Statystyczna Analiza Danych – laboratorium

Wczytywanie danych, przedziały ufności

Dorota Celińska-Kopczyńska

Uniwersytet Warszawski

Zajęcia 3
16/17 marca 2023

O czym będą zajęcia?

- Nauczymy się wczytywać różne zbiory danych do R
- Porozmawiamy o statystyce opisowej danych i podstawowej wizualizacji
- Policzymy i porównamy przedziały ufności dla średnich
- $\hat{\cdot}$ wskazuje, że dana wartość będzie estymatorem (oszacowaniem), $\bar{X}$ będzie średnią

Zadanie 1

1. Wczytaj dane Zadluzenie_gmin.csv
2. Sprawdź podstawowe statystyki opisowe próby (`summary`)
3. Oblicz średnią i odchylenie standardowe zadłużenia
4. Przedstaw zadłużenie gmin na histogramie (ggplot2).
5. Czy zadłużenie gminy Ostrowice wygląda na typowe dla polskiej gminy? Przedyskutujmy, czy powinno się usuwać tę obserwację ze zbioru.

Zadanie 2

1. Przyjrzyj się statystykom opisowym zbioru oraz uzyskanemu histogramowi zadłużenia
2. Oceń czy rozkład zadłużenia odbiega od rozkładu normalnego
   - Sporządź wykres kwantylowy (`ggplot2::stat_qq`)
   - Dodaj prostą obrazującą ogólny trend (`ggplot2::stat_qq_line`)
3. Czy wnioski zmienią się, jeśli zlogarytmujemy zmienną (można dodać ε, żeby uniknąć efektu zera)?

Usuwanie obserwacji – używać BARDZO ostrożnie!

```
# pokazujemy "jak" wykonac, chociaz podczas tego labu nie chcemy usuwac obserwacji!
# ZAWSZE stosowac rozsadnie!

m[-c(2176:2178),] # tak sie usuwa wiersze o podanych numerach

m <- subset(m, m$Zadluzenie.gmin < 110) # przyklad wyboru podzbioru, ktory spelnia warunek

rownames(m[m$Zadluzenie.gmin>110,]) # zwroci numery wierszy dla obserwacji, ktore
                                   # spelniaja heurystyke "co jest podejrzane"
```

Wykres kwantyl-kwantyl w ggplot2

```
ggplot() + stat_qq(aes(sample = m$Zadluzenie.gmin)) + stat_qq_line(aes(sample = m$Zadluzenie.gmin))
 + theme_minimal()

# stat_qq -- warstwa odpowiedzialna za narysowanie wykresu kwantyl-kwantyl
# stat_qq_line -- linia, ktora pomaga zauwazyc odchylenia od zachowania
#                 rozkladu zgodnego z rozkladem normalnym
#                 powstaje przez poprowadzenie linii przez punkty
#                 odpowiadajace Q1 i Q3
```

Zadanie 3+4

- Wczytaj dane iris i wybierz wiersze odpowiadające gatunkowi versicolor
- Sprawdź, czy zmienna Sepal.Width ma rozkład normalny
- Oblicz i porównaj przedziały ufności dla średniej wartości zmiennej Sepal.Width
  - studentyzowany: $(\bar{X} \pm \frac{t_{(1-\alpha/2,n-1)}}{\sqrt{n-1}} \hat{S})$
  - asymptotyczny: $(\bar{X} \pm \frac{q_{1-\alpha/2}}{\sqrt{n}} \hat{S})$
- Jakie konsekwencje dla naszej analizy miałoby niespełnienie założenia o normalności rozkładu?