3-2 자료의 숫자 요약

3-2-1. 자료의 숫자 요약

자료의 숫자요약

자료의 숫자요약이란 자료의 관측값을 대표하는 통계량을 구하여 자료의 특성을 파악하는 통계적 방법

자료분포의 중심위치를 나타내는 통계량

: 평균(mean), 중위수(median), 최빈값(mode)

관측값의 흩어진 정도를 측정하는 통계량

: 분산(variance), 범위(range)

기본개념

변수 (variable)

- 문자를 이용해 자료를 표현하는 방법임
- 일반적으로 X, Y, Z와 같은 영문자로 표현함

모수 (parameter)

- 모집단을 대표하는 값. 일반적으로 알려져 있지 않음
- 그리스 문자인 $\alpha, \beta, \gamma, \mu, \delta$ 등을 이용해 표현

통계량 (statistic)

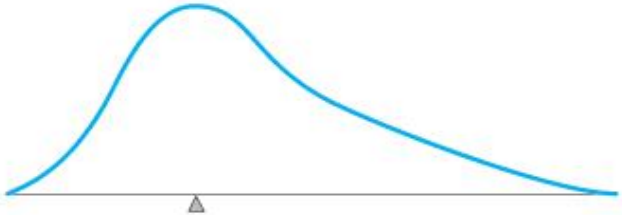
- 표본으로부터 얻은 자료의 대표값
- 통계량 중에서 모수를 추정하는 값을 추정량(estimator)이라 함

3-2-2. 자료의 중심 요약

자료의 중심을 측정하는 통계량

평균 (mean ; $\bar{X}$)

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$



중위수 (median ; $\tilde{X}$)

- 자료를 크기 순서로 정리했을 때 가운데에 위치하는 관측 값
- 소수의 크거나 작은 관측 값(이상치, outlier)에 의해 영향을 받지 않음

최빈값 (mode ; M_0)

- 관측 도수가 가장 많은 값
- 하나 이상의 값 존재 가능
- 소수의 이상치에 영향을 받지 않음

(참고) 평균의 종류

산술평균(mean ; $\bar{X}$) $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

조화평균(harmonic mean; H) $\frac{1}{H} = \frac{1}{n}(\frac{1}{X_1} + \frac{1}{X_2} + \cdots + \frac{1}{X_n})$

두 값 X_1, X_2의 조화평균은 다음과 같음

$\frac{1}{H} = \frac{1}{2}(\frac{1}{X_1} + \frac{1}{X_2}) = \frac{X_1 + X_2}{2X_1X_2}$

$\therefore H = \frac{2X_1X_2}{X_1 + X_2}$

- 일반적으로 역수의 형태로 된 변수를 평균할 경우 조화평균을 사용

(예)자동차가 처음 10km를 시속 30km로 달리고 다음 10km를 시속 60km로 달렸을 경우 평균시속은 얼마인가?

(wrong) $\frac{1}{2}(30+60) = 45\,(\text{km/h})$

(right) $\frac{2}{\frac{1}{30}+\frac{1}{60}} = 40\,(\text{km/h})$

기하평균(geometric mean ; G)

$$G = \sqrt[n]{X_1 X_2 \ldots . X_n} \quad \text{or} \quad logG = \frac{1}{n}\sum_{i=1}^{n} logX_i$$

- 기하평균의 log값(상용로그 또는 자연로그)은 비율로 나타낸 변수의 log값들의 산술평균이며, 이 값의 anti-log가 기하평균 G임
- 기하평균은 비율, 백분율을 평균하는 경우에 적합하며 일반적으로 산술평균보다 작음
- 연평균증가율을 구할 경우 먼저 비율로 나타내고 다음에 비율을 이용하여 G를 구하면 됨

(예) 2016년 연봉이 3,000만원인 K씨의 2017년 연봉인상률이 5%, 2018년 연봉인상률이 15%였다면 2년 평균 연봉인상률은 얼마일까?

•2017년 연봉 : 3,000x(1+0.05)=3,150

•2018년 연봉 : 3,150x(1+0.15)=3,622.5 즉, 3,622.5 = 3,000x(1+0.05)x(1+0.15)

•평균인상률(x) : 3,622.5 = 3,000(1+x)(1+x) =3,000(1+x)^2

∴평균인상률 = $\sqrt{\frac{3,622.5}{3,000}}$ - 1=0.098863(=9.8863%)

$$G = \left(\left(\frac{\text{최종년도의 값}}{\text{최초년도의 값}}\right)^{\left(\frac{1}{n}\right)} - 1\right) \times 100$$

단, n은 경과연도

(예)연도별 인구변화

연 도	인구	전년도대비비율(%)	$\overline{X}$ × 전년도인구	G × 전년도인구
1990	5000			
1991	6000	120	6133	6128
1992	7800	130	7523	7510
1993	9204	118	9228	9204
	G=122. 56	$\overline{X}$=122. 67		

중위수를 구하는 방법

1. n이 홀수인 경우

$\left(\frac{n+1}{2}\right)$ 이 n개의 가운데 위치이므로 관측값 중에서 순서상 $\frac{n+1}{2}$번째 값이 중위수이다. 예를 들어 $n=7$인 경우, $\frac{7+1}{2}=4$ 이므로 관측값을 크기순서로 나열하였을 때 4번째 순서에 있는 값이 중위수이다.

2. n이 짝수인 경우

짝수인 경우에 있어서 가운데 위치는 $\frac{n}{2}+0.5$ 가 된다. 이 경우에는 가운데 위치에 가장 가까운 값인 $\frac{n}{2}$ 번 째 순서값과 $\frac{n}{2}+1$ 번 째 순서값을 구하여 두 값의 평균을 중위수로 한다. 즉, $\tilde{X}=\frac{1}{2}(X_{(\frac{n}{2})}+X_{(\frac{n}{2}+1)})$ 이다.

예를 들어 $n=10$ 인 경우 가운데 위치는 $\frac{10+1}{2}=5.5$ 이며, 중위수는 $\frac{n}{2}=\frac{10}{2}=5$ 번 째 순서값과 $\frac{n}{2}+1=6$ 번 째 순서값의 평균이 된다.

평균의 특징

1. 자료 관측값의 산술평균임
2. 각 자료에 있어서 유일하게 구하여짐
3. 소수의 매우 크거나 작은 값에 의하여 영향을 받음
4. 자료를 몇 개의 작은 집단으로 나누었을 때 각 집단의 평균의 평균은 전체 자료를 이용하여 구한 평균과 같음
5. 수량으로 관측된 자료에만 이용 가능함

중위수의 특징

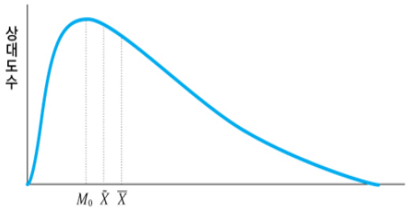
1. 중앙위치의 값으로 관측값의 50%가 왼쪽에, 그리고 나머지 50%가 오른쪽에 존재함
2. 각 자료에 있어서 유일하게 구하여짐
3. 소수의 매우 크거나 작은 값에 의하여 영향을 받지 않음
4. 자료를 몇 개의 작은 집단으로 나누었을 때 각 집단의 중위수의 중위수는 전체자료를 이용하여 구한 중위수와 항상 일치하지는 않음
5. 수량으로 관측된 자료에만 이용 가능함

최빈값의 특징

1. 자료에서 관측빈도의 수가 가장 많은 값임
2. 각 자료에는 하나 이상의 최빈값이 있을 수 있음
3. 소수의 극한값에 영향을 받지 않음
4. 자료를 몇 개의 작은 집단으로 나누었을 때 각 집단의 최빈값에 의하여 전체의 최빈값을 유도할 수 없음
5. 양적으로 측정된 자료와 질적으로 측정된 자료 모두에 이용 가능함

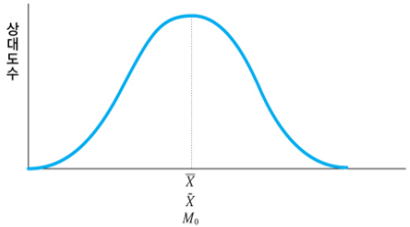
세 통계량의 관계

자료의 분포모양에 따른 평균($\overline{X}$), 중위수($\tilde{X}$), 최빈값(M_0)의 위치



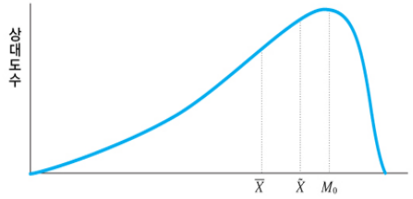
(a) 오른쪽 긴 꼬리 자료

자료의 분포모양에 따른 평균($\overline{X}$), 중위수($\tilde{X}$), 최빈값(M_0)의 위치



(b) 대칭인 자료

자료의 분포모양에 따른 평균($\overline{X}$), 중위수($\tilde{X}$), 최빈값(M_0)의 위치



(c) 왼쪽 긴 꼬리 자료

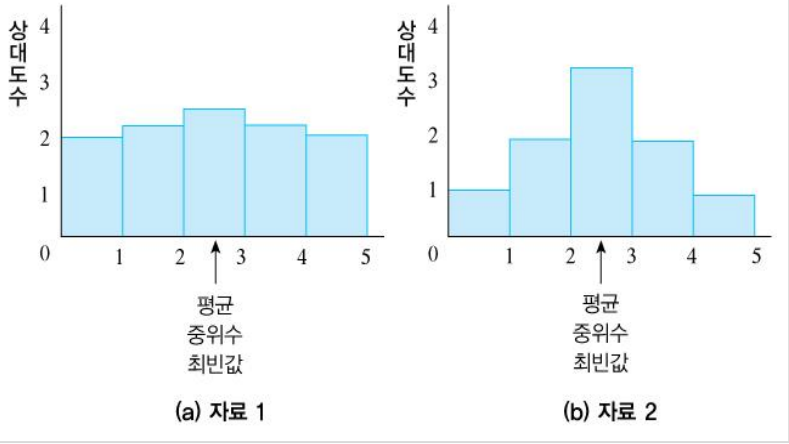
3-2-3. 자료의 변화량 요약

자료의 흩어진 정도를 측정하는 통계량

- 수집된 자료의 특성을 파악하는 대표값은 자료의 중심을 측정하는 통계량과 함께 **자료의 흩어진 정도(변화량 : variability)를 측정하는 통계량**이 있음

- 자료의 변화량의 측정은 자료가 어느 정도 중심에 집중되어 있는가를 측정하여 중심의 대표성에 대한 평가와 함께 자료분포의 구조적 특성을 파악할 수 있도록 함

그림 **동일한 평균, 중위수, 최빈값을 갖는 두 자료**



옆의 그림에서 자료 1 과 자료 2 는 모두 좌우대칭이고 평균, 중위수, 최빈값이 각각 일치하며 두 자료집단이 모두 동일한 값

- ▶ 자료 1 : 거의 모든 구간에서 관측값들의 상대도수가 비슷하게 분포되어 있음.
- ▶ 자료 2 : 중심에 집중적으로 관측값들이 분포되어 있음
- ▶ 중심에 대한 집중도를 고려하면 자료 2 의 평균이 자료 1의 평균보다 중심을 나타내는 대표값으로서의 의미가 크다고 할 수 있음

범위 (range ; R)

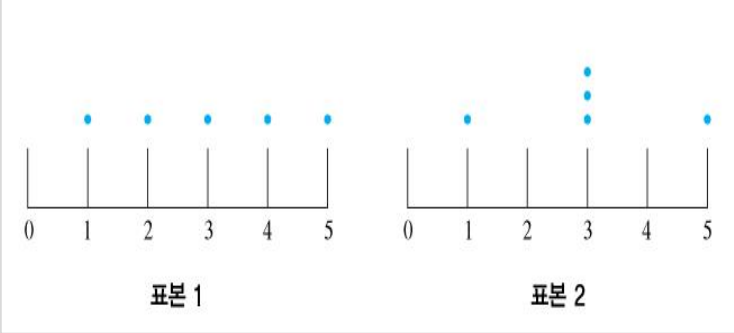
- 관측된 자료들 중에서 가장 큰 값과 가장 작은 값의 차이
- 자료 $\{X_1, X_2, \cdots, X_n\}$ 중에서

$X_{(1)}$ = 가장 작은 값, $X_{(n)}$ = 가장 큰 값 ➡ $R = X_{(n)} - X_{(1)}$

표 두 표본의 분포표

	관측값	평균	중위수
자료 1	1, 2, 3, 4, 5	3	3
자료 2	1, 3, 3, 3, 5	3	3

그림 두 표본의 분포



▶ 표에 나타난 두 자료는 평균과 중위수가 모두 3 으로 동일하고 범위 R=5-1=4 도 서로 동일함

▶ 그러나 자료 1과 자료 2는 분포의 형태가 전혀 상이함을 알 수 있다. 자료 1 과 자료 2 의 범위는 동일하나 자료 1 이 자료 2에 비하여 흩어진 정도가 더 크다고 판단할 수 있음

→ 범위는 쉽게 구할 수 있는 통계량이나 많은 정보를 제공하지 못함

분산 (variance ; S^2)

분산은 자료의 흩어진 정도를 측정하는 가장 일반적으로 쓰이는 통계량

- 자료집단 : $\{X_1, X_2, \ldots, X_n\}$
- 평균 : $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- 편차 (deviation) : $\{X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X}\}$
- 변동 (variation) : $\sum (X_i - \bar{X})^2$
- 분산 (variance) : $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\{\sum_{i=1}^{n} X_i^2 - \frac{1}{n}(\sum_{i=1}^{n} X_i)^2\}$

분산의 특징

자료에서 편차의 합은 항상 0이 되며, 이 편차를 이용하여 자료의 흩어진 정도를 측정하는 통계량이 분산임

- 대부분의 관측값들이 평균에 가까이 있으면(편차가 작으면)
  분산의 값이 작아짐
- 대부분의 관측값들이 평균에서 멀리 떨어져 있으면(편차가 크면)
  분산의 값이 커짐

분산 계산

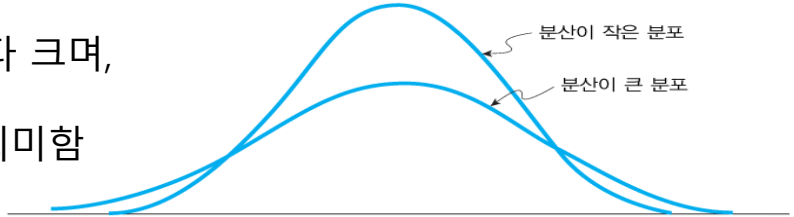
표 두 표본의 분포표

	관측값	평균	중위수
자료 1	1, 2, 3, 4, 5	3	3
자료 2	1, 3, 3, 3, 5	3	3

<자료 1> $S^2 = \frac{1}{4}\{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2\} = \frac{10}{4} = 2.5$

<자료 2> $S^2 = \frac{1}{4}\{(1-3)^2 + (3-3)^2 + (3-3)^2 + (3-3)^2 + (5-3)^2\} = \frac{8}{4} = 2$

즉, <자료 1>의 분산이 <자료 2>의 분산보다 크며,
이는 <자료 1>의 흩어진 정도가
<자료 2>의 흩어진 정도보다 크다는 것을 의미함

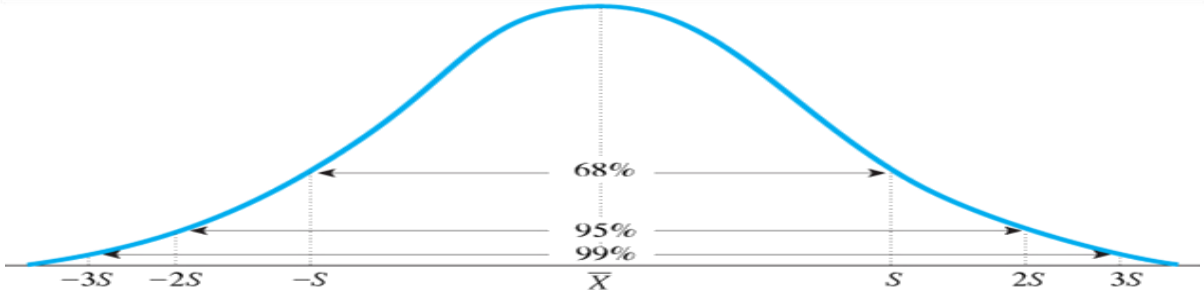


표준편차(standard deviation ; S)

분산의 양의 제곱근 : $S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$

평균을 중심으로 한 범위와 분포의 비율

범위	분포의 빈도비율
$\bar{X} \pm S$	68%
$\bar{X} \pm 2S$	95%
$\bar{X} \pm 3S$	99%



자료의 관측값의 수가 많으며 그 분포가 좌우대칭인 종모양과 같은 경우
경험적으로 관측값들이 평균을 중심으로 분포되어 있음

예

2007년도 대입 모의수능시험을 1,000 명을 대상으로 실시한 결과,
평균이 200 점, 표준편차가 15 점이었음. 모의고사 성적의 분포가
좌우대칭인 종모양과 같다고 할 때 경험적으로 다음과 같이 말할 수 있음

- ▶ 200±15점(185점~215점) 사이에 약 68%(680명 정도)가 있으며,
- ▶ 200±30점(170점~230점) 사이에 약 95%(950명 정도)가 있고,
- ▶ 200±45점(155점~245점) 사이에 약 99%(990명 정도)가 있다고 볼 수 있음

순서통계량

•아래사분위수 (lower quartile; Q_1) : 관측값의 25% 순서에 있는 값

•중위수(median) : 관측값의 50% 순서에 있는 값

•위사분위수 (upper quartile; Q_3) 는 관측값의 75% 순서에 있는 값

중위수와 사분위수의 위치 결정 공식

n개의 관측값 $X_1, \cdots, X_n$ 이 작은 값으로부터 올림차순으로 정리되어 있을 때

중위수와 사분위수의 위치는 다음 공식에 의해 구함. 단, [x]는 x의 가장 큰 정수

$$\text{중위수} = \frac{n+1}{2} \quad Q_1 = \frac{[\text{중위수 위치}] + 1}{2} \quad Q_3 = [\text{중위수 위치}] + \frac{[\text{중위수 위치}] + 1}{2}$$

(예)n=50인 경우

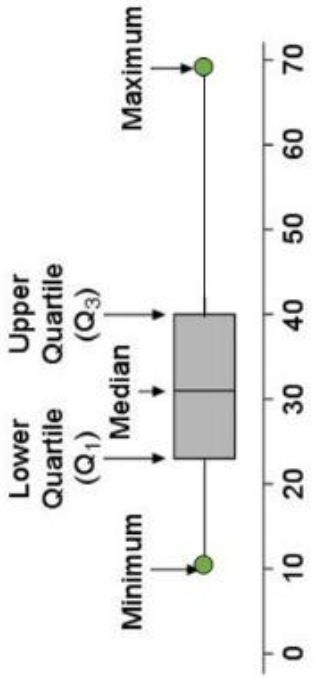
$$Me = \frac{50+1}{2} = 25.5 \quad Q1 = \frac{[25.5]+1}{2} = 13 \quad Q3 = [25.5] + \frac{[25.5]+1}{2} = 38$$

3-2-4. 상자그림

상자 그림(Box Plot)

- 상자그림이란 5가지 순서통계량 (최소값, 아래사분위수,중위수, 위사분위수, 최대값)을 이용하여 자료를 요약 정리하는 그래프적 표현방법

- 즉, 상자그림이란 Q_1과 Q_3를 연결하는 상자를 그리고, 그 상자 안에 중위수를 나타내는 선을 그리며, 최소값과 Q_1, 그리고 Q_3 와 최대값을 선으로 연결하는 표현방법으로 5개의 순서통계량의 위치를 관찰하여 자료분포의 특징을 알 수 있음

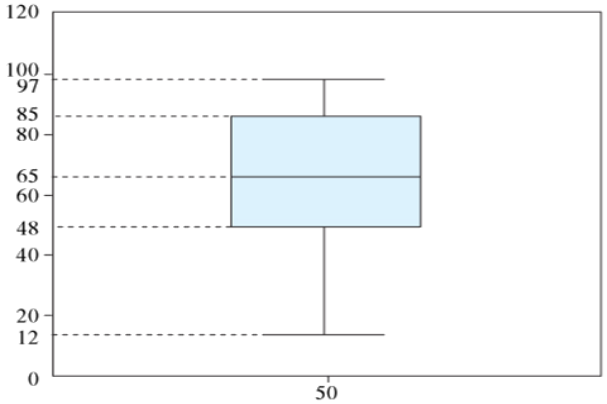


Boxplot.R

```
table2_3<-c(65,62,73,85,65,46,36,49,81,76,
            60,44,43,72,21,33,83,46,64,49,
            12,74,91,78,60,48,24,62,54,97,
            69,31,89,96,96,97,86,88,85,61,
            95,54,85,89,51,77,81,72,47,35)
summary(table2_3)
quantile(table2_3)
(sort(table2_3))
fivenum(table2_3)
boxplot(table2_3)
```

```
> summary(table2_3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.00   48.25   65.00   64.74   84.50   97.00
> quantile(table2_3)
   0%   25%   50%   75%  100%
12.00 48.25 65.00 84.50 97.00
> (sort(table2_3))
 [1] 12 21 24 31 33 35 36 43 44 46 46 47 48 49 49 51 54 54 60 60 61 62 62 64 65 65 69 72 72 73 74 76 77
[34] 78 81 81 83 85 85 85 86 88 89 89 91 95 96 96 97 97
> fivenum(table2_3)
[1] 12 48 65 85 97
> boxplot(table2_3)
```

그림 자료의 상자그림



•해석 : Max-Q3=12, Q1-Min=36이므로 관측치가 아래사분위수(Q1)와 최솟값 사이보다는 위사분위수(Q3)와 최댓값 사이에 밀집되어 분포하고 있음

•또한, 가운데 50%만 고려하면 Q3-Me=20, Me-Q1=17이므로 관측치가 위사분위수(Q3)와 중위수(Me) 사이보다는 중위수(Me)와 아래분위수(Q1) 사이에 밀집되어 분포하고 있음