Projekt ***„Nowa oferta edukacyjna Uniwersytetu Wrocławskiego odpowiedzią na współczesne potrzeby rynku pracy i gospodarki opartej na wiedzy”***

Dane:
Eksploracja (mining)

Problemy:
Jedna zmienna

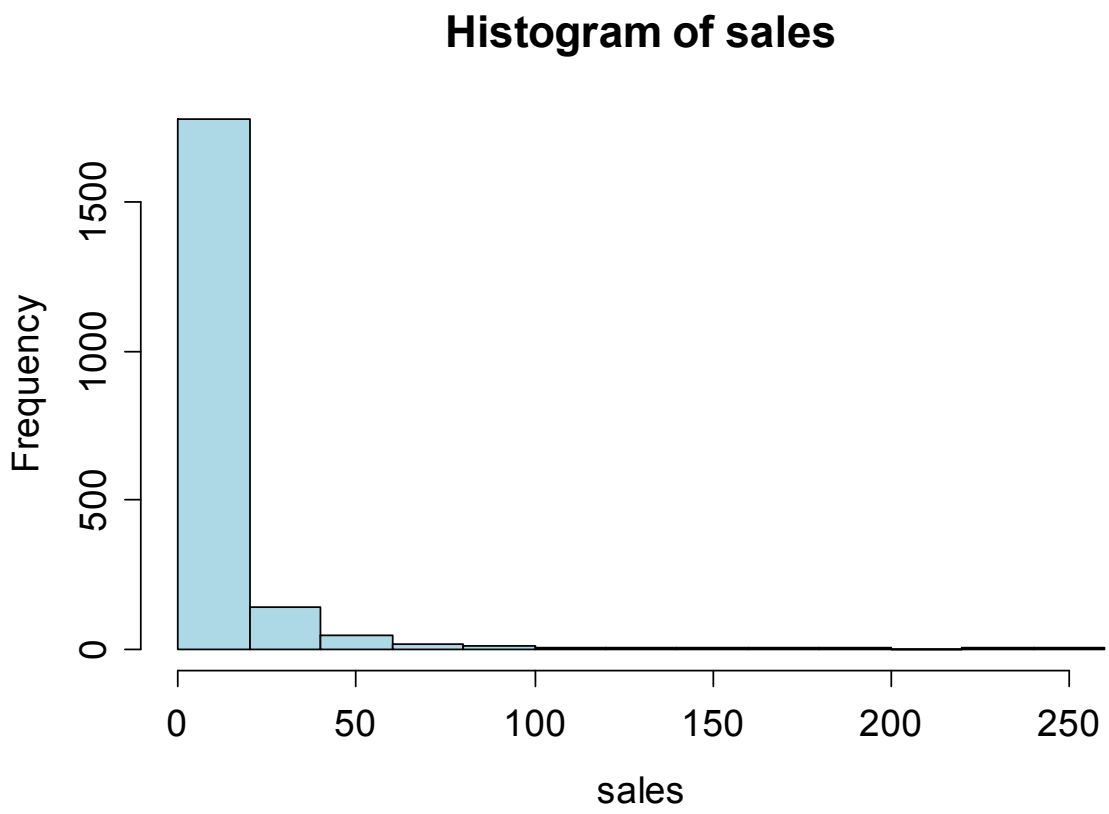
2000 najwi ększych spółek światowych z 2004 (*Forbes Magazine*)

```
data("Forbes2000",package="HSAUR2")
attach(Forbes2000)
```

Czy można stosować klasyczne teorie?

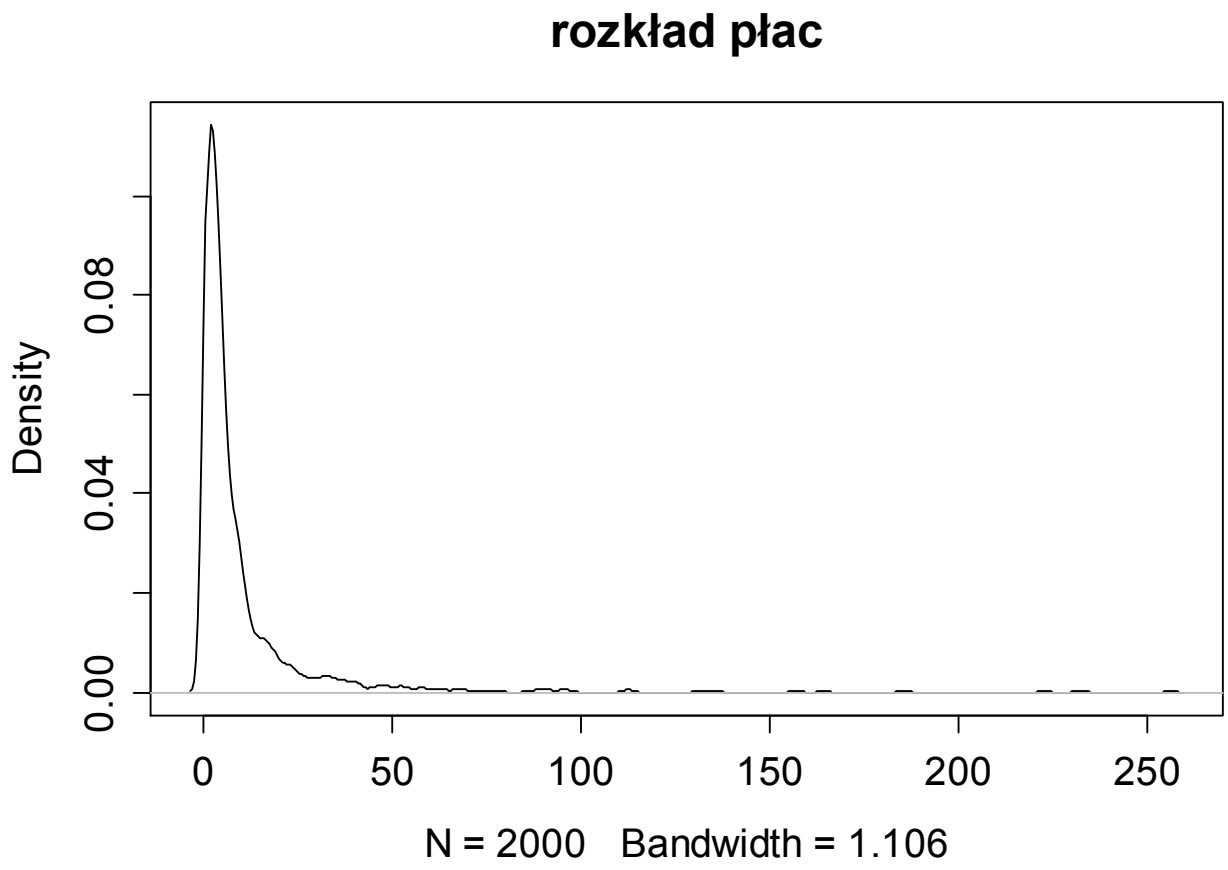
Zgodność ze standardowymi rozkładami

```
hist(sales,col="lightblue")
```



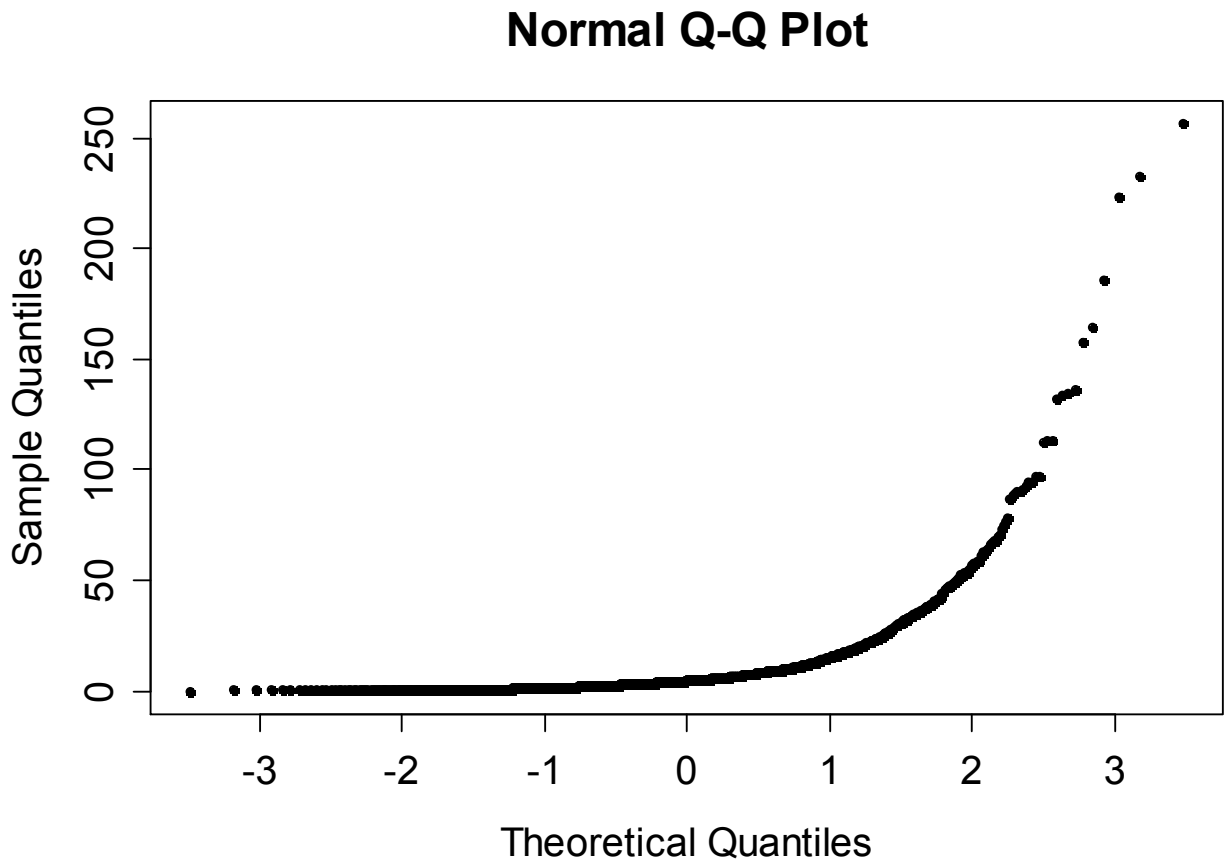
Dobór liczby klas: Sturges: $\lceil \log_2 n + 1 \rceil$

```
plot(density(sales),main="rozkład płac")
```

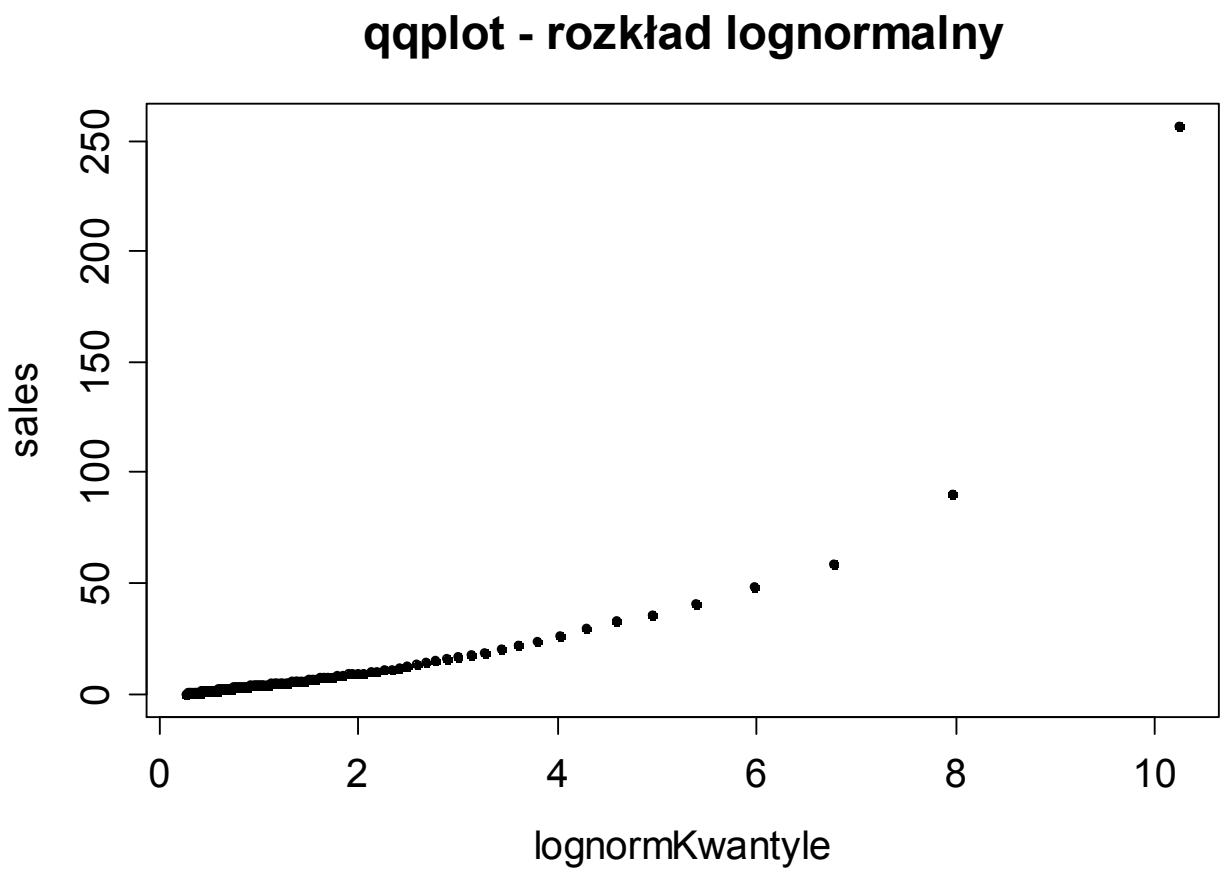


,

qqnorm(sales,pch=20)



```
kwantyle <- seq(0.1,0.99,length.out=100)
lognormKwantyle <- qlnorm(kwantyle)
qqplot(lognormKwantyle,sales,pch=20,main="qqplot - rozkład lognormalny")
```




```
kwantyleSales <- quantile(sales,kwantyle)
summary(lm(kwantyleSales~lognormKwantyle))
```

```
Call:
lm(formula = kwantyleSales ~ lognormKwantyle)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9564 -1.4237  0.0224  1.4422 13.5702

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -3.6361     0.3091  -11.76   <2e-16 ***
lognormKwantyle   7.8287     0.1291   60.66   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.179 on 98 degrees of freedom
Multiple R-squared: 0.9741,    Adjusted R-squared: 0.9738
F-statistic:  3680 on 1 and 98 DF,  p-value: < 2.2e-16
```

Dopasowywanie rozkładów prawdopodobieństwa
http://www.statmethods.net/advgraphs/probability.html

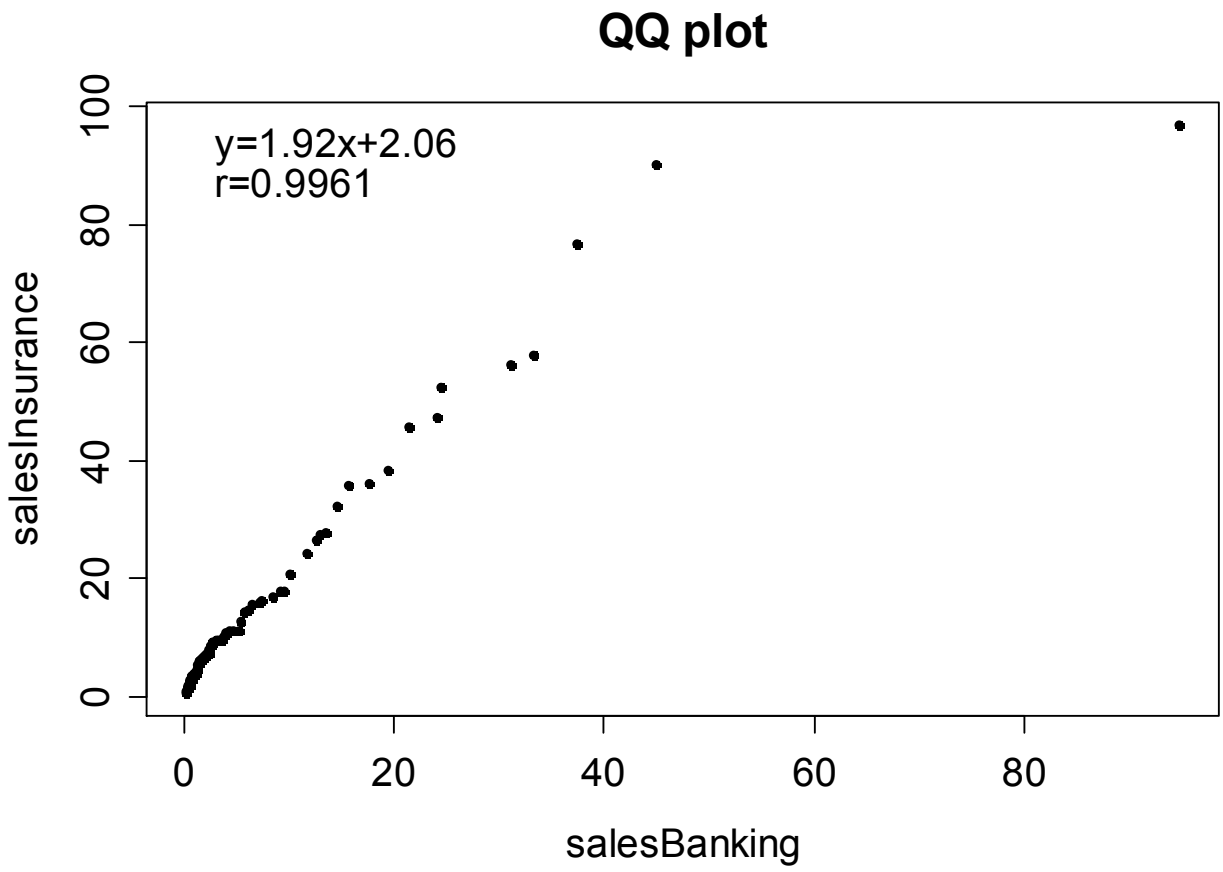
```
salesBanking <- sales[category=="Banking"]
salesInsurance <- sales[category=="Insurance"]
length(salesBanking);length(salesInsurance)
```

```
[1] 313
[1] 112
```

```
qqplot(salesBanking,salesInsurance,pch=20,main="QQ plot")
summary(lm(quantile(salesInsurance,kwantyle)~
                quantile(salesBanking,kwantyle)))
```

```
lm(formula = quantile(salesInsurance, kwantyle) ~ quantile(salesBanking,
    kwantyle))

Residuals:
    Min      1Q  Median      3Q     Max
-7.4786 -0.5699 -0.1926  0.8292  3.5666

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                     2.06342    0.16844   12.25   <2e-16 ***
quantile(salesBanking, kwantyle)  1.91599    0.01721  111.36   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.416 on 98 degrees of freedom
Multiple R-squared: 0.9922,    Adjusted R-squared: 0.9921
F-statistic: 1.24e+04 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
legend("topleft",legend=c("y=1.92x+2.06","r=0.9961"),bty="n")
```



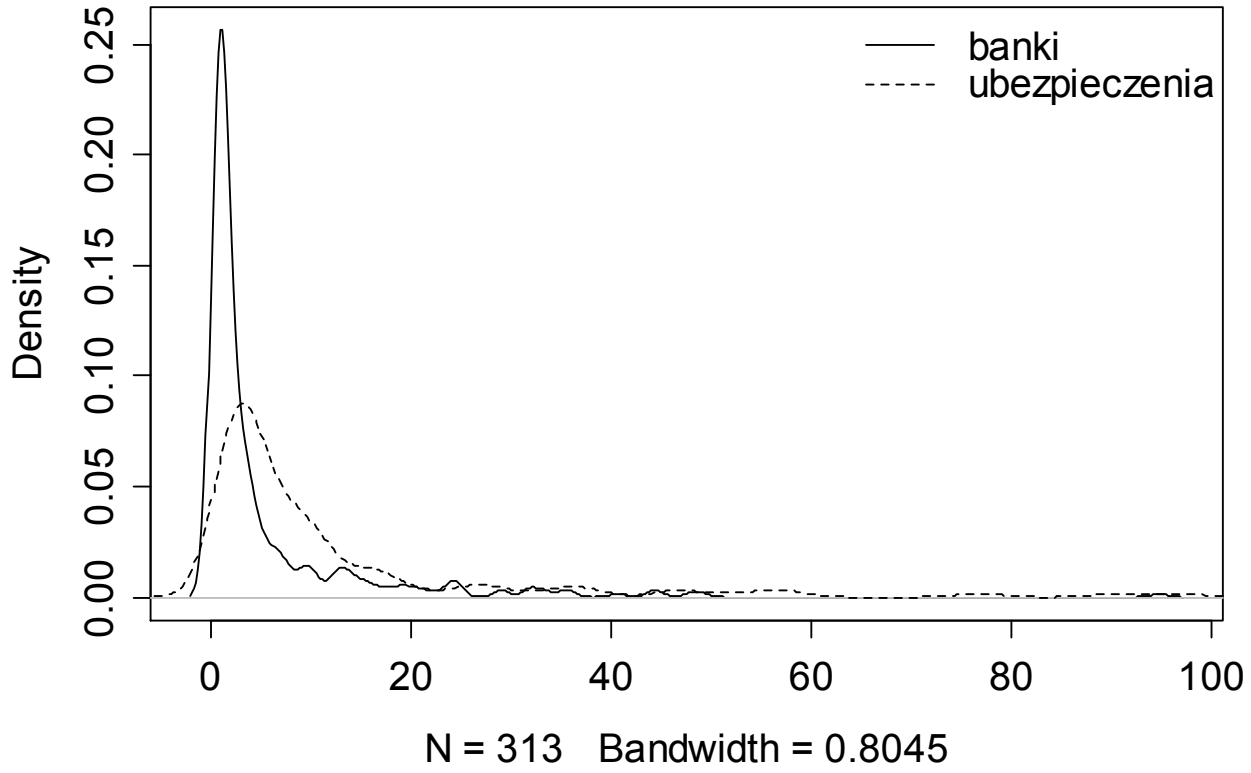

```
summary(salesBanking)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.300   0.700   1.440   5.313   4.480  94.710
```

```
summary(salesInsurance)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.680   2.925   5.305  11.960  11.150  96.880
```

```
c(IQR(salesBanking),IQR(salesInsurance),
  IQR(salesInsurance)/IQR(salesBanking))
```

```
3.78 8.22 2.18
```

```
dBanking <- density(salesBanking)
dInsurance <- density(salesInsurance)
plot(dBanking,lty=1,main="")
lines(dInsurance,lty=2)
legend("topright",lty=1:2,legend=c("banki","ubezpieczenia"),bty="n")
```

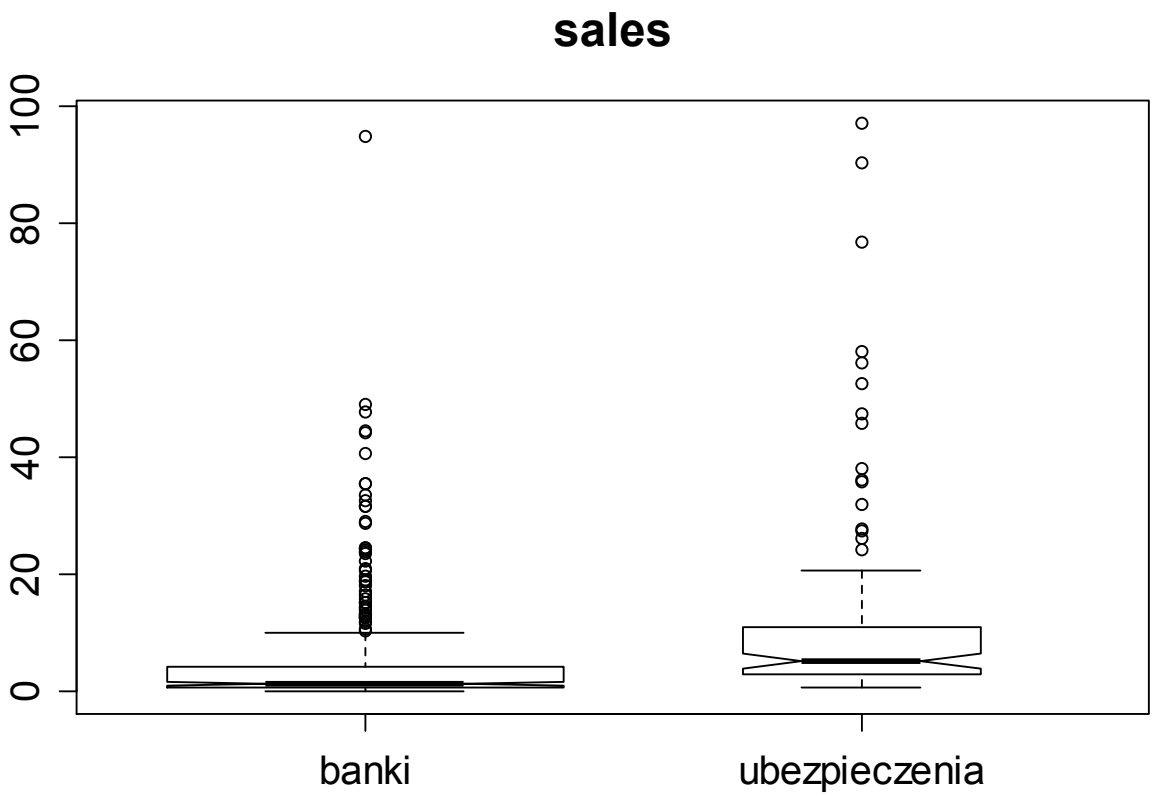



```
simpleTestDifference <- function(q1x,q2x,q3x,q1y,q2y,q3y) {
  (q2x-q2y)/(max(q3x,q3y)-min(q1x,q1y))}

simpleTestDifference(0.7,1.44,4.48,2.925,5.305,11.15)
```

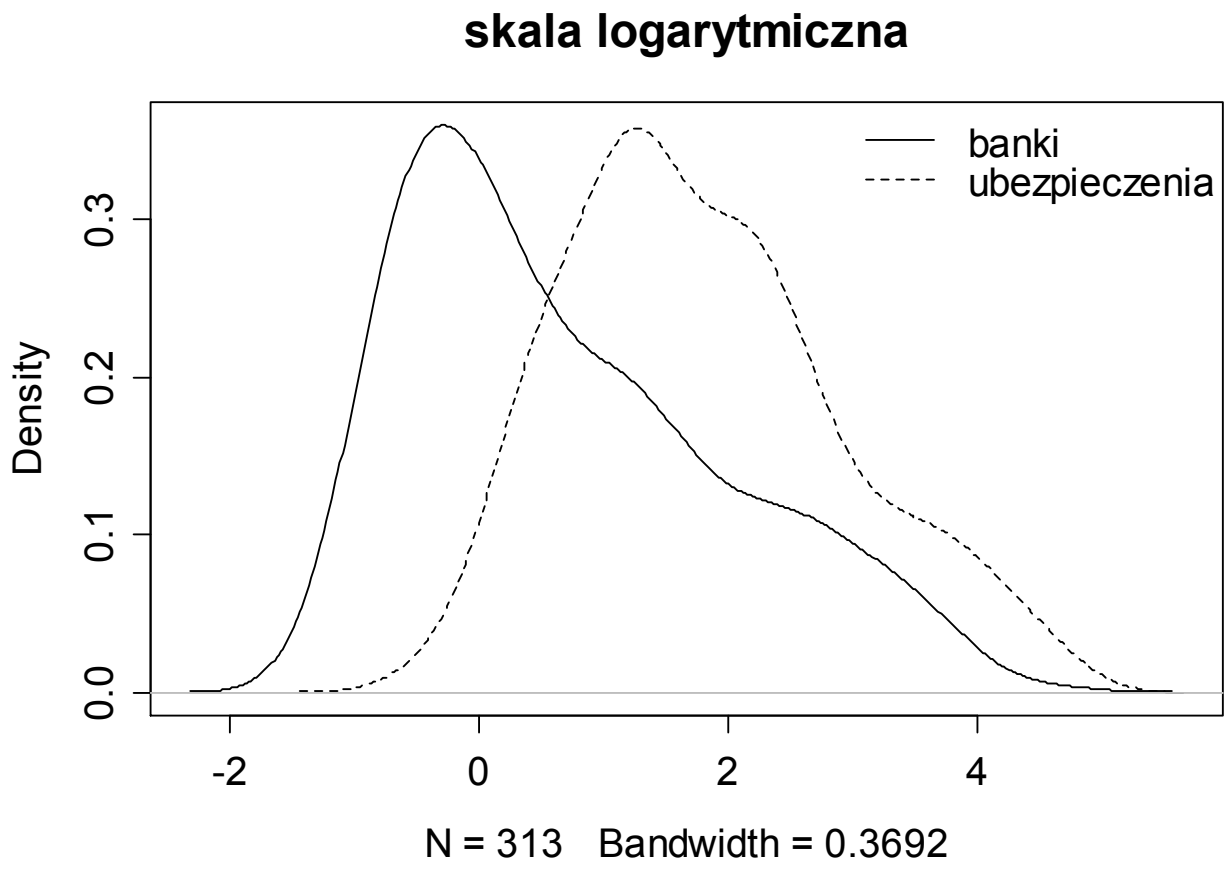
```
-0.37
```

Jeżeli różnica (co do wartości bezwzględnej) jest większa od 1/3 (próba około 40 elementów) lub 1/5 (próba około 100 elementów) to różnica jest istotna

```
boxplot(salesBanking,salesInsurance,varwidth=T,
        names=c("banki","ubezpieczenia"),main="sales",notch=T)
```



```
dBankingLog <- density(log(salesBanking))
dInsuranceLog <- density(log(salesInsurance))
plot(dBankingLog,lty=1,main="skala logarytmiczna")
lines(dInsuranceLog,lty=2)
legend("topright",lty=1:2,legend=c("banki","ubezpieczenia"),bty="n")
```

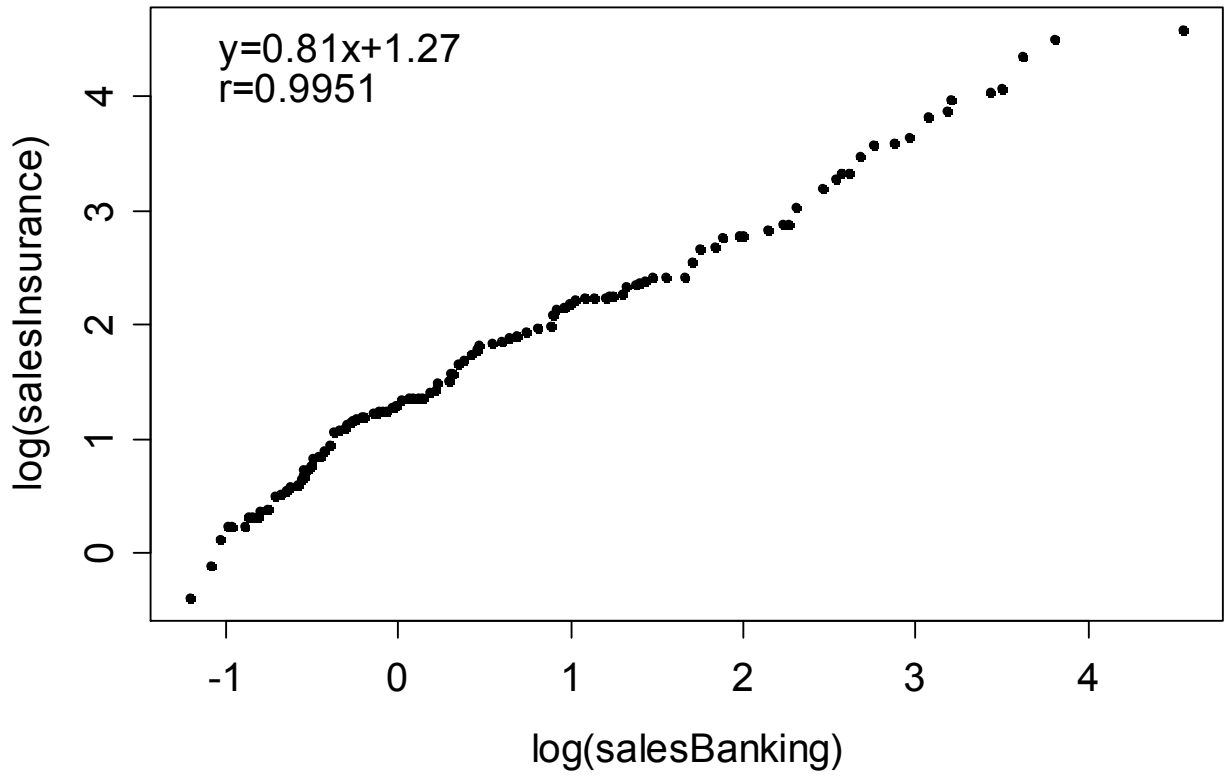


summary(lm(quantile(log(salesInsurance),kwantyle)~
quantile(log(salesBanking),kwantyle)))

```
Call:
lm(formula = quantile(log(salesInsurance), kwantyle) ~
quantile(log(salesBanking),
    kwantyle))

Residuals:
      Min        1Q    Median        3Q       Max
-0.223380 -0.054129  0.006433  0.082437  0.160871

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       1.268528   0.011918   106.4   <2e-16 ***
quantile(log(salesBanking), kwantyle) 0.809984   0.008082   100.2   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09745 on 98 degrees of freedom
Multiple R-squared: 0.9903,    Adjusted R-squared: 0.9902
F-statistic: 1.004e+04 on 1 and 98 DF,  p-value: < 2.2e-16
```

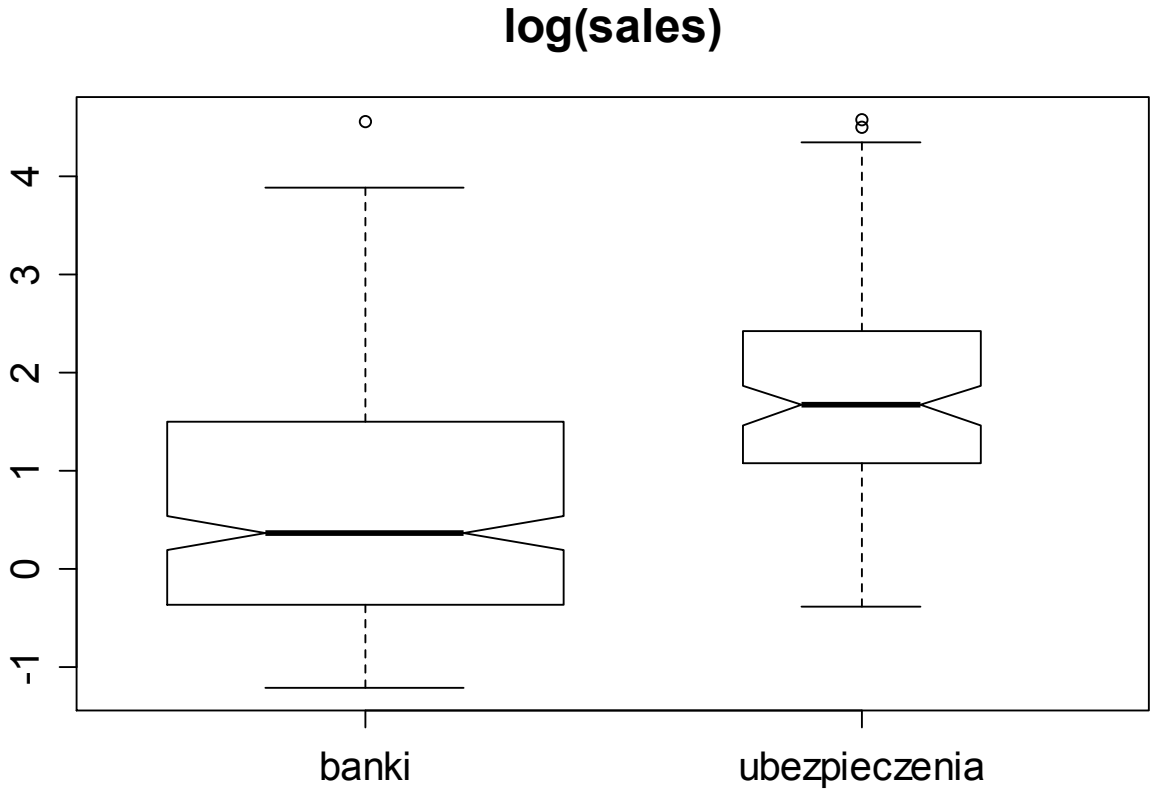
*

QQ plot skala logarytmiczna



Wniosek: wyraźnie widoczna tożsamość wykładów (z dokładnością do wartości typowej i skali)
Uwaga! wsp. kier 0.81 = odchylenia są prawie równe

```
boxplot(log(salesBanking),log(salesInsurance),varwidth=T,
        names=c("banki","ubezpieczenia"),main="log(sales)",notch=T)
```

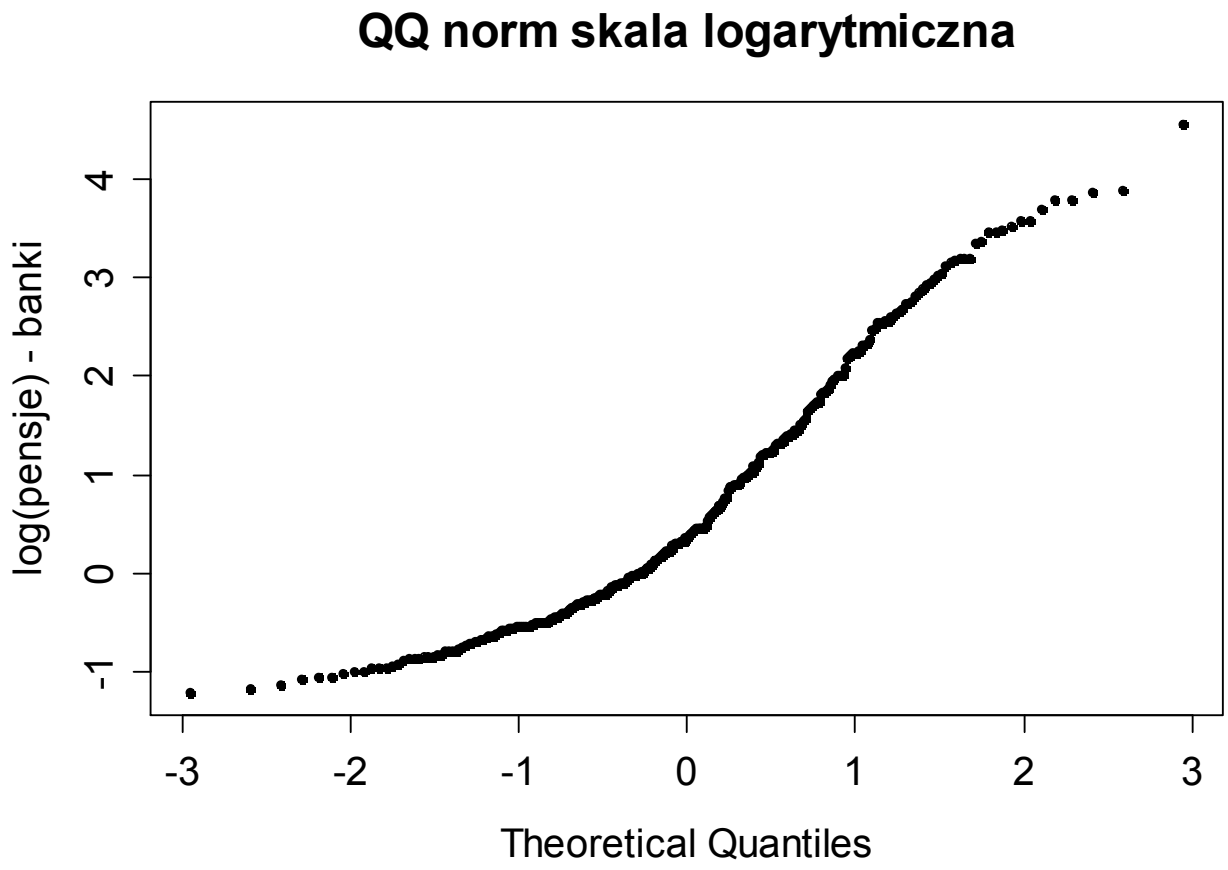



```
summary(log(salesBanking))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.2040 -0.3567  0.3646  0.7025  1.5000  4.5510
summary(log(salesInsurance))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.3857  1.0730  1.6680  1.8070  2.4110  4.5730
```

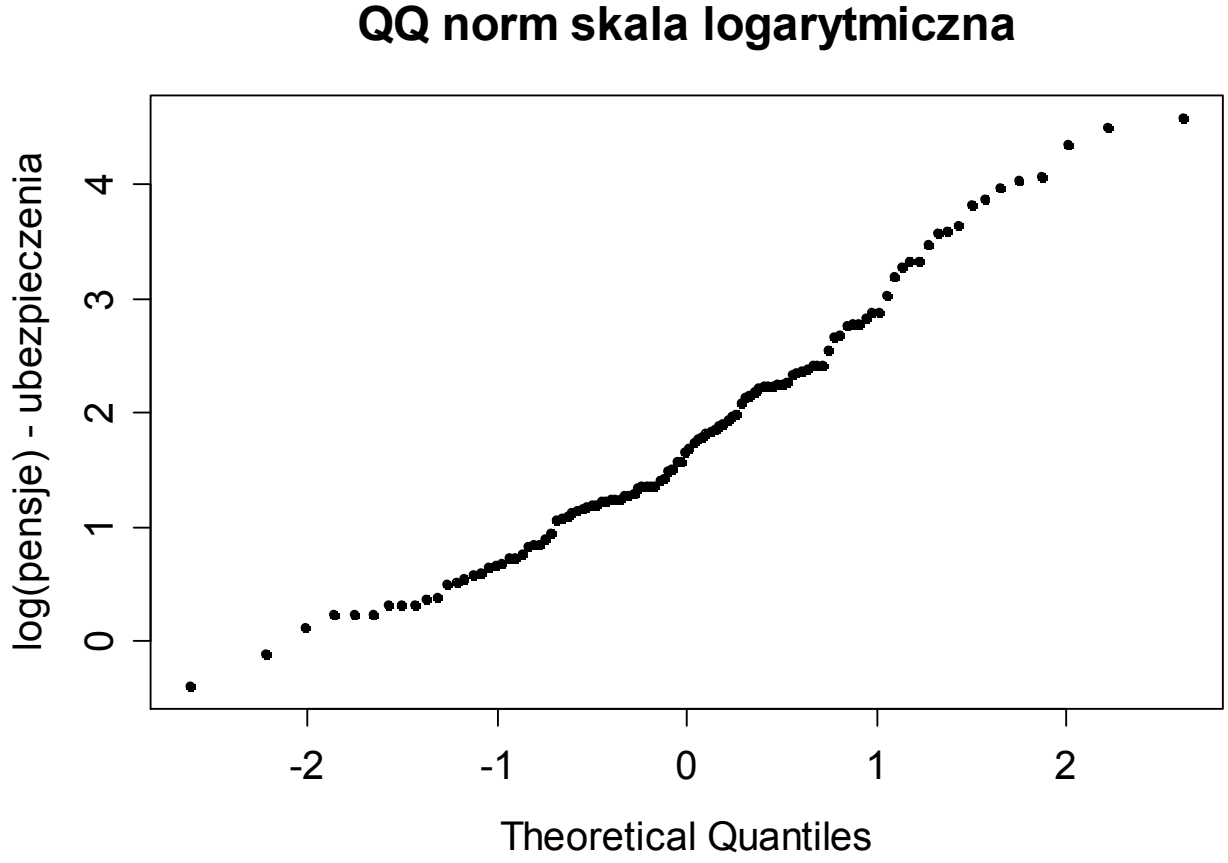
```
simpleTestDifference(-.3567,.3646,1.5,1.073,1.668,2.411)
[1] -0.47
```

```
qqnorm(log(salesBanking),pch=20,
      main="QQ norm skala logarytmiczna",ylab="log(pensje) - banki")
qqnorm(log(salesInsurance),pch=20,
      main="QQ norm skala logarytmiczna",ylab="log(pensje) -
      ubezpieczenia")
```

*



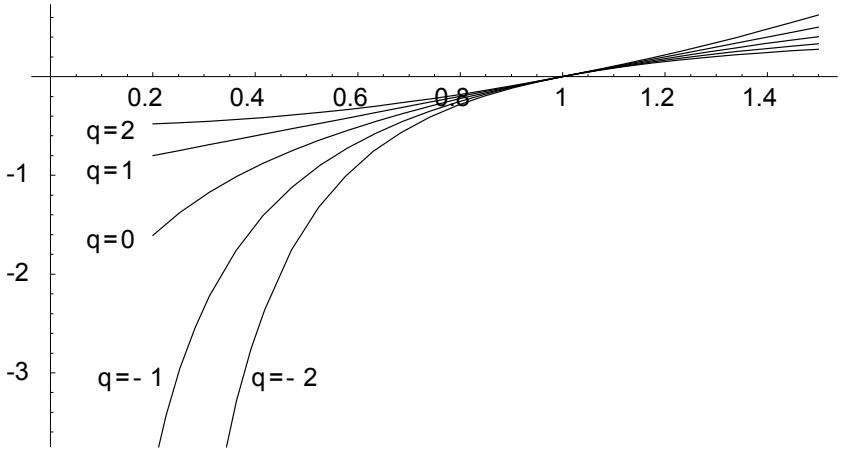
*


QQ norm skala logarytmiczna
log(pensje) - ubezpieczenia
0
1
2
3
4
-2
-1
0
1
2
Theoretical Quantiles

*

Rodzina przekształceń Boxa-Coxa

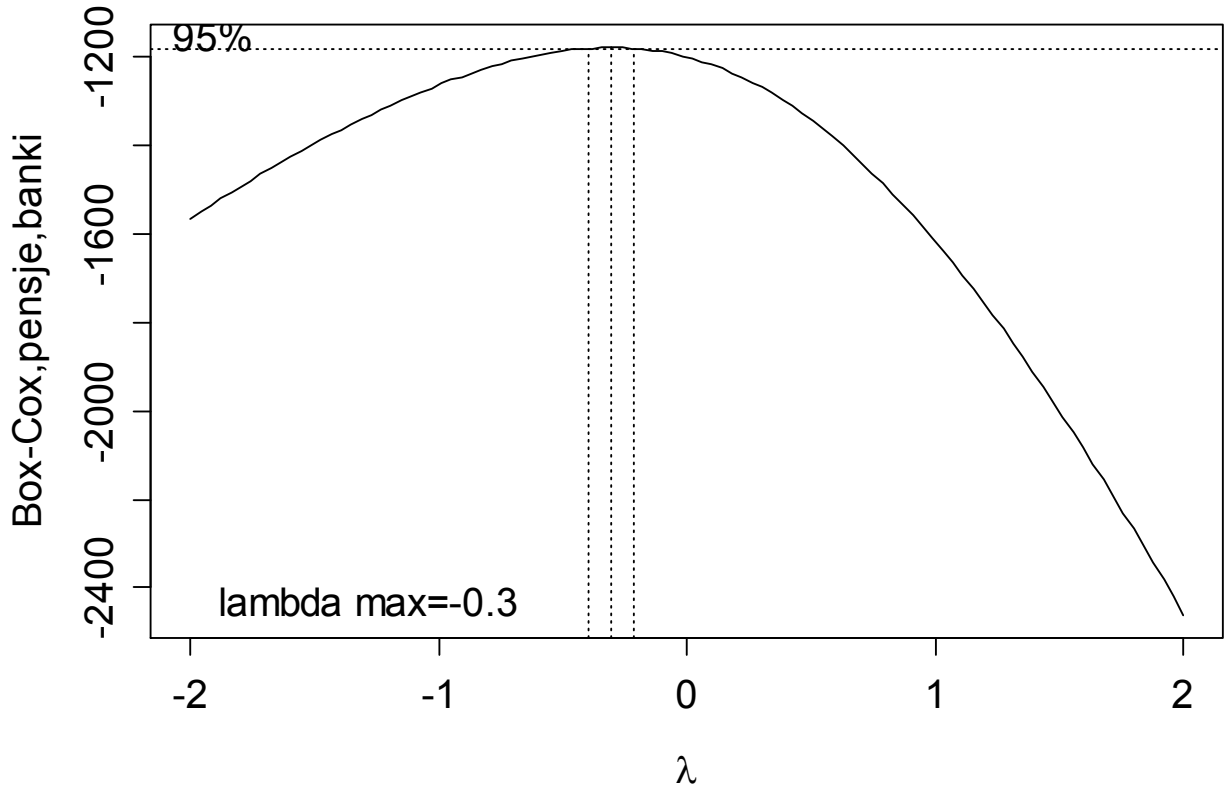
$$h_q(x) = \begin{cases} \dfrac{x^q - 1}{q} & q \neq 0 \\ \ln(x) & q = 0 \end{cases}$$



*

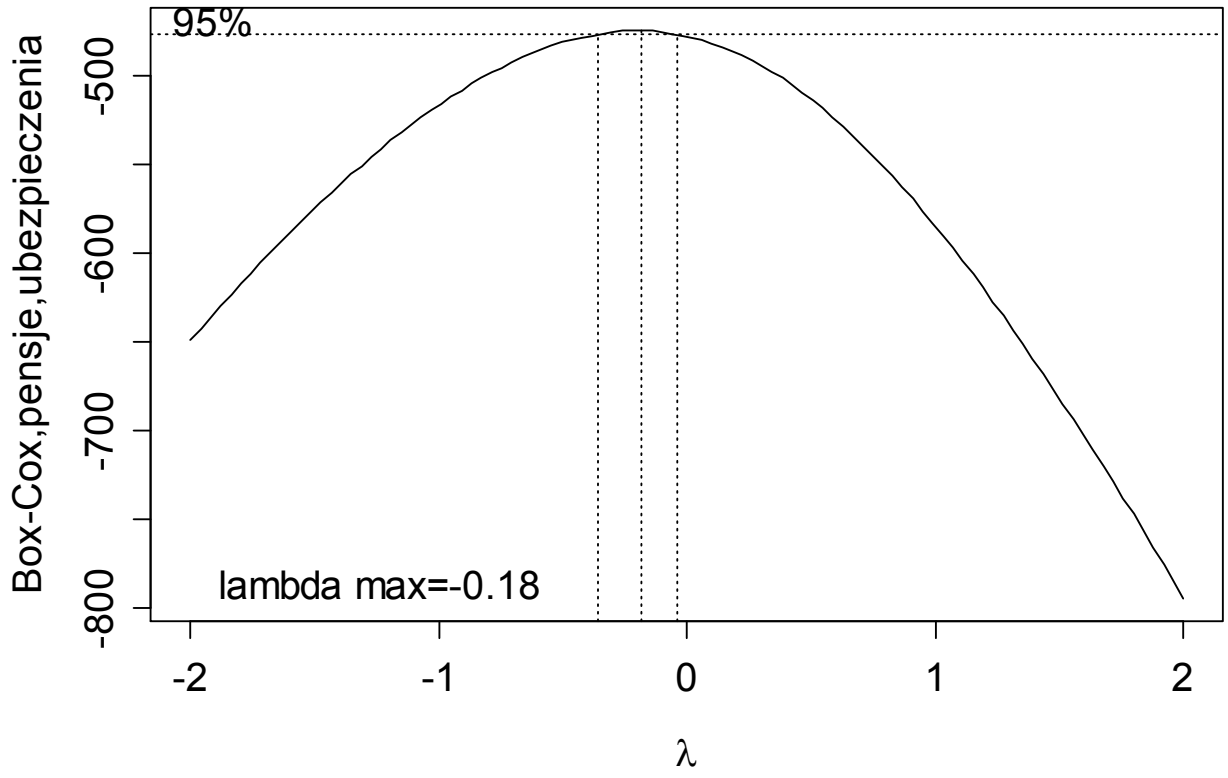
```
library(MASS)
boxcox(salesBanking~salesBanking,ylab="Box-Cox,pensje,banki")
legend("bottomleft",legend="lambda max=-0.3",bty="n")
bb <- boxcox(salesBanking~salesBanking)
bb$x[which.max(bb$y)]
```

```
-0.30
```

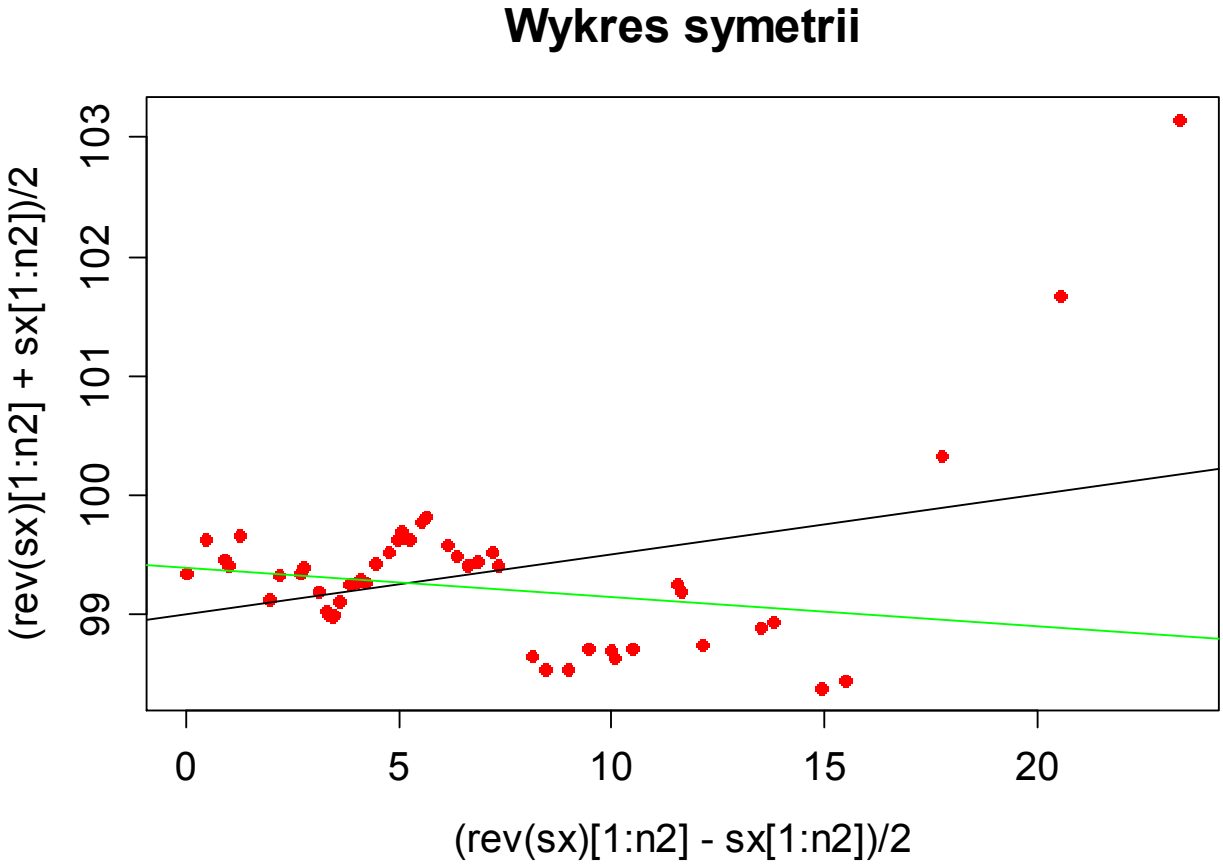


```
boxcox(salesInsurance~salesInsurance,plotit=F)
boxcox(salesInsurance~salesInsurance,ylab="Box-Cox,pensje,ubezpieczenia")
legend("bottomleft",legend="lambda max=-0.18",bty="n")
bl <- boxcox(salesInsurance~salesInsurance)
bl$x[which.max(bl$y)]
```

```
-0.18
```

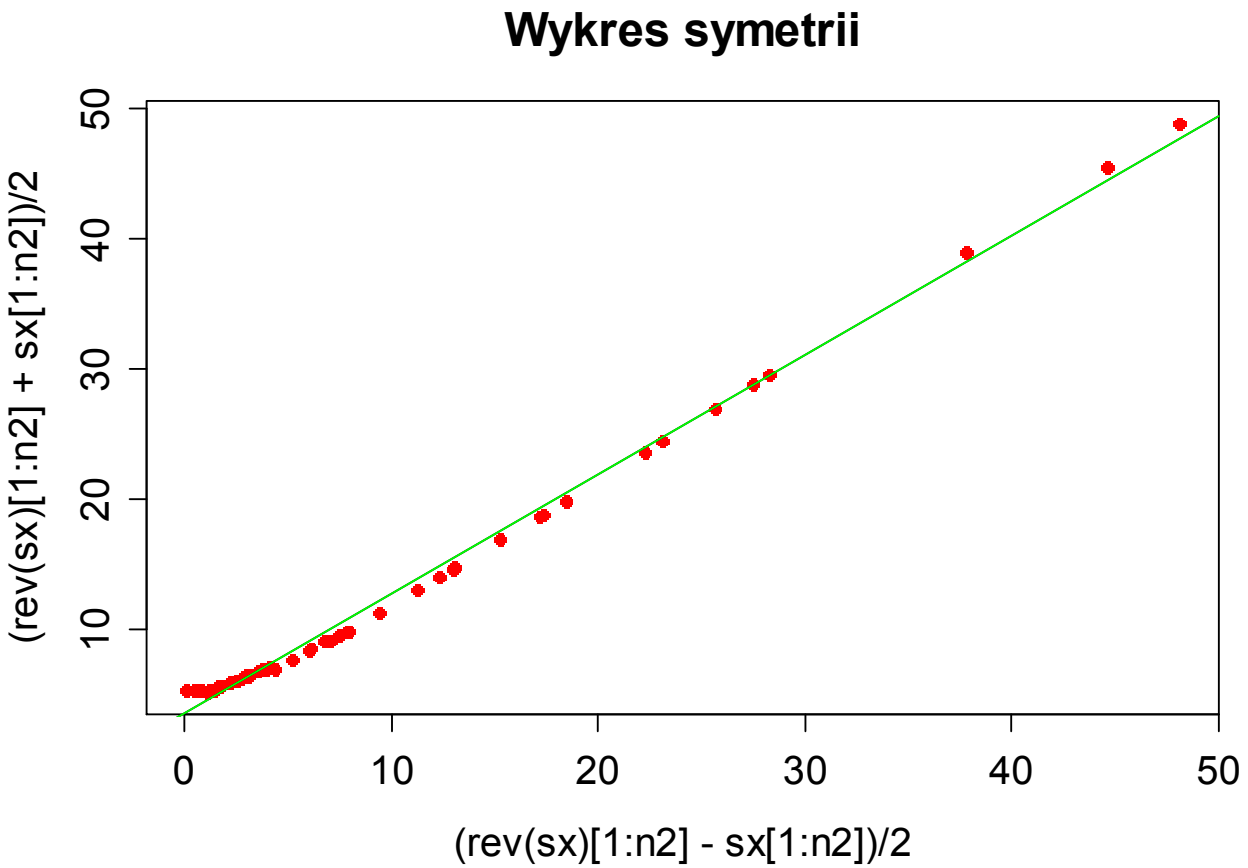


```
normalny <- rnorm(101,mean=100,sd=10)
prostaSymetrii2AD(normalny)
```



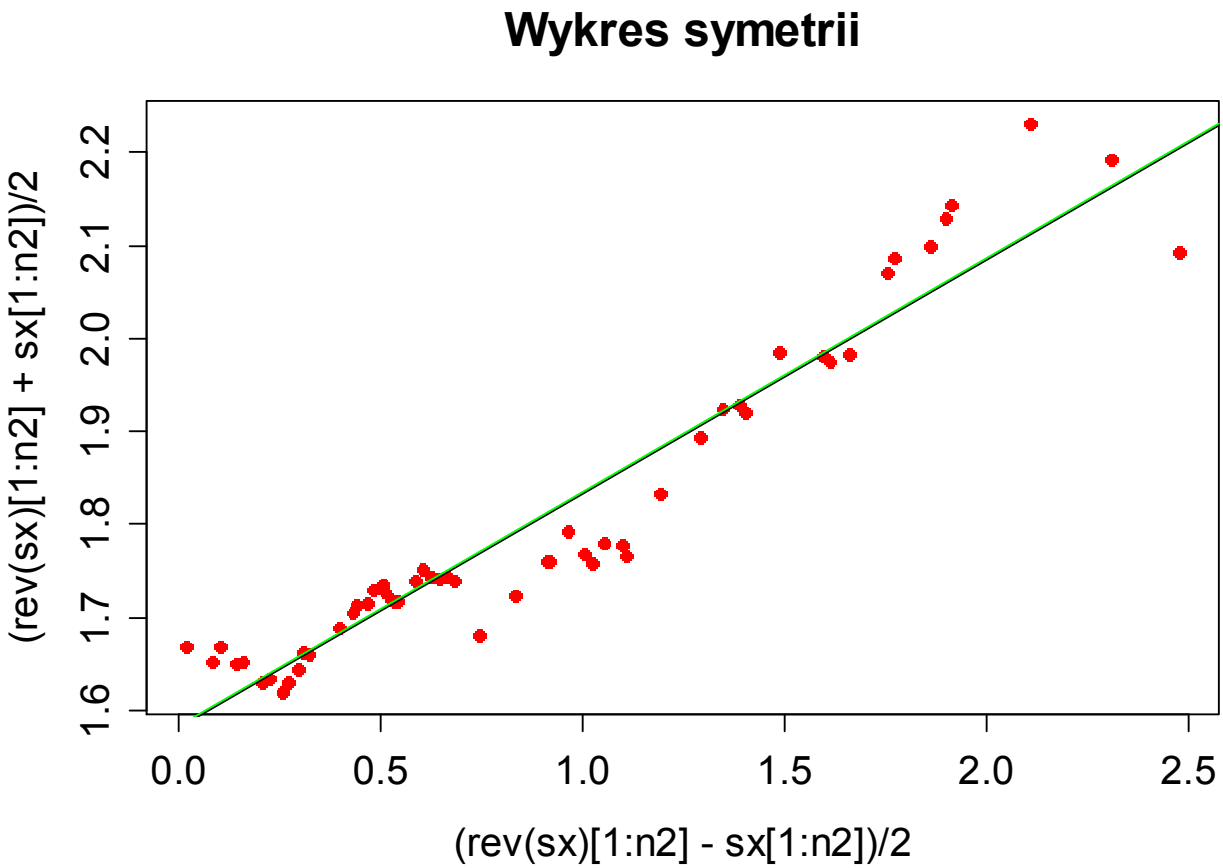

```
                  *
Call: rlm(formula = sympointsAD(x)[, 2] ~ sympointsAD(x)[, 1])
Residuals:
      Min        1Q    Median        3Q       Max
-0.640535 -0.294010  0.001291  0.253899  4.323626

Coefficients:
                   Value     Std. Error t value
(Intercept)          99.3888    0.0932  1066.9154
sympointsAD(x)[, 1]  -0.0243    0.0108    -2.2516

Residual standard error: 0.4227 on 49 degrees of freedom
```

prostaSymetrii2AD(salesInsurance)




```
Call: rlm(formula = sympointsAD(x)[, 2] ~ sympointsAD(x)[, 1])
Residuals:
      Min        1Q    Median        3Q       Max
-0.904989 -0.617975 -0.001559  0.475681  1.730345

Coefficients:
                   Value    Std. Error t value
(Intercept)          3.4875   0.1206    28.9295
sympointsAD(x)[, 1]  0.9175   0.0084   108.9210

Residual standard error: 0.8508 on 54 degrees of freedom
```

prostaSymetrii2AD(log(salesInsurance))




```
Call: rlm(formula = sympointsAD(x)[, 2] ~ sympointsAD(x)[, 1])
Residuals:
      Min        1Q    Median        3Q       Max
-0.112165 -0.018884 -0.001173  0.025178  0.117854

Coefficients:
                   Value    Std. Error t value
(Intercept)          1.5837   0.0114   139.3321
sympointsAD(x)[, 1]  0.2510   0.0104    24.1167

Residual standard error: 0.03574 on 54 degrees of freedom
```