Metody probabilistyczne
Rozwiązania zadań

14. Statystyka 2

23.01.2020

Zadanie 1. *W celu oszacowania średniego czasu przejazdu pociągów na odcinku łączącym pewne dwa miasta pobrano próbę o liczności $n = 100$ pomiarów (tzn. zmierzono czasy przejazdu 100 wybranych losowo pociągów na tym odcinku), uzyskując średnią z próby $\overline{x}_n = 54$ (minuty). Przyjmij, że czasy przejazdy można zamodelować rozkładem normalnym ze znanym odchylenie standardowym $\sigma = 10$ (minut). Wyznacz przedział ufności dla oczekiwanej wartości czasu przejazdu na poziomie ufności $1-\alpha$ dla $\alpha = 0.02$.*

Odpowiedź: Przedział ufności dla wartości oczekiwanej rozkładu normalnego ma postać:

$$[\overline{x}_n - \Delta,\ \overline{x}_n + \Delta]\,, \qquad \text{gdzie } \Delta = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Pozostaje podstawić dane:

$$\begin{aligned} z_{1-\alpha/2} &= z_{0.99} = \Phi^{-1}(0.99) \simeq 2.33, \\ \Delta &= 2.33\frac{10}{\sqrt{100}} = 2.33, \\ [\overline{x}_n - \Delta,\ \overline{x}_n + \Delta] &= [54 - 2.33,\ 54 + 2.33] = [51.67, 56.33]\,. \end{aligned}$$

Zadanie 2*. *Skonstruuj przedział ufności dla parametru p w rozkładzie dwupunktowym $B(p)$. Użyj przybliżenia rozkładem normalnym.*

Odpowiedź: Niech $X \sim B(p)$ ma rozkład dwupunktowy. Znajdziemy przedział ufności dla parametru p. Estymator punktowy parametru p na podstawie próby $X_1, \ldots, X_n$ ma postać:

$$\widehat{p} = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Ponieważ $X_i \in \{0, 1\}$, łatwo zauważyć, że $\widehat{p}$ jest równe liczbie sukcesów podzielonej przez n, czyli empirycznej (wyznaczonej na próbie) częstości sukcesów. Ponieważ $\widehat{p}$ jest również średnią arytmetyczną z próby, mamy:

$$\begin{aligned} E(\widehat{p}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\underbrace{E(X_i)}_{p} = p, \\ D^2(\widehat{p}) &= D^2\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\underbrace{D^2(X_i)}_{p(1-p)} = \frac{p(1-p)}{n}, \end{aligned}$$

gdzie we wzorze na wariancję wykorzystaliśmy niezależność zmiennych $X_1, \ldots, X_n$. Z Centralnego Twierdzenia Granicznego wiemy, że ciąg ustandaryzowanych średnich arytmetycznych zbiega (według dystrybuant) do zmiennej o rozkładzie normalnym standardowym $N(0, 1)$, tzn. że:

$$\frac{\widehat{p} - E(\widehat{p})}{D(\widehat{p})} = \frac{\widehat{p} - p}{\sqrt{p(1-p)}}\sqrt{n} \xrightarrow{D} Z \sim N(0, 1) \tag{1}$$

Z kolei z Prawa Wielkich Liczb wiemy, że średnia arytmetyczna zbiega do wartości oczekiwanej:

$$\widehat{p} \xrightarrow{P} E(X) \;=\; p.$$

Tym samym, możemy[1] w mianowniku (1) zastąpić p przez $\widehat{p}$ i nadal otrzymamy zbieżność według dystrybuant do $Z \sim N(0,1)$:

$$\frac{\widehat{p}-p}{\sqrt{\widehat{p}(1-\widehat{p})}}\sqrt{n} \;\xrightarrow{D}\; Z \;\sim\; N(0,1).$$

Założymy, że n jest wystarczająco duże i użyjemy powyższej własności do przybliżenia rozkładem normalnym, tzn. przybliżymy:

$$\frac{\widehat{p}-p}{\sqrt{\widehat{p}(1-\widehat{p})}}\sqrt{n} \;\simeq\; Z \;\sim\; N(0,1) \tag{2}$$

Konstruujemy przedział ufności na poziomie ufności $1-\alpha$:

$$\begin{aligned}
1-\alpha &= P\left(-\Delta \leqslant \widehat{p}-p \leqslant \Delta\right) \\
&= P\Bigg(-\frac{\Delta\sqrt{n}}{\sqrt{\widehat{p}(1-\widehat{p})}} \leqslant \underbrace{\frac{\widehat{p}-p}{\sqrt{\widehat{p}(1-\widehat{p})}}\sqrt{n}}_{\simeq Z \text{ (używamy (2))}} \leqslant \frac{\Delta\sqrt{n}}{\sqrt{\widehat{p}(1-\widehat{p})}}\Bigg) \\
&= P\Bigg(-\frac{\Delta\sqrt{n}}{\sqrt{\widehat{p}(1-\widehat{p})}} \leqslant Z \leqslant \frac{\Delta\sqrt{n}}{\sqrt{\widehat{p}(1-\widehat{p})}}\Bigg) \\
&\simeq \Phi\left(\frac{\Delta\sqrt{n}}{\sqrt{\widehat{p}(1-\widehat{p})}}\right) - \Phi\left(-\frac{\Delta\sqrt{n}}{\sqrt{\widehat{p}(1-\widehat{p})}}\right) = 2\Phi\left(\frac{\Delta\sqrt{n}}{\sqrt{\widehat{p}(1-\widehat{p})}}\right) - 1,
\end{aligned}$$

gdzie w ostatniej linii wykorzystaliśmy własność dystrybuanty rozkładu normalnego standardowego $\Phi(-x) = 1-\Phi(x)$ dla dowolnych x. Otrzymaliśmy więc równanie:

$$1-\alpha \;=\; 2\Phi\left(\frac{\Delta\sqrt{n}}{\sqrt{\widehat{p}(1-\widehat{p})}}\right) - 1,$$

co po dodaniu stronami 1 i podzieleniu przez 2 daje:

$$\Phi\left(\frac{\Delta\sqrt{n}}{\sqrt{\widehat{p}(1-\widehat{p})}}\right) \;=\; 1-\frac{\alpha}{2}.$$

Ponieważ Φ jest odwracalna, otrzymujemy:

$$\frac{\Delta\sqrt{n}}{\sqrt{\widehat{p}(1-\widehat{p})}} \;=\; \underbrace{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)}_{z_{1-\alpha/2}} \qquad\Longrightarrow\qquad \Delta \;=\; z_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

Zauważmy, że $z_{1-\alpha/2}$ jest kwantylem rzędu $1-\frac{\alpha}{2}$ rozkładu normalnego standardowego. Tym samym wyznaczyliśmy przybliżony przedział ufności na poziomie ufności $1-\alpha$:

$$[\widehat{p}-\Delta, \widehat{p}+\Delta], \qquad \text{gdzie } \Delta \;=\; z_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

[1] Uwaga: stwierdzenie to jest nietrywialne. Wymaga użycia dwóch twierdzeń: (1) jeśli $Y_n \xrightarrow{P} c$ to dla dowolnej funkcji ciągłej f mamy $f(Y_n) \xrightarrow{P} f(c)$; (2) jeśli $X_n \xrightarrow{D} X$ i $Z_n \xrightarrow{P} a$ to $X_n Z_n \xrightarrow{D} aX$. Używamy powyższych twierdzeń biorąc $Y_n = \overline{X}_n = \widehat{p} \xrightarrow{P} p$, $f(x) = \frac{\sqrt{p(1-p)}}{\sqrt{x(1-x)}}$, oraz $Z_n = f(Y_n)$, tym samym $Z_n = \frac{\sqrt{p(1-p)}}{\sqrt{\widehat{p}(1-\widehat{p})}} \xrightarrow{P} f(p) = 1$. Na koniec bierzemy $X_n = \frac{\widehat{p}-p}{\sqrt{p(1-p)}}\sqrt{n} \xrightarrow{D} N(0,1)$, co po wykorzystaniu drugiego twierdzenia daje: $\frac{\widehat{p}-p}{\sqrt{\widehat{p}(1-\widehat{p})}}\sqrt{n} = X_n Z_n \xrightarrow{D} N(0,1)$.

Zadanie 3. *Aby ocenić jakość nowego algorytmu rekomendacji muzyki na pewnym portalu, wybrano losowo próbę $n = 300$ użytkowników, którym nowy algorytm zarekomendował utwór muzyczny zgodnie z ich preferencjami, i sprawdzono ilu spośród użytkowników utwór ten odsłuchało. Okazało się, że utwór odsłuchało 75 osób. Jeśli jakość algorytmu będziemy mierzyć jako prawdopodobieństwo odsłuchania rekomendowanego utworu, wyznacz przedział ufności dla jakości na poziomie ufności $1 - \alpha$, gdzie $\alpha = 0.1$.*

Odpowiedź: Problem ten można zamodelować za pomocą rozkładu dwupunktowego, gdzie $X \sim B(p)$ określa czy losowo wybrany użytkownika odsłucha ($X = 1$), czy nie odsłucha ($X = 0$) rekomendowanego utworu. Celem jest wyznaczenie przedziału ufności dla parametru p, mającego postać

$$[\widehat{p} - \Delta,\ \widehat{p} + \Delta]\,, \qquad \text{gdzie } \Delta = z_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},$$

a $\widehat{p}$ jest częstością sukcesów. Pozostaje podstawić dane:

$$\begin{aligned}
\widehat{p} &= \frac{75}{300} = \frac{1}{4},\\
z_{1-\alpha/2} &= z_{0.95} = \Phi^{-1}(0.95) \simeq 1.64,\\
\Delta &= z_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \simeq 1.64\sqrt{\frac{\frac{1}{4}\cdot\frac{3}{4}}{300}}, = 1.64\frac{1}{\sqrt{16\cdot 100}} = \frac{1.64}{40} \simeq 0.04,\\
[\widehat{p} - \Delta,\ \widehat{p} + \Delta] &= [0.25 - 0.04,\ 0.25 + 0.04] = [0.21,\ 0.29]\,.
\end{aligned}$$

Zadanie 4. *Postanowiono oszacować średnią wartość cen wynajmuj kawalerki w pewnej dzielnicy. W tym celu pobrano próbę o wielkości $n = 25$ uzyskując średnią cenę $\overline{x}_n = 1.4$ (w tys. zł) i wartość estymatora wariancji $\widehat{s}^2_* = 0.09$. Wyznacz przedział ufności dla wartości oczekiwanej ceny na poziomie ufności $1 - \alpha$ dla $\alpha = 0.05$. Załóż na potrzeby zadania, że rozkład cen jest rozkładem normalnym.*

Odpowiedź: Przedział ufności ma postać:

$$[\overline{x}_n - \Delta,\ \overline{x}_n + \Delta]\,, \qquad \text{gdzie } \Delta = t_{1-\alpha/2;n-1}\frac{\widehat{s}_*}{\sqrt{n}}.$$

Pozostaje podstawić dane:

$$\begin{aligned}
t_{1-\alpha/2;n-1} &= t_{0.975;24} = 2.064,\\
\Delta &= 2.064\frac{\sqrt{0.09}}{\sqrt{25}} = 2.064\frac{0.3}{5} = 2.064\cdot 0.06 \simeq 0.12,\\
[\overline{x}_n - \Delta,\ \overline{x}_n + \Delta] &= [1.4 - 0.12,\ 1.4 + 0.12] = [1.28, 1.52]\,.
\end{aligned}$$