

**МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ УКРАИНЫ  
НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ  
«Харьковский политехнический институт»**

**УЧЕБНОЕ ПОСОБИЕ ПО КУРСУ  
«БИОМЕТРИЯ»**

**для студентов специальности 7.092901  
«Промышленная биотехнология»**

**ХАРЬКОВ 2012**

ББК 28.05  
О 39  
УДК 577.352

Рецензенты

, д-р техн. наук, проф. НТУ «ХПИ»

, д-р ? наук, проф.

Горбунов Л.В., Клещев Н.Ф. Учебное пособие по курсу «биометрия» для студентов специальности 7.092901 «Промышленная биотехнология». – Харьков: НТУ «ХПИ», 2012. – 227 с. – На русском языке.

Учебное пособие включает принципы постановки биологического эксперимента и методы статистической обработки экспериментальных данных.

Предназначено для студентов специальности 7.092901 "Промышленная биотехнология" и может быть полезно для всех студентов, как дневного, так и заочного обучения при выполнении курсовых и дипломных работ, а также для аспирантов и инженеров-исследователей, научных работников.

Учебний посібник містить принципи постановки біологічного експерименту та методи статистичної обробки експериментальних даних.

Призначено для студентів спеціальності 7.092901 "Промислова біотехнологія" при виконанні курсових та дипломних робіт, а також для аспірантів, інженерів-дослідників, наукових робітників.

Ил. ??      Табл. ?      Библиогр. 18 назв.

©, Горбунов Л.В., Клещев Н.Ф. 2012 г.

**МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ УКРАИНЫ**  
**Национальный технический университет**  
**«Харьковский политехнический институт»**

*RELEASE* – 10 октября

**УЧЕБНОЕ ПОСОБИЕ ПО КУРСУ**  
**«БИОМЕТРИЯ»**

для студентов специальности 7.092901

«Промышленная биотехнология»

Утверждено

Редакционно-издательским

Советом университета

Протокол № ? от ??.??.2012 г.

Харьков НТУ «ХПИ» 2012

## ВСТУПЛЕНИЕ

Эволюция развития человечества входит в фазу информационного общества, основным ресурсом которого является не только полезные ископаемые (газ, нефть, золото, бриллианты и т.д.), но и информация. Информация как ресурс развития доступна практически каждому, но не все владеют методами извлечения из неё знаний. Стратегия развития передовых стран мира направлена на получение новых знаний с целью разработки новейших технологий. Именно поэтому современный инженер-биотехнолог должен не только хорошо знать свою специальность, но и приобщаться к исследовательской работе, вносить посильный вклад в сокровищницу знаний о природе.

Процесс получения нового знания называют научным исследованием, необходимым условием которого является – воспроизводимость его результатов. Воспроизводимость результатов эксперимента зависит от методов его проведения и обработки, поскольку оценить данную величину без осуществления статистического анализа невозможно.

Большое разнообразие биологических форм, их изменчивость в онтогенезе и в поколениях, сложная многоуровневая организация делают работу биолога – экспериментатора особенно трудной. В этом причина низкой воспроизводимости результатов многих исследований и как следствие этому в биологии недостаточно, по сравнению с другими науками, представлены теоретические обобщения. Основной недостаток существующей экспериментальной парадигмы в биологии заключается в проблеме сопоставимости полученных результатов, что является следствием высокой гетерогенности и низкой точности определения состояния биообъекта. Гетерогенность популяции микроорганизмов может существенно осложнять научно-исследовательскую работу. Генетическая нестабильность штаммов продуцентов – одна из главных причин снижения эффективности биотехнологических процессов.

Сравнение полученных данных возможно только при одинаковых условиях проведения опытов, что обеспечивает воспроизводимость их результатов. Особенно остро данная проблема прослеживается в биотехнологических работах основанных на применении множества химических, физических и биологических методов исследования объекта.

Традиционный способ снижения случайной ошибки заключается в применении теории вероятности. Повтор измерений повышает точность определения среднего результата, но в месте с тем увеличивается разброс текущих значений. Если воспринимать случайность как непознанную закономерность, то возникает вопрос: «каким образом из случайностей извлекать закономерности»?

Анализ опыта работы Института высоких статистических технологий и эконометрики МГТУ им. Н.Э. Баумана показал, что причиной низкой эффективности проведения эксперимента является применение устаревших и частично ошибочных статистических технологий [28]. Экономический эффект от использования статистического контроля в промышленности США составляет 0,8 % от валового национального продукта (20 млрд. долларов в год), что существенно больше, чем от любого иного экономико-математического метода.

Причиной сдерживания темпа развития науки является не только недостаток материальных средств, ограничивающий возможности изучения выбранного объекта, но и инерционность мышления субъекта, реализующего данный процесс. Обзор научных работ в области биологии и медицины показывает [9, 26, 27], что более половины из них содержат ошибочные методы статистической обработки или совсем не имеют анализ представленных данных.

По мнению ведущих исследователей [1-3, 5-9, 25-27, 30] в области статистики, существующие технологии обработки информации можно условно разделить на три уровня: низкий (параметрические критерии –  $F$ ,  $t$ ), средний (непараметрические –  $\chi$ ,  $\chi^2$ ,  $U$ ,  $Z$ ,  $T$  и др., многофакторный анализ и оптимизация) и высокий (стохастические математические модели). Многократное повышение эффективности обработки информации реализуется на основе применения методов стохастического моделирования – Data Mining (получение данных, добыча знаний) [35], задачей которых является поиск не очевидных закономерностей.

Прогресс исследовательской работы основан на применении высоких технологий, как измерения, так и обработки полученных данных. Аналогией является развитие компьютерной техники, которое реализуется за счёт наращивания мощности операционной системы и её программного обеспечения.

Основной задачей данного курса лекций является донесение до сознания студентов методологии структурирования информации в области биотехнологии. Представленная методология включает комплекс методов от постановки задачи на исследование до получения математической модели. Основой курса являются лекции, в которых излагаются положения каждого раздела. Практические занятия и самостоятельная работа студентов является основным способом усвоения материала в свободное от аудиторных занятий время.

Предметом учебной дисциплины «Биометрия» являются принципы постановки биологического эксперимента и статистическая обработка экспериментальных данных. Рассматриваются особенности количественного и качественного анализа с учетом общей оценки ошибок опытов. Излагаются методы планирования экспериментов. Особое внимание обращено на оценку достоверности параметров, распределение случайных величин и проверку статистических гипотез. Представлена структурная схема, отражающая поэтапный процесс проведения моделирования эксперимента, от постановки задачи на исследование до математического обобщения полученного результата. Представлены формулы получения достоверного научного результата при минимальном количестве измерений. На примерах показано, что повышение эффективности исследовательской работы связано с моделированием проведения эксперимента основанном на минимизации затрат времени и материальных средств.

Под проведением комплексного исследования понимается весь объем мероприятий от постановки задачи до получения математической модели изучаемых явлений. Несмотря на кажущуюся простоту такого определения, дать целостное и последовательное изложение этого вопроса нелегко. Это совсем не означает, что материала по предмету нет, наоборот, его слишком много, но он распределен малыми дозами в огромном числе монографий и статей. В данном

пособии сделана попытка, собрать воедино то, что должно помочь начинающему исследователю "войти в работу", понять структуру исследования и узнать о неустранимых трудностях. Особенно ценным является проходящее красной нитью через всю работу положение о необходимости такой постановки экспериментов, какого бы вопроса они ни касались, которая обеспечила бы максимально объективную и статистически достоверную оценку их результатов. Это положение должно войти "в плоть и кровь" всех экспериментаторов в области биотехнологии и отвечало бы его проблемам.

Математический анализ результатов эксперимента, позволяющий установить их статистическую достоверность, которому в данной работе уделено достаточное внимание, окажется, безусловно, весьма полезным как при планировании подобных исследований студентами, так и при анализе их результатов. Его краткое изложение помещено в приложении. Ограниченный объем руководства побудил авторов в отдельных случаях, когда можно дать рекомендации в отношении общедоступной литературы и таблиц, не проводить развернутого изложения некоторых методов.

Самостоятельная работа проводится по всем темам, которые входят в дисциплину, как по тем, по которым читаются лекции, так и по тем, для которых делается только план изучения и ссылки на литературу. В процессе самостоятельной работы студент учится самостоятельно приобретать знания, которые затем используются в ходе практических занятий, при подготовке к выполнению контрольных работ и к зачету.

Настоящее руководство, хотя и рассчитано на студентов, предполагает все же, что читатель имеет первичное представление о методах статистической обработки экспериментальных данных, по крайней мере, в пределах общих разделов одного из наиболее доступных пособий [3-7, 11, 13-15, 17, 18]. Для желающих более глубоко изучить вопросы, затронутые в данной работе, приведен список отечественной и переведенной на русский язык зарубежной литературы.

Поскольку основной акцент в данных источниках делается на выборе адекватных методов анализа уже полученной информации, то в данном пособии диапазон раскрыт от постановки задачи на исследование до получения математической модели. Обобщение большого числа экспериментов позволило в общих чертах создать структурную схему планирования и проведения эксперимента, обработки его данных, хотя, естественно, она не может претендовать на полноту отражения всего многообразия, встречающегося на практике. Ниже с той или иной полнотой рассмотрены и обсуждены отдельные составляющие этой схемы.

## Глава 1. Методология исследования

**Методология исследования** – комплекс методов от постановки задачи до построения математической модели. М.В. Ломоносов утверждал, что: «Математика является царицей наук» и как следствие этому можно сказать, что в каждой дисциплине столько науки, сколько в ней математики. Поскольку только при помощи математического аппарата можно выявить и описать структуру исследуемой системы.

### 1.1 Способы структурирования полученной информации

Термин структура (*structura* – строение) имеет целый спектр значений. В словосочетании структура системы данное понятие трактуется как организация связей и отношений между подсистемами и элементами системы, а также собственно состав этих подсистем и элементов, каждому из которых обычно соответствует определенная функция. Например, структура молекулы воды зависит от многих факторов, ведущим из которых является температура. При снижении температуры вода меняет своё фазовое состояние вследствие изменения структуры её молекул, которые, агрегируясь, образуют вначале кластеры, затем дендриты, сферолиты и кристаллы, имеющие кубическую и гексагональную форму.

Очевидно, что полученная информация должна быть представлена так, чтобы можно было выявить из неё закономерности с целью получения новых знаний. Поскольку знания это выявленные закономерности предметной области (принципы, связи, законы) из полученных данных (отдельные факты, характеризующие объекты, процессы и явления). Знания принято классифицировать по следующим категориям: поверхностные – о видимых взаимосвязях между отдельными событиями и фактами; глубинные – абстракции, аналогии, схемы, модели отображающие структуру и процессы.

Структурирование информации переводит её в качественно новое состояние – данные, агрегация, которых дает новые свойства – знание, на основе которого создается технология. Ярким примером получения качественно новых свойств, в процессе структурирования материала, является процедура получения дамасской стали. Общеизвестно, что булат может иметь лучшие по сравнению со сталью свойства и при определенной обработке переходит в иное качество. Для получения "сверхстали" применяют специальные режимы обработки металла, результатом которых является особая структура материала имеющая характерные узоры на клинке. Следовательно, уникальные особенности булатного клинка зависят не только от состава его материала (наличия химических элементов и их пропорций), но и структуры (которая является следствием выбранной технологии изготовления).

Процедура получения нового знания имеет несколько этапов, зависящие от степени структурированности (упорядоченности, агрегирования) информации исследуемой области познания. В связи с этим различают три типа задач, для которых применяются информационные системы: неструктурированные

(не формализуемые), частично структурированные и полностью структурированные (формализуемые).

Структурированная задача – в которой известны все ее элементы и взаимосвязи между ними. В такой задаче удастся выразить ее содержание в форме математической модели, имеющей точный алгоритм решения. Неструктурированная задача – в которой невозможно выделить элементы и установить связи между ними. Решение этих задач связано с большими трудностями из-за невозможности создания математического описания и разработки алгоритма. Решение в таких случаях принимается человеком на основе своего опыта (интуиции) и косвенной информации из разных источников.

В практике биотехнологических исследований сравнительно немного встречается полностью структурированных или совершенно неструктурированных задач. В большинстве задач известна лишь часть их элементов и связей между ними. Такие задачи называются частично структурированными. В этих условиях необходимо создавать математическую модель.

Характерной особенностью биологического исследования является накопление не структурированных данных и как следствие этому отсутствие математических моделей. Причинами этому есть как объективные, так и субъективные факторы. Объективные – многомерность открытой системы и как следствие этому высокий коэффициент вариации исследуемых параметров. Субъективные – закрытость, инерционность мышления исследователей, что проявляется в игнорировании существующих методов высоких статистических технологий, позволяющих повысить воспроизводимость результатов исследования.

Весь арсенал используемых статистических способов можно распределить на 3 уровня [25, 27]:

- высокие статистические технологии;
- классические статистические технологии,
- низкие статистические технологии.

Под классическими статистическими технологиями понимаем технологии почтенного возраста, сохранившие свое значение для современной статистической практики. Таковы метод наименьших квадратов, статистики Колмогорова, Смирнова, омега-квадрат, непараметрические коэффициенты корреляции Спирмена и Кендалла.

Низкие технологии структурирования информации хорошо применимы при изучении закрытых систем (физика, химия...), для открытых (биология, биотехнология ...) следует применять средние или высокие технологии вследствие высокой гетерогенности (не однородности) и лабильности исследуемых объектов.

Традиционный способ снижения случайной ошибки заключается в применении теории вероятности. Повтор измерений повышает точность определения среднего результата, но в месте с тем увеличивается разброс текущих значений. Если воспринимать случайность как непознанную закономерность, то возникает вопрос: «каким образом из случайностей извлекать закономерности»? Рассмотрим несколько примеров данной экстракции.



## **1.2 История, пути и направления развития статистики и биометрии.**

Статистика прошла путь от количественного описания отдельных свойств биологического объекта до выявления закономерности изучаемого явления в виде математических моделей. В истории статистики можно выделить несколько этапов:

**1 описательный** (начало XVII в.) – наблюдается переход от словесного описания и элементарного количественного учета отдельного биологического объекта к их числовым взаимосвязям;

**2 объяснительный** (начало XIX в.) – целью является не только описание исследуемых явлений, но и их анализ;

**3 формалистический** (конец XIX в.) – создание математического аппарата для обработки экспериментальных данных и применение его к изучению проблемы наследственности и изменчивости организмов;

**4 рационалистический** (начало XX в.) – математические методы применяются как необходимость проведения эксперимента, создаются основы теории малой выборки, теории планирования экспериментов;

**5 стохастическое моделирование** (конец XX начало XXI в.) - многократное повышение эффективности обработки информации, использование методов Data Mining, задачей которых является поиск не очевидных закономерностей [35].

Математическая статистика бурно развивается и в настоящее время. Так, за последние 40 лет можно выделить три принципиально новых направления исследования:

- развитие статистики объектов нечисловой природы как самостоятельного направления в прикладной математической статистике;
- развитие статистических методов используемых в стохастическом моделировании;
- широкое развертывание работ по созданию компьютерных пакетов программ, предназначенных для проведения статистического анализа данных.

Стандартные статистические методы включены в состав популярных электронных таблиц, таких как Excel, Lotus 1-2-3, Quattro Pro, а также в математические пакеты общего назначения, например Mathcad, Maple и др. Однако гораздо большими возможностями обладает программное обеспечение – статистические программные продукты (СПП). Международный рынок насчитывает более 1000 пакетов, решающих задачи статистического анализа данных в среде операционных систем Windows, DOS, OS/2.

СПП можно разделить на:

1. Универсальные пакеты – предлагают широкий диапазон статистических методов. В них отсутствует ориентация на конкретную предметную область.

Из зарубежных универсальных пакетов наиболее распространены BAS, SPSS, Systat, Minilab, Statgraphics, STATISTICA.

2. Специализированные пакеты, как правило, реализуют несколько статистических методов или методы, применяемые в конкретной предметной обла-

сти. Чаще всего это системы, ориентированные на анализ временных рядов, корреляционно-регрессионный, факторный или кластерный анализ.

Основная современная проблема статистических технологий состоит в том, чтобы в конкретных статистических исследованиях использовались только технологий первых двух типов. Например, в США - около 20 миллиардов долларов ежегодно экономится только в области статистического контроля качества. Применение сценарный методов агрегирования информации в знания **Data Mining** (добыча данных) повышает рентабельность производства до 70 раз.

### 1.2.1 Проблемы развития биометрии.

**Биометрия** – раздел биологии, содержанием которой является *планирование* эксперимента и статистическая обработка его результатов. **Математическая статистика** и **теория вероятности** – разделы математики, теоретические фундаментальные науки, рассматривающие массовые явления безотносительно к специфике составляющих их элементов. Основное отличие биометрии от математической статистики и теории вероятности заключено в умении планировать биологический эксперимент с целью получения достоверного результата при минимальных затратах времени и материальных средств.

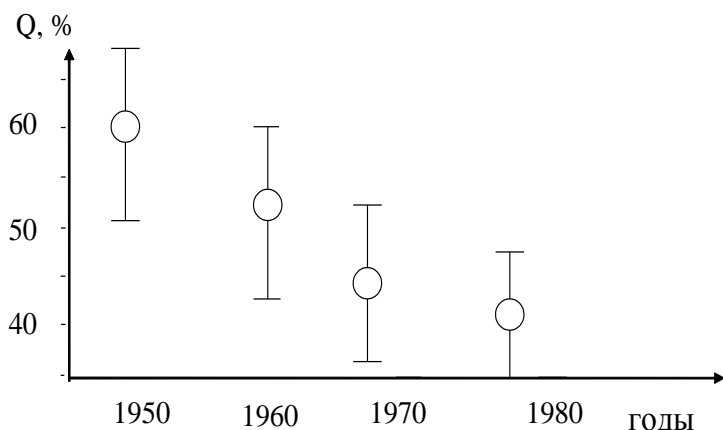
Предубеждение против использования математических методов в биологии было очень сильным до конца XIX века. Биологии отводилась роль пограничной науки, занимающей промежуточное место между “точными науками” и “гуманитарными науками”, которые в Англии традиционно относили к категории “искусств”. В 1901 г. К. Пирсон основал журнал “*Biometrika*”. Его задачей было поощрение использования математических методов в биологии. Начиная с середины XIX века до 30-х годов XX российские медицина и биология занимали передовые позиции в развитии и применении биометрии в научных исследованиях.

В 30-х годах широкое распространение получила точка зрения, согласно которой статистика являлась “... наукой о стихийных явлениях природы и общества ... математические науки так же партийны, как и науки экономические ...” Монополия Т.Д. Лысенко в биологии (“...биология и математика — вещи несовместимые ...”) привела не только к запрету генетики, но и разгрому российской биометрической школы.

“Такие науки, как физика и химия освободились от случайностей. Поэтому они стали точными науками. ... Изживая из нашей науки менделизм-морганизм-вейсманнизм, мы тем самым изгоняем случайность из биологической науки. Нам необходимо твердо запомнить, что наука — враг случайностей”. Последняя фраза, в акцентуации Т.Д. Лысенко имеет следующий смысл: “Где есть случайность, вероятность — там нет науки”.

Следствием августовской сессии ВАСХНИЛ 1948 г. было уничтожение учебников по генетике и биометрии, исключение статистики как предмета из программ подготовки биологов.

В 60-е годы значительно вырос интерес к использованию биометрии в медицине и биологии; появились публикации, аргументированные рекомендации, не потерявшие своей актуальности и по сей день. Анализ научных работ с



1950 по 1980 г. посвященных биологическим и медицинским исследованиям, показывает снижение от 60 до 43 % методических ошибок при использовании статистического анализа (рис. 1).

Рис. 1. Процент биометрических ошибок содержащихся в медико-

биологический статьях напечатанных в период с 1950 по 1980 г. Q — доля статей с ошибками.

Безусловно, невозможно рассмотреть все статьи, публикуемые в биологических журналах, поэтому долю определяли по некоторой случайной выборке. В результате определили оценку истинной доли статей с ошибками, на рисунке эти оценки показаны кружками. Вертикальные отрезки это доверительный интервал, то есть пределы в которых, скорее всего, находится истинная доля статей с ошибками.

В биологии математические методы лишь сравнительно недавно начали приобретать заметное развитие, но все же они приняты здесь значительно меньше, чем, к примеру, в физике или в инженерном деле. Обычно причиной считают большую сложность выявления биологических закономерностей; до известной степени это верно. Лишь тогда, когда мы приходим к сопоставимым явлениям, наступает черед математики. Именно поэтому, т.е. благодаря неправильной трактовке роли математических методов, в биологии нередки случаи, когда математизирование биологической работы лишь дискредитирует идею применения математики в биологии. Результатом явилось то, что за последнее время многие биологи заметно охладели к математике. Правда, немногие решаются полностью отрицать значение и пользу математических методов в биологических исследованиях, но к таким методам существует боязливое, подозрительное отношение [25-27].

Таким образом, анализ научных публикаций показывает [9, 25-27, 35], что отечественные исследователи значительно реже, чем зарубежные, применяют в своих работах методы биометрии. Это является одной из причин низкой эффективности отечественных биомедицинских исследований.

Использование статистических методов в биологических и медицинских журналах в настоящее время показало [9, 25, 27], что в 39 % работ статистические методы не использовались вовсе, в 34 % правильно использовали критерий Стьюдента, дисперсионный анализ или другие методы. В 27 % работ кри-

терий Стьюдента использовали неправильно – для парного сравнения нескольких групп (рис 2).

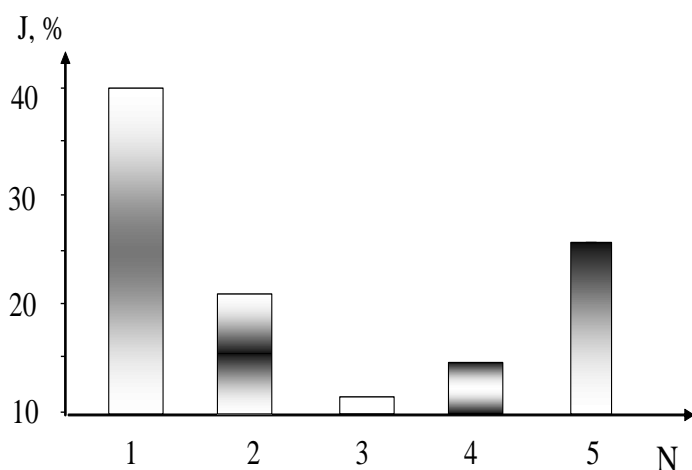


Рис. 2. Процент статей (J), в которых: не использовали статистические методы (1 группа), правильно использовали критерий Стьюдента (2), правильно использовали дисперсионный анализ (3), правильно использовали другие методы (4), неправильно использовали критерий Стьюдента для парного сравнения нескольких групп (5).

Таким образом, анализ научных работ за последние 10 ÷ 20 лет, показывает высокий процент не правильного использования методов статистического анализа и практически полное отсутствие признаков планирования биологического эксперимента, применения методов математического моделирования. Причиной этому является использование одного метода обработки полученной информации (как правило, не парный критерий Стьюдента) при разной постановке задачи на исследование.

В сегодняшнем мире постоянно растет интерес к использованию биометрических параметров (индивидуальных физических свойств, таких, как отпечатки пальцев, лицо, голос, радужная оболочка глаза, геометрия руки) для идентификации человека, они все чаще находят применение в современных системах обеспечения безопасности.

К настоящему времени можно выделить следующие основные направления развития биометрии:

- повышение эффективности биологического исследования;
- повышение воспроизводимости результатов биологического эксперимента;
- подробное изложение принципов планирования эксперимента его проведения и обработки полученных результатов;
- развитие методов математического моделирования как инструмента научного исследования;
- реализация концепции многофакторной биометрической идентификации объекта

Таким образом, структурирование информации основано на выделении из шума скрытых закономерностей изучаемого явления при помощи методов математического абстрагирования.

### **1.3 Повышение эффективности процедуры исследования**

**Научное исследование** - это изучение закономерностей развития явлений объективного мира и их объяснение [3]. Исследование может быть экспериментальным или теоретическим. Опытное решение - основа исследования; оно всегда конкретно и точно, но вместе с тем не имеет характера всеобщности и в некоторых случаях громоздко и трудоемко.

Эксперимент может быть физическим, модельным или мыслимым. В отличие от мыслимого (виртуального) эксперимента он может непосредственно проводиться на объекте или на его модели. Модель обычно отличается от объекта масштабом, а иногда природой. В последнее время наряду с физическими моделями все большее распространение получают абстрактные математические модели. Можно получить новые сведения об объекте, экспериментируя на модели, если она достаточно точно описывает объект. Примером мысленного эксперимента является открытие Эйнштейном теории относительности.

Теоретические исследования обладают свойствами всеобщности. Правильность и надежность зависит от правильности предпосылок, допущений и логического развития. Изучаемые явления настолько многообразны и сложны, что часто трудно, а порой и невозможно получить точное теоретическое решение, поэтому результаты расчета становятся ненадежными.

Можно выделить два полюса в исследовании – эксперимент и теория, которые производят факты и гипотезы. Факты являются фундаментом науки и лежат в основе всякой гипотезы [25]: "Факты – это золото, тогда как теории и гипотезы это кредитные бумаги, которые пускаются в обращение, и которые имеют цену, сообразную с количеством звонкой монеты, которому она соответствует. Факты имеют значение абсолютное, теории и гипотезы – лишь относительное. Факты есть строительный материал, теория – план будущего здания. Во время постройки план может много раз меняться, строительный материал идет в дело, так или иначе, но без плана невозможно вывести здание. Так без теории не может существовать наука". Поэтому для успешного исследования требуется как глубокое понимание особенностей исследуемого объекта (его структуры, функционирования тех или иных подсистем и т.д.), так и хорошее понимание алгоритмов использованных методов статистического анализа.

Эмпирические методы исследуют предмет на уровне явления, а теоретические – сущности. Исследование, состоящее из экспериментальной и теоретической частей, называется комплексным.

Повышение эффективности процедуры исследования заключается в расширении диапазона применяемых методов от его теоретической части до экспериментальной. Раскрытие диапазона исследования обеспечивает высокую эффективность данного процесса только при условии баланса теоретической и практической его части. В противном случае исследование будет сводиться к одной из его крайностей. Это условие становится особенно острым при исследовании биологических объектов. Создание такой дисциплины как биометрия основано на необходимости работы с объектом, параметры которого имеют низкую воспроизводимость вследствие открытости биологических систем.

Многие биологические исследования, традиционным образом, проводятся эмпирическим способом, то есть посредством последовательного перебора возможных вариантов. Проведение эксперимента без должного его планирования приводит исследование к крайности - “ползучему эмпиризму”. Такой подход требует большого количества проведения опытов и как следствие – значимых потерь времени и материальных средств. Кроме того, решение проблемных задач не всегда осуществимо стандартными методами. Применение моделирования, в качестве инструмента познания, повышает эффективность данного процесса посредством перехода к полуэмпирическим поискам. Эффективный процесс исследования, как и обучения, включает в себя модель как элемент теоретической части работы и эксперимент – практической. Моделирование не может и не должно быть альтернативой проведения эксперимента, а является его необходимым дополнением. Построение модели без ее последовательной экспериментальной проверки приводит к другой крайности в исследовании – “слепому академизму”.

Эксперимент занимает центральное место в науке. Однако возникает вопрос, насколько эффективно он используется. Джон Бернал, например, отмечал, что научные исследования организуются и проводятся настолько хаотично, что их коэффициент полезного действия, в отдельных случаях, может быть оценен величиной около 2 % [26]. Для того, чтобы повысить эффективность исследований, требуется нечто совершенно новое. Одним из возможных путей является применение математических методов моделирования эксперимента.

Поскольку исследование начинается с создания модели проведения эксперимента, то естественен и приоритет ее построения над практической частью работы. Проведение моделирования исследования позволяет установить не только возможность решения данной задачи, но и определить условия, обеспечивающие минимальные затраты для разрешения поставленной проблемы. Построение и совершенствование модели осуществляется ступенчато, начиная с формулировки рабочей гипотезы и ее проверки в ходе проведения экспериментальной работы и заканчивая обобщением полученных результатов до уровня математического абстрагирования. Хотя рекомендаций, позволяющих проводить подобного рода обобщения, достаточное количество [1-9, 22, 24, 28-31, 33, 35], алгоритмов связывающих в единую цепь, процесс от постановки задачи исследования до построения аналитической модели, в доступной нам литературе пока ещё нет. В данной работе сделана попытка собрать воедино все то, что может помочь начинающему исследователю понять структуру исследования и реализовать максимально обобщенный результат (таблица 1).

Таблица 1. Структурная схема построения модели эксперимента

№	Этапы выполнения	Способ выполнения	Результат
1	Постановка цели	Литературный поиск	Нерешенность проблемы
		Рабочая гипотеза	Возможность решения
2	Определение параметров	Составление целевой функции	План выполнения
3	Выбор методов	Измерения	Абсолютная, относительная величина
		Вычисления	Кач- количественный статистический анализ
4	Решение задачи	Схема опыта	Таблица
		Поисковый опыт	Оптимизация 1-3 этапов
		Основной эксперимент	Графическая зависимость
5	Анализ и обобщение результатов	Описание, прогноз	Регрессионное выражение
		Поиск причинно-следственной связи	Аналитическое уравнение

В таблице обозначены наиболее важные этапы выполнения работы, способ их реализации и результат. На первом этапе построения модели определяется возможности решения выявленной проблемы, при условии её нерешённости к настоящему времени. Обычно это делается по следующей схеме. В начале проводится анализ литературы с целью определения степени изученности выбранного объекта и предмета исследования.

**Объектом** исследования называют не биообъект или технологию, а исследуемый феномен (явление) со всеми относящимися к делу внешними компонентами, включая пространство (распространение) и время (динамика). Объектом конкретного биотехнологического исследования выступает ограниченная во времени и пространстве биосистема.

**Предмет** исследования – часть, сторона объекта, то, что находится в границах объекта и подлежит непосредственному изучению (значимые свойства, стороны, особенности объекта), это тот аспект объекта исследования, о котором исследователь намерен получить новое знание, какие отношения, свойства, аспекты, функции объекта раскрывает данное исследование (тему исследования определяет предмет). Объект и предмет соотносятся как категории общего и частного.

При анализе предмета исследования выделяются положительные и отрицательные аспекты заданной области. Положительными аспектами является наличие информации о структуре предмета исследования, а отрицательным её отсутствие. Потребность в недостающей информации появляется в том случае, когда определены данные, обрисовывающие границы известного и обнажающие края неизвестного. Аналитический обзор литературы сводится к поиску путей трансформации отрицательных аспектов исследуемого предмета в положительные, а имени способу перевода незнания в знание. Это помогает

определить проблему, разрешение которой обозначит тему исследования.

#### **1.4.1. Определение темы исследования, постановка цели**

Определить проблему и тему – значит выразить основное противоречие, которое должно быть разрешено средствами науки. Специфической чертой проблемы является то, что для ее решения необходимо выйти за рамки старого, уже достигнутого знания, ставя ее, надо ответить на вопрос: «Что надо изучить из того, что раньше не было изучено?». Тема исследования комплиментарна проблеме как вопрос и ответ. Конкретность постановки вопроса, определяет ясность ответа. Так и с определением темы исследования, чем чётче выявлена проблема, тем проще сформулировать название темы. Проблема для исследователя выполняет функцию «трамплина» для его развития, тогда как для обывателя это всегда преграда.

Формулирование темы научной работы «является наиболее важной частью работы исследователя». Название научной работы должно дать возможность однозначно понять, о чем идет речь в работе. Считается, что правильно выбрать и сформулировать тему – значит уже наполовину выполнить научную работу. Окончательно тема может быть сформулирована только после того, как основная часть работы выполнена.

Таким образом, в начале любого исследования необходимо выявить проблему исследуемого предмета, процедура решения которой поможет определить тему.

На основе выбранного, объекта и предмета исследования формулируется цель. Цель ставится непосредственно после проблемы, когда определены названия темы исследования ее объект и предмет.

Цель исследования – решение поставленной (сформулированной) проблемы. Поэтому целью исследования является переформулированная проблема. Цель в обобщенном виде характеризует итог исследования, что проявляется в виде неких закономерностей исследуемых причинно следственных связей и их механизмов. Цель служит постоянным критерием эффективности выполненных действий. Задача в определении цели исследования заключается в выявлении решения выбранной проблемы, которое, по сравнению с другими, обеспечивает более высокую эффективность. Для этого необходимо располагать каким-то количественным показателем или критерием эффективности. Этот критерий выбирается так, чтобы он наилучшим образом отражал целевую направленность исследований.

Выбирая решение, мы стремимся обратить этот критерий в максимум (или в минимум). Например, если целью криоконсервирования является получение высокой сохранности замороженно-оттаянных биообъектов, то естественным критерием эффективности в этом случае будет средняя ожидаемая выживаемость ("средняя", а не просто "выживаемость", поскольку фактическая выживаемость может колебаться в зависимости от случайных обстоятельств), и этот критерий желательно обратить в максимум.

В научных работах часто некорректность формулирования цели исследования возникает, когда вместо определения намечаемого научного результата



– нового научного знания, что должно быть основным итогом любого исследования, авторы подменяют целями практическими. Такие цели, как: «повышение эффективности ...»; и т.д. и т.п. – это не цели научного исследования. Научные результаты в дальнейшем, конечно, при определенных условиях (внедрении и т.п.) могут стать основой для «повышения эффективности» и т.д., но это нельзя ставить как цель исследовательской работы. И даже такая формулировка, как «разработать научно обоснованные рекомендации», может, очевидно, выступать лишь как сопутствующая, вспомогательная, но не основная цель исследования, а, скорее, даже как одна из задач, способствующая повышению практической значимости исследования.

В качестве целей научных исследований могут быть: создание новых концепций в области; разработка новых методик, технологий; разработка новых путей и средств совершенствования управления процессом, форм и методов работы; разработка новых теоретических или прикладных положений, которые стимулируют развитие новых форм, методов.

Цель исследования определяется такими понятиями как актуальность выбранной темы исследования её новизна и практическая значимость.

Актуальна та (и такая) тема, которая отвечает требованию научной новизны, социальным запросам и потребностям теории и практики организации процесса. **Актуальность** характеризует степень расхождения между спросом на научные идеи и практические рекомендации (для удовлетворения той или иной потребности) и предложениями, которые может дать наука и практика в настоящее время. Главное – показать суть проблемной ситуации, из чего и будет видна необходимость и своевременность, актуальность работы.

Актуальность исследований принято классифицировать следующим образом: высокоактуальные исследования – по проблеме, решение которой остро необходимо; актуальные исследования – проблема в науке разработана слабо, имеется достаточно выраженная потребность в ее решении; малоактуальные – потребность в разработке темы незначительна, проблема в целом изучена удовлетворительно; неактуальные исследования – для практики проблема не значима, полученные данные дублируют полученные представления.

**Научная новизна** результатов исследования характеризует одну из основных содержательных сторон – новые теоретические положения, которые ранее не были известны и не зафиксированы в науке и практике; этот компонент показывает насколько решены поставленные задачи. Глаголы «разработаны», «раскрыты», «обоснованы», «определены» и т.д. характерны для верных формулировок научной новизны.

Выделяют следующие варианты формулировок научной новизны исследования: наличие теоретических положений, которые впервые сформулированы и содержательно обоснованы; уточнение определений понятий или новое определение понятий; разработка рекомендаций (теоретических, методических), внедренных в практику и оказавших существенное влияние на достижение новых результатов; новые положения, которые способствуют дальнейшему развитию науки в целом или отдельных ее направлений.

Новизна исследования зависит от степени изученности предмета и надежности применяемых методов. Поэтому новизну научного исследования принято ранжировать в соответствии со значимостью полученных результатов, которая зависит от разработки новых методов исследования (измерения и обработки результатов) и (или) применения старых методов исследования к новому предмету. Использование нового предмета и (или) метода исследования дает возможность вначале выявить и описать закономерности раскрывающие его структуру, а в последствии раскрыть механизм причинно следственной связи исследуемого предмета.

**Практическая значимость** – ответ на вопрос: что дает практике данное исследование. Какие изменения произойдут (произошли) в результате внедрения полученных результатов в практику, в процесс, методику и т.д. Практическая значимость зависит: от числа и состава пользователей результатов работы; масштаба внедрения; степени готовности результатов к внедрению; предполагаемого эффекта от внедрения. Количественные показатели, отражающие практическую значимость исследования – экономический эффект, эффективность, коэффициент полезного действия.

Практическая значимость может быть очень высокая, высокая, удовлетворительная и низкая, соответственно, если: предполагаемый масштаб внедрения общегосударственный; результаты исследования важны для целой области биотехнологии; в результатах заинтересован очень широкий круг потребителей; внедрение полученных результатов в практику экономически целесообразно; результаты исследования могут быть рекомендованы к внедрению, разработке нормативных материалов, программ, учебников, предписаний к деятельности специалиста.

#### **1.4.2. Формулировка рабочей гипотезы и задач исследования**

Формой предвосхищения результата попытки «прорыва» в неизведанную область и определения в основных чертах итога работы является гипотеза. Гипотезой считается предположение, представляемое автором исследования как априорно вероятные результаты работы. Гипотеза отражает опыт и интуицию исследователя. Она не должна содержать логических противоречий, должна иметь характер предсказательности и представлять возможность проверки ее осуществления. В гипотезе находят отражение предполагаемые возможности поиска истины.

**Гипотеза исследования** – совокупность теоретически обоснованных предположений, истинность которых подлежит проверке. Поэтому принято считать, что гипотеза это возможная, но еще не доказанная истина.

Гипотеза считается фокусом исследования, поскольку определяет эпицентр создание гипотетической теории исследуемого объекта. Поэтому гипотеза является прообразом предполагаемой научной теории. В науке различают: малую гипотезу (гипотеза – закономерность) – как предвосхищение закона; большую гипотезу (гипотеза – теория) – как целостное явление универсума (универсум – совокупность всего, что существует).

Гипотеза должна описывать причины и возможные следствия в изучаемых явлениях, а так же отражать их структуру. В связи с этим выделяют прогностические и объяснительные гипотезы. Прогностические гипотезы делают допущения о возможном развитии процесса или явления при реализации исследования, а объяснительные дают пояснение возможных следствий из определенных причин, а также характеризуют условия, при которых эти следствия будут реализованы.

Правильно сформулировать гипотезу исследования можно, если следовать схеме: «Если ... (идея и замысел), то ... (предполагаемый результат), так как ... (объяснение эффекта)». Иногда формулируют и вариативные гипотезы, версии – гипотезы – когда идеи и положения самой гипотезы содержат нечто неочевидное и нуждающееся в проверке. Приведенная выше схема формулирования гипотезы реализует все три основные ее функции: описательную («если ...» – далее идет описание того, что надо делать), прогностическую («то ...» – предполагается, какой последует результат воздействий исследователя), и объяснительную («так как ...» – объясняется почему, вследствие чего достигается эффект происходящего). Иногда авторы исследовательских работ ограничиваются гипотезами, где реализуются не все три функции, а лишь две из них. Это не считается существенным недостатком.

Тщательно изучив все материалы, относящиеся к цели и предмету исследования, как правило, формулируют несколько гипотез, из которых выбирают одну рабочую.

**Рабочая гипотеза** – наиболее вероятное научное предположение о развитии исследуемого явления и его объяснение. Это предположение не доказано, но в большей степени вероятно, чем другие альтернативные гипотезы.

Рабочая гипотеза устанавливает факторы (причины, основные условия, движущие силы), обуславливающие развитие исследуемого явления. В общем виде, это можно представить в виде следующей функциональной зависимости:

$$C(a_j, v_j, d_j), \quad (1)$$

где:  $C$  – целевая функция, численно отражающая поставленную цель исследования;  $a_j, v_j, d_j$  – параметры, определяющие развитие изучаемого явления управляемые –  $a_j$  (поддающиеся изменениям в ходе эксперимента), не управляемые –  $v_j$  и неопределяемые в ходе эксперимента –  $d_j$ .

Целевая функция определяет факторы, обуславливающие развитие изучаемого явления. При разработке целевой функции устанавливают независимые переменные (аргументы  $a_j, b_j$ ), определяющие изменение зависимой переменной (функции  $C$ ). В опытах измеряют характеризующие как факторы (аргументы  $a_j, b_j$ ), так и показатели развития явления (целевую функцию  $C$ ) [14].

Рекомендуется структурировать исследуемые параметры в виде таблицы, распределяя параметры по уровню их значимости (силы влияния на целевую функцию) (см. табл. 2). Способы определения силы влияния см. тему 4. На этом этапе выясняется количество исследуемых параметров и диапазон их варьирования. Эти данные определяют объем массивов собираемой информации,

способы регистрации объектов измерения и факторов среды, схемы опытов и т.п.

Таблица 2. Способ структурирования исследуемых параметров в табличном виде

Фактор	Наименование	Размерность	Диапазон измерения	Ранг
$x_1$	Температура	°С	100÷200	1
$x_2$	Давление	кг/см <sup>2</sup>	2÷6	2
..	..	..	..	..
$x_n$	Плотность	масс. %	20÷40	8

Это весьма важная исходная часть поиска; при неправильном определении параметров исследование может быть безрезультатным, так как пойдет не в том направлении, по которому развивается явление, и поэтому приведет к ложным выводам.

Задачи исследования зачастую формулируют после гипотезы, хотя логически верно их увязывать с проблемой и целью исследования. Взаимосвязь проблемы и задачи исследования можно отразить следующим образом – проблема это задача, для которой не определены условия её решения и наоборот задача это проблема, для которой известны условия её решения.

Задачи обычно связаны с выделенными нами параметрами, которые планируется исследовать (анализировать, оптимизировать, описать графически и аналитически). То есть процедуру изучения выделенных параметров определяют задачи. Если стратегия исследования определяется её целью (целевой функцией), то тактика задачами (параметрами). Задачами («Что сделать?») отмечаются шаги к цели, это мост между ней и конкретными средствами ее достижения. Структуру исследования можно отразить следующим образом: проблема (тема, цель) – задачи – опыты - анализ - выводы – основной результат. Ясность и простота описания структуры исследовательской работы отражается в форме изложения полученных результатов. Примером диалектического единения содержания и формы исследования может, быть сияния света, как результат правильной огранки кристалла, где форма определяется симметрией граней, а содержание чистотой камня.

**Задачи исследования** – это руководства к действию, указания, как делать и что будет, получено в результате, если предпринять такие-то действия. Задачи исследования – формулируются в соответствии с целью, как ее конкретизация (изучить..., описать..., установить..., и т.д.). Формулировать задачи надо тщательно, поскольку описание их решения и есть содержание глав работы, да и заголовки глав рождаются именно из формулировок задач предпринимаемого исследования. Определение задач исследования должно позволить более конкретно, поэтапно реализовать цель исследования. Правильно поставленные задачи исследования помогают автору логически правильно и последовательно проводить теоретическую и экспериментальную часть работы.

В качестве некоторых элементов задач научного исследования, могут быть рекомендованы: решение теоретических вопросов, входящих в проблему исследования (введение в научный оборот новых понятий, раскрытие их сущности и содержания; разработка критериев и показателей оценки эффективности исследуемого процесса; принципов, условий и факторов применения тех или иных методов и методик); выявление тенденций и закономерностей изучаемых процессов; определение путей и средств совершенствования методов; экспериментальное изучение исследуемых вопросов, выявление их типичных компонентов и состояний; экспериментальная проверка разработанных систем.

Задачи исследования должны быть соизмеримыми и взаимосвязаны по своей значимости с силой влияния параметров на целевую функцию (см. тему 4). Обычно их выдвигают не более трех или четырех, относя более частные задачи в качестве подзадач к одной из основных. Единого стандарта в формулировке задач быть не может, но все же чаще всего первая из выдвигаемых задач связана с выявлением сущности, природы, структуры, законов функционирования и развития изучаемого объекта; вторая задача нацелена на раскрытие общих способов преобразования объекта, на построение его моделей; третья задача – на создание, разработку конкретных методик, практических рекомендаций.

Точность определения исследуемых задач и их взаимосвязь определяет структуру построения полученных выводов так же как конкретность и последовательность постановки вопросов ясность и логичность ответов.

#### **1.4.3 Планирование эксперимента, составление методики исследования.**

До начала проведения эксперимента желательно осуществить его планирование. В научно-исследовательской работе рациональная схема исследования (план), естественно, должна составляться таким образом, что бы при минимальной затрате времени и средств получать максимум информации об интересующем нас объекте. Именно в этом смысле можно говорить, что научные методы планирования исследования дают возможность ускоренного достижения поставленной цели исследования, сокращая затраты времени как на собственно эксперимент, так и на статистическую обработку результатов (см. приложение).

*Планирование эксперимента* – это процедура выбора условий проведения опытов, необходимых и достаточных для решения поставленной задачи с требуемой точностью. Планирование эксперимента предполагает активное вмешательство в процесс и возможность определения условия решения поставленной задачи. Поэтому такой эксперимент называется активным. Объект, на котором возможен активный эксперимент, называется управляемым. Это и есть необходимое требование к объекту исследования [7, 8].

Применение планирования эксперимента делает поведение экспериментатора целенаправленным и организованным, существенно способствует повышению производительности его труда и надежности полученных результатов. Важным достоинством метода является его универсальность, пригодность в

огромном большинстве областей исследования, интересующих современного человека.

Планирование эксперимента еще совсем молодая область. Она бурно развивается и вызывает все больший интерес у исследователей. Интерес вполне понятен: перспектива сократить число опытов, найти оптимум, получить количественные оценки влияния факторов и определить ошибки – крайне привлекательна. Но, когда экспериментатор делает попытку познакомиться с планированием эксперимента, он часто сталкивается с серьезными трудностями. Больше того, иногда он просто не верно применяет методы планирования или выбирает не самый оптимальный для данной ситуации путь исследования, или допускает еще какие-нибудь досадные ошибки. При этом снижается эффективность его работы и появляется опасность дискредитации важного и полезного направления.

Поиск оптимальных условий является одной из наиболее распространенных научно-технических задач. Они возникают в тот момент, когда установлена возможность решения поставленной задачи и необходимо найти наилучшие (оптимальные в некотором смысле) условия его реализации.

В настоящее время применение статистических методов в биологии представляет выбор или создание математической модели, проверку её соответствия экспериментальным данным и анализ результатов, вытекающих из её рассмотрения. Выбор той или иной модели в значительной мере определяется биологической природой эксперимента. Планирование эксперимента стало самостоятельным разделом биометрии, что позволяет резко сократить объём эксперимента для получения того же количества информации. Для решения экспериментальных задач наиболее эффективно применение методов многомерной статистики, позволяющих одновременно оценить не только влияние нескольких факторов, но и взаимодействие между ними. Эти методы находят всё большее применение для решения задач систематики. Поэтому математическая модель является как целью, так и средством проведения исследования.

Поиск оптимальных условий, построение интерполяционных формул, выбор существенных факторов, оценка и уточнение констант теоретических моделей (например, кинетических), выбор наиболее приемлемых из некоторого множества гипотез о механизме явлений, исследование диаграмм состав-свойство – вот примеры задач, при решении которых применяется планирование эксперимента. Можно сказать, что там, где есть эксперимент, имеет место и наука о его проведении – планирование эксперимента.

Методы планирования, и в первую очередь статистические методы планирования, существенно отличаются от методов статистической обработки результатов наблюдений, где количественному анализу отводится пассивная роль обработки данных, полученных при случайном сочетании условий. В противоположность этому математический аппарат методов планирования экспериментов играет активную роль, определяя и даже, можно сказать, диктуя исследователю жесткую схему постановки эксперимента и последовательность анализа результатов. Тем не менее, это совершенно не означает, что при использовании методов планирования эксперимента знание и опыт самого экс-

периментатора не играют существенной роли. Наоборот, неточно сформулированная задача или неверно выбранные отправные данные, “пройдя” весь путь математического аппарата планирования эксперимента и анализа полученных результатов, в ходе которого исключается возможность внесения корректив и поправок, приводят хотя и к быстрому, но неполноценному решению. Для этого необходимо определить целевую функцию, численно отражающую поставленную цель исследования и параметры, определяющие развитие изучаемого явления. До тех пор пока биотехнолог достаточно точно не определит цель исследования, природа будет отвечать уклончиво, как дельфийский оракул. Наиболее простой формой модели эксперимента является план его выполнения.

Следующим этапом моделирования исследования является выбор методов измерения и обработки полученных результатов. При всем разнообразии методов исследовательской работы задача моделирования эксперимента сводится к тому, чтобы при возможно минимальных объемах наблюдений получать достаточно полную информацию об изучаемых объектах.

**Методика исследования** – это совокупность способов, приемов исследования, адекватных поставленным задачам. Она отвечает на вопрос: как и какими способами, проводить исследования, направленные на установление причинно-следственных связей между целевой функцией и параметрами, определяющими его величину? Методика, которая относится ко всему исследованию и представляет стержневые способы и приемы, проходящие через все исследования, является общей. Общая методика включает в себя комплекс методов измерения исследуемых параметров и их обработки.

Выбор и обоснование методов работы следует делать, изучив предварительно возможности принятых методов и средств исследования, обратив особое внимание на целесообразность и специфику системного подхода. Разрабатывая методику экспериментальной работы по теме, следует: уяснить специфику эксперимента; описать его возможные этапы, их цели и задачи; указать целесообразность методов и средств организации и проведения каждого этапа эксперимента; учесть взаимосвязь выбранных способов измерения и обработки результатов эксперимента

Для повышения надежности полученных результатов необходимо применять разные методы измерения и обработки полученных данных (проводить параллельные опыты и вычисления), что называется сценарным подходом.

При разработке методики следует стремиться получить в опытах данные, которые не только устанавливают закономерности развития явлений, но и помогают найти объяснение этих закономерностей, их причины, физическую основу явления. Чем полнее, тщательнее и продуманнее составлена методика исследования, тем легче и скорее выполнить исследование. Ясность, точность и достаточно подробное (но краткое) изложение особенно важны в тех случаях, когда в исследовании принимает участие группа работников.

Правильная методика - залог успеха исследования. Не методичность исследования или неверная методика приводят к ошибкам, в результате чего

накапливаются груды "сырого" материала, из которых никаких полноценных выводов сделать нельзя.

Знание новых зависимостей позволяет генерировать новые гипотезы о механизмах исследуемых факторов воздействия на изучаемые объекты – методов работы с биообъектом. Эту информацию можно считать наиболее ценной, поскольку именно она позволяет исследователю решать задачу прогноза. "Знать, чтобы предвидеть, предвидеть – чтобы управлять" – эти слова Огюста Куно как нельзя лучше иллюстрируют ценность информации о изучаемых зависимостях.

Очевидно, что в такой ситуации необходимо использовать статистические методы, которые обеспечивают максимально достижимую надежность результатов анализа. Профессионал, исследующий интересующие его объекты, всегда будет относиться к ним как к системе. А любая система, как известно, включает в себя не только отдельные элементы, но и структуру, объединяющую их в единое целое. При этом структура биологических объектов чаще всего проявляется в виде разнообразных статистических связей. Кроме простых последовательных структур причинно-следственных связей существуют и более сложные структуры, например, параллельно-последовательные, древовидные, матричные, с обратной связью и т.д. В итоге применения адекватных методов статистического анализа появляется возможность отфильтровать случайную составляющую и получить глубинную информацию о структуре исходных признаков и исследуемых групп.

Основные проблемы применения методов обработки полученных результатов подобны сложностям их получения (методам измерения). Например, каковы проблемы применения замораживателей или криопротекторов в криобиологии? Классифицируя их, можно выделить следующие уровни: 1) материальный – наличие необходимого криопротектора или замораживателя; 2) информационный – наличие необходимых знаний, опыта, методик, программ и т.д.; 3) организационный – наличие необходимых условий проведения исследования, заинтересованности руководителей и исполнителей и т.п. Задачи 1-го уровня зависят в большей степени от внешних условий (они объективны), тогда как 2-го и 3-го уровней от внутренних (субъективны). Проблема 2-го уровня стоящая перед профессионалом, работающим в области биологии, заключается в умении применять методы биометрии. Поэтому ведущей причиной ограничения использования математического моделирования в биологии является отождествление дисциплин биометрии и статистики.

Умение определять условия решения поставленной задачи является критерием оценки специалиста в овладении им методов планирования биологического эксперимента. Отсутствие плана исследования приводит к тому, что поставленная задача может, не имеет решений или выводы статистического анализа собранных данных, отвечают на отдельные, узкие задачи исследования, которые оказываются не адекватными ранее поставленной цели. Скудность подобных выводов отражает несостоятельность гипотез и предположений, которыми оперирует исследователь, опираясь только на факты без должного их анализа. Таковыми они становятся лишь после "экстракции" из них, в том чис-



ле и методами статистики, надежных закономерностей, обобщений, отбрасывания случайной, "шумовой" информации.

Неумение применять методы планирования исследования (2 и 3 -я субъективные задачи) сдерживает решение и 1 -й - объективной задачи. До тех пор пока интересы исследователя ограничены сферой только эксперимента, решение этих проблем невозможно. Полагаем, что первопричина данной проблемы лежит в инерционности мышления большинства современных биологов, что предопределено историей развития данной науки.

Таким образом, структурирование информации основано на выделении из шума скрытых закономерностей изучаемого явления при помощи методов математического абстрагирования. Методология исследования определяет направление его проведения от выявления проблемы до её разрешения. Повышение эффективности исследования основано на комплексном использовании экспериментальной и теоретической части.

### ВОПРОСЫ ДЛЯ САМОКОНТРОЛЯ.

- 1 В чем заключена необходимость структурирования полученной информации и каковы способы её реализации?
- 2 Дайте определение понятию биометрия. На основе, каких наук изучается курс биометрия и в чем заключено его основное отличие от статистики?
- 3 Отрадите основные этапы истории и направления развития статистики.
- 4 Изложите основные проблемы биометрии и способы их решения.
- 5 Как осуществляется повышение эффективности процедуры исследования?
- 6 Перечислите приоритеты проведения теоретического и экспериментального исследования.
- 7 Какое исследование называется комплексным?
- 8 Перечислите этапы построения модели эксперимента.
- 9 Какова основная цель исследования?
- 10 Для чего необходима рабочая гипотеза?
- 11 Какой смысл вкладывается в понятие планирование эксперимента?
- 12 Как составляется методика исследования?

## Глава 2. Первичная статистическая обработка данных

### 2.1 Сводка и группировка данных

Знания о природе приобретаются путем наблюдения, сравнения и опыта. При чем под наблюдением в широком смысле подразумевают процесс планомерного добывания и накопления фактов независимо от того, как оно осуществляется - в эксперименте или непосредственным описанием изучаемого предмета. Но факты - это еще не наука. Как груда строительных материалов не является зданием, так и масса накопленных фактов не составляет содержание науки. Только сведенные в некую систему факты приобретают определенный смысл, позволяют извлечь заключенную в них информацию. Эта работа требует от исследователя не только профессионального мастерства, но и умения правильно планировать эксперименты, анализировать их результаты, делать из фактов научно обоснованные выводы. Система таких знаний и составляет содержание биометрии — науки, призванной играть хотя и вспомогательную, но весьма важную роль в биологических исследованиях.

От наблюдения к обобщению и от обобщения к проверке на практике - таков общий путь исследования. Эти три ступени обязательны для законченного исследования. Задача научного исследования состоит в том, чтобы установить закономерности развития явлений между величиной целевой функции и определяющих ее параметров. При определении закономерности мы ищем постоянное и необходимое отношение, связь между явлениями, лежащую в основе их развития.

Наблюдение применяют для предварительной проверки общей правильности рабочей гипотезы, т.е. для установления направленностей развития явлений. Эксперимент - видоизменение условий развития явлений  $a_j$   $\phi(1)$  в желательном направлении, чтобы получить ясные закономерности. Выведенные закономерности оказываются неясными и даже неверными в случае, если полученные значения лежат в пределах возможной ошибки опыта. Ошибка опыта зависит от многих факторов (см. главу 3) в том числе от способа группировки полученной информации.

Простейшая группировка данных проводится на различных этапах исследования: сбор информации (статистическое наблюдение); обработка информации (сводка и группировка); анализ информации (индексный, дисперсионный, корреляционно-регрессионный); выводы.

**2.1.1. Сбор и накопление данных об исследуемых биологических явлениях.** При сборе данных важно помнить правило «единообразия и однородности» собираемых данных, чтобы свести к минимуму грубые и систематические ошибки (см. главу 3), уменьшающие точность измерений. Это условие относится к способу формирования выборок, суть которого заключается в создании одинаковых условий наблюдения и обеспечении равной вероятности получаемых результатов в противном случае состав выборки будет не однородным и статистические законы будут проявляться «неправильно», что сделает невозможным применение точных статистических критериев.

Генеральная совокупность – это множество всех вариантов определенного типа (выборка бесконечного размера). Чаще всего получить все возможные значения в принципе невозможно. Поэтому судить о генеральной совокупности приходится, исследуя выборки – по части составлять представление о целом.

Выборка – множество значений случайной величины, совокупность вариантов, набор чисел; отдельная варианта – это объект, несущий качественный или числовой признак. Термин «выборка» указывает на процесс выбора части из чего-то большего, в данном случае – на процесс получения ограниченного количества значений из генеральной совокупности.

Варианта качественно или количественно выражает признак данного объекта исследования (полученного при данном уровне фактора внешней среды вполне определенным методом).

Признак (свойство, показатель, величина, характеристика, переменная) – любая информация о наблюдаемом объекте, выраженная качественно или определенная количественно. В рамках вариационной статистики признаки выступают в роли случайной величины.

Случайная величина – численная характеристика, принимающая те или иные заранее точно не известные значения. Несмотря на то, что точное описание поведения случайной величины получить нельзя, математическая статистика позволяет выполнить вероятностное описание.

**2.1.2 Форма учета результатов наблюдений.** Наблюдения над биологическими объектами проводят обычно по принятой исследователем программе. Результаты наблюдений фиксируют в дневниках, журналах, бланках, анкетах или других документах учета. Существует много различных форм и способов учета; выбор той или иной формы определяется задачей исследования и теми условиями, в которых оно проводится. Процедура регистрации исследуемых данных осуществляется с учетом её группировки.

**2.1.3 Способы группировки первичных данных.** Зафиксированные в документах учета сведения об изучаемом объекте (или объектах) представляют тот первичный фактический материал, который нуждается в соответствующей обработке. Обработка начинается с упорядочения или систематизации собранных данных. Процесс систематизации результатов массовых наблюдений, объединения их в относительно однородные группы по некоторому признаку называется группировкой.

Группировка — это не просто технический прием, позволяющий представить первичные данные в комплексном виде, но и глубоко осмысленное действие, направленное на выявление связей между явлениями. Ведь от того, как группируется исходный материал, во многих случаях зависят выводы о природе изучаемого явления. Один и тот же материал дает диаметрально противоположные выводы при разных приемах группировки. Нельзя группировать в одну и ту же совокупность неоднородные по составу данные, необдуманно выбирать способ группировки» Группировка должна отвечать требованию поставленной задачи и соответствовать

содержанию изучаемого явления. Наиболее распространенной формой группировки являются статистические таблицы. Для структурирования информации в виде таблицы необходимо познакомиться со свойствами выборки.

Прежде чем приступить к накоплению экспериментально полученных данных необходимо определить категорию, к которой они относятся. Существует целый ряд категорий определяющих исследуемые признаки биологических объектов.

**Качество** (нечисловой дискретный признак) – простой, непосредственный, чувственный способ регистрации фактов; это статус, сезон, таксон, цвет, плотность, тип действия и пр. Значения таких признаков выражаются словами или символами, они не имеют количественного содержания и выражают принадлежность данного объекта к определенной обширной группе объектов (зеленый, январь, ♀, ♀).

**Балл** (оценка) – дискретный полуколичественный признак, численная характеристика объекта, присвоенная в соответствии с внешней заранее принятой шкалой баллов. Во время оценки объект соотносится с этими критериями и ему присваивается соответствующий балл.

**Количество** (число) – дискретный (счетный) количественный признак (число натурального ряда), характеризующий множество однородных объектов, черт, деталей строения, состав (например, число эмбрионов у самки, число жаберных тычинок у рыб, число тычинок в цветке, число деревьев на пробной площадке).

Необходимо учитывать, что качественные и полукачественные показатели, например обычные школьные отметки - баллы не подлежат статистической обработке! Перевод качественных и полуколичественных показателей в количественные есть обязательное условие для применения математической статистики. Путь объективизации данных (измерение доли усвоенной информации) и возможность их последующей статобработки возможен посредством умения классифицировать и трансформировать показатели по шкале измерения.

**2.1.4. Классификация шкал измерения.** Выделяют 4 типа шкал систематизирующие разные категории, которых представлены в таблице 3.

Шкалирование – это приписывание исследуемым объектам соответствующих им цифровых значений, что позволяет использовать количественные (математические) методы при изучении явлений, обработке результатов наблюдений и экспериментов и т.д. Все шкалы измерения делят на две группы - шкалы качественных и количественных признаков.

**Шкалы качественных признаков** - это шкалы **номинальная** и **порядка**. Номинальная шкала (шкала наименований), которую правильнее было бы считать классификацией, а не измерением, делит все объекты на группы по какому-либо признаку (различию). Этим признакам присваиваются определенные числа (код), что создает удобства при дальнейшей обработке экспериментальных данных. Никакого количественного соотношения между объектами в номинальной шкале нет. Примеры. Учащиеся класса делятся на две категории и обозначаются: девочки – 01, мальчики – 02. Группы нарушителей дисциплины

и их обозначение (кодирование): на уроке – 1, на улице – 2, дома – 3. В процессе проверки соответствия подготовки выпускников школ требованиям ГОС появляется группа аттестованных и не аттестованных учеников.

Таблица 3. Классификация шкал измерения

Группы (характер)	Шкалы измерений			
	Качественные, неметрические, концептуальные ( <i>слабые</i> )		Количественные, метрические, материальные ( <i>сильные</i> )	
тип	Классификации (наименований, номинальная)	Порядковая (ранговая)	Интервальная	Отношений
Практи- ческий аналог, пример	Число - лишь для обозначе- ния названия	Твердость минералов, сила ветра и др.	Время, темпе- ратура и др.	Масса, длина
Наличие – точки отсчета	Нет	нет	Условный нуль введен	Обоснован- ный нуль и есть единица измерения
Измери- тельные возмож- ности	Различение присутствия в объектах ис- следования разных качеств, свойств	Тоже + разли- чение превы- шения содер- жания одного и того же ка- чества в од- ном объекте по сравнению с другим	Тоже + оце- нивание вели- чины превы- шения содер- жания иссле- дуемого каче- ства в объек- тах	Тоже + имеет нулевую (•) отсчета, т.е. возможно фиксирование полного от- сутствия ис- следуемого качества в объекте
Какие арифме- тические действия допусти- мы	Арифметиче- ские действия недопустимы (бессмысленны)	тоже, но есть свойство ( $a > b, b > c, a > c$ ) транзи- тивности	Сложение – вычитание, т.е. «на сколь- ко $>$ или $<$ » можно отве- тить, но не «во сколько раз»	Все 4 арифме- тических дей- ствия, т.к. есть нуль шкалы – точка отсчета

Шкала порядка (порядковая, ранговая, ординальная) предназначена для измерения (обозначения) степени различия какого-либо признака или свойства у разных объектов. Самым ярким примером порядковой шкалы является пяти-балльная система оценки учащихся. Для нее разработаны критерии и различные методы измерения. Значительно труднее применять порядковую шкалу для количественных оценок других качеств личности (в воспитательном про-

цессе). Имеется несколько разновидностей порядкового шкалирования (измерения): ранжирование (в ряд), группировка (ранжирование по группам), парное сравнение, метод рейтинга, метод полярных профилей.

**Ранжирование.** Изучаемые объекты располагаются в ряд (упорядочиваются) по степени выраженности какого-либо качества. Первое место в этом ряду занимает объект с наиболее высоким уровнем данного качества, и ему присваивается наивысший **балл** (числовое значение выбирается произвольно). Затем каждому объекту ранжированного ряда присваиваются более низкие оценки, соответствующие занимаемым местам.

**Группировка** всей совокупности объектов наблюдения в несколько **рангов**, достаточно ясно отличающихся друг от друга по степени измеряемого признака. Пример: учащиеся класса согласно пятибалльной системе оценки делятся на отличников, хорошистов и т. д.

**Парное сравнение.** Учащиеся сопоставляются друг с другом (каждый с каждым) по какому-либо качеству. Если они одинаковы, то каждый получает по баллу. Если у одного этого качества больше, чем у другого, первый получает два балла, второй – 0 (как при спортивных играх по круговой системе). Суммируя полученные каждым баллы, получаем количественное выражение уровня развития данного качества у каждого учащегося (его ранг).

**Рейтинг.** В этом приеме оценка объекта производится путем усреднения оценочных суждений группой компетентных экспертов. Имея общие критерии оценки (в порядковой шкале, в баллах), эксперты независимо друг от друга (в устной или письменной форме) выносят свои суждения. Усредненный результат экспертной оценки является достаточно объективным и называется рейтингом.

**Метод полярных профилей.** Этот прием предполагает применение для оценки условной шкалы, крайними точками которой являются противоположные значения признака (например, добрый - злой, теплый - холодный и т. п.). Промежуток между полюсами делится на произвольное количество частей (баллов). Пример. Оценка степени доверия кандидату на выборную должность дается в полярной шкале: (Доверяю полностью) 10 – 9 – 8 – 7 – 6 – 5 – 4 – 3 – 2 – 1 (Совсем не доверяю)

В различных областях человеческой деятельности применяется много других видов порядковых шкал. Так, например, в минералогии используется шкала Мооса, по которому минералы классифицируются согласно критерию твердости. А именно: тальк имеет балл 1, гипс - 2, кальций - 3, флюорит - 4, апатит - 5, ортоклаз - 6, кварц - 7, топаз - 8, корунд - 9, алмаз - 10. Минерал с большим номером является более твердым, чем минерал с меньшим номером, при нажатии царапает его.

Порядковыми шкалами в географии являются - бифортова шкала ветров ("штиль", "слабый ветер", "умеренный ветер" и т.д.), шкала силы землетрясений. Очевидно, нельзя утверждать, что землетрясение в 2 балла (лампа качнулась под потолком - такое бывает и в Москве) ровно в 5 раз слабее, чем землетрясение в 10 баллов (полное разрушение всего на поверхности земли).

В медицине порядковыми шкалами являются - шкала стадий гипертонической болезни (по Мясникову), шкала степеней сердечной недостаточности (по Стражеско-Василенко-Лангу), шкала степени выраженности коронарной недостаточности (по Фогельсону), и т.д. Все эти шкалы построены по схеме: заболевание не обнаружено; первая стадия заболевания; вторая стадия; третья стадия... Иногда выделяют стадии 1а, 1б и др. Каждая стадия имеет свойственную только ей медицинскую характеристику. При описании групп инвалидности числа используются в противоположном порядке: самая тяжелая - первая группа инвалидности, затем - вторая, самая легкая - третья.

Номера домов также измерены в порядковой шкале - они показывают, в каком порядке стоят дома вдоль улицы. Номера томов в собрании сочинений писателя или номера дел в архиве предприятия обычно связаны с хронологическим порядком их создания.

При оценке качества продукции и услуг, в т.н. квалиметрии (буквальный перевод: измерение качества) популярны порядковые шкалы. А именно, единица продукции оценивается как годная или не годная. При более тщательном анализе используется шкала с тремя градациями: есть значительные дефекты - присутствуют только незначительные дефекты - нет дефектов. Иногда применяют четыре градации: имеются критические дефекты (делающие невозможным использование) - есть значительные дефекты - присутствуют только незначительные дефекты - нет дефектов. Аналогичный смысл имеет сортность продукции - высший сорт, первый сорт, второй сорт,...

При оценке экологических воздействий первая, наиболее обобщенная оценка - обычно порядковая, например: природная среда стабильна - природная среда угнетена (деградирует). Аналогично в эколого-медицинской шкале: нет выраженного воздействия на здоровье людей - отмечается отрицательное воздействие на здоровье.

**Шкалы количественных признаков - это шкалы интервалов и отношений. Интервальная** шкала (интервальное намерение) - это такое присвоение чисел объектам, когда определено расстояние между объектами и предусмотрена общая для всех объектов постоянная единица измерения. Иначе говоря, в интервальной шкале вводится единица и масштаб измерения. Нулевая точка шкалы выбирается произвольно. Примеры: температурные шкалы; шкалы стандартизированного тестирования интеллекта.

Интервальная шкала – количественная. В ней возможны все арифметические действия над числами, кроме операции деления. Таким образом, в интервальной шкале нельзя определить во сколько раз один объект больше или меньше другого. Например, если ученик ответил правильно на 10 заданий, то это не означает, что он знает вдвое больше ученика, ответившего на 5 заданий теста.

По шкале интервалов измеряют величину потенциальной энергии или координату точки на прямой. В этих случаях на шкале нельзя отметить ни естественное начало отсчета, ни естественную единицу измерения. Исследователь должен сам задать точку отсчета и сам выбрать единицу измерения. Допустимыми преобразованиями в шкале интервалов являются линейные возрастаю-

щие преобразования, т.е. линейные функции. Температурные шкалы Цельсия и Фаренгейта связаны именно такой зависимостью:  $^{\circ}\text{C} = 5/9 (^{\circ}\text{F} - 32)$ , где  $^{\circ}\text{C}$  - температура (в градусах) по шкале Цельсия, а  $^{\circ}\text{F}$  - температура по шкале Фаренгейта.

Шкала отношений отличается от интервальной только тем, что ее нулевая точка не произвольна, а указывает на полное отсутствие измеряемого свойства. Сюда относятся и все количественные данные, получаемые пересчетом объектов какого-либо множества (число учащихся, уроков и т. п.). Шкалы интервальная и отношений представлены абсолютными величинами.

**Абсолютные статистические величины** Абсолютные статистические величины показывают объем, размеры, уровни различных социально-экономических явлений и процессов. Они отражают уровни в физических мерах объема, веса и т.п. В общем, абсолютные статистические величины – это именованные числа. Они всегда имеют определенную размерность и единицы измерения. Последние определяют сущность абсолютной величины. Абсолютные величины являются основой для расчета разных относительных статистических показателей. Основа построения объяснительных моделей.

**Относительные статистические величины.** Относительные статистические величины выражают количественные соотношения между явлениями, они получаются в результате деления одной абсолютной величины на другую. Знаменатель (основание сравнения, база) – это величина, с которой производится сравнение. Сравнимая (отчетная, текущая) величина – это величина, которая сравнивается. Относительная величина показывает, во сколько раз сравниваемая величина больше или меньше базисной или какую долю первая составляет по отношению ко второй. В ряде случаев относительная величина показывает, сколько единиц одной величины приходится на единицу другой. Важное свойство – относительная величина абстрагирует различия абсолютных величин и позволяет сравнивать такие явления, абсолютные размеры которых непосредственно несопоставимы. Основа получения сопоставимых, а следовательно и воспроизводимых величин (см тему 6).

В процессе развития соответствующей области знания тип шкалы может меняться. Так, сначала температура измерялась по порядковой шкале (холоднее - теплее). Затем - по интервальной (шкалы Цельсия, Фаренгейта, Реомюра). Наконец, после открытия абсолютного нуля температуру можно считать измеренной по шкале отношений (шкала Кельвина). Надо отметить, что среди специалистов иногда имеются разногласия по поводу того, по каким шкалам следует считать измеренными те или иные реальные величины. Другими словами, процесс измерения включает в себя и определение типа шкалы (вместе с обоснованием выбора определенного типа шкалы). Кроме перечисленных четырех основных типов шкал, иногда используют и иные шкалы.

Уровни измерения и математические вычисления, используемые на разных уровнях, показаны в табл. 4. Из этой таблицы видно, что переход от одного уровня к другому сопровождается расширением класса допустимых математико-статистических операций. Как следует из таблицы, наилучшей являет-



ся шкала отношений, которую на сегодняшний день удалось реализовать только в рамках физических измерений.

Таблица 4. Шкалы количественных и качественных признаков

Шкала	Математические и статистические величины, вычисление которых допустимо на данном уровне
Номинальная	Мода, процентные частоты, доли, корреляция
Порядковая	Мода, медиана, квартили, коэффициент корреляции, дисперсионный анализ
Интервальная	Мода, медиана, квартили, коэффициент корреляции, ранговые критерии, средняя, дисперсия, стандартное отклонение, коэффициент корреляции
отношений	Все арифметические операции, все понятия и методы математической статистики

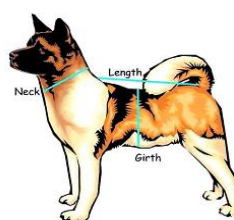
### 2.1.5 Точность измерений. Действия над приближенными числами.

Применяя биометрию к решению практических задач, исследователь имеет дело с измерениями биологических объектов. Обычно измерения проводят с точностью до десятых, сотых или тысячных долей единицы, более точные измерения производят реже. Как показывает опыт, нет необходимости в точности измерений, когда эта точность практически не нужна. Чтобы избежать грубых ошибок в работе и получать сопоставимые результаты, необходимо неукоснительно соблюдать признанные правила записи и округления приближенных чисел. Очень важно, чтобы числа, фиксируемые в документах учета, соответствовали требуемому уровню надежности проводимого исследования (см. тему 3). Так, если измерения проводят с точностью до одного десятичного значимого знака, то результаты измерений нельзя записывать, например, в таком виде: 5,2; 4; 4,69; 4,083 и т. д. Правильная запись этих чисел будет такова: 5,2; 4,0; 4,7; 4,1.

При округлении числа многие исследователи считают более точным такое правило: если за последней сохраняемой цифрой следует цифра 5 (с нулями или без оных после нее), то округление осуществляется с недостатком при условии, что сохраняемая цифра четная. Если же сохраняемая цифра нечетная, то округление осуществляется с избытком. Например, числа 3,585 и 3,575 округляются до двух десятичных знаков таким образом: 3,58 и 3,58.

#### Непрерывные переменные:

1. Если вам почему-то необходимо округлить числа, чётные округляют в меньшую сторону, нечётные – в большую (2.5 в 2, 3.5 в 4);
2. Не нужно писать много знаков после запятой (ориентируются на точность измерений).



## 2.2 Построение вариационного ряда

**2.2.1 Варьирование.** Основная особенность выборки, как множества значений случайной величины – это отличие отдельных вариантов друг от друга, явление изменчивости, варьирования, появления отличий между отдельными вариантами.

Биологу важно знать обычные причины варьирования. Один из источников, эндогенный, – это индивидуальные отличия по статусу и по состоянию. Например, животные одного возраста различны индивидуально, генетически, т. е. по статусу. Кроме того, каждое из них в разные годы, сезоны, время суток имеет разные морфофизиологические характеристики, т. е. отличается по состоянию. В наиболее точных науках (токсикология, биохимия, молекулярная биология) стремятся с помощью химической чистоты постановки опытов и выведения чистых линий подопытных животных убрать все мешающие причины «избыточного» варьирования.

Другой источник отличий между вариантами – факторы внешней среды, т. е. условия проведения наблюдений, среда существования объекта, возможная причина, определяющая текущее состояние объекта. Часто говорят про факторы эндогенные, внутренние (статус, способ существования объекта), и экзогенные, внешние (среда, условия существования объекта). Фактор всегда есть активное, действующее начало, признак – его результат, последствие. Факторы, влияющие на значения вариантов, различаются по своей природе. Если фактор влияет на все варианты выборки постоянно и примерно одинаково, он называется систематическим (или доминирующим). Если фактор непостоянен, влияет на варианты не одинаково, с разной силой, он определяется как случайный.

С методической точки зрения при наблюдениях или в эксперименте самым важным оказывается обязательная регистрация максимально возможного числа факторов (как внешних, так и внутренних). Тогда появляется возможность исследовать их раздельное действие на объект, поскольку существуют методы, которые позволяют из многокомпонентной среды вычленять эффекты действия отдельных факторов (особенно работоспособны дисперсионный, регрессионный и компонентный анализы см. тему 4).

При самом широком варьировании признаков разброс значений выборки не бесконечно широк, каждая из них ограничена неким диапазоном и тяготеет к определенному общему значению. Эти свойства статистических совокупностей – варьирование, в ограниченном диапазоне, – позволяют предложить для описания две группы величин: оценку центрального значения диапазона (среднюю, моду или медиану) и оценку размаха варьирования (лимит, дисперсию, стандартное отклонение). Определение этих значений выполняется после построения вариационного ряда.

Любое статистическое исследование должно начинаться с установления характера распределения изучаемых признаков. Распределение – это соотношение между значениями случайной величины и частотой их встречаемости. Большая повторяемость одних значений по сравнению с другими заставляет

задумываться о причинах наблюдаемых процессов. Если значения признака откладывать по оси абсцисс, а частоты их встречаемости по оси ординат, то можно построить *гистограмму, частотную диаграмму*, удобную для целей иллюстрации и исследования.

**2.2.2 Вариационный ряд.** Основой для построения гистограммы служит вариационный ряд – представленный в виде таблицы ряд значений изучаемого признака, расположенных в порядке возрастания с соответствующими им частотами их встречаемости в выборке.

Полученная в результате статистического наблюдения выборка из  $n$  значений (вариант) изучаемого количественного признака  $X$  образует вариационный ряд. Ранжированный вариационный ряд получают, расположив варианты  $x_j$ , где  $j = 1, 2, \dots, n$ , в порядке возрастания значений, то есть  $x_1 \leq x_2 \leq \dots \leq x_j \leq \dots \leq x_n$ .

Изучаемый признак  $X$  может быть дискретным, то есть его значения отличаются на конечную, заранее известную величину (год рождения, тарифный разряд, число людей), или непрерывным, то есть его значения отличаются на сколь угодно малую величину (время, вес, объем, стоимость).

Частотой  $v_i$  в случае дискретного признака  $X$  называют число одинаковых вариантов  $x_i$ , содержащихся в выборке. В ранжированном вариационном ряду одинаковые варианты очевидно расположены подряд:

$$\overbrace{\underbrace{x_1, x_1, \dots, x_1}_{v_1}, \dots, \underbrace{x_i, x_i, \dots, x_i}_{v_i}, \dots, \underbrace{x_k, x_k, \dots, x_k}_{v_k}}^n.$$

Вариационный ряд для дискретного признака  $X$  принято наглядно и компактно представлять в виде таблицы, в первой строке которой указаны  $k$  различных значений  $x_i$  изучаемого признака, а во второй строке – соответствующие этим значениям частоты  $v_i$ , где  $i = 1, 2, \dots, k$ . Таковую таблицу называют статистическим (выборочным) распределением.

Переход от исходного вариационного ряда дискретного признака  $X$  к соответствующему статистическому распределению поясним на простом примере:

– вариационный ряд, полученный в результате статистического наблюдения (единицы измерения опускаем) – 7, 17, 14, 17, 10, 7, 7, 14, 7, 14;

– ранжированный вариационный ряд  $-x_j$ :  $\underbrace{7, 7, 7, 7}_{v_1=4}, \underbrace{10}_{v_2=1}, \underbrace{14, 14, 14}_{v_3=3}, \underbrace{17, 17}_{v_4=2}$ , где

$j = 1, 2, \dots, n$ ,  $n = 10$ ;

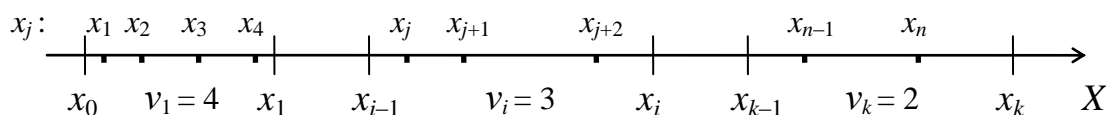
– соответствующее статистическое распределение ( $i = 1, 2, \dots, k$ ,  $k = 4$ ):

$x_i$	7	10	14	17
$v_i$	4	1	3	2

Статистическое распределение для непрерывного признака  $X$  принято представлять интервальным рядом – таблицей, в первой строке которой указаны  $k$  интервалов значений изучаемого признака  $X$  вида  $(x_{i-1} - x_i)$ , а во второй строке – соответствующие этим интервалам частоты  $v_i$ , где  $i = 1, 2, \dots, k$ .

Обозначение  $(x_{i-1} - x_i)$  – указывает не разности, а все значения признака  $X$  от  $x_{i-1}$  до  $x_i$ , кроме правой границы интервала  $x_i$ .

Для непрерывного признака  $X$  частота  $v_i$  – число различных  $x_j$ , попавших в соответствующий интервал:



Переход от исходного вариационного ряда непрерывного признака  $X$  к соответствующему статистическому распределению поясним на простом примере:

- вариационный ряд, полученный в результате статистического наблюдения (единицы измерения опускаем) – 3,14; 1,41; 2,87; 3,62; 2,71; 3,95;
- ранжированный вариационный ряд –  $x_j$ : 1,41; 2,71; 2,87; 3,14; 3,62; 3,95; где  $j = 1, 2, \dots, n$ ,  $n = 6$ ;
- соответствующее статистическое распределение ( $i = 1, 2, \dots, k$ ,  $k = 3$ ):

$x_i$	1–2	2–3	3–4
$v_i$	1	2	3

Если число различных значений дискретного признака очень велико, то для удобства дальнейших вычислений и наглядности статистическое распределение такого дискретного признака также может быть представлено в виде интервального ряда.

Далее показаны формы представления статистических распределений с соответствующими краткими названиями:

Дискретный ряд частот						Интервальный ряд частот				
$x_i$	$x_1$	$x_2$	$\dots$	$x_k$		$x_{i-1}-x_i$	$x_0-x_1$	$x_1-x_2$	$\dots$	$x_{k-1}-x_k$
$v_i$	$v_1$	$v_2$		$v_k$ ,		$v_i$	$v_1$	$v_2$	$\dots$	$v_k$ ,

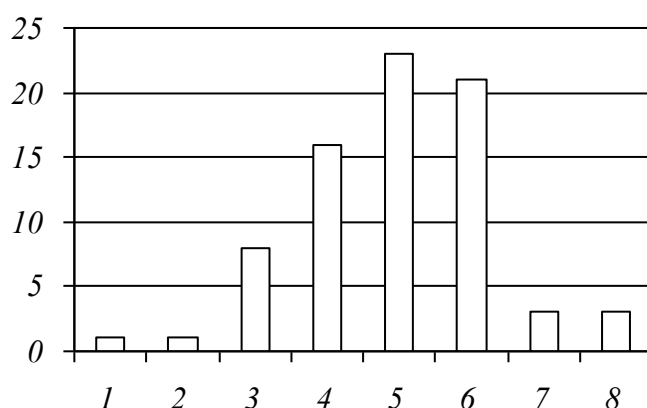
Рассмотрим пример изучения плодовитости серебристо-черных лисиц, которое дало следующие результаты (число щенков на самку): 5 5 6 5 5 6 4 4 4 5 6 4 6 6 4 6 4 5 5 8 5 3 6 5 5 5 5 5 6 3 6 4 6 4 6 2 5 6 5 3 7 6 3 4 6 8 6 3 5 5 6 5 4 3 8 4 7 5 4 3 1 6 5 3 4 5 6 7 4 4 6 5 6 4 6 5.

Для дискретного признака (такова плодовитость) построение вариационного ряда обычно не представляет сложности, достаточно подсчитать встречаемость конкретных значений.

Частота, $v$	1	1	8	16	23	21	3	3
Плодовитость, $x$	1	2	3	4	5	6	7	8

Гистограмма, построенная по данным о плодовитости лисиц (рис. 3), сразу же обнаруживает характерное поведение случайной величины – высокие частоты встречаемости значений в центре распределения и низкие по периферии.

Если же изучаемый признак **непрерывен** (таковы размерно-весовые характеристики), то для построения вариационного ряда сначала весь диапазон изменчивости признака разбивается



на серию **равных интервалов** (классов вариантов), затем подсчитывают, сколько вариантов попало в каждый интервал.

Рис. 3. Распределение плодовитости лисиц.

Число **классов** для больших выборок ( $n > 100$ ) должно быть не менее 7 и не более 12, их оптимальное число можно определить по **эмпирически** полученной формуле:

$$N = 1 + 3.32 \cdot \lg(n), \quad (2)$$

$$\text{или } N \approx \sqrt{n} \quad (2a)$$

где  $N$  – количество классов,  $n$  – объем выборки (число вариантов в выборке).

Составим для примера вариационный ряд для непрерывного признака – по данным о весе 63 взрослых землероек (г):

1) Все операции могут быть выполнены вручную. Вначале следует определить *объем выборки*  $n = 63$ .

9.2	11.6	8.1	9.1	10.1	9.6	9.3	9.7	9.9	9.9	9.6
7.6	10.0	9.7	8.4	8.6	9.0	8.8	8.6	9.3	11.9	9.3
9.2	10.2	11.2	8.1	10.3	9.2	9.8	9.9	9.3	9.1	9.4
9.6	7.3	8.3	8.8	9.2	8.0	8.6	8.8	9.0	9.5	9.1
8.5	8.8	9.7	11.5	10.5	9.8	10.0	9.4	8.7	10.0	7.9
8.6	8.7	9.1	8.2	9.2	9.4	8.8	9.8			

2) Найти *число классов* вариационного ряда по формуле:

$$N = 1 + 3.32 \cdot \lg(63) = 6.973811 \approx 7.$$

Классы	Центр классовой интервала	Частоты, $\nu$
7–7.7	7.35	2
7.8–8.4	8.05	7
8.5–9.1	8.75	18
9.2–9.8	9.45	22
9.9–10.5	10.15	10

3) Найти длину интервала  $\Delta x$  (допустимо округление):

$$\Delta x = (x_{\max} - x_{\min}) / N = (11.9 - 7.3) / 7 \approx 0.7.$$

10.6–11.2	10.85	1
11.3–11.9	11.55	3
Сумма		63

4) Установить границы классов; в качестве первой границы имеет смысл взять округленное минимальное значение:  $x_{\min} = 7$ .

5) Вычислить центральное значение признака в каждом классе; исходным берется значение центра первого интервала; для первого класса 7–7.7, для второго – 7.8–8.4...

Теперь данные можно представить графически, в виде полигона частот (ломаной кривой) или гистограммы (столбиками) (рис. 4).

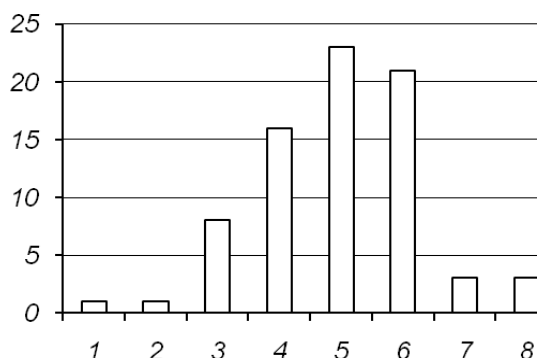


Рис. 4. Распределение бурозубок по весу тела

## 2.3 Вычисление основных статистических показателей

Первым этапом проведения статистической обработки является определение наличие грубых ошибок в выборке и их исключение.

**2.3.1 Грубые ошибки** возникают вследствие нарушения основных условий измерений или в результате небрежности экспериментатора. Они резко отличаются по величине от остальных значений, на чем и основаны некоторые критерии их исключения из рассмотрения. При обнаружении грубой ошибки результат следует отбросить, а само измерение, если это возможно, повторить.

Для выявления резко отклоняющихся величин (промахов), используется ряд критериев, основанных на нормальном распределении признаков. Для малой выборки, когда количество измерений меньше 30, принято использовать Q критерий:

$$Q = \frac{x_{i+1} - x_i}{x_n - x_1} \quad (3),$$

где:  $x_i$  – текущее значение выборки,  $x_1$  и  $x_n$  крайние значения ряда построенного по возрастанию:  $x_1 < x_2 < x_3 \dots x_i < x_n$ .

Наиболее вероятными значениями грубых ошибок являются эти крайние члены ряда, следовательно именно их необходимо проверить на значимость разницы  $x_2 - x_1$  и  $x_n - x_{n-1}$   $\phi(3)$ . Полученное значение Q сравниваем с табличным  $Q_{\text{крит}}$  (таблица 5). Если  $Q > Q_{\text{крит}}$ , то результат является промахом с уровнем вероятности  $P \geq 95\%$ .

Таблица 5. Численные значения Q - критерия от количества измерений – n.

n	3	4	5	6	7	8	9	10
Q <sub>крит</sub>	0,94	0,76	0,64	0,56	0,51	0,47	0,44	0,41

После нахождения и исключения грубых ошибок в исследуемой выборке следует вычислить среднюю величину. Средняя - это один из распространенных приемов обобщений.

### 2.3.2. Первичная статистическая обработка полученных данных.

В биологических исследованиях зачастую встречается ситуация, когда требуется первичная статистическая обработка большого числа выборок, не обязательно с большой точностью. Это может понадобиться для предварительного рассмотрения и оценки материала при проведении поисковых опытов, в частности для оперативного выявления общих тенденций его изменчивости, с тем, чтобы в дальнейшем перейти к специальным методам статистического анализа. Для этих случаев предложен простой экспресс-метод с использованием полученного для данной выборки размаха значений (Lim). В случае нормального распределения средняя арифметическая находится точно по центру (совпадает со значением медианы), т. е. левая и правая границы распределения находятся на одинаковом расстоянии от средней. Исходя из этих соображений, среднюю арифметическую можно рассчитать по формуле медианы:

$$M \approx Me = \frac{x_{\min} + x_{\max}}{2} \quad (4)$$

Для бурозубок эта средняя составит  $M = (7.3 + 11.9) / 2 = 9.6$  г, что вполне соответствует первой точной оценке.

**Медиана** - это варианта, расположенная в середине вариационного ряда. Величина, которая делит численность упорядоченного вариационного ряда на две равные части: одна часть имеет значения варьирующего признака меньше, чем средний вариант, а другая – больше. Если ряд распределения дискретный и имеет нечетное число членов, то медианой будет варианта, находящаяся в середине упорядоченного ряда (упорядоченный ряд - это расположение единиц совокупности в возрастающем или убывающем порядке).

Рассмотрим способы *определения медианы* при различных значениях N. Для нахождения медианы измерения записывают в ряд по возрастанию значений. Если число измерений N нечетное, то медиана численно равна значению этого ряда, стоящему точно в середине, или на  $(N+1)/2$  месте. Например, медиана пяти измерений: 10, 17, 21, 24, 25 – равна 21 – значению, стоящему на третьем месте  $(N+1)/2=(5+1)/2=3$ .

Понятие медианы легко уяснить из следующего примера. На рисунке показаны гистограммы распределения оценок по четырем группам (1, 2, 3, 4). Ме-

диана равна 2 (общее количество групп 4). Средняя успеваемость в четырех группах одинакова. Но независимо от одинаковой средней успеваемости ( $X_1=3,69$ ;  $X=3,69$ ;  $X_3=3,69$ ;  $X_4=3,69$ ) действительная картина успеваемости в четырех группах совершенно различна. В первой группе немного двоек (2) и много троек (16), во второй – наибольшее число (18) – четверки, а пятерок мало (5). В третьей группе больше всего пятерок (10), общее число пятерок и четверок 18. Число пятерок и четверок в четвертой группе 22, но много слабых оценок (10).

Эти гистограммы дают наглядное и полное представление распределения оценок в каждой группе. Гистограммы детальнее и полнее отражают различия в успеваемости в четырех группах, чем показатель средней успеваемости.

Кроме средней оценки ( $\bar{X}$ ) в качестве параметра, характеризующего распределение оценок, используется оценка, встречаемая чаще всего. В теории вероятности и в математической статистике чаще всего встречающееся значение случайной величины (у нас оценка), самое вероятное значение, называется среднелегальным значением, модальным значением, или коротко – модой. В нашем случае модальная успеваемость ( $X_{\text{м}}$ ) для первой группы – три ( $X_{\text{м}1}=3$ ), для второй – четыре ( $X_{\text{м}2}=4$ ), для третьей – пять ( $X_{\text{м}3}=5$ ), четвертая группа имеет две моды ( $X_{\text{м}4}=1$  и  $X_{\text{м}4}=5$ ). Распределение в четвертой группе двумодальное.

**Мода** - это величина признака (варианта), наиболее часто повторяющаяся в изучаемой совокупности. Эта величина легко определяется без вычислений. Она дает возможность произвести качественный анализ состояния и изменений параметров в различных группах. Она является простым способом оценки и сравнения данных.

Анализируя среднюю успеваемость ( $\bar{X}$ ) в группах  $X_1=3,69$ ;  $X_2=3,69$ ;  $X_3=3,69$ ;  $X_4=3,69$  (рис. 5), можно сделать вывод о том, что результаты в четырех группах одинаковы. Анализируя и модальную успеваемость  $X_{\text{м}1}=3$ ;  $X_{\text{м}2}=4$ ;  $X_{\text{м}3}=5$ ;  $X_{\text{м}4}=1$  и  $X_{\text{м}4}=5$ , получаем новые данные для количественного и качественного анализа. Однако мода, в качестве параметра, характеризующего данное распределение, обладает и недостатками. Неточность, с которой определяется средняя вероятная успеваемость ( $X_{\text{в}}$ , больше, чем при определении средней успеваемости ( $\bar{X}$ ). Оценка успеваемости в различных группах затрудняется, когда распределение в одной из них – двумодальное (как в четвертой группе).

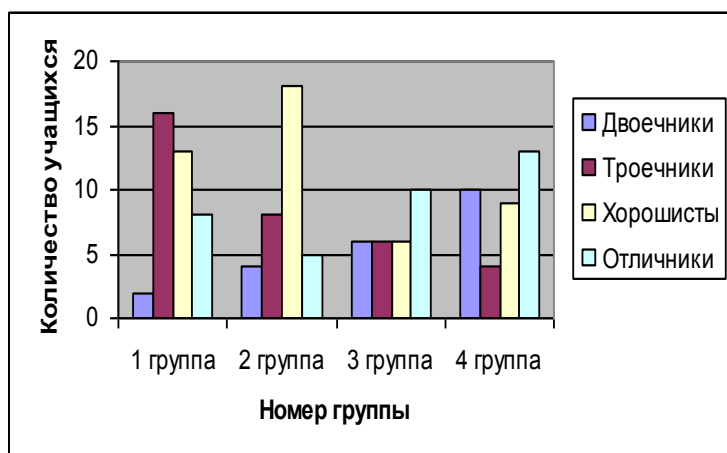


Рис. 5. Распределения количества учащихся имеющих различную успеваемость в группах.

Для дискретных рядов распределения модой будет значение варианта с наибольшей частотой. Мода представляет собой максимально часто



встречающееся значение переменной (иными словами, наиболее «модное» значение переменной), например, популярная передача на телевидении, модный цвет платья или марка автомобиля и т. д. Сложность в том, что редкая совокупность имеет единственную моду. (Например: 2, 6, 6, 8, 9, 9, 9, 10 – мода = 9).

Если распределение имеет несколько мод, то говорят, что оно мультимодально или многомодально (имеет два или более «пика»).

Квартили представляют собой значения, которые делят две половины выборки (разбитые медианой) еще раз пополам (от слова кварта — четверть). Различают верхнюю квартиль, которая больше медианы и делит пополам верхнюю часть выборки (значения переменной больше медианы), и нижнюю квартиль, которая меньше медианы и делит пополам нижнюю часть выборки. Нижнюю квартиль часто обозначают символом 25%, это означает, что 25% значений переменной меньше нижней квартили. Верхнюю квартиль часто обозначают символом 75%, это означает, что 75% значений переменной меньше верхней квартили.

Таким образом, три точки — нижняя квартиль, медиана и верхняя квартиль — делят выборку на 4 равные части.  $\frac{1}{4}$  наблюдений лежит между минимальным значением и нижней квартилью,  $\frac{1}{4}$  — между нижней квартилью и медианой,  $\frac{1}{4}$  — между медианой и верхней квартилью,  $\frac{1}{4}$  — между верхней квартилью и максимальным значением выборки.

Показатели моды, медианы и квартили рекомендуется использовать при проведении поисковых опытов (при малом количестве измерений  $n < 30$ ), для оценки распределения выборки. Если мода попадает в первую или четвертую квартиль, то данное распределение нельзя считать нормальным, а следовательно необходимо использовать особые методы трансформации выборки (см. тему 2.5) или применять непараметрические методы анализа полученных данных (см. тему 4).

Средняя величина — это абстрактная, обобщающая характеристика признака изучаемой совокупности, но она не показывает строения совокупности, которое весьма существенно для ее познания. Средняя величина не дает представления о том, как отдельные значения изучаемого признака группируются вокруг средней, сосредоточены ли они вблизи или значительно отклоняются от нее. В некоторых случаях отдельные значения признака близко примыкают к средней арифметической и мало от нее отличаются. В таких случаях средняя хорошо представляет всю совокупность.

В других, наоборот, отдельные значения совокупности далеко отстают от средней, и средняя плохо представляет всю совокупность. Колеблемость отдельных значений характеризуют показатели вариации. Термин "вариация" произошел от латинского *variatio* — «изменение, колеблемость, различие». Однако не всякие различия принято называть вариацией. Под вариацией в статистике понимают такие количественные изменения величины исследуемого признака в пределах однородной совокупности, которые обусловлены перекрещивающимся влиянием действия различных факторов. Различают вариацию признака: случайную и систематическую.

Анализ систематической вариации позволяет оценить степень зависимости изменений в изучаемом признаке от определяющих ее факторов. Например, изучая силу и характер вариации в выделяемой совокупности, можно оценить, насколько однородной является данная совокупность в количественном, а иногда и качественном отношении. Степень близости данных отдельных единиц  $x_i$  к средней измеряется рядом абсолютных, средних и относительных показателей.

### 2.3.3 Средняя арифметическая.

Средняя величина - это обобщающие показатели, в которых находят выражение действия общих условий, закономерностей изучаемого явления. При помощи средней происходит как бы сглаживание различий в величине признака, которые возникают по тем или иным причинам у отдельных единиц наблюдения. Одной из важнейших обобщающих характеристик вариационного ряда является средняя величина признака (обычно обозначается буквой  $M$ ). Существует несколько видов средних (средняя арифметическая – простая и взвешенная, средняя гармоническая, средняя квадратичная), но в практике биологических исследований наибольшее значение имеет средняя арифметическая – величина, вокруг которой «концентрируются» варианты.

Общая формула для определения величины средней арифметической – это отношение суммы значений всех вариантов ( $x_i$ ) выборки к их числу (объему выборки,  $n$ ):

$$M = \frac{\sum x_i}{n} \quad (4)$$

В нашем примере с определением массы бурозубок средняя величина равна  $M = 9.298412698$  г. При расчетах статистических параметров на ЭВМ следует помнить, что большое количество цифр обычно не имеет никакого биологического смысла. Записывая такие статистические параметры, как средняя и стандартное отклонение, следует оставлять в лучшем случае на одну значащую цифру больше, чем имели значения вариантов, а оценки ошибок – на две значащих цифры (см. тему 3.2).

Средняя арифметическая характеризует действие только систематических факторов, поскольку сумма случайных отклонений влево и вправо от средней в силу симметричности нормального распределения обращается в нуль. Поэтому модель варианты меняется  $x_i = M \pm x_{случ.}$

**2.3.4 Абсолютные и средние показатели вариации и способы их расчета.** Для характеристики совокупностей полученных величин важно знать, какая вариация изучаемого признака скрывается за средним. Для характеристики колеблемости признака используется ряд показателей. Наиболее простой из них - размах вариации, который принято использовать при планировании исследования (уровень надежности составляет  $P > 0,68$  см. главу 3.2).

**Размах вариации** - это разность между наибольшим  $x_{mak}$  и наименьшим  $x_{min}$  значениями вариантов

$$R = x_{\max} - x_{\min} \quad (5)$$

Чтобы дать обобщающую характеристику распределению отклонений, исчисляют среднее линейное отклонение, которое учитывает различие всех единиц изучаемой совокупности. При анализе основных опытов (уровень надежности составляет  $P > 0,95$ ) принято использовать показатели дисперсии, среднеквадратического отклонения и коэффициент вариации.

**Дисперсия** – характеризует вариацию признака, которая зависит от всех условий в данной совокупности, определяемый как средняя из отклонений, введенных в квадрат  $(X - \bar{X})^2$ . На практике меру вариации объективно отражает показатель дисперсии ( $\sigma^2$  – средний квадрат отклонений),  $\sigma = \sqrt{\sigma^2}$  – называется среднеквадратичным отклонением и является мерой вариации (т.е. разброса, рассеяния) признака для генеральной совокупности ( $n \geq 30$ ) и отдельной выборки ( $n < 30$ ):

$$\sigma = \sqrt{\frac{\sum (M - X_i)^2}{n}} \quad (6a) \quad S = \sqrt{\frac{\sum (M - X_i)^2}{n-1}} \quad (6б),$$

где  $x$  – значение признака у каждого объекта в группе,  $M$  – средняя арифметическая признака,  $n$  – число вариантов выборки.

Для примера с массой тела бурозубок стандартное отклонение будет равно:  $S = 0.897216496$ , а после необходимого округления  $S = 0.897$  г.

Средняя величина отражает тенденцию развития, т.е. действие главных причин (факторов), а  $\sigma$  измеряет силу воздействия прочих факторов. «Дидактический смысл среднего квадратичного отклонения ( $\sigma$ ) в приведенном балле успеваемости ( $X$ ) состоит в том, что им характеризуется степень индивидуализации обучения и однородности успеваемости в классе. Этим снимается иллюзорная информативность, а на самом деле дезинформативность средних субъективных оценок, из-за которых и появилась болезнь, называемая процентоманией. Примеры средней температуры в больнице, глубины реки. Наиболее информативным данный показатель становится при использовании двенадцатибалльной шкалы, построенной на основе четырех уровней усвоения».

**Коэффициент вариации.** Колеблемость отдельных значений признака внутри изучаемой совокупности характеризуют показателями вариации, одним из которых (и наиболее распространенным) является коэффициент вариации. Определяется как отношение среднеквадратичного отклонения к средней арифметической в процентах:

$$C_v = \frac{\sigma}{M} 100\% \quad (7)$$

Учитывая, что среднеквадратическое отклонение дает обобщающую характеристику колеблемости всех вариантов совокупности, коэффициент вариации является наиболее распространенным показателем колеблемости, используемым для оценки типичности средних величин. При этом исходят из того,

что если  $C_v$  больше 15 %, то это говорит о большой колеблемости признака в изучаемой совокупности.

Коэффициент вариации оценивает типичность средних величин. В нашем примере с весом тела бурозубок:

$$C_v = \frac{S}{M} \cdot 100\% = \frac{0.89}{9.3} \cdot 100\% = 9.6 \, \%.$$

Индивидуальная изменчивость (варьирование) признаков – одна из наиболее емких характеристик биологической популяции, любого биологического процесса или явления. Коэффициент вариации может считаться вполне адекватным и объективным показателем, хорошо отражающим фактическое разнообразие совокупности независимо от абсолютной величины признака. Данный коэффициент был создан для унификации показателей изменчивости разных или разноразмерных признаков путем приведения их к одному масштабу. Практика показывает, что для многих биологических признаков наблюдается увеличение изменчивости (стандартного отклонения) с ростом их величины (средней арифметической). При этом коэффициент вариации остается примерно на одном и том же уровне.

### 2.3.5 Ошибка репрезентативности выборочных параметров.

По части никогда не удастся полностью охарактеризовать целое, всегда остается вероятность того, что выборочная оценка недостаточна, близка к значению параметра генеральной совокупности, имеет некоторую ошибку. Отличия значений выборочных параметров от генеральных называются ошибкой репрезентативности данного параметра, или просто (статистической) ошибкой. При увеличении объема выборки ошибки репрезентативности стремятся к нулю (следствие закона больших чисел). Численно выраженные статистические ошибки служат мерой тех пределов, в которых выборочные параметры могут отклоняться от значений генеральных параметров. Например, если для нескольких выборок, полученных из одной и той же генеральной совокупности, посчитать средние, а затем оценить изменчивость этих средних, то стандартное отклонение средних ( $S_M$ ) и будет численной мерой ошибки репрезентативности выборочной средней. Она обозначается буквой  $m$ .

Величина ошибки тем больше, чем больше варьирование признака ( $S$ ) и чем меньше выборка ( $n$ ). Ошибку репрезентативности имеют все статистические параметры, рассчитанные по выборке. Для практики статистического оценивания разработаны специальные формулы. Для нормального распределения они имеют следующий вид. Ошибка средней:  $m_M = \frac{S}{\sqrt{n}}$ , (8)

$$\text{ошибка стандартного отклонения: } m_S = \frac{S}{\sqrt{2 \cdot n}}, \quad (9)$$

$$\text{ошибка коэффициента вариации: } m_{CV} = \frac{CV}{\sqrt{2 \cdot n}} \quad (10).$$

Вычисленные значения ошибок подставляют к соответствующим параметрам со знаками плюс-минус (параметр  $\pm$  ошибка) и в такой форме представляют в научных отчетах и публикациях. В примере с бурозубками для разных параметров имеем:

$$m_M = \frac{0.89}{\sqrt{63}} = 0.113039, M = 9.3 \pm 0.11 \text{ г},$$

$$m_S = \frac{0.89}{\sqrt{2 \cdot 63}} = 0.07993, S = 0.89 \pm 0.079 \text{ г},$$

$$m_{CV} = \frac{9.6}{\sqrt{2 \cdot 63}} = 0.859613, CV = 9.6 \pm 0.9\%.$$

Не следует путать статистическую ошибку с методическими ошибками и ошибками точности (точности измерений, анализов, подсчетов и т. д.), хотя методические погрешности и увеличивают ошибку репрезентативности, но другим путем – методические огрехи увеличивают изменчивость признака, стандартное отклонение.

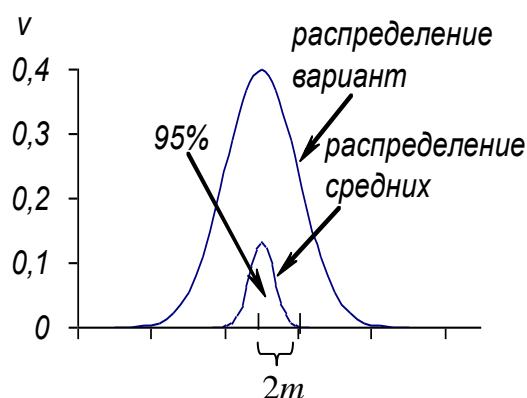
При всей неизбежности статистической ошибки она может быть сведена к минимуму отбором достаточного числа особей (вариант). С ростом объема выборки оценки параметров стабилизируются, а их ошибки репрезентативности уменьшаются.

### 2.3.6 Доверительный интервал.

Параметры генеральной совокупности практически всегда остаются неизвестными, о них судят по выборочным оценкам, используя для этого значения ошибок репрезентативности. Теоретические исследования поведения выборочных средних (как случайных величин) показали, что они подчиняются нормальному закону, большинство из них (95%) находится поблизости от генеральной средней – в диапазоне  $M_{ген.} \pm 1.96t$  (приблизительно  $\pm 2t$ ). Это обстоятельство позволяет делать обратное заключение – генеральная средняя находится в диапазоне  $M_{выбор.} \pm 1.96t$ , т. е. предсказывать ширину интервала, в котором заключен генеральный параметр, давать *интервальную оценку* генеральному параметру.

В соответствии с законом нормального распределения можно ожидать, что генеральный параметр (истинное значение) окажется в интервале от  $M - tm$  до  $M + tm$ , где  $t$  – ошибка средней арифметической,  $t$  – квантиль распределения Стьюдента (табл. 2 ПЗ) при данном числе степеней свободы ( $k$ ) и уровне значимости (обычно  $p = 0.05$ ).

Сказанное можно перефразировать так: с вероятностью  $P$  можно ожидать, что генеральная средняя находится в доверительном интервале  $M \pm tm$ , построенном вокруг выборочной средней арифметической  $M$ . Доверительный интервал



– интервал значений изучаемого признака, в котором с той или иной вероятностью  $P$  находится значение генерального параметра.

Возвращаясь к примеру о весе землероек-бурозубок, мы теперь можем записать доверительные интервалы при разных уровнях вероятности (граничные значения  $t$  взяты для случая  $n = \infty$ ):

для  $P = 0.95$   $M \pm tm = 9.3 \pm 1.96 \cdot 0.11 = 9.3 \pm 0.21$  г;

для  $P = 0.99$   $M \pm tm = 9.3 \pm 2.58 \cdot 0.11 = 9.3 \pm 0.28$  г.

Здесь искомая генеральная средняя величина веса землероек с вероятностью  $P = 95\%$  находится в пределах 9.11-9.53 г, а при  $P = 99\%$  – 9.04-9.6 г.

Если объем выборки, для которой были получены параметры и ошибка репрезентативности  $m$ , был невелик ( $n < 50$ ), то необходимо вводить поправки на объем выборки, расширяя область возможного пребывания генерального параметра. Это понятно, поскольку при дефиците информации любые заключения не могут быть очень точными. Так, для выборки объемом  $n = 20$  экз. ошибка

средней составит  $m_M = \frac{0.89}{\sqrt{20}} = 0.19901$  г, а доверительный интервал:  $M \pm tm = 9.3 \pm 2.09 \cdot 0.2 = 9.3 \pm 0.41$  г – от 8.9 до 9.7 г (при уровне значимости  $p = 0.05$  и числе степеней свободы  $k = n - 1 = 20 - 1 = 19$  табличная величина статистики Стьюдента равна  $t = 2.09$ ).

Аналогичным образом можно построить доверительный интервал для стандартного отклонения ( $S \pm tm_S$ ), коэффициента вариации ( $CV \pm tm_{CV}$ ), а также других статистических параметров (коэффициентов асимметрии, эксцесса, регрессии, корреляции).

## 2.4 Основные типы распределений признаков

Как уже отмечалось, биометрия изучает случайные события, поведение случайных величин. Приступая к наблюдению или даже начиная биологический эксперимент, невозможно точно сказать, каков будет результат – уровень численности животных в данном районе, вес еще не отловленных особей, количество сахара в крови через час после введения препарата и т. п. В этом смысле биологические явления случайны, точно не предсказуемы. Однако любому биологу ясно, что случайность эта не абсолютна. Несмотря на сложность точного прогноза, приблизительный результат можно предугадать, в частности, предсказав, что интересующая нас величина будет находиться в пределах некоторого интервала между конкретными минимальными и максимальными значениями. Ясно, например, что рост человека вряд ли превысит два или будет ниже полутора метров. Вариационная статистика может дать и более точный прогноз, ориентируясь на известные законы поведения случайных величин, относящихся к разным типам распределений. При этом под распределением признаков (случайных величин, объектов) понимается соотношение между их значениями и частотой встречаемости.

Среди многих известных типов распределений мы рассмотрим лишь пять (нормальное, биномиальное, альтернативное, Пуассона, полиномиальное, равномерное). Зная тип распределения, можно воспользоваться разработанными

специально для него приемами математической обработки и получить наиболее полную информацию о явлении, точнее оценить различия между параметрами разных выборок.

**Нормальное распределение.** Наиболее характерный тип распределения непрерывных случайных величин, (соответствует количественной шкале измерения см. табл. 3) из него можно вывести (к нему сводятся) все остальные. Распределение симметрично, причем крайние значения (наибольшие и наименьшие) появляются редко, но чем ближе значения признака к центру (к средней арифметической), тем оно чаще встречается (рис. 6).

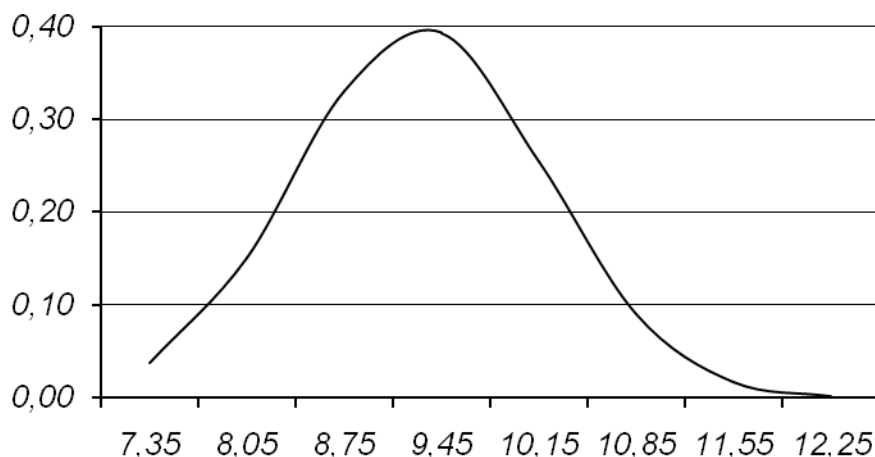


Рис. 6. Нормальное распределение с параметрами  $n = 63$ ,  $M = 9.3$ ,  $S = 0.79$ . По оси абсцисс – вес тела землероек-бурозубок, по оси ординат – табличные значения для нормального распределения.

Рассчитать ординаты нормальной кривой для конкретного значения  $x_i$  можно по формуле:

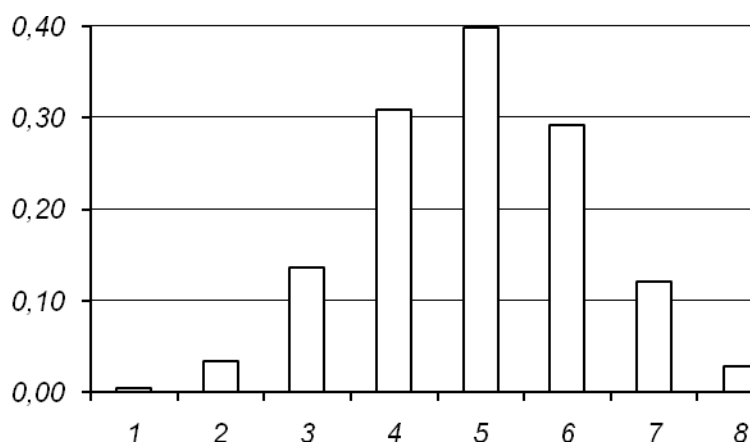
$$y_i = (1/\sqrt{2\pi}) \cdot e^{-(x_i - M)^2 / 2 \cdot S^2} \quad (11).$$

Среднее квадратичное отклонение примерно 4 раза укладывается в размахе изменчивости признака и по величине значительно уступает средней. Геометрически стандартное отклонение равно расстоянию от центра кривой распределения до точки перегиба кривой. Примеры расчета параметров ( $M$ ,  $S$ ) нормального распределения приведены выше.

**Биномиальное распределение.** Во многом близко к нормальному. Отличие состоит лишь в том, что оно характеризует поведение дискретных признаков, выраженных целыми числами (соответствует количественной шкале измерения см. табл. 3). Как правило, для описания биологических признаков подходит симметричное биномиальное распределение, у которого дисперсия много меньше средней. Распределение организуется в процессе обора проб (объемом больше одного,  $h > 1$ ). Число классов больше двух,  $k > 2$ .

Примерами описания признаков с помощью биномиального распределения могут служить число поврежденных участков на листьях, число волосков на единице площади шкурки, количество лучей в плавниках рыб, число хвостовых

щитков у рептилий, плодовитость (размер выводка) самок и т.д. В основе биномиального распределения лежит альтернативное проявление качественного признака: он может присутствовать у единичного объекта или отсутствовать, проявиться или нет. Отдельный корнеплод может быть больным или здоровым (признак качественный), тогда *проба* из нескольких корнеплодов будет содержать некоторое *число* здоровых корнеплодов (признак количественный), а множество равнообъемных проб образует уже выборку чисел, для которой можно построить гистограмму распределения. Вероятность отдельного события (корнеплод больной) составляет  $W$ , а вероятность альтернативного события (корнеплод здоровый) равна  $U = 1 - W$ . При равенстве вероятностей событий  $W = U = 0.5$ , большинство проб (вариант) будет иметь около половины возможных событий (поровну больных и здоровых корнеплодов); распределение примет симметричную форму. В случае неравенства вероятностей наблюдается та или иная степень асимметрии распределения.



Рассмотрим результаты изучения плодовитости серебристо-черных лисиц (число щенков на самку рис. 7).

Рис. 7. Биномиальное распределение ( $n = 76$ ,  $M = 4.95$ ,  $S = 1.33$ ). По оси абсцисс – число щенков лисицы на одну самку, по оси ординат – частоты (относительные частоты).

Для построения вариационного ряда берем  $h=k=8$  классов, классовой интервал для этого дискретного признака составит  $\Delta x = 1$ . Все основные параметры распределения вычисляются по рассмотренным выше формулам:

$$M = \frac{\sum x}{n} = 4.96, \quad S = \sqrt{\frac{\sum (M - X_i)^2}{n-1}} = 1.33, \quad m = \frac{S}{\sqrt{n}} = \frac{1.33}{\sqrt{63}} = 0.1676 \text{ экз./самку},$$

Результаты оказываются идентичными с точностью до ошибки округления. Доверительный интервал для параметров биномиального распределения строится так же, как и для нормального распределения:  $M \pm tm$ .

**Полиномиальное распределение.** Наблюдается для качественных признаков, имеющих не два альтернативных свойства, но несколько возможных проявлений качества (соответствует качественной ранговой шкале измерения см. табл. 3). Примеры полиморфизма популяций – из этой области. В их числе варианты окраски покровов и волос, типы рисунков в определенных областях тела, способы жилкования листьев растений или крыльев насекомых, варианты расположения и формы щитков рептилий и другие проявления множественности



фенотипов особей. Формализуя описание, укажем, что в одной пробе содержится одна варианта ( $h = 1$ ), но типов вариант (морф, фенотипов) больше, чем два ( $k > 2$ ).

Примером полиномиального (иначе – мультиномиального) распределения может служить встречаемость 4 фенотипов головы живородящей ящерицы – 4 вариантов контакта лобно-носового, предлобных и лобного щитков (рис. 8). Лучше всего выборка может быть представлена вариационным рядом – частотами ( $W_i$ ) встречаемости в популяции особей с данным ( $i$ -м) проявлением качественного признака и общим числом морф ( $k$ ). Для более емкого представления ряда и учета характера распределения частот между разными морфами используется величина «среднее число фенотипов»:  $M = \sum(W_i)^2$ , среднее квадратическое отклонение рассчитывается так  $S = \sqrt{M(k - M)}$ .

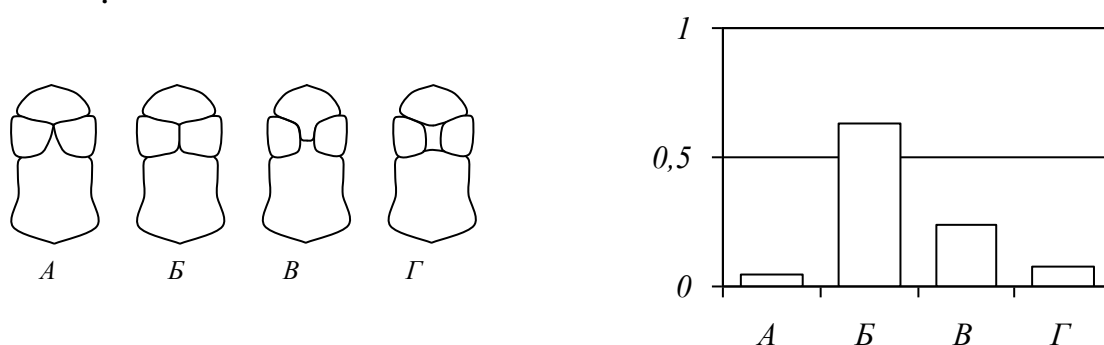


Рис. 8. Полиномиальное распределение (4 фена головы ящерицы). По оси ординат – частоты фенотипов среди 64 сеголетков живородящей ящерицы.

Под Петрозаводском отловлены 64 сеголетка живородящей ящерицы, с фенотипом головы А (рис. 8)  $n_1=4$ ; Б  $n_2=39$ ; В  $n_3=16$  и Г  $n_4=5$ . Вероятность отлова составила по группам  $W_1=0,0625$ ;  $W_2=0,6094$ ;  $W_3=0,25$  и  $W_4=0,0781$ , средняя величина  $M=0,4439$  и её ошибка  $m=0,1570$ , а среднее квадратическое отклонение  $S=1,2563$ .

Среднее число фенотипов ( $M$ ) равно числу фенотипов ( $k$ ) только тогда, когда частоты всех фенотипов одинаковы ( $W_1 = W_2 = \dots = W_j \dots = W_k$ ), и меньше во всех других случаях.

**Распределение Пуассона.** Это вариант описания стохастического поведения дискретных признаков для случаев, когда вероятность элементарных альтернативных событий неодинакова (соответствует ранговой шкале измерения см. табл. 3), одно из них наблюдается заметно чаще другого ( $W \ll U$ ) (классический пример – попадание гитлеровских авиационных бомб в разные кварталы Лондона). Закон Пуассона описывает редкие события, происходящих 1, 2, 3 и т. д. раз на сотни и тысячи обычных событий. Поведение биологических объектов, соответствующее закону Пуассона, наблюдается в том случае, когда по пробам случайно распределены редкие объекты. Примеры таких явлений – частота нарушений хромосомного аппарата на каждую тысячу митозов, встречаемость семян сорняка в большой серии навесок семян культурного растения, число по-

вторных попаданий животных в ловушки, встречаемость животных на отрезках длинных маршрутов (или на пробных площадках обширной территории), отловы животных в отдельные промежутки времени при длительных наблюдениях.

Случайная величина, распределенная по закону Пуассона, определяется при подсчете числа элементарных событий *в пробе* (в группе, в навеске, на участке, на этапе). Число объектов в пробе больше 1 ( $h > 1$ ), число классов больше двух ( $k > 2$ ).

Распределение Пуассона резко асимметрично, причем дисперсия равна средней арифметической, что может служить критерием для оценки характера распределения изучаемого признака.

Пример. В течение одного года (1946) поместили кольцами и выпустили на волю 32 буревестника. В последующие пять лет часть из них отлавливали повторно: 7 экз. по одному разу, 7 – по два, 2 – по три, 1 экз. – четыре раза, 15 экз. окольцованных птиц повторно не попадались. Число классов составляет  $k = 4$ , интервал  $\Delta x = 1$ .

Асимметрия в частотах встречаемости птиц позволяет предполагать распределение Пуассона (рис. 9).

Число повторных отловов, $h$	Число отловленных животных, $v$	Число случаев повторного отлова, $h \cdot v$
0	15	0
1	7	7
2	7	14
3	2	6
4	1	4
$n$	32	31

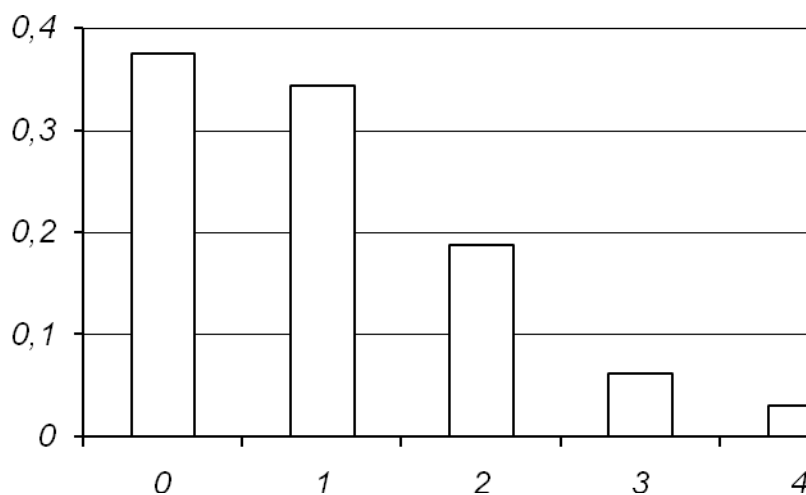


Рис. 9. Распределение Пуассона с параметрами  $n = 32$ ,  $M \approx S^2 = 0.968$ . По оси абсцисс – число повторных отловов, по оси ординат – частоты (относительные частоты).

Средняя арифметическая равна:  $M = \frac{\sum x}{n} = \frac{31}{32} = 0.968$  экз.

Критерий Фишера не выявил достоверных отличий между средней и дисперсией:  $F = 1.257 / 0.968 = 1.157 < F_{(0.05, 31, 31)} = 1.8$ , что свидетельствует о соответствии наблюдаемого распределения закону Пуассона. Доверительный интервал для параметров распределения Пуассона определить несколько сложнее, чем для других типов.

**Альтернативное распределение.** Распределение *дискретной случайной величины* (соответствует номинальной шкале измерения см. табл. 3), имеющей

Категория	Частота
самцы	0,4
самки	0,6

Важнейшей характеристикой является доля ( $W$ ) вариант определенного вида ( $A$ ), представленных общим числом  $n_A$  в пределах выборки объемом  $n$ :

Если исходы отдельных испытаний выразить числами 0 или 1 (что аналогично отбору проб с объемом  $h = 1$ ), доля вариантов совпадает со средней арифметической, рассчитанной для всех значений:

Например, результат отловов полевок из природных популяций показал, что в исследуемой группе (200 особей) было 120 самок и 80 самцов. В данном случае мы имеем дело с альтернативным признаком (самка – самец). Из 200 проб 120 содержат самок (значение 1), 80 – не содержат (значение 0), так получаем выборку  $n = 200$ :

Доля вариантов со значением 1 составляет:

что совпадает со средней арифметической для всего ряда:

Для альтернативного распределения могут применяться те же формулы расчета выборочных параметров, что и для биномиального распределения. Средняя (доля самок):

Стандартное отклонение для больших выборок ( $n > 30$ ) или при проведении

поисковых опытов можно рассчитывать при помощи упрощенных формул:

$$S = \sqrt{W \cdot U} = \sqrt{W(1-W)} \quad (13).$$

$$S = \sqrt{W \cdot U} = \sqrt{0.6 \cdot 0.4} = 0.24.$$

Ошибка средней (ошибка доли самок):

$$m = \frac{S}{\sqrt{n}} = \frac{0.24}{\sqrt{200}} = 0.017.$$

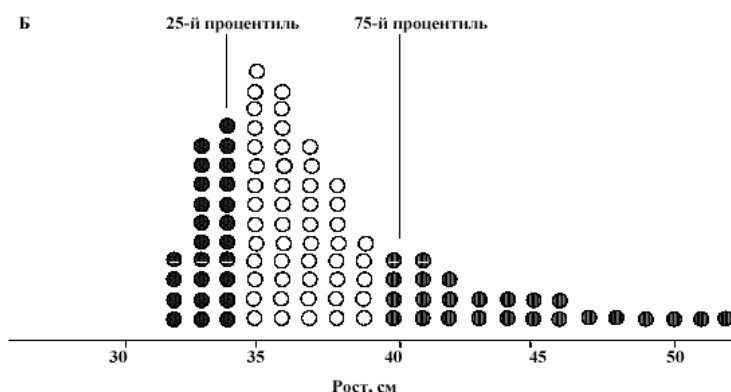
Следует отметить, что формулы описательной статистики  $\phi$  (5 – 11) можно применять при нормальном распределении исследуемого признака в выборке  $\phi$  (11). В противном случае необходимо увеличивать количество опытов или преобразовывать выборку к нормальному виду.

## 2.5 Преобразование выборки к нормальному виду

**Гипотеза нормальности.** В некоторых исследовательских проектах можно сформулировать гипотезы относительно распределения рассматриваемой переменной. Например, переменные, значения которых определяются бесконечным числом независимых факторов, распределены по нормальному закону: можно предположить, что рост индивидуума является результатом воздействия многих независимых факторов, таких как различные генетические предрасположенности, болезни, перенесенные в раннем возрасте и т.д. Как следствие, рост имеет тенденцию к нормальному распределению в населении.

С другой стороны, если наблюдаемые значения переменной являются результатом очень редких событий, то переменная будет иметь распределение Пуассона (которое иногда называется распределением редких событий). Например, несчастные случаи на производстве можно рассматривать как результат пересечения ряда неудачных событий (на житейском языке стечением маловероятных обстоятельств), поэтому их частота приближенно описывается распределением Пуассона.

Другим обычным приложением процедуры подгонки распределения является проверка гипотезы нормальности до того, как использовать какой-либо параметрический тест. Как правило, при малом количестве измерений ( $n < 30$ ) биологического объекта (см. рис. 11) показатели моды близки к величинам первой или третьей квартили (25-й или 75-й процентиля). В таком случае полученное распределение не следует считать нормальным и применять общепринятые формулы описательной статистики. В таком случае рекомендуется использовать методы подгонки распределения выборки к нормальному виду. Универсальным способом преобразования к нормальному распределению является по-



строение выборки не из текущих значений, а из их средних показателей.

Рис. 11. Мода роста исследуемых особей — 35 см. 25-й и 75-й процентиля отсекают чет-

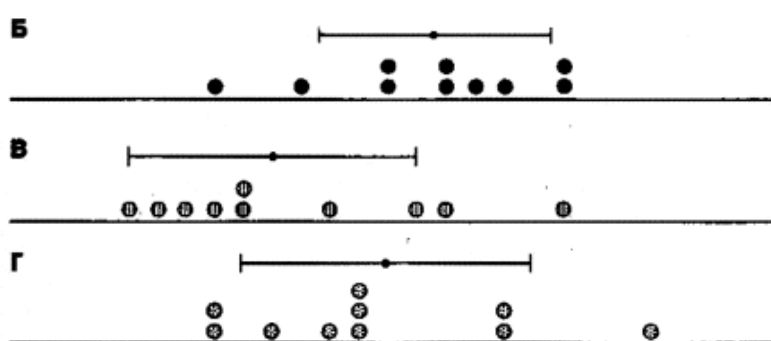
верть самых низких и четверть самых высоких особей 25-й процентиль ближе к медиане, чем 75-й — это говорит об асимметричности распределения.

Для описания асимметричного распределения следует использовать моду и процентиля.

### 2.5.1 Приведение к нормальному виду количественных признаков

Выборочное среднее и выборочное стандартное отклонение есть оценки среднего и стандартного отклонения для совокупности, вычисленные по случайной выборке. Понятно, что разные выборки дадут разные оценки. Для характеристики точности выборочных оценок используют *стандартную ошибку*. Стандартную ошибку можно подсчитать для любого показателя, но сейчас мы остановимся на *стандартной ошибке среднего*, — она позволяет оценить точность, с которой выборочное среднее характеризует значение среднего по всей совокупности.

На рис. 12 представлено распределение особей по росту. Посмотрим, что получится, если оценивать средний рост по выборке объемом, скажем, 10 особей. Из 200 особей наугад выберем 10 и пометим их черными кружками. На рис. 12Б эта выборка изображена в виде, принятом в журнальных публикациях. Точка и два отрезка по бокам от нее изображают выборочное среднее ( $\bar{X} = 41,5$  см) и выборочное стандартное отклонение ( $s = 3,8$  см). Эти значения близки, но не



равны среднему по совокупности ( $\mu = 40$  см) и стандартному отклонению ( $\sigma = 5$  см).

Рис. 12. Три случайные выборки из одной совокупности дают три разных оценки среднего и стандартного отклонения.

Извлечем еще одну случайную выборку того же объема. Результат показан на рис. 12В. Выборочное среднее (36 см) по-прежнему близко к среднему по совокупности, хотя и отличается от него; что касается выборочного стандартного отклонения (5 см), то на этот раз оно совпало со стандартным отклонением по совокупности.

На рис. 12Г представлена третья выборка. Попавшие в нее особи изображены кружками с точками. Среднее и стандартное отклонение для этой выборки составляют соответственно 40 и 5 см. Теперь пора поставить добычу случайных выборок на промышленную основу. Рассмотрим *совокупность средних* для каждой из возможных выборок по 10 особей. Общее число таких выборок превышает  $10^{16}$ . Три из них мы уже обследовали. Средние по этим выборкам представлены на рис. 13 в виде заполненных кружков. Пустые кружки — это средние еще для 22 выборок. Итак, теперь каждому выборочному среднему соответству-

ет кружок, точно так же, как до сих пор кружки соответствовали отдельному объекту.

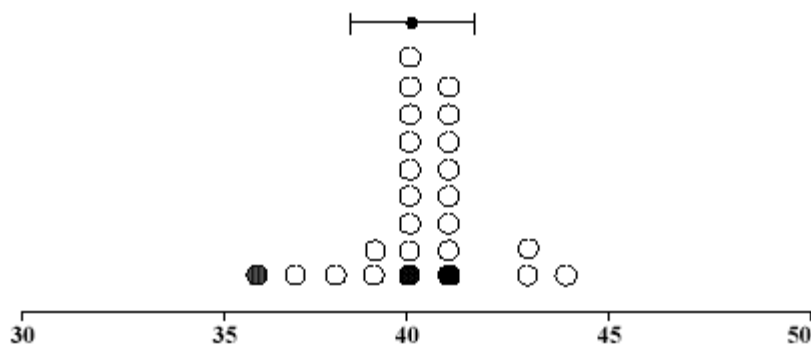


Рис. 13. Такое распределение мы получим, выбрав 25 раз по 10 особей из совокупности представленной на рис 12А, и рассчитав среднее для каждой выборки (средние для трех выборок показаны заполненными кружками).

Посмотрим на рис. 13. Набор из 25 выборочных средних имеет колоколообразное распределение похожее на нормальное. Это не случайно. Можно доказать, что если переменная представляет собой сумму большого числа независимых переменных, то ее распределение стремится к нормальному, какими бы ни были распределения переменных, образующих сумму. Так как выборочное среднее определяется именно такой суммой, его распределение стремится к нормальному, причем, чем больше объем выборок, тем точнее приближение. (Если выборки принадлежат совокупности с нормальным распределением, распределение выборочных средних будет нормальным независимо от объема выборок).

Если построить распределение средних для всех возможных выборок, оно окажется нормальным. Среднее этого распределения будет равно среднему той совокупности, из которой извлекаются выборки. Стандартное отклонение этого распределения называется стандартной ошибкой среднего. Поскольку распределение на рис. 13 нормальное, его можно описать с помощью среднего и стандартного отклонения.

Так как среднее значение для рассматриваемых 25 точек есть среднее величин, которые сами являются средними значениями, обозначим его  $\bar{\bar{x}}$ . Аналогично, стандартное отклонение обозначим  $\bar{s}$ . Величины средневзвешенных средней и среднеквадратического отклонения вычисляются по формулам:

$$\bar{x} = \frac{\sum_{i=1}^v \bar{x}_i n_i}{\sum_{i=1}^v n_i} \quad (13),$$

$$\bar{s} = \sqrt{\frac{\sum_{i=1}^v (\bar{x}_i - \bar{x})^2}{v-1}} \quad (14),$$

где:  $n_i$ ,  $\bar{x}_i$  и  $s_i$  – объем выборки, среднее значение и её среднеквадратическое отклонение для каждой  $i$ -й группы,  $v$  - число независимых выборочных групп.

По формулам для средневзвешенного среднего и стандартного отклонения находим  $\bar{x} = 40$  см и  $\bar{s} = 1,6$  см. Среднее выборочных средних  $\bar{x}$  оказалось равно среднему  $\mu$  всей совокупности из 200 особей. Ничего неожиданного в этом нет. Действительно, если бы мы провели исследования всех возможных выборок, то каждая из 200 особей была бы выбрана равное число раз. Итак, среднее выборочных средних совпадает со средним по совокупности.

Интересно, равно ли средневзвешенное стандартное отклонение  $\bar{s}$ , среднеквадратическому отклонению  $\sigma$  совокупности из 200 особей? Стандартное отклонение для совокупности выборочных средних  $\bar{s}$  равно 1,6 см, а стандартное отклонение самой совокупности —  $\sigma = 5$  см. Почему средневзвешенное  $\bar{s}$  меньше, чем  $\sigma$ ? В общих чертах это можно понять, если учесть, что в случайную выборку редко будут попадать одни только коротышки и одни гиганты. Чаше их будет примерно поровну, и отклонения роста от среднего будут сглаживаться. Даже в выборке, куда попадут 10 самых высоких особей, средний рост составит только 50 см, тогда как рост самой высокой особи — 53 см.

Подобно тому, как стандартное отклонение исходной выборки из 10 особей  $s$  служит оценкой изменчивости роста особей,  $\bar{s}$  является оценкой изменчивости значений средних для выборок по 10 особей в каждой. Таким образом, величина  $\bar{s}$  служит мерой точности, с которой выборочное среднее  $\bar{x}$  является оценкой среднего по совокупности  $\mu$ . Поэтому  $\bar{s}$  носит название **стандартной ошибки среднего**.

Чем больше выборка, тем точнее оценка среднего и тем меньше его стандартная ошибка. Чем больше изменчивость исходной совокупности, тем больше изменчивость выборочных средних, поэтому стандартная ошибка среднего возрастает с увеличением стандартного отклонения совокупности.

Так как возможные значения выборочного среднего стремятся к нормальному распределению, истинное среднее по совокупности примерно в 95% случаев лежит в пределах 2 стандартных ошибок выборочного среднего.

Как уже говорилось, распределение выборочных средних приближенно всегда следует нормальному распределению не зависимо от распределения совокупности, из которой извлечены выборки. В этом и состоит суть утверждения, называемого *центральной предельной теоремой*. Эта теорема гласит следующее.

- □ Выборочные средние имеют приближенно нормальное распределение независимо от распределения исходной совокупности, из которой были извлечены выборки.

- □ Среднее значение всех возможных выборочных средних равно среднему исходной совокупности.

- □ Стандартное отклонение всех возможных средних по выборкам данного объема, называемое стандартной ошибкой среднего, зависит как от стандартного отклонения совокупности, так и от объема выборки.

## 2.5.2 Приведение к нормальному виду качественных признаков

Если бы в наших руках были данные по всем членам совокупности, то не было бы никаких проблем связанных с точностью оценок. Однако нам всегда приходится довольствоваться ограниченной выборкой. Поэтому возникает вопрос, насколько точно доли в выборке соответствуют долям в совокупности. Проведем мысленный эксперимент наподобие того, который мы провели в гл. 5.1, когда рассматривали насколько хорошей оценкой среднего по совокупности является выборочное среднее.

Предположим, что из всех 200 особей случайным образом выбрали 10. Распределение розовых и зеленых особей во всей совокупности неизвестное исследователям изображено в верх ней части рис. 14. Закрашенные кружки соответствуют особям, попавшим в выборку. В нижней части рис. 14 показана информация, которой располагал бы исследователь, получивший такую выборку. Как видим в выборке розовые, и зеленые особи поделились поровну. Основываясь на этих данных, мы решили бы, что розовых особей столько же, сколько и зеленых, то есть их доля составляет 50%.

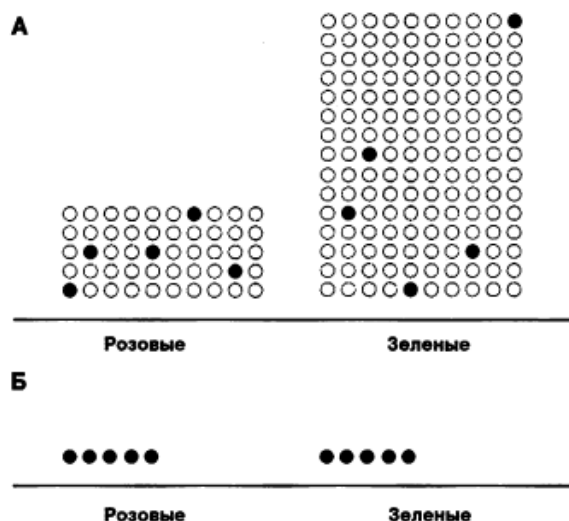
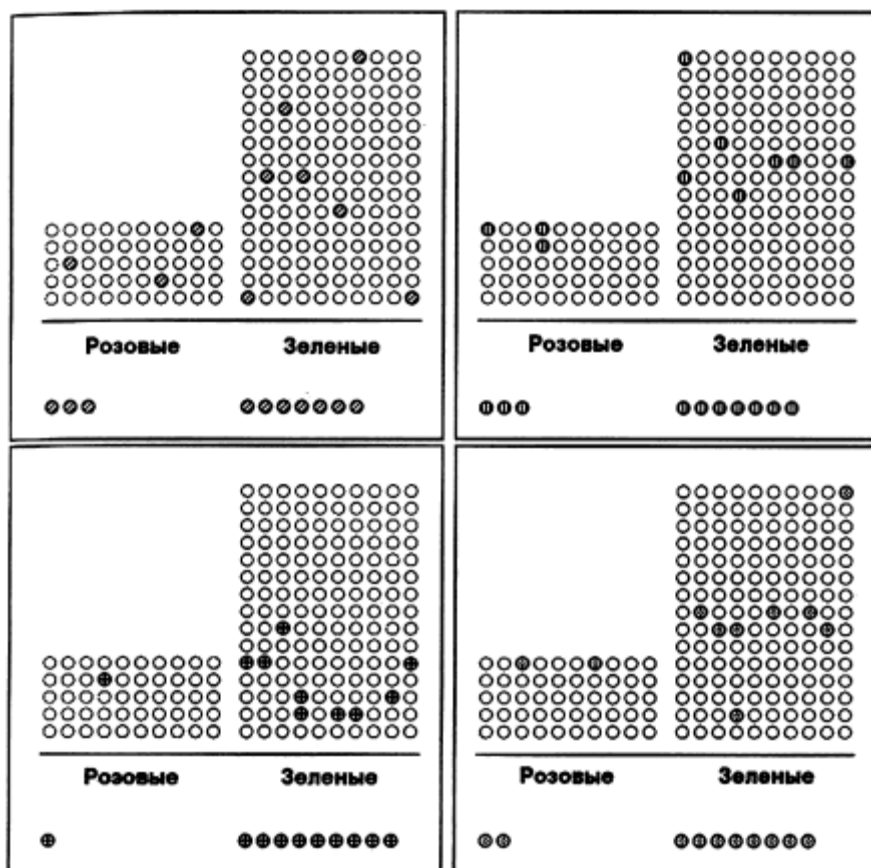


Рис. 14. А. Из совокупности особей, среди которых 150 зеленых и 50 розовых, извлекли случайную выборку из 10 особей. В выборку попало 5 зеленых и 5 розовых, на рисунке они помечены черным. Б. В таком виде данные представят перед исследователем, который не может наблюдать всю совокупность и вынужден судить о ней по выборке. Оценка доли розовых особей  $W = 5/10 = 0,5$ .

Исследователь мог бы извлечь другую выборку, например одну из представленных на рис. 15. Здесь выборочные доли розовых особей равны 30, 30, 10, и 20%. Как любая выборочная оценка, оценка доли (обозначим ее  $W$ ) отражает долю  $W$  в совокупности, но отклоняется от нее в силу случайности. Рассмотрим теперь не совокупность особей, а совокупность всех значений  $W$ , вычисленных по выборкам объемом 10 каждая. (Из совокупности в 200 членов можно получить более  $10^6$  таких выборок). На рис. 5.6 приведены пять значений  $W$ , вычисленных по пяти выборкам с рис. 14 и 15 и еще 20 значений полученных на других случайных выборках того же объема.



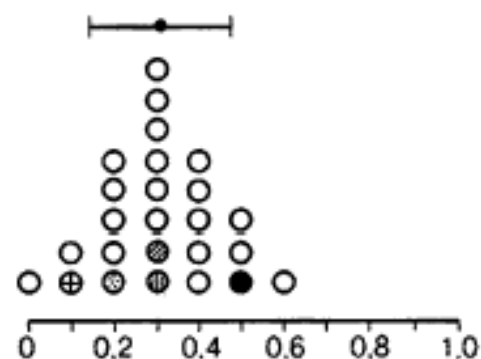
Среднее этих 25 значений составляет 30%. Это близко к истинной доле ро-



зовых марсиан — 25%. По аналогии со стандартной ошибкой среднего найдем *стандартную ошибку доли*. Для этого нужно охарактеризовать разброс выборочных оценок доли, то есть рассчитать стандартное отклонение совокупности  $W$ . В данном случае оно равно примерно 14%.

Рис. 15. Еще 4 случайные выборки из той же совокупности особей. Оценки доли розовых особей: 30, 30, 10 и 20%.

Рис. 16. Нанесем на график оценки доли розовых особей, полученные по выборке с рис. 24 и четырем выборкам с рис. 25. Добавим к ним еще 20 выборочных оценок. Получилось распределение выборочных оценок  $W$ . Стандартное отклонение совокупности средних — это стандартная ошибка доли.



Среднее этих 25 значений составляет 30%. Это близко к истинной доле розовых — 25%. По аналогии со стандартной ошибкой среднего  $\bar{x}$  найдем стандартную ошибку доли. Для этого нужно охарактеризовать разброс выборочных оценок доли, то есть рассчитать стандартное отклонение совокупности  $W$ . В данном случае оно равно 14%.

Из центральной предельной теоремы вытекает, что при достаточно большом объеме выборки выборочная оценка  $W$  приблизительно подчиняется нормальному распределению, имеющему среднее  $W$  и стандартное отклонение  $\sigma_w$ . Однако при значениях  $W$ , близких к 0 или 1, и при малом объеме выборки это не так. При какой численности выборки можно пользоваться приведенным способом оценки? Математическая статистика утверждает, что нормальное распределение служит хорошим приближением, если использовать распределения средних значений малых выборок. Напомним, что примерно 95% всех членов нормально

распределенной совокупности находятся в пределах двух стандартных отклонений от средневзвешенного.

Для ответа на вопрос какое минимальное количество выборок необходимо взять на исследование и сколько объектов должно состоять в каждой группе для получения достоверного результата необходимо познакомиться со свойствами малых выборок.

### 2.5.3 Точность определения выборочных оценок

Для анализа свойств малых и больших выборок рассмотрим пример распределения кроликов по отдельным выборкам по 10 штук в каждом (табл. 6). Всю совокупность кроликов весом от 3,05 до 3,69 кг (генеральное распределение 200 штук) разбили на 13 групп с интервалом в 40 грамм. Основная выборка имела нормальное распределение, тогда как малые выборки отличались от нормального распределения (оказались двухмодальными). Средние значения оказались близкими во всех выборках, тогда как среднеквадратическое отклонение в первой оказалось меньше почти в два раза, чем в третьей.

Усредняя значения средней величины и среднеквадратического отклонения по каждой из малых выборок (общим объемом  $n=30$ ) получаем величины, которые имеют относительную погрешность в сравнении с большой выборкой (объемом  $n=200$ ) по средней величине 1,2 % и среднеквадратическим отклонением 31%.

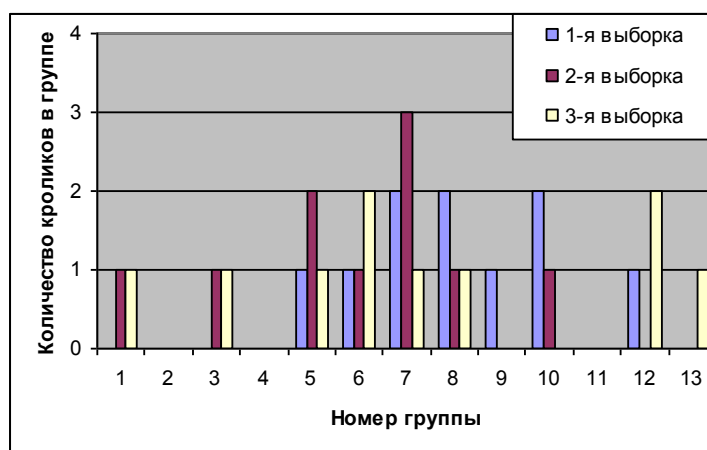
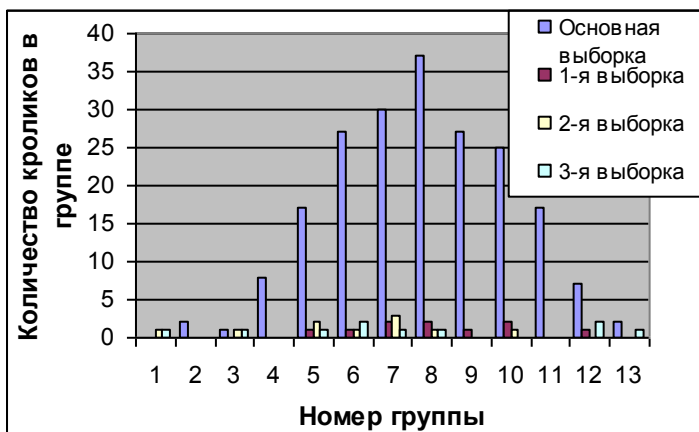


Таблица 6. Распределение кроликов в группах по весу

Вес кроликов по группам, кг	Количество кроликов в группе			
	Основная выборка	1-я выборка	2-я выборка	3-я выборка
3,05—3,09	—	—	1	1
3,10—3,14	2	—	—	—
3,15—3,19	1	—	1	1
3,20—3,24	8	—	—	—
3,25—3,29	17	1	2	1
3,30—3,34	27	1	1	2
3,35—3,39	30	2	3	1
3,40—3,44	37	2	1	1
3,45—3,49	27	1	—	—
3,50—3,54	25	2	1	—
3,55—3,59	17	—	—	—
3,60—3,64	7	1	—	2
3,65—3,69	2	—	—	1
Всего	200	10	10	10
$\bar{x}$	3,416	3,430	3,315	3,385
$s$	0,110	0,105	0,128	0,200

В случае если усредняемые выборки разные по объему, то необходимо использовать формулы средневзвешенного значения.

### Определение нормированного отклонения

Многие наблюдаемые явления подчиняются приблизительно нормальному закону распределения. По этой причине основная часть классической статистической теории предполагает нормальность рассматриваемой случайной величины. Важным теоретическим результатом относительно выборочных распределений является **центральная предельная теорема**. Вот одно из важных применений этой теоремы: если мы получаем случайную выборку объема  $n$  из генеральной совокупности с конечной дисперсией, то независимо от распределения нашей случайной величины  $X$  распределение выборочного среднего  $M$  при больших  $n$  будет приблизительно нормальным [24].

Соблюдения условия нормальности распределения выборки позволяет использовать как все представленные в данной работе формулы, так и установление величины нормированного отклонения. Отклонение той или иной варианты от средней арифметической, отнесенной к величине среднего квадратического отклонения, называют **нормированным отклонением**:

$$t = \frac{\bar{x} - x_i}{\sigma} \quad (15).$$

Этот показатель позволяет "измерять" отклонение отдельных вариантов от среднего уровня и сравнивать их для разных признаков. Иными словами, он характеризует индивидуальность признака и позволяет установить сопряженность связей между аргументами и величиной целевой функции. Так же оценку нахождения и развития любой как физической так и биологической системы можно произвести при помощи нормируемого отклонения (критерия Стьюдента –  $t$ ). При  $t=\pm 1$  состояние системы находится в оптимуме - opt,  $t=+1 \div +2$  - supopt состояние,  $t=-1 \div -2$  - subopt состояние,  $t=+2 \div +3$  - suppat состояние,  $t=-2 \div -3$  - subpat состояние (рис. 7).

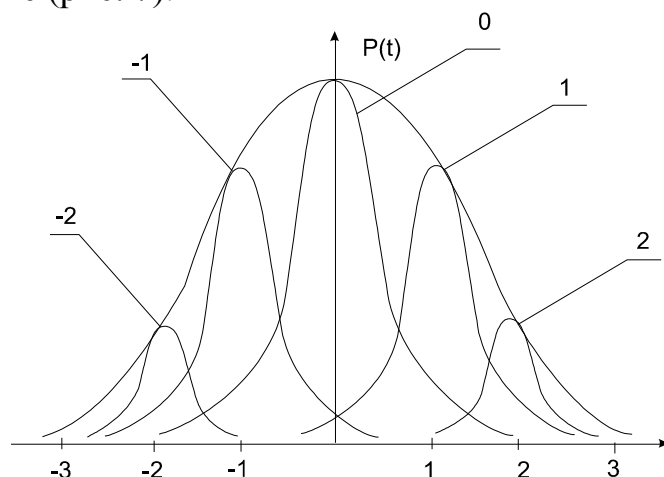


Рис. 7. Стандартизованная форма нормальной кривой генерального распределения, включающего спектр отдельных составляющих его гармоник.  $P(t)$  - функция плотности вероятности распределения изучаемого явления;  $t$  - нормируемое отклонение (критерий Стьюдента); 0,  $\pm 1$ ,  $\pm 2$  - отдельные составляющие гармоник.

Например, проведем сравнительный анализ существующих технологий криоконсервации биообъекта. Полагаем, что центральная теорема статистики утверждающая, что всякое распределение стремится к нормальному при повышении количества измерений, может служить методологическим постулатом целесообразности выбора приоритетности развития изучаемого явления. Различные направления исследований являются теми факторами, которые расширяют диапазон и приводят генеральное распределение к нормальному. Только та система будет развиваться, рост которой сбалансирован противоположно дополняющими началами. Односторонность развития приводит к обратным явлениям: деградации и распаду на отдельные составляющие гармоник общего распределения системы. Условия равновесия развития будут выполняться только тогда, когда система соответствует нормальному закону распределения составляющих её гармоник. И как следствие, приоритетность развития каждой из них должна соответствовать весу её плотности вероятности распределения в общей генеральной совокупности.

Величина вероятности распределения каждой из гармоник  $P(t)$ , в нашем представлении, является прямым аналогом приоритетности направления изучаемого явления для любой из составляющих генеральной совокупности,  $t$  - количественный показатель степени отклонения конкретной гармоники от условно оптимального состояния целостной генеральной совокупности. Отрицательные гармоники ( $t = -1, -2$ ) - показатели субоптимальности режимов, а положительные - супероптимальности. Представленные на графике численные значения не обязательно должны быть целыми величинами, они могут принимать любые значения. Чем ближе к нулевому (равновесному) состоянию расположена гармоника, тем она ближе к оптимуму и, следовательно, большим приоритетом должна обладать.

Адекватная оценка изучаемого явления требует проведения полного и всестороннего анализа с учётом всех его качеств и характеристик. Проведение объективного анализа возможно при наличии обобщающего фактора, который позволяет не только уравновесить, но и сгладить противоречия.

Ускорение процессов развития организма и создание новых видов растений и животных, которые не существуют в живой природе, волнуют не одно поколение ученых и способствуют созданию новых отраслей биологической науки. Выбор приоритетных направлений, способствующих скорейшей реализации поставленной задачи, затруднен установлением критериев оценки значимости их дальнейшего развития, что предопределено как объективными, так и субъективными причинами. Исследователь на различных этапах своей деятельности сталкивается с наличием не только противоположных мнений, но даже направлений работы, которые, в свою очередь, могут, как исключать, так и дополнять друг друга. Ошибка выбора приоритета развития как раз и заключается в выделении какой-либо из особенностей, отрицая остальные и, забывая принцип соотношения Бора о том, что каждая новая теория должна включать в себя элемент старой как предельный случай новой. Даже противоположные направления зачастую имеют общее, объединяющее начало не только при выборе объекта исследования или методологии, а и при постановке задачи в целом. Важно отыскать это начало и попытаться оттолкнуться от него.

Создание алгоритма объективной оценки изучаемого явления с учетом всего спектра составляющих его качеств позволяет получить максимальный эффект при его дальнейшей реализации.

Выполнение эксперимента (см. структурную схему проведения биологического исследования) сопряжено с проведением измерений и их статистической обработкой, начиная от поисковых опытов и кончая выполнением всего эксперимента в целом. По окончании первичной статистической обработки полученных данных можно приступить к планированию эксперимента на основе проведенных поисковых опытов.

**Заключение.** Когда совокупность подчиняется нормальному распределению, она исчерпывающе описывается *параметрами распределения* — средним и стандартным отклонением. Когда же распределение сильно отличается от нормального, более информативны мода и процентиля.

Так как наблюдать всю совокупность удается редко, мы *оцениваем* параметры распределения по выборке, случайным образом извлеченной из совокупности. Стандартная отклонение среднего служит мерой точности, с которой выборочное среднее является оценкой среднего по совокупности.

### ВОПРОСЫ ДЛЯ САМОКОНТРОЛЯ.

1. Определите основную задачу научного исследования? В чем состоит принципиальное отличие наблюдения от эксперимента?
2. Перечислите особенности сбора, накопление и группировки данных отражающих процесс исследуемого явления.
3. Охарактеризуйте особенности шкалы качественных признаков
4. Охарактеризуйте особенности шкалы количественных признаков
5. Опишите основные принципы построения вариационного ряда
6. Перечислите основные ошибки измерения и особенности их появления.
7. Назовите способы устранения грубых, систематических и случайных ошибок измерения.
8. В чем заключено различие качественного и количественного статистического анализа?
9. Какова необходимость применения статистических показателей: медиана, мода, квартиль? Опишите их характерные особенности.
10. Как производится вычисление величин среднего арифметического значения при проведении качественного и количественного анализов?
11. Как производится вычисление значений среднего квадратического отклонения при проведении качественного и количественного анализов?
12. Что характеризует показатель коэффициента вариации выборки.
13. Какую функцию отражает величина средней квадратической ошибки среднего арифметического?
14. Как устанавливается доверительный интервал?
15. При помощи, каких статистических показателей исследуется индивидуальная изменчивость (варьирование) признаков?
16. Какие виды распределений признаков Вы знаете? Опишите их характерные особенности.
17. Каким образом можно привести не нормальное распределение к нормальному?
18. Какова необходимость вычисления средней взвешенной величины и ее ошибки?
19. Перечислите основные функции нормированного отклонения.

### Глава 3. Планирование эксперимента

Экспериментальные исследования являются основным источником получения достоверных сведений об объектах реального мира. Такие исследования проводятся с целью выбора рациональных технологических режимов функционирования или оптимизации параметров систем, оценки степени выполнения заданных требований к создаваемым изделиям, выяснения закономерностей функционирования, анализа влияния факторов на показатели качества систем и т.д. Натурные исследования свойств технических средств или сложных моделей требуют значительных затрат ресурсов. Данное обстоятельство заставляет уделять серьезное внимание рациональной организации экспериментального изучения таких объектов.

Экспериментальные данные формируются путем пассивного наблюдения либо с помощью активного эксперимента. При пассивном наблюдении информация получается путем регистрации необходимых сведений в условиях обычного функционирования объекта. В активном эксперименте производится целенаправленное воздействие на объект по заранее составленной схеме. Активный эксперимент позволяет расширить область исследования, точнее вскрыть закономерности функционирования, сократить потребности в ресурсах на проведение исследования. Но организация и проведение активного эксперимента сложнее пассивного. Кроме того, следует учитывать и принципиальные ограничения в проведении активных экспериментов на действующих объектах, невозможность их осуществления для недоступных объектов.

Планирование экспериментов (ПЭ) охватывает широкий круг вопросов – от учета конкретных особенностей определенных объектов исследования до общих концептуальных проблем. Целесообразность создания и начальные положения специальной теории планирования эксперимента впервые были сформулированы в Англии на агробиологической станции в Ротамстеде в ходе решения практических задач по растениеводству в конце XIX века. Серьезный вклад в становление теории внес Р.А. Фишер, именно его работы в тридцатых годах прошедшего столетия заложили основы теории статистического анализа, начал науки о планировании и анализе сравнительных экспериментов. Среди наших соотечественников существенную роль в становлении отечественной школы в области теории ПЭ и ее практического применения сыграли Ю.П. Адлер, Е.В. Маркова, В.В. Налимов, В.В. Федоров и другие ученые.

В настоящее время ТПЭ выступает как самостоятельное научное направление и находит практическое применение там, где проводятся многофакторные научные и технические экспериментальные исследования. Теория использует аппарат математической статистики, линейной алгебры, комбинаторики и других разделов математики.

Методы теории планирования экспериментов направлены на разработку оптимальных планов проведения экспериментов с целью сокращения объема проводимых исследований при заданной точности и достоверности получения

результатов, извлечения из полученных опытных данных максимума полезных сведений. Составной частью ТПЭ является исследование способов обработки результатов эксперимента, проведенного по выбранному плану, анализ свойств получаемых оценок показателей качества объекта. Экспериментальные данные, полученные с помощью ТПЭ, часто являются основой для применения других математических методов, например градиентных методов оптимизации.

Наличие формальной теории планирования эксперимента не исключает необходимости четкого представления физических основ процессов, протекающих в объекте исследования, факторов, воздействующих на объект. Эти сведения нужны на этапе составления плана эксперимента, при анализе и интерпретации результатов.

### 3.1 Учет ошибок измерения и вычисления.

Перед проведением планирования эксперимента необходимо определить, в ходе выполнения поисковых опытов, ошибки измерения и вычисления, которые зависят от соответствующих методов. В ходе статистического анализа можно выявить ошибки. Их принято делить: по стадии возникновения на ошибки регистрации (измерения) и репрезентативности (статистической обработки); по причинам возникновения на случайные и систематические (рис. 27).

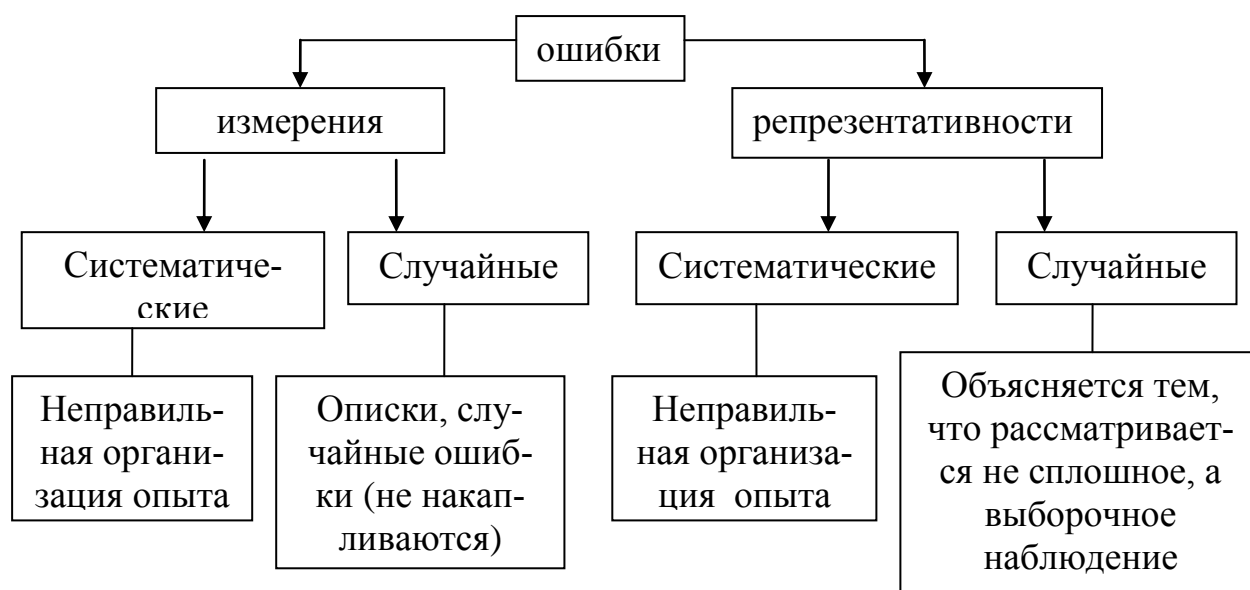


Рис. 27. Составляющие компоненты ошибки выборки.

Ошибки репрезентативности – ошибки выборочного наблюдения: чем больше объем выборки, тем меньше ошибка. Это связано с тем, что структура выборки может не совпадать со структурой генеральной совокупности, откуда производилась выборка.

При сплошном наблюдении ошибка отсутствует. Ведущая ошибка репрезентативности связана с не учетом неоднородностей при анализе нескольких выборок. Это приводит к проблеме не воспроизводимости и как следствие этому не



сопоставимости полученных результатов. Случайные ошибки при большом массиве информации, не имеют направленности, взаимопогашаются и не искажают результаты анализа. Систематические ошибки имеют определенную направленность, приводят к искажению результатов.

Снижение ошибки выборки связано с требованиями, предъявляемые к информации: достоверность, воспроизводимость и сопоставимость. Достоверность - информация должна правдиво отражать явления и точно рассчитываться. Воспроизводимость - результаты отдельных опытов должны иметь одинаковые дисперсии. Сопоставимость - средние значения этих опытов должны совпадать, то есть быть соразмерными.

Например, применяемые методы измерения и обработки полученных данных в криобиологических исследованиях обеспечивают достоверность результата только при сравнении сохранности биообъекта в опытной группе с контролем при одинаковых условиях проведения эксперимента. Методологический приём парных сравнений - “контроль-опыт”, применяемый в одном эксперименте, не даёт возможности сопоставлять результаты, полученные в разное время и тем более при использовании различных способов криоконсервирования. Это приводит к проблеме не сопоставимости абсолютных значений, полученных при реализации разных методов или их модификаций, что создает сложность анализа и последующего обобщения данных.

Проведение криобиологического исследования сопряжено с необходимостью применения комплекса взаимосвязанных методов. При этом результат оказывается неотделимым от методов при помощи, которых он получен. При таком подходе анализ заданного способа криоконсервирования подобен работе с “черным ящиком”. Вместе с тем существует необходимость сравнения результатов опытов полученных в изменяющихся условиях проведения эксперимента (опытов проведенных при разных условиях).

Таким образом, ошибки исследования состоят из ошибок регистрации (измерения) и репрезентативности (обработки, вычисления). Ошибку регистрации можно снизить посредством повышения класса используемого прибора, а репрезентативности применением адекватных методов обработки полученной информации.

Рассмотрим способы снижения ошибок обработки полученной информации.

**3.1.1 Установление ошибки измерения.** Биологические признаки варьируют под влиянием самых различных, в том числе и случайных, причин. Наряду с естественным варьированием на величине признаков сказываются и ошибки, неизбежно возникающие при измерениях изучаемых объектов. Опыт показал, что как бы точно ни были проведены измерения, они всегда сопровождаются отклонениями от действительного значения измеряемой величины, т. е. не могут быть проведены абсолютно точно. Разница между результатами измерений и действительно существующими значениями измеряемой величины называется погрешностью или

ошибкой. Ошибкой измерения называется разность между результатом измерения  $X_i$  и истинным значением  $X$ , измеряемой величины ( $X_i - X$ ).

Ошибки возникают из-за неисправности или неточности измерительных приборов и инструментов (технические ошибки), личных качеств исследователя, его навыков и мастерства в работе (личные ошибки) и от целого ряда других, не поддающихся регулированию и неустраняемых причин (случайные ошибки).

Технические и личные ошибки, объединяемые в категорию **систематических**, т. е. неслучайных ошибок, можно в значительной степени преодолеть, совершенствуя технические средства, условия работы и личный опыт. Эти меры позволяют свести размеры таких ошибок до минимума, которым можно пренебречь. Случайные же ошибки, как независимые от воли человека, остаются и сказываются на результатах наблюдений.

Метод вариационной статистики возник как метод борьбы с ошибками, поскольку все измерения, как бы тщательно они не были выполнены, дают неточное, а лишь приближенное значение. Ошибки бывают **грубые** (см разд), **систематические** и **случайные**.

**Исключение систематических ошибок.** Систематические ошибки могут зависеть от точности приборов, реактивов, применяемых методов или от индивидуальных особенностей экспериментатора. Обнаружение их требует таких специальных исследований, как измерение одной и той же величины разными методами или нескольких эталонов с помощью одного и того же прибора и т.д. К началу математической обработки систематические ошибки должны быть обнаружены и сведены путем соответствующих поправок до ничтожно малых величин.

**Снижение величины случайных ошибок.** Случайные ошибки могут быть следствием влияния большого количества разнообразных факторов на результат измерений, в том числе и биологической изменчивости. Эти ошибки не могут быть учтены и устранены путем введения каких-либо поправок, так как они не определены по величине и знаку. Однако с помощью методов теории вероятности можно учесть их влияние на оценку истинного значения измеряемой величины.

Учет влияния случайных ошибок основан на знании законов их распределения, причем чаще всего принимается закон распределения Гаусса. Совокупность всех ошибок измерений называют ошибкой измерения. При выполнении опыта проводится ряд измерений, а, следовательно, и возникает ошибка опыта.



Рис. 28. Общая вариация выборки, включает в себя межгрупповую и внутригрупповые составляющие дисперсии.

### 3.1.2 Установление ошибки вычислений.

Ошибки репрезентативности (вычисления) по аналогии с ошибками измерения так же зависят от случайных и систематических факторов. К систематическим

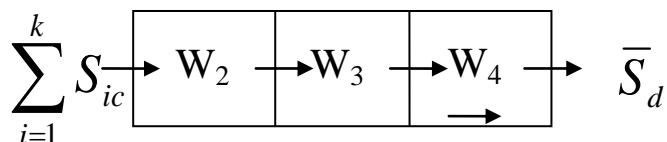
следует отнести выбор схемы проведения исследования и метода статистической обработки. Случайные факторы зависят от естественной изменчивости исследуемых признаков, которые определяются межгрупповой и внутригрупповой вариацией.

До проведения поисковых опытов, как правило, общая, меж- и внутригрупповая вариация выборки неизвестна. Поэтому основной **задачей** планирования эксперимента и есть оценка наиболее вероятного значения данных **величин** в ходе проведения и обработки результатов поисковых опытов. Способы оценки общей, меж- и внутригрупповой вариации выборки см. в разделе дисперсионного анализа. Межгрупповую вариацию можно свести к минимальному значению при использовании прямых методов измерений и правильной группировки данных с целью формирования однородных выборок. Минимальность величины межгрупповой дисперсии является показателем эффективности планирования эксперимента.

### 3.1.3 Установление ошибки обработки результатов опыта.

В соответствии с методами обработки результатов все способы измерения можно разбить на два класса: измерения **прямые**, когда истинную величину измеряют непосредственно, и измерения **косвенные** или опосредованные, когда истинная величина является **функцией** измеряемых величин. *Пример измерения температуры при помощи термометра и термопары.*

**Исключение систематических ошибок вычисления.** При прямом измерении ошибка опыта равна ошибке отдельного измерения  $\Delta \approx \Delta_i$ . В результате искомая величина находится прямо по показаниям измерительного прибора, например, сила тока – по отклонению стрелки амперметра, вес – по растяжению пружинных весов и т.д. Однако гораздо чаще измерения проводят косвенно, например, площадь прямоугольника определяют по измерению длин его сторон, электрическое сопротивление – по измерениям силы тока и напряжения и т.д. Во всех этих случаях искомое значение измеряемой величины получается путем соответствующих расчетов. Поэтому при проведении косвенных измерений ошибка опыта равняется сумме всех ошибок измерения  $\Delta \approx \sum \Delta_i$ .



Пример. В случае криоконсервирования биообъекта, имеющего разное качество, общепринято производить усреднение показателя сохранности биоматериала. Обобщение разнокачественного материала приводит к неоднородности выборки, что увеличивает величину межгрупповой дисперсии. Ошибка опыта увеличивается пропорционально количеству разнородных групп эмбрионов. При проведении последующего эксперимента величина ошибки опыта накапливается пропорционально количеству этапов криоконсервирования.

Таким образом, общая ошибка опыта будет содержать количество долей ошибок равной произведению числа неоднородных групп биообъекта (для эмбрионов удовлетворительное, хорошее, отличное качество) на количество этапов

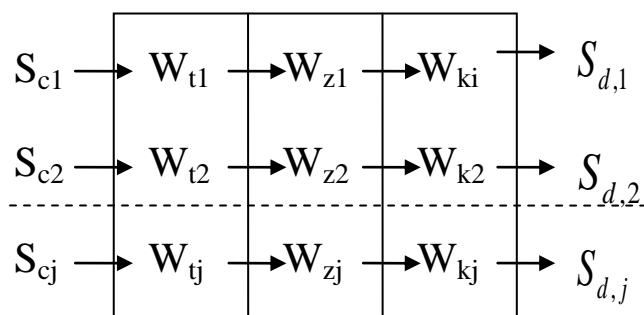
криоконсервирования (применение криопротектора, процедура замораживания-оттаивания, культивирование).

Так как измерения косвенные зависят от ряда прямых, то при прочих равных условиях всегда выгоднее применять тот способ, при котором будет меньше косвенных измерений, а значит и меньше сумма ошибок. Вычисление относительной предельной (наибольшей, случайной ошибки) ошибки опыта оцениваются на основании следующих правил:

а) **ошибка суммы** заключена между наибольшей и наименьшей из относительных ошибок слагаемых; практически берут или наибольшую относительную ошибку, или среднюю арифметическую;

б) **ошибка произведения** и частного от деления равна сумме относительных ошибок сомножителей или соответственно делимого и частного;

в) **ошибка  $n$ -й степени** какого-либо основания (величины) в  $N$  раз больше относительной ошибки основания.



Ошибку опыта можно снизить посредством перехода от косвенных измерений к прямым. Работа с однородной выборкой (состоящей из одинакового качества биообъекта) не создает межгрупповой дисперсии. Тем самым ошибку опыта можно снизить пропорционально количеству разнородных групп содержащих исследуемый биообъект. При необходимости усреднение производится только конечного результата.

Общеизвестно, что сравнивать можно только однородные величины, т. е. имеющие одинаковую размерность. Для оценки вариабельности абсолютной величины используется показатель среднеквадратическое отклонение  $\phi$  (7), а количественное сопоставление однородных показателей производится при помощи отношения дисперсий данных величин (дисперсионный анализ). Сравнение вариабельности разнородных величин (имеющих различные размерности) осуществляется при помощи перевода их в относительные показатели (коэффициент вариации – отношение среднеквадратического отклонения к средней величине  $\phi$  (8)). Поэтому для сопоставления разнородных показателей используют прием перевода их в относительную систему координат. Для описания корреляционной связи стохастических параметров применяют критерий нормированного отклонения ( $t$  критерий Стьюдента – отношение разности средней величины и текущей к среднеквадратическому отклонению), а при оптимизации детерминистических используют относительные величины [4, 5, 6].

Для сопоставимости результатов полученных при разных условиях их реализации используется так же показатель эффективности. Термин "эффективность" означает определенный результат известных усилий. Оценка этого результата проводится на основании количественных показателей, отражающих степень достижения намеченной цели. При этом и усилия, и результат должны быть измерены в сопоставимых единицах, а их величины подвергнуты сравне-

нию. Эффективность выражают в процентном отношении значений полученного результата к общим усилиям. Частным случаем данного показателя является коэффициент полезного действия - отношение полезно использованной энергии к затраченной. Экономическая эффективность (рентабельность) вычисляется как отношение прибыли к затратам.

При использовании относительных показателей – эффективности ошибку опыта можно уменьшить до величины отдельного измерения. В случае с эмбрионами млекопитающих в девять раз, при этом их количество можно сократить до 80 раз при обеспечении достоверного результата.

Таким образом, не учет начального состояния биологического материала и эффективности этапов его криоконсервирования является причиной создания межгрупповой вариации. Усреднение значений неоднородных выборок (от начального до и конечного после всех этапов криоконсервирования разнокачественного биообъекта) приводит к накоплению ошибок опыта. Учет индивидуальных свойств биообъекта и особенностей этапов криоконсервирования дает возможность использовать прямые методы измерения, существенно снижающие погрешность опыта.

Вычисления можно проводить как угодно точно, но результат вычисления не может быть точнее тех данных, на которых оно основано. Установление общей ошибки эксперимента, при проведении и анализе поисковых опытов, является основой определения условия решения поставленной проблемы.

## **3.2 Определение условий решения поставленной задачи**

**3.2.1 Проведение поисковых опытов.** Задача математической статистики, в строгом понимании этого термина, состоит в разработке и применении методов описания реальных явлений вероятностными моделями, исходя из данных, полученных в результате наблюдений за этими явлениями. В более широком смысле математическая статистика понимается как совокупность методов планирования и обработки данных, полученных в результате экспериментов или наблюдений, причем эти методы могут не основываться на вероятностных моделях. При таком широком понимании вместо термина «математическая статистика» часто используют термин «анализ данных».

Исторически вначале сформировались методы обработки данных, не связанные тесно с теорией вероятности, так называемая дескриптивная, или описательная статистика. С начала этого века начали интенсивно развиваться методы анализа данных, основанные на вероятностных моделях, - это, прежде всего, методы статистического оценивания и статистической проверки гипотез. Классические работы Р. Фишера открыли новую страницу в истории биометрии. Они показали, что планирование эксперимента и обработка результатов - это две тесно связанные между собой задачи статистического анализа.

Бурное развитие вычислительной техники вызвало к жизни ряд новых методов анализа. Некоторые из этих методов разработаны на основе подходов, отличных от теоретико-вероятностного (геометрические, оптимизационные и др.). Вероятностное обоснование этих методов либо отсутствует, либо недостаточно развито, что затрудняет количественную оценку степени достоверности получа-

емых выводов аналитическими средствами классической математической статистики. Однако в последние годы, в связи с быстрым ростом производительности вычислительных машин, начали также получать распространение процедуры так называемого стохастического моделирования (пермутационные методы, бутстрэп и др.), позволяющие оценить статистические свойства получаемых решений без использования аналитических методов.

Поисковые опыты проводят для того, чтобы найти условие решения поставленной на исследование проблемы. Поэтому целью поисковых опытов является проверка отдельных частей разработанной методики, приспособленности приборов к тем измерениям, которые определила методика и расчет количества опытов необходимых для получения достоверного результата.

Выбор величины доверительной вероятности является отправной точкой планирования эксперимента. Доверительная вероятность или надежность  $P$  – это вероятность того, что среднее значение выборки  $\bar{x}$  попадает в интервал  $\mu - t_{p,m} \leq \bar{x} \leq \mu + t_{p,m}$ , где  $\mu$  – истинное значение генеральной совокупности ( $\bar{x} \rightarrow \mu$  при  $n \rightarrow \infty$ ). Установив в ходе поискового опыта основные статистические величины изучаемых явлений, как то среднеарифметическую, среднеквадратическое отклонение и др., можно приступить к планированию эксперимента.

Как правило, минимально допустимая доверительная вероятность, рекомендуемая при определении доверительных интервалов в биологических исследованиях, равна 95% ( $P=0,95$ ), а уровень значимости оценок 5% ( $p=0,05$ ):

$$p = 1 - P = \frac{\Delta}{M} \quad (17),$$

где  $\Delta$  - доверительный интервал  $\Delta = t_m$ .

Доверительный интервал оценивается в 5 процентов от выборочной средней  $M$ . Другими словами доверительная вероятность обратно пропорциональна уровню значимости оценок измерений, которая в свою очередь равна ее относительной погрешности. Величина вероятной ошибки не должна превышать величины заданной надежности для проведения данного исследования.

**Установление заданной надежности.** При всем разнообразии методов исследовательской работы задача планирования сводится к тому, чтобы при возможно минимальных объемах наблюдений получать достаточно полную информацию об изучаемых объектах. Планирование эксперимента основывается на установленной надежности –  $P$ , которая обеспечивается за счет увеличения количества измерений и уменьшения погрешностей приборов. В некотором смысле эта задача является обратной той, которая решается при статистической обработке результатов поискового эксперимента.



Применяя статистический анализ к решению практических задач, исследователь имеет дело с измерениями биологических объектов. Обычно измерения проводят с точностью до десятых, сотых или тысячных долей единицы. Более точные измерения проводят реже. Как показывает опыт, нет необходимости в точности измерений, когда эта точность практически не нужна. Данное положение относится и к измеряемым объектам, и к вычислениям, обобщающим статистические характеристики. Точность измерений и их последующих вычислений определяется установленной степенью надежности предполагаемых результатов.

Значения заданных надежностей т.е. вероятностей реализаций данных событий –  $P$  представлены в таблице 7.

Таблица 7. Требуемые уровни надежности полученного значения, связанные с проведением конкретного рода исследования.

Уровень надежности значения – $P_i$	$P_1$	$P_2$	$P_3$	$P_4$
Численное значение вероятности	0,68	0,95	0,99	0,999
Градация уровней надежности	низкий	средний	повышенный	высокий
Уровень значимости оценок – $p_i$	0,32	0,05	0,01	0,001

$P_1$  – вероятность приемлемая для поисковых опытов;

$P_2$  – для биологических исследований;

$P_{3,4}$  – для исследований, связанных со здоровьем человека.

**Определение количества значимых цифр.** Чтобы избежать грубых ошибок в работе и получить сопоставимые результаты, необходимо неукоснительно соблюдать признанные правила записей и округления приближенных чисел. Относительная ошибка при проведении поисковых опытов не должна превышать 0,32 (уровень надежности  $P_1$ ). При округлении трехзначного числа максимальное абсолютное значение погрешности составляет 5 единиц, а относительная ошибка при минимальном среднем значении 100 не превышает  $5/100=0,05$  т.е. относительная погрешность составляет  $p_1<0,1$ . Для среднего уровня надежности  $p_2=5/10^3=0,005$  ( $p_2<0,01$ ); повышенного  $p_3=5/10^4=0,0005$  ( $p_3<0,001$ ) и высокого  $p_4=5/10^5=0,00005$  ( $p_4<0,0001$ ).

Количество **значимых цифр** (минимальный порядок оперируемых чисел) обеспечивающее заданный уровень надежности, представлено в таблице 8.

Таблица 8. Количество значимых цифр, соответствующих заданному уровню надежности.

Вероятность заданного уровня – $P_i$	$P_1$	$P_2$	$P_3$	$P_4$
Количество значимых цифр – $K$	3	4	5	6

Последний знак значимого числа считается сомнительным и округление возможно только до этого знака. Установленный уровень надежности накладывает условия на общую величину погрешности эксперимента, которая не должна превышать требуемый уровень значимости. Величина погрешности эксперимента обусловлена величинами погрешностей проведенных опытов, которые в свою очередь складываются из ошибок отдельных измерений.

Для установления предельно допустимой величины погрешности эксперимента необходимо определить его общий доверительный интервал и, оценив необходимое количество измерений, которое нужно произвести, понизить погрешность до требуемого уровня значимости.

Определение доверительного интервала с учетом случайных и систематических погрешностей:

а) доверительный интервал для случайных погрешностей  $\Delta_{\varepsilon}$  определяется из следующей формулы:

$$\Delta_{\varepsilon} = t_p m \quad (18).$$

б) доверительный интервал для систематических погрешностей состоит из отдельных  $\Delta_{\theta}$  - погрешностей приборов и методов.

Погрешность средства измерения, используемого в нормальных условиях для конкретного прибора, устанавливается посредством данной формулы:

$$\Delta_{\theta} = \frac{h(X_b - X_n)}{100} \quad (19),$$

где:  $h$  – класс точности приборов,  $X_b$  – показание верхней границы прибора и  $X_n$  – нижней.

В случае если  $X_n=0$ , а  $X_b=100$ , то  $\Delta_{\theta} = h$ , то есть класс точности прибора соответствует систематической погрешности в процентах.

Для косвенных измерений общая доверительная вероятность будет состоять из ряда частных:

$$\Delta_{\theta} = k \sqrt{\sum_{i=1}^m \Delta_{\theta}^2} \quad (20),$$

где:  $k$  – коэффициент, определяемый принятой доверительной вероятностью. Коэффициент  $k$  принимают равным 1,1 при доверительной вероятности  $P=0.95$  [9, 10].

В случае если отношение величин систематической погрешности к случайной  $\Delta_{\theta}/m < 0,8$ , то не исключенными систематическими погрешностями, по сравнению со случайными, пренебрегают и принимают, что граница погрешности результата равна случайной погрешности  $\Delta = \Delta_{\varepsilon}$ . Если же  $\Delta_{\theta}/m > 0,8$ , то случайной погрешностью, по сравнению с систематическими, пренебрегают и принимают, что граница погрешности результата равна систематической погрешности  $\Delta = \Delta_{\theta}$ .



При проведении измерений предпочтительно регистрировать не только полученные абсолютные значения исследуемых параметров  $a_i$ , но и в отдельных случаях переводить их в относительные величины. Относительные показатели дают возможность сопоставлять измеренные величины, имеющие различные размерности, как это показано в нижеприведенных формулах, и переходить к относительной системе координат. Например, подобным образом осуществлена оценка влияния индивидуальных свойств деконсервированных ооцитов и эмбрионов млекопитающих на уровень их сохранности и установлен верхний порог данного показателя [8].

Элементарная логика и практический опыт показывают, что не разумно стремиться к неоправданно большому числу испытаний, если убедительный результат можно получить при минимально допустимом объеме выборки.

Установленная надежность накладывает условия на общую величину погрешности эксперимента, которая не должна превышать требуемый уровень значимости. Общая величина погрешности эксперимента состоит из погрешностей, возникающих при проведении отдельных опытов, которые складываются из ошибок отдельных измерений. Для установления предельно допустимой величины погрешности эксперимента необходимо определить его общий доверительный интервал и, оценив необходимое количество измерений, понизить величину погрешности до требуемого уровня значимости. Выбор величины доверительной вероятности является отправной точкой планирования эксперимента. Установив в ходе поискового опыта основные статистические величины изучаемых явлений, такие как среднее арифметическое и среднеквадратическое отклонение, можно приступить к планированию эксперимента.

### **3.2.2 Определение необходимого объема выборки**

**Определение минимального количества измерений, обеспечивающего достоверный научный результат для одной выборки.** При всем разнообразии методов исследовательской работы задача планирования эксперимента сводится к тому, чтобы при возможно минимальных объемах наблюдений получать достаточно полную информацию об изучаемых объектах. Планирование эксперимента основывается на установленной надежности, которая обеспечивается за счет увеличения количества измерений и уменьшения их погрешностей. Представленные в литературе формулы оценки оптимального числа измерений имеют значительные численные расхождения [6, 7, 11]. Очевидно, что данные расхождения объясняются допущениями, вводимыми авторами в ходе вывода конечных формул. Количественная взаимосвязь представленных параметров выражена в предлагаемых формулах.

Оптимальная схема исследования основана на проведении минимального количества измерений при обязательном условии получения достоверного результата. Необходимая численность выборки -  $n$ , отвечающая точности получения среднего результата, зависит от величины ошибки выборочной средней и выводится из формулы вычисления средней квадратической ошибки среднего арифметического  $m$  формула (6) [1, 2, 3].

$$n = \left( \frac{\sigma}{m} \right)^2 \quad (21),$$

где:  $\sigma$  – среднее квадратическое отклонение.

Домножив данное выражение, приведенное в скобках, на среднюю величину –  $M$  и критерий Стьюдента -  $t$ , произвели замену второго сомножителя на  $p^1$ , и третьего сомножителя на коэффициент вариации –  $C_v$ .

$$n = \left( \frac{t}{p} C_v \right)^2 \quad (21a).$$

Для уровня надёжности  $P=0,95$  и величины критерия Стьюдента  $t=2$ , соответствующего выборке ( $n \geq 30$ ), значение формулы упрощается:

$$n_o = 0.16 C_v^2 \quad (21б).$$

Значения численности выборки, основанной на нормальном законе распределения признаков и имеющей количественную градацию в зависимости от коэффициента вариации при уровне надёжности  $P=0,95$ , представлены в таблице 9.

Таблица 9. Значение минимального объема выборки  $n_o$  в зависимости от коэффициента вариации для количественного анализа при  $P=0,95$

$C_v, \%$	5	10	15	20	25	30	40	50	100
$n_o$	4	16	32	64	100	144	256	400	1600

Пример. В случае если величина среднеквадратического отклонения исследуемого нами признака равна пяти ( $\sigma=5$ ), а среднее арифметическое двадцати ( $M=20$ ), коэффициент вариации для данной выборки составляет двадцать пять процентов ( $C_v=(5/20)100=25\%$ ). Тогда минимальное количество измерений, обеспечивающее достоверный результат с уровнем надёжности 0,95, равно сто ( $n_o=0.16 \times 25^2 = 100$ ). Относительная погрешность измерения соответствует поставленному критерию надёжности.

При *альтернативной группировке данных*, основанных на распределении признаков не имеющих количественных градаций, когда численность выборочных групп выражается в долях единицы, планируемый объем наблюдений определяется следующим образом:

$$n_o = 0.16 C_v^2 \quad (21в),$$

где:  $W$ - вероятность, величина, определяемая как отношение числа случаев  $n_l$ , при которых это событие произошло, к общему числу наблюдений  $n$ :

$$W = \frac{n_l}{n}, \text{ и } \sigma = \sqrt{W(1-W)} \quad [3, 7], \text{ тогда:}$$

$$n_o = 1600 \frac{100-W}{W} \quad (21г).$$

Для выборки основанной на *альтернативной группировке данных* признаков, при уровне надёжности 0.95, оптимальное количество измерений представлено в таблице 10.

Таблица 10. Значение минимального объема выборки  $n$  в зависимости от средней вероятности  $W$ , для качественного анализа при  $P=0.95$ .

$W (\%)$	99	95	90	85	80	75	70	60
$n_0$	16	89	178	285	400	533	686	1067

Пример. В случае если вероятность изучаемого нами признака составляет восемьдесят пять процентов ( $W=85\%$ ), то величина среднеквадратического отклонения для данной выборки составит около тридцати шести измеряемых единиц ( $\sigma=35,7$ ). Таким образом, необходимо провести не менее двухсот восьмидесяти пяти измерений, чтобы получить достоверный результат с заданным уровнем надёжности ( $n_0=1600(15/85)=285$ ).

Обобщая полученные результаты, можно сделать следующее заключение: для уменьшения ошибки выборочной средней в  $K$  раз нужно увеличить объем выборки в  $K^2$  раз. Если величина количества измерений, обеспечивающего требуемую достоверность значения, настолько велика, что её невозможно реализовать на практике, то в отдельных случаях следует планировать опыт, основываясь на оценке достоверности различия средних значений контроля и эксперимента.

**2.2.2 Определение минимального количества измерений, обеспечивающего достоверный научный результат при сравнении двух выборок.** Минимальное количество измерений, обеспечивающих достоверный результат, зависит от выбора методов измерения и проведения статистического анализа полученного результата. Если результат планируется сравнивать с данными, полученными в контроле при выполнении данного эксперимента, то необходимое количество измерений ( $n_\Delta$ ) в каждой выборке можно определить из следующей формулы  $\Phi(17)$ :

$$n = 2(t \cdot C_v / \alpha)^2, \quad \alpha = \frac{\Delta M}{M} \cdot 100\%, \quad n_\Delta = 2 (t C_v / \alpha)^2 \quad (22),$$

где:  $C_v$  – коэффициент вариации выборки,  $\alpha$  – относительный показатель расхождения средних величин сравниваемых выборок,  $t$  – критерий Стьюдента  $\{1,00; 1,66; 1,98; 2,63\}$  для уровней надёжности  $P\{0,68; 0,90; 0,95; 0,99\}$  при допустимых величинах погрешности вычисления  $p\{0,32; 0,10; 0,05; 0,01\}$ , соответственно. Данная формула  $\Phi(13)$  позволяет установить минимальное количество измерений по каждой из сравниваемых выборок.

Значимая вариация биологических параметров ( $C_v \geq 15\%$ ) приводит к необходимости проведения большого количества повторов измерений ( $n$ ) с целью получения достоверного результата (рис. 29).

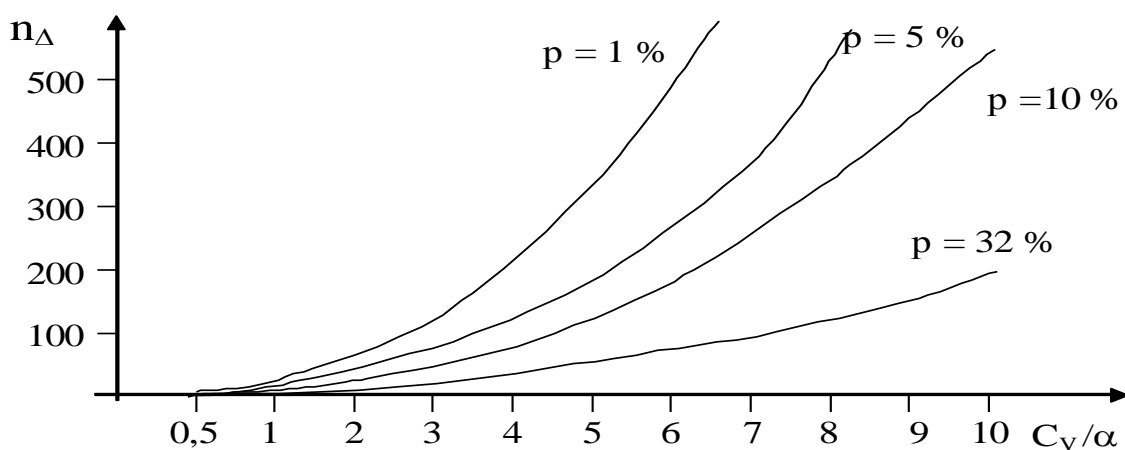


Рис. 29. Зависимость минимального количества измерений, необходимых для получения достоверности различия двух сравниваемых выборок, от соотношения показателей коэффициента вариации ( $C_V$ ) и относительного показателя расхождения средних величин сравниваемых выборок ( $\alpha$ ) при различных допустимых величинах погрешности вычисления ( $p$ ).

Если для количественной оценки исследуемого показателя, как правило, достаточно малой выборки ( $n \leq 30$  при  $C_V \leq 15\%$ ), то для качественной (метод альтернативного варьирования) количество повторов необходимо значительно больше ( $n \geq 100$  при  $25 \leq C_V \leq 300\%$ , для вероятности полученного результата  $W \approx 90(10) \div 50\%$ , соответственно) (рис. 30).

При общепринятом в биологии уровне надежности ( $P \geq 0,95$ ), когда допустимая величина погрешности вычисления не должна превышать пять процентов ( $p < 5\%$ ) формулы упрощаются:

$$n_{\Delta} = 8(C_V/\alpha)^2 \quad (22a), \quad n_o = 0,16C_V^2 \quad (21б).$$

В случае если различие между сравниваемыми выборками составляет больше семи процентов ( $\alpha > 7$ ), то предпочтительно планировать схему проведение опыта на основе различия контроль-опыт, что обеспечивает сокращение количества измерений ( $n_{\Delta} < n_o$ ). Данное заключение получено при решении системы состоящей из двух выражений (22а и 21б).

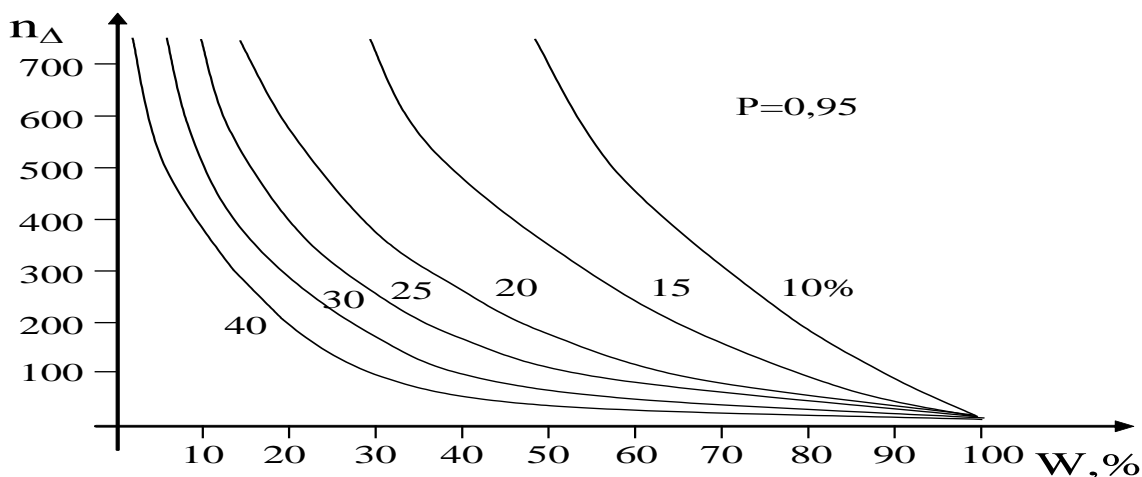


Рис. 30. Зависимость минимального количества измерений, необходимых для получения достоверности различия двух сравниваемых выборок, от относительного показателя расхождения средних величин сравниваемых выборок ( $\alpha$ ) при различной вероятности реализуемого события ( $W$ ) (метод альтернативного варьирования при уровне надежности 0,95).

Поскольку среднееквадратическое отклонение ( $\sigma$ ) при использовании метода альтернативного варьирования зависит от средней величины ( $W$ ), то и необходимое количество измерений связано с данным параметром, который может изменяться от 0 до 100 %. Данная зависимость представлена на рис 31.

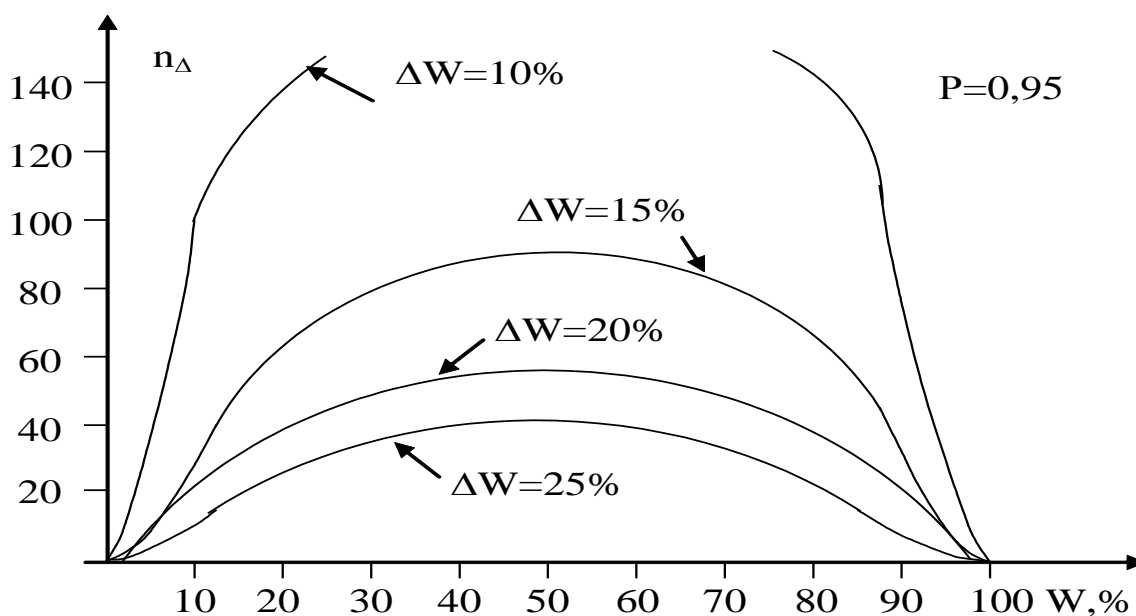


Рис. 31. Зависимость минимального количества измерений, необходимых для получения достоверности различия двух сравниваемых выборок, от показателя расхождения средних величин сравниваемых выборок ( $\Delta W$ ) при различной вероятности реализуемого события ( $W$ ) (метод альтернативного варьирования при уровне надежности 0,95).

### 3.2.3 Формулы приближенной оценки статистических показателей

**Исключение грубых ошибок.** Один из критериев исключения промахов основан на правиле трех сигм. Вероятность того, что ошибка измерения по абсолютной величине выйдет за пределы трех сигм, довольно мала. Поэтому величину  $3\sigma$  условно можно считать предельной границей ошибок измерений. Рассчитывают величину  $3\sigma$  и сравнивают ее с разностью между резко отклоняющейся вариантой и средней величиной  $\bar{x}$  для данного ряда. Если  $x_{\max} - \bar{x}$  или  $\bar{x} - x_{\min}$  больше  $3\sigma$ , то такой результат считают грубой ошибкой и его исключают из выборки:

$$|\bar{x} - x_i| \leq 3\sigma \quad (23).$$

**Определение величины среднеквадратического отклонения.** Рекомендуется следующая формула приблизительного расчета стандартного отклонения (Ашмарин и др., 1975):

$$S = \frac{x_{\max} - x_{\min}}{d} \quad (24),$$

где  $x_{\max}$  и  $x_{\min}$  – минимальное и максимальное значение выборки, величина  $d$  взята из таблицы 11 (против соответствующего объема выборки,  $n$ ).

Таблица 11. Зависимость коэффициента –  $d$  от количества измерений –  $n$  ( $P \geq 0,68$ ).

$n$	$d$	$n$	$d$	$n$	$d$	$n$	$d$
2	1.128	7	2.704	12	3.258	17	3.588
3	1.693	8	2.847	13	3.336	18	3.640
4	2.059	9	2.970	14	3.407	19	3.689
5	2.326	10	3.079	15	3.472	20	3.735
6	2.534	11	3.173	16	3.532	более	4

Выборочное стандартное отклонение веса тела бурозубок ( $n = 63$ ), рассчитанное по приведенной формуле, составляет:  $S = (11.9 - 7.3) / 4 = 1.15$  г, что достаточно близко к точному значению,  $S = 0.89$  г.

Использование экспресс-оценок стандартного отклонения значительно сокращает время расчетов, существенно не сказываясь на их точности при осуществлении планирования эксперимента. Отмечается лишь небольшая тенденция к завышению получаемых этим методом значений стандартного отклонения при небольших объемах выборок.

### **Определение минимального количества измерений, необходимое для приближенной оценки основных статистических показателей**

1. Для количественного метода оценки исследуемых показателей, при проведении поисковых опытов, принято использовать малые выборки ( $n < 30$ ). Таким образом, основные статистические параметры можно выразить следующим образом:

а) среднее значение выборки  $\bar{x} \approx 0,5(x_{\max} + x_{\min})$ ;

б) приближенная величина среднеквадратического отклонения определяется из  $\Phi(15)$  и таблицы (7) для  $n=4 \pm 1$ , если  $\Delta x$  – разность между максимальным и минимальным значениями данной выборки  $\Delta x \approx x_{\max} - x_{\min}$ :

$$\sigma \approx 0,5 \cdot \Delta x;$$

в) коэффициент вариации –  $C_v = 100\sigma / \bar{x}$  из  $\Phi(5)$ , следовательно:

$$C_v \approx 50\Delta x / \bar{x};$$

г) Уровень значимости оценки (относительная ошибка выборки –  $\rho$ ) для уровня надежности  $P = 95\%$ , соответственно для  $n=4 \pm 1$  составляет величину  $\rho \approx C_v / 0,3$ .

д) Минимальное количество измерений, необходимое для получения достоверного результата в одной выборке –  $n_0$  и в двух сравниваемых выборках –  $n_\Delta$ , при относительном показателе расхождения средних величин сравниваемых выборок –  $\alpha = 10\%$ :

$$n_0 \approx 0,5 \cdot C_v^2, \quad n_\Delta \approx 0,1 \cdot C_v^2, \quad \alpha = 100 \cdot \Delta \bar{x} / \bar{x}.$$

2. Для качественного метода оценки исследуемых показателей, при проведении поисковых опытов, применимы большие выборки ( $n \geq 30$ ), из табл. (12):

а)  $\sigma \approx 0,25 \cdot \Delta x$ , где  $\Delta x \approx x_{\max} - x_{\min}$ ; б)  $C_v \approx 25 \Delta x / \bar{x}$ , где  $\bar{x} \approx 0,5(x_{\max} + x_{\min})$ ; в)  $\rho \approx 0,5 C_v$ ; г)  $n_0 \approx 0,2 \cdot C_v^2$ ; д)  $n_\Delta \approx 0,04 \cdot C_v^2$ , при  $\alpha = 10\%$ .

Таблица 12. Значения t-критерия Стьюдента

t(n)	Количество измерений – n					
	3	7	10	20	30	100
t	3,18	2,36	2,23	2,09	2,04	1,98
t/√n	1,84	0,89	0,71	0,47	0,371	0,20
t/5	0,64	0,47	0,45	0,42	0,41	0,4
t/10	0,32	0,24	0,22	0,21	0,2	0,2

Полученные формулы можно структурировать в виде таблицы 13.

Таблица 13. Способы приближенной оценки основных статистических показателей

Статистический анализ	$n_{\min}$	$\bar{x}$	$\Delta x$	$\sigma$	$C_v$	$\rho$	$n_0$ ( $n_\Delta, \alpha=5$ )	$n_\Delta$ , $\alpha=10$
Количественный	$4 \pm 1$			$\frac{\Delta x}{2,5}$	$\frac{\Delta x}{\bar{x}} \cdot 40$	$\frac{C_v}{0,3}$	$\frac{C_v^2}{2,4}$	$\frac{C_v^2}{9,8}$
Полуколичественный	$10 \pm 5$	$\frac{x_{\max} + x_{\min}}{2}$	$x_{\max} - x_{\min}$	$\frac{\Delta x}{3}$	$\frac{\Delta x}{\bar{x}} \cdot 30$	$\frac{C_v}{2}$	$\frac{C_v^2}{4,9}$	$\frac{C_v^2}{22}$
Качественный	$20 \pm 5$	$\frac{n_1}{n} \cdot 100\%$	$100 - \bar{x}$	$\sqrt{\bar{x} \cdot \Delta x}$	$\frac{\sigma}{\bar{x}} \cdot 100$	$\frac{C_v}{2,1}$	$\frac{C_v^2}{5,7}$	$\frac{C_v^2}{23}$

$n_{\min}$  - минимальное количество измерений, необходимое для приближенной оценки основных статистических показателей,  $n_0$  – количество измерений обеспечивающее достоверное значение в одной выборке и  $n_\Delta$  – двух сравниваемых,  $n_1$  – количество измерений с заданным признаком.

В случае если качественную оценку биообъекта можно разбить более чем на две группы и количественно определить их средние показатели, то это дает возможность применить количественный метод статистического анализа, а это в свою очередь позволяет значительно снизить объем выборки. Например, использование количественного метода статистической обработки для оценки жизнеспособности эмбрионов млекопитающих существует возможность сократить количество измерений более чем в десять раз, по сравнению с качественным методом, за счет уменьшения величины ошибки среднеквадратического отклонения при условии получения достоверного результата (табл. 14).

Табл. 14. Способы оценки технологии криоконсервации эмбрионов крупного рогатого скота при помощи различных методов статистического анализа.

N	Показатели состояния биообъекта	Качественный		Количественный			
		Сохранность $C_n=k/n$ , %		Жизнеспособность $S_n = \bar{S}_i$ , %		Эффективность $W_{n+1}=S_{n+1}/S_n$ , %	
		$n_o$	$M \pm m$ для $n=10$	$n_o$	$M \pm m$ для $n=10$	$n_o$	$M \pm m$ для $n=10$
1	Свежеполученного	3	$100 \pm 0$	14	$90 \pm 1,7$	–	–
2	Деконсервированного	178	$90 \pm 9,5$	15	$88,5 \pm 1,8$	16	$98,3 \pm 1,5$
3	Культивированного	400	$80 \pm 12,7$	15	$86 \pm 1,8$	17	$96,6 \pm 1,6$

$n_o$  – количество измерений, необходимое для получения достоверного среднего значения;  $n$  – количество эмбрионов в данном опыте,  $k$  – количество эмбрионов, пригодных к пересадке,  $M$  – средняя величина показателя;  $m$  – ошибка средней величины,  $S_i$  – жизнеспособность эмбрионов различного качества: отличного  $95 \pm 5$ , хорошего  $85 \pm 5$ , удовлетворительного  $70 \pm 5$ .

В случае проведения многофакторного исследования, когда в одном опыте изменяется более одного параметра, принято использовать специальные методы планирования, основанные на применении математического моделирования.

### 3.3 Основы планирования многофакторного эксперимента

Важнейшей задачей методов обработки полученной в ходе эксперимента информации является задача построения *математической модели* изучаемого явления, процесса, объекта. Ее можно использовать и при *анализе* процессов и при *проектировании* объектов. Можно получить хорошо аппроксимирующую математическую модель, если целенаправленно применяется активный эксперимент. Другой задачей обработки полученной в ходе эксперимента информации является задача *оптимизации*, т.е. нахождения такой комбинации влияющих независимых переменных, при которой выбранный показатель оптимальности принимает экстремальное значение.



Опыт – это отдельная экспериментальная часть. План эксперимента – совокупность данных определяющих число, условия и порядок проведения опытов. Планирование эксперимента – выбор плана эксперимента, удовлетворяющего заданным требованиям, совокупность действий направленных на разработку стратегии экспериментирования (от получения априорной информации до получения работоспособной математической модели или определения оптимальных условий). Это целенаправленное *управление экспериментом*, реализуемое в условиях *неполного знания* механизма изучаемого явления.

В процессе измерений, последующей обработки данных, а также формализации результатов в виде математической модели, возникают погрешности и теряется часть информации, содержащейся в исходных данных. Применение методов планирования эксперимента позволяет определить погрешность математической модели и судить о ее адекватности. Если точность модели оказывается недостаточной, то применение методов планирования эксперимента позволяет модернизировать математическую модель с проведением дополнительных опытов без потери предыдущей информации и с минимальными затратами.

Цель планирования эксперимента – нахождение таких условий и правил проведения опытов при которых удастся получить надежную и достоверную информацию об объекте с наименьшей затратой труда, а также представить эту информацию в компактной и удобной форме с количественной оценкой точности.

Пусть интересующее нас свойство ( $Y$ ) объекта зависит от нескольких ( $n$ ) независимых переменных ( $X_1, X_2, \dots, X_n$ ) и мы хотим выяснить характер этой зависимости -  $Y=F(X_1, X_2, \dots, X_n)$ , о которой мы имеем лишь общее представление. Величина  $Y$  – называется “отклик”, а сама зависимость  $Y=F(X_1, X_2, \dots, X_n)$  – “функция отклика”.

Отклик должен быть определен количественно. Однако могут встречаться и качественные признаки  $Y$ . В этом случае возможно применение рангового подхода. Пример рангового подхода - оценка на экзамене, когда одним числом оценивается сложный комплекс полученных сведений о знаниях студента.

Независимые переменные  $X_1, X_2, \dots, X_n$  – иначе факторы, также должны иметь количественную оценку. Если используются качественные факторы, то каждому их уровню должно быть присвоено какое-либо число. Важно выбирать в качестве факторов лишь независимые переменные, т.е. только те которые *можно изменять, не затрагивая другие факторы*. Факторы должны быть однозначными. Для построения эффективной математической модели целесообразно провести предварительный анализ значимости факторов (степени влияния на функцию), их ранжирование и исключить малозначащие факторы.

Диапазоны изменения факторов задают область определения  $Y$ . Если принять, что каждому фактору соответствует координатная ось, то полученное пространство называется факторным пространством. При  $n=2$  область определения  $Y$  представляется собой прямоугольник, при  $n=3$  – куб, при  $n > 3$  - гиперкуб.

При выборе диапазонов изменения факторов нужно учитывать их совместимость, т.е. контролировать, чтобы в этих диапазонах любые сочетания фак-

торов были бы реализуемы в опытах и не приводили бы к абсурду. Для каждого из факторов указывают граничные значения

$$X_{i\min} \leq X_i \leq X_{i\max}, i=1, \dots, n.$$

Регрессионный анализ функции отклика предназначен для получения ее математической модели в виде уравнения регрессии

$$Y = F(X_1, X_2, \dots, X_n; B_1, B_2, \dots, B_m) + e,$$

где  $B_1, \dots, B_m$  – некоторые коэффициенты;  $e$  – погрешность.

Среди основных методов планирования, применяемых на разных этапах исследования, используют:

- планирование отсеивающего эксперимента, основное значение которого – выделение из всей совокупности факторов группы существенных факторов, подлежащих дальнейшему детальному изучению;
- планирование эксперимента для дисперсионного анализа, т.е. составление планов для объектов с качественными факторами;
- планирование регрессионного эксперимента, позволяющего получать регрессионные модели (полиномиальные и иные);
- планирование экстремального эксперимента, в котором главная задача – экспериментальная оптимизация объекта исследования;
- планирование при изучении динамических процессов и т.д.

Инициатором применения планирования эксперимента является Рональд А. Фишер, другой автор известных первых работ – Френк Йетс. Далее идеи планирования эксперимента формировались в трудах Дж. Бокса, Дж. Кифера. В нашей стране – в трудах Г.К. Круга, Е.В. Маркова и др.

В настоящее время методы планирования эксперимента заложены в специализированных пакетах, широко представленных на рынке программных продуктов, например: StatGrafics, Statistica, SPSS, SYSTAT и др.

**Представление результатов экспериментов.** При использовании методов планирования эксперимента необходимо найти ответы на 4 вопроса:

- Какие сочетания факторов и сколько таких сочетаний необходимо взять для определения функции отклика?
- Как найти коэффициенты регрессии  $B_0, B_1, \dots, B_m$ ?
- Как оценить точность представления функции отклика?
- Как использовать полученное представление для поиска оптимальных значений  $Y$ ?

Геометрическое представление функции отклика в факторном пространстве  $X_1, X_2, \dots, X_n$  называется поверхностью отклика. Если исследуется влияние на  $Y$  лишь одного фактора  $X_1$ , то нахождение функции отклика – достаточно простая задача (рис. 32).

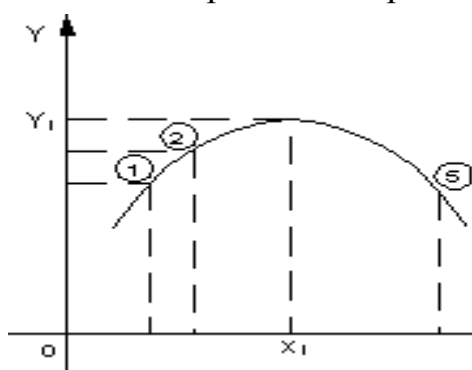
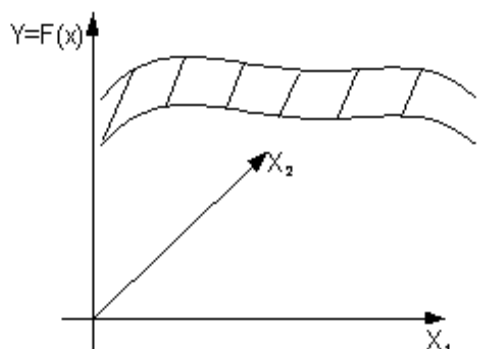


Рис. 32. Поверхность отклика одной переменной.

Задавшись несколькими значениями этого фактора, в результате опытов получаем соответствующие значения  $Y$  и график  $Y = F(X)$ . По его виду можно подобрать математическое выражение функции отклика. Если мы не уверены, что опыты хорошо воспроизводятся, то обычно опыты повторяют несколько раз и получают зависимость с учетом разброса опытных данных.



Если факторов два, то необходимо провести опыты при разных соотношениях этих факторов (рис. 33).

Рис. 33. Построение функции отклика двух переменных по опытным данным

Полученную функцию отклика в  $3^x$ -мерном пространстве (рис. 2) можно анализировать, проводя ряд сечений с фиксированными значениями одного из факторов (рис. 34). Вычлененные графики сечений можно аппроксимировать совокупностью математических выражений.

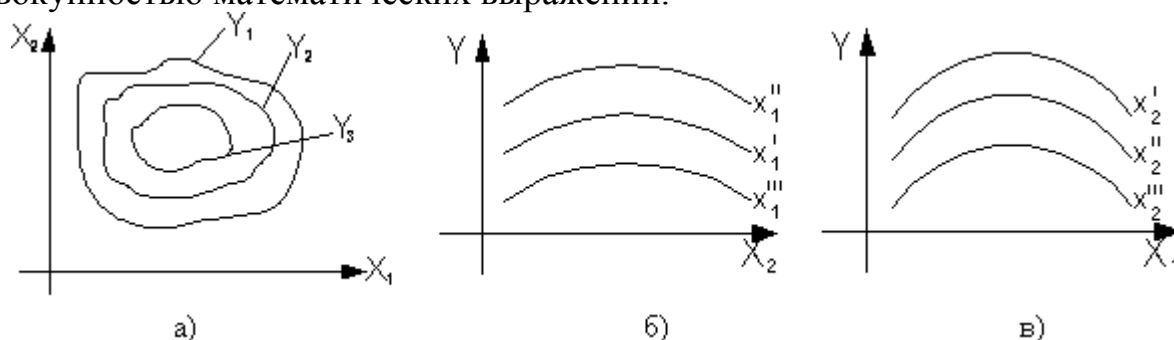
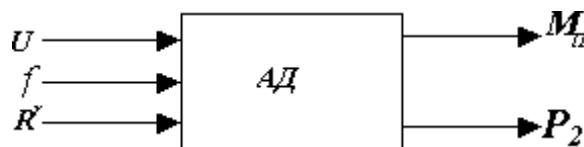


Рис. 34. Сечения поверхности отклика при фиксированных откликах (а) и переменных (б, в).

При трех и более факторах задача становится практически неразрешимой. Если и будут найдены решения, то использовать совокупность выражений достаточно трудно, а часто и не реально. Например, пусть необходимо исследовать влияние  $U, f$  и  $R^r$  на  $M_n$  и  $P_2$  асинхронного двигателя (АД) (рис. 35).

Рис. 35. Исследование влияния  $U, f$  и  $R^r$  на  $M_n$  и  $P_2$  АД



Если в диапазоне изменения каждого фактора взять хотя бы по пять точек (табл. 15), то для того чтобы выполнить опыты при всех возможных сочетаниях значений факторов (их три) необходимо выполнить  $5^3=125$  опытов и сформировать по  $5^2=25$  кривых для каждой из двух функций отклика. Если мы хотим хотя бы продублировать опыты чтобы снизить погрешность, то число опытов пропорционально возрастает, поэтому произвольное выполнение опытов при числе факторов более двух и использование их результатов - практически нереально.

Таблица 15. Вариация параметров определяющих мощность работы асинхронного двигателя

Исследуемые параметры	Численные значения, измеренные в опыте					
	1	2	3	4	5	6
Напряжение - U, В	170	180	190	200	210	220
Частота - f, Гц	40	45	50	55	60	65
Сопротивление - $R^r$ , Ом	0,1	0,2	0,3	0,4	0,5	0,6

В общем случае объект исследования можно представить в виде структурной схемы, показанной на рис.36.

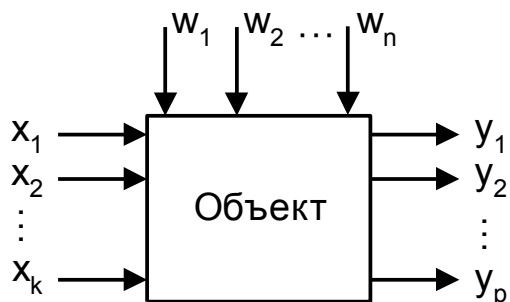


Рис. 36. Представление объекта в виде такой схемы основано на принципе «черного ящика».

Имеем следующие группы параметров: управляющие (входные)  $x_i$ , которые называются *факторами*; выходные параметры  $y_i$ , которые называются *параметрами состояния*;  $w_i$  - *возмущающие воздействия*. Предполагается, что возмущающие воздействия не поддаются контролю и либо являются случайными, либо меняются во времени. Каждый фактор  $x_i$  имеет область определения, которая должна быть установлена до проведения эксперимента. Комбинацию факторов можно представить как точку в многомерном пространстве, характеризующую состояние системы.

На практике целью многофакторного эксперимента является установление зависимости  $y = f(x_1, x_2, \dots, x_k)$ , описывающей поведение объекта. Чаще всего функция строится в виде линейной зависимости

$$y = a_0 + a_1 x_1 + a_2 x_2 \quad (25)$$

или

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_{11} x_1^2 + a_{22} x_2^2 + a_{12} x_1 x_2 \quad (26)$$

Целью эксперимента может быть, например, построение зависимости (6.2.1) при минимальном количестве измерений значений управляющих параметров  $x_i$ .

На первом этапе планирования эксперимента необходимо выбрать область определения факторов  $x_i$ . Выбор этой области производится исходя из априорной информации. Значения  $x_i$  называются *уровнями управляющего параметра*.

Если выбрана линейная модель (6.2.2), то для построения аппроксимирующей функции достаточно выбрать *основной уровень* и *интервал варьирования* управляющего параметра  $x_i$ .

Для линейной модели интервал варьирования можно определить как,

$$I = \frac{x_{\max} - x_{\min}}{2} \quad (27),$$

а основной (нулевой) уровень - как среднее значение

$$x_0 = \frac{x_{\max} + x_{\min}}{2} \quad (28)$$

Для упрощения планирования эксперимента принято вместо реальных (натуральных) уровней  $x_i$  использовать *кодированные* значения факторов. Для факторов с непрерывной областью определения это можно сделать при помощи следующего преобразования

$$x_i = \frac{\bar{x}_i - x_{io}}{I_i} \quad (29),$$

где  $\bar{x}_i$  - натуральное значение фактора;  $I_i$  - интервал варьирования;  $x_{io}$  - основной уровень;  $x_i$  - кодированное значение.

В результате  $x_i$  принимает значения на границах  $x_i = \pm 1$ , на основном уровне  $x_i = 0$ . Основная проблема состоит в выборе области варьирования, поскольку эта задача является неформализованной.

Рассмотрим *полный* факторный эксперимент на примере линейной модели (6.2.2). Если число факторов  $k$ , то для проведения полного факторного эксперимента нужно  $N = 2^k$  опытов, где 2 - число уровней, которого достаточно для построения линейной модели.

Условие проведения этого эксперимента можно зафиксировать в матрице планирования (табл. 16).

Таблица 16

Номер опыта	$x_1$	$x_2$	$y$
1	-1	-1	$y_1$
2	+1	-1	$y_2$
3	-1	+1	$y_3$
4	+1	+1	$y_4$

Таким образом, для двух факторов построение матрицы планирования элементарно. Для большего числа факторов необходимо разработать правила построения таких матриц. Например, при появлении фактора  $x_3$  в табл.16 произойдут следующие изменения (табл. 17):

Таблица 17

Номер опыта	$x_1$	$x_2$	$x_3$	$y$
1	-1	-1	+1	$y_1$
2	+1	-1	+1	$y_2$
3	-1	+1	+1	$y_3$
4	+1	+1	+1	$y_4$
5	-1	-1	-1	$y_5$
6	+1	-1	-1	$y_6$
7	-1	+1	-1	$y_7$
8	+1	+1	-1	$y_8$

При появлении нового столбца каждая комбинация уровней исходной таблицы проявится дважды. Это не единственный способ расширения матрицы планирования. Используют также перемножение столбцов, правило чередования знаков.

Очень важны общие свойства матрицы планирования:

1) симметричность матрицы относительно центра эксперимента:

$$x_i = 0. \text{ Тогда } \sum_{j=1}^N x_{ji} = 0.$$

2) условие нормировки  $\sum_{i=1}^N x_{ij}^2 = N$ , то есть сумма квадратов элементов каждого столбца равна числу опытов.

Первые два свойства относятся к построению отдельных столбцов матрицы

3) совокупность столбцов имеет следующее свойство  $\sum_{i=1}^N x_{ij} x_{in} = 0$ , где  $j \neq n$ .

4) *Ротатабельность*. Это означает, что точки (значения факторов) в матрице планирования подбираются так, что точность предсказания выходного параметра должна быть одинакова на равных расстояниях от центра эксперимента (нулевого уровня) и не зависеть от направления.

### Планирование экспериментов при помощи существующего программного обеспечения *STATISTICA*

Модуль *Планирование экспериментов STATISTICA DOE* предлагает исчерпывающий набор процедур для построения и анализа экспериментальных планов, используемых в промышленных исследованиях. Вниманию пользователя предлагаются  $2^{k-p}$  факторные планы с блоками (для планов, содержащих более 100 факторов, имеются высокоэффективные алгоритмы для нахождения планов с минимальной абберацией и максимально несмешанных планов, в которых пользователь может задавать эффекты взаимодействия, которые должны быть несмешанными), отсеивающие планы (для более 100 факторов предусмотрены планы Плакетта-Бермана),  $3^{k-p}$  факторные планы с блоками (в том

числе планы Бокса-Бенкена), смешанные планы, (малые) центральные композиционные планы (или поверхности отклика), планы на Латинских квадратах, робастные планы Тагучи и ортогональные массивы, планы для смесей и тернарных поверхностей, вершины и центроиды для поверхностей и смесей с ограничениями, *D*- и *A*-оптимальные планы для факторных планов, поверхностей и смесей. Ниже представлено подробное описание доступных планов и методов их построения и анализа.

Таким образом, *STATISTICA* незаменима в сложных исследованиях, где требуется сочетание творческих и новаторских путей, в тех случаях, когда зависимые исследуемые переменные являются категориальными, или когда эффект предсказанных переменных не является линейным.

Убедившись в возможности реализации необходимого количества измерений в данном планируемом эксперименте, в ходе выполнения 1÷3 этапов построения его модели (табл. 1), можно приступать к проведению основных опытов. Моделирование основного эксперимента в ходе проведения поисковых опытов позволяет исследователю получить при завершении работы достоверный научный результат.

#### ВОПРОСЫ ДЛЯ САМОКОНТРОЛЯ.

1. Какова цель проведения поисковых опытов?
2. Из каких соображений устанавливается доверительная вероятность?
3. Какие факторы влияют на выбор уровня надежности?
4. От чего зависит количество значимых цифр в числе, с которым производятся требуемые операции?
5. Как определить доверительный интервал для случайных и систематических погрешностей?
6. Какие существуют особенности соотношения доверительных интервалов для случайных и систематических погрешностей?
7. Из каких составляющих состоит общая величина погрешности эксперимента?
8. Для какой цели производится оценка минимального количества измерений для одной выборки?
9. В чем заключается основное отличие объема выборок полученных при качественном и количественном методах проведения статистического анализа?
10. Какова необходимость оценки минимального количества измерений для двух сравниваемых выборок?
11. Напишите формулы приблизительной и точной оценки величины среднего арифметического и среднего квадратического отклонения.
12. Обоснуйте необходимость проведения 1÷3 этапов моделирования эксперимента.

## Глава 4. Анализ экспериментальных данных

При изучении и анализе сложных и многообразных причинно-следственных отношений между объектами и явлениями биотехнологу приходится учитывать целый комплекс внешних и внутренних факторов, от которых в конечном итоге зависят уровень и ход наблюдаемых процессов, те или иные биологические свойства живых организмов, их динамика и разнообразие. При этом зачастую важно оценивать не только роль одного из многочисленных внешних факторов, но и их взаимодействие при совместном влиянии на технологию или организм.

Идейная база для изучения действия факторов содержится уже в методе сравнения двух выборок. Биотехнологическим содержанием операции сравнения двух выборок, в конце концов, выступает поиск факторов, ответственных за смещение средних арифметических или усиление изменчивости признаков. Развивая это направление биометрического исследования, можно не ограничиваться только двумя «дозами» фактора, но изучить серию ситуаций, в которых фактор проявлял разную силу действия на результативный признак – от самого слабого до самого сильного. При этом каждому уровню фактора будет соответствовать отдельная выборка, и общая задача получит формулировку «сравнить несколько выборок». В терминах факториальной биометрии вопрос о влиянии фактора на признак звучит так: сказывается ли отличие условий получения разных выборок на качестве (значениях) вариант? В терминах статистики вопрос звучит несколько иначе: из одной ли генеральной совокупности отобраны все выборки, оценивают ли выборочные средние арифметические одну и ту же генеральную среднюю? Вариантов ответа может быть только два:

Все выборки отобраны из одной генеральной совокупности, условия возникновения вариант одни и те же.

Выборки отобраны из разных генеральных совокупностей, условия возникновения вариант выборок различаются.

В постановке вопроса можно уловить противоречие. Выше было сказано, что по условию задачи выборки формировались в разных условиях, и тут же предполагается, что условия были одинаковые. На самом деле противоречия нет, поскольку речь идет об определении чувствительности признака к действию фактора. Условия формирования выборок могут отличаться, но они могут никак и не сказаться на величине изучаемого признака, не отразиться на значениях вариант. Смысл статистического сравнения в том и состоит, чтобы оценить эффективность действия фактора на признак, доказать реальность реакции вариант выборок на разные условия их формирования. В сферу исследования можно вовлекать как один, так и два признака, как количественные, так и качественные характеристики. В каждом случае процедура анализа несколько отличается.

### 4.1 Изучение сходства (различий) двух выборок

В любых биологических экспериментах и наблюдениях особое значение имеют различия, на основании которых судят об эффективности действия тех или иных факторов, например, по разности между опытной и контрольной груп-



пами делают заключение о результатах опыта. При этом особенно важно оценить статистическую *достоверность разности*, т. е. определить, можно ли данное различие считать закономерным, *характерным для всей генеральной совокупности* и рассматривать его как результат действия особенных факторов, или же оно случайно и является следствием недостаточного количества данных и в следующих опытах может не проявиться.

#### 4.1.1. Статистическая проверка гипотез

Логика проверки гипотез в математической статистике (она напоминает логику доказательства от противного) состоит в следующем. Вначале предполагается, что проверяемая гипотеза (ее принято называть *нулевой гипотезой* и обозначать  $H_0$ ) верна. В предположении, что  $H_0$  верна, ищется распределение вероятностей некоторой функции  $g(x_1, x_2, \dots, x_n)$  от значений выборки, называемой *статистикой критерия* (правило проверки гипотезы принято называть *критерием*), и в области значений этой статистики выделяется некоторая область, называемая *критической областью*, такая, что вероятность попадания выборочного значения статистики  $g$  в эту область не превосходит заданного малого значения  $p$ , называемого *уровнем значимости критерия* (обычно полагают  $p$  равным 0,05 или 0,01). Если для данной конкретной выборки  $g$  попадает в критическую область, то гипотеза  $H_0$  *отвергается* (говорят – «отвергается на уровне значимости  $p$ »), поскольку вероятность этого события, если в действительности верна  $H_0$ , мала. Если же  $g$  не попадает в критическую область  $W$ , то говорят, что «гипотеза  $H_0$  *не отвергается* на уровне значимости  $p$ » (или – «полученные данные не дают оснований отвергнуть гипотезу  $H_0$  на уровне значимости  $p$ »).

Очевидно, однако, что можно разными способами задать статистику критерия  $g(x_1, x_2, \dots, x_n)$ , а для заданной статистики можно разными способами выбрать критическую область. Поэтому следует выбирать  $g$  и критическую область в некотором смысле наилучшими из возможных, а именно такими, чтобы полученный критерий был *наиболее мощным*.

Для определения понятия *мощности критерия* введем понятие *альтернативной гипотезы*  $H_1$ , т.е. гипотезы, которая выполняется, если не выполняется нулевая гипотеза  $H_0$ . Тогда в терминах правильности или ошибочности принятия  $H_0$  и  $H_1$  можно указать четыре потенциально возможных результата применения критерия к выборке, как следующие

		Принята гипотеза	
		$H_0$	$H_1$
Верна гипотеза	$H_0$	вероятность правильно принять $H_0$ , когда верна $H_0$ (чувствительность критерия)	вероятность ошибочно принять $H_1$ , когда верна $H_0$ (ошибка 1-го рода, уровень значимости)
	$H_1$	вероятность ошибочно принять $H_0$ , когда верна $H_1$ (ошибка 2-го рода)	- вероятность правильно принять $H_1$ , когда верна $H_1$ (мощность критерия)

Как мы видим, мощность критерия - это вероятность принятия при применении данного критерия альтернативной гипотезы  $H_1$  при условии, что она верна. Очевидно, что при фиксированной ошибке 1-го рода (ее мы задаем сами, и она не зависит от свойств критерия) критерий будет тем лучше, чем больше его мощность (т.е. чем меньше ошибка 2-го рода).

Проверка гипотезы о равенстве заданному числу математического ожидания нормально распределенной случайной величины с известной дисперсией. Итак, пусть имеется выборка  $x_1, x_2, \dots, x_n$  значений случайной величины  $\xi$  объема  $n$ . Предположим, что  $H_0$  верна и мы выбираем в качестве статистики критерия стандартизованное выборочное среднее

$$t = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

где:  $\bar{x}$  – средняя величина определенная в опытной группе и  $\mu_0$  – истинное значение (контроль),  $\sigma$  – среднеквадратического отклонение в опыте.

При верной  $H_0$  статистика  $t$  имеет стандартное нормальное распределение представленное на рис. 37. На этом рисунке также указана критическая область уровня  $p=0,05$ , вероятность попадания в каждую из которых статистики  $t$  равна  $p/2=0,025$ .

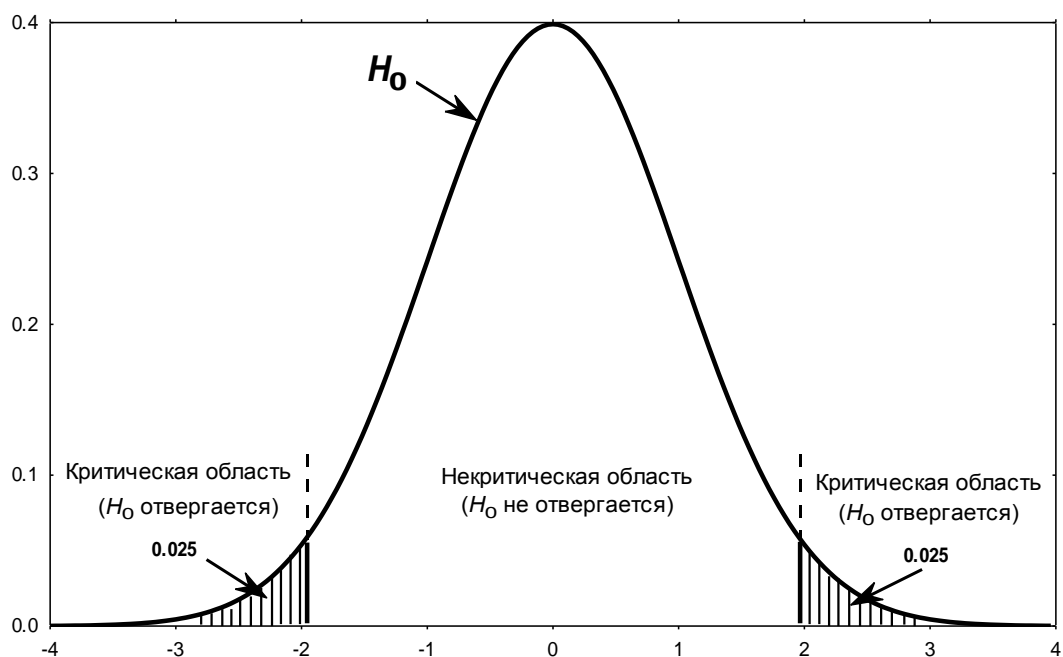


Рис. 37. Пример критической области.

Имеются формализованные подходы к выбору статистик критериев и построению критических областей, приводящие к наиболее мощным критериям, но мы их здесь не рассматриваем. Неформальное же правило состоит в том, чтобы выбирать в качестве статистики величину, характеризующую степень отклонения от нулевой гипотезы. Очевидно, разность  $\bar{x} - \mu_0$  удовлетворяет этому условию, а деление на константу  $\sigma / \sqrt{n}$ , сохраняя это качество, приводит к величине  $t$  с полностью заданным распределением, что позволяет выбрать критическую область с требуемым уровнем значимости. Неформальное же правило выбора критической области состоит в том, чтобы она включала значения статистики, соответствующие наибольшим отклонениям от нулевой гипотезы - на рис. 37 эта рекомендация соблюдена.

До сих пор мы говорили о свойствах критерия в предположении, что верна гипотеза  $H_0$ . А что происходит, когда верна альтернативная гипотеза  $H_1$ ? В этом случае распределение статистики критерия  $t$  изменится. Чтобы его найти, произведем преобразование

$$u = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{\bar{x} - \mu_0 - \mu_1 + \mu_1}{\sigma / \sqrt{n}} = \frac{\bar{x} - \mu_1}{\sigma / \sqrt{n}} + \frac{\mu_1 - \mu_0}{\sigma / \sqrt{n}},$$

из которого следует, что при гипотезе  $H_1$  распределение статистики  $t$  отличается от стандартного нормального сдвигом на величину  $\Delta = \frac{\mu_1 - \mu_0}{\sigma / \sqrt{n}}$ , при выполнении  $H_1$ .

На рис. 38 взаимное расположение плотностей распределения статистики  $t$  при гипотезах  $H_0$  и  $H_1$  показано для случая  $\mu_0 = 0$ ,  $\mu_1 = 2$ ,  $\sigma^2 = 1$ ,  $\alpha = 0,05$  и  $n = 1$ . Вероятности ошибки 2-го рода  $\beta$  соответствует площадь под кривой функции плотности при  $H_1$  на промежутке от  $-1,96$  до  $1,96$ , где не отвергается гипотеза  $H_0$ , а следовательно, ошибочно не принимается гипотеза  $H_1$ . В данном случае ошибка 2-го рода,  $\beta = 0,48$ , довольно велика. Это произошло, главным образом, потому, что мал объем выборки - имеется всего одно наблюдение,  $n = 1$ . При увеличении  $n$  распределение, соответствующее альтернативной гипотезе  $H_1$ , будет сдвигаться вправо, поскольку величина  $\Delta = \frac{\mu_1 - \mu_0}{\sigma / \sqrt{n}}$  будет увеличиваться, что приведет, как легко понять по рис. 39, к уменьшению ошибки  $\beta$ . Очевидно также, что большей величине разности  $\mu_1 - \mu_0$  соответствует большая величина  $\Delta$ , и следовательно, меньшая ошибка 2-го рода. Ошибка 2-го рода уменьшается также при уменьшении дисперсии  $\sigma^2$  наблюдаемой случайной величины. Кроме того,  $\beta$  уменьшается при увеличении  $\alpha$ , однако не принято брать  $\alpha$  больше  $0,05$ . При уменьшении  $\alpha$  ошибка  $\beta$ , напротив, растет, поэтому не следует брать  $\alpha$  слишком малым, если число наблюдений  $n$  мало, разность между  $\mu_1$  и  $\mu_0$  невелика, а дисперсия  $\sigma^2$  - большая.

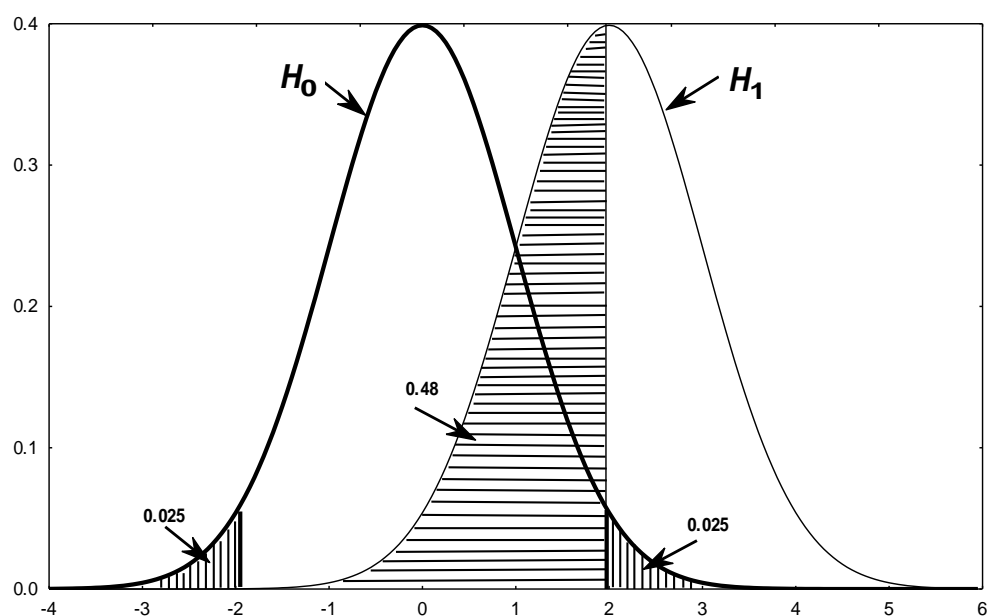


Рис. 38. Взаимосвязь между ошибками 1-го и 2-го рода при двусторонней альтернативе.

Содержательно, ошибка 1-го рода - это ошибка ложного обнаружения несуществующего отклонения от нулевой гипотезы (ложного обнаружения несуществующего эффекта). Ошибка же 2-го рода - это ошибка ложного необнаружения существующего отклонения от нулевой гипотезы (ложного необнаружения существующего эффекта). Мощность критерия - это его способность обнаружить имеющееся отклонение от нулевой гипотезы.

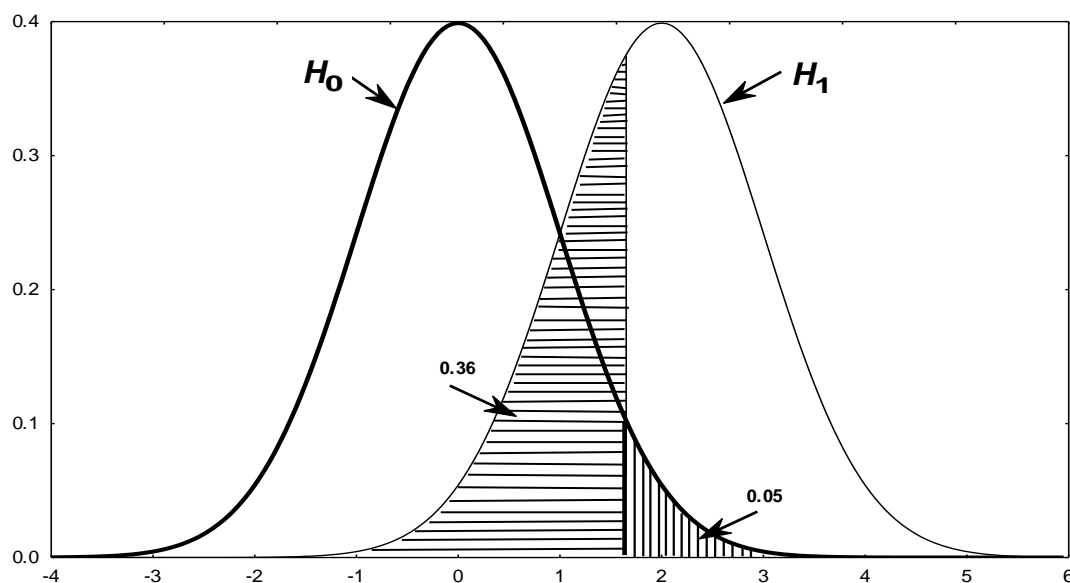


Рис. 39. Взаимосвязь между ошибками 1-го и 2-го рода при односторонней альтернативе.

Пример. Известно, что датчик генерирует случайные числа, нормально распределенные с дисперсией 1, но есть сомнения в том, что математическое ожидание равно 0. Требуется проверить гипотезу о равенстве математического ожидания нулю при двусторонней альтернативе по следующей случайной выборке объема  $n = 25$ :

0,830	0,177	-0,294	0,471	-0,044	0,635	2,209	-0,394	-0,404
1,257	1,137	-0,839	1,668	0,751	0,416	-0,922	1,473	-0,317
0,220	0,414	0,428	1,088	-1,130	-0,015	0,142		

Выборочное среднее равно  $\bar{x} = 0.358$ , следовательно, для статистики критерия получаем

$$u = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{0.358 - 0}{1 / \sqrt{25}} = 1.79$$

Значение 1,79 не выходит за двусторонние 5%-ные критические пределы, поэтому гипотеза не отвергается.

На самом деле математическое ожидание датчика было равным 0,25, т.е. отклонение ошибочно не было обнаружено - при проверке гипотезы была сделана ошибка 2-го рода. Очевидно, мощность критерия при данном числе наблюдений  $n = 25$ , данной разности между гипотетическим и истинным математиче-

скими ожиданиями  $\mu_1 - \mu_0 = 0,25$  и данной дисперсии  $\sigma^2 = 1$  недостаточна. (В другом эксперименте с этим же датчиком была получена выборка значений объема  $n = 100$ . Выборочное среднее оказалось равным  $\bar{x} = 0,218$ , а выборочное значение статистики – равным 2,18, что дало основание отвергнуть нулевую гипотезу.)

#### 4.1.2 Параметрические критерии оценки различия двух выборок

Обнаружение достоверных отличий статистических параметров – первый шаг к познанию новых биологических закономерностей, причем количественно доказанных. Ответ на вопрос о достоверности или случайности отличий дают статистические критерии, среди которых самые распространенные параметрические критерии  $t$  Стьюдента и  $F$  Фишера. Вычисление их ведется по специальным формулам (различным в зависимости от сравниваемых параметров и типов распределения). Полученные этим способом значения критериев (для чего в формулы подставляются экспериментальные данные) сравнивают с табличными при выбранном уровне значимости (обычно 0.05) и числе степеней свободы (объемы выборок без числа ограничений). Результатом такого сравнения должен стать один из двух вариантов следующего статистического вывода. Если полученное значение (величина) критерия больше табличного, значит, различия между параметрами при заданном уровне значимости и установленном числе степеней свободы достоверны, в разных выборках действительно проявилось действие разных факторов или разных уровней одного фактора. Если же полученная величина критерия меньше табличной, то при данном уровне значимости и числе степеней свободы различия между параметрами недостоверны. Последнее говорит о том, что различия случайны, никакого определенного вывода о побудительных причинах отличий сделать нельзя, нулевая гипотеза остается не опровергнутой.

При сравнении выборок по степени выраженности признака говорят о достоверности (недостоверности) отличий средних арифметических и долей, а при сравнении по уровню изменчивости показателей – о достоверности (недостоверности) отличий стандартных отклонений (дисперсий) и коэффициентов вариации. Особый случай представляет сравнение двух выборок по характеру распределения (достоверность отличия частот), а также общее отличие выборок без указания определенных параметров (для признаков в полуколичественных единицах).

##### **Сравнение средних арифметических**

Для выяснения вопроса о случайном или неслучайном *расхождении значений* некоторого параметра проводят две серии экспериментов и для каждой из них подсчитывают средние значения  $M_1$  и  $M_2$ . Если между ними есть различие, то можно допустить два равновероятных объяснения. Первое - предположив, что обнаруженная разность обуславливается случайным варьированием средних

величин исследуемой выборки около некоторой общей, хотя и неизвестной нам в точности генеральной средней. Тогда найденные средние должны считаться принадлежащими к одной и той же качественно однородной совокупности. Второе - рассматриваемые две средние величины принадлежат к двум качественно отличным друг от друга статистическим совокупностям. Оба эти предположения одинаково законны и применимы ко всякой паре сравниваемых величин  $M$ .

**Сравнение двух групп при помощи не парного критерия Стьюдента.** Наиболее распространенный способ проведения технологического исследования основан на сравнении результатов полученных в контрольной и опытной группах с применением не парного критерия Стьюдента, которая справедлива для больших ( $n > 30$ ) или малых равновеликих выборок ( $n_1 \approx n_2$ ):

$$t_d = \frac{|M_1 - M_2|}{\sqrt{m_1^2 + m_2^2}} \geq t_{(p, k)}, \quad m_i = \frac{\sigma_i}{\sqrt{n}} \quad (30)$$

где:  $M_1$  и  $M_2$  – средние величины в контроле и опыте (выражения 5),  $m_1$  и  $m_2$  – ошибки среднеквадратического отклонения в контроле и опыте.

Следует помнить, что разность средних нужно брать по модулю, т. е. без учета знака. Для определения существенности различий необходимо обращаться к таблице Стьюдента (см. прилож. табл. 7). По найденному значению  $t_d$  и числу степеней свободы  $k = n_1 + n_2 - 2$  в таблице 7 находим соответствующее значение вероятности  $P$ . Результатом такого сравнения должен стать один из двух вариантов следующего статистического вывода. Если полученное значение (величина) критерия больше табличного, значит, различия между параметрами при заданном уровне значимости и установленном числе степеней свободы достоверны. Если же полученная величина критерия меньше табличной, то при данном уровне значимости и числе степеней свободы различия между параметрами недостоверны. Последнее говорит о том, что различия случайны, никакого определенного вывода сделать нельзя, нулевая гипотеза остается не опровергнутой.

Как уже отмечалось, в биологических экспериментах принято, чтобы показатель надежности был не менее 95% ( $P \geq 0,95$  или  $p \leq 0,05$ ). Если полученное различие меньше, к примеру, 80...90%, то необходимо высказываться с меньшей категоричностью. Можно указать, что разность может иметь место, но в данной серии экспериментов она не доказана, и необходимо поставить дополнительные опыты.

**Пример.** При сравнении выборочных параметров нормального и биномиального распределений используется одна и та же формула. Например, в процессе специальных исследований было установлено, что у стариков до лечения инсулином среднее содержание белков в крови составляло  $81.04 \pm 1.7$ , а после лечения  $79.33 \pm 1.6$ . Нетрудно видеть, что полученные величины неодинаковы. Но достоверно ли это различие, закономерно ли оно? Можно ли на его основании утверждать, что лечение инсулином понижает содержание белков в крови? Ответ на этот вопрос может дать критерий достоверности различий средних

арифметических. Согласно общей нулевой гипотезе средние не отличаются. Проверим ее с помощью критерия Стьюдента:

$$t = \frac{M_1 - M_2}{\sqrt{m_1^2 + m_2^2}} = \frac{81.04 - 79.33}{\sqrt{1.7^2 + 1.6^2}} = 0.7.$$

По таблице граничных значений критерия (табл. 6II) находим, что для уровня значимости  $p = 0.05$  и числа степеней свободы  $df = 20 + 20 - 2 = 38$  величина критерия составляет  $t_{(0.05, 39)} = 2.03$ . Поскольку полученное значение (0.7) меньше табличного (2.03), нулевая гипотеза сохраняется, различия между средними величинами статистически недостоверны (незначимы). Следовательно, влияние инсулина на содержание белков в крови приведенными выше данными не подтверждается и остается недоказанным, возможно, из-за недостаточного числа определений.

Заметим, что хотя применение  $t$ -критерия требует нормальности исходной случайной величины, он может применяться и при умеренных отклонениях от нормальности и не слишком малых  $n$  ( $n \geq 15$ ).

Не парный критерий Стьюдента может применяться для сравнения связанных выборок

$$t = \frac{(\bar{x} - \bar{y})\sqrt{n}}{\sqrt{s_x^2 + s_y^2 - 2s_{xy}}}, \text{ где } s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

**Сравнение долей.** При сравнении достоверности различия долей или процентов ( $W$ ) признаков, характеризующихся альтернативным распределением, применяют критерий Фишера с  $\varphi$ -преобразованием. Вместо процентов берут фи-значения ( $\varphi = \arcsin \sqrt{W}$  или по таблице 10II) и подставляют их в формулу:

$$F = \frac{(\varphi_1 - \varphi_2)^2 \cdot n_1 \cdot n_2}{n_1 + n_2} \sim F_{(\alpha, df1, df2)},$$

где  $\varphi_1$  и  $\varphi_2$  – преобразованные доли,  $n_1$  и  $n_2$  – объемы выборок.

Полученное значение сравнивают с табличным в соответствии с заданным уровнем значимости,  $\alpha = 0.05$ , и числом степеней свободы:  $df_1 = 1$ ,  $df_2 = n_1 + n_2 - 2$ .

Например, в процессе учетов мелких млекопитающих в двух разных биотопах, где стояло по 200 ловушек, попало соответственно 5 и 15 зверьков. Отличается ли численность животных на этих площадках? Если рассматривать ловушку как вариант, способную принимать два значения – «пустая» и «сработавшая» (со зверьком), то получаем выборку вариант (ловушек) с альтернативным распределением. Число пойманных особей можно пересчитать в процент сработавших ловушек:  $W_1 = 100\% \cdot 5 / 200 = 2.5\%$ ,  $W_2 = 100\% \cdot 15 / 200 = 7.5\%$ . По таблице 10II находим значения  $\varphi$  и вычисляем значение критерия:



$$F = \frac{(0.318 - 1.555)^2 \cdot 200 \cdot 200}{200 + 200} = 5.62. \text{ Полученная величина (5.62) больше крити-}$$

ческой  $F_{(0.05, 1, 398)} = 3.9$ , значит, численность мелких млекопитающих во втором биотопе достоверно выше, чем в первом.

Задача сравнения выборочных средних – это вопрос о том, действовал ли при составлении одной из выборок новый систематический фактор по сравнению с другой выборкой. В терминах статистики отличия между средними могут иметь два противоположных источника:

1. Обе выборки взяты из одной генеральной совокупности, но средние отличаются в силу ошибки репрезентативности.

2. Выборки взяты из разных генеральных совокупностей, отличие средних вызвано, в основном, действием разных доминирующих факторов (а также и случайно).

Статистическая задача состоит в том, чтобы сделать обоснованный выбор. Исходно предполагается (Но): «достоверных отличий между средними нет». Отличить закономерное от случайного можно только на основе знания законов поведения случайной величины. Для исключения чужеродных («выскакивающих») вариант мы применяли закон нормального распределения: в диапазоне четырех стандартных отклонений,  $M \pm 1.96 \cdot S$ , отклонение вариант от средней происходит по случайным причинам; за границами этого диапазона лежат чужеродные для данной выборки значения. Поскольку выборочные средние имеют нормальное распределение, критерий отличия двух выборочных средних также базируется на *свойствах нормального распределения*: в границах  $M_{общ.} \pm 1.96 \cdot m$  (или приблизительно  $M_{общ.} \pm 2 \cdot m$ ) выборочные средние арифметические отличаются от общей (генеральной) средней по случайным причинам.

**Сравнение двух групп при помощи парного критерия Стьюдента.** Для учета индивидуальных свойств биообъекта (в случае, когда исследование сопряжено с парно связанными параметрами) предпочтительно использовать парный критерий Стьюдента:

$$t_{\Delta} = \frac{\Delta M_{12}}{m_{12}}, \quad \Delta M_{12} = \frac{1}{k} \sum_{i=1}^k (M_{i1} - M_{i2}), \quad m_{12} = \sqrt{\frac{\sum_{i=1}^k (\Delta M_{12} - M_{i1} + M_{i2})^2}{k(k-1)}}, \quad (31)$$

где:  $M_{12}$  – средняя разность величин в контроле и опыте,  $m_{12}$  – ошибка среднеквадратического отклонения разности величин в контроле и опыте,  $k$  – количество пар измерений контроль – опыт.

Для наглядности преимущества использования парного критерия Стьюдента рассмотрим следующий пример. На рис. 39 представлены показатели подвижности спермиев до, и после криоконсервирования для каждого эякулята в виде вектора. На рисунке видно, что разброс результатов в выборках до и после криоконсервирования сопоставим с величиной их различия между средними

значениями. Вместе с тем, если соединить линиями отдельные значения, полученные для каждого эякулята, то во всех опытах наблюдается приблизительно одинаковое снижение подвижности спермиев (угол наклона вектора).

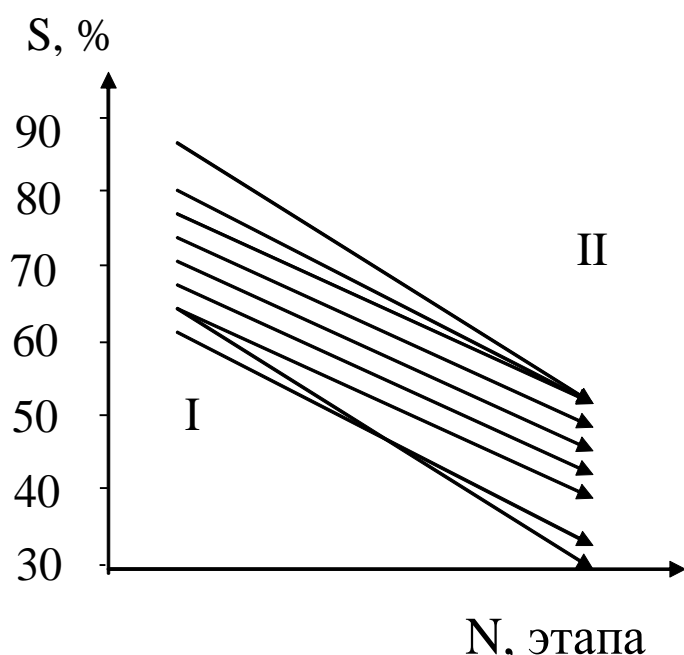


Рис. 39. Процент подвижности спермиев до (I этап) и после замораживания (II этап)

Анализ результатов при помощи не парного критерия Стьюдента (выражение 1.8 а) дает значение  $t_d=1,4$ , тогда как парный (выражение 1.8 b) в три раза выше  $t_d=4,3$ . Причина высокой чувствительности парного критерия основана на учете только межгрупповой вариации, тогда как не парный включает в себя общую дисперсию (внутри- и межгрупповую рис. 1.10 и 1.11). Учет гетерогенности эякулята (начальная подвижность спермиев) дает возможность повысить воспроизводимость результатов анализируемого эксперимента.

Таким образом, для повышения воспроизводимости результатов опытов необходимо чтобы единственным различием между контрольной и экспериментальной группой должно быть воздействие исследуемого, и никакого другого, факторов.

Рассмотрим способы учета составляющих вариаций исследуемого явления.

### Сравнение показателей изменчивости

Изучение суммарного влияния многих количественно неопределенных факторов, сочетающихся в произвольных и непонятных для исследователя соотношениях, часто приводит к неясным закономерностям и ошибочным выводам. В таких случаях следует все факторы, обуславливающие явления, разделить на основные  $a_j$ ,  $b_j$ , оказывающие наибольшее влияние на развитие явления, и дополнительные  $d_j$ , влияющие на развитие явления второстепенно. Тогда в опыте измеряют или отмечают лишь величины и свойства, характеризующие основные факторы.

Чтобы устранить или, по крайней мере, уменьшить ошибку, появляющуюся вследствие деления основных факторов на  $a_j$  - поддающиеся изменениям в ходе эксперимента и  $b_j$  - не поддающиеся, надо при постановке опытов стремиться нейтрализовать вариацию  $b_j$  факторов. То есть создать такие условия, при которых действия  $b_j$  факторов было бы возможно более неизменно и незначительно, а величины и свойства, характеризующие эти факторы, приближались бы к постоянным. При этом исследователь должен стремиться сделать перемен-

ными величинами лишь  $a_j$  факторы. Таким образом, общим принципом исследования является постоянство всех остальных факторов при изменении избранных  $a_j$ . Тестом на соблюдение выше перечисленных условий является критерий равноточности измерений.

Измерения, дающие дисперсии одинаковой величины, называются *равноточными*. Равноточность измерений серии опытов позволяет обобщить обработку результатов экспериментов и уменьшает суммарные ошибки исследования. Практически добиться равноточности измерений можно только в тех случаях, когда измерения будет проводить достаточно опытный человек одним и тем же прибором в одинаковых условиях.

Условием равноточности опытов является *критерий Фишера*  $F$ , который определяется из таблиц (см. приложение, табл. 3.).

$$F = \frac{\sigma_1^2}{\sigma_2^2} \sim F_{(p, df_1, df_2)},$$

где:  $\sigma_1$  и  $\sigma_2$  - среднеквадратические отклонения разных серий опытов,  $\sigma_1 \geq \sigma_2$ .  $df_1 = n_1 - 1$ ,  $df_2 = n_2 - 1$ .

В случае если вычисленное значение критерия Фишера окажется меньше или равным табличному ( $F \leq F_T$ ), то результаты считаются равноточными. В некоторых случаях неравноточность приводит к несравнимым результатам, к ошибочным выводам и заключениям, поэтому надо всегда добиваться равноточности измерений исследуемых величин. Для избегания получения искажения объективной информации также необходимо исключить грубые ошибки, которые возникают при проведении измерений.

Обобщение результатов эксперимента, полученных в различных опытах, проводят при помощи вычисления средневзвешенной величины и ее ошибки. Заметим, что в отличие от  $t$ -критерия  $F$ -критерий чувствителен к отклонениям исходных случайных величин от нормальности. Наиболее точным методом определения достоверности различий между выборочными дисперсиями служит критерий  $F$  Фишера в форме отношения дисперсий (большее значение должно стоять в числителе). При значительных отклонениях от нормальности, особенно при небольшом числе наблюдений, его не следует применять.

Рассмотрим такой пример. При сравнении по показателю плодовитости (число эмбрионов на самку) двух популяций красной полевки с разным уровнем численности (у первой, горной, популяции плотность населения в два раза выше, чем у равнинной) оказалось, что при очень близких средних арифметических (соответственно  $M_1 = 5.8$  и  $M_2 = 5.4$ , разница статистически недостоверна)

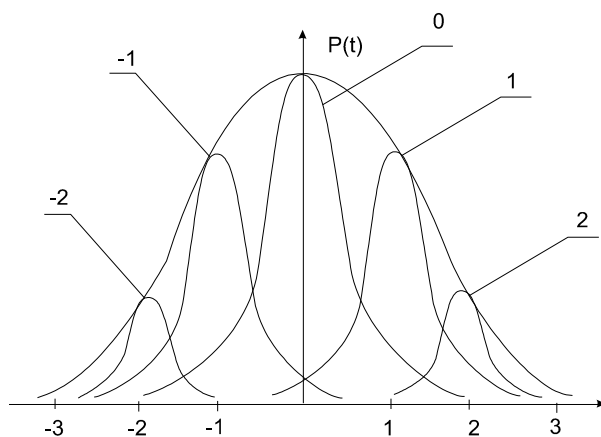


Рис. 40. Общая вариация генеральной выборки и её составляющие, выраженные через  $t$ -критерий, где  $t = (M - X) / \sigma$

стандартные отклонения значительно различаются:  $S_1 = 1.82$ ,  $S_2 = 0.52$  (при  $n_1 = 27$ ,  $n_2 = 12$ ). Отсюда

$$F = \frac{S_1^2}{S_2^2} = \frac{3.3124}{0.2704} = 12.25.$$

Полученное значение критерия ( $F = 12.2$ ) больше табличного  $F_{(0.05, 26, 11)} = 2.6$ , следовательно, нулевую гипотезу о случайности отличий можно отбросить, сделав вывод о том, что показатели изменчивости плодовитости в разных по численности популяциях достоверно отличаются. С биологических позиций это понятно, поскольку генетические отличия между особями практически по всем признакам, включая плодовитость, в больших популяциях выше, чем в малых. Новым фактором, усиливающим изменчивость особей в выборке, становится возможность появления аберрантных форм в условиях более свободной панмиксии.

Коэффициенты вариации также можно использовать для сравнения изменчивости разных показателей. Достоверность отличий коэффициентов оценивается с помощью критерия Стьюдента по формуле:

$$t = \frac{|CV_1 - CV_2|}{\sqrt{m_1^2 + m_2^2}} \sim t_{(0.05, n_1+n_2-2)},$$

где  $CV_1$ ,  $CV_2$  и  $m_1$ ,  $m_2$  – значения и ошибки коэффициентов вариации  $\phi(11)$ .

Вывод о достоверности отличий делается в том случае, если рассчитанное значение превысит табличное при заданном уровне значимости  $p = 0.05$  и числе степеней свободы  $df = n_1 + n_2 - 2$ . Сравним по критерию Стьюдента изменчивость веса тела землероек и плодовитости лисиц:

$CV_1 = 8.6 \pm 0.77\%$ ,  $n_1 = 63$ ;  $CV_2 = 26.7 \pm 2.2\%$ ,  $n_2 = 76$ , отсюда

$$t = \frac{|8.6 - 26.7|}{\sqrt{0.77^2 + 2.2^2}} = 7.76.$$

Поскольку полученное значение (7.8) больше табличного ( $t_{(0.05, 137)} = 1.96$ ), изменчивость плодовитости лисиц достоверно выше, чем изменчивость веса тела землероек.

Описанные выше статистические критерии ( $t$ ,  $F$  и др.) относятся к *параметрическим*, т. к. используют стандартные параметры распределений ( $M$ ,  $S$ ,  $n$ ). Они связаны с законом нормального распределения и применяются для оценки расхождения между генеральными параметрами по выборочным показателям сравниваемых совокупностей. Существенным достоинством параметрических критериев служит их большая статистическая мощность, т. е. широкие разрешающие возможности, а недостатком – трудоемкость расчетов, неприменимость к распределениям, сильно отклоняющимся от нормального, а также при исследовании качественных признаков.

Правильное применение параметрических критериев для проверки статистических гипотез основано на предположении о нормальном распределении совокупно-

стей, из которых взяты сравниваемые выборки. Однако это не всегда имеет место, так как не все биологические признаки распределяются нормально. Немаловажным является и то обстоятельство, что исследователю приходится иметь дело не только с количественными, но и с качественными признаками, многие из которых выражаются порядковыми номерами, индексами и другими условными знаками. В таких случаях необходимо использовать *непараметрические критерии*.

#### **4.1.3 Сравнение выборок с помощью непараметрических критериев (Nonparametric tests )**

**Непараметрические методы проверки гипотез** — это методы проверки нулевой гипотезы, предполагающие, что изучаемые переменные измерены с помощью номинальной или порядковой шкал.

Непараметрические критерии обладают широкой областью применения, устойчивостью выводов, простотой математических средств. Они значительно менее трудоемки, а при распределениях, далеких от нормального, более эффективны и точны, чем параметрические. В основе непараметрических статистических критериев лежит оперирование частотами или рангами эмпирических данных, при этом, в отличие от параметрических критериев, тип распределения данных не обязательно должен соответствовать нормальному. Для расчета непараметрических критериев результаты измерений должны быть представлены в шкале наименований рангов или в шкале интервалов (если распределение интервальных данных значительно отличается от нормального, что бывает довольно часто при малом размере выборки). Тип распределения данных при использовании непараметрических критериев может быть любым, но при этом необходимо учитывать ограничения, специфичные для некоторых критериев.

Вся процедура состоит из трех этапов – упорядочивание и ранжирование вариант, подсчет сумм рангов в соответствии с правилами данного критерия, сравнение полученной величины с табличным значением критерия. При этом с параметрическими критериями их роднит общая идеологическая подоплека. Нулевая гипотеза, как правило, состоит в том, что сравниваемые выборки взяты из одной и той же генеральной совокупности, значит, характер распределения вариант в этих выборках должен быть сходным. Поскольку вместо самих значений вариант используются ранги, все непараметрические методы исследуют один вопрос, насколько равномерно варианты разных выборок «перемешаны» между собой. Если варианты разных выборок более или менее регулярно чередуются в общем упорядоченном ряду, значит, они распределены сходным образом и отличий между совокупностями нет. Если же выборки пересекаются не полно (смешиваются только краями распределений, либо одна поглощает другую), то становится ясно, что эти выборки взяты из разных генеральных совокупностей (со смещенными центрами или разными дисперсиями).

### U-критерий Уилкоксона (Манна—Уитни).

Этот метод сравнения двух выборок признается наиболее чувствительным и мощным среди прочих непараметрических критериев. Согласно нулевой гипотезе, сравниваемые совокупности имеют одинаковые распределения. Техника метода состоит в том, что все варианты сравниваемых совокупностей ранжируют в одном общем ряду: каждому значению присваивают ранг, порядковый номер. При этом одинаковым (повторяющимся) значениям вариантов должен соответствовать один и тот же средний ранг (они как бы «делят места»).

Для расчета  $n_2$ -критерия необходимо:

1. Расположить числовые значения сравниваемых выборок в возрастающем порядке в один общий ряд и пронумеровать члены общего ряда от 1 до  $N = n_1 + n_2$ . (Эти номера и будут «рангами» членов ряда.)

2. Отдельно для каждой выборки найти суммы рангов  $R$  и определить величины, которые отображают связь между суммами рангов первой и второй выборки.

3. В качестве U -критерия использовать меньшую величину  $U_{\text{ф}}$ , которую сравнить с табличным значением  $U_{\text{ст}}$ . Условием для сохранения принятой  $H_0$ -гипотезы служит неравенство  $U_{\text{ф}} > U_{\text{ст}}$ . Критические точки (U -критерия  $U_{\text{ст}}$  для  $n_1$  и  $n_2$  и принимаемого уровня значимости  $\alpha$  содержатся в табл. XI Приложений.

После этого ранги вариант суммируют отдельно по каждой выборке:

$R_1 = \sum r_i$ ,  $R_2 = \sum r_j$ ,  $i = 1, 2, \dots, n_1$ ,  $j = 1, 2, \dots, n_2$  и вычисляют величину критерия:

$$t = \frac{U - 0.5 \cdot n_1 \cdot n_2}{\sqrt{(n_1 \cdot n_2 \cdot (n_1 + n_2 + 1) / 12)}},$$

где  $U = \max(U_1, U_2)$  – максимальное значение из двух величин:

$$U_1 = n_1 \cdot n_2 + 0.5 \cdot n_1 (n_1 + 1) - R_1,$$

$$U_2 = n_1 \cdot n_2 + 0.5 \cdot n_2 (n_2 + 1) - R_2.$$

Если выборка достаточно велика ( $n > 20$ ), величина статистики  $t$  сравнивается с табличным значением критерия Стьюдента для  $df = \infty$  и  $\alpha = 0.1$  (т. е. только для верхней 95% области нормального распределения). Считается, что метод хорошо работает для выборок объемом больше 10. В случае с меньшими выборками нужно пользоваться таблицами Уилкоксона – Манна – Уитни (табл. 11П).

В качестве примера сравним 5- и 35-дневных щенков песцов по активности фермента каталазы в сердце ( $E$ ):

5-дневные: 41, 44, 31, 38, 43, 29, 71, 45,  $M = 42.6$ ,  $S = 12.8$ ,  $n_1 = 8$ ,

35-дневные: 52, 51, 62, 52, 52, 50, 54, 62, 31,  $M = 51.7$ ,  $S = 9.0$ ,  $n_2 = 9$ .

Высокие коэффициенты вариации (30 и 17%) говорят о том, что распределения признаков, скорее всего, не соответствуют нормальному. Поэтому сравнивать средние следует с помощью непараметрического критерия. Ранжируем всю совокупность – упорядочим значения выборок по возрастанию:

$E_5$	29	31	38	41	43	44	45	71	
$E_{35}$	31	50	51	52	52	52	54	62	62

Затем упорядочим все значения вместе, но так, чтобы значения каждой выборки располагались в двух отдельных рядах ( $E_5$ ,  $E_{35}$ ). Такое расположение упрощает назначение рангов (ряды  $r_5$ ,  $r_{35}$ ) и суммирование рангов ( $R$ ):

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	$R$
$E_5$	29		31	38	41	43	44	45									71	
$E_{35}$		31							50	51	52	52	52	54	62	62		
$r_5$	1		2.5	4	5	6	7	8									17	50.5
$r_{35}$		2.5							9	10	12	12	12	14	15.5	15.5		102.5

$$U_1 = 9 \cdot 8 + 0.5 \cdot 9 \cdot (9 + 1) - 50.5 = 66.5,$$

$$U_2 = 9 \cdot 8 + 0.5 \cdot 8 \cdot (8 + 1) - 102.5 = 5.5,$$

$$U = \max(U_1, U_2) = 66.5,$$

$$T = \frac{66.5 - 0.5 \cdot 9 \cdot 8}{\sqrt{(9 \cdot 8 \cdot 10 / 12)}} = 3.81.$$

Полученное значение (3.81) больше табличного ( $t_{(0.1, \infty)} = 1.65$ , табл. 6П), т. е. активность каталазы с возрастом меняется. Раз выборки малы, воспользуемся точными таблицами Уилкоксона – Манна – Уитни (табл. 11П). Получаем  $t_{(0.05, 8, 9)} = 51$ . Полученное значение (66.5) больше табличного (51), следовательно, различия между выборками достоверны.

**Т-критерий Уилкоксона.** Когда члены сравниваемых выборок связаны попарно некоторыми общими условиями (зависимые выборки), различия между ними с достаточной точностью оцениваются с помощью *рангового критерия Уилкоксона*  $T$ . Парный Т-критерий Уилкоксона является более мощным, чем критерий знаков.

Т-критерий рассчитывают следующим образом.

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

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

3. Сравнивают эту величину  $T\phi$  с критическим значением  $T_{\alpha}t$  для принятого уровня значимости  $\alpha$  и числа парных наблюдений  $n$ , которое берут без нулевых разностей. Нулевую гипотезу отвергают, если  $T\phi > T_{\alpha}t$ . Если же  $T \geq T_{\alpha}t$ , нулевую гипотезу не учитывать нельзя. Критические значения парного критерия Уилкоксона  $T_{\alpha}t$  содержатся в табл. XIII Приложений.

*Пример.* Применим парный критерий Уилкоксона для проверки Но-гипотезы относительно данных о годовых удоях коров материнского поколения и их потомства, приведенных в табл. 43. В последней графе этой таблицы помещены ранги абсолютных разностей между парными значениями годового удоя коров. Определя-

ем суммы плюсовых и минусовых разностей:  $T_{(+)}=7+10+9+3+5+11+2+6=53$  и  $T_{(-)}=8+4+12+1=25$ . Меньшая разность дает  $T_f = 25$ . Сравниваем эту величину с критическим значением  $T_{st}=15$  для  $n=12$  и  $\alpha=5\%$  (см. табл. XIII Приложений). Так как  $T_f > T_{st}$ , то нулевая гипотеза остается в силе.

Удой коров, кг		Разница выражена		Ранги разницы
Материнского поколения	потомства	знаками	числами	
3770	2991	-	779	8
3817	4593	+	776	7
2450	3529	+	1076	10
3463	4274	+	811	9
3500	3103	-	397	4
5544	3949	-	1597	12
3112	3491	+	379	3
3150	3559	+	409	5
3118	2918	-	202	1
3018	4580	+	1562	11
4291	4510	+	219	2
3463	4144	+	681	6

**Критерий Q Розенбаума.** Этот критерий, как и предыдущие, оценивает достоверность различий двух эмпирических распределений, но в отличие от них почти не требует вычислений. Сравним два ряда цифр, характеризующих привесы (г) барашков одного возраста при добавлении в корм специальной подкормки (234, 277, 214, 201, 174, 167, 184, 157, 196, 173, 190, 191, 141, 150, 191) и без нее (183, 154, 175, 159, 157, 189, 198, 165, 176, 124, 173, 182, 204, 151, 147). Устанавливаем максимальные (277 и 204) и минимальные (141 и 124) значения и определяем порядковый номер сравниваемых совокупностей. В качестве *первой* следует принять выборку с *наибольшей* *вариантой* 277.

Далее находим число значений первой выборки, *превышающих* максимальное значение второй выборки (204):  $Q_1 = 3$  (это варианты 234, 277, 214). Затем определяем число вариантов второй выборки, *уступающих* по величине минимальному значению первой выборки (141):  $Q_2 = 1$  (варианта 124). Далее определяем критерий Розенбаума как сумму полученных чисел:  $Q = Q_1 + Q_2 = 3 + 1 = 4$ . По таблице 12/II находим критическое значение  $Q_{(0.05, 15, 15)} = 6$ . Поскольку эмпирическое значение (4) меньше табличного (6), приходим к выводу об отсутствии достоверного отличия выборок друг от друга, а значит, и влияния подкормки на привесы барашков. Следует все же иметь в виду, что возможности этого метода ограничены, он дает лишь прикидочный результат и оказывается эффективным только в случае сравнительно больших различий между выборками.

Среди множества известных методов мы рассмотрим два метода: Уилкоксона – Манна – Уитни (довольно точный, но не самый простой для вычислений) и критерий Q Розенбаума. (простой для расчетов, но не очень точный).



**Х-критерий Ван-дер-Вардена.** Этот критерий относится к группе ранговых критериев, его применяют для проверки нулевой гипотезы при сравнении друг с другом независимых выборок. Техника расчетов Х-критерия сводится к следующему. Сравниваемые выборки ранжируют в один общий ряд по возрастающим значениям признака. Затем каждому члену ряда присваивают порядковый номер, отмечающий его место в общем ранжированном строю. Далее по порядковым номерам одной из выборок, обычно меньшей по объему, находят отношение  $R/(N+1)$ , где  $N+1=n_1 + n_2 + 1$ , т. е. сумма всех членов сравниваемых групп, увеличенная на единицу, а  $R$  — порядковый номер членов ряда, их «ранг».

Масса тела мышей, г		Порядковый номер, R	R1	R2	R/(N+1)	Φ(R/(N+1))
опыт	контроль					
	60	1		1		
	60	2		2		
	62	3		3		
64		4	4		0,190	-0,88
	66	5		5		
68		6	6		0,286	-0,57
	68	7		7		
	69	8		8		
70		9	9		0,429	-0,18
	70	10		10		
	71	11		11		
72		12	12		0,571	0,18
	73	13		13		
75		14	14		0	0,43
76		15	15		0	0,57
	78	16		16		
79		17	17		0	0,88
	80	18		18		
80		19	19		0	1,31
83		20	20		0	1,66
n=9	n=11		116	94		3,4

С помощью специальной таблицы (см. табл. IX Приложений) находят значения функции  $\psi[R/(N+1)]$  для каждого значения  $R/(N+1)$ . Суммируя результаты (обязательно с учетом знаков!), получают величину  $X_{\phi} = \sum \psi[R/(N+1)]$ , которую сравнивают с критической точкой этого критерия  $X_{st}$  для принятого уровня значимости  $p$  и общего числа членов сравниваемых выборок, т. е.  $N=n_1 + n_2$ . Критические точки Х-критерия для 5%-ного и 1%-ного уровней значимости и общего числа членов двух выборок  $N=n_1 + n_2$  (с учетом разности  $n_1 - n_2$ ) содержатся в табл. X Приложений.

Нулевая гипотеза сводится к предположению, что сравниваемые выборки извлечены из генеральных совокупностей с одинаковыми функциями распределения. Ес-

ли окажется, что  $X_{\phi} \geq X_{st}$ , нулевая гипотеза должна быть отвергнута на принятом уровне значимости.

*Пример 10.* Вернемся к результатам опыта по проверке действия нового препарата на массу тела лабораторных мышей (см. пример 2), где сравнивали две группы — опытную ( $n_1 = 9$ ) и контрольную ( $n_2 = 11$ ) — с попарно не связанными числовыми значениями признака. Применим X-критерий Ван-дер-Вардена к оценке результатов этого эксперимента. Расчет X-критерия по ранжированным значениям признака сравниваемых групп приведен в табл. 40.

В данном случае  $N = n_1 + n_2 = 9 + 11 = 20$ . Для этого числа ( $N = 20$ ) и 5%-ного уровня значимости с учетом разности  $n_1 - n_2 = 11 - 9 = 2$  в табл. X Приложений находим  $X_{st} = 3,84$ . Так как  $X_{\phi} = 3,40 < X_{st} = 3,84$ , нулевую гипотезу не учитывать нельзя; разница между контролем и опытом оказывается статистически недостоверной.

**Критерий знаков z.** В тех случаях, когда результаты наблюдений выражаются не числами, а знаками плюс (+) и минус (—), различия между попарно связанными членами сравниваемых выборок оценивают с помощью *критерия знаков z*. Конструкция этого критерия базируется на весьма простых соображениях: если попарно сравниваемые значения двух зависимых выборок существенно не отличаются друг от друга, то число плюсовых и минусовых разностей окажется совершенно одинаковым; если же заметно преобладают плюсы или минусы, это будет указывать на положительное или отрицательное действие изучаемого фактора на результативный признак. Большое число однозначных разностей служит в качестве фактически найденной величины z-критерия знаков. При этом нулевые разности, т. е. случаи, не давшие ни положительного, ни отрицательного результата, обозначаемые цифрой 0, в расчет не принимают и число парных наблюдений соответственно уменьшается.

Как и всякий другой выборочный показатель, z-критерий знаков является величиной случайной; он служит для проверки  $H_0$ -гипотезы, т. е. предположения о том, что совокупность или совокупности, из которых взяты сравниваемые выборки, имеют одну и ту же или одинаковые функции распределения.  $H_0$ -гипотеза отвергается, если  $z_{\phi} \geq z_{st}$  для принятого уровня значимости  $\alpha$  и числа парных наблюдений  $n$ , взятых без нулевых разностей. Критические точки  $z_{st}$  для двух уровней значимости и числа парных наблюдений содержатся в табл. XII Приложений.

*Пример 13.* Изучали влияние туберкулина на состав периферической крови низших обезьян. Результаты наблюдений приведены в табл.

Из табл. видно, что после введения туберкулина количество эозинофилов в периферической крови у большинства обезьян оказалось пониженным. Так, из 14 наблюдений два оказались нулевыми, т. е.  $n = 14 - 2 = 12$ . Из этого числа положительных разностей насчитывается 10. Следовательно,  $z_{\phi} = 10$ . По табл. XII Приложений для  $n = 12$  и  $\alpha = 5\%$  находим  $z_{st} = 10$ . Равенство  $z_{\phi} = z_{st}$  дает основание отвергнуть  $H_0$ -гипотезу на 5%-ном уровне значимости. Следовательно, с вероятностью 95% можно утверждать, что введение туберкулина (реакция Манту) вызывает заметное снижение эозинофилов в периферической крови обезьян.

Номер подопытных животных	Эозинофилия в крови при введении туберкулина						Эффект воздействия
	до			после			
1	+	+		+	+		0
2	+	+	+	+	+		+
3	+	+		+			+
4	+	+		+			+
5	+	+	+	+	+		+
6	+	+		+	+	+	-
7	+	+		+			+
8	+			+	+		-
9	+	+	+	+	+		+
10	+	+		+			+
11	+	+		+	+		0
12	+	+	+	+			+
13	+	+		+			+
14	+	+		+			+

### Сравнение двух частотных распределений. Критерий хи-квадрат

В практике биологических исследований часто бывает необходимо проверить ту или иную гипотезу, т. е. выяснить, насколько полученный экспериментатором фактический материал подтверждает теоретическое предположение, насколько анализируемые данные совпадают с теоретически ожидаемыми. Возникает задача статистической оценки разницы между фактическими данными и теоретическим ожиданием, установления того, в каких случаях и с какой степенью вероятности можно считать эту разницу достоверной и, наоборот, когда ее следует считать несущественной, незначимой, находящейся в пределах случайности. В последнем случае сохраняется гипотеза, на основе которой рассчитаны теоретически ожидаемые данные или показатели. Таким вариационно-статистическим приемом проверки гипотезы служит метод *хи-квадрат* ( $\chi^2$ ). Этот показатель часто называют «критерием соответствия» или «критерием согласия» Пирсона. С его помощью можно с той или иной вероятностью судить о степени соответствия эмпирически полученных данных теоретически ожидаемым.

С формальных позиций сравниваются два вариационных ряда, две совокупности: одна – эмпирическое распределение, другая представляет собой выборку с теми же параметрами ( $n$ ,  $M$ ,  $S$  и др.), что и эмпирическая, но ее частотное распределение построено в точном соответствии с выбранным теоретическим законом (нормальным, Пуассона, биномиальным и др.), которому предположительно подчиняется поведение изучаемой случайной величины.

В общем виде формула критерия соответствия может быть записана следующим образом:

$$\chi^2 = \sum \frac{(a - A)^2}{A}, \text{ где } a - \text{ фактическая частота наблюдений, } A - \text{ теоретически ожидаемая частота для данного класса.}$$

Нулевая гипотеза предполагает, что достоверных различий между сравниваемыми распределениями нет. Для оценки существенности этих различий следует обратиться к специальной таблице критических значений хи-квадрат (табл. 9П) и, сравнив вычисленную величину  $\chi^2$  с табличной, решить, достоверно или не достоверно отклоняется эмпирическое распределение от теоретического. Тем самым гипотеза об отсутствии этих различий будет либо опровергнута, либо оставлена в силе. Если вычисленная величина  $\chi^2$  равна или превышает табличную  $\chi^2_{(\alpha, df)}$ , решают, что эмпирическое распределение от теоретического отличается достоверно. Тем самым гипотеза об отсутствии этих различий будет опровергнута. Если же  $\chi^2 < \chi^2_{(\alpha, df)}$ , нулевая гипотеза остается в силе. Обычно принято считать допустимым уровень значимости  $\alpha = 0.05$ , т. к. в этом случае остается только 5% шансов, что нулевая гипотеза правильна и, следовательно, есть достаточно оснований (95%), чтобы от нее отказаться.

Определенную проблему составляет правильное определение числа степеней свободы ( $df$ ), для которых из таблицы берут значения критерия. Для определения числа степеней свободы из общего числа классов  $k$  нужно вычесть число ограничений (т. е. число параметров, использованных для расчета теоретических частот).

В зависимости от типа распределения изучаемого признака формула для расчета числа степеней свободы будет меняться. Для *альтернативного* распределения ( $k = 2$ ) в расчетах участвует только один параметр (объем выборки), следовательно, число степеней свободы составляет  $df = k - 1 = 2 - 1 = 1$ . Для *полиномиального* распределения формула аналогична:  $df = k - 1$ . Для проверки соответствия вариационного ряда распределению *Пуассона* используются уже два параметра – объем выборки и среднее значение (численно совпадающее с дисперсией); число степеней свободы  $df = k - 2$ . При проверке соответствия эмпирического распределения варианту *нормальному* или *биномиальному* закону число степеней свободы берется как число фактических классов минус три условия построения рядов – объем выборки, средняя и дисперсия,  $df = k - 3$ . Сразу стоит отметить, что критерий  $\chi^2$  работает только для выборок *объемом не менее 25 вариантов*, а частоты отдельных классов должны быть *не ниже 4*.

Вначале проиллюстрируем применение критерия хи-квадрат на примере анализа *альтернативной изменчивости*. В одном из опытов по изучению наследственности у томатов было обнаружено 3629 красных и 1176 желтых плодов. Теоретическое соотношение частот при расщеплении признаков во втором гибридном поколении должно быть 3:1 (75% к 25%). Выполняется ли оно? Иными словами, взята ли данная выборка из той генеральной совокупности, в которой соотношение частот 3:1 или 0.75:0.25?

Сформируем таблицу (табл. 4), заполнив значениями эмпирических частот и результатами расчета теоретических частот по формуле:

$$A = n \cdot p,$$

где  $p$  – теоретические частоты (доли вариантов данного типа),  
 $n$  – объем выборки.

Например,  $A_2 = n \cdot p_2 = 4805 \cdot 0.25 = 1201.25 \approx 1201$ .

Таблица 4

Значение (цвет плода), $x_j$	Фактическая частота, $a$	Теоретическая частость, $p$	Теоретическая частота, $A$	$\frac{(a - A)^2}{A}$
Красный	3629	0.75	3604	0.187621
Желтый	1176	0.25	1201	0.5204
Сумма	$n = \Sigma a = 4805$	1	$n = \Sigma A = 4805$	$\chi^2 = 0.71$

Далее вычисляем хи-квадрат  $\chi^2 = 0.71$  и число степеней свободы (при двух классах и одном ограничении, объеме выборки)  $df = k - 1 = 2 - 1 = 1$ . По табл. 9II находим критическое значение  $\chi^2_{(0.05, 1)} = 3.84$ . Поскольку полученная величина (0.71) меньше табличной (3.84), различия сравниваемых распределений статистически недостоверны. Иначе говоря, фактические частоты хорошо согласуются с теоретически ожидаемыми. Полученные данные не отвергают принятую гипотезу о том, что в нашем случае имеется соотношение 3:1.

Здесь следует еще раз обратить внимание читателей на то обстоятельство, что сохранение нулевой гипотезы нельзя считать доказательством справедливости нулевой гипотезы. Результатами представленных вычислений теория о расщеплении по фенотипам в соотношении 3:1 *не доказана*, хотя и не опровергнута. Статистика доказывает только факт отличий, но не их отсутствие. Чтобы доказать теорию, нужно предположить анти-теорию (например, соотношение 1:1) и опровергнуть ее с помощью статистических приемов.

В процессе другого исследования добыты 671 самец и 569 самок. Требуется определить, подтверждают ли эти данные факт преобладания самцов или налицо просто случайное отличие цифр. Теоретическое отношение признаков (соотношение полов) 1:1. Подтверждается ли оно? Находим сумму  $671 + 569 = 1240$ , среднее 620,  $\chi^2 = \frac{(671 - 620)^2}{620} + \frac{(569 - 620)^2}{620} = 8.4$ .

Сравнение вычисленного (8.4) и критического значения (для  $df = 1$  и  $\alpha = 0.05$   $\chi^2_{(0.05, 1)} = 3.84$ ) явно свидетельствует о существенном отклонении фактического соотношения полов от гипотезы – 1:1. Вероятность правильности нулевой гипотезы (т. е. что в данном случае действительно имеет место численное равенство полов) оказалась даже меньше 0.01. Следовательно, есть все основания говорить о достоверном преобладании самцов.

В качестве первого примера решим задачу, соответствует ли *закону Пуассона* распределение числа повторных отловов альбатросов (табл. 5). В этом случае рассматривается процесс, этапами которого выступают события «отлов птицы». В чреде таких событий встречаются редкие – «отлов меченной особи». Биологическая подоплека состоит в следующем: случайны ли повторные отловы птиц или есть факторы, ответственные за нарушение случайности? Например, птицы могут приманиваться и стремиться попасться вновь либо могут стараться избежать повторного отлова. В обоих случаях птицы будут «умышленно» попадаться чаще или реже, нарушая *случайность* повторного отлова и искажая тем

самым форму распределения, которое будет отходить от формы, предписанной законом Пуассона. Согласно нулевой гипотезе, птицы ведут себя случайно, их встречаемость соответствует этому закону. Алгоритм расчетов теоретических частот для этого случая прост и основан на формулах прямого расчета теоретических частот:

$$A_0 = \frac{n}{e^M} \text{ (частота нулевого класса),}$$

$$A_x = \frac{M}{x} \cdot A_{x-1} \text{ (частота прочих классов),}$$

где  $M$  – средняя арифметическая ряда,  
 $x$  – значение ряда (число объектов в пробе),

$A_x$  – теоретическая частота значения  $x$ ,

$n$  – объем выборки (число проб),

$e = 2.7183\dots$  – основание натурального логарифма.

Параметры данного вариационного ряда были рассчитаны выше (с. 23):  
 $M = 0.968$ . Теоретическая частота нулевого значения равна:

$$A_0 = \frac{n}{e^M} = \frac{32}{e^{0.968}} = 11.93803 \approx 12,$$

частота значения  $x = 1$ :

$$A_x = \frac{M}{x} \cdot A_{x-1} = \frac{0.968}{1} \cdot 11.93 = 11.55602 \approx 11$$

и т. д. (табл. 5, графа 3).

Таблица 5

По окончании вычислений получаем два ряда частот, отличающиеся между которыми оцениваются по критерию хи-квадрат.

Перед расчетом значения критерия следует убедиться, что выполнены тре-

бования к данным для расчета критерия  $\chi^2$ :

- объем выборки более 25 вариантов,  $n > 25$ ,
- суммы эмпирических и теоретических частот равны объему выборки  $n = \Sigma a = \Sigma A$  (с точностью не ниже 1-2%),
- все классы эмпирического и теоретического рядов имеют частоты более 4,  $a_j > 4$ ; если какие-либо классы имеют меньше 4 вариантов (у нас значения 3 и 4 имеют частоты 2 и 1), то они должны быть объединены (суммированы) с соседними, что и показано в таблице с помощью фигурных скобок.

Далее вычисляем значения критерия: для первой строки

Число повторных отловов, $x$	Фактическая частота, $a$	Теоретическая частота, $A$	$\frac{(a - A)^2}{A}$
0	15	12	0.75
1	7	11	1.45
2	7	6	
3	2	2	0.17
4	1	1	
Сумма	$n = \Sigma a = 32$	$n = \Sigma A = 32$	$\chi^2 = 2.31$

$$\frac{(a - A)^2}{A} = \frac{(15 - 12)^2}{12} = 0.75$$

и т. д. (графа 4), итого  $\chi^2 = 2.31$ .

Число степеней свободы находим как *число окончательных классов* (3) минус число ограничений:  $df = k - 2 = 3 - 2 = 1$ .

Табличное значение  $\chi^2_{(0.05, 1)} = 3.84$ . Полученная величина (2.31) меньше табличной (3.84), следовательно, нулевая гипотеза не отвергается: эмпирическое распределение достоверно не отличается от распределения Пуассона. Иными словами, у нас нет оснований утверждать, что вероятность повторного отлова изменяется: нельзя утверждать, что сама операция отлова привлекает или пугает птиц.

Соответствие эмпирического ряда *распределению Пуассона* можно проверить и другим способом: сравнив по критерию Фишера величины средней арифметической и дисперсии для числа степеней свободы  $df_1 = n - 1$ ,  $df_2 = n - 1$ . В нашем случае  $M = 0.968$ ,  $S^2 = 1.257$ ,  $F = 1.257 / 0.968 = 1.157$ . Поскольку эта величина меньше табличной ( $F_{(0.05, 31, 31)} = 1.84$ ), сравниваемые показатели достоверно не отличаются, а равенство средней и дисперсии характерно лишь для распределения Пуассона.

При статистическом исследовании непрерывных признаков нужно быть уверенным, что они действительно подчиняются *нормальному закону*, а в случае дискретных признаков – *биномиальному*. Для такой проверки нулевая гипотеза звучит так: «полученное распределение соответствует нормальному (биномиальному)» или «выборка взята из генеральной совокупности, подчиняющейся закону нормального (биномиального) распределения». Все вычислительные операции для случаев нормального и биномиального распределений совпадают. Рассмотрим проверку первой гипотезы.

Расчеты начинаются с построения вариационного ряда и поиска центральных значений для каждого класса (табл. 6 и 7). Далее по формуле  $t = \frac{|x_j - M|}{S}$  вычисляются нормированные отклонения середины каждого классического интервала ( $x_j$ ) от общей средней  $M$  ( $S$  – стандартное отклонение). В нашем случае  $M = 9.29$  г,  $S = 0.897$  г., тогда, например для второго интервала получаем:  $t = |8.05 - 9.27| / 0.897 = 1.38$ . Далее определяем теоретические частоты нормального распределения, или ординаты нормальной кривой (табл. 4П), соответствующие вычисленным нормированным отклонениям. Для  $t = 1.38$  находим  $p = 0.1539 \approx 0.15$  (табл. 6, графа 5). (Заметим, что модуль в формуле нормированных отклонений берется потому, что в таблице 6П приведены частоты  $p$  только для положительных значений  $t$ .) Следующая операция, вычисление теоретических частот, ведется по формуле:

$$A = c \cdot p,$$

где  $p$  – ординаты нормальной кривой,

$c$  – константа ряда, определяемая по формуле  $c = \frac{dx \cdot n}{S}$ ,

$dx$  – классовой интервал (в данном случае он равен 0.7) (см. с. 13),  
 $n$  – объем выборки (63).

Для нашего примера  $c = \frac{0.7 \cdot 63}{0.897} = 49.16$ .

Таблица 6

Классовые интервалы	Центр интервала, $x_j$	Фактическая частота, $a$	Нормированное отклонение, $t$	Ординаты нормальной кривой, $p$	Теоретическая частота, $A$	$\frac{(a - A)^2}{A}$
7 – 7.7	7.35	2	2.16	0.04	2	0.1
7.8 – 8.4	8.05	7	1.38	0.15	8	
8.5 – 9.1	8.75	18	0.60	0.33	17	0.25
9.2 – 9.8	9.45	22	0.18	0.39	19	0.47
9.9 – 10.5	10.15	10	0.96	0.25	12	0.33
10.6 – 11.2	10.85	1	1.74	0.09	4	0.2
11.3 – 11.9	11.55	3	2.52	0.02	1	
$\Sigma$		$n = \Sigma a = 63$			$n = \Sigma A = 63$	$\chi^2 = 1.36$

Теоретическая частота для  $f = 0.15$  составит:

$A = 49.16 \cdot 0.1539 = 7.55 \approx 8$  (графа 6).

В результате вычислений получаем теоретическое нормальное распределение с параметрами  $M = 9.29$  г,  $S = 0.897$  г,  $n = 63$  (см. рис. 4, с. 20).

Перед расчетом критерия хи-квадрат проверяем совпадение суммы эмпирических и теоретических частот (по 63 варианты) и минимальные объемы в отдельных классах. Поскольку в крайних классах частоты были ниже 4, проводим их объединение (отмечено скобками), после чего число классов сократилось до  $k = 5$ . Вычисляем значения  $\chi^2$ : для первого класса  $(9 - 10)^2 / 10 = 0.1$ , для всего ряда  $\chi^2 = 1.36$ . Число степеней свободы  $df = 5 - 3 = 2$ . Табличное значение (табл. 9П)  $\chi^2_{(0.05, 2)} = 5.99$ .

Поскольку полученное значение (1.36) меньше табличного (5.99), нулевая гипотеза сохраняется, распределение бурозубок по массе тела достоверно от нормального не отличается.

Аналогичные расчеты для дискретного признака (плодовитость лисиц), имеющего предположительно биномиальное распределение (дискретный аналог нормального), представлены в табл. 7. Так, при параметрах  $M = 5$  экз.,  $S = 1.33$  экз. для второго интервала получаем:  $t = |8 - 5| / 1.33 = 1.5$ .



Таблица 7

Центр интервала, $x_j$	Фактическая частота, $a$	Нормированное отклонение, $t$	Ординаты нормальной кривой, $p$	Теоретическая частота, $A$	$\frac{(a - A)^2}{A}$
2	1 } 9 8 } 16 } 23 }	2.26	0.031	2 } 9 7 } 17 } 23 }	0
3		1.5	0.129		
4		0.75	0.301		
5		0	0.399		
6	21 } 6 3 } 3 }	0.75	0.301	17 } 9 7 } 2 }	0.94
7		1.5	0.129		
8		2.26	0.031		
Сумма	$n = \Sigma a = 75$			$n = \Sigma A = 75$	$\chi^2 = 2$

Соответствующая ордината нормальной кривой равна  $p = 0.1295$  (графа 4), теоретическая частота составит:

$$A = c \cdot p = 56.38 \cdot 0.129 = 7.3 \approx 7 \text{ (графа 5),}$$

поскольку значение  $c = 1 \cdot 75 / 1.33 = 56.38$ . В результате вычислений получаем частоты (A) распределения (с параметрами  $M = 5$ ,  $S = 1.33$ ,  $n = 75$ ), строго соответствующего биномиальному (см. рис. 5, с. 21). Объединим классы с частотами менее 4 и рассчитаем значение критерия  $\chi^2 = 2$ . Число степеней свободы (при трех ограничениях и пяти классах) равно  $df = 5 - 3 = 2$ . Поскольку это значение ( $\chi^2 = 2$ ) меньше табличного ( $\chi^2_{(0.05, 2)} = 5.99$ ), нулевая гипотеза не может быть отклонена, значит, распределение лисиц по плодовитости в целом соответствует биномиальному закону.

### Краткий анализ непараметрических процедур

По существу, для каждого параметрического критерия имеется, по крайней мере, один непараметрический аналог. Эти критерии можно отнести к одной из следующих групп:

- критерии различия между группами (независимые выборки);
- критерии различия между группами (зависимые выборки);
- критерии зависимости между переменными.

**Различия между независимыми группами.** Обычно, когда имеются две выборки (например, мужчины и женщины), которые вы хотите сравнить относительно среднего значения некоторой изучаемой переменной, вы используете  $t$ -критерий для независимых выборок. Непараметрическими альтернативами этому критерию являются: критерий серий Вальда-Вольфовица,  $U$  критерий Манна-Уитни и двухвыборочный критерий Колмогорова-Смирнова. Если вы имеете несколько групп, то можете использовать дисперсионный анализ. Его непараметрическими аналогами являются: ранговый дисперсионный анализ Краскела-Уоллиса и медианный тест.

**Различия между зависимыми группами.** Если вы хотите сравнить две переменные, относящиеся к одной и той же выборке (например, математические

успехи студентов в начале и в конце семестра), то обычно используется  $t$ -критерий для зависимых выборок. Альтернативными непараметрическими тестами являются: критерий знаков и критерий Вилкоксона парных сравнений. Если рассматриваемые переменные по природе своей категориальны или являются категоризованными (т.е. представлены в виде частот попавших в определенные категории), то подходящим будет критерий хи-квадрат Макнемара.

Если рассматривается более двух переменных, относящихся к одной и той же выборке, то обычно используется дисперсионный анализ (ANOVA) с повторными измерениями. Альтернативным непараметрическим методом является ранговый дисперсионный анализ Фридмана или Q критерий Кохрена (последний применяется, например, если переменная измерена в номинальной шкале). Q критерий Кохрена используется также для оценки изменений частот (долей).

Зависимости между переменными. Для того, чтобы оценить зависимость (связь) между двумя переменными, обычно вычисляют коэффициент корреляции. Непараметрическими аналогами стандартного коэффициента корреляции Пирсона являются статистики Спирмена R, тау Кендалла и коэффициент Гамма.

Если две рассматриваемые переменные по природе своей категориальны, подходящими непараметрическими критериями для тестирования зависимости будут: Хи-квадрат, Фи коэффициент, точный критерий Фишера. Дополнительно доступен критерий зависимости между несколькими переменными так называемый коэффициент конкордации Кендалла.

Этот тест часто используется для оценки согласованности мнений независимых экспертов (судей), в частности, баллов, выставленных одному и тому же субъекту.

**Описательные статистики.** Если данные не являются нормально распределенными, а измерения, в лучшем случае, содержат ранжированную информацию, то вычисление обычных описательных статистик (например, среднего, стандартного отклонения) не слишком информативно. Например, в психометрии хорошо известно, что воспринимаемая интенсивность стимулов (например, воспринимаемая яркость света) представляет собой логарифмическую функцию реальной интенсивности (яркости, измеренной в объективных единицах - люксах).

Выбор статистического метода для анализа результатов. Независимо от конкретных биотехнологических особенностей, для ответа на вопросы, поставленные перед исследованием, с помощью различных критериев приходится проверять нулевую гипотезу об отсутствии эффекта или различиях в эффекте, сравнивая при этом выборочные средние, доли, кривые выживаемости и т.п. Табл. 1 может помочь ориентироваться в множестве *статистических критериев* (параметрических и непараметрических), обычно применяемых для анализа результатов клинических исследований. Чтобы правильно выбрать статистический метод, необходимо учитывать прежде всего характер интересующего нас признака (количественный, порядковый или качественный) и тип распределения (нормальное или какое-либо другое). Для сравнения качественных переменных с числом градаций больше двух применяются статистические методы сравнения распределений, например критерий  $\chi^2$  или таблицы сопряженности соответствующей размерности. Кроме того, необходимо различать ситуации, связанные

со сравнением двух или более групп (при этом применяются различные критерии), а также сравниваются ли независимые группы или одна группа пациентов получает сравниваемые воздействия. 1-2 статистических метода из каждой клетки табл. 1, наиболее приспособленные для решения соответствующей задачи. В результате применения любого из этих критериев наличие статистически значимых различий будут или не будут установлены.

Таблица статистических критериев, обычно применяемых для решения типичных задач в области клинических исследований

Признак	Две независимые группы	Более двух независимых групп	Одна группа, связанные измерения	Одна группа, несколько связанных измерений
<i>Параметрические методы</i>				
Количественный, нормальное распределение	Критерий Стьюдента, дисперсионный анализ, критерий Тьюкки, критерий Шеффе	Дисперсионный анализ, критерии Стьюдента для множественных сравнений, критерий Тьюкки, критерий Даннета, критерий Шеффе, критерий Ньюмена-Кейлса	Критерий Стьюдента для связанных пар, дисперсионный анализ повторных измерений	Дисперсионный анализ повторных измерений, критерий Шеффе для зависимых выборок
<i>Непараметрические методы</i>				
Количественный, распределение отличается от нормального, порядковый	Критерий Уилкоксона—Манна—Уитни, медианный критерии	Критерий Краскела—Уоллиса, медианный критерии	Г-критерий Уилкоксона, критерий знаков, критерий знаковых рангов Уилкоксона	Критерий Фридмана
<i>Методы сравнения долей</i>				
Качественный, альтернативное распределение	Критерий $\chi^2$ , точный критерий Фишера	Критерий $\chi^2$	Критерий Мак-Нимара	Критерий Кокрена

Какой метод использовать. Нелегко дать простой совет, касающийся использования непараметрических процедур. Каждая непараметрическая процедура в модуле имеет свои достоинства и свои недостатки. Например, двухвыборочный критерий Колмогорова-Смирнова чувствителен не только к различию в

положении двух распределений, например, к различиям средних, но также чувствителен и к форме распределения. Критерий Вилкоксона парных сравнений предполагает, что можно ранжировать различия между сравниваемыми наблюдениями.

Если это не так, лучше использовать критерий знаков. В общем, если результат исследования является важным (например, оказывает ли людям помощь определенная очень дорогостоящая и болезненная терапия?), то всегда целесообразно применить различные непараметрические тесты. Возможно, результаты проверки (разными тестами) будут различны. В таком случае следует попытаться понять, почему разные тесты дали разные результаты.

С другой стороны, непараметрические тесты имеют меньшую статистическую мощность (менее чувствительны), чем их параметрические конкуренты, и если важно обнаружить даже слабые отклонения (например, является ли данная пищевая добавка опасной для людей), следует особенно внимательно выбирать статистику критерия.

Большие массивы данных и непараметрические методы. Непараметрические методы наиболее приемлемы, когда объем выборок мал. Если данных много (например,  $n > 100$ ), то не имеет смысла использовать непараметрические статистики.

Главное здесь состоит в том, что когда выборки становятся очень большими, то выборочные средние подчиняются нормальному закону, даже если исходная переменная не является нормальной или измерена с погрешностью.

Таким образом, параметрические методы, являющиеся более чувствительными (имеют большую статистическую мощность), всегда подходят для больших выборок. Большинство критериев значимости многих непараметрических статистик, описанных далее, основываются на асимптотической теории (больших выборок) поэтому соответствующие тесты часто не выполняются, если размер выборки становится слишком малым. Обратитесь к описаниям определенных критериев, чтобы узнать больше об их мощности и эффективности.

## **4.2 Дисперсионный анализ**

### **4.2.1 Однофакторный дисперсионный анализ количественных признаков**

Дисперсионный анализ позволяет оценить степень и достоверность отличия нескольких выборочных средних одновременно, т. е. изучить влияние одного контролируемого фактора на результативный признак путем оценки его относительной роли в общей изменчивости этого признака, вызванной влиянием всех факторов. Сущность дисперсионного анализа заключается в расчленении общей вариации (дисперсии) изучаемого признака, вычисляемой по сумме квадратов отклонений отдельных вариантов ( $x$ ) от средней арифметической всего комплекса наблюдений ( $M$ ), на его составные части – дисперсию, вызванную организованными, учитываемыми в исследовании факторами (факториальную дисперсию), оценивающую межгрупповую изменчивость, и дисперсию, обусловленную остальными, неорганизованными в данном исследовании факторами

(внутригрупповую, или случайную, дисперсию) отклонения отдельных значений от средней в группе.

Общая вариация (сумма квадратов) признака рассчитывается как сумма квадратов отклонений всех вариантов ( $x_i$ ) от общей средней ( $M$ ):

$$C_{\text{общ.}} = \sum (x_i - M)^2.$$

Факториальная (межгрупповая, межвыборочная) сумма квадратов рассчитывается как сумма квадратов отклонений частных средних ( $M_j$ ) для каждой выборки (всего  $k$  выборок) от общей средней:

$$C_{\text{факт.}} = \sum (M_j - M)^2.$$

Остаточная (случайная, внутригрупповая) сумма квадратов есть сумма квадратов отклонений вариантов каждой выборки ( $x_i$ ) от своей средней ( $M_j$ ):

$$C_{\text{случ.}} = \sum (x_i - M_j)^2.$$

Очевидно, что в общем комплексе наблюдений должно выполняться равенство  $C_{\text{общ.}} = C_{\text{факт.}} + C_{\text{случ.}}$ .

Отношение сумм квадратов к соответствующему числу степеней свободы дает оценку величины дисперсии, или средний квадрат, иногда ее именуют варианса. Влияние изучаемого фактора отражает факториальная, или межгрупповая, дисперсия  $S^2_{\text{факт.}}$ , а влияние случайных неорганизованных в данном исследовании причин – случайная  $S^2_{\text{случ.}}$ , или внутригрупповая, остаточная дисперсия  $S^2_{\text{остат.}}$ :

$$S^2_{\text{факт.}} = \sum_{j=1}^k (M_j - M_{\text{общ.}})^2 / df_{\text{факт.}},$$

где  $df_{\text{факт.}} = k - 1, j = 1, 2, \dots, k, k$  – число сравниваемых средних.

$$S^2_{\text{случ.}} = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - M_j)^2 / df_{\text{случ.}},$$

где  $df_{\text{случ.}} = n - 1, i = 1, 2, \dots, n, n$  – число вариантов всех выборок.

Сила влияния фактора определяется как доля частной суммы квадратов в общем варьировании признака. Показатель силы влияния изучаемого фактора составляет:  $\eta^2_{\text{факт.}} = C_{\text{факт.}} / C_{\text{общ.}}$ , неорганизованных (случайных):  $\eta^2_{\text{случ.}} = C_{\text{случ.}} / C_{\text{общ.}}$ ; сумма этих показателей, естественно, равна единице:  $\eta^2_{\text{факт.}} + \eta^2_{\text{случ.}} = 1$ . Заметим, что показатель силы влияния дисперсионного комплекса есть не что иное, как квадрат пирсоновского корреляционного отношения, которым и оценивается относительная доля влияния организованного (изучаемого) фактора в общем суммарном статистическом влиянии всех факторов, определяющих развитие данного результативного признака.

О достоверности оценок влияния факторов судят по уже знакомому нам критерию Фишера:  $F = \frac{S^2_{\text{факт.}}}{S^2_{\text{случ.}}} \sim F_{(\alpha, df_1, df_2)}$ ,

где  $df_1 = k - 1, df_2 = n - k, k$  – число градаций,  $n$  – общий объем всех выборок.

Проверяется нулевая гипотеза: «влияние фактора на признак отсутствует». Влияние считается доказанным, если величина расчетного критерия равна или превышает свое табличное значение с принятым уровнем значимости (обычно  $\alpha = 0.05$ ) ( $F$  определяется по табл. 7II). Все параметры однофакторного дисперсионного анализа и порядок их вычислений представлены в таблице 8.

Таблица 8

Составляющие дисперсии	Суммы квадратов ( $SS$ ), $C$	Сила влияния, $\eta^2$	Степени свободы, $df$	Дисперсии (средний квадрат, $MS$ ), $S^2$	Критерий влияния, $F$
Факториальная	$C_{\text{факт.}} = \sum (M_j - M)^2$	$\frac{C_{\text{факт.}}}{C_{\text{общ.}}}$	$k - 1$	$S^2_{\text{факт.}} = \frac{C_{\text{факт.}}}{df_{\text{факт.}}}$	$F = \frac{S^2_{\text{факт.}}}{S^2_{\text{случ.}}}$
Случайная	$C_{\text{случ.}} = \sum (x_i - M_j)^2$		$n - k$	$S^2_{\text{случ.}} = \frac{C_{\text{случ.}}}{df_{\text{случ.}}}$	
Общая	$C_{\text{общ.}} = \sum (x_i - M)^2$				

Однофакторным называется анализ, изучающий действие на результативный признак только одного организованного фактора  $A$ . Для примера оценим влияние растворенного в воде вещества на плодовитость дафний, используемых в качестве тест-объектов в водно-токсикологических экспериментах. В ходе предварительного исследования были получены четыре выборки, четыре группы значений плодовитости животных, выращенных в средах с разным содержанием химической добавки.

Сначала необходимо сгруппировать выборочный материал в комбинативную таблицу (организовать дисперсионный комплекс). Для этого варианты каждой выборки записываются в отдельные графы, именуемые градациями (табл. 9). Результативным признаком служит средняя плодовитость дафний за неделю (для иллюстративности расчетов она дана в целых числах). В нашем примере организованы 4 градации – чистая вода (контроль, градация  $A1$ ; значения плодовитости 6, 5, 5, 7), слабая концентрация вещества (5 мг/л,  $A2$ ; 8, 7, 6, 6), средняя (15 мг/л,  $A3$ ; 8, 8, 7) и сильная (30 мг/л,  $A4$ ; 8, 7, 9). Предлагаемый ниже алгоритм расчетов позволяет использовать неравное число вариантов в градациях. Расчеты показаны в таблице 9.

Таблица 9

	Градации фактора									
	A1		A2		A3		A4			
	$x$	$x^2$	$x$	$x^2$	$x$	$x^2$	$x$	$x^2$		
	6	36	8	64	8	64	8	64		
	5	25	7	49	8	64	7	49		
	5	25	6	36	7	49	9	81		
7	49	6	36							
									Σ	
Σ $x^2$		135		185		177		194	691	$H1 = \Sigma \Sigma x^2 = 691$
Σ $x$	23		27		23		24		97	$H2 = (\Sigma \Sigma x)^2/n =$
$n$	4		4		3		3		14	$= (97)^2/14 = 672$
Σ $x^2/n$	132		182		176.3		192		682.8	$H3 = \Sigma \Sigma x^2/n =$
M	5.8		6.8		7.67		8		6.93	$= 682.8$

$$C_{\text{факт.}} = H3 - H2 = 682.8 - 672 = 10.76$$

$$C_{\text{случ.}} = H1 - H2 = 691 - 672 = 8.17$$

$$C_{\text{общ.}} = H1 - H3 = 691 - 682.8 = 18.93$$

Полученные значения позволяют вычислить дисперсии, определить силу влияния фактора и критерий достоверности Фишера.

Составляющие дисперсии	Суммы квадратов, $C$	Сила влияния, $\eta^2$	Степени свободы, $df$	Дисперсии, $S$	Критерий, $F$
Факториальная	10.76	57%	3	3.59	4.39
Случайная	8.17		10	0.82	
Общая	18.93			4.39	

Поскольку полученное значение критерия ( $F = 4.39$ ) больше табличного ( $F_{(0.05,3,10)} = 3.7$ ) (табл. 7П), отличие факториальной и случайной дисперсий достоверно, влияние фактора значимо.

Отсюда следует биологический вывод: стимулирующее влияние изучаемого фактора (вещества) на плодовитость дафний относительно велико (57%) и достоверно (с вероятностью  $P > 0.95$ ).

### **Непараметрический однофакторный дисперсионный анализ**

Рассмотренные выше схемы дисперсионного анализа исходили из предположения о нормальном распределении изучаемого результативного признака. Когда для какого-либо признака нет уверенности, что выполняется предположение о его нормальном распределении, когда требуется провести анализ быстро и без особой точности, когда мало данных или они выражены *качественными признаками*, можно использовать схему непараметрического дисперсионного

анализа. Этот метод более неприхотлив, но менее точен, нежели параметрический анализ. Он исследует распределения вариантов в нескольких выборках. Нулевая гипотеза состоит в том, что распределения одинаковы, т. е. выборки взяты из одной генеральной совокупности.

Порядок вычислений состоит в том, что все варианты ранжируются в порядке возрастания. Затем суммируются ранги вариантов по каждой выборке отдельно и рассчитывается критерий:

$$H = \frac{12}{n \cdot (n-1)} \cdot \left( \frac{R_1^2}{n_1} + \dots + \frac{R_j^2}{n_j} + \dots + \frac{R_k^2}{n_k} \right) - 3 \cdot (n+1) \sim \chi^2_{(\alpha, k-1)},$$

где  $n$  – число всех вариантов,

$n_j$  – объем  $j$ -й градации фактора,

$R_j$  – сумма рангов для каждой  $j$ -й градации фактора,

$k$  – число градаций фактора ( $j = 1, 2, \dots, k$ ).

При объеме выборок больше 5 вариант статистика  $H$  имеет распределение хи-квадрат с  $df = k - 1$  степенями свободы и сравнивается со значениями из табл. 9П.

Применим эту схему (табл. 10) к нашим данным из табл. 9, расположив их в строку.

№ п/п	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Градация	1	1	1	1	2	2	2	2	3	3	3	4	4	4
Значение	5	5	6	7	6	6	7	8	7	8	8	7	8	9

Затем упорядочим и ранжируем их. Для нескольких одинаковых значений берется средний ранг.

№ п/п	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Градация	1	1	1	2	2	1	2	3	4	2	3	3	4	4
Значение	5	5	6	6	6	7	7	7	7	8	8	8	8	9
Ранг	1.5	1.5	4	4	4	7.5	7.5	7.5	7.5	11.5	11.5	11.5	11.5	14

Наконец, разнесем ранги по градациям и подсчитаем необходимые суммы.

Таблица 10

Градация	1	1	1	1	2	2	2	2	3	3	3	4	4	4	
Значение	5	5	6	7	6	6	7	8	7	8	8	7	8	9	
Ранг, $R$	1.5	1.5	4	7.5	4	4	7.5	11.5	7.5	11.5	11.5	7.5	11.5	14	
Сумма, $R$				14.5				27				30.5			33
$n$				4				4				3			3
$R^2/n$				52.56				182.3				310.1			363

Общий объем выборки равен  $n = 14$ . Величина критерия  $H$  составит:



$$H = \frac{12}{14 \cdot 13} \cdot (52.56 + 182.3 + 310.1 + 363) - 3 \cdot 13 = \\ = 0.065934 \cdot 907.8958 - 45 = 14.86.$$

По таблице распределения статистики  $\chi^2$  для  $\alpha = 0.05$  и  $df = 4 - 1 = 3$  находим  $\chi^2_{(0.05, 3)} = 7.81$ . Полученное значение критерия (14.86) больше табличного (7.81), значит, отличие выборочных распределений достоверно. Химическая добавка действительно изменяет плодовитость дафний.

#### 4.2.2 Двухфакторный дисперсионный анализ количественных признаков

Двухфакторный дисперсионный анализ исследует влияние на результативный признак двух факторов как порознь, так и совместно. Учет эффекта влияния каждого фактора по отдельности теоретически ничем не отличается от описанных выше схем. И там и тут оценивается изменчивость средних по градациям на фоне случайной изменчивости вариантов внутри градаций, с помощью критерия Фишера устанавливается достоверность отличий межгрупповых дисперсий от внутригрупповых.

Двухфакторный дисперсионный анализ, естественно, требует более сложных вычислительных операций, чем однофакторный, но в принципе ничем не отличается от описанных выше схем. Однако это относится лишь к ортогональным (равномерным, или пропорциональным) комплексам, характеризующимся равной или по крайней мере пропорциональной численностью групп (в градациях содержатся одинаковые или пропорциональные числа вариантов). Что же касается неортогональных многофакторных комплексов, то их анализ принципиально возможен, но имеет свои особенности, существенно усложняющие технику вычислений, и в настоящем пособии не рассматривается.

На практике вполне допустим и такой способ избежать сложностей обработки неравномерных комплексов, как искусственное превращение их в равномерные. Для этого нужно составить выборки одинаковой или пропорциональной численности, используя только часть имеющихся данных. Следует, однако, помнить, что такой отбор не должен быть субъективным. Чтобы не допустить возможной тенденциозности, лучше всего прибегнуть к жеребьевке.

Важным преимуществом двухфакторного дисперсионного анализа перед однофакторным служит то, что с его помощью удастся определить варьирование по сочетанию градаций  $C_{\text{сочет.}} = C_{AB}$ , позволяющее получить новый и весьма ценный в биологическом отношении показатель – оценку влияния сочетанного действия (взаимодействия) факторов.

Общая вариация (сумма квадратов) признака теперь состоит из четырех компонентов за счет более детального разложения факториальной дисперсии.

Правило разложения вариаций предстает как:

$$C_{\text{общ.}} = C_A + C_B + C_{AB} + C_{\text{случ.}}$$

$$C_{\text{факт.}} = C_{\text{общ.}} - C_{\text{случ.}} = C_A + C_B + C_{AB}.$$

Для расчетов используются следующие смысловые формулы:

$$C_{\text{общ.}} = \sum (x_i - M)^2,$$

$C_A = \Sigma(M_{Aj} - M)^2$ ,  $j$  – число градаций фактора  $A$ ,  $M_{Aj}$  – групповые средние по градациям фактора  $A$ ,

$C_B = \Sigma(M_{Bk} - M)^2$ ,  $k$  – число градаций фактора  $B$ ,  $M_{Bk}$  – групповые средние по градациям фактора  $B$ ,

$$C_{случ.} = \Sigma(x_i - M_{xi})^2,$$

$$C_{AB} = C_{общ.} - (C_A + C_B + C_{случ.}).$$

Сочетанное действие (взаимодействие) каждого из двух факторов проявляется в усилении или ослаблении непосредственного действия другого фактора на объект исследования. К примеру, неурожай кормов усугубляет негативное действие зимнего холода на численность популяций мелких млекопитающих.

Рассмотрим числовой пример – испытания стимулятора многоплодия при разной полноценности рационов. Полноценность рациона (первый фактор) представлена двумя градациями:  $A1$  – рацион с недостатком минеральных веществ,  $A2$  – рацион, полностью сбалансированный по всем питательным веществам, включая и минеральные. Стимулятор (второй фактор) был испытан в трех дозах:  $B1$  – одинарная,  $B2$  – двойная,  $B3$  – тройная. Результативный признак – плодовитость самок, измерявшаяся числом детенышей в помете. Для каждого сочетания градаций рациона и стимулятора были подобраны три одновозрастные самки.

Комбинативная таблица двухфакторного равномерного дисперсионного комплекса с трехкратной повторностью ( $n_i = 3$ ) включает две градации по фактору  $A$  и три градации по фактору  $B$  (табл. 11). Варианты размещаются по градациям, определяется объем градации, вычисляются суммы вариантов, частные средние, затем вспомогательные величины ( $H1, H2, H3, H_A, H_B$ ) и суммы квадратов отклонений (дисперсий) по рабочим формулам. В завершение всего заполняют таблицу дисперсионного анализа (табл. 12), находят показатель достоверности влияния Фишера и, сопоставляя его с табличным для соответствующих степеней свободы и принятого уровня значимости, делают статистический вывод.

Таблица 11

Градации факторов		A1		A2		$\Sigma$	Для B		
		$x$	$x^2$	$x$	$x^2$		$M_B$	$\Sigma\Sigma x^2/n$	$\Sigma(\Sigma x^2/n)$
B1		5	25	1	1		4	96	$H_B = \Sigma(\Sigma x^2/n) = 486$
		6	36	4	16				
		7	49	1	1				
	$\Sigma x^2$		110		18	$\Sigma\Sigma x^2 = 128$			
	$\Sigma x$	18		6		$\Sigma\Sigma x = 24$			
	$n$	3		3		$n_{B1} = 6$			
	$\Sigma x^2/n$	108		12		$\Sigma(\Sigma x^2/n) = 120$			
B2		4	16	10	100		7	294	
		3	9	9	81				
		5	25	11	121				
	$\Sigma x^2$		50		302	$\Sigma\Sigma x^2 = 352$			
	$\Sigma x$	12		30		$\Sigma\Sigma x = 42$			
	$n$	3		3		$n_{B2} = 6$			
	$\Sigma x^2/n$	48		300		$\Sigma(\Sigma x^2/n) = 348$			
B3		2	4	7	49		4	96	
		3	9	4	16				
		1	1	7	49				
	$\Sigma x^2$		14		114	$\Sigma\Sigma x^2 = 128$			
	$\Sigma x$	6		18		$\Sigma\Sigma x = 24$			
	$n$	3		3		$n_{B3} = 6$			
	$\Sigma x^2/n$	12		108		$\Sigma(\Sigma x^2/n) = 120$			
$\Sigma\Sigma$	$\Sigma\Sigma x^2$		174		434	$H1 = \Sigma\Sigma\Sigma x^2 = 608$		$H2 = (\Sigma\Sigma\Sigma x)^2/N = 450$	
	$\Sigma\Sigma x$	36		54		$\Sigma\Sigma\Sigma x = 90$			
	$n_A = \Sigma n$	9		9		$N = \Sigma\Sigma n = 18$			
		$\Sigma x^2/n$	168		420		$H3 = \Sigma\Sigma(\Sigma x^2/n) = 588$		
Для A	$M_A = \Sigma\Sigma x/n$	2	6			$j = 2 - \text{число градаций фактора A}$ $k = 3 - \text{число градаций фактора B}$			
	$\Sigma x^2/n$	144	324						
	$H_A = \Sigma(\Sigma x^2/n) = 468$								

$C_{\text{общ.}} = H1 - H2 = 608 - 450 = 158$
$C_{\text{случ.}} = H1 - H3 = 608 - 588 = 20$
$C_{\text{факт.}} = C_{A+B+AB} = H3 - H2 = 588 - 450 = 138$
$C_A = H_A - H2 = 468 - 450 = 18$
$C_B = H_B - H2 = 486 - 450 = 36$
$C_{AB} = C_{\text{факт.}} - C_A - C_B = 138 - 18 - 36 = 84$

В нашем примере все факториальные влияния оказались достоверными с доверительной вероятностью  $P > 0.95$  (табл. 12). Это позволяет сделать определенные выводы относительно действия стимулятора на плодовитость самок. Влияние каждого фактора в отдельности (качества рациона и дозы стимулятора)

и их суммарного эффекта достаточно существенно, но особенно результативно действие стимулятора в сочетании с полноценным рационом (величина  $\eta^2_{AB}$  выше, чем  $\eta^2_A$  и  $\eta^2_B$ ). Более того, при недостатке в корме минеральных веществ двукратные и трехкратные дозы стимулятора могут даже снизить плодовитость животных.

Таблица 12

Составляющие дисперсии	Суммы квадратов, $C$	Сила влияния, $\eta^2$ (%)	Степени свободы, $df$	Дисперсии, $S$	Критерий, $F$ ( $F_{(\alpha, df_1, df_{сл.})}$ )
Фактор $A$	18	11	$j - 1 = 1$	18	10.8 (4.7)
Фактор $B$	36	23	$k - 1 = 2$	18	10.8 (3.9)
Взаимодействие $AB$	84	53	$df_A \cdot df_B = 2$	42	25.2 (3.9)
Факториальная (всего)	138	87	$j \cdot k - 1 = 5$	27.6	16.5 (3.1)
Случайная	20	13	$N - j \cdot k = 12$	1.67	
Общая	158	100	$N - 1 = 17$		

Таблица двухфакторного дисперсионного анализа имеет ту же структуру, что и таблица для однофакторного анализа, только факториальная дисперсия разложена на три компоненты (для факторов  $A$ ,  $B$  и их взаимодействия). Для каждой из них требуется вычислить число степеней свободы с учетом числа градаций фактора  $A$  ( $j$ , количество столбцов) и числа градаций фактора  $B$  ( $k$ , количество рядов), значения дисперсий, а также критерий Фишера. Поскольку каждому из расчетных значений критерия соответствует свое число степеней свободы, табличные значения окажутся разными.

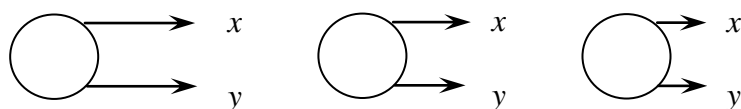
### Оценка зависимости между признаками

Изложенные выше методы статистического анализа дают возможность изучать изменчивость биологических объектов по отдельным признакам – весу, размерам, плодовитости, физиологическим показателям и др. Однако в ряде случаев важно знать, какова зависимость между вариацией двух или нескольких признаков, изменяются ли две переменные самостоятельно, независимо друг от друга, или варьирование одного признака в какой-то степени связано с изменчивостью другого. В качестве второй переменной часто выступает какой-либо фактор среды.

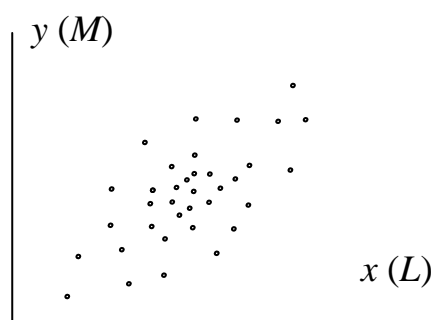
Задачу исследования зависимостей можно рассматривать как развитие метода дисперсионного анализа, решающего задачу сравнения нескольких выборок, т. е. изучающего влияния фактора на признак. Техника дисперсионного анализа имеет две особенности. Фактор (или факториальный признак) задан дискретно, в виде градаций, или «доз». Когда исследуется фактор, заданный *качественно*, то разбиение на градации всего диапазона его действия оказывается очень эффективным способом создания подобия количественной переменной.

Но при изучении количественно заданного фактора в грубой градуальной схеме дисперсионного анализа утрачивается часть информации, которая содержится в исходных выборках и которую можно было бы использовать. Кроме этого, дисперсионный анализ явным образом не учитывает тенденции изменения среднего уровня признака при изменении уровня фактора, не содержит показателя характера (знака) зависимости признака от фактора. Все эти «недостатки» дисперсионного анализа не характерны для методов изучения *сопряженной изменчивости* – корреляционного и регрессионного анализов.

Способ представления отдельных наблюдений здесь меняется: каждая варианта рассматривается как носитель двух численных характеристик объекта измерения, двух *зависимых* значений случайной величины. Если выше мы отождествляли отдельное значение с отдельной вариантой, то теперь мы рассматриваем варианту как некоторое тело, обладающее минимум двумя зарегистрированными качествами, различными у разных вариант:



Например, для любого животного можно определить массу ( $M$ ) и длину ( $L$ ) тела; отдельная варианта будет нести два значения ( $L, M$ ). При этом множество вариант выборки можно отобразить графически как точки на плоскости осей двух признаков  $M$  и  $L$ . Вся выборка предстанет в виде множества точек на плоскости (двумерное рассеяние). Как видно на диаграмме (рис. 10), «облако» вариант вытянуто в направлении диагонали облака точек. Справа вверху находятся варианты с высокими значениями и размеров и массы тела, в левом нижнем углу – с наименьшими значениями. В центре расположены варианты с промежуточными, средними значениями.

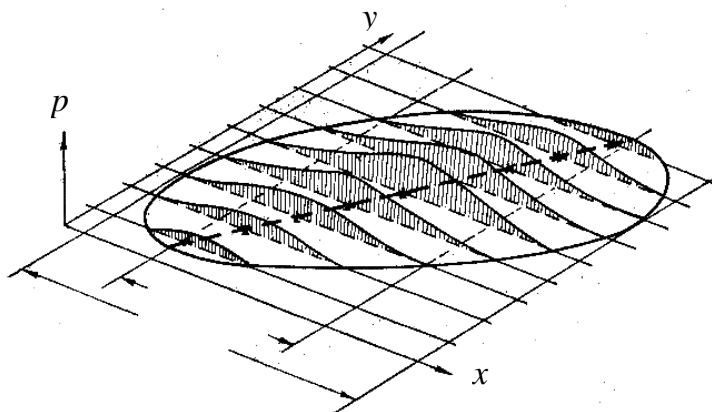


**Рис. 10.** Область рассеяния вариант

В первом приближении можно сказать, что *двумерное распределение* – это *ординация вариант на плоскости осей двух признаков*. Помимо рассеяния на плоскости, в определение двумерного распределения входит и частота встречаемости отдельных значений ( $a$ ). Если признаки  $x$  и  $y$  теоретически подчиняются нормальному закону, тогда скопление вариант в трех осях (оси признаков  $x$ ,  $y$  и частоты  $a$ ) образует весьма странный «гребень», растянутое в пространстве вы-

пуклое нормальное распределение (рис. 11). Однако в реальности такой идеальной картины получить никогда не удастся, приходится ориентироваться только на плоскую фигуру рассеяния немногочисленных вариантов. Если область, занятую вариантами, очертить по периферии плавной линией, мы получим вытянутую фигуру, эллипс, ограничивающий область рассеяния вариант, эллипс рассеяния. *Эллипс рассеяния – это область распространения вариант одной совокупности.*

Можно видеть, что в нашем примере признаки связаны друг с другом – есть общая тенденция: чем больше длина тела, тем больше вес; эта зависимость не очень жесткая, она размыта индивидуальными особенностями объектов (вариант).



*Рис. 11. Двумерное распределение*

В двумерном распределении проявляются два эффекта: синхронное изменение двух признаков и размывание этой синхронности, т. е. действие факторов сопряжения признаков вдоль оси эллипса и действие случайных факторов – поперек нее.

### 4.3 Корреляционный анализ

*Взаимная связь (взаимная зависимость) двух признаков при их изменчивости, т. е. сопряженность их вариации, называется корреляцией.* Корреляция имеет место в тех случаях, когда признаки изменяются не автономно, а согласованно. Если с увеличением одного признака происходит соответствующее увеличение другого, говорят о положительной корреляции, и коэффициент корреляции имеет в этом случае положительный знак (+). Если же по мере увеличения первого признака второй уменьшается, то это отрицательная корреляция, коэффициент корреляции пишется со знаком минус (–). Полная положительная корреляция выражается единицей  $r = 1$ , полная отрицательная  $r = -1$ . В природе такая ситуация встречается редко, и степень связи выражается той или иной долей единицы. При этом о тесной (сильной) корреляции обычно говорят в тех случаях, когда коэффициент корреляции не ниже  $\pm 0.6$ ; значения ниже  $\pm 0.6$  указывают на среднюю связь, а ниже  $\pm 0.3$  – на слабую.

Коэффициент корреляции призван численно выражать долю сопряженной вариации двух признаков в общей их вариации:

$$r = \sqrt{\frac{\text{ковариация}}{\text{изменчивость}}} = \frac{C_{xy}}{\sqrt{C_x \cdot C_y}} = \frac{\sum (y - M_y)(x - M_x)}{\sqrt{\sum (y - M_y) \cdot \sum (x - M_x)}},$$

где  $C_{xy}$  – характеристика сопряженной изменчивости признаков,  
 $C_x, C_y$  – характеристика общей изменчивости признаков.

При большом количестве данных коэффициент корреляции имеет смысл вычислять на компьютере (например, с помощью функции КОРРЕЛ в среде программы Excel), но для небольших выборок его можно быстро найти и при ручном счете. Рабочая формула для расчетов имеет вид:

$$r = \frac{C_{xy}}{\sqrt{C_x \cdot C_y}} = \frac{\sum xy - (\sum x \cdot \sum y) / n}{\sqrt{(\sum x^2 - (\sum x)^2 / n) \cdot (\sum y^2 - (\sum y)^2 / n)}}.$$

Способ вычисления коэффициента корреляции показан в таблице 13 на примере зависимости между живым весом коров ( $x$ ) и их приплода ( $y$ , кг). По таблице рассчитываются квадраты вариантов и их произведения, а также суммы вариантов, квадратов и произведений. Вычисления ведутся по точным рабочим формулам.

Таблица 13

$i$	$y$	$x$	$y^2$	$x^2$	$x \cdot y$
1	25	352	625	123904	8800
2	26	376	676	141376	9776
3	31	402	961	161604	12462
4	32	453	1024	205208	14496
5	34	484	1156	234256	16456
6	38	528	1444	278784	20064
7	38	555	1444	308025	21090
$\Sigma$	224	3150	7330	1453158	103144

Проведем последовательные расчеты. Сначала определим вспомогательные величины:

$$C_{xy} = \Sigma(x \cdot y) - (\Sigma x) \cdot (\Sigma y) / n = 103144 - 3150 \cdot 224 / 7 = 2344,$$

$$C_y = \Sigma y^2 - (\Sigma y)^2 / n = 7330 - 224^2 / 7 = 162,$$

$$C_x = \Sigma x^2 - (\Sigma x)^2 / n = 1453158 - 3150^2 / 7 = 35658;$$

затем – коэффициент корреляции:

$$r = \frac{C_{xy}}{\sqrt{C_x \cdot C_y}} = \frac{2344}{\sqrt{35658 \cdot 162}} = 0.975.$$

Далее найдем его ошибку:

$$m_r = \sqrt{\frac{1 - r^2}{n - 2}} = \sqrt{\frac{1 - 0.975^2}{7 - 2}} = 0.099,$$

и, наконец, критерий  $t$  Стьюдента для проверки значимости коэффициентов:

$$t_r = r / m_r = 0.975 / 0.099 = 9.84.$$

Нулевая гипотеза предполагает отсутствие связи: «коэффициент корреляции значимо от нуля не отличается»,  $r = 0$ . В нашем примере для уровня значи-

мости  $\alpha = 0.05$  и числа степеней свободы  $df = n - 2 = 5$  находим табличное значение критерия Стьюдента  $t_{(0.05, 5)} = 2.57$ . Полученная величина (9.84) значительно превышает табличную (2.57), что говорит о высокой статистической значимости коэффициента корреляции, о достоверности его отличия от нуля. Признаки *положительно* коррелируют, масса тела теленка действительно *возрастает* вслед за ростом массы тела коровы.

Выборный коэффициент корреляции в той или иной степени соответствует генеральному параметру. Определить диапазон, где лежит генеральное значение, можно с помощью доверительного интервала, хотя его *нельзя* построить непосредственно по формуле  $r \pm t_{(\alpha, df)} \cdot m_r$ . Дело в том, что область изменений коэффициента ограничена рамками  $\pm 1$ , поэтому распределение выборочных коэффициентов корреляции в общем не соответствует нормальному (с диапазоном изменчивости  $\pm \infty$ ). Поэтому перед расчетом коэффициент корреляции преобразуют в величину  $z$ , имеющую нормальное распределение, и уже для нее отыскивают границы доверительного интервала, после чего выполняют обратное преобразование.

Доверительный интервал для нашего случая ( $r = 0.975$ ,  $\alpha = 0.05$ ,  $n = 7$ ,  $df = n - 2 = 5$ ,  $t_{(0.05, 5)} = 2.57$ ) рассчитывается так. Преобразуем  $r$ :

$$z = 0.5 \cdot \ln \left( \frac{1+r}{1-r} \right) = 0.5 \cdot \ln \left( \frac{1+0.975}{1-0.975} \right) = 2.184$$

или берем его более точное значение из таблицы 13II, тогда  $z = 2.0923$ .

$$\text{Определяем ошибку } m_z = \sqrt{\frac{1}{n-3}} = \sqrt{\frac{1}{7-3}} = 0.5.$$

Находим верхнюю границу:  $\max z = z + t_{(\alpha, df)} \cdot m_z = 2.09 + 2.57 \cdot 0.5 = 3.375$  и нижнюю границу:  $\min z = z - t_{(\alpha, df)} \cdot m_z = 2.09 - 2.57 \cdot 0.5 = 0.805$ .

Обратное преобразование (по табл. 14II) дает:  $\max r \approx 1.00$ ,  $\min r \approx 0.67$ . Истинное значение коэффициента корреляции находится в диапазоне от 0.67 до 1.00.

### Ложная корреляция

Когда величина коэффициента корреляции определяется в первую очередь способом подбора вариант в выборку, а не реальной зависимостью между изучаемыми признаками, то говорят о «ложной корреляции».

Величина коэффициента корреляции зависит от вытянутости эллипса рассеяния: чем больше длина главной оси эллипса отличается от сечения, тем выше значение коэффициента. Случайные единичные, а тем более парные значения могут резко повысить показатель силы связи признаков. Особенно чувствителен коэффициент корреляции к нулям, которые могут попасть в исходную матрицу при переносе данных между электронными таблицами.

Явление ложной корреляции возникает и в том случае, когда исследуемые показатели имеют в сумме постоянное значение, например 100%. Рассмотрим соотношение численности грызунов и насекомоядных в разных биотопах (табл. 14). Представители и первого, и второго отрядов чаще встречаются в коренных хвойных лесах, нежели в антропогенных стациях, тем более в агроценозах. Синхронность их реакции на трансформацию ландшафтов выражается высоким ко-



эффицентом корреляции их численности  $r = 0.85$ .

Таблица 14

Биотоп	Численность (экз./100 конусо-суток)			Доля, $P$ (%)		
	бурозубок $N_{\text{б}}$	грызунов $N_{\text{г}}$	общая $N_{\text{о}}$	бурозубок $N_{\text{б}} / N_{\text{о}}$	грызунов $N_{\text{г}} / N_{\text{о}}$	общая $N_{\text{о}} / N_{\text{о}}$
Кедровник	25	29	54	0.46	0.54	1
Смешанный	25	32	57	0.44	0.56	1
Экотон	23	21	44	0.52	0.48	1
Сосняк	22	16	38	0.58	0.42	1
Березняк	20	23	43	0.47	0.53	1
Луг	10	9	19	0.53	0.47	1
$r$	0.85			-1.00		

Если же оценить зависимость между долей грызунов ( $P_{\text{г}} = N_{\text{г}}/N_{\text{о}}$ ) и долей бурозубок ( $P_{\text{б}} = N_{\text{б}}/N_{\text{о}}$ ) в этих стациях (между индексами доминирования), она составит  $r = -1.00$ . Дело в том, что эти показатели рассчитываются относительно общей суммы, поэтому доля полевок составляет разницу между 1 и долей бурозубок:  $P_{\text{г}} = 1 - P_{\text{б}}$ . По существу, мы имеем уравнение строго функциональной обратной регрессии ( $y = 1 - 1 \cdot x$ ), которому соответствует, естественно, максимальный отрицательный коэффициент корреляции. Требование неизменности суммы двух показателей (1 или 100%), принятое для вычисления процентов, оказывается причиной постоянной обратной пропорции между этими показателями. Такая корреляция должна быть названа ложной, потому что характеризует не биологическую зависимость показателей, а способ их расчета. Когда общую сумму образуют три и более признаков, ложная корреляция будет отличаться от  $r = -1$ , но от этого не утратит своей природы математического артефакта.

При обработке массивов данных с большим числом производных признаков (индексы доминирования видов в сообществе, морфологические индикаторы) нетрудно пропустить еще один вид ложной корреляции, которая наблюдается между двумя признаками, отнесенными к общей для них третьей переменной. По неосмотрительности коэффициенты связи между индексами могут быть восприняты как оценка зависимости между признаками. Такие корреляции, бессознательно наведенные третьим фактором, по сути являются ложными.

Безусловно, содержательную интерпретацию можно дать как корреляции признаков, так и корреляции индексов, но они будут кардинально отличаться. Например, среди нескольких видов куньих (от ласки до барсука) коэффициент корреляции между длиной тонкого и толстого отделов кишечника ( $r = 0.96$ ) отражает простые морфологические пропорции: у крупного животного кишечник длиннее, чем у мелкого. Однако корреляция между индексами этих органов (размеров, отнесенных к длине тела особи) характеризует уже отличия диеты

разных видов ( $r = 0.78$ ): кишечник относительно меньше у облигатных хищников, нежели у полифагов. Однако в большом массиве производных значений такие отношения между индексами могут восприниматься как зависимости между признаками, что неизбежно приведет к ложным выводам.

Чтобы уйти от подобной двусмысленности, к обработке желательно привлекать только предварительно выверенные реальные исходные показатели, а не связанные методом расчета доли, проценты или индексы.

#### 4.3.1 Множественная корреляция

Разобранные выше примеры корреляционных зависимостей касались главным образом взаимосвязи двух сопряженных процессов, явлений или варьирующих признаков. Между тем в практике биологических исследований нередко приходится сталкиваться с более сложными случаями, например, когда сопряжены не два, а три или более изменчивых фактора (признака). В такой ситуации возникает необходимость изучить множественные связи между большим числом взаимодействующих переменных, выступающих как в виде целой системы коррелированных признаков организма, так и в форме совместного влияния сложной совокупности факторов на определенное явление. Корреляционная зависимость нескольких переменных носит название множественной корреляции и оценивается коэффициентом, определяемым на основе корреляций между всеми парами признаков. Например, коэффициент множественной корреляции между тремя признаками  $A$ ,  $B$  и  $C$  вычисляется по формуле:

$$r_{A.BC} = \sqrt{\frac{r_{AB}^2 + r_{AC}^2 - 2 \cdot r_{AB} \cdot r_{AC} \cdot r_{BC}}{1 - r_{AB}^2}}.$$

Полученная величина характеризует связь первого признака ( $A$ ) с двумя другими ( $B$  и  $C$ ). Покажем этот способ на примере совокупного действия двух факторов,  $B$  и  $C$  (температуры и влажности), на суточную активность травяных лягушек ( $A$ ). Определение парных корреляций дало следующие результаты ( $n = 110$ ):  $r_{AB} = +0.58$ ;  $r_{AC} = +0.80$ ;  $r_{BC} = -0.45$ . Отсюда

$$r_{A.BC} = \sqrt{\frac{0.58^2 + 0.8^2 - 2 \cdot 0.58 \cdot 0.8 \cdot 0.45}{1 - 0.45^2}} = 0.86.$$

Сводный коэффициент корреляции оказался довольно высоким и, как показывает его сопоставление со стандартным значением по таблице 15II, вполне достоверным (при  $\alpha < 0.001$ ).

С другой стороны, если обнаружена значительная корреляция между признаками  $A$  и  $C$  и между  $B$  и  $C$ , то не исключена возможность мнимой корреляционной зависимости между  $A$  и  $B$ , которая создается за счет одновременного влияния на них третьего признака  $C$ . Например, установленная по исследованиям в Карелии корреляция между численностью лесных полевок и урожаем семян сосны, скорее всего, объясняется не значением последних в питании грызунов (т. е. прямой причинной связью), а тем, что оба эти явления (численность полевок и урожай семян) контролируются одними и теми же экологическими факторами (прежде всего метеорологическими) и поэтому изменяются параллельно,

хотя непосредственно между собой не связаны.

В этом и подобных случаях (например, когда настоящие зависимости между признаками животных маскируются влиянием возраста или когда связи между отдельными промерами организма создаются за счет влияния живого веса и т. д.) возникает задача изучить корреляцию между двумя признаками ( $A$  и  $B$ ), исключив влияние на эту связь третьего признака ( $C$ ), как бы элиминировав его.

#### 4.3.2 Частная корреляция

Этой цели служит коэффициент частной корреляции, оценивающий связь между первым и вторым признаками при постоянных значениях третьего и вычисляемый по формуле:

$$r_{A(BC)} = \frac{r_{AB} - r_{AC} \cdot r_{BC}}{\sqrt{(1 - r_{AC}^2) \cdot (1 - r_{BC}^2)}},$$

где  $A$  и  $B$  – факторы, связь которых требуется изучить;

$C$  – фактор, влияние которого необходимо исключить из корреляционной зависимости между  $A$  и  $B$  (реперный признак);

$r_{AB}$ ,  $r_{AC}$ ,  $r_{BC}$  – соответствующие парные коэффициенты корреляции, вычисляемые обычным способом;

$r_{A(BC)}$  – искомый коэффициент частной корреляции, показывающий связь между двумя признаками при исключении влияния третьего.

Этот же метод можно применить и для элиминации двух факторов при четырех переменных и т. д. Формула для расчетов примет в этом случае следующий вид:

$$r_{AB(BD)} = \frac{r_{AB(C)} - r_{AC(B)} \cdot r_{BC(D)}}{\sqrt{(1 - r_{AC(D)}^2) \cdot (1 - r_{BC(D)}^2)}}.$$

Рассмотрим нахождение коэффициента частной корреляции на упрощенном примере (взятом из книги П. Ф. Рокицкого). Получены данные о корреляции между давлением крови ( $A$ ), содержанием в ней холестерина ( $B$ ) и возрастом ( $C$ ) у 142 женщин. Соответствующие коэффициенты корреляции таковы:  $r_{AB} = +0.25$ ;  $r_{AC} = +0.33$ ;  $r_{BC} = 0.51$ .

Известно, что повышенное артериальное давление может быть связано с высоким содержанием холестерина в стенках кровеносных сосудов, однако и давление крови, и концентрации холестерина увеличиваются с возрастом. Поэтому возникает вопрос, создается ли корреляция между давлением крови и содержанием в ней холестерина за счет их общей связи с возрастом или же она реально существует для каждого возраста (и независимо от него). Элиминируя эффект возраста по приведенной выше формуле, получим:

$$r_{A(BC)} = \frac{0.25 - 0.33 \cdot 0.51}{\sqrt{(1 - 0.33^2) \cdot (1 - 0.5^2)}} = 0.12.$$

По таблице 15/II можно установить, что при  $n = 150$  для достоверности коэффициента корреляции даже при уровне значимости  $\alpha = 0.05$  его величина

должна быть не меньше 0.159. В данном же случае полученное значение меньше табличного и, следовательно, коэффициент корреляции от нуля достоверно не отличается. Таким образом, внутри отдельных возрастных групп корреляционной связи между давлением крови и содержанием холестерина, по крайней мере на изученном материале, не обнаруживается. Пока нет оснований отбрасывать нулевую гипотезу.

Второй пример демонстрирует использование коэффициента частной корреляции для более глубокого проникновения в структуру нескольких факторов наведения. Рассмотрим выборку объектов разного статуса (11 видов мелких млекопитающих), взяв в качестве признаков их численность в семи биотопах прибайкальской равнины. Реперным признаком послужила суммарная численность вида во всех биотопах. Здесь коэффициент корреляции отражает сходство между биотопами по соотношениям численности 11 видов. Например, оказалось, что между березняком и экотоном (граница между березняком и коренными лесами) и общая корреляция ( $r = 0.92$ ), и частная ( $r = 0.64$ ) высока и положительна. Можно утверждать, что население животных этих биотопов почти идентично.

В свою очередь, корреляция между кедровником и лугом не проявилась ( $r = -0.08$ ), но коэффициент частной корреляции был велик и отрицателен ( $r = -0.43$ ). Этим оттеняется тот факт, что виды, отсутствующие на лугу, многочисленны в кедровнике (красная полевка, мышь), а обычные в агроценозе – крайне редки в тайге (серые полевки).

Частная корреляция показала, что население этих биотопов во многом диаметрально противоположно. Она выявила два вида факторов наведения. Один из них хорошо известен – это сезонное расселение видов в другие биотопы. В течение периода размножения видовой состав тайги и луга меняется несогласованно (одни виды идут из тайги в агроценозы, другие – в противоположном направлении) и численность всех видов относительно выравнивается,  $r = -0.08$ . Частная корреляция устраняет эффект прироста численности за счет иммигрантов и выдвигает на первый план контраст «базовой» численности, которую формируют характерные обитатели биотопов: в тайге это лесные полевки, на лугу – серые. Так *проявляется* второй фактор: отличие качества среды в разных биотопах. Он обеспечивает формирование принципиально несходных зооценозов, что и показывает высокой частной корреляцией  $r = -0.43$ .

### 4.3.3 Ранговая корреляция

Помимо рассмотренных выше параметрических показателей связи в биометрии применяются и непараметрические. Обычно их используют при сильных отклонениях изучаемого распределения от нормального (или сомнениях на этот счет), а также в тех случаях, когда требуется оценить зависимость между качественными или полуколичественными признаками, точное количественное измерение которых затруднено (оценки в баллах или других условных единицах). Если варианты выборки могут быть упорядочены по степени выраженности их свойств, для измерения степени сопряженности между ними можно воспользо-

ваться *непараметрическим показателем связи* – ранговым коэффициентом корреляции Спирмена:

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

где  $d$  – разность между рангами сопряженных значений признаков  $x$  и  $y$ ;  
 $n$  – объем выборки.

Этой формулой следует пользоваться в тех случаях, когда выборки не содержат повторяющихся вариантов, когда все ранги выражены разными целыми числами. Если же исходные ряды содержат одинаковые значения, расчет корреляции приходится вести по другой формуле, включающей поправку на повторы (при этом одинаковым вариантам присваивается средний ранг):

$$r_s = \frac{\frac{(n^3 - n)}{6} - (T_x + T_y) - \sum d^2}{\sqrt{\left(\frac{(n^3 - n)}{6} - 2 \cdot T_x\right) \left(\frac{(n^3 - n)}{6} - 2 \cdot T_y\right)}},$$

где  $T_x, T_y$  – поправки на серии повторов для каждой выборки:

$$T_x = \frac{\sum_{k=1}^k (t_x^3 - t_x)}{12},$$

где  $t$  – число членов в каждой группе одинаковых вариантов.

Поправки  $T_x, T_y$  учитывают  $k$  групп повторяющихся вариантов.

Рассмотрим технику вычислений на примере изучения связи между оцененными в баллах численностью лисицы ( $x$ ) и обилием мышевидных грызунов ( $y$ ) (по годам наблюдений):

	1957	1958	1959	1960	1961	1962	1963	1964	1965	1966
$x$	2.6	2.1	2.3	2.3	1.6	2.2	3.0	2.1	1.5	2.2
$y$	3.0	2.4	3.6	2.9	3.7	3.3	4.0	2.1	1.0	3.5

Чтобы проверить наличие и определить силу этой связи, нужно упорядочить значения сопряженных признаков по степени их выраженности, затем присвоить им ранги, обозначив значения порядковыми числами натурального ряда, и рассчитать коэффициент корреляции. Техника вычислений показана в таблице 15.

Таблица 15

Численность лисицы в баллах, $x$	Обилие гры- зунов в баллах, $y$	Ранги вариант		Разность между рангами, $d$	$d^2$
		$R_x$	$R_y$		
1.5	1.0	1	1	0	0
1.6	3.7	2	6	-4.0	16.00
2.1	2.4	3.5	3	+0.5	0.25
2.1	2.1	3.5	2	+1.5	2.25
2.2	3.3	5.5	7	-1.5	2.25
2.2	3.6	5.5	8.5	-3.0	9.00
2.3	3.6	7.5	8.5	-1.0	1.00
2.3	2.9	7.5	4	+3.5	12.25
2.6	3.0	9	5	+4.0	16.00
3.0	4.0	10	10	0	0
					$\Sigma = 59$

В ряду значений признака  $x$  есть три пары одинаковых вариантов, поэтому поправка будет равна:  $T_x = \frac{(2^3 - 2) + (2^3 - 2) + (2^3 - 2)}{12} = 1.5$ .

В ряду признака  $y$  всего одна пара одинаковых значений; поправка составит:  $T_y = \frac{(2^3 - 2)}{12} = 0.5$ .

Находим величину  $\frac{(n^3 - n)}{6} = \frac{(10^3 - 10)}{6} = 165$ .

Коэффициент ранговой корреляции составит:

$$r_s = \frac{165 - (1.5 + 0.5) - 59}{\sqrt{(165 - 2 \cdot 1.5)(165 - 2 \cdot 0.5)}} = 0.638.$$

Если воспользоваться формулой без поправок, результат будет несколько иным:

$$r_s = 1 - \frac{6 \cdot \sum d^2}{n \cdot (n^2 - 1)} = 1 - \frac{6 \cdot 59}{10 \cdot (10^2 - 1)} = 0.642.$$

Статистическая ошибка и критерий достоверности отличия коэффициента корреляции от нуля вычисляются по формулам:

$$m_r = \sqrt{\frac{1 - r_s^2}{n - 2}} = \sqrt{\frac{1 - 0.638^2}{10 - 2}} = 0.272,$$

$$t_r = r_s / m_r = 0.638 / 0.272 = 2.34.$$

Величина критерия (2.34) несколько выше критического значения (2.31) для уровня значимости  $\alpha = 0.05$  и числа степеней свободы  $df = n - 2 = 8$  (табл.

6П). Казалось бы, это дает основание отвергнуть нулевую гипотезу ( $r_s = 0$ ) и с вероятностью  $P = 95\%$  констатировать достоверность установленной связи. Однако при небольших выборках статистические свойства коэффициента Спирмена не очень «хороши» и для оценки значимости корреляции лучше воспользоваться специально подготовленной таблицей 16П, аналогичной рассмотренной выше таблице 15П.

Чтобы полученный коэффициент можно было считать достоверно отличным от нуля, он должен превышать табличное значение при данном  $n$ . В нашем случае ( $n = 10$ ,  $\alpha = 0.05$ ) коэффициент  $r = 0.638$  ниже табличного  $r = 0.64$ , следовательно, значимо от нуля не отличается. Зависимость численности лисицы и грызунов по приведенным данным достоверно не прослеживается.

### 4.3.3 Коэффициент контингенции

Степень сопряженности (сочетаемость) двух возможных состояний двух качественных признаков также можно измерить с помощью особого коэффициента корреляции – коэффициента контингенции Шарлье.

У каждой особи отмечают два признака, имеющих альтернативные распределения, и вся выборка разбивается на четыре части:

$a$  – число особей, имеющих оба признака (+ +),

$b$  – число особей, имеющих первый признак, но не имеющих второго (+ –),

$c$  – число особей, не имеющих первого признака, но имеющих второй (– +),

$d$  – число особей, не имеющих обоих признаков (– –).

На схеме это выглядит как четырехклеточная корреляционная решетка:

Признак 1 \ Признак 2	Присутствует (+)	Отсутствует (–)	$\Sigma$
Присутствует (+)	$a$	$c$	$a + c$
Отсутствует (–)	$b$	$d$	$b + d$
$\Sigma$	$a + b$	$c + d$	$n = a + b + c + d$

Степень взаимосвязи определяется по формуле:

$$r = \frac{a \cdot d - b \cdot c}{\sqrt{(a + b) \cdot (c + d) \cdot (a + c) \cdot (b + d)}}.$$

При вычислении коэффициента корреляции между двумя альтернативными признаками выясняется вопрос о том, чаще ли оба признака одновременно присутствуют или отсутствуют у варианты, чем это могло бы быть по случайным причинам. Достоверность отличия от нуля оценивается по критерию Стьюдента:

$$t_r = r / m_r, \text{ где } m_r = \frac{1 - r^2}{\sqrt{n - 1}}.$$

При проверке влияния перекрытий на оплодотворяемость самок песцов получены первичные материалы о численности родивших (+) и неродивших (–) самок из числа хотя бы дважды перекрытых (+) и неперекрываемых (–).

Признак 1 \ Признак 2	Родившие (+)	Неродившие (-)	$\Sigma$
Перекрытые (+)	370	90	460
Неперекрытые (-)	100	120	220
$\Sigma$	470	210	$n = 680$

Коэффициент ассоциации равен:  $r = \frac{370 \cdot 120 - 100 \cdot 90}{\sqrt{470 \cdot 210 \cdot 460 \cdot 220}} = 0.35$ .

Ошибка коэффициента корреляции составит:  $m_r = \frac{1 - 0.35^2}{\sqrt{680 - 1}} = 0.0327$ ,

а критерий Стьюдента  $t_r = 0.35 / 0.0327 = 10.7$ .

Полученное значение (10.7) настолько велико, что превышает табличное даже для доверительной вероятности выше  $P = 0.999$  (уровень значимости  $\alpha < 0.001$ ). Влияние повторных покрытий на оплодотворяемость самок песцов несомненно.

При исследовании связи между белой мастью и красными глазами у кроликов получены следующие данные.

Шерсть \ Глаза	Красные	Некрасные	$\Sigma$
Белая	29	11	40
Окрашенная	1	59	60
$\Sigma$	30	70	100

Подстановка всех значений сумм из таблицы в формулы дает:  $r = 0.76$ ,  $m = 0.04$ ,  $t = 19$ . Достоверность связи не вызывает сомнений.

#### 4.4 Регрессионный анализ

Коэффициент корреляции указывает лишь на степень (тесноту) связи в изменчивости двух переменных величин, но не позволяет судить о том, как меняется одна величина по мере изменения другой. Ответ на этот вопрос дает вычисление коэффициента регрессии, показывающего, на какую величину в среднем изменяется один признак при изменении другого на единицу измерения. Регрессионный анализ, в отличие от корреляционного, изучает эффект *влияния одного признака на другой*, зависимость признака от фактора, характер влияния фактора на признак. Его основные результаты таковы:

1. Таблица дисперсионного анализа, в которой показана сила и достоверность влияния на признак изучаемого фактора или другого признака.
2. Уравнение регрессии, выражающее пропорциональность сопряженного изменения признаков, тенденции их взаимосвязанной изменчивости или динамики.
3. Оценки значимости коэффициентов уравнения регрессии.

Регрессионный анализ методически ориентирован односторонне – на изу-



чение зависимости одного признака от другого (зависимость  $y$  от  $x$  или, напротив, зависимость  $x$  от  $y$ ), хотя может применяться к случаям, когда фактически имеется взаимозависимость двух переменных.

Основную тенденцию взаимосвязанного изменения двух признаков можно отобразить с помощью простого графического приема. Разобьем ось  $x$  на несколько интервалов. Найдем для каждого из них частные средние значения признака  $y$  ( $M_y$ ). Теперь проведем через эти средние точки ломаную линию. Это будет линия регрессии  $Y$  по  $x$ . Регрессия – изменение среднего уровня одного признака при изменении другого (рис. 12).

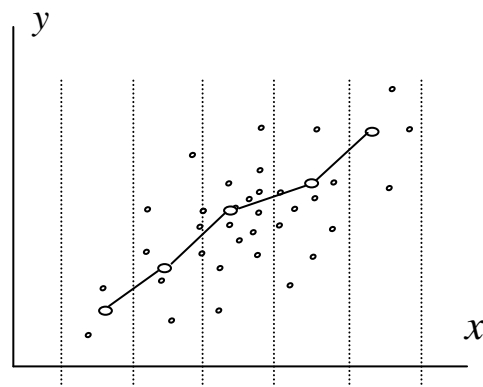


Рис. 12. Эмпирическая линия регрессии

#### 4.4.1 Линейная регрессия

К сожалению, ход ломаной линии нельзя передать простым уравнением, к тому же на нем сказывается способ интервального разбиения оси абсцисс, а также уровень репрезентативности в разных областях распределения. В этом смысле предпочтительнее единственная прямая линия регрессии, подчеркивающая основные тенденции зависимости признаков, которая может быть выражена простым уравнением линии:  $y = ax + b$ .

Судить о том, как меняется одна величина по мере изменения другой, позволяет коэффициент регрессии ( $a$ ), показывающий, на какую величину в среднем изменяется один признак ( $y$ ) при изменении другого ( $x$ ) на единицу измерения (точнее, на какую величину один признак отклоняется от своей средней при некотором отклонении другого признака от своей средней):

$$y - My = a \cdot (x - Mx).$$

Простые преобразования:

$$y = a \cdot x + My - a \cdot Mx, b = My - a \cdot Mx$$

и приводят к уравнению линии:  $y = ax + b$ .

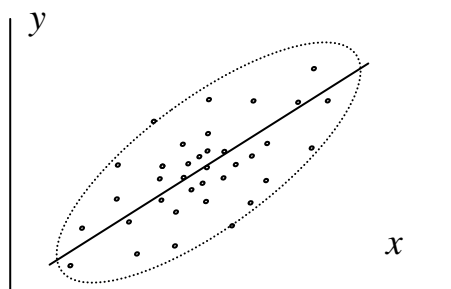


Рис. 13. Линейная регрессия

Рассчитать коэффициенты уравнения регрессии позволяет метод

наименьших квадратов, основная идея которого состоит в том, чтобы линия регрессии прошла на наименьшем удалении от каждой точки, т. е. чтобы сумма квадратов расстояний от всех точек до прямой линии была наименьшей. В математической статистике показано, что для случая двумерного нормального распределения лучшей (эффективной, несмещенной и пр.) линией, описывающей зависимость одного признака от другого, может быть только линия частных средних арифметических.

Вычисления коэффициентов линейной регрессии  $y = ax + b$  ведутся по следующему алгоритму. Сначала найдем вспомогательные величины:

$$C_x = \sum x^2 - (\sum x)^2 / n,$$

$$C_y = \sum y^2 - (\sum y)^2 / n,$$

$$C_{xy} = \sum (x \cdot y) - (\sum x) \cdot (\sum y) / n,$$

$$M_y = \sum y / n, M_x = \sum x / n.$$

Затем рассчитаем коэффициенты:  $a = C_{xy} / C_x$ ,  $b = M_y - a \cdot M_x$ .

Оценить значимость коэффициента регрессии позволяет критерий  $t$  Стьюдента, проверяющий нулевую гипотезу Но:  $a = 0$ , коэффициент регрессии значимо от нуля не отличается. С этой целью рассчитывается ошибка коэффициента регрессии  $m_a$ :

$$m_a = \frac{S_y}{S_x} \cdot m_r, \text{ где } m_r - \text{ошибка коэффициента корреляции (см. с. 62),}$$

и вычисляется значение критерия:

$$t = (a - 0) / m_a = a / m_a \sim t_{(0.05, n-2)}.$$

Смысл этого критерия состоит в следующем. Коэффициент регрессии  $a$  характеризует сопряженность пропорционального изменения двух признаков, т. е. отвечает за то, что линия регрессии имеет некоторый угол относительно оси абсцисс. Значение  $a = 0$  означает, что линия регрессии идет параллельно оси ОХ, что при изменении признака  $x$  признак  $y$  не меняется, т. е. что  $y$  не зависит от  $x$ . Значения коэффициента, отличные от нуля, говорят о том, что взаимосвязь признаков имеет место, при  $a > 0$  зависимость положительная, при  $a < 0$  – отрицательная.

Вернемся к примеру с описанием зависимости между живым весом коров и их приплода (стр. 61). Расчеты для построения уравнения регрессии показаны в таблице 16. Сначала вычисляются квадраты вариантов и их произведения, а также суммы вариантов, квадратов и произведений. Вычисления ведутся по точным рабочим формулам. Проще всего это делать в среде Excel, с помощью команды Сервис \ Анализ данных \ Регрессия.

Таблица 16

$i$	$y$	$x$	$y^2$	$x^2$	$x \cdot y$	$Y$	$(y - Y_i)^2$	$t \cdot m_y$	$\min Y$	$\max Y$
1	25	352	625	123904	8800	25.6	0.31	2.0	23.6	27.5
2	26	376	676	141376	9776	27.1	1.29	1.7	25.5	28.8
3	31	402	961	161604	12462	28.8	4.65	1.4	27.4	30.2
4	32	453	1024	205208	14496	32.2	0.04	1.2	31.0	33.4
5	34	484	1156	234256	16456	34.2	0.06	1.3	32.9	35.5
6	38	528	1444	278784	20064	37.1	0.76	1.7	35.4	38.9
7	38	555	1444	308025	21090	38.9	0.81	2.1	36.8	41.0
$\Sigma$	224	3150	7330	1453158	103144		7.92			

Проведем последовательные расчеты вручную. Сначала определим вспомогательные величины:

$$n = 7,$$

$$C_{xy} = \Sigma(x \cdot y) - (\Sigma x) \cdot (\Sigma y) / n = 103144 - 3150 \cdot 224 / 7 = 2344,$$

$$C_y = \Sigma y^2 - (\Sigma y)^2 / n = 7330 - 224^2 / 7 = 162,$$

$$C_x = \Sigma x^2 - (\Sigma x)^2 / n = 1453158 - 3150^2 / 7 = 35658,$$

затем – параметры:

$$M_y = \Sigma y / n = 224 / 7 = 32,$$

$$M_x = \Sigma x / n = 3150 / 7 = 450,$$

$$S_y = \sqrt{\frac{C_y}{n-1}} = \sqrt{\frac{162}{6}} = 5.2,$$

$$S_x = \sqrt{\frac{C_x}{n-1}} = \sqrt{\frac{35658}{6}} = 77.1,$$

$$r = \frac{C_{xy}}{\sqrt{C_x \cdot C_y}} = \frac{2344}{\sqrt{35658 \cdot 162}} = 0.975,$$

$$a = \frac{C_{xy}}{C_x} = \frac{2344}{35658} = 0.0657,$$

$$b = M_y - a \cdot M_x = 32 - 0.0657 \cdot 450 = 2.419.$$

Получено уравнение линейной регрессии  $Y = 0.0657x + 2.419$ , которое позволяет рассчитать теоретические значения  $Y$  (табл. 16, графа 7).

Далее найдем ошибку коэффициента регрессии:

$$m_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-0.975^2}{7-2}} = 0.099,$$

$$m_a = \frac{S_y}{S_x} \cdot m_r = \frac{5.2}{77.1} \cdot 0.099 = 0.00667,$$

и, наконец, критерий  $t$  Стьюдента для проверки значимости коэффициента регрессии:  $t_a = a / m_a = 0.0657 / 0.00667 = 9.84$ .

Для уровня значимости  $\alpha = 0.05$  и числа степеней свободы  $df = n - 2 = 5$  находим табличное значение критерия Стьюдента  $t_{(0.05,5)} = 2.57$ . Полученная величина (9.84) превышает табличную (2.57), что говорит о статистической значимости коэффициента регрессии ( $a$ ), о достоверности его отличия от нуля. Масса

тела теленка действительно возрастает вслед за ростом массы тела коровы.

Рассчитаем *доверительную зону* (интервал), в которой с той или иной *вероятностью* заключены *теоретические средние значения* веса новорожденных. Критерий Стьюдента (нормированное отклонение) для уровня значимости  $\alpha = 0.05$ , и числа степеней свободы  $df = n - 1 = 6$  составит 2.45. Далее находим границы. Так, для значения  $x = 352$  кг прогноз по уравнению регрессии равен  $Y = 25.56$ , а возможное отклонение средней составит:

$$t \cdot m_Y = t \cdot m_y \cdot \sqrt{\frac{1}{n} + \frac{(x_i - M_x)^2}{C_x}} = 2.45 \cdot 1.2582 \cdot \sqrt{\frac{1}{7} + \frac{(352 - 450)^2}{35658}} = \\ = 2.45 \cdot 0.81 = 1.98.$$

Отсюда находим границу доверительного интервала (табл. 16):

верхнюю:  $\max Y = Y_i + t \cdot m_Y = 25.56 + 1.98 = 27.54$

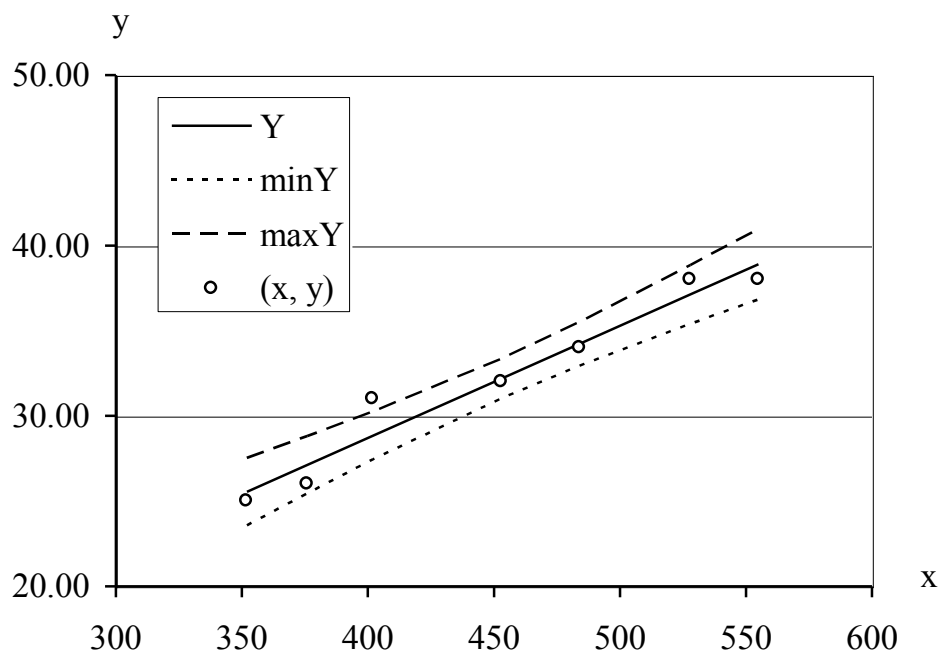
и нижнюю:  $\min Y = Y_i - t \cdot m_Y = 25.56 - 1.98 = 23.58$ .

Средняя масса новорожденного теленка для коров весом 352 кг с вероятностью  $P = 0.95$  должна находиться в диапазоне от 23.6 до 27.5 кг (рис. 14).

Регрессионный анализ позволяет проверить *значимость* и второго коэффициента уравнения регрессии, *свободного члена*  $b$ . Математический смысл свободного члена уравнения линии состоит в том, что этому значению равна функция ( $y$ ) при условии, что аргумент равен нулю ( $x = 0$ ):

$$y = ax + b = a \cdot 0 + b = b.$$

В рамках регрессионного анализа рассматривается именно эта гипотеза Но:  $b = 0$ , т. е. что линия регрессии проходит через начало осей координат, точку пересечения осей координат, через нуль. Если гипотеза опровергается, значит, линия регрессии не пересекает ось ординат. Если гипотеза не опровергается, мы можем считать, что между признаками существует простая пропорция ( $Y = ax$ ) и расчет коэффициента регрессии  $a$  упрощается:  $a = \Sigma(x \cdot y) / \Sigma x^2$ . Нулевая гипотеза Но:  $b = 0$  проверяется по критерию Стьюдента:  $t = (b - 0) / m_b = b / m_b \sim t_{(0.05, n-2)}$ , где  $m_b$  – ошибка коэффициента  $b$ .



**Рис. 14.** Линия регрессии  $Y = 0.0657 \cdot x + 2.1347$  и ее доверительный интервал

Ошибка второго коэффициента регрессии рассчитывается в два этапа. Сначала находим общую ошибку регрессионной средней (или остаточное стандартное отклонение), которая может вычисляться по-разному.

Точная формула для *небольших выборок* дает величину:

$$m_y = S_y \cdot \sqrt{\frac{(n-1) \cdot (1-r^2)}{n-2}} = 5.2 \cdot \sqrt{\frac{(7-1) \cdot (1-0.975^2)}{7-2}} = 1.2582.$$

Общая точная формула показывает практически такой же результат:

$$m_y = \sqrt{\frac{\sum_{i=1}^n (y_i - Y_i)^2}{n-2}} = \sqrt{\frac{C_{остат.}}{n-2}} = \sqrt{S_{остат.}^2} = \sqrt{\frac{7.92}{5}} = \sqrt{1.5832} = 1.2582$$

(величина  $C_{остат.} = \sum_{i=1}^n (y_i - Y_i)^2$  – это сумма квадратов разности между расчетными и реальными значениями признака, она найдена в табл. 16, внизу 7 графы,  $C_{остат.} = 7.92$ ).

Теперь вычисляем ошибку коэффициента  $b$ :

$$m_b = m_y \cdot \sqrt{\frac{1}{n} + \left(\frac{M_x}{C_x}\right)^2} = 1.2582 \cdot \sqrt{\frac{1}{7} + \left(\frac{450}{35658}\right)^2} = 3.0359$$

и критерий  $t$  Стьюдента:  $t_b = b / m_b = 2.419 / 3.0359 = 0.797$ .

Для уровня значимости  $\alpha = 0.05$  и числа степеней свободы  $df = n - 2 = 5$  табличное значение составляет  $t_{(0.05, 5)} = 2.57$ . Анализ показал, что критерий Стьюдента для свободного члена уравнения (0.797) оказался ниже табличного значения (2.57), т. е. коэффициент  $b$  значимо от нуля не отличается (при данном объеме собранных материалов). Это позволяет пересчитать коэффициент ре-

грессии:  $a = \Sigma(x \cdot y) / \Sigma x^2 = 0.071$ . Теперь можно пользоваться уравнением регрессии вида:  $Y = 0.071 \cdot x$ .

Оценить *достоверности взаимодействия признаков* можно и с помощью дисперсионного анализа (табл. 17). В этом случае общая дисперсия зависимого признака  $y$  ( $C_{общ.}$ ) разлагается на две составляющие – регрессионную дисперсию (изменчивость признака  $y$ , связанная с влиянием признака  $x$ ,  $C_{регр.}$ ) и случайную, или остаточную, дисперсию (изменчивость признака  $y$ , связанная с влиянием неучтенных случайных факторов,  $C_{остат.}$  (рис. 14, табл. 17, 18).

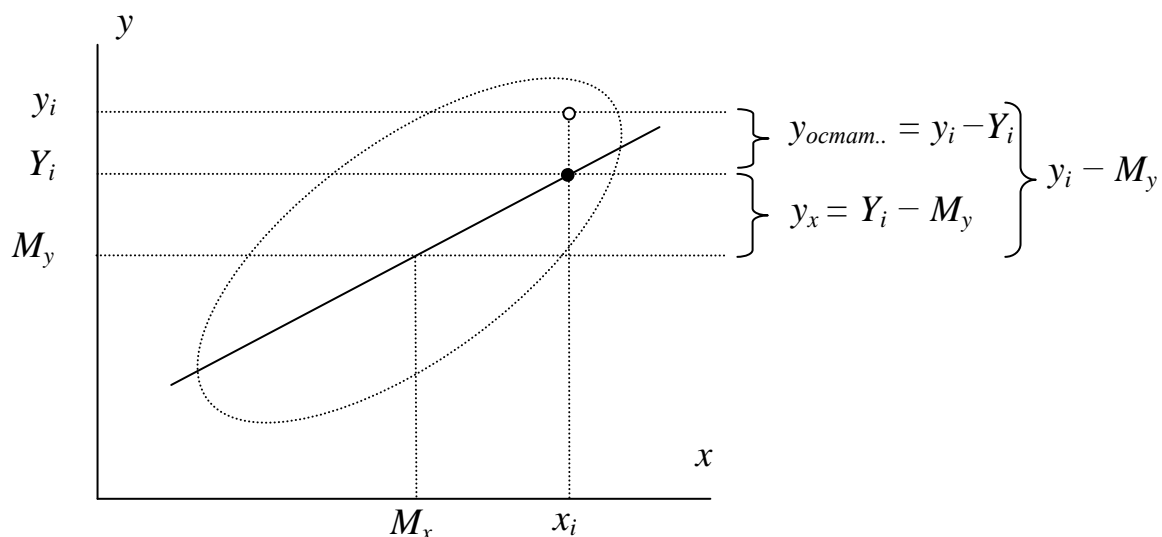
Таблица 17

Составляющие дисперсии	Суммы квадратов, $C$	Формулы расчета сумм квадратов	$df$	$S^2$	$F$
Регрессия	$C_{регр.} = \Sigma(Y_i - M_y)^2$	$C_{общ.} - C_{остат.}$	1	$S^2_{регр.} = \frac{C_{регр.}}{df_{регр.}}$	$\frac{S^2_{регр.}}{S^2_{остат.}}$
Отклонения вариант от линии регрессии	$C_{остат.} = \Sigma(y_i - Y_i)^2$		$n - 2$	$S^2_{остат.} = \frac{C_{остат.}}{df_{остат.}}$	$F_{(0.05, 1, n-2)}$
Общая (всего)	$C_{общ.} = \Sigma(y_i - M_y)^2$	$(\Sigma y_i^2 - \Sigma y_i)^2 / n = C_y$			

Таблица 18

Составляющие дисперсии	$C$		$df$	$S^2$	$F$
Регрессия	$C_{регр.} = \Sigma (Y_i - Y)^2$	154.08	1	$S^2_{регр.} = 154.08$	$F = \frac{154.08}{1.58} = 97.3$
Отклонения вариант от линии регрессии	$C_{остат.} = \Sigma (y_i - Y_{xi})^2$	7.92	5	$S^2_{остат.} = 1.58$	$F_{(0.05, 1, 5)} = 6.6$
Общая (всего)	$C_{общ.} = \Sigma (y_i - Y)^2$	162			

Общую сумму квадратов ( $C_{общ.} = C_y = \Sigma(y_i - M_y)^2 = \Sigma y_i^2 - (\Sigma y_i)^2 / n$ ) находят непосредственно как сумму квадратов отличий между значением  $y_i$  для каждой варианты и общей средней признака  $y$ . Остаточную сумму квадратов ( $C_{остат.} = \Sigma(y_i - Y_i)^2$ ) находят также непосредственно как сумму квадратов отличий между значением  $y_i$  для каждой варианты и значением, предварительно рассчитанным по уравнению регрессии  $Y_i = ax_i + b$  (для соответствующих значений  $x_i$ ). Модельную сумму квадратов ( $C_{мод.} = \Sigma(Y_i - M_y)^2$ ) рассчитывают как разность между общей и остаточной ( $C_{мод.} = C_{общ.} - C_{остат.}$ ).



**Рис. 15.** Модель варианты в регрессионном анализе

Показателем «силы влияния признака на признак» служит *коэффициент детерминации*, отношение регрессионной суммы квадратов к общей сумме квадратов (принимает значения от 0 до 1):  $R^2 = \frac{C_{\text{мод.}}}{C_{\text{общ.}}} = \frac{154.08}{162} = 0.95$ . Между ко-

эффициентом детерминации и коэффициентом корреляции существует простое соответствие:  $r = \sqrt{R} = \sqrt{0.95} = 0.975$ .

Построив таблицу дисперсионного анализа с помощью критерия Фишера можно проверить нулевую гипотезу Но: предсказания регрессионной модели в целом неадекватно описывают исходные данные, зависимости между признаками нет. Конструкция критерия исследует вопрос, превышает ли варьирование, учтенное моделью, случайное (остаточное) варьирование? Критерий Фишера вычисляется как отношение модельной и остаточной дисперсии:

$$F = S^2_{\text{мод.}} / S^2_{\text{остат.}} = 154.08 / 1.58 = 97.3.$$

Табличное значение  $F_{(0.05, 1, 5)} = 6.6$ . Поскольку полученное значение критерия оказалось выше табличного, дисперсия реального признака у приближается по величине к дисперсии расчетных значений признака  $Y$ , т. е. существенно превышает (случайные) отличия между ними. Регрессионная модель в целом адекватно описывает исходные данные.

#### 4.4.2 Криволинейная регрессия

В большинстве случаев связь биологических признаков не бывает линейной, они изменяются либо с разной скоростью, либо в разных масштабах. Соответственно на графике форма такой связи отображается не прямой, а кривой линией. Примерами могут служить геометрическая прогрессия роста численности популяции в оптимальных условиях, различие скоростей роста разных частей тела, определяющее аллометрический характер зависимости признаков (лицевой

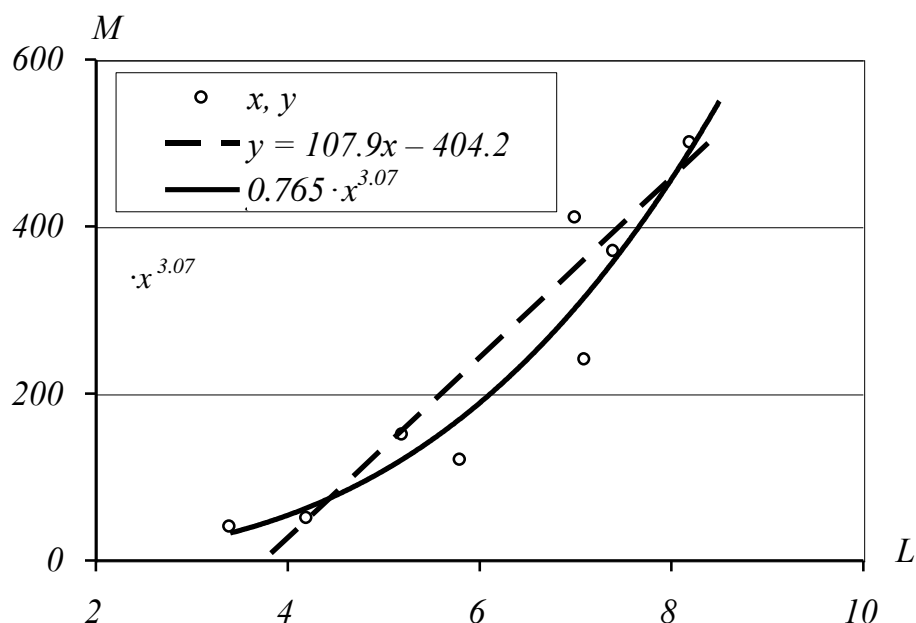
отдел черепа растет более интенсивно, чем мозговой). В подобных случаях эффективнее использовать не уравнения прямой линии ( $y = ax + b$ ), а разнообразные уравнения кривых линий, например, степенной, гиперболической, экспоненциальной, параболической, логистической и др.

Поскольку метод наименьших квадратов исходно ориентирован на линию (поиск уравнения линии, наименее удаленной от всех эмпирических точек), прямой расчет уравнений кривых в рамках регрессионного анализа невозможен. Натурные данные необходимо предварительно «выпрямить», т. е. сделать возможным вычисление *линейного уравнения регрессии* с тем, чтобы потом из него получить уравнение криволинейной связи. Общий порядок регрессионного анализа для криволинейной зависимости следующий:

- преобразование исходных данных, «выпрямляющее» зависимость,
- расчет коэффициентов линейной регрессии преобразованных данных,
- проведение дисперсионного анализа, оценка значимости коэффициентов регрессии,
- обратное преобразование коэффициентов линейной регрессии для конструирования уравнения криволинейной регрессии.

Рассмотрим процесс поиска уравнения криволинейной регрессии на примере изучения зависимости веса печени прыткой ящерицы от длины ее тела (рис. 16).

Рассчитанное по исходным данным уравнение линейной регрессии имеет вид:  $y = 107.9x - 404.2$ . И хотя коэффициент регрессии достоверен ( $t = 7.6$ ,  $\alpha < 0.05$ ) и коэффициент детерминации высок  $R^2 = 0.866$ , это уравнение весьма приблизительно описывает зависимость признаков – для наименьших наблюдаемых значений длины тела оно дает абсурдное (отрицательное) значение массы печени ( $107.9 \cdot 3.4 - 404.2 = -37.3$  мг). Линейная модель не годится даже для интерполяции изучаемых данных. Гораздо успешнее справляется с подобной задачей степенная (аллометрическая) функция  $y = bx^a$ .



**Рис. 16.** Зависимость веса печени (M, мг) от длины тела (L, мм) у ящериц



Для вычисления коэффициентов этого уравнения воспользуемся преобразованием:  $Y = \lg y$ ,  $X = \lg x$ ,  $B = \lg b$ . После логарифмирования степенное уравнение приняло линейный вид:  $\lg y = \lg b + a \cdot \lg x$  или  $Y = B + aX$ . Теперь остается отыскать коэффициенты уравнения  $B$  и  $a$ , используя алгоритм метода наименьших квадратов (табл. 19).

Таблица 19

№	$x$	$y$	$X = \lg x$	$Y = \lg y$	$X^2$	$Y^2$	$X \cdot Y$	$Y'$	$(Y' - Y)^2$	$y'$
1	3.4	40	0.531	1.60	0.282	2.567	0.85	1.517	0.00718	33
2	4.2	50	0.623	1.69	0.388	2.886	1.06	1.799	0.01009	63
3	5.2	150	0.716	2.18	0.513	4.735	1.56	2.085	0.00838	121
4	5.8	120	0.763	2.08	0.583	4.323	1.58	2.23	0.02284	170
5	7.1	240	0.851	2.38	0.725	5.665	2.03	2.5	0.01442	316
6	7.0	410	0.845	2.61	0.714	6.827	2.21	2.481	0.01728	303
7	7.4	370	0.869	2.57	0.756	6.596	2.23	2.556	0.00016	359
8	8.2	500	0.914	2.69	0.835	7.284	2.47	2.693	0.00004	493
9	8.5	610	0.929	2.78	0.864	7.758	2.59	2.741	0.00201	550
$\Sigma$	56.8	2490	7.043	20.6	5.66	48.64	16.6		0.08239	

Далее рассчитаем суммы, необходимые промежуточные значения и коэффициенты (расчеты выполнялись в среде Excel):

$$\Sigma Y = \Sigma \lg y = 20.6, \Sigma Y^2 = \Sigma (\lg y)^2 = 48.64, \Sigma X = \Sigma \lg x = 7.043,$$

$$\Sigma X^2 = \Sigma (\lg x)^2 = 5.659, \Sigma XY = \Sigma (\lg x \cdot \lg y) = 16.577,$$

$$M_Y = \Sigma Y / n = 20.6 / 9 = 2.289, M_X = \Sigma X / n = 7.043 / 9 = 0.7826,$$

$$C_{XY} = \Sigma XY - (\Sigma X) \cdot (\Sigma Y) / n = 16.572 - 7.043 \cdot 20.602 / 9 = 0.45542,$$

$$C_X = \Sigma X^2 - (\Sigma X)^2 / n = 5.655 - (7.04)^2 / 9 = 0.14816,$$

$$C_Y = \Sigma Y^2 - (\Sigma Y)^2 / n = 48.638 - (20.601)^2 / 9 = 1.4823,$$

$$S_Y = \sqrt{C_Y / (n - 1)} = \sqrt{1.4823 / 8} = 0.4305,$$

$$S_X = \sqrt{C_X / (n - 1)} = \sqrt{0.14816 / 8} = 0.1361,$$

$$r = C_{XY} / \sqrt{C_X \cdot C_Y} = 0.45542 / \sqrt{0.14816 \cdot 1.34823} = 0.9718,$$

$$a = C_{XY} / C_X = 0.45541 / 0.14815 = 3.0739,$$

$$B = M_Y - aM_X = 2.289 - 3.0739 \cdot 0.7826 = -0.11643.$$

Линейное уравнение для преобразованных данных имеет вид:

$$\lg y = 3.07 \cdot \lg x + \lg(-0.116) \text{ или } Y' = 3.07 \cdot X - 0.116.$$

Это уравнение дает возможность рассчитать теоретические значения признака  $Y'$  (теоретические значения логарифмов массы печени), квадраты отклонений прогнозных значений от реальных:  $(Y' - Y)^2$ , а также их сумму  $\Sigma(Y' - Y)^2 = 0.08239$ .

Эта величина есть остаточная сумма квадратов; вместе с общей суммой квадратов  $C_y = C_{общ.} = 1.4823$  она позволяет сформировать таблицу дисперсионного анализа (табл. 20):  $C_{мод.} = C_{общ.} - C_{остат.} = 1.4823 - 0.08239 = 1.39993$ .

Таблица 20

Составляющие дисперсии	$C$	$df$	$S^2$	$F$
Наклон модельной линии	$C_{регр.} = \sum (Y'_i - M_Y)^2$	1.399	1	$S^2_{регр.} = 0.39993$ $F = 118.9377$
Отклонения вариант от линии	$C_{остат.} = \sum (y_i - Y'_i)^2$	0.0824	6	$S^2_{остат.} = 0.01177$ $F_{(0.05,1,7)} = 5.6$
Общая (всего)	$C_{общ.} = \sum (y_i - M_Y)^2$	1.482		

Полученное значение  $F = 118$  больше табличного (5.6), следовательно, дисперсия, обусловленная регрессией, достоверно больше случайной, т. е. признак  $Y$  действительно зависит от признака  $X$ , и линия регрессии адекватна исходным данным. Коэффициент детерминации больше, чем у линейной регрессии, и составляет:  $R^2 = C_{регр.} / C_{общ.} = 1.399 / 1.4823 = 0.944$ .

Ошибка коэффициента криволинейной регрессии равна:

$$m_a = \frac{S_y}{S_x} \cdot \sqrt{\frac{1-r^2}{n-2}} = \frac{0.430}{0.136} \cdot \sqrt{\frac{1-0.9718^2}{9-2}} = 0.281,$$

а критерий Стьюдента, проверяющий гипотезу  $H_0: a = 0$ , составляет

$$t = a / m_a = 3.0739 / 0.281 = 10.9.$$

Полученное значение (10.9) больше табличного ( $t_{(0.05, 8)} = 2.31$  для уровня значимости  $\alpha = 0.05$  и числа степеней свободы  $df = n - 2 = 8$ ), коэффициент регрессии  $a$  значимо отличается от нуля; зависимость признака  $Y$  от  $X$  есть, причем очень тесная. Следует помнить, что при расчете ошибки коэффициента криволинейной регрессии используются стандартные отклонения для преобразованных (у нас – прологарифмированных) значений признаков.

В завершение выполним обратное преобразование второго коэффициента регрессии, свободный член равен:

$$b = 10^B = 10^{-0.11643} = 0.764839.$$

Теперь уравнение регрессии принимает вид степенной зависимости:

$$y' = 0.765 \cdot x^{3.07}.$$

Теоретические значения  $y'$ , рассчитанные по этому уравнению, гораздо ближе к исходным данным, что хорошо видно и на графике (рис. 16), и по большей величине коэффициента детерминации ( $0.94 > 0.87$ ) (читателю несложно будет проделать все вычисления в среде Excel с помощью программы Регрессия – как для исходных, так и для преобразованных данных).

Аллометрическое уравнение ( $y' = 0.8x^3$ ) не только лучше описывает зависимость между признаками в статистическом плане, но и придает ей более ясный биологический смысл (масса печени =  $0.8 \cdot \text{длина тела}^3$ ). Как известно, объемные величины (объем, масса тела) пропорциональны кубу линейных промеров

(длина тела). В свою очередь, вес печени и вес тела связаны прямой пропорциональной зависимостью. Так становится понятной наблюдаемая прямая пропорциональность веса печени кубу длины тела.

#### ВОПРОСЫ ДЛЯ САМОКОНТРОЛЯ.

1. Какова цель изучения сходства (различий) двух выборок?
2. Из каких соображений проводится статистическая проверка гипотез?
3. Какие параметрические критерии оценки различия двух выборок Вам известны? Опишите их особенности.
4. Как провести сравнение двух групп при помощи не парного критерия Стьюдента?
5. Как провести сравнение двух групп при помощи парного критерия Стьюдента?
6. Какова необходимость сравнения выборок с помощью непараметрических критериев?
7. Охарактеризуйте необходимость применения U-критерий Уилкоксона (Манна—Уитни).
8. Охарактеризуйте необходимость применения T-критерия Уилкоксона.
9. Охарактеризуйте необходимость применения Критерия Q Розенбаума.
10. Охарактеризуйте необходимость применения X-критерия Ван-дер-Вардена.
11. Охарактеризуйте необходимость применения критерия знаков z.
12. Охарактеризуйте необходимость сравнения двух частотных распределений на основе применения критерия хи-квадрат.
13. Дайте краткий анализ непараметрических процедур.
14. Для какой цели применяется дисперсионный анализ.
15. Охарактеризуйте необходимость применения однофакторного дисперсионного анализа количественных признаков.
16. Охарактеризуйте необходимость применения непараметрического однофакторного дисперсионного анализа.
17. Для какой цели применяется двухфакторного дисперсионного анализа количественных признаков.
18. Какова необходимость проведения оценки зависимости между признаками
19. Для какой цели применяется корреляционный анализ. Охарактеризуйте эффект ложной корреляции.
20. Какова необходимость проведения множественной корреляции.
21. Какова необходимость проведения частной корреляции.
22. Какова необходимость проведения ранговой корреляции.
23. Для какой цели применяется регрессионный анализ.
24. Какова необходимость применения линейной и криволинейной регрессий?

## Глава 5. Оптимизация результатов эксперимента

Основная причина неприменимости существующего многообразия методов оптимизации, используемых при технологических исследованиях, основана на высокой вариации биологических параметров. Это создает сложность построения аналитической модели и, как следствие этому, невозможность взятия производной оптимизируемой целевой функции.

При оптимизации биологических задач чаще всего приходится иметь дело с моделями, в которых нет аналитических зависимостей для первых производных оптимизируемой целевой функции. Поэтому поиск локального максимума или минимума в этом случае приходится вести по результатам экспериментальных измерений, полученных по значениям целевой функции с помощью прямых методов оптимизации.

Для удобства проведения анализа и последующей оптимизации следует создать базу данных. Многофакторный анализ с последующей оптимизацией полученных результатов целесообразно производить при помощи программ: Word, Excel, Access и SPSS 17 в следующей последовательности:

- данные, представленные в рабочем журнале, набирают в таблицы в редакторе Word или в базы Excel, dBASE IV;
- для перевода данных из таблиц выделяют диапазон и переводят “сохранить как” в формат Access (имя до 8 символов).
- значения дат переводят в SPSS 17: →Date (Variable view →String→Numeric);
- проводят фильтрацию данных: Data→Select Cases→If condition is satisfied;
- осуществляют анализ данных: Analyze→Regression (General Linear Model) →Univariate, Dependent variable, Fixed Factor, Covariate (ввод и выход проводится стрелкой), Model →Custom, Factor→Type→Main effects→ Continue;
- выделение отдельного фактора (или их взаимодействия) Interaction Include intercept in model;
- для построения графика выделяют Plots→horizontal axis→separate lines→separate plots→ Continue. Option (выделить первые 4 позиции) → Continue.

### 5.1 Особенности методов однофакторной оптимизации исследования

Среди методов оптимизации нулевого порядка находят применение в химических и физических исследованиях методы Розенброка, конфигураций, деформируемого многогранника, случайного поиска. К методам с использованием производных относятся методы наискорейшего спуска, сопряженных градиентов, переменной метрики. Основной причиной неприменимости этих методов оптимизации в биологии является высокая вариация биологических параметров,

что приводит к сложности построения аналитической модели и, как следствие этому, невозможность взятия производной от целевой функции.

При оптимизации биологических задач чаще всего приходится иметь дело с моделями, в которых нет аналитических зависимостей для первых производных целевой функции. Поэтому поиск локального максимума или минимума в этом случае приходится вести по результатам экспериментальных измерений, полученным по значениям целевой функции – с помощью **прямых методов оптимизации**.

### 5.1.1 Выделение существенных факторов

Для правильного отображения объекта необходимо, чтобы его математическая модель включала все факторы, существенно влияющие на выходную величину [15]. Отсутствие в модели хотя бы одного из существенных факторов может повлечь за собой ошибочную интерпретацию явлений, протекающих в объекте. В свою очередь это может явиться причиной серьезных ошибок в решениях, принимаемых с помощью модели. Например, расчетное положение оптимума окажется значительно удаленным от действительного и т. п. Кроме того, модели, в которые входят не все доминирующие факторы, как правило, являются неадекватными.

При исследовании сложных процессов, на которые воздействует большое число факторов, обычно заранее не известно, какие из этих факторов следует считать доминирующими. Степень их влияния весьма различна, поэтому задача заключается в том, чтобы с помощью соответствующих методов выделить и идентифицировать существенные, доминирующие факторы на „шумовом фоне” всех остальных факторов. Цель этих методов состоит в получении грубых оценок воздействия факторов (рис. 1.13) при возможно меньших затратах (минимальном числе экспериментов).

Для построения математических моделей разработаны специальные методы, которые при выполнении некоторых предпосылок позволяют выявить существенные факторы с помощью небольшого числа экспериментов и при небольших затратах времени.

К числу таких методов относятся, например, следующие: опрос экспертов с целью ранжирования факторов по степени их влияния на выходную величину; дисперсионный, корреляционный, факторный и дискриминантный анализ; регрессионные модели и многомерное шкалирование.

В основу данных методов положено предположение о том, что существенность некоторого (дискретного) фактора характеризуется его вкладом в выход-

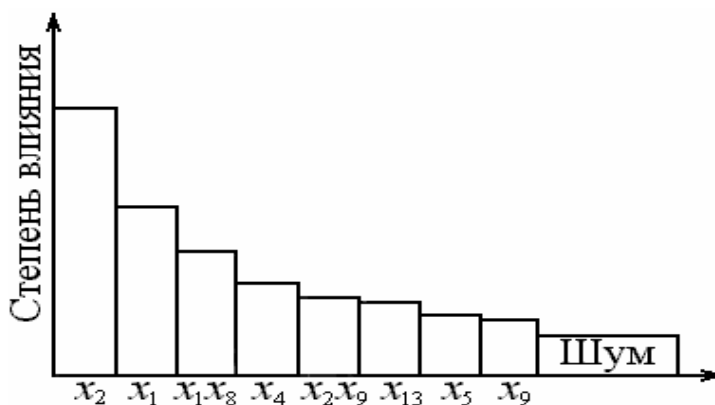


Рис. 1.13. Распределение факторов по степени их влияния на целевую величину.

ной величины которую количественно можно оценить через показатели: дисперсии, коэффициенты корреляции и регрессии.

### 5.1.2 Методы случайного поиска в задачах экспериментальной оптимизации биотехнологических параметров

Стохастическими или случайными методами поиска называют методы, в которых для определения оптимума целевой величины или направления движения к экстремуму сознательно используются случайные элементы (случайные числа). Алгоритм поиска содержит при этом следующие три основных элемента, каждый из которых состоит в свою очередь из ряда последовательных этапов:

1. Алгоритм получения начальной точки поиска.
2. Решающее правило, с помощью которого определяется дальнейший ход поиска на основании данных, полученных путем реализации эксперимента.
3. Накопление информации, которая может быть полезной на последующих этапах поиска. Эта информация может различными способами использоваться на фазе обучения или адаптации, процедуры поиска [15]. Например, на основе накопленных данных может оказаться целесообразным изменить статистические свойства случайных элементов, используемых при поиске, с целью повышения скорости, точности и достоверности поисковой процедуры. Могут быть внесены изменения в значения параметров, характеризующих алгоритм поиска, которые позволят повысить его эффективность.

При применении простейших алгоритмов случайного поиска обычно обходятся без обучения и адаптации.

### 5.1.3 Метод покоординатного спуска (Гаусса-Зейделя).

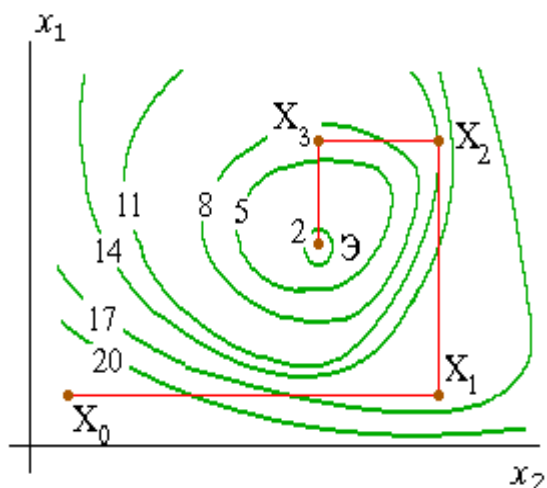


Рис.4. Траектория покоординатного спуска

Этот метод характеризуется выбором направлений поиска поочередно вдоль всех координатных осей, шаг рассчитывается на основе одномерной оптимизации.

Траектория покоординатного спуска для примера двумерного пространства управляемых параметров показана на рис. 4, где  $X_k$  – точки на траектории поиска,  $x_i$  – управляемые параметры. **Целевая функция** представлена своими линиями равного уровня, около каждой линии записано соответствующее ей значение  $F(X)$ . Очевидно, что Э есть точка минимума.

Достоинством **метода Гаусса-Зейделя** является простота проведения эксперимента. Основной недостаток заключается в медленности схождения на **овражных функциях**, в которых овраг не ориентирован в направлении какой-либо из координатных осей.

Оврагом называют часть пространства управляемых параметров, в которой наблюдаются слабые изменения производных целевой функции по одним направлениям и значительные изменения с переменной знака – по некоторым другим направлениям. Знак производной меняется в точках, принадлежащих дну оврага. При использовании метода покоординатного спуска велика вероятность "застревания" поиска на дне **оврага** вдали от точки экстремума. На рис. 5 видно, что после попадания в точку А, расположенную на дне оврага, дальнейшие шаги возможны лишь в направлениях  $aa$  или  $bb$ , но они приводят к ухудшению целевой функции. Следовательно, поиск прекращается в точке А.

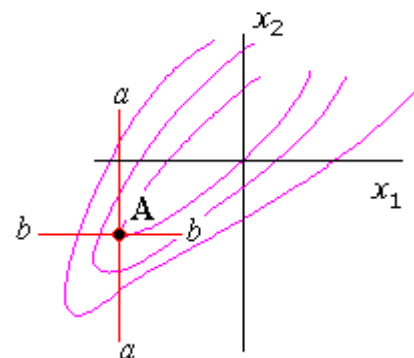


Рис. 5. «Застревание» по координатного спуска на дне оврага



Рис.6. Траектория по координатного спуска при благоприятной ориентации координатных осей

Такой поворот осуществляют на основе данных, полученных после серии из  $n$  шагов по координатного спуска. Положение новых осей  $S_i$  может быть получено линейным преобразованием прежних осей  $x_i$ : ось  $S_1$  совпадает по направлению с вектором  $X_{k+1}-X_k$ ; остальные оси выбирают из условия ортогональности к  $X_1$  и друг к другу.

Другой удачной модификацией по координатного спуска является **метод конфигураций (Хука-Дживса)**. В соответствии с этим методом вначале выполняют обычную серию из  $n$  шагов по координатного спуска, затем делают дополнительный шаг в направлении вектора  $X_{k+1}-$

В то же время при благоприятной ориентации дна оврага, а именно при положении одной из координатных осей, близком к параллельности с дном оврага, поиск оказывается весьма быстрым. Эта ситуация показана на рис. 6.

#### 5.1.4 Метод Розенброка

**Метод Розенброка** является улучшенным вариантом по координатного спуска. Метод заключается в таком повороте координатных осей, чтобы одна из них оказалась квазипараллельной дну оврага. Такой поворот осуществляют на основе данных, полученных после серии из  $n$  шагов по координатного спуска. Положение новых осей  $S_i$  может быть получено линейным преобразованием прежних осей  $x_i$ : ось  $S_1$  совпадает по направлению с вектором  $X_{k+1}-X_k$ ; остальные оси выбирают из условия ортогональности к  $X_1$  и друг к другу.

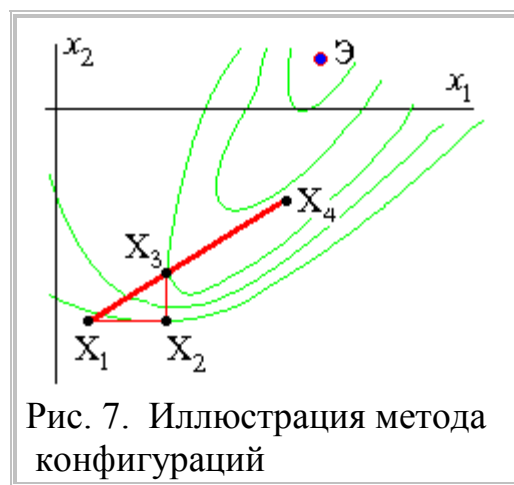


Рис. 7. Иллюстрация метода конфигураций

$X_k$ , как показано на рис. 7, где дополнительный шаг выполняют в направлении вектора  $X_3 - X_1$ , что и приводит в точку  $X_4$ .

Поиск экстремума методом **деформируемого многогранника (Нелдера-Мида)** основан на построении многогранника с  $(n+1)$  вершинами на каждом шаге поиска, где  $n$  – размерность пространства управляемых параметров. В начале поиска эти вершины выбирают произвольно, на последующих шагах выбор подчинен правилам метода.

Эти правила поясняются (рис. 8) на примере двумерной задачи оптимизации. Выбраны вершины исходного треугольника:  $X_1, X_2, X_3$ . Новая вершина  $X_4$  находится на луче, проведенном из худшей вершины  $X_1$  (из вершины с наибольшим значением целевой функции) через центр тяжести ЦТ многогранника, причем рекомендуется  $X_4$  выбирать на расстоянии  $d$  от ЦТ, равном  $|\text{ЦТ} - X_1|$ . Новая вершина  $X_4$  заменяет худшую вершину  $X_1$ .

Если оказывается, что  $X_4$  имеет лучшее значение целевой функции среди вершин многогранника, то расстояние  $d$  увеличивают. На рисунке именно эта ситуация имеет место и увеличение  $d$  дает точку  $X_5$ . В новом многограннике с вершинами  $X_2, X_3, X_5$  худшей является вершина  $X_2$ , аналогично получают вершину  $X_6$ , затем вершину  $X_7$  и т.д. Если новая вершина окажется худшей, то в многограннике нужно сохранить лучшую вершину, а длины всех ребер уменьшить, например вдвое (стягивание многогранника к лучшей вершине). Поиск прекращается при выполнении условия уменьшения размеров многогранника до некоторого предела.



Рис. 8. Иллюстрация метода деформируемого многогранника

## 5.2 Необходимость проведения многофакторного эксперимента

Задачей большинства биотехнологических экспериментов является количественное изучение каких-либо свойств объекта. Для этого проводятся измерения одной или нескольких физических или биологических величин с последующей обработкой полученных данных. Экспериментальные результаты обычно содержат погрешности, связанные с тем, что любые измерения сопровождаются действием и взаимодействием большого числа разнообразных и трудно учитываемых факторов. Конечной целью любого исследования является не только представление наилучшей, по мнению экспериментатора, оценки измеряемой величины, но и максимально достоверной оценки погрешности измерений.



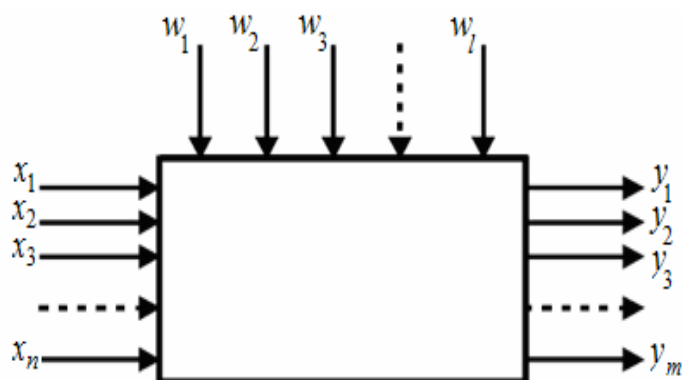


Рис. 1.19. Схема объекта исследования

Любой прибор или устройство для измерения физических величин можно рассматривать в виде объекта исследования (рис. 1.19), для которой  $x_1, \dots, x_n$  – входные измеряемые и регулируемые параметры;  $w_1, \dots, w_l$  – неконтролируемые, случайным образом изменяющиеся параметры («шум» объекта);  $y_1, \dots, y_m$  – выходные параметры. Комплекс параметров  $x_1, \dots, x_n$  называют *ос-*

*новным*, поскольку он определяет условия эксперимента.

Результат опыта зависит не только от основных параметров, но и от «шума» объекта, влияние которого носит случайный характер. Поэтому естественно рассматривать и результат эксперимента, и ошибку измерения как случайные величины, управляемые вероятностными законами, и применять для учета действия случайных факторов теорию вероятностей. Тогда влияние случайных ошибок на результат измерения можно количественно оценить при помощи математической статистики.

Современная биотехнология выпускает несколько десятков тысяч наименований продуктов, в лабораториях разрабатываются сотни новых технологических операций. Экспериментальное изучение механизмов протекания всех этих процессов нереально, между тем задачи их оптимизации и управления необходимо решать. Для этих целей успешно применяются экспериментально статистические методы, с помощью которых составляется математическая модель объекта и при неизвестном механизме протекающих в объекте процессов изучается зависимость отклика системы на изменения основных параметров.

Математической моделью объекта служит функция отклика, связывающая выходной параметр, характеризующий результаты эксперимента, с переменными, которые варьируют при проведении опытов. Независимые переменные  $x_1, x_2, \dots, x_n$  называют факторами, пространство с координатами  $x_1, x_2, \dots, x_n$  – факторным пространством, а геометрическое изображение функции отклика в факторном пространстве – поверхностью отклика:  $y = \varphi(x_1, x_2, \dots, x_n)$ . Многие исследования в области информационных систем и технологий, а также и других областях знаний сводятся к решению задач на экстремум, т.е. к нахождению максимумов или минимумов функций нескольких переменных. С помощью необходимых условий существования экстремума функции нескольких переменных находятся стационарные точки путем решения системы уравнений  $d\varphi/dx_1=0, d\varphi/dx_2=0, \dots, d\varphi/dx_n=0$ . Далее, среди найденных стационарных точек выявляются точки экстремума с помощью достаточного условия локального экстремума функции нескольких переменных. Эта же задача может быть решена численными методами поиска экстремума, например, градиентным, релаксационным и т. д. Однако случай, когда априорно известно аналитическое выраже-

ние, описывающее объект исследования, на практике реализуется не всегда. Чаще оно неизвестно, а указывается лишь область, в которой могут изменяться факторы. Поиск локального экстремума в таких условиях осуществляется с помощью экспериментов “при неполной информации об объекте оптимизации”.

Существующие биотехнологические исследования связаны с изучением совокупности параметров, которые описывают биологическое явление. В целом, их можно разделить на биологические и технологические. Как правило, они находятся в тесной взаимосвязи. Например, на уровень сохранности деконсервированного биообъекта влияет различие криорезистентных свойств биоматериала и многообразие криоповреждающих факторов как следствие выбранной технологии замораживания-оттаивания.

Эффективность экспериментов в большой степени зависит от методов их проведения. Общепринятым способом проведения криобиологического эксперимента является метод Гаусса-Зейделя [12]. Его особенностью является постановка серии опытов с поочередным варьированием каждой из исследуемых переменных. В этом случае принято изменять в каждом эксперименте только один параметр, поддерживая остальные на фиксированном уровне. При таком подходе возможное количество измерений определяется количеством параметров и их уровней градации, что составляет:

$$N_{\text{л}} = n \cdot k, \quad (1.14)$$

где  $N_{\text{л}}$  – количество измерений для нахождения локального оптимума;  
 $n$  – количество параметров, поставленных на исследование;  
 $k$  – количество уровней градаций для каждого параметра.

Недостатком такого подхода является не учет влияния совместного сочетания исследуемых параметров на результирующий показатель и как следствие этому нахождение только локального оптимума. Для отыскания же глобального оптимума необходимо провести значительно больше измерений (например, для 5 факторов, имеющих 5 уровней градаций изменения, количество возможных комбинаций составит 3125), учитывая не только все возможные изменения –  $k$  исследуемых параметров –  $n$ , но и *их взаимное* сочетание, общее число которых составляет:

$$N = k^n, \quad (1.15)$$

где  $N$  – количество измерений, необходимых для нахождения глобального оптимума методом последовательного перебора.

Поскольку состояние открытой системы (биологического объекта) представляет собой взаимосвязанное сочетание большого числа составляющих факторов (химических, физических и биологических), то основной задачей комплексного исследования являются результаты изучения влияния их комбинаций. Однако последовательное проведение ряда опытов по формуле (1.15) для изуче-

ния взаимосвязи всех факторов может затянуться на долгие годы, что приведет к неоправданно большому расходу денежных средств и времени.

Альтернативный подход к изучаемой проблеме заключается в одновременном изменении всех факторов. В этом состоит идея многофакторного эксперимента, в котором одновременно изучаются все возможные комбинации уровней различных факторов. Интерес к многофакторному исследованию обусловлен рядом существенных преимуществ: возможностью значимого сокращения числа опытов, нахождения глобального оптимума, количественного установления влияния всех факторов на изучаемый показатель [1-10].

Первая работа в этой области была опубликована в 1951 г. Боксом и Уилсоном в Англии [14]. Идея метода Бокса-Уилсона крайне проста. Экспериментатору предлагается ставить последовательные небольшие серии опытов, в каждой из которых одновременно варьируются по определенным правилам все факторы. Серии организуются таким образом, чтобы после математической обработки предыдущей можно было выбрать условия проведения (т.е. спланировать) следующую серию. Так последовательно шаг за шагом, достигается область экстремума. Наиболее короткий путь к вершине – направление градиента функции отклика. Градиент непрерывной однозначной функции  $\varphi$  есть вектор

$$\nabla \varphi = \frac{\partial \varphi}{\partial x_1} i + \frac{\partial \varphi}{\partial x_2} j + \dots + \frac{\partial \varphi}{\partial x_k} k, \quad (1.16)$$

где  $\nabla \varphi$  – обозначение градиента,  $\frac{\partial \varphi}{\partial x_i}$  – частная производная функции по  $i$ -му фактору,  $i, j, k$  – единичные векторы в направлении координатных осей.

Следовательно, составляющие градиента суть частные производные функции отклика, оценками которых являются коэффициенты регрессии. Изменяя независимые переменные пропорционально величинам коэффициентов регрессии, возможно перемещение в направлении градиента функции отклика по самому крутому пути. Поэтому процедура движения к почти стационарной области называется крутым восхождением.

Таким образом, задача многофакторной оптимизации сводится к опытному определению такого сочетания уровней  $n$  факторов (координаты точки в  $(n+1)$ -мерном факторном пространстве), при котором достигается максимальное (минимальное) значение выходного параметра  $y$  (или нескольких параметров).

Для наглядности рассмотрим случай, когда на систему оказывают влияние только два фактора ( $x_1$  и  $x_2$  в безразмерном масштабе). Построим контурные сечения  $y = const$  поверхности отклика при  $n = 2$  (рис. 1.20 а). Поиск экстремальной точки поверхности отклика в традиционном эксперименте методом Гаусса-Зейделя проводится следующим образом. В точке  $L$  с известным значением  $y$  фиксируется один из факторов, например  $x_1$ , и начинается движение из этой точки вдоль оси  $x_2$ . Движение по  $x_2$  продолжается до тех пор, пока не прекращается прирост  $y$  (рис. 1.20 б). В точке  $M$  с наилучшим значением выходного параметра фиксируется фактор  $x_2$  и начинается движение в направлении оси  $x_1$ . В точке  $N$

со следующим наилучшим значением  $y$  снова фиксируется  $x_1$  и начинается движение по  $x_2$  и т. д. Очевидно, что путь к экстремуму по ломаной кривой  $LMNR$  (рис. 1.20 б) не является оптимальным. Кратчайшим, наиболее крутым путем достижения экстремума будет движение из точки  $L$  по градиенту перпендикулярно изолиниям  $y = const$  (на рис. 1.20 б этот путь показан пунктирной линией).

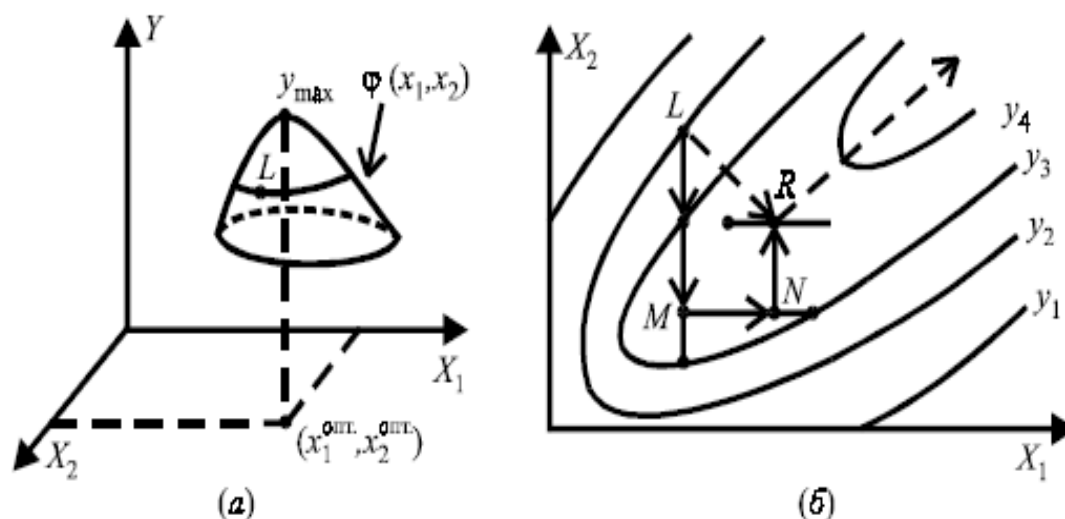
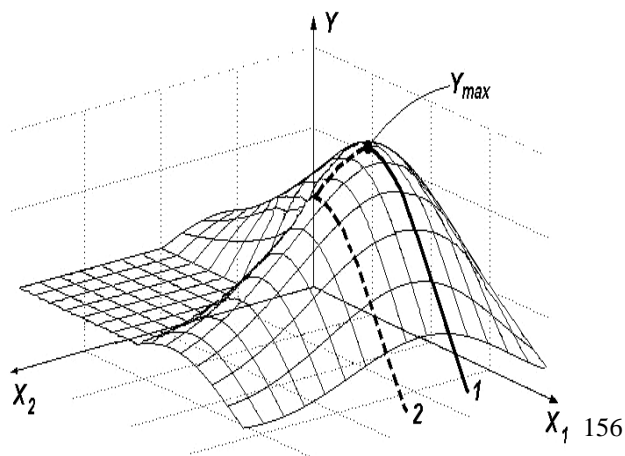


Рис. 1.20. Движение по поверхности отклика (а) к экстремуму в традиционном эксперименте и в методе крутого восхождения (б).

### 5.2.1 Сравнительный анализ существующих методов многофакторной оптимизации исследования.

Планирование эксперимента используют для изучения и математического описания процессов и явлений путём построения математических моделей (например, в форме так называемых уравнений регрессии) – соотношений, связывающих с помощью ряда параметров значения факторов и результаты эксперимента, называемых откликами. Основное требование, предъявляемое к планам факторного эксперимента, в отличие от пассивного эксперимента, – минимизация числа опытов, при которой получают достоверные оценки вычисляемых параметров при соблюдении приемлемой точности математических моделей в заданной области факторного пространства. В этом случае задача обработки результатов факторного эксперимента заключается в определении численных значений указанных параметров.

Рассмотрим задачу нахождения экстремума методом крутого восхождения, который содержит в себе существенные элементы метода полного факторного эксперимента (или дробного факторного эксперимента) и метода наискорейшего спуска. Стратегия



метода крутого спуска (восхождения) состоит из двух основных этапов.

Этап I состоит в построении линейной модели функции отклика методом полного факторного эксперимента и включает в себя следующее: а) полный факторный эксперимент; б) проверку воспроизводимости эксперимента; в) нахождение и проверку значимости выборочных коэффициентов регрессии; г) проверку адекватности математической модели.

Этап II – крутое восхождение. Он состоит из: а) нахождения направления градиента; б) выбора шага движения по направлению градиента; в) движения по направлению градиента до достижения точки условного экстремума.

Рассмотрим на примере зависимости целевой функции  $y$  от двух факторов схемы достижения экстремума поверхности отклика методами крутого восхождения и симплекс-планирования (рис. 1.21).

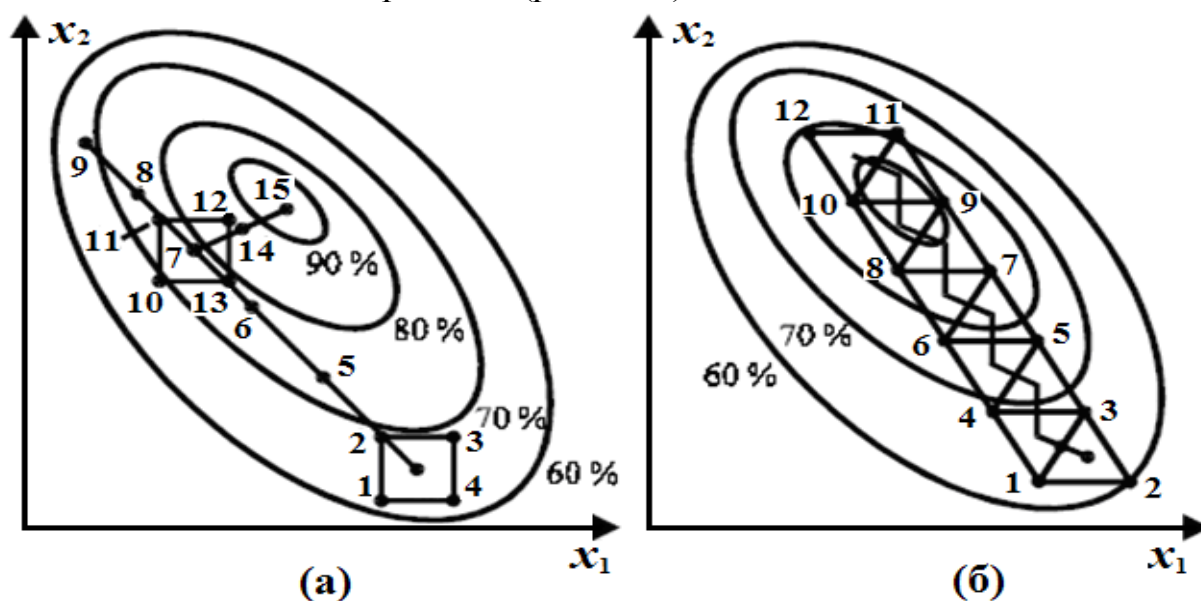


Рис. 1.21. Достижение экстремума поверхности отклика методами крутого восхождения (а) и симплекс-планирования (б).

При оптимизации методом крутого восхождения (рис. 1.21 а) в окрестности точки  $M$  поставлен полный факторный эксперимент – ПФЭ  $2^2$ , движение по градиенту линейного приближения осуществлялось в опытах  $5 \div 9$ . Далее был поставлен новый ПФЭ  $2^2$  (точки  $10 \div 13$ ) с центром в точке 7, в которой было получено наилучшее значение  $y$ . Движение по новому градиенту (точки  $14 \div 15$ ) приводит к экстремуму.

При оптимизации методом симплекс-планирования (рис. 1.21 б) в исходном симплексе (точки  $1 \div 3$ ) худшей точкой оказалась точка 2. Ее зеркальным отражением относительно точки 1 – центра грани  $1 \div 3$  – является точка 4. Область оптимума достигается при реализации симплекса 9, 10, 11. Хотя оба рассмотренных метода требуют проведения примерно одинакового числа опытов, симплекс-планирование имеет ряд важных преимуществ: при использовании этого метода параметр оптимизации  $y$  может измеряться приближенно, достаточно иметь возможность проранжировать его величину; можно одновременно учитывать несколько параметров оптимизации; метод не предъявляет жестких

требований к локальной аппроксимации поверхности отклика уравнением регрессии.

Крутое восхождение – это движение в направлении градиента функции отклика. Градиент задается частными производными, а частные производные функции отклика оцениваются коэффициентами регрессии. В крутом восхождении независимые переменные изменяют пропорционально величинам коэффициентов регрессии и с учетом их знаков. Составляющие градиента однозначно получаются умножением коэффициентов регрессии на интервалы варьирования по каждому фактору.

Применение планирования эксперимента делает проведение экспериментатора целенаправленным и организованным, существенно способствует повышению производительности его труда и надежности полученных результатов.

### 5.2.2 Генетические Алгоритмы (Genetic Algorithms)

Генетический алгоритм является самым известным на данный момент представителем эволюционных алгоритмов, и по своей сути является алгоритмом для нахождения глобального экстремума многоэкстремальной функции. Он представляет собой модель размножения живых организмов и является новейшим, но не единственно возможным способом решения задач оптимизации.

Идею ГА подсказала сама природа и работы Дарвина. Делается предположение, что если взять два вполне хороших решения задачи и каким-либо образом получить из них новое решение, то будет высокая вероятность того, что новое решение получится хорошим или даже более лучшим. Для реализации этого используют моделирование эволюции (естественного отбора). Известны два основных пути решения таких задач – переборный и локально-градиентный. У этих методов свои достоинства и недостатки, и в каждом конкретном случае следует подумать, какой из них выбрать. Переборный метод наиболее прост по своей сути. Для поиска оптимального решения (точки максимума целевой функции) требуется последовательно вычислить значения целевой функции во всех возможных точках, запоминая максимальное из них.

Градиентные методы работают очень быстро, но не гарантируют оптимальности найденного решения. Они идеальны для применения в так называемых *унимодальных* задачах, где целевая функция имеет единственный локальный максимум (он же – глобальный).

Типичная практическая задача, как правило, *мультимодальна* и многомерна, то есть содержит много параметров. Для таких задач не существует ни одного универсального метода, который позволял бы достаточно быстро найти абсолютно точное решение. Однако, комбинируя переборный и градиентный методы, можно надеяться получить хотя бы приближенное решение, точность которого будет возрастать при увеличении количества проведенных опытов.

Преимущества генетических алгоритмов: не требуют дополнительной информации о поверхности отклика; разрывы, существующие на поверхности отклика, незначительно влияют на эффективность оптимизации; устойчивы к попаданию в локальные оптимумы; хорошо работают при решении задач многоце-

левой оптимизации; могут быть использованы для широкого класса задач; просты и прозрачны в реализации; могут быть использованы в задачах с изменяющейся средой.

Недостатки генетических алгоритмов. Не желательно и проблематично использовать ГА: в случае, когда необходимо найти точный глобальный оптимум; время исполнения функции оценки велико; необходимо найти все решения задачи, а не одно из них; конфигурация является не простой (кодирование решения); поверхность ответа имеет слабо изменяющийся рельеф.

Отличительной особенностью данного способа является высокая эффективность поиска оптимального решения при исследовании технических систем, моделируемых на высоких уровнях сложности и иерархичности как, включая последние достижения математического моделирования, 2 -х и 3 -х мерные задачи.

К многофакторной оптимизации можно отнести следующие методы: случайного поиска, симплексный, Бокса-Уилсона и его модификацию – эволюционное восхождение. Для определения условий повышения эффективности криобиологических исследований посредством сокращения расходов на их проведение необходим выбор адекватного метода оптимизации эксперимента. Сравнительный анализ показал (приложение В), что количество опытов существенно сокращается с увеличением числа совместно исследуемых параметров. В случае, когда количество планов равно числу факторов и величина шага изменения параметра равна ошибке опыта, то при увеличении числа одновременно изменяемых параметров количество биообъекта может быть сокращено более чем в несколько раз.

Анализ литературных данных и собственных исследований позволил сделать следующие выводы. Симплексный метод, может быть, применим только для определения оптимума функции отклика. Методы Бокса-Уилсона дают возможность, помимо этого, также построить адекватную описательную модель посредством полученных уравнений регрессии. Построение планов первого порядка при помощи полного (ПФЭ) и дробного факторного эксперимента (ДФЭ) позволяет описать только линейные эффекты и эффекты взаимодействия факторов. Для уменьшения числа опытов, когда априорно известно, что ПФЭ дает избыточную информацию, не существенную при построении линейной модели, применим ДФЭ. Для описания участка поверхности отклика применимы планы второго порядка, ядром которых являются планы первого порядка. Для уточнения причинно-следственной связи с использованием уже существующей математической модели наиболее применим метод Гаусса-Зейделя.

Можно выделить несколько уровней исследования и соответствующие им наиболее приемлемые методы проведения эксперимента, зависящие от определения проблемы и создания рабочей гипотезы её решения до разработки новой теории, органически дополняющей существующую парадигму:

1) симплексный метод наиболее подходит для нахождения глобального оптимума (максимальное значение сохранности найденное при совместном варьировании нескольких параметров), когда проблема не исследована и не по-

ставлены методы ее решения, величина шага изменения параметра минимальна и соответствует ошибке опыта;

2) метод крутого восхождения – для нахождения и описания области глобального оптимума, когда проблема частично изучена, поставлены методы исследования, но ещё не получены статистически достоверные значения, шаг изменения параметра выше величины ошибки опыта;

4) расчётный метод – при обобщении и абстрагировании полученных данных, необходимых для дальнейшего исследования на стыке разных областей знаний, когда проблема исследована в целом, создана модель представленная аналитическими выражениями, параметры которых имеют физические размерности, позволяющая оптимизировать изучаемый процесс, оптимальные значения параметров необходимые для проведения эксперимента определяются на основе математических вычислений.

### **5.2.3 Исследование поверхности отклика на основе уравнений регрессии второго порядка**

Уравнение регрессии, полученное с помощью метода крутого восхождения, позволяет не только предсказать значение функции отклика для заданных условий проведения эксперимента, но и дает информацию о форме поверхности отклика. Исследование этой поверхности необходимо для описания исследуемого явления.

Для изучения формы поверхности отклика уравнение регрессии необходимо привести к так называемой канонической форме, которая имеет вид

$$y - y_s = B_{11}Z_1^2 + B_{22}Z_2^2 + \dots + B_{nn}Z_n^2, \quad (1.17)$$

где  $y$  – функция отклика;  $Z_i$  – новые независимые переменные величины;  $B_{ii}$  – коэффициенты канонической формы.

Приведение уравнения регрессии к канонической форме (1.17) соответствует переносу начала координат в новую точку факторного пространства и повороту координатных осей на некоторый угол.  $y_s$  – это значение функции отклика в новом начале координат.

Все многообразие поверхностей отклика, описываемых уравнением вида (1.17), можно разделить на три класса (рис. 1.22 и 1.23). К первому классу относятся поверхности, имеющие экстремум (рис. 1,22 а, б). В этом случае все коэффициенты канонической формы имеют одинаковые знаки. Анализ таких поверхностей заканчивается после приведения уравнения регрессии к канонической форме. Если центр поверхности находится в допустимой области задания влияющих факторов, то исследователю необходимо поставить несколько контрольных опытов вблизи центра поверхности и убедиться в том, что значения функции отклика, предсказанные уравнением регрессии, достаточно хорошо совпадают с экспериментальными данными. Поверхности отклика основных классов в двухмерной системе координат представляются



в следующем виде (рис. 1.23). Ко второму классу относятся поверхности типа «стационарного возвышения» (см. рис. 1.22 в, г). В этом случае некоторые коэффициенты канонической формы близки к нулю.

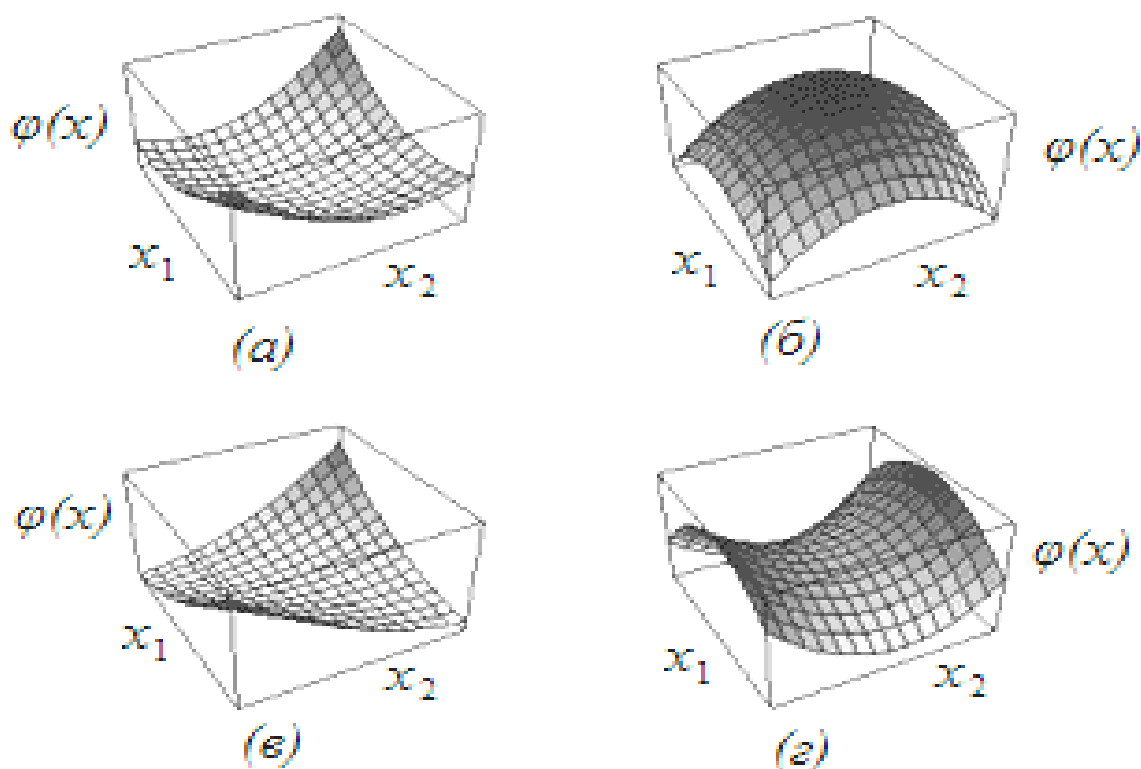


Рис. 1.22. Типы поверхностей отклика в трёхмерной системе координат.

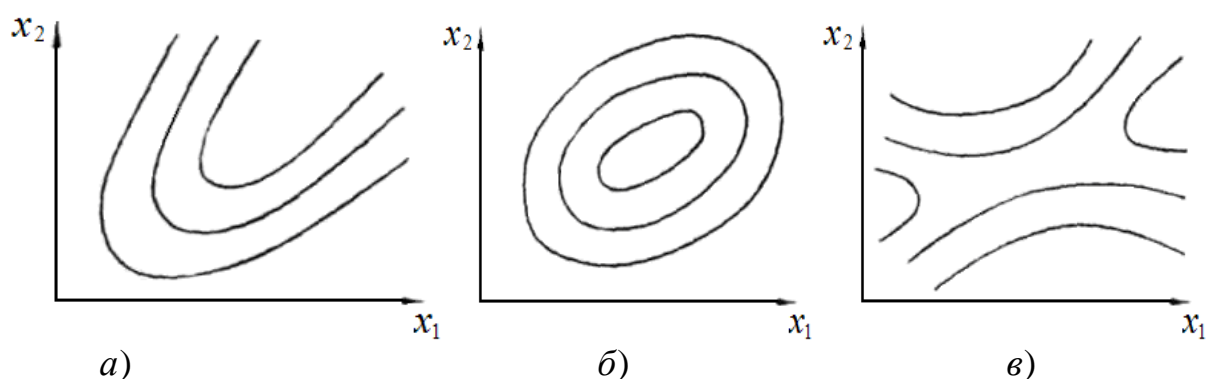


Рис. 1.23. Типы поверхностей отклика в двухмерной системе координат.

Имея дело с поверхностями отклика типа «стационарное возвышение» или «седло», исследователь должен пользоваться методами вычислительной математики и средствами вычислительной техники для нахождения условного экстремума критерия оптимальности с учетом ограничений, наложенных на влияющие факторы.

### 5.3 Контурно-графический анализ

При изучении формы поверхности отклика исследователи в ряде случаев обходятся без составления математического описания функции отклика, используя приемы контурно-графического анализа. Сущность его состоит в определенном расположении опытов в факторном пространстве, получении дополнительной информации линейной интерполяцией экспериментальных данных и построении на факторной плоскости (или на двумерных сечениях) линий постоянного уровня функции отклика. Экспериментальные точки располагают таким образом, чтобы они охватывали всю область факторного пространства, представляющую интерес для исследователя. Схема, предложенная П. Берчем, предполагает постановку шести опытов (рис. 15). Один из них проводится в центре исследуемой области, а остальные в вершинах пятиугольника. Схема В. Клеймана (рис. 16) требует постановки четырех опытов в вершинах прямоугольника и двух опытов на его оси симметрии.

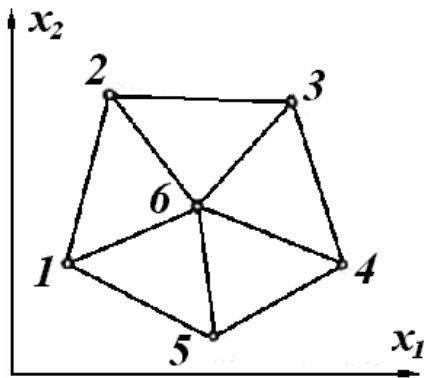


Рис. 15. Схема П. Берча.

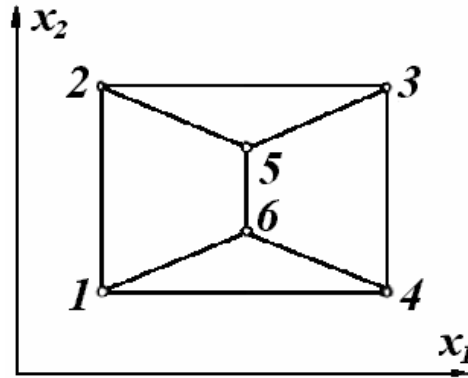


Рис. 16. Схема В. Клеймана.

Разумеется, схемы П. Берча и В. Клеймана не исчерпывают всех возможных вариантов расположения экспериментальных точек в факторном пространстве. Однако следует иметь в виду, что расстояния между этими точками не должны быть слишком велики. В противном случае при большой кривизне и сложной форме поверхности отклика погрешности линейной интерполяции наложат заметный отпечаток на результаты исследований.

На рис. 17 показано, как осуществляется контурно-графический анализ по схеме В. Клеймана. Сначала проводят намеченную серию опытов. Затем соединяют линейными отрезками соседние точки и находят методом линейной интерполяции значения функции отклика в серединах этих отрезков (как средние арифметические значения результатов опытов в соединяемых экспериментальных точках). Наконец, проводят линии через точки с одинаковыми значениями функции отклика.

Контурно-графический анализ по схеме П. Берча проводится аналогично. Рассмотренной методикой можно воспользоваться и при построении двумерных сечений поверхности отклика.

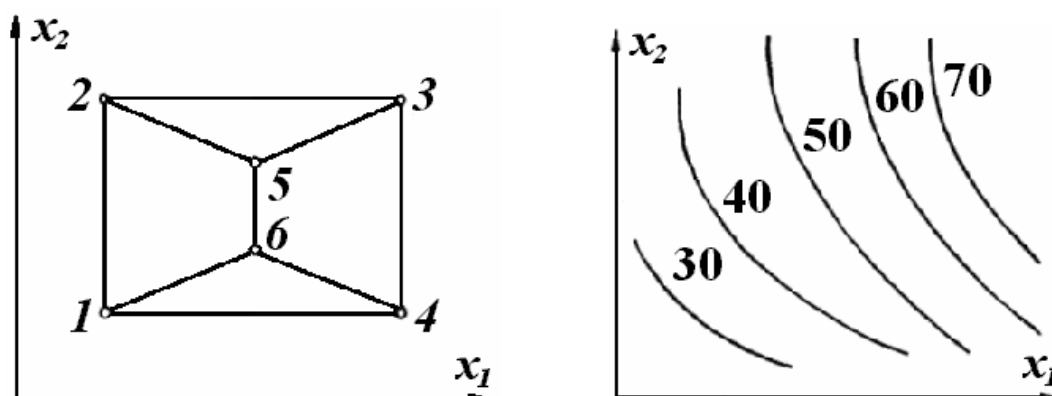


Рис. 17. Контурно-графический анализ по схеме В. Клеймана.

Сначала попарно проводят намеченную серию опытов. Затем соседние точки соединяются линейными отрезками и методом линейной интерполяции определяются расчетные значения функции отклика в серединах этих отрезков. Для построения полной картины поверхности отклика, при необходимости, методом линейной экстраполяции находят расчетные точки на продолжениях линейных отрезков (на расстояниях, равных длинам предшествующих отрезков). Линиями соединяются точки с одинаковыми значениями функции отклика.

В качестве примера эффективности исследования взаимосвязи различных биологических и технологических параметров при помощи поверхностей отклика рассмотрим примеры криоконсервирования черенков плодово-ягодных культур (рис. 18 - 20). Целевой функцией являлся показатель жизнеспособности ( $V$ , %) черенков, в большей степени зависящий от их влажности ( $\eta$ , %) и температуры ( $T$ , °C). Менее значимыми оказались скорости охлаждения и отогрева. Применение комплекса способов высушивания, охлаждения, хранения и оттаивания черенков дает возможность сохранить начальную жизнеспособность образцов. При этом допустимый предел начальной влажности исследуемых образцов индивидуален и не должен быть ниже описанных зависимостей в положительном и в отрицательном диапазоне температур. Показано, что жизнеспособность деконсервированных черенков зависит от комплекса факторов, наиболее значительным из которых является скорость охлаждения. Уровень жизнеспособности изменяется от 0 до 100 % при охлаждении (1 и 0,1 °C/ч) до различных температур

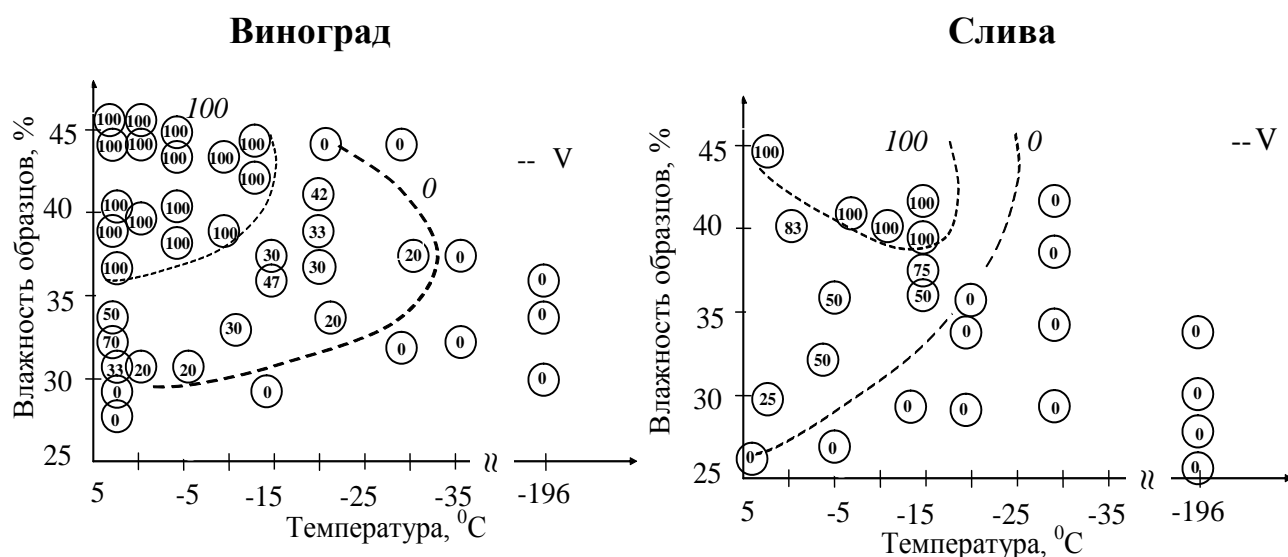


Рис. 18. Жизнеспособность черенков в зависимости от их влажности ( $\eta$ , %) и температуры ( $T$ ,  $^{\circ}\text{C}$ ). Скорость охлаждения –  $0,1^{\circ}\text{C/ч}$ , отогрева –  $70^{\circ}\text{C/мин}$ .

**Примечание:** Номер, очерченный кругом, отражает показатель жизнеспособности черенков (in vitro), а пунктирной линией обозначен контур, ограничивающий одинаковые величины.

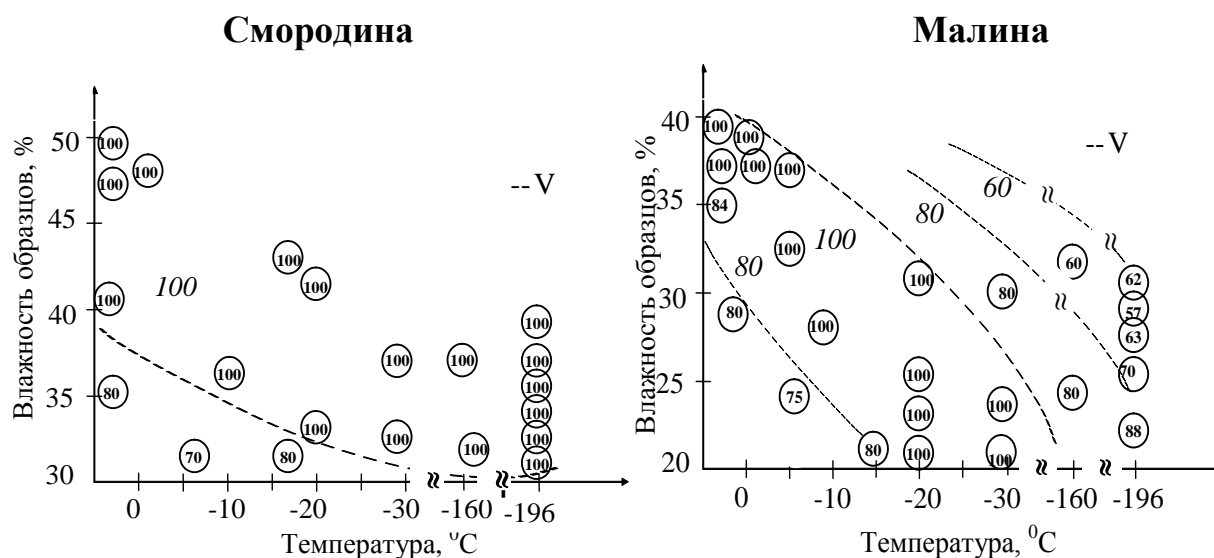


Рис. 19. Жизнеспособность ( $V$ , %) черенков в зависимости от их влажности ( $\eta$ , %) и температуры ( $T$ ,  $^{\circ}\text{C}$ ). Скорость охлаждения  $0,1^{\circ}\text{C/ч}$ , отогрева  $70^{\circ}\text{C/мин}$ .

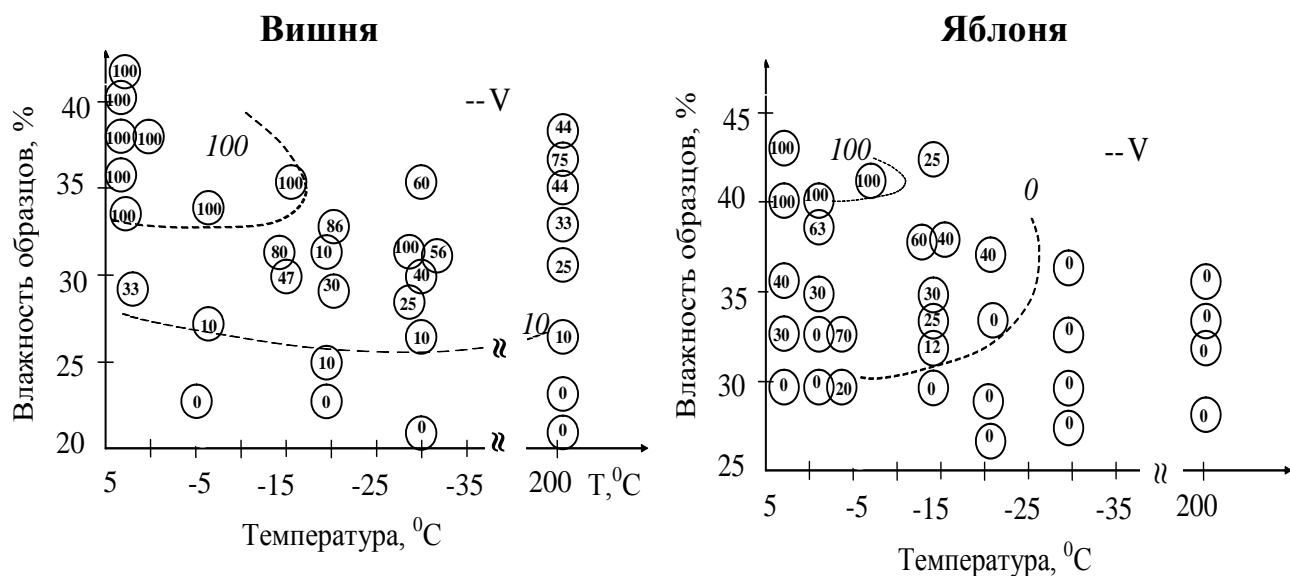


Рис.20. Жизнеспособность черенков в зависимости от их влажности ( $\eta$ , %) и температуры ( $T$ , °C). Скорость охлаждения – 0,1 °C/ч, отогрева – 70 °C/мин.

## 5.4. Оптимизация проведения многофакторного биологического эксперимента векторным методом

### 5.4.1 Принцип графического нахождения оптимума для двухфакторного эксперимента

При отсутствии специальных программ или навыков использования матричного исчисления, процедуру оптимизации проведения биотехнологических исследований можно значительно упростить следующим образом. Необходимо схему решения задачи оптимизации перевести из аналитической формы, требующей наличие регрессионных уравнений, в векторную. Наиболее простой способ представлен в следующем алгоритме:

1. Устанавливаем целевую функцию  $y(x_i)$ , которую необходимо привести к минимуму (или к максимуму) посредством варьирования определяющих её параметров  $x_i$ .

Поиск и построение поверхности отклика производится в системе абсолютных величин, а обобщение параметров, имеющих различную размерность, происходит благодаря вводу относительного показателя – коэффициента крутизны, определяющего шаг и направление изменения каждого параметра:

$$a_{ij} = \frac{x_{ij} - x_{\min j}}{x_{\max j} - x_{\min j}} \quad (8.2),$$

где  $a_{ji}$  – исследуемый параметр, представленный в относительной форме;  $x_{\max j}$  – максимальное значение исследуемого  $j$ -го параметра, выраженного в абсолютных единицах;  $x_{ij}$  – текущее значение исследуемого параметра;  $x_{\min}$  – минимальное значение исследуемого параметра.

2. Проведение опыта начинается после определения исходных значений исследуемых параметров  $a_i$ . Нулевой уровень выбирается по данным литературы, обеспечивающим наилучший результат. Если оптимальные значения ещё не установлены, то базовым уровнем можно считать значение целевой функции  $y_0(\bar{x}_i)$  при усредненных параметрах  $\bar{x}_i$ :

$$\bar{x}_i = \frac{x_{i \max} + x_{i \min}}{2} \quad (8.3),$$

где  $x_{i \max}$  и  $x_{i \min}$  – максимально и минимально допустимые значения исследуемых параметров, которые могут быть заданы в данном опыте.

Предложенный способ также имеет преимущество в сравнении с существующими аналогами (кроме метода слепого перебора), заключающееся в отсутствии ограничений в выборе направления и величины последующего измерения. То есть, следующее измерение может быть как произвольным, так и регламентированным предложенным алгоритмом. В любом случае полученные значения способствуют нахождению искомого оптимума.

4. Выбор направления последующего поиска осуществляется следующим образом. Для осуществления первого шага, можно произвести изменения величины целевой функции, уменьшив  $y_1(0,5\bar{a}_i)$  или увеличив  $y_1(1,5\bar{a}_i)$  её усредненные или выбранные параметры  $\bar{a}_i$  на половину. После проведения соответствующего опыта и установления величины  $y_1$  необходимо вычислить коэффициенты крутизны  $K_i(a_i)$  для каждого исследуемого параметра по следующей формуле:

5.

$$K_i = \frac{y_i - y_{i-1}}{\sqrt{\sum_{j=1}^n (a_{ji} - a_{ji-1})^2}} \quad (8.4),$$

где  $K_i$  – коэффициент крутизны подъема вектора  $i$ -го измерения.  $y_i$  и  $y_{i-1}$  – значения целевых функций, установленных в  $i$ -том опыте и в предшествующем –  $i-1$ ;  $j$  – номер параметра.

Для двухпараметрической функции, когда исследуются только два параметра  $x_1$  и  $x_2$ , формула имеет следующий вид:

$$K_i = \frac{y_i - y_{i-1}}{\sqrt{(a_{1i} - a_{1i-1})^2 + (a_{2i} - a_{2i-1})^2}} \quad (8.4a).$$

Таким образом, величина  $K_i$  показывает степень отклика целевой функции  $y$  на изменение  $x_i$  параметра. Другими словами,  $K_i$  показывает «крутизну подъема» целевой функции при варьировании  $x_i$  параметра. Проранжировав полученные коэффициенты по возрастанию их величин  $K_1 > K_2 > K_3 \dots$ , можно определить приоритет дальнейшего направления поиска.

4. Определение шага изменения параметров производится на каждом этапе процесса оптимизации, причем это изменение происходит пропорционально коэффициенту крутизны каждого вектора (по аналогии с методом крутого восхождения Бокса-Уилсона).

Выбор последующего шага определения  $x_i$  параметра для следующего эксперимента определяется из учета направления и величины вектора подъема:

$$\vec{h}_i = \frac{1}{K_{\max}} \sum_{j=1}^n K_{i-j} \cdot \vec{h}_{i-j} \quad (8.5),$$

где  $\vec{h}_i$  – вектор подъема для  $i$ -того опыта;  $n$  – общее количество исследуемых факторов;  $j$  – текущий параметр и соответствующие ему коэффициенты крутизны  $K_j$  и вектор  $\vec{h}_j$ ;  $K_{\max}$  – максимальное значение  $K_i$  коэффициента.

Кроме этого, возможен вариант произвольного изменения шага последующего измерения, например, по способу золотого сечения (по аналогии с методом Гаусса-Зейделя). Такой подход более рационален при наличии уже имеющихся обобщенных результатов, представленных в виде достоверных значений исследуемых величин, математических моделей, выраженных регрессионными и аналитическими уравнениями, опыта и интуиции экспериментатора и иных данных, прогнозирующих направление возрастания целевой функции.

Произвольный выбор шага дает возможность исследовать параметры, изменяющиеся дискретным образом. Например, в нашем случае величины скоростей замораживания-оттаивания не всегда можно изменять плавно, особенно в диапазоне высоких и сверхвысоких скоростей.

Для того, чтобы векторный метод при каждом новом измерении имел новый план, необходимо суммировать все вектора, кроме одного, имеющего минимальный коэффициент крутизны, за исключением последнего измерения. В случае если последнее измерение имеет минимальный коэффициент крутизны, а величина целевой функции возрастает, то рекомендуется произвести новое измерение в произвольной точке в непосредственной близости от последнего. Полученный таким образом новый план уточнит направление искомого градиента. Если величина целевой функции в новой точке оказывается меньше предыдущей, то этот факт является показателем достижения локального оптимума. Для нахождения глобального оптимума поиск производится параллельно или последовательно из различных максимально удаленных друг от друга точек факторного пространства.

Для двухпараметрической функции вектор подъема представляется следующим образом:

$$\vec{h}_i = K_{i-1} \cdot \vec{h}_{i-1} + K_{i-2} \cdot \vec{h}_{i-2} \quad (8.5a).$$

Вектор подъема  $\vec{h}_i$  для целевой функции, состоящей из 2–6 параметров, при отсутствии программного обеспечения более удобно определять не аналитическим способом, а графическим (рис. 29).

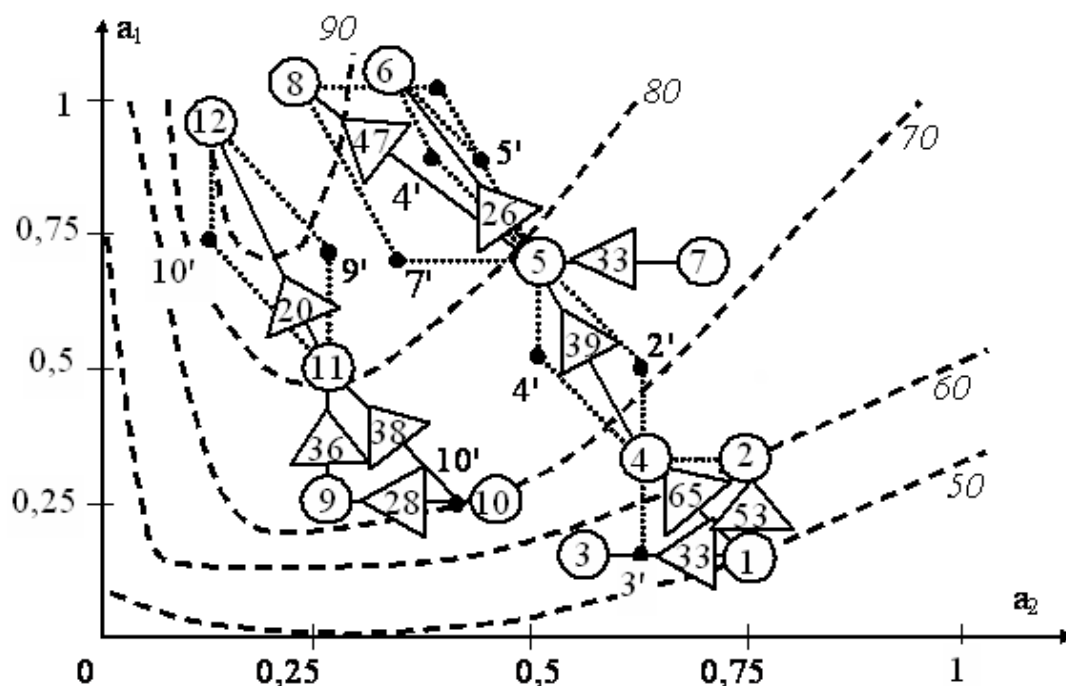


Рис. 29. Построение планов при каждом проведении измерения в процессе движения к экстремуму по поверхности отклика векторным методом.

По осям координат представлены исследуемые параметры  $x_i$ , а уровни поверхности отклика (значения целевой функции  $y$ ), имеющие одинаковые величины, представлены пунктирными линиями по аналогии с построением геодезических карт, отражающих уровень местности. Номер, охваченный кругом, показывает последовательность проведения опыта, а треугольником – показатель «крутизны» подъема вектора  $\phi$ . (8.4). Числа со штрихами обозначают проекции соответствующих векторов при их сложении.

Рассмотрим действие данного алгоритма на следующем примере. Параметры  $x_1$  и  $x_2$  представлены в относительной форме ( $a_1$  и  $a_2$ ) и изменяются от 0 до 1. Тогда для первого опыта параметры  $a_1$  и  $a_2$  составят  $a_1 = 0,15$  и  $a_2 = 0,75$ . Величина  $y_1 = 50$ . Для второго опыта  $y_2 = 60$ , а разница составит  $\Delta y_{12} = 10$ ,  $\Delta a_1 = 0,19$ , и для изменяемого параметра отношение разницы целевых функций  $\Delta y_{12}$  к разнице параметра  $\Delta a_1$  дает коэффициент крутизны  $K_2 = 53 \phi$ . (8.4a). Положительная величина знака указывает направление вектора (треугольник на рисунке повер-



нут в сторону второго опыта). Подобным образом меняем параметр  $a_2$  для третьего опыта и получаем  $K_3=33$ ,  $\Delta y=6$ ,  $\Delta a_2=0,18$ . Графическое сложение векторов по правилу «параллелограмма» с учетом веса коэффициента  $K$  задает величины параметров  $a_1=0,34$  и  $a_2=0,6$  для проведения четвертого опыта. Величина целевой функции при этом составляет  $y_4=65$ , а  $K_4=65$ . Продолжая движение в направлении увеличения градиента, определим величины параметров  $a_1$  и  $a_2$  для пятого опыта. Подобным образом выбираем следующее направление вектора подъема  $\rightarrow h$  с учетом наибольшего коэффициента крутизны –  $K$  при установлении величин параметров  $a_1$  и  $a_2$ . Поиск продолжается до тех пор, пока целевая функция  $y$  не выйдет на ограничения (6 опыт) или не примет максимальное значение, как это показано в восьмом опыте.

Если отдельные оптимальные значения функции отклика уже известны, то в данной области строится план (ПФЭ илиДФЭ) с целью подтверждения и описания поверхности отклика.

В качестве примера рассмотрим поверхности отклика, полученные в результате исследования зависимости показателя эффективности режима замораживания от скорости охлаждения в диапазоне медленных (рис. 30) и высоких скоростей (рис 31).

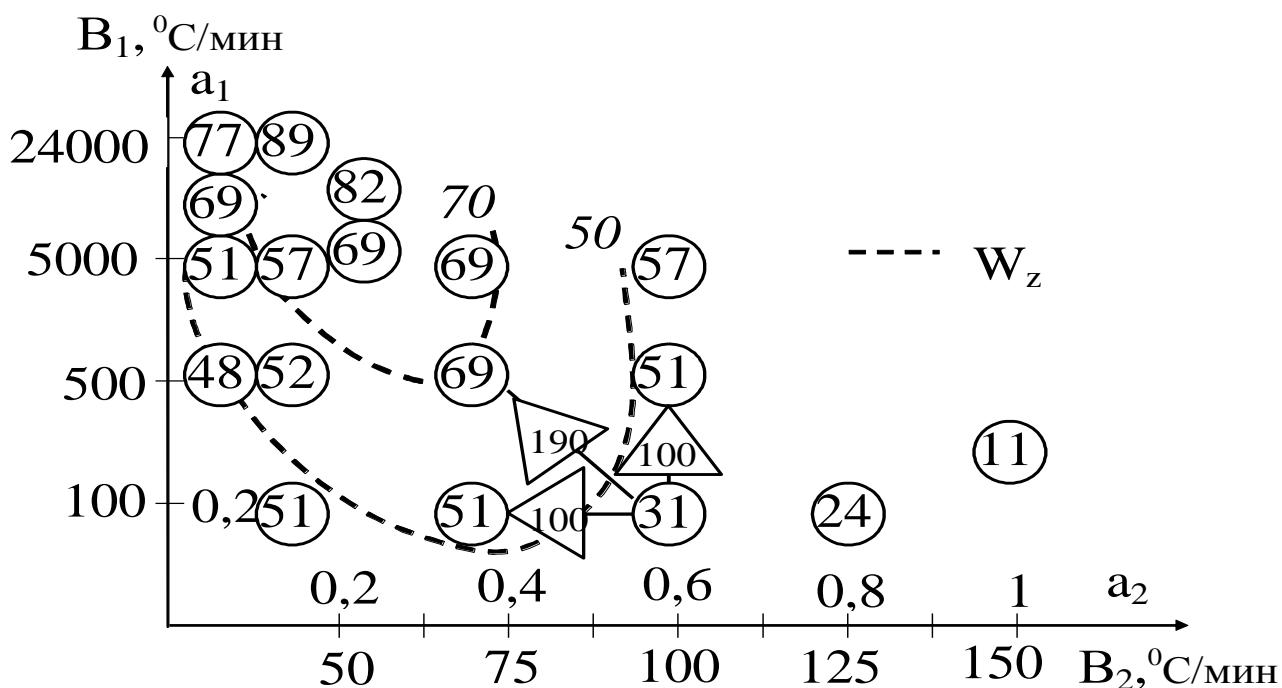


Рис. 30. Поиск максимальной величины сохранности деконсервированных спермиев быков при помощи предложенного векторного метода оптимизации, в процессе которого одновременно изменяются скорости оттаивания –  $V_{от}$  и замораживания –  $V_3$ : для  $V_{от}/V_3 < 10$  и  $V_{от}/V_3 > 70$ . Параметры  $a_1$  и  $a_2$  являются относительными величинами скоростей оттаивания и замораживания.

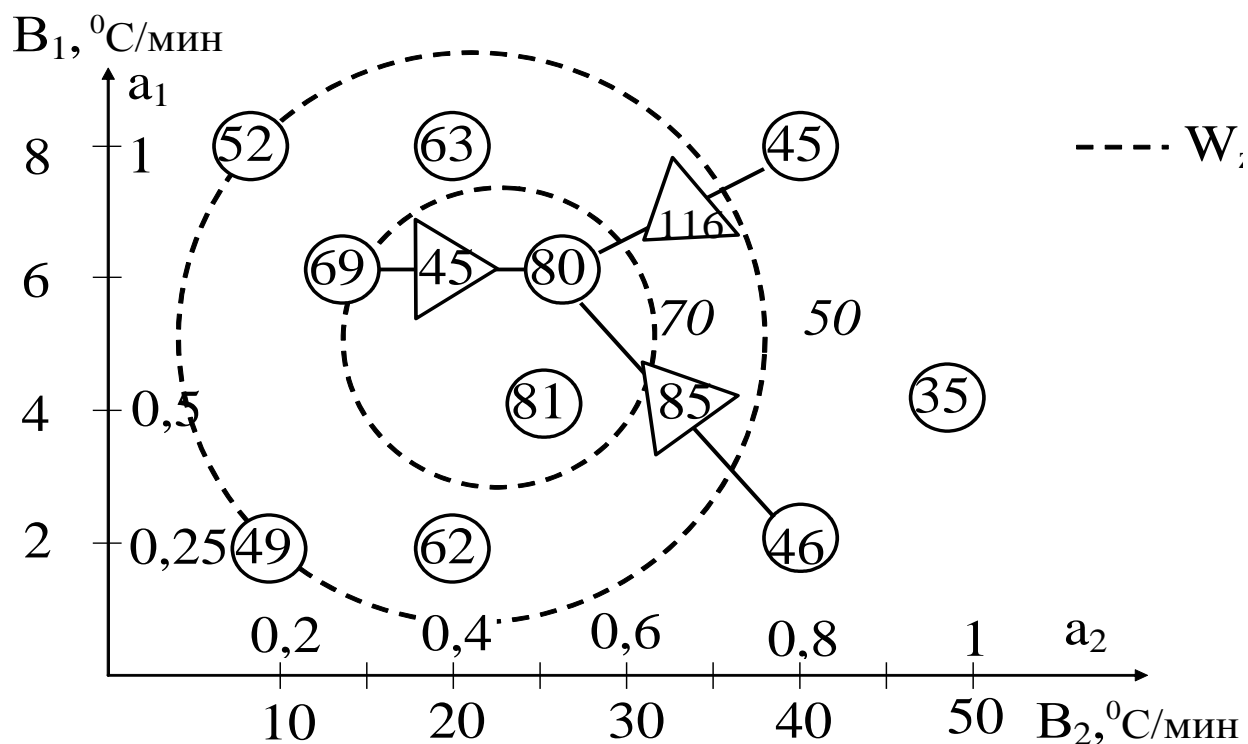


Рис. 31. Контуры эффективности режима замораживания-оттаивания спермиев карпа, отражающие зависимость совместного влияния скорости охлаждения на первой ступени  $B_1$  (от  $5^{\circ}\text{C}$  до  $-20^{\circ}\text{C}$ ) и на второй  $B_2$  (от  $-20^{\circ}\text{C}$  до  $-50$ ). Параметры  $a_1$  и  $a_2$  являются относительными величинами скоростей замораживания на первой и второй ступени, соответственно ф. (8.2). Номер очерченный кругом выделяет показатель относительной сохранности  $W_z$ , а треугольником величину «крутизны» подъема вектора ф. (8.4). Пунктирной линией обозначены одинаковые уровни абсолютных и относительных величин сохранности.

#### 5.4.2. Парная оптимизация параметров многофакторного эксперимента

При графическом решении поиска оптимума многопараметрической функции целесообразно по парно проранжировать ряд параметров, как это представлено на рис. 32 и 33, в соответствии с их рангом приоритетности (силой влияния на целевую функцию).

На первом этапе оптимум ищется для параметров  $x_1$  и  $x_2$  (при фиксированных  $x_3$  и  $x_4$ ) после нахождения, которого фиксируются оптимальные значения для первой пары и производится оптимизация для второй.

Поверхности отклика в графическом виде для четырехфакторного эксперимента ( $x_1, x_2, x_3, x_4$ ) выполненного векторным методом строится в несколько этапов (рис. 34 и 35).

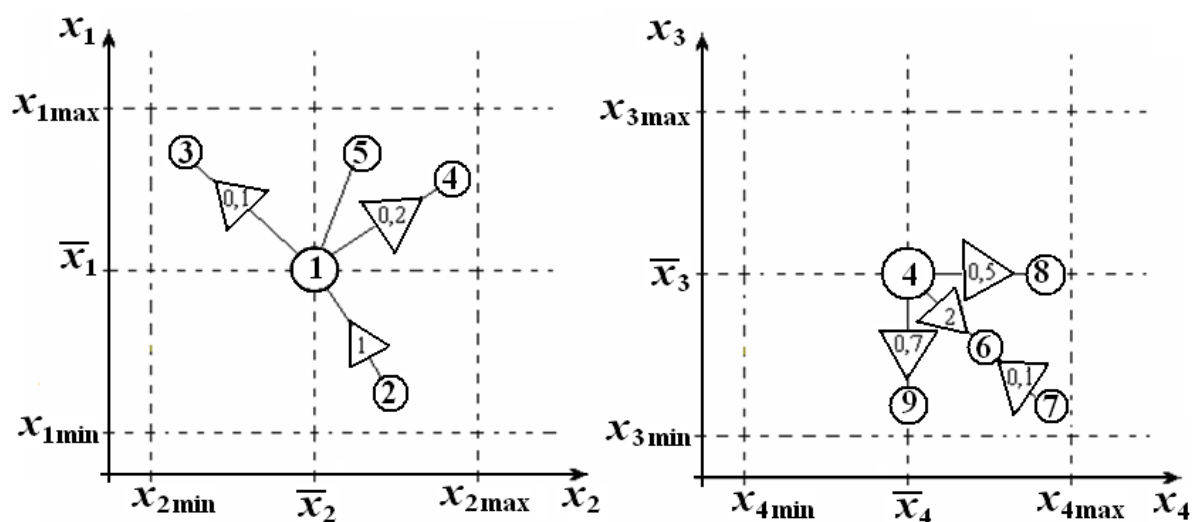


Рис. 32 и 33. Схема последовательности этапов проведения четырехфакторного эксперимента ( $x_1, x_2, x_3, x_4$ ) векторным методом в графической форме.

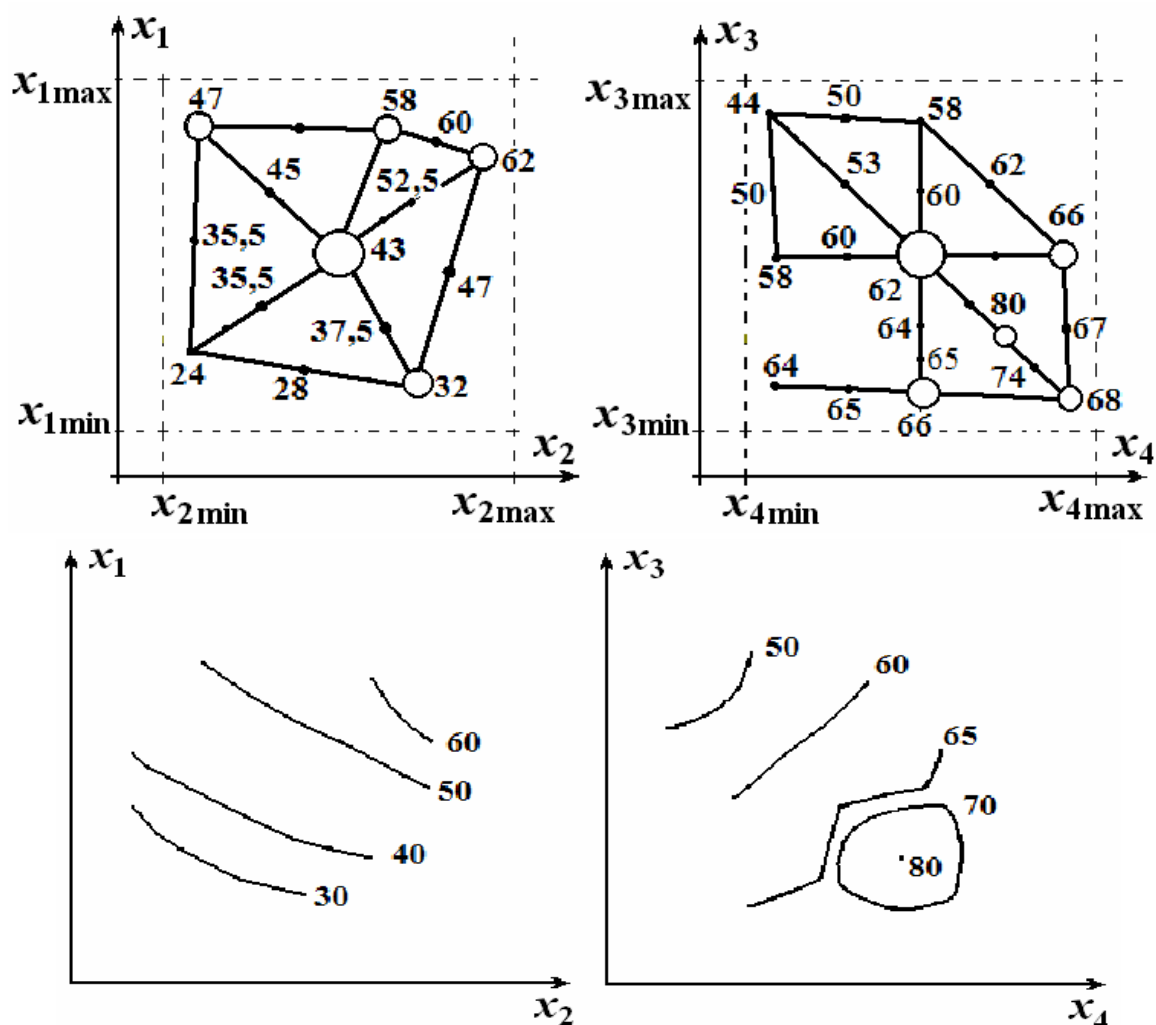


Рис. 34 и 35. Построение поверхности отклика на основе применения векторного метода кругового восхождения в графическом виде.

В качестве примера эффективности оптимизации взаимосвязи различных биологических и технологических параметров при помощи предложенного векторного способа, рассмотрим примеры криоконсервирования эмбрионов мыши,

замороженных при низких (прил. 2), высоких и сверхвысоких скоростях. Проведена оптимизация концентраций криопротекторов, определяющих состав криоконсерванта в пластиковых соломинках с внешним диаметром  $0,5 \div 2,0$  мм [25] (рис. 36-39).

соломинка  $d=2$  мм, экспозиция эмбрионов мыши в растворе:

Рис 36 - глицерина Гл 10 мин

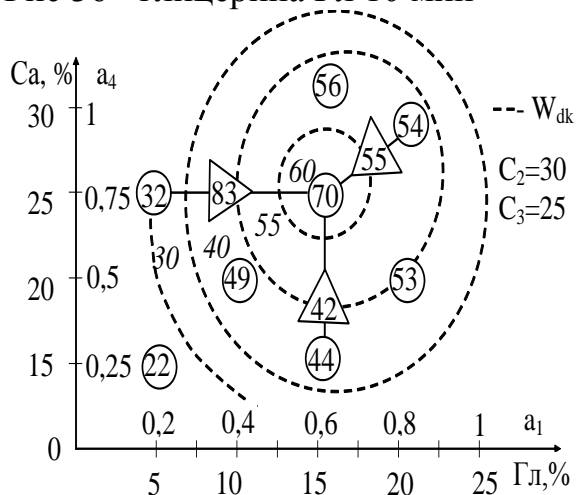
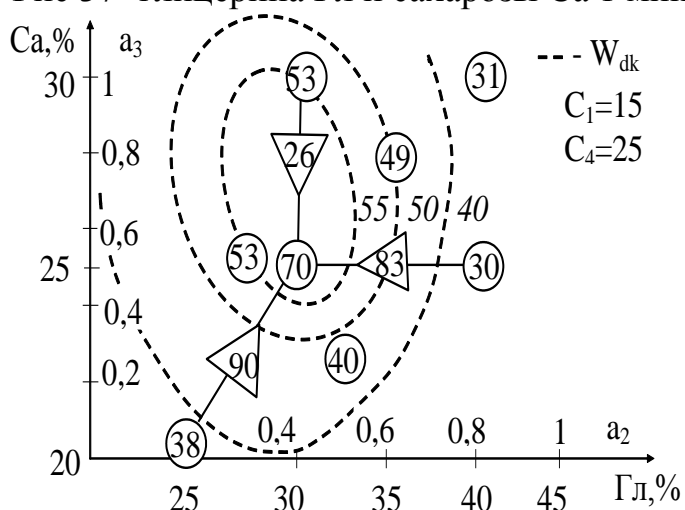


Рис 37- глицерина Гл и сахарозы Са 1 мин



соломинка  $d=0,5$  мм, экспозиция эмбрионов мыши в растворе:

Рис 38 - глицерина Гл 10 мин

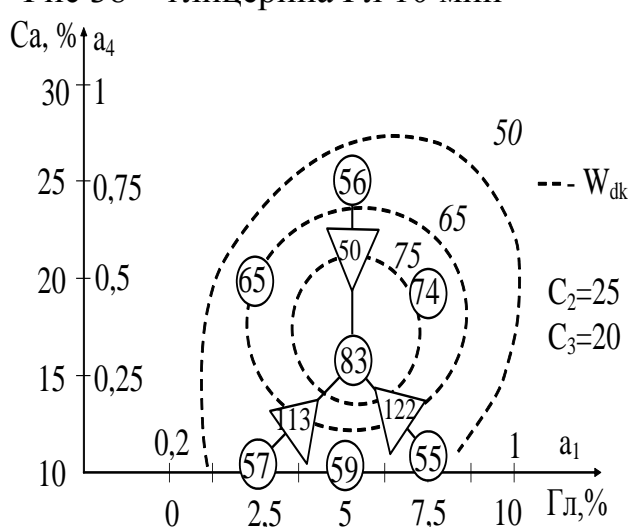


Рис 39 - глицерина Гл и сахарозы Са 1 мин

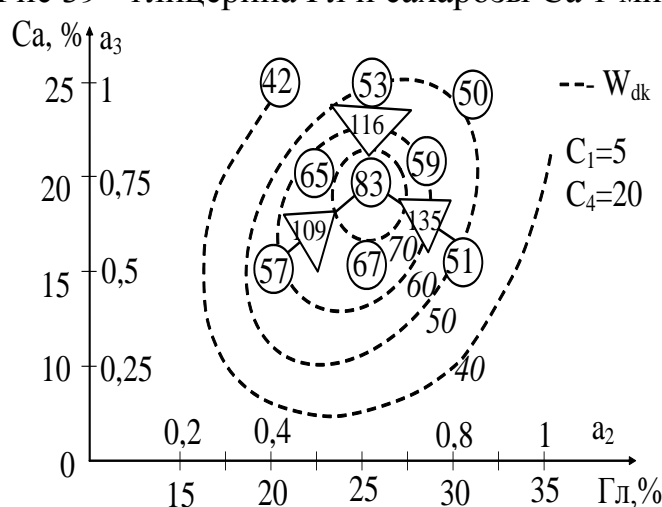


Рис. 36-39. Обобщенная эффективность криоконсервирования  $W_{dk}$  эмбрионов мыши в зависимости от выбора концентрации глицерина Гл и сахарозы Са при замораживании в пластиковой соломинке  $d=0,5$  (2) мм со скоростью  $V_3=1680$  ( $12 \cdot 10^3$ )  $^{\circ}\text{C}/\text{мин}$  и отогревом  $V_0=2520$  ( $14,4 \cdot 10^3$ )  $^{\circ}\text{C}/\text{мин}$ . Параметры  $a_1$  и  $a_2$  – исследуемые концентрации криопротектора (Гл и Са) ф. (8.2).

Полученная графическая зависимость эффективности криоконсервирования эмбрионов от изменения концентрации эквilibрирующего  $C_1=15$  % и выводящего раствора  $C_4=25$  %, вблизи его оптимального значения показывает, что ко-

эффективности «крутизны» составляют 83 и 42 у.е. при замораживании эмбрионов мыши в 2 мм соломинке (рис. 36). Относительно близкие значения 83 и 26 у.е. получены при вариации концентрации глицерина и сахарозы в витрифицирующем растворе (рис. 37).

Уменьшение диаметра пластикового капилляра до 0,5 мм увеличило скорость его охлаждения в семь раз, что позволило снизить концентрацию раствора глицерина на 10 % и повысить эффективность криоконсервирования до 82 %. Сохранность девитрифицированных эмбрионов мыши увеличилась от 52 до 70 %. Эта процедура привела к увеличению силы влияния оптимизируемых параметров на эффективность криоконсервирования эмбрионов приблизительно в  $1,5 \div 2,6$  раза, в первом растворе до 95, а в четвертом до 50 у.е. (рис. 38). Для второго и третьего раствора до 170 и 116 у.е. (рис. 39) при оптимизации составляющих концентраций раствора глицерина (Гл) и сахарозы (Са), с короткой экспозицией в нем эмбрионов мыши. Уровень сохранности деконсервированных эмбрионов мыши при последующем развитии в культуре *in vitro* составил  $70,1 \pm 6,1$  %,  $n=54$ .

Таким образом, проведенное многофакторное исследование дало возможность определить оптимальные параметры криоконсервирования эмбрионов мыши при высоких и сверхвысоких скоростях замораживания. Максимальная эффективность криоконсервирования 83 % получена для скорости  $11,8 \times 10^3$  °С/мин. Уровень сохранности деконсервированных эмбрионов мыши при последующем развитии в культуре *in vitro* составил  $70,1 \pm 6,1$  %,  $n=54$ .

#### **5.4.3 Пример проведения оптимизации криозащитной среды и режима замораживания эмбрионов мыши, основанного на применении низких скоростей**

Охлаждения эмбрионов животных реализуются посредством применения постоянных и переменных скоростей замораживания. Проведение сравнительного анализа эффективности данных способов и их этапов, на основе предложенных показателей, возможно при выполнении предварительной оптимизации параметров, определяющих жизнеспособность деконсервированного биообъекта.

Общеизвестно, что оптимальной скоростью замораживания для эмбрионов мыши и коровы является величина  $0,3$  °С/мин, которая обеспечивает необходимую степень обезвоживания клеток и тем самым препятствует образованию внутриклеточного льда. При этом сами клетки находятся в течение нескольких часов в гипертоническом растворе. Снизить пагубность «эффекта раствора» возможно за счёт оптимизации характера изменения скорости замораживания посредством применения экспоненциальных режимов охлаждения.

Для определения влияния характера изменения скорости охлаждения на сохранность эмбрионов мыши проведено многофакторное исследование, направленное на установление оптимальных значений параметров: концентрации криопротектора, скорости охлаждения и величины её ускорения, конечной

температуры охлаждения и времени выдержки биообъекта перед погружением его в жидкий азот.

Для изучения влияния изменения скорости охлаждения эмбрионов на уровень их сохранности после оттаивания разработан способ, позволяющий реализовывать разные режимы охлаждения до различных конечных температур. Получение необходимого режима охлаждения биообъекта производится на основе пассивного остывания термоблока в горловине сосуда Дьюара. Скорость охлаждения биообъекта изменяется посредством открытия поверхности термоблока, что увеличивает пропускную способность паров азота к упаковкам. Изменяя степень открытости термоблока ( $0 \div 100 \%$ ) и конечную температуру охлаждения ( $+5 \div -70$  °C), зависящую от глубины погружения термоблока в горловину сосуда Дьюара ( $3 \div 27$  см), можно реализовать требуемую величину скорости замораживания ( $0,1$  до  $15$  °C/мин) эмбрионов.

**Оптимизация конечной температуры охлаждения и времени выдержки эмбрионов мыши перед погружением биообъекта в жидкий азот.** Для определения конечной температуры охлаждения и времени выдержки эмбрионов мыши перед погружением в жидкий азот было проведено охлаждение исследуемого биообъекта, реализованное при помощи устройства, в котором различные конечные температуры охлаждения эмбрионов получали посредством изменения глубины погружения термоблока в горловину сосуда Дьюара. В качестве криопротектора использовали  $1$  М раствор глицерина; биообъект был помещен в пробирки Уленгута. Величина конечной температуры охлаждения определялась при помощи полученной регрессионной зависимости.

Для определения силы влияния варьируемых параметров использовался двухфакторный дисперсионный анализ. Исследуемыми факторами являлись: конечная температура ( $T_k$ ), до которой производится охлаждение биообъекта, и время выдержки эмбрионов мыши при данной температуре с последующим погружением биообъекта в жидкий азот. Уровнями градации температурной выдержки эмбрионов является  $-30$  и  $-35$  °C; время экспозиции составляло  $0$  и  $15$  мин (табл. 7).

По показателю сохранности выявлено достоверное влияние времени выдержки эмбрионов  $\eta^2_A=0,40$  перед их погружением в жидкий азот и конечной температуры  $\eta^2_B=0,21$ . Совместное влияние времени выдержки и конечной температуры составило  $\eta^2_{AB}=0,01$ . Показатель жизнеспособности и эффективности дал такую же силу влияния времени выдержки  $\eta^2_A=0,40$  и в два раза больше температуры  $\eta^2_B=0,45$ . Совместное влияние времени выдержки и конечной температуры охлаждения на жизнеспособность деконсервированных эмбрионов составило  $\eta^2_{AB}=0,86$ . Уровень надежности полученных результатов составил  $P \geq 0,95$ .

Для описания контуров эффективности криоконсервирования эмбрионов, зависящих от конечной температуры охлаждения и времени выдержки биообъекта, проведено еще четыре дополнительных опыта  $5 \div 8$  (табл. 7). Поиск оптимальных значений времени и температурной выдержки проводился в расширенной области исследования: конечная температура  $T_k$  °C, до которой производят-

ся охлаждение биообъекта, в интервале от  $-25$  до  $-37,5$  °C; время выдержки эмбрионов мыши при данной температуре ( $T_k$ ) с последующим погружением биообъекта в жидкий азот составляло  $t_k \{0 \div 30 \text{ мин}\}$ .

Таблица 7. Определение конечной температуры охлаждения ( $T_k$ ) и времени выдержки эмбрионов мыши ( $t_k$ ) перед погружением их в жидкий азот по показателям состояния деконсервированного биообъекта: сохранности  $S_{dk}$ , жизнеспособности  $V_{dk}$ , обобщенной эффективности криоконсервирования  $W_{dk}$

Номер опыта	Время выдержки $t_k$ , мин.	Конечная температура $T_k$ , °C	Количество, $n_1/n$	Вероятность развития в культуре $M \pm m$ , %		
				Сохранность $S_{dk}$	Жизнеспособность $V_{dk}$	Эффективность $W_{dk}$
1	0	-30	14/16	$87,5 \pm 6,2$	$71,9 \pm 0,9$	$90,1 \pm 0,6$
2	15	-30	11/16	$68,8 \pm 5,9$	$65,0 \pm 3,2$	$86,3 \pm 4,9$
3	0	-35	11/14	$78,6 \pm 1,7$	$73,2 \pm 0,3$	$90,3 \pm 1,1$
4	15	-35	12/16	$75,0 \pm 4,4$	$65,6 \pm 2,1$	$79,3 \pm 2,6$
5	30	-35	10/16	$80,7 \pm 7,6$	$60,1 \pm 3,3$	$76,4 \pm 3,7$
6	0	-32,5	12/14	$72,2 \pm 5,6$	$73,9 \pm 3,7$	$91,3 \pm 4,7$
7	0	-25	13/18	$63,6 \pm 8,8$	$55,8 \pm 6,3$	$68,6 \pm 7,5$
8	7,5	-30	12/16	$75 \pm 7,3$	$68,8 \pm 4,3$	$84,2 \pm 5,0$

Примечание:

1.  $n$  – общее количество эмбрионов мыши (отличного, хорошего и удовлетворительного качества);

2.  $n_1$  – количество пригодных эмбрионов к дальнейшему использованию, полученных после оттаивания и культивирования,  $S_k=86,7 \pm 8,8$  – сохранность (жизнеспособность  $V_k=81,3 \pm 6,1$ ) нативных эмбрионов в культуре.

Для нахождения градиента роста относительной жизнеспособности деконсервированных эмбрионов проведено три опыта. В четвертом опыте найдена максимальная величина жизнеспособности деконсервированных эмбрионов. Двухфакторная оптимизация при помощи векторного метода дала возможность установить максимальную эффективность криоконсервирования эмбрионов  $W_{dk}=91\%$  (рис.46) при конечной температуре охлаждения  $T_k= -32$  °C и времени выдержки  $t_k=0$ .

Для построения поверхности отклика (контуров эффективности) использовали методы П. Берча и В. Клеймана с определенной модификацией. В отличие от предложенных аналогов экспериментальные точки в нашем случае располагались произвольным образом. При необходимости построения полной картины

поверхности отклика, методом линейной экстраполяции находили расчетные точки на продолжениях линейных отрезков (на расстояниях, равных длинам предшествующих отрезков). Соседние точки соединяли линейными отрезками и методом линейной интерполяции определяли расчетные значения функции отклика (величины эффективности) в серединах этих отрезков. Пунктирными линиями соединяли точки, имеющие одинаковые значения показателей эффективности.

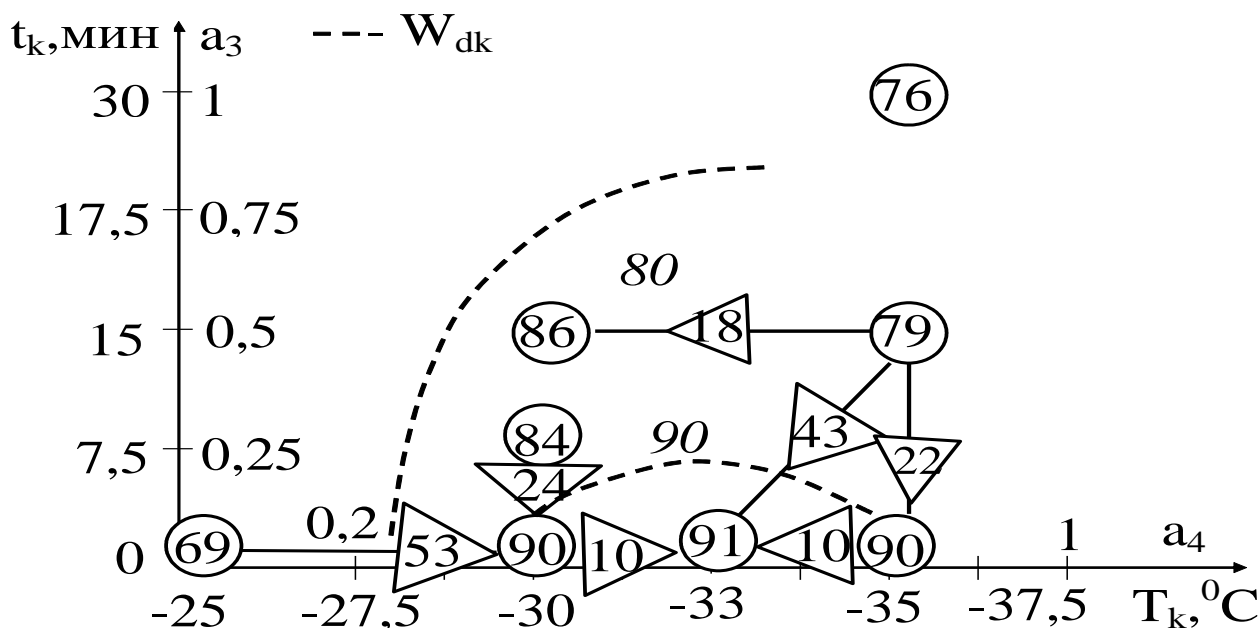


Рис. 46. Эффективность криоконсервирования ( $W_{dk}$ ) эмбрионов мышцы в зависимости от выбора конечной температуры охлаждения ( $T_k$ ) и времени выдержки ( $t_k$ ) перед погружением биообъекта в жидкий азот. Номер, очерченный кругом, отражает показатель эффективности, а треугольником - величину «крутизны» подъема вектора  $\phi$  (8.4). Пунктирной линией обозначены контуры эффективности криоконсервирования, имеющие одинаковые величины. Параметры  $a_3$  и  $a_4$  - исследуемые величины ( $T_k$  и  $t_k$ ), представленные в относительной форме  $\phi$  (8.2).

Анализ контуров обобщенной эффективности криоконсервирования эмбрионов показывает, что оптимизируемые параметры (время выдержки и конечная температура) оказывают значимое влияние на сохранность биообъекта. Коэффициент «крутизны», отражающий изменение относительной жизнеспособности биообъекта и рассчитанный для конечной температуры охлаждения, изменяется от 10 до 53 у.е, а для времени выдержки перед его погружением в жидкий азот -  $22 \div 46$  у.е в зависимости от области факторного пространства, для которого он был рассчитан. Рассчитанные выше показатели силы влияния, выраженные в процентах, дают сопоставимые величины 40 и 45 %. Вместе с тем в области оптимума вариация времени выдержки эмбрионов оказывает более значимое влияние чем температура, коэффициенты «крутизны» которых составляют 45 и 10 у.е. соответственно. Следует обратить внимание на то, что конечная температура охлаждения биообъекта и время его выдержки перед по-



гружением в жидкий азот также зависят от режима охлаждения, то есть величины скорости и характера её изменения.

Применение векторного способа оптимизации конечной температуры охлаждения и времени выдержки эмбрионов мыши перед их погружением в жидкий азот по пяти уровням градации дало возможность описать зависимость изменения сохранности эмбрионов мыши и эффективности их криоконсервирования по 8 проведенных опытах в сравнении с 25 необходимыми для традиционно используемого способа латинского квадрата.

Таким образом, проведенная двухфакторная оптимизация дала возможность сократить количество опытов в 3,1 раза при определении максимальной эффективности криоконсервирования эмбрионов мыши  $W_{dk}=91\%$ , соответствующей конечной температуре охлаждения  $T_k=-32\text{ }^{\circ}\text{C}$  и отсутствии временной выдержки биообъекта перед его погружением в жидкий азот.

**Изучение влияния характера изменения скорости охлаждения на сохранность эмбрионов мыши.** Для получения различных экспоненциальных режимов охлаждения эмбрионов мыши применялся способ «помещение биообъекта в уже охлажденный термоблок». Для этого контейнеры с биообъектом (5) и (6) размещали в предварительно охлажденном термоблоке в отверстия, которые расположены в диске (4). Для получения необходимой конечной температуры охлаждения термоблока (3) его располагали на глубине  $L=11,5\text{ см}$  от верхнего края горловины сосуда Дьюара. Предварительное охлаждения термоблока длилось от 0 до 60 минут, что определяло скорость охлаждения контейнера содержащего биообъект.

Для определения оптимального режима замораживания термоблок погружали на заданную глубину с целью достижения конечной температуры охлаждения биообъекта  $-30\pm 2\text{ }^{\circ}\text{C}$ , при этом время предварительного охлаждения блока составляло от 0 до 60 мин. После охлаждения термоблока в него помещали пробирку Уленгута или соломинку, выдерживая в течение 120 мин для обезвоживания клеток. После выхода режима на заданную (конечную) температуру охлаждения, контейнеры, содержащие эмбрионы, переносили в жидкий азот.

Величина средневзвешенной скорости охлаждения ( $\bar{V}$ ,  $^{\circ}\text{C}/\text{мин}$ ) соломинок и пробирок Уленгута в диапазоне от  $-5$  до  $-25\text{ }^{\circ}\text{C}$  имела разные значения ( $V$ ,  $^{\circ}\text{C}/\text{мин}$ ), а от  $-5$  до  $-30\text{ }^{\circ}\text{C}$  составляла  $0,3\text{ }^{\circ}\text{C}/\text{мин}$  (табл. 8, рис. 47).

Для описания полученных режимов охлаждения термоблока использовалось регрессионное выражение:

$$T = a_0 + a_1 \cdot \ln t, \quad (1)$$

где:  $T$  – температура,  $^{\circ}\text{C}$ ;  $t$  – время выдержки термоблока;  $a_0$ ,  $a_1$  – коэффициенты регрессионного уравнения.

Для удобства обозначения каждого режима ему присвоен номер в соответствии длительности предварительного охлаждения термоблока:

$$\begin{aligned} T &= 27,1 + 12,1 \cdot \ln t & (2) & \quad T = 9,3 + 8,9 \cdot \ln t & (3) & \quad T = 3,3 + 7,6 \cdot \ln t & (4) \\ T &= 0,68 + 9,3 \cdot \ln t & (5) & \quad T = 0,4 + 10,9 \cdot \ln t & (6) \end{aligned}$$

Таблица 8. Зависимость скорости замораживания ( $V$ ,  $^{\circ}\text{C}/\text{мин}$ ) от времени предварительного охлаждения термоблока ( $t_{\text{по.}}$ , мин)

Время предварительного охлаждения термоблока $t_{\text{по.}}$ , мин	Средняя скорость охлаждения в температурном диапазоне от $-5$ до $-25^{\circ}\text{C}$ ( $-5 \div -30^{\circ}\text{C}$ ) $V$ , $^{\circ}\text{C}/\text{мин}$	
	Пробирки Уленгута	Пластиковые соломинки
0	0,4 (0,3)	0,5 (0,3)
20	0,8 (0,3)	1,4 (0,3)
40	1,1 (0,3)	2 (0,3)
60	2 (0,3)	3,3 (0,3)

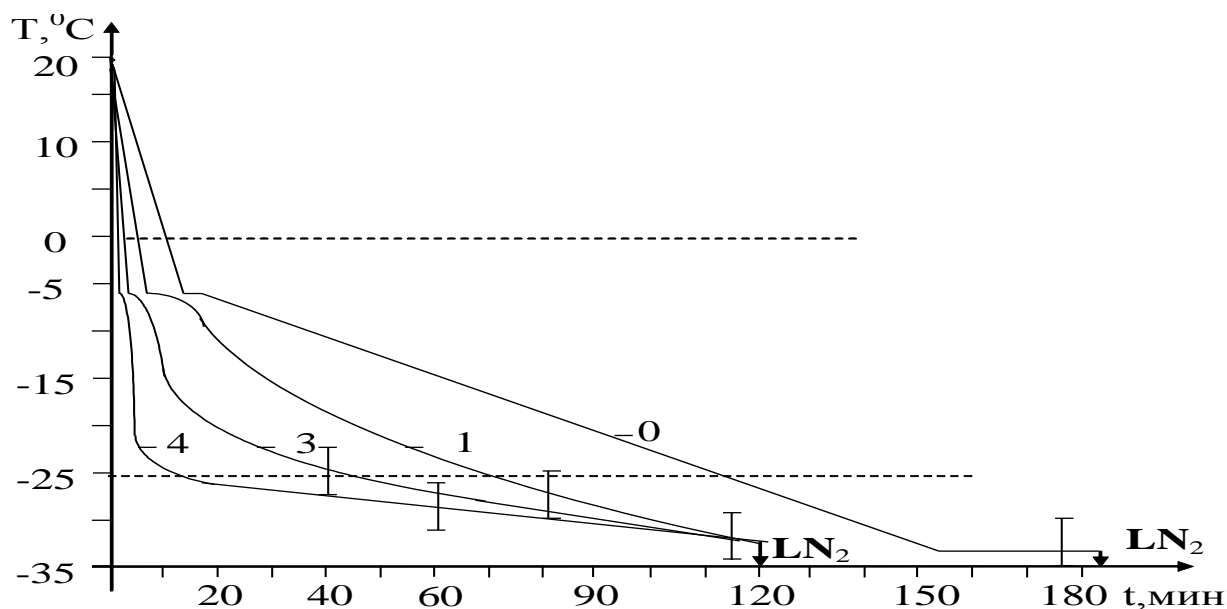


Рис. 47 Режимы охлаждения эмбрионов млекопитающих в горловине сосуда Дьюара: 0 - контролируемое программное замораживание с постоянной скоростью  $0,3^{\circ}\text{C}/\text{мин}$ ; 1 - режим замораживания без предварительного охлаждения термоблока; 3 - в предварительно охлажденном термоблоке в течение 20 минут; 4 - в предварительно охлажденном термоблоке в течение 40 минут.

Для повышения сохранности деконсервированных эмбрионов мыши, исследовалось совместное действие изменения концентрации криопротектора и режима замораживания. Концентрацию криопротектора изменяли в диапазоне от 5 до 15% глицерина. В качестве тестируемых режимов апробировано несколько вариантов охлаждения эмбрионов мыши: с постоянной скоростью  $V=0,3^{\circ}\text{C}/\text{мин}$ , полученной при помощи программного замораживателя ЗЭМ-3 –  $N=0$  [54]; переменной, основанной на пассивном охлаждении термоблока [100] в горловине сосуда Дьюара ф (2) –  $N=1$ ; замораживание эмбрионов в термоблоке, предварительно охлажденном в горловине сосуда в течение 10 мин ф. (3) –  $N=2$ ; с пред-

варительным охлаждением термоблока в течение 20 мин ф. (4) – N=3; с охлаждением 40 мин ф. (5) – N=4; 60 мин – N=5 ф. (6).

На первом этапе исследования для определения силы влияния варьируемых параметров использовался двухфакторный дисперсионный анализ с двумя уровнями градации (опыты 1 ÷ 4 табл. 9). По показателю сохранности выявлено существенное влияние режима замораживания эмбрионов  $\eta^2_A=0,67$  и незначительное по концентрации криопротектора  $\eta^2_B=0,09$ . Совместное влияние режима замораживания и концентрации криопротектора составило силу влияния  $\eta^2_x=0,76$ , а их взаимосвязь  $\eta^2_{AB}=0,01$ .

Таблица 9. Выбор экспоненциального режима замораживания и концентрации криопротектора в комплексе по показателям состояния деконсервированного биообъекта: сохранности  $S_{dk}$ , жизнеспособности  $V_{dk}$ , обобщенной эффективности криоконсервирования  $W_{dk}$

Но- мер опыта	Концентра- ция криопро- тектора, Гл, %	режима охлажде- ния, N	Коли- чество, $n_1/n$	Вероятность развития в культуре $M \pm m$ , %		
				$S_{dk}$	$V_{dk}$	$W_{dk}$
1	10	2	13/15	87,5±6,2	72,2±4,3	89,9±4,6
2	10	4	8/14	57,3±3,3	49,7±1,6	63,1±1,2
3	15	2	12/16	75±7,3	64,4±3,8	76,4±3,0
4	15	4	7/14	50,0±5,8	42,8±2,9	53,5±3,2
5	5	1	10/14	71,4±10,0	64,6±5,1	80,5±4,4
6	10	1	14/16	87,5±6,2	77,2±3,4	92,3±4,5
7	10	0	14/16	87,5±6,2	73,1±3,2	87,7±4,1
8	12,5	3	12/16	74,4±7,3	66,2±3,2	77,2±2,6
9	7,5	3	10/16	62,2±2,2	60,9±2,2	78,8±3,9

**Примечание:**

1.  $C_1$  - концентрация криопротектора с длительной экспозицией ( $t=10$  мин,  $T=20 \pm 2$  °C);
2.  $n$  – общее количество эмбрионов мыши (отличного, хорошего и удовлетворительного качества);
3.  $n_1$  – количество пригодных эмбрионов. Сохранность нативных эмбрионов в культуре –  $S_k=86,7 \pm 8,8$ ;
4. Конечная температура охлаждения биообъекта составляет  $T_k=-32$  °C и время выдержки  $t_k=0$ .

Показатели жизнеспособности и эффективности дали большее значение влияния режима замораживания  $\eta^2_A=0,77$  и  $\eta^2_A=0,71$  соответственно. Сила влияния фактора, отвечающего за концентрацию криопротектора, составила  $\eta^2_B=0,09$  и  $\eta^2_B=0,18$ . Совместное влияние режима замораживания и концентрации криопротектора составило силу влияния  $\eta^2_x=0,86$  и  $\eta^2_x=0,89$  соответственно. Уровень надежности полученных результатов составляет  $P \geq 0,95$ .

Следует отметить, что при использовании показателей сохранности и жизнеспособности биообъекта, влияние концентрации криопротектора на биообъект – 9 %, а эффективности 18 %. Поиск оптимальных значений, определяющих концентрацию криопротектора и режим замораживания, производился на основе применения векторного метода при помощи дополнительно проведенных опытов 5 ÷ 9 (табл. 9). Совместная оптимизация концентрации криопротектора и режима замораживания показала, что максимальная жизнеспособность эмбрионов после оттаивания получена при использовании 10 % раствора глицерина для режимов N 1 и N 2. При использовании скорости охлаждения  $0,3 \div 0,6$  °C/мин в температурном диапазоне от -5 до -25 °C получена жизнеспособность (эффективность криоконсервирования) эмбрионов мыши  $78 \div 80$  % ( $88 \div 93$  %). При увеличении скорости от 0,8 до 3,3 °C/мин показатели жизнеспособности снизились от 75 (76 %) до 57 % (63 %) (рис. 48).

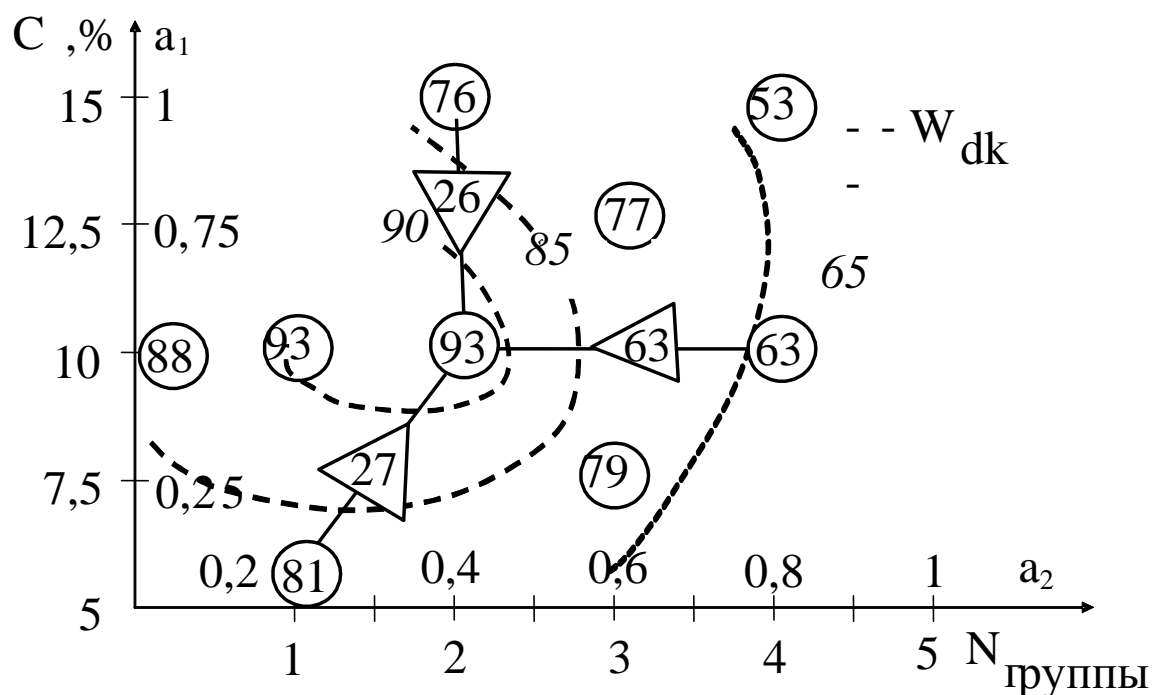


Рис. 48 Обобщенная эффективность криоконсервирования ( $W_{dk}$ ) эмбрионов мыши в зависимости от концентрации глицерина (Гл) и выбора режима замораживания (N). Конечная температура охлаждения биообъекта составляет  $T_k=-32^\circ\text{C}$  и время выдержки  $t_k=0$ . Параметры  $a_1$  и  $a_2$  - исследуемые параметры (C и N) представленные в относительной форме ф. (8.2).

Анализ контуров эффективности криоконсервирования эмбрионов мыши показывает, что совместное изменение концентрации криопротектора и режима охлаждения оказывает значимое влияние на сохранность биообъекта. Влияние режима охлаждения отражает коэффициент «крутизны» равный 63 у.е., а изменение концентрации глицерина в два раза ниже - 26 у.е. Приблизительно такие же величины получены для показателей, отражающих силу влияния данных факторов на эффективность выбранного способа криоконсервирования 71 и 18 % соответственно.

Сохранность деконсервированных эмбрионов мыши  $87,5 \pm 8,3$  % полученная при линейных  $N=0$  и не линейных  $N=1$ ,  $N=2$  режимах замораживания (табл. 9), согласуется с данными  $80 \div 90$  % представленными в литературе [26]. Эффективность режимов замораживания составляет  $96 \div 98$  %. Полученные данные показывают, что однофакторный поиск концентрации криопротектора, режима замораживания, конечной температуры охлаждения биообъекта и времени выдержки перед погружением его в жидкий азот, обеспечивает такие же значения, как и при использовании многофакторной оптимизации.

Таким образом, многофакторная оптимизация режима замораживания эмбрионов мыши дала возможность определить допустимый диапазон изменения скорости охлаждения  $0,3 \div 0,6$  °C/мин в температурном диапазоне от -5 до -25 °C, что позволило реализовать показатель жизнеспособности деконсервированных эмбрионов мыши на уровне 83,3 %.

#### ВОПРОСЫ ДЛЯ САМОКОНТРОЛЯ.

1. Опишите особенности методов однофакторной оптимизации исследования
2. Какова необходимость выделения существенных факторов?
3. Охарактеризуйте методы случайного поиска в задачах экспериментальной оптимизации биотехнологических параметров.
4. Каковы особенности применения метода покоординатного спуска (Гаусса-Зейделя)?
5. Каковы особенности применения метода Розенброка?
6. Какова необходимость проведения многофакторного эксперимента?
7. Дайте сравнительный анализ существующих методов многофакторной оптимизации исследования.
8. Каковы особенности применения метода оптимизации Genetic Algorithms?
9. Как осуществляется исследование поверхности отклика на основе уравнений регрессии второго порядка?
10. Каков порядок проведения оптимизация проведения многофакторного биологического эксперимента векторным методом?
11. Опишите принцип графического нахождения оптимума для двухфакторного эксперимента.

## Глава 6 Математическое моделирование как инструмент научного исследования

### 6.1 Влияние космофизических факторов на случайные процессы исследованные С.Э.Шнолем.

В 1951-1956 г.г. С.Э.Шноль исследуя «разброс результатов» измерения скорости гидролиза АТФ, катализируемой белками актомиозинового комплекса, установил, что параллельные опыты имеют расхождение до 1,5%, а последовательные во времени до 24% [34].

Проведенные исследования в течении 50 лет дали возможность автору определить:

1 «Разброс результатов измерений» свойственен процессам любой природы от биохимических реакций до радиоактивного распада. На рис. 1 представлены спонтанные изменения ферментативной активности в растворе актомиозина, в последовательных пробах. Среднеквадратичная амплитуда «разброса результатов» в % к средней величине равна 23% (коэффициент вариации).

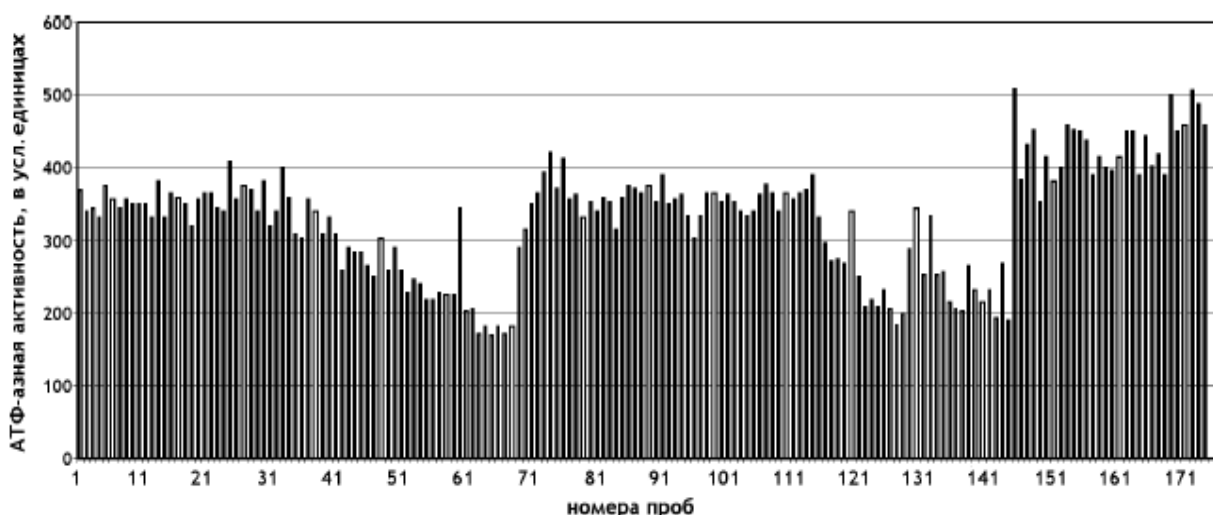


Рис. 1. Изменения АТФ - азной активности в растворе актомиозина. По оси абсцисс – номера 15 секундных интервалов между пробами. По оси ординат – ферментативная (АТФ-азная) активность, в условных единицах [34].

2 Величина разброса относительно измеряемой величины различна для процессов разной природы.

3 Форма гистограмм в одно и то же время, в данном географическом пункте сходна для любых процессов.

На рисунке 2 отражены гистограммы сходной «идеи формы» наблюдаемые при измерениях креатинкиназой и реакции аскорбиновой кислоты (АК) с дихлорфенолиндофенолом (ДХФИФ).

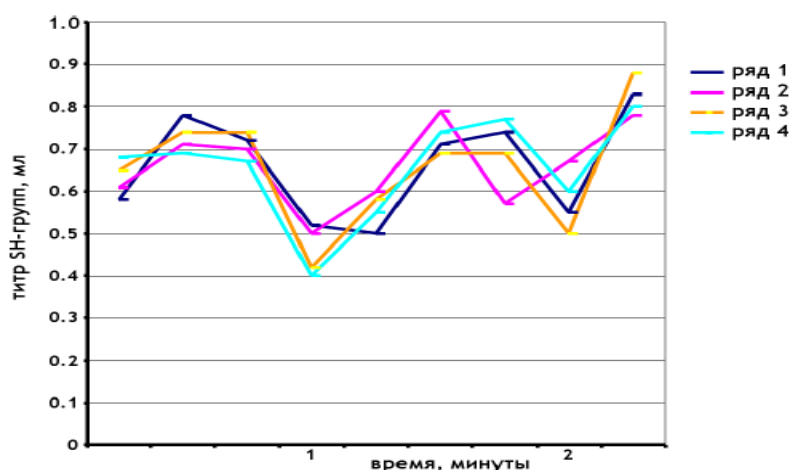


Рис 2. Сходства формы гистограмм в опытах с АК + ДХФИФ трипсином и креатинкиназой. По оси ординат число проб с данной активностью [34].

3 Форма гистограмм закономерно изменяется во времени и эти изменения определяются космофизи-

ческими факторами.

На рисунке 3 представлена иллюстрация синхронного изменения титра SH - групп в разных порциях раствора актина в течении нескольких минут.

Детальный анализ амплитуды флуктуаций реакций в зависимости от времени суток «разброса результатов» измерений скорости реакции АК+ДХФИФ среднеквадратического отклонения  $\sigma$  после нормировки (коэффициент вариации). Каждая точка величины среднеквадратического отклонения амплитуды на графике была вычислена по 1200-1500 измерениям.

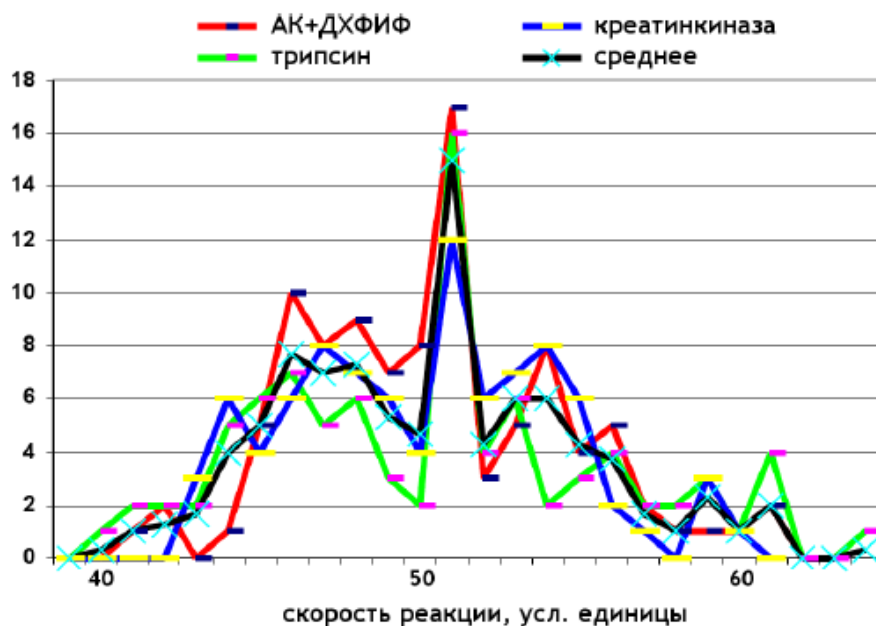


Рис. 3. Изменение титра SH - групп в растворе актина. По оси абсцисс – время. По оси ординат – титр в мл  $1,25 \cdot 10^{-4}$  раствора  $\text{AgNO}_3$  [34].

Сравнение зависимости от времени суток амплитуды «разброса результатов» измерений скорости реакции АК+ДХФИФ ( $\sigma$ ) и проявлений 11 видов заболеваний человека (рис. 4 данные Л.Я. Глыбина).

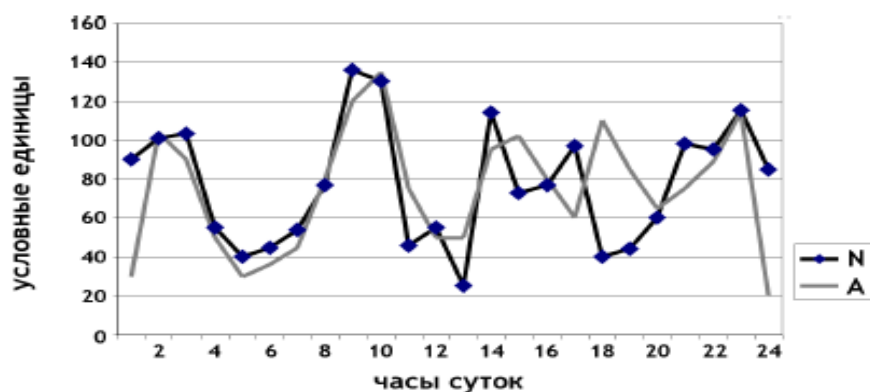


Рис. 4. N – вероятность появления заболевания, A – амплитуда «макроскопических флуктуаций» [34].

Обработка и представление результатов измерений с 1986 по 1992 г.г. альфа активности препарата  $^{239}\text{Pu}$ , включающих 60000 измерений, в виде гистограмм (рис. 5) показала наибольшую вероятность сходства ближайших соседей «эффект ближней зоны» и четкую околосуточную периодичность появления сходных гистограмм.

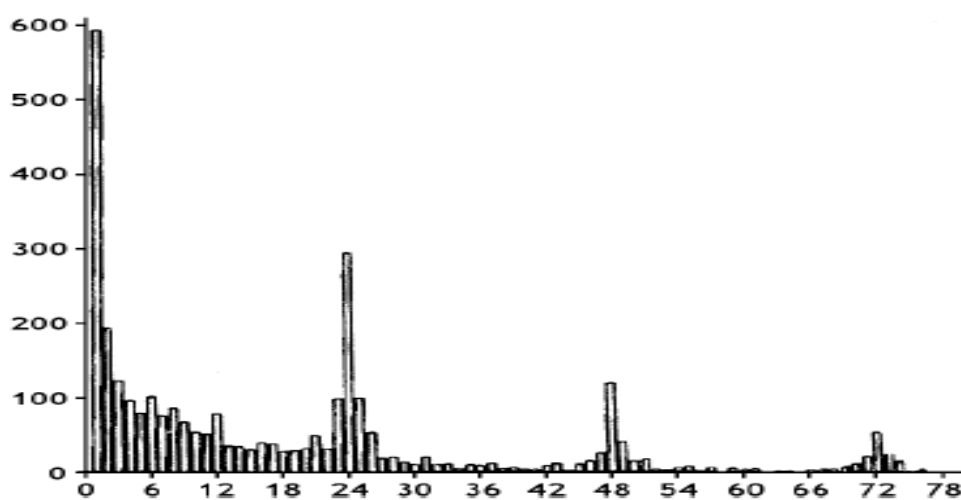


Рис 5. Зависимость вероятности повторного появления гистограмм, построенных по результатам измерений альфа-активности  $^{239}\text{Pu}$ , от величины разделяющего их интервала времени. По абсциссе – интервалы времени (часы). По ординате – число сходных пар, соответствующее данной величине интервала [34].

На рис. 6 видна корреляция изменений содержания гемоглобина крови и отклонения движения Луны по кеплеровской орбите -  $\lambda$ . Аналогичные картины получены для скорости реакции окисления унитиола (метод В.В. Соколовского), скорость оседания эритроцитов, флуктуаций «компьютерного времени» и ряда других показателей.

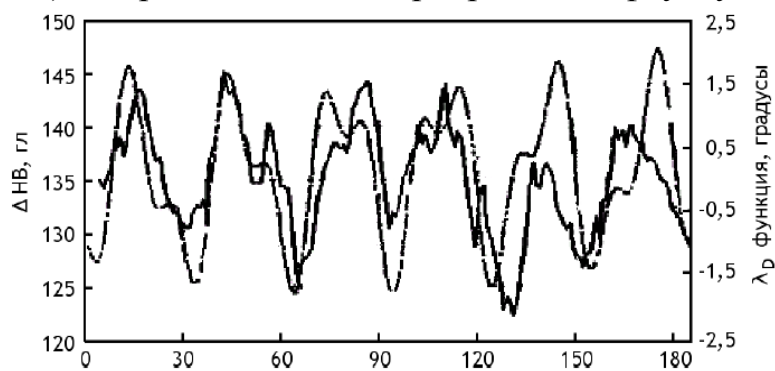


Рис 6. Изменение содержания гемоглобина в крови (автора исследования) при ежедневном измерении с 1 июня по 31 октября 2002 г. в условиях антарктической



экспедиции в сопоставлении с изменениями  $\lambda$  [34].

В центре рисунка 7 Солнце, вращающееся против часовой стрелки с около 27 – суточным периодом. По околосуточной орбите, так же против часовой стрелки, совершает годичный путь Земля. Сторона Земли, обращенная к Солнцу – день. Противоположная сторона (заштрихованная) – ночь. Земля, также против часовой стрелки, вращается вокруг своей оси. Поперек рисунка изображена линия равноденствий. Сверху вниз косыми линиями изображено (условное) направление движения Солнечной системы в Галактике.

Отклонение истинного движения Луны от движения по законам Иоганна Кеплера, вызываемое воздействием Солнца называется эвекцией. Смысл эвекции состоит в периодическом изменении формы лунной орбиты, т.е. в возрастании и уменьшении эксцентриситета орбиты Луны. Данный эффект обусловлен гравитационным воздействием Солнца на Луну. Из совокупности полученных результатов автором сделан вывод - разброс значений измеряемых ве-

личин является следствием флуктуаций пространства-времени.

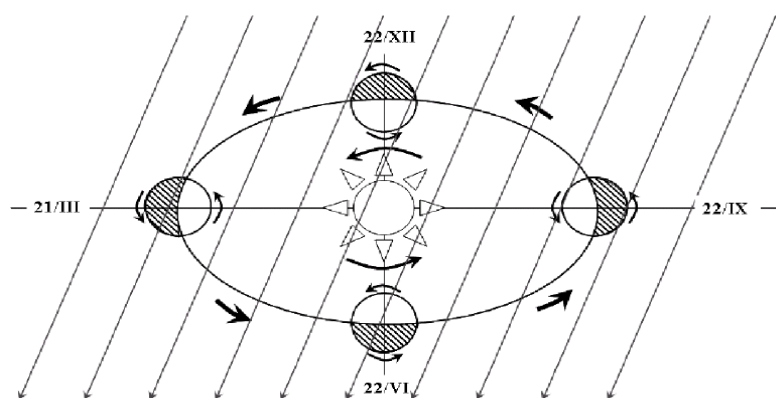


Рис 7. Вращение земли вокруг своей оси и Солнца относительно движения Солнечной системы в Галактике

Таким образом, причиной низкой воспроизводимости результатов последовательных измерений является исследованный С.Э. Шнолем эффект «макроскопических флуктуаций», который выражается в зависимости периодичности изменений исследуемых явлений: суточных, месячных и годовых. Следовательно, альтернативой усреднения результатов последовательных измерений является применение методов математического моделирования процессов имеющих колебательный характер.

Во многих учебниках колебательный характер исследуемых изменений приводят как подтверждение гипотез, положенных в основу модели Вольтерра [31]. На рис. 8. представлены кривые колебаний численности пушных зверей по данным компании Гудзонова залива о числе заготовленных шкурок. Действительно, периоды колебаний численности зайцев (жертв) и рысей (хищников) примерно одинаковы и составляют около 9 – 10 лет. При этом максимум численности зайцев опережает, как правило, максимум численности рысей на один год. Наблюдаются регулярные колебания, осложненные случайными факторами, связанными с погодой и проч.

Модель взаимодействия видов предложена В. Вольтерра в тридцатые годы XX века для объяснения периодических изменений числа особей, так называемую модель «хищник-жертва». Обозначив число зайцев (жертв)  $x$ , а число вол-

ков (хищников)  $y$ . Зайцы размножаются со скоростью, пропорциональной их числу  $\varepsilon_x x$ . Убыль зайцев пропорциональна вероятности встречи с волком и наоборот число волков нарастает тем быстрее, чем чаще происходят их встречи с зайцами, т.е. пропорциональна произведению численностей  $x y$ . Кроме того, имеет место процесс естественной смертности волков, причем скорость

смертности пропорциональна их количеству.

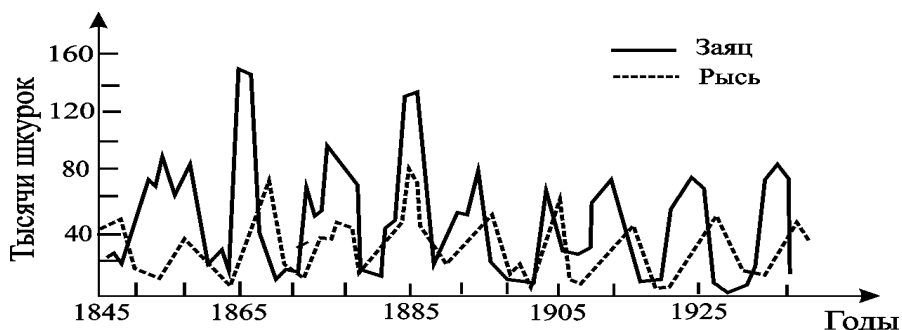


Рис. 8. Кривые численности зайца и рыси в Канаде.

Эти рассуждения приводят к системе уравнений для изменений численности зайцев и волков:

$$\begin{aligned} \frac{dx}{dt} &= x(\varepsilon_x - \gamma_{xy} y), \\ \frac{dy}{dt} &= -y(\varepsilon_y - \gamma_{yx} x). \end{aligned} \quad (1)$$

Где:  $\varepsilon_x, \varepsilon_y, \gamma_{xy}, \gamma_{yx}$  - параметры системы.

Анализ системы (1) показывает, что траектории вблизи особой точки являются концентрическими эллипсами, а сама особая точка – центром (рис 9).

Рассматриваемая модель Вольтерра имеет замкнутые траектории и определяется параметрами системы. Изменения численности жертвы и хищника во времени представляют собой колебания, причем колебания численности хищника отстают по фазе от колебаний жертв. В момент, когда точка находится в положении  $M_1$ , в систему добавляется извне некоторое число особей  $y$  такое, что изображающая точка переходит скачком из точки  $M_1$  в точку  $M_2$ . Если после этого систему предоставить самой себе, колебания  $x(t), y(t)$  уже будут происходить с большими амплитудами, чем прежде, и изображающая точка будет двигаться по траектории 2. Это и означает, что колебания в системе неустойчивы: они навсегда изменяют свои характеристики при внешнем воздействии.

Система имеет особую точку – устойчивый фокус, узел рис. 10.

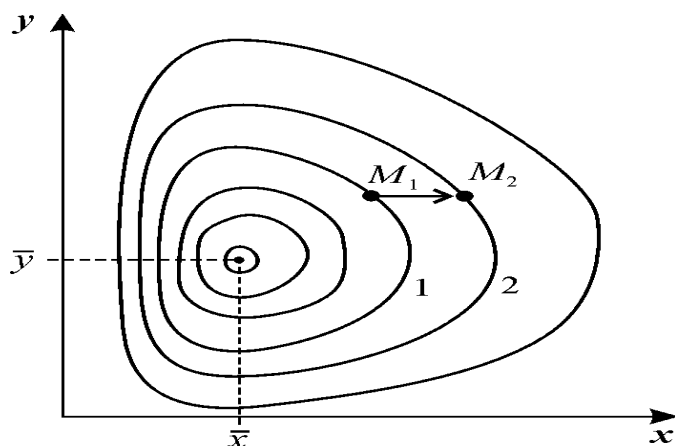


Рис. 9. Фазовый портрет системы (рис. 8). Параметры системы:  $\varepsilon_x = 4$ ,  $\gamma_{xy} = 0,3$ ,  $\varepsilon_y = \gamma_{yx} = 0,4$

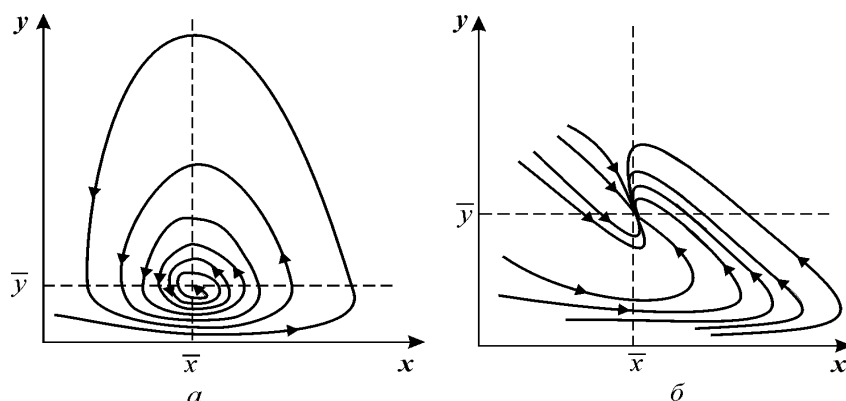


Рис. 10. Фазовый портрет системы. Где: *a* – устойчивый фокус, параметры системы:  $\epsilon x = 2$ ,  $\gamma xy = 18$ ,  $\delta x = 1$ ,  $\epsilon y = 3$ ,  $\gamma ux = 5$ ,  $\delta y = 1$ , *б* – устойчивый узел,  $\epsilon x = 2$ ,  $\gamma xy = 1$ ,  $\delta x = 1$ ,  $\epsilon y = 3$ ,  $\gamma ux = 1$ ,  $\delta y = 1$

Таким образом, применение математического моделирования дает возможность достаточно точно определить текущее значение и существенно снизить многократный разброс результатов полученных в разные временные периоды (от 20 до 150 тысяч шкурок зайцев полученных в год).

В месте с тем остается открытый вопрос: «Что делать в случае, когда нет математической модели»?

Традиционный подход основан на сравнении результатов полученных параллельно в контроле и опыте. Разница данных величин определяет абсолютную погрешность. Величина относительной погрешности дает возможность количественно сравнивать величины, имеющие разные размерности. Относительные величины являются основой проведения корреляционного, регрессионного, дисперсионного и других видов статистического анализа. Примеры применения критериев Стьюдента - *t* и Фишера - *F*.

Следовательно, для сравнения результатов полученных последовательно (при разных условиях) рекомендуется применять относительные показатели.

Рабочая гипотеза решения проблемы не сопоставимости последовательно полученных результатов измерений заключается в применении относительных показателей полученных в контроле и опыте. Отношение параллельно полученных величин в контроле и опыте дает возможность сравнивать результаты, полученные в последовательно проведенных измерениях.

## 6.2 Уникальные возможности Data Mining

Сам термин Data Mining переводится как "извлечение информации" или "добыча данных". Нередко рядом с Data Mining встречаются слова knowledge discovery - "обнаружение знаний" и Data Warehouse - "хранилище данных". Возникновение указанных терминов, которые являются неотъемлемой частью Data Mining, связано с новым витком в развитии средств и методов обработки и хранения данных. Она оперирует усредненными характеристиками выборки, которые часто являются фиктивными величинами (средней жизнеспособностью биообъекта, когда в зависимости от выбора технологии или качества биообъекта, нам необходимо уметь прогнозировать состояние деконсервированного объекта;

средней эффективности технологии, тогда как нам интересны характерные особенности отдельных её этапов и т.д.).

Цель Data Mining состоит в выявлении скрытых закономерностей в полученных данных, образно говоря, в "нахождении золотых песчинок в огромной куче руды". Дело в том, что человеческий разум сам по себе не приспособлен для восприятия больших массивов разнородной информации. В среднем человек, за исключением некоторых индивидуумов, не способен улавливать более двух-трех взаимосвязей даже в небольших выборках. Но и традиционная статистика, долгое время претендовавшая на роль основного инструмента анализа данных, также нередко пасует при решении задач из реальной жизни. Поэтому методы математической статистики оказываются полезными главным образом для проверки заранее сформулированных гипотез, тогда как определение гипотезы иногда бывает достаточно сложной и трудоемкой задачей. Современные технологии Data Mining перелопачивают информацию с целью автоматического поиска шаблонов (паттернов), характерных для каких-либо фрагментов неоднородных многомерных данных. В отличие от оперативной аналитической обработки данных (OLAP) в Data Mining бремя формулировки гипотез и выявления необычных (unexpected) шаблонов переложено с человека на компьютер. Data Mining – это не один, а совокупность большого числа различных методов обнаружения знаний. Выбор метода часто зависит от типа имеющихся данных и от того, какую информацию вы пытаетесь получить. Сфера применения Data Mining ничем не ограничена он нужен везде, где имеются какие-либо данные. Опыт многих таких предприятий показывает, что отдача от использования Data Mining может достигать 1000 %. Например, известны сообщения об экономическом эффекте, в 10 ÷ 70 раз превысившем первоначальные затраты от 350 до 750 тыс. дол. Приводятся сведения о проекте в 20 млн. дол., который окупился всего за 4 месяца. Другой пример – годовая экономия 700 тыс. дол. за счет внедрения Data Mining в сети универсамов в Великобритании. Data Mining представляют большую ценность для руководителей и аналитиков в их повседневной деятельности. Деловые люди осознали, что с помощью методов Data Mining они могут получить ощутимые преимущества в конкурентной борьбе.

Компанией StatSoft была разработана система *STATISTICA Data Miner*, которая спроектирована и реализована как универсальное и всестороннее средство анализа данных – от взаимодействия с различными базами данных до создания готовых отчетов, реализующее так называемый графически – ориентированный подход.

Data Mining включает огромный набор различных аналитических процедур, что делает его недоступным для обычных пользователей, которые слабо разбираются в методах анализа данных. (Дюк В., "Data Mining – состояние, проблемы, новые решения", [http://on.wplus.net/sparm/science/Data\\_mining.html](http://on.wplus.net/sparm/science/Data_mining.html), 1999)

К сожалению научных работ, в области биологии, с применением программ Data Mining, пока еще нет по следующим причинам: инерционность мышления исследователей, множество взаимосвязанных параметров и высокий коэффициент их вариации, отсутствие баз данных.

### 6.1.1 Сценарный подход в исследовании

Для повышения эффективности научного исследования следует применять сценарный подход, обеспечивающий требуемую надежность результатов анализируемой информации (Data Mining). Под сценарностью понимается применение ряда методов проведения параллельных измерений и их обработки (рис 1.25). При правильной организации эксперимента параллельность измерения и обработки результатов является основой объективности полученной информации вследствие её дополнения. Достоверность результатов научного исследования пропорциональна количеству методов, как измерения, так и обработки (структурирования) полученных данных. Основными этапами структурирования информации являются анализ, оптимизация и графическое представление поверхности отклика.

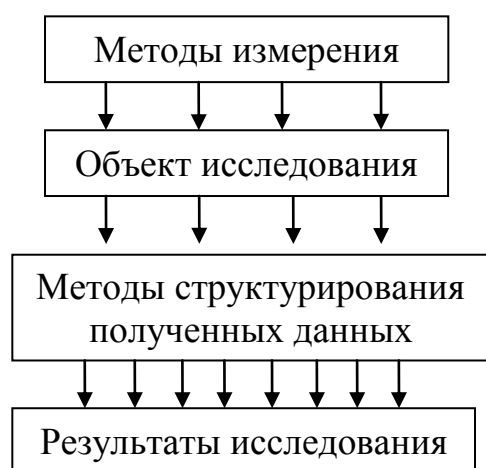


Рис. 1.25. Методы исследования.

Проведенный анализ литературы показывает, что на результаты решения криобиологической задачи оказывает влияние множество факторов. Экспериментальное изучение механизмов протекания всех этих явлений крайне сложны в реализации. Между тем задачи анализа и оптимизации этих процессов необходимо решать. Для этих целей успешно применяются экспериментально статистические методы, с помощью которых строится математическая модель объекта и изучается зависимость отклика системы на изменения варьируемых параметров.

Отсутствие обратной связи при проведении исследования, также является важным препятствием использования существующих способов многофакторной оптимизации, что выражается в возможности получения графических представлений отражающих взаимосвязь интересующих нас параметров, как при проведении эксперимента, так и при анализе полученных результатов. Тем самым теряется возможность отслеживания механизма причинно-следственных связей, являющегося основой построения аналитической модели. Вместе с тем, существенным недостатком общепринятых методов оптимизации исследования является их специфичность, то есть пригодность для определенных областей исследования – решение химических и физических задач. Для исследования биологических объектов существует необходимость создания способа оптимизации, совмещающего достоинства существующих методов и снижающего их недостатки, что позволит ускорить процесс решения криобиологических задач связанных с большим расходом биообъекта и возникающими с этим проблемами как этического, так и экономического характера.

Следующей проблемой является ограниченность совместного применения методов многофакторного анализа и оптимизации в комплексе. Что приводит к необходимости раздельного проведения нескольких повторяющихся опытов.

Хотя, методы оптимизации, основанные на регрессионном анализе, позволяют проводить проверку значимости параметров, однако при большом коэффициенте вариации оптимизируемых параметров они оказываются практически неприемлемыми для этих целей. Такие же трудности возникают при необходимости установления не только значимых факторов, но и их всевозможных комбинаций. Поэтому для определения факторов оказывающих существенное влияние на целевую функцию рекомендуется предварительно использовать дисперсионный анализ с последующим проведением оптимизации выявленных параметров [3, 4, 23, 24, 44 ÷ 49, 61, 81, 82, 85, 87, 90, 96, 112].

Вместе с тем существующие методы многофакторного исследования, как правило, оказываются не применимыми на этапе проведения поисковых опытов. Это зачастую приводит исследователя к отрицательным результатам после выполнения комплекса работ связанных с поэтапным осуществлением многофакторного анализа и оптимизации. В связи с этим существует необходимость разработки алгоритма объединяющего методы многофакторного анализа и оптимизации, с графическим представлением поверхности отклика, работающего на этапе проведения поисковых опытов.

Таким образом, на вероятность получения полноценного потомства от деконсервированных черенков плодово-ягодных культур, спермиев и эмбрионов животных оказывает влияние их начальное состояние и эффективность применяемого способа криоконсервирования. Следовательно, воспроизводимость результатов сохранности биообъекта зависит от вариации его исходного состояния и эффективности выбранного способа криоконсервирования и его составляющих этапов. Для учета взаимосвязи изучаемых параметров криоконсервирования биообъекта рекомендуется комплексно применять градиентные методы оптимизации и дисперсионного анализа на основе реализации относительных величин изменения сохранности биообъекта.

Например, для исследования гетерогенности биологического объекта общая структура исследования разделена на три части – изучение начального состояния, анализ разных методов криоконсервирования и приведение полученных результатов к одинаковым условиям измерения (рис 2). Для изучения неоднородности используют биообъект, имеющий разное начальное состояние. Во второй части, определяют эффективность различных способов криоконсервирования биообъекта замороженного при низких, высоких и сверхвысоких скоростях. Для этого устанавливают этапы, имеющие низкую эффективность, и проводят оптимизацию их параметров по результатам развития биообъекта в условиях *in vitro*.

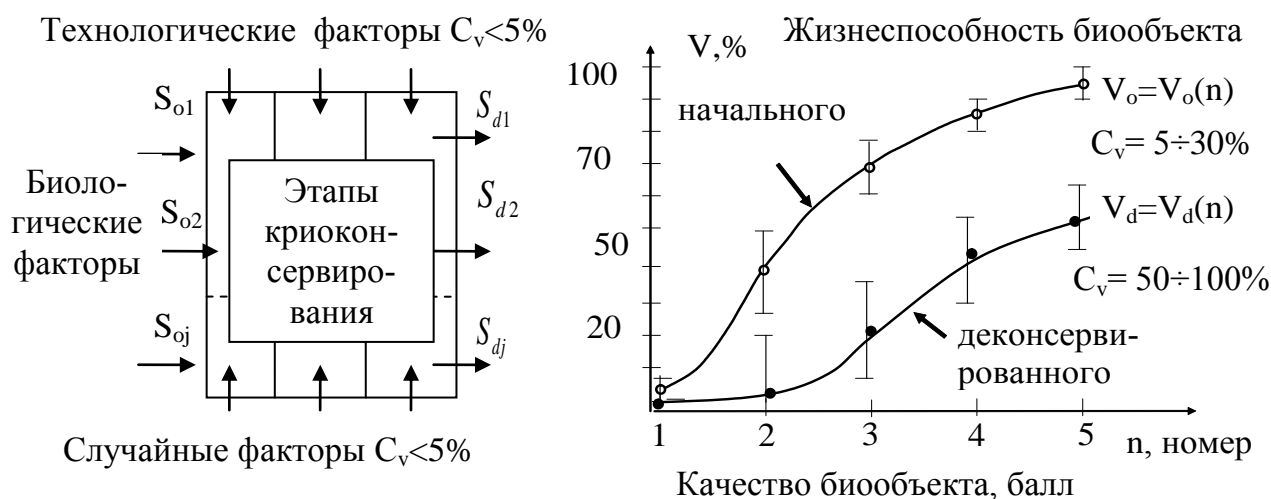


Рис 2. Схема исследования гетерогенности биообъекта на различных этапах его криоконсервирования.  $C_v$  - коэффициент вариации.

В третьей части создается математическая модель, описывающая состояние деконсервированного биообъекта в зависимости от его нативного состояния и эффективности криоконсервирования. Повышение воспроизводимости результатов обеспечивалось за счет устранения межгрупповой вариации обусловленной разным начальным состоянием биообъекта и эффективностью сопряженных этапов криоконсервирования.

## 6.2 Повышение воспроизводимости результатов биотехнологического исследования

Проблема проведения многофакторного биотехнологического исследования связана с низкой воспроизводимостью результатов эксперимента и, как следствие этому, несопоставимостью данных, полученных в отдельных опытах. Нами предложен алгоритм решения, включающий математическую модель и векторный способ проведения многофакторного исследования. Показано, что применение модели дает возможность многократно повысить воспроизводимость результатов за счет учета индивидуальных свойств биообъекта. Экспериментальная проверка модели на разных типах половых клеток растительного и животного происхождения показала возможность многократного сокращения количества биообъекта, взятого для эксперимента с целью получения достоверного результата. Векторный способ предоставляет возможность обеспечить проведение многофакторного анализа и его оптимизацию с графическим описанием полученных зависимостей.

Согласно векторному методу, решение задачи по оптимизации многофакторного исследования можно проводить аналитическим или графическим способом. Данный способ работает на стадии проведения поисковых опытов и практически не имеет ограничивающих условий, в отличие от существующих аналогов. Кроме того, для анализа причинно-следственных связей существует

возможность описания поверхности отклика. Основной особенностью предложенного метода является его применимость для биологических исследований.

Практическую применимость векторного метода рассмотрим на примере проведения исследований по криоконсервированию образцов животного и растительного происхождения.

Выраженная вариация комплекса взаимосвязанных факторов, влияющих на сохранность деконсервированного объекта, является причиной низкого уровня воспроизводимости результатов эксперимента. Поэтому оптимизация исследуемых параметров криоконсервирования биообъекта обычно проводится на различных этапах, а анализ осуществляют по результатам, полученным при одинаковых условиях проведения эксперимента.

Параметры оценки воспроизводимости жизнеспособности биообъекта: - средняя величина жизнеспособности  $V_{cp} = \sum V/n$  для определения общей дисперсии  $\sigma_y^2$ ;

- средняя разность жизнеспособностей в опыте и контроле  $\Delta V = \sum (V_o - V_k)/n$  для учета дисперсии биологических и биотехнологических факторов  $\sigma_i^2$ ;

- среднее отношение жизнеспособностей в опыте и контроле  $W = \sum (V_o/V_k)$  для оценки дисперсии технологических факторов  $\sigma_i^2$ .

Для определения силы влияния исследуемого параметра (рис. 23) применялся показатель  $\eta_i$ :

$$\eta_i = \sigma_i^2 / \sigma_y^2, \quad \sigma_y^2 = \sum \sigma_i^2$$

где:  $\sigma_i^2$  - дисперсии биологического, технологического или их совместного действия (биотехнологические факторы);  $\sigma_y^2$  - общая дисперсия.

Количество эмбрионов, необходимое для получения достоверного результата при комплексной оптимизации всей технологической цепи, можно выразить в виде произведения числа параметров, шагов их варьирования и повторов опытов (объем выборки), необходимых для получения достоверного результата в каждом эксперименте. В данном случае общий объем выборки составит от нескольких сотен до десятков тысяч эмбрионов. При учете эффекта взаимодействия факторов, число шагов варьирования исследуемого параметра будет выражаться в виде показателя степенной зависимости произведения числа параметров и количества биообъекта.

Таким образом, общепринятым критерием тестирования заданного этапа криоконсервирования биообъекта является его сохранность, полученная в результате применения данной процедуры. Однако, высокие значения вариации показателя сохранности исследуемого биообъекта и погрешности его оценки, предполагают необходимость проведения большого количества измерений для оценки эффекта взаимодействия комплекса изучаемых криобиологических параметров.



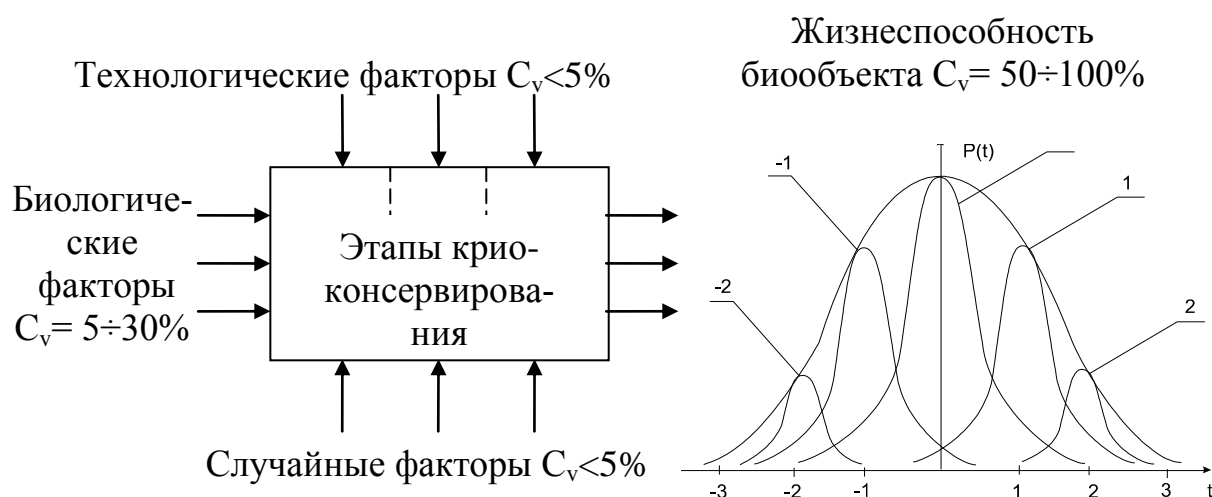


Рис 23. Схема проведения исследования воспроизводимости результатов криоконсервирования образцов животного и растительного происхождения.

$C_v$  - коэффициент вариации всех факторов,  $\sigma_i^2$  - дисперсия факторов:  $i=1$  не учитываемые случайные биологические ( $t=-2$ ) и  $i=2$  технологические ( $t=2$ );  $i=3$  индивидуальные биологические свойства ( $t=-1$ );  $i=4$  биологические и технологические в комплексе ( $t=0$ );  $i=5$  технологические ( $t=1$ ).

Очевидно, что данная задача, является практически не решаемой во всем её объеме. Однако решение этой проблемы возможно по частям при наличии возможности сопоставления полученных результатов. В связи с этим возникает необходимость разработки способа оценки эффективности процесса криоконсервирования биообъекта, его этапов и составных элементов посредством использования относительных показателей, позволяющих проводить сравнительный анализ результатов исследования, полученных при различных условиях выполнения эксперимента.

Зависимости сохранности деконсервированного биообъекта -  $S$  от его начального состояния -  $S_1$  и эффективности этапов криоконсервирования -  $W_i$  представлены на рис. 26, 27. Основной причиной низкой воспроизводимости результатов опытов  $C_v = 0 \div 300\%$  является отсутствие учета начального состояния биологического объекта  $S_1 = 50 \div 100\%$  и эффективности этапов криоконсервирования  $W_2$ ,  $W_3$  и  $W_4$  (Рис. 26).

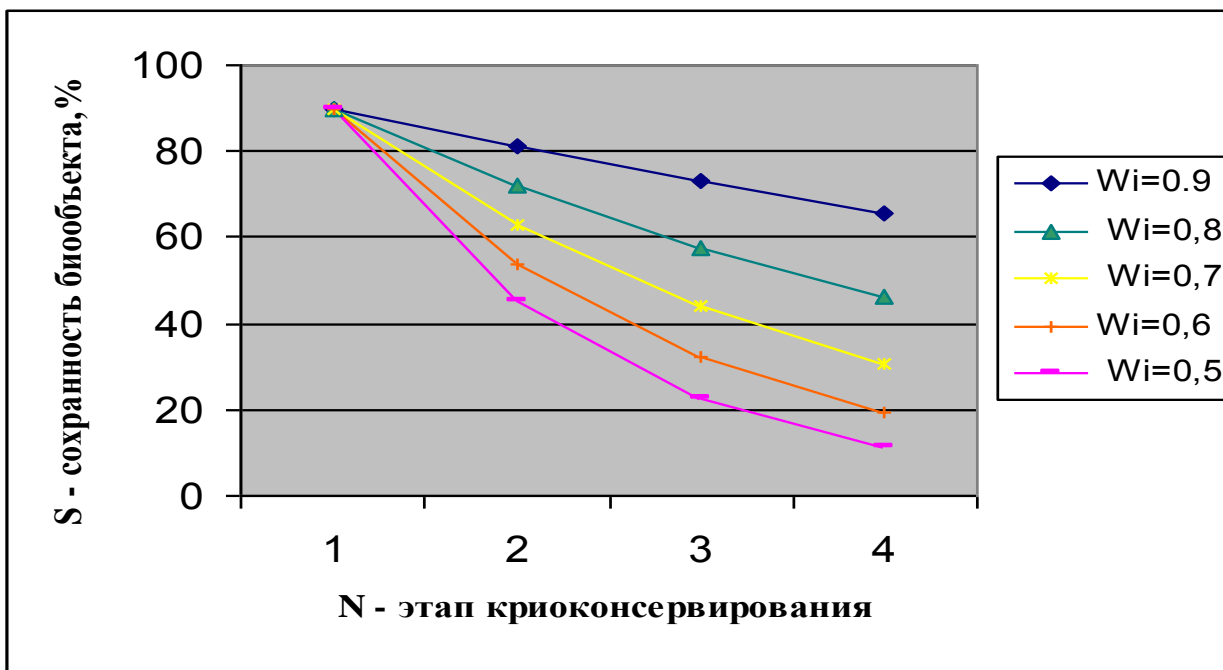


Рис. 26 Сохранность биообъекта в зависимости от эффективности этапов криоконсервирования  $W_i$  {0,5; 0,6; 0,7; 0,8; 0,9}. Начальная сохранность  $S_1=90\%$ .

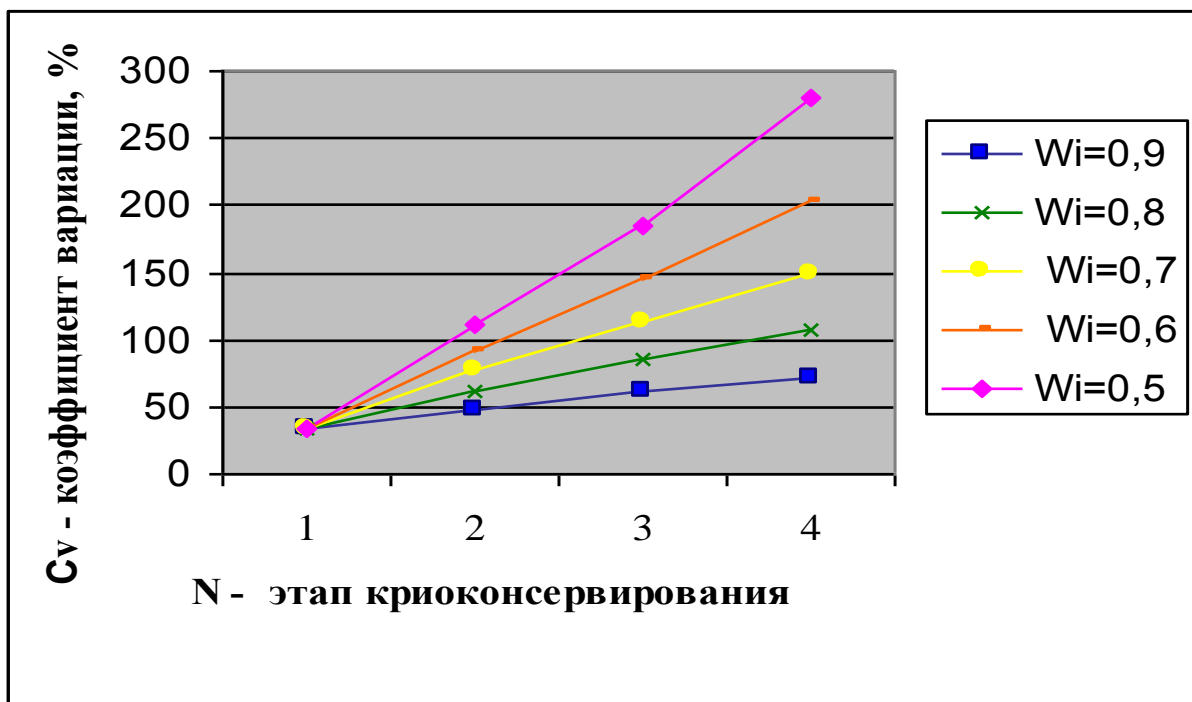


Рис. 27. Воспроизводимость результатов сохранности биообъекта в зависимости от эффективности этапов криоконсервирования  $W_i$  {0,5; 0,6; 0,7; 0,8; 0,9}. Начальная сохранность  $S_1=90\%$ .

Зависимость воспроизводимости результатов криобиологического эксперимента от эффективности этапов криоконсервирования:

$$G = \sqrt{S(1-S)}, \quad C_v = \frac{G}{S} \cdot 100\%, \quad C_v = \sqrt{\frac{1}{S} - 1} \cdot 100\%$$

где:  $G$  – среднеквадратическое отклонение,  $C_v$  – коэффициент вариации (%).

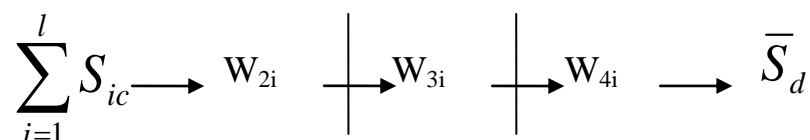
Общеизвестно, что сравнивать можно только однородные величины, т. е. имеющие одинаковую размерность. Для оценки вариабельности абсолютной величины используется показатель среднеквадратического отклонения, а количественное сопоставление однородных показателей производится при помощи отношения дисперсий данных величин (дисперсионный анализ). Сравнение вариабельности разнородных величин (имеющих различные размерности) осуществляется при помощи перевода их в относительные показатели (коэффициент вариации – отношение среднеквадратического отклонения к средней величине). Поэтому для сопоставления разнородных показателей используют прием перевода их в относительную систему координат. Для описания корреляционной связи стохастических параметров применяют критерий нормированного отклонения ( $t$  - критерий Стьюдента - отношение разности средней величины и текущей к среднеквадратическому отклонению), а при оптимизации детерминистических параметров используют относительные величины.

Для сопоставимости результатов, полученных при разных условиях их реализации, используется также показатель эффективности. Термин "эффективность" означает определенный результат известных усилий. Оценка этого результата проводится на основании количественных показателей, отражающих степень достижения намеченной цели. При этом и усилия, и результат должны быть измерены в сопоставимых единицах, а их величины подвергнуты сравнению. Эффективность выражают в процентном отношении значений полученного результата к общим усилиям. Частным случаем данного показателя является коэффициент полезного действия - отношение полезно использованной энергии к затраченной. Экономическая эффективность вычисляется как отношение прибыли к затратам.

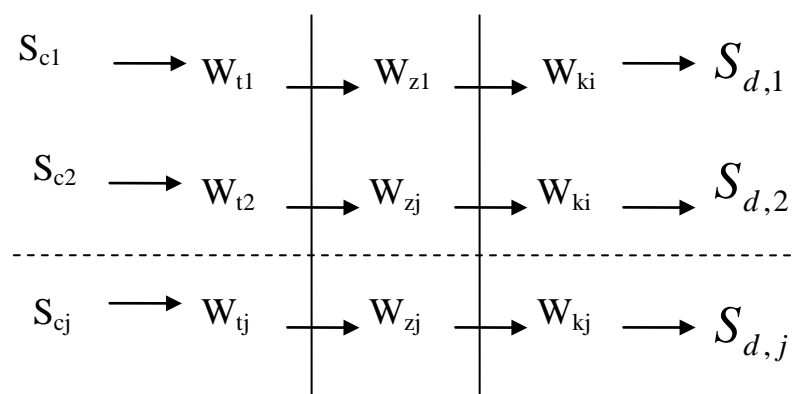
Неучёт начального состояния биологического материала и эффективности этапов криоконсервирования, приводит к накоплению ошибок опыта. Учет индивидуальных свойств биообъекта и особенностей этапов криоконсервирования дает возможность использовать прямые методы измерения, существенно снижающие погрешность эксперимента.

Снижение ошибок опытов посредством перехода от косвенных измерений к прямым:

Косвенные измерения  $\Delta \approx \sum_i^l \Delta_i$



Прямые измерения  $\Delta \approx \Delta_i$



Проблема: низкая воспроизводимость результатов криобиологического исследования  $C_v = 60 \div 300\%$ .

Следствия: отсутствие возможности сравнивать результаты, полученные при разных условиях проведения эксперимента; высокий расход биологического материала необходимого для обеспечения достоверности полученного значения; ограничения возможностей проведения многофакторного исследования.

Причины низкой воспроизводимости криобиологического исследования - неучет систематических погрешностей:

1. Начальное состояние биообъекта  $S_1 = 50 \div 100\%$ ,  $C_{v1} = 0 \div 100\%$ .
2. Эффективности этапов криоконсервирования  $W_i$ ,  $C_{vi} \leq 5\%$ .
3. Применение общих методов статистического анализа при постановке конкретной задачи на исследование.

Задача: снизить количество используемого биообъекта посредством повышения воспроизводимости результатов криобиологического исследования на основе разработанной математической модели.

### **6.2.1. Алгоритм многофакторного анализа и оптимизации исследуемого способа криоконсервирования биообъекта**

Общепринятая оценка выбранного способа криоконсервирования производится по конечному результату - сохранности объекта, который вычисляется по изменению количества объекта пригодного качества «до» и «после» замораживания. Очевидно, что на оценку сохранности деконсервированного материала оказывает влияние начальное состояние биообъекта, особенности выбранных способов применения криопротектора (для черенков растений - сушка), режим замораживания-оттаивания, способ культивирования.

Оценка эффективности выбранного способа криоконсервирования биообъекта в целом и его этапов, затруднена влиянием комплекса неучитываемых факторов. Для решения этой проблемы нами предложен алгоритм проведения исследования, объединяющий методы многофакторного анализа и оптимизации. Алгоритм построен на совместном совершенствовании выбранного способа замораживания и его модели. Модель тестируемого способа криоконсервирования биообъекта строится по результатам оценки изменения его жизнеспособности на каждом этапе проведения исследования.

Абсолютные величины и их разность дают возможность установить влияние начального состояния биообъекта на его жизнеспособность, тогда как переход к относительным величинам способствует устранению этого влияния при оценке эффективности реализации отдельных технологических операций (прил. 1). Полученные значения представляют в виде диаграммы («профиль» выбранного способа криоконсервирования биообъекта для удобства анализа слабых звеньев исследуемой технологической цепи (прил. 2)). Затем осуществляется оптимизация модели цикла криоконсервирования биообъекта предложенным векторным способом ф.(8.3÷8.5), (прил. 1, 3), (рис. 29, 30) и её проверка в ходе выполнения последующего эксперимента. Совершенствование модели и способа замораживания происходит поэтапно: от оценки показателей эффективности этапов выбранного способа криоконсервирования биообъекта в процессе проведения поисковых опытов, до нахождения оптимальных значений совместно исследуемых параметров и последующим их графическим описанием.

Использование методов математического моделирования дает возможность проведения многофакторного исследования, в несколько раз повышающего его эффективность [1, 2, 11, 16, 30, 31]. При этом возможность решения поставленной задачи и условия её реализации определяются в ходе проведения поисковых опытов.

Предложенный векторный способ оптимизации многофакторного исследования может выполняться как в графическом, так и в аналитическом виде. В основе векторного решения лежат два показателя: коэффициент крутизны, определяющий приоритет поиска необходимого параметра, и векторная величина  $k$ , отвечающая за направленность поиска ф. (8.2÷8.5). Графическое решение используется в случае отсутствия вычислительных средств и характеризуется удобством, простотой применения. Аналитическое решение возможно при наличии специального программного обеспечения, основанного на диалоговом режиме

работы (прил. 3).



Рис. 28. Алгоритм проведения многофакторного исследования, повышающий воспроизводимость результатов криоконсервирования спермиев и эмбрионов животных, черенков плодово-ягодных культур.

Данный способ работает на стадии выполнения поисковых опытов и практически не имеет ограничивающих условий при проведении оптимизации, в отличие от существующих аналогов. Кроме того, для изучения причинно-следственных связей существует возможность графического представления зависимости относительной сохранности деконсервированного биообъекта от оптимизируемых параметров. Основной особенностью данного способа является его универсальность в применении на любой из стадий проведения исследования и, что особенно важно, – возможность выполнения дисперсионного анализа при оптимизации биологических параметров.

Сравнительный анализ показывает, что при увеличении количества совместно исследуемых факторов до пяти, потери времени и денежных средств многократно сокращаются в сравнении с традиционно применяемым однофакторным методом (прил. 2). Поиск оптимального значения и построение поверхности отклика производится в системе абсолютных величин, а обобщение параметров, имеющих разную размерность, происходит благодаря вводу относительного показателя - коэффициента «крутизны»  $\phi$ . (8.2), определяющего шаг и направление изменения каждого параметра. Обобщенная векторная величина задаёт координаты последующего измерения. Отличительная особенность предложенного способа от существующих аналогов заключается в объединении их достоинств и устранение ограничений, присущих каждому из них.

Преимущества предложенного способа заключаются в его универсальности,

возможности решения биологических задач; отсутствии необходимости построения регрессионных уравнений; возможности исключать параметры и вводить дополнительные; отсутствии ограничений в изменении шага и направления поиска; применимости в ходе проведения поисковых опытов; возможности использования экспериментальных данных, представленных в дискретном и непрерывном виде; одновременном проведении поиска из разных точек факторного пространства.

Основное отличие предложенного способа от существующих аналогов заключается в его применимости на этапе поисковых опытов при совместном использовании метода многофакторного анализа. При наличии программного обеспечения существует возможность (а точнее необходимость) поддержания диалога исследователя с программой. Наличие обратной связи – значительно повышает эффективность проведения исследования.

### 6.2.3 Определение условия решения поставленной задачи

При всем разнообразии методов исследовательской работы задача планирования эксперимента сводится к тому, чтобы при, возможно, минимальных объемах наблюдений получать достаточно полную информацию об изучаемых объектах. Установив в ходе поискового опыта основные статистические величины изучаемых явлений, такие как среднее арифметическое и среднеквадратическое отклонение, можно приступить к моделированию эксперимента. Моделирование основного эксперимента в ходе проведения поисковых опытов позволяет определить условия решения поставленной задачи и получить при завершении работы достоверный научный результат. Для определения минимального количества измерений в процессе проведения поисковых опытов для одной отдельно взятой выборки  $n_0$  (приложение 4 формула 3) и двух сравниваемых ( $n_\Delta$ ):

$$n_\Delta = (t \cdot C_v / \alpha)^2, \quad \alpha = \frac{\Delta M}{M} \cdot 100\% \quad (8.1),$$

где:  $t$  - критерий Стьюдента;  $C_v$  - коэффициент вариации выборки(ок);  $\alpha$  - относительный показатель расхождения средних величин сравниваемых выборок;  $\Delta M$  – расхождение значений в контроле и в опыте;  $M$  – среднее значение, полученное в контроле.

Вывод формул определения минимального количества измерений, необходимых для получения достоверного результата, представлен [11]. Если различие между выборками составляет больше семи процентов ( $\alpha > 7$ ), то основной эксперимент выгодней проводить на основе сравнения с контролем. В этом случае количество биообъекта, необходимого для получения достоверного резуль-

тата, оказывается значимо меньше ( $n_{\Delta} < n$ ), нежели при сравнении значений по отдельным средним показателям.

Использование количественного метода статистической обработки для оценки жизнеспособности эмбрионов млекопитающих даёт возможность снизить количество измерений более чем в несколько раз по сравнению с качественным, за счет уменьшения величины коэффициента вариации выборок при условии получения достоверного результата. Поэтому там, где это возможно, более предпочтительно использовать количественный статистический анализ.

Количественный учет различного качества эмбрионов дает возможность значимо увеличить точность оценки их жизнеспособности посредством уменьшения дисперсии и тем самым понизить количество биообъекта, необходимого для получения достоверного результата. Для получения достоверных значений средних величин жизнеспособности деконсервированных эмбрионов при использовании количественного метода оценки, необходимо в  $2 \div 5$  раз меньше эмбрионов для каждой из групп, чем для качественного.

При переходе к относительным величинам коэффициент вариации снижается в  $1,3 \div 2,6$  раза, что позволяет уменьшить количество используемого биообъекта в  $1,7 \div 6,8$  раза. Полученные результаты дают подобную закономерность для суспензий клеток. Созданное программное обеспечение апробировано на эмбрионах животных (прил. 2, 3). Применение предложенного метода многофакторного исследования дает возможность сократить объем выборки в несколько раз (прил. 2).

Повысить воспроизводимость криобиологического эксперимента возможно посредством последовательного выполнения следующих функций:

- 1) определение условий решения поставленной задачи – минимальное количество измерений обеспечивающих заданный уровень надежности;
- 2) количественная оценка состояния биообъекта – учет индивидуальных особенностей посредством усреднения разности показателей в контроле и опыте приводит к снижению вариации в  $1,4 \div 2,4$  раза;
- 3) оценка эффективности этапов криоконсервирования посредством усреднения отношения показателей, полученных в контроле и опыте, что снижает вариацию в  $1,5 \div 2$  раза;
- 4) многофакторное исследование – определение взаимосвязи варьируемых параметров, уменьшение количества опытов в  $2 \div 10$  раз;
- 5) сопоставимость полученных результатов при разных условиях проведения эксперимента дает возможность проведения дополнительных опытов.

Таким образом, переход от абсолютных показателей жизнеспособности биообъекта к их разности и отношению дает возможность изучать индивидуальные свойства биообъекта и осуществлять их учет при определении эффективности этапов криоконсервирования. При использовании количественного метода оценки жизнеспособности биообъекта и переходе к относительным величинам повышается точность оценки этапов криоконсервирования в полтора раза по отношению к результатам, полученным по абсолютным качественным показателям - сохранности. Тем самым появляется возможность в несколько раз уменьшить количество биообъекта, необходимого для получения достоверного результата.



### Определение достоверности значений параметров

Уровень надежности прогноза направления поиска оптимальных значений при помощи предложенного векторного способа, определяется минимальной величиной значения надежности средних величин, которые использовались для нахождения градиента роста целевой функции.

Например, при построении векторов использовались значения сохранности деконсервированного биообъекта: для первого  $65 \pm 18$ ,  $70 \pm 17$  % и второго  $65 \pm 18$ ,  $80 \pm 15$  %. Количество биообъекта в каждой выборке составляло 7. Относительная погрешность вычисляется:

$$p = \frac{t \cdot m}{M}, \quad (1)$$

где  $t$  – критерий Стьюдента, для  $n=7$  и заданного уровня надежности  $P \geq 0,95$   $t=2,45$ ;  $m$  – ошибка среднеквадратического отклонения;  $M$  – средняя величина выборки.

Для полученных значений сохранности биообъекта: 65, 70 и 80 %, уровень относительных погрешностей составил 0,66, 0,59, 0,45 соответственно.

Уровень надежности для данных значений вычисляется:

$$P = 1 - p. \quad (2)$$

Для полученных значений сохранности биообъекта: 65, 70 и 80 %, уровень надежности составил 0,34, 0,41, 0,55 соответственно.

Таким образом, уровень надежности прогноза найденного градиента целевой функции составил  $P \geq 0,34$ . Для повышения уровня надежности до величины  $P \geq 0,95$  % необходимо увеличить количество измерений.

Минимальное количество повторов  $n$ , обеспечивающих достоверное среднее значение с уровнем надежности  $P \geq 0,95$ , определяли по [11]:

$$n = \left( \frac{t}{p} C_v \right)^2, \quad (3)$$

где  $p$  – допустимая относительная ошибка  $p=5\%$  для уровня надежности  $P \geq 0,95$ ;

$C_v$  – коэффициент вариации выборки, %.

Для полученных значений сохранности биообъекта: 65, 70 и 80 %, коэффициент вариации составил 73, 70, 65%. Количество биообъекта в каждой пробе

необходимо добавить до 1280, 1188 и 927 соответственно для обеспечения требуемого уровня надежности  $P \geq 0,95$ .

При работе со спермиями получить количество биообъектов в пробе до 1000 и более, возможно, но для эмбрионов и ооцитов животных – нет. Существенное снижение количества измерений возможно при использовании уровня надежности, достаточного для проведения опытов при поиске направления роста целевой функции  $P \geq 0,90$ . Критерий Стьюдента для  $n=7$  и заданного уровня надежности  $P=0,90$  составляет  $t=1,94$ . В этом случае количество биообъекта в каждой пробе из выражения (3) составит: 201, 186 и 146 соответственно.

Применение количественных методов оценки биообъекта снижает приблизительно вдвое коэффициент вариации, а это из выражения (3) в 4 раза сокращает количество биообъекта. Значимое снижение коэффициента вариации возможно при использовании однородного биообъекта (одинакового качества), планирования проведения исследования с использованием особых методов направленных на устранение межгрупповой вариации.

### **6.3 Способы обобщения результатов, полученных в ходе проведения исследования**

Для универсального оперирования полученной информацией необходимо обобщать экспериментально полученные данные до уровня математического абстрагирования. Обобщение результатов эксперимента, полученных в различных опытах, проводят при помощи вычисления величин:  $M$  - среднее арифметическое,  $\sigma$  - среднеквадратическое отклонение,  $C_v$  - коэффициент вариации, которые удобно выражать через моменты  $h$ -го распределения -  $M_h$ , а степень их достоверности через величину ошибки репрезентативности -  $m_h$  (см приложение, табл.1.). Необходимым условием правомерности использования этих параметров является нормальность распределения изучаемых выборок, критериями, определения которых служат: коэффициент асимметрии распределения -  $A_s$  и коэффициент эксцесса -  $E_x$ . Для избежания получения искажения объективной информации также необходимо исключить грубые ошибки, которые возникают при проведении измерений. Численным показателем достоверности текущего значения -  $x_i$ , а также достоверности конкретной выборки -  $t_p$ , является критерий Стьюдента -  $t_{st}(p,n)$ .

В случае, если исследователь наблюдает расхождение двух величин целевых функций  $C$ , полученных при измерении изучаемых факторов  $a_i$ , то достоверность данного расхождения можно подтвердить при помощи критерия Стьюдента -  $t_d$ . Общим принципом чистоты проведения исследования является постоянство всех остальных факторов при изменении избранных  $a_j$ . Тестом на соблюдение выше приведенного условия является критерий равноточности измерений. Равноточность измерений серии опытов позволяет обобщить результаты различных экспериментов и уменьшить суммарные ошибки исследования. Показателем равноточности опытов является критерий Фишера -  $F$ .

Далее экспериментальные данные обобщают в такой последовательности: находят функциональные связи изучаемых явлений, устанавливают основные

закономерности, математически выражают, раскрывают их физическую сущность с целью последующего теоретического развития.

Вывод эмпирических формул проводится на основе экспериментально полученной функциональной зависимости, объясняемой в общем, виде. Подбор типа эмпирической формулы для определения ее параметров используется на основе регрессионного анализа. Построение регрессионной зависимости осуществляется при помощи использования специальных прикладных программ, разработанных на ЭВМ. Полученное эмпирическое выражение можно представить в полуаналитическом или чисто аналитическом виде. Иногда сама структура эмпирической формы указывают на пути превращения в аналитическую формулу. Под аналитическими уравнениями следует понимать такие, в которых все основные члены имеют размерность и логично вытекают из законов физики.

Наличие аналитического выражения зависимости целевой функции от ее аргументов, позволяет использовать его как математическую модель. Очевидно, что исследование считается законченным, если достигнута его цель. Если реализация цели заключена в нахождении максимальной или минимальной величины целевой функции, то построение математической модели, отражающей природу взаимосвязи между целевой функцией  $C$  и ее аргументами  $a_i$ ,  $b_i$  и  $d_i$ , является наиболее приемлемым способом нахождения оптимальных значений изучаемых факторов  $a_i$ ,  $C_{\text{мак}} = F(a_{i_{\text{опт}}}, b_i)$ . Математические модели строятся на основе полученных эмпирических и рациональных формул, а поэтому их удобно разделять на описательные и объяснительные.

Математические модели являются мощным инструментом, как описания так и анализа полученных данных. Ценность модели определяется степенью адекватности описания изучаемых явления и возможностью их объяснения. Математическое абстрагирование полученных результатов способствует разрешению проблемы не повторяемости экспериментальных данных при изменяющихся условиях проведения опытов и служит отправным моментом проведения дальнейших исследований.

Наиболее ярким примером значимости математического моделирования в решении прикладных задач являться работа П. Мейзура [16]. Благодаря данной модели впервые удалось получить жизнеспособное потомство от деконсервированных эмбрионов млекопитающих после неудачно проведенных эмпирических поисков на протяжении трех десятилетий.

Повышение эффективности исследовательской работы связано с моделированием проведения эксперимента от постановки задачи на исследования до математического обобщения полученного результата. Предложенная структурная схема моделирования эксперимента, позволяет получать достоверный научный результат при минимальных затратах времени и средств.

### **6.3.1 Выражение опытных закономерностей в виде формул**

#### ***Вывод эмпирических формул:***

Для вывода эмпирических формул необходимо иметь экспериментально полученную функциональную зависимость, объясняемую в общем виде. Основной метод построения эмпирических формул: подбор типа формулы и нахождение

ние коэффициентов к ней. Подбор типа формулы главное, что определяет результативность и точность эмпирической формулы. Следует учесть также, что кривая изменения функции часто может быть представлена несколькими типами уравнений. В этих случаях надо выбирать такой тип формулы, который в наибольшей мере соответствовал бы природе явлений и мог бы быть улучшен превращением формулы в полуэмпирическую или полурациональную.

Подбор типа эмпирической формулы для определения ее параметров используется на основе регрессионного анализа. Параметры формулы надо вычислять с достаточной степенью точности, особенно, если они относятся к показателям степени. Не надо забывать, что многие математические выражения закономерности, будучи вначале эмпирическими, впоследствии были рационализированы и что иногда сама структура эмпирической формы указывает на пути превращения ее в формулу рациональную.

### ***Получение рациональных формул:***

Под рациональными формулами следует понимать такие, в которых все основные члены имеют размеренность и строго логично вытекают из законов физики. Рациональные формулы не только представляют собой закономерность развития явления, но и объясняют его (хотя бы и в общей форме), показывая факторы и связи между ними. Это важнейший тип формул, так как их всеобщность дает возможность теоретически исследовать все стороны явления.

Для построения рациональной формулы необходимо знать физические основы явления. Не следует увлекаться непродуманными математическими обобщениями. Надо всегда помнить как о необходимости в связи формул с реальными физическими процессами, так и о возможной ограниченности их приложения. Для выведения рациональной формулы вначале следует установить, какой закон или законы физики применимы для объяснения явления. Затем следует решить вопрос о том, какой трансформации или дополнений требуют известные уже выражения законов в связи с особенностями развития явления. Далее решают, как лучше в общей и грубой схеме математически представить найденные зависимости. Наконец, эту схему уточняют, “отшлифовывают” и проверяют на ряде примеров.

Развитие явления обычно может быть представлено дифференциальными уравнениями. При преобразовании первоначальных математических схем в дифференциальные уравнения следует стремиться к простым и ясным выражениям; это предлагает соответствующий выбор системы отсчета (координат), решения вопроса о порядках уравнения и о том, могут ли быть коэффициенты уравнений постоянными величинами.

Не всегда исследователю удастся обогатить науку открытием новых рациональных формул. В ряде случаев бывает полезно представить конкретную экспериментальную кривую, описывая ее чисто математически и преимущественно безразмерными величинами. Такие формулы называются эмпирическими. Они не имеют характера всеобщности, а показывают развитие явления только для условий, в которых проводились опыты. Наличие аналитического выражения зависимости целевой функции от ее аргументов, позволяют использовать его как математическую модель.

Очевидно, что исследование считается законченным, если реализована его цель. Поскольку реализация цели заключена в нахождении максимальной или минимальной величины целевой функции, то построение математической модели, отражающей природу взаимосвязи между целевой функцией  $C$  и ее аргументами  $a_i$ ,  $b_i$  и  $d_i$ , является наиболее приемлемым способом:  $C_{\max} = F(a_{\text{ioпт}}, b_i)$ .

Математические модели строятся на основе полученных эмпирических и рациональных формул, а поэтому их удобно разделять на **описательные** и **объяснительные**. В англоязычной литературе - выделяют модели феноменологические и механистические, критериями состоятельности которых являются информационная ценность и адекватность описания изучаемых явлений [15].

## 6.4 Тестирование полученной модели

### 6.4.1 Установление информационной ценности математической модели

Для проверки **информационной ценности** модели формулируется нуль-гипотеза  $H_0$ :  $\sigma_{об}^2\{C\} = \sigma_{на}^2$ , где  $\sigma_{об}^2\{C\}$  - общее рассеивание результатов измерений по отношению к общему среднему по всему эксперименту.  $\sigma_{об}^2\{C\} = \sigma_{на}^2$  - рассеивание средних экспериментальных значений выходов  $\tilde{C}$  от предсказываемых моделью значений  $\bar{C}$ . Эта величина характеризует неадекватность описания модели эксперимента.

$$\text{Если } F_u = \frac{\sigma_{об}^2}{\sigma_{на}^2} > F_{таб}(\alpha = 0,05) \quad (24),$$

то модель (имеющая  $\sigma_{на}^2$ ) описывает результаты эксперимента лучше, чем простейшая модель. Если  $F_u < F_{таб}$ , то такая модель, несмотря на сложный вид и на ее построение, не имеет информационной ценности.

### 6.4.2 Проверка адекватности описания математической моделью экспериментальных данных

Для проверки **адекватности модели** формируется нуль-гипотеза  $H_0$ :  $\sigma_{на}^2 = \sigma_{э}^2$ . Если она по критерию Фишера (см. приложение, табл. 3)  $F_a$ :

$$F_a = \frac{\sigma_{на}^2}{\sigma_{э}^2} \quad (24a),$$

будет признана правдоподобной ( $F_a < F_{таб}$  обычно при  $\alpha = 0,05$ ), то анализируемая модель адекватно описывает экспериментальные результаты, поскольку  $\sigma_{на}^2 = \sigma_{э}^2$ , и предсказываемые моделью результаты  $\tilde{C}$  будут по точности не хуже экспериментальных.

Существуют и другие способы проверки полученных математических моделей. Выверенные модели позволяют получить объективную информацию с тем же уровнем надежности с которым они протестированы. Однако не следует забывать, что хотя аналитическое представление и сам анализ полученных ре-

зультатов в ходе исследования является самым универсальным способом, но полученные выводы верны с той степенью надежности и при тех условиях, при которых они были получены. Дальнейшее обобщение полученных модельных результатов необходимо проводить с большой осторожностью.

Далеко не перед каждым исследователем предстает необходимость проводить работу в такой последовательности и объеме, как изложено выше, а именно, от постановки задачи до получения результатов математического моделирования. Однако большая часть проблем, затронутых в данной работе, так или иначе, станут актуальными при проведении любого, даже самого простого эксперимента. Многолетний опыт работы в области биотехнологии, где особенно требуется применение комплексных методов исследования и его обобщение, позволил создать выше приведенный алгоритм проведения работы.

### ВОПРОСЫ ДЛЯ САМОКОНТРОЛЯ.

1. В чем заключается уникальность возможностей Data Mining?
2. Опишите особенности применения сценарного подхода в исследовании.
3. Какие способы повышения воспроизводимости результатов биотехнологического исследования Вам известны?
4. Как определить достоверность значений исследуемых параметров?
5. Охарактеризуйте способы обобщения результатов, полученных в ходе проведения исследования.
6. Перечислите основные способы обобщения результатов, полученных в ходе проведения исследования.
7. Какова необходимость получения математических моделей?
8. Охарактеризуйте процедуру получения эмпирических формул.
9. В чем заключено основное отличие эмпирических формул от рациональных?
10. Как производится установление информационной ценности математической модели?
11. Как производится проверка адекватности описания математической моделью экспериментальных данных
12. Назовите причины низкой воспроизводимости биологических экспериментов. В чем выражается влияние космофизических факторов на случайные процессы исследованные С.Э.Шнолем?

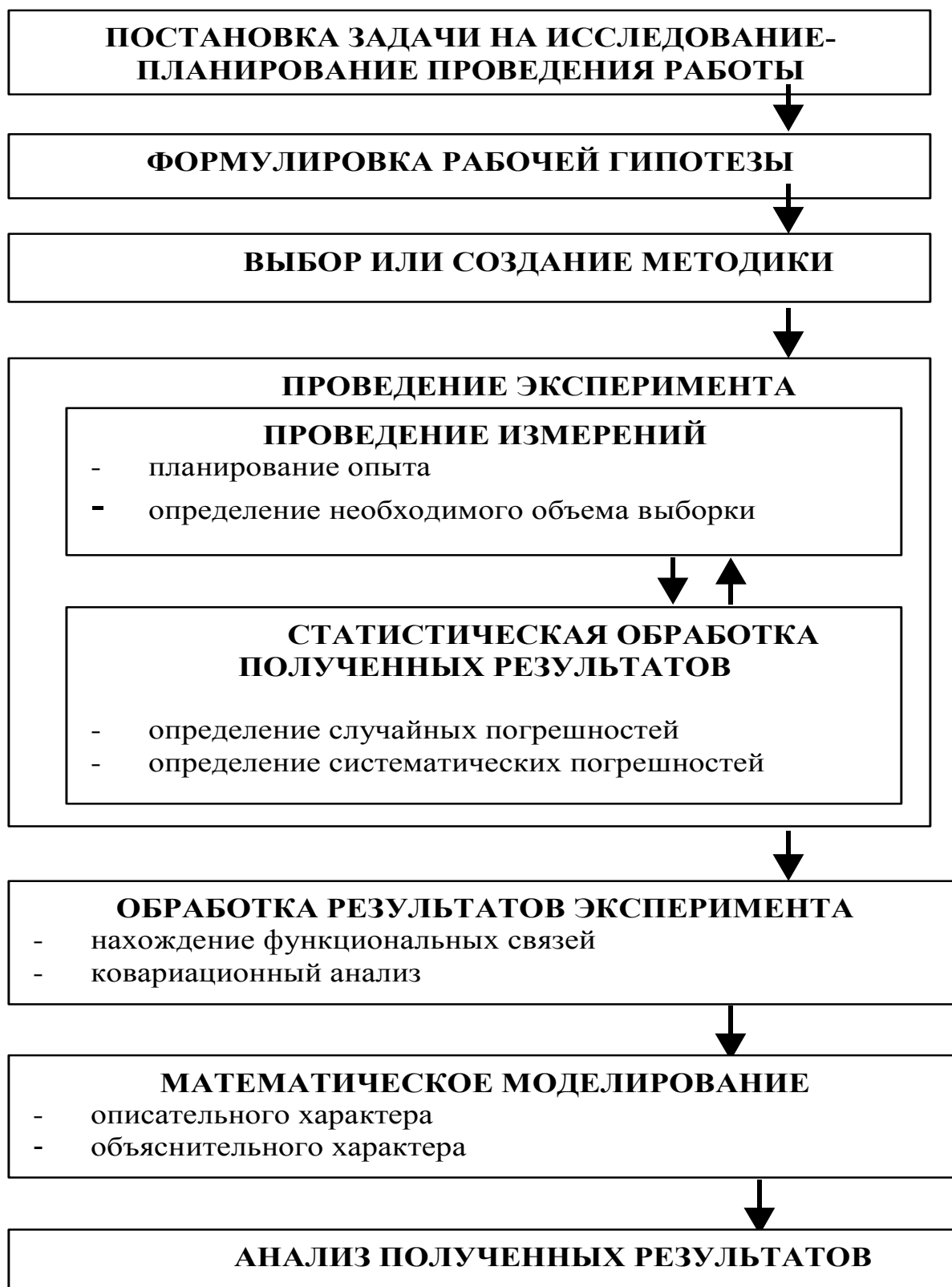
## Заключение

Разработка технологии оптимизации началась примерно 60 лет назад с осознания того факта, что наличие математических моделей (какими бы полными и точными они не были) не является достаточным условием успешного проектирования и модификации сложных технических систем. Для создания конкурентоспособных образцов необходима интеграция математических моделей с такими поисковыми методами исследования, которые бы давали возможность предоставить исследователю для принятия окончательного решения целое множество альтернативных технических решений, не улучшаемых по всей совокупности показателей эффективности (принцип Парето). Применительно к реальным задачам, механическое слияние математических моделей и, например, численных методов оптимизации не является достаточным условием получения практически важных результатов. В этом случае требуется создание некоторой "оптимизационной среды", в которой объединяются различные методы исследования. Собственно эту среду мы и называем технологией оптимизации. В настоящее время она успешно используется для совершенствования технических систем (технологических процессов) и в других областях науки и техники. Однако появляются принципиальные сложности в применении этой технологии в экологии, биотехнологии, экономике и т.п. при условии наличия участия в проведении исследований специалистов из этих областей.

Нами показано, что эффективность научной работы биотехнолога можно значительно повысить посредством выбора адекватного метода проведения многофакторного эксперимента. Установлено, что выбор адекватного метода проведения многофакторного эксперимента зависит от степени изученности исследуемой проблемы и уровня её формализации. При исследовании неизученных процессов более приемлемым оказывается симплексный метод. Для уточнения уже изученного явления и представленного в аналитическом виде, наиболее подходит метод Гаусса-Зейделя. При исследовании процессов, изученных не полностью и представленных в виде регрессионных уравнений, частично описывающих участки поверхности отклика, наиболее подходят методы крутого восхождения. Упрощение процедуры проведения биотехнологического исследования обеспечивается благодаря использованию предложенного векторного способа с графическим представлением получаемых результатов.

Таким образом, сравнительный анализ показывает, что при увеличении количества совместно исследуемых факторов до пяти, потери времени и денежных средств многократно сокращаются в сравнении с общепринятым однофакторным методом исследования. Проведение многофакторного исследования биотехнологических параметров возможно благодаря предложенному векторному способу. Поиск и построение поверхности отклика производится в системе абсолютных величин, а обобщение параметров, имеющих различную размерность, происходит благодаря вводу относительного показателя – коэффициента крутизны, определяющего шаг и направление изменения каждого параметра. Отличительная особенность предложенного способа от существующих аналогов заключается в объединении их преимуществ и отсутствии ограничений, присущих каждому из них.

# СТРУКТУРНАЯ СХЕМА ПРОВЕДЕНИЯ БИОЛОГИЧЕСКОГО ИССЛЕДОВАНИЯ





# ПОЛУЧЕНИЕ ОСНОВНЫХ СТАТИСТИЧЕСКИХ ПАРАМЕТРОВ И КРИТЕРИЕВ ИХ ДОСТОВЕРНОСТИ

Таблица 1. Получение основных статистических параметров и критериев их достоверности, выраженных через моменты распределения

$M_h = \frac{1}{N} \sum_{i=1}^N n_i (x_i - x)^h$ - момент h-го распределения.	$m_H$ - ошибка репрезентативности.	$t_{st}(p,n)$ - критерий Стьюдента (критерий достоверности).  $t_i = \frac{x_i - M}{\sigma}$ - текущего значения. Если $t_i > 3$ , то $x_i$ не принадлежит данной выборке.
$M_1 _{x=0} = \bar{X}$ - среднее арифметическое.	$m_x = \frac{\sigma}{\sqrt{n}}$	
$M_2 _{x=\bar{x}} = \sigma^2$ - дисперсия.	$m_\sigma = \frac{\sigma^2}{\sqrt{2n}}$	
$A_s = \frac{M_3}{\sigma^3}$ (0,2 < $A_s$ < 0,5) - коэффициент асимметрии. $E_x = \frac{M_4}{\sigma^4} - 3$ - коэффициент эксцесса. ( $E_x \approx 3$ - условие нормальности распределения).	$m_A = \frac{\sqrt{\sigma}}{n}$  $m_E = 2m_A$	$t_p = \frac{\alpha M_h}{m_M} = \frac{(1-p)M_h}{m_M}$ - достоверность конкретной выборки. Если $t_p \geq t_{st}$ , то $H_0=0$ данное распределение.  Принадлежит генеральному, при заданном уровне надёжности - $p$ ( $\alpha$ - уровень значимости выборки).
$m = \frac{\sigma}{\sqrt{n}}$ - статистическая ошибка. $C_v = \frac{\sigma}{M_1}$ - коэффициент вариации.	$m_m = \frac{m}{\sqrt{2n}}$  $m_{C_v} = \frac{C_v}{\sqrt{2n}}$	
$F_{ac} = \frac{\sigma_a^2}{\sigma_c^2}$ - критерий Фишера. $r_{ac} = \frac{\sum t_a t_c}{n}$ - коэффициент корреляции. $R_{ac} = r_{ac} \frac{\sigma_a}{\sigma_c}$ - коэффициент регрессии	$m_r = \sqrt{\frac{1-r^2}{n-2}}$  $m_R = \frac{\sigma_a}{\sigma_c} m_r$	

Примечание:  $H_0$  - нулевая гипотеза, (предполагает принадлежность данного распределения генеральному  $H_0=0$ ).

## ТАБЛИЧНЫЕ ДАННЫЕ КРИТЕРИЯ СТЬЮДЕНТА

Таблица 2. Значения коэффициента Стьюдента(t -распределение)

\ P k \	0,95	0,99	0,999
2	4,303	9,925	31,598
3	3,182	5,841	12,924
4	2,776	4,604	8,610
5	2,571	4,032	5,959
6	2,447	3,707	5,408
7	2,365	3,499	5,041
8	2,306	3,355	4,781
9	2,262	3,250	4,587
10	2,228	3,169	4,318
12	2,179	3,055	4,140
4	2,145	2,977	4,015
6	2,120	2,921	3,922
8	2,101	2,878	3,850
20	2,086	2,845	3,792
22	2,074	2,819	3,745
24	2,064	2,797	3,707
26	2,056	2,779	3,674
28	2,048	2,763	3,646
30	2,042	2,750	3,551
40	2,021	2,704	3,460
60	2,000	2,660	3,373
120	1,980	2,617	3,291
$\infty$	1,960	2,576	

Примечание: k - число степеней свобод. Для нахождения критерия Стьюдента, в случае одной выборки  $t_p$ ,  $k=n-1$ . Для установления достоверности различия двух выборок  $t_d$ ,  $k=n_1+n_2-2$ .

## ТАБЛИЧНЫЕ ДАННЫЕ КРИТЕРИЯ ФИШЕРА

Таблица 3. Значения критерия Фишера (F-распределение)

\N1 N2\	2	3	4	5	6	7	8	9	10	20	40	60
2	9,0 9,0	9,2 9,2	9,3 9,3	9,3 9,3	9,3 9,3	9,3 9,4	9,4 9,4	9,4 9,4	9,4 99,4	19,5 99,5	19,5 99,5	9,5 99,5
3	9,5 30,8	9,3 29,5	9,1 8,7	9,0 28,2	8,9 27,9	8,9 27,7	8,9 27,5	8,8 27,4	8,8 27,2	8,7 26,7	8,6 26,4	8,6 26,3
4	6,9 18,0	6,6 16,7	6,4 6,0	6,3 15,5	6,2 15,2	6,1 15,0	6,1 14,8	6,0 14,7	6,0 14,6	5,8 14,0	5,7 13,8	5,7 13,7
5	5,8 13,3	5,4 12,1	5,2 1,4	5,1 11,0	4,9 10,7	4,9 10,5	4,8 10,3	4,8 10,2	4,7 10,1	4,6 9,6	4,5 9,3	4,4 9,2
6	5,1 10,9	4,8 9,8	4,5 9,2	4,4 8,8	4,3 8,5	4,2 8,3	4,2 8,1	4,1 8,0	4,1 7,8	3,9 7,4	3,8 7,1	3,7 7,1
7	4,7 9,6	4,4 8,5	4,1 7,6	4,0 7,5	3,9 7,2	3,8 7,0	3,7 6,8	3,7 6,7	3,6 6,6	3,4 6,2	3,3 5,9	3,3 5,8
8	4,5 8,6	4,1 7,6	3,8 7,0	3,7 6,6	3,6 6,4	3,5 6,2	3,4 6,0	3,4 5,9	3,3 5,8	3,1 5,4	3,0 5,1	3,0 5,0
9	4,3 8,0	3,9 7,0	3,6 6,4	3,5 6,1	3,4 5,8	3,3 5,6	3,2 5,5	3,2 5,4	3,1 5,3	2,9 4,8	2,8 4,6	2,8 4,5
10	4,1 7,6	3,7 6,6	3,5 6,0	3,3 5,6	3,2 5,4	3,1 5,2	3,1 5,1	3,1 4,9	3,0 4,9	2,8 4,4	2,7 4,2	2,6 4,1
20	3,5 5,9	3,1 4,9	2,8 4,4	2,7 4,1	2,6 3,9	2,5 3,7	2,5 3,6	2,4 3,5	2,4 3,4	2,1 2,9	2,0 2,7	2,0 2,6
40	3,2 5,2	2,8 4,3	2,6 3,8	2,5 3,5	2,3 3,3	2,3 3,1	2,2 3,0	2,1 2,9	2,1 2,8	1,8 2,4	1,7 2,1	1,6 2,0
60	3,2 4,5	2,7 4,1	2,5 3,7	2,4 3,3	2,3 3,1	2,2 3,0	2,1 2,8	2,0 2,7	2,0 2,6	1,8 2,2	1,7 1,9	1,6 1,8

Примечание: верхний ряд - при  $P=0,95$ , нижний - при  $P=0,99$ .

. ТАБЛИЧНЫЕ ДАННЫЕ КРИТЕРИЯ  $\chi^2$ Таблица 4. Критические точки распределения  $\chi^2$ 

Число степеней свободы $k$	Уровень значимости $\alpha$					
	0,01	0,025	0,05	0,95	0,975	0,99
1	6,6	5,0	3,8	0,0039	0,00098	0,00016
2	9,2	7,4	6,0	0,103	0,051	0,020
3	11,3	9,4	7,8	0,352	0,216	0,115
4	13,3	11,1	9,5	0,711	0,484	0,297
5	15,1	12,8	11,1	1,15	0,831	0,554
6	16,8	14,4	12,6	1,64	1,24	0,872
7	18,5	16,0	14,1	2,17	1,69	1,24
8	20,1	17,5	15,5	2,73	2,18	1,65
9	21,7	19,0	16,9	3,33	2,70	2,09
10	23,2	20,5	18,3	3,94	3,25	2,56
11	24,7	21,9	19,7	4,57	3,82	3,05
12	26,2	23,3	21,0	5,23	4,40	3,57
13	27,7	24,7	22,4	5,89	5,01	4,11
14	29,1	26,1	23,7	6,57	5,63	4,66
15	30,6	27,5	25,0	7,26	6,26	5,23
16	32,0	28,8	26,3	7,96	6,91	5,81
17	33,4	30,2	27,6	8,67	7,56	6,41
18	34,8	31,5	28,9	9,39	8,23	7,01
19	36,2	32,9	30,1	10,1	8,91	7,63
20	37,6	34,2	31,4	10,9	9,59	8,26
21	38,9	35,5	32,7	11,6	10,3	8,90
22	40,3	36,8	33,9	12,3	11,0	9,54
23	41,6	38,1	35,2	13,1	11,7	10,2
24	43,0	39,4	36,4	13,8	12,4	10,9
25	44,3	40,6	37,7	14,6	13,1	11,5
26	45,6	41,9	38,9	15,4	13,8	12,2
27	47,0	43,2	40,1	16,2	14,6	12,9
28	48,3	44,5	41,3	16,9	15,3	13,6
29	49,6	45,7	42,6	17,7	16,0	14,3
30	50,9	47,0	43,8	18,5	16,8	15,0

## ПРИМЕРЫ

**Пример 1.** Установить основные статистические величины для представленных выборок с уровнем надежности  $P=0,95$  и  $P=0,99$ .

А) **Количественный метод.** Измерение силы тока в первой выборке составило: 28,8; 26,3; 31,4; 26,4; 27,4; 29,4; 25,1 мкА. Во второй выборке измерения силы тока составило: 32,8; 37,4; 36,6; 37,1; 42,3; 33,4; 28,9; 37,6 мкА.

Б) **Качественный метод** - альтернативного варьирования. Определить величину сохранности биообъекта  $C\%$  в каждой из представленных выборок. Общее количество биообъекта в первой выборке составляет  $n_1=56$  и во второй  $n_2=66$ . Количество биообъекта, которое сохранило заданное качество в первой выборке  $n_{01}=36$ , а другой  $n_{02}=48$ .

**РЕШЕНИЕ:**

А) **Количественный метод.**

1) **Определим наличие грубых ошибок.**

а) Для первой выборки: 25,1; 26,3; 26,4; 27,4; 28,8; 29,4; 31,4;  $n_1=7$ .

б) Для второй: 28,9; 32,8; 33,4; 36,6; 37,1; 37,4; 37,6; 42,3;  $n_2=8$ .

$$Q_1 = \frac{x_{i+1} - x_i}{x_n - x_i} = \frac{26,3 - 25,1}{31,4 - 25,1} = \frac{1,2}{6,3} = 0,19 < 0,51; \quad Q_2 = \frac{32,8 - 28,9}{42,3 - 28,9} = \frac{3,9}{13,4} = 0,29 < 0,47.$$

Вывод - грубых ошибок нет.

2) **Вычислим средние величины по каждой из представленных выборок.**

$$\bar{x}_1 = \frac{\sum x_i}{n} = \frac{194,8}{7} = 27,83; \quad \bar{x}_2 = \frac{286,1}{8} = 35,76.$$

3) **Установим величины среднеквадратического отклонения**

$$\sigma_1 = \sqrt{\frac{\sum (\bar{x}_1 - x_i)^2}{n-1}} = \frac{7,45 + 2,34 + 2,04 + 0,18 + 0,94 + 2,46 + 12,34}{6} = 2,17;$$

$$\sigma_2 = \frac{47,06 + 8,76 + 5,57 + 0,71 + 1,8 + 2,69 + 3,39 + 42,77}{7} = 4,01;$$

Вывод - измерения неравноточные.

4) **Определим коэффициенты вариации наших выборок**

$$C_{V_1} = \frac{\sigma_1}{\bar{x}_1} \cdot 100\% = \frac{2,17}{27,83} \cdot 100\% = 7,8\% \text{ - вариабельность средняя;}$$

$$C_{V_2} = \frac{\sigma_2}{\bar{x}_2} \cdot 100\% = \frac{4,01}{35,76} \cdot 100\% = 11,21\% \text{ - вариабельность высокая.}$$

5) **Установим величины ошибок среднеквадратического отклонения**

$$m_1 = \frac{\sigma_1}{\sqrt{n_1}} = \frac{2,17}{2,65} = 0,819; \quad m_2 = \frac{\sigma_2}{\sqrt{n_2}} = \frac{4,01}{2,83} = 1,417.$$

6) **Оценим величины доверительных интервалов при различных уровнях надежности (при  $P=0,95$   $t=2,16$  и  $P=0,99$   $t=3,02$ ).**

а) для уровня надежности  $P=0,95$ :

$$\Delta_1 = t_{p_1} \cdot m_1 = 2,447 \cdot 0,819 = 2,15; \quad \Delta_2 = t_{p_2} \cdot m_2 = 2,365 \cdot 1,417 = 3,35.$$

б) для уровня надежности  $P=0,99$ :

$$\Delta_1 = 3,707 \cdot 0,819 = 3,036; \quad \Delta_2 = 3,499 \cdot 1,417 = 4,96.$$

7) **Установим относительные ошибки измерения** (уровень значимости оценки достоверности) для различных уровней надежности  $P=0,95$  и  $P=0,99$ .

а) для уровня надежности  $P=0,95$ :

$$\rho_1 = \frac{\Delta_1}{x_1} \cdot 100\% = \frac{2,15}{27,83} \cdot 100\% = 7,73\% > 5\%; \quad \rho_2 = \frac{\Delta_2}{x_2} \cdot 100\% = \frac{3,35}{35,76} \cdot 100\% = 9,37\% > 5\%$$

б) для уровня надежности  $P=0,99$ :

$$\rho_1 = \frac{3,036}{27,83} \cdot 100\% = 10,90\% > 1\%; \quad \rho_2 = \frac{4,96}{35,76} \cdot 100\% = 13,87\% > 1\%.$$

Вывод - необходимо большее число измерений.

8) **Установим минимальное количество измерений в каждой из выборок.**

а) для уровня надежности  $P=0,95$ :

$$n_{01} = \left( \frac{t_1}{\rho_1} \cdot C_{v_1} \right)^2 = \left( \frac{2,447}{5} \cdot 7,85 \right)^2 = 14,57; \quad n_{02} = \left( \frac{t_2}{\rho_2} \cdot C_{v_2} \right)^2 = \left( \frac{2,365}{5} \cdot 11,2 \right)^2 = 28,06.$$

б) для уровня надежности  $P=0,99$ :

$$n_{01} = \left( \frac{3,707}{1} \cdot 7,8 \right)^2 = 836; \quad n_{02} = \left( \frac{3,499}{1} \cdot 11,2 \right)^2 = 1536.$$

Б) **Качественный -метод альтернативного варьирования.**

1) **Установим показатели вероятности для каждой из выборок.**

$$W_1 = \frac{n_{c1}}{n_{01}} \cdot 100\% = \frac{36}{56} \cdot 100\% = 64,29\%; \quad W_2 = \frac{n_{c2}}{n_{02}} \cdot 100\% = \frac{48}{66} \cdot 100\% = 72,73\%.$$

2) **Вычислим среднеквадратическое отклонение.**

$$\sigma_1 = \sqrt{W_1 \cdot (100 - W_1)} = \sqrt{64,29 \cdot (100 - 64,29)} = 47,91; \quad \sigma_2 = \sqrt{72,73 \cdot (100 - 72,73)} = 44,53.$$

Вывод - измерения неравноточные.

3) **Определим коэффициенты вариации.**

$$C_{v_1} = \frac{\sigma_1}{W_1} \cdot 100\% = \frac{47,91}{64,29} \cdot 100\% = 74,52 - \text{высокая вариабельность},$$

$$C_{v_2} = \frac{\sigma_2}{W_2} \cdot 100\% = \frac{44,53}{72,73} \cdot 100\% = 61,23 - \text{высокая вариабельность}.$$

4) **Оценим ошибки среднеквадратического отклонения.**

$$m_1 = \frac{\sigma_1}{\sqrt{n_1}} = \frac{47,91}{\sqrt{56}} = 6,41; \quad m_2 = \frac{\sigma_2}{\sqrt{n_2}} = \frac{44,53}{\sqrt{66}} = 5,48.$$

5) **Оценим величины доверительных интервалов при различных уровнях надежности.**

а) для уровня надежности  $P=0,95$ :

$$\Delta_1 = t_{p_1} \cdot m_1 = 2 \cdot 6,41 = 12,82; \quad \Delta_2 = t_{p_2} \cdot m_2 = 2 \cdot 5,48 = 10,96.$$

б) для уровня надежности  $P=0,99$ :

$$\Delta_1 = 2,66 \cdot 6,41 = 17,05; \quad \Delta_2 = 2,66 \cdot 5,48 = 14,58.$$

6) **Установим относительные ошибки измерения** (уровень значимости оценки достоверности) для различных уровней надежности  $P=0,95$  и  $P=0,99$ .

а) для уровня надежности  $P=0,95$ :

$$\rho_1 = \frac{\Delta_1}{W_1} \cdot 100\% = \frac{12,82}{64,29} \cdot 100\% = 19,94\% > 5\% ;$$

$$\rho_2 = \frac{\Delta_2}{W_2} \cdot 100\% = \frac{10,96}{72,73} \cdot 100\% = 15,07\% > 5\% .$$

б) для уровня надежности  $P=0,99$ :

$$\rho_1 = \frac{\Delta_1}{W_1} \cdot 100\% = \frac{17,05}{64,29} \cdot 100\% = 26,52\% > 1\% ;$$

$$\rho_2 = \frac{\Delta_2}{W_2} \cdot 100\% = \frac{14,58}{72,73} \cdot 100\% = 20,05\% > 1\% .$$

Вывод - необходимо увеличить количество измерений.

7) **Установим минимальное количество измерений в каждой из выборок.**

а) для уровня надежности  $P=0,95$ :

$$n_1 = \left( \frac{t_1}{\rho_1} \cdot C_{V_1} \right)^2 = \left( \frac{2}{5} \cdot 74,52 \right)^2 = 889 ; \quad n_2 = \left( \frac{t_2}{\rho_2} \cdot C_{V_2} \right)^2 = \left( \frac{2}{5} \cdot 61,23 \right)^2 = 600 .$$

б) для уровня надежности  $P=0,99$ :

$$n_1 = \left( \frac{2,66}{1} \cdot 74,52 \right)^2 = 39292 ; \quad n_2 = \left( \frac{2,66}{1} \cdot 61,23 \right)^2 = 26527 .$$

Пример 2. Установить достоверность различия сравниваемых выборок с уровнем надежности  $P=0,95$  и  $P=0,99$ .

А) **Количественный метод.** Измерение силы тока в первой выборке составило: 27,8; 28,3; 31,4; 26,4; 29,4; 24,4; 27,1 мкА. Во второй выборке измерения силы тока составило: 32,8; 37,4; 36,6; 38,1; 41,3; 32,4; 28,9; 39,6 мкА.

Б) **Качественный метод** - альтернативного варьирования. Определить величину сохранности биообъекта ( $C\%$ ) в каждой из представленных выборок. Общее количество биообъекта в первой выборке составляет  $n_1=66$  и во второй  $n_2=78$ . Количество биообъекта, которое сохранило заданное качество в первой выборке  $n_{01}=23$  и во второй  $n_{02}=25$ .

### **РЕШЕНИЕ:**

А) **Количественный метод.**

1) **Определим наличие грубых ошибок.**

а) Для первой выборки: 24,4; 26,4; 27,1; 27,8; 28,3; 29,4; 31,4.

$$Q = \frac{x_{i+1} - x_i}{x_n - x_i} ; \quad Q_1 = \frac{26,4 - 24,4}{31,4 - 24,4} = \frac{2}{7} = 0,286 < 0,51 ; \quad Q_2 = \frac{31,4 - 29,4}{31,4 - 24,4} = \frac{2}{7} = 0,286 < 0,51 .$$

Вывод - первая выборка не имеет грубых ошибок.

б) Для второй: 28,9; 32,4; 32,8; 36,6; 37,4; 38,1; 39,6; 41,3.

$$Q_1 = \frac{32,4 - 28,9}{41,3 - 28,9} = \frac{3,5}{12,4} = 0,28 < 0,47; \quad Q_2 = \frac{41,3 - 39,6}{41,3 - 28,9} = \frac{1,7}{12,4} = 0,13 < 0,47.$$

Вывод - вторая выборка не имеет грубых ошибок.

2) **Вычислим средние величины по каждой из представленных выборок.**

$$M_1 = \frac{\sum x_i}{n} = \frac{24,4 + 26,4 + 27,1 + 27,8 + 28,3 + 29,4 + 31,4}{7} = \frac{194,8}{7} = 27,83;$$

$$M_2 = \frac{\sum x_i}{n} = \frac{28,9 + 32,4 + 32,8 + 36,6 + 37,4 + 38,1 + 39,6 + 41,3}{8} = \frac{287,1}{8} = 35,88.$$

3) **Установим величины среднеквадратического отклонения.**

$$\begin{aligned} \sigma_1 &= \sqrt{\frac{\sum (M_1 - x_i)^2}{n-1}} = \sqrt{\frac{(27,83 - 24,4)^2 + (27,83 - 26,4)^2 + (27,83 - 27,1)^2 + (27,83 - 27,8)^2 +}{6} +} \\ &\quad + \frac{(27,83 - 28,3)^2 + (27,83 - 29,4)^2 + (27,83 - 31,4)^2}{6} = \\ &= \sqrt{\frac{11,7649 + 2,0449 + 0,5329 + 0,0009 + 0,2209 + 2,4649 + 12,7449}{6}} = 2,23 \\ \sigma_2 &= \sqrt{\frac{\sum (M_2 - x_i)^2}{n-1}} = \sqrt{\frac{(35,88 - 28,9)^2 + (35,88 - 32,4)^2 + (35,88 - 32,8)^2 + (35,88 - 36,6)^2 +}{7} +} \\ &\quad + \frac{(35,88 - 37,4)^2 + (35,88 - 38,1)^2 + (35,88 - 39,6)^2 + (35,88 - 41,3)^2}{7} = \\ &= \sqrt{\frac{48,7204 + 12,1104 + 9,4864 + 0,5184 + 2,3104 + 4,9284 + 13,8384 + 29,3764}{7}} = 4,16 \end{aligned}$$

4) **Определим коэффициенты вариации наших выборок.**

$$C_{V_1} = \frac{\sigma_1}{M_1} \cdot 100 = \frac{2,23}{27,83} \cdot 100 = 8,01\%; \quad C_{V_2} = \frac{\sigma_2}{M_2} \cdot 100 = \frac{4,16}{35,88} \cdot 100 = 11,59\%.$$

5) **Установим величины ошибок среднеквадратического отклонения.**

$$m_1 = \frac{\sigma_1}{\sqrt{n_1}} = \frac{2,23}{\sqrt{7}} = 0,84; \quad m_2 = \frac{\sigma_2}{\sqrt{n_2}} = \frac{4,16}{\sqrt{8}} = 1,47.$$

6) **Оценим величины доверительных интервалов при различных уровнях надежности (при  $P=0,95$   $t=2,16$  и  $P=0,99$   $t=3,02$ ).**

$$\Delta_1 = t \cdot m_1 = 2,162 \cdot 0,84 = 1,82; \quad \Delta_2 = t \cdot m_1 = 3,016 \cdot 0,84 = 2,53;$$

$$\Delta_3 = t \cdot m_2 = 2,162 \cdot 1,47 = 3,18; \quad \Delta_4 = t \cdot m_2 = 3,016 \cdot 1,47 = 4,43.$$

7) **Установим относительные ошибки измерения для различных уровней надежности  $P=0,95$  и  $P=0,99$ .**

$$\rho_1 = \frac{\Delta_1}{M_1} \cdot 100 = \frac{1,82}{27,83} \cdot 100 = 6,54\%; \quad \rho_2 = \frac{\Delta_2}{M_1} \cdot 100 = \frac{2,53}{27,83} \cdot 100 = 9,09\%;$$

$$\rho_3 = \frac{\Delta_3}{M_2} \cdot 100 = \frac{3,18}{35,88} \cdot 100 = 8,86\%; \quad \rho_4 = \frac{\Delta_4}{M_2} \cdot 100 = \frac{4,43}{35,88} \cdot 100 = 12,35\%.$$

8) **Определим достоверность различия сравниваемых выборок.**



$$t_{d_1} = \frac{M_1 - M_2}{\sqrt{m_1^2 + m_2^2}} = \frac{27,83 - 35,88}{\sqrt{0,84^2 + 1,47^2}} = \frac{-8,05}{\sqrt{0,7056 + 2,1609}} = \frac{-8,05}{1,693} = -4,75.$$

Вывод - при  $t=4,75$  достоверны различия двух выборок для  $P=0,95$  и  $P=0,99$ .

**Б) Метод альтернативного варьирования.**

**1) Установим показатели вероятности для каждой из выборок.**

$$W_1 = \frac{n_{01}}{n_{c1}} \cdot 100 = \frac{23}{66} \cdot 100 = 35\% ; \quad W_2 = \frac{n_{02}}{n_{c2}} \cdot 100 = \frac{25}{78} \cdot 100 = 32\% .$$

**2) Вычислим среднеквадратическое отклонение.**

$$\sigma_1 = \sqrt{W_1 \cdot (100 - W_1)} = \sqrt{35 \cdot (100 - 35)} = 47,7 ; \quad \sigma_2 = \sqrt{W_2 \cdot (100 - W_2)} = \sqrt{32 \cdot (100 - 32)} = 46,6 .$$

**3) Определим коэффициенты вариации.**

$$C_{V_1} = \frac{\sigma_1}{W_1} \cdot 100 = \frac{47,7}{35} \cdot 100 = 136\% ; \quad C_{V_2} = \frac{\sigma_2}{W_2} \cdot 100 = \frac{46,6}{32} \cdot 100 = 145,6\% .$$

**4) Оценим ошибки среднеквадратического отклонения.**

$$m_1 = \frac{\sigma_1}{\sqrt{n_1}} = \frac{47,7}{\sqrt{66}} = 5,87 ; \quad m_2 = \frac{\sigma_2}{\sqrt{n_2}} = \frac{46,6}{\sqrt{78}} = 5,28 .$$

**5) Установим достоверность различия сравниваемых выборок.**

$$t_{d_2} = \frac{W_1 - W_2}{\sqrt{m_1^2 + m_2^2}} = \frac{35 - 32}{\sqrt{5,87^2 + 5,28^2}} = \frac{3}{\sqrt{34,46 + 27,88}} = \frac{3}{7,9} = 0,38.$$

Вывод: при  $t_d=0,38$  различия двух выборок недостоверны при уровнях надежности  $P=0,95$  и  $P=0,99$ .

## ПРИБЛИЖЕННАЯ ОЦЕНКА СТАТИСТИЧЕСКИХ ПОКАЗАТЕЛЕЙ:

### *Количественный анализ* ( $n_{\min}=10\pm 5$ ).

1) Установим среднее арифметическое.

$$\bar{x}_1 = \frac{x_{\max_1} + x_{\min_1}}{2} = \frac{31,4 + 25,1}{2} = 28,25; \quad \bar{x}_2 = \frac{x_{\max_2} + x_{\min_2}}{2} = \frac{42,3 + 28,9}{2} = 35,6.$$

Для точных методов оценки средние значения составляют:

$$\bar{x}_1 = 27,83; \quad \bar{x}_2 = 35,76;$$

Вывод – относительный показатель  $\Delta$  оценки расхождения точных и приближенных значений составил:  $\Delta_1 = 1,5\%$ ;  $\Delta_2 = 0,4\%$ .

2) Установим среднеквадратическое отклонение. Для этого оценим размах выборки.

$$\Delta x_1 = x_{\max_1} - x_{\min_1} = 31,4 - 25,1 = 6,3; \quad \Delta x_2 = x_{\max_2} - x_{\min_2} = 42,3 - 28,9 = 13,4.$$

$$\sigma_1 = \frac{\Delta x_1}{3} = \frac{6,3}{3} = 2,1; \quad \sigma_2 = \frac{\Delta x_2}{3} = \frac{13,4}{3} = 4,47.$$

Для точных значений среднеквадратические отклонение составляют:

$\sigma_1 = 2,17$ ;  $\sigma_2 = 4,04$ , следовательно, относительный показатель  $\Delta$  оценки расхождения точных и приближенных значений составил:  $\Delta_1 = 3,2\%$ ;  $\Delta_2 = 10,6\%$ .

3) Установим коэффициент вариации ( $C_{V_1} = 7,8$ ;  $C_{V_2} = 11,2$ ).

$$C'_{V_1} = \frac{\Delta x_1}{\bar{x}_1} \cdot 30 = \frac{6,3}{28,25} \cdot 30 = 6,69; \quad C'_{V_2} = \frac{\Delta x_2}{\bar{x}_2} \cdot 30 = \frac{13,4}{35,6} \cdot 30 = 11,29.$$

Вывод – относительный показатель  $\Delta$  оценки расхождения точных и приближенных значений составил:  $\Delta_1 = 14,2\%$ ,  $\Delta_2 = 0,8\%$ .

4) Установим уровень значимости оценки достоверности ( $\rho_1 = 7,73$ ,  $\rho_2 = 9,35$ ).

$$\rho'_1 = \frac{C_{V_1}}{1,4} = \frac{6,69}{1,4} = 4,78 - \text{измерений достаточно,}$$

$$\rho'_2 = \frac{C_{V_2}}{1,4} = \frac{11,29}{1,4} = 8,06 - \text{необходимо увеличить число измерений.}$$

Вывод – относительный показатель  $\Delta$  оценки расхождения точных и приближенных значений составил:  $\Delta_1 = 38\%$ ,  $\Delta_2 = 13,6\%$ .

5) Установим минимальное количество измерений ( $\alpha = 5$ ,  $n_{0_1} = 14,57$ ,  $n_{0_2} = 28,06$ ).

$$n'_{0_1} = \frac{C_{V_1}^2}{5} = \frac{6,69^2}{5} = 8,95; \quad n'_{0_2} = \frac{C_{V_2}^2}{5} = \frac{11,29^2}{5} = 25,49.$$

Вывод – относительный показатель  $\Delta$  оценки расхождения точных и приближенных значений составил:  $\Delta_1 = 38,6\%$ ,  $\Delta_2 = 9,2\%$ .

### *Качественный анализ.*

1) Установим вероятность.

$$W_1 = \frac{n_{11}}{n_1} \cdot 100\% = \frac{36}{56} \cdot 100\% = 64,29\%; \quad W_2 = \frac{n_{12}}{n_2} \cdot 100\% = \frac{48}{66} \cdot 100\% = 72,73\%.$$

2) **Определим размах вариации.**

$$\Delta W_1 = 100 - W_1 = 100 - 64,29 = 35,71; \Delta W_2 = 100 - W_2 = 100 - 72,73 = 27,27.$$

3) **Вычислим среднее квадратическое отклонение.**

$$\sigma'_1 = \sqrt{W_1 \cdot \Delta W_1} = \sqrt{64,29 \cdot 35,71} = 47,91; \sigma'_2 = \sqrt{W_2 \cdot \Delta W_2} = \sqrt{72,73 \cdot 27,27} = 44,53.$$

4) **Установим коэффициент вариации** ( $C_{V_1} = 74,52; C_{V_2} = 61,23$ ).

$$C'_{V_1} = \frac{\sigma}{W_1} \cdot 100\% = \frac{47,9}{64,29} \cdot 100\% = 74,52; C'_{V_2} = \frac{\sigma}{W_2} \cdot 100\% = \frac{44,53}{72,73} \cdot 100\% = 61,22.$$

Вывод – относительный показатель  $\Delta$  оценки расхождения точных и приближенных значений составил  $\Delta_1=0\%, \Delta_2=0\%$ .

5) **Определим уровень значимости оценки достоверности** ( $\rho_1=37,26, \rho_2=30,61$ ).

$$\rho'_1 = \frac{C_{V_1}}{2} = \frac{74,52}{2} = 37,26; \rho'_2 = \frac{C_{V_2}}{2} = \frac{61,22}{2} = 30,61; \Delta_1=0\%, \Delta_2=0\%.$$

6) **Вычислим минимальное число измерений** ( $\alpha=5; n_{0_1} = 992; n_{0_2} = 669$ ).

$$n'_{0_1} = \frac{C_{V_1}^2}{5,6} = \frac{74,52^2}{5,6} = 991,62; n'_{0_2} = \frac{C_{V_2}^2}{5,6} = \frac{61,22^2}{5,6} = 669,26; \Delta_1=0\%, \Delta_2=0\%.$$

**Вывод:** упрощенные расчеты условно приемлемы (погрешность не превышает 30%).

## Список условных обозначений

$M$  - среднее арифметическое  
 $W$  - вероятность события  
 $\sigma$  - среднее квадратическое отклонение  
 $C_v$  - коэффициент вариации  
 $m$  - средняя квадратическая ошибка среднего арифметического  
 $\Delta$  - доверительный интервал  
 $\Delta_\varepsilon$  - доверительный интервал для случайных погрешностей  
 $\Delta_\theta$  - доверительный интервал для систематических погрешностей  
 $n$  - объем выборки  
 $P$  - доверительная вероятность  
 $p$  - уровень значимости оценок  
 $\alpha$  - относительный показатель расхождения средних величин сравниваемых выборок  
 $n_0$  - минимальное количество измерений, обеспечивающее достоверный научный результат для одной выборки  
 $n_\Delta$  - минимальное количество измерений, обеспечивающее достоверный научный результат при сравнении двух выборок  
 $t, t_p$  - критерий Стьюдента  
 $t_d$  - критерий достоверности различия двух выборок  
 $F$  - критерий Фишера  
 $A_s$  - коэффициент асимметрии ряда распределения  
 $E_x$  - коэффициент эксцесса распределения  
 $r$  - коэффициент корреляции  
 $R$  - коэффициент регрессии  
 $C$  - целевая функция, численно отражающая поставленную цель  
 $a_j, v_j, d_j$  - параметры, определяющие развитие изучаемого явления:  
 $a_j$  - поддающиеся изменениям в ходе эксперимента.  
 $v_j$  - неподдающиеся изменениям в ходе эксперимента.  
 $d_j$  - неопределяемые в ходе эксперимента изменения.  
 $k$  - число степеней свобод.

## Литература

- 1 Атраментова Л.О. Біометрія Ч. I. Характеристики розподілів: Підручник / Атраментова Л.О., Утєвська О.М. – Х. : Видавництво «Ранок», 2007. – 176.
- 2 Атраментова Л.О. Біометрія Ч. II. Порівняння груп ш аналіз зв'язку: Підручник / Атраментова Л.О., Утєвська О.М. – Х. : Видавництво «Ранок», 2007. – 176.
- 3 Бейли Н. Статистические методы в биологии. М. Мир, 1964 270 с.
- 4 Бирюков В.В. Основы промышленной биотехнологии.– М.: Колос С, 2004. – 296 с.
- 5 Блохин А.В. Теория эксперимента [Электронный курс] / Блохин А.В. - Курс лекций в двух частях: Часть 1.- Электрон. текст. дан. (1,1Мб).– Мн.: Научно-методический центр «Электронная книга БГУ», 2003.
- 6 Блохин А.В. Теория эксперимента [Электронный курс] / Блохин А.В. - Курс лекций в двух частях: Часть 2.- Электрон. текст. дан. (1,0Мб).– Мн.: Научно-методический центр «Электронная книга БГУ», 2003.
- 7 Веденяпин Г.В. Общая методика экспериментального исследования и обработка опытных данных. М. Колос, 1965, 130 с.
- 8 Вознесенский В.А. Статистические методы планирования эксперимента в технико-экономических исследованиях. М. 1981. 250 с.
- 9 Гланц С. Медико-биологическая статистика / Пер. с англ. С. Гланц – М.: Практика, 1998. – 459 с.
- 10 Горбунов Л.В. Способи узагальнення результатів отриманих при проведенні експериментального біологічного дослідження // Рибе господарство. Міжвідомчий тематичний науковий збірник. - Київ. - 2001. - Випуск 59-60. - С. 55-58.
- 11 Горбунов Л.В. Определение минимального количества измерений, обеспечивающего достоверный научный результат // Агроекологічний журнал . – Київ. – 2002. – Випуск 1. –С. 69-71.
- 12 Горбунов Л.В. Кількісний спосіб оцінки життєздатності ембріонів ссавців // Вісник Білоцерківського державного аграрного університету. - Біла Церква.- 2001.- №17.- С. 30-34.
- 13 Горбунов Л.В. Методические рекомендации по моделированию биологического эксперимента: Утверждена на заседании методической комиссии факультета ветеринарной медицины Луганского национального аграрного университета 11.09.03. - Луганск, 2003. – 20 с.
- 14 Горбунов Л.В., Буцацкий Л.П. Метод моделирования биологического эксперимента: Методические рекомендации – Киев, 2004. – 20 с.
- 15 Горбунов Л.В. Оптимизация проведения биологического эксперимента векторным методом / Горбунов Л.В., Саминина М.Г. - Науково-технічний бюлетень ІТ УААН - 2006.- №94.- С. 108-113.
- 16 Горбунов Л.В. Крיוоконсервация половых клеток и эмбрионов/ Горбунов Л.В., Буцацкий Л.П. - Монографія – К.: Издательско-полиграфический центр “Киевский университет”, 2005 – 325 с.

17 Горбунов Л.В. Способы узагальнення результатів отриманих при проведенні експериментального біологічного дослідження/ Горбунов Л.В. - Рибне господарство. Міжвідомчий тематичний науковий збірник. – Київ. – 2001. – Випуск 59-60. – С. 55-58.

18 Горбунов Л.В. Воспроизводимость результатов криоконсервирования черенков различных сортов семечковых плодовых деревьев / Горбунов Л.В.// Проблемы криобиологии. – 2009.-№4. – С.473-480.

19 Горбунов Л.В. Воспроизводимость результатов криоконсервирования эмбрионов млекопитающих / Горбунов Л.В., Безуглый Н.Д., Салина А.С. // Біотехнологія. – 2010. – Т.3, №1. – С. 46 – 51.

20 Горбунов Л.В. Оптимизация криоконсервирования эмбрионов мыши / Горбунов Л.В., Салина А.С., Клещев Н.Ф.// Біотехнологія. – 2010. – Т.3, №5. – С. 100 – 105.

21 Горбунов Л.В. Повышение результатов воспроизводимости результатов исследования криоконсервирования биообъекта / Горбунов Л.В. // Проблемы криобиологии. – 2010.-№3. – С.246-254.

22 Доспехов Б.А. Методика полевого опыта. М. Агропромиздат 1985, 330 с.

23 Ивантер Э.В., Коросов А.В. Элементарная биометрия: Учеб. пособие / Э. В. Ивантер, А. В. Коросов; ПетрГУ. – Петрозаводск, 2005. – 104 с.

24 Лакин Г.Ф. Биометрия. М. “Высшая школа” 1990 г. 350 с.

25 Леонов В.П., Ижевский П.В. Применение статистики в медицине и биологии: анализ публикаций 1990-1997гг.// Сибирский медицинский журнал, 1997, 3-4, с. 64-74.

26 Новиков Д.А., Новочадов В.В. Статистические методы в медико-биологическом эксперименте (типовые случаи) // Волгоград: Изд-во ВолГМУ, 2005. – 84 с.

27 Орлов А.И. Высокие статистические технологии – М. “Заводская лаборатория” – 2003. Т.69. – №11. – С. 55-60.

28 Прикладная статистика. Правила определения оценки и доверительных границ для параметров нормального распределения. ГОСТ. 8.207-76.

29 Прямые измерения с многократными наблюдениями. Методы обработки результатов наблюдений. ГОСТ 8.207-76.

30 Планирование эксперимента в исследовании технологических процессов. / Под ред. канд. техн. наук Э.К. Лецкого. М.: Мир, 1977.–552 с.

31 Ризниченко Г.Ю., Рубин А.Б. Математические модели биологических продукционных процессов. М., изд. МГУ, 1993, 388 с

32 Саминіна М. Г., Горбунов Л. В. Оптимізація багатофакторних біотехнологічних досліджень // Науково-технічний бюлетень ІТ УААН – 2006.– №92.– Харків.– С. 100–104.

33 Методы исследований и организация экспериментов / под ред. проф. К.П. Власова. – Х.: Издательство «Гуманитарный центр», 2002. – 256 с.

34 Шноль С.Э. Космофизические факторы в случайных процессах М., изд. МГУ, 2009, 388 с

35 Чубукова И.А. Data Mining: Учебное пособие. – М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний, 2006. – 382 с.

## СОДЕРЖАНИЕ

Вступление	4
Глава 1. Методология исследования	7
1.1 Способы структурирования полученной информации	7
1.2 История, пути и направления развития статистики и биометрии	9
1.3 Повышение эффективности процедуры исследования	13
Глава 2. Первичная статистическая обработка данных	26
2.1 Сводка и группировка данных	26
2.2 Построение вариационного ряда	35
2.3 Вычисление основных статистических показателей	39
2.4 Основные типы распределений признаков	48
2.5 Преобразование выборки к нормальному виду	54
Глава 3. Планирование и проведение эксперимента	64
3.1 Учет ошибок измерения и вычисления	72
3.2 Определение условий решения поставленной задачи	77
3.3 Основы планирования многофакторного эксперимента	89
Глава 4. Анализ экспериментальных данных	96
4.1 Изучение сходства (различий) двух выборок	96
4.2 Дисперсионный анализ	124
4.3 Корреляционный анализ	134
4.4 Регрессионный анализ	144
Глава 5. Оптимизация результатов эксперимента	155
5.1 Особенности методов однофакторной оптимизации исследования	155
5.2 Необходимость проведения многофакторного эксперимента	160
5.3 Контурно-графический анализ	170
5.4. Оптимизация проведения многофакторного биологического эксперимента векторным методом	174
Глава 6. Математическое моделирование как инструмент научного исследования	189
6.1. Влияние космофизических факторов на случайные процессы исследованные С.Э.Шнолем	9
6.2 Уникальные возможности Data Mining	189
6.3 Повышение воспроизводимости результатов биотехнологического исследования	193
6.4 Способы обобщения результатов, полученных в ходе проведения исследования.	203
6.5 Тестирование полученной модели	206
Заключение	208
Приложение	209
Список условных обозначений	227
Литература	228

Навчальне видання

Горбунов Леонід Володимирович,  
Клещев Миколай Федосовіч

Лекції по курсу  
“Біометрія”  
для студентів спеціальності 7.092901  
“Промислова біотехнологія”

Відповідальний за випуск М.Ф. Клещев

Роботу до видання рекомендував ??

В авторській редакції

План 200\_\_ р., ???/???-??

Підписано до друку ???.??.??.??.? р. Формат 60 x 84 1/16

RISO-друк. Гарнітура Таймс. Ум. друк. арк. ?,?.

Наклад 100 прим. Зам. №

Папір офсетний

Обл. –вид. арк. ?,?.

Ціна договірна

Видавничий центр НТУ “ХПІ”

Свідоцтво про державну реєстрацію ДК № 116 від 10.07.2006 р.  
61002, Харків, вул. Фрунзе, 21

---

Друкарня НТУ “ХПІ”. 61002, Харків, вул. Фрунзе, 21

---