1. Статистический анализ качественных признаков

1.1. Таблицы сопряженности. Критерий хи-квадрат независимости

Предположим, что у нас имеется n наблюдений, характеризующихся значениями двух качественных признаков X и Y с градациями $x_1, \ldots, x_r$ и $y_1, \ldots, y_s$.

ПРИМЕР 1.1. X – это случайный ответ на вопрос „У меня бывают периоды плача и близости к слезам“, допускающий четыре градации x_1 – „совершенно не согласен“, x_2 – „скорее не согласен“, x_3 – „скорее согласен“, x_4 – „совершенно согласен“. Признак Y указывает на пол: y_1 – женский, y_2 – мужской. Из $n = 36$ студентов, участвующих в анкетировании, 14 студентов „совершенно не согласны“, из них 6 девушек и 8 юношей, „скорее не согласны“ пятеро (4 девушки и 1 юноша), 7 человек „скорее согласны“ (1 юноша и 6 девушек) и 10 „совершенно согласны“ (один юноша и 9 девушек).

$X \setminus Y$	y_1	y_2	$\ldots$	y_s	сумма
x_1	n_{11}	n_{12}	$\ldots$	n_{1s}	n_{1*}
x_2	n_{21}	n_{22}	$\ldots$	n_{2s}	n_{2*}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_r	n_{r1}	n_{r2}	$\ldots$	n_{rs}	n_{r*}
сумма	n_{*1}	n_{*2}	$\ldots$	n_{*s}	n

$X \setminus Y$	y_1	y_2	сумма
x_1	6	8	14
x_2	4	1	5
x_3	6	1	7
x_4	9	1	10
сумма	25	11	36

Обозначим через

$$p_{ij} = P(X = x_i, Y = y_j),\ p_{i*} = P(X = x_i),\ p_{*j} = P(Y = y_j).$$

Если X и Y *независимы*, тогда для любых $i = 1, 2, \ldots, r$ и $j = 1, 2, \ldots, s$ справедливо

$$p_{ij} = p_{i*}p_{*j}. \tag{1}$$

При конечном количестве наблюдений оценками вероятностей являются отношения

$$\widehat{p}_{ij} = \frac{n_{ij}}{n}, \quad \widehat{p}_{i*} = \frac{n_{i*}}{n}, \quad \widehat{p}_{*j} = \frac{n_{*j}}{n}, \tag{2}$$

и крайне редко удается получить равенства вида $\frac{n_{ij}}{n} = \frac{n_{i*}}{n} \cdot \frac{n_{*j}}{n}$, например $\widehat{p}_{11} = \frac{6}{36} \neq \frac{14}{36} \cdot \frac{25}{36} = \widehat{p}_{1*}\widehat{p}_{*1}$. Важно определить, не является ли это нарушение случайным.

Для этого используется статистика χ^2 Пирсона (**??**), в выражении которой в качестве вероятностей гипотетического распределения рассматриваются произведения $\widehat{p}_{i*} \cdot \widehat{p}_{*j}$, соответствующие независимости признаков. Вероятности $\widehat{p}_{i*}$ и $\widehat{p}_{*j}$ оцениваются по выборке.

$$\widehat{p}_{i*} = \frac{n_{i*}}{n}, \qquad \widehat{p}_{*j*} = \frac{n_{*j}}{n},$$

статистика χ^2 имеет вид

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - np_{ij})^2}{np_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - n\frac{n_{i*}}{n}\frac{n_{*j}}{n})^2}{n\frac{n_{i*}}{n}\frac{n_{*j}}{n}} =$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{ij}^2 - 2n_{ij}\frac{n_{i*}n_{*j}}{n} + \left(\frac{n_{i*}n_{*j}}{n}\right)^2}{\frac{n_{i*}n_{*j}}{n}} =$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{s} \left(\frac{n \cdot n_{ij}^2}{n_{i*}n_{*j}} - 2n_{ij} + \frac{n_{i*}n_{*j}}{n} \right) = n\left(\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i*}n_{*j}} - 1 \right)$$

и при $n \to \infty$ имеет распределение хи-квадрат с числом степеней свободы, равным

$$rs - 1 - (r-1) - (s-1) = (r-1)(s-1)\,,$$

так как $\sum\limits_{i=1}^{r} p_{i*} = \sum\limits_{j=1}^{s} p_{*j} = 1$, и количество оцениваемых параметров по выборке равно $r + s - 2$.

Получаем значение статистики $\chi^2 = 7.79$ с числом степеней свободы, равным $(4-1)(2-1) = 3$. В электронных таблицах *Excel* для получения P-значения можно воспользоваться встроенной функцией $P = $ ХИ2РАСП$(7.79; 3) = 0.05$. [1] Следовательно, независимость признаков отвергается, то есть юноши и девушки по-разному отвечают на поставленный вопрос, и это отличие значимо.

[1] В *R* используется $1 - pchisq(7.79, 3) = 0.05$ или общий критерий $chisq.test(table(x, y))$.

При использовании этого критерия следует учитывать асимптотический характер распределения статистики и при небольших объемах выборки предпочтительнее обращаться к другим критериям, которые будем рассматривать далее.

1.2. Коэффициент неопределенности

Измерение информации

Всем хорошо известно, что в любом живом существе происходит передача информации. Органы чувств собирают информацию о внешнем мире, нервная система передает ее в мозг, который ее перерабатывает и рассылает приказы на основе поступившей информации по нервным волокнам в органы и мышцы. Аналогично передается информация в любом действующем учреждении: передача информации осуществляется в виде докладных записок, распоряжений, запросов и т.д. Вычислительная техника по заданной программе, т.е. на основе заданной информации, с огромной скоростью перерабатывает большие количества информации. Перечень примеров можно было бы продолжить, но ясно одно, что уровень достижения в большей степени определяется оперативностью работы с информацией, т.е. тем, с какой скоростью и с какой надежностью осуществляется ее передача, хранение (защита) и переработка. Вопросы передачи и хранения информации тесно связаны со способом ее измерения и кодирования. Вопросы переработки информации опираются на линейные статистические методы.

Измерять информацию, подобно расстоянию, массе, количеству теплоты и т.д., стали относительно недавно. Поясним способ измерения информации на основе игры "Бар-Кохба". Один из участников игры должен

отгадать, что задумал другой, задавая ему вопросы, допускающие только два ответа: "да"или "нет". Обычно задумывают какие-либо предметы или строки стихотворений. Количество информации, необходимое для отгадывания, можно измерить количеством вопросов, которые требуется задать при наиболее рациональной тактике дознания.

Запишем полученные ответы, заменив единицами положительные ответы и нулями отрицательные. Получим последовательность нулей и единиц

$$001100011010\ldots,$$

заменяющую последовательность ответов. Процесс замены ответов нулями и единицами называется *кодированием*, а сама последовательность – *кодовым словом*. Длина кодового слова может служить мерой количества информации.

Ответ на любой вопрос, допускающий два равно вероятных ответа "да" или "нет", содержит одну единицу информации, которую называют бит.

Иными словами, количество информации, которое можно с одинаковой вероятностью закодировать нулем или единицей, будем считать единицей информации. Заметим, что при измерении количества информации мы сознательно оставляем без внимания содержание информации.

Формула Хартли

Количество информации H, необходимое для выявления объекта, о котором только известно, что он принадлежит множеству из N объектов, вычисляется при помощи *формулы Хартли*:

$$H = log_2 N. \tag{3}$$

Например, известно, что у одного из 16 студентов сегодня день рождения. Эту неопределенность можно устранить, задав несколько вопросов тому, кто знает. Сколько вопросов нужно задать, если ответы могут быть только "да" или "нет"? – Ответ: $4 = log_2 16$. Для этого лучше всего разделить группу на две части по восемь человек и задать вопрос: "В первой половине счастливчик?". Восьмерку со счастливчиком следует опять разделить пополам и задать аналогичный вопрос, и т.д. Четвертый вопрос устранит неопределенность полностью.

В случае, когда количество N элементов множества не является степенью двойки, количество информации, вычисляемое по формуле Хартли, не является целым, например, $log_2 12 = 3.58$.

Формула Шеннона

Если в множестве элементы не равновероятны, то для вычисления информации, необходимой для отгадывания элемента этого множества, используется *формула Шеннона.* Обозначим через

$$\xi : \begin{pmatrix} x_1 & \dots & x_n \\ p_1 & \dots & p_n \end{pmatrix}$$

случайный эксперимент, а через $x_1, \dots, x_n$ возможные исходы этого эксперимента, которые осуществляются с вероятностями $p_1, \dots, p_n$, $\sum\limits_{k=1}^{n} p_k = 1$. Количество информации $H(\xi)$, которую нужно получить, чтобы узнать, какой исход эксперимента осуществлен, вычисляется по формуле Шеннона:

$$H(\xi) = \sum_{k=1}^{n} p_k \log_2 \frac{1}{p_k}. \tag{4}$$

Например, из перемешанной колоды 36 карт вынимается одна карта. Значениями случайного эксперимента ξ_1 являются все 36 карт, вероятность

появления каждой равна $1/36$:

$$\xi_1 : \begin{pmatrix} x_1 & \dots & x_{36} \\ \frac{1}{36} & \dots & \frac{1}{36} \end{pmatrix}.$$

В этом случае формула Шеннона совпадает с формулой Хартли

$$H(\xi_1) = \sum_{k=1}^{36} \frac{1}{36} \log_2 36 = \log_2 36 = 5.17.$$

Пусть ξ_2 – некий случайный эксперимент, в результате которого из такой же колоды карт вынимается или нет туз:

$$\xi_2 : \begin{pmatrix} \text{'не туз'} & \text{'туз'} \\ \frac{8}{9} & \frac{1}{9} \end{pmatrix}$$

$$H(\xi_2) = \frac{8}{9} \log_2 \frac{9}{8} + \frac{1}{9} \log_2 \frac{9}{1} = 0.352 + 0.151 = 0.503.$$

Необходимая для отгадывания результата случайного эксперимента ξ информация $H(\xi)$ равна неопределенности этого эксперимента, поэтому часто $H(\xi)$ называют *энтропией* распределения случайного эксперимента ξ.

Пусть ξ_0 – некий случайный эксперимент, в результате которого из такой же колоды карт вынимается красная или черная карта. Имеем два равновероятных исхода, поэтому должны получить один бит информации:

$$\xi_0 : \begin{pmatrix} \text{'красная'} & \text{'черная'} \\ 0.5 & 0.5 \end{pmatrix}$$

$$H(\xi_0) = 0.5 \log_2 2 + 0.5 \log_2 2 = 0.5 + 0.5 = 1.$$

Условная энтропия и количество информации

Пусть имеются случайные эксперименты

$$\xi : \begin{pmatrix} A_1 & \dots & A_N \\ p_1 & \dots & p_N \end{pmatrix} \quad \text{и} \quad \eta : \begin{pmatrix} B_1 & \dots & B_n \\ q_1 & \dots & q_n \end{pmatrix}.$$

Каждое событие B_i изменяет вероятности исходов $A_1, \ldots, A_N$. Обозначим через $p_j(B_i) = P(A_j|B_i)$ соответствующие условные вероятности и определим *условную энтропию*

$$H_{B_i}(\xi) = -\sum_{j=1}^{N} p_j(B_i) \log_2 p_j(B_i)\,. \tag{5}$$

Средняя условная энтропия $H_\eta(\xi)$ случайного эксперимента ξ при известном исходе эксперимента η равна

$$H_\eta(\xi) = H(\xi|\eta) = q_1 H_{B_1}(\xi) + \ldots q_n H_{B_n}(\xi). \tag{6}$$

Разность между исходной энтропией $H(\xi)$ и средней условной энтропией $H_\eta(\xi) = H(\xi|\eta)$ называется *количеством информации* $I(\xi, \eta)$ о случайном эксперименте ξ, полученном посредством η,

$$I(\xi, \eta) = H(\xi) - H(\xi|\eta). \tag{7}$$

Убедимся в том, что $H(\eta|\xi) = H(\xi, \eta) - H(\xi)$. Действительно,

$$\begin{aligned} H(\eta|\xi) &= \sum_{k=1}^{r} p_{k\cdot} \left(-\sum_{j=1}^{s} \frac{p_{kj}}{p_{k\cdot}} (\log_2 p_{kj} - \log_2 p_{k\cdot}) \right) = \\ &= -\sum_{k=1}^{r} \sum_{j=1}^{s} p_{kj} \log_2 p_{kj} + \sum_{k=1}^{r} \sum_{j=1}^{s} p_{kj} \log_2 p_{k\cdot} = \\ &= H(\xi, \eta) + \sum_{k=1}^{r} \log_2 p_{k\cdot} \sum_{j=1}^{s} p_{kj} = H(\xi, \eta) - H(\xi). \end{aligned}$$

Отсюда

$$I(\xi, \eta) = H(\xi) + H(\eta) - H(\xi, \eta). \tag{8}$$

Для независимых случайных экспериментов ξ и η количество информации равно нулю, так как $\frac{p_{kj}}{p_{k\cdot}} = p_{\cdot j}$ для любого x_k, $H(\eta|\xi = x_k) = H(\eta)$ и $I(\xi, \eta) = 0$.

С помощью *односторонних коэффициентов неопределенности*

$$J_{X|Y} = \frac{I(X,Y)}{H(Y)} \cdot 100\% \quad \text{и} \quad J_{Y|X} = \frac{I(X,Y)}{H(X)} \cdot 100\% \tag{9}$$

измеряется доля информации одного признака, которую он разделяет с другим. Если интерес представляет не только то, как Y влияет на X, но и то, как сам X влияет на Y, то используют *двусторонний или симметричный коэффициент неопределенности*, равный

$$J = \frac{H(X)}{H(X)+H(Y)} J_{X|Y} + \frac{H(Y)}{H(X)+H(Y)} J_{Y|X} = \\ = \frac{2I(X,Y)}{H(X)+H(Y)} \cdot 100\% \,. \tag{10}$$

ПРИМЕР. Из биологии известно, что за языковые способности отвечает левое полушарие. Рассмотрим таблицу сопряженности двух признаков: X — наличие у больного глиобластомы в правом или в левом полушарии головного мозга; Y — наблюдаемые нарушения речи.

	нарушения речи		
Полушарие	нет	есть	всего
правое	155	12	167
левое	94	59	153
всего	249	71	320

$H(X) = 0.9986$, $H(Y) = 0.7636$, $H(X,Y) = 1.6531$, $I(X,Y) = 0.11$. Коэффициент $J_{X|Y} = 11\%$ указывает на то, какую часть информации о том, какое полушарие повреждено, мы получаем, если узнаем о том, имеются у больного нарушения речи или нет. При отсутствии речевых нарушений у 62 процентов больных поражено правое полушарие, а при нарушении речи у 83 процентов больных глиобластома в левом полушарии. Коэффициент $J_{Y|X} = 14\%$ указывает на то, какую часть информации о том, имеются или нет нарушения речи, мы получаем, если знаем, какое полушарие повреждено: речевые нарушения встречаются у 8 процентов больных с глиобластомой в правом и у 39 процентов больных с глиобластомой в левом полушарии.

1.3. Точный критерий Фишера

Точный критерий Фишера, в основном, используется для проверки

гипотезы однородности категориальных вероятностей в случае таблиц сопряженности размерности 2×2.

ПРИМЕР 1.2. Рассмотрим результаты некоторого наблюдения, согласно которому в 5 из 21 страны уровень сметности от заболеваний системы кровообращения (СК) очень высокий, причем в четырех из них высокий уровень смертности от заболеваний сосудов головного мозга (СГМ). Во странах с высокой смертностью от заболеваний СГМ высокая смертность от заболеваний СК. Составляем таблицу сопряженности по двум признакам: X — фактор смертности от заболеваний СК: $X = 0$ при низкой и $X = 1$ при высокой; Y — фактор смертности от заболеваний СГМ: $Y = 0$ также при низкой и $Y = 1$ при высокой.

$X \setminus Y$	0	1	
0	16	0	16
1	1	4	5
	17	4	21

$X \setminus Y$	0	1	
0	a	b	$a+b$
1	c	d	$c+d$
	$a+c$	$b+d$	n

Обозначим через

$$p_0 = P\{Y = 1|X = 0\} \quad \text{и} \quad p_1 = P\{Y = 1|X = 1\} \tag{11}$$

вероятности высокой смертности от ЗСГМ. Проверяемая гипотеза имеет вид $H_0 : p_0 = p_1$ означает, что вероятность высокой смерности от ЗСГМ не зависит от того, какая смерность от ЗСК.

Вычислим вероятность P^a_{a+c} того, что из $a + c$ стран с низкой смертностью от ЗСГМ в a странах оказалась высокая смертность от ЗСК.

Напомним, что N-множеством в комбинаторике называется множество, состоящее из N элементов. Его K-подмножество называется сочетанием. Упорядоченное K-подмножество — размещением. Число размещений из N элементов по K обозначают через A^K_N. Оно равно произведению из K убывающих сомножителей

$$A^K_N = N \cdot (N-1) \cdot \ldots \cdot (N-k+1),$$

так как на первое место упорядоченного K-множества можно выбрать любой элемент N-множества, на второе — любой из оставшихся $(N-1)$ элементов, и так далее до выбора на K-е место любого из $(N-K+1)$ элементов. Очевидно, что

$$A^K_N = \frac{N!}{(N-K)!}\,.$$

В частности, $A_N^N = N!$. Число сочетаний обозначается через C_N^K. Оно связано с числом размещений выражением $C_N^K \cdot K! = A_N^K$, поэтому для вычисления числа сочетаний используют формулу

$$C_N^K = \frac{N!}{K!(N-K)!}. \tag{12}$$

Для вычисления вероятности P_{a+c}^a сначала требуется вычислить число исходов, удовлетворяющих тому, что $a+c$ стран из $n = a+b+c+d$ имеют низкую сметрность от ЗСГМ. Это число равно

$$C_n^{a+c} = \frac{n!}{(a+c)!(b+d)!}.$$

Затем вычислим число исходов, удовлетворяющих тому, что в a странах из $a+c$ сметрность от ЗСК высокая, а в c странах низкая. При справедливости нулевой гипотезы $H_0 : p_0 = p_1$ это число исходов равно произведению

$$C_{a+b}^a \cdot C_{c+d}^c = \frac{(a+b)!(c+d)!}{a!b!c!d!},$$
$$P_{a+c}^a = \frac{C_{a+b}^a \cdot C_{c+d}^c}{C_n^{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}.$$

Для использования критерия Фишера можно использовать в R функции $table(x, y)$ и $fisher.test(table(x, y))$.

При заданных маргинальных частотах $a+b$, $c+d$, $a+c$, $b+d$ можно перечислить все возможные таблицы с положительными элементами вида

$T \sim$

$X \setminus Y$	0	1	
0	x	$a+b-x$	$a+b$
1	$a+c-x$	$(c+d)-(a+c-x)$	$c+d$
	$a+c$	$b+d$	$a+b+c+d$

и вычислить вероятность получения этой таблицы по формуле

$$P(T) = \frac{(a+b)!(a+c)!(c+d)!(b+d)!}{x!(a+b-x)!(a+c-x)!(d-a+x)!(a+b+c+d)!}.$$

Для каждой из возможных таблиц сопряженности вычисляется характеристика, отражающая степень отклонения от так называемой идеальной таблицы, соответствующей независимым признакам. Для примера предположим, что имеется таблица T_0, будем рассматривать статистику хи-квадрат

$$\chi^2(T_0) = \frac{(ad-bc)^2 \cdot (a+b+c+d)}{(a+b)(c+d)(a+c)(b+d)}.$$

Значимость критерия Фишера в данном случае вычисляется как вероятность получить двумерное распределение (таблицу сопряженности), при котором зависимость между признаками выражена не меньше, чем в случае распределения, соответствующего T_0.

$$\alpha_* = \sum P(T|\chi^2(T) \geq \chi^2(T_0))\,.$$

На основе принципа маловероятных событий при $\alpha_* < \alpha$, где α — уровень значимости, нулевая гипотеза отвергается.

1.4. Практическое занятие по теме проверка независимости качественных признаков

Примеры таблиц сопряженности для независимых признаков

1. Влияние смертности от несчастных случаев и других внешних воздействий (Y) на естественный прирост населения в 1997 году.

ест.прирост в 1997	$Y < 60$	$Y \geq 60$	сумма
низк.	3	4	7
выс.	10	4	14
сумма	13	8	21

Применяем код в R и интерпретируем результат

$$tab < -matrix(c(3, 4, 10, 4), ncol = 2)$$

$$c(p.chi = chisq.test(tab)\$p.value, p.fish = fisher.test(tab)\$p.value)$$

```
     p.chi      p.fish
0.2037365   0.3457516
```

Вывод: высокая смертность от несчастных случаев (≥ 60) в странах, с низким приростом населения $4/7$, с высоким $4/14$, различие незначимо, $p = 0.345$ ($p = 0.204$).

2. Влияние смертности от злокачественных новообразований (Y) на естественный прирост населения в 1997 году.

ест.прирост в 1997	$Y < 180$	$Y \geq 180$	сумма
низк.	3	4	7
выс.	3	11	14
сумма	6	15	21

Вывод: высокая смертность от злокачественных новообразований ($\geq$ 180) в странах, с низким приростом населения 4/7, с высоким 11/14, различие незначимо, $p = 0.345$ ($p = 0.608$).

3. Влияние смертности от ишемической болезни сердца (Y) на естественный прирост населения в 1997 году.

ест.прирост в 1997	$Y < 150$	$Y \geq 150$	сумма
низк.	1	6	7
выс.	7	7	14
сумма	8	13	21

Вывод: высокая смертность от ИБС (≥ 150) в странах, с низким приростом населения 6/7, с высоким 7/14, различие незначимо, $p = 0.173$ ($p = 0.266$).

4. Влияние смертности от заболеваний сосудов головного мозга (Y) на естественный прирост населения в 1997 году.

ест.прирост в 1997	$Y < 190$	$Y \geq 190$	сумма
низк.	3	4	7
выс.	14	0	14
сумма	17	4	21

Вывод: высокая смертность от заболеваний сосудов головного мозга (≥ 190) в странах, с низким приростом населения $4/7$, с высоким 0, различие значимо, $p = 0.006$ ($p = 0.01$).

Таблицы сопряженности для зависимых выборок

Рассмотрим данные такого рода: X_1 и X_2— наличие в ЭКГ эпизода ускоренного идиовентрикулярного ритма (УИР) до и после операции АКШ. У 57 больных этот эпизод не наблюдался ни до, ни после операции; у восьми больных он был только до операции, у двух больных появился после. Не было ни одного больного, у которого эпизод проявлял устойчивость. Проверяется гипотеза о том, что ситуации улучшения или ухудшения равновероятны. Имеется таблица сопряженности в виде матрицы 2×2

$X_1 \setminus X_2$	$-$	$+$	сумма
$-$	a	b	$a+b$
$+$	c	d	$c+d$
сумма	$a+c$	$b+d$	n

$X_1 \setminus X_2$	$-$	$+$	сумма
$-$	57	2	59
$+$	8	0	8
сумма	65	2	67

Нас интересует, насколько значимо различие между частотами b и c. Точная статистика критерия Мак Немара (McNemar's test) вычисляется как

$$\alpha_* = 2 \sum_{i=1}^{\min(b,c)} C_{b+c}^i \frac{1}{2^{b+c}} \,.$$

При малых значениях $\alpha_* < \alpha = 0.05$ гипотеза о равенстве $b = c$ отвергается и различие между ними нельзя объяснить случайностью. В нашем случае $b = 2$, $c = 8$, $\alpha_* = 2(C_{10}^0 + C_{10}^1 + C_{10}^2)/2^{10} = 0.0429 < 0.05$, поэтому число больных, у которых исчез эпизод УИР после операции, значимо больше тех, у кого этот эпизод в ЭКГ появился.

Помимо этой статистики используется выражение

$$\chi^2 = \frac{(b-(b+c)/2)^2}{(b+c)/2} + \frac{(c-(b+c)/2)^2}{(b+c)/2} = \frac{(b-c)^2}{b+c},$$

в котором добавляется поправка на непрерывность

$$\chi^2_* = \frac{(|b-c|-1)^2}{b+c}.$$

При справедливости нулевой гипотезы статистика χ^2 имеет распределение хи-квадрат с одной степенью свободы. Доверительный уровень вероятности равен $p = P\{\chi^2 > \chi^2_* = 3.6\} = 0.058$, следовательно, различие между c и b нельзя объяснить случайностью с уровнем значимости, большим 0.058. С поправкой на непрерывность $p = P\{\chi^2 > \chi^2_* = 2.5\} = 0.11$.

Для категориальных данных, повторяющихся многократно, используется обобщение критерия Мак Немара в виде критерия Кохрена (Cochren's Q test). Пусть имеется s дихотомических признаков у n индивидов. Для определенности закодируем нулем ответы „нет“ и единицей ответы „да“, количество положительных ответов у i-го индивида обозначим через x_{i*}, в j-й момент через x_{*j}, $N = \sum_{j=1}^{s} x_{*j} = \sum_{i=1}^{n} x_{i*}$.

	X_1	X_2	$\dots$	X_s	сумма
1	x_{11}	x_{12}	$\dots$	x_{1s}	x_{1*}
2	x_{21}	x_{22}	$\dots$	x_{2s}	x_{2*}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
n	x_{n1}	x_{n2}	$\dots$	x_{ns}	x_{n*}
сумма	x_{*1}	x_{*2}	$\dots$	x_{*s}	N

При отсутствии изменений в динамике наблюдений статистика

$$T = s(s-1)\frac{\sum_{j=1}^{s}(x_{*j} - \frac{N}{s})^2}{\sum_{i=1}^{n} x_{i*}(s - x_{i*})} \tag{13}$$

имеет распределение хи-квадрат с $s-1$ степенями свободы. Например, исследуется динамика появления эпизода УИР у $n = 52$ больных в $s = 4$

точках: до операции, через две недели после операции, через полгода и через год. Эпизод УИР появлялся у $N = 13$ больных однажды: $x_{j*} \in \{0, 1\}$, до операции у $x_{*1} = 6$ больных, после у $x_{*2} = 2$, затем $x_{*3} = 3$, $x_{*4} = 2$. Значение статистики (13) равно $\chi_* = 3.3$ со значимостью $p = P\{\chi^2 > \chi^2_*\} = 0.35$. Это свидетельствует о том, что снижение числа эпизодов УИР после операции можно объяснить случайностью.