Cwiczenie 4 - Jednorodność ciągów

Michał Marosz

19 grudnia 2015

Spis treści

Metoda różnic i ilorazów **3**

Test t-Studenta **6**

Test Kołmogorowa-Smirnowa **12**

Problem jednorodności ciągów danych jest kluczowa sprawą w przypadku każdej analizy klimatologicznej. W swoim podstawowym założeniu zmienność/wariancja w danych powinna byc odzwierciedleniem zmienności analizowanej zmiennej (np. temperatury powietrza, dobowej/miesięcznej sumy opadu). Potencjalnie, założenie to jest spełnione jedynie w przypadku, kiedy nie oddziaływuja inne czynnik, które mogą wpłynąć na pomiar np.:

- Zmiana przyrządów pomiarowych
- Zmiana lokalizacji przyrządów w obrębie stacji
- Relokacja stacji
- Zmiana obserwatora
- Zmiana otoczenia stacji

Metoda różnic i ilorazów

Najprostszą, graficzną metodą oceny jest metoda różnic i ilorazów. Metode różnic stosujemy do zmiennych ciagłych (np. temperatura powietrza), natomiast metode ilorazów w przypadku zmiennych dyskretnych (np. sumy opadu).
załadujmy dane i przeprowadzdzmy ich analizę wizualną.

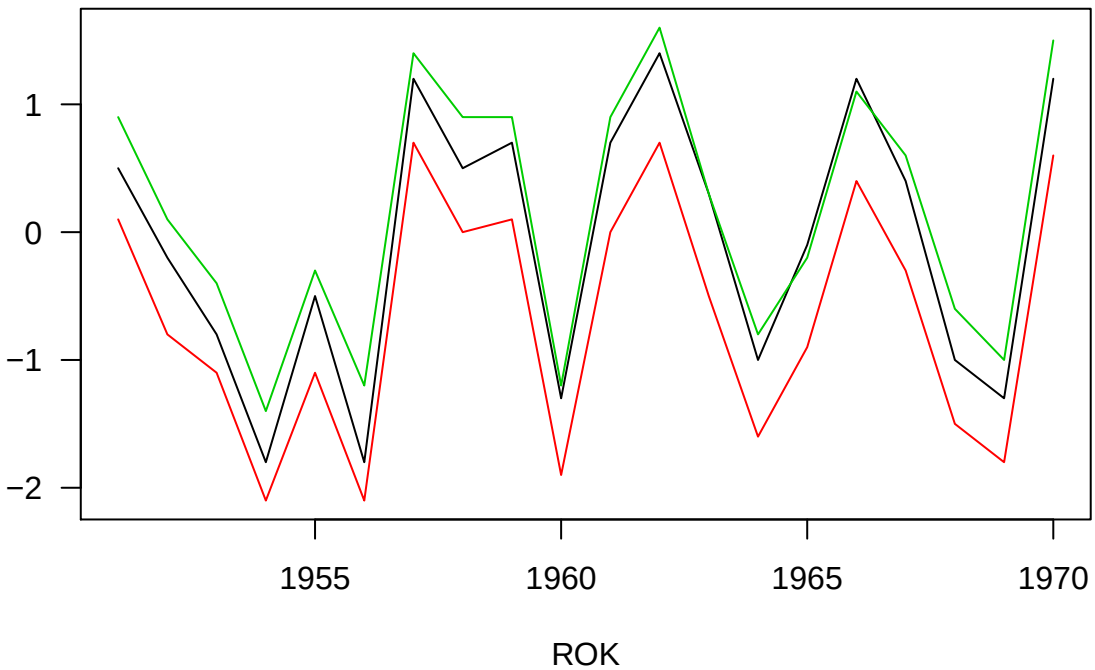
```
dane = read.table("roznice.txt", header = T)
attach(dane)
```

```
summary(dane)
```

```
##       DATA            A                 B                 C
##  Min.   :1951   Min.   :-1.800   Min.   :-2.100   Min.   :-1.400
```

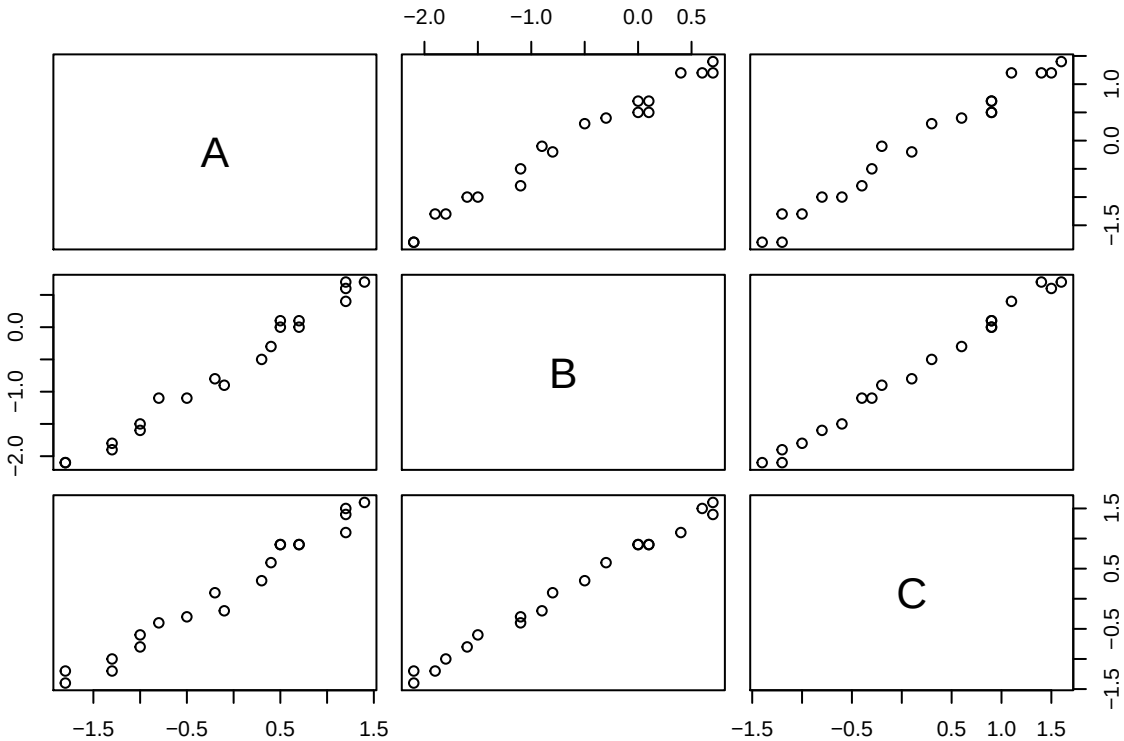
```
##  1st Qu.:1956   1st Qu.:-1.000   1st Qu.:-1.525   1st Qu.:-0.650
##  Median :1960   Median : 0.100   Median :-0.650   Median : 0.200
##  Mean   :1960   Mean   :-0.085   Mean   :-0.655   Mean   : 0.155
##  3rd Qu.:1965   3rd Qu.: 0.700   3rd Qu.: 0.100   3rd Qu.: 0.900
##  Max.   :1970   Max.   : 1.400   Max.   : 0.700   Max.   : 1.600
```

```
plot(DATA, A, type = "l", ylim = range(dane[,2:4]), xlab = "ROK", ylab="",
     las = 1)
lines(DATA, B, col = 2)
lines(DATA, C, col = 3)
```



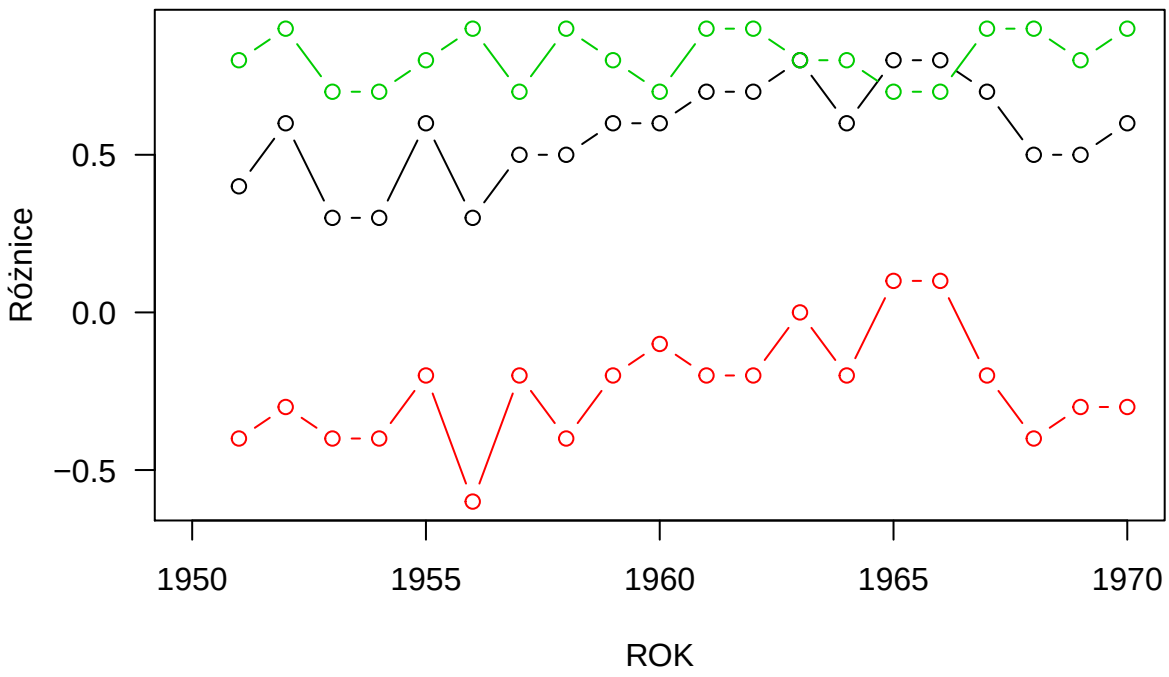
Z wykresu tego niewiele wynika poza faktem, że przebiegi analizowanej zmiennej sa z grubsza podobne a korelacje między stacjami zapewne osiągają wysokie wartości. Spójrzmy zatem

```
pairs(dane[,2:4])
```



Nie posuwa to nas zasadniczo do przodu w kwestii jednorodności. Obliczny zatem różnice między poszczególnymi stacjami i wykreślmy ponownie jako wykres liniowy.

```
plot(DATA, A-B, type = "b", ylim = range(A-B, A-C, C-B), xlim = c(1950,1970),
     xlab = "ROK", ylab="Różnice", las = 1)
lines(DATA, A-C, col = 2, type = "b")
lines(DATA, C-B, col = 3, type = "b")
```



W tym wypadku wyraźnie widać, ze w przypadku linii czerwonej oraz czarnej w okolicach roku 1960 nastepuje zmiana wartości średniej śróżnicy. Jedynie w przypadku linii zielonej takiej zmiany nie notujemy. Podstawowym założeniem metody różnic, jest to, że będa one "stałe" dla stacji położonych względnie niedaleko od siebie. Tak więc jeżeli mamy trzy stacje (a conajmniej tyle potrzebujemy, ponieważ korzystając z dwoch niestety nie będziemy w stanie stwierdzić nic poza faktem zerwania jednorodności na którejś ze stacji, nie wiadoko ktorej) sprawdzamy dla której pary nie nastapiła zmiana wartości średniej różnicy. W naszym przyadku dzieje się tak dla pary stacji C oraz B tak a różnice A-B oraz A-C są wyraźnie obciązone. Tak więc można wysnuć wniosek, że zerwanie jednorodności nastąpiło. Co więcej, możemy jednoznacznie wskazać stację, czyli A.

Test t-Studenta

Test ten jest jednym z częściej stosowanych testów parametrycznych, czyli weryfikujących hipotezy odnoszące się do paramterów. Jednakże, jego aplikacja jest obwarowana dosyć surowymi wymaganiami odnośnie zmiennych (np. normalnością obu prób). Ogólnie, testowaniu

podlega hipoteza zerowa o równości średnich. Wynik testu (na określonym poziomie istotności) pozwala nam ją przyjąc, bądź też odrzucić. Odrzucenie równoznaczne jest ze stwierdzeniem, że średnie nie pochodzą z tej samej populacji generalnej, co w praktyce oznacza, że ciąg jest niejednorodny. Oczywiście w celu zastosowania tego testu niezbędna jest znajomość przynjamniej potencjalnego "miejsca" zerwania jednorodności. W innym przypadku pozostają testy sekwencyjne, ale to już zupełnie oddzielna historia.

Dla przypadku kiedy próby są/nie są jednakowo liczne, ale **mają taką samą wariancję** stosujemy nastepujący wzor w celu obliczenia statystyki testowej t.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

gdzie:

$$s_{X_1X_2} = \sqrt{\frac{(n_1 - 1)s_{X_1}^2 + (n_2 - 1)s_{X_2}^2}{n_1 + n_2 - 2}}$$

liczba stopni swobody wynosi $n_1 + n_2 - 2$

Jeżeli wariancja **nie jest jednakowa** stosujemy *test-t Welch'a* i wówczas statystyka testowa t ma postać:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}}$$

gdzie:

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

w tym przypadku liczba stopni swobody może zostać okreslona następującym wzorem:

$$\text{d.f.} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}$$

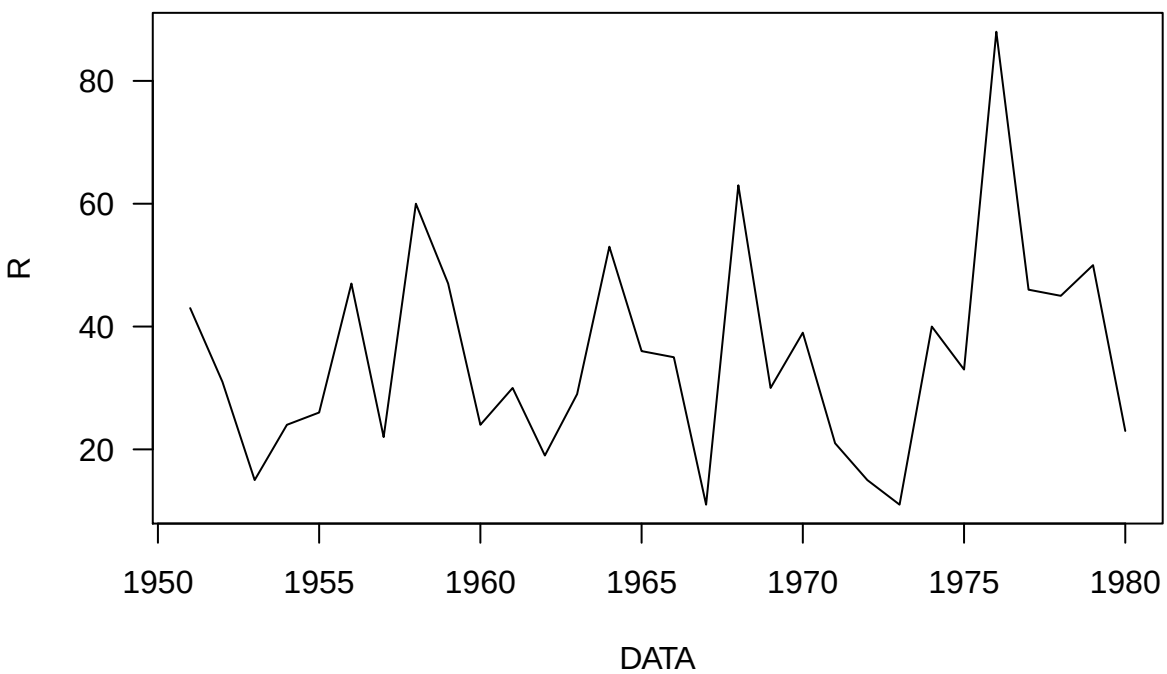
Zweryfikujmy, z wykorzystaniem testu t-studenta hipoteze o jednorodności ciągu danych Wczytujemy dane

```r
tstudent = read.table("t-student.txt", header = T)
attach(tstudent)
```

```
## Następujący obiekt został zakryty z dane:
##
##     DATA
```

wykreślamy wykres

```r
plot(DATA, R, type="l", las = 1)
```



liczba danych nie jest zatrważająca, więc możemy całą ramkę danych wyświetlić w postaci tekstowej

```r
tstudent
```

```
##    DATA  R TYPE
## 1  1951 43    A
```

```
## 2  1952 31    A
## 3  1953 15    A
## 4  1954 24    A
## 5  1955 26    A
## 6  1956 47    A
## 7  1957 22    A
## 8  1958 60    A
## 9  1959 47    A
## 10 1960 24    A
## 11 1961 30    A
## 12 1962 19    A
## 13 1963 29    A
## 14 1964 53    A
## 15 1965 36    A
## 16 1966 35    A
## 17 1967 11    A
## 18 1968 63    A
## 19 1969 30    A
## 20 1970 39    A
## 21 1971 21    B
## 22 1972 15    B
## 23 1973 11    B
## 24 1974 40    B
## 25 1975 33    B
## 26 1976 88    B
## 27 1977 46    B
## 28 1978 45    B
```

```
## 29 1979 50    B
## 30 1980 23    B
```

Podejrzewamy zerwanie jednorodności po roku 1970 i jednoczesnie wariancja w analizowanych podokresach wydaje sie być różna, dlatego też wykorzystamy test w wersji dla danych niezależnych, o różnej liczebności prób oraz nierównej wariancji. Wszystkie te założenia możemy uwzględnić operując argumentami funkcji odpowiedzialnej w R za obliczenie statystyki testowej t i określenie wartości p-value. Podejrzyjmy najpierw *help* do funkcji *t.test*.

```
?t.test
```

W celu zasosowania go do naszych danych kod przyjmie następująca postać:

```
t.test(R[which(TYPE=="A")], R[which(TYPE == "B")],
       alternative = "two.sided",
       paired = FALSE, var.equal = FALSE, conf.level = 0.95       )
```

```
##
##  Welch Two Sample t-test
##
## data:  R[which(TYPE == "A")] and R[which(TYPE == "B")]
## t = -0.3845, df = 12.789, p-value = 0.7069
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -19.88422  13.88422
## sample estimates:
## mean of x mean of y
##      34.2      37.2
```

Wartość statystyki testowej t nie mieści sie w przedziale krytycznym, dlatego też nie ma podstaw do odrzucenia hipotezy zerowej o równości średnich.

Przyprowadzmy jescze jeden test na wygenerowanych danych o rozkładzie normalnym, ale tym razem *wymusimy* istnienie rożnicy.

```
x = rnorm(100, 2.5, 1)
y = rnorm(100, 3, 1)
```

Założyliśmy równą wariancję, więc posłużymy sie odrobine zmodyfikowanym kodem

```
t.test(x, y,
       alternative = "two.sided",
       paired = FALSE, var.equal = TRUE, conf.level = 0.95       )
```

```
##
##  Two Sample t-test
##
## data:  x and y
## t = -5.6424, df = 198, p-value = 5.741e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.9576317 -0.4616098
## sample estimates:
## mean of x mean of y
##  2.380454  3.090075
```

Test Kołmogorowa-Smirnowa

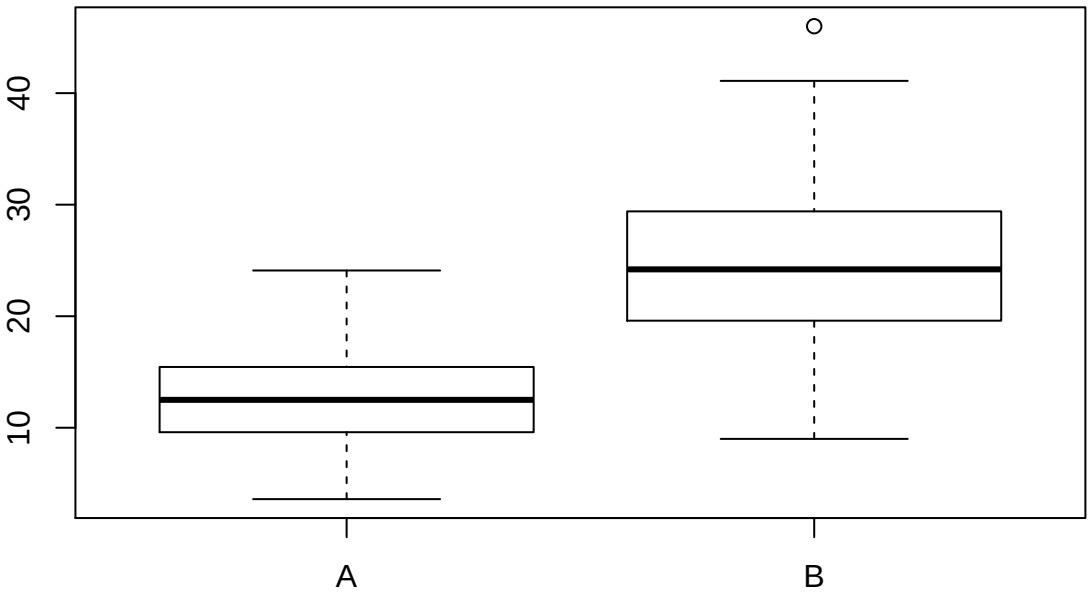
Test t-studenta pozwala na weryfikacje hipotezy o równości średnich, jednak w zasadzie nic nie mówi nam o tym, jak zachowują sie rozkłady. W tym celu posłużymy się testem Kołmogorowa-Smirnova, który pozwala na porównanie dystrybunat dwóch prób. Jeżeli nie będzie podstaw do odrzucenia hipotezy o ich ożsamości, wówczas możemy uznać, ze dane pochodzą z tej samej populacji generalnej, a więc ciąg danych jest jednorodny

Wykorzystajmy w tym celu kolejne dane

```
ks = read.table("ks.txt", header = T)
attach(ks)
```

```
## Następujące obiekty zostały zakryte z dane:
##
##     A, B
```

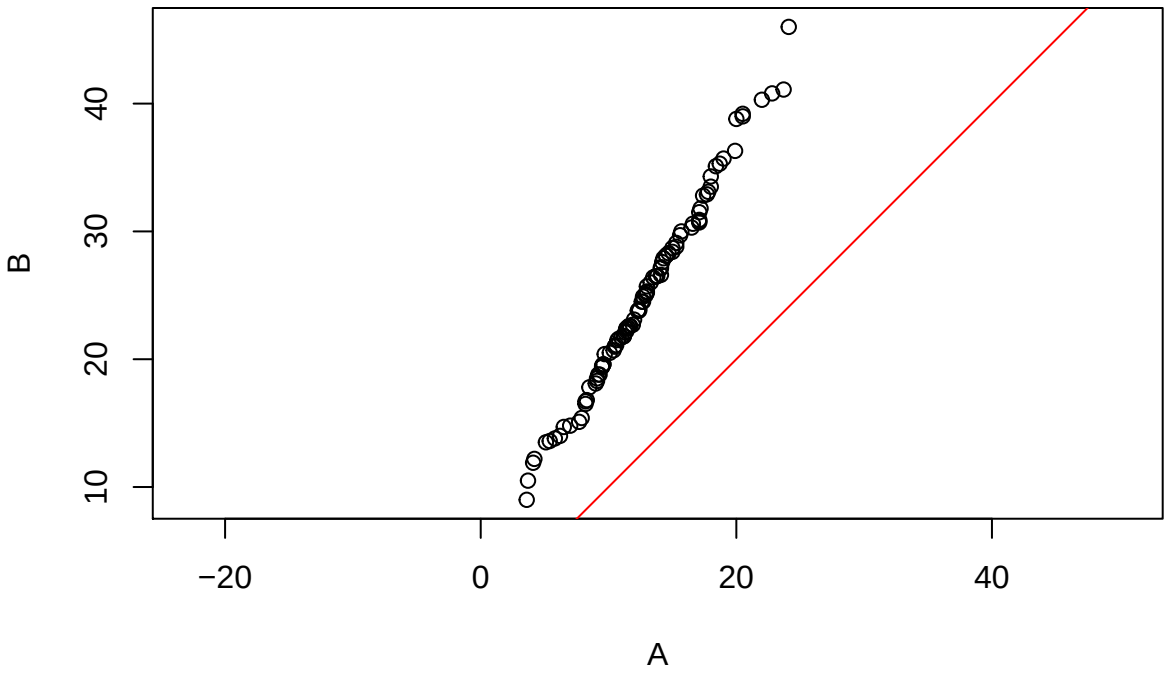
```
boxplot(ks)
```



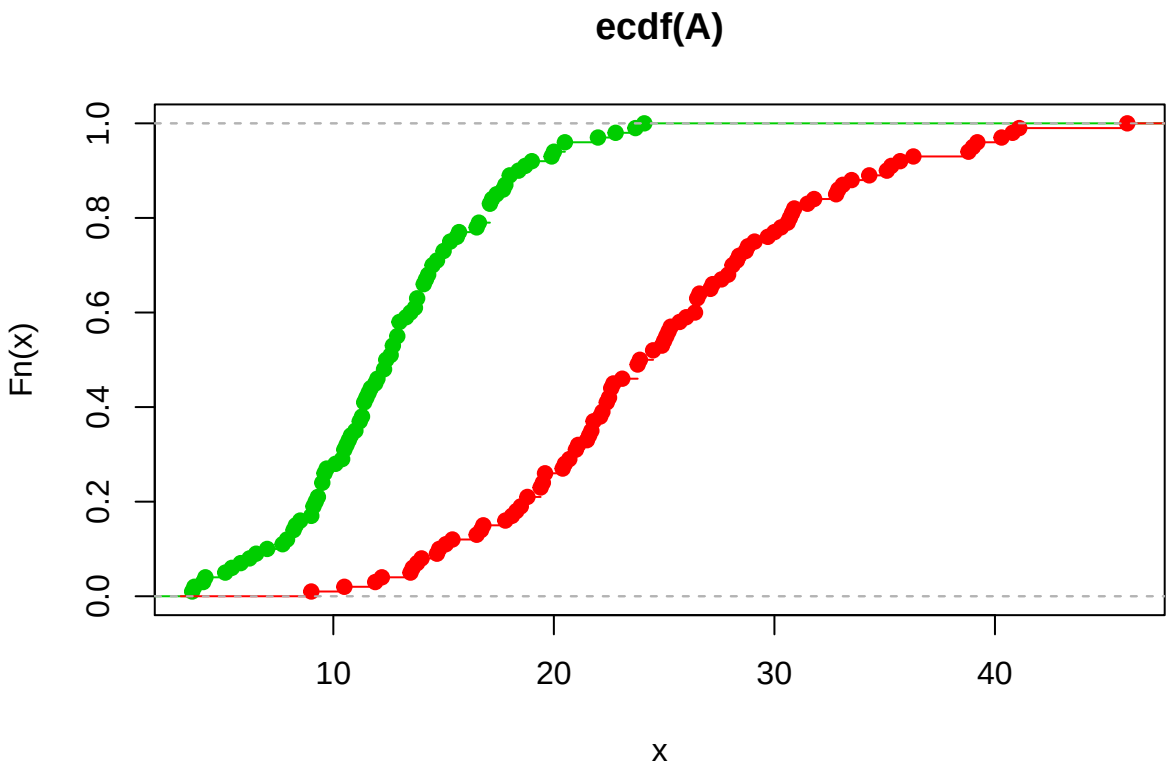
Już na pierwszy rzut oka widac, ze dane różnią sie znacząco od siebie. To na ile jest to istotne stat-

stycznie należy pozostawic formalnemu testowaniu. Zanim przejdziemy do przeprowadzenia testu KS wykreślmy dystrybuanty tempiryczne obu prób oraz wykres qq

```
qqplot(A, B, asp=1)
abline(0,1, col = 2)
```




```
plot(ecdf(A), xlim = range(A, B), col = 3)
plot(ecdf(B), add = T, col = 2)
```



Test

KS jest skonstruowany w ten sposób, że okresla największą rożnicę miedzy dystrybuantami empirycznymi, jeżeli przekroczy ona wartości krytyczną dla okreslonego poziomu istotności to wówczas sa podstawy d odrzucenia hipotezy zerowej o tożsamości dystrybuant.

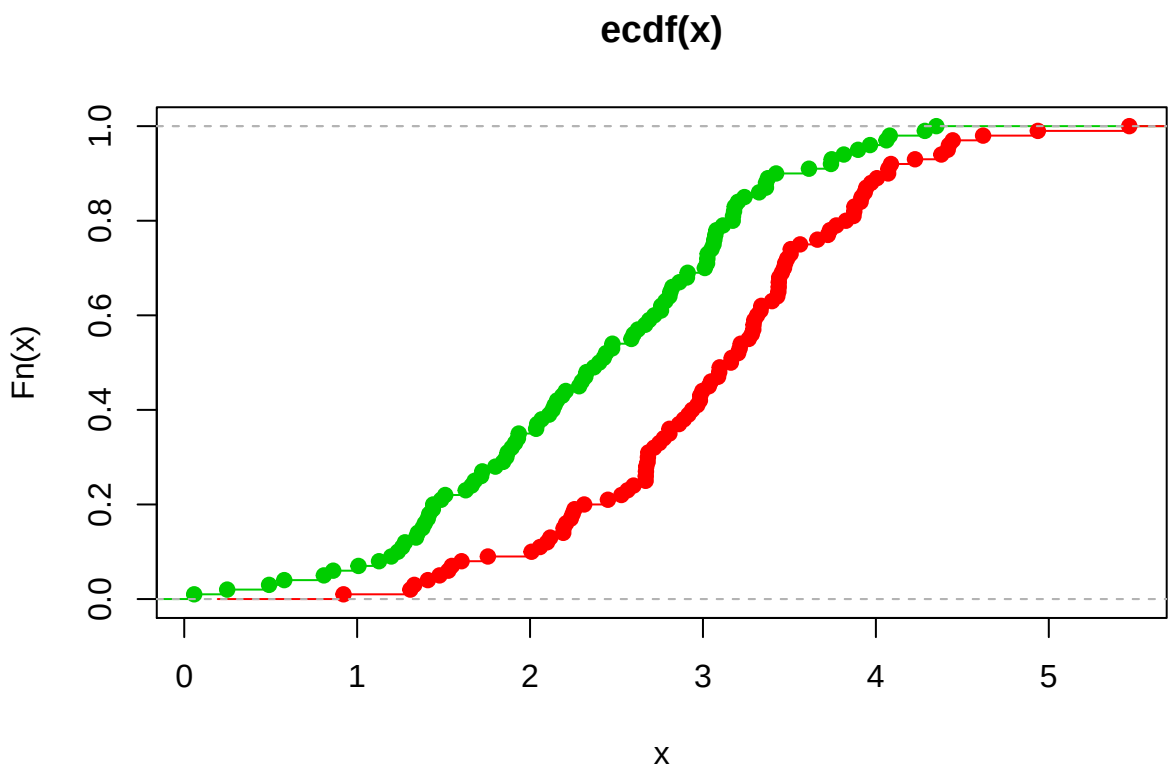
```
ks.test(A, B)
```

```
## Warning in ks.test(A, B): p-value will be approximate in the presence of
## ties

##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  A and B
## D = 0.73, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Dla porównania, przetestujmy nasze zmienne x oraz y z poprzedniego przykładu:

```r
plot(ecdf(x), xlim = range(x, y), col = 3)
plot(ecdf(y), add = T, col = 2)
```



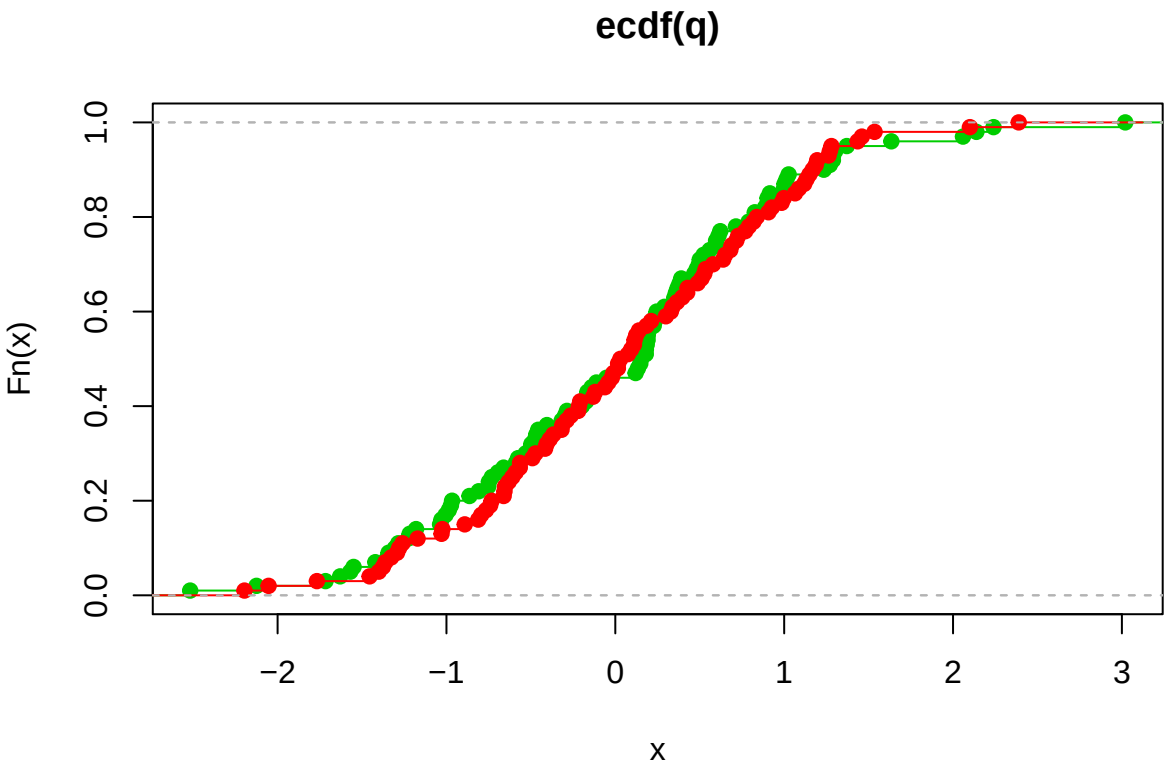

```r
ks.test(x, y)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  x and y
## D = 0.34, p-value = 1.908e-05
## alternative hypothesis: two-sided
```

Dla losowo wygnerowanych zmiennych o takich samych paramterach będzie to wyglądało nastepująco

```r
q = rnorm(100, 0, 1)
w = rnorm(100, 0 , 1)
```

```r
plot(ecdf(q), xlim = range(q, w), col = 3)
plot(ecdf(w), add = T, col = 2)
```




```r
ks.test(q, w)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  q and w
## D = 0.08, p-value = 0.9062
## alternative hypothesis: two-sided
```