Wnioskowanie Statystyczne - Ćwiczenia

Michał Marosz

Monday, February 23, 2015

Zadanie 1

Załaduj do R dane udostepnione na poprzednich zajęciach i wyświetl podstwowe informacje o zawartych tam danych

```
setwd("D:/!CLIMATE/Pulpit")
dane=read.table("dane.txt", header=T)
attach(dane)
summary(dane)
```

```
##       ROK             A               B               C
##  Min.   :1951   Min.   : 2.600   Min.   : 4.000   Min.   : 1.700
##  1st Qu.:1976   1st Qu.: 6.100   1st Qu.: 7.400   1st Qu.: 4.775
##  Median :2000   Median : 6.700   Median : 8.100   Median : 5.800
##  Mean   :2000   Mean   : 6.869   Mean   : 8.207   Mean   : 5.862
##  3rd Qu.:2025   3rd Qu.: 7.825   3rd Qu.: 9.200   3rd Qu.: 6.925
##  Max.   :2050   Max.   :11.300   Max.   :12.600   Max.   :10.300
##        D
##  Min.   :3.700
##  1st Qu.:5.700
##  Median :6.400
##  Mean   :6.323
##  3rd Qu.:6.925
##  Max.   :8.500
```

Zastosowanie funkcji `attach` pozwala na odnoszenie się bezpośrednio do zmiennych wczytanych z nagłówka (np. `A`) zamiast wpisywania `dane$A`. Przypominam, że ten poprzez ten sposób ładowania danych obiekt `dane` jest ramką danych (`dataframe`). Można to zweryfikować nastepującą komendą:

```
is.data.frame(dane)
```

```
## [1] TRUE
```

Zadanie 2

Zapoznaj się z podstawowymi charakterystykami statystycznymi dostepnymi w R a wymienionymi w skrypcie Łukasza Komsty.

Dokonaj obliczenia dla zmiennej `B`. Poniżej podano przykładowe ich zastosowanie dla zmiennej `A`.

`max` — maksymalna wartość z wektora

```
max(A)
```

```
## [1] 11.3
```

`min` — wartość minimalna

```
min(A)
```

```
## [1] 2.6
```

`mean` — średnia arytmetyczna. Jeśli podamy dodatkowy parametr trim, to funkcja policzy średnią po odrzuceniu określonego odsetka wartości skrajnych, np. mean(x,trim=0.1) to średnia z x po odrzuceniu 10% wartości skrajnych

```
mean(A)
```

```
## [1] 6.869
```

`median` — mediana

```
median(A)
```

```
## [1] 6.7
```

`mad` — medianowe odchylenie bezwzględne (median absolute deviation)

```
mad(A)
```

```
## [1] 1.40847
```

`quantile` — dowolny kwantyl, np. `quantile(A, 0.5)` to mediana z x. Poniżej przedstawiono przykładowy sposób okreslenia kwartyli (1-go, 2-go oraz 3-go). Polecenia w jednej linii ale oddzielone od siebie `;` są traktowane osobno.

```
quantile(A, 0.25); quantile(A, 0.5); quantile(A, 0.75)
```

```
## 25%
## 6.1
```

```
## 50%
## 6.7
```

```
##   75%
## 7.825
```

`sd` — odchylenie standardowe `var` — wariancja `length` — długość wektora (liczba elementów) `sum` — suma elementów wektora `sort` — daje wektor z wartościami uporządkowanymi rosnąco `which` — daje wektor zawierający indeksy, przy których argument ma wartość TRUE.

Tutaj wypada na chwilę pochylić się nad funkcją `which`. Jest ona niezmiernie użyteczna w analizach, ponieważ pozwala w latwy sposób pracowac na podzbiorach wyznaczanych przez okreslone warunki (np. na wybranych wieloleciach).

Przykladowo swórzmy nastepujący wektor `x`.

```
x=c(2, 4, 8, 9, 12, 15, 21)
```

Za pomocą funkcji `which` możemy wybrać np. elementy które są większe od 10.

```
which(x>10)
```

```
## [1] 5 6 7
```

Nastepnie, w połaczeniu z indeksowanie za pomocą `[]` można wykonywać funkcje na tylko na wybranym podzbiorze.

```
mean(x[which(x>10)])
```

```
## [1] 16
```

Rozklady teoretyczne

beta `dbeta`

binomial `dbinom`.

For the Cauchy distribution see dcauchy.

For the chi-squared distribution see dchisq.

For the exponential distribution see dexp.

For the F distribution see df.

For the gamma distribution see dgamma.

For the geometric distribution see dgeom. (This is also a special case of the negative binomial.)

For the hypergeometric distribution see dhyper.

For the log-normal distribution see dlnorm.

For the multinomial distribution see dmultinom.

For the negative binomial distribution see dnbinom.

For the normal distribution see dnorm.

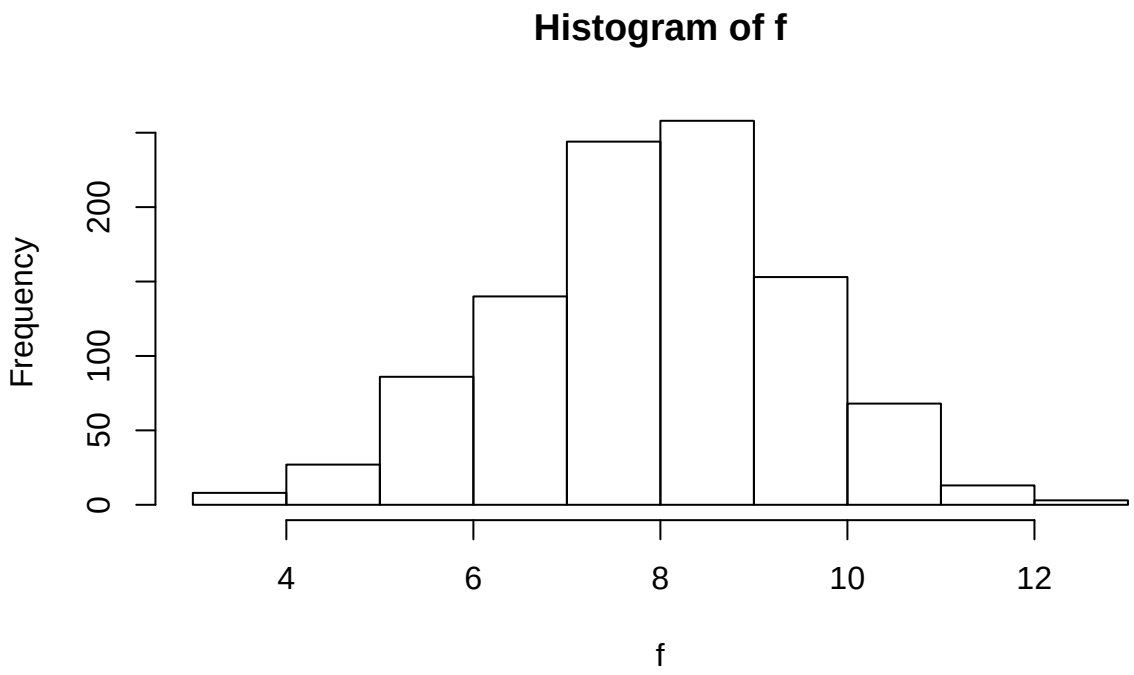
For the Poisson distribution see dpois.

For the Student's t distribution see dt.

For the uniform distribution see dunif.

For the Weibull distribution see dweibull.

```
f=rnorm(1000, 7.9, 1.5)
hist(f)
```




```r
mean(f)
```

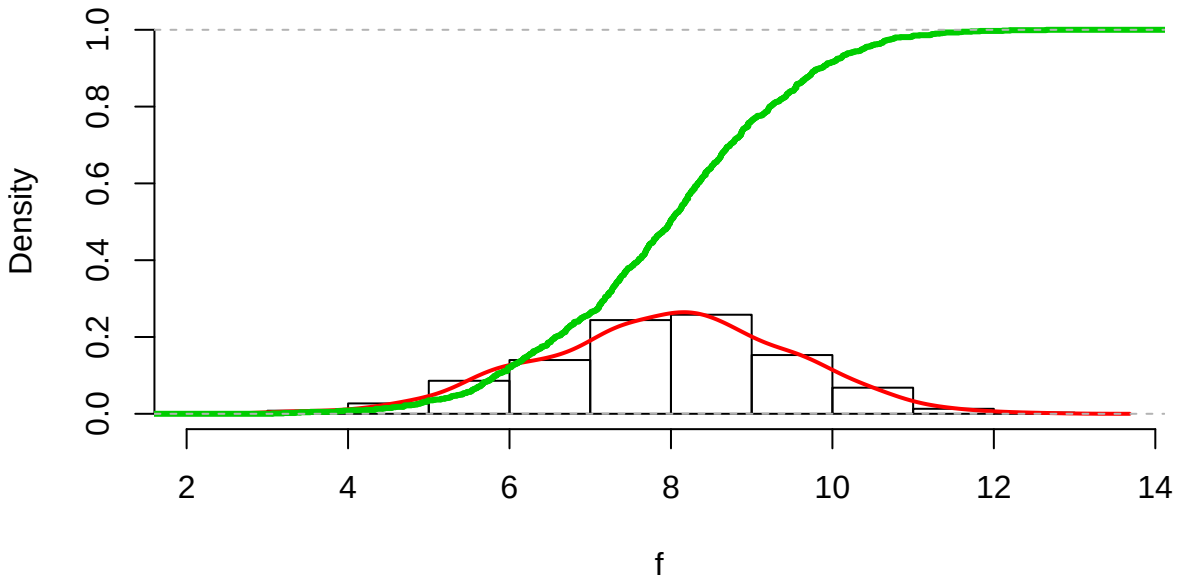
```
## [1] 7.914762
```

```r
sd(f)
```

```
## [1] 1.537669
```

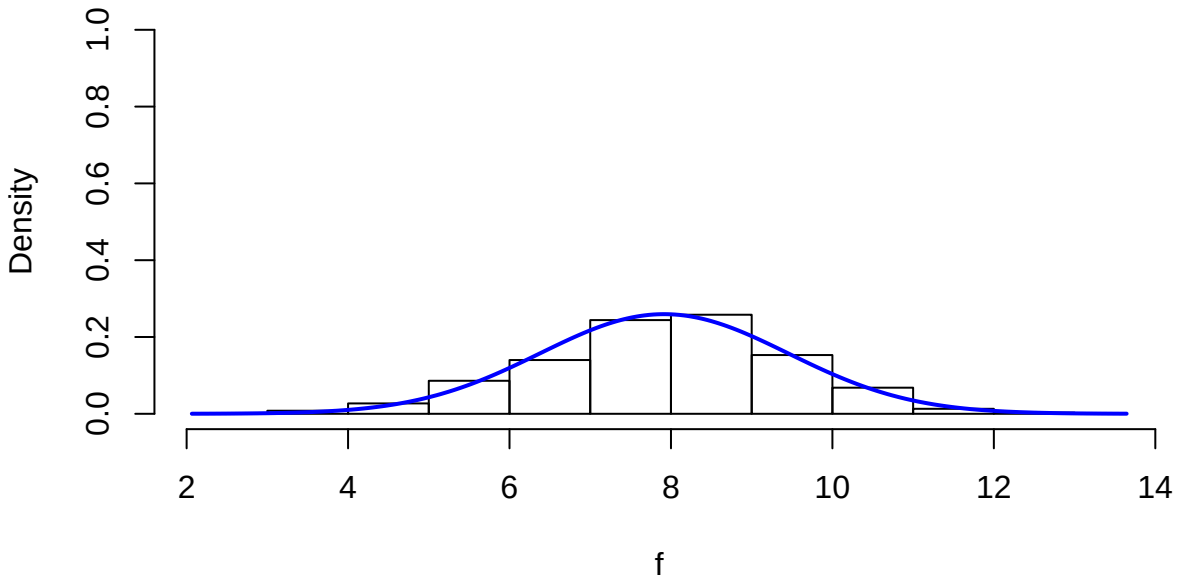
Dystrybuanta empiryczna

```r
hist(f, prob=T, main="", xlim=c(min(f)-1, max(f)+1), ylim=c(0,1))
lines(density(f), col=2, lwd=2)
lines(ecdf(f), col=3, lwd=3)
```



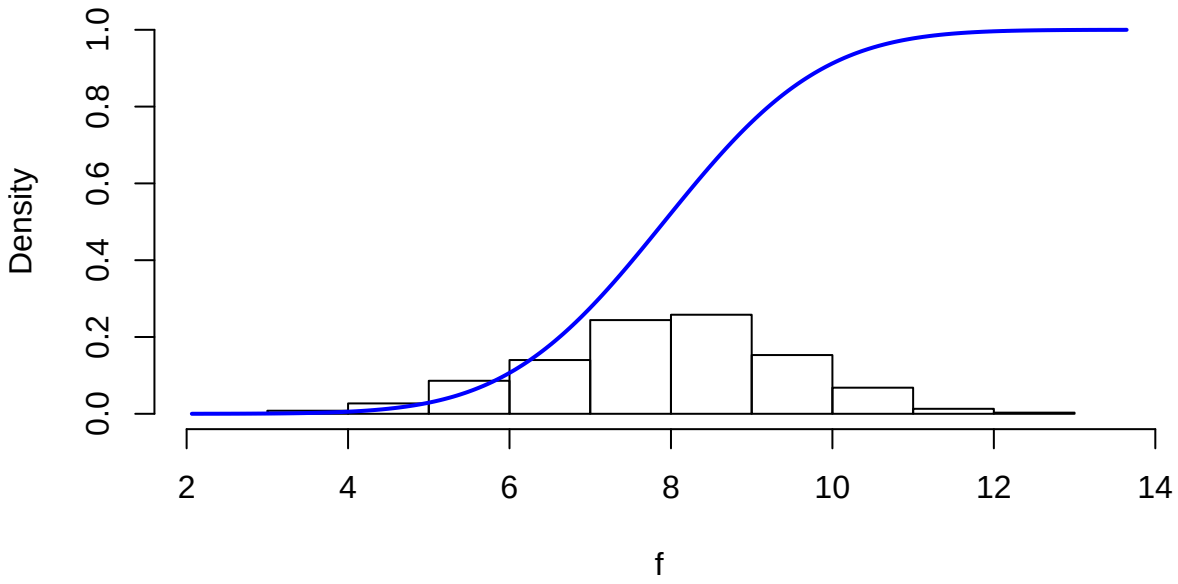
Kreślenie rozkładu teoretycznego

```
hist(f, prob=T, main="", xlim=c(min(f)-1, max(f)+1), ylim=c(0,1))
x=seq(min(f)-1, max(f)+1, 0.01)
lines(x, dnorm(x,mean(f), sd(f)), col=4, lwd=2)
```



Kreślenie dystrybuanty teoretycznej na podstawie oszacowanych parametrów rozkładu

```
hist(f, prob=T, main="", xlim=c(min(f)-1, max(f)+1), ylim=c(0,1))
x=seq(min(f)-1, max(f)+1, 0.01)
lines(x, pnorm(x,mean(f), sd(f)), col=4, lwd=2)
```

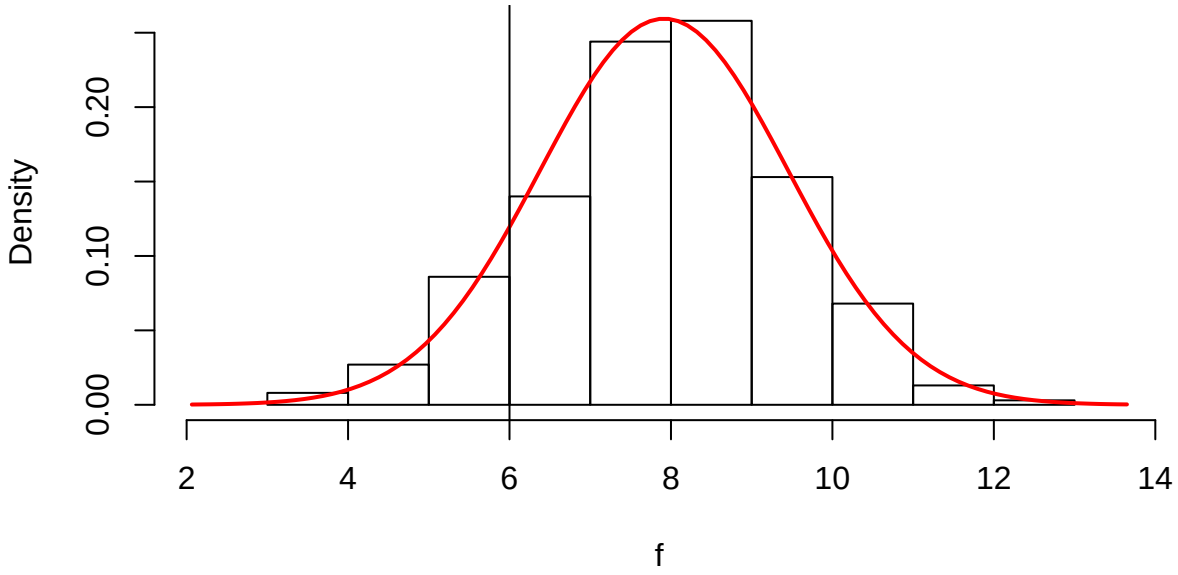


Załóżny, że chcemy sprawdzić jaki odsetek przypadkow w rozkładzie normalnym opisanym parametrami zmiennej `f` będzie mniejszy od 6

```
pnorm(6, mean(f), sd(f))
```

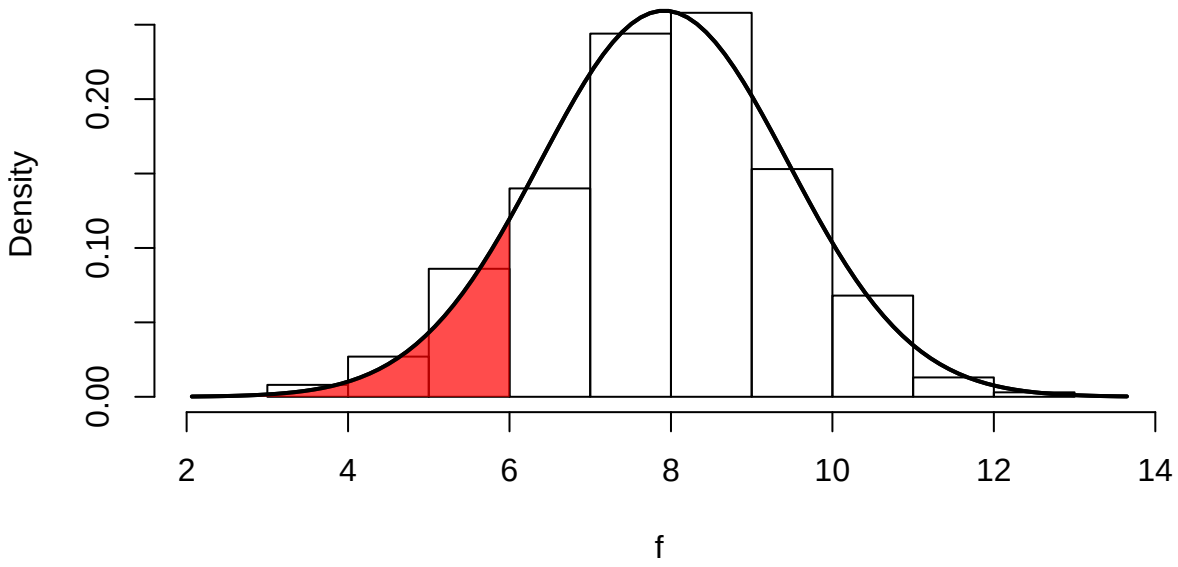
```
## [1] 0.1065223
```

```
hist(f, prob=T, main="", xlim=c(min(f)-1, max(f)+1), )
curve(dnorm(x,mean(f), sd(f)), add=T, col=2, lwd=2)
abline(v=6)
```



Ładniej graficznie wygląda to tak:

```
hist(f, prob=T, main="", xlim=c(min(f)-1, max(f)+1), )
curve(dnorm(x, mean(f), sd(f)), add=T, lwd=2)
x <- seq(min(f)-1, 6, len = 100)
y <- dnorm(x, mean(f), sd(f))
polygon(c(x[1], x, x[100]), c(0, y, 0), col = rgb(1,0,0,0.7), border = NA)
curve(dnorm(x, mean(f), sd(f)), add=T, lwd=2)
```

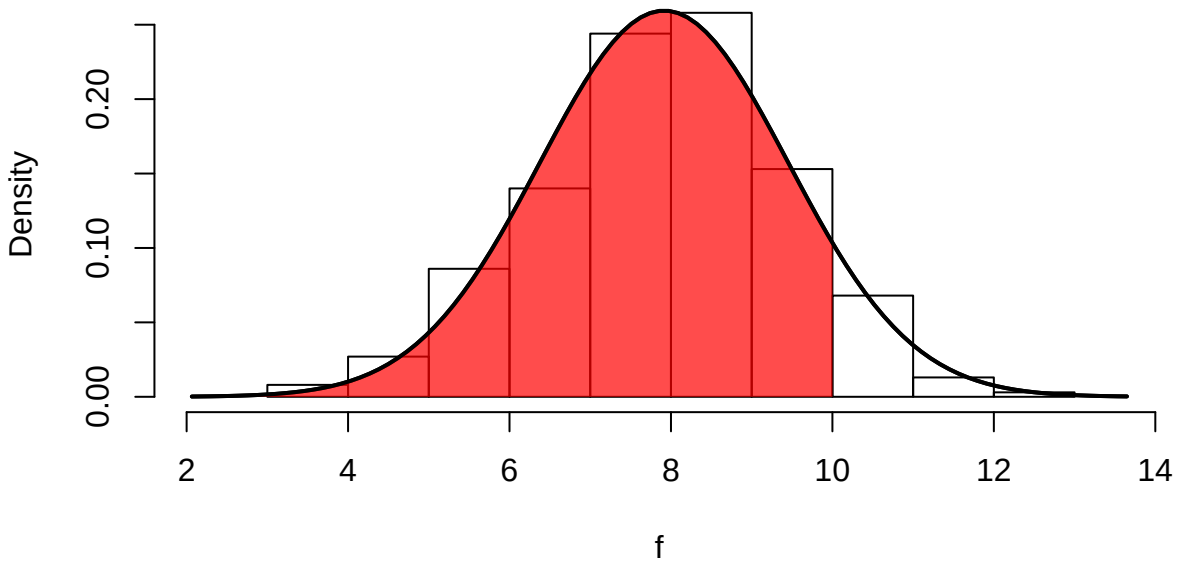


Lub od 10

```
pnorm(10, mean(f), sd(f))
```

```
## [1] 0.9124669
```

```
hist(f, prob=T, main="", xlim=c(min(f)-1, max(f)+1), )
curve(dnorm(x, mean(f), sd(f)), add=T, lwd=2)
x <- seq(min(f)-1, 10, len = 100)
y <- dnorm(x, mean(f), sd(f))
polygon(c(x[1], x, x[100]), c(0, y, 0), col = rgb(1,0,0,0.7), border = NA)
curve(dnorm(x, mean(f), sd(f)), add=T, lwd=2)
```

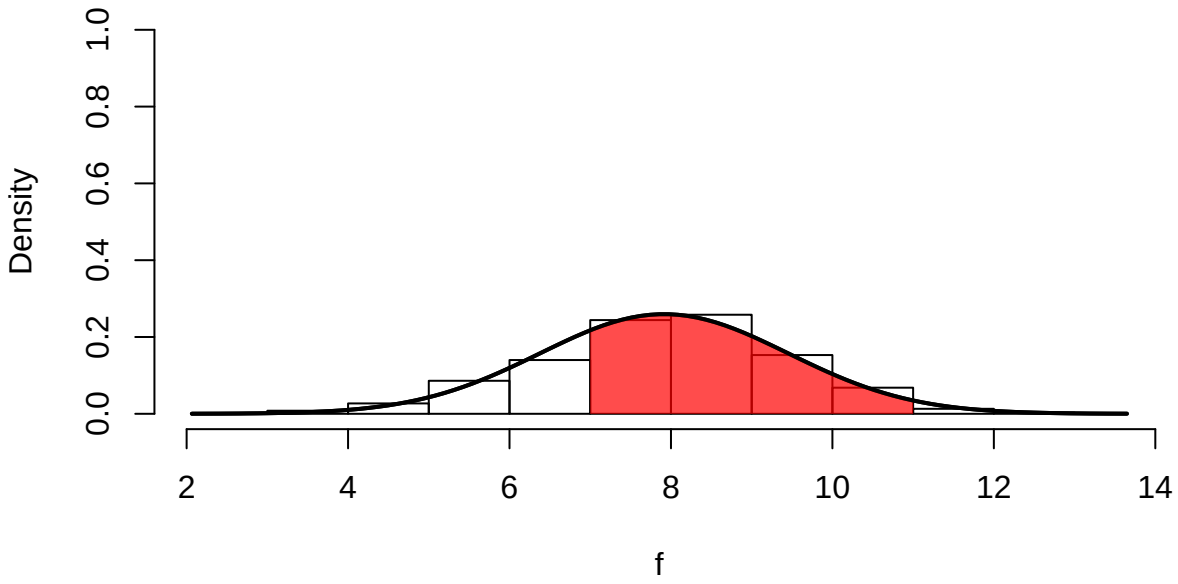


Lub między 7 a 11

```
pnorm(11, mean(f), sd(f))-pnorm(7, mean(f), sd(f))
```

```
## [1] 0.7016408
```

```
hist(f, prob=T, main="", xlim=c(min(f)-1, max(f)+1), ylim=c(0,1))
curve(dnorm(x, mean(f), sd(f)), add=T, lwd=2)
x <- seq(7, 11, len = 100)
y <- dnorm(x, mean(f), sd(f))
polygon(c(x[1], x, x[100]), c(0, y, 0), col = rgb(1,0,0,0.7), border = NA)
curve(dnorm(x, mean(f), sd(f)), add=T, lwd=2)
```


Density
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
10
12
14
f

Rozklady dla zmiennych ciagłych

RozKlad Normalny

Rozklad Weibull'a

Rozkład Gamma

Rozklad χ^2

Rozkład F-Snedecora

Rozkłady dla zmiennych dyskretnych

Rozkład dwumianowy

Rozkład Poisson'a

Rozklad normalny - zadania

Wykonaj poniższe zadania korzystając z `R` i wbudowanych dystrybuant teoretycznych.

Praktyczne korzystanie z rozkładów teoretycznych (na przykładzie normalnego bo w przypadku innych rozkładów należy sprawdzić ich parametry)

`dnorm(x, mean = 0, sd = 1, log = FALSE)` - funkcja gęstości prawdopodobieństwa

`pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)` - dystrybuanta

`qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)` - wartość kwantyla na podstawie prawdopodobieństwa

`rnorm(n, mean = 0, sd = 1)` - generowanie liczb losowych o określonym rozkładzie teoretycznym

`x, q` - kwantyle

`p` - prawdopodobieństwo

`n` - liczba obserwacji

`mean` - średnia

`sd` - odchylenie standardowe

`lower.tail` - wartość logiczna; jeżeli TRUE (default), obliczane jest prawdopodobieństwo z dolnego "ogona rozkładu".

Dopasowywanie parametrów rozkładu - tutaj na przykładzie rozkładu normalnego

```r
#install.packages("fitdistrplus")
library("fitdistrplus", lib.loc="~/R/win-library/3.1")
f=rnorm(1000, 7.9, 1.5)
fitnorm=fitdist(f, distr="norm", method="mle")
fitnorm
```

```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters:
##      estimate Std. Error
## mean 7.935008 0.04650485
## sd   1.470612 0.03288382
```

Gdybyśmy potrzebowali konkretnych wartości do dalszych obliczeń

```r
fitnorm$estimate
```

```
##     mean       sd
## 7.935008 1.470612
```

```r
as.numeric(fitnorm$estimate)
```

```
## [1] 7.935008 1.470612
```

```r
as.numeric(fitnorm$estimate[1])
```

```
## [1] 7.935008
```

```r
as.numeric(fitnorm$estimate[2])
```

```
## [1] 1.470612
```

Zadanie 3

Poszukaj w pomocy R informacji odnośnie rozkładu Weibulla, sprawdz jakie parametry musisz estymować nastepnie wczytaj do R dane odnośnie prędkości wiatru. Wykreśl histogram i dystrybuantę empiryczną rozkładu a następnie dopasuj rozkład teoretyczny. Wykonaj wykres.

Odpowiedz na pytanie a) jakie jest prawdopodobieństwo że prędkość wiatru przekroczy $20ms^{-1}$ b) prędkośc wiatru będzie miała wartośc między 10 a $20ms^{-1}$

Zadanie 4

Wykonaj nastepujące zadania

I.

1) Zmienna losowa Z ma $\mu =$ oraz $\sigma =$.

2) $P(0 < z < 1.53) =$

3) $P(z > -2.18) =$

4) Określ wartość z_o, takie że $P(-z_o < z < z_o) = 0.92$**.**

5) Określ wartość zo, takie że P(z < zo) = 0.3015.

II. Zmienna losowa X ma rozkład normalny ze średnią 80 I odchyleniem standardowym 12

1) Jakie jest prawdopodobieństwo, że wartość zmiennej X będzie między 65 i 95?

2) Jakie jest prawdopodobieństwo, ze wartość losowo wybranej zmiennej X będzie mniejsza od 74?

III. Zmienna losowa X ma rozkład normalny ze średnią 65 I odchyleniem standardowym 15.

Określ x0 takie że P(x > xo) = .6738.

IV. Wyniki testu maja rozkład normalny ze średnia 400 i odchyleniem standardowym 45

1) Jaki odsetek osób podchodzących do egzaminu będzie miała wynik 310 lub wyższy?

2) Jaki odsetek osób podchodzących do egzaminu będzie miało wynik między 445 a 490?

V. Opracowano test, którego zadaniem było zmierzenie poziomu motywacji w liceum. Wyniki poziomu motywacji maja rozkład normalny ze średnią 25 i odchyleniem standardowym 6. Im wyższa wartość tym większa motywacja.

1) Jaki odsetek uczniów biorących udział w badaniu będzie miał wynik poniżej 10?

2) Jan usłyszał, że 35% uczniów ma większą motywację niż on. Jaki jest poziom motywacji Jana?

VI. Rozkład Poisson'a

1) Jeżeli 3% żarówek produkowanych przez fabrykę jest uszkodzonych, określ prawdopodobieństwo że w próbie 100 żarówek dokładnie 5 jest uszkodzonych (e-3 = 0.0498).

2) Wiadomo na podstawie przeszłych doświadczeń, że w fabryce zdarzają się średnio 4 wypadki na miesiąc. Oblicz prawdopodobieństwo, że w miesiącu będą mniej niż 3 wypadki.

VII. Dwumianowy

1) Rzucamy jednocześnie ośmioma monetami. Jakie jest prawdopodobieństwo wyrzucenia co najmniej sześciu orłów

2) Rzucamy dwiema kośćmi do gry 5 razy. Wyrzucenie takiej samej liczby oczek jest uznawane za sukces. Określ prawdopodobieństwo dwóch sukcesów.

3) Prawdopodobieństwo, że student ukończy szkołę wyższą wynosi 0,5. Określ prawdopodobieństwo, że z 7 studentów (i) żaden nie ukończy (ii) jeden ukończy (iii) przynajmniej jeden ukończy studia

4) Równocześnie rzucamy dziesięcioma monetami. Określ prawdopodobieństwo, że otrzymamy:

a. Przynajmniej 7 orłów

b. Dokładnie 7 orłów

c. Co najwyżej 7 orłów

5) Podczas wojny średnio 2 z 10 statków tonęło w trakcie konwoju. Jakie jest prawdopodobieństwo, że przynajmniej 4 z 5 statków bezpiecznie dopłynie do portu przeznaczenia.