

Zadanie 4. Napisz funkcję zamieniającą miarę kąta podaną w stopniach na radiany. Sprawdź działanie tej funkcji dla kątów o mierze: 0°, 30°, 45°, 60°, 90°. Następnie przygotuj ramkę danych, w której zebrane będą informacje o wartościach funkcji sinus, cosinus, tangens i cotangens dla kątów o takich miarach.

```
## [1] 0.0000000 0.5235988 0.7853982 1.0471976 1.5707963

##          sin          cos          tg          ctg
## 1 0.0000000 1.000000e+00 0.000000e+00          Inf
## 2 0.5000000 8.660254e-01 5.773503e-01 1.732051e+00
## 3 0.7071068 7.071068e-01 1.000000e+00 1.000000e+00
## 4 0.8660254 5.000000e-01 1.732051e+00 5.773503e-01
## 5 1.0000000 6.123234e-17 1.633124e+16 6.123234e-17
```

Zadanie 5. Napisz funkcję, której argumentem będzie wektor liczbowy a wynikiem wektor zawierający trzy najmniejsze i trzy największe liczby w tym wektorze. W przypadku argumentu krótszego niż trzy liczby, funkcja ma zwracać komunikat o błędzie z komentarzem „za krótki argument”.

```
# dla
x <- c(2, 6, 1, 5, 7, 3, 4)
```

```
## [1] 1 2 3 5 6 7
```

```
# dla
x <- c(2, 6)
## Error in command 'extreme_3(x)': za krótki argument
```

4 Statystyka opisowa

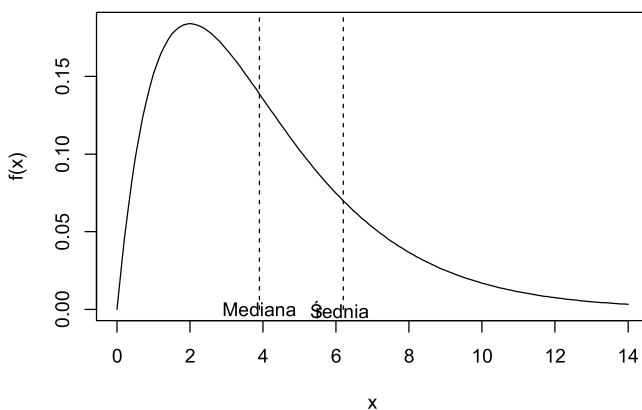
4.1 Miara asymetrii rozkładu

- współczynnik asymetrii (skośności)

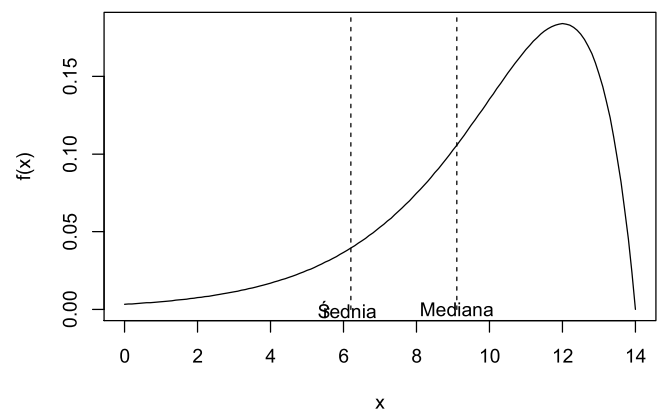
$$A = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{s^3}$$

- Współczynnik asymetrii
 - równy zero oznacza symetrię rozkładu zmiennej.
 - przyjmujący wartość dodatnią oznacza prawostronną asymetrię. Prawy ogon jest dłuższy, a masa rozkładu jest skoncentrowana po lewej stronie.
 - przyjmujący wartość ujemną oznacza lewostronną asymetrię. Lewy ogon jest dłuższy, a masa rozkładu jest skoncentrowana po prawej stronie.

Prawostronna asymetria



Lewostronna asymetria

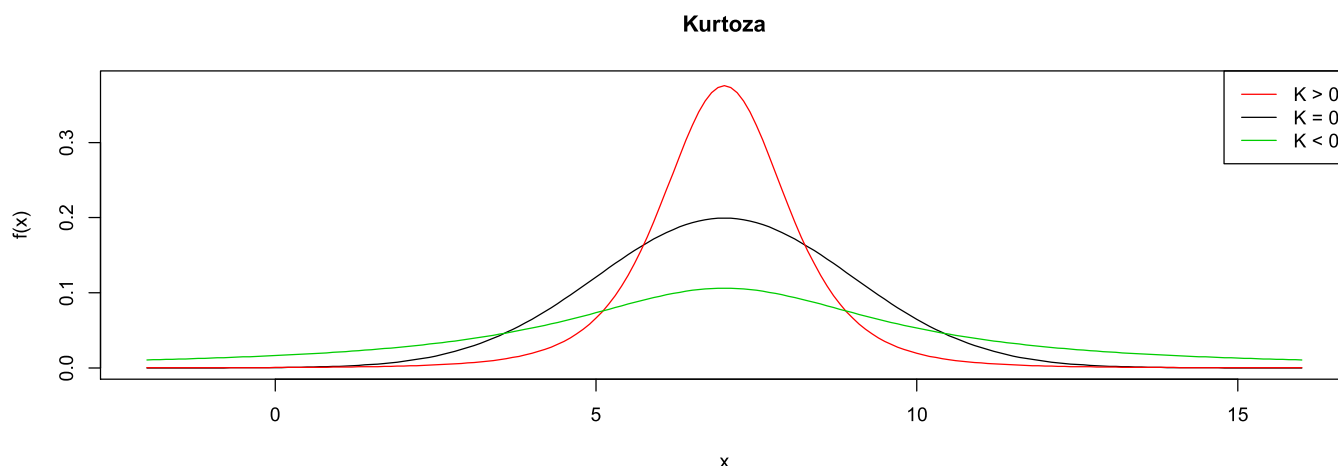


4.2 Miara koncentracji rozkładu

- kurtoza

$$K = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{s^4} - 3$$

- Kurtoza jest miarą skupienia wartości zmiennej wokół średniej.
- Porównuje ona badany rozkład empiryczny z rozkładem normalnym i przyjmuje wartości większe niż -2 .
- Im większa wartość K , tym większe skupienie wartości zmiennej wokół średniej.
- Kurtoza rozkładu normalnego wynosi zero.
- Jeśli $K < 0$, wówczas rozkład jest bardziej spłaszczony niż rozkład normalny, a jeśli $K > 0$ - bardziej smukły.



4.3 Przykłady

Przykład 1. Poniższe dane podają liczbę błędów w grupie 50 osób zdających egzamin testowy. Egzamin składał się z 18 pytań (można dopełnić maksymalnie dwa błędy, aby zdać egzamin).

```
1 1 2 0 1 3 1 4 4 4 0 1 0 0 0 2 3
4 0 1 5 2 3 5 3 2 2 4 0 2 2 0 2 2
3 3 1 3 2 2 0 0 5 4 2 1 5 2 2 0
```

Zmienna X to liczba błędów. Jest to dyskretna zmienna ilościowa.

```
liczba_bledow <- c(1, 1, 2, 0, 1, 3, 1, 4, 4, 4, 0, 1, 0, 0, 0, 2, 3,
                  4, 0, 1, 5, 2, 3, 5, 3, 2, 2, 4, 0, 2, 2, 0, 2, 2,
                  3, 3, 1, 3, 2, 2, 0, 0, 5, 4, 2, 1, 5, 2, 2, 0)
# rozkład empiryczny opisany za pomocą szeregu rozdzielczego
data.frame(cbind(liczebosc = table(liczba_bledow),
                 procent = prop.table(table(liczba_bledow))))
```

```
##   liczebosc procent
## 0         11   0.22
## 1          8   0.16
## 2         14   0.28
## 3          7   0.14
## 4          6   0.12
## 5          4   0.08
```

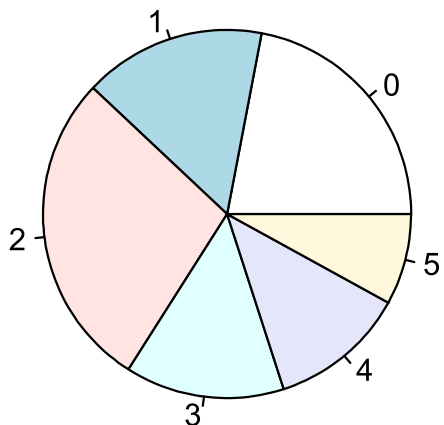
```
# wykres słupkowy
barplot(table(liczba_bledow),
        xlab = "Liczba błędów", ylab = "Liczebność",
        main = "Rozkład empiryczny liczby błędów")
```



```
barplot(prop.table(table(liczba_bledow)),
        xlab = "Liczba błędów", ylab = "Prawdopodobieństwo",
        main = "Rozkład empiryczny liczby błędów")
```



```
# wykres kołowy
pie(table(liczba_bledow))
```



```
# Średnia
mean(liczba_bledow)
```

```
## [1] 2.02
```

```
# mediana
median(liczba_bledow)
```

```
## [1] 2
```

```
# odchylenie standardowe
sd(liczba_bledow)
```

```
## [1] 1.558256
```

```
# współczynnik zmienności
sd(liczba_bledow) / mean(liczba_bledow) * 100
```

```
## [1] 77.14141
```

Przykład 2. Badano czas oczekiwania na tramwaj, który kursuje w jednakowych odstępach czasu. Plik `czas_oczek_tramwaj.RData` zawiera dane dotyczące czasu oczekiwania na tramwaj (wyrażonego w minutach) 100 osób wybranych losowo. Zmienna X to czas oczekiwania na tramwaj. Jest to zmienna ilościowa ciągła.

```
load(url("http://ls.home.amu.edu.pl/data_sets/czas_oczek_tramwaj.RData"))
head(czas_oczek_tramwaj)
```

```
## [1] 4.03 11.04 5.73 12.36 13.17 0.64
```

```
data.frame(cbind(liczebosc = table(cut(czas_oczek_tramwaj, breaks = seq(0, 14, 2))),
                 procent = prop.table(table(cut(czas_oczek_tramwaj, breaks = seq(0, 14, 2))))))
```

```
##      liczebosc procent
## (0,2]         15   0.15
## (2,4]         13   0.13
## (4,6]         15   0.15
## (6,8]         15   0.15
## (8,10]        15   0.15
## (10,12]       12   0.12
## (12,14]       15   0.15
```

```
(czas_oczek_tramwaj_hist <- hist(czas_oczek_tramwaj, plot = FALSE)$breaks)
```

```
## [1] 0 2 4 6 8 10 12 14
```

```
data.frame(cbind(liczebosc = table(cut(czas_oczek_tramwaj, breaks = czas_oczek_tramwaj_hist)),
                 procent = prop.table(table(cut(czas_oczek_tramwaj, breaks = czas_oczek_tramwaj_hist))
```

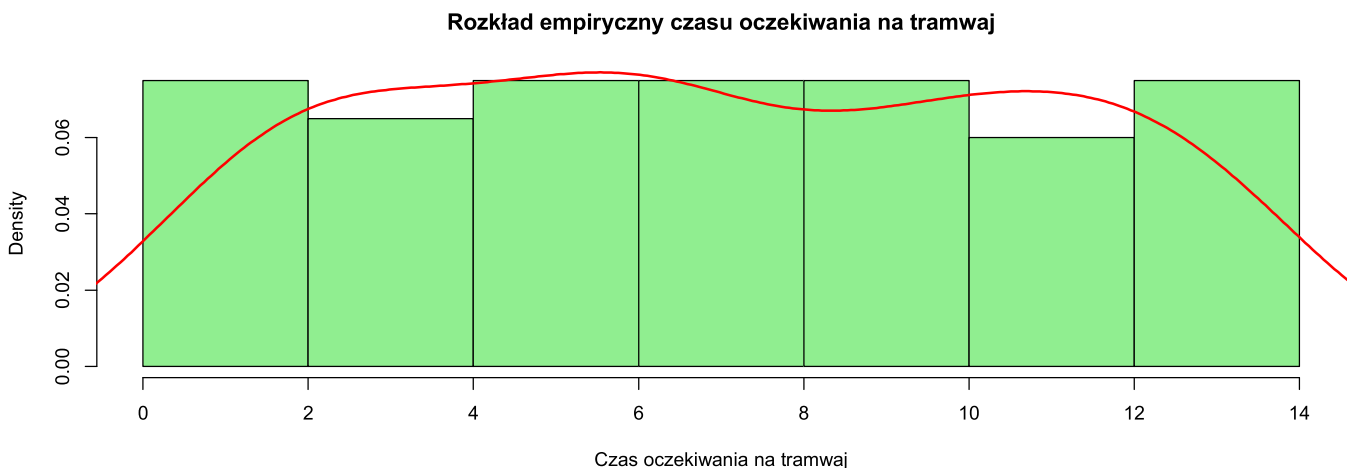
```
##      liczebosc procent
## (0,2]         15   0.15
## (2,4]         13   0.13
## (4,6]         15   0.15
## (6,8]         15   0.15
## (8,10]        15   0.15
## (10,12]       12   0.12
## (12,14]       15   0.15
```

Histogram - zestaw sąsiadujących prostokątów, których podstawy, równe rozpiętości przedziałów klasowych, znajdują się na osi odciętych, a wysokości są liczebnościami przedziałów.

```
# histogram
hist(czas_oczek_tramwaj,
     xlab = "Czas oczekiwania na tramwaj",
     main = "Rozkład empiryczny czasu oczekiwania na tramwaj")
rug(jitter(czas_oczek_tramwaj))
```



```
# histogram z estymatorem jądrowym gęstości
hist(czas_oczek_tramwaj,
     xlab = "Czas oczekiwania na tramwaj",
     main = "Rozkład empiryczny czasu oczekiwania na tramwaj",
     probability = TRUE,
     col = "lightgreen")
lines(density(czas_oczek_tramwaj), col = "red", lwd = 2)
```



Wykres ramkowy to metoda graficznego przedstawienia danych liczbowych za pomocą ich kwantyli. Tworzymy go poprzez umieszczenie na osi pionowej wartości niektórych parametrów rozkładu (kwantyli).

- Wewnątrz prostokąta znajduje się pogrubiona pozioma linia, która określa wartość mediany.
- Nad osią znajduje się prostokąt (ramka), którego dolny bok jest określony przez pierwszy kwartył, a górny bok przez trzeci kwartył. Wysokość pudełka odpowiada wartości rozstępu międzykwartylowego ($Q_3 - Q_1$).
- Pudełko jest uzupełnione od góry i od dołu segmentami (wąsami). Dolny koniec dolnego segmentu

reprezentuje najmniejszą wartość w zestawie danych, zaś górny koniec górnego segmentu jest obserwacją największą. Wartości te muszą spełniać dodatkowy warunek, a mianowicie dolny koniec nie może być mniejszy niż $Q_1 - 1,5 \cdot (Q_3 - Q_1)$, a górny większy niż $Q_3 + 1,5 \cdot (Q_3 - Q_1)$. Jeśli istnieją obserwacje poza tym zakresem, są one zaznaczane na wykresie indywidualnie jako osobne punkty i są traktowane jako obserwacje odstające.

Wykres pudełkowy jako wskaźnik tendencji centralnej, dyspersji, symetrii, skośności i wielkości ogona:

- dyspersja - odstęp między różnymi częściami pudełka
- symetryczny - pogrubiona linia znajduje się blisko środka pudełka, a długości wąsów są takie same
- prawostronnie asymetryczny - górny wąs jest znacznie dłuższy niż dolny wąs, a linia jest bliższa dolnej części pudełka.
- lewostronnie asymetryczny - dolny wąs jest znacznie dłuższy niż górny wąs, a linia jest bliższa górnej części pudełka
- grube ogony - długość wąsów znacznie przekracza długość pudełka
- cienkie ogony - długość wąsów jest krótsza niż długość pudełka
- bardzo krótkie ogony (populacja w kształcie litery U, z zanurzeniem w środku zamiast garbu) - wąsy są nieobecne

```
# wykres ramkowy
boxplot(czas_oczek_tramwaj,
        ylab = "Czas oczekiwania na tramwaj",
        main = "Rozkład empiryczny czasu oczekiwania na tramwaj")
```



- statystyki opisowe

```
# średnia
mean(czas_oczek_tramwaj)
```

```
## [1] 6.9796
```

```
# mediana
median(czas_oczek_tramwaj)
```

```
## [1] 6.525
```

```
# odchylenie standardowe
sd(czas_oczek_tramwaj)
```

```
## [1] 3.989571
```

```
# współczynnik zmienności
sd(czas_oczek_tramwaj) / mean(czas_oczek_tramwaj) * 100
```

```
## [1] 57.16046
```

```
library(e1071)
# współczynnik asymetrii
skewness(czas_oczek_tramwaj)
```

```
## [1] 0.03465377
```

```
# kurtoza
kurtosis(czas_oczek_tramwaj)
```

```
## [1] -1.215931
```

4.4 Zadania

Zadanie 1. Zmienna `wynik` w pliku `ankieta.txt` opisuje wyniki badania działalności prezydenta pewnego miasta. Wybrano losowo 100 mieszkańców miasta i zadano im następujące pytanie: Jak oceniasz działalność prezydenta miasta? Dostępne były następujące odpowiedzi: zdecydowanie dobrze (a), dobrze (b), źle (c), zdecydowanie źle (d), nie mam zdania (e). Jakiego typu jest ta zmienna? Jakie są możliwe wartości tej zmiennej?

1. Zaimportuj dane z pliku `ankieta.txt` do zmiennej `ankieta`.

```
##   plec szkola wynik
## 1    m      p      d
## 2    m      s      e
## 3    m      w      a
## 4    m      s      d
## 5    m      p      c
## 6    m      w      c
## ...
```

2. Przedstaw rozkład empiryczny zmiennej `wynik` za pomocą szeregu rozdzielczego.

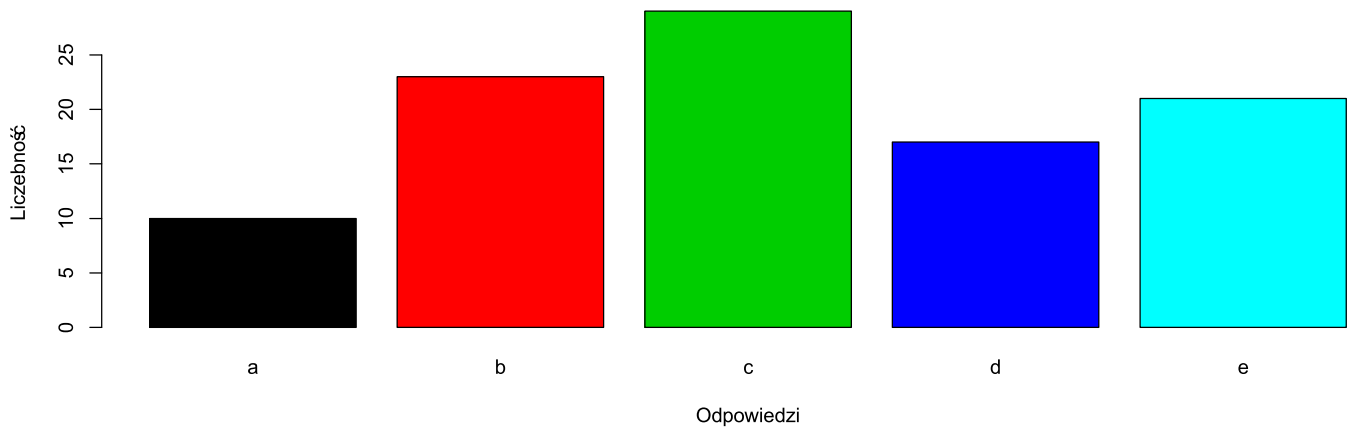
```
##   liczebność  procent
## a           10   0.10
## b           23   0.23
## c           29   0.29
## d           17   0.17
## e           21   0.21
```

3. Przedstaw rozkład empiryczny zmiennej `wynik` tylko dla osób z wykształceniem podstawowym za pomocą szeregu rozdzielczego.

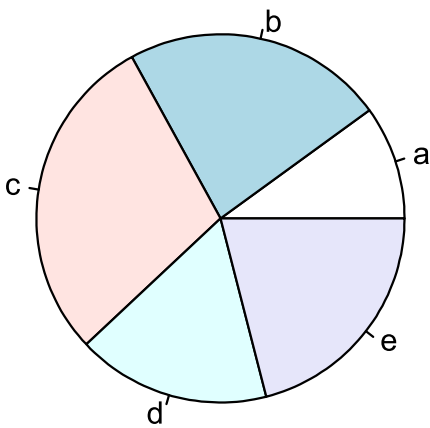
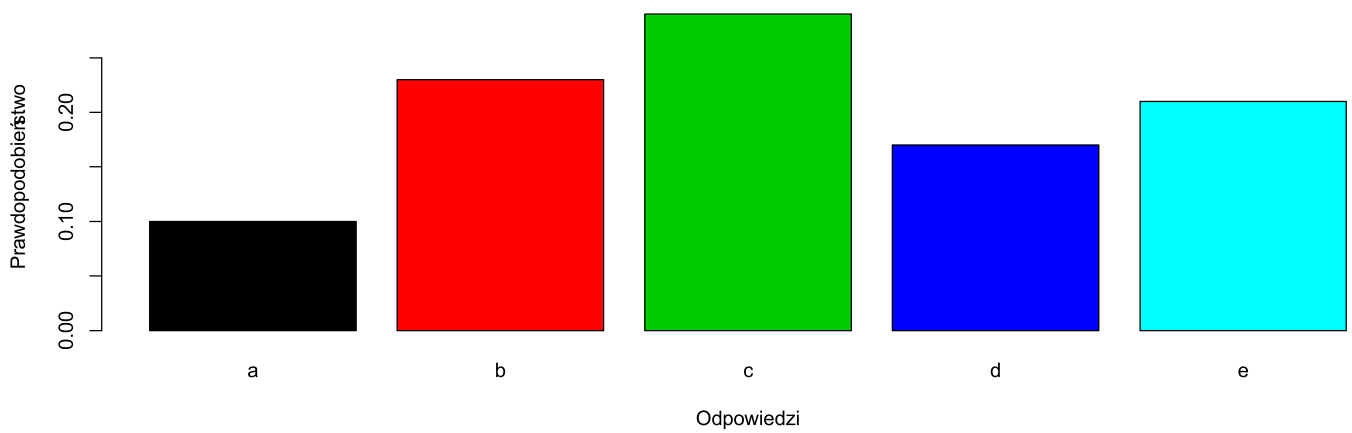
```
##   liczebność  procent
## a           2 0.11764706
## b           3 0.17647059
## c           4 0.23529412
## d           7 0.41176471
## e           1 0.05882353
```

4. Zilustruj wyniki ankiety za pomocą wykresu słupkowego i kołowego.

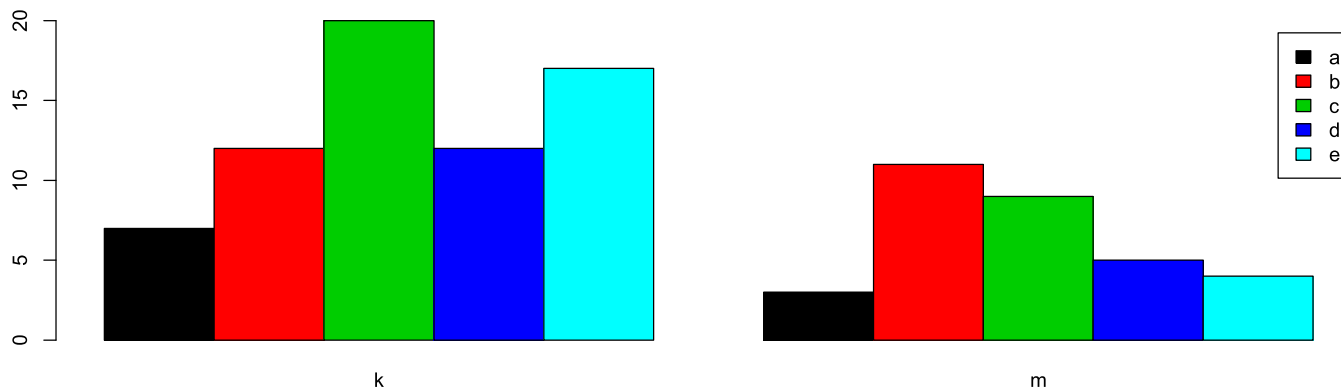
Rozkład empiryczny zmiennej wynik



Rozkład empiryczny zmiennej wynik



5. Zilustruj wyniki ankiety za pomocą wykresu słupkowego z podziałem na kobiety i mężczyzn.



6. Zinterpretuj powyższe wyniki (tabelaryczne i graficzne).

Zadanie 2. Przebadano 200 losowo wybranych 5-sekundowych okresów pracy centrali telefonicznej. Rejestrowano liczbę zgłoszeń. Wyniki są zawarte w pliku Centrala.RData. Jakiego typu jest ta zmienna? Jakie są możliwe wartości tej zmiennej?

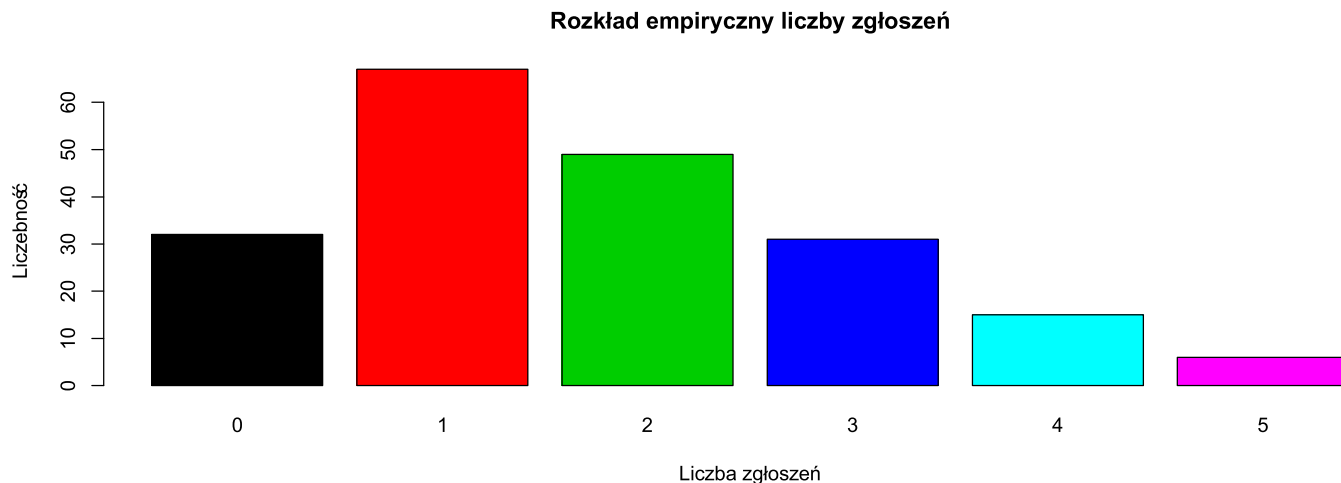
1. Zaimportuj dane z pliku Centrala.RData.

```
## Liczba
## 1      0
## 2      0
## 3      5
## 4      1
## 5      1
## 6      2
## ...
```

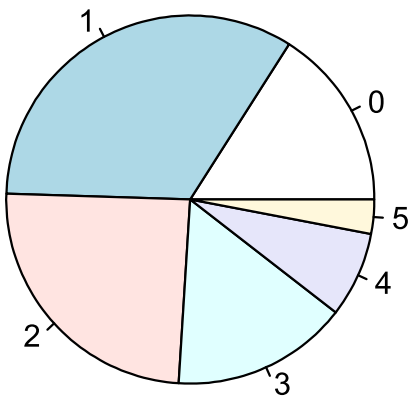
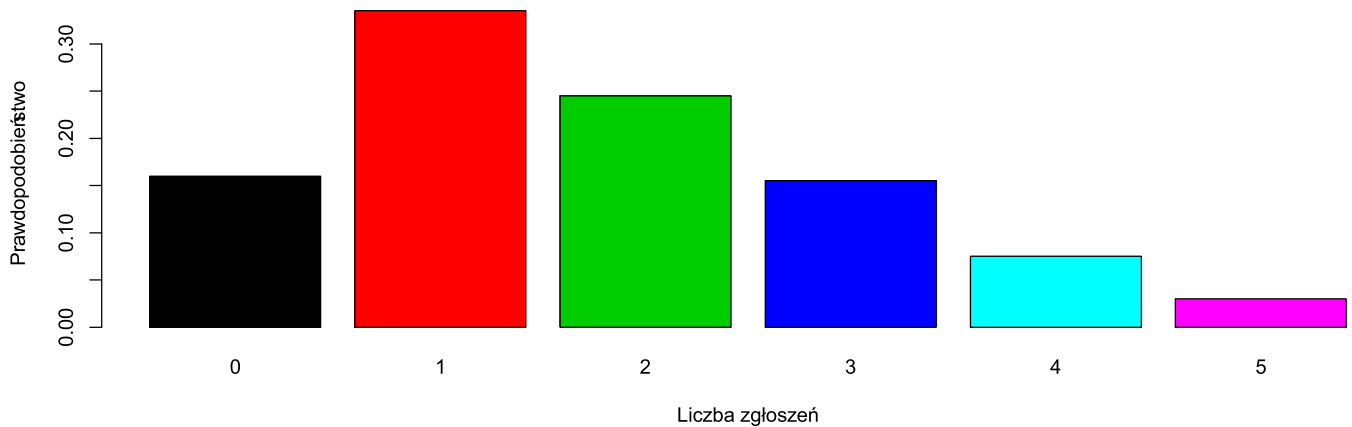
2. Przedstaw rozkład empiryczny liczby zgłoszeń za pomocą szeregu rozdzielczego.

```
## liczebność procent
## 0      32  0.160
## 1      67  0.335
## 2      49  0.245
## 3      31  0.155
## 4      15  0.075
## 5       6  0.030
```

3. Zilustruj liczbę zgłoszeń za pomocą wykresu słupkowego i kołowego.



Rozkład empiryczny liczby zgłoszeń



4. Obliczyć średnią z liczby zgłoszeń, medianę liczby zgłoszeń, odchylenie standardowe liczby zgłoszeń i współczynnik zmienności liczby zgłoszeń.

```
## [1] 1.74
## [1] 2
## [1] 1.28086
## [1] 73.61266
```

5. Zinterpretuj powyższe wyniki (tabelaryczne, graficzne i liczbowe).

Zadanie 3. Notowano pomiary średniej szybkości wiatru w odstępach 15 minutowych wokół nowo powstającej elektrowni wiatrowej. Wyniki są następujące:

0.9	6.2	2.1	4.1	7.3
1.0	4.6	6.4	3.8	5.0
2.7	9.2	5.9	7.4	3.0
4.9	8.2	5.0	1.2	10.1
12.2	2.8	5.9	8.2	0.5

Jakiego typu jest ta zmienna? Jakie są możliwe wartości tej zmiennej?

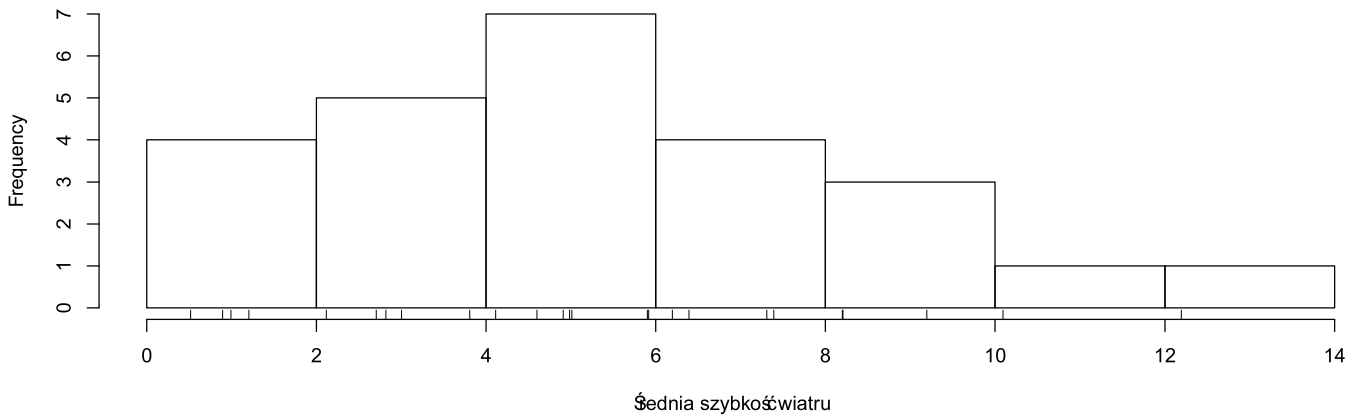
1. Przedstaw rozkład empiryczny badanej zmiennej za pomocą szeregu rozdzielczego.

```
##          liczebność  procent
## (0,2]           4    0.16
## (2,4]           5    0.20
## (4,6]           7    0.28
```

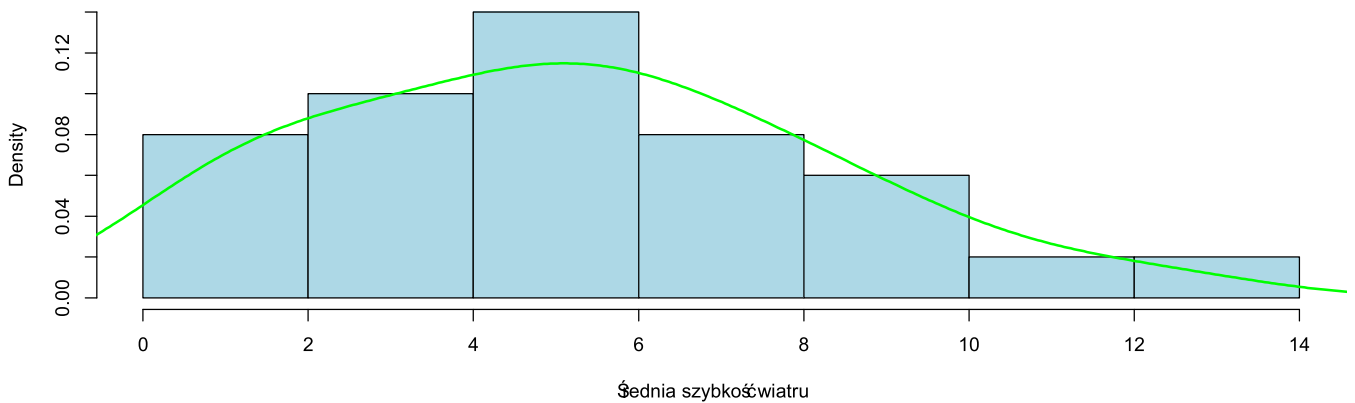
##	(6,8]	4	0.16
##	(8,10]	3	0.12
##	(10,12]	1	0.04
##	(12,14]	1	0.04

2. Zilustruj rozkład empiryczny średniej szybkości wiatru za pomocą histogramu i wykresu pudełkowego. Jakie są zalety i wady tych wykresów?

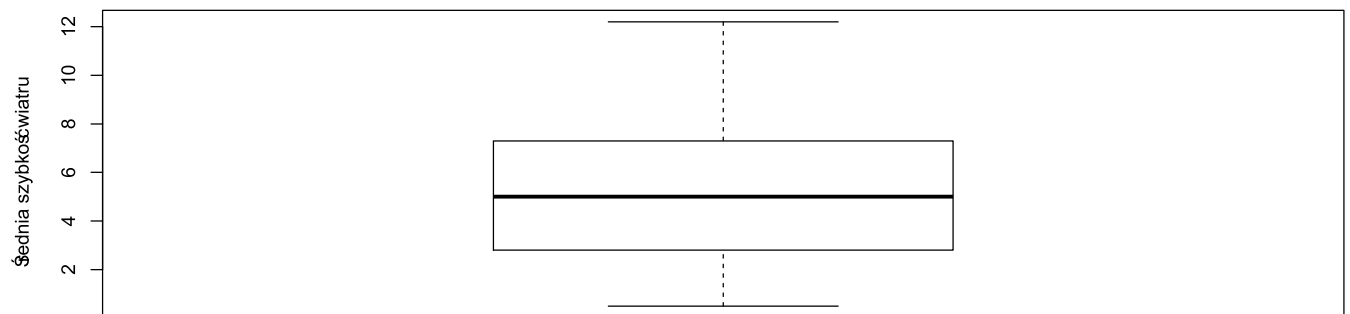
Rozkład empiryczny średniej szybkości wiatru



Rozkład empiryczny średniej szybkości wiatru



Rozkład empiryczny średniej szybkości wiatru



4. Obliczyć średnią, medianę, odchylenie standardowe, współczynnik zmienności, współczynnik asymetrii i kurtozę średniej szybkości wiatru.

[1] 5.144

[1] 5

```
## [1] 3.054925
## [1] 59.38812
## [1] 0.3422838
## [1] -0.665667
```

5. Zinterpretuj powyższe wyniki (tabelaryczne, graficzne i liczbowe).

Zadanie 4. Napisz funkcję `wspolczynnik_zmiennosci()`, która oblicza wartość współczynnika zmienności dla danego wektora obserwacji. Funkcja powinna mieć dwa argumenty:

- `x` - wektor zawierający dane,
- `na.rm` - wartość logiczna (domyślnie `FALSE`), która wskazuje czy braki danych (obiekty `NA`) mają być zignorowane.

Funkcja zwraca wartość współczynnika zmienności wyrażoną w procentach. Ponadto funkcja sprawdza, czy wektor `x` jest wektorem numerycznym. W przeciwnym razie zostanie zwrócony błąd z następującym komunikatem: „argument nie jest liczbą”. Przykładowe wywołania i wyniki funkcji są następujące:

```
x <- c(1, NA, 3)
wspolczynnik_zmiennosci(x)
## [1] NA
wspolczynnik_zmiennosci(x, na.rm = TRUE)
## [1] 70.71068
wspolczynnik_zmiennosci()
## Error in wspolczynnik_zmiennosci() :
## argument "x" is missing, with no default
wspolczynnik_zmiennosci(c("x", "y"))
## Error in wspolczynnik_zmiennosci(c("x", "y")) : argument nie jest liczbą
```

5 Model statystyczny i estymacja punktowa

5.1 Wybrane rozkłady prawdopodobieństwa

1. rozkład dwumianowy $b(m, p)$, $m \in \mathbb{N}$, $p \in (0, 1)$

$$\mathbb{P}(X = k) = \binom{m}{k} p^k (1-p)^{m-k}, \quad k = 0, 1, \dots, m$$

- Funkcja prawdopodobieństwa zmiennej $X \sim b(10, 1/3)$