# Cwiczenie 3 - Rozkłady empiryczne i teoretyczne

*Michał Marosz*

*31 października 2015*

# Spis treści

Analiza właściwości zmiennych jest jednym z podstawowych zadan z jakimi przyjdzie sie Wam zmierzyć, w trakcie analizy danych. Dlatego rozpoczniemy od analizy rozkładów empirycznych a nastepnie wrowadzimy pojęcia i praktyczne zastosowanie dostepnych w R rozkladów teoretycznych, które najczęściej znajduja zastosowanie w analizach z zakresu Klimatologii. W R dostępnych jest cala gama rozkladów teoretycznych. Wystarczy w *pomocy* Rstudio wpisac *distributions* i otrzymamy dostęp do do informacji z tego zakresu.

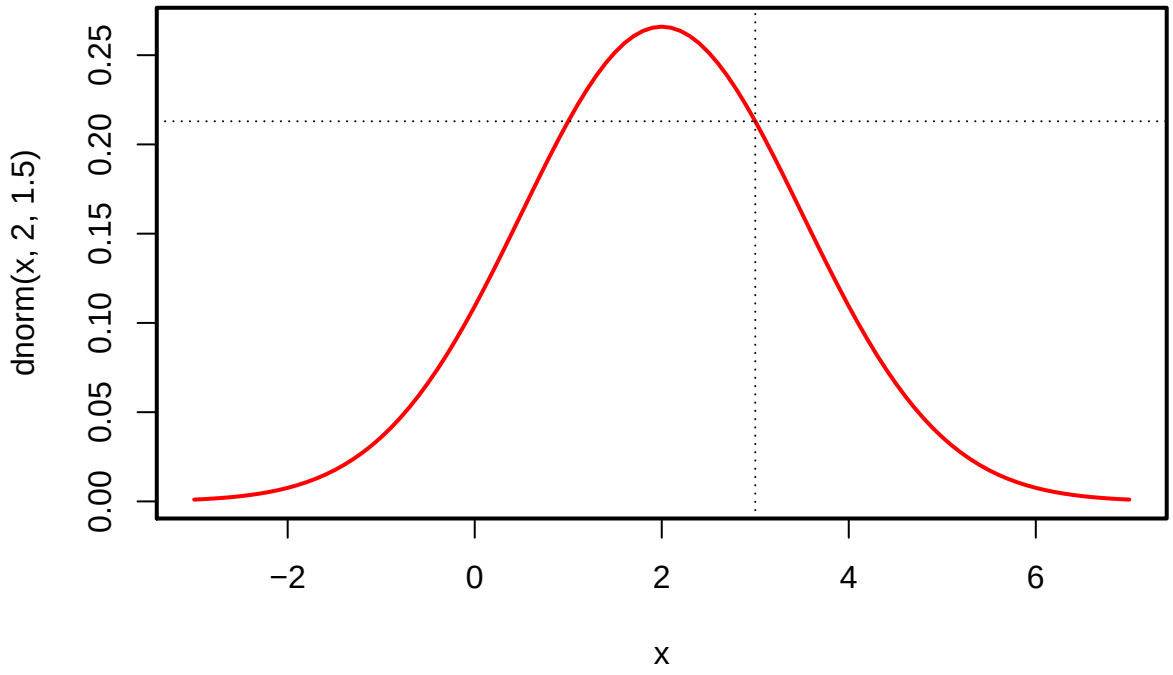
Istotnym jest aby nauczyć sie "zadawać" R odpowiednie *pytania* w analizach rozkladów. I tak, jeżeli chcemy uzyskać informację o wartości gęstości prawdopodobieństwa w określonym rozkładzie i dla konkretnej wartości skrót nazwy rozkładu poprzedzimy literą *d* (od Density) np. wpisując:

```
dnorm(3, 2, 1.5)
```

```
## [1] 0.2129653
```

jako rezultat otrzymamy prawdopodobieństwo wystapienia wartości zmiennej 3, jeżeli ma ona rozklad normalny (Gauss'a) o $\mu = 2$ i $\sigma = 1.5$. Można to graficznie przedstawic w sposob nastepujący:

```
curve(dnorm(x, 2, 1.5), xlim=c(-3, 7), col=2, lwd=2, add=F)
abline(v=3, h=dnorm(3, 2, 1.5), lty=3)
box(lwd=2)
```




Kolejną informację której możemy pożądać stanowi wartość prawdopodobieństwa, że przy założeniu okreslonego rozkładu przekroczona zostanie (lub też nie), określona wartość. Wówczas posłuzymy się przedrostkiem $p$ np.:

```
pnorm(3, 2, 1.5)
```

```
## [1] 0.7475075
```

udzieli nam to odpowiedzi na pytanie jakie jest prawdopodobieństwo że w rozkładzie normalnym o $\mu = 2$ i $\sigma = 1.5$ wartość będzie niższa od 3. Jeżeli natomiast interesuje nas to, czy będzie ona wieksza uzyjemy dodatkowego argumentu *lower.tail=FALSE* np.:

```
pnorm(3, 2, 1.5, lower.tail = FALSE)
```

```
## [1] 0.2524925
```

można oczywiście wykorzystac poprzedni kod ale wynik trzeba odjąc od 1.

```r
1-pnorm(3, 2, 1.5)
```

```
## [1] 0.2524925
```

Np. analizzujemy rozkład wartości średniej dobowej temperatury powietrza dla jednego z miesięcy letnich. Załóżmy, ze jest to rozkład normalny o $\mu = 15$ i $\sigma = 3.2$. jakie jest prawdopodobieństwo, że średnia dobowa temperatura powietrza spadnie poniżej 10°C

```r
pnorm(10, 15, 3.2)
```

```
## [1] 0.05908512
```

albo ze przekroczy 22°C

```r
pnorm(22, 15, 3.2, lower.tail = FALSE)
```

```
## [1] 0.01435302
```

Można oczywiście przemnożyc wynik przez 100 i ladnie zaokrąglić, aby otrzymac wartość w %

```r
round(100*pnorm(22, 15, 3.2, lower.tail = FALSE), digits=1)
```

```
## [1] 1.4
```

Kolejnym z pytań które można zadawać R odnosi się do wartości o konkretnym prawdopodobieństwie przekroczenia, czyli innymi słowy poszukujemy wartości kwantyla. W tym wyapdku nasz predrostek to $q$ a dla rozkładu normalnego o $\mu = 15$ i $\sigma = 3.2$ o wartość której prawdopodobieństwo przekroczenia wynosi 1% można "zapytać się" w sposób nastepujący:

```r
qnorm(0.99, 15, 3.2)
```

```
## [1] 22.44431
```

lub z uwgzlędnieniem argumentu *lower.tail*

```r
qnorm(0.01, 15, 3.2, lower.tail = FALSE)
```

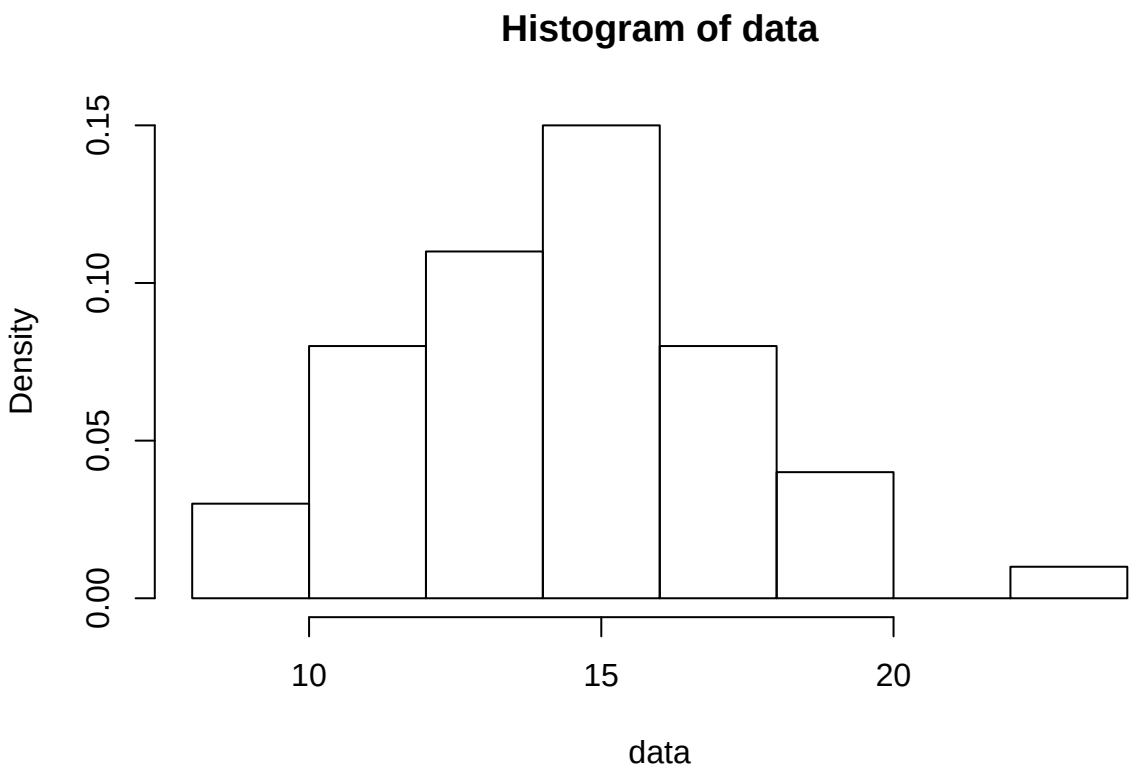
```
## [1] 22.44431
```

## Rozkład empiryczny i dystrybuanta empiryczna

Rozkład empiryczny zazwyczaj przedstawia sie z wykorzystaniem histogramu natomiast dystrybuante empiryczną z wykorzystaniem funkcji *ecdf*

Utwórznmy wektor 100 losowych wartości o rozkładzie normalnym o $\mu = 15$ i $\sigma = 3.2$

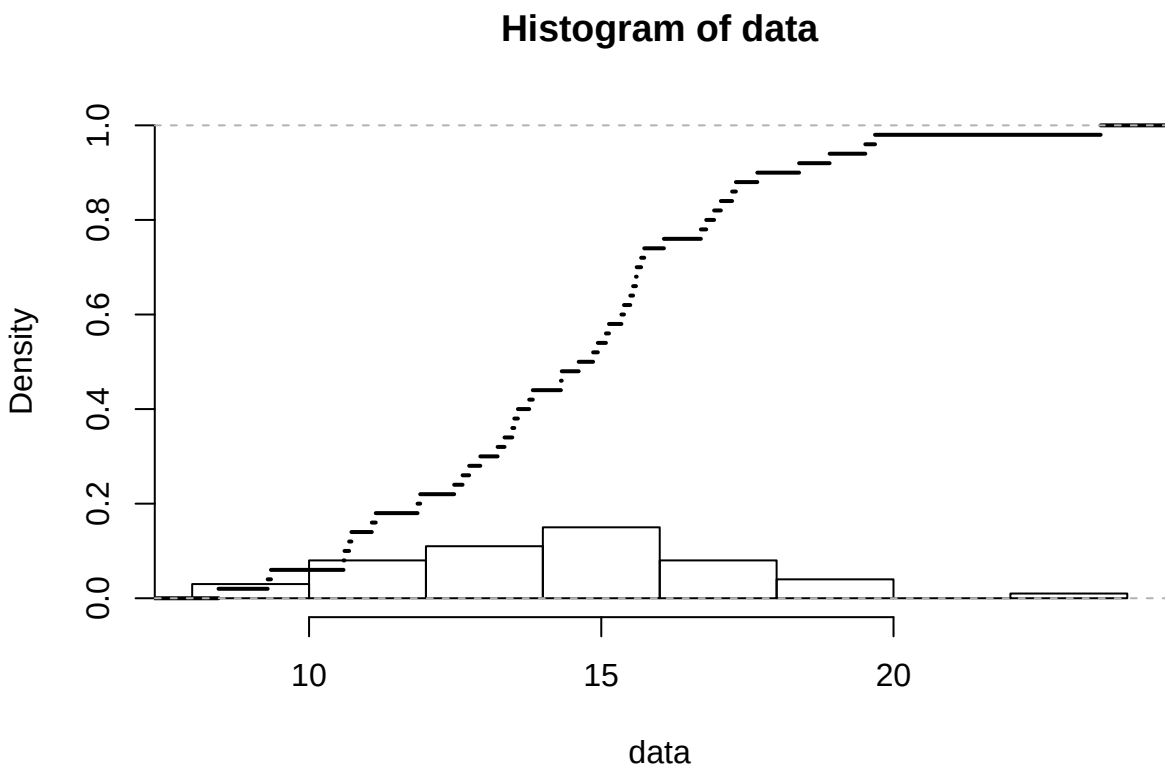
```r
set.seed(1000)
data=rnorm(50, 15, 3.2)
hist(data, prob=T)
```




Aby dodać do powyższego histogramu dystrybuantę empiryczną posłużymy sie funkcją *ecdf*

```
hist(data, prob=T, ylim=c(0,1))
plot(ecdf(data), vertical=FALSE, pch="", add=T, lwd=2)
```

```
## Warning in segments(ti.l, y, ti.r, y, col = col.hor, lty = lty, lwd =
## lwd, : 'vertical' nie jest parametrem graficznym
```




## Estymacja parametrów rozkładów teoretycznych

W przypadku rozkładu normalnego (Gauss'a) estymacja parametrów nie nastręcza większych problemów. Średnia $\mu$ oraz odchylenie standardowe $\sigma$ są wysttarczająco dorymi estymatorami parametrów rozkladu normalnego. Jednak dla pozostałych rozkladów niezbędne jest posłużenie sie dodatkowymi funkcjami pozwalającymi na dopasowanie parametrów rozkładów w tym celu niezbędne jest zainstalowanie *paczki fitdistplus*

Dopasujmy parametry rozkładu Gaussa za pomocą funkcji *fitdist* do temperatur powietrza w kwietniu.

```
dane=read.table("air.txt", header=T)
attach(dane)
temp04=TEMP[which(MC==4)]
hist(temp04)
```


policzmy

wartość średnią $\mu$ oraz odchylenie standardowe $\sigma$

```
mean(temp04)
```

```
## [1] 7.184375
```

```
sd(temp04)
```

```
## [1] 1.564408
```

Teraz spawwdzmy jakie wartości parametrów dopasuje funkcja *fitdis* z wykorzystaniem metody najwiekszej wiarygodności (*mle*), będącej standardem we współczesnych analizach.

Wpiszmy nastepujący kod:

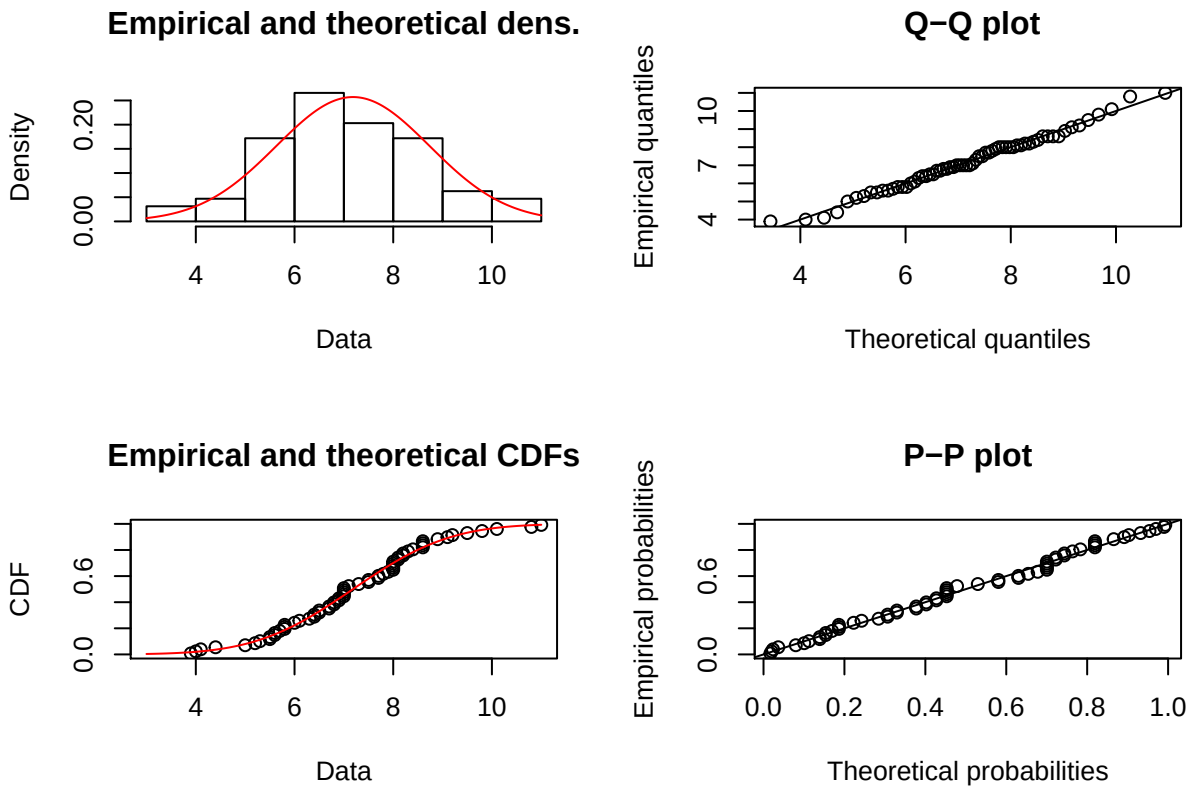
```
library("fitdistrplus", lib.loc="~/R/x86_64-pc-linux-gnu-library/3.2")
```

```
## Loading required package: MASS
```

```
normal_fit = fitdist(temp04, dnorm, method = "mle")
```

Zwróćcie uwagę wynik analiz jest obiektem, ktory można poddać dalszej analizie w celu oceny dopadowania np.:

```
plot(normal_fit)
```




Można również "wyciągnąc" wartości parametrów ze zmiennej *estimate* będącej jedna ze składowych obiektu.

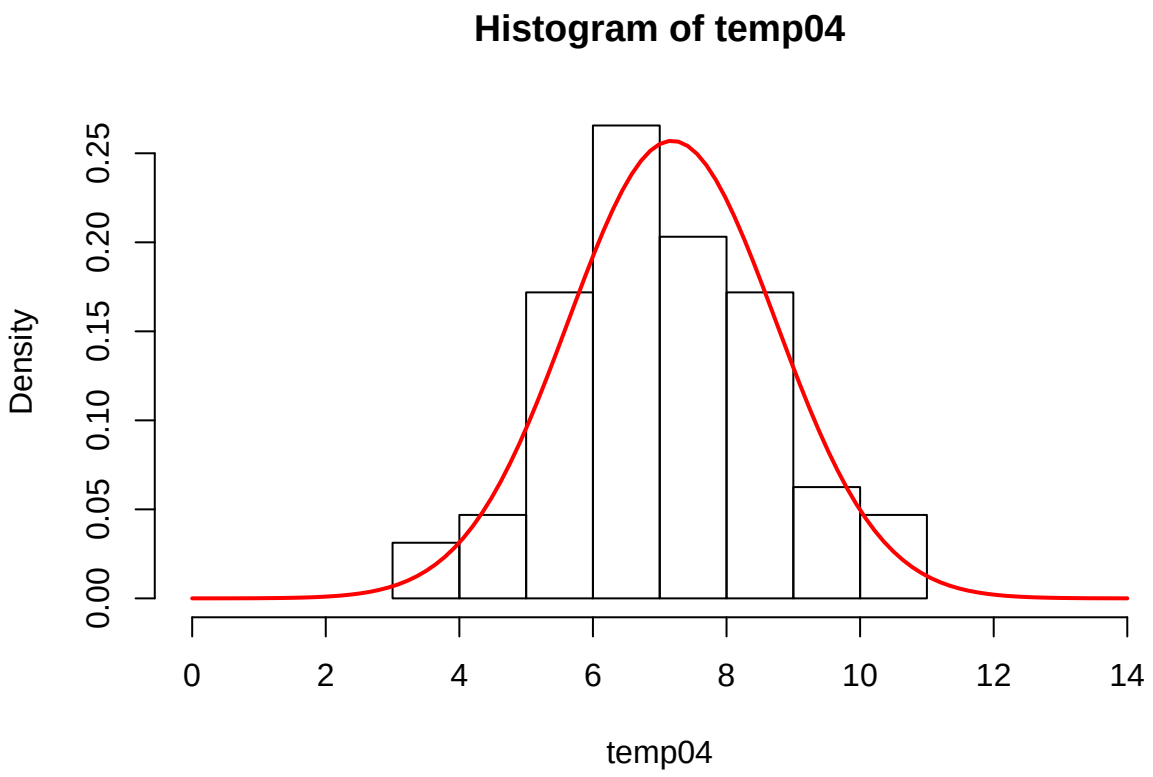
```
normal_fit$estimate
```

```
##     mean       sd
## 7.184375 1.552138
```

Teraz możemy wykreślic ponownie histogram i dodac do niego krzywą rozkładu normalnego wykreslona na podstawie dopasowanych parametrów. “Wypreparujmy” je najpierw z obiektu

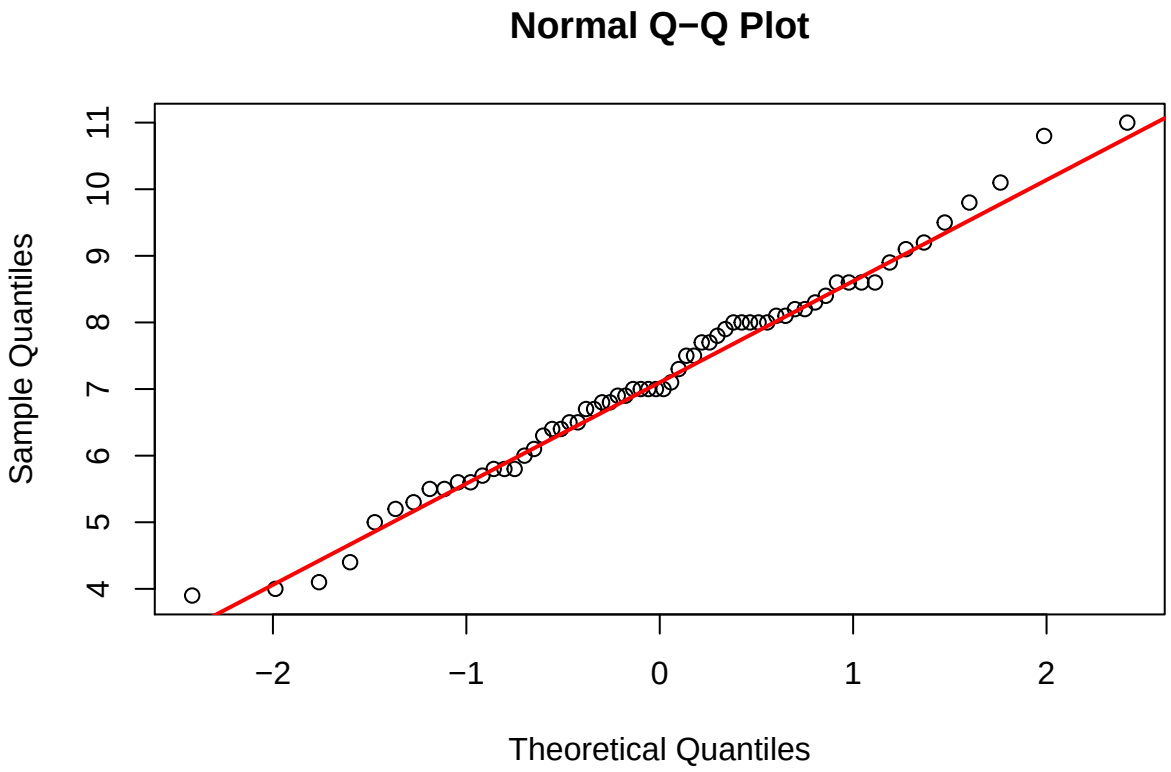
```
mean04=as.numeric(normal_fit$estimate[1])
sd04=as.numeric(normal_fit$estimate[2])
```

```
hist(temp04, prob=T, xlim=c(0,14))
curve(dnorm(x, mean04, sd04), add=T, col=2, lwd=2)
```




Istnieje również klasa wykresów przeznaczona specjalnie do wizualnego porównywania rozkadów empirycznych z teoretycznym normalnym: *qqnorm*

```r
qqnorm(temp04)
qqline(temp04, col=2, lwd=2)
```



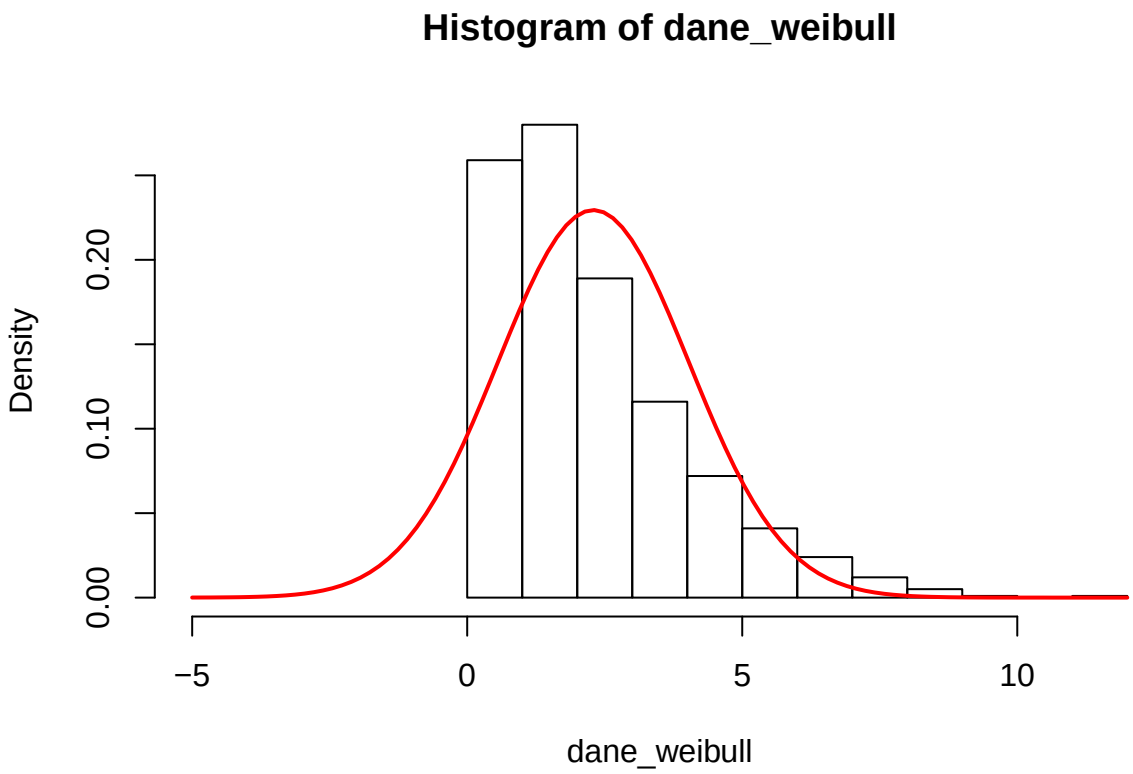

Weryfikację jakości dopasowania można przeprowadzić wizualnie za pomocą uprzednio wywołanej funkcji *plot/qqnorm* lub zaprząc do tego nieco bardziej sformalizowane testy np.: Shapiro-Wilk'a *shapiro.test*

```r
shapiro.test(temp04)
```

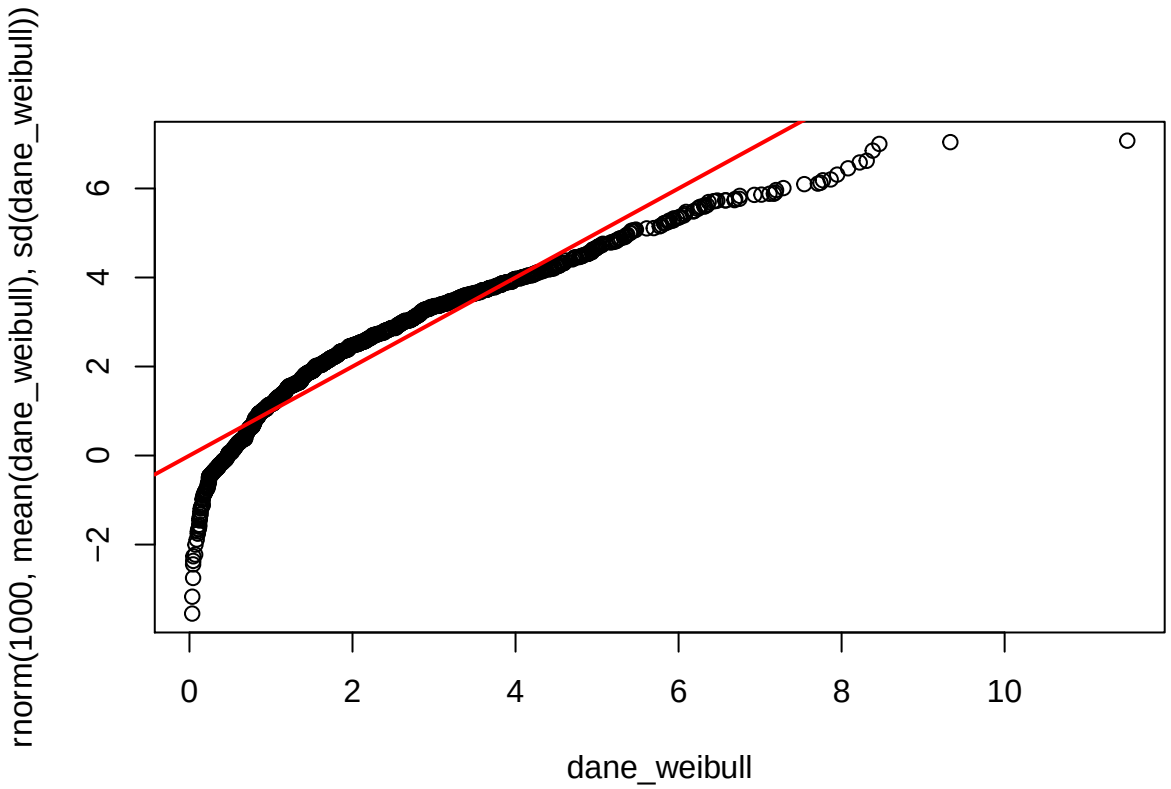
```
##
##  Shapiro-Wilk normality test
##
## data:  temp04
## W = 0.98799, p-value = 0.7913
```

Dla jasności wygenerujmy dane o rozkładzie innym niż normalny i sprawdzmy wykresy oraz wartości z powyższego testu.

```r
set.seed(1000)

dane_weibull=rweibull(1000, 1.3, 2.5)


hist(dane_weibull, prob=T, xlim=c(-5, 12))

curve(dnorm(x, mean(dane_weibull), sd(dane_weibull)), col=2, lwd=2, add=T)
```




```r
qqplot(dane_weibull, rnorm(1000, mean(dane_weibull), sd(dane_weibull)))

abline(0,1, col=2, lwd=2)
```




```
shapiro.test(dane_weibull)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dane_weibull
## W = 0.90707, p-value < 2.2e-16
```

# Zmienne dyskretne - rozkłady teoretyczne

## Rozkład Bernouli'ego

## Rozkład Poisson'a

# Zmienne ciągłe - rozkłady teoretyczne

## Rozkład Gauss'a

## Rozkład Gamma

## Rozkład Weibull'a

## Rozkład GEV (Generalized Extreme Value)

# Rozklady wykorzystywane często we wnioskowaniu statystycznym

## Rozkład t-Studenta

## Rozklad $\chi^2$

## Rozkład Fishera-Snedecora