Joanna Mielniczuk

Metody planowania i analizy eksperymentów

Zadanie domowe nr 2

Dane:

Analizie poddano dane na temat przebiegu pandemii COVID-19 w Polsce w okresie od maja do lipca 2020 roku. Dane pochodzą ze strony *gov.pl*, zawierają dzienne informacje o zakażeniach, zgonach, liczbie ozdrowieńców, aktywnych przypadkach choroby, liczbie osób przebywających na kwarantannie oraz znajdujących się pod nadzorem. Dodatkowo, zawarto w nich kumulatywne wskaźniki dla wybranych wartości.

	Data	Nowe przypadki	Przypadki (kumulatywnie)	Zgony	Zgony kumulatywnie	Ozdrowieńcy	Ozdrowieńcy kumulatywnie	Aktywne przypadki	Kwarantanna	Nadzór
0	01.05.2020	228	13105	7	649	271	3762	8694	95625	18383
1	02.05.2020	270	13375	12	661	183	3945	8769	96612	18306
2	03.05.2020	318	13693	15	676	150	4095	8922	96699	17785
3	04.05.2020	313	14006	19	695	185	4280	9031	100765	17291
4	05.05.2020	425	14431	19	714	375	4655	9062	101395	17081
...	...	...	...	...	...	...	...	...	...	...
87	27.07.2020	337	43402	5	1676	187	33043	8683	94920	7245
88	28.07.2020	502	43904	6	1682	147	33190	9032	95453	6222
89	29.07.2020	512	44416	12	1694	453	33643	9079	97561	8094
90	30.07.2020	615	45031	15	1709	344	33987	9335	97189	8069
91	31.07.2020	657	45688	7	1716	387	34374	9598	98282	8241

92 rows × 10 columns

Narzędzie analizy danych:

Do analizy danych wykorzystano język programowania *Python* oraz biblioteki *Pandas*, *Numpy* i *Matplotlib*. Wyniki przedstawiono w pliku typu *Jupyter Notebook*, a do jego stworzenia wykorzystano *Google Colabolatory*.

Link do *Google Colabolatory*:
https://colab.research.google.com/drive/1CfbrxPkWOE6X1VgArcsZTVHhAkpKhu-R?usp=sharing

Estymacja punktowa:

Przeprowadzono estymację punktową – estymację wartości oczekiwanej dla kolumny z danymi na temat dziennej liczby nowych przypadków zakażenia koronawirusem. Skorzystano w tym celu ze wzoru:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Z estymacji uzyskano wartość 356.641 – jest to oczekiwana dzienna liczba nowych przypadków. Poniżej przedstawiono kod źródłowy obliczeń.

```
[ ] n = 92
    samplesNewCases = data['Nowe przypadki'].values
    samplesNewCases
```

```
array([228, 270, 318, 313, 425, 309, 307, 319, 285, 345, 330, 595, 283,
       411, 401, 241, 272, 356, 383, 471, 404, 476, 312, 395, 305, 443,
       399, 352, 330, 416, 215, 379, 230, 292, 361, 362, 576, 575, 599,
       400, 282, 359, 376, 440, 375, 396, 407, 506, 314, 301, 304, 311,
       296, 300, 294, 298, 276, 319, 193, 247, 239, 382, 371, 259, 314,
       231, 205, 257, 277, 262, 265, 305, 370, 299, 267, 264, 333, 353,
       339, 358, 279, 399, 380, 418, 458, 584, 443, 337, 502, 512, 615,
       657])
```

```
[ ] def myAverage(samples):
        return round(np.average(samples), 3)
```

```
[ ] avg = myAverage(samplesNewCases)
    avg
```

```
356.641
```

Estymacja przedziałowa:
Dla dziennej liczby nowych przypadków przeprowadzono także estymację przedziałową – skonstruowano przedział ufności o poziomie 95%, czyli taki, w którym na 95% znajdzie się wartość oczekiwana. W tym celu skorzystano ze wzoru:

$$\left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Z tablic rozkładu normalnego odczytano dla *alfa* = 0.475 (0.95 / 2) wartość 1.96. Obliczono odchylenie standardowe:

```
[ ] def diffSquare(sample, avg):
        return (avg - sample) ** 2

    def standardDev(samples, avg, n):
        squareDiffs = list(map(lambda s: diffSquare(s, avg), samples))
        return round(np.sqrt(np.sum(squareDiffs) / n), 3)
```

```
[ ] s = standardDev(samplesNewCases, avg, n)
    s
```

```
98.673
```

Przyjmując *n = 92* i podstawiając do wzoru:

```
[ ] ua = 1.96
    sqrtN = np.sqrt(n)
    factor = ua * s / sqrtN

    bottom = round(avg - factor, 3)
    top = round(avg + factor, 3)

    print('Bottom: ', bottom)
    print('Top: ', top)
```

```
Bottom:  336.478
Top:  376.804
```

otrzymano przedział [336.478, 376.804], w którym na 95% znajdzie się wartość oczekiwana dziennej liczby nowych przypadków zakażenia koronawirusem.

Weryfikacja hipotezy statystycznej

Za hipotezę zerową przyjęto, że średnia liczba zgonów z powodu koronawirusa w maju 2020 roku była równa średniej ilości zgonów w czerwcu, a alternatywną – że była mniejsza. Przyjęto w tym celu średnie wartości dziennych zgonów z obu miesięcy:

```
[ ] dataMay = data[data.apply(lambda d: (dt.datetime.strptime(d.Data, "%d.%m.%Y")).month == 5, axis=1)]
    dataJune = data[data.apply(lambda d: (dt.datetime.strptime(d.Data, "%d.%m.%Y")).month == 6, axis=1)]
```

```
[ ] samplesMay = dataMay['Zgony'].values
    samplesJune = dataJune['Zgony'].values
```

```
[ ] avgMay = myAverage(samplesMay)
    avgMay

    13.581
```

```
[ ] avgJune = myAverage(samplesJune)
    avgJune

    13.333
```

a następnie podstawiono do wzoru na wartość statystyki testowej:

$$t = \frac{X - m_0}{\frac{S}{\sqrt{n}}}$$

```
t = round(np.sqrt(30) * (avgJune - avgMay) / standardDev(samplesJune, avgJune, 30), 3)
t

-0.202
```

Otrzymano wartość -0.202, która jest większa od granicznej wartości zbioru krytycznego -1.69726 odczytanej z tablic rozkładu t-Studenta, co sprawia, że nie ma podstaw do odrzucenia hipotezy zerowej.