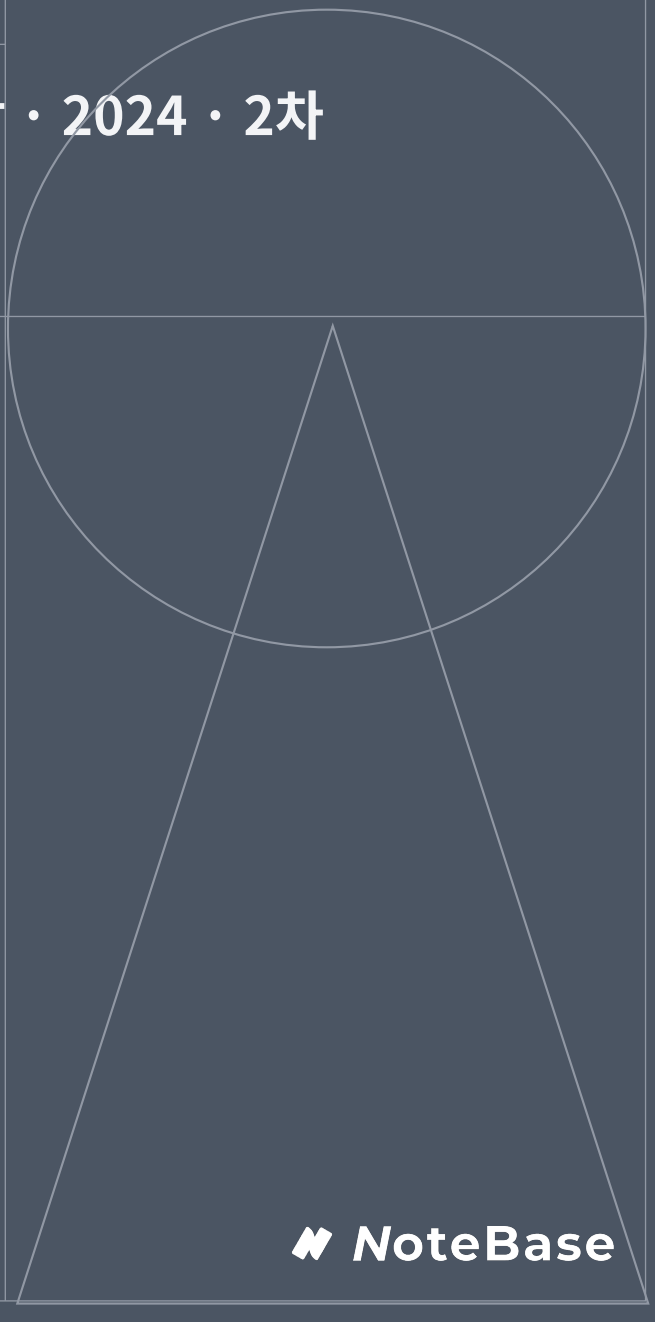
기출OX 해설집

공무원/국가직/7급 · 통계학 · 2024 · 2차



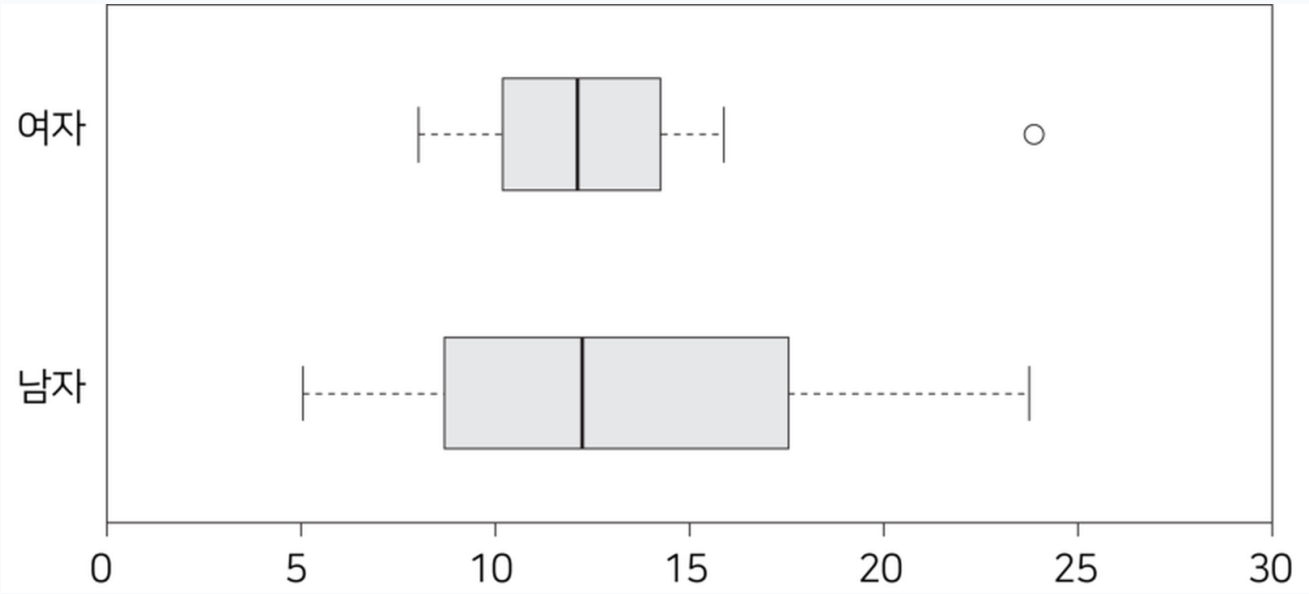
NoteBase

기출OX 해설집

공무원/국가직/7급 · 통계학 · 2024 · 2차

공무원/국가직/7급 · 통계학 · 2024 · 2차

문 1. 다음 그림은 어떤 자료를 성별로 요약한 상자그림을 나타낸다. 이에 대한 설명으로 옳지 않은 것은?



① 최솟값은 남자가 여자보다 작다.

② 여자의 자료는 이상점(outlier)을 가진다.

③ 범위(range)는 남자가 여자보다 크다.

④ 사분위수범위(IQR)는 남자가 여자보다 작다.

정답 4번

출제 포인트

이 문제는 상자그림에서 최소값, 이상점, 범위, 사분위수범위를 각각 읽어 비교할 수 있는지를 묻는다.

자료 읽는 순서

1. 왼쪽 수염 끝과 오른쪽 끝 또는 이상점으로 최소값과 최댓값을 확인한다.
2. 상자의 길이로 사분위수범위를 비교하고, 수염 밖의 점으로 이상점 여부를 확인한다.
3. 각 선지가 최소값, 범위, 사분위수범위 중 무엇을 묻는지 자료 요소와 연결한다.

NoteBase

선지별 해설

1번 O: 남자의 왼쪽 수염 끝이 여자의 왼쪽 수염 끝보다 더 왼쪽에 있으므로 남자의 최소값이 더 작다.
2번 O: 여자 상자그림의 오른쪽에 수염과 떨어진 점이 하나 표시되어 있으므로 이상점이 존재한다.
3번 O: 범위는 최댓값에서 최솟값을 뺀 값이다. 남자는 대략 5에서 24까지이고, 여자는 이상점을 포함해 대략 8에서 24까지이므로 남자의 범위가 더 크다.
4번 X: 사분위수범위는 상자의 길이이다. 남자 상자가 여자 상자보다 더 길므로 남자의 사분위수범위가 더 크다. 틀린 표현은 '남자가 여자보다 작다'이고, 올바르게는 '사분위수범위는 남자가 여자보다 크다'이다.

오답 포인트

범위를 수염 길이만으로 보거나, 사분위수범위를 전체 폭으로 착각하면 틀리기 쉽다. 이 문항은 이상점은 범위 판단에 포함되지만 사분위수범위 판단에는 상자 길이만 쓴다는 점을 구분하는지가 함정이다.

다시 나오면 이렇게 판단

상자그림 문제는 수염 끝은 최소값과 일반적인 최대값, 떨어진 점은 이상점, 상자 길이는 사분위수범위라는 대응을 먼저 확정한 뒤 선지의 비교 대상을 읽으면 된다.

암기 포인트

상자 길이는 사분위수범위이고, 수염 밖의 점은 이상점이며, 범위는 최댓값과 최솟값의 차이이다.

NoteBase

문 2. 어느 학생이 지난 학기에 수강한 4과목의 취득점수, 그리고 과목별 전체 평균과 표준편차가 다음과 같다. 각 과목에서 A학점을 받기 위해서는 상위 16% 안에 들어야 한다고 할 때, 이 학생이 수강한 과목 중 A학점을 받은 과목으로만 나열한 것은? (단, 모든 과목의 성적은 각각 정규분포를 따르고, Z가 $N(0,1)$을 따를 때 $P(-1 \le Z \le 1) = 0.68$이다)

수강과목		통계학	미적분학	영어	경제학
취득점수		85	87	90	92
전체	평균	78	82	85	90
	표준편차	5	6	4	3

① 통계학, 미적분학

② 통계학, 영어

③ 미적분학, 경제학

④ 영어, 경제학

정답 2번

출제 포인트

이 문제는 각 과목의 표준점수 $z = \frac{x-\mu}{\sigma}$를 계산하여 상위 16%에 드는 과목만 고르는 문제이다.

조건 정리

주어진 값은 과목별 취득점수, 평균, 표준편차이다.
구해야 할 값은 A학점을 받은 과목의 목록이다.
상위 16%는 정규분포에서 $z \ge 1$인 구간에 해당한다.

사용할 공식

표준점수 공식은 $z = \frac{x-\mu}{\sigma}$이다.
이 공식을 쓰는 이유는 과목마다 평균과 표준편차가 다르므로 원점수를 그대로 비교하면 안 되고, 평균에서 몇 표준편차만큼 위에 있는지를 같은 기준으로 바꾸어야 하기 때문이다.

NoteBase

계산 과정

통계학은 $z = \frac{85-78}{5} = \frac{7}{5} = 1.4$이므로 A학점이다.
미적분학은 $z = \frac{87-82}{6} = \frac{5}{6} \approx 0.83$이므로 A학점이 아니다.
영어는 $z = \frac{90-85}{4} = \frac{5}{4} = 1.25$이므로 A학점이다.
경제학은 $z = \frac{92-90}{3} = \frac{2}{3} \approx 0.67$이므로 A학점이 아니다.
따라서 A학점을 받은 과목은 통계학과 영어이다.

선지별 해설

① X: 통계학은 $z = 1.4$로 맞지만, 미적분학은 $z \approx 0.83$으로 상위 16%에 들지 못한다. 틀린 표현은 "미적분학"이며, 올바르게는 "영어"이다.
② O: 통계학은 $z = 1.4$, 영어는 $z = 1.25$로 둘 다 1 이상이므로 상위 16%에 해당한다.
③ X: 미적분학은 $z \approx 0.83$, 경제학은 $z \approx 0.67$이므로 둘 다 A학점이 아니다. 틀린 표현은 "미적분학, 경제학"이며, 올바르게는 "통계학, 영어"이다.
④ X: 영어는 $z = 1.25$로 맞지만, 경제학은 $z \approx 0.67$이어서 상위 16%가 아니다. 틀린 표현은 "경제학"이며, 올바르게는 "통계학"이다.

오답 포인트

원점수가 높은 과목을 그대로 고르면 경제학 92점 때문에 흔들리기 쉽다. 이 문항의 함정은 원점수 비교가 아니라 평균과 표준편차를 반영한 상대적 위치 비교라는 점이다.

다시 나오면 이렇게 판단

정규분포에서 상위 몇 퍼센트인지 묻는 문항은 먼저 해당 백분위에 대응하는 z 기준을 정한 뒤, 각 대상을 $z = \frac{x-\mu}{\sigma}$로 바꾸어 기준 이상인지 비교하면 된다.

암기 포인트

A학점처럼 상대평가를 묻는 문항은 원점수가 아니라 z값으로 판단한다.

NoteBase

문 3. 어느 중학교에는 500명의 학생이 재학 중이다. 인터넷 사용 시간에 대한 조사를 수행하는 한 연구자가 성별에 따라 사용 시간에 차이가 큰 것을 알고 전체 학생을 남학생과 여학생으로 구분한 후, 남학생과 여학생의 명부로부터 각각 30명씩 무작위로 추출하였다. 이에 해당하는 확률표본추출방법은?

① 집락추출법(cluster sampling)

② 층화추출법(stratified sampling)

③ 할당추출법(quota sampling)

④ 계통추출법(systematic sampling)

정답 2번

출제 포인트

이 문제는 모집단을 성별이라는 기준으로 층으로 나눈 뒤 각 층에서 무작위 추출하는 층화추출법과 다른 표본추출법을 구별하는 문제이다.

핵심 개념

층화추출법은 모집단을 성질이 비슷한 여러 층으로 나누고 각 층에서 표본을 뽑는 방법이다.
집락추출법은 반, 마을처럼 자연스럽게 묶인 집락을 뽑아 그 집락을 조사하는 방법이다.
할당추출법은 성별·연령별 인원수만 미리 정해 두고 조사자가 비확률적으로 채우는 방법이다.
계통추출법은 명단에서 일정 간격마다 한 명씩 뽑는 방법이다.

선지별 해설

① X: 집락추출법은 몇 개의 집락 자체를 뽑는 방법인데, 이 문제는 남학생 집단과 여학생 집단을 모두 남겨 둔 채 각 집단 안에서 30명씩 무작위로 뽑았다. 틀린 부분은 '집락추출법'이라는 표현이다. 올바르게는 '성별로 층을 나누고 각 층에서 추출하는 층화추출법'이다.
② O: 성별에 따라 인터넷 사용 시간이 다를 수 있다고 보고 남학생층과 여학생층으로 나눈 뒤 각 층에서 무작위로 30명씩 뽑았으므로 층화추출법에 해당한다.
③ X: 할당추출법은 남 30명, 여 30명처럼 인원수만 정해 두고 조사자가 임의로 채우는 비확률표본추출이다. 이 문제는 명부에서 '무작위로' 추출했으므로 할당추출법이 아니다. 틀린 부분은 '할당추출법'이라는 표현이다. 올바르게는 '무작위 표집이 들어간 층화추출법'이다.
④ X: 계통추출법은 명단에서 임의의 시작점을 정한 뒤 매 k번째마다 뽑는 방법이다. 이 문제에는 일정 간격 선택이 없고 성별로 층을 나눈 절차가 핵심이다. 틀린 부분은 '계통추출법'이라는 표현이다. 올바르게는 '각 층에서 무작위 추출하는 층화추출법'이다.

NoteBase

오답 포인트

남학생 30명, 여학생 30명이라는 문구만 보면 할당추출법으로 착각하기 쉽다. 그러나 할당추출법과 층화추출법의 결정적 차이는 '무작위 추출 여부'이다. 또한 성별 집단을 둘 다 조사 대상으로 유지했다는 점에서 일부 집단만 뽑는 집락추출법과도 다르다.

다시 나오면 이렇게 판단

먼저 모집단을 부분집단으로 나누었는지 본다. 나누었다면 모든 부분집단에서 각각 뽑는지, 일부 묶음만 통째로 뽑는지 구분한다. 모든 층에서 무작위로 뽑으면 층화추출법이고, 일부 집락만 뽑으면 집락추출법이며, 인원수만 맞추고 무작위가 없으면 할당추출법이고, 일정 간격마다 뽑으면 계통추출법이다.

암기 포인트

층화는 '나누고 각각 뽑기', 집락은 '묶음을 뽑기', 할당은 '숫자만 채우기', 계통은 '간격대로 뽑기'이다.

NoteBase

문 4. 분산분석에 대한 다음 설명 중 옳은 것만을 모두 고르면?

ㄱ. 일원배치법에서 각 수준에서의 반복수가 반드시 같아야 한다.
ㄴ. 반복이 있는 이원배치법에서 두 인자 A, B 각각의 제곱합(SS_A, SS_B)의 자유도는 교호작용 제곱합($SS_{A\times B}$)의 자유도보다 항상 작다.
ㄷ. 이원배치법에서 교호작용의 존재 여부를 파악하고자 할 경우에는 반복이 있는 이원배치법을 사용해야 한다.

① ㄷ
② ㄱ, ㄴ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 1번

출제 포인트

이 문제는 일원배치법의 반복수 조건, 반복이 있는 이원배치법의 자유도, 교호작용 검정 가능 조건을 구별할 수 있는지를 묻는다.

핵심 개념

일원배치법은 여러 수준의 평균 차이를 비교하는 분산분석이며 각 수준의 표본수가 반드시 같을 필요는 없다.
반복이 있는 이원배치법은 각 조합 셀마다 반복관측이 있어 주효과와 교호작용, 오차를 분리할 수 있다.
반복이 있는 이원배치법의 자유도는 $df_A = a-1$, $df_B = b-1$, $df_{A\times B} = (a-1)(b-1)$이다.

선지별 해설

1번 O: 1번은 ㄷ만 옳다고 본 선택지이다. 반복이 없는 이원배치법에서는 교호작용과 오차가 뒤섞여 교호작용을 따로 검정할 수 없으므로, 교호작용의 존재 여부를 보려면 반복이 있는 이원배치법을 사용해야 한다.
2번 X: 2번은 ㄱ, ㄴ을 옳다고 본 점이 틀리다. ㄱ의 틀린 표현은 "반드시"이다. 일원배치법은 각 수준의 반복수가 같으면 균형설계가 되어 편리하지만 반드시 같아야 하는 것은 아니다. ㄴ도 틀리다. $df_{A\times B} = (a-1)(b-1)$이므로 다른 인자의 수준 수가 2이면 주효과 자유도와 같아질 수 있어 "항상 작다"고 할 수 없다. 올바르게는 "ㄷ만 옳다"이다.
3번 X: 3번은 ㄷ은 옳지만 ㄴ을 함께 옳다고 본 점이 틀리다. ㄴ의 틀린 표현은 "항상 작다"이다. 주효과 자유도는 $a-1, b-1$이고 교호작용 자유도는 $(a-1)(b-1)$이므로 항상 더 작은 것은 아니다. 올바르게는 "ㄷ만 옳다"이다.
4번 X: 4번은 ㄱ, ㄴ, ㄷ을 모두 옳다고 본 점이 틀리다. ㄱ은 반복수가 반드시 같아야 한다고 단정한 부분이 틀리고, ㄴ은 주효과 자유도가 교호작용 자유도보다 항상 작다고 단정한 부분이 틀리다. ㄷ만 옳다. 올바르게는 "ㄷ만 옳다"이다.

NoteBase

오답 포인트

ㄱ의 "반드시"와 ㄴ의 "항상"이 함정이다. 분산분석 문제는 절대적 표현이 붙는 순간 조건식을 다시 확인해야 하는데, 균형설계의 편의성과 필수 조건을 혼동하거나 $(a-1)(b-1)$ 형태를 보지 않으면 ㄴ을 맞는 말처럼 받아들이기 쉽다.

다시 나오면 이렇게 판단

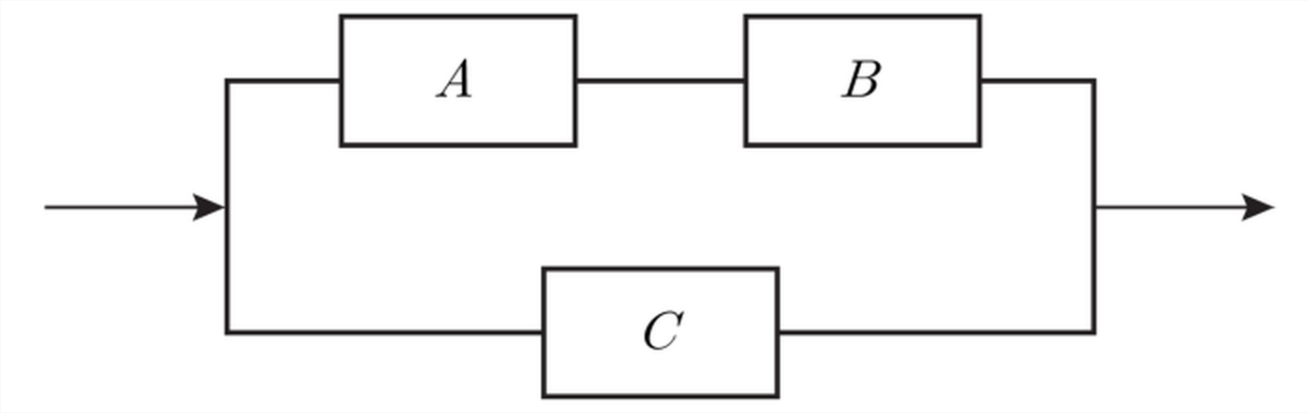
분산분석 진술형 문제는 먼저 절대적 표현이 있는지 확인한 뒤, 반복 유무에 따라 무엇을 분리하여 검정할 수 있는지 본다. 이원배치법의 자유도는 주효과 $a-1$, $b-1$, 교호작용 $(a-1)(b-1)$를 바로 써 보고 항상 성립하는지 판단하면 된다.

암기 포인트

일원배치법은 반복수 같음이 필수는 아니고, 교호작용 검정은 반복이 있어야 가능하며, 교호작용 자유도는 $(a-1)(b-1)$이다.

NoteBase

문 5. 다음 그림은 세 개의 스위치 A, B, C로 구성된 전기전달시스템을 보여 준다. 스위치 A가 ON일 확률은 0.8, 스위치 B가 ON일 확률은 0.7, 스위치 C가 ON일 확률은 0.6이라 하자. 각각의 스위치가 독립적으로 작동한다고 할 때, 이 시스템에서 전기가 전달될 확률은?



① 0.784

② 0.804

③ 0.824

④ 0.844

정답 3번

출제 포인트

직렬 경로의 성공확률을 먼저 구한 뒤 병렬 구조의 전체 전달확률로 결합하는 문제이다.

조건 정리

주어진 값: 스위치 A의 ON 확률은 0.8, B의 ON 확률은 0.7, C의 ON 확률은 0.6이다.
구해야 할 값: 전기가 전달될 전체 확률이다.
주의할 조건: A와 B는 위쪽 직렬 경로이고 C는 아래쪽 단독 경로이며, 세 스위치는 서로 독립이다.

사용할 공식

위쪽 직렬 경로의 성공확률은 $P(A \cap B) = 0.8 \times 0.7 = 0.56$이다. 전체 시스템은 '위쪽 경로 성공' 또는 '아래쪽 경로 성공'이면 전달되므로 병렬 구조이다. 따라서 전체 전달확률은 $1 - (1 - 0.56)(1 - 0.6)$이다.
이 공식을 쓰는 이유: 병렬 구조는 두 경로가 모두 실패할 때만 전체가 실패하므로, 각 경로의 실패확률을 먼저 곱한 뒤 1에서 빼면 전체 성공확률을 바로 구할 수 있다.

NoteBase

계산 과정

1. 위쪽 직렬 경로의 성공확률은 $0.8 \times 0.7 = 0.56$이다.
2. 위쪽 경로의 실패확률은 $1 - 0.56 = 0.44$, 아래쪽 경로의 실패확률은 $1 - 0.6 = 0.4$이다.
3. 전체 실패확률은 $0.44 \times 0.4 = 0.176$이다.
4. 따라서 전체 전달확률은 $1 - 0.176 = 0.824$이다.

선지별 해설

① X: 전기가 전달될 확률은 0.824이므로 0.784는 맞지 않는다. 병렬 구조에서 적어도 하나가 성공하는 확률을 계산할 때 중복 성공 부분을 잘못 처리한 값에 가깝다. 올바르게는 전기가 전달될 확률이 0.824이다.
② X: 전기가 전달될 확률은 0.824이므로 0.804는 맞지 않는다. 직렬 경로의 성공확률 0.56을 구한 뒤 병렬 결합에서 전체 실패확률 0.176을 정확히 반영하지 않으면 이런 오답이 나온다. 올바르게는 전기가 전달될 확률이 0.824이다.
③ O: 위쪽 직렬 경로의 성공확률이 0.56, 아래쪽 단독 경로의 성공확률이 0.6이므로 전체 전달확률은 $1 - (0.44)(0.4) = 0.824$이다. 계산 결과와 일치하므로 옳다.
④ X: 전기가 전달될 확률은 0.824이므로 0.844는 맞지 않는다. 두 경로의 성공확률을 단순 합산하거나 중복 성공 사건을 충분히 빼지 않으면 실제보다 큰 값이 나온다. 올바르게는 전기가 전달될 확률이 0.824이다.

오답 포인트

이 문제의 함정은 A와 B를 각각 따로 보지 말고 먼저 하나의 직렬 경로로 묶어야 한다는 점이다. 위쪽은 A와 B가 모두 켜져야 하므로 곱셈을 먼저 하고, 그다음 C 경로와 병렬 결합으로 처리해야 한다.

다시 나오면 이렇게 판단

회로 신뢰도 문제는 먼저 각 경로가 직렬인지 병렬인지 구분한 뒤, 직렬은 성공확률을 곱하고 병렬은 실패확률을 곱해서 전체 실패를 구한 다음 1에서 빼는 순서로 판단한다.

암기 포인트

직렬은 모두 성공이므로 곱하고, 병렬은 모두 실패를 먼저 구해 1에서 뺀다.

NoteBase

문 6. 연속형 확률변수 X와 Y의 결합확률밀도함수가 다음과 같을 때, 옳은 것은? (단, a는 상수이다)

$f_{X,Y}(x,y) = axy,\ 0 < x < 1,\ 0 < y < 1;\ f_{X,Y}(x,y) = 0$, 그 외

① $P\left(X \le \frac{1}{2}\right) = \frac{3}{4}$
② $P\left(Y \le \frac{1}{3}\right) = \frac{1}{9}$
③ $P\left(X > \frac{1}{4}\right) = \frac{1}{16}$
④ $P\left(Y \le \frac{1}{5}\right) = \frac{24}{25}$

정답 2번

출제 포인트

이 문제는 결합확률밀도함수의 정규화 조건으로 a를 구한 뒤 주변밀도함수로 각 확률을 적분할 수 있는지를 묻는 문제이다.

조건 정리

주어진 결합확률밀도함수는 $0 < x < 1,\ \ 0 < y < 1$에서만 $f_{X,Y}(x,y) = axy$이다.
전체 확률의 합이 1이 되어야 하므로 먼저 a를 정해야 한다.
그다음 X와 Y의 주변밀도함수를 구해 각 선지의 확률을 계산한다.

사용할 공식

정규화 조건은 $\int_0^1 \int_0^1 axy dy dx = 1$이다.
주변밀도함수는 $f_X(x) = \int_0^1 f_{X,Y}(x,y)dy$, $f_Y(y) = \int_0^1 f_{X,Y}(x,y)dx$이다.
이 공식을 쓰는 이유는 결합밀도에서 상수 a와 단일 확률변수의 확률을 정규화와 주변화로 구하기 때문이다.

계산 과정

1. ∫ _0^1 ∫ _0^1 axy dy dx=a≤ft(∫ _0^1 x dx)≤ft(∫ _0^1 y dy)=a· 12· 12=(a)/(4)=1이므로 $a = 4$이다.
2. $f_X(x) = \int_0^1 4xydy = 2x$, $f_Y(y) = \int_0^1 4xydx = 2y$이다.
3. 따라서 $P(X \le c) = \int_0^c 2xdx = c^2$, $P(Y \le c) = \int_0^c 2ydy = c^2$가 된다.

NoteBase

선지 판단

1번 X: P(X≤ 12)= ∫ _0^1/22x dx=14이다. 제시값 34는 틀렸고, 올바르게는 14이다. 34는 여사건처럼 잘못 뒤집으면 나오는 값이다.
2번 O: P(Y≤ 13)= ∫ _0^1/32y dy=19이므로 제시값이 맞다.
3번 X: P(X>14)=1-P(X≤ 14)=1- ∫ _0^1/42x dx=1-116=(15)/(16)이다. 제시값 116은 P(X≤ 14)이고, 올바르게는 $\frac{15}{16}$이다.
4번 X: P(Y≤ 15)= ∫ _0^1/52y dy=125이다. 제시값 $\frac{24}{25}$는 여사건 P(Y>15)에 해당하므로 틀렸고, 올바르게는 125이다.

오답 포인트

이 문항의 함정은 하한 누적확률과 상한 여사건을 서로 바꾸는 데 있다. 1번과 4번은 ≤를 보지 않고 여사건 값으로 착각하면 틀리고, 3번은 >를 보고도 왼쪽 꼬리확률 116을 그대로 쓰면 틀린다.

다시 나오면 이렇게 판단

결합밀도 문제는 먼저 정규화로 상수를 구하고, 다음에 주변밀도함수로 바꾼 뒤, 마지막에 부등호가 ≤인지 >인지 확인하여 누적확률인지 여사건인지 판단한다.

암기 포인트

이 문제처럼 $f_{X,Y}(x,y) \propto xy$가 단위정사각형에서 주어지면 $a=4$, 주변밀도는 $f_X(x)=2x$, $f_Y(y)=2y$, 따라서 $P(X \le c)=P(Y \le c)=c^2$이다.

NoteBase

문 7. 두 확률변수 X, Y의 결합확률분포가 다음과 같다. 확률변수 W와 Z를 $W=-4X+3$, $Z=8Y-3$이라 할 때, 공분산 $\mathrm{Cov}(W,Z)$의 값은?

X \ Y	0	1
1	$\frac{1}{4}$	$\frac{1}{4}$
2	$\frac{1}{2}$	0

① -4

② $-\frac{1}{8}$

③ $\frac{1}{8}$

④ 4

정답 4번

출제 포인트

이 문제는 결합확률분포로부터 $\mathrm{Cov}(X,Y)$를 구한 뒤 선형변환의 공분산 성질 $\mathrm{Cov}(aX+b,cY+d)=ac\mathrm{Cov}(X,Y)$를 적용할 수 있는지를 묻는 문제이다.

조건 정리

주어진 값: $P(X=1,Y=0)=1/4$, $P(X=1,Y=1)=1/4$, $P(X=2,Y=0)=1/2$, $P(X=2,Y=1)=0$ 이다.

구해야 할 값: $W=-4X+3$, $Z=8Y-3$일 때 $\mathrm{Cov}(W,Z)$이다.

주의할 조건: 상수항 $+3$, -3은 공분산에 영향을 주지 않고, 계수 -4, 8만 공분산을 배수로 바꾼다.

사용할 공식

$\mathrm{Cov}(W,Z)=\mathrm{Cov}(-4X+3,8Y-3)=(-4)\cdot 8\mathrm{Cov}(X,Y)$

$\mathrm{Cov}(X,Y)=E(XY)-E(X)E(Y)$

이 공식을 쓰는 이유는 W와 Z가 각각 X, Y의 선형변환이므로 공분산의 선형성으로 바로 연결되기 때문이다.

NoteBase

계산 과정

1. $E(X) = 1 \cdot (1/4 + 1/4) + 2 \cdot (1/2) = 1 \cdot 1/2 + 2 \cdot 1/2 = 3/2$이다.
2. $E(Y) = 0 \cdot (1/4 + 1/2) + 1 \cdot (1/4) = 1/4$이다.
3. $E(XY) = 1 \cdot 1 \cdot 1/4 = 1/4$이다.
4. 따라서 $\mathrm{Cov}(X, Y) = 1/4 - (3/2)(1/4) = 1/4 - 3/8 = -1/8$이다.
5. 그러므로 $\mathrm{Cov}(W, Z) = (-4) \cdot 8 \cdot (-1/8) = 4$이다.

선지별 해설

1번 X: 공분산은 -4가 아니라 4이다. $\mathrm{Cov}(X, Y) = -1/8$인데 부호를 하나만 반영하면 -4가 나오므로, 음수 계수와 음의 공분산이 곱해져 최종값이 양수가 된다는 점을 놓친 값이다. 틀린 값 -4는 올바르게는 4로 고쳐야 한다.
2번 X: $-1/8$은 $\mathrm{Cov}(X, Y)$ 값일 뿐 $\mathrm{Cov}(W, Z)$가 아니다. 선형변환 계수 (-4)와 8을 곱하는 단계를 빠뜨린 값이다. 틀린 값 $-1/8$은 올바르게는 4로 고쳐야 한다.
3번 X: 1/8은 공분산의 부호를 잘못 보거나 변환 계수를 반영하지 않은 값이다. 실제로는 $\mathrm{Cov}(X, Y) = -1/8$이고 여기에 -32를 곱해야 하므로 최종값은 4이다. 틀린 값 1/8은 올바르게는 4로 고쳐야 한다.
4번 O: $\mathrm{Cov}(X, Y) = -1/8$을 먼저 구한 뒤 $\mathrm{Cov}(W, Z) = (-4) \cdot 8\mathrm{Cov}(X, Y)$를 적용하면 4가 된다. 따라서 4번이 옳다.

오답 포인트

이 문항의 함정은 $\mathrm{Cov}(X, Y)$와 $\mathrm{Cov}(W, Z)$를 구분하지 못하게 만드는 데 있다. 특히 2번은 중간값 $-1/8$을 그대로 보기로 제시한 함정이고, 1번과 3번은 부호나 계수 (-4), 8을 하나만 반영한 계산 실수를 노린 보기이다.

다시 나오면 이렇게 판단

공분산 계산형은 먼저 원래 변수의 $E(X)$, $E(Y)$, $E(XY)$로 $\mathrm{Cov}(X, Y)$를 구한 다음, 선형변환이면 마지막에 ac만 곱하는 순서로 판단한다. 상수항은 공분산을 바꾸지 않는다는 점까지 함께 확인하면 된다.

암기 포인트

공분산은 상수항에 불변이고 계수에는 배수로 반응하므로 $\mathrm{Cov}(aX + b, cY + d) = ac\mathrm{Cov}(X, Y)$이다.

NoteBase

문 8. 주사위를 1회 던질 때 2의 배수가 나오는 사건을 A, 3의 배수가 나오는 사건을 B, 2의 배수가 나오지 않는 사건을 C, 3의 배수가 나오지 않는 사건을 D라고 하자. 다음 중 두 사건이 서로 독립인 경우만을 모두 고르면?

ㄱ. A, B
ㄴ. B, C
ㄷ. C, D

① ㄱ, ㄴ
② ㄱ, ㄷ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 4번

출제 포인트

이 문제는 한 번의 시행에서 두 사건의 독립 여부를 $P(E \cap F) = P(E)P(F)$ 또는 여사건의 독립 성질로 판정할 수 있는지를 묻는다.

조건 정리

표본공간은 {1,2,3,4,5,6}이다.
A={2,4,6}, B={3,6}, C={1,3,5}=A의 여사건, D={1,2,4,5}=B의 여사건이다.
각 확률은 P(A)=1/2, P(B)=1/3, P(C)=1/2, P(D)=2/3이다.

사용할 공식

두 사건 E, F가 독립이려면 $P(E \cap F) = P(E)P(F)$를 만족해야 한다.
또한 한 쌍의 사건이 독립이면 각 사건의 여사건과도 독립이다.
이 문제는 A와 B의 독립을 확인한 뒤 B와 C, C와 D를 여사건 관계로 이어서 판단할 수 있는 상황이다.

계산 과정

ㄱ: A\cap B={6}이므로 $P(A \cap B) = 1/6$이다. 한편 $P(A)P(B) = 1/2 \times 1/3 = 1/6$이므로 A와 B는 독립이다.
ㄴ: B\cap C={3}이므로 $P(B \cap C) = 1/6$이다. 한편 $P(B)P(C) = 1/3 \times 1/2 = 1/6$이므로 B와 C도 독립이다.
ㄷ: C\cap D={1,5}이므로 $P(C \cap D) = 2/6 = 1/3$이다. 한편 $P(C)P(D) = 1/2 \times 2/3 = 1/3$이므로 C와 D도 독립이다.

NoteBase

선지 판단

1번: X. 1번은 ㄱ, ㄴ만 독립이라고 보지만 ㄷ도 독립이므로 빠진 표현이 있다. 올바르게는 ㄱ, ㄴ, ㄷ을 모두 포함해야 한다.
2번: X. 2번은 ㄴ을 제외했지만 B와 C는 $P(B \cap C) = P(B)P(C)$를 만족하므로 ㄴ이 틀렸다는 판단이 잘못되었다. 올바르게는 ㄴ도 포함해야 한다.
3번: X. 3번은 ㄱ을 제외했지만 A와 B는 $P(A \cap B) = P(A)P(B)$를 만족하므로 ㄱ이 틀렸다는 판단이 잘못되었다. 올바르게는 ㄱ도 포함해야 한다.
4번: O. ㄱ, ㄴ, ㄷ이 모두 독립이므로 모든 경우를 포함한 4번이 맞다.

오답 포인트

C와 D가 여사건이라는 사실만 보고 배반사건처럼 착각하면 틀린다. 이 문제의 함정은 '여사건'과 '독립'을 반대로 느끼는 데 있다. A와 B가 독립이면 A의 여사건 C, B의 여사건 D와의 조합도 계속 독립이라는 성질을 놓치기 쉽다.

다시 나오면 이렇게 판단

독립 판정 문제는 먼저 각 사건의 원소를 써서 확률을 구한 뒤 $P(E \cap F) = P(E)P(F)$를 확인한다. 한 쌍이 독립으로 확인되면 그 여사건 조합도 독립이라는 성질까지 바로 연결해 판단한다.

암기 포인트

독립은 곱으로 확인하고, 독립이면 여사건끼리도 독립이다.

NoteBase

문 9. 평균이 μ이고 분산이 σ^2인 정규모집단으로부터 추출된 크기 2인 확률표본을 X_1, X_2라고 하자. X_1과 X_2의 표준화된 확률변수를 각각 Z_1과 Z_2라고 할 때, 다음 중 옳지 않은 것은?

① $E(Z_1) = E(Z_2)$

② $\mathrm{Var}(X_1) = 2 \times \mathrm{Var}\left(\frac{X_1 + X_2}{2}\right)$

③ $\mathrm{Var}(Z_1^2) = 1$

④ $E(Z_1^2 + Z_2^2) = 2$

정답 3번

출제 포인트

이 문제는 표준정규변수의 기댓값과 분산, 그리고 Z^2의 분포 성질을 이용해 각 선지를 판단하는 문제이다.

조건 정리

주어진 값: X_1, X_2는 평균이 μ, 분산이 σ^2인 정규모집단에서 나온 크기 2의 확률표본이다.
정의: $Z_1 = \frac{X_1 - \mu}{\sigma}, Z_2 = \frac{X_2 - \mu}{\sigma}$이다.
주의할 조건: 확률표본이므로 X_1, X_2는 서로 독립이고, 따라서 Z_1, Z_2도 서로 독립인 표준정규분포를 따른다.

사용할 공식

표준화하면 $Z_i \sim N(0, 1)$이므로 $E(Z_i) = 0, Var(Z_i) = 1$이다.
또한 독립이면 $Var\left(\frac{X_1 + X_2}{2}\right) = \frac{1}{4}(Var(X_1) + Var(X_2))$이다.
그리고 표준정규변수 Z에 대해 $Z^2 \sim \chi^2(1)$이므로 $E(Z^2) = 1, Var(Z^2) = 2$이다.
이 공식을 쓰는 이유: 각 선지가 모두 표준화, 표본평균의 분산, 카이제곱분포의 성질을 묻고 있기 때문이다.

계산 과정

1. Z_1, Z_2는 각각 표준정규분포를 따르므로 $E(Z_1) = E(Z_2) = 0$이다.
2. $Var\left(\frac{X_1 + X_2}{2}\right) = \frac{1}{4}(\sigma^2 + \sigma^2) = \frac{\sigma^2}{2}$이므로 $2 \times Var\left(\frac{X_1 + X_2}{2}\right) = \sigma^2 = Var(X_1)$이다.
3. $Z_1^2 \sim \chi^2(1)$이므로 $Var(Z_1^2) = 2$이다.
4. $E(Z_1^2 + Z_2^2) = E(Z_1^2) + E(Z_2^2) = 1 + 1 = 2$이다.

선지 판단

1번: O. Z_1, Z_2는 모두 표준정규분포를 따르므로 기댓값이 각각 0으로 같다.
2번: O. $Var(X_1) = \sigma^2$이고 표본평균의 분산은 $\frac{\sigma^2}{2}$이므로 $2 \times Var\left(\frac{X_1 + X_2}{2}\right) = \sigma^2$가 되어 성립한다.
3번: X. $Var(Z_1^2) = 1$이 아니라 2이다. 틀린 표현은 "$Var(Z_1^2) = 1$"이고, 올바르게는 "$Z_1^2 \sim \chi^2(1)$이므로 $Var(Z_1^2) = 2$"이다.
4번: O. $E(Z_1^2) = E(Z_2^2) = 1$이므로 $E(Z_1^2 + Z_2^2) = 2$이다.

오답 포인트

3번에서 Z의 분산 1과 Z^2의 분산을 혼동하면 틀린다. 표준정규변수를 제곱하면 더 이상 정규분포가 아니라 자유도 1의 카이제곱분포가 되므로 분산은 2로 바뀐다.

다시 나오면 이렇게 판단

표준화 문제가 나오면 먼저 $Z \sim N(0, 1)$인지 확인하고, 그다음 $E(Z)$, $Var(Z)$, $E(Z^2)$, $Var(Z^2)$를 구분해서 본다. 표본평균이 보이면 독립성을 이용해 분산부터 계산한다.

암기 포인트

표준정규 Z는 $E(Z) = 0$, $Var(Z) = 1$, $E(Z^2) = 1$, $Var(Z^2) = 2$이다.

NoteBase

문 10. 표준편차가 6인 정규모집단에서 모평균(μ)에 대한 가설 $H_0 : \mu = 10$ 대 $H_1 : \mu > 10$을 검정하고자 한다. 크기 n의 확률표본으로부터 구해진 표본평균을 $\overline{X}$라고 하고, 이 검정의 기각역으로 $R : \overline{X} \geq 11$을 사용하고자 한다. $\mu = 13$일 때, 검정력이 0.95 이상이 되도록 하는 최소 표본의 크기 n을 구하는 부등식은? (단, z_α는 표준정규분포의 제100×(1-α) 백분위수를 나타내고, $z_{0.05} = 1.645$, $z_{0.025} = 1.96$이다)

① $n \geq (3 \times 1.645)^2$

② $n \geq (3 \times 1.96)^2$

③ $n \geq (6 \times 1.645)^2$

④ $n \geq (6 \times 1.96)^2$

정답 1번

출제 포인트

이 문제는 표준편차를 아는 정규모집단의 단측검정에서 기각역이 주어졌을 때 검정력이 0.95 이상이 되도록 하는 최소 표본크기를 구하는 문제이다.

조건 정리

주어진 값은 $\sigma = 6, H_0 : \mu = 10, H_1 : \mu > 10$, 기각역 $R : \overline{X} \geq 11$이다.
검정력을 계산하는 기준 모평균은 $\mu = 13$이다.
구해야 할 값은 $P_{\mu = 13}(\overline{X} \geq 11) \geq 0.95$를 만족하는 최소 n이다.

사용할 공식

표본평균의 분포는 $\overline{X} \sim N(\mu, \frac{36}{n})$이므로 검정력은 $P_{\mu = 13}(\overline{X} \geq 11)$로 계산한다.
즉 $P_{\mu = 13}(\overline{X} \geq 11) = P\left(Z \geq \frac{11 - 13}{6/\sqrt{n}}\right)$ 이다.
이 공식을 쓰는 이유는 기각역이 표본평균 $\overline{X}$로 직접 주어졌고, $\mu = 13$일 때 그 기각역에 들어갈 확률이 곧 검정력이기 때문이다.

NoteBase

계산 과정

1. $\mu = 13$일 때
$P(\overline{X} \geq 11) = P\left(Z \geq \frac{11-13}{6/\sqrt{n}}\right) = P\left(Z \geq -\frac{\sqrt{n}}{3}\right)$
2. 표준정규분포의 대칭성을 쓰면
$P\left(Z \geq -\frac{\sqrt{n}}{3}\right) = \Phi\left(\frac{\sqrt{n}}{3}\right)$
3. 검정력이 0.95 이상이어야 하므로
$\Phi\left(\frac{\sqrt{n}}{3}\right) \geq 0.95$
4. 따라서
$\frac{\sqrt{n}}{3} \geq z_{0.05} = 1.645$
5. 정리하면
$\sqrt{n} \geq 3 \times 1.645 \quad \Rightarrow \quad n \geq (3 \times 1.645)^2$
따라서 ①이 맞다.

선지 판단

① O: 계산 결과가 바로 $n \geq (3 \times 1.645)^2$이므로 맞다.
② X: '1.96'가 틀린 표현이다. 검정력 0.95는 $\Phi(\sqrt{n}/3) \geq 0.95$에 해당하므로 $z_{0.05} = 1.645$를 써야 한다. 1.96은 상위 2.5% 기준인 $z_{0.025}$일 때의 값이다. 올바르게는 $n \geq (3 \times 1.645)^2$이다.
③ X: '6\times1.645'가 틀린 표현이다. 표준화할 때 분자는 기각값과 모평균의 차이인 $13 - 11 = 2$이므로 $2\sqrt{n}/6 = \sqrt{n}/3$가 된다. 이 차이 2를 반영하지 않으면 3이 아니라 6이 남는 오답이 나온다. 올바르게는 $n \geq (3 \times 1.645)^2$이다.
④ X: '6\times1.96'가 틀린 표현이다. 이 선지는 ②의 분위수 선택 오류와 ③의 차이값 반영 누락 오류가 함께 들어간 경우이다. 올바르게는 $z_{0.05} = 1.645$를 쓰고, $13 - 11 = 2$를 반영하여 $n \geq (3 \times 1.645)^2$로 정리한다.

오답 포인트

이 문항의 함정은 두 가지이다. 하나는 단측 검정의 검정력 0.95 조건에서 $z_{0.05}$를 써야 하는데 익숙한 1.96을 기계적으로 대입하는 것이다. 다른 하나는 11과 13의 차이 2를 놓쳐서 $\sqrt{n}/3$ 대신 $\sqrt{n}/6$처럼 처리하는 것이다.

다시 나오면 이렇게 판단

기각역이 주어지면 먼저 정답이 아니라 검정력을 확률식으로 쓴다. 그다음 대입할 모평균에서 기각역 경계값까지의 차이를 분자에 놓고 표준화한다. 마지막으로 단측인지 양측인지 확인해 해당 분위수 z_α를 고르면 된다.

암기 포인트

검정력은 대립가설이 참일 때 기각역에 들어갈 확률이며, 단측 0.95는 $z_{0.05} = 1.645$를 쓴다.

NoteBase

문 11. 평균이 μ이고 분산이 σ^2인 모집단으로부터 추출된 크기 5인 확률표본을 $X_1, X_2, \cdots, X_5$라고 하자. 다음에 제시된 μ에 대한 추정량 중에서 불편추정량이면서 동시에 분산이 제일 작은 추정량은? (단, $\sigma^2 > 0$이다)

$T_1 = \frac{X_1 + X_5}{2}, T_2 = \frac{2X_1 + 2X_2 - X_3}{3}, T_3 = \frac{X_1 + X_2 + X_5}{4}, T_4 = X_3 + X_4 - X_5$

① T_1

② T_2

③ T_3

④ T_4

정답 1번

출제 포인트

이 문제는 선형추정량 $T = \sum a_i X_i$의 불편성 조건과 분산 $\mathrm{Var}(T) = \sigma^2 \sum a_i^2$를 이용해 가장 효율적인 추정량을 고르는 문제이다.

조건 정리

모집단의 평균은 μ, 분산은 σ^2이다. 표본 $X_1, \ \dots, X_5$는 같은 모집단에서 추출되므로 각 X_i의 평균은 μ, 분산은 σ^2로 본다. 선형추정량 $T = \sum a_i X_i$가 μ의 불편추정량이 되려면 계수합 $\sum a_i = 1$이어야 한다.

사용할 공식

불편성은 $E[T] = \sum a_i E[X_i] = \mu \sum a_i$로 판단한다. 분산은 표본들이 서로 독립이므로 $\mathrm{Var}(T) = \sigma^2 \sum a_i^2$로 계산한다. 이 문제는 각 추정량이 X_i의 선형결합이므로 계수합과 계수 제곱합만 비교하면 된다.

계산 과정

1. $T_1 = \frac{X_1 + X_5}{2}$는 계수합이 12+12=1이므로 불편추정량이다. 분산은 σ^2≤ft(14+14)=(σ^2)/(2)이다.
2. $T_2 = \frac{2X_1 + 2X_2 - X_3}{3}$는 계수합이 23+23-13=1이므로 불편추정량이다. 분산은 σ^2≤ft(49+49+19)=σ^2이다.
3. $T_3 = \frac{X_1 + X_2 + X_5}{4}$는 계수합이 14+14+14=34이므로 E[T_3]=34μ가 되어 불편추정량이 아니다.
4. $T_4 = X_3 + X_4 - X_5$는 계수합이 $1 + 1 - 1 = 1$이므로 불편추정량이다. 분산은 $\sigma^2(1 + 1 + 1) = 3\sigma^2$이다.
5. 불편추정량은 T_1, T_2, T_4이고, 이들 중 분산이 가장 작은 것은 $\frac{\sigma^2}{2}$인 T_1이다.

NoteBase

선지별 해설

1번: O. T_1은 계수합이 1이어서 불편추정량이고, 분산이 $\frac{\sigma^2}{2}$로 불편추정량들 중 가장 작다.
2번: X. T_2도 불편추정량이지만 분산이 σ^2이므로 T_1보다 크다. 틀린 표현은 '분산이 제일 작다'이고, 올바르게는 '불편추정량이지만 최소분산은 아니다'이다.
3번: X. T_3는 분산이 $\frac{3\sigma^2}{16}$로 작아 보여도 계수합이 34라서 불편추정량이 아니다. 틀린 표현은 '불편추정량이다'이고, 올바르게는 '기댓값이 34μ이므로 편향추정량이다'이다.
4번: X. T_4는 계수합이 1이어서 불편추정량이지만 분산이 $3\sigma^2$로 가장 크다. 틀린 표현은 '분산이 제일 작다'이고, 올바르게는 '불편추정량이지만 분산이 매우 큰 추정량이다'이다.

오답 포인트

불편성은 계수합이 1인지로 판단하고, 효율성은 계수 제곱합의 크기로 비교해야 한다. T_3처럼 분산이 작아 보여도 계수합이 1이 아니면 바로 탈락한다는 점을 놓치기 쉽다.

다시 나오면 이렇게 판단

선형추정량 비교 문제는 먼저 계수합이 1인지 확인해 불편추정량만 남기고, 그다음 계수 제곱합 $\sum a_i^2$를 비교해 분산이 가장 작은 것을 고르면 된다.

암기 포인트

선형불편추정량 판별은 계수합 1, 최소분산 비교는 계수 제곱합 최소이다.

NoteBase

문 12. $X_1, X_2, \cdots, X_5$**를** $(\theta, \theta+1)$**에서 균일분포(uniform distribution)를 따르는 모집단으로부터 추출된 크기 5인 확률표본이라 하고, 표본평균을** $\overline{X}$**라고 하자. 모수** θ**에 대한 추정량으로** $\hat{\theta} = \overline{X} - \frac{1}{2}$**을 사용할 때, 이 추정량의 편의(bias)와 평균제곱오차(MSE)를 바르게 연결한 것은?**

① 편의: 0, 평균제곱오차: $\frac{1}{60}$

② 편의: 0, 평균제곱오차: $\frac{1}{30}$

③ 편의: $\frac{1}{2}$, 평균제곱오차: $\frac{1}{60}$

④ 편의: $\frac{1}{2}$, 평균제곱오차: $\frac{1}{30}$

정답 1번

출제 포인트

이 문제는 $U(\theta, \theta+1)$ 표본에서 $\hat{\theta} = \bar{X} - \frac{1}{2}$의 편의와 평균제곱오차를 구하는 문제이다.

조건 정리

주어진 값:
- $X_1, \ \dots, X_5$는 $(\theta, \theta+1)$에서의 균일분포 표본이다.
- 표본크기는 5이고, 추정량은 $\hat{\theta} = \bar{X} - \frac{1}{2}$이다.

구해야 할 값:
- 편의 $E(\hat{\theta}) - \theta$이다.
- 평균제곱오차 $MSE(\hat{\theta})$이다.

주의할 조건:
- 균일분포의 평균과 분산을 먼저 구한 뒤 표본평균의 성질을 적용한다.
- 편의가 0이면 평균제곱오차는 분산과 같아진다.

사용할 공식

공식:
- $E[U(a,b)] = \frac{a+b}{2}$이다.
- $Var[U(a,b)] = \frac{(b-a)^2}{12}$이다.
- $Var(\bar{X}) = \frac{Var(X)}{n}$이다.
- $MSE(\hat{\theta}) = Var(\hat{\theta}) + Bias(\hat{\theta})^2$이다.

이 공식을 쓰는 이유:
- 이 문제의 추정량은 표본평균에서 상수 $\frac{1}{2}$를 뺀 형태이므로, 먼저 균일분포의 평균과 분산을 구하면 편의와 평균제곱오차를 바로 계산할 수 있다.

NoteBase

계산 과정

1. $X_i \sim U(\theta, \theta+1)$이므로 $E(X_i) = \frac{\theta+(\theta+1)}{2} = \theta + \frac{1}{2}$, $Var(X_i) = \frac{(\theta+1-\theta)^2}{12} = \frac{1}{12}$이다.
2. 따라서 $E(\bar{X}) = \theta + \frac{1}{2}$이고, $E(\hat{\theta}) = E\left(\bar{X} - \frac{1}{2}\right) = \theta$이므로 편의는 0이다.
3. $Var(\hat{\theta}) = Var\left(\bar{X} - \frac{1}{2}\right) = Var(\bar{X}) = \frac{1/12}{5} = \frac{1}{60}$이다.
4. 편의가 0이므로 $MSE(\hat{\theta}) = Var(\hat{\theta}) + 0^2 = \frac{1}{60}$이다.

선지별 해설

① O: 편의가 0이고 평균제곱오차가 $\frac{1}{60}$으로 계산되므로 옳다.

② X: 편의 0까지는 맞지만 평균제곱오차 $\frac{1}{30}$이 틀리다. 올바르게는 $Var(\bar{X}) = \frac{1/12}{5} = \frac{1}{60}$이므로 평균제곱오차도 $\frac{1}{60}$이다.

③ X: 평균제곱오차 $\frac{1}{60}$은 맞지만 편의 $\frac{1}{2}$가 틀리다. 올바르게는 $E(\bar{X}) = \theta + \frac{1}{2}$이어서 $\bar{X} - \frac{1}{2}$는 정확히 θ를 중심으로 하므로 편의는 0이다.

④ X: 편의 $\frac{1}{2}$와 평균제곱오차 $\frac{1}{30}$이 모두 틀리다. 올바르게는 편의는 0이고 평균제곱오차는 $\frac{1}{60}$이다.

오답 포인트

균일분포 $U(\theta, \theta+1)$의 평균이 $\theta + \frac{1}{2}$라는 점을 놓치면 ③, ④처럼 편의를 $\frac{1}{2}$로 잡기 쉽다. 또 $Var(X) = \frac{1}{12}$에서 표본평균의 분산을 $\frac{Var(X)}{5}$로 줄여야 한다는 점을 놓치면 ②, ④처럼 평균제곱오차를 잘못 계산하기 쉽다.

다시 나오면 이렇게 판단

위치모수 형태의 균일분포 추정량이 나오면 먼저 모평균으로 편의를 확인하고, 다음으로 모분산에서 $Var(\bar{X}) = Var(X)/n$을 계산한 뒤, 불편이면 $MSE = Var$로 정리한다.

암기 포인트

$U(\theta, \theta+1)$에서는 평균이 $\theta + \frac{1}{2}$, 분산이 $\frac{1}{12}$이고, $\bar{X} - \frac{1}{2}$는 불편추정량이므로 표본크기 5에서 평균제곱오차는 $\frac{1}{60}$이다.

NoteBase

문 13. 관측 자료 (x_i, y_i) $(i = 1, 2, \cdots, 50)$를 이용하여 단순선형회귀모형 $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$를 최소제곱법으로 추정한 결과 $\hat{\beta}_0 = 200$, $\hat{\beta}_1 = -0.4$를 얻었다. 설명변수의 평균이 150, 표준편차가 80이고, 추정된 회귀직선의 결정계수(R^2)가 0.81일 때, 이에 대한 설명으로 옳은 것만 모두 고르면? (단, ε_i는 $N(0, \sigma^2)$를 따르고 서로 독립이다)

ㄱ. 설명변수와 반응변수 간의 표본상관계수는 -0.9이다.
ㄴ. 반응변수의 평균은 140이다.
ㄷ. $x = 100$일 때, 반응변수의 추정값은 160이다.

① ㄱ
② ㄱ, ㄴ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 4번

출제 포인트

단순선형회귀에서 R^2와 상관계수의 관계, 회귀직선이 $(\bar{x}, \bar{y})$를 지난다는 성질, 추정식 대입 계산을 함께 묻는 문제이다.

조건 정리

주어진 값은 $\hat{\beta}_0 = 200$, $\hat{\beta}_1 = -0.4$, $\bar{x} = 150$, 설명변수의 표준편차 80, $R^2 = 0.81$이다.
구해야 할 값은 ㄱ의 표본상관계수, ㄴ의 반응변수 평균, ㄷ의 $x = 100$에서의 추정값이다.
주의할 조건은 단순선형회귀에서는 $R^2 = r^2$이고, 상관계수의 부호는 기울기 $\hat{\beta}_1$의 부호와 같으며, 추정 회귀직선은 항상 $(\bar{x}, \bar{y})$를 지난다는 점이다.

사용할 공식

$R^2 = r^2$
$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
이 공식을 쓰는 이유는 문제에서 단순선형회귀의 추정결과와 R^2, 평균값, 특정 x값이 함께 주어졌기 때문이다.

NoteBase

계산 과정

1. ㄱ은 $r^2 = 0.81$이므로 $|r| = 0.9$라는 점에서 출발한다. 기울기 $\hat{\beta}_1 = -0.4 < 0$이므로 상관계수의 부호도 음수여서 $r = -0.9$이다.
2. ㄴ은 추정 회귀직선이 $(\bar{x}, \bar{y})$를 지난다는 성질을 이용한다. 따라서 $\bar{y} = 200 + (-0.4) \times 150 = 200 - 60 = 140$이다.
3. ㄷ은 추정식에 $x = 100$을 대입한다. 따라서 $\hat{y} = 200 + (-0.4) \times 100 = 200 - 40 = 160$이다.
4. 따라서 ㄱ, ㄴ, ㄷ이 모두 옳다.

선지별 해설

① X
①은 ㄱ만 옳다고 본 선택지이다. 그러나 ㄴ도 $\bar{y} = 140$으로 옳고, ㄷ도 $\hat{y} = 160$으로 옳다.
틀린 표현은 "ㄱ만 해당한다"는 취지이다. 올바르게는 "ㄱ, ㄴ, ㄷ이 모두 옳다"이다.

② X
②는 ㄱ, ㄴ만 옳다고 본 선택지이다. 그러나 ㄷ에서 $x = 100$을 대입하면 추정값이 160이므로 ㄷ도 옳다.
틀린 표현은 "ㄷ은 틀리다"는 취지이다. 올바르게는 "ㄷ도 옳다"이다.

③ X
③은 ㄴ, ㄷ만 옳다고 본 선택지이다. 그러나 단순선형회귀에서 $R^2 = r^2$이고 기울기가 음수이므로 상관계수는 -0.9가 되어 ㄱ도 옳다.
틀린 표현은 "ㄱ은 틀리다"는 취지이다. 올바르게는 "ㄱ도 옳다"이다.

④ O
④는 ㄱ, ㄴ, ㄷ을 모두 포함한 선택지이다. ㄱ은 상관계수가 -0.9라는 진술이어서 옳고, ㄴ은 반응변수 평균이 140이라는 진술이어서 옳고, ㄷ은 $x = 100$에서 추정값이 160이라는 진술이어서 옳다.
세 진술이 모두 맞으므로 ④가 정답이다.

오답 포인트

$R^2 = 0.81$에서 바로 $r = 0.81$로 착각하거나, $r = \pm 0.9$에서 부호를 빼먹는 경우가 많다. 또 회귀직선이 $(\bar{x}, \bar{y})$를 지난다는 성질을 잊으면 ㄴ을 빠뜨리기 쉽고, ㄷ은 추정식 대입에서 음수 기울기의 부호를 잘못 처리하면 오답이 된다.

다시 나오면 이렇게 판단

단순선형회귀 문항은 먼저 R^2로 상관계수의 크기를 구하고 기울기 부호로 상관계수 부호를 정한다. 그다음 회귀직선이 $(\bar{x}, \bar{y})$를 지난다는 성질로 평균을 확인하고, 마지막으로 주어진 x값을 추정식에 직접 대입해 각 진술의 진위를 판단한다.

NoteBase

암기 포인트

단순회귀에서는 $R^2 = r^2$이고, 상관계수의 부호는 기울기 부호와 같으며, 추정 회귀직선은 항상 $(\bar{x}, \bar{y})$를 지난다.

문 14. 모수가 λ인 포아송 분포를 따르는 확률변수(X)의 확률질량함수는 다음과 같다. X_1, X_2를 서로 독립이며 각각 모수가 $\lambda_1 = 3, \lambda_2 = 2$인 포아송 분포를 따르는 확률변수라고 할 때, $P(X_1 + 2X_2 = 2)$의 값은?

f(x)=cases(e^-λλ^x)/(x!), & x = 0, 1, 2, ·s \ 0, & 그 외cases (λ > 0)

① $\frac{9}{2}e^{-5}$
② $\frac{11}{2}e^{-5}$
③ $\frac{13}{2}e^{-5}$
④ $\frac{25}{2}e^{-5}$

정답 3번

출제 포인트

이 문제는 독립인 두 포아송확률변수에 대해 조건을 만족하는 가능한 경우를 나누어 확률질량함수를 적용할 수 있는지를 묻는다.

조건 정리

주어진 값: $X_1 \sim \text{Poisson}(3)$, $X_2 \sim \text{Poisson}(2)$, 두 변수는 서로 독립이다.
구해야 할 값: $P(X_1 + 2X_2 = 2)$이다.
주의할 조건: X_2에 2가 곱해져 있으므로 $X_1 + X_2 = 2$로 보면 안 되고, 식을 만족하는 정수 조합만 찾아야 한다.

사용할 공식

포아송분포의 확률질량함수는 $P(X = x) = e^{-\lambda}\lambda^x/x!$이다.
독립이므로 $P(X_1 = a, X_2 = b) = P(X_1 = a)P(X_2 = b)$이다.
이 공식을 쓰는 이유는 $X_1 + 2X_2 = 2$를 만족하는 경우가 $(X_1, X_2) = (2, 0), (0, 1)$ 두 가지뿐이어서 각 결합확률을 곱으로 계산해 더하면 되기 때문이다.

계산 과정

1. $X_1 + 2X_2 = 2$를 만족하는 경우를 찾으면 $(X_1, X_2) = (2, 0), (0, 1)$이다.
2. $P(X_1 = 2, X_2 = 0) = P(X_1 = 2)P(X_2 = 0) = \left(e^{-3} \cdot \frac{3^2}{2!}\right)(e^{-2}) = \frac{9}{2}e^{-5}$이다.
3. $P(X_1 = 0, X_2 = 1) = P(X_1 = 0)P(X_2 = 1) = (e^{-3})(e^{-2} \cdot 2) = 2e^{-5}$이다.
4. 따라서 $P(X_1 + 2X_2 = 2) = \frac{9}{2}e^{-5} + 2e^{-5} = \frac{13}{2}e^{-5}$이다.

NoteBase

선지 판단

1번 X: $\frac{9}{2}e^{-5}$는 $(X_1, X_2) = (2, 0)$ 한 경우만 계산한 값이다. $(0, 1)$ 경우의 $2e^{-5}$를 더해야 하므로 올바르게는 $\frac{13}{2}e^{-5}$이다.
2번 X: $\frac{11}{2}e^{-5}$는 $P(X_2 = 1)$을 e^{-2}로 잘못 계산해 $(0, 1)$ 경우를 e^{-5}로 처리할 때 나오는 값이다. 올바르게는 $P(X_2 = 1) = 2e^{-2}$이므로 전체 확률은 $\frac{13}{2}e^{-5}$이다.
3번 O: 두 가능한 경우 $(2, 0)$, $(0, 1)$의 확률을 더하면 $\frac{9}{2}e^{-5} + 2e^{-5} = \frac{13}{2}e^{-5}$이므로 옳다.
4번 X: $\frac{25}{2}e^{-5}$는 $X_1 + 2X_2$를 잘못 보고 $X_1 + X_2 \sim \text{Poisson}(5)$에서 $P(X_1 + X_2 = 2)$를 계산할 때 나오는 값이다. 이 문제의 올바른 값은 $\frac{13}{2}e^{-5}$이다.

오답 포인트

가장 흔한 실수는 $X_1 + 2X_2 = 2$를 보고 가능한 조합을 끝까지 나누지 않는 것이다. 특히 $(0, 1)$ 경우를 빠뜨리거나, 더 심하게는 $X_1 + X_2 = 2$처럼 합의 계수를 무시하면 1번이나 4번으로 흐르기 쉽다.

다시 나오면 이렇게 판단

포아송 결합확률 계산 문제는 먼저 식을 만족하는 정수 조합을 모두 적고, 그다음 독립이면 각 경우의 확률을 곱해서 더하는 순서로 판단한다.

암기 포인트

독립 포아송의 선형결합 문제는 합분포를 바로 쓰지 말고, 조건식을 만족하는 경우의 수를 먼저 분해하는 것이 핵심이다.

NoteBase

문 15. 다음은 어떤 자료에 대해 단순선형회귀모형 $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ $(i = 1, 2, \cdots, 16)$을 적합한 결과와 그에 따른 해석이다. (가), (나)에 들어갈 내용으로 바르게 연결한 것은? (단, ε_i는 $N(0, \sigma^2)$을 따르고 서로 독립이다. 또한 $t_\alpha(k)$는 자유도가 k인 t분포의 제$100 \times (1-\alpha)$ 백분위수를 나타내고, $t_{0.025}(14) = 2.15$, $t_{0.05}(14) = 1.76$이다)

계수	추정회귀계수	표준오차	t-값	p-값
β_0	120	40	3	0.0096
β_1	1.8	0.5	3.6	0.0029

β_0에 대한 95% 신뢰구간은 (가)이고, 가설 $H_0 : \beta_1 \leq 1$ 대 $H_1 : \beta_1 > 1$에 대한 유의수준 5%의 검정 결과 귀무가설을 기각할 수 (나).

① (가): $(120 - 1.76 \times 40, 120 + 1.76 \times 40)$, (나): 있다

② (가): $(120 - 1.76 \times 40, 120 + 1.76 \times 40)$, (나): 없다

③ (가): $(120 - 2.15 \times 40, 120 + 2.15 \times 40)$, (나): 있다

④ (가): $(120 - 2.15 \times 40, 120 + 2.15 \times 40)$, (나): 없다

정답 4번

출제 포인트

회귀계수의 신뢰구간에서는 양측 임계값을, 단측 가설검정에서는 귀무가설의 경계값과 단측 임계값을 쓰는지를 묻는 문제이다.

조건 정리

표본크기가 16이므로 자유도는 $16 - 2 = 14$이다.
$\hat{\beta}_0 = 120, SE(\hat{\beta}_0) = 40$이다.
$\hat{\beta}_1 = 1.8, SE(\hat{\beta}_1) = 0.5$이다.
95% 신뢰구간은 양측 구간이므로 $t_{0.025}(14) = 2.15$를 쓰고, 유의수준 5% 단측검정은 $t_{0.05}(14) = 1.76$를 쓴다.

사용할 공식

회귀계수의 95% 신뢰구간은 $\hat{\beta}_0 \pm t_{0.025}(14) SE(\hat{\beta}_0)$이다.
가설 $H_0 : \beta_1 \leq 1$ 대 $H_1 : \beta_1 > 1$의 검정통계량은 귀무가설 경계값을 넣어 $t = \frac{\hat{\beta}_1 - 1}{SE(\hat{\beta}_1)}$로 계산한다.
이 공식을 쓰는 이유는 신뢰구간은 β_0의 추정오차를 양측으로 평가하는 문제이고, 검정은 $\beta_1 = 0$이 아니라 $\beta_1 = 1$을 기준으로 하는 단측검정이기 때문이다.

NoteBase

계산 과정

1. β_0의 95% 신뢰구간은 $120 \pm 2.15 \times 40$이므로 $(120 - 2.15 \times 40, 120 + 2.15 \times 40)$이다.
2. 검정통계량은 $t = \frac{1.8 - 1}{0.5} = 1.6$이다.
3. 단측 5% 임계값은 1.76인데 $1.6 < 1.76$이므로 귀무가설을 기각할 수 없다.

선지별 해설

① X

- (가)가 틀리다. 95% 신뢰구간은 양측 구간이므로 1.76이 아니라 2.15를 써야 한다. 틀린 표현은 $(120 - 1.76 \times 40, 120 + 1.76 \times 40)$이고, 올바르게는 $(120 - 2.15 \times 40, 120 + 2.15 \times 40)$이다.
- (나)는 맞다. 검정통계량이 1.6이므로 귀무가설을 기각할 수 있다가 아니라 기각할 수 없다가 맞는다.

② X

- (나)는 맞지만 (가)가 틀리므로 전체 선지는 틀리다. 95% 신뢰구간에서 1.76을 쓰면 단측 5% 또는 양측 90% 수준과 혼동한 것이다.
- 틀린 표현은 $(120 - 1.76 \times 40, 120 + 1.76 \times 40)$이고, 올바르게는 $(120 - 2.15 \times 40, 120 + 2.15 \times 40)$이다.

③ X

- (가)는 맞다. 95% 신뢰구간은 $(120 - 2.15 \times 40, 120 + 2.15 \times 40)$이다.
- (나)가 틀리다. 표의 $t = 3.6, p = 0.0029$는 $H_0 : \beta_1 = 0$ 검정에 대한 값이지 H_0:_1≤1 검정값이 아니다. 틀린 표현은 귀무가설을 기각할 수 있다는 부분이고, 올바르게는 귀무가설을 기각할 수 없다이다.

④ O

- (가)는 $(120 - 2.15 \times 40, 120 + 2.15 \times 40)$로 맞다.
- (나)는 $t = 1.6 < 1.76$이므로 귀무가설을 기각할 수 없어서 맞다.

오답 포인트

표에 이미 적힌 t-값 3과 3.6, p-값 0.0096과 0.0029를 그대로 가져다 쓰면 함정에 걸린다. 그 값들은 각각 $\beta_0 = 0, \beta_1 = 0$ 검정의 결과이고, 이 문항의 (나)는 $\beta_1 = 1$을 기준으로 다시 검정통계량을 계산해야 한다. 또 95% 신뢰구간인데 $t_{0.05}$를 쓰는 실수도 자주 나온다.

다시 나오면 이렇게 판단

회귀 추론 문제는 먼저 자유도와 양측·단측 여부를 정한 뒤, 신뢰구간이면 $\hat{\beta} \pm t \times SE$를 쓰고, 가설검정이면 귀무가설의 경계값을 뺀 검정통계량을 다시 계산해 임계값과 비교하는 순서로 판단한다.

암기 포인트

양측 95% 신뢰구간은 $t_{0.025}$, 단측 5% 검정은 $t_{0.05}$를 쓴다. 검정통계량은 항상 추정치에서 귀무가설 값을 뺀다.

NoteBase

문 16. 남자 110명, 여자 110명을 각각 임의추출하여 가장 지지하는 정당을 조사한 결과가 다음과 같다. 이 조사에서 무응답이나 중복 응답은 없다고 한다. 정당별 지지율이 성별에 따라 다른지를 알아보기 위해 카이제곱검정을 실시하였다. 이 검정에서 p-값이 속하는 범위는? (단, $\chi^2_\alpha(k)$는 자유도가 k인 카이제곱분포의 제 $100 \times (1-\alpha)$ 백분위수를 나타내고, $\chi^2_{0.01}(2) = 9.21$, $\chi^2_{0.05}(2) = 5.99$, $\chi^2_{0.1}(2) = 4.61$, $\chi^2_{0.01}(5) = 15.09$, $\chi^2_{0.05}(5) = 11.07$, $\chi^2_{0.1}(5) = 9.24$이다.)

성별 / 정당: A, B, C, 계
남자: A 55, B 30, C 25, 계 110
여자: A 45, B 50, C 15, 계 110

① p-값 < 0.01
② $0.01 \leq p$-값 < 0.05
③ $0.05 \leq p$-값 < 0.1
④ p-값 ≥ 0.1

정답 2번

출제 포인트

성별과 지지 정당이 독립인지 여부를 2×3 분할표의 카이제곱 검정으로 판단하는 문제이다.

조건 정리

주어진 값:
남자 110명, 여자 110명이고 열합계는 $A = 100, B = 80, C = 40$, 전체는 220명이다.
구해야 할 값:
카이제곱 검정통계량과 그에 따른 p-값의 범위이다.
주의할 조건:
성별 2수준, 정당 3수준이므로 자유도는 $(2-1)(3-1) = 2$이다.

사용할 공식

검정통계량은 $\chi^2 = \sum \frac{(O-E)^2}{E}$ 이다.
기대도수는 E=((행합계)(열합계))/(전체합계) 이다.
이 공식을 쓰는 이유:
두 범주형 변수의 분포가 같은지 보는 문제이므로 각 칸의 관측도수와 독립 가정하의 기대도수를 비교하는 카이제곱 검정을 사용한다.

NoteBase

계산 과정

1. 기대도수를 구하면 남자 행은 50, 40, 20, 여자 행도 50, 40, 20이다.
2. 검정통계량은 $\frac{(55-50)^2}{50}+\frac{(30-40)^2}{40}+\frac{(25-20)^2}{20}+\frac{(45-50)^2}{50}+\frac{(50-40)^2}{40}+\frac{(15-20)^2}{20}$ 이다.
3. 이를 계산하면 $0.5+2.5+1.25+0.5+2.5+1.25=8.5$ 이다.
4. 자유도 2에서 $\chi^2_{0.05}(2)=5.99$, $\chi^2_{0.01}(2)=9.21$ 이므로 $5.99<8.5<9.21$ 이다.
5. 따라서 0.01 ≤ p-값<0.05 이다.

선지별 해설

① X: 검정통계량이 $\chi^2_{0.01}(2)=9.21$ 보다 작으므로 p-값<0.01 이라고 볼 수 없다. 틀린 표현은 "p-값<0.01"이고, 올바르게는 "0.01 ≤ p-값<0.05" 이다.
② O: 검정통계량 8.5는 $\chi^2_{0.05}(2)=5.99$ 보다 크고 $\chi^2_{0.01}(2)=9.21$ 보다 작으므로 p-값은 0.01 이상 0.05 미만이다.
③ X: p-값≥ 0.05 이려면 검정통계량이 5.99 이하이어야 하는데 실제 값은 8.5이다. 틀린 표현은 "0.05 ≤ p-값<0.1"이고, 올바르게는 "0.01 ≤ p-값<0.05" 이다.
④ X: p-값≥ 0.1 이려면 검정통계량이 $\chi^2_{0.1}(2)=4.61$ 이하이어야 하는데 실제 값은 8.5이다. 틀린 표현은 "p-값≥ 0.1"이고, 올바르게는 "0.01 ≤ p-값<0.05" 이다.

오답 포인트

열합계와 기대도수를 구하지 않고 관측도수 차이만 보고 판단하면 오답이 된다. 또 자유도 2를 확인하지 않고 제시된 임계값을 섞어 쓰거나, p-값 범위를 검정통계량의 크기와 반대로 읽으면 ①이나 ③으로 흔들리기 쉽다.

다시 나오면 이렇게 판단

분할표 검정 문제는 행합계·열합계로 기대도수를 먼저 만든 뒤, 자유도 $(r-1)(c-1)$ 를 확인하고, 계산한 χ^2 값을 임계값 사이에 끼워 p-값 범위를 판단하면 된다.

암기 포인트

카이제곱 검정은 기대도수 계산 후 χ^2 값이 클수록 p-값은 작아진다.

NoteBase

문 17. 어느 제약회사에서 코로나 치료제를 개발하기 위해 40명의 코로나 환자를 대상으로 4가지 약품에 대해 임상 시험을 진행하였다. 이 시험에서 얻은 자료로 분산분석표를 작성한 결과가 다음과 같을 때, (가) ~ (다)의 값으로 바르게 연결한 것은?

요인	제곱합	자유도	평균제곱	F-값
처리	30	(가)		(다)
오차		36	(나)	
계	210			

① (가) 3 (나) 5 (다) 2

② (가) 3 (나) 6.7 (다) 1.5

③ (가) 4 (나) 5 (다) 2

④ (가) 4 (나) 6.7 (다) 1.5

정답 1번

출제 포인트

이 문제는 일원분산분석표에서 자유도, 평균제곱, F값의 관계를 이용해 빈칸을 복원하는 문제이다.

조건 정리

주어진 값: 처리 제곱합은 30, 총 제곱합은 210, 처리 수준 수는 4개, 전체 대상자는 40명이다.
구해야 할 값: (가)는 처리 자유도, (나)는 오차 평균제곱, (다)는 F값이다.
주의할 조건: 일원분산분석에서는 처리 자유도와 오차 자유도를 먼저 정한 뒤 제곱합으로 평균제곱과 F값을 계산한다.

사용할 공식

처리 자유도는 $k-1$, 오차 자유도는 $N-k$, 오차 제곱합은 $SSE = SST - SSA$, 처리 평균제곱은 $MST = SSA/(k-1)$, 오차 평균제곱은 $MSE = SSE/(N-k)$, 검정통계량은 $F = MST/MSE$이다.
이 공식을 쓰는 이유는 문제에서 처리 수, 전체 표본수, 처리 제곱합, 총 제곱합이 주어져 있어 분산분석표의 기본 관계만으로 빈칸을 모두 계산할 수 있기 때문이다.

NoteBase

계산 과정

1. 처리 수준이 4개이므로 처리 자유도는 $4 - 1 = 3$이다. 따라서 (가)는 3이다.
2. 오차 자유도는 표에 이미 36으로 주어져 있고, 총 제곱합 210에서 처리 제곱합 30을 빼면 오차 제곱합은 $210 - 30 = 180$이다.
3. 오차 평균제곱은 $180/36 = 5$이므로 (나)는 5이다.
4. 처리 평균제곱은 $30/3 = 10$이고, $F = 10/5 = 2$이므로 (다)는 2이다.

선지별 해설

① O: (가)=3, (나)=5, (다)=2로 계산 결과와 모두 일치한다.
② X: (가)=3은 맞지만 (나)=6.7, (다)=1.5가 틀리다. 오차 평균제곱은 $180/36 = 5$이므로 6.7이 될 수 없고, F 값도 $10/5 = 2$이다. 틀린 표현은 '(나)=6.7, (다)=1.5'이고 올바르게는 '(나)=5, (다)=2'이다.
③ X: (나)=5와 (다)=2는 맞지만 (가)=4가 틀리다. 처리 자유도는 처리 개수 자체가 아니라 처리 개수에서 1을 뺀 값이므로 3이다. 틀린 표현은 '(가)=4'이고 올바르게는 '(가)=3'이다.
④ X: (가)=4, (나)=6.7, (다)=1.5가 모두 계산과 맞지 않는다. 처리 자유도는 3, 오차 평균제곱은 5, F값은 2이다. 틀린 표현은 '(가)=4, (나)=6.7, (다)=1.5'이고 올바르게는 '(가)=3, (나)=5, (다)=2'이다.

오답 포인트

처리 자유도를 처리 수 4로 그대로 쓰는 실수가 자주 나온다. 또 총 제곱합과 오차 제곱합을 구분하지 않으면 MSE를 잘못 계산해 6.7 같은 값이 나오고, 그 오차가 그대로 F값까지 연쇄적으로 틀리게 만든다.

다시 나오면 이렇게 판단

분산분석표 계산형 문제는 처리 수와 표본수로 자유도를 먼저 확정하고, 총 제곱합에서 처리 제곱합을 빼서 오차 제곱합을 구한 뒤 평균제곱과 F값 순서로 계산하면 된다.

암기 포인트

일원분산분석은 df_처리=k-1, df_오차=N-k, $F = MST/MSE$의 흐름으로 정리된다.

NoteBase

문 18. 두 모집단의 평균을 비교하기 위해 가설 $H_0 : \mu_1 = \mu_2$ 대 $H_1 : \mu_1 \neq \mu_2$에 대한 검정을 실시하고자 한다. 두 모집단으로부터 각각 임의추출된 10, 12개의 자료에 대해 두 가지 검정법을 적용하였다. 방법 A로 독립인 두 모집단에 대한 t-검정, 방법 B로 수준이 2인 일원배치법을 적용할 때, 이에 대한 설명으로 옳은 것만을 모두 고르면? (단, 두 모집단은 서로 독립이며 분산이 동일한 정규분포를 따른다)

ㄱ. 방법 A에서의 검정통계량 t-값의 제곱은 방법 B에서의 검정통계량 F-값과 같다.
ㄴ. 두 검정 방법 A와 B에서의 p-값은 같다.
ㄷ. 방법 A에서의 검정통계량 t의 자유도는 방법 B에서의 검정통계량 F의 분자의 자유도와 같다.

① ㄱ, ㄴ
② ㄱ, ㄷ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 1번

출제 포인트

이 문제는 등분산 독립 2표본 t-검정과 수준 2인 일원배치분산분석의 동치 관계, 특히 $F = t^2$와 자유도 대응을 묻는 문제이다.

핵심 개념

독립 2표본 t-검정: 두 독립 집단의 평균 차이를 풀드 표준오차로 나누어 검정하는 방법이다. 키워드는 평균 비교, 등분산 가정, 자유도 $n_1 + n_2 - 2$이다.
수준 2인 일원배치분산분석: 집단 간 변동을 집단 내 변동으로 나눈 F-통계량으로 두 집단 이상의 평균 차이를 검정하는 방법이다. 키워드는 F-검정, 분자 자유도 $k - 1$, 분모 자유도 $N - k$이다.
두 집단만 비교할 때는 두 방법이 같은 귀무가설을 검정하며 $F = t^2$가 성립한다.

NoteBase

선지별 해설

① O: ㄱ, ㄴ만 옳다는 판단이 맞다. 두 집단 비교에서 등분산 독립 2표본 t-검정과 수준 2인 일원배치분산분석은 같은 귀무가설을 검정하므로 검정통계량 사이에 $F = t^2$ 관계가 성립한다. 따라서 ㄱ이 옳고, 같은 검정을 다른 형태로 표현한 것이므로 p-값도 같아 ㄴ도 옳다. ㄷ은 틀리므로 ①이 정답이다.
② X: ㄱ, ㄷ이라고 본 점이 틀리다. 틀린 표현은 ㄷ의 '분자의 자유도'이다. 실제로 방법 A의 t 자유도는 $n_1 + n_2 - 2 = 10 + 12 - 2 = 20$이고, 방법 B의 F에서 분자 자유도는 $k - 1 = 2 - 1 = 1$이므로 같지 않다. 올바르게는 '방법 A의 t 자유도는 방법 B의 F 분모 자유도와 같다'이다.
③ X: ㄱ을 빠뜨린 점이 틀리다. ㄱ에서 방법 A의 t-값을 제곱하면 방법 B의 F-값과 같아지므로 ㄱ은 옳다. 올바른 조합은 ㄴ, ㄷ이 아니라 ㄱ, ㄴ이다.
④ X: ㄷ까지 옳다고 본 점이 틀리다. ㄷ의 자유도 대응이 잘못되었기 때문이다. 올바르게는 'ㄱ, ㄴ만 옳다'이다.

오답 포인트

많이 헷갈리는 지점은 자유도 대응이다. 두 방법이 동치라고 해서 모든 자유도가 같아지는 것은 아니다. 이 문항에서는 t의 자유도 20이 F의 분자 자유도 1과 같다고 착각하면 ㄷ을 옳다고 보게 된다.

다시 나오면 이렇게 판단

두 집단 평균 비교에서 일원배치분산분석의 수준이 2이면 먼저 $F = t^2$를 떠올린다. 그다음 자유도는 t: $n_1 + n_2 - 2$, F: 분자 $k - 1$, 분모 $N - k$로 따로 비교하면 ㄷ 같은 함정을 바로 걸러낼 수 있다.

암기 포인트

두 집단 평균 비교에서는 $F = t^2$, p-값은 같고, t의 자유도는 F의 분모 자유도와 대응한다.

NoteBase

문 19. 어떤 화력발전소에서 배출되는 미세먼지의 일별 배출량은 표준편차가 1인 정규분포를 따르는 것으로 알려져 있다. 임의로 추출한 36일간의 미세먼지 일별 배출량의 평균과 실제 모평균의 차이가 0.1 이상일 확률을 나타내는 식은? (단, Z는 $N(0,1)$을 따르고, 미세먼지의 일별 배출량은 서로 독립이다)

① $P(|Z| \le 0.6)$

② $2 \times P(Z \ge 0.6)$

③ $P(|Z| \le 0.1)$

④ $2 \times P(Z \ge 0.1)$

정답 2번

출제 포인트

이 문제는 표본평균의 분포를 이용해 $|\bar{X} - \mu| \ge 0.1$의 확률을 표준정규분포 확률로 바꾸는 문제이다.

조건 정리

주어진 값: 일별 배출량의 표준편차는 1이고, 36일을 독립적으로 추출하였다.
구해야 할 값: 표본평균과 모평균의 차이가 0.1 이상일 확률이다.
주의할 조건: 개별 관측값이 아니라 표본평균 $\bar{X}$를 표준화해야 하므로 표준오차 $1/\sqrt{36} = 1/6$을 사용한다.

사용할 공식

표본평균의 분포는 $\bar{X} \sim N\left(\mu, \frac{1}{36}\right)$이고, 표준화하면
$Z = \frac{\bar{X} - \mu}{1/6} \sim N(0, 1)$
이다.
이 공식을 쓰는 이유는 문제에서 모표준편차와 표본크기가 주어졌고, 표본평균의 모평균으로부터의 이탈 확률을 묻기 때문이다.

계산 과정

1. 표본평균의 표준오차는 $\sigma/\sqrt{n} = 1/\sqrt{36} = 1/6$이다.
2. 구하려는 확률은 $P(|\bar{X} - \mu| \ge 0.1)$이다.
3. 이를 표준화하면
$P\left(\left|\frac{\bar{X} - \mu}{1/6}\right| \ge \frac{0.1}{1/6}\right) = P(|Z| \ge 0.6)$
이다.
4. 표준정규분포의 대칭성을 쓰면 $P(|Z| \ge 0.6) = 2P(Z \ge 0.6)$이다.

NoteBase

선지별 해설

① X: $P(|Z| \le 0.6)$는 중앙구간에 들어갈 확률이다. 문제는 차이가 0.1 이상일 확률이므로 양쪽 꼬리확률을 구해야 한다. 틀린 표현은 $|Z| \le 0.6$이고, 올바르게는 $|Z| \ge 0.6$이다.
② O: 표준화 결과가 $P(|Z| \ge 0.6)$이고, 대칭성을 이용하면 $2P(Z \ge 0.6)$가 된다. 표본평균의 표준오차를 반영한 식이므로 정답이다.
③ X: $P(|Z| \le 0.1)$은 표준화도 하지 않았고 부등호 방향도 반대이다. 틀린 표현은 기준값 0.1을 그대로 Z에 대입한 점과 $|Z| \le 0.1$이다. 올바르게는 $|Z| \ge 0.6$이다.
④ X: 양측확률 형태는 맞지만 기준값이 틀렸다. 0.1을 그대로 쓰면 표준오차 1/6으로 나누는 과정을 빠뜨린 것이다. 틀린 표현은 $2P(Z \ge 0.1)$이고, 올바르게는 $2P(Z \ge 0.6)$이다.

오답 포인트

출제자가 노린 함정은 0.1을 그대로 Z값으로 착각하게 만드는 점과, '이상'을 '이하'로 바꾸어 중앙확률을 고르게 만드는 점이다. 표본평균 문제에서는 항상 표준오차로 나누는 단계가 먼저이다.

다시 나오면 이렇게 판단

표본평균 확률 문제는 먼저 $\sigma/\sqrt{n}$으로 표준오차를 만든 뒤, 물은 사건이 꼬리확률인지 중앙확률인지 확인해서 Z 식으로 바꾸면 된다.

암기 포인트

표본평균은 σ가 아니라 $\sigma/\sqrt{n}$으로 표준화하고, '이상'은 꼬리확률로 처리한다.

NoteBase

문 20. 변수 X와 Y 사이의 관계를 단순선형회귀모형 $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ $(i = 1, 2, \cdots, n)$로 분석하고자 한다. 이를 위해 β_0, β_1의 최소제곱추정량 $\hat{\beta}_0$, $\hat{\beta}_1$을 각각 구하고 잔차 $e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$를 구하였다. 이와 관련된 설명으로 옳지 않은 것은? (단, ε_i는 $N(0, \sigma^2)$을 따르고 서로 독립이다)

① $\hat{\beta}_0, \hat{\beta}_1$은 각각 β_0, β_1의 불편추정량이다.
② 가설 $H : \beta_1 = 0$은 X와 Y 간에 어떠한 관계성도 없음을 의미한다.
③ 잔차들의 x_i에 대한 가중합 $\sum_{i=1}^{n} x_i e_i$는 0이다.
④ $\hat{\beta}_1$의 분산은 $\frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$이다.

정답 2번

출제 포인트

이 문제는 단순선형회귀모형에서 최소제곱추정량의 성질, 잔차의 직교조건, 그리고 가설 $H : \beta_1 = 0$의 의미를 정확히 구별하는지를 묻는다.

핵심 개념

불편추정량은 추정량의 기대값이 모수와 일치하는 추정량이다.
$H : \beta_1 = 0$은 X가 Y의 조건부 평균에 선형 영향을 주지 않는다는 뜻이다.
절편이 포함된 OLS에서는 잔차가 상수항과 설명변수 x_i에 직교하므로 정규방정식이 성립한다.

선지별 해설

1번: O. 오차항의 평균이 0이고 고전적 선형회귀 가정이 성립하면 $\hat{\beta}_0, \hat{\beta}_1$의 기대값은 각각 β_0, β_1가 되므로 불편추정량이다.
2번: X. 틀린 부분은 "어떠한 관계성도 없음"이다. $\beta_1 = 0$은 $E(Y \mid X) = \beta_0$로 조건부 평균의 선형 기울기가 0이라는 뜻일 뿐이다. 올바르게는 "$\beta_1 = 0$은 X와 Y 사이에 선형적인 평균 관계가 없음을 의미한다"이다.
3번: O. 절편이 포함된 최소제곱법에서는 잔차에 대한 정규방정식이 성립하므로 $\sum_{i=1}^{n} x_i e_i = 0$이다. 이는 잔차가 설명변수와 직교함을 뜻한다.
4번: O. 단순선형회귀에서 $\hat{\beta}_1$의 분산은 설명변수의 산포 $\sum_{i=1}^{n} (x_i - \bar{x})^2$에 반비례하므로 $\mathrm{Var}(\hat{\beta}_1) = \sigma^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2$이다.

오답 포인트

2번의 "$\beta_1 = 0$이면 관계가 전혀 없다"는 과장이 함정이다. 회귀계수 0은 선형 평균관계의 부재를 뜻할 뿐이며, 무상관이나 독립까지 바로 뜻하지는 않는다.

NoteBase

다시 나오면 이렇게 판단

회귀계수에 대한 진술이 나오면 먼저 그것이 평균함수의 기울기를 말하는지, 변수 사이의 모든 관계를 말하는지 구분한다. 이어서 OLS 기본 성질은 불편성, 잔차의 직교성, $\hat{\beta}_1$ 분산식의 세 축으로 정리해 판단한다.

암기 포인트

$\beta_1 = 0$은 무관계가 아니라 평균직선의 기울기 0이다. OLS 잔차는 1과 x에 직교한다.

NoteBase

문 21. 0과 1로만 이루어져 있고 1의 비율이 p인 모집단으로부터 5개의 표본을 임의추출한 결과, 4개가 1로 관측되었다. 가설 $H_0 : p = \frac{2}{3}$ 대 $H_1 : p > \frac{2}{3}$에 대한 검정에서 유의확률은?

① $\frac{1}{3^5}$
② $\frac{11}{3^5}$
③ $\frac{80}{3^5}$
④ $\frac{112}{3^5}$

정답 4번

출제 포인트

귀무가설 아래에서 이항분포의 오른쪽 꼬리확률 $P(X \geq 4)$를 계산하는 단측검정 유의확률 문제이다.

조건 정리

표본크기는 5이고, 귀무가설 아래 성공확률은 $p_0 = 2/3$이다.
관측된 성공횟수는 4이므로 검정통계량 X는 $X \sim Bin(5, 2/3)$를 따른다.
대립가설이 $H_1 : p > 2/3$이므로 오른쪽 꼬리확률을 구한다.

사용할 공식

유의확률은 $P\text{-value} = P(X \geq x_{obs} \mid H_0)$이다.
이항확률은 $P(X = k) = \binom{5}{k}(2/3)^k(1/3)^{5-k}$이다.
이 문제는 0과 1만 나오는 5회의 독립 시행에서 1의 개수를 보므로 이항분포 공식을 사용한다.

계산 과정

1. 유의확률은 $P(X \geq 4) = P(X = 4) + P(X = 5)$이다.
2. $P(X = 4) = \binom{5}{4}(2/3)^4(1/3) = 5 \cdot 16/81 \cdot 1/3 = 80/243$이다.
3. $P(X = 5) = \binom{5}{5}(2/3)^5 = 32/243$이다.
4. 따라서 유의확률은 $80/243 + 32/243 = 112/243 = 112/3^5$이다.

NoteBase

선지별 해설

① X: $1/3^5 = 1/243$는 귀무가설 아래에서 $X \geq 4$가 일어날 확률이 아니다. 틀린 표현은 유의확률을 오른쪽 꼬리확률로 계산하지 않은 부분이다. 올바르게는 $P(X=4)+P(X=5)=112/243$이다.
② X: $11/3^5 = 11/243$는 귀무가설 아래 관측값 이상이 나올 확률과 일치하지 않는다. 틀린 표현은 필요한 경우의 수를 부분적으로만 더한 부분이다. 올바르게는 $X=4$와 $X=5$의 확률을 모두 합한 $112/243$이다.
③ X: $80/3^5 = 80/243$는 $P(X=4)$만 계산한 값이다. 틀린 표현은 '4개가 1로 관측되었다'를 '정확히 4개'로만 해석한 부분이다. 올바르게는 단측검정이므로 그보다 더 극단적인 $X=5$까지 포함해 $112/243$를 써야 한다.
④ O: $112/3^5 = 112/243$는 $P(X=4)+P(X=5)$와 같으므로 유의확률이 맞다.

오답 포인트

유의확률을 관측값의 한 점확률 $P(X=4)$로 착각하면 ③을 고르게 된다. 대립가설이 $p > 2/3$인 오른쪽 단측검정이므로 관측값보다 더 극단적인 $X=5$도 반드시 포함한다.

다시 나오면 이렇게 판단

이항 단측검정은 먼저 귀무가설 아래 분포를 적고, 다음에 대립가설의 방향을 확인한 뒤, 그 방향의 꼬리확률을 관측값 이상 또는 이하로 합산해 판단한다.

암기 포인트

단측 유의확률은 한 점확률이 아니라 대립가설 방향의 꼬리합이다.

NoteBase

문 22. 다음은 단순선형회귀모형 $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ $(i = 1, 2, \cdots, n)$를 적합하기 위해 어떤 자료를 요약한 결과이다. 이에 대한 설명으로 옳지 않은 것은? (단, ε_i는 $N(0, \sigma^2)$을 따르고 서로 독립이다)

$n = 12, \bar{x} = 1.5, \bar{y} = 8$
$\sum_{i=1}^{12}(x_i - \bar{x})^2 = 10, \sum_{i=1}^{12}(y_i - \bar{y})^2 = 90, \sum_{i=1}^{12}(x_i - \bar{x})(y_i - \bar{y}) = 20$

① β_1의 최소제곱추정값은 2이다.

② β_0의 최소제곱추정값은 5이다.

③ 잔차의 제곱합은 40이다.

④ 오차분산의 추정값은 5이다.

정답 3번

출제 포인트

이 문제는 단순선형회귀에서 최소제곱추정량 $\hat{\beta}_1, \hat{\beta}_0$, 잔차의 제곱합 SSE, 오차분산의 추정량 $\hat{\sigma}^2$를 요약통계량만으로 계산할 수 있는지를 묻는 문제이다.

조건 정리

주어진 값은 $n = 12, \bar{x} = 1.5, \bar{y} = 8, S_{xx} = \sum_{i=1}^{12}(x_i - \bar{x})^2 = 10, S_{yy} = \sum_{i=1}^{12}(y_i - \bar{y})^2 = 90,$
$S_{xy} = \sum_{i=1}^{12}(x_i - \bar{x})(y_i - \bar{y}) = 20$이다.
구해야 할 값은 $\hat{\beta}_1, \hat{\beta}_0, SSE, \hat{\sigma}^2$이다.
주의할 조건은 오차분산의 추정량을 구할 때 분모를 n이 아니라 자유도 $n - 2$로 둔다는 점이다.

사용할 공식

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
$SSE = S_{yy} - \hat{\beta}_1 S_{xy} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}, \qquad \hat{\sigma}^2 = \frac{SSE}{n-2}$
이 공식을 쓰는 이유는 단순선형회귀의 최소제곱추정량과 잔차제곱합이 중심화된 요약통계량 S_{xx}, S_{yy}, S_{xy}로 바로 표현되기 때문이다.

계산 과정

1. 기울기는 $\hat{\beta}_1 = \frac{20}{10} = 2$이다.
2. 절편은 $\hat{\beta}_0 = 8 - 2 \times 1.5 = 5$이다.
3. 잔차의 제곱합은 $SSE = 90 - 2 \times 20 = 50$이다. 같은 값으로 $90 - \frac{20^2}{10} = 50$으로 계산해도 된다.
4. 오차분산의 추정량은 $\hat{\sigma}^2 = \frac{50}{12-2} = 5$이다.

NoteBase

선지별 해설

1번 O
$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{20}{10} = 2$이므로 옳은 선지이다.
2번 O
$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 8 - 2 \times 1.5 = 5$이므로 옳은 선지이다.
3번 X
잔차의 제곱합은 $SSE = 90 - 2 \times 20 = 50$이므로 틀린 선지이다. 틀린 표현은 "잔차의 제곱합은 40이다"이고, 올바르게는 "잔차의 제곱합은 50이다"이다.
4번 O
오차분산의 추정량은 $\hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{50}{10} = 5$이므로 옳은 선지이다.

오답 포인트

잔차의 제곱합을 구할 때 $\hat{\beta}_1 S_{xy} = 40$까지만 계산하고 끝내면 ③처럼 오답이 된다. 40은 설명된 제곱합의 크기이고, 잔차의 제곱합은 전체 제곱합 90에서 이를 뺀 50이다. 또한 오차분산의 추정량은 단순회귀에서 자유도 $n-2$를 써야 한다.

다시 나오면 이렇게 판단

단순회귀 요약통계 문제가 나오면 먼저 $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$, 다음에 $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, 그다음 $SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$, 마지막으로 $\hat{\sigma}^2 = \frac{SSE}{n-2}$ 순서로 계산하면 된다.

암기 포인트

단순선형회귀에서는 $\hat{\beta}_1 = S_{xy}/S_{xx}$, $SSE = S_{yy} - S_{xy}^2/S_{xx}$, $\hat{\sigma}^2 = SSE/(n-2)$가 기본식이다.

NoteBase

문 23. 어느 도시에서 특정 지역축제의 개최에 대한 찬성률(p)을 알아보기 위해 시민 100명을 대상으로 작년에 처음 실시한 여론조사에서, 찬성률의 추정에 95% 신뢰수준을 사용하였다. 올해 실시 예정인 동일한 여론조사에서, 오차한계를 포함한 다른 조건은 모두 작년과 동일하게 하되 표본 수를 줄이기 위해 신뢰수준만 90%로 바꾸고자 한다. 올해 조사해야 하는 최소 표본의 수는? (단, z_α는 표준정규분포의 제$100 \times (1-\alpha)$ 백분위수를 나타내고, 계산의 편의상 $z_{0.1} = 1.3$, $z_{0.05} = 1.6$, $z_{0.025} = 2$를 사용하기로 한다)

① 43명

② 64명

③ 87명

④ 92명

정답 2번

출제 포인트

이 문제는 비율의 신뢰구간 오차한계가 같을 때 필요한 표본 수가 임계값의 제곱에 비례한다는 점을 묻는 문제이다.

조건 정리

주어진 값:
- 작년 표본 수는 100명이다.
- 작년 신뢰수준 95%의 양측 신뢰구간에서 $z_{0.025} = 2$를 사용한다.
- 올해 신뢰수준 90%의 양측 신뢰구간에서 $z_{0.05} = 1.6$을 사용한다.

구해야 할 값:
- 올해 조사에 필요한 최소 표본 수이다.

주의할 조건:
- 오차한계와 그 밖의 조건은 작년과 동일하므로 $p(1-p)$는 같게 본다.

사용할 공식

비율 추정의 오차한계는 $E = z_{\alpha/2}\sqrt{p(1-p)/n}$이다.
같은 오차한계에서는 $n \propto z_{\alpha/2}^2$이므로
$n_2 = n_1\left(\frac{z_{0.05}}{z_{0.025}}\right)^2$
이다.
이 공식을 쓰는 이유:
- 올해와 작년은 신뢰수준만 다르고 오차한계와 $p(1-p)$ 구조는 같기 때문이다.

NoteBase

계산 과정

1. 작년 오차한계를 $E = 2\sqrt{p(1-p)/100}$로 둔다.
2. 올해 오차한계를 $E = 1.6\sqrt{p(1-p)/n}$로 둔다.
3. 두 오차한계를 같게 두면 $2\sqrt{p(1-p)/100} = 1.6\sqrt{p(1-p)/n}$이다.
4. 공통항 $p(1-p)$를 소거하면
$n = 100\left(\frac{1.6}{2}\right)^2 = 100 \times 0.64 = 64$
이다.
5. 최소 표본 수는 64명이다.

선지별 해설

① X
- '43명'은 90% 양측 신뢰구간에서 써야 할 $z_{0.05} = 1.6$ 대신 $z_{0.1} = 1.3$을 사용해 $100(1.3/2)^2 \approx 42.25$를 반올림한 값에 해당한다. 올바르게는 64명이다.

② O
- 같은 오차한계에서 표본 수는 임계값의 제곱에 비례하므로 $100(1.6/2)^2 = 64$가 된다. 따라서 64명이 맞다.

③ X
- '87명'은 동일 오차한계 식 $n = 100(1.6/2)^2$를 만족하지 않는다. 신뢰수준을 95%에서 90%로 낮추면 필요한 표본 수는 100명보다 더 크게 줄어 64명이 된다. 올바르게는 64명이다.

④ X
- '92명'도 임계값 제곱비를 적용한 결과와 맞지 않는다. 오차한계가 같을 때는 감소 폭을 선형이 아니라 제곱비로 계산해야 하므로 올바르게는 64명이다.

오답 포인트

90% 신뢰수준의 양측구간인데 $z_{0.1}$을 쓰거나, 표본 수가 z에 비례한다고 보고 제곱을 빠뜨리면 오답이 나온다. 이 문항의 함정은 신뢰수준을 낮출 때 표본 수 감소를 직관적으로만 판단하게 만드는 점이다.

다시 나오면 이렇게 판단

비율 추정 표본 수 문제는 먼저 오차한계 식 $E = z_{\alpha/2}\sqrt{p(1-p)/n}$를 적고, 같게 유지되는 항과 바뀌는 항을 분리해 본다. 오차한계와 $p(1-p)$가 같으면 n은 z의 제곱비로 바로 비교한다.

암기 포인트

같은 오차한계에서 비율 추정 표본 수는 $n \propto z_{\alpha/2}^2$이다. 95%에서 90%로 낮추면 2가 1.6으로 바뀌어 표본 수는 $0.8^2 = 0.64$배가 된다.

NoteBase

문 24. 다음은 어느 대학의 A학과와 B학과 간에 통계학 성적에 차이가 있는지를 알아보기 위해 두 학과에서 각각 11명, 16명을 임의추출하여 조사한 결과이다. 두 학과의 통계학 성적 간에 차이가 있는지를 검정하기 위한 t-검정통계량(t)의 값과 유의수준 5%에서의 기각역으로 바르게 연결한 것은? (단, 각 학과의 통계학 성적은 서로 독립이며 정규분포를 따르고 분산은 같다. 또한, $t_\alpha(k)$는 자유도가 k인 t분포의 제$100 \times (1-\alpha)$ 백분위수이다)

학과	표본 수	평균	표준편차
A	11	70	6
B	16	65	4

① t-검정통계량(t)의 값: $\frac{70-65}{\sqrt{24 \times (\frac{1}{11}+\frac{1}{16})}}$, 기각역: $|t| \geq t_{0.025}(26)$
② t-검정통계량(t)의 값: $\frac{70-65}{\sqrt{\frac{36}{11}+\frac{16}{16}}}$, 기각역: $|t| \geq t_{0.025}(26)$
③ t-검정통계량(t)의 값: $\frac{70-65}{\sqrt{24 \times (\frac{1}{11}+\frac{1}{16})}}$, 기각역: $|t| \geq t_{0.025}(25)$
④ t-검정통계량(t)의 값: $\frac{70-65}{\sqrt{\frac{36}{11}+\frac{16}{16}}}$, 기각역: $|t| \geq t_{0.025}(25)$

정답 3번

출제 포인트

이 문제는 등분산을 가정한 두 독립표본 평균차 검정에서 풀드분산을 사용한 t-검정통계량과 자유도 $n_1 + n_2 - 2$를 정확히 연결하는 문제이다.

조건 정리

주어진 값은 $n_1 = 11, n_2 = 16, \bar{x}_1 = 70, \bar{x}_2 = 65, s_1 = 6, s_2 = 4$이다.
구해야 할 값은 두 학과 평균 차이 검정을 위한 t-검정통계량과 5% 유의수준의 양측 기각역이다.
주의할 조건은 두 집단이 서로 독립이고 정규분포를 따르며 분산이 같다는 점이다.

사용할 공식

등분산 두 독립표본 t-검정에서는

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2(\frac{1}{n_1}+\frac{1}{n_2})}}$$

를 사용하고 자유도는 $n_1 + n_2 - 2$이다.
이 공식을 쓰는 이유는 문제에서 두 집단의 분산이 같다고 주어졌으므로 분리 추정이 아니라 풀드분산 추정을 써야 하기 때문이다.

NoteBase

계산 과정

1. 표본분산은 $s_1^2 = 36$, $s_2^2 = 16$이다.

2. 풀드분산은

$s_p^2 = \frac{10 \cdot 36 + 15 \cdot 16}{11 + 16 - 2} = \frac{360 + 240}{25} = 24$

이다.

3. 따라서 검정통계량은

$t = \frac{70 - 65}{\sqrt{24\left(\frac{1}{11} + \frac{1}{16}\right)}}$

이다.

4. 자유도는 $11 + 16 - 2 = 25$이므로 5% 유의수준 양측검정의 기각역은

$|t| \geq t_{0.025}(25)$

이다.

선지별 해설

① X

이 선지는 분모의 $24\left(\frac{1}{11} + \frac{1}{16}\right)$는 맞지만 자유도를 26으로 두었다는 점이 틀리다. 등분산 두 독립표본 t-검정의 자유도는 $n_1 + n_2 - 2 = 25$이다. 틀린 표현은 $t_{0.025}(26)$이고, 올바르게는 $t_{0.025}(25)$이다.

② X

이 선지는 자유도를 26으로 둔 점도 틀리고, 분모를 $\sqrt{\frac{36}{11} + \frac{16}{16}}$로 쓴 점도 틀리다. 이는 등분산 가정을 무시하고 각 표본분산을 그대로 나눈 형태인데, 이 문제는 풀드분산 $s_p^2 = 24$를 써야 한다. 틀린 표현은 $\sqrt{\frac{36}{11} + \frac{16}{16}}$, $t_{0.025}(26)$이고, 올바르게는 $\sqrt{24\left(\frac{1}{11} + \frac{1}{16}\right)}$, $t_{0.025}(25)$이다.

③ O

분모에 풀드분산 24가 들어가고 자유도도 25로 맞게 연결되어 있다. 문제의 조건인 등분산 가정과 양측검정 5% 유의수준을 정확히 반영한 식이므로 옳다.

④ X

자유도 25는 맞지만 분모가 틀리다. 등분산 가정에서는 $\frac{36}{11} + \frac{16}{16}$처럼 각 표본분산을 직접 더하지 않고, 먼저 풀드분산 24를 구한 뒤 $24\left(\frac{1}{11} + \frac{1}{16}\right)$를 사용한다. 틀린 표현은 $\sqrt{\frac{36}{11} + \frac{16}{16}}$이고, 올바르게는 $\sqrt{24\left(\frac{1}{11} + \frac{1}{16}\right)}$이다.

오답 포인트

이 문항의 함정은 표본표준편차 6과 4를 보자마자 $\frac{36}{11} + \frac{16}{16}$ 형태를 쓰거나, 자유도를 $n_1 + n_2 - 1$로 착각하는 데 있다. 등분산 조건이 보이면 먼저 풀드분산을 만드는지부터 확인해야 한다.

NoteBase

다시 나오면 이렇게 판단

두 독립표본 평균 비교 문제는 먼저 독립성, 정규성, 등분산 가정을 확인한다. 등분산이면 풀드분산을 계산하고 검정통계량 분모에 $s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$를 넣으며, 자유도는 항상 n_1+n_2-2로 판단한다.

암기 포인트

등분산 두 집단 평균차 검정은 풀드분산 사용, 자유도는 n_1+n_2-2이다.

NoteBase

문 25. 새로 제작된 3개의 동전을 동시에 던지는 실험을 독립적으로 80회 실시하였다. 다음 표는 '각 실험에서 앞면이 나온 수(X)'의 빈도(도수)를 나타낸다.

가설 '$H_0 : X$가 이항분포 $B(3, \frac{1}{2})$을 따른다'에 대한 카이제곱검정을 실시할 때, 다음 설명 중 옳지 않은 것은? (단, $\chi^2_\alpha(k)$는 자유도가 k인 카이제곱분포의 제$100 \times (1-\alpha)$ 백분위수이다)

X	H_0하의 기대도수	관측도수
0	E_1	8
1	E_2	27
2	E_3	34
3	E_4	11
계	80	80

① E_2는 30이다.

② H_0하에서 $P(X=2)$는 $\frac{3}{8}$이다.

③ 검정통계량의 값은 $\frac{5}{3}$이다.

④ 유의수준 5%인 검정의 임계값은 $\chi^2_{0.05}(3)$이다.

정답 3번

출제 포인트

이 문제는 $X \sim B\left(3, \frac{1}{2}\right)$ 아래에서 각 값의 기대도수, 카이제곱 검정통계량, 자유도를 계산할 수 있는지를 묻는 문제이다.

조건 정리

주어진 값: 3개의 동전을 동시에 던지는 실험을 80회 시행하였고, 관측도수는 $X=0, 1, 2, 3$에 대해 각각 8, 27, 34, 11이다.
구해야 할 값: 귀무가설 $H_0 : X \sim B\left(3, \frac{1}{2}\right)$ 아래의 기대도수와 카이제곱 검정의 검정통계량, 임곗값의 자유도이다.
주의할 조건: 모수 $p=\frac{1}{2}$를 자료에서 추정한 것이 아니라 귀무가설에서 바로 주었으므로 자유도는 범주 수 4에서 1을 뺀 3이다.

NoteBase

사용할 공식

기대도수는 $E_i = np_i$이다.
카이제곱 검정통계량은 $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$이다.
이 공식을 쓰는 이유: 귀무가설이 각 범주의 확률 p_i를 주고 있으므로, 그 확률로 기대도수를 만든 뒤 관측도수와의 차이를 카이제곱 통계량으로 비교하는 적합도 검정 상황이기 때문이다.

계산 과정

1. $X \sim B\left(3, \frac{1}{2}\right)$이므로 확률은
$P(X=0) = \frac{1}{8}, \quad P(X=1) = \frac{3}{8}, \quad P(X=2) = \frac{3}{8}, \quad P(X=3) = \frac{1}{8}$
이다.
2. 따라서 기대도수는
$E_1 = 80 \cdot \frac{1}{8} = 10, \quad E_2 = 80 \cdot \frac{3}{8} = 30, \quad E_3 = 80 \cdot \frac{3}{8} = 30, \quad E_4 = 80 \cdot \frac{1}{8} = 10$
이다.
3. 검정통계량은
$\chi^2 = \frac{(8-10)^2}{10} + \frac{(27-30)^2}{30} + \frac{(34-30)^2}{30} + \frac{(11-10)^2}{10} = \frac{4}{10} + \frac{9}{30} + \frac{16}{30} + \frac{1}{10} = \frac{4}{3}$
이다.
4. 범주가 4개이고 추가로 추정한 모수가 없으므로 자유도는 $4 - 1 = 3$이다. 따라서 유의수준 5%의 임곗값은 $\chi^2_{0.05}(3)$이다.

선지별 해설

① O
E_2는 $X = 1$일 때의 기대도수이므로 $80 \cdot \frac{3}{8} = 30$이다. 따라서 ①은 옳다.

② O
귀무가설 아래에서 $X \sim B\left(3, \frac{1}{2}\right)$이므로 $P(X=2) = \binom{3}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right) = \frac{3}{8}$이다. 따라서 ②는 옳다.

③ X
검정통계량은 각 범주의 $\frac{(O_i - E_i)^2}{E_i}$를 모두 더한 값이므로 계산 결과는 $\frac{4}{3}$이다. 보기의 $\frac{5}{3}$은 잘못된 값이다. 틀린 표현은 "검정통계량의 값은 $\frac{5}{3}$이다"이고, 올바르게는 "검정통계량의 값은 $\frac{4}{3}$이다"이다.

④ O
적합도 검정에서 자유도는 범주 수 4에서 1을 뺀 3이다. 이 문제에서는 자료로 추정한 모수가 없으므로 추가 차감이 없다. 따라서 유의수준 5%의 임곗값을 $\chi^2_{0.05}(3)$로 두는 ④는 옳다.

NoteBase

오답 포인트

검정통계량을 더할 때 각 항의 분자 제곱이나 분모를 잘못 계산하면 ③처럼 오답을 고르기 쉽다. 또한 자유도를 4로 착각하거나, 범주 수에서 1을 빼야 한다는 점을 놓치면 ④도 흔들리기 쉽다.

다시 나오면 이렇게 판단

적합도 검정 문제는 먼저 귀무가설의 범주별 확률을 구하고, 전체 시행 횟수를 곱해 기대도수를 만든 뒤, $\chi^2 = \sum \frac{(O-E)^2}{E}$를 계산한다. 마지막에는 자유도를 범주 수에서 1을 뺀 값으로 확인하면 된다.

암기 포인트

적합도 검정은 확률 → 기대도수 → 카이제곱합 → 자유도 확인의 순서로 정리된다.

NoteBase