사용자 정의 함수 만들기

- myfunc=function(arg1, arg2,...){
  표현식
  return(객체)
}

```
> y=5
> f(10)
[1] 15
> y=10
> f(10)
[1] 20
> x
Error: object 'x' not found
```

- 함수 안에서 정의된 x는 함수 밖에선 사용할 수 없음
- 함수 밖에서 정의된 y는 함수 안에서 사용할 수 있음

기술통계
Descriptive Statistics

개봉영화자료

- 2012년1월 부터 2013년10월9일 사이에 개봉된 영화
- 변수
  - 영화명
  - 개봉일
  - 첫주매출액
  - 첫주관객수
  - 대표국적
  - 제작사
  - 배급사
  - 등급
  - 장르
  - 총관객수
  - 총매출액

Q: 국내 개봉영화의 평균 관객수는?

통계적 추론(statistical inference)

모집단 - 특정 연구에서 관심의 대상이 되는 모든 요소들의 집합

표본 - 모집단의 부분집합

통계적 추론 - 표본으로부터 얻어진 자료를 분석하여 모집단의 특성을 추정하는 과정

전수조사 - 모집단에 대한 자료를 수집 조사하는 것

표본조사 - 표본에 대한 자료를 수집 조사하는 것

통계적 추론의 과정

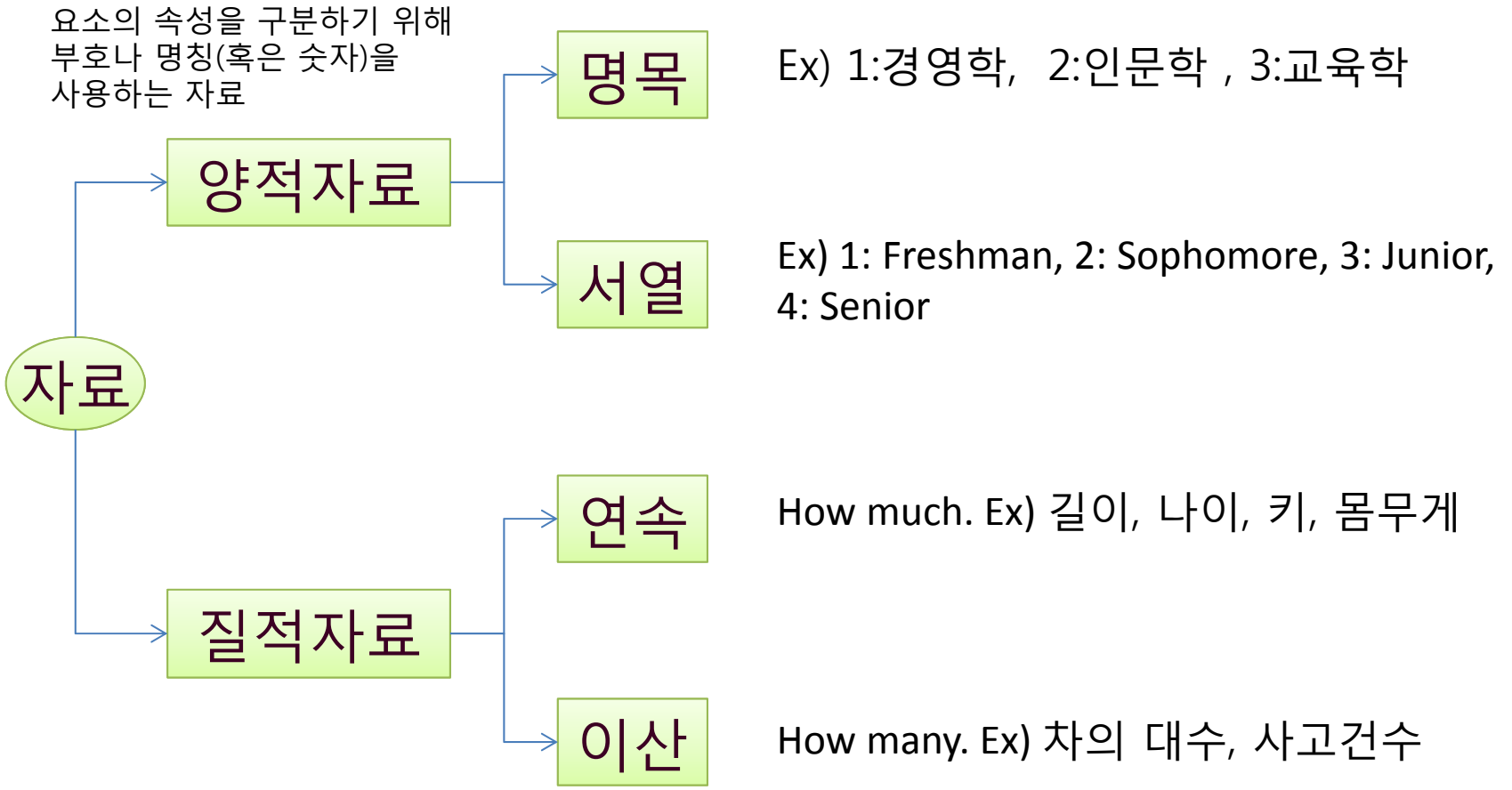
1. 모집단은 모든 국내개봉영화로 구성된다. 평균 관객수는 알려져 있지 않다.

2. 227개의 개봉영화 표본을 조사한다.

3. 표본자료에서 평균관객수는 152만6821명이다.

4. 표본평균으로 모집단의 평균을 추정한다.

측정척도의 종류



Ex) M&M의 색깔, 기온, 눈색깔, 셔츠 사이즈 (S,M,L,XL), 빨간색 M&M 개수

측정척도의 종류 (예)

M&M의 색깔

기온

눈색깔

셔츠 사이즈 (S,M,L,XL)

빨간색 M&M의 개수

양적자료가 아닌것..?

- 범주를 나타내는 숫자
  i.e. 1-male, 0-female

- 측정한 자료가 아니라 label을 나타내는 숫자
  i.e. your University ID#.

Hint: **값들의 평균이 의미가 있는지 체크!**
i.e. 0.5 = (male)과 (female)의 평균??

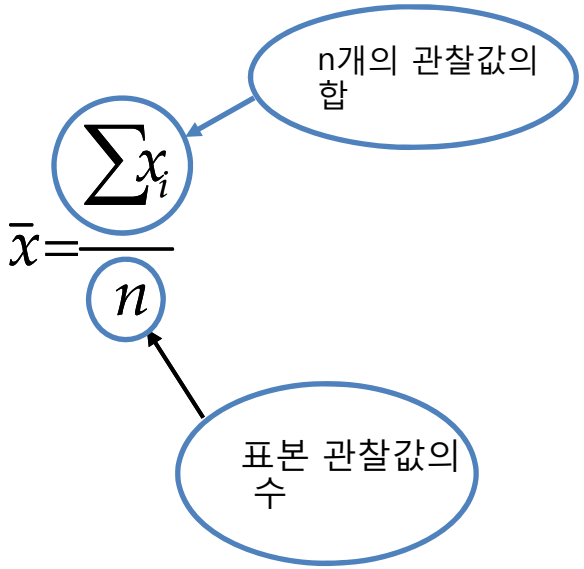
양적자료의 요약

양적자료의 요약

	척도	R 명령어
위치 척도	평균	mean()
	중위수	median()
	사분위수	quantile()
변동성 척도	분산	var()
	표준편차	sd()
	변동계수	sd()/mean()
그래프	boxplot	boxplot()
	히스토그램	hist()
	Q-Q plot	qqnorm(), qqline()

위치척도

- 평균 (Mean)



- 중위수 (Median)
  - 자료가 순서대로(오름차순) 배열되어 있을 때 중앙에 있는 값
  - 자료의 수가 홀수 일 경우 i=n/2를 반올림 한 후 i번째의 값
  - 자료의 수가 짝수 일 경우 i=n/2 번 째와 i+1번째의 평균

위치척도

- 사분위수 (Quartiles)

  자료를 순서대로 나열했을 때
  - Q1: 25%
  - Q2: 50%
  - Q3: 75% 에 위치하는 수

위치척도: Tips data

- tips 데이터 : reshape 패키지
  - 한 레스토랑의 웨이터가 몇 달간 받은 팁을 기록

```
> library(reshape)
> attach(tips)
The following objects are masked from tips (position 3):

    day, sex, size, smoker, time, tip, total_bill
> str(tips)
'data.frame': 244 obs. of  7 variables:
 $ total_bill: num  17 10.3 21 23.7 24.6 ...
 $ tip       : num  1.01 1.66 3.5 3.31 3.61 4.71 2 3.12 1.96 3.23 ...
 $ sex       : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 2 2 2 2 2 ...
 $ smoker    : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ day       : Factor w/ 4 levels "Fri","Sat","Sun",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ time      : Factor w/ 2 levels "Dinner","Lunch": 1 1 1 1 1 1 1 1 1 1 ...
 $ size      : int  2 3 3 2 4 4 2 4 2 2 ...
```

위치척도: Tips data

```
> attach(tips)
The following objects are masked from tips (position 3):

    day, sex, size, smoker, time, tip, total_bill
The following objects are masked from tips (position 4):

    day, sex, size, smoker, time, tip, total_bill
> mean(tip)
[1] 2.998279
> median(tip)
[1] 2.9
> sort(tip)[c(122,123)]
[1] 2.88 2.92
```

오름차순으로 정렬하여
122번째와 123번째
관찰치의 평균이 중위수

위치척도: 평균과 중위수의 비교

```
> tip2=tip
> max(tip)
[1] 10
> tip2[1]=100
> mean(tip2)
[1] 3.403975
> median(tip2)
[1] 2.96
```

첫번째 관찰치를 100으로 바꿈
평균은 커졌으나 중위수는
그대로

- 자료에 극단값이 포함되어 있을 경우, 중위수는 중심위치를 측정하는 데에 있어서 선호
- 연소득이나 재산 자료에서는 중위수가 위치척도로 자주 사용

위치척도: 사분위수- Tips data

```
> quantile(tip)
     0%      25%      50%      75%     100%
 1.0000   2.0000   2.9000   3.5625 10.0000
>
> quantile(tip,seq(0,1,0.1))
    0%    10%    20%    30%    40%    50%    60%    70%    80%    90%   100%
 1.000  1.500  2.000  2.000  2.476  2.900  3.016  3.480  4.000  5.000 10.000
```

함수명	내용
seq(a,b,d)	a 부터 b까지 d 간격씩 떨어진 벡터 생성

변동성 척도

- 분산 (variance) $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$

- 표준편차 (standard deviation): 원래의 자료에서 사용된 단위와 동일한 단위로 측정되므로 분산보다 해석이 용이하다.

$$s = \sqrt{s^2}$$

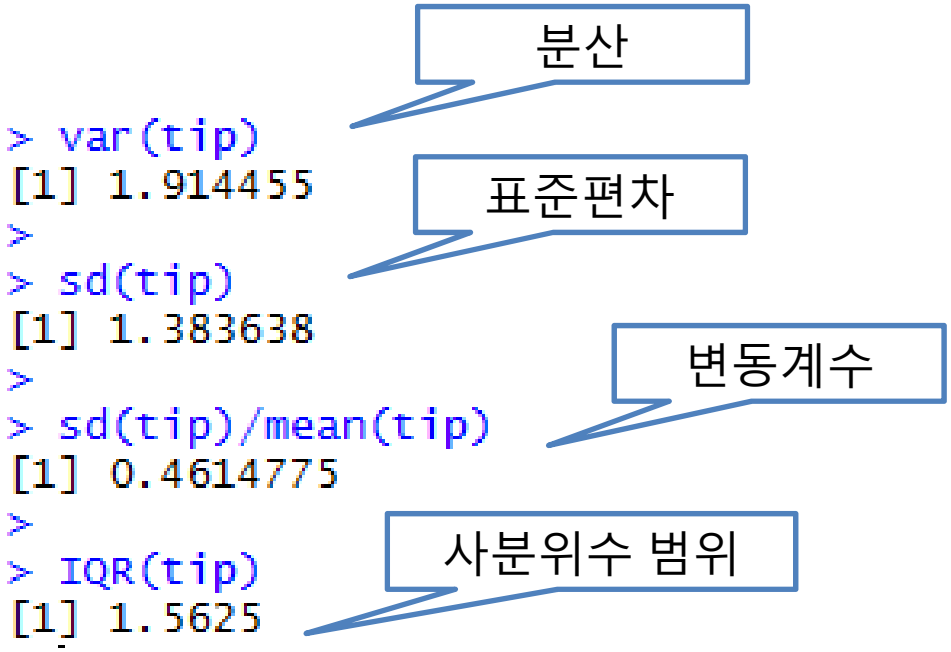
- 변동계수 (Coefficient of Variation): 표준편차가 평균에 비하여 얼마나 큰지를 나타낸다.

$$\frac{s}{\bar{x}}$$

변동성 척도

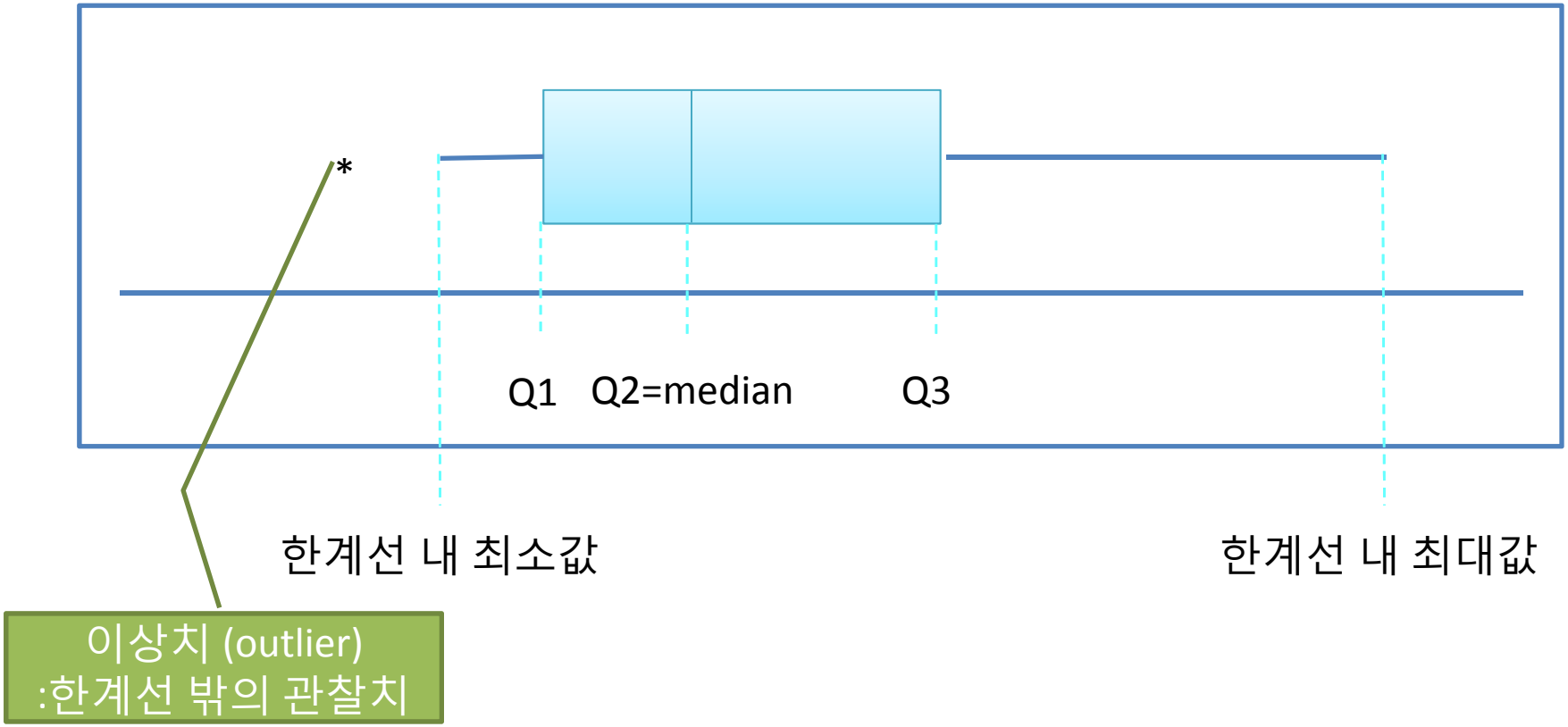
- 사분위수 범위 (interquartile range; IQR)
  - Q1과 Q3의 차이
  - 자료의 중간 50%의 범위
  - 분산, 표준편차 등은 극단값에 민감한데 비해 IQR은 상대적으로 덜 민감

변동성 척도: Tips data



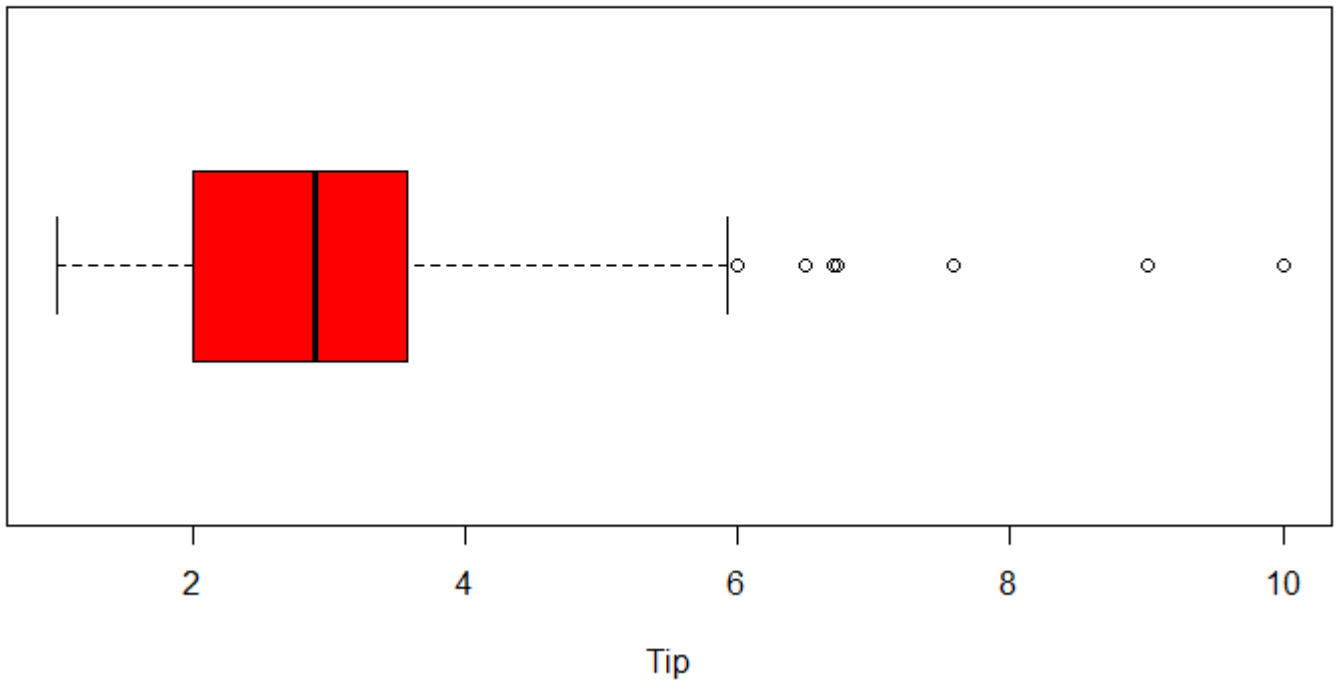
Boxplot

- 하한선: Q1 - 1.5(IQR)
- 상한선: Q3 + 1.5(IQR)



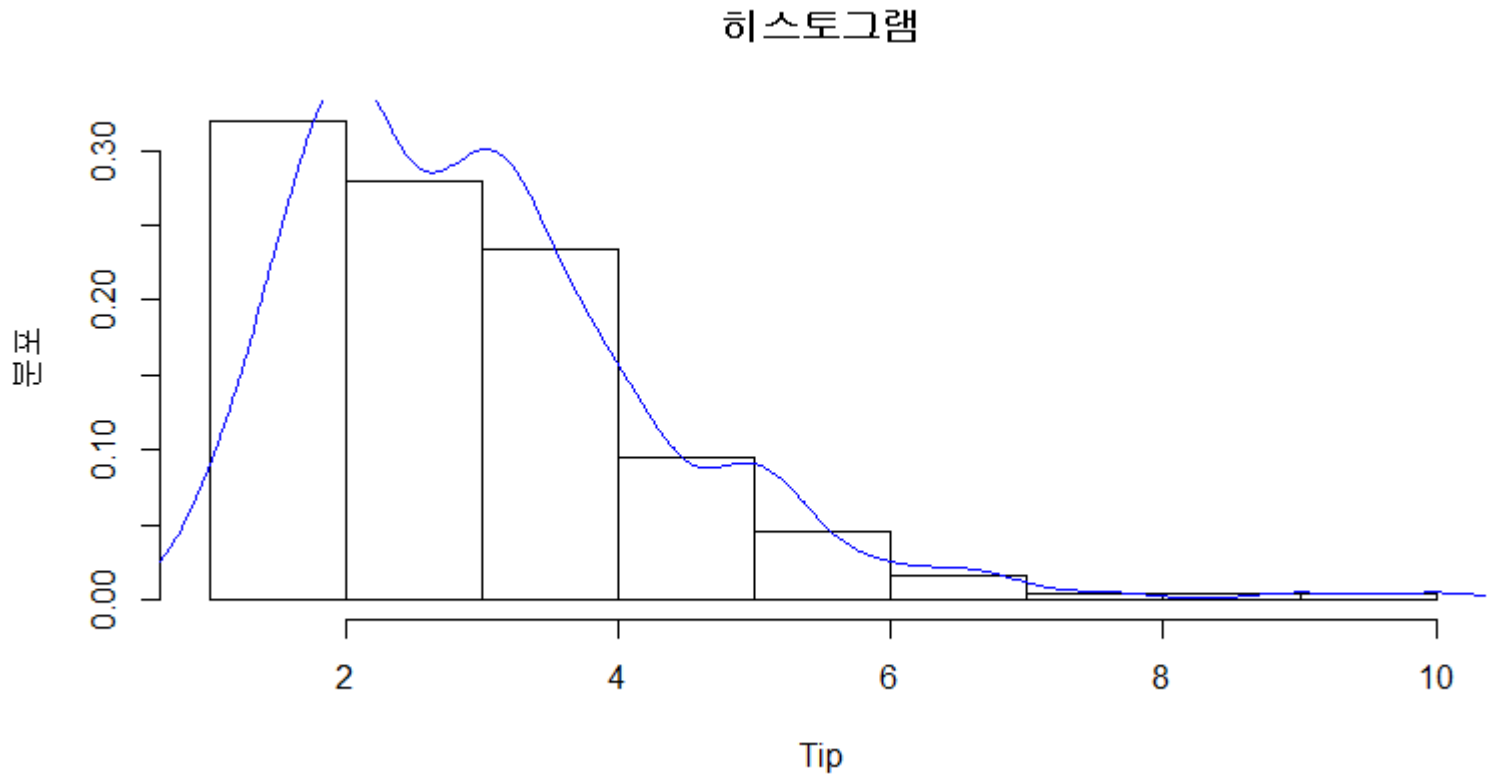
Boxplot: Tips data

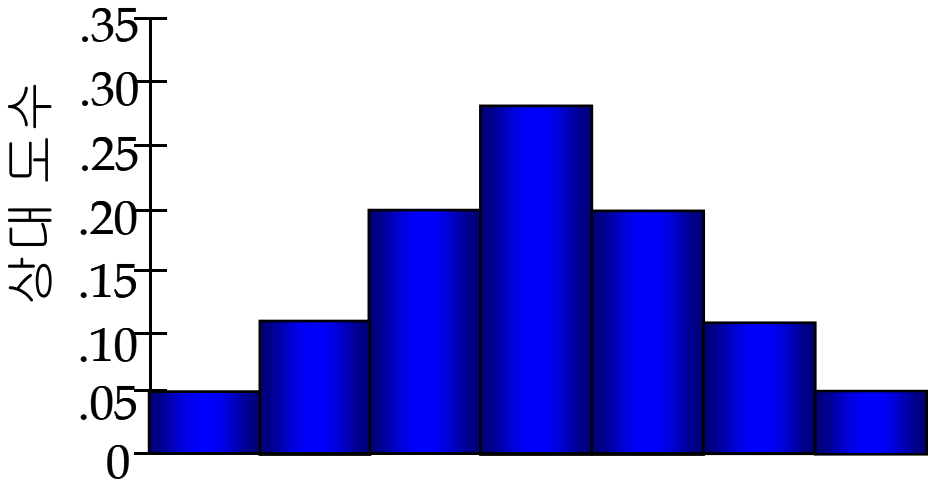
```
> boxplot(tip,col="red",horizontal=TRUE,xlab="Tip")
```



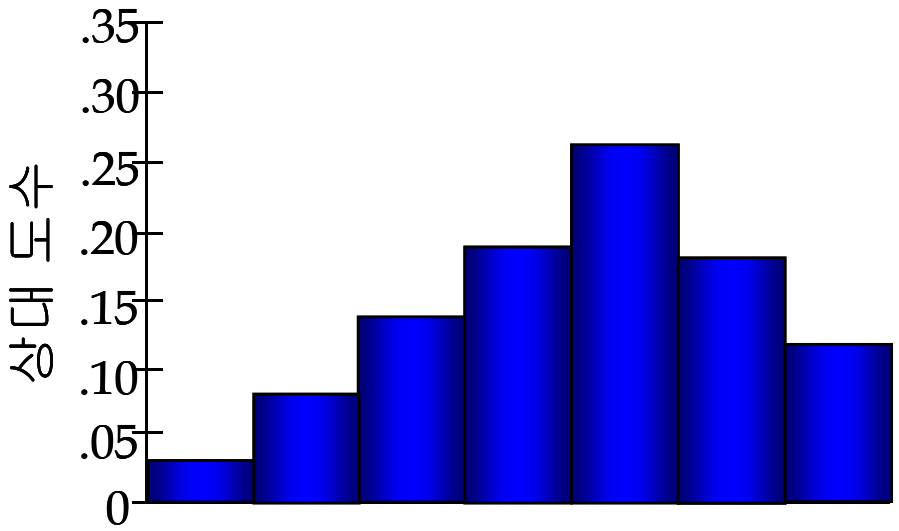
Histogram

```
> hist(tip,probability=TRUE,main="히스토그램",xlab="Tip",ylab="분포")
> lines(density(tip),col="blue")
```

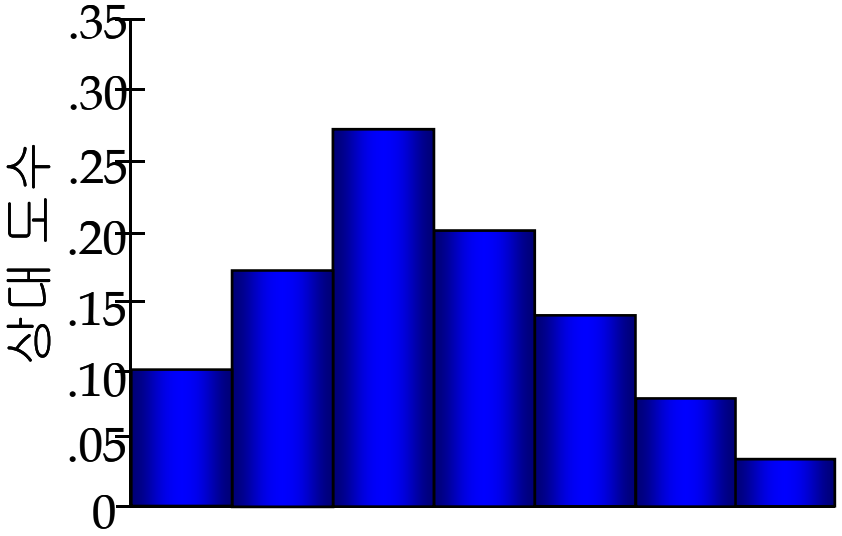




대칭: 정규분포에 가까움



왼쪽으로 경사진 히스토그램

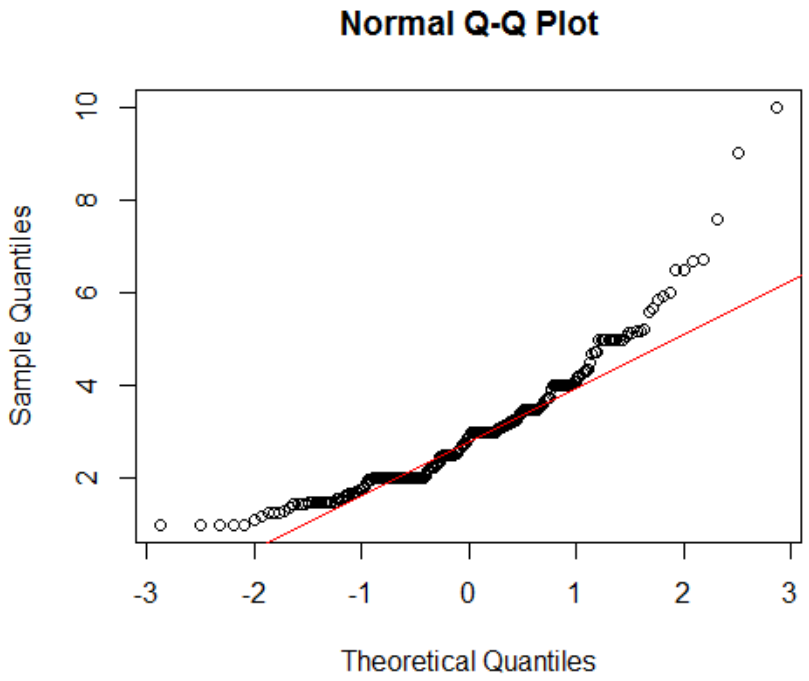


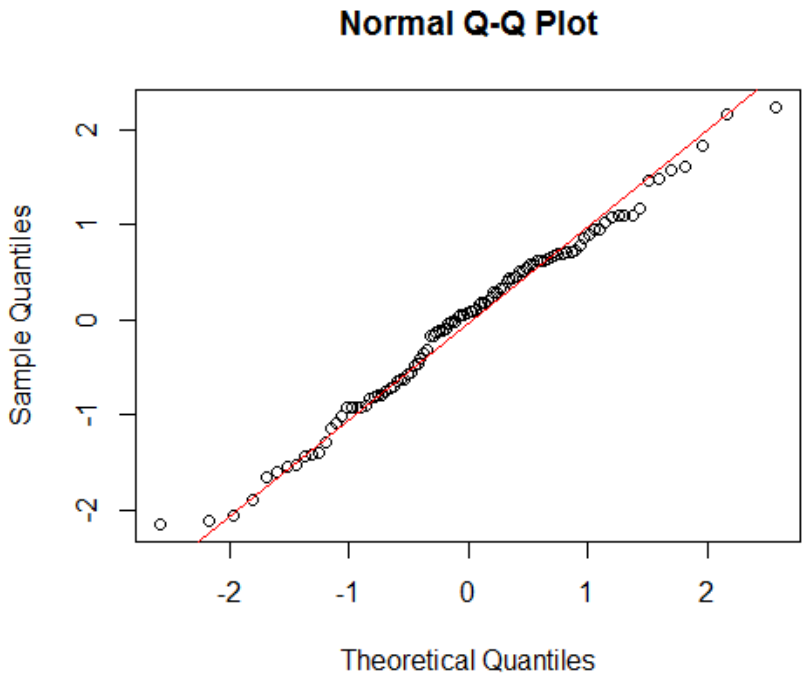
오른쪽으로 경사진 히스토그램

Q-Q Normality Plot

- 자료가 정규분포에 얼마나 근접한지 판단

```
> qqnorm(tip)
> qqline(tip,col=2)
>
> x=rnorm(100)
> qqnorm(x)
> qqline(x,col=2)
```





질적자료의 요약

질적자료의 요약

방법	R 함수명
도수분포표	table()
Bar plot	barplot()
Pie chart	pie()
분할표	xtabs()

예 : Marada Inn

Marada 여관에 투숙한 손님들은 숙박시설에 대하여 평가해줄 것을 요구 받는데, 평가 등급은 ***excellent***, ***above average***, ***average***, ***below average***, ***Poor***이다 . 20명의 표본 손님들에게서 받은 평가 내용이 아래와 같이 나타나 있다:

Below Average	Average	Above Average
Above Average	Above Average	Above Average
Above Average	Below Average	Below Average
Average	Poor	Poor
Above Average	Excellent	Above Average
Average	Above Average	Average
Above Average	Average	Average

도수분포표

등급	도수
Poor	2
Below Average	3
Average	5
Above Average	9
Excellent	1
계	20

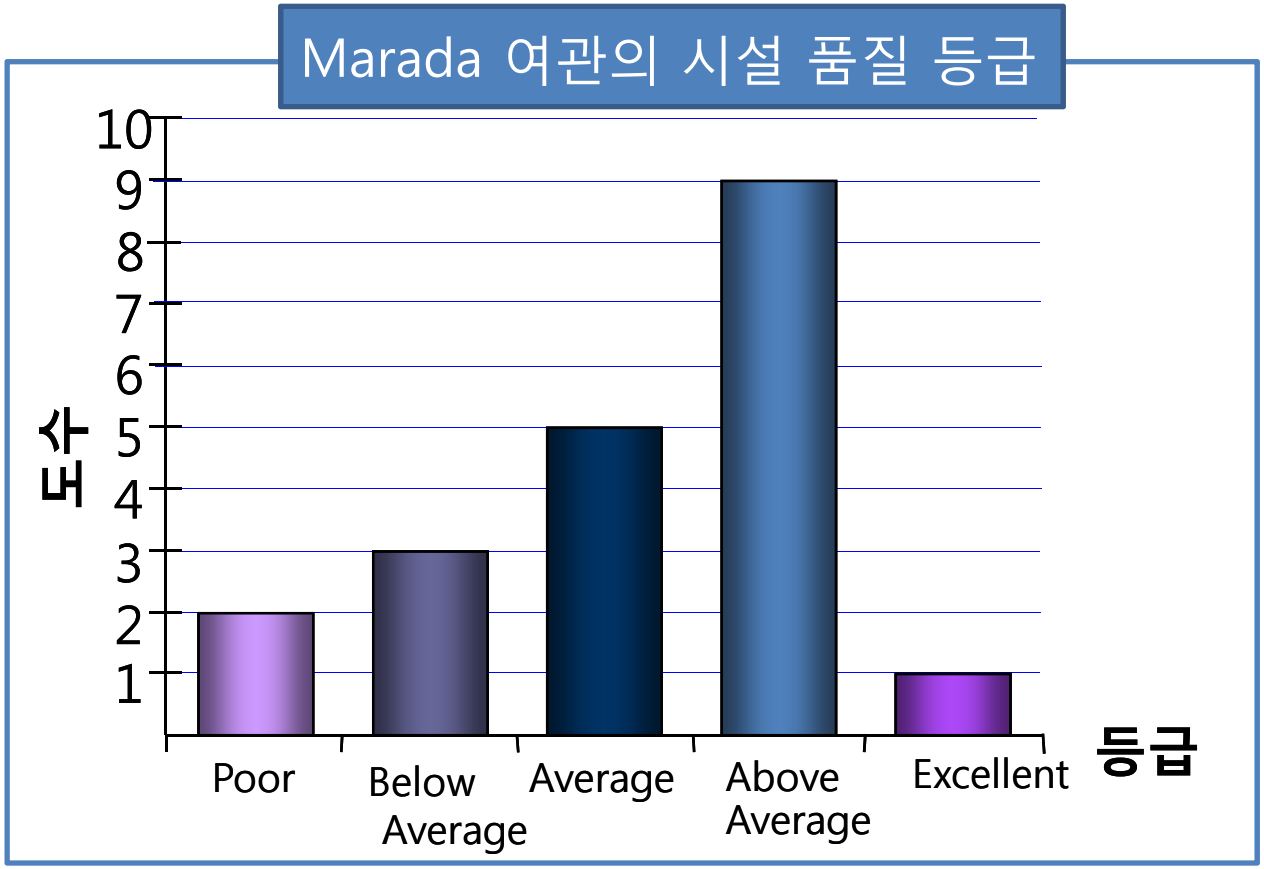
상대 도수와 백분율 도수 분포

등급	상대 도수	백분율도수
Poor	.10	10
Below Average	.15	15
Average	.25	25
Above Average	.45	45
Excellent	.05	5
계	1.00	100

.10(100) = 10

1/20 = .05

막대 그래프



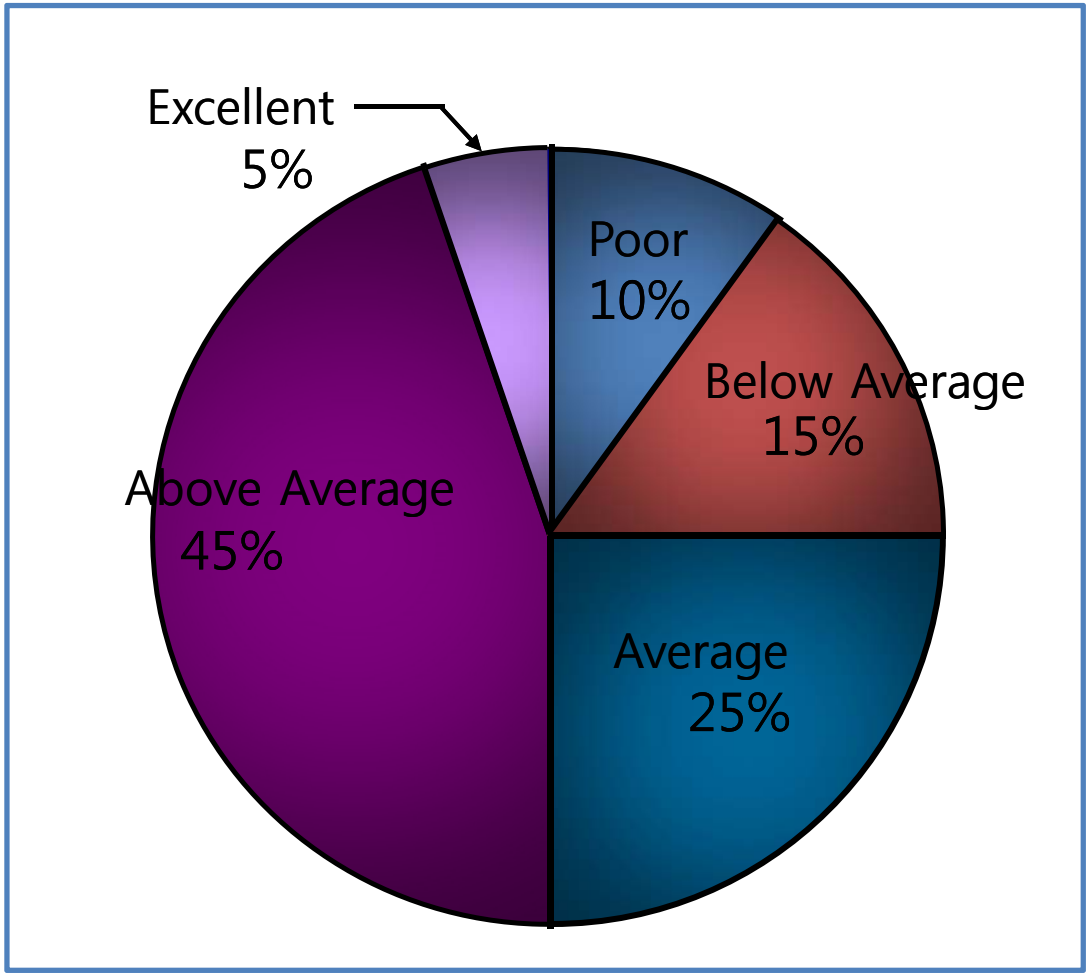
가로축: 계급의 이름

세로축: 도수분포,
상대도수분포,
백분율도수분포의
크기(scale)

각 계급이 분리되어
있다는 것을
강조하기 위해서
막대는 서로
분리되어 있어야
한다

파이차트

Marada 여관의 시설 품질 등급



도수분포표: Tips data

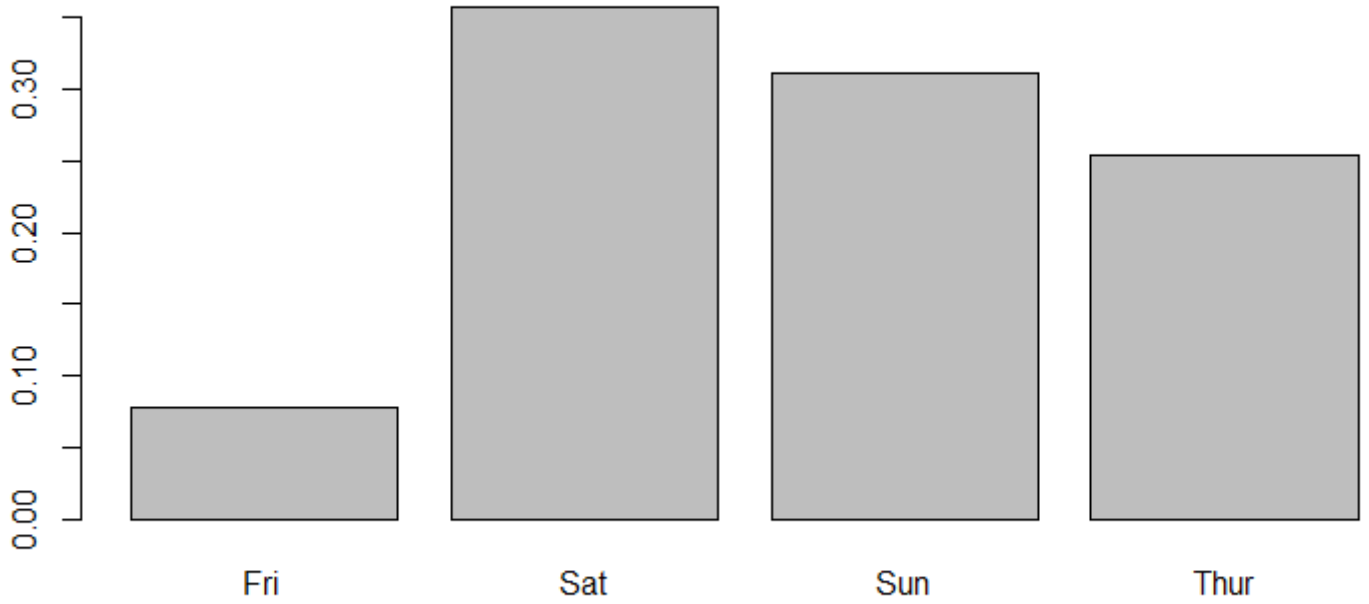
- table(): 질적변수의 도수분포표 출력
- summary(): 질적변수는 도수분포표, 양적변수는 기초통계량 출력

```
> mytable=table(day)
> mytable
day
 Fri  Sat  Sun Thur
  19   87   76   62
> summary(tips)
   total_bill        tip           sex      smoker      day         time
 Min.   : 3.07   Min.   : 1.000   Female: 87   No :151   Fri :19   Dinner:176
 1st Qu.:13.35   1st Qu.: 2.000   Male  :157   Yes: 93   Sat :87   Lunch : 68
 Median :17.80   Median : 2.900                        Sun :76
 Mean   :19.79   Mean   : 2.998                        Thur:62
 3rd Qu.:24.13   3rd Qu.: 3.562
 Max.   :50.81   Max.   :10.000
      size
 Min.   :1.00
 1st Qu.:2.00
 Median :2.00
 Mean   :2.57
 3rd Qu.:3.00
 Max.   :6.00
```

Bar plot: Tips data

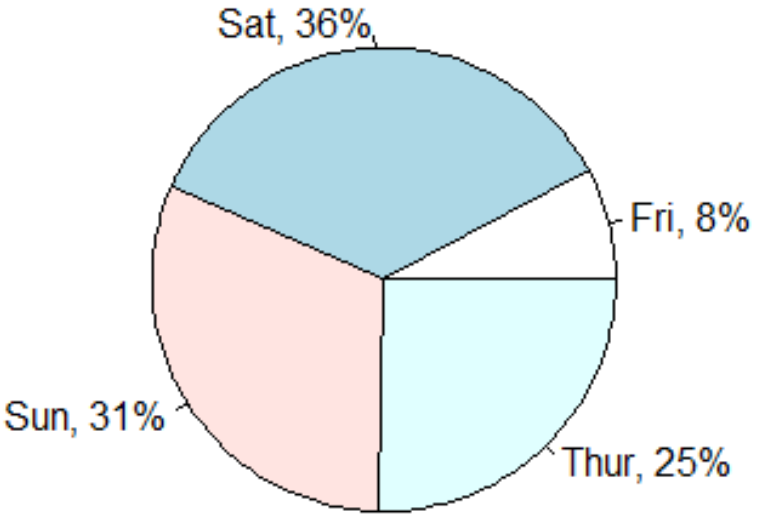
names.arg=c(“name1”,”names2”,...)
옵션으로 bar의 라벨 변경 가능

```
> barplot(mytable/sum(mytable))
```



Pie chart

```
> lbl=paste(names(mytable),", ", round(mytable/sum(mytable)*100),"%",sep="")
>
> lbl
[1] "Fri, 8%"   "Sat, 36%"  "Sun, 31%"  "Thur, 25%"
> pie(mytable,labels=lbl)
```



함수명	내용
paste(a,b,sep=“”)	a와 b를 sep에 지정된 기호로 이어 문자열을 만든다 (여기서 sep=“”를 지정하면 공백 없이 이음)

분할표 (Contingency Table)

- 두 개의 범주형 자료의 요약
- xtabs(~그룹변수1+그룹변수2,data)

```
> head(tips)
  total_bill  tip    sex smoker day   time size
1      16.99 1.01 Female     No Sun Dinner    2
2      10.34 1.66   Male     No Sun Dinner    3
3      21.01 3.50   Male     No Sun Dinner    3
4      23.68 3.31   Male     No Sun Dinner    2
5      24.59 3.61 Female     No Sun Dinner    4
6      25.29 4.71   Male     No Sun Dinner    4
```

```
> mytable2=xtabs(~sex+day,tips)
> mytable2
        day
sex      Fri Sat Sun Thur
  Female   9  28  18   32
  Male    10  59  58   30
```

```
> barplot(mytable2,legend.text=c("Female","Male"))
> barplot(mytable2,legend.text=c("Female","Male"),beside=TRUE)
```