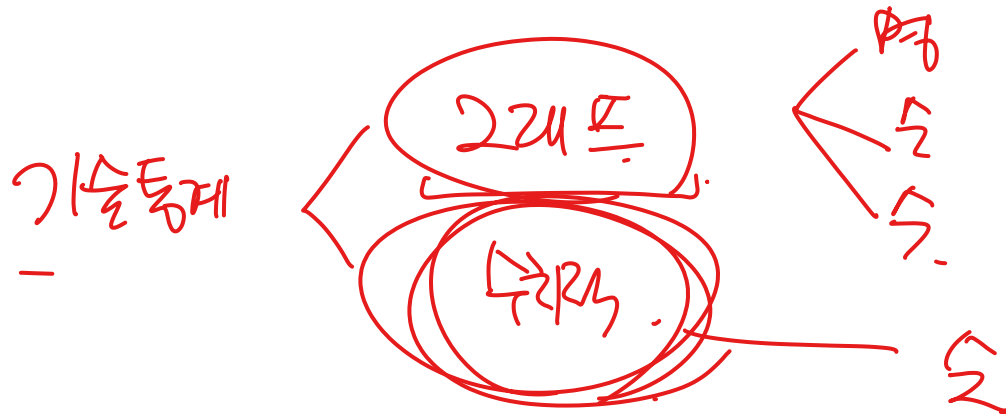


4. 표본 데이터의 요약 (1)



그래프에 의한 기술통계

• 그래프를 이용한 자료의 정리

- 히스토그램, 상자그림, 산점도, 파이그림, 막대그림 등을 이용하여 한눈에 알아볼 수 있게 자료를 정리할 수 있음.

- 질적 자료인 경우 (방법)
 - ◆ 1개 변수 : 막대그림, 파이그림 → 분포
 - ◆ 2개 변수 : 스택트컬럼차트, 히트맵 → 막대그림 (2개)
 - 양적 자료인 경우 (수량적)
 - ◆ 1개 변수 : 히스토그램, 상자그림 → 분포
 - ◆ 2개 변수 : 산점도 → 단상성
- 변량

그래프에 의한 기술통계

■ 도수분포표 (Frequency Distribution Table)

- ♦ 자료의 범주와 각 범주 별 빈도를 정리한 도수분포표를 작성한 뒤, 이를 토대로 자료를 막대그림, 파이그림, 히스토그램 등으로 표현
 - 질적 자료의 경우 범주를 그대로 표기
 - 양적 자료의 경우 각 자료값이 속하는 구간을 이용하여 범주를 정의함.

범주	빈도	상대빈도	누적빈도	누적상대빈도
30~50	30		30	0.3
50~70	50		50	0.5
70~90	30		80	0.8
90~100	20		100	1.0

100

그래프에 의한 기술통계

질량. 1개. 본질 : 양 타 한 시 공간.

■ 막대그림(Bar chart)과 파이그림(Pie chart)

◆ 라이트 맥주 예제

- 어느 맥주회사의 마케팅 관리자는 라이트 맥주 판매를 분석하고자 한다. 어느 마트 고객 중 285명의 표본을 무작위로 추출하여, 다음 7개 중 가장 좋아하는 라이트 맥주를 고르도록 하였다. 응답은 각각 코드 1, 2, 3, 4, 5, 6, 7을 사용하여 기록하였다.

- 1. Budweiser Light
- 2. Busch Light
- 3. Coors Light
- 4. Michelob Light
- 5. Miller Lite
- 6. Natural Light
- 7. Other brand

Number	Brand
1	1
2	1
3	5
4	1
...	...
283	1
284	1
285	5

방향

그래프에 의한 기술통계

- 라이트 맥주 데이터(light_beer_preference_survey.csv)를 이용하여 도수분포표를 작성한 뒤, 이를 막대그림과 파이그림을 이용하여 시각화하여라.

```
> setwd("C:\\...\\")  
> lightbeer <- read.csv('light_beer_preference_survey.csv')  
> head( lightbeer )
```

	Number	Brand	Gender
1	1	1	2
2	2	1	1
3	3	5	2
4	4	1	1
5	5	3	1
6	6	3	1

Handwritten diagram showing a vertical stack of circles containing the numbers 10, 15, and 20, followed by a plus sign and a circle containing the number 3.

Handwritten R code: `cbind(a, b, c)`

Handwritten diagram showing a 3x3 grid with columns labeled a, b, and c, and rows labeled 1, 2, and 3.

그래프에 의한 기술통계

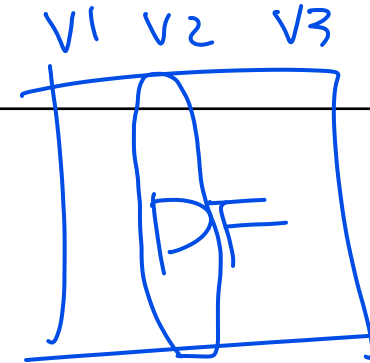
```
> brand <- lightbeer$Brand  
> brand_freq <- table(brand)  
> brand_freq
```

```
brand  
 1  2  3  4  5  6  7  
90 19 62 13 59 25 17
```

```
> freq_dist <- cbind(brand_freq, brand_freq/sum(brand_freq)*100)  
> colnames(freq_dist) <- c('freq', 'relative_freq')  
> freq_dist[,2] <- round(freq_dist[,2], 2)  
> freq_dist
```

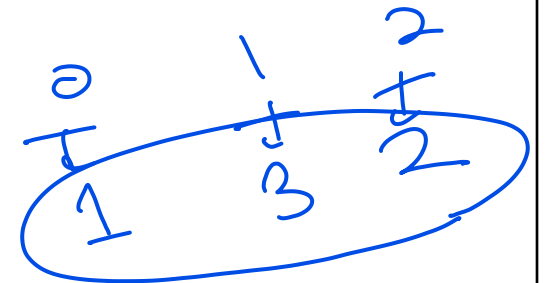
```
freq relative_freq  
1    90          31.58  
2    19           6.67  
3    62          21.75  
4    13           4.56  
5    59          20.70  
6    25           8.77  
7    17           5.96
```

PF \$ V2



MI[.2]

285 .
table(c(1,0,1,2,1))

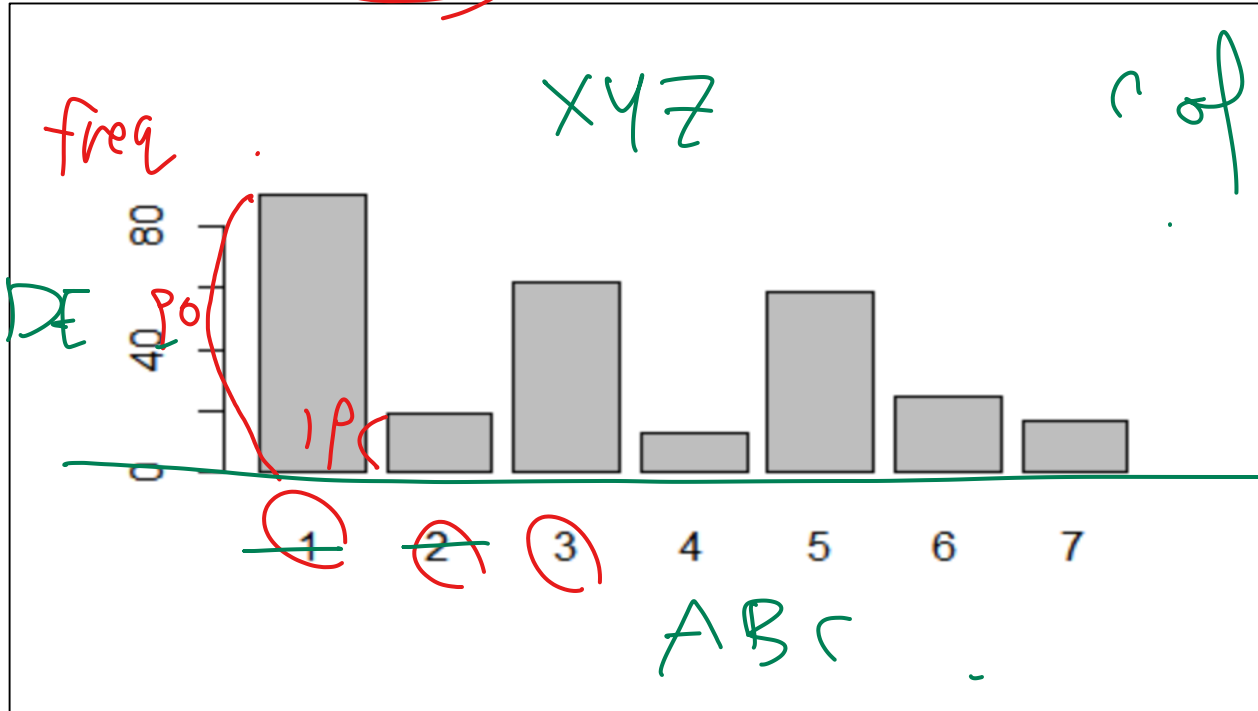


그래프에 의한 기술통계

xlab = 'ABC' , ylab = 'DE'

main = 'XYZ'

```
> barplot(brand_freq)
```



c of = c('red', 'green', ...)

names, arg =

c(" —", " —", ...)

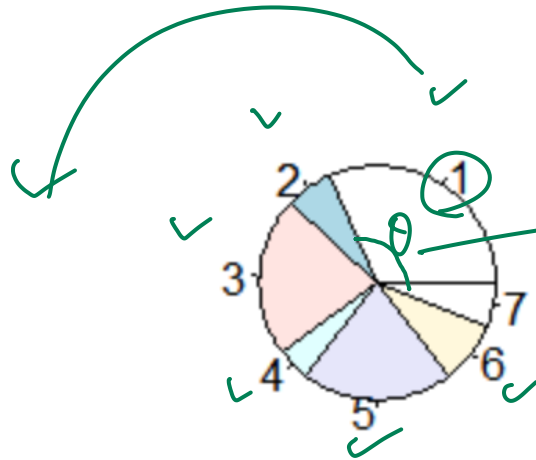
그래프에 의한 기술통계

label s = c(" —", ...)

> pie(brandfreq)

col = c(

상대빈도



$$\frac{8}{360} = 0.3158$$

그래프에 의한 기술통계

☆ 히스토그램 (Histogram)

◆ 통신요금 예제

- 한 통신회사는 적절한 요금체계를 마련하고자 그 회사와 계약한 신규 가입자 중 200명의 표본을 무작위로 추출한 뒤, 각 고객 별 첫 달의 월별 청구액을 다음과 같이 수집하였다.

Subscribers	Bills
1	42.19
2	38.45
3	29.23
4	89.35
5	118.04
...	...



Class Intervals	Frequency
0 ~ 15미만	71
15이상 30미만	37
30이상 45미만	13
45이상 60미만	9
60이상 75미만	10
75이상 90미만	18
90이상 105미만	28
105이상 120미만	14
합계	200

그래프에 의한 기술통계

- 통신요금 데이터(telephone_bills.csv)을 이용하여 월 청구액에 대한 히스토그램을 도출하고 이를 해석하여라.

```
> billdata <- read.csv('telephone_bills.csv')
```

```
> head( billdata )
```

Bills

1 42.19

2 38.45

3 29.23

4 89.35

5 118.04

6 110.46

```
> bills <- billdata$Bills
```

```
> table( cut(bills, breaks=seq(0, 120, 15), include.lowest=TRUE))
```

[0,15]

(15,30]

(30,45]

(45,60]

(60,75]

(75,90]

(90,105]

71

37

13

9

10

18

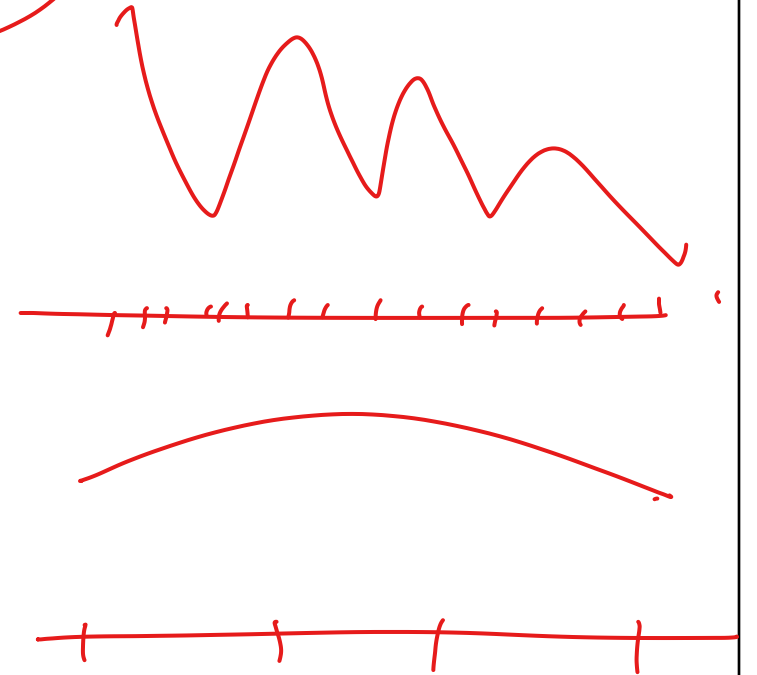
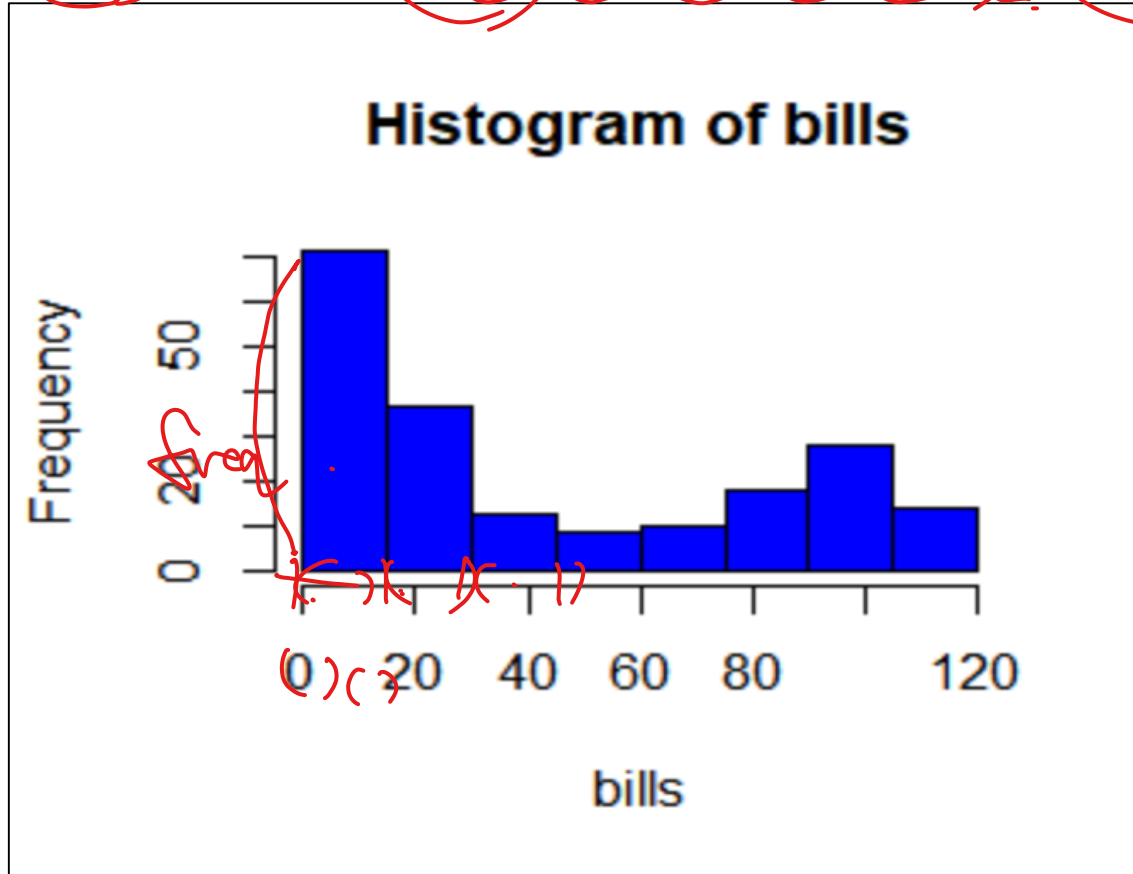
28

(105,120]

14

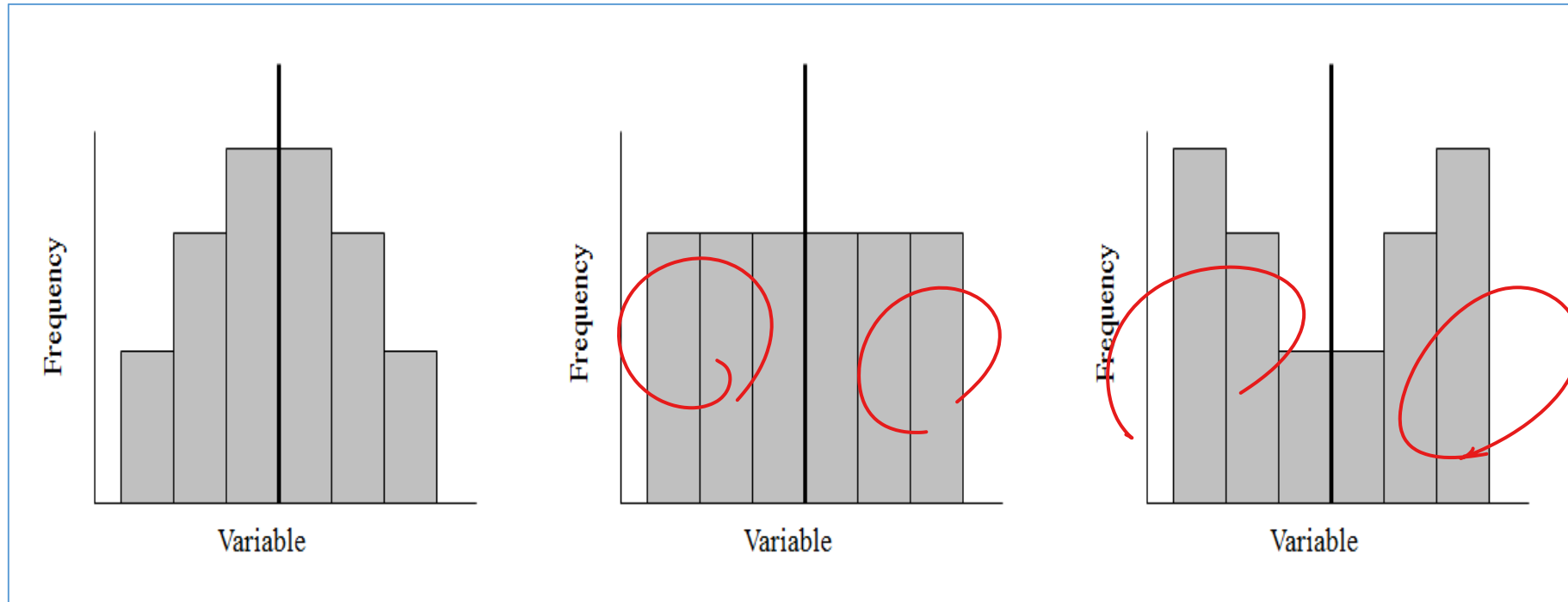
그래프에 의한 기술통계

> `hist(bills, breaks=seq(0, 120, 15), col='blue')`



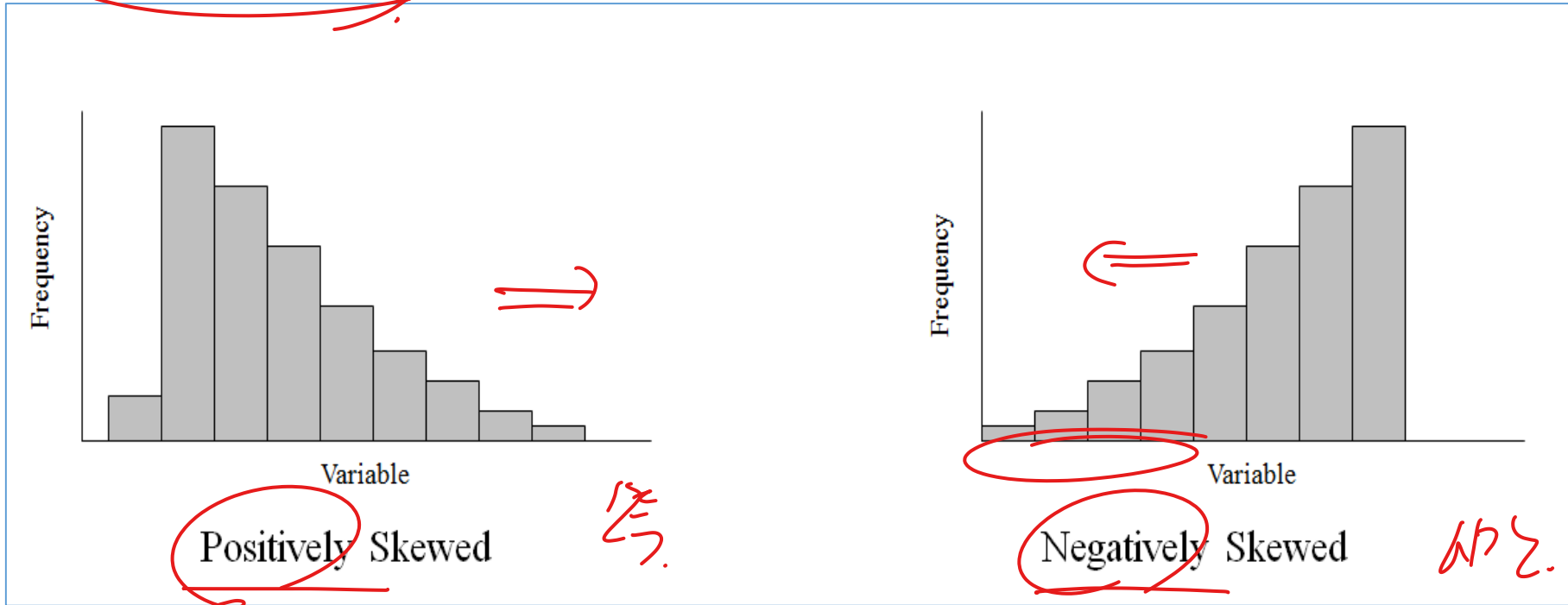
그래프에 의한 기술통계

◆ 대칭 (Symmetry)



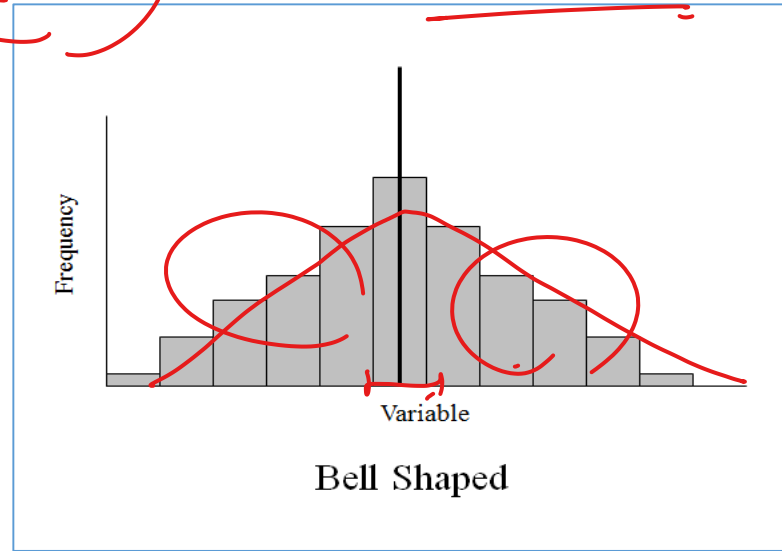
그래프에 의한 기술통계

◆ 치우침 (Skewness)



그래프에 의한 기술통계

- ♦ Symmetric and Unimodal → 종모양 (Bell Shaped)

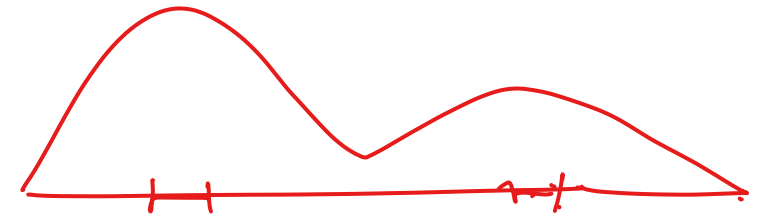


- ♦ 히스토그램 해석 시 주의할 점

- 계급 구간의 수에 따라 히스토그램의 모양이 많이 달라질 수 있음.
- 일반적으로 계급 구간의 수가 많으면 봉우리가 많아지고, 계급구간의 수가 적으면 봉우리의 수가 적어져 평평해짐.

mode : 최빈값

최빈 구간

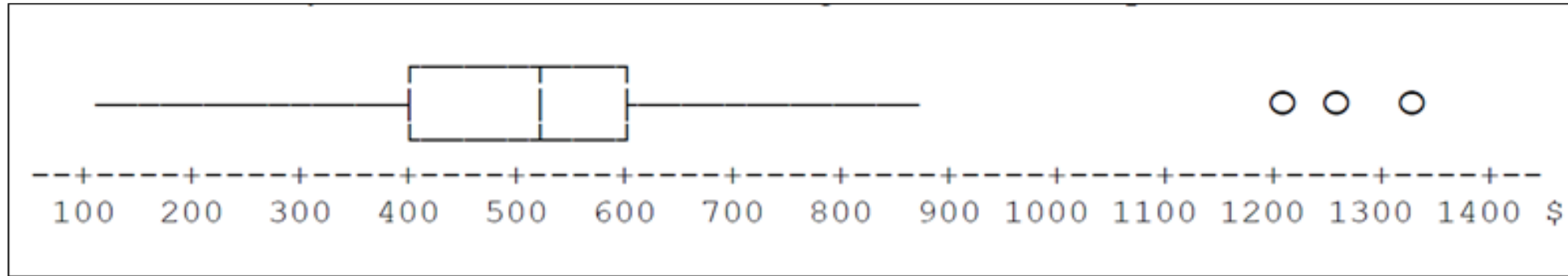


이봉

Bimodal

그래프에 의한 기술통계

▪ 상자수염그림 (Box-and-Whisker Plot)



◆ 상자수염그림 작성방법

- 1) 다섯수치요약값 Q_0 (minimum), Q_1 , Q_2 , Q_3 , Q_4 (maximum)과 IQR을 구한다.
- 2) Q_1 , Q_2 , Q_3 를 이용하여 상자를 작성한다.
- 3) 울타리(fences) 경계값인 $F_1 = Q_1 - 1.5 \times IQR$ 과 $F_2 = Q_3 + 1.5 \times IQR$ 를 구한다.
- 4) 울타리에 안쪽에 있는 자료값 중 가장 극단적인 두 값인 인접값(adjacent values) AV_1 과 AV_2 를 구한다.
- 5) Q_1 부터 AV_1 까지, Q_3 부터 AV_2 까지 수염(whisker)를 그린다.
- 6) 이상치는 양 울타리를 넘어가는 값들로 정의된다. 이상치가 있는 경우 해당 위치에 x, o 등으로 표기한다.

그래프에 의한 기술통계

- 다음은 15명의 학생들의 한달 용돈 자료이다. 이에 대한 상자그림을 작성하여라.
5, 8, 20, 24, 25, 26, 28, 30, 30, 30, 30, 35, 37, 40, 100.

그래프에 의한 기술통계

◆ 드라이브스루 예제

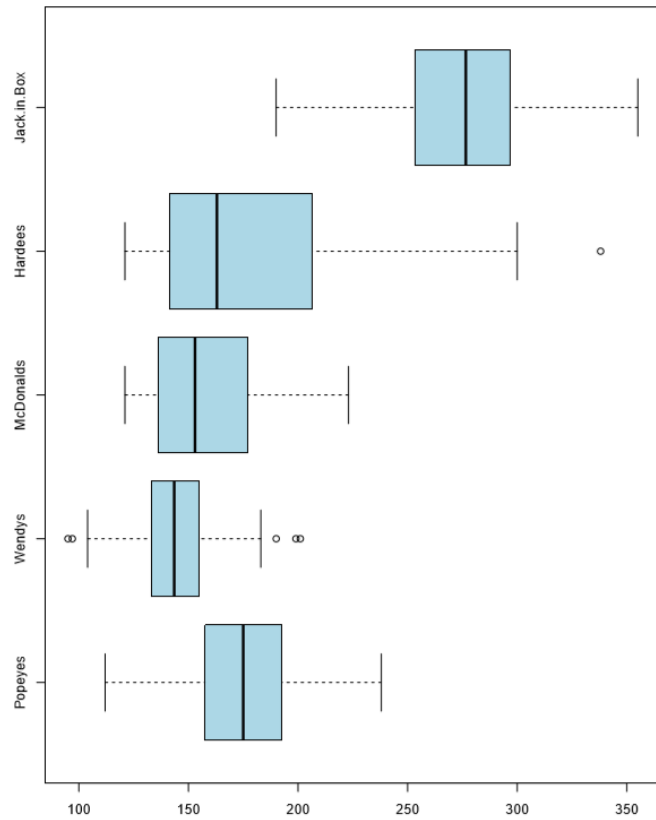
- 드라이브스루 서비스를 제공하는 패스트푸드 레스토랑 다섯 군데(Popeyes, Wendys, McDonalds, Hardees, Jack.in.Box)에서 드라이브스루 고객을 100명씩 무작위로 선택한 뒤 각 고객 별로 주문부터 픽업까지 소요된 시간을 기록하였다. 드라이브스루 데이터 (serving_times_of_drive_throughs.csv)에 기록된 데이터를 토대로 상자그림을 도출한 뒤 5개 레스토랑 별 소요시간이 어떠한가를 비교하여라.

```
> times <- read.csv('serving_times_of_drive_throughs.csv')
> head( times )
```

	Popeyes	Wendys	McDonalds	Hardees	Jack.in.Box
1	150	173	198	158	289
2	197	95	131	195	273
3	186	154	158	177	254
4	166	104	123	147	282
5	196	183	139	181	235
6	192	120	160	192	245

그래프에 의한 기술통계

```
> boxplot( times, horizontal=TRUE, col="lightblue" )
```



그래프에 의한 기술통계

■ 스택트컬럼차트 (Stacked Column Chart)

◆ 신문 예제

- 선호하는 신문과 직업군 간에 연관성이 있는지를 파악하고자 한다. 354명을 무작위로 선택하여 어느 직업군에 속하는지와 어느 신문을 선호하는지를 조사하였다. 선호하는 신문은 Globe and Mail(1), Post(2), Star(3), Sun(4) 중 하나로, 직업군은 blue-collar worker(1), white-collar worker(2), professional(3) 중 하나의 항목으로 입력하였다.

Reader	Occupation	Newspaper
1	2	2
2	1	4
3	2	1
...

그래프에 의한 기술통계

- ◆ 분할표(contingency table) : 두 개의 범주형 변수에 관한 요약

Contingency Table of Frequencies							
Reader	Newspaper	Occupation	Occupation				
			Newspaper	Blue Collar	White Collar	Professional	Total
1	2	2					
2	4	1	G&M	27	29	33	89
3	1	2	Post	18	43	51	112
.	.	.	Star	38	21	22	81
.	.	.	Sun	37	15	20	72
352	2	3					
353	3	1					
354	3	2	Total	120	108	126	354

그래프에 의한 기술통계

- 신문과 직업군에 관한 자료 (newspaper_readership_survey.csv)를 이용하여 분할표를 작성하고 이를 스택트컬럼차트를 이용하여 시각화하여라.

```
> news_reader <- read.csv('newspaper_readership_survey.csv')
> head( news_reader )
```

	Reader	Occupation	Newspaper
1	1	2	2
2	2	1	4
3	3	2	1
4	4	3	2
5	5	1	3
6	6	3	3

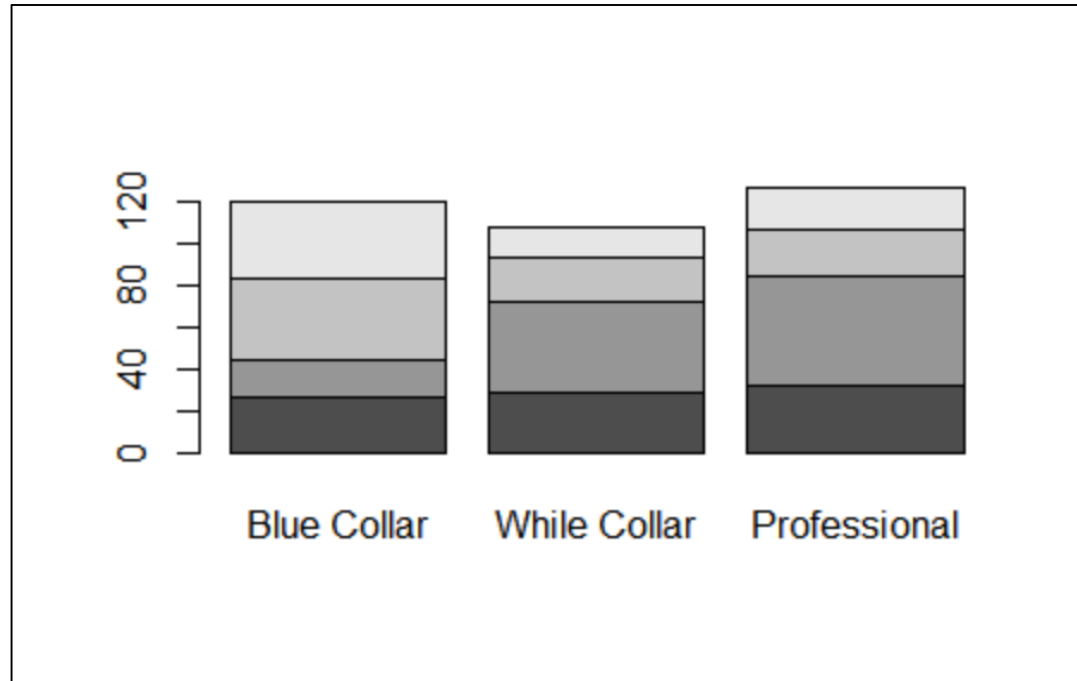
그래프에 의한 기술통계

```
> ctgc_tab <- table( news_reader$Newspaper, news_reader$Occupation )  
> colnames(ctgc_tab)<- c("Blue Collar", "White Collar", "Professional")  
> rownames(ctgc_tab)<- c("G&M", "Post", "Star", "Sun")  
> ctgc_tab
```

	Blue Collar	White Collar	Professional
G&M	27	29	33
Post	18	43	51
Star	38	21	22
Sun	37	15	20

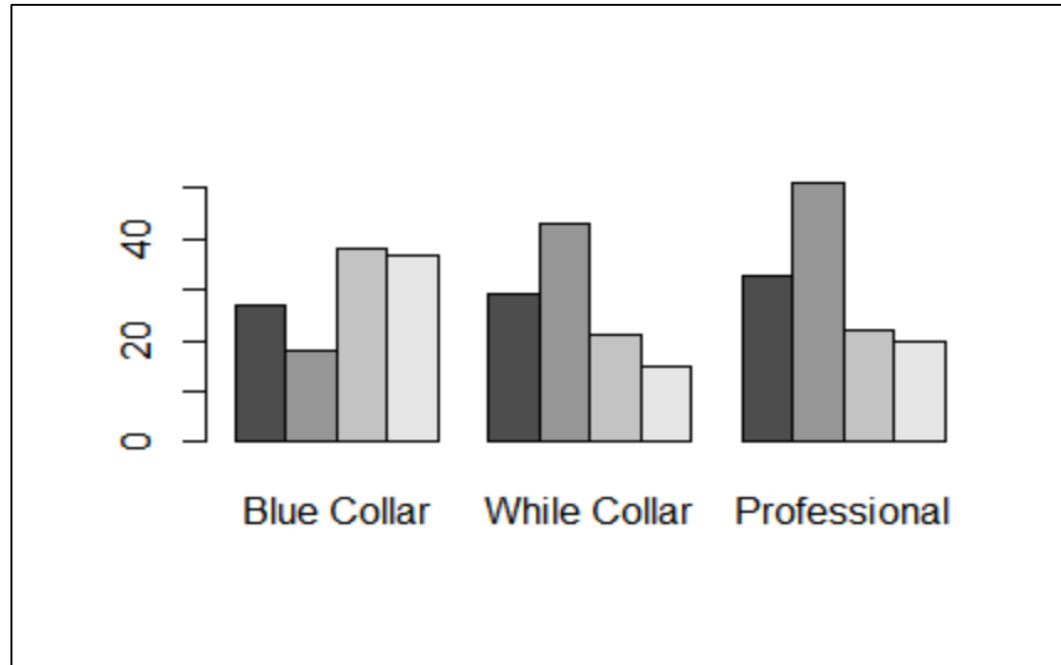
그래프에 의한 기술통계

```
> barplot( ctgc_tab )
```



그래프에 의한 기술통계

```
> barplot( ctgc_tab, beside=TRUE )
```



그래프에 의한 기술통계

▪ 산점도 (Scatter Plot)

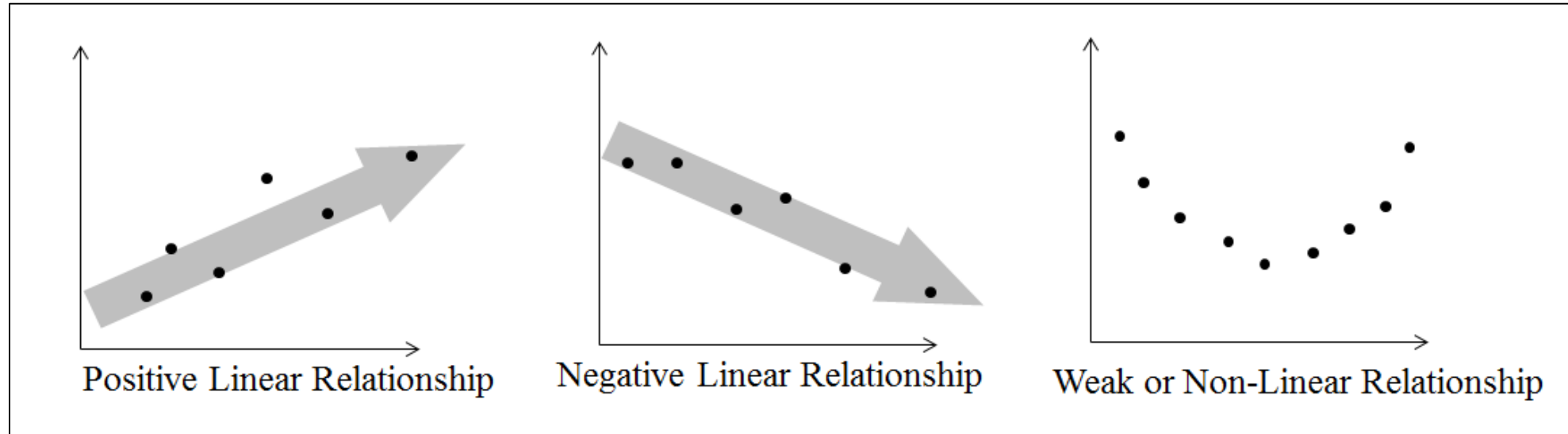
◆ 주택가격 예제

- 한 부동산 중개인은 집의 가격과 크기의 연관성에 대해 알아보고자 한다. 이를 위해 최근에 팔린 12채의 주택을 표본으로 뽑아 각 주택의 가격(단위: 1000달러)과 크기(단위: 100제곱피트)를 기록하였다.

Size	Price
23	315
18	229
26	355
20	261
22	234
14	216
33	308
28	306
23	289
20	204
27	265
18	195

그래프에 의한 기술통계

- ◆ 산점도에서 관찰되는 주요 패턴



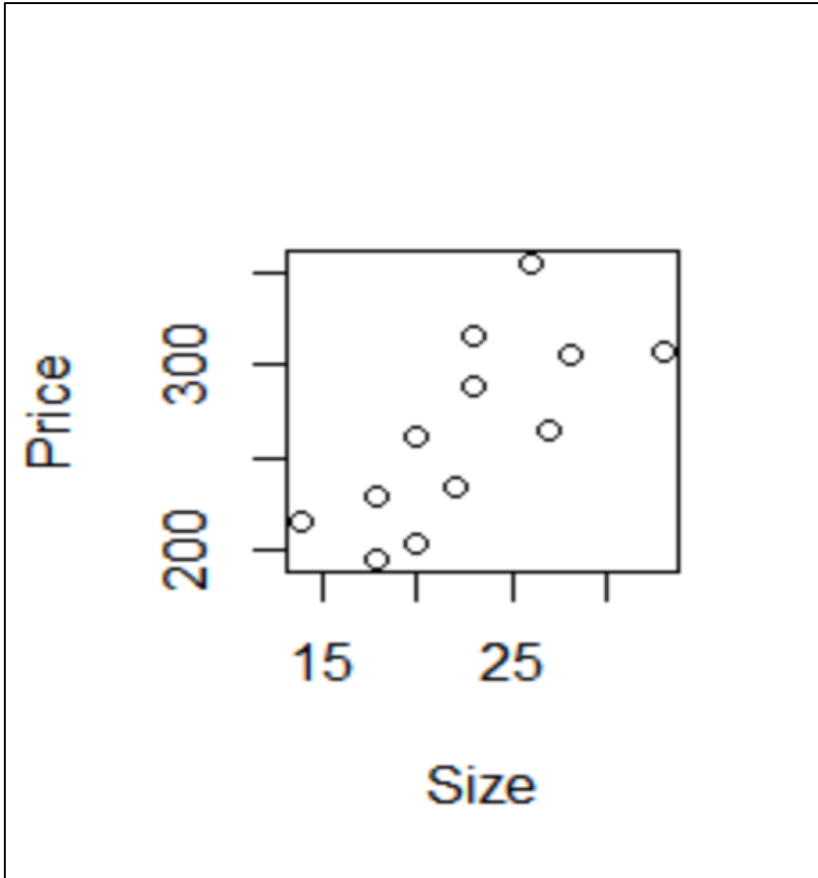
그래프에 의한 기술통계

- ♦ 주택가격 데이터(price_and_size_of_houses.csv)를 이용하여 주택 크기와 가격에 대한 산점도를 도출하여라.

```
> house <- read.csv("price_and_size_of_houses.csv")
> head( house )
  Size Price
1   23   315
2   18   229
3   26   355
4   20   261
5   22   234
6   14   216
```

그래프에 의한 기술통계

```
> plot( house$Size, house$Price, xlab="Size", ylab="Price" )
```



그래프에 의한 기술통계

■ 시계열 그림 (Time Series plot)

◆ 시계열 자료(Time-Series data) vs 횡단면 자료(Cross-Sectional data)

- 시계열자료 : 시간의 순서에 따라 기록한 자료
- 횡단면자료 : 동일한 시점에 수집된 자료

◆ 가솔린 가격 예제

- 1978년부터 매달 순차적으로 수집된 월평균 가솔린 가격(price_of_gasoline.csv) 자료로 시계열 그림을 도출하여라.

```
> gasoline <- read.csv('price_of_gasoline.csv')
> head( gasoline )
  Year Month Price
1 1978     1 0.631
2 1978     2 0.629
3 1978     3 0.629
4 1978     4 0.631
5 1978     5 0.637
6 1978     6 0.645
```

그래프에 의한 기술통계

```
> plot( gasoline$Price, type="l", xlab="Month", ylab="Price of Gasoline")
```

