Sprawozdanie 2

Jakub Markowiak
album 255705

22 kwietnia 2021

Spis treści

1 Krótki opis zagadnienia

W tym sprawozdaniu zajmiemy się analizą danych `Salaries` z pakietu `carData`, zawierających informacje o wysokości wynagrodzenia pracowników na jednym z uniwersytetów w USA oraz wykorzystując podstawowe metody graficzne spróbujemy rozstrzygnąć, czy występuje dyskryminacja płacowa ze względu na płeć. Następnie przeanalizujemy własności histogramu, porównamy różne metody jego konstruowania i sprawdzimy, jak dobrze odpowiada on teoretycznej gęstości. Ostatnim zagadnieniem, które będziemy rozpatrywać, jest estymacja dystrybuanty oraz pojęcie dystrybuanty empirycznej. Napiszemy R-funkcję konstruującą dystrybuantę empiryczną oraz obliczającą statystykę Kołmogorowa D_n, aby kolejno zbadać ich własności.

2 Opis eksperymentów/analiz

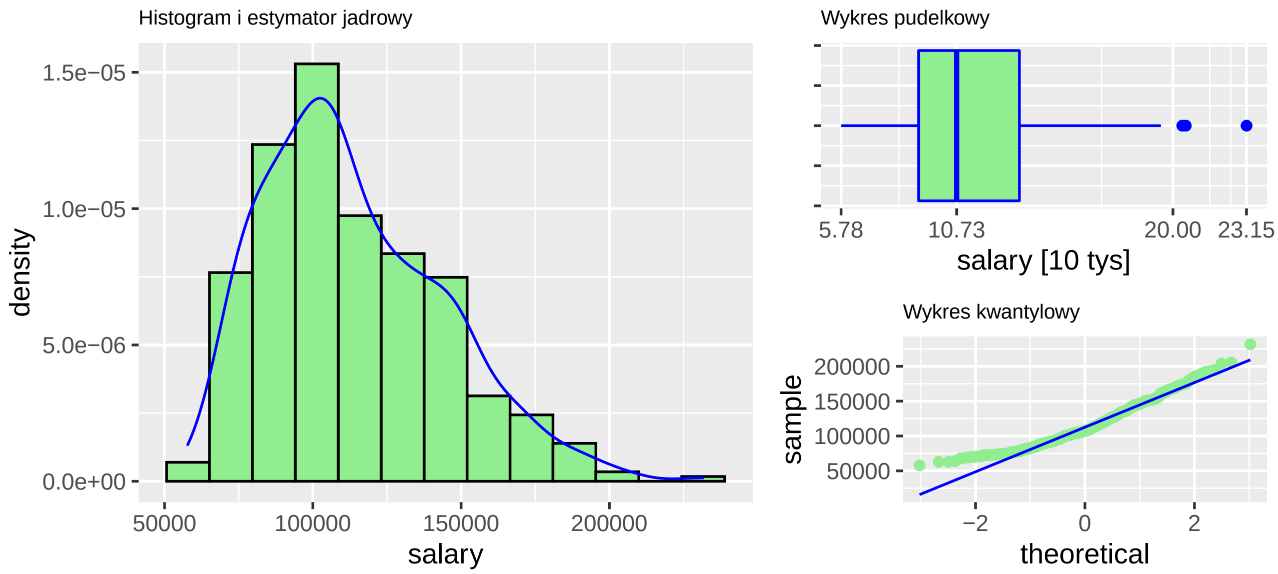
Przeprowadzimy następujące analizy i eksperymenty:

1. analiza danych `Salaries` z pakietu `carData`,
2. estymacja gęstości i badanie własności histogramu,
3. zdefiniowanie i badanie własności dystrybuanty empirycznej.

3 Wyniki

3.1 Analiza danych Salaries z pakietu carData

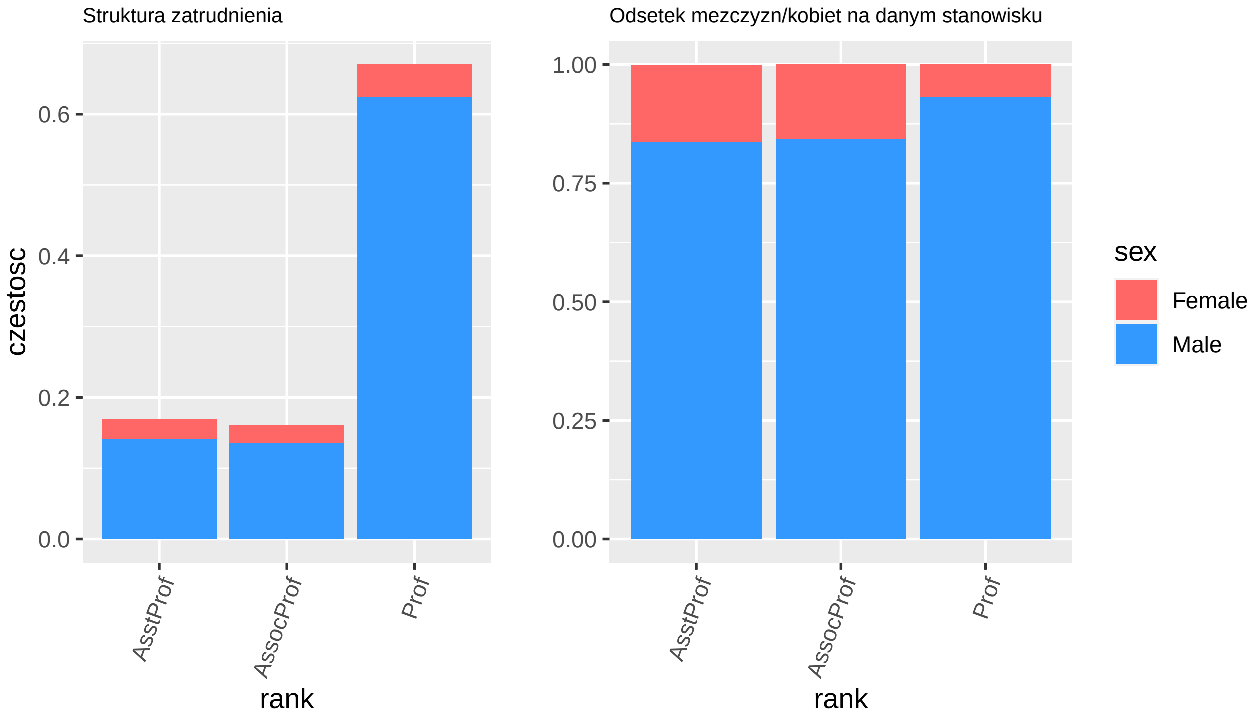
Rozpoczynamy od wczytania danych `Salaries` z pakietu `carData`. Zauważamy, że w danych nie występują brakujące obserwacje, zatem przechodzimy do analizy. Konstruujemy histogram oraz estymator jądrowy dla zmiennej `salary`, która opisuje roczne wynagrodzenie pracownika.



Rysunek 1: Wykresy dla salary

Z histogramu możemy odczytać, że rozkład cechy `salary` jest rozkładem jednomodalnym prawostronnie skośnym. Wykres pudełkowy natomiast wyraźnie wskazał kilka obserwacji odstających (roczna płaca ponad $200,000$). Mediana rocznego wynagrodzenia wynosi natomiast $107,300$, a 75% zatrudnionych zarabia mniej niż $134,200$. Widzimy także z wykresu kwantylowego, że rozkład `salary` nie jest rozkładem normalnym.

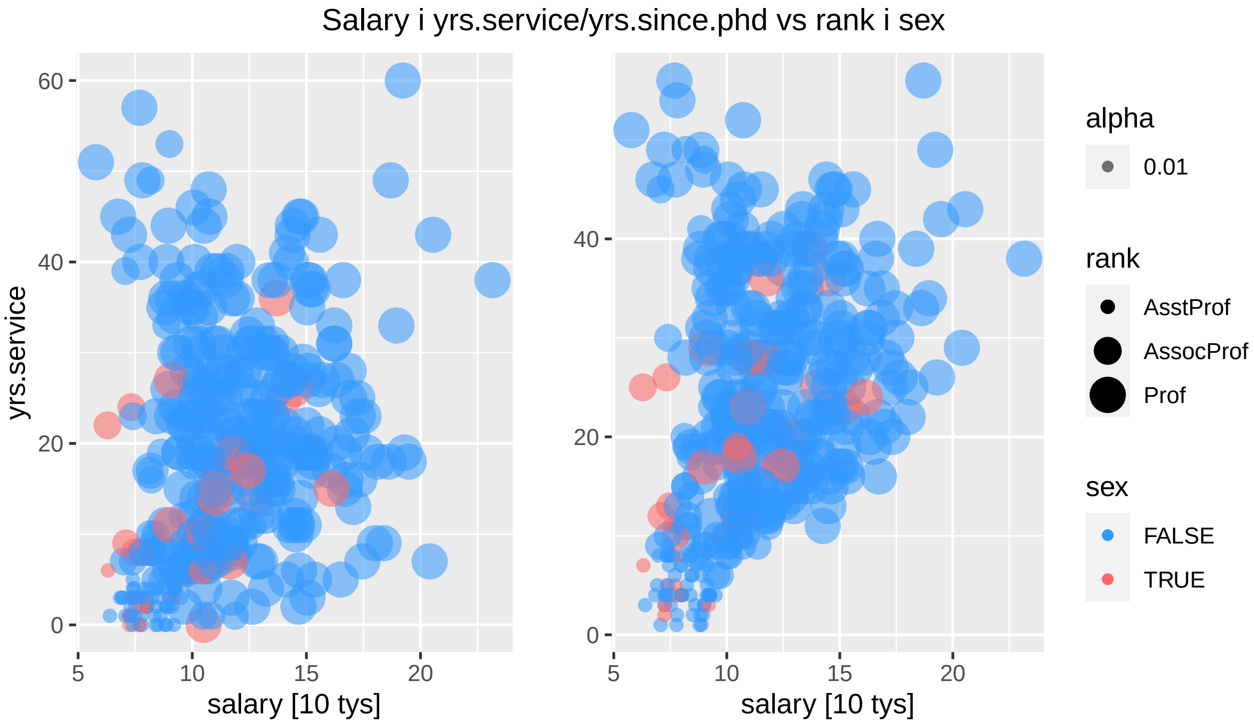
Zajmiemy się teraz analizą rocznych płac ze względu na płeć.



Rysunek 2: Częstość zatrudnienia na danych stanowiskach

Widzimy, że na każdym stanowisku jest zatrudnionych zdecydowanie więcej mężczyzn niż kobiet, natomiast stosunek kobiet do mężczyzn zatrudnionych jako `AsstProf` jest zbliżony do takiego stosunku dla `AssocProf`. Zauważalna różnica występuje natomiast w stosunku kobiet do mężczyzn na stanowisku `Prof`.

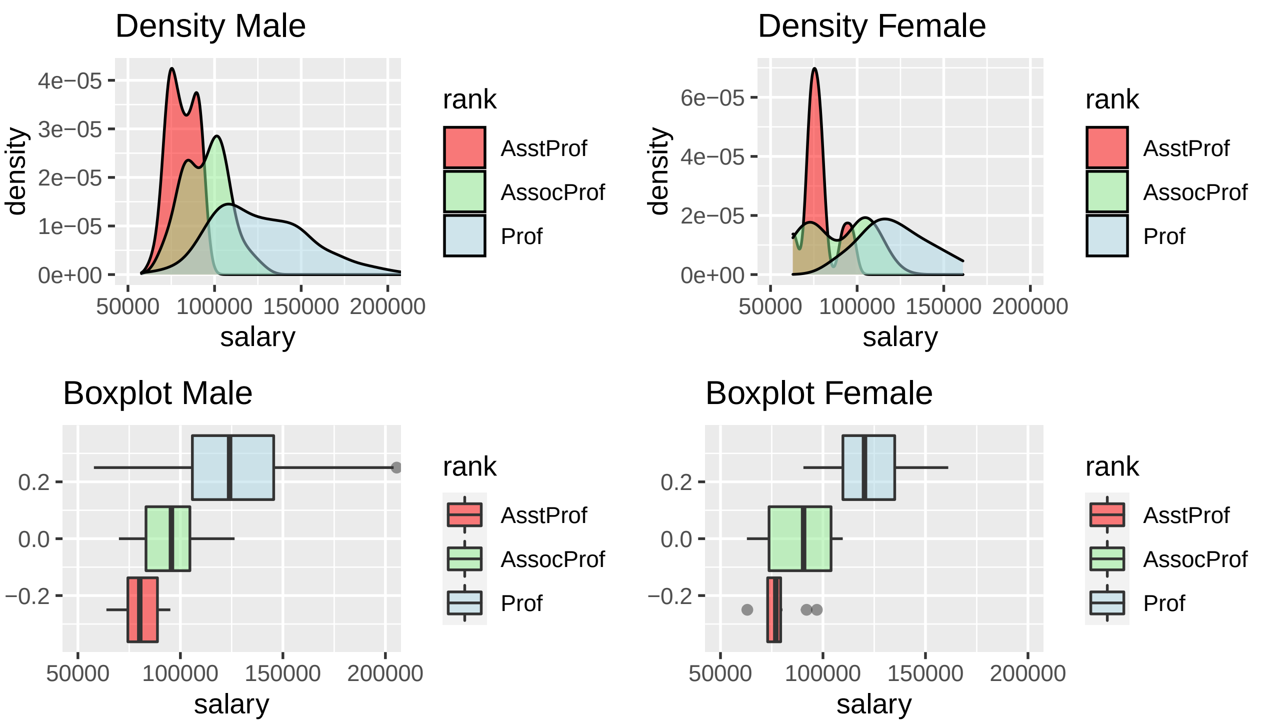
Sprawdzimy również, czy występują jakieś zależności pomiędzy wynagrodzeniem, a stażem pracy i czasem, który upłynął od doktoratu, również w zależności od płci.



Rysunek 3: Wykresy rozrzutu

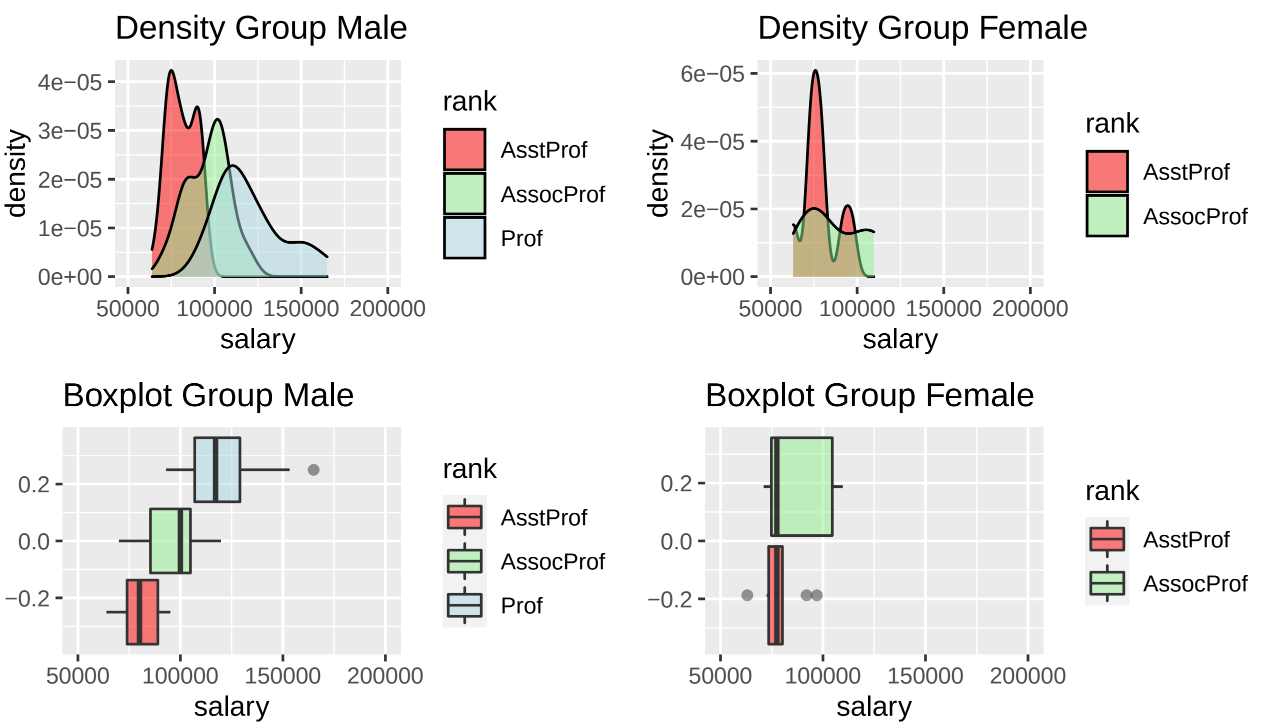
Widzimy, że w przypadku kobiet występują obserwacje odstające – przykładowo na stanowisku `AssocProf` wypłata dla kobiet jest niższa nawet pomimo zbliżonego stażu oraz upłyniętego czasu od uzyskania doktoratu. Nie widać natomiast dużych rozbieżności na stanowisku `AsstProf`. W większości obserwowanych przypadków wypłata kobiet jest zauważalnie bliżej dolnej granicy wynagrodzeń.

Narysujemy teraz wykresy pudełkowe i gęstości empiryczne dla wynagrodzeń w zależności od płci oraz zajmowanego stanowiska.



Rysunek 4: Wykresy pudełkowe i gęstości z podziałem na sex i rank

Rysunek 4 potwierdza nasze przypuszczenia, że występują wyraźne róznice w płacach dla `AssocProf`. Również dla pozostałcyh stanowisk zauważalne są różnice, jednak nie aż tak drastyczne. Warto sprawdzić, czy wpływu na te różnice nie mają inne zmienne. Sprawdzimy więc, jak ma się wynagrodzenie w grupie pracowników, których staż to od 0 do 10 lat oraz uzyskali doktorat od 0 do 20 lat temu.



Rysunek 5: Wkresy pudełkowe i gęstości w grupie: staz: [0,10], doktorat: [0,20]

Odczytujemy, że w tej grupie mediana wynagrodzeń kobiet na stanowisu `AssocProf` wynosi 77500, a mediana wynagrodzeń mężczyzn jest równa 100000. Mediany dla kobiet i mężczyzn dla `AsstProf` wynoszą kolejno 77500 i 80027. Obserwujemy również w tej grupie brak kobiet na stanowisku `Prof`. Możemy zaobserwować, że nawet pomimo zbliżonego stażu, czasu od uzyskania doktoratu oraz stanowiska na uczelni, płace drastycznie się różnią. Szczególnie widać to wśród osób zatrudnionych na stanowisku `AssocProf`. Interesujący jest również fakt, że w grupie kobiet mediana wynagrodzeń `AsstProf` jest identyczna jak dla `AssocProf`, zatem nawet pomimo awansu kobiety nie mogą liczyć na zauważalną podwyżkę.

Na podstawie zebranych danych i powyższych analiz możemy wyciągnąć wstępny wniosek, że na tej uczelni występuje dyskryminacja płacowa pod względem płci. Nawet w przypadku porównywania osób o podobnym doświadczeniu i czasie od uzyskania doktoratu wyraźnie zauważalne są różnice, których nie możemy w inny sposób wytłumaczyć. Warto byłoby również zastanowić się nad ewentualnym występowaniem dyskryminacji w procesie promotorskim – znacznie częściej mężczyźni niż kobiet awansowali na stanowisko `Prof`, co również przyczynia się do różnic płacowych.

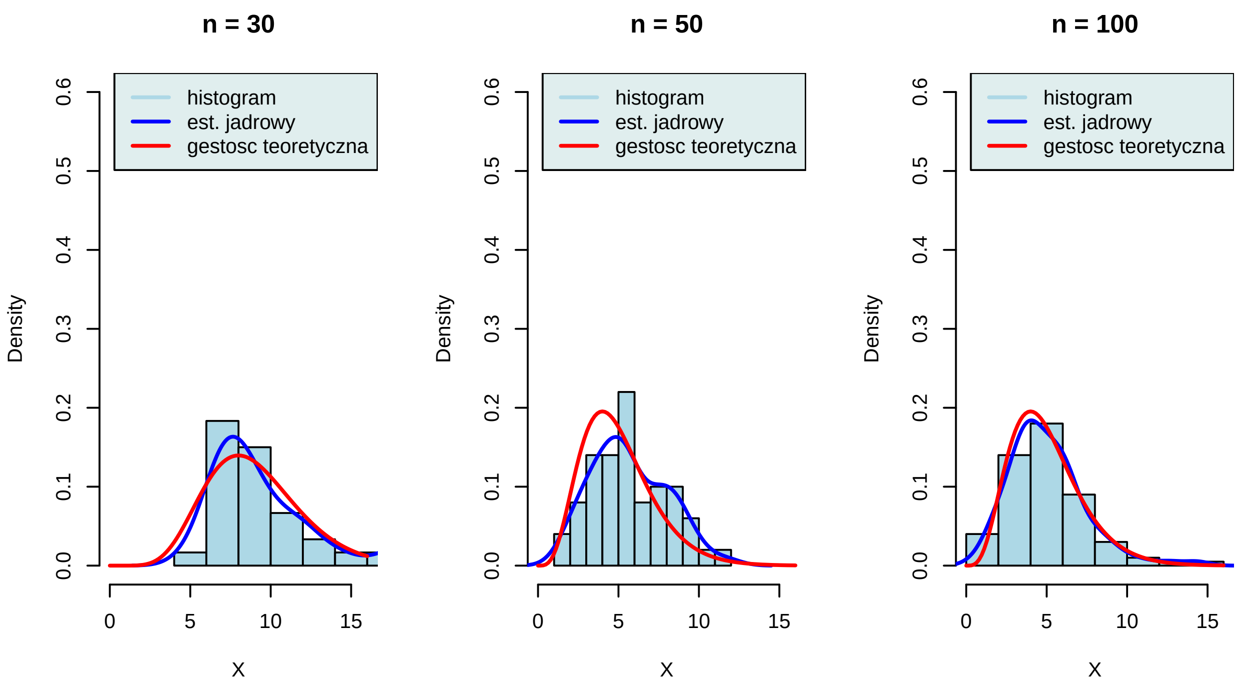
3.2 Estymacja gęstości i badanie własności histogramu

Estymator jądrowy definiujemy jako

$$\hat{f}_n(t) = \frac{1}{n\lambda_n} \sum_{i=1}^{n} K\left(\frac{t - X_i}{\lambda_n}\right), \tag{1}$$

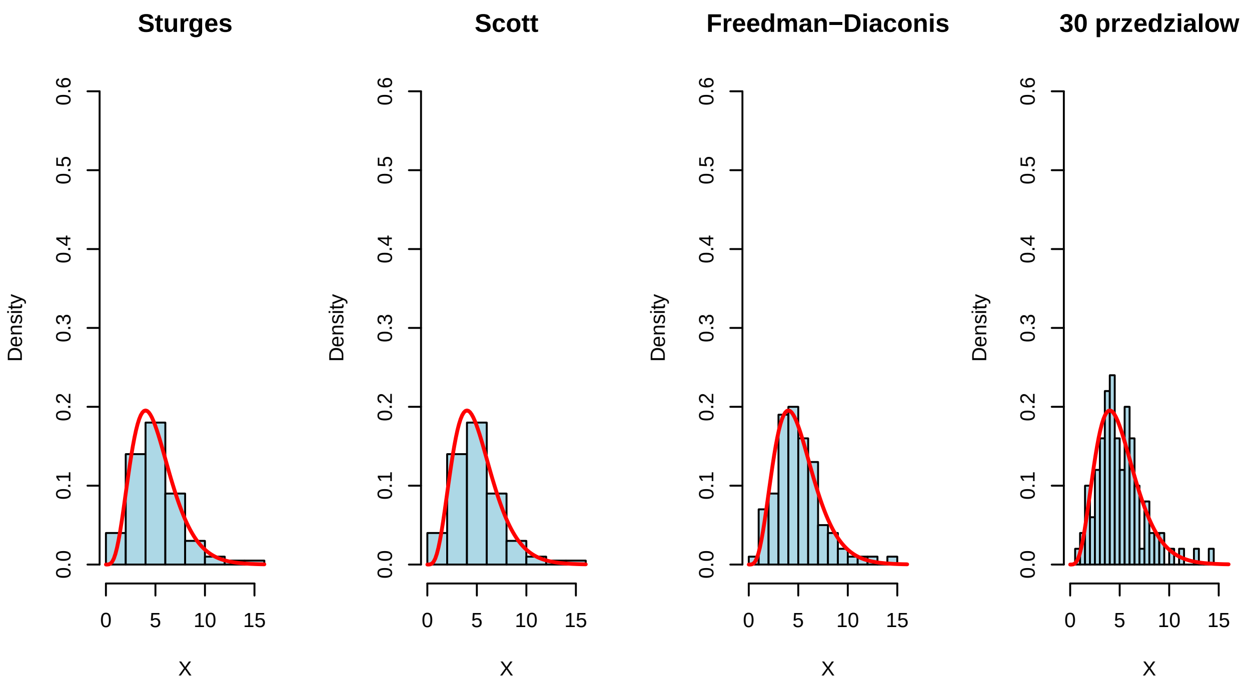
gdzie $K(t)$ to jądro. Zbadamy, jak zachowuje się estymator jądrowy w zależności od liczności próby, wyboru jądra oraz szerokości okna λ_n.

W tym zagadnieniu będziemy rozpatrywali rozkład Gamma $\mathcal{G}(9, 1)$. Wygenerujemy n-elementową próbę z tego rozkładu, gdzie $n \in (30, 50, 100)$, a następnie sporządzimy histogramy oraz estymatory jądrowe i przyrównamy je do gęstości teoretycznej.



Rysunek 6: Histogramy i estymatory jądrowe dla n prób

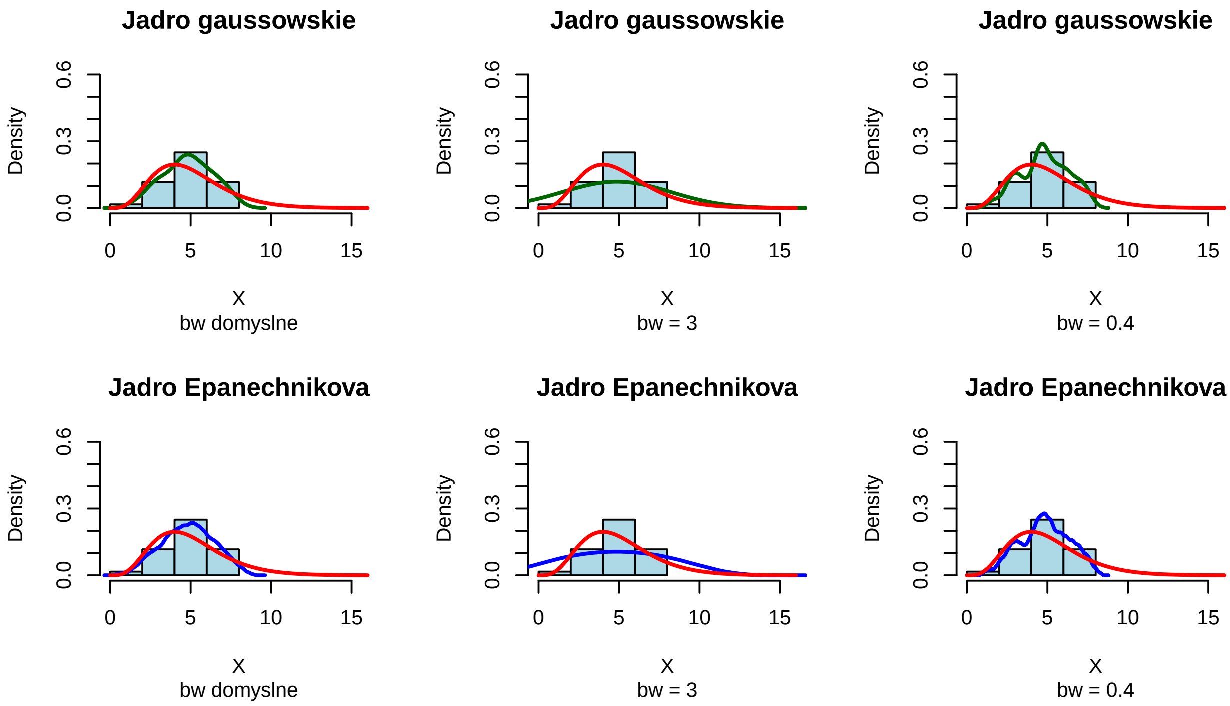
Widzimy, że wraz ze wzrostem obserwacji estymator jądrowy coraz lepiej przybliża gęstość teoretyczną. Zauważamy także, że w zależności od n zmienia się liczba przedziałów klasowych histogramu. Zbadamy więc, jak zachowuje się histogram dla ustalonej, 100-elementowej próby $\mathbf{X}$, w zależności od doboru przedziałów klasowych.



Rysunek 7: Histogramy dla różnych algorytmów wyboru przedziałów klasowych

Wraz ze wzrostem liczby przedziałów klasowych mogą występować gwałtowne skoki w histogramie. Szczególnie dobrze widać to dla przypadku z 30 przedziałami klasowymi. Wybór mniejszej liczby klas pozwala zapobiegać takim skokom kosztem wygładzenia histogramu.

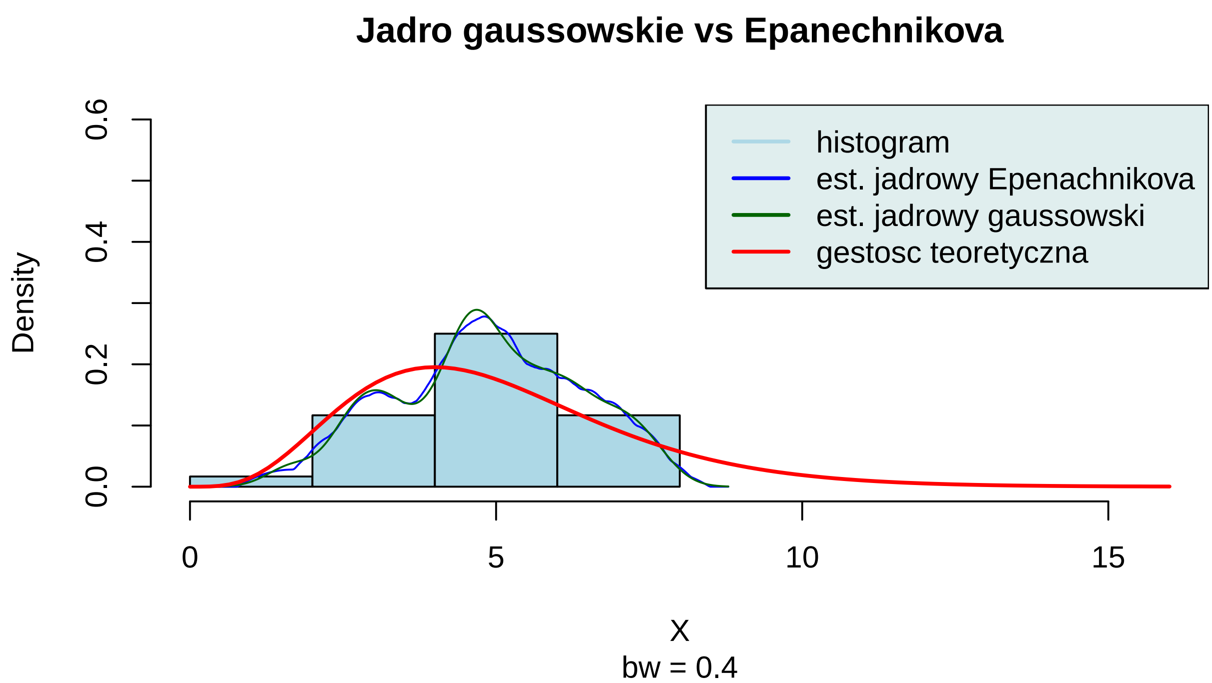
Sprawdzimy teraz jak zachowuje się estymator gęstości dla 30-elementowej próby $\mathbf{X}$ w zależności od szerokości okna (parametr `bw`).



Rysunek 8: Estymatory jądrowe dla różnych wyborów jądra i szerokości okna

Wraz ze wzrostem parametru `bw` (> 1) wykres gęstości spłaszcza się oraz wygładza, natomiast jeśli parametr maleje (< 1), to obserwujemy pojawianie się wielu lokalnych ekstremów oraz punktów przegięcia wykresu.

Sprawdzimy jeszcze, czy wybór jądra również ma tak drastyczny wpływ na estymowany kształt gęstości.



Rysunek 9: Porównanie estymatorów jądrowych

Różnice między estymatorami z różnymi jądrami są niemal niezauważalne. Stąd możemy wywnioskować, że zdecydowanie istotniejszym paramterem przy estymowaniu gęstości jest parametr wygładzenia λ_n.

3.3 Zdefiniowanie i badanie własności dystrybuanty empirycznej

Dystrybuanta empiryczna jest definiowana jako

$$F_n(t, \mathbf{X}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty,t]}(X_i). \tag{2}$$

Napiszemy R-funkcję `demp_plot`, która dla danego wektora $\mathbf{X}$ narysuje dystrybuantę empiryczną, dystrybuantę rozkładu normalnego $\mathcal{N}(0, 1)$, wyznaczy odległość Kołmogorowa $D_n(\mathbf{X})$ oraz zaznaczy ją na wykresie. Do rysowania wykresu wykorzystujemy pakiet `ggplot2`.

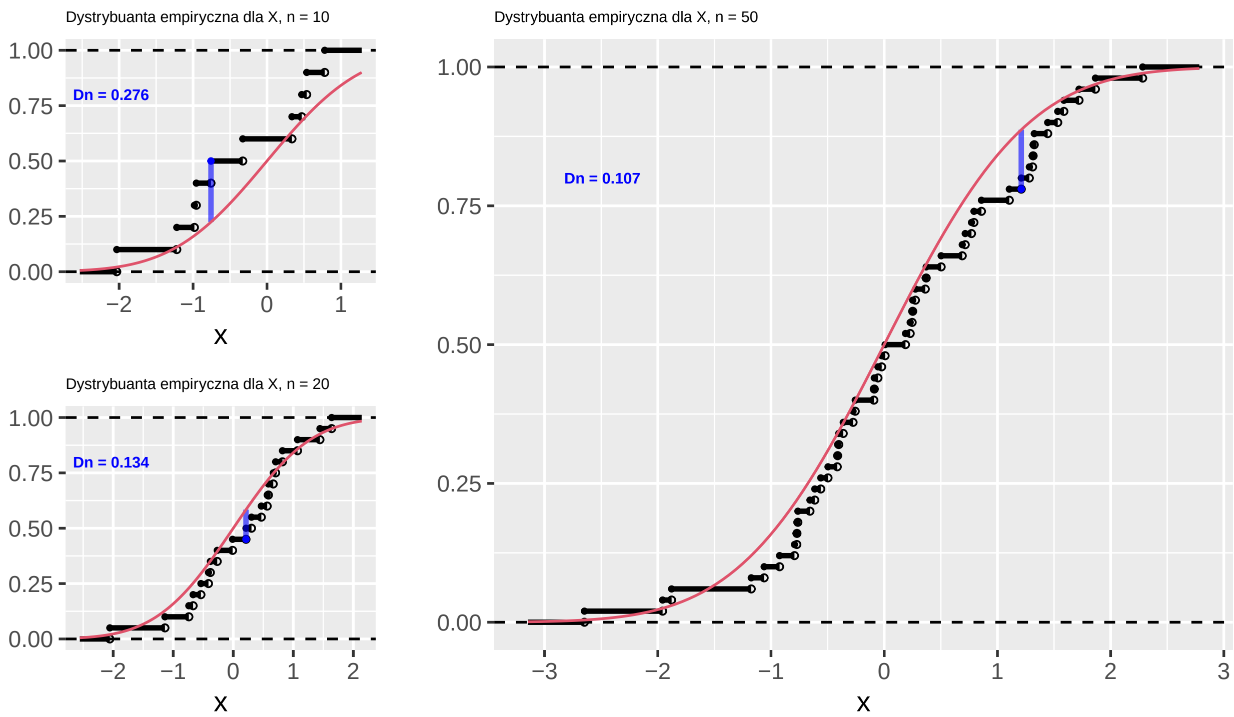
```
demp_plot <- function(X,
                      a = min(X) - 0.5,
                      b = max(X) + 0.5,
                      cc = "black")
{
  n <- length(X) #długość wektora X
  s_poz <- sort(X) #definicja statystyki pozycyjnej
  y1 <- 0 #zdefiniowanie pierwszej wartości y
  V <-
    c() #V, W1, W2 - wektory gromadzące współrzędne punktóW końcowych odcinków
  W1 <- c()
  W2 <- c()
  title <- paste("Dystrybuanta empiryczna dla X, n =", toString(n))
  M <-
    ggplot(data =  data.frame(x = c(-a, b), y = c(-0.1, 1.1)), aes(x = x, y =
                                                                     y))
  for (k in c(1:(n + 1))) {
    if (k == 1) {
      t1 <- a
      t2 <- s_poz[k]
    } else if (k == n + 1) {
      t1 <- s_poz[k - 1]
      t2 <- b
      V <- append(V, t1)
      W1 <- append(W1, y1)
      W2 <- append(W2, y1 + 1 / n)
      y1 <- y1 + 1 / n
    } else {
      t1 <- s_poz[k - 1]
      t2 <- s_poz[k]
      V <- append(V, t1)
      W1 <- append(W1, y1)
      W2 <- append(W2, y1 + 1 / n)
      y1 <- y1 + 1 / n
```

```
  }
  M <- M + #rysowanie linii dla jednego odcinka na wysokości i/n
    annotate(
      "segment",
      x = t1,
      xend = t2,
      y = y1,
      yend = y1,
      colour = cc,
      size = 1
    )
}
geom_graph.1 <- data.frame(V, W1)
geom_graph.2 <- data.frame(V, W2)
M <- M + #rysowanie punktów końcowych odcinków
  geom_point(
    data = geom_graph.1,
    aes(x = V, y = W1),
    colour = cc,
    size = 1,
    shape = 1
  ) +
  geom_point(
    data = geom_graph.2,
    aes(x = V, y = W2),
    colour = cc,
    size = 1,
    shape = 16
  ) +
  geom_hline(yintercept = 0,
             color = "black",
             linetype = "dashed") +
  geom_hline(yintercept = 1,
             color = "black",
             linetype = "dashed") +
  ggtitle(title) +
  theme(plot.title = element_text(size = 6)) +
  xlim(c(a, b))
D1 <- c() #wektory służące do wyznaczenia Dn
D2 <- c()
for (i in c(1:n)) {
  D1 <- append(D1, i / n - pnorm(s_poz[i]))
  D2 <- append(D2, pnorm(s_poz[i]) - (i - 1) / n)
}
Dn <- max(max(D1), max(D2)) #zdefiniowanie Dn
if (Dn %in% D1) {
  #zaznaczenie Dn na wykresie
  k <- which(sapply(
```

```
    D1,
    FUN = function(X)
      Dn %in% X
  ))
  M <-
    M + geom_point(
      aes(x = s_poz[k], y = k / n),
      colour = "blue",
      shape = 16,
      size = 1
    ) +
    annotate(
      "segment",
      x = s_poz[k],
      xend = s_poz[k],
      y = (k) / n,
      yend = pnorm(s_poz[k]),
      colour = "blue",
      size = 1,
      alpha = 0.6
    )
} else {
  k <- which(sapply(
    D2,
    FUN = function(X)
      Dn %in% X
  ))
  M <-
    M + geom_point(
      aes(x = s_poz[k], y = (k - 1) / n),
      colour = "blue",
      shape = 16,
      size = 1
    ) +
    annotate(
      "segment",
      x = s_poz[k],
      xend = s_poz[k],
      y = (k - 1) / n,
      yend = pnorm(s_poz[k]),
      colour = "blue",
      size = 1,
      alpha = 0.6
    )
}
text <- paste("Dn =", toString(round(Dn, 3)))
annotation <- data.frame(x = c(a + abs(b - a) / 9),
                         y = c(0.8),
```

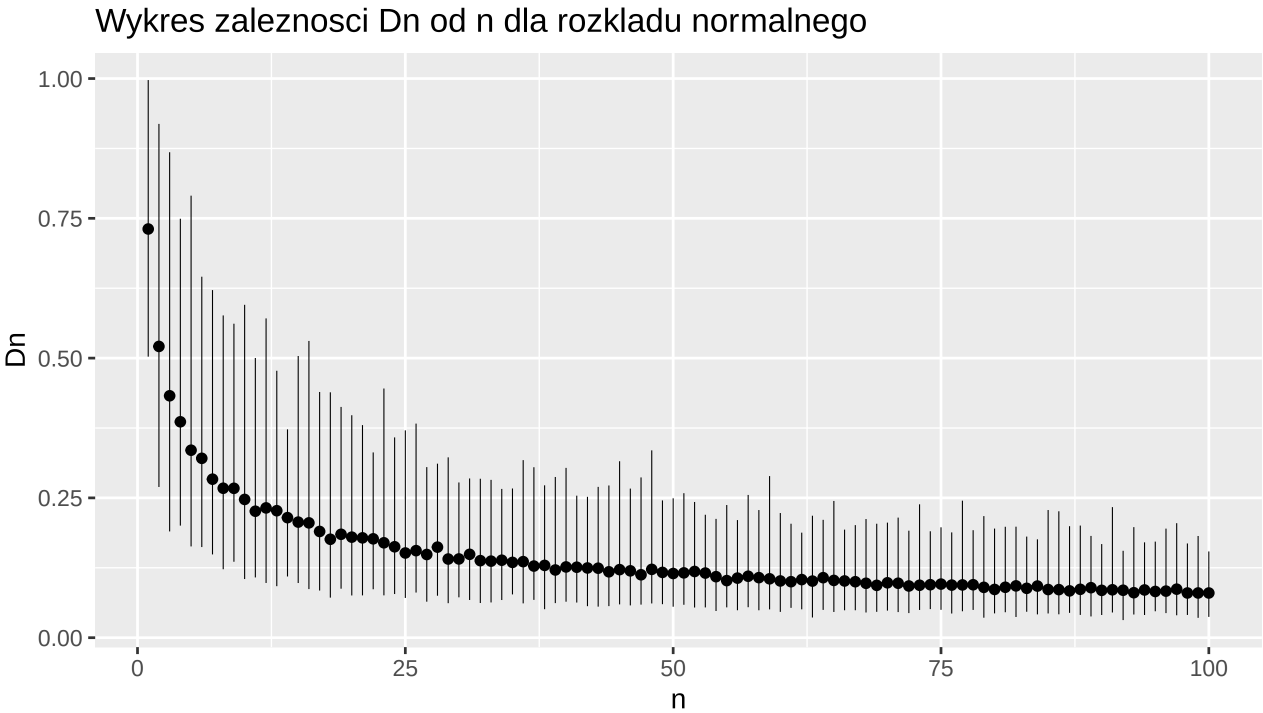
```
                                  label = c(text))
  M <-
    M + geom_text(
      data = annotation,
      aes(x = x, y = y, label = label),
      ,
      color = "blue",
      size = 2,
      fontface = "bold"
    ) + ylab("Fn") + xlab("x") #wyświetlenie Dn na wykresie
  return(M)
}
```

Wygenerujemy teraz trzy próby losowe z rozkładu normalnego. Pierwsza będzie długości 10, druga 20, a trzecia 50. Korzystajmy z funkcji `demp_plot` i otrzymujemy:



Rysunek 10: Porównanie dystrybuant empirycznych i teoretycznych dla n-elementowej próby

Zauważamy, że wraz ze zwiększeniem liczby obserwacji odległość $D_n(\mathbf{X})$ maleje. Sprawdzimy zatem zależność D_n od n. Wyciągając z poprzednio napisanej funkcji fragment kodu odpowiedzialny za wyznaczenie D_n, narysujemy wykres dla badanej zależności. Weźmiemy również pod uwagę błąd pomiaru D_n uzyskany doświadczalnie.



Rysunek 11: Odległość Kołmogorowa i jej błąd w zależności od n

Widzimy, że w przybliżeniu D_n wykładniczo maleje do zera. Wraz ze wzrostem n maleje również błąd pomiaru.

4 Podsumowanie

Poniżej wypunktujemy najważniejsze wnioski, jakie można wyciągnąć z przeprowadzanych analiz:

- narysowanie podstawowych wykresów pozwala nam wstępnie zaobserwować zależności lub ich brak, aby następnie dogłębnie przeanalizować ewentualnie związki między cechami,
- analiza danych w poszczególnych grupach pozwala na dokładniejsze zbadanie problemu, uwzględniając podobieństwa i różnice w wybranych zmiennych,
- wybór jądra w estymatorze jądrowym ma marginalne znaczenie w przeciwieństwie do doboru parametru λ_n, który jest kluczowy dla estymowanego kształtu gęstości,
- wraz ze wzrostem liczby obserwacji statystyka Kołmogorowa D_n znacząco maleje, aż w końcu $\lim_{n\to\infty} D_n = 0$.