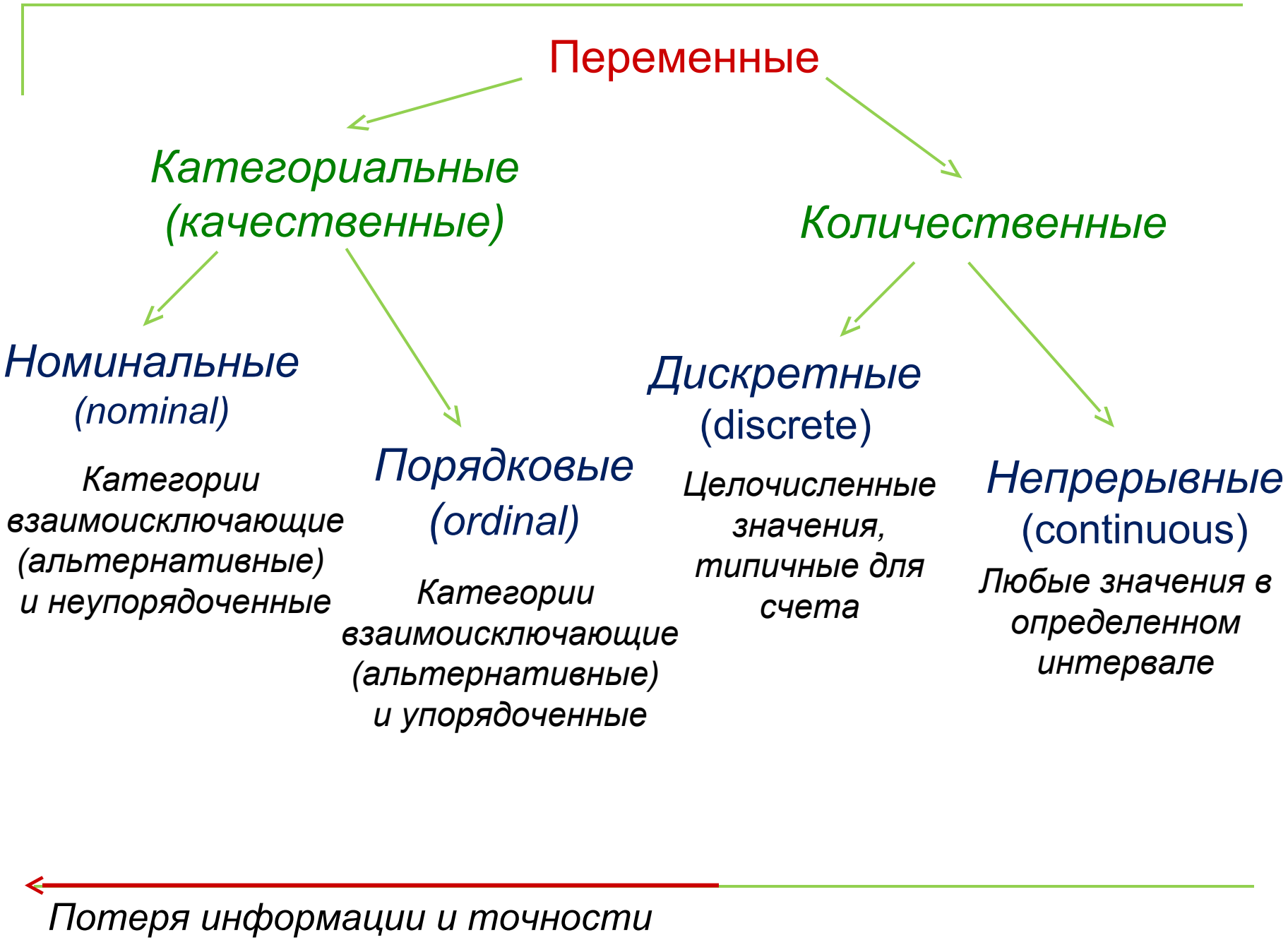
Занятие 4

Анализ качественных признаков


Переменные
Категориальные (качественные)
Количественные
Номинальные (nominal)
Категории взаимоисключающие (альтернативные) и неупорядоченные
Порядковые (ordinal)
Категории взаимоисключающие (альтернативные) и упорядоченные
Дискретные (discrete)
Целочисленные значения, типичные для счета
Непрерывные (continuous)
Любые значения в определенном интервале
Потеря информации и точности

Номинальные признаки

Номинальные признаки часто встречаются в биологии и медицине.

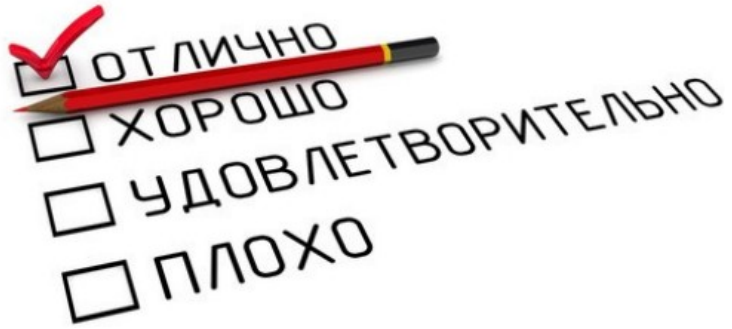
Эти признаки представлены неупорядоченными ***категориями***, например, сельское городское население, пол, группа крови, цвет и марка автомобиля, национальность и т. д.

Порядковые признаки

Порядковые признаки ***отличаются от номинальных*** тем, что ***могут быть размещены в порядке возрастания или убывания*** (например, уровень образования, степень тяжести состояния, балл успеваемости и т.п.).

Порядковые переменные, представляются ***в виде чисел***.

Однако, в отличие от количественных признаков, они ***не дают информации о степени различий*** между находящимися рядом уровнями значений порядковой переменной.



Не могут быть подвергнуты арифметическим операциям.

Описательная статистика качественных данных

Данные представляют собой частоты:
1***) абсолютные*** (в штуках) или
2***) относительные*** (в долях единицы, в процентах и др.).

Принято приводить и абсолютные, и относительные частоты, а последние снабжать 95% ДИ.

95% ДИ можно вычислить разными способами;
лучшие методы:

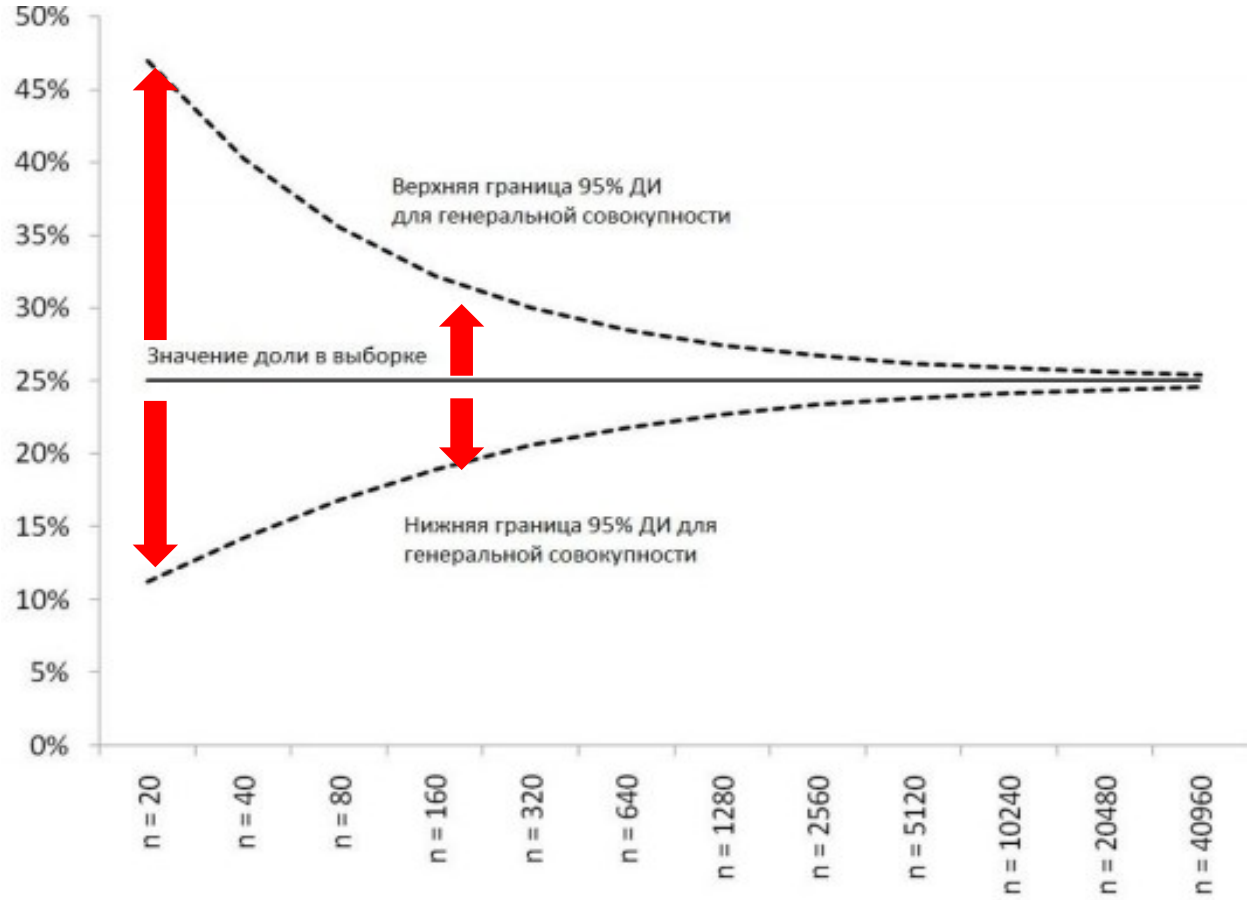
метод ***Джеффриса*** (Jeffreys' CI for proportion) – наиболее рекомендуемый на сегодняшний день,
метод ***Уилсона*** (Wilson…),
метод ***Агрести — Коулла*** (Agresti-Coull…).

Традиционен, но несколько более консервативен точный ***метод Клоппера — Пирсона*** (Clopper-Pearson…).

Особенности использования методов расчёта доверительных интервалов для долей и частот

Способ	Особенности использования
Метод Уилсона	Оптимальный метод для оценки частот: позволяет оценить доверительные интервалы для очень малых и очень больших частот, применим для выборок малого объема
Метод Вальда	Метод не рекомендуется для использования при малых объемах выборок и в случае, если частота встречаемости признака менее 25% или более 75%. Доверительные интервалы в большинстве случаев оказываются слишком узкими
Метод Вальда с коррекцией по Агрести-Коуллу	Метод не рекомендуется для использования при малых объемах выборок и в случае, если частота встречаемости признака приближается к 0% или 100%
Угловое преобразование Фишера	Метод не рекомендуется для использования, если частота встречаемости признака менее 25% или более 75%.
«Точный метод» Клоппера-Пирсона	Доверительные интервалы, полученные с использованием метода, в большинстве случаев слишком широки (степень консервативности метода увеличивается по мере уменьшения объема выборки, особенно при $n < 15$)

Ширина ДИ зависит от объема выборки: чем > объем выборки, тем < будет его ширина



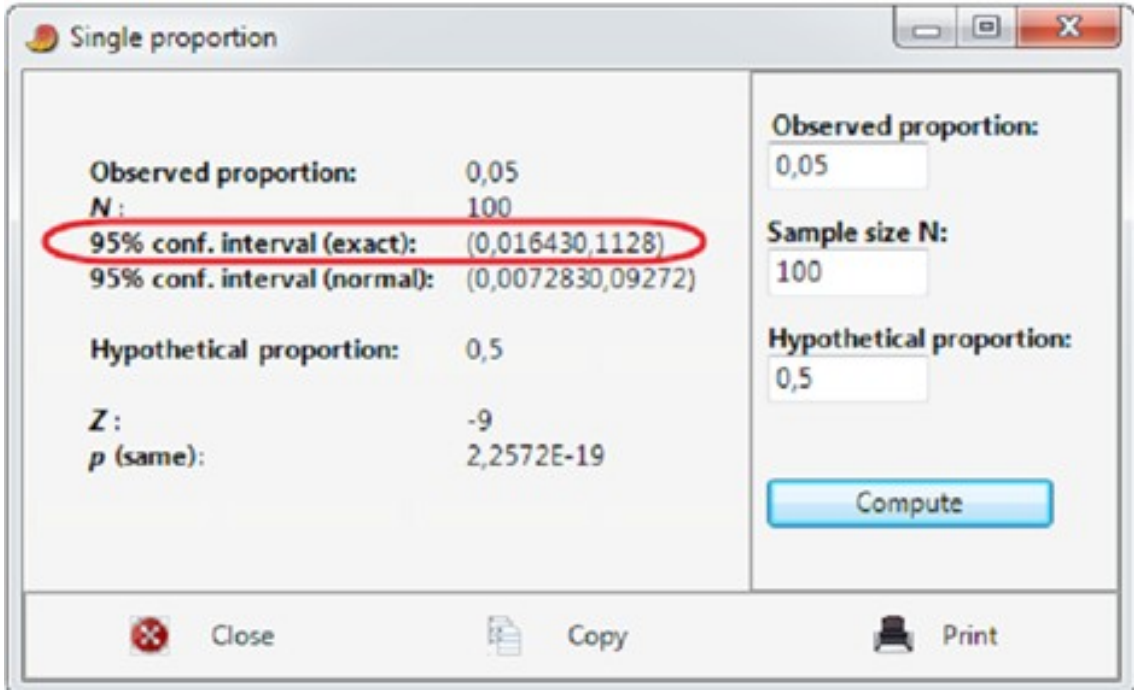
Изменение ***границ 95% ДИ***, рассчитанного ***по методу Уилсона***, в зависимости от объема выборки (значение доли в выборке – 25%).

П р и м е р. Из 100 проанализированных клеток 5 содержали хромосомные аберрации. **Задание:** найти среднюю частоту аберрантных клеток и 95% ДИ для неё.

В пакете PAST

Путь: ***Univariate — Single proportion test.***

Границы 95% ДИ по Клоппер — Пирсону показаны в строке 95% conf. interval (exact)



Средняя частота клеток с аберрациями составила:
5% (95% ДИ: от 1,6 % до 11,3%)

Биномиальные доверительные интервалы | **Доверительный интервал** 95% | ? Помощь

Объём выборки n	Количество k	Частота p	Границы ДИ нижняя	Границы ДИ верхняя	
20	1	5,0%	0,5%	21,1%	Метод Джеффриса (байесовский априорный интервал)
			0,1%	24,9%	Метод Клоппера - Пирсона (точный биномиальный метод)
			-4,6%	14,6%	Метод Вальда (нормальная аппроксимация)
			0,9%	23,6%	Метод Вилсона
			0,1%	27,9%	По распределению Пуассона (через хи-квадрат)
			-0,9%	25,4%	Метод Агрести - Коула (откорректированный метод Вальда)

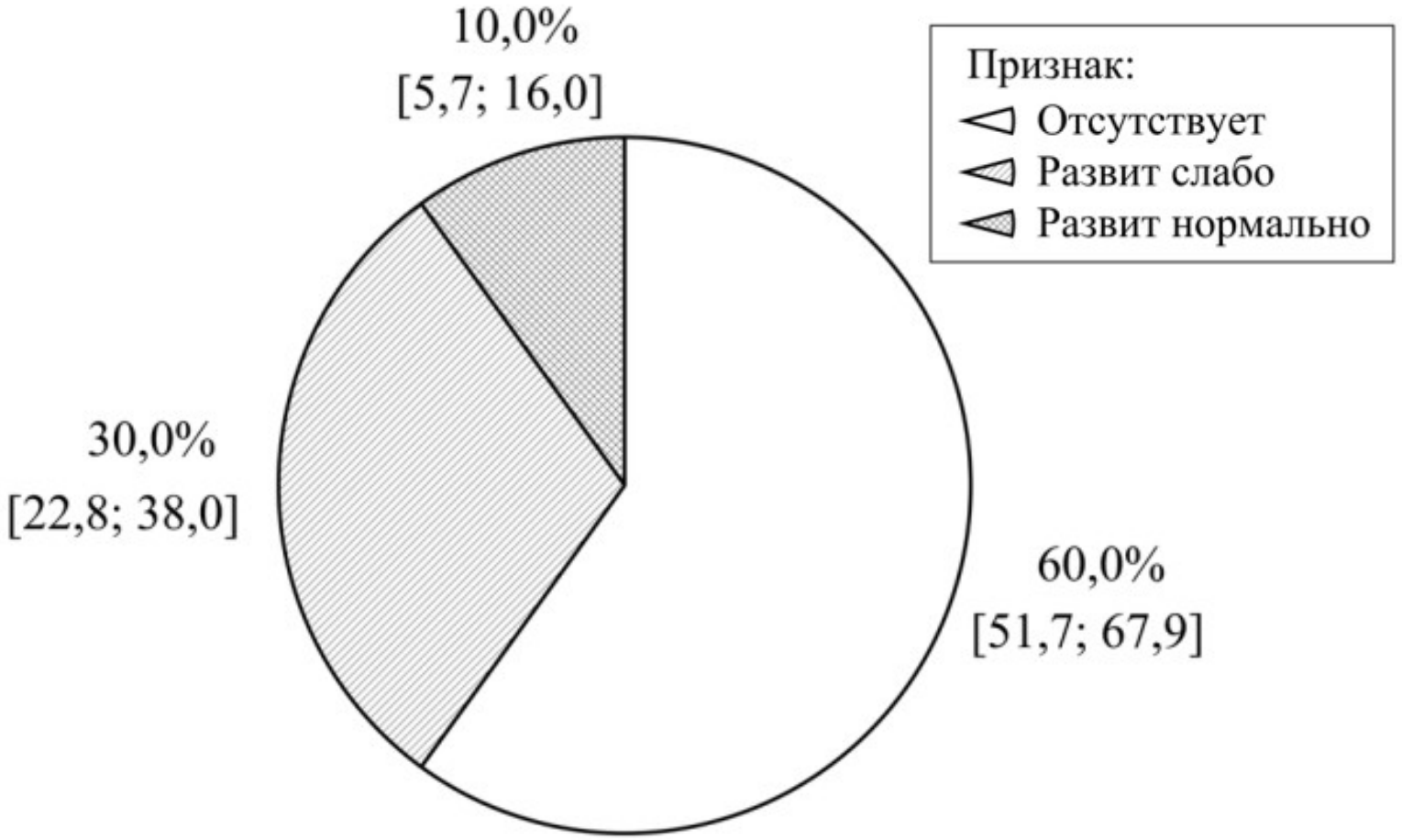
Биномиальные доверительные интервалы | **Доверительный интервал** 95% | ? Помощь

Объём выборки n	Количество k	Частота p	Границы ДИ нижняя	Границы ДИ верхняя	
100	5	5,0%	1,9%	10,6%	Метод Джеффриса (байесовский априорный интервал)
			1,6%	11,3%	Метод Клоппера - Пирсона (точный биномиальный метод)
			0,7%	9,3%	Метод Вальда (нормальная аппроксимация)
			2,2%	11,2%	Метод Вилсона
			1,6%	11,7%	По распределению Пуассона (через хи-квадрат)
			1,9%	11,5%	Метод Агрести - Коула (откорректированный метод Вальда)

Биномиальные доверительные интервалы | **Доверительный интервал** 95% | ? Помощь

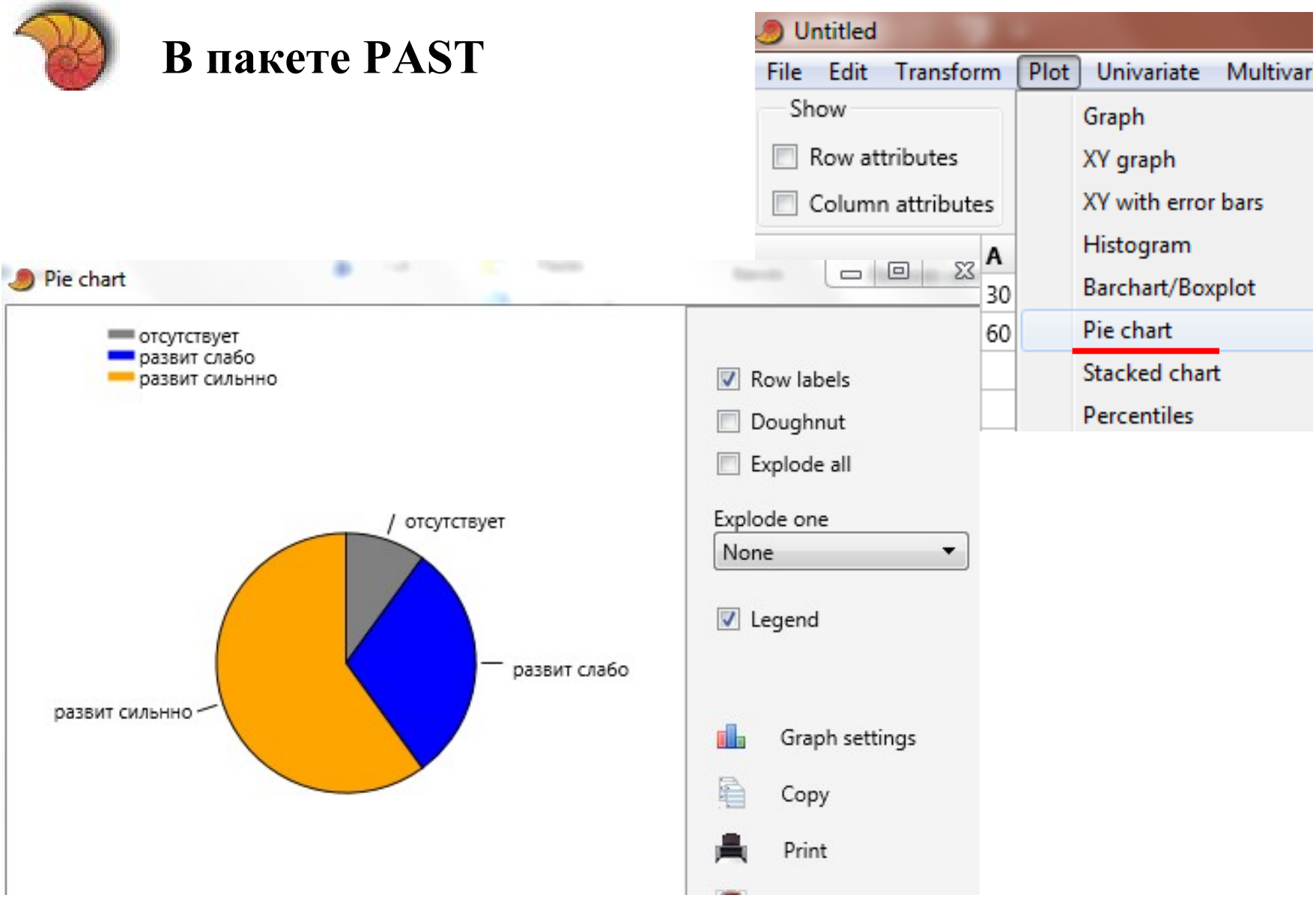
Объём выборки n	Количество k	Частота p	Границы ДИ нижняя	Границы ДИ верхняя	
1000	50	5,0%	3,8%	6,5%	Метод Джеффриса (байесовский априорный интервал)
			3,7%	6,5%	Метод Клоппера - Пирсона (точный биномиальный метод)
			3,6%	6,4%	Метод Вальда (нормальная аппроксимация)
			3,8%	6,5%	Метод Вилсона
			3,7%	6,6%	По распределению Пуассона (через хи-квадрат)
			3,8%	6,5%	Метод Агрести - Коула (откорректированный метод Вальда)

Графическое представление качественных (номинальных) признаков



Круговая диаграмма (Pie chart)

В пакете PAST



Параметры разброса для качественных данных: Индексы разнообразия *(indices of diversity)*

Показывают, насколько равномерно данные распределены по категориям. Разнообразие считается высоким, когда распределение более-менее равномерное, и низким, когда превалирует 1-2 категории

Индекс Шеннона-Винера (или Шеннона-Уивера)

$$H = -\sum_{i=1}^{k} p_i \log p_i$$

p = доля объектов в той или иной категории;
k – число категорий.

Этих индексов много для разных целей; это показатели ОПИСАТЕЛЬНОЙ статистики!

Сравнение двух выборок по качественным показателям

I. Сравнение независимых выборок

Если категории упорядочить нельзя, то есть если данные представлены номинальной шкалой, анализ проводят ***критериями согласия*** или современными ***рандомизационными критериями*** в ходе анализа ***таблиц сопряжённости*** *(ТС, contingency table).*

Методов анализа ТС предложено много; перечислим основные из них:

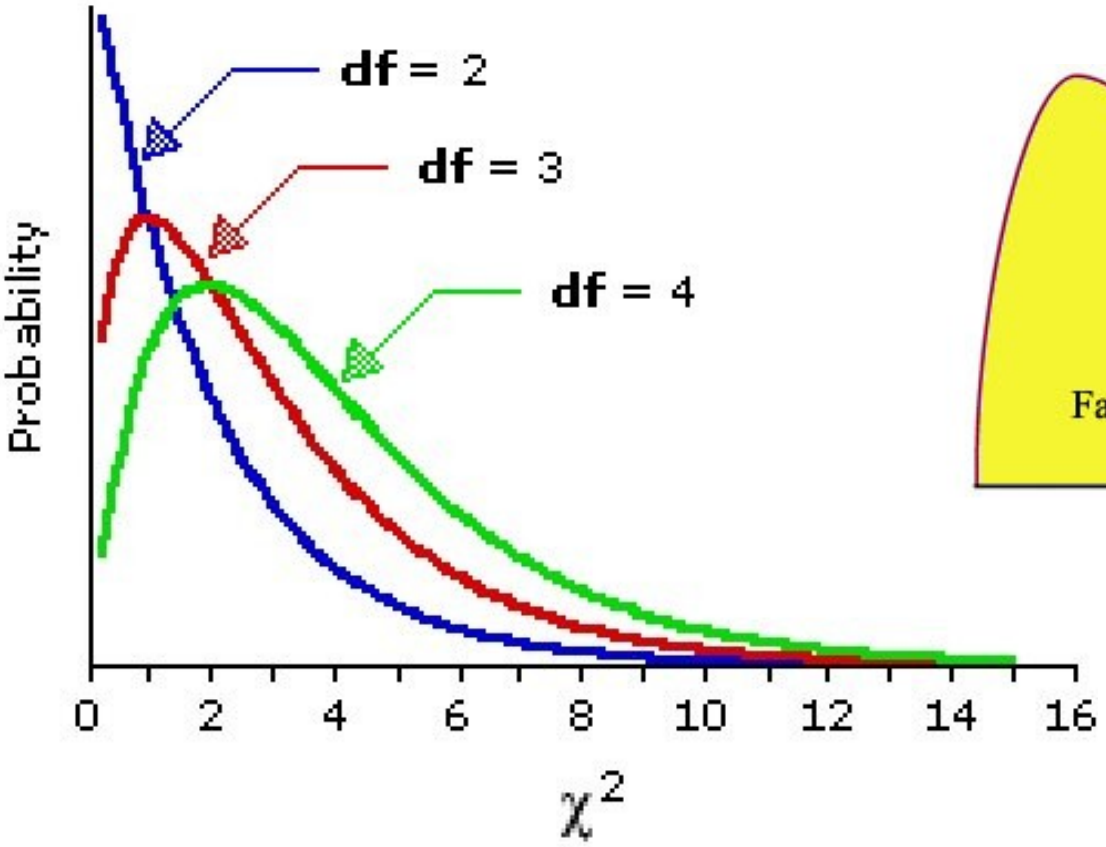
1) ***критерий хи-квадрат Пирсона*** *(Pearson's Chi-square test)*, обозначается $\chi 2$ Пирсона или просто $\chi 2$. Предложен Карлом Пирсоном ещё в 1901 г., но до сих пор популярен. Есть во всех статистических пакетах.

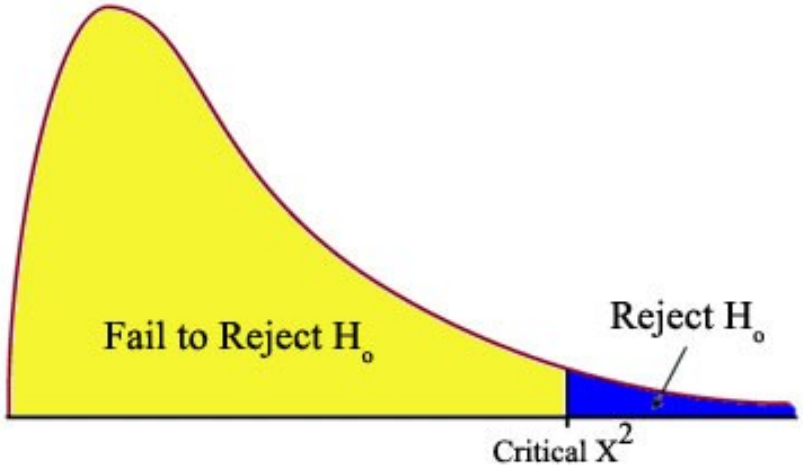
Karl Pearson

Для критерия существует *проблема* ***допустимого минимального значения в ячейке таблицы***. То есть если это *значение от 0 до 5* включительно, то *использовать* этот критерии *некорректно*. Для анализа таких слабонасыщенных таблиц можно применить точный метод Фишера;

Критерии согласия

Распределение статистики χ^2





Условия применения критерия $\chi 2$:

1. Включение в анализ только ***качественных (номинальных или порядковых) данных*** (возможно создание порядковых категорий из непрерывных данных);
2. ***Использование*** только абсолютных фактических и ожидаемых ***частот***;
3. ***Наблюдения*** должны быть ***независимы*** друг от друга;
4. ***Сравниваемые группы*** должны быть также ***независимы*** друг от друга (критерий не может быть использован в случае исследований типа «до – после»);
5. ***Ожидаемое*** (не фактическое) ***число наблюдений в любой из ячеек*** таблицы должно ***быть не менее 5 или 10 (для четырехпольных таблиц)***.
6. ***Доля ячеек*** таблицы с ожидаемым числом наблюдений ***менее 5 не должна превышать 20% (для многопольных таблиц).***

Познакомимся подробнее с критерием χ^2 Пирсона.

Пример. У пациентов определялся уровень общего холестерина в крови. Измерения разбиты на категории: 1) до 6,72 ммоль/л — «норма»; 2) свыше 6,72 ммоль/л — «повышенный» уровень. Параллельно отмечалось наличие заболеваний сердечно-сосудистой системы (ССС).

Уровень холестерина	Заболевания ССС		Всего
	Есть	Нет	
Повышенный	41	245	286
Норма	51	992	1043
Всего	92	1237	1329

Краевые частоты

Общее число наблюдений

Такая простейшая ТС называется таблицей 2×2 («два на два») или четырёхпольной таблицей.

Вопросы: отличаются ли лица с высоким и нормальным холестерином частотами заболеваний ССС? Если отличаются, то насколько сильно?

1. Расчёт относительных частот

Повышенный холестерин (ПХ).
Доля больных равна: 41 / 286 = 0,143, или 14,3 %.
Нормальный холестерин (НХ).
Доля больных равна: 51 / 1043 = 0,049, или 4,9 %.

Уровень холестерина	Заболевания ССС		Всего
	Есть	Нет	
Повышенный	41	245	286
Норма	51	992	1043
Всего	92	1237	1329

Риск заболеваний ССС в группе с ПХ - 0,143, в группе НХ - 0,049.

Необходимо убедиться, что эти два значения различаются статистически значимо

2. Сравнение двух частот с помощью критерия

2.1 Расчёт *ожидаемых частот (expected frequencies)*

По краевым суммам вычислим долю больных людей в популяции как 92/1329. Значит, в группе с ***повышенным холестерином*** должно наблюдаться 286 × 92 / 1 329, а в группе с ***нормальным холестерином*** — 1043 × 92 / 1329 больных людей.

Уровень холестерина	Заболевания ССС		Всего
	Есть	Нет	
Повышенный	41	245	286
Норма	51	992	1043
Всего	92	1237	1329

$$E_{ij} = \frac{\Sigma \text{ по строке} \times \Sigma \text{ по столбцу}}{\Sigma \text{ общая}}.$$

$E_{11} = 286 \times 92 / 1\,329 = 19{,}79 \approx 19{,}8$ (округлим до десятых);

$E_{21} = 1\,043 \times 92 / 1\,329 = 72{,}2$;

$E_{12} = 286 \times 1\,237 / 1\,329 = 266{,}2$;

$E_{22} = 1\,043 \times 1\,237 / 1\,329 = 970{,}8$.

Получим таблицу ожидаемых частот

Уровень холестерина	Заболевания ССС		Всего
	Есть	Нет	
Повышенный	19,8	266,2	286
Норма	72,2	970,8	1043
Всего	92	1237	1329

Таблица ожидаемых частот имеет такую же общую сумму и такие же краевые частоты, как исходная, однако сами *частоты внутри соответствуют нулевой гипотезе* — отсутствию различий между выборками.

2.2 Вычисление критерия χ^2 Пирсона

Критерий оценивает согласие наблюдаемых и ожидаемы

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

O_{ij} – фактическое количество наблюдений в ячейке ij

E_{ij} – ожидаемое число наблюдений в ячейке ij

i – номер строки (от 1 до r)

j – номер столбца (от 1 до c)

Вклад первой ячейки - наибольший

$\chi^2 = (41 – 19,8)^2 / 19,8 + (245 – 266,20)^2 / 246,2 + (51 – 72,2)^2 / 72,2 + (992 – 970,8)^2 / 970,8 =$ **22,70** + 1,69 + 6,22 + 0,46 = 31,07.

У людей с высоким холестерином больных было намного больше, чем ожидалось в соответствии с нулевой гипотезой.

2.3 Расчёт степеней свободы:

$$df = (n_{\text{строк}} - 1)(n_{\text{столбцов}} - 1);$$
$$df = (2 - 1)(2 - 1) = 1.$$

2.4 Оценка статистической значимости.

df	0.05	0.01	0.001
1	3.841	6.635	10.828
2	5.991	9.210	13.816
3	7.815	11.345	16.266
4	9.488	13.277	18.467
5	11.070	15.086	20.515
6	12.592	16.812	22.458
7	14.067	18.475	24.322
8	15.507	20.090	26.125

Полученное значение χ^2 при нужном числе степеней свободы сравнивается с табличным.

$$\chi^2_{cv} = 10.83$$

31.07 >>10.83 Значит p << 0,001 → отвергаем H_0

$\chi^2_{(1)}$ = 31,07; P <<0,001 (различия высоко статистически значимы)

χ^2 распределение непрерывное.

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

Вводиться **поправка Йейтса на непрерывность (Yates' continuity correction)** - уменьшали каждую разность между наблюдаемой и ожидаемой частотами в формуле на 0,5.

Формула для расчета критерия χ^2 с поправкой Йейтса следующая:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(|O_{ij} - E_{ij}| - 0,5)^2}{E_{ij}}$$

Делает тест более консервативным.

3. Оценка величины различий

В качестве показателей ***величины эффекта*** *(effect size)* для различий частот используется несколько мер.

3.1. ***Разность рисков*** *(Risk difference)*.

Показывает, насколько риск события в одной группе больше или меньше по сравнению с риском в другой. Рассчитывается как простая арифметическая разность рисков, рассчитанных в **п. 1**. В нашем случае она равна: 0,143 – 0,049 = 0,094.

п. 1

Повышенный холестерин (ПХ).
Доля больных равна: 41 / 286 = 0,143, или 14,3 %.
Нормальный холестерин (НХ).
Доля больных равна: 51 / 1043 = 0,049, или 4,9 %.

Уровень холестерина	Заболевания ССС		Всего
	Есть	Нет	
Повышенный	41	245	286
Норма	51	992	1043
Всего	92	1237	1329

Риск заболеваний ССС в группе с ПХ - 0,143, в группе НХ - 0,049.

3.2. ***Отношение рисков*** (или относительный риск, *Risk ratio, Relative risk — RR*).

Показывает, во сколько раз риск (частота) события в одной группе больше или меньше по сравнению с риском в другой. Для равных рисков *RR* = 1. В нашем случае *RR* = 0,143 / 0,049 = 2,92.

Повышенный холестерин (ПХ).
Доля больных равна: 41 / 286 = 0,143, или 14,3 %.
Нормальный холестерин (НХ).
Доля больных равна: 51 / 1043 = 0,049, или 4,9 %.

Уровень холестерина	Заболевания ССС		Всего
	Есть	Нет	
Повышенный	41	245	286
Норма	51	992	1043
Всего	92	1237	1329

Риск заболеваний ССС в группе с ПХ - 0,143, в группе НХ - 0,049.

Т.е. с увеличением содержания холестерина в сыворотке крови риск заболеваний ССС увеличиваются в **2,92** раза.

3.3. ***Отношение шансов*** *(Odds ratio — OR)*.

Показывает, во сколько раз шанс события в одной группе больше или меньше по сравнению с шансом в другой. **Шанс** — отношение вероятности события к его альтернативе.

В нашем случае при повышенном холестерине вероятность иметь заболевания ССС составляет 41/1329, а не иметь (альтернатива) — 245/1329. Таким образом, ***шанс*** иметь заболевания ССС ***при высоком холестерине*** составляет: 41/245 = 0,1673

Шанс иметь заболевания ССС ***при нормальном холестерине***: 51/992 = 0,0514

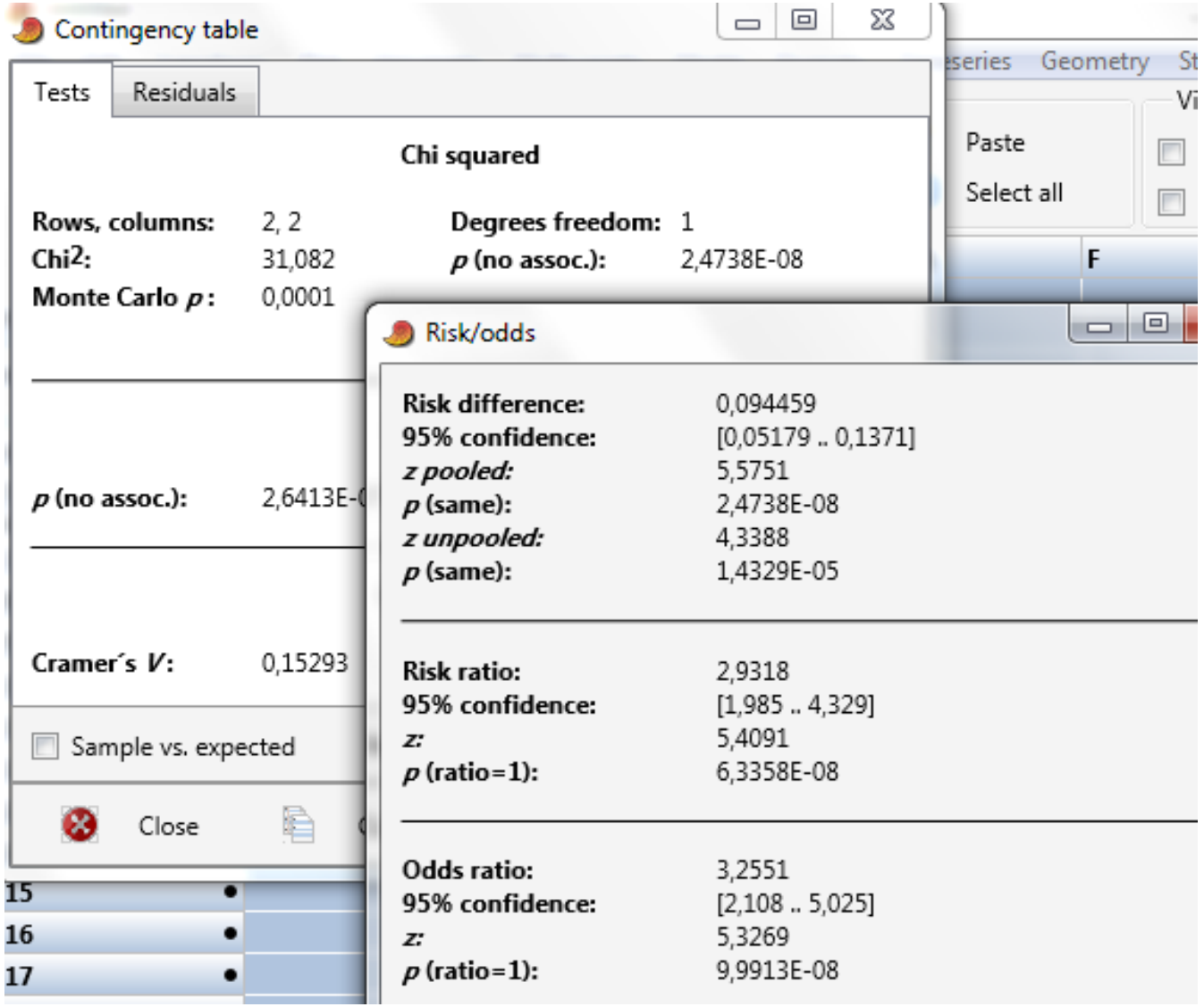
Уровень холестерина	Заболевания ССС		Всего
	Есть	Нет	
Повышенный	41	245	286
Норма	51	992	1043
Всего	92	1237	1329

Следовательно, ***отношение шансов*** составляет:

OR = 0,16735/0,05141 = 3,26.

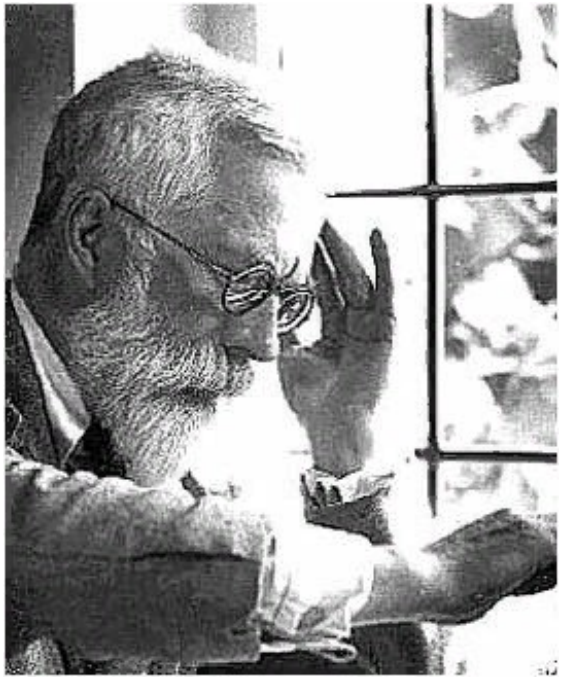
С увеличением содержания холестерина в сыворотке крови шансы заболеваний ССС увеличиваются в **3,26** раза.

В пакете PAST



2) ***точный метод Фишера*** *(Fisher's exact test)* предложен Р. Фишером в 1954 г. для анализа слабонасыщенных таблиц и до сих пор популярен.

Однако теоретически он не очень хорош: критерий основан на гипергеометрическом распределении, хотя используется для анализа ТС с данными, имеющими биномиальное или полиномиальное распределение.



Ronald Fisher

В настоящее время вместо него корректнее пользоваться рандомизационными критериями

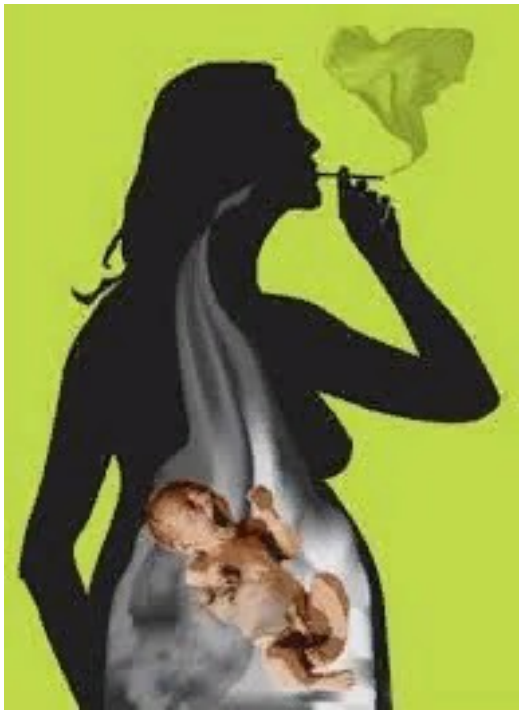
Пример.

Изучается зависимость частоты рождения детей ***с врожденными пороками развития*** (ВПР) ***от курения матери*** во ***время беременности***.

Две группы беременных женщин:

1.Экспериментальная - женщины, курившие в первом триместре беременности,

2.Контрольная - женщины, ведущие здоровый образ жизни на протяжении всей беременности.



	Исход есть (Наличие ВПР)	Исхода нет (Отсутствие ВПР)	Всего
Фактор риска есть (Курящие)	A = 10	B = 70	(A + B) = 80
Фактор риска отсутствует (Некурящие)	C = 2	D = 88	(C + D) = 90
Всего	(A + C) = 12	(B + D) = 158	(A + B + C + D) = 170

Точный критерий Фишера рассчитывается по формуле:

$$P = \frac{(A+B)! \cdot (C+D)! \cdot (A+C)! \cdot (B+D)!}{A! \cdot B! \cdot C! \cdot D! \cdot N!}$$

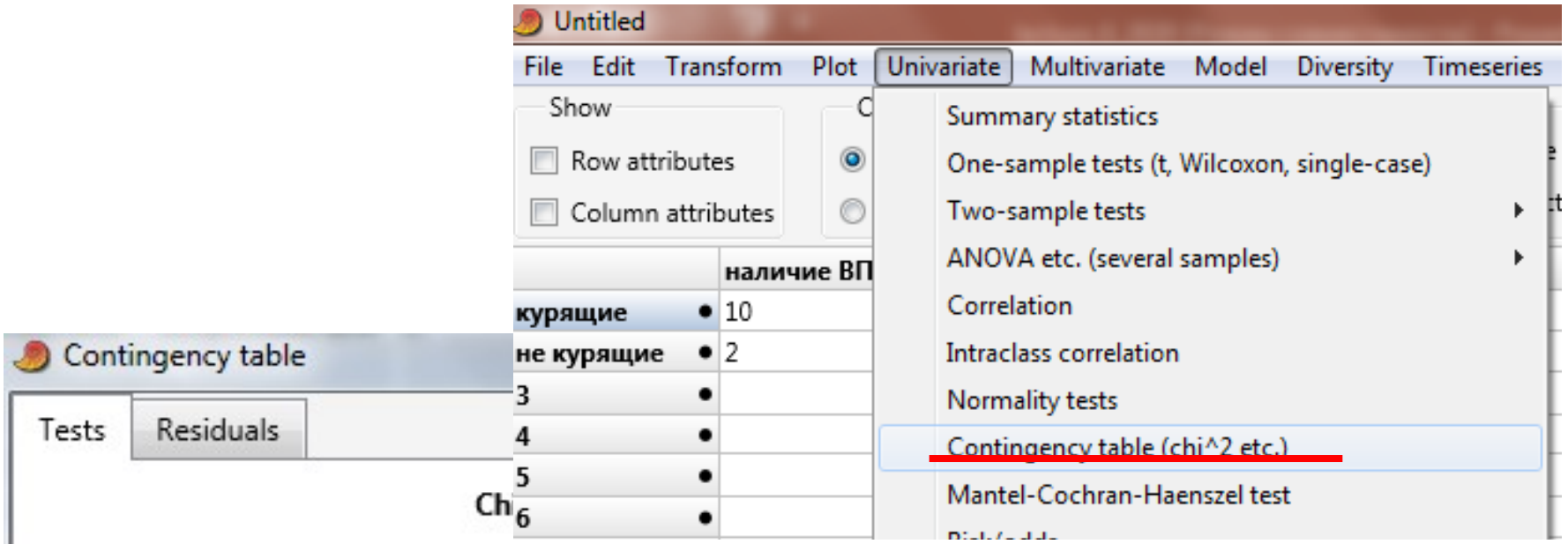
N - общее число исследуемых в двух группах;
! (факториал) - произведение числа на последовательность чисел, каждое из которых меньше предыдущего на 1 (например, 4! = 4 · 3 · 2 · 1). В результате вычислений находим, что P = 0,0137.

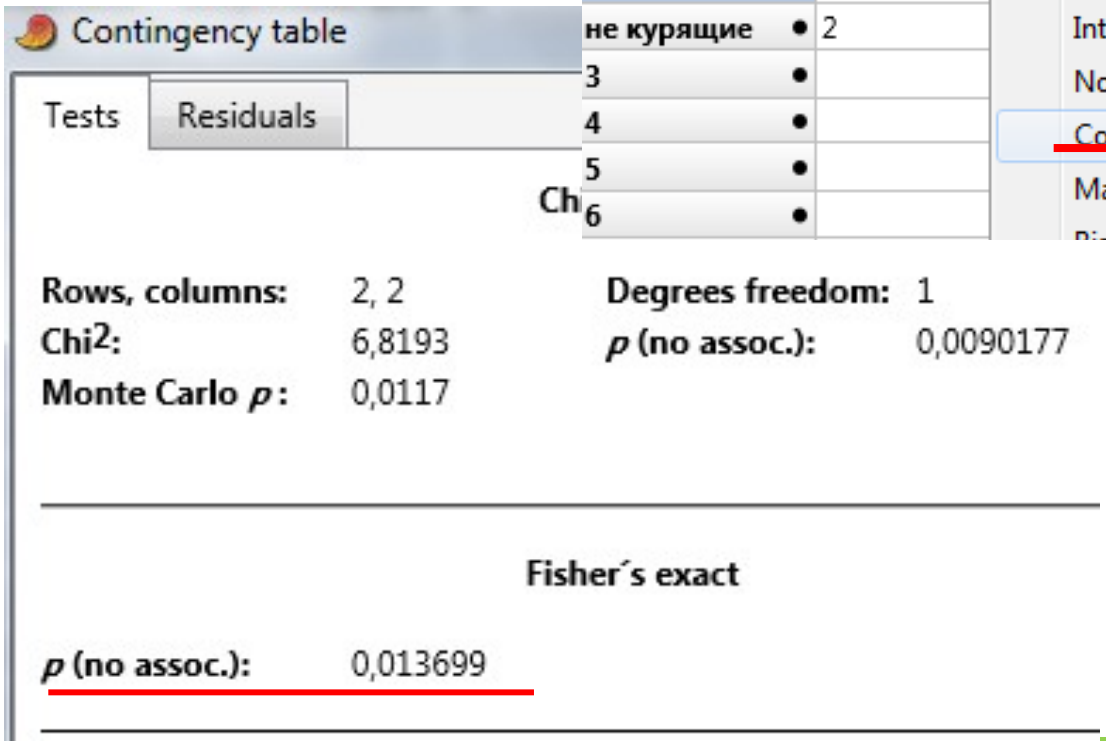
Полученное значение ***0,0137*** и есть уровень значимости различий сравниваемых групп по частоте развития ВПР плода.

Сопоставим данное число с критическим уровнем значимости, обычно принимаемым в медицинских исследованиях за ***0,05.***

В нашем примере ***P < 0,05***, в связи с чем делаем вывод о наличии взаимосвязи курения и вероятности развития ВПР плода. ***Частота возникновения врожденной патологии у детей курящих женщин статистически значимо выше, чем у некурящих***.

В пакете PAST





3) ***рандомизационный критерий Монте-Карло*** *(permutation test, Monte Carlo test)*, случайным образом генерирует большое число (десятки и сотни тысяч) ТС с такими же краевыми частотами, как у исходной.

Годом рождения термина «метод Монте-Карло» считается 1949 год, когда в свет вышла статья Н. Метрополиса и С. Улама «Метод Монте-Карло».

Название метода происходит от района Монте-Карло, известного своими казино.

Стал практически доступен только с появлением компьютеров уровня 1990-х гг.

4) ***точный рандомизационный (перестановочный) критерий (Exact permutation test)***

— похож на 3), но генерируются не случайные таблицы с такими же краевыми частотами, а в точности все возможные. Для ТС с большим числом наблюдений это может быть непосильной задачей даже для современных компьютеров, и тогда можно использовать предыдущий критерий.

Точный рандомизационный критерий — наиболее точный и современный метод, который рекомендуется использовать во всех случаях, а особенно — для анализа слабонасыщенных таблиц.

Не во всех статистических пакетах он есть.

II. Сравнение зависимых выборок

В случае качественных номинальных признаков две зависимые выборки сравнивают ***критерием Макнемара*** *(McNemar testof symmetry)*.

Для не слишком малых выборок статистика критерия имеет распределение хи-квадрат с одной степенью свободы.

В случае малых выборок критерий становится слишком либеральным, поэтому вводится ***поправка Эдвардса на непрерывность*** *(Edwards' continuity correction)*.

Пример. Исследование «до – после».
Оценка наличия изжоги до начала и после окончания курса комплексного лечения язвенной болезни таблице.

Влияние курса лечения язвенной болезни на наличие у пациентов изжоги.

Наличие/отсутствие признака		После лечения		
		Отсутствие изжоги	Наличие изжоги	Всего
До лечения	Наличие изжоги	48 (A)	10 (B)	58 (A + B)
	Отсутствие изжоги	12 (C)	6 (D)	18 (C + D)
	Всего	60 (A + C)	16 (B + D)	54 (N = A + D)

В ***клетке A*** - количество ***благоприятных исходов*** после воздействия фактора (исчезновение изжоги после курса лечения)

В ***клетке D*** – количество ***неблагоприятных исходов*** (после курса лечения изжоги, которой изначально не было).

Для расчета критерия Мак-Нимара используются данные только в этих двух клетках и значение, равное сумме значений этих двух клеток (N = A + D). Значения в клетках B и C, также, как и общий объем выборки, при расчете критерия Мак-Нимара не используются.

Для проверки гипотезы в случае, когда N > 50 (сумма значений в ячейках A и D, но не объем выборки), рассчитывается значение χ2 по упрощенной формуле с числом степеней свободы, равным 1:

$$\chi^2 = \frac{(|A - D| - 1)^2}{A + D}$$

где | A – D | – абсолютное значение (модуль) разности значений соответствующих клеток (модуль разности), а единица вычитается с целью выполнения поправки на непрерывность.

Рассчитываем фактическое значение χ^2:

$$\chi^2 = \frac{(|48 - 6| - 1)^2}{48 + 6} = 31{,}13$$

$\chi^2(31{,}13) >> \chi^2_{cv}$ (10.83)

H_0 отвергается на уровне значимости ($p << 0{,}001$).

Таким образом, ***предложенное комплексное лечение*** язвенной болезни ***статистически значимо уменьшает количество пациентов, страдающих изжогой***.

Оценка силы различий.

В качестве показателя величины эффекта используется отношение шансов.

<u>Рассчитывается как отношение наддиагонального и поддиагонального элементов таблицы</u>: OR = A / D.

В нашем случае OR = 48 / 6 = 8

Шансы выраженного лечебного эффекта предложенного комплексное лечение язвенной болезни в 8 раз выше, чем без него.

Сравнение трёх и более выборок по качественным показателям

Таблицы сопряжённости (ТС) больше, чем таблицы 2×2.

Это ***таблицы* сопряжённости r×c**

r — rows — ряды, строки c — columns — колонки, столбцы)

I. Сравнение независимых выборок

Этап 1. ***Омнибусный критерий***.

Проверяет согласие наблюдаемых и ожидаемых частот для всех ячеек таблицы.

Используются критерии согласия или современные рандомизационные критерии, рассмотренные для таблиц 2×2.

Если H_0 принимается, то констатируем отсутствие различий.

Если H_0 отклоняется ($p \leq 0{,}05$), то далее:

Этап 2. Вместо апостериорных сравнений для таблиц сопряжённости проводят ***выявление ячеек, давших наибольший и неслучайный вклад в отклонение от нулевой гипотезы***.

С помощью расчёта:

отклонений Фримана — Тьюки
(FreemanTukey deviation, FT_{dev})
или
согласованных стандартизованных остатков
(Adjusted residuals, AR), называемых также ***остатками Хабермана***

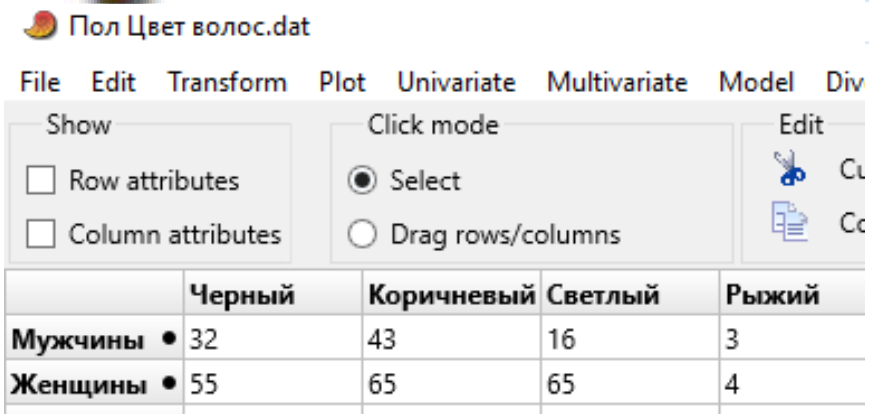
Пример. Среди 282 членов актёрской ассоциации был проведён социологический опрос. При этом отмечался пол и цвет волос респондента.

	Цвет волос			
	Чёрный	Коричневый	Светлый	Рыжий
Мужчины	32	43	16	3
Женщины	55	65	64	4

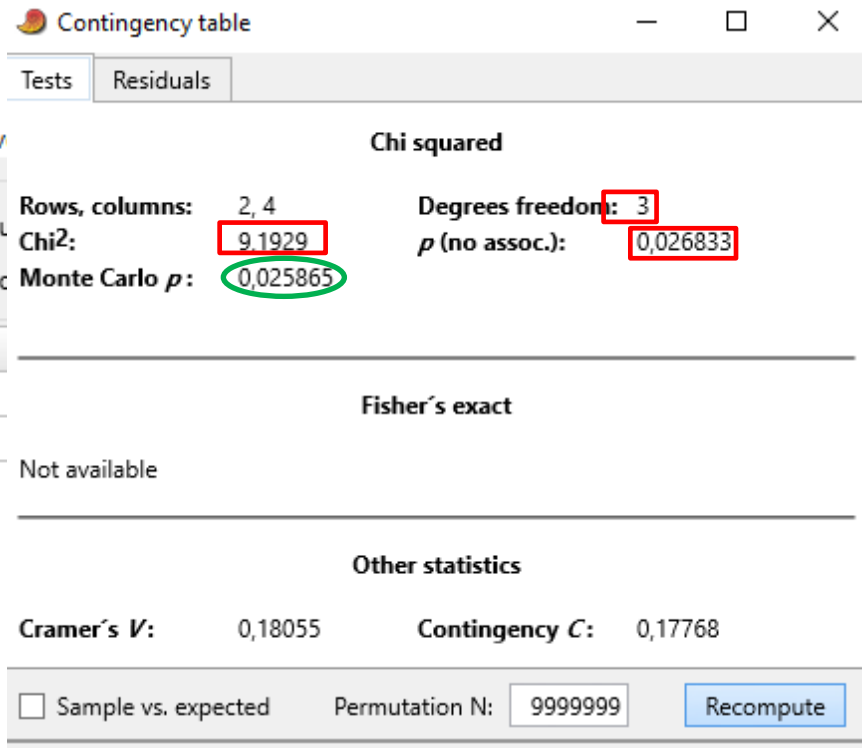
Задание: оценить различия между мужчинами и женщинами по соотношению обладателей волос разного цвета. Если различия есть, то установить, в чём они?

Вопрос можно переформулировать и для задачи сравнения нескольких групп: различаются ли обладатели волос разного цвета соотношением полов?

В пакете PAST



	Черный	Коричневый	Светлый	Рыжий
Мужчины	32	43	16	3
Женщины	55	65	65	4



Если в таблице есть значения 5 и менее (наш случай) — выписываем *p*, вычисленное рандомизационной процедурой Монте-Карло.

Вывод промежуточный: мужчины и женщины статистически значимо различались соотношением обладателей волос разного цвета: критерий хи-квадрат Пирсона $\chi^2_{(3)} = 9{,}19$; $p = 0{,}026$.

Различия обнаружены, но пока непонятно, в чём именно они заключаются.

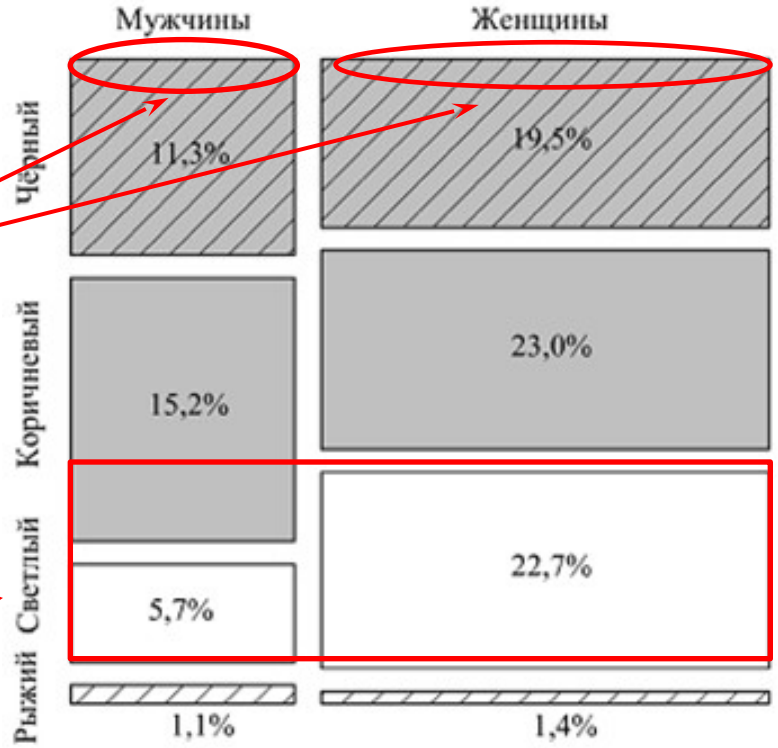
Далее

1)расчёт относительных частоты (в процентах)

2)выявление ячеек, давших неслучайный вклад в статистику критерия

Построим **мозаичный график,** где площадь плитки пропорциональна частоте.

Видно - в выборке почти в 2 раза больше женщин. Наиболее сильные различия между полами наблюдались по светлому цвету волос: женщин-блондинок было заметно больше.



1) Расчёт относительных частот.

Например, для ячейки 11 (r = 1, c = 1 — мужчины с чёрными волосами) имеем: 32 / (32 + 43 + 16 + 3) = 32 / 94 = 0,340, или 3

	Цвет волос			
	Чёрный	Коричневый	Светлый	Рыжий
Мужчины	32	43	16	3
Женщины	55	65	64	4

Итоговая таблица в процентах:

	Цвет волос				
	Чёрный	Коричневый	Светлый	Рыжий	Всего
Мужчины	34,0	45,7	17,0	3,2	100
Женщины	29,3	34,6	34,0	2,1	100

Наиболее сильные различия наблюдаются по доле обладателей светлых волос:

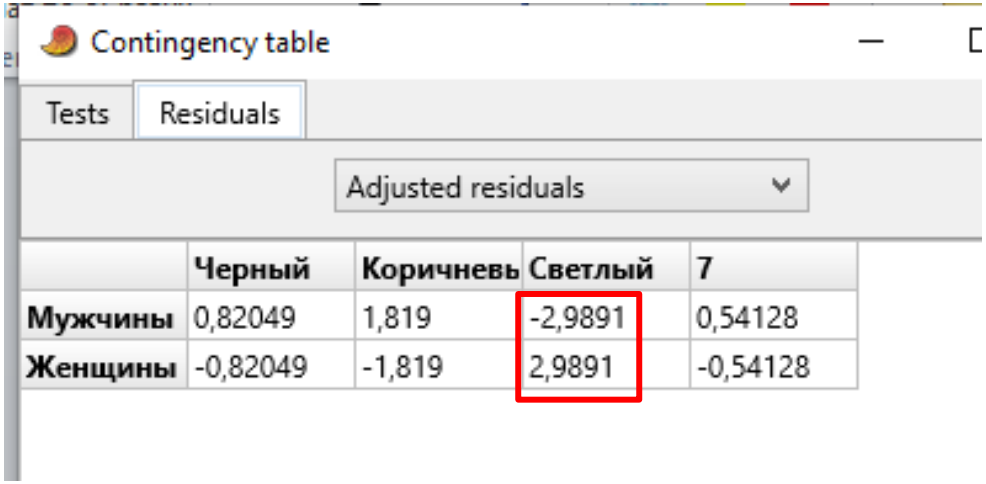
среди мужчин - 17,0 %

среди женщин - 34,0 %

2) Выявление ячеек, давших неслучайный вклад в статистику критерия

Расчёт согласованных стандартизованных остатков.

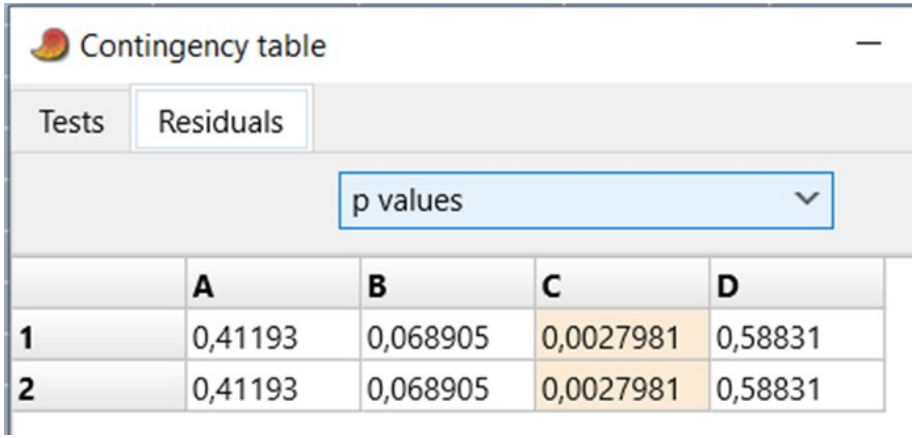
Знак остатков указывает на направление отклонения.



	Черный	Коричневь	Светлый	7
Мужчины	0,82049	1,819	-2,9891	0,54128
Женщины	-0,82049	-1,819	2,9891	-0,54128

плюс - несколько больше, минус - несколько меньше, чем ожидалось в соответствии с нулевой гипотезой.

Статистическая значимость остатков.



	A	B	C	D
1	0,41193	0,068905	0,0027981	0,58831
2	0,41193	0,068905	0,0027981	0,58831

II. Сравнение зависимых выборок

При зависимых выборках в ячейках таблицы два или более раз фигурируют одни и те же объекты исследования (образцы, животные, люди и т. д.).

Анализ обычно проводят с использованием ***критерия симметрии Боукера (Bowker's symmetry test)***, который является обобщением критерия Макнемара на случай нескольких зависимых выборок, и может называться в критерием Макнемара — Боукера.

Реже применяют критерии ***краевой однородности (marginal homogeneity tests) Стюарта — Максвелла (Stuart-Maxwell test) или Бхапкара (Bhapkar's test)***.

Их числовые значения близки и на практике все они обычно приводят к одинаковым выводам.

Пример. Разные типы катаракты - разная этиология.

Вероятно должно наблюдаться соответствие между типами катаракты, развивающейся в левом и правом глазу больного.

В исследования у пациентов глазной клиники, имеющих катаракту обоих глаз, регистрировался её тип в левом и правом глазу. Получены следующее данные:

Левый глаз	Правый глаз			Всего
	Ядерная	Кортикальная	Субкапсулярная	
Ядерная	18	11	6	35
Кортикальная	3	15	7	25
Субкапсулярная	10	9	16	35
Всего	31	35	29	95

Задание: определить, отличаются ли левый и правый глаза по частотам развития катаракты трёх типов.

Критерий Боукера оценивает *нарушение симметрии наддиагональной и поддиагональной частей таблицы*.

Критерии краевой однородности Стюарта — Максвелла и Бхапкара оценивают *различия в краевых частотах.*

***Расчёт критерия Боукера*.**

1. Находим диагональ таблицы, значения в ячейках которой указывают на сходство зависимых выборок. Они не помогают нам вы явить различия между выборками, а потому не участвуют в расчётах: зачеркнём диагональю значения.

Левый глаз	Правый глаз			Всего
	Ядерная	Кортикальная	Субкапсулярная	
Ядерная	~~18~~	11	6	35
Кортикальная	3	~~15~~	7	25
Субкапсулярная	10	9	~~16~~	35
Всего	31	35	29	95

2. Находим пары значений, симметричные относительно диагонали, и подставляем их в формулу критерия Боукера.

$$\chi^2 = \sum \frac{(f_{ij} - f_{ji})^2}{f_{ij} + f_{ji}}. \quad \chi^2 = \frac{(11-3)^2}{11+3} + \frac{(6-10)^2}{6+10} + \frac{(7-9)^2}{7+9} = \underline{4{,}5714} + 1{,}0000 + 0{,}2500 = 5{,}8214.$$

В ходе расчёта критерия подчеркнём слагаемое, давшее максимальный вклад в статистику критерия: 4,5714. Значение статистики округлим до сотых: 5,82.

3. Рассчитываем степени свободы как число слагаемых в критерии Боукера или по формуле *df* = *i* (*i* – 1) / 2, где *i*— число категорий:

$$df = 3 \times (3 - 1) / 2 = 3 \times 1 = 3.$$

Оценка статистической значимости.

Полученное значение χ^2 с табличным. В нашем случае 5,82 <7,81 и p> 0,05, следовательно, различия незначимы.

Наиболее сильные различия наблюдались для пары ядерной и кортикальной катаракты: 4,5714 / 5,8214 = 0,785, или 78,5 % всех различий между правым и левым глазом.

$$\chi^2 = \frac{(11-3)^2}{11+3} + \frac{(6-10)^2}{6+10} + \frac{(7-9)^2}{7+9} = \underline{4,5714} + 1,0000 + 0,2500 = 5,8214.$$

Если бы различия были статистически значимы, то мы бы считали, что при ядерной катаракте в левом глазу в правом чаще развивается кортикальная катаракта: отношение шансов *OR* = 11 / 3 = 3,67.

В нашем случае различия не были статистически значимыми.

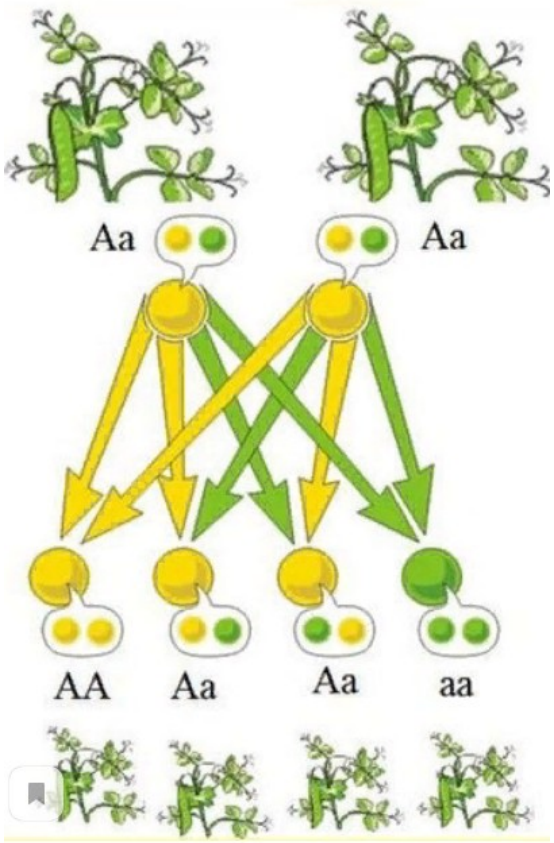
Констатируем отсутствие каких бы то ни было различий между правым и левым глазом в развитии катаракты трёх типов.

Сравнение экспериментальных и теоретических частот

При скрещивании двух гетерозиготных гибридов гороха было получено потомство: 84 желтых и 16 зеленых .

H_0: выборка получена из популяции, где соотношение желтых и зеленых – 3:1.

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



	желтые	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

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

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

p=0.038

H_0 отвергнута, т.е. соотношение гороха не соответствует ожидаемому

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

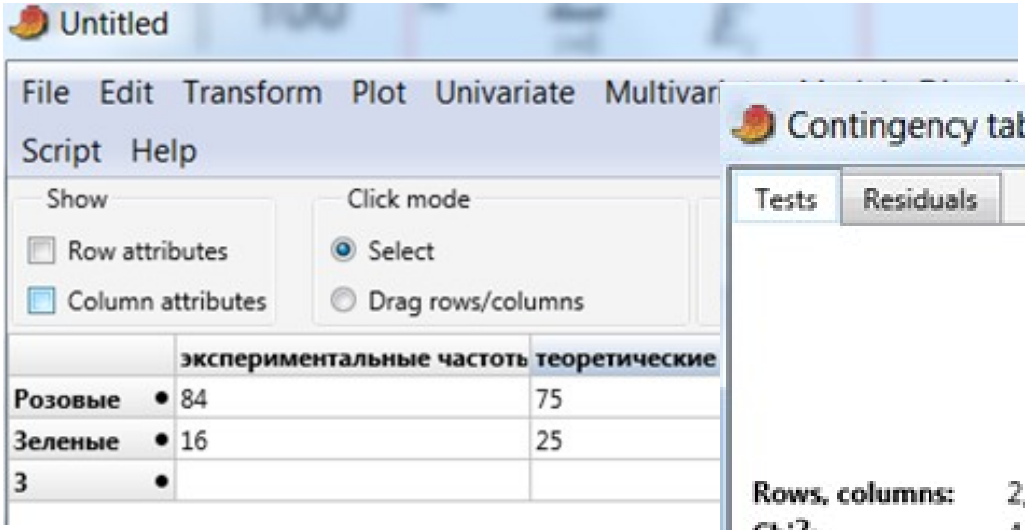
} промежуточных значений не может быть для данных ожидаемых частот

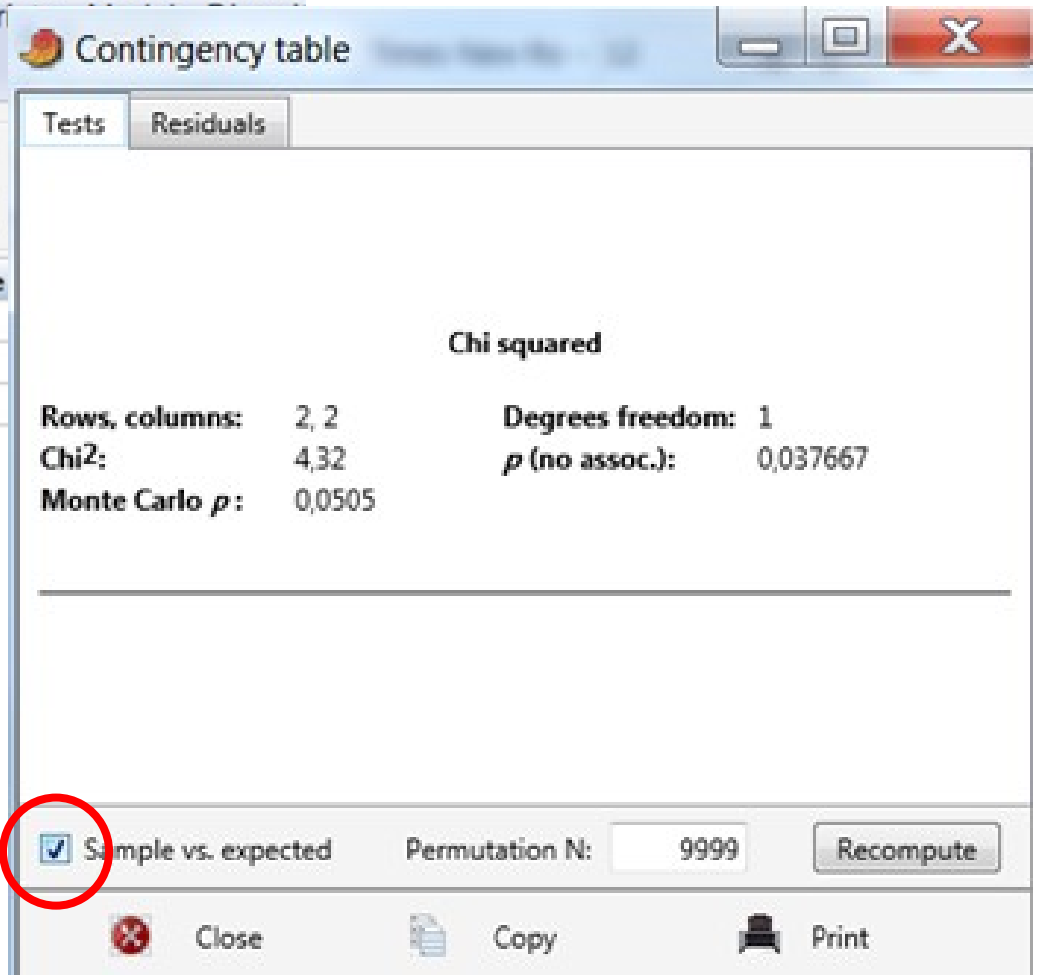
Но χ^2 распределение непрерывное. И для заданного уровня значимости *p* мы не найдём точно соответствующего ему значения χ^2.

χ^2 с поправкой Йейтса:

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

В пакете PAST





Вывод: соотношение мышей в эмпирической выборке не соответствует ожидаемому соотношению 3:1 ($\chi^2_{(1)}$ = 4,32; P = 0,038).

Анализ связей между номинальными показателями

Традиционно связь между двумя и более качественными номинальными показателями — ***ассоциацией*** *(association).*

Для качественных номинальных признаков ***оценка силы связи (ассоциации)*** признаков проводится ***по таблицам сопряжённости***.

Для оценки ***ассоциации*** на
1 этапе мы рассчитываем ***критерии типа хи-квадрат***, **2 этапе** — специфические ***меры ассоциации***.

Наиболее часто используются:
коэффициент сопряжённости Пирсона (в том числе — в модификации Сакоды)
коэффициенты ассоциации Крамера или ***Чупрова***

Коэффициент ассоциации Крамера

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, c-1)}},$$

где min (r – 1, c – 1) — минимальное из двух значений: числа рядов или числа колонок таблицы за вычетом единицы (для таблицы 2×2 это всегда 1); n – общий объем выборки.

Коэффициент сопряжённости Пирсона

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

Основные критерии, используемые для оценки силы связи между номинальными переменными

Наименование критерия	Область применения	Формула расчета
Критерий ϕ («фи»)	Четырехпольные таблицы	$\varphi = \sqrt{\frac{\chi^2}{n}}$ где χ^2 – значение критерия χ^2, n – объем выборки
Критерий Крамера (V)	Четырехпольные и многопольные таблицы	$V = \sqrt{\frac{\chi^2}{n \cdot (r-1) \cdot (c-1)}}$ где χ^2 – значение критерия χ^2, n – объем выборки, r – количество рядов (строк), c – количество столбцов
Коэффициент сопряженности Пирсона (C)	Четырехпольные и многопольные таблицы	$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$ $C' = \frac{C}{\sqrt{\frac{r-1}{r}}}$ где χ^2 – значение критерия χ^2, n – объем выборки, r – количество рядов (или столбцов, так как формула предназначена только для симметричных таблиц). C' рассчитывается для симметричных таблиц (формула Sakoda)
Критерий Чупрова (K)	Многопольные таблицы размером не более 5×5	$K = \sqrt{\frac{\chi^2}{n\sqrt{(r-1) \cdot (c-1)}}}$ где χ^2 – значение критерия χ^2, n – объем выборки, r – количество рядов (строк), c – количество столбцов
Критерий λ Гудмена-Краскела	Четырехпольные таблицы	$\lambda = \frac{\sum f_i - f_t}{n - f_d}$ где f_i – наибольшие числа в ячейках в каждом из классов независимой переменной; f_d – наибольший из маргинальных итогов (сумм) зависимой переменной, n – объем выборки

Все эти коэффициенты изменяются от 0 (отсутствие связи) до +1 (максимально возможная связь).

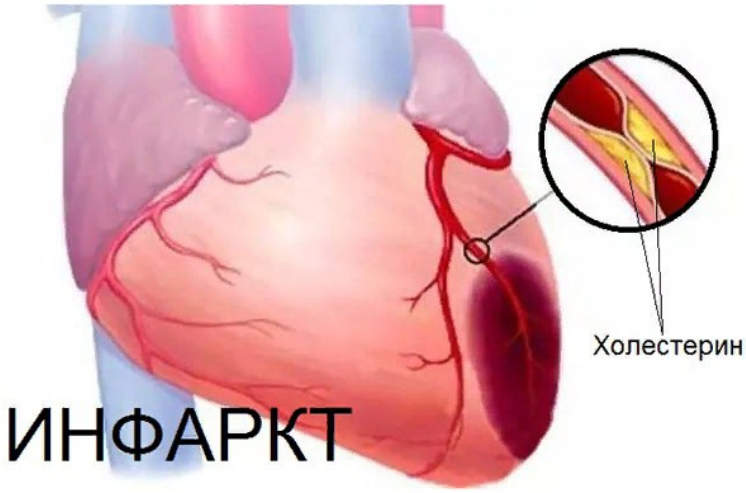
Интерпретация значений критериев φ и V Крамера согласно рекомендациям Rea и Parker

Значение критериев φ или V Крамера	Сила взаимосвязи
<0,1	Несущественная
0,1 – <0,2	Слабая
0,2 – <0,4	Средняя
0,4 – <0,6	Относительно сильная
0,6 – <0,8	Сильная
0,8 –1,0	Очень сильная

Если есть возможность установить направление связи, то о нём судят по частотам ТС.

Пример. Проанализируем уже знакомые данные по холестерину и заболеваниям сердца:

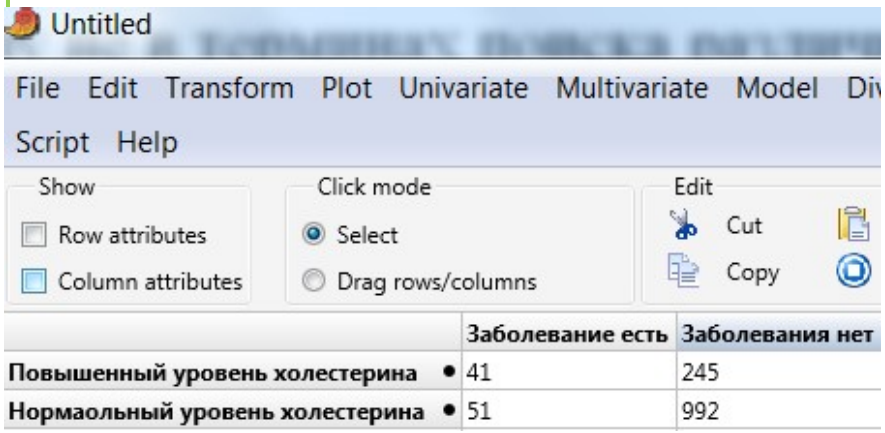
Уровень холестерина	Заболевания ССС		Всего
	Есть	Нет	
Повышенный	41	245	286
Норма	51	992	1043
Всего	92	1237	1329



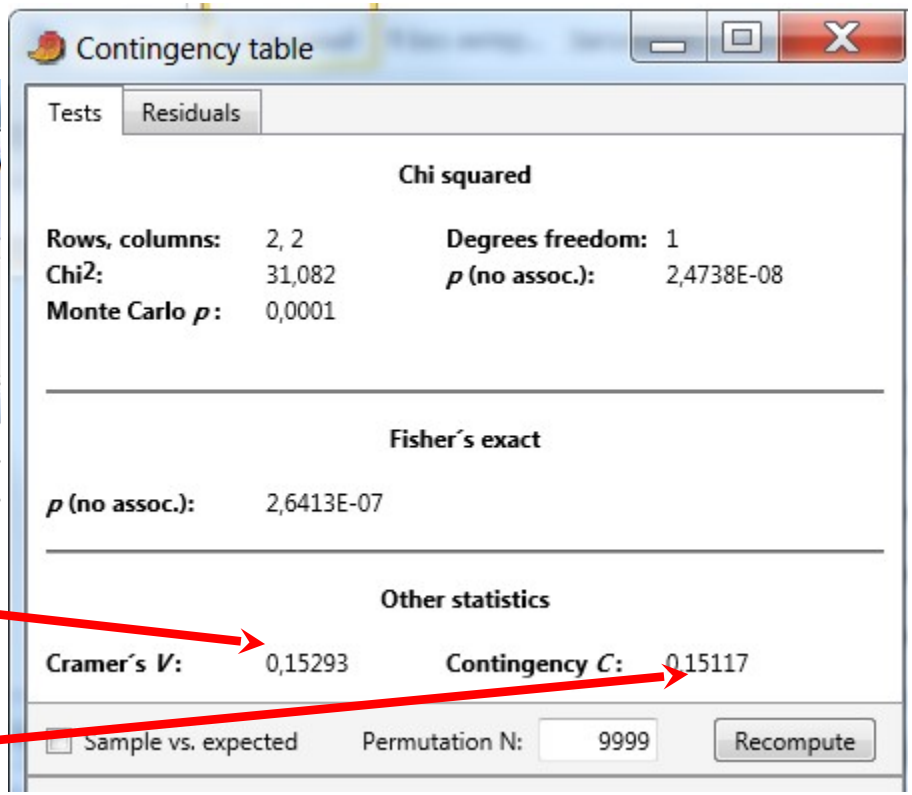
Сформулируем задачу не в терминах поиска различий, а в терминах поиска связи.

Задание: определить, существует ли связь между уровнем холестерина и заболеваниями ССС?

В пакете PAST.



	Заболевание есть	Заболевания нет
Повышенный уровень холестерина	41	245
Нормаольный уровень холестерина	51	992



Коэффициент ассоциации Крамера (*Cramer's V*)
Коэффициент сопряжённости Пирсона (*Contingency C*).

Обе меры близки и с точностью до десятых равны **0,15**. Значения < 0,2 - **«слабая»** по силе ассоциация.

Вывод. Обнаружена слабая, но высоко статистически значимая прямая связь между уровнем холестерина в сыворотке и заболеваниями сердечно-сосудистой системы: коэффициент ассоциации Крамера ***V* = 0,15; *p* <<0,001**.

Спасибо за внимание!