OVER THE

확률과 통계

개념 정복

머리말

가장 먼저, 저는 문과가 아닌 이과라는 점을 꼭 말씀드리고 싶습니다. 저희 학교 외의 다른 좋은 학교들에선 (응용)통계학과가 문과에 있어 저를 문과로 오해하시는 분이 많으십니다. 그러나 저희 학교에서는 통계학과가 수리통계학부에서 분리된 지 얼마 되지도 않았습니다. 또한 저희 통계학과는 수리과학부와 대부분의 활동, 수업을 같이 하는 형제학과나 다름없는 곳입니다. 이 자리를 빌어 저는 엄연히 가형을 응시한 이과이며, 자연과학대 소속임을 밝힙니다.

이 책을 출판하는 과정이 참 고달팠습니다. 모종의 사정으로 〈인투더 시리즈〉를 집필한 일격필살팀과의 협업이 무산되는 바람에 출판사도 바뀌고 공동 집필에서 단독 집필로 바뀌었지요. 이로 인해 계획보다 업무 부담이 배 이상으로 늘었습니다. 게다가 익숙치 않았던 LaTeX을 익히고 작업하는 과정도 매우 힘들었습니다.

만드는 과정에서 우여곡절이 많았던 책이지만, 그만큼 더 훌륭하고 좋은 책을 완성할 수 있도록 최선을 다했습니다. 이 책을 공부하시는 분들도 제가 쏟은 노력만큼 이 책을 열심히 공부하셔서 좋은 결과를 얻으시길 바라겠습니다.

OVER THE 확률과 통계 : 개념 정복은 책 제목 그대로 수능에 필요한 〈확률과 통계〉(이하 확통)에 필요한 모든 개념을 체계적으로 정립하고, 배운 개념을 다양한 상황에서 응용할 수 있도록 구성했습니다. 노베이스도 확통을 배우는 데 어려움이 없을 것이며, 단언컨대 어떤 책과 강의보다도 수능확통에 최적화된 개념서라 확신합니다.

수능에서 확통을 공부할 때에는 〈미적분〉을 공부할 때와는 분명히 다른 방향성을 가지고 공부해야 합니다. 이는 과목 자체의 특성에서 비롯한 것도 있지만, 평가원의 수능 출제 기조가 명백하게 다르기 때문입니다. 〈미적분〉에서는 수능에서 '킬러 문제'라 불리는 극악의 난이도를 자랑하는 문제들이 수도 없이 출제되었고, 앞으로도 이들과 어깨를나란히 할 수 있을 만한 고난도 문항이 등장할 가능성을 배제할 수 없습니다.

이에 반해 확통은 의도적으로 고난도 문항의 출제를 지양하는 모습을 보이고 있습니다. 평가원에서 출제된 확통 문항 중 고난도 문항은 십중팔구 수능이 아닌 6월이나 9월 모의평가에 출제된 문항들입니다. 모의평가에서라면 몰라도, 수능에서는 되도록 확통을 어렵지는 않게 묻는다는 것이지요.

그렇다고 확통을 경시할 수는 없습니다. **어렵지는 않게 낼 뿐**이지, 평가원은 항상 학생들의 약점을 귀신같이 찾아내어 **허를 찌르는 방향**으로 출제해왔기 때문입니다. 그래서 킬러문항도 다 맞춘 학생들이 확통 3점이나 4점 문항을 틀리고 만점 달성에 실패하는 일이 비일비재합니다. 수능에서 이런 참사를 피하기 위해서는 먼저 확고한 개념을 다져야 합니다. 따라서 기초부터 심화까지 개념을 상세히 설명하였습니다.

경우의 수는 가장 기본적인 수형도를 그리는 방법부터 하나하나 가르치고, 경우의 수에서 곱셈과 나눗셈의 의미를 이해하고, 다양한 순열과 조합을 익히도록 하였습니다. 확률에서는 확률의 세계관을 확립시켜주며 경우의 수와의 연관성을 상기시키며 확률에 대한 오개념을 갖지 않도록 노력했습니다. 한편 통계는 진입장벽을 낮추기 위한 친절한 설명과 꼭 외워야 할 것들, 절대 헷갈리지 말아야 할 것들을 언급하였습니다.

또한 앞에서 언급한 내용이 뒤의 복선이고, 학습 과정에서 생길 수 있는 의문은 반드시 주석이나 뒷 내용에서 해결되도록 짜임새 있게 구성하였습니다. 체계적인 학습이 가능할 것이니 기대하셔도 좋습니다.

OVER THE 확률과 통계 : 개념 정복은 개념을 다루는 Basic과 응용을 다루는 Theme로 나뉩니다.

Basic : 확통의 모든 개념을 물샐틈없이 확립합니다.

Basic에서는 오개념과 같은 시행착오를 겪지 않고, 확통에 필요한 개념을 차곡차곡 쌓아갈 수 있도록 개념 설명의 구조를 설계했습니다. 거기에 더해 고등수학에서 다루는 경우의 수 내용부터 차례대로 학습하면서, 확통에서 다루는 경우의 수, 확률, 통계의 내용을 모두 꾹꾹 눌러 담아냈습니다. 이를 통해 수능에 최적화된 방법으로 수능 확통에 필요한 모든 개념을 학습할 방향성을 제시합니다. 또한 Basic에서는 단원별로 저난도 기출문제를 제공하므로, 배운 내용을 적용하며 개념을 재확인할 수 있습니다.

Theme : 다양한 테마별 상황을 접하며 개념을 응용합니다.

Theme에서는 Basic에서 배운 개념을 확장시킨 다양한 상황을 제시합니다. 각각의 Theme는 그 자체로 문항의 역할을 하기도 합니다. 이러한 Theme를 접하며 개념을 응용하여 문항을 분석하고, 배운 개념을 확장해 문제를 해결할 수 있도록 도와줍니다. 이를 통해 단순히 상황별 대처법을 암기하는 것에 그치지 않고, 개념의 다양한 활용까지 익힙니다.

OVER THE 확률과 통계 시리즈를 소개합니다.

OVER THE 확률과 통계 : 개념 정복의 후속 교재는 두 가지가 있습니다. 하나는 평가원 기출문제집 **OVER THE 확률과 통계 : 트윈 기출**이고, 다른 하나는 최상위권용 심화 개념 탐구 & 고난도 문제집 **OVER THE 확률과 통계 : 심화 문풀**입니다.

OVER THE 확률과 통계 : 트윈 기출은 지금까지 세상에 없던 확통 최적화 기출문제집을 표방합니다. 수능 시험지에는 '이 문제는 확통의 어떤 개념을 쓰는 문제이다'라고 문제의 풀이법이나 유형이 직접 적혀 있지 않습니다. 따라서 수능 실전 상황에서 확통 문제를 만났을 때 가장 중요한 것은 문제의 상황을 분석하고, **필요한 개념과 논리가 무엇인지 파악하는 것**입니다.

그러나 기존의 기출문제집들은 이러한 확통의 과목적 특성을 전혀 고려하지 않은 채 '이 문제는 중복조합을 쓰는 문제이다', '이 문제는 조건부확률을 써서 풀이한다'고 문제를 스포일러하는 경우가 많습니다. 따라서 단원별로 1회독하며 **필요한 개념과 논리가 무엇인지 파악하는 훈련**을 할 수 있고, 2회독은 세부 유형별로 풀며 약점을 극복하고, **해를 거듭며 발전하는 평가원 기출문제의 흐름**을 읽을 수 있습니다. 이처럼 같은 문항을 다른 배열로 두 번 풀이할 수 있어 **1권 값에 2권 값을 하는 최상의 가성비**를 자랑합니다.

OVER THE 확률과 통계 : 트윈 기출은 최상위권을 위한 교재이자, 최상위권이 되기 위한 교재입니다. 교육과정에서 생략된 내용들과 교과서가 얼렁뚱땅 넘어가는 개념에 대해서도 철저히 배워보고, 확통에 대한 심화된 발상을 이론적으로 설명합니다. 또한 교육청/사관학교/경찰대에 출제된 **평가원보다 훨씬 어려운 고난도 확통 문제**들을 풀이하며 수능에서 확통에 발목잡히지 않도록 마침표를 찍을 것입니다.

Contents

OVER THE 확률과 통계
개념 정복

Basic

시행착오 없는 노베이스 탈출의 지름길

01 경우의 수 용어 정리

고1 수학

사건

반복할 수 있는 실험이나 관찰에 의하여 일어나는 결과를 **사건**이라고 합니다.

합의 법칙

두 사건 A, B가 동시에 일어나지 않을 때, 사건 A, B가 일어나는 경우의 수가 각각 m, n이면 사건 A 또는 사건 B가 일어나는 경우의 수는 $m+n$입니다. 이를 **합의 법칙**이라고 합니다.[1)]

1) 합의 법칙은 어느 두 사건도 동시에 일어나지 않는 셋 이상의 사건에 대해서도 성립합니다.

곱의 법칙

두 사건 A, B에 대하여 사건 A가 일어나는 경우의 수가 m이고 그 각각에 대하여 사건 B가 일어나는 경우의 수가 n일 때, 두 사건 A, B가 잇달아 일어나는 경우의 수는 $m \times n$입니다. 이를 **곱의 법칙**이라고 합니다.[2)]

2) 곱의 법칙은 잇달아 일어나는 셋 이상의 사건에 대해서도 성립합니다.

계승(팩토리얼)

1부터 n까지의 자연수를 차례로 곱한 것을 n의 **계승**이라 하고, $n!$과 같이 표기합니다.[3)] 한편 자연수가 아닌 수인 0에 대하여 $0!=1$이라 정의합니다.

3) 읽을 때는 '엔 팩토리얼'이라 읽습니다.

순열과 조합

순열

서로 다른 n개에서 r $(0 \le r \le n)$개를 택하여 일렬로 나열하는 것을 'n개에서 r개를 택하는 **순열**'이라 하고, 이 순열의 수를 ${}_n\mathrm{P}_r$과 같이 표기합니다.

조합

서로 다른 n개에서 순서를 생각하지 않고 r $(0 \le r \le n)$개를 택하는 것을 'n개에서 r개를 택하는 **조합**'이라 하고, 이 조합의 수를 ${}_n\mathrm{C}_r$과 같이 표기합니다.

순열과 조합의 식과 관계

순열과 조합에 대하여 다음이 성립합니다.[4)]

$$ {}_n\mathrm{P}_r = \frac{n!}{(n-r)!}, \quad {}_n\mathrm{C}_r = \frac{{}_n\mathrm{P}_r}{r!} = \frac{n!}{r!\,(n-r)!} = {}_n\mathrm{C}_{n-r} $$

4) ${}_n\mathrm{C}_r = {}_n\mathrm{C}_{n-r}$이므로 조합은 대칭성을 갖습니다.

확률과 통계

여러 가지 순열과 조합

원순열

서로 다른 n개를 원형으로 배열하는 순열을 **원순열**이라고 하며, 원순열의 수는 $(n-1)!$입니다.

중복순열

서로 다른 n개에서 중복을 허용하여 r개를 택하는 순열을 **중복순열**이라 하고, 이 중복순열의 수를 ${}_n\Pi_r$과 같이 표기하며, ${}_n\Pi_r = n^r$이 성립합니다.

같은 것이 있는 순열

n개 중에서 서로 같은 것이 각각 p개, q개, $\cdots$, r개 있을 때, 이 n개를 모두 일렬로 배열하는 순열의 수는 $\dfrac{n!}{p!q!\cdots r!}$입니다.[5)]

5) '같은 것이 있는 순열'은 앞으로 이 책에서 같있순이라 줄여 부르기로 합시다.

중복조합

서로 다른 n개에서 중복을 허용하여 r개를 택하는 조합을 **중복조합**이라 하고, 이 중복조합의 수를 ${}_n\mathrm{H}_r$과 같이 표기하며, ${}_n\mathrm{H}_r = {}_{n+r-1}\mathrm{C}_r$이 성립합니다.

이항정리

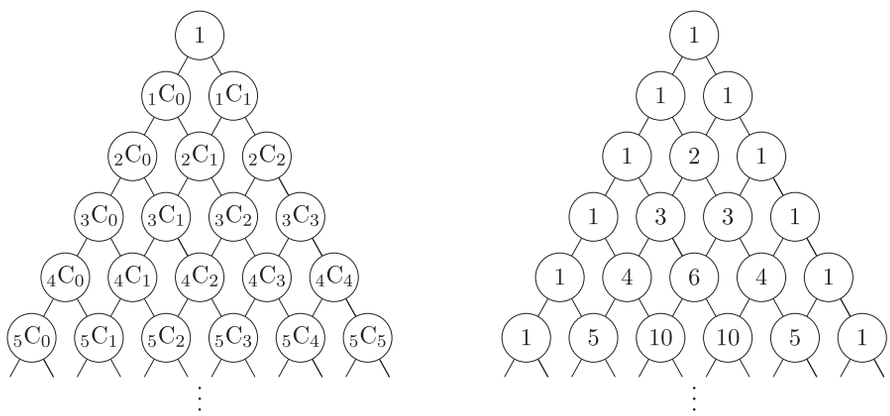
이항정리와 이항계수

다항식 $(a+b)^n$에 대하여 다음이 성립하며, 이를 **이항정리**라 합니다.

$$\begin{aligned}(a+b)^n &= \sum_{r=0}^{n} {}_n\mathrm{C}_r a^r b^{n-r} \\ &= \sum_{r=0}^{n} {}_n\mathrm{C}_r a^{n-r} b^r \\ &= {}_n\mathrm{C}_0 a^n + {}_n\mathrm{C}_1 a^{n-1} b + \cdots + {}_n\mathrm{C}_r a^{n-r} b^r + \cdots + {}_n\mathrm{C}_{n-1} a b^{n-1} + {}_n\mathrm{C}_n b^n \\ &= {}_n\mathrm{C}_0 a^n b^0 + {}_n\mathrm{C}_1 a^{n-1} b^1 + \cdots + {}_n\mathrm{C}_r a^{n-r} b^r + \cdots + {}_n\mathrm{C}_{n-1} a^1 b^{n-1} + {}_n\mathrm{C}_n a^0 b^n\end{aligned}$$

이때 우변의 각 항의 계수 ${}_n\mathrm{C}_0$, ${}_n\mathrm{C}_1$, $\cdots$, ${}_n\mathrm{C}_r$, $\cdots$, ${}_n\mathrm{C}_n$을 **이항계수**라 합니다.

파스칼의 삼각형

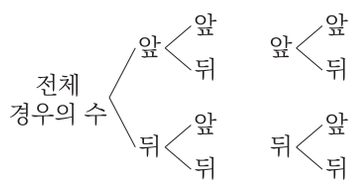


위 그림과 같이 이항계수를 삼각형 모양으로 나타낸 것을 **파스칼의 삼각형**이라고 합니다. 조합의 대칭성에 의해 ${}_n\mathrm{C}_r = {}_n\mathrm{C}_{n-r}$ 이므로 파스칼의 삼각형은 좌우 대칭이며, ${}_n\mathrm{C}_r = {}_{n-1}\mathrm{C}_{r-1} + {}_{n-1}\mathrm{C}_r$ 이 성립함을 시각적으로 확인할 수 있습니다.

02 수형도, 합의 법칙, 곱의 법칙

경우의 수 단원의 실력은 '수형도를 얼마나 잘 그리느냐'에 달려 있습니다. 그리고 수형도를 얼마나 잘 그리느냐는 수형도를 얼마나 덜 그리느냐, 궁극적으로는 수형도를 얼마나 그리지 않느냐에 달려 있습니다. 잘 그리는 방법이 얼마나 덜 그리고 안 그리냐에 달렸다? 언뜻 보기에는 앞뒤가 맞지 않는 이 말이 도대체 무슨 의미인지 차차 알아보도록 합시다.

수형도



(a) (b)

수형도란 경우를 세는 방법의 하나로, '나무 모양의 그림'이라는 의미를 갖고 있습니다. 이름 그대로 나무가 가지를 뻗어나가는 것과 같은 모습으로 경우를 세어나갑니다.[6] 그림 (a)는 동전을 두 번 던졌을 때, 전체 경우의 수 4가지를 수형도를 이용하여 그린 예입니다. 이때 수형도에서 가지가 갈라지는 지점을 마디라 부르도록 합시다. 그림 (a)에는 3개의 마디가 있음을 알 수 있습니다.

그림 (a)를 보면 가장 왼쪽 마디인 '전체 경우의 수'는 모든 수형도에 반드시 등장할 수밖에 없음을 알 수 있습니다. 따라서 그림 (b)와 같이 동일한 상황에서 가장 왼쪽 마디를 생략하고 그리면 마치 두 개의 수형도를 각각 그린 것과 같은 형태를 띱니다. 우리는 앞으로 수형도를 그릴 때 그림 (b)와 같이 불필요한 '전체 경우의 수'는 생략하고 그리기로 약속합시다.

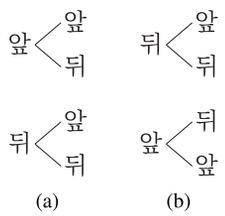
수형도는 경우의 수의 처음이자 끝이라고 해도 과언이 아닙니다. 모든 경우의 수 문제는 수형도를 그리면 반드시 풀립니다. 특히 수능 주관식 문항은 답이 1000 미만이므로, 아무런 이론적 배경을 갖추지 못했더라도 시간 내에 수형도를 그린다면 반드시 맞힐 수 있습니다.[7]

6) 우리가 경우의 수를 셀 때 '가짓수를 센다'의 가지는 결국 수형도의 나뭇가지의 수를 세는 것과 같은 것이지요. 국어사전에서는 이 두 단어를 동음이의어로 보고 있는데, 경우의 수의 관점에서는 다의어로 생각해도 일리가 있겠습니다.

7) 물론 이론상 그렇다는 것이고, 더 나은 길이 있는데도 불구하고 모든 문제를 수형도로 풀이할 필요는 없습니다.

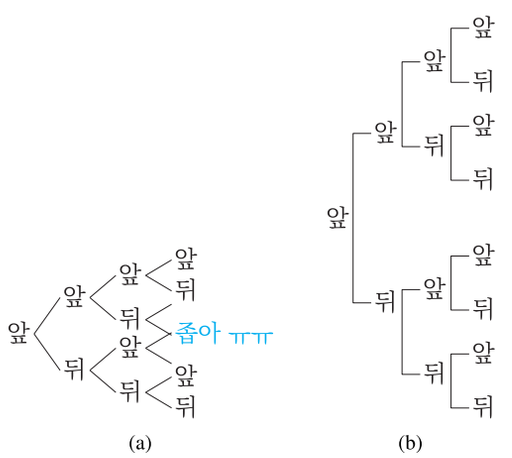
수형도를 잘 그리는 원칙

순서에 대한 규칙을 정하자



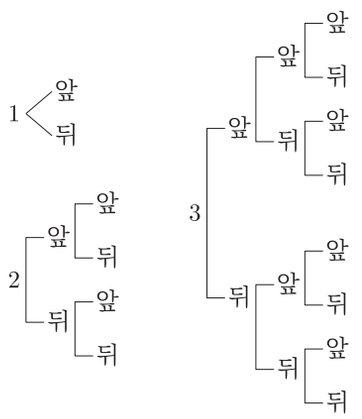
수형도를 혼동 없이 그리기 위해서는 일정한 규칙을 정하는 것이 좋습니다. 가령 그림 (a)와 같은 수형도에서는 항상 '앞'을 마디보다 위쪽에 적고, '뒤'를 마디보다 아래쪽에 적는 규칙을 따라 그리고 있습니다. 그림 (b)와 같이 순서가 뒤죽박죽이라면, 수형도를 그리는 과정에서도 헷갈릴 것이고, 나중에 검토할 때 수형도에 틀린 부분이 있는지 검증하기도 까다로울 것입니다.

첫 마디는 넓게 뻗어나가자



수형도의 특성상 뒤의 가지가 많이 퍼질 수밖에 없습니다. 그래서 그림 (a)와 같이 첫 마디에서 좁게 뻗어나가면 나중에 그림의 한가운데에서 위쪽 가지와 아래쪽 가지가 겹치게 됩니다. 그림 (b)와 같이 첫 마디에서 위아래로 멀리 뻗어나가면 이러한 문제를 예방할 수 있습니다.

원칙 적용의 예



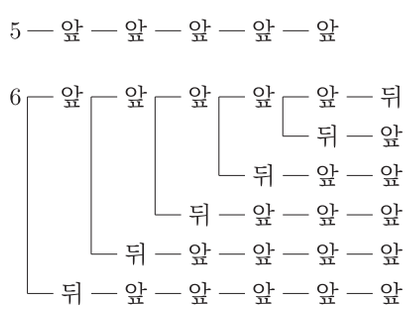
주사위를 던져 나온 눈의 수만큼 동전을 던지는 경우의 수형도를 일부만 그려보면 위 그림과 같습니다. 가장 작은 수인 1부터 위에 적어나가면서, 동전은 앞을 항상 뒤보다 위에 적어 주고 있음을 알 수 있습니다. 4, 5, 6일 때는 일부러 생략했으므로, 반드시 직접 빈 종이에 수형도를 그려보시기 바랍니다.

예제 1. 주사위를 던져 나온 눈의 수만큼 동전을 던진다. 이때 앞면이 나온 횟수가 5인 경우의 수는?

예제 1 풀이

> 주사위를 던져 나온 눈의 수만큼 동전을 던진다. 이때 앞면이 나온 횟수가 5인 경우의 수는?

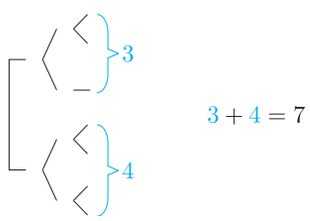
주사위를 던져 나온 눈의 수가 1, 2, 3, 4인 경우는 셀 필요가 없으므로, 수형도는 '첫 마디가 5인 경우'와 '첫 마디가 6인 경우' 두 개만 그리면 됩니다.



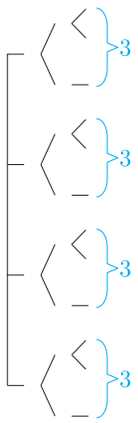
5에서는 앞앞앞앞앞 1가지, 6에서는 앞앞앞앞앞뒤, 앞앞앞앞뒤앞, 앞앞앞뒤앞앞, 앞앞뒤앞앞앞, 앞뒤앞앞앞앞, 뒤앞앞앞앞앞 6가지가 있습니다. 따라서 구하는 경우의 수는 $1+6=7$ 가지입니다.

합의 법칙과 곱의 법칙 : 수형도를 그리는 데 활용되는 기본 원리

합의 법칙과 곱의 법칙은 수형도를 그리는 데 활용되는 기본 원리입니다. 수형도를 그리다 보면 수형도가 매 마디에서 갈라질 때, 갈라져 나온 이후의 구조가 다른지 같은지를 따지게 됩니다. 예상하셨겠듯이, 다르다면 합의 법칙을, 같다면 곱의 법칙을 씁니다.

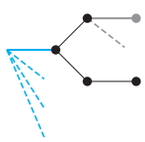


갈라진 각 마디 이후 단계에서 수형도의 구조가 다르다면, 구조가 각기 다른 수형도를 각각 그리고, 각각의 경우의 수를 더하여 셉니다. 이것이 곧 합의 법칙입니다.



$3 \times 4 = 12$

갈라진 각 마디 이후 단계에서 수형도의 구조가 같다면, 하나만 그리고 나머지 수형도는 그리지 않아도 됩니다. 그저 수형도를 한 번만 그려 세고, 구조가 동일한 수형도가 몇 개인지를 파악하여 곱해주면 되는 것입니다. 이것이 곧 곱의 법칙입니다.



$4 \times \{(2 \times 1) + 1\}$

곱의 법칙에서 주어진 수형도를 압축하여 그려봅시다. 합의 법칙과 곱의 법칙을 융합하여 그림과 같이 한 개의 수형도만 그린 후 아래에는 점선으로 표시하거나 곱하기로 표시하여 수형도를 생략할 수 있고, 수형도의 가지가 달라지는 마디는 덧셈으로 표기할 수 있습니다.

결국 합의 법칙과 곱의 법칙을 쓰는 것은, 수형도가 그려지는 구조를 파악하여 가급적 수형도를 덜 그리기 위한 것이라 생각할 수 있습니다. 수형도의 패턴이 반복되면 곱의 법칙을 이용해 수형도를 한 번만 그려 생략하고, 패턴이 다른 경우에만 어쩔 수 없이 각각의 수형도를 그린 후 합의 법칙으로 더하는 것이지요.

지금까지 배운 내용만 이용하여 아래의 문제를 풀어봅시다. 단, 같은 것이 있는 순열(같있순)은 사용하지 마시기 바랍니다.

예제 2. 여섯 개의 자음 ㄱ, ㄱ, ㄴ, ㄴ, ㄷ, ㄷ을 일렬로 나열하여 문자열을 만든다. ㄱ은 ㄱ과 서로 이웃하지 않고, ㄴ은 ㄴ과 서로 이웃하지 않고, ㄷ은 ㄷ과 서로 이웃하지 않도록 배열된 문자열의 개수를 구하시오.

예제 2 풀이

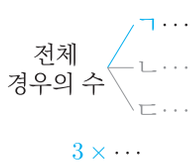
> 여섯 개의 자음 ㄱ, ㄱ, ㄴ, ㄴ, ㄷ, ㄷ을 일렬로 나열하여 문자열을 만든다. ㄱ은 ㄱ과 서로 이웃하지 않고, ㄴ은 ㄴ과 서로 이웃하지 않고, ㄷ은 ㄷ과 서로 이웃하지 않도록 배열된 문자열의 개수를 구하시오.

처음에 ㄱ, ㄴ, ㄷ으로 분류하여 수형도를 3개 그리겠다고 생각했다면, 그것만으로도 절반은 성공한 것입니다. 이때 ㄱ으로 시작하든, ㄴ으로 시작하든, ㄷ으로 시작하든, 이후 수형도의 구조가 같은지 곰곰이 생각해봅시다.

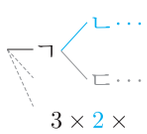
ㄱ의 입장에서 남은 것은 ㄱㄴㄴㄷㄷ이므로, 두 개씩 남은 ㄴ, ㄷ 중 하나를 선택하면 됩니다.
ㄴ의 입장에서 남은 것은 ㄴㄷㄷㄱㄱ이므로, 두 개씩 남은 ㄱ, ㄷ 중 하나를 선택하면 됩니다.
ㄷ의 입장에서 남은 것은 ㄷㄱㄱㄴㄴ이므로, 두 개씩 남은 ㄱ, ㄴ 중 하나를 선택하면 됩니다.

즉 서로 입장만 다를 뿐 완전히 동일한 상황임을 알 수 있습니다. 그래서 ㄱ으로 시작하는 경우의 수인 n을 구한 후, 여기에 3을 곱하면 전체 경우의 수인 $3n$을 구할 수 있습니다.[8)]

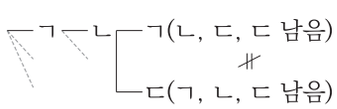
8) 이것이 곧 곱의 법칙입니다.



여기까지 우리가 풀이하며 작성하는 식은 $3\times$ 입니다.

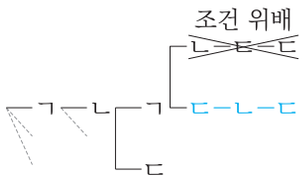


이제 ㄱ에만 집중하여 수형도를 그려나가봅시다. 앞서 말한 것과 같은 원리로, ㄴ을 고르든 ㄷ을 고르든 뒤의 수형도의 구조가 동일합니다. 따라서 ㄱ-ㄴ 이후만 그려나가면 됩니다.
여기까지 우리가 풀이하며 작성하는 식은 $3\times2\times$ 입니다.



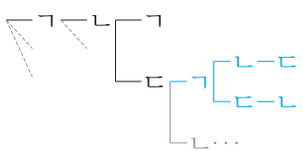
그런데 ㄱ-ㄴ-ㄱ과 ㄱ-ㄴ-ㄷ은 상황이 달라집니다. ㄱ-ㄴ-ㄱ에서는 ㄱ이 이미 다 쓰여 ㄴㄷㄷ만 남고, ㄱ-ㄴ-ㄷ은 ㄱ, ㄴ, ㄷ이 하나씩 남아 있는 상황이기 때문입니다. 따라서 이런 경우는 각각의 경우의 수인 a, b를 구한 후 더하여야 합니다.[9)] 여기까지 우리가 풀이하며 작성하는 식은 $3\times2\times($ $\quad+\quad$ $)$ 입니다. 더하기의 왼쪽에는 a의 값을, 더하기의 오른쪽에는 b의 값을 적으면 됩니다.

9) 이것이 곧 합의 법칙입니다.



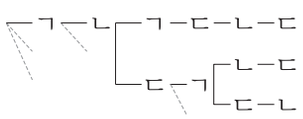
$$3 \times 2 \times (1 + \quad)$$

이제 a를 구하기 위해 ㄱ-ㄴ-ㄱ를 마저 그려봅시다. ㄱ-ㄴ-ㄱ-ㄴ-ㄷ-ㄷ는 ㄷ끼리 이웃하여 문제의 조건에 위배되므로 세지 않습니다. ㄱ-ㄴ-ㄱ-ㄷ-ㄴ-ㄷ는 문제의 조건을 만족시킵니다.[10] 따라서 $a = 1$입니다.



$$3 \times 2 \times (1 + 2 \times 2)$$

이제 b를 구하기 위해 ㄱ-ㄴ-ㄷ를 마저 그려봅시다. ㄱ-ㄴ-ㄷ-ㄱ와 ㄱ-ㄴ-ㄷ-ㄴ는 뒤의 수형도의 구조가 같습니다. 따라서 ㄱ-ㄴ-ㄷ-ㄱ만 풀이하고 2를 곱하면 b를 구할 수 있습니다. ㄱ-ㄴ-ㄷ-ㄱ에서는 ㄱ-ㄴ-ㄷ-ㄱ-ㄴ-ㄷ와 ㄱ-ㄴ-ㄷ-ㄱ-ㄷ-ㄴ의 2가지가 있으므로, $b = 2 \times 2$입니다.



$$3 \times 2 \times (1 + 2 \times 2)$$

따라서 최종적으로 정답은 $3 \times 2 \times (\ 1 \ + \ 2 \times 2\ ) = 30$이 됩니다.

10) 여기서 ㄱ-ㄴ-ㄱ-ㄷ-ㄷ-ㄴ은 왜 그리지 않았는지 의문이 들 수 있지만, 앞서 다룬 예제 1에서 주사위의 눈이 1, 2, 3, 4가 나온 경우를 그리지 않은 것처럼 당연히 조건에 위배되기 때문에 그리지 않은 것입니다.

마치며

이와 같이 경우의 수를 잘 하는 것은, 수형도를 얼마나 잘 그리느냐에 달려 있습니다. 그리고 수형도를 잘 그리는 것은 '몇 개의 수형도를 그릴 것이며', '언제 더하고 언제 곱하는지'를 아는 것과 같습니다.

이에 더하여, 앞으로 다룰 여러 가지 순열과 조합 공식들은 그저 자주 나오는 수형도를 그리지 않고도 한 방에 구하는 테크닉에 불과할 뿐입니다. 결국 이러한 테크닉을 적재적소에 활용한다면, 수형도를 그리지 않고도 수형도의 가짓수를 알 수 있을 것입니다.

자, 이제는 이 단원을 시작하며 말했던 아래의 문장이 다시금 와닿을 것입니다.

> 경우의 수 단원의 실력은 '수형도를 얼마나 잘 그리느냐'에 달려 있습니다. 그리고 수형도를 얼마나 잘 그리느냐는 수형도를 얼마나 덜 그리느냐, 궁극적으로는 수형도를 얼마나 그리지 않느냐에 달려 있습니다.

03 집합으로 경우의 수를 해석하는 관점

집합을 이용하면 경우의 수를 보다 매끄럽게 설명할 수 있습니다. 그런데 교과서에서는 경우의 수를 집합으로 설명하고 있지 않으므로, 당연히 이에 대한 용어도 정의되어 있지 않습니다. 따라서 어쩔 수 없이 교과 외 용어를 먼저 약속하도록 합시다.[11)]

11) 교과서에는 없지만, 일반적인 수학에서 널리 쓰이는 용어입니다.

집합으로 정의하는 여러 가지 용어들

전사건

반복할 수 있는 실험이나 관찰에 의하여 일어날 수 있는 모든 각각의 결과들을 원소로 가지는 집합을 전사건이라 부르기로 합시다.[12)]

예를 들어, 주사위를 한 번 던지는 상황을 생각해보겠습니다. 주사위를 던지면 1, 2, 3, 4, 5, 6이 나올 수 있으므로 $S=\{1,\,2,\,3,\,4,\,5,\,6\}$라 생각할 수 있습니다.

12) 전사건은 뒤의 확률 단원에서 배울 **표본공간(Sample Space)**과 동일한 개념입니다. 따라서 집합의 이름을 주로 S라 짓습니다.

근원사건, 사건, 공사건

전사건 S의 부분집합을 **사건**이라 하고, 특히 원소의 개수가 1인 사건을 **근원사건**이라 부르기로 합시다. 사건 A의 원소의 개수를 $n(A)$라 할 때, $n(A)=0$인 경우, 사건 A를 공사건이라 부르기로 합시다.[13)]

앞서 주사위를 한 번 던지는 상황을 전사건으로 할 때, 각각의 예를 들면 다음과 같습니다.

1. S의 부분집합인 $B=\{1\}$은 '주사위를 한 번 던질 때 1이 나오는 사건'을 의미하며, 동시에 근원사건입니다.
2. S의 부분집합인 $C=\{2,\,4,\,6\}$은 '주사위를 한 번 던질 때 짝수가 나오는 사건'을 의미합니다.
3. $D=\{7\}$은 S의 부분집합이 아니므로 공사건입니다.

13) 본문에서는 A를 S의 부분집합으로 전제하였으므로 $n(A)=0$이면 A를 공사건이라 할 수 있었습니다. A가 S의 부분집합이 아닌 상황도 일반적으로 설명하기 위해서는 '$n(A\cap S)=0$인 집합 A'를 공사건이라 부르는 것이 자연스럽습니다.

합사건, 곱사건, 여사건

두 사건 A와 B에 대하여 다음의 세 사건 C, D, E를 생각할 수 있습니다.

$$C=A\cup B,\quad D=A\cap B,\quad E=A^C$$

이때 C를 'A와 B의 합사건', D를 'A와 B의 곱사건', E를 'A의 **여사건**'이라 부르기로 합시다. 합사건은 'A 또는 B가 일어나는 사건'을 의미하고, 곱사건은 'A와 B가 동시에 일어나는 사건'을 의미하고, 여사건은 'A가 일어나지 않는 사건'을 의미합니다.

배반사건

한편 $A\cap B=\varnothing$인 경우, 즉 곱사건이 공사건인 경우, 두 사건 A, B를 서로 **배반사건**이라고 합니다. 사건 A와 A의 여사건 A^C는 서로 배반사건입니다.

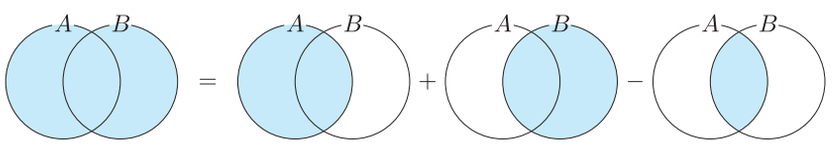
정리

결국 찬찬히 뜯어보면 용어만 집합에서 사건으로 바꾸었을 뿐임을 알 수 있습니다. 따라서 우리는 앞으로 경우의 수를 다룰 때 집합의 연산 체계를 그대로 가져다 쓸 수 있습니다.

합의 법칙의 재해석

합의 법칙은 두 사건이 서로 배반사건인 경우를 논하는 것이다.

앞서 말했듯이 우리는 경우의 수를 다룰 때 집합의 연산 체계를 그대로 활용할 수 있습니다. 이를 이용하여 합의 법칙을 다시 설명해봅시다.



우리는 위의 벤 다이어그램을 통해 집합에서 다음의 식이 성립함을 고1 수학에서 배운 바 있습니다.

$$n(A \cup B) = n(A) + n(B) - n(A \cap B)$$

이는 합사건과 곱사건에도 그대로 적용할 수 있습니다. 이를 확실히 숙지한 상태에서, 우리가 합의 법칙을 어떻게 정의했는지 다시 살펴봅시다.

> 두 사건 A, B가 동시에 일어나지 않을 때, 사건 A, B가 일어나는 경우의 수가 각각 m, n이면 사건 A 또는 사건 B가 일어나는 경우의 수는 $m+n$입니다.

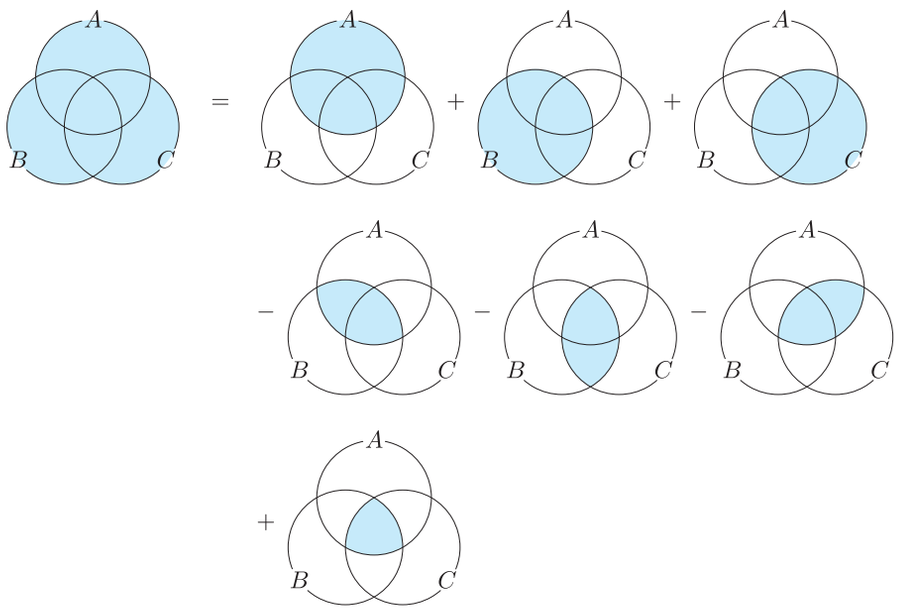
여기서 두 사건 A, B가 동시에 일어나지 않는다는 것은 두 사건이 서로 배반사건임을 의미합니다. 따라서 합의 법칙은 두 사건이 배반사건인 아주 간단한 상황만을 논하고 있음을 알 수 있습니다. 그러나 집합을 이용하면 두 사건이 서로 배반사건이 아니더라도 곱사건의 원소의 수를 빼줌으로써 합의 법칙을 확장할 수 있습니다.[14)]

14) 이를 포함과 배제의 원리라는 이름으로 부르기도 합니다.

세 개 이상의 사건에도 확장된 합의 법칙을 적용할 수 있다.

세 개 이상의 사건에 대해서도 합의 법칙을 일반화할 수 있습니다. 단 보통 네 개 이상의 사건에 대해서 서로가 모두 배반사건이 아니라면 상황이 너무 복잡해집니다. 따라서 네 개 이상의 사건에 대해서는 서로가 모두 배반사건인 경우[15]만 다룹니다. 세 개의 사건에 대해서만 합의 법칙을 확장하여 논해봅시다.

15) 즉 기본적인 합의 법칙으로 논할 수 있는 경우



위의 벤 다이어그램을 통해 다음의 식이 성립함을 알 수 있습니다.

$$\begin{aligned} n(A \cup B \cup C) = {} & n(A) + n(B) + n(C) \\ & - n(A \cap B) - n(B \cap C) - n(C \cap A) \\ & + n(A \cap B \cap C) \end{aligned}$$

이를 이용하여 앞서 풀었던 아래의 문제를 집합의 관점으로 다시 풀어봅시다. 이번에는 풀이 과정에서 같은 것이 있는 순열(같있순)을 사용하세요.

예제 1. 여섯 개의 자음 ㄱ, ㄱ, ㄴ, ㄴ, ㄷ, ㄷ을 일렬로 나열하여 문자열을 만든다. ㄱ은 ㄱ과 서로 이웃하지 않고, ㄴ은 ㄴ과 서로 이웃하지 않고, ㄷ은 ㄷ과 서로 이웃하지 않도록 배열된 문자열의 개수를 구하시오.

예제 1 풀이

> 여섯 개의 자음 ㄱ, ㄱ, ㄴ, ㄴ, ㄷ, ㄷ을 일렬로 나열하여 문자열을 만든다. ㄱ은 ㄱ과 서로 이웃하지 않고, ㄴ은 ㄴ과 서로 이웃하지 않고, ㄷ은 ㄷ과 서로 이웃하지 않도록 배열된 문자열의 개수를 구하시오.

전사건을 S, 구하는 사건을 E, ㄱ끼리 이웃하는 사건을 A, ㄴ끼리 이웃하는 사건을 B, ㄷ끼리 이웃하는 사건을 C라 할 때, 다음이 성립합니다.

$$\begin{aligned} n(A\cup B\cup C) &= n(A)+n(B)+n(C) \\ &\quad -n(A\cap B)-n(B\cap C)-n(C\cap A) \\ &\quad +n(A\cap B\cap C) \end{aligned}$$

$$n(E)=n\left((A\cup B\cup C)^C\right)=n(S)-n(A\cup B\cup C)$$

이때 같있순에 의해 $n(S)=\dfrac{6!}{2!2!2!}=90$임은 쉽게 알 수 있고, $n(A)=x$, $n(A\cap B)=y$, $n(A\cap B\cap C)=z$라 하면 다음이 성립합니다.[16)]

$$n(A) \quad = \quad n(B) \quad = \quad n(C) \quad = x$$

$$n(A\cap B)=n(B\cap C)=n(C\cap A)=y$$

따라서 $n(E)=90-(3x-3y+z)$이므로 x, y, z를 구하면 답을 구할 수 있습니다. x, y는 각각 같있순에 의해 $x=\dfrac{5!}{2!2!}$, $y=\dfrac{4!}{2!}$이고, $z=3!$입니다. 따라서 $n(E)=90-(90-36+6)=30$입니다.

16) 아래의 식이 왜 성립하는지 이해가 되지 않아도, 조금만 머리를 굴려 고민해보세요!

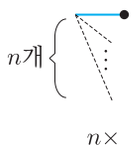
마치며

위 문제를 풀며 같은 문제도 완전히 다른 관점으로 풀이할 수 있으며, 어떻게 풀더라도 답은 동일하다는 것을 알 수 있습니다. 따라서 여러 관점으로 풀이하며 경험을 쌓아나가고, 그 중 자신에게 가장 잘 맞는 방법을 찾아보시기 바랍니다.

04 순열과 조합, 그리고 나눗셈과 곱셈의 이해

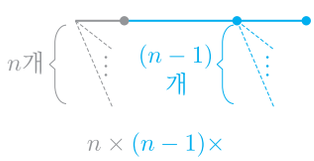
팩토리얼, 순열, 조합

팩토리얼 : n명을 일렬로 세우는 방법의 수는 $n!$이다.



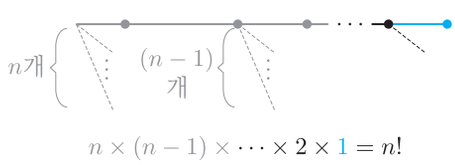
n명을 일렬로 세우려면 먼저 맨 앞인 첫 번째 자리에 설 사람을 정해야 합니다. n명 중 한 사람을 택해 첫 번째 자리에 세우는 방법의 수는 n입니다.[17] 이때 n명 중 누구를 맨 앞에 세우든 수형도의 모양이 같으므로, 이후 곱의 법칙을 적용할 수 있을 것입니다.

17) 이를 ${}_nC_1$로 택하여 1로 배열하는 ${}_nC_1 \times 1$이라 해석할 수도 있습니다.



이제 두 번째로 설 사람을 정해야 하고, 그 방법의 수는 $n-1$입니다.[18] 따라서 n명 중 두 명을 일렬로 세우는 방법의 수는 곱의 법칙에 의해 $n \times (n-1)$입니다.

18) 이를 ${}_{n-1}C_1$로 택하여 1로 배열하는 ${}_{n-1}C_1 \times 1$이라 해석할 수도 있습니다.



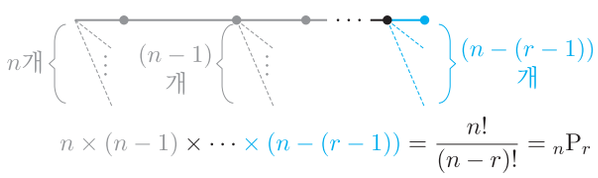
같은 방법으로 곱의 법칙을 반복적으로 적용하면 마지막 n번째 설 사람을 정해야 하고, 그 방법의 수는 1이므로 $n \times (n-1) \times \cdots \times 2 \times 1$입니다. 이처럼 n부터 1까지의 자연수의 곱은 경우의 수에서 자주 쓰이므로 $n!$이라 정의합니다.[19]

19) 6!까지는 자주 쓰이니 수를 외워두는 것도 좋습니다. 각각 다음과 같습니다.

$2! = 2,\ 3! = 6,\ 4! = 24$

$5! = 120,\qquad 6! = 720$

순열 : n명 중에서 r명을 택하여 일렬로 세우는 방법의 수는 ${}_nP_r$이다.



n명 중에서 r명을 택하여 일렬로 세우는 방법의 수 ${}_nP_r$은 $n!$의 수형도를 그리는 과정에서 r번째까지만 그리는 것과 같습니다. 따라서 ${}_nP_r = n \times (n-1) \times \cdots \times (n-r+1)$이고, 이를 팩토리얼로 더 멋있게 나타내면 ${}_nP_r = \dfrac{n!}{(n-r)!}$입니다.[20]

20) 그러나 우리는 이 공식을 되도록 사용하지 않을 것입니다. 그 이유는 이후 차차 설명하겠습니다.

조합 : n명 중에서 r명을 택하는 방법의 수는 ${}_n\mathrm{C}_r$이다.

n개 중에서 r개를 택하는 방법의 수는 ${}_n\mathrm{C}_r = \dfrac{n!}{r!\,(n-r)!}$ 입니다. 공식이 유도된 과정은 나중에 설명할 것이니, 일단은 공식을 잘 외워둡시다.[21] 실제 계산에서 조합을 어떻게 계산하는지 연습해볼 것입니다.

21) 빈칸 문제의 과정에서 저 공식을 정확히 암기해야 제대로 풀이할 수 있습니다.

조합 : ${}_n\mathrm{C}_r = {}_n\mathrm{C}_{n-r}$: 조합의 대칭성

A	B	C	D	E	F
O	X	O	O	X	O

A	B	C	D	E	F
O	X	O	O	X	O

${}_6\mathrm{C}_2$와 ${}_6\mathrm{C}_4$를 생각해봅시다. ${}_6\mathrm{C}_2$는 '6개 중에서 2개를 택하는 방법의 수'를 의미합니다. 이는 곧 6개 중에서 택하기 싫은 4개를 정하는 방법의 수와 동일한 의미를 갖습니다. 그런데 밑줄은 곧 ${}_6\mathrm{C}_4$를 의미합니다. 따라서 ${}_6\mathrm{C}_2 = {}_6\mathrm{C}_4$입니다.

이를 일반화한 것이 ${}_n\mathrm{C}_r = {}_n\mathrm{C}_{n-r}$입니다. 단순히 조합의 정의에 대입하여 계산하면 참임을 알 수 있고, 실제 상황에도 잘 부합함을 알 수 있습니다.[22]

22) 또한 이 내용은 곧 두 종류만 있을 때의 같있순을 해석하며 또 다시 다룰 것입니다.

조합의 공식과 대칭성을 이용한 조합 계산 방법

조합의 공식과 조합의 대칭성을 이용하면 ${}_{11}\mathrm{C}_9$와 ${}_{11}\mathrm{C}_2$에 대하여 다음과 같은 식을 얻을 수 있습니다.

$$ {}_{11}\mathrm{C}_9 = \frac{11!}{9!\,(11-9)!} = \frac{11!}{2!9!} = \frac{11!}{2!\,(11-2)!} = {}_{11}\mathrm{C}_2 $$

이 중 가장 가운데 식에서 눈여겨보아야 할 것은, 분자의 $11!$과 분모의 $9!$이 서로 약분되어 분자에는 11×10만 남고, 분자에는 $2! = 2 \times 1$만 남는다는 것입니다. 따라서 ${}_{11}\mathrm{C}_9 = {}_{11}\mathrm{C}_2 = \dfrac{11 \times 10}{2 \times 1} = 55$라 계산하면 됩니다.

예제 1. 다음 조합을 계산하시오.

① ${}_7\mathrm{C}_5$ ② ${}_8\mathrm{C}_5$ ③ ${}_9\mathrm{C}_6$ ④ ${}_{10}\mathrm{C}_8$

예제 1 풀이

$$① \ {}_7\mathrm{C}_5 = {}_7\mathrm{C}_2 = \frac{7\times 6}{2\times 1} = 21 \qquad ② \ {}_8\mathrm{C}_5 = {}_8\mathrm{C}_3 = \frac{8\times 7\times 6}{3\times 2\times 1} = 56$$

$$③ \ {}_9\mathrm{C}_6 = {}_9\mathrm{C}_3 = \frac{9\times 8\times 7}{3\times 2\times 1} = 84 \qquad ④ \ {}_{10}\mathrm{C}_8 = {}_{10}\mathrm{C}_2 = \frac{10\times 9}{2\times 1} = 45$$

순열($_n\mathrm{P}_r$)은 버린다. 그저 택($_n\mathrm{C}_r$)하고 배열($r!$)할 뿐이다.

순열의 정의와 조합의 정의를 다시 찬찬히 읽어봅시다.

> 서로 다른 n개에서 r $(0 \le r \le n)$개를 택하여 일렬로 나열하는 것을 'n개에서 r개를 택하는 순열'이라 하고, $_n\mathrm{P}_r$이라 나타냅니다.

> 서로 다른 n개에서 순서를 생각하지 않고 r $(0 \le r \le n)$개를 택하는 것을 'n개에서 r개를 택하는 조합'이라 하고, $_n\mathrm{C}_r$이라 나타냅니다.

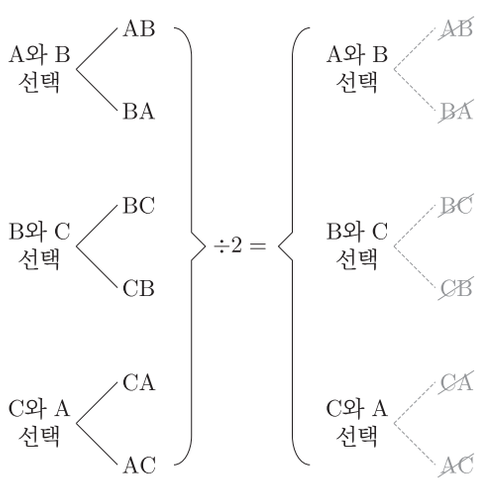
그런데 순열의 정의를 분석해보면, 두 가지의 행위가 잇달아 일어나고 있음을 알 수 있습니다. 바로 n개 중에서 r개를 택하는 행위와 그 r개를 일렬로 나열하는 행위입니다. 이 중 전자는 $_n\mathrm{C}_r$의 정의와 완전히 동일하고, 후자는 $r!$의 상황과 완전히 동일합니다. 즉 곱의 법칙에 의해 $_n\mathrm{P}_r = {}_n\mathrm{C}_r \times r! \cdots ①$라 말할 수 있습니다.

교과서에서는 순열을 먼저 배우고 조합을 나중에 배우므로 $_n\mathrm{C}_r = \frac{{}_n\mathrm{P}_r}{r!} \cdots ②$이라 배우지만, 사실 의미상으로 더 자연스러운 수식은 ②보다는 ①입니다.[23] 따라서 이 책에서는 보다 자연스러운 풀이를 위하여 순열을 사용하지 않을 것입니다.[24]

23) 우리가 이렇게 느끼는 이유는, 곱셈의 법칙은 배웠지만 나눗셈의 법칙은 배운 바가 없기 때문입니다.

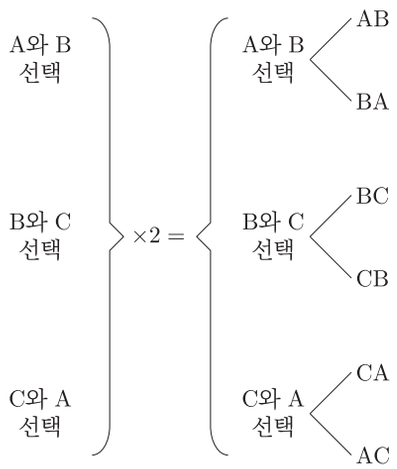
24) 대신 사칙연산, 거듭제곱, 조합, 팩토리얼만을 사용합니다.

$\div n$은 '다른 것 n개'를 '같은 것 1개'로 간주하는 것(축소)이다.



예를 들어, A, B, C 중에서 2개를 택하여 배열하는 경우는 AB, BA, BC, CB, AC, CA로 여섯 가지가 있습니다. 이를 '어떤 두 개를 택하였는가'를 기준으로 수형도로 나타내면 위와 같습니다. 이 상태에서 나눗셈 $\div 2$를 하는 것은, 수형도에서는 세 개의 마디 'A와 B 선택', 'B와 C 선택', 'C와 A 선택'에 각각 달린 두 개(2!개)의 가지를 모두 잘라냄으로써 무엇을 택하였는가만 세겠다는 것으로 생각할 수 있습니다.

$\times n$은 '같은 것 1개'를 '다른 것 n개'로 간주하는 것(증폭)이다.



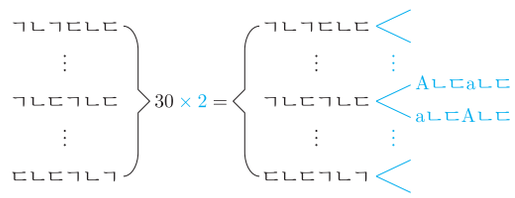
거꾸로 A, B, C 중에서 2개를 택하는 경우만 수형도로 나타내면 위와 같습니다. 이 상태에서 곱셈 $\times 2$를 하는 것은, 수형도의 각 마디에 두 개씩(2!개씩) 가지를 새로 그려 두 개를 택하고, 택한 두 문자를 어떻게 배열하는가까지 고려하여 세는 것으로 생각할 수 있습니다.

이와 같이 순열과 조합의 관계를 통해 ${}_n\mathrm{P}_r$에 $\div r!$함으로써 $r!$배만큼 수형도가 축소되어 ${}_n\mathrm{C}_r$이 되고, ${}_n\mathrm{C}_r$에 $\times r!$함으로써 $r!$배만큼 수형도가 증폭되어 ${}_n\mathrm{P}_r$이 됨을 알 수 있었습니다. 배운 내용을 바탕으로, 앞서 두 번이나 풀이한 문제를 살짝 변형한 다음 문제를 풀어봅시다. 참고로, 원래 문제는 정답이 30이었습니다.

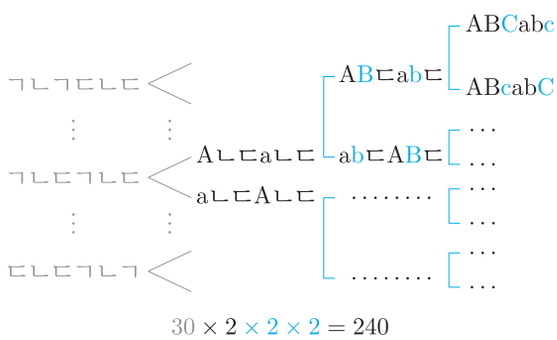
예제 2. 여섯 개의 알파벳 A, a, B, b, C, c를 일렬로 나열하여 문자열을 만든다. A와 a는 서로 이웃하지 않고, B와 b는 서로 이웃하지 않고, C와 c는 서로 이웃하지 않도록 배열된 문자열의 개수를 구하시오.

예제 2 풀이

> 여섯 개의 알파벳 A, a, B, b, C, c를 일렬로 나열하여 문자열을 만든다. A와 a는 서로 이웃하지 않고, B와 b는 서로 이웃하지 않고, C와 c는 서로 이웃하지 않도록 배열된 문자열의 개수를 구하시오.



A, a, ㄴ, ㄴ, ㄷ, ㄷ을 일렬로 나열한다면, 하나의 경우의 수가 두 가지의 경우의 수로 증폭됩니다. 예를 들어, 하나의 경우 ㄱㄴㄷㄱㄴㄷ는 두 가지의 경우 Aㄴㄷaㄴㄷ, aㄴㄷAㄴㄷ로 증폭됩니다. 따라서 ㄱㄱ 대신 Aa로 바뀐 상황에서는 원래 수형도의 끝이 모두 새로운 마디가 되어 각자 2개의 가지를 새로 뻗어냄을 알 수 있습니다. 그러므로 A, a, ㄴ, ㄴ, ㄷ, ㄷ를 일렬로 나열하는 경우의 수는 $30 \times 2 = 60$입니다.



같은 논리로 ㄴㄴ과 ㄷㄷ를 각각 Bb와 Cc로 대체하면 역시 각각 두 개의 가지를 새로 뻗어나가게 됩니다. 따라서 구하는 경우의 수는 $30 \times 2 \times 2 \times 2 = 240$입니다.

이와 같은 곱셈과 나눗셈의 원칙을 정확히 이해해야만 고난도 경우의 수 문제를 잘 풀이할 수 있고, 향후 확률에서 오개념을 피할 수 있습니다.[25)]

25) '같있순을 확률에 쓸 수 있느냐 없느냐'와 같은 논쟁은 확률에 대한 무지와 더불어 수형도에서의 곱셈과 나눗셈이 의미하는 바를 이해하지 못한 데에서 비롯된 것입니다.

11 통계 용어 정리 (1) : 확률변수, E, V, σ

확률변수

표본공간 S의 각 원소를 실수 전체의 집합 R의 한 원소에 대응시키는 함수 X를 **확률변수**라고 합니다. 확률변수 X가 어떤 값 x를 가질 확률을 $\mathrm{P}(X=x)$라 표기하고, X가 가지는 값과 X가 이 값을 가질 확률의 대응 관계를 X의 **확률분포**라 합니다.

이산확률변수

확률변수 X가 가지는 값이 유한개이거나 자연수와 같이 셀 수 있을 때[56] 그 확률변수 X를 **이산확률변수**라고 합니다.

확률분포와 확률질량함수

이산확률변수 X가 가지는 값 $x_i\ (i=1,\ 2,\ 3,\ \cdots,\ n)$와 X가 x_i를 가질 확률 p_i의 대응 관계인 다음의 식이 이산확률변수 X의 확률분포입니다.

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

이때 이 대응 관계를 나타내는 함수를 **확률질량함수**라 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[57] 다음이 성립합니다.

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

② $\displaystyle\sum_{i=1}^{n} p_i=1$

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

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

⑤ $\sigma(X)=\sqrt{\mathrm{V}(X)}$

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

57) 평균, 분산, 표준편차의 정의와 그에 대한 설명은 다음 챕터에 나옵니다. 일단은 공식만 눈에 발라놓도록 합시다.

한편 $\sum$의 성질을 이용하면 확률변수 X, 0이 아닌 상수 a, 상수 b에 대하여 다음이 성립함을 알 수 있습니다.[58]

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

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

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

④ $\sigma(aX+b) = |a|\,\sigma(X)$

58) 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,[59] n번의 독립시행에서 사건 A가 일어나는 횟수를 확률변수 X라 하면, 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 약속하겠습니다.

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

59) 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 1 이상 4 이하의 모든 실수의 값을 가질 수 있을 때, X는 연속확률변수입니다.

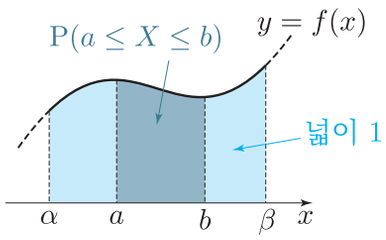
연속확률변수와 이산확률변수의 공통점과 차이점

앞서 언급한 예시에서, 확률변수 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 $\mathrm{P}\left(\sqrt{2} \le X \le \pi\right)$의 값은 논할 수 있습니다.[60]

60) 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 $y = f(x)$가 다음 조건을 모두 만족시킬 때, 함수 f를 X의 **확률밀도함수**라 합니다.

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

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

③ 확률 $\mathrm{P}(a \le X \le b)$는 함수 f의 그래프와 x축 및 두 직선 $x = a$, $x = b$로 둘러싸인 부분의 넓이와 같다.[61)]

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

61) 당연하지만 $\alpha \le a \le b \le \beta$임을 전제해야 이를 만족시킬 수 있습니다.

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

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 $\displaystyle\mathrm{P}(X = a) = \int_{a}^{a} f(x)\,dx$입니다. 이는 정적분의 성질에 의해 0입니다.

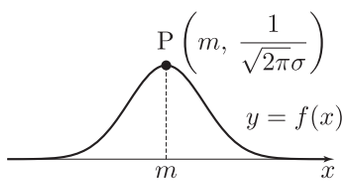
정규분포

실수 전체의 집합에서 정의된 연속확률변수 X의 확률밀도함수 f가 상수 m, 양수 σ와 무리수인 상수 $e = 2.718281\cdots$에 대하여 다음과 같을 때, X의 확률분포를 **정규분포**라고 합니다.[62)]

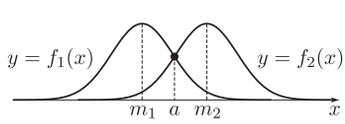
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

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.

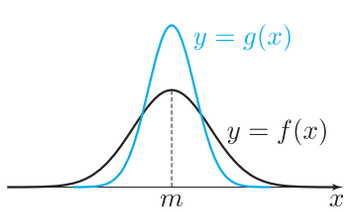
$X \sim \mathrm{N}(m,\ \sigma^2)$일 때 X의 확률밀도함수 $y = f(x)$의 성질

① 

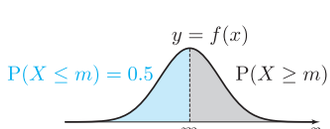
직선 $x = m$에 대하여 대칭인 종 모양의 곡선이며, x축을 점근선으로 하고, $x = m$일 때 최댓값을 갖습니다.[64)]

② 

σ의 값이 일정할 때 m의 값이 달라지면, 대칭축의 위치만 바뀌고 곡선의 모양은 같습니다.[66)]

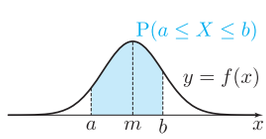
③ 

m의 값이 일정할 때 σ의 값이 달라지면, σ의 값이 커질수록 최댓값이 작아지면서 넓게 퍼지고, σ의 값이 작아질수록 최댓값이 커지면서 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 $\mathrm{P}(a \le X \le b) = \displaystyle\int_a^b f(x)\,dx$입니다.



표준정규분포와 표준화

$Z \sim \mathrm{N}(0,\ 1)$인 확률변수 Z의 확률분포를 **표준정규분포**라 합니다. 한편, $X \sim \mathrm{N}(m,\ \sigma^2)$인 확률변수 X에 대하여 $Z = \dfrac{X-m}{\sigma}$가 성립합니다. 이제 이렇게 확률변수 X를 Z로 변환하는 과정을 표준화라고 부르기로 합시다.

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

64) 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.

66) 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?

12 통계 기초, 이산확률변수, 연속확률변수

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.[67)]

67) 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 성적의 평균을 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
$\mathrm{P}(Y=y)$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{2}{5}$	1

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

Z	80	합
$\mathrm{P}(Z=z)$	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.[68] 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.

분산과 표준편차

평균은 많은 정보를 알려주지만, 모든 정보를 알려주지는 못한다

지금까지 살펴본 세 명의 성적의 평균은 80점으로 동일합니다. 그러나 여러분도 아시다시피, 분명히 세 명의 평균이 같기는 하지만, 세 명의 특성이 동일하다고 말하기에는 뭔가 망설여집니다.

이는 평균이라는 도구가 분명히 무언가 큰 의미[69]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.[70] 따라서 평균이 드러내지 못하는 '보이지 않는 무언가', 즉 각 값들이 평균으로부터 얼마나 떨어져 있는가를 수치화할 수 있는 도구가 필요합니다. 그것이 바로 분산과 표준편차입니다.

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.

68) 이 영단어의 맨 앞글자를 따 평균을 표기하는 것입니다.

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

70) 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 분포**의 의미를 '여러 가지 점수가 골고루 나온다'고 착각하기 쉬운데, 고르게 분포되었다는 것은 각각의 값들이 고만고만하게 비슷비슷하다는 의미입니다.

	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 준영이든 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[71] 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[72]를 구할 수 있는 대안이 절실하게 필요합니다.

71) 다음과 같이 계산됩니다.

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

72) 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 나오지 않도록 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 $\mathrm{E}\left(X^2\right)$, $\mathrm{E}\left(Y^2\right)$, $\mathrm{E}\left(Z^2\right)$

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

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

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

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

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

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

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

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

C	0	합
$\mathrm{P}(C=c)$	$\times 1$	1

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

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

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

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

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

Z^2	80^2	합
$\mathrm{P}(Z^2=z^2)$	$\times 1$	1

$\mathrm{E}(Z^2)=6400$

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

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

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

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

이와 같이 편차²의 평균, 다시 말해 $\mathrm{E}\left(\text{편차}^2\right)$을 **분산**이라 합니다. 분산을 계산할 때에는 정의에 따라 정직하게 구하기보다는, 위 수식에서 $\cdots$로 생략된 유도과정[73]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.

73) 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.[74)]

74) 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 : 특수한 이산확률분포, 표도 그리지 않고, 증명도 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	0	1	2	$\cdots$	$n-1$	n	합
$\mathrm{P}(X=x)$	${}_n\mathrm{C}_0q^n$	${}_n\mathrm{C}_1pq^{n-1}$	${}_n\mathrm{C}_2p^2q^{n-2}$	$\cdots$	${}_n\mathrm{C}_{n-1}p^{n-1}q$	${}_n\mathrm{C}_np^n$	1

비록 이항분포의 확률분포를 표로 잘 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[75)] 이항분포로 번거롭게 계산할 필요 없이 정규분포를 이용하여 쉽게 계산할 수 있음이 알려져 있습니다. 이미 알려져 있으니 우리는 증명할 필요 없이 잘 쓰기만 하면 됩니다.

75) $np\geq5$, $nq\geq5$인 경우

일반적인 연속확률변수

연속확률변수는 단 한 페이지만으로 수능에 필요한 모든 내용을 끝낼 수 있습니다. 일반적인 연속확률변수는 확률밀도함수의 구간을 정적분한 값이 확률값이고, 정의역 전체 구간을 정적분한 값은 1이라는 사실만 알면 됩니다. 현 교육과정 내에서 일반적인 연속확률변수에 대하여 물어볼 수 있는 것은 이게 끝입니다.[76)]

76) 연속확률변수의 평균, 분산, 표준편차를 구해야 하던 시절이 있었습니다만, 이제는 알 필요가 없습니다.

정규분포 : 정적분도 하지 않고, 증명도 없이 꿀빨자

정규분포는 십중팔구 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 기본 개념들을 집요하게 물어볼 수 있습니다.

중학수학, 고등수학, 〈수학 I〉, 〈수학 II〉와 연계될 수 있다

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

13 통계 용어 정리 (2) : 통계적 추정

통계적 추정의 궁극적 목적은 표본을 한 번만 추출해서 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.[77)]

77) 교과서는 이 충분히 큰 n의 기준을 밝히지 않고 있습니다.

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.[78)]

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

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

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, 모표준편차가 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.

모집단이 정규분포를 따르는 경우

모집단이 정규분포 $\mathrm{N}(m,\ \sigma^2)$을 따를 때, 확률변수인 표본평균 $\overline{X}$는 정규분포 $\mathrm{N}\left(m,\ \dfrac{\sigma^2}{n}\right)$를 따릅니다.[79)]

모집단이 정규분포를 따르지 않지만, 표본의 크기가 충분히 큰 경우

모집단의 분포가 정규분포를 따르지 않을 때에도, n이 충분히 크면 확률변수인 표본평균 $\overline{X}$는 근사적으로 정규분포 $\mathrm{N}\left(m,\ \dfrac{\sigma^2}{n}\right)$를 따릅니다.[80)]

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 값 $\overline{x}$를 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.[81)]

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[82)] 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 : $\left[\overline{x}-2.58\dfrac{s}{\sqrt{n}},\ \overline{x}+2.58\dfrac{s}{\sqrt{n}}\right]$

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

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

81) 원래는 100개의 표본평균으로 만든 100개의 신뢰구간 중에서 약 k개가 모평균을 포함한다(표본의 크기가 같은 표본을 여러 번 추출하여 신뢰구간을 만들 때, 이 중 $k\%$가 모평균을 포함할 것으로 기대된다). 라고 말하는 것이 정확합니다. 그러나 수학과나 통계학과에 진학할 것이 아니라면 그냥 본문과 같이 대충 이해하셔도 무방합니다.

82) $n \geq 30$일 때

14 통계적 추정, 딱 필요한 만큼만

솔직히 말해서, 통계적 추정 문제의 대부분은 앞 단원의 '용어 정리'만 외워도 다 풀립니다. 그런데 학생들이 워낙 공식만으로 문제를 풀다보니 개념에 대해 제대로 숙지하지 못하는 부분이 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.[83] 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.[84] 그런데 표본의 크기가 100가 된다면 확률변수 $\overline{X}$가 160 이하일 확률은 급격하게 작아집니다.[85]

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 표본의 크기 n으로 나눈 값인 $\frac{\sigma^2}{n}$이 되고, 표준편차 $\sigma(\overline{X})$는 모표준편차 σ를 $\sqrt{n}$으로 나눈 값인 $\frac{\sigma}{\sqrt{n}}$가 됩니다.[86]

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

84) 0.0013으로, 10000명 중 13명 꼴입니다.

85) 평균으로부터 먼 값들이 나올 확률은 상대적으로 매우 작고, 평균으로부터 가까운 값이 나올 확률은 상대적으로 매우 크기 때문입니다.

86) 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! 표본의 크기가 n으로 동일할 때, 표본표준편차는 크기가 n인 표본을 뽑을 때마다 매번 달라질 수 있는 값이지만, 확률변수 $\overline{X}$의 표준편차인 $\sigma\left(\overline{X}\right)$는 상수입니다. 모표준편차 σ와 표본의 크기 n이 모두 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.

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

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

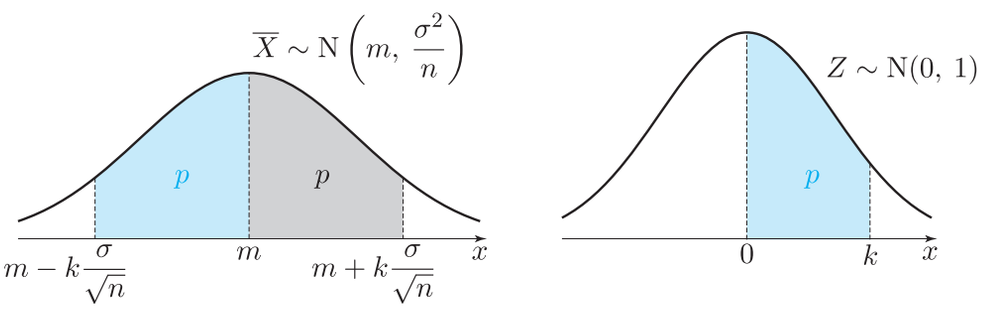
$$s_1 = \sqrt{{s_1}^2} = 7$$

이를 계산해보면 매번 표본을 추출할 때마다 다른 값이 나올 수도 있음을 알 수 있습니다.[87] 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.

87) 예를 들어, 임의추출한 표본이 174, 177, 174일 때, $\overline{X} = 175$, ${s_2}^2 = 3$, $s_2 = \sqrt{3}$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 이용해 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, $\mathrm{P}\left(0 \le Z \le k\right) = p$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. $\overline{x}$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 $\overline{x}$와 m의 위치가 서로 바뀌어야 합니다. 그러면 어떻게 하면 우리가 원하는 바를 얻을 수 있을까요?

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

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

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

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

딱 한 번 구한 $\overline{x}$로 만든 어떤 (신뢰)구간 $\left[\overline{x}-k\frac{\sigma}{\sqrt{n}},\ \overline{x}+k\frac{\sigma}{\sqrt{n}}\right]$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 $\overline{x}$, n, σ를 알아야 합니다. 이때 표본의 크기인 n과 표본평균인 $\overline{x}$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.[88] 여러분이 조심해야 할 것은 2)에서 말했다시피 표본표준편차 s로 대체하는 것이지, 절대로 $\sigma\left(\overline{X}\right)$로 대체하는 게 아니라는 점을 명심하고 헷갈리지 않는 것입니다.

88) 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.