SWD Lab 1 – Statystyka opisowa w R

Zadanie 1

Poniższe dane przedstawiają liczbę nowych kont zarejestrowanych w ciągu kolejnych 10 dni

43, 37, 50, 51, 58, 105, 52, 45, 45, 10.

a) Oblicz średnią, medianę, kwartyle oraz odchylenie standardowe. Zinterpretuj otrzymane wartości.
b) Sprawdź czy są obecne obserwacje odstające (zgodnie z regułą 1,5*IQR).
c) Usuń zidentyfikowane obserwacje odstające i oblicz ponownie średnią, medianę, kwartyle i odchylenie standardowe. Jaki wpływ na wyznaczone statystyki miały obserwacje odstające?

Zadanie 2

W pliku **samochody.csv (UBI)** zamieszczono dane dotyczące parametrów samochodów kilku wybranych marek.

a) Wczytaj dane z pliku do ramki danych – funkcja read.csv2(). Podaj rozmiar ramki danych (liczba obserwacji i liczba zmiennych) – funkcja dim().
b) Jakiego typu danymi w R są poszczególne zmienne? Czy zmienna jakościowa *producent* jest czynnikiem w R? Jeśli nie, zamień jej typ na czynnik – funkcja factor().
c) Usuń braki danych w utworzonej ramce danych – funkcja na.omit().
d) Zmienna *mpg* opisuje zużycie paliwa w liczbie mil przejechanych na 1 galonie. Utwórz zmienną *zp* opisującą zużycie paliwa mierzone w litrach na 100 kilometrów.

   Wskazówka:
   1 mila = 1609 m
   1 galon (amerykański) = 3,785 l

e) Utwórz histogram dla zmiennej *zp* – funkcja hist(). Jak zmienia się kształt histogramu przy różnych liczbach klas (parametr breaks)?
f) Utwórz wykres łodygowo-liściowy – funkcja stem().
g) Oblicz i zinterpretuj podstawowe statystyki próbkowe dla danych opisujących zużycie paliwa (takie jak: średnia, mediana, kwartyle, 10. i 90. percentyl, wartości ekstremalne, wariancja, odchylenie standardowe, rozstęp, rozstęp międzykwartylowy, współczynnik asymetrii, kurtoza, współczynnik zmienności) – np. funkcje mean(), median(), quantile(), min(), max(), range(), var(), sd(), IQR(), skewness(), kurtosis().
h) Utwórz wykres skrzynkowy (ramkowy, pudełkowy) dla zmiennej opisującej zużycie paliwa – funkcja boxplot().

Zadanie 3

Na podstawie zmiennej *zp* z zadania 2:

a) utwórz zmienną jakościową *zp_kat* opisującą zużycie paliwa przez trzy następujące kategorie: (1) mało, gdy *zp* <=7, (2) średnio, gdy 7 < *zp* <= 10, (3) dużo, gdy *zp* > 10,
b) oblicz jaki procent badanych samochodów należy do każdej z kategorii – funkcje table() i prop.table(),
c) dla zmiennej *zp_kat* utwórz wykres słupkowy – barplot() i kołowy – pie().

Zadanie 4

Oblicz przeciętne zużycie paliwa oraz odchylenie standardowe zużycia paliwa oddzielnie dla samochodów produkowanych w Europie, Ameryce i Japonii (zmienne *producent* i *legenda*) – funkcja tapply(). Zestaw wykresy skrzynkowe zużycia paliwa dla samochodów produkowanych w Europie, Ameryce i Japonii.

Zadanie 5

Porównaj zużycie paliwa przez samochody o jednakowej liczbie cylindrów (zmienna *cylindry*).

Zadanie 6

Porównaj przyspieszenie samochodów produkowanych w Ameryce i Japonii (dane dotyczące przyspieszenia znajdują się w zmiennej *przysp*).

Zadanie 7

Oblicz średnie zużycie paliwa, medianę, wariancję, odchylenie standardowe i współczynnik asymetrii zużycia paliwa wyłącznie dla samochodów ważących mniej niż 2500 funtów (wykorzystać zmienną *waga*).

Zadanie 8
Przeprowadź analizę przyspieszenia samochodów o wadze większej niż 2500 funtów, ale mniejszej niż 3000 funtów (zmienne *przysp* i *waga*), a w szczególności:

a) utwórz i opisz szczegółowo wykres skrzynkowy dla wybranej próbki,
b) utwórz histogram,
c) podaj wartość przyspieszenia, którą przekracza 25% wybranych samochodów.

Zadanie 9
Przeprowadź analizę wagi samochodów, które przejeżdżają na jednym galonie więcej niż 26 mil (zmienne *mpg* i *waga*), a w szczególności:

a) utwórz i opisz szczegółowo wykres skrzynkowy dla wagi wybranych samochodów,
b) utwórz histogram,
c) podaj wagę, której nie przekracza 95% wybranych samochodów.

Zadanie 10
W tabeli podana jest wielkość populacji USA (w milionach) w latach 1790 – 2020.

rok	1790	1800	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900
liczba	3.9	5.3	7.2	9.6	12.9	17.1	23.2	31.4	38.6	50.2	63.0	76.2
rok	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000	2010	2020
liczba	92.2	106.0	123.2	132.2	151.3	179.3	203.3	226.5	248.7	281.4	308.7	332.7

a) Utwórz dla nich wykres jako funkcję czasu (szereg czasowy).
b) Oblicz 10-letnie przyrosty liczby ludności, tj. $d_i = x_{i+1} - x_i$ dla i = 1, 2, …, n-1.
c) Oblicz 10-letnie względne przyrosty liczby ludności, tj. $f_i = (x_{i+1} – x_i)/x_i$ dla i = 1, 2, …, n-1.

Zadanie 11
Dzienne liczby blokowanych włamań z 14 dni wynoszą:

56, 47, 49, 37, 38, 60, 50, 43, 43, 59, 50, 56, 54, 58.

Po zmianie ustawień zapory, dzienne liczby blokowanych włamań w ciągu następnych 20 dni wyniosły:

53, 21, 32, 49, 45, 38, 44, 33, 32, 43, 53, 46, 36, 48, 39, 35, 37, 36, 39, 45.

Na podstawie wykresów skrzynkowych i statystyk próbkowych porównaj liczbę blokowanych włamań przed i po zmianie ustawień zapory.

Funkcje w R

x, y – wektory liczbowe w R

Prezentacja graficzna danych:

hist(x) – rysowanie histogramu
stem(x) – rysowanie wykresu łodygowo-liściowego
boxplot(x) – rysowanie wykresu skrzynkowego

Wyznaczanie statystyk liczbowych:

summary(x) – kilka różnych statystyk
mean(x) – średnia
median(x) – mediana
var(x) – wariancja
sd(x) – odchylenie standardowe
quantile(x, c(0.25,0.5,0.75)) – kwartyle
quantile(x, c(0.95)) – 95. percentyl
IQR(x) – rozstęp międzykwartylowy
range(x) – wartości ekstremalne (min, max)
diff(range(x)) – rozstęp

Uwaga: w przypadku występowania braków danych, podajemy dodatkowy argument w powyższych funkcjach na.rm=TRUE, np. mean(auta$mpg, na.rm=TRUE)

Wybieranie podzbioru, tzw. filtrowanie danych:

I sposób: indeksowanie w R przy pomocy [], np. auta$zp[auta$waga < 2500] (zad 6)
II sposób: funkcja subset(), gdzie pierwszym argumentem jest zbiór, z którego wybieramy podzbiór, drugim – warunek wyboru podzbioru, np. auta6 <- subset(auta, waga < 2500); x <- auta6$zp (zad 6)

Wyznaczanie statystyk i rysowanie wykresów skrzynkowych w grupach

funkcja tapply(), dla której pierwszym argumentem jest wektor liczbowy, drugim wektor lub czynnik określający grupy, trzecim funkcja, która zostanie wyznaczona na wektorze liczbowym względem grup, np. średnia i odchylenie standardowe zużycia paliwa w zadaniu 4.

Odpowiedzi

Lab 1

Z4	tapply(auta$zp, auta$producent, mean, na.rm=TRUE) tapply(auta$zp, auta$producent, sd, na.rm=TRUE) boxplot(auta$zp~auta$producent) Uwaga: Znak tylda ~ w R odnosi się do tzw. formuły i oznacza "względem".
Z5	auta5 <- subset(auta, producent==1 | producent==3) boxplot(auta$przysp~auta$producent)

Z6	x <- auta$zp[auta$waga < 2500] # I sposób
	auta6 <- subset(auta, waga < 2500)# II sposób
	x <- auta6$zp

Z7	x <- auta$moc[auta$rok >= 79 & auta$rok <= 81] # I sposób
	auta7 <- subset(auta, rok >= 79 & rok <= 81) # II sposób
	x <- auta7$moc

Z9	auta9 <- subset(auta, mpg > 26) # określenie podzbioru boxplot(auta9$waga) # wykres skrzynkowy summary(auta9$waga) # wartości opisujące wykres skrzynkowy hist(auta9$waga) # histogram quantile(auta9$waga,0.95) # 95.percentyl
Z10	# Wczytanie danych > stacje <- read.csv2("N:/smwd/stacje.csv") #zbior stacje.csv dostepny jest na stronie http://www.ibspan.waw.pl/~pgrzeg/smwd.htm, w nawiasie podajemy ścieżkę do zbioru > fix(stacje) > licznosci <- c(sum(stacje[,1]=="N"), sum(stacje[,1]=="E"), sum(stacje[,1]=="W"), sum(stacje[,1]=="S")) > licznosci > names(licznosci) <- c("N","E","S","W") > pie(licznosci) > barplot(licznosci)

Lab 2 – Przedziały ufności

Ważne pojęcia:

- przedział ufności dla średniej
- przedział ufności dla odchylenia standardowego
- przedział ufności dla odsetka
- wyznaczanie minimalnej liczebności próby do wyznaczania przedziału o zdanej precyzji

Literatura

STATYSTYKA

1. A. Aczel, *Statystyka w Zarządzaniu*, PWN
2. J. Koronacki, J. Mielniczuk, *Statystyka dla studentów kierunków technicznych i przyrodniczych,* WNT
3. Michael Baron, *Probability and statistics for computer scientists*
   https://ww2.ii.uj.edu.pl/~z1099839/naukowe/RP/rps-michael-byron.pdf

PROGRAMOWANIE W R

4. M. Gągolewski, Programowanie w języku R, Warszawa, II, 2016
   https://ksiegarnia.pwn.pl/Programowanie-w-jezyku-R,647767533,p.html

STATYSTYKA W R

5. P. Grzegorzewski, M. Gągolewski, K. Bobecka-Wesołowska, **Wnioskowanie statystyczne z wykorzystaniem środowiska R**, Warszawa 2014

http://www.gagolewski.com/publications/2014wnioskowaniestatystyczne.pdf

6. J. Adler, *R in a Nutshell*. 2nd Edition,

https://visualization.sites.clemson.edu/reu/resources/RText.pdf

7. A. Faraway, *Practical Regression and Anova using R*,

https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf