2-1장. 통계의 기초

3. 모집단, 표본, 표집

1) 모집단

· 모집단(population)이란 추론하여 규명하고자 하는 집단의 총체를 말한다.

· 모집단은 관념적인 수준의 집단이기 때문에 표집에 앞서, 정확하게 모집단을 규정하는 것이 필요하다.

전수조사란?

전수조사란 모집단을 구성하는 모든 사람을 대상으로 하는 조사이다. 모집단 전체를 대상으로 조사하므로 모집단에 존재하는 다양하고 구체적인 특성을 파악할 수 있다. 예컨대 우리나라 인구를 전수조사하면 '가장 고령인 사람이 누구이고 어디에 사는지' '올해 결혼한 사람의 비율은 어떠한지' 등 자세하고 희귀한 사건에 대한 정보를 알 수 있다. 전수조사는 모집단의 지역별 인구분포나 연령별, 세대별 분포 등 목적에 따른 다양한 분류표를 작성할 수 있고 표본조사의 결과와 비교하여 표본조사의 신뢰성과 정확성을 평가할 수도 있다.

이러한 장점이 있음에도 전수조사는 매우 많은 비용과 시간 및 노력의 토입이 필요하며 전수조사가 반드시 표본조사보다 더 정확한 것은 아니다. 많은 기간이 소요되는 전수조사에서는 시간 흐름에 따른 변화, 즉 초기에 수집된 사료와 후기에 수집된 자료 간에 차이가 발생한다. 이런 경우에는 전수조사보다 표본조사가 더 정확할 수 있다.

2) 표본

- 표본(sample)이란 모집단을 대표하는 특성을 지닌 요소(element)들의 집합을 말한다.
- 표본은 일정한 표본추출방법에 의해 추출된 모집단의 부분집합이라 할 수 있다.

3) 표집

- 표집(sampling)이란 모집단에서 표본을 추출하는 과정을 말한다.
- 표본 추출과정을 표집(sampling)이라 하며, 확률표본(probability sample)과 비확률표본(nonprobability sample)으로 구분된다.

표본조사란?

표본조사(sample survey)는 표본을 대상으로 수행되는 조사를 말한다. 대부분의 조사는 표본조사를 기본으로 하고 있으며 표본조사의 결과는 모집단의 특성으로 일반화된다. 표본조사에서 표본의 통계치를 통해 모집단을 추론하기 때문에 오차(error)가 발생한다. 이와 같은 오차를 최소화하기 위해 조사자들은 정교한 표본추출방법을 사용한다.

표본조사에서의 유의점에는 첫째, 최소한의 비용과 시간을 투자하여 보다 정확하게 모집단의 특성을 밝히는 절약의 원리를 전제하고 있다. 최소의 표본을 통해 모집단을 특성을 정확히 기술할 수 있다면, 가장 효율적인 표본조사가 달성된 것이라 할 것이다. 둘째 표집이론에서는 표본수에 따라 발생할 수 있는 오차의 범위(허용오차)가 정해져 있어 합리적인 결론에 도달하는데 필요한 적정수의 표본을 추출하도록 되어 있다. 셋째, 전수조사에 포함되는 시간 및 비용과 같은 투입요소(input)에 대한 결과의 정확성(output)으로 설명될 수 있다. 즉 모집단을 모두 조사하는데 소요되는 시간과 경비가 표본조사의 결과의 큰 차이를 보이지 않는다면, 훨씬 표본조사가 효율적이다.

4. 모수치와 추정치

1) 모수치

· 모수치(parameter)란 모집단의 특성치(numerical characteristic)로 모집단의 구성원들이 소유하고 있는 변인들의 특성이다.

2) 추정치

(1) 추정치

· 추정치(estim

· 이를 통계치(s

본의 속성을 말한다.

(2) 통계치

· 통계치(statist

· 통계치는 표본

· 표본조사를 통해 모집단의 특성을 밝힌다는 것은 표본 통계치를 통해 모수치를 추정한다는 의미이다.

평균에 대한 표현

조사표본에서 구한 평균소득과 연령분포는 통계치로 모수치를 추정하는데 사용한다.

모수치의 평균은 μ(뮤) 표준편차 σ(시그마)

통계치의 평균은 $\bar{x}$(X바) 표준편차 s(에스)

라고 읽는다.

3장. 중심경향값

[학습목표]

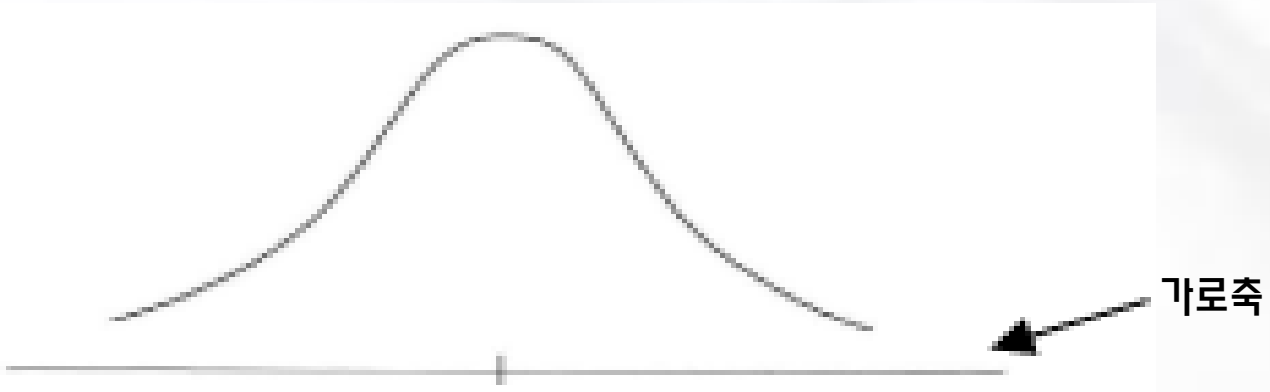
1. 중심경향에 대해 알아본다.
2. 평균을 알아본다.
3. 중앙값을 알아본다.
4. 최빈값을 알아본다.
5. 중심경향값에 따른 분포의 형태를 알아본다.
6. 분산도 개념을 알아본다.
7. 분산도 종류를 알아본다.
8. 분산도에 따른 분포 형태를 알아본다.

1. 중심경향

중심경향(central tendency; 집중경향)은 모집단 혹은 표본으로부터 얻어진 자료를 도표화하면 많은 자료가 어떤 특정한 값으로 몰리는 현상이다.

1) 중심경향값(measure of central tendency; 집중경향값

- 자료를 대표하는 값이다.
- 가로축의 양쪽으로부터 어느 한 점을 중심으로 모이는 경향이다.



2. 평균

- 평균(mean)은 중심경향값에서 가장 흔하게 쓰이는 값이다.
- 전체 사례 수의 값을 더한 다음에 총 사례 수로 나눈 값이다.
- 평균은 산술평균으로 M, X혹은 Y로 표기한다.
- X = 독립변인, Y = 종속변인으로 표기한다.

통계 Tip

※ 1차 함수의 계산 공식 Y = aX+ b

X = 독립변인으로 *Y* = 종속변인으로 표기하는 이유

1차 함수 공식 Y = aX+ b에 의해서 설명할 수 있다.

X 가 증가하면 *Y* 도 증가한다는 정적 1차 함수는 그래프를 통해서도 이해할 수 있다.

평균 계산 공식

$$\overline{X}=\frac{1}{n}(X_1+X_2+X_3+\cdots+X_n)=\frac{\sum X_i}{n}$$

n: 사례수

· 예: 통계 강의를 수강하는 5명의 학생의 연령이 21, 22, 23, 24, 25세라면 평균값은 115/5로 23세이다.

원자료	평균값
21, 22, 23, 24, 25	23
21, 22, 23, 24, 30	24
19, 22, 23, 30, 32	25.2

원자료의 변화에 따라 평균값도 달라진다.

· 평균은 중심경향값 중, 가장 많이 사용한다.

· 평균은 자료의 변화에 따라 민감하게 반응한다.

· 원자료(raw data)를 가지고 있으면 평균의 계산은 간단하다.

· 원자료가 없을 때, 묶음도수표에 의한 정보만으로 평균을 계산하기도 한다.
(원자료가 없이 묶음도수표의 정보에 의존해서 계산하는 것은 바람직하지는 않다.)

3. 중앙값

- 중앙값(median)은 가장 작은 수부터 가장 큰 수까지 크기에 의하여 배열하였을 때 중앙에 위치하는 사례의 값이다.
- $\tilde{M}$, $M_{e,}$ 혹은 M_d로 표기한다.
- 총 사례가 홀수이면 크기의 순서에 의하여 나열한 후 중앙에 위치한 사례의 수치가 중앙값이다.
- 사례가 짝수일 경우에는 크기의 순서에 의하여 나열한 후 가운데 값인 두 값을 더한 후 나누기 2를 하여 중앙값을 찾는다.

중앙값 계산 공식

$$\tilde{M} = \frac{X_{n/2} + X_{(n/2+1)}}{2}$$

n: 총 사례 수

$X_{(n/2)}$: 총 사례 수의 1/2에 해당되는 사례의 점수

$X_{(n/2+1)}$: 총 사례 수의 1/2보다 1이 많은 사례의 점수

- 예: 사례가 홀수인 경우
- 21, 22, 23, 24, 25세, 이들 가운데 세 번째 학생의 연령인 23세가 이들의 중앙값이 된다.
- 예: 사례가 짝수인 경우
- 21, 22, 23, 24, 25, 26세, 이들 가운데 중앙값은 23.5세이다. ➔ (23 + 24)/2 = 23.5

원자료	중앙값
21, 22, 23, 24, 25	23
21, 22, 23, 24, 30	
19, 22, 23, 30, 32	

원자료가 변화해도 중앙값은 변하지 않는다.

- 백분위수의 개념과 비교하면 제50백분위 점수이다.

통계 Tip

백분위수와 사분위수의 이해

백분위수(percentile, percerntile rank) ➔ 얻어진 자료를 크기의 순서로 늘어놓아 100등분하는 값

사분위수(quartile) ➔ 수집된 자료를 크기순으로 배열하여 4등분한 값

제1사분위수 – 제25백분위수 제2사분위수 – 제50백분위수

제3사분위수 – 제75백분위수 제4사분위수 – 제100백분위수

4. 최빈값

- 최빈값(mode)은 분포에서 가장 많은 도수를 갖는 점수이다.
- M_o 로 표기한다.
- 가장 많은 도수를 나타내는 점수가 최빈값이므로 가장 많은 도수 그 자체가 최빈값은 아니다.
- 어떤 분포에는 최빈값이 존재하지 않을 때가 종종 있다.
- 총 사례 수가 적거나 모든 사례가 각기 다양한 값을 가질 때 일어난다.
- 최빈값이 없음, 즉 존재하지 않는다고 해서 최빈값이 0이라는 말과 다르다.
- 원점수의 변화가 최다 도수의 변화에 영향을 주지 않는 한 최빈값은 변하지 않는다.

- 예: 15명의 학생에게 통계 중간고사 10문항을 풀게 하였다. 맞은 문항 수는 다음과 같다.

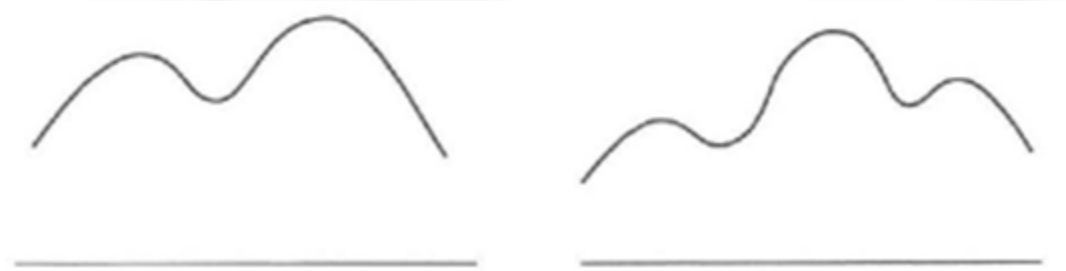
  8, 7, 9, 4, 8, 10, 9, 9, 3, 5, 4, 9, 10, 3, 6

  3, 3, 4, 4, 5, 6, 7, 8, 8, 9, 9, 9, 9, 10, 10 (순서대로 배열)

  ➔ 최빈값은 가장 많은 4명의 점수인 9이다.

원자료	최빈값
1, 2, 3, 4, 5, 6	없음
1, 2, 3, 3, 4, 5	3
1, 2, 3, 3, 4, 4	3, 4

- 최빈값이 여러 개 존재할 수도 있다.
- 최빈값이 여러 개 존재한다면 분포는 이봉분포 혹은 다봉분포의 그래프로 표현될 것이다.



(a) 이봉분포 (b) 다봉분포

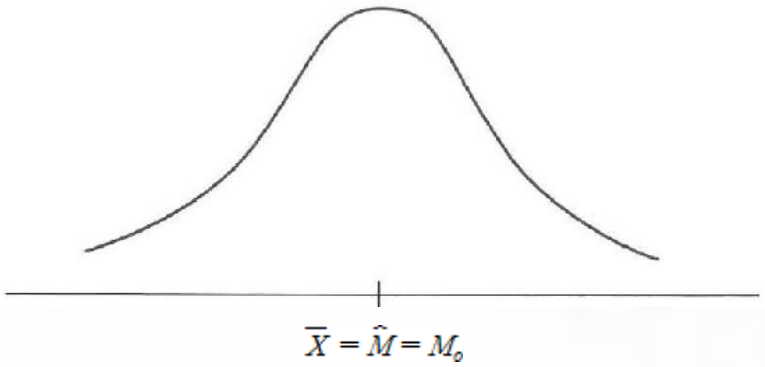
- 최빈값은 분포의 중심경향값을 계산하는데 자주 쓰이지 않는다.
- 쉽게 도수가 가장 많은 수의 값을 알기 위하여 사용한다.
- 최빈값은 자료를 쉽게 파악하거나 어떤 특수한 목적을 위하여 사용한다.

Ex) 의류회사는 의류를 제작할 때 여러 사이즈 중 더 많이 판매되는 사이즈를 고려하여 사이즈 별로 제작 개수를 달리하는 것이 의류 생산업자로서 현명한 방법이다.

5. 중심경향값에 따른 분포의 형태: 정적 편포, 부적 편포, 왜도

1) 정규분포

- 정규분포(normal distribution)는 평균, 중앙값 그리고 최빈값이 일치하며 좌우대칭 형태의 분포이다.
- 특징으로는 세 종류의 중심경향값이 같고 좌우대칭이며 점근선적이다.



- 평균 - 모든 사례 값의 합을 총 사례 수로 나눈 값
- 중앙값 - 가장 작은 수부터 가장 큰 수까지 배열할 때 중앙에 위치한 값
- 최빈값 - 도수가 가장 많은 수의 값

- 많은 모집단의 분포가 항상 정규분포를 나타내지는 않는다.
- 모집단의 특성상 한 점을 중심으로 좌우대칭이 되지 않는 분포도 적지 않다
- 세 종류의 중심경향값이 한 점에 일치할 수도 있고, 각기 다른 위치에 놓일 수도 있다.

2) 편포

- 편포(skewed distribution)는 한 쪽으로 치우친 분포이다.
- 부적 편포(negatively) – 평균, 중앙값, 최빈값 순서
- 정적 편포(positively) – 최빈값, 중앙값, 평균 순서

부적편포 $\overline{X}, \hat{M}, M_o$

정적편포 $M_o, \hat{M}, \overline{X}$

3) 왜도

- 왜도(skewness)는 분포의 대칭 정도를 나타내는 값이다.
- 분포가 기울어진 방향과 정도를 나타낸다.
- 왜도가 0이면 좌우대칭인 정규분포이다.
- 음수이면 부적 편포이고 양수이면 정적 편포이다.

6. 분산도 개념

1) 분산도

- 분산도(variation)는 흩어진 정도를 말한다.
- 중심경향값이 같더라도 흩어진 정도가 다를 수 있다.
- 결국, 중심경향값이 같더라도 집단의 성질이 다름을 의미한다.
- 분산의 범위는 넓은 범위의 분산과 좁은 범위의 분산이 있을 수 있다.

Ex) 평균이 똑같다고 집단의 경향이 같다고 할 수 있는가?

원자료	평균값
21, 22, 23, 24, 25	23
20, 22, 23, 24, 26	23
22, 22, 23, 24, 24	23

➔ 그 답은 아니다.

- 다양성과 그 원인을 분석하는 것이 중요하다.
- 분산의 정도를 파악하기 위하여 범위, 사분위편차, 분산, 표준편차를 계산한다.

7. 분산도 종류

1) 범위

· 범위(range)는 분포의 흩어진 정도를 가장 간단히 알아보는 방법이다.

① 최고값과 최저값을 가지고 파악할 수 있다.

② 연속성을 위한 교정을 고려한다.

③ 최고값 상한계에서 최저값 하한계를 뺀 값이다(계산방법 1).

$$R = (H + u/2) - (L - u/2)$$

$$= (H - L) + u$$

H : 최고값
L : 최저값
u : 측정단위

Ex) 통계 수업을 듣고 있는 학생 10명의 통계 문제 10문항에 대한 결과가 아래에 제시하였다.

3, 4, 5, 5, 6, 6, 7, 8, 9 ,10

계산방법 1.

- 최고점 10점 -> 정확한계 9.5 – 10.5
- 최저점 3점 -> 정확한계 2.5 - 3.5
- 점수의 범위 -> 10.5 – 2.5 = 8

⇒ 점수 범위는 8점

계산방법 2.

- (최고점 10점 – 최저점 3점) + 1 = 8

⇒ 점수 범위는 8점

계산방법 1.과 계산방법 2.의 과정은 다르지만 같은 범위의 값을 제시한다.

- 범위를 구할 때 경우에 따라서 연속성을 위한 교정을 않고 최고값에서 최저값을 빼기도 한다 (계산방법 2).
- 범위를 가지는 두 분포 중 어떤 분포가 보다 넓게 흩어져 있는가에 대한 비교가 가능하다.
- 장점 – 분산도를 간단히 파악 가능하다.
- 단점 – 분산도를 측정하기 위한 정밀성이 없다.
- 최고값과 최저값에 의해서만 범위가 결정된다.
- 최고값과 최저값 사이에 존재하는 값들의 변화에 영향을 받지 않고 그 사이에 존재하는 많은 값을 고려하지 않는다.

집단 A = {1, 2, 3, 5, 8}

집단 B = {1, 2, 3, 4, 5, 6, 7, 8}

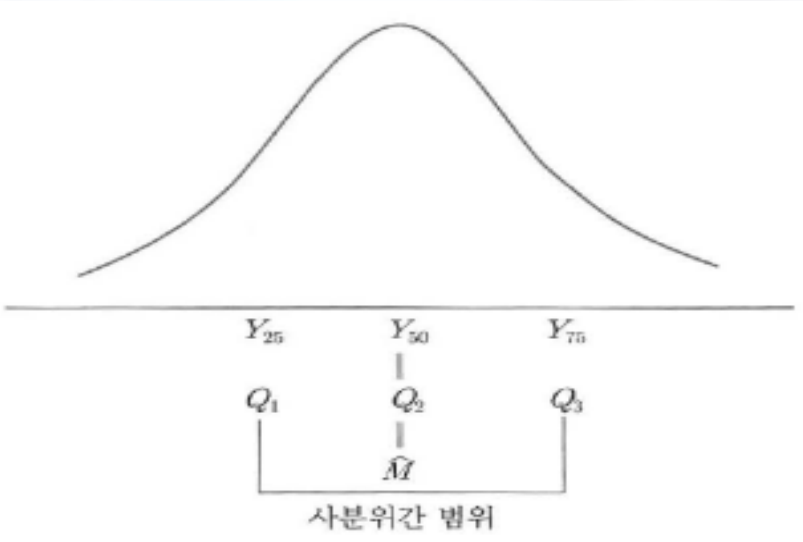
범위는 8로 같지만 사례수가 다르기 때문에 범위만으로는 비교하기 힘들다.

2) 사분위 편차

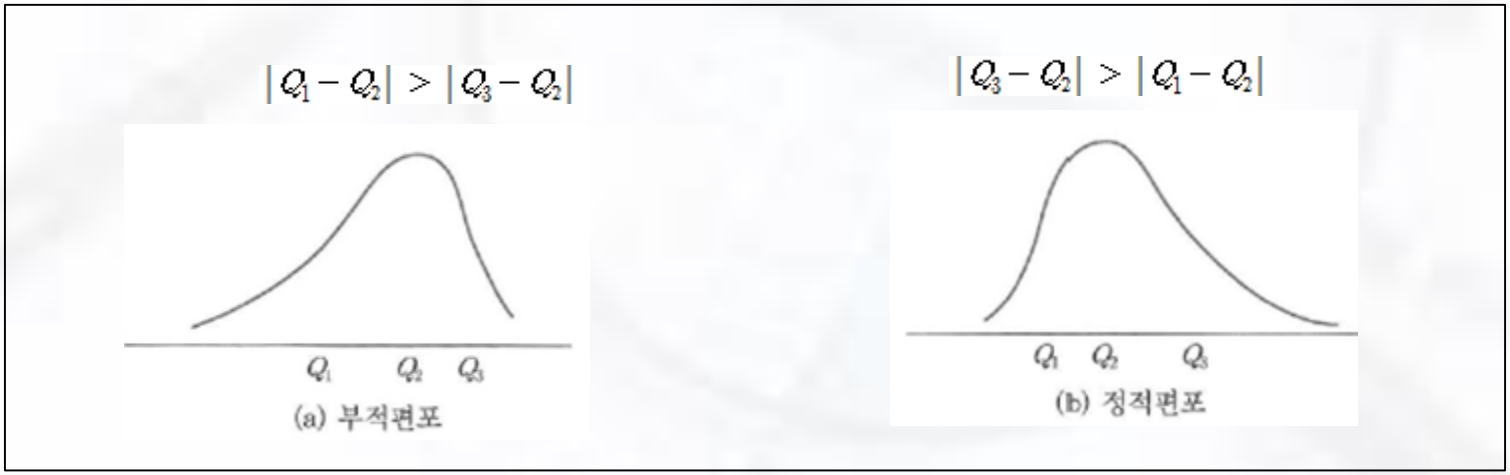
- 사분위 편차(quartile deviation)는 자료의 중앙값을 중심으로 한 분산도를 말한다.
- 사분위 편차에서는 중앙값을 Q_2라고도 한다.
- 이는 분산도의 정도를 분포의 중앙에 위치한 중앙값의 좌우로부터 동일한 백분율을 가진 두 점 간의 거리에 의하여 분석한 결과이다.
- 사분위간 범위가 길면 보다 흩어진 분포이다.
- 사분위간 범위가 짧으면 밀집된 분포이다.

$$Q = \frac{Q_3 - Q_1}{2}$$

사분위 편차



- 사분위 편차와 사분위수에 따라 분포의 형태가 정규분포인지 부적편포 혹은 정적편포인지를 파악 가능하다.
- 정규분포일 때 제2사분위수는 중앙값을 중심으로 제1사분위수와 제3분위수의 거리가 같다.
- 수식으로 좌우대칭 분포를 이루기 위하여 $|Q_1 - Q_2| = |Q_3 - Q_2|$ 성립 되어야 한다



정규분포가 아닌 편포일 경우

- 사분위 편차는 범위와 달리 범위 내에 있는 많은 값을 고려하기에 분산도를 측정하기에 범위보다 정밀하다.
- 하지만 전체 사례수를 고려하지 않아 많이 사용되지는 않는다.

3) 분산

- 모든 자료를 각각 고려하여 분포의 흩어진 정도를 나타내는 것이 분산과 표준편차이다.

(1) 편차

- 편차(deviation)는 각 점수가 평균에서 떨어진 정도를 말한다.
- 편차의 절대값이 크면 그 값은 평균에서 멀리 떨어져 있음을 말한다.
- 편차가 0이면 그 값은 평균과 같다.

$$d_i = Y_i - \overline{Y} \qquad \overline{d_i} = \frac{\sum d_i}{n} = \frac{\sum (Y_i - \overline{Y})}{n}$$

편차를 구하는 공식

- 표준편차는 편차들을 모두 합하여 총 사례 수로 나눈 것이다.
- 그러므로 계산상 모든 편차의 합은 항상 0이 된다.
- 그러나 편차가 0이 되어서는 안 된다.
- 왜냐하면 흩어진 정도가 다르기 때문이다.
- 이로 인해서 편차는 계산상의 한계가 있다.

Ex) 10점 만점의 능력검사에서 5명 학생의 점수가 각기 2, 3, 5, 8, 10이고, 평균 점수는 5점일 때,

	Y_i	$d_i = (Y_i - \overline{Y})$
A	2	-3
B	3	-2
C	5	0
D	5	0
E	10	5

평균 = 5, 편차 = 0 가 된다.

(2) 분산

- 분산(variance)은 편차를 자승하여 그 합을 총 사례 수로 나눈 값을 말한다.
- 각 값으로부터 평균을 뺀 편차를 제곱한 후, 그 수를 모두 더하여 총 사례 수로 나눈 값을 분산이라 한다.
- s_Y^2 혹은 σ_Y^2 로 표기한다.
- 분산을 변량이라고도 말한다.

$$s_Y^2 = \frac{\sum (Y_i - \overline{Y})^2}{n}$$

분산을 계산하는 공식

- 편차부분을 제곱하기 때문에 양수 값인 편차와 음수 값인 편차가 상쇄되지 않으므로 편차의 합은 0이 되지 않는다.

Ex) 10점 만점의 능력검사에서 5명 학생의 점수가 각기 2, 3, 5, 8, 10이고, 평균 점수는 5점일 때

	Y_i	$(Y_i - \overline{Y})$	$(Y_i - \overline{Y})^2$
A	2	-3	9
B	3	-2	4
C	5	0	0
D	5	0	0
E	10	5	25

$$\overline{Y} = \frac{\sum Y_i}{n} = \frac{25}{5} = 5 \qquad \sum (Y_i - \overline{Y})^2 = 38 \qquad s_Y^2 = \frac{(\sum Y_i - \overline{Y})^2}{n} = \frac{38}{5} = 7.6$$

평균 = 5, 분산 = 7.6 가 된다.

(3) 표준편차

- 표준편차(standard deviation)는 분산에 제곱근을 취한 값을 말한다.
- 그래야 편차의 합이 0이 되는 문제를 해결할 수 있다.
- 모수치에서는 σ , 통계치에서는 s_Y 로 표기한다.
- 이 때 분산의 제곱근이므로 $\sqrt{7.6}$ 가 된다.
- 표준편차가 크다면 이는 분포가 넓게 흩어져 있음을 말한다.
- 이때 분산이 7.6이고 표준편차는 $\sqrt{7.6}$ 이므로 2.76가 되어야 한다.
- 하지만 SPSS 프로그램에서는 3.08221의 값이 제시되었다.

기술통계량

	N	최소값	최대값	평균	표준편차	분산
능력점수	5	2.00	10.00	5.0000	3.08221	9.500

$$\bar{Y} = \frac{\sum Y_i}{n} = \frac{25}{5} = 5 \qquad \sum (Y_i - \bar{Y})^2 = 38 \qquad s_Y^2 = \frac{(\sum Y_i - \bar{Y})^2}{n-1} = \frac{38}{4} = 9.5$$

평균 = 5, 분산 = 9.5 가 된다. 표준편차는 $\sqrt{9.5}$ 즉, 3.08221가 된다.

통계 Tip

자유도(degree of freedom)

자유도를 계산하는 공식

df = (n-1)

모집단에서 어떤 수의 표본을 꺼내어 이것을 어떤 조작으로 정리할 때 그 조작 때문에 제약 받는 조건을 없애고 임의로 모집단에서 선출될 수 있다고 생각되는 수를 말한다. df라고 표기한다. 통계적 절차에 의거하여 표준편차를 계산할 때, 표본의 사례 수에서 하나를 빼고 즉, (n-1)로 나누어 계산한다. 쉽게 말해 한 표본을 제외해도 그 집단의 성질은 변하지 않는다는 것을 전제로 한 값이다.