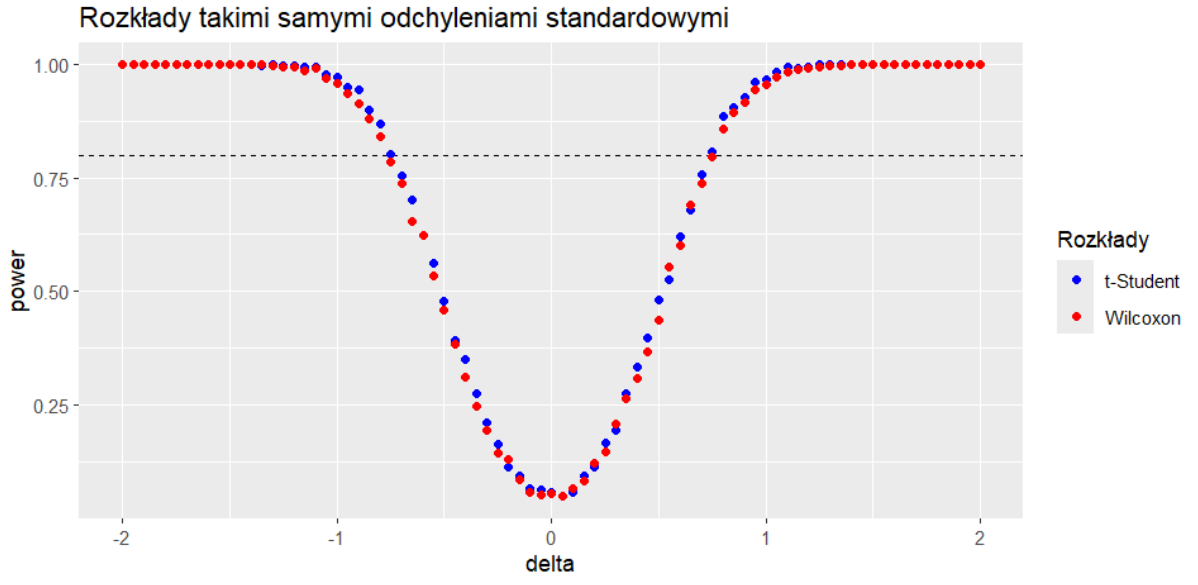
Statystyka w Analizie Danych 2024L

Projekt 2

Jakub Śliwa (335209), Jakub Smela (310900)

Problem 1

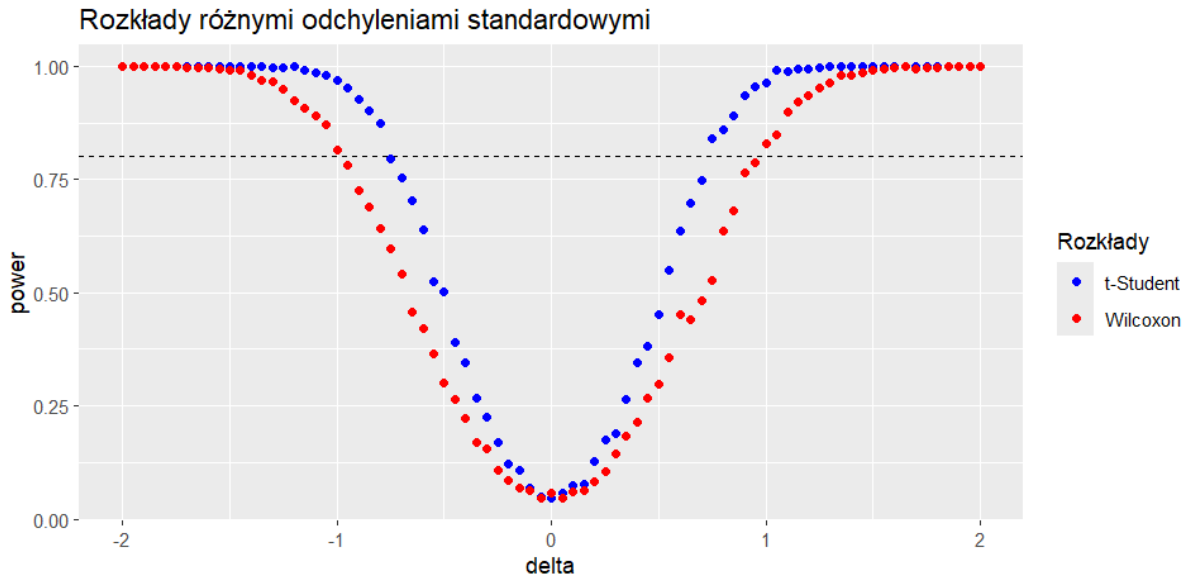
W celu znalezienia różnic pomiędzy dwoma rozkładami danych, zakładając, że mają one rozkłady normalne został wykorzystany test t-Studenta. W drugim przypadku wykonaliśmy test Wilcoxona. W obu sytuacjach przygotowana została lista przesunięć Δ drugiego rozkładu, zawierająca się w przedziale [-2: 2], z wartościami rosnącymi o 0.05. Jako odchylenie standardowe przyjęliśmy wartość 1, dla każdego rozkładu generowaliśmy 30 punktów. Dla każdej wartości Δ wykonaliśmy 1000 testów. Poziom istotności ustaliliśmy jako 0.05 a graniczną moc testu jako 0.8. Moc testu jest liczona jako prawdopodobieństwo odrzucenia hipotezy zerowej, gdy jest ona fałszywa. Dla każdego tysiąca eksperymentów dla danej wartości Δ obliczane jest jaka cześć otrzymanych p-wartości jest mniejsza niż alfa.



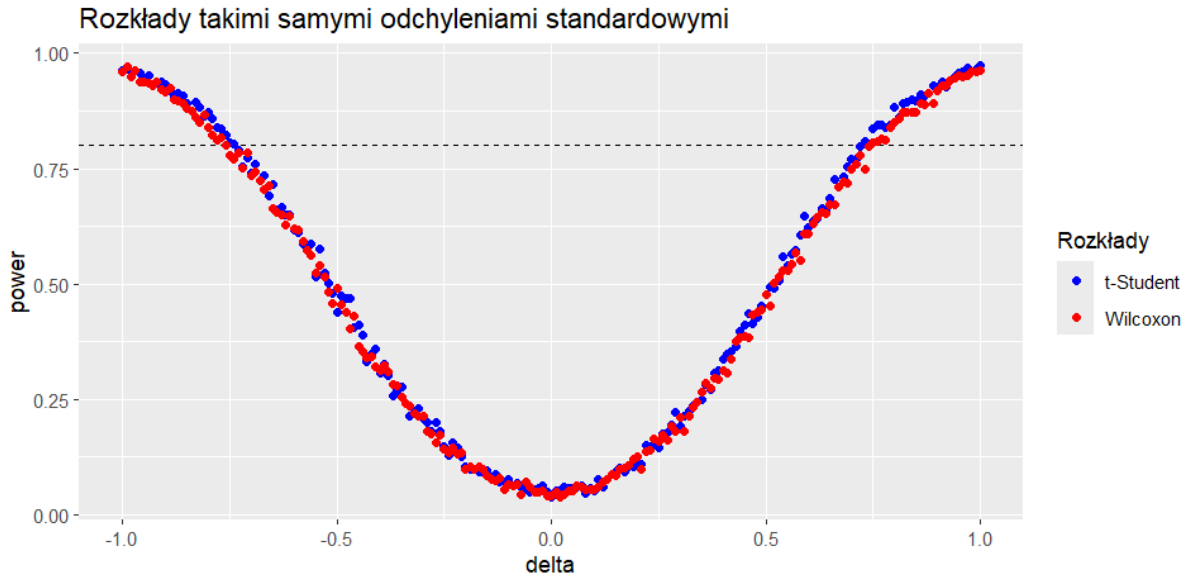
Minimalna wartość bezwzględna Δ w teście t-Studenta, dla której przesunięcie zostało wykryte z mocą co najmniej 0.8 to 0.75, a w teście Wilcoxona 0.8.

Zastosowanie różnych wartości odchylenia standardowego dla rozkładów, rozkład1: 1, rozkład2: 1.5

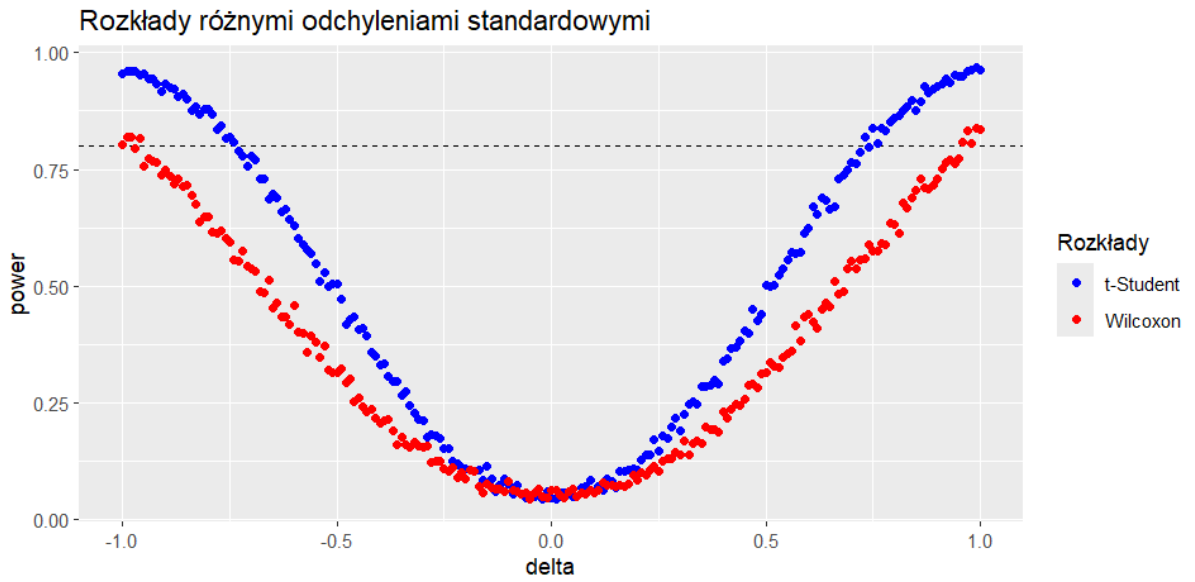
Minimalna wykryta wartość Δ w teście t-Studenta to 0.75, w teście Wilcoxona 0.95.



Po powtórzeniu testów dla przesunięć Δ zawierająca się w przedziale [-1: 1], z wartościami rosnącymi o 0.01 oraz takimi samymi odchyleniami standardowymi, wartość bezwzględna Δ dla której przesunięcie zostało wykryte z mocą co najmniej 0.8 to dla testu t-Studenta to 0.73 a dla testu Wilcoxona 0.75.



Po zastosowaniu odchyleń standardowych równych 1 oraz 1.5, Minimalna wykryta wartość Δ w teście t-Studenta to 0.74, w teście Wilcoxona 0.96.

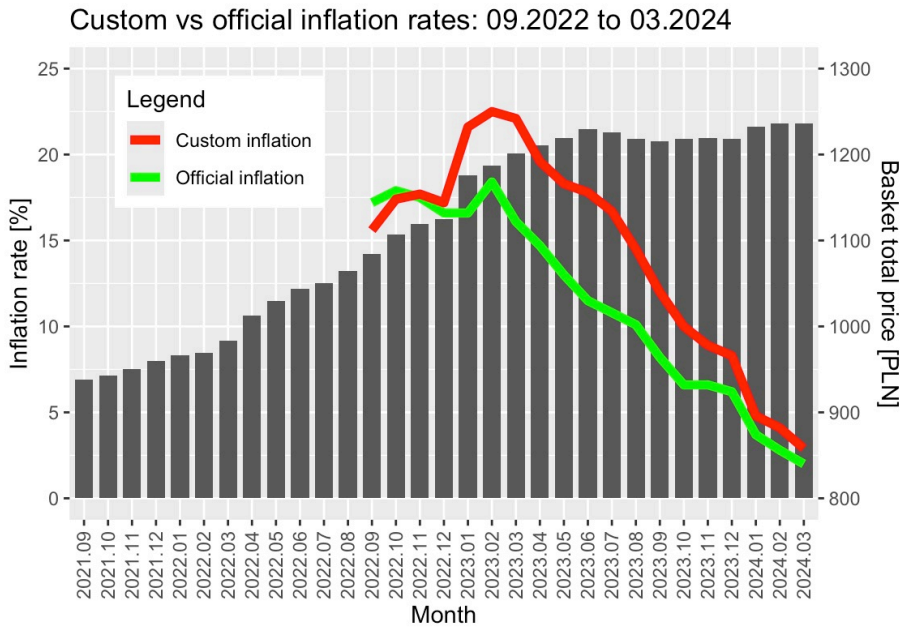


Wnioski

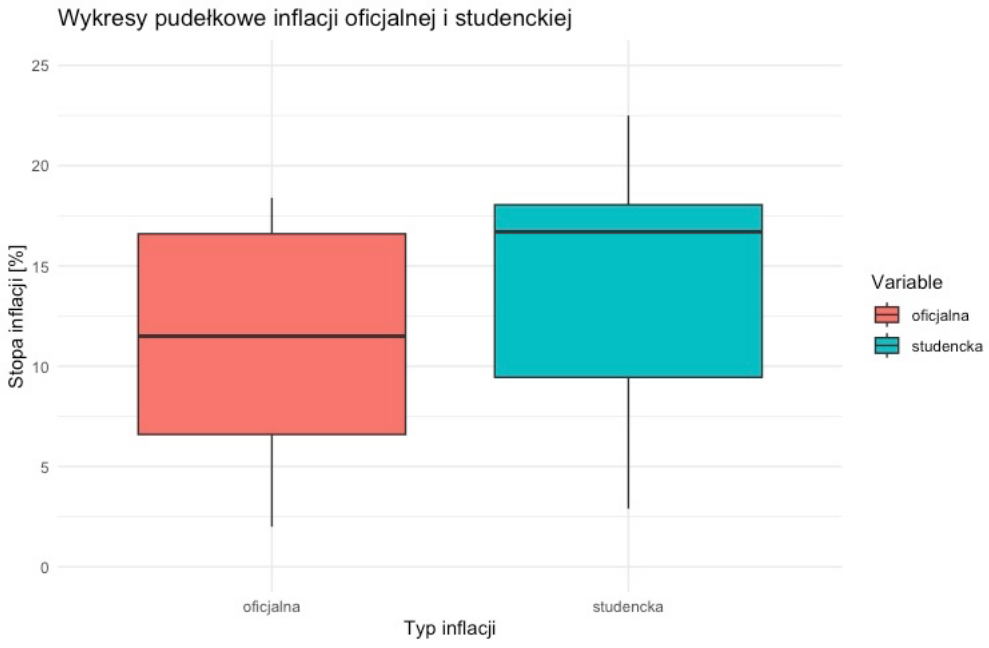
W przypadku testu zakładającego normalność prób losowych, czyli teście t-Studenta, rozbieżności pomiędzy rozkładami wykrywane są przy mniejszych przesunięciach wartości oczekiwanej. Test t-Studenta ma większą moc dla analogicznych rozbieżności pomiędzy wartościami oczekiwanymi rozkładów w porównaniu do testu Wilcoxona. Gdyby dane nie pochodziły z rozkładów normalnych, test t-Studenta mógłby dawać gorsze wyniki. Test Wilcoxona jest odporny na odstępstwa od rozkładów normalnych. Test t-Studenta znacznie lepiej wykrywa różnice w rozkładach, gdy ich odchylenia standardowe różnią się.

Problem 2

Rozwiązanie tego problemu rozpoczęliśmy od wczytania danych dotyczących studenckiej oraz oficjalnej inflacji. Korzystaliśmy z wyników pierwszego projektu Jakuba Śliwy, toteż dane dotyczące inflacji dostępne były dla każdego miesiąca od września 2022 do marca 2024 w odniesieniu do tego samego miesiąca poprzedniego roku (a zatem jedynie 19 obserwacji). Poniżej zamieszczamy wykres przypominający, jak wyglądała wyliczona studencka inflacja.

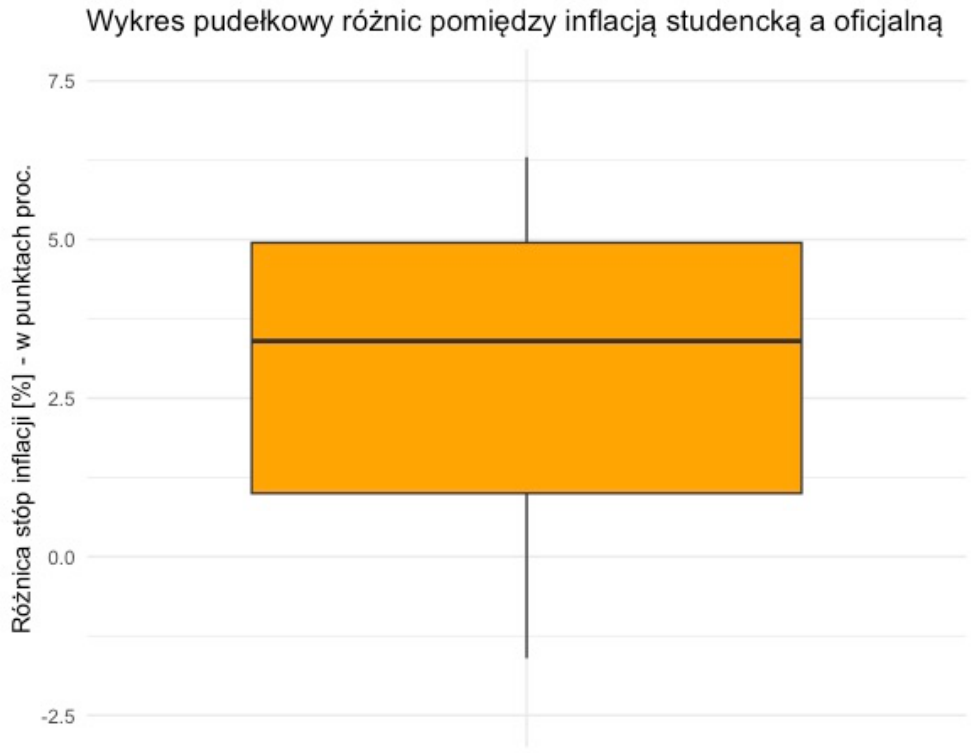


Po zaimportowaniu danych oraz drobnych operacjach modyfikacyjnych, udało nam się narysować wykresy pudełkowe dla obu typów inflacji.



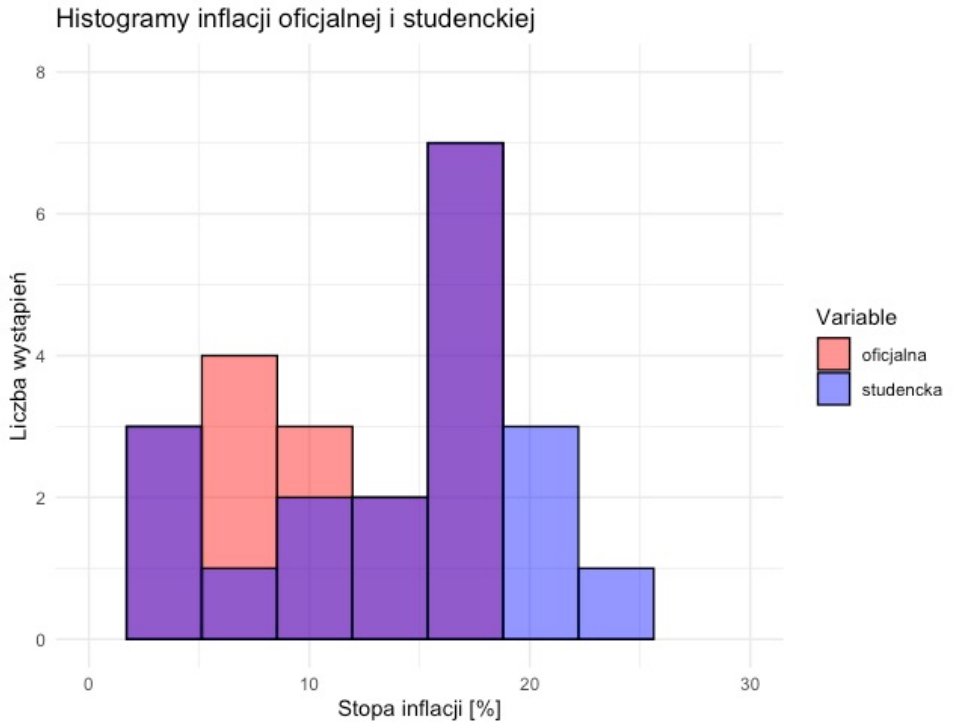
Na podstawie powyższego wykresu można poczynić kilka obserwacji. Po pierwsze, mediana dla studenckiej inflacji znajduje się bardzo blisko jej trzeciego kwartyla, co znaczy, że czwarta część obserwacji dla inflacji studenckiej wpada w bardzo wąski przedział wartości 17 – 18%. Jest ona również położona istotnie wyżej niż mediana inflacji oficjalnej (17% vs 12%), co może sugerować, że inflacja studencka i oficjalna mają istotnie różniące się rozkłady. Po drugie, maksymalna wartość oficjalnej inflacji oraz trzeci kwartyl inflacji studenckiej znajdują się na podobnym poziomie (ok. 18%), co oznacza, że czwarta część obserwacji studenckiej inflacji przyjmuje wartości większe niż największa wartość oficjalnej inflacji. Po trzecie, zakres przyjmowanych wartości, czyli różnica między wartością maksymalną a minimalną, jest istotnie wyższy dla inflacji studenckiej niż oficjalnej, na co wskazuje większa długość wąsów błękitnego pudełka, niźli czerwonego. Po czwarte, rozstęp międzykwartylowy inflacji studenckiej, a więc wielkość błękitnego pudełka, zdaje się być mniejszy niż w przypadku oficjalnej inflacji, co sugeruje nieco większe skupienie danych dla inflacji wyliczonej na własnym koszyku dóbr. Podsumowując, porównanie wykresów pudełkowych daje solidne podstawy by sądzić, iż próby inflacji studenckiej oraz oficjalnej pochodzą z istotnie różniących się rozkładów.

W następnym kroku obliczyliśmy różnice wartości własnej inflacji i inflacji oficjalnej dla każdego miesiąca oraz narysowaliśmy wykres pudełkowy tak zdefiniowanej zmiennej.

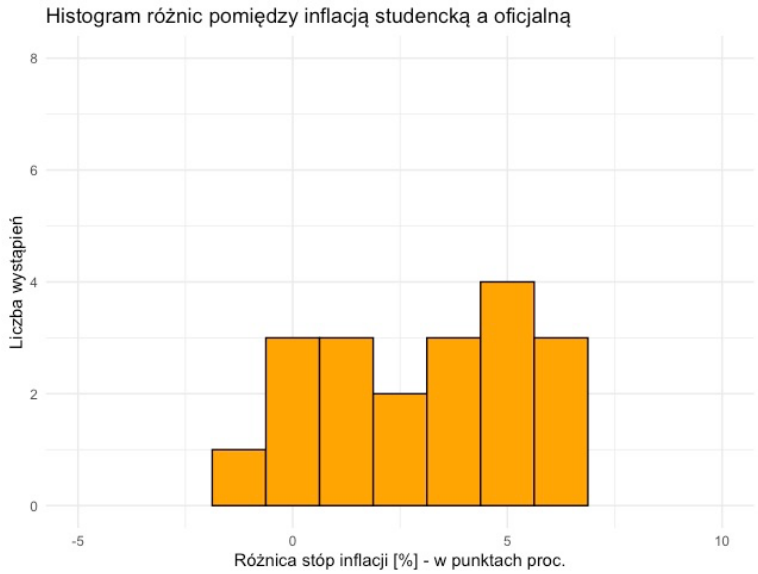


Powyższy wykres zdaje się potwierdzać wysunięte wcześniej przypuszczenia. Zauważmy, że wartość pierwszego kwartyla wynosi ok. 1 punktu procentowego, co oznacza, że w badanym okresie na każde rozpatrywane 4 miesiące, średnio w trzech z nich inflacja studencka była wyższa od oficjalnej o przynajmniej 1 punkt procentowy. Ponadto, mediana znajduje się istotnie bliżej trzeciego niż pierwszego kwartyla. Oznacza to, że większe zagęszczenie danych występuje w przedziale od 3.5 do 5 punktów procentowych, niźli w przedziale od 1 do 3.5 punktu procentowego. Wykres pudełkowy różnic pomiędzy inflacjami studencką a oficjalną wyraźnie nie jest równomiernie rozłożony wokół zera, co wzmacnia nasze przypuszczenia, że próby mogą pochodzić z różniących się rozkładów, gdzie wartość oczekiwana rozkładu inflacji studenckiej jest wyższa od odpowiadającej wartości rozkładu inflacji oficjalnej.

Na podstawie dotychczasowej analizy możemy sformułować hipotezę, że inflacja studencka jest wyższa od inflacji oficjalnej. Chcąc zweryfikować tę hipotezę za pomocą odpowiedniego testu statystycznego, pierwszym krokiem jest wybór pomiędzy testem parametrycznym a nieparametrycznym. Chcąc sprawdzić, czy założenie normalności rozkładu ma rację bytu, narysowaliśmy histogramy obu prób. Wykres poniżej zdaje się rozwiewać wszelkie wątpliwości co do zasadności założenia o gaussowości rozkładu – dla obu typów inflacji rozkład próby nie jest choćby zbliżony do kształtu krzywej gaussowskiej. Rezygnujemy zatem z wykonania testu parametrycznego na rzecz podejścia nieparametrycznego.



Kolejną istotną kwestią jest pytanie, czy test Wilcoxona należy wykonać w wersji dla par, czy też nie. Naszym zdaniem próby X i Y nie są niezależne, gdyż obserwacje kolejnych wartości wykonywane są dla danego miesiąca. Wydaje się zatem, że różnice pomiędzy wartościami dla każdej pary (inflacja studencka i oficjalna w danym miesiącu) nie mają rozkładu normalnego, co zdaje się potwierdzać narysowany histogram obliczonych wcześniej różnic.



Ostatecznie decydujemy się zatem na przeprowadzenie testu Wilcoxona dla par przy jednostronnej hipotezie alternatywnej mówiącej o tym, że różnica pomiędzy inflacją studencką a oficjalną jest większa od zera. Wyniki przeprowadzonego testu obrazuje poniższy zrzut ekranu.

```
> # przeprowadzenie testu Wilcoxona (dla par)
> wilcox.test(X, Y, paired=TRUE, alternative="greater")

        Wilcoxon signed rank exact test

data:  X and Y
V = 181, p-value = 6.294e-05
alternative hypothesis: true location shift is greater than 0
```

Wyznaczona *p-value* jest bardzo niska, toteż na każdym racjonalnym poziomie istotności (e.g. 0.001, 0.01, 0.05) odrzucamy hipotezę zerową na rzeczy hipotezy alternatywnej. Oznacza to, że inflacja studencka jest istotnie wyższa od oficjalnej. Uzyskany wynik zdaje się być zgodny z intuicją, gdyż inflacja zazwyczaj relatywnie bardziej dotyka uboższą część społeczeństwa, do której należy większa część studentów. W skład własnego koszyka dóbr wchodziły przede wszystkim produkty spożywcze, proste usługi (np. fryzjer) oraz opłaty za media (energia, gaz, ciepła woda), które to w okresie wzmożonej inflacji podrożały najbardziej.

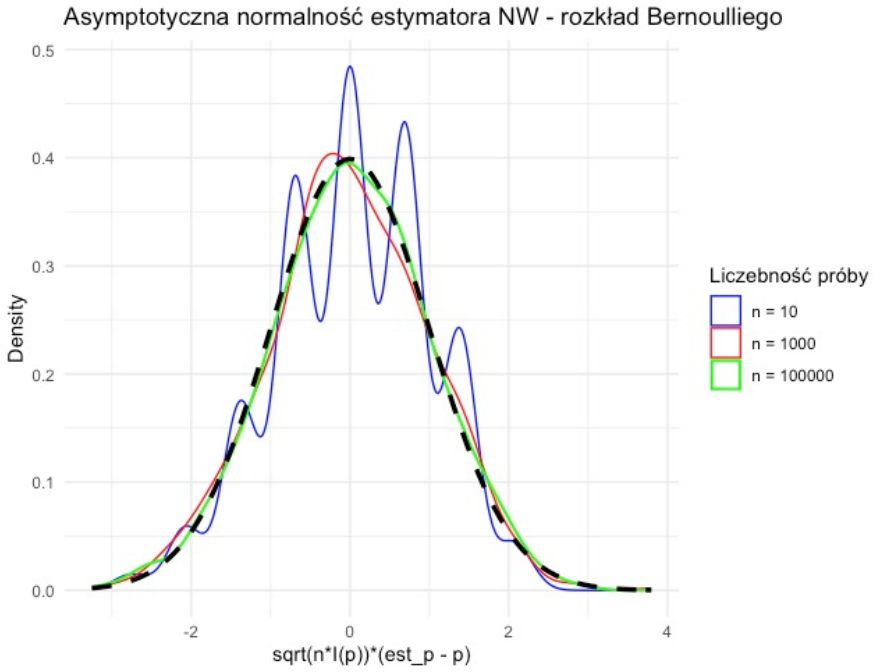
Problem 3

Asymptotyczna normalność estymatorów największej wiarygodności polega na tym, że wraz ze wzrostem rozmiaru próby *n* (do nieskończoności), rozkład estymatora MNW zbiega do rozkładu normalnego (staje się do niego coraz bardziej podobny). Dzieje się to niezależnie od pierwotnego rozkładu, z którego pochodzą dane. Definicję asymptotycznej normalności estymatora MNW można zapisać następująco:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I(\theta)}\right) \text{ gdy } n \to \infty$$

Po przeniesieniu Informacji Fishera *thety* pod pierwiastek widzimy, że znormalizowane wartości różnic pomiędzy estymowanymi MNW wartościami parametru *theta*, a jego rzeczywistą nieznaną wartością w populacji zbiegają do standardowego rozkładu normalnego.

Przedstawienia asymptotycznej normalności estymatorów MNW dokonamy na przykładzie rozkładu Bernoulliego z wartością parametru *p* równą 0.7. Dla różnych liczebności próby – 10, 1 000 i 100 000 wylosowaliśmy po 1000 prób, na podstawie których dokonano estymacji parametru *p* estymatorem MNW, a więc średnią arytmetyczną z próby. Następnie obliczono wartość Informacji Fishera dla parametru *p* jako $\frac{1}{p(1-p)}$, wyznaczono różnice między oszacowaniami zwracanymi przez estymator MNW a rzeczywistą wartością parametru *p*, znormalizowano je poprzez ich pomnożenie przez pierwiastek kwadratowy z iloczynu *n* i Informacji Fishera. Rozkłady tak uzyskanych zmiennych przedstawiono na wykresie poniżej.



Na przedstawionym wyżej wykresie przerywaną czarną linią przedstawiono funkcję gęstości standardowego rozkładu normalnego, do którego to wraz ze wzrostem liczebności próby *n* zbiega rozkład estymatora MNW. Zauważmy, że dla liczebności próby równej 10 (niebieska linia), wykres znacząco odbiega od rozkładu normalnego. W tym przypadku oszacowania estymatora przyjmują tak naprawdę dyskretne wartości ze zbioru {0.0, 0.1, 0.2, …, 0.9, 1.0}, lecz na wykresie w celu zwiększenia czytelności zastosowano wygładzanie. Dla liczebności próby równej 1 000 (czerwona linia) rozkład estymatora MNW znacznie bardziej przypomina już rozkład normalny, lecz wciąż można wskazać miejsca, gdzie odbiega on od czarnej przerywanej linii. Zadanie to staje się bardzo trudne w przypadku zielonej linii (liczebność próby 100 000), która to praktycznie pokrywa się z krzywą funkcji gęstości standardowego rozkładu normalnego. Powyższy wykres demonstruje zatem, że wraz ze wzrostem liczebności próby rozkład estymatora MNW zbiega do rozkładu normalnego, co właśnie mamy na myśli mówiąc o asymptotycznej normalności estymatora MNW.