О критериях хи-квадрат.

Рассмотрим последовательность n независимых испытаний, в результате каждого из которых происходит одно из несовместных событий $A_1, \ldots, A_r$: $A_k \cap A_j = 0$ при $k \neq j$, $\mathbf{P}\left\{\bigcup_{i=1}^{r} A_i\right\} = 1$. Обозначим A_{ij} событие, состоящее в том, что в i-ом испытании осуществилось событие A_j. Тогда случайные величины

$$\nu_j = \sum_{i=1}^{n} \mathbb{1}(A_{ij})$$

равны числу осуществлений событий A_j в n независимых испытаниях, $j = \overline{1, r}$.
Далее, пусть данные $\boldsymbol{\nu} = (\nu_1, \ldots \nu_r)^T$, $\sum_{j=1}^{r} \nu_j = n$, $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_r)^T$, $\pi_j = \mathbf{P}\{A_{ij}\}$, $i = \overline{1, n}$, $\sum_{j=1}^{r} \pi_j = 1$.

Вектор данных $\boldsymbol{\nu}$ имеет полиномиальное распределение

$$\mathbf{P}\left\{\bigcap_{j=1}^{r} (\nu_j = k_j); \theta\right\} = n! \prod_{j=1}^{r} \frac{\pi_j^{k_j}}{k_j!}. \tag{1}$$

При проверке простой гипотезы $\Gamma_1 : \boldsymbol{\theta}_0 = (\pi_{01}, \ldots, \pi_{0r})$ против сложной альтернативы $\Gamma_2 : \boldsymbol{\theta}_0 \neq (\pi_{01}, \ldots, \pi_{0r})$, где π_{0j} – заданные числа, $\sum_{j=1}^{r} \pi_{0j} = 1$, традиционно используется критерий хи-квадрат. Его статистика критерия имеет вид

$$T(y) = \sum_{j=1}^{r} \frac{(\nu_j - n\pi_{0j})^2}{n\pi_{0j}} \tag{2}$$

Для нахождения функции распределения статистики $T(y)$ при гипотезе Γ_1 необходимо, в случае конечных n, просуммировать соответствующие полиномиальные вероятности (1). В результате получим очень сложное выражение, существенно зависящее от $\boldsymbol{\theta}_0$, r и n.

Это обстоятельство исключает возможность табулирования $\mathcal{L}(T(y))$ или составления обозримой компьютерной программы. Асимптотическое же распределение статистики $T(y)$ при $n \to \infty$, в случае справедливости гипотезы Γ_1, обладает замечательным свойством *независимости от конкретных значений компонент вектора* $(\pi_{01}, \ldots, \pi_{0r})^T$.

Приведем предварительно ряд вспомогательных утверждений. Напомним, что квадратную матрицу $\mathbf{B}$ называют *идемпотентной*, если $\mathbf{B}^2 = \mathbf{B}$. Пусть $\dim \mathbf{B} = k \times k$.

Лемма 1. *Предположим, что вещественная симметрическая матрица* $\mathbf{B}$ *является идемпотентной. Тогда все ее собственные числа равны нулю или единице.*

Доказательство. Пусть $\mathbf{C}$ – ортогональная матрица, приводящая $\mathbf{B}$ к диагональному виду

$$\mathbf{C}^T \mathbf{B} \mathbf{C} = \mathbf{diag}(\lambda_1, \ldots, \lambda_k),$$

где $\lambda_1, \ldots, \lambda_k$ – собственные числа матрицы $\mathbf{B}$. Имеем

$$\begin{aligned} \mathbf{diag}(\lambda_1, \ldots, \lambda_k) &= \mathbf{C}^T\mathbf{B}\mathbf{C} = \mathbf{C}^T\mathbf{B}^2\mathbf{C} = \mathbf{C}^T\mathbf{B}\mathbf{B}\mathbf{C} = \\ &= \mathbf{C}^T\mathbf{B}\mathbf{C}\mathbf{C}^T\mathbf{B}\mathbf{C} = \{\mathbf{diag}(\lambda_1, \ldots, \lambda_k)\}^2 = \mathbf{diag}\{\lambda_1^2, \ldots, \lambda_k^2\}, \end{aligned} \tag{3}$$

т. е. $\lambda_i = \lambda_i^2$, $i = \overline{1, k}$. Лемма доказана. ■

Лемма 2. *Пусть* $\mathbf{Z} \stackrel{d}{=} \mathbf{N}_k(0, \mathbf{V})$, *а* $\lambda_1, \ldots, \lambda_k$ *– собственные числа матрицы* $\mathbf{V}$. *Тогда для квадратической формы* $\mathbf{Z}^T\mathbf{Z}$ *справедливо стохастическое представление*

$$\mathbf{Z}^T\mathbf{Z} \stackrel{d}{=} \sum_{i=1}^{k} \lambda_i N_i^2, \tag{4}$$

где N_i^2 *– н.о.р.,* $N_i \stackrel{d}{=} N(0,1)$, $i = \overline{1, k}$.

Доказательство. Пусть $\mathbf{C}$ – ортогональная матрица, приводящая ковариационную матрицу $\mathbf{V}$ к диагональному виду, т. е.

$$\mathbf{C}^T\mathbf{V}\mathbf{C} = diag(\lambda_1, \ldots, \lambda_k)^T. \tag{5}$$

Поскольку $\mathbf{V} \geq 0$, то все $\lambda_i \geq 0$, $i = \overline{1, k}$. Положим $\mathbf{X} = \mathbf{C}^T\mathbf{Z}$. В этом случае

$$\mathbf{X} = \mathbf{C}^T\mathbf{Z} \stackrel{d}{=} \mathbf{N}_k(0, diag(\lambda_1, \ldots, \lambda_k)) \stackrel{d}{=} diag(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_k})\mathbf{N}_k(0, \mathbb{I}). \tag{6}$$

Из (6), поскольку

$$\mathbf{C}^T\mathbf{C} = \mathbb{I}, \tag{7}$$

имеем

$$\begin{aligned} \mathbf{Z}^T\mathbf{Z} &= (\mathbf{C}\mathbf{X})^T(\mathbf{C}\mathbf{X}) = \mathbf{X}^T\mathbf{C}^T\mathbf{C}\mathbf{X} = \mathbf{X}^T\mathbf{X} = \\ &\stackrel{d}{=} (\mathbf{N}_k(0, \mathbb{I}))^T(diag(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_k}))^T diag(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_k})\mathbf{N}_k(0, \mathbb{I}) = \\ &\stackrel{d}{=} \sum_{i=1}^{k} \lambda_i N_i^2, \end{aligned} \tag{8}$$

т. е. из (8) следует справедливость (4). ■

Замечание 1. *Напомним, что следом квадратной матрицы*

$$\mathbf{B} = (b_{ij};\ i, j = \overline{1, k},\ j = \overline{1, k})$$

называют величину $tr\mathbf{B} = \sum_{i=1}^{k} b_{ii}$. *Если* λ_i, $i = \overline{1, k}$ *– собственные значения матрицы* $\mathbf{B}$, *то*

$$tr\mathbf{B} = \sum_{i=1}^{k} \lambda_i. \tag{9}$$

Лемма 3. *Если* $\mathbf{Z} = \mathbf{N}_k(0, \mathbf{V})$, *а* $\mathbf{V}$ *– идемпотентна, то*

$$\mathbf{Z}^T\mathbf{Z} \stackrel{d}{=} \chi_\nu^2, \quad \text{где} \quad \nu = tr\mathbf{V}. \tag{10}$$

Доказательство. Справедливость представления (10) вытекает из Замечания 1 и лемм 1 и 2. ■

Заметим, что непосредственно доказываются следующие соотношения

$$\left\{\begin{array}{rcl} tr(\mathbf{B}_1+\mathbf{B}_2) & = & tr\mathbf{B}_1+tr\mathbf{B}_2, \\ tr(\mathbf{B}_1\mathbf{B}_2) & = & tr(\mathbf{B}_2\mathbf{B}_1) \end{array}\right. \tag{11}$$

для любых квадратных матриц $\mathbf{B}_i$, $i=\overline{1,2}$ (см. Воеводин и Кузнецов (1984)).

Теорема 1. *Пусть гипотеза* Γ_1 *такова, что* $\pi_{0j}>0$, $j=\overline{1,r}$, $\sum\limits_{j=1}^{r}\pi_{0j}=1$. *Тогда при* $n\to\infty$

$$\sum_{j=1}^{r}\frac{(\nu_j-n\pi_{0j})^2}{n\pi_{0j}}\overset{d}{=}\chi^2_{r-1}+0_d(1). \tag{12}$$

Доказательство. Обозначим

$$\begin{array}{rcl} Z_j & = & \sqrt{n}\left(\dfrac{\dfrac{\nu_j}{n}-\pi_{0j}}{\sqrt{\pi_{0j}}}\right),\ j=\overline{1,r}, \\ \mathbf{Z} & = & (Z_1,\ldots,Z_r)^T, \\ \mathbf{L} & = & (l_1,\ldots,l_r)^T,\ l_j=\sqrt{\pi_{0j}},\ j=\overline{1,r}, \end{array}$$

Имеем, $\mathbf{EZ}=0$. Покажем, что

$$\mathbf{COV}(\mathbf{Z},\mathbf{Z})=\mathbf{1}-\mathbf{LL}^T. \tag{13}$$

$$\mathbf{D}Z_j=\frac{\mathbf{D}\nu_j}{n\pi_{0j}}=\frac{n\pi_{0j}(1-\pi_{0j})}{n\pi_{0j}}=1-\pi_{0j}=1-l_jl_j, \tag{14}$$

а внедиагональный (j,k)-ый элемент

$$\mathbf{cov}\{Z_j,Z_k\}=\frac{\mathbf{cov}\{\nu_j,\nu_k\}}{n\sqrt{\pi_{0j}\pi_{0k}}}=\frac{n\mathbf{cov}\{1(A_{ik}),1(A_{ik})\}}{n\sqrt{\pi_{0j}\pi_{0k}}}=-\sqrt{\pi_{0j}\pi_{0k}}=-l_jl_k. \tag{15}$$

Воспользуемся центральной предельной теоремой для полиномиального распределения при $n\to\infty$ получаем

$$\mathbf{Z}\overset{d}{=}\mathbf{N}_r(0,\mathbf{1}-\mathbf{LL}^T)+0_d(1). \tag{16}$$

Осталось показать идемпотентность ковариационной матрицы вектора Z:

$$(\mathbf{COV}(\mathbf{Z},\mathbf{Z}))^2=(\mathbf{1}-\mathbf{LL}^T)^2=\mathbf{1}-\mathbf{LL}^T-\mathbf{LL}^T+\mathbf{LL}^T\mathbf{LL}^T=\mathbf{1}-\mathbf{LL}^T=\mathbf{COV}(\mathbf{Z},\mathbf{Z})$$

Далее, из (11) вытекает, что $tr(\mathbf{1}-\mathbf{LL}^T)=tr\mathbf{1}-tr\mathbf{LL}^T=r-\sum\limits_{i=1}^{e}\pi_{0j}=r-1$. Используя представление (16), а также утверждение леммы 3, убеждаемся в справедливости теоремы 1.

■

Замечание 2. *Соотношение (12) было получено Беньяме еще в 1838 году (см. Джонсон, Котц и Балакришнан (1997)). Но лишь К. Пирсон в 1900 году предложил его использовать для расчета уровней значимости критериев типа хи-квадрат.*

Замечание 3. *Критическая область критерия хи-квадрат определяется как*

$$\{y : T(y) > C\}.$$

Постоянная $C = C(\alpha_1)$ *выбирается так, чтобы критерий имел асимптотическую ошибку первого рода* α_1 *:*

$$\alpha_1 = 1 - \mathcal{K}HI(C(\alpha_1); r-1). \tag{17}$$

Асимптотический наблюденный уровень значимости определяется как

$$\alpha_{\text{набл}} = 1 - \mathcal{K}HI(T(y); r-1). \tag{18}$$

■

Замечание 4. *Пусть* $y = (y_1, \dots y_r)$, y_i *– н.о.р. случайные величины,* $y_i \in R_1$. *Для проверки простой гипотезы* $\Gamma_1 : \mathcal{L}(y_i) = F_0(u)$ *против сложной альтернативы* $\Gamma_2 : \mathcal{L}(y_i) \neq F_0(u)$ *критерий согласия с* Γ_1 *часто строится на основе использования статистики (2) или ей эквивалентных статистик. Для этого прямая* R_1 *разбивается на* r *интервалов:*

$$(-\infty, a_1), [a_1, a_2), \dots, [a_j, a_{j+1}), \dots, [a_{r-1}, \infty),$$

где a_j, $j = \overline{1, r-1}$, *– задаются заранее, до получения* y_0. *Затем*

a) *подсчитываются частоты попадания наблюдений в отдельные интервалы* $\nu_j = n\left(\hat{F}_n(a_{j+1}) - \hat{F}_n(a_j)\right)$, *где* $\hat{F}_n(u) = \dfrac{1}{n}\sum\limits_{i=1}^{n} \mathbb{1}(y_i < u)$ *– эмпирическая функция распределения,*

b) *вычисляются вероятности*

$$\pi_{0j} = F_0(a_{j+1}) - F_0(a_j), \tag{19}$$

где $a_0 = -\infty$, $a_r = +\infty$, $j = \overline{0, r}$.

Вектор $\mathbf{x} = (\nu_1, \dots, \nu_r)^T$ *называют группированными данными. Иногда данные изначально можно получить лишь в группированном виде. В принятых обозначениях для проверки гипотезы* Γ_1 *применяется статистика (2), с естественной заменой символа* y *на* $\mathbf{x}$.

Соответствующий этой статистике асимптотический наблюденный уровень значимости вычисляется на основе приближения (12) по формуле (18).

Обратим внимание, что использование критерия хи-квадрат для проверки простой гипотезы $\Gamma_1 : \mathcal{L}(y_i) = F_0(u)$ *против сложной альтернативы* $\Gamma_2 : \mathcal{L}(y_i) \neq F_0(u)$ *на основе негруппированных одномерных данных типа независимой выборки целесообразно лишь для дискретных данных.*

Дело в том, что для проверки гипотезы Γ_1 *против альтернативы* Γ_2 *для непрерывных данных естественно использовать более мощный критерий Колмогорова (см. раздел "Критерий Колмогорова").* ■

Замечание 5. *Пусть* $y = (\mathbf{y}_1, \dots, \mathbf{y}_n)$, $\mathbf{y}_i$ *– н.о.р. случайные величины,* $\mathbf{y}_i \in R^k$. *В этом случае для проверки простой гипотезы*

$$\Gamma_1 : \mathcal{L}(\mathbf{y}_1) = F_0(u), \qquad u \in R^k, \tag{20}$$

против альтернативы

$$\Gamma_2 : \mathcal{L}(\mathbf{y}_1) \neq F_0(u), \qquad u \in R^k,$$

также можно построить критерий типа хи-квадрат. Для этого рассматривают конечное разбиение пространства R^k:

$$\mathcal{B} = \left\{ B_1, B_2, \dots, B_r;\ \bigcup_{i=1}^{r} B_i = R^k,\ B_i \bigcap B_j = \oslash \text{ при } i \neq j,\ i,j = \overline{1, r} \right\}$$

пространства множества значений случайной величины $\mathbf{y}_1$, *для которого*

$$\pi_{0i} = \mathbf{P}\{B_i\} > 0 \text{ *для* } i = \overline{1, r}.$$

После этого, для проверки гипотезы Γ_1 *можно использовать теорему 1.*

Отметим, что $F_0(u)$ однозначно определяет $\mathbf{P}_0\{B_i\}$, $i = \overline{1, r}$. Вычисление вероятностей $\mathbf{P}_0\{B_i\}$, как правило, сложная проблема даже для достаточно простых B_i. В частности, пусть $B = \left\{\bigcap_{i=1}^{k}\{a_i \le u_i < b_i\}\right\}$ – r-мерный прямоугольник, $\mathbf{u} = (u_1, \ldots, u_k)^T \in R^k$. Тогда

$$\mathbf{P}_0(B) = F_0(b_1, \ldots, b_k) - \sum_{i=1}^{k} q_i + \sum_{i<j} q_{ij} \mp \cdots + (-1)^k F_0(a_1, \ldots, a_k), \tag{21}$$

где $q_{ij\ldots t}$ обозначено значение $F_0(u_1, \ldots, u_k)$ при $u_i = a_i$,
$u_j = a_j \ldots, u_t = a_t$ и при остальных u_s, равных b_s (см. Гнеденко (1988), стр. 128).

Задача 1. Докажите, что статистика

$$\tilde{T}(y) = \sum_{j=1}^{r} \frac{(\nu_j - n\pi_{0j})^2}{\nu_j}$$

эквивалентна при $n \to \infty$ статистике (2), т. е.

$$\tilde{T}(y) = T(y) + o_p(1)$$

■

Задача 2. Докажите, что при фиксированной альтернативе к Γ_1 ошибка второго рода для критерия с асимптотической ошибкой первого рода, рассчитанной по формуле (17), экспоненциально по n убывает при $n \to \infty$.

Задача 3. Рассмотрим использование критерия хи-квадрат для проверки близкой к Γ_1 гипотезы

$$\Gamma_2 : \theta_0 = (\pi_{01} + \frac{b_1}{n^q}, \dots, \pi_{0r} + \frac{b_r}{n^q}), \; q > 0.$$

Обозначим

$$\alpha_2(n; T(y)) = \mathbf{P}\{T(y) < C(\alpha_1); \Gamma_2\}$$

Докажите, что

$$\lim_{n\to\infty} \alpha_2(n, T(y)) = \begin{cases} 1 - \alpha_1, & \frac{1}{2} < q \\ CHI\,(C(\alpha_1);\; r-1, \delta^2)\,, & q = \frac{1}{2} \\ 0, & q < \frac{1}{2} \end{cases},$$

где $\delta^2 = b_1^2 + \cdots + b_r^2$. ■