

중심에 대한 측도

📖 평균

📖 중앙값

📖 최빈값

📖 중심의 이동과 기울어짐의 변화에 따른 값의 변화

Loren ipsum dolor sit amet, ius an molestie facilisi erroribus, mutat nalerum delectus ei vis. Has ornatus conclusionemque id, an videri molestatis sit. In etqui praesent sit. An vel agan porro comprehensan, ad ludus constituto nea, et ius utroque scaevola assuaverit.

Vis cu nodus nulla feugait, oratio facilisi in usu, elit vitae sea te. Ea fabulas accusamus dissentias sea, facete tacinales definitiones et per. Nihil dicant mediocram pro eu, no mei nostro sensibus platonem. Qui id sunno perpetus neglegentur. Vel ipsum novum copiosae ut. Quo et liber detracto probatus. Nam augue scribentur an. Sea oporteat percipitur incidereat at. Qui viris nemore an.



사용할 자료

x의 정의

```
> x <- jitter(seq(1,10))
```

+ seq(1, 10) : 1부터 10사이의 자연수

+ jitter 함수

➤ 원래 값에 임의의 난수를 더해주어 난수로 만들기 때문에 벡터 x의 값은 다음과 같음

```
[1] 1.138562 1.800636 2.977108 4.190691 4.814402 6.025597 6.859326 7.833420  
[9] 9.004932 9.947944
```

* 난수이므로 실행할 때마다 다른 값이 나오에 유의

평균



반응

-  자료의 합을 자료의 개수로 나눈 값으로 지나치게 큰 값이나 작은 값에 빨리 반응함

계산

-  mean 함수로 계산함

```
> mean(x)
[1] 5.459262
```



중앙값



반응

- + 자료를 크기순으로 나열할 때 중앙에 위치한 값으로 지나치게 큰 값이나 작은 값에 늦게 반응함

개념

- + 자료의 개수가 홀수이면 중앙에 한 값이 위치하고 개수가 짝수이면 중앙에 위치한 두 값의 평균값
- + 중위값, 중위수 등으로도 부름

계산

- + median 함수로 계산함

```
> median(x)  
[1] 5.419999
```



중앙값과 평균의 성질



'지나치게 큰 값과 작은 값' 계산 사례 보기



자료 1

12명의 월급여를 조사하여서 얻은 자료를 x1이라고 하자. (c(200, 170, 250, 230, 220, 300, 350, 300, 330, 260, 270, 250))

➤ 이 때의 평균과 중앙값

```
> mean(x1)
[1] 260.8333
> median(x1)
[1] 255
```



자료 2

한 명 더 조사하였는데 우연히 그 직장의 CEO가 포함되었다. 월 5,000인 이 자료가 포함된 경우(다른 값에 비해 지나치게 큰 값)를 x2라고 하자.

➤ x2

```
> x2 <- c(x1, 5000)
```

➤ 이 때 평균과 중앙값

```
> mean(x2)
[1] 625.3846
> median(x2)
[1] 260
```

* 평균이 중앙값에 비해 크게 올라감

평균의 오류

평균의 함정




최빈값



개념

-  자료 중 빈도수가 가장 높은 값

계산

-  R-언어의 내장함수에는 최빈값을 계산하는 함수가 없으므로 몇 개의 함수를 연속적으로 호출하는 방법으로 계산함

```
names(sort(table(x), decr=T))[1]
```



최빈값

계산의 작동

- ① 자료와 빈도표 : 자료가 있을 때 table 함수로 빈도표를 구함

```
> x <- c("A", "B", "C", "D", "B", "C", "D", "C", "D", "D")  
> table(x)  
x  
A B C D  
1 2 3 4
```

- ② 빈도표를 순서대로(desc=T 옵션을 사용하여 역순이므로 빈도가 높은 것이 처음 나옴) Sorting함

```
> sort( table(x), decr=T )  
x  
D C B A  
4 3 2 1
```

*** 결과는 가장 빈도가 높은 순서로 빈도표가 얻어짐**

최빈값

계산의 작동

- ③ 빈도표의 역순에서 가장 처음 나오는 표가 가장 빈도가 높으므로 [1]을 사용하여 첫 번째 표를 얻음

```
> sort( table(x), decr=T )[1]  
D  
4
```

- ④ 빈도표에서 가장 자주 나온 것의 이름을 찾았으므로 이 표에서 표의 header를 찾으면 이 값이 최빈값임

```
> names(sort( table(x), decr=T ))[1]  
[1] "D"
```

* D가 최빈값으로 가장 빈도가 높은 값



‘실자료 최빈값’ 계산 보기

사용할 자료 - 1

```
> BMI <- read.table(url("http://jupiter.hallym.ac.kr/ftpdata/data/bmi.txt"),  
  col.names=c("height", "weight", "year", "religion", "gender", "marriage"))
```

+ 자료출처

- url 함수를 사용하여 인터넷 사이트에서 바로 읽어 올 수 있으며, 다음의 인터넷주소를 브라우저 창에 붙여 넣으면 브라우저에서 볼 수 있음(<http://jupiter.hallym.ac.kr/ftpdata/data/bmi.txt>)

+ 내용 : 2000년, 177명에 대한 조사 결과

- 키, 몸무게, 출생년도
- 종교(Bu=불교, C1=개신교, C2=가톨릭, No=없음)
- 성별(F=여자, M=남자)
- 결혼여부(N=미혼, Y=기혼)

최빈값

'실자료 최빈값' 계산 보기

자료의 일부(처음 부분)

- R의 내장함수인 head 와 tail 함수는 데이터 프레임 등의 첫 부분 및 마지막 부분의 일부를 보여줌(기본 6줄)

```
> head(BMI)
```

```
height weight year religion gender marriage
```

```
1 167 68 1974 No F N
```

```
2 162 49 1974 C2 F N
```

```
3 158 50 1978 C2 F N
```

```
4 165 56 1977 No F N
```

```
5 160 52 1959 No F N
```

```
6 162 52 1972 No F Y
```

- 이 자료에서 종교에 대한 최빈값을 얻어 보면 다음과 같이 '없음'이 얻어짐

```
> names(sort(table(BMI$religion), decreasing=T))[1]
```

```
[1] "No"
```



중심의 이동과 기울어짐의 변화에 따른 값의 변화



중심의 이동에 따른 값의 변화

+ 반응

- 원 자료에 특정한 값을 더하거나 빼주면 더하거나 빼준 값만큼 평균, 중앙값 등도 함께 이동함

+ 계산

- 앞에서 사용한 x에 3을 더해 주어 y라고 하고 y의 평균과 중앙값 계산함

```
> y <- x + 3
> y
[1] 4.138562 4.800636 5.977108 7.190691 7.814402 9.025597 9.859326
[8] 10.833420 12.004932 12.947944
> mean(y)
[1] 8.459262
> median(y)
[1] 8.419999
```

* 원래의 평균과 중앙값보다 정확하게 3만큼 증가



중심의 이동과 기울어짐의 변화에 따른 값의 변화



중심의 이동에 따른 값의 변화



+ 반응

- 자료가 오른쪽 또는 왼쪽으로 기울어진 경우(Skewed to the right/left) 평균이 지나치게 큰 값이나 작은 값에 예민하므로 먼저 반응하고 중앙값은 느리게 반응함
- 여기서는 분포의 기울어짐에 따라 평균과 중앙값이 어떻게 달라지는지 확인함





중심의 이동과 기울어짐의 변화에 따른 값의 변화



기울어짐에 따른 값의 변화

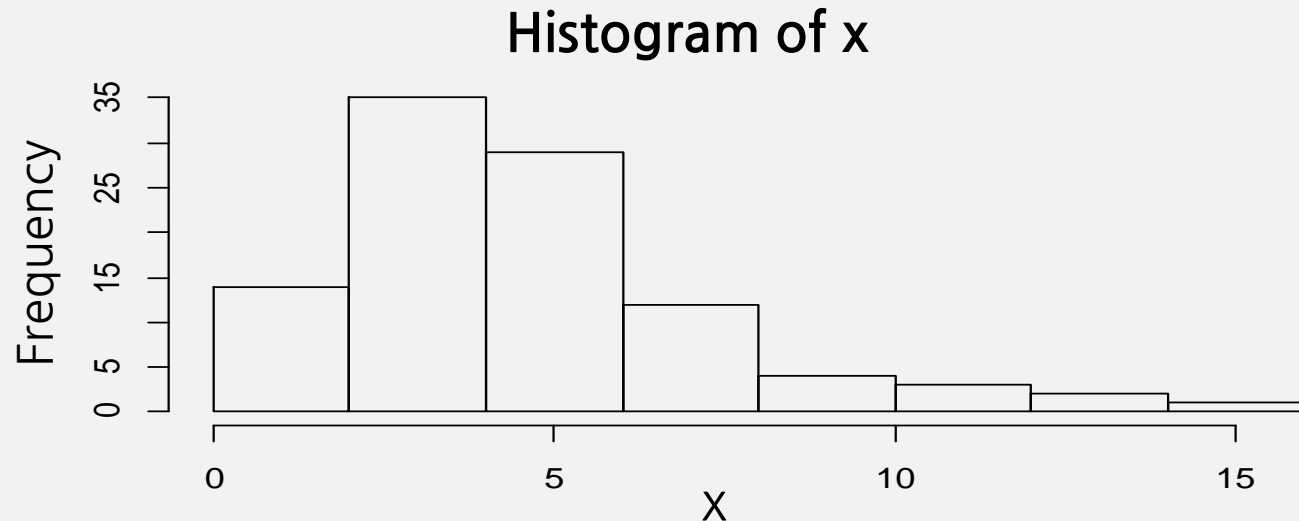


오른쪽으로 기울어진 분포의 경우 : 지나치게 큰 값이 존재하는 경우

```
> x <- rchisq(100,5)
```

* 자유도 5인 카이제곱분포에서 난수 100개를 만들며 이 분포는 오른쪽으로 기울어진 분포임

```
> hist(x)
```



```
> mean(x)
```

```
[1] 4.446424
```

```
> median(x)
```

```
[1] 4.007943
```

* 오른쪽으로 기울어진
(skewed to the right)
분포에서는 평균이
중앙값보다 크게 얻어짐



중심의 이동과 기울어짐의 변화에 따른 값의 변화



기울어짐에 따른 값의 변화

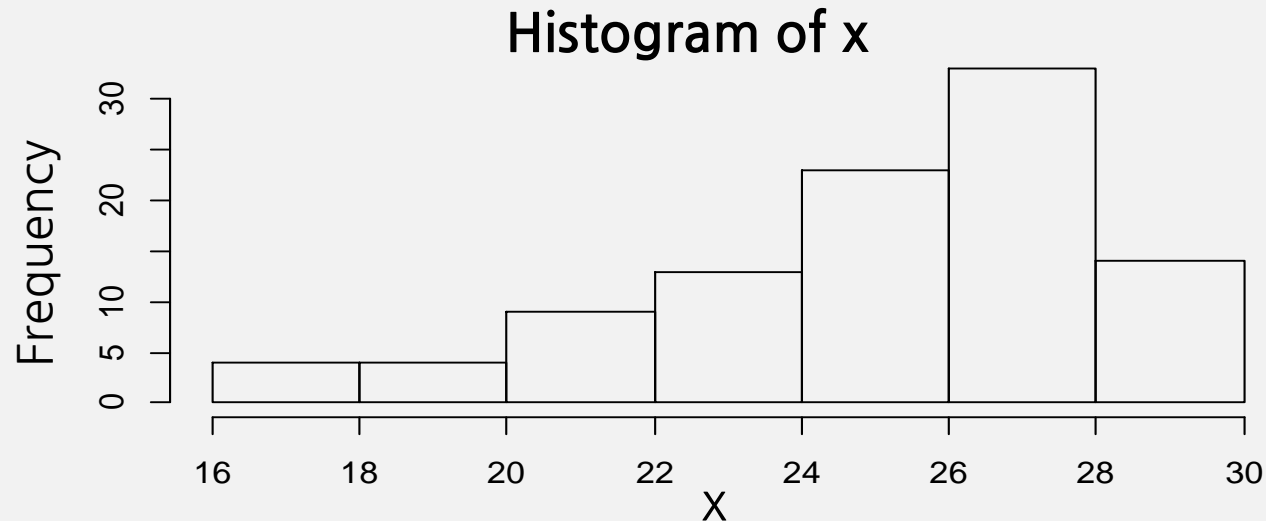


왼쪽으로 기울어진 경우 : 지나치게 작은 값이 존재하는 경우

```
> x <- 30 - rchisq(100,5)
```

* 30- (자유도 5인 카이제곱분포에서 난수) 100개이며 이 분포는 왼쪽으로 기울어진 분포임

```
> hist(x)
```



```
> mean(x)  
[1] 24.98758  
> median(x)  
[1] 25.78768
```

* 왼쪽으로 기울어진
(skewed to the left)
분포에서는 평균이
중앙값보다 작게 얻어짐

분위수 계산

📖 분위수(quantile)

📖 R에서의 분위수 계산

Loren ipsum dolor sit amet, ius an molestie facilisi erroribus, mutat natorum delectus ei vis. Has ornatus conclusionemque id, an vide molestatis sit. In etqui praesent sit. An vel agan porro comprehensan, ad ludus constituto nea, et ius utroque scaevola assuaverit.

Vis cu nodus nulla feugait, oratio facilisi ex usu, eili vitae sea te. Ea fabulas accusamus dissentias sea, facete tacinates definitiones et per. Nihil dicant mediocrem pro eu, no mei nostro sensibus platonem. Qui id sunno perpetua neglegentur. Vel ipsum novum copiosae ut. Quo et liber detracto probatus. Nam augue scriben- tur an. Sea oporteat percipitur incidereit at. Qui viris nemore an.



분위수(quantile)



분위수에 대한 통계학적 엄밀한 정의

- + 어떤 값 x 에 대해 x 보다 같거나 작을 확률은 p 이상, x 보다 같거나 클 확률은 $(1-p)$ 이상인 값 x 를 제 p 번째 분위수라고 함
- + 직관적으로는 자료를 오름차순으로 정리할 때 p 번째에 해당하는 값($0 < p < 1$)



백분위수(percentile)

- + 분위수를 백분위로 바꾼 $100p\%$ 에 해당하는 값





R에서의 분위수 계산



quantile 함수

```
quantile(x, probs = seq(0, 1, 0.25), ...)
```

+ 매개변수

- x : 분위수를 계산할 자료를 포함한 벡터
- probs : 분위의 값으로 기본값은 c(0, 0.25, 0.5, 0.75, 1)로 최소, 25%, 50%(중앙값; 중위수), 75% 및 최댓값을 설정



R에서의 분위수 계산



계산하기



자료

자료가 1부터 100까지의 자연수일 때 기본 분위수 5개를 계산해보자.

```
> x <- seq(1,100)
> quantile(x)
  0%   25%   50%   75%  100%
1.00 25.75 50.50 75.25 100.00
```



5%와 95% 백분위수 계산

```
> quantile(x, probs=c(0.05, 0.95))
  5%   95%
5.95 95.05
```

흠어짐에 대한 측도

📖 분산과 표준편차

📖 범위

📖 IQR과 CV(변동계수; Coefficient of Variation)



분산과 표준편차



var과 sd 함수 사용



자료

자료가 1부터 100까지의 자연수일 때 분산과 표준편차를 계산해보자.

```
> var(x)
[1] 8.986419
> sd(x)
[1] 2.997736
```





범위



범위의 계산

+ 범위는 최대와 최소의 차이로 range 함수나 max와 min 함수를 동시에 사용하여 계산함



자료

자료가 1부터 100까지의 자연수일 때 자료 x의 범위를 range함수로 계산해보자.

```
> x <- seq(1,100)
> range(x)
[1] 1 100
```

range 함수는 최소와 최대를 한 번에 얻어주며 이 차이를 계산하지는 않음



범위



범위의 계산

+ 범위 = 2번째 값(최대) - 첫 번째 값(최소)

```
> range(x)[2]- range(x)[1]  
[1] 99
```

+ min 함수와 max함수의 이용 : 각각 최솟값 및 최댓값을 계산해줌

```
> max(x) - min(x)  
[1] 99
```





IQR과 CV(변동계수; coefficient of variation)



사분위수(quartiles)

사분위수	제1사분위수	제2사분위수	제3사분위수
해당 백분위	제25번째 백분위	제50번째 백분위	제75번째 백분위

➤ 앞에서 본 quantile 함수는 기본값으로 최솟값, 제1사분위수, 제2사분위, 제3사분위수 및 최댓값을 얻게 됨



사분위수 범위(IQR; interquartile range)

+ 제3사분위수와 제1사분위수의 차이

+ IQR 함수를 사용하여 계산함



자료

자료가 1부터 100까지의 자연수일 때 IQR 함수를 사용하여 사분위수 범위를 계산해보자.

```
> IQR(x)
[1] 49.5
```



IQR과 CV(변동계수; coefficient of variation)



변동계수

- + 표준편차를 평균으로 나눈 후 100을 곱한 값
- + 원래의 측정단위가 아예 다르거나 원자료의 크기의 차이가 많이 나는 경우에 사용하면 합리적인 결과를 얻음



변동계수 계산

- + 변동계수를 한 번에 계산하는 R의 기본 내장함수는 없으므로 다음과 같이 계산함

```
> sd(x)/mean(x)*100  
[1] 57.4485
```





IQR과 CV(변동계수; coefficient of variation)



‘주식의 변동계수’ 계산 사례 보기



자료

다음은 두 회사주식의 7영업일간의 종가이다. 어느 주식의 변동성이 높다고 할 수 있는가?

	1일	2일	3일	4일	5일	6일	7일
주식1	100,000	107,000	113,000	95,000	103,000	98,000	95,000
주식2	1,000	900	1,100	1,200	1,000	800	700





IQR과 CV(변동계수; coefficient of variation)



‘주식의 변동계수’ 계산 사례 보기

```
x1 <- c( 100000, 107000, 113000, 95000, 103000, 98000, 95000)
x2 <- c(   1000,    900,   1100,  1200,   1000,   800,   700)
```

① 각각의 표준편차를 계산함

```
> sd(x1)
[1] 6629.659
> sd(x2)
[1] 171.8249
```

* 주식1의 표준편차가 큼

② 두 주식가격의 변동계수를 계산함

```
> sd(x1)/mean(x1)*100
[1] 6.527091
> sd(x2)/mean(x2)*100
[1] 17.95186
```

* 주식2의 변동계수가 큼

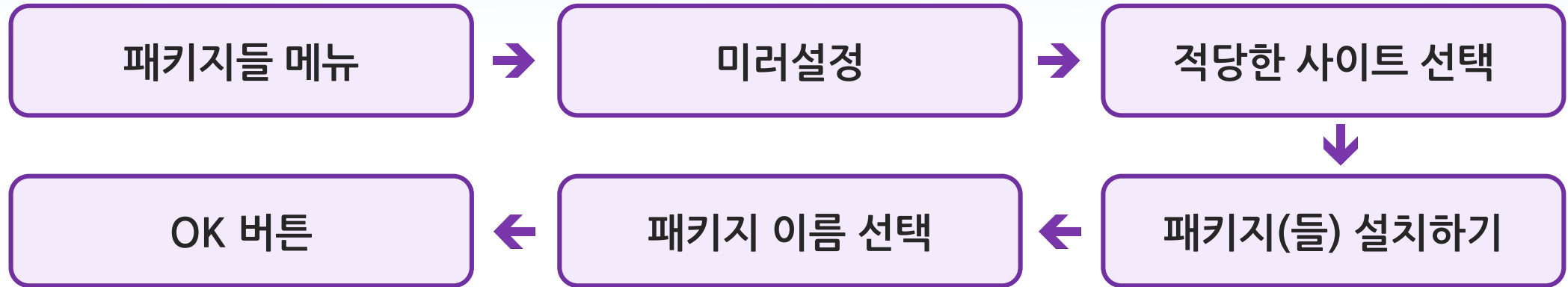
주식1은 5% 내외의 등락을 보이거나 주식2는 10-20% 내외의 등락을 보이므로 **주식2의 변동성이 크다고 할 수 있으며, 이를 잘 반영하는 것이 표준편차보다는 변동계수임**



IQR과 CV(변동계수; coefficient of variation)



raster 패키지에 포함된 함수 cv를 사용하여 변동계수 계산



+ 설치된 패키지를 사용하려면 library 함수로 해당 패키지를 불러옴(R-3.3.* 이후 버전만 가능)

```
> library(raster)
> cv(seq(1,100))
[1] 57.4485
```

결측치에 대한 옵션

📖 Na.rm 옵션

Lorem ipsum dolor sit amet, ius an molestie facilisi erroribus, mutat nalerum delectus ei vis. Has ornatus conclusionemque id, an videri molestatis sit. In etqui praesent sit. An vel agan porro comprehensan, ad ludus constituto nea, et ius utroque scaevola assuaverit.

Vis cu nodus nulla feugait, oratio facilisi in usu, eili vitae sea te. Ea fabulas accusamus dissentias sea, facete tacinates definitiones et per. Nihil dicant mediocram pro eu, no mei nostro sensibus platonem. Qui id sunno perpetua neglegentur. Vel ipsum novum copiosae ut. Quo et liber detracto probatus. Nam augue scribentur an. Sea oporteat percipitur incidere et. Qui viris nemore an.





Na.rm 옵션



자료에 결측치(NA)가 있는 경우

- + na.rm 옵션에 T를 설정하여 결측치를 제외한 값으로 계산함(rm=ReMove)
- + 자료에 NA가 있으면 평균(분산 중앙값 등도 마찬가지임) 결과는 NA가 얻어짐

```
> x <- c(seq(1,100), NA)
> mean(x)
[1] NA
```

- + 결측치 NA를 제외하고 평균을 얻을 수 있음

```
> mean(x, na.rm=T)
[1] 50.5
```

가중평균(Weighted mean)

📖 가중평균

📖 가중평균의 계산

Lorem ipsum dolor sit amet, ius an molestie facilisi erroribus, mutat nalerum delectus ei vis. Has ornatus conclusionemque id, an vide molestatis sit. In etqui praesent sit. An vel agan porro comprehensan, ad ludus constituto nea, et ius utroque scaevola assuaverit.

Vis cu nodus nulla feugait, oratio facilisi ex usu, eili vitae sea te. Ea fabulas accusamus dissentias sea, facete tacinales definitiones et per. Nihil dicant mediocrem pro eu, no mei nostro sensibus platonem. Qui id sunno perpetus neglegentur. Vel ipsum novum copiosae ut. Quo et liber detracto probatus. Nam augue scriben- tur an. Sea oporteat percipitur incidereit at. Qui viris nemore an.

가중평균

가중평균의 사용

산술평균
(Arithmetic mean)

각 자료가 모두 같은 $1/n$ 의
중요도를 가지는 경우

가중평균
(Weighted mean)

물가지수 등과 같은 경우
각 항목의 중요도(가중치)가
다른 경우



가중평균의 계산



자료값 x_i 에 대한 가중치가 w_i 인 경우

$$\frac{\sum x_i \cdot w_i}{\sum w_i}$$

+ R-언어에서 가중 평균은 `weighted.mean` 함수를 사용함

`weighted.mean(x, w,...)`

- x : 자료
- w : 가중치





가중평균의 계산



‘성적의 가중평균’ 계산 사례 보기



자료

각 과목별 점수(100점 만점)로 다음과 같은 점수를 얻은 학생 중 어느 학생의 점수가 높은가?

	국어	영어	수학	사회	과학
과목가중치	25%	25%	25%	12.5%	12.5%
x1	90점	80점	70점	90점	95점
x2	85점	85점	85점	80점	90점

① 단순 평균을 계산함

```

> x1 <- c(90, 80, 70, 90, 95); x2 <- c(85, 85, 85, 80, 90)
> mean(x1)
[1] 85
> mean(x2)
[1] 85

```

* 두 학생의 점수가 같음



가중평균의 계산



‘성적의 가중평균’ 계산 사례 보기

② 가중평균을 사용함

```
> w <- c(25, 25, 25, 12.5, 12.5)/100  
> weighted.mean(x1, w)  
[1] 83.125  
> weighted.mean(x2, w)  
[1] 85
```

* 주요 과목 점수가 더 높은 두 번째 학생의 점수가 높음

③ 검산 : 가중평균계산과 공식에 따른 직접계산을 첫 번째 학생의 점수에 대해 해봄

```
> sum(x1*w)/sum(w)  
[1] 83.125
```

* 두 결과가 일치함