

Вероятности и статистика: Упражнения с R

Ангел Г. Ангелов
agangelov@fmi.uni-sofia.bg

3. I. 2023 г.

Съдържание

1. Въведение в R	4
1.1. Математически функции	4
1.2. Вектори	4
1.3. Генериране на редици	6
1.4. Матрици	7
1.5. Data frame	9
1.6. Полезни функции	10
2. Случайни експерименти	11
3. Случайни величини	15
3.1. Дискретни сл.в.	15
3.1.1. Бернулиево разпределение	15
3.1.2. Биномно разпределение	16
3.1.3. Геометрично разпределение	16
3.1.4. Отрицателно биномно разпределение	17
3.1.5. Хипергеометрично разпределение	18
3.1.6. Поасоново разпределение	18
3.2. Непрекъснати сл.в.	19
3.2.1. Равномерно разпределение	20
3.2.2. Експоненциално разпределение	21
3.2.3. Нормално разпределение	22
4. Данни. Таблицы и графики	25
4.1. Категорни данни	25
4.2. Числови данни	27
5. Числови характеристики на данните	32
6. Многомерни данни	37
7. Доверителни интервали	45
7.1. Увод	45
7.2. Доверителен интервал за средно при известна дисперсия	45
7.3. Доверителен интервал за средно при неизвестна дисперсия	46
7.4. Доверителен интервал за пропорция (вероятност за успех)	48
7.5. Доверителен интервал за медиана	49
7.6. Доверителен интервал за разлика на медиани	50
7.7. Интерпретация на доверителни интервали	52
8. Проверка на хипотези при една извадка	53
8.1. z -тест за средно	53
8.2. t -тест за средно	56
8.3. z -тест за пропорция	59
9. Проверка на хипотези при две извадки	62
9.1. t -тест за разлика на средни	62
9.2. t -тест при зависими извадки	65
9.3. z -тест за разлика на пропорции	68

10. Хи-квадрат тестове	71
10.1. Хи-квадрат тест за съгласуваност	71
10.2. Хи-квадрат тест за независимост	73
11. Линейни модели	76
11.1. Линеен модел с един предиктор	76
11.2. Линеен модел с няколко предиктора	78
Литература	80

1. Въведение в R

*Most good programmers do programming not because
they expect to get paid or get adulation by the public,
but because it is fun to program.*

Linus Torvalds

R е език и среда за статистически изчисления и анализ. Разпространява се свободно, съгласно условията на GNU General Public License. Създаден е първоначално през 1993 от Robert Gentleman и Ross Ihaka от Департамента по статистика на Университета Оукланд (University of Auckland) на основа на езика S. На сайта <https://cran.r-project.org/> може да се намери последната версия.

1.1. Математически функции

```
> (5+7)/(4-1)
[1] 4
```

```
> 9~2
[1] 81
```

```
> sqrt(25)
[1] 5
```

```
> log(exp(1))
[1] 1
```

```
> 28 %% 10
[1] 8
```

```
> 27/1000000
[1] 2.7e-05
```

```
> 5000*5000
[1] 2.5e+07
```

```
> options(scipen=999)
> 27/1000000
[1] 0.000027
```

```
> options(scipen=0)
> 27/1000000
[1] 2.7e-05
```

1.2. Вектори

```
> x <- c(5, 12, 11, 14, 2, 3, 14, 10, 3)

> x[3]
[1] 11
```

```

> x[1:5]
[1]  5 12 11 14  2

> x[c(2,5,9)]
[1] 12  2  3

> x[-4]
[1]  5 12 11  2  3 14 10  3

> x[-c(2,3)]
[1]  5 14  2  3 14 10  3

> x[x>10]
[1] 12 11 14 14

> length(x)
[1] 9

> min(x)
[1] 2

> max(x)
[1] 14

> head(x, 3)
[1]  5 12 11

> tail(x, 3)
[1] 14 10  3

> x>10
[1] FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE

> sum(x>10)
[1] 4

> which(x>10)
[1] 2 3 4 7

> diff(x)
[1]  7 -1  3 -12  1 11 -4 -7

> cumsum(x)
[1]  5 17 28 42 44 47 61 71 74

> sum(x)
[1] 74

> x^2
[1] 25 144 121 196  4  9 196 100  9

> sort(x)
[1]  2  3  3  5 10 11 12 14 14

```

```

> x[ order(x) ]
[1]  2  3  3  5 10 11 12 14 14

> rank(x)
[1] 4.0 7.0 6.0 8.5 1.0 2.5 8.5 5.0 2.5

> rm(x)
> x
Error: object 'x' not found

> x <- c(1,3,5,11,15)
> class(x)
[1] "numeric"

> x <- as.integer(c(1,3,5,11,15))
> class(x)
[1] "integer"

> y <- c("Y", "Y", "N")
> class(y)
[1] "character"

> z <- c(TRUE, TRUE, FALSE)
> class(z)
[1] "logical"

> x <- vector("logical", length=5)
> x
[1] FALSE FALSE FALSE FALSE FALSE

> y <- vector("numeric", length=5)
> y
[1] 0 0 0 0 0

> x <- c(5,5,5,7,7,7)
> y <- c(2,2,1)

> x+y
[1] 7 7 6 9 9 8

> y <- c(2,2,1,1)
> x+y
[1] 7 7 6 8 9 9
Warning message:
In x + y : longer object length is not a multiple of shorter object length

```

1.3. Генериране на редици

```

> rep( 5, times=8 )
[1] 5 5 5 5 5 5 5 5
> rep( c(1,2), times=5 )

```

```

[1] 1 2 1 2 1 2 1 2 1 2
> rep( c(1,2), each=5 )
[1] 1 1 1 1 1 2 2 2 2 2
> rep( c(1,2), length.out=7 )
[1] 1 2 1 2 1 2 1

> rep( c("a","b"), times=5 )
[1] "a" "b" "a" "b" "a" "b" "a" "b" "a" "b"
> rep( c("a","b"), each=3 )
[1] "a" "a" "a" "b" "b" "b"

> 5:12
[1] 5 6 7 8 9 10 11 12
> 10:1
[1] 10 9 8 7 6 5 4 3 2 1
> seq( from=1, to=10, by=2 )
[1] 1 3 5 7 9
> seq( from=10, to=1, by=-2 )
[1] 10 8 6 4 2
> seq( from=0, to=1, by=0.2 )
[1] 0.0 0.2 0.4 0.6 0.8 1.0
> seq( from=0, to=1, length.out=11 )
[1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
>

```

1.4. Матрици

```

> M <- rbind( c(5,3,5,6), c(8,3,7,4) )
> M
      [,1] [,2] [,3] [,4]
[1,]    5    3    5    6
[2,]    8    3    7    4

> M[2, 3]
[1] 7

> M[,3]
[1] 5 7

> M[2, ]
[1] 8 3 7 4

> M <- cbind( c(5,3,5,6), c(8,3,7,4) )
> M
      [,1] [,2]
[1,]    5    8
[2,]    3    3
[3,]    5    7
[4,]    6    4

> t(M)
      [,1] [,2] [,3] [,4]

```

```

[1,] 5 3 5 6
[2,] 8 3 7 4

> M[ c(3,1), ]
      [,1] [,2]
[1,] 5 7
[2,] 5 8

> M[ order( M[,1] ), ]
      [,1] [,2]
[1,] 3 3
[2,] 5 8
[3,] 5 7
[4,] 6 4

> M[ order( M[,1], M[,2] ), ]
      [,1] [,2]
[1,] 3 3
[2,] 5 7
[3,] 5 8
[4,] 6 4

> M <- matrix( c(1:12), nrow=3, ncol=4 )
> M
      [,1] [,2] [,3] [,4]
[1,] 1 4 7 10
[2,] 2 5 8 11
[3,] 3 6 9 12

> M <- matrix( c(1:12), nrow=3, ncol=4, byrow=TRUE )
> M
      [,1] [,2] [,3] [,4]
[1,] 1 2 3 4
[2,] 5 6 7 8
[3,] 9 10 11 12

> head(M, 2)
      [,1] [,2] [,3] [,4]
[1,] 1 2 3 4
[2,] 5 6 7 8

> tail(M, 2)
      [,1] [,2] [,3] [,4]
[2,] 5 6 7 8
[3,] 9 10 11 12

> sqrt(M)
      [,1] [,2] [,3] [,4]
[1,] 1.000000 1.414214 1.732051 2.000000
[2,] 2.236068 2.449490 2.645751 2.828427
[3,] 3.000000 3.162278 3.316625 3.464102

> rownames(M) <- c("a", "b", "c")

```



```

> colnames(M) <- c("X1", "X2", "X3", "X4")
> M
  X1 X2 X3 X4
a  1  2  3  4
b  5  6  7  8
c  9 10 11 12

```

1.5. Data frame

```

> x <- c(5, 8, 11, 3, 2, 9, 4)
> y <- c("Y", "Y", "N", "Y", "N", "N", "Y")
> df <- data.frame(x,y)
> df
  x y
1  5 Y
2  8 Y
3 11 N
4  3 Y
5  2 N
6  9 N
7  4 Y

> str(df)
'data.frame':  7 obs. of  2 variables:
 $ x: num  5 8 11 3 2 9 4
 $ y: chr  "Y" "Y" "N" "Y" ...

> df$x
[1]  5  8 11  3  2  9  4

> df$y
[1] "Y" "Y" "N" "Y" "N" "N" "Y"

> df$x[4]
[1] 3

> df[,1]
[1]  5  8 11  3  2  9  4

> df[5, ]
  x y
5 2 N

> df[, "x"]
[1]  5  8 11  3  2  9  4

> df$z <- seq(from=1, to=14, by=2)
> str(df)
'data.frame':  7 obs. of  3 variables:
 $ x: num  5 8 11 3 2 9 4
 $ y: chr  "Y" "Y" "N" "Y" ...
 $ z: num  1 3 5 7 9 11 13

```

```

> df[ 3, c("x","z") ]
      x z
3 11 5

> df[ c(5,7), c(2,3) ]
      y  z
5 N   9
7 Y 13

> df$x[ df$z <= 5 ]
[1]  5  8 11

> df$x[ df$y == "N" ]
[1] 11  2  9

> df[ df$z <= 5, c("x","z") ]
      x z
1  5 1
2  8 3
3 11 5

```

1.6. Полезни функции

```

getwd()
setwd(dir)
save(...)
save.image(...)
read.table(file)
write.table(x, file)
replace(x, list, values)
ifelse(test, yes, no)
any(...)
all(...)
unique(...)
duplicated(...)
is.element(x, y)
x %in% y
tabulate(...)
substr(x, start, stop)

```

2. Случайни експерименти

*The most important questions of life are indeed,
for the most part, only problems of probability.*

*The theory of probability is only
common sense reduced to calculation.*

Pierre Simon Laplace

Случаен експеримент наричаме експеримент, при който не знаем предварително какъв ще бъде резултата (изхода), но знаем какви са възможните изходи. Пример – при хвърляне на зар знаем, че ще се падне някоя от страните на зара, но не знаем коя.

Нека A е някакво събитие, което може да се случи при извършване на експеримента (или да не се случи). Например, при хвърляне на зар – пада се нечетно число.

На всяко събитие съпоставяме число между 0 и 1, което наричаме *вероятност* на събитието. Вероятността на събитието A означаваме с $\mathbf{P}(A)$.

Повтаряме експеримента n пъти, при едни и същи условия. Да означим с $c_n(A)$ броя случвания на събитието A при n повторения на експеримента, т.е. събитието A се е случило $c_n(A)$ пъти. Тогава за достатъчно големи n е изпълнено

$$\frac{c_n(A)}{n} \approx \mathbf{P}(A).$$

Това твърдение следва от т.нар. *закон за големите числа* в теорията на вероятностите. Например, ако хвърляме зар 1000 пъти и $A = \{\text{пада се нечетно число}\}$ и нечетно число се падне $c_n(A)$ пъти, $c_n(A)/1000$ е приблизително равно на вероятността на A .

Числото $\frac{c_n(A)}{n}$ наричаме *честота* на събитието A . Понякога се нарича относителна честота (*relative frequency*). Законът за големите числа твърди, че при достатъчно повторения на експеримента, честотата на случване на събитието A ще приближава вероятността на A . В известен смисъл вероятността $\mathbf{P}(A)$ е дефинирана така, че $\frac{c_n(A)}{n}$ да клони към $\mathbf{P}(A)$.

Нека B е друго събитие, което може да се случи при извършване на експеримента. Условна вероятност на A при условие B , т.е. вероятността да се случи A , ако имаме информация, че се е случило B , се дефинира така $\mathbf{P}(A|B) = \mathbf{P}(AB)/\mathbf{P}(B)$.

Означаваме с $c_n(AB)$ броя случвания на събитията A и B едновременно. Тогава за достатъчно големи n е изпълнено

$$\frac{c_n(AB)}{c_n(B)} \approx \mathbf{P}(A|B).$$

Числото $\frac{c_n(AB)}{c_n(B)}$ ни показва колко често се е случило събитието A , ако броим само експериментите, при които се е случило събитието B .

В следващите задачи ще намерим приближение за вероятността на дадено събитие като симулираме експеримента достатъчен брой пъти с помощта на \mathbf{R} и използваме горните твърдения.

Функцията `sample(x, size, replace)` генерира определен брой (`size`) случайно избрани елементи от вектора `x`, с връщане (`replace=T`) или без връщане (`replace=F`).

Например, по следния начин генерираме 5 случайни числа от вектора $(1, 2, 3, \dots, 10)$, с връщане:

```
> sample( c(1:10), 5, replace=T )
[1] 7 1 6 2 5
> sample( c(1:10), 5, replace=T )
[1] 5 1 1 8 5
```

По следния начин генерираме случайна пермутация на елементите на вектора $(1, 2, \dots, 7)$:

```
> sample( c(1:7), 7, replace=F )
[1] 5 7 1 6 2 3 4
```

Задача 2.1. В отдел на фирма работят 20 човека. За Коледа те решават да си разменят подаръци. В кутия слагат 20 листчета, на всяко от които има едно име. Всеки тегли листче (без да го връща) и подарява на този, чието име е изтеглил. Каква е вероятността поне един да изтегли своето име?

```
sim.gifts <- function(k) {
  x <- sample( c(1:k), k, replace=F )
  d <- x - c(1:k)
  any(d==0)
}

prob.gifts <- function(Nrep, k) {
  rs <- replicate( Nrep, sim.gifts(k) )
  sum(rs)/length(rs)
}

prob.gifts(100000, 20)
```

Функцията `sim.gifts(k)` симулира един експеримент (за `k` човека) и връща `TRUE` ако се е случило събитието поне един да изтегли своето име.

Функцията `prob.gifts(Nrep, k)` повтаря експеримента `Nrep` пъти и връща честотата на случване на събитието (брой случвания разделен на брой повторения), която използваме като приближение на вероятността.

Задача 2.2. Каква е вероятността в група от 25 човека поне двама да имат рожден ден на един и същи ден от годината?

```
sim.bday <- function(k) {
  x <- sample( c(1:365), k, replace=T )
  anyDuplicated(x) > 0
}

prob.bday <- function(Nrep, k) {
  rs <- replicate( Nrep, sim.bday(k) )
  sum(rs)/length(rs)
}

prob.bday(100000, 25)
```

Задача 2.3. Иван има 5 ключа, но не знае кой е за неговата стая. Той пробва последователно с всеки от тях, като помни кой ключ е пробвал. Каква е вероятността да отключи с петия ключ?

```
sim.keys <- function() {  
  x <- sample( c(1:5), 5, replace=F )  
  x[5]==1  
}
```

```
prob.keys <- function(Nrep) {  
  rs <- replicate( Nrep, sim.keys() )  
  sum(rs)/length(rs)  
}
```

```
prob.keys(100000)
```

Задача 2.4. На всеки от върховете на равностраничен триъгълник има една мравка. Всяка мравка избира произволно един от другите два върха и тръгва към него. За единица време всяка мравка изминава разстоянието от един връх до друг. Две мравки могат да се разминат ако тръгнат една срещу друга. Каква е вероятността след единица време да има по една мравка на всеки връх?

```
sim.ants <- function() {  
  a <- vector("numeric", 3)  
  a[1] <- sample( c(2,3), 1 )  
  a[2] <- sample( c(1,3), 1 )  
  a[3] <- sample( c(1,2), 1 )  
  all( c(1,2,3) %in% a )  
}
```

```
prob.ants <- function(Nrep) {  
  rs <- replicate( Nrep, sim.ants() )  
  sum(rs)/length(rs)  
}
```

```
prob.ants(100000)
```

Задача 2.5. Имаме 3 карти: първата е бяла от двете страни, втората е черна от двете страни, а третата е бяла от едната и черна от другата страна. Всяка карта е поставена в затворена кутия. Избираме произволна кутия, отваряме я и виждаме, че горната страна на картата в нея е бяла. Каква е вероятността другата страна на картата също да е бяла?

```
sim.bw <- function() {  
  card <- sample( c("bb", "ww", "bw"), 1 )  
  side <- sample( c(1,2), 1 )  
  up <- substr( card, start=side, stop=side )  
  c(up, card)  
}
```

```
prob.bw <- function(Nrep) {  
  rs <- replicate( Nrep, sim.bw() )  
  sum(rs[2,]=="ww") / sum(rs[1,]=="w")  
}
```

```
prob.bw(100000)
```

3. Случайни величини

*The human mind treats a new idea the same way
the body treats a strange protein; it rejects it.*

P.B. Medawar

Често с изхода на даден случаен експеримент свързваме някаква числова величина, например сумата от точките при хвърляне на два зара, броя дефектни продукти в дадена партия, времето на живот на батерия. Такава числова величина наричаме *случайна величина* (сл.в.). Предварително не знаем каква ще е стойността на случайната величина, но знаем какви са възможните ѝ стойности. Конкретната стойност на случайната величина при извършване на експеримента се определя еднозначно от изхода му, например ако хвърляме два зара и се падне $\begin{bmatrix} \cdot & \cdot \end{bmatrix}$ и $\begin{bmatrix} \cdot & \cdot \end{bmatrix}$ стойността на сл.в. „сума от точките на двата зара“ ще е 5.

3.1. Дискретни сл.в.

Когато случайната величина може да приема краен брой стойности или стойности от изброимо множество (например целите числа $\{0, \pm 1, \pm 2, \dots\}$ или естествените числа $\{1, 2, 3, \dots\}$), се нарича *дискретна* случайна величина.

Нека $x_1, x_2, \dots, x_k, \dots$ са възможните стойности на дискретната сл.в. X . Вероятността да наблюдаваме стойност x_k при извършване на експеримента означаваме $P(X = x_k)$. Сумата на вероятностите $P(X = x_k)$ е единица: $\sum_k P(X = x_k) = 1$.

Възможните стойности на дискретната случайна величина и техните вероятности обикновено се записват в таблица от вида:

x_1	x_2	\dots	x_k	\dots
$P(X = x_1)$	$P(X = x_2)$	\dots	$P(X = x_k)$	\dots

Дискретната сл.в. е дефинирана, ако знаем възможните ѝ стойности и техните вероятности.

Математическо очакване (средно) на дискретната сл.в. X наричаме числото

$$E(X) = \mu = \sum_k x_k P(X = x_k).$$

Дисперсия на дискретната сл.в. X наричаме числото

$$\text{Var}(X) = \sigma^2 = E(X - \mu)^2 = E(X^2) - \mu^2 = \sum_k (x_k)^2 P(X = x_k) - \mu^2.$$

3.1.1. Бернулиево разпределение

Разглеждаме експеримент (опит), при който може да се случи събитието A , което условно наричаме *успех*, или да не се случи събитието A , т.е. да се случи допълнението $A^c = \bar{A}$, което наричаме *неуспех*. Такъв опит наричаме *Бернулиев опит*. Например, при хвърляне на монета се интересуваме от събитието $A = \{\text{пада се ези}\}$ и го наричаме „успех“, падането на тура ще е „неуспех“. Ако при хвърляне на зар се интересуваме от това дали се е паднала шестлица, за нас „успех“ ще бъде падането на шестлица, а „неуспех“ – падането на число различно от шестлица.

Дефинираме случайна величина X по следния начин:

$$X = \begin{cases} 1, & \text{при } \textit{успех} \\ 0, & \text{при } \textit{неуспех} \end{cases}$$

$$\mathbf{P}(X = 1) = p, \quad \mathbf{P}(X = 0) = 1 - p = q$$

Казваме, че случайната величина X има Бернулиево разпределение с параметър p .

3.1.2. Биномно разпределение

Да разгледаме поредица от n независими Бернулиеви опити с една и съща вероятност за *успех* p и нека $q = 1 - p$. Нека X е броя успехи в тази поредица опити.

$$\mathbf{P}(X = k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, 2, \dots, n.$$

Случайната величина X наричаме биномно разпределена с параметри n, p и означаваме $X \in \text{Bi}(n, p)$ или $X \sim \text{Bi}(n, p)$.

Средно и дисперсия:

$$\mathbf{E}(X) = np, \quad \text{Var}(X) = npq.$$

$$\text{dbinom}(k, n, p) = \mathbf{P}(X = k) = \binom{n}{k} p^k q^{n-k}.$$

$$\text{pbinom}(k, n, p) = \mathbf{P}(X \leq k).$$

`rbinom(N, n, p)` генерира N случайни числа от биномно разпределение с параметри n, p .

Ако X_1 има Бернулиево разпределение с параметър p , то $X_1 \sim \text{Bi}(1, p)$. Ако X_1, X_2, \dots, X_n са независими Бенулиеви сл.в. с параметър p , то сумата им има биномно разпределение: $X_1 + X_2 + \dots + X_n \sim \text{Bi}(n, p)$.

3.1.3. Геометрично разпределение

Разглеждаме поредица от независими Бернулиеви опити с вероятност за *успех* p и нека $q = 1 - p$. Нека X е броя опити до първия успех (включително), с други думи, първият успех е на X -тия опит.

$$\mathbf{P}(X = k) = q^{k-1}p, \quad k = 1, 2, 3, \dots$$

Случайната величина X наричаме геометрично разпределена с параметър p и означаваме $X \sim \text{Ge}(p)$.

Средно и дисперсия:

$$\mathbf{E}(X) = \frac{1}{p}, \quad \text{Var}(X) = \frac{q}{p^2}.$$

$$\text{dgeom}(k - 1, p) = \mathbf{P}(X = k) = q^{k-1}p.$$

$$\text{pgeom}(k - 1, p) = \mathbf{P}(X \leq k).$$

`rgeom(N, p) + 1` генерира N случайни числа от $\text{Ge}(p)$.

⟨!⟩ Понякога случайната величина Y = брой неуспехи преди първия успех в поредица от независими Бернулиеви опити, също се нарича геометрично разпределена. Очевидно $X = Y + 1$. Ще използваме означението $Y \sim \text{Ge}^*(p)$.

$$\mathbf{P}(Y = k) = \mathbf{P}(X = k + 1) = q^k p, \quad k = 0, 1, 2, \dots$$

Средно и дисперсия:

$$\mathbf{E}(Y) = \mathbf{E}(X - 1) = \frac{1}{p} - 1 = \frac{q}{p},$$

$$\text{Var}(Y) = \text{Var}(X - 1) = \text{Var}(X) = \frac{q}{p^2}.$$

$$\text{dgeom}(k, p) = \mathbf{P}(Y = k) = q^k p.$$

$$\text{pgeom}(k, p) = \mathbf{P}(Y \leq k).$$

$$\text{rgeom}(N, p) \quad \text{генерира } N \text{ случайни числа от } \text{Ge}^*(p).$$

3.1.4. Отрицателно биномно разпределение

Разглеждаме поредица от независими Бернулиеви опити с вероятност за *успех* p и нека $q = 1 - p$. Нека X е броя опити до r -тия успех (включително), с други думи, r -тият успех е на X -тия опит (r е фиксирано цяло число).

$$\mathbf{P}(X = k) = \binom{k-1}{r-1} p^r q^{k-r}, \quad k = r, r+1, r+2, \dots$$

Случайната величина X наричаме отрицателно биномно разпределена с параметри r, p и означаваме $X \sim \text{NB}(r, p)$.

Средно и дисперсия:

$$\mathbf{E}(X) = \frac{r}{p}, \quad \text{Var}(X) = \frac{rq}{p^2}.$$

$$\text{dnbinom}(k - r, r, p) = \mathbf{P}(X = k) = \binom{k-1}{r-1} p^r q^{k-r}.$$

$$\text{pnbinom}(k - r, r, p) = \mathbf{P}(X \leq k).$$

$$\text{rnbinom}(N, r, p) + r \quad \text{генерира } N \text{ случайни числа от } \text{NB}(r, p).$$

⟨!⟩ Понякога случайната величина Y = брой неуспехи преди r -тия успех в поредица от независими Бернулиеви опити, също се нарича отрицателно биномно разпределена. Очевидно $X = Y + r$. Ще използваме означението $Y \sim \text{NB}^*(r, p)$.

$$\mathbf{P}(Y = k) = \binom{r+k-1}{r-1} p^r q^k, \quad k = 0, 1, 2, \dots$$

Средно и дисперсия:

$$\mathbf{E}(Y) = \frac{rq}{p}, \quad \text{Var}(Y) = \frac{rq}{p^2}.$$

$$\text{dnbinom}(k, r, p) = \mathbf{P}(Y = k) = \binom{r+k-1}{r-1} p^r q^k.$$

$$\text{pnbinom}(k, r, p) = \mathbf{P}(Y \leq k).$$

$$\text{rnbinom}(N, r, p) \quad \text{генерира } N \text{ случайни числа от } \text{NB}^*(r, p).$$

3.1.5. Хипергеометрично разпределение

В кутия има M бели и $N - M$ черни топки. Вадим n топки без да ги връщаме. Нека X е броя на извадените бели топки.

$$\mathbf{P}(X = k) = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}}, \quad k = 0, 1, 2, \dots, n.$$

Случайната величина X наричаме хипергеометрично разпределена с параметри N, M, n и означаваме $X \sim \text{HG}(N, M, n)$. В горната формула приемаме, че $\binom{n}{k} = 0$ ако $k < 0$ или $n < k$. Вероятностите $\mathbf{P}(X = k)$ са положителни за $\max(0, n - N + M) \leq k \leq \min(M, n)$.

Средно и дисперсия:

$$\mathbf{E}(X) = n \frac{M}{N}, \quad \text{Var}(X) = n \frac{M}{N} \frac{N-M}{N} \frac{N-n}{N-1}.$$

$$\text{dhyper}(k, M, N-M, n) = \mathbf{P}(X = k) = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}}.$$

$$\text{phyper}(k, M, N-M, n) = \mathbf{P}(X \leq k).$$

$\text{rhyper}(R, M, N-M, n)$ генерира R случайни числа от хипергеометрично разпределение с параметри N, M, n .

3.1.6. Поасоново разпределение

Казваме, че случайната величина X има Поасоново разпределение с параметър $\lambda > 0$ и означаваме $X \sim \text{Po}(\lambda)$, ако

$$\mathbf{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \dots$$

Средно и дисперсия:

$$\mathbf{E}(X) = \lambda, \quad \text{Var}(X) = \lambda.$$

$$\text{dpois}(k, \lambda) = \mathbf{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

$$\text{ppois}(k, \lambda) = \mathbf{P}(X \leq k).$$

$\text{rpois}(N, \lambda)$ генерира N случайни числа от Поасоново разпределение с параметър λ .

Поасоновото разпределение може да се използва като апроксимация на биномното за големи стойности на n и малки стойности на p , $np = \lambda$, т.е.

$$\binom{n}{k} p^k q^{n-k} \approx e^{-\lambda} \frac{\lambda^k}{k!}.$$

Разпределение	Функция в R	$\mathbf{P}(X = k)$	k
$\text{Bi}(n, p)$	<code>dbinom(k, n, p)</code>	$\binom{n}{k} p^k q^{n-k}$	$0, 1, 2, \dots, n$
$\text{Ge}(p)$	<code>dgeom($k - 1, p$)</code>	$q^{k-1} p$	$1, 2, 3, \dots$
$\text{Ge}^*(p)$	<code>dgeom(k, p)</code>	$q^k p$	$0, 1, 2, \dots$
$\text{NB}(r, p)$	<code>dnbinom($k - r, r, p$)</code>	$\binom{k-1}{r-1} p^r q^{k-r}$	$r, r + 1, r + 2, \dots$
$\text{NB}^*(r, p)$	<code>dnbinom(k, r, p)</code>	$\binom{r+k-1}{r-1} p^r q^k$	$0, 1, 2, \dots$
$\text{HG}(N, M, n)$	<code>dhyper($k, M, N - M, n$)</code>	$\frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}}$	$0, 1, 2, \dots, n$
$\text{Po}(\lambda)$	<code>dpois(k, λ)</code>	$e^{-\lambda} \frac{\lambda^k}{k!}$	$0, 1, 2, \dots$

3.2. Непрекъснати сл.в.

Случайната величина X наричаме *непрекъсната*, ако съществува неотрицателна функция $f(x)$, такава че

$$\mathbf{P}(a \leq X \leq b) = \int_a^b f(x) dx.$$

Функцията $f(x)$ наричаме *плътност* на случайната величина X . С други думи, вероятността да наблюдаваме стойност в интервала $[a, b]$ е равна на интеграл от плътността в граници от a до b .

Непрекъснатата сл.в. може да приема произволни реални стойности от даден интервал. *Математическо очакване* (средно) на непрекъснатата сл.в. X наричаме числото

$$\mathbf{E}(X) = \mu = \int_{-\infty}^{\infty} x f(x) dx.$$

Дисперсия на непрекъснатата сл.в. X наричаме числото

$$\text{Var}(X) = \sigma^2 = \mathbf{E}(X - \mu)^2 = \mathbf{E}(X^2) - \mu^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2.$$

Стандартно отклонение на сл.в. X наричаме числото $\sigma = \sqrt{\text{Var}(X)}$.

Функцията $F(q) = \mathbf{P}(X \leq q)$ се нарича *функция на разпределение* на сл.в. X .

Обратната функция $Q(p) = F^{-1}(p)$ се нарича *квантилна функция* на сл.в. X . Ако $F(q)$ е строго растяща, $Q(p) = q \iff \mathbf{P}(Y \leq q) = p$. В общия случай квантилната функция се дефинира така: $Q(p) = \inf\{q : F(q) \geq p\}$, $p \in (0, 1)$.

3.2.1. Равномерно разпределение

Случайната величина X наричаме равномерно разпределена в интервала (a, b) и означаваме $X \sim U(a, b)$ ако нейната плътност има вида:

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in (a, b), \\ 0, & x \notin (a, b). \end{cases}$$

Средно и дисперсия:

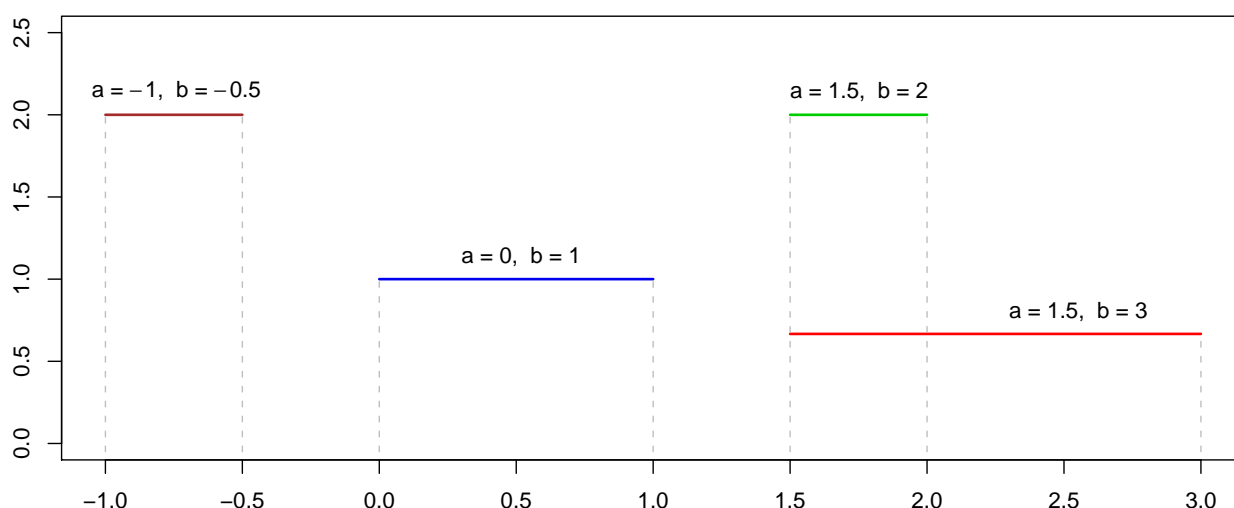
$$E(X) = \frac{a+b}{2}, \quad \text{Var}(X) = \frac{(b-a)^2}{12}.$$

`dunif(x, a, b) = f(x).`

`punif(q, a, b) = P(X ≤ q) = F(q).`

`qunif(p, a, b) = Q(p) = F-1(p).`

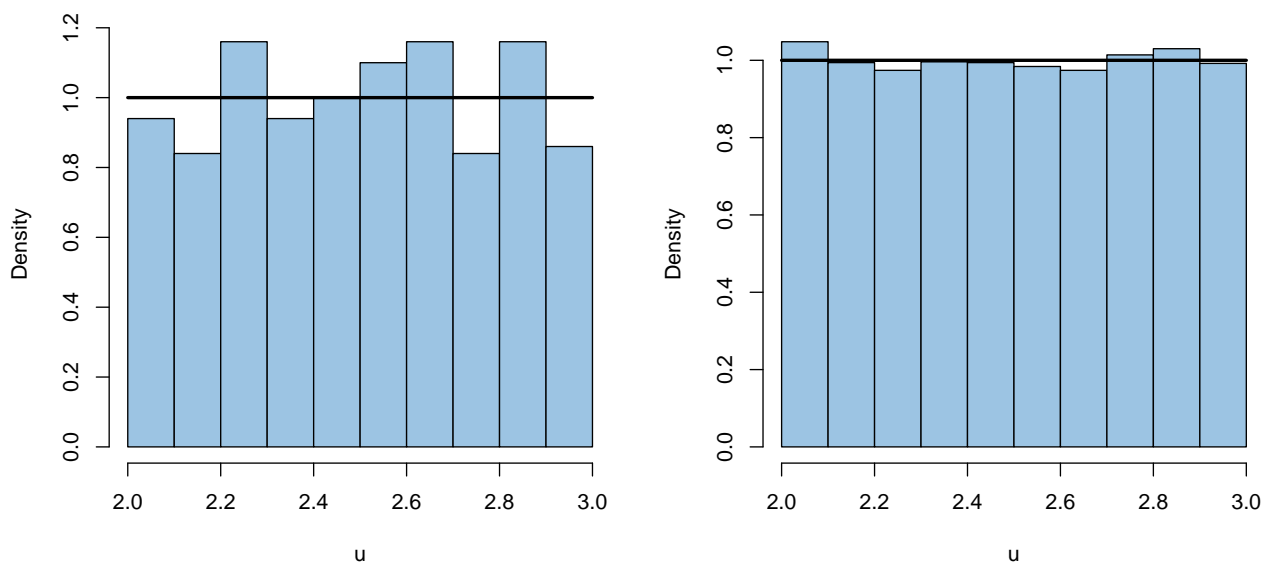
`runif(N, a, b)` генерира N случайни числа от равномерно разпределение в интервала (a, b) .



Фигура 3.1. Равномерно разпределение: графики на плътността при различни стойности на параметрите.

Задача 3.1. Генерирайте 500 случайни числа от равномерно разпределение в интервала $(2, 3)$. Начертайте хистограма на генерираните числа и на същата картинка добавете графика на плътността $f(x)$. Повторете същото с 5000 случайни числа.

```
> u <- runif(500, 2, 3)
> hist( u, probability=T )
> curve( dunif(x, 2, 3), from=2, to=3, add=T, lwd=2.5 )
```



Фигура 3.2.

3.2.2. Експоненциално разпределение

Случайната величина X наричаме експоненциално разпределена с параметър $\lambda > 0$ и означаваме $X \sim \text{Exp}(\lambda)$ ако нейната плътност има вида:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Средно и дисперсия:

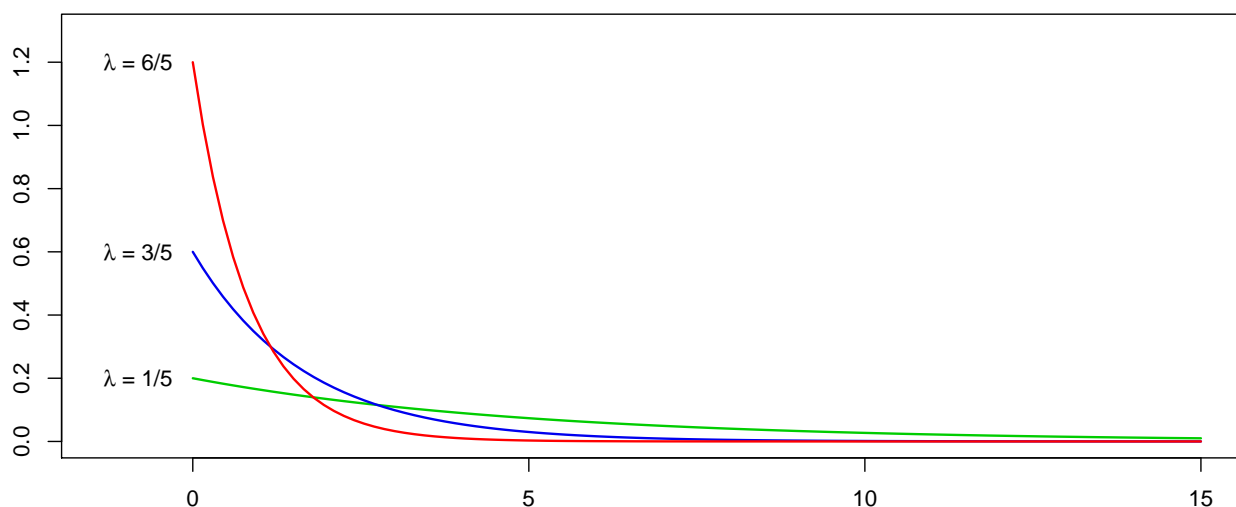
$$E(X) = \frac{1}{\lambda}, \quad \text{Var}(X) = \frac{1}{\lambda^2}.$$

$$\text{dexp}(x, \lambda) = f(x).$$

$$\text{pexp}(q, \lambda) = \mathbf{P}(X \leq q) = F(q).$$

$$\text{qexp}(p, \lambda) = Q(p) = F^{-1}(p).$$

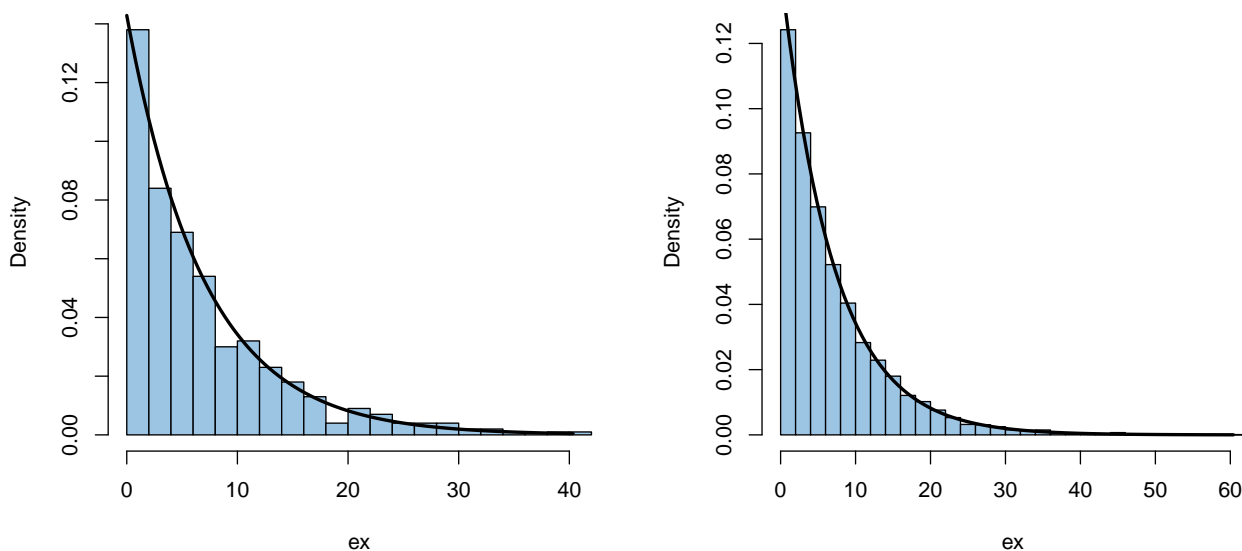
$\text{rexp}(N, \lambda)$ генерира N случайни числа от експоненциално разпределение с параметър λ .



Фигура 3.3. Експоненциално разпределение: графики на плътността при стойности на параметъра $\lambda = 6/5, 3/5, 1/5$.

Задача 3.2. Генерирайте 500 случайни числа от експоненциално разпределение с параметър $\lambda = 1/7$. Начертайте хистограма на генерираните числа и на същата картинка добавете графика на плътността $f(x)$. Повторете същото с 5000 случайни числа.

```
> ex <- rexp(500, rate=1/7)
> hist( ex, probability=T )
> curve( dexp(x, rate=1/7), from=0, to=max(ex), add=T, lwd=2.5 )
```



Фигура 3.4.

3.2.3. Нормално разпределение

Случайната величина X наричаме нормално разпределна с параметри μ , σ^2 и означаваме $X \sim \mathcal{N}(\mu, \sigma^2)$ ако нейната плътност има вида:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}.$$

Средно и дисперсия:

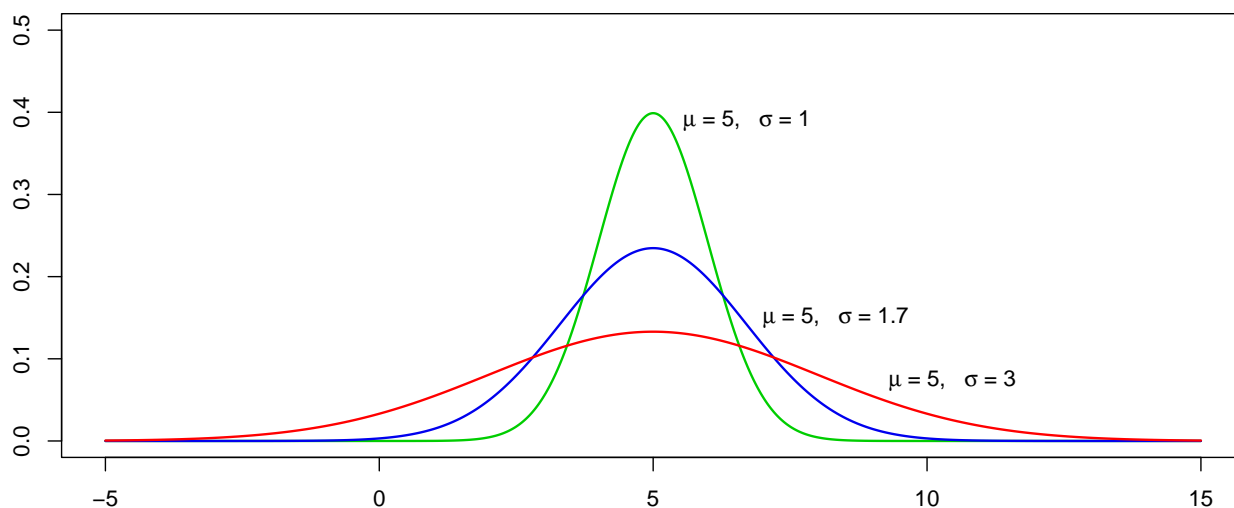
$$E(X) = \mu, \quad \text{Var}(X) = \sigma^2.$$

$$\text{dnorm}(x, \mu, \sigma) = f(x).$$

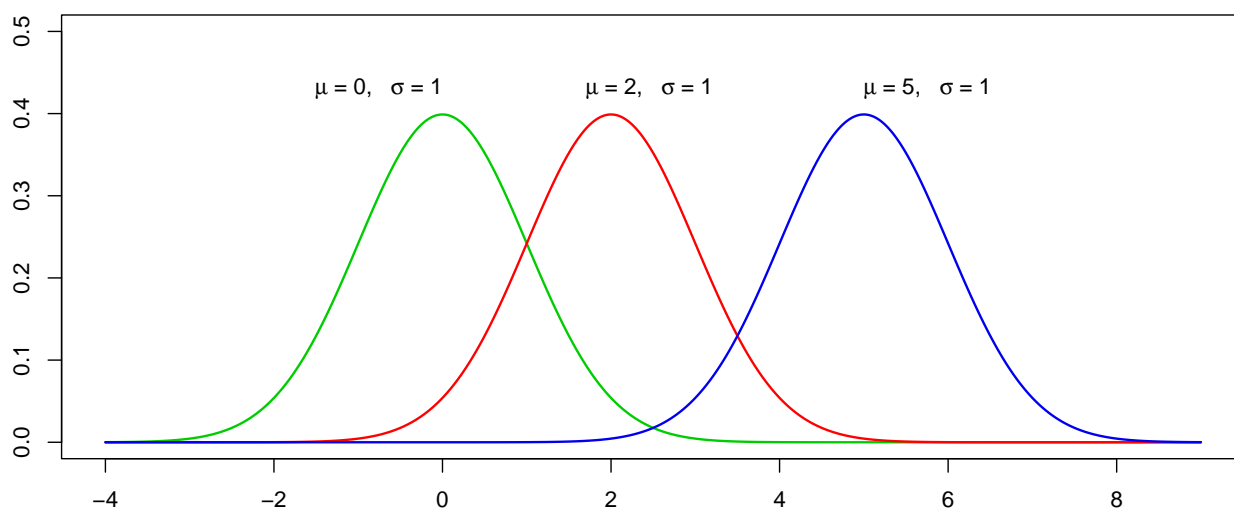
$$\text{pnorm}(q, \mu, \sigma) = \mathbf{P}(X \leq q) = F(q).$$

$$\text{qnorm}(p, \mu, \sigma) = Q(p) = F^{-1}(p).$$

$\text{rnorm}(N, \mu, \sigma)$ генерира N случайни числа от нормално разпределение с параметри μ, σ .



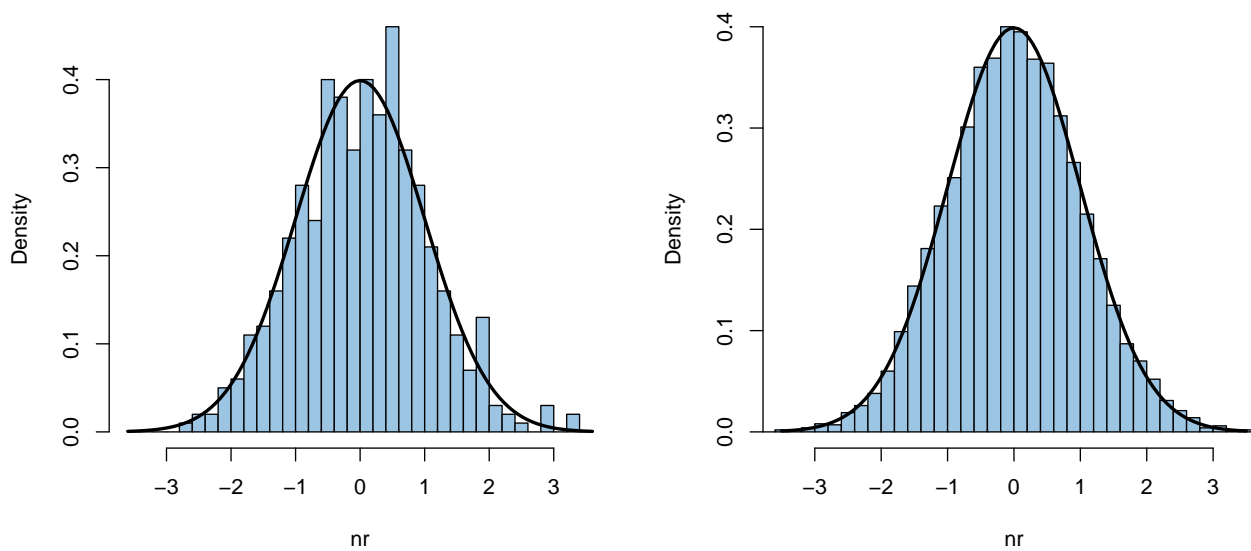
Фигура 3.5. Нормално разпределение: графики на плътността при $\mu = 5$ и няколко стойности на σ .



Фигура 3.6. Нормално разпределение: графики на плътността при няколко стойности на μ и $\sigma = 1$.

Задача 3.3. Генерирайте 500 случайни числа от нормално разпределение с параметри $\mu = 0$, $\sigma = 1$. Начертайте хистограма на генерираните числа и на същата картинка добавете графика на плътността $f(x)$. Повторете същото с 5000 случайни числа.

```
> nr <- rnorm(500, 0, 1)
> hist( nr, probability=T, xlim=c(-3.5,3.5) )
> curve( dnorm(x, 0, 1), add=T, lwd=2.5 )
```



Фигура 3.7.

Разпределение	Функция в R	$f(x)$	x
$U(a, b)$	<code>dunif(x, a, b)</code>	$\frac{1}{b - a}$	(a, b)
$\text{Exp}(\lambda)$	<code>dexp(x, λ)</code>	$\lambda e^{-\lambda x}$	$[0, \infty)$
$\mathcal{N}(\mu, \sigma^2)$	<code>dnorm(x, μ, σ)</code>	$\frac{1}{\sigma\sqrt{2\pi}} e^{-(x - \mu)^2/2\sigma^2}$	$(-\infty, \infty)$

4. Данни. Таблицы и графики

"Data! Data! Data!" he cried impatiently.

"I can't make bricks without clay."

Sherlock Holmes

(Adventures of the Copper Breeches by A.C. Doyle)

Нека X е някаква променлива, от която се интересуваме, например пулса на човек при определена ситуация, времето на безотказна работа на машина, броя пътници в метрото за един ден, съдържанието на калий в един портокал и т.н. Записването на стойността на променливата X наричаме *наблюдение*. Обикновено, за да „изучим“ променливата X правим многократно наблюдения. Съвкупността от наблюдавани стойности на променливата X наричаме *данни* за X и ще означаваме: x_1, x_2, \dots, x_n , като n е броя на наблюденията, които сме направили. Например, ако измерим съдържанието на калий в 30 портокала, ще имаме 30 наблюдения: x_1, \dots, x_{30} над променливата X = съдържание на калий в един портокал.

Множеството от всички възможни наблюдения, които можем да направим над една променлива наричаме *популация* или генерална съвкупност. Понякога популацията се отъждествява с множеството от обекти, които можем да наблюдаваме, например населението на България, персонала на дадена фирма, потребителите на даден продукт; тези популации имат краен брой елементи. Но ако правим лабораторен експеримент, който може да бъде повторен многократно, популацията е съвкупността от всички възможни експерименти и е безкрайна.

Тази част от популацията, която реално наблюдаваме, се нарича *извадка*. Например, ако не можем да наблюдаваме всички потребители на даден продукт, избираме по някакъв (случаен) начин част от тях и правим наблюдения само върху тази част. Когато правим лабораторен експеримент, го повтаряме само краен брой пъти и направените експерименти са нашата извадка. Извадката винаги има краен брой елементи.

Типове данни. Има два основни типа данни – числови (количествени) и категорни (не-количествени).

Числови данни – стойностите на наблюдаваната променлива са числа. Например брой пътници в метрото, температура на въздуха, пулс, брой продадени билети.

Категорни данни – стойностите на променливата (наричаме ги *категории* или *нива*) нямат никакви числови свойства. Например пол, цвят на очите, населено място, кръвна група, майчин език, марка телефон и т.н. Категориите на дадена променлива са взаимно изключващи се.

За удобство при обработка на данните, категориите обикновено се кодират с числа, например 0 = здрав, 1 = болен или 1 = кафяв, 2 = черен, 3 = син, 4 = зелен. Естествено, тези кодове са условни, нямат количествен смисъл.

След като са събрани, данните трябва да се представят в някакъв обобщен вид, за да се добие представа за основните им характеристики. Ще разгледаме често използваните таблични и графични техники за представяне на данни.

4.1. Категорни данни

Попитали сме 20 души кой интернет браузър използват най-често (допитването е направено през юли 2011). Записали сме отговорите във файла `browsers.txt`. Прочитаме данните в R по следния начин:

```
> dt <- read.table("browsers.txt")
```

Записваме ги във вектора `brows`:

```
> brows <- dt$V1
> brows
[1] "IE"      "Firefox" "Firefox" "IE"      "IE"      "Chrome"  "Firefox"
[8] "Firefox" "IE"      "Chrome"  "Firefox" "IE"      "Chrome"  "IE"
[15] "IE"      "Chrome"  "Firefox" "IE"      "Safari"  "Opera"
```

```
> class(brows)
[1] "character"
```

В случая имаме 20 наблюдения над категорна променлива с 5 категории: *Chrome*, *Firefox*, *IE*, *Opera*, *Safari*. Може да представим данните в таблица, показваща колко пъти се среща всяка от категориите:

```
> table(brows)
brows
Chrome Firefox      IE  Opera  Safari
      4       6       8       1       1
```

Виждаме, че *Chrome* ползват 4 човека, *Firefox* ползват 6 човека и т.н.

Ако разделим броя срещания на общия брой наблюдения ($n = 20$) ще получим каква част (процент) от хората са използват съответния браузър:

```
> table(brows)/length(brows)
brows
Chrome Firefox      IE  Opera  Safari
  0.20   0.30   0.40   0.05   0.05
```

От тази таблица разбираме, че *Chrome* ползват 20% от запитаните, *Firefox* ползват 30% от запитаните и т.н.

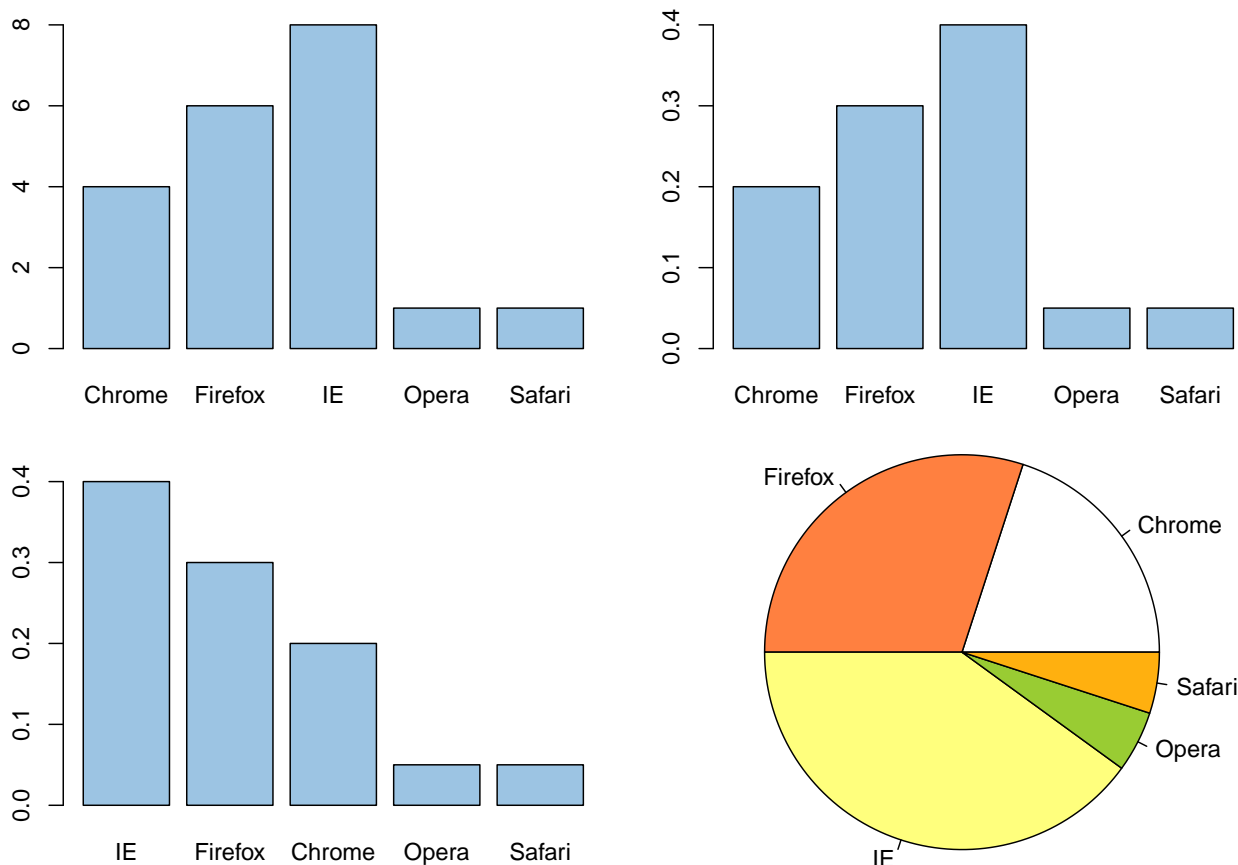
Можем да сортираме таблицата в намаляващ ред:

```
> sort( table(brows)/length(brows), decreasing=T )
brows
      IE Firefox  Chrome  Opera  Safari
  0.40   0.30   0.20   0.05   0.05
```

С помощта на функцията `barplot` представяме съответната таблица във вид на графика, в която всяка категория е представена със стълб с височина равна на съответната стойност от таблицата.

С функцията `pie` получаваме кръгова диаграма – всяка от категориите е представена като сектор от един кръг (с ъгъл, пропорционален на броя срещания).

```
> barplot( table(brows) )
> barplot( table(brows)/length(brows) )
> barplot( sort(table(brows)/length(brows), decreasing=T) )
> pie( table(brows) )
```



Фигура 4.1. Графично представяне на категориални данни

4.2. Числови данни

Иван си е отбелязвал времето (в минути) на чакане на автобуса всяка сутрин в продължение на 25 дни. Записваме данните във вектора `wait`:

```
> wait <- c(2,3,3,5,5,2,7,10,4,3,1,7,11,10,5,6,3,8,5,12,5,3,8,5,7)
```

Отново може да представим данните в таблица:

```
> table(wait)
wait
 1  2  3  4  5  6  7  8 10 11 12
 1  2  5  1  6  1  3  2  2  1  1
```

Таблицата показва, че една минута е чакал само веднъж, 2 минути – 2 пъти, 3 минути – 5 пъти и т.н.

Числовите данни обикновено могат да приемат много на брой стойности. Затова използването на подобна таблица не винаги е удачно. Вместо това, интервалът от възможни стойности се разбива на подинтервали (равни по дължина) и се прави таблица, показваща броя наблюдения във всеки подинтервал. В случая ще разделим интервала $(0, 12]$ на 6 подинтервала.

```
> wait.grp <- cut( wait, breaks=seq(0,12,2) )
> table(wait.grp)
```

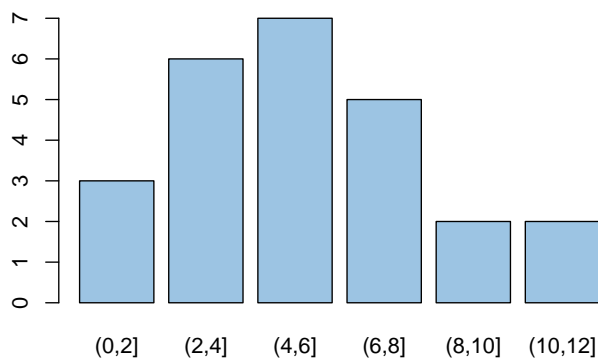
```
wait.grp
(0,2] (2,4] (4,6] (6,8] (8,10] (10,12]
      3      6      7      5      2      2
```

```
> table(wait.grp)/length(wait)
wait.grp
(0,2] (2,4] (4,6] (6,8] (8,10] (10,12]
 0.12  0.24  0.28  0.20  0.08  0.08
```

От първата таблица разбираме, че две или по-малко минути е чакал 3 пъти, от 2 до 4 минути е чакал 6 пъти и т.н. Втората таблица е с проценти: от 2 до 4 минути е чакал в 24% от дните, най-често е чакал между 4 и 6 минути – в 28% от дните.

Като използваме функцията `barplot` може да представим получената таблица във вид на графика, в която всеки подинтервал е представен със стълб с височина равна на броя наблюдения в подинтервала (фиг. 4.2).

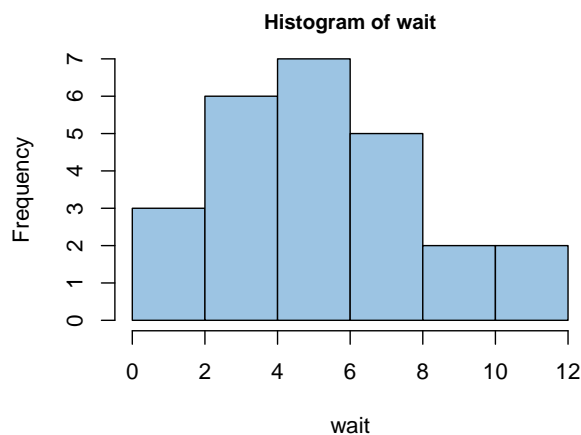
```
> barplot( table(wait.grp) )
```



Фигура 4.2.

Подобна графика получаваме и с функцията `hist` – прилагаме я за вектора `wait`; тя разделя интервала на подинтервали и за всеки подинтервал рисува стълб с височина равна на броя наблюдения в подинтервала (фиг. 4.3). Такава графика се нарича *хистограма*.

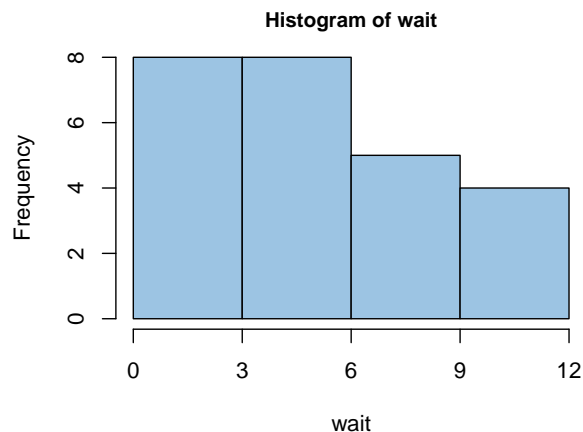
```
> hist(wait)
```



Фигура 4.3.

Функцията `hist` сама определя какви да са подинтервалите. Но ако искаме може да зададем какви да бъдат, например може да разделим интервала $(0, 12]$ на 4 подинтервала:

```
> hist(wait, breaks=seq(0,12,3))
```



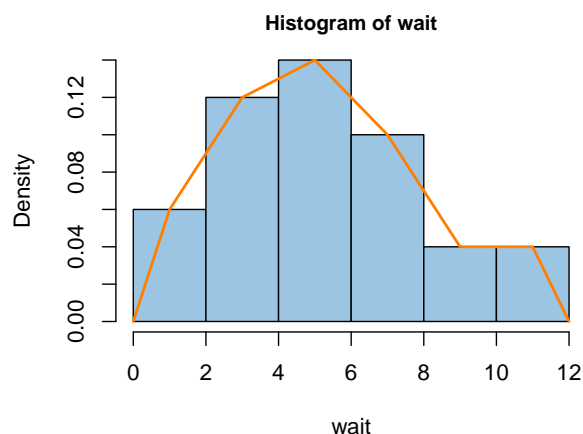
Фигура 4.4.

⟨!⟩ По подразбиране `hist` разделя интервала от стойности $[a, b]$ по следния начин: $[a, c_1]$ $(c_2, c_3]$ $(c_4, c_5]$ \dots $(c_n, b]$, т.е. първият подинтервал е от вида $[,]$, а останалите $(,]$.

Командата `hist(..., probability=T)` чертае хистограма, така че сумата от лицата на всички стълбове (правоъгълници) е единица. Лицето на даден стълб е равно на честотата на данните в съответния интервал. Различава се от хистограмата `hist(...)` само по скалата на оста y .

Оранжевата линия на Фиг. 4.5 се нарича честотен полигон.

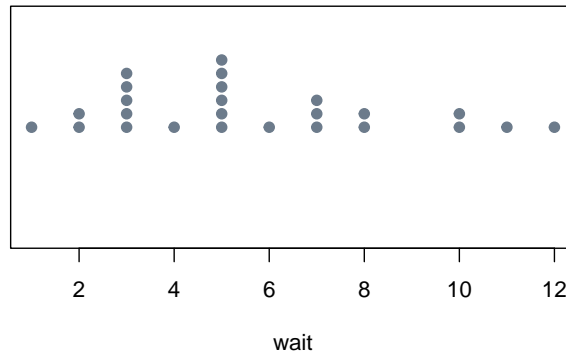
```
h <- hist(wait, probability=T)
lines( x=c( min(h$breaks), h$mids, max(h$breaks) ),
       y=c( 0, h$density, 0 ), type="l", lwd=2, col="darkorange1" )
```



Фигура 4.5.

Друг начин за графично представяне на числови данни е показан на Фиг. 4.6. На графиката всяко наблюдение е представено с кръгче.

```
stripchart(wait, method="stack", pch=20, cex=1.5)
```



Фигура 4.6.

Числови данни могат да бъдат преставени и чрез диаграмата „клон с листа“ (*stem-and-leaf plot*). Подходяща е при малък обем на извадката (малко наблюдения). В известен смисъл е вариант на хистограмата от времената, когато графичните възможности на компютрите са били по-ограничени.

Да се върнем на данните `wait`. При разделяне на интервала на $[0, 4]$ $(4, 9]$ $(9, 12]$ се получава:

```
> stem(wait)
```

```
The decimal point is 1 digit(s) to the right of the |
```

```
0 | 122333334
0 | 555555677788
1 | 0012
```

С параметъра `scale` може да увеличим броя на подинтервалите (например, ако зададем `scale=2` интервалите ще са повече, но не е ясно колко ще са). На следващата диаграма са $[0, 1]$ $(1, 3]$ $(3, 5]$... $(9, 11]$ $(11, 13]$, всяко наблюдение в съответния интервал е представено с 0:

```
> stem(wait, scale=2)
```

```
The decimal point is at the |
```

```
0 | 0
2 | 0000000
4 | 0000000
6 | 0000
8 | 00
10 | 000
12 | 0
```

На следващата картинка, интервалите са $(0, 1]$ $(1, 2]$ $(2, 3]$... $(11, 12]$, отново всяко наблюдение в съответния интервал е представено с 0, т.е. показва ни, че 1 се среща един път, 2 – два пъти, 3 – пет пъти, 4 – един път и т.н.

```
> stem(wait, scale=3)
```

The decimal point is at the |

1		0
2		00
3		00000
4		0
5		000000
6		0
7		000
8		00
9		
10		00
11		0
12		0

От трите получени картинки, при първата и третата данните могат да бъдат възстановени еднозначно, докато при втората (`scale=2`) не могат – например в интервала $[2, 3]$ всички наблюдения са означени по един начин и не знаем колко от тях са '2' и колко '3'.

5. Числови характеристики на данните

*Nothing in life is to be feared,
it is only to be understood.*

Marie Curie

Нека x_1, x_2, \dots, x_n са наблюдения над някаква числова променлива X . Ще означаваме най-малкото по големина наблюдение с $x_{(1)}$, следващото по големина с $x_{(2)}$, и т.н., най-голямото с $x_{(n)}$, т.е. $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$.

Средна стойност (средно) на данните x_1, x_2, \dots, x_n наричаме числото

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_n}{n}.$$

Нарича се още извадъчно средно.

Медиана на данните x_1, x_2, \dots, x_n наричаме числото

$$\widehat{Me} = \begin{cases} x_{(\frac{n+1}{2})}, & \text{при нечетно } n \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}, & \text{при четно } n \end{cases}$$

Нарича се още извадъчна медиана. По-малки или равни на медианата са поне половината от данните и по-големи или равни са също поне половината от данните. Грубо казано, медианата разделя данните на две равни части.

p-квантил на данните x_1, x_2, \dots, x_n наричаме числото, от което са по-малки или равни поне $100p\%$ от данните и са по-големи или равни поне $100(1-p)\%$. Грубо казано, *p*-квантилът разделя данните на две части, съответно от $100p\%$ и останалите $100(1-p)\%$. Нарича се още извадъчен *p*-квантил.

0.5-квантилът е всъщност медианата, 0.25-квантилът се нарича *първи квартил* (Q_1), а 0.75-квантилът се нарича *трети квартил* (Q_3). Разликата между третия и първия квартил ($Q_3 - Q_1$) се нарича *интерквартилен размах* (*IQR*).

Стандартно отклонение на данните x_1, x_2, \dots, x_n наричаме числото

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}}.$$

Нарича се още извадъчно стандартно отклонение.

Неравенство на Чебишев. Нека данните x_1, x_2, \dots, x_n имат средна стойност \bar{x} и стандартно отклонение s . За всяко $k > 0$ е вярно, че поне $100(1 - 1/k^2)$ процента от данните лежат в интервала $[\bar{x} - ks, \bar{x} + ks]$. При $k = 3$ получаваме, че поне 88.9% от данните са в интервала $[\bar{x} - 3s, \bar{x} + 3s]$.

Средната стойност и медианата показват центъра на данните в някакъв смисъл. Медианата е център на данните в смисъл, че е по средата на сортираните данни, т.е. приблизително половината от данните са по-малки и приблизително половината са по-големи от нея.

За да изясним в какъв смисъл средната стойност е център на данните, ще покажем, че $\sum_{i=1}^n (x_i - \bar{x}) = 0$. Наистина

$$\begin{aligned}\sum_{i=1}^n (x_i - \bar{x}) &= \sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x} \\ &= n \frac{1}{n} \sum_{i=1}^n x_i - n\bar{x} \\ &= n\bar{x} - n\bar{x} = 0.\end{aligned}$$

Това твърдение означава, че ако сумираме на разликите на всяко наблюдение от средната стойност ще получим нула. С други думи, сумата на положителните разлики е равна на сумата на отрицателните разлики. В този смисъл средната стойност е център на данните – балансира сумата на положителните и сумата на отрицателните разлики.

Стандартното отклонение характеризира разпръскването на данните около средната стойност. То всъщност е корен от усреднения квадрат на разликата на всяко наблюдение от средната стойност (засега няма да изясняваме, защо усреднява се като се дели на $(n - 1)$ на вместо на n). Друга мярка за разпръскването (разсейването) на данните е разликата между най-голямото и най-малкото наблюдение, $x_{(n)} - x_{(1)}$, нарича се *размах*.

Интервалът от неравенството на Чебишев ни дава представа доколко далече от средното може да се простират данните, ако са ни известни \bar{x} и s .

Нека данните x_1, x_2, \dots, x_n са записани във вектора **x**. Горните числови характеристики се пресмятат в R със следните функции:

```
 $\bar{x}$  = mean(x)  
 $\widehat{Me}$  = median(x)  
 $p$ -квантил = quantile(x, p)  
 $Q_1$  = quantile(x, 0.25)  
 $Q_3$  = quantile(x, 0.75)  
 $IQR$  = IQR(x)  
 $s$  = sd(x)
```

Пример 1. Разглеждаме данните `airquality`. В променливата `airquality$Temp` има наблюдения за температурата (градуси по Фаренхайт) в Ню Йорк от май до септември 1973.

Пресмятаме медианата, средната стойност и стандартното отклонение:

```
> temp <- airquality$Temp  
> median(temp)  
[1] 79  
> mean(temp)  
[1] 77.88235  
> sd(temp)  
[1] 9.46527
```

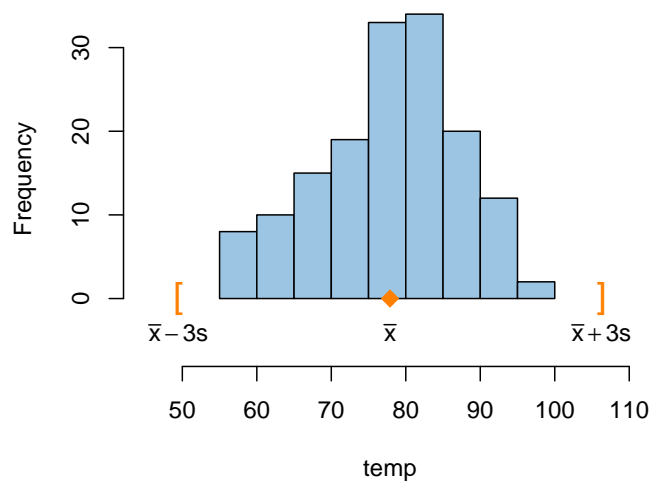
Функцията `summary` ни дава някои от разгледаните числови характеристики:

```
> summary(temp)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 56.00  72.00   79.00   77.88   85.00   97.00
```

Интервалът от неравенството на Чебишев е $[\bar{x} - 3s, \bar{x} + 3s] = [49.5, 106.3]$. Може да забележим, че всички данни са в този интервал.

```
> mean(temp) - 3*sd(temp)
[1] 49.48654
> mean(temp) + 3*sd(temp)
[1] 106.2782
```

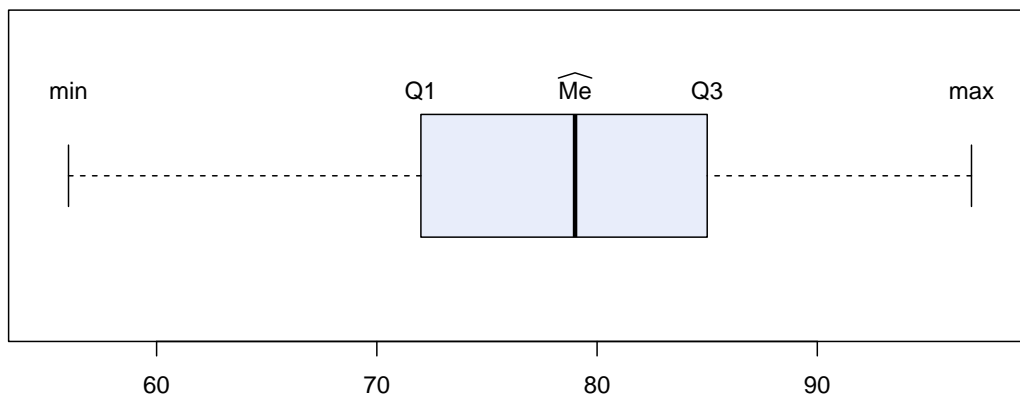
На следващата графика е показана хистограма на температурата и са нанесени средното \bar{x} и интервалът от неравенството на Чебишев:



Фигура 5.1.

Първият квантил, медианата, третият квантил, както и най-малкото и най-голямото наблюдение се изобразяват на графика, наречена *кутия с мустаци*, с помощта на функцията `boxplot`:

```
> boxplot(temp, horizontal=T)
```

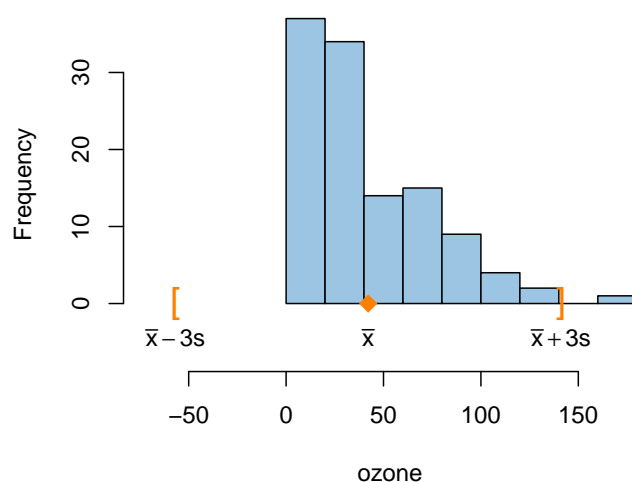


Фигура 5.2. Кутия с мустаци на temp

В данните `airquality` има и измервания на съдържанието на озон във въздуха (променливата `airquality$Ozone`).

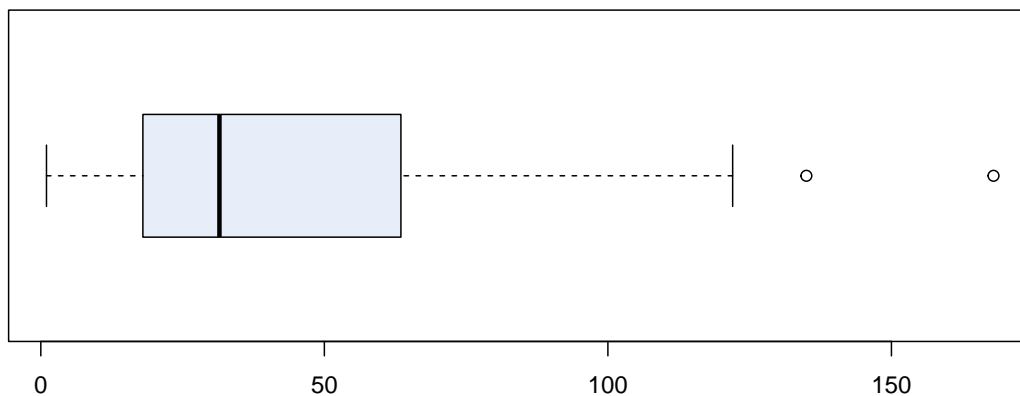
```
> ozone <- airquality$Ozone
> median(ozone, na.rm=T)
[1] 31.5
> mean(ozone, na.rm=T)
[1] 42.12931
> sd(ozone, na.rm=T)
[1] 32.98788
> summary(ozone)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
   1.00  18.00   31.50   42.13  63.25  168.00    37
> mean(ozone, na.rm=T) - 3*sd(ozone, na.rm=T)
[1] -56.83434
> mean(ozone, na.rm=T) + 3*sd(ozone, na.rm=T)
[1] 141.093
```

На следващата графика е дадена хистограма на `ozone` и са нанесени средното \bar{x} и интервалът от неравенството на Чебишев:



Фигура 5.3.

Кутия с мустаци на `ozone`:



Фигура 5.4.

Наблюденията, които не попадат в интервала $[Q_1 - 1.5 * IQR, Q_3 + 1.5 * IQR]$ се изобразяват с кръгче. В случая има две наблюдения извън този интервал. Такива наблюдения се считат за необичайни (*outliers*).

```
> quantile(ozone, 0.25, names=F, na.rm=T) - 1.5*IQR(ozone, na.rm=T)
[1] -49.875
> quantile(ozone, 0.75, names=F, na.rm=T) + 1.5*IQR(ozone, na.rm=T)
[1] 131.125
```

Задача 5.1. Да се намерят (на ръка и с R) средното, медианата, стандартното отклонение, първия и третия квартил на данните: 3, 6, 10, 0, 8, 3, 7, 2, 6.

Средното е:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{3 + 6 + 10 + 0 + 8 + 3 + 7 + 2 + 6}{9} = \frac{45}{9} = 5.$$

За да намерим s , от всяко наблюдение изваждаме \bar{x} (третата колона) и повдигаме на квадрат (четвъртата колона):

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	3	-2	4
2	6	1	1
3	10	5	25
4	0	-5	25
5	8	3	9
6	3	-2	4
7	7	2	4
8	2	-3	9
9	6	1	1
Σ	45	0	82

Стандартното отклонение е:

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{82}{9 - 1}} = 3.20.$$

За да намерим медианата, подреждаме данните по големина:

0, 2, 3, 3, **6**, 6, 7, 8, 10

В средата е числото 6, следователно $\widehat{Me} = 6$.

За да намерим Q_1 и Q_3 разделяме сортираните данни на две половини, като числото в средата се включва и в двете половини:

0, 2, 3, 3, **6**, 6, 7, 8, 10

0, 2, 3, 3, **6**, 6, 7, 8, 10

Медианата на първата половина е първия квартил, т.е. $Q_1 = 3$, а медианата на втората половина е третия квартил, т.е. $Q_3 = 7$.

6. Многомерни данни

The advanced reader who skips parts that appear to him too elementary may miss more than the less advanced reader who skips parts that appear to him too complex.

G. Polya

Когато наблюдаваме (измерваме) повече от една променлива, наричаме данните многомерни. Те обикновено се записват в таблица от вида:

	X	Y	Z	\dots	W
1	x_1	y_1	z_1	\dots	w_1
2	x_2	y_2	z_2	\dots	w_2
3	x_3	y_3	z_3	\dots	w_3
\vdots	\vdots	\vdots	\vdots		\vdots
n	x_n	y_n	z_n	\dots	w_n

където всеки ред отговаря на едно наблюдение (опит, измерване, участник в изследване и т.н.), а всяка колона отговаря на една променлива. Тук ще се запознаем с някои основни техники за боравене с многомерни данни.

Многомерните данни се представят в R чрез обект наречен *data frame*. Той е подобен на матрица, с тази разлика, че колоните могат да бъдат от различен тип.

Ще разгледаме данните `survey` от пакета `MASS`. Тези данни съдържат отговорите на няколко въпроса, зададени на 237 студенти от курса *Статистика I* в Университета на Аделаида (не е ясно кога е направена анкетата).

За да използваме обекти от даден пакет го зареждаме с командата `library(packageName)`.

```
> library(MASS)
> ?survey
> str(survey)
'data.frame': 237 obs. of 12 variables:
 $ Sex : Factor w/ 2 levels "Female","Male": 1 2 2 2 2 1 2 1 2 2 ...
 $ Wr.Hnd: num 18.5 19.5 18 18.8 20 18 17.7 17 20 18.5 ...
 $ NW.Hnd: num 18 20.5 13.3 18.9 20 17.7 17.7 17.3 19.5 18.5 ...
 $ W.Hnd : Factor w/ 2 levels "Left","Right": 2 1 2 2 2 2 2 2 2 2 ...
 $ Fold : Factor w/ 3 levels "L on R","Neither",...: 3 3 1 3 2 1 1 3 3 3 ...
 $ Pulse : int 92 104 87 NA 35 64 83 74 72 90 ...
 $ Clap : Factor w/ 3 levels "Left","Neither",...: 1 1 2 2 3 3 3 3 3 3 ...
 $ Exer : Factor w/ 3 levels "Freq","None",...: 3 2 2 2 3 3 1 1 3 3 ...
 $ Smoke : Factor w/ 4 levels "Heavy","Never",...: 2 4 3 2 2 2 2 2 2 2 ...
 $ Height: num 173 178 NA 160 165 ...
 $ M.I : Factor w/ 2 levels "Imperial","Metric": 2 1 NA 2 2 1 1 2 2 2 ...
 $ Age : num 18.2 17.6 16.9 20.3 23.7 ...
```

(!) По-нататък в тази глава, когато говорим за променлива, ще разбираме наблюденията над тази променлива (стойностите от съответната колона в таблицата по-горе).

`fix(survey)` – извежда данните като таблица;

`summary(survey)` – числови характеристики за всяка променлива;

`survey[, 'Age']` – променливата `Age`;

`survey$Age` – променливата `Age` (друг начин);
`survey[,12]` – дванадесетата променлива;
`survey[5,]` – петото наблюдение;

Може да се обръщаме към променливите от даден data frame директно (например `Age` вместо `survey$Age`), ако напишем `attach(dataFrameName)`.

```
> summary(Age)
Error in summary(Age) : object 'Age' not found
> summary(survey$Age)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
16.75  17.67   18.58   20.37  20.17   73.00
> attach(survey)
> summary(Age)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
16.75  17.67   18.58   20.37  20.17   73.00
```

Понякога има липсващи наблюдения за дадена променлива. Във R те се означават с `NA`. За да се пресметнат някои числови характеристики (например средно, медиана, стандартно отклонение) на променлива, в която има липсващи наблюдения, тези наблюдения трябва да се игнорират при пресмятането (например, при пресмятане на средно се пресмята средното на останалите). Това се задава с параметъра `na.rm=T`.

```
> mean(Age)
[1] 20.37451
> mean(Pulse)
[1] NA
> mean(Pulse, na.rm=T)
[1] 74.15104
> summary(Pulse)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
35.00  66.00   72.50   74.15  80.00  104.00   45.00
```

По подразбиране R показва 7 значещи цифри след десетичната точка. За следващите примери ще ги намалим на 3:

```
> options(digits=3)
```

Когато наблюдаваме няколко променливи, често се интересуваме от някакви връзки между тях. Да разгледаме променливите `Smoke` и `W.Hnd`. И двете са категорни. Във `Smoke` е записана честотата на пушене, има следните категории: *Heavy*, *Regul*, *Occas*, *Never*. Във `W.Hnd` е записано с коя ръка пише студентът (*Left*, *Right*). Може да представим двете променливи в двумерна таблица (крос-таблица):

```
> table(Smoke, W.Hnd)
      W.Hnd
Smoke  Left Right
Heavy    1    10
Never   13   175
Occas    3    16
Regul    1    16
```

От подобна таблица може да разберем, например, колко от студентите пишат с лявата ръка (*Left*) и не пушат (*Never*). В тази таблица са изключени липсващите наблюдения; те могат да бъдат показани с добавяне на `useNA="always"`.

```
> table(Smoke, W.Hnd, useNA="always")
```

```
      W.Hnd
Smoke Left Right <NA>
Heavy    1    10     0
Never   13   175     1
Occas    3    16     0
Regul    1    16     0
<NA>     0     1     0
```

Вместо броя срещания, двумерната таблица може да показва относителния дял (процент). Ще разгледаме три вида такива таблици. Ако разделим първата таблица на общия брой наблюдения, които са 235 (поради изключването на липсващите), ще получим таблицата:

```
> tab.smoke.hand <- table(Smoke, W.Hnd)
```

```
> prop.table(tab.smoke.hand)
```

```
      W.Hnd
Smoke Left  Right
Heavy 0.00426 0.04255
Never 0.05532 0.74468
Occas 0.01277 0.06809
Regul 0.00426 0.06809
```

От нея може да разберем, например, че 74.5% от студентите пишат с дясната ръка и не пушат.

Ако разделим всеки ред от първата таблица на сумата на реда, получаваме таблица с редови процент; за целта използваме командата `prop.table(tab.smoke.hand, 1)`. От тази таблица може да разберем, например, че 84.2% от пушещите „понякога“ (*Occas*) пишат с дясната ръка.

```
> prop.table(tab.smoke.hand, 1)
```

```
      W.Hnd
Smoke Left  Right
Heavy 0.0909 0.9091
Never 0.0691 0.9309
Occas 0.1579 0.8421
Regul 0.0588 0.9412
```

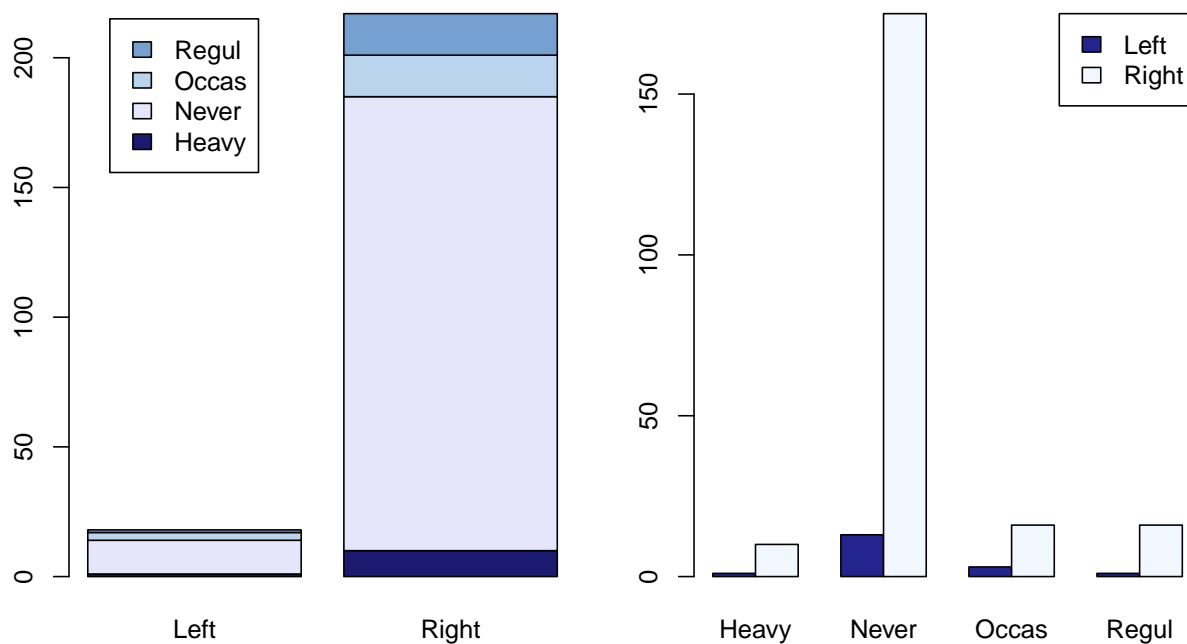
Ако разделим всяка колона от първата таблица на сумата на колоната, получаваме таблица с колонен процент; за целта пишем `prop.table(tab.smoke.hand, 2)`. От тази таблица може да видим, например, че 7.37% от пушещите с дясната ръка пушат „понякога“ (*Occas*).

```
> prop.table(tab.smoke.hand, 2)
```

```
      W.Hnd
Smoke Left  Right
Heavy 0.0556 0.0461
Never 0.7222 0.8065
Occas 0.1667 0.0737
Regul 0.0556 0.0737
```

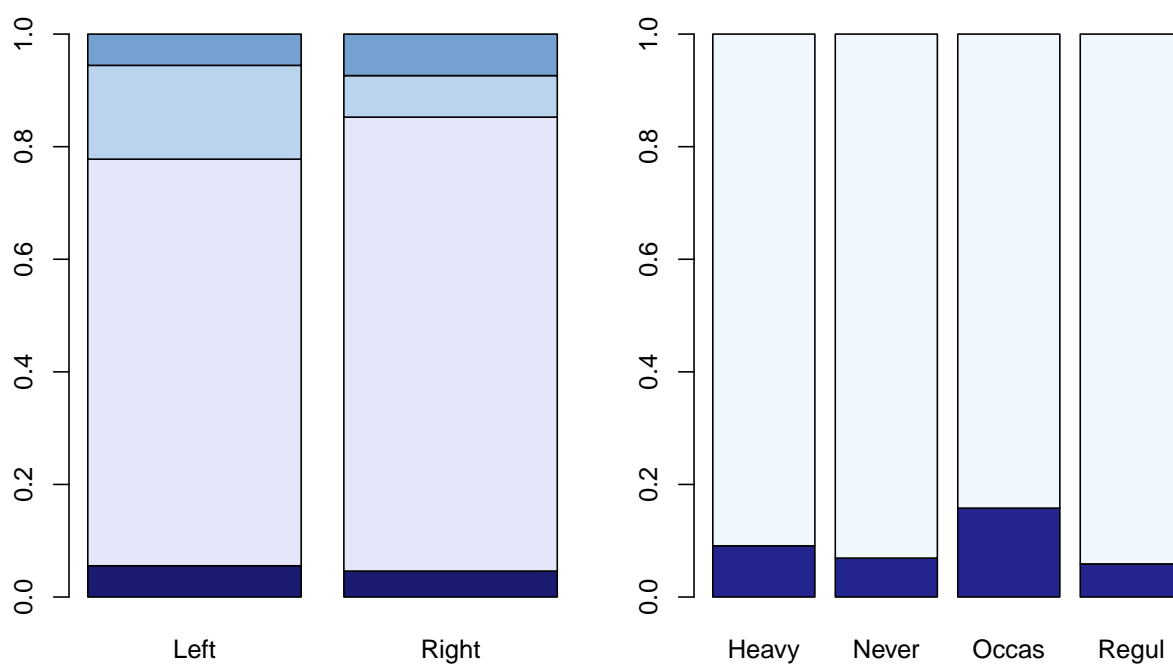
Като използваме функцията `barplot` може да представим горните таблици графично по различни начини:

```
barplot( table(Smoke, W.Hnd), legend=T,
         args.legend=list(x="topleft", inset=0.05) )
barplot( table(W.Hnd, Smoke), beside=T, legend=T,
         args.legend=list(x="topright", inset=0.05) )
```



Фигура 6.1.

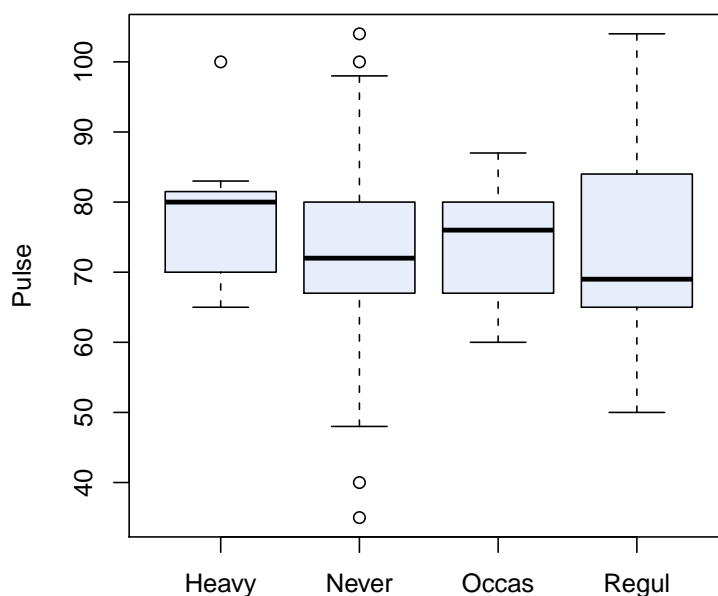
```
barplot( prop.table( tab.smoke.hand, 2 ) )
barplot( t( prop.table( tab.smoke.hand, 1 ) ) )
```



Фигура 6.2.

Може да се интересуваме доколко стойностите на дадена числова променлива се различават при различните нива (категории) на някаква категорна променлива. Да разгледаме числовата променливата `Pulse`, в която е записан пулсът на студента и категорната променлива `Smoke`. Със следната команда получаваме кутия с мустаци на `Pulse` за всяка от категориите на `Smoke` (променливата `Pulse` се разбива по категориите на `Smoke` и за всяка категория се рисува кутия с мустаци).

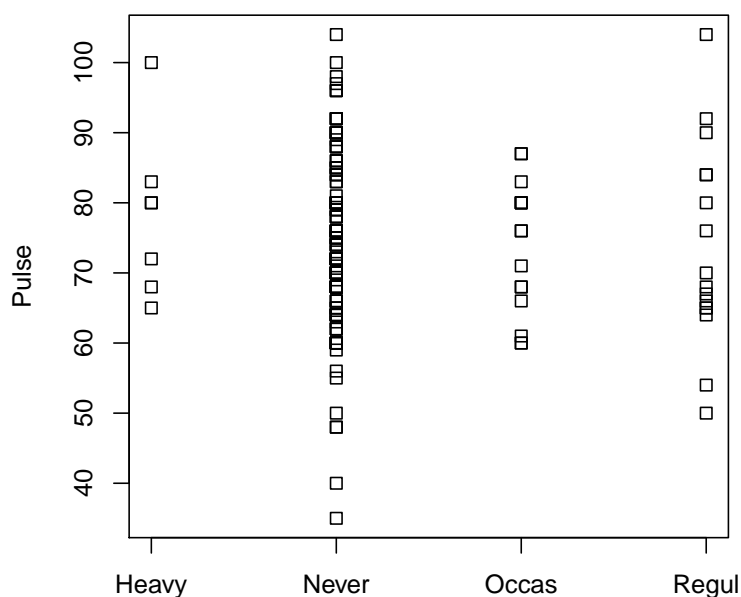
```
> boxplot(Pulse ~ Smoke)
```



Фигура 6.3.

Следната команда представя променливата `Pulse` разбита по категориите на `Smoke`, като всяко наблюдение е изобразено с квадратче:

```
> stripchart(Pulse ~ Smoke, vertical=T)
```



Фигура 6.4.

Нека във вектора **x** са записани наблюденията x_1, \dots, x_n над някаква числова променлива, а във вектора **y** – наблюденията y_1, \dots, y_n над друга числова променлива. Командата **plot(x,y)** изобразява точките $(x_1, y_1), \dots, (x_n, y_n)$ в координатната система Oxy . Изобразяването на две променливи на такава графика, може да ни подсказва за някаква зависимост между тях. Ще изобразим по двойки някои от следните променливи:

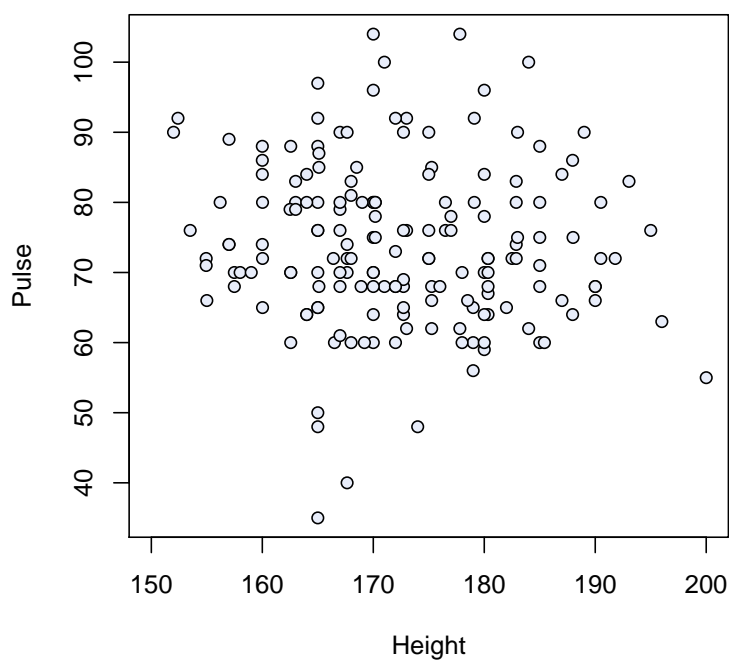
Height – височина на студента (в сантиметри);

Pulse – пулс на студента;

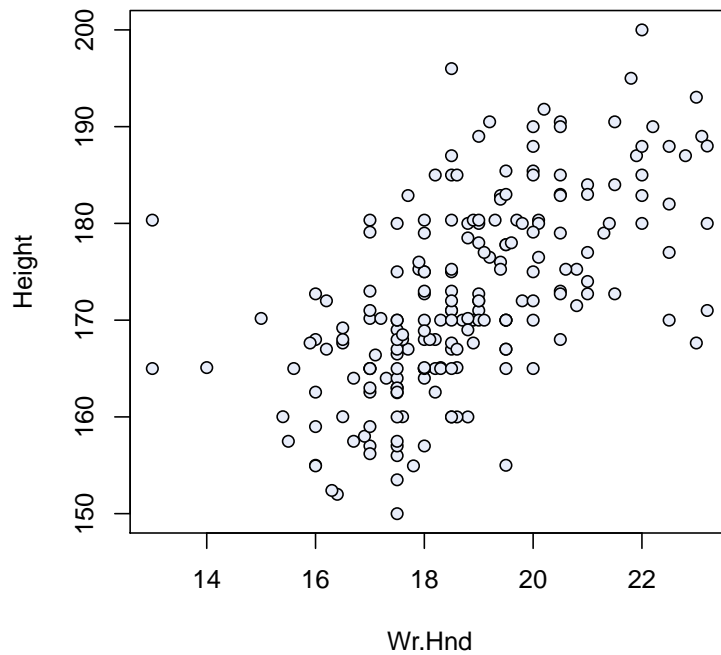
Wr.Hnd – дължина на педята на ръката, с която студентът пише (в сантиметри);

NW.Hnd – дължина на педята на ръката, с която студентът не пише (в сантиметри);

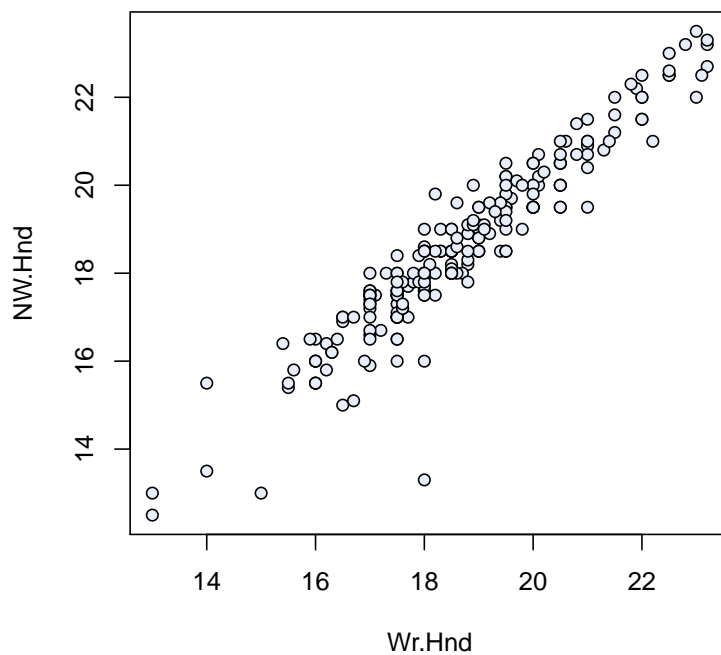
```
> plot(Height, Pulse)
> plot(Wr.Hnd, Height)
> plot(Wr.Hnd, NW.Hnd)
```



Фигура 6.5.



Фигура 6.6.



Фигура 6.7.

Нека x_1, \dots, x_n са наблюдения над някаква числова променлива X , а y_1, \dots, y_n са наблюдения над друга числова променлива Y . Числото

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

се нарича *извадъчна корелация* на променливите X и Y . То е мярка за линейната зависимост между тях. Пресмята се с `cor(x,y)`.

Ако точките $(x_1, y_1), \dots, (x_n, y_n)$ са близо да права линия, то r е близо до 1 или -1 и казваме, че променливите са положително или отрицателно корелирани, съответно. Ако r е близо до 0, казваме, че променливите са некорелирани.

⟨!⟩ Ако точките $(x_1, y_1), \dots, (x_n, y_n)$ са близо до прави от вида $x = \text{const}$ или $y = \text{const}$, извадъчната корелация r ще е близка до 0. (Защо?)

Ще намерим корелациите на двойките променливи, които изобразихме на графики по-горе. Тъй като има липсващи наблюдения, за да се игнорират при пресмятането, добавяме `use="complete.obs"`.

```
> cor(Height, Pulse, use="complete.obs")
[1] -0.0839
> cor(Wr.Hnd, Height, use="complete.obs")
[1] 0.601
> cor(Wr.Hnd, NW.Hnd, use="complete.obs")
[1] 0.948
```

Корелацията между `Height` и `Pulse` е близка до 0. Тази между `Wr.Hnd` и `Height` е по-близко до 1, а между `Wr.Hnd` и `NW.Hnd` е почти 1.

7. Доверителни интервали

*An approximate answer to the right question is far better
than an exact answer to the wrong question.*

John Tukey

7.1. Увод

Наблюденията x_1, x_2, \dots, x_n над променливата X може да разглеждаме като n независими наблюдения над случайна величина X с някакво разпределение. Оказва се удобно да разглеждаме x_1, x_2, \dots, x_n като наблюдения над случайни величини X_1, X_2, \dots, X_n , които са независими и имат същото разпределение като X , т.е. x_1 е наблюдение над сл.в. X_1 , x_2 е наблюдение над сл.в. X_2 и т.н. Случайните величини X_1, X_2, \dots, X_n ще наричаме *извадка* (понякога самите наблюдения x_1, x_2, \dots, x_n се наричат извадка).

Така, средното на данните (извадъчно средно) $\bar{x} = (x_1 + \dots + x_n)/n$ е всъщност наблюдавана стойност над съответната случайна величина $\bar{X} = (X_1 + \dots + X_n)/n$, а стандартното

отклонение на данните (извадъчно стандартно отклонение) $s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$ е наблюдавана стойност над случайната величина $S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$.

Нека случайната величина X има средно $E(X) = \mu$ и дисперсия $\text{Var}(X) = \sigma^2$ (понякога се наричат *популационно средно* и *популационна дисперсия*). Съгласно централната гранична теорема за големи стойности на n (обикновено се приема $n \geq 30$) случайната величина $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ е приблизително нормално разпределена:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

За практически цели може да използваме, че при големи стойности на n случайната величина \bar{X} е приблизително нормално разпределена със средно μ и дисперсия σ^2/n , т.е. $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$.

Медиана на случайната величина X наричаме 0.5-квантилът $Q(0.5)$, където $Q(p)$ е квантилната функция на X .

7.2. Доверителен интервал за средно при известна дисперсия

Нека x_1, x_2, \dots, x_n са наблюдения над случайните величини X_1, X_2, \dots, X_n , които са независими и еднакво разпределени. Средното $E(X_i) = \mu$ е неизвестно, но знаем дисперсията $\text{Var}(X_i) = \sigma^2$.

Означаваме $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$. Нека \bar{x} е наблюдаваната стойност на \bar{X} .

Когато X_1, X_2, \dots, X_n са нормално разпределени, случайната величина

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

има стандартно нормално разпределение. Ако X_1, X_2, \dots, X_n имат произволно разпределение, за големи стойности на n случайната величина Z има приблизително стандартно нормално разпределение.

$$\mathbf{P}(-1.96 \leq Z \leq 1.96) = 0.95$$

$$\mathbf{P}(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96) = 0.95$$

$$\mathbf{P}(-1.96\sigma/\sqrt{n} \leq \bar{X} - \mu \leq 1.96\sigma/\sqrt{n}) = 0.95$$

$$\mathbf{P}(\bar{X} - 1.96\sigma/\sqrt{n} \leq \mu \leq \bar{X} + 1.96\sigma/\sqrt{n}) = 0.95$$

Интервалът

$$\left[\bar{x} - 1.96\sigma/\sqrt{n}, \quad \bar{x} + 1.96\sigma/\sqrt{n} \right]$$

е 95-процентен доверителен интервал за средното μ .

Нека z_α е такова, че $\mathbf{P}(Z > z_\alpha) = \alpha$.

Интервалът

$$\left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \quad \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

е $100(1 - \alpha)$ -процентен доверителен интервал за средното μ .

Задача 7.1. Фирма произвежда електрически крушки. Средното време на живот на една крушка е 2000 часа със стандартно отклонение 300 часа. Предложен е нов тип крушки. Изпробовани са 100 крушки от новия тип. Резултатите показват средно време на живот на новите крушки 2100 часа и същото стандартно отклонение.

а) Намерете 95-процентен доверителен интервал за средното време на живот на новия тип крушки.

б) Ако са изпробовани 200 крушки и останалите данни са същите, намерете 95-процентен доверителен интервал за средното време на живот на новия тип крушки.

```
z1.ci <- function(x.bar, sigma, n, alpha) {
  b1 <- x.bar - qnorm(1-alpha/2)*(sigma/sqrt(n))
  b2 <- x.bar + qnorm(1-alpha/2)*(sigma/sqrt(n))
  c(b1, b2)
}
```

```
> z1.ci(x.bar=2100, sigma=300, n=100, alpha=0.05)
[1] 2041.201 2158.799
> z1.ci(x.bar=2100, sigma=300, n=200, alpha=0.05)
[1] 2058.423 2141.577
```

7.3. Доверителен интервал за средно при неизвестна дисперсия

Нека x_1, x_2, \dots, x_n са наблюдения над случайните величини X_1, X_2, \dots, X_n , които са независими и еднакво разпределени. Средното $E(X_i) = \mu$ и дисперсията $\text{Var}(X_i) = \sigma^2$ са неизвестни.

Да означим

$$\bar{X} = \frac{1}{n}(X_1 + \dots + X_n), \quad S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}.$$

Нека \bar{x} е наблюдаваната стойност на \bar{X} , а s е наблюдаваната стойност на S .

Когато X_1, X_2, \dots, X_n са нормално разпределени, случайната величина

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

има t -разпределение с $n - 1$ степени на свобода (разпределение на Стюдънт). Ако X_1, X_2, \dots, X_n имат произволно разпределение, за големи стойности на n случайната величина T има приблизително t -разпределение с $n - 1$ степени на свобода.

Нека $t_{n-1,\alpha}$ е такава, че

$$\mathbf{P}(T > t_{n-1,\alpha}) = \alpha.$$

Интервалът

$$\left[\bar{x} - t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}, \quad \bar{x} + t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} \right]$$

е $100(1 - \alpha)$ -процентен доверителен интервал за средното μ .

Задача 7.2. В индустриален район е измерена киселинността на 12 валежа през изминалата година. Получени са следните резултати:

6.1 5.4 4.8 5.8 6.6 5.3 6.1 4.4 3.9 6.8 6.5 6.3

От предишни изследвания е известно, че киселинността на валежите има нормално разпределение.

а) Намерете 95-процентен доверителен интервал за средната киселинност.

б) Намерете 99-процентен доверителен интервал за средната киселинност.

```
t1.ci <- function(x.bar, s, n, alpha) {
  b1 <- x.bar - qt(1-alpha/2, df=n-1)*(s/sqrt(n))
  b2 <- x.bar + qt(1-alpha/2, df=n-1)*(s/sqrt(n))
  c(b1, b2)
}

> x <- c(6.1, 5.4, 4.8, 5.8, 6.6, 5.3, 6.1, 4.4, 3.9, 6.8, 6.5, 6.3)

> t1.ci(x.bar=mean(x), s=sd(x), n=length(x), alpha=0.05)
[1] 5.081616 6.251717
> t1.ci(x.bar=mean(x), s=sd(x), n=length(x), alpha=0.01)
[1] 4.841103 6.492230

> t.test(x, conf.level=0.95)$conf.int[1:2]
[1] 5.081616 6.251717
> t.test(x, conf.level=0.99)$conf.int[1:2]
[1] 4.841103 6.492230
```

7.4. Доверителен интервал за пропорция (вероятност за успех)

Нека x_1, x_2, \dots, x_n са наблюдения над случайните величини X_1, X_2, \dots, X_n , които са независими и еднакво разпределени, като X_i приема стойност 1 с вероятност p или 0 с вероятност $1 - p$ (Бернулиево разпределение). Такива данни могат да се получат, например, при въпрос с два възможни отговора в анкетно проучване. По-общо, такива данни може да разгледаме като резултат от n повторения на опит с два изхода: „успех“ и „неуспех“. Основния параметър, който ни интересува е вероятността p (нарича се още пропорция), например, вероятността студент да е пушач, вероятността да се появят странични ефекти при употреба на дадено лекарство, вероятността да се появи дефект в батериите произведени от даден завод.

Нека $X = X_1 + \dots + X_n$ и $x = x_1 + \dots + x_n$. За оценка на вероятността p използваме $\hat{p} = x/n$, т.е. наблюдавания брой успехи разделен на броя опити (всъщност \hat{p} е средното на данните x_1, x_2, \dots, x_n). Случайната величина X е биомно разпределена, $X \sim \text{Bi}(n, p)$.

Разглеждаме случайната величина

$$Z = \frac{X/n - p}{\sqrt{p(1-p)/n}}.$$

Като вземем предвид, че $E(X/n) = p$ и $\text{Var}(X/n) = p(1-p)/n$, от централната гранична теорема, за големи стойности на n случайната величина Z има приблизително стандартно нормално разпределение.

$$\mathbf{P}(-1.96 \leq Z \leq 1.96) = 0.95$$

$$\mathbf{P}\left(-1.96 \leq \frac{X/n - p}{\sqrt{p(1-p)/n}} \leq 1.96\right) = 0.95$$

...

$$\mathbf{P}\left(X/n - 1.96\sqrt{p(1-p)/n} \leq p \leq X/n + 1.96\sqrt{p(1-p)/n}\right) = 0.95$$

Интервалът

$$\left[\hat{p} - 1.96\sqrt{\hat{p}(1-\hat{p})/n}, \quad \hat{p} + 1.96\sqrt{\hat{p}(1-\hat{p})/n}\right]$$

е 95-процентен доверителен интервал за пропорцията p .

Нека z_α е такова, че $\mathbf{P}(Z > z_\alpha) = \alpha$.

Интервалът

$$\left[\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}, \quad \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right]$$

е $100(1 - \alpha)$ -процентен доверителен интервал за пропорцията p .

Задача 7.3. От първите 25000 продадени коли от нов модел, 2700 се оказали с дефект. Намерете 95-процентен доверителен интервал за вероятността кола от новия модел да е дефектна.

```
prop1.ci <- function(x, n, alpha) {  
  p.hat <- x/n  
  b1 <- p.hat - qnorm(1-alpha/2)*sqrt(p.hat*(1-p.hat)/n)  
  b2 <- p.hat + qnorm(1-alpha/2)*sqrt(p.hat*(1-p.hat)/n)  
  c(b1, b2)  
}
```



```
> prop1.ci(x=2700, n=25000, alpha=0.05)
[1] 0.1041526 0.1118474

> prop.test(x=2700, n=25000, conf.level=0.95, correct=F)$conf.int[1:2]
[1] 0.1042126 0.1119078
```

7.5. Доверителен интервал за медиана

Нека x_1, x_2, \dots, x_n са наблюдения над случайните величини X_1, X_2, \dots, X_n , които са независими и еднакво разпределени. Нека X е сл.в. със същото разпределение като X_i и M е медианата на X .

За $b = 1, \dots, B$, нека $x_{1,b}^*, x_{2,b}^*, \dots, x_{n,b}^*$ е извадка с връщане от $\{x_1, x_2, \dots, x_n\}$ и \widehat{M}_b^* е медианата на $x_{1,b}^*, x_{2,b}^*, \dots, x_{n,b}^*$.

$$\begin{array}{cccccc} x_{1,1}^* & x_{2,1}^* & \dots & x_{n,1}^* & \widehat{M}_1^* \\ x_{1,2}^* & x_{2,2}^* & \dots & x_{n,2}^* & \widehat{M}_2^* \\ \vdots & \vdots & & \vdots & \vdots \\ x_{1,B}^* & x_{2,B}^* & \dots & x_{n,B}^* & \widehat{M}_B^* \end{array}$$

Нека $Q_M^{\text{boot}}(\alpha/2)$ е $\alpha/2$ -квантил на $\widehat{M}_1^*, \dots, \widehat{M}_B^*$, а $Q_M^{\text{boot}}(1 - \alpha/2)$ е $(1 - \alpha/2)$ -квантил на $\widehat{M}_1^*, \dots, \widehat{M}_B^*$.

Интервалът

$$[Q_M^{\text{boot}}(\alpha/2), Q_M^{\text{boot}}(1 - \alpha/2)]$$

е $100(1 - \alpha)$ -процентен доверителен интервал за медианата M . Доверителен интервал, получен по този начин, се нарича бутстрап-процентилен доверителен интервал (*bootstrap percentile confidence interval*).

Идеята на метода е да се генерират голям брой извадки с връщане от наблюденията x_1, x_2, \dots, x_n и да се пресметне медианата на всяка извадка. Желателно е броят B на генерираните извадки да е поне 1000.

Ако във вектора **x** са записани наблюденията x_1, x_2, \dots, x_n , извадка с връщане генерираме с командата **sample(x, size=n, replace=TRUE)**.

Със следната команда генерираме B извадки с връщане от $\{x_1, x_2, \dots, x_n\}$, за всяка извадка пресмятаме медианата и получаваме вектор с медианите $\widehat{M}_1^*, \dots, \widehat{M}_B^*$:

```
replicate( B, median( sample( x, size=n, replace=TRUE ) ) )
```

Задача 7.4. Направено е проучване с цел да се изследва ефектът на звука от сърдечния ритъм на майката върху новороденото. Бебетата в родилно отделение са разделени на две групи. Първата група е непрекъснато изложена на звука от сърдечния ритъм на възрастен, а втората група не е изложена на такъв звук. Измерена е промяната в теглото на бебетата от раждането до четвъртия ден. Данните са във файла **salk.txt**. Намерете 95-процентен доверителен интервал за медианата на промяната в теглото на бебетата (за първата и за втората група).

```

med1.ci <- function(x, alpha=0.05, nboot=1000) {
  x <- x[is.finite(x)]
  nx <- length(x)
  est1 <- median(x)
  med1.bt <- replicate( nboot, median( sample( x, size=nx, replace=TRUE ) ) )
  ci <- quantile( med1.bt, probs=c(alpha/2, 1-alpha/2), names=FALSE )
  list( est.med1=est1, ci=ci )
}

> salk <- read.table("salk.txt")
> summary(salk)

> med1.ci( salk[,1], alpha=0.05, nboot=10000 )
$est.med1
[1] 10
$ci
[1] -5 35

> med1.ci( salk[,2], alpha=0.05, nboot=10000 )
$est.med1
[1] -45
$ci
[1] -75 -25

```

7.6. Доверителен интервал за разлика на медиани

Нека x_1, x_2, \dots, x_n са наблюдения над случайните величини X_1, X_2, \dots, X_n , които са независими и еднакво разпределени. Нека X е сл.в. със същото разпределение като X_i и M_X е медианата на X .

Нека y_1, y_2, \dots, y_m са наблюдения над случайните величини Y_1, Y_2, \dots, Y_m , които са независими и еднакво разпределени. Нека Y е сл.в. със същото разпределение като Y_i и M_Y е медианата на Y .

За $b = 1, \dots, B$, нека $x_{1,b}^*, x_{2,b}^*, \dots, x_{n,b}^*$ е извадка с връщане от $\{x_1, x_2, \dots, x_n\}$ и $\widehat{M}_{X,b}^*$ е медианата на $x_{1,b}^*, x_{2,b}^*, \dots, x_{n,b}^*$. За $b = 1, \dots, B$, нека $y_{1,b}^*, y_{2,b}^*, \dots, y_{m,b}^*$ е извадка с връщане от $\{y_1, y_2, \dots, y_m\}$ и $\widehat{M}_{Y,b}^*$ е медианата на $y_{1,b}^*, y_{2,b}^*, \dots, y_{m,b}^*$.

$$\begin{array}{ccccccc}
 x_{1,1}^* & x_{2,1}^* & \dots & x_{n,1}^* & \widehat{M}_{X,1}^* \\
 x_{1,2}^* & x_{2,2}^* & \dots & x_{n,2}^* & \widehat{M}_{X,2}^* \\
 \vdots & \vdots & & \vdots & \vdots \\
 x_{1,B}^* & x_{2,B}^* & \dots & x_{n,B}^* & \widehat{M}_{X,B}^*
 \end{array}$$

$$\begin{array}{cccccc}
y_{1,1}^* & y_{2,1}^* & \cdots & y_{m,1}^* & \widehat{M}_{Y,1}^* \\
y_{1,2}^* & y_{2,2}^* & \cdots & y_{m,2}^* & \widehat{M}_{Y,2}^* \\
\vdots & \vdots & & \vdots & \vdots \\
y_{1,B}^* & y_{2,B}^* & \cdots & y_{m,B}^* & \widehat{M}_{Y,B}^*
\end{array}$$

За $b = 1, \dots, B$ дефинираме $\widehat{D}_b^* = \widehat{M}_{X,b}^* - \widehat{M}_{Y,b}^*$.

Нека $Q_D^{\text{boot}}(\alpha/2)$ е $\alpha/2$ -квантил на $\widehat{D}_1^*, \dots, \widehat{D}_B^*$, а $Q_D^{\text{boot}}(1 - \alpha/2)$ е $(1 - \alpha/2)$ -квантил на $\widehat{D}_1^*, \dots, \widehat{D}_B^*$.

Интервалът

$$[Q_D^{\text{boot}}(\alpha/2), Q_D^{\text{boot}}(1 - \alpha/2)]$$

е $100(1 - \alpha)$ -процентен доверителен интервал за разликата на медианите $M_X - M_Y$. Нарича се бутстрап-процентилен доверителен интервал (*bootstrap percentile confidence interval*).

Задача 7.5. С данните от задача 7.4 намерете 95-процентен доверителен интервал за разликата на медианите на първата и втората група.

```

med2.ci <- function(x, y, alpha=0.05, nboot=1000) {
  x <- x[is.finite(x)]
  y <- y[is.finite(y)]
  nx <- length(x)
  ny <- length(y)
  est1 <- median(x)
  est2 <- median(y)
  est.dif <- est1 - est2
  med1.bt <- replicate( nboot, median( sample( x, size=nx, replace=TRUE ) ) )
  med2.bt <- replicate( nboot, median( sample( y, size=ny, replace=TRUE ) ) )
  dif.bt <- med1.bt - med2.bt
  ci <- quantile( dif.bt, probs=c(alpha/2, 1-alpha/2), names=FALSE )
  list( est.med1=est1, est.med2=est2, est.dif=est.dif, ci=ci )
}

```

```

> med2.ci( salk[,1], salk[,2], alpha=0.05, nboot=10000 )
$est.med1
[1] 10
$est.med2
[1] -45
$est.dif
[1] 55
$ci
[1] 30 90

```

7.7. Интерпретация на доверителни интервали

Нека сме намерили 95-процентен доверителен интервал $[b_1, b_2]$ за средното μ на сл.в. X . Твърдението

$$P(\mu \in [b_1, b_2]) = 0.95$$

е *погрешно*, тъй като в израза няма случайно събитие.

Вярната интерпретация е следната: ако имаме голям брой извадки от дадено разпределение и за всяка от тях намерим доверителен интервал за параметъра μ , около 95% от тези интервали ще съдържат μ . Аналогична е интерпретацията на доверителен интервал за произволен параметър θ (в частност, за пропорция p и медиана M).

8. Проверка на хипотези при една извадка

*If the result confirms the hypothesis,
then you've made a measurement.
If the result is contrary to the hypothesis,
then you've made a discovery.*

E. Fermi

8.1. z -тест за средно

Нека x_1, x_2, \dots, x_n са наблюдения над случайните величини X_1, X_2, \dots, X_n , които са независими и еднакво разпределени. Средното $E(X_i) = \mu$ е неизвестно, но знаем дисперсията $\text{Var}(X_i) = \sigma^2$.

Означаваме $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$. Нека \bar{x} е наблюдаваната стойност на \bar{X} .

Искаме въз основа на данните да отговорим дали имаме основание да твърдим, че μ е равно на някаква предварително зададена стойност μ_0 .

Формално записваме това по следния начин:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

Казваме, че проверяваме хипотезата H_0 срещу алтернативата H_1 . Хипотезата H_0 наричаме *нулева хипотеза*, а H_1 наричаме *алтернативна хипотеза*. Най-общо, отхвърляме H_0 , ако това, което сме наблюдавали е малко вероятно да се случи, когато H_0 е вярна. За да намерим съответната вероятност, разглеждаме случайната величина

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Когато X_1, X_2, \dots, X_n са нормално разпределени, случайната величина Z има стандартно нормално разпределение. Ако X_1, X_2, \dots, X_n имат произволно разпределение, за големи стойности на n случайната величина Z има приблизително стандартно нормално разпределение.

Нека z_{obs} е наблюдаваната стойност на Z , като сме заместили $\mu = \mu_0$:

$$z_{\text{obs}} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Отхвърляме нулевата хипотеза, ако $\mathbf{P}_0(Z \geq |z_{\text{obs}}| \cup Z \leq -|z_{\text{obs}}|) \leq \alpha$. Означаваме с \mathbf{P}_0 вероятността при условие нулевата хипотеза, т.е. при $\mu = \mu_0$.

Вероятността $\mathbf{P}_0(Z \geq |z_{\text{obs}}| \cup Z \leq -|z_{\text{obs}}|)$ наричаме P -стойност (P -value).

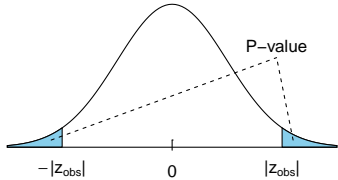
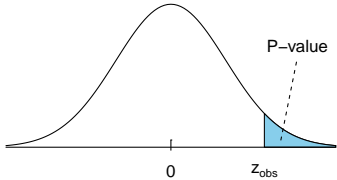
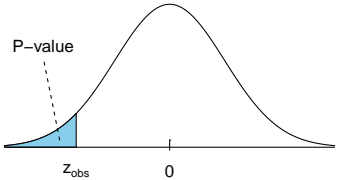
Числото α се нарича *ниво на съгласие* или *ниво на значимост* или *вероятност за грешка от I род*. Нивото на съгласие обикновено се определя предварително, като най-често се взема 0.05 или 0.1 или 0.01.

Описаната процедура наричаме z -тест за средно.

Най-общо, z_{obs} измерва отклонението на данните от нулевата хипотеза. Ако това отклонение е твърде голямо, нулевата хипотеза се отхвърля. По-точно, нулевата хипотеза се отхвърля ако е малко вероятно да има такова отклонение при вярна нулева хипотеза. P -стойността е именно вероятността да се наблюдава такова отклонение (или по-голямо по модул) при вярна нулева хипотеза.

Алтернативната хипотеза може да е от вида $H_1 : \mu > \mu_0$ или $H_1 : \mu < \mu_0$, тогава наричаме теста едностранен, а когато $H_1 : \mu \neq \mu_0$ – двустранен. Как се пресмята P -стойността при различните алтернативни хипотези е дадено в таблицата по-долу.

P-стойност при различни алтернативни хипотези

$H_1 : \mu \neq \mu_0$	$H_1 : \mu > \mu_0$	$H_1 : \mu < \mu_0$
$\mathbf{P}_0(Z \geq z_{\text{obs}} \cup Z \leq - z_{\text{obs}})$ $= 2 \mathbf{P}_0(Z \geq z_{\text{obs}}) = \text{P-value}$	$\mathbf{P}_0(Z \geq z_{\text{obs}}) = \text{P-value}$	$\mathbf{P}_0(Z \leq z_{\text{obs}}) = \text{P-value}$
		

z-тест за средно

Статистика:

Пресмятаме наблюдаваната стойност $z_{\text{obs}} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$.

P-стойност:

H_0	H_1	P-value
$\mu = \mu_0$	$\mu \neq \mu_0$	$2 \mathbf{P}_0(Z \geq z_{\text{obs}}) = 2*(1-\text{pnorm}(\text{abs}(z.\text{obs})))$
$\mu = \mu_0$	$\mu > \mu_0$	$\mathbf{P}_0(Z \geq z_{\text{obs}}) = 1-\text{pnorm}(z.\text{obs})$
$\mu = \mu_0$	$\mu < \mu_0$	$\mathbf{P}_0(Z \leq z_{\text{obs}}) = \text{pnorm}(z.\text{obs})$

Извод:

Ако $\text{P-value} \leq \alpha$, отхвърляме H_0 в полза на H_1 .

Ако $\text{P-value} > \alpha$, нямаме достатъчно основания да отхвърлим H_0 .

Задача 8.1. Фирма произвежда електрически крушки. Средното време на живот на една крушка е 2000 часа със стандартно отклонение 300 часа. Предложен е нов тип крушки. Изпробвани са 100 крушки от новия тип. Резултатите показват средно време на живот на новите крушки 2100 часа и същото стандартно отклонение. Може ли да се твърди, че средното време на живот на новия тип крушки е повече от 2000 часа?

Решение. Нека μ е средното време на живот на новия тип крушки.

Проверяваме хипотезата

$$H_0 : \mu = 2000$$

срещу алтернативата

$$H_1 : \mu > 2000$$

По условие имаме:

$$\bar{x} = 2100$$

$$n = 100$$

$$\sigma = 300$$

```
> x.bar <- 2100
> n <- 100
> sigma <- 300
> mu <- 2000
> z.obs <- (x.bar - mu) / (sigma/sqrt(n))
> z.obs
[1] 3.333333
> p.value <- 1-pnorm(z.obs)
> p.value
[1] 0.0004290603
```

P-стойността е по-малка от 0.05, следователно отхвърляме нулевата хипотеза в полза на алтернативната. *Имаме основание да твърдим, че средното време на живот на новия тип крушки е повече от 2000 часа.*



8.2. t -тест за средно

Нека x_1, x_2, \dots, x_n са наблюдения над случайните величини X_1, X_2, \dots, X_n , които са независими и еднакво разпределени. Средното $E(X_i) = \mu$ и дисперсията $\text{Var}(X_i) = \sigma^2$ са неизвестни.

Означаваме

$$\bar{X} = \frac{1}{n}(X_1 + \dots + X_n), \quad S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}.$$

Нека \bar{x} е наблюдаваната стойност на \bar{X} , а s е наблюдаваната стойност на S .

Искаме въз основа на данните да проверим хипотезата

$$H_0 : \mu = \mu_0$$

срещу алтернативата

$$H_1 : \mu \neq \mu_0$$

Когато X_1, X_2, \dots, X_n са нормално разпределени, случайната величина

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

има t -разпределение с $n - 1$ степени на свобода (разпределение на Стюдънт). Ако X_1, X_2, \dots, X_n имат произволно разпределение, за големи стойности на n случайната величина T има приблизително t -разпределение с $n - 1$ степени на свобода.

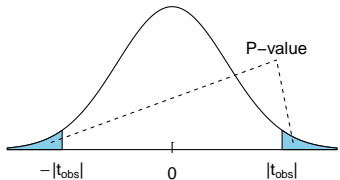
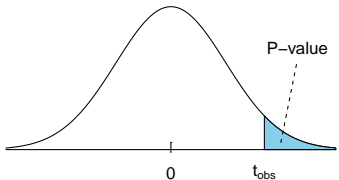
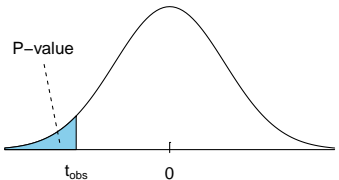
Нека t_{obs} е наблюдаваната стойност на T , като сме заместили $\mu = \mu_0$:

$$t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$

Отхвърляме нулевата хипотеза, ако $\text{P-value} = \mathbf{P}_0(T \geq |t_{\text{obs}}| \cup T \leq -|t_{\text{obs}}|) \leq \alpha$.

Как се пресмята Р-стойността при различните алтернативни хипотези е дадено в таблицата по-долу.

Р-стойност при различни алтернативни хипотези

$H_1 : \mu \neq \mu_0$	$H_1 : \mu > \mu_0$	$H_1 : \mu < \mu_0$
$\mathbf{P}_0(T \geq t_{\text{obs}} \cup T \leq - t_{\text{obs}})$ $= 2 \mathbf{P}_0(T \geq t_{\text{obs}}) = \text{P-value}$	$\mathbf{P}_0(T \geq t_{\text{obs}}) = \text{P-value}$	$\mathbf{P}_0(T \leq t_{\text{obs}}) = \text{P-value}$
		

t-тест за средно

Статистика:

Пресмятаме наблюдаваната стойност $t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$.

P-стойност:

H_0	H_1	P-value
$\mu = \mu_0$	$\mu \neq \mu_0$	$2\mathbf{P}_0(T \geq t_{\text{obs}}) = 2*(1-\text{pt}(\text{abs}(\mathbf{t. obs}), \mathbf{n-1}))$
$\mu = \mu_0$	$\mu > \mu_0$	$\mathbf{P}_0(T \geq t_{\text{obs}}) = 1-\text{pt}(\mathbf{t. obs}, \mathbf{n-1})$
$\mu = \mu_0$	$\mu < \mu_0$	$\mathbf{P}_0(T \leq t_{\text{obs}}) = \text{pt}(\mathbf{t. obs}, \mathbf{n-1})$

Извод:

Ако $\text{P-value} \leq \alpha$, отхвърляме H_0 в полза на H_1 .

Ако $\text{P-value} > \alpha$, нямаме достатъчно основания да отхвърлим H_0 .

Съответната функция в R е `t.test`. Използва се по следния начин (във вектора `x` са записани наблюденията x_1, x_2, \dots, x_n):

```
 $H_1 : \mu \neq \mu_0$     t.test(x, mu =  $\mu_0$ )  
 $H_1 : \mu > \mu_0$     t.test(x, mu =  $\mu_0$ , alternative = 'greater')  
 $H_1 : \mu < \mu_0$     t.test(x, mu =  $\mu_0$ , alternative = 'less')
```

Задача 8.2. Според исторически данни, средната киселинност на дъждовете в определен индустриален район е 5.2. За да се провери има ли изменение в тази стойност е измерена киселинността на 12 валежа през изминалата година. Получени са следните резултати:

6.1 5.4 4.8 5.8 6.6 5.3 6.1 4.4 3.9 6.8 6.5 6.3

От предишни изследвания е известно, че киселинността на валежите има нормално разпределение. Имаме ли основания да твърдим, че киселинността в района се е променила в сравнение с историческите данни.

Решение. Нека μ е средната киселинност на дъждовете в района за изминалата година.

Проверяваме хипотезата

$$H_0 : \mu = 5.2$$

срещу алтернативата

$$H_1 : \mu \neq 5.2$$

```
> x <- c(6.1, 5.4, 4.8, 5.8, 6.6, 5.3, 6.1, 4.4, 3.9, 6.8, 6.5, 6.3)
> n <- length(x)
> mu <- 5.2
> t.obs <- (mean(x) - mu) / (sd(x)/sqrt(n))
> t.obs
[1] 1.75562
> p.value <- 2*(1-pt(abs(t.obs), n-1))
> p.value
```

```
[1] 0.1069226
```

```
> t.test(x, mu=5.2)
```

One Sample t-test

```
data: x
```

```
t = 1.7556, df = 11, p-value = 0.1069
```

```
alternative hypothesis: true mean is not equal to 5.2
```

```
95 percent confidence interval:
```

```
5.081616 6.251717
```

```
sample estimates:
```

```
mean of x
```

```
5.666667
```

```
> t.test(x, mu=5.2)$p.value
```

```
[1] 0.1069226
```

P-стойността е по-голяма от 0.05, следователно нямаме достатъчно основания да отхвърлим нулевата хипотеза. *Нямаме достатъчно основания да твърдим, че средната киселинност на дъждовете в района се е променила в сравнение с историческите данни.*



8.3. z -тест за пропорция

Нека x_1, x_2, \dots, x_n са наблюдения над случайните величини X_1, X_2, \dots, X_n , които са независими и еднакво разпределени, като X_i приема стойност 1 с вероятност p или 0 с вероятност $1 - p$ (Бернулиево разпределение). Такива данни могат да се получат, например, при въпрос с два възможни отговора в анкетно проучване. По-общо, такива данни може да разгледаме като резултат от n повторения на опит с два изхода: „успех“ и „неуспех“. Основния параметър, който ни интересува е вероятността p (нарича се още пропорция), например, вероятността студент да е пушач, вероятността да се появят странични ефекти при употреба на дадено лекарство, вероятността да се появи дефект в батериите произведени от даден завод.

Нека $X = X_1 + \dots + X_n$ и $x = x_1 + \dots + x_n$ (x е наблюдавания брой успехи). Случайната величина X е биомно разпределена, $X \sim \text{Bi}(n, p)$.

Искаме въз основа на данните да проверим хипотезата

$$H_0 : p = p_0$$

срещу алтернативата

$$H_1 : p \neq p_0$$

Разглеждаме случайната величина

$$Z = \frac{X/n - p}{\sqrt{p(1-p)/n}}.$$

Като вземем предвид, че $E(X/n) = p$ и $\text{Var}(X/n) = p(1-p)/n$, от централната гранична теорема, за големи стойности на n случайната величина Z има приблизително стнадартно нормално разпределение.

Нека z_{obs} е наблюдаваната стойност на Z , като сме заместили $p = p_0$, т.е.

$$z_{\text{obs}} = \frac{x/n - p_0}{\sqrt{p_0(1-p_0)/n}}.$$

Отхвърляме нулевата хипотеза, ако $\text{P-value} = \mathbf{P}_0(Z \geq |z_{\text{obs}}| \cup Z \leq -|z_{\text{obs}}|) \leq \alpha$.

Как се пресмята Р-стойността при различните алтернативни хипотези е дадено в таблицата по-долу.

Р-стойност при различни алтернативни хипотези

$H_1 : p \neq p_0$	$H_1 : p > p_0$	$H_1 : p < p_0$
$\mathbf{P}_0(Z \geq z_{\text{obs}} \cup Z \leq - z_{\text{obs}})$ $= 2 \mathbf{P}_0(Z \geq z_{\text{obs}}) = \text{P-value}$	$\mathbf{P}_0(Z \geq z_{\text{obs}}) = \text{P-value}$	$\mathbf{P}_0(Z \leq z_{\text{obs}}) = \text{P-value}$

z-тест за пропорция

Статистика:

Пресмятаме наблюдаваната стойност $z_{\text{obs}} = \frac{x/n - p_0}{\sqrt{p_0(1-p_0)/n}}$.

Р-стойност:

H_0	H_1	P-value
$p = p_0$	$p \neq p_0$	$2\mathbf{P}_0(Z \geq z_{\text{obs}}) = 2*(1-\text{pnorm}(\text{abs}(z.\text{obs})))$
$p = p_0$	$p > p_0$	$\mathbf{P}_0(Z \geq z_{\text{obs}}) = 1-\text{pnorm}(z.\text{obs})$
$p = p_0$	$p < p_0$	$\mathbf{P}_0(Z \leq z_{\text{obs}}) = \text{pnorm}(z.\text{obs})$

Извод:

Ако $\text{P-value} \leq \alpha$, отхвърляме H_0 в полза на H_1 .

Ако $\text{P-value} > \alpha$, нямаме достатъчно основания да отхвърлим H_0 .

За проверка на хипотези за пропорция може да се използва функцията `prop.test`, която прави подобен тест:

$H_1 : p \neq p_0$ `prop.test(x=x, n=n, p=p0, correct=F)`

$H_1 : p > p_0$ `prop.test(x=x, n=n, p=p0, alternative='greater', correct=F)`

$H_1 : p < p_0$ `prop.test(x=x, n=n, p=p0, alternative='less', correct=F)`

Задача 8.3. Известно е, че при 10% от колите от дадена марка се появява сериозен дефект по време на гаранционния срок. От първите 25000 продадени коли от нов модел, 2700 се оказали с дефект. Може ли да се твърди, че вероятността в кола от новия модел да се появи дефект не е 10%?

Решение. Означаваме с p вероятността кола от новия модел да е дефектна.

Проверяваме хипотезата

$$H_0 : p = 0.1$$

срещу алтернативата

$$H_1 : p \neq 0.1$$

По условие имаме извадка от 25000 коли, от които дефектни са 2700, т.е.

$$x = 2700$$

$$n = 25000$$

```
> x <- 2700
> n <- 25000
> p <- 0.1
> z.obs <- ((x/n) - p) / sqrt(p*(1-p)*(1/n))
> z.obs
[1] 4.21637
> p.value <- 2*(1-pnorm(abs(z.obs)))
> p.value
[1] 2.482661e-05
```

За Р-стойността имаме

$$P\text{-value} = 2.482661 * 10^{-5} < 0.05,$$

следователно отхвърляме нулевата хипотеза. *Имаме основание да твърдим, че вероятността в кола от новия модел да се появи дефект не е 10%.*

Същата хипотеза може да проверим с помощта на функцията `prop.test`:

```
> prop.test(x=2700, n=25000, p=0.1, correct=F)
```

```
1-sample proportions test without continuity correction
```

```
data: 2700 out of 25000, null probability 0.1
```

```
X-squared = 17.778, df = 1, p-value = 2.483e-05
```

```
alternative hypothesis: true p is not equal to 0.1
```

```
95 percent confidence interval:
```

```
0.1042126 0.1119078
```

```
sample estimates:
```

```
p
```

```
0.108
```

```
> prop.test(x=2700, n=25000, p=0.1, correct=F)$p.value
```

```
[1] 2.482661e-05
```



9. Проверка на хипотези при две извадки

*We all learn by experience, and the lesson this time is
that you should never lose sight of the alternative.*

Sherlock Holmes

(The Adventures of Black Peter by A.C. Doyle)

9.1. t -тест за разлика на средни

Нека x_1, x_2, \dots, x_n са наблюдения над случайните величини X_1, X_2, \dots, X_n , които са независими и еднакво разпределени. Нека X е сл.в. със същото разпределение като X_i и $E(X) = \mu_X$.

Нека y_1, y_2, \dots, y_m са наблюдения над случайните величини Y_1, Y_2, \dots, Y_m , които са независими и еднакво разпределени. Нека Y е сл.в. със същото разпределение като Y_i и $E(Y) = \mu_Y$.

Предполагаме, че X_1, X_2, \dots, X_n и Y_1, Y_2, \dots, Y_m са независими.

Означаваме

$$\bar{X} = \frac{1}{n}(X_1 + \dots + X_n), \quad \bar{Y} = \frac{1}{m}(Y_1 + \dots + Y_m),$$

$$S_X = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}, \quad S_Y = \sqrt{\frac{\sum_{i=1}^m (Y_i - \bar{Y})^2}{m-1}}.$$

Нека \bar{x} , \bar{y} , s_X и s_Y са наблюдаваните стойности съответно на \bar{X} , \bar{Y} , S_X и S_Y .

Искаме да проверим хипотезата

$$H_0 : \mu_X = \mu_Y$$

срещу алтернативата

$$H_1 : \mu_X \neq \mu_Y$$

Разглеждаме случайната величина

$$T = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}.$$

Ако X и Y са нормално разпределени, случайната величина T има приблизително t -разпределение с ν степени на свобода, където

$$\nu = \frac{\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2}{\frac{(s_X^2/n)^2}{n-1} + \frac{(s_Y^2/m)^2}{m-1}}.$$

В случай, че ν не е цяло число, се закръглява надолу. Ако X и Y имат произволно разпределение, за големи стойности на n и m случайната величина T отново има приблизително t -разпределение с ν степени на свобода. В този случай може да се използва и приближение със стандартно нормално разпределение.

Нека t_{obs} е наблюдаваната стойност на T , като сме заместили $\mu_X - \mu_Y = 0$,

$$t_{\text{obs}} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}.$$

Отхвърляме нулевата хипотеза, ако $\text{P-value} = \mathbf{P}_0(T \geq |t_{\text{obs}}| \cup T \leq -|t_{\text{obs}}|) \leq \alpha$.

Как се пресмята Р-стойността при различните алтернативни хипотези е дадено в таблицата по-долу.

Р-стойност при различни алтернативни хипотези

$H_1: \mu_X \neq \mu_Y$	$H_1: \mu_X > \mu_Y$	$H_1: \mu_X < \mu_Y$
$\mathbf{P}_0(T \geq t_{\text{obs}} \cup T \leq - t_{\text{obs}})$ $= 2 \mathbf{P}_0(T \geq t_{\text{obs}}) = \text{P-value}$	$\mathbf{P}_0(T \geq t_{\text{obs}}) = \text{P-value}$	$\mathbf{P}_0(T \leq t_{\text{obs}}) = \text{P-value}$

t-тест за разлика на средни: независими извадки

Статистика:

Пресмятаме наблюдаваната стойност $t_{\text{obs}} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}.$

Р-стойност:

H_0	H_1	P-value
$\mu_X = \mu_Y$	$\mu_X \neq \mu_Y$	$2 \mathbf{P}_0(T \geq t_{\text{obs}}) = 2*(1-\text{pt}(\text{abs}(\text{t.obs}), \text{df}=\nu))$
$\mu_X = \mu_Y$	$\mu_X > \mu_Y$	$\mathbf{P}_0(T \geq t_{\text{obs}}) = 1-\text{pt}(\text{t.obs}, \text{df}=\nu)$
$\mu_X = \mu_Y$	$\mu_X < \mu_Y$	$\mathbf{P}_0(T \leq t_{\text{obs}}) = \text{pt}(\text{t.obs}, \text{df}=\nu)$

Извод:

Ако $\text{P-value} \leq \alpha$, отхвърляме H_0 в полза на H_1 .

Ако $\text{P-value} > \alpha$, нямаме достатъчно основания да отхвърлим H_0 .

Ако във вектора **x** запишем наблюденията x_1, x_2, \dots, x_n , а във вектора **y** – наблюденията y_1, y_2, \dots, y_m , може да използваме функцията **t.test**:

$H_1: \mu_X \neq \mu_Y$ **t.test(x, y)**

$H_1: \mu_X > \mu_Y$ **t.test(x, y, alternative='greater')**

$H_1: \mu_X < \mu_Y$ **t.test(x, y, alternative='less')**

Задача 9.1. Мениджър обмисля въвеждане на допълнителна 15-минутна почивка за работниците си. За да разбере дали такава почивка ще намали броя на грешките, които правят работниците, избрал случайно 2 групи по 10 души. Първата група имала допълнителна почивка, а втората работила по обичайното работно време. Данните за броя на допуснатите грешки от двете групи са следните:

Група 1: 8, 7, 5, 8, 10, 9, 7, 8, 4, 5
 Група 2: 7, 6, 14, 12, 13, 8, 9, 6, 10, 9

Приемаме, че данните са приблизително нормално разпределени. Може ли да се заключи, че работниците с допълнителна почивка правят средно по-малко грешки?

Решение. Означаваме с μ_X средния брой грешки на работниците с допълнителна почивка и с μ_Y средния брой грешки на работниците без допълнителна почивка.

Проверяваме хипотезата

$$H_0 : \mu_X = \mu_Y$$

срещу алтернативата

$$H_1 : \mu_X < \mu_Y$$

```
> x <- c(8, 7, 5, 8, 10, 9, 7, 8, 4, 5)
> y <- c(7, 6, 14, 12, 13, 8, 9, 6, 10, 9)
> n <- length(x)
> m <- length(y)
> t.obs <- (mean(x) - mean(y)) / sqrt(var(x)/n + var(y)/m)
> t.obs
[1] -2.126351
> df <- (var(x)/n + var(y)/m)^2 / (((var(x)/n)^2)/(n-1) + ((var(y)/m)^2)/(m-1))
> df
[1] 15.77959
> p.value <- pt(t.obs, df)
> p.value
[1] 0.02480789

> t.test(x, y, alt="less")
```

Welch Two Sample t-test

```
data: x and y
t = -2.1264, df = 15.78, p-value = 0.02481
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
 -Inf -0.4099189
sample estimates:
mean of x mean of y
    7.1      9.4
```

Р-стойността е по-малка от 0.05, следователно отхвърляме нулевата хипотеза в полза на алтернативната. *Имаме основание да твърдим, че средният брой грешки на работниците с допълнителна почивка е по-малък от средния брой на тези без почивка.*

■

9.2. t -тест при зависими извадки

Често се налага да сравняваме извадки, които не са независими. Например, измерено е кръвното налягане на 25 души преди и след приема на дадено лекарство. Първата извадка са измерванията *преди*, а втората – *след*. Нека означим данните съответно x_1, x_2, \dots, x_{25} и y_1, y_2, \dots, y_{25} . Двете извадки не са независими, тъй като x_i и y_i са измервания върху един и същи обект (участник в изследването), но при различни условия (преди и след приема на лекарството). Данните в случая са двойки наблюдения (*paired data, paired samples*), които записваме така: $(x_1, y_1), \dots, (x_{25}, y_{25})$.

Нека X_1, X_2, \dots, X_n са независими и еднакво разпределени случайни величини. Нека X е сл.в. със същото разпределение като X_i и $E(X) = \mu_X$. Нека Y_1, Y_2, \dots, Y_n са независими и еднакво разпределени случайни величини. Нека Y е сл.в. със същото разпределение като Y_i и $E(Y) = \mu_Y$. Предполагаме, че за всяко i случайните величини X_i и Y_i са свързани по някакъв начин (не са независими). Нека $(x_1, y_1), \dots, (x_n, y_n)$ са наблюдения над $(X_1, Y_1), \dots, (X_n, Y_n)$.

Искаме да проверим хипотезата

$$H_0: \mu_X = \mu_Y$$

срещу алтернативата

$$H_1: \mu_X \neq \mu_Y$$

Означаваме

$$D_i = X_i - Y_i, \quad i = 1, \dots, n,$$

$$\mu_D = \mu_X - \mu_Y,$$

$$\bar{D} = \frac{1}{n}(D_1 + \dots + D_n), \quad S_D = \sqrt{\frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n-1}}.$$

Нека \bar{d} е наблюдаваната стойност на \bar{D} , а s_D е наблюдаваната стойност на S_D .

Задачата е еквивалентна на проверка на хипотезата

$$H_0: \mu_D = 0$$

срещу алтернативата

$$H_1: \mu_D \neq 0$$

Когато X и Y са нормално разпределени, случайната величина

$$T = \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}}$$

има t -разпределение с $n-1$ степени на свобода. Ако X и Y имат произволно разпределение, за големи стойности на n случайната величина T има приблизително t -разпределение с $n-1$ степени на свобода.

Нека t_{obs} е наблюдаваната стойност на T , като сме заместили $\mu_D = \mu_X - \mu_Y = 0$,

$$t_{\text{obs}} = \frac{\bar{d}}{s_D/\sqrt{n}}.$$

Отхвърляме нулевата хипотеза, ако $P\text{-value} = \mathbf{P}_0(T \geq |t_{\text{obs}}| \cup T \leq -|t_{\text{obs}}|) \leq \alpha$.

Описаната процедура е по същество t -тест приложен за данните d_1, d_2, \dots, d_n , където $d_i = x_i - y_i$.

t-тест за разлика на средни: зависими извадки

Статистика:

Пресмятаме наблюдаваната стойност $t_{\text{obs}} = \frac{\bar{d}}{s_D/\sqrt{n}}$.

P-стойност:

H_0	H_1	P-value
$\mu_X = \mu_Y$	$\mu_X \neq \mu_Y$	$2\mathbf{P}_0(T \geq t_{\text{obs}}) = 2*(1-\text{pt}(\text{abs}(\mathbf{t. obs}), \mathbf{n-1}))$
$\mu_X = \mu_Y$	$\mu_X > \mu_Y$	$\mathbf{P}_0(T \geq t_{\text{obs}}) = 1-\text{pt}(\mathbf{t. obs}, \mathbf{n-1})$
$\mu_X = \mu_Y$	$\mu_X < \mu_Y$	$\mathbf{P}_0(T \leq t_{\text{obs}}) = \text{pt}(\mathbf{t. obs}, \mathbf{n-1})$

Извод:

Ако P-value $\leq \alpha$, отхвърляме H_0 в полза на H_1 .

Ако P-value $> \alpha$, нямаме достатъчно основания да отхвърлим H_0 .

За зависими извадки отново се използва функцията `t.test`, като трябва да се добави `paired=T` (във вектора `x` са наблюденията x_1, x_2, \dots, x_n , а във вектора `y` са y_1, y_2, \dots, y_n):

$H_1: \mu_X \neq \mu_Y$ `t.test(x, y, paired=T)`

$H_1: \mu_X > \mu_Y$ `t.test(x, y, alternative='greater', paired=T)`

$H_1: \mu_X < \mu_Y$ `t.test(x, y, alternative='less', paired=T)`

Командата `t.test(x, y, paired=T)` дава същия резултат като `t.test(x-y)`.

Задача 9.2. За да се изследва ефекта на диета върху нивото на холестерол в кръвта са избрани 15 мъже на възраст между 35 и 50 години. Нивото на холестерола на всеки участник е измерено преди започване на диетата и три месеца след прилагане на диетата. Данните са следните:

Участник	Преди	След
1	265	229
2	240	231
3	258	227
4	295	240
5	251	238
6	245	241
7	287	234
8	314	256
9	260	247
10	279	239
11	283	246
12	240	218
13	238	219
14	225	226
15	247	233

Приемаме, че нивото на холестерол е нормално разпределено. Дали тези данни дават основание да се твърди, че диетата намалява нивото на холестерол в средно?

Решение. Нека μ_X е средното ниво на холестерол преди прилагане на диетата и μ_Y е средното ниво на холестерол три месеца след прилагане на диетата.

Проверяваме хипотезата

$$H_0 : \mu_X = \mu_Y$$

срещу алтернативата

$$H_1 : \mu_X > \mu_Y$$

```
> x <- c(265,240,258,295,251,245,287,314,260,279,283,240,238,225,247)
> y <- c(229,231,227,240,238,241,234,256,247,239,246,218,219,226,233)
> d <- x - y
> n <- length(x)
> t.obs <- mean(d) / (sd(d)/sqrt(n))
> t.obs
[1] 5.465874
> p.value <- 1-pt(t.obs, n-1)
> p.value
[1] 4.157964e-05
> t.test(x, y, alt="greater", paired=T)
```

Paired t-test

```
data: x and y
t = 5.4659, df = 14, p-value = 4.158e-05
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 18.20922      Inf
sample estimates:
mean of the differences
      26.86667
```

Р-стойността е по-малка от 0.05, следователно отхвърляме нулевата хипотеза в полза на алтернативната. *Имаме основание да твърдим, че средното ниво на холестерол три месеца след прилагане на диетата е по-ниско от това преди диетата.*

■

9.3. z -тест за разлика на пропорции

Нека x_1 е наблюдавана стойност на $X_1 \sim \text{Bi}(n_1, p_1)$, а x_2 наблюдавана стойност на $X_2 \sim \text{Bi}(n_2, p_2)$. Искаме да сравним вероятностите p_1 и p_2 , т.е. да проверим хипотезата $H_0 : p_1 = p_2$. Такава задача може да възникне, например, ако искаме да сравним вероятността мъж да е пушач и вероятността жена да е пушач. Други примери: вероятността да се появи дефект в батериите произведени от завод 1 и завод 2, вероятността да се появят странични ефекти при употреба на лекарство 1 и лекарство 2. Данните, които имаме са: от n_1 опита при дадено условие сме наблюдавали x_1 пъти успех и от n_2 опита при друго условие сме наблюдавали x_2 пъти успех.

Означаваме

$$\hat{P}_1 = X_1/n_1, \quad \hat{P}_2 = X_2/n_2,$$

$$\hat{P} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1 \hat{P}_1 + n_2 \hat{P}_2}{n_1 + n_2}.$$

Нека \hat{p}_1 , \hat{p}_2 и \hat{p} са наблюдаваните стойности съответно на \hat{P}_1 , \hat{P}_2 и \hat{P} .

Случайната величина

$$Z = \frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sqrt{\hat{P}(1 - \hat{P})(1/n_1 + 1/n_2)}}$$

има приблизително стандартно нормално разпределение за големи стойности на n_1 и n_2 .

Искаме да проверим хипотезата

$$H_0 : p_1 = p_2$$

срещу алтернативата

$$H_1 : p_1 \neq p_2$$

Нека z_{obs} е наблюдаваната стойност на Z , като сме заместили $p_1 - p_2 = 0$,

$$z_{\text{obs}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})(1/n_1 + 1/n_2)}}.$$

Отхвърляме нулевата хипотеза, ако $\text{P-value} = \mathbf{P}_0(Z \geq |z_{\text{obs}}| \cup Z \leq -|z_{\text{obs}}|) \leq \alpha$.

Как се пресмята Р-стойността при различните алтернативни хипотези е дадено в таблицата по-долу.

Р-стойност при различни алтернативни хипотези

$H_1 : p_1 \neq p_2$	$H_1 : p_1 > p_2$	$H_1 : p_1 < p_2$
$\mathbf{P}_0(Z \geq z_{\text{obs}} \cup Z \leq - z_{\text{obs}})$ $= 2 \mathbf{P}_0(Z \geq z_{\text{obs}}) = \text{P-value}$	$\mathbf{P}_0(Z \geq z_{\text{obs}}) = \text{P-value}$	$\mathbf{P}_0(Z \leq z_{\text{obs}}) = \text{P-value}$

z-тест за разлика на пропорции

Статистика:

Пресмятаме наблюдаваната стойност $z_{\text{obs}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$.

P-стойност:

H_0	H_1	P-value
$p_1 = p_2$	$p_1 \neq p_2$	$2 \mathbf{P}_0(Z \geq z_{\text{obs}}) = 2*(1-\text{pnorm}(\text{abs}(z.\text{obs})))$
$p_1 = p_2$	$p_1 > p_2$	$\mathbf{P}_0(Z \geq z_{\text{obs}}) = 1-\text{pnorm}(z.\text{obs})$
$p_1 = p_2$	$p_1 < p_2$	$\mathbf{P}_0(Z \leq z_{\text{obs}}) = \text{pnorm}(z.\text{obs})$

Извод:

Ако P-value $\leq \alpha$, отхвърляме H_0 в полза на H_1 .

Ако P-value $> \alpha$, нямаме достатъчно основания да отхвърлим H_0 .

Ако във вектора **x** запишем x_1, x_2 , а във вектора **n** запишем n_1, n_2 , може да използваме функцията **prop.test**:

$H_1 : p_1 \neq p_2$ `prop.test(x, n, correct=F)`

$H_1 : p_1 > p_2$ `prop.test(x, n, alternative='greater', correct=F)`

$H_1 : p_1 < p_2$ `prop.test(x, n, alternative='less', correct=F)`

Задача 9.3. В проучване участвали 220 жени и 210 мъже. Според резултатите, 71 жени и 58 мъже отговорили, че предпочитат безкофеиново кафе. Може ли да твърдим, че процентът на жените, предпочитащи безкофеиново кафе, е различен от процентът на мъжете, предпочитащи безкофеиново кафе?

Решение. Нека p_1 е вероятността произволно избрана жена да предпочита безкофеиново кафе, а p_2 е вероятността произволно избран мъж да предпочита безкофеиново кафе. Проверяваме хипотезата

$$H_0 : p_1 = p_2$$

срещу алтернативата

$$H_1 : p_1 \neq p_2$$

По условие имаме:

$$x_1 = 71, \quad x_2 = 58$$

$$n_1 = 220, \quad n_2 = 210$$

```
> x <- c(71,58)
> n <- c(220,210)
> p.hat <- (x[1] + x[2]) / (n[1] + n[2])
> z.obs <- (x[1]/n[1] - x[2]/n[2]) / sqrt(p.hat*(1-p.hat)*(1/n[1] + 1/n[2]))
> z.obs
[1] 1.052625
> p.value <- 2*(1-pnorm(abs(z.obs)))
> p.value
[1] 0.292513
```

```
> prop.test(x, n, correct=F)
```

2-sample test for equality of proportions without continuity correction

```
data:  x out of n
X-squared = 1.108, df = 1, p-value = 0.2925
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.03991226  0.13298585
sample estimates:
   prop 1    prop 2 
0.3227273 0.2761905
```

P-стойността е по-голяма от 0.05, следователно нямаме достатъчно основания да отхвърлим нулевата хипотеза. *Нямаме основания да твърдим, че процентът на жените, предпочитащи безкофеиново кафе, е различен от процентът на мъжете, предпочитащи безкофеиново кафе.*

■

10. Хи-квадрат тестове

*The 'laws of Nature' are only constructs of our minds;
none of them can be asserted to be true or to be false,
they are good in so far as they give good fits to our observations
of Nature, and are liable to be replaced by a better 'fit'. . .*

Karl Pearson

10.1. Хи-квадрат тест за съгласуваност

Тестовите за съгласуваност (*goodness-of-fit tests*) се използват за да се провери доколко данните са съгласувани с даден вероятностен модел (дали този модел описва добре данните).

Разглеждаме експеримент (опит) с k възможни изхода A_1, A_2, \dots, A_k (пълна група събития). Означаваме вероятностите $\mathbf{P}(A_1) = p_1, \dots, \mathbf{P}(A_k) = p_k$. В поредица от n независими опити нека X_i е броя случвания на събитието A_i . Случайният вектор (X_1, X_2, \dots, X_k) има полиномно разпределение (*multinomial distribution*) с параметри n, p_1, \dots, p_k .

Нека x_i е наблюдаваната стойност на X_i , т.е. от n опита, събитието A_i се е случило x_i пъти, $\sum_i x_i = n$. За големи стойности на n , случайната величина

$$\chi^2 = \sum_{i=1}^k \frac{(X_i - np_i)^2}{np_i}$$

има хи-квадрат разпределение с $k - 1$ степени на свобода (χ^2 -разпределение, *chi-square distribution*).

Искаме да проверим хипотезата

$$H_0 : (p_1, p_2, \dots, p_k) = (p_1^o, p_2^o, \dots, p_k^o)$$

срещу алтернативата

$$H_1 : (p_1, p_2, \dots, p_k) \neq (p_1^o, p_2^o, \dots, p_k^o),$$

където p_1^o, \dots, p_k^o са предварително зададени стойности.

Пресмятаме

$$\chi_{\text{obs}}^2 = \sum_{i=1}^k \frac{(x_i - np_i^o)^2}{np_i^o},$$

$$\text{P-value} = \mathbf{P}_0(\chi^2 > \chi_{\text{obs}}^2).$$

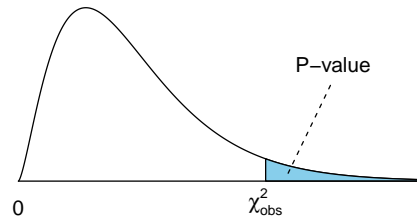
Ако $\text{P-value} \leq \alpha$, отхвърляме H_0 в полза на H_1 .

Ако $\text{P-value} > \alpha$, нямаме достатъчно основания да отхвърлим H_0 .

Р-стойността се пресмята в R по следния начин:

$$\chi_{\text{obs}}^2 = \text{chi2.obs}$$

$$\text{P-value} = \mathbf{P}_0(\chi^2 > \chi_{\text{obs}}^2) = 1 - \text{pchisq}(\text{chi2.obs}, \text{df} = k-1).$$



Ако във вектора **x** запишем x_1, \dots, x_k , а във вектора **probs** запишем p_1^o, \dots, p_k^o , може да използваме функцията `chisq.test(x, p=probs)`.

Задача 10.1. Честотите на срещане на буквите в английския език са следните (в %):

E	T	A	O	I	N	S	R	H	D	other
12.02	9.10	8.12	7.68	7.31	6.95	6.28	6.02	5.92	4.32	26.28

В текст, състоящ се от 2004 букви, броят срещания на съответните букви е:

E	T	A	O	I	N	S	R	H	D	other
221	153	183	111	113	152	103	197	38	104	629

Може ли да се твърди, че текстът е на английски?

Решение. Нека p_E е вероятността да се срещне буква E в текст на същия език, p_T е вероятността да се срещне буква T , и т.н., p_{oth} е вероятността да се срещне някоя от останалите 16 букви в английската азбука.

Проверяваме хипотезата

$$H_0 : (p_E, p_T, \dots, p_{oth}) = (0.1202, 0.0910, \dots, 0.2628)$$

срещу алтернативата

$$H_1 : (p_E, p_T, \dots, p_{oth}) \neq (0.1202, 0.0910, \dots, 0.2628).$$

```
> load("letterFreq.RData")
> probs*100
  E    T    A    O    I    N    S    R    H    D other
12.02 9.10 8.12 7.68 7.31 6.95 6.28 6.02 5.92 4.32 26.28
> x1
  E    T    A    O    I    N    S    R    H    D other
 221  153  183  111  113  152  103  197  38  104  629
> chisq.test(x1, p=probs)
```

Chi-squared test for given probabilities

```
data: x1
X-squared = 160.36, df = 10, p-value < 2.2e-16
```

Р-стойността е по-малка от 0.05, следователно отхвърляме нулевата хипотеза в полза на алтернативната. Нямаме основания да твърдим, че вероятностите за срещане на съответните букви са както в английския език.

■

10.2. Хи-квадрат тест за независимост

Разглеждаме експеримент, чиито изходи могат да бъдат класифицирани по два критерия на A_1, A_2, \dots, A_r или B_1, B_2, \dots, B_c , т.е. изходите могат да бъдат представени като двойки (A_i, B_j) . С други думи, имаме две категорни променливи A и B с възможни стойности, съответно A_1, A_2, \dots, A_r и B_1, B_2, \dots, B_c ; например, цвят на очите и цвят на косата. Нека p_{ij} е вероятността да се случи (A_i, B_j) , т.е. изходът да е класифициран като A_i и същевременно като B_j ; например, вероятността човек да е с руса коса и кафяви очи. Вероятността да се случи A_i е $p_{i\bullet} = \sum_j p_{ij}$, а вероятността да се случи B_j е $p_{\bullet j} = \sum_i p_{ij}$. Вероятностите могат да се представят в следната таблица:

	B_1	B_2	\dots	B_c	
A_1	p_{11}	p_{12}	\dots	p_{1c}	$p_{1\bullet}$
A_2	p_{21}	p_{22}	\dots	p_{2c}	$p_{2\bullet}$
\vdots	\vdots	\vdots		\vdots	\vdots
A_r	p_{r1}	p_{r2}	\dots	p_{rc}	$p_{r\bullet}$
	$p_{\bullet 1}$	$p_{\bullet 2}$	\dots	$p_{\bullet c}$	

В поредица от n независими опити нека X_{ij} е броя случвания на (A_i, B_j) . Случайният вектор (X_{ij}) , $i = 1, \dots, r$, $j = 1, \dots, c$, има полиномно разпределение с параметри n, p_{ij} , $i = 1, \dots, r$, $j = 1, \dots, c$. Нека $X_{i\bullet} = \sum_j X_{ij}$ и $X_{\bullet j} = \sum_i X_{ij}$.

Нека x_{ij} е наблюдаваната стойност на X_{ij} . В сила е $\sum_{i,j} x_{ij} = n$. Данните могат да се представят в таблица от вида:

	B_1	B_2	\dots	B_c	
A_1	x_{11}	x_{12}	\dots	x_{1c}	$x_{1\bullet}$
A_2	x_{21}	x_{22}	\dots	x_{2c}	$x_{2\bullet}$
\vdots	\vdots	\vdots		\vdots	\vdots
A_r	x_{r1}	x_{r2}	\dots	x_{rc}	$x_{r\bullet}$
	$x_{\bullet 1}$	$x_{\bullet 2}$	\dots	$x_{\bullet c}$	

Искаме да проверим хипотезата

$$H_0 : p_{ij} = p_{i\bullet}p_{\bullet j} \text{ за всяка двойка } (i, j)$$

срещу алтернативата

$$H_1 : p_{ij} \neq p_{i\bullet}p_{\bullet j} \text{ за поне една двойка } (i, j)$$

Нулевата хипотеза означава, че променливите A и B са независими, т.е. за всяка двойка (i, j) е вярно $\mathbf{P}(A_i, B_j) = \mathbf{P}(A_i)\mathbf{P}(B_j)$. Алтернативната хипотеза означава, че има някаква връзка (зависимост) между променливите A и B .

Ако нулевата хипотеза е вярна, за големи стойности на n , случайната величина

$$\chi^2 = \sum_{i,j} \frac{(X_{ij} - X_{i\bullet}X_{\bullet j}/n)^2}{X_{i\bullet}X_{\bullet j}/n}$$

има хи-квадрат разпределение с $(r - 1)(c - 1)$ степени на свобода.

Пресмятаме

$$\chi_{\text{obs}}^2 = \sum_{i,j} \frac{(x_{ij} - x_{i\bullet}x_{\bullet j}/n)^2}{x_{i\bullet}x_{\bullet j}/n},$$

$$\text{P-value} = \mathbf{P}_0(\chi^2 > \chi_{\text{obs}}^2).$$

Ако $\text{P-value} \leq \alpha$, отхвърляме H_0 в полза на H_1 .

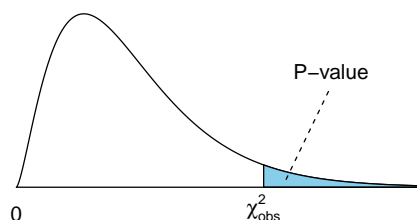
Ако $\text{P-value} > \alpha$, нямаме достатъчно основания да отхвърлим H_0 .

Р-стойността се пресмята в R по следния начин:

$$(r - 1)(c - 1) = \text{df}$$

$$\chi_{\text{obs}}^2 = \text{chi2.obs}$$

$$\text{P-value} = \mathbf{P}_0(\chi^2 > \chi_{\text{obs}}^2) = 1 - \text{pchisq}(\text{chi2.obs}, \text{df}).$$



Ако запишем данните $\{x_{ij}\}$ в матрица или таблица **x**, може да използваме функцията `chisq.test(x)`.

Задача 10.2. Разгледайте данните `HairEyeColor`. Има ли връзка между цвета на косата и цвета на очите?

Решение. Проверяваме хипотезата

H_0 : цвета на косата и цвета на очите са независими

срещу алтернативата

H_1 : има връзка между цвета на косата и цвета на очите.

```
> data(HairEyeColor)
> tb <- HairEyeColor[,1] + HairEyeColor[,2]
> tb
      Eye
Hair   Brown Blue Hazel Green
Black   68   20   15    5
Brown  119   84   54   29
Red     26   17   14   14
Blond    7   94   10   16
> n <- sum(tb)
> df <- (nrow(tb)-1)*(ncol(tb)-1)
> hair <- apply(tb, 1, sum)
```

```

> eyes <- apply(tb, 2, sum)

> expected <- (hair %o% eyes)/n
> observed <- tb
> chi2.obs <- sum( (observed - expected)^2 / expected )
> chi2.obs
[1] 138.2898
> p.value <- 1-pchisq(chi2.obs, df)
> p.value
[1] 0

> chisq.test(tb)

```

Pearson's Chi-squared test

```

data:  tb
X-squared = 138.29, df = 9, p-value < 2.2e-16

> chisq.test(tb)$p.value
[1] 2.325287e-25

```

P-стойността е по-малка от 0.05, следователно отхвърляме нулевата хипотеза в полза на алтернативната. Можем да твърдим, че има връзка между цвета на косата и цвета на очите. ■

11. Линејни модели

*All models are approximations.
Essentially, all models are wrong,
but some are useful.*

George E. P. Box

11.1. Линеен модел с един предиктор

Предполагаме, че случайните величини Y , X и ε са свързани по следния начин:

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

където $E(\varepsilon) = 0$, $\text{Var}(\varepsilon) = \sigma^2$, а β_0 и β_1 са константи. Тогава за средното на Y при условие, че $X = x$, е вярно

$$E(Y | X = x) = \mu_{y|x} = \beta_0 + \beta_1 x.$$

Ще наричаме X *предиктор*, а Y – *отклик*.

Нека (x_i, y_i) , $i = 1, \dots, n$, са независими наблюдения генерирани от горния модел. Тогава за $i = 1, \dots, n$,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Параметрите β_0, β_1 са неизвестни. Въз основа на наблюденията $(x_1, y_1), \dots, (x_n, y_n)$ намираме оценки $\hat{\beta}_0, \hat{\beta}_1$ по метода на най-малките квадрати, т.е. $\hat{\beta}_0$ и $\hat{\beta}_1$ минимизират сумата

$$\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

Така получаваме оценка за средното на Y при условие, че $X = x$:

$$\hat{\mu}_{y|x} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Горното уравнение често се записва във вида

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

и се използва за прогнозиране на стойността на Y при $X = x$, прогнозираната стойност се означава \hat{y} . Това уравнение се нарича *оценено регресионно уравнение* или *оценен модел*.

Разликата между наблюдаваната стойност y_i и прогнозираната стойност \hat{y}_i се нарича *остатък*, $e_i = y_i - \hat{y}_i$.

Коефициент на детерминация (R^2)

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

Доверителен интервал за β_1

$$\left[\hat{\beta}_1 - t_{n-2, \alpha/2} \text{SE}(\hat{\beta}_1), \quad \hat{\beta}_1 + t_{n-2, \alpha/2} \text{SE}(\hat{\beta}_1) \right]$$

Проверка на хипотези за β_1

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

$$t_{\text{obs}} = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)}$$

$$\text{P-value} = 2 \cdot (1 - \text{pt}(\text{abs}(t.\text{obs}), n-2))$$

⟨!⟩ Доверителните интервали и тестовете за β_1 са валидни ако ε има нормално разпределение **или** броят на наблюденията n е достатъчно голям.

Ако във вектора **x** са наблюденията x_1, x_2, \dots, x_n , а във вектора **y** са y_1, y_2, \dots, y_n , намираме оценения модел с `lm(y ~ x)`.

Ако запазим резултата `m1 <- lm(y ~ x)`, следните функции са полезни:

Функция	Резултат
<code>summary(m1)</code>	основна информация за оценения модел
<code>summary(m1)\$coefficients</code>	таблица за оценените коефициенти
<code>summary(m1)\$r.squared</code>	R^2
<code>coef(m1)</code> или <code>coefficients(m1)</code>	оценените коефициенти $\hat{\beta}_0, \hat{\beta}_1$
<code>confint(m1)</code>	доверителни интервали за β_0, β_1
<code>resid(m1)</code> или <code>residuals(m1)</code>	остатъците $e_i = y_i - \hat{y}_i$
<code>fitted(m1)</code> или <code>fitted.values(m1)</code>	\hat{y}_i
<code>predict(m1, new, interval="confidence")</code>	доверителен интервал за $\mu_{y x}$ при $x = x^*$
<code>predict(m1, new, interval="prediction")</code>	интервал за прогноза на y при $x = x^*$
<code>predict(m1, new, interval="none")</code>	\hat{y} за дадено $x = x^*$

11.2. Линеен модел с няколко предиктора

Предполагаме, че случайните величини Y, X_1, \dots, X_k и ε са свързани по следния начин:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon,$$

където $E(\varepsilon) = 0$, $\text{Var}(\varepsilon) = \sigma^2$, а $\beta_0, \beta_1, \dots, \beta_k$ са константи. Тогава за средното на Y при условие, че $X_1 = x_1, \dots, X_k = x_k$, е вярно

$$E(Y | X_1 = x_1, \dots, X_k = x_k) = \mu_{y|x_1 \dots x_k} = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k.$$

Ще наричаме X_1, \dots, X_k *предиктори*, а Y – *отклик*.

Нека $(x_{1i}, \dots, x_{ki}, y_i)$, $i = 1, \dots, n$, са независими наблюдения генерирани от горния модел. Тогава за $i = 1, \dots, n$,

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \varepsilon_i.$$

Параметрите $\beta_0, \beta_1, \dots, \beta_k$ са неизвестни. Въз основа на наблюденията $(x_{1i}, \dots, x_{ki}, y_i)$, $i = 1, \dots, n$, намираме оценки $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ по метода на най-малките квадрати, т.е. $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ минимизират сумата

$$\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{1i} - \dots - \beta_k x_{ki})^2.$$

Така получаваме оценка за средното на Y при условие, че $X_1 = x_1, \dots, X_k = x_k$:

$$\hat{\mu}_{y|x_1 \dots x_k} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k.$$

Горното уравнение често се записва във вида

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k$$

и се използва за прогнозиране на стойността на Y при $X_1 = x_1, \dots, X_k = x_k$, прогнозираната стойност се означава \hat{y} . Това уравнение се нарича *оценено регресионно уравнение* или *оценен модел*.

Разликата между наблюдаваната стойност y_i и прогнозираната стойност \hat{y}_i се нарича *остатък*, $e_i = y_i - \hat{y}_i$.

Коефициент на детерминация (R^2)

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Коригиран R^2 (*adjusted R^2*)

$$R_{\text{adj}}^2 = 1 - \frac{\frac{1}{n-k-1} \sum_i (y_i - \hat{y}_i)^2}{\frac{1}{n-1} \sum_i (y_i - \bar{y})^2}$$

Доверителен интервал за β_j

$$\left[\hat{\beta}_j - t_{n-k-1, \alpha/2} \text{SE}(\hat{\beta}_j), \quad \hat{\beta}_j + t_{n-k-1, \alpha/2} \text{SE}(\hat{\beta}_j) \right]$$

Проверка на хипотези за β_j

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

$$t_{\text{obs}} = \frac{\hat{\beta}_j}{\text{SE}(\hat{\beta}_j)}$$

$$\text{P-value} = 2 * (1 - \text{pt}(\text{abs}(t.\text{obs}), n - k - 1))$$

⟨!⟩ Доверителните интервали и тестовете за β_j са валидни ако ε има нормално разпределение **или** броят на наблюденията n е достатъчно голям.

Ако във вектора **x1** са наблюденията $x_{11}, x_{12}, \dots, x_{1n}$, във вектора **x2** са $x_{21}, x_{22}, \dots, x_{2n}$ и във вектора **y** са y_1, y_2, \dots, y_n , намираме оценения модел с `lm(y ~ x1 + x2)`.

Ако запазим резултата `m1 <- lm(y ~ x1 + x2)`, следните функции са полезни:

Функция	Резултат
<code>summary(m1)</code>	основна информация за оценения модел
<code>summary(m1)\$coefficients</code>	таблица за оценените коефициенти
<code>summary(m1)\$r.squared</code>	R^2
<code>summary(m1)\$adj.r.squared</code>	коригиран R^2
<code>coef(m1)</code> или <code>coefficients(m1)</code>	оценените коефициенти $\hat{\beta}_0, \dots, \hat{\beta}_k$
<code>confint(m1)</code>	доверителни интервали за β_0, \dots, β_k
<code>resid(m1)</code> или <code>residuals(m1)</code>	остатъците $e_i = y_i - \hat{y}_i$
<code>fitted(m1)</code> или <code>fitted.values(m1)</code>	\hat{y}_i
<code>predict(m1, new, interval="confidence")</code>	доверителен интервал за $\mu_{y x_1 \dots x_k}$ при $(x_1, \dots, x_k) = (x_1^*, \dots, x_k^*)$
<code>predict(m1, new, interval="prediction")</code>	интервал за прогноза на y при $(x_1, \dots, x_k) = (x_1^*, \dots, x_k^*)$
<code>predict(m1, new, interval="none")</code>	\hat{y} за дадено $(x_1, \dots, x_k) = (x_1^*, \dots, x_k^*)$

Литература

- [1] Devore J.L.: *Probability and Statistics for Engineering and the Sciences*, 7th ed, Brooks/Cole, 2009.
- [2] Hogg R.V., Craig A.T.: *Introduction to Mathematical Statistics*, 5th ed, Pearson Education, 1995.
- [3] Montgomery D.C., Runger G.C.: *Applied Statistics and Probability for Engineers*, 5th ed, John Wiley & Sons, 2011.
- [4] Ott R.L., Longnecker M.: *An Introduction to Statistical Methods and Data Analysis*, 5th ed, Duxbury/Thomson Learning, 2001.
- [5] Panik M.J.: *Advanced Statistics from an Elementary Point of View*, Elsevier/Academic Press, 2005.
- [6] Peck R., Olsen C., Devore J.L.: *Introduction to Statistics and Data Analysis*, 4th ed, Brooks/Cole, 2012.
- [7] Ramachandran K.M., Tsokos C.P.: *Mathematical Statistics with Applications*, Elsevier/Academic Press, 2009.
- [8] Ross S.M.: *Introduction to Probability and Statistics for Engineers and Scientists*, 3rd ed, Elsevier/Academic Press, 2004.
- [9] Ross S.M.: *Introductory Statistics*, 3rd ed, Elsevier/Academic Press, 2010.
- [10] Tijms H.: *Understanding Probability*, 2nd ed, Cambridge University Press, 2007.
- [11] Trosset M.W.: *An Introduction to Statistical Inference and Its Applications with R*, Chapman & Hall/CRC Press, 2009.
- [12] Venables W.N., Ripley B.D.: *Modern Applied Statistics with S*, 4th ed, Springer, 2002.
- [13] Verzani J.: *Using R for Introductory Statistics*, 2nd ed, Chapman & Hall/CRC Press, 2014.
- [14] Wackerly D.D., Mendenhall W., Scheaffer R.L.: *Mathematical Statistics with Applications*, 7th ed, Brooks/Cole, 2008.
- [15] Walpole R.E., Myers R.H., Myers S.L., Ye K.: *Probability & Statistics for Engineers & Scientists*, 9th ed, Prentice Hall, 2012.
- [16] Wilcox R.R.: *Introduction to Robust Estimation and Hypothesis Testing*, 4th ed, Elsevier/Academic Press, 2017.
- [17] Wilcox R.R.: *Understanding and Applying Basic Statistical Methods Using R*, John Wiley & Sons, 2017.