
Statystyczne Metody Wspomagania Decyzji – Zestaw zadań nr 1

Wydział Informatyki WIT, studia niestacjonarne 2020/2021

Statystyka opisowa w R

Zadanie 1.1. Poniższe dane przedstawiają liczbę nowych kont zarejestrowanych w ciągu kolejnych 10 dni

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

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

Zadanie 1.2. W pliku **samochody.csv (UBI)** zamieszczono dane dotyczące parametrów samochodów kilku wybranych marek.

- Wczytaj dane z pliku do ramki danych – funkcja `read.csv2()`. Podaj rozmiar ramki danych (liczba obserwacji i liczba zmiennych) – funkcja `dim()`.
- 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()`.
- Usuń braki danych w utworzonej ramce danych – funkcja `na.omit()`.
- 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

- Utwórz histogram dla zmiennej *zp* – funkcja `hist()`. Jak zmienia się kształt histogramu przy różnych liczbach klas (parametr *breaks*)?
- Utwórz wykres łądogowo-liściowy – funkcja `stem()`.
- 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()`.
- Utwórz wykres skrzynkowy (ramkowy, pudełkowy) dla zmiennej opisującej zużycie paliwa – funkcja `boxplot()`.

Zadanie 1.3. Wytrzymałość na ciśnienie wewnętrzne szkła butelek jest ich ważną charakterystyką jakościową. W celu zbadania wytrzymałości butelek umieszcza się je w maszynie hydrostatycznej, po czym zwiększa się ciśnienie aż do zniszczenia butelki. Plik **butelki.csv** zawiera dane opisujące graniczną wytrzymałość na ciśnienie wewnętrzne szkła badanej partii butelek (w psi).

- Utworzyć zmienną o nazwie *cisnienie*, opisującą wytrzymałość na ciśnienie wewnętrzne szkła butelek mierzone w MPa

Wskazówka:

1psi = 0,0068947 MPa

- Utworzyć histogram dla danych opisujących wytrzymałość butelek. Prześledzić wpływ liczby klas na kształt histogramu. Porównać różne rodzaje histogramów.
- Utworzyć wykres łądogowo-liściowy.
- Utworzyć i zinterpretować wykres skrzynkowy dla wytrzymałości butelek.
- Wyznaczyć i zinterpretować podstawowe statystyki próbkowe dla danych opisujących wytrzymałość butelek.
- Obliczyć i zinterpretować 5, 10, 25, 50, 75, 90 i 95 percentyl dla rozważanych danych.

Zadanie 1.4. Na podstawie zmiennej *zp* z zadania 2:

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

Zadanie 1.5. 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 1.6. Porównaj zużycie paliwa przez samochody o jednakowej liczbie cylindrów (zmienna *cylindry*).

Zadanie 1.7. Porównaj przyspieszenie samochodów produkowanych w Ameryce i Japonii (dane dotyczące przyspieszenia znajdują się w zmiennej *przysp*).

Zadanie 1.8. 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 1.9. Przeprowadź analizę przyspieszenia samochodów o wadze większej niż 2500 funtów, ale mniejszej niż 3000 funtów (zmienne *przysp* i *waga*), a w szczególności:

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

Zadanie 1.10. Przeprowadź analizę wagi samochodów, które przejeżdżają na jednym galonie więcej niż 26 mil (zmienne *mpg* i *waga*), a w szczególności:

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

Zadanie 1.11. 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

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

Zadanie 1.12. Dienne 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, dienne 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.

Najczęściej wykorzystywane funkcje w R dla wektora liczbowego x:

`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))` – kwantyle
`IQR(x)` – rozstęp międzykwartylowy
`range(x)` – wartości ekstremalne (min, max)
`diff(range(x))` – rozstęp

`hist(x)` – rysowanie histogramu
`stem(x)` – rysowanie wykresu łodygowo-liściowego
`boxplot(x)` – rysowanie wykresu skrzynkowego
`[]` – indeksowanie wektorów

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