Drugie zadanie z kolokwium RPiS - sprawozdanie

Kacper Bajkiewicz

28 kwietnia 2020

1 Wprowadzenie i treść zadania

Tematem tego ćwiczenia jest zbadać współczynnik korelacji miedzy liczbami zakażeń wirusem Sars-CoV-2 wśród kilku krajów wymienionych w danych zadania. Wybrałem opisanie jej dla 5-ciu europejskich krajów. Funkcje przygotowujaca plik CSV do pracy wstawie pod koniec dokumentu, ponadto potrzeba dwóch struktur danych. Dane o całkowitej liczbie zachorowań i nazwy do odszyfrowania krajów beda przechowywane tak:

```
suma_zachorowan = [199414, 156437, 157153, 127008, 112261]
panstwa = {1 : "Włochy", 2: "Niemcy", 3: "Wlk Brytania", 4: "Francja", 5:"Turcja"}
```

2 Co to jest współczynnik korelacji?

Rozpatrzmy dwie dowolne zmienne losowe a i b. W życiu człowieka zdarzaja sie takie sytuacje, kiedy chce zbadać ich wzajemna zależność, i właśnie z odpowiedzia na to pytanie przychodzi współczynnik korelacji. W ogólnym przypadku określa sie go wzorem

$$r_{a,b} = \frac{Cov(a,b)}{\sigma_a \times \sigma_b} \tag{1}$$

Gdzie σ_a to odchylenie standardowe zmiennej a (czyli pierwiastek z wariancji). Kiedy jednak rozpatrujemy dane dotyczace populacji, możemy te wzory rozwinać (niejako *pozbyć* sie prawdopodobieństwa ze wzorów). Zauważmy kilka faktów...

$$EW = \overline{W} \tag{2}$$

Do tego:

$$Cov(x,y) = \sum_{k=i}^{n}(x_i - \overline{x})(y_i - \overline{y}) \tag{3}$$

Ponadto:

$$\sigma_w = \sqrt{\sum_{i=1}^{n}(w_i - \overline{w})^2} \tag{4}$$

Wtedy równanie (1) sprowadza nam sie do postaci:

$$r_{a,b} = \frac{\sum_{k=i}^{n}(a_k - \overline{a})(b_k - \overline{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \overline{a})^2 \times \sum_{i=1}^{n}(b_i - \overline{b})^2}} \tag{5}$$

Co możemy wyliczyć już bezpośrednio, za pomoca nastepujacych funkcji:

```
def srednia(tab): #inputem jest tutaj tablica z liczbą zachorowań dla danego kraju
    return sum(tab[i] for i in range(len(tab))) / len(tab)

def kowariancja(x,y,xsr,ysr): #w zmiennych x i y liczby zachorowań są jako tablice
    suma = 0
    for i in range(len(x)):
        suma += (x[i] - xsr)*(y[i]-ysr)
    return suma

def odchylenie(x,xsr):
    suma = 0
    for el in x:
        suma += (el-xsr)**2
    return sqrt(suma)

def wsp_korelacji(x,y):
    xsr = srednia(x)
    ysr = srednia(y)
    return kowariancja(x,y,xsr,ysr) / (odchylenie(x,xsr) * odchylenie(y,ysr))
```

3 Rozważania o doborze daty startowej

3.1 Jak wybrać punkt poczatkowy?

Wiadomo, że maksima funkcji liczby zakażeń od czasu wystepuja w różnych dniach dla różnych państw (obsuniecie może wynieść nawet ponad miesiac!). Chcac wiec zobaczyć czy istnieje pewne powiazanie w procesie rozprzestrzeniania sie wirusa, musimy osobno dla każdej pary państw wybrać daty, od których zaczniemy szukać korelacji. Analizujac dane doszedłem do wniosku, że, zamiast patrzeć na dane bezwzgledne (np. 50 zakażeń na poczatku epidemii dla Malty jest drastycznie odmienne dla 50 np Włoch), lepiej jest rozpocząć liczyć zakażenia, kiedy suma przypadków w danym kraju przekroczy pewien nieduży procent całkowitej liczby zakazen w badanym kraju (na końcu sprawozdania policze to dla kilku innych wartości, żeby zobaczyć, dla którego odsetka korelacja jest najwieksza).

3.1.1 Algorytm na punkt startowy i tablice z pożadanymi danymi

Metoda znajdowania dat do analizy jest nastepujaca - dla każdych dwóch państw najpierw sprawdzam, kiedy suma zachorowań bedzie wieksza ustalony procent zachorowań do 27 kwietnia. Zdecydowałem sie sprawdzić dwie wersje, jedna polegajaca na sprawdzaniu "tyle, ile tylko sie da"(czyli maksymalna możliwa liczbe dni po osiagnieciu procenta) i druga, która opiera sie na sprawdzaniu przez określona liczbe dni.

```
def znajdz_dzien_pocz(nr):
    suma, dzien, iterator = 0, 0, 0
    while(suma < suma_zachorowan[nr-1] * procent):
        suma += tablica[iterator][nr]
        iterator += 1; dzien += 1
        if dzien + ldni >= max_dzien: dzien = max_dzien - ldni
    return dzien - 1
```

4 Wyznaczanie współczynników korelacji

Tak wyglada przygotowanie liczb zakażen pod obliczenia:

```
def zgromadz_dane(nr, ile, dzien): #ile- tyle dni bierzemy
    zwrot = []
    for i in range(dzien, dzien+ile):    #dodajemy wszystkie dni po osiagnieciu
        zwrot.append(tablica[i-1][nr]) #procenta liczby zachorowań
    return zwrot
```

Pozostaje wiec znaleźć współczynnik korelacji dla dwóch krajów, przy użyciu funkcji wyżej wymienionych (sa one zdefiniowane w *znajdzwspolkorelacji*, dzieki czemu maja dostep do argumentów wywołania i pól nadfunkcji)

```
def znajdz_wspol_korelacji(panstwo1, panstwo2, procent, ldni):
    tablica = wczytaj()
    suma_zachorowan = [199414, 156437, 157153, 127008, 112261]
    panstwa = {1 : "Włochy", 2: "Niemcy", 3: "Wlk Brytania", 4: "Francja", 5:"Turcja"}
    max_dzien = len(tablica) - 1
```

//Tu funkcje które opisałem wcześniej//

```
poczl = znajdz_dzien_pocz(panstwo1)
pocz2 = znajdz_dzien_pocz(panstwo2)
if ldni == True:
    ldni = min(max_dzien - poczl, max_dzien-pocz2)
x = zgromadz_dane(panstwo1, ldni, poczl)
y = zgromadz_dane(panstwo2, ldni, pocz2)
korel = round(wsp_korelacji(x,y), 3)
print(f"Wspolczynnik korelacji dla panstw {panstwa[panstwo1]} i {panstwa[panstwo2]} to : {korel}; ", end='')
print(f"daty startowe: {panstwa[panstwo1]} - {tablica[poczl][0]}, {panstwa[panstwo2]} - {tablica[pocz2][0]}", end='')
print(f"procent startowy: {procent}")
return korel
```

Dzieki temu, że wypisuja sie całe potrzebne dane, od razu opisane sa przeprowadzone testy. Funkcje można wykorzystać do znajdowania najlepszego współczynnika, które wyglada tak:

```
def znajdz_najlepszy_wspolczynnik(panstwo1, panstwo2):
    maks, minn = 0, 0
    print()
    for k in [0.03, 0.07]:
        wsp = znajdz_wspol_korelacji(panstwo1, panstwo2, k, 25)
        maks = max(maks, wsp); minn = min(maks, wsp)
        wsp = znajdz_wspol_korelacji(panstwo1,panstwo2,k, True)
        maks = max(maks, wsp); minn = min(minn, wsp)
    print(f"Dla tych państw minimalny współczynnik to {minn}")
    print()
    return maks
```

Sprawdzane sa 4 możliwości, zależnie od liczby sprawdzanych dni (kiedy inputujemy True, to funkcja sprawdza tyle dni, ile krajowi, który później osiagnał procent zachorowań upływa do 28 kwietnia) i od startowego procenta zachorowań. Widać, że taki rozstrzał pozwala, niespodziewanie, osiagnać czasem bardzo różne współczynniki! Świadczy to raczej o dobrym doborze danych do testów (testowałem jeszcze dla 35, 40, 50 dni ale te wyniki sa uderzajaco podobne do tych z 25 dniami).

4.1 Porównania i współczynniki dla par państw

4.1.1 Włochy i Niemcy

```
Wspolczynnik korelacji dla ITA i GER to: 0.756; start: ITA-mar 9 , GER-mar 16 %: 0.03
Wspolczynnik korelacji dla ITA i GER to: 0.708; start: ITA-mar 9 , GER-mar 16 %: 0.03
Wspolczynnik korelacji dla ITA i GER to: 0.505; start: ITA-mar 13 , GER-mar 20 %: 0.07
Wspolczynnik korelacji dla ITA i GER to: 0.603; start: ITA-mar 13 , GER-mar 20 %: 0.07
Dla tych państw minimalny współczynnik to 0.505
```

4.1.2 Włochy i Wielka Brytania

```
Wspolczynnik korelacji dla ITA i GBR to: 0.792; start: ITA-mar 9 , GBR-mar 22 %: 0.03
Wspolczynnik korelacji dla ITA i GBR to: 0.72; start: ITA-mar 9 , GBR-mar 22 %: 0.03
Wspolczynnik korelacji dla ITA i GBR to: 0.678; start: ITA-mar 13 , GBR-mar 27 %: 0.07
Wspolczynnik korelacji dla ITA i GBR to: 0.594; start: ITA-mar 13 , GBR-mar 27 %: 0.07
Dla tych państw minimalny współczynnik to 0.594
```

4.1.3 Niemcy i Turcja

```
Wspolczynnik korelacji dla GER i TUR to: 0.711; start: GER-mar 16 , TUR-mar 27 %: 0.03
Wspolczynnik korelacji dla GER i TUR to: 0.695; start: GER-mar 16 , TUR-mar 27 %: 0.03
Wspolczynnik korelacji dla GER i TUR to: 0.277; start: GER-mar 20 , TUR-mar 30 %: 0.07
Wspolczynnik korelacji dla GER i TUR to: 0.327; start: GER-mar 20 , TUR-mar 30 %: 0.07
Dla tych państw minimalny współczynnik to 0.277
```

4.1.4 Wielka Brytania i Turcja

```
Wspolczynnik korelacji dla GBR i TUR to: 0.843; start: GBR-mar 22 , TUR-mar 27 %: 0.03
Wspolczynnik korelacji dla GBR i TUR to: 0.763; start: GBR-mar 22 , TUR-mar 27 %: 0.03
Wspolczynnik korelacji dla GBR i TUR to: 0.701; start: GBR-mar 27 , TUR-mar 30 %: 0.07
Wspolczynnik korelacji dla GBR i TUR to: 0.647; start: GBR-mar 27 , TUR-mar 30 %: 0.07
Dla tych państw minimalny współczynnik to 0.647
```

4.1.5 Włochy i Francja

```
Wspolczynnik korelacji dla ITA i FRA to: 0.747; start: ITA-mar 9 , FRA-mar 15 %: 0.03
Wspolczynnik korelacji dla ITA i FRA to: 0.668; start: ITA-mar 9 , FRA-mar 15 %: 0.03
Wspolczynnik korelacji dla ITA i FRA to: 0.644; start: ITA-mar 13 , FRA-mar 19 %: 0.07
Wspolczynnik korelacji dla ITA i FRA to: 0.601; start: ITA-mar 13 , FRA-mar 19 %: 0.07
Dla tych państw minimalny współczynnik to 0.601
```

Na pierwszy rzut oka widać bardzo sporo podobieństw, można też zauważyć, że odkad tempo rozwoju pandemii przekroczy pewien krytyczny punkt, to nastepnie rośnie dość przewidywalnie.

```
Wspolczynnik korelacji dla GER i GBR to: 0.56; start: GER-mar 16 , GBR-mar 22 %: 0.03
Wspolczynnik korelacji dla GER i GBR to: 0.342; start: GER-mar 16 , GBR-mar 22 %: 0.03
Wspolczynnik korelacji dla GER i GBR to: 0.435; start: GER-mar 20 , GBR-mar 27 %: 0.07
Wspolczynnik korelacji dla GER i GBR to: 0.272; start: GER-mar 20 , GBR-mar 27 %: 0.07
Dla tych państw minimalny współczynnik to 0.272
```

4.1.6 Niemcy i Wielka Brytania

Różnica wynika z tego, że, o ile w Niemczech dzienne zachorowania znacznie spadły, to w Wielkiej Brytanii ciagle utrzymuja sie na podobnym poziomie. Kiedy liczymy tylko przez 20 dni widać już wysoka korelacje.

4.1.7 Turcja i Francja

```
Wspolczynnik korelacji dla GER i GBR to: 0.56; start: GER-mar 16 , GBR-mar 22 %: 0.03
Wspolczynnik korelacji dla GER i GBR to: 0.342; start: GER-mar 16 , GBR-mar 22 %: 0.03
Wspolczynnik korelacji dla GER i GBR to: 0.435; start: GER-mar 20 , GBR-mar 27 %: 0.07
Wspolczynnik korelacji dla GER i GBR to: 0.272; start: GER-mar 20 , GBR-mar 27 %: 0.07
Dla tych państw minimalny współczynnik to 0.272
```

4.1.8 Niemcy i Francja

```
Wspolczynnik korelacji dla GER i FRA to: 0.608; start: GER-mar 16 , FRA-mar 15 %: 0.03
Wspolczynnik korelacji dla GER i FRA to: 0.577; start: GER-mar 16 , FRA-mar 15 %: 0.03
Wspolczynnik korelacji dla GER i FRA to: 0.367; start: GER-mar 20 , FRA-mar 19 %: 0.07
Wspolczynnik korelacji dla GER i FRA to: 0.486; start: GER-mar 20 , FRA-mar 19 %: 0.07
Dla tych państw minimalny współczynnik to 0.367
```

4.1.9 Włochy i Turcja

```
Wspolczynnik korelacji dla ITA i TUR to: 0.857; start: ITA-mar 9 , TUR-mar 27 %: 0.03
Wspolczynnik korelacji dla ITA i TUR to: 0.836; start: ITA-mar 9 , TUR-mar 27 %: 0.03
Wspolczynnik korelacji dla ITA i TUR to: 0.711; start: ITA-mar 13 , TUR-mar 30 %: 0.07
Wspolczynnik korelacji dla ITA i TUR to: 0.713; start: ITA-mar 13 , TUR-mar 30 %: 0.07
Dla tych państw minimalny współczynnik to 0.711
```

4.1.10 Wielka Brytania i Francja

```
Wspolczynnik korelacji dla GBR i FRA to: 0.595; start: GBR-mar 22 , FRA-mar 15 %: 0.03
Wspolczynnik korelacji dla GBR i FRA to: 0.469; start: GBR-mar 22 , FRA-mar 15 %: 0.03
Wspolczynnik korelacji dla GBR i FRA to: 0.699; start: GBR-mar 27 , FRA-mar 19 %: 0.07
Wspolczynnik korelacji dla GBR i FRA to: 0.544; start: GBR-mar 27 , FRA-mar 19 %: 0.07
Dla tych państw minimalny współczynnik to 0.544
```

5 Wnioski i zakończenie

Jak widać w poniższej tabelce z maksymalnymi współczynnikami:

	Włochy	Niemcy	Wlk Bryt	Francja	Turcja
Włochy	1.000				
Niemcy	0.756	1.000			
Wlk Bryt	0.792	0.435	1.000		
Francja	0.747	0.608	0.699	1.000	
Turcja	0.857	0.711	0.843	0.77	1.000

Zachodzi wysoka korelacja dla poszczególnych państw. Najlepsze wyniki daje sprawdzanie jak tylko zacznie rozwijać sie epidemia i sprawdzać przez określona liczbe dni. Pokazuje to, że wirus rozprzestrzenia sie mniej wiecej równomiernie i różnice powstaja dopiero w momencie szczytu zachorowań, po nim jakiekolwiek kontrasty wynikaja najpiewniej z różnej kondycji służby zdrowia i wprowadzanych obostrzeń (np. widać, jak świetnie radzi sobie z wirusem RFN)

Dodam jeszcze tylko, że słów *zakażenie* i *zarażenie* używam w takim samym kontekście, ale warto rozróżnić ich dokładne znaczenie. Kiedy pisałem o zakażeniach, miałem na myśli zakażenia wirusem Sars-CoV-2 a kiedy o zachorowaniach - zachorowania na COVID-19.

Kacper Bajkiewicz
Nr indeksu 314438

6 PS

Wstawiam obiecane zdjecie funkcji wczytujacej.

```
def wczytaj():
    zwrot = []
    for linia in open("dane0428.csv", encoding="utf-8"):
        linia = linia.strip("\n") #ususwamy znak nowej linii, zeby nie przeszkadzal
        linia = linia.split(';') #w formacie csv znak ; separuje dane, tutaj sa one naturalnie odseparowane przez tablice
        zwrot.append(linia) #tablica jeszcze w fazie obrobki
    zwrot = (zwrot[4:]) #usuwam puste wiersze z konca i dane o nazwie krajow czy sumie przypadkow
    pom = []
    for dzien in zwrot:
        pom_dzien = [dzien[0]] #data pozwoli nam mniej więcej się orientować, kiedy zaczynamy liczyć
        for i in range(3, 12, 2): #co da nam iterację co 2gi element, śmierci pomijamy
            dzien[i] = dzien[i].replace(' ', '')
            if dzien[i] == '':
                pom_dzien.append(0)
            else:
                pom_dzien.append(int(dzien[i])) #kazdego stringa zamieniam na inta
        pom.append(pom_dzien)
    return pom
```

Jeśli jakkolwiek to Pana zainteresuje, to wrzuciłem jeszcze pierwotna wersje (zrobiona we wtorek) z macierza korelacji dla wszystkich 61 państw, ułożona na podstawie kilku możliwości danych startowych. Plik nazwany *dod.pdf* nie ma żadnego powiazania z rozwiazaniem prezentowanym tutaj.