Wydział Elektroniki i Technik Informacyjnych
Politechnika Warszawska

Statystyka w analizie danych

Raport z projektu 2

Adam Dąbkowski, Jan Kwiatkowski

Warszawa, 2023

Spis treści

I. Problem 1

II. Problem 2

Część I

Problem 1

1. Wstęp

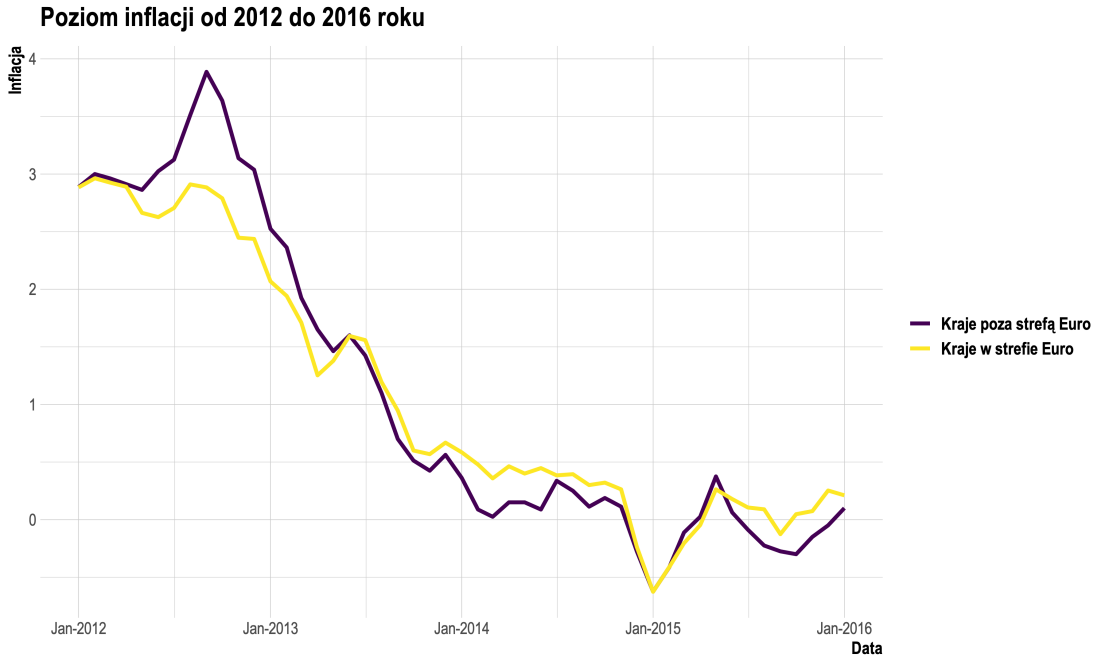
Celem problemu 1 jest zbadanie, czy w wybranym okresie inflacja w krajach strefy euro była niższa niż w krajach UE spoza strefy euro. W przeciwieństwie do projektu 1, w niniejszym problemie należy skupić się na weryfikacji hipotez statystycznych, a co za tym idzie, konieczne będzie wykorzystanie statystyki matematycznej do określenia, czy dane obserwacje potwierdzają lub odrzucają sformułowane hipotezy, które przejawiają istotę badanego problemu. Oprócz samego sformułowania hipotez statystycznych konieczne jest wybranie odpowiedniego testu oraz przeprowadzenie go, jak i zinterpretowanie otrzymanego wyniku. W celach pomocniczych wykorzystane zostaną także różne techniki prezentacji danych. Dane wykorzystywane w problemie 1, podobnie jak wcześniej, pochodzą dane z Europejskiego Banku Centralnego, natomiast badanym okresem będzie okres od 2012 do początku roku 2016.

2. Wizualizacja danych

Analizę problemu 1 rozpoczniemy od wizualizacji danych co, po połączeniu z testami statystycznymi, pozwoli na pełniejsze i bardziej rzetelne rozumienie inflacji w obu strefach. Już po zapoznaniu się z załączonymi wykresami, będziemy w stanie wykryć potencjalne wzorce bądź tendencje, które następnie skonfrontujemy z wynikami weryfikacji hipotez statystycznych.

2.1. Wykres liniowy

Pierwszym z zamieszczonych wykresów jest ten na rysunku rys. 2.1, który przedstawia średni poziom inflacji od 2012 do początku 2016 roku w strefie euro oraz w krajach UE poza strefą euro.

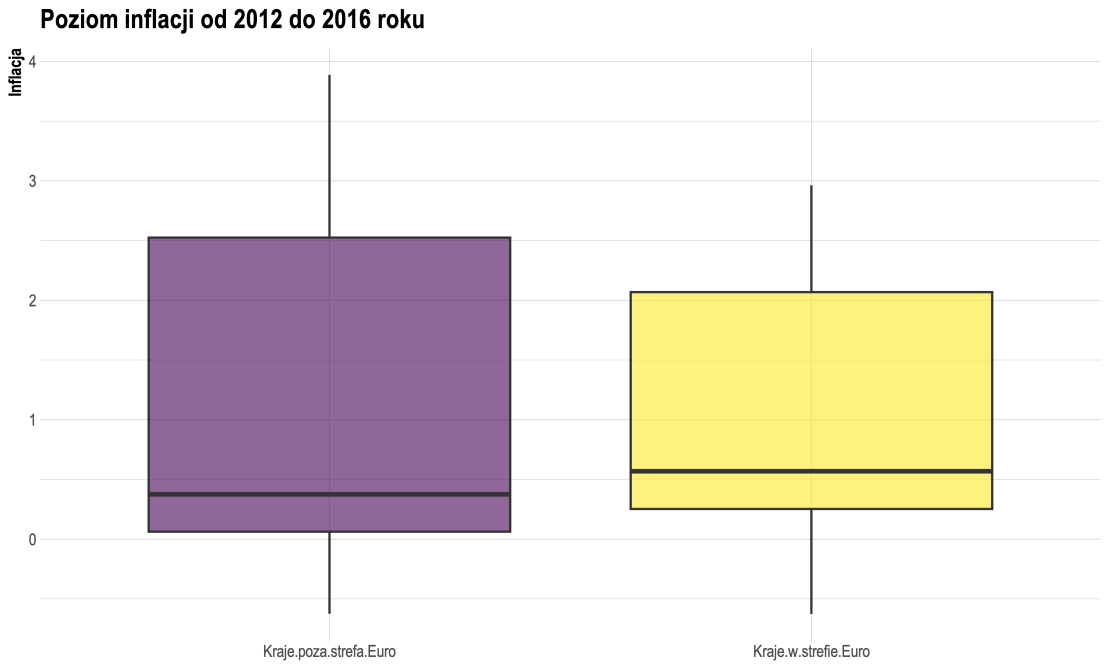


Rys. 2.1. Wykres przedstawiający średni poziom inflacji od 2012 do początku 2016 roku w strefie euro oraz w krajach UE poza strefą euro

Od razu możemy dostrzec, że mamy do czynienia z dość stabilnym okresem, cechującym się stosunkowo niskim poziomem inflacji. Ciężko jest jednak jednoznacznie stwierdzić, czy poziom inflacji w krajach strefy euro jest rzeczywiście niższy niż w krajach spoza strefy. Oczywiście łatwo zauważyć, że do kwietnia 2013 roku poziom inflacji w krajach strefy euro jest definitywnie niższy, jednakże nie możemy tego powiedzieć w przypadku pozostałej części badanego okresu.

2.2. Wykres pudełkowy

Drugą z zastosowanych form prezentacji danych jest wykres pudełkowy, który w zwięzłej formie zarysowuje główne cechy rozkładu badanych przez nas danych. Wykres ten został zamieszony na rysunku rys. 2.2.



Rys. 2.2. Wykres pudełkowy przedstawiający średni poziom inflacji od 2012 do początku 2016 roku w strefie euro oraz w krajach UE poza strefą euro

Analizując wspominany wykres, możemy odczytać kilka istotnych informacji dotyczących poziomu inflacji w poszczególnych strefach.

Zarówno w jednej, jak i drugiej grupie państw Unii Europejskiej poziomy inflacji plasują się na podobnym poziomie, jednakże widzimy, że mediana w strefie euro przyjmuje nieco wyższe wartości. Podobnie jest z pierwszym kwartylem. Sytuacja jest zgoła odmienna w przypadku kwartyla trzeciego. Widzimy zatem, że poziom inflacji w krajach poza strefą cechuje się większą rozbieżnością niż ma to miejsce w drugiej z rozpatrywanych grup. Chcąc sprawdzić dokładne wartości omawianych metryk, załączona zostaje tabela tab. 2.1.

Dzięki przytoczeniu wspomnianej tabeli, nasze dotychczasowe wnioski zostają wzbogacone o wartości średnich w każdej ze stref, a także wartości minimalne i maksymalne. W naszym przypadku szczególnie istotna jest średnia, ponieważ to właśnie ta metryka w głównej mierze decyduje o tym, w której strefie mamy do czynienia z wyższą bądź niższą inflacją. Jak widać średnie w obu strefach przyjmują zbliżone wartości, dlatego aby sprawdzić hipotezę mówiącą o tym, że poziom inflacji w strefie euro jest niższy niż w krajach poza strefą euro, wykorzystamy testy statystyczne.

Tab. 2.1. Podsumowanie danych o poziomie inflacji w obu strefach

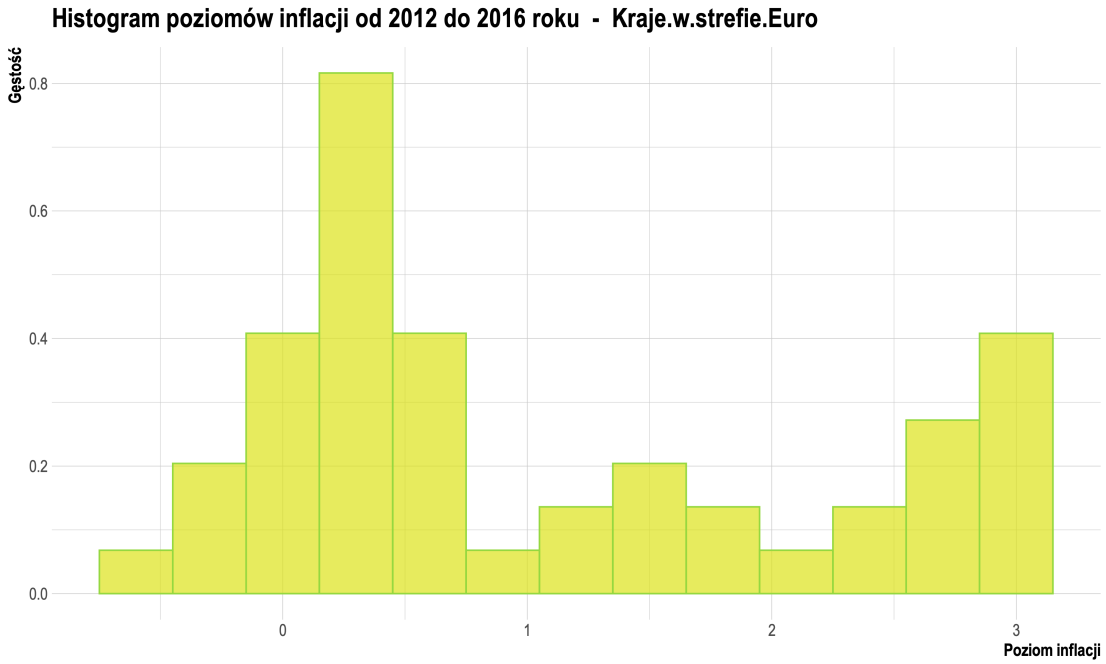
Strefa	Minimum	Dolny kwartyl	Mediana	Średnia	Górny kwartyl	Maximum
Kraje w strefie euro	-0,6263	0,2526	0,5684	1,0729	2,0684	2,9632
Kraje poza strefą euro	-0,6250	0,0625	0,3750	1,1051	2,5250	3,8875

3. Badanie normalności rozkładu

Zanim przejdziemy do głównej części omawianego problemu, w pierwszej kolejności należy ocenić normalność rozkładów wykorzystywanych danych, po to aby móc dobrać odpowiedni test podczas sprawdzania, czy poziom inflacji w strefie euro jest niższy niż w krajach poza strefą euro.

W tym celu wykorzystany został test Kołgomorowa-Smirnowa. Podczas badania normalności rozkładu, hipoteza zerowa zakłada, że dane pochodzą z rozkładu normalnego, natomiast hipoteza alternatywna temu zaprzecza.

Przed wykonaniem wspomnianego testu, warto jednak zobaczyć rozkład badanych danych. Świetnym rozwiązaniem jest tutaj stworzenie histogramów, które pozwalają na dostrzeżenie, które poziomy inflacji były przyjmowane częściej, a które praktycznie nie występowały (rys. 3.1 i rys. 3.2).

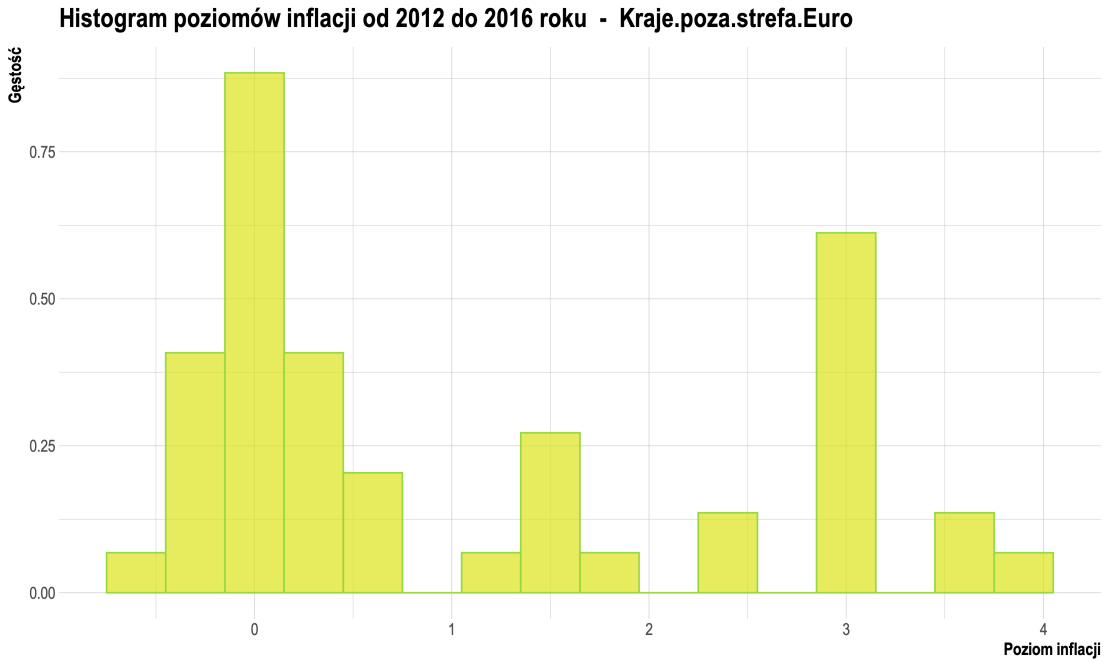


Rys. 3.1. Histogram średnich poziomów inflacji od 2012 do początku 2016 roku w strefie euro

Po przeprowadzeniu testów Kołgomorowa-Smirnowa otrzymaliśmy następujące wyniki (tab. 3.1).

Tab. 3.1. Wyniki badania normalności przy użyciu testu Kołgomorowa-Smirnowa

Strefa	Statystyka testowa D	p-wartość
Kraje w strefie euro	0,39644	$1,859 \cdot 10^{-7}$
Kraje poza strefą euro	0,34127	$2,209 \cdot 10^{-5}$

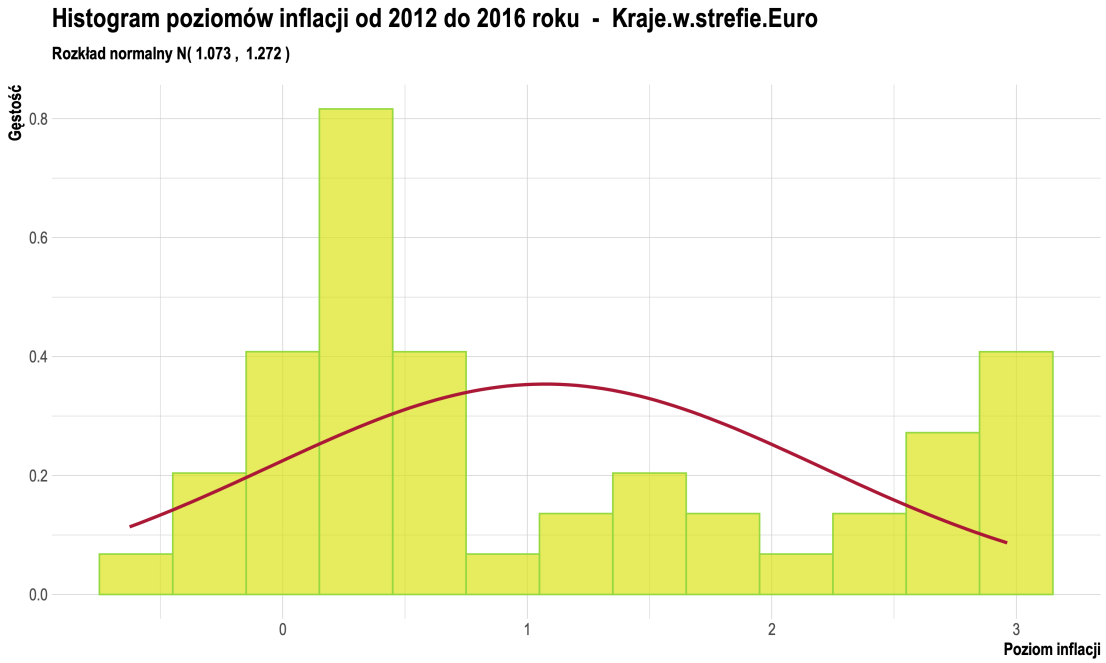


Rys. 3.2. Histogram średnich poziomów inflacji od 2012 do początku 2016 roku w krajach UE poza strefą euro

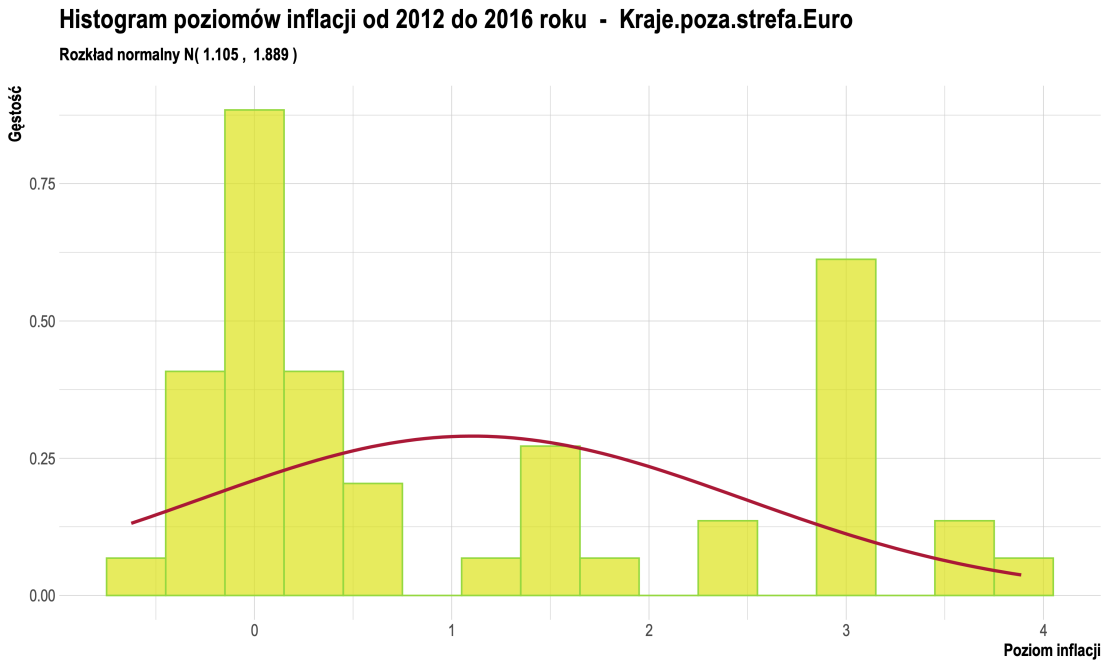
Testy przeprowadzone, zarówno dla strefy euro, jak i dla krajów UE spoza tej strefy odrzucają hipotezę zerową, mówiącą o normalności badanych rozkładów, co sugeruje, że rozkłady poziomów inflacji w obu grupach nie są normalne, na co wskazuje hipoteza alternatywna. Wniosek ten został sformułowany przy założeniu, że poziom istotność wynosi $\alpha = 0,05$, co jest wartością znacznie większą niż uzyskane w obu przypadkach p-wartości ($1,859 \cdot 10^{-7}$ dla strefy euro i $2,209 \cdot 10^{-5}$ dla krajów z poza tej strefy).

Aby utwierdzić się w przekonaniu o słuszności wyniku testów, na rysunkach rys. 3.3 i rys. 3.4 zamieszczone zostały wcześniejsze histogramy, tym razem z nałożonym rozkładem normalnym o parametrach μ i σ bezpośrednio wynikających z poszczególnych danych.

Jak łatwo zauważyć, rozkład normalny nie jest w stanie się dopasować do danych zwizualizowanych za pomocą histogramów, dzięki czemu wiemy, że wyniki testu Kołgomorowa-Smirnowa są słuszne.



Rys. 3.3. Histogram średnich poziomów inflacji od 2012 do początku 2016 roku w strefie euro z zaznaczonym rozkładem normalnym o parametrach $\mu = 1,073$, $\sigma = 1,128$



Rys. 3.4. Histogram średnich poziomów inflacji od 2012 do początku 2016 roku w krajach UE poza strefą euro z zaznaczonym rozkładem normalnym o parametrach $\mu = 1,105$, $\sigma = 1,374$

4. Badanie niezależności zmiennych losowych

Aby dobrać odpowiedni test do sprawdzenia, czy w wybranym okresie inflacja w krajach strefy euro była niższa niż w krajach UE spoza strefy euro, konieczne będzie także zbadanie niezależności zmiennych losowych, w naszym przypadku jest to poziom inflacji w krajach strefy euro oraz w grupie państw do niej nie należących.

Jednym z najczęściej wykorzystywanych testów do sprawdzenia niezależności zmiennych losowych jest test chi-kwadrat χ^2. W wyniku jego zastosowania otrzymujemy następujące wyniki (tab. 4.1).

Tab. 4.1. Wyniki badania niezależności zmiennych losowych przy użyciu testu chi-kwadrat χ^2

Wartość statystyki testowej	Liczba stopni swobody	p-wartość
2021,2	1978	0,2441

Wynik testu chi-kwadrat χ^2 nie pozwala na jednoznaczne stwierdzenie, że badane zmienne są niezależne. W rozpatrywanym przypadku p-wartość wynosi 0,2441, co przy poziomie istotności α równym 0,05 jest wartością, która skłania nas do odrzucenia hipotezy alternatywnej, mówiącej o zależności zmiennych, jednakże pozostanie przy hipotezie zerowej wskazującej na niezależność badanych zmiennych, nie jest tożsame z twierdzeniem, że są niezależne.

Aby przekonać się o wzajemnym związku pomiędzy poziomami inflacji w obu strefach, sprawdzona została korelacja pomiędzy ich wartościami. W rezultacie otrzymaliśmy, że korelacja wynosi aż 0,985, co świadczy o silnym związku, czego można się było spodziewać analizując wykres przedstawiający średni poziom inflacji w strefie euro oraz w krajach UE poza strefą euro (rys. 2.1).

5. Badanie relacji poziomów inflacji w obu strefach UE

5.1. Zdefiniowanie hipotez oraz błędów I i II rodzaju

Po przeprowadzeniu szeregu testów i doświadczeń możemy przejść do głównej części problemu 1, a mianowicie sprawdzenia czy w okresie od 2012 roku do początku roku 2016 inflacja w krajach strefy euro była niższa niż w krajach UE spoza strefy euro. Pierwszym krokiem będzie sformułowanie hipotezy zerowej H_0 oraz hipotezy alternatywnej H_1, które wyglądają następująco:

H_0 – inflacja w krajach strefy euro jest równa inflacji w krajach UE spoza strefy euro

H_1 – inflacja w krajach strefy euro jest niższa niż inflacja w krajach UE spoza strefy euro

Oprócz określenia hipotez, warto zdefiniować błędy I i II rodzaju. Z błędem I rodzaju mamy do czynienia wówczas, gdy odrzuca się hipotezę zerową, kiedy jest ona prawdziwa. Błąd II rodzaju jest to natomiast sytuacja, w której nie odrzuca się hipotezy zerowej, podczas gdy jest ona fałszywa. W naszym przypadku wspomnianymi błędami są:

Błąd I rodzaju – odrzucenie hipotezy, wskazującej na to, że inflacja w krajach strefy euro jest równa inflacji w krajach UE spoza strefy euro, podczas gdy w rzeczywistości tak właśnie jest.

Błąd II rodzaju - brak odrzucenia hipotezy, mówiącej o tym, że inflacja w krajach strefy euro jest równa inflacji w krajach UE spoza strefy euro, podczas gdy w rzeczywistości jest niższa.

Jak w każdym z przeprowadzonych dotąd testów, kluczowe jest określenie poziomu istotności α, który będzie decydował o odrzuceniu, bądź pozostaniu przy hipotezie zerowej. Podobnie jak wcześniej poziom ten będzie wynosił 0,05.

5.2. Niesparowany test Wilcoxona

Mając tak zdefiniowane hipotezy H_0 i H_1, możemy przejść do dobrania odpowiedniego testu. Wiemy, że zmienne nie posiadają rozkładu normalnego. Bazując na rezultacie testu chi-kwadrat, nie możemy odrzucić założenia wskazującego na niezależność zmiennych. Dzięki temu testem, który będzie spełniał wszelkie wymagania będzie nieparametryczny test Wilcoxona. Rezultaty przeprowadzonego testu zamieszczone są w tabeli tab. 5.1.

Tab. 5.1. Wyniki niesparowanego testu Wilcoxona

Statystyka testowa W	p-wartość
1267	0,683

Wyniki testu Wilcoxona wskazują, że nie ma podstaw do tego, aby twierdzić, że inflacja jest statystycznie niższa w strefie euro niż poza nią, gdyż p-wartość (0,683) jest znacznie większa od przyjętego poziomu istotności α (0,05), przez co przyjęcie hipotezy alternatywnej jest niemożliwe.

5.3. Sparowany test Wilcoxona

Analizując wykres przedstawiający średni poziom inflacji w strefie euro oraz w krajach UE poza strefą euro (rys. 2.1), a także po przeprowadzeniu odpowiednich obliczeń możemy stwierdzić, że poziomy inflacji w obu strefach są ze sobą silnie skorelowane, co świadczy o ich silnym związku, czego nie moglibyśmy się spodziewać jedynie po przeprowadzeniu testu chi-kwadrat. Dodatkowo, ze względu na charakter danych możemy dostrzec, że dane są ze sobą sparowane, gdyż w każdej chwili czasowej, posiadamy informacje o poziomie zarówno w strefie euro, jak i w krajach UE, które do niej nie należą. Mając to na uwadze możemy przeprowadzić sparowany test Wilcoxona, który podobnie jak test niesparowany jest testem nieparametrycznym, jednakże ten skupia się na porównaniu pary wartości z każdej strefy. Wyniki przeprowadzonego testu zawarte są w poniższej tabeli (tab. 5.2).

Tab. 5.2. Wyniki sparowanego testu Wilcoxona

Statystyka testowa V	p-wartość
648	0,6387

Pomimo uwzględnienia informacji o sparowaniu danych rezultat nie odbiega znacząco od tego, który uzyskaliśmy w przypadku niesparowanej wersji testu Wilcoxona. Podobnie jak wcześniej, ze względu na dużą p-wartość, w porównaniu do poziomu istotności α, nie możemy przyjąć hipotezy alternatywnej H_1, przez co utwierdzamy się w przekonaniu, że inflacja w krajach strefy euro wcale nie jest niższa niż w przypadku drugiej grupy krajów Unii Europejskiej. Należy jednak podkreślić, że osiągnięty wynik mógłby być inny w przypadku zmiany badanego zakresu, jednakże doświadczenia, które zostały przeprowadzone w ramach niniejszego raportu, stanowią solidną podstawę do przyjęcia wydedukowanego stanowiska.

Część II

Problem 2

6. Wstęp

Celem problemu 2 jest określenie pewnych informacji o próbie losowej będącej wynikiem przeprowadzonego eksperymentu losowego polegającego na dwudziestokrotnym rzucie wybraną monetą. Wyniki kolejnych, niezależnych rzutów zostały zamodelowane zmiennymi losowymi, to znaczy w przypadku wypadnięcia reszki zmienna przyjmuje wartość 0, z kolei dla orła przyjmuje wartość 1.

Na podstawie tak zdefiniowanej próby losowej wyznaczono:

* estymator punktowy prawdopodobieństwa wypadnięcia orła,
* przedział ufności dla wybranego poziomu ufności,
* dodatkową "wiedzę ekspercką" wykorzystując prawdopodobieństwo Bayesa.

7. Eksperyment - wyniki

Pierwszym etapem tej części projektu było przeprowadzenie opisanego we wstępie eksperytmentu, którego wyniki przedstawiono w tabeli 7.1. Uzyskany w ten sposób ciąg zmiennych losowych stanowi próbę losową składającą się z dwudziestu niezależnych od siebie elementów o tym samym rozkładzie.

Tab. 7.1. Uzyskane wyniki eksperymentu rzutu monetą (reszka = 0, orzeł = 1)

Numer rzutu	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Wynik rzutu	0	0	1	1	0	1	0	1	1	1	0	1	0	1	1	1	0	0	1	0

Rozkładem który dobrze opisuje taką przeliczalną próbę, złożoną z dyskretnych zmiennych losowych, jest dwumianowy rozkład Bernoulliego z jednym parametrem p, będącym wartością oczekiwaną dla pojedynczego zdarzenia. Parametr ten stanowi prawdopodobieństwo wylosowania zmiennej o wartości 1, która w odniesieniu do próby oznacza wyrzucenie orła. Przyjmując, że sukcesem jest wylosowanie zmiennej o wartości 1 to w eksperymentalnej próbie wystąpiło 11 sukcesów (orłów) na 20 wszystkich elementów (rzutów).

Dla rozkładu Bernoulliego (gdzie p należy do przedziału [0,1]) funkcja prawdopodobieństwa wylosowania orła tzn. gdy zmienna losowa jest równa 1 określona jest zależnością:

$$f_p(1) = p \tag{7.1}$$

i dla zdarzenia przeciwnego gdy zmienna losowa jest równa 0:

$$f_p(0) = 1 - p \tag{7.2}$$

8. Estymacja punktowa

Analizę uzyskanych wyników rozpoczęto od wyznaczenia estymatora punktowego prawdopodobieństwa wypadnięcia orła, czyli wskazania jednego punktu (jednej wartości).

Estymacja polega na próbie wyznaczenia wartości nieznanego parametru na podstawie próby losowej. Estymator jest funkcją od próby losowej, której używamy do oszacowania (przybliżenia) parametru populacji.

8.1. Metoda największej wiarygodności

Metodą jaką w tym przypadku wyznaczony został estymator jest metoda największej wiarygodności. Badaną przez nas cechą próby jest prawdopodobieństwo wypadnięcia orła. Ze względu na ten parametr maksymalizujemy funkcję wiarygodności określoną wzorem:

$$\mathrm{L_n}(p) = f_p(X_1) \cdot ... \cdot f_p(X_n) \tag{8.1}$$

Funkcja ta jest iloczynem funkcji rozkładu prawdopodobieństwa przeprowadzonej próby losowej, a wartość parametru, dla której uzyskano największy wynik jest wartością estymatora.

W związku z tym, że funkcje prawdopodobieństwa zależą nie tylko od parametru, ale także od próby losowej, to mimo tego samego rozkładu każdej z funkcji należy wyznaczyć ich wartość w każdym możliwym miejscu. Po określeniu takich założeń dla analizowanej próby funkcja wiarygodności prezentuje się następująco:

$$\mathrm{L_n}(p) = (1-p)^{n-(X_1+...+X_n)} \cdot p^{X_1+...+X_n} \tag{8.2}$$

Takie potęgi przy funkcjach prawdopodobieństw w równaniu 8.2 wynikają z faktu, iż sumując kolejne elementy naszej próby uzyskamy liczbę sukcesów czyli ilość wylosowanych orłów. Odejmując natomiast ten wynik od sumy wszystkich elementów w zbiorze otrzymamy liczbę wyrzuconych w eksperymencie reszek. W ten sposób dla analizowanej próby uzależniliśmy funkcję wiarygodności jednocześnie od parametru p oraz próby.

Kolejnym krokiem jest maksymalizowanie funkcji z równania 8.2 ze względu na prawdopodobieństwa wypadnięcia orła (parametr p). Dla uproszczenia obliczeń funkcję wiarygodności logarytmujemy:

$$\mathrm{l_n}(p) = (X_1 + ... + X_n) \cdot \ln p + (n - (X_1 + ... + X_n)) \cdot \ln(1-p) \tag{8.3}$$

Dzięki temu przekształceniu otrzymaliśmy logarytmiczną funkcję wiarygodności, która składa się z sumy dwóch składników, a takie wyrażenie łatwiej zmaksymalizować. Funkcja wiarygodności jest nieujemna i zerowa na krańcach (po podstawieniu $p = 0$ lub $p = 1$), oznacza to że maksimum funkcji znajduje się gdzieś pomiędzy granicami parametru p (p należy do przedziału: [0,1]). Do wyznaczenia największej wartości konieczne jest wyznaczenie pochodnej I rzędu funkcji z równania 8.3, przyrównanie jej do zera i w kolejnym kroku wyznaczenie estymatora parametru p:

$$\frac{\partial l_n(p)}{\partial p} = \frac{X_1 + ... + X_n}{p} - \frac{n - (X_1 + ... + X_n)}{1-p} \tag{8.4}$$

$$\frac{X_1 + ... + X_n}{\hat{p}} - \frac{n - (X_1 + ... + X_n)}{1-\hat{p}} = 0 \tag{8.5}$$

Rozwiązując równanie 8.5 otrzymujemy wzór na estymator parametru p, który jest jednocześnie średnią z próby:

$$\hat{p}_n = \frac{X_1 + ... + X_n}{n} = \overline{X}_n = m_1 \tag{8.6}$$

Dokładnie taki sam estymator uzyskalibyśmy wyznaczając go metodą momentów. Obliczony metodą największej wiarygodności estymator jest zgodny oraz jest pierwszym momentem z próby.

8.2. Wyznaczanie wartości estymatora największej wiarygodności dla wygenerowanych danych

Po podstawieniu do równania 8.6 wartości z naszej eksperymentalnej próby rzutu monetą uzyskujemy wartość estymatora parametru p równą:

$$\hat{p}_n = \frac{0+0+1+1+0+1+0+1+1+1+0+1+0+1+1+1+0+0+1+0}{20} = \frac{11}{20} = 0,55 \tag{8.7}$$

8.3. Wnioski

Wyznaczony w równaniu 8.7 estymator parametru p, to znaczy prawdopodobieństwa wyrzucenia orła badaną przez nas monetą, jest równy 0,55. Oszacowany w ten sposób wynik różni się od spodziewanej wartości 0,5, jednak nie oznacza to, że moneta jest "oszukana". Estymator ten jest asymptotycznie nieobciążony oznacza to, że jego wartość oczekiwana wraz ze wzrostem próby będzie coraz bardziej zbiegała do wartości prawdziwej. Nasza badana próba losowa nie jest wystarczająco liczną próbą, żeby stwierdzić czy wykorzystywana w badaniu moneta jest oszukana, czy jak należy przypuszczać sprawiedliwa (tzn. $p = 0,5$). W związku z tym zauważamy, że dokładność wyznaczonego w tym zadaniu estymatora zależy od próby losowej oraz jej wielkości.

9. Przedział ufności

Ta część projektu polegała na skonstruowaniu przedziału ufności dla danego poziomu ufności równego 0,95. Przedział ufności ma na celu "złapać" nieznany parametr z zadanym prawdopodobieństwem (poziomem ufności) i określony jest nierównością:

$$P(\theta \in [A(X_1, ..., X_n), B(X_1, ..., X_n)]) \geqslant \gamma \tag{9.1}$$

gdzie:

A - dolna granica przedziału ufności

B - górna granica przedziału ufności

θ - parametr

γ - poziom ufności

9.1. Bootstrap parametryczny

W związku z faktem, że posiadamy wiedzę na temat rozkładu badanej próby (rozkład Bernoulliego) oraz wartości estymatora parametru p ($\hat{p} = 0{,}55$ wyliczonego w równaniu 8.7), to do wyznaczenia przedziału ufności zdecydowano się na wykorzystanie techniki "bootstrapingu". W tym przypadku metoda ta polega na odtworzeniu badanej próby poprzez wygenerowanie 1000-ca pakietów, każdy o wielkości 20 elementów (tyle samo co badana próba) wylosowanych z rozkładu Bernoulliego o parametrze $\hat{p}$ równym 0,55. Z każdego z tych pakietów wyznaczone zostały nowe, kolejne estymatory parametru p. Wynik losowania wraz z wygenerowanym dla niego rozkładem normalnym przedstawiony jest na histogramie z rysunku 9.1.

Następnie przy pomocy posiadanej wiedzy na temat realizacji badanej zmiennej losowej wyznaczany jest przedział ufności, będący kwantylem rzędu 2,5% (dolna granica) oraz 97,5% (górna granica).

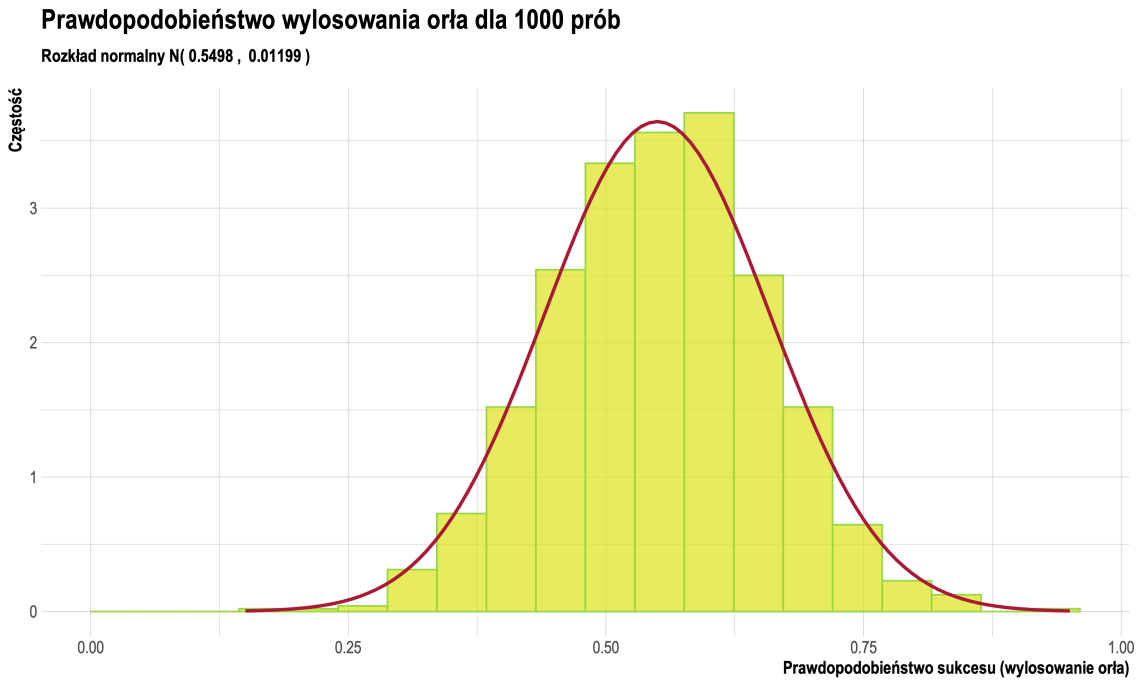
Z przeprowadzonych symulacji wynika, że przedział ufności dla zadanego poziomu ufności γ równego 0,95 wynosi:

$$P(\theta \in [0,35; 0,75]) \geqslant 0,95 \tag{9.2}$$

Wyznaczony w ten sposób przedział ufności stworzony z wykorzystaniem informacji o rozkładzie próby oraz bazujący na obliczonym metodą największej wiarygodności estymatorze zawiera wartość $\frac{1}{2}$.

W tym momencie posiadając informacje o rozkładzie prawdopodobieństwa wylosowania orła w pojedynczym rzucie (rys. 9.1) zauważamy, że rozkład dwumianowy wraz ze wzrostem wielkości próby eksperymentu zbiega do rozkładu normalnego. Oznacza to, że teoretycznie w eksperymencie rzutu monetą, mogą nam się zdarzyć próby w których na 20 rzutów będziemy mieli 20 sukcesów, jednak prawdopodobieństwo takiego zdarzenia jest bardzo małe. Zgadza się to z centralnym twierdzeniem granicznym. Twierdzenie to informuje nas, że suma dużej liczby niezależnych zmiennych losowych o identycznym rozkładzie będzie w przybliżeniu mieć rozkład normalny.

Widać to w tabeli 7.1. Wraz z kolejnymi rzutami monetą ilość 0 i 1 w próbie coraz bardziej zbiega to równej wartości.



Rys. 9.1. Histogram prawdopodobieństwa wylosowania orła w jednokrotnym rzucie (parametr p) wyznaczony z 20 elementowej próby, techniką bootstrapową dla 1000 prób.

Dlatego, gdy wielkość próby eksperymentu dwumianowego jest duża, a prawdopodobieństwo sukcesu jest bliskie 0,5, często aproksymuje się rozkład dwumianowy rozkładem normalnym. Pozwala to na wykorzystanie metod opartych na rozkładzie normalnym do analizy danych i wnioskowania o populacji, z której pobrano próbę.

10. Prawdopodobieństwo Bayesa

W ostatniej części zadania należało przy wykorzystaniu prawdopodobieństwa Bayesa wyznaczyć dodatkową wiedzę na temat badanej próby.

Przed przystąpieniem do eksperymentu lub chwilę po jego przeprowadzeniu posiadamy pewne subiektywne wyobrażenie na przykład na temat wykorzystywanej monety. Taką ogólną wiedzę, która nie musi mieć potwierdzenia w próbie możemy uwzględnić przy wyciąganiu wniosków o monecie.

Aby tego dokonać musimy wyróżnić dwa rozkłady:

* *a priori* - to wyobrażenie o parametrze/próbie przed przeanalizowaniem danych,
* *a posteriori* - wynikający z wyobrażenia (*a priori*) oraz analizy próby.

Rozkład *a posteriori* konstruuje się z rozkładu *a priori* zgodnie z regułą Bayesa. Funkcja rozkładu *a posteriori* $f_{post}(p, x)$ zależy od wartości parametru oraz od próby losowej i opisana jest zależnością:

$$f_{post}(p, x) = \frac{f_p(x) \cdot f_{prior}(p)}{\int\limits_{\theta} (f_p(x) \cdot f_{prior}(p))\, dp} \tag{10.1}$$

gdzie:

$f_p(x)$ - funkcja wiarygodności (wzór 8.1)

$f_{prior}(p)$ - funkcja *a priori*

Równanie 10.1 to iloczyn funkcji wiarygodności i funkcji gęstości. Oznacza to, że wynikiem takiego równania jest pojedyncza wartość. W związku z tym, że rozkład *a posteriori* musi się całkować do wartości 1 to mianownik musi być odwrotnością licznika co znacząco upraszcza obliczenia.

Podejście *a posteriori* uwzględnia naszą wiedzę/doświadczenie na temat takiej próby, które w przypadku błędnego osądu po przeprowadzeniu analizy lub dopływie nowych danych zostanie naprostowane.