Świtalska Iga, Rakus Karolina

Statytyka Stosowana

Analiza wybranych danych rzeczywistych
z wykorzystaniem metod statystyki opisowej

10 grudnia 2021

1. Wstęp

1.1. Opis danych

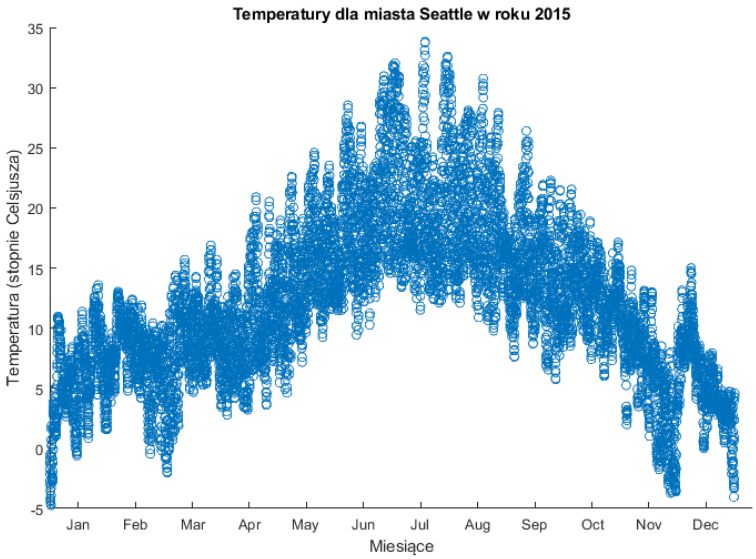
Dane wykorzystane w sprawozdaniu dotyczą składników pogody, takich jak temperatura, wilgotność, ciśnienie powietrza, kierunek oraz prędkość wiatru. Pomiarów dokonywano co godzinę na przestrzeni 5 lat (2013-2017) dla trzydziestu sześciu miast ze Stanów Zjednoczonych, Kanady oraz Izraelu.

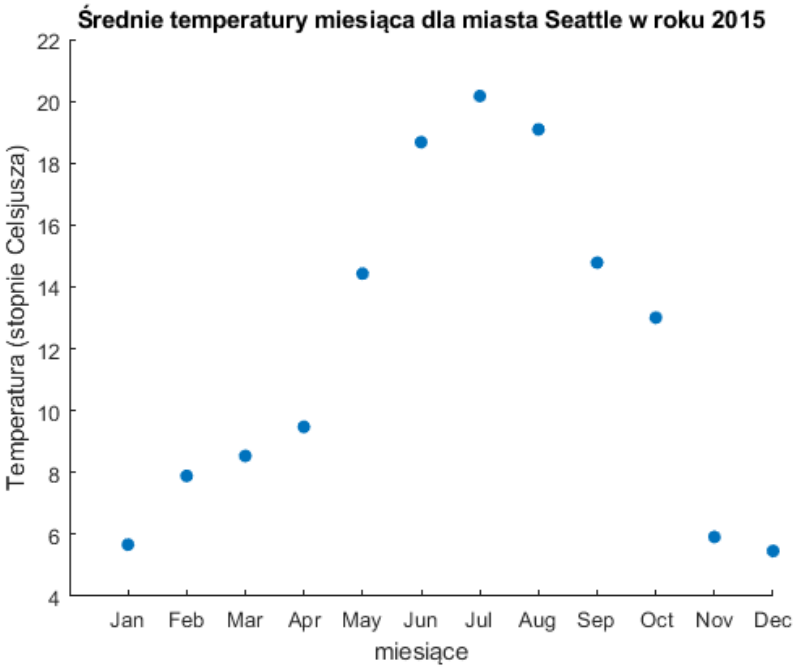
Analizę danych dokonamy na pomiarach temperaturowych dla miasta Seattle w roku 2015. Długość próby wynosi 8760.

Źródło danych: dane zgromadzono przy pomocy aplikacji Weather Api, a następnie umieszczone na stronie kaggle.

https://www.kaggle.com/selfishgene/historical-hourly-weather-data

1.2. Wykresy przedstawiające dane





2. Podstawowe statystyki

2.1. Miary położenia

Opisują umiejscowienie rozkładu na osi.

Wartość średnia

Wartością średnią, oznaczamy $\overline{x}$, nazywamy średnią arytmetyczną wartości cechy w próbie

$$\overline{x} = \frac{1}{n}\sum_{n=1}^{n} x_i. \tag{1}$$

Wartość średnia dla zadanej próby wynosi $11.9559°C$.

Mediana

Medianą w próbie(lub medianą próby) oznaczamy x_{med}, nazywamy następującą wartość

$$x_{med} = \begin{cases} x_{(\frac{n+1}{2})}, & \text{gdy n nieparzyste} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right), & \text{gdy n parzyste.} \end{cases} \tag{2}$$

Mediana dla zadanej próby wynosi $11.4173°C$.

Wartość modalna (moda)

Jest to wartość najczęściej występująca w próbie.

Moda dla zadanej próby wynosi $13°C$.

Średnia ucinana

Średnią ucinaną (z parametrem k), oznaczaną $\overline{x}_{tk}$, nazywamy wielkość

$$\overline{x}_{tk} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i).} \tag{3}$$

Średnia ucinana dla zadanej próby (po ucięciu łącznie 10 % danych odstających) wynosi $11.7911°C$.

Średnia winsorowska

Średnią winsorowską (z parametrem k), oznaczaną $\overline{x}_{wk}$, nazywamy wielkość

$$\overline{x}_{wk} = \frac{1}{n}[(k+1)x_{(k+1)} + \sum_{i=k+2}^{n-k-1} x_{(i)} + (k+1)x_{(n-k)}] \tag{4}$$

Średnia winsorowska dla zadanej próby (po ucięciu łącznie 10 % danych odstających) wynosi $11.9086°C$.

Średnia geometryczna

Średnią geometryczną n dodatnich liczb $a_1, a_2...a_n$, oznaczaną g_n, nazywamy wielkość

$$g_n = \sqrt[n]{a_1 \cdot a_2 \cdot ... \cdot a_n}. \tag{5}$$

(dla nszaej próby nie da się policzyć ze względu na wartości ujemne)

Średnia harmoniczna

Średnią harmoniczną n dodatnich liczb $a_1, a_2...a_n$, oznaczaną h_n, nazywamy wielkość

$$h_n = \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + ... + \frac{1}{a_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{a_i}}. \tag{6}$$

Średnia harmoniczna dla zadanej próby wynosi $7.4958°C$.

Kwartyle

Dolnym (pierwszym) kwartylem próby nazywamy medianę podpróby, składającej się ze wszystkich elementów próby o wartościach mniejszych od mediany całej próby.

Dolny kwartyl dla zadanej próby wynosi $7.3044°C$.

Górnym (trzecim) kwartylem próby nazywamy medianę podpróby, składającej się ze wszystkich elementów próby o wartościach większych od mediany całej próby.

Górny kwartyl dla zadanej próby wynosi $16.1588°C$.

Medianę całej próby nazywamy również **drugim kwartylem** całej próby.

Dolny kwartyl oznaczamy symbolem Q_1, górny symbolem Q_3, medianę oznaczamy niekiedy Q_2.

Maksymalna i minimalna wartość

Maksymalna wartość dla zadanej próby wynosi $33.8000°C$.
Minimalna wartość dla zadanej próby wynosi $-4.7250°C$.

2.2. Miary rozproszenia

Dostarczają informacji jak bardzo zróżnicowana jest populacja pod względem badanej cechy X.

Rozstęp z próby

Rozstępem z próby o liczności n, oznaczanym R, nazywamy wielkość

$$R = x_{(n)} - x_{(1)} \tag{7}$$

gdzie $x_{(1)}$ i $x_{(n)}$ są odpowiednio, najmniejszym i największym elementem w próbie.

Rozstęp z próby dla zadanej próby wynosi $38.5250°C$.

Wariancja

Wariancją w próbie, oznaczaną s^2, nazywamy wielkość

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{(i)} - \overline{x})^2, \tag{8}$$

gdzie $\overline{x}$ oznacza średnią w próbie.

Wariancja dla zadanej próby wynosi $42.5710°C$.

Odchylenie standardowe

Odchylenie standardowe cechy w próbie, oznaczane s to pierwiastek z wariancji

$$s = \sqrt{s^2}. \tag{9}$$

Odchylenie standardowe dla zadanej próby wynosi $6.5246°C$.

Odchylenie przeciętne

Odchyleniem przeciętnym od wartości średniej, oznaczanym d_1 nazywamy wielkość

$$d_1 = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|. \tag{10}$$

Odchylenie przeciętne od wartości średniej dla zadanej próby wynosi $5.2323°C$.

Odchylenie ćwiartkowe

Odchyleniem ćwiartkowym, oznaczanym Q nazywamy połowę różnicy między kwartylem dolnym a górnym

$$Q = \frac{1}{2}(Q_3 - Q_1). \tag{11}$$

Odchylenie ćwiartkowe dla zadanej próby wynosi $4.42715°C$.

Rozstęp międzykwartylowy

Rozstępem międzykwartylowym, oznaczanym IQR, nazywamy wielkość

$$IQR = Q_3 - Q_1. \tag{12}$$

Rozstęp międzykwartylowy dla zadanej próby wynosi $8.8543°C$.

Współczynnik zmienności

Współczynnik zmienności, oznaczany V,

$$V = \frac{s}{\overline{x}}, \tag{13}$$

gdzie s oznacza odchylenie standardowe, $\overline{x}$ średnią arytmetyczną z próby.

Współczynnik zmienność dla zadanej próby wynosi 0.5457.

2.3. Miary asymetrii

Miary asymetrii mówią nam, czy większa część populacji znajduje się powyżej, czy poniżej przeciętnego poziomu badanej cechy X

Współczynnik skośności (asymetrii)

Współczynnikiem skośności, oznaczanym β_1, nazywamy wielkość

$$\beta_1 = \frac{\mu_3}{s^3}, \tag{14}$$

gdzie μ_3 jest trzecim momentem centralnym, a s odchyleniem standardowym z próby.
Moment centralny j-tego rzędy wyrażamy wzorem

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n}(x_{(i)} - \overline{x})^j, \tag{15}$$

gdzie $\overline{x}$ oznacza średnią w próbie.

Współczynnik skośności dla zadanej próby wynosi 0.3825.

2.4. Miary spłaszczenia (koncentracji)

Miary spłaszczenia mówią nam o koncentracji rozkładu wokół wartości oczekiwanej.

Kurtoza Kurtozą, oznaczaną $Kurt$, nazywamy wielkość

$$Kurt = \frac{\mu_4}{s^4}, \tag{16}$$

gdzie μ_4 jest trzecim momentem centralnym, a s odchyleniem standardowym z próby.
Im wyższa jest wartość współczynnika $Kurt$, tym krzywa liczebności wskazuje na tendencję do skupiania się jednostek wokół średniej.

Współczynnik kurtozy dla zadanej próby wynosi 2.9620.

Ekscez Ekscezą, oznaczaną Ex, nazywamy wielkość

$$Ex = Kurt - 3. \tag{17}$$

Mówi nam czy koncentracja wartości badanej zmiennej wokół średniej w danym rozkładzie jest większa czy mniejsza niż w zbiorowości o rozkładzie normalnym. Ex przyjmuje wartość równą 0 gdy rozkład ma kształt normalny.

Współczynnik ekscezy dla zadanej próby wynosi -0.0380.

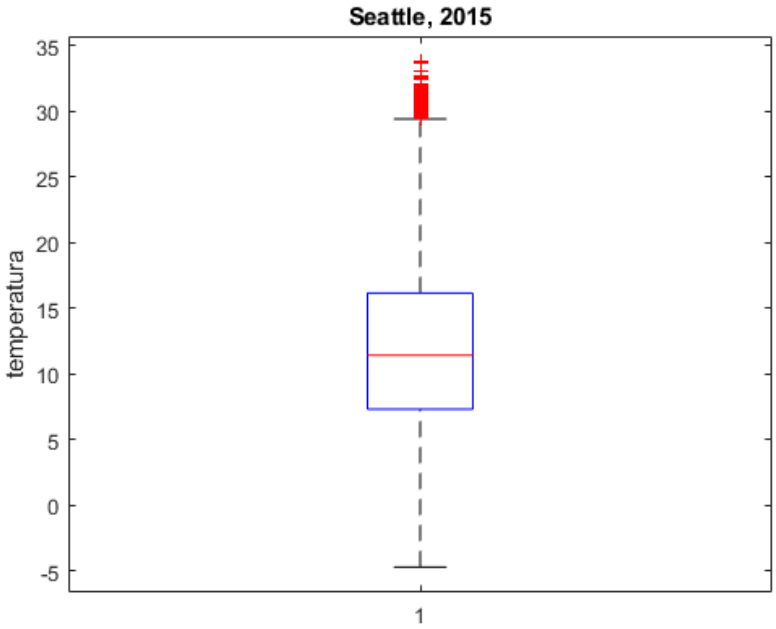
2.5. Interpretacja wyników

Mediana, średnia arytmetyczna, średnia ucinana i średnia winsorowska są do siebie zbliżone, co jest charkterystyczne dla rozkładów symetrycznych. Jednak mediana jest w niewielkim stopniu większa od średniej arytmetycznej, co świadczy o prawostronnej skośności rozkładu badanej próby. Dodatni współczynnik skośności również wskazuje na prawostronną asymetrię. Eksceza w niewielkim stopniu różni się od zera. Zatem rozkład danych pomimo asymetrii jest zbilżony do rozkładu normalnego.

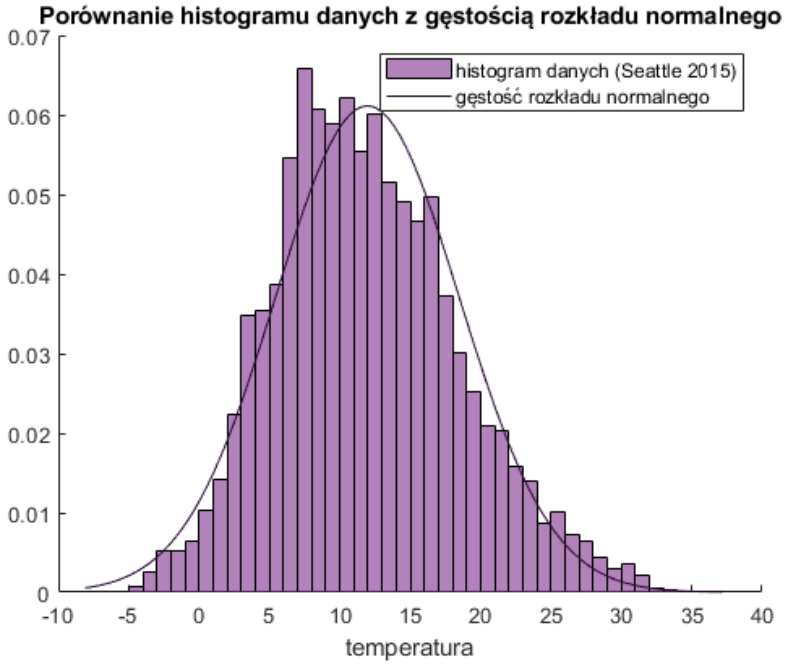
Współczynnik zmienności i rozstęp z próby mają duże wartości, wskazujące na duże zróżnicowanie temperatur w badanej próbie. Może to wynikać ze zmian pór roku.

3. Wizualizacja danych

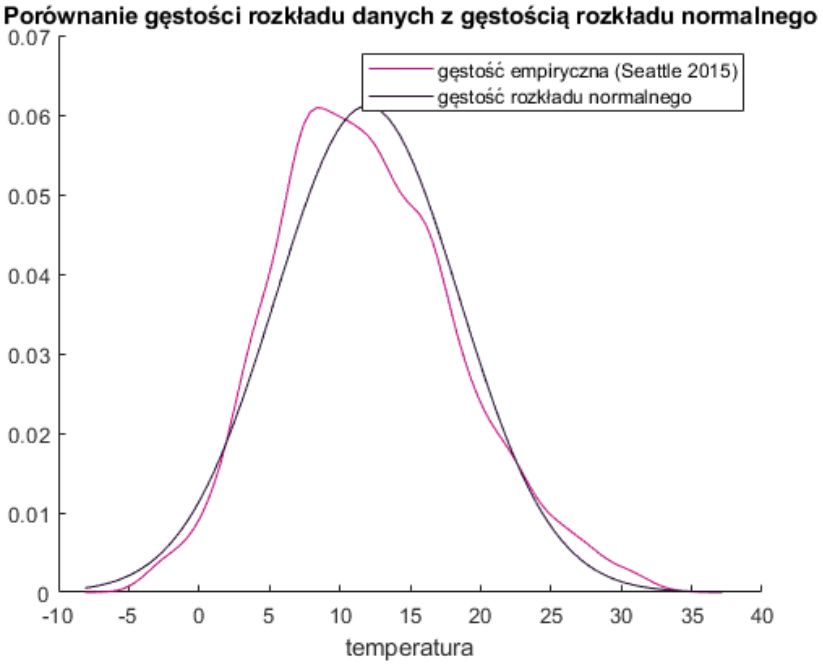
3.1. Wykres pudełkowy



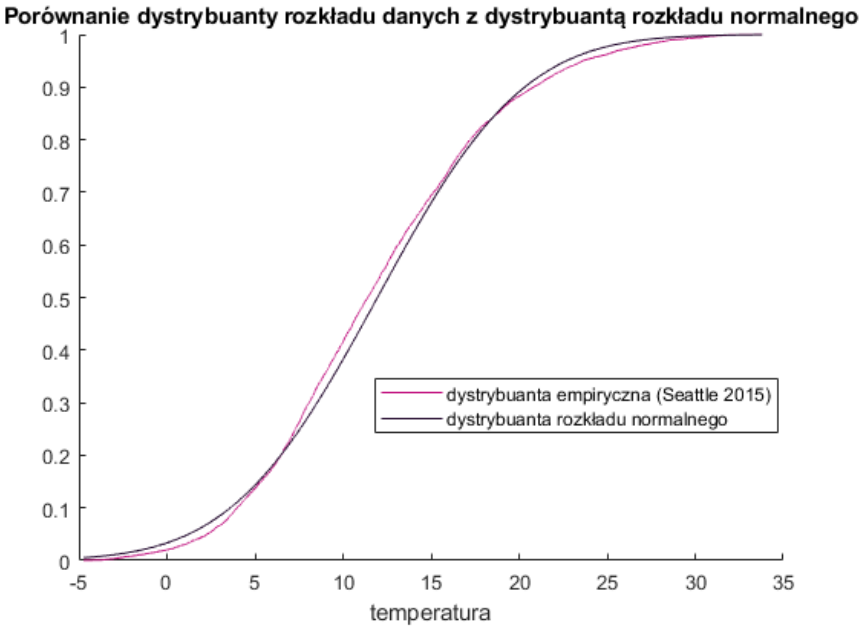
3.2. Histogram



3.3. Gęstość empiryczna



3.4. Dystrybuanta empiryczna



3.5. Interpretacja wyników

Z wykresu pudełkowego widać niewielką prawostronną skośność rozkładu.

4. Dodatkowe