Laboratorium 2 – Miary pozycji i zmienności danych

1. Próbkowanie z rozkładem jednostajnym:

Funkcja `sample(1:<zakres>, <ilość>, replace=TRUE)` losuje z powtórzeniami z przedziału [1,<zakres>] wektor o długości <ilość>.

`sort(<vector>)` sortuje dane liczbowe od najmniejszej

2. Rozkłady empiryczne:

`table(<wektor>)` podaje rozkład empiryczny. Każdej różnej wartości współrzędnej x przypisana jest częstość jej występowania.

3. Szereg rozdzielczy danych:

Szeregi statystyczne stanowią pewną dekompozycję danych. Stanowią jednowymiarową strukturę porządkującą dane. Szeregi dzielimy na:

1. Szczegółowe (wyliczające)
2. Rozdzielcze (strukturalne)
   2.1. Cech mierzalnych
      2.1.1. Punktowe
      2.1.2. Przedziałowe
   2.2. Cech niemierzalnych
3. Przestrzenne (geograficzne)
4. Czasowe (dynamiczne)
   4.1. Momentów
   4.2. Przedziałów

Zasady tworzenia pakietów przedziałowych.

1. Ilość przedziałów powinna zależeć od ilości danych. Przydatne heurystyki: $n = \sqrt{N},\ n = 1 + 3.222\log(N)\,, n \leq 5\log(N)$ gdzie N ilość danych, n ilość przedziałów.
2. Podział całego zakresu danych na równe przedziały klasowe.
3. Skrajne przedziały otwarte, obejmujące wartości odstające.
4. Podział według stałej ilości danych w każdym przedziale

`cut(<dane>,break=c(`a_0`,`a_1`,`a_2`,…,`a_k`))` dzieli dane pomiędzy przedziały $\{(a_i, a_{i+1}]\}$ i=0,k

`cut(<dane>,<n_intervals>)` dzieli dane na n_intervals przedziałów

`table(cut(<dane>,<n_intervals>))/length(<dane>)` podaje rozkład częstotliwości występowania danych w przedziałach

4. Miary pozycji i zmienności:

Tablica 2.1. (Biecek) Statystyki opisowe dla wektora lub macierzy

Z pakietu **base**

`max(x)/min(x)`	Wartość maksymalna/minimalna w próbie x.
`mean(x)`	Średnia arytmetyczna z próby `x`. Opcjonalnym argumentem jest `trim`, jeżeli jest różny od zera, to wyznaczana jest średnia ucięta. Średnią uciętą oblicza się tak jak arytmetyczną po usunięciu 200% * `trim` skrajnych obserwacji.
`length(x)`	Liczba elementów w próbie.
`range(x)`	Przedział zmienności próby, wyznaczony jako [$\min_i x_i$, $\max_i x_i$].
`diff(x,differences=v)`	Oblicza różnice pomiędzy współrzędnymi wektora x, differences podaje rząd różnic, dyfulth to v=1, obliczane są kolejne różnice
`summary(x)`	podaje min, 1 kwantyl medianę, 3 kwantyl i maksimum x. Ma listę etykiet
`fivenum(x)`	podaje same wartości numeryczne summary bez etykiet

Z pakietu **stats**

`weighted.mean(x,w)`	Średnia ważona z próby `x`. Wektor wag jest drugim argumentem. Liczone jest $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} w_i\, x_i$. Użyteczna dla obliczania średniej szeregu rozdzielczego, wtedy wagami są ilości próbek w poszczególnych przedziałach szeregu $w_i = n_i$, natomiast $x_i = \dot{x}_i$ są środkami przedziałów szeregu.
`median(x)`	Mediana (wartość środkowa x).
`quantile()`	Kwantyl wybranego rzędu. Drugim argumentem funkcji `quantile()` jest wektor kwantyli do wyznaczenia. W tej funkcji zaimplementowano 9 różnych algorytmów do wyliczania kwantyli, zobacz opis argumentu `type`. Pierwszy argument to dane, drugi, to wektor granicznych prawdopodobieństw granicznych.
`IQR()`	Rozstęp międzykwartylowy, czyli różnica pomiędzy górnym a dolnym kwartylem IRQ = $q_{0.75} - q_{0.25}$
`var()`	Wariancja w próbie. Wyznaczana jest nieobciążona ocena wariancji $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Dla dwóch wektorów obliczona będzie kowariancja dla tych wektorów, a dla macierzy wynikiem będzie macierz kowariancji kolumn.
`sd()`	Odchylenie standardowe wyznaczone jako $\sqrt{S^2}$, gdzie S^2 to ocena wariancji.
`cor(), cov()`	Macierz korelacji i kowariancji. Argumentami może być para wektorów lub macierz.
`mad(x)`	Medianowe odchylenie bezwzględne, wyznaczone jako $1.4826\, median(y),\ \ y_i = \lvert x_i - median(x)\rvert$

Z innych pakietów

`kurtosis()`	Kurtoza, miara koncentracji, $\frac{n\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^4}-3$. Rozkład normalny ma kurtozę 0. Funkcja z pakietu **e1071**.
`skewness()`	Skośność, miara asymetryczności, $\frac{\sqrt{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{\frac{3}{2}}}$. Rozkład symetryczny ma skośność 0. Funkcja z pakietu **e1071**.
`geometric.mean()`	Średnia geometryczna, wyznaczona jako $\sqrt[n]{(\prod_{i=1}^{n}x_i)}$. Funkcja z pakietu **psych**.
`harmonic.mean()`	Średnia harmoniczna, wyznaczona jako $\frac{n}{\sum_{i=1}^{n}x_i^{-1}}$. Funkcja z pakietu **psych**.
`moda()`	Moda lub dominanta, czyli wartość występująca najczęściej w próbie. Funkcja z pakietu **dprep** (są problemy z dostępnością pakietu) W Linuxie można też użyć funkcji `mod()` z pakietu **RVAideMemoire**.

Odchylenie przeciętne $d_1=\frac{1}{n}\sum_{i=1}^{n}|x_i-\bar{x}|$ można obliczyć przy pomocy funkcji `d1(x)`:

```
d1 <-function(x) mean(abs(x-mean(x)))
```

Zadania

Szeregi statystyczne

Zadanie 1. Dla danych `acme` z pakietu `boot` sporządź szereg rozdzielczy dla kolumny `acme$market` według reguł 1 – 4.

Zadanie 2. Dla danych `acme` z pakietu `boot` sporządź szereg rozdzielczy ilości danych w każdym roku.
Oblicz średnią danych z kolumny `acme$market` dla każdego elementu szeregu rozdzielczego.

Zadanie 3. Napisz funkcję dzielącą dane numeryczne na przedziały o zadanej proporcji zawartości ilości danych (np. 10%, 20%, 30%, 20%,10%,10%).

Zadanie 4. Dla danych `catsM` oblicz dostępne wskaźniki pozycyjne i miary zmienności dla kolumn `catsM$Bwt, catsM$Hwt` i ich różnicy.