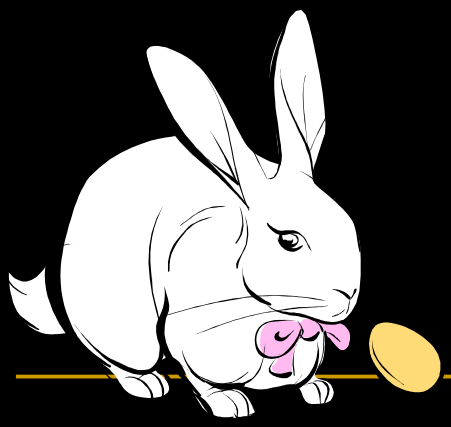
Данные, с которыми мы работаем. Основные понятия.

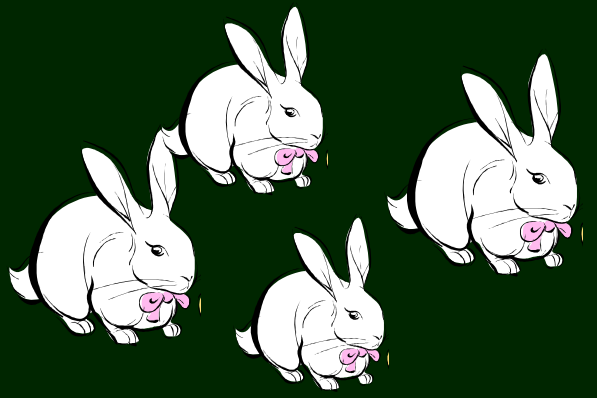
Описательная статистика.



Занятие 1

наблюдения

выборка

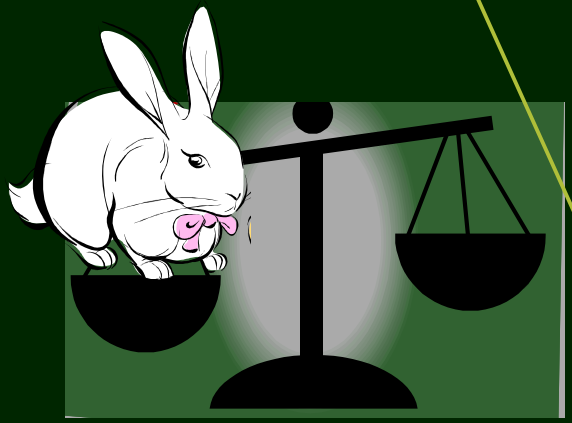


Переменные

Количественные

дискретные

непрерывные



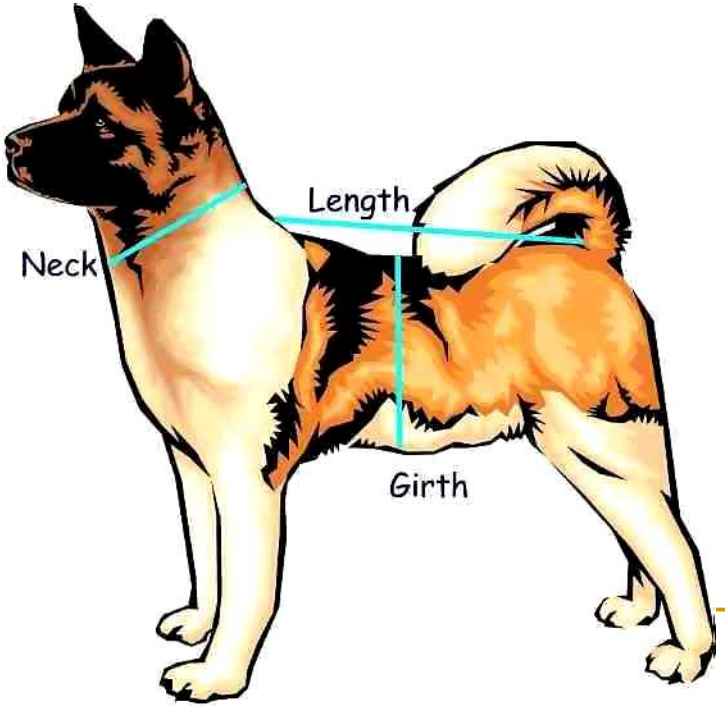
Ранговые (качественные, но могут быть выстроены в последовательность)

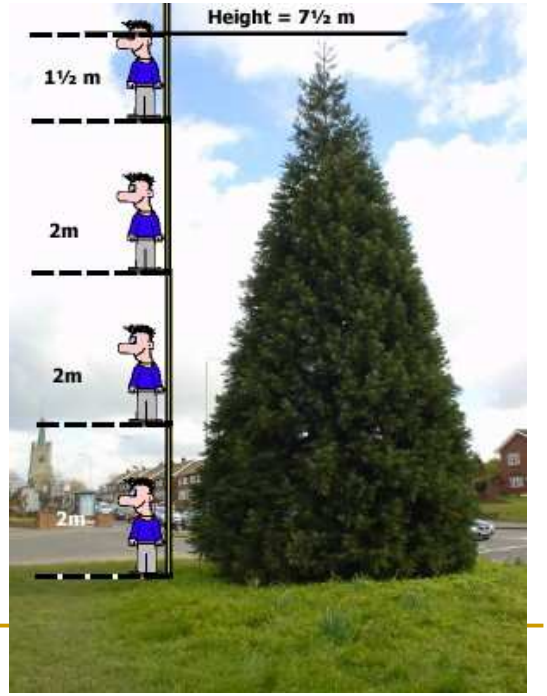
Качественные

descriptive statistics

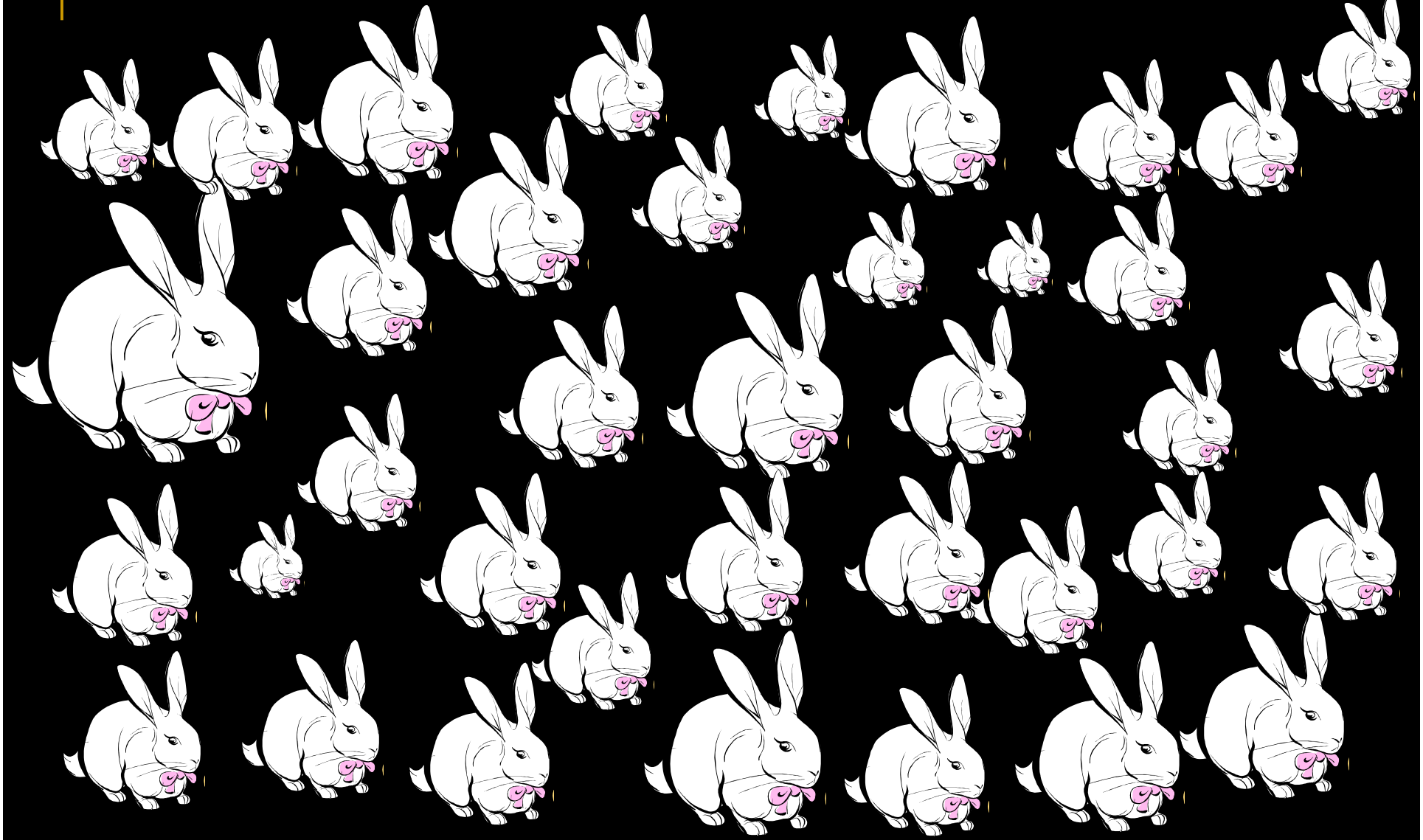
Непрерывные переменные:

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).



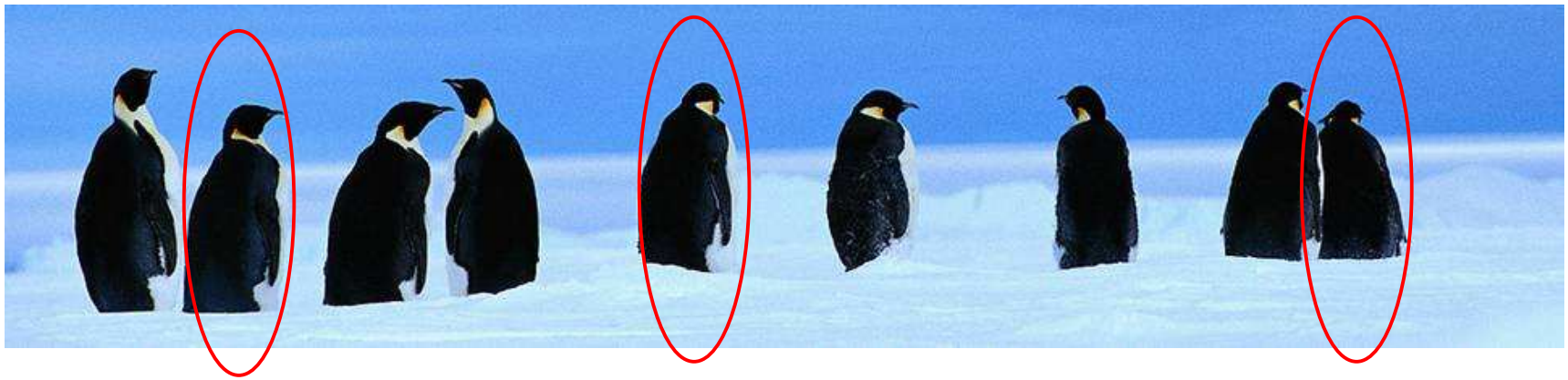


Взвешиваем ***N*** кроликов



Как описать выборку?

ВЫБОРКА ИЗ ПОПУЛЯЦИИ ДОЛЖНА БЫТЬ СЛУЧАЙНОЙ



параметры	Для популяции	Для выборки
Среднее (mean)	$\mu = \frac{\Sigma X}{N}$	$\overline{X} = \frac{\sum_i X_i}{n}$
Дисперсия (variance)	$\sigma^2 = \frac{\sum_i (x_i - \mu)^2}{n}$	$s^2 = \frac{\sum_i (X_i - \overline{X})^2}{n-1}$
Стандартное отклонение (Standard deviation)	$\sigma = \sqrt{\frac{\sum_i (x_i - \mu)^2}{n}}$	$s = \sqrt{\frac{\sum_i (X_i - \overline{X})^2}{n-1}}$

descriptive statistics

Стандартная ошибка используется для характеристики точности параметров для выборки

Стандартная ошибка среднего
Standard error SE ≠ standard deviation SD!!!

Это стандартное отклонение в выборке, полученной из средних значений нескольких выборок объёма n из одной популяции

$$SE = s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

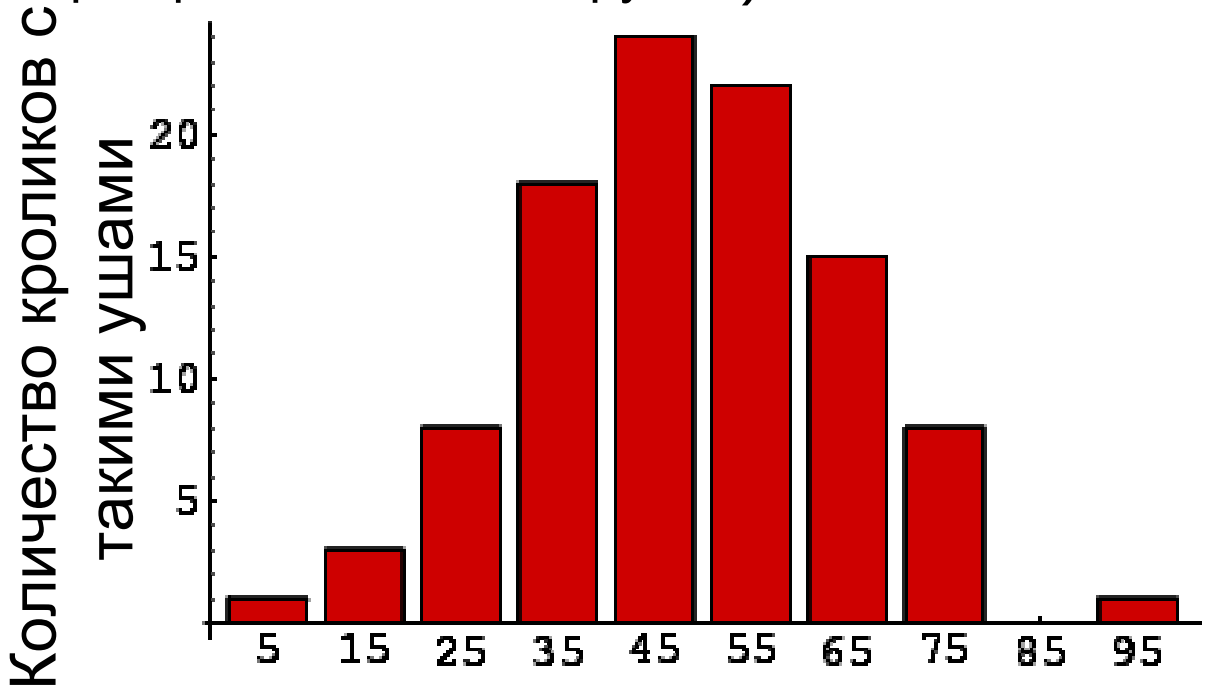
Коэффициент вариации (Coefficient of variation)

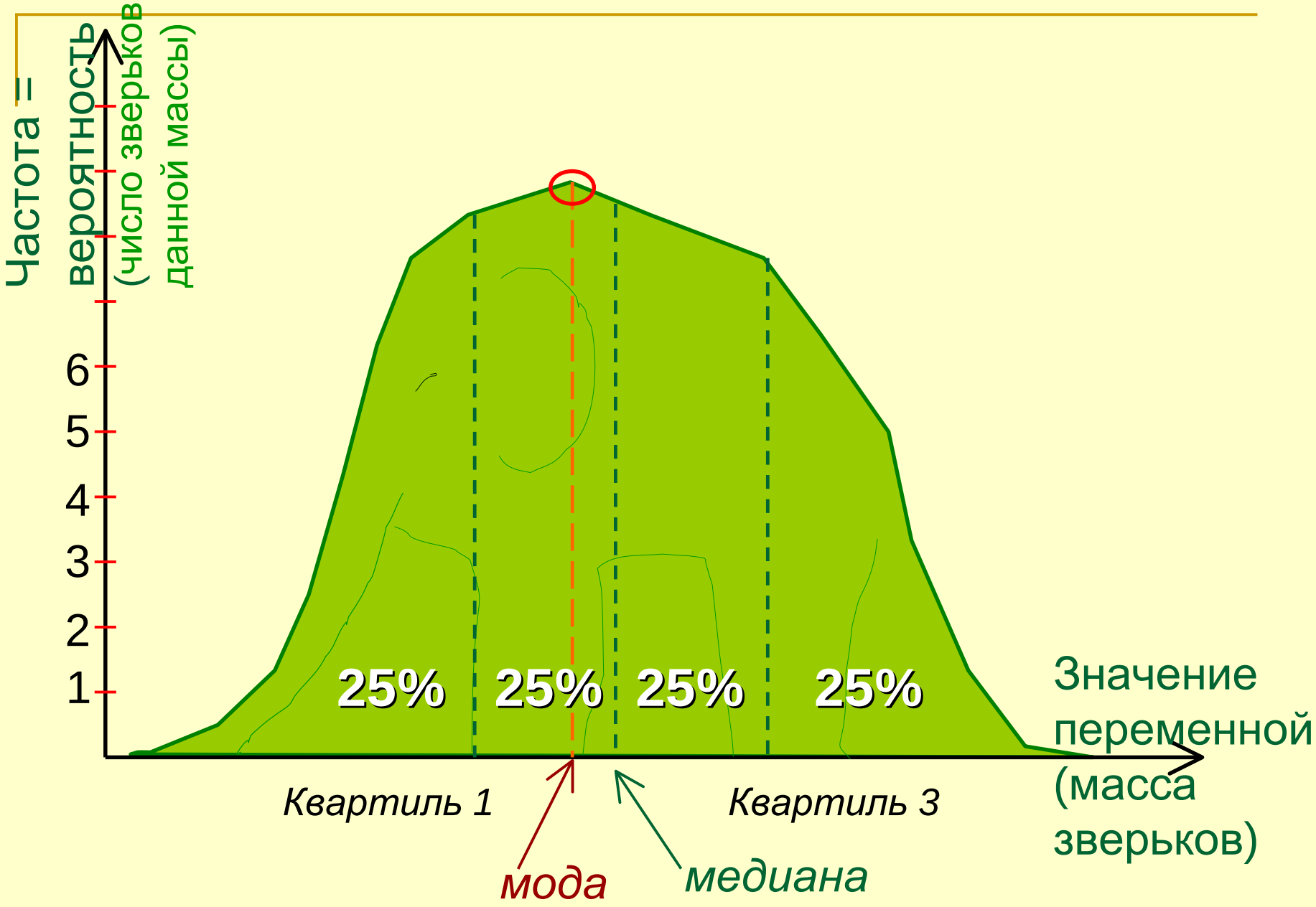
$$CV = \frac{s \cdot 100}{\bar{X}}$$

Частотные распределения

Частотное распределение (frequency distribution) – это соответствие между значениями нашей переменной и их вероятностями

1. *Если возможно*, упорядочим по возрастанию значения переменной (выстроим кроликов от меньшего к большему);
2. Напротив значений напишем, сколько раз они встретились в выборке (для непрерывной переменной сперва разобьём их на группы).




Частота = вероятность (число зверьков данной массы)
6
5
4
3
2
1
25%
25%
25%
25%
Значение переменной (масса зверьков)
Квартиль 1
Квартиль 3
мода
медиана

Квантиль порядка *p*

Медиана (median)– значение, которое делит распределение пополам: половина значений больше медианы, половина – не больше

Квартили (quartiles) делят распределение на четыре части так, что в каждой из них оказывается поровну значений (2-я квартиль = медиана)

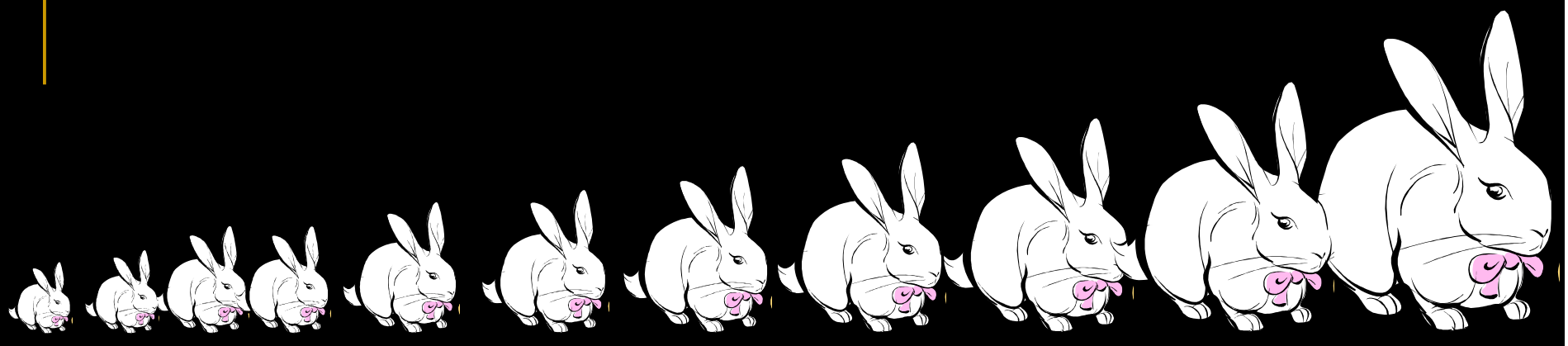
$$\mathrm{M} = X_{(n+1)/2} \qquad Q_1 = X_{(n+1)/4}$$

Это номер измерения, которое будет медианой

Интерквартильный размах – разница между третьей и первой квартилями

Мода (mode) – наиболее часто встречающееся значение [локальный максимум функции плотности], она может быть не одна – би- и мультимодальные распределения. (для качественных переменных тоже существует)

Размах (range) – разность между максимальным и минимальным значениями $= X_n - X_1$



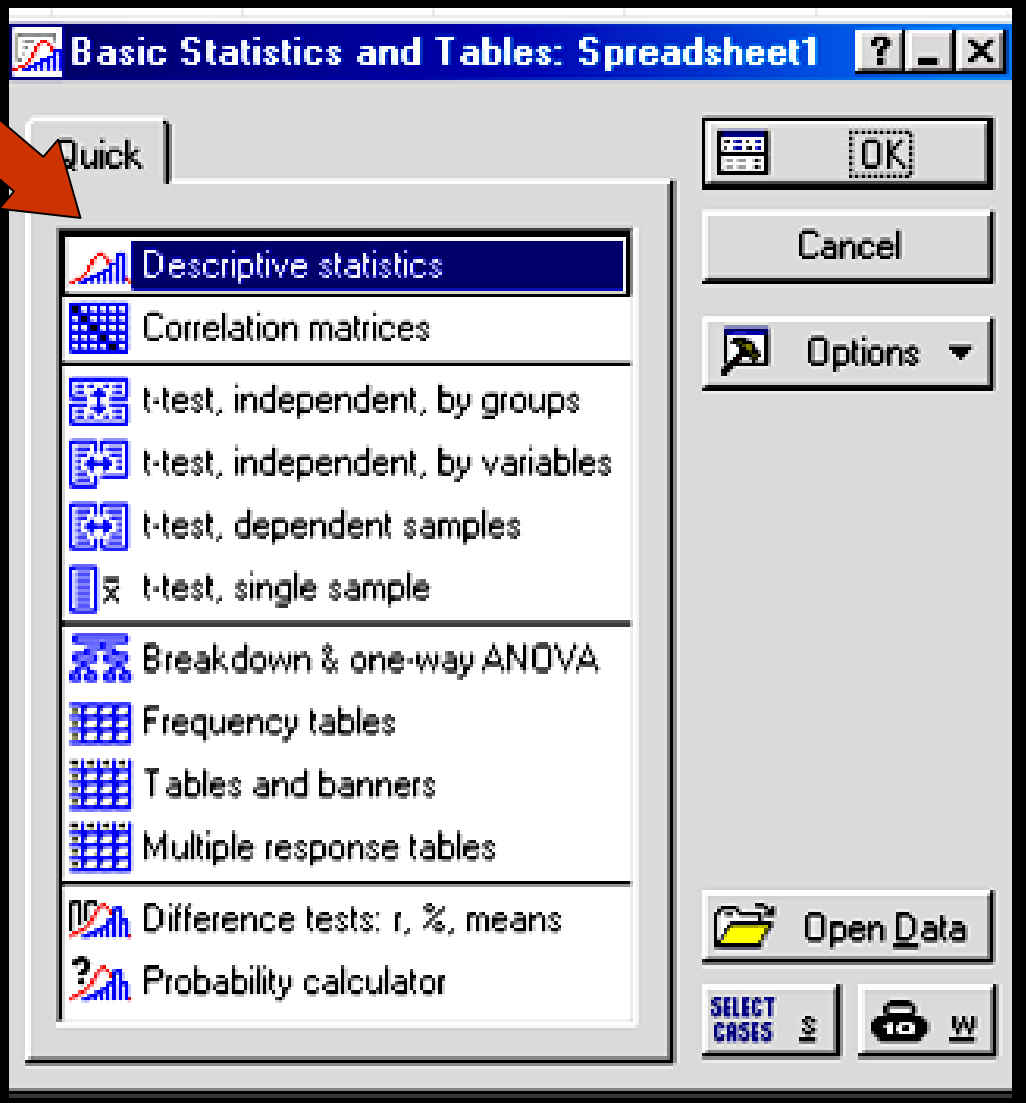
1 2 3 4 5 6 7 8 9 10 11

Номера кроликов

Медиана = (11+1)/2 = 6

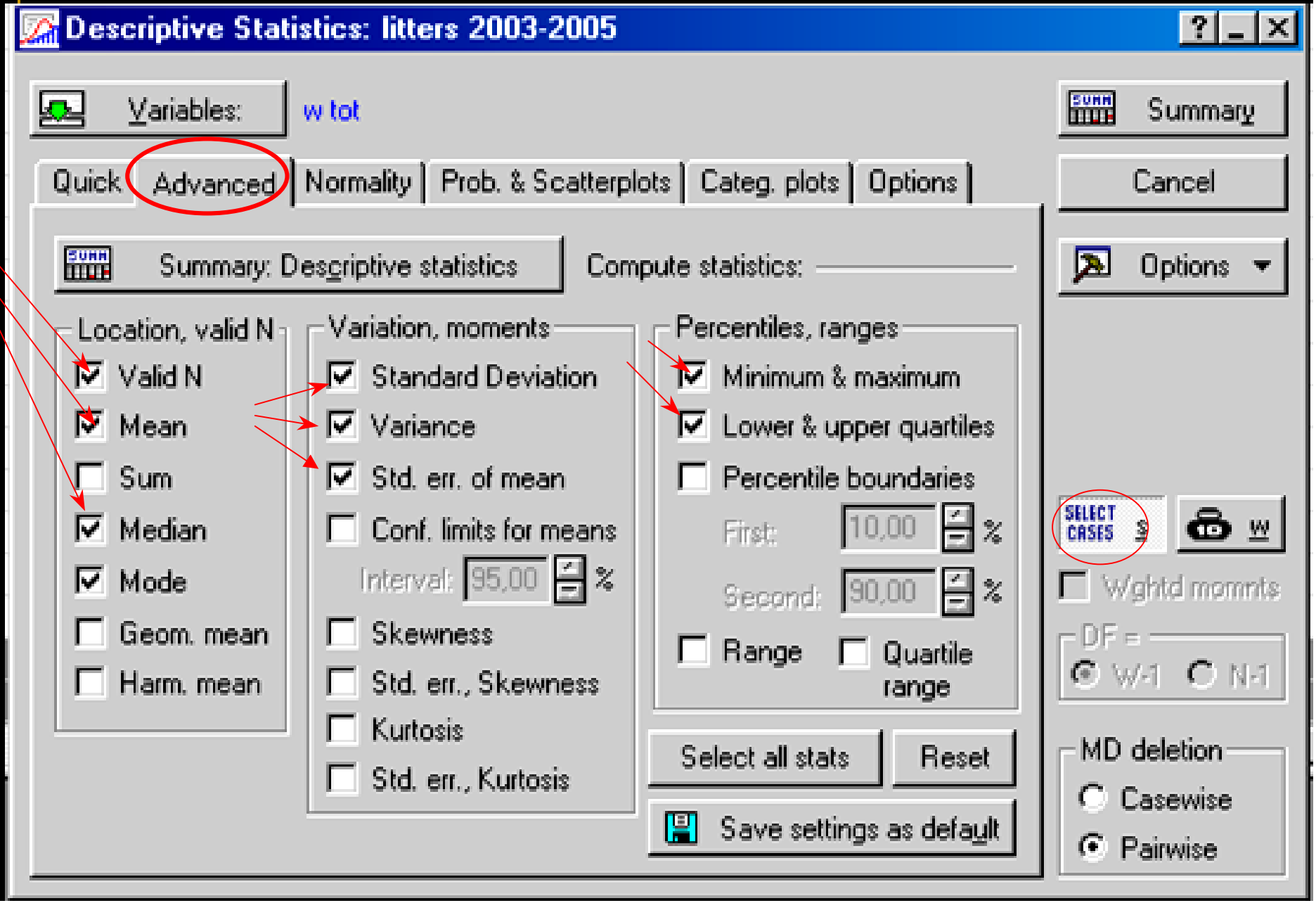
1-я квартиль = (11+1)/4 = 3

Как посчитать параметры распределения

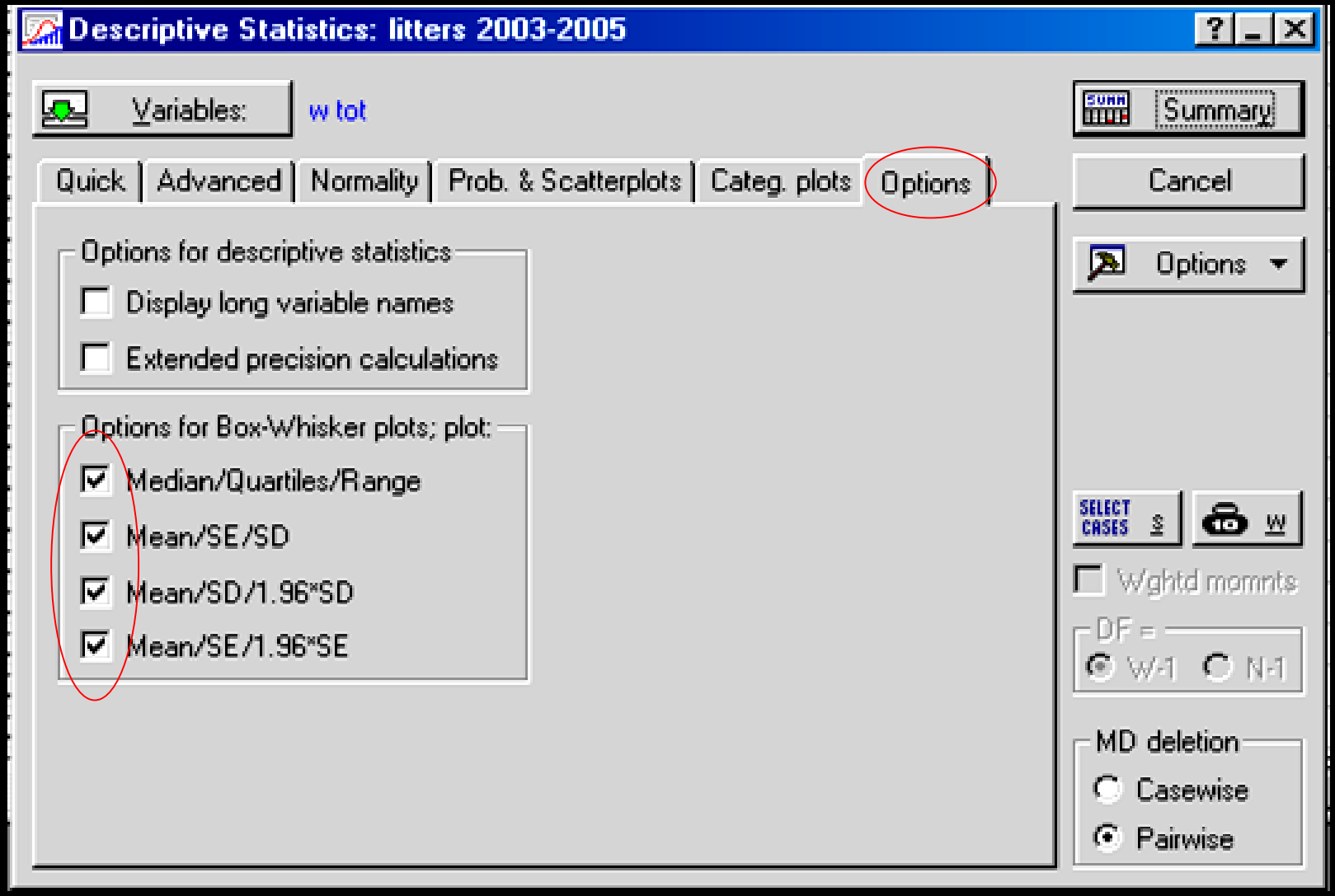


descriptive statistics

Описательная статистика

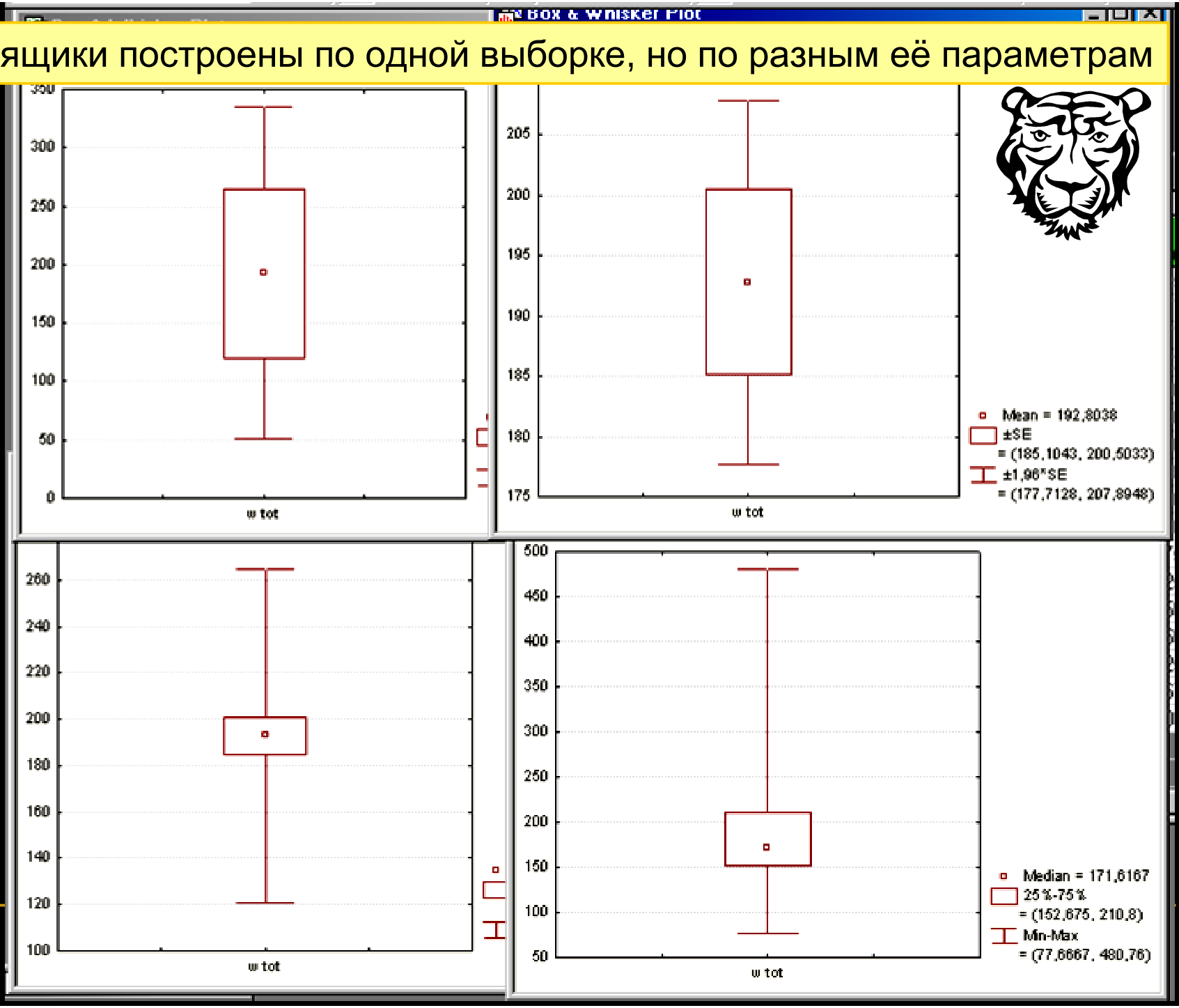


Усатые ящики



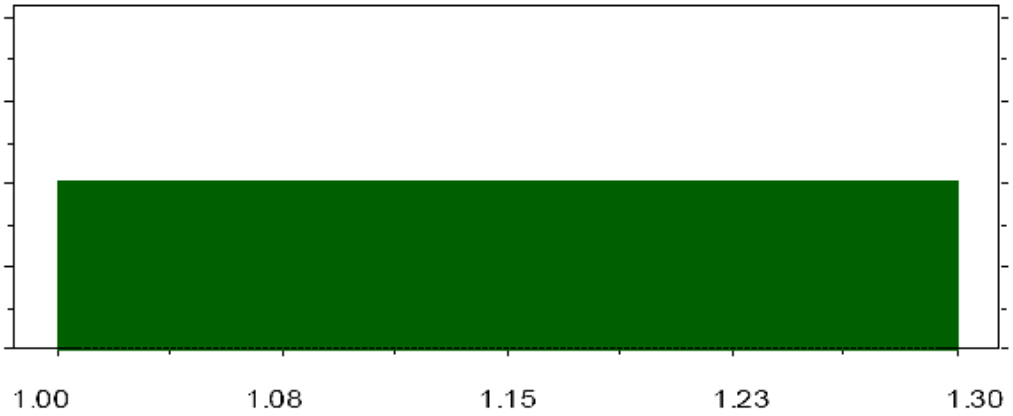
descriptive statistics

Эти ящики построены по одной выборке, но по разным её параметрам

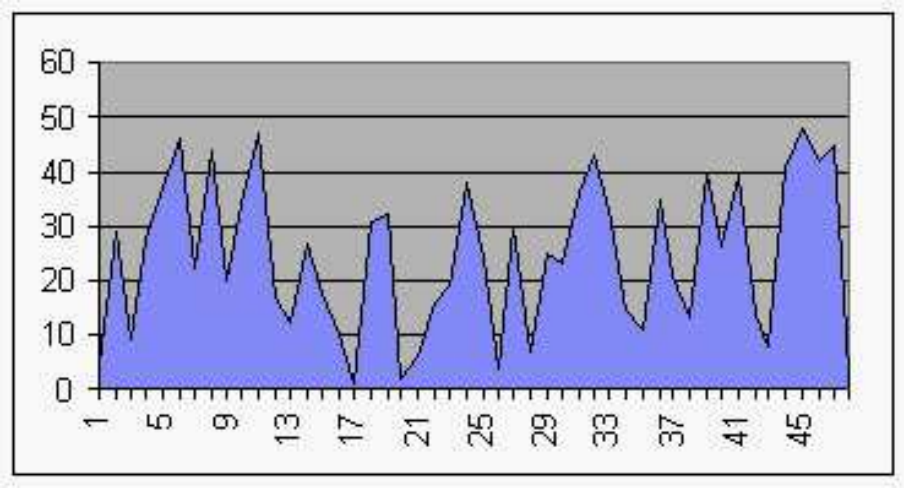


Какие бывают распределения:

1. равномерное

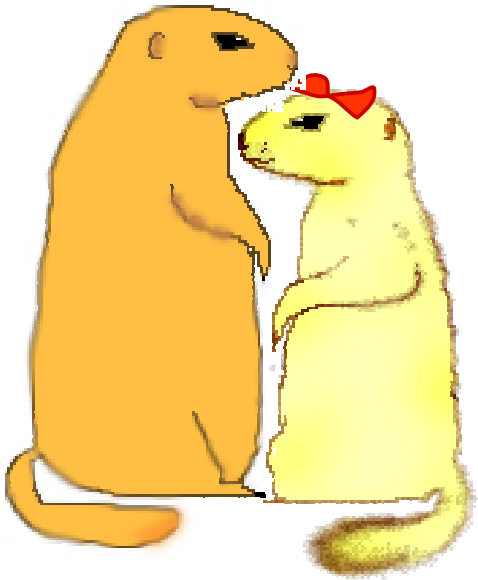


2. случайное

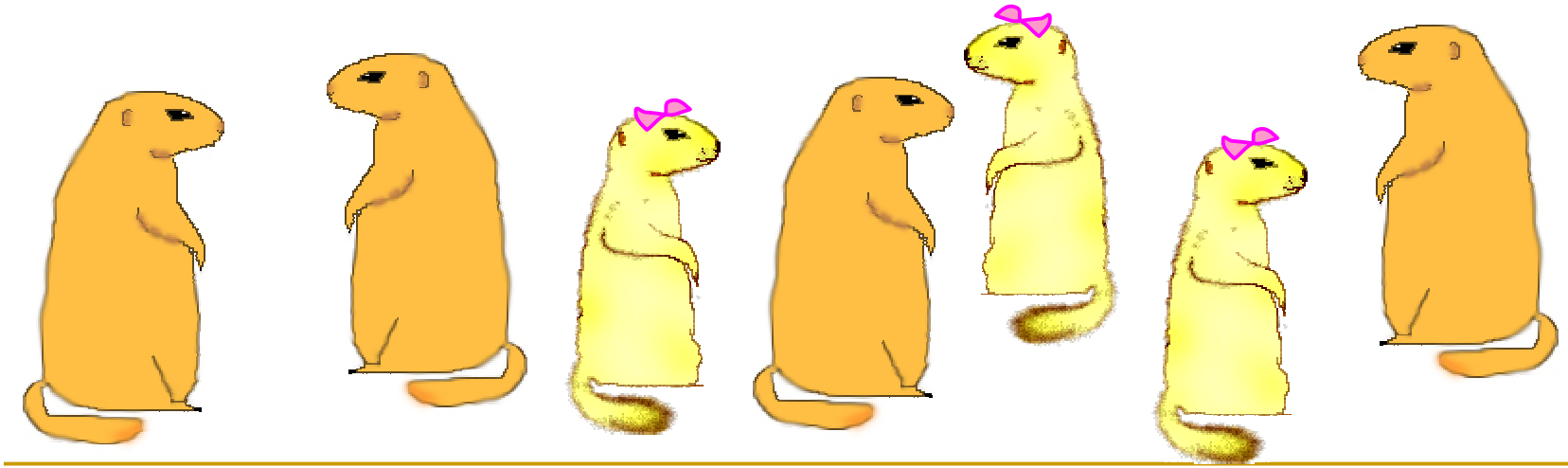


Могут быть и дискретными, и непрерывными

Возможное соотношение самцов и самок в выводке:



7:0; 6:1; 5:2; **4:3**; 3:4; 2:5; 1:3; 0:7



Дискретные распределения

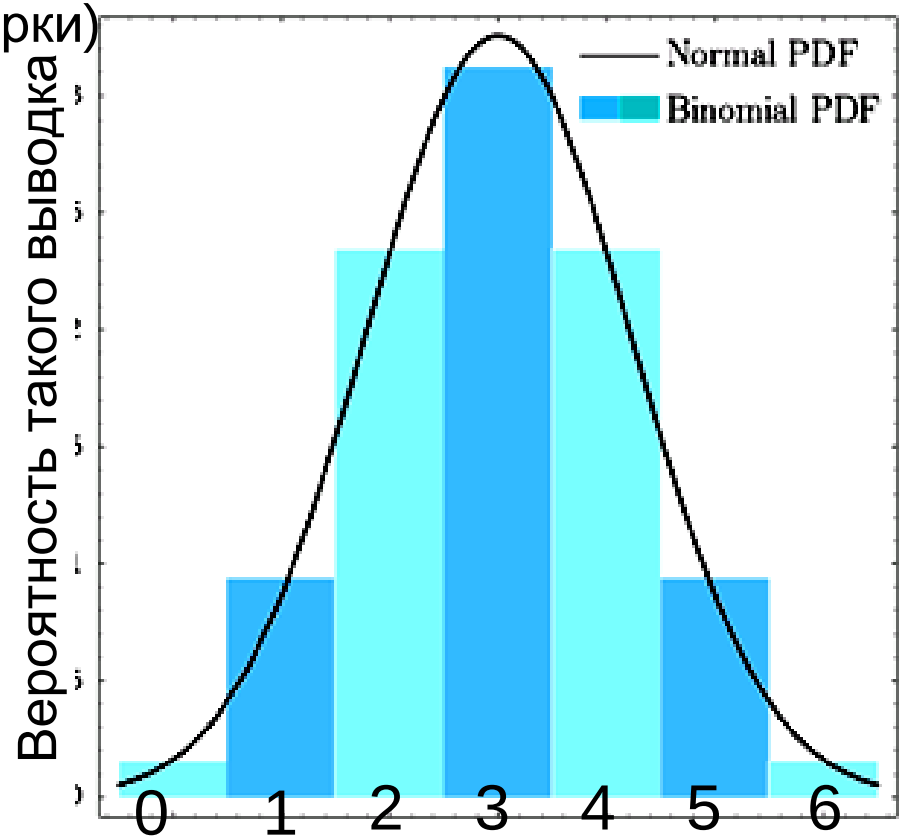
Биномиальное распределение

A – событие с вероятностью p.

q – вероятность того, что A не произойдёт в данном испытании; p+q=1

n – число испытаний (размер выборки)

X – число появления A в этих испытаниях



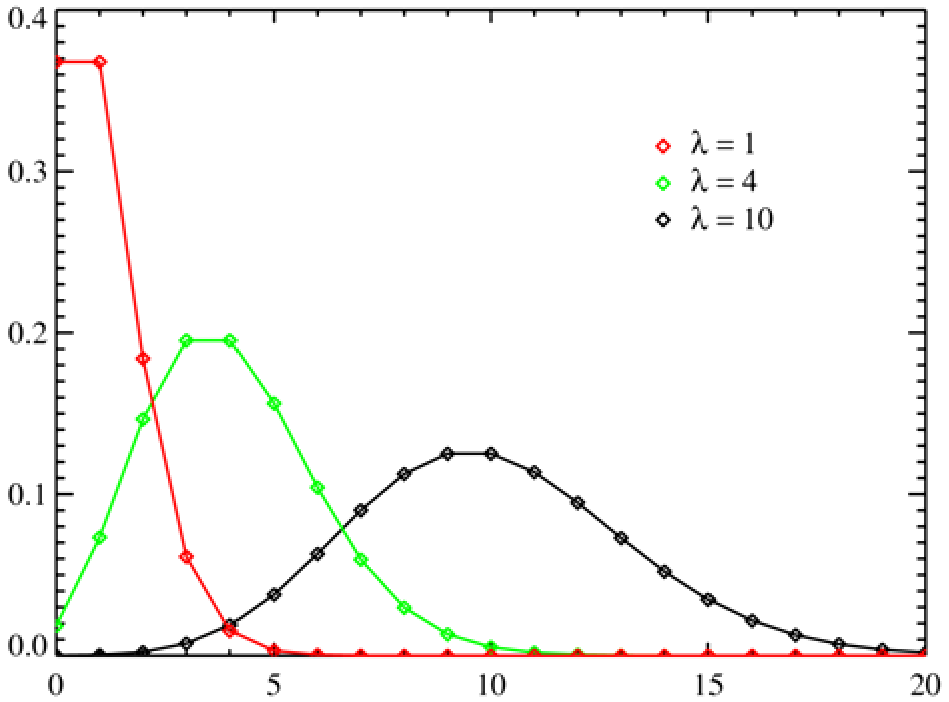
Распределение Пуассона

для редких и случайных событий, когда n очень большое, p<<q

$$np = \lambda$$

$$P_n(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

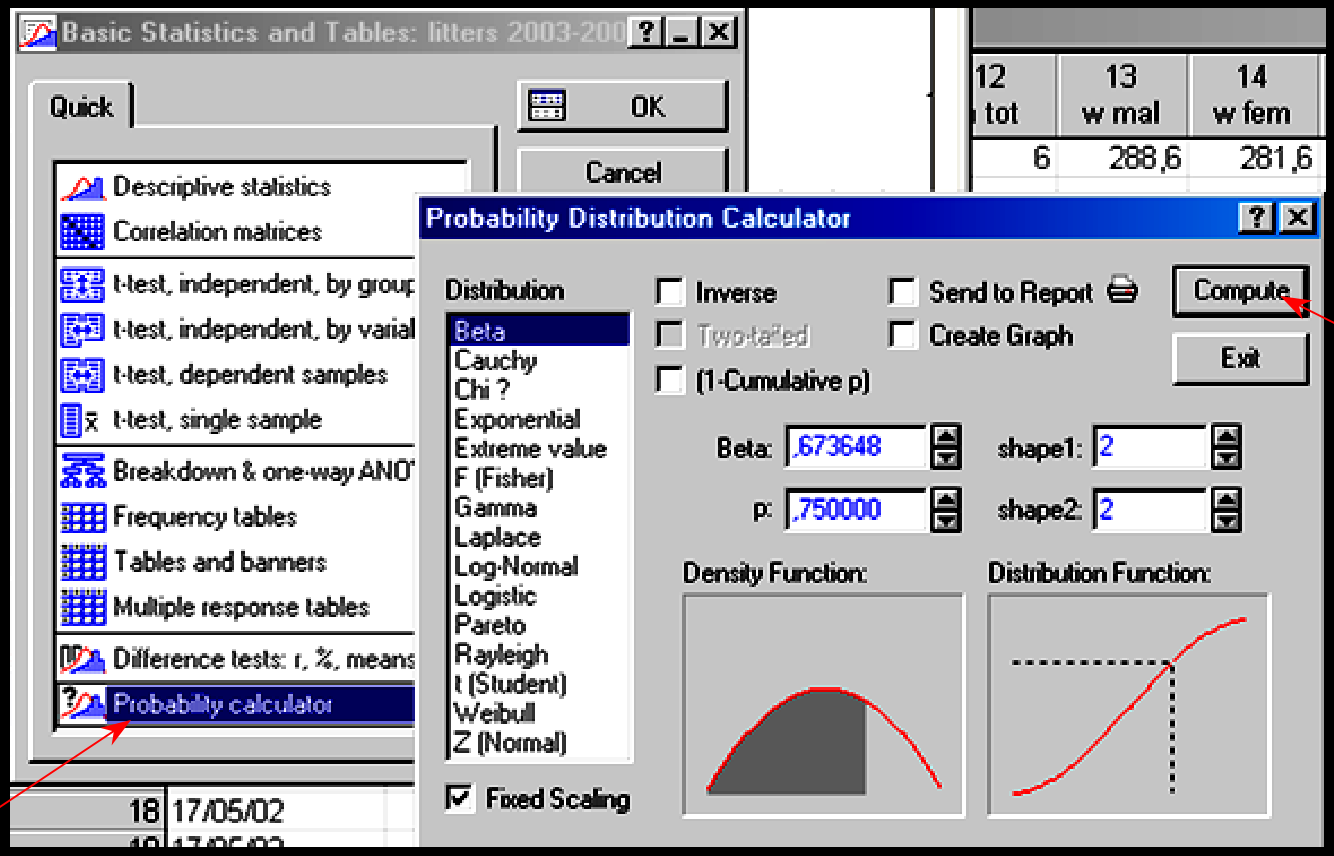
$$\mu = \sigma^2$$



При больших n приближается к нормальному

Другие распределения

Распределение Бернулли, логарифмическое, геометрическое, отрицательное биномиальное, гипергеометрическое и др.



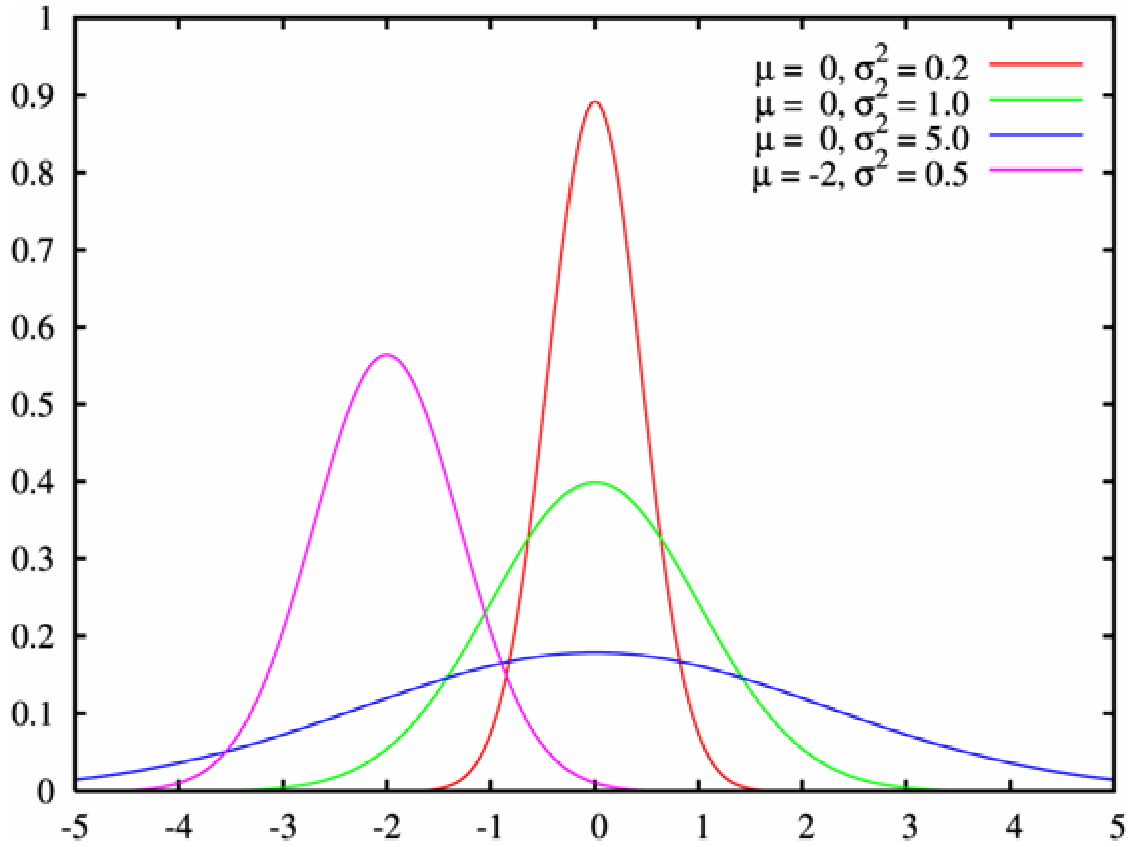
descriptive statistics

Непрерывные распределения

Нормальное распределение (Гауссово)

1. унимодальное;
2. симметричное.

** Стандартное отклонение в НР примерно равняется расстоянию от среднего до точки перегиба*



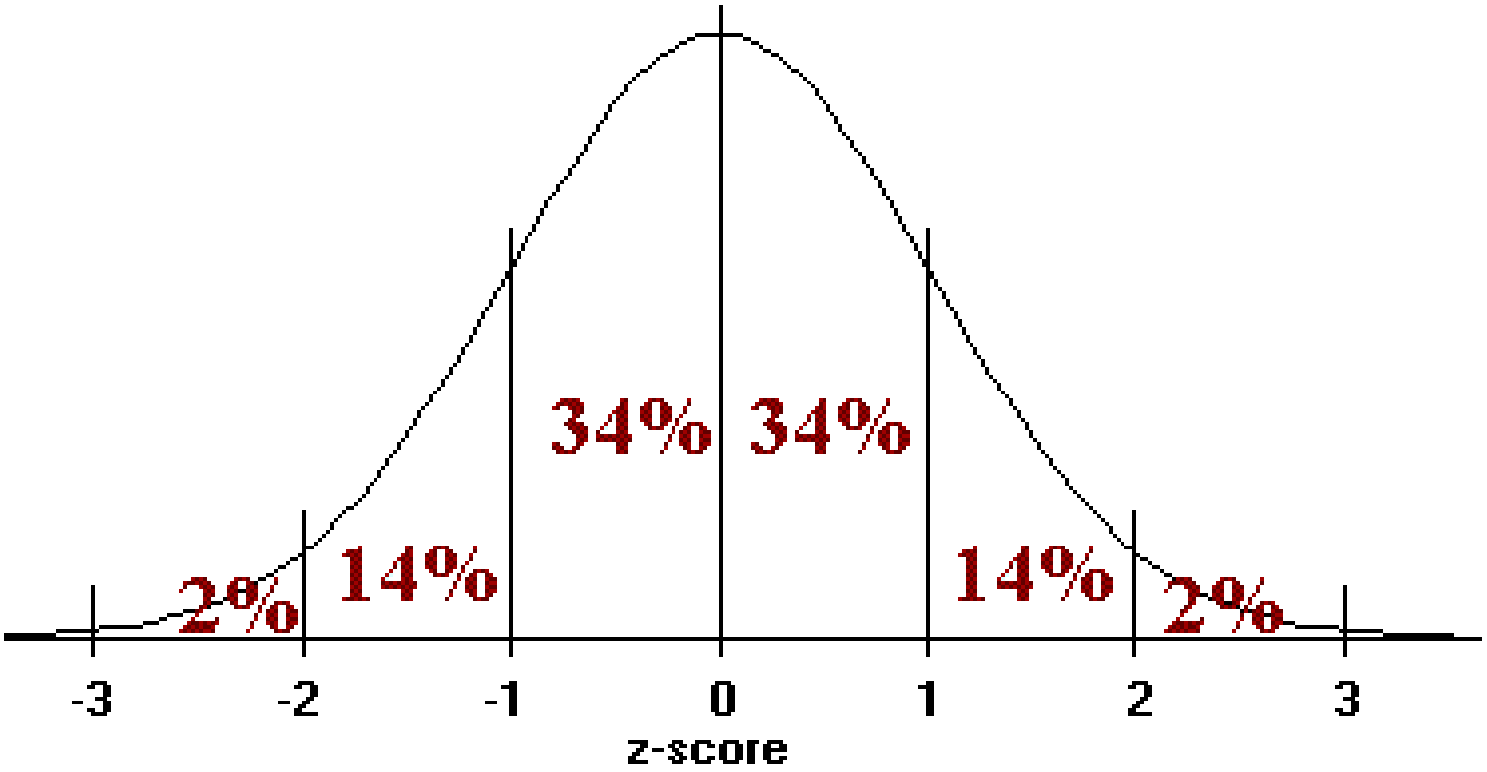
ЦЕНТРАЛЬНАЯ ПРЕДЕЛЬНАЯ ТЕОРЕМА

С увеличением размера выборок распределение *средних значений* выборок из генеральной совокупности, имеющей ЛЮБОЕ распределение, будет стремиться к *нормальному*

Возникает, когда некоторая величина отклоняется от средней под воздействием слабых, независимых друг от друга факторов

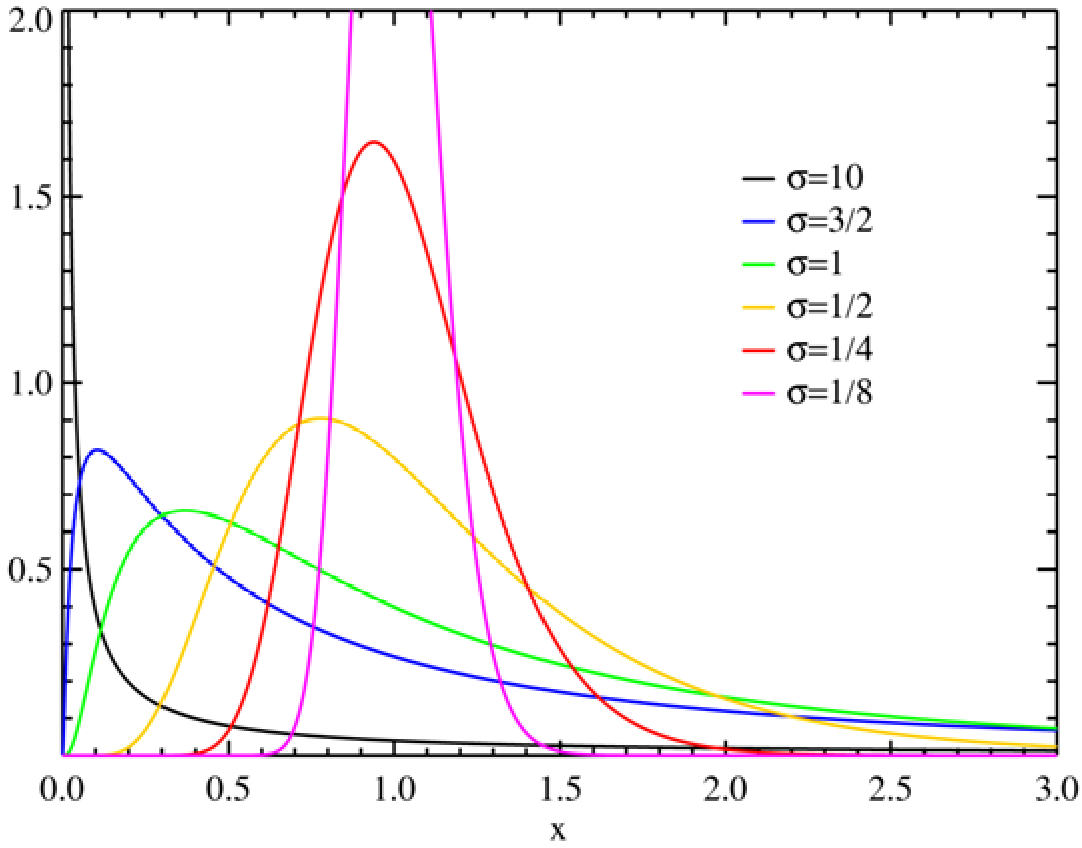
Почему в природе оно так часто встречается?
Из-за Центральной предельной теоремы! Факторы, действующие на зверька, многочисленны и не обязательно имеют нормальное распределение.
Но если они не зависимы друг от друга, то все вместе они дадут именно нормальное распределение переменной, которую мы анализируем (массы, например).

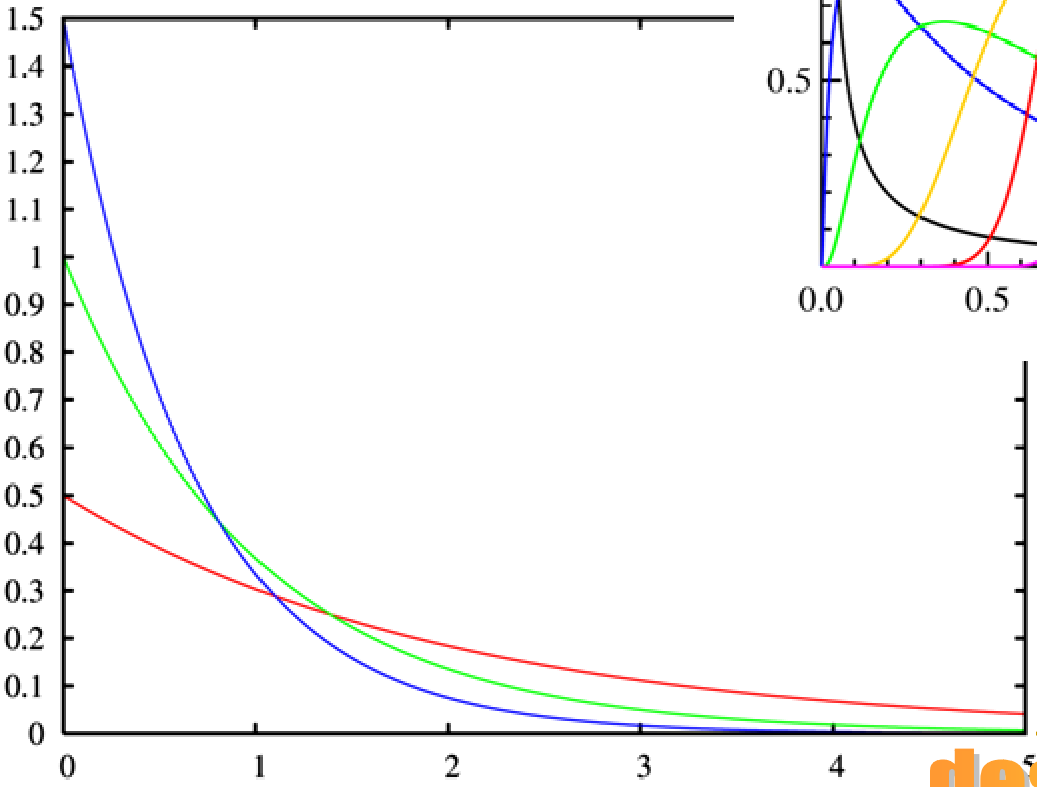
Нормальное распределение



Количество стандартных отклонений

Логнормальное распределение





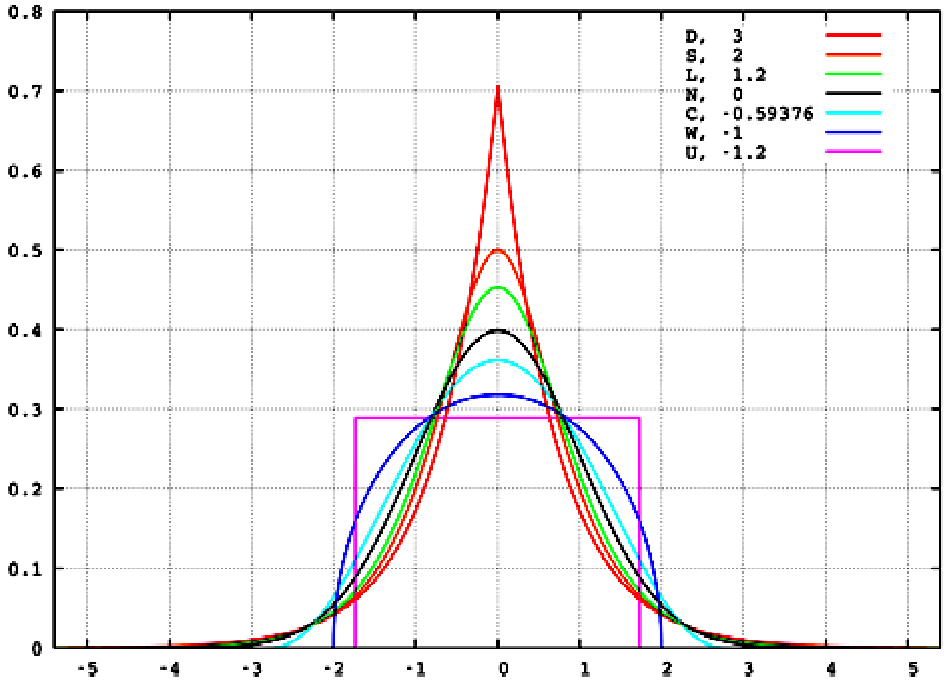
Экспоненциальное распределение

descriptive statistics

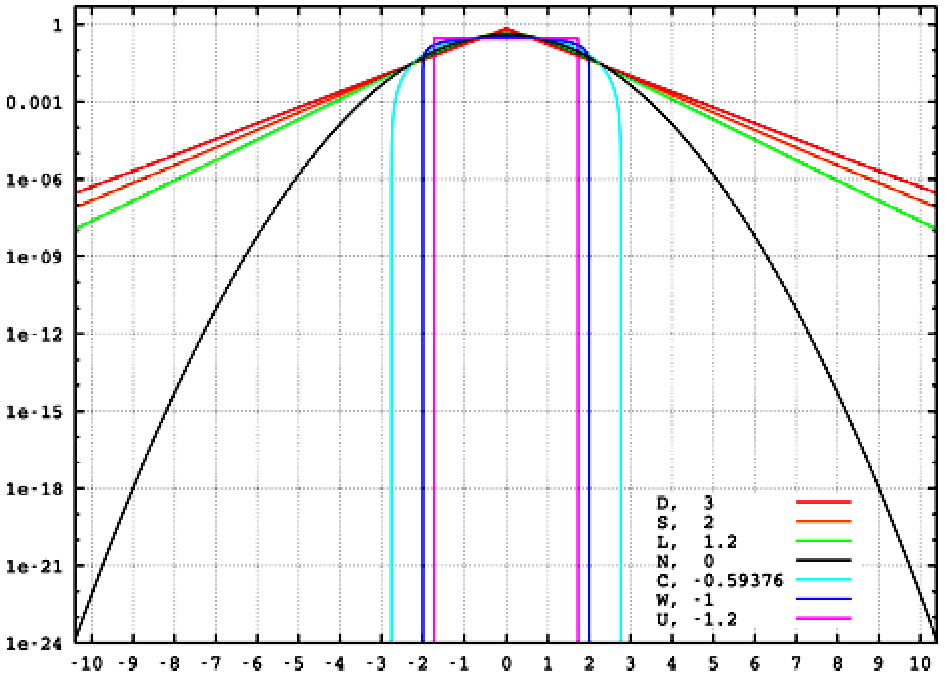
Асимметрия (skewness) и эксцесс (kurtosis)

Асимметрия: g1<0 – распределение будет скошено влево; g1>0 – вправо.

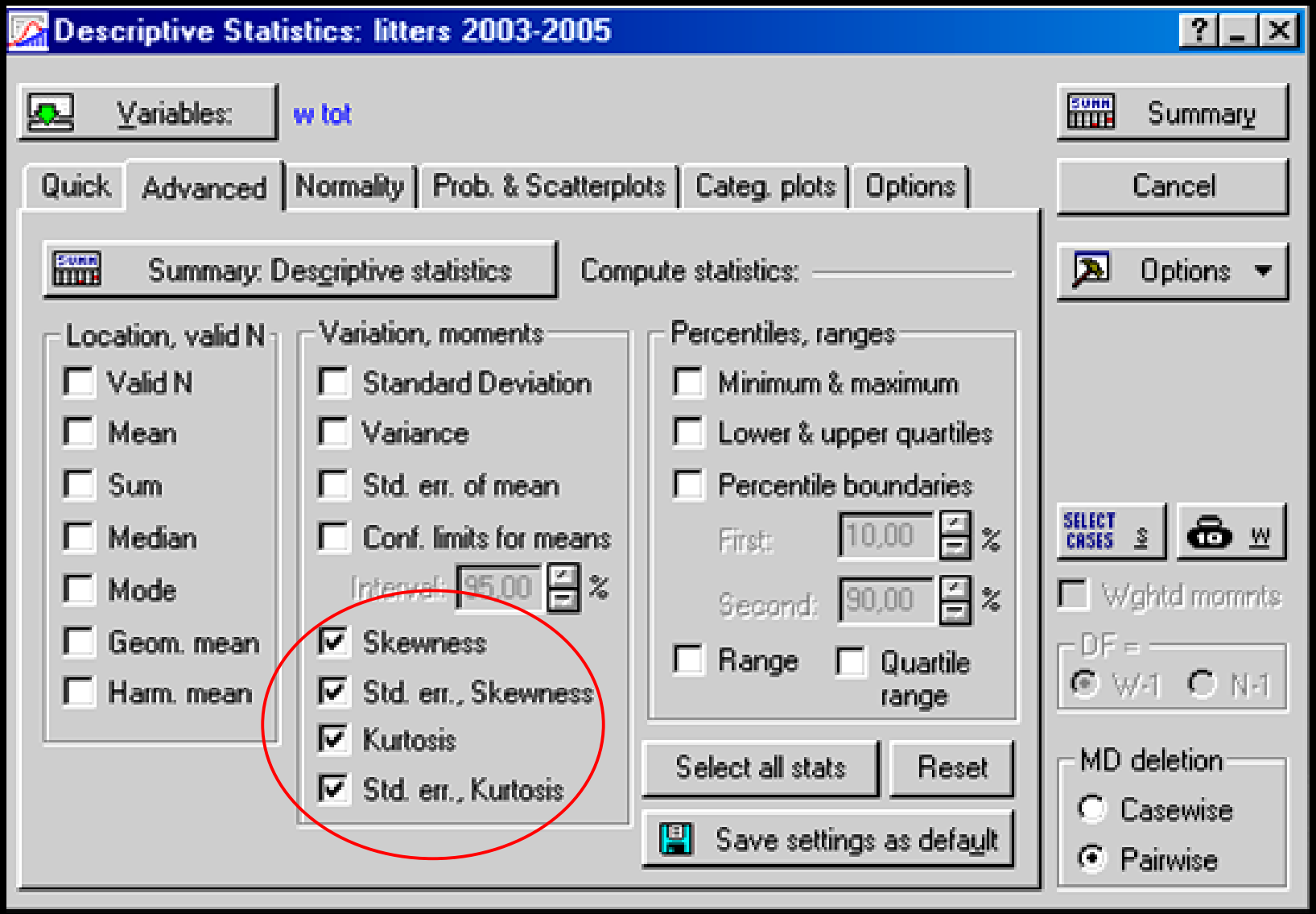
Эксцесс: g2<0 – platykurtosis; g2>0 - leptokurtosis



leptokurtic

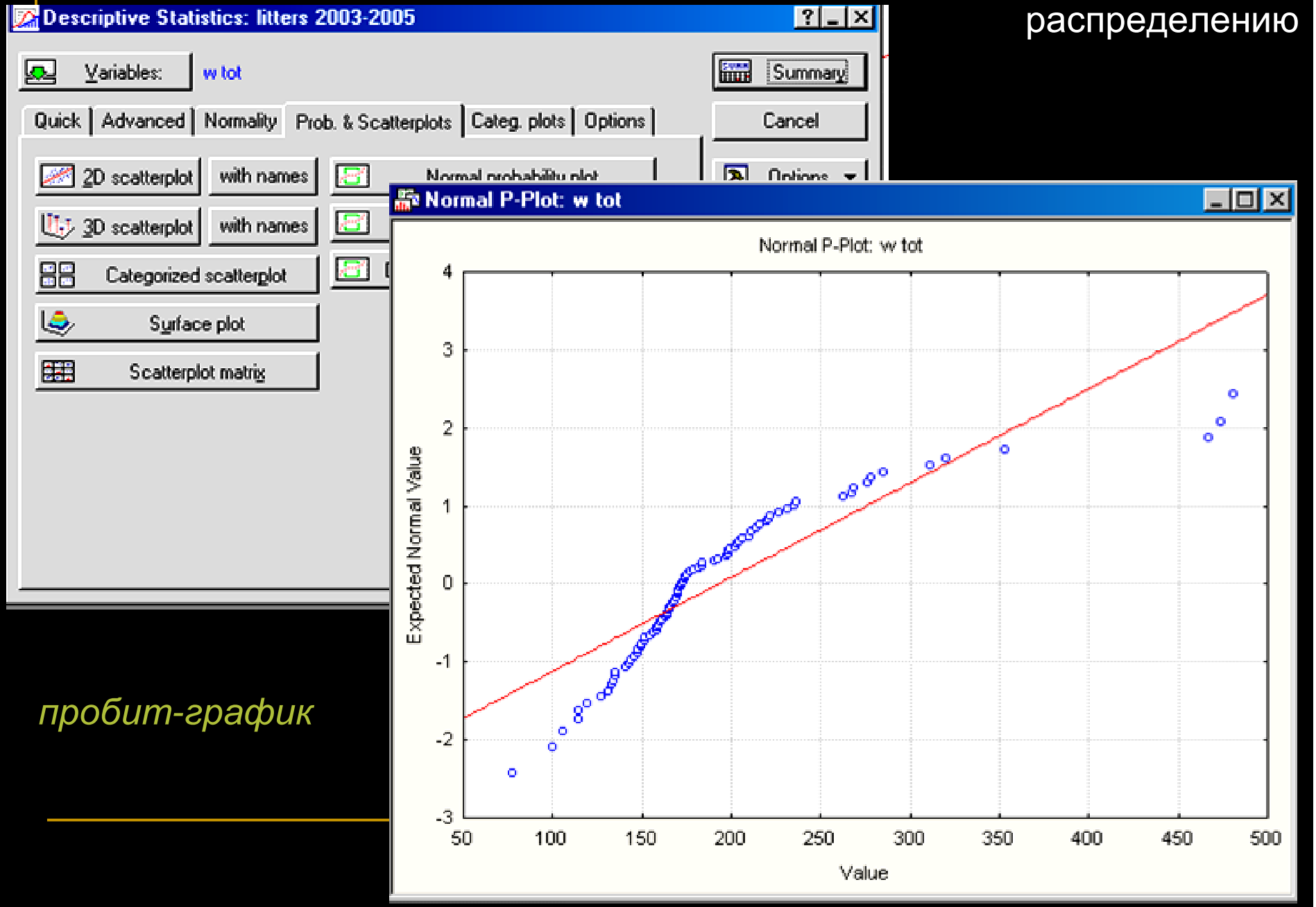


platykurtic



descriptive statistics

Графическое представление соответствия данных нормальному распределению



пробит-график

Тестирование гипотез в статистике

Шаг 1. Формулируем нулевую гипотезу H_0 *и альтернативную ей* H_1

Гипотеза – некоторое предположение об исследуемой популяции и её параметрах.

Именно о ПОПУЛЯЦИИ, а не о выборке из неё!
(предположения о самой выборке легко проверить без статистики)

Тестирование гипотезы – процедура, в которой мы решаем, принять гипотезу или не принимать.

Опровергнуть гипотезу в принципе легче, чем подтвердить.

Тестирование гипотез в статистике

Шаг 1. Формулируем нулевую гипотезу H_0 *и альтернативную ей* H_1

Пример.

Мы хотим узнать, отличается ли средняя масса землероек из США от 1 кг?
Мы никак не поймаем всех землероек в Штатах.
Поймаем 100 зверьков и взвесим.

H_0: $\mu = 1000$ г; (где μ – средняя масса **всех** землероек в США)

H_1 : $\mu \neq 1000$ г

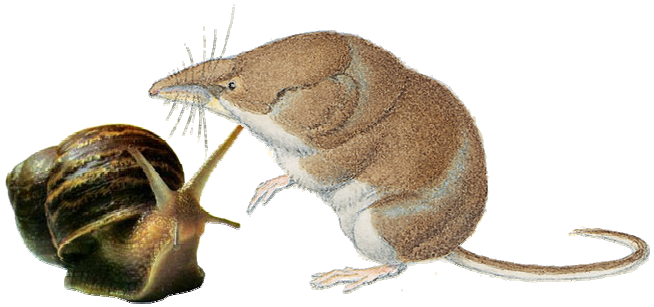
H_0 обычно говорит, что нет различий, нет эффекта, нет изменений…

Шаг 2. Рассчитываем значение статистики соответствующего критерия

Принять нам или отвергнуть H_0 гипотезу мы решаем на основе СТАТИСТИКИ КРИТЕРИЯ.

Она считается на основе нашей выборки.
Она специально разрабатывается для проверки разных типов гипотез.

Весь статистический анализ и представляет собой подбор правильных статистик для проверки разных 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*)

hypothesis testing

Ошибка 1 рода: вероятность найти различия, где их нет.

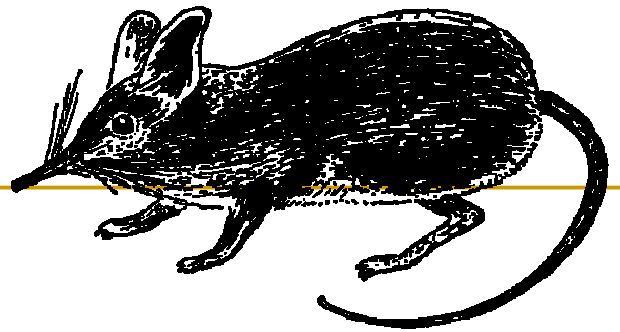
(землеройки и правда весят 1 кг в среднем. Но нам показалось, глядя на выборку, что они отличаются).

Это – нездоровые сенсации, которые могут принести большой вред. Мы её сами определяем – это ***уровень значимости*** *α.*

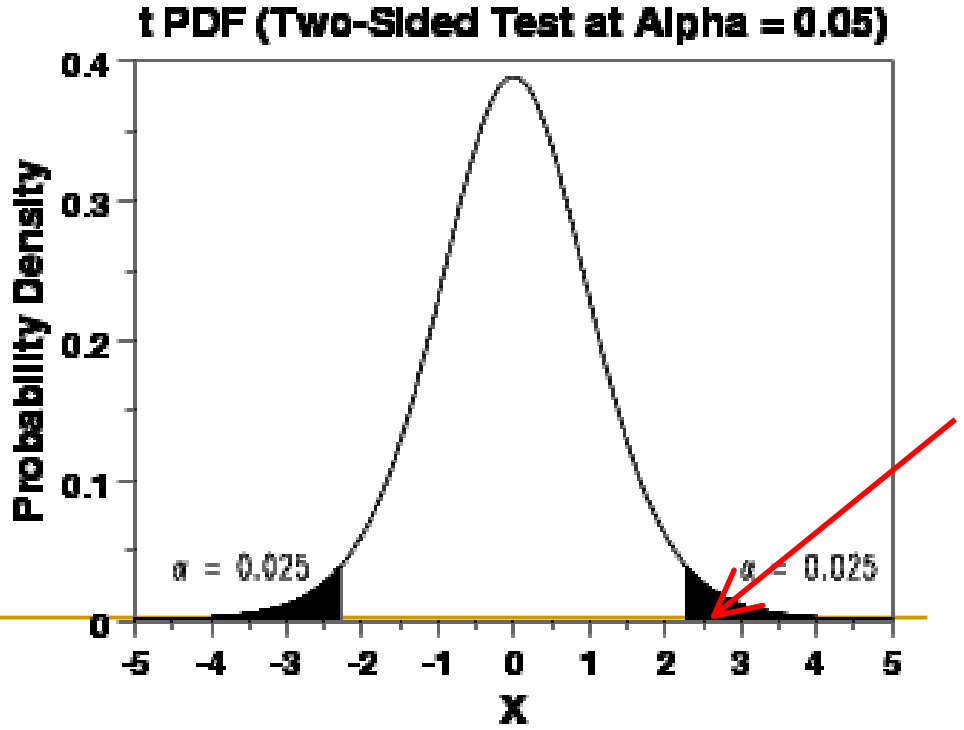
Ошибка 2 рода: *вероятность не увидеть различий, где они есть.*

(на самом деле, конечно, землеройки не весят в среднем 1 кг. Но мы были слишком строги к себе и посчитали, что этих различий недостаточно)

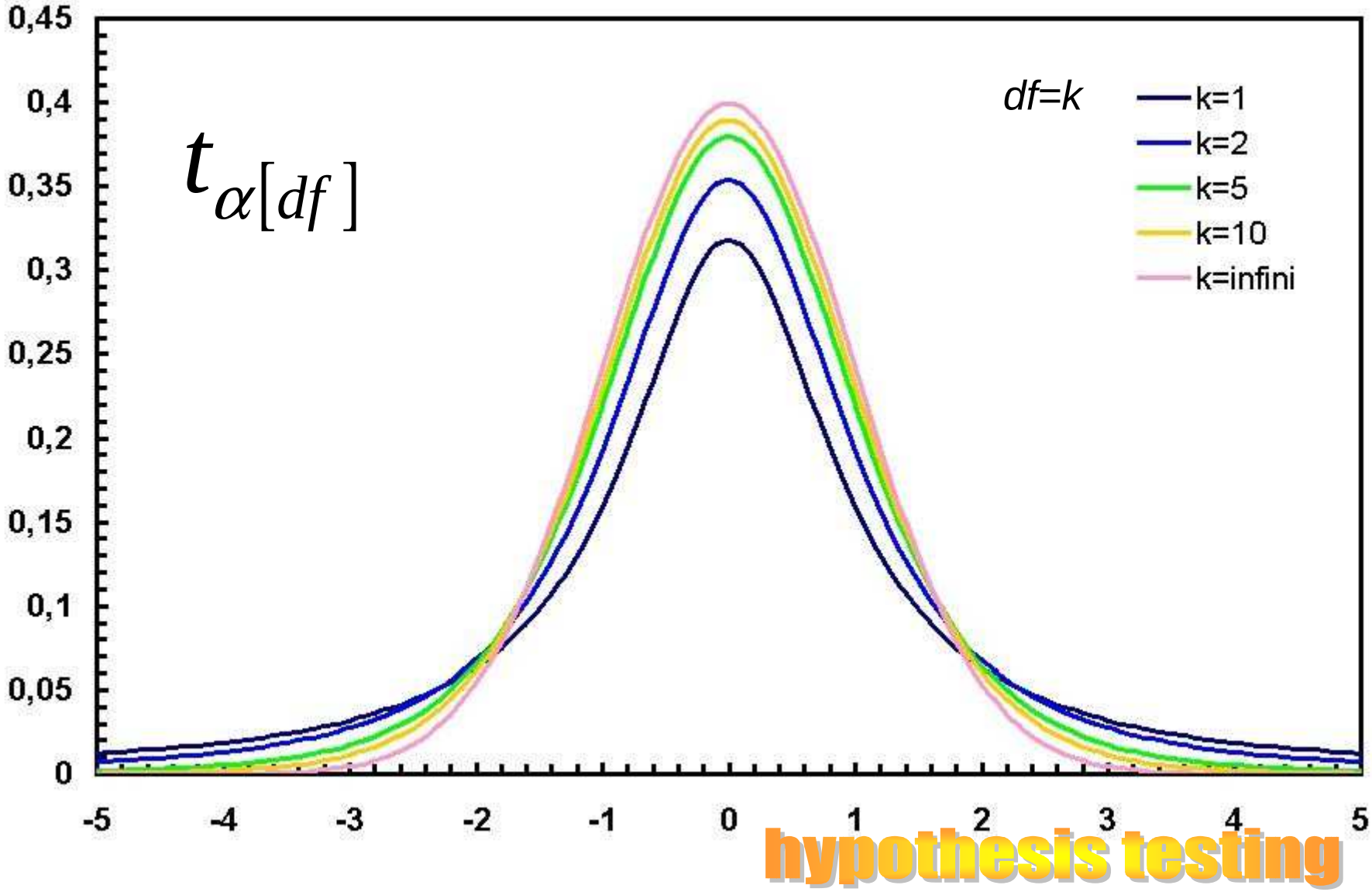
Это «близорукость», или «слепота» критерия, вред от неё не очень большой.



Уже давно придумана статистика критерия для проверки нашей гипотезы.
В нашем случае она имеет t распределение.
Мы выбираем распределение исходя из размера выборки. Выбираем α, находим критическое значение и смотрим, куда попадёт значение статистики для конкретной выборки.

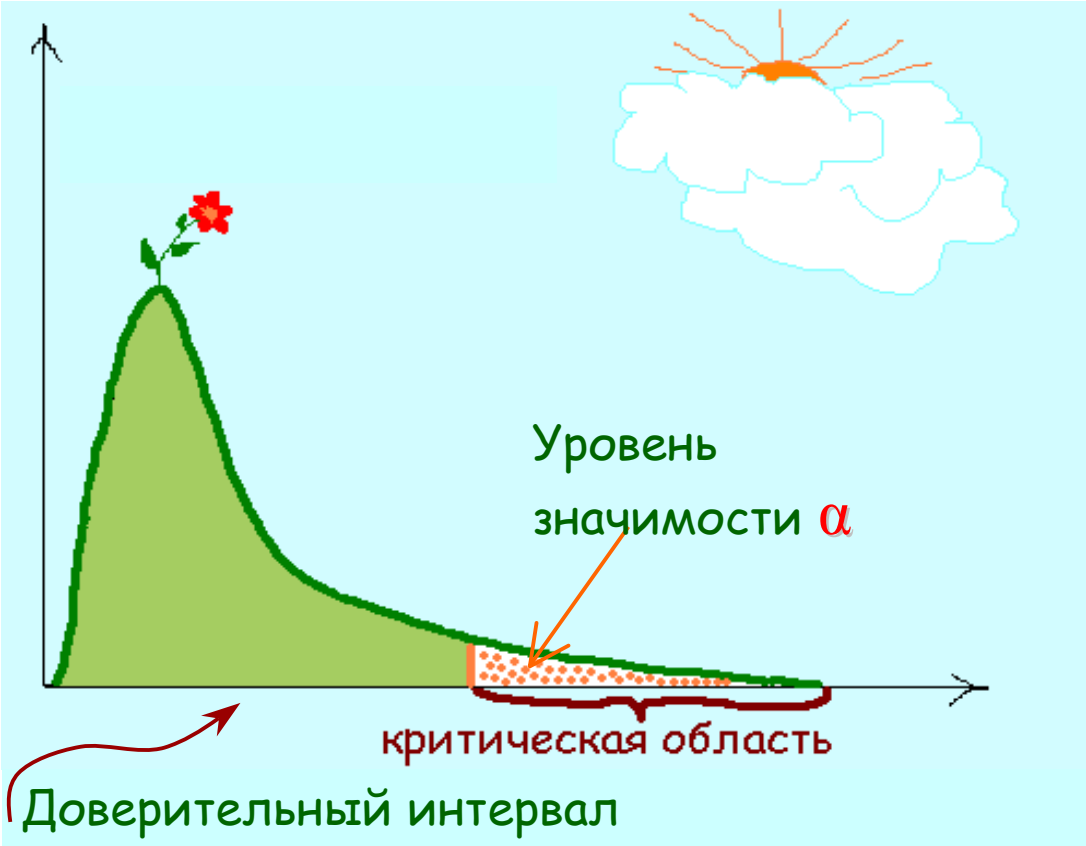


t-распределение (Стьюдента)



Тестирование гипотез в статистике

Шаг 3. Сравниваем полученное значение с критическим значением в соответствии с заданным уровнем значимости



Если наше значение попало в критическую область – отвергаем H_0 это значит, что мы нашли различия

$\alpha = 0{,}05$

Оно равно площади оранжевой фигуры

hypothesis testing

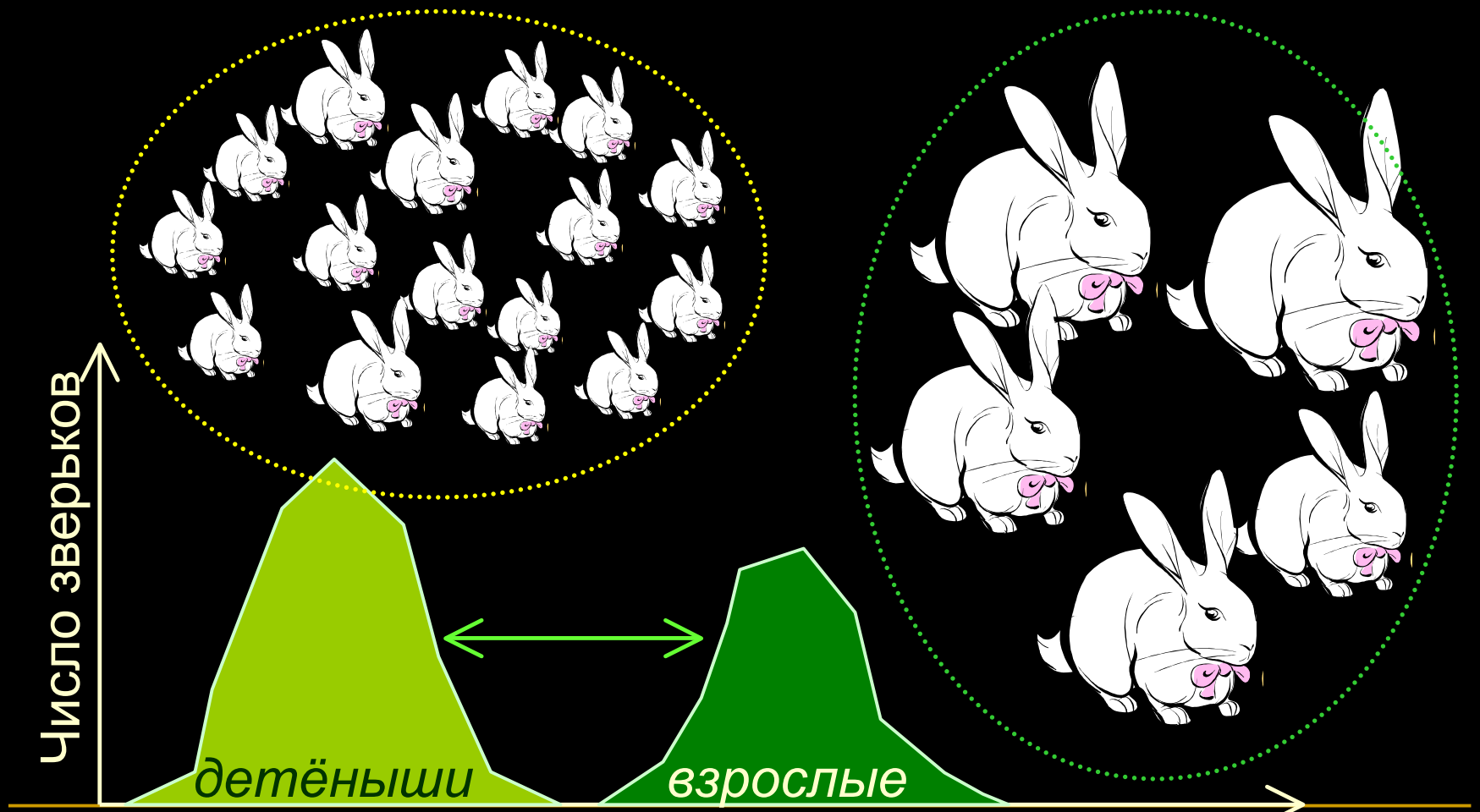
Альтернативы бывают односторонними и двусторонними

1. Односторонние альтернативы – если мы знаем, в какую сторону может быть отклонение от H_0 (например, мы сравниваем массу взрослых зверьков и детёнышей) – *one-tailed test*

2. Двусторонние альтернативы – если мы заранее этого не знаем - *two-tailed test*

H_0 : средняя масса взрослых зверьков НЕ ОТЛИЧАЕТСЯ от средней массы детёнышей

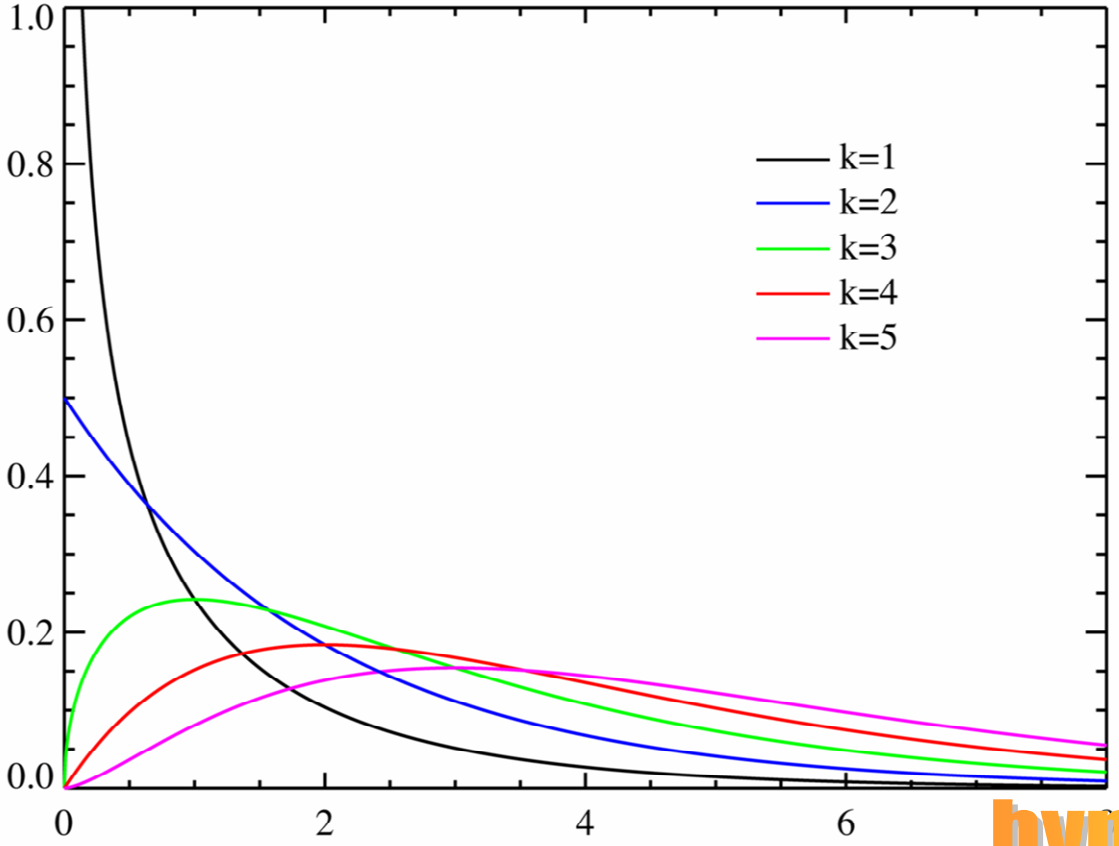
H_1: средняя масса взрослых БОЛЬШЕ средней массы детёнышей



ONE-TAILED TEST

Статистики критериев имеют особые, специальные распределения.

Все распределения получаются с использованием ПАРАМЕТРОВ нашего распределения (среднего значения, дисперсии) и числа СТЕПЕНЕЙ СВОБОДЫ $df = n\text{-}1$ (n – размер выборки)



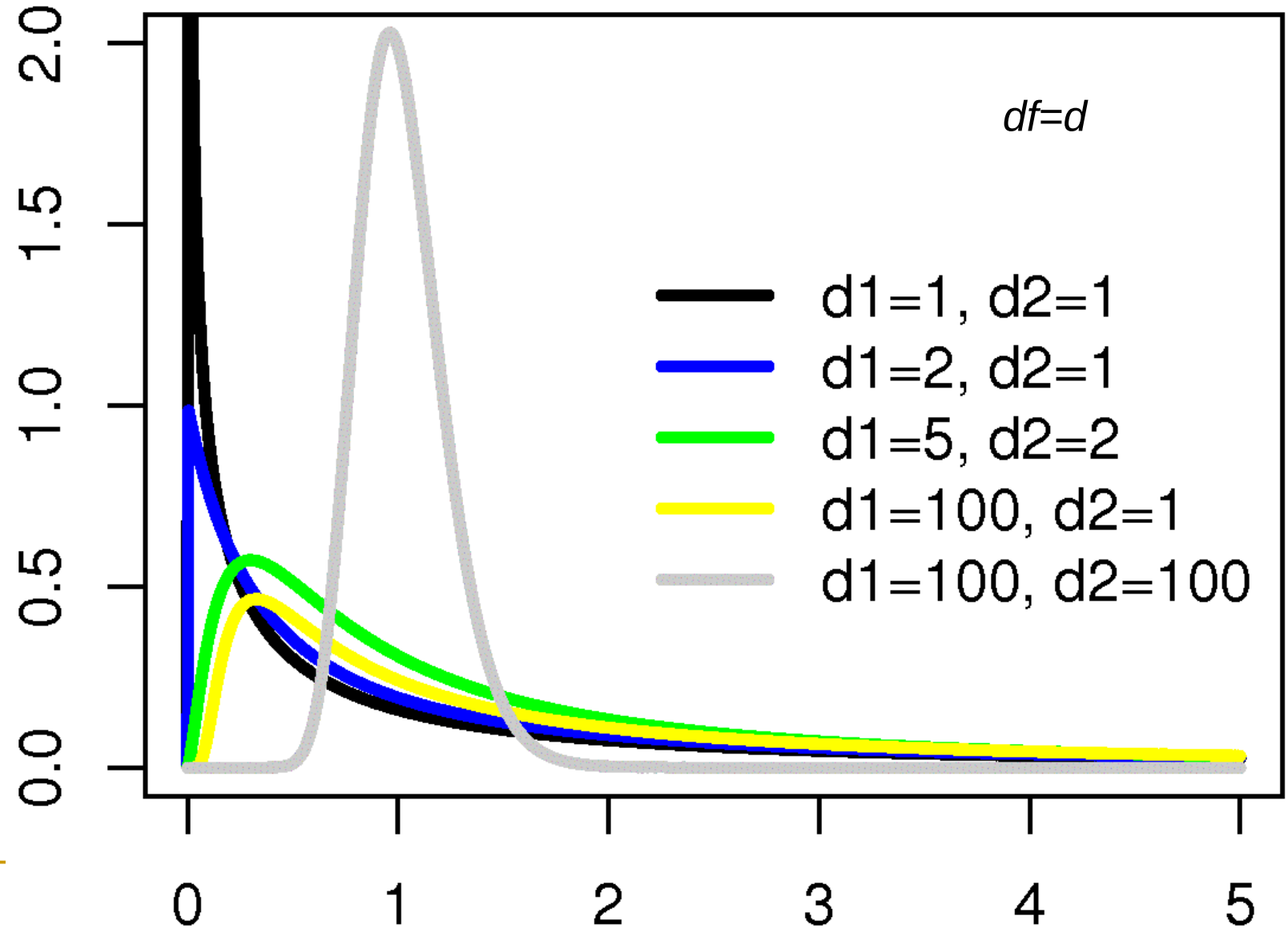
$$\chi^2_{\alpha[df]}$$

На рисунке *df=k*

Распределение χ^2

hypothesis testing

F-распределение (Фишера) $F_{\alpha[df_1, df_2]}$




Основные задачи
Сравнение распределений по форме; качественные признаки
Сравнение групп по средним значениям
Поиск зависимостей и корреляций
непрерывные распределения
дискретные распределения
2 группы: (парные тесты)
связанные
независимые
>2-х групп:
связанных
независимых
корреляции совместное изменение переменных
регрессии причинно-следственная связь!!
✓сравнение с теоретическим р.
✓сравнение наблюдаемых р.
✓сравнение с теоретическим р.
✓сравнение наблюдаемых р.
✓2x2 таблицы
параметрические тесты
непараметрические тесты
КРИТЕРИИ СОГЛАСИЯ; ЧАСТОТНЫЕ КРИТЕРИИ
ANOVA
РЕГРЕССИИ; КОРРЕЛЯЦИИ