PJWSTK Informatyka studia dzienne - SAD Ćwiczenia 15

Testy nieparametryczne

Zadanie 1 Badano, czy istnieje związek między kolorem oczu i kolorem włosów. W tym celu przeprowadzono badanie na losowej grupie 220 osób i otrzymano następujące wyniki:

	niebieski kolor oczu	inny kolor oczu
włosy jasne	67	32
włosy ciemne	53	68

Prowadzący badanie twierdzi, że otrzymane wyniki wskazują na istnienie związku między kolorem oczu i włosów. Czy ma rację? Zweryfikować odpowiednią hipotezę na poziomie istotności 0,01.

Zadanie 2 Badano, czy istnieje związek między nadciśnieniem a nadwagą. W tym celu przeprowadzono badanie na losowej grupie 520 osób i otrzymano następujące wyniki:

	nadciśnienie	ciśnienie w normie
nadwaga	34	162
waga w normie	136	188

Czy na podstawie tych danych można twierdzić, że istnieje zależność między nadwagą i nadciśnieniem? Zweryfikować odpowiednią hipotezę na poziomie istotności 0,05.

Rozw. zad.1.

- (X, Y) - para cech jakościowych, $X -$ kolor włosów, $Y -$ kolor oczu
- Pytanie: czy X, Y są niezależne

	Niebieski kolor oczu	Inny kolor oczu	$n_{i.}$
Włosy jasne	67	32	99
Włosy ciemne	53	68	121
$n_{.j}$	120	100	n =220

1. Hipotezy:
   - H_0: X, Y są niezależnymi zmiennymi losowymi: kolor włosów i kolor oczu są cechami niezależnymi
   - H_1: zaprzeczenie H_0, tzn. zmienne losowe X, Y są zależne

2. Statystyka testowa:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(N_{ij} - \widehat{N_{ij}}\right)^2}{\widehat{N_{ij}}} \sim \chi^2_1,$$

jeśli hipoteza zerowa jest prawdziwa

- liczba stopni swobody rozkładu chi-kwadrat = 1 = (2-1)(2-1), $\widehat{N_{ij}} = \frac{N_{i.}N_{.j}}{n} = estymator\ \ E(N_{ij})$
- $N_{i.} = N_{i1} + N_{i2}$ = liczba elementów próby, których cecha X ma wartość i-tą
- $N_{.j} = N_{1j} + N_{2j}$ = liczba elementów próby, których cecha Y ma wartość j-tą

3. Wartość statystyki testowej

X (kolor włosów)	Niebieski kolor oczu 1	Inny kolor oczu 2	$n_{i.}$
Włosy jasne (=1)	67	32	99
Włosy ciemne (=2)	53	68	121
$n_{.j}$	120	100	n =220

Y (kolor oczu) X (kolor włosów)	Niebieski kolor oczu 1	Inny kolor oczu 2	$n_{i.}$
Włosy jasne (=1)	67 54	32 45	99
Włosy ciemne (=2)	53 66	68 55	121
$n_{.j}$	120	100	n =220

gdzie w tabeli policzono

$$\hat{n}_{ij} = \frac{n_{i.}n_{.j}}{n}, \qquad \text{np. } \hat{n}_{11} = \frac{n_{1.}n_{.1}}{n} = \frac{99 \cdot 120}{220} = 54$$

$$\chi^2_{obs} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(n_{ij} - n_{i.}n_{.j}/220\right)^2}{n_{i.}n_{.j}/220} = \frac{(67-54)^2}{54} + \frac{(32-45)^2}{45} +$$

$$+\frac{(53-66)^2}{66} + \frac{(68-55)^2}{55} = 3{,}13 + 3{,}76 + 2{,}56 + 3{,}07 = 12{,}52$$

5. Zbiór krytyczny

$$C = \left[\chi'^2_{1-\alpha,(k-1,r-1)}, \infty\right) = \left[\chi'^2_{0{,}99,1}, \infty\right) = [6{,}635; \infty)$$

6. $\chi^2_{obs} \in C$, zatem można twierdzić, na poziomie istotności 0,01, że kolor włosów i kolor oczu są cechami zależnymi.

Rozw. zad. 2.

- (X, Y) - para cech jakościowych,
- $X = \begin{cases} 1, & jeśli\ losowo\ wybrana\ osoba\ ma\ nadwagę \\ 2, & jeśli\ losowo\ wybrana\ osoba\ ma\ wagę\ w\ normie \end{cases}$
- $Y = \begin{cases} 1, & jeśli\ losowo\ wybrana\ osoba\ ma\ nadciśnienie \\ 2, & jeśli\ losowo\ wybrana\ osoba\ ma\ ciśnienie\ w\ normie \end{cases}$
- Pytanie: czy X, Y są niezależne

1. Hipotezy:
   - H_0: X, Y są niezależnymi zmiennymi losowymi: ciśnienie i waga są cechami niezależnymi
   - H_1: zaprzeczenie H_0, tzn. zmienne losowe X, Y są zależne

2. Statystyka testowa:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(N_{ij} - \widehat{N_{ij}}\right)^2}{\widehat{N_{ij}}} \sim \chi^2_1,$$

jeśli hipoteza zerowa jest prawdziwa

- liczba stopni swobody rozkładu chi-kwadrat = 1 = (2-1)(2-1), $\widehat{N_{ij}} = \frac{N_{i.}N_{.j}}{n} = estymator\ \ E(N_{ij})$
- $N_{i.} = N_{i1} + N_{i2}$ = liczba elementów próby, których cecha X ma wartość i-tą
- $N_{.j} = N_{1j} + N_{2j}$ = liczba elementów próby, których cecha Y ma wartość j-tą

3. Wartość statystyki testowej

	nadciśnienie	Ciśnienie w normie	$n_{i.}$
Nadwaga	34 64,08	162 131,92	196
Waga w normie	136 105,92	188 218,08	324
$n_{.j}$	170	350	n =520

$$\chi^2_{obs} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(n_{ij} - n_{i.}n_{.j}/520\right)^2}{n_{i.}n_{.j}/520} = \frac{(34 - 64{,}08)^2}{64{,}08} + \frac{(162 - 131{,}92)^2}{131{,}92} +$$

$$+ \frac{(136 - 105{,}92)^2}{105{,}92} + \frac{(188 - 218{,}08)^2}{218{,}08} = 14{,}12 + 6{,}86 + 8{,}54 + 4{,}35 = 33{,}87$$

5. Zbiór krytyczny

$$C = \left[\chi^{2}_{1-\alpha,(k-1,r-1)}, \infty\right) = \left[\chi^{2}_{0,95;1}, \infty\right) = [3{,}8415; \infty)$$

6. $\chi^2_{obs} \in C$, zatem można twierdzić, na poziomie istotności 0,05, że istnieje zależność między wagą i ciśnieniem.

Twierdzenie - sformułowanie CTG dla sumy S_n i dla średniej z próby losowej $\bar{X}$

Jeśli $X_1, X_2, \dots, X_n$ są niezależnymi zmiennymi losowymi o tym samym rozkładzie prawdopodobieństwa z wartością oczekiwaną μ oraz odchyleniem standardowym σ, to dla dużych n rozkład

- sumy $S_n = X_1 + X_2 + \cdots + X_n$ jest bliski rozkładowi normalnemu z wartością oczekiwaną $E(S_n) = n\mu$ i wariancją $Var(S_n) = n\sigma^2$
- średniej z próby losowej

  $$\bar{X} := \frac{S_n}{n}$$

  jest bliski rozkładowi normalnemu z wartością oczekiwaną $E(\bar{X}) = \mu$ i wariancją $Var(\bar{X}) = \frac{\sigma^2}{n}$

Uwaga. Zazwyczaj wystarczy aby $n > 25 \quad (lub\ 30)$.
CTG dla sumy i średniej z próby losowej mówi, że dla każdego x:

$$P(S_n \le x) = P\left(\frac{S_n - n\mu}{\sqrt{n}\cdot\sigma} \le \frac{x - n\mu}{\sqrt{n}\cdot\sigma}\right) \approx \Phi\left(\frac{x - n\mu}{\sqrt{n}\cdot\sigma}\right)$$

$$P(\bar{X} \le x) = P\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \le \frac{x - \mu}{\frac{\sigma}{\sqrt{n}}}\right) \approx \Phi\left(\frac{x - \mu}{\frac{\sigma}{\sqrt{n}}}\right)$$

Notacja:

$$S_n \approx N(n\mu, \sqrt{n}\sigma), \quad \bar{X} \approx N(\mu, \tfrac{\sigma}{\sqrt{n}})$$

Wniosek z twierdzenia (twierdzenie Moivre'a – Laplace'a)

Jeśli $S_n \sim Binomial(n, p)$, $j = 1,2, \dots, n$, to dla $np \ge 5$, $nq \ge 5$

- $P(S_n \le x) = P\left(\frac{S_n - np}{\sqrt{npq}} \le \frac{x + \frac{1}{2} - np}{\sqrt{npq}}\right) \approx \Phi\left(\frac{x + \frac{1}{2} - np}{\sqrt{npq}}\right), gdy\ n \le 100$

wprowadzamy tzw. poprawkę ciągłości (otrzymamy lepsze oszacowanie)

- $P(S_n \leq x) = P\left(\frac{S_n - np}{\sqrt{npq}} \leq \frac{x-np}{\sqrt{npq}}\right) \approx \Phi\left(\frac{x-np}{\sqrt{npq}}\right),\ \ gdy\ n > 100$

Uzasadnienie:

$$S_n \sim X_1 + X_2 + \cdots + X_n,\ \ gdzie\ \ X_j \sim Bin(1,p), \qquad \mu = p, \sigma^2 = pq$$

Zadanie. Wykonano 30 rzutów kostką symetryczną. (a) Obliczyć wartość oczekiwaną i wariancję łącznej liczby wyrzuconych oczek. (b) Znaleźć przybliżoną wartość prawdopodobieństwa, że łączna liczba wyrzuconych oczek przekroczy 100.

Rozw. Rozkład prawdopodobieństwa zmiennej X jest jednostajny na zbiorze {1,2,3,4,5,6}:

P(X=i)=1/6, i =1, 2, 3, 4,5, 6.

E(X)= (1/6)(1+2+3+4+5+6) = 3,5, E(X^2)= (1/6)(1+4+9+16+25+36) =91/6

Var(X)= 91/6 – $(3,5)^2$ = 2,9167

$$E(S_{30}) = E(X_1 + X_2 + \cdots + X_{30}) = 30E(X) = 105$$

$$Var(S_{30}) = Var(X_1 + X_2 + \cdots + X_{30}) = 30Var(X) = 87,5$$

gdzie X_j = liczba oczek w j-tym rzucie kostką, j = 1,2,…,30. Zmienne $X_1, X_2, \ldots, X_{30}$ są niezależne i mają rozkład prawdopodobieństwa taki jak X.

$$P(S_{30} > 100) = 1 - P(S_{30} \leq 100) \approx ?$$

W celu oszacowania $P(S_{30} \leq 100)$ stosujemy CTG (ćwiczenia C10) dla S_{30}, które mówi, że S_{30} ma przybliżony rozkład normalny z parametrami $E(S_{30})$ i $\sqrt{Var(S_{30})}$. Zatem

$$P(S_{30} \leq 100) = P\left(\frac{S_{30} - 105}{\sqrt{87,5}} \leq \frac{100 + \frac{1}{2} - 105}{\sqrt{87,5}}\right) \approx \Phi\left(\frac{-4,5}{\sqrt{87,5}}\right) = dalej\ obliczenia$$

Zadanie. Niech zmienna losowa X oznacza wygraną na loterii, która przyjmuje 2 wartości: P(X = 0) = 0,9, P(X = 20) = 0,1. (a) Obliczyć wartość oczekiwaną i wariancję wygranej w jednej grze. (b) Gracz zagrał na loterii 30 razy. Obliczyć przybliżoną wartość prawdopodobieństwa, że łączna wygrana przekroczy 70.

Rozw.

(a) Rozkład prawdopodobieństwa zmiennej losowej X można przedstawić w tabeli:

x	0	20
$P(X=x)$	0,9	0,1

$$E(X) = 0 \cdot 0{,}9 + 20 \cdot 0{,}1 = 2, \;\; E(X^2) = 400 \cdot 0{,}1 = 40, \;\; Var(X) = 40 - 2^2 = 36$$

(b)

- $n = 30 -$ liczba gier
- $X_j -$ wygrana w j-tej grze, $j = 1,2, \dots ,30$
- $X_1, X_2, \dots , X_{30}$ –niezależne zmienne losowe, o rozkładach takich jak rozkład X
- $S_{30} = X_1 + X_2 + \cdots + X_{30}$ – łączna wygrana
- $E(S_{30}) = 30E(X) = 30 \cdot 2 = 60$
- $Var(S_{30}) = 30Var(X) = 30 \cdot 36 = 1080$

Z CTG rozkład prawdopodobieństwa S_{30} jest bliski rozkładowi $N(60, \sqrt{1080})$

$$P(S_{30} > 70) = 1 - P(S_{30} \leq 70)$$

$$P(S_{30} \leq 70) = P\left(\frac{S_{30} - 60}{\sqrt{1080}} \leq \frac{70 - 60}{\sqrt{1080}}\right) \approx \Phi\left(\frac{70 - 60}{\sqrt{1080}}\right) = \Phi(0{,}3043) = 0{,}6179$$

$$P(S_{30} > 70) \approx 0{,}3821$$