Занятие 8

# Частотный анализ

Все критерии, которые мы изучили, анализировали влияние разных факторов на КОЛИЧЕСТВЕННУЮ (или ранговую) зависимую переменную (или просто взаимосвязь количественных переменных - корреляцию).

Как быть с анализом взаимосвязей между КАЧЕСТВЕННЫМИ ПЕРЕМЕННЫМИ?

*Например, как проверить:*

✓ зависит ли кудрявость волос (кудрявые или нет) от их цвета (рыжие, тёмные, светлые)?

✓ соответствует ли соотношение полов в популяции 1:1?

✓ какие марки машин предпочитают женщины, а какие – мужчины?

✓ зависит ли присутствие антител к токсоплазмозу от вида, пола и возраста у кошачьих?

✓ зависит ли для зверька вероятность дожить до следующего года от массы тела, упитанности, площади участка? (лекция 12)

Очевидно, ни одна из изученных нами моделей для таких переменных **не годится**:

1. У них **нельзя посчитать среднее**, дисперсию, стандартное отклонение и пр.;
2. К шкале качественной переменной обычно **неприменимы понятия «больше» и «меньше»**;
3. Очевидно, распределение качественной переменной какое-то совершенно **не нормальное**.

Принципы анализа качественных данных другие:

✓Наши данные – **количества наблюдений** для каждой комбинации переменных;

✓ удобно представлять в виде **таблицы частот**;

✓ часто невозможно определить, какая переменная зависимая;

✓ основная статистика – **статистика** $\chi^2$, оперирует **частотами**

Для анализа качественных переменных нужен

# частотный анализ

Начнём с КРИТЕРИЕВ СОГЛАСИЯ *(tests for goodness of fit)*

Решают вопрос, соответствует ли распределение в популяции, из которой получена выборка, теоретическому распределению (которое мы сами определяем).

- ✓Соответствует ли соотношение полов в популяции 1:1?
- ✓ можно ли считать, что *M&M's* каждого цвета кладут в пачки поровну?
- ✓ соответствует ли соотношение разных горохов у Менделя 1:3:3:9?

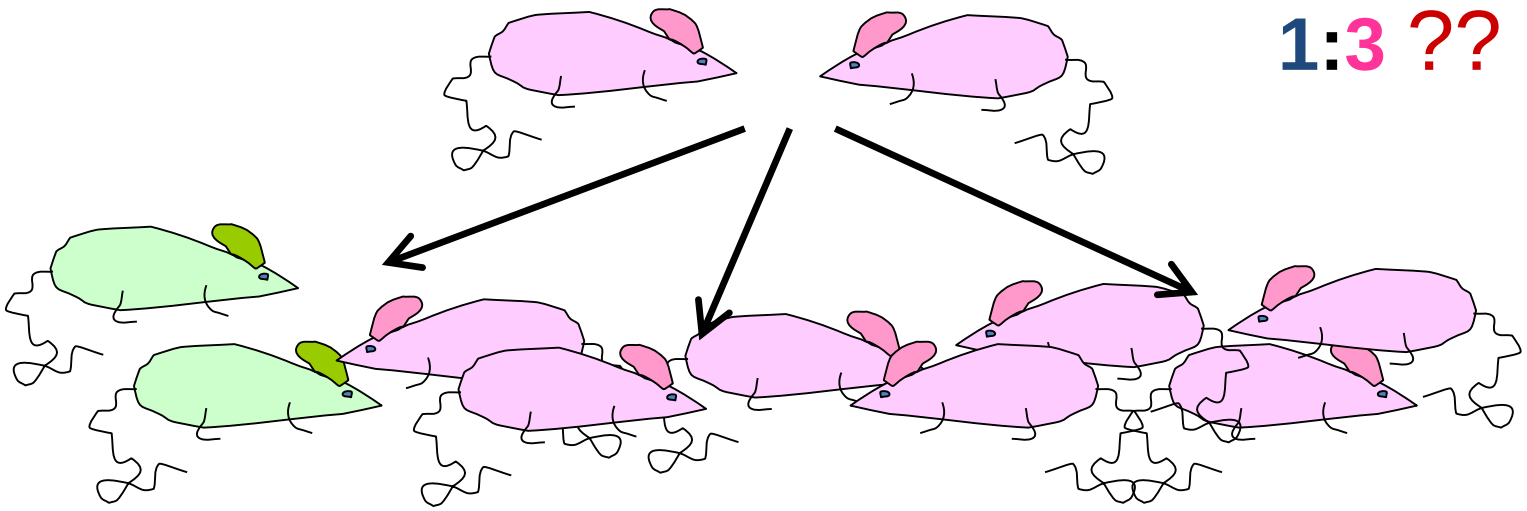
Karl Pearson
Придумал $\chi^2$ статистику ещё в 1900 году!

Пример с игральной костью: как проверить, не кривая ли она? Очевидно, что бросая её 120 раз маловероятно получить ровно по 20 бросков на каждую сторону. **Насколько же допустимы различия?**

Родились в F2:
16 зелёных мыши и 84 розовых.

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

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




Очевидно, практически невозможно, чтобы соотношение фенотипов соответствовало в точности 1:3! Нужно понять, насколько большое отклонение от 1:3 мы можем считать случайным.

# Понятие о частотах

(тесно связано с понятиями о вероятностях):

| | розовые | зелёные | всего |
|---|---|---|---|
| $O_i$ | **84** | **16** | 100 |
| $E_i$ | *75* | *25* | 100 |

**Наблюдаемые частоты** (observed) – просто количества наблюдений (объектов) в каждой категории. *Для розовых это – 84.*

**Ожидаемые частоты** (expected) – какими были бы количества наблюдений в каждой категории, если бы $H_0$ была верна. *Для розовых это – 75, т.к. считается с учётом общего N.*

**Вероятность** (probability) – популяционный параметр; здесь это вероятность, того, что объект будет принадлежать к данной категории. *Для категории розовых это – 0.75, если $H_0$ верна.*

*Заметим, что речь идёт только о частотах, никакие параметры распределения не упоминаются.*

# Критерий $\chi^2$ Пирсона (**Pearson Chi-square test**)

| | розовые | зелёные | всего |
|---|---|---|---|
| $O_i$ | **84** | **16** | 100 |
| $E_i$ | *75* | *25* | 100 |

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

*df* = *k*-1=2-1=1

O – observed
E - expected

$$\chi^2 = \frac{(84-75)^2}{75} + \frac{(16-25)^2}{25} = 1.080 + 3.240 = 4.320$$

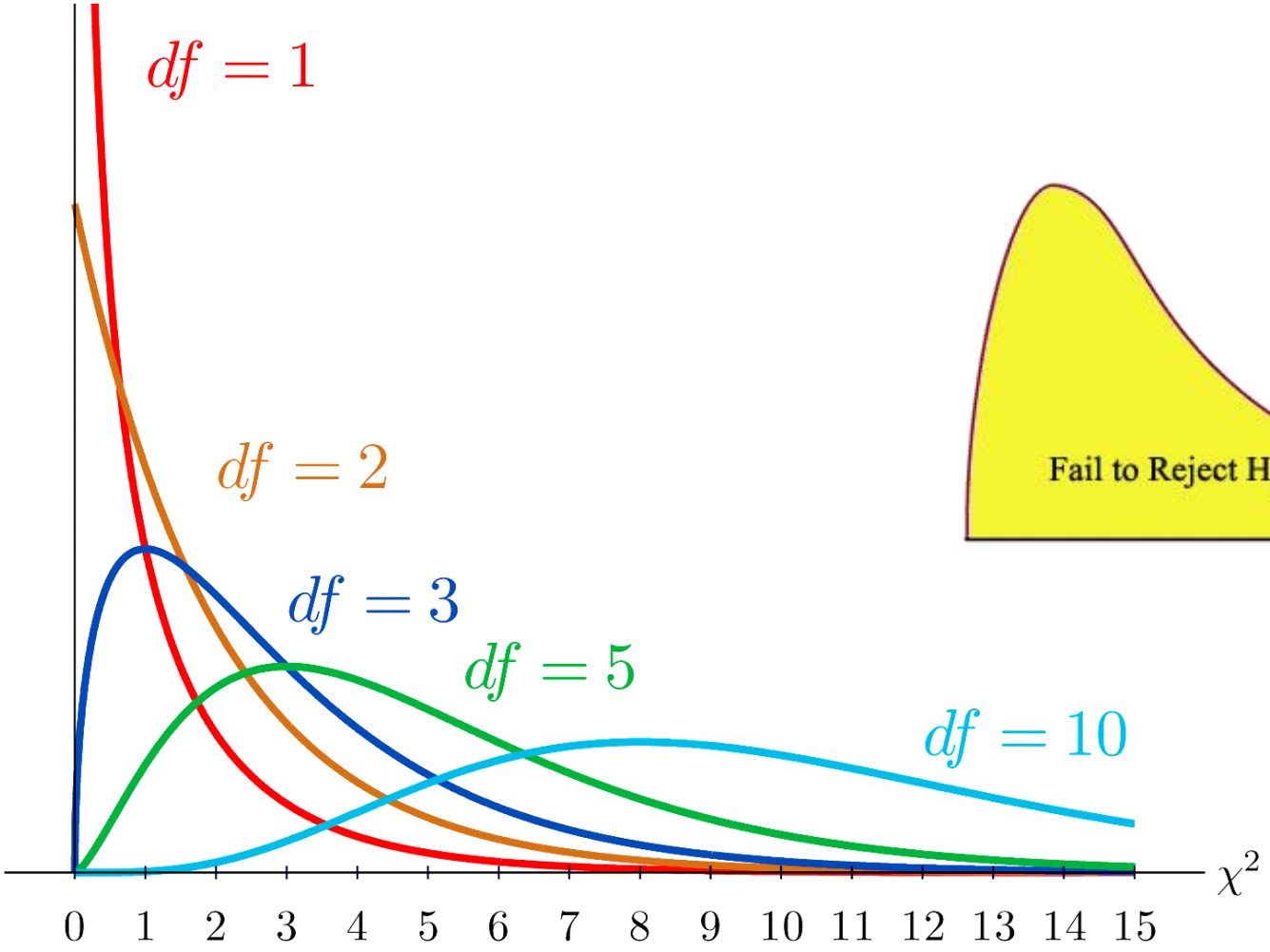
$\chi^2_{cv}$ = 3.841 4.320 ≥ 3.841, → отвергаем $H_0$ *p*=0.038

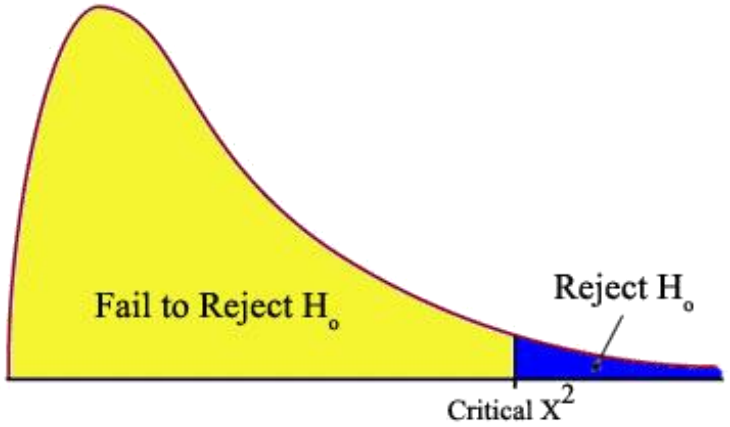
Чем больше значение $\chi^2$, тем хуже наши данные соответствуют теоретическому распределению, тем меньше *p*.

$H_0$ отвергнута, т.е. соотношение мышей не соответствует ожидаемому

*Примечание: недопустимо переводить частоты в проценты ни в одном из частотных критериев!!!*

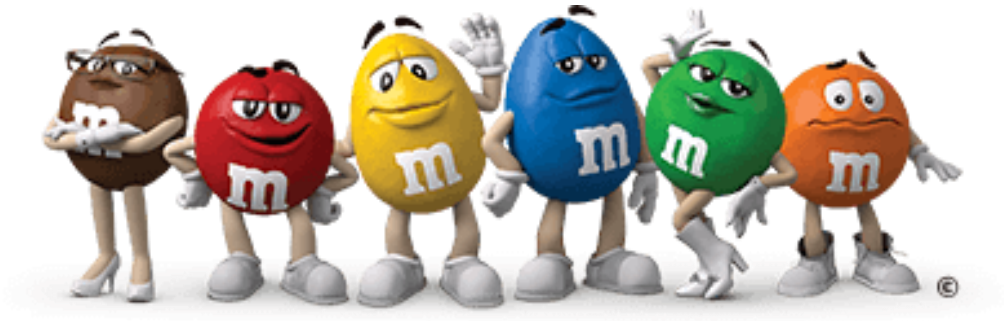
# Распределение статистики $\chi^2$



# Критерии согласия

Категорий может быть больше 2-х ($k \geq 2$). Например, можно открыть пачку M&M's и проверить, соответствует ли распределение конфет разного цвета **равномерному**.

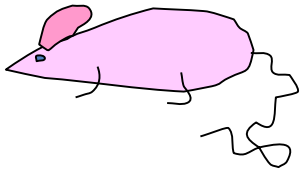


| | коричневые | красные | жёлтые | синие | зелёные | оранжевые | всего |
|---|---|---|---|---|---|---|---|
| Observed | **45** | **39** | **59** | **52** | **49** | **56** | 300 |
| Expected | *50* | *50* | *50* | *50* | *50* | *50* | 300 |

## Важное замечание:

В всех критериях согласия $H_0$ гипотеза – о том, что форма наблюдаемого распределения такая же, как теоретического.

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



Zar, 2010:

Если мы сравнили распределение с теоретическим, получили отличия (!), а теперь хотим 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 только красных меньше, чем нужно, мы можем сравнить:

1. соотношение остальных конфет с 1:1:1:1:1;
2. отношение красных к остальным с 50:250.

Однако, такой анализ допустим скорее для планирования будущих исследований, чем как рутинная процедура, т.к. идёт повторный анализ одной выборки.

*Итак, Критерий $\chi^2$ Пирсона (Pearson Chi-square test):*

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

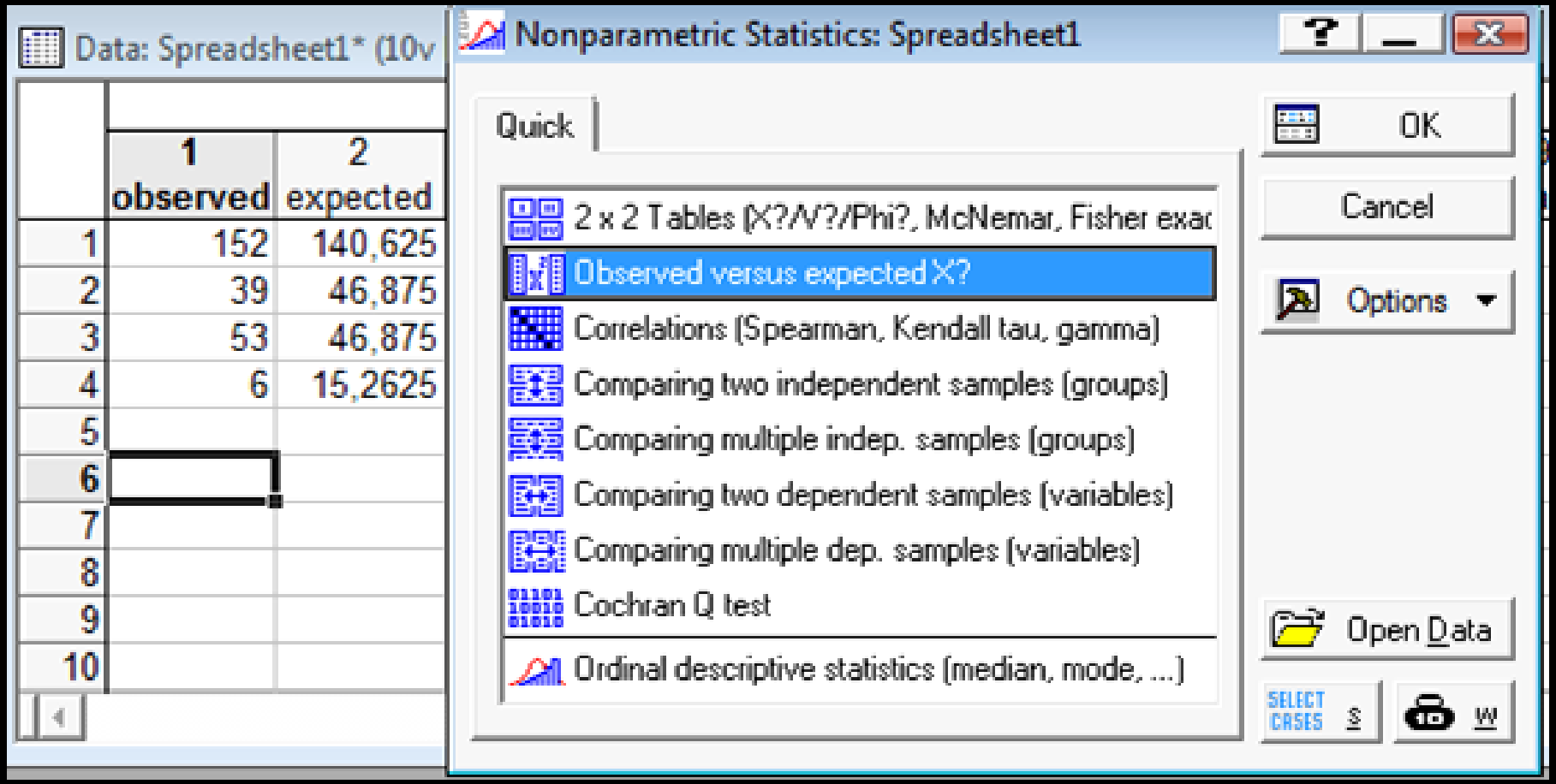
*Требования к выборке:*

1. Наблюдения независимые
2. Ограничения на минимальный размер выборки для критерия $\chi^2$ Пирсона:

О**жидаемые частоты** должны быть **≥5**, иначе непредсказуемо растёт ошибка 1-го рода.

Если частоты малы, для k=2 рекомендуется биномиальный тест.

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




*Недопустимо использовать этот критерий для сравнения 2-х выборок!*

# Критерии согласия

Data: Observed vs. Expected Frequencies (Spreadsheet1)

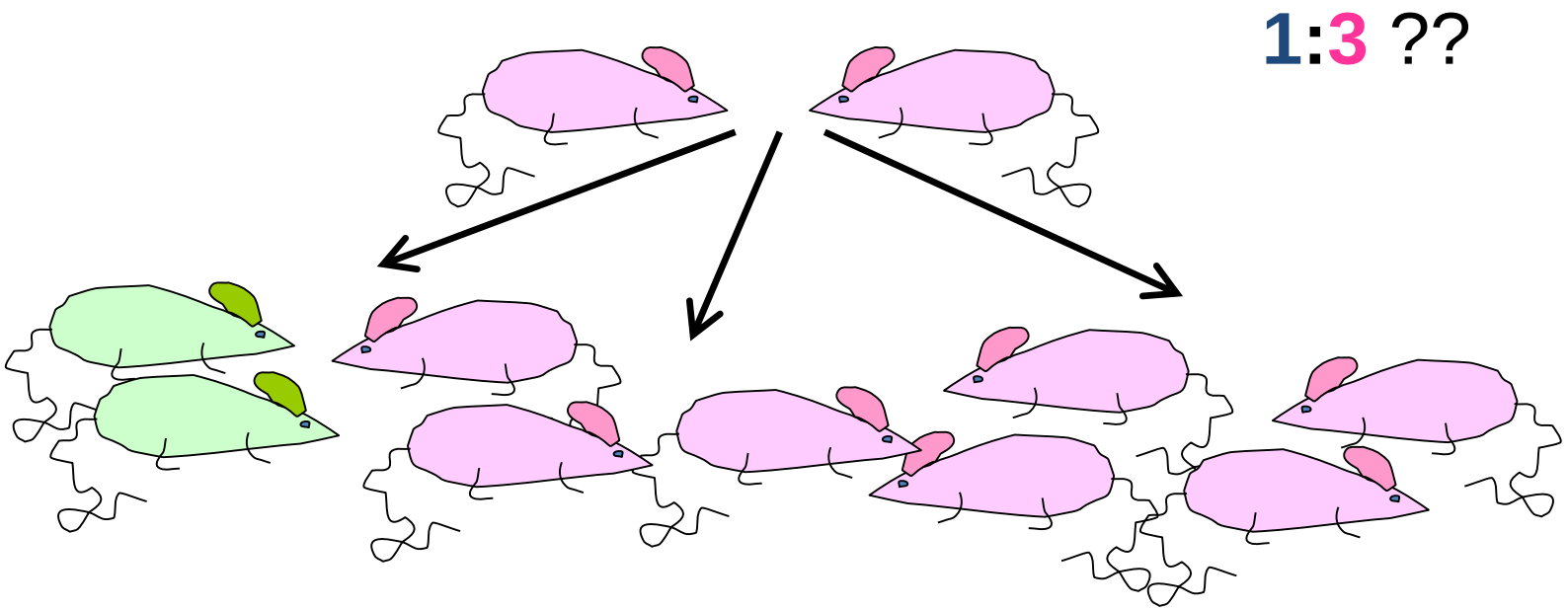
Observed vs. Expected Frequencies (Spreadsheet1)
Chi-Square = 8,664667 df = 3 p < ,034100
NOTE: Unequal sums of obs. & exp. frequencies

| Case | observed<br>observed | expected<br>expected | O - E | (O-E)**2<br>/E |
|---|---|---|---|---|
| C: 1 | 152,0000 | 140,6250 | 11,37500 | 0,920111 |
| C: 2 | 39,0000 | 46,8750 | -7,87500 | 1,323000 |
| C: 3 | 53,0000 | 46,8750 | 6,12500 | 0,800333 |
| C: 4 | 6,0000 | 15,2625 | -9,26250 | 5,621222 |
| Sum | 250,0000 | 249,6375 | 0,36250 | 8,664667 |

В публикациях: приводим $\chi^2$, df, N, $p$

Если у нас только 2 проявления признака:

## Поправка Йейтса для критерия $\chi^2$ (Yates correction for continuity)




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

Например: если ожидаемые частоты – 75 и 25, то значения $\chi^2$ 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 $\chi^2$ распределение непрерывное. И для заданного уровня значимости *p* мы не найдём точно соответствующего ему значения $\chi^2$ (для большего числа групп это тоже справедливо, но там непрерывное распределение хорошо апроксимирует дискретное).

**$\chi^2$ с поправкой Йейтса:**

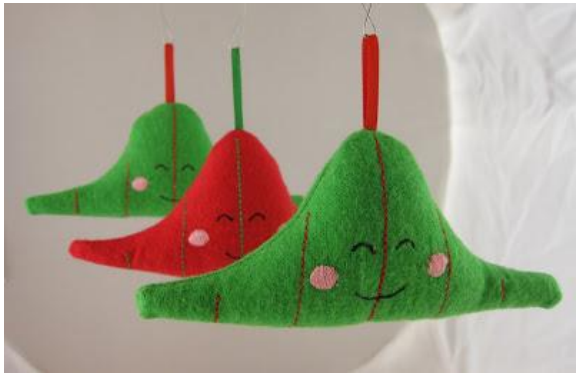
$$\chi^2 = \sum_{i=1}^{k} \frac{\left(|O_i - E_i| - 0.5\right)^2}{E_i}$$

Делает тест более консервативным.

# Тесты на соответствие непрерывным распределениям

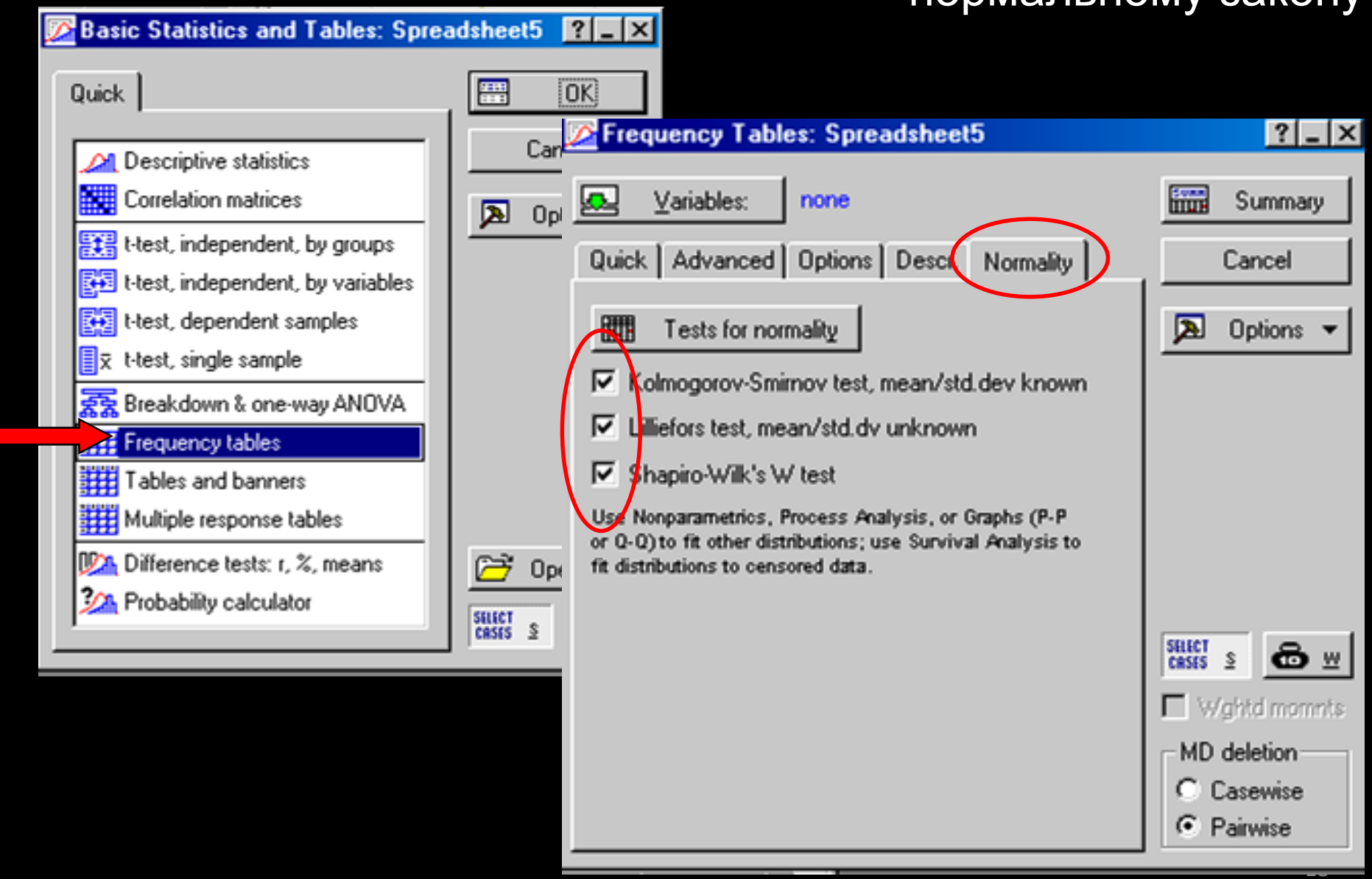
(В том числе, для сравнения с нормальным распределением).

- ✓ Тест Колмогорова-Смирнова (Kolmogorov-Smirnov test) D-статистика.
- ✓ Lilliefors test – «улучшенный К-С тест»
- ✓ **Shapiro-Wilk's W test** (самый мощный, размер выборки до 5000) – наиболее предпочтительный.

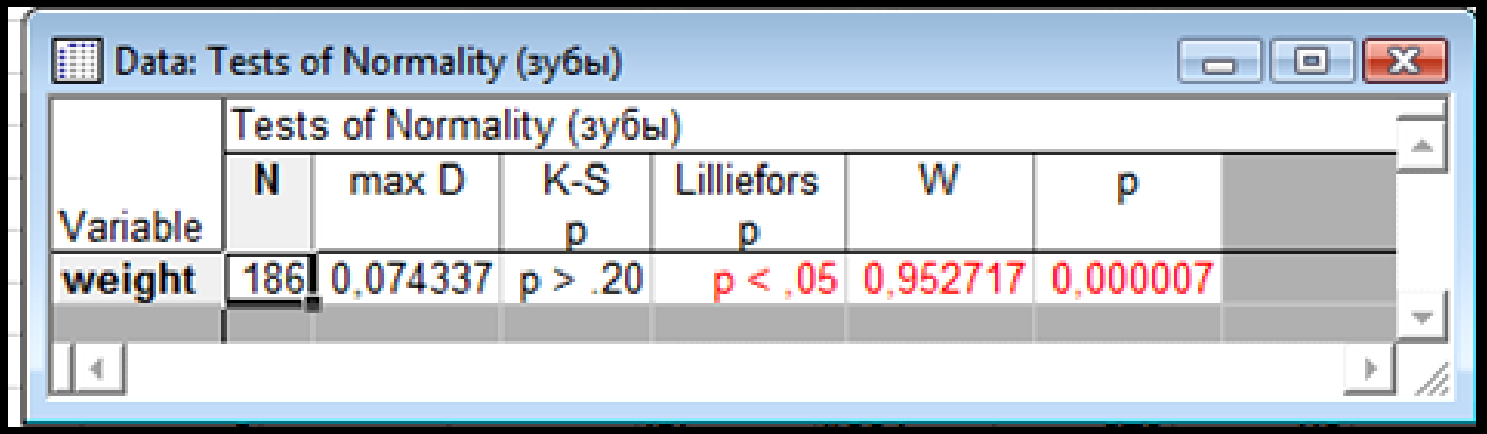


У них (как и у многих других критериев) есть **связь мощности с размером выборки**: они в маленьких выборках отклонения от теоретического распределения видят хуже, чем в больших. *Так что построение гистограмм для маленьких и очень больших выборок необходимо!*

# Проверка распределения на соответствие нормальному закону



# Критерии согласия


Data: Tests of Normality (зубы)

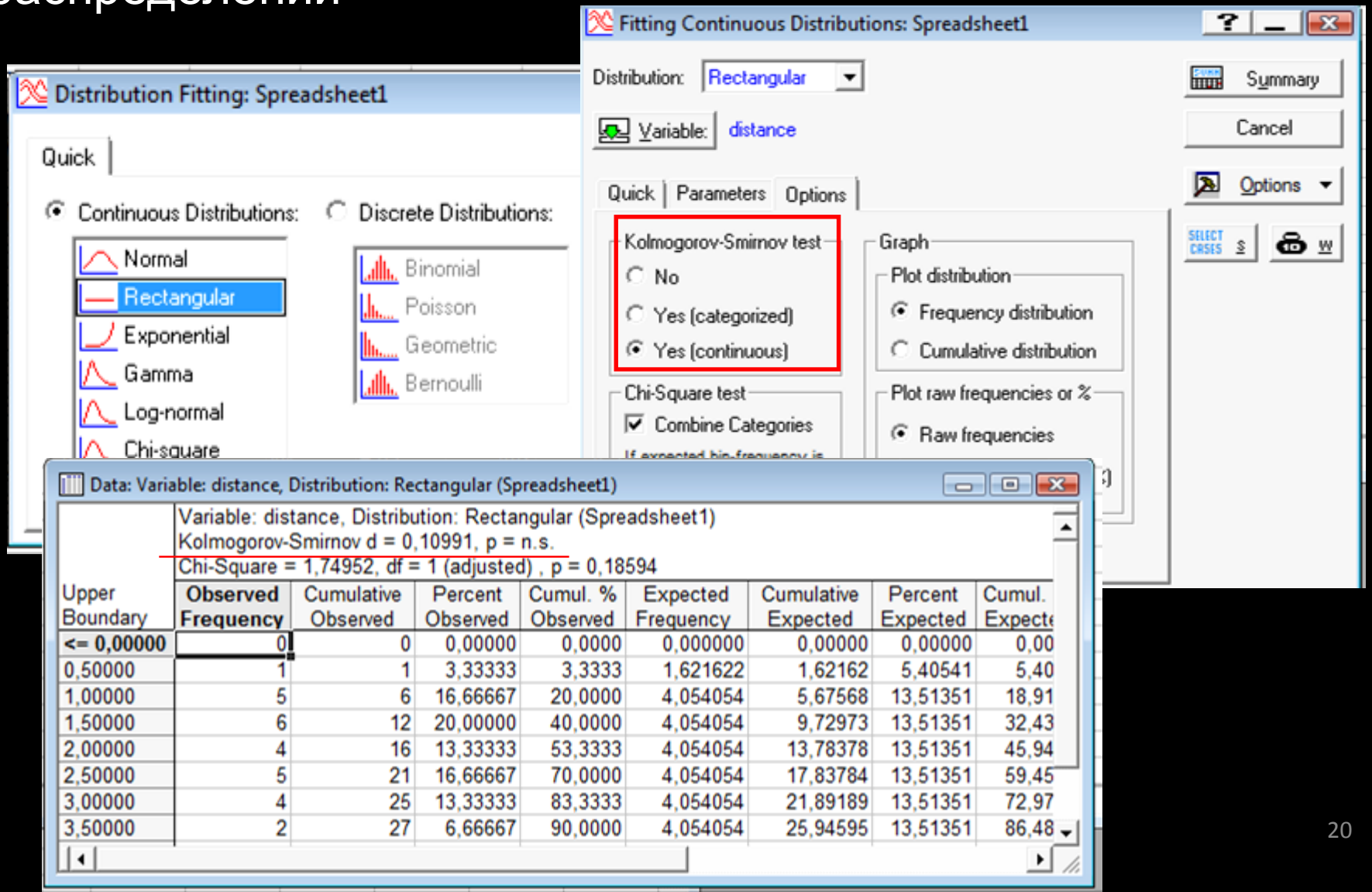
| Variable | Tests of Normality (зубы) | | | | | |
|---|---|---|---|---|---|---|
| | N | max D | K-S p | Lilliefors p | W | p |
| weight | 186 | 0,074337 | p > .20 | p < ,05 | 0,952717 | 0,000007 |

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

Видно, что К-С тест не выявил нарушений закона нормального распределения, тест Лилифорса выявил, но не рассчитал точное p, а тест Шапиро-Уилкса (W) показал серьёзные отклонения от нормального закона.

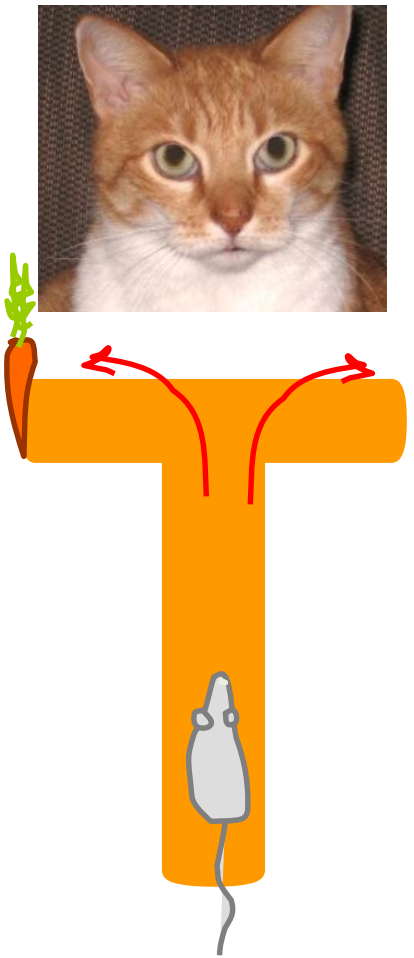
# Сравнение с другими теоретическими распределениями:

## Тест Колмогорова-Смирнова для непрерывных распределений



# Биномиальный тест

Элементарный тест для сравнения двух частот с теоретическими (для маленьких выборок). Большие выборки – задача для теста $\chi^2$.



Пример с котом Гусом: у нас есть подозрение, что он правша. Мы дали ему игрушку на резинке, он ударил по ней 10 раз: 8 - правой, 2 – левой. Справедливо ли наше подозрение?

Пример с Т-образным лабиринтом: 8 мышей пошли налево, 2 – направо.

Нулевая гипотеза о популяционной доле (ударов правой лапой):

$H_0$: $p \leq 0.5$

$H_1$: $p > 0.5$

Zar, 2010 (1999).

http://udel.edu/~mcdonald/statexactbin.htm хорошая книжка для биологов по основам анализа данных; http://www.biostathandbook.com/

# Биномиальный тест

Нам надо на основе биномиального распределения оценить, какова вероятность получить такой результат (и ещё более экстремальный) **случайно** при верной нулевой гипотезе.
Для кота Гуса это будут вероятности результатов 8:2, 9:1 и 10:1 для одностороннего теста

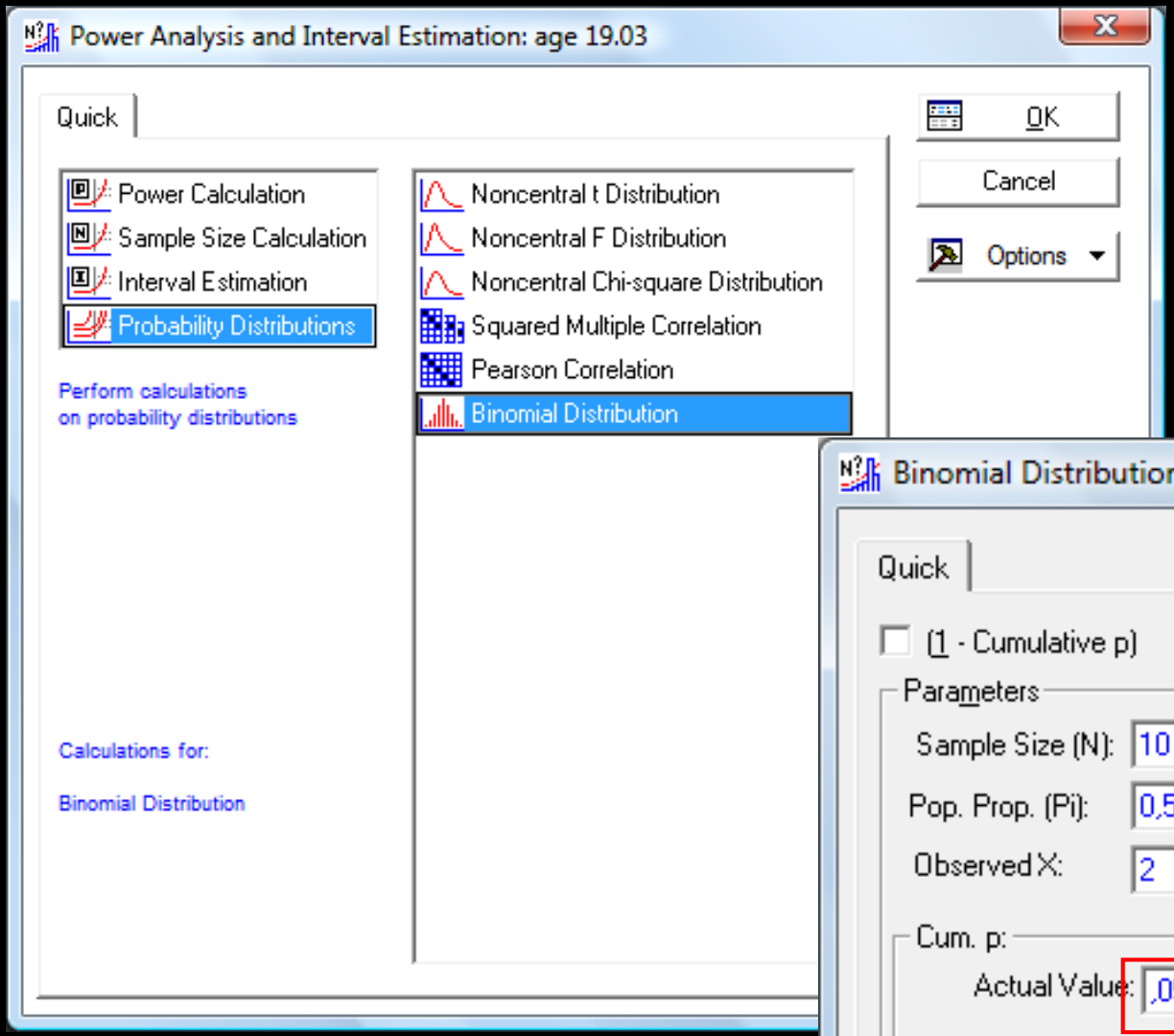
Если она меньше 0.05, отвергаем

Может быть односторонним и двусторонним.

| $X$ | $P(X)$ |
|---|---|
| 0 | 0.00024 |
| 1 | 0.00293 |
| 2 | 0.01611 |
| 3 | 0.05371 |
| 4 | 0.12085 |
| 5 | 0.19336 |
| 6 | 0.22559 |
| 7 | 0.19336 |
| 8 | 0.12085 |
| 9 | 0.05371 |
| 10 | 0.01611 |
| 11 | 0.00293 |
| 12 | 0.00024 |

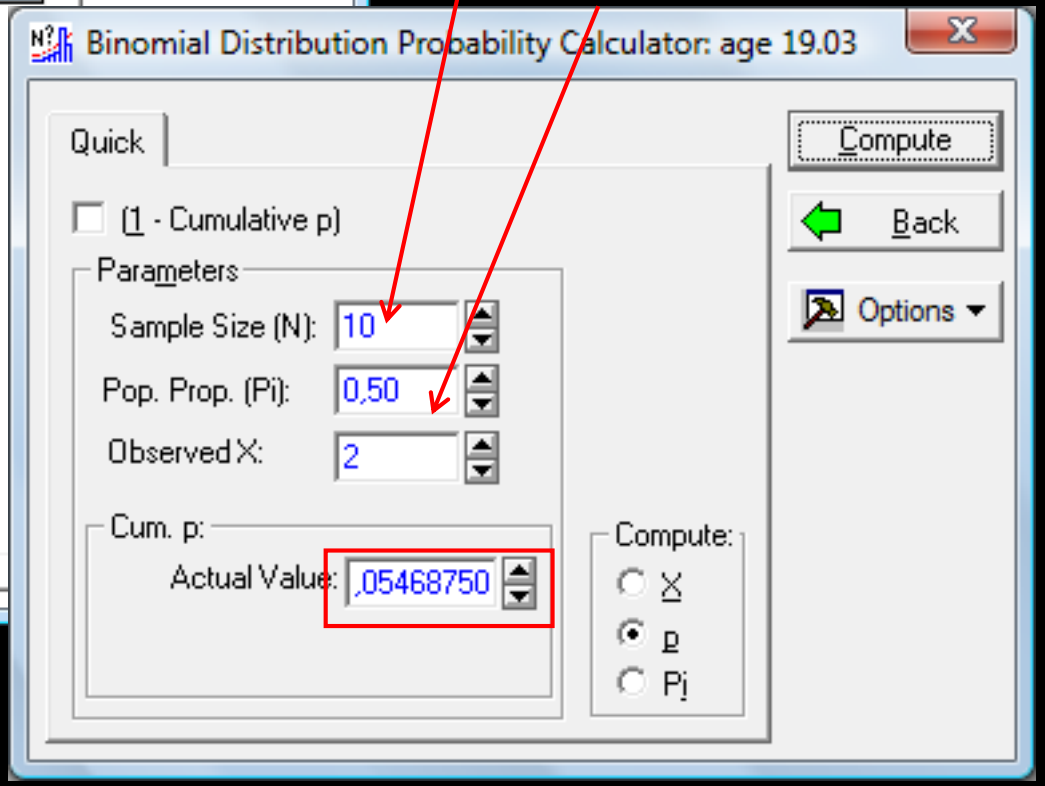
На биномиальном тесте основан Sign test – знаковый тест. По сути дела, он им и является.

# Биномиальный тест




Размер выборки

Теоретическая вероятность события




В программе легко считается только односторонний тест!

$H_0$ не отвергнута, преждевременно утверждать, что кот Гус правша

# Анализ таблиц сопряжённости (contingency tables)

Эти таблицы – классификация данных по нескольким категориальным переменным.

**Цель**: анализ взаимосвязей между ≥2 категориальными переменными.

Обычно **зависимую** переменную выделить нельзя.

**Критерий $\chi^2$** ($\chi^2$ analysis of contingency tables = $\chi^2$ test of independence)

**Пример**: мы хотим проверить, связаны ли цвет волос и пол у людей.

# Таблицы сопряжённости

| пол | брюнеты | шатены | блондины | рыжие | Всего |
|---|---|---|---|---|---|
| мужчины | 32 | 43 | 16 | 9 | 100 |
| женщины | 55 | 65 | 64 | 16 | 200 |
| всего | 87 | 108 | 80 | 25 | 300 |

✓Это – двухмерная табличка, переменных – 2.

✓Здесь, как в корреляции, зависимая переменная и предиктор неразличимы.

✓ все числа в ячейках (cells) – наблюдаемые частоты для разных комбинаций уровней переменных.

✓ суммы в строках и столбцах называются marginal totals.

# Таблицы сопряжённости

| пол | брюнеты | шатены | блондины | рыжие | Всего |
|---|---|---|---|---|---|
| мужчины | 32 | 43 | 16 | 9 | 100 |
| женщины | 55 | 65 | 64 | 16 | 200 |
| всего | 87 | 108 | 80 | 25 | 300 |

Нулевая гипотеза – о независимости частот.

**Смысл гипотезы:** выборка взята из популяции, где для объекта вероятность попасть в ту или иную категорию по одной переменной не зависит о того, в какой он категории по другой переменной (=одинакова для всех категорий по другой переменной).

Точная формулировка несколько сложна. Для ячейки ij из i-той строчки и j-того столбца:

$H_0: \pi_{ij} = \pi_{i.} \cdot \pi_{.j}$

$H_1: \pi_{ij} \neq \pi_{i.} \cdot \pi_{.j}$

На основе $H_0$ считают ожидаемые частоты!

# Таблицы сопряжённости

| пол | брюнеты | шатены | блондины | рыжие | Всего |
|---|---|---|---|---|---|
| мужчины | 32 | 43 | 16 | 9 | 100 |
| женщины | 55 | 65 | 64 | 16 | 200 |
| всего | 87 | 108 | 80 | 25 | 300 |

Мы для каждой ячейки рассчитываем ожидаемую частоту $E_{ij}$:

$$E_{ij} = \frac{(сумма\_i-й\_строки)(сумма\_j-го\_столбца)}{общая\_сумма}$$

Потом считаем статистику $\chi^2$ :

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

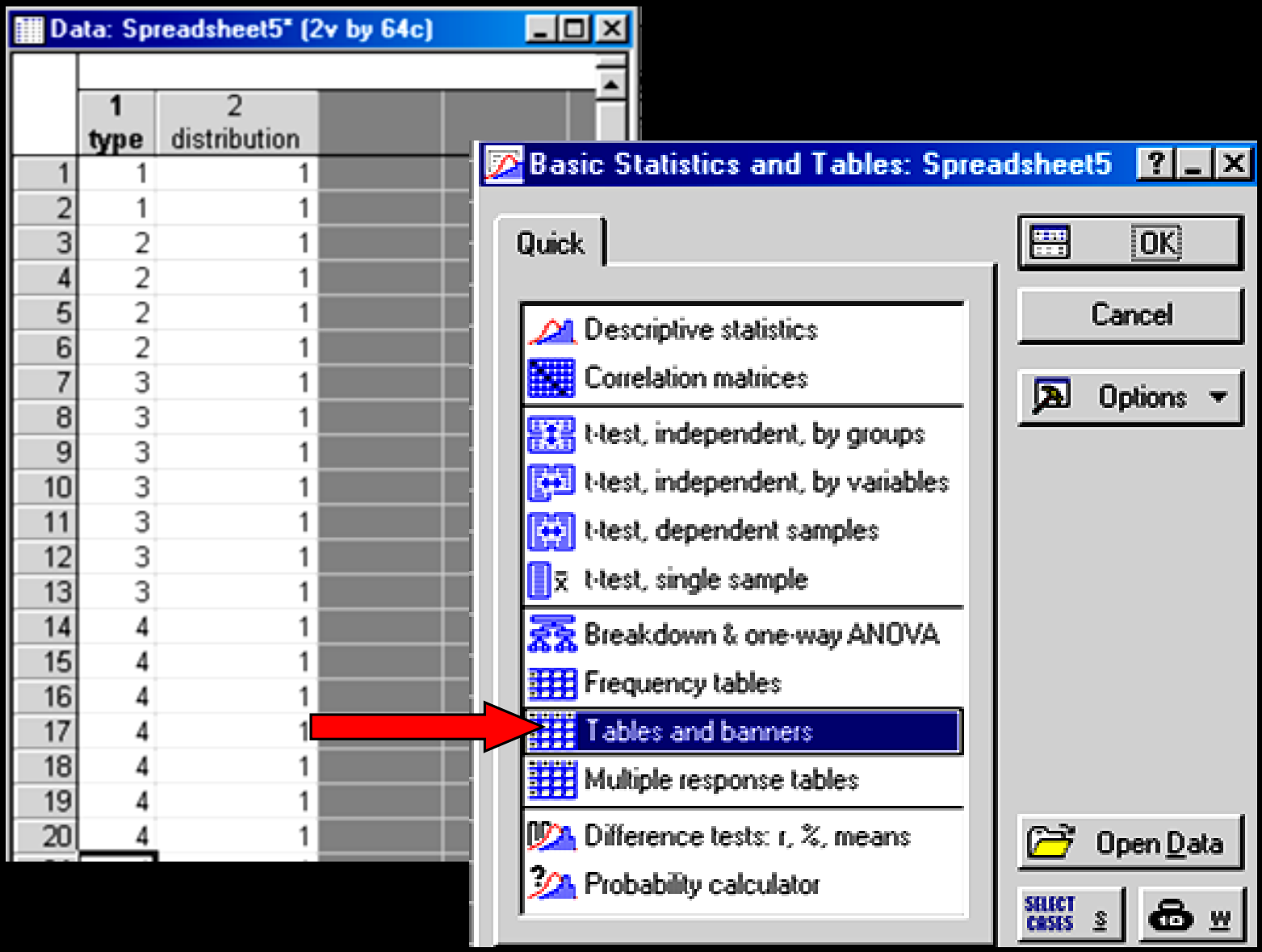
Здесь r=2, c=4. df = (r-1)(c-1)

В точности как в предыдущем $\chi^2$ тесте, только ожидаемые частоты другие

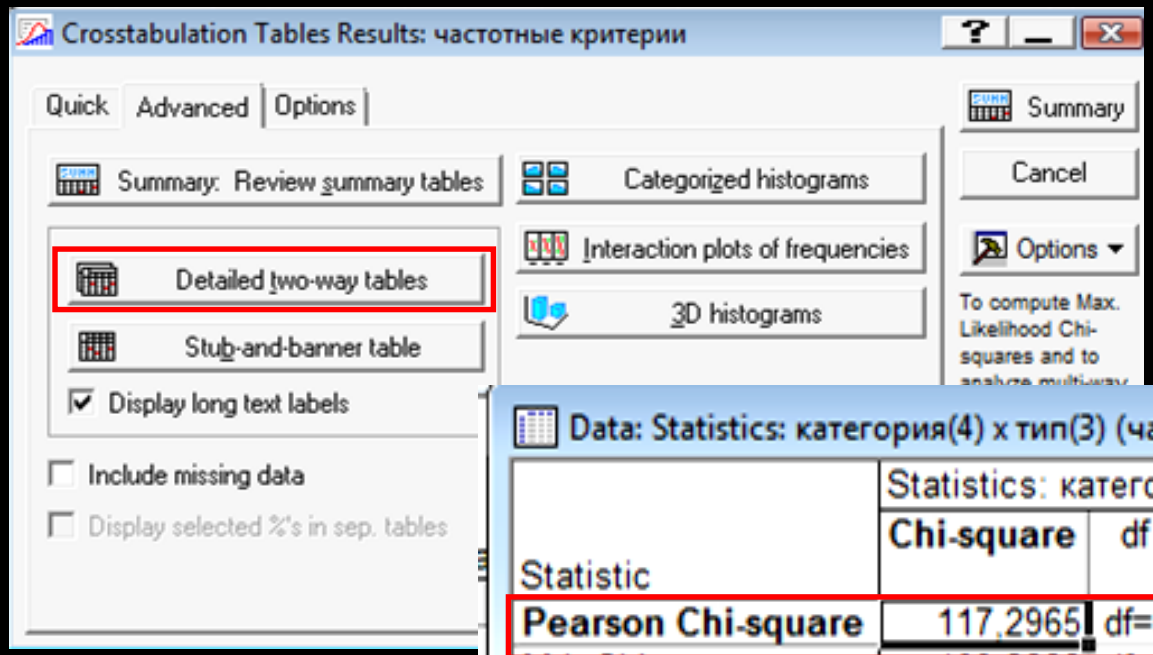
*O – E = residuals*

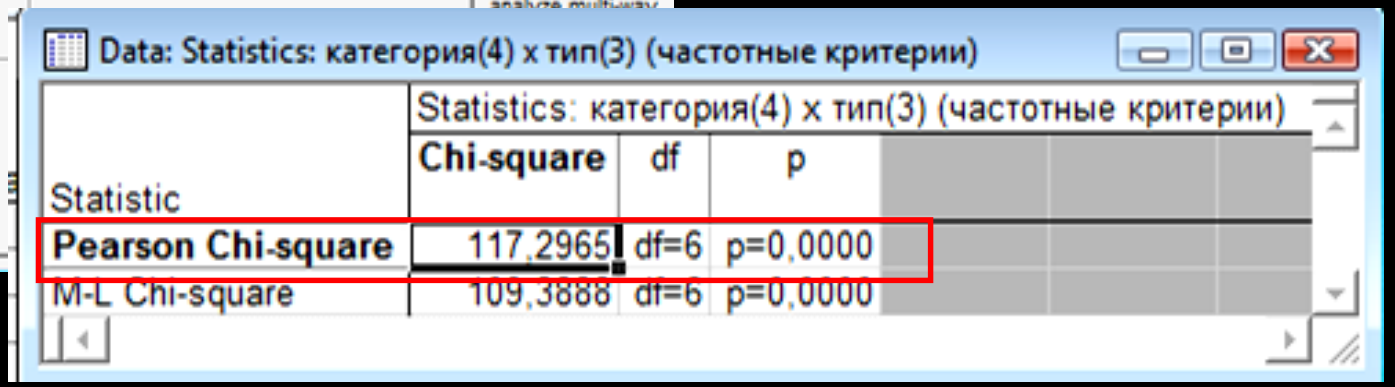
# Таблицы сопряжённости

в таблице должны быть «сырые» данные! Строка – особь (объект)



# Таблицы сопряжённости




Отвергаем нулевую гипотезу: частоты для одной переменной различаются в категориях другой.

В табличке с частотами вида a × b не должно быть значений меньше 5 (по крайней мере, ожидаемых!!). Если это не так, в крайнем случае можно объединить какие-нибудь проявления признака.

В публикациях: приводим $\chi^2$, df, N, $p$

# Таблицы сопряжённости

Zar, 1999:

Если вы отвергли $H_0$, 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.

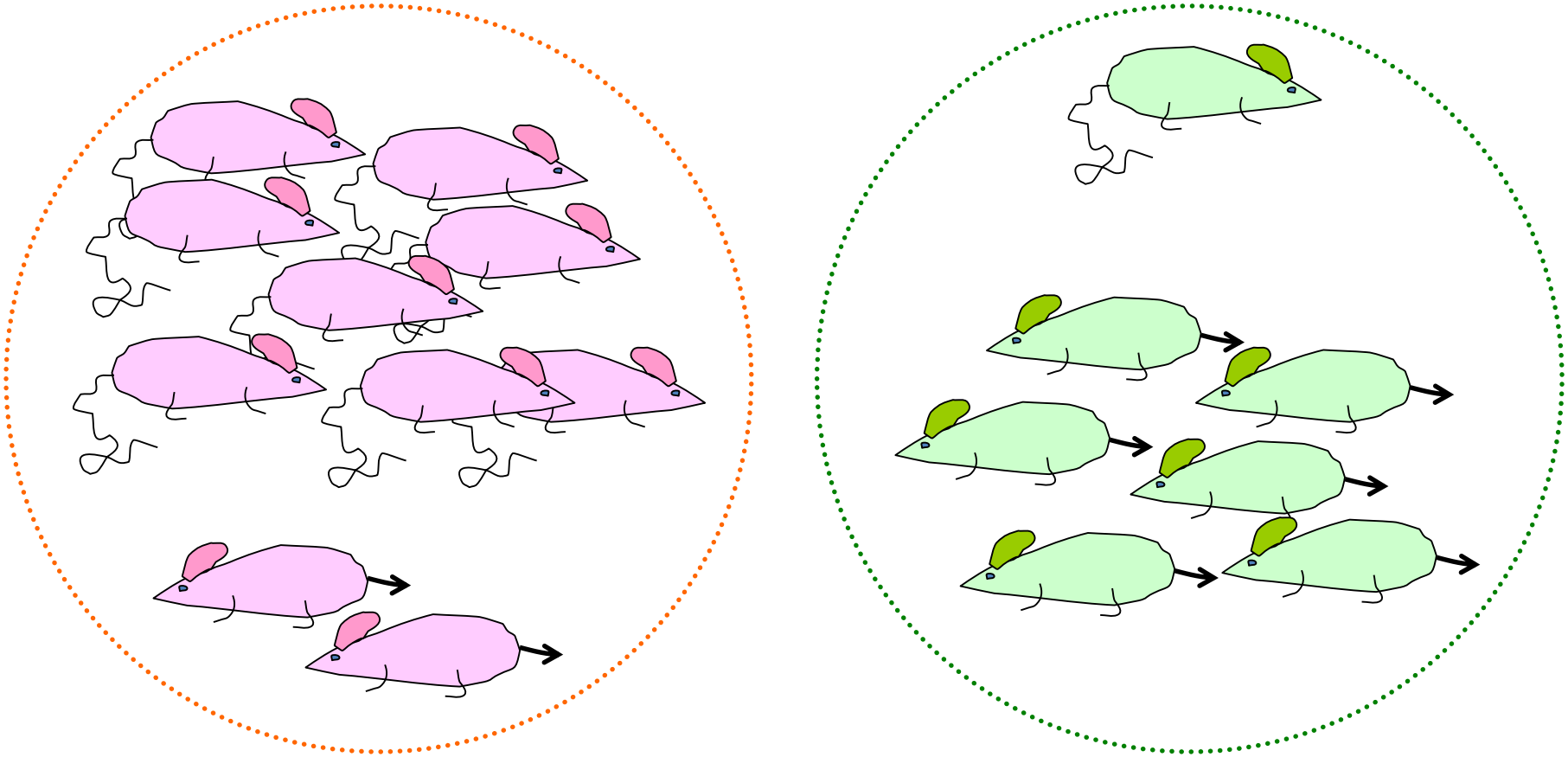
Например, если кажется, что мужчины и женщины отличаются только по соотношению рыжих, можно:

1. исключить рыжих, проверить связь пола и цвета для остальных;
2. проверить связь пола и присутствия рыжего цвета отдельно.

*Идейный аналог пост-хок теста*

# Четырёхпольные таблицы (2 x 2 tables) для независимых выборок: связь бинарных переменных.

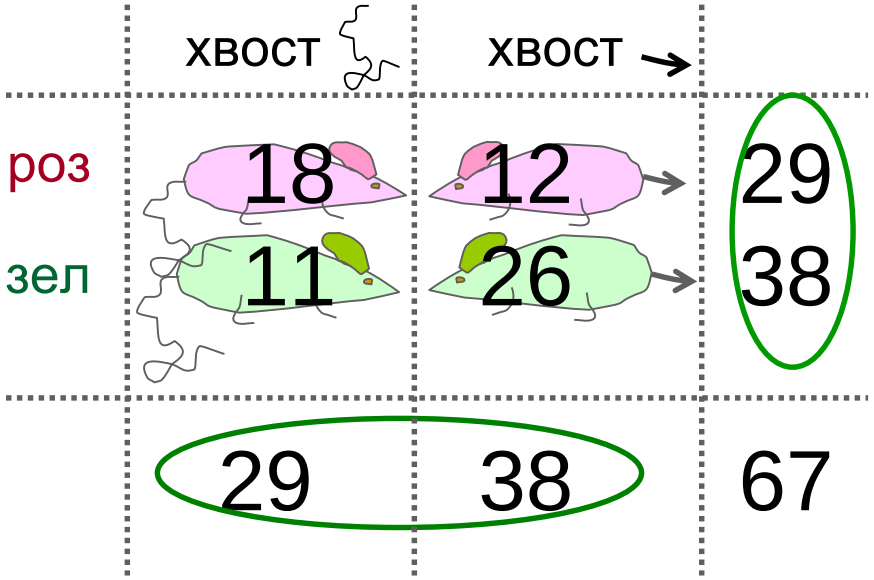
Есть только 2 фактора, у каждого – только по 2 проявления. Очень распространённый дизайн, т.к. в биологии много **БИНАРНЫХ** переменных: самец-самка; выжил-нет; заболел-нет; размножился-нет…



*Связан ли цвет мышей с формой их хвостов??*

# Таблицы сопряжённости

## Четырёхпольные таблицы (2 x 2 table)




Специально для анализа взаимосвязей бинарных переменных разработаны особые тесты.

# Разные модели в таблицах сопряжённости

Модель 1: все особи набираются случайным образом, и количество наблюдений в рядах и строках заранее неизвестно (напр., поставили линию ловушек, поймали зверьков разных видов, хотим сравнить соотношение полов в них). $H_0$ формулируется о независимости двух переменных.

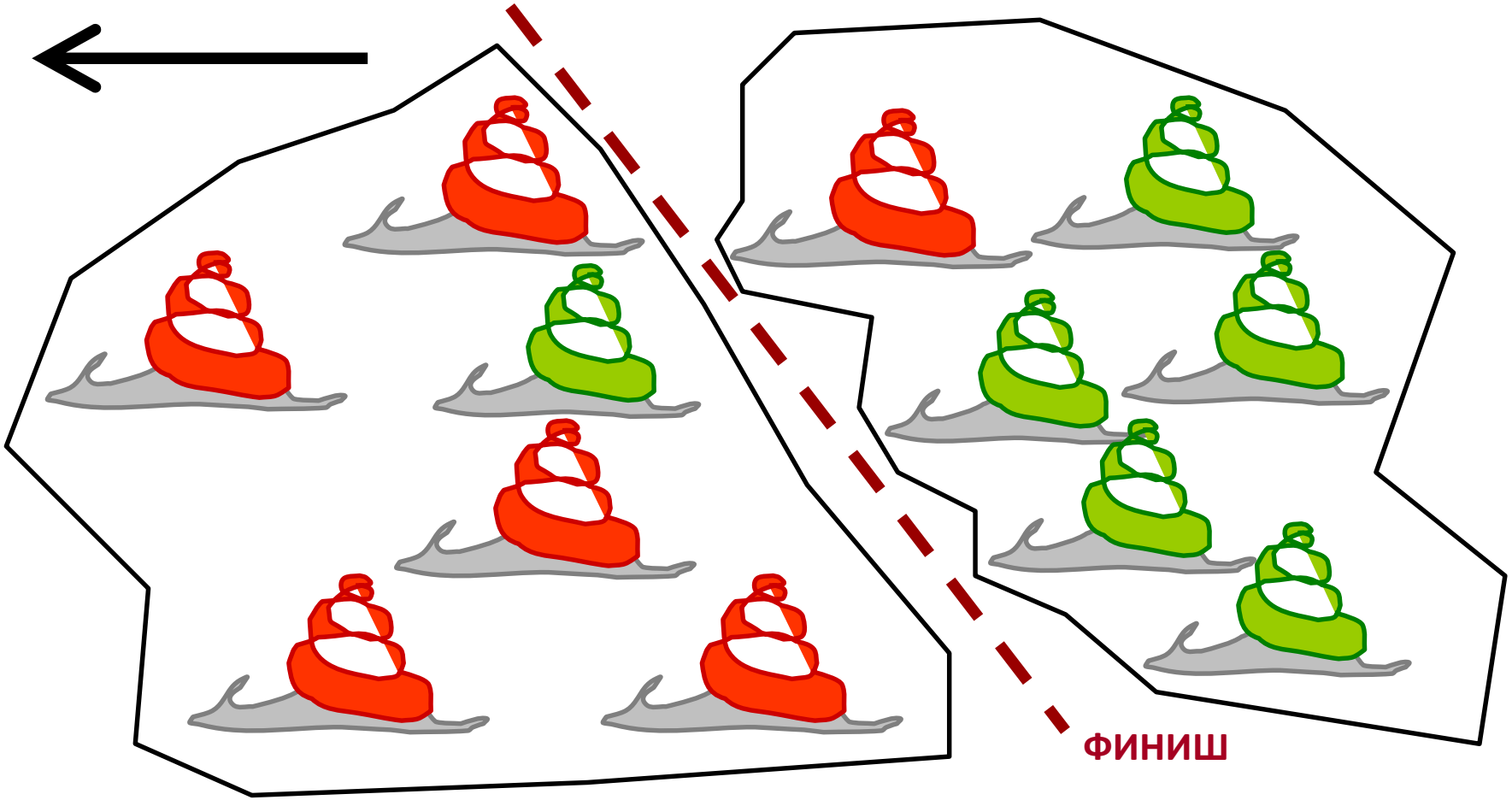
Модель 2: общее число наблюдений задаётся исследователем либо для строк, либо для столбцов (наловили по 100 особей каждого вида и теперь сравниваем соотношение полов) $H_0$ может быть сформулирована о равенстве долей в популяциях.

Модель 3: исследователем задаётся и число наблюдений в строках, и в столбцах (сверхэкзотический вариант).

Почти наверняка мы имеем дело с моделями 1 и 2, для них можно использовать тест $X^2$ и тест Фишера; для 3-й модели – тест Фишера.

# Таблицы сопряжённости

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



## *Критерий $\chi^2$ (Chi-square) с поправкой Йейтса.*

Смысл введения поправки – тот же, что для сравнения наблюдаемых и ожидаемых частот: трудности в апроксимации реальных значений $\chi^2$ непрерывным распределением. Делает тест более консервативным.
**Не обязательна для больших выборок**. В Statistica: поправку вводят, если хотя бы одна частота меньше 10.

**В программе**: если в табличке сырые данные, а не готовая четырёхпольная таблица – Tables and Banners.
Если готовая таблица – 2 x 2 tables.

Нас самом деле, сегодня вместо критерия $\chi^2$ рекомендуется использовать точный критерий Фишера.

# ***Точный критерий Фишера*** (Fisher exact test)

Годится, если одна из частот меньше 5 и вообще, для небольших выборок. Подходит даже для 3-й модели. Вообще, ЛУЧШИЙ из 2x2 тестов!

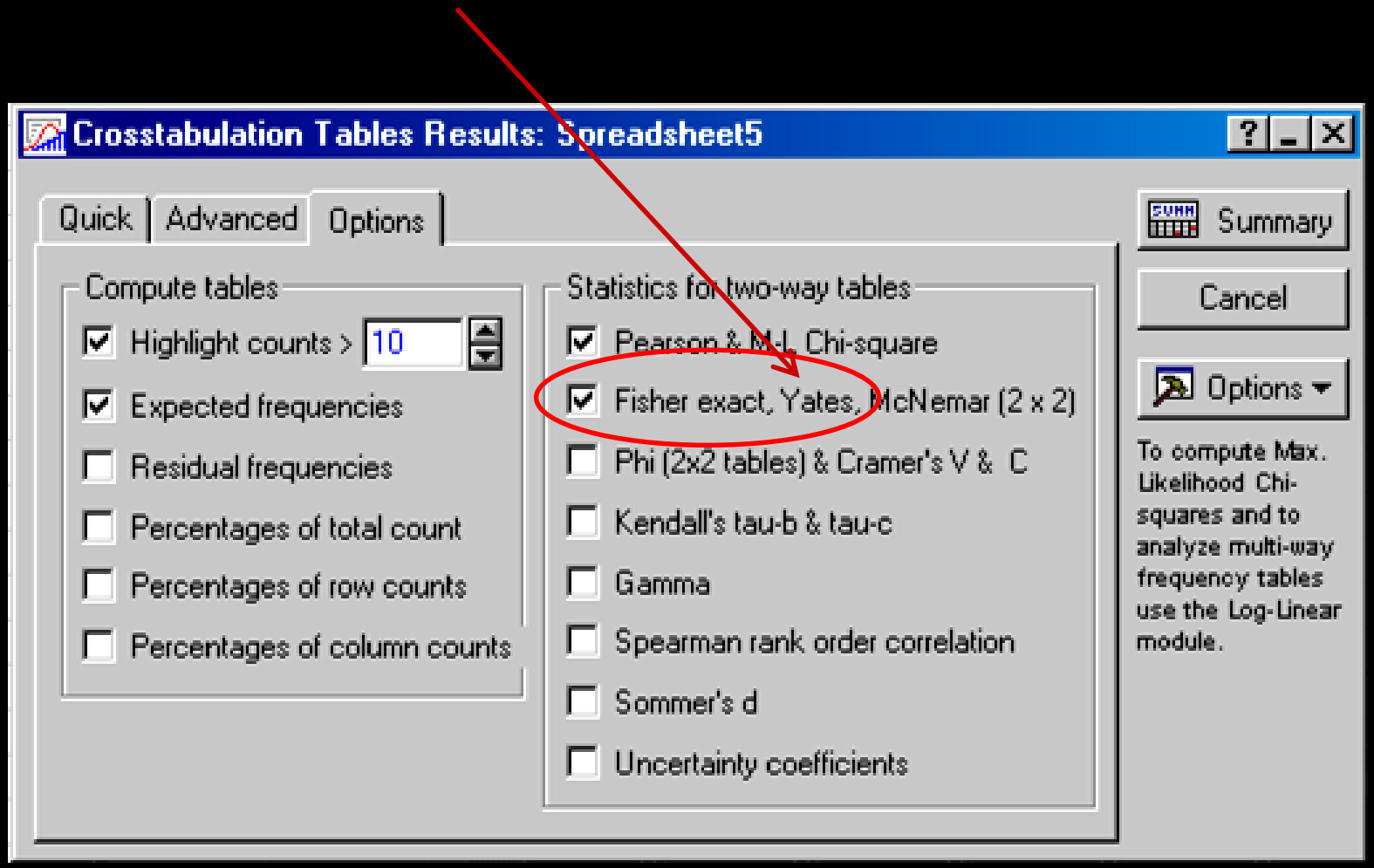
Хотим проверить, есть ли связь между заболеваемостью токсоплазмозом и регионом у манулов

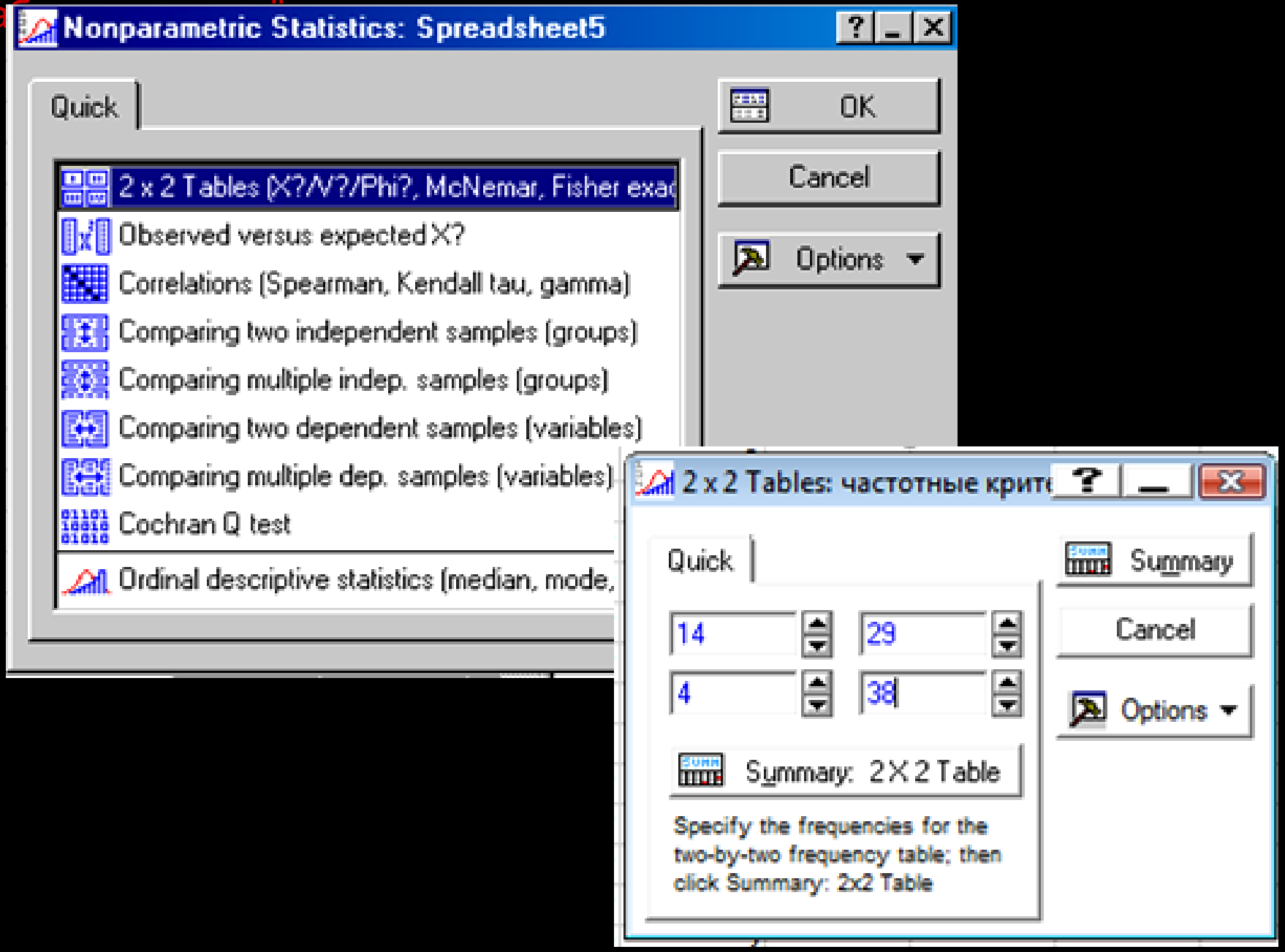
| манулы | Антитела + | Антитела - |
|---|---|---|
| Район A | 14 | 29 |
| Район B | 5 | 38 |

$H_0$: район, где живёт кот, и заболеваемость не связаны;
$H_1$ : между районом и заболеванием есть связь.

Тест основан на частотах гипергеометрического распределения, т.е. его принцип совсем не как у $X^2$, как и биномиальный тест, он рассчитывает точную вероятность получить соотношение частот «не менее экстремальное», чем наблюдаемое.

# Таблицы сопряжённости




Nonparametric Statistics: Spreadsheet5
Quick
2 x 2 Tables (X?/V?/Phi?, McNemar, Fisher exac
Observed versus expected X?
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
2 x 2 Tables: частотные крите
Quick
14
29
4
38
Summary: 2X2 Table
Specify the frequencies for the two-by-two frequency table; then click Summary: 2x2 Table
Summary
Cancel
Options

# Таблицы сопряжённости

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 | |

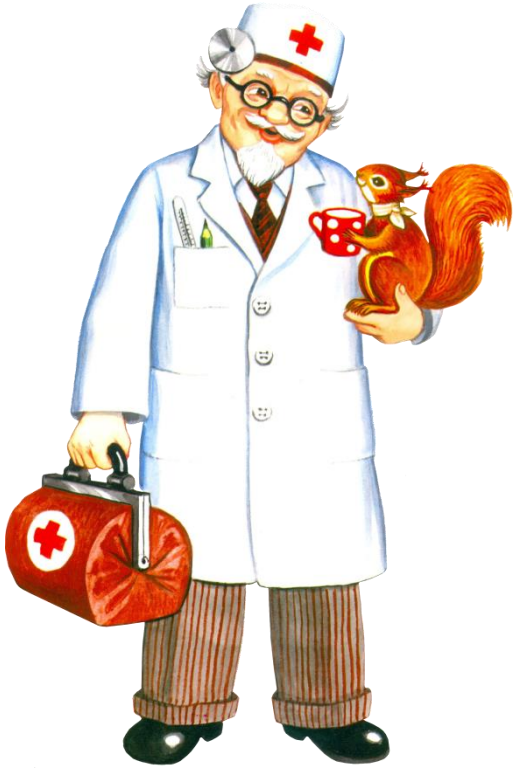
Отвергаем $H_0$

Манулы из разных районов имеют разную заболеваемость. Замечание: тест надо выбирать двусторонний!! В статье для теста Фишера приводится только *p*.

## 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.



# Таблицы сопряжённости

Для бинарных переменных тоже можно посчитать **корреляцию**! В первую очередь, если проверяется вопрос, связанно ли присутствие каких-то признаков у особей (например, антител к 2-м заболеваниям; 2-х разных мутаций и пр.; в обеих переменных присутствие – 1, отсутствие - 0).

**Phi-square** – показатель корреляции между качественными переменными. +1 – признаки встречаются только вместе, -1 – только порознь, 0 – нет связи.

*V*-square – разновидность $\chi^2$ теста.

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

# Измерения бинарной переменной в 2-х связанных выборках:

## **Критерий Мак-Немара** (McNemar Chi-square)

Мы провели в сентябре экзамен по математике. Из 100 учеников 36 сдали экзамен, остальные - провалили.
Потом мы подвергли всех учеников интенсивным занятиям по математике.
Для тех же учеников мы провели экзамен во 2-й четверти.
Повлияли ли занятия на успеваемость?

По сути дела, это просто двухвыборочный тест для связанных выборок – аналог критерия Вилкоксона, только для бинарных переменных.


Требуется специальная организация таблицы

# Бинарная переменная в связанных выборках

| Экзамен второй | Экзамен первый | | Всего |
|---|---|---|---|
| | Не сдали | Сдали | |
| Не сдали | 12 | 6 | 18 |
| Сдали | 52 | 30 | 82 |
| | 64 | 36 | |

$H_0$: доля учеников, которые сдали экзамен в первый раз, такая же, как и во второй раз.

$H_1$ : эти доли различаются.

Рассчитываем ожидаемые частоты для «зелёных» ячеек (=(52+6)/2=29) и сравниваем их с наблюдаемыми частотами тестом $\chi^2$ (df=1). Нельзя менять порядок чисел, когда мы вносим их в Статистику!

Условие применения: сумма частот в сравниваемых ячейках не должна быть меньше 10

# Бинарная переменная в связанных выборках

Критерий Мак-Немара

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 | |

Надо посмотреть в табличку 2-Way Summary Table: A,B,C,D присваиваются по часовой стрелке.

В публикациях: приводим $\chi^2$, N, $p$

Бинарная переменная в связанных выборках

# Измерения бинарной переменной в ≥3-х связанных выборках:

# Cochran's Q test

Сравнивает несколько связанных измерений бинарной переменной.

**Пример**: переменная – наличие/отсутствие укусов у человека, одевающегося в разную одежду.

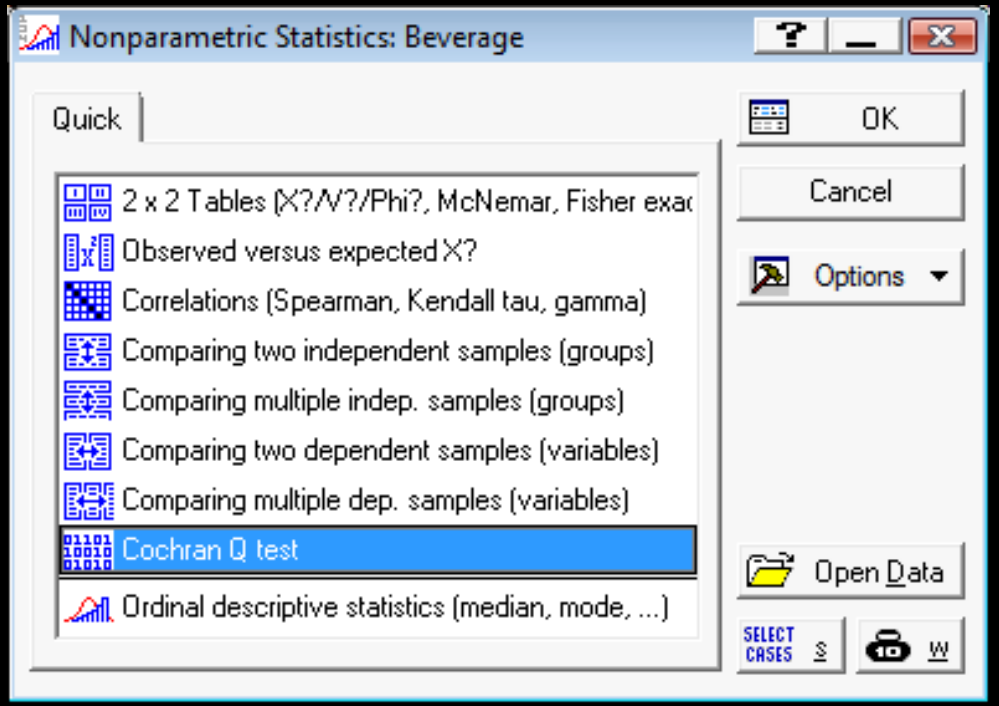
$H_0$: доля покусанных людей одинакова в разной одежде. Исключают строки из одних нулей/единиц, считается $\chi^2$ статистика. Условие: число ненулеых строк должно быть ≥ 6.

| | Clothing Type | | | | | |
|---|---|---|---|---|---|---|
| *Person (block)* | *Light, loose* | *Light, tight* | *Dark, long* | *Dark, short* | *None* | *Totals* ($B_j$) |
| 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| 2* | 1 | 1 | 1 | 1 | 1 | * |
| 3 | 0 | 0 | 0 | 1 | 1 | 2 |
| 4 | 1 | 1 | 0 | 1 | 0 | 3 |
| 5 | 0 | 1 | 1 | 1 | 1 | 4 |
| 6 | 0 | 1 | 0 | 0 | 1 | 2 |
| 7 | 0 | 0 | 1 | 1 | 1 | 3 |
| 8 | 0 | 0 | 1 | 1 | 0 | 2 |
| *Totals** ($G_i$) | 1 | 3 | 3 | 6 | 4 | $\sum_{i=1}^{a} G_i = \sum_{j=1}^{b} B_j = 17$ |

# Бинарная переменная в связанных выборках

everage (16v by 34c)

| | Beverage preferences (see Hoffman & Frank | | | |
|---|---|---|---|---|
| | 1 COKE_Y | 2 COKE_N | 3 DCOKE_Y | 4 DCOKE_N |
| 1 | 1 | 0 | 0 | 1 |
| 2 | 1 | 0 | 0 | 1 |
| 3 | 1 | 0 | 0 | 1 |
| 4 | 0 | 1 | 1 | 0 |
| 5 | 1 | 0 | 0 | 1 |
| 6 | 1 | 0 | 0 | 1 |
| 7 | 0 | 1 | 1 | 0 |
| 8 | 1 | 0 | 1 | 0 |
| 9 | 1 | 0 | 1 | 0 |
| 10 | 1 | 0 | 0 | 1 |
| 11 | 1 | 0 | 0 | 1 |
| 12 | 0 | 1 | 1 | 0 |
| 13 | 0 | 1 | 0 | 1 |
| 14 | 1 | 0 | 0 | 1 |
| 15 | 0 | 1 | 1 | 0 |
| 16 | 0 | 1 | 0 | 1 |
| 17 | 0 | 1 | 1 | 0 |
| 18 | 1 | 0 | 1 | 0 |
| 19 | 1 | 0 | 0 | 1 |




(Здесь недостоверный результат)

ran Q Test (Beverage)

Cochran Q Test (Beverage)
Number of valid cases:34
Q = 1,588235, df = 3, p < ,662061

| Variable | Sum | Percent 0's | Percent 1's |
|---|---|---|---|
| COKE_Y | 20,00000 | 41,17647 | 58,82353 |
| COKE_N | 14,00000 | 58,82353 | 41,17647 |
| DCOKE_Y | 17,00000 | 50,00000 | 50,00000 |
| DCOKE_N | 17,00000 | 50,00000 | 50,00000 |

**В публикациях**: приводим
*Q*, N, *p*

Таблицы сопряжённости могут быть не только двухмерными: анализ взаимосвязей между несколькими качественными переменными.

Сложности:

✓Хи-квадрат тест из Tables& banners показывает только что между всеми переменными где-то **вообще есть взаимосвязь**. Непонятно, где и **между какими**!

✓ между несколькими переменными могут быть взаимодействия **не только парные**, и их неплохо бы проверить;

✓ иногда в данных есть **явная зависимая** переменная, и хорошо бы это учесть.

**Пример**: изучаем связь цвета краба (4 разных цвета), состояния шипов (оба целые – один целый – все сломаны) и наличие сателлита-самца (есть/нет) у самок крабов.

Переменных – 3, **таблица 4 x 3 x 2**

Есть **2 способа** анализа такой модели:

1. Generalized linear models – построение линейной модели, где наличие сателлита – зависимая переменная, остальные – предикторы.
2. Лог-линейный анализ (log-linear model).

# Лог-линейный анализ (log-linear model)

✓Это одна из модификаций анализа **линейных моделей**;

✓он **не выделяет зависимую** переменную, просто тестирует взаимосвязи;

✓ основан на **сравнении качества моделей** методом максимального правдоподобия;

✓ в качестве «зависимой переменной» в линейном уравнении – **частота в ячейке *ij***.

Уравнение для двух переменных (двухмерная табличка):

$$\log f_{ij} = \text{constant} + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

$\log f_{ij}$ – логарифмированная **частота** в ячейке *ij;*

constant – константа;

$\lambda_i^X$ и $\lambda_j^Y$ - влияние i-го столбца переменной X, и j-й строки переменной Y;

$\lambda_{ij}^{XY}$ - **взаимодействие** влияния обеих переменных в ячейке *ij*

# Лог-линейный анализ

✓**Принцип анализа**: по очереди **исключает составляющие** (слагаемые) из модели, и **сравнивает качество** полученной упрощённой (reduced) модели с исходной полной моделью.

✓Сравнение идёт на основе **residuals**: считаются «ожидаемые» частоты для упрощённой модели сравниваются с исходными частотами полной модели.

✓**Цель:** подобрать модель с **наименьшим числом** переменных, чтобы **residuals были минимальны**.

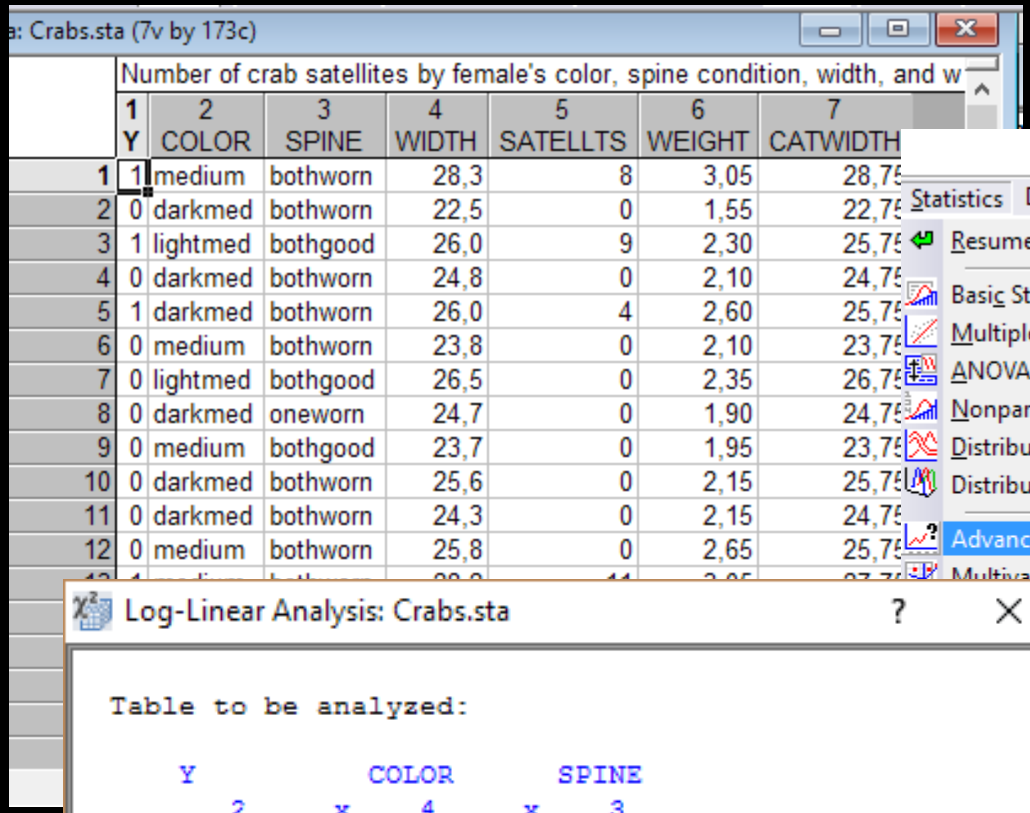
$H_0$: их несколько, по числу взаимодействий между переменными.

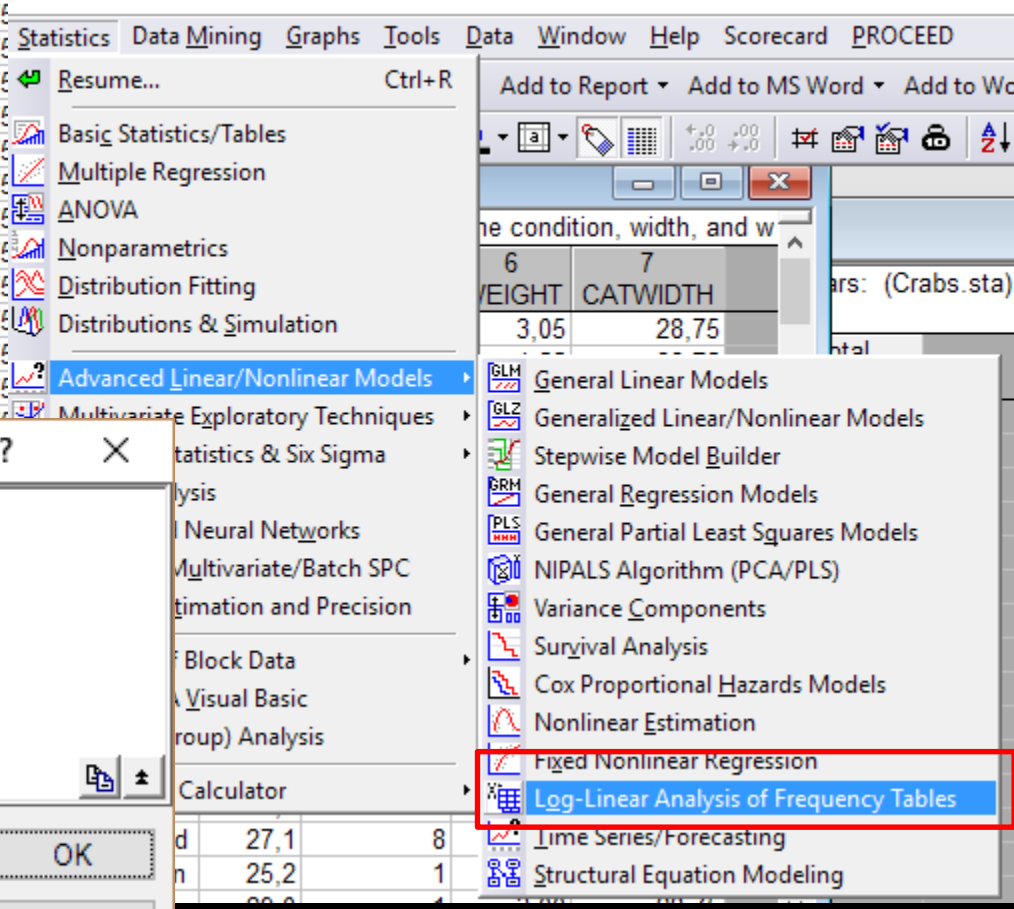
Для 2-х переменных $H_0$: $\lambda_{ij}^{XY} = 0$

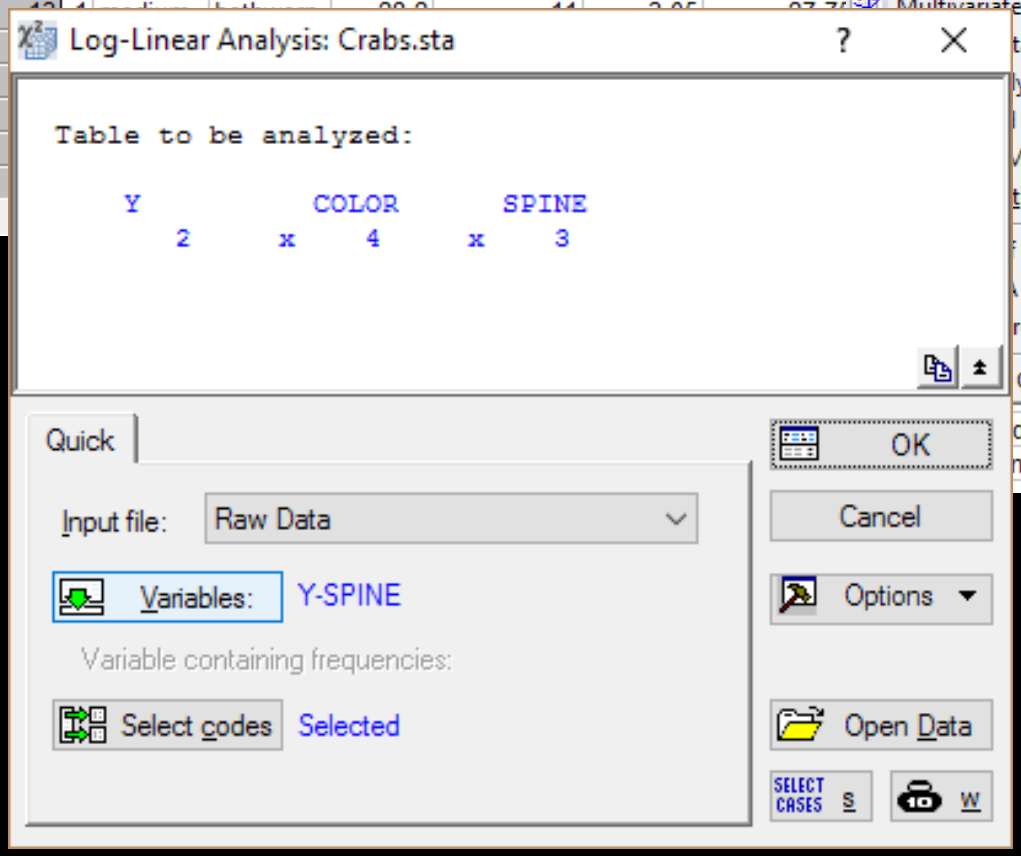
Для 3-х: $\lambda_{ij}^{XY} = 0, \lambda_{ik}^{XZ} = 0, \lambda_{jk}^{YZ} = 0, \lambda_{ijk}^{XYZ} = 0$


Технический момент: если вы включаете 3-ное взаимодействие в модель, все парные взаимодействия для этих переменных включаются по умолчанию.

# Лог-линейный анализ



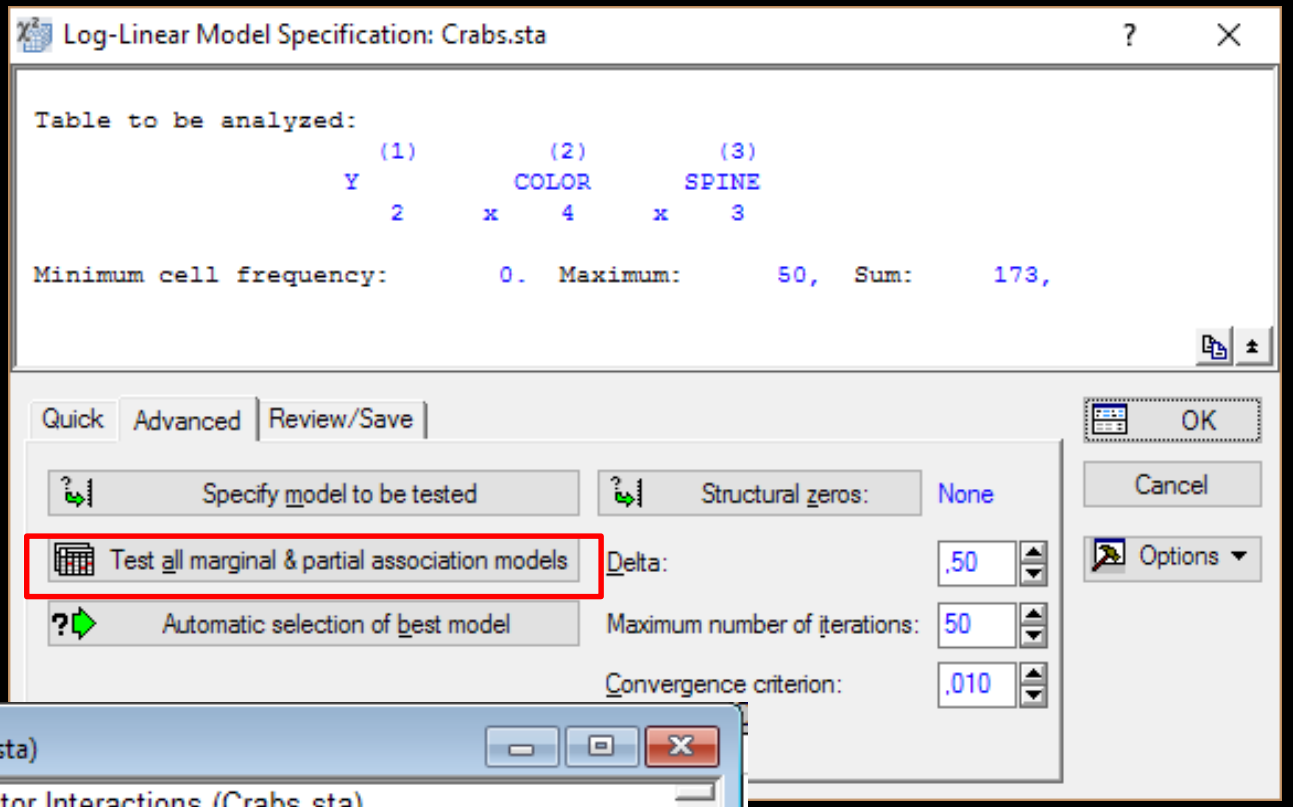



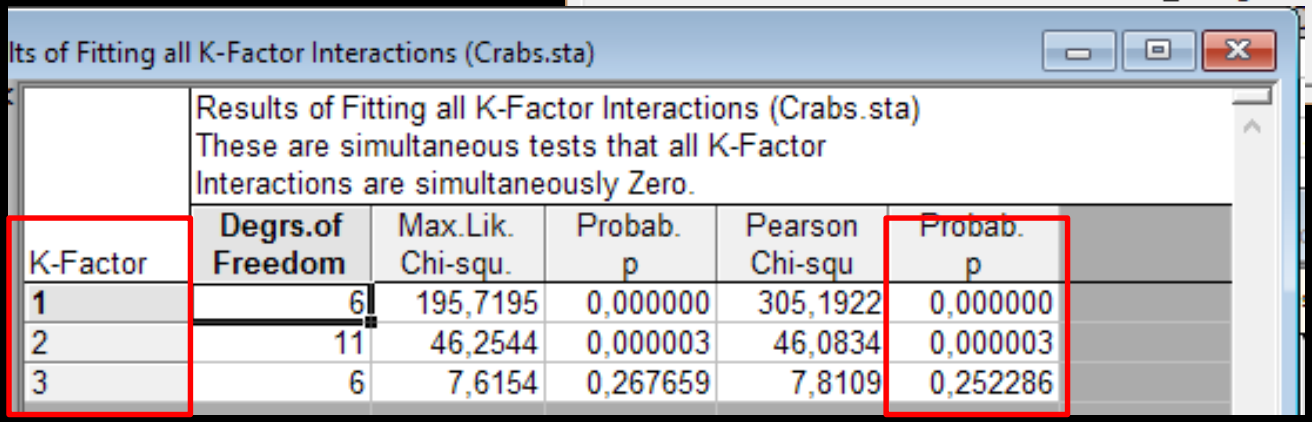




Выбираем переменные

# Лог-линейный анализ

Сперва протестируем все возможные ассоциации




| K-Factor | Degrs.of Freedom | Max.Lik. Chi-squ. | Probab. p | Pearson Chi-squ | Probab. p |
|---|---|---|---|---|---|
| 1 | 6 | 195,7195 | 0,000000 | 305,1922 | 0,000000 |
| 2 | 11 | 46,2544 | 0,000003 | 46,0834 | 0,000003 |
| 3 | 6 | 7,6154 | 0,267659 | 7,8109 | 0,252286 |

На счастье, тройное взаимодействие недостоверно, только парные (отсюда M-L хи-квадрат, df, p – в результаты).

# Лог-линейный анализ

Это нам неинтересно, это вклады отдельных переменных; их трудно обсуждать

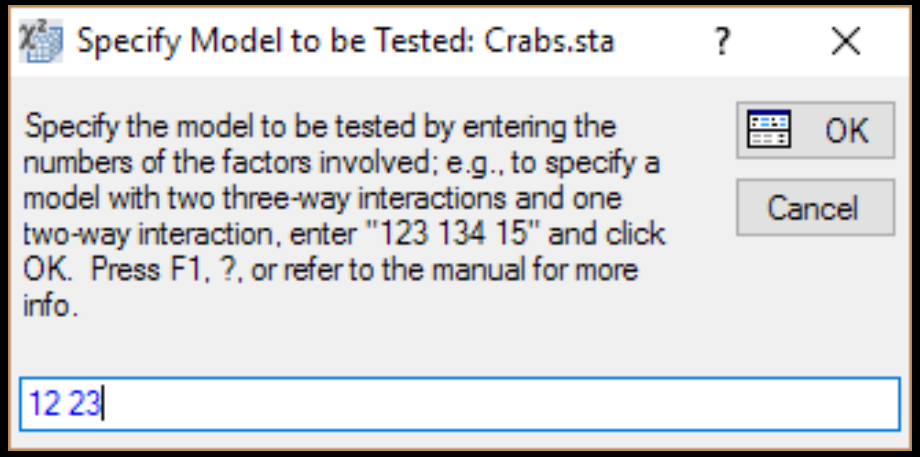
of Marginal and Partial Association (Crabs.sta)

Tests of Marginal and Partial Association (Crabs.sta)

| Effect | Degrs.of Freedom | Pt.Assn. Chi-sqr. | Pt.Assn. p | Mg.Assn. Chi-sqr. | Mg.Assn. p |
|---|---|---|---|---|---|
| 1 | 1 | 13,13457 | 0,000290 | | |
| 2 | 3 | 84,15008 | 0,000000 | | |
| 3 | 2 | 98,43494 | 0,000000 | | |
| 12 | 3 | 13,18467 | 0,004254 | 12,83273 | 0,005013 |
| 13 | 2 | 2,81959 | 0,244193 | 2,46762 | 0,291181 |
| 23 | 6 | 30,95400 | 0,000026 | 30,60206 | 0,000030 |

Зато вот здесь показаны парные взаимодействия: достоверны связи сателлит-цвет и цвет-шипы. Partial Chi-sqr. df. p приводим в результатах.

Теперь мы можем либо сами построить модель только с этими двумя взаимодействиями и проверить её качество, либо попросить программу построить для нас лучшую модель.



# Лог-линейный анализ

Log-Linear Model Specification: Crabs.sta

```
Table to be analyzed:
                    (1)         (2)         (3)
                    Y          COLOR       SPINE
                    2     x     4     x     3

Minimum cell frequency:       0.   Maximum:       50,   Sum:      173,
```

Quick | Advanced | Review/Save

OK

Specify model to be tested

Test all marginal & partial association models

Automatic selection of best model

Struc

Delta:

Maximum numb

Convergence

Results : Crabs.sta

```
Table to be analyzed:
                    (1)         (2)         (3)
                    Y          COLOR       SPINE
                    2     x     4     x     3

Minimum cell frequency:       0.   Maximum:       50,   Sum:      173,

Model to be tested: 21,32

Delta: ,5000 ;   Maximum iterations:  50 ;  Conv. criterion: ,0100
Convergence reached after # of iterations:   2
                                              df        p
Maximum Likelihood Chi-square: 10,435         8    ,23581
          Pearson Chi-square: 10,911          8    ,20677
```

Quick | Advanced | More fit statistics

Summary: Goodness of fit

Observed table

Fitted table

Marginal tables

Residuals (observed-fitted)

Residuals standardized

Plot of observed vs. fitted

Plot of fitted vs. residuals

Summary

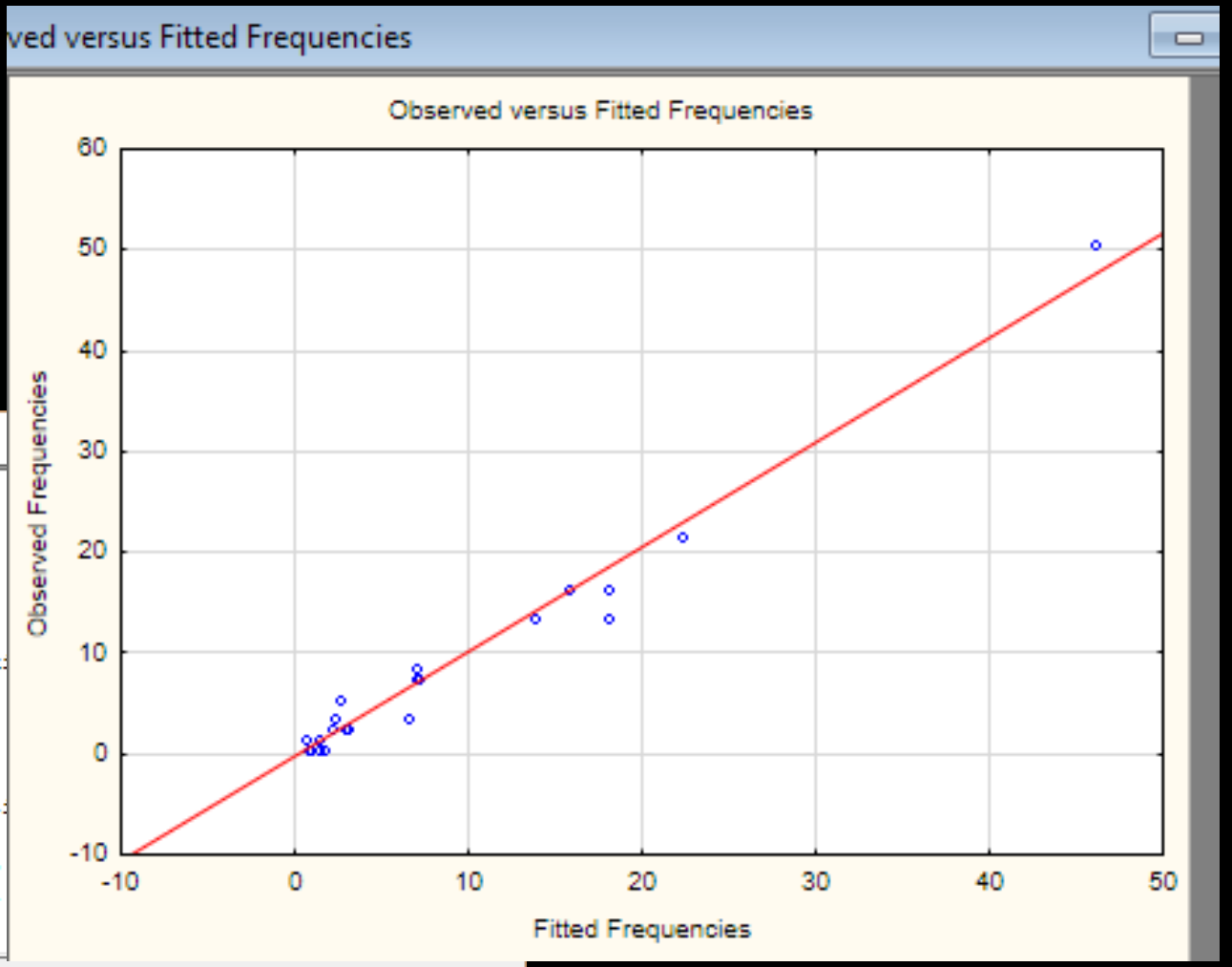
Cancel

Options

Качество этой модели не отличается достоверно от качества исходной полной модели, значит эта модель хорошая (M-L Chi sqr, df, p – в результаты).
Осталось изучить её повнимательнее.

# Лог-линейный анализ

Картинка – диагностика качества модели




Results : Crabs.sta

```
Table to be analyzed:
                    (1)          (2)
                    Y           COLOR
                    2      x     4

Minimum cell frequency:        0.   Max

Model to be tested: 21,32

Delta: ,5000 ;   Maximum iterations:
Convergence reached after # of iterat

Maximum Likelihood Chi-square: 10,435
          Pearson Chi-square: 10,911
```

Quick | Advanced | More fit statistics

- Summary: Goodness of fit
- Observed table
- Fitted table
- Marginal tables
- Residuals (observed-fitted)
- Residuals standardized
- Plot of observed vs. fitted
- Plot of fitted vs. residuals

Summary | Cancel | Options

Рассматривание таблиц с residuals позволяет понять, в каких категориях отличия сильнее всего

# К практическому занятию

- ✓Хи-квадрат - Crabs
- ✓2x2 табл – Activities
- ✓Distribution fitting – тест Колмогорова-Смирнова – Crabs
- ✓Связанные выборки - Beverage
- ✓Лог-линейные модели - Crabs