3

Основни поими на математичката статистика

3.1 Статистички модел

При проучување на соберените статистички податоци во врска со некоја појава, се наметнува барањето да се донесе одреден заклучок во врска со разгледуваната појава. Токму ова е главната задача на математичката статистика. Донесувањето на заклучоци во врска со разгледуваната појава врз основа на конечен број на статистички податоци е познато како **статистичко заклучување**. Постојат главно два пристапа при донесувањето на статистички заклучоци. Првиот, **пристап на честоти** предложен од Кендел, е можеби и најкористениот во пракса и со него се претпоставува дека секој експеримент е бесконечно многу пати повторлив и дека мора да ги земеме во предвид сите можни исходи од експериментот за да може да донесеме некој статистички заклучок. Спротивно од овој пристап, **Бајесовиот пристап** ги темели заклучоците само на набљудуваните податоци и неодреденостите во врска со непознатите параметри се опушуваат преку распределби на веројатност кои зависат од овие податоци. Како и да е постојат и методи на статистичко заклучување кои претставуваат комбинација на овие два пристапа.

Во теоријата на статистичкото заклучување се конструираат математички модели кои овозможуваат егзактно дефинирање на проблемот, потоа со математички методи се пристапува кон решавање на проблемот и целта е да добиените резултати се применат во пракса и во другите научни дисциплини. Општа претпоставка при изградбата на теориските модели е дека низата статистички податоци $x_1, x_2, ..., x_n$ е некоја вредност на одреден случаен вектор $X = (X_1, X_2, ..., X_n)$. Претпоставуваме уште дека распределбата на веројатност на случајниот вектор X е непозната, но припаѓа на некоја фамилија $\mathcal{P}$

од **допустливи распределби на веројатност** за случајниот вектор X. Тогаш, просторот Ω од сите можни исходи на експериментот, σ-алгебрата $\mathcal{F}$ од подмножества од Ω и фамилијата $\mathcal{P}$ од допустливи распределби на веројатност (веројатносни мери) ја формираат тројката $(\Omega, \mathcal{F}, \mathcal{P})$ која се нарекува **статистички модел** за разгледуваниот експеримент. Понекогаш, терминот статистички модел се однесува и на самата фамилија $\mathcal{P}$ од допустливи распределби на веројатност.

За дефинирањето на фамилијата $\mathcal{P}$ не постојат прецизни критериуми, најчесто се изведува врз база на искуство и интуиција. Ако се земе претесна фамилија на допустливи распределби на веројатност, постои можност да вистинската распределба на веројатност остане надвор од таа фамилија, надвор од моделот. Ако за $\mathcal{P}$ се земе преширока класа, тогаш практично ништо нема да може да се заклучи за вистинската распределба врз основа на дадените податоци.

Ако фамилијата $\mathcal{P}$ од допустливи распределби на веројатност може да се опише преку конечен број на параметри $\theta = (\theta_1, ..., \theta_r)$, тогаш станува збор за **параметарски модел** на статистичко заклучување. Тогаш, може да запишеме дека

$$X = (X_1, X_2, ..., X_n) \sim F_\theta,\ \theta \in \Theta,$$

каде F_θ е функцијата на распределба на X и Θ е множеството од сите можни вредности за параметарот θ, познато како **простор на параметри**. Моделите пак чии распределби не можат да се опишат преку конечен број на параметри или начинот на изразување не е едноставен се нарекуваат **непараметарски модели**.

Многу често се претпоставува дека случајните променливи $X_1, X_2, ..., X_n$ се независни и еднакво распределени со заедничка распределба на веројатност P, односно се претопоставува дека податоците $x_1, x_2, ..., x_n$ се добиени како резултат на n независни мерења подложени на една иста статистичка законитост. Тогаш, фамилијата $\mathcal{P}$ од допустливи распределби на веројатност се стеснува на сите можни еднодимензионални распределби на веројатност.

Пример 3.1. (Биномен модел) Несиметрична монета се фрла n пати. Ако со 1 означуваме појава на "писмо", а со 0 појава на "грб", тогаш просторот од сите можни исходи од овој експримент е

$$\Omega = \{(x_1, x_2, ..., x_n) : x_i = 0 \text{ или } x_i = 1,\ i = 1, 2, ..., n\}.$$

Фамилијата од допустливи распределби на веројатност е $\mathcal{P} = \{P_\theta : 0 \leq \theta \leq 1\}$, каде веројатноста P_θ е дадена со

$$P_\theta(x) = \theta^{x_1+x_2+...+x_n}(1-\theta)^{n-(x_1+x_2+...+x_n)},\ x = (x_1, x_2, ..., x_n) \in \Omega,$$

Ирена Стојковска

каде θ е непознатата веројатност за појавување на "писмо" при едно фрлање на монетата. Просторот на параметри е $\{\theta : 0 \leq \theta \leq 1\}$. Овој статистички модел е познат како Биномен модел.

Пример 3.2. (Поасонов модел) Ако треба да се опише случајна променлива која претставува број на реализирани настани во фиксиран временски интервал, како на пример, бројот на телефонски разговори, или бројот на сообраќајни незгоди, или бројот на посетители на еден маркет, се користи Поасоновиот модел определен со множеството $\Omega = \{0, 1, 2, ...\}$ и фамилијата од допустливи распределби на веројатност $\mathcal{P} = \{P_\lambda : \lambda > 0\}$, каде веројатноста P_λ е дадена со

$$P_\lambda(x) = \frac{\lambda^x}{x!} e^{-\lambda},\ x = 0, 1, 2, ...$$

Просторот на параметри е $\{\lambda : \lambda > 0\}$.

Пример 3.3. (Гаусов (нормален) модел) Ако при мерење на некоја физичка величина со непозната вредност m, апаратот за мерење не прави системска грешка, туку на резултатот на мерењето влијаат голем број на случајни фактори, со незначително поединечно влијание, тогаш резултатот на мерењето се опишува со Гаусов (нормален) модел. Тогаш, може да се земе $\Omega = \mathbb{R}$ (сите можни резултати од мерењата), а за фамилијата од сите допустливи распределби на веројатност да се земе

$$\mathcal{P} = \{P_{m,\sigma^2} : -\infty < m < \infty,\ 0 < \sigma^2 < +\infty\},$$

каде веројатноста P_{m,σ^2} е дадена со

$$P_{m,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}},\ x \in \mathbb{R}.$$

Просторот на параметри е $\{(m, \sigma^2) : -\infty < m < \infty,\ 0 < \sigma^2 < +\infty\}$. Доколку претпоставиме дека дисперзијата σ^2 е позната и изнесува σ_0^2, тогаш фамилијата од сите допустливи распределби на веројатност е

$$\mathcal{P} = \{P_{m,\sigma_0^2} : -\infty < m < \infty\},$$

додека просторот на параметри е $\{m : -\infty < m < \infty\}$.

Пример 3.4. Нека $X_1, X_2, ..., X_n$ се независни и еднакво распределнеи случајни променливи со непрекината функција на распределба F која е непозната. Просторот на параметри за овој модел се состои од сите можни непрекинати распределби. И бидејќи овие распределби неможе да се индексираат со коечнодимензионален параметар, затоа овој модел се смета за **непараметарски модел**.

Ирена Стојковска

Исто така може да претпоставиме дека $F(x)$ има густина на распределба $p(x-\theta)$ каде θ е непознат параметар и p е непозната густина на распределба која го задоволува условот $p(x) = p(-x)$. Тогаш, овој модел е исто така непараметарски, но зависи од реално вредносен параметар θ. Затоа, тој може да се смета за **полупараметарски модел**.

Основни проблеми на статистичкото заклучување се **оценувањето на параметри**, односно наоѓање на нумерички вредности со кои се апроксимираат непознатите параметри на претпоставената распределба на веројатност и одредување на точност на таа апроксимација, и **тестирањето на хипотези**, односно дефинирање на постапки за донесување на одлуки за прифаќање, односно отфрлање на однапред поставената хипотеза за непознатите параметри или распределбата на веројатност.

Пример 3.5. Да претпоставиме дека некоја физичка величина m ја мериме n пати и нека резултатот на тие мерења се меѓусебно независни случајни променливи $X_1, X_2, ..., X_n$ со иста распределба која припаѓа на фамилијата допустливи распределби $\mathcal{P}$ дадена во Пример 3.3. Согласно законот на големите броеви имаме дека

$$\frac{X_1 + X_2 + ... + X_n}{n} \xrightarrow{\text{с.с.}} m.$$

Ако $x_1, x_2, ..., x_n$ се конкретните вредности добиени при n-те независни мерења, тогаш природно се наметнува да при големи вредности на n аритметичката средина $m_0 = \frac{x_1+x_2+...+x_n}{n}$ се смета за добра апроксимација на непознатиот параметар m. Оваа апроксимација се смета за оценка на параметарот m. Понатаму може да се применат разни критериуми за проценка колку е добра оваа оценка.

Пример 3.6. Од некои априорни причини, претпоставуваме дека вредноста на непознатиот параметар $\theta = \theta_0$ од Пример 3.1, каде $\theta_0 \in [0, 1]$ е даден број. Со помош на релативната честота $\frac{m}{n}$, каде m е бројот на појавувања на "писмо" при n-те независни фрлања т.е. $m = x_1 + x_2 + ... + x_n$, треба да се одлучиме дали ќе ја прифатиме хипотезата $\theta = \theta_0$ или ќе ја отфрлиме. Бидејќи при големи вредности на n релативната честота $\frac{m}{n}$ е блиска до веројатноста θ, добиваме еден критериум за проверка на хипотезата $\theta = \theta_0$ кој се базира на оценка на растојанието на релативната честота од вистинската вредност на веројатноста т.е. $|\frac{m}{n} - \theta|$. Ако ова растојание е големо, тогаш хипотезата ја отфрламе, ако е мало ја прифаќаме.

Ирена Стојковска

3.2 Популација, обележје и примерок

Доколку при изведување на некој експеримент секој елемент од одредено множество го избираме на случаен начин, тогаш тоа множесто може да се сфати како множество од сите можни исходи на експериментот, се означува со Ω и се нарекува **популација**. Популација (или **целна популација**) се дефинира уште и како целокупноста од еднородни елементи кои се предмет на истражување и за кои потребро е да се набави одредена информација. Набљудуваната заедничка карактеристика за елементите од популацијата која е предмет на истражување се нарекува **обележје**, се означува со X и се смета за случајна променлива чија распеделба на веројатност е непозната.

Пример 3.7. а) Се спроведува испитување за квалитетот на произведените сијалици во една фабрика. Севкупноста на сите произведени сијалици во фабриката ја претставува популацијата, додека обележје може да биде "должината на животот" на сијалицата во часови.

б) Ако се изведува испитување за успехот по математика во едно училиште, популацијата ја претставуваат учениците во училиштето, додека обележје може да биде нивната оценка по математика на звршниот испит.

в) При истражување за климатските услови во една земја, сите календарски години ја сочинуваат популацијата, додека обележје може да бидат вкупните врнежи на единица површина во текот на една година во земјата.

Нека експериментот се состои во избор на елемент од популацијата Ω и забележување на вредноста на обележето X која одговара на избраниот елемент. Резултат од овој експеримент е случајна променлива X. Ако експериментот се повтори n пати, како резултат се добива подредена n-торка од случајни променливи, односно случаен вектор $(X_1, X_2, ..., X_n)$ кој се нарекува **случаен примерок**. Бројот n се нарекува **големина на примерокот** или **обем на примерокот**. Ако случајните променливи $X_1, X_2, ..., X_n$ се независни и еднакво распределени со распределба еднаква на обележјето X, тогаш станува збор за **прост случаен примерок**, кој често се нарекува само **примерок**.

Примерокот треба да биде **репрезентативен** т.е. на правилен начин да ја претставува целата популација. Репрезентативноста може да се постигне ако секој елемент од популацијата има еднакви шанси да биде избран, и изборот на секој елемент да биде **случаен** и **независен**. Постојат повеќе методи, начини за избирање на репрезентативен примерок.

Ако за примерокот $(X_1, X_2, ..., X_n)$ ги забележиме вредностите на набљуваното обележје X за секоја од компонентите на примерокот, добиваме **реализација на примерокот** $(x_1, x_2, ..., x_n)$ која одговара на обележјето X, каде x_i е вредноста на случајната променлива X_i, $i = 1, 2, ..., n$. Често реализацијата на примерокот се нарекува примерок.