Statystyczna Analiza Danych – laboratorium

Estymacja parametrów

Dorota Celińska-Kopczyńska

Uniwersytet Warszawski

Zajęcia 2
10/11 marca 2022

O czym będą zajęcia?

- Poćwiczymy pracę na wektorach i macierzach podczas szacowania wariancji
- Nauczymy się korzystać z rodziny funkcji `apply`
- Nauczymy się również pisać własne funkcje
- $\hat{\cdot}$ wskazuje, że dana wartość będzie estymatorem (oszacowaniem), $\bar{X}$ będzie średnią

Zadanie 1

1. Utwórz zmienną całkowitoliczbową z wybraną przez siebie wartością od 100 do 5000
2. Wylosuj n obserwacji z rozkładu normalnego o średniej 10 a wariancji 1. (`rnorm`)
3. Wykorzystując jedynie funkcję `sum`, oblicz:
   - nieobciążony estymator wariancji
   - estymator największej wiarogodności wariancji
   - estymator największej wiarogodności odchylenia standardowego
4. Porównaj wyniki z wynikami wbudowanych funkcji `var()` i `sd()`

Funkcje-pętle

- Może się zdarzyć, że będziemy chcieli wykonać jakąś operację dla wszystkich kolumn lub wierszy macierzy (lub data frame,...)
- Pętle w stylu C działają w R wolno. Rozwiązaniem są funkcje z rodziny `apply`
- Jeśli w funkcji użytej wewnątrz apply istnieją dodatkowe argumenty, można się do nich odnieść przez nazwę wewnątrz wywołania apply

Apply

```
# ogolna skladnia
apply(obiekt, po czym dzialamy, funkcja)

# niech m bedzie macierza

apply(m, 1, mean)  # obliczy srednia dla wierszy
apply(m, 2, mean)  # obliczy srednia dla kolumn
apply(m, 2, quantile, c(0.25,0.5,0.75)) # obliczy kwartyle dla kolumn

# mozna zapamietac, ze pierwszy indeks to wiersze, drugi to kolumny
```

Sapply

- Uproszczone apply, działa na wektorach a nie macierzach

```
# ogolna skladnia
sapply(wektor, funkcja)

sapply(x, mean) # obliczy srednia dla wektora
sapply(x, function(x) x/2) # podzieli kazdy element wektora
```

Zadanie 2

1. Wylosuj 5000 obserwacji z rozkładu normalnego o średniej 0 i wybranym przez siebie odchyleniu std
2. Przekształć otrzymany wektor w macierz o wymiarach 10 x 500
3. Dla każdej kolumny wyestymuj wariancję na trzy sposoby, korzystając tylko z `apply` i `var`:
   - $\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2$
   - $\hat{S}_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x-\bar{x})^2$
   - $\hat{S}_2^2 = \frac{1}{n+1}\sum_{i=1}^{n}(x-\bar{x})^2$
4. Otrzymaj $\hat{S}_1^2, \hat{S}_2^2$ przemnażając wartości $\hat{S}^2$

Zadanie 2 – cd

1. Oblicz obciążenia dla wszystkich trzech estymatorów wariancji. Który estymator ma najniższe obciążenie?
2. Porównaj błędy średniokwadratowe (RMSE) dla trzech estymatorów

$$\hat{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta)^2}$$

Funkcje

- W R możemy zdefiniować własne funkcje

```
# ogolna skladnia

nazwa_funkcji <- function(parametry) {
  # cialo funkcji -- linijka
  # cialo funkcji -- linijka
  # ...
  # return(obiekt)
}
```

Zadanie 3

- Napisz funkcję o nazwie `sample_sd`, która:
  - przyjmie dwa argumenty: N oraz n
  - wylosuje N prób rozmiaru n z rozkładu normalnego
  - zwróci N estymowanych odchyleń standardowych
- Utwórz wektor `n <- 2:100`. Dla każdej wartości z wektora otrzymaj, korzystając ze swojej funkcji 100 estymowanych odchyleń std
- Przekształć otrzymany wektor do data frame (pierwsza kolumna wartości estymatora, druga liczebność próby)
- Utwórz wykres punktowy dla uzyskanego data frame (ggplot!)