DOI: 10.37317/biul-2011-0058

MAŁGORZATA TARTANUS [1]
MARCIN KOZAK [2]
AGNIESZKA WNUK [2]
[1] Instytut Ogrodnictwa, Skierniewice
[2] Katedra Doświadczalnictwa i Bioinformatyki, Szkoła Główna Gospodarstwa Wiejskiego, Warszawa

Wykorzystanie diagramu łodyga i liście do analizy danych w środowisku R*

Use of stem-and-leaf diagram for data analysis in R environment

Szczególną rolę w analizie danych odgrywa badanie rozkładu zmiennych ilościowych; w tym celu najlepsze okazują się być techniki graficzne. Wśród bogatej oferty różnego rodzaju wykresów wykorzystywanych w tym celu są przede wszystkim histogram, wykres gęstości prawdopodobieństwa i wykres pudełkowy (ang. boxplot). W pracy przedstawiono rzadziej stosowany diagram nazywany łodyga i liście (ang. stem-and-leaf plot), zaliczany do graficzno-tabelarycznych metod wizualizacji danych. Ten wykres wnosi wiele szczegółów do opisu rozkładu zmiennej. Dla zobrazowania przykładowych wykresów wykorzystujemy środowisko R.

Słowa kluczowe: rozkład zmiennej, statystyka, wizualizacja danych

Studying distributions of quantitative traits plays a special role in data analysis; the best tools for that purpose appear to be graphical methods. Among various types of plots, most important ones are histograms, density plots and box-plot. In this paper, we introduce a less frequently applied stem-and-leaf diagram, which is a graphical-tabular technique of data visualization. This simple and interesting technique, despite its simplicity, can provide quite deep information about a variable's distribution. To present how the stem-and-leaf display works, we have used R environment.

Key words: data visualization, statistics, variable distribution

WSTĘP

Diagram łodyga i liście (ang. stem-and-leaf plot) jest graficzno-tabelaryczną formą prezentacji rozkładu zmiennej ilościowej. Diagram ten został zaproponowany przez Tukeya (1977). Niekiedy wykres ten utożsamiany bywa z szeregiem rozdzielczym, który ma na celu podział danych ilościowych na przedziały (klasy) oraz określenie częstości wystąpienia (liczności) w każdej klasie podziału.

* Praca była prezentowana w ramach I Warsztatów Biometrycznych, które odbyły się w IHAR-PIB w Radzikowie w dniach 14-15 września 2010 r.

Diagram zaproponowany przez Tukeya (1977) ma czasem dość skomplikowaną strukturę, a jego czytanie nie zawsze jest proste. Jest to o tyle niewygodne, że zaletą tego typu prezentacji graficzno-tabelarycznej powinna być jego prostota: nawet uczniowie początkowych klas szkoły podstawowej nie powinni mieć problemów ze zrozumieniem jego prostych postaci (Pereira-Mendoza i Dunkels, 1989; Perry i in., 1999). Postać diagramu omawiana przez Beckera i in. (1988) jest bardzo prosta i użyteczna w wielu różnych sytuacjach badawczych, nie wymagając przy tym długiej nauki zasad konstrukcji i czytania wykresu, w przeciwieństwie do klasycznej postaci diagramu, zaproponowanej przez Tukeya (1977).

Celem pracy jest prezentacja użyteczności jednej z prostszych postaci diagramu łodyga i liście, w wersji opisanej przez Beckera i in. (1988). Zaprezentowano zasady tworzenia i czytania tego typu diagramów, a także kod w środowisku R (R Development Core Team, 2010), który może zostać wykorzystany w trakcie analizy danych.

DIAGRAM ŁODYGA I LIŚCIE W R

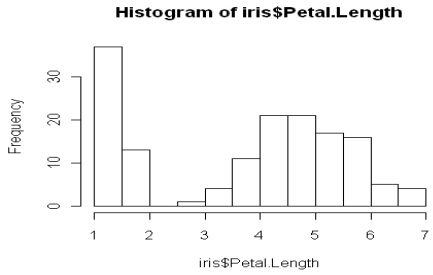
Środowisko R posiada bogatą ofertę funkcji do graficznej prezentacji rozkładu zmiennej. Wiele z nich znajduje się w standardowym pakiecie `graphics`, ale istnieje wiele dodatkowych pakietów (bibliotek), które w dużym stopniu zwiększają możliwości środowiska R. Podstawowymi takimi funkcjami są funkcje `hist` oraz `density` (oba dostępne w pakiecie `graphics`), które tworzą odpowiednio histogram i szacują funkcję gęstości prawdopodobieństwa; funkcja `density` wymaga połączenia z funkcją `plot`, aby można było narysować funkcję gęstości prawdopodobieństwa rozkładu. Poniżej wykorzystanie tych funkcji dla długości płatka kwiatu (ang. petal length) ze słynnego zbioru dotyczącego trzech gatunków irysa (Anderson, Edgar, 1935; Fisher, 1936):

```
> hist(iris$Petal.Length)
```

Histogram ten przedstawiony jest na rys. 1. Wykres gęstości prawdopodobieństwa (rys. 2) uzyskamy poprzez:

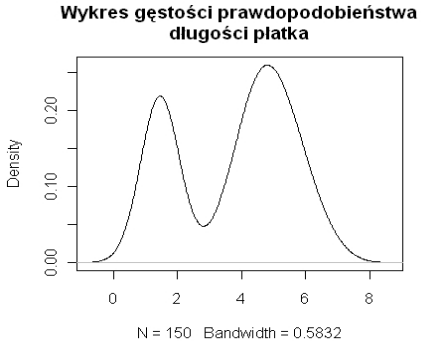
```
> plot(density(iris$Petal.Length), main = "Wykres gęstości prawdopodobieństwa\n długości płatka")
```

(Znak `\n` oznacza przejście do kolejnej linii).



Rys. 1. Histogram przedstawiający rozkład długości płatka dla trzech gatunków irysa (Anderson, 1935; Fisher, 1936)

Fig. 1. Histogram of petal length for three iris species pooled together (Anderson, 1935;Fisher, 1936)



Rys. 2. Wykres gęstości dla długości płatka kwiatu trzech gatunków irysa (Anderson, 1935; Fisher, 1936)
Fig. 2. Density plot of petal length for three iris species pooled together (Anderson, 1935; Fisher, 1936)

Zauważmy, że dane są pogrupowane: mamy po 50 obserwacji dla trzech gatunków irysa: *Iris setosa*, *I. versicolor* i *I. virginica*; powyższe wykresy wyraźnie pokazują, że dane nie są jednorodne. Pakiet `lattice` (Sarkar, 2008) daje doskonałe możliwości tworzenia podobnych wykresów dla danych pogrupowanych (wykresy nieprzedstawione):

```
> library(lattice)
> histogram(~ Petal.Length | Species, data = iris, layout =
c(1, 3), aspect = .7)
> densityplot(~ Petal.Length, data = iris, groups = Species,
auto.key = list(columns = 3), plot.points = FALSE)
```

Łatwo zauważyć, że tworzenie wykresów w pakiecie `lattice` nie jest proste (choć większość z argumentów wykorzystanych powyżej nie jest konieczna, a jedynie ma za zadanie ułatwić czytanie wykresu). Podobne wykresy można stworzyć w pakiecie `ggplot2` (Wickham, 2009).

Przejdźmy do diagramu łodyga i liście. Wywołanie go jest równie proste co histogramu:

```
> stem(iris$Petal.Length)
```

co poskutkuje następującym diagramem:

```
  The decimal point is at the |

  1 | 012233333334444444444444
  1 | 55555555555556666666777799
  2 |
  2 |
  3 | 033
  3 | 55678999
  4 | 000001112222334444
  4 | 55555555666777778888999999
  5 | 000011111111223344
  5 | 55566666677788899
```

```
6 | 0011134
6 | 6779
```

Funkcja `stem` (ze standardowego pakietu `graphics`) tworzy diagram łodyga i liście przedstawiony przez Beckera i in. (1988). Klasyczną postać diagramu, zaproponowaną przez Tukeya (1977) i nieomawianą tutaj, można stworzyć np. przy pomocy funkcji `stem.leaf {aplpack}` i `stem.leaf {QCAGUI}` (w klamrach znajduje się pakiet, w którym dostępna jest dana funkcja).

Powyżej, przy domyślnych ustawieniach parametrów funkcji `stem` (o których wspomniano poniżej) uzyskano podział szeregu rozdzielczego na 12 klas (gałęzi). Interpretacja otrzymanego diagramu jest następująca. Istotnym elementem wykresu jest komunikat: `"The decimal point is at the |"` wyświetlany na początku, w którym określone jest miejsce separatora dziesiętnego. Miejsce to jest wyznacznikiem podziału wykresu na dwie części: łodygę (po lewej stronie symbolu `|`) i gałęzie, złożone z liści (po prawej stronie symbolu `|`). Postać diagramu wymaga, by wartości liczbowe na nim przedstawione zostały zaokrąglone — zależnie od rzędu wartości, może to być zaokrąglenie do wartości dziesiętnych, ale też może się zdarzyć, że liczby zaokrąglane są do dziesiątek (zamiast 1983,5 podana byłaby wartość 198), setek (24 zamiast 2436,3) itp. Dla przykładu, komunikat na powyższym wykresie informuje o tym, że separator dziesiętny znajduje się w miejscu `|`.

Wykres przedstawia tyle obserwacji, ile jest liści na wykresie. Czyli jeden liść informuje o wartości jednej obserwacji. Informację o danej wartości można odczytać na podstawie powyżej wspomnianego komunikatu o położeniu separatora dziesiętnego, ale też zawsze warto spojrzeć, jak wyglądają konkretne wartości badanej cechy, np.:

```
> min(iris$Petal.Length)
> max(iris$Petal.Length)
```

Wartość minimalna wynosi 1, a na diagramie przedstawiona jest jako `1|0` (pierwszy liść na najwyższej gałęzi), zaś maksymalna 6.9 i `6|9` (ostatni liść na najniższej gałęzi). Łatwo więc się domyślić, jak będą wyglądały pozostałe wartości. Np. patrząc na wiersz piąty, widzimy jedną łodygę i 3 liście. Wartość łodygi wynosi 3, a zatem wiersz ten przedstawia następujące wartości cechy: 3,0, 3,3 i 3,3. Tu również możemy zauważyć, że określenie wartości `min` i `max` jest niczym innym jak ustaleniem przedziału (zakresu) dla oryginalnych wartości badanej cechy — w naszym przypadku jest to przedział <1,0, 6,9>. Z podobną łatwością możemy ustalić przedziały dla każdej z łodyg. Na przykład łodyga pierwsza (wiersz pierwszy diagramu) przyjmuje wartość 1, natomiast w kolumnie liści (dla tej łodygi) najmniejsza wartość to 0, a najwyższa 4, zatem ustalony przedział to <1,0, 1,4>. Tabela 1 podaje pełny wykaz wartości rozpatrywanej zmiennej i ich rozpisanie w powyższym diagramie łodyga i liście.

Funkcja `stem` posiada trzy dodatkowe argumenty, spośród których najbardziej przydatne są dwa: `scale` i `width`, które kontrolują odpowiednio wysokość i szerokość diagramu.

Tabela 1

Tabelaryczna interpretacja diagramu łodyga i liście
tabular interpretation of the stem - and - leaf diagram

Cyfra przed separatorem (Stem)	Cyfra po separatorze (Leaf)	Liczba Number	Przedział Range
1	01223333333444444444444	1,0 1,1 1,2 1,2 1,3 1,3 1,3 1,3 1,3 1,3 1,3 1,4 1,4 1,4 1,4 1,4 1,4 1,4 1,4 1,4 1,4 1,4 1,4	<1,0 , 1,4>
1	55555555555556666666777799	1,5 1,5 1,5 1,5 1,5 1,5 1,5 1,5 1,5 1,5 1,5 1,5 1,5 1,5 1,6 1,6 1,6 1,6 1,6 1,6 1,6 1,7 1,7 1,7 1,7 1,9 1,9	<1,5 , 1,9>
2			–
2			–
3	033	3,0 3,3 3,3	<3,0 , 3,4>
3	55678999	3,5 3,5 3,6 3,7 3,8 3,9 3,9 3,9	<3,5 , 3,9>
4	000001112222334444	4,0 4,0 4,0 4,0 4,0 4,1 4,1 4,1 4,2 4,2 4,2 4,2 4,3 4,3 4,4 4,4 4,4 4,4	<4,0 , 4,4>
4	55555555666777778888999999	4,5 4,5 4,5 4,5 4,5 4,5 4,5 4,5 4,6 4,6 4,6 4,7 4,7 4,7 4,7 4,7 4,8 4,8 4,8 4,8 4,9 4,9 4,9 4,9 4,9	<4,5 , 4,9>
5	000011111111223344	5,0 5,0 5,0 5,0 5,1 5,1 5,1 5,1 5,1 5,1 5,1 5,1 5,2 5,2 5,3 5,3 5,4 5,4	<5,0 , 5,4>
5	55566666677788899	5,5 5,5 5,5 5,6 5,6 5,6 5,6 5,6 5,6 5,7 5,7 5,7 5,8 5,8 5,8 5,9 5,9	<5,5 , 5,9>
6	0011134	6,0 6,0 6,1 6,1 6,1 6,3 6,4	<6,0 , 6,4>
6	6779	6,6 6,7 6,7 6,9	<6,5 , 6,9>

Argument `scale` zmienia liczbę klas podziału, w ten sposób modyfikując wysokość diagramu. Domyślnie `scale = 1`, więc zmniejszając (zwiększając) go dwa razy, uzyskamy diagram dwa razy krótszy (dłuższy):

```
> stem(iris$Petal.Length, scale=0.5)

  The decimal point is at the |

  1 | 012233333334444444444445555555555556666666777799
  2 |
  3 | 03355678999
  4 | 000001112222334444555555556667777888899999
  5 | 000011111111223344555666666677788899
  6 | 00111346779
```

W tym przypadku przedstawienie wartości minimalnej (1,0) jak i maksymalnej (6,9) nie uległo zmianie, ale zmniejszyła się liczba gałęzi do 6 oraz zakres każdej z gałęzi (np. gałąź pierwsza ma teraz zakres <1,0, 1,9>.

```
> stem(iris$Petal.Length, scale=2)
```

Oto fragment diagramu, jaki uzyskamy („. . .” reprezentuje nie pokazane gałęzie):

```
The decimal point is 1 digit(s) to the left of the |

  10 | 00
  12 | 000000000
  14 | 00000000000000000000000000
```

```
16 | 00000000000
18 | 00
20 |
22 |
24 |
26 |
28 |
30 | 0
.....
62 | 0
64 | 0
66 | 000
68 | 0
```

Tym razem wartość minimalna (1,0) jest przedstawiona jako `10|0`, zaś maksymalna (6,9) jako `68|0`. Wartości zostały tym razem zaokrąglone tak, by znalazły się w klasach zawierających wartości <1,0, 1.2), <1,2, 1,4) itd.

Argument `width` określa szerokość wykresu. Jego domyślne ustawienie zazwyczaj pozwala na wyświetlenie wszystkich liści w poszczególnych gałęziach. Jednak w niektórych przypadkach (gdy przewidujemy, że w jednej klasie przedziałów może znaleźć się więcej niż 80 wartości, choć takich wykresów raczej należy unikać, gdyż przestają być czytelne), należy odpowiednio zwiększyć wartość tego argumentu. Drugą przydatną wartością dla tego argumentu jest `width = 0`, dzięki której można odczytać liczbę liści w każdej z łodyg:

```
> stem(iris$Petal.Length, width=0)

  The decimal point is at the |
  1 | +24
  1 | +26
  2 |
  2 |
  3 | +3
  3 | +8
  4 | +18
  4 | +25
  5 | +18
  5 | +17
  6 | +7
  6 | +4
```

Jak już wcześniej wspomniano, omawiane dane są pogrupowane, ponieważ dotyczą kwiatów trzech gatunków irysa: *Iris setosa*, *I. versicolor* i *I. virginica.* Widać to było zarówno na rysunkach 1 i 2, jak i na powyższych diagramach. Diagram łodyga i liście można oczywiście tworzyć również dla wybranej grupy danych, jak tutaj dla *I. setosa*:

```
> stem(iris$Petal.Length[iris$Species == "setosa"])
```

```
  The decimal point is 1 digit(s) to the left of the |
 10 | 0
 11 | 0
 12 | 00
 13 | 0000000
 14 | 0000000000000
 15 | 0000000000000
 16 | 0000000
 17 | 0000
 18 |
 19 | 00
```

Aby przy pomocy jednej komendy stworzyć trzy diagramy dla grup, można wykorzystać np. funkcję `tapply` lub `by`:

```
> tapply(iris$Petal.Length, iris$Species, stem)
> by(iris$Petal.Length, iris$Species, stem)
```

Niestety wynik uzyskany przy pomocy tych komend ma pewną wadę: nazwy grup (tutaj gatunków irysa) nie są prezentowane razem z diagramami, tylko ich kolejność jest drukowana pod ostatnim diagramem. Zamiast tego możemy posłużyć się poniższą funkcją:

```
stem.groups <- function(x, group, ...) {
   group <- factor(group)
   for (g in seq_along(levels(group))) {
     cat(levels(group)[g], "\n")
     stem(x[group == levels(group)[g]], ...)
     cat("------------------------\n")
   }
}
```

Zauważmy wykorzystanie argumentu `"..."`, który pozwala na przekazanie odpowiednich argumentów (tutaj `scale` i `width`) do funkcji `stem`:

```
> stem.groups(x = iris$Petal.Length, group = iris$Species,
scale = .3)
setosa

  The decimal point is at the |
  1 | 01223333333444444444444
  1 | 55555555555556666666777799
------------------------
versicolor
  The decimal point is at the |
  3 | 03355678999
  4 | 000001112222334444555555566677778899
  5 | 01
------------------------
virginica
```

```
  The decimal point is at the |
  4 | 588999
  5 | 000111111122334455566666677788899
  6 | 00111346779
------------------------
```

Niestety może się zdarzyć, że dla różnych grup różne będzie miejsce separatora dziesiętnego, o czym można się przekonać, wpisując następujący kod:

```
> stem.groups(iris$Petal.Length, iris$Species, scale = .5)
```

DYSKUSJA I WNIOSKI

Diagram łodyga i liście w pewnym sensie przypomina histogram, gdyż zagregowane wartości realizacji zmiennej przedstawione są w postaci poziomego ciągu znaków numerycznych odpowiadających poziomym kolumnom histogramu o specyficznie ustalonych klasach. Grupowanie obserwacji w klasy zwykle będzie jednak inne, co wynika z różnych metod grupowania danych (optymalna liczba klas histogramu zwykle jest określana przy pomocy jednej z wielu metod do tego służących, podczas gdy liczba gałęzi jest pośrednio określana przez użytkownika). Ponadto na histogramie nie ma możliwości zauważenia różnicy między dwoma wartościami, jeżeli tylko znajdą się w jednej klasie, podczas gdy na diagramie łodyga i liście różnica ta może być zauważalna (chyba że ustalimy `width = 0`).

Przedstawiono kilka uwag, które mogą być bardzo przydatne przy tworzeniu i czytaniu diagramu:

— Argument `scale` odpowiada za wysokość wykresu; dla rozpatrywanych danych należy dobrać najlepszą wysokość, która pozwoli na dobry odczyt wartości i zrozumienie ich rozkładu. Użytkownik nie ma jednak jawnego wpływu na określenie ilości klas podziału: nie może bezpośrednio określić liczby gałęzi.

— Odczytanie wartości zmiennych (zwykle zaokrąglonych) jest najprostsze przy pomocy komunikatu ponad wykresem, informującego o położeniu separatora dziesiętnego, oraz porównania wartości minimalnej i maksymalnej i ich prezentacji na diagramie.

— Argument `width = 0` pokazuje tylko liczbę liści na gałęziach, jest to więc najbardziej kompaktowa forma diagramu.

Z naszej praktyki wynika, że omawiana postać diagramu jest prosta w użyciu. Choć bez odpowiedniego wstępu mogą wystąpić problemy z jego zrozumieniem, to krótkie wprowadzenie i praktyka na kilku różnych zbiorach danych sprawia, że wykres ten okazuje się być łatwym do stworzenia i zrozumienia. Pomimo swej prostoty okazuje się być on bardzo użytecznym narzędziem eksploracji danych i dlatego chcemy zachęcić czytelników do jego używania, zwłaszcza na wstępnych etapach analizy – ale nie tylko, bo np. z naszych doświadczeń wynika, że analiza reszt przy pomocy tego diagramu jest efektywna.

Mogłoby się wydać, że w dobie powszechnej dostępności narzędzi graficznych oferowanych przez programy komputerowe tego typu prezentacja graficzno-tekstowa nie ma zastosowania. Nie jest to jednak prawdą: dzięki swojej prostocie i efektywności w przekazie informacji diagram łodyga i liście może okazać się bardzo przydatnym

narzędziem analizy rozkładu zmiennej, a dzięki programowi komputerowemu jego stworzenie jest bardzo łatwe i szybkie, a jednocześnie czytelne.

LITERATURA

Anderson, E. 1935. The irises of the Gaspe Peninsula. Bulletin of the American Iris Society, 59: 2 — 5.

Becker, R. A., Chambers, J. M., Wilks, A. R. 1988. The New S Language. Wadsworth & Brooks/Cole.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics, **7**, Part II: 179 — 188.

Pereira-Mendoza L., Dunkels A. 1989. Stem-and-leaf plots in the primary grades. Teaching Statistics 11 (2): 34 — 37.

Perry B., Jonem G. A., Thornton C. A., Langrall C. W., Putt I. J., Krat C. 1999. Exploring visual displays involving Beanie Baby data. Teaching Statistics 21 (1): 11 — 13.

R Development Core Team. 2010. 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.

Tukey J. 1977. Exploratory Data Analysis. Addison-Wesley.

Wickham H. 2009. ggplot2: elegant graphics for data analysis. Springer New York.