Lista 3

Martyna Konopacka

Cel

Celem tego sprawozdania jest skonfrontowanie teoretycznych przedziałów ufności średniej, wariancji i proporcji dla nieznanych parametrów z rozkładów: normalnego, Cauchy'ego, logistycznego, wykładniczego i chi-kwadrat z wynikami eksperymentów. W zadaniach rozważane są od razu różne liczebności próby.

Wprowadzenie - estymacja przedziałowa

Załóżmy, że zmienna losowa X ma rozkład w populacji z nieznanym parametrem θ. Z populacji wybieramy próbę losową $X_1, X_2, \ldots, X_n$. *Przedziałem ufności* o współczynniku ufności $1-\alpha$ nazwiemy taki przedział (θ_1, θ_2) który spełnia warunek: $P(\theta_1 < \theta < \theta_2) = 1-\alpha$, gdzie θ_1 i θ_2 są funkcjami wyznaczonymi na podstawie próby losowej. Współczynnik ufności wyraża wtedy prawdopodobieństwo, że rzeczywista wartość θ znajduje się w przedziale θ_1, θ_2.

Zadanie 1 pokazuje procedurę wyznaczania przedziału o zadanym poziomie ufności na przykładzie estymacji średniej w modelu normalnym o znanej wariancji, natomiast w zadaniu 3 przejdziemy do modelu o (bardziej realistyczne) nieznanej wariancji. Estymując średnią oprzemy się na poniższym twierdzeniu:

Centralne Twierdzenie Graniczne (CTG)

Załóżmy, że $X_1, ..., X_n$ są niezależnymi zmiennymi losowymi o tym samym rozkładzie, wartości oczekiwanej μ i wariancji σ^2. Wtedy zmienna $\bar{X} = \frac{1}{n}\sum X_i$ zbiega według rozkładu do rozkładu normalnego $N(\mu, \frac{\sigma^2}{n})$ W praktyce zazwyczaj przyjmuje się, że jeśli $n \geq 30$ to $X \sim N(\mu, \frac{\sigma^2}{n})$.

Uwaga: twierdzenie ma zastosowanie tylko dla rozkładów o znanej wartości oczekiwanej i wariancji! Zobaczymy, że szacowanie przedziałów ufności nie zadziała dla rozkładów niespełniających tego warunku na przykładzie rozkładu Cauchy'ego.

Zadanie 1

Niech $X_1, X_2, ..., X_n$ będzie próbą losową z rozkładu o nieznanej wartości oczekiwanej μ i znanej wariancji σ^2. Na mocy CTG średnia próbkowa $\bar{X}$ ma rozkład $N(\mu, \frac{\sigma^2}{n})$. Wyrazimy standardową zmienną normalną (której przedziały ufności są nam znane) $Z \sim N(1,0)$ jako $Z = \frac{\sqrt{n}}{\sigma}(\bar{X}-\mu)$ i za pomocą przekształceń znajdziemy przedziały ufności dla μ. Niech $\phi^{-1}(\frac{\alpha}{2}) = \mu_1$ i $\phi^{-1}(1-\frac{\alpha}{2}) = \mu_2$, gdzie ϕ jest dystrybuntą rozkładu.

$1-\alpha = P(\mu_1 \leq Z \leq \mu_2)$

$1-\alpha = P(\mu_1 \leq \frac{\sqrt{n}}{\sigma}(\bar{X}-\mu) \leq \mu_2)$

$1-\alpha = P(\frac{\sigma\mu_1}{\sqrt{n}} - \bar{X} \leq -\mu \leq \frac{\sigma\mu_2}{\sqrt{n}} - \bar{X})$

$1-\alpha = P(\bar{X} - \frac{\sigma}{\sqrt{n}}\mu_2 \leq \mu \leq \bar{X} - \frac{\sigma}{\sqrt{n}}\mu_1)$

W ten sposób wyznaczyliśmy przedział ufności na poziomie ufności $1-\alpha$ dla μ.

Zadanie 2

Sprawdzimy doświadczalnie, jaki procent średnich próbkowych rzeczywiście mieści się w obliczonym jak w zadaniu 1 przedziale ufności, przy czym należy zwrócić uwagę że *jako wartości oczekiwanej i odchylenia standardowego nie użyjemy zawsze po prostu parametrów podanych w zadaniu, a odpowiednio* dla rozkładów:

- normalnego: μ, σ
- logistycznego: $\mu, \frac{\sigma\pi}{\sqrt{3}}$
- Cauchy'ego z uwagi na brak wartości oczekiwanej i wariancji: μ, σ
- wykładniczego: $\frac{1}{\lambda}, \frac{1}{\lambda}$
- chi-kwadrat: $\nu, \sqrt{2\nu}$

rozklad	n	mi	sigma	wynik
rnorm	50	0	1.00	0.958
rnorm	50	0	2.00	0.958
rnorm	50	0	3.00	0.958
rlogis	50	0	1.00	0.960
rlogis	50	0	2.00	0.960
rlogis	50	0	3.00	0.960
rcauchy	50	0	1.00	0.184
rcauchy	50	0	2.00	0.184
rcauchy	50	0	3.00	0.184
rexp	50	0	1.00	0.956
rexp	50	0	0.50	0.956
rexp	50	0	0.33	0.956
rchisq	50	0	1.00	0.957
rchisq	50	0	2.00	0.957
rchisq	50	0	3.00	0.957

- zgodnie z przewidywaniami w oczy rzuca się rozkład Cauchy'ego - odsetek średnich rzeczywiście mieszczących się w wyznaczanym przedziale ufności nieakceptowalnie różni się od teoretycznej wartości 0.95. Oczywiście jest tak dlatego, że podstawą konstrukcji przedziału w tym zadaniu było Centralne Twierdzenie Graniczne, które ma zastosowanie tylko do rozkładów o znanej wariancji i wartości oczekiwanej. Pozostałe rozkłady dają prawidłowe odsetki blisko 0.95 i świadczy to o tym, że twierdzenie rzeczywiście działa, a parametry zostały prawidłowo przeliczone

- dla próby rozmiaru 100 wyniki są bardzo zbliżone, a w miarę zwiększania liczebności zbliżają się coraz bardziej do teoretycznej wartości (przykładowo dla rozkładu normalnego i $n = 200$ wynikiem było już dokładnie 0.95). Co ciekawe, nawet dla $n = 20$, które jest mniejsze niż przyjmowany często próg "działania twierdzeń" $n = 30$, nie widać dużej różnicy. Jedynie rozkład Cauchy'ego zwiększył odsetek w wyraźny sposób, ale jak już wiadomo jest on patologiczny.

rozklad	n	mi	sigma	wynik
rnorm	20	0	1.00	0.958
rnorm	20	0	2.00	0.958
rnorm	20	0	3.00	0.958
rlogis	20	0	1.00	0.951
rlogis	20	0	2.00	0.951
rlogis	20	0	3.00	0.951
rcauchy	20	0	1.00	0.255
rcauchy	20	0	2.00	0.255
rcauchy	20	0	3.00	0.255
rexp	20	0	1.00	0.954

rozklad	n	mi	sigma	wynik
rexp	20	0	0.50	0.954
rexp	20	0	0.33	0.954
rchisq	20	0	1.00	0.962
rchisq	20	0	2.00	0.958
rchisq	20	0	3.00	0.966

Zadanie 3

Niech $X_1, X_2, ..., X_n$ będzie próbą losową z rozkładu normalnego o nieznanej wariancji i parametrze μ. Procedura wyznaczania przedziału ufności dla parametru μ będzie przebiegała podobnie do tej w zadaniu 1, z tą różnicą, że zamiast rozkładu normalnego tym razem estymator będzie miał rozkład studenta, a w roli wariancji wystąpi wariancja próbkowa.

Przez $\bar{X}$ oznaczymy średnią próbkową, a przez S^2 wariancję próbkową obliczaną ze wzoru $S^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$. Wiemy, że $T = \frac{\sqrt{n}(\bar{X}-\mu)}{S}$ ma rozkład studenta z $n-1$ stopniami swobody. Niech t_1, t_2 oznaczają odpowiednio wartość odczytaną z tabeli rozkładu studenta dla $p = \frac{\alpha}{2}$ i $p = 1 - \frac{\alpha}{2}$ przy $n-1$ stopniach swobody. Wtedy:

$1 - \alpha = P(t_1 \leq T \leq t_2)$

$1 - \alpha = P(t_1 \leq \frac{\sqrt{n}(\bar{X}-\mu)}{S} \leq t_2)$

$1 - \alpha = P(\bar{X} - \frac{S}{\sqrt{n}}t_2 \leq \mu \leq \bar{X} - \frac{S}{\sqrt{n}}t_1)$

Zadanie 4

rozklad	n	sigma	wynik
rnorm	50	1	0.959
rnorm	50	2	0.959
rnorm	50	3	0.959
rlogis	50	1	0.961
rlogis	50	2	0.961
rlogis	50	3	0.961
rcauchy	50	1	0.987
rcauchy	50	2	0.987
rcauchy	50	3	0.987
rexp	50	1	0.930
rexp	50	0	0.930
rchisq	50	1	0.931
rchisq	50	2	0.934
rchisq	50	3	0.944

- tym razem szerokość przedziału dla rozkładu Cauchy'ego została przeszacowana.
- dla pozostałych rozkładów wyniki nie są znacząco różne od wyników w przypadku znanej wariancji. Największą różnicę można zaobserwować dla rozkładu wykładniczego, najmniejszą dla normalnego.
- dla innych liczebnośc próby wyniki ponownie są dosyć podobne. Dla niektórych rozkładów widać trochę większe różnice dla małego n (wynika to ze sposobu szacowania wariancji)

rozklad	n	sigma	wynik
rexp	20	1	0.917
rexp	20	0	0.917
rchisq	20	1	0.908
rchisq	20	2	0.926
rchisq	20	3	0.955

Zadanie 5

- Załóżmy, że $X_1, X_2, ...X_n$ to niezależne zmienne losowe z rozkładu $N(0,1)$. Wtedy suma ich kwadratów $Y = \sum X_i^2 \sim \chi_n^2$, gdzie χ_{n-1}^2 oznacza rozkład chi-kwadrat o n stoponiach swobody. Jeśli $X \sim N(\mu, \sigma^2)$ to unormowana zmienna $K = \sum_{i=1}^n (\frac{X_i - \mu}{\sigma})^2 \sim \chi_n^2$
- Nieobciążony estymator wariancji wyraża się wzorem $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \mu)^2$. Po przekształceniu otrzymujemy równość $\frac{(n-1)S_n^2}{\sigma^2} = K$
- Niech teraz k_1, k_2 oznaczają odpowiednio kwantyle rzędu $\frac{\alpha}{2}$ i $1 - \frac{\alpha}{2}$ rozkładu χ_n^2. Można zapisać zatem, że $1 - \alpha = P(k_1 \leq \frac{(n-1)S_n^2}{\sigma^2} \leq k_2)$
- Po przekształceniu otrzymamy przedział ufności na poziomie $1 - \alpha$ postaci $\frac{(n-1)S_n^2}{\chi_{1-\frac{\alpha}{2},n}} \leq \sigma^2 \leq \frac{(n-1)S_n^2}{\chi_{\frac{\alpha}{2},n}}$
- W przypadku nieznanej średniej *(zadanie 7)*, przedział wygląda bardzo podobnie, jednak gdy zamiast μ użyjemy średniej próbkowej $\bar{X}$, należy zamienić n stopni swobody na $n - 1$.

Zadanie 6

W dalszej części rozważamy wszystkie rozkłady oprócz Cauchy'ego, który jak już wiadomo jest patologizny. Estymacja przedziału dla wariancji przy znanej średniej dla n = 50 dała następujące wyniki:

rozklad	n	sigma	wynik
rnorm	50	1	0.942
rnorm	50	2	0.942
rnorm	50	3	0.942
rlogis	50	1	0.882
rlogis	50	2	0.882
rlogis	50	3	0.882
rexp	50	1	0.697
rexp	50	0	0.697
rexp	50	0	0.697
rchisq	50	1	0.557
rchisq	50	2	0.719
rchisq	50	3	0.789

- dla rozkładu normalnego wynik jest zbliżony do teoretycznej wartości 0.95
- pozostałe rozkłady wypadły gorzej - wynika to z tego, że wyznaczając przedział w takiej postaci założyliśmy, że próba pochodzi z rozkładu normalnego. Jak widać, gdy założenie nie jest spełnione taka konstrukcja nie działa.
- dla rozkładu chi-kwadrat można zaobserwować wzrost odsetka wariancji mieszczących się w skonstruowanym przedziale ufności wraz ze wzrostem liczby stopni swobody - na podstawie przeprowadzonych dodatkowo testów wydaje się, że dąży on do wartości około 0.95. Jest tak, ponieważ dla dużej liczby stopni swobody chi-kwadrat przypomina rozkład normalny, więc założenie o normalności jest w przybliżeniu spełnione.
- inne rozkłady nie wykazały powyższego zachowania.

Wyniki dla innych liczebności próby są zbliżone, przykładowo:

rozklad	n	sigma	wynik
rnorm	20	1	0.948
rnorm	20	2	0.948
rnorm	20	3	0.948
rlogis	20	1	0.891
rlogis	20	2	0.891
rlogis	20	3	0.891
rexp	20	1	0.733
rexp	20	0	0.733
rexp	20	0	0.733
rchisq	20	1	0.597
rchisq	20	2	0.748
rchisq	20	3	0.820
rnorm	100	1	0.959
rnorm	100	2	0.959
rnorm	100	3	0.959
rlogis	100	1	0.887
rlogis	100	2	0.887
rlogis	100	3	0.887
rexp	100	1	0.713
rexp	100	0	0.713
rexp	100	0	0.713
rchisq	100	1	0.516
rchisq	100	2	0.687
rchisq	100	3	0.772

Z dodatkowych eksperymentów wynika, że wraz ze wzrostem n wyniki dla rozkładu normalnego ponownie były coraz bliższe teoretycznej wartości.

Zadanie 8

rozklad	n	sigma	wynik
rnorm	100	1	0.956
rlogis	100	1	0.886
rexp	100	1	0.715
rchisq	100	1	0.525
rnorm	50	1	0.948
rlogis	50	1	0.892
rexp	50	1	0.701
rchisq	50	1	0.568
rnorm	20	1	0.952
rlogis	20	1	0.893
rexp	20	1	0.723
rchisq	20	1	0.602

Wyniki są bardzo zbliżone do wyników poprzedniego zadania i nie trzeba nic dodawać po analizie tamtych wyników.

Zadanie 9

Rozważmy proporcję p obserwacji większych od 0 w próbie - ma ona rozkład dwumianowy $B(1, p)$. Zmienna $Z \approx \frac{\bar{X}-p}{\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}} \sim N(0,1)$, przy czym jest to asymptotyczne przybliżenie. Stąd poprzez przekształcenia analogiczne do tych w zadaniu 2, otrzymamy przedział ufności korzystający z kwantyli rozkładu standardowego, postaci

$$1-\alpha = P(\bar{X} - \mu\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \leq p \leq \bar{X} + \mu\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}$$

, gdzie μ jest kwantylem normalnym rzędu $1-\frac{\alpha}{2}$, a $\bar{X}$ oznacza odsetek sukcesów w wylosowanej próbie.

rozklad	n	sigma	wynik
rnorm	100	1	0.938
rnorm	100	2	0.938
rnorm	100	3	0.938
rlogis	100	1	0.944
rlogis	100	2	0.944
rlogis	100	3	0.944
rcauchy	100	1	0.944
rcauchy	100	2	0.944
rcauchy	100	3	0.944
rnorm	20	1	0.949
rnorm	20	2	0.949
rnorm	20	3	0.949
rlogis	20	1	0.950
rlogis	20	2	0.950
rlogis	20	3	0.950
rcauchy	20	1	0.950
rcauchy	20	2	0.950
rcauchy	20	3	0.950
rnorm	50	1	0.943
rnorm	50	2	0.943
rnorm	50	3	0.943
rlogis	50	1	0.942
rlogis	50	2	0.942
rlogis	50	3	0.942
rcauchy	50	1	0.942
rcauchy	50	2	0.942
rcauchy	50	3	0.942

- w tym przypadku rodzaj rozkładu ani jego parametry nie mają znaczącego wpływu na wynik eksperymentu - ma to sens, gdyż wszystkie rozważane rozkłady są symetryczne względem 0 i z równym prawdopodobieństwem dają zmienne większe od zera. Co ciekawe najlepiej wypadły próby rozmiaru 20, ale może to być przypadkowe.

Wnioski

Głównym wnioskiem pojawiającym się kilkukrotnie w zadaniach było to, że tworząc model w oparciu o pewne teoretyczne założenia, należy upewnić się czy rzeczwiście są one spełnione. W przeciwnym wypadku wyniki mogą nie mieć sensu, jak było przy wyznaczaniu przedziału ufności dla rozkładu Cauchy'ego w zadaniu 2 czy 4.