Instytut Informatyki Uniwersytetu Wrocławskiego

Cezary Świtała, 316746

# Zadania zamiast egzaminu, część 3.

Rachunek prawdopodobieństwa i statystyka 2020/2021

Wrocław, 28 maja 2021

# Spis treści

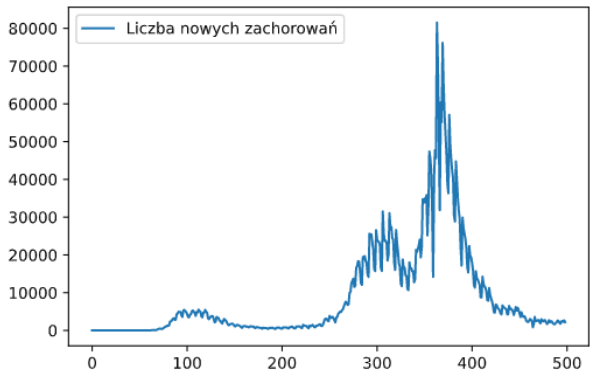
# 1. Regresja liniowa

*Uwaga: Kod dotyczący tej sekcji został umieszczony w pliku regression.py.*

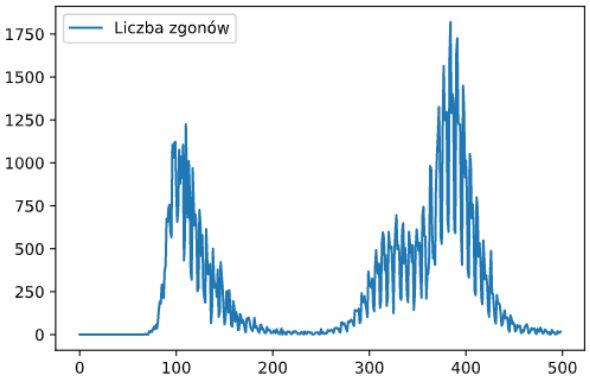
## 1.1. Uśrednienie danych

Otrzymujemy plik z danymi dotyczącymi pandemii koronawirusa. Zawierają one m.in liczby zachorowań oraz zgonów dziennie w okresie od 3 marca 2020r. do 15 maja 2021r. Krajem, któremu będziemy się przyglądać będzie Wielka Brytania.

Dni ponumerujemy kolejno liczbami całkowitymi, począwszy od zera. Aby zniwelować duże różnice w wartościach między kolejnymi dniami, zastąpimy je średnią kroczącą 7-dniową. Najpierw rzućmy okien na dane przed przetworzeniem.



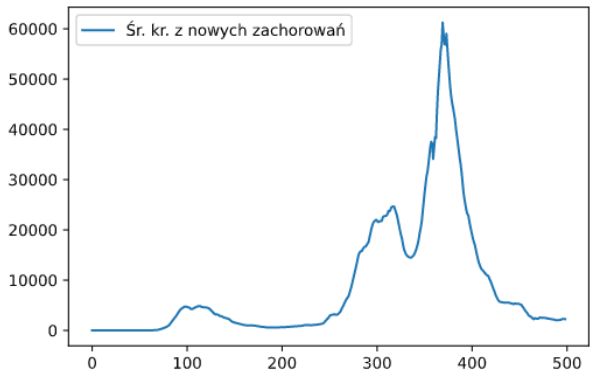

(a) Liczby nowych zachorowań przed uśrednieniem



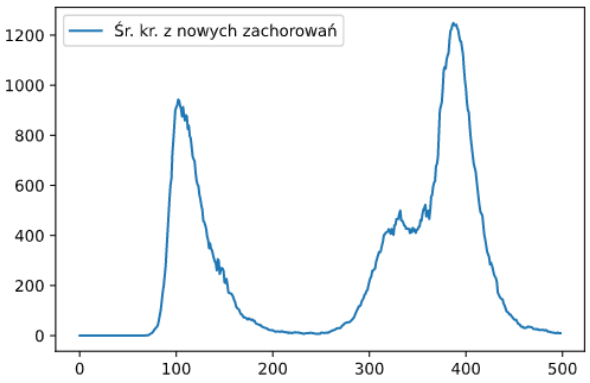

(b) Liczby zgonów przed uśrednieniem

Rysunek 1. Dane przed uśrednieniem

Następnie te same dane, ale z wartościami zastąpionymi przez 7-dniową średnią kroczącą.




(a) Liczby nowych zachorowań po uśrednieniu




(b) Liczby zgonów po uśrednieniu

Rysunek 2. Dane po uśrednieniu

## 1.2. Proste regresji dla całości danych

Wyznaczymy proste regresji na całości danych, a następnie znajdziemy takie przedziały, w których dane da się dobrze przybliżyć taką prostą.

Chcemy zatem prostą w postaci

$$y = \beta_0 + \beta_1 \cdot x,$$

której współczynniki minimalizować będą wartośc funkcji

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i - y_i)^2.$$

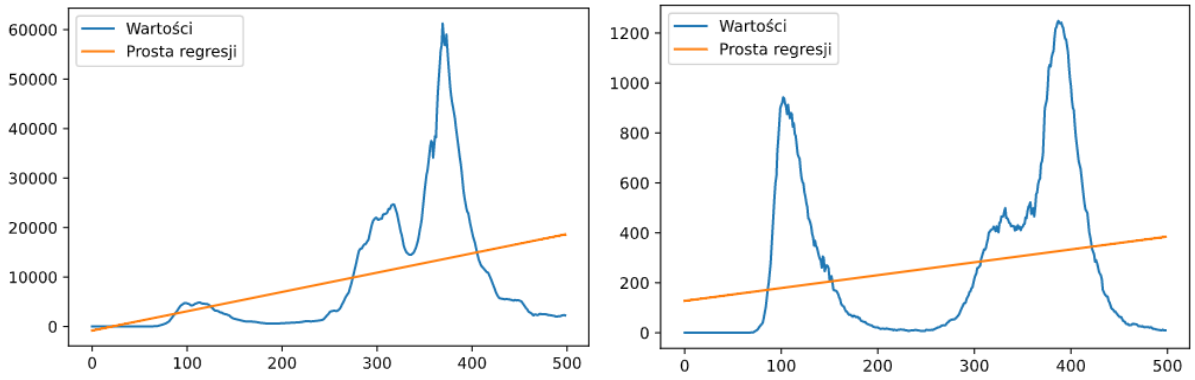
Z wykładu wiemy, że wektor współczynników $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ możemy wyznaczyć z równania

$$\beta = (X^t X)^{-1} X^t Y,$$

gdzie $Y$ to wektor wartości, a $X$ jest macierzą w postaci

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Stosując powyższy wzór otrzymamy proste regresji dla naszych danych.




(a) Prosta regresji nowych zachorowań

(b) Prosta regresji zgonów

Rysunek 3. Proste regresji dla danych

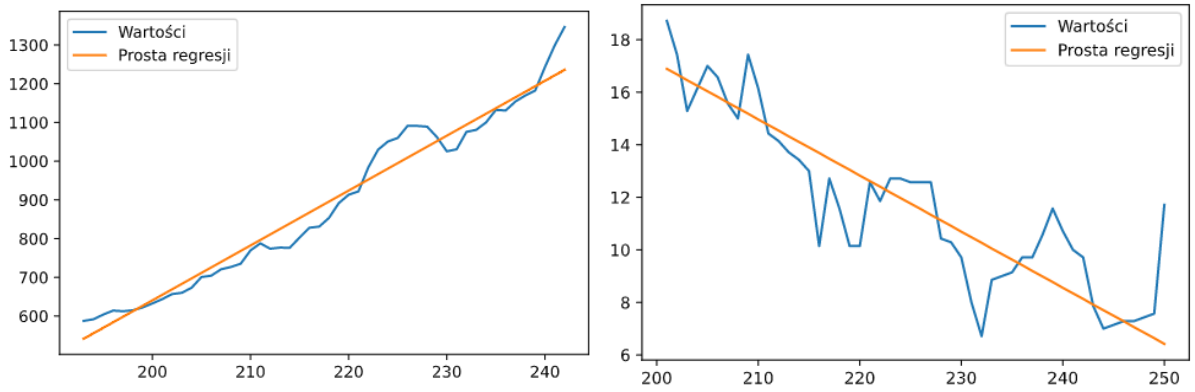
## 1.3. Dobrze przybliżane przedziały

Spróbujemy teraz znaleźć takie krótsze (np. 50-dniowe) przedziały, w których prosta regresji dobrze przybliży wartości. Jako miarę jakości przybliżenia przyjmiemy sumę kwadratów różnic wartości przybliżonej i wartości faktycznej.

$$\sum_{i=1}^{n} (\beta_0 + \beta_1 x_i - y_i)^2,$$

czyli ta sama suma, jak ta któej minimalizowanie leży u podstaw regresji liniowej.

Z poszukiwania takiego przedziału wyłączyłem pierwsze 50 dni, gdyż średnia krocząca w nich nie przekraczała nawet 1, więc zostały świetnie przybliżone przez prostą, ale jednocześnie były mało ciekawe.




(a) Najlepiej przybliżony przedział dla zachorowań

(b) Najlepiej przybliżony przedział dla zgonów

Rysunek 4. Najlepiej przybliżone przedziały

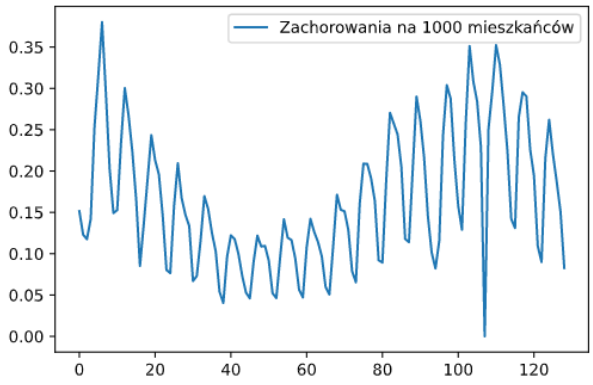
Okazuje się, że nalepiej przybliżony przez prostą regresji 50-cio dniowy przedział dla zachorowań zaczyna się w 193-cim dniu, a dla zgonów w 201-szym dniu.

Moduł *regression.py* udostępnia m.in. funkcję *search_for_the_best*, która przyjmuje cztery argumenty: listę argumentów przybliżanej funkcji, listę wartości, wielkość szukanego przedziału oraz początek i koniec przedziału indeksów argumentów, w którym przedział ma być szukany, a następnie zwraca parę w której pierwszym elementem jest para wyznaczająca najlepiej przybliożony przedział, a drugim wektor współczynników prostej, która ten przedział przybliża.

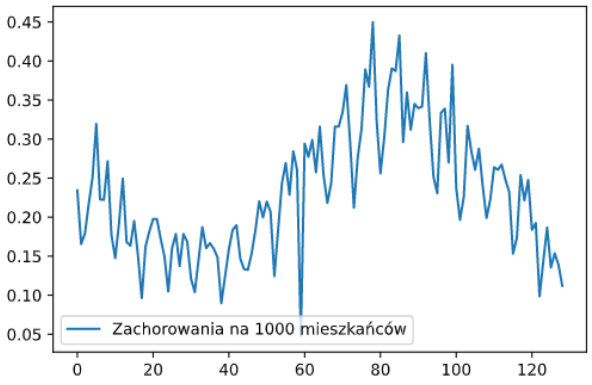
# 2. Równość średnich

W tej sekcji będziemy będziemy testować hipotezy o równości średniej liczby zachorowań na 1000 mieszkańców od 1 stycznia do 10 maja 2021 roku w trzech sąsiednich krajach. Niestety, jako że Wielka Brytania ma tylko jedną granicę lądową, zdecydowałem się rozważać zamiast niej Niemcy, Austrię i Szwajcarię.

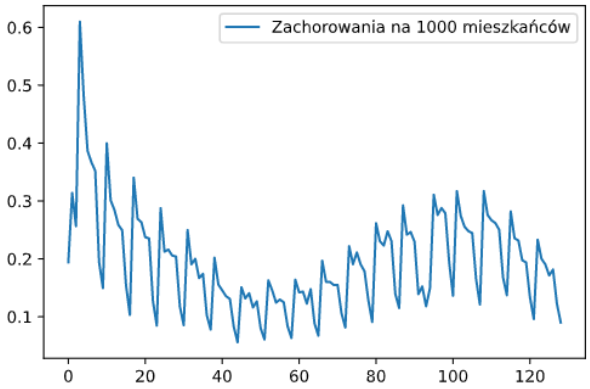
Najpierw znormalizujemy obecne wartości, tak żeby odpowiedały liczbie zachorowań na 1000 mieszkańców.




(a) W Niemczech




(b) W Austrii




(c) W Szwajcarii

Rysunek 5. Liczba dziennych zachorowań na 1000 mieszkańców

## 2.1. Test t-studenta

*Uwaga: Kod dotyczący tej sekcji został umieszczony w pliku t_student.py.*

Najpierw skupimy się na testowaniu hipotezy o równości średnich dwóch krajów. Jest to test na równość średnich dwóch niezależnych zmiennych $X$ i $Y$. Nie znamy ich wariancji, ale możemy spróbować oszacować czy mogą być równe. Wiemy, że wariancję dobrze przybliża wartość $S_X^2$.

$$S_X^2 = \frac{1}{n_X} \sum_{i=1}^{n_X} (x_i - \bar{x})^2.$$

Wyniki zastosowania powyższego wzoru na danych przedstawione są w tabelce poniżej.

| | |
|---|---|
| Niemcy | 0.0065 |
| Szwajcaria | 0.0076 |
| Austria | 0.0066 |

Tablica 1. Szacunkowe wartości wariancji

Nie poznaliśmy sposobów testowania równości wariancji, więc pozostaje nam założyć, że są równe na podstawie bardzo zbliżonych wartości szacujących je. Korzystając z tego założenia, możemy użyć wzoru na test statystyczny, który został podany na wykładzie

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{n_X S_X^2 + n_Y S_Y^2}{n_X n_Y}}}.$$

Przy założeniu, że hipoteza zerowa (o równości średnich) jest prawdziwa, wartość $t$ będzie pochodziła z rozkładu T-studenta z $n_X + n_Y - 2$ stopniami swobody.

Policzmy ją dla każdej pary krajów jaką można utworzyć

| | | |
|---|---|---|
| Niemcy | Austria | -6.45 |
| Niemcy | Szwajcaria | -2.81 |
| Austria | Szwajcaria | -3.40 |

Tablica 2. Przybliżone wartości testu

Następnie możemy przejść do wyznaczenia *p_value*. Przypomnijmy, że dla rozkładów symetrycznych względem osi Y, takich jak rozkład t-studenta, zachodzi

$$p\_value(t) = \begin{cases} 2F(t) & t \leqslant 0 \\ 2(1 - F(t)) & \text{wpp.} \end{cases},$$

gdzie $F(t)$ to dystrubuanta rozkładu. Na podstawie tego wyznaczamy te wartości dla naszych par

| | | |
|---|---|---|
| Niemcy | Austria | $5.4 \cdot 10^{-10}$ |
| Niemcy | Szwajcaria | 0.0052 |
| Austria | Szwajcaria | 0.0007 |

Tablica 3. Przybliżone wartości p_value

Jak widać, poziom istotności $\alpha$ musiałby być bardzo niski żeby chociaż jedna z nich nie została odrzucona. My przyjmiemy poziom istotności $\alpha = 0.1$ i odrzucimy każdą z nich, zatem średnie liczby zachorowań w tych parach sąsiadujących państw nie są równe.

## 2.2. ANOVA

*Uwaga: Kod dotyczący tej sekcji został umieszczony w pliku anova.py.*

Na koniec przetestujemy hipotezę o równości średniej zachorowań we wszystkich trzech krajach jednocześnie. Posłużymy się w tym celu jednoczynnikową analizą wariancji. Jedynym czynnikiem będzie kraj z którego pochodzą obserwacje i będziemy mieli trzy grupy odpowiadające Niemcom, Szwajcarii i Austrii.

Zaczniemy od ustalenia wartość zmienności międzygrupowej $SSA$

$$SSA = \sum_{i=1}^{I} \sum_{j=1}^{J} (x_{i\bullet} - \bar{x})^2 = J \sum_{i=1}^{I} (x_{i\bullet} - \bar{x})^2,$$

gdzie $I$ oznacza liczbę grup, $J$ obserwacji na grupę, $x_{i\bullet}$ średnią wewnątrz grupy i-tej, czyli

$$x_{i\bullet} = \frac{1}{J} \sum_{j=1}^{J} x_{ij},$$

a $\bar{x}$ średnią wszystkich obserwacji, czyli

$$\bar{x} = \frac{1}{IJ} \sum_{i,j} x_{ij}.$$

Z obliczeń wynika że $SSA \approx 0.28$. Z wykładu wiemy, że przy założeniu prawdziwości hipotezy zerowej, z dokładnością do stałej

$$SSA \sim \chi^2(I-1)$$

Teraz możemy przejść do wyznaczenia zmienności wewnątrzgrupowej $SSE$

$$SSE = \sum_{i,j} (x_{ij} - x_{i\bullet})^2$$

Jej wartość okazała się być równa około 2.68. Z wykładu wiemy, że

$$SSE \sim \chi^2(J(I-1))$$

Kolejnym krokiem jest wyznaczenia wartości $MSA$ i $MSE$

$$MSA = \frac{SSA}{I-1}$$

$$MSE = \frac{SSE}{I(J-1)},$$

czyli odpowiednio $SSA$ i $SSE$ podzielone przez ich stopnie swobody.

Wyniki to $MSA \approx 0.14$, $MSE \approx 0.007$. To wszystko robiliśmy po to żeby ostatecznie policzyć wartość $f$

$$f = \frac{MSA}{MSE} \approx 19.74,$$

Jeśli założenie o prawdziwości hipotezy zerowej było prawdziwe to

$$MSA \sim \frac{\chi^2(I-1)}{I-1},\ MSE \sim \frac{\chi^2(I(J-1))}{I(J-1)},$$

czyli wartość $f$ pochodziła z rozkładu Fishera

$$f \sim F(I-1, I(J-1)),$$

bo rozkład Fishera-Snedecore'a może być przedstawiony jako stosunek rozkładów $\chi^2$ podzielonych przez ich stopnie swobody.

Przyjmijmy poziom istotności $\alpha = 0.1$. Pierwszym parametrem rozkładu Fishera, z którego pochodzi $f$ będzie 2, a drugim 384. Jest on ściśle malejący, więc w tym przypadku

$$p\_value = 1 - F(f)$$

Wychodzi ono równe około $6.93 \cdot 10^{-9}$, zatem przy tym poziomie istotności odrzucamy hipotezę o równości średnich zachorowań w tych trzech krajach.