100문제를 푸는 것보다
100문제를 푸는 1가지 원리를 가르칩니다!

원리탐구 중등수학

중3수학 (하)

원리이해 및 내신대비 문제 **선행편**

Mathmatics The discovery of dharma

최경호 지음

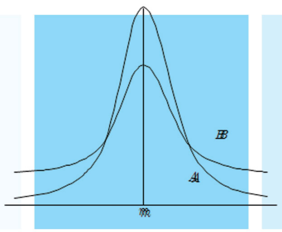
홈페이지(www.m1239.co.kr) 동영상 강의 및 풀이

원리탐구

중3(하) 선행 편 차례

제 1강

대푯값과 산포도



통계학의 이론적·역사적 배경

3000년 전 경의 고대 바빌로니아(Babylonia), 중국(China), 이집트(Egypt) 사람들 유적을 살펴보면, 그 시대에 실시된 인구조사(Cencus)에서 찾아볼 수 있다. 영어로 나라를 State라고 하는 데 상태라는 의미도 갖고 있다. 그래서 통계학을 나라의 상태를 조사하는 것이라는 의미에서 Statistics이라고 한다. 사회적인 구조가 복잡해지고 다양해짐에 따라 국가 전반에 걸친 통계적 조사를 하게 되었고 통치권자들은 국가 전반에 걸친 통계적 지식이 필요하였다.

통계학이 학문으로 확립되기 시작한 것은 17세기부터이다. 영국에서는 그랜트(Graunt, J. ; 1620~1674), 페티(Petty, W. ; 1623~1687) 등에 의하여 「정치학술회의(political arithmetic)」가 형성되었고, 독일에서는 콘링(Conring, H. ; 1606~1681)과 아헨발(Achenwall, G. ; 1719~1772) 등이 「국상학파(Stadenkunde)」를 형성하였다. 정치산술학파는 수치에 기초를 두고 이로부터 어떤 법칙을 찾아내려 하였고, 국상학파는 수치의 사용을 피하고 관념적으로 현상을 파악하려고 하였다.

통계학은 18세기에 이론적으로 발전한다. 당시에 파스칼(Pascal, B. ; 1623~1662), 페르마(Fermat, P. ; 1601~1665), 드 무아브르(De Moivre, A. ; 1667~1754), 라플라스(Laplace, P, S. ; 1749~1827) 등에 의하여 확률론이 수립되고 발전하였으며 통계학의 발전도 가능하게 되었다. 당시 국상학파에서 사용한 이론에 정치산술학파가 사용한 자료를 사용하여 확률론을 기초로 한 통계학이 싹트기 시작하였다.

19세기에 와서 통계학이 여러 분야에서 관심을 모으고 중요시되었다. 19세기 후반에 피어슨(Pearson, K. ; 1857~1936)은 표본이론을 도입하여 모집단의 특성을 추정하는 추측통계학을 확립시켰다.

근대 통계학의 주류인 추측통계학은 피셔(Fisher, R, A. ; 1890~1962)에 의하여 더욱 발전하였다. 그는 새로운 통계학의 기초개념, 모수추정론, 가설검정론 및 실험계획법 등 많은 창의적인 업적으로 추측통계학의 기초를 마련함과 동시에 집단유전학의 발전에도 많은 공헌을 하였다. 1926년에 발행된 그의 저서 「Statistical Methods for Research Workers」는 근대 통계학의 원전으로 불리우고 있다.

통계학은 2차 세계대전 후 급속히 발전, 보급되어 피어슨의 아들인 E. S. 피어슨의 대량생산에서의 품질관리, 폰 노이만(Von Neumann, J. ; 1903~1957)의 가설 검정론, 시장조사, 여론조사 등 표본조사가 일반적으로 사용되었다. 또한 OR (Operation Research), LP(Linear Programming : 선형계획법), 게임이론 분야 등의 연구, 컴퓨터의 발달, 사무의 자동화 등과 결부하여 통계적 방법은 사회 각 분야에서 매우 중요한 역할을 하고 있다.

1. 대표 값과 평균

(1) 대푯값: 자료 전체의 중심적 경향이나 특징을 하나의 수로 나타낸 것으로 평균, 중앙값, 최빈값 등이 있으나 주로 평균을 사용한다.

① (평균)$=\dfrac{\text{(변량)의 총합}}{\text{(변량)의 개수}}$

② 중앙값: 자료를 작은 값에서부터 크기순으로 나열했을 때, 중앙에 위치한 값

자료의 개수 n일 때, 중앙값은 다음과 같다.

· 자료의 개수 n이 홀수이면, $\dfrac{n+1}{2}$번째 자료의 값

· 자료의 개수 n이 짝수이면, $\dfrac{n}{2}$번째와 $\dfrac{n}{2}+1$번째 자료의 값의 평균

③ 최빈값: 자료의 값 중에서 가장 많이 나타난 값, 즉 도수가 가장 많은 자료의 변량 또는 계급값

· 자료의 값 중에서 도수가 가장 큰 값이 한 개 이상 있으면 그 값이 모두 최빈값이다.

· 각 자료의 도수가 모두 같으면 최빈값은 없다.

(2) 도수분포표에서의 평균

$(\text{평균})=\dfrac{\{(\text{계급값})\times(\text{도수})\}\text{의 총합}}{(\text{도수})\text{의 총합}}$

※ 가평균을 이용한 평균 구하기

미리 정한 대략의 평균을 가평균이라 한다.

$(\text{평균})=(\text{가평균})+\dfrac{[\{(\text{변량})-(\text{가평균})\}\times(\text{도수})]\text{의 총합}}{(\text{도수})\text{의 총합}}$

2. 산포도와 표준편차

(1) 산포도

① 산포도: 자료의 분포 상태, 즉 변량들이 흩어져 있는 정도를 하나의 수로 나타낸 값으로 주로 분산과 표준편차를 사용한다.

② 편차: 어떤 자료의 각 변량에서 평균을 뺀 값, 즉 (편차)=(변량)-(평균)

· 편차의 합은 항상 0이다.

(2) 분산과 표준편차

① 분산: 편차들의 제곱의 평균, $(\text{분산})=\dfrac{(\text{편차})^2\text{의 총합}}{\text{변량의 개수}}$

② 표준편차: 분산의 양의 제곱근, $(\text{표준편차})=\sqrt{(\text{분산})}$

자료의 분산과 표준편차가 크면 클수록 그 자료의 분포상태는 평균을 중심으로 흩어져 있는 정도가 더 심하다.

n 개의 변량 $x_1, x_2, \cdots, x_n$ 의 평균을 m 이라고 하면,

$$\text{분산: } S^2=\frac{(x_1-m)^2+(x_2-m)^2+\cdots+(x_n-m)^2}{n}$$

$$=\frac{(x_1^2+x_2^2+\cdots+x_n^2)-2m(x_1+x_2+\cdots+x_n)+m^2n}{n}$$

$$=\frac{(x_1^2+x_2^2+\cdots+x_n^2)-2m\times mn+m^2n}{n}=\frac{x_1^2+x_2^2+\cdots+x_n^2}{n}-m^2$$

즉, (분산)=(변량의 제곱의 평균)-(평균)^2

$$\text{표준편차: } S=\sqrt{\frac{(x_1-m)^2+(x_2-m)^2+\cdots+(x_n-m)^2}{n}}$$

(참고) 분산을 편차의 제곱으로 하는 이유: 편차의 총합이 0이 되므로 0이 되는 것을 방지하여 자료의 분석 자료로 쓰기 위함

(3) 도수분포표에서의 분산과 표준편차

① $(\text{분산})=\dfrac{\{(\text{편차})^2\times(\text{도수})\}\text{의 총합}}{(\text{도수})\text{의 총합}}$ ② $(\text{표준편차})=\sqrt{(\text{분산})}$

다음 도수분포표에서 평균을 m 이라고 하면, 분산S^2 과 표준편차S

계급값	x_1	x_2	$\cdots$	x_n	합
도수	f_1	f_2	$\cdots$	f_n	N

$$\text{분 산: } S^2=\frac{(x_1-m)^2f_1+(x_2-m)^2f_2+\cdots+(x_n-m)^2f_n}{N}$$

$$\text{표준편차: } S=\sqrt{\frac{(x_1-m)^2f_1+(x_2-m)^2f_2+\cdots\cdots+(x_n-m)^2f_n}{N}}$$

1 선행예제

원리탐구

다음 중 옳은 것을 모두 고르면? (정답 3개)
① 최빈값은 없을 수도 있다.
② 대푯값에서 평균, 중앙값, 최빈값 등이 있다.
③ 중앙값은 여러 개일 수도 있다.
④ 평균은 극단적인 값에 영향을 받지 않는다.
⑤ 대푯값은 자료들이 흩어져 있는 정도를 하나의 수로 나타낸 값이다.
⑥ 평균은 자료를 작은 값부터 크기순으로 나열했을 때 중앙에 위치한 값과 일치한다.
⑦ 평균은 변량의 총합을 변량의 개수로 나눈 값이다.

풀이 답: ①, ②, ⑦
① 각 자료의 도수가 모두 같으면 최빈값은 없다. 옳다.
② 옳다. ③ 중앙값은 1개만 존재한다.
④ 극단적인 값에 영향을 받는 대푯값은 평균이다.
⑤ 자료의 전체의 특징을 대표하는 값은 대푯값이라 한다.
⑥ 자료를 작은 값부터 크기순으로 나열했을 때 중앙에 위치한 값을 중앙값이라 하며, 평균값과 일치하는 것은 아니다.
⑦ 옳다.

유제 1

다음 중 옳지 않은 것을 모두 고르면? (정답 2개)
① 평균은 극단적인 값에 영향을 받는다.
② 대푯값에서 평균, 중앙값, 표준편차 등이 있다.
③ 중앙값은 1개다.
④ 최빈값은 여러 개일 수도 있다.
⑤ 자료의 전체의 특징을 대표하는 값은 대푯값이라 한다.
⑥ 최빈값은 자료를 작은 값부터 크기순으로 나열했을 때 중앙에 위치한 값이다.
⑦ 평균은 변량의 총합을 변량의 개수로 나눈 값이다.

2 선행예제

원리탐구

다음 자료의 중앙값, 최빈값, 평균을 각각 a, b, c라 할 때, a, b, c의 크기를 바르게 비교한 것은?

6,	7,	4,	5,	6,	4,	1,	6,	3,	9

① $a<b<c$ ② $c<b<a$ ③ $a<c<b$ ④ $b<c<a$ ⑤ $c<a<b$

풀이 답: ⑤
순서대로 나열하면 1, 3, 4, 4, 5, 6, 6, 6, 7, 9이므로 중앙값, 최빈값, 평균은 각각 다음과 같다.
· 중앙값은 5와 6의 평균인 $a=5.5$이다.
· 최빈값은 가장 많이 나온 $b=6$이다.
평균 $c=\dfrac{1+3+4+4+5+6+6+6+7+9}{10}=5.1$이다.
따라서 $c<a<b$이다.

유제 2

다음은 수돌이네 반 학생 30명의 사회 수행평가 점수를 조사하여 나타낸 도수분포표이다. 수행평가 점수의 평균, 중앙값, 최빈값을 각각 구하여라.

점수(점)	10	15	20	25	30	합계
도수(명)	3	4	12	6	5	30

3 선행예제

A, B 두 분단 학생의 수학성적의 평균이 표와 같은 때, 두 분단 학생 20 명에 대한 수학 성적의 평균을 구하여라.

구분	학생 수(명)	평균(점)
A 분단	11	90
B 분단	9	70

풀이 답: 81점

$(평균)=\dfrac{(변량)의\ 총합}{(변량)의\ 개수}$ 이므로 A, B분단의 수학 성적의 총합을 A, B 두 분단 학생 20명으로 나눈 값이다.

A분단의 수학 성적의 총합은 $90\times11=990$(점), B분단의 수학 성적의 총합은 $70\times9=630$(점)

따라서 두 분단 학생 20명에 대한 수학 성적의 $(평균)=\dfrac{990+630}{11+9}=\dfrac{1620}{20}=81$(점)이다.

유제 3

다음 표는 A, B, C 세 동아리의 회원 수와 나이의 평균이다. 이 때, 세 동아리 전체 회원의 나이의 평균을 구하여라.

동아리	회원 수(명)	평균(세)
A	8	20
B	12	15
C	10	17

4 선행예제

다음 표는 A, B, C, D, E 다섯 명의 수학 성적의 편차를 나타낸 것이다. 평균이 76 점일 때, 학생 E 의 성적을 구하여라.

학 생	A	B	C	D	E
편 차	-3	-4	3	6	x

풀이 답: 74점

편차의 합은 0이므로 $-3-4+3+6+x=0$, $x=-2$이다.

따라서 학생 E의 점수를 y라 하면 $y-76=-2$, $y=76-2=74$(점)이다.

유제 4

다음 표는 A, B, C, D, E 5명의 중간고사 수학성적에 대한 편차를 나타낸 것이다. |보기|의 설명 중 옳은 것을 모두 골라라.

학생	A	B	C	D	E
편차(점)	1	-2	-3	0	4

|보기|

① 점수가 가장 높은 학생은 E이다.
② C와 E의 점수의 차는 1점이다.
③ A의 점수는 평균과 같다.
④ 분산은 6 점이다.

5 선행예제

원리탐구

다음 표는 수순이의 5 회에 걸친 양궁 성적(10 점 만점)을 나타낸 것이다. 평균과 표준편차를 구하여라.

회수	1 회	2 회	3 회	4 회	5 회
점수	5	8	7	6	9

풀이 답: 평균: 7(점), 표준편차: $\sqrt{2}$

$(\text{평균}) = \dfrac{5+8+7+6+9}{5} = \mathbf{7}(\text{점})$

$(\text{분산}) = \dfrac{1}{5}\{(5-7)^2+(8-7)^2+(7-7)^2+(6-7)^2+(9-7)^2\} = \dfrac{10}{5} = 2$

$\therefore$ $(\text{표준편차}) = \sqrt{\text{분산}} = \sqrt{\mathbf{2}}$

유제 5

아래 표는 수순이의 5회에 걸친 수학점수를 나타낸 것이다. 이 때, 평균과 표준편차를 구하여라.

79,	83,	75,	91,	87

6 선행예제

원리탐구

아래 표는 수순이 학급 50명의 학생이 영국에서 활약하고 있는 기성용 선수에 대한 인기투표 결과의 성적을 나타낸 것이다. 다음을 구하여라.

계급(점)	50 이상 ~60 미만	60 이상 ~70 미만	70 이상 ~80 미만	80 이상 ~90 미만	90 이상 ~100 미만	합 계
도수(명)	4	8	12	16	10	50

(1) 평균 (2) 분산 (3) 표준편차

풀이 답: (1) 79(점) (2) 144 (3) 12

각 계급의 계급값을 차례로 나타내면 55, 65, 75, 85, 95 이다.

(1) 가평균을 이용하여 평균을 구하면 다음과 같다.

$(\text{평균}) = 75+\dfrac{1}{50}[(-20\times 4)+(-10\times 8)+(0\times 12)+(10\times 16)+(20\times 10)] = 75+\dfrac{200}{50} = \mathbf{79}\ (\text{점})$

(2) $(\text{분산}) = \dfrac{1}{50}\{(55-79)^2\times 4+(65-79)^2\times 8+(75-79)^2\times 12+(85-79)^2\times 16+(95-79)^2\times 10\}$

$= \dfrac{7200}{50} = \mathbf{144}$

(3) $(\text{표준편차}) = \sqrt{(\text{분산})} = \sqrt{144} = \mathbf{12}$

유제 6

오른쪽 도수분포표에서 다음을 구하여라.

(1) 평균

(2) 분산

(3) 표준편차

계급값	도수
10	1
20	2
30	3
40	4
합 계	10

7 선행예제

원리탐구

변량 3, x, y, 2의 평균이 1, 분산이 5일 때, x^2+y^2의 값을 구하여라.

풀이 답: 11

다음 도수분포표에서 평균을 m이라고 하면, 분산S^2을 구하는 방법

변량	x_1	x_2	$\cdots$	x_n

분 산: $S^2=\dfrac{(x_1-m)^2+(x_2-m)^2+\cdots+(x_n-m)^2}{n}$

$=\dfrac{(x_1^2+x_2^2+\cdots+x_n^2)-2m(x_1+x_2+\cdots+x_n)+m^2n}{n}$

$=\dfrac{(x_1^2+x_2^2+\cdots+x_n^2)-2m\times mn+m^2n}{n}=\dfrac{x_1^2+x_2^2+\cdots+x_n^2}{n}-m^2$

즉, (분산)=(변량의 제곱의 평균)-(평균)^{2}이다.

(해1) $\dfrac{3^2+x^2+y^2+2^2}{4}-1^2=5$, $\dfrac{x^2+y^2+13}{4}=6$ $\quad\therefore x^2+y^2=11$

(해2) $\dfrac{3+x+y+2}{4}=1$, $x+y=-1$이고 $\dfrac{(3-1)^2+(x-1)^2+(y-1)^2+(2-1)^2}{4}=5$이므로

$x^2-2(x+y)+y^2+7=20$, $x^2+y^2=11$이다.

유제 7

변량 a, b, c, d의 평균이 8, 표준편차가 5일 때, a^2, b^2, c^2, d^2의 평균을 구하여라.

종합문제

1. 4회에 걸친 수학 시험 성적이 92점, 80점, 86점, x점이고, 최빈값이 86점이다. 5회까지의 평균은 4회까지 평균보다 2점이 올랐다. 5회의 성적은?

① 96점 ② 95점 ③ 94점 ④ 93점 ⑤ 92점

2. 다음 중 옳은 것은?

① 최빈값은 없을 수도 있다.
② 중앙값은 여러 개일 수도 있다.
③ 중앙값은 극단적인 값에 의해 영향을 받는다.
④ 대푯값에는 평균, 중앙값, 도수분포표 등이 있다.
⑤ 자료의 전체의 특징을 대표하는 값을 변량이라고 한다.

3. 다음 자료의 최빈값을 구하여라.

(1) 13, 12, 15, 12, 13, 7, 11, 8, 12

(2) 10, 11, 17, 20, 29, 14, 12, 21

(3) 8, 6, 17, 9, 10, 71 10, 15, 6, 5, 12

4. 4개의 변량이 있다. 이 변량에 대하여 중앙값은 48이고 세 개의 변량이 42, 51, 52일 때, 나머지 변량을 구하면?

① 45 ② 46 ③ 47 ④ 48 ⑤ 49

5. 다음은 수돌이네 반 학생 10명의 등교하는 데 걸리는 시간을 조사하여 나타낸 것이다. 등교하는 데 걸리는 시간의 평균이 21분일 때, 최빈값을 구하여라.(단위: 분)

10	19	18	22	27	25	x	20	35	15

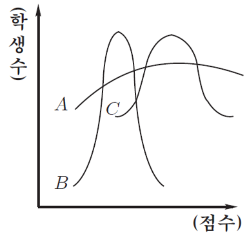
6. 다음 표는 다섯 학생 A, B, C, D, E의 수학 성적에 대한 편차를 나타낸 것인데, E 학생의 편차가 찢어져 보이지 않는다. 이 다섯 학생의 분산을 α라 할 때, 10α의 값은 얼마인가?

학생	A	C	B	D	E
편차	-2	-3	-1	2	

7. 다음은 학돌이네 반 학생 20명의 몸무게를 조사하여 나타낸 도수분포표이다. 몸무게의 평균과 최빈값을 각각 구하여라.

몸무게(kg)	35이상 ~ 45미만	45 ~ 55	55 ~ 65	65 ~ 75	75 ~ 85	합계
도수(명)	2	3	6	8	1	20

8. 그림은 A, B, C 세 학급의 성적을 나타낸 것이다. 다음 설명 중 가장 옳은 것은?



① B 반의 성적이 가장 높다.
② C 반은 B 반보다 평균이 높다.
③ B 반이 A 반보다 평균이 높다.
④ B 반의 표준편차가 가장 크다.
⑤ A 반은 중간층의 학생이 가장 많으므로 표준편차가 가장 적다.

9. 다음은 수순이네 반 학생 20명의 과학 성적을 조사하여 줄기와 잎 그림으로 나타낸 것이다. 과학 성적의 중앙값을 구하여라.

줄기	잎					
5	2	9				
6	0	3	5	7	8	
7	1	3	4	6	7	8
8	2	4	7	6		
9	0	5	8			

10. A, B, C 세 학생의 키의 평균은 $158m$ 이다. A 와 B, B 와 C 의 키의 평균이 각각 $153cm$, $161cm$ 일 때, B 의 키를 구하여라.

11. 다음 중 옳지 않은 것은?

① 편차의 합은 항상 0이다.
② 어떤 자료의 각 변량에서 평균을 뺀 값을 편차라고 한다.
③ 표준편차가 작을수록 변량의 값들이 고르다.
④ 분산은 편차의 제곱의 평균이고, 표준편차는 분산의 양의 제곱근이다.
⑤ 산포도에는 분산, 표준편차, 평균 등이 있다.

12. 다음은 10개의 자료를 작은 값부터 크기순으로 나열한 것이다. 중앙값이 20이고 최빈값이 23일 때, $a+b$의 값은?

5	9	12	13	a
22	b	23	27	30

① 36 ② 38 ③ 40 ④ 41 ⑤ 44

13. 변량 $a, b, c, 9, 2, 4, 3, 2, 9$ 의 중앙값이 7이고, 최빈값이 9일 때, a, b, c의 값을 구하면?(단, a, b, c는 $a<b<c$인 자연수이다.)

① $a=5, b=7, c=8$ ② $a=7, b=8, c=10$ ③ $a=7, b=8, c=9$
④ $a=5, b=6, c=8$ ⑤ $a=4, b=5, c=7$

14. 5개의 변량 16, 14, 11, 15, x의 평균과 최빈값이 같을 때, 상수 x의 값은?

① 11 ② 13 ③ 14 ④ 15 ⑤ 16

15. 다음은 10명의 수학 성적을 나타낸 도수분포표이다. 이 때, 분산을 구하여라.

계급(점)	40 이상 ~ 50 미만	50 이상 ~ 60 미만	60 이상 ~ 70 미만	70 이상 ~ 80 미만	80 이상 ~ 90 미만
도수(명)	1	1	2	4	2

16. 아래 표는 3학년 5반 학생 30명의 턱걸이 횟수에 대한 도수분포표이다. 다음 중 가장 잘못된 것은?

턱걸이 횟수	학생수
3	2
4	5
5	8
6	A
7	6
8	2
50	1
합계	30

① 평균은 7이다.
② 중앙값은 5.5이다.
③ 최빈값은 8이다.
④ 평균은 자료의 중심경향을 잘 나타내었다고 할 수 없다.
⑤ 편차의 합은 0이다.

17. 다음 표는 수돌이네 반 학생 A, B, C, D, E 5명의 수학 성적의 평균이 88점일 때의 편차를 나타낸 것이다. 다음을 구하여라.

학생	A	B	C	D	E
편차(점)	-2	-5	x	4	1

(1) 학생 B의 성적　　(2) 수학 성적의 표준편차

18. 표는 수순이의 5회에 걸친 시험 결과에 따른 편차를 나타낸 것이다. 분산을 구하면?

	-3	x	-1	4
2				

① 6.5　② 6.8　③ 6.9　④ 7　⑤ 7.2

19. 다음 자료들 중에서 표준편차가 가장 큰 것은?

① 3, 3, 3, 3, 3, 3　② 1, 9, 3, 7, 1, 9　③ 3, 5, 5, 5, 9, 9
④ 3, 7, 3, 7, 3, 7　⑤ 1, 5, 1, 5, 1, 5

20. 세 수 a, b, c의 평균이 5이고 분산이 3일 때, a^2, b^2, c^2의 평균을 구하여라.

① 15　② 28　③ 30　④ 33　⑤ 42

21. 아래 표는 A, B 두 학급의 수학 성적의 통계 자료이다. 표에 대한 설명으로 옳은 것은?

학급	평균	표준편차
A반 성적	67	5.7
B반 성적	67	8.9

① 두 학급의 학생수와 분산은 같다.
② B반이 A반보다 성적이 더 우수하다.
③ 수학 성적의 차이가 많은 학급은 A반이다.
④ 두 학급의 성적의 고른 정도를 비교할 수 없다.
⑤ 두 학급 중에서 수학 성적이 고른 학급은 A반이다.

생활과 수학

1. 당선자 알아맞히기

선거 때만 되면 신문 및 방송사들마다 어떤 후보가 당선할지 여론조사 결과를 내보낸다. 놀라운 것은 누가 누구를 얼마나 앞설지 예상한 것이 실제 결과와 크게 다르지 않다는 점이다. 여론조사는 통계와 확률을 이용한 예측이다. 통계와 확률을 이해하려면 우선 모집단, 무작위 추출 등 몇 가지 용어를 알아야 한다.

조리실에서 우리학교 학생 모두를 먹일 수 있는 국물을 솥에 끓이고 있을 때, 국물 전체가 모집단이다. 국물 맛을 보려고 한 국자 떠냈다면 그것이 표본 집단이 된다. 국을 뜨기 전에 국자로 국을 휘저었을 텐데, 무작위 추출에 필요한 표본의 동질성 유지를 위한 과정이라고 할 수 있다.

이제 선거 이야기로 돌아가 보자. 우리나라 전체 국민 가운데 선거권자는 3천만 명이 넘는다. 그렇다면 누가 당선될지 알기 위해 얼마나 많은 사람들을 표본으로 추출해야 할까?

여론조사기관에서는 지난 대선 때 1천5백 명(0.005%)만 뽑아 투표결과를 예측했다. 그러면 3천5백만 명 중 1천5백 명을 어떻게 선정할까. 전국의 지역별로 선거권자 수를 알아보고 해당 지역마다 1천5백 명에 비례하는 숫자만큼 선정한다. 표본을 추출할 때 누구나 뽑힐 가능성이 동일하도록 여론조사기관에서는 특별한 방법을 동원한다. 가령 전국의 집이 모두 일렬로 늘어서 있다고 가정하고 일정 간격을 정해 뽑는 식이다. 이렇게 추출한 1천5백 명이 어떤 후보에 투표할지 조사했다고 가정하자. 실제 선거 결과도 이와 일치할까?

표본 집단을 정확하게 선택하고 선정한 사람들이 솔직하게 답했다면 결과는 정확할 수 밖에 없다. 하지만 그 결과가 예측과 늘 일치하지는 않는다. 이런 가능성을 오차라고 한다. 그래서 예측 결과를 발표할 때 오차범위를 “K후보 지지도 40.3%%, M후보 지지도 39.9%로 표본오차 95% 신뢰수준에서 ±2.0%”라는 형태로 제시한다.

즉, 표본오차가 “±2.0%” 라는 말은 K후보 지지도가 38.3%에서 42.3%사이에 있을 수 있다는 의미로 결과가 예측한 값보다 2% 작을 수도 있고, 2% 클 수도 있다는 말이다. 또한, “95% 신뢰수준”이란 뜻은 같은 일을 100번 반복했을 경우 95번(95%)은 오차 범위(±2.0%)안에 있다는 뜻이다.

오차를 고려하면 위의 결과는 당선자가 바뀔 수도 있고, 애초의 예측보다 더 큰 차이가 날 가능성도 있다. 그런데 이렇게 부정확한 예측을 왜 할까?
그 이유는 많은 비용과 시간을 절약하기 위해서다. 이렇게 우리생활에 많은 도움을 주는 것이 통계이다.

읽을거리

1. 대표값과 대표적인 대표값의 사용경우

통계 자료를 정리하고 도수분포표를 작성함으로써 통계 계열의 변수의 크기에 대하여 하나의 질서가 있음을 이해하게 된다. 이와 같은 일은 통계 자료 중에서 무엇인가 특별한 성격을 파악하기 위하여 취해진 방법으로 아무리 큰 자료라도 어떤 대표적인 측정값으로 압축하는 경우, 변수의 크기를 대표하는 하나의 값을 얻을 수가 있게 된다.

실험, 측정, 여론 조사 등을 통하여 수집된 자료를 정리하여 그 특성을 파악하고자 할 때, 자료의 수가 아무리 방대하더라도 그 집단을 대표할 수 있는 간략한 수치를 생각할 수 있다. 이와 같이 한 집단의 성격을 나타내는 하나의 수치를 대표값이라고 하고, 대표값에는 산술평균, 중앙값, 최빈수 등이 있으며 주어진 집단의 어떤 성질을 알고 싶은 가에 따라 구하는 대표값의 종류가 결정된다.

이 때, 자료 각각의 수치를 전부 사용하여 계산으로 구한 대표값, 즉 산술평균, 기하평균, 조화평균을 계산적 평균이라 한다.

한편, 다수의 자료 중에서 자료의 위치로부터 구하여진 대표값은 위치적 평균이라 하며 중앙값, 최빈수가 여기에 속한다.

(가) 산술평균(Arithmetic Mean) : n 개의 자료 $x_1, x_2, x_3, \cdots, x_n$ 의 총합을 자료의 개수 n으로 나눈 값을 **산술평균**이라고 하며, 일상적으로 가장 많이 사용되는 평균값.

$$\overline{X} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

(나) 기하평균(Geometric Mean) : 자료전체의 곱의 양의 n제곱근을 기하평균이라 하고 주로 인구증가율이나 경제성장률의 평균값 등을 구할 때 많이 사용하며, 또 산술평균이 극단적인 수치에 의하여 좌우되는 폐단을 막기 위해 많이 사용한다.

$$G = \sqrt[n]{x_1 \times x_2 \times \cdots \times x_n}$$

(다) 조화평균(Harmonic Mean) : 각 자료의 역수의 산술평균값의 역수를 조화평균이라 하고, 일정량의 상품을 생산하는데 요하는 시간의 평균을 구하는 경우나 일정 금액을 가지고 구입할 수 있는 상품 수량의 평균 등을 구할 때 사용한다.

$$H = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}\right)} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

(라) 중앙값(Median) : n개의 자료를 크기순으로 배열하였을 때, 그 중앙에 위치한 자료의 값을 중앙값이라고 한다.

자료의 총개수 n이 홀수일 때는 $\frac{n+1}{2}$ 번째의 자료값, n이 짝수일 때는 $\frac{n}{2}$ 번째

읽을거리

와 $\frac{n}{2}+1$ 번째의 자료값의 산술평균을 취한다. 누적도수(累積度數) 그래프에서 세로좌표가 $\frac{n}{2}$ 이 되는 점의 가로좌표에 해당한다.

(마) **최빈값(Mode)** : 도수분포에서 최대의 도수를 가지는 변량의 값이다.

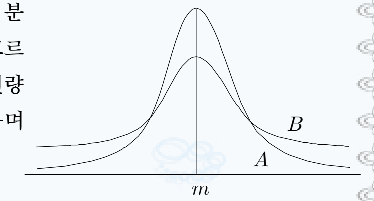
변량의 분포형태가 대칭이면 평균, 중앙값, 최빈값은 일치하나, 비대칭이면 보통 다음과 같은 경험식(經驗式)이 성립한다.(Pearson의 정리)

$$(\text{평균})-(\text{최빈값}) \fallingdotseq 3\times\{(\text{평균})-(\text{중앙값})\}$$

2. 산포도와 표준편차

측정하고자 하는 자료들이 흩어져 있는 정도를 하나의 수로 나타낸 값을 산포도라 한다. 산포도에는 분산, 표준편차, 평균편차, 사분편차, 범위 등이 있으나, 이 중에서 가장 많이 쓰이는 것이 표준편차이다.

표준편차가 작다는 것은 변량들이 평균에 가까이 분포되어 있고, 자료들이 안정성이 강하고 분포가 고르다는 것을 말한다. 또한, 표준편차가 크다는 것은 변량들이 평균에서 멀리 분포되어 있고 안정성이 약하며 분포가 고르지 못하다고 한다.



위 그림에 A, B 두 분포를 비교하여 보면 평균값은 m으로 같다. 그러나 A의 자료들이 B의 자료들 보다 평균을 중심으로 집중되어 있으므로 표준편차는 $s_a < s_b$ 임을 알 수 있다.

분산은 편차의 제곱의 평균이고 표준편차는 분산의 양의 제곱근을 정의한다.

> 변량 $x_1, x_2, \cdots, x_n$ 의 평균을 m, 표준편차를 s 라 하면
>
> (1) 분산 : $s^2 = \dfrac{(x_1-m)^2+(x_2-m)^2+\cdots+(x_n-m)^2}{n} = \dfrac{x_1^2+x_2^2+\cdots+x_n^2}{n} - m^2$
>
> (2) 표준편차 : $\sqrt{\text{분산}} = s$