельных факторов и их роль в общей изменчивости результативного признака.

Результативный признак – признак (физиологический морфологический или другой) растений, животных, человека, изменяющийся под влиянием различных причин: масса тела, содержание сахара в клеточном соке, процент белка в семенах, артериальное давление, количество (%) хромосомных перестроек.

Факторы, вызывающие изменение признака: дозы лекарственных препаратов, удобрений, физические или умственные упражнения, обозначаются прописными буквами латинского алфавита ( $A, B, C$ ), результативные (учитываемые) признаки –  $X, Y, Z$ .

Организованные факторы (регулируемые) испытывают серийно, в виде нескольких независимых друг от друга доз (градаций). В опыте регулируются лишь некоторые факторы, другие не подвергаются регулированию, хотя и влияют на величину результативного признака.

Градации обозначают теми же буквами, что и факторы. Так, фактор  $A$  с градациями:  $A_1, A_2, A_3$ ; градации фактора  $B$ :  $B_1, B_2, B_3$ .

Типы (виды) дисперсионных комплексов:

- однофакторные;
- двух-, трех-, многофакторные;
- равномерные; ортогональные; пропорциональные,
- неравномерные – неортогональные.

## **9.2. Анализ однофакторных дисперсионных комплексов количественных признаков**

Техника дисперсионного анализа однофакторных комплексов заключается в расчете показателей варьирования – дисперсий (сумма квадратов центральных отклонений). При этом рассчитываются групповые средние и общая средняя арифметическая для всего комплекса. Первичные данные, подлежащие дисперсионному анализу, группируют в виде комбинационной таблицы (дисперсионный комплекс). Рассмотрим пример расчета равномерного однофакторного дисперсионного комплекса.

**Пример.** Изучается действие различных доз рентгеновского облучения на высоту стебля растений пшеницы на 10-й день после всходов.

Градации фактора: дозы облучения (в рентгенах):

$A_0$  – контроль (без облучения);

$A_1$  – 10 рентген;

$A_2$  – 20 рентген;

$A_3$  – 30 рентген.

Допустим, что измерение высоты стебля проведено всего у трех растений по каждому варианту опыта. Необходимо установить силу и достоверность влияния рентгеновского облучения на изменение высоты стебля растений.

В однофакторных комплексах определяются три дисперсии, соответствующие трем типам влияния и трем типам разнообразия (факториальное, случайное и общее):

$$C = \sum d^2 -$$

сумма квадратов центральных отклонений: дат от общей и частной средних, частных средних от общей средней.

Таблица 9.1

Структура однофакторного дисперсионного комплекса

Структура комплекса		Градации				$r = 4$ число градаций
		$A_0$	$A_1$	$A_2$	$A_3$	
Даты $V$		7	5	5	0	$n = 3$ объем градаций
		9	9	3	2	
		8	7	1	4	
Объем градаций		3	3	3	3	$N = 12$ объем комплекса
Сумма дат $\sum V$		24	21	9	6	$\sum \sum V = 60$
Частные средние $M_i$		8	7	3	2	$M_\Sigma = 5$ общая средняя
$X$	$M_i - M_\Sigma$	+3	+2	-2	-3	Факториальная дисперсия $C_x = \sum n(V - M_i)^2 = 78$
	$(M_i - M_\Sigma)^2$	9	4	4	9	
	$n(M_i - M_\Sigma)^2$	27	12	12	27	
$Z$	$V - M_i$	-1, +1, 0	2, 2, 0	-2, 0, -2	-2, 0, +2	Случайная дисперсия $C_z = \sum (V - M_i)^2 = 26$
	$(V - M_i)^2$	1, 1, 0	4, 4, 0	4, 0, 4	4, 0, 4	
$Y$	$V - M_\Sigma$	2, +4, +3	0, +4, +2	0, -2, -4	-5, -3, -1	Общая дисперсия $C_y = \sum (V - M_\Sigma)^2 = 104$
	$(V - M_\Sigma)^2$	4, 16, 9	0, 16, 4	0, 4, 16	25, 9, 1	

1. Факториальная (межгрупповая) дисперсия равна сумме взвешенных квадратов центральных отклонений частных средних  $M_i$  по градациям комплекса от общей средней  $M_\Sigma$  :

$$C_x = \sum n(M_i - M_\Sigma)^2.$$

2. Случайная дисперсия (внутригрупповая) равна сумме квадратов центральных отклонений дат ( $V$ ) от своих частных средних ( $M_i$ ) по градациям комплекса:

$$C_z = \sum (V - M_i)^2.$$

4. Общая дисперсия равна сумме квадратов центральных отклонений дат ( $V$ ) от общей средней:

$$C_y = \sum (V - M_{\Sigma})^2.$$

Структура однофакторного дисперсионного комплекса и расчет дисперсий, показателей силы влияния фактора, вариантов, эмпирических критериев достоверности приводятся в таблице 9.1.

Действует закон аддитивности:

$$C_x + C_z = C_y \quad (78 + 26) = 104$$

В любом дисперсионном комплексе сумма частных дисперсий (факториальной и случайной) равна общей.

Ряд частных средних  $M_i = 8, 7, 3, 2$  свидетельствует о заметном влиянии рентгеновского облучения на высоту стебля растений пшеницы: при увеличении дозы высота стебля растений закономерно уменьшается. На основе сопоставления частных средних и градаций фактора строятся графики, которые дают наглядное представление о действии фактора. Используя три полученные дисперсии, можно рассчитать:

1. Основной показатель силы влияния:

$$\eta_x^2 = \frac{C_x}{C_y} = \frac{78}{104} = 0,75.$$

2. Эмпирический критерий достоверности силы влияния ( $F$ ):

$$F = \frac{C_x}{C_z} \cdot \frac{N - r}{r - 1} = \frac{78}{26} \cdot \frac{8}{3} = 8 > F_{st} \text{ (приложение, табл. 5).}$$

Расчет вариантов в современном дисперсионном анализе не требуется, однако он необходим в данном случае для сравнения с дисперсионным анализом другими методами. Р. Фишером доказано, что варианты нужно вычислять путем деления дисперсий на число степеней свободы: вариант  $\delta_i^2 = \frac{C_i}{\gamma_i}$  (дисперсия, деленная на число степеней свободы).

Рассчитываются три вида вариантов:

1. Факториальный вариант:

$$\delta_x^2 = \frac{C_x}{\gamma_x} = \frac{C_x}{r - 1} = \frac{78}{3} = 26.$$

Случайный вариант:

$$\delta_z^2 = \frac{C_z}{\gamma_z} = \frac{C_z}{N-r} = \frac{26}{8} = 3,25.$$

Общая вариация:

$$\delta_y^2 = \frac{C_y}{\gamma_y} = \frac{C_y}{N-1} = \frac{104}{11} = 9,45.$$

Вариансы неаддитивны.

Уравнение Пирсона:

$$\frac{C_x}{C_y} + \frac{C_z}{C_y} = \eta_x^2 + \eta_z^2 = 1,$$

$\eta_x^2$  – основной показатель силы влияния отражает долю разнообразия (варьирования) признака, обусловленную действием исследуемого фактора.

Эмпирический критерий достоверности силы влияния ( $F$ ) может быть вычислен:

$$F = \frac{\delta_x^2}{\delta_z^2} = \frac{26,00}{3,25} = 8 > F_{st} \text{ (при } \gamma_1 = 3, \gamma_2 = 8).$$

Эмпирический критерий достоверности превышает второе стандартное значение  $F_{st}$ . Таким образом, действие рентгеновского облучения достоверно уменьшает высоту стебля растений пшеницы ( $B = 0,99$ ).

Кроме рассмотренного выше, существует рабочий алгоритм расчета дисперсий (Н.А. Плохинского).

Таблица 9.2

Рабочий алгоритм расчета дисперсий

	Градации				$r = 3$
	$A_0$	$A_1$	$A_2$	$A_3$	
$V$	7, 9, 8	5, 9, 7	5, 3, 1	0, 2, 4	$n = 3$
$n$	3	3	3	3	$N = 12$
$\sum V$	24	21	9	6	$\sum V = 60$
$H_i = \frac{(\sum V)^2}{n}$	192	147	27	12	$\sum H_i = 378$
$\sum V^2$	194	155	85	20	$\sum V^2 = 404$
$M_i = \frac{\sum V}{n}$	8	7	3	2	$H_\Sigma \frac{3600}{12} = 300$

$$C_x = \sum H_i - H_\Sigma = 378 - 300 = 78.$$

$$C_z = \sum V^2 - \sum H_i = 404 - 378 = 26.$$

$$C_y = \sum V^2 - H_\Sigma = 404 - 300 = 104.$$



Все три дисперсии можно рассчитать, используя экспресс-метод Н.А. Плохинского:

$$\eta_x^2 = \frac{C_x}{C_y} = \frac{\sum Hi - H_\Sigma}{\sum V^2 - H_\Sigma} = \frac{378 - 300}{404 - 300} = 0,75.$$

$$F = \frac{\eta_x^2}{1 - \eta_x^2} \cdot \frac{N - r}{r - 1} = \frac{0,75}{0,25} \cdot \frac{8}{3} = 8.$$

*Вывод.* Действие рентгеновского облучения на высоту стебля растений пшеницы будет аналогичным при этих дозах в генеральной совокупности.

*Ошибка репрезентативности* в однофакторных комплексах рассчитывается только для одного показателя факториального влияния:

$$m_{\eta_x^2} = 1 - \eta_x^2 \cdot \frac{r - 1}{N - r}.$$

*Доверительные границы* определяются обычным способом по известной ранее формуле, в которой вместо t-критерия Стьюдента используется F-критерий Фишера:

$$\eta_x^2 \pm F_{st} \cdot m_{\eta_x^2} = \eta_x^2 \pm \Delta.$$

Если одна из доверительных границ выходит за максимально допустимый предел при определении доверительных границ, верхняя граница приравнивается к 1,0. При выходе за минимальный предел нижняя граница приравнивается к нулю. Так, если доверительные границы генерального параметра равны  $\eta_x^2 = 0,14 - 1,14$ , то это означает, что сила влияния исследуемого фактора в генеральной совокупности может составить не менее 14 % от всех факторов, определяющих величину (степень) развития признака. Основным показателем силы влияния всегда больше нуля, он не может быть отрицательным. Наибольшая величина показателя  $\eta_x^2 = 1$  может быть в том случае, если даты внутри каждой градации одинаковы и равны своей частной средней. Показатель силы влияния может быть меньше нуля и больше единицы только при определении доверительных границ генерального параметра при малочисленных выборках, при большом разнообразии дат.

Сравнение эмпирического критерия достоверности силы влияния ( $F$ ) со стандартным значением может дать два результата:

- влияние недостоверно,
- влияние достоверно.

Таблица 9.3

**Дисперсионный анализ однофакторных комплексов  
для количественных признаков для малых групп**

	Градации					число градаций $g = 5$	Факториальная дисперсия $C = \sum H_i - H^2 = 52$
	1	2	3	4	5		
Дата $V$	2	4	5	9	3	$H_x = \frac{(\sum V)^2}{N} = \frac{100^2}{20} = 500$	Случайная дисперсия $C_x = \sum V^2 - \sum H_i = 586 - 552 = 34$
	3	3	6	7	6		
	3	6	4	6	5		
	1	3	6	6	6		
П	3	4	5	4	4	объем комплекса: $N = \sum n = 20$	Общая дисперсия $C_y = \sum V^2 - H_x = 586 - 500 = 86$
$\sum V$	6	16	30	28	20	$\sum \sum V = 100$	Факториальная дисперсия $\sigma_x^2 = C_x / g - 1 = 52 / 4 = 13,0$
$H_i = \frac{(\sum V)^2}{n}$	12	64	180	196	100	$\sum H_i = 552$	
$\sum V^2$	14	70	194	202	106	$\sum V^2 = 586$	Случайная дисперсия $\sigma_x^2 = C_x / N - g = 34 / 15 = 2,27$
частные средние $M_i$	2	4	6	7	5	общая средняя $M_x = 5$	
Показатель силы влияния $\eta_x^2 = C_x / C_y = 52 / 86 = \underline{0,605}$							
Его ошибка $m_{\eta_x^2} = (1 - \eta_x^2) \frac{g - 1}{n - g} = 0,395 \cdot \frac{4}{15} = 0,105$							
Его достоверность $F = \frac{\eta_x^2}{m_{\eta_x^2}} = \frac{0,605}{0,105} = \underline{5,76}$ $\nu_1 = g - 1 = 4; \quad \nu_2 = N - g = 15$ $F_{\alpha} = \{3,1 - 4,9 - 8,3\}$							
Доверительные границы генерального показателя (приближенное значен.) $\Delta F_{\alpha} \cdot m_{\eta_x^2} = 3,1 \cdot 0,105 = 0,33;$ $(\beta = 0,95)$ $\bar{\eta}_x^2 = \begin{cases} \bar{\eta}_x^2 + \Delta = 0,61 + 0,33 = 0,94 \\ \bar{\eta}_x^2 - \Delta = 0,61 - 0,33 = 0,28 \end{cases}$							
Достоверность по Фишеру $F = \frac{\sigma_x^2}{\sigma_z^2} = \frac{13,00}{2,27} = \underline{5,74}$					ОБЩИЙ ВЫВОД Влияние фактора достоверно с вероятностью $\beta > 0,99$ . Для всех объектов данной категории влияние изучаемого фактора может составить не менее 28 % от общего влияния всей суммы факторов		
ФОРМА ИТОГОВОЙ ЗАПИСИ							
Разнообразие	дисперсии (суммы квадратов) $C$	числа степеней свободы $\nu$	вариансы (средние квадраты) $\sigma^2$	$\eta_x^2 = 0,605 \pm 0,105$			
факториальное (межгрупповое)	52	4	13,00	$F = \frac{0,605}{0,105} = 5,76$			
случайное (внутригрупповое)	34	15	2,27	$F = \frac{13,00}{2,27} = 5,74$			
общее	86	19	4,53	$F_{\alpha} = \{3,1 + 4,9 - 8,3\}$			

*Влияние не достоверно*, если эмпирический критерий Фишера не достигает стандартного значения, соответствующего установленному порогу вероятности безошибочного прогноза. Получение недостоверного показателя силы влияния ни отрицает, ни подтверждает влияния фактора в генеральной совокупности. В таких случаях нельзя дать определенный прогноз о генеральном влиянии фактора: остается не выясненным, можно ли ожидать (с установленной вероятностью), что в генеральной совокупности при действии данного фактора получатся результаты, близкие к тем, которые получены при выборочном исследовании.

*Влияние достоверно* – это значит, что эмпирический критерий равен или превышает стандартное значение критерия Фишера с требуемой вероятностью. Изучаемый фактор при его массовом применении будет оказывать влияние на результативный признак.

### **9.3. Анализ однофакторных дисперсионных комплексов для качественных признаков**

Если в исследуемой группе из  $n$  особей данный признак проявляется у  $m$  особей, то вероятность проявления признака в группе особей будет:

$$p = \frac{m}{n}.$$

Это показатель имеет значение средней арифметической и принимается за частную и общую среднюю в частных и общих группах. Дисперсия качественного признака (% завязываемости плодов, всхожесть семян, укоренение черенков в % и другие) вычисляется:

$$C = n \cdot p \cdot q \cdot q = 1 - p = n \frac{m}{n} \left(1 - \frac{m}{n}\right) Hi = \frac{m^2}{n}.$$

Таблица 9.4

**Пример расчета однофакторного дисперсионного комплекса  
для качественных признаков**

для ка 100-балльного приложения

X	Градации r = 4				$C_x = \sum H_i - H_\Sigma$
	1 %	2 %	3 %	4 %	
n: ( $\sum n = 80$ )	10	20	30	20	$C_x = \sum m - \sum H_i$ . Случайная дисперсия $C_H = \sum m - H_\Sigma$ . Общая дисперсия: $H_\Sigma = \frac{(\sum m)^2}{N} = \frac{840}{80} = 10,5$
n: ( $\sum m = 29$ )	5	8	6	10	
$m^2$	26	64	36	100	
$H_i = \frac{m^2}{n}$	2,5	3,2	1,2	5,0	$\sum H_i = \sum \frac{m^2}{n} = 11,$ $N = \sum n$
$p = \frac{m}{n}$	0,5	0,4	0,2	0,5	
$C_x = 11,9 - 10,5 = 1,4,$ $C_z = 29,0 - 11,9 = 17,1,$ $C_y = 29,0 - 10,5 = 18,5,$ $F = 2,14; F_{st}: 2,7 - 4,0 - 6,0$					
	x		z		y
C	1,4		17,1		18,5
$\eta_x^2 = C_1/C_y$	0,076		0,924		1,0
$\gamma_{(ню)}$	$3_{(r-i)}$		$76_{(N-r)}$		79 $\Sigma_{n-1}$
$\sigma^2 = C_1/v_1$	0,47		0,22		-

Дисперсионный анализ однофакторных комплексов для качественных признаков							
	Градации					Число градаций $g = 5$	Факториальная дисперсия
	1	2	3	4	5	$H_{\Sigma} = (\Sigma m)^2 / N = 48^2 / 160 = 14,4$	$C_x = \sum H_i - H_{\Sigma} = 19,6 - 14,4 = 5,2$
$n$	20	30	40	30	40	$N = \sum n = 160$	Случайная дисперсия $C_z = \sum m - \sum H_i = 48,0 - 19,6 = 28,4$
$m$	2	3	8	15	20	$\sum m = 48$	Общая дисперсия $C_y = \sum m - H_z = 48,0 - 14,4 = 33,6$
$H_i = \frac{m^2}{n}$	0,2	0,3	1,6	7,5	10,0	$\sum H_i - 19,6$	Факториальная вариация $\sigma_x^2 = \frac{C_x}{g - 1} = \frac{5,2}{4} = 1,300$
$P = \frac{m}{n}$	0,1	0,1	0,2	0,5	0,5	$P_z = 0,3$	Случайная вариация $\sigma_x^2 = \frac{C_z}{N - g} = \frac{28,4}{155} = 0,183$
Показатель силы влияния							
$\eta_x^2 = \frac{C_x}{C_y} = \frac{5,2}{33,6} = \underline{\underline{\underline{0,155}}}$							
Его ошибка							
$m_{\eta_x^2} = (1 - \eta_x^2) \frac{g - 1}{N - g} = 0,845 \cdot \frac{4}{155} = \pm 0,0218$							

Продолжение табл. 9.5

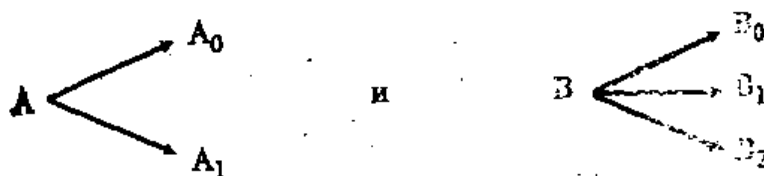
Его достоверность		$F = \eta_x^2 / m_{\eta_x^2} = \frac{0,155}{0,0128} = \underline{\underline{7,1}}$		$V_1 = g - 1 = 4,$ $V_2 = N - g = 155,$ $F_{st} = \{2,4 - 3,4 - 4,9\}$
Доверительные границы генерального показателя (приближенные значения)				
$\Delta F_{st} \cdot m_{\eta_x^2} = 2,4 \cdot 0,0218 = 0,052$ $(\beta = 0,95)$		$\eta_x^2 \begin{cases} \tilde{\eta}_x^2 + \Delta = 0,155 + 0,052 = 0,207 \\ \tilde{\eta}_x^2 - \Delta = 0,155 - 0,052 = 0,103 \end{cases}$		
$F = \frac{\sigma_x^2}{\sigma^2} = \frac{1,300}{0,183} = \underline{\underline{7,1}}$	Общий вывод: влияние фактора достоверно в высшей степени для всех объектов данной категории. Влияние данного фактора может составить $(\beta = 0,95)$ не менее 10 % и не более 21 % от общего влияния всей суммы факторов			
ФОРМА ИТОГОВОЙ ЗАПИСИ				
Разнообразие	дисперсии (суммы квадратов), $C$	числа степеней свободы, $V$	вариансы (средние квадраты), $\sigma^2$	$\eta_x^2 = 0,155 \pm 0,0128,$ $F = \frac{0,155}{0,0128} = \underline{\underline{7,1}},$ $F = \frac{1,300}{0,183} = 7,1,$ $F_{st} = \{2,4 - 3,4 - 4,9\}$
Факториальное (межгрупповое)	5,2	4	1,300	
Случайное (межгрупповое)	28,4	155	0,183	
Общее	33,6	159	0,211	

#### 9.4. Анализ двухфакторных дисперсионных комплексов для качественных признаков

Содержание исследования (пример): изучается влияние двух стимуляторов роста на высоту стебля растений сои;

- первый стимулятор (A) кислотный;
- второй стимулятор (B) щелочной

Каждая доза первого стимулятора исследовалась вместе с каждой дозой второго стимулятора (табл. 9.6).



Результативный признак (РП) – высота стебля исследуемого объекта.

Из графиков (табл. 9.6) видно, что действие второго стимулятора (B) оптимально (до определенного предела). Действие второго стимулятора (B) нейтрализует действие первого, так как в ситуации A1B0 (отсутствие второго) первый стимулятор действует. Второй стимулятор действует в отсутствие первого. Действие второго стимулятора снижает действие первого. Совместное действие стимуляторов снижает результативный признак.

Далее следует рассчитать шесть дисперсий:  $C_A, C_B, C_{AB}, C_x, C_z, C_y$ :

A – разнообразие частных средних по фактору A при усредненном влиянии фактора B; B – разнообразие частных средних по фактору B при усредненном влиянии фактора A; AB – разнообразие, внесенное сочетанием действия обоих факторов.

Таблица 9.6

Пример расчета двухфакторного дисперсионного комплекса (2 способа)

A	A <sub>0</sub>			A <sub>1</sub>			r <sub>A</sub> = 2	<div>Основной способ — 1</div>
B	B <sub>0</sub>	B <sub>1</sub>	B <sub>2</sub>	B <sub>0</sub>	B <sub>1</sub>	B <sub>2</sub>	r <sub>B</sub> = 3	
V	1 4 1	10 9 11	7 4 7	5 6 7	4 3 5	2 3 1	n = 3	
n	3	3	3	3	3	3	N = 18	
ΣV	6	30	18	18	12	6	Σ = 90	
M <sub>1</sub> част. сред.	2	10	6	6	4	2	M <sub>Σ</sub> = 5	

	n	ΣV	M <sub>1</sub>	D	D <sup>2</sup>	n D <sup>2</sup> взвеш.	<div>2-й способ</div>
A <sub>0</sub>	9	54	6	+1	1	9	
A <sub>1</sub>	9	36	4	-1	1	9	
Σ	18	C <sub>A</sub> = Σ n D <sub>A</sub> <sup>2</sup> = 18					
B <sub>0</sub>	6	24	4	-1	1	6	
B <sub>1</sub>	6	42	7	+2	4	24	
B <sub>2</sub>	6	24	4	-1	1	6	
Σ	18	C <sub>B</sub> = Σ n D <sub>B</sub> <sup>2</sup> = 36					

	A <sub>0</sub> B <sub>0</sub>	A <sub>0</sub> B <sub>1</sub>	A <sub>0</sub> B <sub>2</sub>	A <sub>1</sub> B <sub>0</sub>	A <sub>1</sub> B <sub>1</sub>	A <sub>1</sub> B <sub>2</sub>	n — общая повторность комплекса
D <sub>x</sub> = M <sub>1</sub> - M <sub>Σ</sub>	-3	+5	+1	+1	-1	-3	C <sub>X</sub> = Σ n (M <sub>1</sub> - M <sub>Σ</sub> ) <sup>2</sup> = 46 × 3 = 138
D <sub>Z</sub>	-1+2-1	0-1+1	+1-2+1	-1-0+1	0-1+1	0+1-1	D <sub>Z</sub> = V - M <sub>1</sub>
Σ D <sub>Z</sub> <sup>2</sup>	6	2	6	2	2	2	C <sub>Z</sub> = Σ (V - M <sub>1</sub> ) <sup>2</sup> = 20

N = 18 (объем комплекса)      r<sub>A</sub> = 2,      r<sub>B</sub> = 3 — градации факторов  
 n = 3 — объем градаций, D — отклонение дат от частного среднего.



Таблица 9.7

## Изучение силы и достоверности влияния стимуляторов

	A	B	AB	X	Z	Y
	при усредненном влиянии					
	B	A				
C	18	36	84	138	20	158
$\eta^2 = C_i/C_y$	0,11	0,23	0,53	0,87	0,13	$\neq 1,0$
$\gamma$	1	2	2	5	12	17
$\sigma^2 = \frac{C_i}{\gamma}$	18	18	42	27,6	1,67	
$\sigma_1^2/\sigma_z^2$	<u>10,8</u>	<u>10,8</u>	<u>25,2</u>	<u>16,5</u>		
$F_{\text{ит}}$	18,6	12,3	12,3	8,9		
	9,3	6,9	6,9	5,1		
	4,8	3,9	3,9	3,1		

Кроме частных дисперсий, рассчитываются:

$$C_x + C_z = C_y$$

а)  $C_x$  суммарное факториальное разнообразие:

$$C_x = \sum n(M_i - M_{\Sigma})^2 = 138,$$

$$C_x = C_A + C_B + C_{AB};$$

б)  $C_z$  – случайная дисперсия:

$$C_z = \sum (V - M_i)^2;$$

в)  $C_{AB}$  – разнообразие, определяемое сочетанием градаций двух факторов.

После расчета дисперсий необходимо **найти силу влияния каждого из двух исследуемых факторов ( $\eta_x^2$ )**: для A и B в отдельности, а также действие сочетания градаций этих факторов ( $\eta_{AB}^2$ ) на результирующий признак и силу влияния случайных факторов:

$$\frac{\eta_z^2}{\eta_i^2} = \frac{C_i}{C_y}.$$

В двухфакторном дисперсионном комплексе, кроме самостоятельного влияния двух факторов, *существует дисперсия, определяемая сочетанием градаций этих факторов*. Она составляет в данном случае 53 % от всего разнообразия. На долю случайных факторов приходится 13 % от общего разнообразия.

Таблица 9.8

**Рабочий способ решения двухфакторных равномерных комплексов (пример тот же)**

A	A <sub>0</sub>			A <sub>1</sub>			г <sub>A</sub> =2		n	ΣV	H <sub>1</sub>
B	B <sub>0</sub>	B <sub>1</sub>	B <sub>2</sub>	B <sub>0</sub>	B <sub>1</sub>	B <sub>2</sub>	г <sub>B</sub> =3	A <sub>0</sub>	9	54	324
V	1,4,1	10,9,11	7,4,7	5,6,7	4,3,5	2,3,1	$\frac{n=3}{N=18}$	A <sub>1</sub>	9	36	144
ΣV	6	30	18	18	21	6	ΣV=90	$H_1 = \frac{(\Sigma V)^2}{n}$		H <sub>A</sub> =468	
H <sub>1</sub>	12	300	108	108	48	12	ΣH <sub>1</sub> =588	B <sub>0</sub>	6	24	96
ΣV <sup>2</sup>	18	302	114	110	50	14	ΣV <sup>2</sup> =608	B <sub>1</sub>	6	42	294
M <sub>1</sub>	2	10	6	6	4	2		B <sub>2</sub>	6	24	96
$H_{\Sigma} = \frac{(\Sigma V)^2}{N} = \frac{90^2}{18} = \frac{8100}{18} = 450 \qquad \Sigma H_B = 486$											
$C_{\gamma} = \Sigma V^2 - H_{\Sigma} = 608 - 450 = 158 \qquad C_A = \Sigma H_A - H_{\Sigma} = 468 - 450 = 18$											
$C_{\delta} = \Sigma V^2 - \Sigma H_1 = 608 - 588 = 20 \qquad C_B = \Sigma H_B - H_{\Sigma} = 486 - 450 = 36$											
$C_{\epsilon} = \Sigma H_1 - H_{\Sigma} = 588 - 450 = 138 \qquad C_{AB} = C_{\gamma} - C_A - C_B = 158 - 18 - 36 = 104$											

Суммарное действие факторов:  $A + B + AB$  составляет 87 %.

*Заключительный этап:* проведение сравнения эмпирических критериев достоверности со стандартными значениями критериев Фишера (с учетом числа степеней свободы), которое показало, что влияние сочетания градаций достоверно по третьему порогу вероятности безошибочного прогноза. Достоверно влияние на результативный признак каждого из факторов,  $A$  и  $B$ .

*Вопрос:* как можно использовать эти два стимулятора?

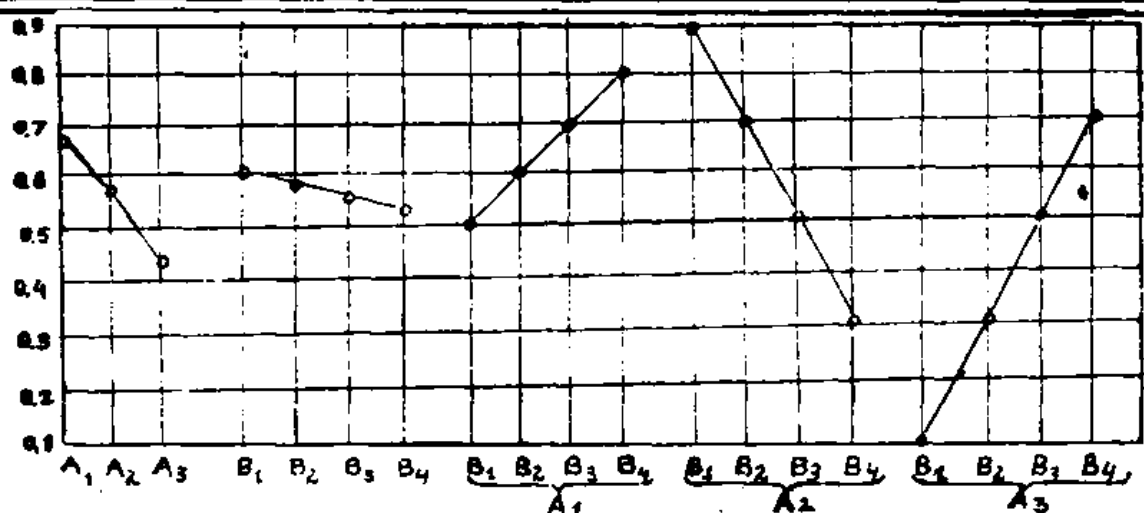
*Ответ:* два стимулятора использовать нет смысла. Целесообразно использовать стимулятор  $B$ , так как в отсутствие действия фактора  $A$  он дает наибольший эффект (в дозе  $B_1$ ).

Производится расчет шести дисперсий:  $C_A, C_B, C_{AB}, C_X, C_Z, C_Y$ .

Таблица 9.9

**Дисперсионный анализ двухфакторных пропорциональных комплексов  
для качественных признаков**

	$A_1$				$A_2$				$A_3$				$\varepsilon_A = 3$ $\varepsilon_B = 4$		$\Sigma n$	$\Sigma m$	$H_i = \frac{(\Sigma m)^2}{\Sigma n}$	$P_i$
	$B_1$	$B_2$	$B_3$	$B_4$	$B_1$	$B_2$	$B_3$	$B_4$	$B_1$	$B_2$	$B_3$	$B_4$						
$n$	10	20	10	20	20	40	20	40	10	20	10	20	$N = \Sigma n = 240$	$A_1$	60	40	26,67	0,67
$m$	5	12	7	16	18	28	10	12	6	5	14		$\Sigma m = 134$	$A_2$	120	68	38,53	0,57
$H_i = \frac{m^2}{n}$	2,5	7,2	4,9	12,8	16,2	19,6	5,0	3,6	0,1	1,8	2,5	9,8	$\Sigma H_i = 86,6$	$A_3$	60	26	11,27	0,43
$P_x = \frac{m}{n}$	0,50	0,60	0,70	0,80	0,90	0,70	0,50	0,30	0,10	0,30	0,50	0,70	$H_\Sigma = 74,8$	$H_A = 76,47$				
	$A$	$B$	$AB$		$X$		$Z$		$Y$		$B_1$	40	24	14,40	0,60			
$C_i$	$H_A - H_\Sigma$ 1,65	$H_B - H_\Sigma$ 0,18	$C_X - C_A - C_B$ 9,35		$\Sigma H_i - H_\Sigma$ 11,18		$\Sigma m - \Sigma H_i$ 48,0		$\Sigma m - H_\Sigma$ 59,18		$B_2$	80	46	26,45	0,58			
$\eta^2 = \frac{C_i}{C_y}$	0,028	0,003	0,158		0,189		0,811		1,000		$B_3$	40	22	12,10	0,55			
	$A$	$B$	$AB$		$X$		$Z$		$Y$		$B_4$	80	42	22,05	0,53			
	$H_B = 75,00$																	
$\nu$	$\varepsilon_A - 1$ 2	$\varepsilon_B - 1$ 3	$(\varepsilon_A - 1)(\varepsilon_B - 1)$ 6		$\varepsilon_A \varepsilon_B - 1$ 11		$N - \varepsilon_A \varepsilon_B$ 228		$N - 1$ 239		$\nu_1$	2	3	6	11			
$\sigma_i^2 = \frac{C_i}{\nu_i}$	0,83	0,06	1,56		1,02		1,02		0,25		$\nu_2$	7,2	5,6	3,9	3,1			
												4,7	3,9	2,9	2,3			
$F$	<u>4,0</u>	<u>0,3</u>	<u>7,4</u>		<u>4,9</u>		—		—		228	3,0	2,6	2,1	1,8			
	$F = \frac{\sigma_i^2}{\sigma_A^2}$																	



При планировании двухфакторных дисперсионных комплексов следует учесть их существующую классификацию (табл. 9.10).

Таблица 9.10

**Структура двухфакторных дисперсионных комплексов**

	Ортогональные								Неравномерные			
	Равномерные				Пропорциональные				Имеют внутриком- плексные корреляции			
	Без внутрикомплексных корреляций											
	$A_1$		$A_2$		$A_1$		$A_2$		$A_1$		$A_2$	
	$B_1$	$B_2$	$B_1$	$A_2$	$B_1$	$B_2$	$B_1$	$B_2$	$B_1$	$B_2$	$B_1$	$B_2$
$n$	15	15	15	15	5	15	10	30	10	10	10	30
$B_1/B_2$	1	1	1	1	1	3	1	3	1	1	1	3

Исследуемые факторы должны быть независимыми друг от друга, *например*; температура и влажность, возраст и пол. Отбор дат для исследования должен производиться рандомизированно.

**9.5. Примеры заданий для самостоятельного изучения раздела**

1. Исследовалась доля хромосомных aberrаций в клетках меристемы корешков в зависимости от высоты местности у трех видов *Crepis* L. (скерда). Фактор  $A$  – высота местности  $A_1$  – Астрахань,  $A_2$  – Кавказ (Нальчик). Результативный признак – доля хромосомных aberrаций в клетках меристемы зародышевого корня *Crepis capillaris*.

Фактор  $A$  – высота местности

$A_1$  – Астрахань

$A_2$  – Кавказ

Фактор  $B$  – вид креписа

$B_1$  – вид I

$B_2$  – вид II

$B_3$  – вид III

Результативный признак – доля хромосомных aberrаций в анафазе.

Исходные данные:

	I вид		II вид		III вид	
	$n$	$n'$	$n$	$n'$	$n$	$n'$
$A_1$	50	15	40	20	10	8
$A_2$	10	1	10	2	80	48

Примечание:  $n$  – общее число исследованных анафаз;  $n'$  – число анафаз с aberrациями.

Установить, влияние высоты местности на частоту возникновения хромосомных перестроек по каждому виду.

Определить силу и достоверность влияния факторов:  $A$ ,  $B$ ,  $AB$ .

2. В течение суток проводили 4 серии определений содержания каротиноидов в листьях лотоса орехоносного (*Nelumbo nucifera*). Определить

силу и достоверность влияния времени суток на содержание каротиноидов в листьях лотоса:

Время суток, ч	Номера проб									
	1	2	3	4	5	6	7	8	9	10
06.00	1,48	1,42	1,58	1,67	1,50	1,80	1,35	1,36	1,67	1,49
12.00	1,43	1,38	1,47	1,33	1,22	1,35	1,10	1,08	1,34	1,11
18.00	1,42	1,30	1,68	1,59	1,49	1,62	1,36	1,26	1,58	1,66
24.00	1,45	1,38	1,49	1,71	1,54	1,57	1,34	1,32	1,66	1,39

5. В течение суток изучали содержание каротиноидов в листьях шелковицы черной (*Morus nigra* L.). Результаты оказались следующими:

Время суток, ч	Номера проб									
	1	2	3	4	5	6	7	8	9	10
06.00	0,58	0,59	0,66	0,46	0,47	0,56	0,70	0,60	0,60	0,64
12.00	0,70	0,57	0,67	0,87	0,52	0,66	0,64	0,59	0,65	0,66
18.00	0,60	0,64	0,69	0,52	0,65	0,52	0,58	0,63	0,69	0,48
24.00	0,61	0,72	0,72	0,58	0,46	0,52	0,72	0,59	0,71	0,79

Определить силу и достоверность влияния изменения условий в разное время суток на содержание каротиноидов в листьях шелковицы черной.

4. При исследовании содержания хлорофилла  $a$  (мг на 100 г сырой массы) в листьях лофанта анисового (*Lophanthus anisatus* Benth.) в разное время суток были получены следующие результаты:

Время суток, ч	Номера проб				
	1	2	3	4	5
06.00	2,76	1,26	1,46	1,30	1,31
09.00	2,78	2,70	2,49	1,66	2,71
12.00	2,41	3,22	1,90	2,00	1,93
15.00	3,06	2,88	2,83	2,41	3,33
18.00	3,20	2,97	2,50	3,03	3,00
21.00	1,82	1,73	1,33	2,25	1,39
24.00	1,67	1,26	1,52	1,3	1,22

Определить силу и достоверность влияния времени суток на содержание хлорофилла  $a$  в листьях лофанта анисового.

5. Исследовали плодовитость самок дрозофилы (*Drosophila melanogaster*) в условиях рентгеновского облучения. Получены следующие данные о плодовитости самок дрозофилы:

Варианты	Число потомков за одну кладку яиц на одну пробирку			
Контроль	100	120	110	100
Доза 100 р	80	100	78	98
Доза 200 р	77	91	67	49

Определить силу и достоверность влияния рентгеновского облучения на плодовитость самок дрозофилы.

6. Исследовали зависимость годовых удоев коров черно-пестрой породы от количества отёлов. Годовые удои (литров) отдельных коров распределились в зависимости от количества отелов следующим образом:

№	Количество отелов	Годовые удои отдельных коров				
		1	2	3	4	5
1	1	2238	2364	2310	2314	2361
2	2	2462	2381	2236	2327	2239
3	3	2381	2472	2415	2389	2428
4	4	2430	2375	2402	2405	2370
5	5	2504	2471	2371	2400	2628
6	6	2439	2508	2439	2784	2538
7	7	2115	2290	2230	2231	22820

Определить силу и достоверность влияния количества отелов на годовые удои коров.

7. При исследовании содержания углеводов в созревающем зерне ржи было отмечено (В.Л Кретович. 1971) высокая динамичность этих процессов (табл. 9.5.7.1).

Таблица 9.5.7.1

**Превращение углеводов в созревающем зерне ржи (по А.Р. Кизелю и В.Л. Кретовичу, 1971)**

Углеводы	Содержание, в % от сухого вещества по данным на:			
	25 июня	3 июля	15 июля	28 июля
Моносахариды	6,1	2,1	0,4	2,1
Сахароза	6,0	4,4	3,1	2,8
Леволёзаны	31,8	12,2	3,0	0,4
Мальтоза	0,0	0,0	0,0	0,0
Крахмал	9,0	25,9	37,5	41,2
Гемицеллюлозы	5,7	12,8	16,2	17,5
Клетчатка	2,0	2,0	2,0	2,4

Проанализируйте полученные результаты методом дисперсионного анализа. Определите силу и достоверность влияния сроков отбора проб на содержание углеводов в созревающем зерне ржи.

8. На опытной станции проводились опыты по изучению влияния внесения фосфорных удобрений на урожай томатов сорта Волгоградский 5/95:

Варианты		Урожай на делянках в пересчете ц/га							
1	Контроль, без удобрений	35	33	31	37	32	35	30	
2	Двойной суперфосфат (46–49 % P <sub>2</sub> O <sub>5</sub> )	43	48	54	49	51	55	42	

3	Простой суперфосфат (16–20 % $P_2O_5$ )	36	31	42	36	38	44	43
4	Щелочной плавный фосфат (25 % $P_2O_5$ )	37	34	40	31	33	40	33
5	Магнийевый плавный фосфат (20 % $P_2O_5$ )	31	31	40	35	34	40	41

Проанализируйте полученные результаты методом дисперсионного анализа. Определите силу и достоверность влияния варианта внесения фосфорных удобрений на урожай томатов сорта Волгоградский 5/95.

9. Исследовали живой вес (масса тела) новорожденных ягнят (кг), выношенных разное количество суток:

Продолжительность беременности, сут	Живой вес (масса тела) отдельных ягнят, кг									
	1	2	3	4	5	6	7	8	9	10
145	4,1	5,1	3,5	2,8	4,2	4,1	4,0	3,9	4,6	3,5
146	4,2	4,4	4,0	2,9	4,1	4,2	4,4	4,1	4,0	5,1
147	4,1	5,0	2,8	3,9	4,2	4,3	4,4	4,1	4,1	5,1
148	4,4	5,7	3,9	4,5	4,4	4,3	3,8	4,1	4,5	4,4
149	4,3	5,6	3,0	3,9	4,1	4,2	4,3	4,7	4,5	4,4
150	4,5	5,0	5,2	4,6	4,3	3,0	4,7	4,6	4,0	5,1
151	4,6	5,3	5,5	4,4	4,3	3,2	4,0	4,5	5,0	5,2
152	4,6	5,4	6,1	4,8	4,4	3,2	4,8	4,7	4,0	4,2
153	4,8	5,5	5,2	4,9	4,5	3,4	4,9	4,4	5,1	5,3

Определить силу и достоверность влияния продолжительности беременности овец на живой вес ягнят.

### **9.6. Возможности многомерного анализа в биологии**

При системном анализе явлений в биологических совокупностях исследователь нередко сталкивается с проблемой многомерности их описания. Это типично для статистической обработки информации в области биологии, психологии (Глас, Стэнли, 1976), техники (Пугачев, 1979), экономики (Дубров, 2003) и др. при анализе объектов по большому числу признаков. Методы многомерного анализа наиболее эффективный количественный метод исследования биологических явлений, описываемых большим числом характеристик. К ним относятся: метод ранжирования многомерных величин, кластерный анализ, таксономия, факторный анализ. Возможно, дисперсионный анализ не является элементом этой группы методов, однако в нем заложены сходные с перечисленными методами идеи (Орехов и др., 2004). Кластерный анализ наиболее полно отражает черты многомерного анализа в классификации, факторный анализ в исследовании связей. Дисперсионный анализ ориентирован на выявление зависимости изменения числовых характеристик наблюдаемой величины от некоторого качественного влияния. Анализ и обработка больших числовых массивов статистического исследования не мыслим без

использования компьютерных программ. Существует значительное количество программ, специализированных для решения таких задач. К ним относятся пакеты “Stadia” и “Olymp”, SAS, “Statgraphics”, SPSS, “Statistica”, “S-Plus” и др. Перечень специализированных программ значительно шире. Методам многомерного анализа посвящено большое количество рекомендаций и разработок. Это касается математического обоснования методов и их компьютерной реализации (например: Боровиков, 2003; Вуколов, 2004; Дубнов, 2004; Дюк, 1998 и др.). В компактной и удобной форме изложены математические основы методов многомерного анализа в пособии М.Г. Близорукова (2008), где приводятся их разновидности, особенности применения. В учебно-методическом пособии представлена последовательная реализация многомерных методов в рамках пакетов SPSS, “Statistica” и “Stadia” (в версиях SPSS 11.5, “Statistica 6.0”, “Stadia 6.2”). Выбор этих программ связан с их популярностью в различных областях деятельности, доступностью этих пакетов в версиях SPSS 11.5, “Statistica 6.0” и “Stadia 6.2”. Следующим этапом является самостоятельная реализация однофакторного дисперсионного анализа средствами “Excel” и др. (в пакете “Stadia”, “ANOVA” – реализация дисперсионного анализа в пакете “Statistica”, изучение особенностей реализации дисперсионного анализа в пакете SPSS). Далее, в зависимости от направления и целей исследования: многофакторный дисперсионный анализ, компьютерная реализация двухфакторного дисперсионного анализа.



## ЗАКЛЮЧЕНИЕ

### *Взаимосвязь методов исследования в современной биологии*

Методы изучения явлений в биологических совокупностях можно условно разделить на три группы: эмпирические, экспериментальные и теоретические. При эмпирическом подходе проводят обширные исследования в течение определенного времени. При этом учитывают такие факторы, как влияние среды на изменчивость признаков организмов в группе. Эти данные могут отражать связь между уровнем и характером генетической изменчивости и другими факторами воздействия. Гипотезы, вытекающие из эмпирических данных, нуждаются в экспериментальной проверке. В таких экспериментах по изучению влияния среды на популяцию (или другую группу организмов) данную популяцию перемещают в новую среду обитания и сравнивают с местной популяцией. Используя данные эмпирических и экспериментальных исследований, можно построить обобщенную теоретическую модель, учитывающую все эти наблюдения. Такие теоретические модели служат основой при исследовании сходных явлений и способствуют пониманию влияния различных факторов на уровень и характер генетической и фенотипической изменчивости.

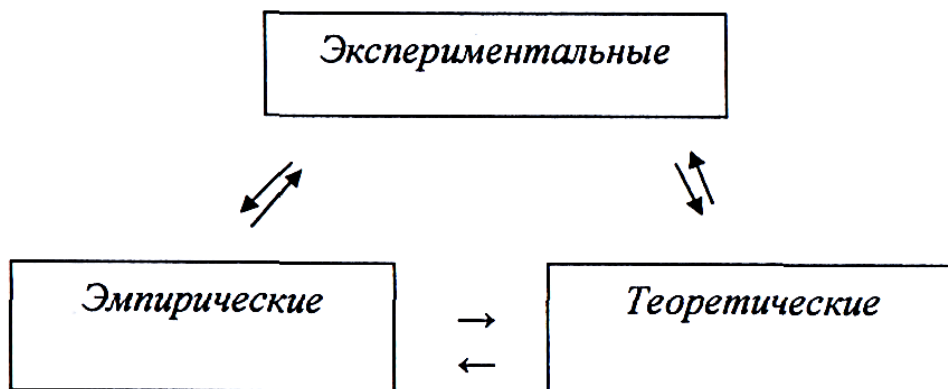


Рис. 3.1. Взаимосвязь между эмпирическими, экспериментальными и теоретическими методами исследования

Взаимосвязи между этими тремя группами методов исследований в популяционной генетике, эволюционной биологии в целом иллюстрирует рисунок 3.1. Как правило, эмпирическая информация поступает первой. Затем развивается гипотеза, объясняющая полученные результаты. После этого проводятся эксперименты, подтверждающие или опровергающие данную гипотезу. Теоретическая модель разрабатывается для объяснения эмпириче-

ских наблюдений, а затем накапливаются данные в пользу биологического соответствия данной модели. Обратная связь между этими тремя подходами позволяет развивать и совершенствовать конкретную гипотезу. Во многих случаях каждое отдельное явление в биологических совокупностях обусловлено множеством причин, и это затрудняет планирование и выполнение экспериментов. При разработке и выполнении экспериментов биологу-исследователю приходится учитывать факторы, лежащие в основе экспериментальной работы:

1. Для чистоты эксперимента необходимо соответствующее количество независимых повторных экспериментов.

2. Одновременно нужно провести адекватные контрольные эксперименты. Следует убедиться в том, что повторные эксперименты действительно независимы, а контрольные эксперименты подходят к данной задаче.

3. Количество наблюдений должно быть достаточным для того, чтобы вероятность статистических отклонений была мала. При определении числа повторных экспериментов и размеров выборки, необходимых для обнаружения конкретного воздействия, надо вычислить уровень статистической силы – вероятности отклонения нулевой гипотезы, если она ложная (Sokal, Rohlf, 1995; Zar, 1999). Эти факторы следует учитывать даже при сравнении последовательностей ДНК у различных видов организмов. Например, родственны или независимы исследуемые последовательности, только ли сравниваемые группы последовательностей имеют интересующие исследователя различия, какова статистическая сила полученных в результате такого сравнения данных? «Философский подход к пониманию научных явлений выглядит привлекательным, простым и, как показано, продуктивным, но распространять его на любые ситуации следует с осторожностью» (Хедрик, 2003). Многие фундаментальные закономерности впервые были получены на модельных организмах в сериях контролируемых экспериментов. К ним относятся законы расщепления и независимого комбинирования признаков, открытые и продемонстрированные Г. Менделем на горохе, а также спонтанные мутации, впервые обнаруженные у *Drosophila melanogaster* Х.Д. Мёллером. Экспериментальные данные, полученные на модельных организмах в лабораторных или в искусственно созданных условиях, могут не отражать сложные и важные факторы, действующие в естественных природных системах. С другой стороны, при слишком строгом эксперименте с множеством ограничений можно получить лишь теоретическую модель, которая подтверждает наши наблюдения и не дает исчерпывающей информации о природных процессах. Однако при опытной проверке того или иного явления в естественных, не контролируемых условиях результаты бывают не достаточно четкими для выбора альтернативных гипотез. Так, при проведении традиционных экспериментов дей-

ствие отбора на молекулярную изменчивость может быть слишком слабым. Поэтому такое действие можно обнаружить, только используя обширные данные по кумулятивному действию отбора на протяжении нескольких поколений. Используя данные эмпирических и экспериментальных исследований, можно построить обобщенную теоретическую модель, учитывающую все эти наблюдения. Такие теоретические модели служат основой при исследовании сходных явлений и способствуют пониманию влияния различных факторов на уровень и характер генетической изменчивости. Для описания концепций современной биологии используются различные модели. *Модель* представляет собой словесное (вербальное), графическое или математическое описание реальных событий. Преимущество модели перед простым описанием в том, что она охватывает как сам процесс, так и его составные части. Модель должна соответствовать описанию природных процессов и согласовываться с реальными наблюдениями. Однако приближенные математические модели нередко дают неточную картину биологических явлений. Степень детальности модели отчасти зависит от интересов исследователей: интересуется их общие принципы или же точное описание частных случаев. Во многих случаях модели могут быть стохастическими, или случайными. В модель можно ввести факторы, которые происходят с некоторой вероятностью или имеют переменную величину. В результате введения изменчивых параметров нет возможности точно прогнозировать будущий результат, как это можно сделать при точно заданных параметрах. Кроме того, могут возникнуть и нестандартные или уникальные факторы, которые не описаны в основной модели.

Теоретическая модель разрабатывается для объяснения эмпирических наблюдений, а затем накапливаются данные в пользу биологического соответствия данной модели. Обратная связь между этими тремя подходами позволяет развивать и совершенствовать конкретную гипотезу. Для расширения научных познаний необходимы альтернативные гипотезы, объясняющие конкретные наблюдения, а затем критические эксперименты. Такие продуманные и тщательно выполненные эксперименты с альтернативными результатами позволяют исключить одну или более из предлагаемых гипотез. Затем оставшиеся гипотезы согласуют с экспериментальными или эмпирическими данными. Из-за трудности построения критических экспериментов сложно оставить только одну гипотезу. Во многих случаях отдельное явление в биологических совокупностях обусловлено множеством (или несколько) причин, и это затрудняет планирование и проведение экспериментов. В настоящее время биометрия, как дисциплина, как область биологической науки, приобрела основательность, ее методы стали более совершенными. В области биометрии в конце 1990-х гг. отдельные близкие научные направления постепенно объединились в единую дисциплину. Биометрия стала областью

исследований, она имеет не только научные, технические, прикладные, но и социальные аспекты. Научное общество постепенно приходит к пониманию ее роли и значения.

Мы отдали предпочтение разделам, биометрии, которые являются общими для различных направлений биологической науки. На протяжении последних пятидесяти лет активно развиваются технологии распознавания лица и голоса по идентификационным признакам. Сегодня объединяются многие актуальные направления биометрии, которые появились в рамках других дисциплин. Научно-технический прогресс, исследования в области медицины и биологии, в дальнейшем приведут к тому, что в будущем будет все шире применяться биометрическая идентификация. Почти любая физическая особенность человеческого организма (химический состав, плотность, отражение, поглощение, выделение) может выступать в качестве биометрического параметра, если есть достоверные результаты ее измерения, особенно, если эту особенность рассматривать как пространственную переменную или повторяющийся сигнал. Технологический прогресс будет способствовать распространению использования биометрических параметров. В будущем автоматизированные биометрические системы будут использоваться для крупных аутентификационных приложений. Однако биометрическая аутентификация не может гарантировать стопроцентную достоверность принимаемого решения. Лишь автоматизированная система, использующая не биометрические удостоверяющие данные, при помощи точного сопоставления может гарантировать абсолютную достоверность результатов.

## ТЕРМИНОЛОГИЧЕСКИЙ СЛОВАРЬ

**Биометрия** – наука о статистическом анализе групповых свойств биологических объектов.

**Биометрии предмет** – 1. Любой биологический объект, изучаемый с количественной стороны с целью оценки его качественного состояния с определенной точностью и надежностью; 2. Статистическая совокупность.

**Больших чисел закон** – частность любого события будет сколь угодно близкой к его вероятности, если число испытаний неограниченно возрастает.

**Варианса** – средний квадрат, равен сумме квадратов центральных отклонений, деленной на число степеней свободы; это дисперсия, отнесенная к числу степеней свободы.

**Варианты** – отдельные количественные значения варьирующего признака.

**Варьирование** – колебания величины определенного признака, наблюдаемые в статистической совокупности.

**Выборка** – результат конечного числа измерений в некоторой идеальной совокупности, состоящей из бесконечного числа измерений.

**Дисперсионный анализ** – метод сравнения параметров нескольких распределений. Сравнение средних производится путем разложения на компоненты общей дисперсии.

**Дисперсия (Д.)** – наличие разнообразия в группе. Это мера, которая определяет степень данного разнообразия, это сумма квадратов центральных отклонений вариантов от средней арифметической и других, это характеристика размаха изменчивости признака, «ширины» кривой плотности распределения.

**Д. общая** – равна сумме квадратов центральных отклонений дат от общей средней. Это дисперсия дат около общей средней.

**Д. случайная (внутригрупповая)** – равна сумме квадратов центральных отклонений дат от частных средних по градациям дисперсионного комплекса.

**Д. факториальная (межгрупповая)** – равна сумме взвешенных квадратов центральных отклонений частных средних от общей средней.

**Достоверность** – соответствие выборочных показателей генеральным параметрам.

**Достоверности критерии (Д.к.)** – величины, функции распределения которых известны. Это величины, позволяющие установить, удовлетворяют ли выборочные показатели принятой гипотезе.

**Д.к. параметрические** – критерии, построенные на основе параметров данной совокупности и представляющие функции этих параметров. Имеют предпочтение при нормальном распределении сравниваемых совокупностей.

**Д.к. непараметрические** – критерии, представляющие функции, зависящие от вариантов данной совокупности с их частотами. Предпочтительны при больших отличиях распределений признака от нормального.

**Критерий Стьюдента (t-критерий)** – параметрический критерий, используемый для сравнительной оценки средних величин.

**Критерий Фишера (F-критерий)** – используется для сравнительной оценки дисперсий нормально распределяющихся генеральных совокупностей.

**Нулевая гипотеза ( $H^0$ )** – проверяемая гипотеза, в результате которой может быть получено согласие с ней, тогда нулевая гипотеза будет принята. Если эмпирические результаты не согласуются с теоретически ожидаемыми, то нулевая гипотеза отвергается и будет принята альтернативная ( $H_a$ ) гипотеза, которая формулируется заранее, одновременно с  $H^0$ .

**Рандомизация** – случайный отбор вариант из генеральной совокупности, обеспечивающий равную возможность для любого члена совокупности быть включенным в состав выборки.

**Распределение (P.)** – явление неодинаковой вероятности встречаемости вариант в совокупности:

**P. нормальное, асимметричное, эксцессивное** – распределение единичных объектов, попадающих в разные классы.

**P. биномиальное** – распределение групп, одинаковых по числу объектов, изучаемых на предмет наличия или отсутствия определенного качественного признака.

**P. Пуассона** – распределение больших групп по встречаемости в них редких событий.

**Случайная переменная величина** – 1) дискретная – альтернативы, которые можно занумеровать, т.к. результаты их исследования распадаются на отдельные легко различимые классы (длинный – короткий, черный – белый, гладкий – морщинистый); 2) недискретная – бесконечное множество возможных результатов: вес, рост, умственное развитие, длина...; 3) величина любого варьирующего признака.

**Случайных величин закон распределения** – функция  $f(x)$ , связывающая значения  $x_1$  случайной переменной величины  $x$ , с их вероятностями.

**Совокупность (C.)** – статистическое множество относительно однородных, но индивидуально различных единиц, объединенных для совместного исследования:

**С. генеральная** – идеальная совокупность, состоящая из бесконечного числа измерений, это бесконечно большая совокупность (или приближающаяся к бесконечности) совокупность всех единиц или элементов, которые могут быть к ней отнесены.

**С. статистическая** – множество, для описания которого используется статистический метод.

**С. стохастическая** – теоретически мыслимая совокупность, совокупность всех возможных наблюдений, проведение которых не всегда возможно.

**С. выборочная** – см. выборка.

**Среднее квадратическое отклонение (сигма)** – основной показатель степени варьирования значений признака в группе, дающей значение среднего размаха варьирования.

**Теорема** – центральная предельная: если случайная величина представляет собой сумму большого числа независимых случайных величин, влияние каждой из которых ничтожно мало, то это величина будет иметь распределение, близкое к нормальному.

**Точности оценок показатель** – отношение ошибки репрезентативности к своей средней.

**Уровень значимости** – та вероятность ошибки, которой решено пренебречь в данном исследовании при заданном пороге вероятности безошибочного прогноза.

**Условной средней (способ)** – вычисление статистических характеристик упрощенным способом, при котором за среднюю величину принимают одну из вариантов, обычно имеющую наибольшую частоту встречаемости.

**Хи-квадрат (критерий К. Пирсона)** – представляет собой сумму квадратов отклонений эмпирических частот  $f$  от вычисленных или ожидаемых частот отнесенную к теоретическим частотам.

## СПИСОК ЛИТЕРАТУРЫ

1. Бейли, Н. Т. Дж. Математика в биологии и медицине : пер. с англ. / Н. Т. Дж. Бейли. – М. : Мир, 1970. – 326 с.
2. Биометрика. – Режим доступа: <http://www.biometrica.tomsk.ru/lib2.htm>, свободный. – Заглавие с экрана. – Яз. рус.
3. Близоруков, М. Г. Статистические методы анализа рынка : учеб.-метод. пос. / М. Г. Близоруков. – Екатеринбург : Ин-т управления и предпринимательства Уральского гос. ун-та, 2008. – 75 с.
4. Болл, Р. М. Руководство по биометрии / Р. М. Болл [и др.] ; пер. с англ. Н. Е. Агаповой. – М. : Техносфера, 2007. – 368 с. – ISBN 978-5-94836-109-3: 102-75. – (Мир цифровой обработки).
5. Боровиков, В. STATISTICA: искусство анализа данных на компьютере. Для профессионалов / В. Боровиков. – СПб : Питер, 2003. – 688 с.
6. Ван-дер-Варден, Б. Л. Математическая статистика / Б. Л. Ван-дер-Варден ; пер. с нем. Л. Н. Большева ; под ред. Н. В. Смирнова. – М. : Иностранная лит-ра, 1960. – 431 с.
7. Васильева, Л. А. Статистические методы в биологии : учеб. пос. / Л. А. Васильева. – Новосибирск : ИЦиГ СО РАН, 2009. – 128 с.
8. Вуколов, Э. А. Основы статистического анализа / Э. А. Вуколов // Практикум по статистическим методам и исследованию операций с использованием пакетов STATISTICA и EXCEL. – М. : ИНФРА. М, 2004.
9. Геронтология in silico: Становление новой дисциплины: математические модели, анализ данных и вычислительные эксперименты : сб. науч. тр. / под ред. Г. И. Марчука, В. Н. Анисимова, А. А. Романюхин, А. И. Яшина. – М. : Бином. Лаборатория знаний, 2007. – 535 с.
10. Гершкович, И. Генетика / И. Гершкович. – М. : Наука, 1968. – 702 с.
11. Гланц, С. Медико-биологическая статистика / С. Гланц ; пер. с англ. д-ра физ.-мат. наук Ю. А. Данилова ; под ред. Н. Е. Бузикашвили и Д. В. Самойлова. – М. : Практика, 1998. – 459 с.
12. Глас, Дж. Статистические методы в педагогике и психологии / Дж. Глас, Дж. Стэнли. – М. : Прогресс, 1976.
13. Глотов, Н. В. Биометрия : учеб. пос. / Н. В. Глотов, Л. А. Животовский, Н. В. Хованов, Н. Н. Хромов-Борисов ; под ред. М. М. Тихомировой. – Л. : Ленинградский ун-т, 1982. – 264 с.
14. Динамика популяционных генофондов при антропогенных воздействиях / под ред. Ю. П. Алтухова. – М. : Наука, 2004. – 619 с.
15. Дубров, А. М. Многомерные статистические методы / А. М. Дубров. – М. : Финансы и статистика, 2003.



16. Дюк, В. Обработка данных на ПК в примерах / В. Дюк. – СПб : Питер, 1997. – 240 с.
17. Дюран, Б. Кластерный анализ / Б. Дюран, П. Одел. – М. : Статистика, 1977.
18. Зайцев, В. М. Прикладная медицинская статистика / В. М. Зайцев, В. Г. Лифляндский, В. И. Маринкин. – СПб : Фолиант, 2003. – 432 с. – Режим доступа: <http://www.booksmed.com/zdravooxranenie/2120-prikladnaya-medicinskaya-statistika-zajcev-uchebno-prakticheskoe-posobie.html>, свободный. – Заглавие с экрана. – Яз. рус.
19. Закс, Л. Статистическое оценивание / Л. Закс ; пер. с нем. В. И. Варыгина ; под ред. Ю. П. Адлера, В. Г. Горского. – М. : Статистика, 1976. – 598 с.
20. Иберла, К. Факторный анализ / К. Иберла. – М. : Статистика, 1980. – 398 с.
21. Козак, М. Ф. Биометрия : учеб. пос. / М. Ф. Козак. – Астрахань : Астраханский пед. ин-т, 1995. – 160 с.
22. Козак, М. Ф. Генетика. Полевая практика : метод. рекомендации / М. Ф. Козак. – Астрахань : Астраханский ун-т, 2006. – 15 с.
23. Козак, М. Ф. Дрозофила – модельный объект генетики / М. Ф. Козак. – Астрахань : Астраханский ун-т, 2007. – 87 с.
24. Козлов, Н. Н. Математический анализ генетического кода / Н. Н. Козлов. – М. : Бином. Лаборатория знаний, 2010. – 215 с.
25. Компьютерная биометрика / [Ю. М. Барабашева, Г. Н. Девяткова, Н. Г. Микешина и др.] ; под ред. В. Н. Носова. – М. : МГУ, 1990. – 232 с.
26. Кретович, В. Л. Основы биохимии растений : учеб. / В. Л. Кретович. – Изд. 5-е, испр. и доп. – Москва : Высшая школа, 1971. – 464 с.
27. Крюков, В. И. Статистические методы изучения изменчивости / В. И. Крюков. – Орёл : Орёл-ГАУ, 2006. – 208 с.
28. Лакин, Г. Ф. Биометрия / Г. Ф. Лакин. – М. : Высшая школа, 1990. – 352 с.
29. Макаров, А. А. Статистический анализ данных на компьютере / А. А. Макаров, Ю. Н. Тюрин ; ред. В. Э. Фигурнов. – М. : ИНФРА, 1998. – 528 с.
30. Медик, В. А. Математическая статистика в медицине : учеб. пособие / В. А. Медик, М.С. Токмачев. – М. : Финансы и кредит, 2007. – 800 с.
31. Мюррей, Дж. Математическая биология / Мюррей, Дж. ; пер. с англ. Л. С. Ванаг, А. Н. Дьяконовой ; под ред. Г. Ю. Ризниченко. – М. – Ижевск : Регулярная и хаотическая динамика. Ин-т компьютерных исследований, 2009. – Т. 1. Введение. – 776 с.
32. Мюррей, Дж. Математическая биология. / Мюррей, Дж. Пер. с англ. А. Н. Дьяконовой [и др.] ; под ред. Г. Ю. Ризниченко. – М. – Ижевск : Регулярная и хаотическая динамика, Институт компьютерных исследований, 2011. –

Т. 2. Пространственные модели и их приложения в биомедицине. – 1104 с. – (Биофизика. Математическая биология).

33. Орехов, Н. А. Математические методы и модели в экономике / Н. А. Орехов, А. Г. Левин, Е. А. Горбунова. – М. : ЮНИТИ, 2004.

34. Плохинский, Н. А. Алгоритмы биометрии / Н. А. Плохинский. – М. : МГУ, 1980. – 81 с.

35. Плохинский, Н. А. Биометрия / Н. А. Плохинский. – Новосибирск : СО АН СССР, 1961. – 367 с.

36. Плохинский, Н. А. Дисперсионный анализ / Н. А. Плохинский ; под ред. Н. П. Дубинина ; СО АН СССР. – Новосибирск : СО АН СССР, 1960. – 121 с.

37. Плохинский, Н. А. Математические методы в биологии : учебно.-методическое. пос. / Н. А. Плохинский. – М. : МГУ, 1978. – 365 с.

38. Плохинский, Н. А. Руководство по биометрии для зоотехников / Н. А. Плохинский. – М. : Колос, 1969. – 352 с.

39. Прикладная медицинская статистика : учеб. пос. / В. М. Зайцев, В. Г. Лифляндский, В. И. Маринкин. – Изд. 2-е. – СПб : Фолиант, 2006. – 432 с.

40. Пугачев, В. С. Теория вероятностей и математическая статистика / В. С. Пугачев. – М. : Наука, 1979 ; Физматлит, 2002.

41. Ратнер, В. А. Математическая генетика как наука / В. А. Ратнер // Известия РАН. Сер. Биологическая. – 1993. – № 2. – С. 323–327.

42. Ригер, Ю. Р. Генетический и цитогенетический словарь : пер. с нем. / Ю. Р. Ригер, А. Михаэлис ; под ред. д-ра биол. наук Я. Л. Глембоцкого, акад. Белорусской АН П. Ф. Рокицкого. – М. : Колос, 1967. – 607 с.

43. Ризниченко, Г. Ю. Лекции по математическим моделям в биологии : учеб. пос. / Г. Ю. Ризниченко. – изд. 2-е, испр. и доп. – Ижевск : Регулярная и хаотическая динамика, 2011. – 560 с.

44. Рокицкий, П. Ф. Биологическая статистика / П. Ф. Рокитский. – М. : Высшая школа, 1978. – 319 с.

45. Романовский, Ю. М. Математическое моделирование в биофизике. Введение в теоретическую биофизику / Ю. М. Романовский, Н. В. Степанова, Д. С. Чернавский. – 2-е изд., доп. – М. – Ижевск : Ин-т компьютерных исследований, 2004. – 472 с.

46. Рубин, А. Б. Кинетика биологических процессов : учеб. пос. / А. Б. Рубин, Н. Ф. Питьева, Г. Ю. Ризниченко. – 2-е изд., испр. и доп. – М. : Московский гос. ун-т, 1987. – 299 с.

47. Славин, М. Б. Методы системного анализа в медицинских исследованиях / М. Б. Славин. – М. : Медицина, 1989. – 303 с.

48. Снедекор, Дж. У. Статистические методы в применении к исследованиям в сельском хозяйстве и биологии / Дж. У. Снедекор. – М. : Сельхозгиз, 1964. – 503 с.

49. Статистический анализ данных в экологии и природопользовании (с использованием программы STATGRAPHICS Plus) : учеб.-методич. пос. / К. А. Мальцев, С. С. Мухарамова. – Казань, 2011. – 50 с.
50. Терентьев, П. В. Практикум по биометрии : учеб. пос. / П. В. Терентьев, Н. С. Ростова. – Л. : Ленинградский гос. ун-т, 1977. – 152 с.
51. Трухачева, Н. В. Математическая статистика в медико-биологических исследованиях с применением пакета Statistica / Н. В. Трухачева. – М. : ГЭОТАР-Медиа, 2012. – 379 с.
52. Урбах, В. Ю. Статистический анализ в биологических и медицинских исследованиях / В. Ю. Урбах. – М. : Медицина, 1975. – 295 с.
53. Филипченко, Ю. А. Изменчивость и методы ее изучения / Ю. А. Филипченко. – Изд. 4-е. – М. – Л. : Госиздат, 1929. – 275 с.
54. Фишер, Р. Э. Статистические методы для исследователей / Р. Э. Фишер. – М. : Наука, 1958. – 268 с.
55. Фролов, А. В. Методы описательной статистики в Excel. : метод. рекомендации / А. В. Фролов // Статистические методы в управлении качеством. – Бийск : Алтайский гос. тех. ун-т им. И.И. Ползунова, 2015.
56. Харман, Г. Современный факторный анализ. Финансы и экономика / Г. Харман. – М. : Статистика, 1972. – 489 с.
57. Хедрик, Ф. В. Генетика популяций : пер. с англ. / Ф. В. Хедрик. – М. : Техносфера, 2003. – 592 с. – (Мир биологии).
58. Хёлыгье, Х. Д. Молекулярное моделирование. Теория и практика : пер. с англ. / Х. Д. Хёлыгье, В. Зиппл, Д. Роньян, Г. Фолькерс ; под ред. В. А. Палюлина и Е. В. Радченко. – М. : БИНОМ. Лаборатория знаний, 2009.
59. Четвериков, С. С. Проблемы общей биологии и генетики (Воспоминания, статьи, лекции) / С. С. Четвериков ; отв. ред. З.С. Никоро. – Новосибирск : Наука, 1983. – 273 с.
60. Юл, Д. Э. Теория статистики / Д. Э. Юл, М. Д. Кендэл. – М. : Госстатиздат ЦСУ СССР. 1960. – 780 с.
61. Sokal, R. Introduction to Biostatistics / R. Sokal, F. Rohlf. – 2<sup>nd</sup> ed. – Dover Publ., 2009. – 374 p. – ISBN 0486469611, 9780486469614.
62. Sokal, R. R. Biometry. The principles and practice of statistics in biological research / R. R. Sokal, F. J. Rohlf. – 3<sup>rd</sup> ed. – New York : W. H. Freeman, 1995. – 887 p.
63. Sokal, R. R. Biometry: the principles and practice of statistics in biological research / R. R. Sokal, F. J. Rohlf. – New York, 1969. – 776 p.
64. StatSoft. Электронный учебник по статистике. – Режим доступа: <http://statsoft.ru/home/textbook/default.htm>, свободный. – Заглавие с экрана. – Яз. рус.

65. "Student". The probable Error of Mean. // Biometrical. – 1908. – Vol. 6. – P. 1–25.
66. TIBCO Products. – Режим доступа: <https://www.tibco.com/products/>, свободный. – Заглавие с экрана. – Яз. англ.
67. XLSTAT. – Режим доступа: <http://www.xlstat.com/>, свободный. – Заглавие с экрана. – Яз. англ.
68. Zar, J. H. Biostatistical analysis / J. H. Zar. – 4<sup>th</sup> ed. – New Jersey : Prentice-Hall Inc., Englewood Cliffs, 1999. – 663 p.

# **ПРИЛОЖЕНИЕ** **МАТЕМАТИЧЕСКИЕ ТАБЛИЦЫ**

Таблица 1

**Значение интеграла вероятностей для разных значений  $t$**

$t$	Сотые доли $t$									
	0	1	2	3	4	5	6	7	8	9
0,0	0000	0080	0160	0239	0319	0399	0478	0558	0638	0717
0,1	0797	0876	0955	1034	1114	1192	1271	1350	1428	1507
0,2	1585	1663	1741	1819	1897	1974	2051	2128	2205	2282
0,3	2358	2434	2510	2586	2661	2737	2813	2886	2961	3035
0,4	3109	3182	3255	3328	3401	3478	3545	3616	3688	3759
0,5	3829	3899	3969	4039	4108	4177	4245	4313	4381	4448
0,6	4515	4581	4647	4713	4778	4843	4907	4971	5035	5098
0,7	5161	5223	5285	5346	5407	5467	5527	5587	5646	5705
0,8	5763	5821	5878	5935	5991	6047	6102	6157	6211	6265
0,9	6319	6372	6424	6476	6528	6579	6628	6679	6729	6778
1,0	6827	6875	6923	6970	7017	7063	7109	7154	7199	7243
1,1	7287	7330	7373	7415	7457	7499	7540	7580	7620	7660
1,2	7699	7737	7775	7813	7850	7887	7923	7959	7995	8030
1,3	8064	8098	8182	8165	8198	8230	8262	8293	8324	8355
1,4	8385	8415	8444	8473	8501	8529	8557	8584	8611	8638
1,5	8664	8690	8715	8740	8764	8788	8812	8836	8859	8882
1,6	8904	8926	8948	8969	8990	9011	9031	9051	9070	9089
1,7	9108	9127	9146	9164	9182	9199	9216	9233	9248	9165
1,8	9281	9297	9312	9327	9342	9357	9371	9385	9399	9412
1,9	9425	9439	9451	9464	9476	9488	9500	9512	9523	9534
2,0	9545	9556	9566	9576	9586	9596	9608	9615	9625	9634
2,1	9643	9652	9660	9668	9676	9684	9692	9700	9707	9715
2,2	9722	9729	9736	9743	9749	9755	9762	9768	9774	9780
2,4	9836	9840	9845	9849	9853	9857	9861	9866	9869	9872
2,6	9907	9909	9912	9915	9917	9920	9922	9924	9926	9929
2,8	9949	9950	9952	9953	9955	9956	9956	9959	9960	9961
3,0	9973	9974	9975	9976	9976	9977	9978	9979	9979	9980

*Примечание:* значения вероятности даны числами после запятой.

Таблица 2

Первая функция нормированного отклонения (ординаты нормальной кривой)

$$f_x = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}$$

	0	1	2	3	4	5	6	7	8	9
0,0	399	399	399	399	399	398	398	398	398	397
0,1	397	397	396	396	395	394	394	393	393	392
0,2	391	390	389	389	387	387	386	385	384	383
0,3	381	380	379	378	377	375	374	373	371	370
0,4	368	367	365	364	362	361	359	357	356	354
0,5	352	350	348	347	345	343	341	339	337	335
0,6	333	331	329	327	325	323	321	319	317	314
0,7	312	310	308	306	303	301	299	297	294	292
0,8	290	287	285	283	280	278	276	273	271	268
0,9	266	264	261	259	256	254	252	249	247	244
1,0	218	215	213	21	208	206	204	201	199	197
1,1	194	192	190	187	185	183	180	178	176	174
1,2	194	192	190	187	185	183	180	178	176	174
1,3	171	169	167	165	163	160	158	156	154	152
1,4	150	147	146	144	141	139	137	135	133	131
1,5	130	128	126	124	122	120	118	116	115	113
1,6	111	109	107	106	104	102	100	098	097	096
1,7	094	092	091	089	088	086	085	083	082	080
1,8	079	078	076	075	073	072	071	069	068	067
1,9	066	064	063	062	061	060	058	057	056	055
2,0	054	053	052	051	050	049	048	047	046	045
2,1	044	043	042	041	040	040	039	038	037	036
2,2	035	035	034	033	032	031	031	030	030	029
2,3	028	028	027	026	026	025	025	024	023	023
2,4	022	022	021	021	020	020	019	019	018	018
2,5	018	017	017	016	016	015	015	015	014	014
2,6	014	013	013	013	012	012	012	011	011	011
2,7	010	010	010	010	009	009	009	009	008	008
2,8	008	008	008	008	007	007	007	006	006	006
2,9	006	006	006	005	005	005	005	005	005	005
3,0	004	004	004	004	004	004	004	004	003	003
3,5	001	001	001	001	001	001	001	001	001	001

Примечание: знаки после запятой.

Таблица 3

Значение вероятности

$$P_{n(m)} = \frac{a_m}{m!} \cdot e^{-a}$$

$\alpha \backslash m$	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
0	9048	8187	7408	6703	6065	5488	4966	7793	4066	3979
1	0905	1637	2222	2681	3033	3293	3476	3595	3659	3679
2	0045	0164	0333	0536	0758	0988	1217	1438	1647	1839
3	002	0011	0033	0072	0126	0198	0284	0383	0494	0613
4	0000	0001	0003	0007	0016	0030	0050	0077	0111	0153
5				0001	0002	0004	0007	0012	0020	0031
6							0001	0002	0003	0005
7										0001
$\alpha \backslash m$	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2,0
0	3329	3012	2725	2466	2231	2019	1827	1653	1496	1353
1	3662	3614	3543	3452	3347	3230	3106	2975	2842	2707
2	2014	2169	2303	2417	2510	2584	2640	2675	2700	2707
3	0738	0867	0998	1128	1255	1378	1496	1607	1710	1804
4	0203	0260	0324	0395	0471	0551	0636	0723	0812	0902
5	0045	0063	0084	0111	0141	0176	0216	0260	0309	0361
6	0008	0013	0018	0026	0035	0047	0061	0078	0098	0120
7	0001	0002	0003	0005	0008	0011	0015	0020	0027	0034
8			0001	0001	0001	0002	0003	0005	0006	0009
9							0001	0001	0001	0002
$\alpha \backslash m$	2,5	3,0	3,5	4,0	5,0	6,0	7,0	8,0	9,0	10,0
0	0821	0498	0302	0183	0067	0025	0009	0003	0001	0000
1	2052	1494	1057	0733	0337	0149	0064	0027	0011	0005
2	2565	2240	1850	1467	0842	0446	0223	0107	0050	0023
3	2138	2240	2153	1954	1404	0892	0521	0286	0150	0076
4	1336	1680	1888	1954	1755	1339	0912	0573	0337	0189
5	0668	1008	1327	1563	1755	1606	1277	0916	0607	0378
6	0278	0504	0771	1042	1462	1606	1490	1221	0911	0631
7	0099	0216	0386	0595	1044	1377	1490	1396	1171	0901
8	0031	0081	0169	0298	0653	1033	1304	1396	1318	1126
9	0009	0027	0066	0132	0363	0688	1014	1241	1318	1251
10	0002	0008	0023	0053	0181	0413	0710	0993	1186	1251

Примечание: значения вероятности даны числами после запятой.

Таблица 4

Стандартные значения критерия Стьюдента ( $t$ )

$\nu$	$B_0 = 0,90$	$B_1 = 0,95$	$B_2 = 0,99$	$B_3 = 0,999$
1	6,3	12,7	63,7	637,0
2	2,9	4,3	9,9	31,6
3	2,4	3,2	5,8	12,9
4	2,1	2,8	4,6	8,6
5	2,0	2,6	4,0	6,9
6	1,9	2,4	3,7	6,0
7	1,9	2,4	3,5	5,3
8	1,9	2,3	3,4	5,0
9	1,8	2,3	3,3	4,8
10	1,8	2,2	3,2	4,6
11	1,8	2,2	3,1	4,4
12	1,8	2,2	3,1	4,2
13	1,8	2,2	3,0	4,1
14–15	1,8	2,1	3,0	4,1
16–17	1,7	2,1	2,9	4,0
18–20	1,7	2,1	2,9	3,9
21–24	1,7	2,1	2,8	3,8
25–28	1,7	2,1	2,8	3,7
29–30	1,7	2,0	2,8	3,7
31–34	1,7	2,0	2,7	3,7
35–42	1,7	2,0	2,7	3,6
43–62	1,7	2,0	2,7	3,5
63–175	1,6	2,0	2,6	3,4
176– $\infty$	1,6	2,0	2,6	3,3



Таблица 5

Стандартные значения критерия Фишера  $F = \frac{\sigma_1^2}{\sigma_2^2}$ 

$\begin{smallmatrix} v_1 \\ v_2 \end{smallmatrix}$	1	2	3	4	5	6	7	8	9	10	11	12
3	167,5 34,1 10,1	148,5 30,8 9,6	141,1 29,5 9,3	137,1 8,7 9,1	134,6 28,2 9,0	132,9 27,9 8,9	131,8 27,7 8,9	130,6 27,5 8,8	130,0 27,4 8,8	129,5 27,2 8,8	128,9 27,1 8,8	128,3 27,1 8,7
4	74,1 21,2 7,7	61,2 18,8 6,9	56,1 16,7 6,6	53,4 16,0 6,4	51,7 15,5 6,3	50,5 15,2 6,2	49,8 15,0 6,1	49,0 14,8 6,0	48,6 14,7 6,0	48,2 14,7 6,0	47,8 14,5 5,9	47,4 14,4 5,9
5	47,0 16,3 6,6	36,6 13,3 5,8	33,2 12,1 5,4	31,1 11,4 5,2	29,8 11,0 5,1	28,8 10,7 5,0	28,2 10,5 4,9	27,6 10,3 4,8	27,3 10,2 4,8	27,0 10,1 4,7	26,7 10,0 4,7	26,7 9,9 4,7
6	35,5 13,4 6,0	27,0 10,9 5,1	23,7 9,8 4,8	21,9 9,2 4,5	20,8 8,8 4,4	20,0 8,5 4,3	19,5 8,3 4,2	19,0 8,1 4,1	18,8 8,0 4,1	18,5 7,9 4,1	18,3 7,8 4,0	18,0 7,7 4,0
7	29,2 12,3 5,6	21,7 9,6 4,7	18,8 8,5 4,4	17,2 7,9 4,1	16,2 7,5 4,0	15,5 7,2 3,9	15,1 7,0 3,8	14,6 6,8 3,7	14,4 6,7 3,7	14,2 6,6 3,6	13,9 6,5 3,6	13,7 6,4 3,6
8	25,4 11,3 5,3	18,5 8,7 4,6	15,8 7,6 4,1	14,4 7,0 3,8	13,5 6,6 3,7	12,9 6,4 3,6	12,5 6,2 3,5	12,0 6,0 3,4	11,8 5,9 3,4	11,6 5,8 3,3	11,4 5,7 3,1	11,2 5,7 3,3
9	22,9 10,6 5,1	16,4 8,0 4,8	13,9 7,0 3,6	12,6 6,4 3,6	11,7 6,1 3,5	11,1 5,8 3,4	10,8 5,6 3,3	10,4 5,5 3,2	10,2 5,4 3,2	10,0 5,3 3,1	9,8 5,2 3,1	9,6 5,1 3,1
10	21,0 10,0 5,0	14,9 7,9 4,1	12,3 6,6 3,7	11,3 6,0 3,5	10,5 5,6 3,3	9,9 5,4 3,2	9,6 5,2 3,1	9,2 5,1 3,1	9,0 5,0 3,0	8,9 4,9 2,9	8,7 4,8 2,9	8,5 4,7 2,9
11	19,7 9,7 4,8	13,8 7,2 4,0	11,6 6,2 3,6	10,4 5,7 3,4	9,6 5,3 3,2	9,1 5,1 3,1	8,8 4,9 3,0	8,4 4,7 3,0	8,2 4,6 2,9	8,0 4,5 2,9	7,8 4,5 2,8	7,6 4,4 2,8
12	18,6 9,3 4,8	12,3 6,9 3,9	10,8 6,0 3,5	9,6 5,4 3,3	8,9 5,1 3,1	8,4 4,8 3,0	8,1 4,7 2,9	7,7 4,5 2,9	7,5 4,4 2,8	7,4 4,3 2,8	7,2 4,2 2,7	7,0 4,2 2,7
13	17,8 9,1 4,7	12,3 6,7 3,8	10,2 5,7 3,4	9,1 5,2 3,2	8,4 4,9 3,0	7,9 4,6 2,9	7,6 4,4 2,8	7,2 4,3 2,8	7,0 4,2 2,7	6,9 4,1 2,7	6,7 4,0 2,6	6,5 4,0 2,6
14	17,1 8,9 4,6	11,8 6,5 3,7	9,7 5,6 3,3	8,6 5,0 3,1	7,9 4,7 3,0	7,4 4,5 2,9	7,1 4,3 2,8	6,8 4,1 2,7	6,6 4,0 2,7	6,5 3,9 2,6	6,3 3,9 2,6	6,1 3,8 2,5
15	16,6 8,7 4,5	11,3 6,4 3,7	9,3 5,4 3,3	8,3 4,9 3,1	7,6 4,6 2,9	7,1 4,3 2,8	6,8 4,1 2,7	6,5 4,0 2,6	6,3 3,9 2,6	6,2 3,8 2,6	6,0 3,7 2,5	5,8 3,6 2,5
$\begin{smallmatrix} v_2 \\ v_1 \end{smallmatrix}$	1	2	3	4	5	6	7	8	9	10	11	12

$v_1 \backslash v_2$	14	16	20	24	30	40	50	75	100	200	500	$\infty$
3	127,7 26,9 8,7	127,1 26,8 8,7	126,5 26,7 8,7	125,9 26,6 8,6	125,6 26,5 8,6	125,3 26,4 8,6	125,0 26,4 8,6	124,7 26,3 8,5	124,4 26,2 8,5	124,1 26,2 8,5	123,8 26,1 8,5	123,5 26,1 8,5
4	47,0 14,2 5,9	46,6 14,1 5,8	46,2 14,0 5,8	45,8 13,9 5,8	45,6 13,8 5,7	45,4 13,7 5,7	45,2 13,7 5,7	45,0 13,6 5,7	44,7 13,5 5,7	44,5 13,5 5,7	44,3 13,5 5,6	44,1 13,5 5,6
5	26,4 9,8 4,6	26,1 9,7 4,6	25,8 9,6 4,6	25,4 9,5 4,5	24,1 9,4 4,5	24,9 9,3 4,5	24,6 9,2 4,4	24,5 9,1 4,4	24,3 9,1 4,4	24,1 9,1 4,4	24,0 9,0 4,4	23,8 9,0 4,4
6	17,7 7,6 3,9	17,5 7,5 3,9	17,2 7,4 3,8	16,9 7,3 3,8	16,8 7,2 3,8	16,6 7,1 3,8	16,5 7,1 3,7	16,4 7,0 3,7	16,2 7,0 3,7	16,1 6,9 3,7	15,9 6,9 3,7	15,9 6,9 3,7
7	13,5 6,3 3,5	13,2 6,2 3,5	13,0 6,1 3,4	12,7 6,0 3,4	12,6 5,9 3,4	12,5 5,9 3,3	12,3 5,9 3,3	12,2 5,8 3,3	12,1 5,8 3,3	12,0 5,7 3,3	11,8 5,7 3,2	11,7 5,7 2,2
8	11,0 5,6 3,2	10,8 5,5 3,2	10,5 5,4 3,2	10,3 5,3 3,1	10,2 5,2 3,1	10,1 5,1 3,1	10,0 5,1 3,0	9,9 5,0 3,0	9,7 5,0 3,0	9,6 4,9 3,0	9,5 4,9 2,9	9,4 4,9 2,9
9	9,4 5,0 3,0	9,2 4,9 3,0	8,9 4,8 2,9	8,7 4,7 2,9	8,6 4,6 2,9	8,5 4,5 2,8	8,4 4,5 2,8	8,3 4,4 2,8	8,1 4,4 2,8	8,0 4,4 2,7	7,9 4,3 2,7	7,9 4,3 2,7
10	8,3 4,6 2,8	8,1 4,5 2,8	7,8 4,4 2,7	7,6 4,3 2,7	7,5 4,3 2,7	7,4 4,2 2,6	7,3 4,1 2,6	7,2 4,1 2,6	7,1 4,0 2,6	7,0 4,0 2,6	6,9 3,9 2,6	6,8 3,9 2,5
11	7,4 4,3 2,7	7,3 4,2 2,7	7,1 4,1 2,7	6,9 4,0 2,6	6,8 3,9 2,6	6,7 3,9 2,5	6,6 3,8 2,5	6,5 3,7 2,5	6,3 3,7 2,5	6,2 3,7 2,4	6,1 3,6 2,4	6,0 3,6 2,4
12	6,8 4,1 2,6	6,7 4,0 2,5	6,5 3,9 2,5	6,3 3,8 2,5	6,2 3,7 2,5	6,1 3,6 2,4	6,0 3,6 2,4	5,9 3,5 2,4	5,7 3,5 2,4	5,6 3,4 2,3	5,5 3,4 2,3	5,4 3,4 2,3
13	6,3 3,9 2,6	6,2 3,8 2,5	6,0 3,7 2,5	5,8 3,6 2,4	5,7 3,5 2,4	5,6 3,4 2,3	5,5 3,4 2,3	5,4 3,3 2,3	5,3 3,3 2,3	5,2 3,2 2,2	5,1 3,2 2,2	5,0 3,2 2,2
14	5,9 3,7 2,5	5,8 3,5 2,4	5,6 3,5 2,4	5,4 3,4 2,4	5,3 3,3 2,3	5,2 3,3 2,3	5,1 2,2 2,2	5,0 3,1 2,2	4,9 3,1 2,2	4,8 3,1 2,2	4,7 3,0 2,1	4,6 3,0 2,1
15	5,6 3,5 2,4	5,5 3,4 2,4	5,3 3,3 2,3	5,1 3,2 2,3	5,0 3,2 2,3	4,9 3,1 2,2	4,8 3,1 2,2	4,7 3,0 2,2	4,6 3,0 2,1	4,5 2,9 2,1	4,4 2,9 2,1	4,3 2,9 2,1
$v_2 \backslash v_1$	14	16	20	24	30	40	50	75	100	200	500	$\infty$

Таблица 6

Стандартные значения критерия Фишера  $\chi^2$  (хи-квадрат)

$\nu$	$\chi_1^2$	$\chi_2^2$	$\chi_3^2$	$\nu$	$\chi_1^2$	$\chi_2^2$	$\chi_3^2$
1	3,8	6,6	10,8	26	38,9	45,6	54,1
2	6,0	9,2	13,8	27	40,1	47,0	55,5
3	7,8	11,3	16,3	28	41,3	48,3	56,9
4	9,5	13,3	18,5	29	42,6	49,6	58,3
5	11,1	15,1	20,5	30	43,8	50,9	59,7
6	12,6	16,8	22,5	32	46,2	53,5	62,4
7	14,1	18,5	24,3	34	48,6	56,0	65,2
8	15,5	20,1	26,1	36	51,0	58,6	67,9
9	16,9	21,7	27,9	38	53,4	61,1	70,7
10	18,3	23,2	29,6	40	55,8	63,7	73,4
11	19,7	24,7	31,3	42	58,1	66,2	76,1
12	21,0	26,2	32,9	44	60,5	68,7	78,7
13	22,4	27,7	34,5	46	62,8	71,2	81,4
14	23,7	29,1	36,1	48	65,2	73,7	84,0
15	25,0	30,6	37,7	50	67,5	76,2	86,7
16	26,3	32,0	39,3	55	73,3	82,3	93,2
17	27,6	33,4	40,8	60	79,1	88,4	99,6
18	28,9	34,8	42,3	65	84,8	94,4	106,0
19	30,1	36,2	43,8	70	90,5	100,4	112,3
20	31,4	37,6	45,3	75	96,2	106,4	118,5
21	32,7	38,9	46,8	80	101,9	112,3	124,8
22	33,9	40,3	48,3	85	107,5	118,2	131,0
23	35,2	41,6	49,7	90	113,1	124,1	137,1
24	36,4	43,0	51,2	95	118,7	130,0	143,3
25	37,7	44,3	52,6	100	124,3	135,8	149,4

Таблица 7

**Достаточная численность выборки**

К \ В	Изучаются: средние $M$ , доли $P$			Изучаются разности средних и разность долей					
				выборки одинакового объема $n_1 = n_2$			Первая выборка меньше второй $n_1 < n_2 \cdot n_1 = \frac{n_2}{e}$		
	в таблице дается объем выборки, $n$			в таблице дается одинаковый объем каждой выборки $n_1 = \hat{n}, n_2 = \hat{n}_1$			в таблице дается объем каждой меньшей выборки $\hat{n}_1$ ( $n_2 = en_1$ )		
	0,95	0,99	0,999	0,95	0,99	0,999	0,95	0,99	0,999
0,20	99	170	278	97	168	274	97	167	273
0,21	90	155	253	89	153	251	88	152	248
0,22	82	141	231	80	139	227	80	138	226
0,23	75	130	212	74	128	208	74	127	207
0,24	69	119	195	68	118	192	68	117	191
0,25	64	111	180	63	110	177	63	109	176
0,26	59	102	167	58	101	164	58	100	163
0,27	55	96	155	54	94	152	54	93	151
0,28	52	89	145	51	87	142	50	87	141
0,29	48	83	135	47	81	133	47	81	132
0,30	45	78	127	44	76	124	44	76	123
0,32	40	69	112	39	67	110	39	66	108
0,34	36	62	100	35	60	97	34	59	96
0,36	32	55	90	31	54	87	31	53	86
0,38	29	50	81	28	48	79	28	48	77
0,40	27	47	74	26	44	71	25	43	70
0,42	25	42	68	23	40	65	23	39	64
0,44	23	38	62	21	37	60	21	36	58
0,46	21	35	57	20	34	55	19	33	54
0,48	19	33	53	18	31	51	18	30	49
0,50	18	31	50	17	29	47	17	28	46
0,60	14	23	36	12	21	34	12	20	32
0,70	11	18	27	10	16	24	9	15	23
0,80	9	14	23	8	13	20	7	12	19
0,90	8	12	19	5	11	17	6	10	16
1,00	7	11	17	7	9	14	5	8	13

Таблица 8

**Критические значения z-критерия знаков  
при разных уровнях значимости  $\alpha$  и объеме выборки  $n$**

$n$	$\alpha, \%$		$n$	$\alpha, \%$		$n$	$\alpha, \%$		$n$	$\alpha, \%$	
	5	1		5	1		5	1		5	1
6	6	–	30	21	23	54	35	37	78	49	51
7	7	–	31	22	24	55	36	38	79	49	52
8	8	8	32	23	24	56	36	39	80	50	52
9	8	9	33	23	25	57	37	39	81	50	53
10	9	10	34	24	25	58	37	40	82	51	54
11	10	11	35	24	26	59	38	40	83	51	54
12	10	11	36	25	27	60	39	41	84	52	55
13	13	12	37	25	27	61	39	41	85	53	55
14	14	13	38	26	28	62	40	42	86	53	56
15	12	13	39	27	28	63	40	43	87	54	56
16	13	14	40	27	29	64	41	43	88	54	57
17	13	15	41	28	30	65	41	44	89	55	58
18	14	15	42	28	30	66	42	44	90	55	58
19	15	16	43	29	31	67	42	45	91	56	59
20	15	17	44	29	31	68	43	46	92	56	59
21	16	17	45	30	32	69	44	46	93	57	60
22	17	18	46	31	33	70	44	47	94	57	60
23	17	19	47	31	33	71	45	47	95	58	61
24	18	19	48	32	34	72	45	48	96	59	62
25	18	20	49	32	34	73	46	48	97	59	62
26	19	20	50	33	35	74	46	49	98	60	63
27	20	21	51	33	36	75	47	50	99	60	63
28	20	22	52	34	36	76	48	50	100	61	64
29	21	22	53	35	37	77	48	51	–	–	–
P	0,05	0,01	–	0,05	0,01	–	0,05	0,01	–	0,05	0,01

Таблица 9

**Количество пар значений  $N$ , достаточное для достоверности  
выборочного коэффициента корреляции, %**

$r_i$	$\hat{N}$			$r_i$	$\hat{N}$			$r_i$	$\hat{N}$		
	$B_1 - 0,95$	$B_2 - 0,99$	$B_3 - 0,999$		$B_1 - 0,95$	$B_2 - 0,99$	$B_3 - 0,999$		$B_1 - 0,95$	$B_2 - 0,99$	$B_3 - 0,999$
05	1539	2263	4359	35	32	53	85	65	9	14	21
06	1069	1850	3028	36	30	50	80	66	9	14	20
07	787	1360	2225	37	28	47	75	67	9	13	20
08	604	1042	1704	38	27	44	71	68	9	13	19
09	477	824	1347	39	26	42	67	69	8	12	18
10	383	661	1081	40	24	40	64	70	8	12	18
11	317	548	896	41	23	38	60	71	8	11	17
12	267	462	754	42	22	36	57	72	8	11	16
13	228	392	640	43	21	34	55	73	7	11	16
14	196	337	550	44	20	33	52	74	7	10	15
15	171	295	481	45	19	31	49	75	7	10	15
16	151	259	422	46	19	30	47	76	7	10	14
17	133	228	373	47	18	29	45	77	7	9	14
18	119	204	332	48	17	27	43	78	7	9	13
19	107	183	297	49	16	26	41	79	7	9	13
20	97	165	270	50	16	25	39	80	6	9	12
21	87	149	242	51	15	24	37	81	6	8	11
22	80	136	211	52	15	23	36	82	6	8	11
23	73	124	202	53	14	22	34	83	6	8	11
24	68	114	185	54	14	21	33	84	6	7	10
25	62	105	170	55	13	20	32	85	5	7	10
26	57	97	157	56	13	20	30	86	5	7	10
27	53	90	145	57	12	19	39	87	5	7	9
28	49	83	135	58	12	18	28	88	5	7	9
29	46	78	125	59	11	18	27	89	5	6	8
30	43	73	117	60	11	17	26	90	5	6	8

Таблица 10

**Основные термины и символы, применяемые в биометрии**

В данном пособии	В работах других авторов
ПРИЗНАК (элементарная особенность каждого живого объекта в экстерьере, конституции, анатомии, гистологии, физиологии, продуктивности)	Величина случайная, переменная
ДАТА (результат измерения признака, его значение, величина), $V, x_i$ ОБЪЕМ ГРУППЫ (число объектов в группе), $n, N$	Значение, приобретаемое случайной переменной, вариант – $V, X, x, y, a$ . Численность, объем группы – $n, N$
Средняя величина признака: $\bar{X} = \bar{M} = \frac{\sum V}{n} = \frac{\sum f_i x_i}{n}.$ Выборочная средняя, $\bar{M}, \bar{X}$	Среднее значение случайной переменной – $M, m, a, \beta, \bar{x}$
РАЗНООБРАЗИЕ (наличие неодинаковых объектов в группе)	Изменчивость, колеблемость, рассеяние, вариабельность, разброс
Сумма квадратов, ДИСПЕРСИЯ: $C = \sum (V - M^2) = \sum (X - \bar{X})^2$ или $C = \sum f(V - M^2),$ или $\sum f(X - \bar{X})^2$	Сумма квадратов центральных отклонений, сумма квадратов, дисперсия: $\sum (V - M^2), \sum (X - \bar{X})^2,$ $\sum x^2, S, SS, SO, SA, SAQ, CQ$
ВАРИАНСА, средний квадрат: $\sigma^2 = \frac{C}{n-1}$	Средний квадрат, дисперсия, девиата, варианса – $\sigma^2, s^2, v^2, e, M, MQ, ES$
СРЕДНЕЕ КВАДРАТИЧЕСКОЕ ОТКЛОНЕНИЕ, сигма: $\sigma = \sqrt{\frac{C}{n-1}}.$ Выборочная сигма $\sigma, s$	Среднее квадратическое отклонение, стандарт – $\sigma, S$
РАЗНОСТЬ ДОСТОВЕРНА: между генеральными средними можно ожидать такое же различие, какое найдено между выборочными средними, – различие по знаку, а величину разности – по доверительным границам: $(\tilde{M}_1 > \tilde{M}_2) \rightarrow (\bar{M}_1 > \bar{M}_2)$	Разность существенна, надежна, значима, реальна, разница есть; разность достоверна, т.е. реальна; выборки из разных генеральных совокупностей
РАЗНОСТЬ НЕДОСТОВЕРНА (получены неопределенные результаты): $(\tilde{M}_1 > \tilde{M}_2) \rightarrow (\bar{M}_1 \geq \bar{M}_2)$	Разность несущественна. Выборки из одной генеральной совокупности

**Маргарита Федоровна КОЗАК,  
Михаил Владимирович КОЗАК**

**БИОМЕТРИЧЕСКИЕ МЕТОДЫ  
В НАУЧНЫХ ИССЛЕДОВАНИЯХ**

***Монография***

Редактирование, компьютерная правка,  
верстка С.Н. Лычагиной

Заказ № 3851. Тираж 500 экз. (первый завод – 40 экз.)  
Уч.-изд. л. 10,5. Усл. печ. л. 9,8.

---

Издательский дом «Астраханский университет»  
414056, г. Астрахань, ул. Татищева, 20а  
тел. (8512) 48-53-47 (отдел планирования и реализации),  
48-53-44, тел./факс (8512) 48-53-46  
E-mail: asupress@yandex.ru