# 촌철살인

# 통계만

## 제 2 권 확률과 통계 4쇄

저자 : 최준호



티칭과 코칭

# 촌철살인 확률과 통계

# 수학자는 수학을 고민하고, 교육자는 철학을 고민한다.
# 이것은 완전히 다른 고민이다.

세상에는 너무 많은 수학이 있다. 한 분야의 수학을 정복하는 것도 불가능하다. 평생을 한 문제와 싸우는 교수님도 계신다. 〈많은 내용〉이야 〈구글 검색〉만 하더라도 앞의 교수님이 가진 지식보다 훨씬 더 많은 지식을 찾을 수 있다.

하지만 앞의 교수님이 가진 〈생각의 깊이〉와 넓고 얕은 지식을 비교할 수는 없는 일이다.

수학교육자가 알아야 하는 수학은 생각보다 많지 않다. 하지만 이 많지 않은 내용을 깊이 고민해야 한다. 교육자는 배우는 이유와 목적, 그리고 방향을 제시하고 학습자가 올바로 학습할 수 있도록 설득해 줄 수 있는 철학을 연구해야 한다.

이 책은 그런 생각을 담고 있다. 즉, 수학자로서의 역할을 기대하지 말고, 수학교육자로서의 역할을 기대하며 이 책을 읽는다면 좋은 책이 될 것이다. 어려운 대학수학들도 잠깐잠깐 다루긴 하지만 그것은 온전히 고등수학을 더 제대로 이해하기 위해서 보조 도구로 다루는 것뿐이다. 이 책은 〈온전한 이해〉만이 가장 빠른 길이라고 믿고 있다.

## 장님 코끼리 만지기?

이 책을 학습할 때 주의사항이 있다. 적어도 이 책으로 공부하기로 마음을 먹었다면 〈장님 코끼리 만지기식〉은 하지 않았으면 한다. 이 책은 확률의 시작인 〈시행과 사건〉부터 시중에 나온 어떤 책도 다루지 않은 부분에 대해서 다루고 있고, 이것을 논리적으로 엄밀하게 연결시켜 〈확률의 연산과 통계〉 끝까지 연결되는 개념을 수록하였다.

한 페이지도 〈장인정신〉없이 쓰지 않았다. 공들여 쓴 몇 페이지의 내용을 고민 끝에 삭제하고 한 두 줄로 요약한 경우도 종종 있었다. 깊은 이해를 위해 필요할 수도 있지만 그 이해가 반드시 필요한 것이 아닌 '알면 좋은 것'이라면 과감하게 삭제하였다.

앞에서부터 제대로 학습한다면 생각보다 훨씬 많은 시간이 절약될 것이다.

## 촌철살인을 읽는 방법

- 제목을 집중하여 읽는다.

: 제목은 본문의 내용을 함축한다. 한 단원을 읽기 전에 제목을 먼저 쭉 읽고 머릿속으로 왜 그렇게 제목을 정했는지 고민하면서 〈제목으로 먼저 뼈대〉를 잡는다.

- 빠르게 읽지 않는다.

: 한 부분 한 부분 모두 고민하여 썼기 때문에 한 부분도 놓치면 안 된다. 특히 문제를 풀 수 있다고 대충 넘어가면 안 되고 해설과 〈티칭〉 그리고 〈코칭〉을 빠짐없이 읽는다.

〈공부는 예제를 가지고 하는 거야.〉라는 부분은 〈뼈대 개념〉과 〈개념의 외연〉을 제대로 학습한 후 풀어본다.

단, 풀고 나서 답만 맞추고 맞았다고 넘어가는 것이 아니라 〈강의를 듣는 심정으로〉 해설을 정독한다.

- 독학재수하는 학생들에게

: 프랑스에는 걷기 프로그램인 랑도네 프로그램이라는 게 있습니다. 최근 발간된 〈1만 시간의 재발견〉이라는 책 제 3장에 〈의욕보다 중요한 연습의 방법〉부분에서 신체활동이 정신과 연결되어 있음과 그 중요성에 대해 서술하고 있습니다. 이 부분도 정독해 볼만합니다. 〈유대인 도서관〉도 참조해 볼만합니다. 〈매우 저명한 수학자, 과학자〉들이 〈움직이며〉 어떤 내용을 공부하고 고민하는 것은 아주 쉽게 접할 수 있습니다.

이 책은 〈런닝머신〉을 걸으면서 공부할 수 있는 유일한 수학책입니다. 수험생활의 성공에 가장 중요한 것은 〈움직이는 것〉입니다. 단지 건강을 위해서가 아니라 움직이며 공부하면 훨씬 좋은 아이디어를 많이 떠올릴 수 있습니다. 꼭 런닝머신을 하며 이 책을 완독하고 (물론 책상에서도 추가공부를 해야겠지요.) 후기를 남기는 학생들이 있었으면 좋겠습니다.

# 5회독으로 만드는 사고의 틀 : 반복의 방법

## 1회독 : 정독 (선생님의 설명이 있을 시 약 한 달 - 하루 90분 기준)

- 한 글자도 놓치지 않고 의미를 생각해가며 꼼꼼히 읽는다. 정독의 습관이 안 된 친구들은 선생님의 도움을 얻어 수업형태로 진행해 보는 것을 추천한다. 아마 처음 읽어서는 정독을 하더라도 놓치는 부분이 많을 것이다.

## 2회독 : 문풀독 (약 20일 - 하루 90분 기준)

- 〈이해를 위한 예제〉는 저자의 설명의 의도가 담긴 문제가 많기 때문에 설명을 이해하기 위해 〈풀기보다는 읽어보는 것〉이다. 하지만 그 의도를 명확하게 이해했다면 2회독 할 때는 〈이해를 위한 예제〉를 풀 수 있을 것이라고 생각한다. 〈~~을 보여라.〉등의 서술식 예제도 본인이 한 번 설명해보고 답지를 봤을 때 논지전개방식이 일치했다면 이 책의 사고방식을 어느 정도 받아들였다고 볼 수 있다. 이것이 안 된다면 다시 한 번 그 부분을 정독하여 이해하면 된다.

## 3회독 : 정리독 (약 20일 - 하루 90분 기준)

- 처음 공부하면서 바로 정리하는 것은 위험하다. 어느 정도 책 전체의 구조가 머릿속에 들어왔을 때 책의 구조도를 그리면서 정리해 본다. 일명 마인드맵이라고 불리는 이 방법은 일단 한 번 정리해 두면 까먹었을 때 다시 복습하기 매우 유용하다.

- 구조도를 그리면서 일명 〈keyword〉를 선별하게 되는데 이 단어를 통해서 이 책의 많은 부분을 연상할 수 있게 된다. 이렇게 단어를 선별하는 과정에서 〈글의 요지〉를 더 정확하게 이해하게 되며 만들고 수정하는 과정에서 〈1회독, 2회독〉에서도 놓친 부분을 다시 한 번 점검하게 된다. 검사해 줄 수 있는 선생님이나 친구가 있으면 더욱 좋다.

## 4회독 : 예측독 (약 7일 - 하루 90분 기준)

- 위에 까지만 해도 충분히 잘했다. 〈4회독〉과 〈5회독〉은 절대로 오래 걸리지 않으므로 주기적으로 반복하길 바란다. 〈예측독〉이란 말 그대로 서문을 읽으면서 글의 내용과 주제가 예측되어지는 상황이다. 한 번 더 꼼꼼히 점검하되 본인의 생각과 일치함을 확인하고 넘어가면 된다.

## 5회독 + : 스킵독 (약 20분)

- 최종 단계이다. 〈예측독〉까지 해서 어느 정도 실력의 확신이 들 때, 〈주 1회 or 격주 1회〉정도 20분 정도 투자하여 전체를 훑어보는 것이다. 여기까지 오면 어떤 〈확률과 통계〉의 문제를 만나더라도 접근에서 막히는 일은 없을 것이다. 그리고 감에 의존해서 문제를 푸는 느낌도 없을 것이라 확신한다.

위의 시간은 개인의 집중력과 역량 차이에 따라 변동 될 수 있다.
본인의 실력에 따라 기간을 조금 더 길게 혹은 조금 더 짧게 설정하여 공부하길 바란다.

티칭과 코칭

: 단 한 권의 책도 단 한 페이지도 장인정신 없이는 책을 쓰지 않습니다.
세계적 수학교육 출판기업이 될 때까지 끊임없이 진보합니다.

# 촌철살인 : 확률과 통계

<table>
<tr><th colspan="4">확률과 통계</th></tr>
<tr><td rowspan="12">#2.<br>확<br>률</td><td rowspan="4">Part 1. 확률의<br>기본 이론</td><td>수학적 확률 - 직관에 의존한 논리</td><td></td></tr>
<tr><td>통계적 확률</td><td></td></tr>
<tr><td>기하학적 확률</td><td></td></tr>
<tr><td>확률의 덧셈 정리 (확률 덧셈의 정당성 부여)</td><td></td></tr>
<tr><td rowspan="5">Part 2. 확률의<br>계산 1</td><td>조건부 확률 - 직관에 의존한 논리</td><td></td></tr>
<tr><td>시행의 독립과 종속 VS 사건의 독립과 종속</td><td></td></tr>
<tr><td>확률의 곱셈정리 (확률 곱셈의 정당성 부여)</td><td></td></tr>
<tr><td>걍 풀 조건부와 계산적 조건부</td><td></td></tr>
<tr><td>결국 효율성의 문제</td><td></td></tr>
<tr><td rowspan="3">Part 3. 확률의<br>계산 2</td><td>독립시행의 확률</td><td></td></tr>
<tr><td>확률부분과 경우의 수 부분</td><td></td></tr>
<tr><td>종속시행의 확률?</td><td></td></tr>
<tr><td rowspan="7">#3.<br>통<br>계</td><td rowspan="2">Part 1.<br>이산확률분포</td><td>이산확률분포</td><td></td></tr>
<tr><td>이항분포</td><td></td></tr>
<tr><td rowspan="3">Part 2.<br>연속확률분포</td><td>연속확률분포</td><td></td></tr>
<tr><td>정규분포</td><td></td></tr>
<tr><td>이항분포와 정규분포의 관계</td><td></td></tr>
<tr><td rowspan="2">Part 3. 통계적<br>추정</td><td>모표균 추정</td><td></td></tr>
<tr><td>모비율 추정</td><td></td></tr>
</table>

이 책이 두꺼워진 이유는 다른 책이 두꺼워진 이유와는 조금 다릅니다.
아는 것을 정리한 것이 아닌 알아야 하는 것을 정리하였습니다.
엄청나게 많은 내용을 넣어놔서 두꺼워진 것이 아니라 알아야 하는 것을 최대한
자세하게 설명해서 두꺼워졌습니다.

알면 좋은 것들을 넣어둔 것이 아닙니다.
반드시 알아야 하는 것을 넣어놨습니다.
이 책은 사전이 아닙니다.

## 아는 만큼 보인다.

이 책의 내용 중 모르는 것이 있다면 반드시 학습해야 합니다.
만약 많이 보지 못했다고 생각되는 내용이 있다면 본인이 많이 하지 않았거나 아니면
알지 못해서 지금까지 수없이 봤음에도 불구하고 인지 못하고 넘어갔음이 분명합니다.
이 책은 공식이 아닌 공식에 담긴 사고방식을 설명하는 책입니다.

꽃 - 김춘수 -

내가 그의 이름을 불러주기 전에는
그는 다만
하나의 몸짓에 지나지 않았다.

내가 그의 이름을 불러주었을 때
그는 나에게로 와서
꽃이 되었다.

내가 그의 이름을 불러준 것처럼
나의 이 빛깔과 향기에 알맞은
누가 나의 이름을 불러다오
그에게로 가서 나도
그의 꽃이 되고 싶다.

우리들은 모두
무엇이 되고 싶다.
나는 너에게 너는 나에게
잊혀지지 않는 하나의 의미가 되고 싶다.

잊혀지지 않는 의미가 되기 위해서 어떤 개념에는 반드시 이름이 있어야 한다.
이름이 없으면 잠깐 스쳐가는 하나의 몸짓에 지나지 않을 것이다.
비단 한국의 수학만 그런 것은 아니라고 생각한다.
공식에는 이름이 있지만 그 공식이 담고 있는 사고방식에는 이름이 없기 때문에 결국 우리는 머릿속에 공식만 남는다.
수학이 실제 생활에서 필요한 이유는 〈공식〉을 쓰기 때문이 아니다.
〈그 공식이 담고 있는 사고방식〉이 우리가 어떤 정보를 받아들이고 판단하는 데 중요한 역할을 하기 때문이다.
앞으로 배우게 되는 경우의 수에서 전체를 생각하여 부분으로 가는 사고방식이나
기준을 설정해서 상황을 고정하고 변하는 것들만 관찰하는 사고방식은 비단 수학에서만 필요한 사고방식이 아니다.
과학에서도 원인과 결과를 정확히 관찰하기 위해 관찰의 초점인 〈조작변인〉을 제외하고 전부 통제한다.
우리가 실생활에서 기계가 작동이 되지 않을 때 무조건 서비스 센터를 찾아가는 것은 개인에게도 기업에게도 굉장히 비효율적이다.
이때 우리는 몇 가지 실험을 거쳐 안 되는 이유를 찾게 되는데 이런 작은 행동들에도 너의 사고방식이 영향을 준다.
수학은 이런 사고방식을 아주 단순화해서 문제를 통해 반복적으로 연습하는 것이다.
사고방식도 마치 근육처럼 단련하고 연습해야 잘 사용할 수 있기 때문이다.
이것이 바로 사고력(力)에 〈힘 력〉자를 쓰는 이유일 것이다.

공식은 단지 그 사고방식을 사용하는 과정에서 계산을 조금 줄여준다거나 조금 빠르게 해줄 뿐이다.

# 촌철살인 #0. 이 책의 학습방향

## 수학은 직관에 의존한 논리이다.

- 직관이란 〈경험을 통한 예측〉이다. 즉, 경험 없이는 아무런 직관도 생기지 않는다. 우리가 수많은 문제를 풀어보면서 직접 연습하는 이유도 바로 이러한 직관력을 기르기 위함이다. 필자는 〈수학〉에 있어서 정말 중요한 〈직관〉이라는 부분이 적절하고 체계적으로 설명되어 있는 학습서를 보지 못했기 때문에 적절한 비유를 통해서 그것을 설명하고자 한다.
  (저명한 물리학자이자 하버드대 교수인 'Gerald Holton'이 "은유는 미지의 세계로 가는 유일한 다리"라고 언급한 것처럼, 적절한 비유는 난해한 개념을 확실히 이해시켜 주는 것에 있어서 매우 중요하다.)
  또한 이런 직관에 해당하는 개념들은 보통 〈구체적 용어화〉가 되어있지 않기 때문에 〈인지〉하는 것 또한 어렵다.
  그래서 그냥 〈헷갈린다, 어렵다, 실수했다〉 등으로 생각하기 마련인데 이러한 부분은 구체적으로 〈용어화〉시킴으로서 이 책을 공부하고 난 후에는 왜 틀렸고 무엇에 의해 풀어야 하는지 정확하게 말할 수 있을 것이다.

## 생각의 방향과 바른 고정관념

- 모든 것이 그렇듯 완벽한 이론이란 없다. 어차피 완벽에 가까운 이론일수록 너무 추상적이어서 쓸모가 없다.
  그러므로 우리는 〈생각의 방향〉을 정해야 한다. 그 〈생각의 방향〉이 모든 문제를 풀 수 없음은 당연하다.
  수학에서 〈고정관념〉을 버려야 한다고 하지만 이것은 공허한 구호에 불과하다. 오히려 수학은 〈올바른 고정관념〉을 가져야 하고 그런 〈고정관념〉으로 풀리지 않는 도전적인 문제를 통해서 〈개념의 외연〉을 넓혀가야 한다.
  그래서 10문제 중 9문제에 적용할 수 있는 〈생각〉이라면 개념화시켜 학습할 충분한 가치를 가지고 있다.

- 이 책은 완벽하지 않다. 완벽을 추구하는 책이 아니다. 올바른 고정관념과 직관을 기르도록 하는 것이 이 책의 목표이다.
  추상적이지만 근본적인 이론들을 적절한 비유를 통해 제시하고 구체적인 언어로 만들어 학습할 수 있도록 이 책을 구성하였다.

1권을 집필한 후... 누군가의 리뷰

다음 책 리뷰 - <제목 : 확률통계 교과과정의 출발 경우의 수에 대한 고민>

제목에 끌려 선택한 책입니다.
현장에서 쌓은 실전의 지식으로 깊은 사유 끝에 만들어진 수학책입니다.
단순하게 확률과 통계 부분만 정리하는 문제집이나 개념서가 아니며
"풀이"를 넘어서 정독하게 하는 수학책의 의미가 있습니다.
확률에 대해 고민하는 경험 있는 강사나 교사들에게
자신의 공부 방법이 풀이보다는 근본적인 개념을 파고드는 학생이나
다른 풀이에 대해 관심 있는 분들에게 권합니다.
(라엘과생명나무 님)

네이버 책 리뷰 - <제목 : 좋아요 진짜로>

책이 너무 좋네요 저는 현재 고2 자연계학생인데요 처음 공부할때 헷갈리거나 잘못생각되기 쉬운 부분들을 정확하게 집어주면서 충분한 예제들과 친절한 해설들이 개념들이 머리속에 박히도록 해줘서 너무 좋아요 예를들면 책에 나와있는 분할적 사고는 보통 다른책에서는 분명하게 집어주지 않아서 익숙해지는데 시간이 오래 걸리는데 자세하게 써져있어서 시행착오하면서 시간을 낭비하는걸 많이줄여주는게 정말 좋습니다. 처음에 확통 공부할때 개인적으로 경우의 수를 논리적으로 세는것이 많이 헷갈려서 몸에 익힐때 까지 시간이 정말 오래걸렸었는데 이책은 그부분을 용어도 붙여서 자세히 알려줘서 인상적이었어요 또 가르쳐만 주고 마는것이 아니라 풍부한 예제들로 개념을 몸에 익혀주는것도 정말 좋았어요. 처음 공부할때 이 책으로 공부했다면 얼마나 좋았을까 하는 생각이 가득합니다ㅠㅠ 그렇지만 물론 한번 공부했었다고 해도 얻어갈게 적다는건 절대 아닙니다ㅎㅎ 책 구성에대한 이야기도 하고 싶습니다. 저는 내용도 마음에 들지만 구성도 정말 마음에 들었거든요 정석이나 개념원리에 길들여진 저는 처음 책을 펼쳤을때는 약간 어색한 느낌도 없잖아 있었습니다. 하지만 ㅎㅎ 익숙해지는데 걸린 시간은 정말 짧았습니다ㅎㅎ 처음에 개념설명과 공식 증명부분으로 뼈대를 쌓고 충분히 많은 예제들로 몸에 익숙하게 해주는데 예제 하나하나가 자세하게 설명해주고 있어서 너무 좋습니다. 또 너무 오바해서 지저분하게 설명하거나 하는것도 없이 필요한 부분들을 이해하기 쉽게 설명하고 있다는 점도 좋았습니다. 개념의 외연 부분도 단원을 이해하는데 큰도움이 되네요 또 마지막에 기출문제들로 다시 싹 정리해줘서 그것도 좋았습니다. 마지막으로 분량에 대한 이야기를 하자면요 저도 사실 처음에는 너무 많은 거 아닌가 하는 생각이 들었습니다. 하지만 직접 풀어 보니 결코 아니더라구요 ㅎ 다른책에서는 조금 직관적으로 이해하고 넘어가야 해서 이해하는데 오래 걸릴수 있는 부분을 용어를 붙여서 친절하게 설명하다보니 길어질수 밖에 없었던거 같습니다. 또 예제들이 정말 풍부하고 하나하나의 예제들이 정말 친절해서 분량이 길어진거였습니다. 책을 직접 풀어 보시면 아시겠지만 집중해서 책을 따라가다보면 시간가는줄을 모르고 꽤 많은 분량을 해내게 합니다. 친절하고 세세한 설명으로 막히는 부분이 없어서 가능했습니다. 그래서 오히려 분량이 많아 보이지만 시간적인 측면에서 훨씬 크게 절약됩니다. 시간이 부족한 고등학생 입장에서 정말 큰 이점일수 밖에 없습니다. 또 페이지수는 많지만 어떤 페이지도 최선을 다해 집필하지 않은 부분이 없었습니다. 정말 모든 예제에 최선을 다해 풀이를 써주신게 느껴집니다. 또 그렇다고 막 불필요하게 많은 교육과정을 벗어나는 지저분한 공식들 같은 것은 없습니다. 정말 알짜배기 책입니다. 만약 확통 기본서를 찾고 계신다면 저는 망설임없이 촌철살인 확률통계를 추천드립니다. 그리고 10월 출판예정인 미적분2 정말 기대됩니다 ㅎㅎ 개인적으로 마음에 쏙드는 책이라서 승승장구하면 좋겠습니다. 화이팅 ㅎㅎ
(오르비 닉네임 : 이불 님 / 경상북도 포항시)

네이버 책 리뷰 - 〈제목 : 수학공부에 정말 도움 많이 받았습니다.〉

처음엔 단순히 경우의 수에서 애매모호한 개념정립이 안되서 공부한 교재였는데 하다보니까 확통단원 뿐만이 아니라 수학공부 전반에 대해 많이 생각하게 해주는 책이었습니다.

책에는 문제보다 많은 정말 다양한 예제들이 나와있는데 그 풀이과정을 따라가다 보면 경우의 수에서 사용되는 논리의 전개에 익숙해지고 조건들을 파악하는 힘이 길러지는데 2회독쯤 하게 되면 문제를 보는 직관이 생기게 됩니다. 이게 순열과 조합 단원에서 헤메면서 잡아먹는 시간을 효율적으로 줄여줍니다. 단순히 양치기를 통해서 '감'을 잡는 것보다 훨씬 구체적이고 확실하게 개념이 어떻게 활용되어지는 지를 알려준다는 게 이 책의 장점이라고도 볼 수 있습니다.

책을 공부하다보면 중간중간에 저자의 수학에 대한 철학과 나름의 공부법 등이 나오는데 모든 책에나 나오는 말들이지만 곱씹어볼만한 가치가 있는 조언들도 많습니다. 특히 수학을 독학하시는 분들이 참고할만한 글들이라고 생각합니다. 단순히 양치기가 아닌 수학에 접근하는 길을 알려준다고 할까요.
마지막으로, 인강이나 몇 권의 개념서를 정독하고 나서도 뜬구름 잡는 거 같던 개념들을 이제 하나의 논리적 도구로서 사용할 수 있게 만들어준 저자에게 감사하단 인사를 남깁니다.
(오르비 닉네임 : TheAnecdote(수능전과목만점) 님 / 경상남도 진주시)

네이버 책 리뷰 - 〈제목 : 정말 좋은 책입니다.〉

일단 자세한 해설과 이미지로 개념을 확실하게 이해하는데 큰 도움을 줍니다.
다른 개념서 같은 경우에는 어느 정도의 적절한 설명과 관련된 문제를 통한 개념 이해 위주라면
이 책은 개념 설명에서부터 이미 문제를 풀 실력이 거의 완벽하게 길러진다고 볼 수 있습니다.
개념이 머리 속에 확실하게 정립되기 때문에 문제를 푸는 것은 그냥 단순한 복습 밖에 안된다고 할 수 있겠네요.
그렇다고 문제 수준이 낮은 건 아닙니다. 제일 중요한 개념이 확실하게 이해되기 때문에
문제가 그만큼 상대적으로 쉬워지는 거겠죠.

다양한 문제를 통해서 단순한 개념 이해가 아니라 고차원의 문제를 풀 실력이 향상되는게 느껴집니다.
이 교재를 본 후 다른 기출 문제집을 풀어보니 확실하게 실력이 향상되었음을 느꼈습니다.
촌철살인을 공부하기 전에는 기출을 풀어봐도 이해가 안되거나 풀었어도 애매한 경우가 많았는데
지금은 너무나 쉽고 재미있게 문제를 풀고 있는 제 자신을 발견하게 되네요.

다른 여러 교재로 개념도 제대로 이해하지 못하게 되어 문제를 풀어도 애매해 하는 불행한 일을 겪지 마시고
이 책으로 처음부터 확실한 개념을 정립하여 확통에 자신감을 갖게 되시길 바랍니다.
차차 다른 단원에 관련된 책도 출판된다고 하니 책이 출판되면 모두 구입하여 공부할 생각입니다.
앞으로 출판될 책도 이런 식으로 수학을 확실하게 이해하는데 도움이 되는 교재이길 바랍니다.

(오르비 닉네임 : 쿡리 님 / 대전광역시 서구)

## 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.

MAP

## #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 에서는 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 |

$\Rightarrow$ 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 |

$\Rightarrow$ 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 |

$\Rightarrow$

| $X$ | 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 <$x$에 시간, $y$에 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 |

$\Rightarrow$ 상대도수에 가중치라는 의미를 부여해보자. - 이때 평균은 가중 평균(가중치가 반영된 평균)이라는 의미를 가지게 된다.
$\Rightarrow$ 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 $5$점이라면 이 학생의 정확한 점수를 알 수는 없지만 이 학생이 평균점수보다 $5$점 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 $0$이다.

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

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

$$\Rightarrow x_1 + x_2 + \cdots + x_n = \overbrace{m + m + \cdots + m}^{n\text{개}}$$

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

이와 같이 편차의 합은 $0$이다.

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

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

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

⇨ 즉, $m = M + \frac{k}{n}$ 이므로

즉, 우리는 <가편차>를 모두 더한 값 $k$를 총 도수인 $n$으로 나눈 값인 $\frac{k}{n}$를 가평균 $M$에 더함으로서 실제평균 $m$을 구할 수 있다.

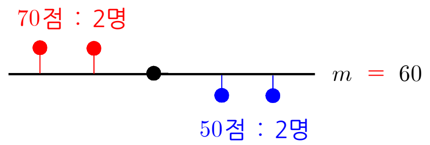
⇨ 예를 들어 이번 시험점수가 $47.8$, $78.5$, $92.1$, $82$, $88.5$, $65.5$, $100$라고 가정하고 가평균을 이용하여 평균을 구해보자. 가평균을 대략 $80$점 정도로 잡으면 가편차는 $-32.2$, $-1.5$, $12.1$, $2$, $8.5$, $-14.5$, $20$라고 볼 수 있다. 이때 가편차의 합은 $-5.6$이다. 즉, $\frac{-5.6}{7} = -0.8$이므로 우리가 구하는 실제 평균은 $80 + (-0.8) = 79.2$

## 2) 분산과 표준편차의 이해

- 다음 두 분포를 비교해 보자

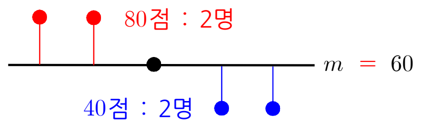
분포 1 :

| $X$ | 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〉을 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

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

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.

## #2 이산확률분포와 평균, 분산, 표준편차

- <상대도수라는 비율의 개념>이 <확률이라는 비율의 개념>으로 바뀌었을 뿐이다.
- <변량과 그에 따른 상대도수>의 관계를 <확률변수와 그에 따른 확률>의 관계로 바꾸어 표현하고 <함수>라는 용어도 사용하기 시작한다. 물론 상대도수분포에서 변량과 상대도수의 관계도 사실상 함수라고 표현할 수 있지만 중학교 때는 함수라는 표현을 쓰지 않았을 뿐이다.
- 이산확률분포에서 평균, 분산, 표준편차는 상대도수분포에서 평균, 분산, 표준편차와 <의미, 계산적인 측면>에서 완전히 같다. 단지 $\sum$ 와 수열의 표현을 활용해서 조금 더 일반적인 식으로 나타낼 뿐이다.

# 1. 이산확률분포

표를 매개로 모든 내용을 떠올린다. 표는 <이산확률분포와 그 공식>에서 필요한 모든 정보를 담고 있으므로 표를 매개로 공식과 식의 의미를 떠올리면 그냥 외우는 방식의 비논리, 비 확신의 느낌에서 벗어날 수 있다.
사실 처음 도수분포표 역시 표를 매개로 공부했기 때문에 이 방식이 친근하기도 하다.

## 1) 확률변수와 확률

| $X$ | $x_1$ | $x_2$ | $\cdots$ | $x_i$ | $\cdots$ | $x_n$ | 총합 |
|---|---|---|---|---|---|---|---|
| $P(X)$ | $p_1$ | $p_2$ | $\cdots$ | $p_i$ | $\cdots$ | $p_n$ | 1 |

<이산확률분포 표> - 편의 상 $x_1 < x_2 < \cdots < x_n$라고 하자.

① 이산확률변수 : <이산>이라는 말은 <셀 수 있다>정도로 이해하면 된다. 뒤에 나오는 <연속>이라는 말은 <셀 수 없다>는 뜻이다.
즉, 확률변수(구 변량)가 몇 개로 주어져 셀 수 있는 경우 우리는 이산확률변수라 하고 이러한 확률변수를 취하는 확률분포를 이산확률분포라고 한다. (확률변수가 무한개라도 위와 같이 각 확률변수를 자연수에 대응시켜 나열할 수 있는 경우는 셀 수 있는 것이다.)

② 확률 : 확률변수에 따른 확률로 함숫값에 해당된다. 즉, 확률변수가 결정되면 그에 따른 확률 값이 결정되는 함수의 구조이다. 그래서 기호 역시 함수의 기호를 차용하여 $P(X)$와 같이 표현한다.

③ 성질
- 확률의 총합은 항상 1이다. 이것은 거의 항상 쓰는 관계식이므로 반드시 기억해야 한다. ⇨ $\sum p_i = 1$
- 확률의 여러 가지 표현법 :
$p_1 + p_2$ ⇨ $P(X = x_1$또는 $X = x_2)$ or $P((X - x_1)(X - x_2) = 0)$ or $P(x_1 \le X \le x_2)$ or $P(x_1 \le X < x_3)$

티칭 통계의 $\sum$ 에서 아랫 끝과 윗 끝을 생략하는 이유
: 앞으로 나오는 통계의 모든 공식은 항상 확률변수를 <처음부터 마지막까지> 다 더하는 구조로 되어 있다.
위의 표를 참조하면 $i = 1$일 때가 처음이고, $i = n$일 때가 마지막이 된다. 이처럼 자주 쓰이는 식은 의미만 통한다면 생략하여 간단히 표시하는 것이 일반적이다. 그래서 $\sum_{i=1}^{i=n} p_i = 1$의 의미가 있는 식을 $\sum p_i = 1$와 같이 쓴다.
꼭 수학이 아니더라도 많이 쓰는 단어나 용어들은 의미만 통하면 되기 때문에 생략하여 짧게 표현하는 경우가 많다. 심지어 요즘 인터넷을 들어가면 이런 생략하는 용어들 때문에 의미전달이 안 될 정도다. 그런 것에 비하면 이정도의 생략은 귀요미정도 밖에 안 된다.

## 2) 확률질량함수 - 표를 함수적으로 표현하는 방법

<확률분포>를 함수처럼 다루겠다는 강력한 의지를 보여주는 부분이다. 사실 이산확률분포는 <표>만 있으면 충분하다.
하지만 우리는 이제 곧 <연속확률분포>를 다루게 될 것이고 이 경우에는 <확률분포>를 <표>를 이용해서 표현할 수 없으므로 항상 함수적인 표현을 이용할 수밖에 없다.
즉, 이산확률분포에서 확률질량함수는 <표를 표현하는 또 다른 방법>이라고 가볍게 이해하고 넘어가면 된다.

| $X$ | $x_1$ | $x_2$ | $\cdots$ | $x_i$ | $\cdots$ | $x_n$ | 총합 |
|---|---|---|---|---|---|---|---|
| $P(X)$ | $p_1$ | $p_2$ | $\cdots$ | $p_i$ | $\cdots$ | $p_n$ | 1 |

<이산확률분포 표>

- 확률질량함수 : $P(X=x_i) = p_i \ (i=1,2,3,\cdots,n)$

⇨ 수열과 같이 표현하면 된다. 위의 표에서 일반항을 계산하여 표현하고 이 일반항을 함수로 나타낸다.
그리고 간단하게 일반항의 아랫 끝의 제한범위를 표시해 주면 확률질량함수는 완성된다.

- 사실 이산확률분포 표를 확률질량함수로 나타내는 문제는 거의 나오지 않는다.
중요한 것은 <확률질량함수>로 <확률분포>가 주어진 경우에도 <표>를 매개로 공식이나 성질을 떠올리는 것이다.

코칭 <질량>이라는 말이 왜 붙었는지 궁금한 사람들은 뒤에 <개념의 외연>에서 다루므로 조금 참길 바란다.

코칭 <표를 완성>하는 문제는 <확률>문제

: 사실 표를 완성하는 문제는 <통계의 형식을 띤 확률 문제>일 가능성이 높다. 물론 $\sum p_i = 1$을 활용하여 빈칸의 일부를 채울 수는 있지만 궁극적으로는 상황이 주어지고 그 상황에서 확률 값을 구해서 빈칸을 채워야 한다.
그래서 통계를 시작하는 가장 처음 부분에서 <통계를 어렵다고 하는 두 번째 이유>로 통계를 가장한 확률 문제이기 때문이라는 사실을 언급했던 것이다.
표를 완성하는 문제에서 <확률변수>는 <시행에 따른 모든 가능한 결과>이다. 즉, 확률에서 표본공간의 모든 원소가 확률변수가 된다. <확률>은 확률변수에 해당하는 각 <근원사건의 확률>이다.

**094** 이해를 위한 예제

확률변수 $X$의 확률분포가 다음과 같다. 이때 다음 값을 구하여라.

| $X$ | 0 | 1 | 2 | 3 | 4 | 합 |
|---|---|---|---|---|---|---|
| $P(X)$ | 0.1 | 0.25 | 0.3 | 0.25 | 0.1 | 1 |

(1) $P(X=2 \text{ or } X=4)$ (2) $P(X<2)$ (3) $P(2 \leq X \leq 4)$ (4) $P(X \geq 1)$

(1) $P(X=2 \text{ or } X=4) = P(2) + P(4) = 0.3 + 0.1 = 0.4$
(2) $P(X<2) = P(0) + P(1) = 0.1 + 0.25 = 0.35$
(3) $P(2 \leq X \leq 4) = P(2) + P(3) + P(4) = 0.3 + 0.25 + 0.1 = 0.65$
(4) $P(X \geq 1) = P(1) + P(2) + P(3) + P(4)$
$= 1 - P(0) = 1 - 0.1 = 0.9$ - 1 에서 빼는 것이 낫다.

정답은 (1) $0.4$ (2) $0.35$ (3) $0.65$ (4) $0.9$

095 이해를 위한 예제 2007. 4. 가형(78%). 8번. 3점

확률변수 $X$가 취하는 모든 값이 $1,\ 2,\ 3,\ \cdots,\ 99$일 때, $X=k$일 확률은 $\mathrm{P}(X=k)=\dfrac{a}{\sqrt{k+1}+\sqrt{k}}\quad (k=1,2,3,\cdots,99)$ 이다. $\mathrm{P}(X=16)+\mathrm{P}(X=17)+\mathrm{P}(X=18)+\ \cdots\ +\mathrm{P}(X=99)=b$라 할 때, $a+b$의 값은?

① $\dfrac{5}{9}$ ② $\dfrac{7}{9}$ ③ $1$ ④ $\dfrac{11}{9}$ ⑤ $\dfrac{13}{9}$

항상 머릿속에 <이산확률분포 표>를 띄워놓고 표를 들여다보는 느낌으로 식을 적어가는 것이 좋다.

$\displaystyle\sum_{k=1}^{99}\frac{a}{\sqrt{k+1}+\sqrt{k}} = a\times\sum_{k=1}^{99}\left(\sqrt{k+1}-\sqrt{k}\right)$ (분모와 분자에 합 · 차 공식 / 확률의 총합은 1이므로 확률을 모두 더한 식이다.)

$= a\times\{(\cancel{\sqrt{2}}-\sqrt{1})+(\cancel{\sqrt{3}}-\cancel{\sqrt{2}})+\cdots+(\sqrt{100}-\cancel{\sqrt{99}})\} = 9a = 1 \Leftrightarrow a = \dfrac{1}{9}$

즉, $P(X=16)+P(X=17)+\cdots+P(X=99)$

$= \dfrac{1}{9}\times\displaystyle\sum_{k=16}^{99}\left(\sqrt{k+1}-\sqrt{k}\right) = \dfrac{1}{9}\times\{(\sqrt{17}-\sqrt{16})+(\sqrt{18}-\sqrt{17})+\cdots+(\sqrt{100}-\sqrt{99})\}$

$= \dfrac{1}{9}\times\left(\sqrt{100}-\sqrt{16}\right) = \dfrac{6}{9} = \dfrac{2}{3}$

$a = \dfrac{1}{9},\ b = \dfrac{2}{3}$이므로 $a+b = \dfrac{7}{9}$이다. 정답은 ②

**096** 이해를 위한 예제 2007. 6. 가형(42%). 23번. 4점

검은 공 3개, 흰 공 2개가 들어 있는 주머니가 있다. 이 주머니에서 한 개의 공을 꺼내어 색을 확인한 후 다시 넣지 않는다. 이와 같은 시행을 반복할 때, 흰 공 2개가 나올 때까지의 시행 횟수를 $X$라 하면 $P(X>3)=\frac{q}{p}$이다. $p+q$의 값을 구하시오. (단, $p$와 $q$는 서로소인 자연수이다.)

통계를 가장한 확률문제

시행 검은 공 3개, 흰 공 2개가 들어 있는 주머니가 있다. 이 주머니에서 한 개의 공을 꺼내어 색을 확인한 후 다시 넣지 않는다. 이와 같은 시행을 반복할 때, 흰 공 2개가 나올 때까지의 시행 횟수를 $X$

⇨ 시행에 따른 모든 가능한 결과는 2회, 3회, 4회, 5회

표의 작성 종속인 사건의 곱셈정리를 이용하여 표를 완성한다.

$P(2) = P(2\text{번만에 흰공 모두}) = P(\text{흰흰}) = \frac{2}{5}\times\frac{1}{4} = \frac{1}{10}$

$P(3) = P(3\text{번만에 흰공 모두}) = P(\text{검흰흰}) + P(\text{흰검흰}) = \frac{3}{5}\times\frac{2}{4}\times\frac{1}{3} + \frac{2}{5}\times\frac{3}{4}\times\frac{1}{3} = \frac{1}{5}$

$P(4) = P(4\text{번만에 흰공 모두}) = P(\text{검검흰흰}) + P(\text{검흰검흰}) + P(\text{흰검검흰})$

$= \frac{3}{5}\times\frac{2}{4}\times\frac{2}{3}\times\frac{1}{2} + \frac{3}{5}\times\frac{2}{4}\times\frac{2}{3}\times\frac{1}{2} + \frac{2}{5}\times\frac{3}{4}\times\frac{2}{3}\times\frac{1}{2} = \frac{3}{10}$

$P(5) = 1-\{P(2)+P(3)+P(4)\} = 1 - \frac{3}{5} = \frac{2}{5}$ - 앞부분의 계산의 확신이 있다면 마지막은 확률분포의 성질을 이용하여 구한다.

| $X$ | 2 | 3 | 4 | 5 | 총합 |
|---|---|---|---|---|---|
| $P(X)$ | $\frac{1}{10}$ | $\frac{1}{5}$ | $\frac{3}{10}$ | $\frac{2}{5}$ | 1 |

결국 표는 위와 같이 완성된다.

즉, $P(X>3) = P(4) + P(5) = \frac{3}{10} + \frac{2}{5} = \frac{7}{10}$

코칭 물론 이 문제는 표를 전부 작성할 필요가 없지만 연습 삼아 작성해본 것이다.

보통은 시행을 확인하여

| $X$ | 2 | 3 | 4 | 5 | 총합 |
|---|---|---|---|---|---|
| $P(X)$ | | | | | 1 |

까지 생각한 후 ( )에 해당하는 부분의 확률만 구하는 것이 일반적이다.

# 2. 평균 (= 기댓값)

## 1) $X$의 평균, 기댓값

| $X$ | $x_1$ | $x_2$ | $\cdots$ | $x_i$ | $\cdots$ | $x_n$ | 합계 |
|---|---|---|---|---|---|---|---|
| $P(X)$ | $p_1$ | $p_2$ | $\cdots$ | $p_i$ | $\cdots$ | $p_n$ | 1 |

- 평균을 구하는 과정은 앞에서 중등수학을 통해 자연스럽게 습득했을 것이므로 생략한다.
  처음 평균을 계산하는 연습을 할 때는 항상 표를 상상하면서 연습하길 바란다.

  $m = E(X) = \sum_{i=1}^{n} x_i p_i = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n$ (시그마는 아랫 끝과 윗 끝을 생략하여 $\sum x_i\, p_i$이라고 쓰기도 한다.)

  <확률변수>와 <그에 따른 확률>을 각각 곱하여 더한다. 빨리 계산하는 방법 없으니 진짜 곱하고 진짜 더하길 바란다.

- <평균>은 $mean$의 첫 글자를 따서 $m$이라고 쓰고, <기댓값>은 $\mathrm{exp}ectation$의 첫 글자를 따서 $E(X)$라고 표현한다.
  앞에서도 언급했듯이 평균과 기댓값은 완전히 같은 말이라고 생각해도 좋다. ⇨ 평균 = 기댓값
  $E(X)$에서 $X$는 <평균을 구하는 대상>을 의미한다. 즉, $X$가 몸무게를 의미한다면 $E(X)$는 몸무게의 평균이 될 것이고, $X$가 점수를 의미한다면 $E(X)$는 점수의 평균이 될 것이다.
  일반적으로 우리가 <평균>이나 <기댓값>이라고 말하는 것은 사실 <$X$의 평균>, <$X$의 기댓값>을 말하는 것이다.
  단지 평균을 주로 $X$에 대해서 구하기 때문에 관용적으로 $X$를 생략하여 부르는 것뿐이다.

- 평균을 $E(X)$와 같이 표현한다는 것은 $X$가 얼마든지 변형될 수 있다는 것을 암시한다.
  실제로 <분산>은 <편차 제곱의 평균>을 낸 것이므로 $E\left((X-m)^2\right)$와 같이 표현될 수 있다.
  이상하게 <편차 제곱의 기댓값>이라는 말을 잘 쓰진 않는데 특별한 이유는 없다. 틀린 표현이 아니며 그냥 습관처럼 굳어진 것뿐이다.

## 2) ~~의 평균, 기댓값

만약 $X^2$의 평균(또는 기댓값)을 구하라고 한다면 표를 다시 작성한다.

| $X$ | $x_1$ | $x_2$ | $\cdots$ | $x_i$ | $\cdots$ | $x_n$ | 합계 |
|---|---|---|---|---|---|---|---|
| $P(X)$ | $p_1$ | $p_2$ | $\cdots$ | $p_i$ | $\cdots$ | $p_n$ | 1 |

⇨

| $X^2$ | $x_1^2$ | $x_2^2$ | $\cdots$ | $x_i^2$ | $\cdots$ | $x_n^2$ | 합계 |
|---|---|---|---|---|---|---|---|
| $P(X)$ | $p_1$ | $p_2$ | $\cdots$ | $p_i$ | $\cdots$ | $p_n$ | 1 |

이때 역시 각 변형된 <확률변수 $X^2$>과 <그에 따른 확률>을 각각 곱해서 더하면 된다. ⇨ $E(X^2) = \sum_{i=1}^{n} x_i^2\, p_i$

이처럼 확률변수 $X$를 변형한 ★의 평균인 경우 ★은 생략할 수 없다. 반드시 <★의 평균>이라 하고 $E(★)$라고 표시한다.

예를 들어 ⇨ ($X$의) 평균, 기댓값 $m = E(X) = \sum x_i\, p_i$

- $m$은 <$X$의 평균>만 나타낼 수 있는 기호이다. 일반적으로 <$X$의 평균, 분산, 표준편차>는 <$X$의>를 생략하고 <평균, 분산, 표준편차>라고 부를 수 있다.

⇨ $X^2$의 평균, 기댓값 $E(X^2) = \sum x_i^2\, p_i$

⇨ $(X-m)^2$의 평균, 기댓값 $E\left((X-m)^2\right) = \sum (x_i - m)^2\, p_i$

코칭 $\sum p_i = 1$(확률의 총합은 1)은 문제 풀 때마다 반드시 고려한다. 문제의 미정계수를 구하기 위해서도 필요하지만 우리가 표를 작성하는 경우에 제대로 작성했는지 확인하기 위해서도 반드시 필요하다.
본인의 실력에 확신을 가지기 전까지는 반드시 확인해보는 것이 좋고, 실력이 뛰어나도 순간적으로 직관을 통해 그리고 습관적으로 확인해보는 것이 좋다.

# 3. 분산과 표준편차

## 1) 분산의 정의(편차 제곱의 평균)와 표준편차

- 앞에서 중등수학을 통해 자연스럽게 이해했을 것이므로 구하는 과정을 한 번 짚고 넘어가자.

  1 단계 : 먼저 평균을 구한다.

| $X$ | $x_1$ | $x_2$ | $\cdots$ | $x_i$ | $\cdots$ | $x_n$ | 합계 |
|---|---|---|---|---|---|---|---|
| $P(X)$ | $p_1$ | $p_2$ | $\cdots$ | $p_i$ | $\cdots$ | $p_n$ | 1 |

즉, $\sum_{i=1}^{n} x_i p_i$을 계산하여 $m$을 얻는다. ($m$은 $\sum_{i=1}^{n} x_i p_i$을 계산하여 얻은 결과라고 가정하자.)

  2 단계 : <편차 제곱>을 확률변수로 갖는 새로운 확률분포 표를 작성한다.

| $(X-m)^2$ | $(x_1-m)^2$ | $(x_2-m)^2$ | $\cdots$ | $(x_i-m)^2$ | $\cdots$ | $(x_n-m)^2$ | 합계 |
|---|---|---|---|---|---|---|---|
| $P(X)$ | $p_1$ | $p_2$ | $\cdots$ | $p_i$ | $\cdots$ | $p_n$ | 1 |

  3 단계 : <편차 제곱의 평균>을 구한다.

즉, $\sigma^2 = V(X) = E((X-m)^2) = \sum_{i=1}^{n} (x_i - m)^2 p_i$

- <표준편차>라고 하는 것은 영어로 $standard\ \ deviation$이고, 약자로 쓸 때는 $s$에 해당하는 그리스 알파벳인 $\sigma$(시그마)를 이용하여 $\sigma$ 또는 $\sigma(X)$라고 표현한다. <분산>은 $\sigma^2$이라고 표현하거나 $variance$의 첫 글자를 따서 $V(X)$라고 표현한다.

- <평균>과 마찬가지로 <$X$의 분산, $X$의 표준편차>는 $X$를 생략하고 <분산, 표준편차>라고 부르기로 한다.

⇨ 즉, ($X$의) 분산 : $V(X) = \sigma^2$ ($X$의) 표준편차 : $\sigma(X) = \sigma$

⇨ 확률변수의 언급 없이 $\sigma$, $\sigma^2$이라고 표현되어 있다면 이것은 무조건 <$X$의 표준편차>, <$X$의 분산>이다.

## 2) 분산의 계산 공식

$\sum_{i=1}^{n} (x_i - m)^2 p_i$을 조금 빨리 계산하는 공식이다. 항상 효율적인 것은 아니지만 고등학교 과정에서는 거의 효율적이다.

- 역시 앞에서 중등수학을 통해 자연스럽게 이해했을 것이므로 구하는 과정을 한 번 짚고 넘어가자.

  1 단계 : 먼저 평균을 구한다.

| $X$ | $x_1$ | $x_2$ | $\cdots$ | $x_i$ | $\cdots$ | $x_n$ | 합계 |
|---|---|---|---|---|---|---|---|
| $P(X)$ | $p_1$ | $p_2$ | $\cdots$ | $p_i$ | $\cdots$ | $p_n$ | 1 |

⇨ $\sum_{i=1}^{n} x_i p_i$을 계산하여 $m$을 얻는다.

  2 단계 : <$X$의 제곱>을 확률변수로 갖는 새로운 확률분포 표를 작성한다.

| $X$ | ${x_1}^2$ | ${x_2}^2$ | $\cdots$ | ${x_i}^2$ | $\cdots$ | ${x_n}^2$ | 합계 |
|---|---|---|---|---|---|---|---|
| $P(X)$ | $p_1$ | $p_2$ | $\cdots$ | $p_i$ | $\cdots$ | $p_n$ | 1 |

  3 단계 : <$X$ 제곱의 평균>을 구한 후 앞서 구한 <평균>을 제곱하여 뺀다.

즉, $\sigma^2 = V(X) = E(X^2) - \{E(X)\}^2 = \sum_{i=1}^{n} {x_i}^2 p_i - \left(\sum_{i=1}^{n} x_i p_i\right)^2$

- 어떤 방식으로 구하든 분산을 구하기 위해서는 항상 먼저 평균을 구해야 한다. 평균을 구하고 분산을 구하고 표준편차를 구한다. 앞의 것을 빼먹고 뒤의 것을 구할 수 없다. <평 · 분 · 표> ⇨ 평균 → 분산 → 표준편차
- 서편제의 후속편 제평제 (제곱의 평균 − 평균의 제곱)

## 3) 공식의 증명

물론 <#1. 중등통계에서 배우다.>를 통해서 한 번 언급되었던 내용이지만 조금 더 일반적인 증명을 해본다.

- $V(X) = E((X-m)^2)$ <분산의 정의> $= \sum \boxed{(x_i-m)^2} p_i$ <평균의 $\sum$ 표현> $= \sum \boxed{(x_i^2-2mx_i+m^2)} p_i$ <전개>

$= \sum (x_i^2 p_i - 2m x_i p_i + m^2 p_i)$

$= \sum x_i^2 p_i - 2m \sum x_i p_i + m^2 \sum p_i$ <$\sum$ 의 성질>

$= E(X^2) - 2mE(X) + m^2$ <$\sum p_i = 1$>

$= E(X^2) - 2m \cdot m + m^2$ <$m = E(X)$>

$= E(X^2) - m^2$

$= E(X^2) - E(X)^2$

- $\sigma(X) = \sqrt{V(X)}$

티칭 분산과 표준편차 정리

① 분산의 정의 : 분산 = 편차 제곱의 평균 ⇨ $\sigma^2 = V(X) = E((X-m)^2)$

② 분산의 공식 : $\sigma^2 = V(X) = E(X^2) - E(X)^2$ ⇨ 보통은 $E((X-m)^2)$보다 $E(X^2) - E(X)^2$이 편하다.

③ 표준편차 : $\sigma = \sigma(X) = \sqrt{V(X)}$ ⇨ 분산을 계산한 뒤 $\sqrt{\ }$를 씌운다.

티칭 <분산>과 <평균>이 주어진 경우 <제곱의 평균> 구하기

⇨ 분산 공식을 살짝 변형한 $E(X^2) = V(X) + E(X)^2$은 꽤 자주 나오는 공식이다.

(후에 이항분포에서 특히나 많이 나온다. 이유는 후에 언급한다.)

코칭 평균과 분산을 구하는 계산은 원래 길다. 빨리 계산하는 방법 없으니 진짜 곱하고 진짜 더하길 바란다.

코칭 확률변수가 직접적으로 언급되지 않는 경우 <문제에서 요구하는 평균과 분산>을 통해 <확률변수>를 확인한다.
예를 들어 주사위의 눈의 평균을 물어본다면 주사위의 눈을 확률변수 $X$로 하여 표를 작성한다.
평균과 분산을 구하려면 어차피 표를 완전체로 작성해야 한다.

# 4. 평균, 분산, 표준편차의 성질 - 확률변수의 변형

## 1) 성질과 느낌

① $E(aX+b) = aE(X)+b$　　② $V(aX+b) = a^2V(X)$　　③ $\sigma(aX+b) = |a|\sigma(X)$

① $E(aX+b) = aE(X)+b$의 느낌

: 우리 반의 <모든 학생>들의 점수를 10점씩 더했다고 생각해보면 당연히 우리 반의 <평균점수>는 10점만큼 오른다.
우리 반의 <모든 학생>들의 점수를 두 배 했다고 생각해보면 당연히 우리 반의 <평균점수>는 두 배가 된다.
그래도 감이 잘 오지 않는다면 확률변수를 세 개 정도로 두고 실제 계산을 통해 느껴본다.

② $V(aX+b) = a^2V(X)$의 느낌

: 우리 반의 <모든 학생>들의 점수를 10점씩 더했다고 가정해 보자. 이 경우 <모든 학생>들의 <편차점수>는 변함이 없다.
내 점수도 10점이 올랐지만 평균점수도 10점이 올랐기 때문에 <변량 − 평균>은 그대로일 수밖에 없다.
<편차점수>가 그대로라면 <편차제곱의 평균>인 분산 역시 변함이 없을 수밖에 없다.
⇨ 즉, 분산은 확률변수에 더한 값 $b$에는 영향을 받지 않는다.

우리 반의 <모든 학생>들의 점수를 두 배 했다고 가정해보자. 이 경우 <편차점수>는 두 배가 된다.
원래 편차였던 <변량 − 평균>이 새로운 편차 <$2\times$변량 − $2\times$평균>이 되었으니 $2$로 묶어준다면 <원래 편차점수>에 $2$를 곱한 값이 새로운 편차가 되는 것이다. 결국 <새로운 편차의 제곱>은 <원래 편차의 제곱>에 $4$배가 될 수밖에 없다.
⇨ 즉, 분산은 확률변수에 곱한 값 $a$에는 제곱 배 만큼 영향을 받는다.

③ $\sigma(aX+b) = |a|\sigma(X)$ - 표준편차는 $\sqrt{\text{분산}}$ 일 뿐이므로 분산을 통해서 추론하면 된다.

## 2) 성질의 증명 : Σ 의 성질로 증명한다.

항상 표를 생각하면서 식을 쓰길 바란다.

| $X$ | $x_1$ | $x_2$ | $\cdots$ | $x_i$ | $\cdots$ | $x_n$ | 합계 |
|---|---|---|---|---|---|---|---|
| $P(X)$ | $p_1$ | $p_2$ | $\cdots$ | $p_i$ | $\cdots$ | $p_n$ | 1 |

⇨

| $aX+b$ | $ax_1+b$ | $ax_2+b$ | $\cdots$ | $ax_i+b$ | $\cdots$ | $ax_n+b$ | 합계 |
|---|---|---|---|---|---|---|---|
| $P(X)$ | $p_1$ | $p_2$ | $\cdots$ | $p_i$ | $\cdots$ | $p_n$ | 1 |

① $E(aX+b) = \sum(ax_i+b)p_i = a\sum x_ip_i + b\sum p_i$ <Σ 의 성질>

$= a\sum x_ip_i + b = aE(X)+b$

② $V(aX+b) = \sum(ax_i+b)^2p_i - \{E(aX+b)\}^2$ <분산 공식 : 제평제>

$= \sum(ax_i+b)^2p_i - \{aE(X)+b\}^2$ <①의 공식 - 수학에서 증명은 이미 앞에서 증명된 공식은 사용하는 것이다. >

$= a^2\sum x_i^2p_i + 2ab\sum x_ip_i + b^2 - [a^2\{E(X)\}^2 + 2abE(X) + b^2]$ <전개와 시그마 분배 과정 생략>

$= a^2E(X^2) + 2abE(X) + b^2 - [a^2\{E(X)\}^2 + 2abE(X) + b^2]$

$= a^2\{E(X^2)-E(X)^2\} = a^2V(X)$

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

코칭 위의 증명이 문제화 될 가능성은 낮으므로 증명에 집착할 필요는 없다. 하지만 위의 식은 <계산이 복잡>할 뿐이지 <의미구조가 복잡>한 식은 아니다. 즉, 위의 식은 마음만 먹으면 누구나 이해할 수 있고, 스스로 증명을 써볼 수도 있다. 위의 증명을 써봄으로서 얻을 수 있는 것은 <통계의 기호의 의미를 읽어가는 자연스러운 시각>이 생긴다는 것이다. 즉, 저자는 반드시 스스로 <자연스럽게> 증명할 수 있게 된 후 넘어가기를 권한다.

티칭 성질 외 - 분산을 구하는 과정에서 잠시 사용된 바 있다.

예를 들어 $E(aX+b) = aE(X)+b$의 식은 배웠지만 $E(aX^2+bX+c)$의 식은 공식으로서 다루지 않는다.

(이유는 $E(aX^2+bX+c)$은 간단히 증명되어 사용할 수 있는 반면에 $V(aX^2+bX+c)$, $\sigma(aX^2+bX+c)$에 대해서는 일반적인 식이 너무 복잡하고 중요도가 떨어지기 때문이다.)

하지만 평균의 가장 기본적인 원리로부터 $E(aX^2+bX+c)$의 식을 추론할 수 있기 때문에 종종 나오는 경우가 있다.

| $X$ | $x_1$ | $x_2$ | $\cdots$ | $x_i$ | $\cdots$ | $x_n$ | 합계 |
|---|---|---|---|---|---|---|---|
| $P(X)$ | $p_1$ | $p_2$ | $\cdots$ | $p_i$ | $\cdots$ | $p_n$ | 1 |

⇨

| $aX^2+bX+c$ | $ax_1^2+bx_1+c$ | $ax_2^2+bx_2+c$ | $\cdots$ | $ax_i^2+bx_i+c$ | $\cdots$ | $ax_n^2+bx_n+c$ | 합계 |
|---|---|---|---|---|---|---|---|
| $P(X)$ | $p_1$ | $p_2$ | $\cdots$ | $p_i$ | $\cdots$ | $p_n$ | 1 |

각 변형된 〈확률변수 $aX^2+bX+c$〉와 〈그에 따른 확률〉을 각각 곱해서 더한다.

$$\Rightarrow E(aX^2+bX+c) = \sum_{i=1}^{n}\left(ax_i^2+bx_i+c\right)\cdot p_i = a\sum_{i=1}^{n}x_i^2 p_i + b\sum_{i=1}^{n}x_i p_i + c\sum_{i=1}^{n}p_i$$
$$= aE(X^2) + b\sum_{i=1}^{n}x_i p_i + c\sum_{i=1}^{n}p_i$$
$$= aE(X^2)+bE(X)+c$$

**097** 이해를 위한 예제 2010. 11. 나형(83%). 8번. 3점

확률변수 $X$의 확률분포 표는 다음과 같다. $P(0 \le X \le 2) = \dfrac{7}{8}$일 때, 확률변수 $X$의 평균 $E(X)$의 값은?

| $X$ | $-1$ | 0 | 1 | 2 | 계 |
|---|---|---|---|---|---|
| $P(X=x)$ | $\dfrac{3-a}{8}$ | $\dfrac{1}{8}$ | $\dfrac{3+a}{8}$ | $\dfrac{1}{8}$ | 1 |

① $\dfrac{1}{4}$ ② $\dfrac{3}{8}$ ③ $\dfrac{1}{2}$ ④ $\dfrac{5}{8}$ ⑤ $\dfrac{3}{4}$

- 먼저 조건 $P(0 \le X \le 2) = \dfrac{7}{8}$을 이용하여 $a$부터 구한다.

$$\Rightarrow P(0 \le X \le 2) = \frac{7}{8} \Leftrightarrow P(0) + P(1) + P(2) = \frac{7}{8} \Leftrightarrow \frac{5+a}{8} = \frac{7}{8} \Leftrightarrow a=2$$

| $X$ | $-1$ | 0 | 1 | 2 | 계 |
|---|---|---|---|---|---|
| $P(X=x)$ | $\dfrac{1}{8}$ | $\dfrac{1}{8}$ | $\dfrac{5}{8}$ | $\dfrac{1}{8}$ | 1 |

즉, $E(X) = \sum_{i=1}^{n}x_i p_i = (-1)\times\dfrac{1}{8} + 0\times\dfrac{1}{8} + 1\times\dfrac{5}{8} + 2\times\dfrac{1}{8} = \dfrac{3}{4}$

정답은 ⑤

**098** 이해를 위한 예제 2004. 9. 나형(73%). 8번. 4점

이산확률변수 $X$의 확률분포 표는 다음과 같다. 확률변수 $X$의 평균이 5일 때, $X$의 분산은?

| $X$ | 1 | 2 | 4 | 8 | 합계 |
|---|---|---|---|---|---|
| $P(X=x)$ | $\frac{1}{4}$ | $a$ | $\frac{1}{8}$ | $b$ | 1 |

① 9.75 ② 8.5 ③ 7.25 ④ 6.5 ⑤ 4.25

- $\sum p_i = 1$을 이용해서 $a, b$의 관계식을 하나 세운다. ⇨ $\frac{1}{4} + a + \frac{1}{8} + b = 1 \quad \Leftrightarrow \quad a+b = \frac{5}{8}$
- 확률변수 $X$의 평균이 5일 때,

⇨ $E(X) = \sum_{i=1}^{n} x_i p_i = 1 \times \frac{1}{4} + 2 \times a + 4 \times \frac{1}{8} + 8b = 5 \quad \Leftrightarrow \quad 2a + 8b = \frac{17}{4}$

이 두 식을 연립하여 $a, b$를 구하면 $a = \frac{1}{8}$, $b = \frac{1}{2}$ 즉,

| $X$ | 1 | 2 | 4 | 8 | 합계 |
|---|---|---|---|---|---|
| $P(X=x)$ | $\frac{1}{4}$ | $\frac{1}{8}$ | $\frac{1}{8}$ | $\frac{1}{2}$ | 1 |

- 평균이 주어져 있으므로 $E(X^2)$을 구한다. ⇨ $\sum_{i=1}^{n} x_i^2 p_i = 1^2 \times \frac{1}{4} + 2^2 \times \frac{1}{8} + 4^2 \times \frac{1}{8} + 8^2 \times \frac{1}{2} = \frac{139}{4}$

$V(X) = E(X^2) - \{E(X)\}^2 = \frac{139}{4} - 25 = \frac{39}{4} = 9.75$이므로 정답은 ①

**099** 이해를 위한 예제 2011. 11. 나형(84%). 6번. 3점

확률변수 $X$의 확률분포를 표로 나타내면 다음과 같다. $E(4X+10)$의 값은?

| $X$ | 0 | 1 | 2 | 합계 |
|---|---|---|---|---|
| $P(X=x)$ | $\frac{1}{4}$ | $a$ | $2a$ | 1 |

① 11 ② 12 ③ 13 ④ 14 ⑤ 15

- $\sum p_i = 1$을 이용해서 $a$를 구한다. ⇨ $\frac{1}{4} + a + 2a = 1 \quad \Leftrightarrow \quad a = \frac{1}{4}$

| $X$ | 0 | 1 | 2 | 합계 |
|---|---|---|---|---|
| $P(X=x)$ | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{2}$ | 1 |

- $E(X) = \sum_{i=1}^{n} x_i p_i = 0 \times \frac{1}{4} + 1 \times \frac{1}{4} + 2 \times \frac{1}{2} = \frac{5}{4}$이므로

⇨ $E(4X+10) = 4E(X)+10 = 4 \times \frac{5}{4} + 10 = 15$

정답은 ⑤

**100** 이해를 위한 예제 2009. 11. 나형(77%). 8번. 3점

확률변수 $X$의 확률분포 표는 다음과 같다. 확률변수 $7X$의 분산 $\mathrm{V}(7X)$ 의 값은?

| $X$ | 0 | 1 | 2 | 계 |
|---|---|---|---|---|
| $\mathrm{P}(X=x)$ | $\frac{2}{7}$ | $\frac{3}{7}$ | $\frac{2}{7}$ | 1 |

① 14 ② 21 ③ 28 ④ 35 ⑤ 42

- $E(X) = \sum_{i=1}^{n} x_i p_i = 0 \times \frac{2}{7} + 1 \times \frac{3}{7} + 2 \times \frac{2}{7} = 1$
- $E(X^2) = \sum_{i=1}^{n} x_i^2 p_i = 0^2 \times \frac{2}{7} + 1^2 \times \frac{3}{7} + 2^2 \times \frac{2}{7} = \frac{11}{7}$ 이므로

⇨ $V(X) = E(X^2) - \{E(X)\}^2 = \frac{11}{7} - 1^2 = \frac{4}{7}$이고 $V(7X) = 7^2 V(X) = 7^2 \times \frac{4}{7} = 28$

정답은 ③

## 학습조언. 절대 빨리 하려고 하지 말라.

- 수학의 이론은 반드시 연결된다.

수학이 엄청나게 많은 이론을 배우는 것 같지만.. 사실은 몇 개의 이론밖에 다루지 않는다. (사실 그놈이 그놈이다.)
앞 단원을 충분히 학습하다보면 뒤의 다른 단원에서 앞서 공부했던 느낌과 비슷한 느낌을 받을 수 있는데 이것을 수학에서는 〈동형사상〉이라고 표현한다. 한 가지 이론을 제대로 이해하기 위해 많은 시간을 소비했다면 절대 시간낭비가 아니다.
수학은 반드시 연결된다. 오히려 이해가 가지 않는 부분을 계속 넘어왔기 때문에 수학이 계속 약점으로 남아 있는 것이다.
수학은 천천히 하나하나 〈납득〉하고 넘어가는 것이 무조건 시간을 절약하는 것이다.

빨리 하고 싶다면 많이 해라. 저절로 빨라진다.

- 〈아는 것〉과 〈이해한 것〉은 다르다.

수학자 비너는 수학에서 개념을 이해했다는 것은 〈개념의 이미지와 느낌〉이 떠오르는 것이라고 말했다.
저자 역시 이 부분에 대해서 동의하며 이것이 곧 직관이라고 생각한다. 앞서 〈통계의 공식〉을 중등 통계로부터 작은 숫자를 두어 설명한 것도 〈평균, 분산, 표준편차〉의 이미지와 느낌을 심어주기 위함이었다. 〈직관은 경험에 의한 결과〉이고 이러한 구체적 경험은 예시와 증명을 통해서 쌓는다. 만약 증명이 와 닿지 않는다면 더 구체적인 숫자로 몇 번 해본 다음 증명해보면 된다.

만약 이 책을 통해서 〈평균, 분산, 표준편차〉의 느낌을 조금 더 구체적으로 이해했다고 느꼈다면 이 책의 서술 방식을 공부에 적용해 보길 바란다. 절대 시간이 더 걸리거나 어려운 공부가 아니다.
이해가 안 되면 몇 가지 구체적인 예를 통해서 기억하고 공부한다. 그리고 증명을 다시 이해한다. 증명을 안보고 쓸 수 있으면 식을 외운 것이 아니라 식의 의미 구조를 읽어나가는 눈이 생긴 것이다.

수학은 〈오직 단순 암기〉로 증명과정을 헷갈리지 않고 모두 쓰는 것은 불가능하다.

## #3 이항분포

- 이항분포도 이산확률분포의 일부이므로 마찬가지로 표를 매개로 모든 내용을 떠올린다. 이항분포는 자칫하면 공식에 매몰 될 수 있다. (물론 공식이 매우 중요하지만) 이항분포도 결국 이산확률분포라는 것을 잊으면 안 된다.

- 이항분포의 <평균, 분산, 표준편차>는 증명을 이해하고 공식을 외운다면 오히려 <이항분포가 아닌 이산확률분포>의 <평균, 분산, 표준편차>를 구하는 것보다 계산이 훨씬 쉽다. 즉, 증명은 다소 복잡하지만 공식은 쉽다.

- <이항분포의 상황>은 <독립시행의 확률의 공식>을 쓸 수 있는 상황과 같다. 즉, <독립시행의 확률공식>에 대해서 완전히 익숙해야 하고 자연스럽게 쓸 수 있어야 한다. - 이것은 이미 앞에서도 여러 번 강조했다.

- 이항분포의 <평균, 분산, 표준편차>를 증명하는 과정은 <경우의 수 : 이항정리>에서 공부한 과정을 그대로 따른다.
  이 부분을 제대로 학습하지 않고 넘어왔다면 어려울 수 있다.
  반드시 <촌철살인 확률과 통계 제 1권>의 <Part 6. 조합의 연속 합>을 학습하고 넘어와야 한다.

# 1. 이항분포

### 1) 확률변수와 확률

| $X$ | 0 | 1 | $\cdots$ | $r$ | $\cdots$ | $n$ | 총합 |
|---|---|---|---|---|---|---|---|
| $P(X)$ | ${}_nC_0p^0(1-p)^n$ | ${}_nC_1p^1(1-p)^{n-1}$ | $\cdots$ | ${}_nC_rp^r(1-p)^{n-r}$ | $\cdots$ | ${}_nC_np^n(1-p)^0$ | 1 |

① 이항분포의 확률변수

: 이항분포의 확률변수는 <독립시행의 결과>로 항상 정해져 있다. 즉, <$n$회의 독립시행에서 사건 $A$가 일어나는 횟수>가 확률변수이다. $n$회의 독립시행에서 확률변수는 항상 $0, 1, \cdots, n$으로 정해져 있으며 (사건 $A$는 아예 안 일어나거나 최대 $n$회 모두 일어날 수도 있다.) 관용적으로 일반항에 $x_i$대신 $r$을 쓴다.

② 확률 : 확률변수에 따른 확률이므로 <독립시행의 확률>이다. 예를 들어 $P(X=1) = {}_nC_1p^1(1-p)^{n-1}$

⇨ 이런 이유로 이항분포는 표도 확률질량함수도 거의 정해져 있다.
문제에 따라서 시행횟수 $n$과 1회 시행에서 사건 $A$가 일어날 확률 $P(A)=p$정도만 달라질 뿐이다.

③ 성질 : 확률의 총합이 1 ($\sum p_i = 1$)은 이항정리를 통해서 확인할 수 있다.

⇨ $\displaystyle\sum_{r=0}^{n} {}_nC_rp^r(1-p)^{n-r} = \{p+(1-p)\}^n = 1$

### 2) 확률질량함수 - 표를 함수적으로 표현하는 방법

| $X$ | 0 | 1 | $\cdots$ | $r$ | $\cdots$ | $n$ | 총합 |
|---|---|---|---|---|---|---|---|
| $P(X)$ | ${}_nC_0p^0(1-p)^n$ | ${}_nC_1p^1(1-p)^{n-1}$ | $\cdots$ | ${}_nC_rp^r(1-p)^{n-r}$ | $\cdots$ | ${}_nC_np^n(1-p)^0$ | 1 |

- 확률질량함수 : $P(X=r) = {}_nC_rp^r(1-p)^{n-r}$ (단, $r=0,1,2,\cdots,n$)

⇨ <확률질량함수>로 <이항분포>가 주어진 경우에도 항상 <표>를 매개로 생각한다.

## 2. 이항분포표의 모든 의미를 담은 기호 : $B(n, p)$

### 1) $B(n, p)$의 의미

- $n$과 $p$만 결정되면 이항분포의 확률변수와 확률은 모두 결정된다. 즉, $n$과 $p$만 알면 <표>가 결정된다.
 다시 한 번 강조하면 $B(n,\ p)$는 표의 정보를 요약해서 보여주는 것으로 <이항분포 표>가 주어진 것이나 마찬가지다.
 $B(n, p)$에서 $B$는 이항분포의 영문인 binomial distribution의 이니셜이다.

- 잠시 후 배우는 <이항분포의 평균, 분산, 표준편차>도 $n$과 $p$만 알면 공식을 이용해서 쉽게 구할 수 있다.
 즉, <이항분포>는 <총 시행횟수 $n$과 1회 시행에서 사건 $A$가 발생할 확률 $p$>만 알면 된다.

예를 들어 $B\left(10,\ \frac{1}{3}\right)$의 정보로 표를 작성하면

| $X$ | $0$ | $1$ | $\cdots$ | $10$ | 총합 |
|---|---|---|---|---|---|
| $P(X)$ | ${}_{10}C_0\left(\frac{1}{3}\right)^0\left(\frac{2}{3}\right)^{10}$ | ${}_{10}C_1\left(\frac{1}{3}\right)^1\left(\frac{2}{3}\right)^9$ | $\cdots$ | ${}_{10}C_{10}\left(\frac{1}{3}\right)^{10}\left(\frac{2}{3}\right)^0$ | $1$ |

- 만약 $P(0 \le x \le 2)$를 구하라고 한다면 <위의 표>를 상상해서 식을 쓴다.
 즉, $P(0 \le x \le 2) = P(X=0) + P(X=1) + P(X=2)$

$$= {}_{10}C_0\left(\frac{1}{3}\right)^0\left(\frac{2}{3}\right)^{10} + {}_{10}C_1\left(\frac{1}{3}\right)^1\left(\frac{2}{3}\right)^9 + {}_{10}C_2\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^8$$ - 계산하지 말고 의미만 생각

### 2) $B(n, p)$로 표현하기

- 주어진 조건에서 독립시행의 확률이 나오는 확률분포의 상황이라면 이항분포라는 확신을 가질 수 있다.
 이때 고정수가 총 시행횟수이고, 변동수와 같은 지수를 갖는 확률이 1회 시행에서 발생할 확률이다.

$$즉,\ P(X=r) = {}_nC_r p^r(1-p)^{n-r}인\ 경우$$

 $B(n, p)$의 분포를 가진다.

티칭 당장은 구할 수 없는 이항분포의 확률

나중에 배우는 내용이긴 하지만 미리 한 번 느껴보자. 현재 이항분포의 확률을 물어 볼 때는 기본적으로 $n$이 작은 경우만 다룬다. 그도 그럴 것이 예를 들어 생각해보면 $B\left(100,\ \frac{1}{6}\right)$에서 $P(0 \le x \le 2)$을 물어본다면

$$P(0 \le x \le 2) = {}_{100}C_0\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^{100} + {}_{100}C_1\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^{99} + {}_{100}C_2\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^{98}$$

인데 어떤 사람이 구할 수 있겠는가? (자손 대대로 계산해도 못 구한다.)
 그래서 이처럼 $n$이 큰 경우 <이항분포의 확률>은 <오성이 콩의 개수를 센 아이디어>처럼 <면적>으로 환산해서 구하는데 이것은 후에 <정규분포>을 학습한 후 공부하기로 한다.

**101** 이해를 위한 예제

확률변수 $X$가 이항분포 $B\left(2,\ \frac{1}{3}\right)$을 따를 때, $X$의 확률분포 표를 만들어라.

- 총 2회의 시행 중 사건 $A$가 일어날 수 있는 모든 결과는 0회, 1회, 2회이다.
- 1회 시행에서 사건 $A$가 일어날 확률 $P(A) = \frac{1}{3}$이다. 이것으로 독립시행의 확률까지 구해서 표를 완성하면

정답은

| $X$ | 0 | 1 | 2 |
|---|---|---|---|
| $P(X)$ | ${}_2C_0\left(\frac{1}{3}\right)^0\left(\frac{2}{3}\right)^2$ | ${}_2C_1\left(\frac{1}{3}\right)^1\left(\frac{2}{3}\right)^1$ | ${}_2C_2\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^0$ |

**102** 이해를 위한 예제 2007. 10. 나형(40%). 18번. 3점

확률변수 $X$는 이항분포 $\mathrm{B}\left(n,\ \frac{1}{2}\right)$을 따른다. $P(X=2) = 10P(X=1)$이 성립할 때, $n$의 값을 구하시오.

당시 학생들은 이항분포 $\mathrm{B}\left(n,\ \frac{1}{2}\right)$가 가진 의미를 모르고 뒤에 나오는 <평균, 분산, 표준편차의 공식>만 외웠기 때문에 이런 문제의 정답률이 낮다. 표를 보면서 쓰는 느낌으로 식을 쓴다.

| $X$ | 0 | 1 | 2 | $\cdots$ |
|---|---|---|---|---|
| $P(X)$ | ${}_nC_0\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^n$ | ${}_nC_1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^{n-1}$ | ${}_nC_2\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^{n-2}$ | $\cdots$ |

$P(X=2) = 10P(X=1) \iff {}_nC_2\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^{n-2} = 10\,{}_nC_1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^{n-1}$ - 간단한 지수법칙 : $\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^{n-2} = \left(\frac{1}{2}\right)^n$

$\iff {}_nC_2\cancel{\left(\frac{1}{2}\right)^n} = 10\,{}_nC_1\cancel{\left(\frac{1}{2}\right)^n}$ - 0이 아니라는 확신이 약분을 가능케 한다.

$\iff {}_nC_2 = 10\,{}_nC_1$

$\iff \frac{\cancel{n}(n-1)}{2} = 10\cancel{n} \iff \frac{n-1}{2} = 10$ - 0이 아니라는 확신이 약분을 가능케 한다.

즉, 계산하면 $n = 21$ 이다.

# 3. 이항분포의 평균, 분산, 표준편차

| 1) 공식 (먼저 외운다.) | 2) 평균, 분산, 표준편차의 성질 |
|---|---|
| $B(n,\ p)$ (단, $q=1-p$)이라고 하면<br>① $E(X)\ =\ np$<br>② $V(X)\ =\ npq$<br>③ $\sigma(X)\ =\ \sqrt{npq}$ | $B(n, p)$ (단, $q=1-p$)이라고 하면<br>① $E(aX+b)\ =\ aE(X)+b\ =\ anp+b$<br>② $V(aX+b)\ =\ a^2V(X)\ =\ a^2np(1-p)$<br>③ $\sigma(aX+b)\ =\ \lvert a\rvert\sqrt{np(1-p)}$ |

티칭 이처럼 이항분포는 $n$과 $p$만 알면 간단한 공식을 통해서 〈평균〉과 〈분산〉을 매우 쉽게 구할 수 있다.
그래서 원래 일반적인 〈이산확률분포〉는 $V(X)$를 구하기 위해서 $E(X^2)$을 구했지만 〈이항분포〉는 $V(X)$를 공식으로 매우 쉽게 구할 수 있기 때문에 반대로 $E(X^2)$을 구하기 위해 $V(X)$를 구한다.
즉, $E(X^2)\ =\ \{E(X)\}^2\ +\ V(X)\ =\ (np)^2\ +\ npq$

## 3) 이항분포에서 평균, 분산의 시그마 표현

| $X$ | 0 | 1 | $\cdots$ | $r$ | $\cdots$ | $n$ | 총합 |
|---|---|---|---|---|---|---|---|
| $P(X)$ | ${}_nC_0\,p^0(1-p)^n$ | ${}_nC_1\,p^1(1-p)^{n-1}$ | $\cdots$ | ${}_nC_r\,p^r(1-p)^{n-r}$ | $\cdots$ | ${}_nC_n\,p^n(1-p)^0$ | 1 |

- 이항분포에서는 $x_i$ 라는 기호 대신에 $r$ 이라는 기호를 쓴다.
  $p_i$ 라는 기호 대신에 ${}_nC_r\,p^r(1-p)^{n-r}$ 이라는 기호를 쓴다.

- 이항분포에서 시그마로 표현된 과정이 실제로 〈평균〉이나 〈분산〉을 구하기 위해서 쓰는 과정이 아니다.
〈평균과 분산〉은 이미 공식을 통해서 쉽게 구할 수 있다. 즉, 〈이항분포〉에서 시그마 표현은 그 의미를 읽어준 다음 결국 공식을 쓴다. 역시 〈표를 들여다보는 느낌〉으로 식을 이해한다면 〈식의 의미〉를 헷갈리지 않을 것이다.

- $\sum r\times{}_nC_rp^r(1-p)^{n-r}\ =\ \sum x_i\,p_i\ =\ E(X)\ =\ np$

  ⇨ 독립시행의 확률 ${}_nC_r\,p^r(1-p)^{n-r}$을 통해서 $B(n,\ p)$임을 안다.

- $\sum r^2\times{}_nC_rp^r(1-p)^{n-r}\ =\ \sum x_i^2\,p_i = E(X^2)\ =\ \{E(X)\}^2\ +\ V(X)\ =\ (np)^2+npq$

- $\sum(ar+b)\times{}_nC_rp^r(1-p)^{n-r} = \sum(ax_i+b)\,p_i\ =\ E(aX+b)\ =\ aE(X)+b\ =\ anp+q$

티칭 이항분포에서 $E(a^X)$과 같이 확률변수가 지수에 있는 경우는 공식을 쓸 수 없다.
대신 기본원리인 〈평균의 정의〉를 이용하여 식을 표현하고 〈이항정리의 계산과정〉을 이용하면 간단히 할 수 있다.

⇨ $E(a^X)\ =\ \sum_{r=0}^{n}a^r\times{}_nC_r\,p^r\,(1-p)^{n-r}\ =\ \sum_{r=0}^{n}{}_nC_r\,(ap)^r\,(1-p)^{n-r}$ <지수법칙> $=\ (ap+1-p)^n$ <이항정리>

# 4. 이항분포의 평균, 분산, 표준편차의 증명

## 1) $B(n,\ p)$에서 $E(X)$의 증명

1 단계) $\sum_{r=0}^{n} r \times {}_{n}C_{r}\, p^{r}\, q^{n-r} = \sum_{r=1}^{n} r \times {}_{n}C_{r}\, p^{r}\, q^{n-r}$ 〈어차피 r =0이면 어차피 0 / 즉, 1부터 시작해도 상관없다.〉

.

2 단계) 조합의 수식적 정의 활용

$$\sum_{r=1}^{n} r \times {}_{n}C_{r}\, p^{r}\, q^{n-r} = \sum_{r=1}^{n} r \times \frac{n!}{r!(n-r)!}\, p^{r}\, q^{n-r}$$ 〈조합의 정의식〉 $$= \sum_{r=1}^{n} \frac{n!}{(r-1)!(n-r)!}\, p^{r}\, q^{n-r}$$ 〈r!에서 r을 약분〉

$$= \sum_{r=1}^{n} \frac{n \times (n-1)!}{(r-1)!(n-r)!}\, p^{r}\, q^{n-r}$$ 〈n!을 n×(n-1)!으로 조작 - 분모의 합이 분자에 있어야 조합으로 바꿀 수 있다.〉

$$= n \sum_{r=1}^{n} {}_{n-1}C_{r-1}\, p^{r}\, q^{n-r}$$ 〈조합의 수식적 정의〉

- (복습) 이항전개식 간단히 하기 : $r=1$부터 $n$까지 대입하면 〈변동수〉는 0부터 끝까지(고정수 $n-1$까지)를 모두 표현

변동수와 같은 지수가 없으므로 $p^{r}$을 $p \times p^{r-1}$으로 변형

$$= n \sum_{r=1}^{n} {}_{n-1}C_{r-1}\, p \times p^{r-1} q^{n-r} = np \sum_{r=1}^{n} {}_{n-1}C_{r-1}\, p^{r-1} q^{n-r}$$ - 지수의 합이 고정수 $n-1$을 맞췄다.

$$= np \times \left( {}_{\boxed{n-1}}C_{0}\, p^{0}\, q^{n-1} + {}_{\boxed{n-1}}C_{1}\, p^{1}\, q^{n-2} + \cdots + {}_{\boxed{n-1}}C_{n-1}\, p^{n-1} q^{0} \right)$$ 〈나열된 식의 구조 관찰 : 연속 합〉

$$= np \times (p+q)^{\boxed{n-1}} = np$$ 〈$p+q=1$이므로〉

## 2) $E(aX+b) = anp+q$의 증명

$$\sum_{r=0}^{n} (ar+b) \times {}_{n}C_{r}\, p^{r}\, q^{n-r} = a \sum_{r=0}^{n} r \times {}_{n}C_{r}\, p^{r}\, q^{n-r} + b \sum_{r=0}^{n} {}_{n}C_{r}\, p^{r}\, q^{n-r}$$ 〈Σ 의 성질〉

$$= anp + b \times 1$$ 〈 - 바로 앞에서 증명한 내용 / - 이항전개식 간단히 하기 or 확률의 총합은 1〉

코칭 위의 과정 중 본인에게 당연한 내용은 생략될 수 있다. 위의 증명과정이 시험에 자주 등장하지는 않으나 내신에서 등장 할 수 있다.

코칭 증명과정에서 중간에 들어가는 회색글씨는 식을 변형하면서 자연스럽게 머릿속에서 드는 생각이다.
수학은 아무리 복잡해 보이는 식이더라도 의미를 알고 이해하면서 쓴다는 생각보다 그리 복잡하지 않다.
특히 자연수 · 정수 범위의 공식들은 암기가 아니라 〈Reading〉의 느낌으로 공부해야 한다는 것은 이미 〈촌철살인 확률과 통계 1 : 경우의 수〉에서도 다룬 바 있다.

코칭 이 증명과정과 아래 나오는 분산의 증명과정을 학습함으로서 〈평균, 분산의 시그마 표현〉이 조금 더 익숙해짐을 기대할 수 있다. 또한 〈엄밀한 식 변형 논리〉에 대한 사고력 증진도 기대하고 있다.

식이 복잡해 보인다고 피하지 말라.

코칭 표준편차의 공식 $\sigma(X) = \sqrt{npq}$과 $\sigma(aX+b) = |a|\sqrt{np(1-p)}$은 바로 뒤에 나오는 분산의 공식에 $\sqrt{\ }$를 취한 것뿐이므로 증명을 생략한다.

## 3) $B(n,\ p)$에서 $V(X)$의 증명

1 단계) $V(X) = E(X^2) - \{E(X)^2\}$에서 $E(X) = np$임은 이미 증명되었으므로 $E(X^2)$을 표현하고 간단히 한다.

$$E(X^2) = \sum_{r=0}^{n} r^2 \times {}_nC_r p^r q^{n-r} = \sum_{r=1}^{n} r^2 \times {}_nC_r p^r q^{n-r}$$ 〈어차피 r =0이면 어차피 0 / 즉, 1부터 시작해도 상관없다.〉

2 단계) 조합의 수식적 정의 활용

$$\sum_{r=1}^{n} r^2 \times {}_nC_r p^r q^{n-r} = \sum_{r=1}^{n} r^2 \times \frac{n!}{r!(n-r)!} p^r q^{n-r} = \sum_{r=1}^{n} r \times \frac{n!}{(r-1)!(n-r)!} p^r q^{n-r}$$ 〈조합의 정의식 r의 약분〉

$$= \sum_{r=1}^{n} \{(r-1)+1\} \frac{n!}{(r-1)!(n-r)!} p^r q^{n-r}$$ 〈식 변형 - 이미 경우의 수에서 다룬 바 있다.〉

$$= \sum_{r=1}^{n} \left\{(r-1) \times \frac{n!}{(r-1)!(n-r)!} p^r q^{n-r} + 1 \times \frac{n!}{(r-1)!(n-r)!} p^r q^{n-r}\right\}$$ 〈분배〉

$$= \sum_{r=1}^{n} (r-1) \frac{n!}{(r-1)!(n-r)!} p^r q^{n-r} + \sum_{r=1}^{n} \frac{n!}{(r-1)!(n-r)!} p^r q^{n-r}$$ 〈$\Sigma$ 의 성질〉

$$= \sum_{r=2}^{n} (r-1) \frac{n!}{(r-1)!(n-r)!} p^r q^{n-r} + \sum_{r=1}^{n} \frac{n!}{(r-1)!(n-r)!} p^r q^{n-r}$$ 〈어차피 r =1이면 어차피 0 / 즉, 2부터 시작해도 상관없다.〉

$$= \sum_{r=2}^{n} \frac{n!}{(r-2)!(n-r)!} p^r q^{n-r} + \sum_{r=1}^{n} \frac{n!}{(r-1)!(n-r)!} p^r q^{n-r}$$ 〈$(r-1)$ 약분〉

$$= n(n-1) \sum_{r=2}^{n} \frac{(n-2)!}{(r-2)!(n-r)!} p^r q^{n-r} + n \sum_{r=1}^{n} \frac{(n-1)!}{(r-1)!(n-r)!} p^r q^{n-r}$$ 〈조합으로 식 변형 - 분모의 합이 분자에 있도록〉

$$= n(n-1) \sum_{r=2}^{n} {}_{n-2}C_{r-2}\, p^r q^{n-r} + n \sum_{r=1}^{n} {}_{n-1}C_{r-1}\, p^r q^{n-r}$$ 〈조합으로 식 변형〉

$$= n(n-1)p^2 \sum_{r=2}^{n} {}_{\boxed{n-2}}C_{r-2}\, p^{r-2} q^{n-r} + np \sum_{r=1}^{n} {}_{\boxed{n-1}}C_{r-1}\, p^{r-1} q^{n-r}$$ 〈이항전개식 간단히 하기 - 변동수와 같은 지수가 있도록〉

$$= n(n-1)p^2 \times (p+q)^{\boxed{n-2}} + np \times (p+q)^{\boxed{n-1}}$$ 〈나열된 식의 구조를 매개로 결과를 떠올리는 것이 좋다. 나열된 식은 생략한다.〉

$$= n(n-1)p^2 + np$$

3 단계) 결국 $V(X) = E(X^2) - E(X)^2 = n(n-1)p^2 + np - (np)^2$

$$= \cancel{n^2p^2} - np^2 + np - \cancel{(np)^2} = np(1-p) = npq$$ 〈$1-p$를 보통 $q$라 한다.〉

## 4) $V(aX+b) = a^2(npq)$의 증명

1 단계) $V(X) = E((aX+b)^2) - \{E(aX+b)\}^2$에서 $\{E(aX+b)\}^2 = \{a(np) + b\}^2$임은 이미 증명되었으므로 $E((aX+b)^2)$을 표현하고 간단히 한다.

2 단계) $E((aX+b)^2) = \sum_{r=0}^{n} (ar+b)^2 \times {}_nC_r p^r q^{n-r} = \sum_{r=1}^{n} (a^2r^2 + 2abr + b^2) \times {}_nC_r p^r q^{n-r}$

$$= a^2 \sum_{r=1}^{n} r^2 \times {}_nC_r p^r q^{n-r} + 2ab \sum_{r=1}^{n} r \times {}_nC_r p^r q^{n-r} + b^2 \sum_{r=1}^{n} {}_nC_r p^r q^{n-r}$$ 〈$\Sigma$ 의 성질〉

$$= a^2\{n(n-1)p^2 + np\} + 2ab(np) + b^2 \times 1$$

3 단계) $V(X) = E((aX+b)^2) - \{E(aX+b)\}^2 = a^2\{n(n-1)p^2 + np\} + 2ab(np) + b^2 - \{a(np) + b\}^2$

$$= a^2\{n(n-1)p^2 + np\} - a^2(np)^2 = a^2np(1-p) = a^2npq$$

**103** 이해를 위한 예제

확률변수가 $X$인 이항분포 $B\left(90,\ \frac{1}{3}\right)$에서 $E(X)$, $V(X)$, $\sigma(X)$, $E(X^2)$을 구하여라.

증명은 복잡했지만 공식은 쉽다.

간단히 공식을 이용하면 $E(X) = 90\times\frac{1}{3} = 30$, $V(X) = 90\times\frac{1}{3}\times\frac{2}{3} = 20$, $\sigma(X) = \sqrt{20}$ 이다.

그리고 $V(X) = E(X^2) - \{E(X)\}^2$ 이므로 $E(X^2) = \{E(X)\}^2 + V(X) = 30^2 + 20 = 920$

$E(X) = 30$, $V(X) = 20$, $\sigma(X) = \sqrt{20}$, $E(X^2) = 920$

**104** 이해를 위한 예제 2011. 11. 가형(94%). 3번. 2점

확률변수 $X$가 이항분포 $B(200.\ p)$를 따르고 $X$의 평균이 40일 때, $X$의 분산은?

① 32 ② 33 ③ 34 ④ 35 ⑤ 36

증명은 복잡했지만 공식은 쉽다.

- 확률변수 $X$가 이항분포 $B(200.\ p)$를 따르고 $X$의 평균이 40일 때 ⇨ $X$의 평균은 $np = 200p = 40 \Leftrightarrow p = \frac{1}{5}$

- $X$의 분산 ⇨ $npq = 200\times\frac{1}{5}\times\frac{4}{5} = 32$

정답은 ①

**105** 이해를 위한 예제 2006. 9. 나형(65%). 29번. 4점

이산확률변수 $X$가 값 $x$를 가질 확률이

$$P(X=x) = {}_nC_x\, p^x (1-p)^{n-x} \text{ (단, } x=0,\ 1,\ 2,\ \cdots,\ n\text{이고 } 0<p<1\text{이다.)}$$

$E(X)=1$, $V(X)=\frac{9}{10}$일 때, $P(X<2)$의 값은?

① $\frac{19}{10}\left(\frac{9}{10}\right)^9$ ② $\frac{17}{9}\left(\frac{8}{9}\right)^8$ ③ $\frac{15}{8}\left(\frac{7}{8}\right)^7$ ④ $\frac{13}{7}\left(\frac{6}{7}\right)^6$ ⑤ $\frac{11}{6}\left(\frac{5}{6}\right)^5$

독립시행의 확률을 통해 이항분포임을 확신한다.

- $P(X=x) = {}_nC_x\, p^x (1-p)^{n-x}$ ⇨ 독립시행의 확률에서 <고정수>가 총 시행횟수를 의미하고, <변동수>와 같은 지수를 가지고 있는 항의 <밑>이 1회 시행에서 발생할 확률을 의미한다. 즉, $B(n,\ p)$이다.

- $E(X) = 1$, $V(X) = \frac{9}{10}$ ⇨ 평균을 표현하면 $np = 1$이고, 분산을 표현하면 $np(1-p) = \frac{9}{10}$이다.

이 두 식을 연립하면 $p = \frac{1}{10}$ 즉, $n = 10$

- $P(X<2) = P(0) + P(1) = {}_{10}C_0\left(\frac{1}{10}\right)^0\left(\frac{9}{10}\right)^{10} + {}_{10}C_1\left(\frac{1}{10}\right)^1\left(\frac{9}{10}\right)^9$

$$= \left(\frac{9}{10}\right)^{10} + \left(\frac{9}{10}\right)^9 = \left(\frac{9}{10}\right)\times\left(\frac{9}{10}\right)^9 + \left(\frac{9}{10}\right)^9 = \frac{19}{10}\left(\frac{9}{10}\right)^9$$

정답은 ①

**106** 이해를 위한 예제 2006. 4. 가형(86%). 21번. 3점

어떤 책을 임의로 펼쳤을 때, 그림이 나올 확률이 $\frac{1}{3}$이라고 한다. 이 책을 임의로 180번 펼쳐 그림이 나오는 횟수를 $X$라고 할 때, $X$의 분산을 구하시오.

1회 시행 어떤 책을 임의로 펼쳤을 때, 그림이 나올 확률이 $\frac{1}{3}$

⇨ 즉, 그림이 나오는 경우를 사건 $A$라 하면 $P(A) = \frac{1}{3}$

독립시행 이 책을 임의로 180번 펼쳐 그림이 나오는 횟수를 $X$

⇨ $P(X=r) = {}_{180}C_r\left(\frac{1}{3}\right)^r\left(\frac{2}{3}\right)^{180-r}$ 이므로 확률변수 $X$는 $B\left(180,\ \frac{1}{3}\right)$을 따른다.

즉, $V(X) = npq = 180\times\frac{1}{3}\times\frac{2}{3} = 40$

정답은 40

**107** 이해를 위한 예제 2008. 4. 가형(60%). 19번. 3점

한 번의 시행에서 일어날 확률이 $\frac{1}{4}$인 사건 $A$가 있다. 80번의 독립시행에서 사건 A가 일어나는 횟수를 확률변수 $X$라 할 때, $X^2$의 평균 $\mathrm{E}(X^2)$을 구하시오.

1회 시행 한 번의 시행에서 일어날 확률이 $\frac{1}{4}$인 사건 $A$가 있다. ⇨ 즉, $P(A) = \frac{1}{3}$

독립시행 80번의 독립시행에서 사건 A가 일어나는 횟수를 확률변수 $X$ ⇨ 확률변수 $X$는 $B\left(80,\ \frac{1}{4}\right)$을 따른다.

즉, $V(X) = E(X^2) - \{E(X)\}^2$에서 $E(X^2) = V(X) + \{E(X)\}^2$이므로

$E(X^2) = V(X) + \{E(X)\}^2 = npq + (np)^2 = 80\times\frac{1}{4}\times\frac{3}{4} + \left(80\times\frac{1}{4}\right)^2 = 15 + (20)^2 = 415$

정답은 415

**108** 이해를 위한 예제 2005. 11. 나형(73%). 5번. 3점

확률변수 $X$가 이항분포 $B\left(100,\ \frac{1}{5}\right)$을 따를 때, 확률변수 $3X-4$의 표준편차는?

① 12 ② 15 ③ 18 ④ 21 ⑤ 24

간단한 공식과 성질을 이용한다.

$B\left(100,\ \frac{1}{5}\right)$에서 $V(X) = 100 \times \frac{1}{5} \times \frac{4}{5} = 16$ 이므로 $\sigma(X) = 4$이다.

즉, $\sigma(3X-4) = 3\sigma(X) = 12$

정답은 ①

**109** 이해를 위한 예제 2005. 사관학교. 문·이과. 29번. 4점

이산확률변수 $X$ 는 이항분포 $B\left(120,\ \frac{1}{121}\right)$을 따른다. 함수 $f(x)=\sum_{k=0}^{120}(x-ak)^2P(X=k)$ 의 최솟값이 1 이 되도록 하는 양수 $a$ 에 대하여 $120a$ 의 값을 구하시오.

조건 함수 $f(x)=\sum_{k=0}^{120}(x-ak)^2P(X=k)$ 의 최솟값이 1 이 되도록 하는 양수 $a$

- 주어진 함수는 $x$에 대한 함수이므로 우변의 식을 $x$에 대해서 정리해야 한다.

$$\sum_{k=0}^{120}(x-ak)^2P(X=k) = \sum_{k=0}^{120}(x^2-2ax\,k+a^2k^2)\cdot P(X=k) = x^2 - 2ax\sum_{k=0}^{120}k\cdot P(X=k) + a^2\sum_{k=0}^{120}k^2\cdot P(X=k)$$

- 그런데 여기에서 확률변수 $X$ 는 이항분포 $B\left(120,\ \frac{1}{121}\right)$를 따르므로

$$\Rightarrow \sum_{k=0}^{120}k\cdot P(X=k) = E(X) = 120\cdot\frac{1}{121} = \frac{120}{121}$$

$$\Rightarrow \sum_{k=0}^{120}k^2\cdot P(X=k) = E(X^2) = V(X)+\{E(X)\}^2 = 120\cdot\frac{1}{121}\cdot\frac{120}{121}+\left(\frac{120}{121}\right)^2 = 2\left(\frac{120}{121}\right)^2$$ 이다.

즉, $f(x) = x^2-2\left(\frac{120}{121}\right)ax+2\left(\frac{120}{121}\right)^2a^2$이다.

구할 것 최솟값이 1 이 되도록 하는 양수 $a$

- 주어진 이차함수 $f(x) = x^2-2\left(\frac{120}{121}\right)ax+2\left(\frac{120}{121}\right)^2a^2$는 <대칭축 $x = \left(\frac{120}{121}\right)a$>에서 최솟값 1을 가진다.

이것을 통해서 관계식을 만들면 $f\left(\left(\frac{120}{121}\right)a\right) = \left(\frac{120}{121}\right)^2a^2 = 1$이므로 <양수 $a$>는 $\frac{121}{120}$이다.

$120a = 121$이므로 정답은 121

# 예제를 통해 개념의 외연을 넓혀라



| 개념의 외연 | |
|---|---|
| | #1. 표 작성 문제는 확률문제 |
| | #2. 이항분포의 확신 |
| | #3. 이항정리의 연산 - 이항전개식 간단히 하기 |
| | #4. 큰 수의 법칙 |

〈개념의 외연〉이라는 부분은 현재 배우고 있는 경우의 수의 국한된 내용만 다루는 것은 아니다.
현재 단원에 해당하는 문제를 풀기 위해서 필요한 수학적 개념이나 유형에 필요한 사고방식을 두루 다룬다.

수학에는 순서가 없다. 하지만 배움에는 순서가 있다.

그래서 순서가 없는 수학에 순서를 정한 것이 교육과정이다. 어떤 내용은 너무 어렵거나 복합적인 생각들을 담고 있어서 여러 학년에 걸쳐 배운다. 이런 경우 보통은 문제를 통해서 익힌다는 느낌이 강한데 이런 개념들을 집약적으로 정리해 주는 부분이 될 것이다.

## #1 표 작성 문제는 확률문제

- 표가 주어진 상태로 〈평균, 분산, 표준편차〉을 구하는 것은 공식만 정확하게 알고 있으면 어렵지 않다.
  하지만 우리가 표를 완성해야 하는 상황이라면 결국 〈확률〉을 정확히 구할 수 있어야 한다.

**110** 이해를 위한 예제 2010. 3. 가형(43%). 23번. 4점

그림과 같이 숫자 $1,\ 2,\ 3$이 각각 하나씩 적혀 있는 흰 공 $3$개와 검은 공 $3$개가 들어있는 주머니가 있다. 이 주머니에서 임의로 $2$개의 공을 동시에 꺼낼 때, 꺼낸 공에 적혀 있는 숫자의 최솟값을 확률변수 $X$라 하자. $X$의 평균이 $\dfrac{q}{p}$일 때, $p+q$의 값을 구하시오. (단, $p$, $q$는 서로소인 자연수이다.)

시행 임의로 2개의 공을 동시에 꺼낼 때, ⇨ 시행에 따른 모든 경우의 수는 ${}_6C_2 = 15$
확률변수 꺼낸 공에 적혀 있는 숫자의 최솟값을 확률변수 $X$
⇨ 가능한 모든 결과는 $X = 1,\ 2,\ 3$이다. (3번 흰 공과 3번 검은 공이 나온다면 최솟값은 3이 된다.)

| $X$ | 1 | 2 | 3 | 합계 |
|---|---|---|---|---|
| $P(X=x)$ | | | | 1 |

⇨

| $X$ | 1 | 2 | 3 | 합계 |
|---|---|---|---|---|
| $P(X=x)$ | $\frac{9}{15}$ | $\frac{5}{15}$ | $\frac{1}{15}$ | 1 |

- $n(\text{최솟값}\,1) = n(X=1) = n(1\,\text{두 개}) + n(1\,\text{한 개},\ 2\text{와}\,3\text{중 한 개}) = 1 + {}_2C_1 \times {}_4C_1 = 9$ ⇨ $P(X=1) = \dfrac{9}{15}$
- $n(\text{최솟값}\,2) = n(X=2) = n(2\,\text{두 개}) + n(2\,\text{한 개},\ 3\text{한 개}) = 1 + {}_2C_1 \times {}_2C_1 = 5$ ⇨ $P(X=2) = \dfrac{5}{15}$
- $n(\text{최솟값}\,3) = n(X=3) = n(3\,\text{두 개}) = 1$ ⇨ $P(X=3) = \dfrac{1}{15}$ 위와 같이 표를 완성한다.

$E(X) = \sum_{i=1}^{n} x_i p_i = 1\times\dfrac{9}{15} + 2\times\dfrac{5}{15} + 3\times\dfrac{1}{15} = \dfrac{22}{15}$이다. 즉, $p+q = 37$

코칭 확률의 총합이 $1$인지 확인하여 본인이 확률은 제대로 구했는지 확인한다.

**111** 이해를 위한 예제  2010. 4. 가형(55%). 10번. 4점

표는 세 개의 주사위를 던져서 나온 눈의 수들 중에서 두 수의 차의 최댓값을 확률변수 $X$라 할 때, 확률변수 $X$의 확률분포표이다. 이때, 확률변수 $Y=12X+5$의 평균 $E(Y)$의 값은?

| $X$ | 0 | 1 | 2 | 3 | 4 | 5 | 계 |
|---|---|---|---|---|---|---|---|
| $P(X=x)$ | $\frac{1}{36}$ | $a$ | $\frac{2}{9}$ | $b$ | $\frac{2}{9}$ | $\frac{5}{36}$ | 1 |

① 40 ② 44 ③ 48 ④ 52 ⑤ 56

시행 세 개의 주사위를 던져서 나온 눈의 수 ⇨ 시행에 따른 모든 경우의 수는 $6\times6\times6 = 6^3$

확률변수 두 수의 차의 최댓값을 확률변수 $X$라 할 때, ⇨ 나올 수 있는 모든 결과는 이미 표로 주어져 있다.

| $X$ | 0 | 1 | 2 | 3 | 4 | 5 | 계 |
|---|---|---|---|---|---|---|---|
| $P(X=x)$ | $\frac{1}{36}$ | $a$ | $\frac{2}{9}$ | $b$ | $\frac{2}{9}$ | $\frac{5}{36}$ | 1 |

- 위의 표에서 $\sum_{i=1}^{n} p_i = 1$이므로 $a$, $b$에 대한 관계식을 하나 얻을 수 있다.

남은 하나는 주어진 상황에서 $P(X=1)$를 진짜로 구해야 한다.

사건의 분석 <두 수의 차의 최댓값이 1이 나오는 상황>은 <세 수 중 두 수는 같고 남은 한 수는 1만 차이가 나는 상황>이다. 아래와 같이 모든 상황을 일일이 따지는 것이 어렵지 않다.

| | | | | |
|---|---|---|---|---|
| 112 | 223 | 334 | 445 | 556 |
| 122 | 233 | 344 | 455 | 566 |

: 이 각각의 경우를 배열하는 3가지의 경우를 고려

$n(\text{차의 최댓값 }1) = 10\times3$이므로 $P(\text{차의 최댓값 }1) = P(X=1) = \frac{30}{6\times6\times6} = \frac{5}{36}$

- 또한 표에서 먼저 $\sum p_i = 1$을 이용해서 $b$를 구한다.

⇨ $\frac{1}{36} + \frac{5}{36} + \frac{2}{9} + b + \frac{2}{9} + \frac{5}{36} = 1 \quad \Leftrightarrow \quad b = \frac{1}{4}$ 즉, 표를 완성하면 다음과 같다.

| $X$ | 0 | 1 | 2 | 3 | 4 | 5 | 계 |
|---|---|---|---|---|---|---|---|
| $P(X=x)$ | $\frac{1}{36}$ | $\frac{5}{36}$ | $\frac{2}{9}$ | $\frac{1}{4}$ | $\frac{2}{9}$ | $\frac{5}{36}$ | 1 |

- $E(X) = \sum_{i=1}^{n} x_i p_i = 0\times\frac{1}{36} + 1\times\frac{5}{36} + 2\times\frac{2}{9} + 3\times\frac{1}{4} + 4\times\frac{2}{9} + 5\times\frac{5}{36} = \frac{35}{12}$

- $E(Y) = E(12X+5) = 12E(X)+5 = 12\times\frac{35}{12}+5 = 40$

정답은 ①

코칭 <두 수의 차의 최댓값이 1이 나오는 상황>을 <세 수 중 두 수는 같고 남은 한 수는 1만 차이가 나는 상황>으로 정리하는 과정에서 사용된 사고가 바로 추론이다.

즉, 몇 가지 구체적인 숫자를 두어 <두 수의 차의 최댓값이 1이 나오는 상황>은 <세 수 중 두 수는 같고 남은 한 수는 1만 차이가 나는 상황>일 수밖에 없음을 확신하는 과정이 필요하다.

코칭 위에서는 <세 수 중 두 수는 같고 남은 한 수는 1만 차이가 나는 상황>의 경우의 수를 구하기 위해 나올 수 있는 모든 경우를 일일이 나열하여 보여줬다. 머릿속에 나올 수 있는 모든 경우에 대한 확신이 있다면 곱의 법칙을 이용하여 $5\times2\times3$라고 바로 풀어도 될 것이다. 하지만 이것은 중요한 문제가 아니다.

## #2 이항분포의 확신

- 〈독립시행의 확률〉을 통해 이항분포와 관련된 문제임을 확신할 수 있고, $B(n,\ p)$에서 $n,\ p$의 값을 결정할 수 있다.

- 이항분포에서 확률변수는 변형될 수 있으나 〈확률 값〉이 〈독립시행의 확률〉로 표현된다는 사실은 변할 수 없다.
  즉, 상황이 조건으로 주어져 이항분포를 활용하는 문제임을 판단하는 과정에서
  〈시행에 따른 모든 가능한 결과〉인 확률변수가 〈$0,\ 1,\ 2,\ \cdots,\ n$〉의 형태로 표현되지 않더라도
  확률 값이 ${}_nC_r p^r(1-p)^{n-r}$의 형태로 표현된다면 이항분포와 관련된 문제임을 알 수 있다.

**112** 이해를 위한 예제

형원, 유원, 성준 세 사람과 ♠, ♥, ♣가 새겨진 같은 모양의 카드가 각각 1장씩 있다. 성준은 세 장의 카드를 모양이 보이지 않도록 뒤집어 마음대로 섞고, 형원이가 세 장의 카드 중 하나를 선택하여 뒤집는다. 뒤집은 카드가 ♥ 모양일 때, 형원이는 유원이에게 300원을 받고, 다른 모양이면 형원이는 유원이에게 200원을 주기로 하였다. 이 게임을 90번 반복했을 때, 유원이가 받을 수 있는 금액의 기댓값은? (단, 뒤집어진 세 카드는 서로 모양을 구분할 수 없다.)

① 2500원 ② 2600원 ③ 2800원 ④ 2900원 ⑤ 3000원

시행 이 게임을 90번 반복했을 때, ⇨ 동일 시행을 90번 반복하는 것에서 이항분포라는 힌트를 얻을 수 있다.

확률변수 유원이가 받을 수 있는 금액의 기댓값 ⇨ 금액을 확률변수로 하여 표를 작성한다.

- 형원이가 뒤집은 카드가 90번 모두 ♥이라면 유원이가 받는 금액은 $-27000$원이다.
  또한 형원이가 뒤집은 카드가 89번이 ♥, 1번이 다른 카드라면 유원이가 받는 금액은 $-26500$원이다.
  이렇게 가능한 모든 경우를 확률변수로 하여 표를 작성한다.

| $X$ | $-27000$ | $-26500$ | $\cdots$ | $18000$ | 계 |
|---|---|---|---|---|---|
| $P(X=x)$ | | | | | 1 |

- 1회의 시행에서 유원이가 돈을 받는 사건을 $A$라 할 때 $P(A)=\dfrac{2}{3}$이다.

- ( ) : 유원이가 $-27000$의 상태가 될 확률은 90회의 독립시행 중 사건 $A$가 0회($A^c$이 90회) 발생할 확률과 같다.
  즉, 유원이가 형원이에게 300원씩 90회 주는 상황이므로 유원이가 받는 금액을 $-27000$라고 표현 할 수 있다.
  이 상황은 〈독립시행의 확률〉을 구하는 상황과 같으므로 ( )에 들어갈 확률은 ${}_{90}C_0\left(\dfrac{2}{3}\right)^0\left(\dfrac{1}{3}\right)^{90}$이다.

- ( ) : 유원이가 $-26500$의 상태가 될 확률은 90회의 독립시행 중 사건 $A$가 1회($A^c$이 89회) 발생할 확률과 같다.
  이 상황은 (독립시행의 확률)을 구하는 상황과 같으므로 ( )에 들어갈 확률은 ${}_{90}C_1\left(\dfrac{2}{3}\right)^1\left(\dfrac{1}{3}\right)^{89}$이다.

⇨ 여기까지에서 우리는 사건이 발생할 확률이 〈독립시행의 확률〉이라는 것을 알았다.
  즉, 어떤 방식으로든 〈이항분포 : $\mathrm{B}\left(90,\ \dfrac{2}{3}\right)$〉의 공식을 이용해서 $E(X)$를 구한다는 것을 생각해야 한다.
  (이렇게 추론을 통해서 감을 잡은 후 일반화 시키는 것이 수학이다. 스스로 자연스럽게 과정을 생략할 수 없다면 생략하지 말길 바란다.)

- 유원이가 게임에서 이긴 횟수를 $Y$라고 할 경우 확률변수 $Y$는 이항분포 $\mathrm{B}\left(90,\ \dfrac{2}{3}\right)$를 따르므로 $E(Y)=90\times\dfrac{2}{3}=60$
  확률변수 $X$는 $500Y-27000$이므로 우리가 구하고자 하는 금액의 기댓값은 $E(X)$이므로

$$E(X)=E(500Y-27000)=500E(Y)-27000\text{(평균의 성질)}=500\times60-27000=3000$$

정답은 ⑤

티칭 유원이가 $0$회를 이긴 경우 $Y=0$이고, 그것에 대응하는 금액 $X$는 $-27000$이다.
유원이가 $1$회를 이긴 경우 $Y=1$이고, 그것에 대응하는 금액 $X$는 $-26500$이다.

$\vdots$

유원이가 $1$회를 이긴 경우 $Y=r$이고, 그것에 대응하는 금액 $X$는 $500r-27000$이다.

$\vdots$

유원이가 $90$회를 이긴 경우 $Y=90$이고, 그것에 대응하는 금액 $X$는 $18000$이다.
즉, $X = 500Y-27000$ ($Y = 0, 1, 2, \cdots, 90$)이고 확률변수 $Y$는 이항분포를 따르게 된다.
- 일반항을 구하는 과정에서 〈등차수열〉은 〈초항〉이 $n=1$에 대응하는 값을 가지지만
〈이항분포〉는 〈초항〉이 $r=0$에 대응하는 값을 가지므로 주의해야 한다.

코칭 앞에서 공부했던 〈평균, 분산, 표준편차의 성질〉은 $X$에 대응하는 $aX+b$의 〈평균, 분산, 표준편차〉를 찾는 과정이었다. 하지만 이 문제는 $aX+b$에 대한 상황이 제시되고, 이것에 대응하는 $X$의 분포를 이용해야 한다.
문제를 푸는 방식은 어렵지 않으나 처음 공부할 때에는 애먹는 경우가 종종 있는데, 문제를 통해서 푸는 방식을 다시 한 번 상기하고 넘어가도록 하자.

## #3 이항정리의 연산

이항전개식 간단히 하기

- <경우의 수 : 이항정리>와 <확률 : 독립시행의 확률> 그리고 <통계 : 이항분포>는 하나로 연결된 개념이다.
- <촌철살인 확률과 통계 1>의 내용 중 일부를 발췌하여 복습한다.

# 1. 이항전개식 간단히 하기

: 공식 전체 구조를 통째로 외워야 한다. 완벽하게 식의 구조적 특징을 기억한다.

### 1) 첫 번째 전개식 : <조합 × 등비수열>이 규칙적으로 더해진 모양 1

- (관찰1) 조합을 보고 힌트를 얻는다. (조합의 하단의 수 ( ) 은 건드리지 않는다.)

$$\sum_{r=0}^{n} {}_nC_r a^{n-r} b^r = {}_nC_0\, a^n b^0 + {}_nC_1\, a^{n-1} b^1 + \cdots + {}_nC_r\, a^{n-r} b^r + \cdots + {}_nC_n a^0 b^n = (a+b)^n$$

전개식을 간단히 만들기 위해 ( ) 부분에 맞게 ( )을 조작한다.

- (관찰2) 어떻게 맞춰주는가?

편의상 ( )의 왼쪽 부분을 <고정수>, 오른쪽 부분을 <변동수>라고 하자.

$$\sum_{r=0}^{n} {}_{\underset{\text{고정수}}{n}}C_{\underset{\text{변동수}}{r}} a^{n-r} b^r$$

| | |
|---|---|
| (1) 변동수 확인 ⇨ <변동수>는 항상 0부터 끝까지(고정수까지) 더한다. | (빠진 게 있으면 채운다.) |
| $\sum_{r=0}^{n} {}_nC_r a^{n-r} b^r = {}_nC_0 a^n b^0 + {}_nC_1 a^{n-1} b^1 + \cdots + {}_nC_r a^{n-r} b^r + \cdots + {}_nC_n a^0 b^n = (a+b)^n$ | |
| (2) 지수 확인 ⇨ <변동수>와 <같은 지수>가 있는지 확인한다. | (없으면 곱하고 나눠서 있게 한다.) |
| $\sum_{r=0}^{n} {}_nC_r a^{n-r} b^r = {}_nC_0 a^n b^0 + {}_nC_1 a^{n-1} b^1 + \cdots + {}_nC_r a^{n-r} b^r + \cdots + {}_nC_n a^0 b^n = (a+b)^n$ | |
| (3) 지수 합 확인 ⇨ <지수의 합>은 <고정수>이다. | (아니라면 곱하고 나눠서 있게 한다.) |
| $\sum_{r=0}^{n} {}_nC_r a^{n-r} b^r = {}_nC_0 a^n b^0 + {}_nC_1 a^{n-1} b^1 + \cdots + {}_nC_r a^{n-r} b^r + \cdots + {}_nC_n a^0 b^n = (a+b)^n$ | |

- 보통 (2)과 (3)은 동시에 맞춰진다. ⇨ 결과적으로 $(a+b)^{\text{고정수}}$로 간단하게 나타내어진다.

$$\sum_{r=0}^{n} {}_nC_r a^{n-r} b^r = {}_nC_0\, a^n b^0 + {}_nC_1\, a^{n-1} b^1 + \cdots + {}_nC_r\, a^{n-r} b^r + \cdots + {}_nC_n\, a^0 b^n = (a+b)^n$$

| 외워라!!!<br><조합 X 등비수열>이 규칙적으로 더해진 모양 | 변동수 확인 : 0부터 끝까지<br>지수 확인 : 변동수와 같은 지수<br>지수 합 확인 : 지수 합 = 고정수 | ⇨ $(a+b)^{\text{고정수}}$ |
|---|---|---|

## 2) 두 번째 전개식 : <조합 × 등비수열>이 규칙적으로 더해진 모양 2

- (관찰1) 조합을 보고 힌트를 얻는다. (조합의 하단의 수 ▬ 은 건드리지 않는다.)

$$\sum_{r=0}^{n} {}_nC_r\, x^r = {}_nC_0 + {}_nC_1\, x + \cdots + {}_nC_r\, x^r + \cdots + {}_nC_n x^n = (1+x)^n$$

전개식을 간단히 만들기 위해 ▬ 부분에 맞게 ▬을 조작한다.

- (관찰2) 어떻게 맞춰주는가?

편의상 ▬의 왼쪽 부분을 <고정수>, 오른쪽 부분을 <변동수>라고 하자.

$$\boxed{\sum_{r=0}^{n} {}_{\underset{\text{고정수}}{n}}C_{\underset{\text{변동수}}{r}}\, x^r}$$

(1) 변동수 확인 ⇨ <변동수>는 항상 0부터 끝까지(고정수까지) 더한다. (빠진 게 있으면 채운다.)

$$\sum_{r=0}^{n} {}_nC_r\, x^r = {}_nC_0 + {}_nC_1\, x + \cdots + {}_nC_r\, x^r + \cdots + {}_nC_n x^n = (1+x)^n$$

(2) 지수 확인 ⇨ <변동수>와 <같은 지수>가 있는지 확인한다. (없으면 곱하고 나눠서 있게 한다.)

$$\sum_{r=0}^{n} {}_nC_r\, x^r = {}_nC_0 + {}_nC_1\, x^1 + \cdots + {}_nC_r\, x^r + \cdots + {}_nC_n x^n = (1+x)^n$$

(3) 보이지 않는 1의 존재를 까먹으면 안 된다.

$$\sum_{r=0}^{n} {}_nC_r\, x^r = {}_nC_0\, 1^n + {}_nC_1\, x^1 1^{n-1} + \cdots + {}_nC_r\, x^r 1^{n-r} + \cdots + {}_nC_n x^n 1^0 = (1+x)^n$$

$$\sum_{r=0}^{n} {}_nC_r\, x^r = {}_nC_0 + {}_nC_1\, x + \cdots + {}_nC_r\, x^r + \cdots + {}_nC_n x^n = (1+x)^n$$

<table>
<tr><td rowspan="3">외워라!!!<br><조합 X 등비수열>이 규칙적으로 더해진 모양</td><td>변동수 확인 : 0부터 끝까지</td><td rowspan="3">⇨ $(1+x)^{\text{고정수}}$</td></tr>
<tr><td>지수 확인 : 변동수와 같은 지수</td></tr>
<tr><td>보이지 않는 1의 존재</td></tr>
</table>

**113** 이해를 위한 예제 2006. 사관학교. 문과. 15번. 4점

사건 $A$가 $1$회의 시행에서 일어날 확률이 $p$일 때, $n$회의 독립시행에서 사건 $A$가 일어나는 횟수를 확률변수 $X$라 하자. 확률변수 $X$의 평균이 $80$이고 분산이 $64$라 할 때, $\sum_{r=0}^{n} 5^r \mathrm{P}(X=r)$의 값은? (단, $\mathrm{P}(X=r)$은 $X=r$일 때의 확률이다.)

① $\left(\frac{9}{5}\right)^{400}$ ② $\left(\frac{7}{5}\right)^{450}$ ③ $\left(\frac{9}{5}\right)^{399}$ ④ $2^{399}$ ⑤ $2^{400}$

이항전개식 간단히 하기

- 사건 $A$가 $1$회의 시행에서 일어날 확률이 $p$일 때, $n$회의 독립시행에서 사건 $A$가 일어나는 횟수를 확률변수 $X$으로부터 $E(X) = np = 80$, $V(X) = np(1-p) = 64$ ⇨ $p=\frac{1}{5}$, $n=400$ 이다.
- 따라서 확률변수 $X$는 $B\left(400, \frac{1}{5}\right)$을 따르고, $P(X=r) = {}_{400}C_r \left(\frac{1}{5}\right)^r \left(\frac{4}{5}\right)^{400-r}$ 이 된다.
- $\sum_{r=0}^{n} 5^r P(X=r)$은 $E(5^X)$으로 우리가 앞서 <성질>에서 공부한 식의 모양과 다르므로 <공식으로 한 번에 처리>할 수 없음을 유의한다. <촌철살인 확률과 통계 1. Part 6>에서 <이항전개식 간단히 하기>를 이미 공부한 바 있다.

$$\sum_{r=0}^{n} 5^r P(X=r) = \sum_{r=0}^{n} 5^r \cdot {}_{400}C_r \left(\frac{1}{5}\right)^r \left(\frac{4}{5}\right)^{400-r} = \sum_{r=0}^{n} {}_{400}C_r \left(\frac{5}{5}\right)^r \left(\frac{4}{5}\right)^{400-r}$$

- $5^r \times \left(\frac{1}{5}\right)^r$을 $\left(\frac{5}{5}\right)^r$으로 표현했다.

$$= \sum_{r=0}^{n} {}_{400}C_r (1)^r \left(\frac{4}{5}\right)^{400-r}$$

⇨ 0부터 끝까지 변동수와 같은 지수 지수의 합이 고정수 5

$$= \left(1+\frac{4}{5}\right)^{400} = \left(\frac{9}{5}\right)^{400}$$

정답은 ①

## #4 큰 수의 법칙

예제없이 식의 의미 정도만 이해하고 넘어간다.

수학에서는 다음과 같은 〈추론식 증명법〉도 있다. 이과는 앞에서 한 번 정도 봤지만 (자연상수 $e$의 정의에서) 문과는 한 번도 못 봤을 것이다. 이항분포를 직접 계산 후 근삿값을 표로 만든 이항분포 표를 이용하여 다음 식을 증명한다. 이것은 교과서에는 있지만 출제될 가능성이 거의 없다.

$$\lim_{n\to\infty} P\left(\left|\frac{X}{n}-p\right|<\epsilon\right) = 1$$

주사위 한 개를 던질 때, 1의 눈이 나오는 횟수 $X$의 이항분포는 $B\left(n,\ \frac{1}{6}\right)$을 따른다.

이 중 $n$을 $10, 30, 50$으로 두어 위의 확률을 계산해 보기로 한다.

우리는 수학적 확률로 알고 있는 $\frac{1}{6}$과 실제 $n$번 던졌을 때, 1이 나오는 눈의 횟수 $X$에 대한 통계적 확률 $\frac{X}{n}$와의 차이 $\left|\frac{X}{n}-\frac{1}{6}\right|$을 계산해 본다. 일단은 $\epsilon = 0.1$로 두고 계산한다.

(1) $n=10$일 때, $\lim_{n\to\infty} P\left(\left|\frac{X}{10}-\frac{1}{6}\right|<0.1\right) = P(0.66\cdots < X < 2.66\cdots) = P(X=1)+P(X=2)$

$= 0.6137$이다. (이것은 표를 찾아서 쓴 것.)

(2) $n=30$일 때, $\lim_{n\to\infty} P\left(\left|\frac{X}{30}-\frac{1}{6}\right|<0.1\right) = P(2 < X < 8) = P(X=3)+P(X=4)+\cdots+P(X=7)$

$= 0.7835$이다. (이것은 표를 찾아서 쓴 것.)

(3) $n=50$일 때, $\lim_{n\to\infty} P\left(\left|\frac{X}{50}-\frac{1}{6}\right|<0.1\right) = P(3.3 < X < 13.3) = P(X=4)+P(X=5)+\cdots+P(X=13)$

$= 0.9455$이다. (이것은 표를 찾아서 쓴 것.)

$P(X=r)$의 값은 이항분포 표를 이용하여 찾아보면 된다. 이처럼 $n$이 커질수록 통계적 확률과 수학적 확률의 차이가 작아진다. 이것은 $\epsilon$을 $0.01$, $0.001$로 두어도 같은 결과가 나온다. (찾는 방법이 중요한 것이 아니므로 표는 생략했다.)

결과적으로 〈수학적 확률〉을 〈통계적 확률〉로 대체할 수 있는 근거가 마련된 것이다. 이제는 100번 중 사건 $A$가 3번 정도 발생했다는 조건을 $P(A)=\frac{3}{100}$ 으로 써도 좋다. 이것의 결과는 중요하지만 어렵지 않고, 과정은 별로 중요하지 않다.

# 공부는 예제를 가지고 하는 거야

촌철살인 경우의 수 | PART 4 이산확률분포와 이항분포

일단 풀고 맞았다고 하더라도 뒤의 해설을 정독하여 읽는다. 왜냐하면 풀이과정이 암기되어서 문제가 그냥 풀리거나 그냥 직관적으로 판단이 되어서 문제가 풀리는 경우도 많기 때문이다. 우리가 문제를 푸는 이유는 앞에서 배운 개념을 확인하기 위해서다. 이 문제를 푸는 것이 중요한 게 아니라 다른 어려운 문제에까지 적용할 수 있는 개념을 익히는 게 중요한 것이다.

**예제 050**

2006. 9. 가형(75%), 나형(69%). 13번. 4점

이산확률변수 $X$의 확률분포 표는 다음과 같다. 집합 $\{x \square 0 \le x \le 10\}$에서 정의된 두 함수 $F(x), G(x)$가

| $X$ | 0 | 1 | 2 | $\cdots$ | 10 | 계 |
|---|---|---|---|---|---|---|
| $P(X=x)$ | $p_0$ | $p_1$ | $p_2$ | $\cdots$ | $p_{10}$ | 1 |

(단, $p_i > 0$이고 $i=0,\ 1, 2,\ \cdots,\ 10$이다.)

$F(x)=P(0 \le X \le x)$, $G(x)=P(X > x)$일 때, 〈보기〉에서 옳은 것을 모두 고른 것은?

〈보 기〉

ㄱ. $G(3)=1-F(3)$ ㄴ. $P(3 \le X \le 8)=F(8)-F(3)$ ㄷ. $P(3 \le X \le 8)=G(2)-G(8)$

① ㄱ ② ㄷ ③ ㄱ, ㄴ ④ ㄱ, ㄷ ⑤ ㄴ, ㄷ

**예제 051**

한 개의 주사위를 두 번 던져서 나오는 눈의 수의 합을 확률변수 $X$라고 할 때, $X$의 확률질량함수를 구하여라.

예제 052

2016. 09. 고3. 가형. 17번. 4점

$1$ 부터 $n$ 까지의 자연수가 하나씩 적혀 있는 $n$ 장의 카드가 있다. 이 카드 중에서 임의로 서로 다른 $4$ 장의 카드를 선택할 때, 선택한 카드 $4$ 장에 적힌 수 중 가장 큰 수를 확률변수 $X$ 라 하자. 다음은 $E(X)$ 를 구하는 과정이다. (단, $n \geq 4$ )

> 자연수 $k\,(4 \leq k \leq n)$ 에 대하여 확률변수 $X$ 의 값이 $k$ 일 확률은 $1$ 부터 $k-1$ 까지의 자연수가 적혀 있는 카드 중에서 서로 다른 $3$ 장의 카드와 $k$ 가 적혀 있는 카드를 선택하는 경우의 수를 전체 경우의 수로 나누는 것이므로 $\mathrm{P}(X=k) = \dfrac{\boxed{(가)}}{{}_n\mathrm{C}_4}$ 이다.
>
> 자연수 $r\,(1 \leq r \leq k)$ 에 대하여 ${}_k\mathrm{C}_r = \dfrac{k}{r} \times {}_{k-1}\mathrm{C}_{r-1}$ 이므로 $k \times \boxed{(가)} = 4 \times \boxed{(나)}$ 이다.
>
> 그러므로 $\mathrm{E}(X) = \displaystyle\sum_{k=4}^{n}\{k \times \mathrm{P}(X=k)\} = \dfrac{1}{{}_n\mathrm{C}_4}\sum_{k=4}^{n}\left(k \times \boxed{(가)}\right) = \dfrac{4}{{}_n\mathrm{C}_4}\sum_{k=4}^{n}\boxed{(나)}$ 이다.
>
> $\displaystyle\sum_{k=4}^{n}\boxed{(나)} = {}_{n+1}\mathrm{C}_5$ 이므로 $\mathrm{E}(X) = (n+1) \times \boxed{(다)}$ 이다.

위의 (가), (나)에 알맞은 식을 각각 $f(k)$ , $g(k)$ 라 하고, (다)에 알맞은 수를 $a$ 라 할 때, $a \times f(6) \times g(5)$ 의 값은?

① $40$ ② $45$ ③ $50$ ④ $55$ ⑤ $60$

예제 053

2005. 11. 가형(79%), 나형(61%). 22번. 3점

다음은 확률변수 $X$의 확률분포표이다. $\frac{4}{7}$, $a$, $b$가 이 순서로 등비수열을 이루고 $X$의 평균이 $24$일 때, $k$의 값을 구하시오.

| $X$ | $k$ | $2k$ | $4k$ | 계 |
|---|---|---|---|---|
| $P(X=x)$ | $\frac{4}{7}$ | $a$ | $b$ | $1$ |

예제 054

2005. 9. 나형(66%). 27번. 4점

이산확률변수 $X$의 확률분포 표는 다음과 같다.
$X$의 분산이 1이 되는 $p$와 $q$에 대하여 $3p+q$의 값은?

| $X$ | 0 | 1 | 2 | 3 | 계 |
|---|---|---|---|---|---|
| $P(X=x)$ | $p$ | $\frac{1}{4}$ | $q$ | $\frac{1}{12}$ | 1 |

① $\frac{1}{2}$ ② $\frac{3}{4}$ ③ $1$ ④ $\frac{3}{2}$ ⑤ $2$

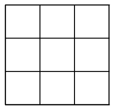
예제 055

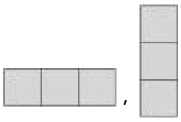
어떤 상품의 가격은 매달 0.5의 확률로 10% 상승하거나 0.5의 확률로 10% 하락한다. 이 상품의 현재 가격은 500원이다. 두 달 후 이 상품의 가격이 500원 이하이면 500원에서 두 달 후 상품가격을 뺀 금액을 받고, 500원 이상이면 받지 않기로 하였다. 두 달 후 받을 수 있는 금액의 기댓값을 $m$이라고 할 때, $4m$의 값을 구하여라.
(단, 첫 번째 달의 가격 변동과 두 번째 달의 가격 변동은 독립이다.)

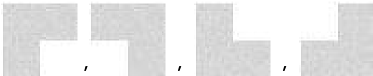
**예제 056**

2006. 6. 가형(58%). 20번. 3점

오른쪽 그림과 같이 한 변의 길이가 3인 정사각형을 한 변의 길이가 1인 정사각형 9개로 나누고, 이 중에서 3개를 색칠할 때 나타나는 모양은 다음과 같이 세 가지 유형으로 분류할 수 있다.



(가) 유형 1 :  와 같은 모양

(나) 유형 2 :  와 같은 모양

(다) 유형 3 : 유형 1도 아니고 유형 2도 아닌 모양

한 변의 길이가 1인 위의 정사각형 9개 중에서 임의로 3개를 색칠하여 얻은 모양의 유형에 따라 확률변수 $X$는 다음과 같다고 하자.

$$X = \begin{cases} 1 & (\text{유형 } 1\text{인 경우}) \\ 2 & (\text{유형 } 2\text{인 경우}) \\ 3 & (\text{유형 } 3\text{인 경우}) \end{cases}$$

$E(42X)$의 값을 구하시오.

예제 057

2008. 10. 가형(35%), 나형(24%). 22번. 3점

확률변수 $X$는 이항분포 $\mathrm{B}(3, p)$를 따르고 확률변수 $Y$는 이항분포 $\mathrm{B}(4, 2p)$를 따른다고 한다.

이때, $10\mathrm{P}(X=3)=\mathrm{P}(Y\geq 3)$을 만족시키는 양수 $p$의 값은 $\frac{n}{m}$이다. $m+n$의 값을 구하시오.

(단, $m$, $n$은 서로소인 자연수이다.)

예제 058

2009. 6. 가형(44%). 13번. 4점

어느 창고에 부품 $S$가 3개, 부품 $T$가 2개 있는 상태에서 부품 2개를 추가로 들여왔다. 추가된 부품은 $S$ 또는 $T$이고, 추가된 부품 중 $S$의 개수는 이항분포 $B\left(2,\ \frac{1}{2}\right)$을 따른다. 이 7개의 부품 중 임의로 1개를 선택한 것이 $T$일 때, 추가된 부품이 모두 $S$였을 확률은?

① $\frac{1}{6}$ ② $\frac{1}{4}$ ③ $\frac{1}{3}$ ④ $\frac{1}{2}$ ⑤ $\frac{3}{4}$

예제 059

2010. 9. 가형(79%), 나형(73%). 13번. 4점

두 사람 $A$와 $B$가 각각 주사위를 한 개씩 동시에 던지는 시행을 한다. 이 시행에서 나온 두 주사위의 눈의 수의 차가 $3$보다 작으면 $A$가 $1$점을 얻고, 그렇지 않으면 $B$가 $1$점을 얻는다. 이와 같은 시행을 15회 반복할 때, $A$가 얻는 점수의 합의 기댓값과 $B$가 얻는 점수의 합의 기댓값의 차는?

① $1$ ② $3$ ③ $5$ ④ $7$ ⑤ $9$

예제 060

2007. 11. 나형(52%). 23번. 4점

한 개의 주사위를 $20$번 던질 때 1의 눈이 나오는 횟수를 확률변수 $X$라 하고, 한 개의 동전을 $n$번 던질 때 앞면이 나오는 횟수를 확률변수 $Y$라 하자. $Y$의 분산이 $X$의 분산보다 크게 되도록 하는 $n$의 최솟값을 구하시오.

예제 061

2009. 10. 나형(46%). 9번. 4점

표는 $k=0, 1, 2, 3, 4$ 일 때, $p_k = {}_{30}\mathrm{C}_k\left(\frac{1}{6}\right)^k\left(\frac{5}{6}\right)^{30-k}$ 의 값을 소수점 아래 셋째자리까지 나타낸 것이다.

| $k$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $p_k$ | 0.004 | 0.025 | 0.073 | 0.137 | 0.185 |

주사위를 30 번 던져 1 의 눈이 나오는 횟수를 확률변수 $X$ 라 할 때, 위의 표를 이용하여 $\sum_{r=3}^{30} r\,P(X=r)$ 의 값을 구한 것은?

① 4.765 ② 4.829 ③ 4.902 ④ 4.946 ⑤ 4.971

예제 062

2010. 10. 나형(45%). 30번. 4점

10 이하의 음이 아닌 정수 $r$ 에 대하여 함수 $f$ 를 $f(r) = {}_{10}C_r\left(\frac{1}{2}\right)^{10}$ 이라 할 때, $2\sum_{r=0}^{10} r^2 f(r)$ 의 값을 구하시오.

예제 063

한 개의 동전을 $n$회 던질 때, 앞면이 나오는 횟수를 $X$라 한다. $E((2X-n)^2)=10$을 만족시키는 $n$의 값은?

☐ 8 ☐ 9 ☐ 10 ☐ 11 ☐ 12

예제 064

한 개의 주사위를 3회 던져서 1의 눈이 나오는 횟수 $X$에 대하여 $4^X$원의 상금을 받기로 할 때, 상금의 기댓값을 구하여라.

**예제 065**

원점 O를 출발하여 수직선 위를 움직이는 점 P가 있다. 동전 한 개를 던져서 앞면이 나오면 양의 방향으로 3만큼, 뒷면이 나오면 음의 방향으로 2만큼 점 P를 이동시킨다. 동전을 10번 던진 후의 점 P의 좌표를 확률변수 $X$라 할 때, $\mathrm{E}(X)$는?

① 1 ② 3 ③ 4 ④ 5 ⑤ 10

# 정답과 해설

예제 050

REVIEW 이산확률분포의 확률표현

2006. 9. 가형(75%), 나형(69%). 13번. 4점

이산확률변수 $X$의 확률분포 표는 다음과 같다.

| $X$ | 0 | 1 | 2 | $\cdots$ | 10 | 계 |
|---|---|---|---|---|---|---|
| $P(X=x)$ | $p_0$ | $p_1$ | $p_2$ | $\cdots$ | $p_{10}$ | 1 |

(단, $p_i > 0$이고 $i=0,\ 1, 2,\ \cdots,\ 10$이다.)

집합 $\{x□0 \le x \le 10\}$에서 정의된

두 함수 $F(x), G(x)$가

$F(x)=P(0 \le X \le x)$, $G(x)=P(X>x)$일 때, <보기>에서 옳은 것을 모두 고른 것은?

<보 기>

ㄱ. $G(3)=1-F(3)$  ㄴ. $P(3 \le X \le 8)=F(8)-F(3)$  ㄷ. $P(3 \le X \le 8)=G(2)-G(8)$

① ㄱ  ② ㄷ  ③ ㄱ, ㄴ  ④ ㄱ, ㄷ  ⑤ ㄴ, ㄷ

치환과 일대일 대응 (후에 중복조합에서 자주 사용되는 스킬)

- $F(x)=P(0 \le X \le x)$, $G(x)=P(X>x)$ - 각 함수 $F(x)$와 $G(x)$는 확률을 의미하는 함수이다.
  이산확률분포의 확률의 표현을 알고, 표만 읽어 줄 수 있으면 어려운 문제가 아니다.

ㄱ. $G(3) = 1 - F(3)$ ⇨ (좌변) $G(3) = P(X>3) = p_4 + p_5 + \cdots + p_{10}$

(우변) $1 - F(3) = 1 - P(0 \le X \le 3) = p_4 + p_5 + \cdots + p_{10}$ … (참)

ㄴ. $P(3 \le X \le 8) = F(8) - F(3)$ ⇨ (좌변) $P(3 \le X \le 8) = p_3 + p_4 + \cdots + p_8$

(우변) $F(8) - F(3) = P(0 \le X \le 8) - P(0 \le X \le 3)$

$= (p_0+p_1+\cdots+p_8) - (p_0+p_1+p_2+p_3)$

$= p_4 + \cdots + p_8$ … (거짓)

ㄷ. $P(3 \le X \le 8) = G(2) - G(8)$ ⇨ (좌변) $P(3 \le X \le 8) = p_3 + p_4 + \cdots + p_8$

(우변) $G(2) - G(8) = P(X>2) - P(X>8)$

$= (p_3+p_4+\cdots+p_{10}) - p_9 + p_{10}$

$= p_3 + p_4 + \cdots + p_8$ … (참)

정답은 ④

예제 051

REVIEW 확률질량함수 표현하기

한 개의 주사위를 두 번 던져서 나오는 눈의 수의 합을 확률변수 $X$라고 할 때, $X$의 확률질량함수를 구하여라.

- 이것은 내신전용이다. 확률질량함수를 구하기 위해서는 수열의 규칙성을 발견해야 한다.
  확률적으로 어렵게 나올 수가 없기 때문에 확률문제라기보다는 수열 문제에 가깝다. 먼저 표를 작성하고 규칙성을 발견한 후에 식을 세운다. (정형화된 풀이가 있는 것은 아니므로 가볍게 패스)

| $X$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 총합 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $P(X)$ | $\frac{1}{36}$ | $\frac{2}{36}$ | $\frac{3}{36}$ | $\frac{4}{36}$ | $\frac{5}{36}$ | $\frac{6}{36}$ | $\frac{5}{36}$ | $\frac{4}{36}$ | $\frac{3}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ | 1 |

결국 위의 표를 보고 식을 세우되 제한범위를 정확히 써준다.

$$P(X=x)=\begin{cases} \dfrac{x-1}{36} & (2 \le x \le 7) \\ \dfrac{13-x}{36} & (8 \le x \le 12) \end{cases} \quad \text{(단, } x\text{는 자연수)}$$

예제 052

REVIEW 확률질량함수 표현하기

2016. 09. 고3. 가형. 17번. 4점

$1$부터 $n$까지의 자연수가 하나씩 적혀 있는 $n$장의 카드가 있다. 이 카드 중에서 임의로 서로 다른 $4$장의 카드를 선택할 때, 선택한 카드 $4$장에 적힌 수 중 가장 큰 수를 확률변수 $X$라 하자. 다음은 $E(X)$를 구하는 과정이다. (단, $n \geq 4$)

> 자연수 $k\,(4 \leq k \leq n)$에 대하여 확률변수 $X$의 값이 $k$일 확률은 $1$부터 $k-1$까지의 자연수가 적혀 있는 카드 중에서 서로 다른 $3$장의 카드와 $k$가 적혀 있는 카드를 선택하는 경우의 수를 전체 경우의 수로 나누는 것이므로 $\mathrm{P}(X=k) = \dfrac{\boxed{(가)}}{{}_n\mathrm{C}_4}$ 이다.
>
> 자연수 $r\,(1 \leq r \leq k)$에 대하여 ${}_k\mathrm{C}_r = \dfrac{k}{r} \times {}_{k-1}\mathrm{C}_{r-1}$이므로 $k \times \boxed{(가)} = 4 \times \boxed{(나)}$이다.
>
> 그러므로 $\mathrm{E}(X) = \sum_{k=4}^{n}\{k \times \mathrm{P}(X=k)\} = \dfrac{1}{{}_n\mathrm{C}_4}\sum_{k=4}^{n}\left(k \times \boxed{(가)}\right) = \dfrac{4}{{}_n\mathrm{C}_4}\sum_{k=4}^{n}\boxed{(나)}$ 이다.
>
> $\sum_{k=4}^{n}\boxed{(나)} = {}_{n+1}\mathrm{C}_5$ 이므로 $\mathrm{E}(X) = (n+1) \times \boxed{(다)}$ 이다.

위의 (가), (나)에 알맞은 식을 각각 $f(k)$, $g(k)$라 하고, (다)에 알맞은 수를 $a$라 할 때, $a \times f(6) \times g(5)$ 의 값은?

① 40 ② 45 ③ 50 ④ 55 ⑤ 60

---

조건 $n$장의 카드 + 서로 다른 $4$장의 카드를 선택할 때, 선택한 카드 $4$장에 적힌 수 중 가장 큰 수를 확률변수 $X$

⇨ 조건에서 가능한 확률변수는 $4$, $5$, $6$, $\cdots$, $n$이라는 사실을 알 수 있다.

$E(X)$를 구하기 위해 <확률>을 구해야 한다는 예측을 하며 보기를 읽어나간다. (다음 장 티칭을 참조)

과정 1 주어진 조건에 의해 $P(X=k)$를 구해보면 $n(S) = {}_nC_4$이고, 최댓값이 $k$인 경우의 수는 $k$를 선택 한 후 $1$부터 $k-1$까지의 자연수 중 $3$개를 선택하면 되므로 ${}_{k-1}C_3$ 따라서 $P(X=k) = \dfrac{{}_{k-1}C_3}{{}_nC_4}$ 즉, $\boxed{(가) = {}_{k-1}C_3}$

과정 2 $E(X)$를 구하기 위하여 $\sum_{k=4}^{n} k \cdot \dfrac{{}_{k-1}C_4}{{}_nC_4}$을 구해야 한다. <과정 2>가 $\sum_{k=4}^{n} k \cdot \dfrac{{}_{k-1}C_4}{{}_nC_4}$의 식을 처리하기 위한 과정임을 예측하며 읽어야 자연스럽게 풀 수 있다.

- 자연수 $r\,(1 \leq r \leq k)$에 대하여 ${}_k\mathrm{C}_r = \dfrac{k}{r} \times {}_{k-1}\mathrm{C}_{r-1}$이므로

⇨ $k \times \boxed{(가)\ {}_{k-1}C_3} = 4 \times \boxed{(나)}$에서 <주어진 식> ${}_kC_r = \dfrac{k}{r} \times {}_{k-1}C_{r-1}$을 활용한다.

${}_kC_r = \dfrac{k}{r} \times {}_{k-1}C_{r-1} \Leftrightarrow r \times {}_kC_r = k \times {}_{k-1}C_{r-1}$에서 $k \times {}_{k-1}C_3 = 4 \times {}_kC_4$이므로 $\boxed{(나) = {}_kC_4}$

(주어진 식을 활용하여 식을 바꾸는 과정이 궁금하다면 다음 장의 티칭을 참조하라.)

과정 3 <과정 2>가 왜 필요했는지 자연스럽게 읽으면서 이해해본다. (다시 처음으로 돌아가 평균을 구하는 과정이다.)

$E(X) = \sum_{k=4}^{n}\{k \times P(X=k)\} = \dfrac{1}{{}_nC_4}\sum_{k=4}^{n}\left(k \times \boxed{(가)\ {}_{k-1}C_3}\right)$ - <과정 1>에서 서술한 내용을 활용한 것이다.

$= \dfrac{4}{{}_n\mathrm{C}_4}\sum_{k=4}^{n}\boxed{(나)\ {}_nC_4}$ 이다. - <과정 2>에서 서술한 내용을 활용한 것이다.

- $\underline{\sum_{k=4}^{n} \boxed{㈏}\, {}_{n}C_{4} = {}_{n+1}C_{5}}$ 이므로 ⇨ <파스칼 삼각형 공식>을 이용해서 간단히 처리할 수 있지만 문제에서 이미 계산해주었으므로 넘어간다.

- $\underline{E(X) = (n+1) \times \boxed{㈐}}$ 이다.

⇨ 최종식인 $\frac{4}{{}_{n}C_{4}} \times {}_{n+1}C_{5}$ 의 정리는 <조합의 정의식>을 이용한다. (촌철살인 확률과 통계 1. - Part 6. - #2 참조)

$$\frac{4}{{}_{n}C_{4}} \times {}_{n+1}C_{5} = 4 \cdot \frac{4! \cdot (n-4)!}{n!} \cdot \frac{(n+1)!}{5! \cdot (n-4)!} = \frac{4}{5}(n+1)$$ 이므로 $\boxed{(다) = \frac{4}{5}}$

따라서 $a = \frac{4}{5}$, $f(6) = {}_{5}C_{3} = 10$, $g(5) = {}_{5}C_{4} = 5$

$a \times f(6) \times g(5) = 40$이므로 정답은 ①

티칭 빈칸 채우기 문제

- 빈칸 채우기 문제는 수학적 서술과정을 물어보기 위한 문제로서 <배운 발상에 대한 과정의 생략>과 <생각해내기 어려운 발상에 대한 조건>이 섞여있는 문제이다. 즉, <배운 발상에 대한 과정의 생략>되어 있으므로 우리가 공부한 <내용>을 **예측**하면서 보기를 읽어야 서술과정에서 <생략된 부분>을 매끄럽게 넘어갈 수 있다.

- 또한 <배우지 않은 공식(항등식), 식 변형 발상>은 반드시 제공되므로 반드시 <제공된 식>을 통해서 정리해야 하는 복잡한 식이 나올 것이라는 예측하며 풀어야 한다.

티칭 <촌철살인 확률과 통계 1>에서 잠깐 언급했지만 <자연수 범위의 공식(항등식)>은 보통 식의 구조가 복잡하므로 외워서 대입하는 것이 아니라 의미를 만들어 읽는 것(Reading)이다.
참고로 조합 ${}_{n}C_{r}$에서 <$n$을 고정수> <$r$을 변동수>라고 부르기로 하자.

<제공된 식> ${}_{k}C_{r} = \frac{k}{r} \times {}_{k-1}C_{r-1}$ ⇔ $k \times {}_{k-1}C_{r-1} = r \times {}_{k}C_{r}$ (좌 · 우변의 위치를 바꾸었을 뿐)

역시 <주어진 식이 정의되는 임의의 자연수 $k$, $r$에 대해서 성립하는 식이라는 뜻>이므로 $n$과 $r$에 대한 항등식이다.
이 식을 읽는 방법은 $k \times {}_{k-1}C_{r-1}$ <고정수>와 <계수 : 고정수+1>에서 <고정수+1> ⇨ 새로운 고정수 $r \times {}_{k}C_{r}$

$k \times {}_{k-1}C_{r-1}$ <변동수> + 1 ⇨ 새로운 변동수와 계수 $r \times {}_{k}C_{r}$

연습) $k \times {}_{k-1}C_{r-1} = r \times {}_{k}C_{r}$의 식을 이용하여 다음을 간단히 해보자.

1) $(k+1) \times {}_{k}C_{r-3}$ ⇨ 새로운 고정수는 $k+1$이고, 새로운 변동수와 계수는 $r-2$이므로 $(r-2) \times {}_{k+1}C_{r-2}$

2) $k \times {}_{k-1}C_{3}$ (앞의 문제 과정 2) ⇨ 새로운 고정수는 $k$이고, 새로운 변동수와 계수는 4이므로 $4 \times {}_{k}C_{4}$

예제 053

REVIEW 이산확률분포와 평균

2005. 11. 가형(79%), 나형(61%). 22번. 3점

다음은 확률변수 $X$의 확률분포표이다. $\frac{4}{7}$, $a$, $b$가 이 순서로 등비수열을 이루고 $X$의 평균이 24일 때, $k$의 값을 구하시오.

| $X$ | $k$ | $2k$ | $4k$ | 계 |
|---|---|---|---|---|
| $P(X=x)$ | $\frac{4}{7}$ | $a$ | $b$ | 1 |

조건 주어진 확률분포 표에서 $\sum p_i = 1$ ⇨ $\frac{4}{7}+a+b = 1 \Leftrightarrow a+b = \frac{3}{7}$ ⋯ ㉠

조건 $\frac{4}{7}$, $a$, $b$가 이 순서로 등비수열 ⇨ $a^2 = \frac{4}{7}b$ ($a, b$는 확률 값이므로 $a, b > 0$) ⋯ ㉡

- 이 두 식을 연립하여 $a, b$를 구한다. ㉠ $b = \frac{3}{7}-a$을 ㉡ $a^2 = \frac{4}{7}b$에 대입하면

$$a^2 = \frac{4}{7}\left(\frac{3}{7}-a\right) \Leftrightarrow 49a^2 = 4(3-7a) \Leftrightarrow 49a^2+28a-12 = 0$$

$\Leftrightarrow (7a+6)(7a-2) = 0$이므로 $a = \frac{2}{7}$, $b = \frac{1}{7}$

| $X$ | $k$ | $2k$ | $4k$ | 계 |
|---|---|---|---|---|
| $P(X=x)$ | $\frac{4}{7}$ | $\frac{2}{7}$ | $\frac{1}{7}$ | 1 |

조건 $X$의 평균이 24이므로 ⇨ $E(X) = \sum_{i=1}^{n} x_i p_i = k\times\frac{4}{7} + 2k\times\frac{2}{7} + 4k\times\frac{1}{7} = \frac{12k}{7} = 24 \Leftrightarrow k = 14$

정답은 14

예제 054

REVIEW 이산확률분포와 분산

2005. 9. 나형(66%). 27번. 4점

이산확률변수 $X$의 확률분포 표는 다음과 같다.

$X$의 분산이 1이 되는 $p$와 $q$에 대하여 $3p+q$의 값은?

| $X$ | 0 | 1 | 2 | 3 | 계 |
|---|---|---|---|---|---|
| $P(X=x)$ | $p$ | $\frac{1}{4}$ | $q$ | $\frac{1}{12}$ | 1 |

① $\frac{1}{2}$ ② $\frac{3}{4}$ ③ 1 ④ $\frac{3}{2}$ ⑤ 2

조건 주어진 확률분포 표에서 $\sum p_i = 1$ ⇨ $p + \frac{1}{4} + q + \frac{1}{12} = 1$ ⋯ ㉠

조건 $X$의 분산이 1 ⇨ $E(X^2) = \sum_{i=1}^{n} x_i^2 p_i = 0^2\cdot p + 1^2\cdot\frac{1}{4} + 2^2\cdot q + 3^2\cdot\frac{1}{12} = 4q+1$

$E(X) = \sum_{i=1}^{n} x_i p_i = 0\cdot p + 1\cdot\frac{1}{4} + 2\cdot q + 3\cdot\frac{1}{12} = 2q+\frac{1}{2}$이므로

$$V(X) = E(X^2) - \{E(X)\}^2 = 4q+1 - \left(2q+\frac{1}{2}\right)^2 = -4q^2+2q+\frac{3}{4}$$

그러므로 $-4q^2+2q+\frac{3}{4} = 1 \Leftrightarrow 16q^2-8q+1 = 0 \Leftrightarrow q = \frac{1}{4}$

이것을 ㉠에 대입하면 $p = \frac{5}{12}$이다. 즉, $3p+q = \frac{3}{2}$이다.

정답은 ④

예제 055

REVIEW 표 작성하기 문제 - 이산확률분포와 평균

어떤 상품의 가격은 매달 0.5의 확률로 10% 상승하거나 0.5의 확률로 10% 하락한다. 이 상품의 현재 가격은 500원이다. 두 달 후 이 상품의 가격이 500원 이하이면 500원에서 두 달 후 상품가격을 뺀 금액을 받고, 500원 이상이면 받지 않기로 하였다. 두 달 후 받을 수 있는 금액의 기댓값을 $m$이라고 할 때, $4m$의 값을 구하여라.
(단, 첫 번째 달의 가격 변동과 두 번째 달의 가격 변동은 독립이다.)

표 작성 문제는 확률문제

조건 두 달 후 받을 수 있는 금액의 기댓값을 $m$

⇨ 이 조건을 통해 확률변수가 〈두 달 후 받을 수 있는 금액〉이라는 사실을 알 수 있다.
문제의 상황을 표로 정리하면 다음과 같다.

| 두 달 후 | 두 달 후 가격 | 확률변수 | 확률 |
|---|---|---|---|
| 상승 → 상승 | $500(1.1)(1.1)=605$ | $0$ | $\frac{1}{2}\times\frac{1}{2}=\frac{1}{4}$ |
| 상승 → 하락 | $500(1.1)(0.9)=495$ | $5$ | $\frac{1}{2}\times\frac{1}{2}=\frac{1}{4}$ |
| 하락 → 상승 | $500(0.9)(1.1)=495$ | $5$ | $\frac{1}{2}\times\frac{1}{2}=\frac{1}{4}$ |
| 하락 → 하락 | $500(0.9)(0.9)=405$ | $95$ | $\frac{1}{2}\times\frac{1}{2}=\frac{1}{4}$ |

⇨ 이 결과로 표를 완성하면

| $X$ | 0 | 5 | 95 | 합계 |
|---|---|---|---|---|
| $P(X=x)$ | $\frac{1}{4}$ | $\frac{1}{2}$ | $\frac{1}{4}$ | 1 |

이므로 금액 $X$의 기댓값은

$$E(X)=\sum_{i=1}^{n}x_ip_i=0\times\frac{1}{4}+5\times\frac{1}{2}+95\times\frac{1}{4}=\frac{105}{4}$$

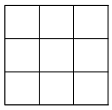
즉, $4m=105$ 이므로 정답은 $105$

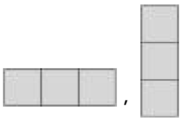
예제 056

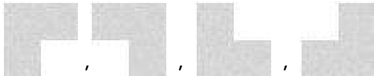
REVIEW 표 작성하기 문제 - 이산확률분포와 평균의 성질

2006. 6. 가형(58%). 20번. 3점

오른쪽 그림과 같이 한 변의 길이가 3인 정사각형을 한 변의 길이가 1인 정사각형 9개로 나누고, 이 중에서 3개를 색칠할 때 나타나는 모양은 다음과 같이 세 가지 유형으로 분류할 수 있다.



(가) 유형 1 :  와 같은 모양

(나) 유형 2 :  와 같은 모양

(다) 유형 3 : 유형 1도 아니고 유형 2도 아닌 모양

한 변의 길이가 1인 위의 정사각형 9개 중에서 임의로 3개를 색칠하여 얻은 모양의 유형에 따라 확률변수 $X$는 다음과 같다고 하자.

$$X = \begin{cases} 1 & (\text{유형 1인 경우}) \\ 2 & (\text{유형 2인 경우}) \\ 3 & (\text{유형 3인 경우}) \end{cases}$$

$E(42X)$의 값을 구하시오.

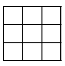
표 작성 문제는 확률문제

- 결국 확률 값을 모두 구해서 표를 작성해야 하는 문제이다.

조건 한 변의 길이가 3인 정사각형을 한 변의 길이가 1인 정사각형 9개로 나누고, 이 중에서 3개를 색칠할 때

⇨ 시행에 다른 모든 경우의 수는 ${}_9C_3 = 84$이다.

- $n$(유형 1)은 하나하나 센다. 즉, $n$(유형 1) $= 6$이므로 $P$(유형 1) $= \frac{6}{84}$

- $n$(유형 2)는 안에서 4개씩 존재하고 안에서 은 총 4개가 존재하므로 $n$(유형 2) $= 4\times 4$

즉, $P$(유형 2) $= \frac{16}{84}$

- $n$(유형 3)은 전체에서 위의 경우를 뺀다. 즉, $n$(유형 3) $= 84 - 6 - 16 = 62$ 이므로 $P$(유형 3) $= \frac{62}{84}$

이것을 표로 만들면 오른쪽과 같다.

| $X$ | 1 | 2 | 3 | 계 |
|---|---|---|---|---|
| $P(X=x)$ | $\frac{6}{84}$ | $\frac{16}{84}$ | $\frac{62}{84}$ | 1 |

$E(X) = \sum_{i=1}^{n} x_i p_i = 1\times\frac{6}{84} + 2\times\frac{16}{84} + 3\times\frac{62}{84} = \frac{224}{84}$

즉, $E(42X) = 42E(X) = 42\times\frac{224}{84} = 112$

정답은 112

예제 057

REVIEW 이항분포와 확률

2008. 10. 가형(35%), 나형(24%). 22번. 3점

확률변수 $X$는 이항분포 $\mathrm{B}(3, p)$를 따르고 확률변수 $Y$는 이항분포 $\mathrm{B}(4, 2p)$를 따른다고 한다.

이때, $10\mathrm{P}(X=3)=\mathrm{P}(Y\geq 3)$을 만족시키는 양수 $p$의 값은 $\frac{n}{m}$이다. $m+n$의 값을 구하시오.

(단, $m$, $n$은 서로소인 자연수이다.)

선분할 개념 : <분할 후 분배>의 논리가 아닌 <분할적 사고>로 논리적 접근

- 확률변수 $X$는 이항분포 $B(3, p)$에서 $10\mathrm{P}(X=3) = 10\times{}_3C_3p^3(1-p)^0 = 10p^3$
- 확률변수 $Y$는 이항분포 $B(4, 2p)$에서 $\mathrm{P}(Y\geq 3) = {}_4C_3p^3(1-p)^1 + {}_4C_4p^4(1-p)^0 = 16p^3(2-3p)$이므로
- $10\mathrm{P}(X=3) = \mathrm{P}(Y\geq 3) \Leftrightarrow 10p^3 = 16p^3(2-3p) \Leftrightarrow 10 = 16(2-3p)$ 따라서 $p = \frac{11}{24}$

$m+n=35$

코칭 정답률에 너무 연연하지 말자. 2008년도는 매 시험이 너무 어려워 앞에서 시간을 너무 많이 빼앗겼고 어려운 시험에 대한 피로감 누적으로 쉬운 문제도 어렵게 보는 경향이 있었다.

코칭 이항분포 $B(n,\ p)$는 표를 요약해서 보여주는 기호일 뿐이다. 반드시 머릿속에는 <표>가 그려져야 한다.

예제 058

REVIEW 이항분포와 확률

2009. 6. 가형(44%). 13번. 4점

어느 창고에 부품 $S$가 3개, 부품 $T$가 2개 있는 상태에서 부품 2개를 추가로 들여왔다. 추가된 부품은 $S$ 또는 $T$이고, 추가된 부품 중 $S$의 개수는 이항분포 $B\left(2,\ \frac{1}{2}\right)$을 따른다. 이 7개의 부품 중 임의로 1개를 선택한 것이 $T$일 때, 추가된 부품이 모두 $S$ 였을 확률은?

① $\frac{1}{6}$ ② $\frac{1}{4}$ ③ $\frac{1}{3}$ ④ $\frac{1}{2}$ ⑤ $\frac{3}{4}$

조건 추가된 부품은 $S$ 또는 $T$이고, 추가된 부품 중 $S$의 개수는 이항분포 $B\left(2,\ \frac{1}{2}\right)$을 따른다.

⇨ $B\left(2,\ \frac{1}{2}\right)$을 표로 나타내보면 오른쪽과 같다.

| $X$ | 0 ($S0$ $T2$) | 1 ($S1$ $T1$) | 2 ($S2$ $T0$) |
|---|---|---|---|
| $P(X)$ | ${}_2C_0\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^2=\frac{1}{4}$ | ${}_2C_1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^1=\frac{1}{2}$ | ${}_2C_2\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^0=\frac{1}{4}$ |

즉, 이항분포를 통해서 <부품이 추가될 확률>을 준 것이다. <이항분포>의 역할은 여기까지이고, 이제부터는 확률문제이다.

조건 7개의 부품 중 임의로 1개를 선택한 것이 $T$일 때, 추가된 부품이 모두 $S$ 였을 확률

⇨ <선 시행 : 부품을 들여오는 시행>에 따라 <후 시행>이 영향을 받으므로 종속적 선후시행이다.

또한 <미래의 결과, 과거의 확률>을 물어보는 조건부 확률이다.

$$P(\text{들여온 부품이 모두 } S \mid T\text{ 선택}) = \frac{P(\text{들여온 부품이 모두 } S \cap T \text{ 선택})}{P(T\text{ 선택})}$$

- 여기에서 분모에 해당하는 <조건부 사건 : $T$선택>은 종속적 선후시행에서 <후행사건>이므로 다음과 같이 3가지 상황으로 분할 할 수 있다.

| $S2$, $T0$ → $T$ 선택 | $S1$, $T1$ → $T$ 선택 | $S0$, $T2$ → $T$ 선택 |
|---|---|---|
| 상황 1 | 상황 2 | 상황 3 |

$$P(T\text{ 선택}) = P(S2,\ T0 \to T\text{ 선택}) + P(S1,\ T1 \to T\text{ 선택}) + P(S0,\ T2 \to T\text{ 선택})$$

$$= P(S2,\ T0)\cdot P(T\text{ 선택} \mid S2,\ T0) + P(S1,\ T1)\cdot P(T\text{ 선택} \mid S1,\ T1) + P(S0,\ T2)\cdot P(T\text{ 선택} \mid S0,\ T2)$$

$$= \frac{1}{4}\cdot\frac{2}{7} + \frac{1}{2}\cdot\frac{3}{7} + \frac{1}{4}\cdot\frac{4}{7}$$

⇨ $$P(S2,\ T0 \mid T\text{ 선택}) = \frac{P(S2,\ T0 \to T\text{ 선택})}{P(S2,\ T0 \to T\text{ 선택}) + P(S1,\ T1 \to T\text{ 선택}) + P(S0,\ T2 \to T\text{ 선택})}$$

$$= \frac{\frac{1}{4}\cdot\frac{2}{7}}{\frac{1}{4}\cdot\frac{2}{7} + \frac{1}{2}\cdot\frac{3}{7} + \frac{1}{4}\cdot\frac{4}{7}} = \frac{1}{6}$$

정답은 ①

예제 059

REVIEW 이항분포의 평균

2010. 9. 가형(79%), 나형(73%). 13번. 4점

두 사람 $A$와 $B$가 각각 주사위를 한 개씩 동시에 던지는 시행을 한다. 이 시행에서 나온 두 주사위의 눈의 수의 차가 3보다 작으면 $A$가 1점을 얻고, 그렇지 않으면 $B$가 1점을 얻는다. 이와 같은 시행을 15회 반복할 때, $A$가 얻는 점수의 합의 기댓값과 $B$가 얻는 점수의 합의 기댓값의 차는?

① 1 ② 3 ③ 5 ④ 7 ⑤ 9

조건 15회 반복할 때, $A$가 얻는 점수의 합의 기댓값과 $B$가 얻는 점수의 합의 기댓값의 차

⇨ $A$가 얻는 점수를 확률변수 $X$, $B$가 얻는 점수를 확률변수 $Y$라 할 수 있다.

15회의 시행에서 $X$가 얻을 수 있는 가능한 모든 점수는 0, 1, 2, ⋯ 15

15회의 시행에서 $Y$가 얻을 수 있는 가능한 모든 점수는 0, 1, 2, ⋯ 15 (아래와 같이 표를 떠올린다.)

| $X$ | 0 | 1 | ⋯ | 15 | 합계 |
|---|---|---|---|---|---|
| $P(X)$ | | | | | 1 |

와

| $Y$ | 0 | 1 | ⋯ | 15 | 합계 |
|---|---|---|---|---|---|
| $P(Y)$ | | | | | 1 |

에서 확률부분을 채운다.

- $A$가 1회 시행에서 점수를 얻을 확률은 $\frac{n(\text{차 }0)+n(\text{차 }1)+n(\text{차 }2)}{36} = \frac{6+5\cdot 2+4\cdot 2}{36} = \frac{2}{3}$이므로

$B$가 1회 시행에서 점수를 얻을 확률은 $1-P(A\text{가 점수얻을}) = 1-\frac{2}{3} = \frac{1}{3}$

- 위의 표에서 각 빈칸에 채워져야 하는 확률이 <독립시행의 확률>임을 눈치 챘다면 ⇨ $X : B\left(15, \frac{2}{3}\right)$ $Y : B\left(15, \frac{1}{3}\right)$

즉, $E(X) = 15\cdot\frac{2}{3} = 10$, $E(Y) = 15\cdot\frac{1}{3} = 5$이므로 기댓값의 차는 5이다.

정답은 ③

예제 060

REVIEW 이항분포의 분산

2007. 11. 나형(52%). 23번. 4점

한 개의 주사위를 20번 던질 때 1의 눈이 나오는 횟수를 확률변수 $X$라 하고, 한 개의 동전을 $n$번 던질 때 앞면이 나오는 횟수를 확률변수 $Y$라 하자. $Y$의 분산이 $X$의 분산보다 크게 되도록 하는 $n$의 최솟값을 구하시오.

조건 한 개의 주사위를 20번 던질 때 1의 눈이 나오는 횟수를 확률변수 $X$

⇨ 확률변수 $X$에 대한 이항분포는 $B\left(20, \frac{1}{6}\right)$

조건 한 개의 동전을 $n$번 던질 때 앞면이 나오는 횟수를 확률변수 $Y$

⇨ 확률변수 $Y$에 대한 이항분포는 $B\left(n, \frac{1}{2}\right)$이다.

- $V(X) = 20\cdot\frac{1}{6}\cdot\frac{5}{6} = \frac{25}{9}$, $V(Y) = n\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{n}{4}$이므로

구할 것 $Y$의 분산이 $X$의 분산보다 크게 되도록 하는 $n$의 최솟값

- $\frac{n}{4} > \frac{25}{9} \Leftrightarrow n > \frac{100}{9} \Leftrightarrow n > 11.\times\times$ 그런데 $n$은 자연수이므로 $n$의 최솟값은 12이다.

정답은 12

REVIEW 이항분포의 평균의 시그마 표현

예제 061

2009. 10. 나형(46%). 9번. 4점

표는 $k=0, 1, 2, 3, 4$ 일 때, $p_k = {}_{30}\mathrm{C}_k\left(\frac{1}{6}\right)^k\left(\frac{5}{6}\right)^{30-k}$ 의 값을 소수점 아래 셋째자리까지 나타낸 것이다.

| $k$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $p_k$ | 0.004 | 0.025 | 0.073 | 0.137 | 0.185 |

주사위를 30 번 던져 1 의 눈이 나오는 횟수를 확률변수 $X$ 라 할 때, 위의 표를 이용하여 $\sum_{r=3}^{30} r\,P(X=r)$ 의 값을 구한 것은?

① 4.765 ② 4.829 ③ 4.902 ④ 4.946 ⑤ 4.971

조건 주사위를 30번 던져 1의 눈이 나오는 횟수를 확률변수 $X$라 할 때

⇨ 확률변수 $X$에 대한 이항분포는 $B\left(30, \frac{1}{6}\right)$

조건 표는 $k=0, 1, 2, 3, 4$ 일 때, $p_k = {}_{30}\mathrm{C}_k\left(\frac{1}{6}\right)^k\left(\frac{5}{6}\right)^{30-k}$ 의 값을 소수점 아래 셋째자리까지 나타낸 것

⇨ $p_k = {}_{30}\mathrm{C}_k\left(\frac{1}{6}\right)^k\left(\frac{5}{6}\right)^{30-k}$ 은 위의 이항분포 $B\left(30, \frac{1}{6}\right)$의 확률 값을 의미하므로 주어진 표는 이항분포 $B\left(30, \frac{1}{6}\right)$의 일부를 나타낸 것이다.

구할 것 $\sum_{r=3}^{30} r\,P(X=r)$

⇨ 식의 모양을 보면 <확률변수와 확률 곱해서 일부를 더한 것>이므로 <평균>과 관련 있다는 사실을 눈치 채야 한다. 즉, <평균의 시그마 식>을 이용하여 식을 정리하면 <이항분포의 평균공식>을 사용할 수 있다.

- $\sum_{\boxed{r=3}}^{30} rP(X=r) = \sum_{\boxed{r=0}}^{30} rP(X=r) - \{P(1)+2P(2)\} = E(X) - \{P(1)+2P(2)\}$

$= 30\times\frac{1}{6} - \{0.025 + 2\times 0.073\} = 5 - 0.171 = 4.829$

정답은 ②

티칭 원래는 $P(1) = {}_{30}C_1\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^{29}$, $P(2) = {}_{30}C_2\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^{28}$ 이지만 계산하는 것이 불가능하기 때문에 문제에서 근삿값을 조건으로 준 것이다.

코칭 시그마로 표현된 식을 보고 <평균, 분산>의 의미를 발견해야 한다.

예제 062

REVIEW 시그마 표현

2010. 10. 나형(45%). 30번. 4점

10 이하의 음이 아닌 정수 $r$ 에 대하여 함수 $f$ 를 $f(r) = {}_{10}C_r\left(\frac{1}{2}\right)^{10}$ 이라 할 때, $2\sum_{r=0}^{10} r^2 f(r)$ 의 값을 구하시오.

조건 $f(r) = {}_{10}C_r\left(\frac{1}{2}\right)^{10}$ ⇨ $f(r) = {}_{10}C_r\left(\frac{1}{2}\right)^r\left(\frac{1}{2}\right)^{10-r}$ 라고 표현할 수 있다. 즉, $f(r)$은 이항분포 $B\left(10, \frac{1}{2}\right)$에서 $P(X=r)$을 나타내는 값이다.

구할 것 $2\sum_{r=0}^{10} r^2 f(r)$ ⇨ 이항분포 $B\left(10, \frac{1}{2}\right)$에서 ⟨$2\times$제곱의 평균⟩ 즉, $2E(X^2)$를 구하는 것이고, 이것은 ⟨이항분포의 평균과 분산의 공식⟩을 이용하여 계산할 수 있다.

- $V(X) = E(X^2) - \{E(X)\}^2 \Leftrightarrow E(X^2) = V(X) + \{E(X)\}^2$
- $E(X) = 10 \cdot \frac{1}{2} = 5$, $V(X) = 10 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{5}{2}$이므로

⇨ $2\sum_{r=0}^{10} r^2 f(r) = E(X^2) = 2\left\{\frac{5}{2} + 25\right\} = 55$

정답은 55

예제 063

REVIEW 확률변수의 변형

한 개의 동전을 $n$회 던질 때, 앞면이 나오는 횟수를 $X$라 한다. $E((2X-n)^2) = 10$을 만족시키는 $n$의 값은?

☐ 8 ☐ 9 ☐ 10 ☐ 11 ☐ 12

조건 한 개의 동전을 $n$회 던질 때, 앞면이 나오는 횟수를 $X$

⇨ $X$는 $B\left(n, \frac{1}{2}\right)$를 따르므로 $E(X) = \frac{n}{2}$, $V(X) = \frac{n}{4}$

⇨ $E((2X-n)^2) = 10 \Leftrightarrow E(4X^2 - 4nX + n^2) = 10 \Leftrightarrow 4E(X^2) - 4nE(X) + n^2 = 10$

$\Leftrightarrow 4\{V(X) + E(X)^2\} - 4nE(X) + n^2 = 10$

$\Leftrightarrow 4\left\{\frac{n}{4} + \frac{n^2}{4}\right\} - 4n \cdot \frac{n}{2} + n^2 = 10$

$\Leftrightarrow n = 10$

정답은 ③

예제 064

REVIEW 확률변수의 변형

한 개의 주사위를 3회 던져서 1의 눈이 나오는 횟수 $X$에 대하여 $4^X$원의 상금을 받기로 할 때, 상금의 기댓값을 구하여라.

조건 한 개의 주사위를 3회 던져서 1의 눈이 나오는 횟수 $X$

⇨ $X$는 $B\left(3, \frac{1}{6}\right)$을 따른다.

조건 $4^X$원의 상금을 받기로 할 때, 상금의 기댓값

⇨ $E(4^X) = \sum_{r=0}^{3} 4^r \times {}_3C_r \left(\frac{1}{6}\right)^r \left(\frac{5}{6}\right)^{3-r} = \sum_{r=0}^{3} {}_3C_r \left(\frac{4}{6}\right)^r \left(\frac{5}{6}\right)^{3-r}$ - $4^r \times \left(\frac{1}{6}\right)^r$을 $\left(\frac{4}{6}\right)^r$으로 표현했다.

⇨ 0부터 끝까지 변동수와 같은 지수 지수의 합이 고정수 5

$= \left(\frac{4}{6} + \frac{5}{6}\right)^3 = \left(\frac{3}{2}\right)^3 = \frac{27}{8}$

정답은 $\frac{27}{8}$원

**예제 065**

REVIEW 확률변수의 변형

원점 O를 출발하여 수직선 위를 움직이는 점 P가 있다. 동전 한 개를 던져서 앞면이 나오면 양의 방향으로 3만큼, 뒷면이 나오면 음의 방향으로 2만큼 점 P를 이동시킨다. 동전을 10번 던진 후의 점 P의 좌표를 확률변수 $X$라 할 때, $\mathrm{E}(X)$는?

① 1 ② 3 ③ 4 ④ 5 ⑤ 10

조건 동전 한 개를 던져서 앞면이 나오면 양의 방향으로 3만큼, 뒷면이 나오면 음의 방향으로 2만큼 점 P를 이동

\+ 동전을 10번 던진 후의 점 P의 좌표를 확률변수 $X$

⇨ 동일 시행을 10번 반복하는 것에서 이항분포라는 힌트를 얻을 수 있다.

- 나올 수 있는 모든 확률변수를 <표>로 나타낸다.

| $X$ | $-20$ | $-15$ | $-10$ | $\cdots$ | $30$ | 계 |
|---|---|---|---|---|---|---|
| $P(X=x)$ | | | | | | 1 |

- 1회의 시행에서 앞면 사건을 $A$라 할 때 $P(A) = \frac{1}{2}$이다.

- : $P(-20)$은 10회의 독립시행 중 사건 $A$가 0회($A^c$이 10회) 발생할 확률과 같다.

즉, 이 상황은 <독립시행의 확률공식>을 쓸 수 있는 상황이므로 에 들어갈 확률은 ${}_{10}C_0\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^{10}$

- : $P(-15)$는 10회의 독립시행 중 사건 $A$가 1회($A^c$이 9회) 발생할 확률과 같다.

즉, 이 상황은 <독립시행의 확률공식>을 쓸 수 있는 상황이므로 에 들어갈 확률은 ${}_{10}C_0\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^9$

⇨ 이와 같은 과정을 통해 주어진 확률분포는 이항분포 $B\left(10,\ \frac{1}{2}\right)$과 관련이 있다는 사실은 알았다.

이항분포 $B\left(10,\ \frac{1}{2}\right)$의 확률변수를 $K$라 하고 <표>를 생각해보면 $X$와 $K$의 관계를 찾을 수 있다.

| $K$ | $0$ | $1$ | $2$ | | $10$ | |
|---|---|---|---|---|---|---|
| $X$ | $-20$ | $-15$ | $-10$ | $\cdots$ | $30$ | 계 |
| $P(X=x)$ | | | | | | 1 |

이 표에서 $X = 5K-20$임을 알 수 있다.

⇨ 즉, $E(X) = E(5K-20) = E(5K-20) = 5E(K)-20 = 5\cdot 10\cdot \frac{1}{2}-20 = 5$

정답은 ④

티칭 $X$와 $K$의 관계를 찾을 때 $X$는 <등차수열의 규칙>임을 알 수 있다.

일반적인 <등차수열>은 <초항>이 $n=1$에 대응하는 값임에 비해 이 문제에서 $X$는 $K=0$에 대응하는 값이다.

이 부분은 이런 류의 문제에서 항상 주의해야 하는 부분이다.

| 존철살인 확·통 | 3. 통계 |
|---|---|
| Part 5. 연속확률분포와 정규분포 | #1. 연속확률분포 |
| | #2. 정규분포 |
| | #3. 이항분포의 정규분포 화 |

## 촌철살인 #5. 빙산의 일각

- 연속확률분포는 아주 작은 빙산의 일각을 공부하는 것으로 조금만 학습하면 매우 쉽게 다가온다.
  단지 그 과정이나 근본적인 것들에 대한 물음을 시작하면 한도 끝도 없이 어려울 수 있다. 그래서 이 파트에서는 교과과정에서 벗어나지 않은 범위 내에서 배우는 이유와 느낌을 최대한 설명한다.
  이해간다고 느꼈다면 문제는 매우 쉬움을 알 수 있다.

MAP

## #1 연속확률분포

- 연속확률분포와 확률밀도함수는 그냥 정의되는 것이다. 우리가 공부한 내용으로부터 자연스럽게 이끌어내는 것은 불가능하다고 봐야 한다.
- 그럼에도 불구하고 <연속확률분포>를 좀 더 친근하게 받아들이기 위해 고등학교 지식의 수준에서 직관적으로 이해할 수 있는 것들은 설명하기로 한다.

# 1. 연속확률분포

### 1) 용어와 의미

① 연속확률변수 : 문제에서 정의되는 변수 (이산확률변수는 <시행에 따른 모든 결과>였다.)

② 확률밀도함수 : 문제에서 정의되는 함수 (이산확률분포에서는 확률질량함수가 있었다.)

- 연속확률분포에서 확률변수와 확률밀도함수는 반드시 주어진다.

③ 확률 : 확률밀도함수의 <정적분 값>

- 문 · 이과 모두 미적분을 학습하기 때문에 정확한 정의를 이해하는 데 무리가 없다.
  단지 적분의 표현을 통해 확률 값을 표현하는 것은 교육과정에서 삭제되었으므로 큰 시험에서 나오기 어렵다.

### 2) 정적분 : 실수계의 시그마

- 정적분의 의미는 실수계의 시그마이다.

⇨ 시그마 : $\sum_{k=a}^{b} A_k = A_a + A_{a+1} + A_{a+2} + \cdots + A_b$

시그마는 이처럼 <아래 끝 $a$>를 시작으로 <$1$씩 변화>시키며 <윗 끝 $b$>까지 대입하여 더한다는 것을 의미한다.

⇨ 정적분(인티그럴) : $\int_a^b f(x)dx = \sum_{x=a}^{b} f(x)dx = f(a)dx + f(a+dx)dx + f(a+2dx)dx + \cdots + f(b)dx$

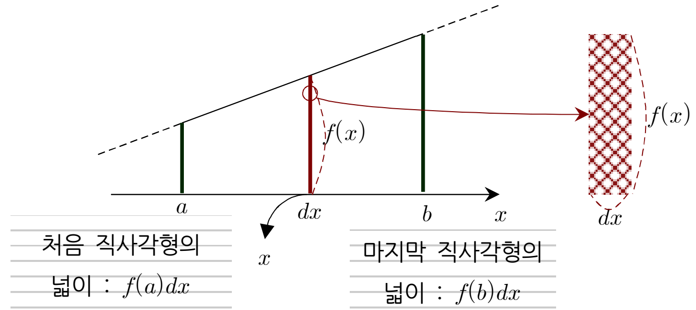
인티그럴도 의미상 <아래 끝 $a$>를 시작으로 <$dx$씩 변화>시키며 <윗 끝 $b$>까지 대입하여 더한다는 것을 의미한다.
단지 $dx$라는 기호는 <무한소>를 기호화 시킨 것으로 실제로는 <거의 $0$>이기 때문에 $a$에서 $b$까지 <연속적으로 더한다>는 의미를 가지고 있다.

정적분을 <실수계의 시그마>로 인지하는 것은 <정적분에 대한 올바른 관점>이며 많은 부분을 자연스럽게 이해시켜준다.

⇨ $dx$의 역할 : $\int_a^b f(x)dx$에서 $\int_a^b$가 <시그마의 역할>을 한다면 $f(x)dx$는 <일반항의 역할>을 한다.

이때 $f(x)$는 함숫값으로서 <부호를 가진 길이>의 의미를 갖는다. 수학에서 길이는 <넓이가 $0$>인 값으로 아무리 많이 더해도 심지어 $\infty$번 더하더라도 넓이는 영원히 $0$이다.

$\int_a^b f(x)dx$가 <부호를 가진 넓이>의 값을 갖는다는 사실을 안다고 가정하면 이 식이 <넓이의 의미>를 갖는 이유가 바로 이 $dx$에 있다. $dx$는 <아주 작은 밑변>을 의미하고 그래서 $f(x)$는 길이이지만 $f(x)dx$는 <아주 작은 넓이>의 의미를 갖는 것이다. (다음 장의 그림을 참조하여라.)



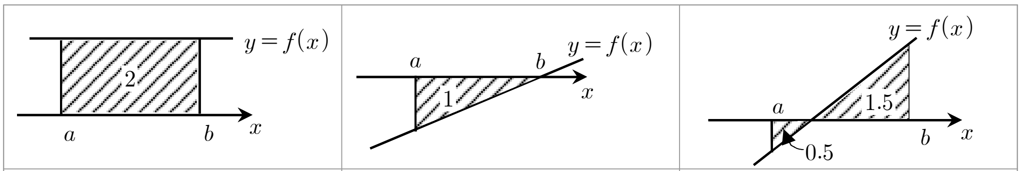

즉, $\int_a^b f(x)dx = f(a)dx + f(a+dx)dx + f(a+2dx)dx + \cdots + f(b)dx$의 의미를 다시 한 번 생각해 보아라.

- $0$과 <무한소>는 완전히 다른 것이다. $0$은 아무리 많이 모아도 $0$이지만 <$0$에 가깝지만 $0$은 아닌 무한소>의 경우에는 <아주 많이 모으면> <어떤 값>이 되거나 심지어 <무한대>가 될 수도 있다.
- 정적분이 <넓이>라는 것을 인정한다면 $f(x)$는 <넓이의 입장>에서 <$0$>을 의미하지만 $f(x)dx$는 <넓이의 입장>에서 <무한소>를 의미한다.

## 3) 확률밀도함수의 특징

① 정의된 구간에서 <정적분 값 $=1$>이다.
② 정의된 구간에서 함숫값은 항상 양수이다.
 - <확률 값>은 음수일 수 없다. 만약 그래프가 정의된 구간에서 음수가 나오게 되면 <확률 값>인 <정적분 값>이 음수가 나오게 되므로 잘못된 함수이다.

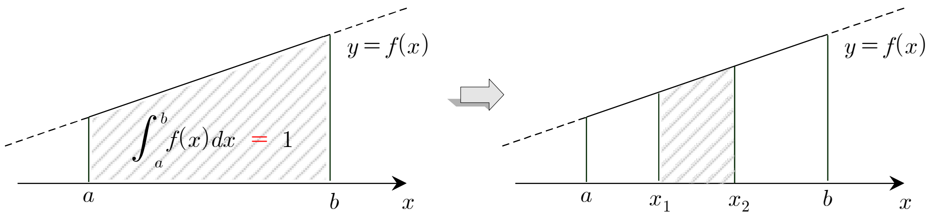
⇨ 정의된 구간이 $[a,\ b]$일 때, 확률밀도함수가 될 수 없는 함수를 관찰하면 정확하게 알 수 있다.




[그림 1] [그림 2] [그림 3]

- [그림 1]은 정의된 구간에서 <정적분 값 $=2$>이므로 확률밀도함수가 될 수 없다. 확률 값은 $1$을 넘을 수 없다.
- [그림 2]는 정의된 구간에 <넓이는 $1$>이지만 <정적분 값 $=-1$>이므로 확률밀도함수가 될 수 없다.
- [그림 3]은 정의된 구간에 <정적분 값 $=1$>이지만 함숫값이 음수인 부분이 포함되어 있으므로 확률밀도함수가 될 수 없다. 확률 값은 음수가 될 수 없다.

⇨ 구간 $[a,\ b]$에서 정의된 확률밀도함수 $y=f(x)$가 다음 그림과 같다고 해보자.




- $P(x_1 \le X \le x_2) = P(x_1 \le X < x_2) = P(x_1 < X \le x_2) = P(x_1 < X < x_2) = \int_{x_1}^{x_2} f(x)dx$

## 4) 질량과 밀도의 차이

- 〈거리〉는 그 자체로 느껴지는 〈실제 값〉인데 반해 〈속력〉은 그 자체로는 느낄 수 없는 〈비율 값〉이다.
  〈속력〉을 말로 풀어보면 〈단위시간 당 움직인 거리〉로서 역시 〈거리〉라는 말을 이용하여 정의하지만 〈시간〉이라는 물리량이 없다면 그 자체로 〈거리〉를 느낄 수가 없다.
  예를 들어 속력이 500인 물체와 속력이 5인 물체가 움직이고 있을 때, 어느 순간 사진을 찍으면 우리는 두 물체의 속력을 느낄 수가 없다. 〈시간이 0〉인 상태에서는 둘 다 움직이지 않기 때문이다. 하지만 〈1초 동안 관찰〉한다면 각 물체는 〈실제 움직인 거리〉를 가지게 되고, 이것을 통해 우리는 〈속력〉을 느낄 수 있다.
  이처럼 〈$\dfrac{\text{거리}}{\text{시간}}$ : 비율 값〉으로 정해지는 〈속력〉은 〈시간〉이라는 물리량이 있어야만 〈실제 값〉이 된다.

- 〈질량〉과 〈밀도〉 역시 이와 같은 관계를 가지고 있다.
  〈$\dfrac{\text{질량}}{\text{부피}}$〉으로 정의되는 〈밀도〉는 (〈속력〉이 〈시간〉이라는 물리량을 곱해야 〈실제 값〉인 거리의 의미를 갖는 것처럼) 〈부피〉라는 물리량이 곱해져야만 〈실제 값〉인 〈질량〉의 의미를 갖게 된다.
  ⇨ 이산확률분포에서 $P(X=x_i) = p_i \ (i=1,\ 2,\ \cdots,\ n)$은 〈함숫 값〉 자체가 〈실제 확률 값〉의 의미를 가지기 때문에 〈확률질량함수〉라고 하기로 했다.
  ⇨ 연속확률분포에서 $y=f(x) \ (\alpha \le x \le \beta)$는 〈함숫 값〉 자체가 〈실제 확률 값〉이 되지 못하고 $dx$라는 〈양〉을 곱해야만 〈실제 확률 값〉이 되기 때문에 〈확률밀도함수〉라고 하기로 했다.
  예를 들어 $f(x) = 2x \ (0 \le x \le 1)$인 확률밀도함수에서 $f(1) = 2$인데 이것은 당연히 〈확률 값〉이 아니다. 〈확률 값〉은 반드시 $0 \le p \le 1$이어야 한다. 앞에서도 언급했듯이 $dx$는 〈무한소 : 거의 $0$〉을 표현한 기호이고 $f(1)\,dx$는 $0$으로 수렴한다. 즉, $P(1)$은 사실 $f(1)$이 아닌 $f(1)\,dx$의 의미이고 그 값은 $0$이다.
  그래서 연속확률분포에서 확률은 반드시 특정한 수가 아닌 구간으로 표시되어야 한다.

- 이것이 이산확률분포와 연속확률분포에서 각 함수에 〈밀도〉와 〈질량〉을 붙인 이유이다.

## 5) 평균, 분산, 표준편차와 성질

- 정적분이 〈실수계의 시그마〉라는 사실만 정확히 이해하면 사실 〈이산확률분포〉와 전혀 다르지 않은 내용임을 알 수 있다.
  확률밀도함수 $y=f(x)$ 구간 $[a,\ b]$에서 정의 될 때, $\left(\text{단, } \displaystyle\int_a^b f(x)dx = 1\right)$

① $E(X) = \displaystyle\int_a^b x\,f(x)dx$은
$= \sum x_i p_i$ 과 같은 의미이다.

② $V(X) = \displaystyle\int_a^b (x-m)^2 f(x)dx$ (편차제곱의 평균) $= \displaystyle\int_a^b x^2 f(x)dx - \left(\int_a^b x f(x)dx\right)^2$ (제평 − 평제)
$= \sum (x_i - m)^2 p_i$ (편차제곱의 평균) $= \sum {x_i}^2 p_i - \left(\sum x_i p_i\right)^2$ (제평 − 평제)

③ $\sigma(X) = \sqrt{V(X)}$

- 적분을 이용하여 〈평균, 분산, 표준편차〉를 구하는 것은 나오지 않는다. 이 부분을 직관적으로 이해할 수 있다면 교과서 상에서 그냥 외워야 하는 다음 내용을 자연스럽게 받아들일 수 있다.

- 이산확률분포에서 <평균, 분산, 표준편차의 성질>이 간단한 시그마의 성질을 통해서 증명되듯이
연속확률분포에서도 <평균, 분산, 표준편차의 성질>은 간단한 정적분의 성질을 통해서 증명된다.
(시그마의 성질과 정적분의 성질은 같다.)

이것은 교육과정에 벗어나지 않으며 반드시 알아야 한다. 교과서에서 아래 식은 <증명하지 않고 사용>하기로 한다.
이 세 식은 연속확률분에서도 그대로 성립한다.

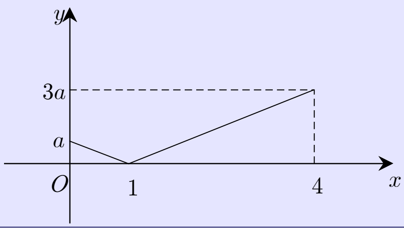
⇨ ① $E(aX+b) = aE(X)+b$ ② $V(aX+b) = a^2V(X)$ ③ $\sigma(aX+b) = |a|\sigma(X)$

- 문 · 이과 모두 적분의 개념을 배우므로 <개념의 외연에서 증명>하기로 한다. 필수는 아니다.

코칭 교과과정을 벗어났다는 시비에서 자유로울 순 없지만.
<정적분을 이용하여> 확률밀도함수를 설명한 것은 <당연히 옳은 설명>이다. 이것을 설명한 이유는 공부를 하다보면 <자연스럽게 드는 의문>을 해결하기 위함이지 <적분> 이용해서 확률을 구하기 위함이 아님을 알아두자.
2)과 4)는 문제를 풀기 위한 개념이 아니며 실제로 이 부분의 문제는 매우 쉽다.

**114** 이해를 위한 예제 2009. 11. 나형(62%). 21번. 4점

연속확률변수 $X$가 갖는 값의 범위는 $0 \leq X \leq 4$이고 $X$의 확률밀도함수의 그래프는 다음과 같다. $100P(0 \leq X \leq 2)$의 값을 구하시오.



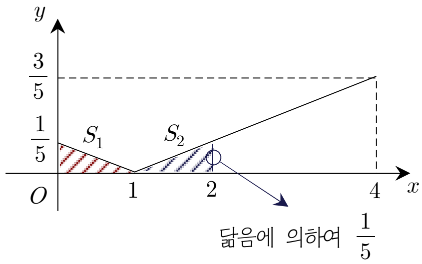

조건 연속확률변수 $X$가 갖는 값의 범위는 $0 \leq X \leq 4$

⇨ 정의된 구간에서 총 넓이는 1이어야 하므로 $\frac{1}{2}a + \frac{1}{2} \cdot 3 \cdot 3a = 1 \Leftrightarrow 5a = 1 \Leftrightarrow a = \frac{1}{5}$

구할 것 $P(0 \leq X \leq 2)$는 오른쪽 빗금친 부분의 넓이이므로

$S_1 + S_2 = \frac{1}{10} + \frac{1}{10} = \frac{1}{5}$이다.

즉, $P(0 \leq X \leq 2) = 100 \times \frac{1}{5} = 20$




정답은 20

# #2 정규분포와 표준정규분포

- 세상의 모든 분포는 정규분포를 따른다. 정규분포는 세상을 구성하는 분포이다.
- 정규분포의 이론은 유도되어 나오지 않는다. 실제로는 유도되어 나오지만 이것은 수학과 전공과정에서 해야 한다. 즉, 정규분포는 〈정의되는 것〉으로 이해하고 그 특징과 성질을 정확히 기억하여 문제에 적용하면 그만이다.

## 1. 정규분포와 확률밀도함수

### 1) 정규분포 뜻과 기호

① 연속확률변수 : 연속확률변수

② 확률밀도함수 : $f(x) = \dfrac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}\ (-\infty < x < \infty)$

- 이 함수를 외울 필요는 전혀 없다. $\pi = 3.14\times\times$를 의미하는 〈상수〉인 것처럼 $e = 2.71\times\times$를 의미하는 상수이다. 결국 이 그래프를 결정하는 것은 평균 $m$과 표준편차 $\sigma$이다.

③ $N(m,\ \sigma^2)$ : 여기에서 $N$은 Normal distribution의 약자이다. 바로 위에서 언급한 것처럼 정규분포의 확률밀도함수의 그래프는 $m$과 $\sigma$에 의해서 결정되므로 $N(m,\ \sigma^2)$의 기호를 통해 $m$과 $\sigma$를 조건으로 준다.

④ 확률 : 역시 확률밀도함수의 〈정적분 값〉으로 정의 된다. 하지만 위의 함수를 적분하는 것은 고등과정에서 한참 벗어난 너무 어려운 것이다. (이 적분을 가우스 적분이라고 하는 데 심지어 켈빈은 이 적분이 당연해 보이면 그 사람을 수학자라고 부른다고 했다.)
그래서 우리는 〈이미 계산된 적분의 의미를 가지고 있는 표〉를 보는 규칙만 공부한다.

티칭 알고 보면 간단한 $f(x) = \dfrac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}\ (-\infty < x < \infty)$의 식의 구조

⇨ $\dfrac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}$ 의 식에서 〈상수〉로 처리되는 부분을 잘 살표보면 식의 구조는 생각보다 간단하다.

- $\pi$, $e$은 수학에서 항상 불변의 상수(무리수)이다.
- $m$, $\sigma$는 〈정규분포〉에 따라 결정되는 값으로 〈상수〉라고 생각할 수 있다. 즉, $x$ 이외에는 모두 상수이다.

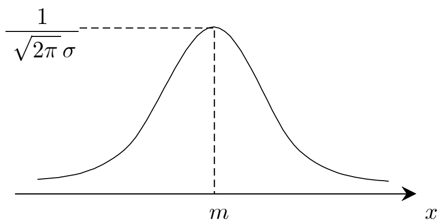
⇨ 그러므로 $y = \dfrac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}$ 은 $y = e^{-x^2}$과 비슷한 〈그래프의 개형〉을 가진다.

$y = e^{-x^2} = \dfrac{1}{e^{x^2}}$이 최댓값을 가지는 순간은 〈분모〉가 〈가장 작은 값〉을 가지는 순간이므로 $x=0$인 순간이다.

마찬가지로 $y = \dfrac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}$ 이 최댓값을 가지는 순간은 $x=m$인 순간이고 그때의 값은 $\dfrac{1}{\sqrt{2\pi}\,\sigma}$이다.

## 2) 정규분포의 그래프

- 정규분포 문제는 항상 그래프를 도구로 푸는 것이 좋다. 그래서 모든 문제의 해설에는 그래프가 들어간다.
  아래 그래프의 성질들은 한 번 정독한 후 <문제>를 통해 자연스럽게 익히면 된다.




그래프의 기본 성질 (함수적 성질)

1. 정의역이 실수 전체
2. $x$축을 점근선으로 갖는다.
3. 좌우 대칭이다. - 중요

⇨ 모든 $x$에 대하여 $f(m-x) = f(m+x)$

$N(m, \sigma^2)$의 확률밀도함수는

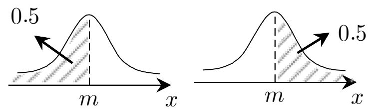
$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}$ $(-\infty < x < \infty)$이고

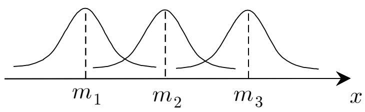
이 함수에 대한 그래프이다. 그래프의 대칭성은 식의 구조를 관찰하면 고등학교 범위 내에서도 충분히 추론이 가능하지만 그냥 외워도 상관없다.

전체 넓이는 1이다. 즉, $\int_{-\infty}^{\infty} f(x)dx = 1$

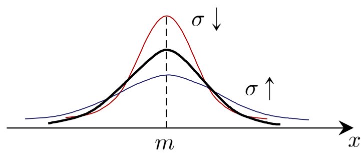
그런데 평균에 대해 좌우 대칭이므로

$P(x \ge m) = P(x \le m) = 0.5$




평균은 그래프의 모양에 영향을 주지 못한다.
표준편차는 같고 $m_1 < m_2 < m_3$인 경우




표준편차가 그래프의 모양을 결정한다.
- 표준편차가 작아지면 평균 주위로 몰린다.
- 표준편차가 커지면 평균 밖으로 흩어진다.

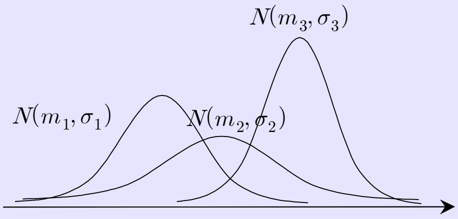
- 함수적 설명 : 정규분포의 그래프는 항상 최댓값 $\frac{1}{\sqrt{2\pi}\,\sigma}$을 가진다. (참고이므로 이해가 안 가면 넘겨도 좋다.)

⇨ $\sigma\downarrow$ 이면 $\frac{1}{\sqrt{2\pi}\,\sigma}\uparrow$ 이다. 즉, 그래프가 높아진다. $\sigma\uparrow$ 이면 $\frac{1}{\sqrt{2\pi}\,\sigma}\downarrow$ 이다. 즉, 그래프가 낮아진다.

티칭 이항분포 $B(n,\ p)$와 기호가 비슷해 보이지만 안에 들어가는 수치는 전혀 상관없다.
$N(m,\ \sigma^2)$과 $B(n,\ p)$에 들어가는 각 수치의 의미를 정확하게 기억한다.
- $B(n,\ p)$ : $B$(총 시행횟수, 1회 시행에서 사건 $A$가 발생할 확률)
- $N(m,\ \sigma^2)$ : $N$(평균, 분산(보통은 표준편차$^2$으로 표현))

**115** 이해를 위한 예제

다음 세 정규분포에 대하여, $m_1, m_2, m_3$와 $\sigma_1, \sigma_2, \sigma_3$를 각각 대소비교 하여라.




- 당연히 $m$은 오른쪽에 있을수록 크다. 즉, $m_1 < m_2 < m_3$이다.
- $\sigma$는 그래프가 높이가 높을수록 작다. 즉, $\sigma_3 < \sigma_1 < \sigma_2$이다.

정답은 $m_1 < m_2 < m_3$이고, $\sigma_3 < \sigma_1 < \sigma_2$

**116** 이해를 위한 예제 2006. 4. 가형(70%). 5번. 3점

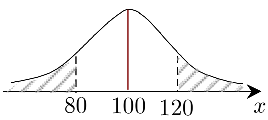
확률변수 $X$는 정규분포 $\mathrm{N}(m, \sigma^2)$을 따른다. $\frac{1}{5}X$의 분산이 1이고 $\mathrm{P}(X \leqq 80) = \mathrm{P}(X \geqq 120)$일 때, $m + \sigma^2$의 값은?

① 105 ② 110 ③ 115 ④ 120 ⑤ 125

조건 $\frac{1}{5}X$의 분산이 1 - $V\left(\frac{1}{5}X\right) = 1 \Leftrightarrow \frac{1}{25}V(X) = 1 \Leftrightarrow V(X) = 25$ 즉, $\sigma^2 = 25$이다.

조건 $\mathrm{P}(X \leqq 80) = \mathrm{P}(X \geqq 120)$일 때,

- 정규분포는 평균에 대해서 대칭이므로 오른쪽 그림과 같이 확률 값이 같아지기 위해서는 80과 120이 평균을 기준으로 대칭적으로 존재해야 한다. 즉, 평균은 100이다.




$m + \sigma^2 = 100 + 25$ 이므로 정답은 ⑤

**117** 이해를 위한 예제 2004. 9. 나형(38%). 29번. 4점

어느 회사에서 만든 휴대전화 배터리의 지속 시간은 평균 60시간인 정규분포를 따른다고 한다. 이 회사에서 만든 8개의 배터리 중에서 지속 시간이 60시간 이상인 배터리가 2개 이상일 확률은?

① $\frac{101}{256}$ ② $\frac{129}{256}$ ③ $\frac{197}{256}$ ④ $\frac{219}{256}$ ⑤ $\frac{247}{256}$

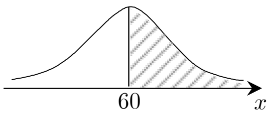
조건 휴대전화 배터리의 지속 시간은 평균 60시간인 정규분포

⇨ $N(60,\ \sigma^2)$인 정규본포의 확률변수는 <배터리 지속시간 : $X$>이다.

구할 것 8개의 배터리 중에서 지속 시간이 60시간 이상인 배터리가 2개 이상일 확률

⇨ 구하고자 하는 확률의 확률변수는 <60시간 이상인 배터리의 개수>이다.
8회의 독립시행에서 <특정결과>가 2번 나올 확률이므로 <독립시행의 확률 공식>을 쓸 수 있는 구조이다.
즉, <60시간 이상인 베터리의 개수>를 확률변수 $Y$라 하면 $Y$는 이항분포를 따른다.

- 정규분포의 대칭성을 통해서 <1회 시행에서 발생할 확률>준 것뿐이다.

$N(60,\ \sigma^2)$ :  에서 <60시간 이상일 확률은 $\frac{1}{2}$>이다.



즉, $Y$는 $B\left(8,\ \frac{1}{2}\right)$인 이항분포를 따른다.

- 1회 시행에서 사건이 발생할 확률이 $\frac{1}{2}$이고, 8회의 독립시행에서 이 사건이 2회 이상 발생할 확률이므로

| $X$ | 0 | 1 | 2 | $\cdots$ | 10 | 총합 |
|---|---|---|---|---|---|---|
| $P(X)$ | ${}_8C_0\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^8$ | ${}_8C_1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^7$ | | $\cdots$ | | 1 |

결국 우리가 구하는 $P(Y\geq 2) = 1 - \{P(Y=0)+P(Y=1)\}$

$$= 1 - \left\{{}_8C_0\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^8 + {}_8C_1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^7\right\} = 1 - \frac{9}{256} = \frac{247}{256}$$

정답은 ⑤

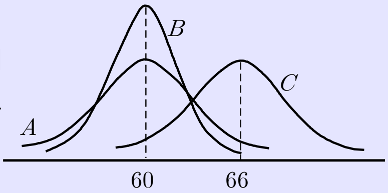
코칭 생각도 습관이다.

위의 해설에서 왜 굳이 표를 넣어놨는지 의문을 갖는 사람도 있을 것이다. 물론 실제로 문제를 풀 때는 표를 그리지 않는다. (머릿속에 있기 때문이다.)

이항분포가 나왔을 때 $B(n,\ p)$는 표의 정보를 요약해서 보여준 것뿐이므로 자연스럽게 표를 매개로 생각해야 한다. 이것을 강조하기 위해 표를 한 번 더 실어놓았다.

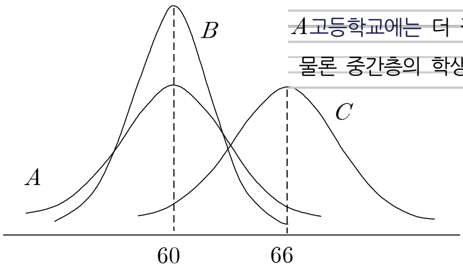
**118** 이해를 위한 예제 고르다는 말

3학년 학생의 수가 각각 300명인 같은 지역의 세 고등학교($A, B, C$) 3학년 학생의 수학 성적 분포가 각각 정규분포를 이루고 그 정규분포 곡선은 그림과 같다고 한다. 다음 중에서 옳은 것은?




ㄱ. 성적이 우수한 학생들이 $B$고등학교보다 $A$고등학교에 더 많이 있다.
ㄴ. $B$고등학교 학생들은 평균적으로 $A$고등학교 학생들보다 성적이 더 우수하다.
ㄷ. $C$고등학교 학생들보다 $B$고등학교 학생들의 성적이 더 고른 편이다.

☐ ㄱ ☐ ㄴ ☐ ㄷ ☐ ㄱ, ㄷ ☐ ㄴ, ㄷ




$A$고등학교에는 더 잘하는 학생도 많고, 더 못하는 학생도 많다.
물론 중간층의 학생들은 $B$고등학교에 더 많다. ㄱ은 참

$C$고등학교는 $B$고등학교보다 그래프가 낮으므로 표준편차가 더 크다.
(평균 밖으로 더 많이 흩어져 있다.)
즉, $B$고등학교학생들의 성적이 더 고르다. ㄷ은 참

$A$고등학교는 $B$고등학교와 평균은 같고 $B$고등학교보다 표준편차는 더 크다.
두 학교는 평균적으로 성적이 같다. ㄴ은 거짓

정답은 ④

코칭 고르다는 말. 국어의 문제

- 여기에서 고르다는 말은 <학생들의 점수>가 들쑥날쑥하지 않는 말로 평균 주위에 점수가 모여 있다는 뜻이다.
  (적어도 문제집들에서는 그렇게 본다.)

- 일부 학생들은 정수가 고르면 <다양하게 분포해야 하는 것>아니냐? 라는 의문을 제기한다.
  점수가 고르게 존재하면서 <한 가지 점수>에만 치우쳐 있으면 안 된다고 생각한 듯하다.
  만약에 나온다면 타당한 의문제기이며 그래서 논란의 소지가 있으므로 <큰 시험>에는 다른 표현을 쓸 것이다.
  어차피 이런 내용들이 수학에서 일일이 정의되어 있는 것이 아니다. 오해의 소지가 없도록 문제를 내야 한다.

- 이 문제를 실어놓은 이유는 이런 <답없는 물음>으로 시간을 허비하지 말라는 뜻이다. 문제집을 풀다보면 반드시 한 번은 부딪힐 것이다.

# 2. 정규분포의 확률 표현

- 정규분포의 확률 값(정적분 값)은 어떤 규칙을 갖는다. 그 규칙에 대해 학습하기만 하면 정규분포는 그 자체로는 참 쉬운 주제이다.
- 이것은 원리적으로 이해하기 위해서는 <가우스 적분>과 <치환적분>을 알아야 한다. 원리가 궁금하다면 후에 대학에서 공부하기로 하고 일단 <규칙>을 받아들이고 시작한다.

## 1) 정규분포의 확률을 표현하는 규칙

- 확률을 구하는 경계되는 값을 <평균>을 기준으로 <표준편차>의 <몇 배>를 통해서 표현한다. 정규분포에서 <특정 확률 변수>에서 확률 값은 <0으로 수렴>한다. 즉, 항상 확률은 <확률변수의 범위>를 통해서 표현되어야 한다. 이때 이 <범위의 경계 값>을 <$m \pm k\sigma$>의 형태로 표현하면 <미리 계산되어 있는 표>에서 확률 값의 정보를 찾을 수 있게 된다.

⇨ 예를 들어 $X : N(50,\ 2^2)$에서 $P(48 \le X \le 53)$라면 이 범위의 경계 값인 $48$은 $m-\sigma$라고 표현할 수 있고 $53$은 $m+1.5\sigma$라고 표현할 수 있다.

즉, $P(48 \le X \le 53)$ ⇨ $P(m-\sigma \le X \le m+1.5\sigma)$

예를 들어 $Y : N(15,\ 3^2)$에서 $P(Y \le 21)$라면 이 범위의 경계 값인 $21$은 $m+2\sigma$라고 표현할 수 있다.

즉, $P(Y \le 21)$ ⇨ $P(Y \le m+2\sigma)$

이때 $m$과 $\sigma$는 그 확률변수가 속한 집단의 $m$과 $\sigma$이면 된다.

## 2) 정규분포의 확률 값을 주는 방법

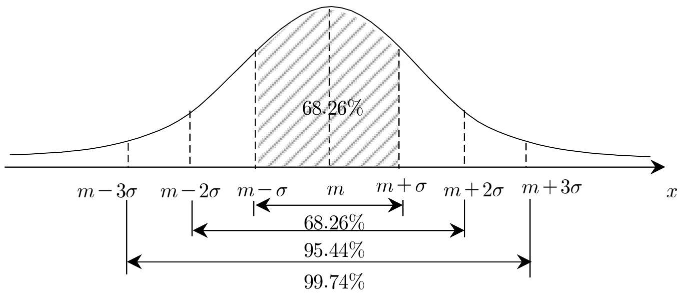
- 식을 통해서 확률 값을 주는 경우

① $P(m-\sigma \le X \le m+\sigma) = 0.6826$
② $P(m-2\sigma \le X \le m+2\sigma) = 0.9544$
③ $P(m-3\sigma \le X \le m+3\sigma) = 0.9974$

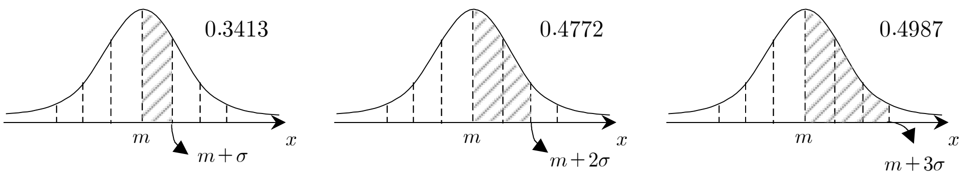
- 표를 통해서 확률 값을 주는 경우

| $x$ | $P(m \le X \le x)$ |
|---|---|
| $m+\sigma$ | $0.3413$ |
| $m+2\sigma$ | $0.4772$ |
| $m+3\sigma$ | $0.4987$ |

- 그래프와 <정적분 값>을 통해서 확률 값을 주는 경우




- 결국 어떤 조건이 나오더라도 다음과 같이 <정규분포의 그래프>를 통해서 <확률 값>을 인지한다. 항상 그래프를 매개로 생각해야 <논리적으로> 문제를 푼다는 느낌을 가질 수 있다.

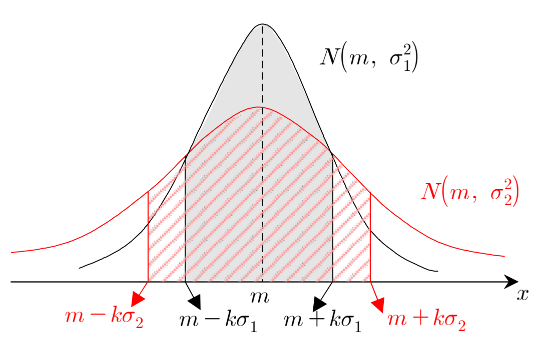


## 3) 같을 수밖에 없는 이유

- 정규분포 집단마다 $m$과 $\sigma$가 달라도 경계 값이 <같은 식>으로 표현되면 <확률 값>이 같다.
  정확한 이유는 <수식적으로 설명>되는 부분이지만 <직관적>으로 어느 정도까지는 납득할 수 있다.
  (수식적 설명은 고등범위를 많이 넘어간다.)

- 평균이 다르고 표준편차가 같은 두 정규분포는 <그래프가 평행이동> 되는 것뿐이므로 $k$값이 같으면 당연히 같은 <정적분 값 = 확률 값>을 갖는다.

- 평균이 같고, 표준편차가 다른 경우는 그래프의 모양자체가 변하므로 아래 그림을 통해 직관적으로만 이해하고 넘어가도록 한다.




평균이 같은 두 정규분포 $N(m,\ \sigma_1^2)$, $N(m,\ \sigma_2^2)$에 대하여

- $\sigma_1 < \sigma_2$라면 표준편차가 $\sigma_1$인 정규분포에 비해 표준편차가 $\sigma_2$인 정규분포는 그래프가 낮다.
- 대신 동일한 $k$값에 대하여 $m+k\sigma_1$에 비해 $m+k\sigma_2$가 더 크기 때문에 결국 왼쪽 그래프에서 <정적분 값>은 같다.
- 의 면적과 과 의 면적 합이 서로 상쇄된다고 생각하면 된다.

## 4) 확률변수의 위치 (내 점수의 위치)

- 특정한 확률변수의 상대적 위치를 파악하기 위해서 <편차>와 <표준편차>의 관계를 알면 된다.
  편차란 $X-m$의 값으로서 나의 상대적 위치를 대략적으로 알려주는 값이다.
  참조 (Part 4. #1. 중등통계에서 배우다.) - P. 237
  예를 들어 <나의 편차점수>가 양수라면 나는 평균보다 잘하는 것이고, <나의 편차점수>가 음수라면 나는 평균보다 못하는 것이다. 이렇게 대략적으로 알 수 있었던 <나의 상대적 위치>에 대해 <정규분포>를 활용하면 좀 더 구체적으로 알 수 있게 된다.

- <백분위>라는 것은 <크기가 클수록> 상위에 위치할 수 있도록 <의미가 변환된 백분율>을 뜻한다.
  즉, 상위 1%는 백분위 99%에 대응된다. 백분위는 클수록 상위에 위치함을 알 수 있다.

- 오른쪽 표를 참조하여 좀 더 구체적인 수치를 통해서 이해해보자.

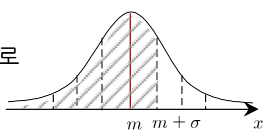
| $x$ | $P(m \le X \le x)$ |
|---|---|
| $m+\sigma$ | 0.3413 |
| $m+2\sigma$ | 0.4772 |
| $m+3\sigma$ | 0.4987 |

예시 1) 오른쪽 표에서

$P(X \le m+\sigma) = 0.5 + 0.3413 = 0.8413$이므로

$X_1 = m+\sigma$로 표현되는 확률변수 $X_1$은

약 백분위 $84.13\%$라는 것을 알 수 있다.



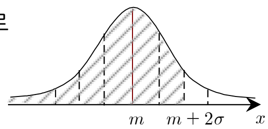

⇨ 즉, <$X_1 - m$_(편차) = $\sigma$_(표준편차)>에서 <나의 편차점수>가 표준편차의 <1배>로 표현되면 백분위 $84.13\%$로 <3등급>임을 알 수 있다.

예시 2) 오른쪽 표에서

$P(X \le m+2\sigma) = 0.5 + 0.4772 = 0.9772$이므로

$X_2 = m+2\sigma$로 표현되는 확률변수 $X_2$는

약 백분위 $97.72\%$라는 것을 알 수 있다.




| $x$ | $P(m \le X \le x)$ |
|---|---|
| $m+\sigma$ | 0.3413 |
| $m+2\sigma$ | 0.4772 |
| $m+3\sigma$ | 0.4987 |

⇨ 즉, <$X_2 - m$_(편차) $= 2\sigma$_(표준편차)>에서 <나의 편차점수>가 표준편차의 <2배>로 표현되면 백분위 $97.72\%$로 <1등급>임을 알 수 있다.

- 앞으로 <모의고사 성적표>를 받는다면 <자신의 백분위 점수>를 스스로 계산해 보자.
  교육청과 평가원사이트를 들어가면 <전국 평균과 표준편차>를 알 수 있기 때문에 <나의 편차점수>와 <표준편차>를 비교하고 <인터넷에 올라와 있는 정규분포 표>를 이용해보면 나의 백분위 점수를 거의 정확히 알 수 있다.
  요즘은 원점수를 공개하지 않지만 이 방법을 이용하면 <계산결과>와 <성적표>를 비교하여 자신의 원점수도 추론해 볼 수 있다. (가채점에 비해 점수가 더 나오거나 덜 나오는 경우도 있기 때문이다.)

## 119 이해를 위한 예제

오른쪽 표를 참고하여 다음을 구하여라.

(1) $P(m-2\sigma \le X \le m+3\sigma)$

(2) $P(X \ge m+2\sigma)$

(3) $P(X \ge m-\sigma)$

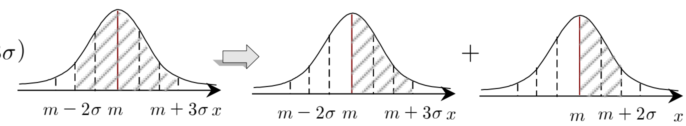
| $x$ | $P(m \le X \le x)$ |
|---|---|
| $m+\sigma$ | 0.3413 |
| $m+2\sigma$ | 0.4772 |
| $m+3\sigma$ | 0.4987 |

그래프로 사고하라.

(1) $P(m-2\sigma \le X \le m+3\sigma)$

$= P(m \le X \le m+2\sigma) + P(m \le X \le m+3\sigma)$

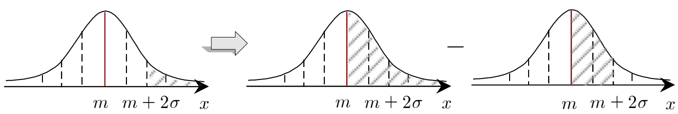
$= 0.4772 + 0.4987 = 0.9759$




(2) $P(X \ge m+2\sigma)$

$= 0.5 - P(m \le X \le m+2\sigma)$

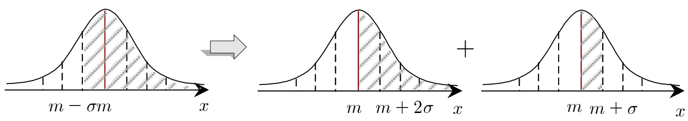
$= 0.5 - 0.4772 = 0.0228$




(3) $P(X \ge m-\sigma)$

$= 0.5 + P(m \le X \le m+\sigma)$

$= 0.5+0.3413 = 0.8413$




정답은 (1) $0.9759$, (2) $0.0228$, (3) $0.8413$

## 120 이해를 위한 예제

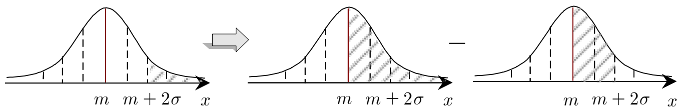
어느 학생들의 몸무게가 평균이 50, 분산이 4인 정규분포를 따른다고 한다.
그 중 한 명을 임의로 뽑을 때, 몸무게가 54이상일 확률을 오른쪽 표를 참고하여 구하여라.

| $x$ | $P(m \leq X \leq x)$ |
|---|---|
| $m+\sigma$ | 0.3413 |
| $m+2\sigma$ | 0.4772 |
| $m+3\sigma$ | 0.4987 |

조건 몸무게가 평균이 50, 분산이 4인 정규분포

⇨ 몸무게를 확률변수 $X$라고 할 경우 $N(50,\ 2^2)$이다. 경계가 되는 54를 〈평균 ± 표준편차의 몇 배〉로 표현한다.

$$P(X \geq 54) = P(X \geq 50+2\times 2)$$
$$= P(X \geq m+2\sigma)$$
$$= 0.5 - 0.4772 = 0.0228$$




정답은 0.0228

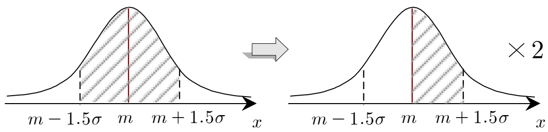
## 121 이해를 위한 예제

확률변수 $X$ 가 정규분포 $N(m,\sigma^2)$을 따를 때,
다음 표를 이용하여 $P(|X-m| \leqq 1.5\sigma)$를 구하면?

| $x$ | $P(m \leq X \leq x)$ |
|---|---|
| $m+\sigma$ | 0.3413 |
| $m+1.5\sigma$ | 0.4332 |

① 0.0668 ② 0.4332 ③ 0.5668 ④ 0.8664 ⑤ 0.9332

$$P(|X-m| \leqq 1.5\sigma) = P(-1.5\sigma \leq X-m \leq 1.5\sigma)$$
$$= P(m-1.5\sigma \leq X \leq m+1.5\sigma)$$
$$= 2\times P(m \leq X \leq m+1.5\sigma)$$
$$= 2\times 0.4332 = 0.8664$$

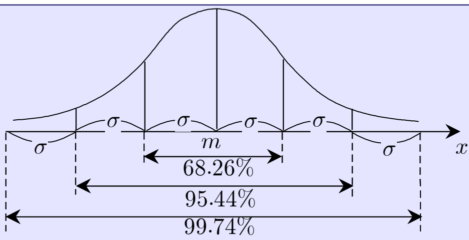



정답은 ④

## 122 이해를 위한 예제

정규분포곡선을 이용하여 확률변수 $X$가 정규분포
$N(80,\ 5^2)$을 따를 때, 다음 확률을 구하여라.

(1) $P(75 \leq X \leq 85)$ (2) $P(80 \leq X \leq 95)$




조건 확률변수 $X$ + $N(80,\ 5^2)$ 우리가 구하는 확률의 경계 값을 〈평균 ± 표준편차의 몇 배〉로 표현해야 한다.

(1) $P(75 \leq X \leq 85) = P(80-5 \leq X \leq 80+5)$
$= P(m-\sigma \leq X \leq m+\sigma) = 0.6826$ 〈위의 그래프에서 68.26%이므로〉

(2) $P(80 \leq X \leq 95) = P(80 \leq X \leq 80+3\times 5)$
$= P(m \leq X \leq m+3\sigma) = \dfrac{1}{2}\times 0.9974 = 0.4987$

정답은 (1) 0.6826 (2) 0.4987

**123** 이해를 위한 예제

확률변수 $X$가 정규분포 $N(53,\ 7^2)$을 따를 때, 다음 표를 이용하여 $P(X \le a) = 0.1587$을 만족하는 상수 $a$의 값을 구하면?
(단, $m$은 평균, $\sigma$는 표준편차)

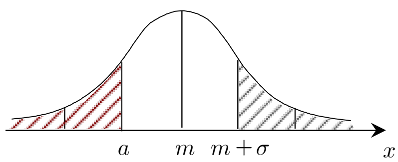
| $x$ | $P(m \leqq X \leqq x)$ |
|---|---|
| $m+\sigma$ | 0.3413 |
| $m+2\sigma$ | 0.4772 |
| $m+3\sigma$ | 0.4987 |

① 44 ② 46 ③ 48 ④ 50 ⑤ 52

조건 확률변수 $X$가 정규분포 $N(53,\ 7^2)$을 따를 때,

구할 것 $P(X \le a) = 0.1587$을 만족하는 상수 $a$

⇨ 표를 통해 $P(X \ge m+\sigma) = 0.1587$라는 사실을 알 수 있다.
오른쪽 그림과 같이 대칭이므로 $a = m-\sigma = 53-7 = 46$




정답은 ②

# 3. 정규분포의 표준화와 표준정규분포

## 1) 표준화

$N(m,\ \sigma^2)$을 따르는 한 확률변수 $X_1$에 대하여 : $X_1 \rightarrow \dfrac{X_1 - m}{\sigma}$

- 표준화란 <주어진 확률변수를 어떤 규칙에 의해 바꾸는 것>을 의미한다. 이 <바꾼 수>가 가지는 의미를 생각해보자. 어떤 정규분포를 따르는 <특정한 확률변수의 편차>를 <표준편차>로 나누면 이 값은 <특정한 확률변수>를 <평균 ± 표준편차의 몇 배>로 표현했을 때, 결국 <확률변수의 상대적 위치>를 결정하는 <몇 배>의 의미를 갖게 된다.

- 바로 앞에서 공부한 <4) 확률변수의 위치>에서 <편차>와 <표준편차>를 비교하여 나의 상대적 위치를 파악할 수 있다고 설명하였는데, 그냥 <표준화>를 시키면 바로 <몇 배>가 나오기 때문에 나의 상대적 위치를 더 쉽게 알 수 있다.

⇨ 예를 들어 $N(60,\ 5^2)$인 정규분포에서 확률변수 $X_1 = 70$이라면 이 확률변수를 표준화한 값 $\dfrac{70-60}{5} = 2$가 의미하는 것은 바로 <몇 배>이다. 즉, <표준화한 값이 $2$>라는 말은 <$X_1 = m + 2\sigma$>의 위치에 있다는 뜻이고, 이것은 아래 표를 참조하면 $X_1$의 백분위가 $97.72\%$에 해당한다는 사실을 알 수 있다.

| $x$ | $P(m \leq X \leq x)$ |
|---|---|
| $m+\sigma$ | $0.3413$ |
| $m+2\sigma$ | $0.4772$ |
| $m+3\sigma$ | $0.4987$ |

- 표준화를 이용하면 서로 다른 정규분포의 확률변수끼리도 쉽게 비교할 수 있다.

⇨ 예를 들어 어떤 학교의 수학점수가 정규분포 $N(60,\ 5^2)$을 따르고, 영어점수는 $N(80,\ 2^2)$인 정규분포를 따른다고 할 때, 민제의 수학점수는 $75$점, 영어점수는 $78$점이라고 해보자.
민제는 어떤 과목을 더 잘 봤는가?

: 민제의 수학점수를 표준화시키면 $\dfrac{75-60}{5} = 3$이고,

영어점수를 표준화시키면 $\dfrac{78-80}{2} = -1$이므로

위의 표를 참조하면 <수학점수의 백분위>는 $99.87\%$로 최상위권인데 반해
<영어점수의 백분위>는 $15.87\%$로 최하위권에 속한다.
즉, <확률변수의 상대적인 위치>를 알려줄 수 있는 수치는 <확률변수의 실제 값>이 아닌 <표준화 된 값>이다.

**124** 이해를 위한 예제

어떤 학급의 국어, 영어, 수학의 성적은 표와 같고 성적의 분포는 정규분포를 이룬다고 한다. $A$군의 성적을 학년 전체와 비교할 때, 3과목 중 상대적으로 가장 성적이 좋은 과목은?

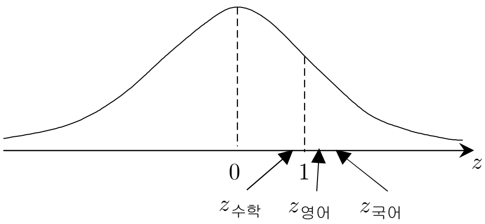
| 학과 | 국어 | 영어 | 수학 |
|---|---|---|---|
| 평균 | 50 | 64 | 62 |
| 표준편차 | 13 | 17 | 14 |
| A군 | 65 | 82 | 75 |

3과목 중 상대적으로 가장 성적이 좋은 과목 - 상대적으로 점수가 가장 좋은 과목은 표준화를 했을 때 가장 큰 값을 가지는 과목이다. 그러므로

$A$군의 국어점수 $x = 65 \Rightarrow z_{국어} = \frac{65-50}{13} = \frac{15}{13}$

$A$군의 영어점수 $y = 82 \Rightarrow z_{영어} = \frac{82-64}{17} = \frac{18}{17}$

$A$군의 수학점수 $y = 75 \Rightarrow z_{수학} = \frac{75-62}{14} = \frac{13}{14}$




그러므로 각 변량들을 표준정규분포에 표시하면 오른쪽과 같다. 상대적으로 가장 성적이 좋은 과목은 국어이다.

정답은 국어

## 2) 표준정규분포

$N(m,\ \sigma^2)$을 따르는 확률변수 $X$에 대하여 : $\dfrac{X-m}{\sigma} = Z$

- 이 부분을 직관적으로 부드럽게 받아들이는 것은 〈함수적 해석능력〉에 달려있다.
  간단하게만 설명하면 〈변수〉의 입장에서 표준화는 〈치환〉에 해당한다.

⇨ 예를 들어 $y = f(2x+3)$이라는 함수에서 $2x+3 = t$라고 치환할 경우 $y = f(t)$와 $t = 2x+3$이라는 두 함수가 탄생하게 된다. 이것이 〈수학 2〉에서 배운 〈합성함수〉를 사고하는 하나의 흐름이다.
복잡해 보이는 함수 $y = f(2x+3)$을 비교적 간단한 두 함수 〈$y = f(x)$와 $y = 2x+3$〉을 통해서 분석하는 〈사고방식〉이 이 안에 숨겨있다.

- 여기에서 $y = f(t)$와 $t = 2x+3$을 $x$와 $y$만 사용하여 다시 $y = f(x)$와 $y = 2x+3$으로 표현하는 것은 일종의 관용이다.

- 두 함수 $y = f(x)$와 $y = g(x)$의 관계를 따질 때도 원래 〈같은 $x$, 같은 $y$〉가 아니기 때문에 $y = f(x)$, $y' = g(x')$의 방식으로 표현해야 한다. 하지만 이러면 오히려 문자가 너무 많아 더 복잡하게 느껴지기 때문에 〈같은 문자 $x$, $y$〉를 쓰는 대신에 〈두 그래프에서 각각 따로 따로 의미를 생각하자.〉라는 사고방식을 관용적으로 인정하고 그렇게 사고해 왔다.
  그런데 수학에서 어떤 경우들은 그 관계를 명확하게 하기 위해서 〈다른 문자 그대로 사용해서〉 함수를 표현하기도 한다.
  표준정규분포 또한 그렇다. ($x$와 $z$의 관계를 시각적으로 명확하게 하기 위해서 그 문자 그대로 사용한다.)

- 표준정규분포도 이와 같이 $y = \dfrac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}} = \dfrac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$ $(-\infty < x < \infty)$의 복잡한 식에서

$\dfrac{x-m}{\sigma} = z$라고 치환함으로서 비교적 간단한 두 함수 $y = \dfrac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}$과 $z = \dfrac{x-m}{\sigma}$으로 분석하고자 하는 생각이 깔려있다. (치환하는 과정에서 $\sigma$가 $1$이 되어 사라지는 현상은 고등과정에서 함수적으로는 설명이 어렵다.)
이와 같이 〈치환〉의 과정을 통해 만들어진 〈$z$에 대한 함수〉는 〈주어진 정규분포〉에 상관없이 항상 같은 그래프를 가지게 된다. (각 집단마다 $m$과 $\sigma$가 다르겠지만 그에 맞게 치환할 것이기 때문에 치환된 함수의 식은

항상 $y = \dfrac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}$의 모양이고 당연히 그래프도 항상 일정한 모양이 나온다.)

그리고 여기에서는 $z$를 다시 $x$로 표현하지 않는다. 〈두 변수의 관계〉를 정확히 시각적으로 볼 필요가 있기 때문이다.

## 3) 표준정규분포의 평균, 분산, 표준편차

- 새로운 확률변수 $Z$의 평균, 분산, 표준편차를 〈성질〉을 이용하여 간단히 증명해 본다.
  연속확률분포에서 〈평균, 분산, 표준편차의 성질〉은 일단 증명하지 않고 그대로 인정하고 사용하기로 했다.

- 〈이전의 확률변수 $X$〉의 평균과 표준편차를 통해 치환한 〈새로운 확률변수 $Z$〉는 항상 평균이 $0$이고 표준편차가 $1$인 정규분포를 따른다.

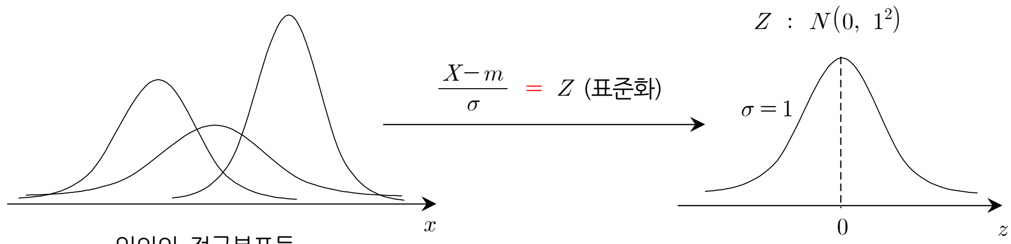
- $E(Z) = 0$, $V(Z) = 1$, $\sigma(Z) = 1$의 증명 : $E(X) = m$, $V(X) = \sigma^2$이라 하면,

① $E(Z) = E\left(\dfrac{X-m}{\sigma}\right) = E\left(\dfrac{1}{\sigma}X - \dfrac{m}{\sigma}\right) = \dfrac{1}{\sigma}E(X) - \dfrac{m}{\sigma} = \dfrac{m}{\sigma} - \dfrac{m}{\sigma} = 0$ ⇨ 결국 $E(Z) = 0$

② $V(Z) = V\left(\dfrac{X-m}{\sigma}\right) = V\left(\dfrac{1}{\sigma}X - \dfrac{m}{\sigma}\right) = \dfrac{1}{\sigma^2}V(X) = \dfrac{1}{\sigma^2}\times\sigma^2 = 1$ ⇨ 결국 $V(Z) = 1$

③ $\sigma(Z) = \sqrt{V(Z)} = 1$ ⇨ 결국 $\sigma(Z) = 1$

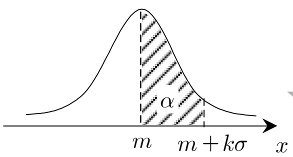
티칭 <정규분포>를 <표준화>하여 <표준정규분포>로 표현하는 것을 요약하면 아래와 같다.



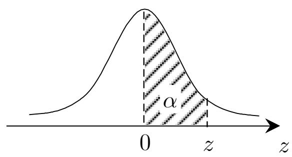

임의의 정규분포들　　　　　　　　　　　　　　　　표준정규분포

## 4) 정규분포 표 → 표준정규분포 표

| $k$ | $P(m \leq X \leq m+k\sigma)$ |
|---|---|
| 1.5 | 0.4332 |
| 2.0 | 0.4772 |
| 2.5 | 0.4938 |
| 3.0 | 0.4987 |




| $z$ | $P(0 \leq Z \leq z)$ |
|---|---|
| 1.5 | 0.4332 |
| 2.0 | 0.4772 |
| 2.5 | 0.4938 |
| 3.0 | 0.4987 |




- 결국 $z$가 바로 $m+k\sigma$에서 $k$의 의미를 갖는다는 것을 알 수 있다.
  앞으로는 <정규분포 표>의 의미를 그대로 가진 <표준정규분포 표>를 통해서 <확률 값>의 정보가 주어질 것이다.

⇨ 예를 들어 $P(m-2\sigma \leq X \leq m+2\sigma) = 0.9544$의 식을 표준화하면
(가운데에 있는 변수를 표준화함은 물론이고, 경계가 되는 $m-2\sigma$와 $m+2\sigma$도 $X$의 확률변수 중 일부이므로 모두 표준화하여 나타낸다.)

$$P\left(\frac{m-2\sigma-m}{\sigma} \leq \frac{X-m}{\sigma} \leq \frac{m+2\sigma-m}{\sigma}\right) = 0.9544 \quad \Leftrightarrow \quad P(-2 \leq Z \leq 2) = 0.9544$$

- 표준화 된 문자 $z$가 사실은 원래 문자 $x$를 $m+k\sigma$의 형태로 표현했을 때 $k$와 같은 의미를 가진다.
  이 사실을 있는 그대로 받아들인다면 <확률 값>을 통해 <확률변수>를 찾는 문제에서,
  주어진 $z$를 통해서 $x$값을 다시 찾는 <암기된 무의미한 수식적 계산>을 통해서 푼다는 느낌을
  <논리적으로 식의 의미를 파악>하면서 푼다는 느낌으로 바꿀 수 있다.
  주어진 확률 값과 $z$를 통해서 확률변수를 찾아가는 이런 문제를 앞으로는 <역표준화 문제>라고 부르기로 하자.

티칭 <정규분포를 따른다.>는 말 ⇨ 앞에서도 종종 등장했던 이 말의 의미를 생각해보자.

예를 들어 우리 반 50명의 수학점수가 $N(60,\ 10^2)$인 <정규분포>를 따른다고 해보자. 사실 잘 생각해보면 이상한 점을 발견할 수 있다. 확률변수로 취급되는 <우리 반 50명의 수학점수>는 기껏 해봐야 50개밖에 안 되고 이것은 연속적이지 않으므로 <연속확률분포>인 정규분포로 표현될 리가 없다.

그래서 <정규분포를 따른다>는 조건이 나오는 것이다. 확률변수가 50개인 <이산확률분포>이지만 <정규분포>를 통해 <확률의 근삿값>을 구할 수 있게 된다.

| $z$ | $P(0 \leq Z \leq z)$ |
|---|---|
| 1 | 0.3413 |
| 2 | 0.4772 |
| 3 | 0.4987 |

이 정규분포에서 <우리 반에서 한 명을 뽑았을 때 그 사람의 점수가 70점 이상일 확률>을 구하라고 한다면 $P(X \geq 70) = P(z \geq 1) = 0.1587$이라고 자연스럽게 구하게 될 텐데 이때의 확률 값은 당연히 근삿값이 된다.

예컨대 그래서 <70점이 넘는 사람의 수>를 물어본다면 <0.1587을 만족하는 확률 값>이 곧 <70점을 넘는 사람의 비율 값>이므로 $50(\text{명}) \times 0.1587 = 7.935$로 계산되어 나온 7.935 역시 8명의 근삿값으로 도출된 것이다. (수능 정도의 시험에서는 곱했을 때 소수 점이 나오지 않도록 숫자를 잘 맞춰서 줄 것이다.)

**125** 이해를 위한 예제 2006. 11. 나형(73%). 9번. 3점

어느 세차장에서 승용차 한 대를 세차하는 데 걸리는 세차 시간은 평균 30분, 표준편차 2분인 정규분포를 따른다고 한다. 한 대의 승용차를 이 세차장에서 세차할 때, 세차 시간이 33분 이상일 확률을 다음 표준정규분포 표를 이용하여 구한 것은?

| $z$ | $\mathrm{P}(0 \le Z \le z)$ |
|---|---|
| 0.5 | 0.1915 |
| 1.0 | 0.3413 |
| 1.5 | 0.4332 |
| 2.0 | 0.4772 |

① 0.0228 ② 0.0668 ③ 0.1587
④ 0.2708 ⑤ 0.3085

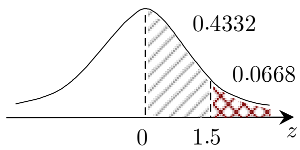
표준정규분포와 확률

조건 세차하는 데 걸리는 세차 시간은 평균 30분, 표준편차 2분인 정규분포를 따른다고 한다.

- 세차시간을 $X$라 하면 $X : N(30,\ 2^2)$

구할 것 세차 시간이 33분 이상일 확률

- 즉, $P(X \ge 33) = P\left(\frac{X-30}{2} \ge \frac{33-30}{2}\right) = P(Z \ge 1.5) = 0.0668$




정답은 ②

코칭 물론 표준화를 하지 않고 <$z$와 $k$의 관계>를 통해서도 얼마든지 문제를 풀 수 있으며 이런 방식으로 문제를 푸는 것은 좋은 방법이다. 표준화를 그냥 암기한 것이 아니라 <정확한 의미>를 알고 있다는 뜻이기 때문이다.
하지만 <표준화> 역시 중요한 사고방식이기 때문에 해설은 <표준화>를 통해서 하도록 하겠다.

**126** 이해를 위한 예제 2009. 11. 가형(87%). 나형(70%) 9번. 3점

어느 공장에서 생산되는 병의 내압강도는 정규분포 $N(m,\ \sigma^2)$ 을 따르고, 내압강도가 40 보다 작은 병은 불량품으로 분류한다. 이 공장의 공정능력을 평가하는 공정능력지수 $G$는 $G = \frac{m-40}{3\sigma}$ 으로 계산한다. $G = 0.8$ 일 때, 임의로 추출한 한 개의 병이 불량품일 확률을 다음 표준정규분포 표를 이용하여 구한 것은?

| $z$ | $\mathrm{P}(0 \le Z \le z)$ |
|---|---|
| 2.2 | 0.4861 |
| 2.3 | 0.4893 |
| 2.4 | 0.4918 |
| 2.5 | 0.4938 |

① 0.0139 ② 0.0107 ③ 0.0082 ④ 0.0062 ⑤ 0.0038

표준정규분포와 확률

조건 내압강도는 정규분포 $N(m,\ \sigma^2)$ 을 따르고, 내압강도가 40보다 작은 병은 불량품으로 분류한다.

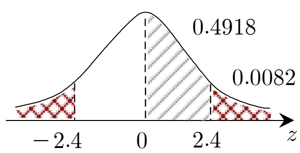
조건 공정능력지수 $G$는 $G = \frac{m-40}{3\sigma}$ 으로 계산한다. $G = 0.8$ 일 때,

- $0.8 = \frac{m-40}{3\sigma} \Leftrightarrow m = 40 + 2.4\sigma$ 이다.

구할 것 임의로 추출한 한 개의 병이 불량품일 확률

- $P(\text{불량품}) = P(X < 40) = P\left(\frac{X-m}{\sigma} < \frac{40-m}{\sigma}\right)$

$= P\left(Z < \frac{40-(40+2.4\sigma)}{\sigma}\right) = P(Z < -2.4) = 0.0082$

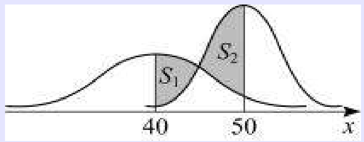



정답은 ③

코칭 표준화를 하지 않고 계산하면 헷갈릴 소지가 있도록 문제를 겨냥하고 내는 경우도 많다.
이 문제도 $P(X < 40)$에서 40을 <평균 + 표준편차의 몇 배>로 표현할 수 있고, <몇 배>로 $z$를 대체할 수 있지만 그것을 암산하는 과정이 헷갈릴 수 있도록 문제가 디자인된 것이다.

**127** 이해를 위한 예제 2009. 3. 가형(74%). 27번. 3점

그림은 정규분포 $\mathrm{N}(40,\ 10^2)$, $\mathrm{N}(50,\ 5^2)$을 따르는 두 확률변수 $X$, $Y$의 정규분포곡선을 나타낸 것이다. 그림과 같이 $40 \le x \le 50$인 범위에서 두 곡선과 직선 $x=40$으로 둘러싸인 부분의 넓이를 $S_1$, 두 곡선과 직선 $x=50$으로 둘러싸인 부분의 넓이를 $S_2$라 할 때, $S_2 - S_1$의 값을 다음 표준정규분포 표를 이용하여 구한 것은?




| $z$ | $\mathrm{P}(0 \le Z \le z)$ |
|---|---|
| 1 | 0.3413 |
| 2 | 0.4772 |
| 3 | 0.4987 |

① 0.1248 ② 0.1359 ③ 0.1575 ④ 0.1684 ⑤ 0.1839

정규분포와 확률, 표준정규분포와 확률

조건 정규분포 $\mathrm{N}(40,\ 10^2)$, $\mathrm{N}(50,\ 5^2)$을 따르는 두 확률변수 $X$, $Y$의 정규분포곡선을 나타낸 것

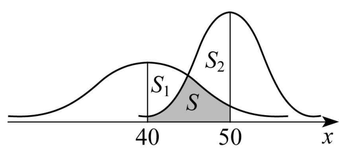
- $X$의 평균보가 $Y$의 평균이 크므로 주어진 그림에서
  왼쪽의 그림이 $X$의 정규분포이고, 오른쪽의 그림의 $Y$의 정규분포이다.

조건 그림에서 어두운 부분의 넓이를 $S$라 하면

$S$는 $P(40 \le Y \le 50)$와 $P(40 \le X \le 50)$에 동시에 포함된 영역이다.

그러므로 $P(40 \le Y \le 50) - P(40 \le X \le 50) = S_2 - S_1$

(어두운 부분은 상쇄되어 사라지므로 결국 $S_2 - S_1$의 의미가 된다.)




구할 것 $S_2 - S_1$의 값

⇨ $S_2 - S_1 = P(40 \le Y \le 50) - P(40 \le X \le 50)$

$= P\left(0 \le Z \le \dfrac{60-50}{5}\right) - P\left(0 \le Z \le \dfrac{50-40}{10}\right)$

$= P(0 \le Z \le 2) - P(0 \le Z \le 1) = 0.4772 - 0.3413 = 0.1359$

정답은 ②

코칭 <정적분 값>이 곧 <확률 값>이라는 사실을 다시 한 번 명심하자.

**128** 이해를 위한 예제  2008. 10. 나형(83%). 8번. 3점

어떤 특산품 과일을 재배하는 과수원에서는 해마다 수확량의 일부를 해외로 수출한다. 이 과수원에서 올해 수확한 과일 30000개의 무게는 평균 400 g, 표준편차 20 g인 정규분포를 따른다고 한다. 이 30000개의 과일 중 무게가 400 g 이상이고 440 g 이하인 과일을 선별하여 수출하였다. 이 과수원에서 올해 수출한 과일의 개수를 아래 표준정규분포 표를 이용하여 구한 것은?

| $z$ | $\mathrm{P}(0 \leq Z \leq z)$ |
|---|---|
| 1.0 | 0.34 |
| 1.5 | 0.43 |
| 2.0 | 0.48 |
| 2.5 | 0.49 |

① 10200 ② 11600 ③ 12900 ④ 14400 ⑤ 14700

확률과 비율 : 표준정규분포와 확률 → 확률 값을 통한 개수 구하기

조건 올해 수확한 과일 30000개의 무게는 평균 400 g, 표준편차 20 g인 정규분포를 따른다고 한다.

- 무게를 확률변수 $X$라 하면 $X$는 이산확률변수이지만 그 분포가 근사적으로 $N(400,\ 20^2)$을 따른다는 뜻이다.

조건 이 30000개의 과일 중 무게가 400 g 이상이고 440 g 이하인 과일을 선별하여 수출

구할 것 수출한 과일의 개수

- 〈확률 값〉이 곧 〈비율 값〉이므로 〈확률 값〉을 구하여 〈전체 대상의 개수 : 30000개〉에 곱한다.

⇨ $P(400 \leq X \leq 440) = P(0 \leq Z \leq 2) = 0.48$이므로

수출한 과일은 $30000 \times 0.48 = 14400$(개)이다.

정답은 ④

**129** 이해를 위한 예제 2007. 11. 가형(88%), 나형(66%). 13번. 4점

어느 회사의 전체 신입 사원 $1000$명을 대상으로 신체검사를 한 결과, 키는 평균 $m$, 표준편차 $10$인 정규분포를 따른다고 한다. 전체 신입 사원 중에서 키가 $177$ 이상인 사원이 $242$명이었다. 전체 신입 사원 중에서 임의로 선택한 한 명의 키가 $180$이상일 확률을 다음 표준정규분포 표를 이용하여 구한 것은? (단, 키의 단위는 cm이다.)

① $0.1587$ ② $0.1841$ ③ $0.2119$ ④ $0.2267$ ⑤ $0.2420$

| $z$ | $P(0 \leq Z \leq z)$ |
|---|---|
| 0.7 | 0.2580 |
| 0.8 | 0.2881 |
| 0.9 | 0.3159 |
| 1.0 | 0.3413 |

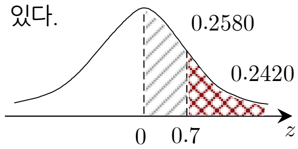
확률과 비율 : 비율을 통해서 주어진 확률 값

조건 신입 사원 $1000$명을 대상으로 신체검사를 한 결과, 키는 평균 $m$, 표준편차 $10$인 정규분포를 따른다고 한다.

- 신입사원의 키를 확률변수 $X$라 하면 $X$는 이산확률변수이지만 그 분포가 근사적으로 $N(m,\ 10^2)$을 따른다는 뜻이다.

조건 전체 신입 사원 중에서 키가 $177$이상인 사원이 $242$명이었다.

- $P(X \geq 177) = \dfrac{242}{1000} = 0.242$ (비율 값을 통해서 확률 값을 준 것이다.)
- 이런 방식으로 확률 값을 주는 것은 오직 <이산확률분포>가 <정규분포>를 따른다는 가정이 있을 때 가능하다.
- 이미 구한 관계식 $P(X \geq 177) = 0.242$을 표준화함으로서 $N(m,\ 10^2)$에서 $m$을 구할 수 있다.

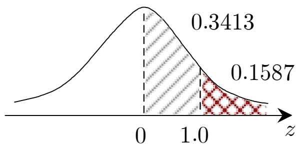



즉, $P\left(\dfrac{X-m}{10} \geq \dfrac{177-m}{10}\right) = 0.242 \Leftrightarrow P\left(Z \geq \dfrac{177-m}{10}\right) = 0.242$

위의 표를 통해 그린 오른쪽 그래프를 참조하면 결국 $\dfrac{177-m}{10} = 0.7$이므로 $m = 170$

구할 것 임의로 선택한 한 명의 키가 $180$이상일 확률

- $N(170,\ 10^2)$에서 $P(X \geq 180) = P\left(\dfrac{X-170}{10} \geq \dfrac{180-170}{10}\right) = P(Z \geq 1) = 0.1587$




정답은 ①

코칭 표준화를 하면 $P(X \geq 177) = 0.242$을 통해 $N(m,\ 10^2)$의 $m$을 간단히 도출할 수 있다.
이렇게 <표준화>를 하지 않으면 헷갈릴 수 있도록 문제를 디자인한다.

**130** 이해를 위한 예제 2006. 9. 가형(87%), 나형(76%). 10번. 3점

어느 농장의 생후 7개월 된 돼지 200마리의 무게는 평균 110 kg, 표준편차10 kg인 정규분포를 따른다고 한다. 이 200마리의 돼지 중 무거운 것부터 차례로 3마리를 뽑아 우량 돼지 선발대회에 보내려고 한다. 우량 돼지 선발대회에 보낼 돼지의 최소 무게를 아래 표준정규분포 표를 이용하여 구한 것은?

| $z$ | $P(0 \leq Z \leq z)$ |
|---|---|
| 2.12 | 0.4830 |
| 2.17 | 0.4850 |
| 2.29 | 0.4890 |

① 121.6 kg ② 126.7 kg ③ 130.7 kg ④ 131.7 kg ⑤ 132.9 kg

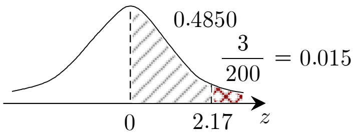
확률과 비율 : 비율을 통해서 주어진 확률 값 + 역 표준화 문제 (경계 값 구하기)

조건 돼지 200마리의 무게는 평균 110 kg, 표준편차 10 kg인 정규분포를 따른다고 한다.

- 돼지의 무게를 확률변수 $X$라 하면 $X$는 이산확률변수이지만 그 분포가 근사적으로 $N(110, 10^2)$인 정규분포를 따른다.

조건 무거운 것부터 차례로 3마리를 뽑아 우량 돼지 선발대회

- <비율 값>이 곧 <확률 값>이므로 비율적으로 상위 $\frac{3}{200}$에 들어가는 무게를 구한다.




$P(\text{대회출전}) = P(X \geq k) = \frac{3}{200} = 0.015$

이때의 $k$값이 대회에 출전할 수 있는 최소무게이다. 오른쪽 그림에서
우량 돼지 선발대회에 보낼 돼지의 최소 무게는 $z = 2.17$인 순간의 $X$값이므로
$X = 110 + 2.17 \cdot 10 = 131.7$

정답은 ④

티칭 위의 해설은 $z$의 의미를 생각해서 $X$값을 바로 구한 것이다. <표준화>를 통해서 구하는 것은

$$P(X \geq k) = \frac{3}{200} \Leftrightarrow P\left(Z \geq \frac{k-110}{10}\right) = \frac{3}{200} \text{ 즉, } \frac{X-110}{10} = 2.17 \Leftrightarrow X = 131.7$$

# 4. 서로 다른 확률 변수로 표현된 확률 관계식

- 표준화를 하지 않으면 굉장히 헷갈릴 수 있는 계산문제이다.

⇨ 예를 들어 $P(X \geq \alpha) = P(Y \geq \beta)$와 같이 <서로 다른 확률 변수>에 대한 <확률 관계식> 비교하는 문제는 변수를 통일시키기 위해서 무조건 표준화를 시키는 것이 현명하다.

- 이것 말고도 문제에서 <함수를 정의>할 때 표준화를 해야만 <정의된 함수>가 변수로 표현되는 문제들도 <표준화>를 겨냥하고 낸 문제라고 할 수 있다.

**131** 이해를 위한 예제 2011. 9. 나형(66%). 16번. 4점

어느 공장에서 생산되는 제품 A의 무게는 정규분포 $N(m, 1)$을 따르고, 제품 B의 무게는 $N(2m, 4)$를 따른다. 이 공장에서 생산된 제품 A와 제품 B에서 임의로 제품을 1개씩 선택할 때, 선택된 제품 A의 무게가 $k$이상일 확률과 선택된 제품 B의 무게가 $k$이하일 확률이 같다. $\frac{k}{m}$의 값은?

① $\frac{11}{9}$ ② $\frac{5}{4}$ ③ $\frac{23}{18}$ ④ $\frac{47}{36}$ ⑤ $\frac{4}{3}$

서로 다른 확률변수의 확률관계식 : 변수통일 표준화

조건 제품 A의 무게는 정규분포 $N(m, 1)$을 따르고, 제품 B의 무게는 $N(2m, 4)$를 따른다.

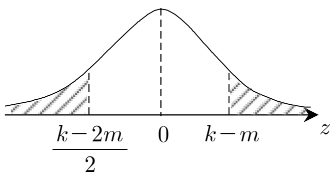
- $A$제품의 무게를 확률변수 $X$라 하고, $B$제품의 무게를 확률변수 $Y$라 하자.

조건 제품 A와 제품 B에서 임의로 제품을 1개씩 선택할 때, 선택된 제품 A의 무게가 $k$이상일 확률과 선택된 제품 B의 무게가 $k$이하일 확률이 같다.

- 이것을 기호로 나타내면 $P(X \geq k) = P(Y \leq k)$이다. 이와 같이 서로 다른 확률변수에 대한 확률 관계식이 나오면 표준화를 통해서 식을 정리한다.

⇨ $P(X \geq k) = P(Y \leq k) \Leftrightarrow P\left(\frac{Y-2m}{2} \leq \frac{k-2m}{2}\right) = P\left(\frac{Y-2m}{2} \leq \frac{k-2m}{2}\right)$

$\Leftrightarrow P(Z \geq k-m) = P\left(Z \leq \frac{k-2m}{2}\right)$




표준정규분포의 대칭성에 따라서 $k-m = -\left(\frac{k-2m}{2}\right) \Leftrightarrow 4m = 3k \Leftrightarrow \frac{k}{m} = \frac{4}{3}$

정답은 ⑤

**132** 이해를 위한 예제 2009. 9. 나형(56%). 29번. 4점

양의 실수 전체의 집합을 정의역으로 하는 함수 $H(t)$는 평균 20, 표준편차 $t$인 정규분포를 따르는 확률변수 $X$에 대하여 $H(t)=P(X\le 15)$이다. 옳은 것만을 <보기>에서 있는 대로 고른 것은?
(단, 표준정규분포를 따르는 $Z$에 대하여 $P(0\le Z\le 1)=0.3413$, $P(0\le Z\le 2)=0.4772$이다.)

<보 기>

ㄱ. $H(2.5)=P(Z\ge 2)$ ㄴ. $H(2)<H(2.5)$ ㄷ. $H(5)<5H(2)$

① ㄱ ② ㄷ ③ ㄱ, ㄴ ④ ㄴ, ㄷ ⑤ ㄱ, ㄴ, ㄷ

문제에서 정의된 함수 : 변수생성 표준화

조건 양의 실수 전체의 집합을 정의역으로 하는 함수 $H(t)$는 평균 20, 표준편차 $t$인 정규분포를 따르는 확률변수 $X$에 대하여 $H(t)=P(X\le 15)$

- $X$ : $N(20,\ t^2)$이고, $H(t)\ =\ P(X\le 15)$으로 정의되어 있다. 주어진 식 $P(X\le 15)$에 <변수 $t$>가 존재하지 않으므로 <표준화>를 통해서 변수 $t$가 존재하는 식으로 바꾸어 줄 수 있다. ⇨ $H(t)\ =\ P\left(Z\le\dfrac{15-20}{t}\right)$

ㄱ. 좌변의 식 $H(2.5)\ =\ P\left(Z\le\dfrac{15-20}{2.5}\right)\ =\ P(Z\le -2)$이므로 우변의 식 $P(Z\ge 2)$과 같다. (참)

ㄴ. 우변의 식 $H(2)\ =\ P\left(Z\le\dfrac{15-20}{2}\right)\ =\ P(Z\le -2.5)$이고, 좌변의 식 $H(2.5)\ =\ P\left(Z\le\dfrac{15-20}{2.5}\right)\ =\ P(Z\le -2)$ 이다. 즉, $P(Z\le -2.5)<P(Z\le -2)$이므로 $H(2)<H(2.5)$은 (참)

ㄷ. 우변의 식 $H(5)\ =\ P\left(Z\le\dfrac{15-20}{5}\right)\ =\ P(z\le -1)\ =\ 0.1587$이고

좌변의 식 $5H(2)\ =\ 5P\left(Z\le\dfrac{15-20}{2}\right)\ =\ 5P(Z\le -2.5)\ <\ 5P(Z\le -2)\ =\ 0.1140$이므로

$H(5)<5H(2)$는 (거짓) - $0.1140$보다 작은 값이 $0.1587$보다 크다는 것은 당연히 잘못된 것이다.

정답은 ③

## #3 이항분포의 정규분포 화

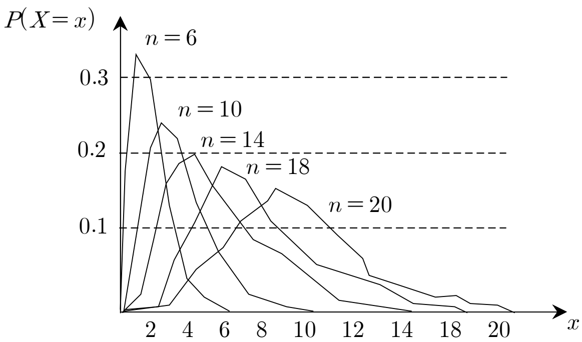
- 사실 정규분포는 이항분포를 통해서 유도된 것으로 그 과정은 상당히 복잡하다.
  (네이버에 정규분포를 검색하면 <네이버 캐스트>가 나오는데 반드시 한 번 읽어보는 것을 추천한다.)

- <이항분포>의 일부의 확률을 구하는 것이 <불가능함>을 느끼고, 그래서 자연스럽게 <이항분포의 정규분포 화>를 통해 확률의 근삿값을 구하는 과정이다.

- 꼭 이항분포가 아니더라도 거의 모든 이산확률분포는 근사적으로 <정규분포>를 따른다.
  이것은 라플라스가 증명한 내용으로 그 과정에 대해서는 알 필요가 없다. 우리가 앞에서 문제의 조건으로 나왔던 정규분포를 따르는 <이산확률분포>들이 여기에 해당한다.

# 1. 이항분포의 정규분포 화

### 1) 라플라스의 정리




18세기 라플라스는 이항분포의 계산을 $n$을 점차 크게 늘려가며 계산해서 그래프 상에 표시해 보았다. 놀랍게도 이항분포로 계산한 확률은 $n$을 늘려 가면 늘려갈수록 <종 모양>에 가까워졌고, 가운데의 값이 평균과 비슷해진다는 결론을 얻었다. 이것이 라플라스의 정리이다. (사실 가우스-라플라스정리, 또는 드무아브르-라플라스의 정리라고도 한다.)
<라플라스의 정리>의 의의는 <이항분포>를 <정규분포>로 바꾸어 계산해도 수치가 크게 차이나지 않는다는 것을 밝힌 것에 있다.

### 2) $X : B(n, p) \quad \rightarrow \quad X : N\left(np, (\sqrt{npq})^2\right)$인 경우 확률 구하기

- $B\left(180,\ \frac{1}{6}\right)$에서 $P(X \geq 35)$을 구하라고 한다면 계산이 불가능하다고 여기는 게 당연하다.

  ${}_{180}C_{35}\left(\frac{1}{6}\right)^{35}\left(\frac{5}{6}\right)^{145} + \cdots + {}_{180}C_{180}\left(\frac{1}{6}\right)^{180}\left(\frac{5}{6}\right)^{0}$인데, 더하기는커녕 하나의 항만 계산하는 것도 불가능하다.

- 결국 $n$이 충분히 큰 $B\left(180,\ \frac{1}{6}\right)$은 근사적으로 정규분포를 따르므로

  자명한 식 $m = 180 \cdot \frac{1}{6} = 30,\ \sigma^2 = 180 \cdot \frac{1}{6} \cdot \frac{5}{6} = 25$를 통해서 구한 <평균과 분산>을

  정규분포의 <평균과 분산>으로 사용하여 (라플라스의 정리에 의하여) 확률변수 $X$는 근사적으로 $N(30,\ 5^2)$을 따른다.
  그러므로 위의 $P(X \geq 35)$은 표준화시켜서 표준정규분포 표를 이용하면 바로 구할 수 있다.

### 3) $X : B(n, p) \quad \rightarrow \quad X : \cancel{N\left(np, (\sqrt{npq})^2\right)}$~~인~~ 경우 확률 구하기

$n$이 충분히 크지 않은 상태에서 확률을 구하라고 한다면 (10 미만으로 주어질 것)
일전에 학습했듯이 표를 떠올리면서 <독립시행의 확률>로 확률 값을 계산하면 된다.

코칭 모든 $n$에 대하여 증명된 자명한 식

$X : B(n,\ p)$ (단, $q=1-p$)이라고 하면 ① $E(X) = np$ ② $V(X) = npq$ ③ $\sigma(X) = \sqrt{npq}$

〈이항분포의 평균, 분산, 표준편차〉를 나타내는 위의 식은 〈자명한 식〉이며 단지 $n$이 충분히 커야 여기에서 구한 평균과 분산을 〈정규분포 화〉에 활용할 수 있다는 뜻이다.

코칭 이제 간단한 정규분포의 확률을 구하는 과정에서 〈정규분포의 그래프〉와 〈표준화에 대한 간단한 계산과정〉은 생략하기로 한다.

## 2. 이항분포의 확률표현 : $\sum$를 통한 표현에 익숙해지자

예시 1 $\displaystyle\sum_{r=35}^{40} {}_{180}C_r\left(\frac{1}{6}\right)^r\left(\frac{5}{6}\right)^{180-r}$ : 일단 ${}_{180}C_r\left(\frac{1}{6}\right)^r\left(\frac{5}{6}\right)^{180-r}$을 통해서 $B\left(180,\ \frac{1}{6}\right)$라는 사실을 안다.

$r$이 확률변수 $X$를 대신하므로 우리가 구하는 확률은 $P(35 \le X \le 40)$이고,

정규분포 $N(30,\ 5^2)$을 활용하여 푼다.

예시 2 $\displaystyle\sum_{r=1}^{7} {}_7C_r\left(\frac{1}{3}\right)^r\left(\frac{2}{3}\right)^{7-r}$ : 일단 ${}_7C_r\left(\frac{1}{3}\right)^r\left(\frac{2}{3}\right)^{7-r}$을 통해서 $B\left(7,\ \frac{1}{3}\right)$라는 사실을 안다. 이때는 $n$이 작은 수이므로

그대로 계산한다. 확률의 총합은 1 이므로 $\displaystyle\sum_{r=1}^{7} {}_7C_r\left(\frac{1}{3}\right)^r\left(\frac{2}{3}\right)^{7-r} = 1 - {}_7C_0\left(\frac{1}{3}\right)^0\left(\frac{2}{3}\right)^7$

예시 3 $\displaystyle\sum_{r=0}^{180} r\,{}_{180}C_r\left(\frac{1}{6}\right)^r\left(\frac{5}{6}\right)^{180-r}$ 와 $\displaystyle\sum_{r=0}^{180} r^2\,{}_{180}C_r\left(\frac{1}{6}\right)^r\left(\frac{5}{6}\right)^{180-r}$

: $B\left(180,\ \frac{1}{6}\right)$에서 $E(X)$와 $E(X^2)$을 구하는 것으로 〈공식〉을 이용할 수 있다.

예시 4 $\displaystyle\sum_{r=0}^{10} r\,{}_{10}C_r\left(\frac{1}{6}\right)^r\left(\frac{5}{6}\right)^{10-r}$ 와 $\displaystyle\sum_{r=0}^{10} r^2\,{}_{10}C_r\left(\frac{1}{6}\right)^r\left(\frac{5}{6}\right)^{10-r}$

: $B\left(10,\ \frac{1}{6}\right)$에서 $E(X)$와 $E(X^2)$을 구하는 것으로 역시 〈공식〉을 이용할 수 있다.

$n$이 작아도 $E(X)$와 $E(X^2)$의 〈이항분포의 평균, 분산공식〉은 얼마든지 쓸 수 있는 자명한 식임을 다 시 한번 상기한다. $n$의 크기에 따라 (〈그냥 구할지〉 〈정규분포 화〉해서 구할지) 고민해야 하는 것은 〈확률 값〉이다.

**133** 이해를 위한 예제 2009. 사관학교. 문·이과. 11번. 3점

서류전형 후 필기시험을 실시하는 어느 시험에서 720 명이 서류전형에 합격하였다. 서류전형 합격자는 필기시험에서 A, B, C, D 4 과목 중 2 과목을 반드시 선택해야 하고, 각 과목을 선택할 확률은 모두 같다고 한다. 4 과목 중 A, B 를 선택한 서류전형의 합격자의 수가 110 명 이상 145 명 이하일 확률을 오른쪽 표준정규분포표를 이용하여 구한 것은?

| $z$ | $P(0 \leq Z \leq z)$ |
|---|---|
| 1.0 | 0.3413 |
| 1.5 | 0.4332 |
| 2.0 | 0.4772 |
| 2.5 | 0.4938 |

① 0.0166 ② 0.1359 ③ 0.1525
④ 0.8351 ⑤ 0.9104

조건 720 명이 서류전형에 합격 + A, B, C, D 4 과목 중 2 과목을 반드시 선택해야 하고, 각 과목을 선택할 확률은 모두 같다고 한다.

구할 것 A, B 를 선택한 서류전형의 합격자의 수가 110 명 이상 145 명 이하일 확률

- 서류전형에 합격한 사람이 필기시험에서 A, B, C, D 4 과목 중 2 과목을 선택하는 방법의 수는 ${}_4C_2 = 6$가지

즉, 임의로 고른 한 사람이 A, B 를 선택할 확률은 $\frac{1}{6}$

그러므로 720명 중 $A$, $B$를 선택한 사람 수를 확률변수 $X$라 하면 $X$는 이항분포 $B\left(720,\ \frac{1}{6}\right)$를 따른다.

⇨ $B\left(720,\ \frac{1}{6}\right)$은 근사적으로 정규분포 $N(120,\ 10^2)$을 따르므로

$P(110 \leq X \leq 145) = P(-1 \leq Z \leq 2.5) = 0.3413 + 0.4938 = 0.8351$

정답은 ④

**134** 이해를 위한 예제 2012. 10. 가형(78%). 11번. 3점

어느 과수원에서 수확한 사과의 무게는 평균 400$g$, 표준편차 50$g$인 정규분포를 따른다고 한다. 이 사과 중 무게가 442$g$ 이상인 것을 1등급 상품으로 정한다. 이 과수원에서 수확한 사과 중 100 개를 임의로 선택할 때, 1등급 상품이 24 개 이상일 확률을 다음 표준정규분포 표를 이용하여 구한 것은?

| $z$ | $P(0 \leq Z \leq z)$ |
|---|---|
| 0.64 | 0.24 |
| 0.84 | 0.30 |
| 1.00 | 0.34 |
| 1.28 | 0.40 |

① 0.10 ② 0.16 ③ 0.20 ④ 0.26 ⑤ 0.34

조건 사과의 무게는 평균 400$g$, 표준편차 50$g$인 정규분포를 따른다고 한다.

- 사과의 무게를 $X$라 하면 $X$ : $N(400, 50^2)$

조건 무게가 442$g$ 이상인 것을 1등급 상품

- $P(\text{일등급}) = P(X \geq 442) = P(Z \geq 0.84) = 0.2$

구할 것 수확한 사과 중 100 개를 임의로 선택할 때, 1등급 상품이 24 개 이상일 확률

- 확률변수가 〈개수〉로 바뀌었음을 인지해야 한다.
- 사과 100개 중 1등급 상품의 개수를 확률변수 $Y$라 하면 $Y$ : $B(100,\ 0.2)$를 따르고 근사적으로 $N(20, 4^2)$을 따른다.

⇨ $P(Y \geq 24) = P\left(Z \geq \frac{24-20}{4}\right) = P(Z \geq 1) = 0.16$

정답은 ②

**135** 이해를 위한 예제

확률변수 $X$에 대하여 $P(X=r) = {}_{100}C_r\left(\frac{1}{5}\right)^r\left(\frac{4}{5}\right)^{100-r}$ $(r = 0,\ 1,\ 2,\ \cdots,\ 100)$ 일 때, $P(16 \le X \le 24)$의 값은?

(단, $P(0 \le Z \le 1) = 0.34$)

① 0.16 ② 0.34 ③ 0.51 ④ 0.66 ⑤ 0.68

조건 $P(X=r) = {}_{100}C_r\left(\frac{1}{5}\right)^r\left(\frac{4}{5}\right)^{100-r}$ $(r = 0,\ 1,\ 2,\ \cdots,\ 100)$

- 주어진 확률질량함수를 통해서 $X$는 이항분포 $B\left(100,\ \frac{1}{5}\right)$을 따르고 근사적으로 $N(20, 4^2)$를 따른다.

⇨ $P(16 \le X \le 24) = P(-1 \le Z \le 1) = 0.68$

답. ⑤

**136** 이해를 위한 예제

표준정규분포 표를 이용하여

$${}_{100}C_{100}\left(\frac{9}{10}\right)^{100} + {}_{100}C_{99}\left(\frac{9}{10}\right)^{99}\left(\frac{1}{10}\right) + {}_{100}C_{98}\left(\frac{9}{10}\right)^{98}\left(\frac{1}{10}\right)^2 + {}_{100}C_{97}\left(\frac{9}{10}\right)^{97}\left(\frac{1}{10}\right)^3 + {}_{100}C_{96}\left(\frac{9}{10}\right)^{96}\left(\frac{1}{10}\right)^4$$

의 근삿값을 구하여라.

| $z$ | $P(0 \le Z \le z)$ |
|---|---|
| 1.5 | 0.4332 |
| 2.0 | 0.4772 |
| 2.5 | 0.4938 |

조건 $${}_{100}C_{100}\left(\frac{9}{10}\right)^{100} + {}_{100}C_{99}\left(\frac{9}{10}\right)^{99}\left(\frac{1}{10}\right) + {}_{100}C_{98}\left(\frac{9}{10}\right)^{98}\left(\frac{1}{10}\right)^2 + {}_{100}C_{97}\left(\frac{9}{10}\right)^{97}\left(\frac{1}{10}\right)^3 + {}_{100}C_{96}\left(\frac{9}{10}\right)^{96}\left(\frac{1}{10}\right)^4$$

- 주어진 식을 통해서 이항분포 $B\left(100,\ \frac{9}{10}\right)$을 따르는 확률변수를 $X$라 할 때 $P(96 \le X \le 100)$을 구하는 것임을 알 수 있다. 이 확률변수 $X$는 근사적으로 $N(90,\ 3^2)$을 따르고, $P(96 \le X \le 100) = P(2 \le Z \le 3.5)$이다.
그런데 위의 표를 참조하면 $z = 3.5$일 때 값은 주어지지 않았다.

- 정규분포는 평균 주위에 대부분의 확률 값이 몰려있기 때문에 $z = 3$이후의 값은 모두 $\infty$라고 생각해도 좋다.
예를 들어 $P(0 \le Z \le 3)$이나 $P(0 \le Z \le 4)$나 $P(0 \le Z \le 5)$나 모두 비슷한 값을 가진다.
물론 대소비교를 한다면 $P(0 \le Z \le 3) < P(0 \le Z \le 4) < P(0 \le Z \le 5)$이겠으나 <값을 구하는 문제>에서는 $z$값이 주어지지 않은 영역은 모두 $\infty$로 처리해도 좋다.
즉, $P(2 \le Z \le 3.5) = P(2 \le Z) = 0.0228$

정답은 0.0228

# 예제를 통해 개념의 외연을 넓혀라

촌철살인 경우의 수 | PART 5 연속확률분포와 정규분포

개념의 외연 #1. 연속확률분포와 적분의 관계

<개념의 외연>이라는 부분은 현재 배우고 있는 경우의 수의 국한된 내용만 다루는 것은 아니다.
현재 단원에 해당하는 문제를 풀기 위해서 필요한 수학적 개념이나 유형에 필요한 사고방식을 두루 다룬다.

수학에는 순서가 없다. 하지만 배움에는 순서가 있다.

그래서 순서가 없는 수학에 순서를 정한 것이 교육과정이다. 어떤 내용은 너무 어렵거나 복합적인 생각들을 담고 있어서 여러 학년에 걸쳐 배운다. 이런 경우 보통은 문제를 통해서 익힌다는 느낌이 강한데 이런 개념들을 집약적으로 정리해 주는 부분이 될 것이다.

## #1 연속확률분포와 적분의 관계

교과과정 외 - 넘어가도 좋다.

### 1. 연속확률분포의 평균, 분산, 표준편차

- 확률밀도함수 $y=f(x)$ 구간 $[a,b]$에서 정의 될 때, $\left(\text{단, } \int_a^b f(x)dx=1\right)$

① $E(X) = \int_a^b xf(x)dx$ ② $V(X) = \int_a^b x^2f(x)dx - \left(\int_a^b xf(x)dx\right)^2$ ③ $\sigma(X) = \sqrt{V(X)}$

### 2. 연속확률분포의 평균, 분산, 표준편차의 성질 증명

① $E(aX+b) = aE(X)+b$ ② $V(aX+b) = a^2V(X)$ ③ $\sigma(aX+b) = |a|\sigma(X)$

증명 ① $E(aX+b) = \int_a^b (ax+b)f(x)dx = \int_a^b \{axf(x)+bf(x)\}dx = \int_a^b axf(x)dx + \int_a^b bf(x)dx$

$$= a\int_a^b xf(x)dx + b\int_a^b f(x)dx = aE(X)+b$$

② $V(aX+b) = \int_a^b (ax+b)^2f(x)dx - \{aE(X)+b\}^2$

$$= \int_a^b (a^2x^2+2abx+b^2)f(x)dx - \{aE(X)+b\}^2$$

$$= \int_a^b \{a^2x^2f(x)+2abxf(x)+b^2f(x)\}dx - \{aE(X)+b\}^2$$

$$= a^2\int_a^b x^2f(x)dx + 2ab\int_a^b xf(x)dx + b^2\int_a^b f(x)dx - \{aE(X)+b\}^2$$

$$= a^2E(X^2) + 2abE(X) + b^2 - \{a^2E(X)^2+2abE(X)+b^2\}$$

$$= a^2\{E(X^2)-E(X)^2\} = a^2V(X)$$

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

**137** 이해를 위한 예제 2011. 9. 가형(75%). 22번. 3점 (교육과정 삭제 - 어렵진 않지만 넘어가도 좋다.)

연속확률변수 $X$의 확률밀도함수가 $f(x) = ax + b\ (0 \le x \le 1)$이다. $E(X) = \dfrac{7}{12}$일 때, 두 상수 $a, b$에 대하여 $80ab$의 값을 구하시오.

조건 연속확률변수 $X$의 확률밀도함수가 $f(x) = ax + b\ (0 \le x \le 1)$

- 확률밀도 함수는 주어진 구간에서 정적분 값이 1이므로

$$\int_0^1 (ax+b)dx = 1 \Leftrightarrow \left[\frac{a}{2}x^2 + bx\right]_0^1 = 1 \Leftrightarrow \frac{a}{2} + b = 1 \Leftrightarrow a + 2b = 2$$

조건 $E(X) = \dfrac{7}{12}$일 때,

$$- \int_0^1 x(ax+b)dx = \frac{7}{12} \Leftrightarrow \left[\frac{a}{3}x^3 + \frac{b}{2}x^2\right]_0^1 = \frac{7}{12} \Leftrightarrow \frac{a}{3} + \frac{b}{2} = \frac{7}{12} \Leftrightarrow 4a + 6b = 7$$

이 두 일차방정식 $a+2b = 2$과 $4a+6b = 7$을 연립하면 $a = 1$, $b = \dfrac{1}{2}$이다. 즉, $80ab = 40$

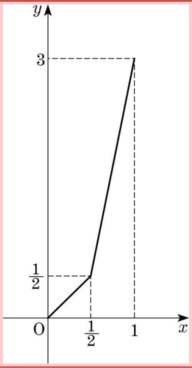
정답은 $40$

**138** 이해를 위한 예제 2011. 10. 나형(45%). 28번. 4점 (교육과정 삭제 - 어렵진 않지만 넘어가도 좋다.)

연속확률변수 $X$가 갖는 값의 범위가 $0 \le X \le 1$이고 확률밀도함수의 그래프는 그림과 같다.

확률변수 $X$의 평균이 $\mathrm{E}(X) = \dfrac{q}{p}$일 때, $p+q$의 값을 구하시오.

(단, $p$와 $q$는 서로소인 자연수이다.)



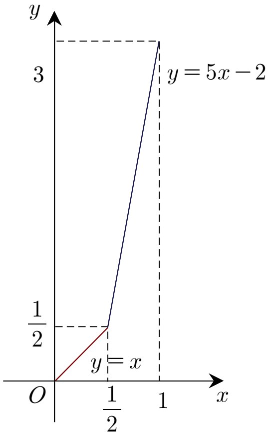

주어진 그래프의 함수식을 세우면 오른쪽과 같다.

$$\text{즉, } f(x) = \begin{cases} 5x - 2 & \left(\frac{1}{2} \le x \le 1\right) \\ x & \left(0 \le x \le \frac{1}{2}\right) \end{cases} \text{이다.}$$

$$E(X) = \int_0^1 x f(x)dx = \int_0^{\frac{1}{2}} x f(x)dx + \int_{\frac{1}{2}}^1 x f(x)dx$$

$$= \int_0^{\frac{1}{2}} x^2 dx + \int_{\frac{1}{2}}^1 x(5x-2)dx$$

$$= \left[\frac{1}{3}x^3\right]_0^{\frac{1}{2}} + \left[\frac{5}{3}x^3 - x^2\right]_{\frac{1}{2}}^1 = \frac{1}{24} + \frac{17}{24} = \frac{3}{4}$$




정답은 7

**139** 이해를 위한 예제 2011. 11. 나형(57%). 27번. 4점 (교육과정 삭제 - 어렵진 않지만 넘어가도 좋다.)

구간 [0, 1]에서 정의된 연속확률변수 $X$의 확률밀도 함수가 $f(x)$이다. $X$의 평균이 $\frac{1}{4}$이고, $\int_0^1 (ax+5)f(x)dx = 10$일 때, 상수 $a$의 값을 구하시오.

조건 구간 [0, 1]에서 정의된 연속확률변수 $X$의 확률밀도 함수가 $f(x)$이다. $X$의 평균이 $\frac{1}{4}$

- $E(X) = \int_0^1 xf(x)dx = \frac{1}{4}$이라는 뜻이다.

조건 $\int_0^1 (ax+5)f(x)dx = 10$

- $\int_0^1 (ax+5)f(x)dx = 10 \Leftrightarrow a\int_0^1 xf(x)dx = 5\int_0^1 f(x)dx = 10 \Leftrightarrow aE(X) + 5 = 10$

$\Leftrightarrow \frac{1}{4}a = 5$ 〈문제의 조건에서 $E(X) = \frac{1}{4}$이므로〉 $\Leftrightarrow a = 20$이다.

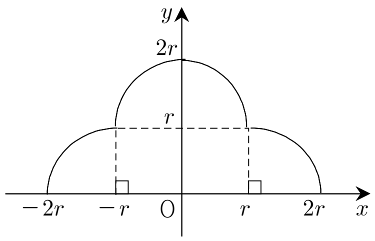
정답은 20

일단 풀고 맞았다고 하더라도 뒤의 해설을 정독하여 읽는다. 왜냐하면 풀이과정이 암기되어서 문제가 그냥 풀리거나 그냥 직관적으로 판단이 되어서 문제가 풀리는 경우도 많기 때문이다. 우리가 문제를 푸는 이유는 앞에서 배운 개념을 확인하기 위해서다. 이 문제를 푸는 것이 중요한 게 아니라 다른 어려운 문제에까지 적용할 수 있는 개념을 익히는 게 중요한 것이다.

예제 066

2009. 4. 가형(66%). 10번. 3점

연속확률변수 $X$의 확률밀도함수 $y=f(x)$ $(-2r \le x \le 2r)$의 그래프가 그림과 같이 반지름의 길이가 $r$인 사분원들의 호로 이루어져 있을 때, $P\left(0 \le X \le \dfrac{1}{\sqrt{\pi+2}}\right)$의 값은?




① $\dfrac{1}{4}$ ② $\dfrac{1}{3}$ ③ $\dfrac{\pi}{7}$ ④ $\dfrac{1}{\pi+2}$ ⑤ $\dfrac{\pi+4}{4\pi+8}$

예제 067

2004. 3. 가형(%). 14번. 4점

2005학년도 대학수학능력시험 수리영역의 원점수 $X$의 평균을 $m$, 표준편차를 $\sigma$라 할 때, 표준점수 $T$는 $T=a\left(\dfrac{X-m}{\sigma}\right)+b$ (단, $a>0$)꼴로 나타내어진다. 수리영역의 표준점수 $T$가 평균이 $100$, 표준편차가 20인 분포를 이룬다고 할 때, 두 상수 $a, b$의 합 $a+b$의 값은?

① 80 ② 90 ③ 100 ④ 110 ⑤ 120

예제 068

2011. 사관학교. 문과. 12번. 3점

평균이 $m$ , 표준편차가 $\sigma$ 인 정규분포를 따르는 확률변수 $X$와 표준정규분포를 따르는 확률변수 $Z$가 다음 두 조건을 만족시킨다. $m+\sigma$ 의 값은?

| | |
|---|---|
| (가) $\mathrm{P}(X \geq 58) = \mathrm{P}(Z \geq -1)$ | (나) $\mathrm{P}(X \leq 55) = \mathrm{P}(Z \geq 2)$ |

① 62 ② 63 ③ 64 ④ 65 ⑤ 66

예제 069

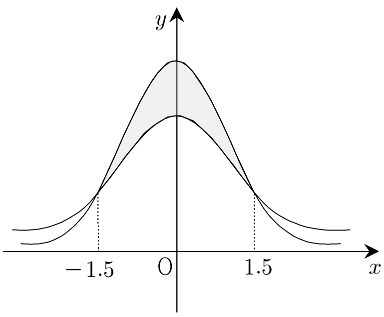
2008. 사관학교. 문·이과. 16번. 3점

확률변수 $X$는 정규분포 $\mathrm{N}(0,\ \sigma^2)$을 따르고, 확률변수 $Z$는 표준정규분포 $\mathrm{N}(0,\ 1^2)$ 을 따른다. 두 확률변수 $X$, $Z$의 확률밀도함수를 각각 $f(x)$, $g(x)$라 할 때, 다음 조건이 모두 성립한다.

[조건]

(가) $\sigma > 1$ (나) 두 곡선 $y=f(x)$, $y=g(x)$는 $x=-1.5$, $x=1.5$ 일 때 만난다.

두 곡선 $y=f(x)$, $y=g(x)$로 둘러싸인 부분의 넓이가 0.096 일 때, $X$의 표준편차 $\sigma$의 값을 아래 표준정규분포표를 이용하여 구한 것은?




| $z$ | $\mathrm{P}(0 \leq Z \leq z)$ |
|---|---|
| 1.2 | 0.385 |
| 1.5 | 0.433 |
| 2.0 | 0.477 |

① 1.20 ② 1.25 ③ 1.50 ④ 1.75 ⑤ 2.00

예제 070

2013. 9. B형(78%). 20번. 4점

양의 실수 전체의 집합에서 정의된 함수 $G(t)$ 는 평균이 $t$, 표준편차가 $\frac{1}{t^2}$ 인 정규분포를 따르는 확률변수 $X$에 대하여 $G(t)=P\left(X \leq \frac{3}{2}\right)$ 이다. 함수 $G(t)$ 의 최댓값을 표준정규분포표를 이용하여 구한 것은?

| $z$ | $P(0 \leq Z \leq z)$ |
|---|---|
| 0.4 | 0.1554 |
| 0.5 | 0.1915 |
| 0.6 | 0.2257 |
| 0.7 | 0.2580 |

① 0.3085 ② 0.3446 ③ 0.6915
④ 0.7257 ⑤ 0.7580

예제 071

2005. 9. 가형(61%), 나형(50%). 16번. 4점

세 확률변수 $X$, $Y$, $W$는 각각 다음과 같다.

$X$는 이항분포 $B\left(100, \frac{1}{5}\right)$을 따른다. $Y$는 이항분포 $B\left(225, \frac{1}{5}\right)$을 따른다.

$W$는 이항분포 $B\left(400, \frac{1}{5}\right)$을 따른다.

<보기>에서 옳은 것을 모두 고른 것은?

<보 기>

ㄱ. $P\left(\left|\frac{X}{100}-\frac{1}{5}\right| < \frac{1}{10}\right) < P\left(\left|\frac{W}{400}-\frac{1}{5}\right| < \frac{1}{10}\right)$

ㄴ. $P\left(\left|\frac{X}{100}-\frac{1}{5}\right| < \frac{1}{10}\right) < P\left(\left|\frac{Y}{225}-\frac{1}{5}\right| < \frac{1}{25}\right)$

ㄷ. $P\left(\left|\frac{Y}{225}-\frac{1}{5}\right| < \frac{1}{25}\right) < P\left(\left|\frac{W}{400}-\frac{1}{5}\right| < \frac{1}{25}\right)$

① ㄱ ② ㄴ ③ ㄱ, ㄷ ④ ㄴ, ㄷ ⑤ ㄱ, ㄴ, ㄷ

예제 072

2004. 11. 가형(76%), 나형(54%). 16번. 3점

다음은 어느 백화점에서 판매하고 있는 등산화에 대한 제조회사별 고객의 선호도를 조사한 표이다.

| $z$ | $P(0 \leq Z \leq z)$ |
|---|---|
| 0.5 | 0.1915 |
| 1.0 | 0.3413 |
| 1.5 | 0.4332 |
| 2.0 | 0.4772 |

| 제조회사 | A | B | C | D | 계 |
|---|---|---|---|---|---|
| 선호도(%) | 20 | 28 | 25 | 27 | 100 |

192명의 고객이 각각 한 켤레씩 등산화를 산다고 할 때, C 회사 제품을 선택할 고객이 42명 이상일 확률을 위에 표준정규분포 표를 이용하여 구한 것은?

① 0.6915　② 0.7745　③ 0.8256　④ 0.8332　⑤ 0.8413

예제 073

2007. 3. 가형(71%). 13번. 4점

$\sum_{k=351}^{369} {}_{400}\mathrm{C}_k \left(\frac{9}{10}\right)^k \left(\frac{1}{10}\right)^{400-k}$ 의 값을 아래 표준정규분포 표를 이용하여 구한 것은?

① 0.1587　② 0.3085　③ 0.6826　④ 0.8664　⑤ 0.9544

< 표준정규분포 표 >

| $z$ | $P(0 \leq Z \leq z)$ |
|---|---|
| 0.5 | 0.1915 |
| 1.0 | 0.3413 |
| 1.5 | 0.4332 |
| 2.0 | 0.4772 |

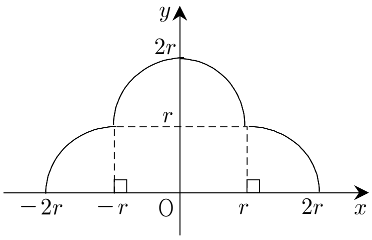
# 정답과 해설

예제 066

REVIEW 연속확률분포

2009. 4. 가형(66%). 10번. 3점

연속확률변수 $X$의 확률밀도함수 $y=f(x)$ $(-2r \le x \le 2r)$의 그래프가 그림과 같이 반지름의 길이가 $r$인 사분원들의 호로 이루어져 있을 때, $P\left(0 \le X \le \frac{1}{\sqrt{\pi+2}}\right)$의 값은?




① $\frac{1}{4}$　② $\frac{1}{3}$　③ $\frac{\pi}{7}$　④ $\frac{1}{\pi+2}$　⑤ $\frac{\pi+4}{4\pi+8}$

- 함수 $f(x)$의 그래프와 $x$축으로 둘러싸인 부분의 넓이는 1이므로 ⇨ $\pi r^2 + 2r^2 = 1 \Leftrightarrow r = \frac{1}{\sqrt{\pi+2}}$

$\mathrm{P}(0 \le X \le r) = \frac{\pi r^2}{4} + r^2 = \frac{\pi+4}{4\pi+8}$

정답은 ⑤

예제 067

REVIEW 연속확률분포의 평균, 분산, 표준편차의 성질

2004. 3. 가형(%). 14번. 4점

2005학년도 대학수학능력시험 수리영역의 원점수 $X$의 평균을 $m$, 표준편차를 $\sigma$라 할 때, 표준점수 $T$는 $T=a\left(\frac{X-m}{\sigma}\right)+b$ (단, $a>0$)꼴로 나타내어진다. 수리영역의 표준점수 $T$가 평균이 $100$, 표준편차가 20인 분포를 이룬다고 할 때, 두 상수 $a, b$의 합 $a+b$의 값은?

① 80 ② 90 ③ 100 ④ 110 ⑤ 120

조건 수리영역의 원점수 $X$의 평균을 $m$, 표준편차를 $\sigma$라 할 때,

- $X$는 정규분포 $N(m, \sigma^2)$을 따른다.

조건 표준점수 $T$는 $T=a\left(\frac{X-m}{\sigma}\right)+b$ (단, $a>0$)

- $T = a\left(\frac{X-m}{\sigma}\right) + b$에서 $\frac{X-m}{\sigma} = Z$라 하면 $T = aZ + b$이고, $E(Z) = 0$, $\sigma(Z) = 1$이다.

조건 표준점수 $T$가 평균이 $100$, 표준편차가 20인 분포를 이룬다고 할 때,

- $E(T) = 100 \Leftrightarrow E(aZ+b) = 100 \Leftrightarrow aE(Z) + b = 100 \Leftrightarrow b = 100$
- $\sigma(T) = 20 \Leftrightarrow \sigma(az+b) = 20 \Leftrightarrow |a|\sigma(Z) = 20 \Leftrightarrow |a| = 20 \Leftrightarrow a>0$이므로 $a = 20$

$a+b = 120$이므로 정답은 ⑤

코칭 정규분포라는 말이 없지만 상황 상 정규분포를 따른다고 가정해야 한다. (보통은 주어지므로 너무 고민하지 마라.)

코칭 연속확률분포에서도 평균, 분산, 표준편차의 성질이 성립한다.

예제 068

REVIEW 표준화와 표준정규분포

2011. 사관학교. 문과. 12번. 3점

평균이 $m$, 표준편차가 $\sigma$인 정규분포를 따르는 확률변수 $X$와 표준정규분포를 따르는 확률변수 $Z$가 다음 두 조건을 만족시킨다. $m+\sigma$ 의 값은?

| | |
|---|---|
| (가) $\mathrm{P}(X \geq 58) = \mathrm{P}(Z \geq -1)$ | (나) $\mathrm{P}(X \leq 55) = \mathrm{P}(Z \geq 2)$ |

① 62 ② 63 ③ 64 ④ 65 ⑤ 66

조건 평균이 $m$, 표준편차가 $\sigma$인 정규분포를 따르는 확률변수 $X$와 표준정규분포를 따르는 확률변수 $Z$

- $X : N(m, \sigma^2)$, $Z : N(0, 1^2)$

조건 (가) $\mathrm{P}(X \geq 58) = \mathrm{P}(Z \geq -1)$

- $z$와 $k$의 관계를 안다면 표준화하지 않아도 바로 관계를 구할 수 있다. ⇨ $58 = m-\sigma$

조건 $\mathrm{P}(X \leq 55) = \mathrm{P}(Z \geq 2)$

- 역시 표준화하지 않고 구할 수 있다. $\mathrm{P}(X \leq 55) = P(Z \geq 2)$을 $\mathrm{P}(X \leq 55) = P(Z \leq -2)$라고 생각하면
⇨ $55 = m-2\sigma$
- 두 식을 연립해서 풀면 $m=61$, $\sigma=3$이므로 $m+\sigma=64$

정답은 ③

예제 069

REVIEW 정규분포의 그래프와 표준정규분포

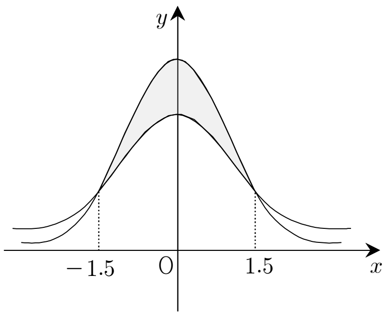
2008. 사관학교. 문·이과. 16번. 3점

확률변수 $X$는 정규분포 $\mathrm{N}(0,\ \sigma^2)$을 따르고, 확률변수 $Z$는 표준정규분포 $\mathrm{N}(0,\ 1^2)$을 따른다.
두 확률변수 $X$, $Z$의 확률밀도함수를 각각 $f(x)$, $g(x)$라 할 때, 다음 조건이 모두 성립한다.

[조건]

(가) $\sigma > 1$　　(나) 두 곡선 $y=f(x)$, $y=g(x)$는 $x=-1.5$, $x=1.5$일 때 만난다.

두 곡선 $y=f(x)$, $y=g(x)$로 둘러싸인 부분의 넓이가 $0.096$일 때, $X$의 표준편차 $\sigma$의 값을 아래 표준정규분포표를 이용하여 구한 것은?




| $z$ | $\mathrm{P}(0 \leq Z \leq z)$ |
|---|---|
| 1.2 | 0.385 |
| 1.5 | 0.433 |
| 2.0 | 0.477 |

① 1.20　② 1.25　③ 1.50　④ 1.75　⑤ 2.00

조건 확률변수 $X$는 정규분포 $\mathrm{N}(0,\ \sigma^2)$을 따르고, 확률변수 $Z$는 표준정규분포 $\mathrm{N}(0,\ 1^2)$을 따른다.

- $X$와 $Z$는 평균이 같다.

조건 (가) $\sigma > 1$ - $\sigma > 1$이므로 상대적으로 위에 있는 그래프가 $Z$의 그래프이다.

조건 (나) 두 곡선 $y=f(x)$, $y=g(x)$는 $x=-1.5$, $x=1.5$일 때 만난다. + $y=f(x)$, $y=g(x)$로 둘러싸인 부분의 넓이가 $0.096$

- 주어진 그래프는 평균에 대한 대칭이므로 $P(0 \leq Z \leq 1.5) - P(0 \leq X \leq 1.5) = 0.048$

$$P(0 \leq Z \leq 1.5) - P(0 \leq X \leq 1.5) = 0.048 \Leftrightarrow P(0 \leq Z \leq 1.5) - P\left(0 \leq Z \leq \frac{1.5}{\sigma}\right) = 0.048$$

$$\Leftrightarrow 0.433 - P\left(0 \leq Z \leq \frac{1.5}{\sigma}\right) = 0.048$$

$$\Leftrightarrow P\left(0 \leq Z \leq \frac{1.5}{\sigma}\right) = 0.385 \Leftrightarrow \frac{1.5}{\sigma} = 1.2 \Leftrightarrow \sigma = 1.25$$

정답은 ②

예제 070

REVIEW 문제에서 정의된 함수 + 표준화

2013. 9. B형(78%). 20번. 4점

양의 실수 전체의 집합에서 정의된 함수 $G(t)$ 는 평균이 $t$, 표준편차가 $\frac{1}{t^2}$ 인 정규분포를 따르는 확률변수 $X$ 에 대하여 $G(t)=P\left(X\leq\frac{3}{2}\right)$ 이다. 함수 $G(t)$ 의 최댓값을 표준정규분포표를 이용하여 구한 것은?

| $z$ | $P(0\leq Z\leq z)$ |
|---|---|
| 0.4 | 0.1554 |
| 0.5 | 0.1915 |
| 0.6 | 0.2257 |
| 0.7 | 0.2580 |

① 0.3085 ② 0.3446 ③ 0.6915
④ 0.7257 ⑤ 0.7580

조건 평균이 $t$, 표준편차가 $\frac{1}{t^2}$ 인 정규분포를 따르는 확률변수 $X$ - $X : N\left(t,\ \frac{1}{t}\right)$을 따른다.

조건 $G(t)=P\left(X\leq\frac{3}{2}\right)$

- $t$에 대한 함수 $G(t)$를 $t$에 대한 식으로 표현하기 위해 <표준화>를 한다.

$$G(t) = P\left(X\leq\frac{3}{2}\right) = P\left(Z\leq\frac{\frac{3}{2}-t}{\frac{1}{t^2}}\right) = P\left(Z\leq\frac{3}{2}t^2-t^3\right)$$

이 값이 최댓값을 가지려면 $\frac{3}{2}t^2-t^3$이 최댓값을 가지면 된다. (간단한 삼차함수의 최대 · 최소이다.)

$f(t)=\frac{3}{2}t^2-t^3$이라 하면 $f'(t)=3t-3t^2$이므로 $t=1$일 때 최댓값 $f(1)=\frac{1}{2}$을 가진다.

⇨ 즉, $G(1) = P\left(Z\leq\frac{1}{2}\right) = 0.6915$

정답 ③

REVIEW 이항분포의 정규분포 화, 서로 다른 확률 변수의 확률 관계식 - 표준화

예제 071

2005. 9. 가형(61%), 나형(50%). 16번. 4점

세 확률변수 $X$, $Y$, $W$는 각각 다음과 같다.

| | |
|---|---|
| $X$는 이항분포 $B\left(100, \frac{1}{5}\right)$을 따른다. | $Y$는 이항분포 $B\left(225, \frac{1}{5}\right)$을 따른다. |
| $W$는 이항분포 $B\left(400, \frac{1}{5}\right)$을 따른다. | |

<보기>에서 옳은 것을 모두 고른 것은?

<보 기>

ㄱ. $P\left(\left|\frac{X}{100}-\frac{1}{5}\right| < \frac{1}{10}\right) < P\left(\left|\frac{W}{400}-\frac{1}{5}\right| < \frac{1}{10}\right)$

ㄴ. $P\left(\left|\frac{X}{100}-\frac{1}{5}\right| < \frac{1}{10}\right) < P\left(\left|\frac{Y}{225}-\frac{1}{5}\right| < \frac{1}{25}\right)$

ㄷ. $P\left(\left|\frac{Y}{225}-\frac{1}{5}\right| < \frac{1}{25}\right) < P\left(\left|\frac{W}{400}-\frac{1}{5}\right| < \frac{1}{25}\right)$

① ㄱ ② ㄴ ③ ㄱ, ㄷ ④ ㄴ, ㄷ ⑤ ㄱ, ㄴ, ㄷ

- 주어진 이항분포의 확률변수 $X$, $Y$, $Z$는 모두 $n$이 충분히 크므로 <정규분포>를 따른다.

| | | |
|---|---|---|
| $X : B\left(100, \frac{1}{5}\right) \Rightarrow X : N(20, 4^2)$ | $Y : B\left(225, \frac{1}{5}\right) \Rightarrow Y : N(45, 6^2)$ | $W : B\left(400, \frac{1}{5}\right) \Rightarrow W : N(80, 8^2)$ |

- 서로 다른 확률 변수의 확률을 비교하려면 표준화를 하는 것이 좋다.

ㄱ. $P\left(\left|\frac{X}{100}-\frac{1}{5}\right| < \frac{1}{10}\right) < P\left(\left|\frac{W}{400}-\frac{1}{5}\right| < \frac{1}{10}\right) \Leftrightarrow P\left(\left|\frac{X-20}{100}\right| < \frac{1}{10}\right) < P\left(\left|\frac{W-80}{400}\right| < \frac{1}{10}\right)$ <통분>

<마치 표준화 하라고 모양을 맞추어 준 것 같다.> $\Leftrightarrow P\left(\left|\frac{X-20}{100}\right| \cdot 25 < \frac{1}{10} \cdot 25\right) < P\left(\left|\frac{W-80}{400}\right| \cdot 50 < \frac{1}{10} \times 50\right)$

$\Leftrightarrow P\left(\left|\frac{X-20}{4}\right| < 2.5\right) < P\left(\left|\frac{W-80}{8}\right| < 5\right) \Leftrightarrow P(|Z| < 2.5) < P(|Z| < 5)$이므로 (참)

ㄴ. $P\left(\left|\frac{X}{100}-\frac{1}{5}\right| < \frac{1}{10}\right) < P\left(\left|\frac{Y}{225}-\frac{1}{5}\right| < \frac{1}{25}\right) \Leftrightarrow P\left(\left|\frac{X-20}{100}\right| < \frac{1}{10}\right) < P\left(\left|\frac{Y-45}{225}\right| < \frac{1}{25}\right)$

$\Leftrightarrow P\left(\left|\frac{X-20}{100}\right| \times 25 < \frac{1}{10} \times 25\right) < P\left(\left|\frac{Y-45}{225}\right| \times \frac{225}{6} < \frac{1}{25} \times \frac{225}{6}\right)$

$\Leftrightarrow P\left(\left|\frac{X-20}{4}\right| < 2.5\right) < P\left(\left|\frac{Y-45}{6}\right| < 1.5\right) \Leftrightarrow P(|Z| < 2.5) < P(|Z| < 1.5)$이므로 (거짓)

ㄷ. $P\left(\left|\frac{Y}{225}-\frac{1}{5}\right| < \frac{1}{25}\right) < P\left(\left|\frac{W}{400}-\frac{1}{5}\right| < \frac{1}{25}\right) \Leftrightarrow P\left(\left|\frac{Y-45}{225}\right| < \frac{1}{25}\right) < P\left(\left|\frac{W-80}{400}\right| < \frac{1}{25}\right)$ <통분>

$\Leftrightarrow P\left(\left|\frac{Y-45}{225}\right| \cdot \frac{225}{6} < \frac{1}{25} \cdot \frac{225}{6}\right) < P\left(\left|\frac{W-80}{400}\right| \cdot 50 < \frac{1}{25} \cdot 50\right)$

$\Leftrightarrow P\left(\left|\frac{Y-45}{6}\right| < 1.5\right) < P\left(\left|\frac{W-80}{8}\right| < 2\right) \Leftrightarrow P(|Z| < 1.5) < P(|Z| < 2)$이므로 (참)

정답은 ③

예제 072

REVIEW 이항분포의 정규분포 화

2004. 11. 가형(76%), 나형(54%). 16번. 3점

다음은 어느 백화점에서 판매하고 있는 등산화에 대한 제조회사별 고객의 선호도를 조사한 표이다.

| 제조회사 | A | B | C | D | 계 |
|---|---|---|---|---|---|
| 선호도(%) | 20 | 28 | 25 | 27 | 100 |

| $z$ | $\mathrm{P}(0 \le Z \le z)$ |
|---|---|
| 0.5 | 0.1915 |
| 1.0 | 0.3413 |
| 1.5 | 0.4332 |
| 2.0 | 0.4772 |

192명의 고객이 각각 한 켤레씩 등산화를 산다고 할 때, C 회사 제품을 선택할 고객이 42명 이상일 확률을 위에 표준정규분포 표를 이용하여 구한 것은?

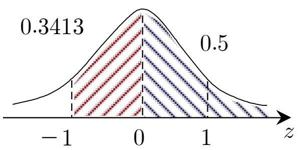
① 0.6915 ② 0.7745 ③ 0.8256 ④ 0.8332 ⑤ 0.8413

조건 192명의 고객이 각각 한 켤레씩 등산화를 산다고 할 때, $C$회사 제품을 선택할 고객이 42명 이상일 확률

- $P(C\text{제품 살}) = \frac{1}{4}$ (선호도 표를 통해서 확인)
- $C$제품을 선택한 고객의 인원수를 확률변수 $X$라고 하면 $X$는 $B\left(192, \frac{1}{4}\right)$를 따른다.

여기에서 $m = 192 \cdot \frac{1}{4} = 48$이고, $\sigma^2 = 192 \cdot \frac{1}{4} \cdot \frac{3}{4} = 36$이므로

$X$는 근사적으로 $N(48,\ 6^2)$을 따른다.




⇨ $P(X \ge 42) = P\left(\frac{X-48}{6} \ge \frac{42-48}{6}\right) = P(Z \ge -1) = 0.3413 + 0.5 = 0.8413$

정답은 ⑤

예제 073

REVIEW 시그마 표현

2007. 3. 가형(71%). 13번. 4점

$\sum_{k=351}^{369} {}_{400}\mathrm{C}_k \left(\frac{9}{10}\right)^k \left(\frac{1}{10}\right)^{400-k}$ 의 값을 아래 표준정규분포 표를 이용하여 구한 것은?

〈 표준정규분포 표 〉

| $z$ | $\mathrm{P}(0 \le Z \le z)$ |
|---|---|
| 0.5 | 0.1915 |
| 1.0 | 0.3413 |
| 1.5 | 0.4332 |
| 2.0 | 0.4772 |

① 0.1587 ② 0.3085 ③ 0.6826 ④ 0.8664 ⑤ 0.9544

조건 $\sum_{k=351}^{369} {}_{400}\mathrm{C}_k \left(\frac{9}{10}\right)^k \left(\frac{1}{10}\right)^{400-k}$

- 확률변수 $K$는 $B\left(400, \frac{9}{10}\right)$이고 근사적으로 $N(360, 6^2)$을 따른다.

$$\sum_{k=351}^{369} {}_{400}\mathrm{C}_k \left(\frac{9}{10}\right)^k \left(\frac{1}{10}\right)^{400-k} = P(351 \le K \le 369) = P\left(\frac{351-360}{6} \le \frac{K-360}{6} \le \frac{369-360}{6}\right)$$

$$= P(-1.5 \le Z \le 1.5) = 0.8664$$

정답은 ④

## 존철살인 #6. 표본평균과 표본비율의 차이.

### 누구나 이해 할 수 있다.

- 표본평균과 표본비율의 차이는 누구나 이해할 수 있는 직관적이고 쉬운 것이다.
  표본평균과 표본비율은 관찰의 초점만 다를 뿐 모든 것이 같다. 즉, 조사하고자 하는 집단, 조사하는 방식 모든 것이 같다.
  단지 같은 집단에서 같은 방식으로 조사하되 <관찰의 초점>이 다른 것뿐이다.

- 예컨대 우리 반 50명의 수학점수를 확률변수로 가지는 집단이 있다고 하자.
  우리의 관찰의 초점이 우리 반의 <수학점수의 평균>이라면 이 50명의 평균을 <모평균>이라 부른다.
  우리의 관찰의 초점이 우리 반의 <평균점수를 넘는 사람들의 비율>이라면 $\frac{\text{평균을 넘는 사람 수}}{50}$ 를 <모비율>이라 부른다.

- 우리가 조사하고자 하는 집단의 대상이 너무 많아 <일부만 뽑아서 조사>하는 경우 이 일부에 해당하는 집단이 <표본>이다.
  이렇게 표본을 조사해서 <모평균>과 <모비율>을 예측하는 것을 <추정>이라고 한다.

가령 위의 집단에서 10명의 표본을 뽑아서 표본조사를 하는 경우 규칙은 다음과 같다.
1단계 : 50명이 핸드폰이 있다고 가정하고 50명 중에 무작위로 한 명을 뽑아 전화를 걸어 점수를 물어본다.
2단계 : 방금 전화를 건 사람을 포함하여 다시 50명 중에 무작위로 한 명을 뽑아 전화를 걸어 점수를 물어본다.
3단계 : 이러한 시행을 10번하여 총 10명의 점수를 수집한다.

이렇게 조사한 10명의 점수를 가지고 평균을 내면 <표본평균>이 된다. 또한 10명의 점수 중 평균점수를 넘는 사람의 수만 조사하여 $\frac{\text{평균을 넘는 사람 수}}{10}$ 를 구하면 이것이 <표본비율>이다.
<표본평균, 표본비율>은 당연히 <모평균, 모비율>과 다를 가능성이 높다. 하지만 비슷할 것이다.
그래서 <표본평균, 표본비율>을 가지고 <모평균, 모비율>을 예측하는 과정과 그에 따른 식을 <통계적 추정>이라고 한다.

### 논리의 흐름을 이해하면 어렵지 않다.

- 위와 같이 직관적으로 이해하는 것은 매우 쉽지만 교과서 상의 추정을 이해하는 과정은 다음의 과정을 따르기 때문에 학습자의 입장에서는 헷갈릴 수 있다.

<표본평균의 이해 → 모평균의 추정> <표본비율의 이해 → 모비율의 추정>

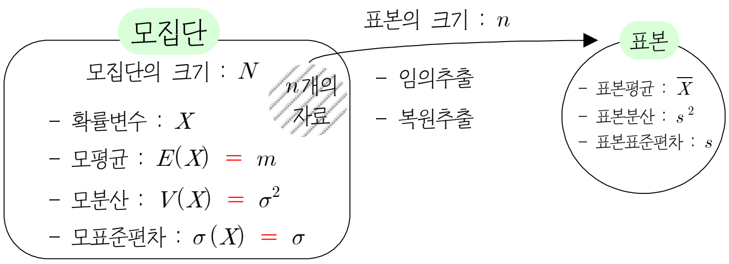
위와 같은 과정에서 <표본평균(비율)의 이해>는 <모평균(비율)의 추정>과는 관련 없어 보이는 내용을 다룬다.
하지만 <표본평균(비율)의 이해>는 누구나 느낄 수 있는 <표본평균(비율)>과 <모평균(비율)>의 대략적인 관계가 아닌 수식적으로 표현되는 명확한 관계를 나타내기 위해 반드시 필요한 내용이며 느낌상으로는 <추정>보다 <정규분포>에 가깝다.
이 과정을 받아들이고 이해하면 <표본평균(비율)>을 통해서 <모평균(비율)의 예측>하는 과정을 수식적으로 다룰 수 있게 된다. 여기까지 공부하면 우리의 기나긴 <고등 확률과 통계의 여정>은 끝이 나게 된다.

## #1 표본평균의 이해

- 용어의 이해가 반이다. 수학적 내용을 익히기 전에 용어를 익히고 또 익힌다.
- 〈통계적 추정〉을 위한 내용적 근거를 공부하는 부분이다. 내용상 〈추정〉보다는 〈확률의 곱셈정리, 정규분포〉문제에 가깝다. 〈복원추출〉의 의미를 정확히 알고, 집단을 형성하는 〈표본의 평균〉들의 평균, 분산, 표준편차를 구하는 것의 복잡함을 느낀다.
- 〈모집단〉의 평균, 분산, 표준편차를 통해서 〈표본의 평균〉들의 평균, 분산, 표준편차를 구하는 공식을 익힌다.
  (모집단의 자료가 너무 많을 때 추정하는 것이지만 역설적이게도 모집단의 평균을 구하는 것이 표본 집단의 평균을 구하는 것보다 훨씬 간단하다.)

# 1. 표본평균($\overline{X}$)의 이해

### 1) 이해 1단계 : 용어와 의미




① 확률변수 $X$ : 모집단의 확률변수를 의미한다.

⇨ 예컨대 모집단이 우리 반 50명이라면 확률변수 $X$는 우리 반 50명의 수학점수라고 생각할 수 있다.

② 모집단의 크기 ($N$) : 모집단 안의 자료의 개수

⇨ 예컨대 모집단이 우리 반 50명이라면 모집단의 크기는 $50$이다. (즉, $N = 50$)

③ 추출(sampling) : 모집단의 자료 중 일부를 뽑는 것으로 이 자료를 표본이라고 한다.

⇨ 반드시 〈임의 추출〉해야 한다. 표본을 통해 모집단의 성질을 예측하는 것이 추정의 본질인데 〈의도를 가지고 추출〉할 경우 우리의 원래 목적인 〈모집단〉과는 성질이 완전히 다른 〈표본〉이 나올 수 있기 때문이다.

④ 표본의 크기 ($n$) : 표본 1개 안에 있는 자료의 개수이다.

⇨ 예컨대 우리 반 50명 중 10명을 추출하여 조사한다고 할 때, 표본의 크기는 $10$이다. (즉, $n = 10$)

⑤ 〈표본의 크기〉가 $n$인 표본을 〈복원추출〉

⇨ 예컨대 우리 반 50명 중 〈표본의 크기가 $10$〉인 표본을 〈복원 추출〉하여 조사한다고 할 때 〈복원 추출로 한 번에 10명을 뽑는 것〉이 아니라 〈복원추출로 1명씩 10번 뽑아〉 조사하는 것이다.

⇨ 이것이 전제가 되어야 〈추정의 공식〉이 성립한다. - 공식은 잠시 후에 다루기로 한다. (뒷장에 ⑦까지 있습니다.)

⑥ $m$ (모평균), $\sigma^2$(모분산), $\sigma$(모표준편차)

: 모집단의 확률변수에 대한 평균, 분산, 표준편차를 말하는 것이다. <표본평균, 표본분산, 표본표준편차>와 구분하기 위하여 앞에 <모>라는 글자를 추가한 것이다.

⇨ 예컨대 모집단이 우리 반 50명이고, 확률변수가 50명의 수학점수라면 모평균은 50명의 점수의 평균이고, 모분산은 50명의 점수의 분산이고, 모표준편차는 50명의 점수의 표준편차이다.

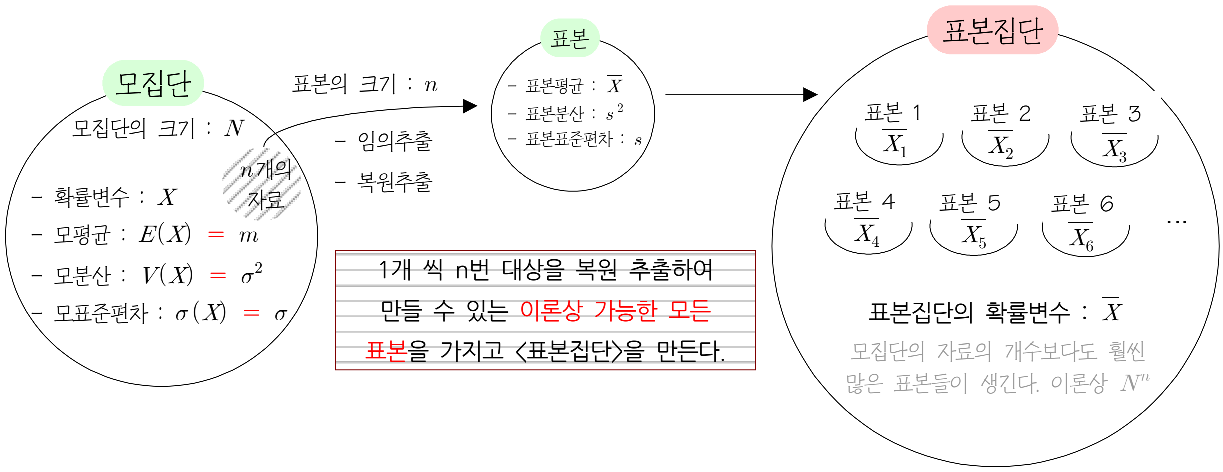
⑦ $\overline{X}$ (표본평균), $s^2$(표본분산), $s$(표본표준편차)

: $\overline{X}$(표본평균), $s^2$(표본분산), $s$(표본표준편차)는 모두 <표본 1개 안에 있는 $n$개의 자료>의 평균, 분산, 표준편차를 말하는 것이다. 후에 나오는 <$\overline{X}$의 평균, 분산, 표준편차>와는 분명히 구분을 해야 하는 개념이다.

(이미 모 표준편차에서 $\sigma$를 사용했기 때문에 표본표준편차는 그리스 알파벳 $\sigma$에 해당하는 알파벳인 $s$라고 표현한다.)

⇨ 예컨대 모집단이 우리 반 50명이고, 확률변수가 50명의 수학점수라 하자. 여기에서 10명을 추출하여 조사한다고 할 때 (임의추출, 복원추출) 표본평균은 10명의 점수의 평균이고, 표본분산은 10명의 점수의 분산이고, 표본표준편차는 10명의 점수의 표준편차이다. <앞장의 그림을 보며 용어의 의미를 다시 생각해본다. - 5회 이상>

## 2) 이해 2단계 : $\overline{X}$를 확률변수로 인정하는 과정




- 표본 집단이란 이론상 가능한 모든 표본들을 모아 놓은 집단이다. 이렇게 모인 <표본들의 개수>는 이론상 모집단의 자료의 개수보다 훨씬 많다. 이런 <표본들의 개수>를 <표본 집단의 크기>라고 하는데 별로 중요하지 않다. 단지 <표본집단의 자료의 개수>가 <모집단의 자료의 개수>보다도 **훨씬 많음을 눈치 채는 정도면 충분**하다.

⇨ 예컨대 모집단이 우리 반 50명이고, 확률변수가 50명의 수학점수라 하자. 여기에서 10명을 추출하여 조사한다고 할 때, 이론상 가능한 모든 <표본의 개수>는 $50^{10}$이다. (곱의 법칙으로 계산)

⇨ <표본의 크기>는 표본 1개 안에 들어있는 자료의 개수이므로 10이고, <표본 집단의 크기>는 <표본>이 모여 있는 집단의 크기. 즉, 이론상 가능한 <표본의 개수>이므로 $50^{10}$이다.

- 새로운 확률변수 $\overline{X}$

: <표본 집단의 표본의 개수>가 <모집단의 자료의 개수>보다 이론상 훨씬 많으므로 <표본의 평균 - $\overline{X}$> 역시 모집단의 자료 $X$보다도 훨씬 많다. 즉, $\overline{X}$를 하나의 확률변수로 인정하여 분포를 따질 수 있다.

# 2. $\overline{X}$의 평균(기댓값), 분산, 표준편차 (선 암기, 후 이해)

## 1) 공식과 의미 Reading

- 모평균, 모분산, 모표준편차를 통해서 ⇨ $\overline{X}$의 평균, $\overline{X}$의 분산, $\overline{X}$의 표준편차를 구하다.

① 공식 : $E(\overline{X}) = E(X) = m$

의미 : 〈$\overline{X}$의 평균〉은 〈$X$의 평균 즉, 모평균〉과 같다.

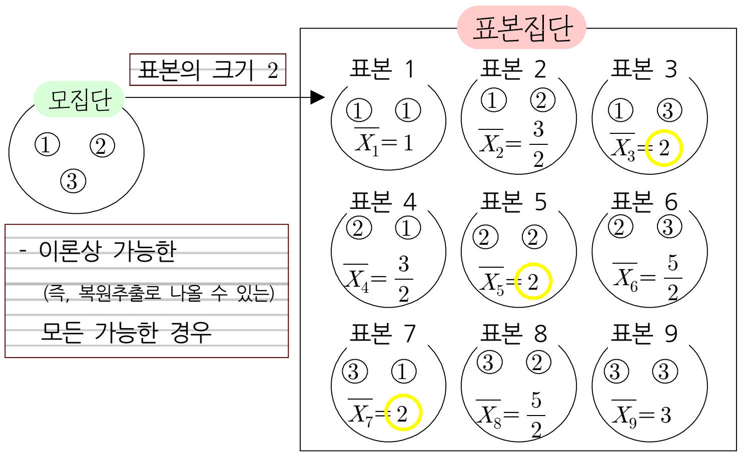
② 공식 : $V(\overline{X}) = \dfrac{V(X)}{n} = \dfrac{\sigma^2}{n}$

의미 : 〈$\overline{X}$의 분산〉은 〈$X$의 분산 즉, 모 분산〉을 표본의 크기($n$)로 나눈 것과 같다.

③ 공식 : $\sigma(\overline{X}) = \dfrac{\sigma(X)}{n} = \dfrac{\sigma}{\sqrt{n}}$

의미 : 〈$\overline{X}$의 표준편차〉는 〈$X$의 표준편차 즉, 모 표준편차〉를 $\sqrt{n}$으로 나눈 것과 같다.

## 2) 공식 설명 : $n$을 충분히 작은 수로 두어 경험적으로 안다.




① $X$의 평균, 분산을 구해본다.

| $X$ | 1 | 2 | 3 | 합계 |
|---|---|---|---|---|
| $P(X)$ | $\frac{1}{3}$ | $\frac{1}{3}$ | $\frac{1}{3}$ | 1 |

⇨ - 모평균 : $E(X) = 2$ - 모분산 : $V(X) = \dfrac{2}{3}$

② $\overline{X}$를 확률변수로 가지는 새로운 확률분포를 만들고 평균과 분산을 구해본다.

| $\overline{X}$ | 1 | $\frac{3}{2}$ | 2 | $\frac{5}{2}$ | 3 | 합계 |
|---|---|---|---|---|---|---|
| $P(\overline{X})$ | $\frac{1}{9}$ | $\frac{2}{9}$ | $\frac{3}{9}$ | $\frac{2}{9}$ | $\frac{1}{9}$ | 1 |

⇨ - 표본평균의 평균 : $E(\overline{X}) = 2$ - 표본평균의 분산 : $V(\overline{X}) = \dfrac{1}{3}$

결국 ⇨ 표본평균의 평균 = 모평균, 표본평균의 분산 = $\dfrac{\text{모분산}}{\text{표본의 크기}}$ 임이 확인되었다.

⇨ 이 공식만 보면 〈모집단의 자료가 너무 많아 추정〉한다는 의미와는 반대로 〈모집단의 자료〉를 조사하는 것이 훨씬 간단하다.

⇨ 수학적으로 일반화된 증명을 하는 것도 어렵지는 않지만 위의 사실을 직관적으로 받아들이고 이해하면 그만이다. 교과서도 일반화된 증명을 하지 않고 있다. 추정의 핵심이 아니기 때문이다.

## 3) 공식을 사용하는 근거

(1) <복원 추출>하는 경우 : 표본의 크기 $n$에 관계없이 $E(\overline{X}) = m$과 $V(\overline{X}) = \frac{\sigma^2}{n}$이 성립한다.

⇨ 모집단의 크기(모집단의 자료의 개수)와 표본의 크기(뽑는 자료의 개수)가 작다면 반드시 <복원추출>이라는 조건을 줘야 한다. <복원추출>이라는 조건이 없으면 <모집단의 정보>를 통해 <$\overline{X}$의 정보>를 알아내는 추정문제가 아니다.

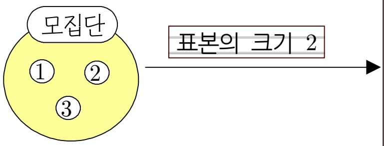
(2) 비 복원 추출 하는 경우 : 모집단의 크기가 충분히 크면 복원 추출하는 경우와 근사적으로 같다.

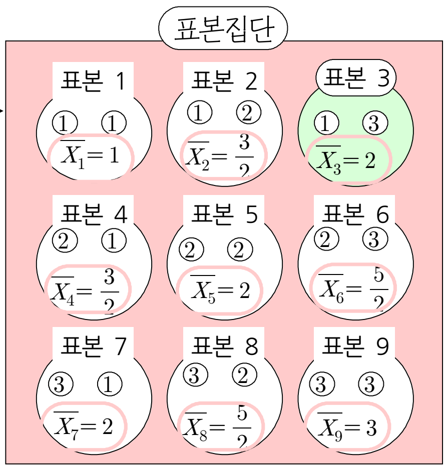
⇨ 결국 <복원, 비복원>에 대한 언급이 없어도 $E(\overline{X}) = m$과 $V(\overline{X}) = \frac{\sigma^2}{n}$의 식이 성립한다.

사실 그냥 생각해봐도 <5천만명>중에 표본의 크기가 10인 표본을 추출한다면 <1명씩 뽑아 조사할 때> 조사한 사람을 조사할 대상에 다시 포함시키든 포함시키지 않든 결과에 미치는 영향은 미미할 것이다. 즉, 앞으로 굳이 <복원추출>하지 않아도 상관없는 경우에는 <임의추출>이라는 말만 나와도 위의 공식을 쓰는 것에 무리가 없다.

코칭 앞에서 보는 바와 같이 $\overline{X}$의 집단(표본집단)이 $X$의 집단(모집단)보다 훨씬 더 많은 자료를 가지고 있다.
즉, <$\overline{X}$의 평균(표본평균의 평균)>을 구하는 것은 <$X$의 평균(모평균)>을 구하는 것보다 계산상 훨씬 복잡하다.
그렇기 때문에 <$\overline{X}$의 평균>을 구하는 문제는 진짜 <$\overline{X}$의 분포>를 구해서 평균을 구하는 대신 <모평균>을 구하는 것이 합리적이다. 또한 <$\overline{X}$의 분산(표본평균의 분산)>을 구하는 문제 역시 진짜 <$\overline{X}$의 분포>를 구해서 분산을 구하는 대신 <모분산>을 구해서 표본의 크기 $n$으로 나누는 것이 합리적이다.

코칭 용어 재정립 **(각 집단의 색과 설명박스의 색이 같도록 표현하였다. 반드시 같은 색끼리 함께 확인하라.)**




이 집단에서
<평균, 분산, 표준편차>를
<표본평균, 표본분산, 표본표준편차>
라 하고, 다음과 같은 기호를 쓴다.
- 표본평균 : $\overline{X}$
- 표본분산 : $s^2$
- 표본표준편차 : $s$

이 집단에서
<평균, 분산, 표준편차>를
<모평균, 모분산, 모표준편차>
라 하고, 다음과 같은 기호를 쓴다.
- 모평균 : $E(X) = m$
- 모분산 : $V(X) = \sigma^2$
- 모표준편차 : $\sigma(X) = \sigma$

이 집단($\overline{X}$를 확률변수로 갖는 집단)에서 <평균, 분산, 표준편차>를
<표본평균의 평균, 표본평균의 분산, 표본평균의 표준편차>라 하고, 다음과 같은 기호를 쓴다.
- 표본평균의 평균 : $E(\overline{X})$ ⇨ (모평균과의 관계) $E(\overline{X}) = E(X) =$ 모평균
- 표본평균의 분산 : $V(\overline{X})$ ⇨ (모분산과의 관계) $V(\overline{X}) = \frac{\sigma^2}{n} = \frac{\text{모분산}}{n}$
- 표본평균의 표준편차 : $\sigma(\overline{X})$ ⇨ (모표준편차와의 관계) $\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}} = \frac{\text{모표준편차}}{\sqrt{n}}$

- 후에 <추정>을 할 때 부분의 $s^2$(표본분산)과 $s$(표본표준편차)로 부분의 $\sigma^2$(모분산)과 $\sigma$(모표준편차)를 대체하기도 한다. 이것은 자연스럽게 받아들이는 것이고 정밀한 증명은 필요하지 않다. **(용어학습이 완전해져야 한다.)**

# 3. $\overline{X}$의 분포

## 1) 이산확률분포

- 모집단의 자료의 개수도 충분히 작고 (3개~4개) 표본의 크기도 충분히 작다면 (2~3개) $\overline{X}$는 정규분포를 따르지 않는다.
  이때 <크기가 $n$인 표본을 복원추출>한다는 조건이 반드시 주어진다. 이 말속에는 <임의로 1개씩 n번 뽑는 복원추출>이라는 뜻이 숨겨져 있는데 이 말 뜻만 이해하면 <독립인 사건의 곱셈정리>문제 그 이상도 이하도 아니다.
  실제로 요즘은 <임의로 1개를 뽑아 확인하고 넣는 시행을 n번 반복>이라고 상황을 구체적으로 제시하는데 이렇게 되면 추정을 전혀 학습하지 않은 학습자도 충분히 맞출 수 있는 매우 쉬운 문제가 되어버린다.

- 이 문제에서도 <$\overline{X}$의 평균, 표준편차>는 <모평균, 모표준편차>를 통해서 구한다. 물론 <복원추출>을 반드시 문제에서 명시할 것이다.

## 2) 정규분포

- 사실 <모집단>이 어쨌든 자료의 개수를 셀 수 있는 집단이므로 이 중 몇 개를 뽑아 만든 <표본의 평균>들이 모인 집단도 자료의 개수를 셀 수 있는 <이산적인 집단>임이 분명하다.
  하지만 <Part5의 정규분포 - 개념의 외연>에서 설명했듯이 - 라플라스(P.-S. Laplace, 1749-1827)에 의하여 이항분포가 아닌 확률분포도 정규분포를 따름은 이미 밝혀진 사실이다.

(1) 모집단이 정규분포를 따를 때 : 표본의 크기 $n$에 상관없이 $\overline{X}$의 분포는 정규분포 $N\left(m,\left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$을 따른다.

(2) 모집단이 정규분포를 따르지 않을 때

: 표본의 크기 $n$이 충분히 크면 $\overline{X}$의 분포는 근사적으로 정규분포 $N\left(m,\left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$을 따른다.

(n은 충분히 크거나 충분히 작은 경우만 나오므로 염려할 필요 없다. 예제를 보면 별 고민할 필요 없다는 사실을 알게 된다.)

⇨ 결국 <모집단의 분포>에 대한 언급이 없어도 $\overline{X}$의 분포는 정규분포 $N\left(m,\left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$라고 생각하고 문제를 푼다.

새로운 $\overline{X}$ 확률분포인 $N\left(m,\left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$를 만들었다면 정규분포 문제라고 생각해도 좋다.

티칭 $\overline{X}$에서 성질문제 : 표본 안에 있는 $n$개의 자료의 총합 ⇨ $n\overline{X}$

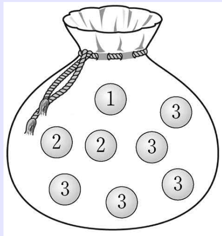
$\overline{X}$가 원래 평균이었다는 사실을 잊으면 안 된다. 우리 반이 $30$명인데 평균점수가 $60$점이라고 하자. 우리 반의 모든 학생들의 총 점수 합은 얼마인가? 당연히 $30\times60$이다.

즉, 표본의 크기가 $n$이고, 표본의 평균이 $\overline{X}$일 때, 표본 안에 있는 $n$개의 자료의 총합 $=n\overline{X}$이다.

(증명) $\frac{x_1+x_2+\cdots+x_n}{n}=\overline{X}$이라 하면 $x_1+x_2+\cdots+x_n=n\times\overline{X}$

**140** 이해를 위한 예제 2014. 11. B형(85%). 18번. 4점

주머니 속에 1의 숫자가 적혀 있는 공 1개, 2의 숫자가 적혀 있는 공 2개, 3의 숫자가 적혀 있는 공 5개가 들어 있다. 이 주머니에서 임의로 1개의 공을 꺼내어 공에 적혀 있는 수를 확인한 후 다시 넣는다. 이와 같은 시행을 2번 반복할 때, 꺼낸 공에 적혀 있는 수의 평균을 $\overline{X}$라 하자. $\mathrm{P}(\overline{X}=2)$의 값은?



① $\frac{5}{32}$ ② $\frac{11}{64}$ ③ $\frac{3}{16}$ ④ $\frac{13}{64}$ ⑤ $\frac{7}{32}$

시행횟수가 적은 경우 <독립인 사건의 곱셈정리> 그 이상도 이하도 아니다.

- 모집단을 확률분포 표로 표현할 경우 다음과 같이 표현된다.

| $X$ | 1 | 2 | 3 | 합계 |
|---|---|---|---|---|
| $P(X=x)$ | $\frac{1}{8}$ | $\frac{2}{8}$ | $\frac{5}{8}$ | 1 |

조건 이 주머니에서 임의로 1개의 공을 꺼내어 공에 적혀 있는 수를 확인한 후 다시 넣는다. 이와 같은 시행을 2번 반복할 때,

⇨ 문제에서 복원시행의 상황을 말로 풀어 알려주었다. 모집단의 자료의 개수가 3개, 표본의 크기가 2이므로 $\overline{X}$의 분포는 〈이산확률분포〉를 따른다. 즉, $\overline{X}$의 분포를 표로 나타낼 경우 (문제와 상관없이 표를 완성시켜 보았다.)

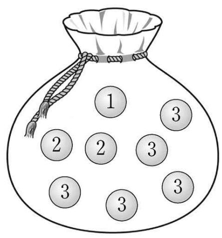
| $\overline{X}$ | 1 | $\frac{3}{2}$ | 2 | $\frac{5}{2}$ | 3 | 합계 |
|---|---|---|---|---|---|---|
| $P(\overline{X}=x)$ | $\frac{1}{64}$ | $\frac{4}{64}$ | $\frac{14}{64}$ | $\frac{20}{64}$ | $\frac{25}{64}$ | 1 |

- 위의 표에서 부분에 들어갈 확률 값은 뽑은 두 수의 평균이 2일 확률이므로 $(1\to3)$, $(3\to1)$, $(2\to2)$ 세 가지가 있다. 즉, $\mathrm{P}(\overline{X}=2) = \frac{1}{8}\cdot\frac{5}{8} + \frac{5}{8}\cdot\frac{1}{8} + \frac{2}{8}\cdot\frac{2}{8} = \frac{5+5+4}{64} = \frac{7}{32}$
- 이 문제는 〈복원시행〉이라는 말에 함의된 내용을 〈말로 풀어 설명〉했기 때문에 추정을 아예 몰라도 푸는데 지장이 없다.

정답 ⑤

티칭 이 문제의 상황을 〈표본의 크기가 2인 표본을 복원추출〉한다고 표현해도 완전히 같은 문제이다.
단지 〈복원추출〉의 의미를 모르는 학생에게는 〈한 번에 2개를 뽑고 다시 넣는다.〉라고 인지할 가능성이 매우 크기 때문에 이 부분을 정확하게 기억 해야 한다.

티칭 2014. 11. B형(85%). 18번. 4점 - 표현 살짝 바꾸기
주머니 속에 1의 숫자가 적혀 있는 공 1개, 2의 숫자가 적혀 있는 공 2개, 3의 숫자가 적혀 있는 공 5개가 들어 있다. 이 주머니에서 크기가 2인 표본을 복원추출 할 때, 꺼낸 공에 적혀 있는 수의 평균을 $\overline{X}$라 하자. $\mathrm{P}(\overline{X}=2)$의 값은?



사건의 분석 한 개씩 2번 뽑아서 나온 수의 평균이 2가 된다는 의미만 알면 매우 쉬운 문제이다.
두 수의 평균이 2가 되려면 $(1\to3)$ 또는 $(3\to1)$ 또는 $(2\to2)$이 되어야 한다.

⇨ 즉, 확률의 곱셈정리와 덧셈정리에 의하여

$\mathrm{P}(\overline{X}=2) = \frac{1}{8}\cdot\frac{5}{8} + \frac{5}{8}\cdot\frac{1}{8} + \frac{2}{8}\cdot\frac{2}{8} = \frac{5+5+4}{64} = \frac{7}{32}$ (추정을 몰라도 문제를 푸는데 지장이 없다.)

이처럼 〈표본의 크기가 2인 표본을 복원추출〉이라는 말의 의미만 알면 사실 〈추정〉의 개념을 생각할 필요도 없다.

**141** 이해를 위한 예제 2010. 9. 나형(70%). 29번. 4점

다음은 어느 모집단의 확률분포표이다. 이 모집단에서 크기가 16인 표본을 임의 추출할 때, 표본평균 $\overline{X}$의 표준편차는? (단, $a$는 상수이다.)

| $X$ | $-2$ | $0$ | $1$ | 합계 |
|---|---|---|---|---|
| $P(X=x)$ | $\frac{1}{4}$ | $a$ | $\frac{1}{2}$ | $1$ |

① $\frac{\sqrt{6}}{8}$ ② $\frac{\sqrt{6}}{6}$ ③ $\frac{\sqrt{6}}{4}$ ④ $\frac{\sqrt{6}}{2}$ ⑤ $\sqrt{6}$

- 공식 $E(\overline{X}) = m$, $V(\overline{X}) = \frac{\sigma^2}{n}$, $\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$이 성립한다고 가정하고 푼다.

 사실은 <복원시행>을 해야만 같지만 이 말을 생략한다고 하더라도 고등과정에서는 $\overline{X}$가 등장하는 순간 항상 위의 공식을 사용할 수 있다는 가정이 깔린다고 생각해도 좋다.

⇨ <$\overline{X}$의 표준편차>는 <모 표준편차>를 구해서 $\sqrt{n}$으로 나누는 방식으로 구한다.
($\overline{X}$의 분포는 모집단인 $X$의 분포보다 훨씬 많고 복잡하기 때문이다.)

- 주어진 표에서 확률의 총합이 $1$이므로 일단 $a$를 구하면 $\frac{1}{4}$이다. 이제 표를 완성하면 다음과 같다.

| $X$ | $-2$ | $0$ | $1$ | 합계 |
|---|---|---|---|---|
| $P(X=x)$ | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{2}$ | $1$ |

조건 크기가 16인 표본을 임의 추출할 때, 표본평균 $\overline{X}$의 표준편차는 ⇨ $n = 16$이므로 $\sigma(\overline{X}) = \frac{\sigma(X)}{\sqrt{16}} = \frac{\sigma(X)}{4}$

- 즉, 모 표준편차를 구하면

⇨ $E(X^2) = (-2)^2 \times \frac{1}{4} + 0^2 \times \frac{1}{4} + 1^2 \times \frac{1}{2} = \frac{3}{2}$이고, $E(X) = (-2) \times \frac{1}{4} + 0 \times \frac{1}{4} + 1 \times \frac{1}{2} = 0$이므로

즉, $\sigma(X) = \sqrt{V(X)} = \sqrt{E(X^2) - E(X)^2} = \sqrt{\frac{3}{2}} = \frac{\sqrt{6}}{2}$

$\sigma(\overline{X}) = \frac{\sigma(X)}{4} = \frac{\sqrt{6}}{8}$이므로 정답은 ①

**142** 이해를 위한 예제 2008. 11. 나형(28%). 29번. 4점

다음은 어떤 모집단의 확률분포표이다. 이 모집단에서 크기가 2인 표본을 복원 추출하여 구한 표본평균을 $\overline{X}$라 하자. $\overline{X}$의 평균이 18일 때, $P(\overline{X}=20)$의 값은?

| $X$ | 10 | 20 | 30 | 계 |
|---|---|---|---|---|
| $P(X=x)$ | $\frac{1}{2}$ | $a$ | $\frac{1}{2}-a$ | 1 |

① $\frac{2}{5}$ ② $\frac{19}{50}$ ③ $\frac{9}{25}$ ④ $\frac{17}{50}$ ⑤ $\frac{8}{25}$

조건 $\overline{X}$의 평균이 18일 때, ⇨ $E(\overline{X}) = E(X)$이므로 결국 $E(X)=18$이라는 조건을 준 것이나 마찬가지이다.

즉, $\sum x_i p_i = 10\times\frac{1}{2}+20\times a+\left(\frac{1}{2}-a\right) = 20-10a = 18$이므로 $a = \frac{1}{5}$이다.

이것을 바탕으로 표를 완성하면 다음과 같다.

| $X$ | 10 | 20 | 30 | 계 |
|---|---|---|---|---|
| $P(X=x)$ | $\frac{1}{2}$ | $\frac{1}{5}$ | $\frac{3}{10}$ | 1 |

조건 크기가 2인 표본을 복원 추출 ⇨ 이 말 뜻만 이해하면 매우 쉬운 문제가 된다. 즉, <1개씩 2번 뽑는 복원추출>을 하여 $P(\overline{X}=20)$을 구한다. 두 수의 평균이 20이 되려면 $(10\rightarrow30)$ 또는 $(30\rightarrow10)$ 또는 $(20\rightarrow20)$이 되어야 한다.

즉, $P(\overline{X}=20) = \frac{1}{2}\cdot\frac{3}{10}+\frac{3}{10}\cdot\frac{1}{2}+\frac{1}{5}\cdot\frac{1}{5} = \frac{17}{50}$이므로 정답은 ④

코칭 이 문제가 정답률이 낮은 이유는 <1개씩 2번 뽑는다는 표현>대신 <크기가 2인 표본을 복원 추출>이라는 표현으로 주었기 때문이다. 말의 뜻만 안다면 이 문제 역시 <독립인 사건의 곱셈정리> 그 이상도 이하도 아니다.

코칭 <$\overline{X}$의 평균(표본평균의 평균)>, <$\overline{X}$의 분산(표본평균의 분산)>, <$\overline{X}$의 표준편차(표본평균의 표준편차)>를 직접 구하는 경우는 없다. <모평균, 분산, 표준편차>를 직접 구한 후 이를 공식에 대입해서 얻어지는 결과라는 것을 명심 또 명심하자.

**143** 이해를 위한 예제

확률변수 $X$가 정규분포 $N(60,10^2)$을 따르는 모집단에서 크기 20인 표본을 임의추출 할 때, 표본평균 $\overline{X}$가 이루는 분포를 구하여라.

모집단의 크기와 표본의 크기가 매우 작은 경우를 제외하면 <$\overline{X}$의 분포>는 항상 <정규분포>를 따른다.

공식에 의하여 $E(\overline{X}) = 60$이고 $V(\overline{X}) = \frac{100}{20} = 5$이므로 정규분포 $N(60,\sqrt{5}^{\,2})$를 따른다.

정답은 $N(60,\sqrt{5}^{\,2})$

**144** 이해를 위한 예제 2008. 11. 가형(87%), 나형(69%). 8번. 3점

세계핸드볼연맹에서 공인한 여자 일반부용 핸드볼 공을 생산하는 회사가 있다. 이 회사 에서 생산된 핸드볼 공의 무게는 평균 $350\mathrm{g}$, 표준편차 $16\mathrm{g}$인 정규분포를 따른다고 한다. 이 회사는 일정한 기간 동안 생산된 핸드볼 공 중에서 임의로 추출된 핸드볼 공 64개의 무게의 평균이 $346\mathrm{g}$ 이하이거나 $355\mathrm{g}$ 이상이면 생산 공정에 문제가 있다고 판단한다. 이 회사에서 생산 공정에 문제가 있다고 판단할 확률을 아래 표준정규분포 표를 이용하여 구한 것은?

| $z$ | $\mathrm{P}(0 \le Z \le z)$ |
|---|---|
| 2.00 | 0.4772 |
| 2.25 | 0.4878 |
| 2.50 | 0.4938 |
| 2.75 | 0.4970 |

① 0.0290 ② 0.0258 ③ 0.0184 ④ 0.0152 ⑤ 0.0092

결국 정규분포

조건 핸드볼 공의 무게는 평균 $350\mathrm{g}$, 표준편차 $16\mathrm{g}$인 정규분포

- 핸드볼의 무게를 확률변수 $X$라 하면 $X$는 $N(350, 16^2)$인 정규분포를 따른다.

조건 임의로 추출된 핸드볼 공 64개의 무게의 평균

- 공식에 따라 $E(\overline{X}) = 350$이고, $\sigma(\overline{X}) = \dfrac{16}{\sqrt{64}} = 2$이므로 $\overline{X}$는 $N(350, 2^2)$인 정규분포를 따른다.

구할 것 무게의 평균이 $346\mathrm{g}$ 이하이거나 $355\mathrm{g}$ 이상이면 생산 공정에 문제가 있다고 판단할 확률

- $P(\text{문제가 있다고 판단}) = P(\overline{X} \le 346) + P(\overline{X} \ge 355)$

$$= P\left(\frac{\overline{X}-350}{2} \le \frac{346-350}{2}\right) + P\left(\frac{\overline{X}-350}{2} \ge \frac{355-350}{2}\right) = P(Z \le -2) + P(Z \ge 2.5)$$

$$= 0.0228 + 0.0062 = 0.0290$$ (문제의 표 참조)

정답은 ①

**145** 이해를 위한 예제 2010. 11. 나형(64%). 27번. 3점

어느 도시에서 공용 자전거의 1회 이용 시간은 평균이 60분, 표준편차가 10분인 정규분포를 따른다고 한다. 공용 자전거를 이용한 25회를 임의 추출하여 조사할 때, 25회 이용 시간의 총합이 1450분 이상일 확률을 다음 표준정규분포 표를 이용하여 구한 것은?

| $z$ | $P(0 \le Z \le z)$ |
|---|---|
| 1.0 | 0.3413 |
| 1.5 | 0.4332 |
| 2.0 | 0.4772 |
| 2.5 | 0.4938 |

① 0.8351 ② 0.8413 ③ 0.9332 ④ 0.9772 ⑤ 0.9938

조건 공용 자전거의 1회 이용 시간은 평균이 60분, 표준편차가 10분인 정규분포

- 공용 자전거 1회 이용시간을 확률변수 $X$라고 하면 $X$는 $N(60, 10^2)$인 정규분포를 따른다.

조건 공용 자전거를 이용한 25회를 임의 추출하여 조사할 때,

- 이용시간의 평균을 $\overline{X}$라고 할 때, 공식에 따라 $E(\overline{X}) = 60$이고, $\sigma(\overline{X}) = \dfrac{10}{\sqrt{25}} = 2$이므로 $\overline{X}$는 $N(60, 2^2)$인 정규분포를 따른다.

구할 것 25회 이용 시간의 총합이 1450분 이상일 확률

- 〈25회의 이용시간의 총합〉은 〈평균 $\times 25$〉라고 할 수 있으므로 $25\overline{X}$라고 생각할 수 있다.

⇨ $P(25\overline{X} \ge 1450) = P(\overline{X} \ge 58) = P\left(\dfrac{\overline{X}-60}{2} \ge \dfrac{58-60}{2}\right) = P(Z \ge -1) = 0.3413 + 0.5 = 0.8413$

정답은 ②

## #2 모평균의 추정

# 1. 모평균의 추정 식의 도출

### 1) 1단계 : 신뢰계수와 신뢰도

- 새로운 확률변수 $\overline{X}$가 새로운 정규분포 $N\left(m,\ \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$을 따른다는 사실을 알고 있고,

정규분포인 $\overline{X}$의 확률을 $P\left(m-k\frac{\sigma}{\sqrt{n}}\ \le\ \overline{X}\ \le\ m+k\frac{\sigma}{\sqrt{n}}\right)$와 같이 표현했을 때, 확률 값은 $k$에만 영향을 받는다는 사실은 이미 정규분포에서 학습한 바 있다.

$P(|Z|\le 1.96)\ =\ 0.95$이고, $P(|Z|\le 2.58)\ =\ 0.99$이므로

(문제의 조건으로 주어지며 표준정규분포 표를 찾으면 쉽게 알 수 있는 내용이기도 하다.)

⇨ $k\ =\ 1.96$인 경우 $P\left(m-1.96\frac{\sigma}{\sqrt{n}}\ \le\ \overline{X}\ \le\ m+1.96\frac{\sigma}{\sqrt{n}}\right)\ =\ 0.95$

⇨ $k\ =\ 2.58$인 경우 $P\left(m-2.58\frac{\sigma}{\sqrt{n}}\ \le\ \overline{X}\ \le\ m+2.58\frac{\sigma}{\sqrt{n}}\right)\ =\ 0.99$

- 이때 $k$값을 <신뢰계수>라 하고 그에 따른 확률 값을 백분율로 표현한 것을 <신뢰도>라고 한다.
즉, 신뢰계수가 $1.96$인 경우 신뢰도는 $95\%$로 정해져 있고,
신뢰계수가 $2.58$인 경우 신뢰도는 $99\%$로 정해져 있다. 신뢰도는 오직 신뢰계수에만 영향을 받는다.

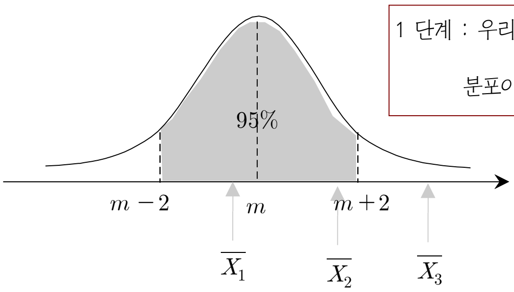
### 2) 2단계 : 식의 표현 바꾸기

$$P\left(m-1.96\frac{\sigma}{\sqrt{n}}\ \le\ \overline{X}\ \le\ m+1.96\frac{\sigma}{\sqrt{n}}\right)\ =\ 0.95\quad\Leftrightarrow\quad P\left(|\overline{X}-m|\ \le\ 1.96\frac{\sigma}{\sqrt{n}}\right)\ =\ 0.95$$

- 와 같이 식을 변형하면 ⇨ 표본평균($\overline{X}$)과 모평균의 차이($m$)가 $1.96\frac{\sigma}{\sqrt{n}}$ 이하일 확률이 $0.95$라고 읽을 수 있다.

- 조금 더 실생활적으로 말을 바꾸어 표현하면 ⇨ 표본평균과 모평균의 차이의 차이가 $1.96\frac{\sigma}{\sqrt{n}}$이하일 가능성은 $95\%$라고 읽을 수 있다.

- 구체적인 상황을 설정하여 그래프 상에서의 느낌을 확인해보자.
어떤 모집단의 표준편차가 알려져 있고 표본의 크기가 $100$인 표본을 임의 추출했다고 가정해보자.
이때, $\sigma$는 어차피 문제에서 정해지는 값이므로 $1.96\frac{\sigma}{\sqrt{100}}$의 계산된 결과가 $2$라고 가정해보자.




1 단계 : 우리가 이론상 가능한 $\overline{X}$를 모두 조사했다고 가정하여 수학적으로 정해진 $\overline{X}$의 분포이다. 이 분포는 평균이 $m$(모평균)이고, 표준편차가 $\frac{\sigma}{\sqrt{100}}$인 정규분포이다.

$P(|\overline{X}-m|\ \le\ 2)\ =\ 0.95$

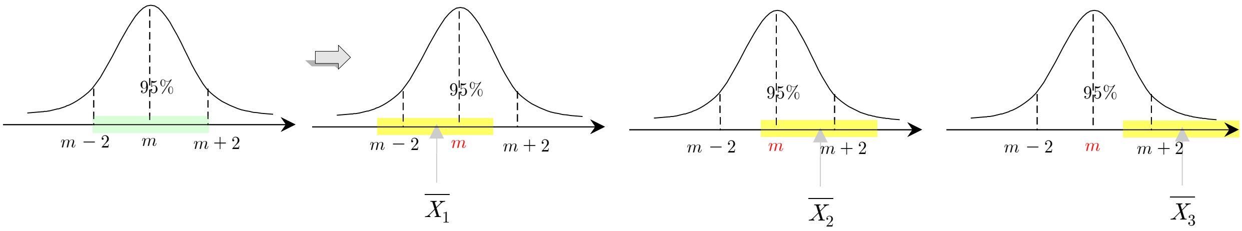
① $\overline{X_1}$, $\overline{X_2}$와 같이 표본을 딱 1개만 조사할 경우 그 **표본평균과 모평균의 차이**가 왼쪽 그림과 같이 2이하가 될 가능성이 $95\%$이다.

② 물론 $\overline{X_3}$처럼 우리가 조사한 표본평균과 모평균의 차이가 2이상일 가능성도 $5\%$는 된다.

## 3) 3단계 : 최종식

- 최종식은 $\overline{X}$의 입장에서 식을 다시 바라본 것이다.

$P\left(\overline{X}-2 \le m \le \overline{X}+2\right) = 0.95$ ⇨ $\overline{X}$를 기준으로 좌우로 길이가 2인 띠를 만든다.




① $\overline{X_1}$, $\overline{X_2}$와 같이 **표본을 딱 1개만 조사**할 경우 그 <표본평균>의 좌우로 2인 구간 안에 모평균이 존재할 가능성이 95%이다.

② 물론 $\overline{X_3}$처럼 우리가 조사한 <표본평균>의 좌우로 2인 구간 안에 모평균이 존재하지 않을 가능성도 5%는 된다.

- 위의 결론에 대하여 수학적인 기호를 제하고 일반적인 용어를 통해서 재정립한다. 통계치를 수학적 용어로 대중에게 전달할 수는 없는 일이다.

신뢰도 $\alpha$%의 추정 식 : $\overline{X}-k\dfrac{\sigma}{\sqrt{n}} \le m \le \overline{X}+k\dfrac{\sigma}{\sqrt{n}}$ (단, $P(|Z|\le z) = \dfrac{\alpha}{100}$)

① 표본평균 $(\overline{X})$ : 물론 이 식을 유도하는 과정까지는 <확률변수>의 역할을 했지만 위의 식을 실제로 사용하는 과정에서는 <상수>의 역할을 한다. 이제는 <표본은 1개만 조사>할 것이고, 이 한 개의 표본의 <표본평균>을 기준으로 좌우로 $k\dfrac{\sigma}{\sqrt{n}}$만큼 떨어진 곳 사이에 모평균이 존재할 것이라고 말할 수 있다.

② 신뢰계수($k$)와 신뢰도($\alpha$%) : $\alpha$는 **확률의 의미**를 가지므로 $k$값에만 영향을 받는다.
즉, $k$값이 결정됨에 따라 ①의 결론이 **<얼마만큼 믿을 만한지>** 알 수 있다.

- $k=1.96$인 경우 $\alpha=95$ ⇨ 보통 표준정규분포를 활용해서 $P(|Z|\le 1.96) = 0.95$처럼 주므로 외울 필요는 없다
- $k=2.58$인 경우 $\alpha=99$ ⇨ 보통 표준정규분포를 활용해서 $P(|Z|\le 2.58) = 0.99$처럼 주므로 외울 필요는 없다

③ 신뢰구간 $\left[\overline{X}-k\dfrac{\sigma}{\sqrt{n}},\ \overline{X}+k\dfrac{\sigma}{\sqrt{n}}\right]$ : 모평균이 존재할 것이라고 예측되는 구간이다. 위의 그래프에서 에 해당하는 구간을 식으로 표현한 것뿐이다. (물론 $\overline{X_3}$와 같이 신뢰구간 안에 모평균이 존재하지 않을 가능성도 존재하고 이 가능성은 신뢰계수에 의해 결정된다.)

④ 신뢰구간의 길이 $2k\dfrac{\sigma}{\sqrt{n}}$ : 위의 신뢰구간에서 $\left(\overline{X}+k\dfrac{\sigma}{\sqrt{n}}\right)-\left(\overline{X}-k\dfrac{\sigma}{\sqrt{n}}\right) = 2k\dfrac{\sigma}{\sqrt{n}}$로 계산된 것이다.
이것은 <정확도>를 측정하는 기준이 된다. 위의 그래프에서 의 길이를 의미한다.
의 길이가 짧을수록 <정확한 추정>이 된다.

- 예를 들어 어느 지역 몸무게의 평균이 <65와 69사이>라고 말하는 것보다 <66과 68사이>라고 말하는 것이 모평균을 조금 더 <정확하게 예측했다>라고 말할 수 있을 것이다.

⑤ 표본오차, 표본오차의 최댓값 : 말 그대로 우리가 뽑은 <한 개의 표본의 평균>과 <모평균>의 차이가 오차이다.
추정치를 신뢰한다고 가정했을 때 (물론 틀릴 가능성도 있지만 확률적으로 작기 때문에)
표본오차의 범위는 $\left|\overline{X}-m\right| \le k\dfrac{\sigma}{\sqrt{n}}$이고 표본오차의 최댓값은 당연히 $k\dfrac{\sigma}{\sqrt{n}}$이다.

⑥ $\sigma$ : 원래는 <모 표준편차 : $\sigma$>의 의미이지만 우리가 표본조사를 1회 실시하여 얻은 $n$개의 자료에서 구한 표준편차인 <표본표준편차 : $s$>로 **대신할 수 있다.** 문제에서 모 표준편차가 주어지는 경우도 있고, 표본표준편차가 주어지는 경우도 있다. (모평균을 몰라서 추정하는데 어떻게 모 표준편차를 알 수 있느냐고 반문하는 학생들도 꽤 많은데 실생활에서도 모평균은 모르는데 모 표준편차는 경험상 대략적으로 알 수 있는 경우가 많다. 사실 수학적으로 오차분포라는 내용이 있다. 즉, 표준편차도 추정할 수 있다는 뜻이다. - 이해할 필요 없다. 넘어가라.)

<표본평균의 표준편차 : $\sigma(\overline{X}) = \dfrac{\sigma}{\sqrt{n}}$>와는 구분해야 하는 다른 개념이다.

⑦ $n$ : 표본의 크기이고 문제에서 주어진 대로 대입한다.

티칭 정확한 추정을 하는 법

- 신뢰도 $95\%$의 추정 식 : $\overline{X}-1.96\dfrac{\sigma}{\sqrt{n}} \le m \le \overline{X}+1.96\dfrac{\sigma}{\sqrt{n}}$에 대해 생각해보자.

⇨ 정확한 추정을 위해 <신뢰구간의 길이>인 $2\times1.96\dfrac{\sigma}{\sqrt{n}}$를 작게 하면 된다. 그런데 $1.96$은 거의 건드리지 않는다.

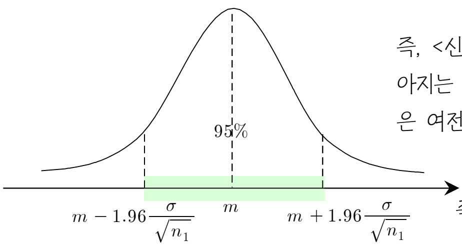
예를 들어 실제로 $1.96\dfrac{\sigma}{\sqrt{n}}$을 작게하기 위해서 $k=0.5$로 두고 계산하면 $P\left(|\overline{X}-m| \le 0.5\dfrac{\sigma}{\sqrt{n}}\right) = 0.383$이다.

이 말은 <신뢰구간의 길이>은 $2\times0.5\dfrac{\sigma}{\sqrt{n}}$로 짧아졌으나 그럴 가능성이 $38.3\%$밖에 안 된다는 뜻이다.

즉, 아무리 표본평균과 모평균이 가깝다고 해도 신뢰도가 낮으면 믿을 수가 없으므로 무가치한 추정이다.
그래서 실제로 $k$는 일상적으로 가장 효율적인 수치인 $k = 1.96$과 $k = 2.58$로 정해놓은 것이다.

⇨ $\sigma$는 우리가 정하는 값이 아니라 <조사결과 도출되는 값, 문제에서는 조건>일 뿐이기 때문에 결국 정확한 추정을 위해 우리가 할 수 있는 일은 $n$을 크게 하는 것뿐이다. 즉, 표본조사를 할 때 그 표본 안의 자료($n$)를 많을수록 정확한 추정이 된다. - 많이 조사하라는 뜻이다.

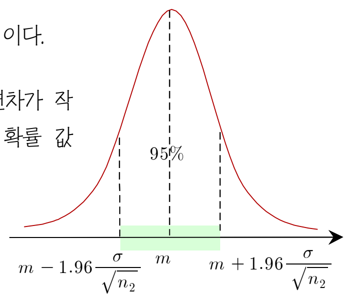
⇨ 그래프로 생각해보면 $\overline{X}$를 다시 확률변수로 생각했을 때 $n$을 크게 할 경우 $\sigma(\overline{X})$인 $\dfrac{\sigma}{\sqrt{n}}$가 작아지므로 가운데로 몰려있는 그래프가 될 것이고, 그래서 신뢰구간의 길이는 짧아지지만 신뢰도(확률값)에는 변함이 없게 된다.
아래 그림을 통해 다시 한 번 이해해본다.




이 그림과 같이 $n_1 < n_2$일 경우 $\dfrac{\sigma}{\sqrt{n_1}} > \dfrac{\sigma}{\sqrt{n_2}}$이다.

즉, <신뢰구간은 짧아지지만> $\overline{X}$의 입장에서는 표준편차가 작아지는 효과가 있어 그래프가 가운데로 몰리게 되어 확률 값은 여전히 $0.95$가 된다.

즉, 신뢰도의 변화없이 신뢰구간의 길이를 짧게 하고 싶다면 $n$을 크게 하는 수밖에 없다.




코칭 다음 용어들을 명확하게 구분하고 있다면 이 단원을 이해하는 데에 성공한 것.

⇨ 모평균 $VS$ 표본평균 $VS$ 표본평균의 평균

⇨ 모 표준편차 $VS$ 표본표준편차 $VS$ 표본평균의 표준편차

**146** 이해를 위한 예제 2011. 11. 가형(80%). 9번. 3점

어느 회사에서 생산하는 음료수 1병에 들어 있는 칼슘 함유량은 모평균이 $m$, 모 표준편차가 $\sigma$인 정규분포를 따른다고 한다. 이 회사에서 생산한 음료수 16병을 임의 추출하여 칼슘 함유량을 측정한 결과 표본평균이 12.34이었다. 이 회사에서 생산한 음료수 1병에 들어 있는 칼슘 함유량의 모평균 $m$에 대한 신뢰도 95%의 신뢰구간이 $11.36 \le m \le a$일 때, $a+\sigma$의 값은? (단, $Z$가 표준정규분포를 따를 때 $P(0 \le Z \le 1.96) = 0.4750$이고, 칼슘 함유량의 단위는 $mg$이다.)

① 14.32 ② 14.82 ③ 15.32 ④ 15.82 ⑤ 16.32

모 표준편차로 추정

조건 음료수 1병에 들어 있는 칼슘 함유량은 모평균이 $m$, 모 표준편차가 $\sigma$인 정규분포를 따른다.

⇨ 모집단에 대한 설명일 뿐이다.

조건 음료수 16병을 임의 추출하여 칼슘 함유량을 측정한 결과 표본평균이 12.34

⇨ 표본의 크기 $n = 16$, 표본 1개의 평균 $\overline{X} = 12.34$

조건 $P(0 \le Z \le 1.96) = 0.4750$ ⇨ 95%추정에서 신뢰계수 $k$는 $1.96$이라는 뜻이다.

조건 모평균 $m$에 대한 신뢰도 95%의 신뢰구간이 $11.36 \le m \le a$

⇨ 모평균을 추정하면 $12.34 - 1.96\dfrac{\sigma}{\sqrt{16}} \le m \le 12.34 + 1.96\dfrac{\sigma}{\sqrt{16}}$이고, 신뢰구간이 $11.36 \le m \le a$이므로

$$12.34 - 1.96\frac{\sigma}{\sqrt{16}} = 11.36 \quad \Leftrightarrow \quad 1.96 \times \frac{\sigma}{4} = 0.98 \quad \Leftrightarrow \quad \sigma = 2\text{이고}$$

$$a = 12.34 + 1.96\frac{\sigma}{\sqrt{16}} = 12.34 + 1.96\frac{2}{\sqrt{16}} = 12.34 + 0.98 = 13.32\text{이다.}$$

즉, $a+\sigma = 15.32$이므로 정답은 ③

**147** 이해를 위한 예제 2006. 11. 가형(69%), 나형(57%). 10번. 3점

어느 공장에서 생산되는 탁구공을 일정한 높이에서 강철 바닥에 떨어뜨렸을 때 탁구공이 튀어 오른 높이는 정규분포를 따른다고 한다. 이 공장에서 생산된 탁구공 중 임의 추출한 $100$개에 대하여 튀어 오른 높이를 측정하였더니 평균이 $245$, 표준편차가 $20$이었다. 이 공장에서 생산되는 탁구공 전체의 튀어 오른 높이의 평균에 대한 신뢰도 $95\%$의 신뢰구간에 속하는 정수의 개수는? (단, 높이의 단위는 $\mathrm{mm}$이고, $Z$가 표준정규분포를 따를 때 $P(0 \le Z \le 1.96) = 0.4750$이다.)

① 5 ② 6 ③ 7 ④ 8 ⑤ 9

표본표준편차로 추정

조건 탁구공이 튀어 오른 높이는 정규분포를 따른다고 한다. ⇨ 모집단에 대한 설명일 뿐이다

조건 임의 추출한 $100$개에 대하여 튀어 오른 높이를 측정하였더니 평균이 $245$, 표준편차가 $20$

⇨ 표본의 크기 $n = 100$, 표본 1개의 평균 $\overline{X} = 245$, 표본표준편차 $s = 20$

조건 $P(0 \le Z \le 1.96) = 0.4750$ ⇨ 95%추정에서 신뢰계수 $k$는 $1.96$이라는 뜻이다.

구할 것 신뢰도 $95\%$의 신뢰구간에 속하는 정수의 개수 - $95\%$의 신뢰도에서 모평균

⇨ 모평균을 추정하면 $245 - 1.96\dfrac{20}{\sqrt{100}} \le m \le 245 + 1.96\dfrac{20}{\sqrt{100}} \quad \Leftrightarrow \quad 241.08 \le m \le 248.92$ 이므로

$242$부터 $248$까지 정수의 개수가 정답이다.

즉, $248-242+1 = 7$이므로 정답은 ③

**148** 이해를 위한 예제 2010. 4. 가형(71%). 13번. 4점

분산이 $\sigma^2$인 정규분포를 따르는 모집단에서 크기 $n$인 표본을 임의 추출하여 모평균 $m$을 추정한 후 신뢰구간의 길이를 구하고자 한다. 아래 표준정규분포 표를 이용하여 구한 모평균 $m$에 대한 신뢰도 $79.6\%$의 신뢰구간의 길이가 $l$이고, 모평균 $m$에 대한 신뢰도 $\alpha\%$의 신뢰구간의 길이는 $2l$이다. 이때, $\alpha$의 값은?

| $z$ | $P(0 \le Z \le z)$ |
|---|---|
| 1.27 | 0.3980 |
| 1.69 | 0.4545 |
| 1.96 | 0.4750 |
| 2.54 | 0.4945 |
| 3.29 | 0.4995 |

① 87.3 ② 90.9 ③ 95.0 ④ 98.9 ⑤ 99.9

사건 분산이 $\sigma^2$인 정규분포를 따르는 모집단에서 크기 $n$인 표본을 임의 추출하여 모평균 $m$을 추정한 후 신뢰구간의 길이를 구하고자 한다. ⇨ 신뢰구간의 길이는 $2k\dfrac{\sigma}{\sqrt{n}}$이다.

사건 신뢰도 $79.6\%$의 신뢰구간의 길이가 $l$이고, 모평균 $m$에 대한 신뢰도 $\alpha\%$의 신뢰구간의 길이는 $2l$이다.

⇨ 위의 표에서 $P(-1.27 \le Z \le 1.27) = 0.7960$이므로 신뢰도 $79.6\%$에서 신뢰계수는 $1.27$이다.

즉, 신뢰구간의 길이는 $l = 2\times 1.27\dfrac{\sigma}{\sqrt{n}}$이다.

구할 것 모평균 $m$에 대한 신뢰도 $\alpha\%$의 신뢰구간의 길이는 $2l$

⇨ $n$와 $\sigma$가 고정된 상황에서 신뢰구간의 길이가 2배가 되려면 신뢰계수가 $2$배가 되어야 한다.

즉, 신뢰계수가 $2.54$인 경우에 신뢰구간의 길이가 $2l$이 된다.

이때, $P(-2.54 \le Z \le 2.54) = 0.9890$이므로 신뢰도는 $98.9\%$이다.

정답은 ④

## #3 표본비율의 이해

- 용어의 이해가 반이다. 수학적 내용을 익히기 전에 용어를 익히고 또 익힌다.

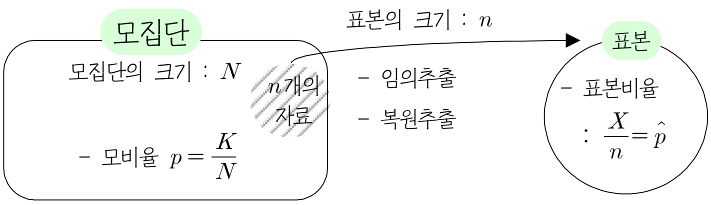
- <통계적 추정>을 위한 내용적 근거를 공부하는 부분이다. 내용상 <추정>보다는 <이항분포, 정규분포>문제에 가깝다.
  <$\overline{X}$의 평균>과는 다르게 <$\hat{p}$ 의 평균>은 일반적인 경우에 대한 증명이 가능하다. - 이항분포를 공부했다면 간단하다.

- <이항분포>에서 <평균, 분산, 표준편차>가 간단히 공식으로 구해졌듯이 <$\hat{p}$ 의 평균, 분산, 표준편차>도 간단히 공식으로 구해진다. <#3. 표본비율의 이해>는 <#1. 표본평균의 이해>와도 비교해야 하지만 <이항분포>와도 비교해야한다.

### 1. 표본비율($\hat{p}$)의 이해

#### 1) 이해 1단계 : 용어와 의미

<표본평균의 이해>에서 $X$가 우리 반 $50$명의 점수였다면 이제 $X$는 평균점수를 넘는 사람의 수이다.




① 확률변수 $X$, 상수 $K$ : 특정 성질을 만족하는 자료의 개수 $X$
⇨ 예컨대 모집단이 우리 반 50명이라면 상수 $K$는 우리 반 50명 중 평균을 넘는 사람의 수이다.
이때 $K$는 항상 정해지는 값으로 이것이 변수의 역할을 하는 경우는 없다.
⇨ 예컨대 50명 중 10명을 추출하여 조사했다면 확률변수 $X$는 뽑힌 10명 중 평균을 넘는 사람의 수이다.
$X$가 변수인 이유는 <1명씩 10번 뽑은 결과> $X$는 $0,\ 1,\ 2,\ \cdots,\ 10$까지 나올 수 있기 때문이다.

② 모집단의 크기 ($N$) : 모집단의 자료의 개수
⇨ 예컨대 모집단이 우리 반 50명이라면 모집단의 크기는 $50$이다. (즉, $N = 50$)

③ 추출(sampling) : 모집단의 자료 중 일부를 뽑는 것으로 이 자료를 표본이라고 한다.
⇨ 반드시 <임의 추출>해야 한다. 표본을 통해 모집단의 성질을 예측하는 것이 추정의 본질인데 <의도를 가지고 추출> 할 경우 우리의 원래 목적인 <모집단>과는 성질이 완전히 다른 <표본>이 나올 수 있기 때문이다.

④ 표본의 크기 ($n$) : 표본 1개 안에 있는 자료의 개수이다.
⇨ 예컨대 우리 반 50명 중 10명을 추출하여 조사한다고 할 때, 표본의 크기는 $10$이다. (즉, $n = 10$)

⑤ <표본의 크기>가 $n$인 표본을 <복원추출>
⇨ 예컨대 우리 반 50명 중 <표본의 크기가 $10$>인 표본을 <복원 추출>하여 조사한다고 할 때 <복원 추출로 한 번에 10명을 뽑는 것>이 아니라 <복원추출로 1명씩 10번 뽑아> 조사하는 것이다.
⇨ 이것이 전제가 되어야 <$\hat{p}$은 이항분포>를 따른다. - 이 내용은 잠시 후에 다루기로 한다. (뒷장에 ⑦까지 있습니다.)

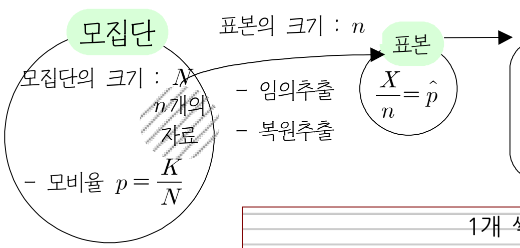
⑥ 모비율 $p=\dfrac{K}{N}$ : 모집단에서 특정한 성질을 만족하는 자료의 비율이다. 위에서도 언급했듯이 $K$는 상수이다.

⇨ 예컨대 모집단이 우리 반 $50$명이고, 평균을 넘는 사람의 수를 $20$명이라고 해보자. 이때 모비율은 $\dfrac{2}{5}$이다.

⑦ 표본비율 $\hat{p}=\dfrac{X}{n}$ : 우리가 조사한 $n$개의 자료 안에서 특정한 성질을 만족하는 자료의 비율이다.

⇨ 예컨대 모집단이 우리 반 $50$명이고, 여기에서 10명을 추출하여 조사한다고 할 때 (임의추출, 복원추출) 평균점수를 넘는 사람이 3명이었다면 표본비율은 $\hat{p}=\dfrac{3}{10}$이다. 이 경우 $X$는 변수로서 $0,\ 1,\ 2,\ \cdots,\ 10$으로 정해져 있고 표본비율 역시 변수로서 $\dfrac{0}{10},\ \dfrac{1}{10},\ \dfrac{2}{10},\ \cdots,\ \dfrac{10}{10}$으로 정해져 있다.

## 2) 이해 2단계




표본집단

| $\hat{p}$ | $\frac{0}{n}$ | $\frac{1}{n}$ | $\cdots$ | $\frac{n}{n}$ | 총합 |
|---|---|---|---|---|---|
| $P(\hat{p})$ | | | | | 1 |

1개 씩 n번 대상을 복원 추출하여 만들 수 있는
이론상 가능한 모든 표본을 가지고 <표본집단>을 만든다.
- <표본평균의 집단>에서 확률변수는 문제의 조건에 따라 상황에 맞게 계산해야 하지만
- <표본비율의 집단>은 무조건 위와 같이 확률변수가 $\dfrac{r}{n}$ $(r=0,\ 1,\ 2,\ \cdots)$로 정해진 <이산확률분포>를 따른다. 모집단의 크기 $N$과도 상관이 없다.
어차피 표본의 크기가 $n$이라면 표본비율은 $\dfrac{0}{n}$ 부터 $\dfrac{n}{n}$까지로 딱 정해져 있다.

- 표본 집단이란 이론상 가능한 모든 표본들을 모아 놓은 집단이다. <표본평균의 집단>과는 다르게 <표본비율의 집단>은 이론상 나올 수 있는 결과가 딱 정해져 있다.

- 새로운 확률변수 $\hat{p}$
: 위의 표와 같이 $\hat{p}$은 $\dfrac{0}{n}$ 부터 $\dfrac{n}{n}$까지로 딱 정해져 있는 <이산확률변수>이다.
사실 <복원시행>의 규칙으로 표본을 만들 수 있는 <경우의 수>는 엄청나게 많지만 그 <결과>가 전부 중복되어 <서로 다른 경우의 결과>로 나올 수 있는 모든 표본비율은 $\dfrac{0}{n}$ 부터 $\dfrac{n}{n}$까지로 정해진다.

## 2. $\hat{p}$의 평균(기댓값), 분산, 표준편차 (선 이해, 후 암기)

### 1) 공식의 이해와 증명 : 이항분포의 상황임을 쉽게 이해할 수 있다.

1단계 : <복원추출>의 규칙으로 표본의 크기가 $n$인 표본을 뽑는 표본조사에서 가능한 모든 표본비율은 $\frac{0}{n}, \frac{1}{n}, \cdots, \frac{n}{n}$ 이다. 이론상 $0$명이 <특정성질>을 만족할 수도 있고, $n$명 모두 <특정성질>을 만족할 수도 있다.

2단계 : 모비율은 곧 확률이다. (원래 비율과 확률은 같은 것이다.)
모비율이 $p$라고 하자. 모집단에서 $1$명을 뽑았을 때 <특정성질>을 만족할 확률이 $p$임을 이해한다.

⇨ 우리 반에 평균을 넘는 학생의 비율이 $\frac{2}{5}$라면 우리 반에서 임의로 한 명을 뽑았을 때 그 학생이 평균을 넘을 확률은 당연히 모비율과 같은 $\frac{2}{5}$가 될 수 밖에 없음을 이해한다.

⇨ <복원시행>의 규칙에서 결국 표본비율이 $\frac{r}{n}$일 확률은 결국 <$n$회의 독립시행> 중 <특정성질을 만족한 횟수>가 $r$회일 확률임을 이해한다. 그리고 이것은 독립시행의 확률이므로 $P\left(\frac{r}{n}\right) = {}_nC_r p^r(1-p)^{n-r}$임을 이해한다.

(독립시행의 확률은 확률에서 사고방식이 아닌 유일한 공식이라고 설명한 바 있다. 특정상황, 특정구조를 만족할 때 이것이 독립시행의 확률을 구하는 상황임을 눈치 챌 수 있어야 한다. 그만큼 많이 나오기 때문에 공식화 시켜놓은 것이다.)

3단계 : 위의 내용을 근거로 이산확률분포 표를 완성한다.

| $\hat{p}$ | $\frac{0}{n}$ | $\frac{1}{n}$ | $\cdots$ | $\frac{r}{n}$ | $\cdots$ | $\frac{n}{n}$ | 총합 |
|---|---|---|---|---|---|---|---|
| $P(\hat{p})$ | ${}_nC_0p^0(1-p)^n$ | ${}_nC_1p^1(1-p)^{n-1}$ | $\cdots$ | ${}_nC_rp^r(1-p)^{n-r}$ | $\cdots$ | ${}_nC_np^n(1-p)^0$ | 1 |

⇨ 이 표에서 $\hat{p}$을 이항분포 $B(n,\ p)$를 따르는 확률변수 $X$를 변형한 확률변수인 $\frac{X}{n}$라고 생각할 수 있다.
이때 $X$는 총 시행횟수가 $n$이고, 1회 시행에서 사건이 발생할 확률이 $p$이므로 $B(n,\ p)$를 따른다.

4단계 : <이항분포의 공식>과 <평균, 분산, 표준편차의 성질>을 이용하여 공식을 완성한다.

① $E(\hat{p}) = \sum \frac{r}{n} \times {}_nC_rp^r(1-p)^{n-r} = \frac{1}{n}\sum r\times {}_nC_rp^r(1-p)^{n-r} = \frac{1}{n}\times np$ - 이항분포에서 증명

⇨ 물론 간단히 성질을 이용하면 증명된다. $E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}\times np = p$

② $V(\hat{p}) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X) = \frac{1}{n^2}\times npq = \frac{pq}{n}$

③ $\sigma(\hat{p}) = \sqrt{V\left(\frac{X}{n}\right)} = \sqrt{\frac{pq}{n}}$

- 성질이 와 닿지 않는다면 ①과 같이 가장 기본적인 원리로 다시 증명해 보길 바란다.

## 2) 공식과 의미 Reding

- <이항분포의 평균, 분산, 표준편차>를 증명하는 과정은 꽤 복잡했지만 <결과>는 간단한 <공식>으로 도출되어 실제로는 <일반적인 이산확률분포의 평균, 분산, 표준편차>를 구하는 것에 비해 간단히 계산됨을 기억할 것이다.

- <표본비율의 평균, 분산, 표준편차>도 증명을 이해할 때는 약간의 혼동이 올 수 있지만 <결과>는 간단한 <공식>으로 도출되어 실제로는 <표본평균의 평균, 분산, 표준편차>를 구하는 것에 비해 간단히 계산된다.
  (물론 <표본평균의 평균, 분산, 표준편차>도 구하는 공식 $E(\overline{X}) = m$, $V(\overline{X}) = \frac{\sigma^2}{n}$, $\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$가 있지만 이 공식을 사용하기 위해서는 <모평균 $m$과 모분산 $\sigma^2$>을 구해서 대입해야 하므로 결국 이 과정이 <일반적인 이산확률분포의 평균, 분산, 표준편차>를 구하는 과정이다.)

- **상쇄효과** : $\hat{p}$과 $p$는 다르지만 ⇨ $\hat{p}\,\hat{q}$과 $pq$는 같다고 본다.
  : $\hat{p}$과 $p$는 비슷할 뿐 같지 않으므로 $\hat{p}$으로 $p$를 대체할 수는 없다. (그래서 후에 $\hat{p}$으로 $p$를 추정한다.)
  하지만 $\sqrt{\frac{pq}{n}}$에서 $pq$ 대신에는 $\hat{p}\,\hat{q}$을 쓸 수 있는데 이것을 **상쇄효과**라고 한다.
  $p$ 보다 $\hat{p}$이 조금 클 경우 $1-p$인 $q$보다 $1-\hat{p}$인 $\hat{q}$이 조금 작아져서 $pq$와 $\hat{p}\,\hat{q}$은 거의 같아진다.

① 공식 : $E(\hat{p}) = p$
의미 : <$\hat{p}$의 평균>은 <$p$ = 모비율>과 같다.

② 공식 : $V(\hat{p}) = \frac{pq}{n} = \frac{\hat{p}\,\hat{q}}{n}$ (단, $p+q=1$, $\hat{p}+\hat{q}=1$)
의미 : <$\hat{p}$의 분산>은 $p$와 $(1-p)$를 곱하여 $n$으로 나눈 것과 같다.

③ 공식 : $\sigma(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{\hat{p}\,\hat{q}}{n}}$ (단, $p+q=1$, $\hat{p}+\hat{q}=1$)
⇨ <$\overline{X}$의 표준편차>는 분산에 $\sqrt{\ }$를 씌우면 되므로 간단히 넘어간다.

## 3) 공식을 사용하는 근거

**(1) <복원 추출>하는 경우** : 표본의 크기 $n$에 관계없이 $E(\hat{p}) = p$와 $V(\hat{p}) = \frac{pq}{n} = \frac{\hat{p}\,\hat{q}}{n}$ (단, $p+q=1$) 성립

⇨ 모집단의 크기(모집단의 자료의 개수)와 표본의 크기(뽑는 자료의 개수)가 작다면 반드시 <복원추출>이라는 조건을 줘야 한다.

**(2) 비 복원 추출 하는 경우** : 모집단의 크기가 충분히 크면 복원 추출하는 경우와 근사적으로 같다.

⇨ 결국 <복원, 비복원>에 대한 언급이 없어도 $E(\hat{p}) = p$와 $V(\hat{p}) = \frac{pq}{n} = \frac{\hat{p}\,\hat{q}}{n}$ (단, $p+q=1$) 성립

사실 그냥 생각해봐도 <5천만명>중에 표본의 크기가 10인 표본을 추출한다면
<1명씩 10번 뽑는 과정>에서 조사한 사람을 조사할 대상에 다시 포함시키든, 다시 포함시키지 않든
결과에는 거의 영향을 주지 않을 것이다. (누구나 직관적으로 느낄 수 있다.)
즉, 굳이 <복원추출>하지 않아도 상관없는 경우에는 <임의추출>이라는 말만 나와도 위의 <공식>을 쓰는 것에
무리가 없고, 그러므로 고등학교 범위의 문제 상에서는 <복원추출>이라는 말이 언급되지 않아도 <공식>을 써도 된다.
(고등학교 문제도 자료의 개수와 표본의 크기가 매우 작은 상황을 설정해서 줬다면 조건을 꼼꼼히 확인해야 한다.)

# 3. $\hat{p}$의 분포

## 1) 무조건 이항분포

- 모집단의 자료의 개수에 상관없이 <표본의 크기>가 충분히 작아도 (10개 이하) $\hat{p}$은 이항분포의 성격을 가진다.
  이때 <크기가 $n$인 표본을 복원추출>한다는 조건이 반드시 주어진다. 이 말속에는 <임의로 1개씩 n번 뽑는 복원추출>이라는 뜻이 숨겨져 있는데 이 말 뜻만 이해하면 <독립사건의 확률>문제 그 이상도 이하도 아니다.

- 이 문제에서도 <$\hat{p}$의 평균, 분산, 표준편차>공식은 항상 성립한다. 물론 <복원추출>을 반드시 문제에서 명시할 것이다.

## 2) 이항분포의 정규분포 화

- 표본의 크기에 상관없이 $\hat{p} = \dfrac{X}{n}$에서 $X$의 분포는 항상 이항분포 $B(n,\ p)$를 따른다.

- 단지 $n$이 작으면 그대로 ⇨ $\hat{p} = \dfrac{X}{n}$에서 $X$의 분포는 이항분포 $B(n,\ p)$로서 인정하여 확률 값을 계산할 수 있고
  $n$이 충분히 크면 ⇨ 이항분포를 <정규분포 화> 할 수 있다. 물론 $X$에 대해서 <정규분포 화>하면
  $N\left(np,\ (\sqrt{npq})^2\right)$이겠지만 $\hat{p} = \dfrac{X}{n}$이므로 $\hat{p}$은 $N\left(p,\ \left(\sqrt{\dfrac{pq}{n}}\right)^2\right)$ 또는 $N\left(p,\ \left(\sqrt{\dfrac{\hat{p}\hat{q}}{n}}\right)^2\right)$인
  정규분포를 따른다. 여기까지 왔다면 정규분포 문제 그 이상도 이하도 아니다.

**149** 이해를 위한 예제

학교 알리미 교육정보공개서비스에 따르면 어느 고등학교의 4년제 대학교의 진학률이 $60\%$라 한다. 이 고등학교 학생 120명을 대상으로 4년제 대학교 진학률을 조사하였을 때, 4년제 대학교 진학률의 분산은?

① 0.002 ② 0.003 ③ 0.004 ④ 0.005 ⑤ 0.006

조건 어느 고등학교의 4년제 대학교의 진학률이 $60\%$ ⇨ 모비율을 $p$라고 할 때 $p = 0.6$

구할 것 학생 120명을 대상으로 4년제 대학교 진학률을 조사하였을 때, 4년제 대학교 진학률의 분산

⇨ 표본의 크기가 120인 추출에서 표본비율(표본에서 대학 진학률)을 $\hat{p}$이라고 할 때, $V(\hat{p})$를 구하는 문제이다.

공식에 의하여 $V(\hat{p}) = \dfrac{pq}{n} = \dfrac{0.6 \times 0.4}{120} = 0.002$

정답은 ①

티칭 <복원추출>이라는 전제 + <이항분포의 내용>만 정확히 알아도 충분히 풀 수 있는 문제다.

- 1단계 : 이 학교에서 임의로 1명을 뽑았을 때 그 학생이 4년제 대학교의 진학했을 확률을 0.6이라고 생각해도 상관없다.
- 2단계 : 120명 조사 ⇨ <1명씩 120번 복원추출>로 조사했다는 전제가 있다. 이때 <4년제 대학을 진학한 학생 수>는 0부터 최대 120까지이고, 조사한 120명 중 $r$명이 진학했을 확률은 ${}_{120}C_r(0.6)^r(0.4)^{n-r}$이다.
  즉, <4년제 대학을 진학한 학생 수>를 확률변수 $X$라고 하면 $X$는 $B(120,\ 0.6)$을 따른다.
- 3단계 : 즉, $V(X) = 120 \times 0.6 \times 0.4$이고 <120명의 진학률 - $\dfrac{X}{120}$>의 분산은 $V\left(\dfrac{X}{120}\right) = \dfrac{120 \times 0.6 \times 0.4}{120^2}$이다.
  이제 계산만 하면 위와 같은 결론을 얻는다. 위의 공식은 이 과정을 생략한 것뿐이다. (이해를 기반으로 암기하길..)

**150** 이해를 위한 예제 2007. 9. 나형(32%). 24번. 4점 - 변형

어느 스포츠 용품 가게에서는 별(★) 모양이 그려져 있는 야구공 한 개를 포함되거나 포함되지 않은 20개의 야구공을 한 상자에 담아 판매한다고 한다. 이때 별(★) 모양이 그려져 있는 야구공이 포함된 상자를 선택할 경우 당첨되어 축구공을 경품으로 준다고 할 때, 이 가게의 당첨율이 $0.25$이라고 한다. 이 가게에서 임의로 표본의 크기가 $3$인 표본을 복원추출 할 때, 당첨율이 $0.5$이상일 $0.8$ 이하일 확률은 $\frac{q}{p}$이다. $p+q$의 값을 구하시오. (단, $p$, $q$는 서로소인 자연수이다.)

조건 이 가게의 당첨율이 $0.25$ ⟹ 결국 모비율을 $p$라고 할 때 $p = 0.25$

조건 임의로 표본의 크기가 $3$인 표본을 복원추출 ⟹ $n = 3$

구할 것 당첨율이 $0.5$이상일 확률은 $\frac{q}{p}$ ⟹ 결국 표본비율 $p$ 라고 할 때 $P(0.5 \le \hat{p} \le 0.8)$을 구하는 것이다.

이것을 표로 작성하면

| $\hat{p}$ | $\frac{0}{3}$ | $\frac{1}{3}$ | $\frac{2}{3}$ | $\frac{3}{3}$ | 총합 |
|---|---|---|---|---|---|
| $P(\hat{p})$ | ${}_3C_0(0.25)^0(0.75)^3$ | ${}_3C_1(0.25)^1(0.75)^2$ | ${}_3C_2(0.25)^2(0.75)^1$ | ${}_3C_3(0.25)^3(0.75)^0$ | 1 |

즉, $P(0.5 \le \hat{p} \le 0.8) = P\left(\hat{p} = \frac{2}{3}\right) = {}_3C_2(0.25)^2(0.75)^1 = \frac{9}{64}$이므로 정답은 $73$

티칭 이 문제는 원래 아래와 같았다. (앞의 이해를 위한 예제 075)

- 어느 스포츠 용품 가게에서는 별(★) 모양이 그려져 있는 야구공 한 개를 포함하여 모두 20개의 야구공을 한 상자에 넣어 상자 단위로 판매한다. 한 상자에서 5개의 야구공을 임의 추출하여 별(★) 모양이 그려져 있는 야구공이 있으면 축구공 한 개를 경품으로 준다. 어느 고객이 이 가게에서 야구공 3상자를 구입하여 경품 당첨 여부를 모두 확인할 때, 축구공 2개를 경품으로 받을 확률은 $\frac{q}{p}$이다. $p+q$의 값을 구하시오. (단, $p$, $q$는 서로소인 자연수이다.)

1회 시행에서 경품을 받을 확률 $\frac{{}_{19}C_4}{{}_{20}C_5} = \frac{1}{4}$이다. 이것을 <당첨율 - 모비율>로 문제를 바꾼 것뿐이다.

독립시행의 확률 3상자를 구입하여 경품을 2개 받을 확률은 ${}_3C_2\left(\frac{1}{4}\right)^2\frac{3}{4} = \frac{9}{64}$

이 부분을 <표본비율>을 이용하여 <통계적>으로 표현했을 뿐이다.

- 이렇게 문제의 본질은 같지만 표현을 다르게 함으로서 문제를 완전히 달라보이게 할 수 있다. 아직 수능과 평가원에 등장한 적이 없으므로 유념하여 보길 바란다. (2016년 9월 기준)

티칭 결국 <시행횟수>가 적을 때 <표본비율의 분포>와 그 확률을 물어보는 문제는 <이항분포와 독립시행의 확률> 그 이상도 이하도 아니다. 이 부분에 대한 추가 설명은 개념의 외연에서 하기로 한다.

**151** 이해를 위한 예제

어느 지역에 거주하는 60대 이상의 주민들 중에서 백내장 환자의 비율은 36%라고 한다. 이 지역에 거주하는 60대 이상의 주민들 중에서 100명을 임의로 추출하였을 때, 그 중에 백내장 환자의 비율이 0.48 이상일 확률을 다음 표준정규분포 표를 이용하여 구한 것은?

① 0.0062 ② 0.0228 ③ 0.0668 ④ 0.1525 ⑤ 0.1587

| $z$ | $P(0 \le Z \le z)$ |
|---|---|
| 1.0 | 0.3413 |
| 1.5 | 0.4332 |
| 2.0 | 0.4772 |
| 2.5 | 0.4938 |

조건 60대 이상의 주민들 중에서 백내장 환자의 비율은 36% ⇨ 결국 모비율을 $p$라고 할 때 $p = 0.36$

조건 60대 이상의 주민들 중에서 100명을 임의로 추출 ⇨ $n = 100$ (복원이든 비복원이든 공식을 써도 상관없다.)

구할 것 백내장 환자의 비율이 0.48 이상일 확률을 ⇨ 결국 $n = 100$일 때, $P(\hat{p} \ge 0.48)$을 구하는 것이다.

- 공식에 의하여 $E(\hat{p}) = p = 0.36$이고, $V(\hat{p}) = \dfrac{pq}{n} = \dfrac{0.36 \times 0.64}{100} = \left(\dfrac{0.6 \times 0.8}{10}\right)^2 = (0.048)^2$이므로

$\hat{p}$은 $N(0.36, (0.048)^2)$인 정규분포를 따른다. (이항분포가 정규분포 화 된 것이다.)

즉, $P(\hat{p} \ge 0.48) = P\left(\dfrac{\hat{p} - 0.36}{0.048} \ge \dfrac{0.48 - 0.36}{0.048}\right) = P(Z \ge 2.5) = 0.0062$이다. 정답은 ①

**152** 이해를 위한 예제 2010. 11. 가형(선택 확률과 통계 30번). 4점

우리나라 성인을 대상으로 특정 질병에 대한 항체 보유비율을 조사하려고 한다. 모집단의 항체보유비율을 $p$, 모집단에서 임의로 추출한 $n$명을 대상으로 조사한 표본의 항체 보유 비율을 $\hat{p}$이라고 할 때, $|\hat{p} - p| \le 0.16\sqrt{\hat{p}(1-\hat{p})}$ 일 확률이 0.9544 이상이 되도록 하는 $n$의 최솟값을 구하시오. (단 $Z$가 표본정규분포를 따르는 확률변수 일 때, $P(0 \le Z \le 2) = 0.4772$이다.)

조건 모집단의 항체보유비율을 $p$, 모집단에서 임의로 추출한 $n$명을 대상으로 조사한 표본의 항체 보유 비율을 $\hat{p}$

⇨ $\hat{p}$은 $N\left(p, \left(\sqrt{\dfrac{p(1-p)}{n}}\right)^2\right)$ 또는 $N\left(p, \left(\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}\right)^2\right)$인 정규분포를 따른다. **(상쇄효과)**

조건과 구할 것 $|\hat{p} - p| \le 0.16\sqrt{\hat{p}(1-\hat{p})}$ 일 확률이 0.9544 이상이 되도록 하는 $n$의 최솟값

⇨ $P\left(|\hat{p} - p| \le 0.16\sqrt{\hat{p}(1-\hat{p})}\right) \ge 0.9544$

- 이 식을 표준화 시킬 때, 식의 구조상 $\hat{p}$의 분포는 $N\left(p, \left(\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}\right)^2\right)$라고 보는 것이 좋다.

즉, $P\left(|\hat{p} - p| \le 0.16\sqrt{\hat{p}(1-\hat{p})}\right) \ge 0.9544 \iff P\left(\left|\dfrac{\hat{p} - p}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}}\right| \le \dfrac{0.16\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}}\right) \ge 0.9544$

$\iff P(|Z| \le 0.16\sqrt{n}) \ge 0.9544$ - 복잡해 보이지만 표준화를 했을 뿐이다.

- 이 식이 만족하려면 $P(|Z| \le 2) = 0.4772$이므로 $0.16\sqrt{n} \ge 2$이어야만 한다. 즉, $n \ge 156.25$이므로

정답은 157

코칭 의미를 모르고 외운 자들에게는 너무 복잡해 보이는 식이지만 <식의 의미를 아는 자>에게는 단순한 표준화의 과정일 뿐이고, 이후로는 간단한 정규분포 문제일 뿐이다.

## #4 모비율의 추정

# 1. 모비율의 추정 식의 도출 - 모평균 추정과 거의 같다.

### 1) 1단계 : 신뢰계수와 신뢰도

- 새로운 확률변수 $\hat{p}$이 새로운 정규분포 $N\left(p,\ \left(\sqrt{\frac{pq}{n}}\right)^2\right)$을 따른다는 사실을 알고 있고, - 항상 $\sqrt{\frac{pq}{n}} = \sqrt{\frac{\hat{p}\,\hat{q}}{n}}$ 은 염두한다.

  정규분포인 $\hat{p}$의 확률을 $P\left(p-k\sqrt{\frac{pq}{n}} \le \hat{p} \le p+k\sqrt{\frac{pq}{n}}\right)$와 같이 표현했을 때, 확률 값은 $k$에만 영향을 받는다는 사실은 이미 정규분포에서 학습한 바 있다.

$$P(|Z| \le 1.96) = 0.95\text{이고, } P(|Z| \le 2.58) = 0.99\text{이므로}$$

(문제의 조건으로 주어지며 표준정규분포 표를 찾으면 쉽게 알 수 있는 내용이기도 하다.)

- $k = 1.96$인 경우 $P\left(p-1.96\sqrt{\frac{pq}{n}} \le \hat{p} \le p+1.96\sqrt{\frac{pq}{n}}\right) = 0.95$
- $k = 2.58$인 경우 $P\left(p-2.58\sqrt{\frac{pq}{n}} \le \hat{p} \le p+2.58\sqrt{\frac{pq}{n}}\right) = 0.99$

- 이때 $k$값을 <신뢰계수>라 하고 그에 따른 확률 값을 백분율로 표현한 것을 <신뢰도>라고 한다.
  즉, 신뢰계수가 $1.96$인 경우 신뢰도는 $95\%$로 정해져 있고,
  신뢰계수가 $2.58$인 경우 신뢰도는 $99\%$로 정해져 있다. 신뢰도는 오직 신뢰계수에만 영향을 받는다.

### 2) 2단계 : 식의 표현 바꾸기

- 먼저 상쇄효과에 의하여 $pq$를 대신해 $\hat{p}\,\hat{q}$으로 고치고 시작한다. 이 단원의 목적이 $\hat{p}$을 통해서 $p$를 추정하는 것이므로 $p$가 주어지는 일이 없기 때문이다.

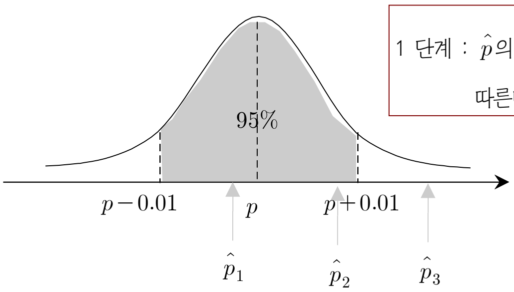
$$P\left(p-1.96\sqrt{\frac{\hat{p}\,\hat{q}}{n}} \le \hat{p} \le p+1.96\sqrt{\frac{\hat{p}\,\hat{q}}{n}}\right) = 0.95 \quad \Leftrightarrow \quad P\left(|\hat{p}-p| \le 1.96\sqrt{\frac{\hat{p}\,\hat{q}}{n}}\right) = 0.95$$

- 와 같이 식을 변형하면 ⇨ 표본비율($\hat{p}$)과 모비율의 차이($p$)가 $1.96\sqrt{\frac{\hat{p}\,\hat{q}}{n}}$ 이하일 확률이 $0.95$라고 읽을 수 있다.

- 조금 더 실생활적으로 말을 바꾸어 표현하면 ⇨ 표본평균과 모평균의 차이가 $1.96\sqrt{\frac{\hat{p}\,\hat{q}}{n}}$ 이하일 가능성은 $95\%$라고 읽을 수 있다.

- 구체적인 상황을 설정하여 그래프 상에서의 느낌을 확인해보자.
  어떤 모집단의 표준편차가 알려져 있고 표본의 크기가 $100$인 표본을 임의 추출했다고 가정해보자.
  이때, $\sigma$는 어차피 문제에서 정해지는 값이므로 $1.96\sqrt{\frac{\hat{p}\,\hat{q}}{100}}$ 의 계산된 결과가 $0.01$라고 가정해보자.

1 단계 : $\hat{p}$의 분포는 평균이 $p$(모비율)이고, 표준편차가 $1.96\sqrt{\frac{\hat{p}\,\hat{q}}{100}}$ 인 정규분포를 따른다. (원래는 이항분포이지만 $n$이 $100$정도면 정규분포로 근사된다.)




$$P(|\hat{p}-p| \le 0.01) = 0.95$$

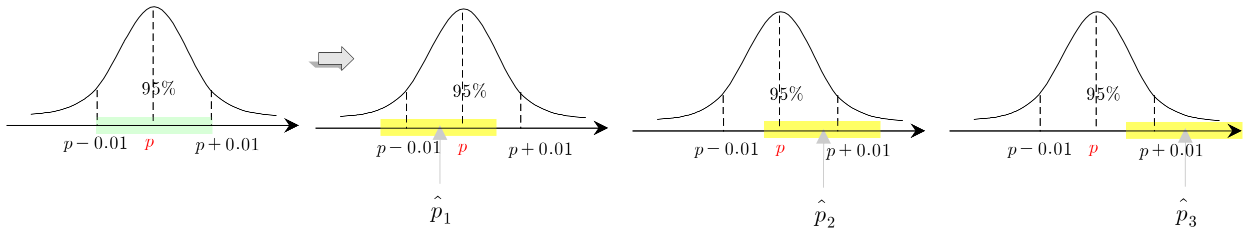
① $\hat{p}_1$, $\hat{p}_2$와 같이 표본을 딱 1개만 조사할 경우 그 표본비율과 모비율의 차이가 왼쪽 그림과 같이 $0.01$이하가 될 가능성이 $95\%$이다.

② 물론 $\hat{p}_3$처럼 우리가 조사한 표본비율과 모비율의 차이가 $0.01$이상일 가능성도 $5\%$는 된다.

## 3) 3단계 : 최종식

- 최종식은 $\hat{p}$의 입장에서 식을 다시 바라본 것이다.

$P(\hat{p}-0.01 \le p \le \hat{p}+0.01) = 0.95$ ⇨ $\hat{p}$를 기준으로 좌우로 길이가 $0.01$인 띠를 만든다.




① $\hat{p}_1$, $\hat{p}_2$와 같이 **표본을 딱 1개만 조사**할 경우 그 〈표본비율〉의 좌우로 $0.01$인 구간 안에 〈모비율〉이 존재할 가능성이 $95\%$이다.

② 물론 $\hat{p}_3$처럼 우리가 조사한 〈표본비율〉의 좌우로 $0.01$인 구간 안에 모비율이 존재하지 않을 가능성도 $5\%$는 된다.

- 위의 결론에 대하여 수학적인 기호를 제하고 일반적인 용어를 통해서 재정립한다. 통계치를 수학적 용어로 대중에게 전달할 수는 없는 일이다.

신뢰도 $\alpha\%$의 추정 식 : $\hat{p}-k\sqrt{\dfrac{\hat{p}\hat{q}}{n}} \le p \le \hat{p}+k\sqrt{\dfrac{\hat{p}\hat{q}}{n}}$ (단, $P(|Z| \le z) = \dfrac{\alpha}{100}$)

① 표본비율 $(\hat{p})$ : 물론 이 식을 유도하는 과정까지는 〈확률변수〉의 역할을 했지만 위의 식을 실제로 사용하는 과정에서는 〈상수〉의 역할을 한다. 이제는 〈표본은 1개만 조사〉할 것이고, 이 한 개의 표본의 〈표본비율〉을 기준으로 좌우로 $k\sqrt{\dfrac{\hat{p}\hat{q}}{n}}$ 만큼 떨어진 곳 사이에 모비율이 존재할 것이라고 말할 수 있다.

② 신뢰계수$(k)$와 신뢰도$(\alpha\%)$ : $\alpha$는 **확률의 의미**를 가지므로 $k$값에만 영향을 받는다. 즉, $k$값이 결정됨에 따라 ①의 결론이 **〈얼마만큼 믿을 만한지〉** 알 수 있다.

- $k=1.96$인 경우 $\alpha=95$ ⇨ 보통 표준정규분포를 활용해서 $P(|Z| \le 1.96) = 0.95$처럼 주므로 외울 필요는 없다
- $k=2.58$인 경우 $\alpha=99$ ⇨ 보통 표준정규분포를 활용해서 $P(|Z| \le 2.58) = 0.99$처럼 주므로 외울 필요는 없다

③ 신뢰구간 $\left[\hat{p}-k\sqrt{\dfrac{\hat{p}\hat{q}}{n}},\ \hat{p}+k\sqrt{\dfrac{\hat{p}\hat{q}}{n}}\right]$ : 모비율이 존재할 것이라고 예측되는 구간이다. 위의 그래프에서 ＿＿＿＿에 해당하는 구간을 식으로 표현한 것뿐이다. (물론 $\hat{p}_3$와 같이 신뢰구간 안에 모비율($p$)이 존재하지 않을 가능성도 존재하고 이 가능성은 신뢰계수에 의해 결정된다.)

④ 신뢰구간의 길이 $2k\sqrt{\dfrac{\hat{p}\hat{q}}{n}}$ : 위의 신뢰구간에서 $\left(\hat{p}+k\sqrt{\dfrac{\hat{p}\hat{q}}{n}}\right)-\left(\hat{p}-k\sqrt{\dfrac{\hat{p}\hat{q}}{n}}\right) = 2k\sqrt{\dfrac{\hat{p}\hat{q}}{n}}$으로 계산된 것이다. 이것은 〈정확도〉를 측정하는 기준이 된다. 위의 그래프에서 ＿＿＿＿의 길이를 의미한다. ＿＿＿＿의 길이가 짧을수록 〈정확한 추정〉이 된다.

- 예를 들어 $A$ 후보를 지지하는 사람의 비율이 〈$0.5$와 $0.6$사이〉라고 말하는 것보다 〈$0.54$과 $0.56$사이〉라고 말하는 것이 모비율을 조금 더 〈정확하게 예측했다〉라고 말할 수 있을 것이다

⑤ 표본오차, 표본오차의 최댓값 : 말 그대로 우리 뽑은 〈한 개의 표본의 비율〉과 〈모비율〉의 차이가 오차이다. 추정치를 신뢰한다고 가정했을 때 (물론 틀릴 가능성도 있지만 확률적으로 작기 때문에) 표본오차의 범위는 $|\hat{p}-p| \le k\sqrt{\dfrac{\hat{p}\hat{q}}{n}}$이고 표본오차의 최댓값은 당연히 $k\dfrac{\sigma}{\sqrt{n}}$이다.

⑥ $\hat{p}\,\hat{q}$ : 원래는 $\langle pq \rangle$가 있어야 할 자리지만 상쇄효과에 의하여 $\hat{p}\,\hat{q}$ 으로 대체되었다.

(모비율을 몰라서 추정하는데 모비율을 줄 리는 없다.)

⑦ $n$ : 표본의 크기이고 문제에서 주어진 대로 대입한다.

티칭 정확한 추정을 하는 법

- 신뢰도 $95\%$의 추정 식 : $\hat{p}-k\sqrt{\dfrac{\hat{p}\,\hat{q}}{n}} \le p \le \hat{p}+k\sqrt{\dfrac{\hat{p}\,\hat{q}}{n}}$ 에 대해 생각해보자.

⇨ 정확한 추정을 위해 <신뢰구간의 길이>인 $2\times 1.96\sqrt{\dfrac{\hat{p}\,\hat{q}}{n}}$ 을 작게 하면 된다. 그런데 $1.96$은 거의 건드리지 않는다.

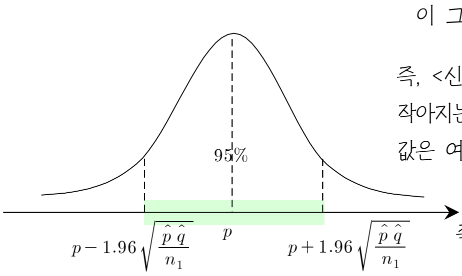
예를 들어 실제로 $1.96\sqrt{\dfrac{\hat{p}\,\hat{q}}{n}}$ 을 작게 하기 위해서 $k=0.5$로 두고 계산하면 $P\left(|\hat{p}-p| \le 0.5\sqrt{\dfrac{\hat{p}\,\hat{q}}{n}}\right) = 0.383$

이다. 이 말은 <신뢰구간의 길이>는 $2\times 0.5\sqrt{\dfrac{\hat{p}\,\hat{q}}{n}}$ 로 짧아졌으나 그럴 가능성이 $38.3\%$밖에 안 된다는 뜻이다.

즉, 아무리 표본비율과 모비율이 가깝다고 해도 신뢰도가 낮으면 믿을 수가 없으므로 무가치한 추정이다.
그래서 실제로 $k$는 일상적으로 가장 효율적인 수치인 $k = 1.96$과 $k = 2.58$로 정해놓은 것이다.

⇨ $\hat{p}$는 우리가 정하는 값이 아니라 <조사결과 도출되는 값, 문제에서는 조건>일 뿐이기 때문에 결국 정확한 추정을 위해 우리가 할 수 있는 일은 $n$을 크게 하는 것뿐이다. 즉, 표본조사를 할 때 그 표본 안의 자료$(n)$를 많을수록 정확한 추정이 된다. - 많이 조사하라는 뜻이다.

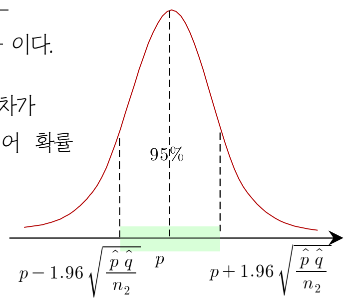
⇨ 그래프로 생각해보면 $\hat{p}$을 다시 확률변수로 생각했을 때 $n$을 크게 할 경우 $\sigma(\hat{p})$인 $\sqrt{\dfrac{\hat{p}\,\hat{q}}{n}}$ 이 작아지므로 가운데로 몰려있는 그래프가 될 것이고, 그래서 신뢰구간의 길이는 짧아지지만 신뢰도(확률값)에는 변함이 없게 된다.
아래 그림을 통해 다시 한 번 이해해본다.




이 그림과 같이 $n_1 < n_2$일 경우 $\sqrt{\dfrac{\hat{p}\,\hat{q}}{n_1}} > \sqrt{\dfrac{\hat{p}\,\hat{q}}{n_2}}$ 이다.

즉, <신뢰구간은 짧아지지만> $\hat{p}$의 입장에서는 표준편차가 작아지는 효과가 있어 그래프가 가운데로 몰리게 되어 확률값은 여전히 $0.95$가 된다.

즉, 신뢰도의 변화없이 신뢰구간의 길이를 짧게 하고 싶다면 $n$을 크게 하는 수밖에 없다.




코칭 혹시 모비율과 표본비율을 백분율(%)를 통해서 준다고 해도 당황하지 말고 일단 100으로 나눈 후 차근차근 푼다.

**153** 이해를 위한 예제 2016. 09. 가형. 28번. 4점

어느 고등학교에서 대중교통을 이용하여 등교하는 학생의 비율을 알아보기 위하여 이 고등학교 학생 중 $n$명을 임의추출하여 조사한 결과 $50\%$의 학생이 대중교통을 이용하여 등교하는 것으로 나타났다. 이 결과를 이용하여 구한 이 고등학교 전체학생 중에서 대중교통을 이용하여 등교하는 학생의 비율 $p$에 대한 신뢰도 $95\%$의 신뢰구간이 $a \le p \le b$이다. $b-a=0.14$일 때, $n$의 값을 구하시오. (단, $Z$가 표준정규분포를 따르는 확률변수일 때, $\mathrm{P}(|Z| \le 1.96)=0.95$로 계산한다.)

조건 이 고등학교 학생 중 $n$명을 임의추출하여 조사한 결과 $50\%$의 학생이 대중교통을 이용하여 등교

⇨ 추출된 표본에서의 비율이므로 표본비율이다. 즉, $\hat{p} = \dfrac{1}{2}$

조건 $\mathrm{P}(|Z| \le 1.96)=0.95$로 계산 ⇨ $95\%$에서 신뢰계수를 $1.96$으로 인정하라는 것이다.

구할 것 학생의 비율 $p$에 대한 신뢰도 $95\%$의 신뢰구간이 $a \le p \le b$이다. $b-a=0.14$일 때, $n$의 값

⇨ $b-a$가 신뢰구간을 의미하는 것은 이미 알지만 주어진 그대로 풀어보자. 신뢰도 $95\%$의 신뢰구간은

$$\frac{1}{2}-1.96\sqrt{\frac{\frac{1}{2}\cdot\frac{1}{2}}{n}} \le p \le \frac{1}{2}+1.96\sqrt{\frac{\frac{1}{2}\cdot\frac{1}{2}}{n}} \Leftrightarrow \frac{1}{2}-1.96\sqrt{\frac{1}{4n}} \le p \le \frac{1}{2}+1.96\sqrt{\frac{1}{4n}}$$

따라서 $b-a = 2\times1.96\times\sqrt{\dfrac{1}{4n}} = 1.96\sqrt{\dfrac{1}{n}} = 0.14$이므로

$\sqrt{n}=14 \Leftrightarrow n=196$이다. 정답은 $196$

**154** 이해를 위한 예제

임의로 추출된 고등학생 $400$명을 대상으로 계열별 선호도를 조사하였더니 인문계열을 선호하는 학생이 $256$명, 자연계열을 선호하는 학생이 $144$명이었다. 이 고등학교 학생 중에서 자연계열을 선호하는 학생의 비율을 신뢰도 $99\%$로 추정할 때, 모비율 $p$와 표본비율 $\hat{p}$의 최대오차는? (단, $P(|Z| \le 3)=0.99$로 계산한다.)

① $0.048$ ② $0.072$ ③ $0.084$ ④ $0.123$ ⑤ $0.144$

조건 임의로 추출된 고등학생 $400$명을 대상으로 계열별 선호도를 조사하였더니 인문계열을 선호하는 학생이 $256$명, 자연계열을 선호하는 학생이 $144$ ⇨ 추출된 표본에서의 비율이므로 표본비율이다.

자연계열을 선호학생의 비율을 $p$라고 할 경우 $\hat{p} = \dfrac{144}{400}$

조건 단, $P(|Z| \le 3)=0.99$로 계산한다. - $99\%$에서 신뢰계수를 $1.96$으로 인정. (큰 시험에서 2.58로 줄 가능성이 크다.)

구할 것 모비율 $p$와 표본비율 $\hat{p}$의 최대오차 ⇨ $|p-\hat{p}| \le k\sqrt{\dfrac{\hat{p}\hat{q}}{n}}$에서 (추정치가 맞는다고 가정했을 때)최대오차는 $k\sqrt{\dfrac{\hat{p}\hat{q}}{n}}$

$$k=3,\ n=400,\ \hat{p}=\frac{144}{400},\ \hat{q}=\frac{256}{400}\text{이므로 } 3\sqrt{\frac{\left(\frac{144}{400}\right)\left(\frac{256}{400}\right)}{400}} = 3\times\frac{\left(\frac{12}{20}\right)\left(\frac{16}{20}\right)}{20} = \frac{36}{500}\text{이다.}$$

정답은 ②

**155** 이해를 위한 예제 2008. 11. 가형(선택 확률과 통계 30번). 4점

어떤 모집단에서 임의로 100명을 추출하여 구한 모비율에 대한 신뢰도 $95\%$의 신뢰구간이 $\left[\frac{1}{10}-c,\ \frac{1}{10}+c\right]$이었다. 같은 모집단에서 $n$명을 임의로 추출하여 구한 모비율에 대한 신뢰도 $95\%$의 신뢰구간이 $\left[\frac{1}{9}-s(n), \frac{1}{9}+s(n)\right]$이고 $s(n)=\frac{50}{81}c$이다. $n$의 값을 구하시오.

조건 어떤 모집단에서 임의로 100명을 추출하여 구한 모비율에 대한 신뢰도 $95\%$의 신뢰구간이 $\left[\frac{1}{10}-c,\ \frac{1}{10}+c\right]$

⇨ 즉, $n = 100$, $\hat{p}_1 = \frac{1}{10}$이다. 신뢰도 $95\%$에 대한 신뢰계수가 주어지지 않았으므로 $k$라고 하자.

이때, $c = k\sqrt{\frac{\hat{p}_1\hat{q}_1}{100}} = k\sqrt{\frac{\frac{1}{10}\times\frac{9}{10}}{100}} = \frac{3}{100}k$이다.

조건 같은 모집단에서 $n$명을 임의로 추출하여 구한 모비율에 대한 신뢰도 $95\%$의 신뢰구간이 $\left[\frac{1}{9}-s(n), \frac{1}{9}+s(n)\right]$

⇨ 같은 모집단이므로 모비율이 같다. 즉, $\hat{p}_2 = \frac{1}{9}$이다. 신뢰도 $95\%$에 대한 신뢰계수는 역시 $k$이다.

이때, $s(n) = k\sqrt{\frac{\hat{p}_2\hat{q}_2}{100}} = k\sqrt{\frac{\frac{1}{9}\times\frac{8}{9}}{n}} = \frac{2\sqrt{2}}{9\sqrt{n}}k$

구할 것 $s(n)=\frac{50}{81}c$이다. $n$의 값을 구하시오.

⇨ 결국 $\frac{2\sqrt{2}}{9\sqrt{n}}k = \frac{50}{81}c \Leftrightarrow \frac{2\sqrt{2}}{9\sqrt{n}}k = \frac{50}{81}\times\frac{3}{100}k \Leftrightarrow \sqrt{n} = 12\sqrt{2} \Leftrightarrow n = 288$

정답은 $288$

티칭 예리한 학생들은 의문을 갖는다.

⇨ 상쇄효과에 의하여 $pq = \hat{p}\hat{q}$이고, 위의 두 번의 추정에서 〈같은 모집단〉이므로 모비율은 당연히 같은 것인데 이 문제에서는 $\hat{p}_1\hat{q}_1$과 $\hat{p}_2\hat{q}_2$를 구분해야하기 때문이다. 앞에서 배운 이론에 따르면 $pq = \hat{p}_1\hat{q}_1 = \hat{p}_2\hat{q}_2$일 것 같은데 실제로는 같지 않다.

⇨ 〈추정〉 이전에 〈정규분포〉 자체가 〈확률의 근삿값〉을 구하는 것이다. 어차피 (고등수학이 아닌 역사적으로는)〈정규분포〉는 〈이항분포〉를 통해 유도되어 나온 것으로 〈이항분포〉에서 나오는 〈독립시행의 확률〉의 근삿값을 구하기 위해 나온 것이다. 그리고 라플라스에 의해 이것은 〈이항분포〉가 아닌 분포에 대해서도 활용할 수 있는 방법이 되었다. 즉, 통계는 정확한 값을 구하는 것이 아니다. (근사적으로 비슷한 값을 구하는 것이다.)

⇨ 〈추정〉 역시 정규분포로부터 자연스럽게 나오는 것으로 어차피 〈근사적인 수치들〉로 이루어져 있다.
$pq = \hat{p}\hat{q}$ 역시 $pq$가 주어지지 않기 때문에 어쩔 수 없이 $\hat{p}\hat{q}$을 대입하는 것이다.
- 이 문제처럼 구체적으로 나온 경우 $\hat{p}_1\hat{q}_1$을 $\hat{p}_2\hat{q}_2$을 구분해서 대입해야 이론상 조금 더 〈근접한 수치〉로 추정할 수 있게 된다. 그리고 적어도 시험을 보는 입장에서는 반드시 그래야만 한다. $\hat{p}_2\hat{q}_2$대신 $\hat{p}_1\hat{q}_1$을 대입하여 $n$값을 구해보면 $262.44$명이 나온다. - 이것은 당연히 잘못된 답이다.

# 예제를 통해 개념의 외연을 넓혀라

촌철살인 경우의 수 | PART 5 연속확률분포와 정규분포

개념의 외연 #1. 표본비율의 확률

<개념의 외연>이라는 부분은 현재 배우고 있는 경우의 수의 국한된 내용만 다루는 것은 아니다.
현재 단원에 해당하는 문제를 풀기 위해서 필요한 수학적 개념이나 유형에 필요한 사고방식을 두루 다룬다.
수학에는 순서가 없다. 하지만 배움에는 순서가 있다.
그래서 순서가 없는 수학에 순서를 정한 것이 교육과정이다. 어떤 내용은 너무 어렵거나 복합적인 생각들을 담고 있어서 여러 학년에 걸쳐 배운다. 이런 경우 보통은 문제를 통해서 익힌다는 느낌이 강한데 이런 개념들을 집약적으로 정리해 주는 부분이 될 것이다.

## #1 표본비율의 확률

### 1. 표본비율은 사실 이항분포

- <표본비율을 확률변수>로 인정하여 확률을 구하는 문제는 <이항분포>에서 <독립시행의 확률>을 구하는 과정과 같다.
  ⇨ 이항분포에서 독립시행의 확률은 계산이 불가능한 경우(n이 충분히 큰 경우) <정규분포>로 근사해서 구한다.
  계산이 가능한 경우는(n이 충분히 작은 경우) <독립시행의 확률> 그대로 계산한다.

- <표본비율의 분포>에서 <모비율>은 1회 시행에서 발생할 확률의 역할을 하고
  <표본의 크기 n + 복원시행>은 n회의 독립시행과 같은 의미를 가지므로
  결국 표현이 약간 바뀐 <이항분포>, <독립시행의 확률>을 풀게 된다.

- 2016년 현재 <표본비율>의 표현을 통해 <이항분포와 독립시행의 확률> 문제가 나온 적이 없어 그 동안의 기출문제를 변형하여 문제를 만들어 넣었음을 알린다.

**156** 이해를 위한 예제 2004. 4. 가형(49%). 14번. 3점

어느 대학의 2004학년도 합격자 1차 등록 비율은 75%이었다. 그 합격자들 중에서 임의로 192명을 뽑아 등록 여부를 조사하였을 때, 132명 이상이 등록했을 확률을 아래의 표준정규분포 표를 이용하여 구하면?

[ 표준정규분포 표 ]

| $z$ | $\mathrm{P}(0 \leq Z \leq z)$ |
|---|---|
| 0.5 | 0.1915 |
| 1.0 | 0.3413 |
| 1.5 | 0.4332 |
| 2.0 | 0.4772 |

① 0.6915 ② 0.7745 ③ 0.8413 ④ 0.9332 ⑤ 0.9772

<이항분포의 정분분포 화>문제를 <표본비율>처럼 풀어보기

표현의 차이일 뿐 본질적 차이가 아님을 알리기 위해 〈표본비율의 형식〉으로 풀어본다.

조건 어느 대학의 2004학년도 합격자 1차 등록 비율은 75% ⇨ 모비율을 $\frac{3}{4}$이라고 생각한다.

구할 것 그 합격자들 중에서 임의로 192명을 뽑아 등록 여부를 조사하였을 때, 132명 이상이 등록했을 확률

⇨ $n = 192$일 때, $P\left(\hat{p} \geq \frac{132}{192}\right)$을 물어본다고 생각해도 된다. 이때 $\hat{p}$은 $N\left(\frac{3}{4}, \left(\sqrt{\frac{\frac{3}{4}\times\frac{1}{4}}{192}}\right)^2\right)$

정규분포를 따른다. 즉, $P\left(\hat{p} \geq \frac{132}{192}\right) = P\left(\frac{\hat{p}-\frac{3}{4}}{\sqrt{\frac{\frac{3}{4}\times\frac{1}{4}}{192}}} \geq \frac{\frac{132}{192}-\frac{3}{4}}{\sqrt{\frac{\frac{3}{4}\times\frac{1}{4}}{192}}}\right) = P(Z \geq -2) = 0.9772$

정답은 ⑤

티칭 원래 풀이 - 이미 학습한 내용이므로 과정을 조금 생략한다.

⇨ 주어진 분포는 이항분포 $B\left(192, \frac{3}{4}\right)$를 따른다. 또한 $B\left(192, \frac{3}{4}\right)$은 근사적으로 정규분포 $N(144, 6^2)$을 따른다.

⇨ $P(X \geq 132) = P\left(\frac{X-144}{6} \geq \frac{132-144}{6}\right) = P(Z \geq -2) = 0.9772$

코칭 〈이항분포〉와 〈표본비율의 분포〉가 원리적으로 같다는 것을 설명하기 위해 〈표본비율〉처럼 풀어본 것.

**157** 이해를 위한 예제 2012. 10. 가형(78%). 11번. 3점

어느 과수원에서 수확한 사과의 무게는 평균 $400\,g$, 표준편차 $50\,g$인 정규분포를 따른다고 한다. 이 사과 중 무게가 $442\,g$ 이상인 것을 1등급 상품으로 정한다. 이 과수원에서 수확한 사과 중 100개를 임의로 선택할 때, 1등급 상품이 24개 이상일 확률을 다음 표준정규분포 표를 이용하여 구한 것은?

| $z$ | $\mathrm{P}(0 \le Z \le z)$ |
|---|---|
| 0.64 | 0.24 |
| 0.84 | 0.30 |
| 1.00 | 0.34 |
| 1.28 | 0.40 |

① 0.10 ② 0.16 ③ 0.20 ④ 0.26 ⑤ 0.34

<이항분포의 정분분포 화>문제를 <표본비율>처럼 풀어보기

표현의 차이일 뿐 본질적 차이가 아님을 알리기 위해 〈표본비율의 형식〉으로 풀어본다.

조건 사과의 무게는 평균 $400\,g$, 표준편차 $50\,g$인 정규분포를 따른다고 한다. 이 사과 중 무게가 $442\,g$ 이상인 것을 1등급 상품으로 정한다.

⇨ 사과의 무게 $X$라 하면 $X$는 $N(400,\ 50^2)$을 따른다. 사과 한 개를 뽑았을 때 일등급일 확률은

$$P(X \ge 442) = P\left(Z \ge \frac{442-400}{50}\right) = P(Z \ge 0.84) = 0.2$$

이것을 이 집단에서 1등급인 사과의 비율 즉, 모비율이라고 생각해보자. 즉, 모비율은 $\frac{1}{5}$

구할 것 수확한 사과 중 100개를 임의로 선택할 때, 1등급 상품이 24개 이상일 확률

⇨ $n = 100$일 때, $P\left(\hat{p} \ge \frac{24}{100}\right)$을 물어본다고 생각해도 된다.

이때 $\hat{p}$은 $N\left(\frac{1}{5},\left(\sqrt{\frac{\frac{1}{5}\times\frac{4}{5}}{100}}\right)^2\right)$ 즉, $N\left(\frac{1}{5},\left(\frac{1}{25}\right)^2\right)$인 정규분포를 따른다.

$$즉,\ P\left(\hat{p} \ge \frac{24}{100}\right) = P\left(\frac{\hat{p}-\frac{1}{5}}{\frac{1}{25}} \ge \frac{\frac{24}{100}-\frac{1}{5}}{\frac{1}{25}}\right) = P(Z \ge 1) = 0.16$$

정답은 ②

코칭 〈이항분포〉와 〈표본비율의 분포〉가 원리적으로 같다는 것을 설명하기 위해 〈표본비율〉처럼 풀어본 것.
왜 굳이 그렇게 풀어야 하냐고 반문하는 자들은 하수다. 완전한 이해를 위해서 〈원리적으로 같음〉을 이해하기 위해서 좀 더 복잡한 방식으로 풀어본 것이다.

**158** 이해를 위한 예제　이해를 위한 예제 151번 문제이다.

어느 지역에 거주하는 60대 이상의 주민들 중에서 백내장 환자의 비율은 36%라고 한다. 이 지역에 거주하는 60대 이상의 주민들 중에서 100명을 임의로 추출하였을 때, 그 중에 백내장 환자의 비율이 0.48 이상일 확률을 다음 표준정규분포 표를 이용하여 구한 것은?

| $z$ | $P(0 \le Z \le z)$ |
|---|---|
| 1.0 | 0.3413 |
| 1.5 | 0.4332 |
| 2.0 | 0.4772 |
| 2.5 | 0.4938 |

① 0.0062　② 0.0228　③ 0.0668　④ 0.1525　⑤ 0.1587

<표본비율>문제를 <이항분포의 정분분포 화>처럼 풀어보기

조건 60대 이상의 주민들 중에서 백내장 환자의 비율은 36%

⇨ 결국 모비율을 $p$라고 할 때 $p = 0.36$인데 이것은 1회 시행에서 발생할 확률이라고 인정해도 된다.

조건 60대 이상의 주민들 중에서 100명을 임의로 추출

⇨ $n = 100$은 〈표본의 크기〉가 100이라는 뜻인데 이것은 〈총 시행횟수가 100회이 독립시행〉으로 인정해도 좋다.

구할 것 백내장 환자의 비율이 0.48 이상일 확률

⇨ $P(\hat{p} \ge 0.48)$이지만 백내장 환자의 사람 수($X$)가 48명 이상일 확률을 구한다고 생각해도 좋다.

결국 $X$는 $B\left(100, \frac{36}{100}\right) \to N\left(36, \left(\frac{24}{5}\right)^2\right)$인 정규분포를 따르고, $P(X \ge 48)$을 구하는 것이므로

⇨ $P(X \ge 48) = P\left(\frac{X-36}{\frac{24}{5}} \ge \frac{48-36}{\frac{24}{5}}\right) = P(Z \ge 2.5) = 0.0062$이다.

정답은 ①

**159** 이해를 위한 예제 2007. 10. 나형(40%). 18번. 3점 - 변형

모비율이 $\frac{1}{2}$인 어떤 집단에서 크기가 $n$은 표본을 복원추출 하였다고 한다.

표본비율은 $\hat{p}$이라고 할 때 $\hat{p}$은 다음을 만족한다. $P\left(\hat{p}=\frac{2}{n}\right) = 10P\left(\hat{p}=\frac{1}{n}\right)$일 때, $n$의 값을 구하시오.

<이항분포> 문제를 <표본비율>의 표현을 이용하여 바꿔보았다.

조건 모비율이 $\frac{1}{2}$인 어떤 집단에서 크기가 $n$은 표본을 복원추출

⇨ $\hat{p} = \frac{X}{n}$에서 $X$는 $B(n,\ p)$를 따른다.

조건 표본비율은 $\hat{p}$이라고 할 때 $\hat{p}$은 다음을 만족한다. $P\left(\hat{p}=\frac{2}{n}\right) = 10P\left(\hat{p}=\frac{1}{2}\right)$

⇨ $\hat{p}$은 무조건 〈이항분포〉의 성격을 가진다. 하지만 경우에 따라 계산이 불가능한 경우 〈정규분포 화〉해서 확률을 근사적으로 구할 뿐이다. 조건에서 만약 $\hat{p}$을 정규분포 화 한다면 $P\left(\hat{p}=\frac{2}{n}\right)$와 $P\left(\hat{p}=\frac{1}{2}\right)$은 모두 $0$으로 계산된다.

(연속확률분포의 이론 상 특정 값에 대한 확률은 무조건 $0$이다.)

〈독립시행의 확률〉로 실제 값을 구할 경우 실제로도 거의 $0$에 가까울 정도로 작은 값이 나온다.

하지만 이 문제는 〈참 값〉을 구하는 관계식을 통해 $n$값을 구하라고 물어보고 있다.

즉, 이항분포를 〈정규분포 화〉시키지 않는다.

⇨ $\hat{p}$을 이산확률변수이므로 이를 통해 표를 그려보면

| $\hat{p}$ | $\frac{0}{n}$ | $\frac{1}{n}$ | $\cdots$ | $\frac{n}{n}$ | 총합 |
|---|---|---|---|---|---|
| $P(\hat{p})$ | ${}_nC_0\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^n$ | ${}_nC_1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^{n-1}$ | $\cdots$ | ${}_nC_n\left(\frac{1}{2}\right)^n\left(\frac{1}{2}\right)^0$ | 1 |

⇨ 즉, $P\left(\hat{p}=\frac{2}{n}\right) = 10P\left(\hat{p}=\frac{1}{2}\right) \Leftrightarrow {}_nC_2\left(\frac{1}{2}\right)^n = 10\,{}_nC_1\left(\frac{1}{2}\right)^n \Leftrightarrow {}_nC_2 = 10\,{}_nC_1$

$\Leftrightarrow \frac{n(n-1)}{2} = 10n$ 이므로 $n=21$

$n=21$

코칭 그냥 외우는 것이 아니라 〈표본비율〉의 상황과 〈이항분포〉의 상황을 정확히 이해하면 어떻게 변형되어 나와도 풀 수 있을 것이라고 확신한다. 〈표본비율〉의 상황은 곧 〈이항분포〉의 상황이다.

# 공부는 **예제**를 가지고 하는 거야

촌철살인 경우의 수 | PART 2 확률의 계산 1

일단 풀고 맞았다고 하더라도 뒤의 해설을 정독하여 읽는다. 왜냐하면 풀이과정이 암기되어서 문제가 그냥 풀리거나 그냥 직관적으로 판단이 되어서 문제가 풀리는 경우도 많기 때문이다. 우리가 문제를 푸는 이유는 앞에서 배운 개념을 확인하기 위해서다. 이 문제를 푸는 것이 중요한 게 아니라 다른 어려운 문제에까지 적용할 수 있는 개념을 익히는 게 중요한 것이다.

예제 074

2011. 11. 나형(67%). 16번. 4점

어느 공장에서 생산되는 제품의 길이 $X$는 평균이 $m$이고, 표준편차가 4인 정규분포를 따른다고 한다. $P(m \le X \le a) = 0.3413$일 때, 이 공장에서 생산된 제품 중에서 임의추출한 제품 16개의 길이의 표본평균이 $a-2$ 이상일 확률을 다음 표준정규분포 표를 이용하여 구한 것은?

(단, $a$는 상수이고, 길이의 단위는 cm이다.)

| $z$ | $P(0 \le Z \le z)$ |
|---|---|
| 1.0 | 0.3413 |
| 1.5 | 0.4332 |
| 2.0 | 0.4772 |

① 0.0228 ② 0.0668 ③ 0.0919 ④ 0.1359 ⑤ 0.1587

예제 075

2005. 11. 가형(52%), 나형(33%). 14번. 3점

어느 공장에서 생산되는 제품의 무게가 정규분포 $\mathrm{N}(11, 2^2)$을 따른다고 하자. $A$와 $B$두 사람이 크기가 $4$인 표본을 각각 독립적으로 임의 추출하였다. $A$와 $B$가 추출한 표본의 평균이 모두 $10$ 이상 $14$ 이하가 될 확률을 다음 표준정규분포 표를 이용하여 구한 것은?

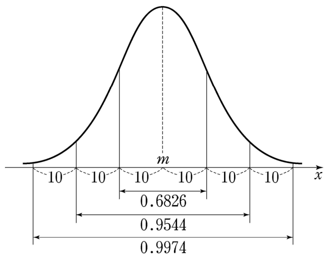
| $z$ | $\mathrm{P}(0 \le Z \le z)$ |
|---|---|
| 1 | 0.3413 |
| 2 | 0.4772 |
| 3 | 0.4987 |

① $0.8123$ ② $0.7056$ ③ $0.6587$ ④ $0.5228$ ⑤ $0.2944$

예제 076

2008. 9. 가형(72%). 나형(49%). 13번. 4점

어떤 모집단의 분포가 정규분포 $N(m,\ 10^2)$을 따르고, 이 정규분포의 확률밀도 함수 $f(x)$의 그래프와 구간별 확률은 다음과 같다. 확률밀도함수 $f(x)$는 모든 실수 $x$에 대하여 $f(x)=f(100-x)$를 만족한다. 이 모집단에서 크기 25인 표본을 임의추출 할 때의 표본평균을 $\overline{X}$라 하자. $P(44 \le \overline{X} \le 48)$의 값은?




① 0.1359 ② 0.1574 ③ 0.1965 ④ 0.2350 ⑤ 0.2718

예제 077

2004. 3. 가형(-%). 18번. 4점

$A$고등학교 학생의 몸무게는 평균이 60kg, 표준편차가 6kg인 정규분포를 이룬다고 한다. 적재중량이 549kg이상이 되면 경고음을 내도록 설계되어 있는 엘리베이터에 $A$고등학교 학생 중 임의 추출한 9명이 탑승하였을 때, 경고음이 울릴 확률은?

| $z$ | $P(0 \le Z \le z)$ |
|---|---|
| 0.5 | 0.1915 |
| 1.0 | 0.3413 |
| 1.5 | 0.4332 |
| 2.0 | 0.4772 |

① 0.1587 ② 0.1915 ③ 0.3085 ④ 0.3413 ⑤ 0.4332

예제 078

2011. 경찰대. 6번

어느 대민 봉사 센터의 전화상담의 통화 시간은 평균이 $8$분이고 표준편차가 $2$분인 정규분포를 따른다고 한다. 이 봉사 센터에 걸려오는 상담 전화 중 임의로 선택한 $4$통의 통화시간의 합이 $30$분 이상일 확률은?
(단, 다음 표준정규분포 표를 이용한다.)

| $z$ | $P(0 \le Z \le z)$ |
|---|---|
| 0.5 | 0.192 |
| 1.0 | 0.341 |
| 1.5 | 0.433 |

① $0.690$ ② $0.691$ ③ $0.692$ ④ $0.693$ ⑤ $0.694$

예제 079

어느 고등학교의 학생들이 한 달 동안 받는 스팸 메일의 개수는 표준편차가 10개인 정규분포를 따른다고 한다. 이 고등학교 학생 중 임의로 $n$명을 뽑아 한 달 동안 받는 스팸 메일의 개수를 조사 하였더니 평균이 15개이었을 때, 이 학교 학생들이 한 달 동안 받는 스팸 메일의 평균 개수 $m$을 신뢰도 95%로 추정한 신뢰구간이 $13 \le m \le 17$이었다. 이때, $n$의 값을 구하여라. (단, $\mathrm{P}(|Z| \leqq 2) = 0.95$)

예제 080

어떤 공장에서 생산되는 정수기의 필터의 수명은 표준편차가 350시간인 정규분포를 따른다고 한다.
여기서 생산한 필터의 평균 수명 시간을 신뢰도 95%로 추정한다고 할 때, 신뢰구간의 길이를 $200$이하로 하려면 표본의 크기를 얼마 이상으로 해야 하는지 구하여라. (단, $P(|Z| \le 1.96) = 0.95$이다.)

예제 081

2007. 4. 가형(40%). 21번. 3점
어떤 도시에 있는 전체 고등학교 학생들의 몸무게는 표준편차가 $5\mathrm{kg}$인 정규분포를 따른다고 한다. 이 도시의 고등학교 학생 전체에 대한 몸무게의 평균을 신뢰도 $95\%$로 추정할 때, 신뢰구간의 길이를 $1\mathrm{kg}$이하가 되도록 하려고 한다. 조사하여야 할 표본의 크기의 최솟값을 구하시오. (단, $\mathrm{P}(0 \le Z \le 1.96) = 0.4750$)

예제 082

2004. 11. 나형(69%). 13번. 3점

다음은 신뢰구간, 신뢰도, 표본의 크기의 관계를 설명한 것이다.

> 정규분포 $N(m, \sigma^2)$을 따르는 모집단이 있다. 이 모집단에서 크기 $n$인 표본을 임의추출하면 표본평균은 정규분포 $\boxed{(가)}$을 따른다. 이 표본평균의 분포를 이용하여 추정한 모평균 $m$에 대한 신뢰도 $\alpha$의 신뢰구간을 $a \le m \le b$라 하자. 표본의 크기를 $n$으로 고정하고 신뢰도를 $\alpha$보다 높게 한 신뢰구간을 $c \le m \le d$라 할 때, $d-c$는 $b-a$보다 $\boxed{(나)}$. 한편, 신뢰도를 $\alpha$로 고정하고 표본의 크기를 $2n$으로 한 신뢰구간을 $e \le m \le f$라 할 때, $f-e$는 $b-a$의 $\boxed{(다)}$배가 된다.

위의 과정에서 (가), (나), (다)에 알맞은 것은?

| | (가) | (나) | (다) |
|---|---|---|---|
| ① | $N(m, \sigma^2)$ | 크다 | $\frac{1}{2}$ |
| ② | $N(m, \sigma^2)$ | 작다 | $\frac{1}{2}$ |
| ③ | $N\left(m, \frac{\sigma^2}{n}\right)$ | 크다 | $\frac{1}{\sqrt{2}}$ |
| ④ | $N\left(m, \frac{\sigma^2}{n}\right)$ | 크다 | $\sqrt{2}$ |
| ⑤ | $N\left(m, \frac{\sigma^2}{n}\right)$ | 작다 | $\frac{1}{\sqrt{2}}$ |

예제 083

2008. 10. 가형(35%), 나형(24%). 22번. 3점 - 변형

모비율이 $p$인 집단에서 표본의 크기가 3이 표본을 임의로 복원 추출하였을 때, 표본비율을 $\hat{p}_1$이라고 하고, 모비율이 $2p$인 집단에서 표본의 크기가 4이 표본을 임의로 복원 추출하였을 때, 표본비율을 $\hat{p}_2$라고 하자.

이때, $10\mathrm{P}(0.8 \le \hat{p}_1 \le 1) = \mathrm{P}(\hat{\mathrm{p}}_2 \ge 0.75)$ 을 만족시키는 양수 $p$의 값은 $\dfrac{n}{m}$ 이다. $m+n$의 값을 구하시오. (단, $m$, $n$은 서로소인 자연수이다.)

예제 084

어느 도시는 전국체전을 유치하기 위해 임의로 시민 400명을 추출하여 여론을 조사하였더니 다음 표와 같은 결과를 얻었다. 이 도시의 전국체전 유치에 대해 찬성하는 사람의 비율 $p$를 신뢰도 95%로 추정한 신뢰구간이 $[a,\ b]$일 때, $b-a$의 값은? (단, $Z$가 표준정규분포를 따를 때, $P(|Z| \le 2) = 0.95$로 계산한다.)

| 조사결과 | | |
| --- | --- | --- |
| 찬성 | 반대 | 기권 |
| 360 | 34 | 6 |

① 0.052 ② 0.054 ③ 0.056 ④ 0.058 ⑤ 0.06

예제 085

어느 지역에서 300명을 임의추출하여 간염 검사를 하였더니 75명이 양성 반응을 일으켰다. 이 지역의 간염 양성 반응률의 신뢰도 95%의 신뢰구간은? (단, $P(|Z| \le 1.96) = 0.95$)

① $[0.229, 0.271]$ ② $[0.225, 0.275]$ ③ $[0.221, 0.279]$

④ $[0.216, 0.284]$ ⑤ $[0.201, 0.299]$

# 정답과 해설

예제 074

REVIEW 표본평균의 이해

2011. 11. 나형(67%). 16번. 4점

어느 공장에서 생산되는 제품의 길이 $X$는 평균이 $m$이고, 표준편차가 4인 정규분포를 따른다고 한다. $P(m \le X \le a) = 0.3413$일 때, 이 공장에서 생산된 제품 중에서 임의추출한 제품 16개의 길이의 표본평균이 $a-2$ 이상일 확률을 다음 표준정규분포 표를 이용하여 구한 것은?

(단, $a$는 상수이고, 길이의 단위는 cm이다.)

| $z$ | $P(0 \le Z \le z)$ |
|---|---|
| 1.0 | 0.3413 |
| 1.5 | 0.4332 |
| 2.0 | 0.4772 |

① 0.0228 ② 0.0668 ③ 0.0919 ④ 0.1359 ⑤ 0.1587

조건 제품의 길이 $X$는 평균이 $m$이고, 표준편차가 4인 정규분포를 따른다고 한다. $P(m \le X \le a) = 0.3413$

- $N(m, 4^2)$에서 $P(m \le X \le a) = 0.3413 \Leftrightarrow P\left(\frac{m-m}{4} \le \frac{X-m}{4} \le \frac{a-m}{4}\right) = 0.3413$

$\Leftrightarrow P\left(0 \le Z \le \frac{a-m}{4}\right) = 0.3413$ 즉, $\frac{a-m}{4} = 1 \Leftrightarrow a = m+4$

구할 것 임의 추출한 제품 16개의 길이의 표본평균이 $a-2$ 이상일 확률

- $\overline{X}$는 $N(m, 1^2)$인 정규분포를 따르므로

$$P(\overline{X} \ge a-2) = P(\overline{X} \ge (m+4)-2) = P\left(\frac{\overline{X}-m}{1} \ge \frac{m+2-m}{1}\right) = P(Z \ge 2) = 0.0228$$

정답은 ①

예제 075

REVIEW 표본평균의 이해, 확률의 곱셈정리 - 독립인 경우

2005. 11. 가형(52%), 나형(33%). 14번. 3점

어느 공장에서 생산되는 제품의 무게가 정규분포 $\mathrm{N}(11, 2^2)$을 따른다고 하자. $A$와 $B$ 두 사람이 크기가 4인 표본을 각각 독립적으로 임의 추출하였다. $A$와 $B$가 추출한 표본의 평균이 모두 10 이상 14 이하가 될 확률을 다음 표준정규분포 표를 이용하여 구한 것은?

| $z$ | $\mathrm{P}(0 \le \mathrm{Z} \le \mathrm{z})$ |
|---|---|
| 1 | 0.3413 |
| 2 | 0.4772 |
| 3 | 0.4987 |

① 0.8123 ② 0.7056 ③ 0.6587 ④ 0.5228 ⑤ 0.2944

조건 무게가 정규분포 $N(11, 2^2)$을 따른다고 하자. $A$와 $B$ 두 사람이 크기가 4인 표본을 각각 독립적으로 임의 추출

- $\overline{X}_A$와 $\overline{X}_B$는 각각 $N(11, 1^2)$인 정규분포를 따른다.

구할 것 $A$와 $B$가 추출한 표본의 평균이 모두 10이상 14이하가 될 확률

- $P(10 \le \overline{X}_A \le 14) = P\left(\frac{10-11}{1} \le \frac{\overline{X}_A - 11}{1} \le \frac{14-11}{1}\right) = P(-1 \le Z \le 3) = 0.3413 + 0.4987 = 0.84$

그런데 $\overline{X}_B$도 같은 분포를 따르므로 $P(10 \le \overline{X}_B \le 14)$ 역시 같은 값 0.84를 갖는다.

⇨ 사건 $10 \le \overline{X}_A \le 14$과 사건 $10 \le \overline{X}_B \le 14$는 독립이므로

$$P(\overline{X}_A, \overline{X}_B \text{ 모두 10이상 14이하}) = P(10 \le \overline{X}_A \le 14) \times P(10 \le \overline{X}_B \le 14) = 0.84 \times 0.84 = 0.7056$$

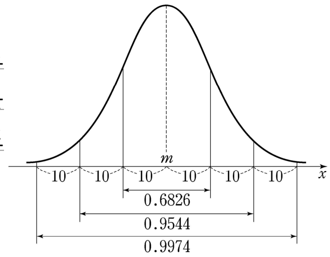
정답은 ②

예제 076

REVIEW 표본평균의 이해, 대칭함수 표현

2008. 9. 가형(72%). 나형(49%). 13번. 4점

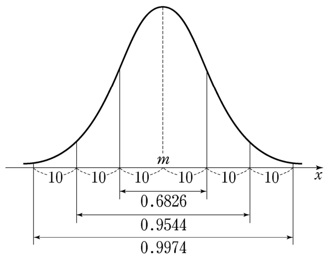
어떤 모집단의 분포가 정규분포 $N(m,\ 10^2)$을 따르고, 이 정규분포의 확률밀도함수 $f(x)$의 그래프와 구간별 확률은 다음과 같다. 확률밀도함수 $f(x)$는 모든 실수 $x$에 대하여 $f(x)=f(100-x)$를 만족한다. 이 모집단에서 크기 25인 표본을 임의추출 할 때의 표본평균을 $\overline{X}$라 하자. $P(44 \le \overline{X} \le 48)$의 값은?




① 0.1359 ② 0.1574 ③ 0.1965 ④ 0.2350 ⑤ 0.2718

조건 모집단의 분포가 정규분포 $N(m,\ 10^2)$을 따르고,
이 정규분포의 확률밀도함수 $f(x)$의 그래프와 구간별 확률은 다음과 같다.

- 그래프를 통해서 <정규분포 표>를 대신한다.
 이 문제는 굳이 표준화 할 필요가 없다.




조건 모든 실수 $x$에 대하여 $f(x)=f(100-x)$를 만족

- $m = 50$이라는 뜻이다. 즉, $X$는 $N(50, 10^2)$인 정규분포를 따른다.

구할 것 모집단에서 크기 25인 표본을 임의추출 할 때의 표본평균을 $\overline{X}$라 하자. $P(44 \le \overline{X} \le 48)$

- $P(44 \le \overline{X} \le 48) = P\left(\frac{44-50}{2} \le \frac{\overline{X}-50}{2} \le \frac{48-50}{2}\right) = P(-3 \le Z \le -1) = \frac{0.9974-0.6826}{2} = 0.1574$

정답은 ②

예제 077

REVIEW 표본평균의 이해, 추출된 대상의 총합

2004. 3. 가형(-%). 18번. 4점

$A$고등학교 학생의 몸무게는 평균이 60kg, 표준편차가 6kg인 정규분포를 이룬다고 한다. 적재중량이 549kg이상이 되면 경고음을 내도록 설계되어 있는 엘리베이터에 $A$고등학교 학생 중 임의 추출한 9명이 탑승하였을 때, 경고음이 울릴 확률은?

| $z$ | $P(0 \le Z \le z)$ |
|---|---|
| 0.5 | 0.1915 |
| 1.0 | 0.3413 |
| 1.5 | 0.4332 |
| 2.0 | 0.4772 |

① 0.1587 ② 0.1915 ③ 0.3085 ④ 0.3413 ⑤ 0.4332

조건 $A$고등학교 학생의 몸무게는 평균이 60kg, 표준편차가 6kg인 정규분포

- 모집단의 분포는 $N(60, 6^2)$ 인 정규분포를 따른다.

구할 것 적재중량이 549kg이상이 되면 경고음을 내도록 설계되어 있는 엘리베이터에 $A$고등학교 학생 중 임의 추출한 9명이 탑승하였을 때, 경고음이 울릴 확률

- A 고등학교에서 임의로 뽑은 9명의 학생의 몸무게의 평균을 $\overline{X}$라 할 때 9명의 몸무게의 합은 $9\overline{X}$이다.
 즉, $P(9\overline{X} \ge 549)$을 구하는 것이다.

⇨ $E(\overline{X}) = 60$, $\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{9}} = 2$이므로 $\overline{X}$는 정규분포 $N(60, 2^2)$ 을 따른다.

$$P(9\overline{X} \ge 549) = P(\overline{X} \ge 61) = P\left(Z \ge \frac{61-60}{2}\right) = P(Z \ge 0.5) = 0.5 - P(0 \le Z \le 0.5)$$
$$= 0.5 - 0.1915 = 0.3085$$

정답은 ③

예제 078

REVIEW 표본평균의 이해, 추출된 대상의 총합

2011. 경찰대. 6번

어느 대민 봉사 센터의 전화상담의 통화 시간은 평균이 8분이고 표준편차가 2분인 정규분포를 따른다고 한다. 이 봉사 센터에 걸려오는 상담 전화 중 임의로 선택한 4통의 통화시간의 합이 30분 이상일 확률은?
(단, 다음 표준정규분포 표를 이용한다.)

| $z$ | $P(0 \le Z \le z)$ |
|---|---|
| 0.5 | 0.192 |
| 1.0 | 0.341 |
| 1.5 | 0.433 |

① 0.690 ② 0.691 ③ 0.692 ④ 0.693 ⑤ 0.694

조건 통화 시간은 평균이 8분이고 표준편차가 2분인 정규분포

- 모집단의 분포는 $N(60, 6^2)$ 인 정규분포를 따른다.

구할 것 이 봉사 센터에 걸려오는 상담 전화 중 임의로 선택한 4통의 통화시간의 합이 30분 이상일 확률

- 상담 전화 중 임의로 선택한 4통의 통화시간의 평균을 $\overline{X}$ 라 할 때 임의로 선택한 4통의 통화시간의 합은 $4\overline{X}$ 이다.
  즉, $P(4\overline{X} \ge 30)$을 구하는 것이다.

⇨ $E(\overline{X}) = 8$, $\sigma(\overline{X}) = \dfrac{\sigma}{\sqrt{n}} = \dfrac{2}{\sqrt{4}} = 1$이므로 $\overline{X}$는 정규분포 $N(8, 1^2)$ 을 따른다.

$$P(4\overline{X} \ge 30) = P(\overline{X} \ge 7.5) = P\left(Z \ge \frac{7.5-8}{1}\right) = \mathrm{P}(Z \ge -0.5) = 0.5 + P(0 \le Z \le 0.5)$$
$$= 0.5 + 0.192 = 0.692$$

정답 ③

예제 079

REVIEW 모평균의 추정, 신뢰구간과 표본의 크기

어느 고등학교의 학생들이 한 달 동안 받는 스팸 메일의 개수는 표준편차가 10개인 정규분포를 따른다고 한다. 이 고등학교 학생 중 임의로 $n$명을 뽑아 한 달 동안 받는 스팸 메일의 개수를 조사 하였더니 평균이 15개이었을 때, 이 학교 학생들이 한 달 동안 받는 스팸 메일의 평균 개수 $m$을 신뢰도 95%로 추정한 신뢰구간이 $13 \le m \le 17$이었다. 이때, $n$의 값을 구하여라. (단, $\mathrm{P}(|Z| \le 2) = 0.95$)

조건 학생들이 한 달 동안 받는 스팸 메일의 개수는 표준편차가 10개인 정규분포 - 모 표준편차는 10이다.

조건 임의로 $n$명을 뽑아 한 달 동안 받는 스팸 메일의 개수를 조사 하였더니 평균이 15개

- 표본의 크기는 $n$이고, $\overline{X} = 15$이다.

조건 $\mathrm{P}(|Z| \le 2) = 0.95$ - 95%추정에서 신뢰계수 $k$는 2이라는 뜻이다.

구할 것 평균 개수 $m$을 신뢰도 95%로 추정한 신뢰구간이 $13 \le m \le 17$이었다. 이때, $n$의 값

⇨ 주어진 정보를 이용하여 신뢰도 95%에서 모평균을 추정하여 신뢰구간을 구하면

$$15 - 2\frac{10}{\sqrt{n}} \le m \le 15 + 2\frac{10}{\sqrt{n}} \Leftrightarrow 13 \le m \le 17\text{이므로 } 15 - 2\frac{10}{\sqrt{n}} = 13 \Leftrightarrow n = 100$$

정답은 100

REVIEW 모평균의 추정, 신뢰구간의 길이와 표본의 크기

예제 080

어떤 공장에서 생산되는 정수기의 필터의 수명은 표준편차가 350시간인 정규분포를 따른다고 한다.
여기서 생산한 필터의 평균 수명 시간을 신뢰도 95%로 추정한다고 할 때, 신뢰구간의 길이를 $200$이하로 하려면 표본의 크기를 얼마 이상으로 해야 하는지 구하여라. (단, $P(|Z| \leq 1.96) = 0.95$이다.)

조건 정수기의 필터의 수명은 표준편차가 350시간인 정규분포 - 모 표준편차는 $350$이다.

조건 $P(|Z| \leq 1.96) = 0.95$ - $95\%$추정에서 신뢰계수 $k$는 $1.96$이라는 뜻이다.

구할 것 여기서 생산한 필터의 평균 수명 시간을 신뢰도 95%로 추정한다고 할 때, 신뢰구간의 길이를 $200$이하로 하려면 표본의 크기를 얼마 이상

⇨ 신뢰구간의 길이는 신뢰구간의 양 끝을 뺀 값이므로 $2k\frac{\sigma}{\sqrt{n}}$이다.

즉, $2 \times 1.96\frac{350}{\sqrt{n}} \leq 200 \quad \Leftrightarrow \quad 6.86 \leq \sqrt{n} \quad \Leftrightarrow \quad 47.0596 \leq n$ 결국 $n$은 $48$이상이어야 한다.

정답은 $48$

REVIEW 모평균의 추정, 신뢰구간의 길이와 표본의 크기

예제 081

2007. 4. 가형(40%). 21번. 3점

어떤 도시에 있는 전체 고등학교 학생들의 몸무게는 표준편차가 $5\mathrm{kg}$인 정규분포를 따른다고 한다. 이 도시의 고등학교 학생 전체에 대한 몸무게의 평균을 신뢰도 $95\%$로 추정할 때, 신뢰구간의 길이를 $1\mathrm{kg}$이하가 되도록 하려고 한다. 조사하여야 할 표본의 크기의 최솟값을 구하시오. (단, $\mathrm{P}(0 \leq Z \leq 1.96) = 0.4750$)

조건 고등학교 학생들의 몸무게는 표준편차가 $5\mathrm{kg}$인 정규분포 - 모 표준편차는 $5$이다.

조건 $\mathrm{P}(0 \leq Z \leq 1.96) = 0.4750$ - $95\%$추정에서 신뢰계수 $k$는 $1.96$이라는 뜻이다.

구할 것 몸무게의 평균을 신뢰도 $95\%$로 추정할 때, 신뢰구간의 길이를 $1\mathrm{kg}$이하가 되도록 하려고 한다. 조사하여야 할 표본의 크기의 최솟값

⇨ 신뢰구간의 길이는 신뢰구간의 양 끝을 뺀 값이므로 $2k\frac{\sigma}{\sqrt{n}}$이다.

즉, $3.92 \times \frac{5}{\sqrt{n}} \leq 1 \quad \Leftrightarrow \quad 19.60 \leq \sqrt{n} \quad \Leftrightarrow \quad 384.16 \leq n$ 결국 $n$은 $385$이상이어야 한다.

즉, 표본의 크기의 최솟값은 $385$

예제 082

REVIEW 신뢰도, 신뢰구간, 신뢰구간의 크기의 의미

2004. 11. 나형(69%). 13번. 3점

다음은 신뢰구간, 신뢰도, 표본의 크기의 관계를 설명한 것이다.

정규분포 $N(m, \sigma^2)$을 따르는 모집단이 있다. 이 모집단에서 크기 $n$인 표본을 임의추출하면 표본평균은 정규분포 $\boxed{(가)}$을 따른다. 이 표본평균의 분포를 이용하여 추정한 모평균 $m$에 대한 신뢰도 $\alpha$의 신뢰구간을 $a \le m \le b$라 하자. 표본의 크기를 $n$으로 고정하고 신뢰도를 $\alpha$보다 높게 한 신뢰구간을 $c \le m \le d$라 할 때, $d-c$는 $b-a$보다 $\boxed{(나)}$. 한편, 신뢰도를 $\alpha$로 고정하고 표본의 크기를 $2n$으로 한 신뢰구간을 $e \le m \le f$라 할 때, $f-e$는 $b-a$의 $\boxed{(다)}$배가 된다.

위의 과정에서 (가), (나), (다)에 알맞은 것은?

| | (가) | (나) | (다) |
|---|---|---|---|
| ① | $N(m, \sigma^2)$ | 크다 | $\frac{1}{2}$ |
| ② | $N(m, \sigma^2)$ | 작다 | $\frac{1}{2}$ |
| ③ | $N\left(m, \frac{\sigma^2}{n}\right)$ | 크다 | $\frac{1}{\sqrt{2}}$ |
| ④ | $N\left(m, \frac{\sigma^2}{n}\right)$ | 크다 | $\sqrt{2}$ |
| ⑤ | $N\left(m, \frac{\sigma^2}{n}\right)$ | 작다 | $\frac{1}{\sqrt{2}}$ |

조건 (가)는 공식으로 기억하는 것이다. ⇨ $\overline{X}$는 $N\left(m, \frac{\sigma^2}{n}\right)$인 정규분포를 따른다. 즉, (가)에 들어갈 말은 $N\left(m, \frac{\sigma^2}{n}\right)$

조건 모평균 $m$에 대한 신뢰도 $\alpha$의 신뢰구간을 $a \le m \le b$라 하자. 표본의 크기를 $n$으로 고정하고 신뢰도를 $\alpha$보다 높게 한 신뢰구간을 $c \le m \le d$라 할 때, $d-c$는 $b-a$보다 $\boxed{(나)}$.

⇨ $d-c$와 $b-a$은 신뢰구간의 길이를 의미한다. 신뢰도를 높이면 $k$값이 커지므로 신뢰구간의 길이도 길어진다.
즉, (나)에 들어갈 말은 <크다>

조건 신뢰도를 $\alpha$로 고정하고 표본의 크기를 $2n$으로 한 신뢰구간을 $e \le m \le f$라 할 때, $f-e$는 $b-a$의 $\boxed{(다)}$배가 된다.

⇨ 신뢰도가 $\alpha$일 때의 신뢰계수를 $k$라 하면 $b-a = 2k\frac{\sigma}{\sqrt{n}}$이다. $f-e = 2k\frac{\sigma}{\sqrt{2n}}$이므로 $b-a$의 $\frac{1}{\sqrt{2}}$배이다.

즉, (다)에 들어갈 말은 $\frac{1}{\sqrt{2}}$

정답은 ③

예제 083

REVIEW 표본비율의 이해 - 이항분포임을 이해 (기출 변형 문제)

2008. 10. 가형(35%), 나형(24%). 22번. 3점 - 변형

모비율이 $p$인 집단에서 표본의 크기가 3이 표본을 임의로 복원 추출하였을 때, 표본비율을 $\hat{p}_1$이라고 하고,

모비율이 $2p$인 집단에서 표본의 크기가 4이 표본을 임의로 복원 추출하였을 때, 표본비율을 $\hat{p}_2$라고 하자.

이때, $10\mathrm{P}(0.8 \le \hat{p}_1 \le 1) = \mathrm{P}(\hat{p}_2 \ge 0.75)$을 만족시키는 양수 $p$의 값은 $\frac{n}{m}$이다. $m+n$의 값을 구하시오.

(단, $m$, $n$은 서로소인 자연수이다.)

조건 모비율이 $p$인 집단에서 표본의 크기가 3이 표본을 임의로 복원 추출하였을 때, 표본비율을 $\hat{p}_1$

⇨ 표본의 크기가 작으므로 <이항분포 표>를 그대로 활용하며 문제를 푼다. - 머릿속에 있다면 굳이 그리지 않아도 좋다.

| $\hat{p}_1$ | $\frac{0}{3}$ | $\frac{1}{3}$ | $\frac{2}{3}$ | $\frac{3}{3}$ | 총합 |
|---|---|---|---|---|---|
| $P(\hat{p}_1)$ | ${}_3C_0p^0(1-p)^3$ | ${}_3C_1p^1(1-p)^2$ | ${}_3C_2p^2(1-p)^1$ | ${}_3C_3p^3(1-p)^0$ | 1 |

조건 모비율이 $2p$인 집단에서 표본의 크기가 4이 표본을 임의로 복원 추출하였을 때, 표본비율을 $\hat{p}_2$

⇨ 표본의 크기가 작으므로 <이항분포 표>를 그대로 활용하며 문제를 푼다.

| $\hat{p}_2$ | $\frac{0}{4}$ | $\frac{1}{4}$ | $\frac{2}{4}$ | $\frac{3}{4}$ | $\frac{4}{4}$ | 총합 |
|---|---|---|---|---|---|---|
| $P(\hat{p}_1)$ | ${}_4C_02p^0(1-2p)^4$ | ${}_4C_12p^1(1-2p)^3$ | ${}_4C_22p^2(1-2p)^2$ | ${}_4C_32p^3(1-2p)^1$ | ${}_4C_42p^4(1-2p)^0$ | 1 |

구할 것 $10\mathrm{P}(0.8 \le \hat{p}_1 \le 1) = \mathrm{P}(\hat{p}_2 \ge 0.75)$을 만족시키는 양수 $p$의 값은 $\frac{n}{m}$

⇨ 위의 표에서 $10P(0.8 \le \hat{p}_1 \le 1)$은 결국 $10P\left(\hat{p}_1 = 1\right)$을 구하는 것과 같다. 즉, $10P\left(\hat{p}_1 = 1\right) = 10\,{}_3C_3p^3(1-p)^0$

위의 표에서 $P(\hat{p}_2 \ge 0.75)$은 결국 $P\left(\hat{p}_2 = \frac{3}{4}\right) + P\left(\hat{p}_2 = \frac{4}{4}\right)$을 구하는 것과 같다.

즉, $P(\hat{p}_2 \ge 0.75) = P\left(\hat{p}_2 = \frac{3}{4}\right) + P\left(\hat{p}_2 = \frac{4}{4}\right) = {}_4C_32p^3(1-2p)^1 + {}_4C_42p^4(1-2p)^0$이므로

$10\,{}_3C_3p^3(1-p)^0 = {}_4C_32p^3(1-2p)^1 + {}_4C_42p^4(1-2p)^0 \quad \Leftrightarrow \quad 10p^3 = 16p^3(2-3p) \quad \Leftrightarrow \quad p = \frac{11}{24}$

즉, $m+n=35$ 이므로 정답은 35

예제 084

REVIEW 모비율 추정

어느 도시는 전국체전을 유치하기 위해 임의로 시민 $400$명을 추출하여 여론을 조사하였더니 다음 표와 같은 결과를 얻었다. 이 도시의 전국체전 유치에 대해 찬성하는 사람의 비율 $p$를 신뢰도 $95\%$로 추정한 신뢰구간이 $[a,\ b]$일 때, $b-a$의 값은? (단, $Z$가 표준정규분포를 따를 때, $P(|Z| \le 2) = 0.95$로 계산한다.)

| 조사결과 | | |
|---|---|---|
| 찬성 | 반대 | 기권 |
| 360 | 34 | 6 |

① $0.052$ ② $0.054$ ③ $0.056$ ④ $0.058$ ⑤ $0.06$

조건 전국체전을 유치하기 위해 임의로 시민 $400$명을 추출하여 여론을 조사했더니 찬성하는 사람 $360$명

- 표본비율을 조건으로 준 것이다. 즉, $\hat{p} = \dfrac{360}{400} = 0.9$이다.

조건 $P(|Z| \le 2) = 0.95$로 계산 - $95\%$추정에서 신뢰계수 $k$는 $2$라는 뜻이다.

구할 것 찬성하는 사람의 비율 $p$를 신뢰도 $95\%$로 추정한 신뢰구간이 $[a,\ b]$일 때, $b-a$의 값

- 결국 신뢰구간의 길이를 구하는 것이므로 $2k\sqrt{\dfrac{\hat{p}\,\hat{q}}{n}}$ 을 구한다.

⇨ 위의 조건에서 $k = 2$, $n = 400$, $\hat{p} = 0.9$이므로 $2k\sqrt{\dfrac{\hat{p}\,\hat{q}}{n}} = 4\sqrt{\dfrac{0.9 \times 0.1}{400}} = 4 \times \dfrac{0.3}{20} = 0.06$

정답은 ⑤

예제 085

REVIEW 모비율 추정

어느 지역에서 $300$명을 임의추출하여 간염 검사를 하였더니 $75$명이 양성 반응을 일으켰다. 이 지역의 간염 양성 반응률의 신뢰도 $95\%$의 신뢰구간은? (단, $P(|Z| \le 1.96) = 0.95$)

① $[0.229,\ 0.271]$ ② $[0.225,\ 0.275]$ ③ $[0.221,\ 0.279]$
④ $[0.216,\ 0.284]$ ⑤ $[0.201,\ 0.299]$

조건 어느 지역에서 $300$명을 임의추출하여 간염 검사를 하였더니 $75$명이 양성 반응

- 표본비율을 조건으로 준 것이다. 즉, $\hat{p} = \dfrac{75}{100} = 0.25$이다.

조건 $P(|Z| \le 1.96) = 0.95$ - $95\%$추정에서 신뢰계수 $k$는 $2$라는 뜻이다.

구할 것 이 지역의 간염 양성 반응률의 신뢰도 $95\%$의 신뢰구간

⇨ 위의 조건에서 $k = 1.96$, $n = 300$, $\hat{p} = 0.25$이므로

$$\left[0.25 - 1.96\sqrt{\frac{0.25 \times 0.75}{300}},\ 0.25 + 1.96\sqrt{\frac{0.25 \times 0.75}{300}}\right] = [0.25 - 1.96 \times 0.025,\ 0.25 + 1.96 \times 0.025]$$

$$= [0.201,\ 0.299]$$

정답 ⑤

# MAP

# PART 02 확률의 계산 1



# PART 02 확률의 계산 1

## PART 03 확률의 계산 2



## PART 03 확률의 계산 2

# PART 04 이산확률분포와 이항분포



# PART 04 분할의 상황

# PART 05 연속확률분포와 정규분포



# PART 05 연속확률분포와 정규분포

## PART 06 통계적 추정



## PART 06 통계적 추정

존철살인 경우의 수

본 책의 PDF파일의 가격은 25000원으로 책정 되었습니다.

# 촌철살인

## 확률과 통계

수학은 직관에 의존한 논리다.

값 25000원
55410
9 788968 194498
ISBN 978-89-6819-449-8 (PDF)

티칭과 코칭