Kolokwium zadanie 2.

Szymon Kosakowski 309980
Wiktor Hamberger 308982

4 maja 2020

Wstęp

Koronawirus COVID-19 i jego pokłosie w postaci zapaści gospodarki, może być największym wyzwaniem, z jakim będzie musiało poradzić nasze pokolenie. Na chwilę pisania tej pracy zarażonych tym wirusem jest 3,45 mln ludzi na całym świecie, natomiast ofiar śmiertelnych jest 244 tysięcy. W celu minimalizowania liczby zachorowań państwa na całym świecie wprowadzają wszelkiego rodzaju obostrzenia – w jednych krajach surowsze, natomiast w innych łagodniejsze. Pytanie brzmi, czy pomimo różnic w podjętych krokach, można jakoś skorelować przyrost liczby chorych w różnych krajach. Spróbujemy to zrobić dla Rosji, Szwajcarii, Izraela, Ukrainy i Czech.

Narzędzia

Jako materiału źródłowego użyjemy danych WHO, opisujących liczbę zachorowań w danych krajach z podziałem na dni. Jako, że w każdym kraju koronawirus zaczął się rozprzestrzeniać w innym czasie, będziemy szukać równolicznych przedziałów czasowych, przesuniętych pomiędzy sobą w czasie, ale takich, żeby zmaksymalizować korelację pomiędzy nimi. Dla rozpatrywanych przedziałów będziemy rozważać liczbę zachorowań każdego dnia parami tj. jeżeli rozpatrujemy kraj X i kraj Y na przedziałach n-dniowych, przy czym w kraju X zaczynamy pierwszego kwietnia, a w kraju Y dziesiątego kwietnia, to odpowiednio x_1 będzie oznaczało liczbę zachorowań w państwie X pierwszego kwietnia, a y_1 liczbę zachorowań w państwie Y dziesiątego kwietnia. Mamy więc obserwacje:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

Zakładamy, że to są realizacje dwuwymiarowej zmiennej, której pierwszą współżędną jest zmienna X, a drugą zmienna Y. O tych obserwacjach zakładamy że są one niezależne, a ich prawdopodobieństwo to:

$$P((x_i, y_k)) = \begin{cases} 0 & i \neq k, \\ \frac{1}{n} & i = k \end{cases}$$

Współczynnik korelacji wygląda następująco:

$$\rho_{X,Y} = \frac{CoV(X,Y)}{\sqrt{V(X), V(Y)}}$$

Możemy zatem wyznaczyć dwie zmienne brzegowe: X, Z, których tabelki wyglądają następująco:

X/Y	x_1/y_1	...	x_n/y_n
P_i	$\frac{1}{n}$	...	$\frac{1}{n}$

Dzięki temu możemy policzyć:

$$V(X) = E(X^2) - (E(X))^2 = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - (\frac{1}{n}\sum_{k=1}^{n} x_k)^2$$

$$V(Y) = E(Y^2) - (E(Y))^2 = \frac{1}{n}\sum_{k=1}^{n} y_k^2 - (\frac{1}{n}\sum_{k=1}^{n} y_k)^2$$

$$CoV(X,Y) = E(X \cdot Y) - E(X) \cdot E(Y) =$$
$$= \frac{1}{n}\sum_{k=1}^{n} x_k y_k - (\frac{1}{n}\sum_{k=1}^{n} x_k) \cdot (\frac{1}{n}\sum_{k=1}^{n} y_k)$$

Co daje nam wszystkie potrzebne wzory do policzenia współczynnika korelacji.

Program

Program to implementacja powyższych wzorów jako funkcji w języku Python3. Co do szczegółów technicznych, najpierw usunęliśmy z arkusza wszystkie niepotrzebne dane tj. zostawiliśmy tylko liczbę zachorowań z podziałem na dni w każdym z pięciu wybranych krajów (Rosja, Szwajcaria, Izrael, Ukraina i Czechy), a następnie wczytaliśmy dane do programu za pomocą biblioteki pandas. Później tylko wykonywaliśmy poniższe funkcje na różnych datach i okresach w danych krajach.

```
def Cov(Y, dolY, Z, dolZ):
  suma = 0.0
  for i in range(okres):
    suma += dane[colNames[Y]][dolY + i]*dane[colNames[Z]][dolZ + i]
  suma*=1.0/okres

  sumY = 0.0
  for i in range(okres):
    sumY += dane[colNames[Y]][dolY + i]
  sumY *= 1.0/okres

  sumZ = 0.0
  for i in range(okres):
    sumZ += dane[colNames[Z]][dolZ + i]
  sumZ *= 1.0/okres

  return suma - sumY*sumZ

def V(Y, dol):
  suma = 0.0
  for i in range(okres):
    suma += dane[colNames[Y]][dol + i]*dane[colNames[Y]][dol+ i]
  suma*=1.0/okres

  suma2 = 0.0
  for i in range(okres):
    suma2 += dane[colNames[Y]][dol + i]
  suma2*=1.0/okres

  suma2 = suma2*suma2

  return suma - suma2

def P(Y, poczY, Z, poczZ):
  return Cov(Y, poczY, Z, poczZ)/math.sqrt(V(Y, poczY)*V(Z, poczZ))
```

Wyniki

Na podstawie wstępnej obserwacji danych stwierdzono, że początek epidemii nastąpił w wybranych państwach w następujących dniach:

państwo	początek pandemii
RUS	03.04
SUI	02.28
ISR	02.28
UKR	03.17
CZE	03.01

Tablica 1: data początku pandemii w danym państwie

Poniższe tabelki opisują współczynniki korelacji między ilością zakażeń koronawirusem w 5-ciu państwach. Sekcja jest podzielona na dwie części, względem długości okresu branego do badań. W każdej znajdują się dwie tabelki, przedstawiające wyniki korelacji danych na początku epidemii w danych krajach, a także znalezione okresy z maksymalną korelacją pomiędzy państwami.

Dla okresów 15-sto dniowych

okres	państwo 1	okres	państwo 2	współczynnik korelacji
3.4 - 3.19	RUS	2.28 - 3.14	SUI	0.**814**5540265934349
3.4 - 3.19	RUS	2.28 - 3.14	ISR	0.**758**8789011924836
3.4 - 3.19	RUS	3.17 - 4.1	UKR	0.**756**996102475508
3.4 - 3.19	RUS	3.1 - 3.16	CZE	0.**368**85033050141086
2.28 - 3.14	SUI	2.28 - 3.14	ISR	0.**609**7487176485175
2.28 - 3.14	SUI	3.17 - 4.1	UKR	0.**768**1247273560939
2.28 - 3.14	SUI	3.1 - 3.16	CZE	0.**644**2048875707921
2.28 - 3.14	ISR	3.17 - 4.1	UKR	0.**723**3607090625718
2.28 - 3.14	ISR	3.1 - 3.16	CZE	-0.**015**26351206753281
3.17 - 4.1	UKR	3.1 - 3.16	CZE	0.**566**9982552001599

Tablica 2: współczynniki korelacji na początkach pandemii w danych krajach

okres	państwo 1	okres	państwo 2	współczynnik korelacji
3.24 - 4.8	RUS	2.29 - 3.15	SUI	0.**945**3390637690183
3.24 - 4.8	RUS	3.9 - 3.24	ISR	0.**946**515845551321
3.21 - 4.5	RUS	3.28 - 4.12	UKR	0.**952**5103255373809
4.5 - 4.20	RUS	3.2 - 3.17	CZE	0.**964**9158173374197
2.29 - 3.15	SUI	3.9 - 3.24	ISR	0.**963**2179283569918
2.28 - 3.14	SUI	3.26 - 4.10	UKR	0.**939**8409292145742
3.5 - 3.20	SUI	3.5 - 3.20	CZE	0.**905**7811674952464
3.5 - 3.20	ISR	3.18 - 4.2	UKR	0.**900**8269988299719
3.4 - 3.19	ISR	3.5 - 3.20	CZE	0.**897**3660971235348
3.24 - 4.8	UKR	3.11 - 3.26	CZE	0.**913**1087754353004

Tablica 3: maksymalne współczynniki korelacji i w jakim okresie wystąpiły

Dla okresów 20-sto dniowych

okres	państwo 1	okres	państwo 2	współczynnik korelacji
3.4 - 3.24	RUS	2.28 - 3.19	SUI	0.**754**7370115556089
3.4 - 3.24	RUS	2.28 - 3.19	ISR	0.**589**0144942715078
3.4 - 3.24	RUS	3.17 - 4.6	UKR	0.**676**7782072826606
3.4 - 3.24	RUS	3.1 - 3.21	CZE	0.**438**06251719919886
2.28 - 3.19	SUI	2.28 - 3.19	ISR	0.**581**8523695452746
2.28 - 3.19	SUI	3.17 - 4.6	UKR	0.**756**5517644880192
2.28 - 3.19	SUI	3.1 - 3.21	CZE	0.**659**6203925133872
2.28 - 3.19	ISR	3.17 - 4.6	UKR	0.**548**7836378333103
2.28 - 3.19	ISR	3.1 - 3.21	CZE	0.**664**6011188816967
3.17 - 4.6	UKR	3.1 - 3.21	CZE	0.**621**7546777712852

Tablica 4: współczynniki korelacji na początkach pandemii w danych krajach

okres	państwo 1	okres	państwo 2	współczynnik korelacji
3.27 - 4.16	RUS	3.3 - 3.23	SUI	0.**915**8528845474712
3.19 - 4.8	RUS	3.4 - 3.24	ISR	0.**944**2508937147224
3.16 - 4.5	RUS	3.23 - 4.12	UKR	0.**957**3890154907134
3.19 - 4.8	RUS	3.6 - 3.26	CZE	0.**921**0324679883152
3.1 - 3.21	SUI	2.29 - 3.20	ISR	0.**939**7894045133509
3.3 - 3.23	SUI	3.23 - 4.12	UKR	0.**918**0072020274089
3.3 - 3.23	SUI	3.8 - 3.28	CZE	0.**904**6920336838308
3.3 - 3.23	ISR	3.28 - 4.17	UKR	0.**908**723142993114
3.3 - 3.23	ISR	3.2 - 3.22	CZE	0.**876**9912988464675
3.18 - 4.7	UKR	3.5 - 3.25	CZE	0.**928**482272504913

Tablica 5: maksymalne współczynniki korelacji i w jakim okresie wystąpiły

Podsumowanie

Każdy kraj zastosował różne formy zapobiegania rozprzestrzeniania się koronawirusa. Te formy przyniosły różne efekty, co potwierdzają dane. Pierwsze, co zwraca uwagę, to bardzo niska korelacja danych z pierwszych 15 dni epidemii w Czechach w porównaniu z resztą krajów. Jak pokazuje Tablica 2, wsp. korelacji danych z Czech z żadnym innym krajem nie przekracza **0.6**, co ma odzwierciedlenie w danych – Czechy, w stosunku do reszty, miały bardzo łagodny początek epidemii. W Tablicy 3 warto zwrócić uwagę na fakt, że Rosja ma maksymalną korelację około 2-3 tygodnie po początku epidemii, z danymi z początków epidemii w innych krajach. To może oznaczać, że Rosja nie poradziła sobie z wypłaszczeniem krzywej zachorowań i bardzo gwałtowny wzrost chorych (jakim charakteryzują się początkowe dni pandemii) w trakcie epidemii może spowodować przepełnienie szpitali. W Tablicy 4 ciekawe wydaje się

unormowanie współczynnika korelacji pomiędzy Czechami, a Izraelem. Wystarczyło zebrać dane o pięć dni dłużej, żeby współczynnik urósł z **-0.015** do **0.66**. To pokazuje, jak łatwo zbyt mała ilość danych może prowadzić do poważnych przekłamań statystycznych. Ostatnia tabela pokazuje, że pomimo pojawienia się wirusa w różnych krajach w różnych terminach, to zwykle, poza Rosją oraz parą UKR i ISR, najwyższa korelacja jest pomiędzy danymi na początku epidemii w każdym z krajów, zaczynając kilka dni po pierwszych przypadkach (te pierwsze zazwyczaj są mocno zaburzone).

Na temat koronawirusa, jego rozprzestrzeniania się i jego skutków powstanie z pewnością wiele prac i opracowań naukowych, ale mamy nadzieję, że ta praca chociaż trochę rozgrzebała temat koraleacji pomiędzy zachorowaniami oraz zrealizowała postawione zadanie.