5.3 Тестирање на непараметарски хипотези

Нека X е обележје со непозната функција на распределба F_X. Непараметарските хипотези се однесуваат на функцијата на распределба на обележјето X и не зависат од непознатите параметри. Постојат неколку типа на непараметарски тестови. **Тестовите на согласност** ја тестирааат нултата хипотеза $H_0 : F_X = F$, каде F е дадена функција на распределба, против алтернативната хипотеза $H_1 : F_X \neq F$. Ако разгледуваме две обележја X и Y со непознати функции на распределба F_X и F_Y соодветно, **тестовите за хомогеност** ја тестирааат нултата хипотеза $H_0 : F_X = F_Y$, против алтернативната хипотеза $H_1 : F_X \neq F_Y$. **Тестовите за независност** ја тестирааат нултата хипотеза $H_0 : F = F_X F_Y$, каде F е функцијата на распределба на случајниот вектор (X, Y), против алтернативната хипотеза $H_1 : F \neq F_X F_Y$.

5.3.1 Пирсонов χ^2-тест на согласност

Пирсоновиот χ^2-тест на согласност е еден од првите предложени тестови кој се темели на едноставен математички модел. Со него се тестира хипотезата $H_0 : F_X = F$, против алтернативната хипотеза $H_1 : F_X \neq F$, каде F_X е непознатата функција на распределба на обележјето X и F е дадена функција на распределба. Главните карактеристики на овој тест кои му обезбедуваат широка примена се тие што тој може да се применува за произволна функција на распределба F и реализацијата на тест статистиката е лесно пресметлива.

Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето X. Дефинирањето на Пирсоновата χ^2 тест статистика е доста едноставно. Најнапред го разбиваме просторот од релани броеви на r дисјунктни подмножества, односно $\mathbb{R} = S_1 \cup S_2 \cup \cdots \cup S_r$, $S_i \cap S_j = \emptyset$, $i \neq j$. За секој $i \in \{1, 2, ..., r\}$ го означуваме со M_i бројот на случајни променливи од примерокот $(X_1, X_2, ..., X_n)$ кои примаат вредности од множеството S_i, и нека $p_i = P\{X \in S_i | H_0\}$. Тогаш, за секој $i \in \{1, 2, ..., r\}$, под претпоставка H_0 да е точна, случајната променлива M_i има $\mathcal{B}(n, p_i)$ распределба. И бидејќи $E(M_i) = np_i$, статистиката

$$\chi^2 = \sum_{i=1}^{r} \frac{(M_i - np_i)^2}{np_i} \tag{5.10}$$

добро го опишува отстапувањата на случајните променливи $M_1, M_2, ..., M_r$ од нивните математички очекувања. Статистиката χ^2 дефинирана со (5.10) се нарекува **Пирсонова χ^2 статистика**. Имено, се покажува дека распределбата на случајната променлива χ^2 асимптотски се стреми кон χ^2 распределба со $r - 1$ степени на слобода, односно

$$\chi^2 \xrightarrow{dist} \chi^2_{r-1}. \tag{5.11}$$

Ирена Стојковска

Знаејќи ја асимптотската распределба на тест статистиката χ^2, може да ја определиме критичната област со големина α (ниво на зачајност, веројатност за грешка од прв вид) за тестирање на хипотезата H_0. Имено, бидејќи

$$P\{\chi^2 > \chi^2_{r-1,\alpha} | H_0\} \approx \alpha, \tag{5.12}$$

каде бројот $\chi^2_{r-1,\alpha}$ се чита од таблица, имаме дека критичната област е

$$C = \{x \mid \overline{\chi}^2 = \sum_{i=1}^{r} \frac{(m_i - np_i)^2}{np_i} > \chi^2_{r-1,\alpha}\},$$

каде m_i е бројот на компоненти од реализацијата $x = (x_1, ..., x_n)$ на примерокот $(X_1, ..., X_n)$ кои припаѓаат во множеството S_i, $i \in \{1, ..., r\}$. Ова значи дека, ако врз база на податоците $x_1, ..., x_n$ добиеме дека $\overline{\chi}^2 > \chi^2_{r-1,\alpha}$, тогаш хипотезата H_0 ја отфрламе, во спротивно, ако $\overline{\chi}^2 \leq \chi^2_{r-1,\alpha}$, хипотезата H_0 ја прифаќаме. Да забележиме дека во пракса, апроксимацијата (5.12), а со тоа и самиот Пирсонов χ^2 - тест, дава задоволителни резултати за $n \geq 50$ и $m_i \geq 5$, за секој $i \in \{1, 2, ..., r\}$.

Пирсоновиот χ^2-тест може да се модифицира и за случај кога хипотезата $H_0 : F_X = F$ не е проста, односно кога функцијата на распределба F зависи од непознати праметри. Во тој случај нултата хипотеза е $H_0 : F_X \in \{F(x, \theta) | \theta \in \Theta\}$, па веројатностите $p_i(\theta) = P\{X \in S_i | H_0\}$, $i \in \{1, 2, ..., r\}$ зависат од непознатиот параметар θ. Тогаш, и Пирсоновата χ^2 статистика исто така ќе зависи од непознатиот параметар, односно

$$\chi^2(\theta) = \sum_{i=1}^{r} \frac{(M_i - np_i(\theta))^2}{np_i(\theta)}. \tag{5.13}$$

Нека $\theta = (\theta_1, ..., \theta_j)$, каде $j \leq r - 1$, односно функцијата на распределба F има j непознати параметри. Се покажува дека ако $\widehat{\theta}$ е максимално подобен оценувач за θ, тогаш важи

$$\chi^2(\widehat{\theta}) \xrightarrow{dist} \chi^2_{r-j-1}. \tag{5.14}$$

Имено, ова значи дека во случај кога функцијата на распределба F има непознати параметри (вкупно j непознати параметри), најнапред тие се заменуваат со нивни оценки врз база на дадените податоци $x_1, ..., x_n$ и со тоа распределбата F е потполно одредена. Па, заради (5.14), во овој случај критичната област ќе биде

$$C = \{x \mid \overline{\chi}^2 = \sum_{i=1}^{r} \frac{(m_i - np_i)^2}{np_i} > \chi^2_{r-j-1,\alpha}\},$$

па ако $\overline{\chi}^2 > \chi^2_{r-j-1,\alpha}$, тогаш хипотезата H_0 ја отфрламе, во спротивно, ако $\overline{\chi}^2 \leq \chi^2_{r-j-1,\alpha}$, хипотезата H_0 ја прифаќаме.

Ирена Стојковска

5.3.2 Колмогоров тест на согласност

Кога при тестирање на хипотезата $H_0 : F_X = F$, функцијата на распределба F е непрекината, за тест статистика може да се земе **Колмогоровата тест статистика** дефинирана со

$$D_n = D_n(X_1, ..., X_n) = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|, \tag{5.15}$$

каде $(X_1, ..., X_n)$ е примерок кој одговара на обележето X и $F_n(x)$ е емпириската функција на распределба на примерокот.

Се покажува дека при големи вредности на n, случајната променлива $\sqrt{n}D_n$ има асимптотска функција на распределба

$$K(x) = \lim_{n\to\infty} P\{\sqrt{n}D_n \leq x\} = 1 + 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2x^2}, \ x > 0,$$

која се нарекува **Колмогорова распределба**.

Тогаш, знаејќи ја распределбата на тест статистиката, може да се определи критичната вредност c_0 од критичната област C при тестирање на $H_0 : F_X = F$ со ниво на значајност α, од условот

$$P\{D_n(X_1, ..., X_n) > c_0 | H_0\} = \alpha.$$

Се изведува дека критичната област го има обликот

$$C = \{x \mid D_n(x) > d_{n,\alpha}\},$$

каде вредноста $d_{n,\alpha}$ е таква да $P\{D_n > d_{n,\alpha}\} = \alpha$ и се чита од таблица. Додека пак, при големи вредности на $n \geq 100$, критичната област го има обликот

$$C = \{x \mid D_n(x) > c_0\},$$

каде критичната вредност c_0 може да се пресметува со примена на следната таблица.

α	0,10	0,05	0,01
c_0	$\frac{1,22}{\sqrt{n}}$	$\frac{1,36}{\sqrt{n}}$	$\frac{1,63}{\sqrt{n}}$

Во пракса, вредноста на статистиката $D_n(x) = d_n$ за дадена низа од статистички податоци $x = (x_1, ..., x_n)$ со подреден облик $x_{(1)} \leq ... \leq x_{(n)}$ се пресметува според формулата

$$d_n = \max_{1\leq i\leq n} |F_n(x_{(i)}) - F(x_{(i)})|,$$

Ирена Стојковска

односно според формулата

$$d_n = \max_{1 \le i \le r} |F_n(a_i) - F(a_i)|,$$

каде a_i, $i = 1, ..., r$ се десните граници на интервалите при интервало зададени статистички податоци. Па, ако $d_n > d_{n,\alpha}$ (односно $d_n > c_0$ при вредности $n \ge 100$), тогаш хипотезата H_0 ја отфрламе, во спротивно, ако $d_n \le d_{n,\alpha}$ (односно $d_n \le c_0$ при вредности $n \ge 100$), хипотезата H_0 ја прифаќаме.

5.3.3 Тест за хомогеност на Колмгоров-Смирнов

Колмогоровата распределба наоѓа примена и при конструкција на статистичкиот тест за тестирање на хипотезата $H_0 : F_X = F_Y$, каде F_X и F_Y се функциите на распределба на непрекинатите обележја X и Y соодветно. Нека $(X_1, .., X_{n_1})$ и $(Y_1, .., Y_{n_2})$ се два независни примерока кои одговараат на обележјата X и Y соодветно. Во овој случај се користи **тест статистиката на Колмогоров-Смирнов** дефинирана со

$$D_{\frac{n_1 n_2}{n_1+n_2}} = D_{\frac{n_1 n_2}{n_1+n_2}}(X_1, ..., X_{n_1}, Y_1, ..., Y_{n_2}) = \sup_{x \in \mathbb{R}} |F_{n_1}(x) - F_{n_2}(x)|, \tag{5.16}$$

каде $F_{n_1}(x)$ и $F_{n_2}(x)$ се емпириските функции на распределба на примероците $(X_1, .., X_{n_1})$ и $(Y_1, .., Y_{n_2})$ соодветно. Ако хипотезата H_0 е точна, тогаш емпириските функции на распределба $F_{n_1}(x)$ и $F_{n_2}(x)$ оценуваат една иста функција на распределба $F_X = F_Y = F$, па за големи вредности на n_1 и n_2 природно е да се очекуваат мали реализирани вредности на тест статистиката $D_{\frac{n_1 n_2}{n_1+n_2}}$.

Се покажува дека при големи вредности на n_1 и n_2, случајната променлива $\sqrt{\frac{n_1 n_2}{n_1+n_2}} D_{\frac{n_1 n_2}{n_1+n_2}}$ има асимптотски Колмогорова распределба. Тогаш, критичната вредност c_0 од критичната област C при тестирање на H_0 со ниво на значајност α, се определува од условот

$$P\{D_{\frac{n_1 n_2}{n_1+n_2}}(X_1, ..., X_{n_1}, Y_1, ..., Y_{n_2}) > c_0 | H_0\} = \alpha.$$

Се добива дека критичната област го има обликот

$$C = \{(x, y) \mid D_{\frac{n_1 n_2}{n_1+n_2}}(x, y) > d_{\frac{n_1 n_2}{n_1+n_2}, \alpha}\},$$

каде вредноста $d_{n,\alpha}$ е таква да $P\{D_n > d_{n,\alpha}\} = \alpha$ и се чита од таблица. И ако ја означиме со $d_{\frac{n_1 n_2}{n_1+n_2}}$ реализацијата на статистиката $D_{\frac{n_1 n_2}{n_1+n_2}}(x, y)$ за дадени низи статистички податоци $x = (x_1, ..., x_{n_1})$ и $y = (y_1, ..., y_{n_2})$, тогаш ако $d_{\frac{n_1 n_2}{n_1+n_2}} > d_{\frac{n_1 n_2}{n_1+n_2}, \alpha}$, хипотезата H_0 ја отфрламе, во спротивно, ако $d_{\frac{n_1 n_2}{n_1+n_2}} \le d_{\frac{n_1 n_2}{n_1+n_2}, \alpha}$, тогаш хипотезата H_0 ја прифаќаме.

Ирена Стојковска

5.3.4 χ^2-тест за независност

Претходно зборувавме за зависноста меѓу обележјата X и Y, врз основа дводимензионалните статистички податоци $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, изразена преку вредностите на одредени параметри (криви на регресија, прави на регресија, коефициент на корелација, отстапување од статистичка независност, степен на статистичка зависност) и нивната геометриска интерпретација (види Дескриптивна статистика, Дводимензионални обележја). Овде ќе се задржиме на тестирање на хипотезата $H_0 : X$ и Y се независни случајни променливи, која симболички може да се запише како $H_0 : F = F_X F_Y$, каде F_X и F_Y се непознатите функции на распределба на обележјата X и Y соодветно, и F е функцијата на распределба на случајниот вектор (X, Y).

Нека $((X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n))$ е дводимензионален примерок кој одговара на дводимензионалното обележје (X, Y). Врз основа на овој примерок ќе конструираме статистички тест за тестирање на хипотезата H_0. Нека $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ е реализација на дводимензионалниот примерок дадена со табелата на контингенција на честотите, Табела 2.6, и соодветната табела на контингенција на релативните честоти, Табела 2.7. Во случај кога X и Y се непрекинати обележја, по групирањето на податоците во интервали, ја користиме Табела 2.11.

Ако хипотезата H_0 е точна, тогаш од условот за независност на две случајни променливи од дискретен тип важи

$$p_{i,j} = q_i r_j, \ i = 1, ..., r, \ j = 1, ..., s \tag{5.17}$$

(при тоа ги користиме ознаките од горе споменатите табели). Па, ако го означиме со $t = (q_1, ..., q_r, r_1, ..., r_s)$ векторот од непознати параметри, заради (5.17) и условите

$$\sum_{i=1}^{r} \sum_{j=1}^{s} p_{ij} = 1, \ \sum_{i=1}^{r} q_i = 1, \ \sum_{j=1}^{s} r_j = 1,$$

не се сите негови компоненти независни, имено постојат две функционални зависности, и тогаш димензијата на векторот t е $r + s - 2$.

Понатаму, со M_{ij} го означуваме бројот на случајни парови од примерокот $((X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n))$ кои ја примаат вредноста (a_i, b_j), повторно се користиме со ознаките од горе споменатите табели, и нека

$$G_i = M_{i1} + M_{i2} + ... + M_{is}, \ i = 1, ..., r, \ H_j = M_{1j} + M_{2j} + ... + M_{rj}, \ j = 1, ..., s. \tag{5.18}$$

Тогаш, слично како Пирсоновата χ^2 статистика (5.10), под претпоставка H_0 да е точна, статистиката

$$D = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(M_{ij} - np_{ij})^2}{np_{ij}} = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(M_{ij} - nq_i r_j)^2}{nq_i r_j} \xrightarrow{dist} \chi^2_{rs-1}, \tag{5.19}$$

Ирена Стојковска

го карактеризира отстапувањето на случајните променливи M_{ij} од нивните математички очекувања, односно од теориската распределба и таа има χ^2 распределба со $rs-1$ степени на слобода, според (5.11).

И слично како во (5.13), бидејќи p_{ij} зависат од векторот од непознати параметри t, тогаш и статистиката (5.19) ќе зависи од t, односно

$$D(t)=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(M_{ij}-np_{ij}(t))^2}{np_{ij}(t)}=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(M_{ij}-nq_i(t)r_j(t))^2}{nq_i(t)r_j(t)}. \tag{5.20}$$

Нека $\widehat{T}$ е максимално подобен оценувач за t, и бидејќи димензијата на t е $r+s-2$, слично како во (5.14), статистиката $D(\widehat{T})$ ќе има χ^2 распределба со $rs-(r+s-2)-1=rs-r-s+1=(r-1)(s-1)$ степени на слобода, односно

$$D(\widehat{T})=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(M_{ij}-np_{ij}(\widehat{T}))^2}{np_{ij}(\widehat{T})}=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(M_{ij}-nq_i(\widehat{T})r_j(\widehat{T}))^2}{nq_i(\widehat{T})r_j(\widehat{T})}\xrightarrow{dist}\chi^2_{(r-1)(s-1)}. \tag{5.21}$$

Од равенствата

$$q_i(\widehat{T})=\frac{G_i}{n}, i=1,...,r,\ r_j(\widehat{T})=\frac{H_j}{n},\ j=1,...,s,$$

каде G_i и H_j се дефинирани со (5.18), статистиката $D(\widehat{T})$ може да ја запишеме во еквивалентен облик

$$D(\widehat{T})=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(M_{ij}-\frac{G_iH_j}{n})^2}{\frac{G_iH_j}{n}}=n\Big(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{M_{ij}^2}{G_iH_j}-1\Big). \tag{5.22}$$

Тогаш, за определување на критичната област со големина α за тестирање на хипотезата H_0 се користиме со приближното равенство

$$P\{D(\widehat{T})>\chi^2_{(r-1)(s-1),\,\alpha}\}\approx\alpha,$$

и заклучуваме дека критичната област го има обликот

$$C=\{(x,y)\mid d=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(f_{ij}-\frac{g_ih_j}{n})^2}{\frac{g_ih_j}{n}}=n\Big(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{f_{ij}^2}{g_ih_j}-1\Big)=nf^2>\chi^2_{(r-1)(s-1),\,\alpha}\},$$

каде f_{ij},g_i,h_j се честотите кои за низата податоци $(x_1,y_1),(x_2,y_2),...,(x_n,y_n)$ дадени се со Табела 2.6, и f^2 е отстапувањето од статитичката независност дефинирано со (2.13). Тоа значи дека, ако $d>\chi^2_{(r-1)(s-1),\,\alpha}$, тогаш хипотезата H_0 ја отвфрламе, и ако $d\leq\chi^2_{(r-1)(s-1),\,\alpha}$, тогаш хипотезата H_0 ја прифаќаме.