

R을 이용한 통계 기초와 데이터 분석

남현진

한성대학교

2020

Section 1

통계 이론

강의 내용

- 모평균, 모분산, 표본 평균, 표본 분산
- 백분위수, 상자그림
- 확률분포
- 행렬, 행렬의 연산, 역행렬, 전치행렬
- R 설치 및 인터페이스 설명
- 실습: R로 평균과 분산 계산

Definition

$$\mu = \frac{\sum x}{N} \quad (1)$$

where:

- μ denotes the population mean.
- x denotes any value.
- N denotes the number of the values in the population.

Examples

When $x = (1, 2, 3, 4, 5)$ then $N = 5$, $\mu = 3$

Definition

$$\bar{x} = \frac{\sum x}{n} \quad (2)$$

where:

- \bar{x} denotes the sample mean.
- n denotes the number of values in the sample.
- x denotes any value.

Examples

When $x = (1, 2, 3, 4, 5)$ then $n = 5$, $\bar{x} = 3$

모분산 Population variance

Definition

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N} \quad (3)$$

where:

- σ^2 is the population variance.
- x is the value of a particular observation in the population.
- μ is the arithmetic mean of the population.
- N is the number of observations in the population.

Examples

When $x = (1, 2, 3, 4, 5)$ then $N = 5$, $\mu = 3$, $\sigma^2 = 2$

표본 분산 Sample variance

Definition

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} \quad (4)$$

where:

- s^2 is the sample variance.
- x is the value of a each observation in the sample.
- \bar{x} is the mean of the sample.
- n is the number of observations in the sample.
- The denominator $(n - 1)$ corrects its tendency for underestimation.

Examples

When $x = (1, 2, 3, 4, 5)$ then $n = 5$, $\bar{x} = 3$, $s^2 = 2.5$

백분위수(Percentile) 와 사분위수(Quantile)

Definition

백분위수란 크기 순서로 나열한 자료를 100등분 했을때, $x\%$ 인 관측값을 의미한다. x 분위값이란 자료 값 중 $x\%$ 가 그 값보다 작거나 같게 되는 값이다.[2]

Important theorem

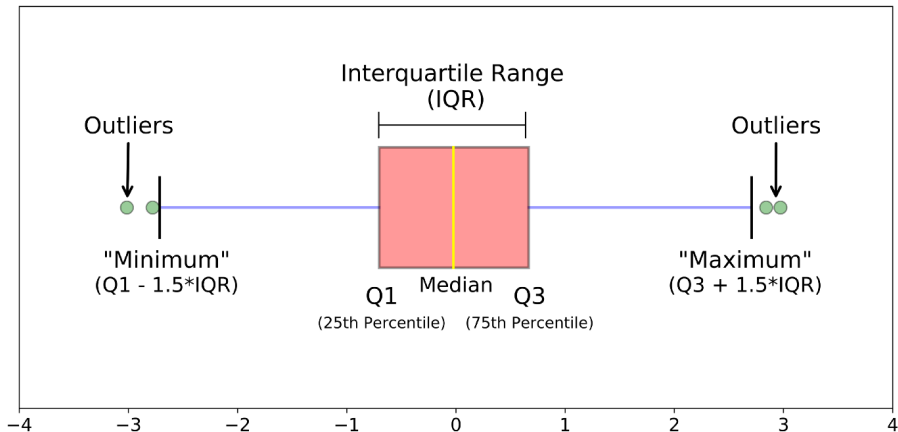
- $Q_1 = 25\text{분위수} = \text{자료의 } 25\% \text{에 상응하는 } x \text{ 분위값}$
- $Q_2 = 50\text{분위수} = \text{중앙값} = \text{자료의 } 50\% \text{에 상응하는 } x \text{ 분위값}$
- $Q_3 = 75\text{분위수} = \text{자료의 } 75\% \text{에 상응하는 } x \text{ 분위값}$
- $Q_4 = 100\text{분위수} = \text{자료의 } 100\% \text{에 상응하는 } x \text{ 분위값}$
- $IQR = Q_3 - Q_1$

Examples

When $y = (1, 3, 3, 4, 5, 6, 6, 7, 8, 8)$ then $Q_1 = 3, Q_2 = 5.5, Q_3 = 7, IQR = 4$

상자그림 Box-plot

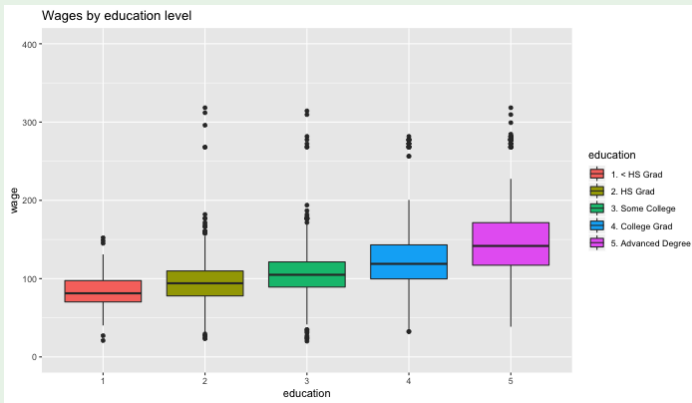
Definition



[3]

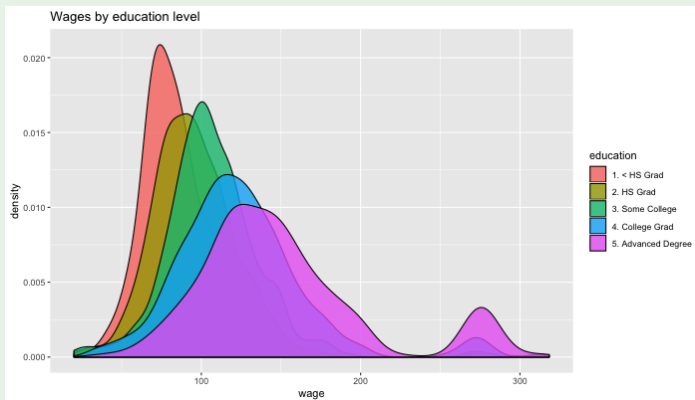
상자그림 Box-plot

Examples



Wage data for a group of 3000 male workers in the Mid-Atlantic region. Data was manually assembled by Steve Miller, of Open BI (www.openbi.com), from the March 2011 Supplement to Current Population Survey data.

Examples



정규분포 Normal Probability Distribution

Definition

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left[\frac{(x-\mu)^2}{2\sigma^2}\right]} \quad (5)$$

where:

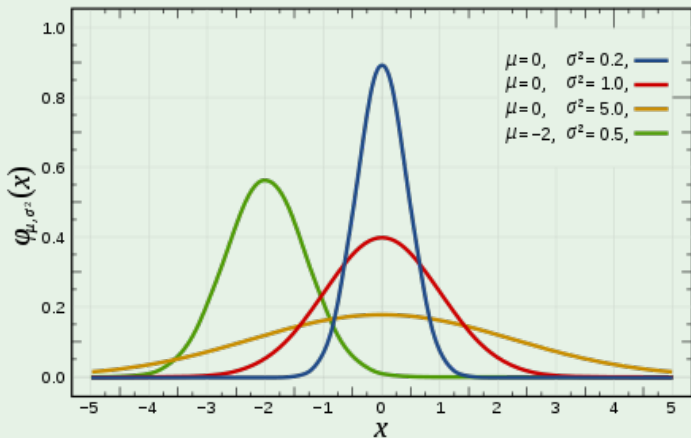
- σ refers to the standard deviation.
- μ refers to the mean.
- e is a constant, respectively, the base of the natural log system and approximately equals to 2.718.
- π a constant with an approximate value of $\frac{22}{7}$ or 3.1416.
- x refers to the value of the random variable.

Important theorem

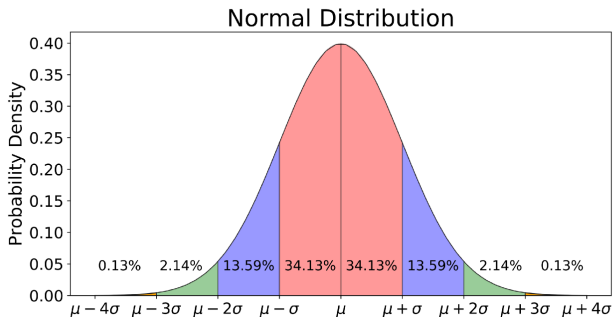
When $\mu = 0$ and $\sigma = 1$, it is known as the standard normal distribution.

확률분포 Probability Distribution

Examples



Examples



행렬 Matrix

Definition

행렬은 수나 수를 나타내는 문자를 괄호 안에 직사각형 꼴로 배열한 것이다. 행렬의 가로 줄을 행, 세로 줄을 열이라 한다.

Examples

$A = \begin{pmatrix} 1 & -2 & 2 \\ 3 & -1 & 3 \end{pmatrix}$ 일 때, A 는 2행과 3열을 가진 행렬이다.

행렬의 연산 Matrix calculation

Definition

주어진 두 $m \times n$ 행렬 A와 B에 대해, 덧셈과 뺄셈은 각각 성분별 덧셈과 뺄셈으로 정의된다.

Examples

$$\begin{pmatrix} 1 & 3 & 7 \\ 1 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 5 \\ 7 & 5 & 0 \end{pmatrix} = \begin{pmatrix} 1+0 & 3+0 & 7+5 \\ 1+7 & 0+5 & 0+0 \end{pmatrix} = \begin{pmatrix} 1 & 3 & 12 \\ 8 & 5 & 0 \end{pmatrix}$$

행렬의 곱셈 Matrix multiplication

Definition

주어진 $m \times n$ 행렬 A와 $n \times p$ 행렬 B의 곱은 $m \times p$ 행렬이며, 각 (i, j) 성분은 A의 i 행벡터와 B의 j 열벡터의 점곱으로 정의된다.

Important theorem

$$(\mathbf{AB})_{ij} = \sum_{k=1}^n \mathbf{A}_{ik} \mathbf{B}_{kj} = A_{i1}B_{1j} + A_{i2}B_{2j} + \cdots + A_{in}B_{nj}$$

이다.

Examples

$$\begin{pmatrix} 1 & 0 & 2 \\ -1 & 3 & 1 \end{pmatrix} \cdot \begin{pmatrix} 3 & 1 \\ 2 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} (1 \cdot 3 + 0 \cdot 2 + 2 \cdot 1) & (1 \cdot 1 + 0 \cdot 1 + 2 \cdot 0) \\ (-1 \cdot 3 + 3 \cdot 2 + 1 \cdot 1) & (-1 \cdot 1 + 3 \cdot 1 + 1 \cdot 0) \end{pmatrix} = \begin{pmatrix} 5 & 1 \\ 4 & 2 \end{pmatrix}$$

역행렬 Inverse Matrix

Definition

2개의 행렬 A, B 에서 $AB=I$ 이 되는 B 를 A 의 역행렬이라 하고, 그 때의 관계를 $B=A^{-1}$ 로 나타내어진다. 이 때 I 은 단위 행렬이라 한다.

Important theorem

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \text{ 일 때 } AA^{-1} = I \text{ 가 되려면 } A^{-1} = \begin{pmatrix} \frac{d}{ad-bc} & \frac{-b}{ad-bc} \\ \frac{-c}{ad-bc} & \frac{a}{ad-bc} \end{pmatrix}$$

이다.

Examples

$$A = \begin{pmatrix} 1 & -2 \\ 3 & -1 \end{pmatrix} \text{ 일때 } A^{-1} = \begin{pmatrix} -1/5 & 2/5 \\ -3/5 & 1/5 \end{pmatrix} \text{ 이다.}$$

전치행렬 Transposed Matrix

Definition

임의의 행렬 A 가 주어졌을 때 그 행렬 A 에서 행과 열을 바꾼 행렬을 행렬 A 의 전치행렬이라 하고, 보통 A^T 혹은 A' 로 나타낸다.

Important theorem

$$A = \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix} \text{ 일 때 } A^T = \begin{pmatrix} a & d \\ b & e \\ c & f \end{pmatrix} \text{ 이다.}$$

Section 2

R 실습

기본적인 R언어 문법

```
print("Hello, world!")  
## [1] "Hello, world!"
```

기본적인 R언어 문법

```
2 + 2
```

```
## [1] 4
```

- “[]” 는 그 줄의 첫 element의 위치 값을 보여준다.

```
1:10
```

```
## [1] 1 2 3 4 5 6 7 8 9 10
```

- “:”는 앞부터 뒤까지를 포함하는 연속적인 숫자들을 출력한다.

기본적인 R언어 문법

```
x <- 2 + 2
```

```
y <- 3 * 6
```

- “<-” 는 해당 값을 부여하는 연산자이다.

```
y/x #y divided by x
```

```
## [1] 4.5
```

- 저장된 값을 이용하여 원하는 계산을 실행할 수 있다.
- 실행하지 않는 코드는 “#”를 사용하여 적을 수 있다. 해당 코드를 설명하는 문장을 적을때 유용하게 쓸 수 있다.

데이터 구조의 이해

Data Structure

- 1차원 데이터: Vector (동질성), List (이질성)
- 2차원 데이터: Matrix (동질성), Data frame (이질성)
- 3차원 데이터: Array

Vector

```
vec1 <- c(1, 2, 3)
```

```
vec1
```

```
## [1] 1 2 3
```

- R에서 벡터는 동질적인 값을 가지고 있는 숫자의 집합이다.
- 벡터를 만들기 위해서는 c()를 이용하여 해당 값을 부여한다.

```
vec2<- c("How", "are", "you", "?")
```

```
vec2
```

```
## [1] "How" "are" "you" "?"
```

- 따옴표안에 숫자를 넣으면 R은 입력값을 문자열로 인식한다.

List

```
list1 <- list(vec1, vec2)
list1
## [[1]]
## [1] 1 2 3
##
## [[2]]
## [1] "How" "are" "you" "?"
```

- 리스트는 이질적인 변수들이 모여있는 데이터 형태이다.

Matrix

```
matrix1 <- matrix(1:12, nrow=2, ncol = 6)
```

```
matrix1
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]  
## [1,]    1    3    5    7    9   11  
## [2,]    2    4    6    8   10   12
```

```
matrix2 <- matrix(1:12, nrow=2, ncol = 6, byrow = TRUE)
```

```
matrix2
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]  
## [1,]    1    2    3    4    5    6  
## [2,]    7    8    9   10   11   12
```

Matrix

```
vec <- c(1,2,3)
```

```
cbind(vec,vec)
```

```
##      vec vec
```

```
## [1,]  1  1
```

```
## [2,]  2  2
```

```
## [3,]  3  3
```

```
rbind(vec,vec)
```

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

```
## vec    1    2    3
```

```
## vec    1    2    3
```

- cbind를 사용하면 column(열)을 기준으로 두 벡터/행렬/데이터프레임이 결합한다.
- rbind를 사용하면 row(행)기준으로 두 벡터/행렬/데이터프레임이 결합한다.

Data Frame

```
df1 <- data.frame(item = c('pencil', 'pen', 'eraser'),  
                  stock = c(T,T,F),  
                  price = c(1000,1300,500))
```

df1

```
##      item stock price  
## 1 pencil  TRUE  1000  
## 2   pen   TRUE  1300  
## 3 eraser FALSE   500
```

- 데이터 프레임은 가장 많이 쓰이는 데이터 형식이다.

Data Frame

```
df2 <- as.data.frame(matrix2)
```

```
df2
```

```
##      V1 V2 V3 V4 V5 V6
```

```
## 1    1  2  3  4  5  6
```

```
## 2    7  8  9 10 11 12
```

- as.data.frame 함수를 사용하여 행렬을 데이터프레임 형식으로 바꿀 수 있다.

데이터 구조의 이해

Data Type

- Integer: 실수
- Numeric: 정수
- Character(string): 문자열
- Factor: 요인형
- Logical(boolean): 논리값

데이터 타입

```
head(sample_data)
```

```
## # A tibble: 6 x 5
##   name.first registered.age gender location.country marital.status
##   <chr>           <dbl> <fct>   <fct>           <lgl>
## 1 Janique             21 male    Brazil          FALSE
## 2 Khaled              24 male    Norway          FALSE
## 3 Maja               30 female  Denmark         FALSE
## 4 Latife              28 female  Turkey          FALSE
## 5 Sayenne            32 female  Netherlands     TRUE
## 6 Linda              38 female  Ireland         TRUE
```

```
str(sample_data)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':   100 obs. of  5 variables:
```


평균과 분산 구하기

1부터 1000까지의 100개의 난수를 생성한 후 평균과 분산을 구하기.

```
set.seed(1)
x <- sample(1:1000, 100, replace=T)
mean(x)
## [1] 534.38
sd(x)
## [1] 289.6076
var(x)
## [1] 83872.54
```

참고 문서



Lind, Douglas A. et. al. (2015): Statistical Techniques in Business and Economics, Sixteenth edition, New York, NY 2015.



BioinformaticsAndMe : 분위수(Quantile)
<https://bioinformaticsandme.tistory.com/246>



Gravatar, How To Manually Order Boxplot in Seaborn?, Data Viz with Python and R
<https://datavizpyr.com/how-to-manually-order-boxplot-in-seaborn/>



Wikipedia: Normal distribution
https://en.wikipedia.org/wiki/Normal_distribution



전자용어사전, 월간전자기술 편집위원회, 성안당
<https://terms.naver.com/entry.nhn?docId=753859&cid=42341&categoryId=42341>