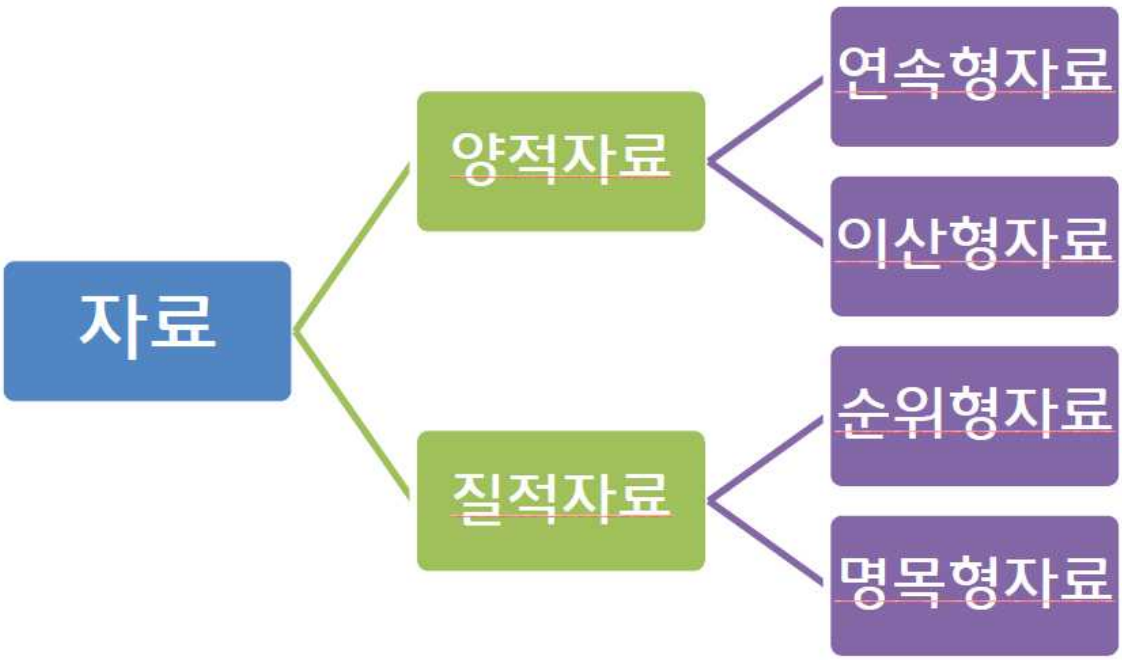
제12주: 자료와 그래프

통계 자료를 그래프로 요약하기 위한 함수를 소개합니다.

1. 자료의 종류

통계학에서 다루는 자료의 유형은 다음의 네가지 입니다.



양적자료: 사칙연산을 통해 계산할 수 있는 자료들

- 연속형자료: 값의 증감이 연속적으로 발생하는 자료
  (예: 키, 체중)
- 이산형자료: 값의 증감이 계단형태로 발생하는 자료
  (예: 안경을 쓴 학생의 수)

질적자료: 자료의 양적인 크기에 상관없이 값 자체에 의미를 부여하는 자료

- 순위형 자료: 순위가 있는 범주를 나타내는 자료
  (예: 설문지의 5점 척도:
  1=전혀 그렇지 않음, 3=보통, 5=매우 그러함)
- 명목형자료: 순위가 없는 범주를 나타내는 자료
  (예: 성별, 왼손잡이 여부)

2. 예제 자료 준비하기: ldeaths

- ldeaths: R의 datasets패키지의 ldeaths 자료
- ?ldeaths

  Monthly Deaths from Lung Diseases in the UK
- 데이터구조 및 자료형

```
> ldeaths
     Jan  Feb  Mar  Apr  May  Jun  Jul  Aug  Sep  Oct  Nov  Dec
1974 3035 2552 2704 2554 2014 1655 1721 1524 1596 2074 2199 2512
1975 2933 2889 2938 2497 1870 1726 1607 1545 1396 1787 2076 2837
1976 2787 3891 3179 2011 1636 1580 1489 1300 1356 1653 2013 2823
1977 3102 2294 2385 2444 1748 1554 1498 1361 1346 1564 1640 2293
1978 2815 3137 2679 1969 1870 1633 1529 1366 1357 1570 1535 2491
1979 3084 2605 2573 2143 1693 1504 1461 1354 1333 1492 1781 1915
> str(ldeaths)
Time-Series [1:72] from 1974 to 1980: 3035 2552 2704 2554 2014 ...
```

- 시계열 자료(벡터)를 행렬로 바꾸기
  - matrix() 함수를 이용
  - 자료값이 입력되는 순서에 유의 (열방향? 행방향?)
  - 행이름과 열이름을 추가

```
> ld <- matrix(ldeaths, nrow = 6, ncol = 12, byrow = TRUE)
> rownames(ld) <- 1974:1979    #행이름을 1974,...,1979로 변경
> colnames(ld) <- month.abb    #열이름=미국식 달 이름의 약자
> ld
> str(ld)
```

- 범주형 변수 만들기
  - ⊙ ldeaths자료에 대하여, 각 값들이 어떤 달에 속하는지를 나타내는 범주형 변수를 만들자.
  - ⊙ 자료형: 팩터(factor). 순서가 있으므로 ordered=TRUE

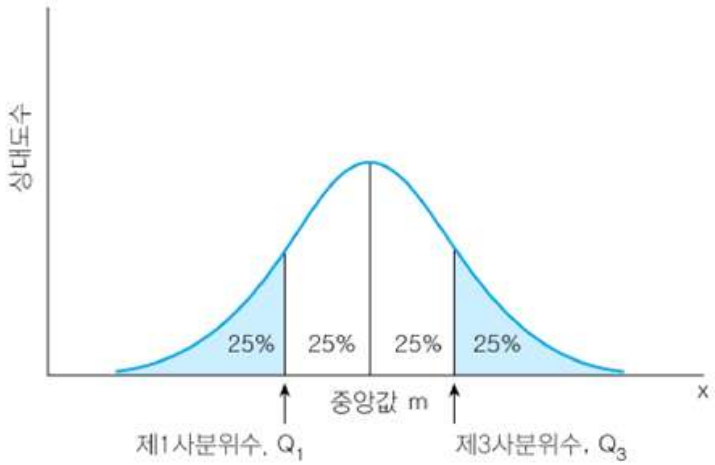
```
> month <- factor(rep(month.abb, 6), ordered = TRUE, levels = month.abb)
> ld_df <- data.frame(ldeaths, month)
> ld_df
```

3. 상자그림: boxplot

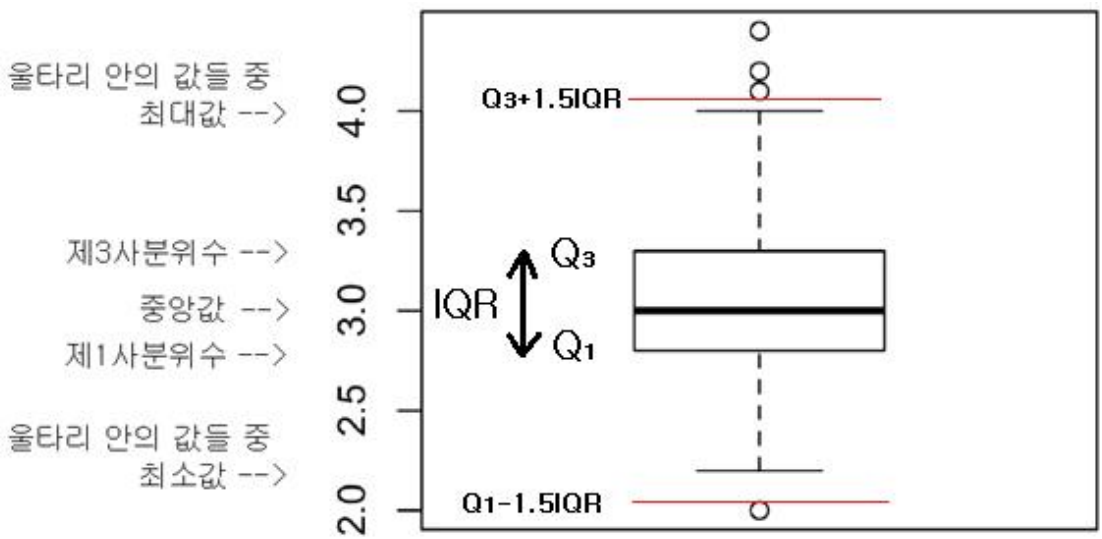
- 데이터의 분포를 보여주는 그림
- 상자는 제1사분위수(Q_1), 중앙값, 제3사분위수(Q_3)를 나타낸다.

⊙ 사분위수(quartile)란?

- 측정값들을 크기순서대로 정렬한다. 관측값을 4개로 동일한 비율로 나누는 중요한 3개의 수를 사분위수라고 한다.



[그림 2.9] 사분위수의 위치



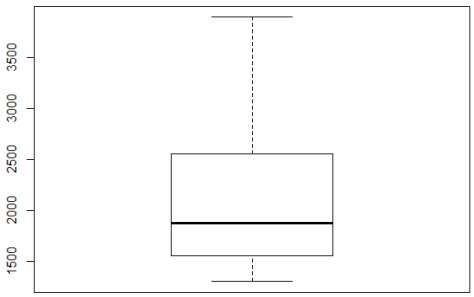
- 5개의 요약통계량 계산하기: summary() 함수

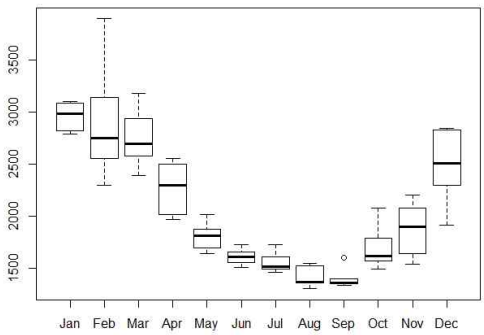
```
> summary(ldeaths)   #1. 자료가 벡터일 경우
> summary(ld)        #2. 자료가 행렬일 경우
```

```
> summary(ldeaths)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1300    1552    1870    2057    2552    3891
> summary(ld)
      Jan             Feb             Mar             Apr             May             Jun             Jul             Aug             Sep             Oct             Nov             Dec
 Min.   :2787   Min.   :2294   Min.   :2385   Min.   :1969   Min.   :1636   Min.   :1504   Min.   :1461   Min.   :1300   Min.   :1333   Min.   :1492   Min.   :1535   Min.   :1915
 1st Qu.:2844   1st Qu.:2565   1st Qu.:2600   1st Qu.:2044   1st Qu.:1707   1st Qu.:1560   1st Qu.:1491   1st Qu.:1356   1st Qu.:1348   1st Qu.:1566   1st Qu.:1675   1st Qu.:2342
 Median :2984   Median :2747   Median :2692   Median :2294   Median :1809   Median :1606   Median :1514   Median :1364   Median :1356   Median :1612   Median :1897   Median :2502
 Mean   :2959   Mean   :2895   Mean   :2743   Mean   :2270   Mean   :1805   Mean   :1609   Mean   :1551   Mean   :1408   Mean   :1397   Mean   :1690   Mean   :1874   Mean   :2478
 3rd Qu.:3072   3rd Qu.:3075   3rd Qu.:2880   3rd Qu.:2484   3rd Qu.:1870   3rd Qu.:1650   3rd Qu.:1588   3rd Qu.:1484   3rd Qu.:1386   3rd Qu.:1754   3rd Qu.:2060   3rd Qu.:2745
 Max.   :3102   Max.   :3891   Max.   :3179   Max.   :2554   Max.   :2014   Max.   :1726   Max.   :1721   Max.   :1545   Max.   :1596   Max.   :2074   Max.   :2199   Max.   :2837
```

- 상자그림 그리기: boxplot()

```
> boxplot(ldeaths)    #1. 자료가 벡터일 경우
> boxplot(ld)         #2. 자료가 행렬일 경우
> boxplot(ldeaths ~ month, data = ld_df)  #3. 자료가 데이터프레임
```

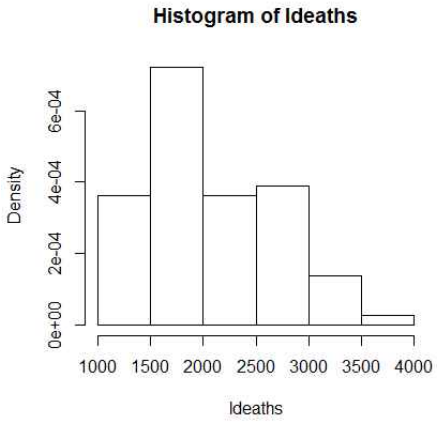




4. 히스토그램: hist

- 데이터의 분포를 보여주는 그림
- 각 구간 별 데이터의 개수를 막대높이로 그린다.

```
> hist(ldeaths)
> hist(ldeaths, breaks = 10)     #계급의 개수를 지정
> hist(ldeaths, freq = FALSE)  #데이터 개수가 아닌 확률밀도를 그린다. 즉, 막대의 면적의 합은 1과 같다.
> x = hist(ldeaths)
> x                                #그림에 관한 정보를 얻을 수 있다.
```

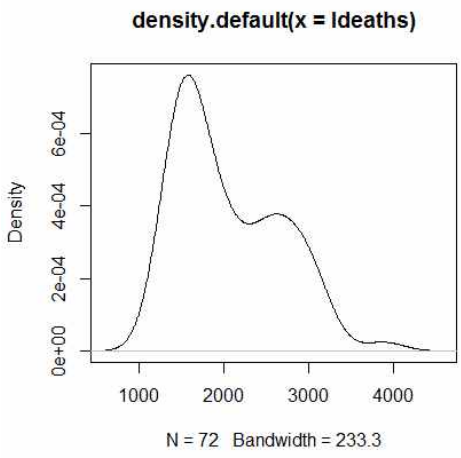


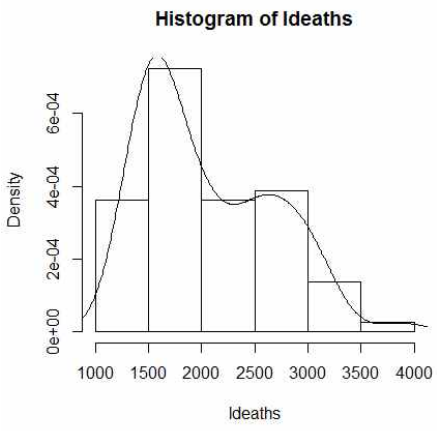

```
> x = hist(ldeaths)
> x
$breaks
[1] 1000 1500 2000 2500 3000 3500 4000

$counts
[1] 13 26 13 14  5  1

$density
[1] 3.611111e-04 7.222222e-04 3.611111e-04 3.888889e-04

$mids
[1] 1250 1750 2250 2750 3250 3750

$xname
[1] "ldeaths"

$equidist
[1] TRUE

attr(,"class")
[1] "histogram"
```

5. 밀도그림: density

- histogram의 경우 구간의 너비에 따라 그림이 크게 바뀐다는 단점이 있다.
- density()함수는 kernel density estimation을 수행한다.

```
> plot(density(ldeaths))
# adjust= 옵션은 곡선의 부드러운 정도를 조절. 1은 기준값
> plot(density(ldeaths, adjust =0.5))
# 히스토그램과 밀도함수그림을 함께 그릴 수 있다.
> hist(ldeaths, freq = FALSE)
> lines(density(ldeaths))
```





5. 막대그림: barplot

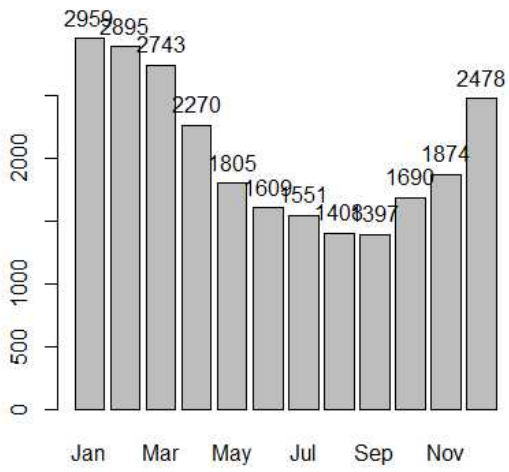
- 막대그림은 범주별 통계를 표현
- 막대 높이를 barplot()함수의 입력으로 전달한다.

  ==> 자료를 일단 표로 요약할 필요가 있다.
- ldeath자료에 대하여, 월별 평균값을 계산하여 막대그림을 그리자.

```
# 행렬 자료의 각 열별로 평균을 계산하기
> ldmean <- apply(ld, 2, mean)
> ldmean
> barplot(ldmean)

# 벡터 자료의 경우, tapply()로 범주별 평균을 계산하기
> ldmean <- tapply(ldeaths, month, mean)
> barplot(ldmean)
```

- 막대 위 끝부분에 막대의 높이를 글씨로 직접 써줄 수 있을까? text() 함수를 이용한다.

```
# 글씨가 그림의 위로 넘어가도 글씨를 써준다. (xpd=NA)
> bp = barplot(ldmean)                  # 막대의 가로축 위치 저장
> text(x = bp,                          # 글씨의 가로 위치
       y = ldmean,                      # 글씨의 세로 높이
       labels = round(ldmean, 0),       # 막대 위에 쓸 글씨
       pos = 3, xpd = NA)               # 지정한 위치의 위(pos=3)
```



6. 파이 그래프: pie

- 파이 그래프는 범주별 통계를 표현
- 데이터의 비율을 알아보는데 적합

```
> pie(ldmean)

> pct = paste(names(ldmean),
                round(ldmean/sum(ldmean)*100,1), "%")
> pie(ldmean, labels = pct)
```