Dzień 2 - Testy statystyczne i rozkłady - zadania

Patryk Czortek, Marcin K. Dyderski

11 stycznia 2022

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://raw.githubusercontent.com/mkdyderski/BSS/BSS2019/datasety/lichenes1.csv]. Możesz również ściągnąć go do R za pomocą funkcji `read.csv()`:

```
lichenes<-read.csv('https://raw.githubusercontent.com/mkdyderski/BSS
                   /BSS2019/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 nie reprezentują rozkładu normalnego, 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?

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://raw.githubusercontent.com/mkdyderski/BSS/BSS2019/datasety/wylezyska.csv]. Możesz również ściągnąć go do R za pomocą funkcji `read.csv()`:

```
wylezyska<-read.csv('https://raw.githubusercontent.com/mkdyderski/BSS
                    /BSS2019/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://raw.githubusercontent.com/mkdyderski/BSS/BSS2019/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/BSS2019/datasety/freq.epiphytes.csv',sep=';')
```

4. Wczytaj plik 'prunus.csv' dostepny na githubie, link: [https://raw.githubusercontent.com/mkdyderski/BSS/BSS2019/datasety/prunus.csv]. Możesz również ściągnąć go do R za pomocą funkcji `read.csv()`:

```
prunus<-read.csv('https://raw.githubusercontent.com/mkdyderski/BSS/
                 BSS2019/datasety/prunus.csv', sep=';')
```

W tabeli kolumna typ opisuje typ roślinności (Car-Aln to ols, Fra-Aln to łęg olszowo-jesionowy, transit to zbiorowisko przejściowe - miedzy olsem a łęgiem, LZZ - skrajnie zdegenerowany, przesuszony i brzydki łeg), a kolumna L - wskaźnik świetlny Ellenberga. Za pomocą jednoczynnikowej analizy wariancji sprawdź czy są różnice w L pomiędzy typami roślinności. Jeśli są, za pomocą testu Tukeya sprawdź pomiędzy którymi.

Propozycje do pracy z własnym zbiorem danych

5. Obejrzyj *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.