

Статистика в биологии

Зачем она нужна?

Основные понятия и допущения.

Как правильно выбрать
критерий?

Доказательная и ...иная биология и медицина

- Теоретическая
- Описательная
- Народная [медицина]
- «Эмпирическая»
- Интуитивная
- **Сомнительная**
- ...
- Народная [агрономия, селекция]

Зачем она нужна?

Черты доказательного подхода

- Элемент редукционистского, аналитического подхода к познанию;
- Наличие этапа планирования, целенаправленный поиск;
- Наличие ясных формулировок проверяемых гипотез;
- Ясное описание условий проводимых исследований и установленных эффектов;
- Частая сопряженность с экспериментальным подходом;
- Повторные исследования и повторяющиеся эффекты;
- **Статистическое подтверждение выводов;**
- **Осторожность, критичность, скептицизм.**

Зачем она нужна?

Статистика – инструмент генерализации заключений

- НЕ формальное требование; НЕ причуда высоколобых; НЕ требование «рецензентов» ; НЕ средство «давления» редколлегий журналов на авторов;
 - НЕ признак «научности», «современности» etc .;
 - НЕ способ «запутать» результаты там, где «и так все ясно»;
 - НЕ способ «впихнуть невпихуемое, чтобы STATISTICA там сама все посчитала»;
 - НЕ средство установления «достоверности»;
 - НЕ способ доказательства ЗНАЧИМОСТИ или ОБЪЕКТИВНОСТИ полученных данных.
- Зачем она нужна?

Законы природы

- Динамические (100%)

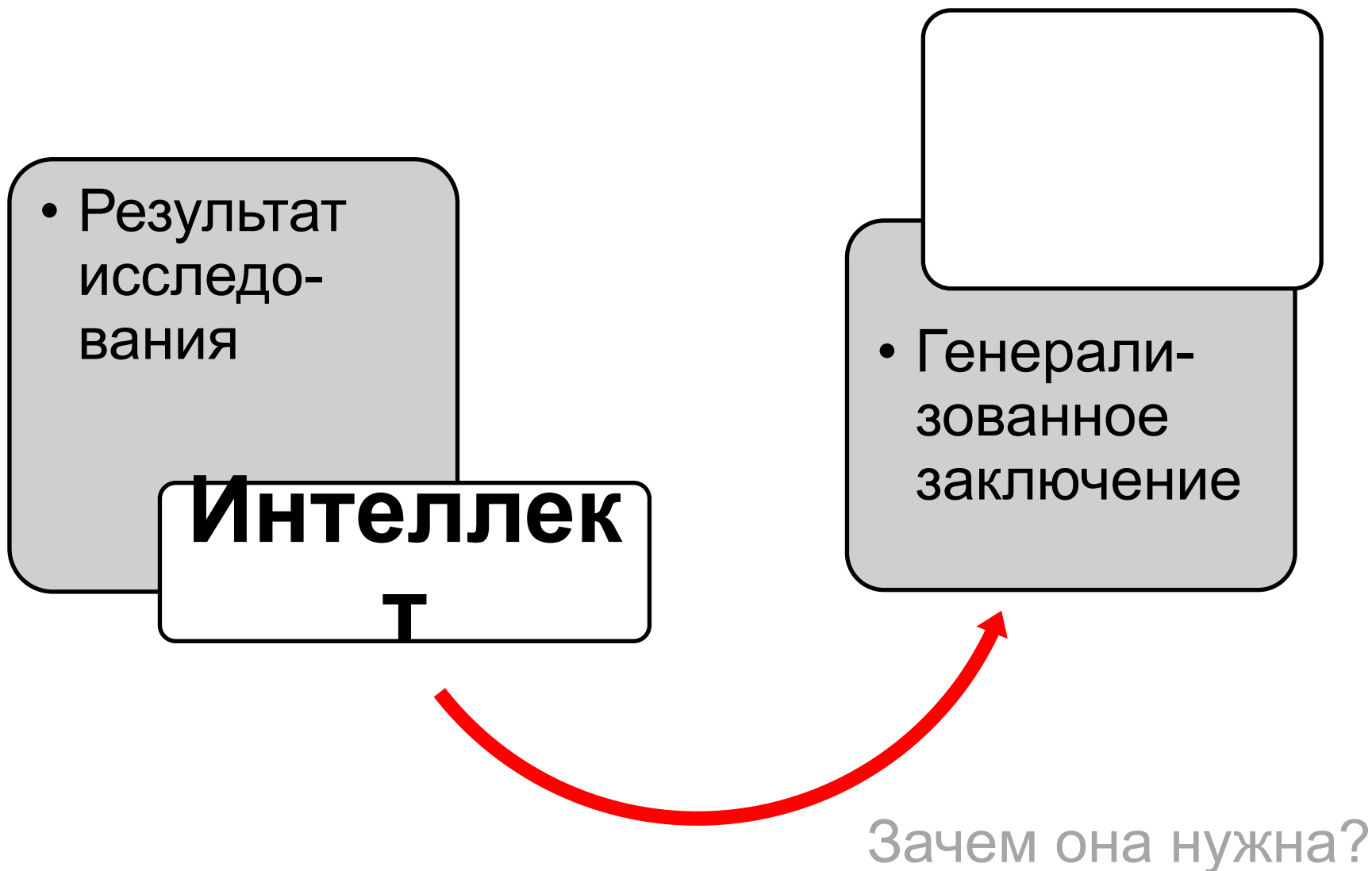
...

- Динамические : Статистические (75:25 %)
- **Динамические : Статистические (50:50 %)**
- *Динамические : Статистические (25:75 %)*

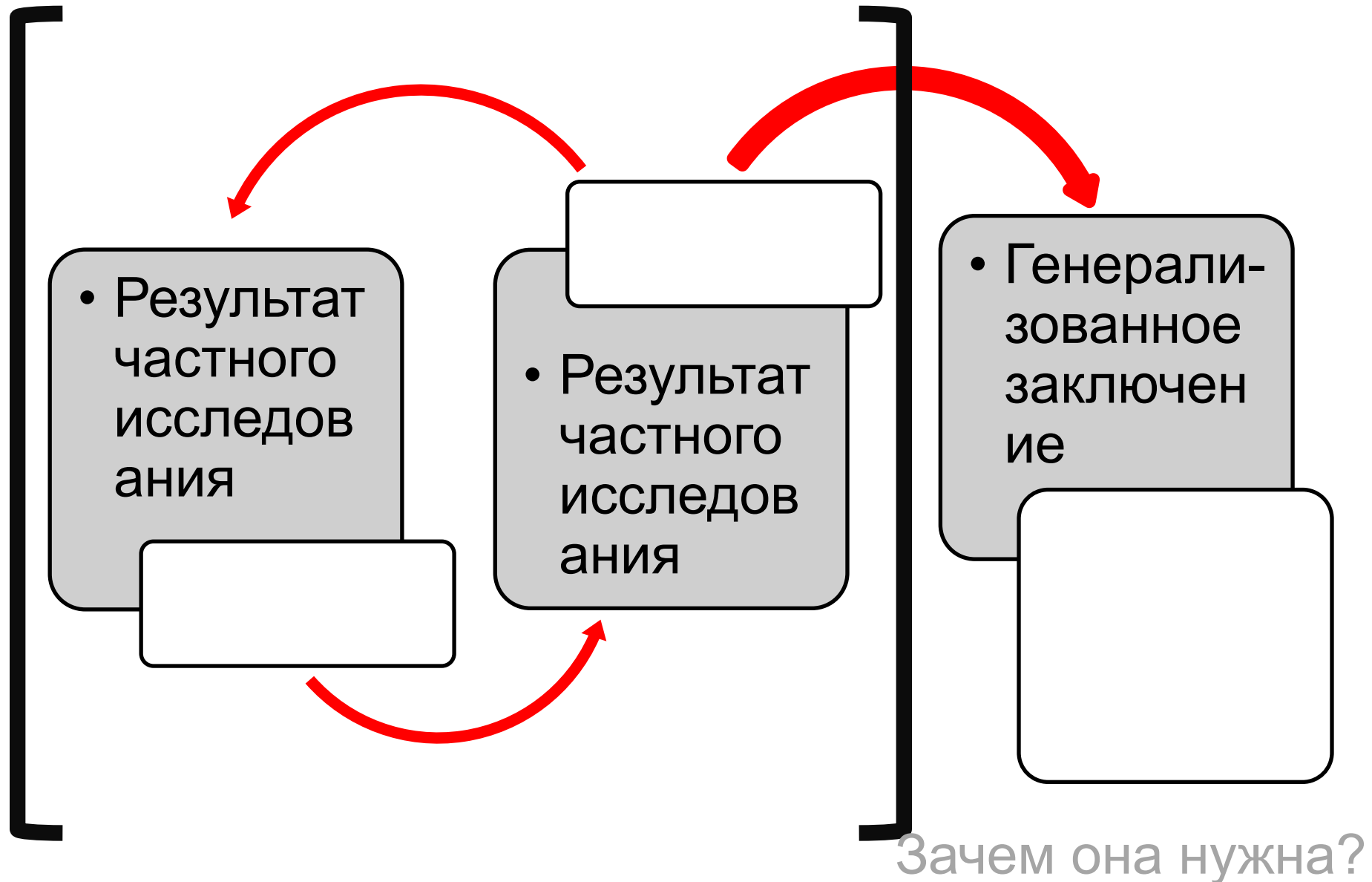
...

- *Статистические (100%) ?* Зачем она нужна?

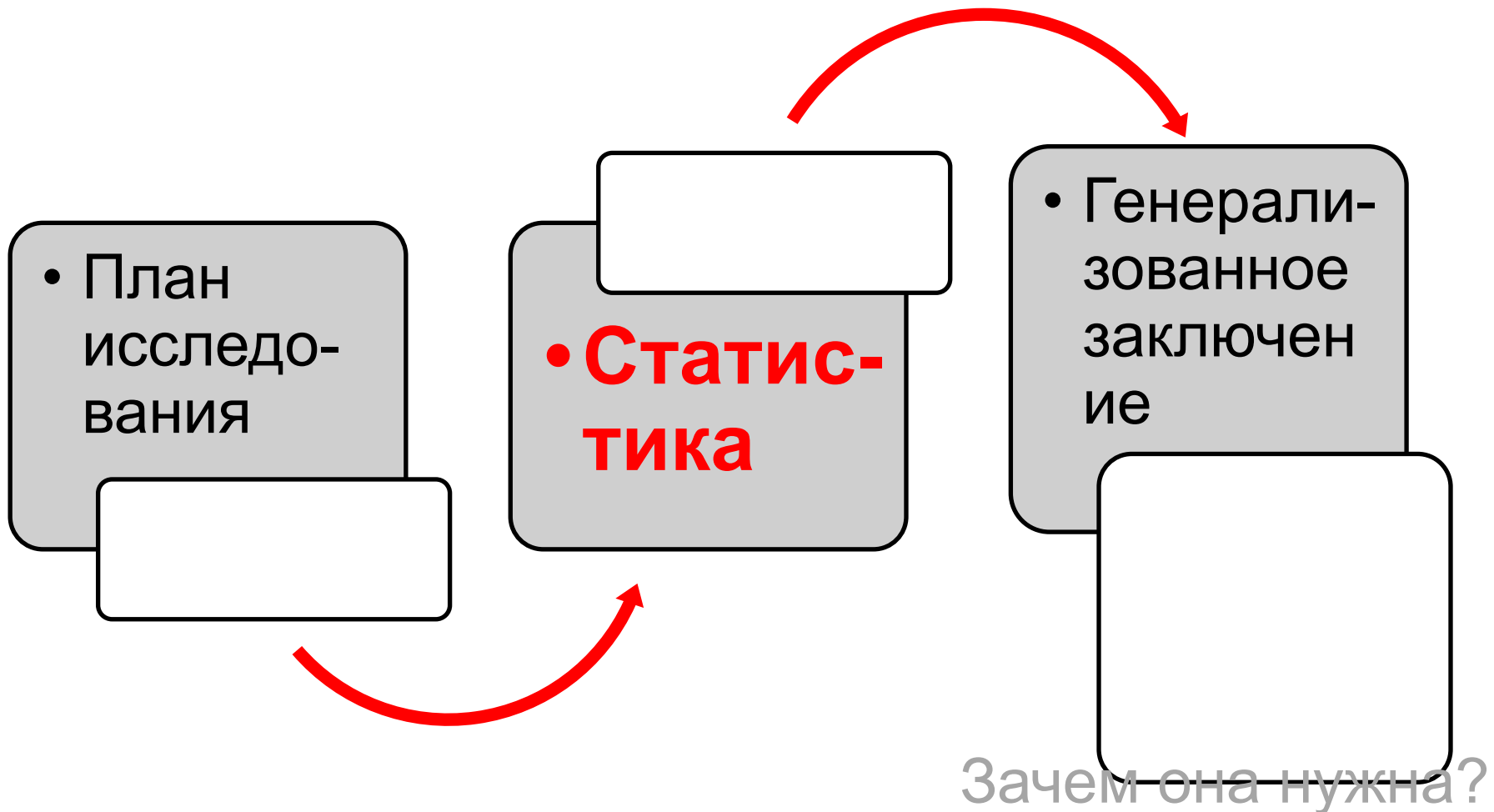
Познание динамических или сильных статистических законов



Традиционный эмпирический путь познания статистических законов



Статистика – инструмент генерализации заключений при познании «слабых» закономерностей



Биологические законы – статистические законы

Общие источники изменчивости систем:

- 1) Ошибки измерения;
- 2) Систематические ошибки;

Специфические источники изменчивости живых систем:

- 3) **Изменчивость живого на всех уровнях организации;**
- 4) Суммирование изменчивости на более высоких уровнях организации.

Зачем она нужна?

Статистика – способ вынесения надежного суждения об общем (о «генеральной совокупности») на основании анализа части («выборки»)

Генеральная совокупность – все реально существующее или воображаемое количество изучаемых объектов:

- генеральная совокупность может быть ограниченной / конечной или практически бесконечной величиной;
- как предмет изучения, генеральная совокупность – идеальная конструкция;

Выборка – реально анализируемое (измеряемое) количество объектов:

- выборка – реальное подмножество;
- выборка – в идеале – случайное подмножество из генеральной совокупности;
- разные выборки из одной генеральной совокупности – разные;
- надежные выборки – большие и случайные; **ненадежные** – маленькие и **неслучайные**.

Зачем она нужна?

Статистика – это:

1. Способ вынесения надежного суждения об общем на основании анализа части этого общего; средство оценки надежности заключений;
2. Стандарт представления результатов в доказательной науке;
3. Средство коммуникации (язык) науки и исследователей;
4. Средство регуляции доверия качеству исследования и уровню исследователя (но это неоднозначно);



Кардинал (Лэнгдон **Статистика** **ГИЯ**
несовершенна. Но только потому, что
человек не совершенен. Любой человек,
включая и этого.

Как это возможно?

1. Создание гипотезы.

**2. Формирование
выборки из генеральной
совокупности.**

**3. Измерения; расчет и
анализ средних.**

**4. Расчет статистик,
проверка гипотезы .**

Гипотезы

Гипотезы – утверждение, предполагающее доказательство.

Научная гипотеза – утверждение, которое потенциально может быть проверено критическим экспериментом.

Статистические гипотезы

Статистическое оценивание, по сути, это проверка нескольких статистических гипотез, одна из которых называется нулевой, а другая, конкурирующая с первой, называется альтернативной.

Гипотезы относятся к ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ!

Первая гипотеза, обычно, предполагает ОТСУТСТВИЕ различий или эффекта или связи. Например, для случая сравнения средних эти гипотезы принято записывать следующим образом:

$$H_0: \mu_1 = \mu_2 \text{ («нулевая», т.к. } \mu_1 - \mu_2 = 0)$$

$$H_1: \mu_1 \neq \mu_2$$

Критичный этап: определение или формирование генеральной совокупности и взятие выборки

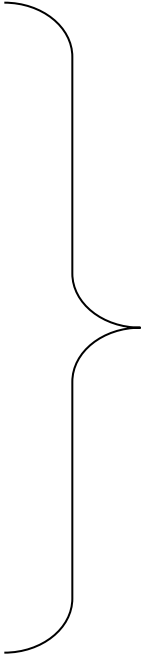
- Никакая статистика не компенсирует ошибки, совершенные на этапе планирования или осуществления наблюдения и экспериментов;
- Статистику как инструмент планирования надо использовать уже на этом этапе;

Проверка статистических гипотез – это расчет и оценка критериев (статистик)

Критерии рассчитываются на основании
ВЫБОРКИ!

Для разных типов данных и для разных
типов задач критерии **РАЗНЫЕ**.

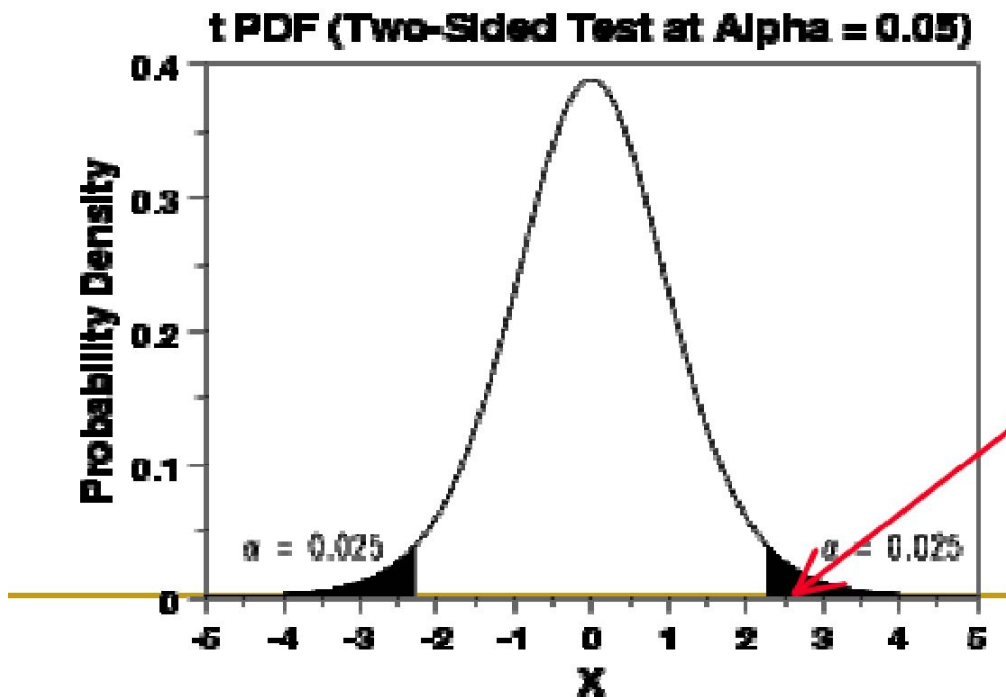
В зависимости от типа критерия и объема
наблюдений (размера выборки)
определяется **ЗНАЧИМОСТЬ** критерия,
на основании которого принимаются или
отвергаются предварительные гипотезы.



Знание условий
использования
критериев, умение
их рассчитывать и
интерпретировать
является умением
«выполнять
статистический
анализ»

Суть статистических критериев

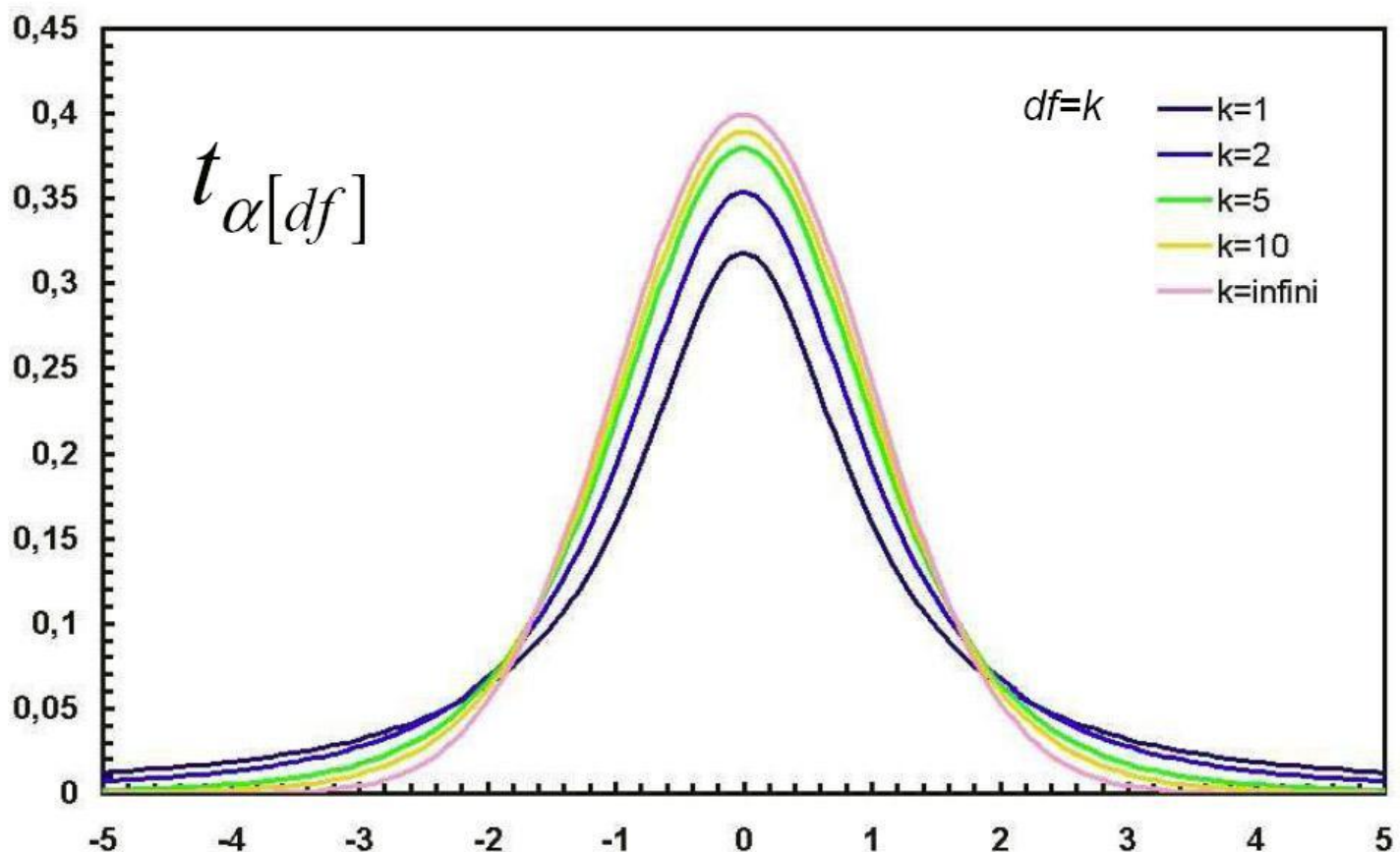
Статистическое критерий – некоторая теоретическая функция (распределение), которая используется для описания анализируемых фактических данных и при заданном числе наблюдений (числе степеней свободы) получает СТРОГОЕ



Это число сравнивается с эталонным (табличным) и на основании этого сравнения делается вывод о значимости/существенности/надежности искомого/оцениваемого эффекта.

Оценка надежности критерия ЖЕСТКО связана с количеством наблюдений

***t*-распределение (Стьюдента)**



Возможные исходы при проверке статистических гипотез

		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)

Возможные исходы при проверке статистических гипотез

		Принята гипотеза	
		H_0	H_1
Верна гипотеза	H_0		Ошибка 1 рода: вероятность найти различия, где их нет. Это – нездоровые сенсации, которые могут принести большой вред. Вероятность ошибки первого рода – это уровень значимости (α или P).
	H_1	Ошибка 2 рода: вероятность не увидеть различий, где они есть. Это «близорукость», или «слепота» критерия, вред от неё не очень большой.	

Условия принятия / отвержения гипотез $P > 0,05$ и $P < 0,05$ – в чем разница и что это значит?

Достигнутый уровень значимости (P) – это вероятность получить такое же (или более экстремальное) значение критерия в длинной серии повторных выборок при условии справедливости H_0 .

Проще: " P " – это вероятность ошибочно отвергнуть нулевую гипотезу при отсутствии различий.

Еще проще: " P " – это вероятность справедливости нулевой гипотезы при условии ее отвержения.

Совсем просто, но не совсем корректно: " P " – это вероятность найти различия или другой эффект при их реальном отсутствии.

Граничные условия P – ВНЕ ЛОГИКИ, просто результат договора (заговора?) специалистов: $P \leq 0,05$; $P \leq 0,01$; $P \leq 0,001$.

Свойства «идеального» статистического анализа

- Большие, случайно сформированные выборки;
- Ограниченное число независимых друг от друга или, наоборот, значительно коррелирующих между собой признаков биологических объектов;
- Ограниченное число ярко выраженных и независимых «факторов», в соответствие с которыми варьирует строение биологических объектов;
- Нормальное распределение количественных признаков;
- Сильные (разницы более $0,5 - 1,0 \sigma$) значимые ($P < 0,001$) эффекты и связи, легко интерпретируемые биологически.

Статистика – это язык: необходимый минимум описания результатов статистического анализа

Пример: как описать результаты применения критерия МАННА-УИТНИ при сравнении средних?

$Me_1=12,5; Me_2=14,0; U_{(n1=12; n2=11)}=41,5; P=0,1316.$

Характеристика центральных тенденций распределения значений признаков; здесь д.б. медианы

Вид критерия

Описание объема наблюдений: размер выборки или число степеней свободы (в зависимости от критерия)

Полученное (рассчитанное) значение критерия

Достигнутый уровень значимости критерия

ОСНОВНЫЕ ПОНЯТИЯ

Типы статистических задач

Задачи	Инструменты
Описание совокупностей объектов	Анализ одной выборки; расчет параметров распределений (положения, формы); проверка нормальности распределений; построение доверительных интервалов
Сравнение параметров	Парные и множественные сравнения средних; сравнение распределений; сравнение частот; t-критерий; тест Манна-Уитни или Краскела-Уоллеса; дисперсионный анализ;
Анализ зависимостей	Установление взаимосвязи между двумя переменными или между многими переменными; установление силы влияния одной или многих переменных на одну результирующую; корреляционный анализ, парная и множественная регрессия, логит-регрессия;
Снижение размерности, ординация, классификация	Кластерный, факторный, дискриминантный анализ; анализ соответствий; многомерное шкалирование и др.

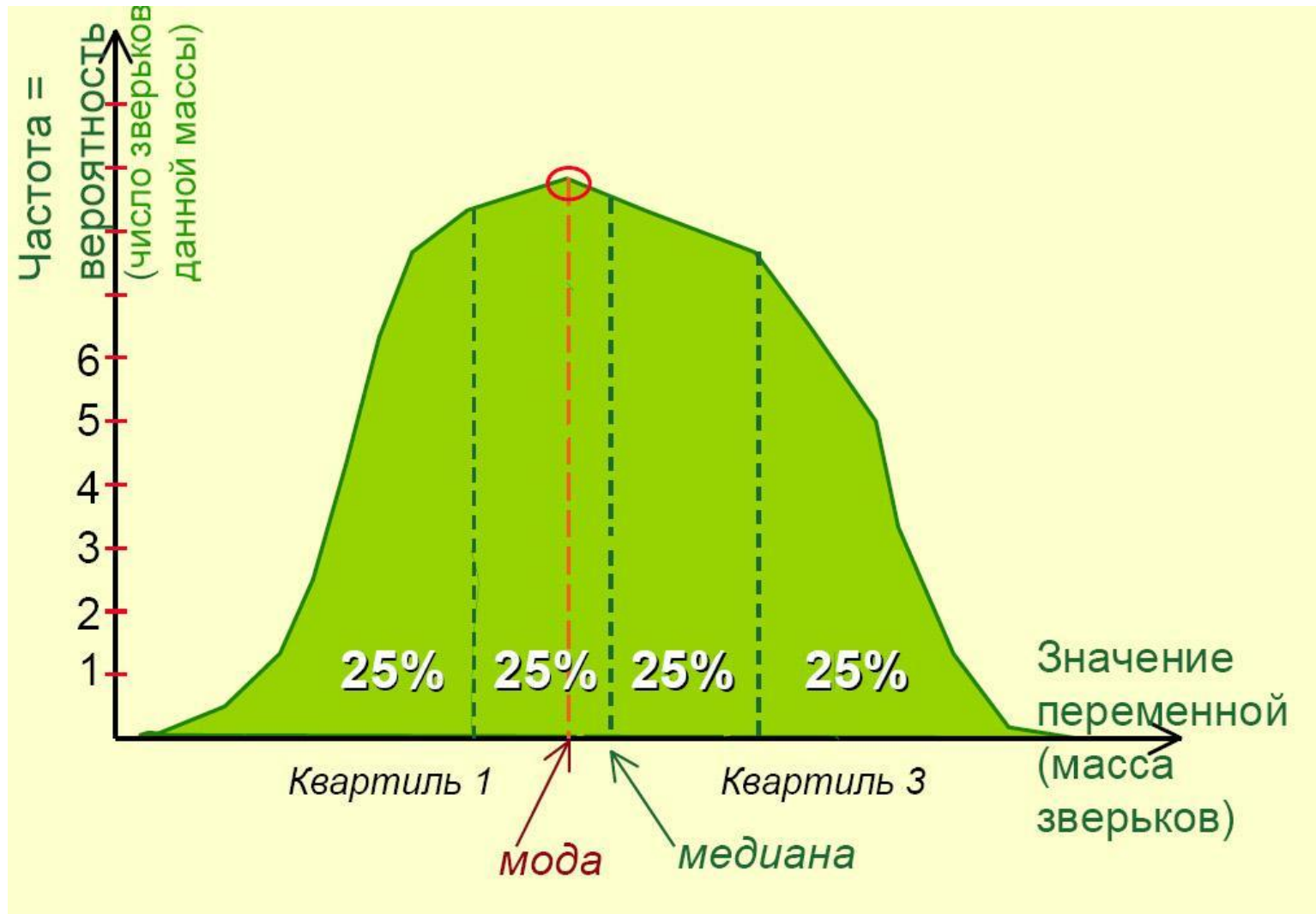
Как выполнить «статистический анализ»?

1. Определить задачу (в т.ч. в статистических терминах);
2. Провести исследование; собрать данные; выполнить измерения;
- 3(1) Сформулировать нулевую и альтернативную статистические гипотезы;
- 4(2) Выбрать адекватный критерий;
- 5(3) Выполнить расчеты и принять одну из статистических гипотез;
- 6(4) Опубликовать результаты применения статистики и интерпретировать статистические вводы в биологических терминах.

Выбор статистического теста при сравнении распределений (сравнении центральных тенденций)

Задача	Количественная шкала, нормальное распределение	Порядковая шкала или отклонение от нормального распределения	Номинальная шкала
Сравнить одну группу с гипотетическим значением	t-тест Стьюдента для одной выборки	Тест Вилкоксона	Тест хи-квадрат
Сравнить две не связанные совокупности	t-тест Стьюдента для не связанных совокупностей	Тест Манна-Уитни	Тест Фишера (тест хи-квадрат)
Сравнить две связанные совокупности	t-тест Стьюдента для связанных совокупностей	Тест Вилкоксона	Тест Мак-Неймера
Сравнить более двух не связанных совокупностей	Однофакторный дисперсионный анализ	Тест Краскела-Уоллиса	Тест хи-квадрат
Сравнить более двух связанных совокупностей	Дисперсионный анализ с повторными измерениями	Тест Фридмана	Тест Кохрана

Основные центральные тенденции распределений



Как правильно выбрать критерий?

Типы биологических данных. Шкалы

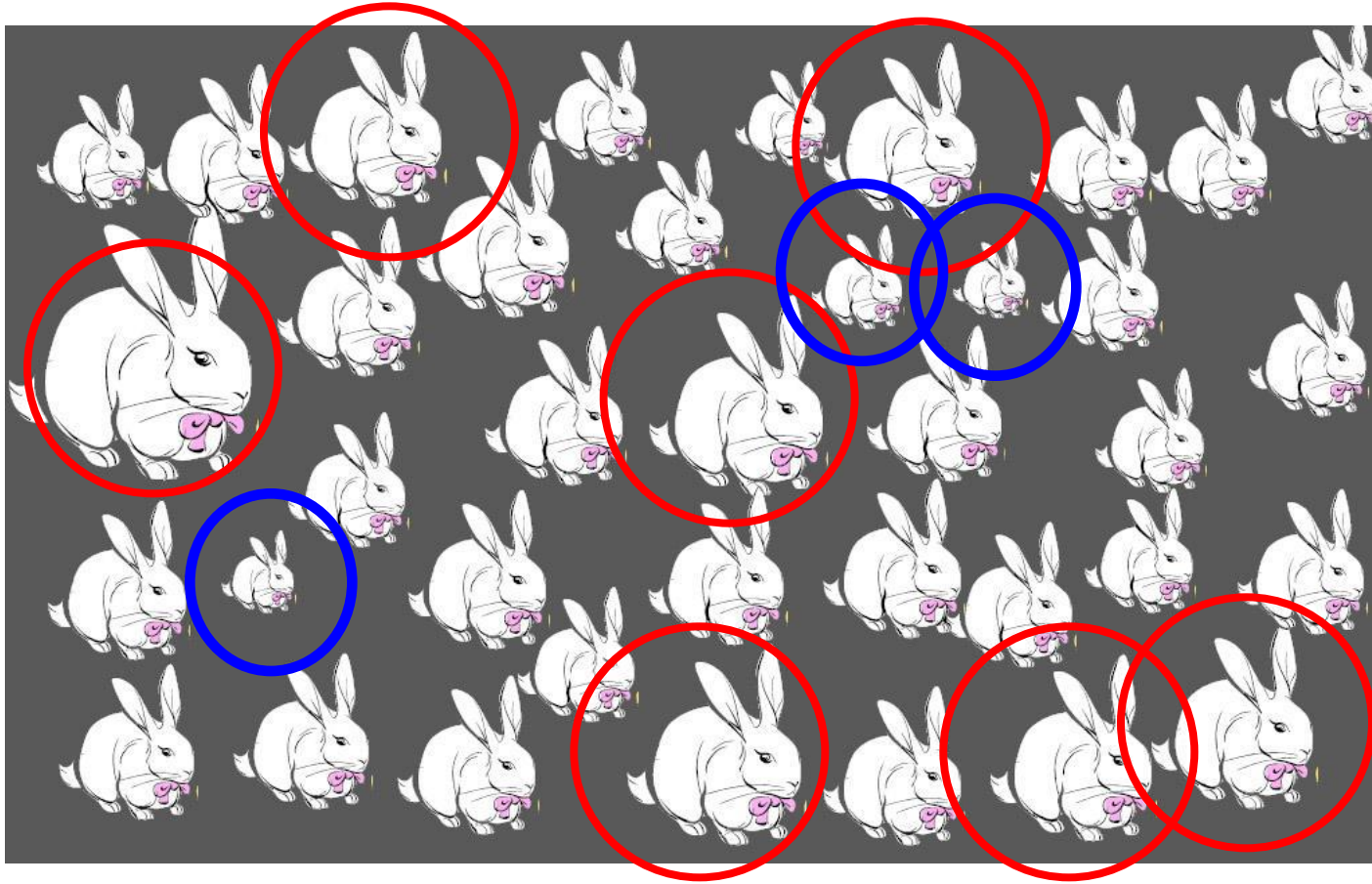
Шкала	Свойства	Пример	Характеристики центральных тенденций
Наименований	Используется при классификации объектов	Вид организма; пол	Мода
Порядковая	Используется при упорядочивании объектов, но не оценивать интервалы	мало – много – очень много; неокрашенный – средняя окраска – меланист;	Мода; <i>медиана</i>
Интервальная	Используется для измерения разницы между объектами	Температура в градусах Цельсия или Фаренгейта	Мода, медиана, среднее (арифметическое)
Отношений	Используется для измерения в категориях (единицы измерения), а ноль пункт задан естественно	Рост в см, масса в килограммах; число организмов или их частей в штуках	Мода, медиана, среднее (арифметическое)

Качественные

Количественные

Как правильно выбрать критерий?

Типы биологических данных. Шкалы



Номинальная:
самец или
самка;
родитель или
потомок.

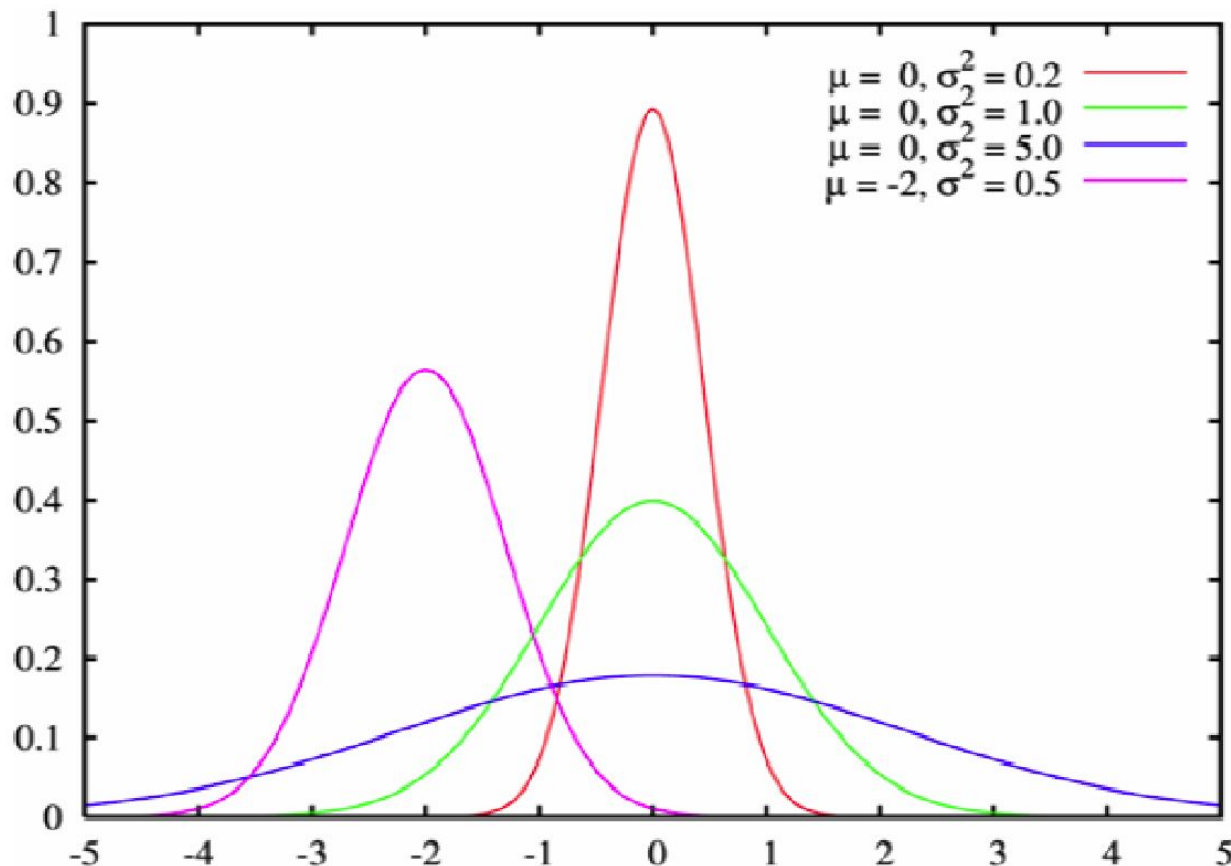
Ординальная:
крупный,
средний или
мелкий
кролик;
зараженный
или
незараженный
гельминтами.

Интервальная шкала : температура тела; масса кролика,
выраженная в единицах массы новорожденного кролика.

Шкала отношений: масса в граммах; длина уха в см; количество
волосков в вибриссах.

Основные понятия

Нормальное распределение (Гауссово)



Нормальное
распределение;

1) унимодальное;

2) симметричное.

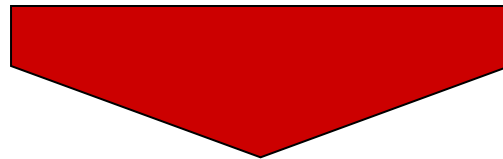
Описывает случайно
изменяющуюся
непрерывно
варьирующую
величину,
измеренную в
номинальной шкале
или шкале
отношений.

Нормальное распределение (Гауссово)



Правильный выбор статистического критерия зависит от:

1. Задачи;
2. Шкалы, в которой измерены данные;
3. Соответствия/несоответствия количественных данных (т.е. измеренных в шкале отношений или шкале интервалов) нормальному распределению.



Проверка «нормальности» – первый и обязательный этап анализа количественных данных!



**Соответствие «нормальному»
распределению:
параметрические статистики**



**Не соответствие «нормальному»
распределению:
непараметрические статистики**

Выбор статистического теста при сравнении распределений (сравнении центральных тенденций)

Задача	Количественная шкала, нормальное распределение	Порядковая шкала или отклонение от нормального распределения	Номинальная шкала
Сравнить одну группу с гипотетическим значением	t-тест Стьюдента для одной выборки	Тест Вилкоксона	Тест хи-квадрат
Сравнить две не связанные совокупности	t-тест Стьюдента для не связанных совокупностей	Тест Манна-Уитни	Тест Фишера (тест хи-квадрат)
Сравнить две связанные совокупности	t-тест Стьюдента для связанных совокупностей	Тест Вилкоксона	Тест Мак-Неймера
Сравнить более двух не связанных совокупностей	Однофакторный дисперсионный анализ	Тест Краскела-Уоллиса	Тест хи-квадрат
Сравнить более двух связанных совокупностей	Дисперсионный анализ с повторными	Тест Фридмана	Тест Кохрана

**Учебники –
вещь сугубо
полезная**

**А. Афифи,
С. Эйзен**

**Статистический
анализ**

**Подход
с использованием
ЭВМ**

Перевод с английского
И. С. Енюкова и
И. Д. Новикова
под редакцией
Г. П. Башарина

Москва «Мир» 1982

Почему «Statistica» «лучшая»?

(из SPSS, SAS, Statistica, NCSS97, S-Plus, STATA/StatTransfer, SYSTAT, MINITAB, STATGRAPHICS+)

- Большой набор тестов.
- «Интуитивный» кнопочный интерфейс с разворачивающимися подменю и взаимодействием «компьютер – пользователь» по типу «вопрос –ответ».
- Высококачественная графика с автоматическим предложением адекватных данным типов иллюстраций.
- Модульный принцип организации меню.
- Развитая система подсказки(!).
- Достаточно развиты возможности экспорта и импорта данных.

«Таким образом, Statistica является одной из наиболее простых для неподготовленного пользователя систем, с наименьшим периодом овладения ее возможностями и удачным набор графических возможностей».

К недостаткам системы можно отнести ее малую расширяемость, отсутствие модулей третьих фирм и пользователей, а также недостаточно эффективный командный язык»

STATISTICA™

Том I: ОСНОВНЫЕ СОГЛАШЕНИЯ И СТАТИСТИКИ I

1. Основные соглашения	1001
2. Панели инструментов и Строка состояния	1049
3. Окно таблицы с исходными данными	1123
4. Окно таблицы результатов Scrollsheet	1245
5. Окно текста/вывода	1273
6. Кнопки автозадач	1293
7. Управление данными	1315
8. Основные статистические понятия	1425
9. Основные статистики и таблицы	1439
10. Быстрые основные статистики	1573
11. Непараметрическая статистика и распределения	1601
12. Множественная регрессия	1653
13. Общая ANOVA/MANOVA	1709
Приложения	1827
Литература	1897

Том I: СОГЛАШЕНИЯ И СТАТИСТИКИ I
Том II: ГРАФИКА
Том III: СТАТИСТИКИ II
Том IV: ПРОМЫШЛЕННЫЕ СТАТИСТИКИ
Том V: ЯЗЫКИ: BASIC и SCL



StatSoft®

**Вместе с
официальным
и копиями
программы
поставляется
ОЧЕНЬ
ПРИЛИЧНОЕ
руководство**

STATISTICA™

Том III: СТАТИСТИКИ II

1. Нелинейное оценивание	3001
2. Анализ дискриминантных функций	3065
3. Надежность и позиционный анализ.....	3111
4. Каноническая корреляция	3137
5. Кластерный анализ	3165
6. Факторный анализ	3197
7. Многомерное шкалирование.....	3239
8. Анализ временных рядов.....	3261
9. Лог-линейный анализ	3441
10. Анализ выживаемости	3473
11. Моделирование структурными уравнениями	3535
12. Менеджер мегафайлов	3699
Литература.....	3757

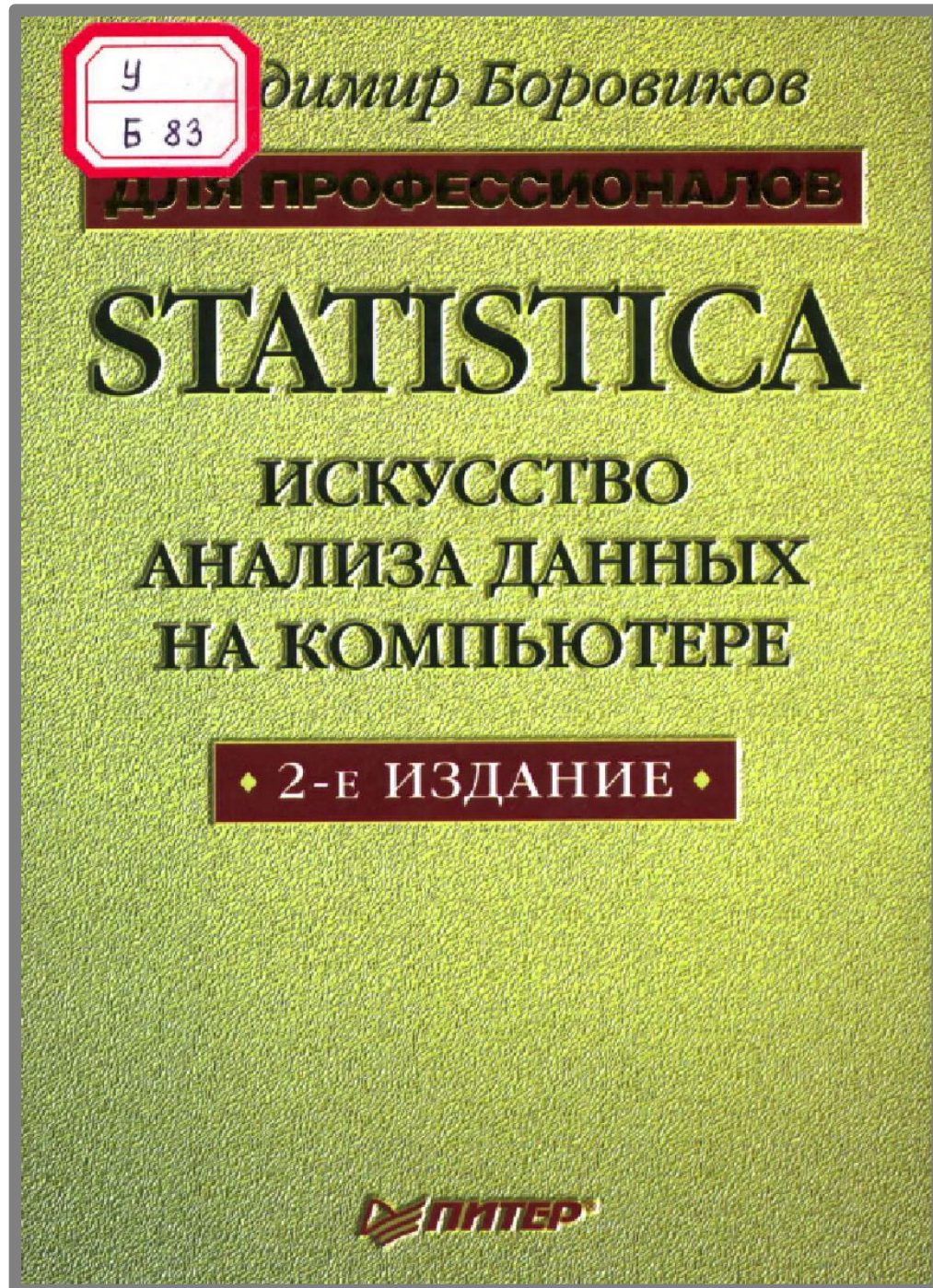
Том I:	СОГЛАШЕНИЯ И СТАТИСТИКИ I
Том II:	ГРАФИКА
Том III:	СТАТИСТИКИ II
Том IV:	ПРОМЫШЛЕННЫЕ СТАТИСТИКИ
Том V:	ЯЗЫКИ: BASIC и SCL



StatSoft®

Вместе с
официальны
ми копиями
программы
поставляетс
я **ОЧЕНЬ**
ПРИЛИЧНОЕ
руководство

**Не жалеете
время на
изучение
учебников**



А.А. Халафян

— Учебник —

STATISTICA 6

Статистический
анализ
данных



**Не жалеете
время на
изучение
учебников**