Duomenų analizės įvadas

2.2. dalis - R programavimas

Justas Mundeikis

VU EVAF

2019-04-04

Turinys

Loop funkcijos

Loop funkcijos

Rašant skriptus, `for`, `while` ir kiti loopai yra tinkami, bet jeigu norima parašyti kodą tiesiog konsolėje, tada susiduriama su daug problemų.

- `lapply`: loopina per list ir paleidžia funkciją kiekvienam elementui
- `sapply`: veikia kaip ir `lapply` tik supaprastina rezultatus
- `apply`: taiko funkciją masyvo stulepliams / eilutėms
- `tapply`: taiko funkciją vektoriaus dalims
- `mapply`: multivariatinė `lapply` versija

apply

apply

`apply` naudojama taikyti funkcijas dataframe, matricų eilutėms ar stulpeliams. `apply` iš esmės supaprastina `for loop` naudojimą.

```
args(apply)
## function (X, MARGIN, FUN, ...)
## NULL
# kur X yra array
# MARGIN=1 eilutėms
# MARGIN=2 stulpeliams
# FUN taikoma funkcija
```

apply

```
x <- matrix(1:4,2,2)
x
##      [,1] [,2]
## [1,]    1    3
## [2,]    2    4
apply(x, 1, mean) #1 - eiltuėms
## [1] 2 3
apply(x, 2, mean) #2 - stulpeliams
## [1] 1.5 3.5
apply(x, 1, sum)
## [1] 4 6
apply(x, 2, sum)
## [1] 3 7
```

apply

Norint pritaikyti apply funkciją daugiau dimensijų turinčiam duomenų masyvui, būtina nurodyti vektorių, kurios dimensijos išlaikomos

```
x <- array(data=rnorm(40), dim = c(2,2,10))
apply(x, c(1,2), mean)
##              [,1]       [,2]
## [1,] -0.07282579 0.01602339
## [2,] -0.09809258 0.21179569
```

apply

Exercice:

- create a sales dataframe
- calculate column then rowsums

```
#generate sales dataframe
set.seed(101)
sales <- data.frame(row.names = month.abb[1:12],
                    Marry=sample(1:12,12, replace = TRUE),
                    John=sample(1:12,12, replace = TRUE),
                    Felix=sample(1:12,12, replace = TRUE),
                    Jenny=sample(1:12,12, replace = TRUE))
```

apply

Solution:

```
# sukuriame 5x6 matricą
set.seed(101)
sales <- data.frame(row.names = month.abb[1:12],
                    Marry=sample(1:12,12, replace = TRUE),
                    John=sample(1:12,12, replace = TRUE),
                    Felix=sample(1:12,12, replace = TRUE),
                    Jenny=sample(1:12,12, replace = TRUE))



apply(sales, MARGIN=1, sum)
## Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
##  29  33  17  33  24  21  29  14  25  24  22  39
apply(sales, MARGIN=2, sum)
## Marry  John Felix Jenny
##    78    86    74    72
```

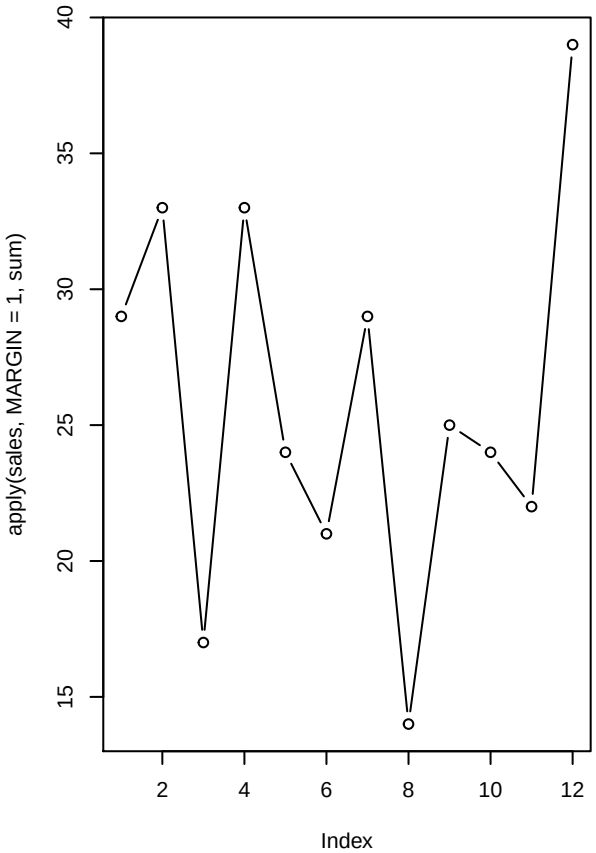
apply

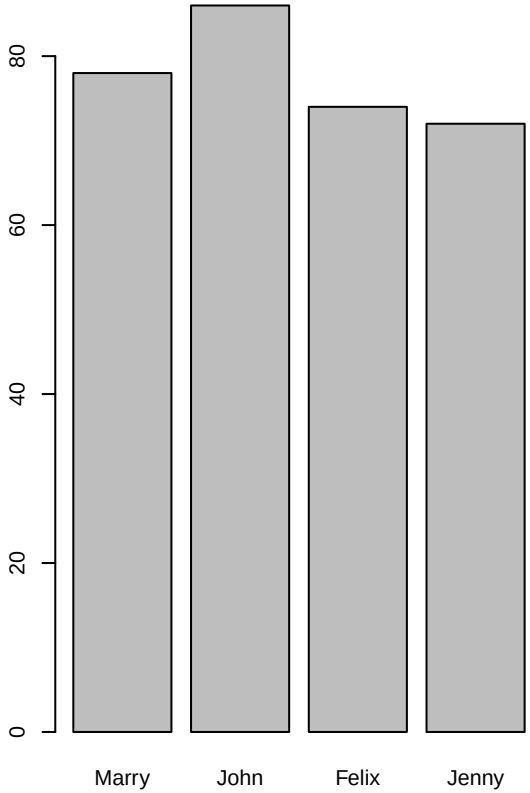
Exercice 1:

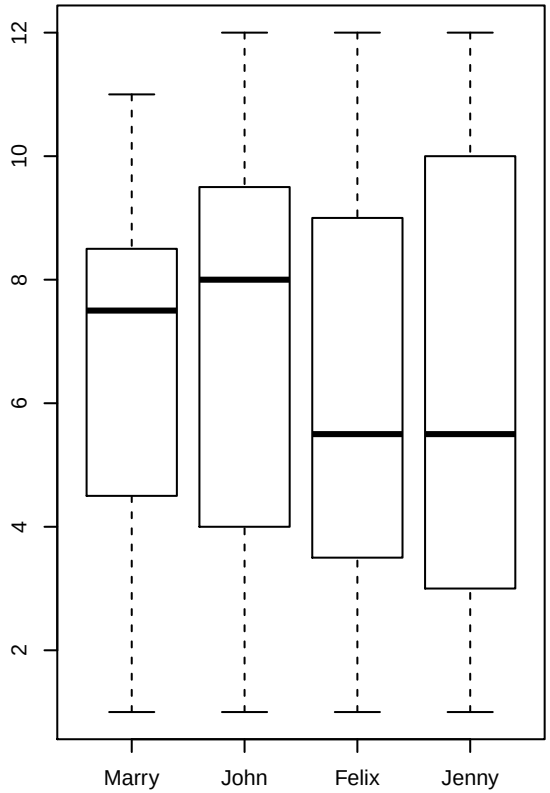
- nubraižykite linijinę diagramą visų pardavimų pamėnesiui
- nubraižykite stulpelinę diagramą kiekvienam pardarvėjui
- nubraižykite boxplotus kiekvienam pardavėjui

apply

Sollutions







apply

Jeigu norima apskaičiuoti dataframe / matricų eilučių ar stuleplių sumas / vidurkius, galima naudoti jau supaprastintas funckijas, jos veikia dar greičiau, nei originalas.

- `rowSums=apply(x,1,sum)`
- `rowMeans=apply(x,1,mean)`
- `colSums=apply(x,2,sum)`
- `colMeans=apply(x,2,mean)`

apply

```
args(apply)
## function (X, MARGIN, FUN, ...)
## NULL
# for ... arguments of quantile function
apply(sales[,-1], 2, quantile, probs=c(0,0.25 ,0.5, 0.75,1))
##       John Felix Jenny
## 0%    1.00  1.00   1.0
## 25%   4.50  3.75   3.0
## 50%   8.00  5.50   5.5
## 75%   9.25  8.50  10.0
## 100% 12.00 12.00  12.0
apply(sales[,-1], 2, summary, digits=0)
##         John Felix Jenny
## Min.       1     1     1
## 1st Qu.    4     4     3
## Median     8     6     6
## Mean       7     6     6
## 3rd Qu.    9     8    10
## Max.      10    10    10
```

rep

rep

```
# Funkcija rep priima objektus ir replikuoja juos times arba each
args(rep)
## function (x, ...)
## NULL
?args
rep(c("a","b"), c(1,3))
## [1] "a" "b" "b" "b"
```

sweep

sweep

Return an array obtained from an input array by sweeping out a summary statistic.

```
args(sweep)
## function (x, MARGIN, STATS, FUN = "-", check.margin = TRUE, ...)
## NULL
```

sweep

- Kartais reikia naudotis normalizuotais duomenimis
- $= \frac{X-\mu}{\sigma}$ PVZ:

```
sales
boxplot(sales)
sales_mean <- apply(sales,2 ,mean)
sales_sd <- apply(sales,2, sd)
df1 <- sweep(sales,2, sales_mean, "-")
df2 <- sweep(df1,2, sales_sd, "/")
boxplot(df2)
```

lapply

lapply

`lapply` priima 3 argumentus: (1) list objektą, (2) funkciją arba funkcijos pavadinimą, (3) galimus funkcijos papidlomus argumentus

Jeigu `X` nėra list, tada R bando paversti `X` list objektu.

```
args(lapply)
## function (X, FUN, ...)
## NULL
```

lapply

`lapply` visad grąžina `list` klasės objektą

```
x <-  list(a=1:10, b=rnorm(10), c=seq(from=100, to=200, by=2))
lapply(x, mean)
## $a
## [1] 5.5
##
## $b
## [1] 0.07560161
##
## $c
## [1] 150
```

lapply

```
x <- 1:3
as.list(1:3) #taip lapply mato vektorių x konvertuus jį į list objektą
## [[1]]
## [1] 1
##
## [[2]]
## [1] 2
##
## [[3]]
## [1] 3
lapply(x, runif)
## [[1]]
## [1] 0.882948
##
## [[2]]
## [1] 0.9413468 0.2344180
##
## [[3]]
## [1] 0.9366658 0.5667020 0.8427904
```

lapply

Išnaudojant `...` galime perleisti papildomus argumentus `runif` funckijai:

```
x <- 1:3
lapply(x, runif, min=5, max=10)
## [[1]]
## [1] 9.106506
##
## [[2]]
## [1] 6.399864 5.236506
##
## [[3]]
## [1] 6.124703 8.365463 9.794654
```

lapply

`lapply` ir kitos apply funkcijos gali naudotis `USER DEFINED FUNCTION`, t.y. niekur kitur nedefinuotomis funkcijomis

```
x <- list(a=matrix(1:9, nrow=3, ncol = 3),
          b=matrix(1:4, nrow = 2, ncol=2))
lapply(x, function(elt) elt[,1, drop=FALSE]) #elt yra anoniminė funkcija
## $a
##      [,1]
## [1,]    1
## [2,]    2
## [3,]    3
##
## $b
##      [,1]
## [1,]    1
## [2,]    2
```

lapply

Klausimas: ką generuoja ši `lapply` funkcija?

```
A <- matrix(sample(1:10,9),3,3)
B <- matrix(sample(1:10,9),3,3)
C <- matrix(sample(1:10,9),3,3)
MAT_LIST <- list(A,B,C)

lapply(MAT_LIST, "[", ,1)
## [[1]]
## [1]  7 10  9
##
## [[2]]
## [1] 1 6 8
##
## [[3]]
## [1]  1 10  2
```

lapply

- Ar galima pritaikyti `lapply` funkciją `sales` dataframe?
- Jeigu taip, kokia objekto klasė bus grąžinta?
- Išbandykite

lapply

Solution:

```
## $Marry
## [1] 78
##
## $John
## [1] 86
##
## $Felix
## [1] 74
##
## $Jenny
## [1] 72
```

sapply

sapply

`sapply` bando supaprastinti `lapply` rezultatus (jeigu įmanoma)

- jeigu `lapply` grąžintų list, kurių kiekvienas elementas yra 1 ilgumo, tada `sapply` grąžina vektorių
- jeigu `lapply` grąžintų list, kurių kiekvienas elementas yra >1 ir vienodo ilgumo, tada `sapply` grąžina matricą
- jeigu netinka pirma du variantai, grąžina list

sapply

```
x <-  list(a=1:10, b=rnorm(10), c=seq(from=100, to=200, by=2))
lapply(x, mean)
## $a
## [1] 5.5
##
## $b
## [1] 0.2591564
##
## $c
## [1] 150
```

sapply

```
x <-  list(a=1:10, b=rnorm(10), c=seq(from=100, to=200, by=2))
sapply(x, mean)
##           a           b           c
##   5.5000000  -0.2375948 150.0000000
```

sapply

```
sapply(MAT_LIST,"[",1,1, simplify = TRUE)  # nepriima tuščių argumentų
## [1] 7 1 1
sapply(MAT_LIST,"[",1,1, simplify = FALSE) # nepriima tuščių argumentų
## [[1]]
## [1] 7
##
## [[2]]
## [1] 1
##
## [[3]]
## [1] 1
```

sapply

- Ar galima pritaikyti `sapply` funkciją `sales` dataframe?
- Jeigu taip, kokia objekto klasė bus grąžinta?
- Išbandykite

sapply

Solution:

```
## Marry  John Felix Jenny
##    78    86    74    72
```

mapply

mapply

- `mapply` taiko paraleliai (vienu metu) funckiją skirtingiems argumentams arba list arba vektoriams
- m - multivariate

```
str(mapply)
## function (FUN, ..., MoreArgs = NULL, SIMPLIFY = TRUE, USE.NAMES = TRUE)
```

- `FUN` yra funkcija, kuri bus taikoma
- `...` argumentai, kuriais naudojamasi funkcijoje
- `MoreArgs` kiti `FUN` argumentai
- `SIMPLIFY` ar rezultatas turėtų būti simplifikuotas kaip `sapply`

mapply

Jeigu norime sukurti tokį list objektą, 4 kartus rašome rep(), su argumentais 1-4 ir 4-1

```
list(rep(1,4), rep(2,3), rep(3,2), rep(4,1))
## [[1]]
## [1] 1 1 1 1
##
## [[2]]
## [1] 2 2 2
##
## [[3]]
## [1] 3 3
##
## [[4]]
## [1] 4
```

mapply

Supaprastinant galima naudoti `mapply` funkciją, kurios argumentai `rep` funckija ir du vektoriai 1:4 ir 4:1

```
mapply(rep, 1:4, 4:1)
## [[1]]
## [1] 1 1 1 1
##
## [[2]]
## [1] 2 2 2
##
## [[3]]
## [1] 3 3
##
## [[4]]
## [1] 4
```

mapply

Funkcija `noise` generuoja `n` atsitiktinių normaliojo skirstinio skaičių su vidurkiu `mean` ir standaritiniu nuokyrpiu `sd`

```
noise <- function(n, mean, sd){
        rnorm(n, mean, sd)
}

noise(4,1,2) # veikia kaip tikėtasi
## [1] 1.7826946 0.6749809 0.7936188 4.4184063
# list(noise(1,1,0.1),noise(2,2,0.1),noise(3,3,0.1),noise(4,4,0.1))
noise(1:4,1:4,0.01) # veikia ne kaip tikėtasi
## [1] 0.9925361 2.0074224 3.0118019 4.0049726
```

mapply

Šioje vietoje galima naudotis `mapply` tam kad funkcija primtų argumentus iš vektorių

```
noise <- function(n, mean, sd){
        rnorm(n, mean, sd)
}
# list(noise(1,1,0.1),noise(2,2,0.1),noise(3,3,0.1),noise(4,4,0.1))
mapply(noise, 1:4, 1:4, 0.1)
## [[1]]
## [1] 1.015169
##
## [[2]]
## [1] 2.318956 2.130100
##
## [[3]]
## [1] 3.024953 3.212102 3.065213
##
## [[4]]
## [1] 4.226947 4.091738 4.044441 4.071369
```

mapply

Exercise Generate following matrix using `mapply` (hint: use `rep()`)

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    2    3    4
## [2,]    1    2    3    4
## [3,]    1    2    3    4
## [4,]    1    2    3    4
```

rapply

rapply

`rapply` yra *recursive apply* ir taikoms `list` objektams

```
args(rapply)
## function (object, f, classes = "ANY", deflt = NULL, how = c("unlist",
##     "replace", "list"), ...)
## NULL

#pvz:
x <- list(1,2,3,4)
rapply(x, function(x){x^2})
## [1]  1  4  9 16
```

rapply

Exercise: * Calculate the bonus for each salesperson * Bonus =0.15 * average sales of the year (use `rapply` and either `lapply` or `sapply`)

```
## Marry  John Felix Jenny
## 0.975 1.075 0.925 0.900
```

rapply

Papildomi `rapply` pvz., savistudijoms
https://www.r-bloggers.com/rapply-function-explanation-and-examples

tapply

tapply

Apply a function to each cell of a ragged array, that is to each (non-empty) group of values given by a unique combination of the levels of certain factors.

```
str(tapply)
## function (X, INDEX, FUN = NULL, ..., default = NA, simplify = TRUE)
```

- `X` yra vektorius
- `INDEX` faktorius arba faktorių list
- `FUN` taikoma funkcija
- `...` papildomi `FUN` argumentai
- `simplify` ar supaprastinti rezultatus

tapply

```
x <- c(rnorm(10), runif(10), rnorm(10,1))
x
##  [1] -0.67383319 -2.38267820 -1.24680374 -0.82496053 -2.13228424
##  [6] -0.61403748 -0.20734092  1.28162501 -0.09783121  0.15755788
## [11]  0.22614243  0.98355758  0.09838715  0.87955170  0.23345948
## [16]  0.77239607  0.47185934  0.40884825  0.82322317  0.53497093
## [21]  0.97796802  1.36601687  1.08631503 -0.74613833  0.97936851
## [26]  0.84474013  0.10847413 -0.05660892  1.89043338  0.52272172
# Generate factors by specifying the pattern of their levels.
#gl(n, k, length = n*k, labels = seq_len(n), ordered = FALSE)
f <- gl(3,10)
f
##  [1] 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3
## Levels: 1 2 3
tapply(x, f, mean)
##          1          2          3
## -0.6740587  0.5432396  0.6973291
```

tapply

```
tapply(x, f, mean, simplify = FALSE)
## $`1`
## [1] -0.6740587
##
## $`2`
## [1] 0.5432396
##
## $`3`
## [1] 0.6973291
```

tapply

```
tapply(x, f, summary)
## $`1`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.3827 -1.1413 -0.6439 -0.6741 -0.1252  1.2816
##
## $`2`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.09839 0.27731 0.50342 0.54324 0.81052 0.98356
##
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.7461  0.2120  0.9114  0.6973  1.0596  1.8904
```

split

split

`split` padalina vektorių arba kitą objektą į grupes priklausomai nuo faktorių arba faktorių list

```
str(split)
## function (x, f, drop = FALSE, ...)
```

- `x` vektorius / list / dataframe
- `f` faktorius arba faktorių list
- `drop` indikuoja, ar tušti faktoriai turėtų būti panaikinti

split

```
split(x, f)
## $`1`
##  [1] -0.67383319 -2.38267820 -1.24680374 -0.82496053 -2.13228424
##  [6] -0.61403748 -0.20734092  1.28162501 -0.09783121  0.15755788
##
## $`2`
##  [1] 0.22614243 0.98355758 0.09838715 0.87955170 0.23345948 0.77239607
##  [7] 0.47185934 0.40884825 0.82322317 0.53497093
##
## $`3`
##  [1]  0.97796802  1.36601687  1.08631503 -0.74613833  0.97936851
##  [6]  0.84474013  0.10847413 -0.05660892  1.89043338  0.52272172
# dabar galima naudoti lapply / sapply
```

split

Taigi galime suskaidyti `x` į 3 `list` objektus ir tada kiekvienam atlikti `lapply` arba

```
lapply(split(x, f),mean)
## $`1`
## [1] -0.6740587
##
## $`2`
## [1] 0.5432396
##
## $`3`
## [1] 0.6973291
tapply(x, f, mean)
##          1          2          3
## -0.6740587  0.5432396  0.6973291
```

split

```
head(airquality)
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

split

```
s <- split(airquality, airquality$Month)
lapply(s, function(x) colMeans(x[,1:4]))
## $`5`
##    Ozone  Solar.R     Wind     Temp
##       NA       NA 11.62258 65.54839
##
## $`6`
##     Ozone   Solar.R      Wind      Temp
##        NA 190.16667  10.26667  79.10000
##
## $`7`
##      Ozone    Solar.R       Wind       Temp
##         NA 216.483871   8.941935  83.903226
##
## $`8`
##     Ozone   Solar.R      Wind      Temp
##        NA        NA  8.793548 83.967742
##
## $`9`
##    Ozone  Solar.R     Wind     Temp
```

split

```
s <- split(airquality, airquality$Month)
lapply(s, function(x) colMeans(x[,1:4], na.rm=TRUE))
## $`5`
##     Ozone   Solar.R      Wind      Temp
##  23.61538 181.29630  11.62258  65.54839
##
## $`6`
##     Ozone   Solar.R      Wind      Temp
##  29.44444 190.16667  10.26667  79.10000
##
## $`7`
##      Ozone    Solar.R       Wind       Temp
##  59.115385 216.483871   8.941935  83.903226
##
## $`8`
##      Ozone    Solar.R       Wind       Temp
##  59.961538 171.857143   8.793548  83.967742
##
## $`9`
##     Ozone   Solar.R      Wind      Temp
```

split

```
s <- split(airquality, airquality$Month)
sapply(s, function(x) colMeans(x[,1:4], na.rm=TRUE))
##                  5         6          7          8         9
## Ozone     23.61538  29.44444  59.115385  59.961538  31.44828
## Solar.R 181.29630 190.16667 216.483871 171.857143 167.43333
## Wind      11.62258  10.26667   8.941935   8.793548  10.18000
## Temp      65.54839  79.10000  83.903226  83.967742  76.90000
```

split

```
regionas <- as.factor(rep(c("Vilnius", "Kaunas", "Klaidpėda"), each=10 ))
lytis <- as.factor(c("M", "V"))
x <- data.frame(metai=rep(2013:2017),
                regionas=rep(c("Vilnius", "Kaunas", "Klaidpėda"), each=10 ),
                lytis=c("M", "V"),
                bvp=rep(runif(5,100,200),3),
                vartojimas=rnorm(30)
                )
s <- split(x, list(regionas, lytis))
```

split

```
head(x,15)
##    metai regionas lytis      bvp vartojimas
## 1   2013  Vilnius     M 152.1847  0.9810854
## 2   2014  Vilnius     V 120.3190 -0.6616053
## 3   2015  Vilnius     M 165.3160 -0.7724177
## 4   2016  Vilnius     V 127.7222 -2.0184735
## 5   2017  Vilnius     M 140.8674 -0.5335854
## 6   2013  Vilnius     V 152.1847  0.4347283
## 7   2014  Vilnius     M 120.3190 -0.7711673
## 8   2015  Vilnius     V 165.3160 -0.7539408
## 9   2016  Vilnius     M 127.7222 -0.2993578
## 10  2017  Vilnius     V 140.8674  1.6639664
## 11  2013   Kaunas     M 152.1847 -1.2443298
## 12  2014   Kaunas     V 120.3190 -0.7831344
## 13  2015   Kaunas     M 165.3160  0.2448306
## 14  2016   Kaunas     V 127.7222 -0.1438872
## 15  2017   Kaunas     M 140.8674 -1.6086314
```

split

```
head(s,2)
## $Kaunas.M
##    metai regionas lytis      bvp vartojimas
## 11  2013   Kaunas     M 152.1847 -1.2443298
## 13  2015   Kaunas     M 165.3160  0.2448306
## 15  2017   Kaunas     M 140.8674 -1.6086314
## 17  2014   Kaunas     M 120.3190 -1.8191317
## 19  2016   Kaunas     M 127.7222  1.8871394
##
## $Klaidpėda.M
##    metai  regionas lytis      bvp vartojimas
## 21  2013 Klaidpėda     M 152.1847 -0.3805995
## 23  2015 Klaidpėda     M 165.3160 -0.3380941
## 25  2017 Klaidpėda     M 140.8674  0.2175429
## 27  2014 Klaidpėda     M 120.3190 -0.2878594
## 29  2016 Klaidpėda     M 127.7222 -0.4700714
```

split

```
sapply(s, function(x) mean(x[,4]))
##    Kaunas.M Klaidpėda.M   Vilnius.M    Kaunas.V Klaidpėda.V   Vilnius.V
##    141.2819    141.2819    141.2819    141.2819    141.2819    141.2819
sapply(s, function(x) colMeans(x[,4:5]))
##                Kaunas.M Klaidpėda.M   Vilnius.M    Kaunas.V Klaidpėda.V
## bvp         141.2818817 141.2818817 141.2818817 141.2818817  141.281882
## vartojimas   -0.5080246  -0.2518163  -0.2790886   0.6597898   -0.284219
##               Vilnius.V
## bvp         141.281882
## vartojimas   -0.267065
```

aggregate

agregate

Splits the data into subsets, computes summary statistics for each, and returns the result in a convenient form.

```
args(aggregate)
## function (x, ...)
## NULL
?aggregate
```

- x - an R object.
- by- a list of grouping elements, each as long as the variables in the data frame x. The elements are coerced to factors before use.

agregate

```
aggregate(airquality,
          list(airquality$Month), #jeigu nebus list() mes klaid, nes tada b
          mean, na.rm=TRUE)
##   Group.1    Ozone  Solar.R      Wind     Temp Month  Day
## 1       5 23.61538 181.2963 11.622581 65.54839     5 16.0
## 2       6 29.44444 190.1667 10.266667 79.10000     6 15.5
## 3       7 59.11538 216.4839  8.941935 83.90323     7 16.0
## 4       8 59.96154 171.8571  8.793548 83.96774     8 16.0
## 5       9 31.44828 167.4333 10.180000 76.90000     9 15.5
# tik vienam kintamajam
aggregate(airquality$Ozone, list(airquality$Month), FUN=mean, na.rm=TRUE)
##   Group.1        x
## 1       5 23.61538
## 2       6 29.44444
## 3       7 59.11538
## 4       8 59.96154
## 5       9 31.44828
```

agregate

```
aggregate(iris,
          list(iris$Species),
          mean, na.rm=TRUE)
## Warning in mean.default(X[[i]], ...): argument is not numeric or logical
## returning NA

## Warning in mean.default(X[[i]], ...): argument is not numeric or logical
## returning NA

## Warning in mean.default(X[[i]], ...): argument is not numeric or logical
## returning NA
##      Group.1 Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1     setosa        5.006       3.428        1.462       0.246      NA
## 2 versicolor        5.936       2.770        4.260       1.326      NA
## 3  virginica        6.588       2.974        5.552       2.026      NA
```

Distribucijos

Distribucijos

Ne retai atliekant įvarius tyrimus ar skaičiuojant tikimybes statistikoje, reikės remtis tam tikrais skirstiniais. R gali generuoti įvarius skirstinius (*distributions*) `?distributions`

- dnorm
- dgamma
- beta
- dpois

ir t.t.

Distribucijos

Šioje dalyje aptarsime

- Binomial Distribution
- Poisson Distribution
- Continuous Uniform Distribution
- Exponential Distribution
- Normal Distribution

Distribucijos

Visos distribucijos galimos su 4 funkcijomis:

```
# ?dnorm

dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)
```

- d density
- p cumulative distribution
- q quantile function
- r random number generation

set.seed(...)

Tyrimuose naudojant sugeneruotus atsitiktinius skaičius iš tam tikro skirstinio, būtina naudoti `set.seed()`, tam, kad tyrimas būtų atkartojamas.

```
set.seed(1)
rnorm(n=5, mean=5, sd=2)
## [1] 3.747092 5.367287 3.328743 8.190562 5.659016
rnorm(n=5, mean=5, sd=2)
## [1] 3.359063 5.974858 6.476649 6.151563 4.389223
set.seed(1)
rnorm(n=5, mean=5, sd=2)
## [1] 3.747092 5.367287 3.328743 8.190562 5.659016
```

Binominis skirstinys

Binominis skirstinys

Dichotomine matavimų skale matuojamų požymių reikšmių skirstinys. Skirstinys yra diskretus ir apibūdinamas parametrais n ir p. Parametras $n \geq 0$ reiškia bandymų skaičių, o p – požymio tikimybę įgyti vieną iš dviejų galimų reikšmių.

Binominio skirstinio pasiskirstymo tankio funkcija (tikimybė gauti x reikmšę su n bandymų ir p tikimybės reikmše):

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ kur } x = 1, 2, 3, ..., n$$

Binominis skirstinys

Tarkime duomenų analizės teste yra 10 klausimų, kurių kiekvienas turi 4 galimus atsakymus, iš kurių tik vienas yra teisingas. Tarkime studentas atėjo visiškai nepasiruošęs ir visiškai atsitiktinai pasirinks atsakymus. Norint išlaikyti testą, reikia teisingai astakyti į ne mažiau kaip 5 klausimus. Kokia tikimybė, jog studentas neišlaikys testo?

- p = 1/4=0.25 ir (1-p)=1-0.25=0.75
- n = 10
- x=4

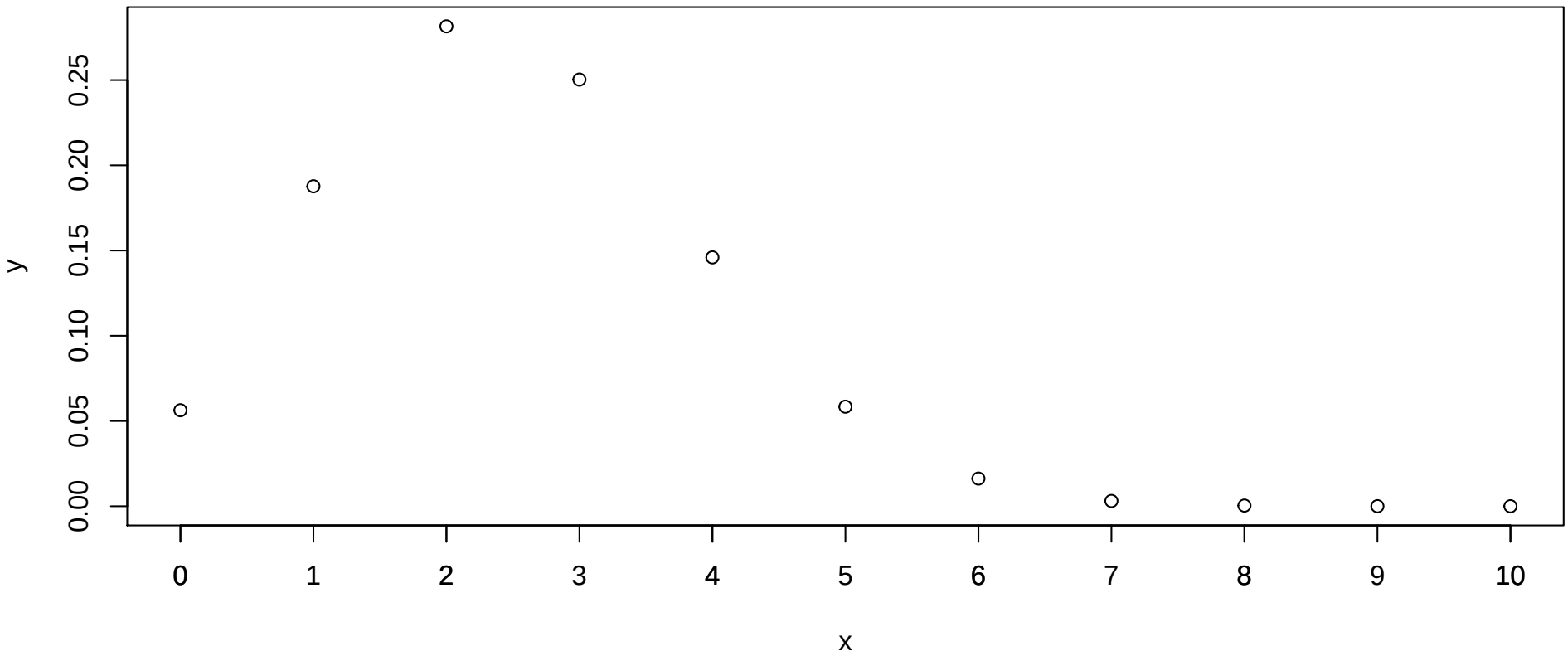
Binominis skirstinys

- p = 1/4=0.25 ir (1-p)=1-0.25=0.75
- n = 10
- x=4

```
# tikimybė jog studentas atsakys lygiai 4 teisingai
dbinom(x=4, size = 10, prob = 0.25)
## [1] 0.145998
```

Binominis skirstinys

```
x <- seq(from=0, to=10, by=1)
y <- dbinom(x, size=10, prob=0.25)
plot(x,y, type = "p")
axis(side = 1, at = x,labels = T)
```



Binominis skirstinys

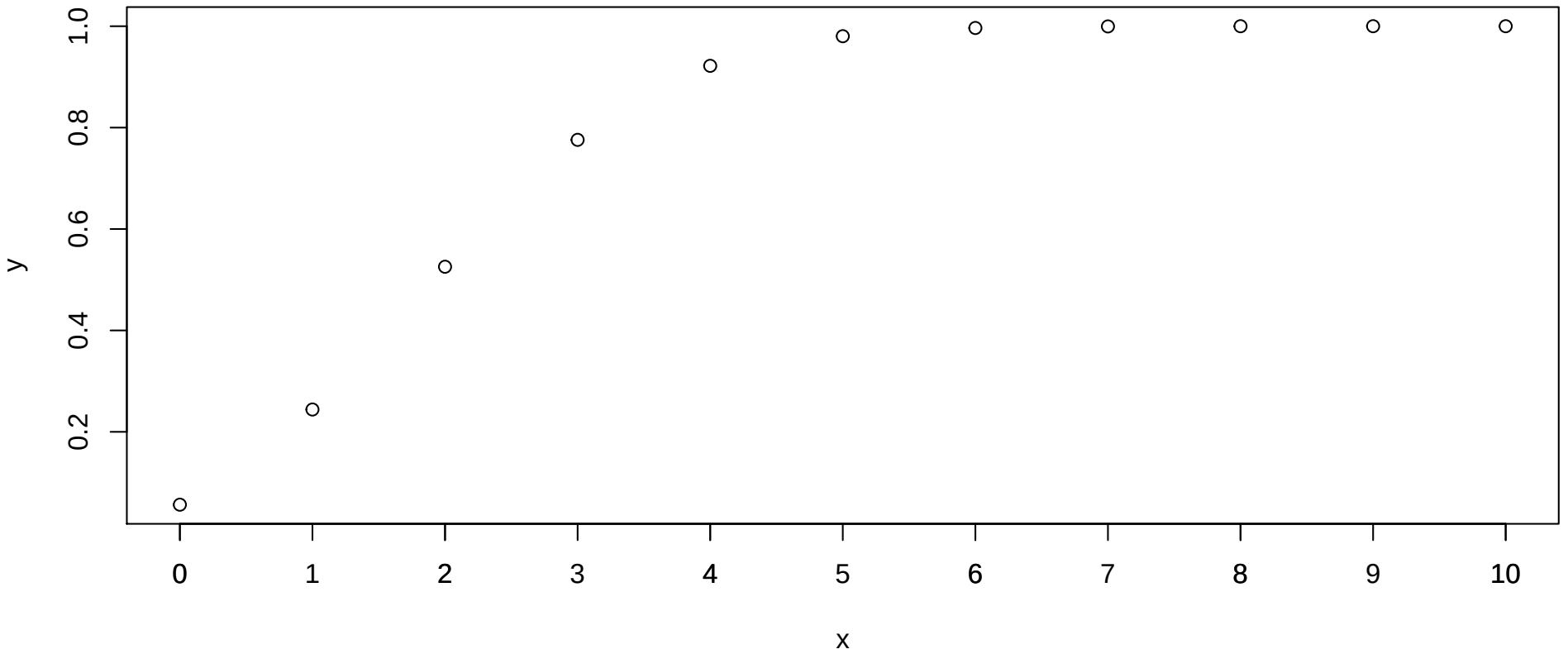
Tačiau norint žinoti visas vertes iki 4

```
# todėl norint žinoti tikimybę jog studentas atsakys į 4 arba mažiau
dbinom(x=0, size = 10, prob = 0.25)+
        dbinom(x=1, size = 10, prob = 0.25)+
        dbinom(x=2, size = 10, prob = 0.25)+
        dbinom(x=3, size = 10, prob = 0.25)+
        dbinom(x=4, size = 10, prob = 0.25)
## [1] 0.9218731

# alternatyviai galima pasinaudoti pbinom()
pbinom(q=4, size= 10, prob = 0.25, lower.tail = TRUE)
## [1] 0.9218731
# tačiau piktąjį dėstytoją domina,
# kokia tikimybė, jog studentas "praslys":
pbinom(q=4, size= 10, prob = 0.25, lower.tail = FALSE)
## [1] 0.07812691
```

Binominis skirstinys

```
x <- seq(from=0, to=10, by=1)
y <- pbinom(x, size=10, prob=0.25)
plot(x,y, type = "p")
axis(side = 1, at = x,labels = T)
```



Binominis skirstinys

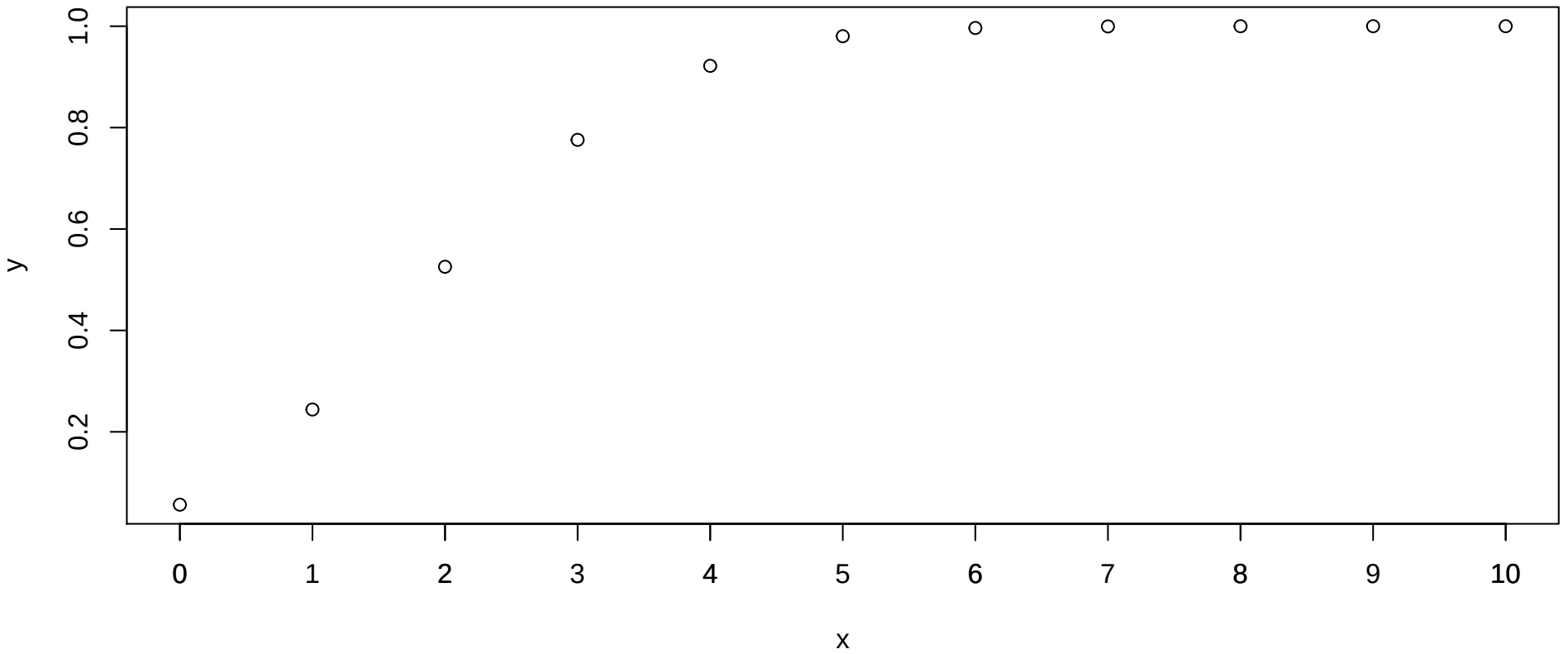
Tarkime dėstytojas nori nustatyti ribą, į kiek klausimų turi teisingai atsakyti studentai, kai:

- studentai turėdami 4 galimus pasirinkimus (daugiau alternatyvių atsakymų dėstytojas nenori sugalvoti, nes tingi)
- destyjas nenori, kad studentai praslystų pro testą didesne nei 10% tikimybe
- destytojas tingi galvoti daugiau nei 10 klausimų

```
qbinom(0.1, 10, 0.25, lower.tail = FALSE)
## [1] 4
```

Binominis skirstinys

```
x <- seq(from=0, to=10, by=1)
y <- pbinom(x, size=10, prob=0.25)
plot(x,y, type = "p")
axis(side = 1, at = x,labels = T)
```



Poisson skirstinys

Poisson skirstinys

Dichotomine matavimų skale matuojamų požymių reikšmių skirstinys. Skirstinys yra diskretus ir apibūdinamas parametrais n ir p. Parametras $n \geq 0$ reiškia bandymų skaičių, o p – požymio tikimybę įgyti vieną iš dviejų galimų reikšmių.

Poisson skirstinio pasiskirstymo tankio funkcija:

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!} \text{ kur } x = 1, 2, 3, ..., n$$

Poisson skirstinys

Poisson distribucija

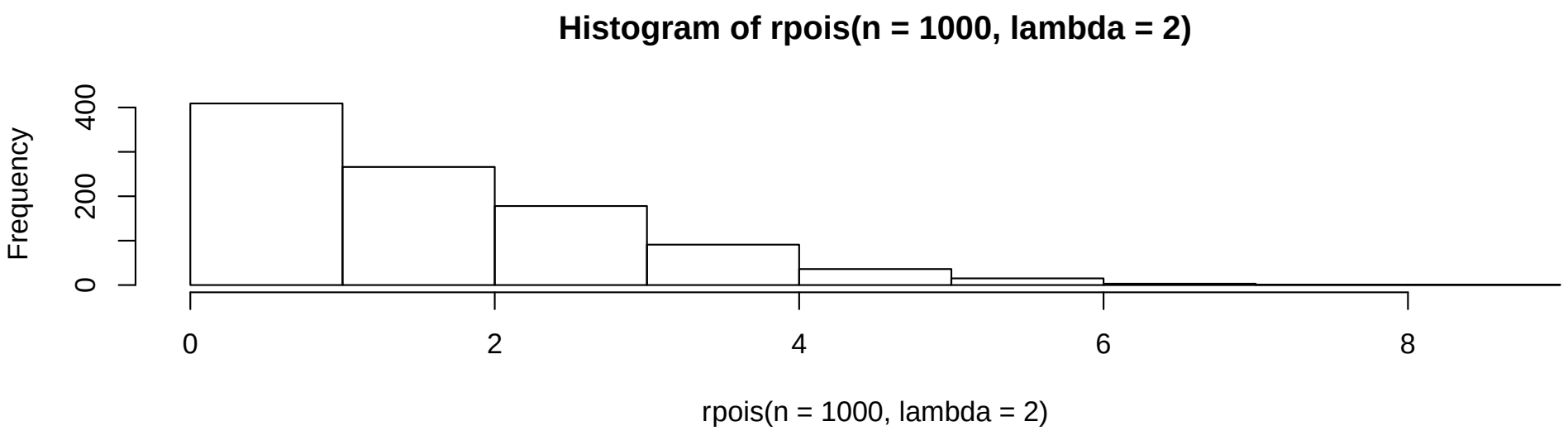
```
# ?dpois

dpois(x, lambda, log = FALSE)
ppois(q, lambda, lower.tail = TRUE, log.p = FALSE)
qpois(p, lambda, lower.tail = TRUE, log.p = FALSE)
rpois(n, lambda)
```

Poisson skirstinys

Poisson distribucija, kur λ yra vidutinė įvykio tikimybė per tam tikrą laikotarpį

```
rpois(n=10, lambda = 1)
##  [1] 0 0 1 1 2 1 1 4 1 2
rpois(n=10, lambda=2)
##  [1] 4 1 2 0 1 1 0 1 4 1
hist(rpois(n=1000, lambda=2))
```



Poisson skirstinys

Skambučių centras per valandą sulaukia 50 skambučių. *Maximum capacity* yra 65 skambučiai per valandą. Tada skambučiai nukreipiami į alternatyvų skambučių centrą, kuriame dirba beždžionėlės, tad klientai visad lieka nepatenkinti. Klausimas, kokia yra tikimybė, jog per sekančią valandą skambučių centras sulauks: 5, 30, 60 (arba mažiau skambučių):

```
dpois(5, 50) ## 5 Pr(x=5), lambda=50
## [1] 5.022786e-16
dpois(30 , 50) ## 30  Pr(x=30), lambda=50
## [1] 0.0006771985
dpois(60, 50) ## 60  Pr(x=50), lambda=50
## [1] 0.02010487

ppois(5, 50) ## 5 arba mažiau skambučių Pr(x<=5), lambda=50
## [1] 5.567756e-16
ppois(30, 50) ## 30 arba mažiau skambučių Pr(x<=30), lambda=50
## [1] 0.001594027
ppois(60, 50) ## 60 arba mažiau skambučių Pr(x<=50),lambda=50
## [1] 0.9278398
```

Poisson skirstinys

Kokia tikimybė, jog skambučių centras sulauks daugiau skaimbučių nei skabučių centro maksimalus aptarnavimo limitas? Jeigu įmonės išsikeltas tiklsas, jog nepatenkintų klientų būtų mažiau nei 0.1%, ar patartumėte vadovybei plėsti skambučių centro galimybes? Kiek papildomų darbuotojų reikia nusamdyti skambučiui centrui, jeigu 1 darbuotojas gali priimti po 5 skambučius per valandą?

```
ppois(q=65, lambda = 50, lower.tail = TRUE)
## [1] 0.9827354
ppois(q=65, lambda = 50, lower.tail = FALSE)
## [1] 0.01726457

# Kiek papildomų skambučių reiktų papildomai galėti priimti?
qpois(p=0.001, lambda = 50, lower.tail = FALSE)
## [1] 73
ceiling(
        (qpois(p=0.001, lambda = 50, lower.tail = FALSE) - 65) / 5
        )
## [1] 2
```

Tolygusis skirstinys (Continuous uniform distribution)

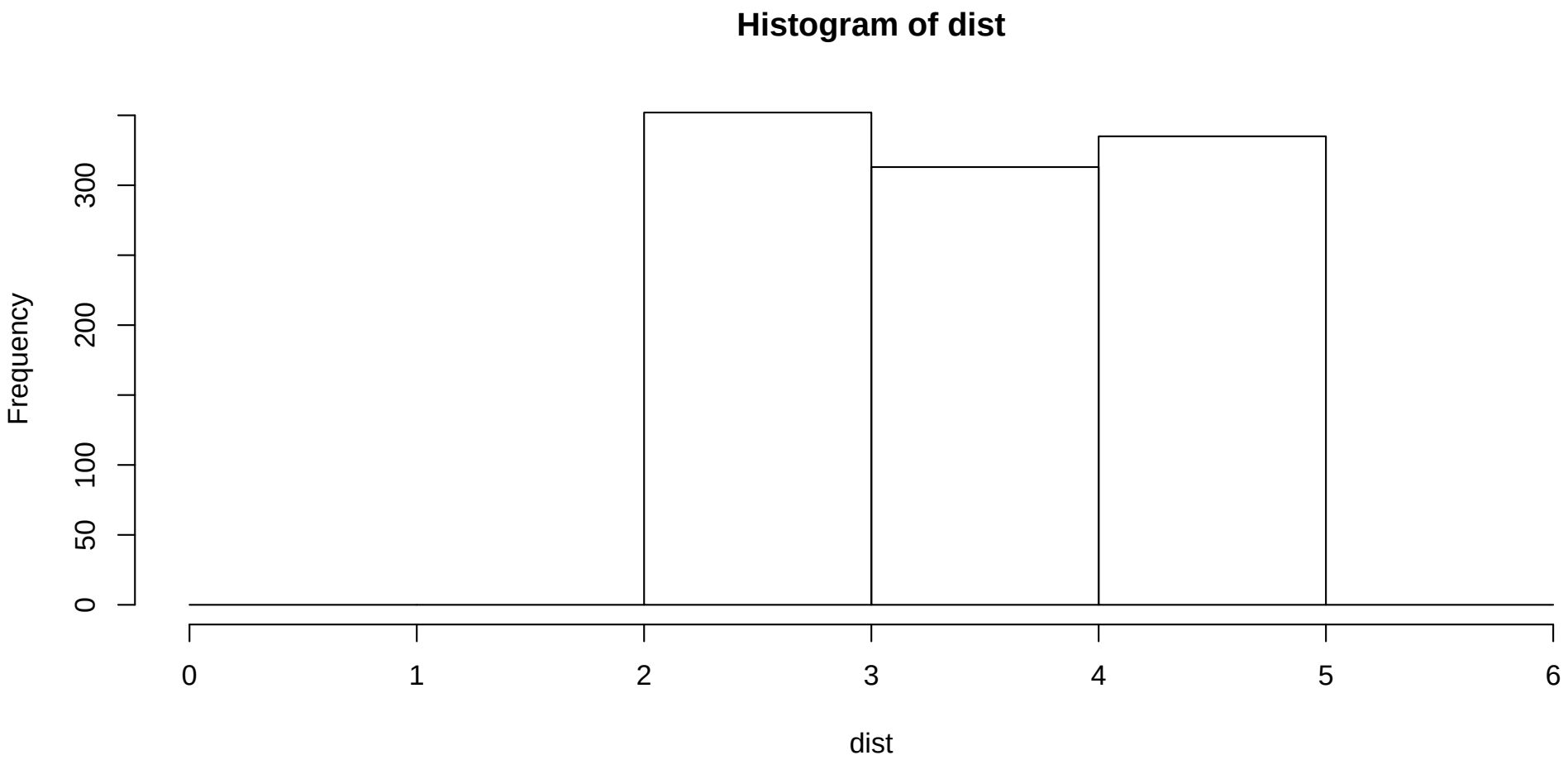
Tolygusis skirstinys (Continuous uniform distribution)

Skirstinys su vienoda tikimybe visiems skaičiams tarp a ir b. Visais kitais atvejais tikimybė =0.

$$f(x) = \begin{cases} \frac{1}{b-a} \text{ when } a \leq x \leq b \\ 0, else \end{cases}$$

Tolygusis skirstinys (Continuous uniform distribution)

```
dist <- runif(n=1000, min=2, max=5)
hist(dist, breaks = seq(from=0, to=6, by=1))
```



Eksponentinis skirstinys

Eksponentinis skirstinys

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

$$f(x, \mu) = \begin{cases} \frac{1}{\mu} e^{-x/\mu}, \text{ when } x \geq 0 \\ 0, x < 0 \end{cases}$$

Eksponentinis skirstinys

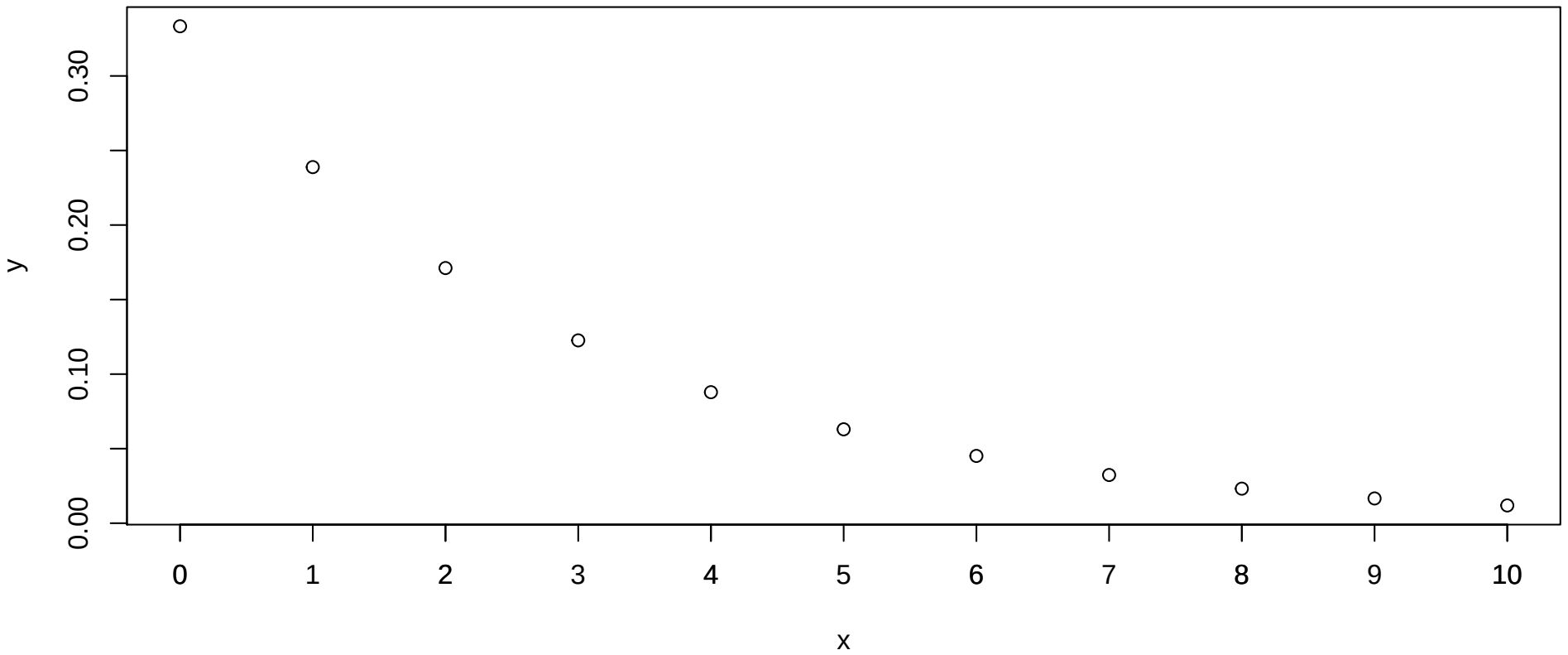
PVZ: Tarkime kasininkas aptarnauja vieną klientą per vidutiniškai 3 minutes. Žinoma, kad aptarnavimo laikas turi eksponentinį skirstinį. Kokia tikimybe sekantis klientas bus aptarnautas per mažiau nei 2 minutes

- vidutinis aptarnavimo greitis: 1/3=0.333 klientų per minutę

```
pexp(2, rate=1/3)
## [1] 0.4865829
```

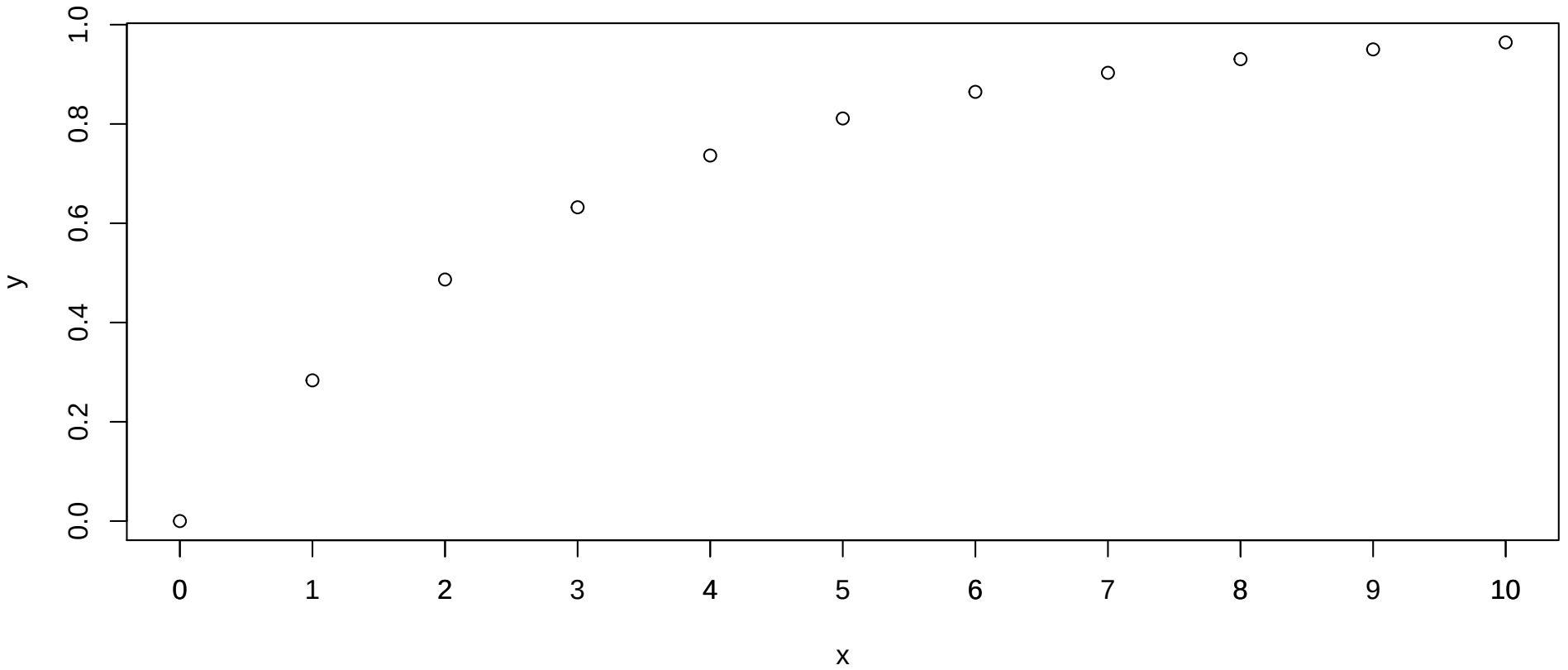
Eksponentinis skirstinys

```
x <- seq(from=0, to=10, by=1)
y <- dexp(x, rate=1/3 )
plot(x,y, type = "p")
axis(side = 1, at = x,labels = T)
```



Eksponentinis skirstinys

```
x <- seq(from=0, to=10, by=1)
y <- pexp(x, rate=1/3 )
plot(x,y, type = "p")
axis(side = 1, at = x,labels = T)
```



Normalusis skirstinys

Normalusis skirstinys

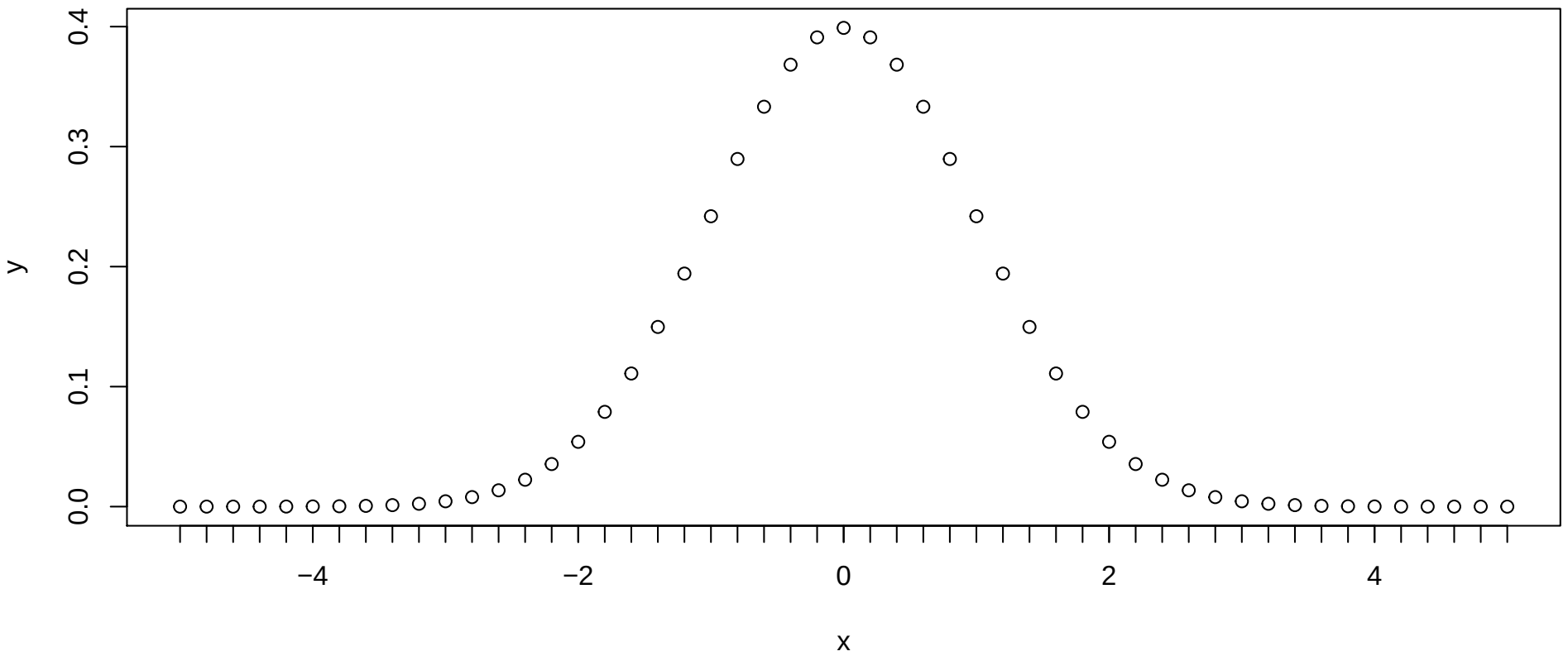
Sakysime, kad atsitiktinis dydis x turi normalųjį skirstinį, jei jo tankis

$$\varphi_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} \text{ for } -\infty < x < \infty; -\infty < \mu < \infty, \sigma^2 > 0$$

Sakysime, kad atsitiktinis dydis x turi standartinį normalųjį skirstinį, jeigu $\mu = 0, \sigma^2 = 1$

Normalusis skirstinys

```
x <- seq(from=-5, to=5, by=0.2)
y <- dnorm(x)
plot(x,y, type = "p")
axis(side = 1, at = x,labels = F)
```



Normalusis skirstinys

```
x <- seq(from=-5, to=5, by=0.2)
y <- pnorm(x)
plot(x,y, type = "p")
axis(side = 1, at = x,labels = F)
```