Course Code 009959 (Spring 2018)

Probability and Statistics with Programming

Probability Distributions in R

Prof. S. M. Riazul Islam, Dept. of Computer Engineering, Sejong University, Korea

E-mail: riaz@sejong.ac.kr

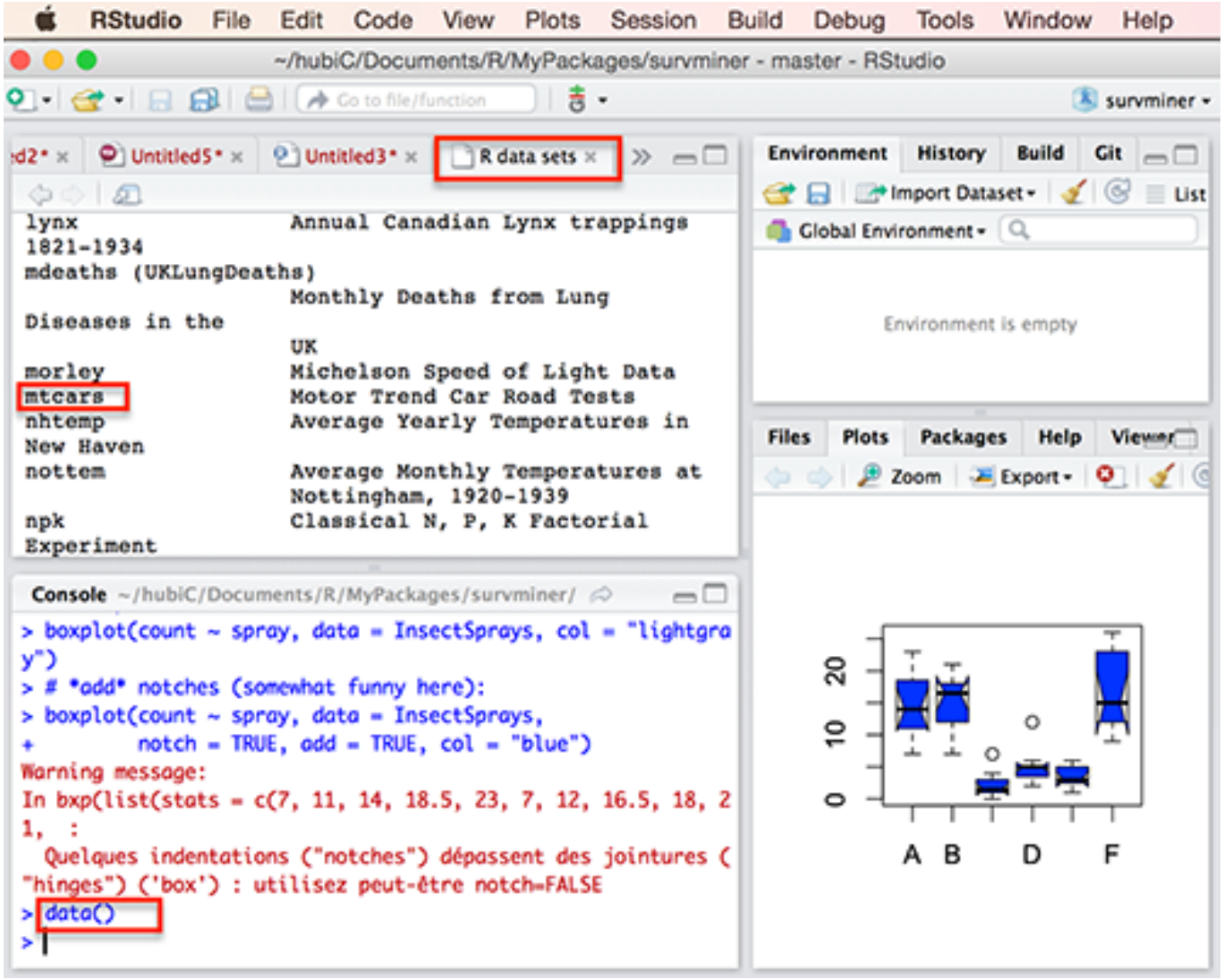
Probability Distributions in R

- Built-in Data sets
- Making Histogram
- Numerical Measures: Mean and Variance
- Probability Distributions

Probability Distributions in R

Built-in Datasets in R

>data()



Probability Distributions in R

mtcars: Motor Trend Car Road Tests

- View the content of *mtcars* data set:

```
# 1. Loading
data("mtcars")
# 2. Print
head(mtcars)
```

- It contains 32 observations and 11 variables:

```
# Number of rows (observations)
nrow(mtcars)
```

```
# Number of columns (variables)
ncol(mtcars)
```

```
[1] 11
```

- Description of variables:

1. mpg: Miles/(US) gallon
2. cyl: Number of cylinders
3. disp: Displacement (cu.in.)
4. hp: Gross horsepower
5. drat: Rear axle ratio
6. wt: Weight (1000 lbs)
7. qsec: 1/4 mile time
8. vs: V/S
9. am: Transmission (0 = automatic, 1 = manual)
10. gear: Number of forward gears
11. carb: Number of carburetors

Probability Distributions in R

Faithful: Old Faithful Geyser Data

```
> data("faithful")
> head(faithful)
  eruptions waiting
1     3.600      79
2     1.800      54
3     3.333      74
4     2.283      62
5     4.533      85
6     2.883      55
> 
```

```
> faithful$eruptions
  [1] 3.600 1.800 3.333 2.283 4.533 2.883 4.700 3.600 1.950 4.350 1.833 3.917 4.200 1.750 4.700 2.167 1.750 4.800 1.600
 [20] 4.250 1.800 1.750 3.450 3.067 4.533 3.600 1.967 4.083 3.850 4.433 4.300 4.467 3.367 4.033 3.833 2.017 1.867 4.833
 [39] 1.833 4.783 4.350 1.883 4.567 1.750 4.533 3.317 3.833 2.100 4.633 2.000 4.800 4.716 1.833 4.833 1.733 4.883 3.717
 [58] 1.667 4.567 4.317 2.233 4.500 1.750 4.800 1.817 4.400 4.167 4.700 2.067 4.700 4.033 1.967 4.500 4.000 1.983 5.067
 [77] 2.017 4.567 3.883 3.600 4.133 4.333 4.100 2.633 4.067 4.933 3.950 4.517 2.167 4.000 2.200 4.333 1.867 4.817 1.833
 [96] 4.300 4.667 3.750 1.867 4.900 2.483 4.367 2.100 4.500 4.050 1.867 4.700 1.783 4.850 3.683 4.733 2.300 4.900 4.417
[115] 1.700 4.633 2.317 4.600 1.817 4.417 2.617 4.067 4.250 1.967 4.600 3.767 1.917 4.500 2.267 4.650 1.867 4.167 2.800
[134] 4.333 1.833 4.383 1.883 4.933 2.033 3.733 4.233 2.233 4.533 4.817 4.333 1.983 4.633 2.017 5.100 1.800 5.033 4.000
[153] 2.400 4.600 3.567 4.000 4.500 4.083 1.800 3.967 2.200 4.150 2.000 3.833 3.500 4.583 2.367 5.000 1.933 4.617 1.917
[172] 2.083 4.583 3.333 4.167 4.333 4.500 2.417 4.000 4.167 1.883 4.583 4.250 3.767 2.033 4.433 4.083 1.833 4.417 2.183
[191] 4.800 1.833 4.800 4.100 3.966 4.233 3.500 4.366 2.250 4.667 2.100 4.350 4.133 1.867 4.600 1.783 4.367 3.850 1.933
[210] 4.500 2.383 4.700 1.867 3.833 3.417 4.233 2.400 4.800 2.000 4.150 1.867 4.267 1.750 4.483 4.000 4.117 4.083 4.267
[229] 3.917 4.550 4.083 2.417 4.183 2.217 4.450 1.883 1.850 4.283 3.950 2.333 4.150 2.350 4.933 2.900 4.583 3.833 2.083
[248] 4.367 2.133 4.350 2.200 4.450 3.567 4.500 4.150 3.817 3.917 4.450 2.000 4.283 4.767 4.533 1.850 4.250 1.983 2.250
[267] 4.750 4.117 2.150 4.417 1.817 4.467
> 
```

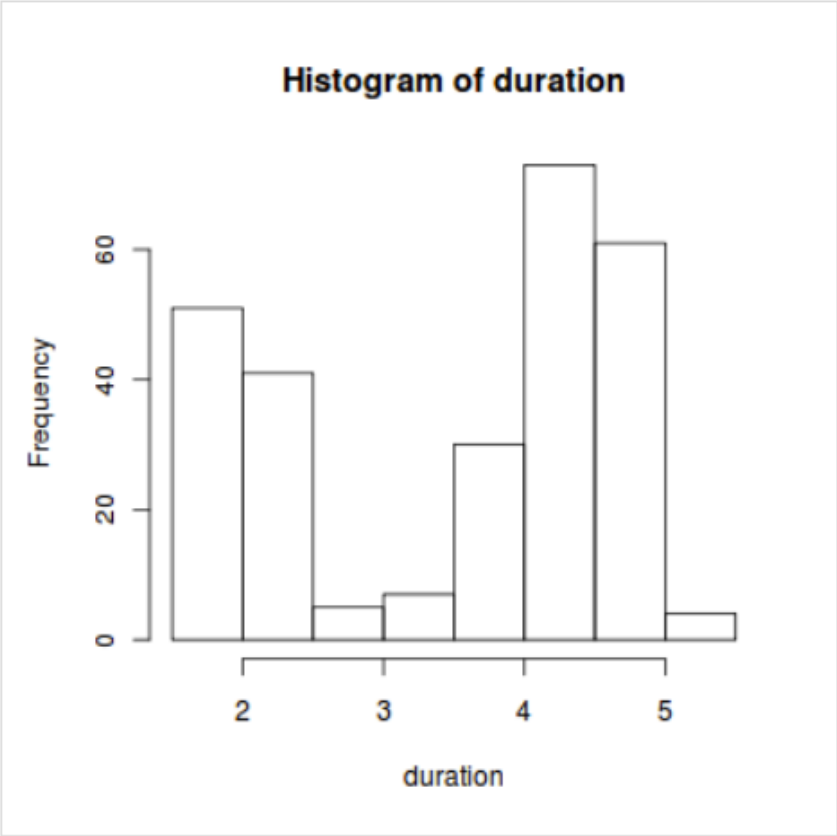
Probability Distributions in R

Histogram: Faithful Eruptions

```
> duration = faithful$eruptions
> hist(duration,    # apply the hist function
+   right=FALSE)    # intervals closed on the left
```

Answer

The histogram of the eruption durations is:



Probability Distributions in R

Faithful: Old Faithful Geyser Data

Problem

Find the mean eruption duration in the data set faithful.

Solution

We apply the mean function to compute the mean value of eruptions.

```
> duration = faithful$eruptions      # the eruption durations
> mean(duration)                     # apply the mean function
[1] 3.4878
```

Probability Distributions in R

Faithful: Old Faithful Geyser Data

Problem

Find the variance of the eruption duration in the data set faithful.

Solution

We apply the var function to compute the variance of eruptions.

```
> duration = faithful$eruptions    # the eruption durations
> var(duration)                    # apply the var function
[1] 1.3027
```

Probability Distributions in R

Faithful: Old Faithful Geyser Data

Problem

Find the standard deviation of the eruption duration in the data set faithful.

Solution

We apply the sd function to compute the standard deviation of eruptions.

```
> duration = faithful$eruptions    # the eruption durations
> sd(duration)                     # apply the sd function
[1] 1.1414
```

Probability Distributions in R

Faithful: Old Faithful Geyser Data

Problem

Find the covariance of eruption duration and waiting time in the data set faithful. Observe if there is any linear relationship between the two variables.

Solution

We apply the cov function to compute the covariance of eruptions and waiting.

```
> duration = faithful$eruptions   # eruption durations
> waiting = faithful$waiting      # the waiting period
> cov(duration, waiting)          # apply the cov function
[1] 13.978
```

Probability Distributions in R

Faithful: Old Faithful Geyser Data

Problem

Find the correlation coefficient of eruption duration and waiting time in the data set faithful. Observe if there is any linear relationship between the variables.

Solution

We apply the cor function to compute the correlation coefficient of eruptions and waiting.

```
> duration = faithful$eruptions    # eruption durations
> waiting = faithful$waiting       # the waiting period
> cor(duration, waiting)           # apply the cor function
[1] 0.90081
```

Probability Distributions in R

Binomial Distribution

Problem

Suppose there are twelve multiple choice questions in an English class quiz. Each question has five possible answers, and only one of them is correct. Find the probability of having four or less correct answers if a student attempts to answer every question at random.

```
> dbinom(4, size=12, prob=0.2)
[1] 0.1329
```

Probability Distributions in R

Binomial Distribution

```
> dbinom(0, size=12, prob=0.2) +
+ dbinom(1, size=12, prob=0.2) +
+ dbinom(2, size=12, prob=0.2) +
+ dbinom(3, size=12, prob=0.2) +
+ dbinom(4, size=12, prob=0.2)
[1] 0.9274
```

```
> pbinom(4, size=12, prob=0.2)
[1] 0.92744
```

Probability Distributions in R

Poisson Distribution

Problem

If there are twelve cars crossing a bridge per minute on average, find the probability of having seventeen or more cars crossing the bridge in a particular minute.

Solution

The probability of having *sixteen or less* cars crossing the bridge in a particular minute is given by the function ppois.

```
> ppois(16, lambda=12)   # lower tail
[1] 0.89871
```

```
> ppois(16, lambda=12, lower=FALSE)   # upper tail
[1] 0.10129
```

Probability Distributions in R

Uniform Distribution

Problem

Select ten random numbers between one and three.

Solution

We apply the generation function runif of the uniform distribution to generate ten random numbers between one and three.

```
> runif(10, min=1, max=3)
[1] 1.6121 1.2028 1.9306 2.4233 1.6874 1.1502 2.7068
[8] 1.4455 2.4122 2.2171
```

Probability Distributions in R

Normal Distribution

Problem

Assume that the test scores of a college entrance exam fits a normal distribution. Furthermore, the mean test score is 72, and the standard deviation is 15.2. What is the percentage of students scoring 84 or more in the exam?

```
> pnorm(84, mean=72, sd=15.2, lower.tail=FALSE)
[1] 0.21492
```

Probability Distributions in R

Normal Distribution

How to generate normally distributed random numbers?

rnorm(n,mean=0,sd=1)

```
> rnorm(20,10,2)
 [1]  9.466577  9.830326  7.384981 11.491899 12.768284
 [6] 11.590528  9.602543  9.081311 10.391586  7.127413
[11]  8.792176  8.753303 11.418330 10.822846 11.317697
[16]  9.176370 10.798532  9.070099  9.422551  7.723528
>
```

Q&A