Graficzne przedstawienie danych — szereg rozdzielczy, histogram

dr Mariusz Grządziel

Katedra Matematyki, Uniwersytet Przyrodniczy we Wrocławiu

Wykład 2; r. akad. 2022/2023

Przykład

Zbiór danych *trees* należący do pakietu *trees* systemu R składa się z pomiarów wykonanych dla 31 drzew czeremchowych trzech zmiennych (wektorów):

- Girth — średnica drzewa na wysokości (jednostka: cal);
- Height — wysokość drzewa (jednostka: stopa);
- Volume — objętość drzewa wyrażona w stopach sześciennych

Dane dotyczące średnicy drzew (w calach):

```
8.3 8.6 8.8 10.5 10.7
10.8 11.0 11.0 11.1 11.2
11.3 11.4 11.4 11.7 12.0
12.9 12.9 13.3 13.7 13.8
14.0 14.2 14.5 16.0 16.3
17.3 17.5 17.9 18.0 18.0 20.6
```

Histogram i szereg rozdzielczy

Dla zbioru danych liczbowych $y_1, y_2 \ldots, y_n$ niech:
$MIN1$ oznacza liczbę mniejszą lub równą niż najmniejsza z liczb $y_1, y_2 \ldots, y_n$;
$MAX1$ oznacza liczbę większą lub równą niż największa z liczb $y_1, y_2 \ldots, y_n$; $MIN1 \leqslant MIN$ i $MAX1 \geqslant MAX$ mogą być odpowiednimi „zaokrągleniami” wartości, odpowiednio, minimalnej i maksymalnej naszego zbioru danych. (MIN i MAX oznaczają, odpowiednio, wartość minimalną i maksymalną dla $\{y_1, y_2 \ldots, y_n\}$).
Podzielmy odcinek $[MIN1, MAX1]$ na k przedziałów (zwanych klasami) o równej długości:

$$[x_0, x_1], (x_1, x_2], \ldots, (x_{k-1}, x_k], \text{ gdzie } x_0 = MIN1, x_k = MAX1$$

Funkcję przyporządkowującą poszczególnym przedziałom liczbę elementów naszego zbioru danych do nich należących będziemy nazywać szeregiem rozdzielczym.

Ustalenie liczby klas w szeregu rozdzielczym

Istnieje kilka reguł ustalania liczby klas k szeregu rozdzielczego w zależności od liczby obserwacji n. Oto niektóre z nich:

$$k \approx \log_2 n + 1; \quad k \approx \sqrt{n}.$$

Szereg rozdzielczy dla danych dotyczących średnic pni drzew czeremchowych

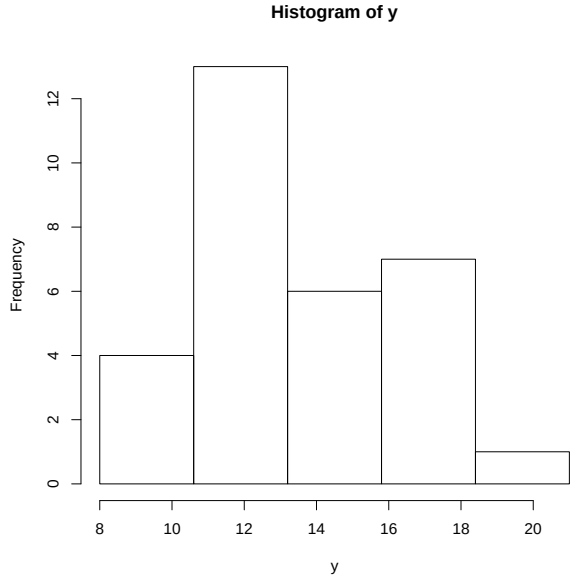
Szereg rozdzielczy dla danych dotyczących średnich pni drzew czeremchowych

```
8.3 8.6 8.8 10.5 10.7
10.8 11.0 11.0 11.1 11.2
11.3 11.4 11.4 11.7 12.0
12.9 12.9 13.3 13.7 13.8
14.0 14.2 14.5 16.0 16.3
17.3 17.5 17.9 18.0 18.0 20.6
```

obliczamy: $MIN = 8.3$, $MAX = 20.6$
Przyjmujemy: $MIN1 = 8$; $MAX1 = 21$ oraz $k = 5$.
Otrzymujemy szereg rozdzielczy, przedstawiony w postaci tabeli:

klasa	$(8.0, 10.6]$	$(10.6, 13.2]$	$(13.2, 15.8]$	$(15.8, 18.4]$	$(18.4, 21.0]$
liczność	4	13	6	7	1

Histogram liczebności dla danych dotyczących średnic pni drzew czeremchowych



Wykres został wykonany przy użyciu sekwencji poleceń systemu R:

```
y=trees$Girth
max(y)
# 20.6
min(y)
# 8.3
hist(y,seq(from=8,to=21,length.out=6))
```

Histogram częstości

Jeśliby otrzymay histogram liczebności przeskalować w ten sposób, że wysokości słupków odpowiadałyby ilorazom liczebności klas i liczby wszystkich obserwacji n, wtedy otrzymalibyśmy histogram częstości. Wysokości słupków tego histogramu byłyby równe:

$$\frac{4}{31} \approx 0{,}13; \quad \frac{13}{31} \approx 0{,}42 \text{ itd.}$$

Histogram probabilistyczny

Jeśliby histogram przeskalować tak, aby suma pól wszystkich prostokątów („słupków”) była równa 1, otrzymamy tzw. histogram probabilistyczny (od *probability* (ang.) - prawdopodobieństwo).
Wtedy

$$h_i = \frac{n_i}{nh},$$

gdzie h_i oznacza wysokość i-tego słupka w histogramie probabilistycznym, n_i liczebność i-tej klasy, n liczebność próby, h szerokość klasy.
Histogram probabilistyczny: oszacowanie rozkładu jedności prawdopodobieństwa dla danej cechy.
Zdefiniujmy funkcję H, określoną wzorem:

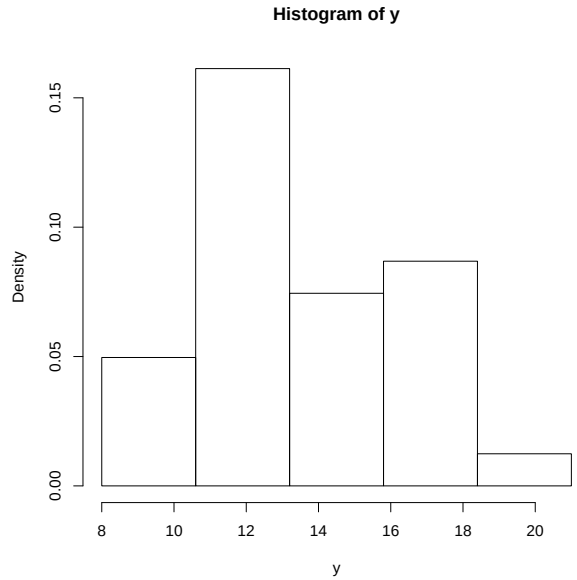
$$H(x) = \begin{cases} h_i, & \text{jeżeli } x \quad \text{należy do i-tej klasy ;} \\ 0, & \text{jeżeli } x < MIN1 \quad \text{lub} \quad x > MAX1; \end{cases}$$

dziedziną funkcji H jest zbiór liczb rzeczywistych.

Histogram probabilistyczny i pojęcie prawdopodobieństwa

Pole pod wykresem funkcji H nad odcinkiem $[a, b]$ — oszacowanie prawdopodobieństwa, że wartość cechy reprezentowanej przez obserwacje $y_1, y_2, \ldots$ należy do $[a, b]$.

Histogram probabilistyczny dla danych dotyczących średnic pni drzew czeremchowych



Wykres uzyskano przy użyciu sekwencji poleceń:

```
y=trees$Girth
max(y)
# 20.6
min(y)
# 8.3
hist(y,seq(from=8,to=21,length.out=6),prob=TRUE)
```