Метод минимума хи-квадрат

Теорема Пирсона. Случай неизвестных параметров

$$\chi^2 = \sum_{i=1}^{r} \frac{(\nu_i - np_i(\alpha_1, \ldots, \alpha_s))^2}{np_i(\alpha_1, \ldots, \alpha_s)} \tag{1}$$

p_i — функции от параметров, которые оцениваются по выборке, и свойства распределения зависят от метода оценивания параметров.

Нейманом и Пирсоном было изучено предельное распределение величины χ^2 при оценке параметров по методу минимума хи-квадрат. Относительно $\alpha_1, \ldots, \alpha_s$ решается система уравнений для $j = 1, 2, \ldots, s$:

$$-\frac{1}{2}\frac{\partial \chi^2}{\partial \alpha_j} = \sum_{i=1}^{r} \left(\frac{\nu_i - np_i}{p_i} + \frac{(\nu_i - np_i)^2}{2np_i^2} \right) \frac{\partial p_i}{\partial \alpha_j} = 0\,,$$

Видоизмененный метод минимума хи-квадрат

$$-\frac{1}{2}\frac{\partial\chi^2}{\partial\alpha_j} = \sum_{i=1}^{r}\left(\frac{\nu_i - np_i}{p_i} + \frac{(\nu_i - np_i)^2}{2np_i^2}\right)\frac{\partial p_i}{\partial\alpha_j} = 0\,,$$

Если при больших n вторым слагаемым можно пренебречь, считая $2np_i^2$ постоянным, то получим систему

$$\sum_{i=1}^{r}\frac{\nu_i - np_i}{p_i}\frac{\partial p_i}{\partial\alpha_j} = 0\,. \qquad (2)$$

Метод оценки параметров из этой системы называется видоизмененным методом минимума хи-квадрат.

Theorem

Пусть заданы функции $p_i(\alpha_1, \ldots, \alpha_s)$ от $s < r$ переменных $\alpha_1, \ldots, \alpha_s$ из невырожденного интервала A в s-мерном пространстве, $i = 1, 2, \ldots, r$, удовлетворяющие условиям:

1. $\sum_{i=1}^{r} p_i(\alpha_1, \ldots, \alpha_s) = 1$;
2. $p_i(\alpha_1, \ldots, \alpha_s) > c^2$;
3. все p_i имеют непрерывные производные $\frac{\partial p_i}{\partial \alpha_j}$, $\frac{\partial^2 p_i}{\partial \alpha_j \partial \alpha_k}$;
4. матрица первых производных имеет ранг s.

Тогда уравнения (2) имеют одно решение $\alpha = (\alpha_1, \ldots, \alpha_s)$, сходящееся по вероятности к $\alpha_0 = (\alpha_1^0, \ldots, \alpha_s^0)$ при $n \to \infty$, где α_0 внутренняя точка интервала A. Значение статистики (1) при таких α асимптотически распределено по $\chi^2(r - s - 1)$.

Проверка согласия с нормальным распределением

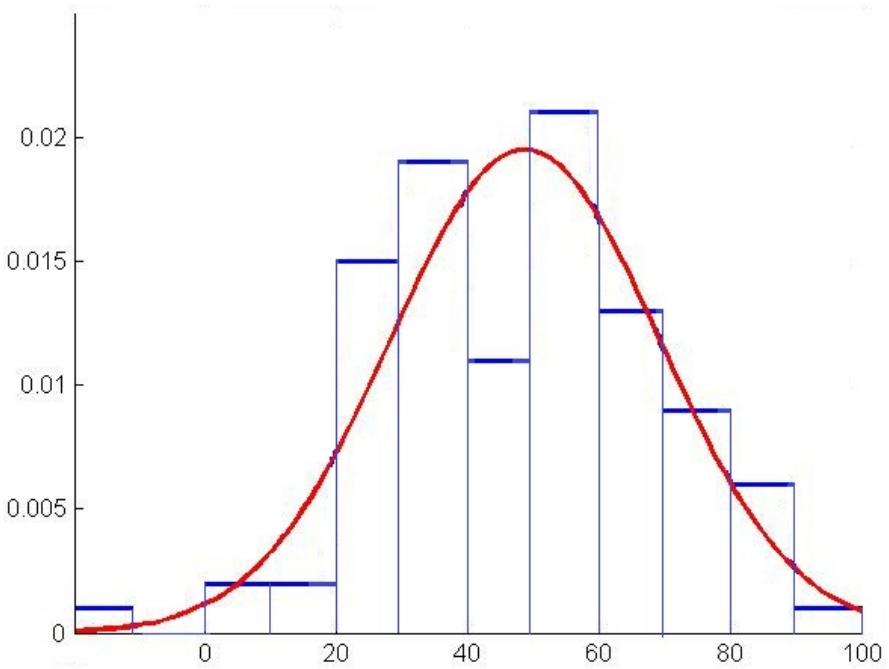

Рис.: Выборочная гистограмма и плотность $\mathcal{N}(\mu = 50, \sigma = 20)$.

Оценки параметров $\hat{\mu} = \bar{\mathrm{x}} = 48.72$, $\hat{\sigma} = \mathrm{S} = 20.47$.

Вычисление эмпирических и теоретических частот

Обозначим через $(z_{i-1}; z_i]$ интервал S_i, $i = 1, \ldots, r$, $z_0 = -\infty$, $z_r = +\infty$, через ν_i количество элементов выборки x_k, таких что $z_{i-1} < x_k \leq z_i$.
Для вычисления вероятностей p_i воспользуемся функцией $\Phi(x)$ стадартного нормального распределения[1].

$$p_i = \Phi\left(\frac{z_i - \hat{\mu}}{\hat{\sigma}}\right) - \Phi\left(\frac{z_{i-1} - \hat{\mu}}{\hat{\sigma}}\right). \tag{3}$$

Объединим наблюдения в крайних ячейках так, чтобы все np_i были не меньше пяти.
Результаты вычислений представлены в следующей таблице.

[1] В электронных таблицах $\Phi(x)$ можно получить при помощи функции НОРМСТРАСП(x), в R функция pnorm

Таблица для вычисления статистики χ^2

i	$(z_{i-1}; z_i]$	ν_i	$\Phi(\frac{z_i-\hat{\mu}}{\hat{\sigma}})$	p_i	$\frac{(\nu_i-np_i)^2}{np_i}$
1	$(-\infty; 20]$	5	0.080	0.080	1.144
2	$(20; 30]$	15	0.180	0.100	2.510
3	$(30; 40]$	19	0.335	0.155	0.799
4	$(40; 50]$	11	0.525	0.190	3.360
5	$(50; 60]$	21	0.709	0.184	0.359
6	$(60; 70]$	13	0.851	0.142	0.094
7	$(70; 80]$	9	0.937	0.086	0.018
8	$(80; +\infty)$	7	1	0.063	0.072
		$n = 100$			

Суммируя элементы в последнем столбце, получаем значение статистики $\chi^2_* = 8.356$. Число степеней свободы равно $df = 8 - 1 - 2 = 5$.

Проверка гипотез

Проверка гипотез основана на принципе маловероятных событий, согласно которому события, вероятность которого мала, считаются невозможными.
Сложилось так, что в качестве „малой“ вероятности принимают вероятность $\alpha = 0.05$ выигрыша в игре в кости.

Вероятность выигрыша в игре в кости

Рис.: Игральные кости astrogalos или "бабки".

Обозначим стороны астрагала A, B, C, D.
$P(A) = 0.39, P(B) = 0.37, P(C) = P(D) = 0.12$. В одной из игр в древней Греции бросали одновременно четыре астрагала. Выигрышным броском считался тот, при котором выпадали разные стороны; такой бросок назывался „Венерой“.
При подбрасывании четырех костей одновременно общее число исходов равно $4^4 = 256$, но вероятности их появления разные.
$P(ABCD) = 0.39 \cdot 0.37 \cdot 0.12 \cdot 0.12 = 0.002$.
$24 = 4 \cdot 3 \cdot 2 \cdot 1$ выигрышных варианта

ABCD	ABDC	ACBD	ACDB	ADBC	ADCB
BACD	BADC	BCAD	BCDA	BDAC	BDCA
CABD	CADB	CBAD	CBDA	CDAB	CDBA
DABC	DACB	DBAC	DBCA	DCAB	DCBA

Суммируя соответствующие вероятности, получаем, что вероятность броска „Венера“ равна $24 \cdot 0.002 \approx 0.05$.

Статистический критерий

- $x_1, \ldots, x_n$ выборочные наблюдения
- Нулевая гипотеза H_0 относительно параметров θ или других свойств генеральной совокупности с функцией распределения $F(x|\theta)$ случайной величины ξ
- H_1 альтернативная гипотеза
- Правило, согласно которому отвергается гипотеза, называется статистическим критерием, а используемая для проверки гипотезы функция от выборочных наблюдений $x_1, \ldots, x_n$ называется статистикой критерия.
- Гипотезу отвергают тогда, когда наблюдаемое значение статистики попадает в критическую область V, которую выбирает так, чтобы вероятность $P(V|H_0)$ была мала, $P(V|H_0) \leq \alpha$.

Пример (продолжение)

Критическое значение, соответствующее уровню значимости $\alpha = 0.05$, вычисляется при помощи специальных таблиц, вероятностных калькуляторов или электронных таблиц Excel и равно $\chi^2_{0.95} = \text{ХИ2ОБР}(0.05; 5) = 11.07$. Наблюдаемое значение статистики $\chi^2_* = 8.356$ меньше критического, следовательно, гипотеза о согласии эмпирического распределения с нормальным не отвергается с уровнем значимости $\alpha = 0.05$. Доверительный уровень вероятности можно вычислить

$$\text{p} = \text{P}\{\chi^2 > \chi^2_*\} = \text{ХИ2РАСП}(8.356; 5) = 0.14 > \alpha = 0.05.$$

При $\text{p} < \alpha$ мы бы сказали, что эмпирическое распределение плохо согласуется с нормальным.

Критерий хи-квадрат независимости

X и Y – качественные признаки 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).

Таблицы сопряженности

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

Представим эти данные в виде таблицы сопряженности, где строки определяются градациями признака X, столбцы – градациями признака Y, а элементы таблицы равны количеству n_{ij} наблюдений с соответствующими градациями: $X = x_i$, $Y = y_j$, $i = 1, 2, \dots, r$, $j = 1, 2, \dots, s$. Через n_{i*} и n_{*j} обозначены соответствующие суммы:

$$n_{i*} = \sum_{j=1}^{s} n_{ij}, \quad n_{*j} = \sum_{i=1}^{r} n_{ij}.$$

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

$$p_{ij} = p_{i*}p_{*j}. \tag{4}$$

При конечном количестве наблюдений оценками вероятностей являются отношения

$$\widehat{p}_{ij} = \frac{n_{ij}}{n}, \quad \widehat{p}_{i*} = \frac{n_{i*}}{n}, \quad \widehat{p}_{*j} = \frac{n_{*j}}{n}, \tag{5}$$

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, насколько случайным является это нарушение.

Статистика Пирсона χ^2 при независимости X и Y

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

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

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

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 $r + s - 2$.

Пример

X \ Y	y_1	y_2	сумма	P
соверш. не согл.	6	8	14	0.57
скорее не согл.	4	1	5	0.20
скорее согл.	6	1	7	0.14
соверш. согл.	9	1	10	0.10
сумма	25	11	36	0.31

$\chi^2_* = 7.79$ с числом степеней свободы, равным $(4-1)(2-1) = 3$.

В электронных таблицах Excel для получения P-значения можно воспользоваться встроенной функцией $\text{P} = \text{ХИ2РАСП}(7.79; 3) = 0.05$. В R функция pchisq.

Следовательно, независимость признаков отвергается, и то, что есть юноши и девушки по-разному отвечают на поставленный вопрос.

Частный случай r = 2, s = 2

При частотах $n_{11} = a$, $n_{12} = b$, $n_{21} = c$, $n_{22} = d$

$$\chi^2 = \frac{n(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}. \qquad (6)$$

$$\frac{a^2}{(a + b)(a + c)} + \frac{b^2}{(a + b)(b + d)} + \frac{c^2}{(a + c)(c + d)} + \frac{d^2}{(b + d)(c + d)} - 1$$

Отдельно рассмотрим числитель этого выражения.

$$a^2(b + d)(c + d) + b^2(a + c)(c + d) + c^2(a + b)(b + d) +$$
$$+d^2(a + b)(a + c) - (bc + cd + bd + d^2)(a^2 + ab + ac + bc) =$$
$$\left\{ \begin{array}{c} a^2(bc + cd + bd + d^2) \\ +b^2(ac + c^2 + ad + cd) \\ +c^2(ab + b^2 + ad + bd) \\ +d^2(a^2 + ab + ac + bc) \end{array} \right\} -$$

$$-(bc+cd+bd+d^2)(a^2+ab+ac+bc)=$$
$$\left\{\begin{array}{c} a^2(bc+cd+bd+d^2) \\ +b^2(ac+c^2+ad+cd) \\ +c^2(ab+b^2+ad+bd) \end{array}\right\}-$$
$$-(bc+cd+bd)(a^2+ab+ac+bc)=$$
$$\left\{\begin{array}{c} a^2d^2+ \\ +b^2(ac+c^2+ad+cd) \\ +c^2(ab+b^2+ad+bd) \end{array}\right\} - \left\{\begin{array}{c} bc(ab+ac+bc) \\ +cd(ab+ac+bc) \\ +bd(ab+ac+bc) \end{array}\right\}=$$
$$=a^2d^2-2abcd+b^2c^2=(ad-bc)^2\,.$$

Измерение информации

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:

$$H = \log_2 N. \tag{7}$$

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: $4 = \log_2 16$.
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:

$$H(\xi) = \sum_{k=1}^{n} p_k \log_2 \frac{1}{p_k}. \tag{8}$$

Существенные частные случаи

Пусть из перемешанной колоды 36 карт вынимается одна карта. Значениями случайного эксперимента ξ_1 являются все 36 карт, вероятность появления каждой равна $1/36$:

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

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_0 эксперимент, в результате которого 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 $A_1, \dots, A_N$. Обозначим через $p_j(B_i) = P(A_j|B_i)$ соответствующие условные вероятности и определим условную энтропию

$$H_{B_i}(\xi) = -\sum_{j=1}^{N} p_j(B_i) \log_2 p_j(B_i). \tag{9}$$

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

$$H_\eta(\xi) = H(\xi|\eta) = q_1 H_{B_1}(\xi) + \dots q_n H_{B_n}(\xi). \tag{10}$$

$\mathrm{H}(\eta|\xi) = \mathrm{H}(\xi,\eta) - \mathrm{H}(\xi)$

$\xi\|\eta$	y_1	$\cdots$	y_s	
x_1	p_{11}	$\cdots$	$\mathrm{p}_{1\mathrm{s}}$	p_{1*}
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_r	$\mathrm{p}_{\mathrm{r}1}$	$\cdots$	p_{rs}	$\mathrm{p}_{\mathrm{r}*}$
	p_{*1}	$\cdots$	$\mathrm{p}_{*\mathrm{s}}$	

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

$$= -\sum_{\mathrm{k}=1}^{\mathrm{r}} \sum_{\mathrm{j}=1}^{\mathrm{s}} \mathrm{p}_{\mathrm{kj}} \log_2 \mathrm{p}_{\mathrm{kj}} + \sum_{\mathrm{k}=1}^{\mathrm{r}} \sum_{\mathrm{j}=1}^{\mathrm{s}} \mathrm{p}_{\mathrm{kj}} \log_2 \mathrm{p}_{\mathrm{k}\cdot} =$$

$$= \mathrm{H}(\xi,\eta) + \sum_{\mathrm{k}=1}^{\mathrm{r}} \log_2 \mathrm{p}_{\mathrm{k}\cdot} \sum_{\mathrm{j}=1}^{\mathrm{s}} \mathrm{p}_{\mathrm{kj}} = \mathrm{H}(\xi,\eta) - \mathrm{H}(\xi).$$

Количество информации

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

$$\mathrm{I}(\xi, \eta) = \mathrm{H}(\xi) - \mathrm{H}(\xi|\eta) = \mathrm{H}(\eta) - \mathrm{H}(\eta|\xi). \tag{11}$$

Из $\mathrm{H}(\xi|\eta) = \mathrm{H}(\xi, \eta) - \mathrm{H}(\eta)$ следует

$$\mathrm{I}(\xi, \eta) = \mathrm{H}(\xi) + \mathrm{H}(\eta) - \mathrm{H}(\xi, \eta). \tag{12}$$

Совместная информация при независимых величинах

Если ξ и η независимы, то $\mathrm{I}(\xi, \eta) = 0$.

Покажем, что при $\frac{\mathrm{p_{kj}}}{\mathrm{p_{k\cdot}}} = \mathrm{p_{\cdot j}}$ для любого $\mathrm{x_k}$ имеет место $\mathrm{H}(\eta|\xi = \mathrm{x_k}) = \mathrm{H}(\eta)$. Действительно,

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

$$\mathrm{H}(\eta|\xi) = \sum_{\mathrm{k}=1}^{\mathrm{r}} \mathrm{p_{k\cdot}} \mathrm{H}(\eta|\xi = \mathrm{x_k}) = \mathrm{H}(\eta), \iff$$

$$\iff \mathrm{I}(\xi, \eta) = \mathrm{H}(\eta) - \mathrm{H}(\eta|\xi) = 0\,.$$

С помощью односторонних коэффициентов неопределенности

$$J_{X|Y} = \frac{I(X,Y)}{H(Y)} \cdot 100\% \quad \text{и} \quad J_{Y|X} = \frac{I(X,Y)}{H(X)} \cdot 100\% \tag{13}$$

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

$$J = \frac{H(X)}{H(X)+H(Y)} J_{X|Y} + \frac{H(Y)}{H(X)+H(Y)} J_{Y|X} = \\ = \frac{2I(X,Y)}{H(X)+H(Y)} \cdot 100\% \,. \tag{14}$$

ПРИМЕР

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	P
правое	155	12	167	0.07
левое	94	59	153	0.39
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.

Основы комбинаторики

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

$$A_N^K = N \cdot (N-1) \cdot \ldots \cdot (N-k+1) = \frac{N!}{(N-K)!}$$

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

$$C_N^K = \frac{N!}{K!(N-K)!}. \tag{15}$$

Точный критерий Фишера

Урновая схема

Чему равна вероятность того, что среди n шаров, вынутых из урны, в которой среди N шаров M красных, окажется m красных?

$$P_n^m = \frac{C_M^m C_{N-M}^{n-m}}{C_N^n}$$

X и Y — смертность от заболеваний СК и СГМ: 0 низкая, 1 высокая.

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

Точный критерий Фишера

$X \setminus Y$	0	1	
0	$a = m$	$b = M - m$	$a + b = M$
1	$c = n - m$	$d = N - M - (n - m)$	$c + d = N - M$
	$a + c = n$	$b + d$	$n_0 = N$

$$P_n^m = \frac{C_M^m C_{N-M}^{n-m}}{C_N^n} =$$

$$= \frac{M!}{m!(M-m)!} \cdot \frac{(N-M)!}{(n-m)!!(N-M-(n-m))!} \cdot \frac{n!(N-n)!}{N!} =$$

$$= \frac{M!(N-M)!n!(N-n)!}{m!(M-m)!(n-m)!!(N-M-(n-m))!N!},$$

$$P_{a+c}^a = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n_0!}.$$

Обозначим через

$$p_0 = P\{Y = 1|X = 0\} \quad \text{и} \quad p_1 = P\{Y = 1|X = 1\} \tag{16}$$

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.

Вероятность получения таблиц

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

$T_x \sim$

$X \setminus Y$	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

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

Значимость точного критерия Фишера

Пусть имеется таблица $\mathrm{T_a}$ с элементами $\mathrm{a, b, c, d}$,

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

Рассматриваются таблицы $\mathrm{T_x}, 0 \leq \mathrm{x} \leq \min(\mathrm{a} + \mathrm{c}, \mathrm{a} + \mathrm{b})$, вычисляется $\chi^2(\mathrm{T_x})$. Значимость критерия Фишера вычисляется как вероятность получить таблицу сопряженности $\mathrm{T_x}$, при которой зависимость между признаками выражена не меньше, чем в случае наблюдаемого распределения, соответствующего $\mathrm{T_a}$.

$$\alpha_* = \sum \mathrm{P}(\mathrm{T_x} | \chi^2(\mathrm{T_x}) \geq \chi^2(\mathrm{T_a}))\,.$$

При $\alpha_* < \alpha$, где α — уровень значимости, нулевая гипотеза о независимости признаков отвергается. Для использования критерия Фишера можно использовать в R функции table(x, y) и fisher.test(table(x, y)).

Значимость точного критерия Фишера, пример

Влияние смертности от несчастных случаев и других внешних воздействий (Y) на естественный прирост населения в 1997 году. Значение статистики $\chi^2 = 3.5$, значимость критерия p = 0.1587.

ест.прирост в 1997	Y < 60	Y ≥ 60	сумма
низк.	2	5	7
выс.	10	4	14
сумма	12	9	21

χ^2	14.000	7.875	3.500	0.875
P(T)	0.0003	0.0087	0.0715	0.2384
T	0 | 7 12 | 2	1 | 6 11 | 3	2 | 5 10 | 4	3 | 4 9 | 5

χ^2	0.000	0.875	3.500	7.875
P(T)	0.3576	0.2452	0.0715	0.0068
T	4 | 3 8 | 6	5 | 2 7 | 7	6 | 1 6 | 8	7 | 0 7 | 9

Точный критерий Фишера для таблиц более чем 2 на 2

- Вместо статистики хи-квадрат Пирсона используется отношение правдоподобия.
- При увеличении числа строк или столбцов приходит сообщение о невозможности точного вычисления, тогда используется метод Монте-Карло.
- В R опция simulate.p.value = TRUE.

Точный критерий Мак-Немара (McNemar's test) для зависимых выборок

X_1 и X_2— наличие в ЭКГ эпизода ускоренного идиовентрикулярного ритма (УИР) до и после операции.

$X_1 \setminus X_2$	$-$	$+$	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

$H_0 : b = c$, то есть ситуации улучшения или ухудшения равновероятны. Точная статистика критерия Мак Немара

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

При $\alpha_* < \alpha = 0.05$ гипотеза H_0 отвергается, различие между ними нельзя объяснить случайностью. $b = 2$, $c = 8$, $\alpha_* = 2(C_{10}^0 + C_{10}^1 + C_{10}^2)/2^{10} = 0.0429 < 0.05$, поэтому число улучшения значимо больше числа ухудшений.

Асимптотический критерий Мак-Немара

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

добавляется поправка на непрерывность

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

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

Критерий Кохрена для s > 2 зависимых выборок

Пусть x_{ij} ответ i -го индивида в j-й момент, $x_{ij} \in \{0, 1\}$,
x_{i*} — количество единиц у i-го индивида,
x_{*j} — в j-й момент, $N = \sum_{j=1}^{s} x_{*j} = \sum_{i=1}^{n} x_{i*}$.

	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

$$T = s(s-1)\frac{\sum_{j=1}^{s}\left(x_{*j} - \frac{N}{s}\right)^2}{\sum_{i=1}^{n} x_{i*}(s - x_{i*})} \sim \chi^2(s-1) \qquad (17)$$

Пример применения критерия Кохрена

Например, динамика 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: до операции у $x_{*1} = 6$ больных, после операции у $x_{*2} = 2$, затем $x_{*3} = 3$, $x_{*4} = 2$. Значение статистики (17) 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.