#3. 통계

PART 04 이산확률분포와 이항분포

#1. 중등통계에서 배우다.

#2. 이산확률분포와 평균, 분산, 표준편차

#3. 이항분포

촌철살인 #4 질적 개념과 양적 개념

춥다. 덥다. 정당하다. 이런 것이 질적 개념이다. 온도가 23도다 21도다 이런 것이 양적개념이다.
질적 개념은 누구나 느끼는 동물적 감각이다. 개인마다 차이도 매우 크고 주변 환경에 따라 스스로도 일관성이 없다.
(추운 날 꽁꽁 언 손을 미지근한 물에만 넣어도 뜨겁다고 느끼는 것이 그렇다.)
그렇기 때문에 상황을 질적인 개념으로만 이해할 경우 상황 대응력도 떨어지고 바른 대책도 세울 수 없다.

● 양적개념으로 세상을 바라보는 A씨는 친구들과 음식점을 갔다. 그때 A씨의 친구는 이렇게 말했다.

친구 : 이거(음식점) 하나 차리면 편할 것 같아. 일이야 알바생 고용해서 시키면 되고 손님도 많으니까 이 정도면 정말 많이 벌 수 있을 것 같아. (많이? – 이런 게 질적 개념이다.)

A씨 : 친구야, 너는 절대 차리면 안 되겠어. 사업을 한다면 임대료, 가맹비, 재료비, 알바생 시급, 초기 투자비, 광고비, 돈을 빌린다면 이자비용까지 모두 수학적인 계산을 해야 되고 예상 매출과 그에 따른 영업이익과 순이익을 모두 계산해야 돼. 그리고 그 사업의 지속성도 생각해야 해. 3억 투자하고 1000만원씩 2년 동안 순수익을 보고 망하거나 급격하게 수익이 악화된다면 이자 등을 따지지 않아도 6000만원 손해거든. 게다가 거기에 쏟아 부은 시간과 노동의 가치를 비용으로 환산해서 계산해 보면 정말 큰 손해일 수 있지.

수학은 세상에 존재하는 수많은 질적인 개념을 양적인 개념으로 이해하고 계산하는 논리도구이다. 확률과 통계도 마찬가지이다.

– 확률은 특정한 사건이 일어날 가능성을 수치화시킨 것이다.
– 통계는 모든 사건이 일어날 가능성을 따져 평균과 분산을 계산함으로서 그 집단이 가진 특성을 수치화시킨 것이다.
(따지자면 확률은 나무를 보는 것이고 통계에서 평균과 분산은 숲을 보는 것이라고 할 수 있다.)

이렇게 세상에서 일어나는 많은 현상들을 양적개념으로 이해하고 이것을 활용하려고 노력하여 바른 판단능력을 갖추기를 바란다.

PART 01 수학적 확률

#1. 시행과 사건 상황을 설명하는 용어의 정립
1. 단순시행과 복합시행
2. 독립적 선후시행과 종속적 선후시행
3. 사건의 종류, 관계, 선후사건
4. 사건의 개수

#2. 수학적 확률 결국엔 경우의 수
1. 수학적 확률의 계산
2. 같은 것을 포함한 확률

#3. 통계적 확률과 기하학적 확률
통계적 확률은 〈문제의 조건〉이고, 기하학적 확률을 중요도가 떨어진다.

Part 1. 개념의 외연
#1. 정수론 2, 3탄
#2. 원주각과 중심각
#3. 해석기하

상황을 분석하기 위해 〈상황을 설명하는 용어〉를 반드시 정립해야 한다.

PART 02 확률의 계산 1

#1. 비율의 면적화와 영역분할
각 시행에서 비율의 면적화가 가지는 의미를 생각한다.

#2. 확률의 덧셈정리
1. 덧셈정리
2. 여사건의 아이디어

#3. 조건부 확률
1. 바뀐 표본공간
2. 의미상 조건부와 계산적 조건부

#4. 독립과 종속
1. 시행의 독립과 종속
2. 사건의 독립과 종속
3. 독립조건

Part 2. 개념의 외연
#1. 가비의 리
#2. 명제의 판단
#3. 배타적 고리문제

〈비율의 면적화〉를 통해서 애매할 수 있는 〈확률의 연산〉을 정확히 설명하다.

PART 03 확률의 계산 2

#1. 비율의 마법과 확률의 곱셈정리
1. 확률의 곱셈정리
2. 독립사건과 종속사건에서 A∩B
3. 분할과 사건의 고정
4. 비율을 느껴라
5. 확률조건에서 두 사건의 독립과 종속

#2. 곱셈정리와 조건부 확률
1. 조건부 사건의 분할
2. 베이즈 정리

#3. 독립시행의 확률
1. 독립시행의 확률의 상황
2. 반복시행의 확률

Part 3. 개념의 외연
#1. 확률의 기본연산
#2. 이항정리와 독립시행의 확률
#3. 수열과 확률
#4. 경우의 수의 맹점
#5. 경우의 수로 풀기

PART 04 이산확률분포와 이항분포

#1. 중등통계에서 배우다.
1. 도수분포와 평균의 이해
2. 편차와 분산의 이해

#2. 이산확률분포
1. 이산확률분포
2. 평균 (=기댓값)
3. 분산과 표준편차
4. 성질 (확률변수의 변형)

#3. 이항분포
1. 이항분포
2. $B(n,\ p)$
3. 평균, 분산, 표준편차
4. 증명

Part 4. 개념의 외연
#1. 표 작성 문제는 확률문제
#2. 이항분포의 확신
#3. 이항정리의 연산
#4. 큰 수의 법칙

PART 05 연속확률분포와 정규분포

#1. 연속확률분포
기본정의를 정확히 알고 넘어가면 된다.

#2. 정규분포와 표준정규분포
1. 정규분포와 확률밀도함수
2. 정규분포의 확률 표현
3. 정규분포의 표준화와 표준정규분포
4. 서로 다른 확률분포로 표현된 확률 관계식

#3. 이항분포의 정규분포화
1. 이항분포의 정규분포화
2. 이항분포의 확률표현

Part 5. 개념의 외연
#1. 연속확률분포와 적분

PART 06 통계적 추정

#1. 표본평균의 이해
1. 표본평균($\overline{X}$)의 이해
2. 표본평균($\overline{X}$)의 평균, 분산, 표준편차
3. 표본평균($\overline{X}$)의 분포

#2. 모평균의 추정
모평균의 추정 식 도출의 과정을 이해한다.

#3. 표본비율의 이해
1. 표본비율($\hat{p}$)의 이해
2. 표본비율($\hat{p}$)의 평균, 분산, 표준편차
3. 표본비율($\hat{p}$)의 분포

#4. 모비율의 추정
모비율의 추정 식 도출의 과정을 이해한다.

Part 6. 개념의 외연
#1. 표본비율의 확률

뼈대가 되는 기본 개념

#1 중등통계에서 배우다.

- 통계란 확률적인 모든 자료를 모아서 경향성을 분석하는 것으로 나무를 보기보다는 숲을 보고 큰 흐름을 알아낸 것.

- 놀랍게도 모든 선생님이 쉽다고 하는 통계를 어렵다고 하는 학생들도 있다. 저자는 그 이유를 두 가지라고 생각한다.

  첫 번째 : 느낌 없이 외운 공식은 휘발성이 강하기 때문이다.

  사실 교육과정은 중학교에서 <평균, 분산, 표준편차>의 계산법과 용어가 가진 느낌을 구체적 예시를 통해서 충분히 이해한 상태로 고등학교 과정을 학습하도록 되어 있다. 고등학교 과정은 중학교 과정에서 용어를 살짝 바꾸고 일반화된 표현(이것이 살짝 복잡하다.)을 사용할 뿐이기 때문이다. 이 과정을 생략하고 바로 <정의>라는 딱딱한 용어와 의미를 모르고 외우는 <복잡해 보이는 공식>으로 시작하면 어렵게 느낄 수 있다.

  두 번째 : 통계를 가장한 확률 문제이기 때문이다. 사실 확률과 경우의 수가 어려운 것이지 통계가 어려운 것이 아니다.

1. 도수분포와 평균의 이해

1) 도수분포

- 도수분포는 우리가 공부할 확률분포를 직관적으로 이해하기에 앞서 조금 더 실생활적으로 다루는 분포로서 중학교 때 배우는 과정이다. 어떤 변량(변하는 양)과 그에 따른 도수(사람 수, 과목 수)의 분포를 말한다. 이런 말들은 보통 정확한 정의가 있어서 이것을 암기하기보다는 간단한 예를 통해서 어떤 뜻인지를 이해하면 된다.

- 예를 들어 A, B, C, D, E 총 5과목의 시험을 봐서 50점이 두 과목, 60점이 두 과목, 70점이 한 과목이라면 이것을 다음과 같이 도수분포표로 나타낼 수 있다. 이때 점수가 변량이고, 과목 수가 도수가 된다.

점수	50	60	70	총합
과목 수	2	2	1	5

사실 실생활에서 위와 같이 자료가 적은 경우는 도수분포를 이용하여 나타내지 않는다. 위의 표는 그저 도수분포의 의미와 용어를 알려줄 뿐이다.

- 아래 표는 학생수가 500명인 어떤 학교의 키의 분포를 도수분포표를 이용하여 나타낸 것이다.

키	160 ~ 165	165 ~ 170	170 ~ 175	175 ~ 180	180이상 ~ 185미만	총합
사람 수	30	90	230	120	30	500

도수 분포는 보통 이렇게 범위(계급)와 그 범위에 해당하는 도수를 통해서 나타낸다. 도수분포를 이와 같이 나타내는 이유는 통계라는 것이 어차피 개개인의 키를 보고자 하는 것이 아니라 학교 전체 학생들의 키의 분포를 대략적으로 알고 싶은 것이기 때문이다. 계급은 도수분포를 관찰하고자 하는 사람이 그 목적에 맞게 조정하면 된다. 더 정확한 자료가 필요하다면 변량의 구간을 더 작게 잡아 조사하면 된다.

이 경우 구간의 중앙값을 계급값이라 한다. 평균, 분산, 표준편차를 주어진 자료에서 계산하는 방법은 학생들의 키를 계급값이라 치고 계산하는 것이다. (예를 들어 165 ~ 170에서는 키가 167.5인 학생이 90명이라고 치고 평균을 계산한다.) 당연히 실제 평균과 다를 수 있지만 크게 차이가 나지는 않을 것이고 이 집단의 대략적인 특성을 보여줄 수 있다.

(정확한 자료가 필요하다면 일일이 학생들 키를 조사해서 정확한 평균을 내면 그만이다.)

2) 평균의 이해

- 평균이란 모든 변량을 다 더해서 총 자료의 개수로 나누는 것이다. (시험점수의 평균을 내봤다면 이 정도는 알 것이라 믿는다.)

- 다음 두 개의 표를 통해 평균을 계산하는 느낌을 이해해 보자.

X	50	60	70	총합
N	1	2	3	6

⇨ 평균 $= \dfrac{50\times1 + 60\times2 + 70\times3}{6} = 50\times\dfrac{1}{6} + 60\times\dfrac{2}{6} + 70\times\dfrac{3}{6}$

X	50	60	70	총합
N	10	20	30	60

⇨ 평균 $= \dfrac{50\times10 + 60\times20 + 70\times30}{60} = 50\times\dfrac{1}{6} + 60\times\dfrac{2}{6} + 70\times\dfrac{3}{6}$

: 여기에서 우리는 평균을 구함에 있어서 중요한 것은 <도수>가 아니라 <도수의 비율>이라는 것을 알게 된다.
<도수>가 아무리 달라도 <도수의 비율>만 같으면 같은 평균값을 가지게 된다.

- <상대도수분포>란 이런 평균의 성질을 반영하여 <도수> 대신 <도수의 비율>을 나타낸 분포이다.

X	50	60	70	총합
N	1	2	3	6

⇨

X	50	60	70	총합
$\frac{n}{N}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	1

3) 가중 평균, 기댓값 - 같은 계산에 다른 의미

- 숫자에 의미를 부여하여 읽어나가는 것은 수학에서 매우 일반적이고 자연스러운 과정 중 하나이다.
예를 들어 $y=2x+1$이라는 간단한 일차함수에서 <x에 시간, y에 인구>라는 의미를 부여하면 우리가 흔히 알고 있는 <기울기 2>는 <시간에 따른 인구 증가율>이라는 의미가 부여된다.

- 평균도 마찬가지다. 아래 표에서 각 <상대도수>에 어떤 의미를 부여하면 그에 따라 평균도 다른 의미가 부연된다.

X	50	60	70	총합
$\frac{n}{N}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	1

⇨ 상대도수에 가중치라는 의미를 부여해보자. - 이때 평균은 가중 평균(가중치가 반영된 평균)이라는 의미를 가지게 된다.

⇨ 상대도수에 확률이라는 의미를 부여해보자. - 이때 평균은 기댓값(확률적으로 가장 나올 가능성이 높아서 기대되는 값)이라는 의미를 가지게 된다.

즉, <평균, 가중평균, 기댓값>은 구하는 계산의 과정이 모두 같다. 그냥 위의 표에서 주어진 비율에 다른 의미를 부여했기 때문에 느낌의 차이가 존재하는 용어들이 탄생했을 뿐이다.

티칭 대푯값에는 평균, 중앙값, 최빈값이 있다.
최빈값과 중앙값은 고등수학에서 다루지 않으므로 넘어간다. 중학교 때 다루고 나중에 대학 때 다시 다룬다.
즉, 우리가 다루는 <유일한 대푯값>은 <평균>이다. 평균을 대푯값이라고 부르는 이유는 자연스럽게 느낄 수 있다. 반 평균점수가 50점인 반이 있고 70점인 반이 있다면 우리는 70점인 반이 <전반적으로 우수>하다는 것을 느낄 수 있다. (오히려 가장 잘하는 친구는 평균이 50점인 반에 있을 수도 있다.) 이처럼 우리는 평균을 가지고 집단을 평가하고 비교한다. 그러니 평균이 그 집단을 대표하는 값이라고 할 수 있는 것이다.
다시 한 번 말하지만, <평균>은 고등학교에서 다루는 유일한 <대푯값>이다.

2. 편차와 분산의 이해

1) 편차

① 편차의 정의 : 어떤 집단에 각 대상들의 차이(점수, 키, 몸무게.. 등등)가 심한 경우 <편차가 심하다.>라는 말을 자연스럽게 쓴다. 수학에서도 의미가 비슷하지만 정확한 정의를 알아야 한다.

⇨ 편차 = X - m 즉, 편차는 사실 <차이>가 아니라 변수 - 평균이며 당연히 음수가 나올 수 있다.

② 편차를 통해 알 수 있는 것.

: 우리는 <편차>를 통해 평균으로부터 각 변량의 상대적 위치를 알 수 있다. 예를 들어 어떤 학교의 수학점수의 분포를 생각할 때 어떤 학생의 점수의 편차가 5점이라면 이 학생의 정확한 점수를 알 수는 없지만 이 학생이 평균점수보다 5점 높은 점수를 받았다는 사실을 알 수 있다.

추가로 나중에 정규분포까지 배운다면 <편차와 표준편차의 관계>를 통해 이 학생의 수학점수가 전체집단에서 상위 몇 %(상대적 위치)에 해당하는 지까지 알 수 있다.

③ 편차의 합은 0이다.

: 참고로 이런 성질을 이용하여 평균을 내는 방법이 있는데 이런 방법은 <가평균>이라고 한다. 사실 생각해보면 이런 방법은 실생활에서도 많이 쓰는 방법이다.

$\frac{x_1 + x_2 + \cdots + x_n}{n} = m$ (평균의 계산) ⇨ $x_1 + x_2 + \cdots + x_n = nm$ (식의 변형)

⇨ $x_1 + x_2 + \cdots + x_n = \overbrace{m + m + \cdots + m}^{n개}$

⇨ $(x_1 - m) + (x_2 - m) + \cdots + (x_n - m) = 0$

이와 같이 편차의 합은 0이다.

우리가 m을 모르는 경우 m을 구하는 방법으로 <가평균(임시평균)>을 설정하는 방법이 있다. 이때 대략적으로 평균값에 가까울 것이라고 예측되는 가평균을 M이라고 정한 후 <가편차>의 합을 통해서 <평균>을 구할 수 있다.

$(x_1 - M) + (x_2 - M) + \cdots + (x_n - M) = k$ ⇨ $x_1 + x_2 + \cdots + x_n = nM + k$

⇨ $\frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{nM + k}{n} = M + \frac{k}{n}$

⇨ 즉, $m = M + \frac{k}{n}$ 이므로

즉, 우리는 <가편차>를 모두 더한 값 k를 총 도수인 n으로 나눈 값인 $\frac{k}{n}$를 가평균 M에 더함으로서 실제평균 m을 구할 수 있다.

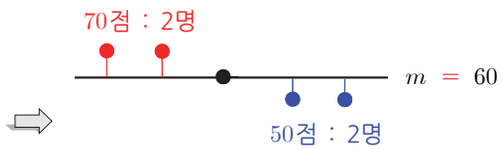
⇨ 예를 들어 이번 시험점수가 47.8, 78.5, 92.1, 82, 88.5, 65.5, 100라고 가정하고 가평균을 이용하여 평균을 구해보자. 가평균을 대략 80점 정도로 잡으면 가편차는 -32.2, -1.5, 12.1, 2, 8.5, -14.5, 20라고 볼 수 있다. 이때 가편차의 합은 -5.6이다. 즉, $\frac{-5.6}{7} = -0.8$이므로 우리가 구하는 실제 평균은 $80 + (-0.8) = 79.2$

2) 분산과 표준편차의 이해

- 다음 두 분포를 비교해 보자

분포 1 :

X	50	60	70	총합
$\frac{n}{N}$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{2}{5}$	1



분포 2 :

X	40	60	80	총합
$\frac{n}{N}$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{2}{5}$	1

위의 두 분포는 같은 평균값을 가지고 있지만 분산도(자료들이 평균을 기준으로 흩어진 정도)는 다르다.

꼭 분산의 개념을 모르더라도 위의 두 분포에서 <분포1>보다 <분포2>가 더 많이 흩어져 있어 분산된 정도가 크다는 것을 알 수 있다. (즉, 분산도가 크다.) 이런 분산도를 나타낼 수 있는 대푯값이 바로 <분산>이다. 분산을 <대푯값>이라고 부를 수 있는 이유는 <분산> 역시 <평균>이기 때문이다. - 물론 우리가 일반적으로 말하는 평균과 계산하는 변량이 다르다.

- 분산도 결국엔 평균이다.

위와 같은 분포에서 <|편차|들의 평균>을 내본다면 두 집단의 분산된 정도를 비교할 수 있을 것 같다.

(편차의 합은 0이므로 편차의 평균을 내봤자 두 집단 모두 0이 나온다. 즉, 편차의 평균은 의미가 없다.)

하지만 실제로는 그렇게 하지 않고 <$(\text{편차})^2$의 평균>을 분산이라고 정의한다. 이렇게 해도 훨씬 많이 흩어져 있는 <분포2>의 분산이 더 큰 값을 가질 것은 당연한 일이다. - 왜 |편차| 대신 $(\text{편차})^2$을 쓰는지는 중요하지 않다.

분포 1 :

X	50	60	70	총합
$\frac{n}{N}$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{2}{5}$	1

⇨

$(X-m)^2$	$(50-60)^2$	$(60-60)^2$	$(70-60)^2$	총합
$\frac{n}{N}$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{2}{5}$	1

<분포1>의 분산 $= 100\times\frac{2}{5} + 0\times\frac{1}{5} + 100\times\frac{2}{5} = 80$

분포 2 :

X	40	60	80	총합
$\frac{n}{N}$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{2}{5}$	1

⇨

$(X-m)^2$	$(40-60)^2$	$(60-60)^2$	$(80-60)^2$	총합
$\frac{n}{N}$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{2}{5}$	1

<분포2>의 분산 $= 400\times\frac{2}{5} + 0\times\frac{1}{5} + 400\times\frac{2}{5} = 320$

분산이란 위의 두 분포처럼 먼저 <편차의 제곱>을 구해서 <각 변량이 상대적으로 평균과 떨어진 정도>를 수치로 나타낸 다음 <이들의 평균>을 내는 것이다.

- $\sqrt{\text{분산}} = \text{표준편차}$

<분산>과 <표준편차>는 같은 의미를 가지지만 분산을 조금 더 사실적인 값으로 만드는 과정이 표준편차이다.

위의 분포가 점수라고 가정해본다면 <분포1>의 분산은 80점이 나오고 <분포2>의 분산은 무려 320점이 나온다. 물론 두 집단의 흩어진 정도를 비교하는 것에는 문제가 없지만 분산으로 나온 숫자가 너무 비현실적이다.

(이 수치를 통해서 <집단1>의 각 변량들이 평균 60점과 대략 어느 정도 떨어져 있다고 할 수 있는지는 감이 잘 오지 않는다.)

그래서 분산에 양의 제곱근($\sqrt{\ }$)을 취함으로서 조금 더 사실적인 값을 얻을 수 있다.

<분포1>의 표준편차는 $\sqrt{80}$의 근삿값은 대략 8.94정도 된다. (이 참에 상용로그도 복습해 보길...) 즉, 각 변량들의 평균과의 차이가 대략 9정도라고 생각할 수 있고, 실제 <|편차|에 대한 평균>은 8이므로 대략 비슷하다. <분포2> 역시 $\sqrt{320}$이 17.9정도이므로 실제 <|편차|에 대한 평균>인 16과 대략 비슷하다.

티칭 실제로 <|편차|에 대한 평균>은 <분산>과 <표준편차>의 느낌을 설명하기 위해 보조적으로 끌어온 지표일 뿐 어떤 교과서나 책에 나오는 문제를 풀기 위해 의미를 가지는 값이 아니다. 즉, 기억에서 지워도 되지만

$$\sqrt{\text{분산}} = \text{표준편차} \neq \text{|편차|에 대한 평균}$$

이라는 사실은 한 번 더 짚고 넘어가도록 하자. (<|편차|에 대한 평균>을 설명하는 평균편차라는 말이 있기는 하지만 필요 없다.)

티칭 만약 $\sqrt{\text{분산}}$ = <|편차|에 대한 평균>이라고 착각하는 사람들은 다음을 확인하라.

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

좌변의 $\sqrt{\ }$를 분자에만 분배를 해야 우변의 식이 나오는데... 그냥 아예 말이 안 되는 것이다.

<사실 간단히 표준편차와 평균편차는 다르다.>라고 말하면 되지만 평균편차는 고등과정에서 다루지 않는다.

티칭 표준편차는 분산의 양의 제곱근인가? 음 아닌 제곱근인가? - 고등학생은 그냥 넘겨도 좋다.

고등에서 중요한 내용은 아니지만 중등에서는 중요한 문제가 되기도 한다. 여기에서 <음 아닌 제곱근>이라고 하는 입장은 분산이 0 일 수도 있어서 <양의 제곱근>이라는 말이 틀렸다고 한다. 이런 착각이 생길 수 있음을 이해한다.

하지만 $\sqrt{\ }$을 <양의 제곱근>이라 하고 $-\sqrt{\ }$을 <음의 제곱근>이라고 읽는다. 즉, 0의 양의 제곱근 $\sqrt{0}$이든 음의 제곱근 $-\sqrt{0}$이든 모두 0이라고 생각할 수 있으므로

$\sqrt{\text{분산}}$을 <분산의 양의 제곱근> 이라고 읽는 것에 문제가 없다.

그래서 구 교육과정에서는 표준편차를 <분산의 양의 제곱근>이라고 했으나 <양>이라는 말의 느낌 때문에 자꾸 오류라고 지적하는 사람이 생겨 신 교육과정에서는 <분산의 음 아닌 제곱근>으로 개정되었으니 참조 바란다.

(별로 중요하지 않다. 결국 수학도 언어로 표현하기 때문에 중·고등 과정에서도 논란의 소지가 있는 것들이 꽤 있다.)

3) 분산을 구하는 공식

- 분산을 구하는 가장 기본적인 공식이면서 정의가 <$(\text{편차})^2$의 평균>이라는 사실을 이미 다룬 바 있다.

- 분산을 구하는 또 다른 공식은 <제곱의 평균 − 평균의 제곱>이다.

1단계 : $\dfrac{x_1+x_2+\cdots+x_n}{n} = m$이라고 해보자.

이때 분산은 <$(\text{편차})^2$의 평균>인 $\dfrac{(x_1-m)^2+(x_2-m)^2+\cdots+(x_n-m)^2}{n}$이다.

2단계 : $\dfrac{(x_1-m)^2+(x_2-m)^2+\cdots+(x_n-m)^2}{n}$ (그냥 전개시키면 아래의 식을 얻을 수 있다.)

$$= \frac{(x_1)^2+(x_2)^2+\cdots+(x_n)^2}{n} + \frac{-2mx_1-2mx_2-\cdots-2mx_n}{n} + \frac{\overbrace{m^2+m^2+\cdots+m^2}^{n\text{개}}}{n}$$

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

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

$$= \frac{(x_1)^2+(x_2)^2+\cdots+(x_n)^2}{n} - m^2$$ 이것을 <제곱의 평균 − 평균의 제곱>이라고 읽는다.

말 그대로 <각 변량을 제곱한 값들의 평균>에서 <그냥 평균값을 제곱한 값>을 뺀다는 뜻이다.

결론 : 결국 분산을 구하는 식은 아래와 같이 2가지가 있고, 이 중에서 보통은 우변의 식이 계산이 더 간단하다.

어차피 <분산>을 구하기 위해서는 반드시 <평균>을 먼저 계산해야 한다.

이때 굳이 각 변량에서 평균을 뺀 후 이것을 또 제곱해서 평균을 내는 것보다는(즉, 편차 제곱의 평균)

그냥 각 변량을 제곱해서 평균을 낸 다음 이미 구한 <평균>을 제곱해서 빼면 되기 때문이다.

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

티칭 분산을 구하는 공식

분산을 구하는 공식을 설명하는 과정에서 우리에게 조금 더 익숙한 느낌의 식을 다루기 위해 <상대도수분포>가 아닌 <도수가 전부 1인 도수분포>를 가정하고 설명했다. 사실은 다음의 식이 조금 더 일반화된 식이다.

X	x_1	x_2	$\cdots$	x_n	총합
$\frac{n}{N}$	p_1	p_2	$\cdots$	p_n	1

⇨

X^2	$(x_1)^2$	$(x_2)^2$	$\cdots$	$(x_n)^2$	총합
$\frac{n}{N}$	p_1	p_2	$\cdots$	p_n	1

<평균> $= x_1 \cdot p_1 + x_2 \cdot p_2 + \cdots + x_n \cdot p_n$

<(편차)^{2}의 평균> $= (x_1-m)^2 \cdot p_1 + (x_2-m)^2 \cdot p_2 + \cdots + (x_n-m)^2 \cdot p_n$

<제곱의 평균 − 평균의 제곱> $= (x_1)^2 \cdot p_1 + (x_2)^2 \cdot p_2 + \cdots + (x_n)^2 \cdot p_n - (x_1 \cdot p_1 + x_2 \cdot p_2 + \cdots + x_n \cdot p_n)^2$

코칭 평균, 분산, 표준편차를 빨리 세는 방법은 없다.

위와 같이 일일이 계산해야 한다. 식으로 익히기보다는 예제를 반복해서 풀면서 자연스럽게 익히는 것이 좋다.

특히 분산을 구하려면 반드시 먼저 평균을 구해야 한다. 평균을 구하지 않고 분산을 구하는 방법은 없다.

코칭 표준편차는 분산을 구할 수 있다면 따로 공부하는 것이 아니다. 그냥 분산에 $\sqrt{\ }$를 씌우면 된다.