Новый критерий оценки качества классификационных моделей

Максим Гончаров
Декабрь 2013

В статье предлагается новый критерий оценки качества моделей классификации: увеличение вероятности совпадения значения целевой переменной со значением прогноза по сравнению с вероятностью их совпадения в случае независимости этих переменных. Получена формула для вычисления асимптотического доверительного интервала для этого критерия. Также получен статистический тест для проверки нулевой гипотезы, заключающейся в предположении, что вероятность совпадения целевой переменной и ее прогноза не превышает вероятности их совпадения если бы они были независимы. Получена асимптотическая значимость этого теста.

Пусть $(X_1, Y_1, Z_1), (X_2, Y_2, Z_2), \ldots$ – последовательность одинаково распределенных независимых случайных величин, где множества значений Y_i и Z_i конечны и совпадают, т.е. $Y_i, Z_i : \Omega \to \{1, \ldots, s\}$. Мы будем считать, что Y_i – это прогнозируемая целевая переменная, а Z_i – ее прогноз, зависящий только от набора входных переменных (предикторов) X_i, т.е. $Z_i = f(X_i)$. Далее, нас не будут интересовать предикторы X_i и метод получения прогноза f. Наша задача будет заключаться в оценке качества прогноза.

Введем обозначения:

$$\begin{aligned} p_k &:= \mathrm{P}(Y_i = k) \\ q_k &:= \mathrm{P}(Z_i = k) \\ r_{kl} &:= \mathrm{P}(Y_i = k, Z = l) \end{aligned} \quad \textbf{(1)}$$

Определим случайные величины:

$$\begin{aligned} U_i^{(k)} &:= 1_{\{Y_i = k\}} \\ V_i^{(k)} &:= 1_{\{Z_i = k\}} \\ W_i^{(k)} &:= 1_{\{Y_i = k, Z_i = k\}} \end{aligned} \quad \textbf{(2)}$$

Из (2) и (1) получаем для математических ожиданий:

$$\begin{aligned} EU_i^{(k)} &= E1_{\{Y_i = k\}} = \mathrm{P}(Y_i = k) = p_k \\ EV_i^{(k)} &= E1_{\{Z_i = k\}} = \mathrm{P}(Z_i = k) = q_k \\ EW_i^{(k)} &= E1_{\{Y_i = k, Z_i = k\}} = \mathrm{P}(Y_i = k, Z_i = k) = r_{kk} \end{aligned} \quad \textbf{(3)}$$

Далее получим выражения для дисперсий и ковариаций:

$$
\begin{aligned}
&Var\left(U_i^{(k)}\right) = E1_{\{Y_i=k\}}{}^2 - \left(E1_{\{Y_i=k\}}\right)^2 = p_k - p_k{}^2 \\
&Cov\left(U_i^{(k)}, U_i^{(l)}\right) = E1_{\{Y_i=k\}}1_{\{Y_i=l\}} - E1_{\{Y_i=k\}}E1_{\{Y_i=l\}} = -p_k p_l, k \neq l \\
&Var\left(V_i^{(k)}\right) = E1_{\{Z_i=k\}}{}^2 - \left(E1_{\{Z_i=k\}}\right)^2 = q_k - q_k{}^2 \\
&Cov\left(V_i^{(k)}, V_i^{(l)}\right) = E1_{\{Z_i=k\}}1_{\{Z_i=l\}} - E1_{\{Z_i=k\}}E1_{\{Z_i=l\}} = -q_k q_l, k \neq l \\
&Var\left(W_i^{(k)}\right) = E1_{\{Y_i=k,Z_i=k\}}{}^2 - \left(E1_{\{Y_i=k,Z_i=k\}}\right)^2 = r_{kk} - r_{kk}{}^2 \\
&Cov\left(W_i^{(k)}, W_i^{(l)}\right) = E1_{\{Y_i=k,Z_i=k\}}1_{\{Y_i=l,Z_i=l\}} - E1_{\{Y_i=k,Z_i=k\}}E1_{\{Y_i=l,Z_i=l\}} = -r_{kk} r_{ll}, k \neq l \\
&Cov\left(U_i^{(k)}, V_i^{(l)}\right) = E1_{\{Y_i=k\}}1_{\{Z_i=l\}} - E1_{\{Y_i=k\}}E1_{\{Z_i=l\}} = r_{kl} - p_k q_l \\
&Cov\left(U_i^{(k)}, W_i^{(l)}\right) = E1_{\{Y_i=k\}}1_{\{Y_i=l,Z_i=l\}} - E1_{\{Y_i=k\}}E1_{\{Y_i=l,Z_i=l\}} = -p_k r_{ll}, k \neq l \\
&Cov\left(U_i^{(k)}, W_i^{(k)}\right) = E1_{\{Y_i=k\}}1_{\{Y_i=k,Z_i=k\}} - E1_{\{Y_i=k\}}E1_{\{Y_i=k,Z_i=k\}} = r_{kk} - p_k r_{kk} \\
&Cov\left(V_i^{(k)}, W_i^{(l)}\right) = E1_{\{Z_i=k\}}1_{\{Y_i=l,Z_i=l\}} - E1_{\{Z_i=k\}}E1_{\{Y_i=l,Z_i=l\}} = -q_k r_{ll}, k \neq l \\
&Cov\left(V_i^{(k)}, W_i^{(k)}\right) = E1_{\{Z_i=k\}}1_{\{Y_i=k,Z_i=k\}} - E1_{\{Z_i=k\}}E1_{\{Y_i=k,Z_i=k\}} = r_{kk} - q_k r_{kk}
\end{aligned}
\qquad \textbf{(4)}, \textbf{(5)}
$$

Рассмотрим последовательность случайных величин при i=1,2,...

$$
T_i := \begin{pmatrix} U_i^{(1)} \\ \vdots \\ U_i^{(s)} \\ V_i^{(1)} \\ \vdots \\ V_i^{(s)} \\ W_i^{(1)} \\ \vdots \\ W_i^{(s)} \end{pmatrix} \qquad \textbf{(6)}
$$

Так как каждая T_i является измеримой функцией от (Y_i,Z_i) (см. формулу (2)), а последовательность (Y_1,Z_1), (Y_2,Z_2),.. представляет собой последовательность одинаково распределенных независимых случайных величин, то и последовательность T_1, T_2,... также является последовательностью одинаково распределенных независимых случайных величин с конечными вторыми моментами, определяемыми формулами (4) и (5). Таким образом мы можем применить центральную предельную теорему:

$$
\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} T_i - ET_1\right) \xrightarrow{d} N\left(0, \Sigma\right), \qquad \textbf{(7)}
$$

где $N\left(0,\Sigma\right)$ многомерно нормально распределенная случайная величина с нулевым вектором математического ожидания и матрицей ковариаций $\Sigma = Var\left(T_1\right)$.

Далее, с учетом (3):

$$ET_1 = \begin{pmatrix} EU_1^{(1)} \\ \vdots \\ EU_1^{(s)} \\ EV_1^{(1)} \\ \vdots \\ EV_1^{(s)} \\ EW_1^{(1)} \\ \vdots \\ EW_1^{(s)} \end{pmatrix} = \begin{pmatrix} p_1 \\ \vdots \\ p_s \\ q_1 \\ \vdots \\ q_s \\ r_{11} \\ \vdots \\ r_{ss} \end{pmatrix} \quad \textbf{(8)}$$

И с учетом (4) и (5):

$$\Sigma = Var\left(T_1\right) =$$

$$= \begin{pmatrix} Var\left(U_1^{(1)}\right) & \cdots & Cov\left(U_1^{(1)},U_1^{(s)}\right) & Cov\left(U_1^{(1)},V_1^{(1)}\right) & \cdots & Cov\left(U_1^{(1)},V_1^{(s)}\right) & Cov\left(U_1^{(1)},W_1^{(1)}\right) & \cdots & Cov\left(U_1^{(1)},W_1^{(s)}\right) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ Cov\left(U_1^{(s)},U_1^{(1)}\right) & \cdots & Var\left(U_1^{(s)}\right) & Cov\left(U_1^{(s)},V_1^{(1)}\right) & \cdots & Cov\left(U_1^{(s)},V_1^{(s)}\right) & Cov\left(U_1^{(s)},W_1^{(1)}\right) & \cdots & Cov\left(U_1^{(s)},W_1^{(s)}\right) \\ Cov\left(V_1^{(1)},U_1^{(1)}\right) & \cdots & Cov\left(V_1^{(1)},U_1^{(s)}\right) & Var\left(V_1^{(1)}\right) & \ldots & Cov\left(V_1^{(1)},V_1^{(s)}\right) & Cov\left(V_1^{(1)},W_1^{(1)}\right) & \cdots & Cov\left(V_1^{(1)},W_1^{(s)}\right) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ Cov\left(V_1^{(s)},U_1^{(1)}\right) & \cdots & Cov\left(V_1^{(s)},U_1^{(s)}\right) & Cov\left(V_1^{(s)},V_1^{(1)}\right) & \cdots & Var\left(V_1^{(s)}\right) & Cov\left(V_1^{(s)},W_1^{(1)}\right) & \cdots & Cov\left(V_1^{(s)},W_1^{(s)}\right) \\ Cov\left(W_1^{(1)},U_1^{(1)}\right) & \cdots & Cov\left(W_1^{(1)},U_1^{(s)}\right) & Cov\left(W_1^{(1)},V_1^{(1)}\right) & \cdots & Cov\left(W_1^{(1)},V_1^{(s)}\right) & Var\left(W_1^{(1)}\right) & \cdots & Cov\left(W_1^{(1)},W_1^{(s)}\right) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ Cov\left(W_1^{(s)},U_1^{(1)}\right) & \cdots & Cov\left(W_1^{(s)},U_1^{(s)}\right) & Cov\left(W_1^{(s)},V_1^{(1)}\right) & \cdots & Cov\left(W_1^{(s)},V_1^{(s)}\right) & Cov\left(W_1^{(s)},W_1^{(1)}\right) & \cdots & Var\left(W_1^{(s)}\right) \end{pmatrix} =$$

$$= \begin{pmatrix} p_1 - p_1^{\,2} & \cdots & -p_1 p_s & r_{11} - p_1 q_1 & \cdots & r_{1s} - p_1 q_s & r_{11} - p_1 r_{11} & \cdots & -p_1 r_{ss} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ -p_s p_1 & \cdots & p_s - p_s^{\,2} & r_{s1} - p_s q_1 & \cdots & r_{ss} - p_s q_s & -p_s r_{11} & \cdots & r_{ss} - p_s r_{ss} \\ r_{11} - p_1 q_1 & \cdots & r_{s1} - p_s q_1 & q_1 - q_1^{\,2} & \ldots & -q_1 q_s & r_{11} - q_1 r_{11} & \cdots & -q_1 r_{ss} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ r_{1s} - p_1 q_s & \cdots & r_{ss} - p_s q_s & -q_s q_1 & \cdots & q_s - q_s^{\,2} & -q_s r_{11} & \cdots & r_{ss} - q_s r_{ss} \\ r_{11} - p_1 r_{11} & \cdots & -p_s r_{11} & r_{11} - q_1 r_{11} & \cdots & -q_s r_{11} & r_{11} - r_{11}^{\,2} & \cdots & -r_{11} r_{ss} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ -p_1 r_{ss} & \cdots & r_{ss} - p_s r_{ss} & -q_1 r_{ss} & \cdots & r_{ss} - q_s r_{ss} & -r_{ss} r_{11} & \cdots & r_{ss} - r_{ss}^{\,2} \end{pmatrix} \quad \textbf{(9)}$$

Далее рассмотрим функцию:

$\varphi : \mathbb{R}^{3s} \to \mathbb{R},$

$$\varphi\left(u_1,...,u_s,v_1,...,v_s,w_1,...,w_s\right) := \sum_{k=1}^{s} w_k - \sum_{k=1}^{s} u_k v_k \quad \textbf{(10)}$$

Очевидно, φ дифференцируемая функция, для которой справедливо:

$$\frac{\partial\varphi}{\partial u_j}(p_1,...,p_s,q_1,...,q_s,r_{11},...,r_{ss}) = -q_j$$

$$\frac{\partial\varphi}{\partial v_j}(p_1,...,p_s,q_1,...,q_s,r_{11},...,r_{ss}) = -p_j$$

$$\frac{\partial\varphi}{\partial w_j}(p_1,...,p_s,q_1,...,q_s,r_{11},...,r_{ss}) = 1$$

Отсюда следует:

$$D\varphi(p_1,...,p_s,q_1,...,q_s,r_{11},...,r_{ss}) = D\varphi(ET_1) = (-q_1,...,-q_s,-p_1,...,-p_s,1,...,1) \quad \textbf{(11)}$$

Используя дельта-метод, получаем из (7):

$$\sqrt{n}\left(\varphi\left(\frac{1}{n}\sum_{i=1}^{n}T_i\right)-\varphi(ET_1)\right)\xrightarrow{d} D\varphi(ET_1)\cdot N(0,\Sigma) = N\left(0, D\varphi(ET_1)\cdot\Sigma\cdot\left[D\varphi(ET_1)\right]^t\right) \quad \textbf{(12)}$$

Далее,

$$\hat{T}_n := \varphi\left(\frac{1}{n}\sum_{i=1}^{n}T_i\right) = \varphi\begin{pmatrix}\frac{1}{n}\sum_{i=1}^{n}U_i^{(1)}\\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n}U_i^{(s)}\\ \frac{1}{n}\sum_{i=1}^{n}V_i^{(1)}\\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n}V_i^{(s)}\\ \frac{1}{n}\sum_{i=1}^{n}W_i^{(1)}\\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n}W_i^{(s)}\end{pmatrix} = \sum_{k=1}^{s}\frac{1}{n}\sum_{i=1}^{n}W_i^{(k)} - \sum_{k=1}^{s}\left(\frac{1}{n}\sum_{i=1}^{n}U_i^{(k)}\right)\left(\frac{1}{n}\sum_{j=1}^{n}V_j^{(k)}\right) =$$

$$= \sum_{k=1}^{s}\sum_{i=1}^{n}\frac{1}{n}1_{\{Y_i=k,Z_i=k\}} - \sum_{k=1}^{s}\left(\frac{1}{n}\sum_{i=1}^{n}1_{\{Y_i=k\}}\right)\left(\frac{1}{n}\sum_{j=1}^{n}1_{\{Z_j=k\}}\right) = \sum_{k=1}^{s}\frac{\nu_{kk}}{n} - \sum_{k=1}^{s}\frac{\mu_k}{n}\frac{\eta_k}{n} \quad \textbf{(13)}$$

Здесь

$\nu_{kk} := \sum_{i=1}^{n}1_{\{Y_i=k,Z_i=k\}}$ – число записей, у которых Y и Z равны k,

$\mu_k := \sum_{i=1}^{n}1_{\{Y_i=k\}}$ – число записей, у которых Y равно k,

$\eta_k := \sum_{i=1}^{n}1_{\{Z_i=k\}}$ – число записей, у которых Z равно k.

$$\theta := \varphi\left(ET_i\right) = \varphi\begin{pmatrix} p_1 \\ \vdots \\ p_s \\ q_1 \\ \vdots \\ q_s \\ r_{11} \\ \vdots \\ r_{ss} \end{pmatrix} = \sum_{k=1}^{s} r_{kk} - \sum_{k=1}^{s} p_k q_k =$$

$$= \sum_{k=1}^{s} \mathrm{P}\left(Y_i = k, Z_i = k\right) - \sum_{k=1}^{s} \mathrm{P}\left(Y_i = k\right)\mathrm{P}\left(Z_i = k\right) = \quad \textbf{(14)}$$

$$= \mathrm{P}\left(Y_i = Z_i\right) - \sum_{k=1}^{s} \mathrm{P}\left(Y_i = k\right)\mathrm{P}\left(Z_i = k\right) = \mathrm{P}\left(Y_i = Z_i\right) - \mathrm{P}_{\perp}\left(Y_i = Z_i\right)$$

Здесь $\mathrm{P}_{\perp}\left(Y_i = Z_i\right) := \sum_{k=1}^{s} \mathrm{P}\left(Y_i = k\right)\mathrm{P}\left(Z_i = k\right)$. **(15)**

Эта величина равна вероятности совпадения значений Y и Z если бы они были независимы. Действительно, в случае независимости Y и Z получаем:

$$\sum_{k=1}^{s} \mathrm{P}\left(Y_i = k\right)\mathrm{P}\left(Z_i = k\right) = \sum_{k=1}^{s} \mathrm{P}\left(Y_i = k, Z_i = k\right) = \mathrm{P}\left(Y_i = Z_i\right)$$

Итак, согласно (14), $\varphi\left(ET_i\right)$ есть разность вероятности совпадения Y и Z и вероятности их совпадения, если бы они были независимы.

Получим теперь выражение дисперсии для предельного нормального распределения в (12):

$$S := D\varphi\left(ET_1\right)\cdot\Sigma\cdot\left[D\varphi\left(ET_1\right)\right]^t =$$

$$= \left(-q_1,...,-q_s,-p_1,...,-p_s,1,...,1\right)\cdot$$

$$\cdot\begin{pmatrix} p_1 - p_1^{\,2} & \cdots & -p_1p_s & r_{11} - p_1q_1 & \cdots & r_{1s} - p_1q_s & r_{11} - p_1r_{11} & \cdots & -p_1r_{ss} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ -p_sp_1 & \cdots & p_s - p_s^{\,2} & r_{s1} - p_sq_1 & \cdots & r_{ss} - p_sq_s & -p_sr_{11} & \cdots & r_{ss} - p_sr_{ss} \\ r_{11} - p_1q_1 & \cdots & r_{s1} - p_sq_1 & q_1 - q_1^{\,2} & \cdots & -q_1q_s & r_{11} - q_1r_{11} & \cdots & -q_1r_{ss} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ r_{1s} - p_1q_s & \cdots & r_{ss} - p_sq_s & -q_sq_1 & \cdots & q_s - q_s^{\,2} & -q_sr_{11} & \cdots & r_{ss} - q_sr_{ss} \\ r_{11} - p_1r_{11} & \cdots & -p_sr_{11} & r_{11} - q_1r_{11} & \cdots & -q_sr_{11} & r_{11} - r_{11}^{\,2} & \cdots & -r_{11}r_{ss} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ -p_1r_{ss} & \cdots & r_{ss} - p_sr_{ss} & -q_1r_{ss} & \cdots & r_{ss} - q_sr_{ss} & -r_{ss}r_{11} & \cdots & r_{ss} - r_{ss}^{\,2} \end{pmatrix}\cdot$$

$$\cdot\left(-q_1,...,-q_s,-p_1,...,-p_s,1,...,1\right)^t = \sum_{k=1}^{s}\left(p_k^2q_k + p_kq_k^2 - 4p_k^2q_k^2 + 6p_kq_kr_{kk} + r_{kk} - r_{kk}^2 - 2p_kr_{kk} - 2q_kr_{kk}\right) -$$

$$-\sum_{k=1}^{s}\sum_{l=k+1}^{s}\left(2r_{kk}r_{ll} + 8p_kp_lq_kq_l\right) + \sum_{k=1}^{s}4p_kq_k\sum_{l=1,l\neq k}^{s} r_{ll} + \sum_{k=1}^{s}\sum_{l=1,l\neq k}^{s} 2p_kq_lr_{lk}$$

Итак,

$$S = D\varphi\left(ET_1\right)\cdot\Sigma\cdot\left[D\varphi\left(ET_1\right)\right]^t =$$

$$= \sum_{k=1}^{s}\left(p_k^2 q_k + p_k q_k^2 - 4p_k^2 q_k^2 + 6p_k q_k r_{kk} + r_{kk} - r_{kk}^2 - 2p_k r_{kk} - 2q_k r_{kk}\right) - \quad (16)$$

$$-\sum_{k=1}^{s}\sum_{l=k+1}^{s}\left(2r_{kk}r_{ll} + 8p_k p_l q_k q_l\right) + \sum_{k=1}^{s}4p_k q_k \sum_{l=1,l\neq k}^{s} r_{ll} + \sum_{k=1}^{s}\sum_{l=1,l\neq k}^{s} 2p_k q_l r_{lk}$$

Из (12) с учетом (13), (14) и (16) получаем:

$\sqrt{n}\left(\hat{T}_n - \theta\right)\xrightarrow{d} N\left(0,S\right)$, и, следовательно

$$\frac{1}{\sqrt{S}}\sqrt{n}\left(\hat{T}_n - \theta\right)\xrightarrow{d} N\left(0,1\right) \quad (17)$$

Определим

$$\hat{p}_0^{(n)} := \frac{1}{n}\sum_{i=1}^{n} 1_{\{Y_i=0\}},\ \hat{p}_1^{(n)} := \frac{1}{n}\sum_{i=1}^{n} 1_{\{Y_i=1\}},$$

$$\hat{q}_0^{(n)} := \frac{1}{n}\sum_{i=1}^{n} 1_{\{Z_i=0\}},\ \hat{q}_1^{(n)} := \frac{1}{n}\sum_{i=1}^{n} 1_{\{Z_i=1\}},$$

$$\hat{r}_{00}^{(n)} := \frac{1}{n}\sum_{i=1}^{n} 1_{\{Y_i=0,Z_i=0\}},\ \hat{r}_{01}^{(n)} := \frac{1}{n}\sum_{i=1}^{n} 1_{\{Y_i=0,Z_i=1\}},\ \hat{r}_{10}^{(n)} := \frac{1}{n}\sum_{i=1}^{n} 1_{\{Y_i=1,Z_i=0\}},\ \hat{r}_{11}^{(n)} := \frac{1}{n}\sum_{i=1}^{n} 1_{\{Y_i=1,Z_i=1\}}$$

Из усиленного закона больших чисел следует при $n\to\infty$ почти наверняка:

$$\hat{p}_0^{(n)}\to p_0,\ \hat{p}_1^{(n)}\to p_1,\ \hat{q}_0^{(n)}\to q_0,\ \hat{q}_1^{(n)}\to q_1,$$

$$\hat{r}_{00}^{(n)}\to r_{00},\ \hat{r}_{01}^{(n)}\to r_{01},\ \hat{r}_{10}^{(n)}\to r_{10},\ \hat{r}_{11}^{(n)}\to r_{11}$$

Следовательно, с вероятностью 1:

$$\hat{S}_n :=$$

$$= \sum_{k=1}^{s}\left(\left(\hat{p}_k^{(n)}\right)^2\hat{q}_k^{(n)} + \hat{p}_k^{(n)}\left(\hat{q}_k^{(n)}\right)^2 - 4\left(\hat{p}_k^{(n)}\right)^2\left(\hat{q}_k^{(n)}\right)^2 + 6\hat{p}_k^{(n)}\hat{q}_k^{(n)}\hat{r}_{kk}^{(n)} + \hat{r}_{kk}^{(n)} - \left(\hat{r}_{kk}^{(n)}\right)^2 - 2\hat{p}_k^{(n)}\hat{r}_{kk}^{(n)} - 2\hat{q}_k^{(n)}\hat{r}_{kk}^{(n)}\right) - \quad (18)$$

$$-\sum_{k=1}^{s}\sum_{l=k+1}^{s}\left(2\hat{r}_{kk}^{(n)}\hat{r}_{ll}^{(n)} + 8\hat{p}_k^{(n)}\hat{p}_l^{(n)}\hat{q}_k^{(n)}\hat{q}_l^{(n)}\right) + \sum_{k=1}^{s}4\hat{p}_k^{(n)}\hat{q}_k^{(n)}\sum_{l=1,l\neq k}^{s}\hat{r}_{ll}^{(n)} + \sum_{k=1}^{s}\sum_{l=1,l\neq k}^{s}2\hat{p}_k^{(n)}\hat{q}_k^{(n)}\hat{r}_{lk}^{(n)} \to S,(n\to\infty)$$

Отсюда $\frac{S}{\hat{S}_n}\xrightarrow{P}1,\left(n\to\infty\right)$

Используя теорему Слуцкого, получаем из (17) и (18):

$$\frac{1}{\sqrt{\hat{S}_n}}\sqrt{n}\left(\hat{T}_n - \theta\right) = \underbrace{\sqrt{\frac{S}{\hat{S}_n}}}_{\xrightarrow{P}1}\underbrace{\frac{1}{\sqrt{S}}\sqrt{n}\left(\hat{T}_n - \theta\right)}_{\xrightarrow{d}N(0,1)}\xrightarrow{d} N\left(0,1\right) \quad (19)$$

Мы получили, что случайная величина $\frac{1}{\sqrt{\hat{S}_n}}\sqrt{n}\left(\hat{T}_n - \theta\right)$ имеет асимптотическое стандартное нормальное распределение. Это позволяет получить асимптотический доверительный интервал для $\theta = \mathrm{P}\left(Y_i = Z_i\right) - \mathrm{P}_{\perp}\left(Y_i = Z_i\right)$:

$$1-\alpha = \mathrm{P}\left(y_{\frac{\alpha}{2}} \le N(0,1) \le y_{1-\frac{\alpha}{2}} \right) \approx \mathrm{P}\left(y_{\frac{\alpha}{2}} \le \frac{1}{\sqrt{\hat{S}_n}} \sqrt{n}\left(\theta - \hat{T}_n\right) \le y_{1-\frac{\alpha}{2}} \right) =$$

$$= \mathrm{P}\left(\hat{T}_n + y_{\frac{\alpha}{2}} \sqrt{\frac{\hat{S}_n}{n}} \le \theta \le \hat{T}_n + y_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{S}_n}{n}} \right) \quad \textbf{(20)}$$

Здесь y_β – это β-ый квантиль стандартного нормального распределения, т.е. $\mathrm{P}\left(N(0,1) \le y_\beta\right) = \beta$.

Далее рассмотрим следующие гипотезы:

H_0: Вероятность совпадения Y и Z не превышает вероятности их совпадения в случае независимости, т.е. $\theta = \mathrm{P}\left(Y_i = Z_i\right) - \mathrm{P}_\perp\left(Y_i = Z_i\right) \le 0$

H_1: Вероятность совпадения Y и Z больше вероятности их совпадения в случае независимости, т.е. $\theta = \mathrm{P}\left(Y_i = Z_i\right) - \mathrm{P}_\perp\left(Y_i = Z_i\right) > 0$

Пусть задано (малое) число $0 < \alpha < 1$.

Определим следующий статистический тест: если $\sqrt{\frac{n}{\hat{S}_n}} \hat{T}_n \ge y_{1-\alpha}$, то отвергаем гипотезу H_0 и принимаем альтернативную гипотезу H_1. Этот тест имеет асимптотический уровень значимости $1-\alpha$.

Действительно: пусть справедлива гипотеза H_0, следовательно $\theta \le 0$, и при этом $\sqrt{\frac{n}{\hat{S}_n}} \hat{T}_n \ge y_{1-\alpha}$. Тогда получаем $\sqrt{\frac{n}{\hat{S}_n}} \left(\hat{T}_n - \underbrace{\theta}_{\le 0} \right) \ge \sqrt{\frac{n}{\hat{S}_n}} \hat{T}_n \ge y_{1-\alpha}$, следовательно, мы получили событие $\sqrt{\frac{n}{\hat{S}_n}} \left(\hat{T}_n - \theta\right) \ge y_{1-\alpha}$.

Но согласно (19) вероятность этого события в пределе равна $\mathrm{P}\left(\sqrt{\frac{n}{\hat{S}_n}} \left(\hat{T}_n - \theta\right) \ge y_{1-\alpha} \right) \to \mathrm{P}\left(N(0,1) \ge y_{1-\alpha}\right) = \alpha$. Т.е. в случае справедливости гипотезы H_0, мы ошибочно отвергаем ее с вероятностью, асимптотически равной α, что доказывает асимптотический уровень значимости теста $1-\alpha$.