Непараметрические методы анализа качественных признаков

1 Энтропия дискретного распределения

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.

1.1 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:

$$H = log_2 N. \tag{1}$$

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, $log_2 12 = 3.58$.

1.2 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:

$$H(\xi) = \sum_{k=1}^{n} p_k \log_2 \frac{1}{p_k}. \tag{2}$$

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:

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

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

1.3 Условная энтропия и количество информации

Пусть имеются случайные эксперименты

$\xi\|\eta$	y_1	$\dots$	y_s	
x_1	p_{11}	$\dots$	p_{1s}	$p_{1\cdot}$
$\vdots$	$\vdots$	$\dots$	$\vdots$	
x_r	p_{r1}	$\dots$	p_{rs}	$p_{r\cdot}$
	$p_{\cdot 1}$	$\dots$	$p_{\cdot s}$	1

Каждое событие y_j изменяет вероятности исходов $x_i,\ i = 1, \dots, r$ и наоборот. Определим условную энтропию $H(\eta|\xi)$ как среднюю энтропию условных распределений

$$H(\eta|\xi) = \sum_{k=1}^{r} p_{k\cdot} H(\eta|\xi = x_k) = \sum_{k=1}^{r} p_{k\cdot}\left(-\sum_{j=1}^{s}\frac{p_{kj}}{p_{k\cdot}}\left(\log_2\frac{p_{kj}}{p_{k\cdot}}\right)\right)$$

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,

$$I(\xi, \eta) = H(\xi) - H(\xi|\eta). \tag{3}$$

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

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

Отсюда

$$I(\xi,\eta) = H(\xi) + H(\eta) - H(\xi,\eta). \tag{4}$$

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

$$J_{X|Y} = \frac{I(X,Y)}{H(Y)}\cdot 100\% \quad \text{и} \quad J_{Y|X} = \frac{I(X,Y)}{H(X)}\cdot 100\% \tag{5}$$

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

$$J = \frac{H(X)}{H(X)+H(Y)}J_{X|Y} + \frac{H(Y)}{H(X)+H(Y)}J_{Y|X} =$$
$$= \frac{2I(X,Y)}{H(X)+H(Y)}\cdot 100\%\,. \tag{6}$$

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.

```
>       z<-c(155, 12, 94, 59)
>       tab<-matrix( z,ncol=2)
>       tab
```

```
     [,1] [,2]
[1,]  155   94
[2,]   12   59
```

```
> # Условные вероятности возникновения нарушений речи
>     #  при поражении правого и левого полушария"
> round(apply(cbind(tab,rowSums(tab)),1,function(x)x[2]/x[3]),digits=2)
```

```
[1] 0.38 0.83
```

```
> Entropy<-function(xx)  -sum(xx[xx!=0]*log(xx[xx!=0],2))
```

Энтропия совместного распределения равна

```
> H12<-Entropy(z/sum(z))
> H12
```

```
[1] 1.653096
```

Энтропия распределения случайной величины X равна

```
> H1<-Entropy(rowSums(tab)/sum(z))
> H1
```

```
[1] 0.7635764
```

Энтропия распределения случайной величины Y равна

```
> H2<-Entropy(colSums(tab)/sum(z))
> H2
```

```
[1] 0.9986189
```

Совместная информация имеет вид

```
> I<-H1+H2-H12
> I
```

```
[1] 0.1090988
```

Коэффициенты неопределенности

```
> Jy.x<-I/H1*100
> Jx.y<-I/H2*100
> J<-2*I/(H1+H2)
> c(Jx.y,Jy.x,J)
```

```
[1] 10.9249739 14.2878751  0.1238215
```

2 Точный критерий Фишера

Точный критерий Фишера используется для проверки гипотезы однородности категориальных вероятностей в случае таблиц сопряженности размерности 2×2.

$T_0 \sim$

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

Например, X — фактор смертности от заболеваний системы кровообращения (ЗСК): $X = 0$ при низкой и $X = 1$ при высокой; Y — фактор смертности от заболеваний сосудов головного мозга (ЗСГМ): $Y = 0$ также при низкой и $Y = 1$ при высокой.

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

Обозначим через

$$p_0 = P\{Y = 1|X = 0\} \;\text{ и }\; p_1 = P\{Y = 1|X = 1\} \tag{7}$$

условные 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.

$$P^a_{a+c} = \frac{C^a_{a+b} \cdot C^c_{c+d}}{C^{a+c}_n} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} = P(T_0).$$

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 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 по таблице T_0 вычисляется как вероятность получить двумерное распределение, 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.

Применим эти вычисления для проверки зависимости смертности от ЗСГМ и ЗСК.

```
> z<-c(16,1,0,4)
> tab<-matrix(c(16,1,0,4),ncol=2)
> tab
```

```
     [,1] [,2]
[1,]   16    0
[2,]    1    4
```

```
> fisher.test(tab)
```

```
        Fisher's Exact Test for Count Data

data:  tab
p-value = 0.0008354
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 3.707758      Inf
sample estimates:
odds ratio
       Inf
```

2.1 Критерий для проверки равенства частот

Пусть имеется парная выборка $(x_1, y_1), \ldots, (x_n, y_n)$ наблюдений над бернуллиевскими случайными величинами X и Y. Обозначим через $p = P\{X = 1\}$ вероятность успеха для X. Предположим независимость X и Y. Тогда будут одинаковыми условные вероятности успеха $p_1 = p_2 = p$ из (7). Очевидно,

$$\mathbf{E}\bar{X} = p, \mathbf{D}\bar{X} = p(1-p).$$

Представим наблюдения в виде таблицы сопряженности

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

Предложение 1. Пусть справедлива нулевая гипотеза

$$H_0 : p_1 = p_2 = p. \tag{8}$$

Тогда стандартное нормальное распределение имеет статистика вида

$$Z = \frac{(bc - ad)\sqrt{n}}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \sim \mathcal{N}(0,1). \tag{9}$$

Доказательство. В качестве оценок будем рассматривать

$$\hat{p}_1 = \frac{c}{a+c}, \quad \hat{p}_2 = \frac{d}{b+d}, \quad \hat{p} = \frac{c+d}{n}, \quad \text{где} \quad n = a+b+c+d.$$

Заметим, что $\hat{p}_1$ и $\hat{p}_1$ являются выборочными средними выборки X при условиях $Y = 0$ и $Y = 1$, т.е. $\hat{p}_1 = \bar{x}_1$ и $\hat{p}_2 = \bar{x}_2$. С учетом предположения (8) их математические ожидания равны

$$\mathbf{E}\bar{x}_1 = \mathbf{E}\bar{x}_2 = p, \;\Rightarrow\; \mathbf{E}(\bar{x}_1 - \bar{x}_2) = 0$$

а дисперсии на основе независимости элементов выборки имеют вид:

$$\mathbf{D}\bar{x}_1 = \frac{p(1-p)}{a+c}, \qquad \mathbf{D}\bar{x}_2 = \frac{p(1-p)}{b+d}, \;\Rightarrow$$
$$D = \mathbf{D}(\bar{x}_1 - \bar{x}_2) = p(1-p)\left(\frac{1}{a+c} + \frac{1}{b+d}\right).$$

Согласно ЦПТ,

$$\bar{x}_1 - \bar{x}_2 = \frac{c}{a+c} - \frac{d}{b+d} = \frac{bc-ad}{(a+c)(b+d)} \sim \mathcal{N}(0, \sqrt{D}),$$

следовательно, при справедливости нулевой гипотезы (8) статистика $Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{D}}$ имеет стандартное нормальное распределение. При $\hat{p} = \frac{c+d}{n}$ алгебраически приходим к (9) следующим образом:

$$Z = \frac{\frac{bc-ad}{(a+c)(b+d)}}{\sqrt{\frac{(c+d)}{n}\frac{(a+b)}{n}\left(\frac{1}{a+c} + \frac{1}{b+d}\right)}} = \frac{(bc-ad)\sqrt{n}\sqrt{(a+c)(b+d)}}{(a+c)(b+d)\sqrt{(c+d)(a+b)}}.$$

Значимость критерия вычисляется как вероятность получить значение статистики Z, больше по абсолютной величине наблюдаемого значения статистики Z_*.

Если вычислить коэффициент корреляции между признаками X и Y по формуле

$$R = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_i x_i^2 - n\bar{x}^2}\sqrt{\sum_i y_i^2 - n\bar{y}^2}},$$

то используя то, что

$$\sum_i x_i y_i = d, \quad \sum_i x_i^2 = c+d, \quad \sum_i y_i^2 = b+d,$$
$$\bar{x} = \frac{c+d}{n}, \quad \bar{y} = \frac{b+d}{n},$$

можно получить выражение

$$R = \frac{d - (c+d)(b+d)/n}{\sqrt{(c+d) - (c+d)^2/n}\sqrt{(b+d) - (b+d)^2/n}} =$$
$$= \frac{d(a+b+c+d) - (c+d)(b+d)}{\sqrt{(c+d)(b+d)}\sqrt{n - (c+d)}\sqrt{n - (b+d)}} =$$
$$= \frac{ad - bc}{\sqrt{(c+d)(b+d)(a+b)(a+c)}}.$$

Пример.

```
     [,1] [,2]
[1,]   16    0
[2,]    1    4
```

Значимость равна

```
[1] 6.996611e-05
```

Отсюда следует, что условные вероятности $P(Y = 1|X = 0)$, $P(Y = 1|X = 1)$

```
> c(tab[1,2]/rowSums(tab)[1], tab[2,2]/rowSums(tab)[2])
```

```
[1] 0.0 0.8
```

отличаются значимо, так гипотеза о равенстве условных частот отвергается.

3 Таблицы сопряженности для зависимых выборок

3.1 Критерий McNemar

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

	n	Hx	Hy	Hxy	Jx	Jy	J	p_Fish	p_Freq	p_Chisq
Irr	1	0.61	0.95	1.54	2.63	1.70	2.07	0.63	0.41	0.41
Crav	2	0.68	0.77	1.40	8.71	7.70	8.18	0.21	0.15	0.15
Weak	3	0.64	1.00	1.46	27.44	17.50	21.38	0.04	0.02	0.02
Head	4	0.54	0.93	1.47	0.77	0.45	0.57	1.00	0.67	0.67
Trem	5	0.61	0.95	1.53	4.87	3.16	3.83	0.33	0.23	0.23
Anor	6	0.71	0.98	1.65	6.09	4.40	5.11	0.21	0.17	0.17
Thirst	7	0.64	0.99	1.56	11.08	7.11	8.66	0.15	0.09	0.09
Hyper	8	0.61	0.89	1.36	22.74	15.77	18.63	0.02	0.01	0.01
weak...irr	9	1.00	0.96	1.95	0.83	0.86	0.84	0.73	0.53	0.53
weak...crav	10	0.94	0.94	1.85	2.01	2.01	2.01	0.47	0.35	0.35
weak..poly	11	1.00	0.73	1.73	0.13	0.18	0.15	1.00	0.80	0.80
weak..swea	12	1.00	0.75	1.74	0.58	0.78	0.67	0.69	0.61	0.61

Таблица 1: Ответы к 12 заданиям по проверке независимости качественных признаков.

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

$$\alpha_* = 2 \sum_{i=0}^{\min(b,c)} C_{b+c}^i \frac{1}{2^{b+c}}\,.$$

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

$$p = P\{\chi^2 > \chi^2_* = 2.5\} = 0.11.$$

```
> a<-matrix(c(57,8,2,0),ncol=2)
> X<-rbind(
+   matrix(rep(c(0,0),a[1,1]),ncol=2,byrow=TRUE),
+   matrix(rep(c(1,0),a[2,1]),ncol=2,byrow=TRUE),
+   matrix(rep(c(0,1),a[1,2]),ncol=2,byrow=TRUE),
+   matrix(rep(c(1,1),a[2,2]),ncol=2,byrow=TRUE))
> mcnemar.test(X[,1],X[,2])
```

```
        McNemar's Chi-squared test with continuity
        correction

data:  X[, 1] and X[, 2]
McNemar's chi-squared = 2.5, df = 1, p-value = 0.1138
```

3.2 Критерий Кохрена

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}, $N = \sum\limits_{j=1}^{s} x_{*j} = \sum\limits_{i=1}^{n} x_{i*}$.

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

При отсутствии изменений в динамике наблюдений статистика

$$T = s(s-1)\frac{\sum\limits_{j=1}^{s}(x_{*j} - \frac{N}{s})^2}{\sum\limits_{i=1}^{n} x_{i*}(s - x_{i*})} \tag{10}$$

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 (10) 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.

```
> day1 <- read.csv("~/Documents/share/R/lessions/A_Fitting/days_AAC/day1.csv",
+                  sep=";", dec=",")
> day2 <- read.csv("~/Documents/share/R/lessions/A_Fitting/days_AAC/day2.csv",
+                  sep=";", dec=",")
> day3 <- read.csv("~/Documents/share/R/lessions/A_Fitting/days_AAC/day3.csv",
+                  sep=";", dec=",")
> day10 <- read.csv("~/Documents/share/R/lessions/A_Fitting/days_AAC/day10.csv",
+                  sep=";", dec=",")
> colnames(day1)<-paste()
> XX<-lapply(c(1,2,3,10),
+       function(x){
+         NameFolder<-"~/Documents/share/R/lessions/A_Fitting/days_AAC"
+         file<-paste("day",x,".csv",sep="")
+         X<-read.csv(paste(NameFolder,file, sep="/"),sep=";", dec=",")
+         colnames(X)<-paste(colnames(X),x,sep=".")
+         return(G=data.frame(X))
+         })
> XXX<-data.frame(XX)
> v<-c("depressed.mood",
+      "anxiety",
+      "suspiciousness",
+      "irritability",
+      "craving.to.alcohol",
+      "weakness",
+      "insomia",
+      "headache" ,
+      "tremor",
+      "polyuria",
+      "sweating",
+      "transient.hallusinations",
+      "vomiting" ,
+      "anoreksia",
+      "diarrhea",
+      "thirst" ,
+      "chest.pain",
+      "hyperemia" )
> k<-1
```

```
> data<-XXX[,paste(v[k],c(1,2,3,10),sep=".")]
> data<-apply(data,2,function(x)sapply(x,function(y){y[y>1]<-1;y}))
> head(data)
```

```
     depressed.mood.1 depressed.mood.2 depressed.mood.3
[1,]                1                1                1
[2,]                1                1                1
[3,]                1                1               NA
[4,]                1                1                0
[5,]                1                1                1
[6,]                1                1                0
     depressed.mood.10
[1,]                 1
[2,]                 1
[3,]                 1
[4,]                 0
[5,]                 0
[6,]                 0
```

```
> data<-data.frame(na.omit(data))
> Data<-cbind(stack(data),subject=rep(seq(nrow(data)),ncol(data)))
> library("RVAideMemoire")
> cochran.qtest(values~ind|subject, Data, alpha = 0.05, p.method = "fdr")
```

```
        Cochran's Q test

data:  values by ind, block = subject
Q = 42.8824, df = 3, p-value = 2.607e-09
alternative hypothesis: true difference in probabilities is not equal to 0
sample estimates:
 proba in group depressed.mood.1
                      1.00000000
proba in group depressed.mood.10
                      0.09090909
 proba in group depressed.mood.2
                      0.90909091
 proba in group depressed.mood.3
                      0.63636364

        Pairwise comparisons using Wilcoxon sign test

                  depressed.mood.1 depressed.mood.10
depressed.mood.10        1.144e-05                 -
```

```
depressed.mood.2          5.000e-01          2.289e-05
depressed.mood.3          1.172e-02          9.766e-04
                  depressed.mood.2
depressed.mood.10                 -
depressed.mood.2                  -
depressed.mood.3            0.08438

P value adjustment method: fdr
```

```
> rowSums(data)
```

```
 [1] 4 4 2 3 2 2 2 2 3 2 3 3 3 3 2 3 1 3 2 3 3 3
```

```
> colSums(data)
```

```
 depressed.mood.1  depressed.mood.2  depressed.mood.3
               22                20                14
depressed.mood.10
                2
```

```
> TT<-ncol(data)*(ncol(data)-1)*sum((colSums(data)-sum(sum(data))/ncol(data))^2)
> TT<-TT/sum(rowSums(data)*(ncol(data)-rowSums(data)))
> 1-pchisq(TT,ncol(data)-1)
```

```
[1] 2.606692e-09
```

```
>
```

4 Ранговый критерий Вилкоксона

```
> wilcox.test(data[,1],data[,2],  paired = TRUE)
```

```
        Wilcoxon signed rank test with continuity correction

data:  data[, 1] and data[, 2]
V = 3, p-value = 0.3458
alternative hypothesis: true location shift is not equal to 0
```

```
> x<-data[,1];y<-data[,2]
> x <- c(1.83,  0.50,  1.62,  2.48, 1.68, 1.88, 1.55, 3.06, 1.30)
> y <- c(0.878, 0.647, 0.598, 2.05, 1.06, 1.29, 1.06, 3.14, 1.29)
> wilcox.test(x, y, paired = TRUE, alternative = "greater")
```

```
        Wilcoxon signed rank test

data:  x and y
V = 40, p-value = 0.01953
alternative hypothesis: true location shift is greater than 0
```

```
> wilcox.test(y - x, alternative = "less")    # The same.
```

```
        Wilcoxon signed rank test

data:  y - x
V = 5, p-value = 0.01953
alternative hypothesis: true location is less than 0
```

```
> wilcox.test(y - x, alternative = "less",
+             exact = FALSE, correct = FALSE)
```

```
        Wilcoxon signed rank test

data:  y - x
V = 5, p-value = 0.01908
alternative hypothesis: true location is less than 0
```

```
> df<-data.frame(z=x-y,zz=abs(x-y))
> df
```

```
      z    zz
1  0.952 0.952
2 -0.147 0.147
3  1.022 1.022
4  0.430 0.430
5  0.620 0.620
6  0.590 0.590
7  0.490 0.490
8 -0.080 0.080
9  0.010 0.010
```

```
> df<-cbind(df,r=rank(df$zz))
> df
```

```
      z     zz r
1  0.952 0.952 8
2 -0.147 0.147 3
3  1.022 1.022 9
4  0.430 0.430 4
5  0.620 0.620 7
6  0.590 0.590 6
7  0.490 0.490 5
8 -0.080 0.080 2
9  0.010 0.010 1
```

```
> S_m<-sum(df[df$z<0,"r"])
> N<-nrow(df)
> Z<-(S_m-N*(N+1)/4)/sqrt(N*(N+1)*(2*N+1)/24)
> pnorm(-abs(Z))
```

```
[1] 0.01907586
```

```
> df<-cbind(df,r=rank(df[,2]))
> N<-nrow(df)
> df
```

```
      z     zz r r
1  0.952 0.952 8 8
2 -0.147 0.147 3 3
3  1.022 1.022 9 9
4  0.430 0.430 4 4
5  0.620 0.620 7 7
6  0.590 0.590 6 6
7  0.490 0.490 5 5
8 -0.080 0.080 2 2
9  0.010 0.010 1 1
```

```
> N<-nrow(df)
> p.p<-length(df$z[df$z>0])/nrow(df)
> p.m<-length(df$z[df$z<0])/nrow(df)
> theta<-p.p+p.m-(p.p-p.m)^2
```

```
> S<-min(S.p<-nrow(df[df$z>0,]),S.m<-nrow(df[df$z<=0,]))
> W<-max(S.p<-nrow(df[df$z>0,]),S.m<-nrow(df[df$z<=0,]))
> W<-W-S
> W<-S
> Z<-(W-N*(N+1)/4)/sqrt(N*(N+1)*(2*N+1)/24)
> pnorm(-abs(Z))*2
```

```
[1] 0.01515597
```

```
> W<-S.p-S.m
```

Таблицы сопряженности

```
> colnames(day1)
```

```
 [1] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
[19] NA NA NA NA NA NA NA NA NA NA NA NA
```

```
> table(day1[,c(7,12)])
```

```
   NA.1
NA  0 1 2
  0 0 2 1
  1 3 8 4
  2 4 8 4
```

```
>
```