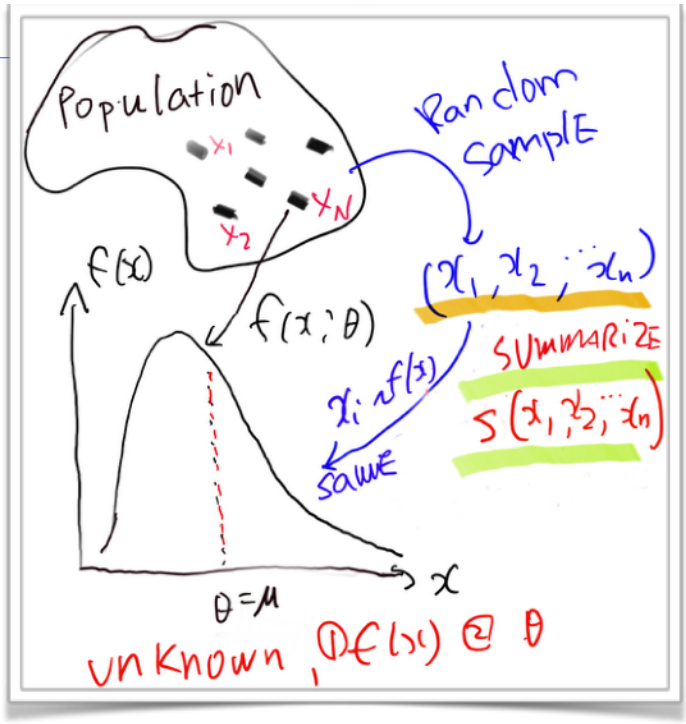
데이터 요약 개념



- 모집단으로부터($X \sim f(x;\theta)$) 얻은 확률표본(데이터)은 $(x_1, x_2, \ldots, x_n)$는 모집단의 분포와 동일하다.
- 모집단에서 궁금한 것은 확률분포함수 $f(x)$와 모수(θ)이다.
- 모수에 대한 정보는 확률표본, 데이터의로부터 계산될 수 있는데, f(x)는 그래프 graphical 요약으로부터 모수에 대한 정보는 숫자 numerical 요약으로 얻는다.
- 통계적 방법론에서 데이터 종류는 "숫자형, 정량변수", "분류, 범주형, 정성변수", 2개로 나뉘고, 변수의 종류에 따라 요약방법이 정해진다.

정성적 변수 qualitative, categorical

정성(범주)형 변수(데이터)는 명목형, 순서형(순서가 있는 범주형)으로 세분화 되지만 통계적 분석에서는 동일한 방법이 적용된다.

정성적 데이터는 가질 수 있는 값의 수준(범주 category)이 유한하므로 각 범주의 빈도 frequency로 숫자로 표현하거나 그래프로 나타내면 된다.

- 빈도 frequency : 데이터에서 동일한 범주 값이 반복된 개수
- 상대빈도 relative frequence : 빈도를 데이터 크기로 나눈 값 => **비율 proportion ($p, \hat{p}$)**

1) 숫자 요약 = 빈도

- 각 범주의 빈도, 상대빈도(비율 ratio)가 정리한다.
- 빈도표, 상대빈도 (= 비율 = 확률분포함수) 표 정리한다.

PET Type	Count	Rel. Freq
Dog	16	.29
Cat	28	
Fish	8	
Other	4	
Totals	56	

16/56

- [**예제 데이터**] 소속팀의 선수 빈도표와 상대빈도를 구하시오. [엑셀 이용]

```
    Atlanta     Baltimore
         11            15
 Cincinnati     Cleveland
         12            12
Los Angeles     Milwaukee
         14            14
    Oakland  Philadelphia
         12            12
    Seattle      St Louis
         12            11
```

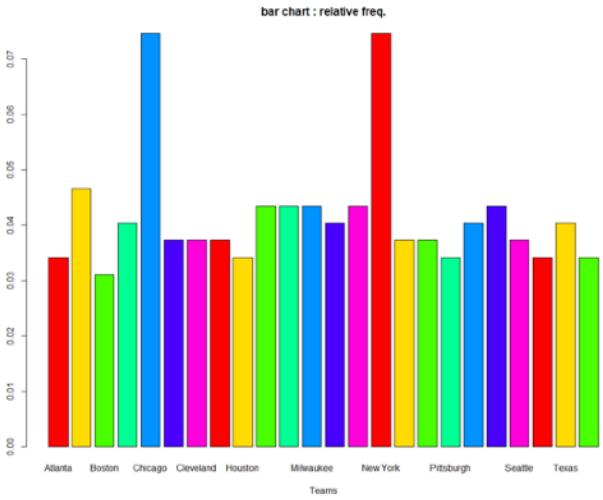
```
    Atlanta     Baltimore
0.03416149    0.04658385
 Cincinnati     Cleveland
0.03726708    0.03726708
Los Angeles     Milwaukee
0.04347826    0.04347826
    Oakland  Philadelphia
0.03726708    0.03726708
    Seattle      St Louis
0.03726708    0.03416149
```

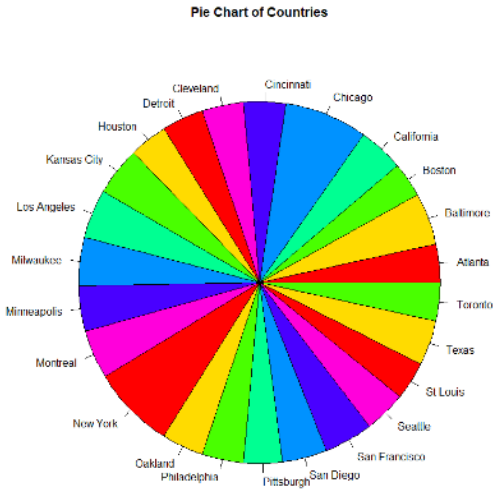
2) 그래프 요약

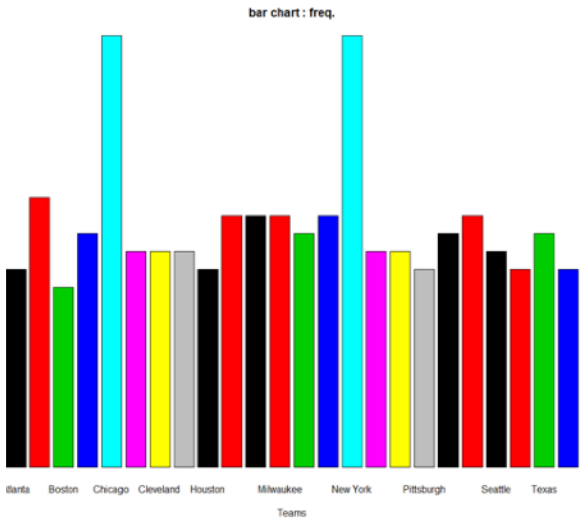
빈도표(상대빈도표)를 막대 그래프나 파이(상대빈도만 가능, 원 전체=100%) 차트로 나타내면 된다.

(상대)빈도가 **가장 큰 범주를 최빈값 (mode)**라 한다.

Chicago, New York 최빈값 —>







정량적 변수 quantitative, numeric

크기를 가진 숫자 데이터이므로 데이터 관측값을 활용하여 1)모집단의 확률분포함수 형태를 알 수 있는 표본 확률분포함수(그래프 요약), 2)모수에 대한 정보를 알 수 있는 통계량을 구할 수 있다.

1) 그래프 요약

데이터 관측값을 BIN (계급 구간 class interval) 8~12개(Thumb's Rule) (Sturge's rule - $K = 1 + 3.322 * log_{10}(n)$)로 구성하여 막대 그래프로 나타낸 히스토그램 histogram이나, 5개 주요 통계량을 그래프에 나타낸 나무상자그림으로 요약한다.

- 표본 확률분포함수의 형태를 알 수 있다. $f(x)$
- 중앙 위치, 산포(데이터 흩어짐)에 대한 정보를 얻는다.

(1) 순서통계량 order statistic

데이터 관측값($x_1, x_2, \ldots, x_n$)을 크기 순으로 정렬한 통계량 - $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$

- 최대값 maximum value : 데이터 관측값 중 가장 큰 값, $x_{(n)}$
- 최소값 minimum value : 데이터 관측값 중 가장 작은 값, $x_{(1)}$
- 범위 range : 최대값-최소값, $R = x_{(n)} - x_{(1)}$
- 중위값 median : 데이터 관측값 중 크기 순서에서 가운데 있는 관측값, $x_{(MD)}$, MD=중위값 위치, $MD = (n+1)/2$, 만약 MD가 정수가 아닌 경우 (예: 11.5) 11번째 순서통계량과 12번째 순서통계량의 평균값을 중위값 - $\frac{(x_{(11)} + x_{(12)})}{2}$
- 사분위값 quartile : 데이터 크기 순 25%(일 first 사분위, $x_{(QD)}$), 50%(이사분위, 중위값), 75%(삼사분위) 값, 사분위 위치 Quartile Depth, $QD = \frac{MD_integer + 1}{2}$ (예 : MD=11.5 인 경우 MD_integer=11임)
- 분위값 percentile : 데이터를 크기순으로 정렬 했을 때 백분위 위치에 있는 관측값, 상위 20% = 80% 분위, $x_{(0.8*n)}$ 보간법 (예 : n=22이면, 0.8*22=17.6 위치, $x_{(17)}$와 $x_{(18)}$을 활용한 보간법으로 (0.4:0.6) 배분

(예제 데이터) n=12

8 17 9 10 9 11 7 13 12 3 10 4

(순서 통계량)

3 4 7 8 9 9 10 10 11 12 13 17

(최대값) $x_{(12)} = 17$, (최소값) $x_{(1)} = 3$ —> 범위 range = 최대값-최소값=14

(중위값) 중위 위치 $MD = (12+1)/2 = 6.5$, $\frac{x_{(6)} + x_{(7)}}{2} = (9+10)/2 = 9.5$

(제일사분위값) 사분위 위치 $QD = (MD_i nteger + 1)/2 = 3.5$

$$\frac{x_{(3)} + x_{(4)}}{2} = (7+8)/2 = 7.5$$

(제삼사분위값) $\frac{x_{(9)} + x_{(10)}}{2} = (11+12)/2 = 11.5$

(80% 분위값) 분위위치=0.8*12=9.6 ->

$x_{(10)} - x_{(9)} = 1 * 0.6 =$

$x_{(9)} + 0.6 = 11.6$

✳ 위의 방법은 수작업 방법으로 사분위, 백분위, 통계 소프트웨어 산식(아래 산식)은 복잡하고 정확하여 위의 계산 결과와 상이함

최소값	3
최대값	17
중위값	9.5
제일사분위	7.75
제3사분위	11.25
80%분위	11.8

```
> min(x);max(x);median(x);
[1] 3
[1] 17
[1] 9.5
> quantile(x,0.25);quantile(x,0.75)
 25%
7.75
  75%
11.25
> quantile(x,0.8)
 80%
11.8
```

$$\frac{c_\ell + 0.5 f_i}{N} \times 100\%$$

N = 데이터 크기, c_l = 해당 분위 전 누적 빈도, f_i = 해당 분위 빈도

(2) 나무상자그림 box whisker plot

5개 주요 순서통계량 (최소, 1사분위, 중위, 3사분위, 최대) 그리고 이상치, 상자와 수염 크기로 나타낸다.

- IQR 사분위 범위 Inter-Quratile Range : $IQR = Q_3 - Q_1$
- 이상치 outlier : $(Q_1 - 1.5 * IQR, Q_3 + 1.5 * IQR)$ 벗어난 관측치
- 극심 severe 이상치 : $(Q_1 - 3 * IQR, Q_3 + 3 * IQR)$ 벗어난 관측치

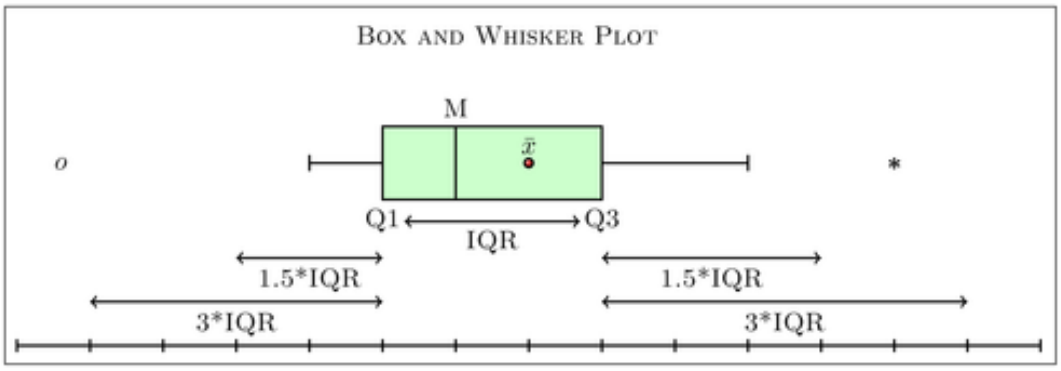
4개의 구간에 데이터 개수가 1/4씩 분포되어 있음

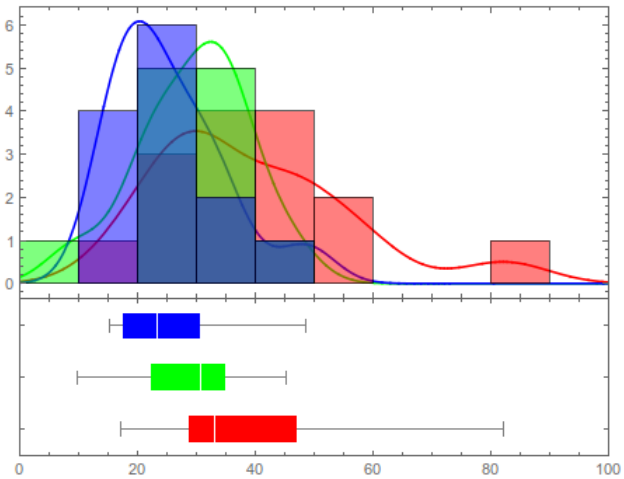
- 좌우 대칭인 경우 상자 정확하게 이등분, 수염 길이 동일

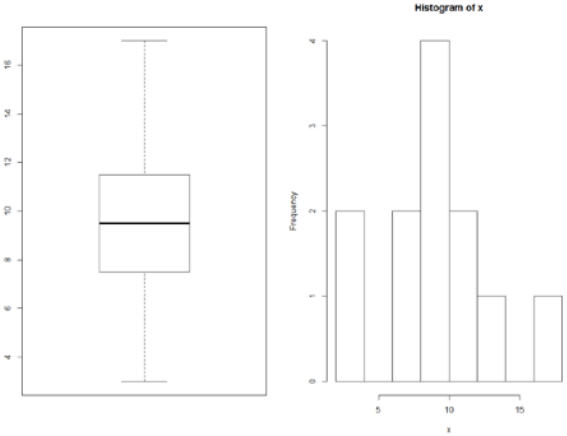
(진단 내용) a) 확률분포함수 형태 - 모집단 확률분포함수 (치우침, 좌우 대칭) b) 이상치 진단

(단점) 봉우리 개 수 알 수 없음, 정확한 통계량 값 알 수 없음

[이상치 해결] 오류 검증 후 제거 후 통계적 분석 방법 적용







(3) 히스토그램 histogram

정량적 데이터의 바 그래프와 동일, 범주를 구간으로 설정

a) 데이터를 크기 순으로 정렬한 후 최대값과 최소값을 구하고 범위 range을 구한다. (예제 데이터 : 범위=14)

b) 빈(bin 계급 class) 개수를 결정하고 범위를 빈 개수로 나누어 계급 구간 넓이(폭) interval width을 결정한다. (예 : 빈 개수를 5개로 하는 경우 구간 폭은 14/5=2.8)

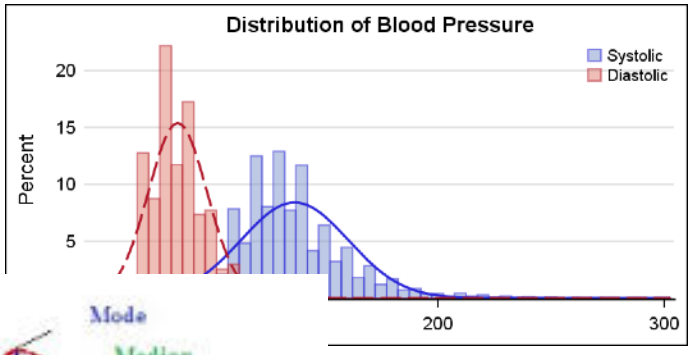
c) 구간 폭이 의미를 갖도록 가능하면 정수 단위로 조정한다. (예: 3~17, 최초 폭 2.8이므로 구간 폭은 최종 2 혹은 5로 결정)

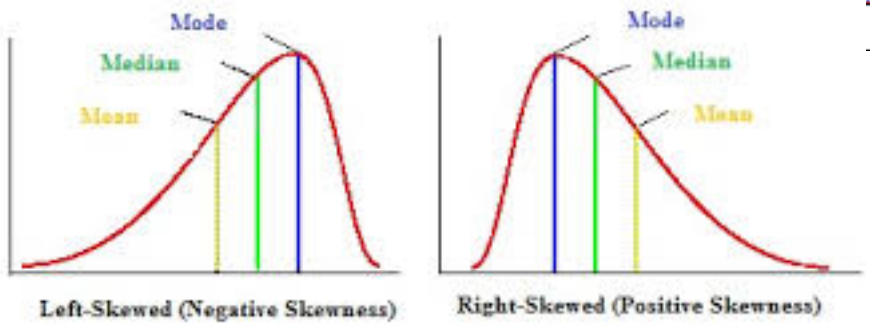
d) 범위를 구간으로 나눈다. [2, 4), [4, 6), [6, 8), ⋯, [16, 18) 혹은 [0, 5), [5, 10), [10, 15), [15, 20)

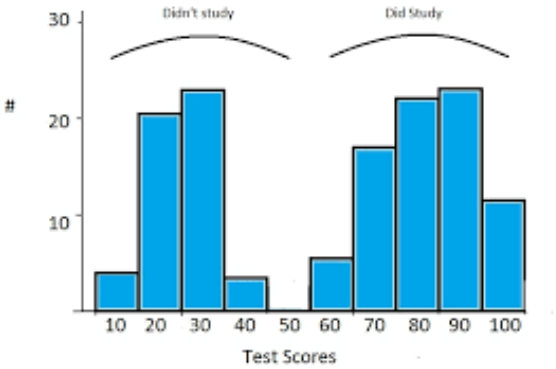
e) 각 구간의 범위에 속한 데이터 빈도를 구하고 빈도 크기를 막대로 표현하면 된다.

빈의 중앙 값을 연결하면 확률분포함수이다.

(진단 내용) a) 확률분포함수 형태 - 모집단 확률분포함수와 동일 (좌로 치우침, 우로 치우침, 죄우대칭) b) 봉우리 (최빈값)







두 개의 서로 다른 집단 데이터의 히스토그램을 그리는 경우 봉우리가 2개 나타나는 경우가 발생 - 상자 수염 그림으로는 판별 불가, 하여 정량적 데이터의 그래프 요약은 나무 상자 그림과 히스토그램 동시에 요약하는 것을 권한다. 옆의 히스토그램은 시험 공부한 집단과 그렇지 않은 집단의 시험 성적 히스토그램이다. -> 데이터 분리하여 분석

2) 숫자 요약

(1) 중앙 위치 center location measure

	크기 magnitude	순서 order
통계량	평균 mean	중위값 median
기호	$\mu, \bar{x}$	MD
공식	$\mu = \frac{\sum_{i=1}^{N} X_i}{N}, \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	$X_{(MD)}, x_{(MD)}$
장점	표본평균 샘플링 확률분포함수를 알 수 있어(중심극한정리) 신뢰구간 및 가설검정 추론이 가능하다. -집단 평균 비교 가능	치우침의 영향이 적어 중앙 위치 통계량으로 가장 적절
단점	좌우 대칭이 아닌 치우침 데이터는 중앙 위치의 왜곡이 있음	중위값 샘플링 확률분포함수를 구하기 어려워 모수 추론이 불가능 - 비모수추론

✴ 데이터의 분포를 좌우 대칭(정규변환)으로 만든 후 평균을 이용하는 것이 가장 적절

(예제 데이터) n=12

8 17 9 10 9 11 7 13 12 3 10 4

표본 평균 : $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{8 + 17 + \ldots + 4}{12} = 9.42$

중위값 (수작업 계산) : $\frac{x_{(6)} + x_{(7)}}{2} = (9 + 10)/2 = 9.5$ 엑셀 - MD=9.5 (동일)

(2) 산포 척도 spread measure

	크기 magnitude	순서 order
통계량	분산 variance 표준편차 standard deviation	범위 range, 사분위 범위 IQR
기호	σ^2, σ, s^2, s	R, IQR
공식	$\sigma^2 = \frac{\sum_i^N (X_i - \mu)^2}{N}, s^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n-1}$	$R = X_{(n)} - X_{(1)}$, $IQR = Q_3 - Q_1$
장점	표본분산 샘플링 확률분포함수를 알 수 있어 가설검정 추론이 가능하다.	치우침의 영향이 적어 산포 통계량으로 적절
단점	좌우 대칭이 아닌 치우침 데이터는 분산 크기가 왜곡이 있음	모수 추론이 불가능 - 비모수추론

✶ 평균의 크기가 다른 두 집단 분산의 비교 시 변동계수 CV 통계량 이용 : $CV = \frac{s}{\bar{x}}$*100 (%)

✶ 표본 표준편차의 분모에 n 대신, (n-1) 사용한 이유는 가장 좋은 추정치(MVUE 최소분산불편추정량)이기 때문임

(예제 데이터) n=12

8 17 9 10 9 11 7 13 12 3 10 4

표본 분산 = 3.8

$$s^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n-1} = \frac{(8-9.42)^2 + (17-9.42)^2 + \ldots + (9-9.42)^2}{11}$$

범위 : $R = 17 - 3 = 14$

사분위범위 : $IQR = 11.5 - 7.5 = 4$, (엑셀) 엑셀 - IQR=11.25-7.75=3.5

변동계수 variance coefficient : $CV = \frac{s}{\bar{x}} = \frac{3.8}{9.42} \times 100 = 0.4036$

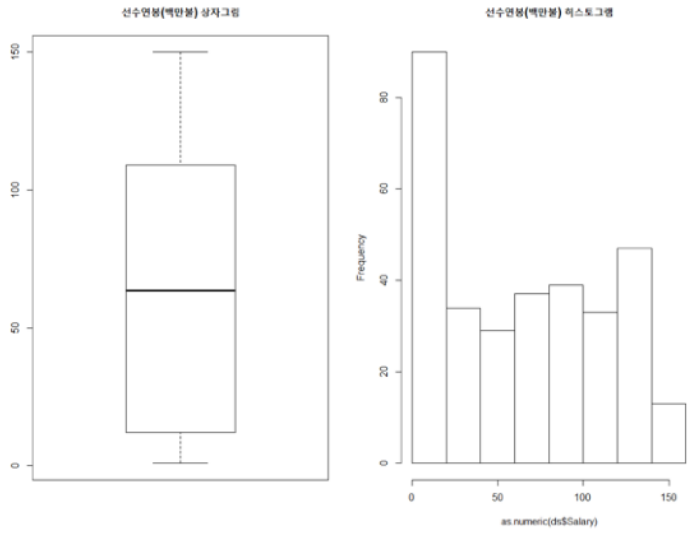
일변량 통계적 방법 요약

Variables	Numerical description	Graphical description	Parametric test	Non-parametric test
nominal	Frequencies (one-dimensional contingency table)	Bar plot Pie chart	---	Chi-square for a one-dimensional contingency table
scale	Descriptive statistics	Histogram Boxplot	Student's t for one variable	Sign test

nominal 명목, scale 측정, descriptive statistic : 기술 통계량

[**예제 데이터**] 선수들의 연봉(salary) 데이터 이용하여 다음을 구하시오. [엑셀 이용]

1) 나무상자그림, 히스토그램을 그리고 해석하시오.

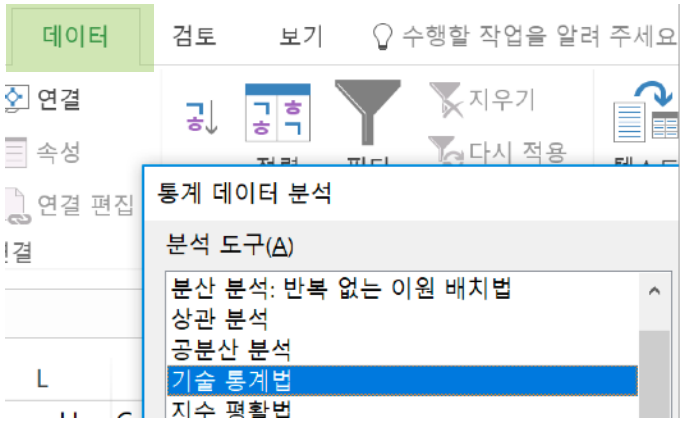


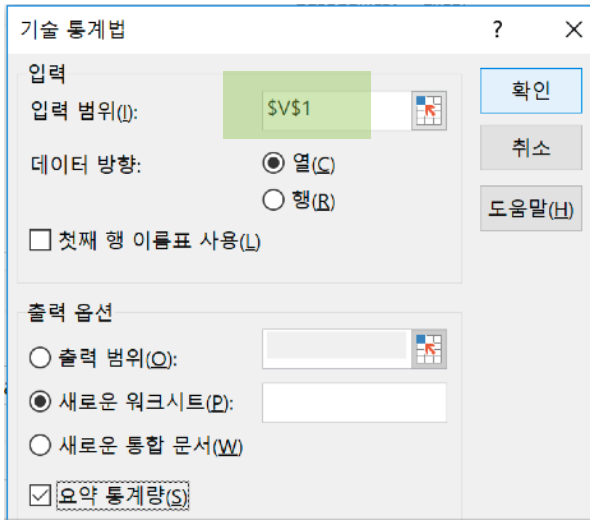
2) 주요 통계량을 구하시오.

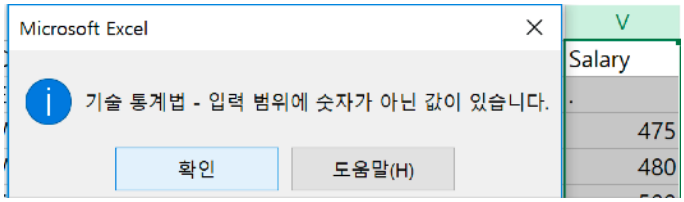
```
> summary(as.numeric(ds$Salary))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   12.25   63.50   64.17  108.75  150.00
```

```
> sd(as.numeric(ds$Salary))
[1] 49.20608
```

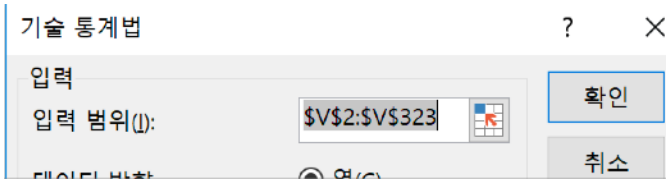
[엑셀]

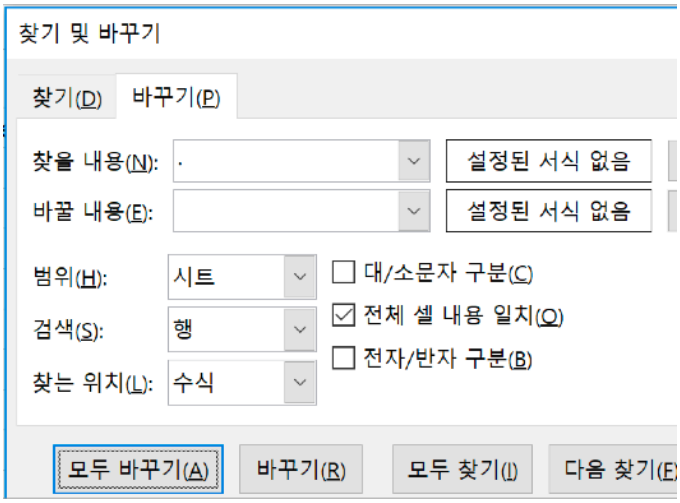




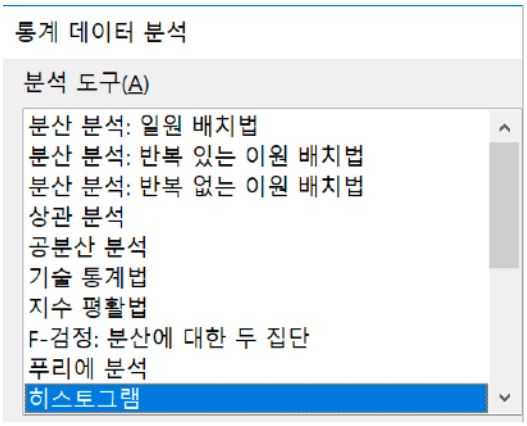


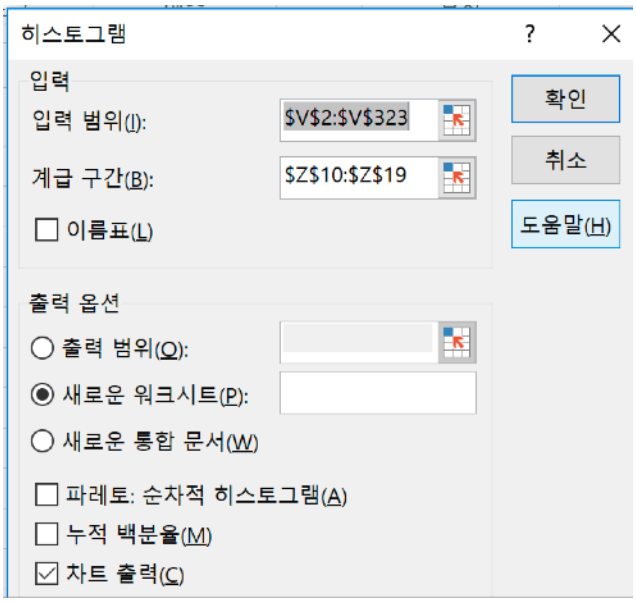
V 열 첫 행(v1)에 숫자가 아닌 변수명이 있고 결측치는 .으로 되어 있으므로 입력범위를 "$V$2:$V$323" 수정하고 . 을 찾기 바꾸기 메뉴(핫키 : ctrl+H)를 이용하여 . 을 공백으로 바꾼 후 "기술통계법" 확인 버튼을 누르면 실행된다.

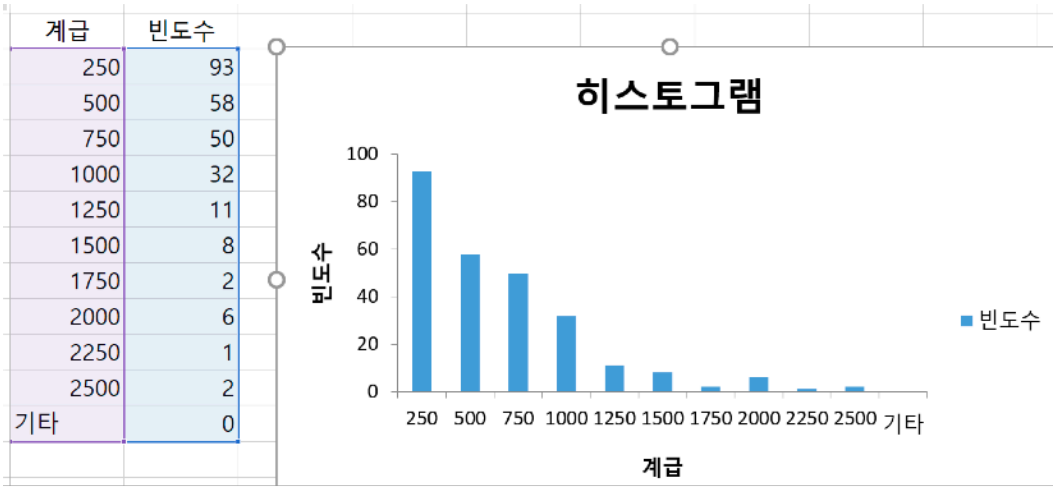




평균	535.97
표준 오차	27.82
중앙값	425.00
최빈값	750.00
표준 편차	451.10
분산	203494.85
첨도	3.06
왜도	1.59
범위	2392.00
최소값	68.00
최대값	2460.00
합	140959.00
관측수	263.00







나무 상자 그리기 : 엑셀 2016에서는 통계차트 삽입에 포함되어 있음