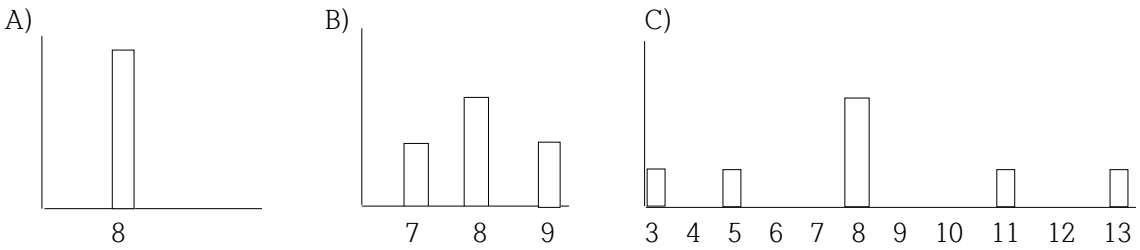
분산도에 관한 보충자료

왜 중심화경향(central tendency)[평균, 중앙값, 최빈치] 이외에 **분산도**를 알아야 하는가?

A)분포 8 8 8 8 8 8 8 의 경우 평균=8, 중앙값=8, 최빈치=8
B)분포 7 7 8 8 8 9 9 의 경우 평균=8, 중앙값=8, 최빈치=8
C)분포 3 5 8 8 8 11 13 의 경우 평균=8, 중앙값=8, 최빈치=8

위의 세 분포도는 아래처럼 분명히 다른데 중심화 경향지수로 보면 다 같아 구분할 수 없다.



<연습문제풀이>
2 2 3 3 4 4 4 5 5 5 5 6 6 6 6 6 7 7 7 7 사례수 N=20
다 더하면 $\sum f.X_i$ = 2x2 + 2x3 + 3x4 + 4x5 + 5x6 + 4x7 =100 (Xi: i번째 X값)
그러므로 평균은 $\overline{X}$ = $\sum f.X_i$/N = 100/20 = 5 (일렬로 나열한 수의 경우 빈도(f) 불필요)

사례수(N)가 짝수 혹은 묶은 도표의 경우 중앙값을 구하는 공식 **Median= TLL+[(N/2-cf)/f(class)] x i**
TLL: 중앙값이 있을 것이라 여겨지는 급간의 실제 하한(True Lower Limit)
cf: 중앙값이 있을 것이라고 여겨지는 급간을 제외한 바로 전까지의 누적빈도(cumulative frequency)
i: 그 급간의 간격(interval)의 크기 f(class): 중앙값이 있을 것으로 추정되는 급간의 빈도
만일 사례수(N)가 **홀수**일 경우는 가운데 위치한 숫자가 중앙값임 (시각 중심이므로), 즉 **(N+1)/2**번째 값
그러나 사례수(N)가 짝수이거나 빈도분포로 묶은 도표일 경우에는 위의 공식을 활용한다.

위의 사례에서 N=20이므로 10번째 혹은 11번째가 중앙값으로 추정되므로 5에 해당됨
5의 실제 하한(TLL)은 4.5, 누적빈도(cf)는 7이며, 또 f(class)에 해당되는 5의 빈도는 네(4)개임. 그러므로 위의 공식에 대입하면 중앙값**(median) = 4.5 + [(20/2 - 7)/4] x 1 = 4.5 + 0.75 = 5.25**
그러나 **컴퓨터 상에는 단순히 5**로 나온다. 위의 경우처럼 계산하지 않고 컴퓨터는 단순하게 처리함.

값	빈도	(Xi값 - 평균)	f 절대값(Xi값 - 평균)	f(Xi값 - 평균)자승
2	2	2-5=-3	2 x 3 = 6	2 x -3제곱 = 18
3	2	3-5=-2	2 x 2 = 4	2 x -2제곱 = 8
4	3	4-5=-1	3 x 1 = 3	3 x -1제곱 = 3
5	4	5-5=0	4 x 0 = 0	4 x 0제곱 = 0
6	5	6-5=1	5 x 1 = 5	5 x 1제곱 = 5
7	4	7-5=2	4 x 2 = 8	4 x 2제곱 = 16
	20		26	합= $\sum$ f (Xi-$\overline{X}$)제곱 = 50 그러므로 분산(V)=50/20= 2.5

평균편차 = 26/20=1.3 표준편차= 분산의 제곱근 = 1.58, 실제분산=50/(N-1)=2.63, 실제표준편차=1.62