Лекция 6. Критерии согласия. Проверка независимости двух номинальных признаков

Буре В.М., Грауэр Л.В.

ШАД

Санкт-Петербург, 2013

Содержание

Критерий согласия Пирсона

Определение 1

Статистической гипотезой называется любое предположение о законе распределения генеральной совокупности.

Определение 2

Гипотеза называется простой, если в ней единственным образом определяется закон распределения генеральной совокупности. В противном случае гипотеза называется сложной.

Один из типов гипотез — *гипотезы согласия*. Методы проверки этих гипотез — *критерии согласия*.

Пусть задана генеральная совокупность ξ, функция распределения F_ξ, которой взаимно однозначно соответствует распределению генеральной совокупности P_ξ, и выборка $X_{[n]} = (X_1, \ldots, X_n)$.
Пусть проверяется гипотеза согласия $H_0 : F_\xi = F_0$ (нулевая гипотеза), при этом предполагается, что $F_0(x)$ **известна**. Очевидно, что данная гипотеза — простая.
Сформулируем альтернативную гипотезу: $H_1 : F_\xi \neq F_0$.

Для нас важна гипотеза H_0, необходимо решить: принять ее или отклонить. Решение принимается по имеющейся выборке $X_{[n]}$, т. е. проверяется «хорошо» ли выборка $X_{[n]}$ согласуется с F_0. Мы можем принять гипотезу, но при этом она на самом деле может быть неверной.

Рассмотрим критерий Пирсона (критерий χ^2).

Числовую ось разбиваем на r промежутков $-\infty = a_0 < a_1 < \ldots < a_r = \infty$, $\Delta_i = (a_{i-1}, a_i]$, $r = 1, \infty, r$, и построим статистику χ^2:

$$\chi^2(X_{[n]}) = \sum_{i=1}^{r} \frac{(n_i - np_i^{(0)})^2}{np_i^{(0)}}.$$

где $p_i^{(0)} = F_0(a_i) - F_0(a_{i-1})$.
Если H_0 верна, тогда по теореме 1 Л4:

$$\chi^2(X_{[n]}) \xrightarrow[n\to\infty]{\mathrm{d}} \zeta,$$

где ζ подчиняется распределению хи-квадрат с $r-1$ степенями свободы.

В лекции 4 было показано, что

$$\chi^2(X_{[n]}) = nS(X_{[n]}) = nh\left(\frac{1}{n}g(X_i)\right),$$

где $h(t_1, \dots, t_r) = \sum\limits_{i=1}^{r}\left(t_i - \sqrt{p_i^{(0)}}\right)^2$, $g(x) = (g_1(x), \dots, g_r(x))$ и $g_i(x) = \frac{1}{\sqrt{p_i^{(0)}}} I\left\{x \in \Delta_i\right\}$, $i = 1, \dots, r$.

Статистика $S(X_{[n]}) = h\left(\frac{1}{n}\sum\limits_{j=1}^{n} g(X_j)\right)$ является статистикой первого типа.

При гипотезе H_0 имеем $a = \left(\sqrt{p_1^{(0)}}, \dots, \sqrt{p_r^{(0)}}\right)$ и $h(a) = 0$.

Если H_0 верна, тогда $h\left(\frac{1}{n}\sum\limits_{j=1}^{n} g(X_j)\right) \xrightarrow[n\to\infty]{\text{п.н.}} h(a) = 0.$

Если верна гипотеза H_1, то $a \neq \left(\sqrt{p_1^{(0)}}, \ldots, \sqrt{p_r^{(0)}}\right)$, тогда $h(a) \neq 0$,

но статистика $h\left(\frac{1}{n}\sum_{j=1}^{n} g(X_j)\right) \xrightarrow[n\longrightarrow\infty]{\text{п.н.}} h(a) \neq 0$, следовательно,

$$\chi^2 = nh\left(\frac{1}{n}\sum_{j=1}^{n} g(X_j)\right) \xrightarrow[n\longrightarrow\infty]{\text{п.н.}} \infty.$$

Поведение статистики χ^2 зависит от того, верна нулевая гипотеза или нет.
Нетрудно понять, как выбрать критическую область для гипотезы H_0.

Если $\chi^2(X_{[n]}) \geqslant C$, то H_0 отклоняется, если $\chi^2(X_{[n]}) < C$, то нет оснований для отклонения H_0.

Выберем вероятность $\alpha \in (0, 1)$. В качестве α часто выбирают 0,05.
Константу $C(r-1, \alpha)$ выберем из условия

$$P\{\eta \geqslant C\} = \alpha,$$

где случайная величина η подчиняется распределению хи-квадрат с $r-1$ степенью свободы.
Константа $C(r-1, \alpha)$ представляет собой квантиль уровня $1-\alpha$ распределения хи-квадрат с $r-1$ степенью свободы.
Для практического нахождения квантили можно использовать статистические таблицы.

Область $(C(r-1, \alpha), \infty)$ является критической для гипотезы H_0.
Если $\chi^2(X_{[n]}) > C(r-1, \alpha)$, то H_0 отклоняется, а
если $\chi^2(X_{[n]}) \leq C(r-1, \alpha)$, то для отклонения нет оснований.

В результате применения критерия могут возникать следующие ошибки:

- Ошибка первого рода, если мы отбросили гипотезу H_0, а она на самом деле верна.
- Ошибка второго рода, если мы принимаем гипотезу H_0, а она на самом деле не верна.

Учитывая доказанную асимптотику, вероятность ошибки первого рода приближенно совпадает с заданной вероятностью α. Вероятность ошибки первого рода называют часто уровнем значимости критерия.

Возможен альтернативный подход.
Найдем вероятность

$$P\{\eta > \chi^2(X_{[n]})\} = 1 - F_{\chi^2_{r-1}}(\chi^2(X_{[n]}),$$

где $F_{\chi^2_{r-1}}(\cdot)$ — функция распределения хи-квадрат с $r-1$ степенью свободы.
Вероятность вычисляется при условии справедливости H_0.
$P\{\eta > \chi^2(X_{[n]})\}$ вероятность называется p-значением (p-value).

Большие значения статистики χ^2 свидетельствуют против H_0.
Интервал $(\chi^2(X_{[n]}); +\infty)$ — критическая область значений статистики χ^2, там находятся еще более «худшие» значения статистики для гипотезы H_0.

$$P_{H_0}(\chi^2(X_{[n]}); +\infty) = 1 - F_{\chi^2_{r-1}}(\chi^2(X_{[n]})) = p.$$

Далее можно выбрать значение α (например, 0,05) и применить следующий критерий.
Если $p \leqslant \alpha$, то мы отвергаем H_0 и принимаем альтернативу H_1.
Если $p > \alpha$, то гипотеза H_0 согласуется с наблюдениями.

Указанный критерий полностью эквивалентен критерию в первом подходе.

Можно отметить некоторые общие особенности статистических критериев, которые обсуждались выше. Любой статистический критерий устроен следующим образом:

1. Выбираем статистику критерия, закон распределения которой известен, если справедлива H_0.
2. Определяем критическую область для гипотезы H_0, попадание в которую маловероятно, если гипотеза H_0 верна.
3. Уровень значимости или вероятность ошибки первого рода — вероятность попадания в критическую область при условии истинности H_0, обычно она выбирается малой, например, 0,05. Существует правило p-значения: находится вероятность получения значений статистики, которые еще «хуже», чем полученное значение при условии истинности H_0. Если p-значение оказывается меньше заданного уровня значимости α, то нулевую гипотезу отвергают.

Критерий согласия Колмогорова

Рассмотрим другой критерий согласия — критерий Колмогорова.
Пусть задана генеральная совокупность ξ, функция распределения F_ξ, которой взаимно однозначно соответствует распределение P_ξ, и выборка $X_{[n]} = (X_1, \ldots, X_n)$.
Выдвинем нулевую гипотезу $H_0 : F_\xi = F_0$, $H_1 : F_\xi \neq F_0$.
Дополнительно наложим ограничение: функция $F(x)$ непрерывна на $\mathbb{R}$.

Рассмотрим статистику Колмогорова:

$$D_n(X_{[n]}) = \sup_{x \in R} \left| F_n^*(x) - F_0(x) \right|. \qquad (1)$$

Если верна гипотеза H_0, то $D_n(X_{[n]}) \xrightarrow[n\to\infty]{\text{п.н.}} 0$.
Если верна гипотеза H_1, т. е. $F_\xi \equiv G \neq F_0$, тогда

$$D_n(X_{[n]}) \xrightarrow[n\to\infty]{\text{п.н.}} \sup_{x \in R} \left| G(x) - F_0(x) \right| > 0.$$

Можно показать, что при условии справедливости гипотезы H_0 распределение статистики $D_n(X_{[n]})$ не зависит от конкретного вида F_0.

Лемма 1

Если гипотеза H_0 верна, и $F_0(x)$ — непрерывная функция на $\mathbb{R}$, тогда распределение статистики

$$D_n = \sup_{x \in R} \left| F_n^*(x; x_{[n]}) - F_0(x) \right|$$

не зависит от закона распределения генеральной совокупности.

Доказательство

Предположим дополнительно, что существует (a, b): $a \geq -\infty$, $b \leq +\infty$, что $F_0(x)$ строго монотонна на (a, b), при этом $F_0(a) = 0$, $F_0(b) = 1$.

Подставим в качестве аргумента в $F_0(x)$ случайную величину ξ.
Заметим, что $\xi \overset{\text{п.н.}}{\in} (a, b)$ и $X_i \overset{\text{п.н.}}{\in} (a, b)$ для всех i.
Тогда из строгой монотонности F_0 следуют равенства:

$$P\{F_0(\xi) \le y\} = P\{\xi \le F_0^{-1}(y)\} = F_0(F_0^{-1}(y)) = y.$$

Таким образом, получили соотношение: $P\{F_0(\xi) \le y\} = y$, если $y \in [0, 1]$, то есть, $F_0(\xi)$ подчиняется равномерному распределению.
Обозначим $F_0(X_i)$ через Y_i, $i = 1, \dots, n$, тогда выборка $Y_{[n]} = (Y_1, \dots, Y_n)$ — выборка из равномерного распределения на отрезке $[0, 1]$. Очевидно, что

$$F_n^*(x, X_{[n]}) = \frac{1}{n}\sum_{j=1}^{n} I\{X_j \le x\} = \frac{1}{n}\sum_{j=1}^{n} I\{F_0(X_j) \le F_0(x)\} =$$

$$= \frac{1}{n}\sum_{j=1}^{n} I\{Y_j \le y\} = F_n^*(y, Y_{[n]}),$$

где $F_n^*(y, Y_{[n]})$ — эмпирическая функция распределения, построенная по выборке из равномерного распределения, $y = F_0(x)$.

Тогда имеет место равенство:

$$\sup_{x\in\mathbb{R}} \left| F_n^*(x, X_{[n]}) - F_0(x) \right| = \sup_{y\in[0,1]} \left| F_n^*(y, Y_{[n]}) - y \right| .$$

Отсюда следует равенство:

$$P\{\sup_{x\in\mathbb{R}} \left| F_n^*(x, X_{[n]}) - F_0(x) \right| \leq z\} = P\{\sup_{y\in[0,1]} \left| F_n^*(y, Y_{[n]}) - y \right| \leq z\}.$$

Последнюю вероятность можно вычислять для различных значений z как многомерный интеграл от плотности, тождественно равной единице внутри гиперкуба $[0,1]^n$. Интегрирование проводится по множеству точек в $[0,1]^n$, удовлетворяющих неравенству, содержащемуся под знаком вероятности. Таким образом, для рассматриваемого случая можно построить точные критические области с заданным уровнем значимости независимо от конкретного вида $F_0(x)$.

Можно найти $z_{1-\alpha}$ для некоторого α. Рассмотрим критическую область $(z_{1-\alpha}, 1]$.
Если статистика (1) попадает в данную область, тогда отвергаем H_0 и принимаем H_1.
Если $D_n \in [0, z_{1-\alpha})$, тогда принимаем гипотезу H_0.
Лемма доказана.

Замечание 1

Доказательство леммы 1 можно распространить на случай, когда от $F_0(x)$ не требуется строгой монотонности, для доказательства необходимо рассмотреть обобщенную обратную функцию:

$$F_0^{-1}(y) = \sup\{x : F_0(x) \leq y\}.$$

Таким образом, возможно построение точного критерия согласия для фиксированного n. Однако, при больших n возникают серьезные вычислительные трудности. В связи с этим применяется важный асимптотический результат. Справедлива следующая теорема А.Н. Колмогорова.

Теорема 1 (А.Н. Колмогорова)

Если гипотеза H_0 верна, и $F_0(x)$ — непрерывная функция на $\mathbb{R}$, тогда имеет место сходимость:

$$P\{\sqrt{n}D_n(X_{[n]}) \leqslant z\} \xrightarrow[n\to\infty]{} K(z) = 1 + 2\sum_{m=1}^{\infty}(-1)^m e^{-2m^2z^2}.$$

Находим константу $d_{1-\alpha}$ как решение следующего уравнения:

$$K(d_{1-\alpha}) = 1 - \alpha.$$

Правило проверки гипотез будет следующим.

Если $\sqrt{n}D_n(X_{[n]}) \in (d_{1-\alpha}, \infty)$, тогда гипотеза H_0 отвергается,
если $\sqrt{n}D_n(X_{[n]}) \notin (d_{1-\alpha}, \infty)$, тогда гипотеза H_0 принимается.

Для практической проверки гипотез согласия по критерию Колмогорова можно воспользоваться таблицами математической статистики.
Статистику $D_n(X_{[n]})$ можно вычислить с помощью простого вычислительного алгоритма:

$$D_n(X_{[n]}) = \max_{1 \leqslant i \leqslant n} \left[\frac{i}{n} - F_0(X_{(i)}), F_0(X_{(i)}) - \frac{i-1}{n} \right],$$

где $X_{(1)} < \ldots < X_{(n)}$ — вариационный ряд, построенный по выборке $X_{[n]}$.

Критерий $\overline{\omega}^2$

Замечание 2 (Критерий $\overline{\omega}^2$)

Пусть задана генеральная совокупность ξ с функцией распределения F_ξ и выборка $X_{[n]} = (X_1, \dots, X_n)$ из этой генеральной совокупности. Выдвинем нулевую гипотезу $H_0 : F_\xi = F_0$, при конкурирующей гипотезе $H_1 : F_\xi \neq F_0$.
Статистика критерия имеет вид:

$$\overline{\omega}_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left\{ F_0(X_{(i)}) - \frac{2i-1}{2n} \right\}^2,$$

где $X_{(1)} < \dots < X_{(n)}$ — вариационный ряд, построенный по выборке $X_{[n]}$.
При справедливости гипотезы H_0 и непрерывности функции F_0 распределение статистики омега-квадрат зависит только от n и не зависит от F_0.

При малых n имеются таблицы критических точек, а для больших значений n следует использовать предельное (при $n \to \infty$) распределение статистики $\overline{\omega}_n^2$. Для него составлены подробные таблицы и вычислительные программы.

Важное с теоретической точки зрения свойство критериев, основанных на D_n и $\overline{\omega}_n^2$: они состоятельны против любой альтернативной гипотезы $F_\xi \neq F_0$.
Статистический критерий для проверки гипотезы H_0 называют состоятельным против альтернативной гипотезы H_1, если вероятность отвергнуть H_0, когда на самом деле верна H_1, стремится к 1 при неограниченном увеличении объема наблюдений.

Критерий согласия хи-квадрат для сложных гипотез

В критерии согласия хи-квадрат реализуется следующая схема. Выдвигаются гипотезы:

- H_0: $F_\xi(x) \equiv F(x)$ — нулевая гипотеза.
- H_1: $F_\xi(x) \neq F(x)$ — альтернативная гипотеза.

В прикладных задачах, как правило, известна не сама функция распределения, а параметрическое семейство, которому она принадлежит:

$$\left\{F(\cdot/\theta) : \theta \in \Theta \subset \mathbb{R}^l\right\}.$$

Таким образом, проверяемая гипотеза принимает вид:

$$H_0 : F_\xi \in \left\{F(\cdot/\theta) : \theta \in \Theta \subset \mathbb{R}^l\right\}.$$

Альтернативная гипотеза H_1 примет вид: гипотеза H_0 не верна.

Разобьем числовую ось на k промежутков: $\Delta_1, \ldots, \Delta_k$ таким образом, что $\bigcup_i \Delta_i = \mathbb{R}$, $\Delta_i \cap \Delta_j = \emptyset$, $i \neq j$.
Получаем набор частот: $n_1, \ldots, n_k$, $\sum_{i=1}^{k} n_i = n$.
Каждому промежутку $\Delta_1, \ldots, \Delta_k$ сопоставим вероятности: $p_1(\theta), \ldots, p_k(\theta)$. В монографии [4] приведена теорема Фишера.

Теорема 2 (Теорема Фишера)

Пусть Θ – открытое множество в $\mathbb{R}^l$. Пусть выполнены условия:

1. *Для любого $\theta \in \Theta$: $\sum_{i=1}^{k} p_i(\theta) = 1$.*
2. *Для любого $\theta \in \Theta$: $p_i(\theta) > c > 0$ для любого $i = \overline{1, k}$.*
3. *Для любого $\theta \in \Theta$ существуют и непрерывны производные: $\partial p_i(\theta)/\partial\theta_j$, $\partial^2 p_i(\theta)/(\partial\theta_u \partial\theta_v)$ для любого $i = 1, \ldots, k$, $u, v, j = 1, \ldots, l$.*
4. *Для любого $\theta \in \Theta$ матрица $\left(\dfrac{\partial p_i(\theta)}{\partial\theta_j}\right)_{i,j=\overline{1,k}}$ имеет ранг l.*

Пусть $\hat{\theta}$ *— оценка, найденная методом максимального правдоподобия по выборке* $n_1, \ldots, n_k$, *т. е.* $\hat{\theta} = \arg\max_{\theta \in \Theta} L(\{n_i\}, \theta)$, *где*

$$L(\{n_i\}, \theta) = \frac{n!}{n_1! \cdot \ldots \cdot n_k!} \prod_{i=1}^{k} p_i^{n_i}(\theta),$$

или $\hat{\theta}$ *— оценка по методу минимума хи-квадрат:*

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{k} \frac{(n_i - np_i(\theta))^2}{np_i(\theta)}.$$

Тогда, если гипотеза H_0 *верна, то*

$$\chi^2(\hat{\theta}) = \sum_{i=1}^{k} \frac{(n_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})} \xrightarrow[n \to \infty]{d} \zeta \sim \chi^2_{k-l-1}.$$

Критическая область для гипотезы H_0 при использовании статистики $\chi^2(\hat{\theta})$ имеет вид: $S = (u_{1-\alpha,k-l-1}, \infty)$, где $u_{1-\alpha,k-l-1}$ — квантиль уровня $1-\alpha$ распределения хи-квадрат с $k-l-1$ степенями свободы. Вероятность ошибки первого рода приближенно равна α.

Таблицы сопряженности

Определение 3

В номинальных шкалах измерения представляют собой метки, обозначающие принадлежность измерения определенной градации измеряемого признака. Никаких содержательных соотношений кроме $x = y$ или $x \neq y$ между значениями в этих шкалах нет.

Для проверки независимости качественных признаков A и B, то есть, признаков, измеряемых в номинальных шкалах, применяются таблицы сопряженности.

Пусть имеется два качественных признака A и B. Признак A имеет r градаций: $A_1, \ldots, A_r$, признак B имеет s градаций $B_1, \ldots, B_s$. По выборке из n случайно выбранных объектов можно составить таблицу сопряженности:

	B_1	B_2	$\ldots$	B_s	
A_1	n_{11}	n_{12}	$\ldots$	n_{1s}	m_1
A_2	n_{21}	n_{22}	$\ldots$	n_{2s}	m_2
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	
A_r	n_{r1}	n_{r2}	$\ldots$	n_{rs}	m_r
	n_1	n_2	$\ldots$	n_s	n

где n_{ij} — количество элементов в выборке, обладающих одновременно свойствами A_i и B_j. Приведенная таблица называется *таблицей сопряженности*.

Справедливы равенства:

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

$$\sum_{j=1}^{s} n_j = \sum_{i=1}^{r} m_i = n.$$

Пусть $p_i = P(A_i)$, $i = 1, \ldots, r$ и $q_j = P(B_j)$, $j = 1, \ldots, s$.
При этом, $\sum_{i=1}^{r} p_i = 1$, $\sum_{j=1}^{s} q_j = 1$.

Признаки A и B называются независимыми, если при любых i и j выполняется равенство:

$$p_{ij} = P(A_i \cap B_j) = p_i q_j.$$

Очевидно, что $\sum_{i=1}^{r} p_{ij} = q_j$, $\sum_{j=1}^{s} p_{ij} = p_i$.

Сформулируем гипотезу независимости и альтернативную ей гипотезу:

- H_0: $P(A_i \cap B_j) = p_i q_j$ для любых i, j.
- H_1: существует пара (i, j) такая, что $P(A_i \cap B_j) \neq p_i q_j$.

Гипотеза H_0 представляет собой гипотезу независимости двух признаков.

Построим функцию правдоподобия выборки:

$$L = \frac{n!}{\prod\limits_{\substack{i=\overline{1,r} \\ j=\overline{1,s}}} n_{ij}!} \prod_{\substack{i=\overline{1,r} \\ j=\overline{1,s}}} (p_i q_j)^{n_{ij}}.$$

Найдем оценки максимального правдоподобия по выборке частот и воспользуемся теоремой 2.

Нетрудно заметить, что, максимизируя $\ln L$ по $p_1, \ldots, p_r$, $q_1, \ldots, q_s$ при ограничениях $\sum_{i=1}^{r} p_i = 1$ и $\sum_{j=1}^{s} q_j = 1$ методом неопределенных множителей Лагранжа, получим следующие оценки:

$$\hat{p}_i = \frac{m_i}{n}, \; i = 1, \ldots, r, \quad \hat{q}_j = \frac{n_j}{n}, \; j = 1, \ldots, s.$$

Статистика χ^2 для данной задачи после подстановки оценок метода максимального правдоподобия имеет вид:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - \frac{m_i n_j}{n})^2}{\frac{m_i n_j}{n}} \xrightarrow{\text{d}} \zeta \sim \chi^2_{(s-1)(r-1)}.$$

Число степеней свободы в предельном распределении хи-квадрат в соответствии с теоремой Фишера 2 вычисляется как $rs - (r-1) - (l-1) - 1 = (r-1)(s-1)$. Большие значения статистики хи-квадрат свидетельствуют против нулевой гипотезы H_0.

Получаем критерий для проверки гипотезы H_0:

- Если $\chi^2 > \chi^2_{\text{кр}}$, то отвергаем гипотезу H_0 в пользу альтернативной гипотезы H_1, где $\chi^2_{\text{кр}}$ представляет собой квантиль уровня $1-\alpha$ распределения хи-квадрат с $(s-1)(r-1)$ степенями свободы.
- Если $\chi^2 \leqslant \chi^2_{\text{кр}}$, то принимаем гипотезу H_0.

Точный критерий Фишера

Пусть имеется два качественных признака A и B. Признаки A и B имеют по 2 градации: A_1, A_2 и B_1, B_2, соответственно. По выборке из n случайно выбранных объектов можно составить таблицу сопряженности 2x2 T_0:

	B_1	B_2	сумма
A_1	n_{11}^0	n_{12}^0	m_1
A_2	n_{21}^0	n_{22}^0	m_2
сумма	n_1	n_2	n

Требуется проверить нулевую гипотезу о независимости признаков A и B.

Критерий χ^2 применим, если ожидаемые значения в любой из клеток таблицы сопряженности не меньше 5 ($n_j \geq 5$, $m_k \geq 5$). Когда это условие не выполняется, например, число наблюдений невелико, применяют *точный критерий Фишера*.

Точный критерий Фишера основан на переборе всех возможных вариантов заполнения таблицы сопряжженности T_i при зафикисрованных значениях m_1, m_2, n_1, n_2.

	B_1	B_2	сумма
A_1	n^i_{11}	n^i_{12}	m_1
A_2	n^i_{21}	n^i_{22}	m_2
сумма	n_1	n_2	n

Вероятность получить некоторую таблицу T_i равна

$$P_i = \frac{C_{m_1}^{n^i_{11}} C_{m_2}^{n^i_{21}}}{C_n^{n_1}} = \frac{m_1! m_2! n_1! n_2!}{n! n^i_{11}! n^i_{12}! n^i_{21}! n^i_{22}!} \tag{2}$$

Двусторонний вариант точного критерия Фишера основан на расчете вероятности

$$P = \sum_{T_i : P_i \leq P_0, i \neq 0} P_i + P_0.$$

где P_0 — вероятность получить исходную таблицу T_0, P_i — таблицу T_i при зафикисрованных значениях m_1, m_2, n_1, n_2.

Если вероятность P не превосходит уровень значимости α, тогда нулевую гипотезу о независмости признаков отклоняют. В противном случае нет оснований отклонять нулевую гипотезу.

Правила использования точного критерия Фишера:

1. Вычислите вероятность получить исходную таблицу
2. Постройте все возможные варианта заполнения таблицы при неизменных суммах по строкам и столбцам. (Для этого в одной из клеток проставляют все целые числа от 0 до максимально возможного, пересчитывая числа в остальных клетках так, чтобы суммы по столбцам и строкам оставались неизменными)
3. Вычислите вероятности для полученных таблиц
4. Просуммируйте вероятность получить исходную таблицу и все вероятности, которые не превышают ее.
5. Примите статистическое решение

- Большев Л. Н., Смирнов Н. В.Таблицы математической статистики. — М.: Изд. Наука, 1983.
- Тюрин Ю. Н., Макаров А. А.Статистический анализ опытных данных на компьютере. — Под ред. В.Э. Фигурнова. М.: ИНФРА-М, 1998.
- Холлендер М., Вулф Д.Непраметрические методы статистики. — М.: Финансы и статистика, 1983. — 518 с.
- Greenwood P. E., Nikulin M. S.A Guide to Chi-Squared Testing. New York, John Wiley & Sons, Inc., 1996.