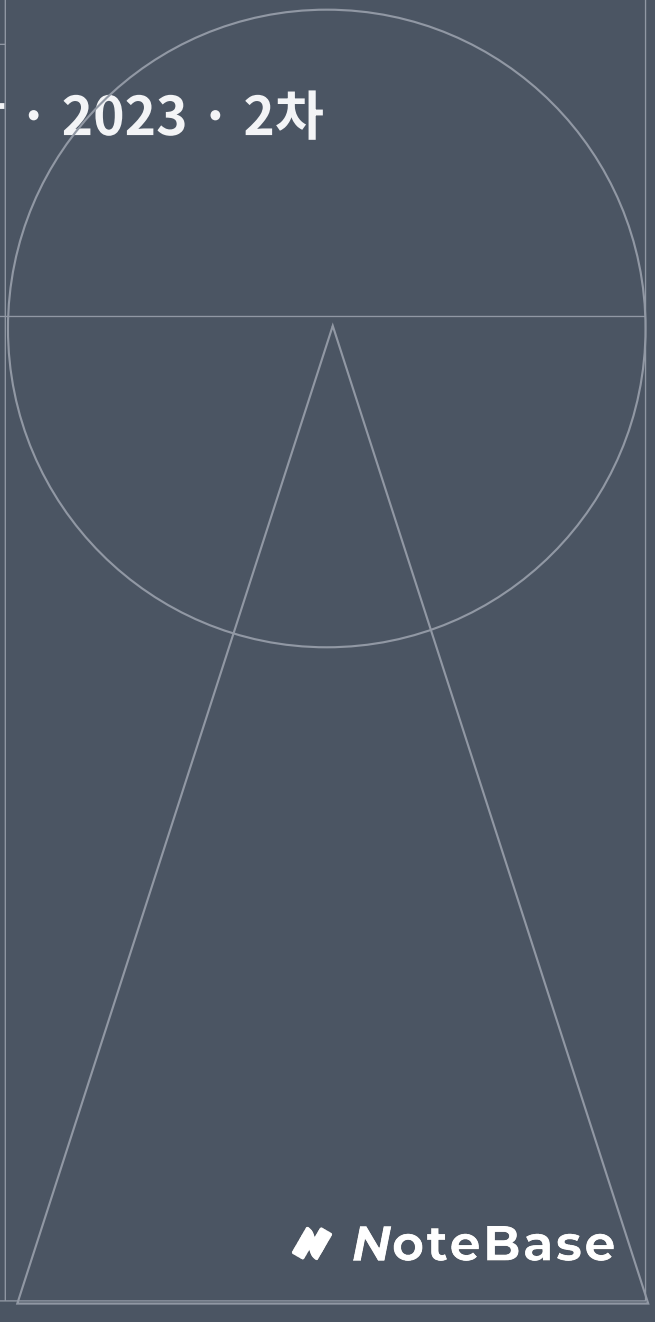
기출OX 해설집

공무원/국가직/7급 · 통계학 · 2023 · 2차



NoteBase

기출OX 해설집

공무원/국가직/7급 · 통계학 · 2023 · 2차

공무원/국가직/7급 · 통계학 · 2023 · 2차

문 1. 다음은 남학생과 여학생 각 15명의 통계학 성적을 줄기-잎 그림으로 나타낸 것이다. 이에 대한 설명으로 옳은 것만을 모두 고르면?



ㄱ. 중앙값은 남학생이 여학생보다 3점 작다.
ㄴ. 사분위수범위는 남학생과 여학생이 같다.
ㄷ. 표준편차는 여학생이 남학생보다 크다.

① ㄱ
② ㄱ, ㄴ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 2번

출제 포인트

이 문제는 줄기-잎 그림에서 두 집단의 중앙값, 사분위수범위, 표준편차를 정확히 읽어 참인 설명의 조합을 고르는 문제이다.

NoteBase

자료 읽는 순서

1. 줄기와 잎을 실제 점수로 복원한다.
2. 각 집단의 중앙값과 사분위수를 구해 사분위수범위를 비교한다.
3. 두 자료가 전체적으로 같은 모양인지, 일정한 값만큼 평행이동한 자료인지 확인해 산포를 판단한다.

핵심 근거

남학생 점수는 51, 52, 63, 64, 65, 71, 72, 73, 74, 78, 81, 82, 93, 94, 95이고 여학생 점수는 54, 55, 66, 67, 68, 74, 75, 76, 77, 81, 84, 85, 96, 97, 98이다.
중앙값은 각각 73점, 76점이므로 남학생이 3점 작다.
사분위수는 남학생 $Q_1 = 64$, $Q_3 = 82$, 여학생 $Q_1 = 67$, $Q_3 = 85$이므로 두 집단의 사분위수범위는 모두 18이다.
여학생의 각 점수는 대응하는 남학생 점수에 3점을 더한 형태이므로 자료 전체가 평행이동한 것과 같아 표준편차는 서로 같다.

선지별 해설

1번 X: ㄱ만 고른 것이 틀리다. ㄱ은 중앙값이 남학생 73점, 여학생 76점이어서 옳지만, ㄴ도 사분위수범위가 둘 다 18이어서 함께 옳다. 올바르게는 'ㄱ, ㄴ'이다.
2번 O: ㄱ, ㄴ만 옳다. ㄱ은 중앙값 차이가 3점이라는 점에서 맞고, ㄴ은 사분위수범위가 두 집단 모두 18이어서 맞다. ㄷ은 여학생 자료가 남학생 자료 전체를 3점씩 평행이동한 것이므로 표준편차가 커지지 않아 틀리다.
3번 X: ㄴ은 옳지만 ㄷ을 포함한 것이 틀리다. ㄷ의 '여학생이 남학생보다 크다'는 표현이 잘못되었고, 올바르게는 '여학생과 남학생의 표준편차는 같다'이다. 따라서 올바른 조합은 'ㄱ, ㄴ'이다.
4번 X: ㄱ, ㄴ은 옳지만 ㄷ까지 모두 옳다고 본 것이 틀리다. 일정한 수를 모든 자료에 더해도 표준편차는 변하지 않으므로, 올바르게는 'ㄱ, ㄴ만 옳다'이다.

오답 포인트

표준편차를 평균과 함께 이동한다고 착각하면 ㄷ을 참으로 보기 쉽다. 모든 점수에 같은 수를 더하면 중심은 이동하지만 퍼짐의 정도는 변하지 않는다. 또 사분위수범위는 중앙값 차이와 별개이므로 사분위수를 따로 계산해야 한다.

다시 나오면 이렇게 판단

줄기-잎 그림 비교 문제는 먼저 자료를 순서대로 복원한 뒤 중앙값과 Q_1, Q_3를 직접 찾는다. 두 자료가 항목별로 일정한 값만큼만 차이 나면 평균과 중앙값은 달라져도 표준편차와 사분위수범위 같은 산포 척도는 그대로라고 판단할 수 있다.

NoteBase

암기 포인트

모든 자료에 같은 수를 더해도 표준편차와 사분위수범위는 변하지 않는다.

NoteBase

문 2. 모집단 $\{2, 4\}$로부터 임의추출한 확률변수 X와 모집단 $\{1, 5\}$로부터 임의추출한 확률변수 Y에 대해 $X + Y$의 분산은? (단, X와 Y는 독립이다)

① 4

② 5

③ 8

④ 10

정답 2번

출제 포인트

이 문제는 서로 독립인 두 확률변수의 합에 대한 분산을 구하는 문제이다.

조건 정리

주어진 값:
- 확률변수 X는 {2, 4}에서 각각 같은 확률로 추출된다.
- 확률변수 Y는 {1, 5}에서 각각 같은 확률로 추출된다.
- X와 Y는 독립이다.

구해야 할 값:
- $Var(X + Y)$

주의할 조건:
- 독립이면 $Var(X + Y) = Var(X) + Var(Y)$가 성립한다.

사용할 공식

$Var(X + Y) = Var(X) + Var(Y)$

$Var(X) = E(X^2) - \{E(X)\}^2, \quad Var(Y) = E(Y^2) - \{E(Y)\}^2$

이 공식을 쓰는 이유:
- 문제에서 X와 Y가 독립이라고 했으므로 공분산이 0이 되어 두 분산을 그대로 더할 수 있다.

계산 과정

1. X의 평균은 $E(X) = (2 + 4)/2 = 3$이다.
2. X의 분산은 $\{(2 - 3)^2 + (4 - 3)^2\}/2 = (1 + 1)/2 = 1$이다.
3. Y의 평균은 $E(Y) = (1 + 5)/2 = 3$이다.
4. Y의 분산은 $\{(1 - 3)^2 + (5 - 3)^2\}/2 = (4 + 4)/2 = 4$이다.
5. 따라서 $Var(X + Y) = Var(X) + Var(Y) = 1 + 4 = 5$이다.

NoteBase

선지별 해설

① X
- 4는 Y의 분산값만 반영한 수치이므로 틀리다.
- 틀린 표현 '분산이 4이다' → 올바르게는 '합의 분산은 $Var(X) + Var(Y) = 5$이다.'

② O
- $Var(X) = 1$, $Var(Y) = 4$이고 두 변수는 독립이므로 합의 분산은 5가 된다.

③ X
- 8은 평균을 기준으로 한 분산 계산이 아니라 편차를 잘못 합치거나 범위만 보고 판단한 값이므로 틀리다.
- 틀린 표현 '분산이 8이다' → 올바르게는 '평균 3을 기준으로 편차제곱의 평균을 구하면 합의 분산은 5이다.'

④ X
- 10은 X와 Y의 편차제곱합 $2 + 8$을 평균내지 않고 그대로 더한 값이어서 틀리다.
- 틀린 표현 '분산이 10이다' → 올바르게는 '편차제곱합을 각 경우의 수로 나누어 분산을 구하면 결과는 5이다.'

오답 포인트

분산을 구할 때 편차제곱합만 더하고 평균을 내지 않거나, Y의 분산 4만 보고 X의 분산 1을 빠뜨리는 실수가 자주 나온다.

다시 나오면 이렇게 판단

두 확률변수의 합의 분산 문제는 먼저 독립 여부를 확인하고, 각 변수의 평균과 분산을 따로 구한 뒤 마지막에 더하는 순서로 판단한다.

암기 포인트

독립이면 $Var(X + Y) = Var(X) + Var(Y)$이고, 분산은 편차제곱의 평균이다.

NoteBase

문 3. 정규분포를 따르는 확률변수 X, Y, W의 평균과 분산은 다음과 같다. 다음 확률값을 작은 것부터 순서대로 바르게 나열한 것은?

확률변수	평균	분산
X	2	4
Y	0	9
W	8	16

ㄱ. $P(X \geq 1)$

ㄴ. $P(Y \leq -2)$

ㄷ. $P(W \leq 12)$

① ㄴ, ㄱ, ㄷ

② ㄴ, ㄷ, ㄱ

③ ㄷ, ㄱ, ㄴ

④ ㄷ, ㄴ, ㄱ

정답 1번

출제 포인트

이 문제는 정규분포를 표준정규분포로 변환해 각 확률을 계산한 뒤 그 크기를 비교하는 문제이다.

조건 정리

주어진 값은 $X \sim N(2, 4), Y \sim N(0, 9), W \sim N(8, 16)$이다.
분산이 각각 4, 9, 16이므로 표준편차는 순서대로 2, 3, 4이다.
구해야 할 값은 $P(X \geq 1), P(Y \leq -2), P(W \leq 12)$의 대소관계이다.

사용할 공식

정규분포 $N(\mu, \sigma^2)$는 $Z = \frac{X-\mu}{\sigma}$로 표준화한다.
또한 $P(Z \geq -a) = P(Z \leq a)$와 $P(Z \leq -a) = 1 - P(Z \leq a)$를 이용한다.
이 공식을 쓰는 이유는 평균과 분산이 서로 다른 세 정규분포를 같은 표준정규분포 기준으로 비교해야 하기 때문이다.

NoteBase

계산 과정

ㄱ. $P(X \geq 1) = P\left(Z \geq \frac{1-2}{2}\right) = P(Z \geq -0.5) = P(Z \leq 0.5) \approx 0.6915$이다.
ㄴ. $P(Y \leq -2) = P\left(Z \leq \frac{-2-0}{3}\right) = P\left(Z \leq -\frac{2}{3}\right) \approx 0.2525$이다.
ㄷ. $P(W \leq 12) = P\left(Z \leq \frac{12-8}{4}\right) = P(Z \leq 1) \approx 0.8413$이다.
따라서 작은 것부터 나열하면 $0.2525 < 0.6915 < 0.8413$이므로 ㄴ, ㄱ, ㄷ 순서이다.

선지별 해설

1번 O: 계산 결과가 ㄴ ≈ 0.2525, ㄱ ≈ 0.6915, ㄷ ≈ 0.8413이므로 작은 것부터 ㄴ, ㄱ, ㄷ 순서가 맞다.
2번 X: "ㄴ, ㄷ, ㄱ"이 틀렸다. 올바르게는 "ㄴ, ㄱ, ㄷ"이다. ㄷ는 약 0.8413이고 ㄱ은 약 0.6915이므로 ㄷ가 ㄱ보다 뒤에 와야 한다.
3번 X: "ㄷ, ㄱ, ㄴ"이 틀렸다. 올바르게는 "ㄴ, ㄱ, ㄷ"이다. 가장 큰 값은 ㄷ이고 가장 작은 값은 ㄴ이므로 순서가 완전히 거꾸로 되어 있다.
4번 X: "ㄷ, ㄴ, ㄱ"이 틀렸다. 올바르게는 "ㄴ, ㄱ, ㄷ"이다. ㄷ를 가장 작은 값으로 둔 것이 오류이며 실제로는 ㄷ가 가장 크다.

오답 포인트

분산을 표준편차로 바꾸지 않고 4, 9, 16을 그대로 분모에 넣으면 표준화 값이 잘못된다. 또 $P(Z \geq -0.5)$를 대칭성으로 $P(Z \leq 0.5)$로 바꾸지 못하면 ㄱ의 확률을 작게 잘못 판단하기 쉽다.

다시 나오면 이렇게 판단

정규분포 비교 문제는 먼저 분산을 표준편차로 고친 뒤 각 값을 Z값으로 표준화하고, 마지막에 표준정규분포표의 누적확률로 바꾸어 크기를 비교하면 된다.

암기 포인트

정규분포 비교는 분산이 아니라 표준편차로 표준화하고, 음수 구간은 대칭성으로 양수 누적확률로 바꾸면 계산이 빨라진다.

NoteBase

문 4. 사건 A, B, C는 서로 독립이고 $P(A) = 0.1$, $P(B) = 0.2$, $P(C) = 0.3$일 때, $P(A \cup B \cup C)$의 값은?

① $\frac{61}{125}$
② $\frac{62}{125}$
③ $\frac{63}{125}$
④ $\frac{64}{125}$

정답 2번

출제 포인트

이 문제는 서로 독립인 세 사건의 합집합 확률을 여사건 또는 포함배제 원리로 구할 수 있는지를 묻는 문제이다.

조건 정리

주어진 값: $P(A) = 0.1, P(B) = 0.2, P(C) = 0.3$이다.
구해야 할 값: $P(A \cup B \cup C)$이다.
주의할 조건: 세 사건이 서로 독립이므로 $P(A^c \cap B^c \cap C^c) = P(A^c)P(B^c)P(C^c)$가 성립한다.

사용할 공식

$P(A \cup B \cup C) = 1 - P(A^c \cap B^c \cap C^c)$
또는
$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$
이 공식을 쓰는 이유:
문제가 세 사건 중 적어도 하나가 일어날 확률을 묻고 있으므로, 아무 일도 일어나지 않을 확률을 먼저 구해 1에서 빼면 된다.

계산 과정

1. $P(A^c) = 0.9, P(B^c) = 0.8, P(C^c) = 0.7$이다.
2. 독립이므로 $P(A^c \cap B^c \cap C^c) = 0.9 \times 0.8 \times 0.7 = 0.504 = \frac{63}{125}$이다.
3. 따라서 $P(A \cup B \cup C) = 1 - 0.504 = 0.496 = \frac{62}{125}$이다.

NoteBase

선지별 해설

① X
$\frac{61}{125} = 0.488$은 계산 결과와 다르다.
틀린 표현은 값 자체이다. 올바르게는 $\frac{62}{125}$이다.

② O
여사건을 이용하면 $1 - 0.9 \times 0.8 \times 0.7 = \frac{62}{125}$이므로 맞다.

③ X
$\frac{63}{125}$는 $P(A^c \cap B^c \cap C^c)$, 즉 세 사건이 모두 일어나지 않을 확률이다.
틀린 표현은 구하려는 확률과 여사건의 확률을 바꾼 점이다. 올바르게는 $P(A \cup B \cup C) = 1 - \frac{63}{125} = \frac{62}{125}$이다.

④ X
$\frac{64}{125} = 0.512$는 계산 결과와 다르다.
틀린 표현은 최종 값 자체이다. 올바르게는 $\frac{62}{125}$이다.

오답 포인트

합집합 확률을 구하면서 $P(A^c \cap B^c \cap C^c)$를 바로 답으로 적는 실수가 가장 흔하다. 독립이라는 조건이 보이면 여사건의 곱을 구한 뒤 마지막에 1에서 빼는 단계까지 확인해야 한다.

다시 나오면 이렇게 판단

세 사건의 합집합 확률은 '적어도 하나 발생'으로 읽고, 독립이면 먼저 '모두 일어나지 않음'의 확률을 곱으로 구한 뒤 1에서 빼는 순서로 판단한다.

암기 포인트

독립인 사건의 '적어도 하나'는 1 − '모두 안 일어남'이다.

NoteBase

문 5. 다음은 벼의 품종(A)과 비료(B)가 쌀 생산량에 영향을 미치는지를 알아보기 위하여 벼의 품종은 a개의 수준, 비료는 b개의 수준에서 반복이 없는 이원배치법을 적용하여 얻은 분산분석표의 일부이다. a와 b의 값을 바르게 연결한 것은?

요인	제곱합	자유도	평균제곱	F-값
인자 A	20		4	2
인자 B				3
오차				
전체	180			

① a: 5, b: 10

② a: 5, b: 11

③ a: 6, b: 10

④ a: 6, b: 11

정답 4번

출제 포인트

반복이 없는 이원배치법에서 자유도, 평균제곱, F-값의 관계를 이용해 요인 수준 수 a, b를 복원하는 문제이다.

조건 정리

주어진 값: 인자 A의 제곱합은 20, 평균제곱은 4, $F_A = 2$, $F_B = 3$, 전체 제곱합은 180이다.
구해야 할 값: 품종 수준 수 a와 비료 수준 수 b이다.
주의할 조건: 반복이 없는 이원배치법이므로 자유도는 인자 A가 $a-1$, 인자 B가 $b-1$, 오차가 $(a-1)(b-1)$, 전체가 $ab-1$이다.

사용할 공식

$MS_A = \frac{SS_A}{a-1}, \quad F_A = \frac{MS_A}{MS_E}, \quad F_B = \frac{MS_B}{MS_E}$
$SS_T = SS_A + SS_B + SS_E$
반복이 없는 이원배치법에서는 위 관계만으로 자유도와 제곱합을 연결할 수 있으므로 a와 b를 역산할 수 있다.

NoteBase

계산 과정

1. 인자 A에서 $MS_A = SS_A/(a-1)$ 이므로 $4 = 20/(a-1)$, 따라서 $a-1=5$, $a=6$이다.
2. $F_A = MS_A/MS_E = 2$ 이므로 $MS_E = 4/2 = 2$이다.
3. $F_B = MS_B/MS_E = 3$ 이므로 $MS_B = 3\times 2 = 6$이다. 따라서 $SS_B = MS_B(b-1) = 6(b-1)$이다.
4. 오차 자유도는 $(a-1)(b-1) = 5(b-1)$ 이므로 $SS_E = MS_E \times df_E = 2\times 5(b-1) = 10(b-1)$이다.
5. 전체 제곱합 180을 쓰면 $180 = 20 + 6(b-1) + 10(b-1) = 20 + 16(b-1)$이다. 따라서 $16(b-1) = 160$, $b-1=10$, $b=11$이다.

선지별 해설

① X: $a=5$, $b=10$은 둘 다 틀리다. 인자 A의 자유도는 $20/4 = 5$이므로 수준 수는 자유도보다 1 큰 $a=6$이다. 또 $b=10$이면 $b-1=9$가 되어 전체 제곱합이 $20 + 6\times 9 + 10\times 9 = 164$로 맞지 않는다. 틀린 표현은 $a=5$, $b=10$이고, 올바르게는 $a=6$, $b=11$이다.
② X: $a=5$가 이미 틀리다. a는 자유도 5 자체가 아니라 그보다 1 큰 수준 수이므로 6이다. 또한 $b=11$만 맞고 a가 틀리므로 정답이 될 수 없다. 틀린 표현은 $a=5$이고, 올바르게는 $a=6$이다.
③ X: $a=6$은 맞지만 $b=10$이 틀리다. b를 구할 때도 자유도와 수준 수를 구분해야 하며, 전체 제곱합 식을 맞추면 $b-1=10$이므로 $b=11$이다. 틀린 표현은 $b=10$이고, 올바르게는 $b=11$이다.
④ O: $a=6$이고 $b=11$이면 $df_A=5$, $MS_E=2$, $MS_B=6$, $SS_B=60$, $SS_E=100$이 되어 전체 제곱합이 $20+60+100=180$으로 정확히 맞는다. 따라서 정답 선택지는 ④이다.

오답 포인트

가장 흔한 실수는 자유도와 수준 수를 같은 값으로 보는 오류이다. 이 문항에서는 $a-1=5$를 보고 곧바로 $a=5$로 적거나, $b-1=10$을 보고 $b=10$으로 적으면 오답이 된다.

다시 나오면 이렇게 판단

이원분산분석표 복원 문제는 먼저 $MS = SS/df$로 자유도를 찾고, 다음에 $F = MS/MS_E$로 오차평균제곱을 구한 뒤, 마지막에 전체 제곱합 $SS_T = SS_A + SS_B + SS_E$를 맞춰 남은 수준 수를 결정하면 된다.

암기 포인트

이원배치법에서는 수준 수 = 자유도 + 1이고, 반복이 없으면 오차 자유도는 $(a-1)(b-1)$이다.

NoteBase

문 6. 통계부서에 근무하는 직원 9명의 지난 한 해 동안 출장 일수가 다음과 같이 조사되었다. 이 관측값을 $x_1, x_2, \cdots, x_9$라고 할 때, $\sum_{i=1}^{9}(x_i - a)^2$을 최소로 하는 a의 값과 $\sum_{i=1}^{9}|x_i - b|$를 최소로 하는 b의 값을 바르게 연결한 것은?

10, 20, 30, 40, 60, 70, 70, 70, 80

① a: 50, b: 60

② a: 50, b: 70

③ a: 60, b: 60

④ a: 60, b: 70

정답 1번

출제 포인트

이 문제는 제곱편차합을 최소로 하는 값은 평균, 절대편차합을 최소로 하는 값은 중앙값이라는 성질을 묻는 문제이다.

조건 정리

주어진 자료는 10, 20, 30, 40, 60, 70, 70, 70, 80이다.
구해야 할 값은 $\sum_{i=1}^{9}(x_i - a)^2$를 최소로 하는 a와 $\sum_{i=1}^{9}|x_i - b|$를 최소로 하는 b이다.
자료가 이미 크기순으로 정렬되어 있으므로 평균과 중앙값을 바로 찾을 수 있다.

사용할 공식

$a = \bar{x} = \frac{1}{9}\sum_{i=1}^{9} x_i$
$b = \text{median}$
제곱편차합은 평균에서 최소가 되고, 절대편차합은 중앙값에서 최소가 되므로 이 공식을 사용한다.

계산 과정

1. 평균을 구하면 $\bar{x} = \frac{10+20+30+40+60+70+70+70+80}{9} = \frac{450}{9} = 50$이다.
2. 자료가 9개이므로 중앙값은 가운데 있는 5번째 값이다.
3. 5번째 값은 60이므로 $b = 60$이다.
4. 따라서 $a = 50, b = 60$이다.

NoteBase

선지별 해설

① O

$a = 50$, $b = 60$이므로 맞다.

② X

$a = 50$은 맞지만 $b = 70$은 틀리다. 절대편차합을 최소로 하는 값은 최빈값이 아니라 중앙값이므로, b는 70이 아니라 60이다.

③ X

$b = 60$은 맞지만 $a = 60$은 틀리다. 제곱편차합을 최소로 하는 값은 중앙값이 아니라 평균이므로, a는 60이 아니라 50이다.

④ X

$a = 60$, $b = 70$이 모두 틀리다. a는 평균 50으로 고쳐야 하고, b는 중앙값 60으로 고쳐야 한다.

오답 포인트

70이 세 번 나와 최빈값이 눈에 먼저 들어오므로 $\sum|x_i - b|$의 최소값을 최빈값으로 착각하기 쉽다. 또 제곱편차합 최소 조건과 절대편차합 최소 조건을 같은 대표값으로 묶어 외우면 a와 b를 모두 60으로 고르는 실수가 나온다.

다시 나오면 이렇게 판단

대표값을 고르는 최적화 문제가 나오면 먼저 식의 모양을 본다. 제곱이 붙으면 평균, 절댓값이 붙으면 중앙값으로 연결하면 빠르게 판단할 수 있다.

암기 포인트

제곱편차합 최소는 평균, 절대편차합 최소는 중앙값이다.

NoteBase

문 7. 다음 자료에 단순선형회귀모형 $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ $(i = 1, 2, 3, 4, 5)$를 최소제곱법으로 적합하여 얻은 β_0와 β_1의 추정값 a와 b의 값을 바르게 연결한 것은? (단, ε_i는 $N(0, \sigma^2)$을 따르고 서로 독립이다)

X	1	2	3	4	5
Y	2	3	5	6	9

① a: $\frac{1}{10}$, b: $-\frac{17}{10}$

② a: $\frac{1}{10}$, b: $\frac{17}{10}$

③ a: $-\frac{1}{10}$, b: $-\frac{17}{10}$

④ a: $-\frac{1}{10}$, b: $\frac{17}{10}$

정답 4번

출제 포인트

이 문제는 최소제곱법으로 단순선형회귀식의 기울기 추정값 $b = \hat{\beta}_1$와 절편 추정값 $a = \hat{\beta}_0$를 구하는 문제이다.

조건 정리

주어진 값: $(x_i, y_i) = (1, 2), (2, 3), (3, 5), (4, 6), (5, 9)$이다.
구해야 할 값: 회귀식 $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$의 최소제곱추정량 $a = \hat{\beta}_0, b = \hat{\beta}_1$이다.
주의할 조건: 보기의 핵심은 절편의 부호와 기울기의 부호를 함께 정확히 연결하는 데 있다.

사용할 공식

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \qquad a = \bar{y} - b\bar{x}$$

이 공식을 쓰는 이유: 단순선형회귀모형을 최소제곱법으로 적합할 때 기울기는 공분산형 합 S_{xy}를 분산형 합 S_{xx}로 나눈 값이고, 절편은 표본평균점 $(\bar{x}, \bar{y})$을 회귀직선이 지난다는 성질로 바로 구해진다.

계산 과정

1. 평균은 $\bar{x} = (1 + 2 + 3 + 4 + 5)/5 = 3$, $\bar{y} = (2 + 3 + 5 + 6 + 9)/5 = 5$이다.
2. $S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = (-2)(-3) + (-1)(-2) + 0 \cdot 0 + 1 \cdot 1 + 2 \cdot 4 = 17$이다.
3. $S_{xx} = \sum (x_i - \bar{x})^2 = (-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2 = 10$이다.
4. 따라서 $b = S_{xy}/S_{xx} = 17/10$이다.
5. 절편은 $a = \bar{y} - b\bar{x} = 5 - (17/10) \cdot 3 = 50/10 - 51/10 = -1/10$이다.

NoteBase

선지별 해설

① X: 틀린 표현은 $a = 1/10$과 $b = -17/10$이다. 계산 결과 기울기는 $b = 17/10$로 양수이고, 절편은 $a = -1/10$이다. 부호가 둘 다 뒤바뀐 보기이다.
② X: 틀린 표현은 $a = 1/10$이다. $b = 17/10$은 맞지만 절편은 $a = 5 - (17/10) \cdot 3 = -1/10$이므로 올바르게는 $a = -1/10$이어야 한다.
③ X: 틀린 표현은 $b = -17/10$이다. 절편 $a = -1/10$은 맞지만, x가 증가할수록 y가 커지므로 기울기는 음수가 아니라 $17/10$이다.
④ O: $b = 17/10$, $a = -1/10$으로 최소제곱추정값과 정확히 일치한다.

오답 포인트

기울기 b를 먼저 구한 뒤 절편 $a = \bar{y} - b\bar{x}$를 계산하지 않고 감으로 부호를 정하면 ②나 ③으로 흔들리기 쉽다. 특히 자료가 전반적으로 증가하므로 기울기는 양수라는 점을 놓치면 ①이나 ③의 음의 기울기에 속기 쉽다.

다시 나오면 이렇게 판단

회귀계수 계산 문제는 먼저 $\bar{x}, \bar{y}$를 구하고, 다음으로 S_{xy}와 S_{xx}의 부호와 값을 계산해 b를 정한 뒤, 마지막에 $a = \bar{y} - b\bar{x}$로 절편까지 확인하는 순서로 판단한다.

암기 포인트

단순회귀 최소제곱추정량은 $b = S_{xy}/S_{xx}$, $a = \bar{y} - b\bar{x}$이다.

NoteBase

문 8. 검은 공이 6개, 흰 공이 2개 들어 있는 상자에서 갑과 을이 번갈아 가며 임의로 하나씩 공을 꺼낼 때 흰 공을 먼저 꺼낸 사람이 이기는 게임을 한다. 이때 꺼낸 공을 다시 집어넣고 게임을 하는 방법(게임 A)과 꺼낸 공을 다시 집어넣지 않고 게임을 하는 방법(게임 B)을 고려할 수 있다. 첫 번째 공을 갑이 꺼낸다고 했을 때 게임 A에서 갑이 이길 확률(P_A)과 게임 B에서 갑이 이길 확률(P_B)의 차 $P_A - P_B$의 값은?

① $-\frac{1}{7}$

② 0

③ $\frac{1}{7}$

④ $\frac{2}{7}$

정답 2번

출제 포인트

복원추출과 비복원추출에서 모두 '첫 흰 공이 홀수 번째에 나오면 갑이 이긴다'는 사건의 확률을 비교하는 문제이다.

조건 정리

주어진 값: 검은 공 6개, 흰 공 2개이다.
구해야 할 값: 첫 번째 공을 갑이 꺼낼 때 게임 A와 게임 B에서 갑이 이길 확률의 차 $P_A - P_B$이다.
주의할 조건: 갑은 1, 3, 5, 7번째처럼 홀수 번째 차례에 공을 꺼내므로 첫 흰 공이 홀수 번째에 나오면 갑이 이긴다.

사용할 공식

게임 A는 처음 흰 공이 나올 때까지의 시행을 기하급수적으로 합한다.
P_A=14+≤ft(34)^214+≤ft(34)^414+·s
게임 B는 첫 흰 공의 위치가 1, 3, 5, 7번째인 경우를 더한다.
이 공식을 쓰는 이유: 두 게임 모두 승패는 첫 흰 공의 등장 위치로 결정되기 때문이다.

NoteBase

계산 과정

1. 게임 A에서는 한 번 뽑을 때마다 흰 공이 나올 확률이 항상 14이다. 따라서 갑이 이길 확률은
P_A=14≤ft(1+≤ft(34)^2+≤ft(34)^4+·s)=(14)/(1-≤ft(34)^2)=47
이다.
2. 게임 B에서는 첫 흰 공이 1, 3, 5, 7번째에 나오면 갑이 이긴다.
P_B=28+68·57·26+68·57·46·35·24+68·57·46·35·24·13
3. 이를 계산하면
P_B=14+(5)/(28)+(3)/(28)+(1)/(28)=47
이므로
P_A-P_B=47-47=0
이다.

선지별 해설

① X
P_A-P_B=-17이라는 표현이 틀렸다. 실제로는 두 게임 모두 갑의 승률이 47이므로 차는 음수가 아니라 0이다. 이 값은 게임 B의 확률을 과소계산했을 때 나올 수 있다.

② O
$P_A - P_B = 0$이 맞다. 게임 A와 게임 B에서 모두 갑이 이길 확률이 47이므로 차는 0이다.

③ X
P_A-P_B=17이라는 표현이 틀렸다. 올바르게는 $P_A = P_B$이므로 차는 17이 아니라 0이다. 게임 A만 계산하고 게임 B의 홀수 번째 경우를 하나 빠뜨리면 이런 오답이 나오기 쉽다.

④ X
P_A-P_B=27이라는 표현이 틀렸다. 올바르게는 두 확률이 같아서 차는 27이 아니라 0이다. 복원추출과 비복원추출이라는 말에 끌려 두 확률이 크게 다를 것이라고 생각하면 생기는 오답이다.

오답 포인트

게임 방식이 다르다는 이유만으로 P_A와 P_B가 다를 것이라고 성급하게 판단하기 쉽다. 이 문항의 핵심은 복원 여부가 아니라 첫 흰 공이 홀수 번째에 나오면 갑이 이긴다는 사건을 정확히 세는 데 있다.

다시 나오면 이렇게 판단

번갈아 뽑는 승부 문제는 먼저 누가 이기려면 첫 성공이 몇 번째에 나와야 하는가를 홀수·짝수 위치로 바꿔 본다. 그다음 각 게임에서 그 위치의 확률만 계산해 비교하면 된다.

NoteBase

암기 포인트

번갈아 뽑기 승부는 첫 성공의 위치가 홀수면 선공 승, 짝수면 후공 승으로 바꾸면 정리가 쉽다.

NoteBase

문 9. $Z_1, Z_2, \cdots, Z_6$**이 표준정규분포로부터의 임의표본일 때, 다음 식을 만족하는 상수** c**의 값은? (단,** $F_\alpha(k_1, k_2)$**는 분자와 분모의 자유도가 각각** k_1**과** k_2**인** F**분포의 제**$100 \times (1-\alpha)$ **백분위수를 나타낸다)**

$$P\left(\frac{Z_1^2+Z_2^2}{Z_3^2+Z_4^2+Z_5^2+Z_6^2} < c\right) = 0.05$$

① $\frac{1}{2} \times \frac{1}{F_{0.05}(4,2)}$
② $\frac{1}{2} \times F_{0.05}(2,4)$
③ $2 \times \frac{1}{F_{0.05}(4,2)}$
④ $2 \times F_{0.05}(2,4)$

정답 1번

출제 포인트

이 문제는 독립인 카이제곱변수의 비를 F분포로 바꾸고, 주어진 확률 0.05를 하측 5% 분위수로 해석하는 문제이다.

조건 정리

주어진 값:
$Z_1^2+Z_2^2 \sim \chi^2(2), Z_3^2+Z_4^2+Z_5^2+Z_6^2 \sim \chi^2(4)$이다.
구해야 할 값:
$P\left(\frac{Z_1^2+Z_2^2}{Z_3^2+Z_4^2+Z_5^2+Z_6^2} < c\right) = 0.05$를 만족하는 c이다.
주의할 조건:
문제의 $F_\alpha(k_1, k_2)$는 $F(k_1, k_2)$분포의 상위 α점, 즉 제 $100 \times (1-\alpha)$ 백분위수이다.

사용할 공식

공식:
$\frac{U/\nu_1}{V/\nu_2} \sim F(\nu_1, \nu_2) \quad \text{if } U \sim \chi^2(\nu_1),\ V \sim \chi^2(\nu_2)$
이 공식을 쓰는 이유:
분자 $U = Z_1^2+Z_2^2$는 자유도 2의 카이제곱분포이고 분모 $V = Z_3^2+Z_4^2+Z_5^2+Z_6^2$는 자유도 4의 카이제곱분포이며 서로 독립이므로, 비율을 $F(2, 4)$분포로 바로 바꿀 수 있다.

NoteBase

계산 과정

1. $U = Z_1^2 + Z_2^2, V = Z_3^2 + Z_4^2 + Z_5^2 + Z_6^2$라고 두면 $U \sim \chi^2(2), V \sim \chi^2(4)$이다.
2. 따라서
$\frac{U/2}{V/4} \sim F(2, 4)$
이고
$\frac{U/2}{V/4} = \frac{4U}{2V} = 2\frac{U}{V}$
이므로
$2\frac{U}{V} \sim F(2, 4), \quad \frac{U}{V} = \frac{1}{2}F$
로 볼 수 있다.
3. 문제 조건은
$P\left(\frac{U}{V} < c\right) = 0.05$
이므로
$P(F < 2c) = 0.05$
이다.
4. $F_{0.05}(2, 4)$는 상위 5%점이므로 하위 5%점은 그대로 쓸 수 없고, 역수 성질 $1/F \sim F(4, 2)$를 이용하면
$P\left(F < \frac{1}{F_{0.05}(4, 2)}\right) = 0.05$
이다.
5. 따라서
$2c = \frac{1}{F_{0.05}(4, 2)}$
이므로
$c = \frac{1}{2} \times \frac{1}{F_{0.05}(4, 2)}$
이다.

NoteBase

선지별 해설

① O

- 위 계산에서 $c = \frac{1}{2} \times \frac{1}{F_{0.05}(4, 2)}$가 직접 나온다. 분자 자유도 2, 분모 자유도 4를 가진 F분포를 만든 뒤 하측 5%점을 역수 성질로 바르게 바꾼 식이다.

② X

- $\frac{1}{2} \times F_{0.05}(2, 4)$는 $F(2, 4)$의 상위 5%점을 그대로 사용한 값이므로 $P(F < 2c) = 0.05$가 아니라 큰 확률에 대응한다.
- 틀린 표현은 $F_{0.05}(2, 4)$를 하측 5%점처럼 쓴 부분이다. 올바르게는 $2c = \frac{1}{F_{0.05}(4, 2)}$이어야 한다.

③ X

- $2 \times \frac{1}{F_{0.05}(4, 2)}$는 역수 성질은 맞게 썼지만, $\frac{U}{V} = \frac{1}{2}F$를 $\frac{U}{V} = 2F$처럼 뒤집은 값이다.
- 틀린 표현은 앞의 계수 2이다. 올바르게는 $\frac{1}{2} \times \frac{1}{F_{0.05}(4, 2)}$이다.

④ X

- $2 \times F_{0.05}(2, 4)$는 상위 5%점과 배율을 모두 반대로 쓴 값이다. 하측 5%점으로 바꾸는 역수 처리도 빠졌고, 앞의 계수도 2가 아니라 $\frac{1}{2}$이어야 한다.
- 틀린 표현은 $2 \times F_{0.05}(2, 4)$ 전체이다. 올바르게는 $\frac{1}{2} \times \frac{1}{F_{0.05}(4, 2)}$이다.

오답 포인트

가장 흔한 실수는 두 가지이다. 첫째, $F_{0.05}(2, 4)$를 하측 5%점으로 오해하는 실수이다. 이 문제의 표기는 상위 5%점이다. 둘째, $\frac{U/2}{V/4} = 2\frac{U}{V}$인데 이를 거꾸로 보고 앞의 계수를 2로 쓰는 실수이다.

다시 나오면 이렇게 판단

카이제곱합의 비가 나오면 먼저 $\frac{U/\nu_1}{V/\nu_2}$ 꼴로 F분포를 만든다. 그다음 문제에서 준 F_α가 상측점인지 하측점인지 확인하고, 하측 확률이면 $1/F \sim F(\nu_2, \nu_1)$를 써서 역수로 바꾼다.

암기 포인트

$\frac{U}{V}$는 바로 F가 아니고 자유도 보정이 붙으며, 하측점은 역수 성질 $1/F \sim F(\nu_2, \nu_1)$로 처리한다.

NoteBase

문 10. 연속형 확률변수 X와 Y의 결합확률밀도함수가 다음과 같을 때 $P\left(Y<\frac{1}{2} \mid X>\frac{1}{2}\right)$의 값은?

f_X,Y(x,y)=cases12xy(1-x), & 0<x<1, 0<y<1 \ 0, & 그 외cases

① $\frac{1}{8}$
② $\frac{1}{4}$
③ $\frac{1}{2}$
④ 1

정답 2번

출제 포인트

주어진 결합확률밀도함수에서 X와 Y의 독립성을 확인하고 조건부확률 $P(Y<1/2 \mid X>1/2)$를 계산하는 문제이다.

조건 정리

주어진 밀도는 $f_{X,Y}(x,y)=12xy(1-x)$이고 범위는 $0<x<1, 0<y<1$이다.
구해야 할 값은 $P(Y<1/2 \mid X>1/2)$이다.
$f_{X,Y}(x,y)=\{6x(1-x)\}\{2y\}$로 분리되므로 X와 Y는 독립이다.

사용할 공식

$P(Y<1/2 \mid X>1/2)=\frac{P(Y<1/2, X>1/2)}{P(X>1/2)}$
또는 X와 Y가 독립이면 $P(Y<1/2 \mid X>1/2)=P(Y<1/2)$를 사용한다.
이 문제는 결합밀도함수가 x에 대한 함수와 y에 대한 함수의 곱으로 분리되므로 독립성 공식을 바로 적용할 수 있다.

계산 과정

1. 주변밀도함수는 $f_X(x)=\int_0^1 12xy(1-x)dy=6x(1-x)$, $f_Y(y)=\int_0^1 12xy(1-x)dx=2y$이다.
2. 따라서 $f_{X,Y}(x,y)=f_X(x)f_Y(y)$이므로 X와 Y는 독립이다.
3. 그러면 $P(Y<1/2 \mid X>1/2)=P(Y<1/2)=\int_0^{1/2} 2ydy=1/4$이다.
직접 계산해도 분자는 1/8, 분모는 1/2이므로 비는 1/4이다.

NoteBase

선지 판단

1번 X: 1/8은 $P(Y < 1/2, X > 1/2)$인 분자만 계산한 값이다. 조건부확률은 이를 $P(X > 1/2) = 1/2$로 나누어야 하므로 올바르게는 1/4이다.
2번 O: X와 Y가 독립이므로 $P(Y < 1/2 \mid X > 1/2) = P(Y < 1/2) = 1/4$가 되어 맞다.
3번 X: 1/2는 조건부확률 값이 아니라 $P(X > 1/2)$인 분모에 해당한다. 올바르게는 1/4이다.
4번 X: 1은 $Y < 1/2$가 $X > 1/2$일 때 반드시 일어난다는 뜻인데, 실제 조건부확률은 1/4이므로 틀리다. 올바르게는 1/4이다.

오답 포인트

분자 1/8과 분모 1/2를 각각 구해 놓고 마지막 나눗셈을 빠뜨리면 1번이나 3번으로 흔들린다. 결합밀도함수가 곱으로 분리되면 독립을 먼저 확인하는 습관이 필요하다.

다시 나오면 이렇게 판단

결합밀도 문제는 먼저 $f_{X,Y}(x, y)$가 $g(x)h(y)$로 분리되는지 본다. 분리되면 독립성을 이용해 조건부확률을 주변확률로 바꾸고, 분리되지 않으면 정의식 $\frac{P(A \cap B)}{P(B)}$로 계산한다.

암기 포인트

결합밀도함수가 $g(x)h(y)$로 분리되면 독립이고, 독립이면 $P(A \mid B) = P(A)$이다.

NoteBase

문 11. 혈압약 A와 B의 효과를 비교하기 위해 고혈압 환자에서 임의추출한 100명 중 임의로 50명을 선택하여 혈압약 A를 투여하고, 나머지 50명에게는 혈압약 B를 투여한 결과는 다음과 같다. "두 혈압약의 효과는 차이가 없다."는 가정하에서 혈압약 A를 투여한 후 효과가 있는 고혈압 환자수의 기대도수는?

혈압약 \ 효과	있음	없음
A	37	13
B	27	23

① 30

② 32

③ 34

④ 36

정답 2번

출제 포인트

독립성 가정 아래 2 × 2 분할표의 기대도수를 행합계, 열합계, 전체합계로 계산하는 문제이다.

조건 정리

A를 투여한 환자 수는 50명이다.
효과가 있는 전체 환자 수는 37 + 27 = 64명이다.
전체 환자 수는 100명이다.
구해야 할 값은 귀무가설 "두 혈압약의 효과는 차이가 없다." 아래에서 A를 투여한 뒤 효과가 있는 칸의 기대도수이다.

사용할 공식

기대도수는 행합계×열합계전체합계이다.
두 혈압약의 효과 차이가 없다는 가정은 혈압약 종류와 효과 유무가 독립이라는 뜻이므로 각 칸의 기대도수는 해당 행합계와 열합계의 곱을 전체합계로 나누어 구한다.

NoteBase

계산 과정

A 행의 합계는 50이다.
효과 있음 열의 합계는 64이다.
따라서 A를 투여한 뒤 효과가 있는 환자 수의 기대도수는 $\frac{50 \times 64}{100} = 32$이다.

선지별 해설

① X: 기대도수는 30이 아니라 32이다. 30이라는 값은 A 행합계 50, 효과 있음 열합계 64, 전체합계 100을 이용한 기대도수 계산과 맞지 않는다. 올바르게는 기대도수는 32이다.
② O: 기대도수는 $\frac{50 \times 64}{100} = 32$이므로 옳다.
③ X: 기대도수는 34가 아니라 32이다. 34라는 값은 기대도수 공식에 표의 합계를 정확히 대입한 결과가 아니다. 올바르게는 기대도수는 32이다.
④ X: 기대도수는 36이 아니라 32이다. 36은 관측도수 37과 비슷한 값에 끌려 판단한 오답일 수 있지만, 기대도수는 관측도수가 아니라 행합계와 열합계로 계산한다. 올바르게는 기대도수는 32이다.

오답 포인트

기대도수와 관측도수를 혼동하기 쉽다. 이 문항은 표에 적힌 37을 그대로 쓰거나 대충 중간값을 고르는 함정을 두었지만, 기대도수는 반드시 행합계와 열합계의 곱을 전체합계로 나누어 계산해야 한다.

다시 나오면 이렇게 판단

독립성 또는 효과 차이 없음 가정이 보이면 먼저 해당 칸의 행합계와 열합계를 찾고, 전체합계로 나누어 기대도수를 계산하는 순서로 판단한다.

암기 포인트

분할표 기대도수는 행합계 곱하기 열합계 나누기 전체합계이다.

NoteBase

문 12. $X_1, X_2, \cdots, X_n$은 구간이 $(2, 4)$인 균일분포(uniform distribution)로부터의 임의표본이다. n이 무한히 커질 때 통계량 $\theta_n = \frac{n}{\sum_{i=1}^{n} X_i}$이 확률적으로 수렴하는 값은?

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

정답 3번

출제 포인트

이 문제는 큰수의 법칙으로 표본평균의 확률수렴값을 구한 뒤, 그 역수로 통계량의 확률수렴값을 판단하는 문제이다.

조건 정리

주어진 값: $X_1, X_2, \ldots, X_n$은 구간 $(2, 4)$의 균일분포에서 나온 임의표본이다.
구해야 할 값: $\theta_n = \frac{n}{\sum_{i=1}^{n} X_i}$의 확률수렴값이다.
주의할 조건: $\frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}_n}$ 이므로 먼저 표본평균 $\bar{X}_n$의 극한을 본다.

사용할 공식

큰수의 법칙에 따라 $\bar{X}_n \to E[X]$ 이다.
균일분포 $U(2, 4)$의 평균은 $E[X] = \frac{2+4}{2} = 3$ 이다.
따라서 연속함수 $g(x) = 1/x$를 적용하면 $\hat{\theta}_n = 1/\bar{X}_n \to 1/3$ 이다.
이 공식을 쓰는 이유는 문제의 통계량이 표본평균의 역수 형태이기 때문이다.

계산 과정

1. 표본평균을 $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ 로 둔다.
2. $X_i \sim U(2, 4)$ 이므로 $E[X] = 3$ 이고, 큰수의 법칙에 따라 $\bar{X}_n \to 3$ 이다.
3. $\theta_n = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}_n}$ 이므로 $\hat{\theta}_n \to \frac{1}{3}$ 이다.

NoteBase

선지별 해설

① X: 1이 되려면 $\bar{X}_n$이 1로 수렴해야 하는데, 실제로는 $\bar{X}_n \to 3$ 이다. 틀린 표현은 수렴값을 1로 본 부분이다. 올바르게는 $\hat{\theta}_n$은 1/3으로 수렴한다.
② X: 1/2는 구간의 왼쪽 끝점 2만 보고 역수를 취한 값이다. 그러나 확률수렴은 끝점이 아니라 표본평균의 극한인 $E[X] = 3$으로 판단한다. 틀린 표현은 수렴값을 1/2로 본 부분이다. 올바르게는 1/3이다.
③ O: $U(2, 4)$의 평균이 3이므로 $\bar{X}_n \to 3$ 이고, 따라서 $1/\bar{X}_n \to 1/3$ 이다. 선택지 ③이 맞다.
④ X: 1/4는 구간의 오른쪽 끝점 4를 그대로 역수 처리한 값이다. 문제의 통계량은 극단값이 아니라 표본평균의 역수이므로 이렇게 판단하면 틀린다. 틀린 표현은 수렴값을 1/4로 본 부분이다. 올바르게는 1/3이다.

오답 포인트

구간 $(2, 4)$의 양 끝값 2와 4를 바로 가져와 1/2나 1/4를 고르기 쉽다. 이 문항의 함정은 $\frac{n}{\sum X_i}$를 낯선 형태로 제시해도 실제로는 $1/\bar{X}_n$이라는 점을 놓치게 만드는 데 있다.

다시 나오면 이렇게 판단

표본합과 n이 함께 보이면 먼저 표본평균 형태로 바꾼다. 그다음 큰수의 법칙으로 표본평균의 극한을 구하고, 마지막에 필요한 함수형태를 적용해 최종 수렴값을 정한다.

암기 포인트

$\frac{n}{\sum X_i} = \frac{1}{\bar{X}_n}$, 균일분포 $U(a, b)$의 평균은 $(a + b)/2$이다.

NoteBase

문 13. 나무 30그루에서 측정한 나무둘레(Y), 나무높이(x_1), 나무부피(x_2)의 자료에 다중선형회귀모형 $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$를 적합한 결과의 일부는 다음과 같다. 이에 대한 설명으로 옳은 것만을 모두 고르면? (단, ε_i는 $N(0, \sigma^2)$을 따르고 서로 독립이다)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.53105	1.87281	5.623	5.74e-06
x_1	−0.04515	0.02675	−1.688	0.103
x_2	0.20578	0.01157	17.781	< 2e-16

Residual standard error: 0.7481 on ( ) degrees of freedom

Multiple R-squared: 0.9369, Adjusted R-squared: 0.9323

F-statistic: 200.5 on ( ) and ( ) DF, p-value: < 2.2e-16

ㄱ. 유의수준 5 %에서 나무부피(x_2)가 주어졌을 때, 나무둘레(Y)에 대해 나무높이(x_1)는 유의한 설명변수이다.

ㄴ. 유의수준 5 %에서 나무높이(x_1)가 주어졌을 때, 나무부피(x_2)가 증가할수록 나무둘레(Y)는 증가한다.

ㄷ. 잔차제곱합의 자유도는 29이다.

① ㄴ

② ㄱ, ㄷ

③ ㄴ, ㄷ

④ ㄱ, ㄴ, ㄷ

정답 1번

출제 포인트

이 문제는 다중선형회귀 출력표에서 각 회귀계수의 유의성, 계수의 부호, 잔차자유도를 읽어 ㄱ, ㄴ, ㄷ의 참거짓을 판단하는 문제이다.

NoteBase

자료 읽는 순서

1. x_1과 x_2의 p값을 유의수준 5%와 비교한다.
2. 유의한 설명변수의 추정계수 부호를 확인해 Y의 증가 방향을 판단한다.
3. 표본 수 30, 설명변수 2개, 절편 1개이므로 잔차자유도는 $30-3=27$임을 확인한 뒤 ㄱ, ㄴ, ㄷ을 조합한다.

선지별 해설

① O

- ㄱ은 x_1의 p값이 0.103으로 0.05보다 커서 유의하지 않으므로 거짓이다.
- ㄴ은 x_2의 추정계수 0.20578이 양수이고 p값이 $< 2 \times 10^{-16}$이므로, x_1가 주어졌을 때 x_2가 증가할수록 Y는 증가한다고 볼 수 있어 참이다.
- ㄷ은 잔차자유도가 $30-(2+1)=27$이므로 거짓이다.
- 따라서 ㄴ만 옳아 ①이 맞다.

② X

- ②는 ㄱ과 ㄷ이 옳다고 본 선택지인데, ㄱ은 p값 0.103 때문에 유의하지 않고 ㄷ도 잔차자유도를 29로 잘못 쓴 것이므로 틀리다.
- 올바르게는 ㄴ만 옳다.

③ X

- ③은 ㄴ과 ㄷ이 옳다고 본 선택지인데, ㄴ은 참이지만 ㄷ은 잔차자유도가 27이므로 거짓이다.
- 올바르게는 ㄴ만 옳다.

④ X

- ④는 ㄱ, ㄴ, ㄷ이 모두 옳다고 본 선택지인데, ㄱ은 p값 0.103으로 유의하지 않고 ㄷ도 잔차자유도를 29로 잘못 판단했으므로 틀리다.
- 올바르게는 ㄴ만 옳다.

오답 포인트

다중회귀에서는 다른 변수를 통제한 상태의 p값을 봐야 하는데, x_1의 추정치 부호만 보고 유의하다고 착각하기 쉽다. 또 잔차자유도를 총자유도 $30-1=29$와 혼동하면 ㄷ을 참으로 잘못 판단하게 된다.

다시 나오면 이렇게 판단

회귀표 문제는 먼저 각 설명변수의 p값으로 유의성부터 가르고, 그다음 유의한 계수의 부호를 해석한다.
자유도는 총자유도와 잔차자유도를 구분해 잔차자유도 =n-(설명변수 수+1)로 계산하면 된다.

NoteBase

암기 포인트

다중회귀 해석은 p값으로 유의성 확인, 계수 부호로 방향 확인, 잔차자유도는 $n - p - 1$ 확인의 순서이다.

NoteBase

문 14. 절삭 방법 A, B, C가 가공 속도에 영향을 미치는지 알아보기 위해 절삭 방법 A, B, C에서 각각 6, 7, 8회 실험한 결과를 분석하여 얻은 분산분석표의 일부는 다음과 같다. 이에 대한 설명으로 옳지 않은 것은?

요인	제곱합	자유도	$F-$값	$p-$값
처리		(가)	(나)	(다)
오차	36	(라)		
전체	40			

① (가)의 값은 2이다.

② (나)의 값은 1이다.

③ (라)의 값은 18이다.

④ (나)의 값이 커지면 (다)의 값도 커진다.

정답 4번

출제 포인트

이 문제는 일원분산분석표에서 처리자유도, 오차자유도, F값을 계산하고 F값과 p값의 방향 관계를 해석하는 문제이다.

조건 정리

주어진 값은 절삭 방법이 3개이므로 집단 수 $k=3$, 표본수는 6, 7, 8이므로 전체 표본수 $N=21$, 오차제곱합은 36, 전체제곱합은 40이다.
구해야 할 값은 처리 자유도 (가), 오차 자유도 (라), F값 (나), 그리고 F값과 p값의 관계이다.
주의할 조건은 분산분석의 p값은 관측된 F값보다 크거나 같은 값이 나올 확률이므로 F값이 커질수록 p값은 작아진다는 점이다.

사용할 공식

일원분산분석에서는 처리 자유도 $=k-1$, 오차 자유도 $=N-k$이다.
또한 처리제곱합 =전체제곱합-오차제곱합, 평균제곱은 MS = SS/df, 검정통계량은 F=MS_처리/MS_오차이다.
이 공식을 쓰는 이유는 표가 일원분산분석표의 일부이며, 빈칸 (가), (나), (라)는 모두 분산분석의 기본 관계식으로 바로 구할 수 있기 때문이다.

NoteBase

계산 과정

1. 처리 자유도는 집단 수가 3개이므로 (가)=k-1=3-1=2이다.
2. 전체 표본수는 $N = 6 + 7 + 8 = 21$이므로 오차 자유도는 (라)=N-k=21-3=18이다.
3. 처리제곱합은 $40 - 36 = 4$이다.
4. 처리평균제곱은 $4/2 = 2$, 오차평균제곱은 $36/18 = 2$이므로 F값은 (나)=2/2=1이다.
5. p값은 $P(F_{2,18} \geq 1)$인 상측확률이므로 1이 아니며, F값이 커질수록 더 극단적인 값이 되어 p값은 작아진다.

선지별 해설

① O
처리 자유도는 집단 수에서 1을 뺀 값이므로 $3 - 1 = 2$이다. “(가)의 값은 2이다”는 맞는 설명이다.

② O
처리제곱합은 4, 처리자유도는 2이므로 처리평균제곱은 2이다. 오차평균제곱은 $36/18 = 2$이므로 $F = 2/2 = 1$이다. “(나)의 값은 1이다”는 맞는 설명이다.

③ O
오차 자유도는 전체 표본수 21에서 집단 수 3을 뺀 값이므로 18이다. “(라)의 값은 18이다”는 맞는 설명이다.

④ X
분산분석의 p값은 관측된 F값보다 크거나 같은 값이 나올 상측확률이다. 따라서 “(나)의 값이 커지면 (다)의 값도 커진다”가 틀린 표현이다. 올바르게는 “(나)의 값이 커지면 (다)의 값은 작아진다”가 된다.

오답 포인트

이 문항의 함정은 F값과 p값의 방향을 같은 방향으로 착각하는 점이다. F값이 커질수록 귀무가설에서 멀어지므로 p값은 커지는 것이 아니라 작아진다. 또 오차 자유도를 전체 표본수에서 1을 빼는 값으로 착각하면 20으로 잘못 계산할 수 있다.

다시 나오면 이렇게 판단

분산분석표 문제는 먼저 집단 수와 전체 표본수로 자유도를 채운 뒤, 제곱합에서 평균제곱과 F값을 계산한다. 마지막에는 F값이 커질수록 p값은 작아진다는 검정통계량의 방향 관계를 확인하면 된다.

암기 포인트

일원분산분석에서는 처리 자유도 $k - 1$, 오차 자유도 $N - k$, 그리고 큰 F값은 작은 p값으로 이어진다.

NoteBase

문 15. 유의수준과 p-값(유의확률)에 대한 설명으로 옳지 않은 것은?

① 유의수준 α의 검정에서 제1종 오류를 범할 확률이 α보다 작거나 같다.

② p-값은 검정통계량의 관측값을 기준으로 제2종 오류를 범할 확률이다.

③ p-값이 주어진 유의수준보다 작으면 귀무가설을 기각한다.

④ 같은 자료를 사용하여 동일한 가설검정을 하더라도 유의수준의 값에 따라 검정결과가 달라질 수 있다.

정답 2번

출제 포인트

이 문제는 유의수준, 제1종 오류, 제2종 오류, p-값의 정의를 정확히 구분할 수 있는지를 묻는 문제이다.

핵심 개념

유의수준 α는 귀무가설이 참일 때 이를 잘못 기각하는 제1종 오류의 허용 상한이다.
p-값은 귀무가설이 참이라는 가정 아래, 관측된 검정통계량과 같거나 더 극단적인 값이 나올 확률이다.
제2종 오류는 귀무가설이 거짓인데도 기각하지 못하는 오류로, p-값과는 전혀 다른 개념이다.

선지별 해설

① O: 유의수준 α로 검정을 정하면 제1종 오류 확률은 보통 α 이하가 되도록 기각역을 잡는다. 이산분포에서는 정확히 α가 아니라 그보다 작을 수도 있으므로 옳은 설명이다.
② X: 틀린 부분은 '제2종 오류를 범할 확률'이다. p-값은 제2종 오류 확률이 아니라, 귀무가설이 참일 때 관측된 검정통계량보다 같거나 더 극단적인 결과가 나올 확률이다. 제2종 오류 확률은 대립가설이 참일 때 귀무가설을 기각하지 못할 확률로 정의된다.
③ O: 같은 검정에서 p-값이 주어진 유의수준보다 작으면 귀무가설을 기각한다. 검정 결과를 p-값과 유의수준 비교로 판단하는 표준 규칙에 맞는 설명이다.
④ O: 자료와 검정방법이 같아도 유의수준을 0.01로 둘지 0.05로 둘지에 따라 기각 여부가 달라질 수 있다. 예를 들어 $p = 0.03$이면 0.05에서는 기각하고 0.01에서는 기각하지 않는다.

오답 포인트

②는 p-값을 '오류 확률'이라고 막연히 외운 수험생이 가장 많이 헷갈리는 선지이다. p-값은 귀무가설이 참일 때의 데이터 극단성이고, 제2종 오류는 대립가설이 참일 때의 기각 실패 확률이라는 점에서 기준 가설부터 다르다.

다시 나오면 이렇게 판단

검정 문제에서는 먼저 '어느 가설이 참이라고 가정한 확률인가'를 본다. 귀무가설이 참이라고 두고 극단적 관측값의 확률을 묻으면 p-값이고, 귀무가설이 거짓인데 놓치는 확률을 묻으면 제2종 오류이다.

NoteBase

암기 포인트

유의수준은 제1종 오류의 기준, p-값은 귀무가설 기준의 극단확률, 제2종 오류는 대립가설 기준의 기각 실패이다.

NoteBase

문 16. 변수 X와 Y에 대한 자료 (x_i, y_i) $(i = 1, 2, \cdots, 6)$에서 X와 Y의 표본평균은 각각 3과 4이고, X와 Y의 표본표준편차는 1로 동일하다. 이 자료에 단순선형회귀모형 $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$를 적합하여 얻은 β_1의 최소제곱추정값이 0.5일 때, 이에 대한 설명으로 옳은 것만을 모두 고르면? (단, ε_i는 $N(0, \sigma^2)$을 따르고 서로 독립이다)

ㄱ. β_0의 최소제곱추정값은 2.5이다.
ㄴ. 잔차제곱합의 자유도는 4이다.
ㄷ. 결정계수(R^2)는 0.25이다.

① ㄱ
② ㄱ, ㄴ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 4번

출제 포인트

단순선형회귀에서 $\hat{\beta}_1 = r\frac{s_y}{s_x}$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, 잔차제곱합의 자유도 $n - 2$, 결정계수 $R^2 = r^2$의 연결을 묻는 문제이다.

조건 정리

주어진 값은 $\bar{x} = 3, \bar{y} = 4, s_x = s_y = 1, n = 6, \hat{\beta}_1 = 0.5$이다.
판단 대상은 ㄱ, ㄴ, ㄷ의 참거짓과 이를 조합한 ①~④ 선택지이다.

사용할 공식

$\hat{\beta}_1 = r\frac{s_y}{s_x}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \mathrm{df}(SSE) = n - 2, \quad R^2 = r^2$
이 문제는 절편이 있는 단순선형회귀모형이고, s_x와 s_y가 같으므로 $\hat{\beta}_1$에서 바로 상관계수 r를 복원할 수 있다.

계산 과정

1. $\hat{\beta}_1 = r\frac{s_y}{s_x}$에서 $s_y = s_x = 1$이므로 $0.5 = r$이다.
2. 절편은 $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 4 - 0.5 \times 3 = 2.5$이다.
3. 잔차제곱합의 자유도는 단순회귀의 모수 2개를 추정하므로 $n - 2 = 6 - 2 = 4$이다.
4. 결정계수는 $R^2 = r^2 = 0.5^2 = 0.25$이다.
5. ㄱ, ㄴ, ㄷ이 모두 옳으므로 정답 선택지는 ④이다.

NoteBase

선지별 해설

① X: ①은 ㄱ만 옳다고 본 선택지이다. 실제로 ㄱ은 $\hat{\beta}_0 = 2.5$이므로 옳고, ㄴ도 자유도 4로 옳고, ㄷ도 $R^2 = 0.25$로 옳다. 틀린 표현은 'ㄱ만 옳다'이고, 올바르게는 'ㄱ, ㄴ, ㄷ이 모두 옳다'이다.

② X: ②는 ㄱ, ㄴ만 옳다고 본 선택지이다. ㄱ과 ㄴ은 맞지만 $R^2 = r^2 = 0.25$이므로 ㄷ도 옳다. 틀린 표현은 'ㄷ은 틀리다'이고, 올바르게는 'ㄷ도 옳다'이다.

③ X: ③은 ㄴ, ㄷ만 옳다고 본 선택지이다. ㄴ과 ㄷ이 옳은 것은 맞지만, 절편도 2.5이므로 ㄱ 역시 옳다. 틀린 표현은 'ㄱ은 틀리다'이고, 올바르게는 'ㄱ도 옳다'이다.

④ O: ㄱ은 $\hat{\beta}_0 = 4 - 0.5 \times 3 = 2.5$로 옳고, ㄴ은 잔차제곱합 자유도 $6 - 2 = 4$로 옳고, ㄷ은 $R^2 = 0.5^2 = 0.25$로 옳다. 따라서 ㄱ, ㄴ, ㄷ을 모두 포함한 ④가 맞다.

오답 포인트

표본표준편차가 둘 다 1이면 $\hat{\beta}_1$과 상관계수 r가 같아진다는 점을 놓치기 쉽다. 또 잔차제곱합의 자유도를 $n-1$이나 n으로 착각하면 ㄴ을 틀리게 된다. 조합형 문항이라 개별 문장 하나만 맞히고 선택지를 고르면 ①~③에 걸리기 쉽다.

다시 나오면 이렇게 판단

단순회귀 요약값 문제는 먼저 $\hat{\beta}_1 = r\frac{s_y}{s_x}$로 상관계수를 복원하고, 다음에 $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$, 자유도 $n-2$, 결정계수 r^2를 순서대로 연결하면 안정적으로 판단할 수 있다.

암기 포인트

$s_x = s_y$이면 단순회귀의 기울기와 상관계수가 같고, 절편은 $\bar{y} - \hat{\beta}_1\bar{x}$, 결정계수는 r^2이다.

NoteBase

문 17. 지역 A, B, C에서 특정 정책에 대한 지지 여부를 조사한 결과가 다음과 같을 때, 지역과 지지 여부의 독립성을 검정하기 위한 카이제곱 검정통계량의 값과 유의수준 5%에서의 검정결과를 바르게 연결한 것은? (단, $\chi^2_\alpha(k)$는 자유도가 k인 카이제곱 분포의 제$100 \times (1-\alpha)$백분위수를 나타내고, $\chi^2_{0.05}(1) = 3.84,\ \chi^2_{0.05}(2) = 5.99,\ \chi^2_{0.05}(6) = 12.59$이다)

지역 \ 지지 여부	지지함	지지하지 않음
A	30	20
B	40	10
C	50	50

① 카이제곱 검정통계량의 값: 12.5, 검정결과: 지역과 지지 여부는 독립이다.

② 카이제곱 검정통계량의 값: 12.5, 검정결과: 지역과 지지 여부는 독립이 아니다.

③ 카이제곱 검정통계량의 값: 16.5, 검정결과: 지역과 지지 여부는 독립이다.

④ 카이제곱 검정통계량의 값: 16.5, 검정결과: 지역과 지지 여부는 독립이 아니다.

정답 2번

출제 포인트

지역과 지지 여부의 독립성 검정에서 기대도수, 자유도 $(r-1)(c-1)$, 검정통계량, 임계값 비교를 연결하는 문제이다.

조건 정리

관측도수는 A가 (30, 20), B가 (40, 10), C가 (50, 50)이다.
행합은 A 50, B 50, C 100이고 열합은 지지함 120, 지지하지 않음 80, 전체합은 200이다.
유의수준은 5%이고 자유도 2의 임계값은 $\chi^2_{0.05}(2) = 5.99$이다.

사용할 공식

독립성 검정이므로 기대도수는 E_ij=(행합)(열합)전체합, 검정통계량은 $\chi^2 = \sum \frac{(O-E)^2}{E}$를 쓴다.
이 공식을 쓰는 이유는 범주형 변수 두 개의 독립 여부를 관측도수와 기대도수의 차이로 판단하는 문제이기 때문이다.

NoteBase

계산 과정

1. 기대도수를 구하면 A는 (30, 20), B는 (30, 20), C는 (60, 40)이다.
2. 검정통계량은
$\chi^2 = \frac{(30-30)^2}{30} + \frac{(20-20)^2}{20} + \frac{(40-30)^2}{30} + \frac{(10-20)^2}{20} + \frac{(50-60)^2}{60} + \frac{(50-40)^2}{40}$
$= 0 + 0 + \frac{100}{30} + \frac{100}{20} + \frac{100}{60} + \frac{100}{40} = 12.5$
이다.
3. 자유도는 $(3-1)(2-1) = 2$이고, $12.5 > 5.99$이므로 귀무가설을 기각한다. 따라서 지역과 지지 여부는 독립이 아니다.

선지별 해설

① X: 검정통계량 12.5는 맞지만 $12.5 > 5.99$이므로 '독립이다'는 틀리다. 틀린 표현은 '독립이다'이고, 올바르게는 '독립이 아니다'이다.
② O: 검정통계량이 12.5이고 자유도 2에서 임계값 5.99보다 크므로 독립성 귀무가설을 기각한다. 따라서 '12.5, 독립이 아니다'의 연결이 맞다.
③ X: 검정통계량을 16.5로 본 것이 틀리고 검정결과도 틀리다. 틀린 표현은 '16.5'와 '독립이다'이고, 올바르게는 '12.5'이며 '독립이 아니다'이다.
④ X: 검정결과인 '독립이 아니다'는 방향이 맞지만 검정통계량 16.5가 틀리므로 전체 선지는 틀리다. 틀린 표현은 '16.5'이고, 올바르게는 '12.5'이다.

오답 포인트

기대도수를 구하지 않고 관측도수만 바로 비교하거나, 검정통계량을 구한 뒤 자유도 2의 임계값 5.99와 비교하지 않아 결론을 뒤집는 경우가 많다. 이 문항은 값 계산과 기각 여부 판단을 따로 확인하는지가 핵심 함정이다.

다시 나오면 이렇게 판단

독립성 검정 표가 나오면 행합, 열합, 전체합으로 기대도수를 먼저 만든다. 그다음 $\sum \frac{(O-E)^2}{E}$를 계산하고 자유도 $(r-1)(c-1)$의 임계값과 비교해 독립 여부를 판단한다.

암기 포인트

독립성 검정은 '기대도수 작성 → χ^2 계산 → 자유도 $(r-1)(c-1)$ 확인 → 임계값 비교' 순서로 정리된다.

NoteBase

문 18. 확률변수 X에 대하여 $E(X)=\frac{1}{p}$이고 $Var(X)=\frac{1-p}{p^2}$일 때, $Var(X)$의 불편추정량은?

① $\frac{X(X-1)}{2}$
② $\frac{X^2-1}{2}$
③ $\frac{X^2}{2}$
④ $\frac{X}{2}$

정답 1번

출제 포인트

이 문제는 $E(X)=1/p, Var(X)=(1-p)/p^2$를 만족하는 기하분포의 적률을 이용해 $Var(X)$의 불편추정량을 찾는 문제이다.

조건 정리

주어진 값은 $E(X)=1/p, Var(X)=(1-p)/p^2$이다.
구해야 할 값은 기대값이 $Var(X)=(1-p)/p^2$와 정확히 같은 통계량이다.
보기는 X 한 개로 만든 함수들이므로 각 보기의 기대값을 계산해 비교하면 된다.

사용할 공식

불편추정량은 그 기대값이 모수와 같아야 하므로 $E[T(X)]=Var(X)$를 확인한다.
또한 $E[X(X-1)]=E(X^2)-E(X), E(X^2)=Var(X)+\{E(X)\}^2$를 사용한다.
이 공식을 쓰는 이유는 보기들이 모두 $X, X^2, X(X-1)$ 꼴이라서 기대값을 적률로 바로 비교할 수 있기 때문이다.

계산 과정

먼저 $E(X^2)=Var(X)+\{E(X)\}^2=\frac{1-p}{p^2}+\frac{1}{p^2}=\frac{2-p}{p^2}$이다.
따라서
$E[X(X-1)]=E(X^2)-E(X)=\frac{2-p}{p^2}-\frac{1}{p}=\frac{2-p}{p^2}-\frac{p}{p^2}=\frac{2(1-p)}{p^2}$
이다.
그러므로
$E\left(\frac{X(X-1)}{2}\right)=\frac{1}{2}E[X(X-1)]=\frac{1-p}{p^2}=Var(X)$
가 되어 1번이 불편추정량이다.

NoteBase

선지 판단

1번 O: $E\left(\frac{X(X-1)}{2}\right)=\frac{1-p}{p^2}=Var(X)$이므로 불편추정량이 맞다.
2번 X: $E\left(\frac{X^2-1}{2}\right)=\frac{E(X^2)-1}{2}=\frac{(2-p)-p^2}{2p^2}$로 $(1-p)/p^2$와 같지 않다. 틀린 표현은 "불편추정량"이고, 올바르게는 "\(Var(X)의 불편추정량이 아니다"\)이다.
3번 X: $E\left(\frac{X^2}{2}\right)=\frac{2-p}{2p^2}$로 $(1-p)/p^2$보다 크다. X^2만 쓰면 $E(X)$ 항이 빠져 분산으로 맞춰지지 않는다. 올바르게는 "\(X^2/2는 불편추정량이 아니고, $X(X-1)/2$여야 한다"\)이다.
4번 X: $E\left(\frac{X}{2}\right)=\frac{1}{2p}$로 분산 $(1-p)/p^2$와 전혀 다른 값이다. 평균을 절반으로 줄인다고 분산이 되지 않는다. 올바르게는 "\(X/2는 $Var(X)$의 불편추정량이 아니다"\)이다.

오답 포인트

분산의 불편추정량을 찾을 때 $E(X^2)$만 보거나 $E(X)$를 따로 빼지 않으면 2번과 3번 같은 함정에 걸린다. 이 문항은 $Var(X)=E(X^2)-\{E(X)\}^2$를 그대로 대입하는 문제가 아니라, 보기의 기대값이 분산과 일치하는지를 묻는 문제이다.

다시 나오면 이렇게 판단

불편추정량 문제는 각 보기를 $T(X)$로 두고 먼저 $E[T(X)]$를 계산한다. 보기 안에 X^2와 X가 함께 있으면 $E(X^2)$, $E(X)$, $E[X(X-1)]$로 바꿔 본 뒤 목표 모수와 정확히 일치하는지만 판단하면 된다.

암기 포인트

$E[X(X-1)]=E(X^2)-E(X)$이고, 이 값이 기하분포에서는 $2(1-p)/p^2$가 되므로 $Var(X)$의 불편추정량은 $X(X-1)/2$이다.

NoteBase

문 19. 평균이 μ이고 분산은 $\sigma^2 = 144$인 모집단으로부터 64개의 표본을 임의추출하여 가설 $H_0 : \mu = 30$ 대 $H_1 : \mu > 30$를 유의수준 5%에서 검정하고자 한다. $\mu = 39$에서 제2종 오류를 범할 확률은? (단, z_α는 표준정규분포의 제$100 \times (1-\alpha)$ 백분위수이다)

① $P(Z > z_{0.05} + 4)$

② $P(Z < z_{0.05} - 4)$

③ $P(Z > z_{0.05} + 6)$

④ $P(Z < z_{0.05} - 6)$

정답 4번

출제 포인트

이 문제는 모분산을 아는 단측 Z검정에서 기각역을 먼저 구한 뒤, $\mu = 39$일 때 제2종 오류 확률 β를 표준정규확률로 바꾸는 문제이다.

조건 정리

주어진 값: $H_0 : \mu = 30, H_1 : \mu > 30, \sigma^2 = 144, n = 64, \alpha = 0.05$이다.
구해야 할 값: $\mu = 39$일 때 제2종 오류 확률이다.
주의할 조건: 모분산을 아니까 표본평균의 표준오차는 $\sigma/\sqrt{n} = 12/8 = 1.5$이고, 제2종 오류는 대립가설이 참인데도 H_0를 기각하지 못하는 확률이다.

사용할 공식

기각역은 $\frac{\bar{X} - 30}{1.5} > z_{0.05}$ 이므로 $\bar{X} > 30 + 1.5z_{0.05}$이다.
제2종 오류 확률은 $\beta = P(\bar{X} \le 30 + 1.5z_{0.05} \mid \mu = 39)$ 이다.
이 공식을 쓰는 이유: 단측 상측검정에서는 임계값보다 큰 경우에만 H_0를 기각하므로, $\mu = 39$에서의 제2종 오류는 임계값 이하에 머무는 비기각 확률로 계산된다.

계산 과정

1. 표본평균의 분포는 $\bar{X} \sim N(\mu, 144/64) = N(\mu, 2.25)$이므로 표준오차는 1.5이다.
2. 유의수준 5%의 기각역은 $\bar{X} > 30 + 1.5z_{0.05}$이다.
3. $\mu = 39$에서 제2종 오류 확률은
$P(\bar{X} \le 30 + 1.5z_{0.05} \mid \mu = 39) = P\left(\frac{\bar{X} - 39}{1.5} \le \frac{30 + 1.5z_{0.05} - 39}{1.5}\right) = P(Z \le z_{0.05} - 6)$ 이다.

NoteBase

선지별 해설

① X: $P(Z > z_{0.05} + 4)$는 꼬리 방향도 반대이고 이동량도 틀리다. 제2종 오류는 기각역 바깥이 아니라 비기각역에 머무는 확률이므로 $<$가 와야 하고, 표준화된 평균 차이는 $(39 - 30)/1.5 = 6$이다. 틀린 표현은 `> z_{0.05}+4`이고, 올바르게는 `$P(Z < z_{0.05} - 6)$`이다.
② X: 꼬리 방향은 비기각확률이라서 맞지만, 이동량 4가 틀리다. $\mu = 39$와 귀무가설 평균 30의 차이 9를 표준오차 1.5로 나누면 6이므로 `-4`가 아니라 `-6`이어야 한다. 틀린 표현은 `z_{0.05}-4`이고, 올바르게는 `$P(Z < z_{0.05} - 6)$`이다.
③ X: 이동량 6은 맞지만 $P(Z > z_{0.05} + 6)$는 상측 꼬리여서 검정력의 일부 방향을 잘못 잡은 식이다. 제2종 오류는 기각확률이 아니라 비기각확률이므로 `>`가 아니라 `<`가 와야 한다. 틀린 표현은 `> z_{0.05}+6`이고, 올바르게는 `$P(Z < z_{0.05} - 6)$`이다.
④ O: 기각역이 $\bar{X} > 30 + 1.5z_{0.05}$이므로 $\mu = 39$에서 제2종 오류는 그 임계값 이하일 확률이다. 이를 표준화하면 정확히 `$P(Z < z_{0.05} - 6)$`이 된다.

오답 포인트

제2종 오류를 검정력과 혼동해 꼬리 방향을 반대로 잡기 쉽다. 또 평균 차이 $39 - 30 = 9$를 표준편차 12가 아니라 표준오차 1.5로 나누어야 하는데, 이 표준화를 놓치면 `4` 같은 오답으로 흐르기 쉽다.

다시 나오면 이렇게 판단

단측 검정의 제2종 오류 문제는 `기각역 설정 → 참인 평균에서 비기각확률로 바꾸기 → 표준오차로 표준화` 순서로 판단한다. 제2종 오류는 항상 `대립가설이 참일 때 기각하지 못할 확률`이라는 문장을 먼저 붙이면 꼬리 방향이 흔들리지 않는다.

암기 포인트

모분산을 아는 상측 Z검정에서는 `$\beta$ = 참인 $\mu$에서 임계값 이하 확률`이고, 이동량은 `평균 차이 ÷ 표준오차`이다.

NoteBase

문 20. n개의 자료 (x_i, y_i)에 단순선형회귀모형 $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$를 최소제곱법으로 적합하여 얻은 회귀식은 $\hat{y}_i = 2 + 3x_i$이었다. 이 회귀식에 대한 회귀제곱합은 640이고 잔차제곱합은 360이었다. 자료 (x_i, y_i)를 $x_i^* = 1 + 3x_i$, $y_i^* = 5 - \frac{y_i}{10}$로 변환하여 단순선형회귀모형 $Y_i^* = \gamma_0 + \gamma_1 x_i^* + \eta_i$를 적합하였을 때, γ_1의 최소제곱추정값과 (x_i^*, y_i^*)의 표본상관계수를 바르게 연결한 것은? (단, ε_i는 $N(0, \sigma_\varepsilon^2)$를 따르고 서로 독립이며, η_i는 $N(0, \sigma_\eta^2)$를 따르고 서로 독립이다)

① γ_1의 최소제곱추정값: -0.3, (x_i^*, y_i^*)의 표본상관계수: -0.8

② γ_1의 최소제곱추정값: -0.3, (x_i^*, y_i^*)의 표본상관계수: 0.8

③ γ_1의 최소제곱추정값: -0.1, (x_i^*, y_i^*)의 표본상관계수: -0.8

④ γ_1의 최소제곱추정값: -0.1, (x_i^*, y_i^*)의 표본상관계수: 0.8

정답 3번

출제 포인트

이 문제는 단순회귀에서 설명변수와 반응변수를 선형변환했을 때 기울기 추정량과 표본상관계수가 어떻게 바뀌는지를 묻는 계산형 문제이다.

조건 정리

주어진 값: 원래 회귀식은 $\hat{y}_i = 2 + 3x_i$이고 회귀제곱합은 640, 잔차제곱합은 360이다.
구해야 할 값: 변환된 자료 $x_i^* = 1 + 3x_i$, $y_i^* = 5 - \frac{y_i}{10}$에 대한 γ_1의 최소제곱추정값과 (x_i^*, y_i^*)의 표본상관계수이다.
주의할 조건: 평행이동은 기울기와 상관계수에 영향을 주지 않고, 배율과 부호만 영향을 준다.

사용할 공식

$R^2 = \frac{SSR}{SST} = r^2$, $SST = SSR + SSE$이다.
또한 $x^* = a + bx$, $y^* = c + dy$이면 기울기는 $\hat{\gamma}_1 = \frac{d}{b}\hat{\beta}_1$, 표본상관계수는 $r_{x^*y^*} = \mathrm{sgn}(bd) r_{xy}$이다.
이 공식을 쓰는 이유는 이 문항이 원자료의 회귀계수와 제곱합을 주고, 선형변환 뒤의 회귀계수와 상관계수 변화를 직접 묻기 때문이다.

계산 과정

1. 원자료의 총제곱합은 $SST = 640 + 360 = 1000$이다. 따라서 $R^2 = \frac{640}{1000} = 0.64$이고, 원래 기울기 $\hat{\beta}_1 = 3$이 양수이므로 $r_{xy} = +\sqrt{0.64} = 0.8$이다.
2. 변환식에서 $b = 3$, $d = -\frac{1}{10}$이므로 $\hat{\gamma}_1 = \frac{d}{b}\hat{\beta}_1 = \frac{-1/10}{3} \times 3 = -0.1$이다.
3. 표본상관계수는 평행이동의 영향은 없고 배율의 부호만 반영하므로 $r_{x^*y^*} = \mathrm{sgn}(3 \cdot -\frac{1}{10}) \times 0.8 = -0.8$이다.

NoteBase

선지별 해설

① X

- 표본상관계수 -0.8은 맞지만 $\gamma_1 = -0.3$은 틀리다. 기울기는 원래 기울기 3에 $\frac{-1/10}{3}$을 곱하므로 -0.1이 된다.
- 틀린 표현은 "$\gamma_1 = -0.3$"이고, 올바르게는 "$\gamma_1 = -0.1$"이다.

② X

- $\gamma_1 = -0.3$도 틀리고 상관계수 0.8도 틀리다. y에 음의 배율이 곱해졌으므로 상관계수의 부호가 반드시 바뀐다.
- 틀린 표현은 "$\gamma_1 = -0.3$, 표본상관계수 0.8"이고, 올바르게는 "$\gamma_1 = -0.1$, 표본상관계수 -0.8"이다.

③ O

- 계산 결과 $\hat{\gamma}_1 = -0.1$, $r_{x^*y^*} = -0.8$이므로 제시된 값이 모두 맞다.

④ X

- $\gamma_1 = -0.1$은 맞지만 표본상관계수 0.8은 틀리다. x의 양의 배율은 부호를 바꾸지 않지만 y의 음의 배율이 상관계수의 부호를 뒤집는다.
- 틀린 표현은 "표본상관계수 0.8"이고, 올바르게는 "표본상관계수 -0.8"이다.

오답 포인트

①과 ②의 -0.3은 기울기 변환에서 y의 배율만 보고 x의 배율 3으로 다시 나누는 과정을 빠뜨릴 때 나오기 쉽다. ②와 ④의 0.8은 상관계수의 크기만 유지된다고 외우고, $y^* = 5 - \frac{y}{10}$의 음의 부호가 상관계수 부호를 뒤집는다는 점을 놓칠 때 고르기 쉽다.

다시 나오면 이렇게 판단

회귀 변환 문제는 먼저 $R^2 = SSR/SST$로 원래 상관계수의 크기를 구하고, 원래 기울기의 부호로 상관계수 부호를 정한다. 그다음 $x^* = a + bx, y^* = c + dy$ 꼴에서 기울기는 $\frac{d}{b}$배, 상관계수는 bd의 부호만 반영된다고 판단하면 된다.

암기 포인트

$x^* = a + bx,\ y^* = c + dy$이면 $\hat{\gamma}_1 = (d/b)\hat{\beta}_1, r^* = \mathrm{sgn}(bd)r$이다.

NoteBase

문 21. $X_1, X_2, \cdots, X_{10}$**이 정규분포** $N(\mu, \sigma^2)$**으로부터의 임의표본일 때,** μ**와** σ^2**에 대한 95 % 신뢰구간에 대한 설명으로 옳은 것만을 모두 고르면?**

ㄱ. σ^2이 알려져 있는 경우 정규분포를 사용한 μ에 대한 신뢰구간의 길이는 $X_1, X_2, \cdots, X_{10}$의 값에 관계없이 일정하다.
ㄴ. σ^2이 알려져 있지 않은 경우 μ에 대한 신뢰구간의 길이는 $X_1, X_2, \cdots, X_{10}$의 값에 관계없이 일정하다.
ㄷ. μ가 알려져 있지 않은 경우 σ^2에 대한 신뢰구간의 길이는 $X_1, X_2, \cdots, X_{10}$의 값에 관계없이 일정하다.

① ㄱ
② ㄱ, ㄴ
③ ㄱ, ㄷ
④ ㄴ, ㄷ

정답 1번

출제 포인트

이 문제는 모평균 μ와 모분산 σ^2의 95% 신뢰구간 길이가 표본값에 의존하는지 여부를 구분하는 문제이다.

핵심 개념

모분산 σ^2를 알 때 모평균 μ의 95% 신뢰구간은 $\bar{X} \pm z_{0.975}\sigma/\sqrt{10}$이므로 길이가 $2z_{0.975}\sigma/\sqrt{10}$로 고정된다.
모분산 σ^2를 모를 때 모평균 μ의 95% 신뢰구간은 $\bar{X} \pm t_{0.975,\,9}S/\sqrt{10}$이므로 길이가 표본표준편차 S에 따라 달라진다.
모평균 μ를 모를 때 모분산 σ^2의 95% 신뢰구간은 $\left(\frac{9S^2}{\chi^2_{0.975,\,9}}, \frac{9S^2}{\chi^2_{0.025,\,9}}\right)$ 형태이므로 길이가 S^2에 비례하여 달라진다.

NoteBase

선지별 해설

1번 O
- ㄱ만 옳다.
- ㄱ은 모분산을 아는 경우이므로 신뢰구간 길이가 $2z_{0.975}\sigma/\sqrt{10}$로 정해져 있고, $\bar{X}$만 이동할 뿐 길이는 표본값에 따라 변하지 않는다.

2번 X
- ㄴ이 틀렸으므로 ㄱ, ㄴ은 옳지 않다.
- 틀린 부분은 '값에 관계없이 일정하다'이다.
- 실제로는 $\bar{X} \pm t_{0.975,\,9}S/\sqrt{10}$이므로 길이가 $2t_{0.975,\,9}S/\sqrt{10}$가 되어 표본표준편차 S가 바뀌면 길이도 바뀐다.

3번 X
- ㄷ이 틀렸으므로 ㄱ, ㄷ은 옳지 않다.
- 틀린 부분은 '값에 관계없이 일정하다'이다.
- 실제로는 모분산 신뢰구간의 양 끝점이 모두 S^2에 비례하므로 구간 길이도 표본값에 따라 달라진다.

4번 X
- ㄴ과 ㄷ이 모두 틀렸으므로 옳지 않다.
- ㄴ은 표본표준편차 S가, ㄷ은 표본분산 S^2가 신뢰구간 길이를 결정하므로 둘 다 '일정하다'로 고치면 안 된다.

오답 포인트

신뢰구간의 중심이 $\bar{X}$인지, 길이까지 S 또는 S^2가 들어가는지를 구분하지 않으면 ㄴ과 ㄷ을 함께 맞다고 착각하기 쉽다. 이 문항의 함정은 '정규분포에서 나온 표본'이라는 공통 조건 때문에 모든 구간 길이도 고정된다고 오해하게 만드는 데 있다.

다시 나오면 이렇게 판단

신뢰구간 길이 공식을 먼저 본다. 길이 식에 $\bar{X}$만 없고 상수만 있으면 길이는 고정되고, S나 S^2가 들어가면 길이는 표본값에 따라 달라진다고 판단하면 된다.

암기 포인트

알려진 σ는 길이 고정, 모르는 σ는 S가 들어가 길이 가변이다.

NoteBase

문 22. 당뇨병 환자 중 임의로 15명(집단 A)과 10명(집단 B)을 선택하여 집단 A에게 신약을 처방한 후에 얻은 혈당 개선 효과 자료 $\{x_1, x_2, \cdots, x_{15}\}$와 집단 B에게 위약을 처방한 후에 얻은 혈당 개선 효과 자료 $\{y_1, y_2, \cdots, y_{10}\}$를 다음과 같이 요약하였다. 집단 A와 B의 평균 혈당 개선 효과 μ_A와 μ_B에 대한 가설 $H_0 : \mu_A = \mu_B$ 대 $H_1 : \mu_A > \mu_B$를 유의수준 α에서 검정하고자 한다. 이에 대한 설명으로 옳은 것은? (단, 집단 A와 B의 혈당 개선 효과는 분산이 동일한 정규분포를 따르고, $t_\alpha(k)$는 자유도가 k인 t 분포의 제100 $\times (1-\alpha)$ 백분위수이며, $s_P > 0$이다)

$$\bar{x} = \frac{1}{15}\sum_{i=1}^{15} x_i$$
$$\bar{y} = \frac{1}{10}\sum_{i=1}^{10} y_i$$
$$s_P = \sqrt{\frac{1}{23}\left(\sum_{i=1}^{15}(x_i - \bar{x})^2 + \sum_{i=1}^{10}(y_i - \bar{y})^2\right)}$$

① $\frac{\sqrt{6}(\bar{x}-\bar{y})}{s_P^2}$가 $t_\alpha(23)$보다 크면 귀무가설을 기각한다.
② $\frac{\sqrt{6}(\bar{x}-\bar{y})}{s_P^2}$가 $-t_\alpha(23)$보다 작으면 귀무가설을 기각한다.
③ $\frac{\sqrt{6}(\bar{x}-\bar{y})}{s_P}$가 $t_\alpha(23)$보다 크면 귀무가설을 기각한다.
④ $\frac{\sqrt{6}(\bar{x}-\bar{y})}{s_P}$가 $-t_\alpha(23)$보다 작으면 귀무가설을 기각한다.

정답 3번

출제 포인트

이 문제는 등분산을 가정한 두 독립표본 평균 비교에서 pooled t 검정통계량과 우측단측 기각역을 묻는 문제이다.

조건 정리

주어진 값: 집단 A의 표본크기는 15, 집단 B의 표본크기는 10이다.
구해야 할 값: $H_0 : \mu_A = \mu_B, H_1 : \mu_A > \mu_B$에서의 올바른 검정통계량과 기각역이다.
주의할 조건: 두 집단은 분산이 동일한 정규분포를 따르므로 pooled 표준편차 s_P를 쓰고 자유도는 $15 + 10 - 2 = 23$이다.

NoteBase

사용할 공식

$T = \frac{\bar{x} - \bar{y}}{s_P\sqrt{1/15 + 1/10}}$

또한

$\frac{1}{15} + \frac{1}{10} = \frac{1}{6}$

이므로

$T = \frac{\sqrt{6}(\bar{x} - \bar{y})}{s_P}$

이다.

이 공식을 쓰는 이유는 두 독립표본이 등분산 정규모집단에서 왔고, 공통분산을 모를 때 pooled 표준편차 s_P로 표준오차를 구성하는 두 표본 t 검정을 사용하기 때문이다.

계산 과정

1. 문제에서 제시한 s_P는 두 표본의 편차제곱합을 자유도 23으로 나눈 뒤 제곱근을 취한 pooled 표준편차이다.
2. 표준오차는 $s_P\sqrt{1/15 + 1/10} = s_P/\sqrt{6}$이다.
3. 따라서 검정통계량은 $T = \sqrt{6}(\bar{x} - \bar{y})/s_P$가 된다.
4. 대립가설이 $\mu_A > \mu_B$이므로 큰 양의 값에서 귀무가설을 기각하며, 기각역은 $T > t_\alpha(23)$이다.

선지별 해설

① X

검정통계량의 분모는 s_P^2가 아니라 표준오차 $s_P\sqrt{1/15 + 1/10}$이다. 틀린 표현은 "s_P^2"이며, 올바르게는 $T = \sqrt{6}(\bar{x} - \bar{y})/s_P$이다.

② X

①과 같이 분모를 s_P^2로 둔 것부터 틀렸고, 대립가설이 $H_1 : \mu_A > \mu_B$이므로 좌측기각이 아니라 우측기각을 써야 한다. 틀린 표현은 "s_P^2"와 "$-t_\alpha(23)$보다 작으면"이며, 올바르게는 $T > t_\alpha(23)$이다.

③ O

등분산 두 표본 t 검정에서 $T = \sqrt{6}(\bar{x} - \bar{y})/s_P$이고 자유도는 23이다. 또한 $H_1 : \mu_A > \mu_B$이므로 $T > t_\alpha(23)$일 때 귀무가설을 기각하므로 옳다.

④ X

분모를 s_P로 둔 형태는 맞지만 기각역 방향이 반대이다. 틀린 표현은 "$-t_\alpha(23)$보다 작으면"이며, 올바르게는 "$t_\alpha(23)$보다 크면" 귀무가설을 기각한다.

오답 포인트

이 문항의 함정은 pooled 분산 s_P^2와 pooled 표준편차 s_P를 뒤섞는 점, 그리고 $H_1 : \mu_A > \mu_B$의 부호를 놓쳐 우측단측 검정을 좌측단측으로 바꾸는 점이다. 분산을 그대로 분모에 넣으면 단위부터 맞지 않게 된다.

NoteBase

다시 나오면 이렇게 판단

등분산 두 표본 평균 검정은 먼저 대립가설의 방향으로 우측검정인지 좌측검정인지 정하고, 다음으로 s_P가 표준편차인지 s_P^2가 분산인지 구분한 뒤, $T = (\bar{X} - \bar{Y}) / \{ s_P \sqrt{1/n_1 + 1/n_2} \}$ 형태로 정리하면 된다.

암기 포인트

등분산 두 표본 t 검정은 "pooled는 s_P, 검정통계량은 표준오차로 나눈다, H_1이 $>$이면 우측기각"으로 기억하면 된다.

NoteBase

문 23. 변수 X와 Y의 자료 (x_i, y_i) $(i = 1, 2, \ldots, 20)$에서 X의 표본평균은 1이고 표본분산은 3이다. 이 자료에 단순선형회귀모형 $Y_i = \beta x_i + \varepsilon_i$를 적합시킨 β의 최소제곱추정량의 분산은? (단, ε_i는 $N(0, 1)$을 따르고 서로 독립이다)

① $\frac{1}{57}$
② $\frac{1}{60}$
③ $\frac{1}{77}$
④ $\frac{1}{80}$

정답 3번

출제 포인트

이 문제는 절편이 없는 단순회귀모형에서 최소제곱추정량의 분산 $Var(\hat{\beta}) = \sigma^2 / \sum x_i^2$를 구하기 위해 표본평균과 표본분산으로 $\sum x_i^2$를 복원할 수 있는지를 묻는다.

조건 정리

표본크기는 20이다.
$\bar{x} = 1$이다.
표본분산은 3이므로 $s_x^2 = \frac{1}{19} \sum_{i=1}^{20} (x_i - \bar{x})^2 = 3$이다.
오차분산은 $Var(\varepsilon_i) = 1$이다.
구해야 할 값은 $Var(\hat{\beta})$이다.

사용할 공식

절편이 없는 회귀모형 $Y_i = \beta x_i + \varepsilon_i$에서 최소제곱추정량은 $\hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2}$이고, 분산은 $Var(\hat{\beta}) = \frac{\sigma^2}{\sum x_i^2}$이다.
이 공식을 쓰는 이유는 모형에 절편항이 없으므로 분모가 $\sum (x_i - \bar{x})^2$가 아니라 $\sum x_i^2$가 되기 때문이다.

계산 과정

표본분산 정의에서
$\sum_{i=1}^{20} (x_i - \bar{x})^2 = (20 - 1) \times 3 = 57$
또한 항등식 $\sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$를 쓰면
$\sum x_i^2 = 57 + 20 \times 1^2 = 77$
따라서
$Var(\hat{\beta}) = \frac{1}{\sum x_i^2} = \frac{1}{77}$
이므로 정답은 3번이다.

NoteBase

선지 판단

1번 X: $\frac{1}{57}$은 $\sum(x_i-\bar{x})^2=57$만 분모에 넣은 값이다. 그러나 절편이 없는 회귀에서는 분모가 $\sum x_i^2$이어야 하므로 틀렸다. 올바르게는 57에 $n\bar{x}^2=20$을 더해 77을 써야 한다.
2번 X: $\frac{1}{60}$은 표본분산을 $\frac{1}{n}\sum(x_i-\bar{x})^2$로 잘못 써서 $\sum(x_i-\bar{x})^2=20\times3=60$으로 계산한 값이다. 문제의 표본분산은 보통 $n-1$로 나누는 정의이므로 올바르게는 57을 써야 한다.
3번 O: 표본분산으로부터 $\sum(x_i-\bar{x})^2=57$을 구하고, 여기에 $n\bar{x}^2=20$을 더하면 $\sum x_i^2=77$이 된다. 따라서 $Var(\hat{\beta})=1/77$이 맞다.
4번 X: $\frac{1}{80}$은 $\sum x_i^2$를 구하는 구조는 맞지만, 표본분산을 n으로 나누는 분산으로 착각해 $60+20=80$으로 계산한 값이다. 올바르게는 $57+20=77$이다.

오답 포인트

이 문항의 함정은 두 가지이다. 첫째, 절편이 없는 회귀의 분산 분모를 $\sum(x_i-\bar{x})^2$로 착각하기 쉽다. 둘째, 표본분산을 n이 아니라 $n-1$로 나누는 정의를 놓치면 57 대신 60을 쓰게 된다.

다시 나오면 이렇게 판단

회귀계수 분산 문제는 먼저 절편 포함 여부를 확인한 뒤 분모가 $\sum x_i^2$인지 $\sum(x_i-\bar{x})^2$인지 구분한다. 그다음 표본분산 정의가 $n-1$ 기준인지 확인하고, 마지막에 $\sum x_i^2=\sum(x_i-\bar{x})^2+n\bar{x}^2$를 적용하면 된다.

암기 포인트

절편 없음이면 분산 분모는 중심화된 제곱합이 아니라 $\sum x_i^2$이고, 표본분산은 기본적으로 $n-1$ 기준이다.

NoteBase

문 24. 확률밀도함수가 다음과 같은 임의표본 X_1, X_2, X_3의 중앙값을 M이라 할 때, $P\left(M < \frac{1}{2}\right)$의 값은?

f(x)=cases2-2x,&0<x<1\\0,&그 외cases

① $\frac{5}{32}$
② $\frac{7}{16}$
③ $\frac{9}{16}$
④ $\frac{27}{32}$

정답 4번

출제 포인트

이 문제는 표본 3개의 중앙값이 어떤 기준보다 작을 확률을, 먼저 $F(1/2)$를 구한 뒤 이항확률로 계산하는 문제이다.

조건 정리

주어진 밀도함수는 $f(x) = 2 - 2x$ $(0 < x < 1)$이다.
구해야 할 값은 표본 X_1, X_2, X_3의 중앙값 M에 대하여 $P(M < 1/2)$이다.
$M < 1/2$가 되려면 세 표본 중 적어도 2개가 1/2보다 작아야 한다.

사용할 공식

$p = F(1/2) = P(X < 1/2)$를 먼저 구한 뒤, $M < 1/2$ 사건을 '성공확률 p인 시행 3번 중 성공이 2번 이상'으로 본다.
$P(M < 1/2) = \binom{3}{2}p^2(1-p) + p^3$
이 공식을 쓰는 이유는 표본이 3개일 때 중앙값이 1/2보다 작으려면 정렬했을 때 둘째 값이 1/2보다 작아야 하고, 이는 곧 세 개 중 두 개 이상이 1/2보다 작은 경우와 같기 때문이다.

계산 과정

1. 누적확률을 구하면
p=P(X<1/2)= ∫ _0^1/2(2-2x) dx=[2x-x^2]_0^1/2=1-14=34
2. 따라서 각 표본이 1/2보다 작을 확률은 $p = 3/4$이다.
3. 중앙값이 1/2보다 작을 확률은
P(M<1/2)=32p^2(1-p)+p^3=3·≤ft(34)^2≤ft(14)+≤ft(34)^3=(27)/(64)+(27)/(64)=(27)/(32)
따라서 정답값은 27/32이다.

NoteBase

선지별 해설

① X: 5/32는 계산 결과와 맞지 않는다. $P(X < 1/2) = 3/4$를 구하지 못하거나 경우의 수를 잘못 센 값이다. 올바르게는 $P(M < 1/2) = 27/32$이다.
② X: 7/16은 세 표본 중 정확히 2개만 1/2보다 작은 경우만 보거나 적분값을 잘못 처리했을 때 나올 수 있는 값이다. 중앙값이 1/2보다 작으려면 정확히 2개뿐 아니라 3개 모두 작은 경우도 포함해야 한다. 올바르게는 27/32이다.
③ X: 9/16은 $(3/4)^2$처럼 두 개가 작을 확률을 단순 제곱으로 본 값에 가깝다. 그러나 중앙값 조건은 경우의 수 $\binom{3}{2}$와 3개 모두 작은 경우 p^3까지 함께 세어야 한다. 올바르게는 27/32이다.
④ O: $P(X < 1/2) = 3/4$를 구한 뒤, 세 표본 중 적어도 2개가 1/2보다 작은 확률을 더하면 27/32가 된다. 계산 결과와 일치하므로 옳다.

오답 포인트

중앙값 조건을 '정확히 2개'로 축소하거나 경우의 수 $\binom{3}{2}$를 빠뜨리기 쉽다. 또 먼저 $F(1/2) = 3/4$를 구하지 않고 점값 $f(1/2)$를 확률처럼 다루면 오답이 된다.

다시 나오면 이렇게 판단

표본 3개의 중앙값 확률은 먼저 '기준보다 작은 표본이 몇 개 필요한가'로 사건을 바꾼다. 그다음 한 개 표본의 확률 $p = F(a)$를 적분으로 구하고, 마지막에 $Y \sim Bin(3, p)$로 바꾸어 $P(Y \geq 2)$를 계산하면 된다.

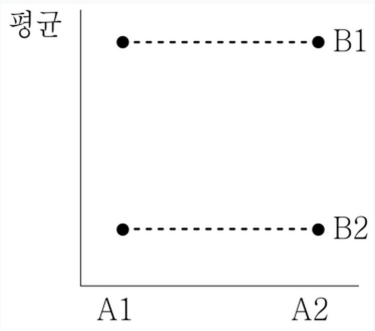
암기 포인트

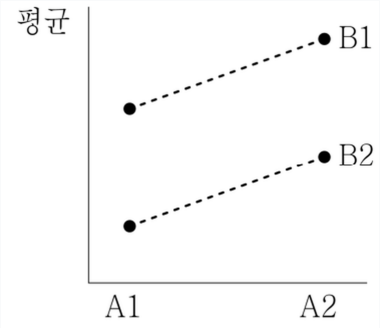
표본 3개의 중앙값은 '세 개 중 두 개 이상'으로 바꾸고, 연속형 확률은 밀도값이 아니라 누적확률 $F(a)$로 계산한다.

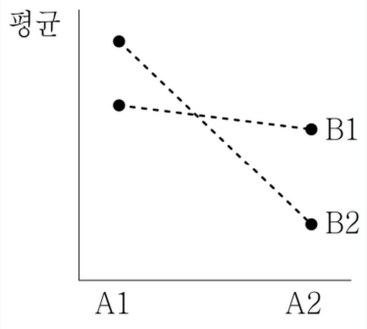
NoteBase

문 25. 반복수가 동일한 이원배치법에서 인자 A의 수준이 A1, A2이고 인자 B의 수준이 B1, B2인 실험을 통해 얻은 결과의 분산분석표는 다음과 같다. 인자의 수준 조합에서 반응변수 평균을 나타낸 그림 중 분산분석표의 결과와 부합하는 것은? (단, 유의수준은 5%이다)

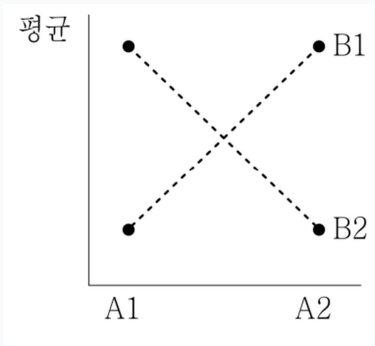
요인	제곱합	자유도	평균제곱	$F-$값	$p-$값
인자 A	784	1	784	26.13	0.0003
인자 B	16	1	16	0.53	0.4792
교호작용$(A\times B)$	400	1	400	13.3	0.0033
오차	360	12	30		
전체	1560	15			

① 

② 

③ 

NoteBase

④ 

정답 3번

출제 포인트

이 문제는 반복수가 같은 이원배치 분산분석표에서 인자 A, 인자 B, 교호작용의 유의성을 읽고 수준평균 그래프의 모양으로 연결하는 문제이다.

선지별 해설

1번: X. 두 선이 수평이고 서로 평행하므로 인자 A 효과가 없고 교호작용도 없다. B1과 B2의 높이 차만 일정하므로 인자 B 주효과만 있는 그림이다. 이 문항의 결과에 맞게 고치면 두 선이 평행이 아니어야 하고 A1과 A2의 열평균도 달라져야 한다.
2번: X. 두 선이 평행하게 함께 상승하므로 인자 A와 인자 B의 주효과는 모두 있지만 교호작용은 없다. 이 문항은 인자 B의 p값이 0.4792로 유의하지 않고 교호작용의 p값이 0.0033으로 유의하므로, B의 전체 평균 차이가 뚜렷하면 안 되고 평행선이어도 안 된다. 맞는 그림으로 고치면 비평행이면서 B의 선평균은 같아야 한다.
3번: O. 두 선이 평행하지 않고 서로 교차하므로 교호작용이 있다. 또한 A1 쪽 두 점의 평균이 A2 쪽 두 점의 평균보다 커 보여 인자 A 주효과가 있고, B1과 B2 각 선의 전체 평균 높이는 비슷하여 인자 B 주효과는 없다. 유의수준 5%에서 인자 A와 교호작용만 유의한 분산분석표와 일치한다.
4번: X. 두 선이 교차하므로 교호작용은 있지만 대칭적인 X 모양이라 A1과 A2의 열평균이 같아 인자 A 주효과가 없다. 이 문항은 인자 A의 p값이 0.0003으로 유의하므로 A1과 A2의 전체 평균 차이가 나타나야 한다. 맞는 그림으로 고치면 한쪽 열평균이 다른 쪽보다 크도록 비대칭 교차가 되어야 한다.

오답 포인트

교호작용이 유의하다는 이유만으로 교차하는 그림만 찾으면 4번에도 끌리기 쉽다. 이 문제는 교호작용뿐 아니라 인자 A 주효과도 유의하고 인자 B 주효과는 유의하지 않으므로, 선의 평행 여부와 함께 열평균 차이와 선평균 차이를 같이 봐야 한다.

NoteBase

다시 나오면 이렇게 판단

먼저 p값으로 유의한 효과를 고른다. 교호작용이 유의하면 비평행선만 남기고, 인자 A 주효과는 A1과 A2의 열평균 차이로, 인자 B 주효과는 B1과 B2의 선평균 차이로 확인한다. 반복수가 같으면 이 평균을 단순평균으로 읽으면 된다.

암기 포인트

이원배치 평균도표는 교호작용은 비평행, A 주효과는 열평균 차이, B 주효과는 선평균 차이로 읽는다.

NoteBase