Wykład V: Wnioskowanie statystyczne - estymacja przedziałowa

Alicja Janic

Politechnika Wrocławska *alicja.janic@pwr.edu.pl*

10 grudzień 2020

Estymacja przedziałowa

Idea

Zadaniem estymacji przedziałowej jest skonstruowanie na podstawie próby losowej przedziału, o którym można z dużą dozą przekonania powiedzieć, iż zawiera prawdziwą wartość szacowanego parametru. Konstrukcja przedziału jest oczywiście równoznaczna z podaniem jego dwóch końców. Jeżeli próba losowa nie została jeszcze zaobserwowana, jest to przedział o losowych końcach będących funkcjami tej próby. Z kolei, zaobserwowana wartość estymatora przedziałowego, powstała na podstawie zaobserwowanej realizacji próby losowej, jest wyznaczona przez dwie liczby (dwa końce przedziału)

Konstrukcja przedziałów ufności

Zadanie

Na podstawie próby losowej $X_1, X_2 \ldots, X_n$ skonstruować przedział

$$[\underline{g}, \overline{g}]$$

w którym końce $\underline{g} = \underline{g}(X_1, X_2 \ldots, X_n)$ oraz $\overline{g} = \overline{g}(X_1, X_2 \ldots, X_n)$ będą statystykami (zmiennymi losowymi) i który z zadawalającym prawdopodobieństwem $1 - \alpha$, nazywanym poziomem ufności, zawiera prawdziwą wartość szacowanego parametru θ:

$$P_\theta\left(\underline{g}(X_1, X_2 \ldots, X_n) \leqslant \theta \leqslant \overline{g}(X_1, X_2 \ldots, X_n)\right) \geqslant 1 - \alpha$$
$$= 1 - \alpha$$

Najczęściej próba pochodzi z rozkładu normalnego. A dla rozkładów ciągłych mamy równość

Konstrukcja przedziałów ufności

Konstrukcja

1. Szukamy **funkcji centralnej** - funkcja zależna od próby losowej $X_1, X_2..., X_n$ i estymowanego parametru θ, natomiast jej znany rozkład (może być asymptotyczny) nie zależy od wartości nieznanego parametru θ
2. Jeżeli $Q = Q(X_1, X_2..., X_n; \theta)$ jest fukcją centralną, to szukamy przedziału spełniającego warunek

$$P_\theta\left(a \leqslant Q \leqslant b\right) = 1 - \alpha$$

Dodatkowo szukamy przedziałów symetrycznych, to znaczy takich, że

$$P_\theta\left(Q \leqslant a\right) = \frac{\alpha}{2} \quad \text{oraz} \quad P_\theta\left(Q \geqslant b\right) = \frac{\alpha}{2}$$

Przypadek rozkładu normalnego o znanym odchyleniu standardowym

Przypadek 1

Rozważmy próbę losową $X_1, X_2 ..., X_n$ z rozkładu normalnego $\mathcal{N}(m, \sigma)$ ze znanym odchyleniem standardowym σ. Zadanie polega na wyznaczeniu przedziału ufności dla nieznanej wartości średniej m. Wiadomo, że średnia w próbie $\overline{X}$ ma rozkład normalny $\mathcal{N}\left(m, \frac{\sigma}{\sqrt{n}}\right)$. Stąd, zmienna losowa

$$Z = \frac{\overline{X} - m}{\sigma / \sqrt{n}}$$

ma standardowy rozkład normalny $\mathcal{N}(0, 1)$ i jest funkcją centralną

Przypadek rozkładu normalnego o znanym odchyleniu standardowym

Przypadek 1

Wyznaczamy przedział, do którego wartości zmiennej losowej Z należą z prawdopodobieństwem $1-\alpha$, gdzie α jest zadaną liczbą z przedziału $(0,1)$. Mianowicie

$$P(z_{\alpha/2} \leqslant Z \leqslant z_{1-\alpha/2}) = 1-\alpha$$

gdzie z_p jest kwantylem rzędu p standardowego rozkładu normalnego. Ze względu na symetrię gęstości standardowego rozkładu normalnego mamy przy tym

$$z_{\alpha/2} = -z_{1-\alpha/2}$$

Przypadek rozkładu normalnego o znanym odchyleniu standardowym

Przypadek 1

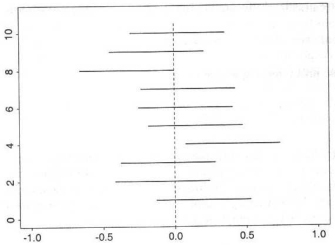
Po dokonaniu prostych przekształceń otrzymujemy

$$P\left(-z_{1-\alpha/2} \leqslant \frac{\overline{X}-m}{\sigma/\sqrt{n}} \leqslant z_{1-\alpha/2}\right) =$$
$$P\left(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leqslant m \leqslant \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

W ten sposób otrzymaliśmy przedział losowy, zawierający z zadanym prawdopodobieństwem $1-\alpha$ nieznaną wartość średnią m. Zaobserwowawszy próbę losową $X_1, X_2..., X_n$, czyli mając realizację tej próby $x_1, x_2..., x_n$, możemy obliczyć realizację średniej w próbie $\overline{x}$ i podać przedział ufności dla m na poziomie ufności $1-\alpha$

$$\left[\overline{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Przypadek rozkładu normalnego o znanym odchyleniu standardowym



Rys. 3.6. Przedziały ufności dla μ dla 10 prostych prób losowych o liczności 25 z rozkładu $N(0, 1)$

Oczekujemy, że zdarzenie

$$m \in \left[\overline{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

zachodzi z częstością $1 - \alpha$

Przypadek rozkładu normalnego o znanym odchyleniu standardowym

Błąd średniej próbkowej

Ponieważ

$$P\left(|\overline{X} - m| \leqslant z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Skąd wynika, że błąd średniej próbkowej $\overline{x}$ nie przekracza na poziomie ufności $1-\alpha$ wartości

$$z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Długość przedziału ufności

Długość przedziału ufności jest równa podwojonej wartości podanego błędu

Przypadek rozkładu normalnego o znanym odchyleniu standardowym

Długość przedziału ufności

Długość przedziału ufności jest tym mniejsza im większa jest liczność próby n. Wynika stąd, że dobierając odpowiednio dużą liczność próby, możemy uzyskać przedział ufności o dowolnie małej, ustalonej długości. Jeżeli chcemy by przedział ufności nie był dłuższy od zadanej wartości np. $2d$

$$2z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leqslant 2d$$

to liczność próby musi spełniać warunek

$$n \geqslant \left(\frac{z_{1-\alpha/2}\sigma}{d}\right)^2$$

Przykłady

Przykład 1

Zmierzono czas życia, czyli czas działania, próby losowej 16 żarówek o ustalonej mocy. Średni czas życia w próbie wyniósł 3000 godzin, natomiast odchylenie standardowe całej populacji żarówek wynosi 20 godzin. Przy założeniu, że czas życia żarówki jest zmienną losową o rozkładzie normalnym, podać przedział ufności dla wartości średniej tego rozkładu na poziomie ufności 0,9

Rozwiązanie

Przypadek rozkładu normalnego o nieznanym odchyleniu standardowym

Wariancja w prostej próbie losowej

Wariancją w prostej próbie losowej $X_1, ..., X_n$ o liczności n nazywamy statystykę

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

Stwierdzenie

Niech S^2 będzie wariancją w prostej próbie losowej o liczności n z rozkładu o wartości średniej m i wariancji σ^2. Wówczas

$$E\left(S^2\right) = \sigma^2 \qquad \text{oraz} \qquad Var\left(S^2\right) = \frac{2\sigma^4}{n-1}$$

Przypadek rozkładu normalnego o nieznanym odchyleniu standardowym

Przypadek 2

Najczęściej odchylenie standardowe rozkładu populacji nie jest znane. Nasuwa się zatem myśl zastąpienia zmiennej Z zmienną losową

$$T = \frac{\overline{X} - m}{S/\sqrt{n}}$$

Rozkład zmiennej losowej T nie zależy od nieznanego parametru m i jest znany (**funkcja centralna**). Można mianowicie udowodnić, że jest to tzw. rozkład t (zwany też rozkładem Studenta) z $n-1$ stopniami swobody oznaczany symbolem t_{n-1}

Przypadek rozkładu normalnego o nieznanym odchyleniu standardowym

Przypadek 2

Mając zmienną losową T i jej rozkład t_{n-1} możemy przedział ufności dla m zbudować w sposób zupełnie analogiczny do poprzedniego przypadku. Przedział ufności na poziomie $1-\alpha$ przyjmuje postać

$$\left[\overline{x} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}, \overline{x} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}\right],$$

gdzie $t_{1-\alpha/2,n-1}$ jest kwantylem rzędu $1-\alpha/2$ rozkładu t_{n-1}

$$P(T \leqslant t_{1-\alpha/2,n-1}) = 1-\alpha/2$$

Wartości kwantyli odczytujemy z tablic statystycznych dla rozkładu t Studenta

Przypadek rozkładu normalnego o nieznanym odchyleniu standardowym

Przybliżenie

Ponieważ rozproszenie estmatora S^2 maleje wraz ze wzrostem liczności próby n, estymator ten dąży w pewnym probabilistycznym sensie do prawdziwej wartości wariancji rozkładu σ^2. Stąd, zmienne losowe Z i T stają się przy rosnacym n nierozróżnialne, zaś gęstość rozkładu t_{n-1} dąży do gęstości rozkładu $\mathcal{N}(0,1)$, czyli kwantyle rozkładu t dążą do kwantyli tego samego rzędu rozkładu $\mathcal{N}(0,1)$. Zatem, dla dostatecznie dużej liczności próby, można w przypadku nieznajomości odchylenia standardowego traktować przedział

$$\left[\overline{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \overline{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\right],$$

jako dobre przybliżenie przedziału ufności na poziomie ufności $1-\alpha$ dla wartości średniej m. W praktyce $n \geqslant 30$

Przykłady

Przykład 1 raz jeszcze

Zmierzono czas życia, czyli czas działania, próby losowej 16 żarówek o ustalonej mocy. Średni czas życia w próbie wyniósł 3000 godzin, natomiast odchylenie standardowe w próbie wyniosło 20 godzin. Przy założeniu, że czas życia żarówki jest zmienną losową o rozkładzie normalnym, podać przedział ufności dla wartości średniej tego rozkładu na poziomie ufności 0,98

Rozwiązanie

Porównanie dwóch wartości średnich

Niezależne próby losowe

Możemy mieć do czynienia z dwiema niezależnymi prostymi próbami losowymi (o niekoniecznie tej samej liczności), $X_1, X_2, \dots, X_{n_1}$ oraz $Y_1, Y_2, \dots, Y_{n_2}$, z wartościami średnimi, odpowiednio, m_1 i m_2. Nadal zakładamy, że próby losowe pochodzą z rozkładów normalnych. Dodatkowo załóżmy narazie, że są znane odchylenia standardowe obydwu rozkładów σ_1 i σ_2. Mamy zatem do czynienia z dwiema próbami, z których pierwsza pochodzi z rozkładu $\mathcal{N}(m_1, \sigma_1)$, natomiast druga z rozkładu $\mathcal{N}(m_2, \sigma_2)$

Porównanie dwóch wartości średnich

Przypadek 1 - znane odchylenia standardowe

Niech $\overline{X}$ i $\overline{Y}$ oznaczają, odpowiednio, średnią w pierwszej i drugiej próbie losowej. Łatwo wykazać, że statystyka

$$Z = \frac{(\overline{X} - \overline{Y}) - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

ma standardowy rozkład normalny (**funkcja centralna**).
Dwustronny przedział ufności dla różnicy $m_1 - m_2$ na poziomie ufności $1 - \alpha$ ma postać

$$\left[(\overline{x} - \overline{y}) - z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\overline{x} - \overline{y}) + z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

Porównanie dwóch wartości średnich

Przypadek 2 - nieznane odchylenia standardowe

Przypadek nieznanych odchyleń standardowych σ_1 i σ_2 rozważymy jedynie przy założeniu równości obydwu odchyleń standardowych $\sigma_1 = \sigma_2 = \sigma$ (przypadek nierównych odchyleń standardowych jest bardziej złożony, nieznany jest bowiem wówczas dokładny rozkład statystyki testowej). Wariancja różnicy $\overline{X} - \overline{Y}$ jest równa

$$\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

Porównanie dwóch wartości średnich

Przypadek 2 - nieznane odchylenia standardowe

Można wykazać, że oparta na obydwu próbach statystyka

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

gdzie S_i^2, $i = 1, 2$, jest wariancją w i-tej próbie, jest nieobciążonym estymatorem wariancji σ^2 tzn. $E(S_p^2) = \sigma^2$. Co więcej statystyka

$$T = \frac{(\overline{X} - \overline{Y}) - (m_1 - m_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

ma rozkład t Studenta $n_1 + n_2 - 2$ stopniami swobody

Porównanie dwóch wartości średnich

Przypadek 2 - nieznane odchylenia standardowe

Ograniczając się do dwustronnego przedziału ufności dla różnicy $m_1 - m_2$ na poziomie ufności $1 - \alpha$, otrzymujemy przedział

$$\left[(\overline{x} - \overline{y}) - t_{1-\alpha/2, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \right.$$
$$\left. (\overline{x} - \overline{y}) + t_{1-\alpha/2, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right],$$

gdzie $t_{1-\alpha/2, n_1+n_2-2}$ jest kwantylem rzędu $1 - \alpha/2$ rozkładu t Studenta z $n_1 + n_2 - 2$ stopniami swobody

Porównanie dwóch wartości średnich

Pary obserwacji

Jakościowo inna sytuacja, gdy mamy do czynienia z **parami obserwacji**

$$(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n),$$

gdzie pary mają taki sam dwuwymiarowy rozkład normalny i są wzajemnie niezależne, ale zmienne w parze mogą być zależne. Na przykład, gdy pacjentowi z nadciśnieniem tętniczym badamy ciśnienie skurczowe przed zastosowaniem terapii i po jej zastosowaniu. Każda para obserwacji odpowiada wówczas konkretnemu pacjentowi i zmienne w parze nie są oczywiście niezależne

Porównanie dwóch wartości średnich

Pary obserwacji

Zauważmy, że nawet jeżeli znamy wariancje zmiennych losowych X_i oraz Y_i, dla $i = 1, 2, ..., n$, to ze względu na zależność między każdą taką parą zmiennych nie możemy na tej podstawie podać wariancji różnic $D_i = X_i - Y_i$. Możemy jednak podać oczywisty estymator tej wariancji, a mianowicie wariancję w próbie

$$S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2,$$

gdzie $\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$

Porównanie dwóch wartości średnich

Pary obserwacji

Zauważmy dalej, że różnice D_i tworzą próbę niezależnych zmiennych losowych o rozkładzie normalnym z nieznaną wartością średnią $m_D = m_1 - m_2$, i że statystyka:

$$T = \frac{\overline{D} - m_D}{S_D/\sqrt{n}}$$

ma rozkład t Studenta z $n-1$ stopniami swobody. W ten sposób zadanie konstrukcji przedziału ufności dla różnicy wartości średnich, gdy obserwacje występują w parach, sprowadza się do zadania wcześniej już omówionego. Na przykład dwustronny przedział ufności dla m_D na poziomie $1-\alpha$ ma postać

$$\left[\overline{d} - t_{1-\alpha/2,n-1}\frac{s_D}{\sqrt{n}}, \overline{d} + t_{1-\alpha/2,n-1}\frac{s_D}{\sqrt{n}}\right]$$

Przykłady

Przykład 2

Siłownia reklamuje program odchudzający twierdząc, że ćwiczący zmiejsza swój obwód w talii w ciągu 5 dni ćwiczeń średnio o 2 cm. Zmierzono obwody w talii 6 w mężczyzn biorących udział w programie przed rozpoczęciem ćwiczeń oraz po upływie 5 dni. W przypadku pierwszego mężczyzny uzyskano 95,5 cm przed i 93,9 cm po 5-dniowym cyklu ćwiczeń. W przypadku drugiego uzyskano 98,7 i 97,4 cm. W przypadku kolejnych uczestników badania uzyskano odpowiednio przed i po cyklu zajęć: 90,4 i 91,7 cm; 115,9 i 112,8 cm; 104,0 i 101,3 cm; 85,6 i 84,0 cm. Założyć normalny rozkład różnic obwodów przed i po 5 dniach ćwiczeń, znaleźć przedział ufności dla średniego zmniejszenia obwodu na poziomie ufności 0,95. Czy otrzymany wynik świadczy, że twierdzenie siłowni jest uzasadnione?

Rozwiązanie

Przedziały ufności dla wariancji rozkładu normalnego

Przedziały ufności dla wariancji

Punktowym estymatorem wariancji, w szczególności wariancji rozkładu normalnego, jest oczywiście wariancja w próbie

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

W przypadku, gdy niezależne zmienne losowe X_i pochodzą ze standardowego rozkładu normalnego, zmienna losowa

$$\sum_{i=1}^{n} X_i^2$$

ma tzw. rozkład χ^2 o n stopniach swobody. Zatem zmienna losowa

$$\sum_{i=1}^{n}\left(\frac{X_i - m}{\sigma}\right)^2,$$

gdzie X_i, $i = 1, 2, ..., n$ są niezależnymi zmiennymi losowymi o rozkładzie $\mathcal{N}(m, \sigma)$, ma także rozkład χ^2 z n stopniami swobody

Przedziały ufności dla wariancji rozkładu normalnego

Przedziały ufności dla wariancji

Naturalnym odpowiednikiem jest zmienna losowa

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 = \frac{(n-1)S^2}{\sigma^2},$$

która ma rozkład χ^2 z $n-1$ stopniami swobody. Rozkład zmiennej χ^2 nie zależy od nieznanych parametrów i jest znany, a zatem może być funkcją centralną

Przedziały ufności dla wariancji rozkładu normalnego

Przedziały ufności dla wariancji

Zatem

$$P\left(\chi^2_{\alpha/2,n-1} \leqslant \frac{(n-1)S^2}{\sigma^2} \leqslant \chi^2_{1-\alpha/2,n-1}\right) = P\left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} \leqslant \sigma^2 \leqslant \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}\right) = 1-\alpha \quad ,$$

gdzie α jest ustaloną liczbą z przedziału $(0,1)$ oraz $\chi^2_{\alpha/2,n-1}$ i $\chi^2_{1-\alpha/2,n-1}$ są kwantylami odpowiednio rzędu $\alpha/2$ i $1-\alpha/2$ z rozkładu χ^2 z $n-1$ stopniami swobody

$$P(\chi^2 \leqslant \chi^2_{\alpha/2,n-1}) = \alpha/2 \quad \text{oraz} \quad P(\chi^2 \leqslant \chi^2_{1-\alpha/2,n-1}) = 1-\alpha/2$$

Przedziały ufności dla wariancji rozkładu normalnego

Przedziały ufności dla wariancji

Szukany przedział ufności na poziomie ufności $1-\alpha$ dla wariancji rozkładu normalnego ma postać

$$\left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}\right]$$

Na tej samej podstawie otrzymujemy przedział ufności na poziomie ufności $1-\alpha$ dla odchylenia standardowego rozkładu normalnego

$$\left[\sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}}\right]$$

Rozwiązanie

Przedziały ufności dla proporcji

Przedziały ufności dla proporcji

Punktowym estymatorem proporcji p jest oczywiście częstość $\hat{p}$. Częstość jest nieobciążonym estymatorem proporcji Opierając się na częstości $\hat{p}$ skonstruujemy przedziały ufności dla proporcji p. Zakładamy, że próba losowa niezależnych zmiennych ma rozkład dwupunktowy $\quad P(X_i = 1) = 1 - P(X_i = 0) = p, \quad i = 1, 2, ..., n$ jest dostatecznie liczna, by móc skorzystać z przybliżenia rozkładu statystyki

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

standardowym rozkładem normalnym

Przedziały ufności dla proporcji

Przedziały ufności dla proporcji

Zatem

$$P\left(-z_{1-\alpha/2} \leqslant \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leqslant z_{1-\alpha/2}\right) \approx 1-\alpha,$$

gdzie $z_{1-\alpha/2}$ oznacza kwantyl rzędu $1-\alpha/2$ z rozkładu $\mathcal{N}(0,1)$

Przedziały ufności dla proporcji

Przedziały ufności dla proporcji

Stąd dla dostatecznie dużej liczności próby losowej oraz gdy $n\hat{p} \geqslant 5$ i $n(1-\hat{p}) \geqslant 5$, przybliżony dwustronny przedział ufności na poziomie ufności $1-\alpha$ dla proporcji p ma postać

$$\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Obliczanie minimalnej liczności próby, przy której długość tego przedziału nie przekracza zadanej wielkości l:

$$n \geqslant \frac{4z^2_{1-\alpha/2}\hat{p}(1-\hat{p})}{l^2}$$

Przykłady

Przykład 3

Jedna z agencji badających opinię publiczną ogłosiła w czerwcu 2000 r., że przebadała reprezentatywną próbę 1000 dorosłych obywateli polskich, z których 57% poparło starania ich państwa o wejście do Unii Europejskiej. Uznając, że mamy do czynienia z rozkładem dwupunktowym (popieranie lub nie starań o wejście do UE) możemy skonstruować 95% przedział ufności dla proporcji obywateli popierających wejście Polski do UE

Rozwiązanie