Statystyka - wykład III

Pomoce do samodzielnej pracy

TESTOWANIE HIPOTEZ STATYSTYCZNYCH

Teoria weryfikacji hipotez statystycznych jest istotnym działem wnioskowania statystycznego.

Hipoteza statystyczna – dowolne przypuszczenie co do rozkładu populacji generalnej (jego postaci funkcyjnej lub wartości parametrów). Prawdziwość tego przypuszczenia jest oceniania na podstawie wyników próby losowej.

Hipoteza parametryczna – jeśli sformułowane przypuszczenie dotyczy wartości parametrów rozkładu.

Pozostałe hipotezy nazywane są nieparametrycznymi.

Zbiór hipotez dopuszczalnych Ω – jest zbiorem rozkładów, które mogą charakteryzować populację. W zależności od tego, co wiemy o populacji, rozkłady należące do zbioru hipotez dopuszczalnych mogą różnić się postacią funkcyjną i wartością parametrów.

Każda hipotez statystyczna jest podzbiorem zbioru hipotez dopuszczalnych, co można zapisać

$$H: F(x) \in \omega, \quad gdzie\ \omega \in \Omega$$

natomiast F(x) jest dystrybuantą rozkładu zmiennej losowej w populacji generalnej.

Hipoteza prosta (hipoteza jednoznacznie specyfikuje rozkład populacji generalnej) – jeśli podzbiór ω zbioru hipotez dopuszczalnych Ω składa się z jednego elementu.

Hipoteza złożona – jeśli podzbiór ω zbioru hipotez dopuszczalnych Ω zawiera więcej niż jeden rozkład.

Po sformułowaniu odpowiedniej hipotezy dotyczącej populacji generalnej niezbędne jest określenie zasad weryfikacji tej hipotezy. To znaczy zasad postępowania umożliwiającego stwierdzenie na podstawie wyników próby, czy hipotezę tę można uznać za słuszną, czy nie.

Test statystyczny – reguła postępowania, która każdej możliwej próbie przyporządkowuje decyzję przyjęcia lub odrzucenia hipotezy. Jest to reguła rozstrzygająca, jakie wyniki próby pozwalają uznać sprawdzaną hipotezę za prawdziwą, jakie – za fałszywą.

Test parametryczny – służy do weryfikacji hipotezy parametrycznej.

Test nieparametryczny – służy do weryfikacji hipotezy nieparametrycznej.

Wyróżnia się kilka etapów konstruowania testów statystycznych.

1. Formułuje się hipotezę, która podlega weryfikacji. Tę sprawdzaną hipotezę nazywa się zerową i zapisuje jako:

$$H_0: F(x) \in \omega_0, \quad gdzie\ \omega_0 \subset \Omega$$

Następnie formułuje się hipotezę alternatywną i zapisuje jako:

$$H_1: F(x) \in \omega_1, \quad gdzie\ \omega_1 \subset \Omega$$

2. Określamy zbiór W wszystkich możliwych wyników n-elementowej próby (przestrzeń próby).

$W_n = (x_1, x_2, \dots, x_n)$ jest pewną próbą (punktem w przestrzeni próby). Konstrukcja testu polega na określeniu takiego obszaru przestrzeni próby w, że jeśli $W_n \in w$ (to znaczy wynik próby znajdzie się w tym obszarze), to sprawdzaną hipotezę zerową odrzucamy. Jeśli natomiast $W_n \in W\text{-}w$, to hipotezę H_0 przyjmujemy.

3. Obszar w jest to obszar odrzucenia hipotezy lub obszar krytyczny testu. Obszar $W\text{-}w$ jest to obszar przyjęcia hipotezy H_0.

4. W praktyce zwykle określa się odpowiednią statystykę z próby Z_n, której wartość z próby jest podstawą do podjęcia decyzji o przyjęciu lub odrzuceniu H_0 i dla której określa się obszar krytyczny. Statystyka taka jest nazywana sprawdzianem hipotezy.

Błąd I rodzaju – uznanie za fałszywą i odrzucenie hipotezy H_0, która w istocie jest prawdziwa.

Błąd II rodzaju – przyjęcie hipotezy H_0, która jest fałszywa.

TEST ISTOTNOŚCI DLA WARTOŚCI ŚREDNIEJ POPULACJI GENERALNEJ

Przyjmujemy, że populacja generalna ma rozkład normalny o nieznanej wartości średniej *m* i znanym odchyleniu standardowym σ. Hipoteza zerowa jest przypuszczeniem, że średnia *m* ma wartość m_0, czyli

$H_0 : m = m_0$

wobec hipotezy alternatywnej

$H_1 : m \neq m_0$

Hipotezę H_0 należy zweryfikować na podstawie wyników n-elementowej próby $(X_1, X_2, \ldots, X_n)$. Za sprawdzian tej hipotezy przyjmujemy średnią arytmetyczną $\bar{X}$ z próby.

Średnia arytmetyczna z próby pobranej z populacji o rozkładzie normalnym $N(m, \sigma)$ ma rozkład $N(m, \frac{\sigma}{\sqrt{n}})$.

Jeśli prawdziwa jest hipoteza zerowa, to $\bar{X}$ ma rozkład $N(m_0, \frac{\sigma}{\sqrt{n}})$. A zatem jeśli H_0 jest prawdziwa, to statystyka o postaci

$$U = \frac{\bar{X} - m_0}{\sigma}\sqrt{n}$$

ma rozkład $N(0, 1)$.

Jeśli prawdziwa jest hipoteza zerowa, to wartość średniej arytmetycznej otrzymana z próby nie powinna zbytnio różnić się od hipotetycznej wartości m_0 i moduł statystyki U nie powinien przyjmować zbyt dużych wartości. To znaczy nie powinien przekraczać pewnej wartości u_α zwanej wartością krytyczną, która dla ustalonego poziomu istotności α jest określona w rozkładzie N(0, 1) w taki sposób, aby zachodziła relacja

$$P(|U| \geq u_\alpha) = \alpha$$

α – poziom istotności

Wartości zmiennej U spełniające nierówność $|U| \geq u_{\alpha}$ tworzą obszar krytyczny testu. Jeśli z próby otrzymamy taką wartość u tej statystyki, że $|u| \geq u_{\alpha}$, to hipotezę zerową odrzucamy na korzyść hipotezy alternatywnej. Oznacza to bowiem, że U przyjęła (co do modułu) zbyt dużą wartość, tzn. różnica między średnią z próby $\bar{x}$ a hipotetyczną wartością m_0 jest zbyt duża, aby można było przyjąć hipotezę zerową. Uznajemy, że różnica między wartością średniej arytmetycznej $\bar{x}$ a m_0 jest statystycznie istotna. Odrzucając hipotezę zerową musimy mieć świadomość popełnienia błędu. Prawdopodobieństwo tego, że odrzucimy hipotezę, która jest prawdziwa wynosi α. W przeciwnym przypadku, tzn. gdy wynik z próby jest taki, że $|u| < u_{\alpha}$ stwierdzamy, że nie ma podstaw do odrzucenia hipotezy zerowej (tzn. wyniki z próby nie dają podstaw do odrzucenia H_0.

ROZKŁAD POISSONA

Zastosowanie tego rozkładu umożliwia w sposób przybliżony charakteryzować takie zjawiska, jak liczba usterek, liczba awarii, liczba błędów itp. Zatem rozkład ten jest często wykorzystywany w analizie niezawodnościowej systemów technicznych (np. wodociągowych lub kanalizacyjnych).

Definicja

Zmienna losowa X przyjmująca wartości $k = 0, 1, 2, \ldots$ ma rozkład Poissona o parametrze λ, jeśli jej funkcja prawdopodobieństwa opisana jest wzorem:

$$P(X = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$$

dla $k = 0, 1, 2, \ldots$

gdzie λ jest dodatnią stałą ($\lambda > 0$)

Dystrybuantę rozkładu Poissona określa wzór:

$$F(x) = \sum_{k \leq x} \frac{\lambda^k}{k!} \cdot e^{-\lambda}, \quad x \in R$$

Opierając się na definicji wartości oczekiwanej i wariancji zmiennej losowej dla rozkładu Poissona otrzymujemy:

$E(X) = \lambda$

$D^2(X) = \lambda$

A zatem parametr λ jest średnią i wariancją zmiennej losowej o rozkładzie Poissona.

TEST ZGODNOŚCI CHI-KWADRAT

Obszerną klasę testów istotności stanowią testy nieparametryczne, służące do sprawdzania hipotez nieparametrycznych. Zalicza się do nich tzw. testy zgodności.

Za ich pomocą sprawdza się zgodność rozkładu empirycznego z próby z rozkładem hipotetycznym lub też zgodność dwóch lub więcej rozkładów z próby.

Popularnym testem zgodności jest test skonstruowany na podstawie statystyki χ^2. Niech hipotezą zerową będzie przypuszczenie, że populacja generalna ma rozkład określony pewną dystrybuantą $F_0(x)$:

$$H_0 : F(x) = F_0(x),$$

wobec hipotezy alternatywnej:

$$H_1 : F(x) \neq F_0(x)$$

W celu weryfikacji tej hipotezy należy wylosować z populacji dużą próbę (dużą, bo korzystamy z granicznego rozkładu pewnej statystyki). Wyniki z próby należy przedstawić w postaci rozkładu empirycznego poprzez utworzenie *r* rozłącznych klas wartości badanej zmiennej w próbie. Liczebność w *i*-tej klasie oznaczamy symbolem n_i ($i = 1, \ldots, r$). Przyjmując, że prawdziwa jest hipoteza H_0, tzn. że rozkład populacji generalnej jest opisany dystrybuantą $F_0(x)$, należy obliczyć prawdopodobieństwo p_i tego, że badana zmienna losowa przyjmie wartość z *i*-tej klasy. Gdyby prawdziwa była hipoteza zerowa, to liczebności w poszczególnych klasach powinny wynosić np_i ($i = 1, \ldots, r$), gdzie *n* jest liczebnością próby.

Podstawą do konstrukcji miary zgodności rozkładu empirycznego z hipotetycznym jest różnica między liczebnościami zaobserwowanymi n_i a liczebnościami teoretycznymi (hipotetycznymi) np_i. Do oceny zgodności tych rozkładów stosuje się statystykę o postaci:

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

która przy założeniu prawdziwości hipotezy zerowej ma asymptotyczny rozkład χ^2 o r-k-1 stopniach swobody, gdzie r jest liczbą klas wartości zmiennej, natomiast k oznacza liczbę parametrów rozkładu. Jeśli prawdziwa jest hipoteza zerowa, to statystyka χ^2 nie powinna przyjmować dużych wartości.

Obszar krytyczny jest określony przez:

$$P(\chi^2 \geq \chi^2_{\alpha,r-k-1}) = \alpha$$

gdzie α jest poziomem istotności, zaś $\chi^2_{\alpha,r-k-1}$ jest wartością krytyczną wyznaczoną z rozkładu chi-kwadrat.

Jeśli wartość statystyki χ^2 z próby jest taka, że:

$$\chi^2 \geq \chi^2_{\alpha,r-k-1},$$

to oznacza to, że różnica między rozkładem empirycznym a hipotetycznym (tzn. między odpowiednimi liczebnościami n_i oraz np_i) jest statystycznie istotna i hipotezę zerową należy odrzucić.

INFORMACJE POMOCNICZE

Ważone średnie arytmetyczne – gdy dane są w postaci szeregu rozdzielczego.

x_i – wyróżnione wartości w rozkładzie

n_i – liczebności klasowe

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i, \quad i = 1, \dots, k$$