

Raport trzeciej listy z laboratorium Statystyki

Kamil Zdancewicz

November 25, 2025

Zadanie 1

Cel

Celem zadania było przedziału ufności dla **średniej** dla współczynnika ufności $1 - \alpha$ przy **znanej wariancji**.

Obliczenia

Oznaczenia:

θ – średnia, σ_T^2 – znana wariancja $\sqrt{n}T$

Będziemy estymować średnią za pomocą pewnej statystyki $T = T(X_1, \dots, X_n)$, gdzie X_1, \dots, X_n jest próbą z rozkładu o pewnej gęstości $f(x; \theta)$, $\theta \in \Theta$ i skończonej wariancji. Oczywiście nie znamy prawdziwej wartości θ , oznaczmy ją przez θ_0 .

Założmy, że T jest esymtorem spełniającym:

$$\sqrt{n}(T - \theta_0) \xrightarrow{d} N(0, \sigma_T^2)$$

Przy definicji $Z = \sqrt{n}(T - \theta_0)/\sigma_T$ mamy asymptotycznie $Z \sim N(0, 1)$ lub dokładnie, jeżeli $X_i \sim N(\theta, \sigma^2)$, dodatkowo oznaczmy kwantyle rozkładu $N(0, 1)$ jako $z_{1-\alpha/2} - 1 - \alpha/2$ - ty kwantyl, wtedy:

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

Przekształcając:

$$P\left(T - z_{1-\alpha/2} \frac{\sigma_T}{\sqrt{n}} \leq \theta_0 \leq T + z_{1-\alpha/2} \frac{\sigma_T}{\sqrt{n}}\right) = 1 - \alpha$$

Gdy t to zrealizowana wartość statystyki T otrzymujemy przedział ufności:

$$\left(t - z_{1-\alpha/2} \frac{\sigma_T}{\sqrt{n}}, t + z_{1-\alpha/2} \frac{\sigma_T}{\sqrt{n}}\right)$$

Zadanie 2

Cel

Celem zadania było oszacowanie prawdopodobieństwa pokrycia nieznanej średniej ze **znaną wariancją** przez przedział ufności z poziomem ufności 0.95, oraz jego długość dla trzech rozkładów: (a) normalnego, (b) logistycznego i (c) Cauchy'ego z różnymi parametrami każdy.

Stosowane metody

Dla każdego podpunktu (a) - (c) wygenerowałem 10 000 razy próbę wielkości $n = 50$. Dla każdej z tych prób obliczyłem przedział ufności (różnie dla podpunktów) oraz jego długość. Następnie obliczyłem prawdopodobieństwo, że średnia (równa przesunięciu dla (a) i (b), nieistniejąca dla (c)) znajduje się w tym przedziale, oraz jego średnią długość. Przedziały zostały obliczone w następujący sposób:

- (a) Dzięki rozkładowi normalnemu, dla tego podpunktu możemy dokładnie obliczyć przedział ufności używając wzorów z [zadania 1](#) (nie musimy korzystać z centralnego twierdzenia granicznego przy rozkładzie Z w [obliczeniach](#) zadania 1)
- (b) Stosujemy asymptotyczny przedział ufności dla średniej, oparty na centralnym twierdzeniu granicznym. Dodatkowo musimy obliczyć dodatkowo wariancję ze skali: $\sigma = \text{skala}\pi/3$, lecz podążając wzorami z [zadania 1](#) ciągle możemy je obliczyć.
- (c) Średnia oraz wariancja są nieokreślone w rozkładzie Cauchy’ego, więc nie możemy ich szacować.

Wyniki

(a) Rozkład normalny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	94.730003	0.554362
$\mu = 0, \sigma = 2$	95.320000	1.108723
$\mu = 0, \sigma = 3$	94.750000	1.663085

(b) Rozkład logistyczny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	95.029999	1.005501
$\mu = 0, \sigma = 2$	94.760002	2.011001
$\mu = 0, \sigma = 3$	94.949997	3.016502

(c) Rozkład Cauchy’ego Nieokreślone

Wnioski

Dla rozkładu normalnego rzeczywiste pokrycie jest bardzo bliskie oczekiwanemu, a długości przedziałów są małe, co zgadza się z teorią z [zadania 1](#) – dla rozkładu normalnego mamy dokładny wynik.

Dla rozkładu logistycznego używamy przedziału normalnego asymptotycznego, przez co teoretycznie powinny być większe odchylenia od rozkładu normalnego, jednak już przy $n = 50$ rzeczywiste pokrycie jest bliskie oczekiwanemu. Jednak przez ciężkie ogony tego rozkładu, długość przedziału jest dłuższa.

W rozkładzie Cauchy’ego:

- nie istnieje średnia
- nie istnieje wariancja
- nie działa prawo wielkich liczb
- nie działa centralne twierdzenie graniczne

Średnia z próby nie stabilizuje się i nie koncentruje. Dlatego przedziały oparte na rozkładzie Cauchy’ego nie mają żadnego sensu i nie mogą zapewnić pokrycia.

Załączniki

Generacja danych, obliczanie przedziału i dalsza logika znajduje się w załączonym pliku `main.cpp`.

Zadanie 3

Cel

Celem zadania było przedział ufności dla **średniej** dla współczynnika ufności $1 - \alpha$ przy **nieznanej wariancji**.

Obliczenia

Na początku będziemy rozważać zmienne losowe należące do rozkładu normalnego, później dla rozkładu o pewnej gęstości $f(x; \theta)$, $\theta \in \Theta$ i skończonej wariancji. Będę również używał tych samych oznaczeń co w poprzednich zadaniach na przedziały ufności.

Na początek niech:

$$X_i \sim N(\mu, \sigma^2), \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim N(\mu, \frac{\sigma^2}{n}), S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

odpowiednio średnia i wariancja próbna (nieobciążona). Ważne jest to, że \bar{X} i S są niezależne. Używając ponownie Z jak w zadaniu 1, oraz definiujemy dodatkowo U :

$$Z \sim N(0, 1), U = \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

Definiujemy kolejną statystykę:

$$T = \frac{Z}{\sqrt{U/(n-1)}} = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t_{n-1}$$

Z definicją $t_{1-\alpha/2, n-1}$ jako $1 - \alpha/2$ kwantyl rozkładu t_{n-1} :

$$P\left(-t_{1-\alpha/2, n-1} \leq \frac{\sqrt{n}(\bar{X} - \mu)}{S} \leq t_{1-\alpha/2, n-1}\right) = 1 - \alpha$$
$$P\left(\bar{X} - t_{1-\alpha/2, n-1} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{1-\alpha/2, n-1} \frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

Czyli dla zrealizowanych wartości \bar{X} i S :

$$P\left(\bar{x} - t_{1-\alpha/2, n-1} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{1-\alpha/2, n-1} \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

Dla przypadku ogólnego, czyli X_i z rozkładu o pewnej gęstości $f(x; \theta)$, $\theta \in \Theta$ i skończonej wariancji. Ponieważ w naszym modelu estymator S^2 wariancji σ^2 jest spójny, to:

$$S \xrightarrow{p} \sigma$$

Więc z twierdzenia Slutsky'ego mamy:

$$\frac{\sqrt{n}(\bar{X} - \mu)}{S} \xrightarrow{d} N(0, 1)$$

Więc bardzo podobnie jak w zadaniu 1, jednak z estymatorem wariancji przedział dla zrealizowanych wartości to:

$$\left(\bar{x} - z_{1-\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2} \frac{s}{\sqrt{n}}\right)$$

Zadanie 4

Cel

Celem zadania było oszacowanie prawdopodobieństwa pokrycia nieznanej średniej z **nieznaną wariancją** przez przedział ufności z poziomem ufności 0.95, oraz jego długość dla trzech rozkładów: (a) normalnego, (b) logistycznego i (c) Cauchy'ego z różnymi parametrami każdy.

Stosowane metody

Dla każdego podpunktu (a) - (c) wygenerowałem 10 000 razy próbę wielkości $n = 50$. Dla każdej z tych prób obliczyłem przedział ufności oraz jego długość. Następnie obliczyłem prawdopodobieństwo, że średnia (równa przesunięciu dla (a) i (b), nieistniejąca dla (c)) znajduje się w tym przedziale, oraz jego średnią długość. Przedziały zostały obliczone w następujący sposób:

- (a) Pomimo możliwości implementacji dokładnej wersji przedziału ufności, wybrałem użycie metody asymptotycznej, ponieważ jest ona prostsza w implementacji. Obliczamy S^2 tak samo jak w [zadaniu 3](#).
- (b) Ponownie używamy asymptotycznie poprawny przedział. Nie musimy dodatkowo skalować, ponieważ S oblicza wariancję, nie skalę rozkładu.
- (c) Średnia oraz wariancja są nieokreślone w rozkładzie Cauchy'ego, więc nie możemy ich szacować.

Wyniki

(a) Rozkład normalny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	93.879997	0.551119
$\mu = 0, \sigma = 2$	94.370003	1.102754
$\mu = 0, \sigma = 3$	94.059998	1.652058

(b) Rozkład logistyczny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	94.519997	0.995998
$\mu = 0, \sigma = 2$	94.769997	1.995133
$\mu = 0, \sigma = 3$	94.459999	2.996672

(c) Rozkład Cauchy'ego Nieokreślone

Wnioski

Dla rozkładu normalnego rzeczywiste pokrycie jest minimalnie mniejsze od oczekiwanego, wynika to z użycia przybliżenia zamiast dokładnej wartości σ , a długości przedziałów mają takie same długości co w [zadaniu 2](#), co zgadza się z teorią z [zadania 3](#).

Dla rozkładu logistycznego pokrycie okazuje się mniej więcej tak samo precyzyjne jak dla rozkładu normalnego, przez stabilną estymację wariancji pokrycie pozostaje w małej odległości od oczekiwanego.

Tak samo jak w [zadaniu 2](#), obliczanie średniej nie ma sensu dla rozkładu Cauchy'ego.

Załączniki

Generacja danych, obliczanie przedziału i dalsza logika znajduje się w załączonym pliku `main.cpp`.

Zadanie 5

Cel

Celem zadania było przedziału ufności dla **wariancji** dla współczynnika ufności $1 - \alpha$ przy **znanej średniej**.

Obliczenia

Na początku dla $X_i \sim N(\mu, \sigma^2)$ dla znanego μ , następnie przypadek ogólny (asymptotyczny).

$$Q = \sum_{i=1}^n (X_i - \mu)^2$$

Korzystając z rozkładu chi-kwadrat:

$$\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 = \frac{Q}{\sigma^2} \sim \chi_n^2$$

Z oznaczeniem kwantyla tak jak w poprzednich zadaniach mamy przedział:

$$P(\chi_{n,\alpha/2}^2 \leq \frac{Q}{\sigma^2} \leq \chi_{n,1-\alpha/2}^2) = 1 - \alpha$$

Przekształcając względem σ^2 :

$$P\left(\frac{Q}{\chi_{n,1-\alpha/2}^2} \leq \sigma^2 \leq \frac{Q}{\chi_{n,\alpha/2}^2}\right) = 1 - \alpha$$

Czyli dla zrealizowanej wartości Q mamy przedział:

$$\left(\frac{q}{\chi_{n,1-\alpha/2}^2}, \frac{q}{\chi_{n,\alpha/2}^2}\right)$$

Teraz **dla przypadku ogólnego**, z dodatkowym założeniem $E[(X_i - \mu)^4] < \infty$ dodatkowo $S^2 = Q/n$ oraz $Y_i = (X_i - \mu)^2$ mamy:

$$E[Y_i] = \sigma^2, \tau^2 := \text{Var}(Y_i) < \infty$$

τ jednak nie jest nam dokładnie znane, więc estymujemy je spójnie przez $\hat{\tau} = \frac{1}{n} \sum_{i=1}^n (Y_i^2 - (S^2)^2)$, używając centralnego twierdzenia granicznego:

$$\sqrt{n}(S^2 - \sigma^2) \xrightarrow{d} N(0, \hat{\tau}^2)$$

Dodatkowo używając twierdzenia Slutsky'ego:

$$\frac{\sqrt{n}(S^2 - \sigma^2)}{\hat{\tau}} \xrightarrow{d} N(0, 1)$$

W końcu możemy użyć kwantyli rozkładu normalnego i otrzymać:

$$P\left(-z_{1-\alpha/2} \leq \frac{\sqrt{n}(S^2 - \sigma^2)}{\hat{\tau}} \leq z_{1-\alpha/2}\right) \approx 1 - \alpha$$

I przedział dla zrealizowanych wartości:

$$\left(s^2 - z_{1-\alpha/2} \frac{\hat{\tau}}{\sqrt{n}}, s^2 + z_{1-\alpha/2} \frac{\hat{\tau}}{\sqrt{n}} \right)$$

Zadanie 6

Cel

Celem zadania było oszacowanie prawdopodobieństwa pokrycia nieznanej wariancji ze **znaną średnią** przez przedział ufności z poziomem ufności 0.95, oraz jego długość dla trzech rozkładów: (a) normalnego, (b) logistycznego i (c) Cauchy’ego z różnymi parametrami każdy.

Stosowane metody

Dla każdego podpunktu (a) - (c) wygenerowałem 10 000 razy próbę wielkości $n = 50$. Dla każdej z tych prób obliczyłem przedział ufności oraz jego długość. Następnie obliczyłem prawdopodobieństwo, że średnia (równa przesunięciu dla (a) i (b), nieistniejąca dla (c)) znajduje się w tym przedziale, oraz jego średnią długość. Przedziały zostały obliczone w następujący sposób:

- (a) Pomimo możliwości implementacji dokładnej wersji przedziału ufności, wybrałem użycie metody asymptotycznej, ponieważ jest ona prostsza w implementacji. Obliczamy S^2 i $\hat{\tau}$ tak samo jak w [zadaniu 5](#).
- (b) Ponownie używamy asymptotycznie poprawny przedział. Znowy musimy obliczyć dodatkowo wariancję ze skali: $\sigma = \text{skala}\pi/3$ aby otrzymać poprawne wartości.
- (c) Średnia oraz wariancja są nieokreślone w rozkładzie Cauchy’ego, więc nie możemy ich szacować.

Wyniki

(a) Rozkład normalny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	91.220001	0.752604
$\mu = 0, \sigma = 2$	91.550003	2.995736
$\mu = 0, \sigma = 3$	91.519997	6.775370

(b) Rozkład logistyczny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, s = 1$	88.779999	3.013078
$\mu = 0, s = 2$	88.559998	11.977070
$\mu = 0, s = 3$	88.790001	26.949993

(c) Rozkład Cauchy’ego Niekreślone

Wnioski

Dla rozkładu normalnego rzeczywiste pokrycie jest mniejsze od oczekiwanego, wynika to z użycia przybliżenia zamiast dokładnej wartości σ , długości przedziałów różnią się od zadań ze średnimi. Widocznie zbieżność wariancji jest wolniejsza od średniej.

Dla rozkładu logistycznego pokrycie jest gorsze, co wynika z ciężkich ogonów tego rozkładu, wynika to między innymi z tego, że τ^2 jest większe, przez co przedziały są bardziej niestabilne, oraz są zdecydowanie większe.

Tak samo jak w poprzednich zadaniach, obliczanie wariancji nie ma sensu dla rozkładu Cauchy’ego.

Załączniki

Generacja danych, obliczanie przedziału i dalsza logika znajduje się w załączonym pliku `main.cpp`.

Zadanie 7

Cel

Celem zadania było przedziału ufności dla **wariancji** dla współczynnika ufności $1 - \alpha$ przy **nieznanej średniej**.

Obliczenia

Początkowo dla $X_i \sim N(\mu, \sigma^2)$, następnie przejdziemy do ogólnego przypadku. Używając takich samych definicji \bar{X} jak w [zadaniu 3](#), oraz $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ zaskakująco szybciej dostajemy szukany rozkładu chi-kwadrat, jednak z $n - 1$ stopniami swobody zamiast n :

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2$$

Co prosto prowadzi nas do prawdopodobieństwa kwantylowego:

$$P\left(\chi_{n-1, \alpha/2}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{n-1, 1-\alpha/2}^2\right) = 1 - \alpha$$

Końcowo dając nam przedział (dla zrealizowanych wartości):

$$\left(\frac{(n-1)s^2}{\chi_{n-1, 1-\alpha/2}^2}, \frac{(n-1)s^2}{\chi_{n-1, \alpha/2}^2}\right)$$

Teraz **dla przypadku ogólnego**, ponownie musimy założyć $E[(X_i - \mu)^4] < \infty$, dodatkowo przyjmujemy definicje Y_i oraz $\tilde{S}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{X})^2$. Korzystając z rozkładu:

$$(X_i - \bar{X})^2 = (X_i - \mu)^2 - 2(\bar{X} - \mu)(X_i - \mu) + (\bar{X} - \mu)^2$$

Możemy zapisać:

$$\tilde{S}^2 - \sigma^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \sigma^2) - 2(\bar{X} - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2$$

Gdzie część poza sumą możemy zapisać jako reszta r_n , która przez posiadanie jedynie elementów typu $(\bar{X} - \mu)^2$ zachowuje się następująco:

$$(\bar{X} - \mu)^2 = O(n^{-1}) \implies \sqrt{n}(\bar{X} - \mu)^2 = O(n^{-1/2})$$

Zgodnie z [Big O notation](#), czyli asymptotycznie r_n dąży do zera. Chcemy skorzystać z centralnego twierdzenia granicznego i twierdzenia Slutsky'ego:

$$\sqrt{n}(\tilde{S}^2 - \sigma^2) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (Y_i - \sigma^2) + \sqrt{n}r_n$$

Więc z $\tau^2 := \text{Var}(Y_i) < \infty$ oraz jego spójnym estymatorem $\hat{\tau} = \frac{1}{n} \sum_{i=1}^n (Y_i^2 - (\tilde{S}^2)^2)$:

$$\sqrt{n}(\tilde{S}^2 - \sigma^2) \xrightarrow{d} N(0, \tau^2), \frac{\sqrt{n}(\tilde{S}^2 - \sigma^2)}{\hat{\tau}} \xrightarrow{d} N(0, 1)$$

Otrzymujemy prawdopodobieństwo:

$$P\left(-z_{1-\alpha/2} \leq \frac{\sqrt{n}(\tilde{S}^2 - \sigma^2)}{\hat{\tau}} \leq z_{1-\alpha/2}\right) \approx 1 - \alpha$$

I przedział:

$$\left(\tilde{s}^2 - z_{1-\alpha/2} \frac{\hat{\tau}}{\sqrt{n}}, \tilde{s}^2 + z_{1-\alpha/2} \frac{\hat{\tau}}{\sqrt{n}}\right)$$

Zadanie 8

Cel

Celem zadania było oszacowanie prawdopodobieństwa pokrycia nieznaną wariancji z **nieznaną średnią** przez przedział ufności z poziomem ufności 0.95, oraz jego długość dla trzech rozkładów: (a) normalnego, (b) logistycznego i (c) Cauchy'ego z różnymi parametrami każdy.

Stosowane metody

Dla każdego podpunktu (a) - (c) wygenerowałem 10 000 razy próbę wielkości $n = 50$. Dla każdej z tych prób obliczyłem przedział ufności oraz jego długość. Następnie obliczyłem prawdopodobieństwo, że średnia (równa przesunięciu dla (a) i (b), nieistniejąca dla (c)) znajduje się w tym przedziale, oraz jego średnią długość. Przedziały zostały obliczone w następujący sposób:

- (a) Pomimo możliwości implementacji dokładnej wersji przedziału ufności, wybrałem użycie metody asymptotycznej, ponieważ jest ona prostsza w implementacji. Obliczamy \tilde{S}^2 i $\hat{\tau}$ tak samo jak w [zadaniu 7](#).
- (b) Ponownie używamy asymptotycznie poprawny przedział. Ze skalowaną skalą jako docelowa wariancja.
- (c) Średnia oraz wariancja są nieokreślone w rozkładzie Cauchy'ego, więc nie możemy ich szacować.

Wyniki

(a) Rozkład normalny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	89.349998	0.735663
$\mu = 0, \sigma = 2$	89.889999	2.960889
$\mu = 0, \sigma = 3$	90.029999	6.631015

(b) Rozkład logistyczny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, s = 1$	86.849998	2.905790
$\mu = 0, s = 2$	87.029999	11.646937
$\mu = 0, s = 3$	86.930000	26.246204

(c) Rozkład Cauchy'ego Niekreślone

Wnioski

Dla rozkładu normalnego rzeczywiste pokrycie ponownie jest mniejsze od oczekiwanego, co znowu jest spowodowane użyciem asymptotycznie poprawnego modelu. Przy nieznaną średniej pokrycie jest nieznacznie gorsze niż przy znanej, co w oczywisty sposób wynika z kolejnej warstwy przybliżenia. Długości przedziałów są minimalnie krótsze od modelu ze znaną średnią. Pokrycie pozostaje stabilne i poprawne.

Dla rozkładu logistycznego pokrycie również jest gorsze od poprzedniego modelu, co wynika z ciężkich ogonów tego rozkładu, między innymi z tego, że $\hat{\tau}^2$ jest większe, przez co przedziały są bardziej niestabilne, oraz przedziały są zdecydowanie większe. Przedziały i pokrycie pozostają zdecydowanie gorsze od rozkładu normalnego.

Tak samo jak w poprzednich zadaniach, obliczanie wariancji nie ma sensu dla rozkładu Cauchy'ego.

Załączniki

Generacja danych, obliczanie przedziału i dalsza logika znajduje się w załączonym pliku `main.cpp`.

Zadanie 9

Cel

Celem zadania było przedział ufności dla **wariancji** dla współczynnika ufności $1 - \alpha$ przy **nieznanej średniej**.

Nie jestem 100% pewny czy o to chodzi, ale zrozumiałem zadanie jako ustalenie pewnego warunku, który ma zostać spełniony przez zmienną losową z jakiegoś rozkładu (np. $X > 0$) i wyznaczenie przedziału, w którym znajduje się prawdopodobieństwo spełniania warunku przez zmienną losową, z prawdopodobieństwem $1 - \alpha$

Obliczenia

U nas tym warunkiem będzie $P(X > 0)$, definiujemy p – prawdopodobieństwo, że pojedyncza obserwacja X jest dodatnia oraz \hat{p} – estymator proporcji w następujący sposób:

$$p := P(X > 0), \hat{p} := \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{X_i > 0\}$$

Ponieważ I_i są i.i.d. z rozkładu Bernoulliego z $E[I_i] = p$, $Var[I_i] = p(1-p)$, zatem $\hat{p} \xrightarrow{p} p$, z centralnego twierdzenia granicznego:

$$\sqrt{n}(\hat{p} - p) \xrightarrow{d} N(0, p(1-p))$$

Z klasycznej własności wariancji $Var(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n}$, więc ze Slutsky'ego:

$$\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \xrightarrow{d} N(0, 1)$$

Ponownie używając kwantyli rozkładu normalnego:

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

Otrzymujemy w końcu przedział:

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

Zadanie 10

Cel

Celem zadania było oszacowanie prawdopodobieństwa pokrycia nieznanej proporcji dodatnich obserwacji przez przedział ufności z poziomem ufności 0.95, oraz jego długość dla trzech rozkładów: (a) normalnego, (b) logistycznego i (c) Cauchy'ego z różnymi parametrami każdy.

Stosowane metody

Dla każdego podpunktu (a) - (c) wygenerowałem 10 000 razy próbę wielkości $n = 50$. Dla każdej z tych prób obliczyłem przedział ufności oraz jego długość. Następnie obliczyłem prawdopodobieństwo, że proporcja znajduje się w tym przedziale, oraz jego średnią długość. Przedziały zostały obliczone w następujący sposób:

- (a) W tym przypadku normalność rozkładu nie ma znaczenia, liczymy \hat{p} tak samo jak w [zadaniu 9](#).
- (b) Ponownie używamy asymptotycznie poprawny przedział.
- (c) Po raz pierwszy możemy obliczyć parametr dla rozkładu Cauchy'ego, liczymy tak samo jak w poprzednich podpunktach.

Wyniki

(a) Rozkład normalny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	93.239998	0.274313
$\mu = 0, \sigma = 2$	93.410004	0.274373
$\mu = 0, \sigma = 3$	93.360001	0.274383

(b) Rozkład logistyczny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	93.940002	0.274429
$\mu = 0, \sigma = 2$	93.019997	0.274294
$\mu = 0, \sigma = 3$	93.470001	0.274342

(c) Rozkład Cauchy'ego

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	93.449997	0.274335
$\mu = 0, \sigma = 2$	93.160004	0.274314
$\mu = 0, \sigma = 3$	93.529999	0.274318

Wnioski

Dla wszystkich podpunktów pokrycie jest nieco poniżej ustalonego 95%, co wynika z asymptotycznej poprawności wzoru. Oczekujemy że poprawi się to z większym n .

Pokrycie oraz średnia długość mają bardzo bliskie wartości we wszystkich podpunktach i przy wszystkich parametrach. Są takie same między parametrami, ponieważ wartości X_i nie wpływają na boki przedziałów, oraz między rozkładami, ponieważ \hat{p} również nie zależy od wartości, indykator $\mathbf{1}\{X > 0\}$ ma rozkład Bernoulliego. Czyli ciężkie ogony lub nieokreślona średnia lub wariancja nie mają wpływu na wynik.

Załączniki

Generacja danych, obliczanie estymatora i dalsza logika znajduje się w załączonym pliku `main.cpp`.

Zadanie 11

Cel

Przez zmianę wielkości próby na $n = 20$ albo $n = 100$ dla zadań 2, 4, 6, 8 i 10 należy zauważyć różnicę między wynikami dla oryginalnej wielkości próby a zmodyfikowanej. (Tu będzie **BARDZO** dużo danych)

Wyniki dla $n = 20$

Zadanie 2

(a) Rozkład normalny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	94.930000	0.876523
$\mu = 0, \sigma = 2$	94.989998	1.753045
$\mu = 0, \sigma = 3$	94.839996	2.629567

(b) Rozkład logistyczny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	95.040001	1.589836
$\mu = 0, \sigma = 2$	94.779999	3.179672
$\mu = 0, \sigma = 3$	94.949997	4.769508

(c) Rozkład Cauchy'ego Nieokreślone

Zadanie 4

(a) Rozkład normalny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	93.629997	0.865866
$\mu = 0, \sigma = 2$	93.699997	1.730083
$\mu = 0, \sigma = 3$	93.379997	2.593298

(b) Rozkład logistyczny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	93.400002	1.555475
$\mu = 0, \sigma = 2$	93.739998	3.104293
$\mu = 0, \sigma = 3$	93.639999	4.669601

(c) Rozkład Cauchy'ego Nieokreślone

Zadanie 6

(a) Rozkład normalny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	87.230003	1.129791
$\mu = 0, \sigma = 2$	86.910004	4.504707
$\mu = 0, \sigma = 3$	86.949997	10.144330

Parametry	Pokrycie [%]	Średnia długość
-----------	--------------	-----------------

(b) Rozkład logistyczny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	83.150002	4.367570
$\mu = 0, \sigma = 2$	83.260002	17.490837
$\mu = 0, \sigma = 3$	83.040001	39.165217

(c) Rozkład Cauchy’ego Nieokreślone

Zadanie 8

(a) Rozkład normalny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	83.389999	1.060738
$\mu = 0, \sigma = 2$	84.010002	4.304702
$\mu = 0, \sigma = 3$	83.870003	9.562838

(b) Rozkład logistyczny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	79.309998	4.019426
$\mu = 0, \sigma = 2$	79.290001	16.177702
$\mu = 0, \sigma = 3$	79.639999	36.549187

(c) Rozkład Cauchy’ego Nieokreślone

Zadanie 10

(a) Rozkład normalny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	95.779999	0.426970
$\mu = 0, \sigma = 2$	95.949997	0.427039
$\mu = 0, \sigma = 3$	95.940002	0.426856

(b) Rozkład logistyczny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	95.580002	0.426754
$\mu = 0, \sigma = 2$	95.949997	0.426778
$\mu = 0, \sigma = 3$	96.050003	0.427044

Parametry	Pokrycie [%]	Średnia długość
-----------	--------------	-----------------

(c) Rozkład Cauchy’ego

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	95.720001	0.426942
$\mu = 0, \sigma = 2$	95.900002	0.426744
$\mu = 0, \sigma = 3$	95.529999	0.426505

Wyniki dla $n = 100$

Zadanie 2

(a) Rozkład normalny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	95.360001	0.391993
$\mu = 0, \sigma = 2$	94.989998	0.783986
$\mu = 0, \sigma = 3$	95.089996	1.175978

(b) Rozkład logistyczny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	95.269997	0.710996
$\mu = 0, \sigma = 2$	94.489998	1.421993
$\mu = 0, \sigma = 3$	95.120003	2.132989

(c) Rozkład Cauchy’ego Nieokreślone

Zadanie 4

(a) Rozkład normalny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	94.800003	0.390895
$\mu = 0, \sigma = 2$	94.779999	0.782562
$\mu = 0, \sigma = 3$	95.040001	1.173047

(b) Rozkład logistyczny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	94.419998	0.707865
$\mu = 0, \sigma = 2$	94.370003	1.416054
$\mu = 0, \sigma = 3$	94.639999	2.125839

(c) Rozkład Cauchy’ego Nieokreślone

Zadanie 6

(a) Rozkład normalny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	93.190002	0.541758
$\mu = 0, \sigma = 2$	93.199997	2.170244
$\mu = 0, \sigma = 3$	93.180000	4.888465

(b) Rozkład logistyczny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	91.339996	2.196907
$\mu = 0, \sigma = 2$	91.260002	8.765365
$\mu = 0, \sigma = 3$	91.519997	19.785853

(c) Rozkład Cauchy’ego Nieokreślone

Zadanie 8

(a) Rozkład normalny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	92.309998	0.536705
$\mu = 0, \sigma = 2$	91.989998	2.149012
$\mu = 0, \sigma = 3$	92.120003	4.831196

(b) Rozkład logistyczny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	89.790001	2.165302
$\mu = 0, \sigma = 2$	89.709999	8.635128
$\mu = 0, \sigma = 3$	90.169998	19.464339

(c) Rozkład Cauchy’ego Nieokreślone

Zadanie 10

(a) Rozkład normalny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	94.040001	0.195003
$\mu = 0, \sigma = 2$	94.410004	0.195009
$\mu = 0, \sigma = 3$	94.379997	0.195004

(b) Rozkład logistyczny

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	94.330002	0.195018
$\mu = 0, \sigma = 2$	94.230003	0.195005
$\mu = 0, \sigma = 3$	94.190002	0.195009

(c) Rozkład Cauchy’ego

Parametry	Pokrycie [%]	Średnia długość
$\mu = 0, \sigma = 1$	94.760002	0.195035
$\mu = 0, \sigma = 2$	94.230003	0.195016
$\mu = 0, \sigma = 3$	94.349998	0.195006

Wnioski

Zadanie 2 Dla (a) i (b) zmiana n nie ma widocznego wpływu na pokrycie, jednak zwiększanie n powoduje zmniejszenie przedziału ufności, co zgadza się z wzorami.

Zadanie 4 Ponownie z oboma podpunktami, zwiększanie n ponownie zmniejsza przedział ufności, chociaż teraz minimalnie można zauważyć również zwiększenie pokrycia. Ogółem wyniki są mniej dokładne niż z [zadania 2](#).

Zadanie 6 W obu podpunktach zwiększanie n drastycznie poprawia pokrycie, zbliżając te wartości do siebie (co wskazuje że przy większych n zwiększanie daje mniejszy zysk) oraz ponownie skraca przedział ufności zgodnie z wzorami. Ciężkie ogony mogą wyciągać przedział poza pokrycie prawdziwej wartości.

Zadanie 8 Znowu zwiększanie n drastycznie poprawia pokrycia oraz zmniejsza przedział ufności. Ponownie przy nieznaney wartości drugiego parametru, wyniki są nieco gorsze niż te z [zadania 6](#).

Zadanie 10 Możemy omówić wszystkie podpunkty, z wzrostem n pokrycie się nieco zmniejsza, przedział ufności również się zmniejsza. Spadek pokrycia wynika ze zmniejszenia przedziału szybciej niż asymptotyka wzoru przybliża nas do prawdziwej wartości, przez w tych wartościach n nieco tracimy precyzję, chociaż testy dla wartości około $n = 10000$ wskazują na to, że pokrycie stabilizuje się bardzo blisko oczekiwanego $1 - \alpha$.

Źródła

- *Introduction to Mathematical Statistics (6th edition)* - Hogg, McKean, Craig
- *Statystyka dla studentów kierunków technicznych i przyrodniczych* Koronacki; J., Mielniczuk, J. (2009)
- [Student’s t-distribution](#)