



گزارش پروژه دوم

دانیال سعیدی

بخش اول

۱) مطابق زیر فایل countries.csv را می خوانم: (فیلدهایی که خالی هستن NA گذاشتم که بعدا جایگزین کنم)

```
2 elements <- read.csv("./countries.csv", head = TRUE, ",")  
3  
4 df <- data.frame(elements)  
5 |  
6 df[df==""] <- NA
```

۲) برای حل مشکل NA بودن قسمتی از دیتا راه حل های زیادی وجود دارد. به طور مثال:

.i. حذف که دیتای ناقص

اگر بیش از ۵۰٪ اطلاعات در دسترس باشد حذف آن روشی بهینه نیست.

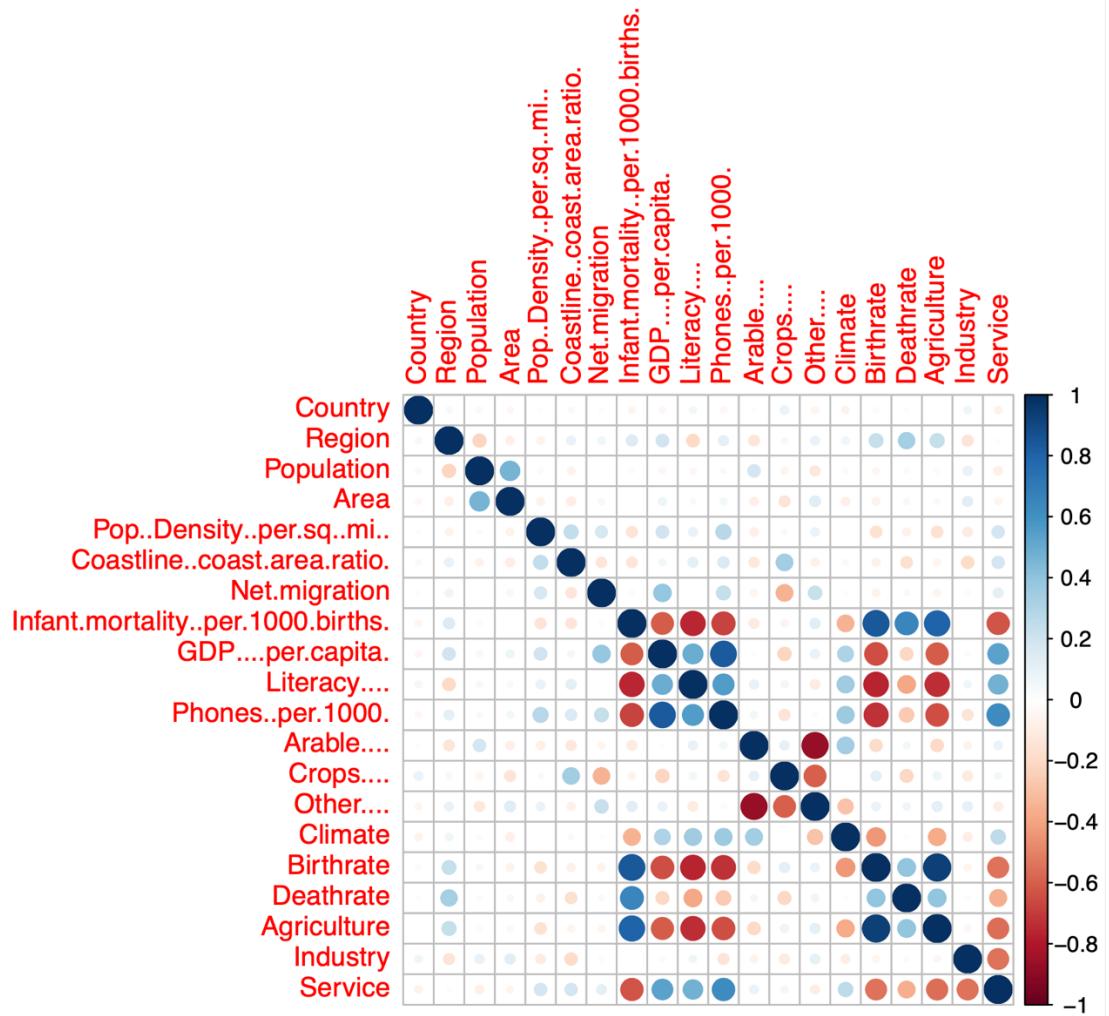
.ii. قرار دادن Mean/Median/Mode هر ستون بجای مقادیر NA

حساب کردن مقادیر میانگین و میانه و مد و قرار دادن آن به جای اطلاعات گم شده می تواند راه حلی سریع باشد. اما یکی از بدی هایی که می توان برای این روش اشاره کرد، کاهش پراکندگی اطلاعات(Variance) است.

.iii. استفاده از Linear Regression

با استفاده از ماتریس همبستگی(Correlation Matrix) بهترین پیش بینی کننده یک مقدار انتخاب و مقدار گم شده را با آن پیش بینی می کنیم.
به طور مثال مقدار Agriculture یک کشور مشخص نیست. طبق نمودار رسم شده در صفحه بعد، مشاهده می شود که ضریب همبستگی Birthrate و Agriculture بسیار نزدیک به یک است. یعنی Agriculture و Birthrate نسبت به هم رفتار خطی دارند.

حال می توان با روش های ریاضی یک رابطه خطی بین این دو متغیر به دست آورد. و اگر در بخشی از اطلاعات که یکی از این دو معلوم بودند، میشود دیگری را حدس زد. البته با توجه داشت که استفاده از این روش، می تواند این توهمندی را به ما بدهد که داده ها خطی اند.



منبع:

<https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4>

۳) مقادیر ناقص را با میانگین هر سطر جایگزین می کنیم تا مشکل NA بودن رفع شود:

```

7 #replacing NA values with mean of their column
8 for(i in 1:ncol(df))
9 {
10 | df[is.na(df[,i]), i] <- mean(df[,i], na.rm = TRUE)
11 }
```

سپس اطلاعات dataframe را به یک ماتریس تبدیل می کنیم و به تابع cor که ضریب همبستگی را محاسبه می کند، ورودی می دهیم:

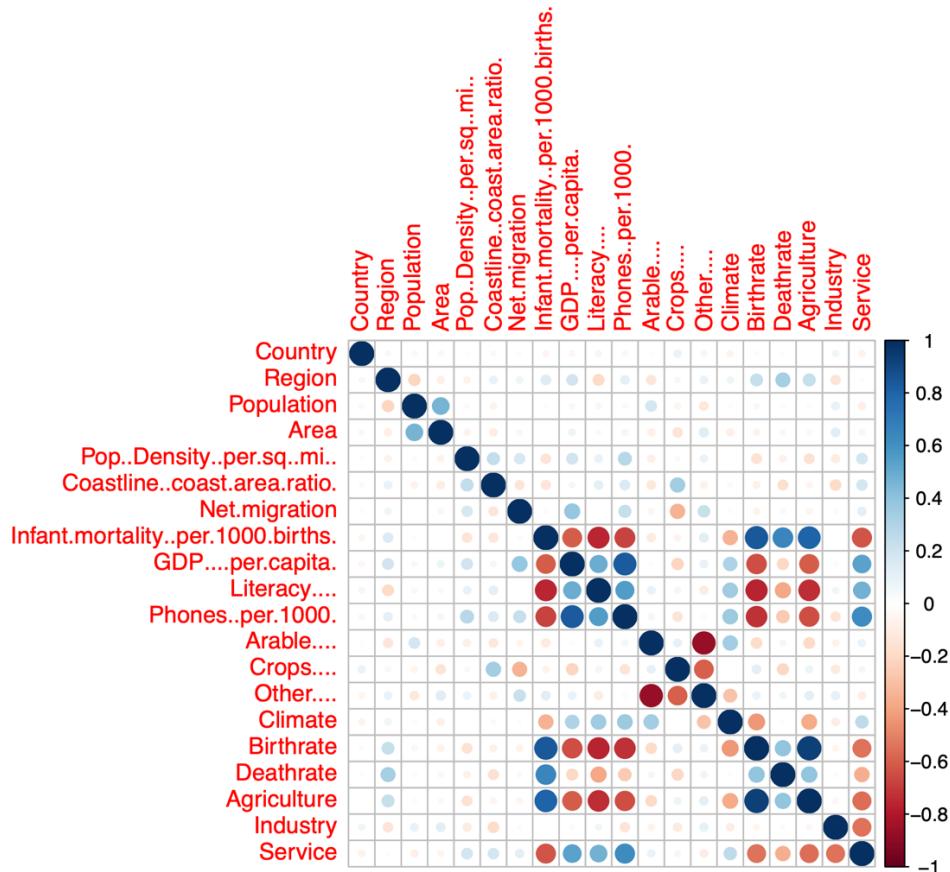
```

12 matrix <- data.matrix(df);
13 correlation_matrix <- cor(matrix);
```

با استفاده از لایبری corrplot ماتریس همبستگی را رسم کردم:

```
16 corrplot(correlation_matrix)
```

که خروجی آن: (عکس واضح تر در part1.pdf قرار داده شده است.)



خروجی آن فضای زیادی میخواهد و در terminal به صورت زیر نمایش داده میشود:

```

CA2 -- zsh -- 118x31
~/Documents/GitHub/EngineeringProbabilityStatistics/CA2 -- R

3: Setting LC_MESSAGES failed, using "C"
4: Setting LC_MONETARY failed, using "C" read.csv("./countries.csv", head = TRUE, ",")
corrrplot 0.84 loaded
Warning messages:
1: In mean.default(df[, i], na.rm = TRUE) :
  argument is not numeric or logical: returning NA
2: In mean.default(df[, i], na.rm = TRUE) :
  argument is not numeric or logical: returning NA
  df <- data.frame(elements)
Country           1.00000000  0.04290017 -3.994767e-02
Region            0.042900171 1.00000000 -2.111468e-01
Population         0.039947675 -0.21114677 1.000000e+00
Area              0.042163369 -0.08247139  4.696981e-01
Pop..Density..per.sq..mi.. 0.019513858 -0.06901122 -2.865918e-02
Coastline..coast.area.ratio. 0.033397148  0.09313424 -6.816810e-02
Net.migration      0.019759813  0.06951539  1.087094e-05
Infant.mortality..per.1000.births. -0.052510695  0.14915959  2.300002e-02
GDP....per.capita. 0.048962141  0.19052297 -4.001134e-02 l = "blue"
Literacy....       0.051854518 -0.19378426 -4.573171e-02
Phones..per.1000.  -0.045534389  0.11554756 -3.140330e-02
Arable....        0.010904273 -0.13962244  1.870149e-01
Crops....         0.088045553  0.04314646 -5.949969e-02
Other....          0.053985967  0.09074419 -1.206485e-01
Climate            -0.061142409  0.06403621 -2.880542e-02
Birthrate          0.028342478  0.23262311 -4.508733e-02
Deathrate          -0.019445219  0.33015175 -2.836095e-02
Agriculture        -0.008915703  0.23926231 -2.933534e-02
Industry           0.062644102 -0.14312219  9.913460e-02
Service            -0.063869168  0.01426257 -8.208339e-02
                           Area Pop..Density..per.sq..mi...
Country           -0.04216337   0.019513858

```

```

12 matrix <- data$Area Pop..Density..per.sq..mi...
13 correl<-0.04216337 <- cor(matrix); 0.019513858
14 correl<-0.08247139                  -0.069011220
15               0.46969809                  -0.028659177
16 corrrplot(matrix, correlation_matrix) -0.067428997
17 plot(d = 1.00000000, df$Birthrate, ylab = "Population", xlab = "Agriculture vs Birthrate",
18      x=-0.06742900, y=0.09313424, l = "blue")
19 coastline.coast.area.ratio. <- 0.09540818
20 Net.migration <- 0.09540818
21 lines( 0.04745779, df$Birthrate, ylab = "Population", xlab = "Agriculture vs Birthrate",
22      x=-0.06742900, y=0.09313424, l = "blue")
23 Infant.mortality..per.1000.births. <- 0.00716167
24 GDP....per.capita. <- 0.07137038
25 Literacy.... <- 0.03327233
26 Phones..per.1000. <- 0.05262450
27 Arable.... <- 0.08190439
28 Crops.... <- 0.14268408
29 Other.... <- 0.13939111
30 Climate <- 0.08400734
31 Birthrate <- 0.06641087
32 Deathrate <- 0.04006184
33 Agriculture <- 0.03731443
34 Industry <- 0.12094677
35 Service <- 0.05516121

```

```

Country
Region
Population
Area
Pop..Density..per.sq.mi...
Coastline..coast.area.ratio.
Net.migration
Infant.mortality..per.1000.births.
GDP....per.capita.
Literacy....
Phones..per.1000.
Arable....
Crops....
Other....
Climate
Birthrate
Deathrate
Agriculture
Industry
Service
  (base) daniel@Daniels-MacBook-Pro CA2 %

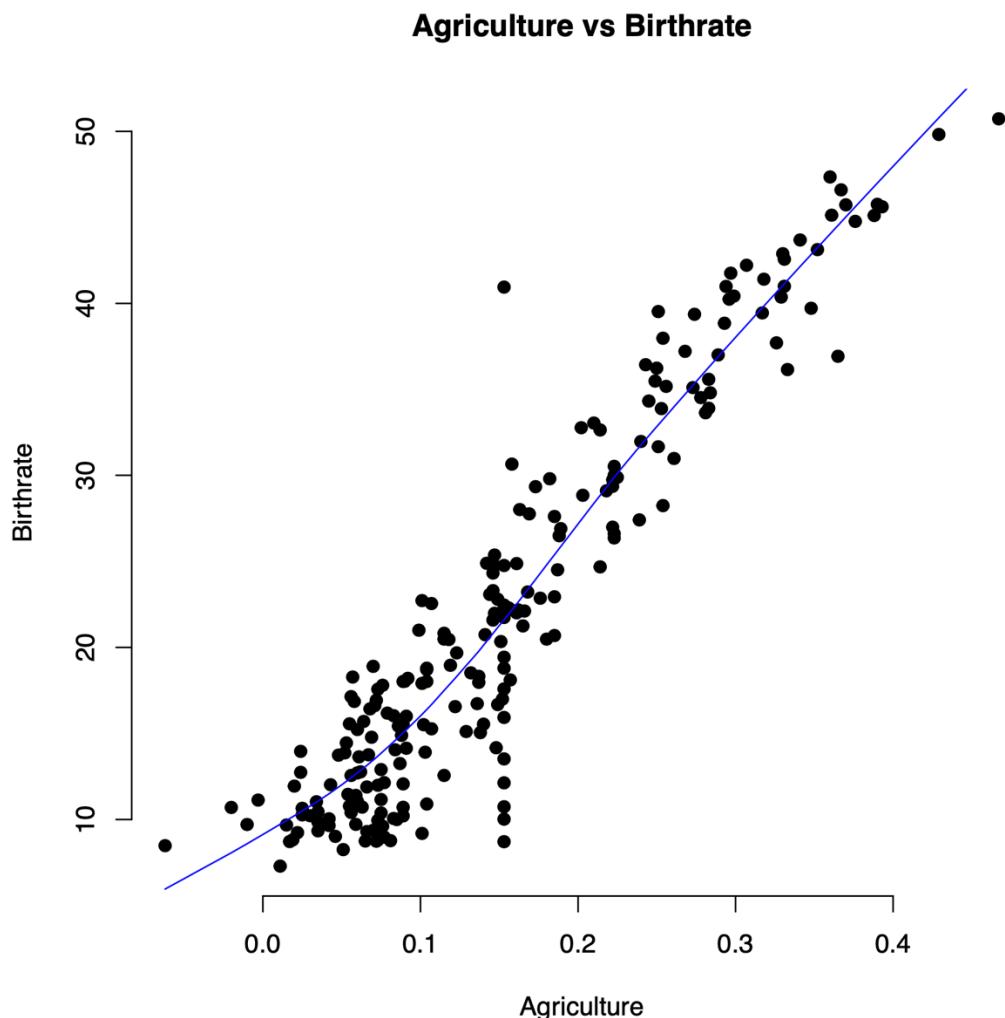
Country
Region
Population
Area
Pop..Density..per.sq.mi...
Coastline..coast.area.ratio.
Net.migration
Infant.mortality..per.1000.births.
GDP....per.capita.
Literacy....
Phones..per.1000.
Arable....
Crops....
Other....
Climate
Birthrate
Deathrate
Agriculture
Industry
Service
  (base) daniel@Daniels-MacBook-Pro CA2 %

Country
Region
Population
Area
Pop..Density..per.sq.mi...
Coastline..coast.area.ratio.
Net.migration
Infant.mortality..per.1000.births.
GDP....per.capita.
Literacy....
Phones..per.1000.
Arable....
Crops....
Other....
Climate
Birthrate
Deathrate
Agriculture
Industry
Service
  (base) daniel@Daniels-MacBook-Pro CA2 %

```


(۴) با استفاده از تابع plot نمودار scatter plot را رسم کردم. و برای نشان دادن رابطه خطی داشتن کشاورزی و نرخ تولد خط های آبی را رسم کردم.

```
17 plot(df$Agriculture, df$Birthrate, main = "Agriculture vs Birthrate"  
18     xlab = "Agriculture", ylab = "Birthrate",  
19     pch = 19, frame = FALSE)  
20 lines(lowess(df$Agriculture, df$Birthrate), col = "blue")  
21
```



(۵) با توجه به ماتریس همبستگی که به دست آوردهیم، مقادیر گم شده را با متغیر دیگری که ضریب همبستگی بسیار نزدیک به ۱ یا -۱ دارد، از طریق روش Linear Regression محاسبه می کنیم.

به طور مثال در فایل countries.csv مقدار Agriculture برحی از کشورها مشخص نیست. طبق نمودار رسم شده در صفحه بعد، مشاهده می شود که ضریب همبستگی Agriculture و Birthrate بسیار نزدیک به یک است. یعنی ضریب همبستگی Birthrate و Agriculture نسبت به هم رفتار خطی دارند. حال می توان با روش های ریاضی یک رابطه خطی بین این دو متغیر به دست آورد. و اگر در بخشی از اطلاعات که یکی از این دو معلوم بودند، میشود دیگری را حدس زد.

با استفاده تابع lm مقدار شبیب و عرض از مبدا خط پیش بینی کننده را محاسبه می کنیم:

```
22 relation <- lm(Agriculture~Birthrate,df)
23 relation
```

که مشاهده می شود که شبیب خط ۰.۰۰۸۵ و عرض از مبدا -۰.۰۳۶ می باشد:

```
Call:
lm(formula = Agriculture ~ Birthrate, data = df)

Coefficients:
(Intercept)    Birthrate
-0.036378      0.008565
```

بخش دوم

(۱) مطابق زیر متغیر تصادفی با توزیع یکنواخت استاندارد ساختم:

```
9      #U~Uniform(0,1)
10     U <- runif(1, min = 0, max = 1)|  
p=0.6 (۲)
```

```
11    #X~Ber(p)
12    X <- (U < p)
```

(۳) فرض کنیم Y متغیر تصادفی با توزیع دو جمله‌ای باشد. میتوان Y را برحسب مجموع متغیرهای برنولی قسمت ۲ نوشت. یعنی:

$$Y = X_1 + X_2 + X_3 + \dots + X_n$$

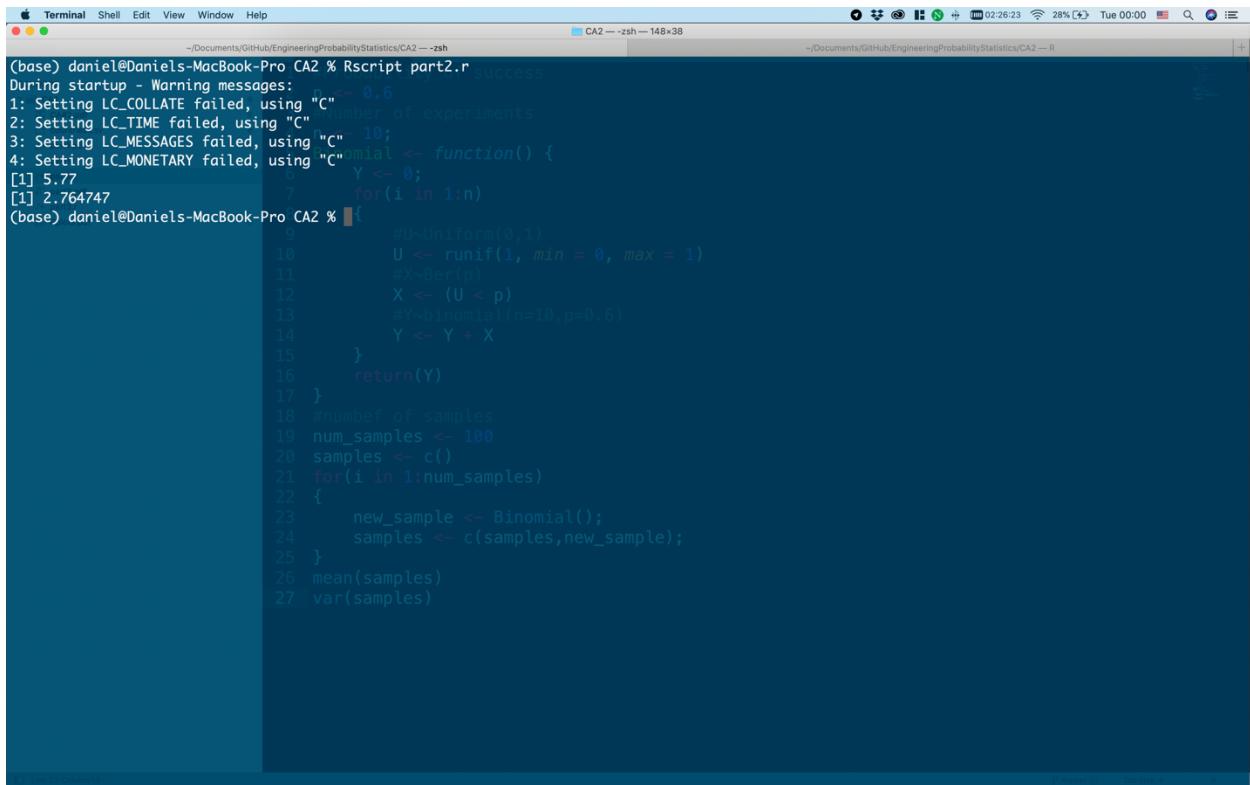
یک تابعی نوشتیم که خروجی آن متغیر تصادفی Y است:

```
1 #Probability of success
2 p <- 0.6
3 #Number of experiments
4 n <- 10;
5 Binomial <- function() {
6   Y <- 0;
7   for(i in 1:n)
8   {
9     #U~Uniform(0,1)
10    U <- runif(1, min = 0, max = 1)
11    #X~Ber(p)
12    X <- (U < p)
13    #Y~binomial(n=10,p=0.6)
14    Y <- Y + X
15  }
16  return(Y)
17 }
```

سپس ۱۰۰ نمونه از این متغیر را در وکتور samples ذخیره می‌کنم:

```
18 #number of samples
19 num_samples <- 100
20 samples <- c()
21 for(i in 1:num_samples)
22 {
23   new_sample <- Binomial();
24   samples <- c(samples,new_sample);
25 }
26 mean(samples)
27 var(samples)
```

که خروجی آن:



The screenshot shows a macOS Terminal window with two tabs open. The left tab displays the R script code for generating samples from a binomial distribution. The right tab shows the results of running the script, including warning messages about locale settings and numerical values for mean and variance.

```
(base) daniel@Daniels-MacBook-Pro CA2 % Rscript part2.r
During startup - Warning messages:
1: Setting LC_COLLATE failed, using "C" [+]
2: Setting LC_TIME failed, using "C" 02:26:23
3: Setting LC_MESSAGES failed, using "C" 28% [4]
4: Setting LC_MONETARY failed, using "C" Tue 00:00
[1] 5.77 USA
[1] 2.764747 R
(base) daniel@Daniels-MacBook-Pro CA2 % 
 9      #U-Uniform(0,1)
10     U <- runif(1, min = 0, max = 1)
11     #X-Ber(p)
12     X <- (U < p)
13     #Y-binomial(n=10,p=0.6)
14     Y <- Y + X
15   }
16   return(Y)
17 }
18 #number of samples
19 num_samples <- 100
20 samples <- c()
21 for(i in 1:num_samples)
22 {
23   new_sample <- Binomial();
24   samples <- c(samples,new_sample);
25 }
26 mean(samples)
27 var(samples)
```

مقدار واریانس 2.76 و میانگین 5.77 می باشد. که این مقدار با افزایش نمونه ها، دقیق تر میشود.

مقدار میانگین و واریانس متغیر Y با جایگذاری در فرمول ۶ و ۲.۴ می باشد.

بخش سوم

کد مربوط به بخش سوم:

```
1 #inverse transform sampling
2 #number of samples
3 samples <- 1000
4 U      <- runif(samples)
5 X      <- -log(U)/2
6 #plot
7 hist(X, freq=F, xlab='X', main='(Frequency distribution)')
8 curve(dexp(x, rate=2) , 0, 3, lwd=2, xlab = "", ylab = "", add = T)
```

نمودار توزیع فراوانی:

