Dzień 2 - Rozkłady, testy statystyczne i regresja - zadania

Patryk Czortek, Marcin K. Dyderski

23 kwietnia 2024

Zadania do wykonania

1. Dane zawarte w pliku `lichenes1.csv` reprezentują bogactwo (kolumna `Rich`) i różnorodność gatunkową (kolumna `Shan`) oraz proporcję gatunków porostów epifitycznych o różnych wymaganiach względem zasobności podłoża w azot (kolumna `EIV_N`) w Puszczy Białowieskiej na 144 powierzchniach historycznych z 1992 roku (kolumna `time=='h'`) oraz na 144 powierzchniach powtórnie przebadanych w roku 2014 (kolumna `time=='n'`) wraz z danymi odnośnie typu zbiorowiska leśnego dla każdej powierzchni (kolumna `habitat`). link: [https://github.com/mkdyderski/BSS/blob/BSS2024/datasety/lichenes1.csv]. Możesz również ściągnąć go do R za pomocą funkcji `read.csv()`:

```
lichenes<-read.csv('https://raw.githubusercontent.com/mkdyderski/
                   BSS/BSS2024/datasety/lichenes1.csv', sep=';')
```

   a) Korzystając z funkcji `hist()` lub `ggplot2::geom_histogram()` ocenić, czy bogactwo i różnorodność gatunkowa prób historycznych i powtórnie przebadanych reprezentują rozkład normalny
   b) Zakładając, że dane reprezentują rozkład normalny, zaproponować rodzaj testu statystycznego, odpowiedniego do zbadania różnic w różnorodności gatunkowej pomiędzy dwoma typami zbiorowisk leśnych. W którym zbiorowisku różnorodność gatunkowa była większa? Czy różnice były istotne statystycznie? A biologicznie?
   c) Zakładając, że dane reprezentują rozkład normalny, zaproponować rodzaj testu statystycznego, odpowiedniego do zbadania różnic w średnich wartościach wskaźnika zasobności podłoża w azot (`EIV_N`) pomiędzy danymi historycznymi i powtórnie przebadanymi. Kiedy średni udział porostów o wyższych wymaganiach względem azotu był większy – w 1992 roku, czy w roku 2014? Czy różnice były istotne statystycznie? Ocenić, czy różnice w czasie były duże, czy niewielkie.

2. Po ponad 90 latach od pierwszych obserwacji florystycznych badano zmiany w bogactwie gatunkowym wyleżysk (plik `wylezyska.csv`; kolumna `rich`). Zakładając, że zarówno dane historyczne (kolumna `time=='k'`), jak i powtórnie przebadane (kolumna `time=='n'`) nie reprezentują rozkładu normalnego, oraz że dane w 2015 roku były pobrane dokładnie z tych samych lokalizacji, co w 1927 roku, zaproponować rodzaj testu statystycznego, odpowiedniego do zbadania różnic w bogactwie gatunkowym pomiędzy dwoma okresami badawczymi. Czy różnice w bogactwie gatunkowym były istotne statystycznie? link: [https://github.com/mkdyderski/BSS/blob/BSS2024/datasety/wylezyska.csv]. Możesz również ściągnąć go do R za pomocą funkcji `read.csv()`:

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

3. W pliku `freq.epiphytes.csv` zawarto zmiany we frekwencji 10 gatunków porostów epifitycznych po 30 latach od pierwszych badań. Ile gatunków istotnie zwiększyło/zmniejszyło częstość występowania w porównaniu do stanu sprzed 30 lat? link: [https://github.com/mkdyderski/BSS/blob/BSS2024/datasety/freq.epiphytes.csv]. Możesz również ściągnąć go do R za pomocą funkcji `read.csv()`:

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

4. 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 analizę wariancji i test Tukeya dla zależności wskaźnika światła (kolumna `L`) od grup historyczno-geograficznych (`hg`)
b. oblicz z niej średnie brzegowe za pomocą funkcji `emmeans()`, wynik pokaż na wykresie kolumnowym z słupkami błędów
c. wykonaj model liniowy dla zależności wskaźnika światła (kolumna `L`) od grup historyczno-geograficznych (`hg`) oraz SLA, sprawdź `summary(model)`
d. wykonaj wizualizację modelu za pomocą `ggpredict()`
e. dla chętnych: co by było gdyby usunąć obserwację odstającą (duże SLA)? Czy wychodzi interakcja między zmiennymi?

##Propozycje do pracy z własnym zbiorem danych

5. Wczytaj *własny zbiór danych* i sprawdź rozkłady zmiennych - zastanów się jakie to będzie miało znaczenie dla modelowania
6. Sprawdź czy badane cechy różnią sie pomiędzy grupami za pomocą testów t-Studenta/chi-kwadrat lub analizy wariancji. Jeśli wykonujesz analizę wariancji, pamiętaj o testach post-hoc (Tukeya).
7. Przygotuj wykres i tabelę z analizą wariancji dla wybranej zmiennej. Wzoruj się na publikacjach ze swojej działki lub zapytaj co musi się tam znaleźć.