Wykład 15. Metody nieparametryczne. Elementy analizy wielowymiarowej.

dr Mariusz Grządziel

semestr letni 2012

Weryfikacja założenia o normalności rozkładu populacji
Dane są obserwacje $x_1, x_2, \ldots, x_n$. Czy można założyć, że $x_1, x_2, \ldots, x_n$ jest realizacją próby prostej z rozkładu normalnego $N(\mu, \sigma)$ dla pewnych μ i σ?
Metody weryfikacji założenia o normalności:

- Sporządzenie histogramu i porównanie jego kształtu z krzywą dzwonową (wykresem funkcji ϕ.)
- Sporządzenie wykresu ramkowego i odpowiednia jego interpretacja
- Sporządzenie wykresu kwantylowego i odpowiednia jego interpretacja.
- Testy zgodości: Shapiro–Wilka i inne.

Wykres kwantylowy
Uporządkowany ciąg obserwacji $x_1, \ldots, x_n$ oznaczamy przez $x_{1:n}, \ldots, x_{n:n}$. Niech

$$x'_{1:n} = \frac{x_{1:n} - m}{s}, \ldots, x'_{n:n} = \frac{x_{n:n} - m}{s},$$

gdzie m onacza średnią z próby, a s odchylenie standardowe z próby. Jeśli założymy, że $x_1, \ldots, x_n$ pochodzą z rozkładu normalnego, to $x'_{1:n}, x'_{2:n}, \ldots$, powinny być sensownymi przybliżeniami kwantyli $q_{1/n}, q_{2/n}, \ldots$ (rzędu $\frac{1}{n}$, rzędu $\frac{2}{n}$ itd.) rozkładu $N(0,1)$; stąd punkty $(x'_{1:n}, q_{1/n}), (x'_{2:n}, q_{2/n})$ powinny się układać wokół (pewnej) prostej ; analogicznie punkty:

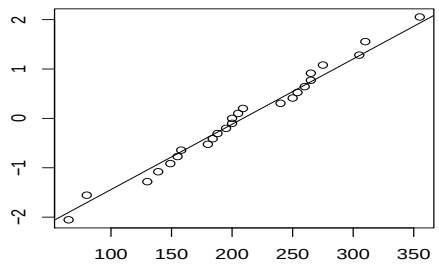
$$(x_{1:n}, q_{1/n}), (x_{2:n}, q_{2/n}), \ldots \tag{1}$$

również powinny się układać wokół prostej.Wykres przedstawiający punkty (1) jest nazywany wykresem kwantylowym.
W praktyce zamiast kwantyli rzędu i/n wyznacza się np. kwantyle rzędu $i/(n+1)$; w ten sposób unikamy problemu z n-tym punktem!
Wykres kwantylowy „odbiegający od prostej": założenie o normalności rozkładu, z którego pochodzą $x_1, \ldots, x_n$, należy odrzucić.
Szczegóły dotyczące sprządzania rozkładu kwantylowego można znaleźć w [KM01,par.3.4.2].

Wykres kwantylowy— obliczenia
Rozważmy dane dotyczące cen mieszkań w dzielnicy A. Chcemy zweryfikować założenie o normalności rozkładu zmiennej losowej, którą jest cena losowo wybranego mieszkania w dzielnicy A. Wykres kwantylowy dla rozważanych danych można otrzymać wydając polecenia:

```
y=c(65,80,139,180,355,158,240,205,265,305,200,155,
209,310,149,254,188,265,275,200,184,130,260,250,195)
# dane dotyczace cen mieszkan zostaly zapisane do zm. y
qqnorm(y,datax=TRUE,main="", xlab="",ylab="")# wykres...
qqline(y,datax=TRUE) # dodaje odpowiednia prosta
```

Ceny mieszkań w A



Rysunek 1: Wykres kwantylowy dla danych dotyczących cen mieszkań w A

Test Shapiro–Wilka i inne testy zgodności
Zgodność z rozkładem normalnym można weryfikować za pomocą testu Shapiro-Wilka.
Załóżmy, że $X_1, X_2, \ldots, X_n$ są losową próbą prostą z rozkładu ciągłego. Jesteśmy zainteresowani weryfikacją hipotezy

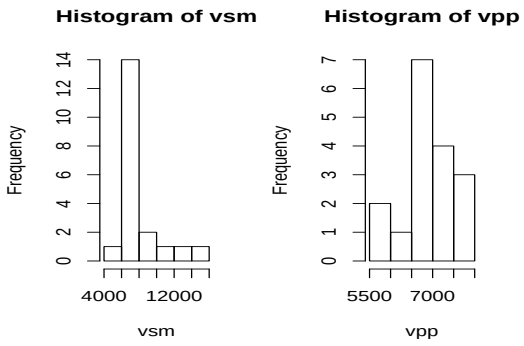
$$H_0 : X_i, i = 1, \ldots, n \text{ mają rozkład } N(\mu, \sigma) \text{ dla pewnych } \mu i \sigma;$$

przeciw

$$H_1 : \text{ rozkład zmiennych } X_i, i = 1, \ldots, n \text{ nie jest rozkładem normalnym.}$$

Do weryfikacji tej hipotezy można zastosować test Shapiro-Wilka. Test ten bazuje na idei wykresu kwantylowego. Więcej informacji o teście Shapiro-Wilka; można znaleźć w [KM01,par. 3.4.2].
Inne testy zgodności z rozkładem normalnym: w pakiecie R-owskim **nortest**.

Histogramy dla cen mieszkań w PP i SM



Test Shapiro–Wilka: przykład obliczeń
W zmiennej y zapisane są wartości cen mieszkań w dzielnicy A.

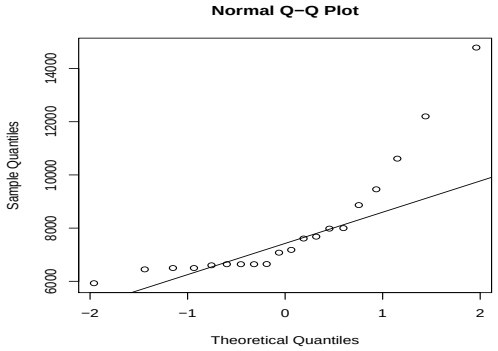
```
> y
 [1]  65  80 139 180 355 158 240 205 265 305 200 155 209
      310 149 254 188 265 275
[20] 200 184 130 260 250 195
> shapiro.test(y)

        Shapiro-Wilk normality test

data:  y
W = 0.9849, p-value = 0.962
```

p-wartość jest równa $0{,}962$– nie ma podstaw do odrzucenia hipotezy o zgodności z rozkładem normalnym.

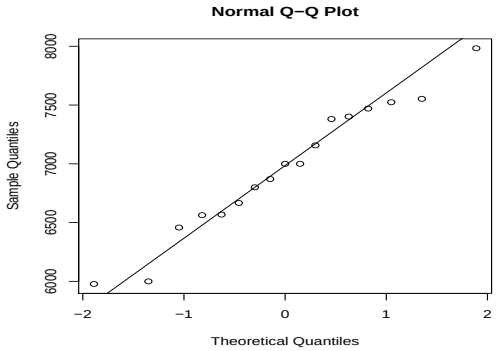
Porównanie rozkładów dwóch populacji— sytuacja nieparametryczna
Rozważmy następujące dane dotyczące ceny metra kwadratowego mieszkań 2-pokojowych w dzielnicach SM i PP miasta W:

```
vsm<-c(6602,10612,9459,6644,7679,6644,14787,8000,6450,
5929,8865,6650,6500,7077,7182,6500,7609,12200,7981,6644)
vpp<-c(7402,7000,7470,7524,6000,7381,6568,6800,6870,6667,
7553,5978,6458,7000,6563,7157,7983)
```

Ceny są podane w zł.; dane zostały zapisane do wektorów **vsm** i **vpp**.



Rysunek 2: Wykres kwantylowy dla danych dotyczących cen mieszkań 2-pokojowych w dzielnicy SM



Rysunek 3: Wykres kwantylowy dla danych dotyczących cen mieszkań 2-pokojowych w dzielnicy PP

Weryfikacja założenia normalności— test Shapiro-Wilka

```
> shapiro.test(vsm)

        Shapiro-Wilk normality test

data:  vsm
W = 0.7597, p-value = 0.0002297
> shapiro.test(vpp)

        Shapiro-Wilk normality test

data:  vpp
W = 0.9691, p-value = 0.8029
```

Test Wilcoxona
Założenie o normalności rozkładu cen mieszkań oferowanych do sprzedaży w dzielnicy SM należy odrzucić.

Z dostępnych informacji dotyczących dzielnic SM i PP można wnioskować, że średnia cena mieszkań w SM „nie może" być niższa niż średnia cena mieszkań w PP.
Odpowiednim narzędziem, które można zastosować do porówania średnich dla populacji odpowiadających cenom mieszkań w dzielnicy SM i PP, byłby nieparametryczny odpowiednik testu t-Studenta dla dwóch prób, dla jednostronnej hipotezy alternatywnej.

Test Wilcoxona
Jesteśmy zainteresowani weryfikacją hipotezy

$$H_0 : F = G$$

przeciwko hipotezie H_1 :

$$H_1 : G(x) \leqslant F(x) \text{ dla każdego } x \in \mathbb{R} \text{ oraz } G \neq F,$$

gdzie F i G są dystrybuantami porównywanych populacji. Zauważmy, że przy założeniu, że F odpowiada rozkładowi $N(\mu_1, \sigma)$ a G rozkładowi $N(\mu_2, \sigma)$ hipotezy H_0 i H_1 odpowiadają hipotezom $\mathcal{H}_0$ i $\mathcal{H}_1$:

$$\mathcal{H}_0 : \mu_1 = \mu_2 \text{ oraz } \mathcal{H}_1 : \mu_1 \geqslant \mu_2.$$

Niech F_{SM} oznacza dystrybuantę cen mieszkań (oferowanych do sprzedaży) w dzielnicy SM a F_{PP} oznacza dystrybuantę cen mieszkań oferowanych do sprzedaży w dzielnicy PP. Aby zweryfikować hipotezę $H_0 : F_{SM} = F_{PP}$ przeciwko

$$H_1 : F_{SM}(x) \leqslant F_{PP}(x) \text{ dla każdego } x \text{ i } F_{SM}(x) \neq F_{PP}$$

można użyć polecenia **wilcox.test**

Ceny mieszkań w dzielnicach...—obliczenia

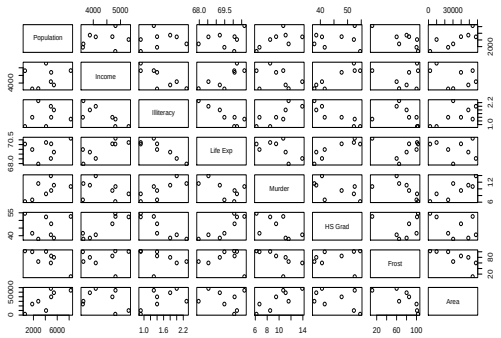
```
> wilcox.test(vsm,vpp,alternative="greater")

        Wilcoxon rank sum test with continuity correction

data:  vsm and vpp
W = 204, p-value = 0.1536
alternative hypothesis: true mu is greater than 0

Warning message:
cannot compute exact p-value with ties in:
wilcox.test.default(vsm, vpp, alternative = "greater")
```

Wniosek: nie ma podstaw do odrzucenia hipotezy o równości rozkładów cen mieszkań oferowanych do sprzedaży w dzielnicach SM i PP.

Metody analizy wielowymiarowej— wzmianka
W obiekcie (macierzy) SouthAtlantic zapisane są, w kolumnach *Population*, *Income*, itd... , dane dotyczące populacji (tys. mieszkańców,1975), dochodu *per capita* (1974), średniej długości życia (1969–1971), liczby morderstw przypadajacych na 100 tys. osób (1976), procentu osób ze średnim wykszt. (1970), średniej liczby dni, w których temperatura była ujemna (1931– 1960), powierzchni (w milach kw.).

```
> SouthAtlantic
               Population Income Illiteracy Life Exp Murder HS Grad Frost  Area
Florida              8277   4815        1.3    70.66   10.7    52.6    11 54090
Georgia              4931   4091        2.0    68.54   13.9    40.6    60 58073
North Carolina       5441   3875        1.8    69.21   11.1    38.5    80 48798
South Carolina       2816   3635        2.3    67.96   11.6    37.8    65 30225
Virginia             4981   4701        1.4    70.08    9.5    47.8    85 39780
West Virginia        1799   3617        1.4    69.48    6.7    41.6   100 24070
Maryland             4122   5299        0.9    70.22    8.5    52.3   101  9891
Delaware              579   4809        0.9    70.06    6.2    54.6   103  1982
```

Diagramy rozrzutu dla par zmiennych— polecenie pairs

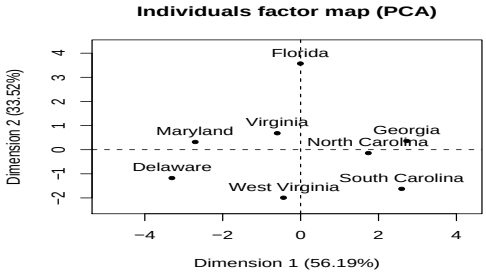


Rysunek 4: Wykres otrzymany po użyciu polecena pairs(SouthAtlantic)

Analiza skladowych głównych

Metodą pozwalającą na przedstawienie danych „wielowymiarowych" dotycących analizowanych obiektów (w naszym przykładzie stanów USA) w przestrzeni dwuwymiarowej jest analiza składowych głównych.

Idea leżąca u podstaw tej metody jest następująca. Utożsamiamy dane dotyczące kolejnych obiektów z punktami w przestrzeni k-wymiarowej, gdzie k jest równe liczbie cech. Następnie dokonujemy zmiany układu współrzędnych tak, aby pierwszy wektor („pierwsza składowa główna") w nowej bazie odpowiadał „możliwie dużej zmienności danych", drugi wektor („druga składowa główna") odpowiadał możliwie dużej zmienności danych w przestrzeni prostopadłej do pierszego wektora itd. Wykres odpowiadający podprzestrzeni rozpiętej na pierszych dwóch składowych głównych prezentuje realtywnie dużo informacji dotyczących interesującego nas układu punktów w przestrzeni k-wymiarowej.



Rysunek 5: Dane dotyczace stanów leżących w regionie South Atlantic prezdstawione na wykresie odpowiadającym dwóm pierszym składowym głównym. Składowa pierwsza, reprezentowana przez oś poziomą, odpowiada $56{,}19\%$ zmienności danych. Składowa druga, reprezentowana przez oś pionową, odpowiada $33{,}52\%$ zmienności danych.

Lektura uzupełniająca

[KM01] Koronacki, J., Mielniczuk, J. Statystyka dla studentów kierunków technicznych i przyrodniczych. WNT. Warszawa 2001, s. 111–118.