Zastosowanie pakietu R w kontekście lokalizacji genów głównych sprzężonych z cechą ilościową (lokalizacja QTLi)

Przemyslaw.Biecek@gmail.com, MIM Uniwersytet Warszawski

Plan

1. R wprowadzenie
2. Przegląd wybranych możliwości statystycznych
3. Wybrane pakiety wspierające analizy danych QTL
4. Podsumowanie

Kilka słów o R

R to

- R to język programowania, środowisko oraz platforma.
- Powstał z inicjatywy Roberta Gentlemana i Ross Ihake (znani jako R&R) w 1995.
- Od roku 1997 rozwojem R kieruje liczny zespół osób nazwanych „core team".
- Język R jest podobny do języka S opracowanego w laboratoriach Bell'a.
- R jest projektem GNU opartym o licencje GPL GNU (darmowy do wszystkich zastosowań).

Wykorzystując go w pracy naukowej, warto zaznaczyć to w referencjach

```
R Development Core Team (2008). R: A language and environment for
statistical computing. R Foundation for Statistical Computing,
Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.
```

Kilka słów o R

Najważniejsze zalety R:

- to najszybciej rozwijający się pakiet statystyczny,
- otrzymujemy powtarzalne wyniki, możemy publikować kody w R opisujące jak uzyskaliśmy te wyniki,
- możliwość tworzenia i upowszechnianie pakietów, przez co coraz więcej zaawansowanych metod statystycznych jest dostępnych dla zwykłego użytkownika,
- możliwość łatwej komunikacji z innymi programami,
- zupełnie darmowy,
- tworzy grafiki o wysokiej jakości.

Kilka słów o R

- około 1700 pakietów, większość statystycznych metod analizy danych
- wszystkie platformy: Windows, Linuxy, Mac itp
- pakiety wspierające obliczenia w klastrach obliczeniowych (zrównoleglanie)
- szczególnie chętnie używany przez naukowców
- autorzy artykułów często dołączają do swoich prac pakiety lub kody w R
- wiele interfejsów okienkowych
- komunikacja z bazami danych
- obsługa plików danych innych pakietów

Kilka słów o R

- R działa interaktywnie, komendy podawane są w konsoli, umożliwia to:
- używanie powtarzalnych wyników
- kontrola nad tym „co ja właściwie zrobiłem rok temu”
- stroma krzywa uczenia
- możliwa integracja z innymi programami, językami (R, Java itp) pakietami (Matlab, SAS itp)
- R można nasić ze sobą na płycie/pendrive

Instalacja

Skąd ściągnąć pakiet R?

- Wpisać w wyszukiwarce R i wybrać pierwszy link.
- Otworzyć stronę `http://cran.r-project.org/mirrors.html`.
- Ściągnąć plik `R-2.8.2-win32.exe` z wersją R 2.8.2 (lub wybrać nowszą jeżeli jest).
- Domyślnie R instaluje się w katalogu `c:/Program Files/R/R-2.8.2`.
- Uruchomić poleceniem `Rgui.exe` (wersja dla Windows).

Pierwszy kontakt

Po uruchomieniu platformy pojawi się linia poleceń. Możemy teraz wpisywać i wykonywać kolejne komendy.

Przykład w R

```
> # dziubek to symbol gotowości
> cat("Hello word! \n")
Hello word!
> 2^13
[1] 8192
> plot(data)
```

Wybrane ciekawe pakiety

- seqinr - Pakiet do analiz na sekwencjach, umożliwia łatwe wczytywanie i zapisywanie danych o sekwencjach zapisanych w różnych formatach, oraz wiele przydatnych operacji na tych danych
- qtl - Pakiet do analiz QTLowych, lokalizacji QTLi i zagadnień z tym związanych
- kinship - pakiet do wyznaczania macierzy pokrewieństwa
- GO - Pakiet do operacji na ontologii Gene Ontology
- Rgraphviz - Pakiet do wizualizacji grafów (np struktury GO)
- ape - Analyses of Phylogenetics and Evolution
- genefilter - Pomocniczy pakiet z funkcjami ułatwiającymi filtrowanie genów.
- Biobase - Zbiór podstawowych funkcji do zarządzania danymi.
- multtest - Zbiór funkcji wspomagających wielokrotne testowanie.
- ALL - Acute Lymphoblastic Leukemia, zbiór danych dotyczących białaczki.

Analizy wielowymiarowe

Statystyka

Podsumowania danych, funkcja: summary(base)

Funkcja summary(base) wyświetla podsumowania. Dla zmiennej jakościowej pokaże liczebności obserwacji.

Przykład w R

```
> summary(wyksztalcenie)
podstawowe    srednie     wyzsze   zawodowe
        22         34         93         55
```

Dla zmiennych ilościowych wynikiem jest wektor z wartościami minimum, maksimum, średniej, mediany i kwartyle.

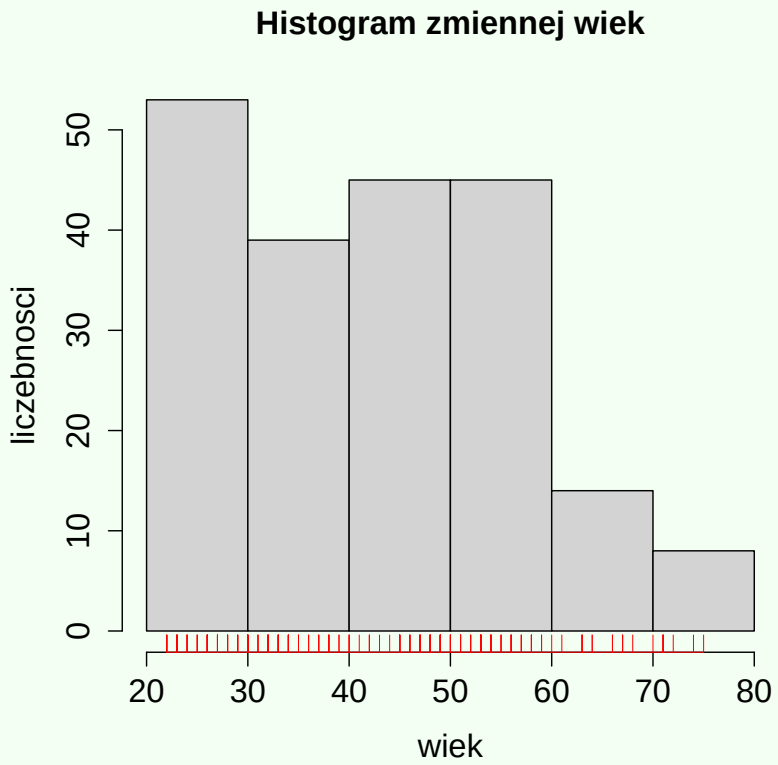
Przykład w R

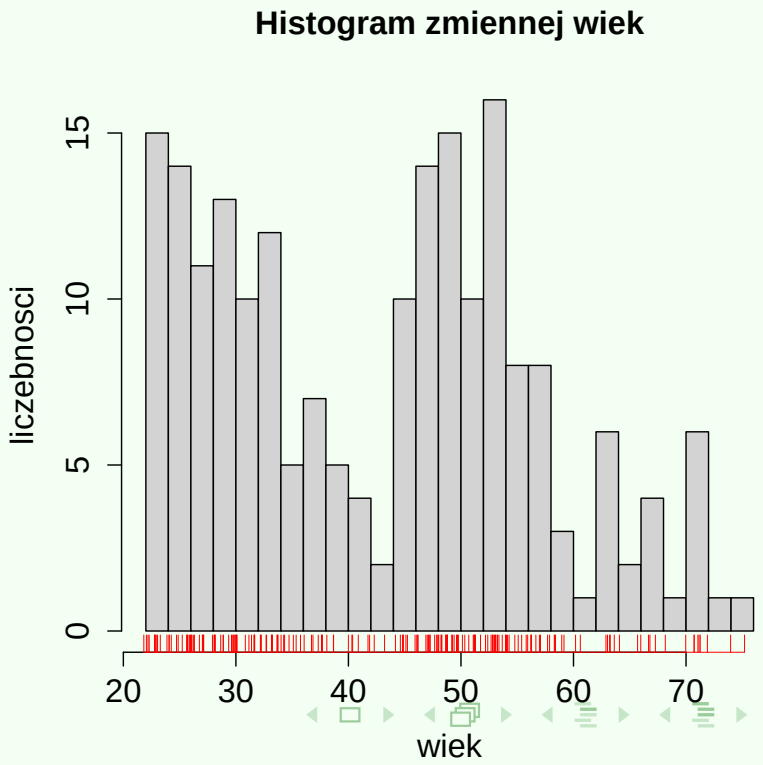
```
> summary(wiek)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  22.00   30.00   45.00   43.16   53.00   75.00
```

Histogram, funkcja: hist(graphics)

Przykład w R

```
> hist(wiek,5,main="Histogram zmiennej wiek",ylab=" liczebnosci"
> rug(wiek,side=1,ticksize=0.03,col=" red")
```

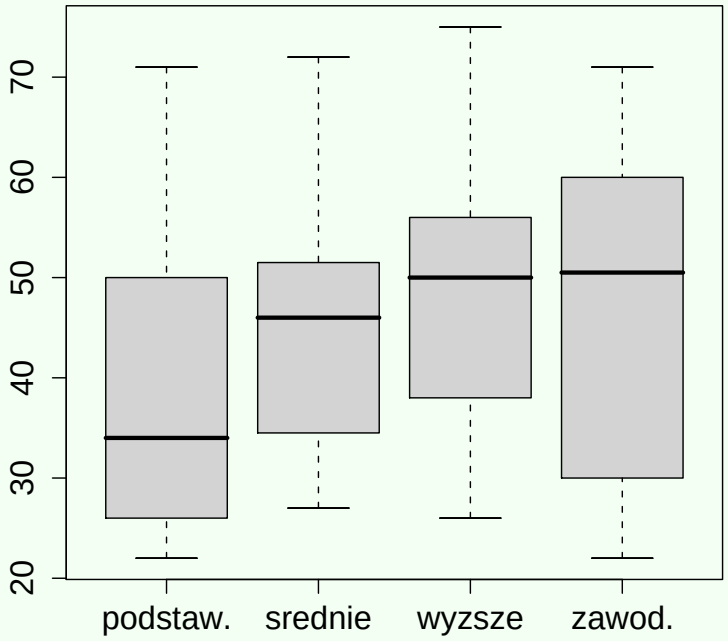


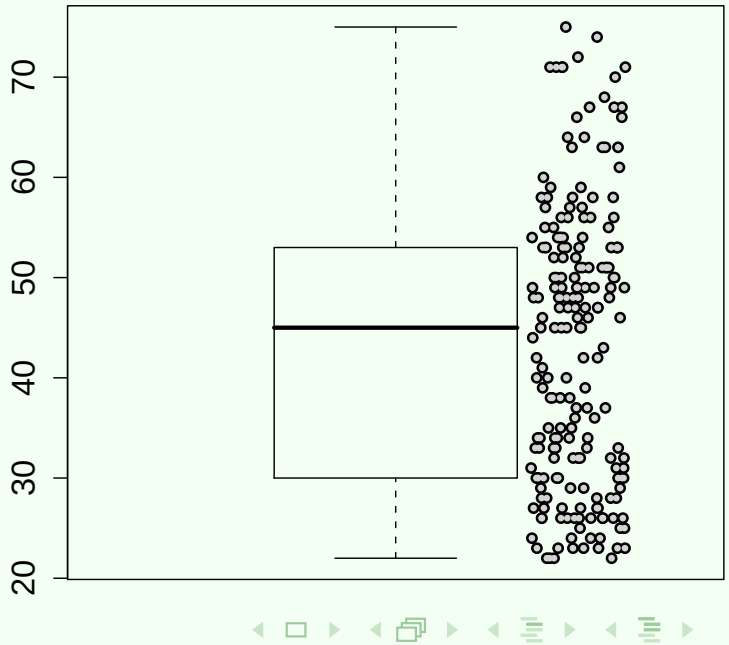


Wykresy pudełkowe: boxplot(graphics)

Przykład w R

```
# wykres pudełkowy w rozbiciu na podpopulacje
boxplot(wiek~wyksztalcenie, data = dane, col=" lightgrey")
boxplot(wiek)
```

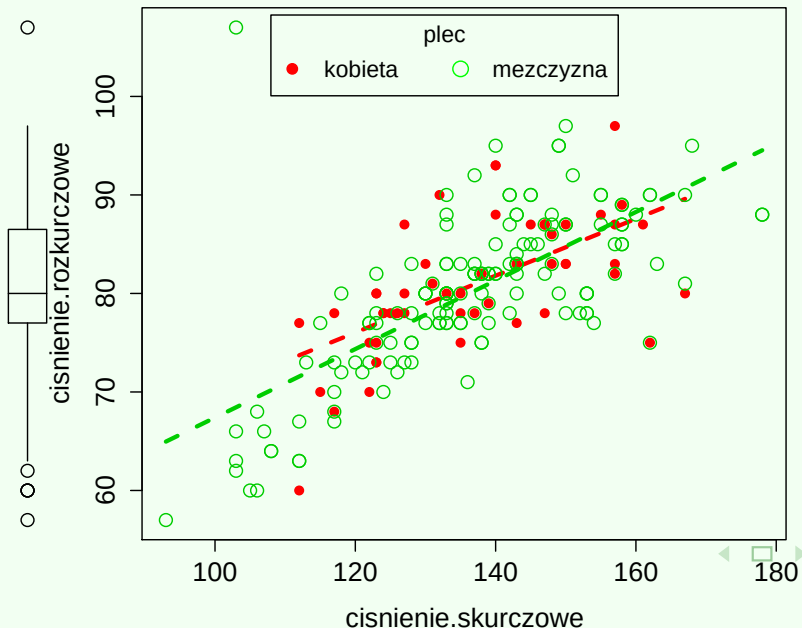




Wykres rozrzutu, funkcja: scatterplot(car)

Przykład w R

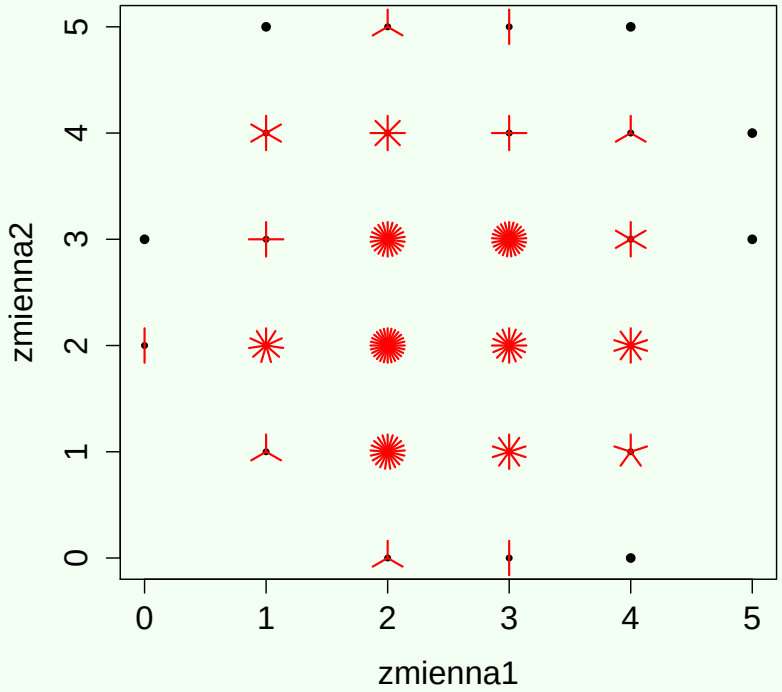
```
> sp(cisnienie.skurczowe, cisnienie.rozkurczowe, groups=plec,
      smooth=F, lwd=3, pch=c(20,21), cex=1.5)
> sp(cisnienie.skurczowe, cisnienie.rozkurczowe, smooth=F)
```

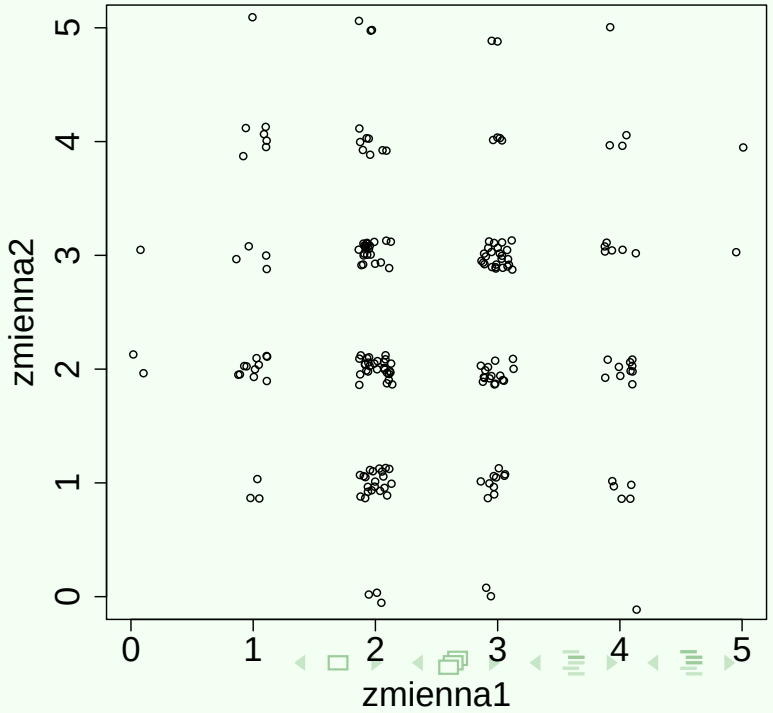


Wykres słonecznikowy, funkcja: sunflowerplot(graphics)

Przykład w R

```
# wykres słonecznikowy dla dwóch zmiennych
sunflowerplot(zmienna1, zmienna2)
```

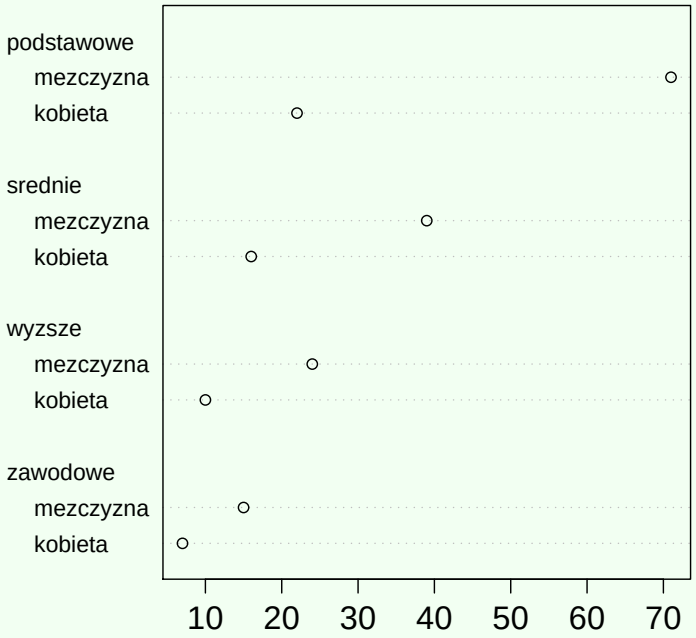


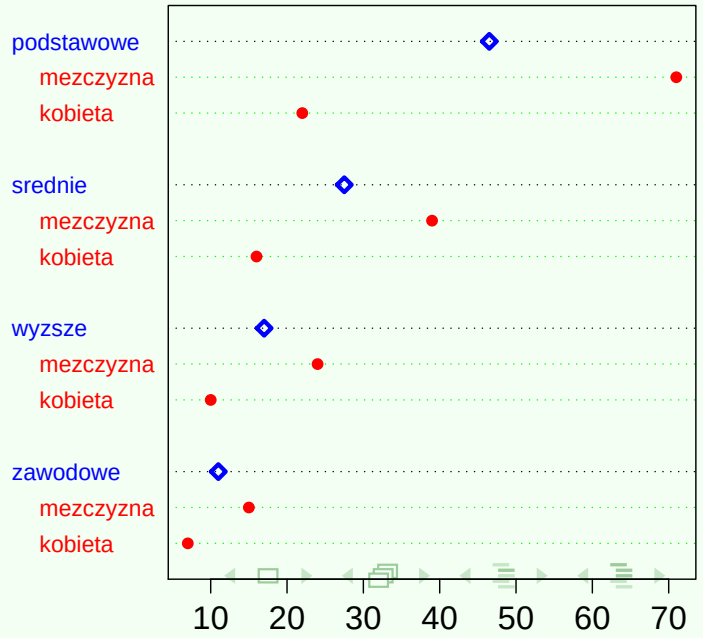


Wykres kropkowy, funkcja: dotchart(graphics)

Przykład w R

```
dotchart(tab)
dotchart(tab, gdata=apply(tab,2,mean), pch=19, gpch=5,
    color=" red", gcolor="blue", lcolor="green", lwd=3)
```

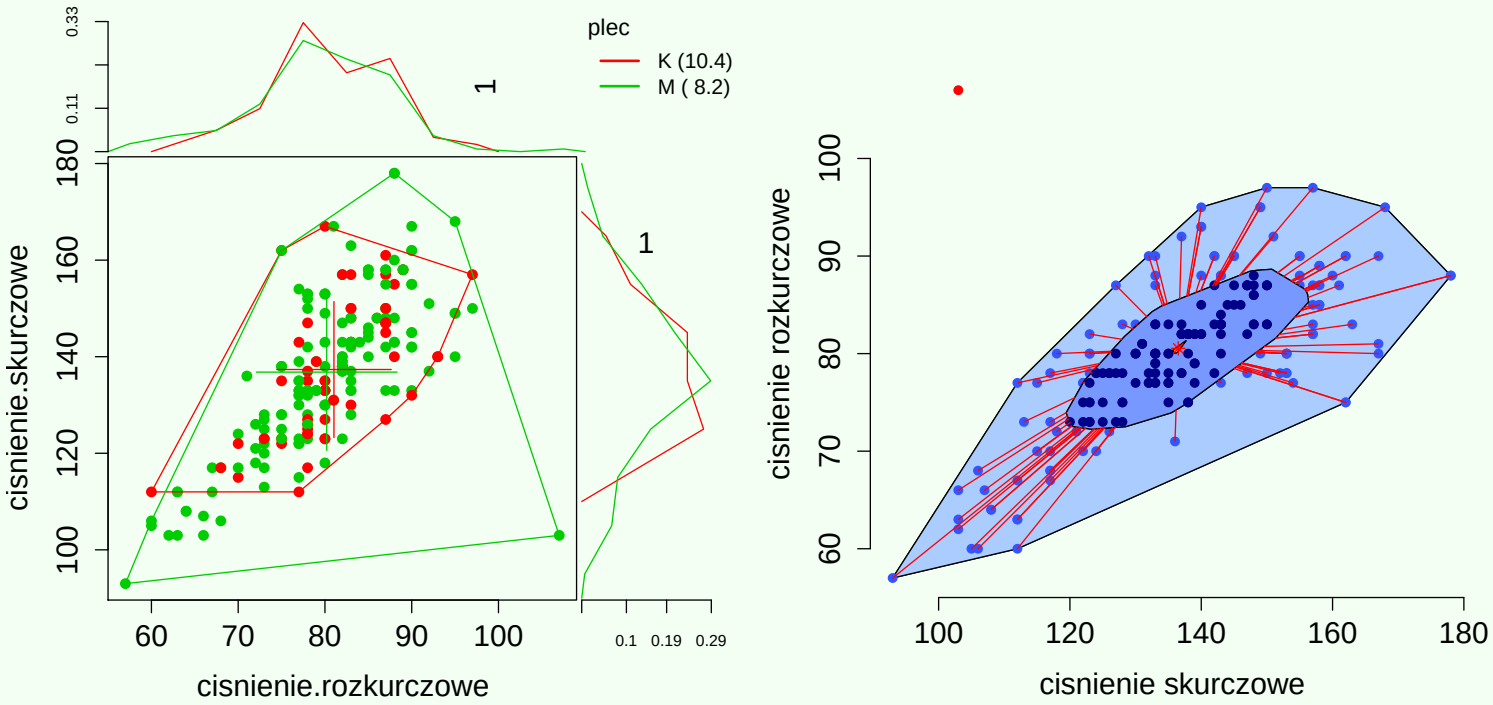




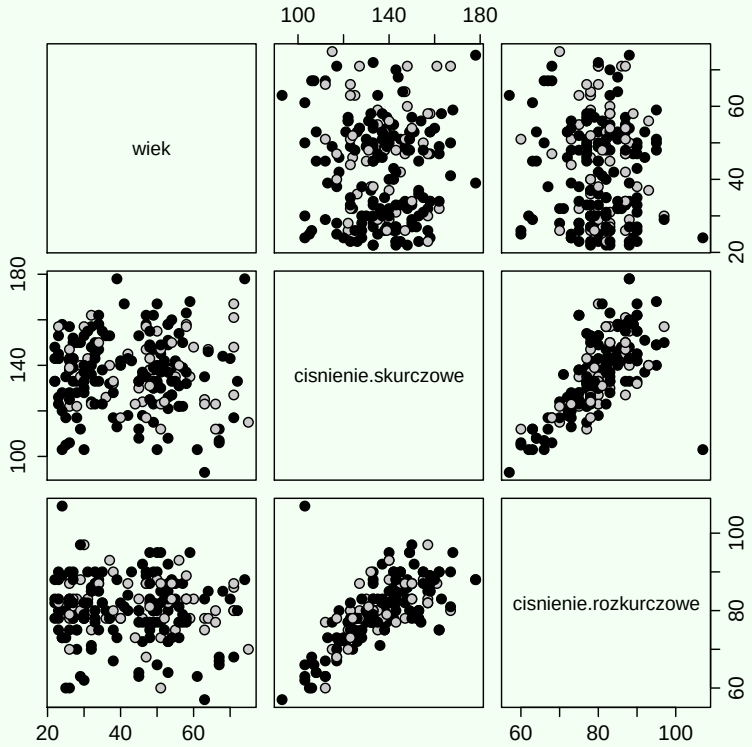
Wykres torbowy, funkcja: bagplot(aplpack)

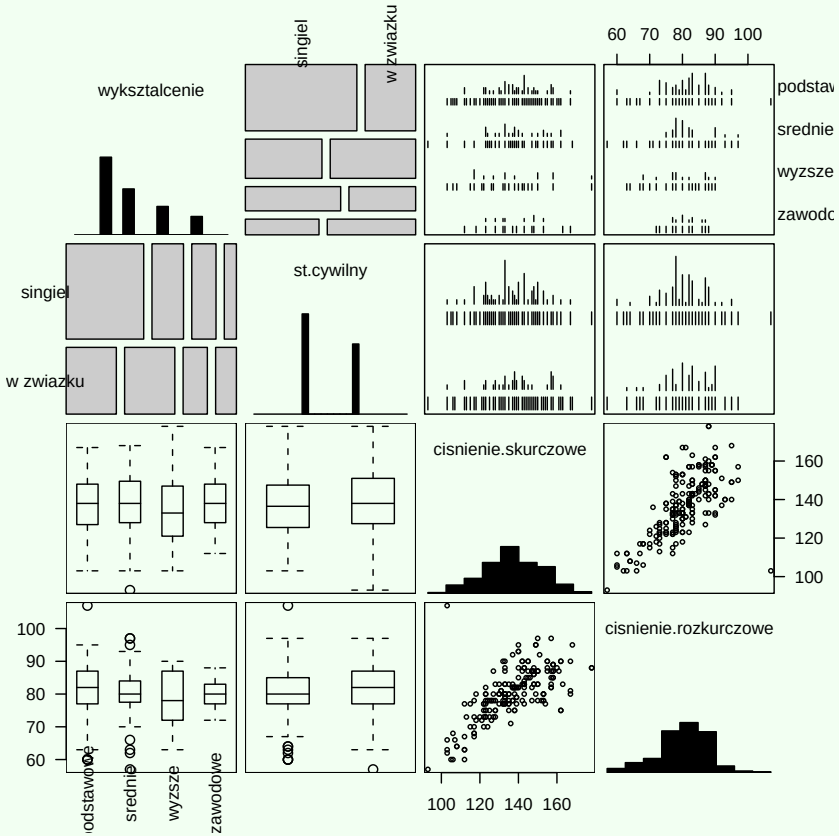
Przykład w R

```
bagplot(cisnienie.skurczowe, cisnienie.rozkurczowe)
```

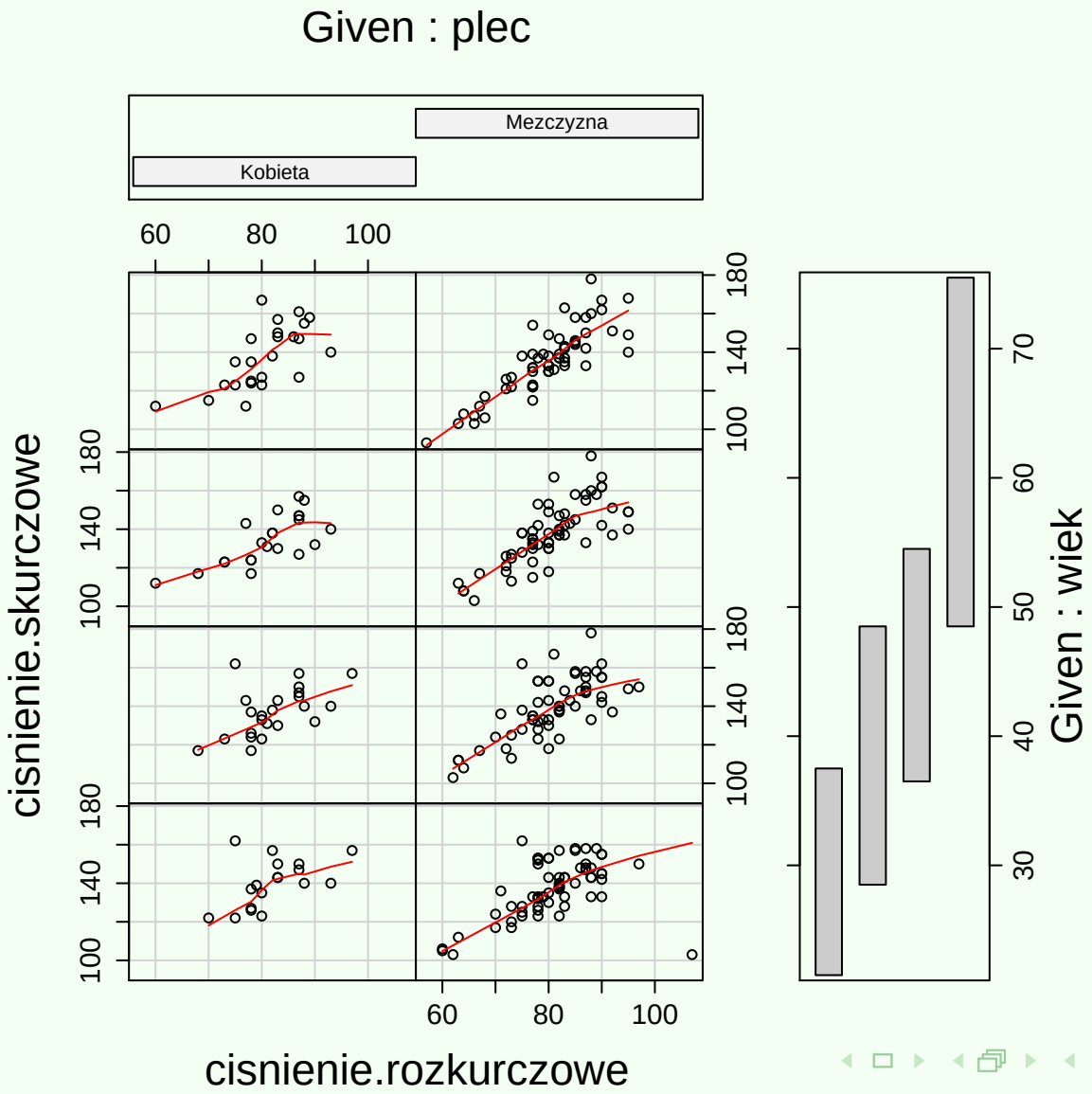


Wykresy rozrzutu, funkcje: pairs(graphics), scatterplot.matrix(car) i gpairs(YaleToolkit)

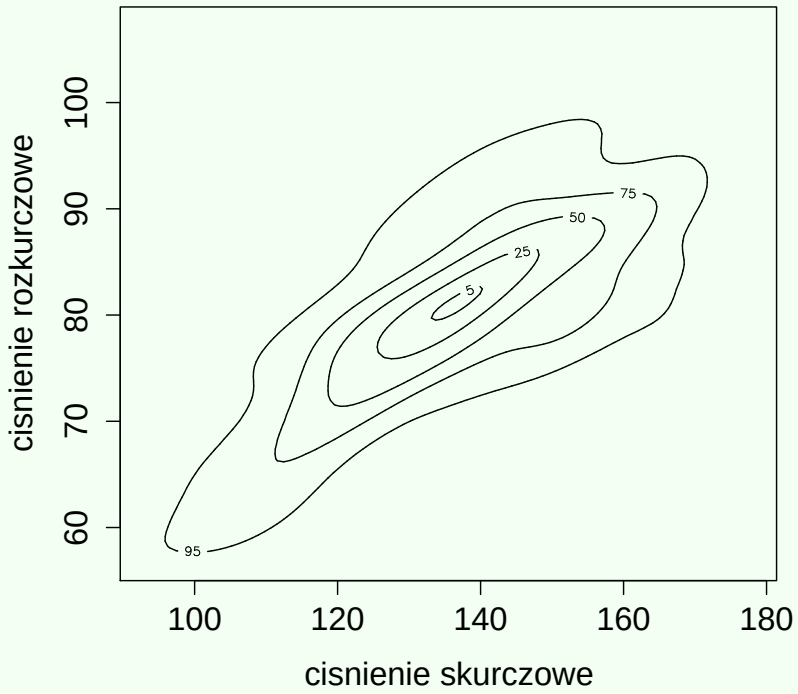


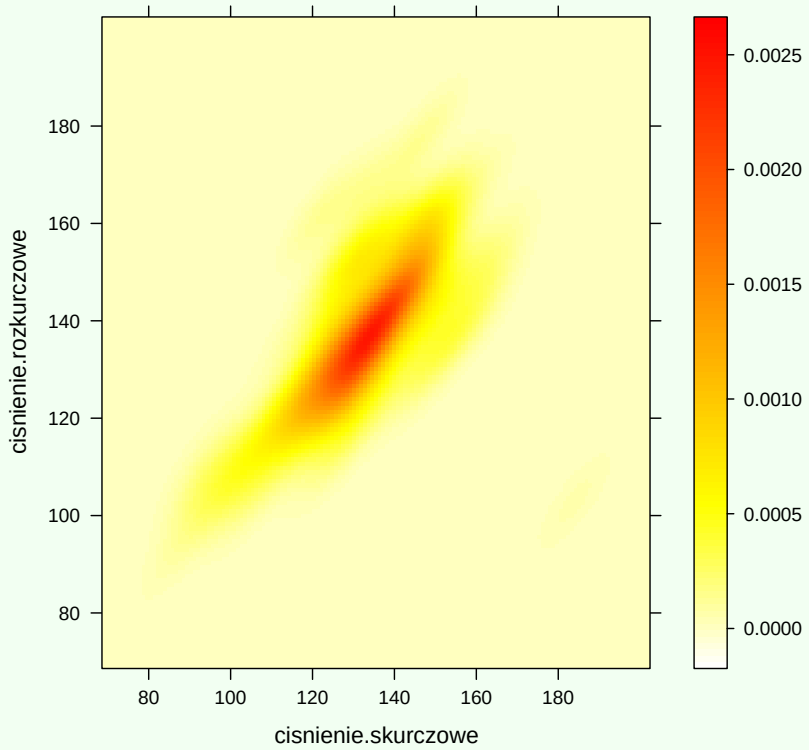


Warunkowy wykres rozrzutu, funkcja:coplot(graphics)



Wykresy konturowe, funkcje: contour(graphics), filled.contour(graphics), levelplot(lattice)

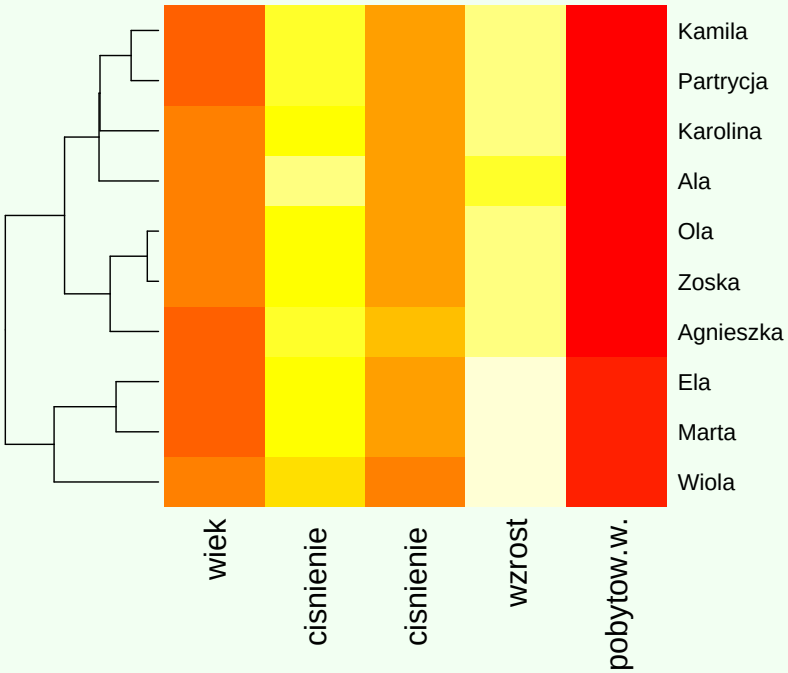


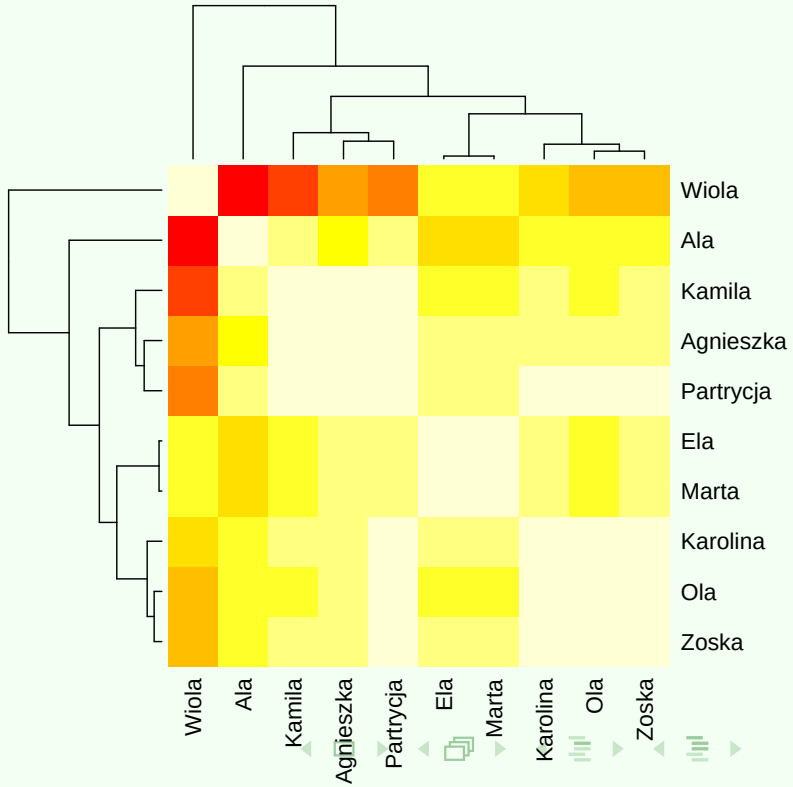


Mapa ciepła, funkcja: heatmap(stats)

Przykład w R

```
heatmap(osoby, margins=c(7,7), Colv=NA) # macierz liczb
heatmap(cor(t(osoby)), margins=c(7,7), symm=T) # macierz korelac
```





Jak to zrobić w pakiecie R?

W pakiecie R test na równość średnich można wykonać funkcją

```
t.test(x, y, alternative = c("two.sided", "less", "greater"),
       paired = FALSE, var.equal = FALSE)
```

- argument **x** określa pierwszy wektor obserwacji,
- argument **y** określa drugi wektor obserwacji,
- argument **alternative** określa jaka hipoteza alternatywna jest testowana,
- argument **paired** określa czy obserwacje są sparowane, czy nie,
- argument **var.equal** określa czy wariancje są równe w obu grupach.

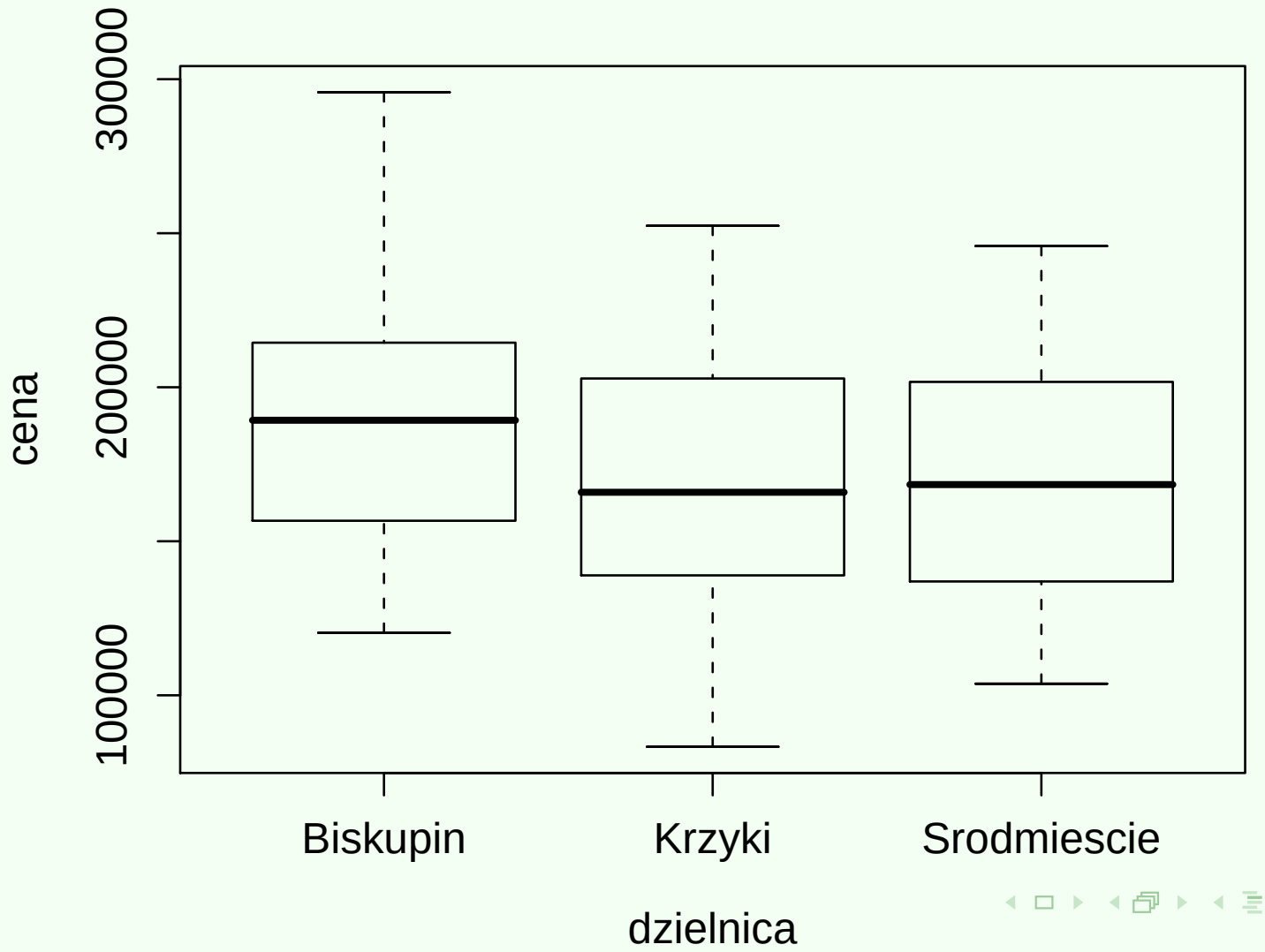
Jak to zrobić w pakiecie R?

Zacznijmy od przykładu z dwustronną alternatywą.

Przykład w R

```
> x
[1] 350 287 393 69 98 276 238 121 315 276
> y
[1] 334 253 339 313 364 292 302 409 351 476
> t.test(x, y)
Welch Two Sample t-test
data: x and y
t = -2.513, df = 14.334, p-value = 0.0245
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
-187.01365 -14.98635
sample estimates:
mean of x mean of y
242.3 343.3
```

Przykład dotyczący pieniędzy



Przykład

Interesuje nas weryfikacja hipotezy, czy średnie ceny mieszkań, w różnych dzielnicach, są równe.

Przykład w R

```
> mieszkania = read.table("http://www.biecek.pl/R/dane/daneMiesz
> (a1 = anova(lm(cena~dzielnica, data = mieszkania)))

Analysis of Variance Table
Response: cena
            Df     Sum Sq    Mean Sq F~value   Pr(>F)
dzielnica    2 1.7995e+10 8.9977e+09  5.0456 0.007294 **
Residuals  197 3.5130e+11 1.7833e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> pairwise.t.test(cena,dzielnica)
```

Testowanie hipotezy o równości parametrów skali

var.test(stats)	Test F dla dwóch prób do weryfikacji hipotezy o jednorodności wariancji. Można również testować hipotezę zerową, że iloraz wariancji wynosi **ratio** (domyślnie **ratio=1**).
ansari.test(stats)	Test Ansariego-Bradleya dla dwóch prób do weryfikacji hipotezy o równości parametrów skali.
bartlett.test(stats)	Test Bartletta dla wielu prób do weryfikacji hipotezy o jednorodności wariacji.
fligner.test(stats)	Test Flingera-Killeena dla wielu grup do weryfikacji hipotezy o jednorodności wariancji.
mood.test(stats)	Rangowy test Mooda dla dwóch prób do weryfikacji hipotezy o równości parametrów skali.
levene.test(lawstat)	Test Levene'a dla wielu prób do weryfikacji hipotezy o jednorodności wariancji.

Testowanie zgodności z rozkładem normalnym

cvm.test(...) Test Craméra-von Misesa.

ad.test(...) Test Andersona-Darlinga. W porównaniu do Cramera-von Misesa większą uwagę zwraca na ogony.

lillie.test(...) Test Lillieforsa, czyli test bazujący na statystyce Kolmogorova-Smirnova. Sprawuje się średnio gorzej niż dwa przedstawione powyżej testy.

pearson.test(...) Test χ^2 Pearsona. Liczbę klas, na które mają być dzielone obserwacje wyznaczana jest domyślnie ze wzoru $2n^{2/5}$, można też tę liczbę określić argumentem **n.klas**. Wartości krytyczne dla statystyki testowej są domyślnie wyznaczane z rozkładu $\chi^2_{n.klas-3}$.

shapiro.test(...) Test Shapiro-Wilka. Jeden z najbardziej popularnych i jednocześnie jeden z lepszych testów normalności.

sf.test(...) Test Shapiro-Francia jest modyfikacją testu Shapiro-Wilka.

Regresja liniowa w R

```
> summary(lm(cena~dzielnica+ powierzchnia+pokoi))

Call:
lm(formula = cena ~ dzielnica + powierzchnia + pokoi)

Residuals:
     Min       1Q   Median       3Q      Max
-30501.4  -8480.2   -144.1   7346.0  35729.6

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          94222.02    2320.36  40.607  < 2e-16 ***
dzielnicaKrzyki     -20934.86    1842.79 -11.360  < 2e-16 ***
dzielnicaSrodmiescie -12722.60    2008.03  -6.336 1.60e-09 ***
powierzchnia          2022.99     116.31  17.393  < 2e-16 ***
pokoi                   34.36    2157.52   0.016    0.987
---
```

Analizy wielowymiarowe

R oferuje bogaty zestaw funkcji do wykonywania analiz wielowymairowych

Metoda PCA

W metodzie PCA wyznaczany jest nowy układ współrzędnych w bazie oryginalnych zmiennych.
Przekształcone zmienne są więc kombinacjami liniowymi oryginalnych zmiennych.
Pierwsza nowa zmienna jest tak wyznaczana, by wariancja wyznaczona dla niej była możliwie największa. Kolejne zmienne są wyznaczane tak by były ortogonalne do poprzednich i również maksymalizowały wariancję.
Ta konstrukcja zmiennych powoduje, że nowe zmienne odpowiadają wektorom własnym kolejnych wartości własnych macierzy korelacji pomiędzy oryginalnymi zmiennymi.

Analiza skupień

- Analiza skupień to zbiór metod pozwalających na wyróżnienie zbiorów obserwacji (nazywanych skupieniami lub klastrami) podobnych do siebie.
- Proces szukania podziału na grupy, nazywany jest czasem klastrowaniem.
- W pakiecie R dostępnych jest bardzo wiele metod do przeprowadzania analizy skupień.
- Analiza skupień jest często wykorzystywana do wykrywania ukrytej struktury w zbiorze danych, np. na podstawie właściwości aminokwasów chcemy je podzielić na grupy tych najbardziej podobnych.

Metoda hierarchiczna

- AGNES (Agglomerative Nesting)
  - W pierwszym kroku każdy obiekt jest osobną grupą,
  - W kolejnym kroku znajdź dwie najbardziej podobne grupy i połącz je w jedną grupę,
  - Powtarzaj powyższy krok aż do otrzymania jednej grupy.
- DIANA (Divisive Analysis)
  - W pierwszym kroku wszystkie obiekty tworzą jedną grupę,
  - W kolejnym kroku znajdź najlepszy podział tej grupy na dwie podgrupy,
  - Powtarzaj powyższy krok aż każdy obiekt nie znajdzie się w jednoelementowej grupie.
- CLARA (Clustering Large Applications), metoda służąca do analizy dużych zbiorów danych, wywołuje metodę PAM na podzbiorach a następnie wybiera najlepszy zbiór medoidów na całym zbiorze
- FANNY (Fuzzy Analysis), klastrowanie rozmyte.

Analizy wielowymiarowe

Genetyka

Pakiet kinship, macierze pokrewieństwa i efekty losowe

Stopień pokrewieństwa pomiędzy dwoma osobnikami (kinship coefficient) to współczynnik określający prawdopodobieństwo, iż losowo wybrany allel od jednego osobnika jest identyczny przez pochodzenie z losowo wybranym allelem drugiego osobnika.

$$\Phi_{xy} = \sum_i \Phi_{ii} \left(\frac{1}{2}\right)^{n_i - 1} + \sum_j \sum_{j \neq k} \Phi_{jk} \left(\frac{1}{2}\right)^{n_{jk} - 2}$$

gdzie indeksy i, j i k przebiegają zbiór wszystkich wspólnych przodków osobników x i y, a n_i to długość ścieżki od osobnika x do osobnika y przez osobnika i, oraz odpowiednio n_{jk} to suma ścieżek od osobnika x do y poprzez j i k (którzy są spowinowaceni).

- Model poligeniczny opisuje wpływ wielu (K) małych genów (poligenów) mających niewielki addytywny wpływ na badana cechę.
- Estymacja wpływu każdego z tych genów przy obecnych rozmiarach próby jest niemożliwa, chcemy jednak uwzględnić te efekty w przeprowadzanych analizach.
- Jeżeli wiec model opisujący wpływ genów na cechę jest postaci

$$Y_i = \mu + \sum_l x_{i,l}\beta_l + \sum_k (\alpha_i^{k,1} + \alpha_i^{k,2}) + \varepsilon_i$$

gdzie i to numer osobnika, $x_{i,l}$ oznacza genotyp $l$tego silnego QTLa, β_l oznacza efekt $l$tego silnego QTLa (zakładamy, ze te QTLe jesteśmy w stanie zlokalizować), $\alpha_i^{k,1}$ oraz $\alpha_i^{k,2}$ oznaczają wpływy alleli $k$tego słabego QTLa, ε_i to szum środowiskowy, zmienne niezależne o rozkładzie normalnym.

Modele poligeniczne

Korelacje pomiędzy zmienna Y_i i Y_j można wyznaczyć i wynosi ona (dla populacji outbred przy założeniu równowagi Hardyego-Weinberga)

$$Cov(Y_i, Y_j) = 2\Phi^{i,j}\sigma_a^2 + \Delta_7^{i,j}\sigma_d^2$$

gdzie $\Phi^{i,j}$ to współczynnik pokrewieństwa pomiędzy osobnikami i i j, a $\Delta_7^{i,j}$ to skondensowany współczynnik Jacquarda (patrz rysunek 2). Współczynniki σ_a^2, σ_d^2 to dwie (ortogonalne) składowe kowariancji. Wyprowadzenie można znaleźć miedzy innymi w `http://nitro.biosci.arizona.edu/Notes/Lecture10.pdf`. `Współczynniki te można oszacować za pomocą modeli mieszanych wprowadzając jako macierz wariancji macierz V`

$$V = 2\Phi\sigma_a^2 + \Delta_7\sigma_d^2 + \mathcal{I}\sigma_e^2.$$

Co jest do zrobienia?

Możemy teraz badać

- estymacja współczynników σ_a^2, σ_d^2, σ_e^2. Tutaj wykorzystać można ML lub REML przy czym ML jest szybsze co ma znaczenie przy dużych populacjach (problemem jest odwracanie dużych macierzy),
- predykcja efektów losowych z wykorzystaniem mixed model equations (Henderson 1984). Dla modelu

$$y = X\beta + Z\gamma + \varepsilon$$

$$\begin{bmatrix} X^T\hat{R}^{-1}X & X^T\hat{R}^{-1}Z \\ Z^T\hat{R}^{-1}X & Z^T\hat{R}^{-1}Z + \hat{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} X^T\hat{R}^{-1}y \\ Z^T\hat{R}^{-1}y \end{bmatrix}$$

gdzie

$$R = \sigma_e^2 I,$$

$$G = \begin{bmatrix} \sigma_a^2 & 0 \\ 0 & \sigma_d^2 \end{bmatrix}.$$

Nadciśnienie u myszy a pakiet qtl

Dane dotyczą 250 męskich osobników wyhodowanych w krzyżówce backcross.
Zgenotypowano 174 markery dla każdego osobnika, w tym 4 na chromosomie X
Badano jedną cechę fenotypową - ciśnienie krwi.
Dla większości markerów genotypty są dostępne tylko dla osobników z skrajnymi wartościami fenotypów.
Sugiyama, F., Churchill, G. A., Higgens, D. C., Johns, C., Makaritsis, K. P., Gavras, H. and Paigen, B. (2001) Concordance of murine quantitative trait loci for salt-induced hypertension with rat and human loci. Genomics 71, 70–77.

Warto zobaczyć

- **„Przewodnik po pakiecie R”**, GiS 2008, Przemysław Biecek,
- **„The R Book”**, Michael Crawley. Wiley-Blackwell, ISBN: 9780470510247.
- Książka **„R Graphics”**, Paul Murrell. Chapman & Hall/CRC Computer Science & Data Analysis, ISBN: 9781584884866.
- Strona z przykładami ciekawych wykresów wykonanych w R http://addictedtor.free.fr/graphiques/.
- Opis pakietu **ggplot2** wraz z wieloma przykładami http://had.co.nz/ggplot2/.
- **„S Programming”**, William Venables i B.D. Ripley. Springer, ISBN: 9780387989662.
- **„Using R for Introductory Statistics”**, John Verzani. Chapman & Hall/CRC Computer Science & Data Analysis, ISBN: