Analiza szeregów czasowych - raport 1.

Miłosz Kubera(249823)
Patryk Krukowski(249824)

31 marca 2021

Spis treści

1 Wstęp i opis eksperymentów

W tym sprawozdaniu zajmiemy się symulacyjną analizą własności rozkładów asymptotycznych estymatorów średniej, autokowariancji i autokorelacji (w pierwszej części sprawozdania) oraz testowaniem białoszumowości (w drugiej części sprawozdania). Wnioski i spostrzeżenia będą zamieszczane na bieżąco w sprawozdaniu, pełniąc jednocześnie rolę podsumowania. Przedstawiamy spis najważniejszych użytych w analizie narzędzi oraz eksperymentów:

- Histogramy
- Wykresy estymatorów jądrowych, dystrybuant empirycznych
- Testy zgodności Kołmogorowa oraz Shapiro-Wilka, służące do badania rozkładu normalnego w próbie
- Testy białoszumowości *Ljungi-Boxa* oraz *Boxa-Pierce'a*
- Analiza funkcji autokorelacji

2 Symulacyjna analiza własności rozkładów asymptotycznych estymatorów średniej, autokowariancji i autokorelacji

Rozważmy trzy następujące estymatory parametrów rozkładu stacjonarnego szeregu czasowego drugiego rzędu:

- Estymator próbkowy wartości oczekiwanej μ (średnią próbkową)

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

- Estymator funkcji autokowariancji $\gamma(h)$

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h} \left(X_{t+h} - \overline{X}_n\right)\left(X_t - \overline{X}_n\right)$$

  dla $h = 0, 1, \ldots, n-1$

- Estymator funkcji autokorelacji $\rho\,(h)$

$$\hat{\rho}\,(h) = \frac{\hat{\gamma}(h)}{\hat{\sigma}^2}$$

  dla $h = 0, 1, \ldots, n-1$

2.1 Symulacyjna analiza rozkładów asymptotycznych przy użyciu metod graficznych

Wykonajmy 80 realizacji szeregu czasowego typu biały szum, przy czym musimy oczywiście uwzględnić fakt, że jeśli losujemy obserwacje z rozkładu wykładniczego (lub ogólnie rozkładu o niezerowej średniej), to od każdej realizacji wektora obserwacji wpisanego do odpowiedniej kolumny macierzy, musimy odjąć średnią tego wektora.

Dla ustalenia uwagi:

- Jeśli rozważamy rozkład normalny $N(\mu, \sigma)$, to σ oznacza **wariancję**.
- Jeśli rozważamy rozkład wykładniczy $Exp\,(\lambda)$, to parametryzacja λ jest taka, że $\mathbb{E}X = \frac{1}{\lambda}$, gdzie X to zmienna losowa ze wspomnianego rozkładu wykładniczego.

1. Szereg ma długość 100 oraz obserwacje pochodzą z rozkładu $N(0, 1)$.

```
k <- 80
n_1 <- 100 #d■ługość szeregu
realizacje_1 <- matrix(rnorm(n_1*k),n_1,k) #losujemy realizacje do macierzy
```

2. Szereg ma długość 200 oraz obserwacje pochodzą z rozkładu $N(0, 1)$.

```
n_2 <- 200
realizacje_2 <- matrix(rnorm(n_2*k),n_2,k)
```

3. Szereg ma długość 50 oraz obserwacje pochodzą z rozkładu $Exp(5)$.

```
n_3 <- 50
realizacje_3 <- matrix(rexp(n_3*k,rate=5), n_3, k)
realizacje_3 <- realizacje_3-colMeans(realizacje_3)
```

4. Szereg ma długość 150 oraz obserwacje pochodzą z rozkładu $Exp(5)$.

```
n_4 <- 150
realizacje_4 <- matrix(rexp(n_4*k,rate=5), n_4,k)
realizacje_4 <- realizacje_4-colMeans(realizacje_4)
```

5. Szereg ma długość 300 oraz obserwacje pochodzą z rozkładu $Exp(5)$.

```
n_5 <- 300
realizacje_5 <- matrix(rexp(n_5*k, rate=5),n_5,k)
realizacje_5 <- realizacje_5-colMeans(realizacje_5)
```

Przejdźmy teraz do badania konkretnych estymatorów za pomocą wybranych narzędzi graficznych

2.1.1 Średnia próbkowa

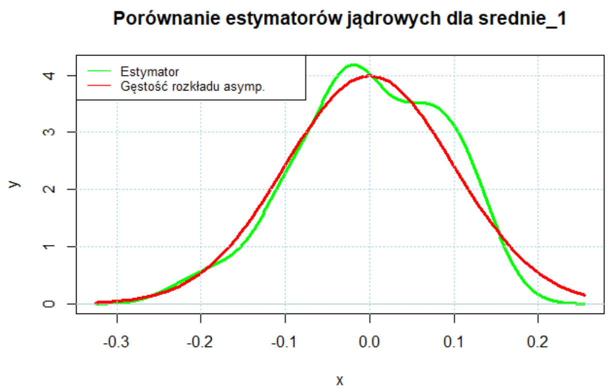
1. Szereg ma długość $n_1 = 100$ oraz obserwacje pochodzą z rozkładu $N(0,1)$.

```
#Wyznaczenie wartości estymatora wartości oczekiwanej - średnia próbkowa
srednie_1 <- apply(realizacje_1, MARGIN=2, FUN=mean)
#wgląd w wartości średniej próbkowej dla 10. pierwszych realizacji
head(srednie_1[1:10])

## [1]  0.15242679 -0.05042811 -0.08194214 -0.05439455 -0.11543385  0.02903591
```

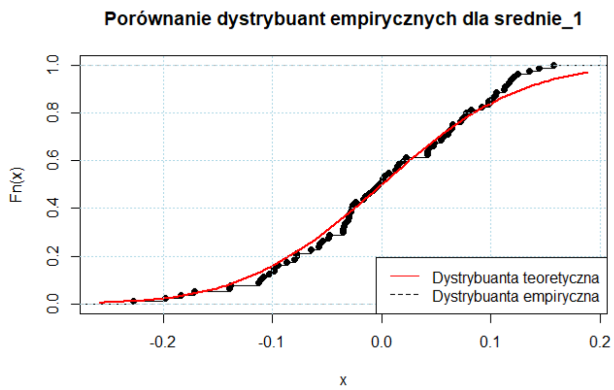
Rozważany rozkład teoretyczny, na mocy twierdzenia z wykładu, to dokładnie $N\left(0, \frac{1}{n_1}\right)$. Wykresy estymatora jądrowego i dystrybuanty empirycznej na rysunkach (1) i (2) mówią nam o tym, że rozkład empiryczny dobrze przybliża rozkład teoretyczny w sensie słabej zbieżności.

2. Szereg ma długość $n_2 = 200$ oraz obserwacje pochodzą z rozkładu $N(0,1)$.

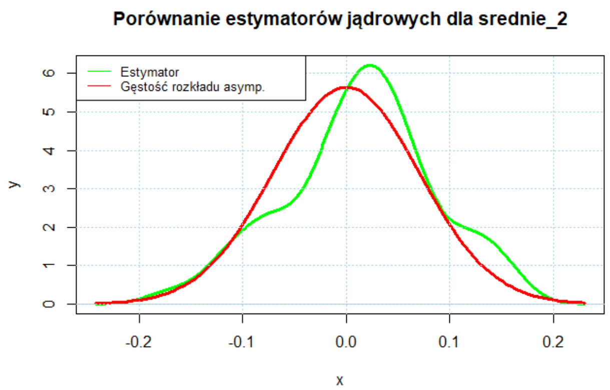
```
srednie_2 <- apply(realizacje_2, MARGIN=2, FUN=mean)
head(srednie_2[1:10])

## [1] -0.032864650  0.070213567 -0.004588498  0.081202483 -0.078137113
## [6]  0.032465306
```



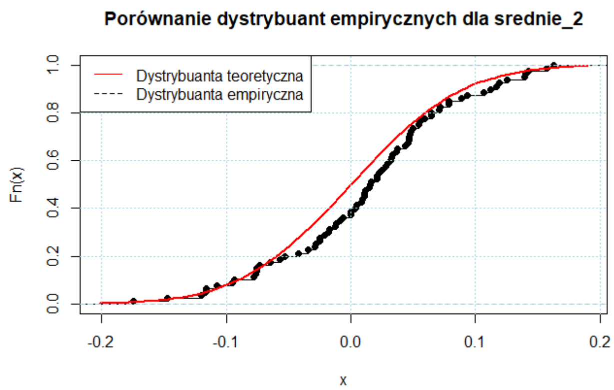
Rysunek 1: Porównanie estymatora jądrowego dla *srednie_1* oraz gęstości rozkładu $N\left(0, \frac{1}{n_1}\right)$



Rysunek 2: Porównanie dystrybuanty empirycznej dla *srednie_1* oraz teoretycznej rozkładu $N\left(0, \frac{1}{n_1}\right)$



Rysunek 3: Porównanie estymatora jądrowego dla *srednie_2* oraz gęstości rozkładu $N\left(0, \frac{1}{n_2}\right)$



Rysunek 4: Porównanie dystrybuanty empirycznej dla *srednie_2* oraz teoretycznej rozkładu $N\left(0, \frac{1}{n_2}\right)$

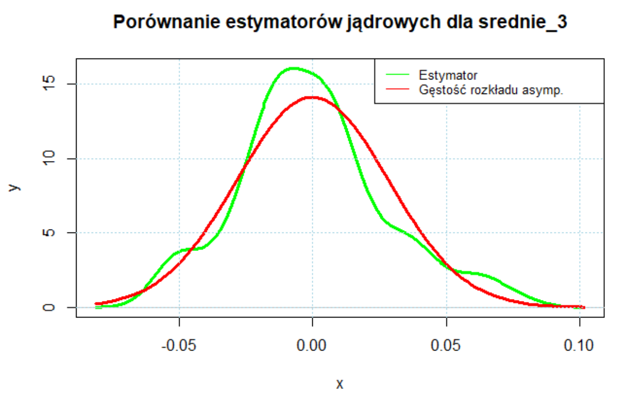
Rozważany rozkład teoretyczny, na mocy twierdzenia z wykładu, to dokładnie $N\left(0, \frac{1}{n_2}\right)$.

Wniosek 1 *Z porównania wykresów (1), (2), (3) i (4) wnosimy, że im długość szeregu czasowego jest większa, tym rozkład empiryczny lepiej przybliża rozkład teoretyczny. W szczególności, możemy zauważyć formowanie charakterystycznego "dzwonu" na rysunku (3). Oczywiście wniosek ten odnosi się jedynie (jeszcze) do rodziny rozkładów normalnych.*

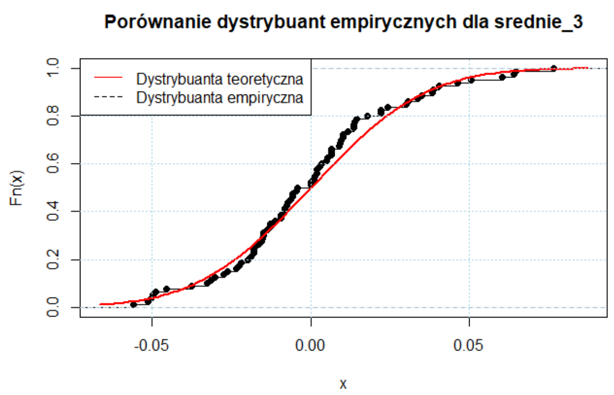
3. Szereg ma długość $n_3 = 50$ oraz obserwacje pochodzą z rozkładu $Exp(5)$.

```
srednie_3 <- apply(realizacje_3, MARGIN=2, FUN=mean)
head(srednie_3[1:10])

## [1]  0.007570508  0.012684451 -0.057015254  0.035015229 -0.008023315
## [6] -0.047481279
```

Rozważany rozkład teoretyczny jest asymptotycznie normalny z parametrami $AN\left(0, \frac{1}{5^2 n_3}\right)$, gdzie $n_3 = 50$ to długość szeregu.



Rysunek 5: Porównanie estymatora jądrowego dla *srednie_3* oraz gęstości rozkładu $N\left(0, \frac{1}{5^2 n_3}\right)$

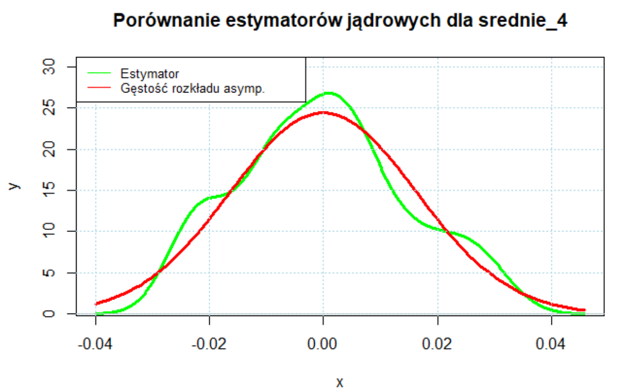


Rysunek 6: Porównanie dystrybuanty empirycznej dla *srednie_3* oraz teoretycznej rozkładu $N\left(0, \frac{1}{5^2 n_3}\right)$

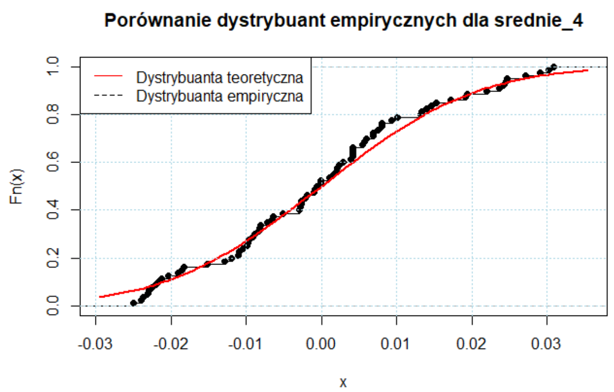
Pomimo dość małej długości szeregu, rozkład empiryczny dobrze przybliża rozkład teoretyczny. Zobaczmy w przykładach 4. i 5., jak będą wyglądać własności rozkładu średniej próbkowej dla większych długości szeregu.

4. Szereg ma długość $n_4 = 150$ oraz obserwacje pochodzą z rozkładu $Exp(5)$

```
srednie_4 <- apply(realizacje_4, MARGIN=2, FUN=mean)
head(srednie_4[1:10])

## [1] -0.006758488  0.001690032  0.022757574 -0.016571571  0.004670700
## [6] -0.002893414
```



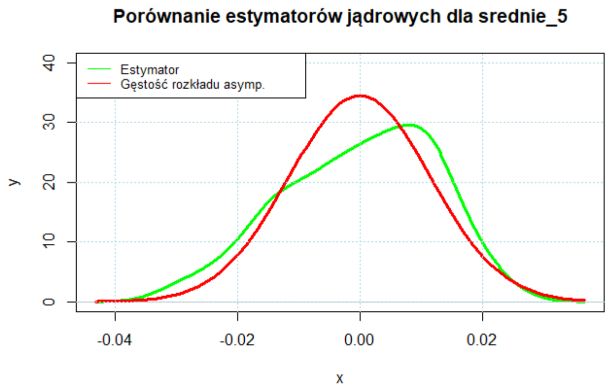
Rysunek 7: Porównanie estymatora jądrowego dla *srednie_4* oraz gęstości rozkładu $N\left(0, \frac{1}{5^2 n_4}\right)$



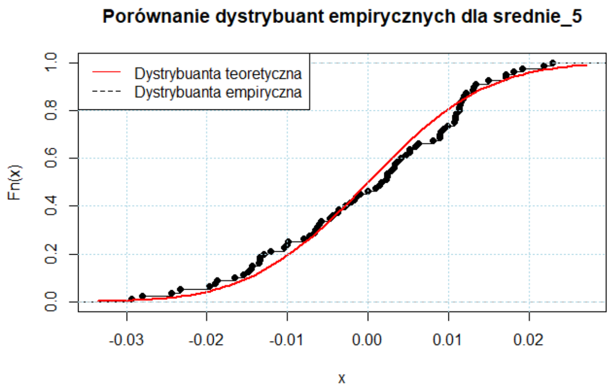
Rysunek 8: Porównanie dystrybuanty empirycznej dla *srednie_4* oraz teoretycznej rozkładu $N\left(0, \frac{1}{5^2 n_4}\right)$

5. Szereg ma długość $n_5 = 300$ oraz obserwacje pochodzą z rozkładu $Exp(5)$

```
srednie_5 <- apply(realizacje_5, MARGIN=2, FUN=mean)
head(srednie_5[1:10])

## [1]  0.003219674 -0.006428710 -0.011790776  0.005797311 -0.010174875
## [6]  0.036578126
```



Rysunek 9: Porównanie estymatora jądrowego dla *srednie_5* oraz gęstości rozkładu $N\left(0, \frac{1}{5^2 n_5}\right)$



Rysunek 10: Porównanie dystrybuanty empirycznej dla *srednie_5* oraz teoretycznej rozkładu $N\left(0, \frac{1}{5^2 n_5}\right)$

Wniosek 2 *Z porównania wykresów (5), (6), (7), (8), (9), (10) wnosimy, że im długość szeregu czasowego jest większa, tym rozkład empiryczny lepiej przybliża rozkład teoretyczny (w sensie słabej zbieżności), gdy obserwacje są losowane z rozkładu wykładniczego.*

Przejdźmy teraz do zbadania własności asymptotycznych estymatora autokorelacji.

2.1.2 Estymator autokorelacji

We wszystkich poniższych eksperymentach przyjęliśmy, że maksymalne opóźnienie h występujące w definicji funkcji (i jej estymatora) autokorelacji wynosi $\lfloor n \rfloor$, gdzie n to długość szeregu czasowego, w wyborze posłużyliśmy się metodą **Boxa-Jenkinsa**. Pod koniec tej podpodsekcji sprawdzimy, jak się zachowuje estymator autokorelacji dla opóźnień h bliskich n. Ponadto z rozważań usuwamy $ACF(0)$, ponieważ z definicji wartość ta wynosi 1 i nie wnosi nic do analizy.

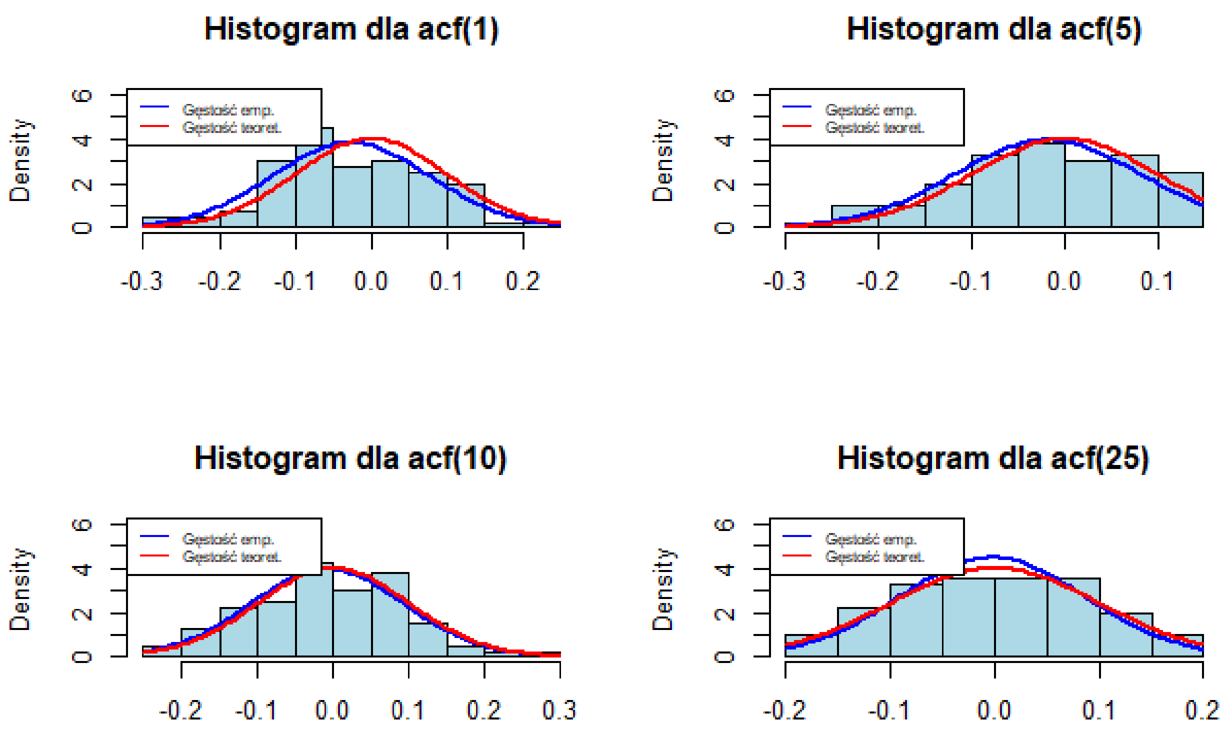
Przed przystąpieniem do analizy zauważmy, że mamy do czynienia z białym szumem, zatem $\rho(h) = 0$ dla $h = 1, 2, \ldots, h.max$ oraz $\rho(h) = 1$ dla $h = 0$.

1. Szereg ma długość $n_1 = 100$ oraz obserwacje pochodzą z rozkładu $N(0, 1)$. Do graficznej analizy wybieramy opóźnienia $h \in \{1, 5, 10, 25\}$, maksymalne opóźnienie wynosi 25.

```
# wyznaczamy macierz ACF (w wierszach mamy realizacje ACF(h) dla h=1,2,3,...,h.ma
h.max_1 <- floor(n_1/4) # maksymalne opóźnienie
acf.matrix_1 <- apply(realizacje_1, 2, function(x) acf(x, lag.max=h.max_1,
```

```
                                        type="correlation", plot=FALSE)$a

# usuwamy ACF(0)
acf.matrix_1 <- acf.matrix_1[-1,]
h.wybrane_1 <- c(1, 5, 10, 25)
```

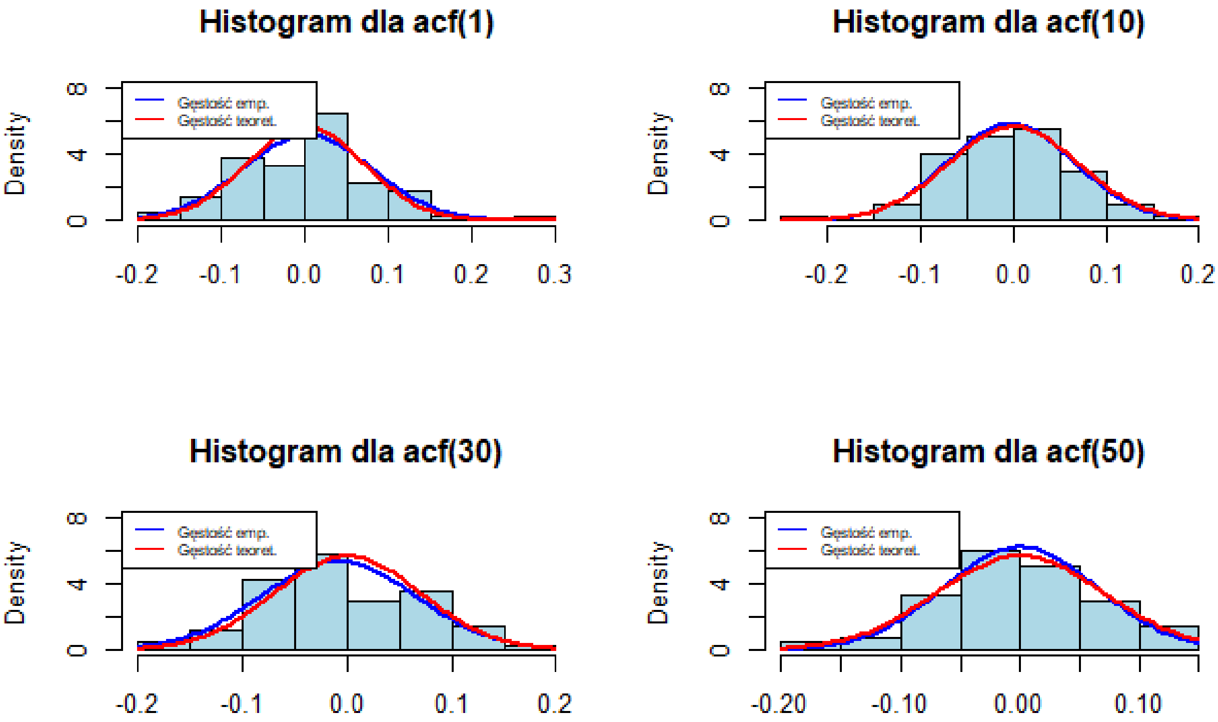
Spójrzmy teraz na narysowane histogramy i gęstości rozkładu empirycznego oraz teoretycznego ($N\left(0, \frac{1}{n_1}\right)$).



Rysunek 11: Histogram oraz wykres estymatora jądrowego wraz z gęstością rozkładu teoretycznego ($N\left(0, \frac{1}{n_1}\right)$

2. Szereg ma długość $n_2 = 200$ oraz obserwacje pochodzą z rozkładu $N(0,1)$. Do graficznej analizy wybieramy opóźnienia $h \in \{1, 10, 30, 50\}$, maksymalne opóźnienie wynosi 50.

```
h.max_2 <- floor(n_2/4) # maksymalne opóźnienie
acf.matrix_2 <- apply(realizacje_2, 2, function(x) acf(x, lag.max=h.max_2,
                                    type="correlation", plot=FALSE)$acf)

# usuwamy ACF(0)
acf.matrix_2 <- acf.matrix_2[-1,]
h.wybrane_2 <- c(1, 10,30 ,50 )
```

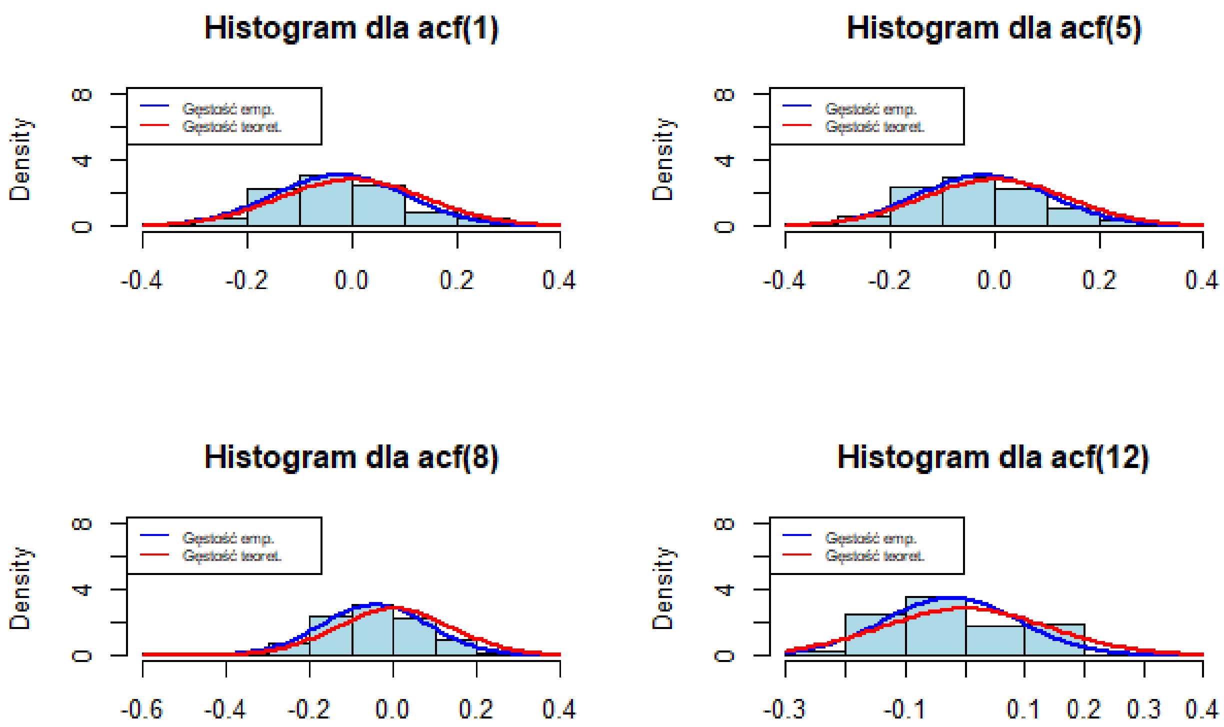


Rysunek 12: Histogram oraz wykres estymatora jądrowego wraz z gęstością rozkładu teoretycznego ($N\left(0, \frac{1}{n_2}\right)$

Wniosek 3 *Estymator autokorelacji funkcji autokorelacji szeregu o obserwacjach wylosowanych z rozkładu* $N(0,1)$ *ma bardzo dobre własności asymptotyczne, pomimo że* $\boldsymbol{n}$ *nie jest "bardzo duże".*

3. Szereg ma długość $n_3 = 50$ oraz obserwacje pochodzą z rozkładu $Exp(5)$. Do graficznej analizy wybieramy opóźnienia $h \in \{1, 5, 8, 12\}$, maksymalne opóźnienie wynosi 12.

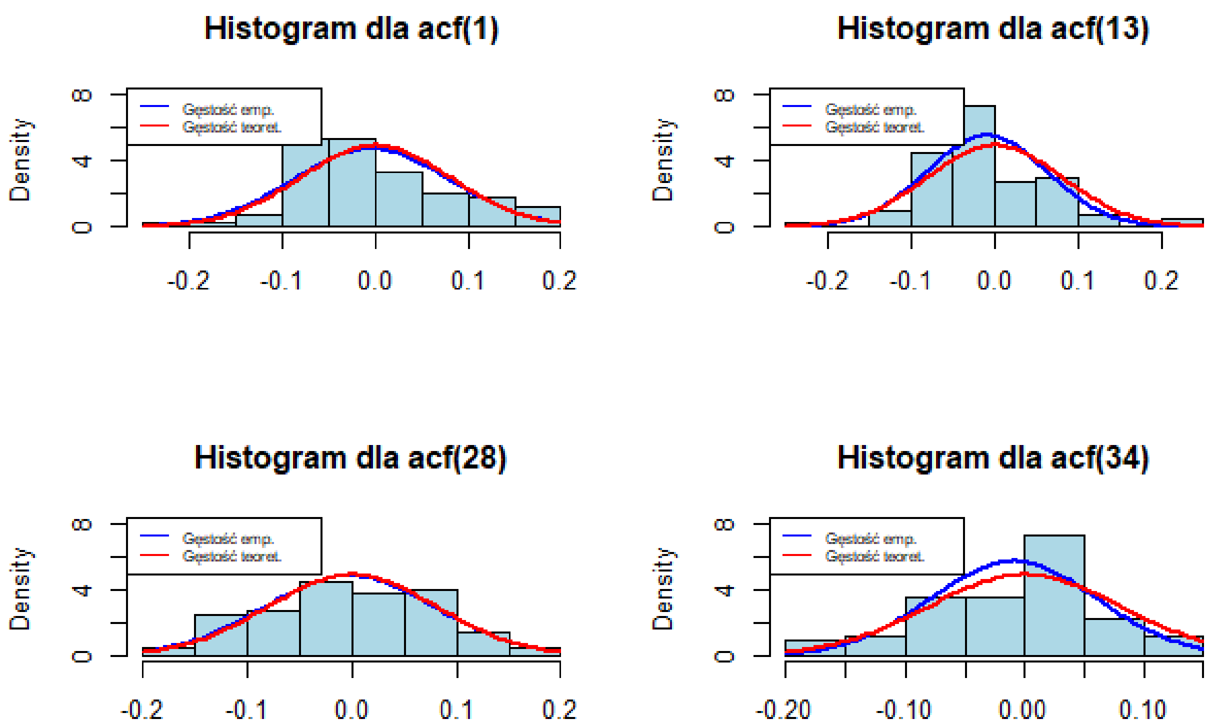
```
h.max_3 <- floor(n_3/4) # maksymalne opóźnienie
acf.matrix_3 <- apply(realizacje_3, 2, function(x) acf(x, lag.max=h.max_3,
                                                    type="correlation", plot=FALSE)$acf

# usuwamy ACF(0)
acf.matrix_3 <- acf.matrix_3[-1,]

h.wybrane_3 <- c(1, 5,8 ,12)
```



Rysunek 13: Histogram oraz wykres estymatora jądrowego wraz z gęstością rozkładu teoretycznego $(N\left(0, \frac{1}{n_3}\right)$

4. Szereg ma długość $n_4 = 150$ oraz obserwacje pochodzą z rozkładu $Exp(5)$. Do graficznej analizy wybieramy opóźnienia $h \in \{1, 13, 28, 34\}$, maksymalne opóźnienie wynosi 37.

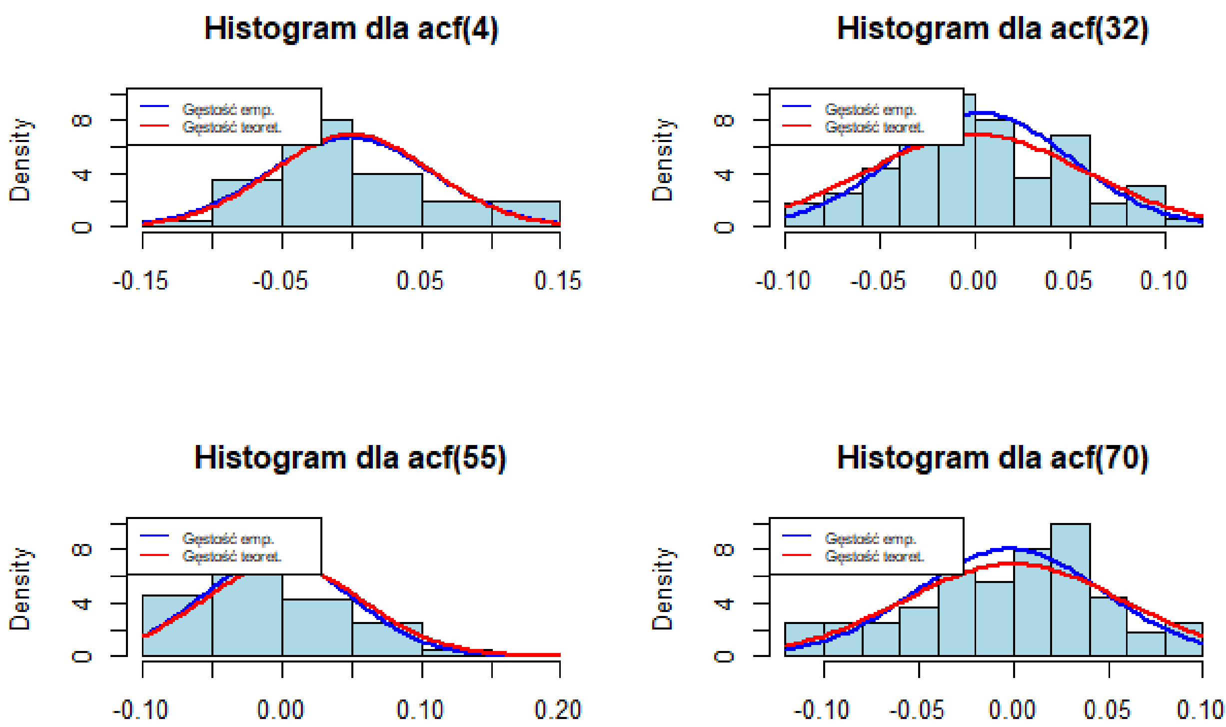
```
h.max_4 <- floor(n_4/4) # maksymalne opóźnienie
acf.matrix_4 <- apply(realizacje_4, 2, function(x) acf(x, lag.max=h.max_4,
                                            type="correlation", plot=FALSE)$acf)

# usuwamy ACF(0)
acf.matrix_4 <- acf.matrix_4[-1,]
h.wybrane_4 <- c(1, 13,28 ,34)
```



Rysunek 14: Histogram oraz wykres estymatora jądrowego wraz z gęstością rozkładu teoretycznego ($N\left(0, \frac{1}{n_4}\right)$

5. Szereg ma długość $n_5 = 300$ oraz obserwacje pochodzą z rozkładu $Exp(5)$. Do graficznej analizy wybieramy opóźnienia $h \in \{4, 32, 55, 70\}$, maksymalne opóźnienie wynosi 75.

```
h.max_5 <- floor(n_5/4) # maksymalne opóźnienie (mo■na uzale■nić od n)
acf.matrix_5 <- apply(realizacje_5, 2, function(x) acf(x, lag.max=h.max_5,
                                                   type="correlation", plot=FALSE)$a
# usuwamy ACF(0)
acf.matrix_5 <- acf.matrix_5[-1,]

h.wybrane_5 <- c(4, 32,55 ,70)
```

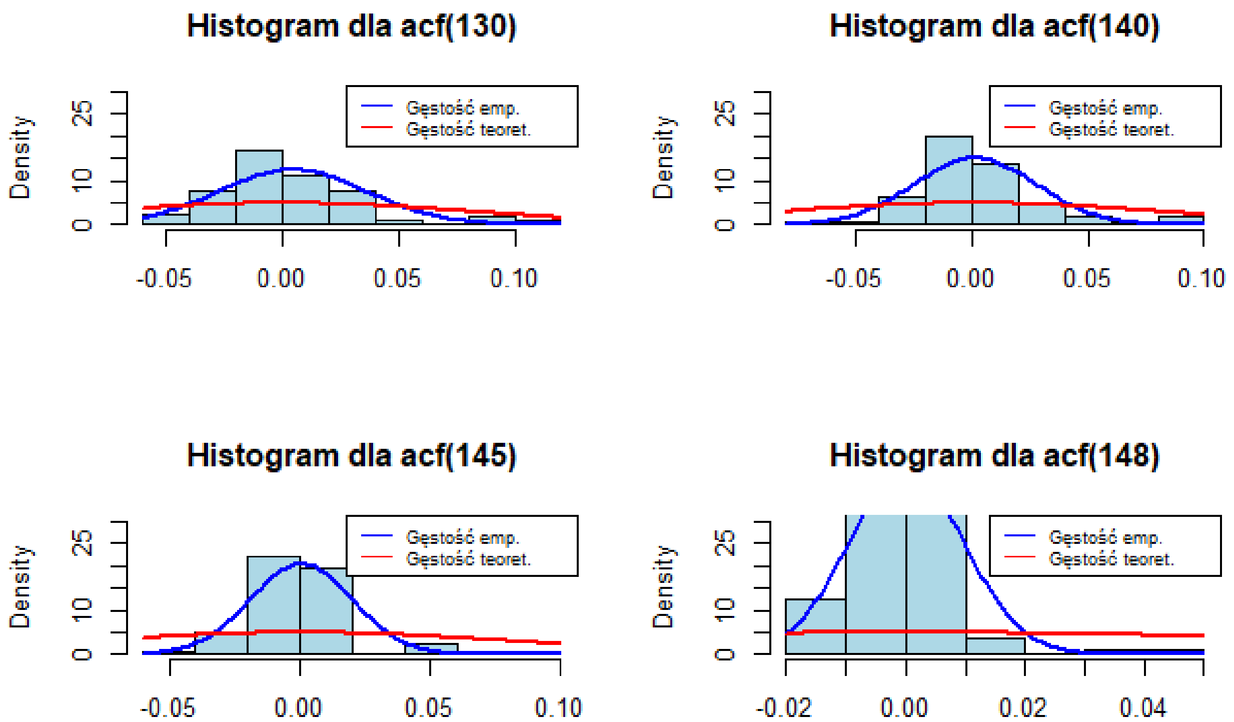


Rysunek 15: Histogram oraz wykres estymatora jądrowego wraz z gęstością rozkładu teoretycznego $(N\left(0, \frac{1}{n_5}\right)$

Wniosek 4 *Im długość n szeregu (w przypadku, gdy realizacje szeregu są wylosowane z $Exp(5)$) rośnie, tym lepiej empiryczny rozkład estymatora autokorelacji przybliża teoretyczny asymptotyczny rozkład mormalny $AN\left(0, \frac{1}{n}\right)$. Estymator ten ma także najlepsze własności asymptotyczne dla opóźnień h znajdujących się "pośrodku" zbioru $\{1, 2, \ldots, h.max\}$.*

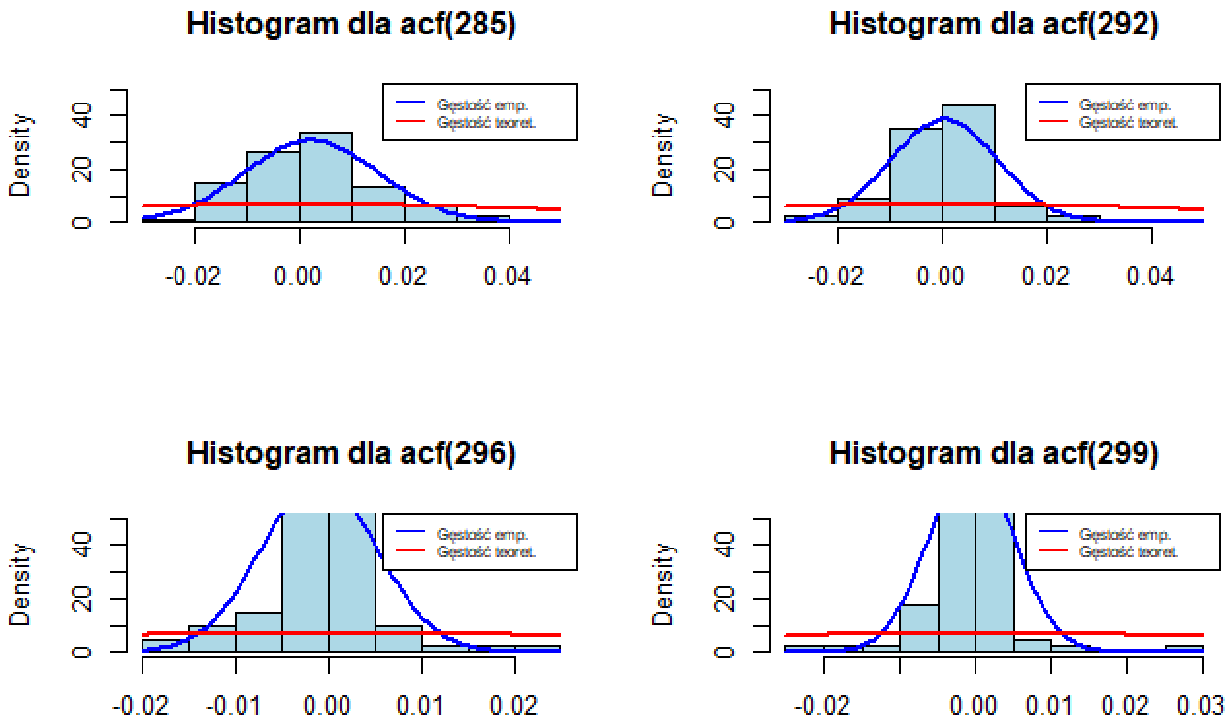
Zastanówmy się jeszcze, jak zachowuje się estymator autokorelacji dla opóźnień h bliskich n. Bez utraty ogólności tę analizę możemy wykonać jedynie dla tego estymatora - własności asymptotyczne estymatora autokowariancji będą takie same, ponieważ estymator autokowariancji to przeskalowany estymator autokorelacji. W tym celu rozważmy odpowiednie realizacje pochodzące z rozkładu wykładniczego $Exp(5)$.

1. Szereg ma długość $n_4 = 150$. Do graficznej analizy wybieramy opóźnienia $h \in \{130, 140, 145, 148\}$, maksymalne opóźnienie wynosi 149.



Rysunek 16: Histogram oraz wykres estymatora jądrowego wraz z gęstością rozkładu teoretycznego ($N\left(0, \frac{1}{n_4}\right)$ dla dużych opóźnień h

2. Szereg ma długość $n_5 = 300$. Do graficznej analizy wybieramy opóźnienia $h \in \{285, 292, 296, 299\}$, maksymalne opóźnienie wynosi 299.



Rysunek 17: Histogram oraz wykres estymatora jądrowego wraz z gęstością rozkładu teoretycznego ($N\left(0, \frac{1}{n_5}\right)$ dla dużych opóźnień h

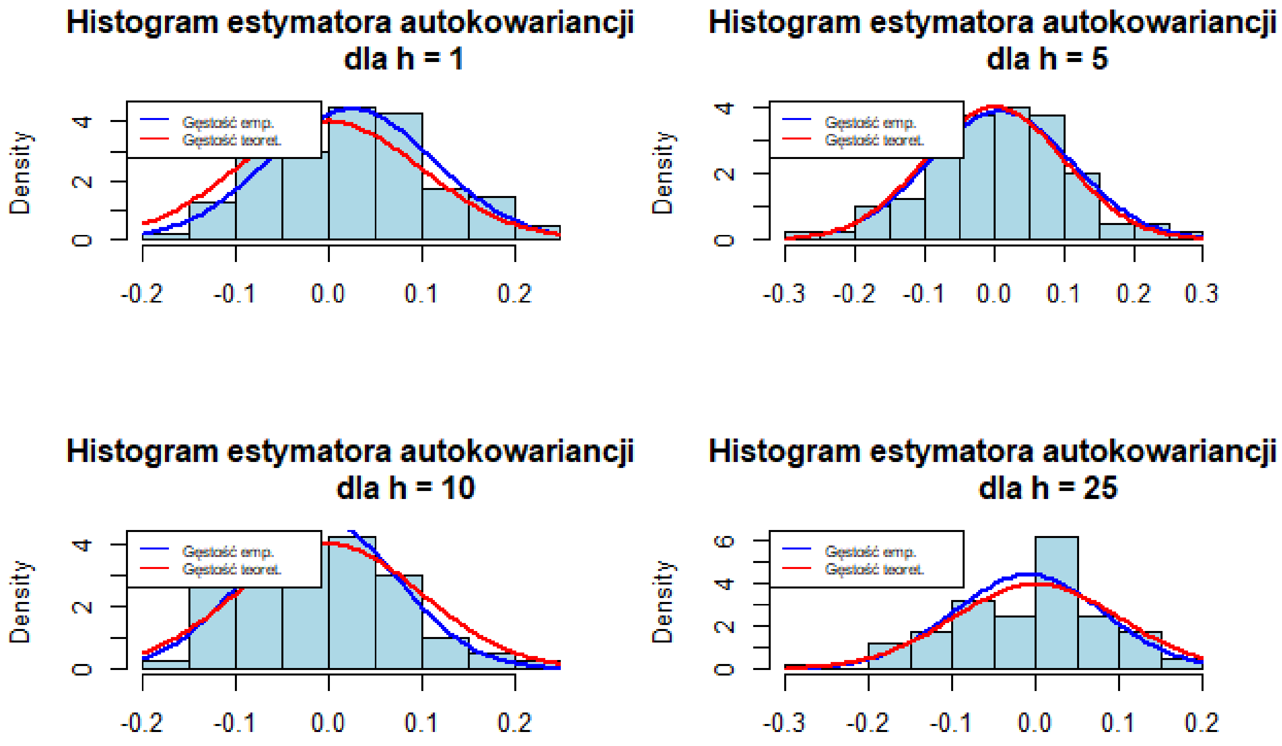
Wniosek 5 *Dla dużych opóźnień h własności estymatora autokorelacji (i także autokowariancji) nie są dobre. Pokrywa się to z definicjami tych estymatorów, ponieważ, im n jest bliższe h, tym mniej obserwacji z próby bierzemy do wyznaczenia tych estymatorów.*

2.1.3 Estymator autokowariancji

Teraz zajmiemy się własnościami asymptotycznymi estymatora autokowariancji. Zauważmy, że estymator ten jest równy estymatorowi autokorelacji pomnożonego przez estymator wariancji próbkowej. W związku z tym, wspomniane własności powinny pokrywać się z własnościami estymatora autokorelacji. Sprawdźmy, czy rzeczywiście tak jest.

1. Szereg ma długość $n_1 = 100$ oraz obserwacje pochodzą z rozkładu $N(0,1)$. Do graficznej analizy wybieramy opóźnienia $h \in \{1, 5, 10, 25\}$, maksymalne opóźnienie wynosi 25.

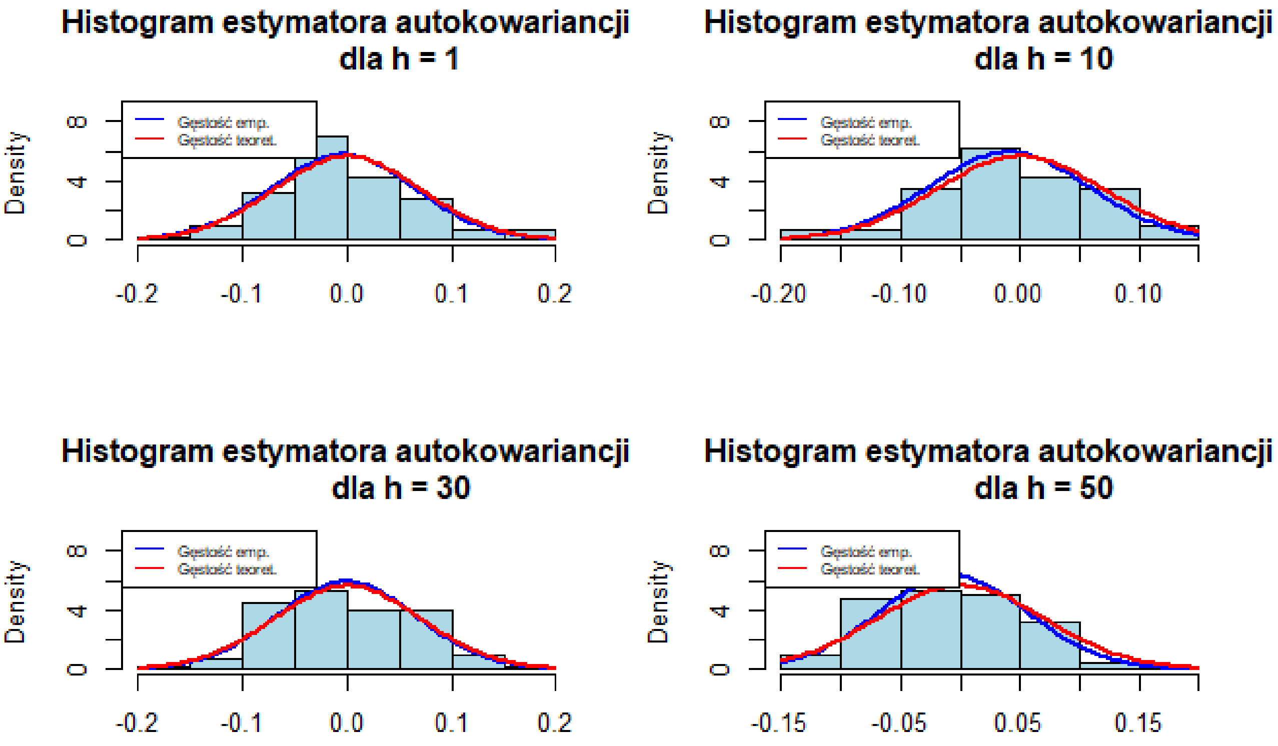
```
matrix_1_autokow <- apply(realizacje_1, 2, function(x) acf(x, lag.max=h.max_1,
                                        type="covariance", plot=FALSE)$acf)
matrix_1_autokow <- matrix_1_autokow[-1,]
```



Rysunek 18: Histogram oraz wykres estymatora jądrowego wraz z gęstością rozkładu teoretycznego $(N\left(0,\frac{1^4}{n_1}\right)$

2. Szereg ma długość $n_2 = 200$ oraz obserwacje pochodzą z rozkładu $N(0,1)$. Do graficznej analizy wybieramy opóźnienia $h \in \{1, 10, 30, 50\}$, maksymalne opóźnienie wynosi 50.

```
matrix_2_autokow <- apply(realizacje_2, 2, function(x) acf(x, lag.max=h.max_2,
                                            type="covariance", plot=FALSE)$acf)
matrix_2_autokow <- matrix_2_autokow[-1,]
```

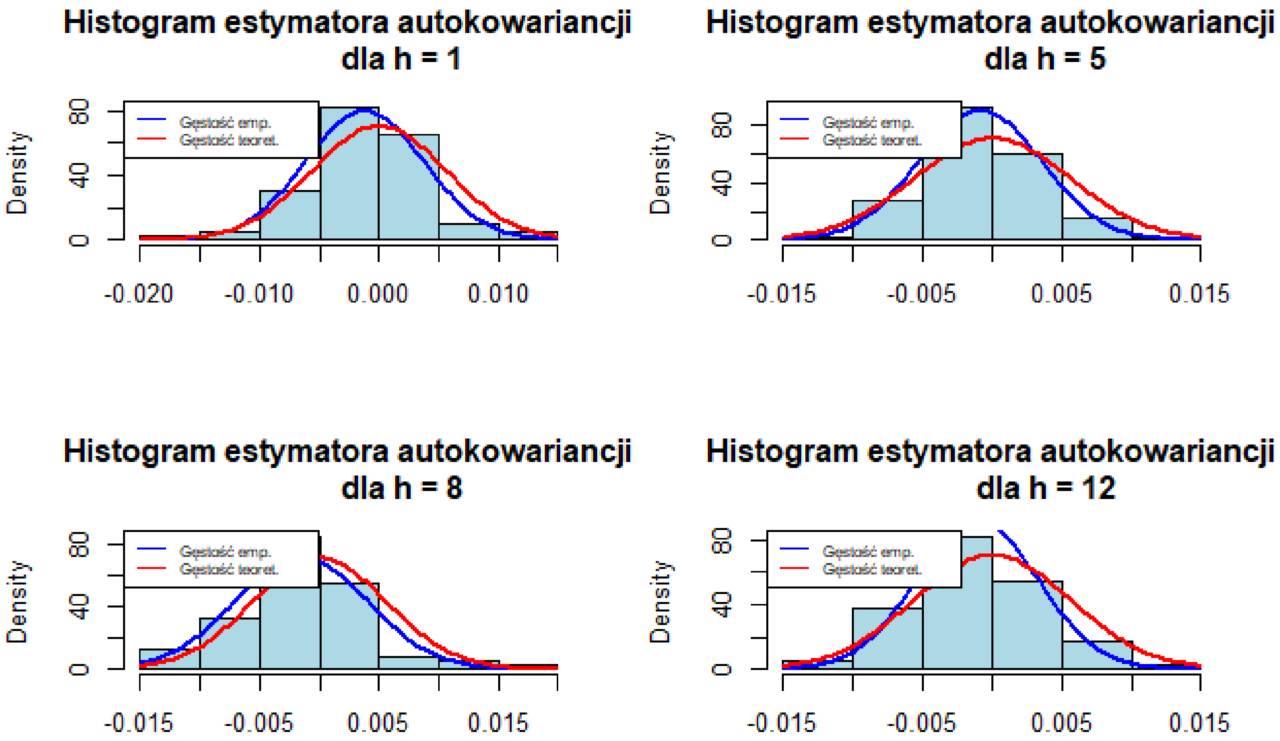


Rysunek 19: Histogram oraz wykres estymatora jądrowego wraz z gęstością rozkładu teoretycznego ($N\left(0, \frac{1^4}{n_2}\right)$

Wniosek 6 *Wniosek jest taki sam, jak przy estymatorze autokorelacji dla rozkładu normalnego (wniosek nr 1).*

3. Szereg ma długość $n_3 = 50$ oraz obserwacje pochodzą z rozkładu $Exp(5)$. Do graficznej analizy wybieramy opóźnienia $h \in \{1, 5, 8, 12\}$, maksymalne opóźnienie wynosi 12.

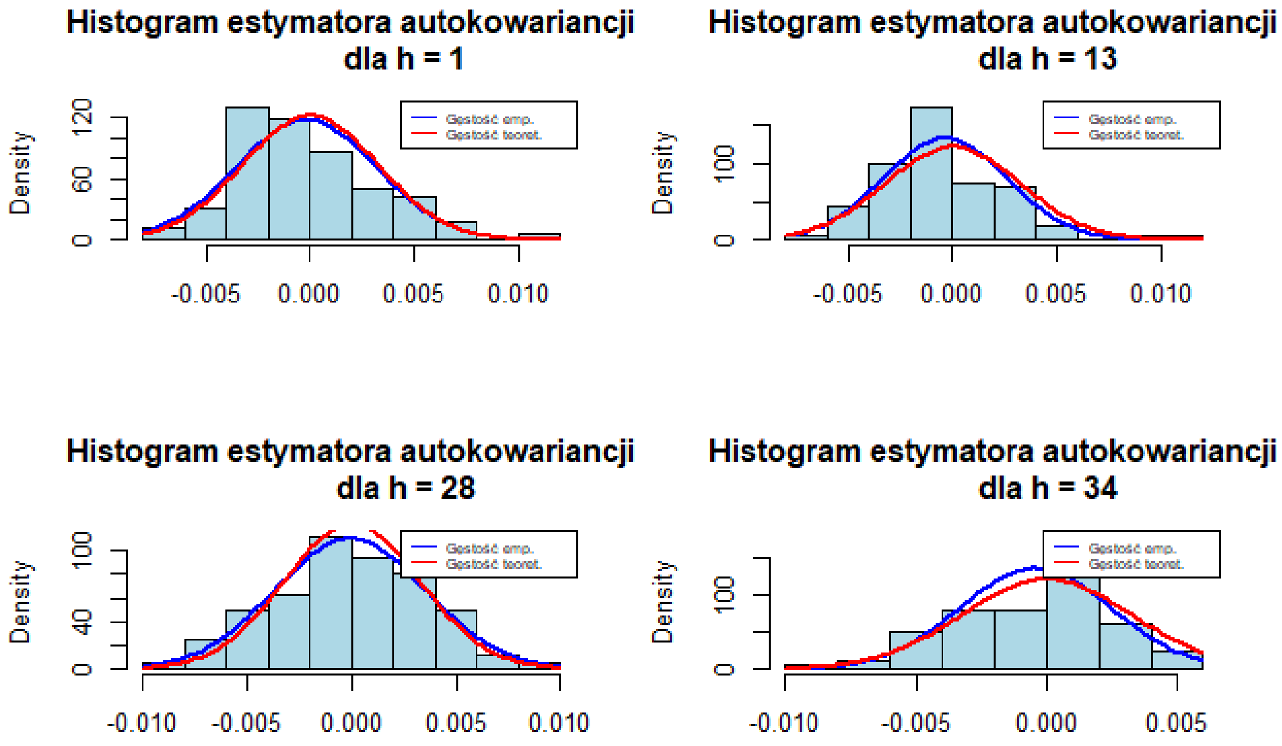
```
matrix_3_autokow <- apply(realizacje_3, 2, function(x) acf(x, lag.max=h.max_3,
                                                           type="covariance", plot=FAL

matrix_3_autokow <- matrix_3_autokow[-1,]
```



Rysunek 20: Histogram oraz wykres estymatora jądrowego wraz z gęstością rozkładu teoretycznego ($N\left(0, \frac{1}{5^4 n_3}\right)$

4. Szereg ma długość $n_4 = 150$ oraz obserwacje pochodzą z rozkładu $Exp(5)$. Do graficznej analizy wybieramy opóźnienia $h \in \{1, 13, 28, 34\}$, maksymalne opóźnienie wynosi 37.

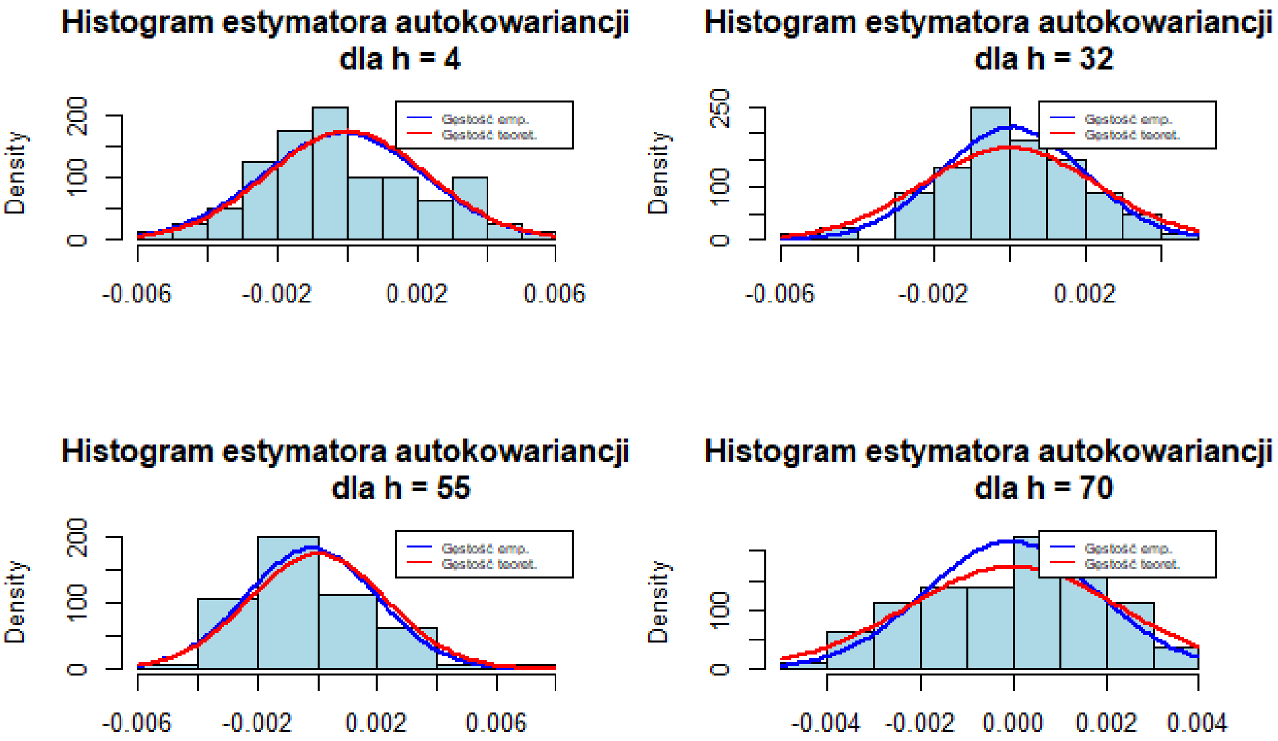
```
matrix_4_autokow <- apply(realizacje_4, 2, function(x) acf(x, lag.max=h.max_4,
                                                           type="covariance", plot=FAL

matrix_4_autokow <- matrix_4_autokow[-1,]
```



Rysunek 21: Histogram oraz wykres estymatora jądrowego wraz z gęstością rozkładu teoretycznego $(N\left(0, \frac{1}{5^4 n_4}\right)$

5. Szereg ma długość $n_5 = 300$. Do graficznej analizy wybieramy opóźnienia $h \in \{285, 292, 296, 299\}$, maksymalne opóźnienie wynosi 299.

```
matrix_5_autokow <- apply(realizacje_5, 2, function(x) acf(x, lag.max=h.max_5,
                                                        type="covariance", plot=FAL

matrix_5_autokow <- matrix_5_autokow[-1,]
```



Rysunek 22: Histogram oraz wykres estymatora jądrowego wraz z gęstością rozkładu teoretycznego ($N\left(0, \frac{1}{5^4 n_5}\right)$

Wniosek 7 *Tak jak oczekiwaliśmy, rozkład estymatora funkcji autokowariancji, wraz ze wzrostem długości szeregu n o obserwacjach wylosowanych z rozkładu wykładniczego, coraz lepiej przybliża rozkład normalny. Najlepsze przybliżenia obserwujemy dla opóźnień h nie za małych, i jednocześnie nie za dużych, $h \in \{1, 2, \ldots, h_{max}\}$.*

2.2 Symulacyjna analiza rozkładów asymptotycznych przy użyciu testów zgodności

Do symulacji wykorzystamy testy zgodności *Kołmogorowa-Smirnowa* oraz *Shapiro-Wilka*. Przy używaniu testu *Kołmogorowa-Smirnowa* musimy jednak uważać, czy aby przypadkiem rozważany rozkład nie zależy od nieznanych parametrów (estymowanych z próby). Wówczas nie możemy użyć statystyki testowej Kołmogorowa. Wtedy lepiej sprawdzi się test *Shapiro-Wilka*, którego statystyka testowa opiera się na ekscesie i skośności rozkładu. Ponadto przy symulacjach uwzględnimy losowość, powtarzając doświadczenia 1000 razy w przypadku średniej próbkowej oraz 100 razy w przypadku estymatora autokowariancji i autokorelacji. Wszystkie wyniki skompletujemy w postaci tabelki pod koniec każdej z podpodsekcji.

Wszystkie poniższe testy wyznaczamy na poziomie istotności 0.05. Testujemy hipotezę H_0, mówiącą o zgodności rozkładów vs H_1, która mówi o tym, że hipoteza H_0 jest fałszywa.

2.2.1 Średnia próbkowa

Do testowania zgodności w przypadku średniej próbkowej, używamy jedynie testu *Shapiro-Wilka*.

1. Szereg ma długość $n_1 = 100$ oraz obserwacje pochodzą z rozkładu $N(0,1)$.

```
ile.powtorz <- 1000

# deklaracja wektorow wynikowych
wynik.sw_srednie_1 <- numeric(ile.powtorz)

for (i in 1:ile.powtorz) {
  # generujemy k realizacji bia■lego szumu i wyznaczamy średnie
  realizacje_losowe_1 <- matrix(rnorm(n_1*k, mean=0, sd=1),n_1,k)
  srednie_losowe_1    <- apply(realizacje_losowe_1, 2, mean)

  # testujemy zgodność z rozk■ladem normalnym
  wynik.sw_srednie_1[i] <- shapiro.test(srednie_losowe_1)$p.value>0.05

}
# częstości przyjęcia H0
sum(wynik.sw_srednie_1)/ile.powtorz

## [1] 0.943
```

2. Szereg ma długość $n_2 = 200$ oraz obserwacje pochodzą z rozkładu $N(0,1)$. Kod ”ukryty” jest analogiczny do wcześniejszego przypadku.

```
# częstości przyjęcia H0
sum(wynik.sw_srednie_2)/ile.powtorz

## [1] 0.953
```

3. Szereg ma długość $n_3 = 50$ oraz obserwacje pochodzą z rozkładu $Exp(5)$.

```
#dla realizacje_3
wynik.sw_srednie_3 <- numeric(ile.powtorz)

for (i in 1:ile.powtorz) {
  # generujemy k realizacji bia■lego szumu i wyznaczamy średnie
  realizacje_losowe_3 <- matrix(rexp(n_3*k, rate=5),n_3,k)
  realizacje_losowe_3 <- realizacje_losowe_3 - colMeans(realizacje_losowe_3)
  srednie_losowe_3    <- apply(realizacje_losowe_3, 2, mean)

  # testujemy zgodność z rozk■ladem normalnym
  wynik.sw_srednie_3[i] <- shapiro.test(srednie_losowe_3)$p.value>0.05
}
```

```r
# częstości przyjęcia H0
sum(wynik.sw_srednie_3)/ile.powtorz

## [1] 0.852
```

Poniższe 2 przypadki są analogiczne, więc dla czytelności ukrywamy kod.

4. Szereg ma długość $n_4 = 150$ oraz obserwacje pochodzą z rozkładu $Exp(5)$.

```r
# częstości przyjęcia H0
sum(wynik.sw_srednie_4)/ile.powtorz

## [1] 0.917
```

5. Szereg ma długość $n_5 = 300$ oraz obserwacje pochodzą z rozkładu $Exp(5)$.

```r
# częstości przyjęcia H0
sum(wynik.sw_srednie_5)/ile.powtorz

## [1] 0.933
```

Przedstawmy wyniki w postaci tabeli.

Podpunkt	1.	2.	3.	4.	5.
Częstość przyjęcia H_0 w teście S-W	0.951	0.946	0.881	0.919	0.933

Tabela 1: Wyniki testu zgodności dla średniej próbkowej

Wniosek 8 *Test zgodności dla rozkładu normalnego wypadł bardzo dobrze, mimo że drugi wynik jest nieco gorszy od pierwszego. Natomiast jeśli chodzi o rozkład wykładniczy, to wyniki pokrywają się w zasadzie z analizą graficzną. Im większa długość szeregu, tym hipoteza zerowa, mówiąca o zgodności rozkładów, jest częściej przyjmowana.*

2.2.2 Estymator autokorelacji

Końcowe wyniki oraz komentarz w postaci wniosku zamieścimy na końcu tej podpodsekcji.

1. Szereg ma długość $n_1 = 100$ oraz obserwacje pochodzą z rozkładu $N(0,1)$. Do testowania zgodności wybieramy opóźnienia $h \in \{1,5,10,25\}$, maksymalne opóźnienie wynosi 25.

```r
ile.powtorz_1 <- 100

#dla realizacje_1
wynik.sw_autokor_1 <- matrix(rep(0, ile.powtorz_1*h.max_1), nrow=ile.powtorz_1,
                             ncol=h.max_1)
wynik.ks_autokor_1 <- matrix(rep(0, ile.powtorz_1*h.max_1), nrow=ile.powtorz_1,
```

```r
                              ncol=h.max_1)

for (h in h.wybrane_1) {
  for (j in 1:ile.powtorz_1) {
    realizacje_losowe_1 <- matrix(rnorm(n_1*k),n_1,k)
    matrix_losowe_1 <- apply(realizacje_losowe_1, 2, function(x) acf(x, lag.max=h.m
                                          type="correlation", plot=FALSE)$acf)
    matrix_losowe_1 <- matrix_losowe_1[-1,]
    h_losowe_1 <- matrix_losowe_1[h,]
    wynik.sw_autokor_1[j,h] <- shapiro.test(h_losowe_1)$p.value>0.05
    wynik.ks_autokor_1[j,h] <- ks.test(h_losowe_1, 'pnorm',
                                   mean=0, sd=1/sqrt(n_1))$p.value>0.05
  }
}

#Ile razy przyjeliśmy H0 dla poszczególnych opóźnień h?
sum(wynik.sw_autokor_1[,1])/ile.powtorz_1

## [1] 0.94

sum(wynik.sw_autokor_1[,5])/ile.powtorz_1

## [1] 0.96

sum(wynik.sw_autokor_1[,10])/ile.powtorz_1

## [1] 0.97

sum(wynik.sw_autokor_1[,25])/ile.powtorz_1

## [1] 0.99

sum(wynik.ks_autokor_1[,1])/ile.powtorz_1

## [1] 0.92

sum(wynik.ks_autokor_1[,5])/ile.powtorz_1

## [1] 0.88

sum(wynik.ks_autokor_1[,10])/ile.powtorz_1

## [1] 0.88

sum(wynik.ks_autokor_1[,25])/ile.powtorz_1

## [1] 0.89
```

2. Szereg ma długość $n_2 = 200$ oraz obserwacje pochodzą z rozkładu $N(0,1)$. Do testowania zgodności wybieramy opóźnienia $h \in \{1, 10, 30, 50\}$, maksymalne opóźnienie wynosi 50.

```
#Ile razy przyjeliśmy H0 dla poszczególnych opóźnień h?
sum(wynik.sw_autokor_2[,1])/ile.powtorz_1

## [1] 0.94

sum(wynik.sw_autokor_2[,10])/ile.powtorz_1

## [1] 0.93

sum(wynik.sw_autokor_2[,30])/ile.powtorz_1

## [1] 0.96

sum(wynik.sw_autokor_2[,50])/ile.powtorz_1

## [1] 0.95

sum(wynik.ks_autokor_2[,1])/ile.powtorz_1

## [1] 0.9

sum(wynik.ks_autokor_2[,10])/ile.powtorz_1

## [1] 0.92

sum(wynik.ks_autokor_2[,30])/ile.powtorz_1

## [1] 0.92

sum(wynik.ks_autokor_2[,50])/ile.powtorz_1

## [1] 0.92
```

3. Szereg ma długość $n_3 = 50$ oraz obserwacje pochodzą z rozkładu $Exp(5)$. Do testowania zgodności wybieramy opóźnienia $h \in \{1, 5, 8, 12\}$, maksymalne opóźnienie wynosi 12.

```
#Ile razy przyjeliśmy H0 dla poszczególnych opóźnień h?
sum(wynik.sw_autokor_3[,1])/ile.powtorz_1

## [1] 0.75
```

```
sum(wynik.sw_autokor_3[,5])/ile.powtorz_1

## [1] 0.83

sum(wynik.sw_autokor_3[,8])/ile.powtorz_1

## [1] 0.81

sum(wynik.sw_autokor_3[,12])/ile.powtorz_1

## [1] 0.91

sum(wynik.ks_autokor_3[,1])/ile.powtorz_1

## [1] 0.68

sum(wynik.ks_autokor_3[,5])/ile.powtorz_1

## [1] 0.74

sum(wynik.ks_autokor_3[,8])/ile.powtorz_1

## [1] 0.65

sum(wynik.ks_autokor_3[,12])/ile.powtorz_1

## [1] 0.68
```

4. Szereg ma długość $n_4 = 150$ oraz obserwacje pochodzą z rozkładu $Exp(5)$. Do testowannia zgodności wybieramy opóźnienia $h \in \{1, 13, 28, 34\}$, maksymalne opóźnienie wynosi 37.

```
#Ile razy przyjeliśmy H0 dla poszczególnych opóźnień h?
sum(wynik.sw_autokor_4[,1])/ile.powtorz_1

## [1] 0.86

sum(wynik.sw_autokor_4[,13])/ile.powtorz_1

## [1] 0.9

sum(wynik.sw_autokor_4[,28])/ile.powtorz_1

## [1] 0.89
```

```
sum(wynik.sw_autokor_4[,34])/ile.powtorz_1

## [1] 0.86

sum(wynik.ks_autokor_4[,1])/ile.powtorz_1

## [1] 0.86

sum(wynik.ks_autokor_4[,13])/ile.powtorz_1

## [1] 0.81

sum(wynik.ks_autokor_4[,28])/ile.powtorz_1

## [1] 0.91

sum(wynik.ks_autokor_4[,34])/ile.powtorz_1

## [1] 0.81
```

5. Szereg ma długość $n_5 = 300$. Do testowania zgodności wybieramy opóźnienia $h \in \{285, 292, 296, 299\}$, maksymalne opóźnienie wynosi 299.

```
wynik.sw_autokor_5 <- matrix(rep(0, ile.powtorz_1*h.max_5), nrow=ile.powtorz_1,
                             ncol=h.max_5)
wynik.ks_autokor_5 <- matrix(rep(0, ile.powtorz_1*h.max_5), nrow=ile.powtorz_1,
                             ncol=h.max_5)


for (h in h.wybrane_5) {
  for (j in 1:ile.powtorz_1) {
    realizacje_losowe_5 <- matrix(rexp(n_5*k, 5),n_5,k)
    acf.matrix_losowe_5 <- apply(realizacje_losowe_5, 2, function(x) acf(x,
                               lag.max=h.max_5, type="correlation", plot=FALSE)$acf)
    acf.matrix_losowe_5 <- acf.matrix_losowe_5[-1,]
    acf.h_losowe_5 <- acf.matrix_losowe_5[h,]
    wynik.sw_autokor_5[j,h] <- shapiro.test(acf.h_losowe_5)$p.value>0.05
    wynik.ks_autokor_5[j,h] <- ks.test(acf.h_losowe_5, 'pnorm', mean=0,
                                        sd=1/sqrt(n_5))$p.value>0.05

  }
}
```

```r
#Ile razy przyjeliśmy H0 dla poszczególnych opóźnień h?
sum(wynik.sw_autokor_5[,4])/ile.powtorz_1

## [1] 0.89

sum(wynik.sw_autokor_5[,32])/ile.powtorz_1

## [1] 0.86

sum(wynik.sw_autokor_5[,55])/ile.powtorz_1

## [1] 0.92

sum(wynik.sw_autokor_5[,70])/ile.powtorz_1

## [1] 0.84

sum(wynik.ks_autokor_5[,4])/ile.powtorz_1

## [1] 0.94

sum(wynik.ks_autokor_5[,32])/ile.powtorz_1

## [1] 0.86

sum(wynik.ks_autokor_5[,55])/ile.powtorz_1

## [1] 0.89

sum(wynik.ks_autokor_5[,70])/ile.powtorz_1

## [1] 0.87
```

Przedstawmy teraz wyniki w formie tabeli i wyciągnijmy odpowiednie wnioski. Załóżmy, że tabelę robimy zawsze jedynie dla trzeciej wartości opóźnienia h spośród wszystkich wektorów *h_wybrane_1*, *h_wybrane_2*, *h_wybrane_3*, *h_wybrane_4*, *h_wybrane_5*. Dlaczego takie h? Ponieważ w tym przypadku zwykle wtedy estymatory miały dobre własności asymptotyczne, co mogliśmy zauważyć szybciej, rysując odpowiednie histogramy. Sprawdźmy, czy testy potwierdzają analizę graficzną.

Wniosek 9 *Testy zgodności potwierdzają przeprowadzoną analizę graficzną przy "optymalnym" wyborze opóźnienia* h*. Dla pozostałych wartości* h *wniosek pozostaje w zasadzie bez zmian, gdy rozważamy coraz to większe długości szeregu* n*. W szczególności możemy zaobserwować wzrost poprawnie przyjętych hipotez* H_0 *w obu testach, gdy rozważamy rozkład wykładniczy, co także odpowiada szybszym analizom i wnioskom.*

Podpunkt	1.	2.	3.	4.	5
Częstość przyjęcia H_0 w teście *Kołmogorowa-Smirnowa*	0.93	0.93	0.64	0.84	0.92
Częstość przyjęcia H_0 w teście *Shapiro-Wilka*	0.91	0.88	0.77	0.89	0.9

Tabela 2: Wyniki testów zgodności dla estymatora autokorelacji dla trzecich opóźnień h

2.2.3 Estymator autokowariancji

Z racji tego, że estymator autokowariancji jest równy estymatorowi autokorelacji przemnożonemu przez estymator wariancji, wszystkie wnioski i wyniki z poprzedniej podsekcji stosują się i w tym przypadku.

3 Część druga sprawozdania

Wykorzystując własności funkcji autokorelacji białego szumu, zaimplementuj i przeprowadź test graficzny weryfikujący hipotezę, że obserwowany szereg czasowy jest szeregiem czasowym typu biały szum.

3.1 Test graficzny

Test graficzny hipotezy o białym szumie polega na zweryfikowaniu dwóch warunków dla wartości funkcji autokorelacji:

- 95% wartości znajduje się wewnątrz przedziału ufności
- Żadna z wartości nie przekracza przedziału ufności w sposób 'istotny'

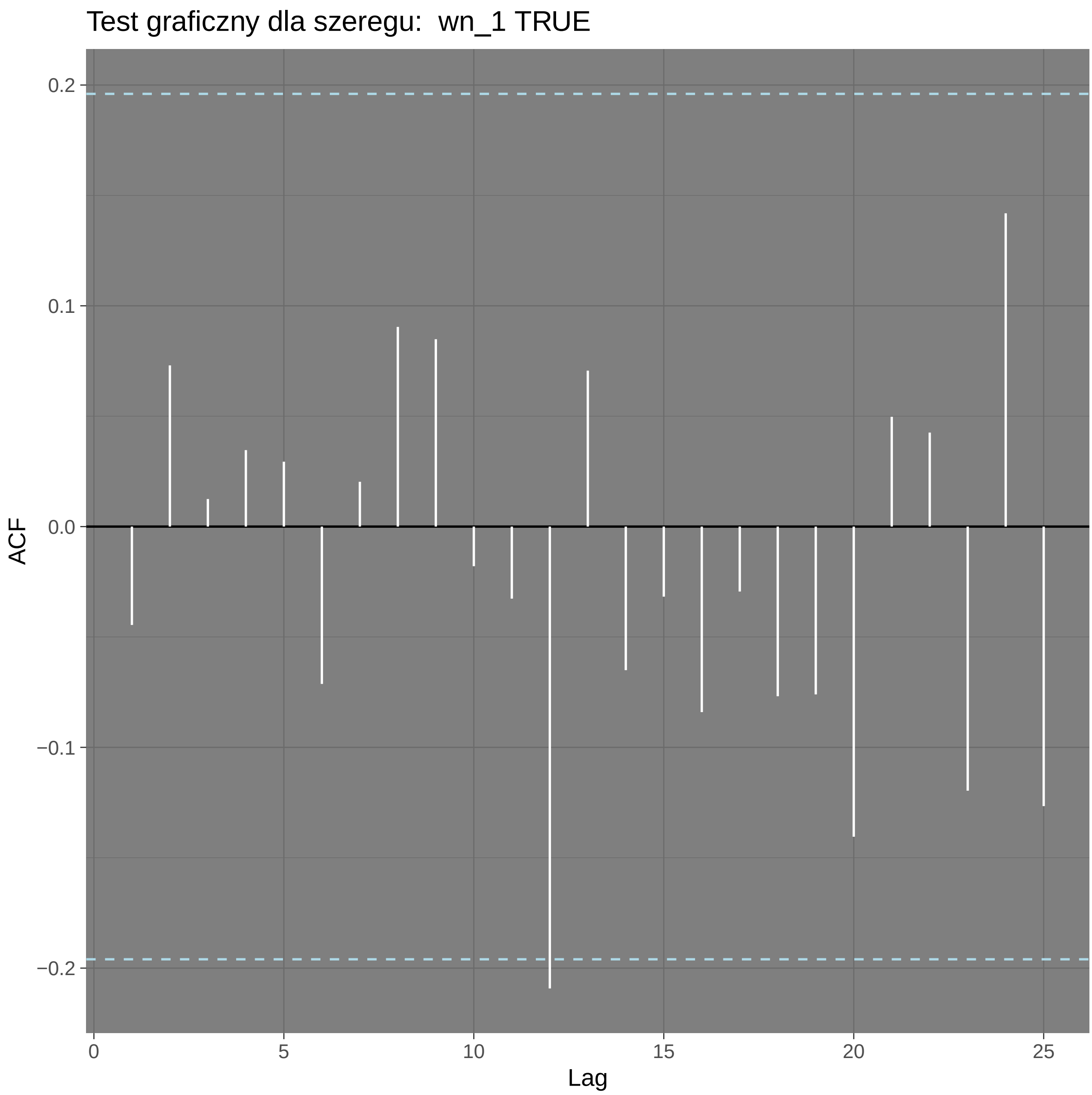
Za przedział ufności przyjmujemy $(-\frac{1.96}{\sqrt{n}}, \frac{1.96}{\sqrt{n}})$, gdzie n - długość szeregu czasowego, natomiast za 'istotne' przekroczenie przedziału uznajemy wartości większe lub równe (co do modułu) $\frac{3}{\sqrt{n}}$.
Wartość *lag.max* przyjmujemy zgodnie z wykładem $\lfloor \frac{n}{4} \rfloor$.
W celu implementacji tego testu korzystamy z funkcji *ggAcf* z paczki *forecast*. Dodatkowo w celu przeprowadzenia symulacji (skrócenie czasu działania) korzystamy z argumentu *plot=FALSE*.
W celu ewaluacji wyżej wymienionych warunków stosujemy poniższą implementacje w R:

- mean(abs(acf)<= conf.int)
- any(abs(acf)>=conf.int)

Pierwszy z warunków zwraca ilość elementów znajdujących się wewnątrz przedziałów ufności, w celu ewaluacji chcemy by była ona wyższa niż 0.95, natomiast drugi z warunków zwraca wartość *TRUE* gdy istnieje przynajmniej jedna wartość przekraczająca 105% wartości przedziału ufności (co do modułu), zatem w celu ewaluacji chcemy uzyskać wartość *FALSE*.

W celu zobrazowania działania funkcji generujemy przykładowy biały szum długości 100, a następnie przywołujemy zaimplementowaną funkcję na nim:

```
## Warning: Ignoring unknown parameters: main

## [[1]]
## [1] 0.96
##
## [[2]]
## [1] FALSE
##
## [[3]]

## Warning: Removed 2 rows containing missing values (geom_hline).
```




```
##
## Autocorrelations of series 'wn_1', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000 -0.045  0.073  0.012  0.035  0.029 -0.071  0.020  0.090  0.085 -
0.018
##     11     12     13     14     15     16     17     18     19     20
## -0.033 -0.209  0.071 -0.065 -0.032 -0.084 -0.029 -0.077 -0.076 -0.140
```

3.2 Formalny test statystyczny

Formalny test statystyczny jest zaimplementowany w R pod postacią funkcji *Box.test* z paczki *stats*. Zwraca ona statystykę testową Boxa-Pierce'a lub Ljungi-Boxa. Po przywołaniu tej statystyki testowej sprawdzamy p.value: Jeżeli jest ona mniejsza niż poziom istotności α, to odrzucamy hipotezę zerową, która mówi o tym, że zadany szereg czasowy jest białym szumem. Za parametr lag przyjmujemy $\frac{1}{4}$ długości szeregu czasowego.
Jako przykład przywołamy wyżej wymienioną funkcję na tym samym szeregu co przykład dla testu graficznego:

```
##
##  Box-Pierce test
##
## data:  wn_1
## X-squared = 17.989, df = 25, p-value = 0.8429
```

Widzimy, że otrzymane p.value jest dużo wyższe od 0.05, zatem nie ma podstaw by odrzucić hipotezę zerową o białoszumowości.

Dodatkowo zainteresowaliśmy się doborem odpowiednich wartości *H(lag)*, sugerując się artykułem Selecting optimal lag order in Ljung-Box test, robimy dodatkowy test na poziomie alpha=0.05. Artykuł sugeruje, że optymalna wartość H zależy zarówno od długości szeregu czasowego oraz poziomu istotności testu. Najbardziej popularnymi 'standardowymi' wartościami są min(20,T-1) oraz ln(T), gdzie T- długość szeregu czasowego. Dla odpowiednio krótkich (<=500) szeregów wartość ln(T) jest stosunkowo zgodna z badaniami symulacyjnymi z artykułu (są to małe wartości, niedalekie 4), dlatego z niej skorzystamy, jednak dla większych wartości są one zbyt restrykcyjne. Skorzystamy zatem z podziału H=min{10,T/5}, co dla $T \in \{501, ..., 750\}$ daje H=10 sugerowane przez Hyndmana i Athanasopoulosa oraz H=20 sugerowane przez Shumwaya i Stoffera dla T >= 751. (Sugerowane wartości nie dotyczą tych przedziałów, są ogólne. Dodatkowo artykuł dla szeregów o długości w okolicy 1000 dla $\alpha = 0.05$ proponuje H=50, zatem zaniżamy tę wartość).

W celu przykładu przywołujemy zaimplementowaną funkcję znów na tym samym szeregu. Jest on długości 100, zatem przyjęta wartość parametru *lag* jest dużo mniejsza niż w przykładzie wyżej.

```
##
##  Box-Ljung test
##
## data:  ts
## X-squared = 0.90333, df = 4.6052, p-value = 0.9563
```

Otrzymana wartość p.value przy tym doborze wartości *lag* jest jeszcze większa.

3.3 Symulacje

W celu przeprowadzenia symulacji generujemy różne typy szeregów, a następnie ewaluujemy ilość nie odrzuconych hipotez zerowych (o białoszumowości). Dla każdego typu generujemy 500 realizacji o zadanych długościach: 50/100/600/1000

Ewaluację skuteczności testów na podstawie białych szumów realizujemy dla ciągów i.i.d. z rozkładów normalnych o średnich 0 oraz odchyleniach standardowych: 0.5/1/2
Z kolei ewaluację na podstawie innych szeregów przeprowadzimy poprzez generowanieszeregów funkcją *arima.sim* przymując w miejscu parametru *model*:

- pojedyncze oraz podwójne wartości *ar*, np. 0.5 lub (-0.2, 0.5)
- pojedyncze oraz podwójne wartości *ma*, np. -2.5 lub (3, 0.6)
- parę pojedynczych wartości *ar* oraz *ma*, np ar=0.6, ma=2.7

Korzystając z funkcji *apply* sprawdzamy spełnienie odpowiednich warunków w testach: Dla testu graficznego odrzucamy hipotezę zerową o białoszumowości gdy jeden z warunków nie jest spełniony, natomiast dla testów formalnych hipoteza zerowa jest odrzucana gdy p.value jest mniejsze od poziomu istotności testu, zatem dla nas jest to 0.05.

Wnioski na temat mocy testu patrząc na rozpoznawalność białego szumu patrząc na pierwsze trzy tabele (Symulacje na białych szumach) są bardzo czytelne: testy formalne są dużo skuteczniejsze, widzimy także, że po dopasowaniu wartości H test jest jeszcze lepszy. Test graficzny natomiast ma mniejszą dokładność wśród krótkich szeregów, następnie ona rośnie, lecz znów maleje wraz z wydłużeniem szeregu. Testy formalne są zdecydowanie lepsze od testów graficznych.

Skupiając się natomiast na szeregach AR z pojedynczymi parametrami *ar* (tabele 3 oraz 5) możemy zaobserwować, że testy formalne przymują znacznie więcej hipotez zerowych niż test graficzny. Zmodyfikowanie wartości H poprawiło jednak klasyfikacje dla dłuższych szeregów.

Dla szeregów MA z pojedynczymi parametrami sytuacja jest analogiczna jak dla szeregów AR.

Zarówno dla szeregów AR jak i MA z podwójnymi parametrami szereg graficzny przyjmuje zero (lub prawie zero) hipotez zerowych, natomiast testy formalne dla krótkich szeregów w zależności od wartości parametrów przyjmują zauważalną ilość, lub (prawie) wcale.

Patrząc na szeregi ARMA jako ciekawy fakt można uznać, że ujemna wartość parametru *ar* skutkuje zwiększoną klasyfikacją szeregów jako biały szum (w przeciwieństwie do wartości dodatniej, gdzie otrzymujemy praktycznie zerowe klasyfikacje). Dodatkowo zwiększając wartość parametru *ma* pogarszamy wynik jeszcze bardziej.

Podsumowując, test graficzny ma gorszą moc patrząc na wyniki w symulacjach na białych szumach, jednak również dla innych szeregów odrzuca on więcej hipotez zerowych niż zaimplementowane testy formalne. W celu jednoznacznej oceny testóW należałoby przede wszystkim przeanalizować jaki wpływ mają zmiany wartości parametrów dla danych typów szeregów oraz także konkretne typy wartości (np. ujemne vs dodatnie).

	T.50	T.100	T.600	T.1000
graphic	0.68	0.78	0.71	0.62
LB_test	0.94	0.93	0.93	0.91
LB_adjusted_H	0.97	0.98	0.97	0.94

Tabela 3: WN, N(0,0.5)

	T.50	T.100	T.600	T.1000
graphic	0.68	0.79	0.70	0.61
LB_test	0.92	0.90	0.90	0.92
LB_adjusted_H	0.98	0.97	0.95	0.93

Tabela 4: WN, N(0,1)

	T.50	T.100	T.600	T.1000
graphic	0.73	0.81	0.64	0.64
LB_test	0.94	0.94	0.90	0.92
LB_adjusted_H	0.99	0.96	0.96	0.95

Tabela 5: WN, N(0,2)

	T.50	T.100	T.600	T.1000
graphic	0.50	0.55	0.03	0.00
LB_test	0.83	0.78	0.45	0.27
LB_adjusted_H	0.86	0.73	0.05	0.01

Tabela 6: ar=-0.2

	T.50	T.100	T.600	T.1000
graphic	0.07	0.02	0.00	0.00
LB_test	0.37	0.12	0.00	0.00
LB_adjusted_H	0.29	0.02	0.00	0.00

Tabela 7: ar=0.5

	T.50	T.100	T.600	T.1000
graphic	0.12	0.04	0.00	0.00
LB_test	0.57	0.33	0.00	0.00
LB_adjusted_H	0.46	0.05	0.00	0.00

Tabela 8: ma=0.5

	T.50	T.100	T.600	T.1000
graphic	0.20	0.15	0.00	0.00
LB_test	0.64	0.48	0.00	0.00
LB_adjusted_H	0.60	0.20	0.00	0.00

Tabela 9: ma=-2.5

	T.50	T.100	T.600	T.1000
graphic	0.02	0.00	0.00	0.00
LB_test	0.20	0.03	0.00	0.00
LB_adjusted_H	0.17	0.01	0.00	0.00

Tabela 10: ar=(0.2, 0.5)

	T.50	T.100	T.600	T.1000
graphic	0.00	0.00	0.00	0.00
LB_test	0.01	0.00	0.00	0.00
LB_adjusted_H	0.00	0.00	0.00	0.00

Tabela 11: ar=(-0.7,0.1)

	T.50	T.100	T.600	T.1000
graphic	0.05	0.01	0.00	0.00
LB_test	0.36	0.17	0.00	0.00
LB_adjusted_H	0.32	0.02	0.00	0.00

Tabela 12: ma=(3,0.6)

	T.50	T.100	T.600	T.1000
graphic	0.06	0.01	0.00	0.00
LB_test	0.33	0.10	0.00	0.00
LB_adjusted_H	0.21	0.01	0.00	0.00

Tabela 13: ma=(-0.4,1.6)

	T.50	T.100	T.600	T.1000
graphic	0.22	0.13	0.00	0.00
LB_test	0.66	0.43	0.00	0.00
LB_adjusted_H	0.69	0.16	0.00	0.00

Tabela 14: ar=-0.4, ma=1

	T.50	T.100	T.600	T.1000
graphic	0.68	0.77	0.49	0.27
LB_test	0.93	0.92	0.82	0.79
LB_adjusted_H	0.96	0.95	0.63	0.61

Tabela 15: ar=-0.4, ma=2.7

	T.50	T.100	T.600	T.1000
graphic	0.00	0.00	0.00	0.00
LB_test	0.00	0.00	0.00	0.00
LB_adjusted_H	0.00	0.00	0.00	0.00

Tabela 16: ar=0.6, ma=1

	T.50	T.100	T.600	T.1000
graphic	0.00	0.00	0.00	0.00
LB_test	0.02	0.00	0.00	0.00
LB_adjusted_H	0.00	0.00	0.00	0.00

Tabela 17: ar=0.6, ma=2.7