R 데이터 관리

저장 및 로드

R 데이터 파일 저장

```
save(x,y,
file="mydata.RData)
```

```
load("mydata.RData")
```

데이터 목록 및 제거

```
ls()
```

```
rm(list=ls())
```

csv 파일 불러오기 및 저장하기

```
read.csv("mydata.csv",
stringsAsFactors=FALSE,
header=FALSE)
```

```
write.csv(pt_data,
file="pt_data.csv")
```

데이터 탐색과 이해

usedcars 데이터 셋 설명 :
2012년에 미국의 유명 웹 사이트에서 최근에 판매하고자 광고를 했던
중고차의 실제 데이터

no	year	model	price	mileage	color	transmission
1	2011	SEL	21992	7413	Yellow	AUTO
2	2011	SEL	20995	10926	Gray	AUTO
…						
…						
150	2000	SE	3800	109259	Red	AUTO

데이터 구조 탐색

```
#데이터 불러오기
usedcars <- read.csv("usedcars.csv", stringsAsFactors = FALSE)


#데이터 구조 탐색 : 데이터 셋이 어떻게 구성돼 있는가?
str(usedcars)
```

```
> str(usedcars)
'data.frame':   150 obs. of  6 variables:
 $ year        : int  2011 2011 2011 2011 2012 2010 2011 2010 2011 2010 ...
 $ model       : chr  "SEL" "SEL" "SEL" "SEL" ...
 $ price       : int  21992 20995 19995 17809 17500 17495 17000 16995 16995 16995 ...
 $ mileage     : int  7413 10926 7351 11613 8367 25125 27393 21026 32655 36116 ...
 $ color       : chr  "Yellow" "Gray" "Silver" "Gray" ...
 $ transmission: chr  "AUTO" "AUTO" "AUTO" "AUTO" ...
```

데이터 수치 탐색

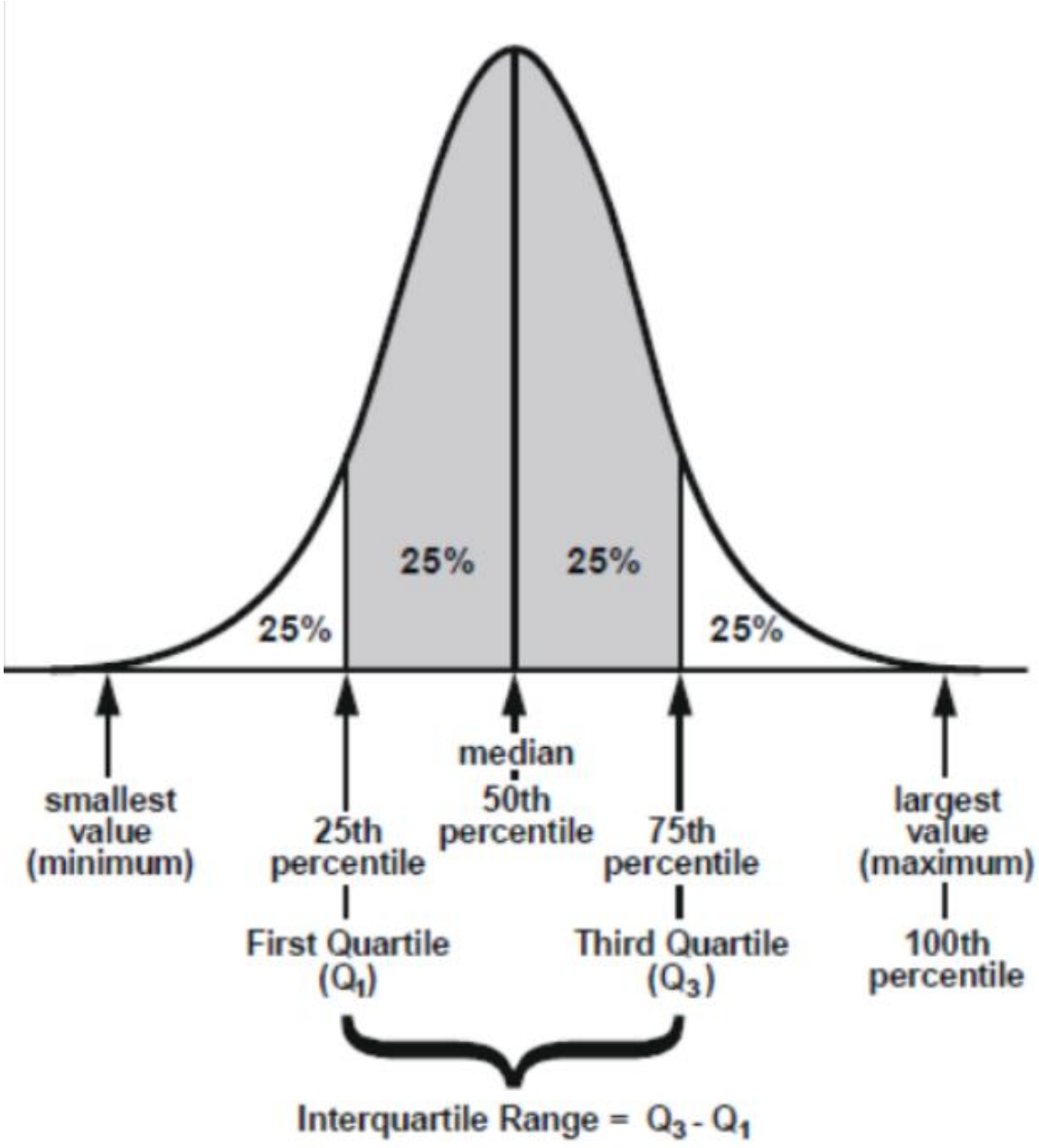
```
#수치변수탐색
#요약통계량
summary(usedcars$year)
#여러 열 벡터의 요약통계량
summary(usedcars[c("price","mileage")])

#중심경향 측정 : 평균과 중앙값
#평균
mean(usedcars$price)
mean(usedcars$mileage)
중앙값
median(usedcars$price)
median(usedcars$mileage)

#퍼짐 측정 : 사분위수와 다섯 숫자 요약
range(usedcars$price) # 최소값, 최대값
diff(range(usedcars$price)) # 최대값 - 최소값
21992-3800
IQR(usedcars$price) # 사분위수 범위
quantile(usedcars$price) # 사분위수
quantile(usedcars$price,probs = c(0.01,0.99)) # 절단점
quantile(usedcars$price,seq(from=0, to=1, by=0.2))
```

수치요약

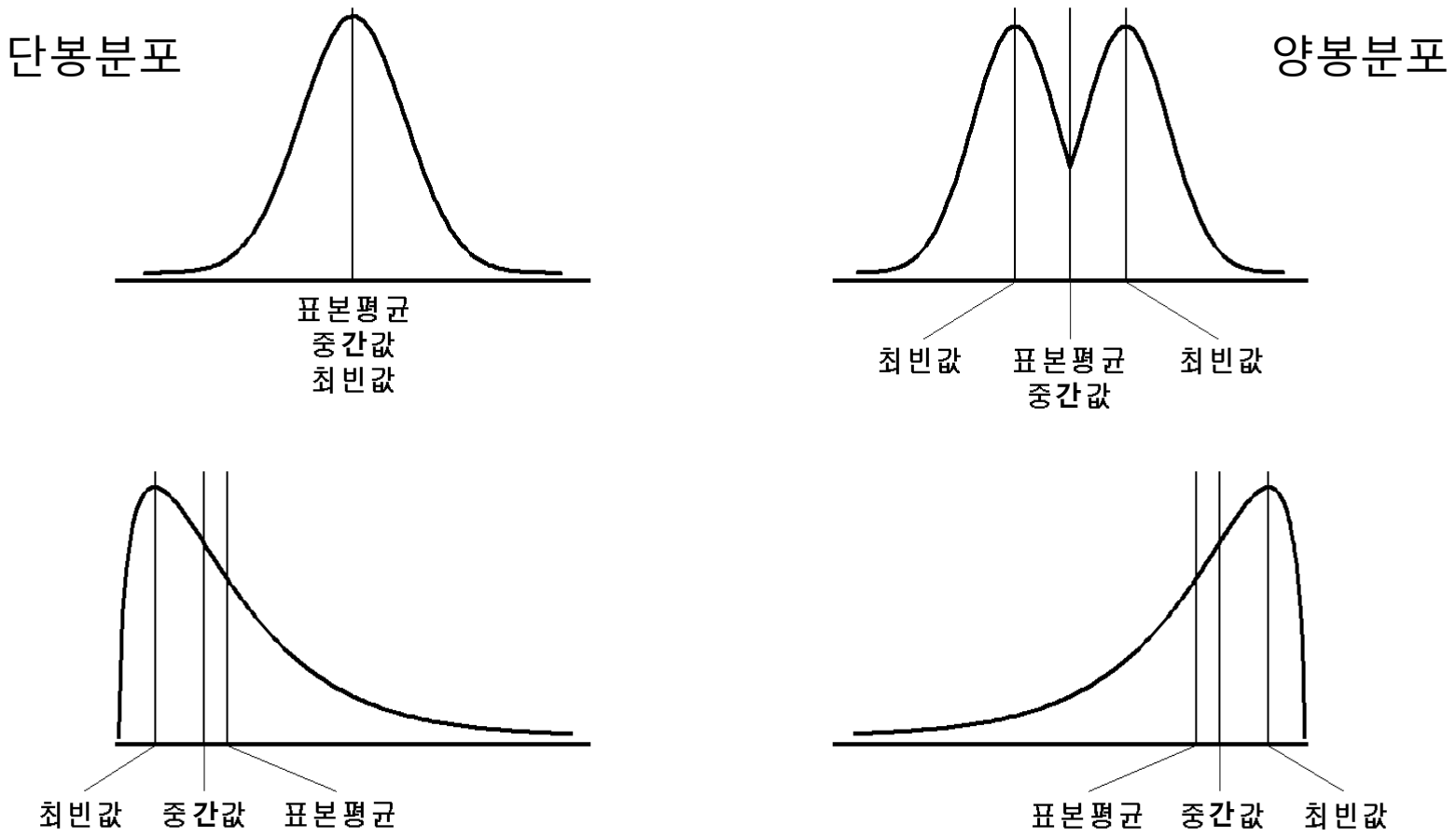
- 중심위치의 측도
  - 평균, 중앙값, 최빈값
  - 평균은 극단적으로 아주 큰 값이나 아주 작은 값의 영향을 받을 수 있다.(not robust)
  - 중앙값은 평균과 달리 관측값의 변화에 민감하지 않고 큰 값이나 작은 값의 영향을 받지 않는다.(robust)

- 산포(퍼진 정도)의 측도
  - 표준편차,분산, 분위수(백분위수, 사분위수), 범위(max-min)
  - 표준편차는 편차의 제곱으로부터 나오므로 저항성이 떨어진다.
  - 표준편차보다 사분위수가 더 강한 저항성을 갖기 때문에 사분위수의 차이를 산포의 측도로 사용

- 다섯수치요약
  - 최소값, 제1사분위수, 중앙값, 제3사분위수, 최대값


25%
25%
25%
25%
smallest value (minimum)
25th percentile
median 50th percentile
75th percentile
largest value (maximum)
First Quartile (Q_1)
Third Quartile (Q_3)
100th percentile
Interquartile Range = $Q_3 - Q_1$

분포모양에 따른 중심위치의 측도

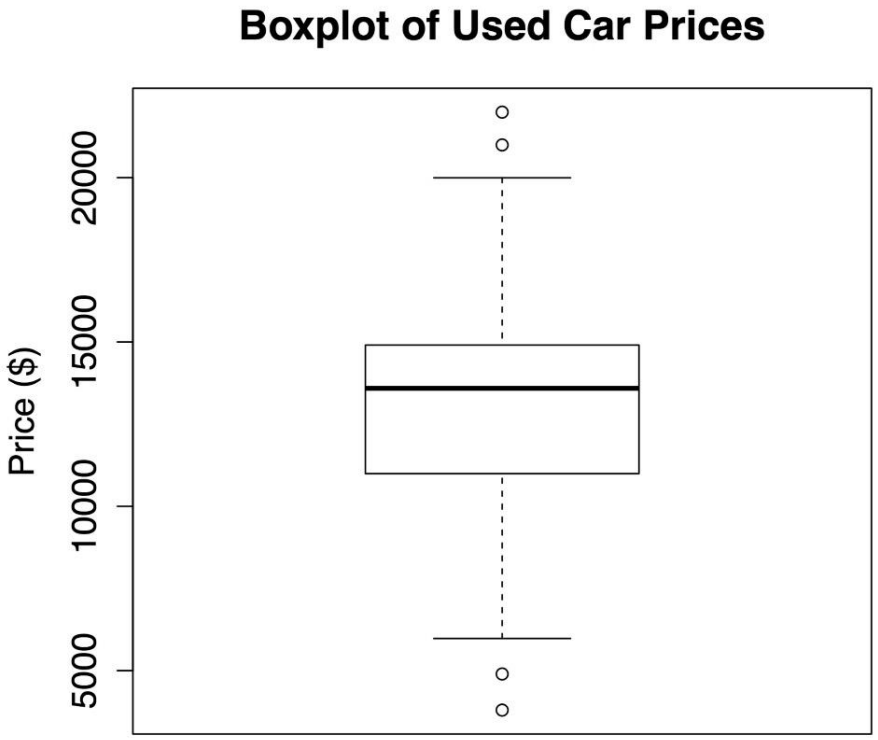


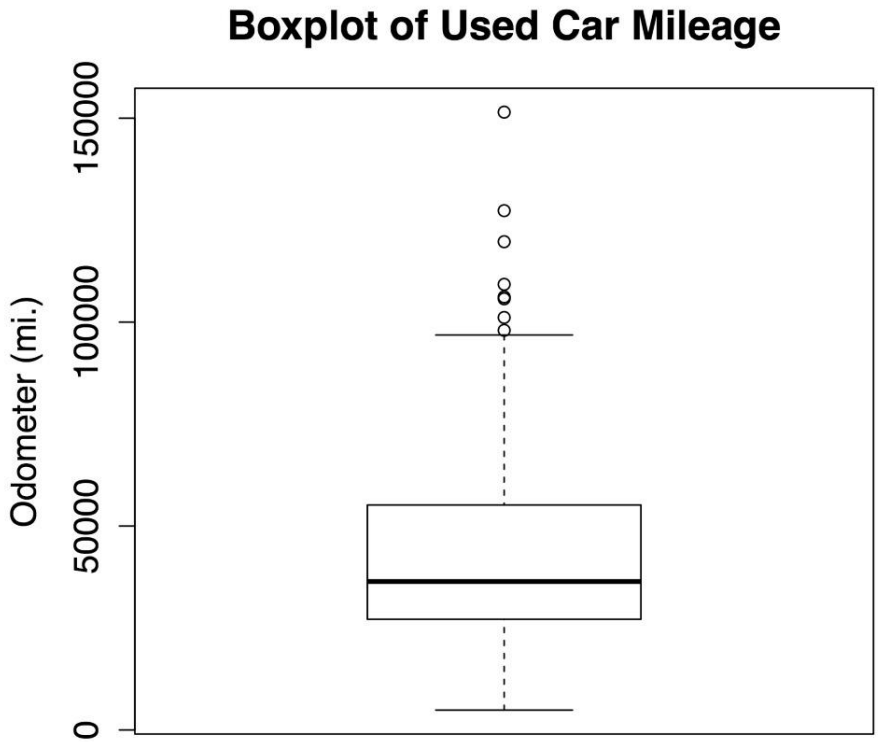
편중된 자료(skewed data)에 대해서는 중간값이 표본평균보다 중심위치의 측도로 적절할 수 있다

수치 변수 시각화

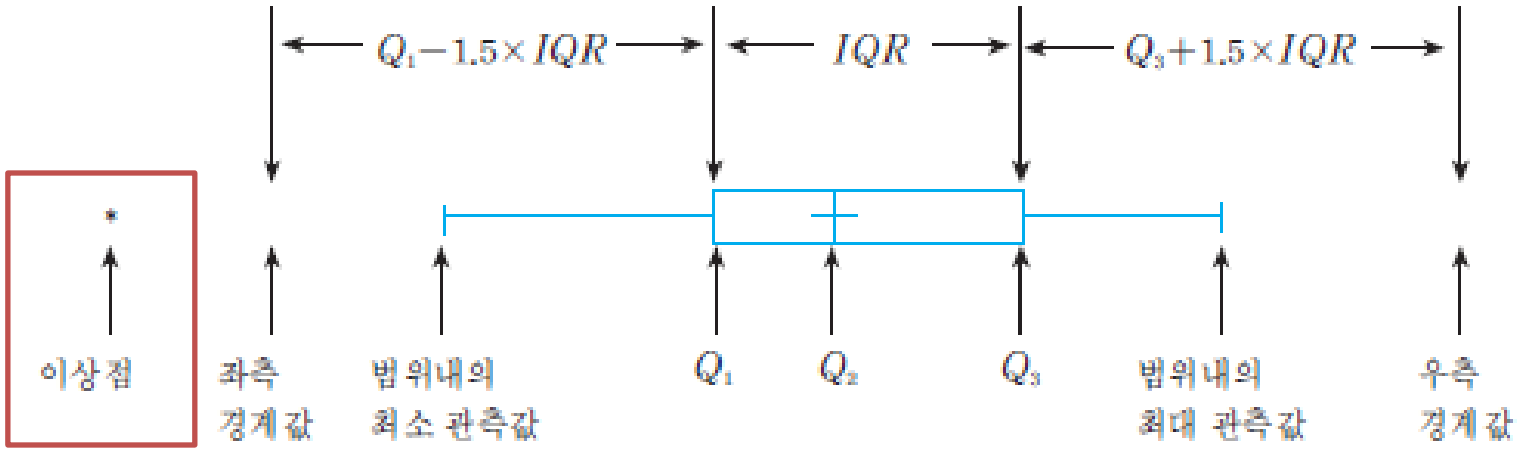
- 수치변수 시각화
  - 상자그림
  - 히스토그램

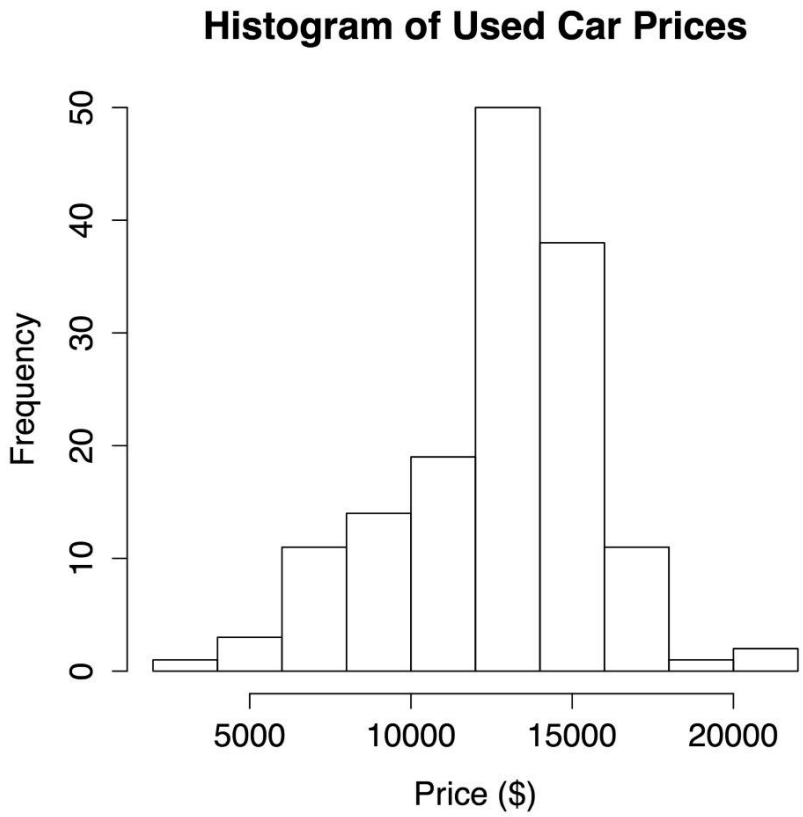
```
#수치 변수 시각화

boxplot(usedcars$price, main="Boxplot of Used Car Prices", ylab="Price ($)")

boxplot(usedcars$mileage, main="Boxplot of Used Car Mileage", ylab="Odometer (mi.)")

hist(usedcars$price, main = "Histogram of Used Car Prices", xlab="Price ($)")

hist(usedcars$mileage, main="Boxplot of Used Car Mileage", ylab="Odometer (mi.)")
```

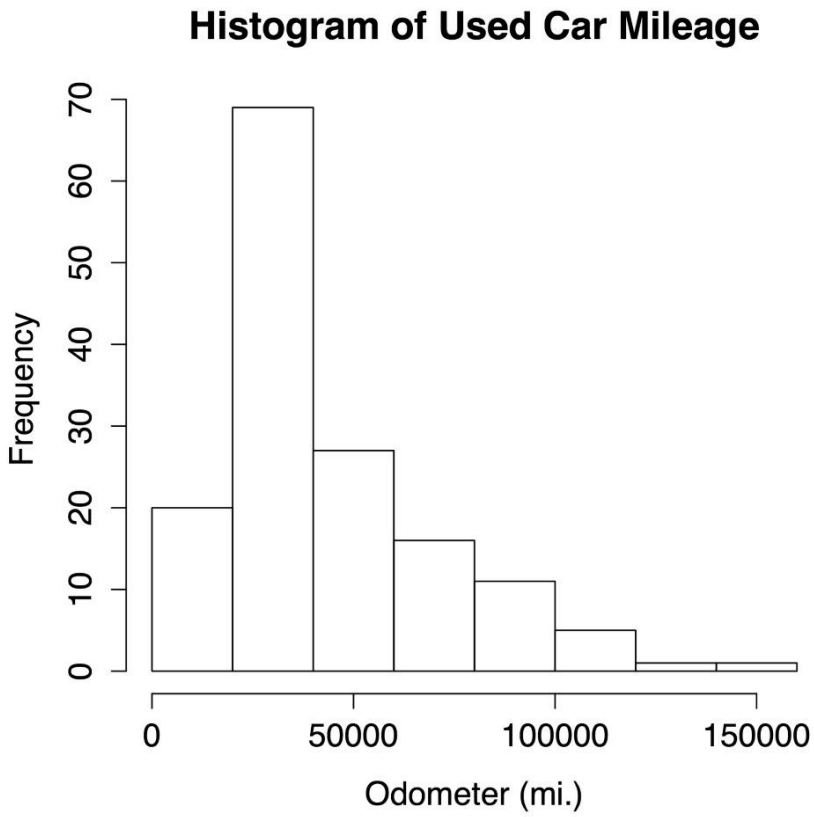

Boxplot of Used Car Prices
Price ($)
5000
10000
15000
20000

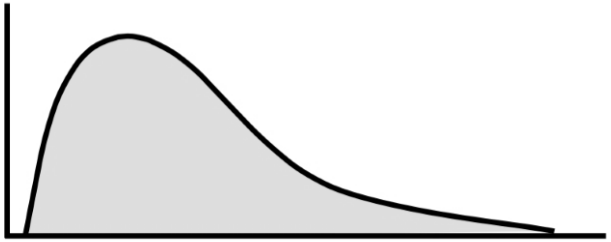

Boxplot of Used Car Mileage
Odometer (mi.)
0
50000
100000
150000

- Box and whisker plot
  - 다섯 수치 요약을 시각화하는 방법.

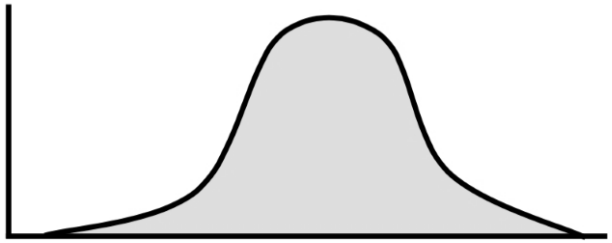



Histogram of Used Car Prices
Frequency
0
10
20
30
40
50
5000
10000
15000
20000
Price ($)

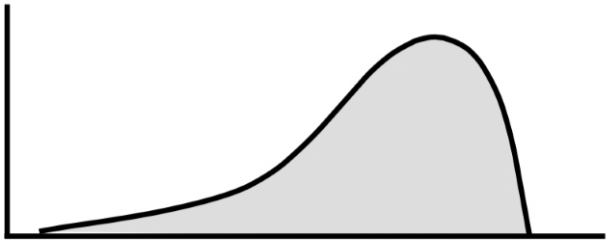

Histogram of Used Car Mileage
Frequency
0
10
20
30
40
50
60
70
0
50000
100000
150000
Odometer (mi.)



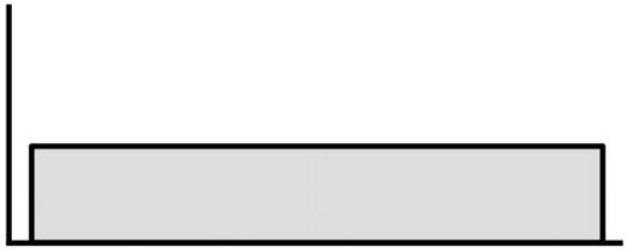
Right Skew



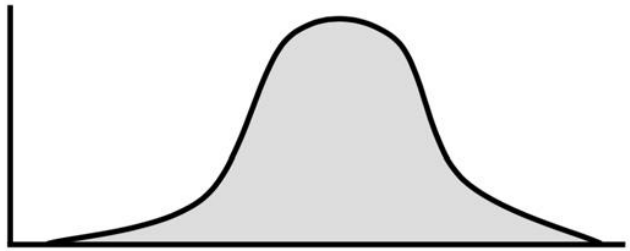
No Skew



Left Skew



Uniform Distribution



Normal Distribution

- 퍼짐 측정
  - 분산과 표준편차

$$\mathrm{Var}(X) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\mathrm{StdDev}(X) = \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

```
#정규분포의 중심위치와 퍼짐정도

var(usedcars$price)

sd(usedcars$price)

var(usedcars$mileage)

sd(usedcars$mileage)
```

68-95-99.7 규칙
정규 분포에서 값의 68%는 평균의 1표준 편차 내에 포함되는 반면 95%는 2표준 편차, 99.7%는 3표준 편차 내에 각각 포함되는 것을 말한다.

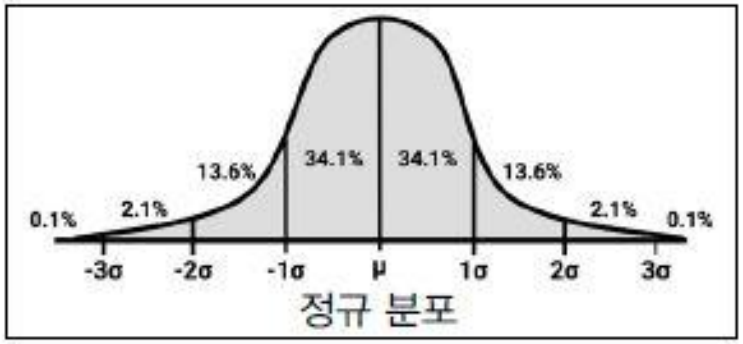
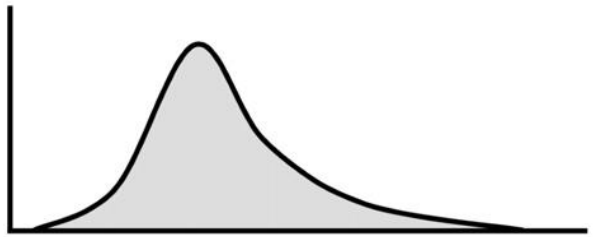


그림 2.6: 정규 분포 평균으로부터 1, 2, 3 표준 편차만큼 떨어진 부분의 %

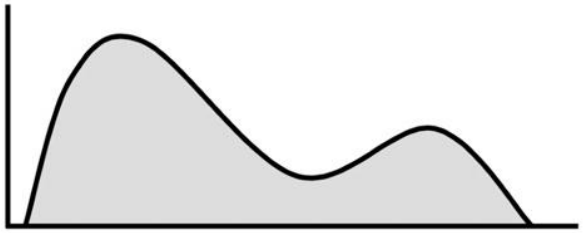
범주 변수 탐색

- 범주 변수 탐색
  - 중심 경향 측정 : 최빈값

```
#도수분포표
table(usedcars$year)
table(usedcars$model)
table(usedcars$color)
#상대도수분포표
model_table <- table(usedcars$model)
prop.table(model_table)
color_table <- table(usedcars$color)
color_pct <- prop.table(color_table)*100
round(color_pct,digits = 1)
```



Unimodal Distribution



Bimodal Distribution

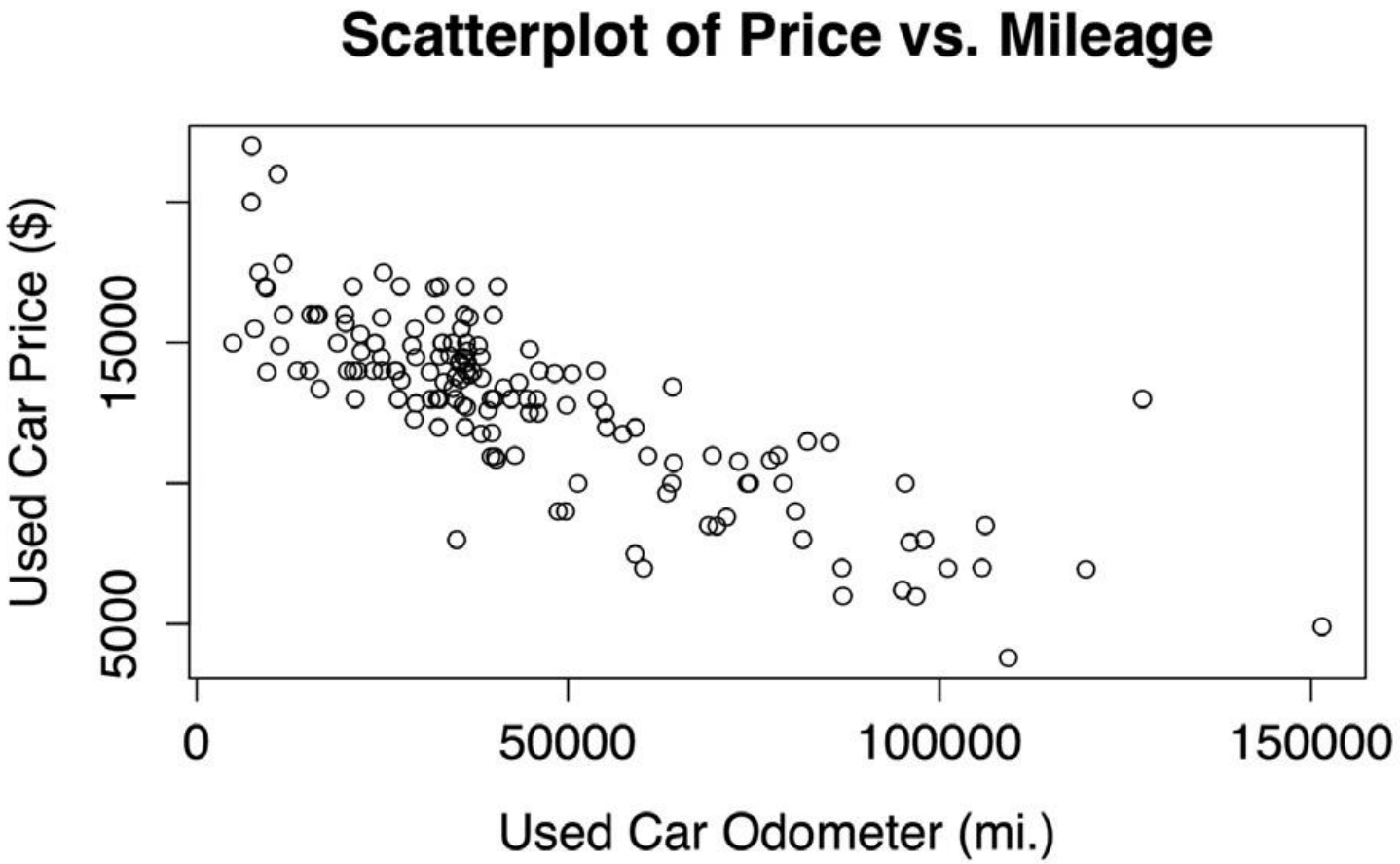
범주 변수 시각화

- 범주 변수 시각화
  - 평균막대도표
    - 데이터를 요약한 평균표를 먼저 만들고 평균표를 이용해 그래프 생성
    - + geom_col()
  - 빈도막대도표
    - 원자료를 이용해 바로 그래프 생성
    - + geom_bar()

변수 간의 관계 탐색

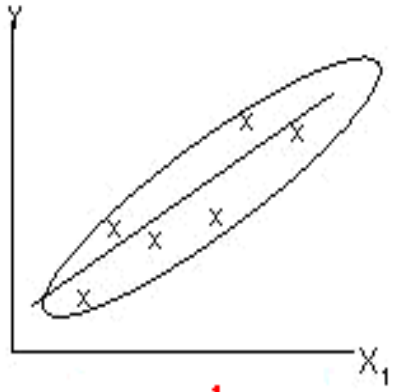
- 이변량 자료의 관계 분석
  - 관계 시각화
    - 산포도 : 두 연속형 변수의 관계를 시각화하는 다이어그램
    - 예제 : 주행거리에 따른 중고차 가격의 변화

```
#변수관계탐색 #관계시각화 #산포도

plot(x=usedcars$mileage, y=usedcars$price,

     main="Scatterplot of Price vs.Mileage",

     xlab = "Used Car Odometer (mi.)",

     ylab = "Used Car Price ($)")
```


Scatterplot of Price vs. Mileage
Used Car Price ($)
5000
15000
0
50000
100000
150000
Used Car Odometer (mi.)

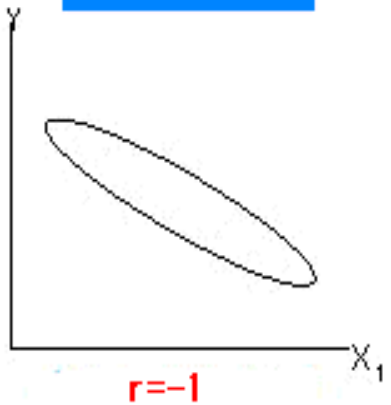
산점도의 유형

양의 선형관계



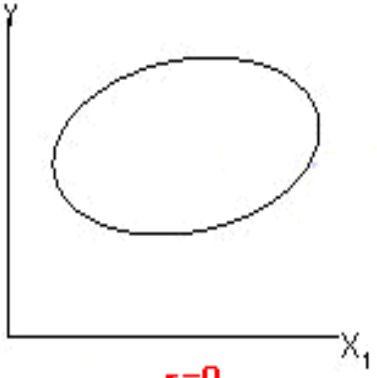
r=1

음의 선형관계



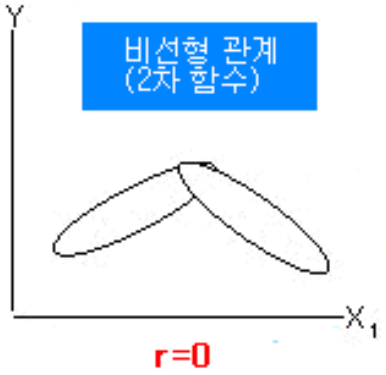
r=-1

아무 관계 없음



r=0

비선형 관계
(2차 함수)



r=0

변수 간의 관계 탐색

- 이변량 자료의 관계 분석
  - 관계 관찰
    - 이원교차표(분할표) : 두 범주형 변수의 관계를 나타내는 표
    - 하나의 변수 값이 다른 변수 값에 의해 어떻게 변하는지 관찰할 수 있다.
    - CrossTable()의 교차표 이용
      - gmodels 팩키지

```
#이원교차표

#install.packages("gmodels")

library(gmodels)

usedcars$conservative <- usedcars$color %in% c("Black","Gray","Silver","White")

table(usedcars$conservative)

CrossTable(x=usedcars$model, y=usedcars$conservative, chisq = TRUE)
```

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  150

                | usedcars$conservative
usedcars$model  |     FALSE |      TRUE | Row Total |
----------------|-----------|-----------|-----------|
             SE |        27 |        51 |        78 |
                |     0.009 |     0.004 |           |
                |     0.346 |     0.654 |     0.520 |
                |     0.529 |     0.515 |           |
                |     0.180 |     0.340 |           |
----------------|-----------|-----------|-----------|
            SEL |         7 |        16 |        23 |
                |     0.086 |     0.044 |           |
                |     0.304 |     0.696 |     0.153 |
                |     0.137 |     0.162 |           |
                |     0.047 |     0.107 |           |
----------------|-----------|-----------|-----------|
            SES |        17 |        32 |        49 |
                |     0.007 |     0.004 |           |
                |     0.347 |     0.653 |     0.327 |
                |     0.333 |     0.323 |           |
```

Pearson's Chi-squared test : 두 범주형 변수의 독립성 검정

```
#데이터 불러오기
usedcars <- read.csv("usedcars.csv", stringsAsFactors = FALSE)

#데이터 구조 탐색
str(usedcars)

#수치변수탐색
summary(usedcars$year)
summary(usedcars[c("price","mileage")])

#중심경향 측정 : 평균과 중앙값
mean(usedcars$price)
mean(usedcars$mileage)
median(usedcars$price)
median(usedcars$mileage)

#퍼짐 측정 : 사분위수와 다섯 숫자 요약
range(usedcars$price)
diff(range(usedcars$price))
21992-3800
IQR(usedcars$price)
quantile(usedcars$price)
quantile(usedcars$price,probs = c(0.01,0.99))
quantile(usedcars$price,seq(from=0, to=1, by=0.2))
```

```
#수치 변수 시각화
boxplot(usedcars$price, main="Boxplot of Used Car Prices", ylab="Price ($)")
boxplot(usedcars$mileage, main="Boxplot of Used Car Mileage", ylab="Odometer (mi.)")
hist(usedcars$price, main = "Histogram of Used Car Prices", xlab="Price ($)")
hist(usedcars$mileage, main="Boxplot of Used Car Mileage", ylab="Odometer (mi.)")


#정규분포의 중심위치와 퍼짐정도
var(usedcars$price)
sd(usedcars$price)
var(usedcars$mileage)
sd(usedcars$mileage)


#범주변수탐색
table(usedcars$year)
table(usedcars$model)
table(usedcars$color)
model_table <- table(usedcars$model)
prop.table(model_table)
color_table <- table(usedcars$color)
color_pct <- prop.table(color_table)*100
round(color_pct,digits = 1)
```

```
#변수관계탐색
#관계시각화
#산포도
plot(x=usedcars$mileage, y=usedcars$price,
     main="Scatterplot of Price vs.Mileage",
     xlab = "Used Car Odometer (mi.)",
     ylab = "Used Car Price ($)")

#이원교차표
#install.packages("gmodels")
library(gmodels)
usedcars$conservative <- usedcars$color %in% c("Black","Gray","Silver","White")
table(usedcars$conservative)
CrossTable(x=usedcars$model, y=usedcars$conservative, chisq = TRUE)
```

문제[3]

[ggplot2 :: mpg 이용]

1. Class(자동차 종류)가 “compact”, “subcompact”, “suv”인 자동차의 cty(도시 연비)가 어떻게 다른지 비교해보려고 한다. 세 차종의 cty를 나타낸 상자 그림을 작성하고 해석하시오. (class를 세 종류로 전처리 후 상자그림을 그리시오)

2. cty(도시 연비)와 hwy(고속도로 연비) 간에 어떤 관계가 있는지 알아보려고 한다. 산점도를 그리고 해석하시오.

3. 어떤 회사에서 생산한 "suv" 차종의 도시 연비가 높은지 알아보려고 한다. "suv" 차종을 대상으로 평균 cty(도시 연비)가 가장 높은 회사 다섯 곳을 막대 그래프로 작성하고, 막대는 연비가 높은 순으로 정렬하시오.

4. 자동차 중에서 어떤 class(자동차 종류)가 가장 많은지 알아보려고 한다. 자동차 종류별 빈도를 표현한 막대 그래프를 작성하시오.