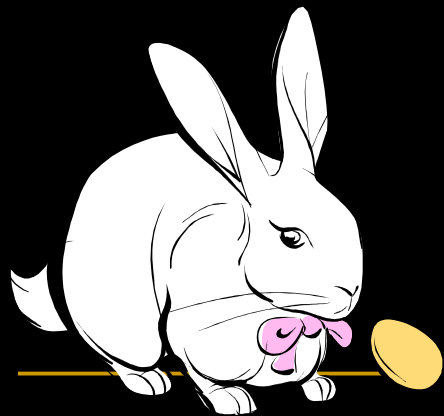


Издания расположены по порядку в зависимости от моего уважения и любви к ним.

- *Sokal R.R., Rolf F.J.* 1995. Biometry. 3rd ed., W.H. Freeman & Co., New York. 887 pp. – *все издания хороши.*
- *Zar J.H.*, 1999. Biostatistical analysis. 4th ed., Prentice Hall, New Jersey.
- *Hurlburt R.T.*, 2003. Comprehending behavioral statistics. 3rd ed., Wadsworth/Thomson Learning, Belmont. 588 pp. *самый простой из англоязычных, понятные примеры, свежий взгляд.*
- *Гланц С.*, 1998. Медико-Биологическая Статистика. McGraw-Hill, 1994; М.: Практика. 459 с. *Простой, доступный, для понимания основ. Но про конкретные тесты – мало.*
- *Lehner P.N.*, 1996. Handbook of ethological methods. Cambridge University press. 672p.
- *Sheskin D.J.*, 2000. Handbook of parametric and nonparametric STATISTICAL PROCEDURES 2nd ed., Chapman & Hall/CRC. *Это огромный суперподробный манускрипт.*
- **Гмурман В.Е.*, 1979. Руководство к решению задач по теории вероятностей и математической статистике. М.: Высш. школа, , 400 стр. *Если кто Очень хочет понять математическую суть вопроса...*

Частотный анализ

Занятие 2

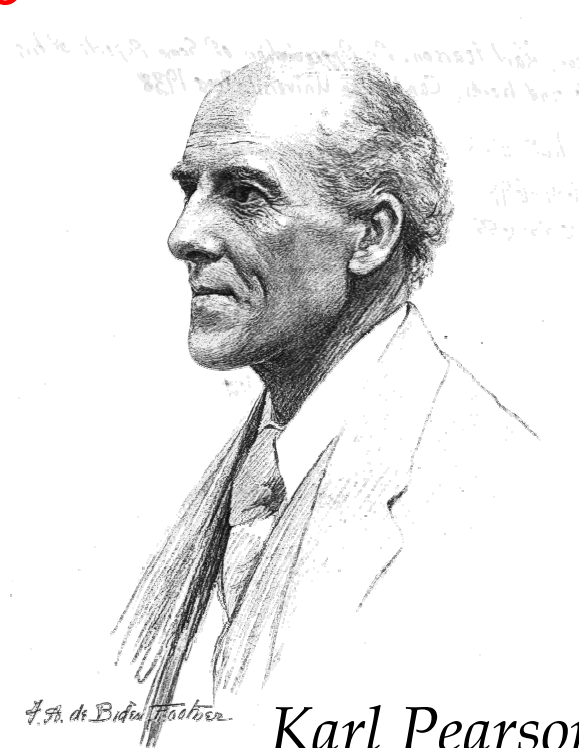


Критерии согласия (Tests for goodness of fit):

1. качественные и ранговые данные

У нас есть **выборка**. Данные — **качественные**.

Вопрос: соответствует ли распределение в популяции, из которой получена выборка, теоретическому распределению? (которое мы сами определяем)



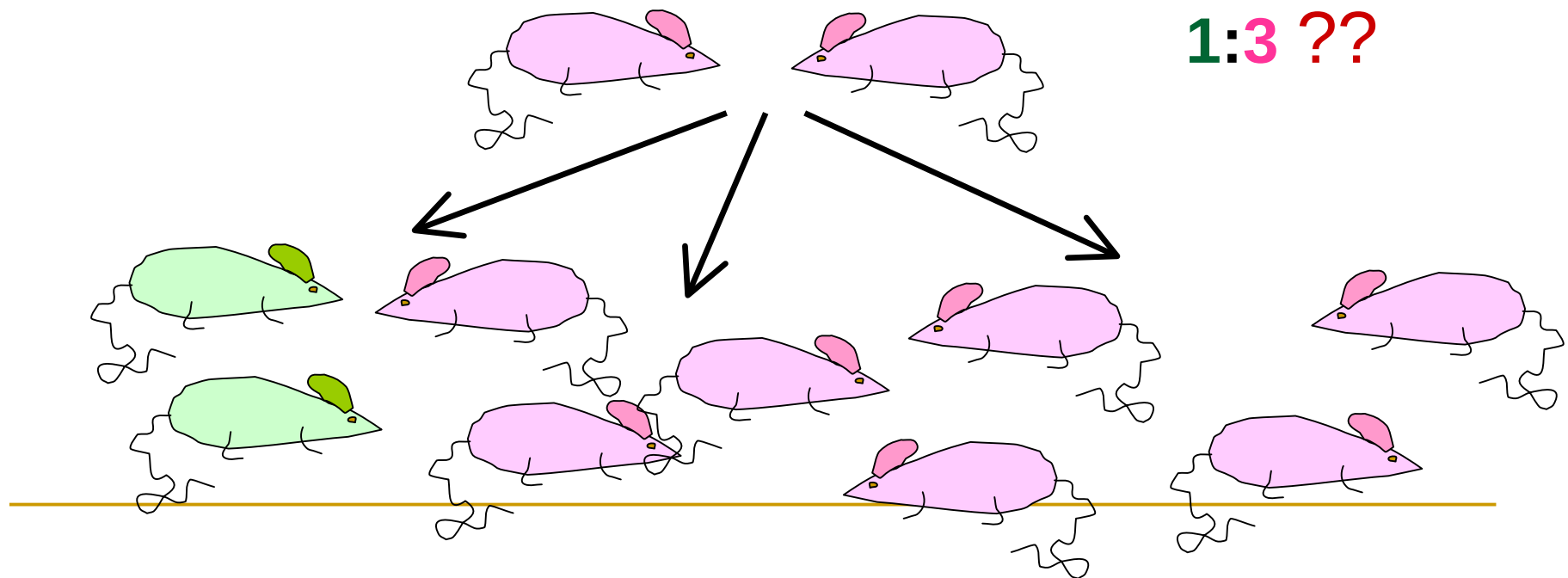
Karl Pearson
Придумал χ^2
статистику ещё
в 1900 году!

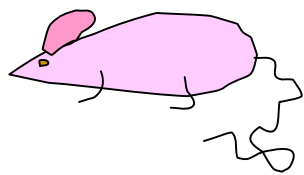
Родились:

84 розовых мыши и 16 зелёных.

H_0 : выборка получена из популяции, где соотношение розовых и зелёных – 3:1.

H_1 : выборка получена из популяции, где соотношение розовых и зелёных не равно 3:1





	розовые	зелёные	всего
f_i	84	16	100
$\bar{f}_i$	75	25	

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - \bar{f}_i)^2}{\bar{f}_i} = \frac{(84 - 75)^2}{75} + \frac{(16 - 25)^2}{25} =$$
$$= 1.080 + 3.240 = 4.320$$

$$\chi^2_{0.05,1} = 3.841 > 4.320$$

$$p = 0.038$$

H_0 отвергаем

Чем больше значение χ^2 , тем хуже наши данные соответствуют теоретическому распределению – тем меньше p

Критерии согласия (Tests for goodness of fit):

1. качественные и ранговые данные

Категорий может быть сколько угодно.

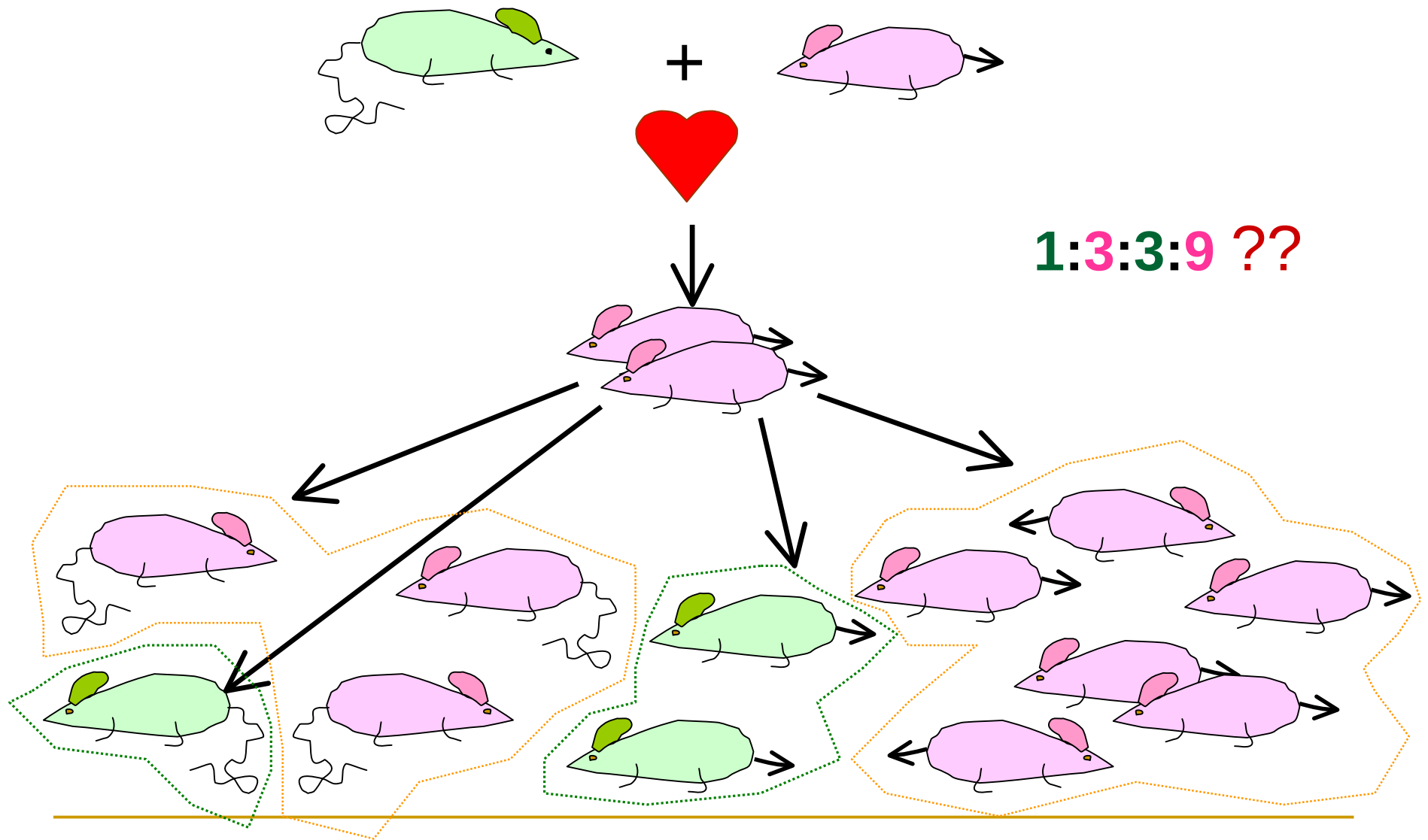
Родились:

152 розовых мыши с острым хвостом; 39 розовых с курчавым хвостом; 53 зелёных с острым, 6 зелёных с курчавым.

H_0 : выборка получена из популяции, где соотношение фенотипов – 9:3:3:1.

H_1 : выборка получена из популяции, где соотношение фенотипов не равно 9:3:3:1

Критерии согласия (Tests for goodness of fit)



Важное замечание:

В всех критериях согласия H_0 гипотеза – о том, что форма распределений **ОДИНАКОВА**.

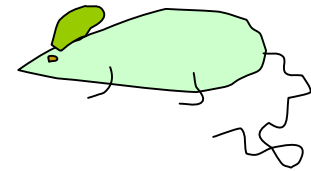
То есть, когда мы ищем подтверждение тому, что наши данные удовлетворяют некоторому распределению, мы должны радоваться, получив $p \gg 0.05$!



Zar, 1999:

Если вы сравнили распределение с теоретическим, получили отличия (!), а теперь хотите показать, из-за какой именно категории эти отличия возникли, можно отдельно сравнить с теоретическим распределением остальные категории, а затем – отношение этой категории к остальным.

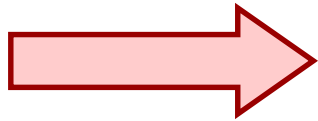
Т.е., если нам кажется, что всё портят зелёные мыши с курчавыми хвостами, сравним:



1. соотношение остальных мышей с 9:3:3;
 2. отношение зелёных-курчавых к остальным с 1:15.
-

Итак:

1. *у нас одна выборка*
2. мы сравниваем наблюдаемые частоты с ожидаемыми (**observed and expected**)



Критерий χ^2 Пирсона (**Pearson Chi-square test**)

Сравнение *нашего распределения с теоретическим*
(нужна таблица с посчитанными частотами)

Data: Spreadsheet1* (10v)

	1	2
	observed	expected
1	152	140,625
2	39	46,875
3	53	46,875
4	6	15,2625
5		
6		
7		
8		
9		
10		

Nonparametric Statistics: Spreadsheet1

Quick

- 2 x 2 Tables (X²/V²/Phi², McNemar, Fisher exact)
- Observed versus expected X²**
- Correlations (Spearman, Kendall tau, gamma)
- Comparing two independent samples (groups)
- Comparing multiple indep. samples (groups)
- Comparing two dependent samples (variables)
- Comparing multiple dep. samples (variables)
- Cochran Q test
- Ordinal descriptive statistics (median, mode, ...)

OK

Cancel

Options

Open Data

SELECT CASES

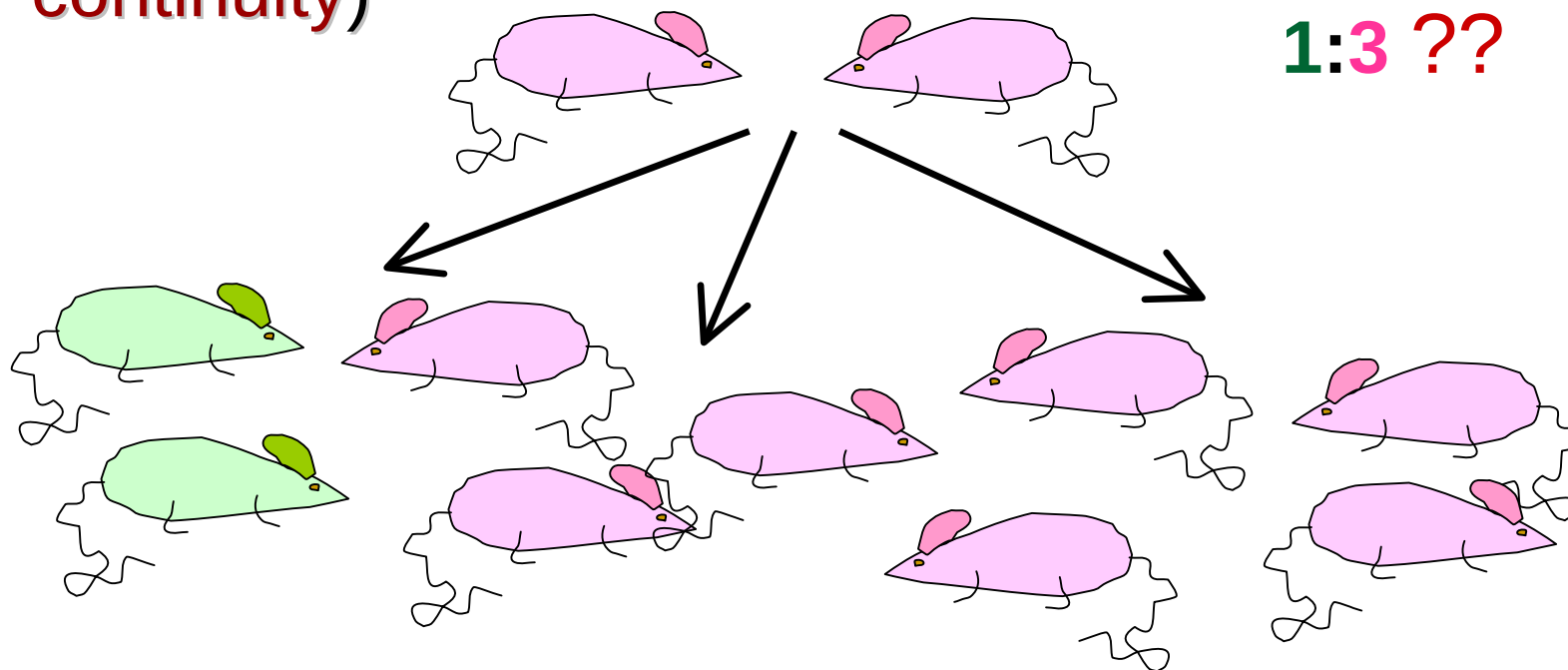
W

результаты

Data: Observed vs. Expected Frequencies (Spreadsheet1)						
Observed vs. Expected Frequencies (Spreadsheet1)						
Chi-Square = 8,664667 df = 3 p < ,034100						
NOTE: Unequal sums of obs. & exp. frequencies						
Case	observed	observed	expected	O - E	(O-E)**2 /E	
C: 1	152,0000	152,0000	140,6250	11,37500	0,920111	
C: 2	39,0000	39,0000	46,8750	-7,87500	1,323000	
C: 3	53,0000	53,0000	46,8750	6,12500	0,800333	
C: 4	6,0000	6,0000	15,2625	-9,26250	5,621222	
Sum	250,0000	250,0000	249,6375	0,36250	8,664667	

ЕСЛИ У НАС ТОЛЬКО 2 ПРОЯВЛЕНИЯ ПРИЗНАКА

Поправка Йейтса для критерия χ^2 (Yates correction for continuity)



Для заданного теоретического распределения χ^2 может принимать только строго определённые значения для разных наблюдаемых распределений.

Например: если ожидаемые частоты – 75 и 25, то

значения χ^2 будут

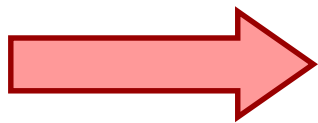
для 84 и 16 – 4.32,

для 83 и 17 – 3.14,

для 82 и 18 – 2.61

} промежуточных значений
не может быть для данных
ожидаемых частот

Но χ^2 распределение непрерывное. И для заданного уровня значимости p мы не найдём точно соответствующего ему значения χ^2 .



χ^2 с поправкой Йейтса:
(для больших N не нужен)

Делает тест более консервативным.

$$\chi^2 = \sum_{i=1}^k \frac{(|f_i - \bar{f}_i| - 0.5)^2}{\bar{f}_i}$$

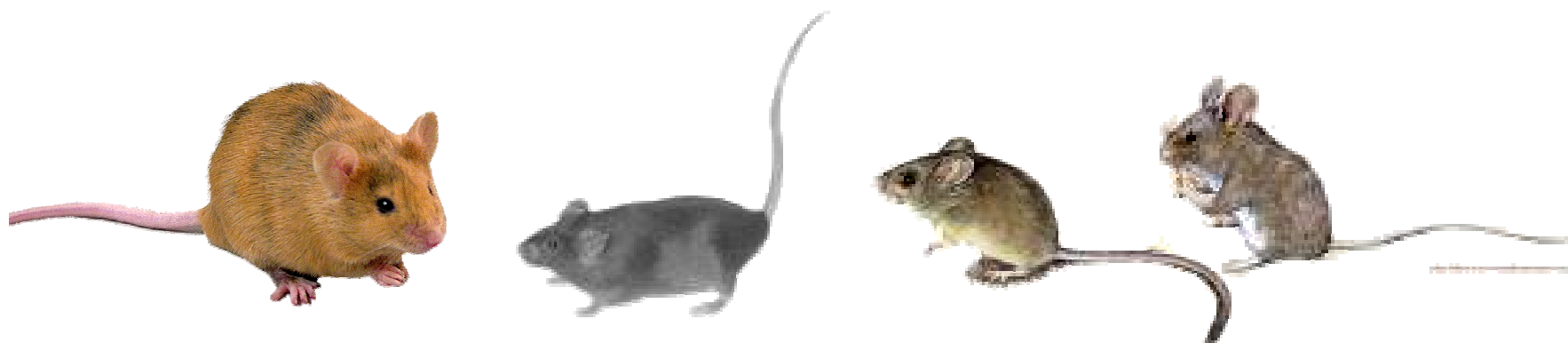
Критерии согласия (Tests for goodness of fit):

1. качественные и ранговые данные:

то, чего нет в Statistica 6.0

Если у вас несколько выборок, а гипотеза для них одинаковая, рекомендуют протестировать их на **гетерогенность**, прежде чем суммировать.

Считаются значения χ^2 для всех выборок и для общей выборки, складываются и сравниваются с табличкой.



Критерии согласия (Tests for goodness of fit):

1. качественные и ранговые данные:

то, чего нет в Statistica 6.0

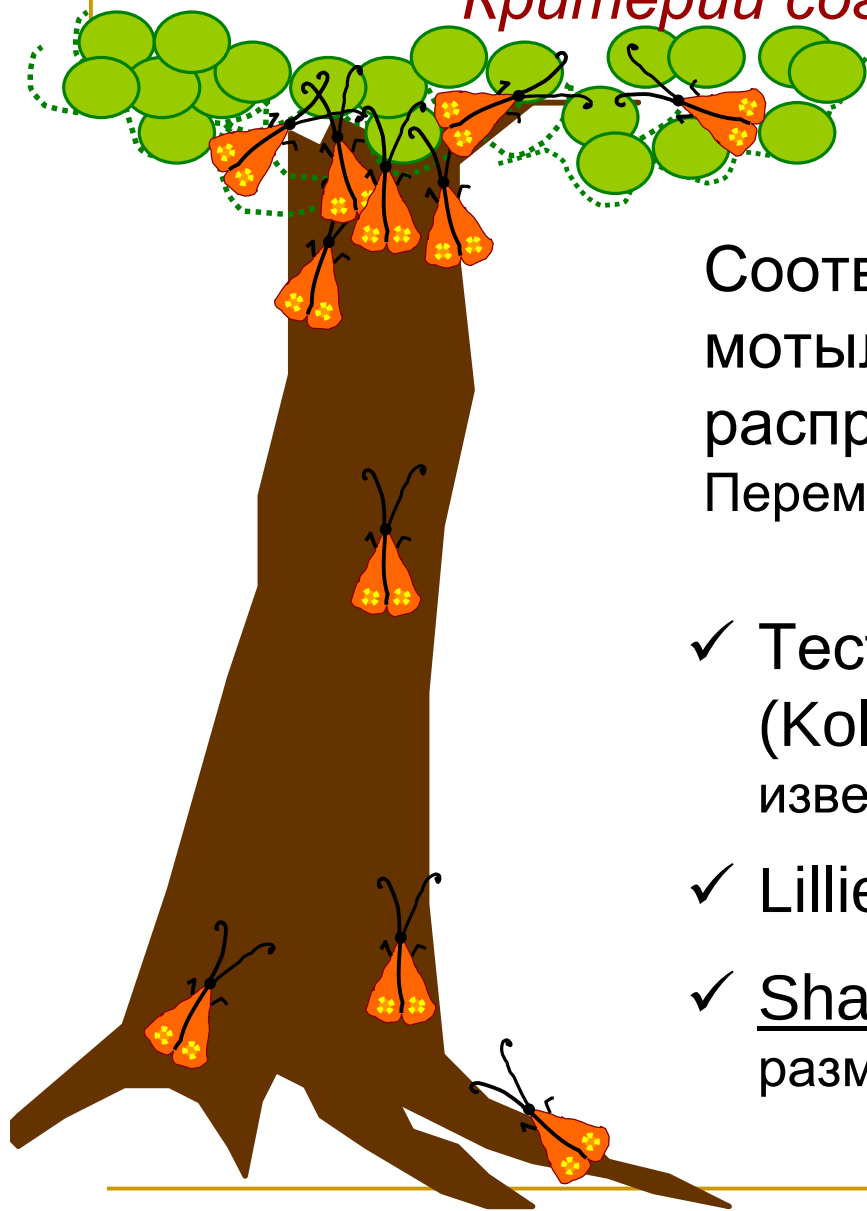


Замечательный тест Колмогорова-Смирнова для ранговых данных (Kolmogorov-Smirnov goodness of fit for discrete ordinal scale data).

35 кошек выбирают из 5 типов корма, различающихся по влажности. Случаен ли выбор или есть предпочтения?



Критерии согласия (Tests for goodness of fit): 2. непрерывное распределение

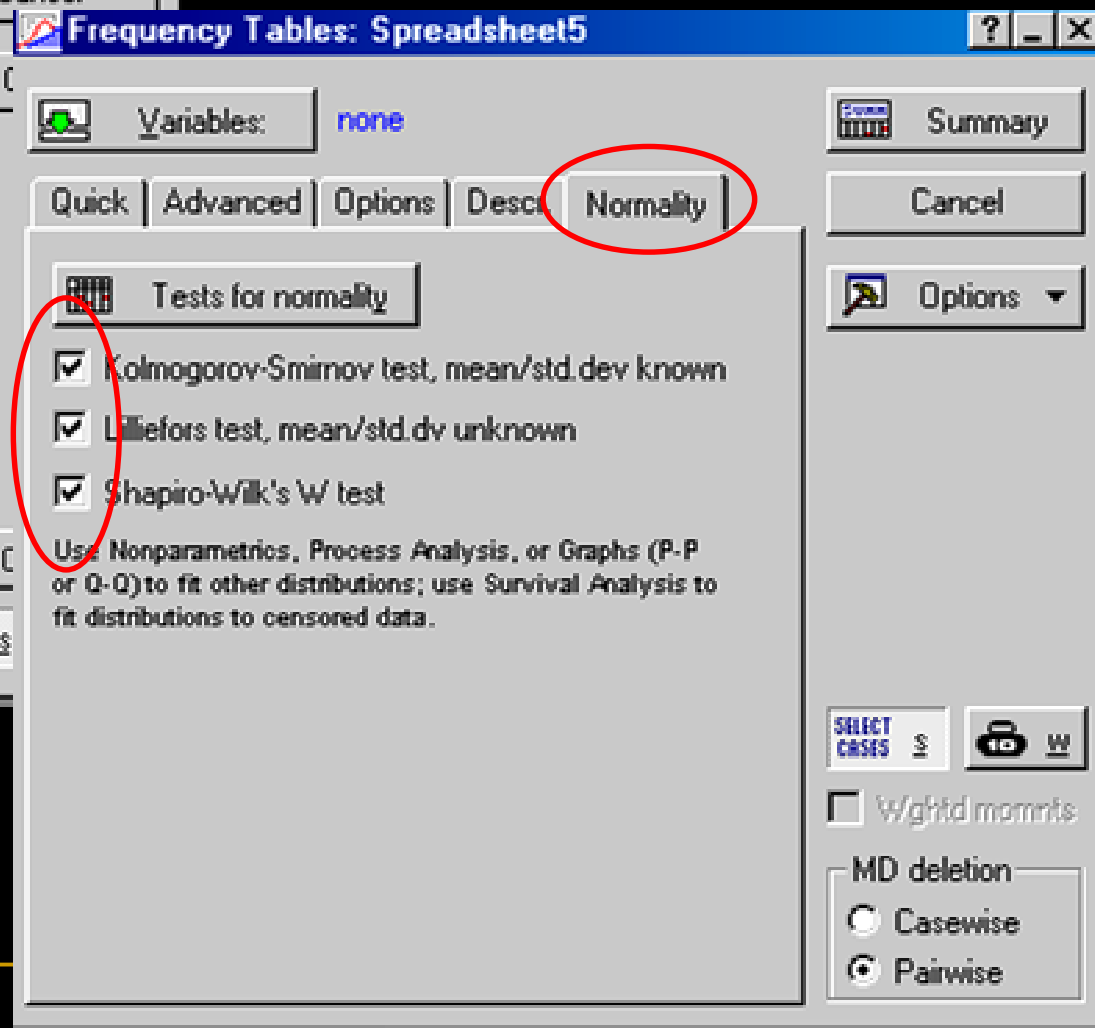
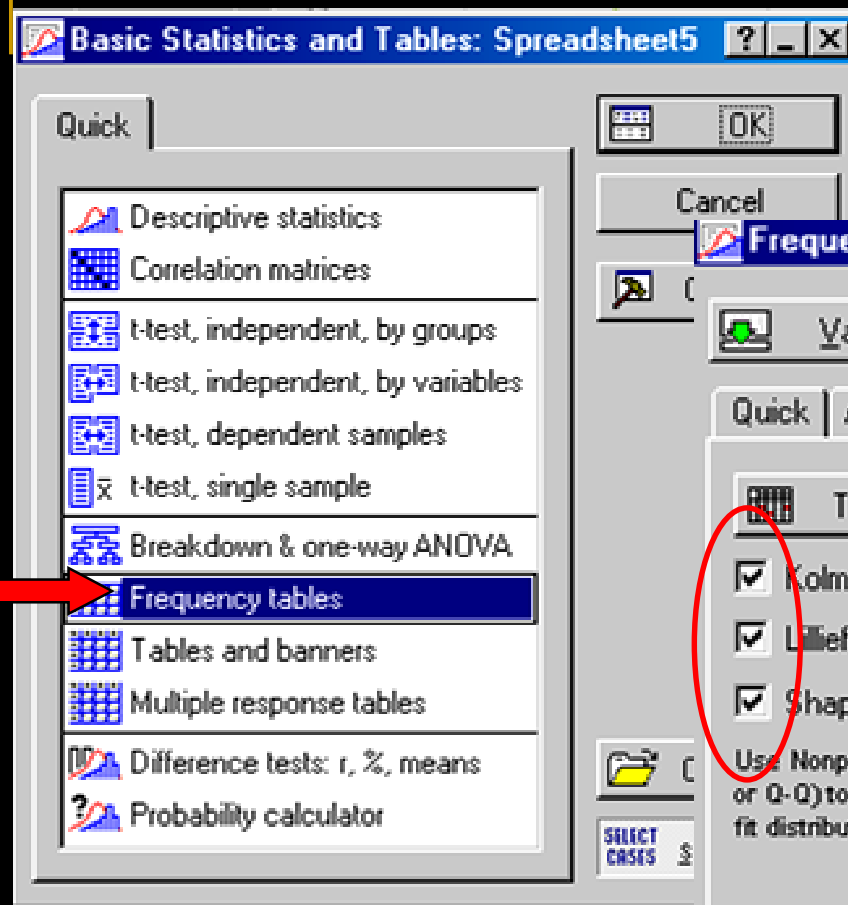


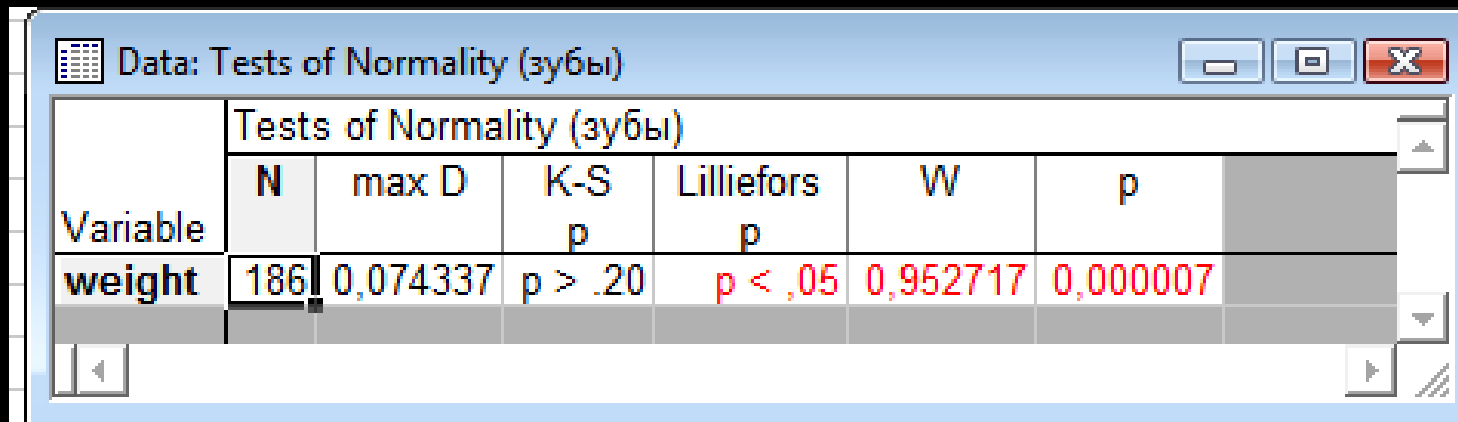
Соответствует ли распределение мотыльков на дереве нормальному распределению?

Переменная – высота от земли в метрах

- ✓ Тест Колмогорова-Смирнова (Kolmogorov-Smirnov test) (если известны дисперсия и среднее в популяции)
- ✓ Lilliefors test
- ✓ Shapiro-Wilk's W test (самый мощный, размер выборки до 5000)

Проверка распределения на нормальность





The screenshot shows a window titled "Data: Tests of Normality (зубы)". Inside, there is a table with the following data:

Variable	N	max D	K-S p	Lilliefors p	W	p
weight	186	0,074337	p > .20	p < ,05	0,952717	0,000007

маленькое p говорит о том, что данные не соответствуют нормальному распределению.

Сравнение с другими теоретическими распределениями:

Тест Колмогорова-Смирнова для непрерывных распределений

The image shows the SPSS 'Distribution Fitting: Spreadsheet1' dialog box and the resulting 'Data' table. The dialog box is set to 'Rectangular' distribution for the variable 'distance'. The 'Kolmogorov-Smirnov test' is selected with the 'Yes (continuous)' option. The 'Data' table displays the observed and expected frequencies for the rectangular distribution.

Distribution Fitting: Spreadsheet1

Quick | Parameters | Options

Distribution: Rectangular

Variable: distance

Kolmogorov-Smirnov test

- ☐ No
- ☐ Yes (categorized)
- ☒ Yes (continuous)

Chi-Square test

- ☒ Combine Categories

Graph

Plot distribution

- ☒ Frequency distribution
- ☐ Cumulative distribution

Plot raw frequencies or %

- ☒ Raw frequencies

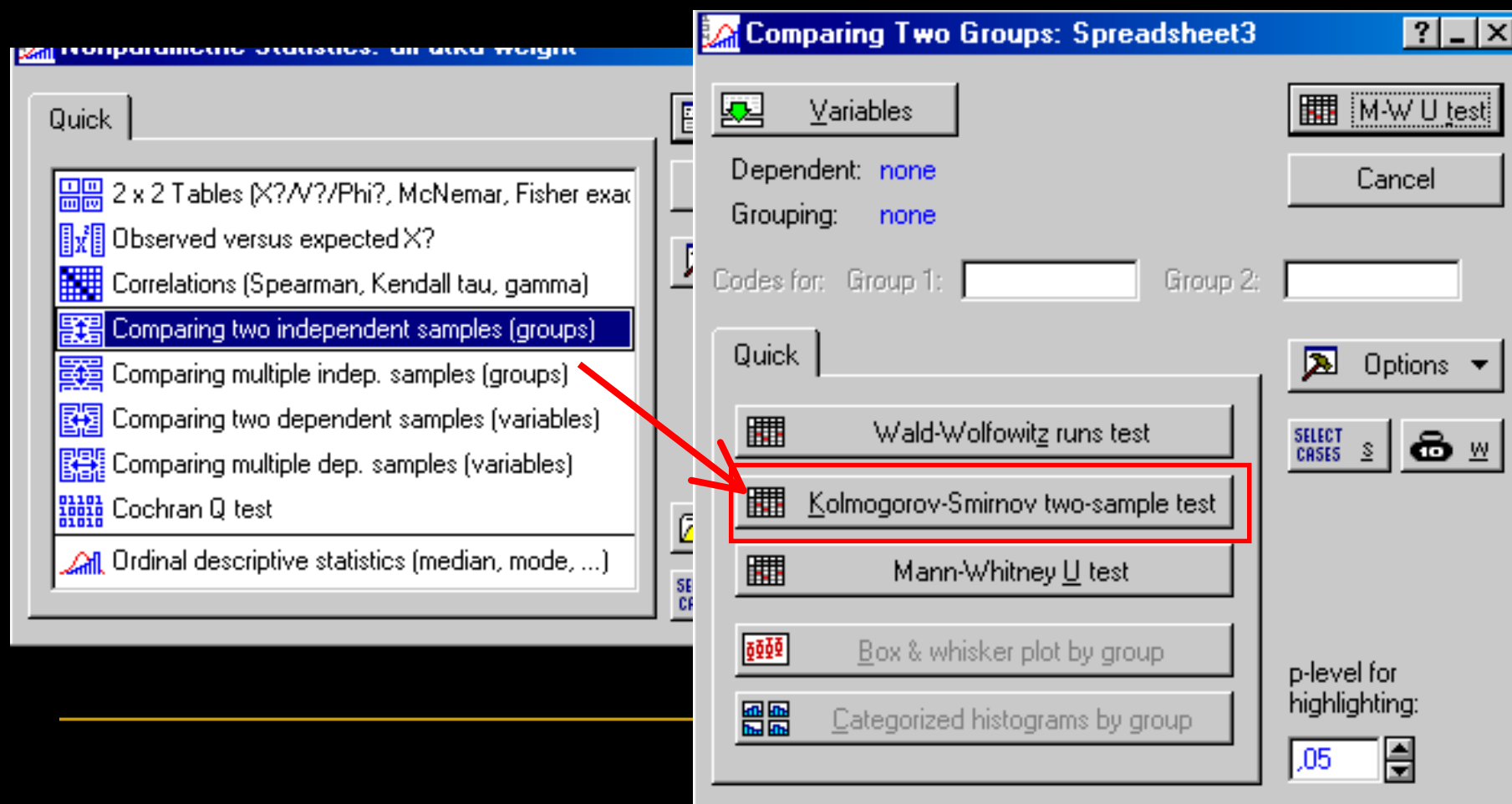
Data: Variable: distance, Distribution: Rectangular (Spreadsheet1)

Variable: distance, Distribution: Rectangular (Spreadsheet1)
Kolmogorov-Smirnov d = 0,10991, p = n.s.
Chi-Square = 1,74952, df = 1 (adjusted), p = 0,18594

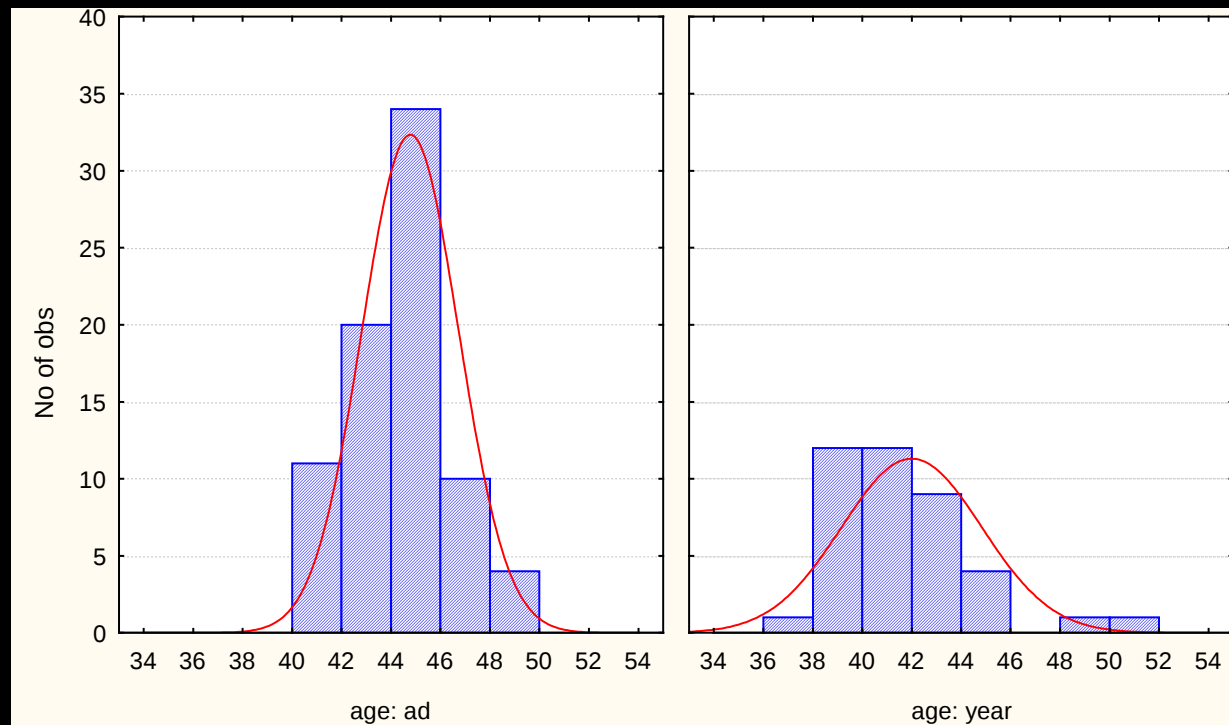
Upper Boundary	Observed Frequency	Cumulative Observed	Percent Observed	Cumul. % Observed	Expected Frequency	Cumulative Expected	Percent Expected	Cumul. Expected
<= 0,00000	0	0	0,00000	0,0000	0,000000	0,00000	0,00000	0,00
0,50000	1	1	3,33333	3,3333	1,621622	1,62162	5,40541	5,40
1,00000	5	6	16,66667	20,0000	4,054054	5,67568	13,51351	18,91
1,50000	6	12	20,00000	40,0000	4,054054	9,72973	13,51351	32,43
2,00000	4	16	13,33333	53,3333	4,054054	13,78378	13,51351	45,94
2,50000	5	21	16,66667	70,0000	4,054054	17,83784	13,51351	59,45
3,00000	4	25	13,33333	83,3333	4,054054	21,89189	13,51351	72,97
3,50000	2	27	6,66667	90,0000	4,054054	25,94595	13,51351	86,48

Если мы хотим сравнить *между собой два непрерывных распределения*

Тест Колмогорова-Смирнова
(Kolmogorov-Smirnov two-sample test)



Этот тест учитывает и форму распределения, и разницу в средних значениях



Data: Kolmogorov-Smirnov Test (зубы)*

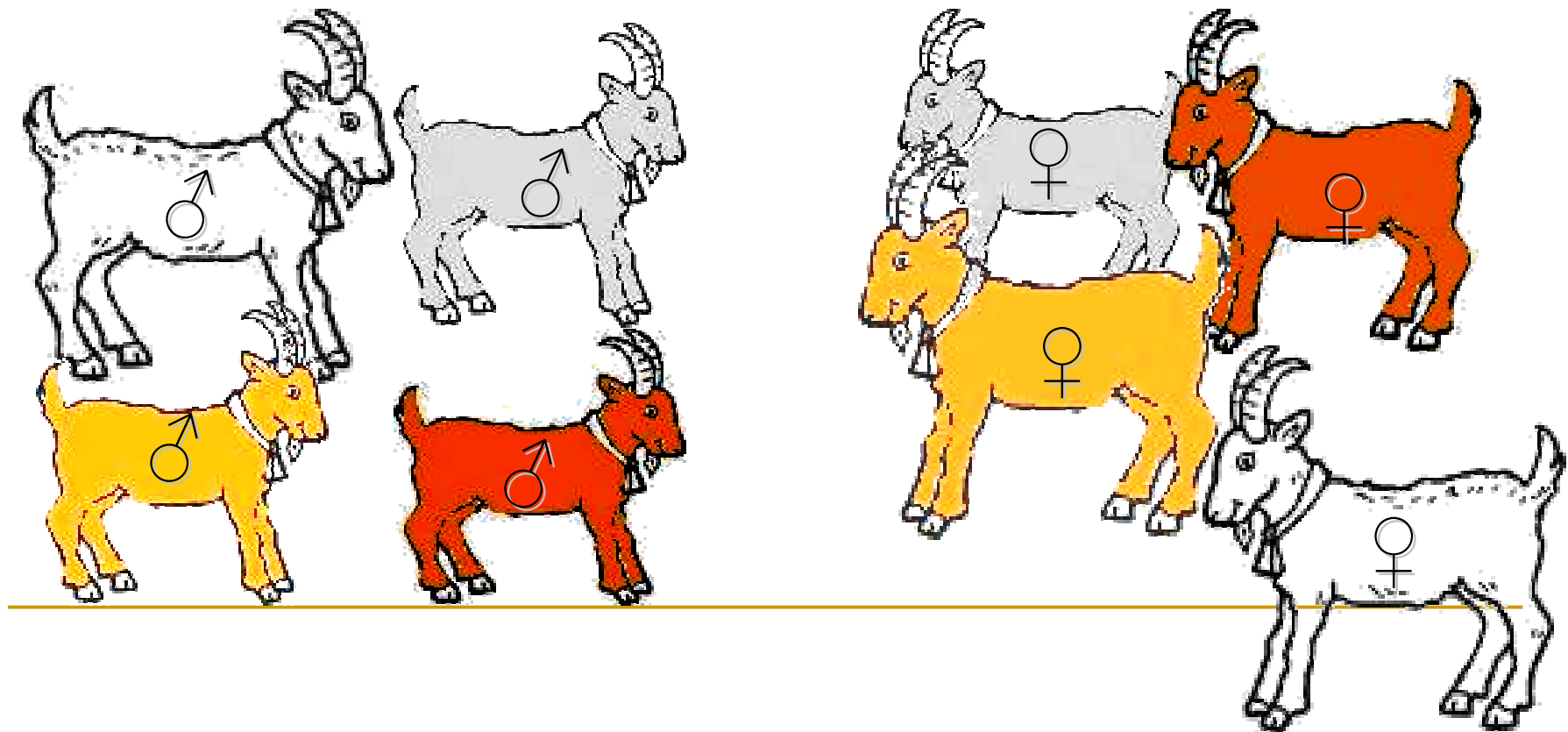
Kolmogorov-Smirnov Test
By variable age
Marked tests are significant at $p < .05000$

variable	Max Neg Differnc	Max Pos Differnc	p-level	Mean ad	Mean year	Std.Dev. ad	Std.Dev. year	Valid N ad	Valid N year
zygomatic	-0,025000	0,549684	$p < .001$	44,71772	41,95250	1,948948	2,822436	79	40

Проверка независимости действия нескольких категориальных факторов

Таблицы вида $a \times b$. Общая H_0 гипотеза: частоты в строчках не зависят от частот в столбцах.

Связаны ли пол и цвет у коз?



п о л	б е л ы е	к р а с н ы е	ж ё л т ы е	с е р ы е	В с е г о
с а м ц ы	32	43	16	9	100
с а м к и	55	65	64	16	200
в с е г о	87	108	80	25	300

H_0 : цвет меха у коз не зависит от пола в популяции коз;

H_1 : цвет меха у коз зависит от пола в популяции коз.

Критерий χ^2 (Chi-square analysis of contingency tables)

Пример из жизни сусликов:

Связаны ли категории социальных контактов (как контактирует) с типом партнёра?



в таблице должны быть сырые данные

Data: Spreadsheet5* (2v by 64c)

	1 type	2 distribution
1	1	1
2	1	1
3	2	1
4	2	1
5	2	1
6	2	1
7	3	1
8	3	1
9	3	1
10	3	1
11	3	1
12	3	1
13	3	1
14	4	1
15	4	1
16	4	1
17	4	1
18	4	1
19	4	1
20	4	1

Basic Statistics and Tables: Spreadsheet5

Quick

- Descriptive statistics
- Correlation matrices
- t-test, independent, by groups
- t-test, independent, by variables
- t-test, dependent samples
- t-test, single sample
- Breakdown & one-way ANOVA
- Frequency tables
- Tables and banners**
- Multiple response tables
- Difference tests: r, %, means
- Probability calculator

OK

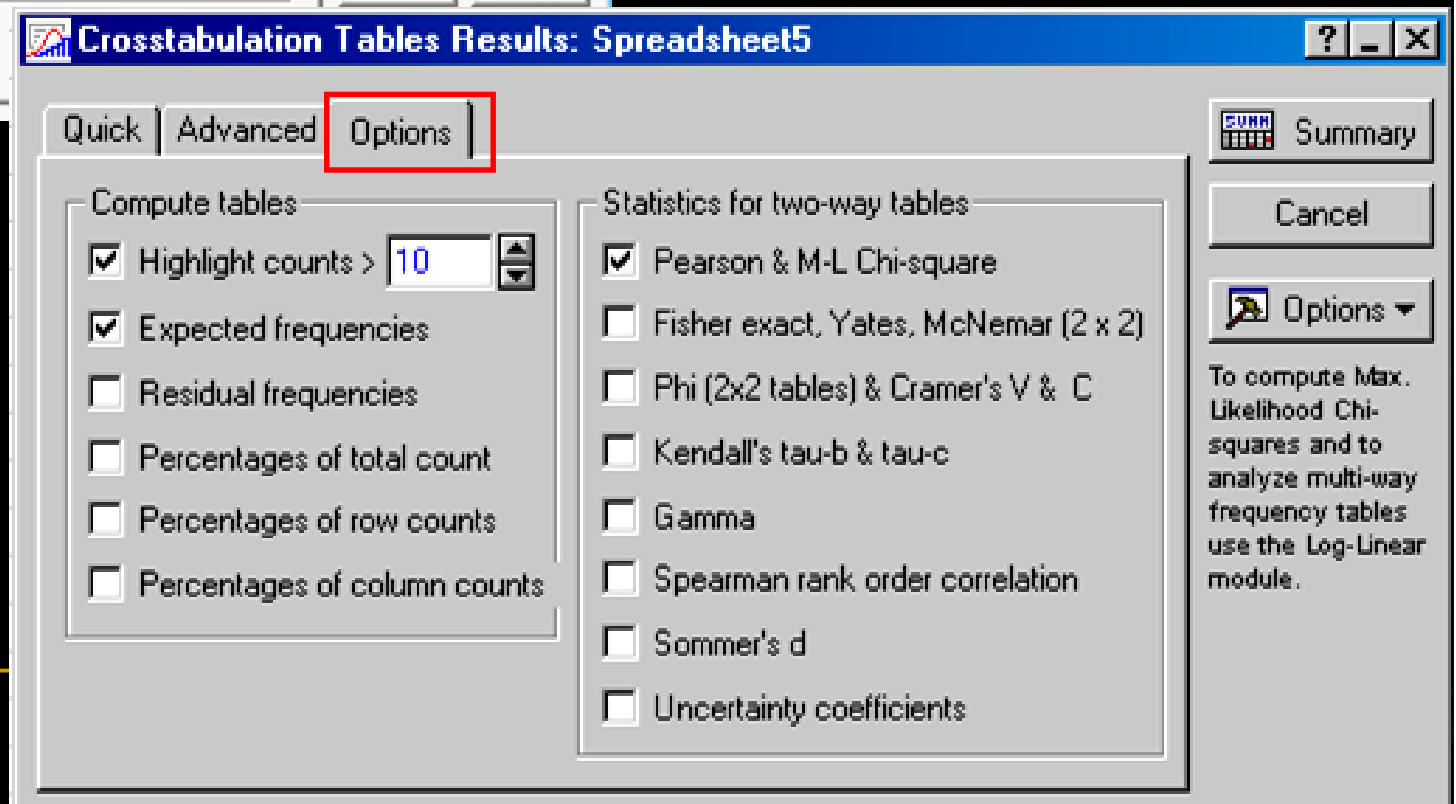
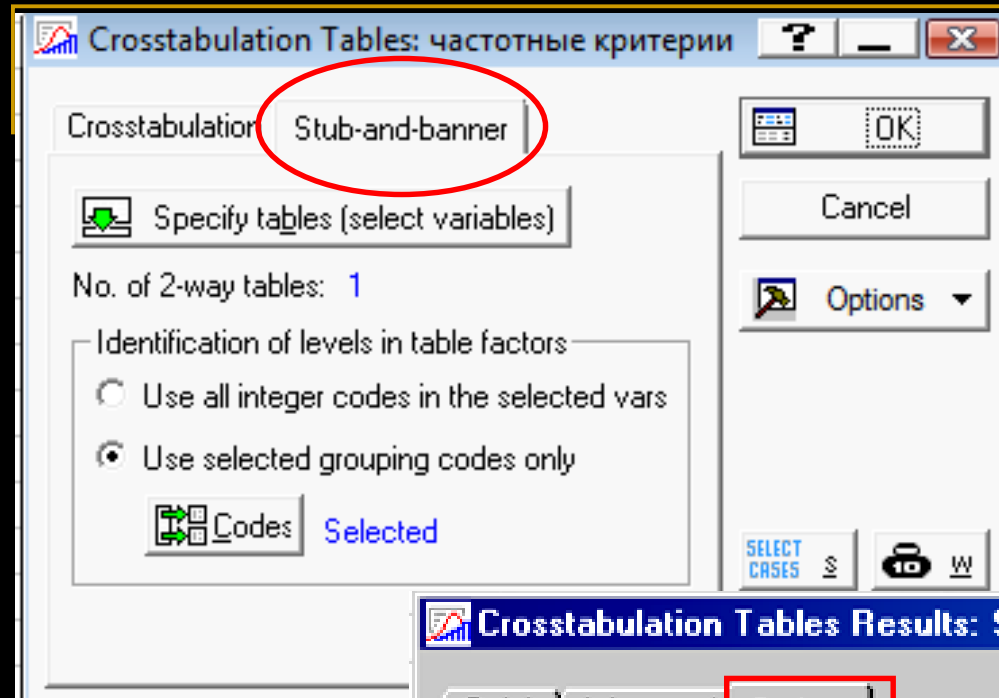
Cancel

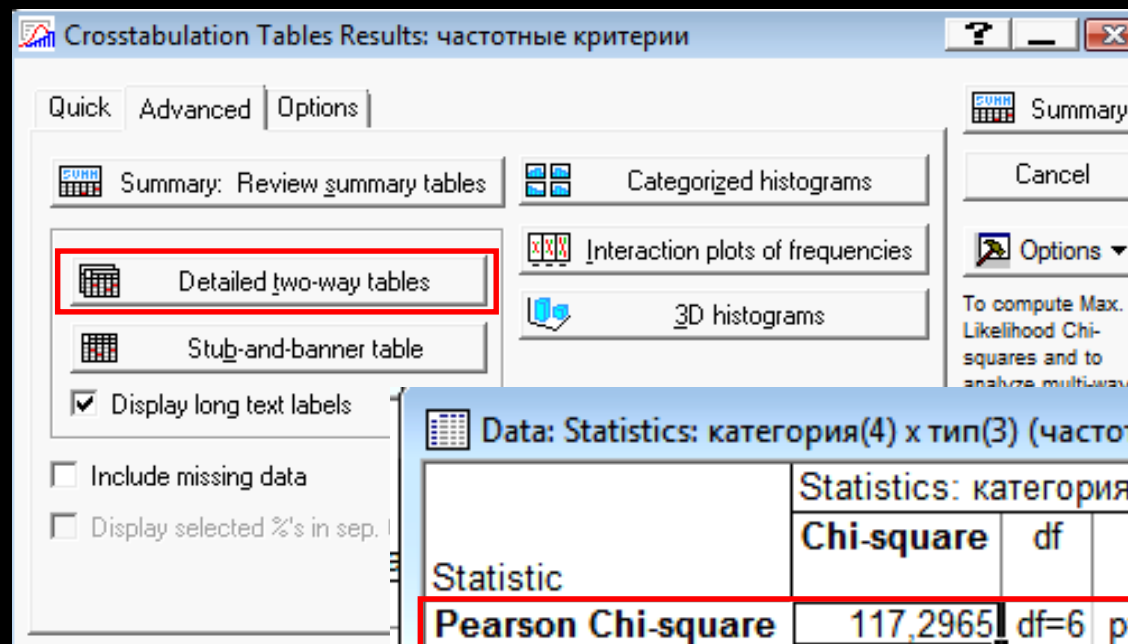
Options

Open Data

SELECT CASES

10 W





Statistics: категория(4) x тип(3) (частотные критерии)				
Statistic	Chi-square	df	p	
Pearson Chi-square	117,2965	df=6	p=0,0000	
M-L Chi-square	109,3888	df=6	p=0,0000	

Отвергаем нулевую гипотезу об отсутствии взаимодействия между переменными

В табличке с частотами вида $a \times b$ не должно быть значений меньше 5. Если это не так, следует объединить какие-нибудь проявления признака.

Zar, 1999:

Если вы не отвергли связь переменных (!), а теперь хотите показать, из-за какой именно категории есть связь, можно отдельно проверить связь переменных на остальных категориях, а затем – отношение этой категории к остальным.

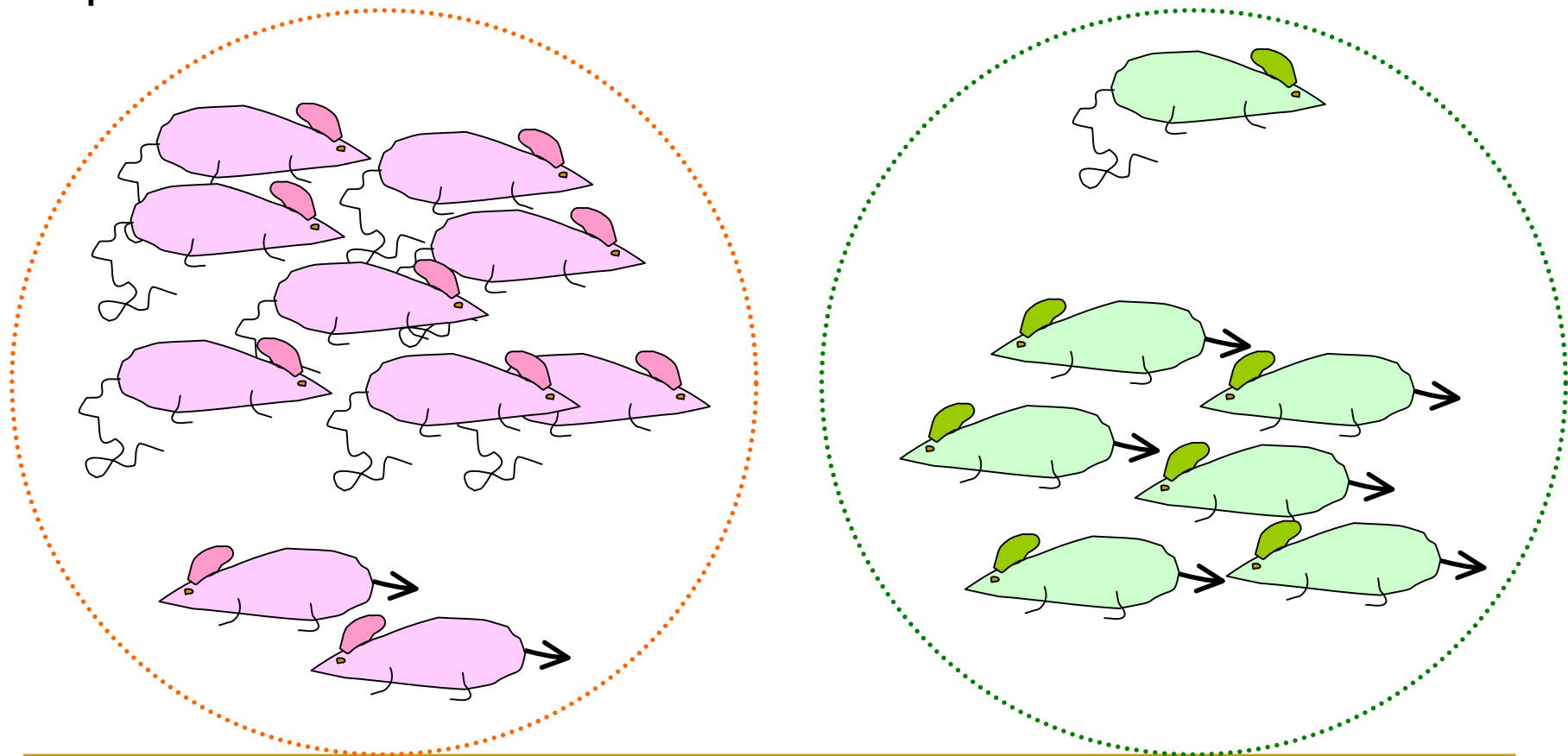
Например, если самцы и самки коз отличаются, по-видимому, только по соотношению белых коз, можно:

1. исключить белых, проверить связь пола и цвета для остальных;
 2. проверить связь пола и присутствия белого цвета у козы.
-

Проверка независимости действия категориальных факторов:

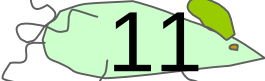
Четырёхпольные таблицы (2 x 2 tables).

Есть только 2 фактора, у каждого – только по 2 проявления.



Связан ли цвет мышей с формой их хвостов??

Четырёхпольные таблицы (2 x 2 table)

	хвост 	хвост → 	
роз 	18	12	29
зел 	11	26	38
	29	38	67

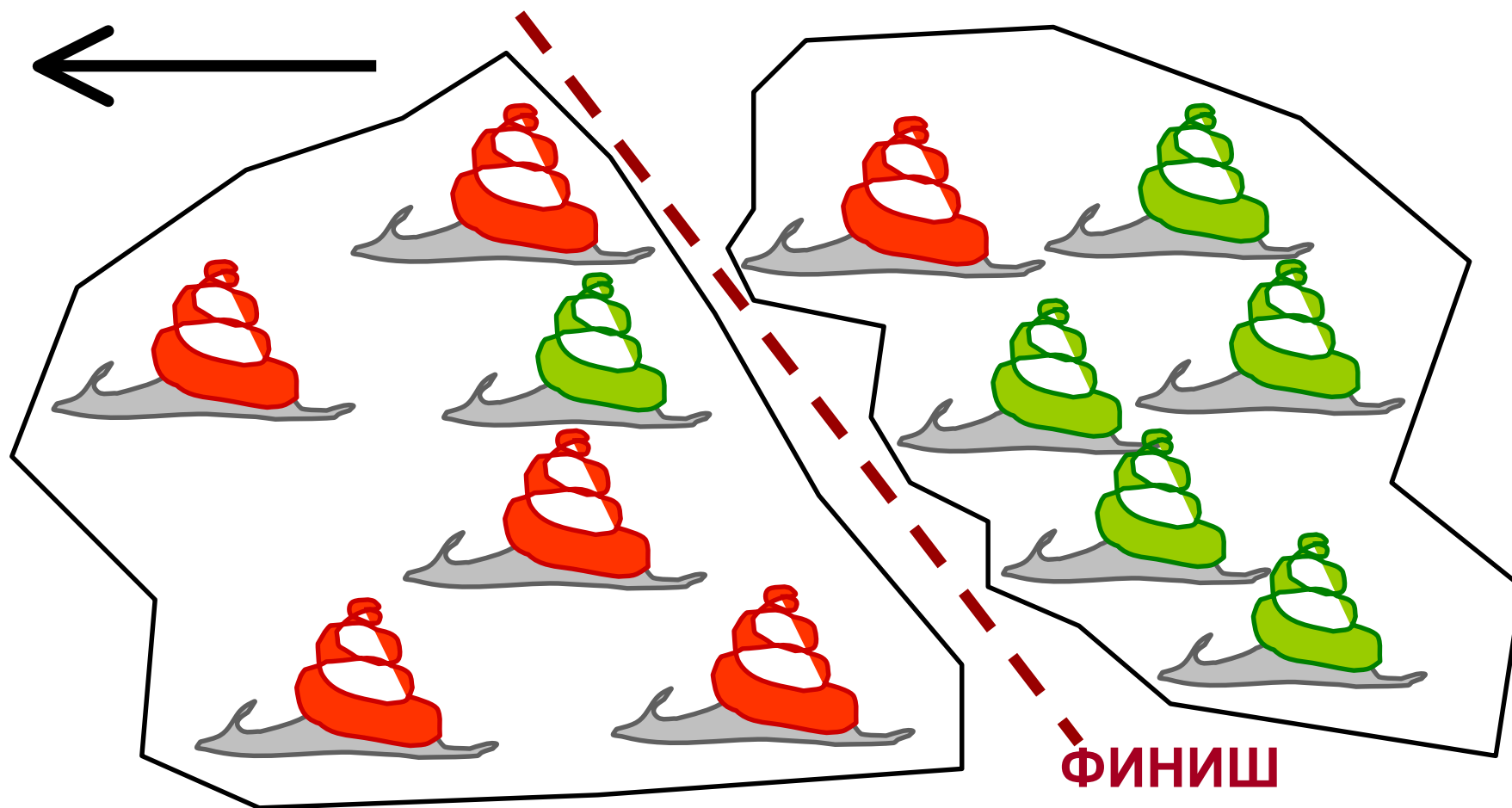
Модель 1: мы задаём только общий размер выборки

Модель 2: одна из сумм фиксирована (половине крыс дали яд, половине – нет и сравниваем выживаемость)

Модель 3: фиксированы обе суммы (про улиток)

Обычно мы имеем дело с моделями 1 и 2.

Пояснение к Модели 3 – красных и зелёных улиток по 6 штук, соревнование продолжалось до тех пор, пока половина улиток не перешла линию финиша



Для независимых выборок

Критерий χ^2 (Chi-square) с поправкой Йейтса.

Если в табличке сырые данные, а не готовая четырёхпольная таблица – Tables and Banners.
Если готовая таблица – 2 x 2 tables.

Принцип введения поправки – тот же, что для сравнения наблюдаемых и ожидаемых частот, делает тест более консервативным.

Не нужна для больших выборок. В Statistica: поправку вводят, если хотя бы одна частота меньше 10.

Точный критерий Фишера (Fisher exact test)

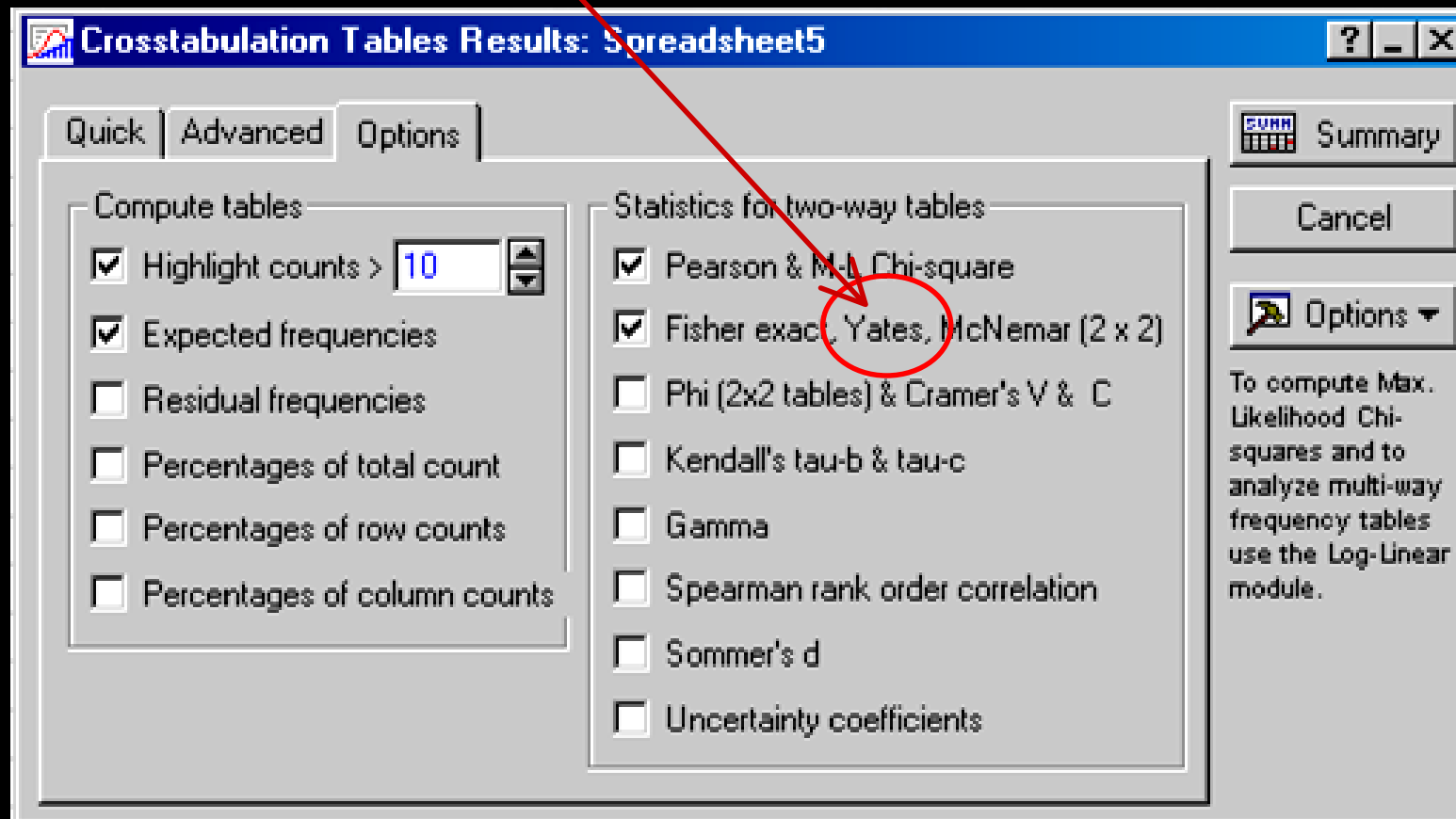
Годится, если одна из частот меньше 5 и вообще, для небольших выборок. Подходит для 3-й модели.

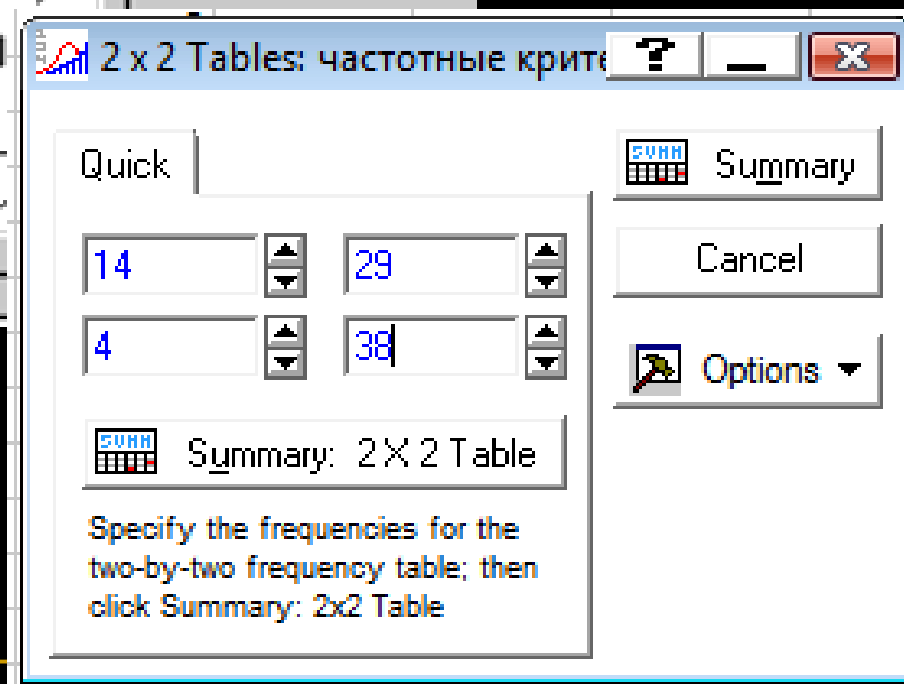
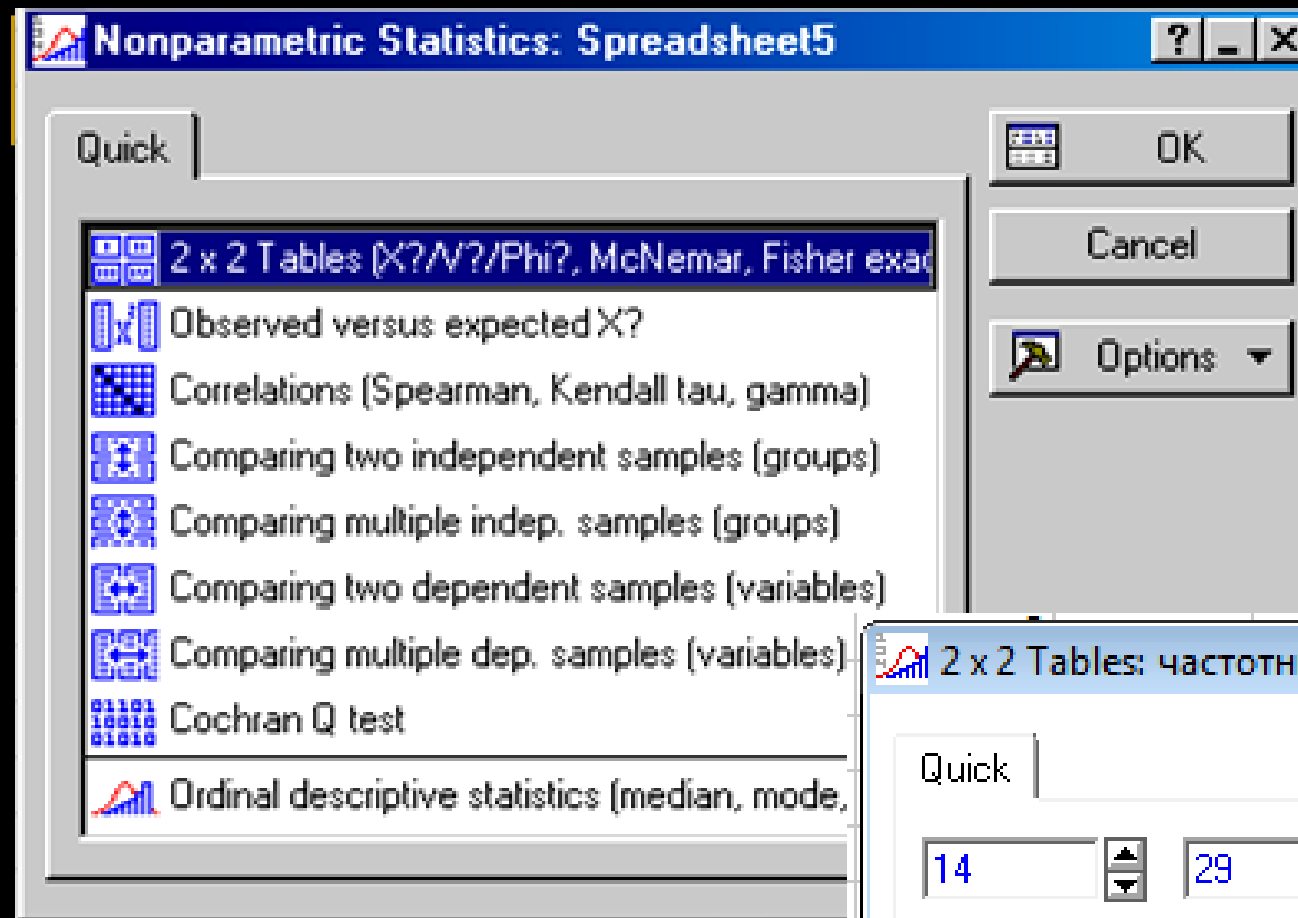
<i>скунсы</i>	<i>с бешенством</i>	<i>без бешенства</i>
восточные	14	29
западные	5	38

H_0 : район, где живёт скунс, и заболеваемость, не связаны друг с другом;

H_1 : между районом и заболеванием есть связь.







Отвергаем H_0

Data: 2 x 2 Table (частотные критерии)			
	2 x 2 Table (частотные критерии)		
	Column 1	Column 2	Row Totals
Frequencies, row 1	14	29	43
Percent of total	16,471%	34,118%	50,588%
Frequencies, row 2	4	38	42
Percent of total	4,706%	44,706%	49,412%
Column totals	18	67	85
Percent of total	21,176%	78,824%	
Chi-square (df=1)	6,75	p= ,0094	
V-square (df=1)	6,67	p= ,0098	
Yates corrected Chi-square	5,44	p= ,0196	
Phi-square	,07946		
Fisher exact p, one-tailed		p= ,0089	
two-tailed		p= ,0155	
McNemar Chi-square (A/D)	10,17	p= ,0014	
Chi-square (B/C)	17,45	p= ,0000	

Скунсы из разных районов имеют разную заболеваемость.

Замечание: тест в данном случае двусторонний!!

Односторонний тест Фишера:

Для случаев, когда мы заранее знаем, куда может отклониться соотношение частот.

Например, мы даём лекарство больным зверям и сравниваем, сколько из них выздоровело по сравнению с контрольной группой.

Предполагается, что лекарство не может ухудшить состояние зверей, а только может либо вылечить, либо нет.



Phi-square – показатель корреляции между качественными переменными.

V-square – разновидность χ^2 теста.

Все эти тесты подразумевали, что выборки независимы (например, каждая особь входит только в одну из ячеек).

Для связанных выборок:

Критерий Мак-Немара (McNemar Chi-square)

Например: мы провели в сентябре контрольную по математике. Из 100 учеников, 18 написали на положительную оценку, остальные – двоечники. Для тех же учеников мы провели контрольную во 2-й четверти. Нас интересует, как изменилась успеваемость.

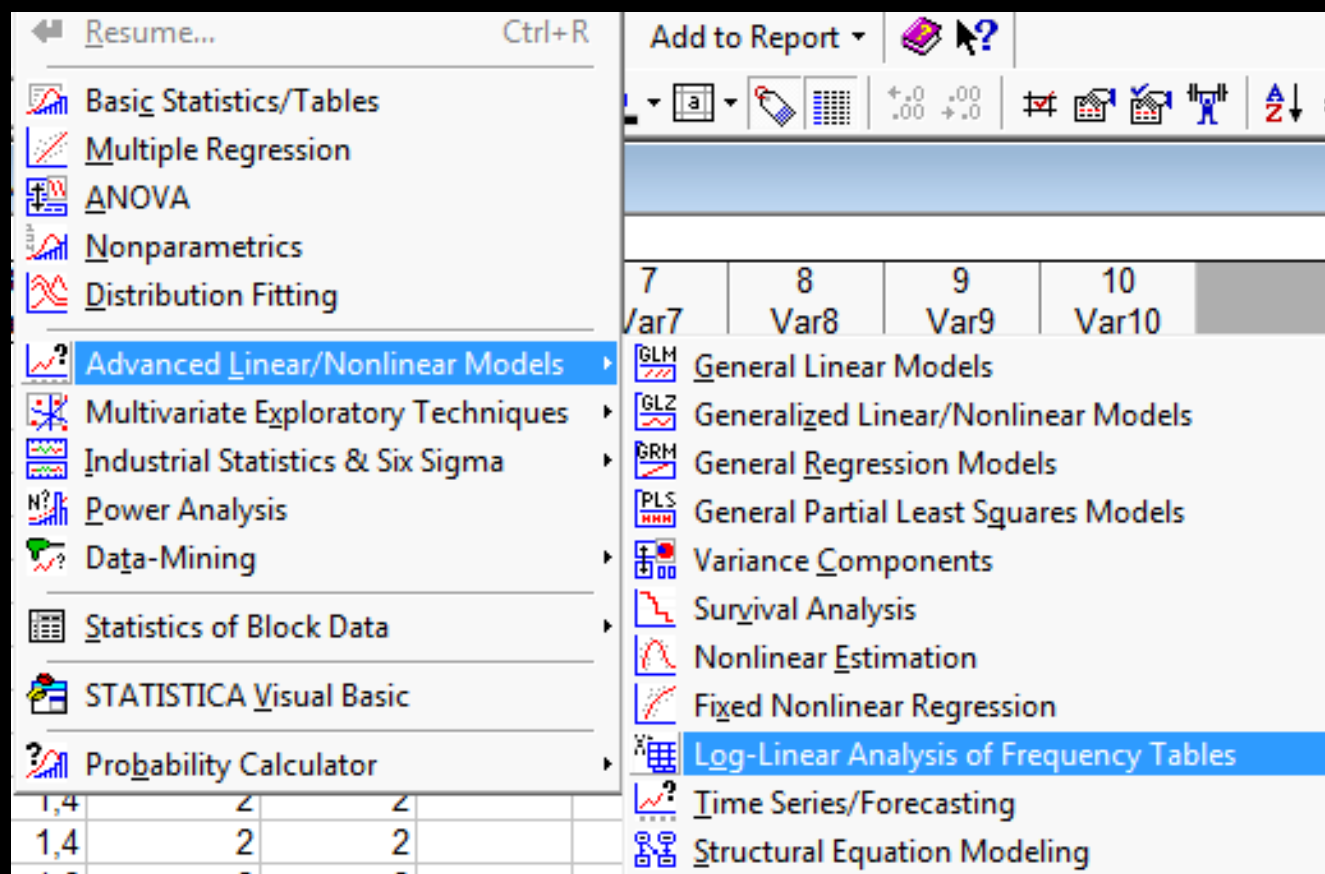


Требуется специальная организация таблицы!!!

Контрольная в сентябре	Контрольная весной	
	Получили 2	Получили 3 и выше
Получили 2 (82)	30	52
Получили 3 и выше (18)	6	12

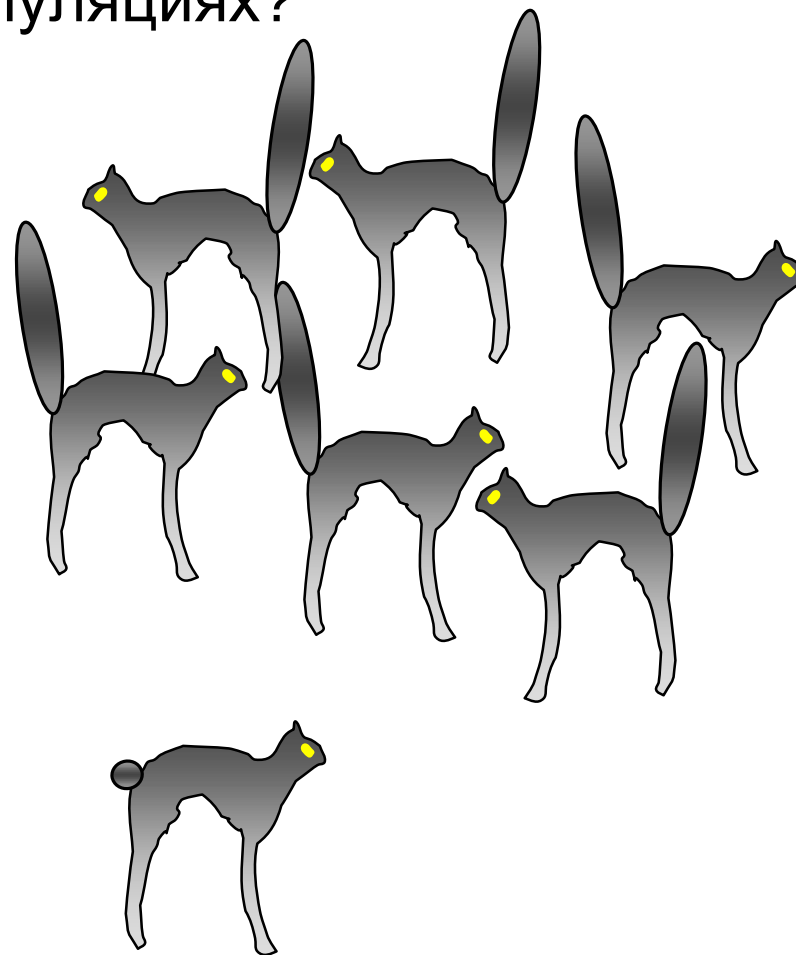
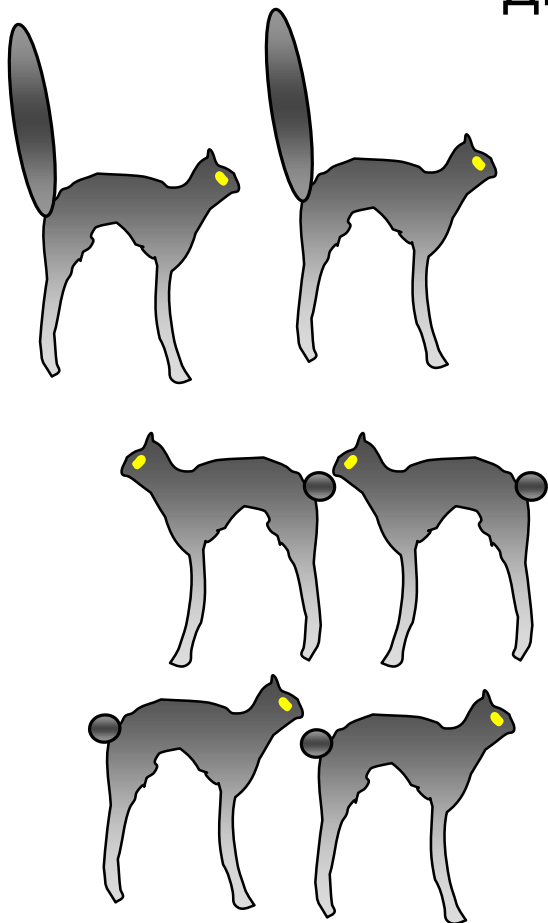
Тест МакНемара сравнивает числа под и над главной диагональю. Поэтому нельзя менять порядок чисел, когда мы вносим их в Статистику.

Сложные многомерные варианты, продвинутый частотный анализ



Сравнение пропорций

Отличается ли доля короткохвостых кошек в двух разных популяциях?



Basic Statistics and Tables: Spreadsheet5

Quick

- Descriptive statistics
- Correlation matrices
- t-test, independent, by groups
- t-test, independent, by variables
- t-test, dependent samples
- t-test, single sample
- Breakdown & one-way ANOVA
- Frequency tables
- Tables and banners
- Multiple response tables
- Difference tests: r , $\%$, means**
- Probability calculator

OK

Cancel

Options

Difference tests: r , $\%$, means: Spreadsheet5

☐ Send/print results for each Compute to Report window

Cancel

Difference between two correlation coefficients

r_1 : 0,00

N_1 : 10

r_2 : 0,00

N_2 : 10

p : 1,0000

☐ One-sided

☒ Two-sided

Compute

Difference between two means (normal distribution)

M_1 : 0,

StDv 1: 1,

N_1 : 10

p : 1,0000

M_2 : 0,

StDv 2: 1,

N_2 : 10

☐ One-sided

☒ Two-sided

Compute

☐ Single mean 1 vs .population mean 2

Difference between two proportions

Pr_1 : ,50

N_1 : 10

Pr_2 : ,50

N_2 : 10

p : 1,0000

☐ One-sided

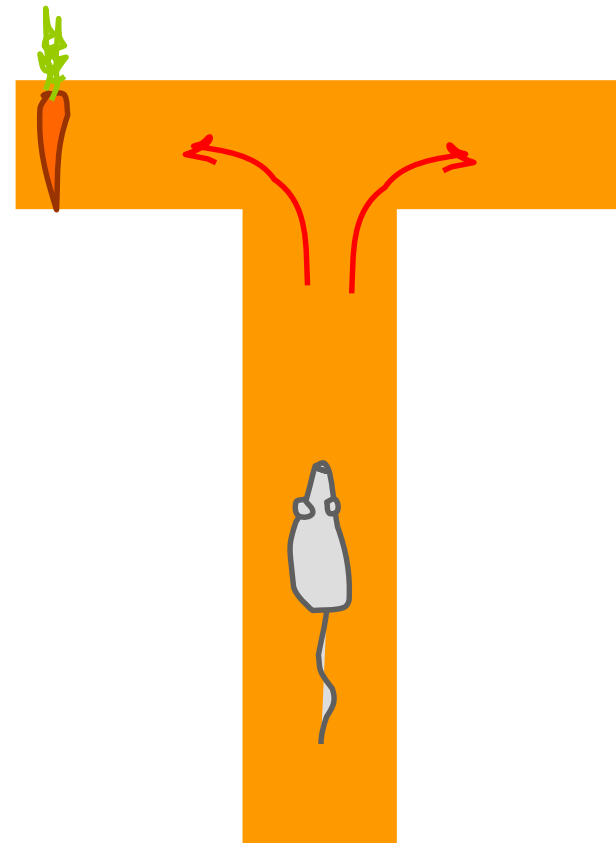
☒ Two-sided

Compute

Биномиальный тест

У нас было 10 мышей;
7 пошли налево, 3
направо.

Действительно ли мышь
распознаёт морковку на
расстоянии?



*У нас всего 2 числа, и мы должны их сравнить с
теоретическими частотами при случайном выборе*