01 확률변수 용어 정리

확률변수

표본공간 S의 각 원소를 실수 전체의 집합 R의 한 원소에 대응시키는 함수 X를 **확률변수**라고 합니다.

이산확률변수

확률변수 X가 가지는 값이 유한개이거나 자연수와 같이 셀 수 있을 때[1] 그 확률변수 X를 **이산확률변수**라고 합니다.

1) 이 짧은 서술에 내포된 내용이 세 가지 있습니다.
① 자연수는 유한하지 않다(무한하다).
② 자연수는 셀 수 있다.
③ 셀 수 없는 무한도 있다.
즉 교과서는 '셀 수 있는 무한'과 '셀 수 없는 무한'의 존재를 넌지시 알려주고 있는 것입니다.

확률분포와 확률질량함수

이산확률변수 X가 어떤 값 x를 가질 확률을 $\mathrm{P}(X=x)$라 표기합니다. 이때 X가 가지는 값 $x_i\ (i=1,\ 2,\ 3,\ \cdots,\ n)$와 X가 x_i를 가질 확률 p_i의 대응관계인 다음의 식을 이산확률변수 X의 확률분포라 합니다.

$$\mathrm{P}(X=x_i)=p_i \quad (i=1,\ 2,\ 3,\ \cdots,\ n)$$

또한 이 대응 관계를 나타내는 함수를 확률질량함수라 합니다.

이산확률변수의 평균, 분산, 표준편차

$x_i\ (i=1,\ 2,\ 3,\ \cdots,n)$의 값을 가질 수 있는 이산확률변수 X의 대응 관계를 아래와 같이 표로 나타낼 수 있습니다.

X	x_1	x_2	x_3	$\cdots$	x_n	합
$\mathrm{P}(X=x)$	p_1	p_2	p_3	$\cdots$	p_n	1

이때 확률의 기본 성질과 **평균**, **분산**, **표준편차**의 정의에 의해[2] 다음이 성립합니다.

2) 평균, 분산, 표준편차의 정의와 그에 대한 설명은 다음 챕터에 나옵니다. 일단은 공식만 눈에 발라놓으세요.

① $0\le p_i\le 1 \quad (i=1,\ 2,\ 3,\ \cdots,n)$

② $\displaystyle\sum_{i=1}^{n}p_i=1$

③ $\mathrm{E}(X)=\displaystyle\sum_{i=1}^{n}x_ip_i=m$

④ $\mathrm{V}(X)=\mathrm{E}\left((X-m)^2\right)=\displaystyle\sum_{i=1}^{n}(x_i-m)^2p_i$

⑤ $\sigma(X)=\sqrt{\mathrm{V}(X)}$

한편 $\sum$의 성질을 이용하면 확률변수 X, 0이 아닌 상수 a, 상수 b에 대하여 다음이 성립함을 알 수 있습니다.[3)]

① $\mathrm{V}(X) = \mathrm{E}(X^2) - m^2$

② $\mathrm{E}(aX+b) = a\mathrm{E}(X) + b$

③ $\mathrm{V}(aX+b) = a^2\mathrm{V}(X)$

④ $\sigma(aX+b) = |a|\,\sigma(X)$

3) ①에서
$$\mathrm{E}(X^2) - \{\mathrm{E}(X)\}^2$$
이라 쓰면 혼동하기 쉽고 표기도 깔끔하지 않으므로 뒤의 $\mathrm{E}(X)$를 m으로 대체하였습니다.

이항분포

한 번의 시행에서 어떤 사건 A가 일어날 확률이 p, 일어나지 않을 확률이 $q = 1-p$일 때,[4)] n번의 독립시행에서 사건 A가 일어나는 횟수를 확률변수 X라 할 때, X의 확률질량함수는 독립시행의 확률에 의해 다음과 같습니다.

$$\mathrm{P}(X=r) = {}_n\mathrm{C}_r p^r q^{n-r} \quad (\text{단, } r\text{는 } 0 \le r \le n \text{인 정수이다.})$$

이와 같은 확률분포를 **이항분포**라 하고, $\mathrm{B}(n,\ p)$라 표기하며, 이러한 상황을 확률변수 X가 이항분포 $\mathrm{B}(n,\ p)$를 따른다고 합니다. 우리는 앞으로 이를 간단히 $X \sim \mathrm{B}(n,\ p)$라 나타내기로 합시다.

독립시행의 확률에서 배웠듯이 $\displaystyle\sum_{r=0}^{n}\mathrm{P}(X=r) = \sum_{r=0}^{n}{}_n\mathrm{C}_r p^r q^{n-r} = 1$입니다. 한편 이항분포를 따르는 확률변수 X에 대하여 $\mathrm{E}(X) = np$, $\mathrm{V}(X) = npq$, $\sigma(X) = \sqrt{npq}$가 성립합니다.

4) 앞으로 이항분포를 논할 때 $p+q=1$임은 굳이 언급하지 않아도 기본적으로 전제하도록 합시다.

연속확률변수

확률변수 X가 가지는 값이 어떤 범위에 속한 모든 실수의 값일 때, X를 **연속확률변수**라 합니다. 예를 들어 X가 1과 4 사이의 모든 실수의 값을 가질 수 있을 때, X는 연속확률변수입니다.

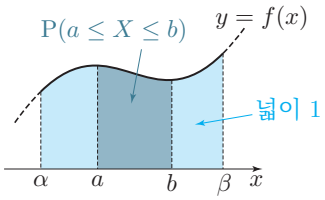
연속확률변수와 이산확률변수의 공통점과 차이점

앞서 언급한 예시에서, X가 1 이상 4 이하의 값을 가질 확률인 $\mathrm{P}(1 \le X \le 4)$의 값은 1입니다. 즉 연속확률변수도 이산확률변수와 동일하게 X가 가질 수 있는 모든 값에 대한 확률을 모두 더한 값은 1이다라는 공통점이 있습니다.

그러나 연속확률변수는 특정 실수값을 가질 확률이 0이라는 점에서 이산확률변수와 구별됩니다. 예를 들어 $\mathrm{P}(X=3) = 0$, $\mathrm{P}\left(X=\sqrt{2}\right) = 0$, $\mathrm{P}(X=\pi) = 0$입니다. 연속확률변수의 확률은 X의 값이 특정 범위에 속할 확률일 때에만 비로소 의미를 갖습니다. 예를 들어 $\mathrm{P}(\sqrt{2} \le X \le \pi)$의 값은 논할 수 있습니다.[5)]

5) 이산확률변수에서 내포되었던 내용을 이용하면, '어떤 범위에 속한 모든 실수'가 셀 수 없는 무한이기 때문이 아닐까 조심스럽게 추측할 수 있습니다.

확률밀도함수와 확률분포



$\alpha \le X \le \beta$의 모든 실수의 값을 가지는 연속확률변수 X에 대하여 어떤 함수 $y = f(x)$의 그래프가 다음 조건을 만족시킬 때, 함수 f를 X의 **확률밀도함수**라 합니다.

① $\alpha \le x \le \beta$에서 $f(x) \ge 0$

② f의 그래프와 x축 및 두 직선 $x = \alpha$, $x = \beta$로 둘러싸인 부분의 넓이는 1이다.

③ x축 및 두 직선 $x = a$, $x = b$로 둘러싸인 부분의 넓이는 확률 $\mathrm{P}(a \le X \le b)$와 같다.

이렇게 확률밀도함수 f를 이용하여 X가 가지는 값의 범위에 속하는 구간에 확률을 대응시키는 것을 연속확률변수 X의 확률분포라 합니다.

교과서가 닫힌구간과 정적분을 쓰지 못하는 이유

①, ②, ③을 읽으며 〈수학 II〉에서 배운 구간표기법이나 정적분을 쓰면 간단하게 쓸 수 있을 법한 내용들을 대체 왜 문장으로 길게 늘어뜨리는지 의아한 학생들이 있을 것입니다. 이는 〈확률과 통계〉 교과서가 〈수학 II〉를 배우지 않은 학생들을 대상으로 서술하느라 생긴 문제입니다. 그러나 우리는 모두 〈수학 II〉를 배우므로, 앞으로 이 책에서는 간결한 서술을 위하여 다음과 같이 닫힌구간과 정적분 표기를 사용하도록 하겠습니다.

① $[\alpha, \beta]$에서 $f(x) \ge 0$

② $\displaystyle\int_{\alpha}^{\beta} f(x)\,dx = 1$

③ $\displaystyle\int_{a}^{b} f(x)\,dx = \mathrm{P}(a \le X \le b)$

연속확률변수가 특정 실수의 값을 가질 확률이 0인 이유는 정적분으로 설명할 수 있다

확률밀도함수가 f인 연속확률변수 X가 가질 수 있는 범위에 포함된 임의의 실수 a에 대하여 $\mathrm{P}(X = a) = \displaystyle\int_{a}^{a} f(x)\,dx$입니다. 이는 정적분의 성질에 의해 0입니다.

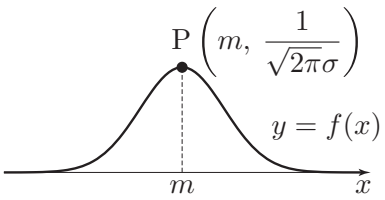
정규분포

실수 전체의 집합에서 정의된 연속확률변수 X의 확률밀도함수 f가 두 상수 m, σ와 무리수인 상수 $e = 2.718281\cdots$에 대하여 다음과 같을 때, X의 확률분포를 **정규분포**라고 합니다.[6)]

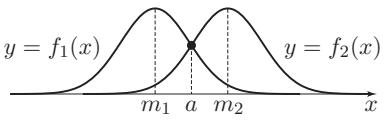
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

이때 확률변수 X의 평균과 표준편차는 각각 m, σ임이 알려져 있습니다. 평균이 m, 표준편차가 σ인 정규분포를 $\mathrm{N}(m, \sigma^2)$라 표기하고, 확률변수 X는 정규분포 $\mathrm{N}(m, \sigma^2)$을 따른다고 합니다. 우리는 앞으로 이를 간단히 $X \sim \mathrm{N}(m, \sigma^2)$이라 나타내기로 합시다.

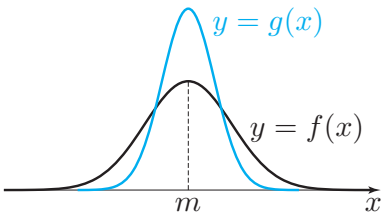
$X \sim \mathrm{N}(m, \sigma^2)$일 때 X의 확률밀도함수 $y = f(x)$의 성질

① 

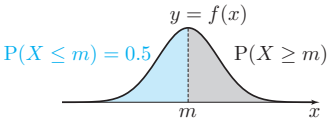
$x = m$에 대하여 대칭인 종 모양의 곡선이며, x축을 점근선으로 하며, $x = m$일 때 최댓값을 갖습니다.[8)]

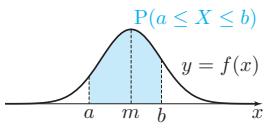
② 

σ의 값은 일정하고 m의 값이 바뀌면, 대칭축의 위치만 바뀌고 곡선의 모양은 같습니다.[10)]

③ 

m의 값은 일정하고 σ의 값이 바뀌면, σ의 값이 커질수록 최댓값이 작아지면서 양옆으로 넓게 퍼지고, σ의 값이 작아질수록 최댓값이 커지면서 양옆으로 좁게 모입니다.

④ 



곡선 $y = f(x)$와 x축 사이의 넓이는 1이고,
$\mathrm{P}(X \le m) = \mathrm{P}(X \ge m) = 0.5$입니다.
또한 $\mathrm{P}(a \le X \le b) = \int_a^b f(x)\,dx$입니다.

6) 함수식을 두려워하지는 마세요. 이를 암기할 필요는 거의 없습니다! 교과서가 알려주고 있어서 차마 외우지 말라고 단정짓지는 못하겠지만, 대부분의 문제에서 함수식 자체를 중요하게 여기지는 않습니다.

8) 이 값이 $\frac{1}{\sqrt{2\pi}\sigma}$이라고 일부 교과서가 언급을 하고 있기는 합니다만, 함수식도 외우지 않는 마당에 최댓값을 외우고 있기에는 좀$\cdots$ 그렇다고 함수식을 모든 교과서가 언급했고 $x = m$에서 최대인 것까지는 언급을 했는데 이걸 아예 안 적기는 또 그렇고$\cdots$ 미묘합니다.

10) 이때 두 곡선 $y = f_1(x)$와 $y = f_2(x)$의 교점의 x좌표를 a라 할 때, a의 값은 몇이고, 두 곡선은 직선 $x = a$에 대하여 어떤 성질을 가질까요?

정규분포곡선의 성질을 조금만 더 수학적으로 서술해보자

f의 수식을 분석해봅시다. 이차함수 $g(x) = -\frac{1}{2\sigma^2}(x-m)^2$와 지수함수 $h(x) = \frac{1}{\sqrt{2\pi}\sigma}e^x$에 대하여 $f = h \circ g$입니다. ①에서 $x = m$에 대하여 대칭임은 $f(x-m) = f(x+m)$이 성립함을 통해 수식으로 보일 수 있고, 점근선은 아래와 같이 함수의 극한으로 설명할 수 있습니다.

$$\lim_{x\to\infty} f(x) = \lim_{x\to\infty} h(g(x)) = \lim_{t\to-\infty} h(t) = 0$$

$$\lim_{x\to-\infty} f(x) = \lim_{x\to-\infty} h(g(x)) = \lim_{t\to-\infty} h(t) = 0$$

①은 $f = h \circ g$에서 h가 증가함수이므로 g가 최대일 때 f가 최대이고, g가 $x = m$에서 최대이므로 $x = m$에서 f가 최대임을 통해 설명할 수 있습니다. ②는 평행이동으로 설명할 수 있습니다. ③은 최댓값 $f(m) = \frac{1}{\sqrt{2\pi}\sigma}$을 σ에 대한 함수 $M(\sigma)$로 볼 수 있으며, $M(\sigma)$가 감소함수이므로 σ가 커질수록 M이 작아지는 것을 통해 설명할 수 있습니다.

①의 뒷문장을 정적분으로 나타낼 수는 없을까?

교육과정에서 배우는 정적분은 위끝과 아래끝에 모두 실수를 적어야 합니다. 그래서 $\mathrm{P}(X \le m) = 0.5$, $\mathrm{P}(X \ge m) = 0.5$를 정적분으로 나타내려 시도해도 각각 다음과 같은 난관에 봉착합니다.

$$\int_{?}^{m} f(x)\,dx = 0.5, \qquad \int_{m}^{?} f(x)\,dx = 0.5$$

그런데 물음표의 자리에 어떤 실수를 넣어도 우리가 원하는 바를 제대로 표현할 수가 없습니다. 그래서 교과서가 정적분을 쓰지 않고 $\mathrm{P}(X \le m) = \mathrm{P}(X \ge m) = 0.5$ 라고만 두루뭉술하게 표현하고 넘어가는 것입니다.

이미 아는 분들도 계시겠지만, 물음표에 들어갈 기호는 각각 $-\infty$와 ∞입니다.

교과외이기는 하지만, 우리가 정적분의 위끝부터 아래끝을 '적분구간'이라 부르는 점과, 구간표기법에서 $(-\infty, a]$, $[b, \infty)$와 같은 표현을 써왔음을 고려할 때, 이 둘을 섞은 것이라 생각하면 자연스럽게 받아들일 수 있을 것입니다.

마찬가지로 $(-\infty, \infty)$라는 표현에 착안하여 곡선 $y = f(x)$와 x축 사이의 넓이는 1임을 정적분으로 나타내면 $\int_{-\infty}^{\infty} f(x)\,dx = 1$입니다.

표준정규분포와 표준화

$Z \sim \mathrm{N}(0,\ 1)$인 확률변수 Z의 확률분포를 **표준정규분포**라 합니다. $X \sim \mathrm{N}\left(m,\ \sigma^2\right)$인 확률변수 X에 대하여 다음의 관계식이 성립합니다.

$$Z = \frac{X - m}{\sigma}$$

이렇게 X를 Z로 변환하는 과정을 표준화라고 부르기로 합시다.

나형 TMI : 표준화의 의미는 '평균으로부터 시그마의 몇 배만큼 떨어졌는가'이다

$Z = \dfrac{X - m}{\sigma}$의 의미를 잘 생각해봅시다. 확률변수 X의 값 x와 확률변수 Z의 값 z에 대하여 다음이 성립합니다.

$$x = m + z\sigma$$

따라서 z가 의미하는 것은 x가 평균 m으로부터 시그마의 z배만큼 떨어졌다는 것입니다. 즉 평균과 표준편차의 값에 관계없이 확률변수가 가질 수 있는 값을 $m + k\sigma$ 꼴로 나타내었을 때의 k값이 정규분포에서의 확률값을 결정함을 알 수 있습니다. 이 내용을 단번에 이해하기 어렵다면, 아예 이 박스의 내용을 읽은 적도 없는 것으로 생각하고, 그냥 평소처럼 표준화하여 문제를 푸세요. 이 내용을 단번에 이해하였다면, 모든 정규분포 문제를 이렇게 풀어보세요. 정규분포 문제를 푸는 것이 퍼즐을 풀 듯 즐거울 것입니다.

가형 TMI : 표준화는 치환적분이다

표준화는 사실 $x = m + z\sigma$로 치환하여 다음의 정적분이 서로 같음을 보이는 것과 같습니다.

$$\int_{x_1}^{x_2} f(x)\,dx = \int_{z_1}^{z_2} f(z)\,dz$$

왜 이러한 결과가 성립하는지는 정규분포를 따르는 확률변수의 확률밀도함수를 이용하여 직접 치환적분해보면 알 수 있습니다. 그러나 확률과 통계에서 연속확률변수의 확률밀도함수를 정적분하는 것을 요구하지 않으므로, 지적 호기심에 의한 것이 아닌 이상 굳이 계산할 필요는 없습니다.

02 확률변수 (1) : 기초부터 이산확률변수까지

통계는 깊은 이해가 필요한 단원이 아닙니다. 수능에서도 통계 개념과 그 유도 과정에 대한 깊은 이해를 요구하지 않습니다. 따라서 우리는 이번 챕터에서 수능 통계를 쉽고 빠르게 다 맞기 위해 문제풀이에 필요한 사고만 컴팩트하게 정리할 것입니다. 수식으로 증명하는 과정이 생략된 것에 대해 크게 의문을 품지 않는 것이 좋습니다.[11)]

11) 논술을 준비하는 것이 아니라면, 수능 대비로는 헛수고입니다.

우리에게 익숙한 평균

우리에게 익숙한 평균 이야기를 먼저 해봅시다. 민렬, 준영, 관호의 국영수사과 성적이 각각 다음과 같다고 해봅시다.

	국어	영어	수학	사회	과학
민렬	80	90	80	70	80
준영	100	60	80	100	60
관호	80	80	80	80	80

세 명의 평균성적을 구하면 각각 다음과 같습니다.

$$(\text{민렬의 성적의 평균}) = \frac{80+90+80+70+80}{5} = 80$$

$$(\text{준영의 성적의 평균}) = \frac{100+60+80+100+60}{5} = 80$$

$$(\text{관호의 성적의 평균}) = \frac{80+80+80+80+80}{5} = 80$$

우리가 익숙한 평균은 이런 평균입니다. 이제 우리는 확률변수(그 중 이산확률변수)가 무엇인지, 우리가 알던 평균이 이산확률변수의 평균과 어떻게 이어지는지 알아볼 것입니다.

평균을 확률변수로 보는 관점

확률변수 : 무슨 상황이든 확률로 바라본다

관점을 약간 비틀어서, 각 학생의 과목중 임의로 하나를 골랐을 때, 그 과목의 점수가 몇일 확률이 어떤지를 따져봅시다.

① 민렬 : 70점이 나올 확률은 $\frac{1}{5}$, 80점이 나올 확률은 $\frac{3}{5}$, 90점이 나올 확률은 $\frac{1}{5}$입니다.

② 준영 : 60점이 나올 확률은 $\frac{2}{5}$, 80점이 나올 확률은 $\frac{1}{5}$, 100점이 나올 확률은 $\frac{2}{5}$입니다.

③ 관호 : 80점이 나올 확률은 $\frac{5}{5}=1$입니다.

이때 민렬의 과목 중 임의로 하나를 선택할 때, 선택한 과목의 성적을 X라 하면, X가 가질 수 있는 값은 70, 80, 90이며, X와 X가 특정한 값 x를 가질 확률인 $\mathrm{P}(X=x)$ $(x=70,\ 80,\ 90)$가 대응됩니다. 이를 표로 나타내면 다음과 같습니다.

X	70	80	90	합
$\mathrm{P}(X=x)$	$\frac{1}{5}$	$\frac{3}{5}$	$\frac{1}{5}$	1

이처럼 X가 여러 값을 가질 수 있는 변수이고, X가 특정한 값을 가질 확률이 대응될 때, X를 (이산)확률변수라고 하며, 위와 같은 표를 통하여 확률변수 X의 확률분포를 알 수 있습니다. 이처럼 확률과는 전혀 무관해보이는 상황들도 임의로 선택하는 상황을 강제로 설정하면 확률변수로 바라볼 수 있습니다.

이산확률변수의 평균(기댓값)

그렇다면 확률변수 X의 **평균** $\mathrm{E}(X)=m_1$에 대하여 왜 $m_1=\sum_{i=1}^{n}x_ip_i$이 성립할까요?

우리는 앞서 (민렬의 성적의 평균) $=\frac{80+90+80+70+80}{5}=80$이라는 식으로 평균을 구했습니다. 이 익숙한 수식을 변형하여 표의 구성 요소로 등장하는 수들이 나타나도록 변형하면 다음과 같습니다.

$$\begin{aligned}\frac{80+90+80+70+80}{5}&=\frac{70\times1+80\times3+90\times1}{5}\\&=\left(70\times\frac{1}{5}\right)+\left(80\times\frac{3}{5}\right)+\left(90\times\frac{1}{5}\right)\\&=80\end{aligned}$$

X	70	80	90	합
$\mathrm{P}(X=x)$	$\times\frac{1}{5}$	$\times\frac{3}{5}$	$\times\frac{1}{5}$	1

$\mathrm{E}(X)=m_1=80$

이는 표에서 '각각의 값'과 '각각의 값이 나올 확률'을 서로 곱한 것, 즉 표에서 세로로 적힌 값들을 서로 곱한 후, 그 값들을 서로 더한 것과 같음을 알 수 있으며, $\sum_{i=1}^{n}x_ip_i$가 의미하는 바와 정확히 일치합니다.

마찬가지로 준영의 과목 중 임의로 하나를 선택할 때, 선택한 과목의 점수를 확률변수 Y라 하고, 관호의 과목 중 임의로 하나를 선택할 때, 선택한 과목의 점수를 Z라 하면, Y와 Z의 확률분포를 표로 나타내고 $\mathrm{E}(Y)=m_2$와 $\mathrm{E}(Z)=m_3$를 계산하면 다음과 같습니다.

Y	60	80	100	합
$\mathrm{P}(Y=y)$	$\times\frac{2}{5}$	$\times\frac{1}{5}$	$\times\frac{2}{5}$	1

$\mathrm{E}(Y)=m_2=80$

Z	80	합
$\mathrm{P}(Z=z)$	$\times 1$	1

$\mathrm{E}(Z)=m_3=80$

$$\mathrm{E}(Y)=\left(60\times\frac{2}{5}\right)+\left(80\times\frac{1}{5}\right)+\left(100\times\frac{2}{5}\right)=80$$

$$\mathrm{E}(Z)=80\times 1=80$$

한편 확률변수의 관점에서는 평균을 기댓값(Expectation)이라는 용어로도 부릅니다.[12] 이는 '1회 시행하면 대략 결괏값이 어느 정도라고 기대할 수 있는가'를 의미하는 것이지요. 그리고 지금까지 알아본 바와 같이, 기댓값은 우리가 알고 있던 평균과 동일한 개념입니다.

12) 이 영단어의 맨 앞글자를 따 평균을 표기하는 것입니다.

분산과 표준편차

평균은 많은 정보를 알려주지만, 모든 정보를 알려주지는 못한다

지금까지 살펴본 세 명의 평균성적은 80점으로 동일합니다. 그러나 여러분도 아시다시피, 비록 분명히 세 명의 평균성적이 같기는 하지만, 세 명의 특성이 동일하다고 말하기에는 뭔가 망설여집니다.

이는 평균이라는 도구가 분명히 무언가 큰 의미[13]를 나타내기는 하지만, 평균이라는 도구만으로는 담아내지 못하는 '보이지 않는 무언가'가 있다는 것을 의미합니다. 그것은 각 과목 성적의 분포 양상입니다.

준영이는 각 과목별 성적이 들쭉날쭉하므로, 평균점수에 비해 멀리 떨어진 값들(100, 60)이 나타납니다. 그에 반해 민렬이는 각 과목별 성적이 평균점수에 비해 멀리 떨어진 정도가 준영보다는 덜합니다. 관호는 아예 모든 점수가 평균점수와 동일합니다.[14] 따라서 평균이 드러내지 못하는 '보이지 않는 무언가', 즉 각 값들이 평균으로부터 얼마나 떨어져 있는가를 수치화할 수 있는 도구가 필요합니다. 그것이 바로 분산과 표준편차입니다.

13) 평균은 자료 전체의 특징을 하나의 수로 나타낸다는 의미를 갖습니다. 이러한 역할을 하는 값을 **대푯값**이라고 합니다. 우리는 중학교에서 평균 뿐만 아니라 중앙값, 최빈값 등의 대푯값을 배운 바 있습니다. 그러나 수능에서는 평균만 알면 됩니다.

14) 이러한 민렬, 준영, 관호의 성적 분포 양상을 비교할 때 쓰이는 표현이 있습니다. 관호의 점수가 민렬의 점수보다, 민렬의 점수가 준영의 점수보다 비교적 고르게 분포되어 있다고 하는 것이죠. 여기서 고르게 분포의 의미를 '여러 가지 점수가 골고루 나온다'고 착각하기 쉬운데, 고르게 분포되었다는 것은 각각의 값들이 고만고만하게 비슷비슷하다는 의미입니다.

편차 : 얼마나 퍼졌는지 대강은 알 수 있지만, 평균적인 추세는 알 수 없는 불완전한 개념

분산과 표준편차를 공부하기 전, **편차**라는 개념을 알 필요가 있습니다. 편차는 다음과 같이 정의됩니다.

$$(\text{편차})=(\text{항목의 값})-(\text{평균})$$

편차를 이용하면 각 값들이 평균으로부터 얼마나 떨어져 있는지가 눈에 띌 것입니다. 민렬, 준영, 관호의 편차를 구해보면 각각 다음과 같습니다.

	국어	영어	수학	사회	과학
민렬	0	10	0	−10	0
준영	20	−20	0	20	−20
관호	0	0	0	0	0

그럼 이제 이 편차의 분포를 이용하면 각 값들이 평균으로부터 떨어진 정도가 대략 어느 정도 되는지를 구할 수 있을 것입니다. 이럴 때 유용한 것이 바로 평균입니다. 즉 편차의 평균인 E(편차)를 구해보려는 것이지요.

그러나 안타깝게도 민렬이든 준영든 관호든 관계없이 모든 경우에서 편차를 모두 더하면 0이 되어버립니다. 따라서 편차의 평균인 E(편차)는 항상 0일 수밖에 없습니다. 그 이유는 $\sum$의 성질과 편차의 정의를 생각하면[15] 쉽게 알 수 있습니다. 따라서 편차의 평균을 구하여 '각 항목들이 개략적으로 얼마나 퍼져있는가'를 추정하려는 시도는 보기 좋게 실패하고 말았습니다. 따라서 우리는 편차의 기본 정신은 살리면서도, 각 항목들의 평균적인 분포 추세[16]를 구할 수 있는 대안이 절실하게 필요합니다.

15) 다음과 같이 계산됩니다.

$$\begin{aligned}\sum_{i=1}^{5} \text{편차} &= \sum_{i=1}^{5}(x_i - m)\\ &= \sum_{i=1}^{5} x_i - \sum_{i=1}^{5} m\\ &= 5m - 5m\\ &= 0\end{aligned}$$

16) 이를 산포도라고 합니다. 중학교 때 분명히 배운 단어인데 잘 기억이 나지 않으실테니 설명드리자면, 산포도는 각각의 값들이 얼마나 흩어져 있는지 그 정도를 하나의 수로 나타낸 값을 뜻합니다.

분산 : 제곱을 이용하여 편차가 음숫값을 갖지 않도록 한다

편차의 합과 평균이 항상 0인 이유는 (관호의 경우와 같이 정말 모든 값이 동일하여 모든 편차값이 0인 경우를 제외하고는) 편차의 값 중 음수인 값이 나타나기 때문입니다. 따라서 음수가 아니도록 편차의 값을 적절히 조작해야 합니다. 그런데 마이너스 부호를 없애는 가장 쉬운 방법은 절댓값을 씌우는 것이지만, 여러분은 모두 절댓값을 싫어하실 것입니다. 따라서 절댓값을 대신하여 각각의 값에서 부호를 없애는 가장 만만한 대안을 생각해봅시다. 그것은 바로 제곱입니다.

각각의 편차를 제곱한 값을 구하면 다음과 같습니다.

	국어	영어	수학	사회	과학
민렬	0	100	0	100	0
준영	400	400	0	400	400
관호	0	0	0	0	0

이제 '편차의 제곱'의 평균을 구하면 각각 다음과 같습니다.

$$\text{민렬} : \frac{0+100+0+100+0}{5} = 40$$

$$\text{준영} : \frac{400+400+0+400+400}{5} = 320$$

$$\text{관호} : \frac{0+0+0+0+0}{5} = 0$$

'편차의 제곱'을 확률변수로 보더라도 같은 결과를 얻습니다. 민렬, 준영, 관호의 '성적의 편차의 제곱'을 각각 확률변수 A, B, C라 하고, 민렬, 준영, 관호의 '성적의 제곱'인 세 확률변수 X^2, Y^2, Z^2에 대하여 다음을 알아봅시다.

① A, B, C의 확률분포를 나타낸 표와 각각의 평균인 $\mathrm{E}(A)$, $\mathrm{E}(B)$, $\mathrm{E}(C)$

② X^2, Y^2, Z^2의 확률분포를 나타낸 표와 각각의 평균인 $\mathrm{E}(X^2)$, $\mathrm{E}(Y^2)$, $\mathrm{E}(Z^2)$

③ A와 X 사이의 관계, B와 Y 사이의 관계, C와 Z 사이의 관계

④ $\mathrm{V}(X)$, $\mathrm{E}(A)$, $\mathrm{E}(X^2)$ 사이의 관계

⑤ $\mathrm{V}(Y)$, $\mathrm{E}(B)$, $\mathrm{E}(Y^2)$ 사이의 관계

⑥ $\mathrm{V}(Z)$, $\mathrm{E}(C)$, $\mathrm{E}(Z^2)$ 사이의 관계

A	0	100	합
$\mathrm{P}(A=a)$	$\frac{3}{5}$	$\frac{2}{5}$	1

$\mathrm{E}(A) = 40$

B	0	400	합
$\mathrm{P}(B=b)$	$\frac{1}{5}$	$\frac{4}{5}$	1

$\mathrm{E}(B) = 320$

C	0	합
$\mathrm{P}(C=c)$	1	1

$\mathrm{E}(C) = 0$

X^2	70^2	80^2	90^2	합
$\mathrm{P}(X^2=x^2)$	$\frac{1}{5}$	$\frac{3}{5}$	$\frac{1}{5}$	1

$\mathrm{E}(X^2) = 6440$

Y^2	60^2	80^2	100^2	합
$\mathrm{P}(Y^2=y^2)$	$\frac{1}{5}$	$\frac{3}{5}$	$\frac{1}{5}$	1

$\mathrm{E}(Y^2) = 6720$

Z^2	80^2	합
$\mathrm{P}(Z^2=z^2)$	1	1

$\mathrm{E}(Z^2) = 1600$

$$A = (X-m_1)^2, \qquad B = (Y-m_2)^2, \qquad C = (Z-m_3)^2$$

$$\mathrm{V}(X) = \mathrm{E}(A) = \mathrm{E}\left((X-m_1)^2\right) = \cdots = \mathrm{E}\left(X^2\right) - (m_1)^2 = 6440 - 6400 = 40$$

$$\mathrm{V}(Y) = \mathrm{E}(B) = \mathrm{E}\left((Y-m_2)^2\right) = \cdots = \mathrm{E}\left(Y^2\right) - (m_2)^2 = 6720 - 6400 = 320$$

$$\mathrm{V}(Z) = \mathrm{E}(C) = \mathrm{E}\left((Z-m_3)^2\right) = \cdots = \mathrm{E}\left(Z^2\right) - (m_3)^2 = 6400 - 6400 = 0$$

이와 같이 편차^{2}의 평균, 다시 말해 $\mathrm{E}(\text{편차}^2)$을 **분산**이라 합니다. 분산을 계산할 때에는 정의에 따라 정직하게 구하기보다는, 위 수식에서 $\cdots$로 생략된 유도과정[17]을 통해 얻어지는 가장 오른쪽 변의 식을 이용하여 구합니다.

이 여러 과정을 거쳐 수고스럽게 계산해낸 분산의 의미를 살펴봅시다. 분산의 개념을 이용하면 우리가 개략적으로 느꼈던 다음의 개념을 명확한 수치로 나타낼 수 있습니다.

준영의 점수는 들쭉날쭉하고,
민렬이는 준영이보다는 덜하지만 점수가 퍼져 있기는 하고,
관호는 아예 모든 점수가 퍼져 있지 않고 같은 값을 갖는다

준영의 분산은 320, 민렬의 분산은 40, 관호의 분산은 0이라고 하는 것이지요.

17) 생략된 부분은 $\sum$, 평균(기댓값), 편차, 분산의 정의와 성질에 따른 단순 계산에 불과하므로 직접 유도해볼 필요는 없습니다. 머릿속으로 암산해보시거나, 암산이 어렵다면 교과서의 유도 과정을 눈으로 따라가보는 것만으로도 충분합니다.

표준편차 : 제곱으로 인한 분산값의 뻥튀기를 루트를 씌워 보정한다

분산을 통해 각각의 확률변수의 분포가 어떤지를 수치로 나타낼 수 있었지만, 계산 과정에서 제곱이 쓰이다보니 그 값이 너무 과장되는 면이 있습니다. 이렇게 제곱으로 인해 뻥튀기된 값을 보정하기 위해 분산에 루트를 씌운 값을 **표준편차**라 합니다. 준영의 표준편차는 $17.88\cdots$, 민렬의 표준편차는 $6.32\cdots$, 관호의 표준편차는 0이므로, 분산을 비교할 때보다는 값들이 작아져서 다루기 편할 것입니다.[18)]

18) 사실 표준편차는 이산확률분포에서 존재감이 별로 없습니다. 그러나 연속확률분포, 그 중에서도 정규분포에서 매우 중요한 역할을 하며, 이후 통계적 추정에서도 아주 중요한 역할을 할 것입니다.

$aX+b$의 평균, 분산, 표준편차

$\mathrm{E}(aX+b)=a\mathrm{E}(X)+b$는 $\sum$의 성질로 쉽게 증명할 수 있고, $\mathrm{V}(aX+b)=a^2\mathrm{V}(X)$는 $\mathrm{V}(X)=\mathrm{E}(X^2)-m^2$으로 쉽게 증명할 수 있고, $\sigma(aX+b)=|a|\sigma(X)$는 정의에 의해 자명합니다. 이 수식이 무엇을 의미하는지 민렬, 준영, 관호의 성적에 a와 b의 값을 구체적으로 넣어 직접 계산해보는 것도 좋습니다.

이항분포 : 표를 그리지 않는 이산확률분포, 증명 없이 꿀빨자

지금까지 알아본 바와 같이, 이산확률변수는 주로 표를 그려 해결합니다. 표를 그려야 평균, 분산, 표준편차를 구할 수 있기 때문입니다. 그런데 표를 그리지 않는 특이한 이산확률분포가 있습니다. 바로 이항분포입니다.

한 번의 어떤 사건 A가 일어날 확률이 p, 일어나지 않을 확률이 q일 때, n번의 독립시행에서 사건 A가 몇 번 일어났는지를 확률변수로 X라 하면, 직관적으로 $\mathrm{E}(X)=np$임을 알 수 있습니다. 또한 $\mathrm{V}(X)=npq$, $\sigma(X)=\sqrt{npq}$임을 증명 없이 받아들입시다.

어떤 이항분포는 표나 확률질량함수를 줍니다!

X^2	1	2	3	$\cdots$	$n-1$	n	합
$\mathrm{P}(X=x)$	${}_n\mathrm{C}_0p^n$	${}_n\mathrm{C}_1p^{n-1}q$	${}_n\mathrm{C}_2p^{n-2}q^2$	$\cdots$	${}_n\mathrm{C}_{n-1}pq^{n-1}$	${}_n\mathrm{C}_nq^n$	1

비록 이항분포의 확률분포를 표로 나타내지 않는다고 하더라도, 이항분포 또한 태생이 이산확률분포임을 잊지 말아야 합니다. 따라서 위와 같은 표나 $\mathrm{P}(X=x)={}_n\mathrm{C}_xp^xq^{n-x}$와 같은 확률질량함수를 이용하여 $X\sim\mathrm{B}(n,\,p)$라는 정보를 간접적으로 제시할 수 있습니다.

n이 충분히 크면 $X\sim\mathrm{B}(n,\,p)$인 X는 근사적으로 $X\sim\mathrm{N}(np,\,npq)$이다.

n이 충분히 큰 경우[19)] 이항분포로 번거롭게 계산할 필요 없이 정규분포를 이용하여 쉽게 계산할 수 있음이 알려져 있습니다. 이미 알려져 있으니 우리는 증명할 필요 없이 잘 쓰기만 하면 됩니다.

19) $np\geq5$, $nq\geq5$인 경우

03 확률변수 (2) : 한 페이지로 연속확률변수 끝내기

연속확률변수는 단 한 페이지만으로 수능에 필요한 모든 내용을 끝낼 수 있습니다.

일반적인 연속확률변수

연속확률변수는 확률밀도함수의 정의역 전체 구간을 정적분한 값이 1이라는 사실만 알면 됩니다. 현 교육과정 내에서 일반적인 연속확률변수에 대하여 물어볼 수 있는 것은 이게 전부입니다.[20)]

20) 연속확률변수의 평균, 분산, 표준편차를 구해야 하던 시절이 있었습니다만, -틀- 시절 이야기이므로 우리는 알 필요 없습니다.

정규분포

정규분포는 십중팔구 단순히 표준화를 하는 것만으로도 간단히 풀리는 일반적인 문제가 출제됩니다. 그런데 이는 달리 말하면 변칙적인 문제도 출제된다는 것입니다. 따라서 먼저 일반적인 문제에 대한 해법을 간단히 정리한 후, 변칙적인 문제에 대처하는 방법을 배워봅시다.

일반적인 문제 : 그냥 표준화하자

대부분의 학생들이 풀어왔듯이, 모든 정규분포를 $Z=\dfrac{X-m}{\sigma}$로 표준화하여 풀면 풀립니다. 이게 전부입니다.

변칙적인 문제 : 별 걸 다 물어본다

기출문제집에서 정규분포 문제 중 변칙적인 문제만 골라 찾아보면 정말 별의 별 것을 다 끌어와서 문제화시킨다는 것을 느낄 수 있을 것입니다. 따라서 정규분포 문제가 항상 쉽게 풀리는 것은 아닐 수도 있음을 유의하기 바랍니다.

정규분포의 성질을 숙지하자

정규분포의 대칭성, 증감성, 점근선을 숙지해야 합니다. $Z\sim \mathrm{N}(0,\,1)$인 확률변수 Z의 확률밀도함수 f와 $a<b$인 두 양수 a, b에 대하여 다음이 성립합니다.

$$\mathrm{P}(0\le Z\le a)=\mathrm{P}(-a\le Z\le 0),\quad f(a)=f(-a),\quad f(a)>f(b)$$

평소에는 당연하게 느껴지는 성질이지만, 문제화되었을 때 이 기본개념을 집요하게 물어볼 수 있습니다.

중학수학, 고등수학, 〈수학 I〉, 〈수학 II〉와 연계될 수 있다

정규분포 자체만으로는 어렵게 출제될 수 없다보니 다른 단원과의 연계를 통해 난이도 향상을 꾀할 수 있습니다. 정규분포 문제에 타 단원을 접목하여 특이하고 생소한 표현이 나올 수 있음을 잊지 말아야 합니다.

04 통계적 추정 용어 정리

통계적 추정의 궁극적 목적은 표본평균 하나만 구해서 모평균의 추정 범위를 개략적으로 구하는 것입니다. '모집단과 표본'은 그저 복선에 불과하며, '정규분포'의 개념과 '모집단의 표본'의 개념을 엮어 궁극적 목적을 달성하게 됩니다. 이에 유의하며 용어를 정리해봅시다.

모집단과 표본

통계조사에서 조사하고자 하는 대상 전체를 **모집단**이라고 하며, 모집단 전체를 조사하는 것을 **전수조사**라 합니다. 모집단에서 일부를 추출한 일부분을 **표본**이라 하고, 표본을 조사하는 것을 **표본조사**라고 합니다. 이때 추출된 표본에 포함된 대상의 개수를 **표본의 크기**라고 합니다.

모집단의 각 자료가 같은 확률로 독립적으로 추출하는 것을 **임의추출**이라고 합니다. 한 개의 자료를 추출하고 되돌려 놓고 다시 추출하는 것을 **복원추출**이라고 하는데, 이러한 복원추출은 임의추출입니다. 한편, 한 개의 자료를 추출한 후 되돌려 놓지 않고 다시 추출하는 것을 **비복원추출**이라고 하는데, 표본의 크기가 충분히 크면 비복원추출도 임의추출로 볼 수 있습니다.[21)]

모집단에서 조사하고자 하는 성질을 나타내는 확률변수를 X라 할 때, X의 평균, 분산, 표준편차를 각각 **모평균**, **모분산**, **모표준편차**라고 하며, 각각 m, σ^2, σ라 표기합니다.

모집단에서 임의추출한 크기가 n인 표본을 X_1, X_2, $\cdots$, X_n이라 할 때, 이들의 평균, 분산, 표준편차를 각각 **표본평균**, **표본분산**, **표본표준편차**라고 하며, 각각 $\overline{X}$, S^2, S라 표기합니다.[22)]

$$\overline{X}=\frac{1}{n}\sum_{k=1}^{n}X_k,\quad S^2=\frac{1}{n-1}\sum_{k=1}^{n}\left(X_k-\overline{X}\right),\quad S=\sqrt{S^2}$$

모평균 m은 고정된 상수이지만, 모집단에서 크기가 같은 표본을 임의추출했을 때 표본평균 $\overline{X}$는 추출된 표본에 따라 값이 정해지는 확률변수입니다. 따라서 $\overline{X}$의 확률분포, 평균, 표준편차 등을 구할 수 있습니다.

표본평균의 분포

일반적인 경우

모평균이 m, 모표분편차가 σ인 모집단에서 크기가 n인 표본을 임의추출할 때, 확률변수인 표본평균 $\overline{X}$에 대하여 다음이 성립합니다.

$$\mathrm{E}\left(\overline{X}\right)=m,\quad \mathrm{V}\left(\overline{X}\right)=\frac{\sigma^2}{n},\quad \sigma\left(\overline{X}\right)=\frac{\sigma}{\sqrt{n}}$$

이는 모집단이 이산확률변수인지 연속확률변수인지 관계없이, 모집단의 확률분포가 어떤지에 관계없이 항상 성립합니다.

21) 교과서는 이 충분히 큰 n의 기준을 밝히지 않고 있습니다.

22) 표본분산에서 뜬금없이 n이 아니라 $n-1$로 나누는 것은 일단 '그런가 보다' 하고 받아들입시다.

모집단이 정규분포를 따르는 경우

모집단이 정규분포 $\mathrm{N}(m,\ \sigma^2)$을 따를 때, 확률변수인 표본평균 $\overline{X}$는 $\overline{X} \sim \mathrm{N}\left(m,\ \dfrac{\sigma^2}{n}\right)$ 입니다.[23)]

23) 이는 일반적인 경우에서의 평균과 표준편차를 그대로 가져다 쓴 것입니다.

모집단이 정규분포를 따르지 않지만, 표본의 크기가 충분히 큰 경우

모집단의 분포가 정규분포를 따르지 않을 때, n이 충분히 크면 확률변수인 표본평균 $\overline{X}$는 근사적으로 $\overline{X} \sim \mathrm{N}\left(m,\ \dfrac{\sigma^2}{n}\right)$ 입니다.[24)]

24) n이 충분히 큰 경우는 $n \geq 30$일 때이며, $n < 30$인 경우는 함부로 $\overline{X}$의 분포를 정규분포로 근사하면 안 됩니다.

통계적 추정

표본조사에서 모집단의 일부인 표본을 조사하여 얻은 정보로부터 모집단의 성질을 확률적으로 추측하는 것을 **추정**이라고 합니다. 확률변수인 표본평균 $\overline{X}$을 단 한 번 구해 얻은 $\overline{x}$의 값을 이용하여 모평균 m을 추정할 때, 모평균 m이 특정 범위에 포함될 확률이 $k\%$가 되도록 어떤 닫힌구간을 정할 수 있습니다. 이때 이 닫힌구간을 모평균 m에 대한 신뢰도 $k\%$의 **신뢰구간**이라고 합니다.[25)]

25) 원래는 100개의 표본평균으로 만든 100개의 신뢰구간 중에서 약 k개가 모평균을 포함한다(수없이 많이 만들어 낸 신뢰구간 중에서 모평균을 포함하는 신뢰구간은 전체의 $k\%$ 이다).라고 말하는 것이 올바릅니다. 교과서도 정확히 이렇게 서술하고 있습니다. 그러나 수학과나 통계학과에 진학할 것이 아니라면 그냥 본문과 같이 대충 이해하셔도 무방합니다.

탁상공론형 통계적 추정

정규분포 $\mathrm{N}(m,\ \sigma^2)$을 따르는 모집단에서 크기가 n인 표본을 임의추출하여 구한 표본평균 $\overline{X}$의 값이 $\overline{x}$일 때, 모평균 m의 신뢰구간은 다음과 같습니다.

1. 신뢰도 95%의 신뢰구간 : $\left[\overline{x} - 1.96\dfrac{\sigma}{\sqrt{n}},\ \overline{x} + 1.96\dfrac{\sigma}{\sqrt{n}}\right]$
2. 신뢰도 99%의 신뢰구간 : $\left[\overline{x} - 2.58\dfrac{\sigma}{\sqrt{n}},\ \overline{x} + 2.58\dfrac{\sigma}{\sqrt{n}}\right]$

현실적인 통계적 추정

표본의 크기 n이 충분히 클 때[26)] 표본표준편차 S의 값 s를 모표준편차 σ 대신 쓸 수 있음이 알려져 있습니다. 이를 이용하여 구한 모평균 m의 신뢰구간은 다음과 같습니다.

1. 신뢰도 95%의 신뢰구간 : $\left[\overline{x} - 1.96\dfrac{s}{\sqrt{n}},\ \overline{x} + 1.96\dfrac{s}{\sqrt{n}}\right]$
2. 신뢰도 99%의 신뢰구간 : $\left[\overline{x} - 2.58\dfrac{s}{\sqrt{n}},\ \overline{x} + 2.58\dfrac{s}{\sqrt{n}}\right]$

26) $n \geq 30$일 때

05 통계적 추정, 딱 필요한 만큼만

솔직히 말해서, 통계적 추정의 대부분의 문제는 앞 단원의 '용어 정리'만 외워도 다 풀립니다. 그런데 학생들이 개념에 대한 이해 없이 워낙 공식만으로 문제를 풀다보니 개념에 대해 제대로 숙지하지 못하는 면이 있어, 이를 해소하고자 개념을 정확히 설명해드리고자 합니다.

1) 표본평균은 확률변수다.

모집단에서 크기가 1인 표본을 임의추출할 때, 표본의 값은 확률적으로 정해집니다. 크기가 n인 표본 $X_1, X_2, X_3, \cdots, X_n$을 생각하면, X_1의 값도 확률적으로 정해지고, X_2의 값도, X_3의 값도, $\cdots$, X_n의 값도 확률적으로 정해집니다. 따라서 이들을 모두 더하고 n으로 나누어 얻는 값인 표본평균 $\overline{X}$는 확률적으로 정해집니다. 따라서 표본평균은 확률변수입니다. 이 개념이 통계적 추정에서 가장 중요하므로 절대로 잊어서는 안 됩니다. 이를 상기시키기 위하여 표본평균 $\overline{X}$가 아니라 의도적으로 확률변수 $\overline{X}$라 부르도록 하겠습니다.

확률변수 $\overline{X}$의 평균, 분산, 표준편차는 어떠한가?

확률변수 $\overline{X}$의 평균 $\mathrm{E}(\overline{X})$의 값은 m이다.

확률변수 $\overline{X}$의 평균인 $\mathrm{E}(\overline{X})$의 값은 모평균인 m과 같습니다. 이는 표본평균의 값을 만들어낼 때 쓰이는 모든 값들은 결국 모두 모집단에서 임의추출된 값들이기 때문에, 모집단의 평균이 곧 $\overline{X}$의 평균일 수밖에 없다고 받아들이면 됩니다.

확률변수 $\overline{X}$의 분산 $\mathrm{V}(\overline{X})$와 표준편차 $\sigma(\overline{X})$는 표본의 크기 n이 커질수록 작아진다.

n이 커진다는 것은 곧 임의추출하는 횟수가 많아진다는 것을 뜻합니다. 그러면 n이 커지면 커질수록 확률변수 $\overline{X}$가 나타내는 값은 평균에 가까워지는 경향성을 띄게 됩니다.[27] 반대로 n이 작으면 작을수록 확률변수 $\overline{X}$가 나타내는 값은 평균에 가까워지는 경향성이 상대적으로 덜합니다.

예를 들어 생각해봅시다. 2017년 대한민국 19세~24세 남성의 키는 평균이 174, 표준편차가 5.7입니다. 계산의 편의를 위해 평균이 175, 표준편차가 5라고 두겠습니다. 표본의 크기가 1일 때에는 비교적 평균으로부터 먼 값인 '160 이하'가 나올 확률이 비교적 높습니다.[28] 그런데 표본의 크기가 100가 된다면 확률변수 $\overline{X}$가 160 이하일 확률은 급격하게 작아집니다.[29]

따라서 평균으로부터 먼 값이 나오는 경향성은 점점 줄어듭니다. 이와 반대로 평균과 비슷한 값이 나오는 경향성은 점점 커집니다. 그리고 이러한 양상은 n이 커지면 커질수록 더욱 심해집니다. 따라서 분산 $\mathrm{V}(\overline{X})$는 모분산 σ^2을 n으로 나눈 값인 $\frac{\sigma^2}{n}$이 되고, 표준편차 $\sigma(\overline{X})$는 모표준편차 σ를 $\sqrt{n}$으로 나눈 값 $\frac{\sigma}{\sqrt{n}}$이 됩니다.[30]

27) 앞서 이러한 상황을 '값이 고르게 분포한다'고 부른다고 배웠습니다.

28) 0.0013으로, 10000명 중 13명 꼴입니다.

29) 평균으로부터 먼 값들이 나올 확률은 상대적으로 매우 적고, 평균으로부터 가까운 값이 나올 확률은 상대적으로 매우 크기 때문입니다.

30) 이에 대해 수식으로 증명할 필요는 없습니다.

2) 표본표준편차 S는 확률변수 $\overline{X}$의 표준편차 $\sigma\left(\overline{X}\right)$와 전혀 무관하다.

많은 학생들이 둘을 혼동하는 경우가 많습니다. 절대로 헷갈리면 안 됩니다! 표본표준편차는 상수 n에 대하여 크기가 n인 표본을 뽑을 때마다 매번 달라질 수 있는 값이지만, 확률변수 $\overline{X}$의 표준편차 $\sigma\left(\overline{X}\right)$는 상수입니다. 모표준편차 σ가 상수이고 표본의 크기 n이 상수이기 때문입니다.

예를 들어, 대한민국 19세~24세 남성의 키가 평균이 175, 표준편차가 5일 때, 크기가 3인 표본을 임의추출해봅시다. 임의추출한 표본이 169, 183, 176일 때, 표본평균 $\overline{X}$, 표본분산 ${s_1}^2$, 표본표준편차 s_1은 각각 다음과 같습니다.

$$\overline{X} = \frac{1}{3}(169+183+176) = 176$$

$${s_1}^2 = \frac{1}{3-1}\left\{(169-176)^2+(183-176)^2+(176-176)^2\right\} = 49$$

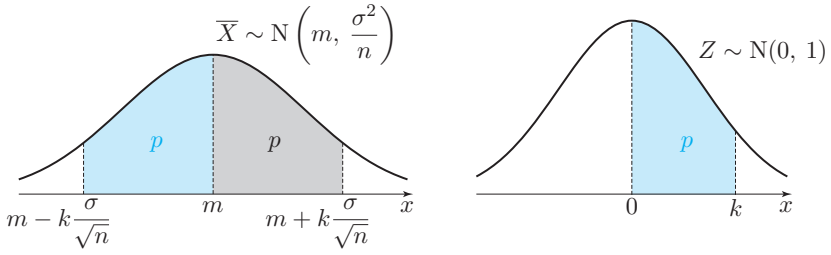
$$s_1 = \sqrt{{s_1}^2} = 7$$

이는 매번 표본을 추출할 때마다 계산해보면 다른 값이 나올 수도 있음을 알 수 있습니다.[31] 이에 반해, 확률변수 $\overline{X}$의 표준편차인 $\sigma\left(\overline{X}\right)$는 $\sigma=5$, $n=3$이므로 $\sigma\left(\overline{X}\right) = \frac{5}{\sqrt{3}} = \frac{5\sqrt{3}}{3}$임을 알 수 있습니다. 이처럼 두 개념은 전혀 다르다는 사실을 명심하기 바랍니다.

31) 예를 들어, 크기가 3인 표본 174, 177, 174에 대하여 $\overline{X}=175$, ${s_2}^2=3$, $s_2=\sqrt{3}$입니다.

3) 신뢰구간을 구하는 탁상공론 : $\overline{X} \sim \mathrm{N}\left(m, \frac{\sigma^2}{n}\right)$임을 이용

용어 정리에서 일부러 설명을 생략한 내용입니다. 신뢰구간은 확률변수 $\overline{X}$가 정규분포를 따른다는 점을 통하여 구합니다. 탁상공론이라는 한계는 있지만, 적어도 틀린 내용은 없으니 한 번 찬찬히 따라가봅시다. 우리는 $\overline{X}$의 값을 단 하나($\overline{x}$)만 구한 후, 우리가 구한 $\overline{x}$를 이용하여 만든 어떤 구간 $[\overline{x}-\bigstar, \overline{x}+\bigstar]$이 모평균 m을 포함할 확률이 몇이다라는 식을 얻기 위해 노력할 것입니다.



확률변수 $\overline{X}$가 정규분포 $\mathrm{N}\left(m, \frac{\sigma^2}{n}\right)$을 따르므로, $\mathrm{P}(0 \le Z \le k) = p$를 만족시키는 두 실수 k, p에 대하여 다음이 성립합니다.

$$\mathrm{P}\left(m-k\frac{\sigma}{\sqrt{n}} \le \overline{x} \le m+k\frac{\sigma}{\sqrt{n}}\right) = 2p$$

이 식은 정규분포의 성질에 의해 당연한 말을 하고 있습니다. $\bar{x}$가 특정한 구간에 포함될 확률이 $2p$임을 뜻하고 있죠. 그러나 옳은 말이기는 해도, 이 식은 우리가 원래 바라던 식의 꼴은 아닙니다. 우리가 원하는 꼴은 위 식에서 $\bar{x}$와 m의 위치가 서로 바뀌어야 합니다. 그러면 어떻게 하면 우리가 원하는 바를 얻을 수 있을까요?

$$\mathrm{P}\left(m-k\frac{\sigma}{\sqrt{n}} \le \bar{x} \le m+k\frac{\sigma}{\sqrt{n}} \right) = 2p$$

$$\mathrm{P}\left(-\bar{x}-k\frac{\sigma}{\sqrt{n}} \le -m \le -\bar{x}+k\frac{\sigma}{\sqrt{n}} \right) = 2p$$

$$\mathrm{P}\left(\bar{x}-k\frac{\sigma}{\sqrt{n}} \le m \le \bar{x}+k\frac{\sigma}{\sqrt{n}} \right) = 2p$$

바로 세 변에 $m+\bar{x}$를 빼준 후, 각 변에 -1을 곱해주는 것입니다. 그러면 식 변형 과정에서 부등식의 성립 여부가 달라지지 않으므로 마지막 식 역시 성립함을 알 수 있습니다. 그리고 이는 바로 우리가 원하던 바로 그 상황입니다.

딱 한 번 구한 $\bar{x}$로 만든 어떤 (신뢰)구간 $\left[\bar{x}-k\frac{\sigma}{\sqrt{n}},\ \bar{x}+k\frac{\sigma}{\sqrt{n}}\right]$이
모평균 m을 포함할 확률은 $2p$이다.

이때 $p=0.475$이면 $2p=0.95$이므로 95%의 확률이 되고, 이에 해당하는 k의 값은 1.96이므로 교과서에서 신뢰도 95%의 신뢰구간에서 1.96이라는 수를 제시했던 것입니다. 신뢰도 99%의 신뢰구간을 구할 때 2.58이라는 수를 제시하는 것 또한 마찬가지입니다. 이 원리를 이해하면 신뢰도가 몇%이든 관계없이 $\frac{\sigma}{\sqrt{n}}$의 계수를 정하여 신뢰구간을 구할 수 있습니다.

4) 현실과의 타협 : 알지도 못할 모표준편차 σ는 표본표준편차 s로 대체되었다

3)의 내용은 이론적으로는 흠잡을 데 없지만, 치명적인 단점이 있습니다. 모평균 m에 대한 신뢰구간을 구하려면 $\bar{x}$, n, σ를 알아야 할 것입니다. 표본의 크기인 n과 표본평균인 $\bar{x}$는 직접 조사했기 때문에 우리가 알고 있는 값입니다.

그러나 모표준편차 σ는 그렇지 못합니다. 우리는 지금 모평균조차 알지 못해서 그나마도 확률적으로 추정하려고 애쓰고 있는 상황입니다. 모평균도 모르는데 모표준편차를 알 도리가 있을 턱이 없습니다. 즉 3)은 이론적으로는 완벽했을지 몰라도, 현실적으로는 아무 짝에도 쓸모가 없는 공식이라는 것입니다.

그래서 통계학자들은 현실적인 타협안을 찾았습니다. 그것은 바로 알지도 못할 σ 대신, 정확히 알고 있는 표본표준편차 s를 이용하여 σ를 대체하는 것입니다.[32] 여러분이 조심해야 할 것은 2)에서 말했다시피 표본표준편차 s로 대체하는 것이지, 절대로 $\sigma\left(X\right)$로 대체하는 게 아니라는 점을 명심하고 헷갈리지 않는 것입니다.

32) 이 부분에서 갑자기 너무 뜬금없이 대충 끼워맞추는 것 아니냐고 어리둥절할 수 있습니다. 그러나 수학자들은 $n \ge 30$일 때 σ의 값이나 s의 값이나 큰 차이가 없다는 사실을 이론적으로 증명했습니다. 우리는 비록 그 원리는 알지 못하더라도, 결과만 잘 이용하여 모평균을 추정하는 데 활용만 할 수 있으면 됩니다.