경영통계 데이터 요약 및 정리

측정척도의 종류

요소의 속성을 구분하기 위해 부호나 명칭(혹은 숫자)을 사용하는 자료

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

중심위치측도: 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 ...
```

중심위치측도: 평균, 중위수

- 평균: $\bar{x} = \sum x_i / n$
- 중위수: 자료를 오름차순으로 정렬했을 때 가장 가운데 값

```
> 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

중심위치측도 : 평균, 중위수

```
> 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

중심위치측도 : 사분위수

```
> 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
```

평균의 함정

뉴스 > 경제

삼성전자 평균 연봉 1억…현대차 9천600만원

송욱 기자
입력 : 2016.04.01 04:57

520

0 0

삼성전자의 1인당 직원 평균 연봉이 국내 10대 기업 중 유일하게 1억원을 넘어선 것으로 나타났습니다.

연합뉴스

대졸 신입 초임 월 290만원…작년보다 4.5%↑

본문듣기 | 설정

기사입력 2015-10-25 11:00 | 최종수정 2015-10-25 14:04 기사원문 301 13

평균 임금인상률 5.0%…3.2% 포인트 하락

(서울=연합뉴스) 김윤구 기자 = 올해 4년제 대졸 신입사원 초임은 상여금 포함 월 290만9천원인 것으로 조사됐다. 이는 지난해의 278만4천원보다 4.5% 증가한 금액이다.

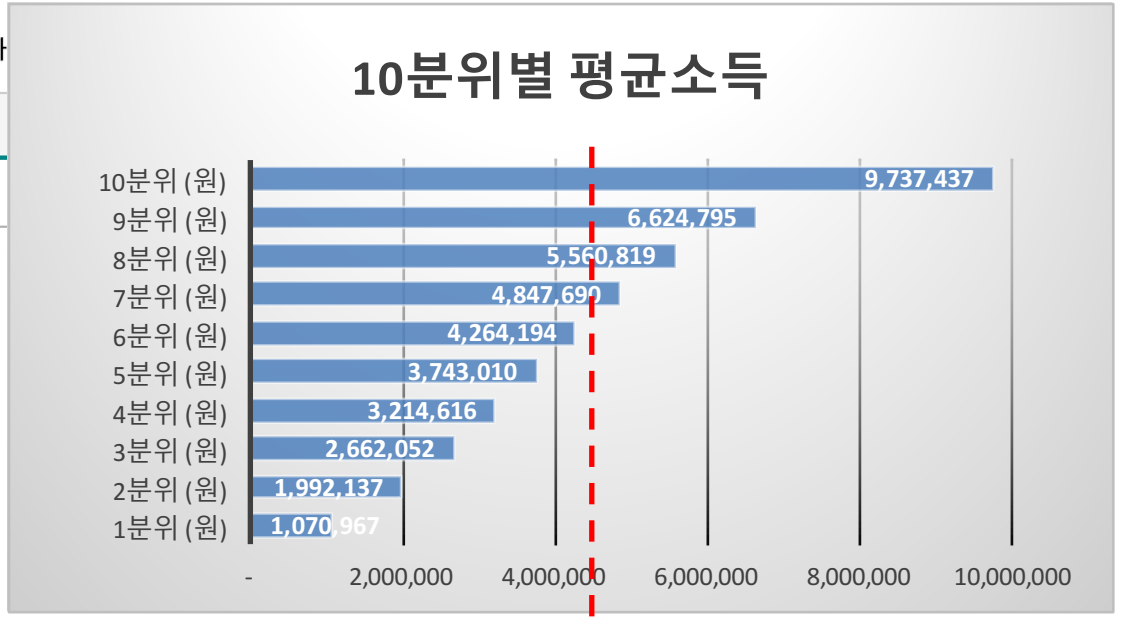
한국경영자총협회가 414개 기업을 대상으로 최근 실시한 '2015년 임금조정 실태조사'에서 이같은 결과가 나왔다.

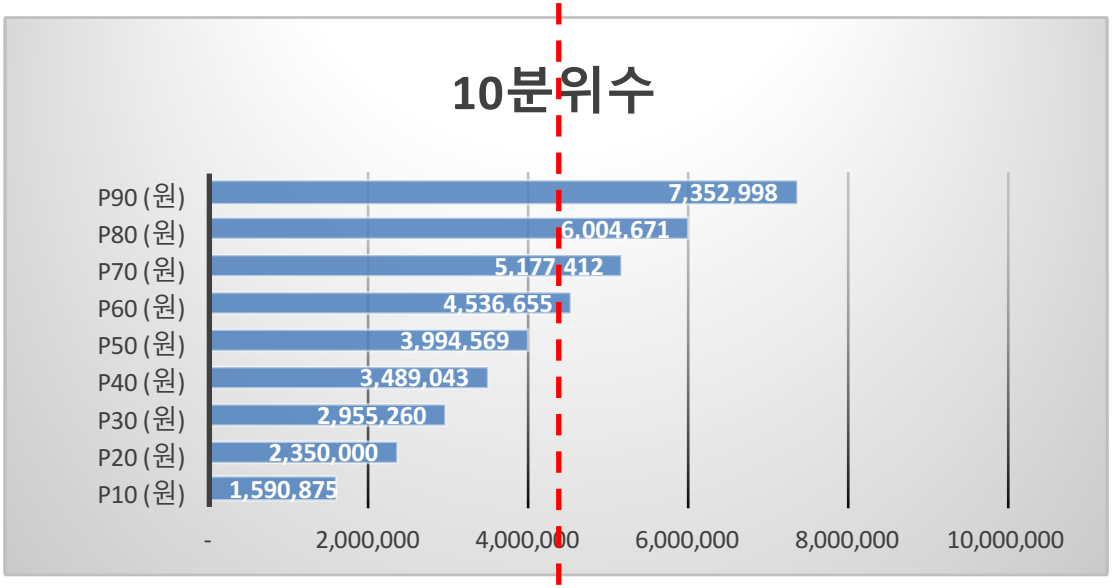
1) 가구당 월평균 분위경계값 및 적자가구비율(2인이상,가구기준)

자료갱신일 : 2016-05-27 / 수록기간 : 분기, 년 1990 1/4 ~ 2016 1/4 / 자

일괄설정 + | 항목[1/3] | 지표별(22/36)

지표별	2015 전국(2인이상 비농가)
10분위별평균소득 (원)	
전체 (원)	4,373,116
1분위 (원)	1,070,967
2분위 (원)	1,992,137
3분위 (원)	2,662,052
4분위 (원)	3,214,616
5분위 (원)	3,743,010
6분위 (원)	4,264,194
7분위 (원)	4,847,690
8분위 (원)	5,560,819
9분위 (원)	6,624,795
10분위 (원)	9,737,437
경계값 (원)	
p10 (원)	1,590,875
p20 (원)	2,350,000
p30 (원)	2,955,260
p40 (원)	3,489,043
p50 (원)	3,994,569
p60 (원)	4,536,655
p70 (원)	5,177,412
p80 (원)	6,004,671
p90 (원)	7,352,998





변동성 측도

- 분산 (variance) $$S^2 = \frac{\sum(x - x_i)^2}{n - 1}$$
- 표준편차 (standard deviation): 원래의 자료에서 사용된 단위와 동일한 단위로 측정되므로 분산보다 해석 용이

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

```
> var(tip)
[1] 1.914455
>
> sd(tip)
[1] 1.383638
```

분산

표준편차

변동성 측도

- 변동계수 (Coefficient of Variation): 표준편차가 평균에 비하여 얼마나 큰지 나타냄

$$\frac{S}{\overline{X}}$$

  - 평균이 클 수록 표준편차가 큰 경향이 있으므로 다른경우 표준편차로 변동성 비교가 곤란
  - 두 지역 강수량 비교

A지역	B지역
6 8 10 12 16 18	56 58 60 62 66 68
평균= 11.67 표준편차 = 4.23 변동계수= 0.36	평균= 61.67 표준편차 = 4.23 변동계수= 0.069

```
> sd(tip)/mean(tip)
[1] 0.4614775
```

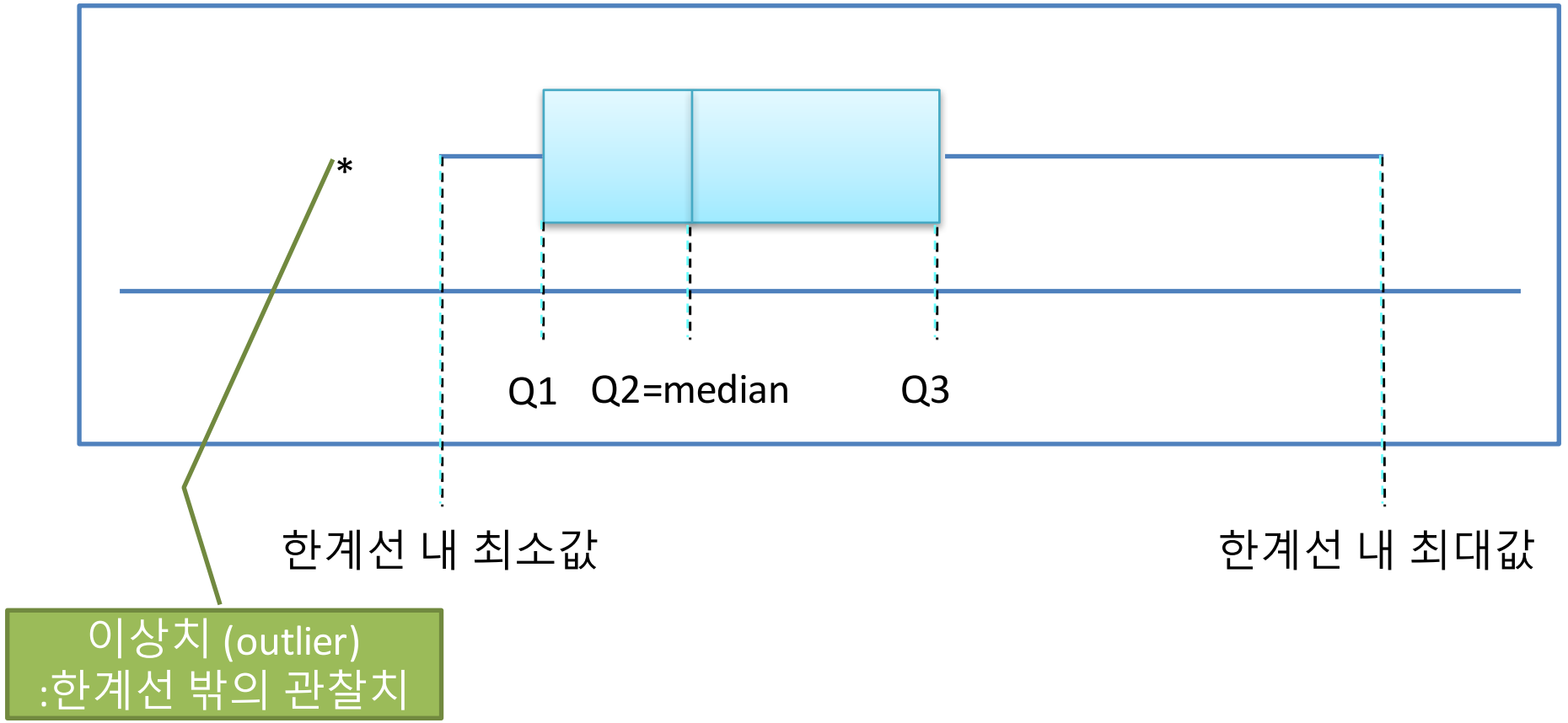
변동성 측도

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

```
> IQR(tip)
[1] 1.5625
```

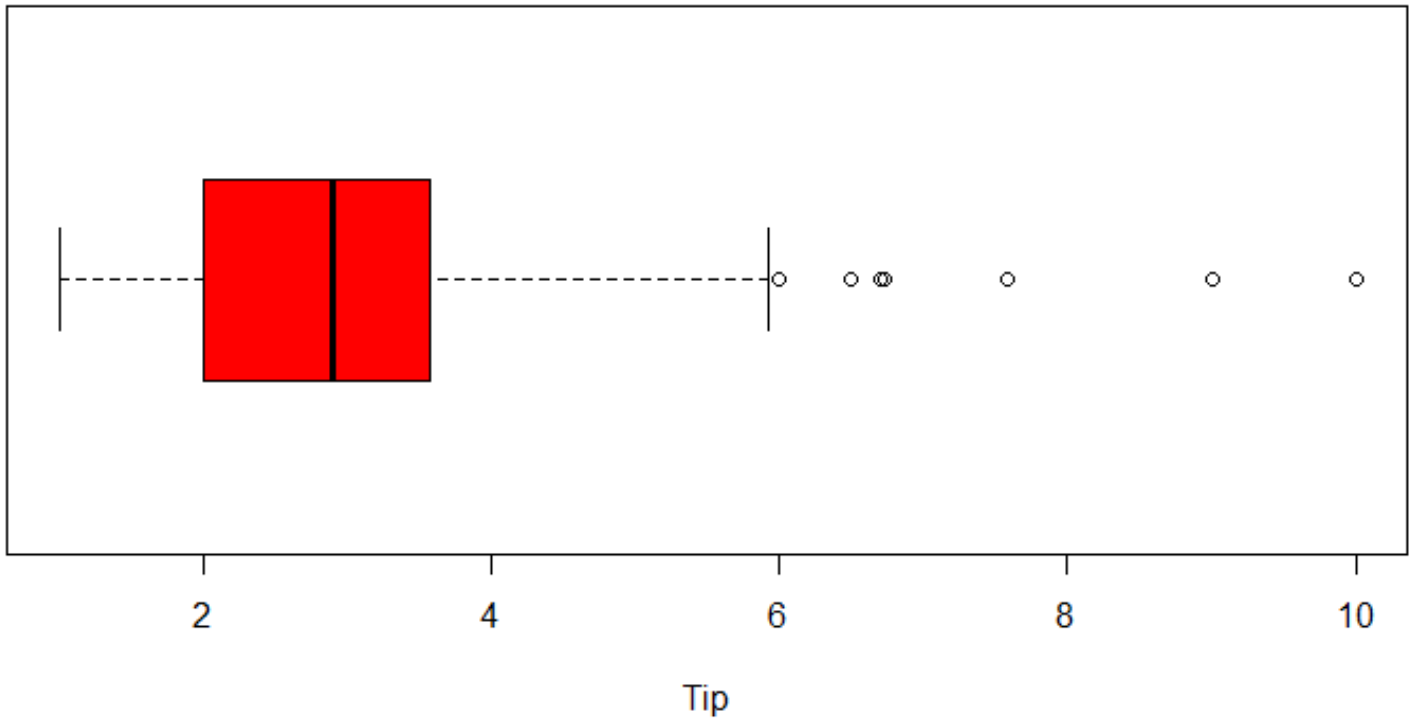
Boxplot

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



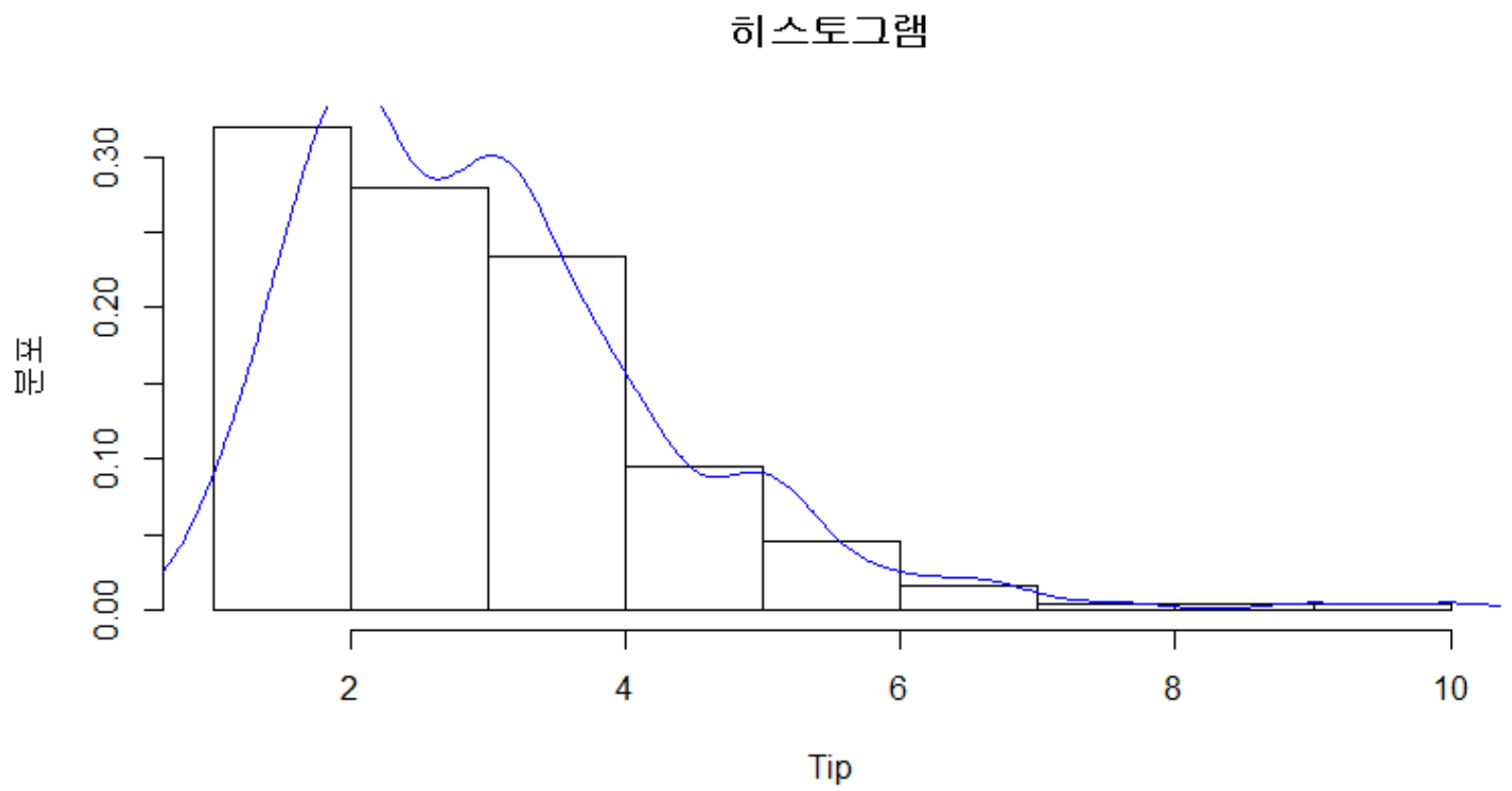
Boxplot: Tips data

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



Histogram

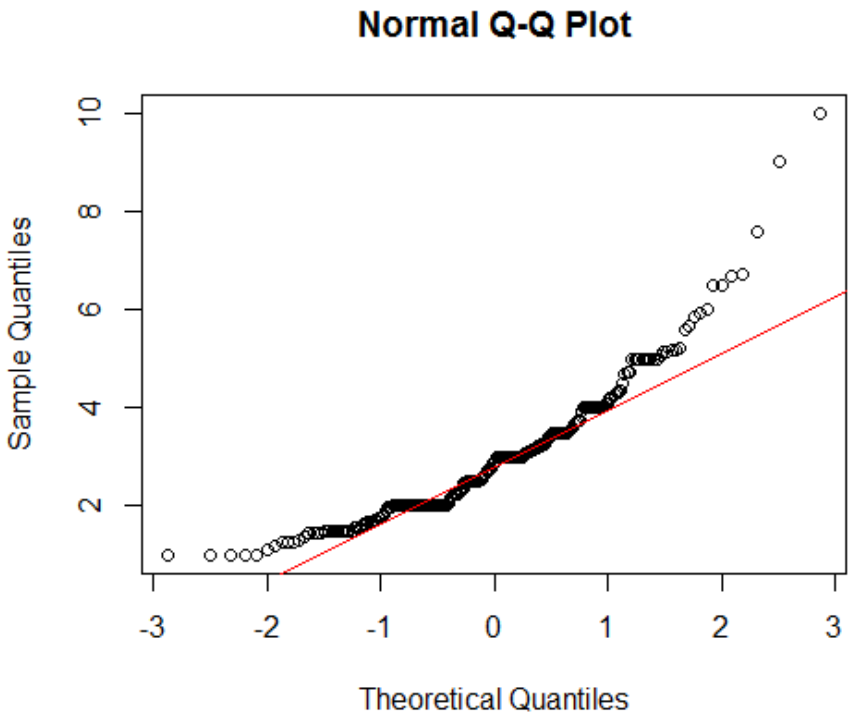
```
> hist(tip,probability=TRUE,main="히스토그램",xlab="Tip",ylab="분포")
> lines(density(tip),col="blue")
```

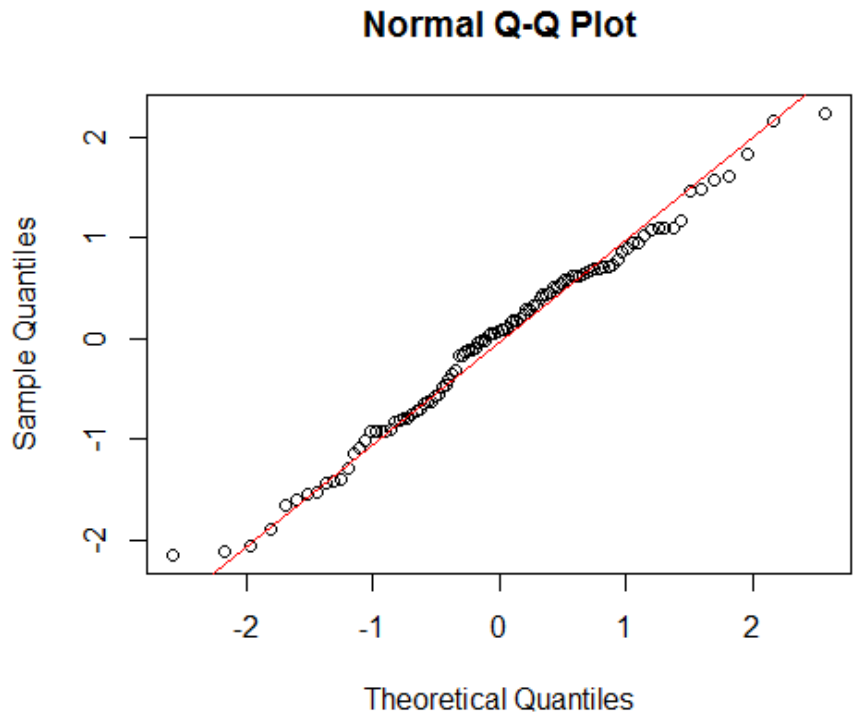


Q-Q Normality Plot

- 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

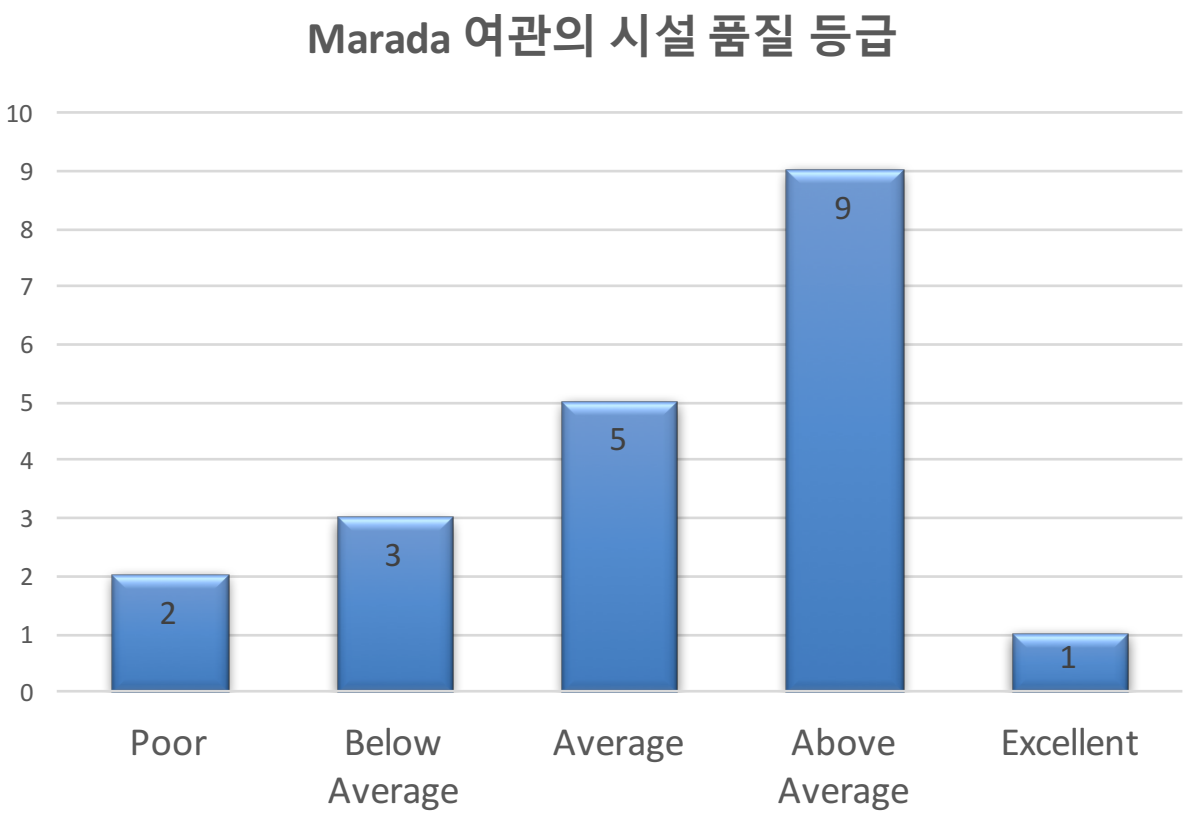
상대 도수와 백분율 도수 분포

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

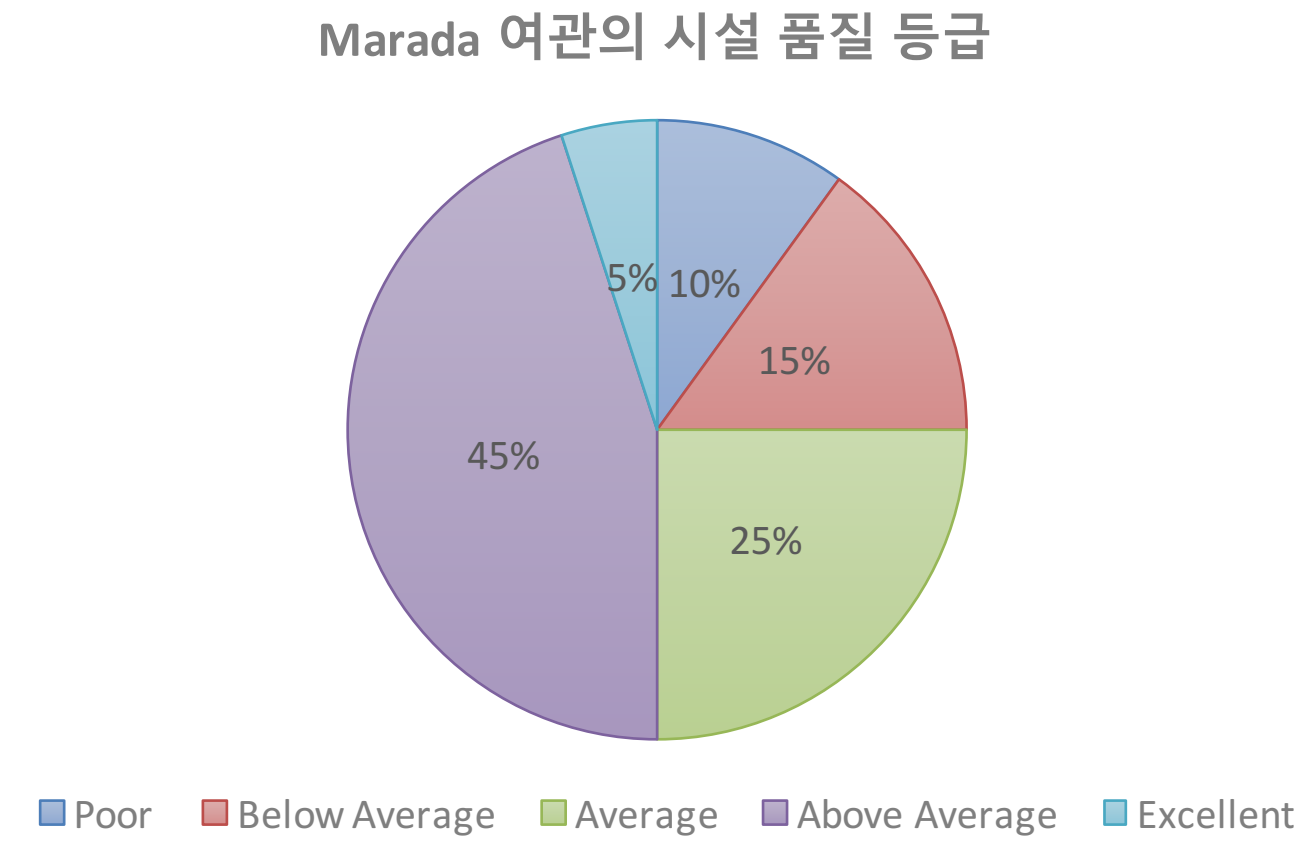
1/20 = .05

막대 그래프



- 가로축: 계급의 이름
- 세로축: 도수분포, 상대도수분포, 백분율도수분포의 크기(scale)
- 각 계급이 분리되어 있다는 것을 강조하기 위해서 막대는 서로 분리

파이차트


Marada 여관의 시설 품질 등급
5%
10%
15%
45%
25%
Poor
Below Average
Average
Above Average
Excellent

도수분포표: Tips data

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

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

Thur  Fri  Sat  Sun
  62   19   87   76
```

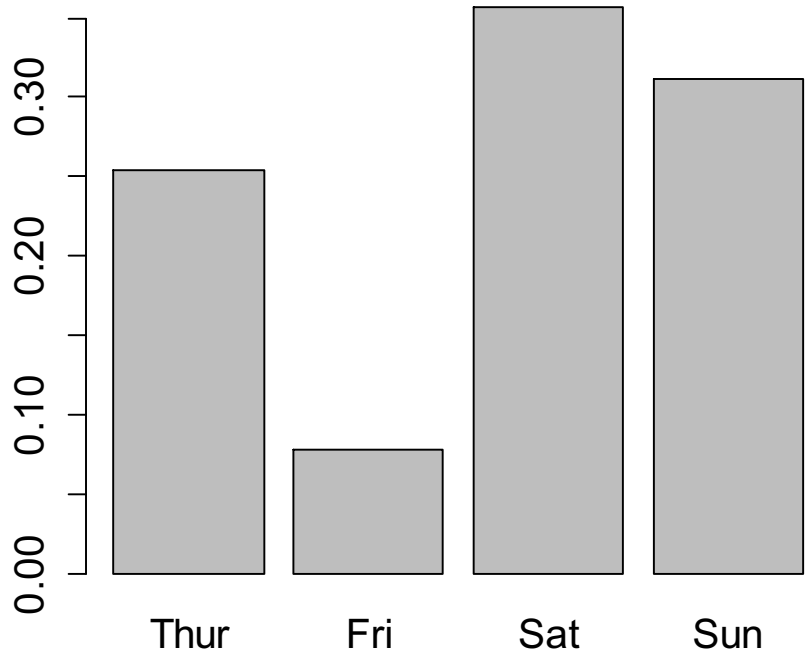
- summary vs. describe(psych package)

```
> library(psych)
> describe(tips)
           vars   n  mean   sd median trimmed  mad  min   max range  skew kurtosis   se
total_bill    1 244 19.79 8.90   17.8   18.73 7.46 3.07 50.81 47.74  1.12     1.14 0.57
tip           2 244  3.00 1.38    2.9    2.84 1.33 1.00 10.00  9.00  1.45     3.50 0.09
sex*          3 244  1.64 0.48    2.0    1.68 0.00 1.00  2.00  1.00 -0.60    -1.65 0.03
smoker*       4 244  1.38 0.49    1.0    1.35 0.00 1.00  2.00  1.00  0.49    -1.77 0.03
day*          5 244  2.74 0.93    3.0    2.78 1.48 1.00  4.00  3.00 -0.06    -1.02 0.06
time*         6 244  1.28 0.45    1.0    1.22 0.00 1.00  2.00  1.00  0.98    -1.04 0.03
size          7 244  2.57 0.95    2.0    2.42 0.00 1.00  6.00  5.00  1.43     1.63 0.06
```

Bar plot: Tips data

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

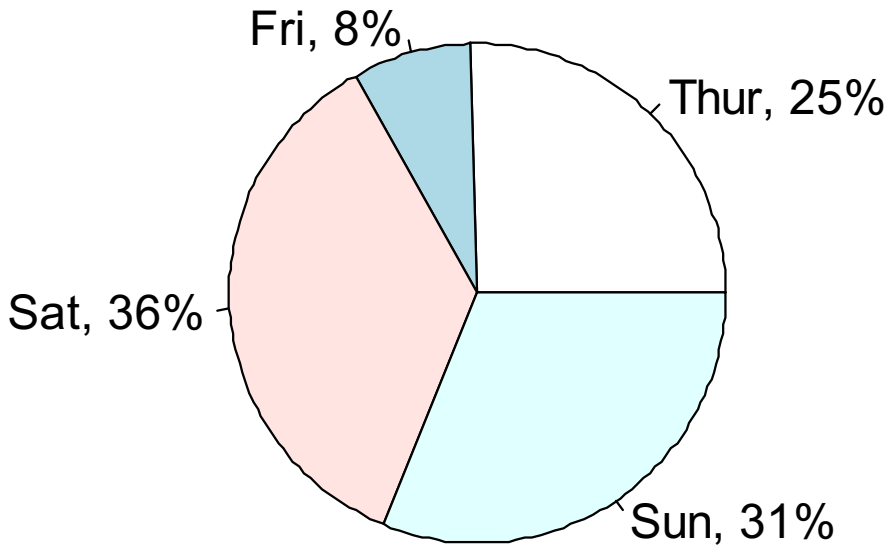
```
> barplot(mytable/sum(mytable))
```



Pie chart

```
> lbl=paste(names(mytable),", ",round(mytable/sum(mytable))*100,"%",sep="")
> lbl
[1] "Thur, 0%" "Fri, 0%"  "Sat, 0%"  "Sun, 0%"
> mytable

Thur  Fri  Sat  Sun
  62   19   87   76
> lbl=paste(names(mytable),", ",round(mytable/sum(mytable)*100),"%",sep="")
> lbl
[1] "Thur, 25%" "Fri, 8%"   "Sat, 36%"  "Sun, 31%"
> pie(mytable,labels=lbl)
```



두 변수의 요약

분할표 (Contingency Table)

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

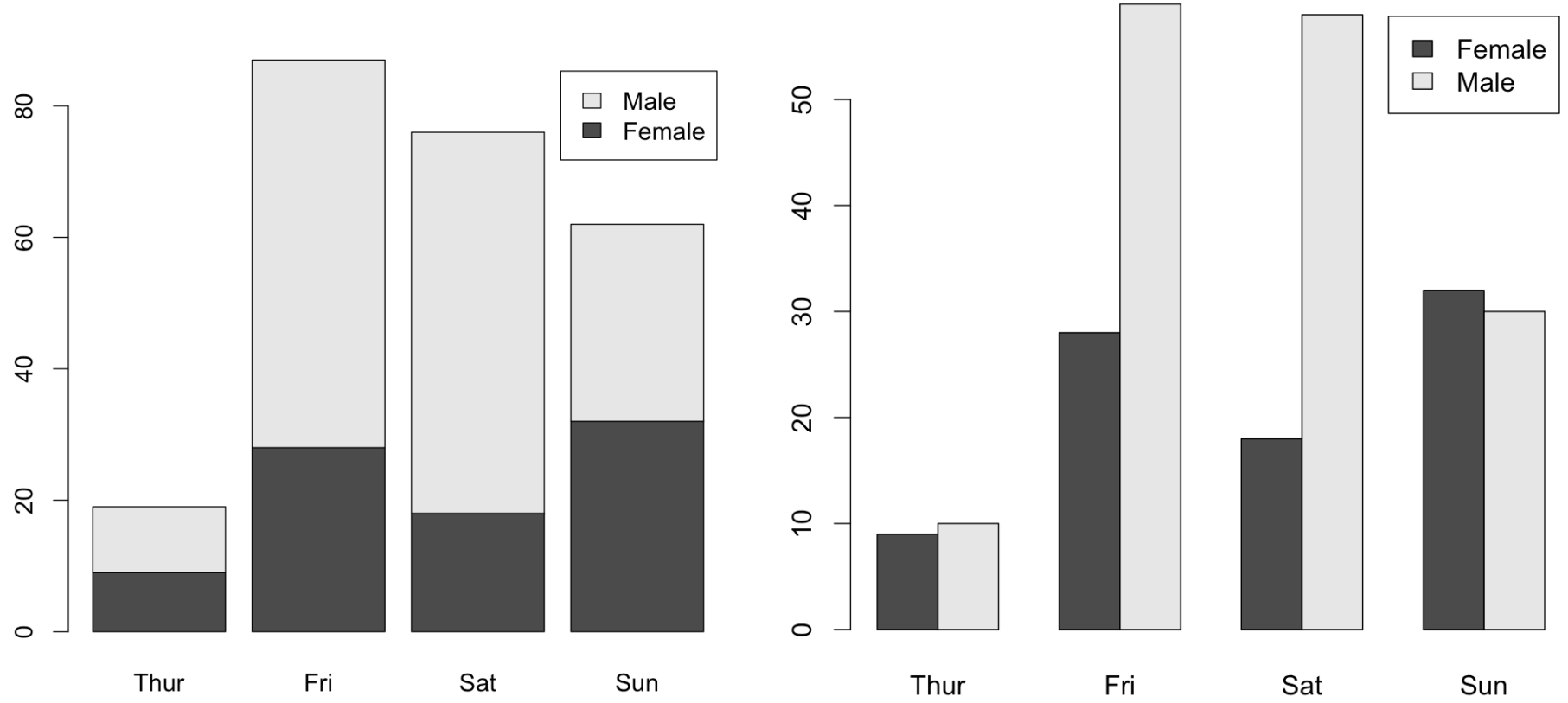
```
> head(tips)
  total_bill  tip    sex smoker day   time size
1      16.99 1.01 Female     No Sun Dinner    2
2      10.34 1.66   Male     No Sun Dinner    3
3      21.01 3.50   Male     No Sun Dinner    3
4      23.68 3.31   Male     No Sun Dinner    2
5      24.59 3.61 Female     No Sun Dinner 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

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

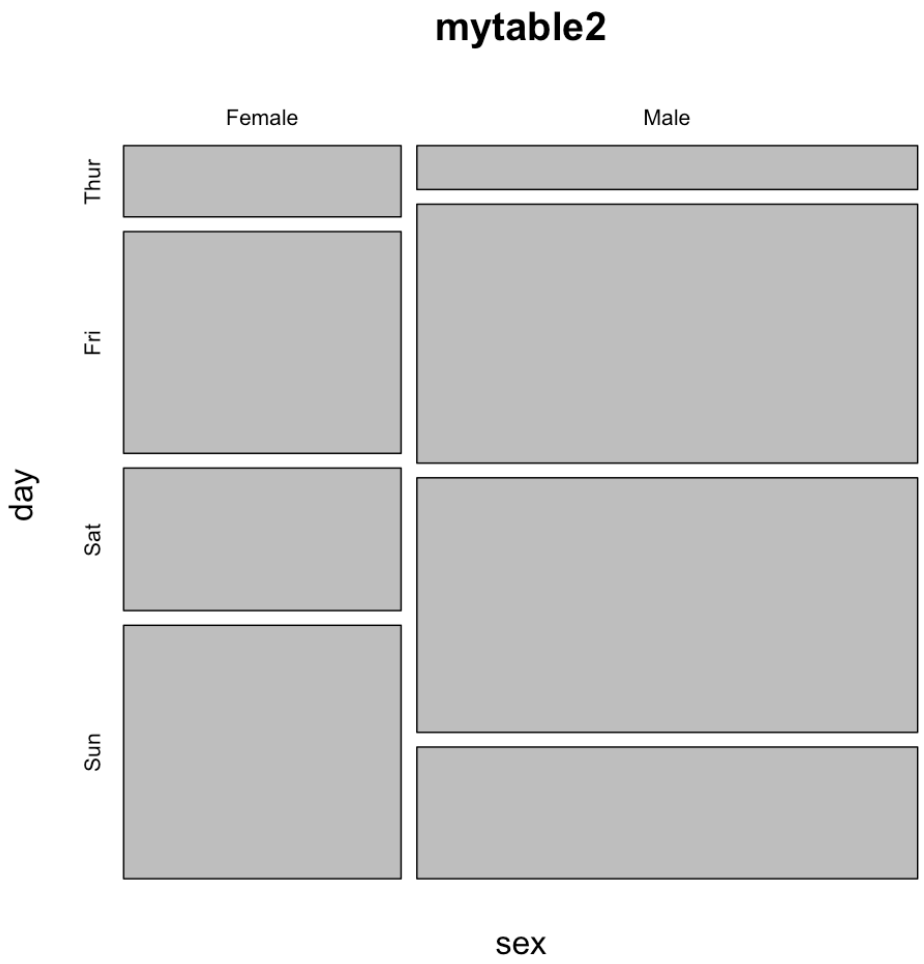
Bar Plot

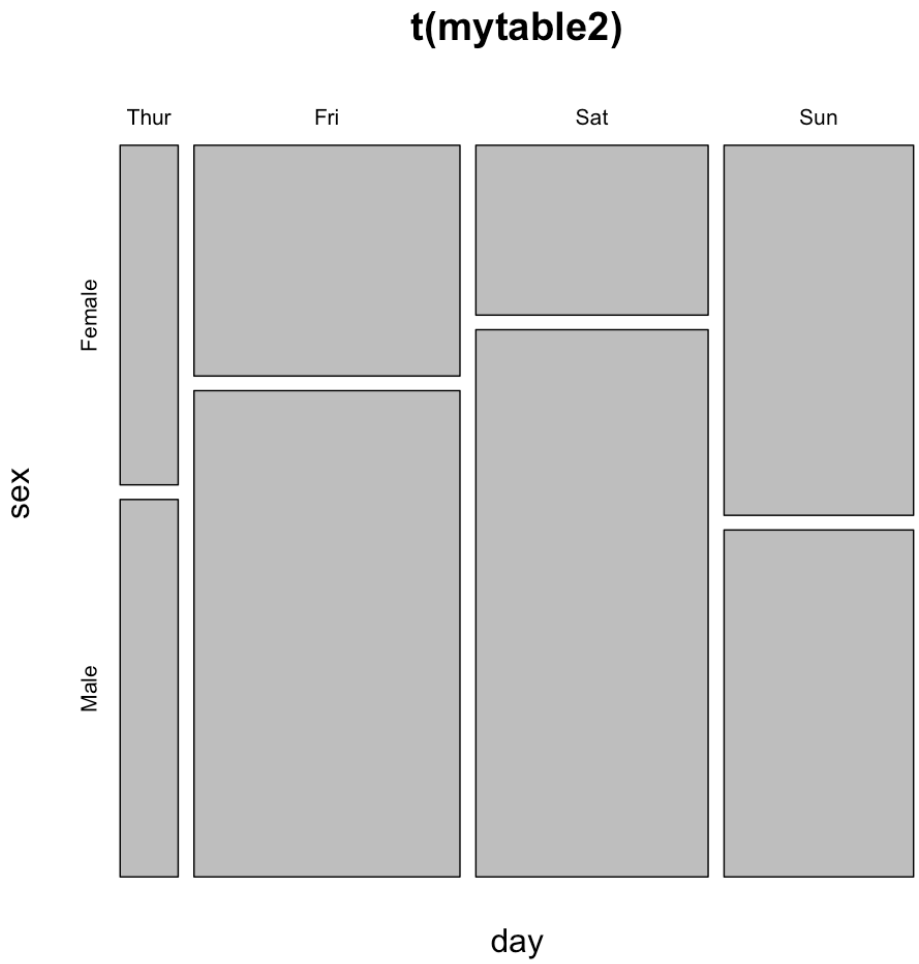
```
> barplot(mytable2,legend.text=c("Female","Male"))
> barplot(mytable2,legend.text=c("Female","Male"),beside=TRUE)
```



Mosaic Plot

```
> mosaicplot(mytable2)
> mosaicplot(t(mytable2))
```

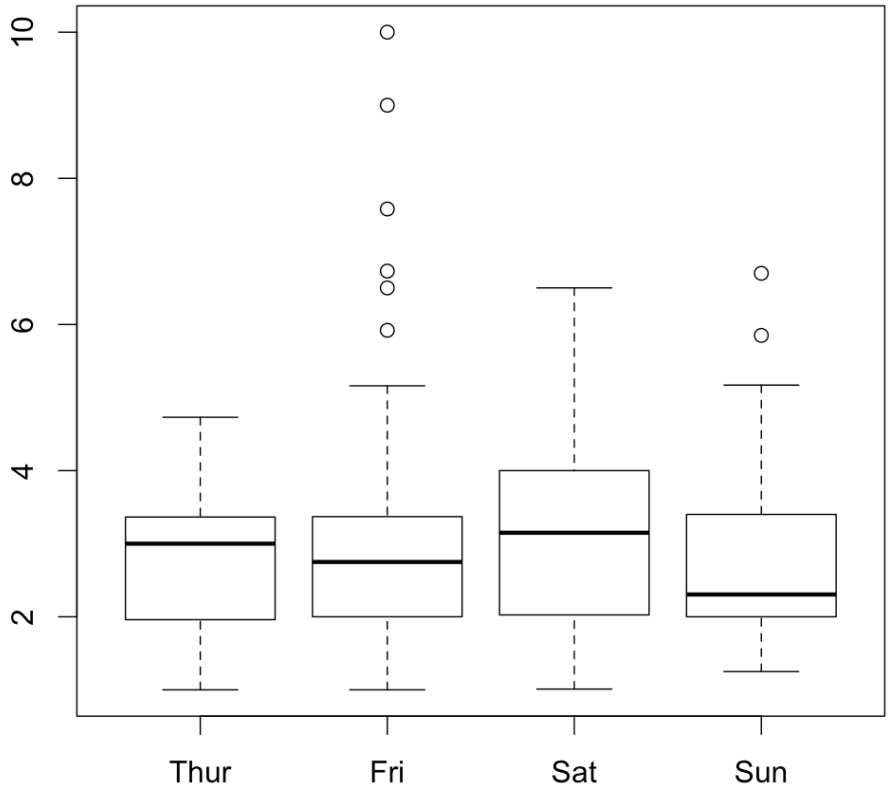




상자그림

- 범주형 변수와 양적 변수의 요약

```
> boxplot(tip~day,data=tips)
```



산점도

- 두 양적변수의 요약

```
> plot(tip~total_bill,tips)
```