

Raport 4

Magdalena Potok

2023-12-05

Zadanie 1

Niech X_1, X_2, \dots, X_{n_1} będą niezależnymi zmiennymi losowymi z rozkładu $N(\mu_1, \sigma_1^2)$ oraz niech Y_1, Y_2, \dots, Y_{n_2} będą niezależnymi zmiennymi losowymi z rozkładu $N(\mu_2, \sigma_2^2)$. Wtedy przedział ufności dla różnicy średnich tych rozkładów, przy założeniu, że znamy ich wariancję, wygląda następująco:

$$\bar{X} - \bar{Y} \pm z_{(1-\frac{\alpha}{2})}^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

gdzie $z_{(1-\frac{\alpha}{2})}^*$ jest wartością krytyczną z rozkładu normalnego na poziomie istotności α , \bar{X} i \bar{Y} to średnia próbkowa zmiennych niezależnych.

Zadanie 2

Wygeneruję n_1 oraz n_2 obserwacji z rozkładu:

(a) normalnego z parametrem przesunięcia μ_1 i μ_2 oraz skali σ_1 i σ_2

(i) $\mu_1 = 0, \sigma_1 = 1, \mu_2 = 0, \sigma_2 = 1$,

(ii) $\mu_1 = 0, \sigma_1 = 1, \mu_2 = 1, \sigma_2 = 2$.

(b) logistycznego z parametrem przesunięcia μ_1 i μ_2 oraz skali σ_1 i σ_2

(i) $\mu_1 = 0, \sigma_1 = 1, \mu_2 = 0, \sigma_2 = 1$,

(ii) $\mu_1 = 0, \sigma_1 = 1, \mu_2 = 1, \sigma_2 = 2$.

Na podstawie powyższych rozkładów wyznaczę przedział ufności z zadania 1. dla parametru $\mu_1 - \mu_2$ na poziomie ufności $1 - \alpha = 0.95$ oraz jego długość. Doświadczenie powtórzę 10 000 razy, oszacuję prawdopodobieństwo pokrycia nieznanego parametru przez przedział ufności oraz jego długość.

Table 1: Wyniki testów dla różnych przypadków

	śr. długość	Prawdopodobieństwo	Mu1 - Mu2	Lewy PU	Prawy PU
Norm (i)	0.784	0.951	0	-0.393	0.391
Norm (ii)	1.240	0.950	-1	-1.621	-0.382
Logis (i)	0.784	0.714	0	-0.386	0.398
Logis(ii)	1.240	0.721	-1	-1.624	-0.385

Można zauważyć dla rozkładu normalnego, że prawdopodobieństwo nie zależy od przyjętych n, μ, σ i oscyluje dla obu przypadków około 0.95, czyli teoretycznej wartości prawdopodobieństwa. Można za to zauważyć, że dla Norm (ii) przedział ufności jest dłuższy, wynika to z zwiększonej wariancji dla obserwacji z drugiego rozkładu, obserwacje są bardziej rozproszone, co wpływa na jakość estymacji.

Dla rozkładu logistycznego również prawdopodobieństwa są podobne dla obu podpunktów i wynoszą około 0.72, czyli można powiedzieć, że zmienne z polecenia parametry nie mają wpływu na pokrycie prawdziwej wartości z przedziałami ufności. Wpływ ma za to zmienienie σ_2 na średnią długość naszego przedziału, ponieważ ze zwiększeniem się tej wartości zwiększyła też się średnia długość. Wyniki i wnioski są zatem podobne dla obu tych rozkładów.

Zadanie 5

Niech X_1, X_2, \dots, X_{n_1} będą niezależnymi zmiennymi losowymi z rozkładu $N(\mu_1, \sigma_1^2)$ oraz niech Y_1, Y_2, \dots, Y_{n_2} będą niezależnymi zmiennymi losowymi z rozkładu $N(\mu_2, \sigma_2^2)$. Wtedy przedział ufności dla różnicy średnich tych rozkładów, przy założeniu, że nie znamy ich wariancji, wygląda następująco:

$$\bar{X} - \bar{Y} \pm t_{(1-\frac{\alpha}{2}, df)}^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

gdzie \bar{X} i \bar{Y} to średnia próbkowa zmiennych niezależnych, s_1^2 i s_2^2 to odpowiednio estymatory wariancji próbek 1 i 2 populacji, $t_{(1-\frac{\alpha}{2}, df)}^*$ to wartość krytyczna z rozkładu t-Studenta dla zadanej liczby swobody (tutaj $df = n_1 + n_2 - 2$) i poziomu istotności.

Zadanie 6

Powtórzę eksperyment numeryczny z zadania 2., na jego podstawie oszacuję prawdopodobieństwo pokrycia nieznanego parametru przez przedział ufności z zadania 5. na poziomie ufności 0.95 oraz jego długość

Table 2: Wyniki testów dla różnych przypadków

	śr. długość	Prawdopodobieństwo	Mu1 - Mu2	Lewy PU	Prawy PU
Norm (i)	0.792	0.947	0	-0.397	0.395
Norm (ii)	1.252	0.951	-1	-1.627	-0.375
Logis (i)	0.794	0.735	0	-0.387	0.407
Logis(ii)	1.255	0.727	-1	-1.621	-0.366

Dla rozkładu normalnego wyniki są bardzo podobne, jak w przypadku, gdy znaleźliśmy wariancję. Średnia długość przedziałów jest nieco dłuższa, co wynika z faktu, że estymujemy wariancję. Prawdopodobieństwo jest minimalnie mniejsze, ale wciąż utrzymuje się na oczekiwanym poziomie, czyli 0.95.

Dla rozkładu logistycznego delikatnie większa różnica jeżeli chodzi o zwiększenie średniej długości, ale wyniki również są bardzo podobne, jak w zadaniu 2. Prawdopodobieństwo jest nieco wyższe, plasuje się w obu przypadkach bliżej okolicy 0.73.

W obu przypadkach zwiększenie parametru σ wpłynęło na zwiększenie długości przedziału, tak jak w zadaniu 2., ale nie wpływa na prawdopodobieństwo.

Zadanie 9

Niech X_1, X_2, \dots, X_{n_1} będą niezależnymi zmiennymi losowymi z rozkładu $N(\mu_1, \sigma_1^2)$ oraz niech Y_1, Y_2, \dots, Y_{n_2} będą niezależnymi zmiennymi losowymi z rozkładu $N(\mu_2, \sigma_2^2)$. Wtedy przedział ufności dla ilorazu dwóch wariancji o nieznanach średniach na poziomie ufności $1 - \alpha$ ma postać:

$$\left[\frac{s_1^2}{b \cdot s_2^2} ; \frac{s_1^2}{a \cdot s_2^2} \right]$$

gdzie $s_1^2 = \frac{\sum_{i=1}^{n_1} (X_i - \bar{X})^2}{n_1 - 1}$, $s_2^2 = \frac{\sum_{i=1}^{n_2} (Y_i - \bar{Y})^2}{n_2 - 1}$, $a = F^{-1}(\frac{\alpha}{2}, n_1 - 1, n_2 - 1)$ oraz $b = F^{-1}(1 - \frac{\alpha}{2}, n_1 - 1, n_2 - 1)$.

Zadanie 10

Powtórzę eksperyment numeryczny z zadania 2. na jego podstawie oszacuję prawdopodobieństwo pokrycia nieznanego parametru przez przedział ufności z zadania 9. na poziomie ufności 0.95 oraz jego długość.

Table 3: Wyniki testów dla różnych przypadków

	śr. długość	Prawdopodobieństwo	var1/var2	Lewy PU	Prawy PU
Norm (i)	1.247	0.953	1.00	0.592	1.839
Norm (ii)	0.312	0.953	0.25	0.148	0.461
Logis (i)	1.273	0.889	1.00	0.605	1.878
Logis(ii)	0.318	0.885	0.25	0.151	0.469

Dla rozkładu normalnego zastosowanie przedziału ufności dało oczekiwany efekt w obu przypadkach, ponieważ prawdopodobieństwo wynosi około 0.95. To co można ciekawego zauważyć, to fakt, że dla rozkładu o różnych wariancjach średnia długość wyszła o wiele mniejsza. Oznacza to, że istnieje tendencja do tego, że przedział ufności dla rozkładów o mniejszym stosunku wariancji ($0.25 < 1$) będą węższe niż te, dla rozkładów o większym stosunku. Dla rozkładu logistycznego również ta reguła działa. Średnia długość przedziałów wyszła 4 razy mniejsza, to tyle, ile razy stosunek wariancji jest mniejszy. W przypadku logistycznego rozkładu prawdopodobieństwo wynosi nieco mniej, bo około 0.89, dla obu przypadków (i) i (ii).

Zadanie 12

Lemat (Metoda delta)

Jeżeli dla ciągu zmiennych losowych T_n mamy $\sqrt{n}(T_n - \mu) \xrightarrow{d} N(0, \sigma^2)$ przy $n \rightarrow \infty$ i $h: \mathbb{R} \rightarrow \mathbb{R}$ jest funkcją różniczkowalną w punkcie μ , to

$$\sqrt{n}(h(T_n) - h(\mu)) \xrightarrow{d} N(0, \sigma^2 (h'(\mu))^2)$$

Metoda δ polega na wnioskowaniu o rozkładach prawdopodobieństwa próbek statystycznych, gdy liczba obserwacji dąży do nieskończoności. Można to zrobić poprzez analizę zbieżności różnych rodzajów dystrybucji prawdopodobieństwa do innych rozkładów obserwując co się dzieje, gdy wielkość próby rośnie. Jednym z przykładów wykorzystania tej metody jest teoria asymptotycznej zbieżności, pozwala ona na wnioskowanie o własnościach estymatorów statystycznych w granicy, gdy liczba obserwacji wzrasta do nieskończoności. Główne zastosowanie metody δ w konstrukcji przedziałów ufności jest uzasadnione centralnym twierdzeniem granicznym, które mówi nam, że dla dużego zestawu zmiennych losowych (iid) dowolnego rozkładu jest on przybliżony rozkładem normalnym. Korzystając z tego twierdzenia możemy opracowywać przedział ufności dla różnych parametrów populacji, takich jak w poprzednich zadaniach.