R Lab I

Zeren Li

9/2/2019

```
# remove all objects
rm(list = ls())
```

R Markdown

Seeing Theory

"Seeing Theory is a project designed and created by Daniel Kunin with support from Brown University's Royce Fellowship Program. The goal of the project is to make statistics more accessible to a wider range of students through interactive visualizations."

Check this: https://seeing-theory.brown.edu/basic-probability/index.html

Importing dataset

Here are various ways of importing data:

```
library(readr)
library(tidyverse)
```

```
## Registered S3 method overwritten by 'rvest':
##   method            from
##   read_xml.response xml2

## -- Attaching packages ------------------------------------------ tidyverse 1.2.1 --

## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v ggplot2 3.2.1     v forcats 0.4.0

## -- Conflicts --------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(haven)

# RData (R)
load("UNpop.RData")
# csv
UNpop <- read_csv("./UNpop.csv") # readr package
```

```
## Parsed with column specification:
## cols(
##   year = col_double(),
##   world.pop = col_double()
## )
```

```
#dta (Stata)
UNpop_stata_new <- read_dta("UNpop.dta") # haven package (new)
```

Read CEO data

```
ceo = read_dta("CEOSAL2.DTA") # read CEO dataset
class(ceo)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```
summary(ceo$salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   100.0   471.0   707.0   865.9  1119.0  5299.0
```

Mean and Variance

$$\mathrm{E}[X] = \sum_{x\in\mathcal{X}} xP(x)$$

```
m_salary = sum(ceo$salary)/length(ceo$salary)
mean(ceo$salary)
```

```
## [1] 865.8644
```

sample variance:

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}[X])^2]$$

```
sum( (ceo$salary - m_salary)^2 )/ (length(ceo$salary)-1)
```

```
## [1] 345261.2
```

```
var(ceo$salary)
```

```
## [1] 345261.2
```

Sample Covariance & Correlation

$$Cov(X,Y) = E[(X - E(X)E(Y - E(Y))]$$

$$Corr(X,Y) = \frac{E[(X - E(X)E(Y - E(Y))]}{\sqrt{Var(X)Var(Y)}}$$

We would like look at the covariance and coreelation between CEO's salary and firm peformance measured by profit margins

```
cov(ceo$salary,ceo$profmarg,) # covariance
```

```
## [1] -303.6705
```

```
m_profmarg = sum(ceo$profmarg)/length(ceo$profmarg)
sum((ceo$salary - m_salary) * (ceo$profmarg - m_profmarg ))/(length(ceo$profmarg) - 1)
```

```
## [1] -303.6705
```

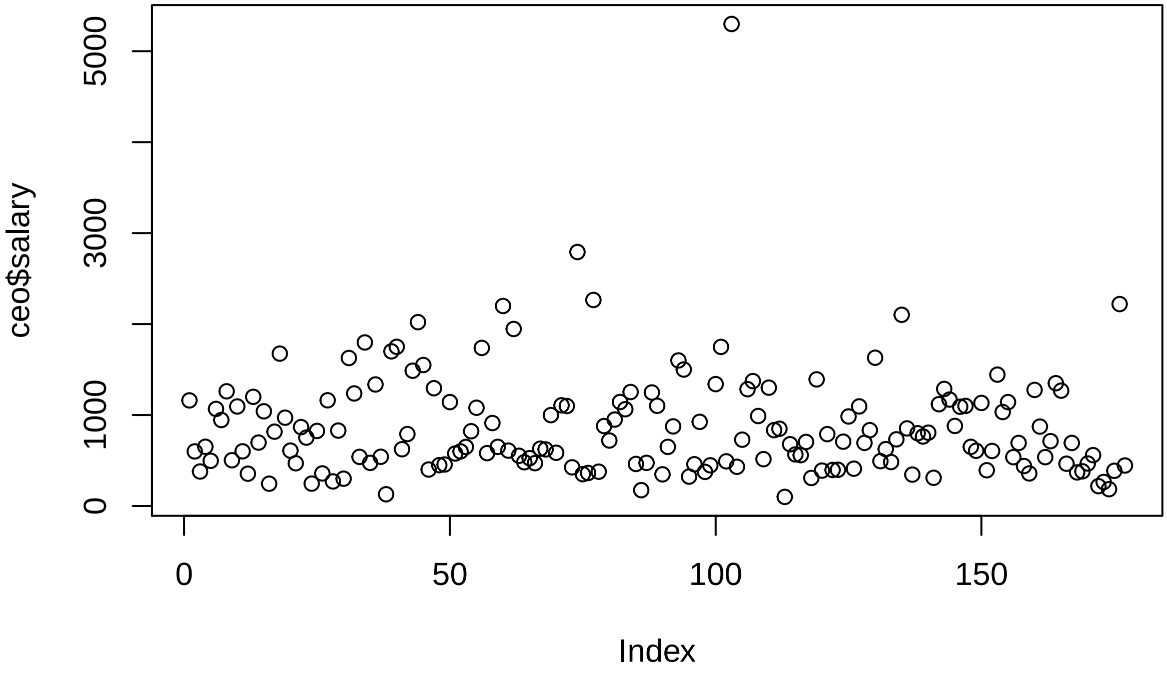
```
cor(ceo$salary,ceo$profmarg) # correlation
```

```
## [1] -0.02893538
```

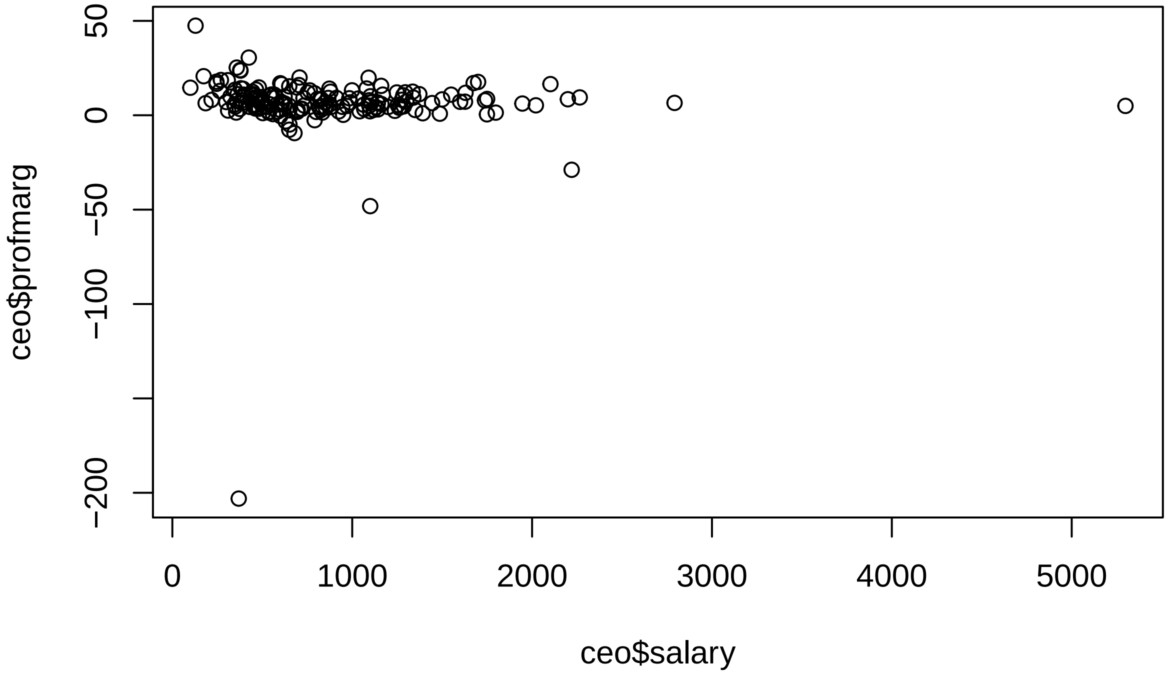
```
# how to compute manually?
```

CDF and PDF of Normal Distribution

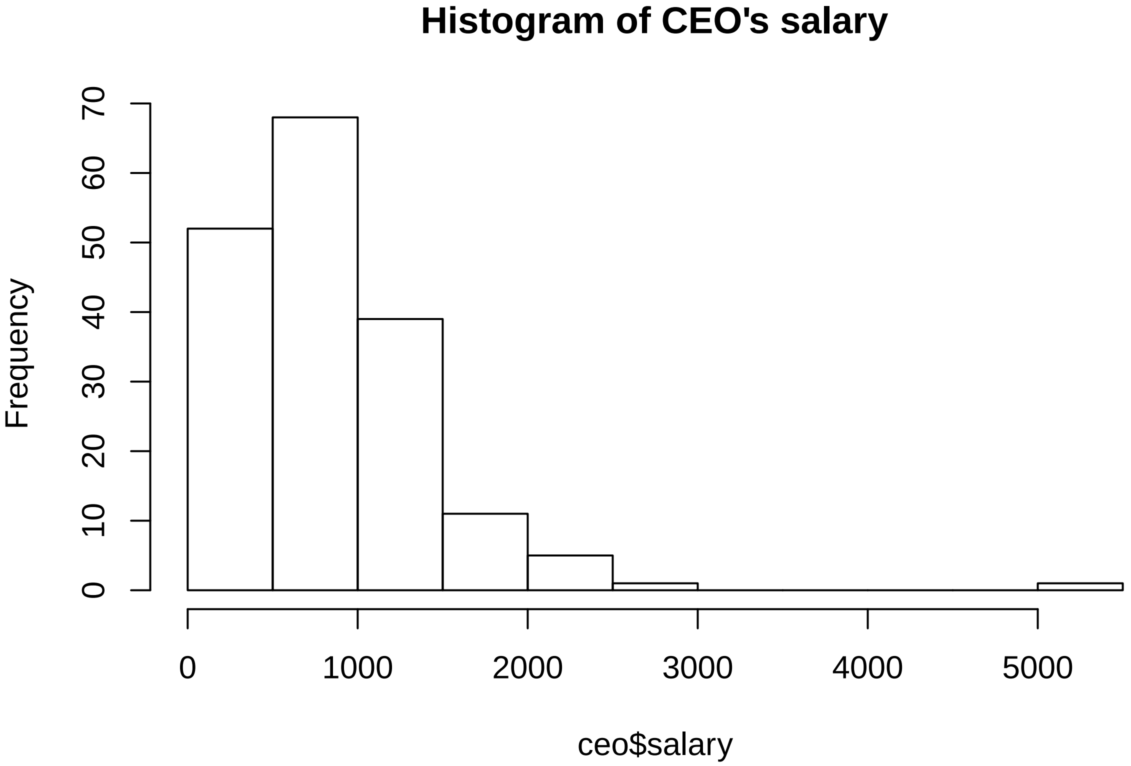
```
plot(ceo$salary) # one-way scatterplot
```



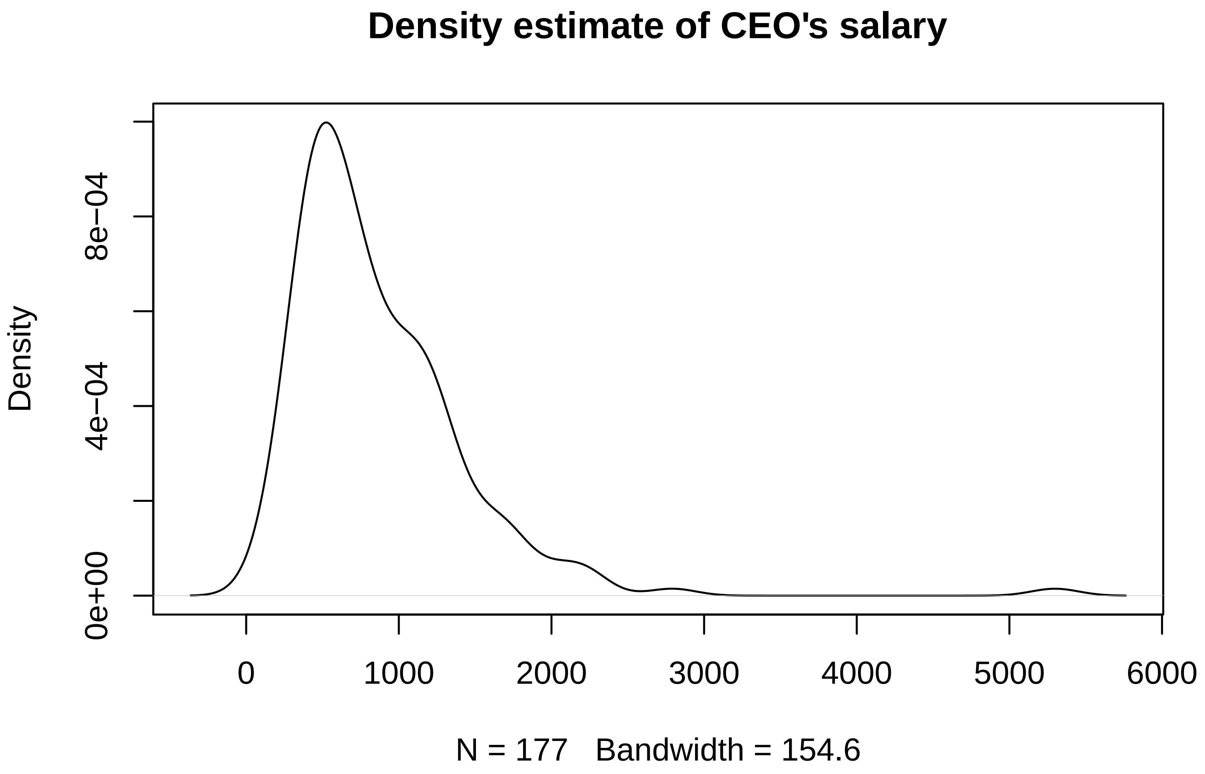
```
plot(ceo$salary, ceo$profmarg)
```



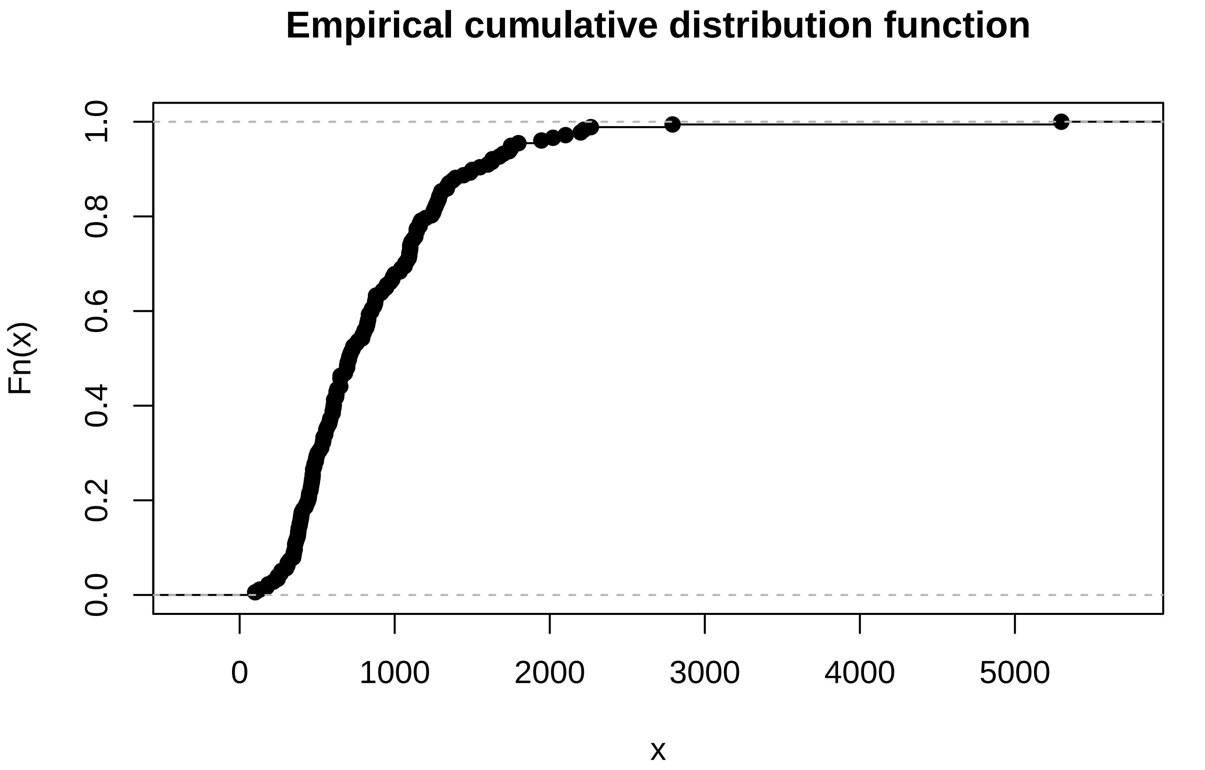
```
hist(ceo$salary,main="Histogram of CEO's salary")
```



```
plot(density(ceo$salary),main="Density estimate of CEO's salary")
```



```
plot(ecdf(ceo$salary),main=  "Empirical cumulative distribution function")
```



Exploring Real Data

```
# load package
library(wbstats)
library(readr)
library(foreign)
library(haven)
```

View Data

Exploratory Analysis

```
class(ceo) # the type of data structure
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```
dim(ceo) # dimension
```

```
## [1] 177  15
```

```
names(ceo) # variable names (column)
```

```
##  [1] "salary"   "age"      "college"  "grad"     "comten"   "ceoten"  
##  [7] "sales"    "profits"  "mktval"   "lsalary"  "lsales"   "lmktval" 
## [13] "comtensq" "ceotensq" "profmarg"
```

```
nrow(ceo) # number of row
```

```
## [1] 177
```

```
ncol(ceo) # number of column
```

```
## [1] 15
```

```
# filter
  ceo_1to5 <- ceo[c(1:5), ]
  # View(data_s1)

# select variables
# after viewing the data, we decide we need only the "country" and "value" column
  names(ceo_1to5)
```

```
##  [1] "salary"   "age"      "college"  "grad"     "comten"   "ceoten"  
##  [7] "sales"    "profits"  "mktval"   "lsalary"  "lsales"   "lmktval" 
## [13] "comtensq" "ceotensq" "profmarg"
```

```
  ceo_1to5 <- ceo_1to5[, c("salary", "profmarg")]

# rename variable
# rename the second column to "GDP_PAP_2016"
  names(ceo_1to5)[2] <- "profit_margin"
  # View(data_s2)
# Remove
  rm(ceo_1to5)
```