

2.1 Statystyki opisowe

Opisywać będziemy próby proste - czyli wektory obserwacji wylosowanych z tego samego rozkładu. W pakiecie `coin` dostępny jest zbiór danych `glioma`. Zawiera on dane dotyczące 37 pacjentów. Dla każdego pacjenta mamy wybrane dane socjodemograficzne, czyli wiek i płeć, dane kliniczne, czyli informacje czy pacjent dożył zakończenia badania, informacje o czasie życia pacjenta, informacje o wyniku histologii. Poniżej wymienione statystyki opisowe pozwalają na podsumowanie tego zbioru w przejrzysty sposób.

2.1.1 Proste podsumowania

Funkcja: `summary`

Funkcja wyświetlająca podsumowania to `summary()`. W przypadku zmiennej występującej na kilku poziomach, funkcja ta wyznaczy licznosc obserwacji posiadających danych poziom.

```
> summary(glioma$sex)
Female   Male
      16    21
```

W przypadku zmiennych liczbowych funkcja ta zwraca informacje o minimum, maksimum, obu kwartylach, średniej i medianie.

```
> summary(glioma$age)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
19.00  40.00   47.00  48.49  57.00   83.00
```

Argumentem funkcji `summary()` może być również obiekt `data.frame`, w tym przypadku podsumowanie zostanie wyznaczone dla każdej kolumny.

Funkcja: `table`

Jeżeli interesuje nas częstość występowania pewnego czynnika w danych w rozbiu na inny czynnik, powinniśmy wykorzystać funkcję `table()` wyznaczającą tablicę kontyngencji dla dwóch zmiennych jakościowych

```
> table(glioma$sex, glioma$histology)
      GBM Grade3
Female    10     6
Male     10    11
```

Funkcja: `range`

Funkcja `range()` wyznacza zakres zmienności (czyli minimum i maksimum) wybranej zmiennej

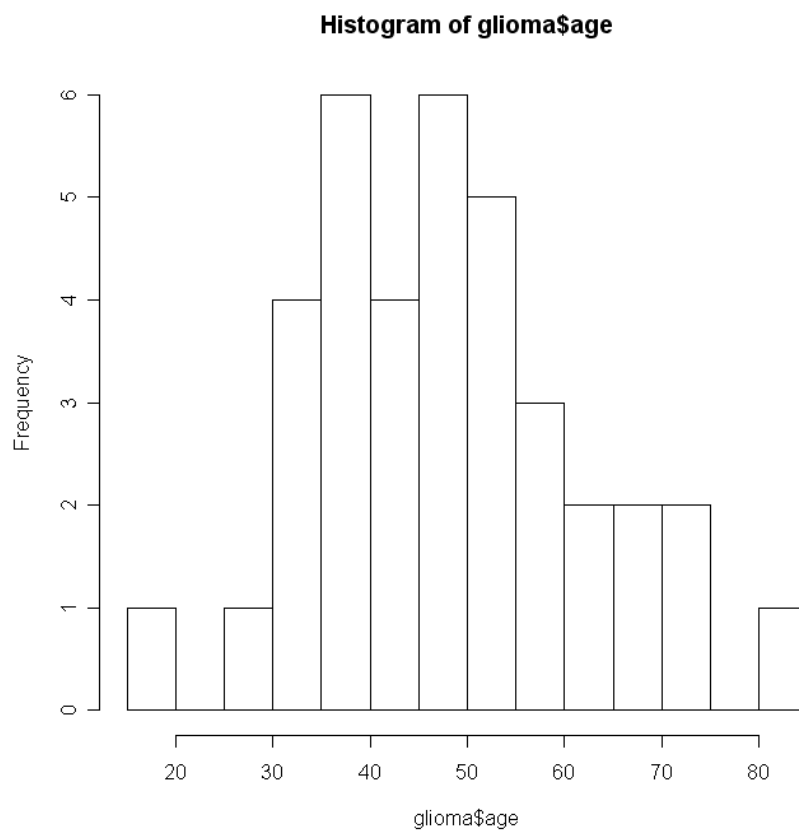
```
> range(glioma$age)
[1] 19 83
```

2.1.2 Podsumowania graficzne

Funkcja: hist

Najpopularniejsza statystyka graficzna. Przedstawia licznosci pacjentów w poszczególnych przedziałach (nazywanych też kubełkami) danej zmiennej. Domyślnie w funkcji histogram liczba kubełków dobierana jest w zależności od liczby obserwacji jak i ich zmienności. Możemy jednak subiektywnie wybrać interesującą nas liczbę kubełków.

```
> hist(glioma$age, 10)
```



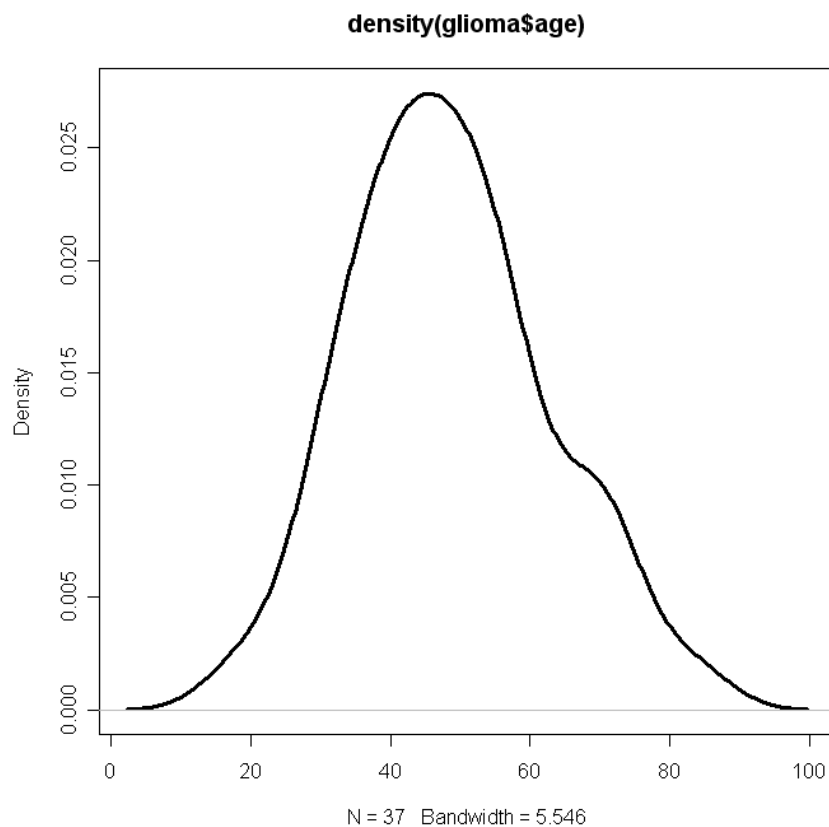
Rysunek 2.1: Histogram dla zmiennej wiek.

Funkcja: density

Jeżeli przeszkadzają komuś kanciaste brzegi histogramu, to może wykorzystywać jego wygładzoną wersję - jądrowy estymator gęstości. Pomimo zakręconej nazwy, z funkcji `density()` korzysta się naprawdę prosto, poniżej przykład. Stopień wygładzenia jest dobierany automatycznie (tak jak liczba kubeków dla histogramu), można go zmieniać manipulując argumentem `bw`.

Domyślnie wyznaczane jest wygładzanie jądrem gausowskim, przeglądając pomoc dla tej funkcji zainteresowany czytelnik odkryje jak wykorzystywać inne jądra i czym one się różnią.

```
> plot(density(glioma$age))
```



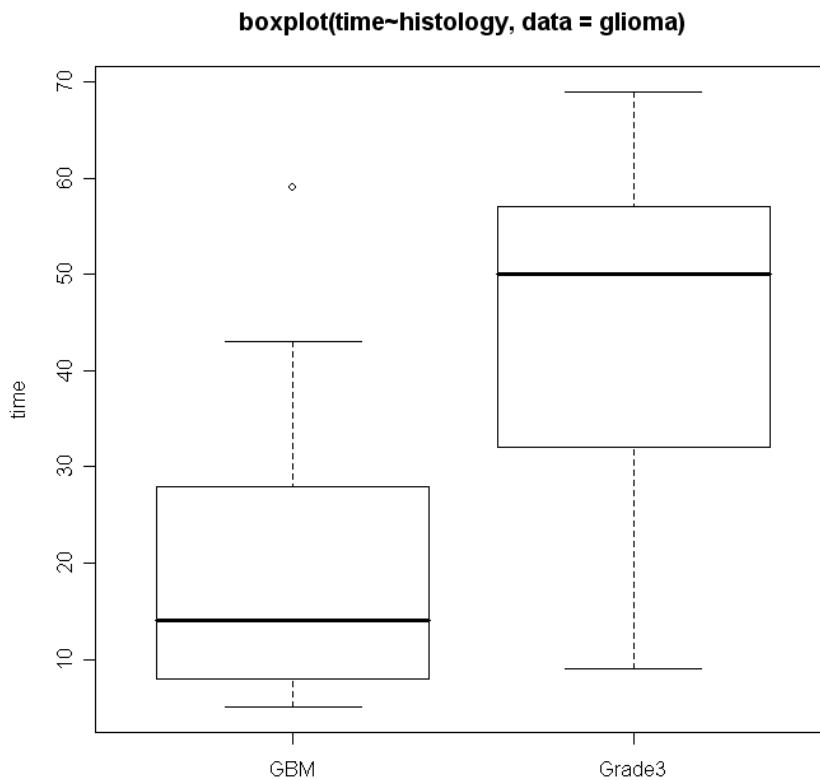
Rysunek 2.2: Estymator jądrowy gęstości dla zmiennej wiek.

Funkcja: boxplot

Wykres pudełkowy można wyznaczać dla pojedynczej zmiennej, dla kilku zmiennych lub dla pojedynczej zmiennej w rozbiciu na grupy. Wykres przedstawia medianę (środek pudełka), kwartyły (dolna i górna granica pudełka), obserwacje odstające (zaznaczone kropkami) oraz maksimum i minimum po usunięciu obserwacji odstających.

Wykres pudełkowy jest bardzo popularną metodą prezentacji zmienności pojedynczej zmiennej.

```
> boxplot(time~histology, data = glioma)
> boxplot(glioma$time, glioma$age)
```



Rysunek 2.3: Wykres pudełkowy dla zmiennej time w rozbiciu na grupy zmiennej histology.

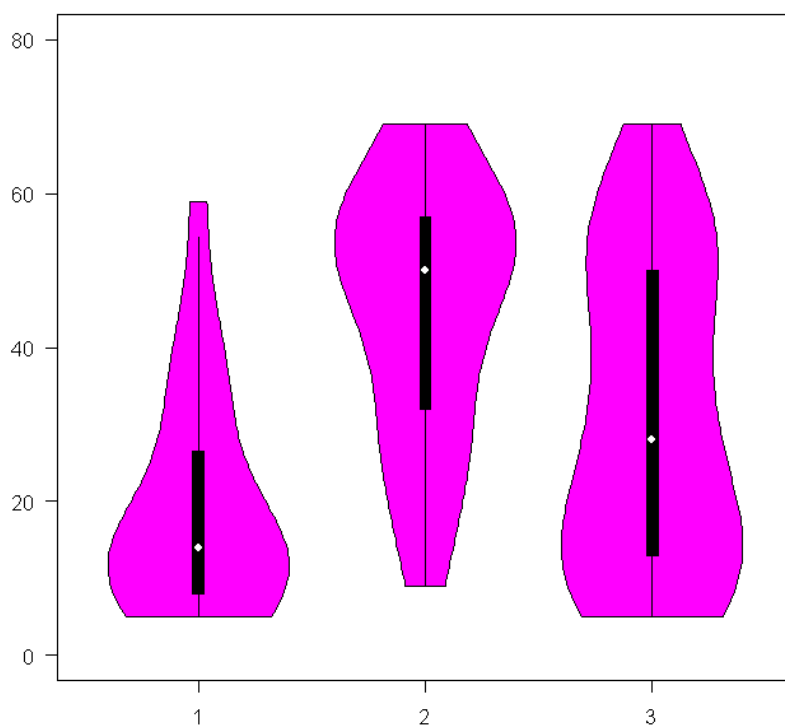
Funkcja: vioplot

Funkcja `vioplot()` znajduje się w pakiecie o tej samej nazwie. Z uwagi na wygląd wyników nazywany jest wykresem skrzypcowym.

W środku każdego „skrzypiec” przedstawione są wykresy pudełkowe, a szerokość skrzypiec w punkcie x odpowiada natężeniu obserwacji o wartości cechy zbliżonej do x .

Można go traktować jako wygładzoną wersję wykresu pudełkowego. Przydatna szczególnie w przypadku danych o wielomodalnym rozkładzie (patrz prawy obrazek, widocznej tu dwumodalności nie było by widać na wykresie pudełkowym)

```
> vioplot(glioma$time[glioma$histology=="GBM"],  
+ glioma$time[glioma$histology=="Grade3"], glioma$time)
```



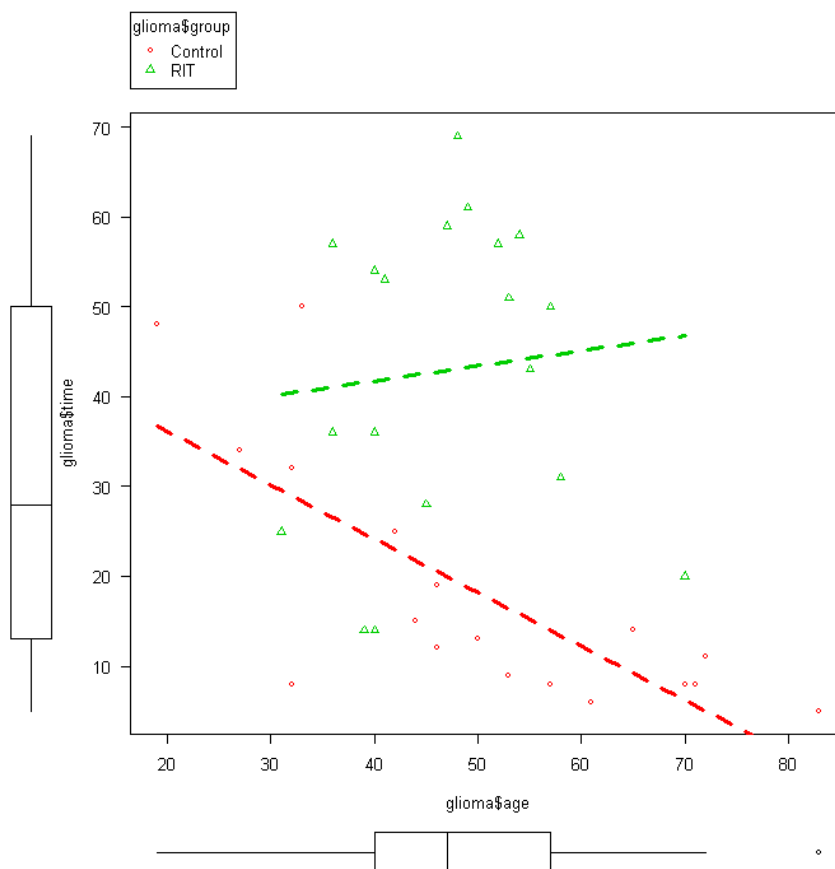
Rysunek 2.4: Wykres skrzypcowy w rozbiciu na grupy dla zmiennej histology.

Funkcja: scatterplot

Wykres rozrzutu z biblioteki `car` pozwala na przedstawienie zależności pomiędzy parą zmiennych. Domyślnie na osiach rysowane są wykresy pudełkowe dla poszczególnych zmiennych, a dla danych wyznaczana jest prosta regresji. W prezentowanym przypadku wykres rozrzutu został rozbity ze względu na zmienną grupą.

Wykresy rozrzutu są bardzo przydatne do wykrywania i opisywania liniowych zależności pomiędzy parą zmiennych (nieliniowe też będą widoczne).

```
> scatterplot(glioma$age, glioma$time, groups=glioma$group, smooth=F)
```



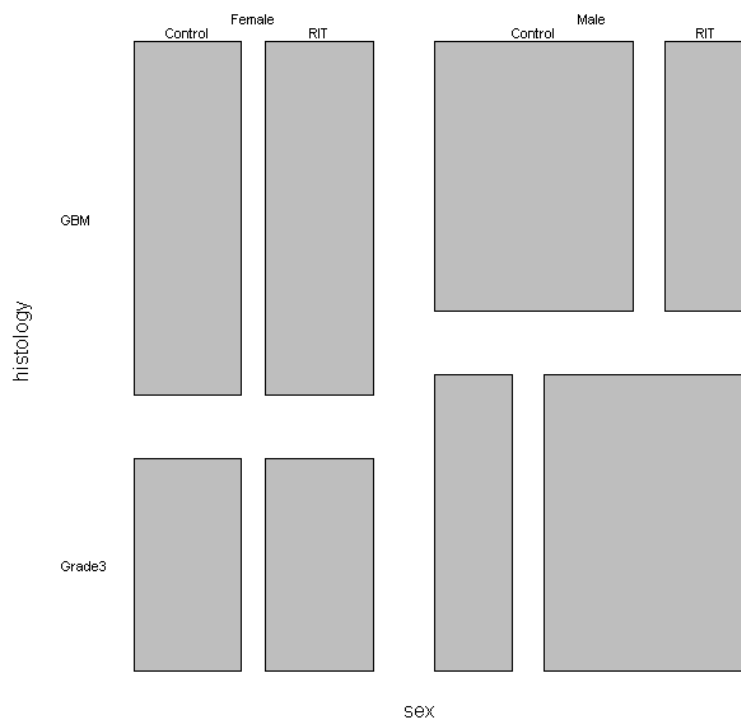
Rysunek 2.5: Wykres rozrzutu dla zmiennych time i age w rozbiciu na grupy.

Funkcja: mosaicplot

Odpowiednikiem wykresu rozrzutu dla zmiennych jakościowych jest wykres mozaikowy. Na tym wykresie pola obszarów są proporcjonalne do liczby przypadków w poszczególnych grupach czynników.

Warto zaznaczyć się z tego typu wykresami, na jednej mozaice można umieścić dużo informacji.

```
> mosaicplot(~sex+histology+group, glioma)
```



Rysunek 2.6: Wykres mozaikowy.