Dzień 1 - Wprowadzenie do R i wizualizacja - zadania

Patryk Czortek, Marcin K. Dyderski

22 kwietnia 2024

Zadania do wykonania

1. Załóżmy, że dysponujemy próbkami masy ciała, koloru sierści oraz struktury płci z populacji ryjówki aksamitnej. Dane reprezentujące próbę można wpisać w skrypt w następujący sposób:

```
shrews.mass<-c(26,29,41,24,28,56,74,35,68,95,45,
67,89,35,67,88,75,34)

fur.color<-c('gray','gray','gray','gray','brown','brown',
'brown','brown','brown','black','black','black',
'black','black','black','gray','brown','brown')

male<-c('TRUE','TRUE','TRUE','TRUE','FALSE','FALSE',
        'FALSE','FALSE','FALSE','TRUE','TRUE','TRUE','TRUE',
        'FALSE','FALSE','FALSE','FALSE','FALSE')
```

   a) Korzystając z funkcji `mean()` obliczyć średnią arytmetyczną z masy ciała ryjówek Przyjmując założenie, że dane masy ciała reprezentują rozkład normalny:
   b) Korzystając z funkcji `var()` obliczyć wariancję
   c) Korzystając z funkcji `sd()` obliczyć odchylenie standardowe
   d) Dane masy ciała ryjówek zapisać jako ciąg liczb (funkcja `as.integer()`), kolor sierści jako ciąg znaków (funkcja `as.character()`), a strukturę płci jako wartość logiczną (funkcja `as.logical()`)
   e) Z trzech wektorów stworzonych w podpunkcie (d) stworzyć listę (funkcja `list()`). Każdemu poziomowi listy nadać nazwę (funkcja `names()`)
   f) Listę przekształcić w ramkę danych (funkcja `as.data.frame()`)
   g) Korzystając z funkcji `summary()` wyświetlić podsumowanie ramki danych - co daje nam ta funkcja?

2. W plikach `molinion1.csv` oraz `molinion2.csv` zawarto dane procentowego pokrycia gatunków roślin naczyniowych na n powierzchniach badawczych zlokalizowanych na łąkach zmiennowilgotnych w okolicach Dąbrowy Górniczej. Linki: [https://github.com/mkdyderski/BSS/blob/BSS2024/datasety/molinion1.csv] oraz [https://github.com/mkdyderski/BSS/blob/BSS2024/datasety/molinion2.csv]. Możesz również ściągnąć go do R za pomocą funkcji `read.csv()`:

```
moli1<-read.csv('https://raw.githubusercontent.com/mkdyderski/BSS/BSS2024/datasety/molinion1.csv',
                sep=';')
moli2<-read.csv('https://raw.githubusercontent.com/mkdyderski/BSS/BSS2024/datasety/molinion2.csv',
                sep=';')
```

   a) Oba pliki wczytać do R

b) Sprawdzić, czy dane zostały wczytane poprawnie (funkcja `head()`, `str()`, `dim()` oraz `length()`). Na każdym etapie sprawdzaj obiekty w oknie środowiska (podgląd tabeli).
c) Dokonać transpozycji obu ramek danych (funkcja `t()`)
d) Korzystając z biblioteki `reshape2` obie ramki danych przekształcić do wąskich tabelek (funkcja `melt()`)
e) Obie wąskie tabelki skleić wierszami w jeden obiekt (funkcja `rbind()`)
f) Zmienić nazwy kolumn. Kolumna pierwsza reprezentuje id powierzchni, druga gatunek (zmień na 'species'), a trzecia procentowe pokrycie gatunku
g) Korzystając z biblioteki `reshape2` tabelkę wąską przekształcić w szeroką (funkcja `dcast()`). W kolumnach powinny być nazwy powierzchni, a w wierszach nazwy gatunków. Dlaczego dla niektórych wierszy (gatunków) wyprodukowane zostały wartości `NA`?
h) Szeroką tabelkę wyeksportować z R do pliku .csv (funkcja `write.table()`) i otworzyć w Excelu. Wartości NA przekształcić w zera (wtedy powiemy R, że te dane nie są brakujące, ale że w danej próbie dany gatunek osiągnął zerowe pokrycie, czyli w jednej próbie go mogło nie być, a np. w 25 innych próbach notowany był z dużym pokryciem). Pierwszy wiersz przesunąć o jedną kolumnę w prawo. Kolumna A pozostanie wtedy pusta. Należy ją usunąć.
i) Szeroką tabelkę załadować ponownie do R
j) Przekształcić tabelkę szeroką w wąską
k) Po załadowaniu do R pliku `cechy.all.csv`, link: [https://raw.githubusercontent.com/mkdyderski/BSS/BSS2024/datasety/cechy.all.csv']. Możesz również ściągnąć go do R za pomocą funkcji `read.csv()`:

```
cechy.all<-read.csv('https://raw.githubusercontent.com/mkdyderski/BSS/BSS2024/datasety/cechy.all.csv',
                    sep=';')
```

l) Stwórz obiekt `nasiona.liscie`, zawierającą gatunek (kolumna `species`), wysokość pędów (`canopy_height`) masę liści (`leaf_mass`), rozmiar liści (`leaf_size`) i masę nasion (`seed_mass`). Wykorzystaj mechanizm indeksowania - podaj numery odpowiednich kolumn w nawiasie kwadratowych. Numery kolumn sprawdź za pomocą funkcji `colnames(cechy.all)`

```
nowa.tabela<-cechy.all[,c(1,3,23,45)]#numery trzeba sprawdzić
```

m) Złącz tabelę wąską z nazwami gatunków z tabelą `nasiona.liscie` W tym celu należy wykorzystać funkcję `left_join()` z pakietu `dplyr`.

3. Z GitHuba pobierz dataset z cechami roślin, link: [https://raw.githubusercontent.com/mkdyderski/BSS/BSS2024/datasety/vege_1517_traits.csv]. Możesz również ściągnąć go do R za pomocą funkcji `read.csv()`:

```
baza<-read.csv('https://raw.githubusercontent.com/mkdyderski/BSS/BSS2024/datasety/vege_1517_traits.csv'
               sep=';')
```

a. Wykonaj histogram dla `seed_mass`.
b. Wykonaj wykres rozrzutu dla `canopy_height` i `SLA`, `SLA` i `seed_mass` oraz `seed_mass` i `canopy_height`. Dodaj linie trendu używając funkcji `geom_smooth(method='lm')` - sprawdź jak zmieni się wynik po dodaniu skal logarytmicznych.
c. Wykonaj boxploty dla `seed_mass` w ramach `strategy`, użyj skali log10, dodaj skalę barwną z palety ColorBrewer używając `scale_fill_brewer()`
d. Podaj srednie wartosci kilku wybranych cech dla grup hg (historyczno-geograficznych) i wykonaj wykres słupkowy, pokazujący srednie i SE (kod w prezentacji))
e. Narysuj wykres na którym pokażesz zależnośc pomiędzy `SLA` i `canopy_height` a wielkość punktów (`aes(...size=...)`) zależeć będzie od `seed_mass`. Dopasuj skale i linię trendu.

Zadbaj o estetykę wszystkich wykresów - zmień tło, opisy osi i elementy graficzne. Skorzystaj z linków do dodatkowych materiałów.

##Propozycje do pracy z własnym zbiorem danych##

5. Wczytaj *własny zbiór danych* i sprawdź poprawność wczytanych danych
6. Spróbuj obliczyć podstawowe statystyki na zmiennych które badasz. Wypróbuj użycie niektórych funkcji.
7. Skonsultuj z prowadzącymi koniecznosc przekształceń danych - czy będzie trzeba zmienić format danych?
8. Jeśli część danych wymaga złączenia to teraz jest na to najlepszy czas. Nawet jeśli dane mają się zmienić, warto przygotować sobie kod który pozwoli łączyć tabele wg określonego klucza już teraz.
9. Policz średnie wartości i miary dyspersji (SD lub SE, szczegóły poniżej) dla grup w ramach swojego zbioru danych. Mając średnie wartości i miary dyspersji możesz zastanawiać się nad różnicami pomiędzy grupami.

wariancja, SD i SE: wariancja to SD^2, SD jest tutaj estymatorem (przybliżeniem) dyspersji danych SE to SD/sqrt(), czyli jest to SD dzielone przez pierwiastek kwadratowy z liczby prób. W bazowym R nie ma funkcji na SE, więc można ją napisac samemu i wrzucić do konsoli. Funkcje w R to również obiekty - jeśli jakiegoś nie ma, możesz stworzyć go samemu:

```
se<-function(x)sd(x,na.rm=T)/sqrt(length(x[!is.na(x)]))
```

10. Poznaj strukturę i zakresy zmiennych w Twoim zbiorze danych. Czy są obserwacje odstające? Eksploracja danych pozwala wykryć wartości nielogiczne z biologicznego punktu widzenia biologicznego i naprawić je przed właściwymi analizami.
11. Wykonaj wykres punktowy pokazujący relacje pomiędzy cechami dla których zakładasz występowanie pewnych zależności, dodaj linię trendu i oceń czy jest w miarę sensownie dopasowana. W przypadku wątpliwosci poproś prowadzących o podpowiedź odnośnie typu linii trendu. Możesz zestawić np. sukces reprodukcyjny z cechami środowiska czy występowanie gatunków (0-1) z dostępnością zasobów. Dla zmiennych binarnych zamiast `+geom_smooth(method='lm')` daj `+geom_smooth(method='glm', method.args=c(family="binomial"))`.
12. Sprawdź czy zmienne liczbowe różnią sie pomiędzy grupami (bez testów, na razie tylko wizualnie). Możesz np. sprawdzić bogactwo gatunkowe w różnych wariantach, wartości odbicia widma dla różnych gatunków, masę czy wielkość prób dla różnych terminów lub indeks heterotermii dla różnych osobników.
13. Sprawdź, czy proste przeliczenie pozwoli dać Ci kolejną cechę do analiz. Być może wystarczy podzielić masę przez objętość by zyskać gęstość? Albo określić udział gildii/grup funkcjonalnych organizmów? Spróbuj wykorzystać do tego pakiet dplyr.
14. Przygotuj zestawienie liczebności prób w różnych wariantach za pomocą wykresów słupkowych - takie aby móc łatwo opowiedzieć o układzie badań.