Rachunek Prawdopodobieństwa i Elementy Statystyki Matematycznej

Anna Dembińska

Wydział Matematyki i Nauk Informacyjnych

Wykład 9

Projekt „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca” współfinansowany jest ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Zadanie 10 pn. „Modyfikacja programów studiów na kierunkach prowadzonych przez Wydział Matematyki i Nauk Informacyjnych”, realizowane w ramach projektu „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca”, współfinansowanego ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

9. ESTYMACJA PRZEDZIAŁOWA

Definicja. *Przedziałem ufności* dla parametru $\theta \in \Theta \subset \mathbb{R}$ na poziomie ufności $1-\alpha$, gdzie $\alpha \in (0,1)$, nazywamy przedział (θ_1, θ_2), gdzie $\theta_1 = \theta_1(X_1, X_2, \ldots, X_n)$ i $\theta_2 = \theta_2(X_1, X_2, \ldots, X_n)$ to mierzalne funkcje próby takie, że $\theta_1 \leqslant \theta_2$ i

$$P\left(\theta \in (\theta_1(X_1, X_2, \ldots, X_n), \theta_2(X_1, X_2, \ldots, X_n)\right) = 1-\alpha \quad \text{dla każdego} \quad \theta \in \Theta. \tag{1}$$

Końce przedziału ufności $\theta_1(X_1, X_2, \ldots, X_n)$ i $\theta_2(X_1, X_2, \ldots, X_n)$ to zmienne losowe. Dla niektórych realizacji próby losowej otrzymamy przedział

$$(\theta_1(x_1, x_2, \ldots, x_n), \theta_2(x_1, x_2, \ldots, x_n))$$

taki, że

$$\theta \in (\theta_1(x_1, x_2, \ldots, x_n), \theta_2(x_1, x_2, \ldots, x_n)),$$

dla innych realizacji - taki, że

$$\theta \notin (\theta_1(x_1, x_2, \ldots, x_n), \theta_2(x_1, x_2, \ldots, x_n)).$$

Warunek (1) gwarantuje nam, że dla dużej liczby realizacji procent otrzymanych przedziałów takich, że zachodzi $\theta \in (\theta_1(x_1, x_2, \ldots, x_n), \theta_2(x_1, x_2, \ldots, x_n))$ będzie w przybliżeniu równy $(1-\alpha)100\%$.

Podstawowe statystyki próbkowe i funkcje w R wyliczające te statystyki	
średnia z próby: $\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$, > mean(dane)	wariancja z próby: $s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x})^2$ > var(dane)

Oznaczenia podstawowych kwantyli i funkcje w R wyliczające te kwantyle		
dla rozkładu normalnego	dla rozkładu t-Studenta	dla rozkładu chi-kwadrat
u_α	$t_{\alpha,n}$	$\chi^2_{\alpha,n}$
> qnorm(α)	> qt(α, n)	> qchisq(α, n)

Przedziały ufności na poziomie ufności $1-\alpha$	
dla wartości średniej μ	**dla wariancji** σ^2 **(odchylenia standardowego** σ**)**
Model I. Cecha $X \sim N(\mu, \sigma^2), \mu$ - nieznane, σ - znane	
$\overline{x} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$	
Model II. Cecha $X \sim N(\mu, \sigma^2), \mu$ - nieznane, σ - nieznane	
$\overline{x} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}} < \mu < \overline{x} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}$ `> t.test(x,conf.level)$conf.int` gdzie: `x` określa wektor z danymi, `conf.level` określa poziom ufności	$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}$
Model III. Cecha X ma rozkład dowolny o skończonej wariancji, próba jest duża ($n \geqslant 100$)	
$\overline{x} - u_{1-\alpha/2}\frac{s}{\sqrt{n}} < \mu < \overline{x} + u_{1-\alpha/2}\frac{s}{\sqrt{n}}$	
Model IV. Cecha X ma rozkład dwupunktowy $P(X=1)=p, P(X=0)=q=1-p$, p - nieznane	

$\hat{p} = \frac{k}{n} = \frac{\text{ilość sukcesów}}{\text{ilość prób}}$, $\hat{q} = 1-\hat{p}$
`> binom.test(x,n,conf.level)$conf.int`
gdzie: `x` i `n` określają liczbę sukcesów i liczbę prób,
`conf.level` określa poziom ufności.
Jeśli n jest duże i $n\hat{p} > 5$ i $n\hat{q} > 5$, to można wyznaczyć przybliżony przedział asymptotyczny używając:
`> prop.test(x,n,conf.level)$conf.int`	

Wyznaczanie niezbędnej ilości pomiarów do próby do oszcowania wartości średniej μ z maksymalnym błędem d na poziomie ufności (1-α)
Model I. Cecha $X \sim N(\mu, \sigma^2), \mu$ **-** nieznane**,** σ **-** znane $n \geqslant \left(u_{1-\alpha/2}\frac{\sigma}{d}\right)^2$
Model II. Cecha $X \sim N(\mu, \sigma^2), \mu$ **-** nieznane**,** σ **-** nieznane $n \geqslant \left(t_{1-\alpha/2,n_0-1}\frac{s_0}{d}\right)^2$ gdzie n_0 i s_0 to liczność i odchylenie standardowe pobranej próby wstępnej
Model III. Cecha X ma rozkład dowolny o skończonej wariancji, próba jest duża ($n \geqslant 100$) $n \geqslant \left(u_{1-\alpha/2}\frac{s_0}{d}\right)^2$ gdzie s_0 jest odchyleniem standardowym pobranej próby wstępnej
Model IV. Cecha X ma rozkład dwupunktowy $P(X=1)=p, P(X=0)=q=1-p$, p - nieznane $n \geqslant u^2_{1-\alpha/2}\frac{p_0 q_0}{d^2}$ jeżeli znany jest szacunkowy procent p_0 (wtedy $q_0 = 1-p_0$) $n \geqslant u^2_{1-\alpha/2}\frac{1}{4d^2}$ jeżeli nie jest znany szacunkowy procent p_0

Przykład 9.1. Dział kontroli jakości w zakładach chemicznych chce oszacować średnią wagę proszku do prania sprzedawanego w pudełkach o nominalnej wadze 3 kg. Pobrano w tym celu próbkę losową 7 pudełek i otrzymano wyniki (w kg): 2.93, 2.97, 3.05, 2.91, 3.02, 2.87, 2.92. Wiadomo, że rozkład wagi pudełka do prania jest normalny.
(a) Na poziomie ufności 0,95 zbudować przedział ufności dla średniej wagi pudełka proszku do prania.
(b) Na poziomie ufności 0,95 zbudować przedział ufności dla odchylenia standardowego wagi pudełka proszku do prania.
(c) Jak liczną próbkę pudełek proszku należy pobrać, aby z maksymalnym błędem 50 g wyznaczyć przedział ufności na pozimoie ufności $0,95$ dla średniej wagi pudełka proszku do prania?

Rozwiązanie przykładu 9.1:

Oznaczmy: X-waga proszku do prania. Z treści zadania wiemy, że X ma rozkład normalny, ale parametrów tego rozkładu nie znamy. Zapisujemy to natępująco

$$X \sim \mathcal{N}(\mu, \sigma^2), \text{ gdzie } \mu \text{ i } \sigma \text{ są nieznane.}$$

Stąd widzimy, że będzie nam pasować model oznaczony jako model II w tabeli Przedziały ufności *a także model II w tabeli* Wyznaczanie niezbędnej ilości pomiarów do próby.

Wpisujemy dane do R:

```
> waga.proszku <- c(2.93, 2.97, 3.05, 2.91, 3.02, 2.87, 2.92)
```

(a) Aby wyznaczać przedział ufności dla średniej wagi pudełka patrzymy na tabelę Przedziały ufności, *dokładniej na jej lewą kolumnę (bo interesuje nas przedział ufności dla średniej, a nie dla wariancji) i na wiersz z modelem II. Widzimy, że możemy użyć funkcji* `t.test()`.

```
> t.test(x=waga.proszku,conf.level=0.95)$conf.int
```

Otrzymujemy następujący przedział: $(2.89; 3.01)$

(b) Aby wyznaczyć przedział ufności dla odchylenia standardowego wagi pudełka patrzymy na tabelę Przedziały ufności, *dokładniej na jej prawą kolumnę (bo interesuje nas przedział ufności nie dla średniej ale dla odchylenia standardowego) i na wiersz z modelem II. Widzimy tam następujący wzór*

$$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}.$$

Musimy jeszcze wszystkie jego strony spierwiastkować, aby otrzymać wzór na przedział ufności dla odchylenia standardowego (czyli dla σ):

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}}.$$

Rachunki przeprowadzamy wykorzystując R. Policzmy najpierw licznik wyrażeń znajdujących się pod pierwiastkiem:
> `licznik <- (7-1)*var(waga.proszku)`
a następnie mianowniki tych wyrażeń (poziom ufności wynosi $1-\alpha = 0,95$, więc $\alpha = 0,05$):
> `mianownik.lewy <- qchisq(1-0.05/2,7-1)`
> `mianownik.prawy <- qchisq(0.05/2,7-1)`
Ostatecznie otrzymujemy przedział o następujących końcach
> `sqrt(licznik/mianownik.lewy)`
> `sqrt(licznik/mianownik.prawy)`
czyli $(0.04121; 0.1408)$.

(c) Teraz patrzymy na tabelę Wyznaczanie niezbędnej ilości pomiarów do próby, *na model II. Szukamy n a mamy dane $d = 50g = 0,05kg$ i $1-\alpha = 0,95$ (zatem $\alpha = 0,05$). Korzystamy ze wzoru*

$$n \geqslant \left(t_{1-\alpha/2,n_0-1}\frac{s_0}{d}\right)^2,$$

gdzie liczność próby wstępnej to $n_0 = 7$ (to ta próba podana w treści zadania) a jej odchylenie standardowe s_0 policzymy używając funkcji `sd(waga.proszku)`.
Prawa strona powyższego wzoru jest równa
> `(qt(1-0.05/2,7-1)*sd(waga.proszku)/0.05)^2`
czyli 9.79649. *Zatem potrzebujemy próbę składającą się z co najmniej* 10 *pomiarów wagi proszku do prania.*

Przykład 9.2. 88 spośród 400 osób, które w 2019 roku nabyły nowy samochód sportowy, stwierdziło, że jakość samochodu jest lepsza niż oczekiwali.
(a) Na poziomie ufności 0,95 zbudować przedział ufności dla badanej frakcji.
(b) Jak liczną próbę należałoby pobrać, aby z maksymalnym błędem ±2% móc oszacować badaną frakcję na poziomie ufności 0,95?

Rozwiązanie przykładu 9.2:

Mamy tutaj do czynienia z rozkładem dwupunktowym: osoba, która w 2019 roku nabyła nowy samochód sportowy może stwierdzić, że jakość zakupionego samochodu jest lepsza niż oczekiwała lub tego nie stwierdzić. Oznaczmy:

$$X = \begin{cases} 1 & \textit{jeśli osoba stwierdzi, że jakość zakupionego samochodu jest lepsza niż tego oczekiwała,} \\ 0 & \textit{jeśli osoba nie stwierdzi, że jakość zakupionego samochodu jest lepsza niż tego oczekiwała,} \end{cases}$$

zaś p niech będzie prawdopodobieństwem trafienia na osobę, która uzna, że jakość zakupionego samochodu jest lepsza niż tego oczekiwała, tzn. $p = P(X = 1)$. Inaczej mówi"sc p jest frakcją osób, które uznają, że jakość zakupionego samochodu jest lepsza niż tego oczekiwali.

(a) Patrzymy na tabelę Przedziały ufności, *na wiersz z modelem IV. Zauważamy, że n jest na tyle duże, że możemy użyć* `prop.test()`. *Rzeczywiście*

- *liczba sukcesów* $n\hat{p} = 88 > 5$,
- *liczba porażek* $n\hat{q} = 400 - 88 > 5$.

Zatem piszemy
> `prop.test(x=88,n=400,conf.level=0.95)$conf.int`
i otrzymujemy szukany przedział ufności: $(0.1810175; 0.2644723)$.

(a) Patrzymy na tabelę Wyznaczanie niezbędnej ilości pomiarów do próby, *na wiersz z modelem IV. Szacunkowy procent* p_0 *możemy wyznaczyć na podstawie danych z zadania, więc użyjemy wzoru pierwszego z tam podanych:*

$$n \geqslant u^2_{1-\alpha/2}\frac{p_0 q_0}{d^2},$$

gdzie p_0 *to szacunkowy procent a* $q_0 = 1 - p_0$. *Oczywiście* $p_0 = 88/400$. *Ponato* $1 - \alpha$ *to poziom ufności, zatem* $1 - \alpha = 0,95$, *co daje* $\alpha = 0,05$. *Maksymalny błąd to* $d = 2\% = 0,02$. *Reasumując*
```
> p0 <- 88/400
> q0 <- 1-p0
> d <- 0.02
```
i prawa strona powyższej nierówności wynosi
```
> qnorm(1-0.05/2)^2*p0*q0/(d^2)
```
czyli 1647.986. *Potrzebna jest próba o liczności co najmniej* 1648.