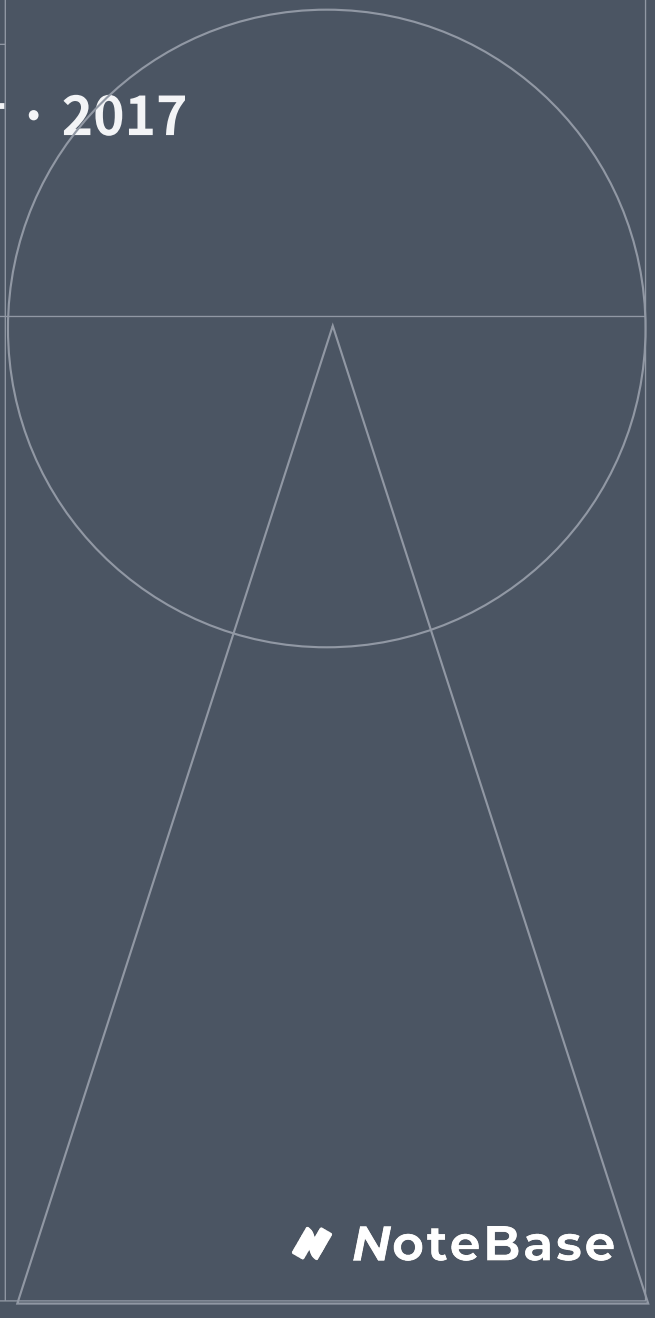
기출OX 해설집

공무원/국가직/7급 · 통계학 · 2017



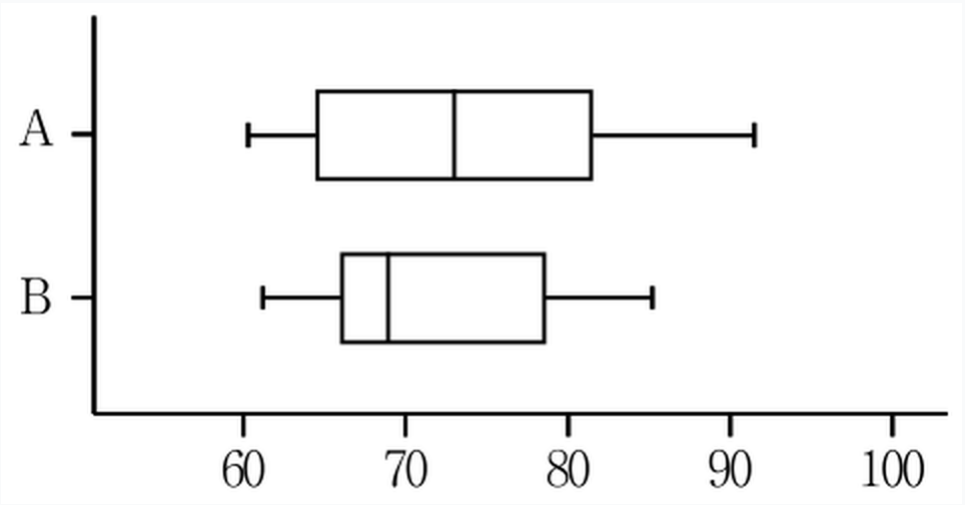
NoteBase

기출OX 해설집

공무원/국가직/7급 · 통계학 · 2017

공무원/국가직/7급 · 통계학 · 2017

문 1. 다음은 어느 대학교의 학과 A와 B에 대한 통계학 성적의 상자그림(box plot)이다. 이에 대한 설명으로 옳지 않은 것은?



① 성적이 가장 좋은 학생은 학과 A에 속해 있다.

② 학과 A의 사분위범위(interquartile range)는 학과 B의 사분위범위보다 크다.

③ 학과 A가 학과 B보다 70점 이하인 학생의 비율이 높다.

④ 학과 A의 제2사분위수는 학과 B의 제2사분위수보다 크다.

정답 3번

출제 포인트

상자그림(box plot)에서 최솟값·제1사분위수·중앙값(제2사분위수)·제3사분위수·최댓값의 위치를 읽어 두 집단의 분포를 비교하는 문제이다. 특히 상자 안에서 특정 점수(70점)가 차지하는 위치를 보고 그 점수 이하인 자료의 비율을 어림하는 능력을 확인한다.

NoteBase

자료 읽는 순서

먼저 상자의 양 끝(제1사분위수, 제3사분위수)과 상자 내부 세로선(중앙값)의 위치를 가로축 눈금에 맞추어 읽는다. 학과 A는 제1사분위수 약 65점, 중앙값 약 73점, 제3사분위수 약 81점, 최솟값 약 64점, 최댓값 약 91점이다. 학과 B는 제1사분위수 약 66점, 중앙값 약 69점, 제3사분위수 약 78점, 최솟값 약 61점, 최댓값 약 85점이다. 다음으로 70점이라는 특정 값이 각 상자에서 어느 구간(하위 25% 미만, 25~50%, 50~75%, 75% 초과)에 위치하는지를 확인한다. 학과 A에서 70점은 제1사분위수(65)와 중앙값(73) 사이에 있으므로 70점 이하인 학생의 비율은 25%~50% 사이이다. 학과 B에서 70점은 중앙값(69)과 제3사분위수(78) 사이에 있으므로 70점 이하인 학생의 비율은 50%~75% 사이이다. 따라서 70점 이하 비율은 학과 A보다 학과 B가 더 높다.

선지별 해설

① 성적이 가장 좋은 학생은 두 학과의 최댓값을 비교하면 알 수 있다. 학과 A의 최댓값은 약 91점, 학과 B의 최댓값은 약 85점으로 학과 A가 더 높으므로 옳은 설명이다.
② 사분위범위는 제3사분위수에서 제1사분위수를 뺀 값, 즉 상자의 가로 길이이다. 학과 A는 약 81−65=16, 학과 B는 약 78−66=12로 학과 A의 사분위범위가 더 크므로 옳은 설명이다.
③ 학과 A에서 70점은 제1사분위수와 중앙값 사이(하위 약 25~50% 구간)에 위치하여 70점 이하 비율이 50% 미만이다. 반면 학과 B에서 70점은 중앙값과 제3사분위수 사이(하위 약 50~75% 구간)에 위치하여 70점 이하 비율이 50%를 넘는다. 따라서 70점 이하인 학생의 비율은 학과 A가 아니라 학과 B가 더 높으므로 틀린 설명이다 → 옳은 설명은 '학과 B가 학과 A보다 70점 이하인 학생의 비율이 높다'이다.
④ 제2사분위수는 중앙값을 의미한다. 학과 A의 중앙값은 약 73점, 학과 B의 중앙값은 약 69점으로 학과 A가 더 크므로 옳은 설명이다.

오답 포인트

상자그림에서 '어떤 점수 이하의 비율'을 판단할 때는 그 점수가 상자의 어느 구간에 속하는지를 봐야 하는데, 단순히 상자의 위치가 오른쪽에 있다고 해서(즉 전반적으로 성적이 높아 보인다고 해서) 특정 점수 이하 비율도 낮을 것이라고 성급히 판단하면 안 된다. 학과 A는 전반적으로 학과 B보다 성적이 높은 분포이지만, 70점이라는 특정 기준점에서는 오히려 학과 A의 제1사분위수와 중앙값 사이에 걸쳐 있어 70점 이하 비율이 학과 B보다 낮다는 점이 이 문제의 함정이다.

다시 나오면 이렇게 판단

특정 점수 X 이하(또는 이상) 비율을 묻는 선지가 나오면, 먼저 그 학과의 상자그림에서 X가 최솟값·제1사분위수·중앙값·제3사분위수·최댓값 중 어느 두 값 사이에 있는지 확인한다. 최솟값과 제1사분위수 사이면 0~25%, 제1사분위수와 중앙값 사이면 25~50%, 중앙값과 제3사분위수 사이면 50~75%, 제3사분위수와 최댓값 사이면 75~100%로 어림한 뒤 두 집단을 비교한다.

NoteBase

암기 포인트

사분위범위(IQR)=제3사분위수−제1사분위수(상자의 가로 길이), 제2사분위수=중앙값(상자 내부 세로선), 특정 점수 이하 비율은 그 점수가 속한 사분위 구간(0~25/25~50/50~75/75~100%)으로 판단한다.

NoteBase

문 2. 다음은 표본의 크기가 17인 어느 자료에 대하여 단순선형회귀모형 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ i = 1, 2, \cdots, 17$을 적용하여 얻은 분산분석표의 일부이다. 이에 대한 설명으로 옳지 않은 것은?

요인	제곱합	자유도	평균제곱	$F-$값	$p-$값
회귀	20			(가)	0.0064
오차(잔차)			(나)		
합계	50				

① Y의 총변동 중 40%는 X로 설명할 수 있다.

② (가)는 $\frac{32}{3}$이다.

③ (나)는 2이다.

④ 유의수준(significance level) 5%에서 X가 Y에 영향을 준다고 할 수 있다.

정답 2번

출제 포인트

단순선형회귀모형의 분산분석표(ANOVA table)에서 제곱합만 주어졌을 때 자유도·평균제곱·F값을 직접 계산하고, 이를 이용해 결정계수와 유의성 검정 결과를 해석할 수 있는지를 묻는 문제이다.

조건 정리

표본크기는 $n = 17$이고 단순선형회귀이므로 설명변수는 X 하나이다. 회귀(설명) 제곱합 $SSR = 20$, 총제곱합 $SST = 50$이 주어져 있고, 오차(잔차)의 평균제곱이 (나), 회귀의 F값이 (가)이다. 회귀의 p-값은 0.0064이다.

사용할 공식

총자유도는 $n - 1$, 회귀자유도는 설명변수 개수인 1, 오차자유도는 총자유도에서 회귀자유도를 뺀 값이다. 각 평균제곱은 $MS = \frac{SS}{df}$이며, F값은 $F = \frac{MSR}{MSE}$이다. 결정계수는 $R^2 = \frac{SSR}{SST}$이다.

계산 과정

총자유도는 $17 - 1 = 16$이고 회귀자유도는 1이므로 오차자유도는 $16 - 1 = 15$이다. 오차제곱합은 $SSE = SST - SSR = 50 - 20 = 30$이므로 오차의 평균제곱(나)은 $MSE = \frac{30}{15} = 2$이다. 회귀의 평균제곱은 $MSR = \frac{20}{1} = 20$이므로 (가)는 $F = \frac{MSR}{MSE} = \frac{20}{2} = 10$이다. 한편 결정계수는 $R^2 = \frac{20}{50} = 0.4$로 40%에 해당한다.

NoteBase

선지 판단

① Y의 총변동 중 40%는 X로 설명할 수 있다. → O, 결정계수 $R^2 = SSR/SST = 20/50 = 0.4$로 총변동의 40%를 X가 설명하므로 옳은 설명이다.
② (가)는 32/3이다. → X, 계산 결과 (가)는 F값으로 $F = MSR/MSE = 20/2 = 10$이다. 32/3은 실제 계산값과 다르므로 옳지 않은 설명이며 이것이 정답이다.
③ (나)는 2이다. → O, (나)는 오차의 평균제곱이며 MSE=SSE/df_오차=30/15=2로 계산되므로 옳은 설명이다.
④ 유의수준 5%에서 X가 Y에 영향을 준다고 할 수 있다. → O, 회귀의 p-값이 0.0064로 유의수준 0.05보다 작으므로 귀무가설(회귀계수=0, 즉 X가 Y에 영향을 주지 않는다)을 기각하고 X가 Y에 영향을 준다고 판단할 수 있다.

오답 포인트

이 문제는 표에 빈칸이 많아 자유도부터 순서대로 채워야 하는데, 오차자유도를 구하지 않고 바로 오차제곱합을 오차자유도가 아닌 다른 값으로 나누면 (나)와 (가)가 모두 틀어진다. 또한 F값과 p-값을 혼동해 p-값이 작다는 사실만으로 F값의 크기를 임의로 추정하면 32/3 같은 오답에 속을 수 있으므로 반드시 공식에 따라 순서대로 계산해야 한다.

다시 나오면 이렇게 판단

분산분석표 문제는 총자유도(=n-1) → 회귀자유도(=설명변수 개수) → 오차자유도(=총자유도-회귀자유도) 순으로 자유도를 먼저 채우고, 그다음 제곱합의 빈칸(합계=회귀+오차)을 채운 뒤 각 평균제곱과 F값을 계산하는 순서로 접근한다. 결정계수는 항상 SSR/SST로 구한다는 것도 함께 확인한다.

암기 포인트

단순선형회귀에서 회귀자유도는 항상 1이고, 오차자유도는 $n-2$이다(총자유도 $n-1$에서 회귀자유도 1을 뺀 값과 같다). $MS = SS/df$, $F = MSR/MSE$, $R^2 = SSR/SST$의 세 공식은 세트로 함께 외워 둔다.

NoteBase

문 3. 이산확률변수 X의 확률분포(probability distribution)는 다음과 같다. $E(X)=1.6$일 때, 상수 c_1과 c_2의 값은?

x	-1	0	1	2	3	4	5	계
$P(X=x)$	0.05	c_1	0.30	c_2	0.10	0.10	0.05	1

① c_1: 0.25, c_2: 0.15

② c_1: 0.2, c_2: 0.2

③ c_1: 0.3, c_2: 0.1

④ c_1: -0.2, c_2: 0.6

정답 2번

출제 포인트

이산확률변수의 확률질량함수는 모든 확률의 합이 1이 되어야 하고, 기댓값 $E(X)=\sum xP(X=x)$로 정의된다는 두 조건을 연립방정식으로 세워 미지수 c_1, c_2를 구하는 문제이다.

조건 정리

확률표는 $x=-1,0,1,2,3,4,5$에 대해 각각 $0.05,\ c_1,\ 0.30,\ c_2,\ 0.10,\ 0.10,\ 0.05$이고 합은 1이다. 추가로 $E(X)=1.6$이라는 조건이 주어져 있으므로 미지수 두 개(c_1, c_2)에 방정식 두 개가 생겨 값을 유일하게 결정할 수 있다.

사용할 공식

전확률의 합 조건: $\sum_x P(X=x)=1$.
기댓값 정의: $E(X)=\sum_x xP(X=x)$.

계산 과정

① 전확률 합 조건: $0.05+c_1+0.30+c_2+0.10+0.10+0.05=1$이므로 $c_1+c_2=1-0.60=0.40$.

② 기댓값 조건: $E(X)=(-1)(0.05)+0\cdot c_1+1(0.30)+2c_2+3(0.10)+4(0.10)+5(0.05)$
$=-0.05+0.30+2c_2+0.30+0.40+0.25=1.20+2c_2$.
이 값이 1.6이므로 $1.20+2c_2=1.6$, 즉 $2c_2=0.4$에서 $c_2=0.2$이다.

③ $c_1+c_2=0.40$에 $c_2=0.2$를 대입하면 $c_1=0.40-0.2=0.2$이다.

따라서 $c_1=0.2,\ c_2=0.2$이다.

NoteBase

선지 판단

① $c_1 = 0.25,\ c_2 = 0.15 \rightarrow$ 이 값을 대입하면 합은 $0.25 + 0.15 = 0.40$으로 전확률 조건은 맞지만, 기댓값을 계산하면 $1.20 + 2(0.15) = 1.50 \neq 1.6$이므로 틀린 선지이다. 옳지 않음.
② $c_1 = 0.2,\ c_2 = 0.2 \rightarrow$ 합 $0.2 + 0.2 = 0.40$이고 기댓값은 $1.20 + 2(0.2) = 1.60$으로 두 조건을 모두 만족한다. 옳음.
③ $c_1 = 0.3,\ c_2 = 0.1 \rightarrow$ 합은 0.40으로 맞지만 기댓값은 $1.20 + 2(0.1) = 1.40 \neq 1.6$이므로 틀린 선지이다. 옳지 않음.
④ $c_1 = -0.2,\ c_2 = 0.6 \rightarrow$ 합은 0.40으로 맞지만 확률은 음수가 될 수 없으므로($c_1 = -0.2 < 0$) 애초에 확률분포의 조건을 위반하며, 기댓값도 $1.20 + 2(0.6) = 2.40 \neq 1.6$이다. 옳지 않음.

오답 포인트

이 문제에서 자주 하는 실수는 전확률 합 조건 하나만 세우고 $c_1 + c_2 = 0.4$를 만족하는 선지 중 아무거나 고르는 것이다. 그러나 선지 ①③④ 모두 합은 0.40으로 동일하게 맞아떨어지므로, 반드시 기댓값 조건까지 함께 확인해야 유일한 정답을 가려낼 수 있다. 또한 $x = 0$에 해당하는 c_1은 기댓값 계산에서 $0 \times c_1 = 0$이 되어 사라진다는 점을 놓치면 계산 과정에서 혼동이 생길 수 있다.

다시 나오면 이렇게 판단

미지수가 포함된 확률분포표 문제는 먼저 전확률 합이 1이라는 식을 세우고, 기댓값(또는 분산 등 추가 조건)이 주어지면 반드시 두 번째 방정식으로 활용해 연립한다. 미지수가 2개면 조건도 2개가 필요하다는 점을 기준으로 삼아 식을 세운다.

암기 포인트

확률분포표의 두 가지 기본 조건, 즉 $\sum P(X = x) = 1$과 $E(X) = \sum xP(X = x)$를 항상 함께 확인하며, $x = 0$인 항은 기댓값 계산에서 자동으로 0이 된다는 점과 확률값은 음수가 될 수 없다는 점을 기억해 둔다.

NoteBase

문 4. 어떤 회사에서 생산한 건전지의 수명은 평균이 140시간, 표준편차가 20시간인 정규분포를 따른다고 한다. 가설 $H_0 : \mu \le 140$ 대 $H_1 : \mu > 140$을 검정하기 위해 이 회사에서 100개의 건전지를 임의추출하여 조사한 결과 수명의 표본평균이 150시간이었다. 이에 대한 설명으로 옳지 않은 것은? (단, μ는 모평균이고, $z_{0.025} = 1.96$, $z_{0.05} = 1.645$이다)

① 검정통계량의 값은 5이다.

② 위 가설의 검정은 단측검정이다.

③ 위 가설에 대한 유의확률(p-값)은 0.05보다 작다.

④ 유의수준 5%에서 귀무가설을 기각할 수 없다.

정답 4번

출제 포인트

모평균에 대한 단측 가설검정에서 검정통계량을 계산하고, 그 값을 기각역(임계값)과 비교하여 귀무가설의 기각 여부와 유의확률의 범위를 올바르게 판단할 수 있는지를 묻는다.

조건 정리

모표준편차 $\sigma = 20$, 표본크기 $n = 100$, 표본평균 $\bar{x} = 150$, 귀무가설의 경계값 $\mu_0 = 140$이다. 가설은 H_0:μ≤140 대 $H_1 : \mu > 140$으로 오른쪽 단측검정이며, 유의수준 5%에서의 임계값은 $z_{0.05} = 1.645$이다.

사용할 공식

모표준편차가 알려진 경우의 검정통계량은 $Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$이며, 오른쪽 단측검정에서는 Z가 임계값 $z_{0.05}$보다 크면 귀무가설을 기각한다.

계산 과정

표준오차는 $\sigma/\sqrt{n} = 20/\sqrt{100} = 20/10 = 2$이다. 검정통계량은 $Z = \frac{150 - 140}{2} = \frac{10}{2} = 5$이다. 유의수준 5%에서의 기각역 경계값 $z_{0.05} = 1.645$와 비교하면 $5 > 1.645$이므로 검정통계량이 기각역에 포함되어 귀무가설을 기각한다. 이때 유의확률은 $P(Z > 5)$로서 1.96이나 1.645에 대응하는 확률(각각 0.025, 0.05)보다 한참 작은, 사실상 0에 가까운 매우 작은 값이 되어 0.05보다 작다.

NoteBase

선지 판단

① 검정통계량의 값은 5이다 → 옳다. $Z = (150 - 140)/(20/\sqrt{100}) = 5$로 계산되어 문항 설명과 일치한다.
② 위 가설의 검정은 단측검정이다 → 옳다. $H_1 : \mu > 140$과 같이 한쪽 방향만을 대립가설로 설정하였으므로 오른쪽 단측검정이다.
③ 위 가설에 대한 유의확률(p-값)은 0.05보다 작다 → 옳다. 검정통계량 5는 유의수준 5%의 임계값 1.645는 물론 유의수준 2.5%의 임계값 1.96보다도 훨씬 크므로, 이에 대응하는 유의확률은 0.05보다 훨씬 작다.
④ 유의수준 5%에서 귀무가설을 기각할 수 없다 → 틀린 설명 → 올바른 설명: 검정통계량 5가 임계값 1.645보다 크므로 귀무가설은 기각된다. 즉 유의수준 5%에서 귀무가설을 기각할 수 있다.

오답 포인트

검정통계량의 값이 임계값보다 크게 나왔음에도 기각 여부를 반대로 판단하는 실수가 흔하다. 오른쪽 단측검정에서는 검정통계량이 임계값보다 '크면' 기각, '작거나 같으면' 채택(정확히는 기각하지 못함)이라는 방향을 혼동하지 않아야 한다. 또한 유의확률과 유의수준을 비교할 때, 유의확률이 유의수준보다 작으면 기각한다는 규칙과 검정통계량과 임계값을 비교하는 규칙이 서로 같은 결론을 내야 한다는 점도 함께 확인해야 한다.

다시 나오면 이렇게 판단

먼저 가설의 방향(단측인지 양측인지, 어느 쪽 단측인지)을 확인한 뒤 검정통계량 $Z = (\bar{x} - \mu_0)/(\sigma/\sqrt{n})$을 계산한다. 그 값을 주어진 임계값(예: $z_{0.05} = 1.645$, $z_{0.025} = 1.96$)과 직접 비교하여 기각역 포함 여부를 판단하고, 이 결론이 유의확률과 유의수준을 비교한 결론과 반드시 일치하는지 검산한다.

암기 포인트

오른쪽 단측검정 기각 조건: 검정통계량 $Z >$임계값이면 귀무가설 기각. 유의수준 5%의 오른쪽 단측 임계값은 1.645, 양측(또는 유의수준 2.5%)의 임계값은 1.96이다. 검정통계량이 임계값보다 클수록 유의확률(p-값)은 더 작아진다.

NoteBase

문 5. A회사 휴대전화기의 결함률은 0.01, B회사는 0.02, C회사는 0.03, D회사는 0.02라고 한다. 각 회사별 휴대전화기의 시장 점유율은 A회사는 0.4, B회사는 0.3, C회사는 0.2, D회사는 0.1이다. 어떤 휴대전화기가 결함을 보였을 때, 이 휴대전화기가 B회사 또는 C회사의 제품일 확률은?

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

정답 2번

출제 포인트

결함이 발견되었다는 조건 아래에서 특정 회사(들) 제품일 확률을 구하는 문제로, 베이즈 정리(조건부 확률)를 적용하는 유형이다. 시장 점유율은 사전확률, 회사별 결함률은 조건부확률에 해당하며, 이를 결합해 결함 발생이라는 사건의 전체확률과 그 안에서 B 또는 C가 차지하는 몫을 구해야 한다.

조건 정리

시장 점유율(사전확률): $P(A) = 0.4,\ P(B) = 0.3,\ P(C) = 0.2,\ P(D) = 0.1$. 회사별 결함률(조건부확률): P(결함 A)=0.01, P(결함 B)=0.02, P(결함 C)=0.03, P(결함 D)=0.02. 구해야 하는 값은 결함이 발생한 휴대전화기가 B회사 또는 C회사 제품일 확률, 즉 P(B C 결함)이다.

사용할 공식

전체확률의 법칙: P(결함) = Σ_i P(i) P(결함 i) (단, $i = A, B, C, D$). 베이즈 정리: P(B C 결함) = P(B)P(결함 B) + P(C)P(결함 C)P(결함).

계산 과정

각 회사가 결함 제품을 만들 결합확률을 구하면 다음과 같다. A: $0.4 \times 0.01 = 0.004$. B: $0.3 \times 0.02 = 0.006$. C: $0.2 \times 0.03 = 0.006$. D: $0.1 \times 0.02 = 0.002$. 이를 모두 더하면 결함이 발생할 전체확률은 P(결함)=0.004+0.006+0.006+0.002=0.018이다. B 또는 C에서 결함이 발생할 확률은 $0.006 + 0.006 = 0.012$이므로, 구하는 조건부확률은 $\frac{0.012}{0.018} = \frac{2}{3}$이다.

NoteBase

선지 판단

① 1/3 → 틀린 표현. A와 D만의 몫인 $\frac{0.004 + 0.002}{0.018} = \frac{1}{3}$에 해당하는 값으로, 문제에서 묻는 B 또는 C의 몫이 아니라 A 또는 D의 몫이다.
② 2/3 → 옳은 표현. B와 C의 결합확률의 합 0.012를 전체확률 0.018로 나눈 값과 일치한다.
③ 1/4 → 틀린 표현. 단순히 점유율만으로 B와 C의 비중(0.3+0.2=0.5)을 어림잡거나 다른 값끼리 나눈 결과로, 결함률을 반영하지 않은 계산에서 나올 수 있는 오답이다.
④ 3/4 → 틀린 표현. ①의 여사건에 해당하는 값($1 - \frac{1}{3} = \frac{2}{3}$이 아니라 다른 조합으로 계산했을 때 나오는 값)으로, B 또는 C의 올바른 몫이 아니다.

오답 포인트

이 문제는 시장 점유율(사전확률)과 결함률(조건부확률)을 혼동하기 쉽다. 단순히 점유율만 더하거나(0.3+0.2=0.5) 결함률만 더하는 실수를 하면 안 되고, 반드시 두 값을 곱해 결합확률을 만든 뒤 전체확률로 나누어야 한다. 또한 분모는 전체 회사(A, B, C, D)의 결함 결합확률을 모두 더한 값이어야 하며, B와 C만의 결함확률을 분모로 삼는 실수도 흔하다.

다시 나오면 이렇게 판단

먼저 각 항목의 사전확률과 조건부확률을 표처럼 정리해 곱한 결합확률을 모두 구한다. 그 다음 조건(여기서는 결함 발생)에 해당하는 전체 결합확률의 합을 분모로 놓고, 문제에서 묻는 대상(B 또는 C)의 결합확률의 합을 분자로 놓아 나누면 실수 없이 답을 구할 수 있다.

암기 포인트

베이즈 정리 계산 순서: (1) 사전확률과 조건부확률을 곱해 각 항목의 결합확률을 만든다. (2) 모든 결합확률을 더해 분모(전체확률)를 만든다. (3) 묻는 대상의 결합확률 합을 분자로 하여 나눈다. 이 3단계를 순서대로 적용하면 헷갈리지 않는다.

NoteBase

문 6. 다음의 두 자료 A, B에 대한 각각의 표본상관계수는?

자료 A	(1, 0.1), (2, 0.2), (3, 0.3), (4, 0.4), (5, 0.5)
자료 B	(1, 0.01), (2, 0.02), (3, 0.03), (4, 0.04), (5, 0.05)

① A: 0.1, B: 0.01

② A: 0.1, B: 1

③ A: 1, B: 0.1

④ A: 1, B: 1

정답 4번

출제 포인트

표본상관계수는 두 변수 사이의 선형관계의 방향과 강도를 나타내는 척도이며, 각 변수에 상수를 곱하거나 더하는 선형변환을 하더라도 그 값이 변하지 않는다는 성질(척도·이동 불변성)을 이해하는지 확인하는 문항이다.

조건 정리

자료 A는 $x = 1, 2, 3, 4, 5$에 대해 $y = 0.1x$로 주어지며, 자료 B는 같은 x에 대해 $y = 0.01x$로 주어진다. 두 자료 모두 x값이 커질수록 y값이 정확히 비례해서 커지는 완전한 직선 관계이다.

사용할 공식

표본상관계수는 $$r=\dfrac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum (x_i-\bar{x})^2}\sqrt{\sum (y_i-\bar{y})^2}}$$로 정의된다. 만약 $y_i = cx_i$ ($c > 0$인 상수)의 관계가 성립하면 $\bar{y} = c\bar{x}$이므로 $y_i - \bar{y} = c(x_i - \bar{x})$가 되고, 이를 대입하면 $$r=\dfrac{\sum c(x_i-\bar{x})^2}{\sqrt{\sum (x_i-\bar{x})^2}\sqrt{c^2\sum (x_i-\bar{x})^2}}=\dfrac{c\sum (x_i-\bar{x})^2}{|c|\sum (x_i-\bar{x})^2}=1$$이 되어, y가 x의 양의 상수배이기만 하면 상관계수는 항상 1이다.

계산 과정

자료 A는 $y = 0.1x$로 x의 양의 상수배($c = 0.1$)이므로 위 공식에 의해 표본상관계수는 1이다. 자료 B도 $y = 0.01x$로 역시 x의 양의 상수배($c = 0.01$)이므로 표본상관계수는 마찬가지로 1이다. 즉 두 자료의 y값 크기(0.1 단위인지 0.01 단위인지)는 상관계수 값에 아무 영향을 주지 않으며, 오직 x와 y가 완전한 직선 관계라는 사실만이 상관계수를 결정한다.

NoteBase

선지 판단

① A: 0.1, B: 0.01 → X, 상관계수를 y값의 크기(기울기)와 혼동한 오답이다.
② A: 0.1, B: 1 → X, A만 기울기 값을 상관계수로 착각하고 B는 맞게 처리한 형태로 일관성이 없는 오답이다.
③ A: 1, B: 0.1 → X, A는 맞게 처리했지만 B의 기울기(0.01)를 상관계수로 착각한 오답이다.
④ A: 1, B: 1 → O, 두 자료 모두 x와 y가 완전한 양의 직선 관계이므로 표본상관계수가 각각 1로 같다.

오답 포인트

이 문항에서 가장 흔한 착각은 자료표에 나온 y값의 배율(0.1배, 0.01배)을 그대로 상관계수 값으로 읽어버리는 것이다. 상관계수는 두 변수의 '값 크기'가 아니라 '관계의 형태(직선인지 아닌지, 방향이 양인지 음인지)'만을 측정하므로, $y = cx$ 꼴의 완전한 비례 관계라면 c가 얼마이든 상관계수는 항상 1이 된다는 점을 놓치면 안 된다.

다시 나오면 이렇게 판단

두 변수 사이에 $y = a + bx$ 형태의 완전한 직선 관계가 보이면, b의 부호만 확인한다. $b > 0$이면 상관계수는 무조건 1, $b < 0$이면 무조건 -1이며, a나 b의 구체적인 크기는 상관계수 값과 무관하다는 점만 확인하면 바로 답을 낼 수 있다.

암기 포인트

표본상관계수는 척도·이동 변환에 불변이다: $y = a + bx$이면 $r = 1 (b > 0)$ 또는 $r = -1 (b < 0)$로, a, b의 절대적 크기와는 무관하게 오직 완전한 직선 관계 여부와 기울기의 부호만으로 결정된다.

NoteBase

문 7. 체비셰프 부등식(Chebyshev's inequality)은 "확률변수 X의 평균이 μ, 분산이 $\sigma^2 < \infty$일 때, 0보다 큰 모든 k에 대해 $P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$"이다. $E(X) = 17$이고 $E(X^2) = 298$일 때, 체비셰프 부등식을 이용한 $P(11 < X < 23)$의 하한값은?

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

정답 4번

출제 포인트

체비셰프 부등식은 분포의 형태를 몰라도 평균과 분산만으로 확률의 하한(또는 상한)을 구할 수 있게 해주는 부등식이다. 이 문제는 $E(X)$와 $E(X^2)$로부터 분산 σ^2을 구한 뒤, $P(11 < X < 23)$을 $P(|X-\mu| < k\sigma)$ 꼴로 바꾸어 체비셰프 부등식의 하한값을 계산할 수 있는지를 묻는다.

조건 정리

$E(X) = \mu = 17$, $E(X^2) = 298$이므로 $\sigma^2 = E(X^2) - \mu^2 = 298 - 17^2 = 298 - 289 = 9$, 따라서 $\sigma = 3$이다. 구간 $11 < X < 23$은 $17 - 6 < X < 17 + 6$, 즉 $|X - 17| < 6$이다. $6 = k\sigma = 3k$이므로 $k = 2$이다.

사용할 공식

체비셰프 부등식 $P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$의 여사건을 취하면 $P(|X-\mu| < k\sigma) \geq 1 - \frac{1}{k^2}$이 되며, 이것이 곧 $P(11 < X < 23)$의 하한값이다.

계산 과정

$k = 2$를 대입하면 $P(|X - 17| < 6) \geq 1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4}$이다. 따라서 $P(11 < X < 23)$의 하한값은 $\frac{3}{4}$이다.

선지 판단

① $\frac{1}{4}$: $1 - \frac{1}{k^2}$이 아니라 $\frac{1}{k^2}$ 자체(상한 쪽 값)를 답으로 착각한 경우에 나오는 값이므로 틀렸다.
② $\frac{1}{3}$: k를 3으로 잘못 놓거나 계산을 잘못한 경우에 나오는 값으로, 조건과 맞지 않아 틀렸다.
③ $\frac{2}{3}$: $k = 2$를 제대로 구했더라도 $1 - \frac{1}{k^2}$ 계산에서 k^2 대신 k를 사용해 $1 - \frac{1}{2} = \frac{1}{2}$가 아닌 다른 실수 계산으로 나올 수 있는 값이나, 정확한 계산 결과와는 다르므로 틀렸다.
④ $\frac{3}{4}$: $\sigma = 3$, $k = 2$로 정확히 구한 뒤 $1 - \frac{1}{2^2} = \frac{3}{4}$로 계산한 값과 일치하므로 옳다.

NoteBase

오답 포인트

이 문제에서 자주 하는 실수는 두 가지이다. 첫째, $\sigma^2 = E(X^2) - \mu^2$ 공식 대신 $E(X^2) - E(X)$ 등으로 잘못 계산해 σ 값을 틀리게 구하는 것이다. 둘째, 체비셰프 부등식이 원래 $P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$ 형태이므로, 구하려는 사건이 여사건 관계($<$)임을 놓치고 $\frac{1}{k^2}$를 그대로 답으로 쓰는 것이다. 하한값을 구하는 문제이므로 반드시 $1 - \frac{1}{k^2}$로 바꾸어야 한다.

다시 나오면 이렇게 판단

구간이 주어지면 먼저 그 구간의 중심이 평균과 일치하는지 확인하고, 중심에서 양쪽으로 벌어진 폭을 $k\sigma$로 놓아 k를 구한다. 이후 문제가 "이하/미만(부등호에 등호 없는 구간 안쪽)"의 확률 하한을 묻는지, 아니면 "바깥쪽(꼬리)" 확률의 상한을 묻는지를 구분해 $1 - \frac{1}{k^2}$ 또는 $\frac{1}{k^2}$ 중 맞는 것을 적용한다.

암기 포인트

체비셰프 부등식: $P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$, 여사건으로 $P(|X-\mu| < k\sigma) \geq 1 - \frac{1}{k^2}$. 분산 계산: $\sigma^2 = E(X^2) - \{E(X)\}^2$.

NoteBase

문 8. 다음은 도시 A와 도시 B에서 지하철 이용 비율에 차이가 있는지 알아보기 위해 두 지역에서 각각 5000명, 3000명을 임의추출하여 조사한 결과이다. 도시 A의 지하철 이용 비율이 도시 B보다 높은지 검정하기 위한 Z 검정통계량의 값과 유의수준 5%에서의 임계치(critical value)는?

도시	출근자 표본 수	지하철 이용 출근자 수
A	5000	4000
B	3000	2000

① Z 검정통계량의 값: $\frac{\frac{4}{5}-\frac{2}{3}}{\sqrt{\frac{3}{4}(1-\frac{3}{4})(\frac{1}{5000}+\frac{1}{3000})}}$, 임계치: $z_{0.05}=1.645$
② Z 검정통계량의 값: $\frac{\frac{4}{5}-\frac{2}{3}}{\sqrt{\frac{4}{5}(1-\frac{4}{5})(\frac{1}{5000}+\frac{1}{3000})}}$, 임계치: $z_{0.05}=1.645$
③ Z 검정통계량의 값: $\frac{\frac{4}{5}-\frac{2}{3}}{\sqrt{\frac{3}{4}(1-\frac{3}{4})(\frac{1}{5000}+\frac{1}{3000})}}$, 임계치: $z_{0.025}=1.96$
④ Z 검정통계량의 값: $\frac{\frac{4}{5}-\frac{2}{3}}{\sqrt{\frac{4}{5}(1-\frac{4}{5})(\frac{1}{5000}+\frac{1}{3000})}}$, 임계치: $z_{0.025}=1.96$

정답 1번

출제 포인트

두 모비율의 차이에 대한 단측 검정을 다룬다. 검정통계량의 분모에 들어갈 표준오차를 각 집단의 개별 표본비율로 계산하는지 아니면 귀무가설(두 모비율이 같다) 하의 합동(pooled) 표본비율로 계산하는지, 그리고 대립가설의 방향성(한쪽이 더 높은지)에 따라 임계치를 양측값 대신 단측값으로 써야 하는지를 동시에 확인하는 문제이다.

조건 정리

도시 A: 표본수 5000, 지하철 이용자수 4000이므로 표본비율 $\hat{p}_A = 4000/5000 = 4/5$.
도시 B: 표본수 3000, 지하철 이용자수 2000이므로 표본비율 $\hat{p}_B = 2000/3000 = 2/3$.
귀무가설은 $H_0 : p_A = p_B$, 대립가설은 "A의 이용 비율이 B보다 높다"는 $H_1 : p_A > p_B$로 한쪽 방향만 검정하는 단측검정이다. 유의수준은 5%이다.

사용할 공식

귀무가설이 $p_A = p_B$라고 전제하므로, 검정통계량의 분산을 추정할 때는 두 표본을 합쳐서 하나의 공통 비율(합동표본비율) $\hat{p}$를 사용한다.

$$\hat{p} = \frac{x_A + x_B}{n_A + n_B}$$

$$Z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_A}+\frac{1}{n_B})}}$$

대립가설이 한쪽 방향(A가 B보다 높다)이므로 이는 단측검정이며, 유의수준 5%에서의 임계치는 $z_{0.05} = 1.645$를 사용한다. 양측검정에서 쓰는 $z_{0.025} = 1.96$은 여기서는 해당되지 않는다.

NoteBase

계산 과정

합동표본비율은 $\hat{p} = (4000 + 2000)/(5000 + 3000) = 6000/8000 = 3/4$이다.
따라서 검정통계량의 분모는 $\sqrt{(3/4)(1 - 3/4)(1/5000 + 1/3000)}$의 형태가 되어야 한다.
분자는 두 표본비율의 차이 그대로 $4/5 - 2/3$이다.
대립가설이 "A가 B보다 높다"는 단측 방향이므로 유의수준 5%에서의 임계치는 $z_{0.05} = 1.645$이다.

선지 판단

① 분모에 합동표본비율 3/4을 사용하고 임계치로 단측값 $z_{0.05} = 1.645$를 사용한다 → 옳다.
② 분모에 도시 A의 개별 표본비율 4/5를 그대로 사용했다 → 틀린 표현. 귀무가설이 두 비율이 같다고 전제하므로 개별 비율이 아니라 합동표본비율 3/4을 써야 한다.
③ 분모는 합동표본비율 3/4으로 옳게 썼지만 임계치로 양측값 $z_{0.025} = 1.96$을 사용했다 → 틀린 표현. 대립가설이 한쪽 방향(A > B)이므로 단측 임계치 $z_{0.05} = 1.645$를 써야 한다.
④ 분모에 개별 표본비율 4/5를 사용하고 임계치도 양측값 1.96을 사용했다 → 두 가지 모두 틀린 표현이다.

오답 포인트

가장 흔한 실수는 두 가지이다. 첫째, 검정통계량의 분모를 만들 때 귀무가설이 두 모비율이 같다고 가정하고 있음을 잊고 각 집단의 개별 표본비율(4/5, 2/3)을 그대로 대입하는 것인데, 이 경우 표본을 합쳐서 만든 합동표본비율 3/4을 써야 한다. 둘째, 대립가설이 방향성을 가진 단측검정임을 놓치고 양측검정의 임계치인 1.96을 쓰는 것인데, "A가 B보다 높은지"라는 표현은 한쪽 방향만 보는 단측검정이므로 임계치는 1.645이어야 한다.

다시 나오면 이렇게 판단

두 모비율 차이 검정 문제를 보면 먼저 대립가설이 "차이가 있는지"(양측)인지 "한쪽이 더 큰지/작은지"(단측)인지부터 확인한다. 단측이면 5% 유의수준의 임계치는 1.645, 양측이면 1.96이다. 그 다음 분모의 표준오차 계산에서는 귀무가설이 두 비율이 같다고 가정하므로 반드시 두 표본을 합쳐 만든 합동표본비율을 사용하고, 각 집단의 개별 표본비율을 분모에 쓰지 않는다.

암기 포인트

두 모비율 차이 검정: 분모의 표준오차는 합동표본비율 $\hat{p} = (x_A + x_B)/(n_A + n_B)$로 계산한다. 대립가설이 한쪽 방향(단측)이면 5% 유의수준 임계치는 1.645, 양쪽 방향(양측)이면 1.96이다.

NoteBase

문 9. 다음과 같은 확률분포를 따르는 확률변수 X에 대하여 3개의 임의표본(random sample)으로 2, 2, 3을 얻었을 때, 모수 p에 대한 최대우도추정치(maximum likelihood estimate)는? (단, $0 < p < 1$이다)

x	1	2	3	계
$P(X=x)$	p^2	$2p(1-p)$	$(1-p)^2$	1

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

정답 2번

출제 포인트

이산확률분포에서 표본이 주어졌을 때 우도함수를 세우고 이를 최대화하는 모수의 최대우도추정치(MLE)를 구하는 문제이다. 확률변수 X는 1, 2, 3 세 값만 가지며 각 확률이 p^2, $2p(1-p)$, $(1-p)^2$로 주어져 있으므로, 이는 시행 횟수 2인 이항분포와 같은 구조를 가진 분포임을 알아채는 것이 핵심이다.

조건 정리

확률변수 X의 확률질량함수는 $P(X=1)=p^2$, $P(X=2)=2p(1-p)$, $P(X=3)=(1-p)^2$이다. 임의표본 3개로 2, 2, 3을 얻었으므로, 이 세 관측값이 서로 독립이라고 보고 우도함수를 구성한다. 관측값 중 $X=2$가 두 번, $X=3$이 한 번 나왔다.

사용할 공식

독립인 표본 x_1, x_2, x_3에 대한 우도함수는 $L(p)=\prod_{i=1}^{3} P(X=x_i)$이다. 최대우도추정치는 로그우도함수 $\ln L(p)$를 p에 대해 미분하여 0이 되는 지점에서 구한다: $\frac{d}{dp}\ln L(p)=0$.

NoteBase

계산 과정

관측값이 2, 2, 3이므로 우도함수는 다음과 같다.
$L(p) = P(X=2) \cdot P(X=2) \cdot P(X=3) = [2p(1-p)]^2 \cdot (1-p)^2 = 4p^2(1-p)^4$
양변에 로그를 취하면
$\ln L(p) = \ln 4 + 2\ln p + 4\ln(1-p)$
이를 p에 대해 미분하면
$\frac{d}{dp}\ln L(p) = \frac{2}{p} - \frac{4}{1-p}$
이를 0으로 놓으면
$\frac{2}{p} = \frac{4}{1-p} \Rightarrow 2(1-p) = 4p \Rightarrow 2 - 2p = 4p \Rightarrow 6p = 2$
따라서 $p = \frac{1}{3}$이며, 이 값에서 $\ln L(p)$는 최댓값을 가지므로(이차미분 확인 시 $p = 1/3$ 근방에서 위로 볼록) 최대우도추정치는 $\frac{1}{3}$이다.

선지 판단

① 1/4 — X, 우도함수를 최대화하는 값이 아니다. 미분값을 0으로 만드는 p는 1/3이지 1/4가 아니다.
② 1/3 — O, 우도함수 $L(p) = 4p^2(1-p)^4$를 최대화하는 값으로 계산 결과와 일치한다.
③ 1/2 — X, p=1/2를 대입하면 $\ln L$의 미분값이 $2/(1/2) - 4/(1/2) = 4 - 8 = -4 \neq 0$이므로 최댓값이 아니다.
④ 2/3 — X, p=2/3을 대입하면 미분값이 $2/(2/3) - 4/(1/3) = 3 - 12 = -9 \neq 0$이므로 최댓값이 아니다.

오답 포인트

이 문제에서 자주 하는 실수는 관측값 2, 2, 3을 단순 평균 내어 $\bar{x}$를 구하거나, 표본비율처럼 성공 횟수를 표본 크기로 나누어 답을 내는 것이다. 하지만 이 분포는 확률변수의 값(1, 2, 3) 자체가 이항분포의 결과 개수(0, 1, 2)와 어긋나게 매겨져 있으므로, 반드시 각 관측값에 대응하는 확률식을 정확히 대입해 우도함수를 세운 뒤 미분해야 한다. 우도함수를 세울 때 지수를 잘못 세거나(예: $P(X=2)$가 두 번 곱해지는 것을 놓치는 경우) 로그미분 과정에서 부호를 실수하면 다른 값이 나올 수 있으므로 계산 과정을 검산하는 것이 중요하다.

다시 나오면 이렇게 판단

확률변수의 확률질량함수가 표로 주어지고 임의표본이 제시되면, 먼저 표본에 등장한 각 값의 확률식을 표에서 그대로 가져와 곱해 우도함수를 구성한다. 이후 로그를 취해 미분을 쉽게 만들고, 미분값을 0으로 두어 방정식을 풀면 최대우도추정치를 얻을 수 있다. 지수 형태의 확률식이 나오면 로그를 취하는 것이 계산을 단순화하는 표준적인 접근이다.

암기 포인트

최대우도추정법의 절차는 우도함수 구성 → 로그우도함수 변환 → 모수에 대해 미분 → 0으로 놓고 풀기의 순서로 고정되어 있다. 로그를 취하면 곱셈이 덧셈으로, 거듭제곱이 계수 곱셈으로 바뀌어 미분이 간단해진다는 점을 기억해 둔다.

NoteBase

문 10. 어느 학급의 영어 성적이 정규분포를 따른다고 할 때, 크기가 9인 임의표본에서 계산된 모평균의 95% 신뢰구간 길이는 2였다. 신뢰구간의 길이를 0.5 이하가 되도록 하기 위한 최소 표본크기는? (단, 모분산은 알려져 있다)

① 36

② 64

③ 100

④ 144

정답 4번

출제 포인트

모분산을 아는 정규모집단에서 모평균의 신뢰구간 길이가 표본크기와 어떤 관계를 갖는지를 이용해, 원하는 신뢰구간 길이를 얻기 위한 최소 표본크기를 구하는 문제이다. 신뢰구간 길이가 표본크기의 제곱근에 반비례한다는 사실만 알면 표준편차 값을 직접 구하지 않고도 답을 낼 수 있다.

조건 정리

모분산을 아는 정규분포에서 표본크기 $n = 9$일 때 모평균의 95% 신뢰구간 길이가 2이다. 신뢰구간 길이를 0.5 이하로 줄이기 위한 최소 표본크기 n'을 구한다.

사용할 공식

모분산 σ^2을 아는 경우 모평균의 95% 신뢰구간은 $\bar{X} \pm z_{0.025}\frac{\sigma}{\sqrt{n}}$이므로 신뢰구간의 길이는 $L = 2z_{0.025}\frac{\sigma}{\sqrt{n}}$ 이다. 신뢰수준과 σ가 같을 때 두 상황의 길이 비는 $\frac{L'}{L} = \sqrt{\frac{n}{n'}}$ 로 표본크기에만 의존한다.

계산 과정

$n = 9$일 때 $L = 2$이고, 원하는 길이는 $L' = 0.5$이다. 길이 비 공식에 대입하면 $\frac{0.5}{2} = \sqrt{\frac{9}{n'}}$ 이므로 $\frac{1}{4} = \sqrt{\frac{9}{n'}}$ 양변을 제곱하면 $\frac{1}{16} = \frac{9}{n'}$ 따라서 $n' = 9 \times 16 = 144$ 이다.

선지 판단

① 36 → X: $n' = 36$이면 $\sqrt{9/36} = 1/2$이 되어 길이가 $2 \times (1/2) = 1$에 그쳐 목표 0.5를 만족하지 못한다.
② 64 → X: $\sqrt{9/64} = 3/8$이므로 길이가 $2 \times (3/8) = 0.75$로 여전히 0.5를 넘는다.
③ 100 → X: $\sqrt{9/100} = 3/10$이므로 길이가 $2 \times (3/10) = 0.6$으로 아직 0.5보다 크다.
④ 144 → O: $\sqrt{9/144} = 1/4$이므로 길이가 $2 \times (1/4) = 0.5$가 되어 목표치인 0.5 이하를 정확히 만족하는 최소 표본크기이다.

NoteBase

오답 포인트

신뢰구간 길이를 원래 길이의 몇 분의 1로 줄이려면 표본크기는 그 제곱만큼 늘어나야 한다는 점을 놓치면 표본크기를 단순 비례(4배)로 계산해 36 같은 오답을 고르게 된다. 길이를 1/4로 줄이려면 표본크기는 $4^2 = 16$ 배가 되어야 한다.

다시 나오면 이렇게 판단

표본크기 변화 문제에서는 z값과 σ를 굳이 따로 구하지 않고, 신뢰구간 길이가 $1/\sqrt{n}$에 비례한다는 관계식 $L'/L = \sqrt{n/n'}$만 세워 비율로 풀면 계산이 빠르고 실수가 줄어든다.

암기 포인트

모분산을 아는 정규모집단의 신뢰구간 길이는 $L = 2z_{\alpha/2}\sigma/\sqrt{n}$이며, 길이를 k배로 줄이려면 표본크기는 $1/k^2$ 배가 아니라 k^2배로 늘려야 한다는 관계를 기억해 둔다.

NoteBase

문 11. 어느 대학교에서 재학생의 의식구조를 조사하기 위해 전체 학과 중 10개 학과를 임의추출한 후, 추출된 학과의 모든 학생들을 표본으로 추출하였다. 여기서 사용한 표본추출방법은? (단, 학과 간의 의식구조는 동질적이고 학과 내에서는 이질적이다)

① 군집(집락)추출법(cluster sampling)

② 단순임의추출법(simple random sampling)

③ 계통추출법(systematic sampling)

④ 층화추출법(stratified sampling)

정답 1번

출제 포인트

표본추출방법 중 군집(집락)추출법의 절차와 적용 조건을 구분할 수 있는지를 묻는다. 핵심은 '어떤 단위를 임의추출하는가'와 '군집 내부·군집 간의 동질성·이질성 조건'이다.

핵심 개념

군집(집락)추출법(cluster sampling)은 모집단을 여러 군집(학과 등)으로 나눈 뒤 일부 군집을 임의추출하고, 선택된 군집에 속한 모든 개체를 전수 조사하는 방법이다. 이 방법이 효율적이려면 군집 간에는 서로 비슷한 특성을 가지고(군집 간 동질적) 군집 내부에는 다양한 특성이 섞여 있어야(군집 내 이질적) 한다. 반대로 층화추출법(stratified sampling)은 모집단을 층으로 나눈 뒤 각 층에서 표본을 추출하는데, 이때는 층 내부가 동질적이고 층 간에 이질적이어야 효율적이다. 단순임의추출법은 모집단 전체에서 개체를 하나씩 임의로 뽑는 방법이고, 계통추출법은 모집단에 일련번호를 매긴 뒤 일정한 간격으로 표본을 추출하는 방법이다.

선지별 해설

① 군집(집락)추출법(cluster sampling): O. 전체 학과 중 10개 학과를 임의추출한 뒤 그 학과에 속한 모든 학생을 표본으로 삼았으므로 학과라는 군집을 단위로 추출한 것이다. 또한 학과 간 의식구조는 동질적이고 학과 내에서는 이질적이라는 조건은 군집추출법이 효율적으로 작동하는 전형적인 조건과 정확히 일치한다.
② 단순임의추출법(simple random sampling): X. 모집단(전체 재학생) 개개인을 대상으로 임의로 뽑는 방법이 아니라 학과 단위로 먼저 추출하였으므로 해당하지 않는다.
③ 계통추출법(systematic sampling): X. 일련번호를 매겨 일정 간격으로 추출하는 절차가 아니므로 해당하지 않는다.
④ 층화추출법(stratified sampling): X. 층화추출법은 층 내부가 동질적이고 층 간에 이질적일 때 사용하는 방법인데, 이 문제의 조건(학과 간 동질적, 학과 내 이질적)은 오히려 정반대이므로 층화추출법의 조건과 맞지 않는다.

NoteBase

오답 포인트

군집추출법과 층화추출법을 혼동하는 경우가 가장 많다. 두 방법 모두 모집단을 여러 집단으로 나눈다는 점은 같지만, 군집추출법은 나눈 집단(군집) 중 일부만 뽑아 그 안을 전수 조사하고, 층화추출법은 나눈 집단(층) 전부에서 일부씩 뽑는다는 점이 다르다. 또한 동질성·이질성 조건이 서로 반대라는 점을 놓치면 ④번을 답으로 고르는 실수를 하게 된다.

다시 나오면 이렇게 판단

먼저 '무엇을 임의추출했는가'를 확인한다. 개별 학생이 아니라 학과 같은 집단 단위를 임의추출하고 그 안을 통째로 조사했다면 군집추출법이다. 그다음 동질성·이질성 조건 문구를 확인하여 '집단 간 동질, 집단 내 이질'이면 군집추출법, '집단 내 동질, 집단 간 이질'이면 층화추출법이라는 대응 관계로 최종 확인한다.

암기 포인트

군집추출법: 집단을 뽑아 그 안을 전수 조사, 집단 간 동질·집단 내 이질일 때 효율적. 층화추출법: 모든 집단에서 일부씩 추출, 집단 내 동질·집단 간 이질일 때 효율적. 단순임의추출법: 개체별 임의추출. 계통추출법: 일정 간격 추출.

NoteBase

문 12. 어느 실험결과에 대해 일원배치 분산분석법(one-way analysis of variance)을 적용하여 얻은 분산분석표의 일부가 다음과 같을 때, 옳지 않은 것은?

요인	제곱합	자유도	평균제곱	$F-$값
처리	8	(나)	4	(라)
오차(잔차)	(가)	6	(다)	
합계	26			

① (가)는 18이다.

② (나)는 2이다.

③ (다)는 3이다.

④ (라)는 2이다.

정답 4번

출제 포인트

일원배치 분산분석표에서 처리(요인)와 오차(잔차)의 제곱합·자유도·평균제곱·F값 사이의 관계식을 이용해 표의 빈칸 (가)~(라)를 채우고, 그 값이 맞는지를 판단하는 문제이다. 분산분석표의 기본 구조를 정확히 알아야 빈칸을 순서대로 복원할 수 있다.

조건 정리

처리 제곱합 $SS_{Tr}=8$, 처리 평균제곱 $MS_{Tr}=4$, 오차 자유도 $df_E=6$, 총 제곱합 $SS_T=26$이 주어져 있다. 구해야 할 값은 오차 제곱합 (가), 처리 자유도 (나), 오차 평균제곱 (다), F값 (라)이다.

사용할 공식

평균제곱은 제곱합을 자유도로 나눈 값이다: $MS=\frac{SS}{df}$. 총 제곱합은 처리 제곱합과 오차 제곱합의 합이다: $SS_T=SS_{Tr}+SS_E$. F값은 처리 평균제곱을 오차 평균제곱으로 나눈 값이다: $F=\frac{MS_{Tr}}{MS_E}$.

계산 과정

(나) 처리 자유도: $MS_{Tr}=\frac{SS_{Tr}}{df_{Tr}}$이므로 $4=\frac{8}{df_{Tr}}$, 따라서 $df_{Tr}=2$이다.
(가) 오차 제곱합: $SS_T=SS_{Tr}+SS_E$이므로 $26=8+SS_E$, 따라서 $SS_E=18$이다.
(다) 오차 평균제곱: $MS_E=\frac{SS_E}{df_E}=\frac{18}{6}=3$이다.
(라) F값: $F=\frac{MS_{Tr}}{MS_E}=\frac{4}{3}$이다. 이는 2가 아니다.

NoteBase

선지 판단

① (가)는 18이다 → O. 계산 과정에서 $SS_E = 26 - 8 = 18$로 확인된다.
② (나)는 2이다 → O. $df_{Tr} = 8 \div 4 = 2$로 확인된다.
③ (다)는 3이다 → O. $MS_E = 18 \div 6 = 3$으로 확인된다.
④ (라)는 2이다 → X. $F = 4 \div 3 = \frac{4}{3}$이며 2가 아니다. → 틀린 표현 "(라)는 2이다" → 올바른 표현 "(라)는 $\frac{4}{3}$이다".

오답 포인트

F값을 처리 평균제곱과 오차 자유도를 나누는 식으로 잘못 계산하거나(예: $4 \div 2 = 2$), 오차 평균제곱을 구하지 않고 임의의 값으로 나누면 F값을 2로 잘못 계산하기 쉽다. F값은 반드시 두 평균제곱의 비, 즉 $MS_{Tr} \div MS_E$로 계산해야 하며, 이때 분모인 오차 평균제곱을 먼저 정확히 구하는 것이 핵심이다.

다시 나오면 이렇게 판단

분산분석표 문제는 항상 같은 순서로 접근한다. 먼저 평균제곱과 제곱합·자유도의 관계로 알려지지 않은 자유도나 제곱합을 구하고, 다음으로 총 제곱합에서 각 요인의 제곱합을 빼거나 더해 나머지 제곱합을 구하며, 마지막으로 F값은 항상 처리 평균제곱을 오차 평균제곱으로 나누어 계산한다는 순서를 지킨다.

암기 포인트

평균제곱=제곱합/자유도, 총 제곱합=처리 제곱합+오차 제곱합, F값=처리 평균제곱/오차 평균제곱. 이 세 관계식만 정확히 적용하면 분산분석표의 어떤 빈칸도 순서대로 채울 수 있다.

NoteBase

문 13. 중심극한정리(central limit theorem)에 대한 설명으로 옳은 것은?

① 이항분포에서 표본크기 n이 크고 성공확률 p가 작으면 근사적으로 모수가 $\lambda = np$인 포아송분포를 따른다.
② 임의의 모집단에서 표본크기 n이 충분히 커지면 표본평균 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$는 근사적으로 정규분포를 따른다.
③ 확률실험의 횟수가 증가하면 어떤 사건이 관측되는 비율은 그 사건의 확률로 수렴한다.
④ 표본평균 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$와 표본분산 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2$은 독립이다.

정답 2번

출제 포인트

중심극한정리(CLT)가 정확히 무엇을 주장하는지를 다른 극한정리(포아송근사, 대수의 법칙)나 표본통계량의 성질(표본평균과 표본분산의 독립성)과 구별할 수 있는지를 묻는다. 선지 각각이 실제로 CLT를 설명하는지, 아니면 이름만 비슷한 다른 정리를 설명하는지를 가려야 한다.

핵심 개념

중심극한정리는 모집단의 분포 형태와 관계없이(정규분포가 아니어도) 표본크기 n이 충분히 크면 표본평균 $\bar{X}$의 분포가 근사적으로 정규분포를 따른다는 정리이다. 즉 $\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$로 근사할 수 있다는 것이 핵심이며, '임의의 모집단'이라는 조건이 이 정리의 특징이다.
포아송근사(포아송 극한정리)는 이항분포에서 n이 크고 p가 작을 때(np가 일정할 때) 이항분포가 모수 $\lambda = np$인 포아송분포로 근사된다는 별개의 정리로, CLT와는 다른 정리이다.
대수의 법칙(큰 수의 법칙)은 시행 횟수가 늘어날수록 사건의 관측 비율(상대도수)이 그 사건의 확률로 수렴한다는 정리로, 이 역시 CLT와는 별개이다.
표본평균과 표본분산의 독립성은 모집단이 정규분포일 때에 한해 성립하는 성질(정규모집단에서의 표본분포 이론)로, 임의의 모집단에 대해 일반적으로 성립하지 않으며 CLT의 내용도 아니다.

NoteBase

선지별 해설

① X → 이항분포에서 n이 크고 p가 작을 때 모수 $\lambda = np$인 포아송분포로 근사되는 것은 포아송 극한정리(포아송근사)이지 중심극한정리가 아니다. 진술 자체는 통계학적으로 알려진 근사이지만 CLT에 대한 설명이 아니므로 정답이 될 수 없다.
② O → 모집단의 분포 형태에 관계없이(임의의 모집단) 표본크기 n이 충분히 커지면 표본평균 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$가 근사적으로 정규분포를 따른다는 것은 중심극한정리의 정확한 내용이다.
③ X → 확률실험 횟수가 증가할 때 어떤 사건의 관측 비율(상대도수)이 그 사건의 확률로 수렴한다는 것은 대수의 법칙(큰 수의 법칙)의 내용이며, 표본평균의 분포 형태를 다루는 중심극한정리와는 다른 정리이다.
④ X → 표본평균 $\bar{X}$와 표본분산 S^2의 독립성은 모집단이 정규분포를 따를 때에만 성립하는 성질이다. 이 선지는 모집단에 대한 아무런 조건도 제시하지 않은 채 일반적으로 독립이라고 서술하고 있어 옳지 않고, 설령 정규모집단이라는 전제를 추가하더라도 이는 중심극한정리의 내용이 아니라 정규모집단에서의 표본분포 성질일 뿐이다.

오답 포인트

①③④는 모두 통계학에서 실제로 다루는 정리·성질을 언급하고 있어 얼핏 그럴듯해 보이지만, 문제가 요구하는 것은 '중심극한정리'에 대한 설명이라는 점을 놓치면 안 된다. 이름이 비슷하거나 같은 극한 개념(근사, 수렴)을 다룬다고 해서 CLT와 동일시하면 안 되고, 각 정리가 실제로 무엇을 주장하는지를 정확히 매칭해야 한다.

다시 나오면 이렇게 판단

선지에 '표본평균', '정규분포로 근사', '임의의 모집단', 'n이 충분히 크면'이라는 표현이 함께 나오면 중심극한정리를 가리킬 가능성이 높다고 판단한다. 반대로 '포아송분포로 근사'가 나오면 포아송 극한정리, '관측 비율이 확률로 수렴'이 나오면 대수의 법칙, '정규모집단에서 독립'이 나오면 정규분포의 표본분포 성질임을 구분해서 소거한다.

암기 포인트

중심극한정리 = 임의의 모집단 + n이 충분히 크면 + 표본평균이 근사적으로 정규분포. 포아송근사(n크고 p작음) → 포아송분포, 대수의 법칙 → 관측 비율이 확률로 수렴, 표본평균·표본분산의 독립성 → 정규모집단이라는 전제가 있어야 성립한다는 점을 각각 구분하여 암기해 둔다.

NoteBase

문 14. 두 이산확률변수 X와 Y의 결합확률분포(joint probability distribution)가 다음과 같을 때, 이에 대한 설명으로 옳은 것만을 모두 고른 것은?

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

ㄱ. $E(X)$는 0이다.

ㄴ. $E(X \mid Y=1)$은 0이다.

ㄷ. X와 Y는 독립이다.

① ㄱ

② ㄱ, ㄴ

③ ㄴ, ㄷ

④ ㄱ, ㄴ, ㄷ

정답 2번

출제 포인트

결합확률분포표가 주어졌을 때 주변확률(marginal probability), 기댓값 $E(X)$, 조건부 기댓값 $E(X \mid Y=1)$, 독립성 여부를 표에서 직접 계산할 수 있는지를 확인하는 문제이다. 표에서 확률이 0이 아닌 칸만 골라내어 정확히 합산하는 것이 핵심이다.

NoteBase

조건 정리

표에서 확률이 0이 아닌 결합확률은 다음 네 칸이다.
$P(X=-1,Y=1)=\frac{1}{4}$, $P(X=0,Y=0)=\frac{1}{4}$, $P(X=0,Y=2)=\frac{1}{4}$, $P(X=1,Y=1)=\frac{1}{4}$
네 확률의 합이 1이므로 표는 정상적인 결합확률분포이다.
X의 주변확률: $P(X=-1)=\frac{1}{4}$, $P(X=0)=\frac{1}{2}$, $P(X=1)=\frac{1}{4}$
Y의 주변확률: $P(Y=0)=\frac{1}{4}$, $P(Y=1)=\frac{1}{2}$, $P(Y=2)=\frac{1}{4}$

사용할 공식

기댓값 $E(X)=\sum xP(X=x)$, 조건부확률 $P(X=x\mid Y=y)=\frac{P(X=x,Y=y)}{P(Y=y)}$, 조건부 기댓값 $E(X\mid Y=y)=\sum xP(X=x\mid Y=y)$, 독립 조건 $P(X=x,Y=y)=P(X=x)P(Y=y)$ (모든 x,y에서 성립해야 함)

계산 과정

ㄱ. $E(X)=(-1)\cdot\frac{1}{4}+0\cdot\frac{1}{2}+1\cdot\frac{1}{4}=-\frac{1}{4}+\frac{1}{4}=0$이므로 참이다.
ㄴ. $Y=1$인 칸은 $X=-1$일 때 $\frac{1}{4}$, $X=1$일 때 $\frac{1}{4}$이고 $P(Y=1)=\frac{1}{2}$이다. 따라서 $P(X=-1\mid Y=1)=\frac{1}{2}$, $P(X=1\mid Y=1)=\frac{1}{2}$이고, $E(X\mid Y=1)=(-1)\cdot\frac{1}{2}+1\cdot\frac{1}{2}=0$이므로 참이다.
ㄷ. $P(X=0,Y=0)=\frac{1}{4}$인데 $P(X=0)P(Y=0)=\frac{1}{2}\times\frac{1}{4}=\frac{1}{8}$로 서로 다르다. 결합확률과 주변확률의 곱이 일치하지 않으므로 X와 Y는 독립이 아니다.

선지 판단

① ㄱ — ㄱ만 옳으나 ㄴ도 옳으므로 불완전한 선택이라 틀린 선지이다.
② ㄱ, ㄴ — $E(X)=0$이고 $E(X\mid Y=1)=0$이 모두 성립하며 ㄷ은 성립하지 않으므로 정확히 옳은 조합이다.
③ ㄴ, ㄷ — ㄴ은 옳지만 ㄷ이 독립이 아니므로 틀린 선지이다.
④ ㄱ, ㄴ, ㄷ — ㄷ이 독립이 아니므로 틀린 선지이다.

오답 포인트

이 문제에서 가장 흔한 실수는 X와 Y의 결합확률표에서 0이 아닌 칸이 대각선 형태로 배치되어 있다고 해서 곧바로 독립이라고 오판하거나, 반대로 $E(X)=0$과 같은 대칭적인 결과만 보고 독립이라고 단정하는 것이다. 독립 여부는 반드시 결합확률과 두 주변확률의 곱을 직접 비교해서 확인해야 하며, 이 표에서는 P(X=0,Y=0)=14 ≠ P(X=0)P(Y=0)=18이므로 독립이 아니다. 또한 조건부 기댓값을 구할 때 전체 확률 14 대신 조건부 확률(전체를 $P(Y=1)$로 나눈 값)을 사용해야 한다는 점도 놓치기 쉽다.

NoteBase

다시 나오면 이렇게 판단

결합확률분포표 문제는 먼저 확률이 0이 아닌 칸만 골라 표시하고, 그 합이 1인지 확인한 뒤 X와 Y 각각의 주변확률을 구해 둔다. 기댓값은 주변확률로, 조건부 기댓값은 조건에 해당하는 열(또는 행)의 확률을 그 열의 합으로 나눈 조건부확률로 계산한다. 독립성은 느낌으로 판단하지 않고 적어도 한 칸에서 결합확률과 주변확률의 곱이 다르면 독립이 아니라고 결론짓는다.

암기 포인트

독립 판정은 "모든 칸에서" $P(X=x, Y=y) = P(X=x)P(Y=y)$가 성립해야 하며, 한 칸이라도 어긋나면 독립이 아니다. 조건부 기댓값은 조건부확률(전체 확률을 조건이 되는 사건의 확률로 나눈 값)로 가중평균해야 한다.

NoteBase

문 15. 피어슨(Pearson)의 표본상관계수r에 대한 설명으로 옳은 것만을 모두 고른 것은?

ㄱ. r의 범위는 $-1 \le r \le 1$이다.
ㄴ. r은 두 변수 사이에 존재하는 직선관계 또는 곡선관계의 정도를 측정한다.
ㄷ. r이 0이면 두 변수는 항상 독립이다.

① ㄱ
② ㄱ, ㄴ
③ ㄱ, ㄷ
④ ㄴ, ㄷ

정답 1번

출제 포인트

피어슨 표본상관계수 r의 정의와 성질을 정확히 알고 있는지를 묻는 문제이다. r의 값의 범위, r이 측정하는 관계의 종류(직선관계 여부), 그리고 $r = 0$과 독립의 관계라는 세 가지 핵심 성질을 각각 옳고 그름으로 구분할 수 있어야 한다.

핵심 개념

피어슨 표본상관계수는 두 변수 X, Y에 대해 다음과 같이 정의된다.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

이 값은 코시-슈바르츠 부등식에 의해 항상 $-1 \le r \le 1$의 범위를 가진다. r은 오직 두 변수 사이의 직선(선형) 관계의 강도와 방향만을 측정하는 지표이며, 곡선관계처럼 비선형 관계는 반영하지 못한다. 예를 들어 $Y = X^2$ 처럼 완벽한 곡선관계가 있어도 r은 0에 가까운 값이 나올 수 있다. 또한 상관계수가 0이라는 것은 두 변수 사이에 선형관계가 없다는 뜻일 뿐, 두 변수가 서로 독립이라는 것을 의미하지는 않는다. 두 변수가 독립이면 상관계수는 반드시 0이 되지만, 그 역은 성립하지 않는다(비선형 관계가 존재하면서도 상관계수는 0이 되는 경우가 있다).

선지별 해설

ㄱ. r의 범위는 -1≤r≤1이다. → O. 피어슨 상관계수는 정의상 항상 이 범위 안의 값을 가진다.
ㄴ. r은 두 변수 사이에 존재하는 직선관계 또는 곡선관계의 정도를 측정한다. → X. r은 직선(선형)관계의 정도만을 측정하며, 곡선(비선형)관계는 측정하지 못한다. 따라서 "직선관계 또는 곡선관계"라는 서술이 틀렸다. 올바른 표현은 "r은 두 변수 사이에 존재하는 직선관계의 정도만을 측정한다"이다.
ㄷ. r이 0이면 두 변수는 항상 독립이다. → X. r=0은 두 변수 사이에 선형관계가 없다는 것을 의미할 뿐이며, 비선형관계로 인해 종속관계가 존재할 수 있으므로 항상 독립이라고 단정할 수 없다. 올바른 표현은 "r이 0이면 두 변수 사이에 선형관계가 없다는 것을 의미할 뿐, 항상 독립인 것은 아니다"이다.
따라서 옳은 것은 ㄱ뿐이므로 정답은 ①이다.

NoteBase

오답 포인트

ㄴ을 고르는 학생은 상관계수가 '두 변수 사이의 관계 전반'을 측정한다고 넓게 생각해 곡선관계까지 포함시키는 오류를 범한다. 상관계수는 어디까지나 직선(선형) 관계만을 대상으로 한다는 점을 기억해야 한다. ㄷ을 고르는 학생은 '상관이 없다=독립이다'로 곧바로 연결하는 오류를 범한다. 독립이면 무상관(r=0)이 성립하지만, 무상관이라고 해서 항상 독립인 것은 아니라는 방향성을 헷갈리지 않아야 한다.

다시 나오면 이렇게 판단

상관계수 관련 문제가 나오면 먼저 범위(-1~1)는 항상 옳은 진술로 처리하고, '직선관계'라는 단어가 '곡선관계' 또는 '모든 관계'로 바뀌어 있지 않은지를 확인한다. 그다음 'r=0이면 독립이다'라는 문장이 나오면 곧바로 틀린 진술로 판단한다. 역방향인 '독립이면 r=0이다'는 옳은 진술이므로 방향을 혼동하지 않는다.

암기 포인트

피어슨 상관계수 3원칙: 1) 범위는 $-1 \le r \le 1$이다. 2) r은 직선(선형)관계만 측정한다(곡선관계는 못 잡는다). 3) 독립이면 r=0이지만, r=0이라고 항상 독립인 것은 아니다.

NoteBase

문 16. 다중선형회귀(multiple linear regression)에 대한 설명으로 옳지 않은 것은?

① 결정계수(coefficient of determination)는 총변동 중에서 회귀변동이 차지하는 비율이다.

② 결정계수의 값은 독립변수가 추가될수록 커진다.

③ 다중공선성(multicollinearity)은 독립변수와 종속변수 간의 상관관계가 높은 경우에 발생한다.

④ 다중공선성 유무는 분산확대인자(variance inflation factor)로 검토할 수 있다.

정답 3번

출제 포인트

다중선형회귀의 핵심 개념인 결정계수와 다중공선성의 정의를 정확히 구분할 수 있는지를 묻는 문항이다. 특히 다중공선성이 '어떤 변수들 사이의' 상관관계를 의미하는지가 쟁점이다.

핵심 개념

결정계수(R^2)는 총변동(SST) 중 회귀에 의해 설명되는 변동(SSR)이 차지하는 비율로 정의되며, $R^2 = SSR/SST$로 나타낸다. 다중공선성은 회귀모형에 포함된 여러 독립변수들 간에 상관관계가 높아 회귀계수 추정이 불안정해지는 현상을 말한다. 즉 다중공선성은 독립변수와 종속변수 사이가 아니라 독립변수들끼리의 관계에서 생기는 문제이다. 다중공선성 여부는 분산확대인자(VIF)를 계산하여 판단하며, 일반적으로 VIF 값이 크면(흔히 10 이상) 다중공선성이 심한 것으로 본다.

선지별 해설

① 결정계수는 총변동 중에서 회귀변동이 차지하는 비율이라는 설명은 정의와 일치한다. (O)
② 다중선형회귀에서 결정계수는 독립변수를 추가하면 설명력이 줄어들지 않고 그대로 유지되거나 커지는 성질을 가진다. 새 변수가 종속변수와 전혀 관련이 없어도 우연에 의한 상관까지 반영되어 결정계수 값은 감소하지 않고 커지는 경향을 보인다. (O)
③ 다중공선성은 '독립변수와 종속변수 간의 상관관계'가 아니라 '독립변수들 간의 상관관계'가 높을 때 발생한다. 종속변수와 독립변수 사이의 강한 상관관계는 오히려 모형의 설명력을 높이는 바람직한 특성이므로 이 선지의 설명은 틀렸다. 틀린 표현 → 올바른 표현: '독립변수와 종속변수 간의 상관관계' → '독립변수들 간의 상관관계'. (X)
④ 분산확대인자(VIF)는 다중공선성의 정도를 진단하는 대표적 지표로, 각 독립변수를 나머지 독립변수들로 회귀시켰을 때의 결정계수를 이용해 계산하며 VIF 값이 클수록 다중공선성이 심함을 의미한다. (O)

오답 포인트

다중공선성을 '독립변수와 종속변수의 관계'로 잘못 이해하면 안 된다. 다중공선성은 어디까지나 여러 독립변수들 사이의 상관관계 문제이며, 독립변수와 종속변수 사이의 관계는 오히려 모형의 설명력(결정계수)과 관련된 별개의 개념이다. 이 둘을 혼동하면 ②와 ③을 반대로 판단하기 쉽다.

NoteBase

다시 나오면 이렇게 판단

'다중공선성'이라는 단어가 나오면 항상 '독립변수들 간의' 상관관계인지부터 확인한다. 독립변수와 종속변수 사이의 상관관계 이야기가 나오면 그것은 결정계수·설명력의 문제이지 다중공선성의 정의가 아니다.

암기 포인트

결정계수 = 회귀변동/총변동, 독립변수 추가 시 감소하지 않고 증가하는 경향. 다중공선성 = 독립변수들끼리의 상관관계가 높을 때 발생, 진단 도구는 분산확대인자(VIF).

NoteBase

문 17. 새로운 정책에 대해 산업별로 찬반에 차이가 있는지 알아보기 위하여 중공업분야 전문가 130명, 농업분야 전문가 130명을 임의추출하여 조사한 결과가 다음과 같을 때, 이에 대한 설명으로 옳은 것은? (단, 중공업에서의 찬성률은 P_1, 농업에서의 찬성률은 P_2, 비율차는 $P_1 - P_2$, 상대비율은 $\frac{P_1}{P_2}$, 오즈비는 $\frac{P_1/(1-P_1)}{P_2/(1-P_2)}$ 이다)

정책 찬반 / 산업 분야	찬성	반대	계
중공업	100	30	130
농업	70	60	130
계	170	90	260

① 표본오즈비는 $\frac{100 \times 70}{30 \times 60} = \frac{35}{9}$이다.

② 비율차가 0에 가까울수록 두 분야의 찬성률 차이는 커질 것이다.

③ 상대비율이 1에 가까울수록 두 분야의 찬성률 차이는 작아질 것이다.

④ 오즈비의 값이 0.5이면 두 분야의 찬성률은 동일하다고 할 수 있다.

정답 3번

출제 포인트

2×2 분할표에서 두 집단(중공업, 농업)의 찬성률 P_1, P_2를 비교하는 세 가지 지표인 비율차 $P_1 - P_2$, 상대비율 $\frac{P_1}{P_2}$, 오즈비 $\frac{P_1/(1-P_1)}{P_2/(1-P_2)}$의 정의와 각 지표가 나타내는 '두 비율이 같다'는 기준값(비율차는 0, 상대비율과 오즈비는 1)을 정확히 아는지 묻는다.

조건 정리

중공업: 찬성 100명, 반대 30명, 계 130명 → $P_1 = \frac{100}{130}$. 농업: 찬성 70명, 반대 60명, 계 130명 → $P_2 = \frac{70}{130}$. 두 찬성률이 같으면(즉 산업별 차이가 없으면) 비율차는 0, 상대비율과 오즈비는 1이 되어야 한다는 기준을 세워 각 선지를 판단한다.

NoteBase

사용할 공식

비율차 $P_1 - P_2$: 값이 0이면 두 찬성률이 같고, 0에서 멀어질수록(양수든 음수든 절댓값이 커질수록) 차이가 커진다.
상대비율 $\frac{P_1}{P_2}$: 값이 1이면 두 찬성률이 같고, 1에서 멀어질수록 차이가 커진다.
오즈비 $\frac{P_1/(1-P_1)}{P_2/(1-P_2)}$: 값이 1이면 두 찬성률이 같고, 1에서 멀어질수록(1보다 훨씬 크거나 훨씬 작을수록) 차이가 커진다.
표본오즈비 계산 시 대각 원소끼리 곱해 비교한다: $\frac{100 \times 60}{30 \times 70}$.

계산 과정

$P_1 = \frac{100}{130} \approx 0.769, P_2 = \frac{70}{130} \approx 0.538.$
비율차: $P_1 - P_2 = \frac{100-70}{130} = \frac{30}{130} \approx 0.231$로 0보다 크므로 두 찬성률에 차이가 있다.
상대비율: $\frac{P_1}{P_2} = \frac{100/130}{70/130} = \frac{100}{70} = \frac{10}{7} \approx 1.43$으로 1보다 크며, 이 값이 1에 가까울수록 P_1과 P_2가 비슷해져 차이가 작아진다.
표본오즈비: $\frac{100/30}{70/60} = \frac{100 \times 60}{30 \times 70} = \frac{6000}{2100} = \frac{20}{7} \approx 2.86$이며, 이는 1에서 멀리 떨어진 값이므로 두 산업 간 찬반 비율에 차이가 있음을 나타낸다.

선지 판단

① X: 표본오즈비는 대각 원소인 (중공업 찬성×농업 반대)와 (중공업 반대×농업 찬성)을 곱해 비교해야 하므로 $\frac{100 \times 60}{30 \times 70} = \frac{20}{7}$이 되어야 한다. 선지의 식 $\frac{100 \times 70}{30 \times 60} = \frac{35}{9}$는 대각이 아닌 같은 행끼리 곱한 잘못된 식이며 값도 실제 오즈비(약 2.86)와 다르다.
② X: 비율차는 두 찬성률이 같을 때 0이 되고, 0에서 멀어질수록(0으로부터의 거리가 커질수록) 두 분야의 찬성률 차이가 커진다. 즉 '0에 가까울수록 차이가 커진다'가 아니라 '0에 가까울수록 차이가 작아진다'가 맞다.
③ O: 상대비율 $\frac{P_1}{P_2}$은 두 찬성률이 같을 때 1이 되므로, 1에 가까울수록 P_1과 P_2의 값이 비슷해져 두 분야의 찬성률 차이는 작아진다.
④ X: 오즈비는 두 찬성률이 같을 때 1이 되는 지표이다. 오즈비가 0.5이면 1이 아니므로 두 찬성률은 서로 다르며(오히려 P_1이 P_2보다 작은 방향으로 차이가 있음을 뜻한다), 동일하다고 말할 수 없다.

오답 포인트

비율차·상대비율·오즈비는 모두 '두 비율이 같다'를 나타내는 기준값이 다르다(비율차는 0, 상대비율과 오즈비는 1). 이 기준값을 헷갈려 0과 1을 서로 바꿔 쓰면 ②, ④처럼 반대로 판단하게 된다. 또한 오즈비를 계산할 때 분할표에서 어느 칸끼리 곱해야 대각(교차) 곱인지 헷갈리면 ①처럼 엉뚱한 값을 만들게 된다.

NoteBase

다시 나오면 이렇게 판단

먼저 표를 보고 '비율차=0', '상대비율=1', '오즈비=1'이 각각 무엇을 의미하는지(두 집단의 비율이 같다는 뜻) 적어 둔 뒤, 각 선지가 그 기준값에 가까워질 때 차이가 커지는지 작아지는지를 하나씩 대입해 확인한다. 오즈비를 직접 계산해야 하는 선지가 나오면 분할표의 좌상-우하를 곱한 값과 우상-좌하를 곱한 값을 나눈 형태인지부터 확인한다.

암기 포인트

비율차 0, 상대비율·오즈비 1 = '차이 없음' 기준값. 오즈비 계산은 대각(교차) 곱셈 $\frac{ad}{bc}$ 형태임을 기억해 둔다.

NoteBase

문 18. 다음은 어느 지역에서 600명을 임의추출하여 흡연 여부와 치주질환 유무를 조사한 결과이다. 흡연 여부와 치주질환 유무에 대한 독립성 검정을 하고자 할 때, 이에 대한 설명으로 옳은 것만을 모두 고른 것은?

흡연 여부 \ 치주질환 유무	있음	없음	계
과거흡연	50	100	150
현재흡연	50	100	150
비흡연	100	200	300
계	200	400	600

ㄱ. 귀무가설은 “흡연 여부와 치주질환 유무는 서로 관계가 있다.”이다.

ㄴ. 카이제곱 검정통계량의 값은 0이다.

ㄷ. 검정통계량 값이 임계치보다 작으면 귀무가설을 기각한다.

ㄹ. 귀무가설 하에서 카이제곱 검정통계량의 자유도는 2이다.

① ㄱ, ㄹ

② ㄴ, ㄷ

③ ㄱ, ㄷ

④ ㄴ, ㄹ

정답 4번

출제 포인트

2×3 분할표(흡연 여부 3범주 × 치주질환 유무 2범주)에서 독립성 검정의 귀무가설 설정, 카이제곱 검정통계량의 계산(기대도수와 관측도수 비교), 기각역 판단 기준, 자유도 산출을 종합적으로 이해하고 있는지를 묻는다.

조건 정리

표본 크기는 총 600명이며, 행(흡연 여부)은 과거흡연 150명·현재흡연 150명·비흡연 300명, 열(치주질환 유무)은 있음 200명·없음 400명이다. 각 칸의 관측도수는 과거흡연-있음 50, 과거흡연-없음 100, 현재흡연-있음 50, 현재흡연-없음 100, 비흡연-있음 100, 비흡연-없음 200이다.

NoteBase

사용할 공식

독립성 검정의 귀무가설은 H_0: 두 변수(흡연 여부와 치주질환 유무)는 서로 독립이다(관계가 없다)이며, 대립가설은 H_1: 두 변수는 서로 관계가 있다이다. 기대도수는 E_ij=(행 합계_i)×(열 합계_j)n로 구하고, 검정통계량은 $\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$이며, 자유도는 (행의 수-1)×(열의 수-1)이다. 검정통계량 값이 자유도에 대응하는 임계치보다 크면 귀무가설을 기각한다.

계산 과정

각 칸의 기대도수를 구하면 다음과 같다.
과거흡연-있음: $150 \times 200/600 = 50$
과거흡연-없음: $150 \times 400/600 = 100$
현재흡연-있음: $150 \times 200/600 = 50$
현재흡연-없음: $150 \times 400/600 = 100$
비흡연-있음: $300 \times 200/600 = 100$
비흡연-없음: $300 \times 400/600 = 200$
모든 칸에서 기대도수와 관측도수가 정확히 일치하므로(각 흡연 범주 내 '있음:없음' 비율이 200:400=1:2로 전체 비율과 동일), $\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{0^2}{50} + \frac{0^2}{100} + \cdots = 0$이 된다. 행의 수가 3(과거흡연·현재흡연·비흡연), 열의 수가 2(있음·없음)이므로 자유도는 $(3-1) \times (2-1) = 2$이다.

선지 판단

ㄱ. 틀림 — 독립성 검정의 귀무가설은 "두 변수는 서로 관계가 없다(독립이다)"이다. "서로 관계가 있다"는 대립가설의 내용이므로 귀무가설 설명이 뒤바뀌었다.
ㄴ. 맞음 — 계산 과정에서 확인했듯이 모든 칸의 관측도수가 기대도수와 정확히 일치하므로 카이제곱 검정통계량의 값은 0이다.
ㄷ. 틀림 — 카이제곱 검정은 검정통계량 값이 임계치보다 클 때 귀무가설을 기각한다. 임계치보다 작으면 기각하지 못하고 귀무가설을 채택(유지)한다.
ㄹ. 맞음 — 행이 3개(과거흡연·현재흡연·비흡연), 열이 2개(있음·없음)이므로 자유도는 $(3-1)(2-1) = 2$이다.
따라서 옳은 설명은 ㄴ, ㄹ이므로 4번이 정답이다.

오답 포인트

①·③번은 ㄱ을 옳다고 본 경우인데, 귀무가설과 대립가설의 문구를 바꿔 읽으면 안 된다. 독립성 검정의 귀무가설은 항상 "관계가 없다(독립이다)" 쪽이다. ②·③번은 ㄷ을 옳다고 본 경우인데, 카이제곱 검정의 기각 방향(큰 값일 때 기각)을 반대로 착각하면 안 된다.

NoteBase

다시 나오면 이렇게 판단

분할표 문제가 나오면 먼저 각 칸의 기대도수(행 합계×열 합계÷전체)를 구해 관측도수와 비교하는 습관을 들인다. 모든 칸에서 두 값이 같으면 검정통계량은 0이 되고, 이는 자료가 완전히 독립적인 구조(각 행에서 열 비율이 전체 비율과 동일)임을 뜻한다. 그리고 귀무가설의 방향(관계 없음)과 기각 기준(임계치보다 클 때 기각)은 항상 함께 점검한다.

암기 포인트

독립성 검정 귀무가설: "두 변수는 서로 관계가 없다(독립이다)." 자유도: (행의 수-1)×(열의 수-1). 기각 규칙: 검정통계량 > 임계치이면 귀무가설 기각. 관측도수=기대도수이면 카이제곱 통계량은 0이다.

NoteBase

문 19. 관측쌍 $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$의 대응(쌍체)비교 t-검정(paired t-test)에 대한 설명 중 옳지 않은 것은?

① 각 관측쌍들 간은 독립이고, 쌍 내의 자료들은 독립이 아닌 경우에 사용된다.

② 쌍으로 관측된 자료의 차에 대해 일표본 t-검정을 적용하는 방법이다.

③ 귀무가설 하에서 검정통계량의 자유도는 $2(n-1)$이다.

④ n이 충분히 큰 경우에는 정규분포로 근사하여 검정할 수 있다.

정답 3번

출제 포인트

대응(쌍체)비교 t-검정(paired t-test)의 적용 조건과 검정통계량의 자유도를 정확히 알고 있는지를 묻는 문제이다. 대응비교 검정은 같은 대상에서 처리 전후 또는 짝지어진 두 관측값을 비교할 때 사용하며, 두 관측값의 차이를 새로운 확률변수로 만들어 일표본 t-검정으로 환원하는 것이 핵심이다.

핵심 개념

대응비교 t-검정은 관측쌍 (X_i, Y_i)마다 차이 $D_i = X_i - Y_i$를 구하고, 이 차이들 $D_1, D_2, \ldots, D_n$이 서로 독립이며 정규분포를 따른다고 가정한다. 각 관측쌍(즉 i번째 쌍과 j번째 쌍)은 서로 독립이지만, 같은 쌍 안의 X_i와 Y_i는 독립이 아니어도 된다(오히려 상관되어 있는 경우에 이 검정을 쓰는 이유가 된다). 검정통계량은 D_i들의 표본평균 $\bar{D}$와 표본표준편차 S_D를 이용해 $t = \frac{\bar{D}}{S_D/\sqrt{n}}$로 계산되며, 이는 표본 하나(차이값들)에 대한 일표본 t-검정과 완전히 동일한 형태이다. 표본의 개수가 n개(관측쌍이 n쌍)이므로 차이값 D_i도 n개이고, 표본평균과 표본분산을 계산하는 과정에서 자유도가 하나 줄어들어 검정통계량의 자유도는 $n-1$이 된다. $2(n-1)$은 두 독립표본 t-검정에서 등분산을 가정할 때 각 표본의 자유도 $n_1 - 1, n_2 - 1$를 합쳐 얻는 자유도(예: $n_1 = n_2 = n$이면 $2(n-1)$)와 혼동한 표현으로, 대응비교 검정에는 해당하지 않는다.

선지별 해설

① 옳다. 대응비교 검정은 관측쌍들 사이(서로 다른 i) 독립을 가정하지만, 같은 쌍 내의 X_i, Y_i는 독립이 아닌(상관된) 상황에서 사용하는 방법이다. 만약 쌍 내 자료도 독립이라면 애초에 대응비교를 할 이유 없이 독립표본 t-검정을 쓰면 된다.

② 옳다. 각 쌍의 차이 $D_i = X_i - Y_i$를 구한 뒤, 이 차이들의 모평균이 0인지를 검정하는 일표본 t-검정을 적용하는 것이 대응비교 t-검정의 정의이다.

③ 옳지 않다 → 귀무가설 하에서 검정통계량은 자유도 $n-1$인 t-분포를 따른다. $2(n-1)$은 두 독립표본(등분산 가정, 표본크기가 같은 경우)의 합동분산 t-검정에서 나오는 자유도이며, 관측쌍 개수가 n개인 대응비교 검정의 자유도와는 다르다.

④ 옳다. 표본크기 n이 충분히 크면 중심극한정리에 의해 표본평균 $\bar{D}$의 분포가 정규분포에 근사하므로, t-분포 대신 표준정규분포를 이용해 근사적으로 검정할 수 있다.

NoteBase

오답 포인트

대응비교 검정의 자유도를 두 독립표본 비교(자유도 $n_1 + n_2 - 2$, 표본크기가 같으면 $2(n-1)$)의 자유도와 혼동하기 쉽다. 대응비교는 애초에 표본이 하나(차이값들의 표본)뿐이므로 자유도도 그 표본크기에서 1을 뺀 $n-1$이라는 점, 즉 '몇 쌍을 관측했는가'가 자유도를 결정한다는 점을 놓치지 않아야 한다.

다시 나오면 이렇게 판단

선지에 자유도가 등장하면 먼저 어떤 검정통계량인지(일표본, 대응비교, 독립 두 표본 등분산, 독립 두 표본 이분산 근사) 확인하고, 그에 맞는 표본 개수와 자유도 공식을 떠올린다. 대응비교는 '쌍의 개수 n에서 1을 뺀 값'이 자유도라는 것을 기준점으로 삼는다.

암기 포인트

대응비교 t-검정: 차이 $D_i = X_i - Y_i$로 환원 → 일표본 t-검정과 동일 → 자유도 $n-1$. 쌍 간은 독립, 쌍 내는 독립 아니어도 됨(오히려 그런 상황에 사용). n이 크면 정규분포 근사 가능.

NoteBase

문 20. X_1과 X_2가 다음과 같은 확률밀도함수(probability density function)로부터 얻어진 임의표본일 때, 확률 $P(\min(X_1, X_2) < \frac{1}{2})$의 값은? (단, $\min(X_1, X_2)$는 X_1과 X_2 중에서 더 작은 확률변수이다)

$f(x) = 4x^3,\ 0 < x < 1$

① $\frac{3}{8}$
② $\frac{1}{4}$
③ $\frac{21}{128}$
④ $\frac{31}{256}$

정답 4번

출제 포인트

두 독립인 연속확률변수 중 최솟값의 분포를 구하는 문제이다. $P(\min(X_1, X_2) < c)$를 직접 구하기보다 여사건인 $P(\min(X_1, X_2) \geq c)$를 이용해 $1 - [P(X > c)]^2$ 형태로 계산하는 것이 핵심이다.

조건 정리

X_1, X_2는 서로 독립이고 동일한 확률밀도함수 $f(x) = 4x^3,\ 0 < x < 1$을 따르는 임의표본이다. 구하는 값은 $P(\min(X_1, X_2) < \frac{1}{2})$이다.

사용할 공식

누적분포함수는 $F(x) = \int_0^x 4t^3 dt = x^4,\ 0 < x < 1$이다. 두 확률변수 중 최솟값이 특정 값보다 작을 확률은 여사건을 이용해 $P(\min(X_1, X_2) < c) = 1 - P(X_1 \geq c)P(X_2 \geq c) = 1 - [1 - F(c)]^2$로 구한다(독립이므로 곱으로 분리된다).

계산 과정

$c = \frac{1}{2}$를 대입하면 $F\left(\frac{1}{2}\right) = \left(\frac{1}{2}\right)^4 = \frac{1}{16}$이다. 따라서 $P\left(X \geq \frac{1}{2}\right) = 1 - \frac{1}{16} = \frac{15}{16}$이다. 두 표본이 독립이므로 $P\left(\min(X_1, X_2) \geq \frac{1}{2}\right) = \left(\frac{15}{16}\right)^2 = \frac{225}{256}$이다. 구하는 확률은 $P\left(\min(X_1, X_2) < \frac{1}{2}\right) = 1 - \frac{225}{256} = \frac{31}{256}$이다.

선지 판단

① 3/8: 여사건을 이용하지 않고 단순히 $2F(1/2)$ 형태로 잘못 계산했을 때 나올 수 있는 값으로 틀린 값이다.
② 1/4: 두 표본의 독립성을 반영하지 않고 $F(1/2)$를 단순히 조작한 값으로 틀린 값이다.
③ 21/128: 제곱 계산 과정에서 부호나 항을 잘못 처리했을 때 나오는 값으로 틀린 값이다.
④ 31/256: 앞선 계산 과정에서 도출한 값과 정확히 일치하므로 옳다.

NoteBase

오답 포인트

이 문제에서 가장 흔한 실수는 $P(\min(X_1, X_2) < 1/2)$를 $P(X_1 < 1/2)$와 $P(X_2 < 1/2)$를 단순히 더하거나 하나만 계산해서 답으로 삼는 것이다. 최솟값이 어떤 값보다 작다는 것은 두 표본 중 적어도 하나가 그 값보다 작다는 의미이므로, 그 반대인 두 표본이 모두 그 값 이상일 확률을 구해 1에서 빼는 여사건 접근이 필요하다. 또한 두 표본이 독립이라는 조건을 놓치고 곱셈 대신 다른 연산을 적용하면 오답이 나온다.

다시 나오면 이렇게 판단

확률변수 여러 개 중 최솟값이나 최댓값에 관한 확률을 구하는 문제는 먼저 누적분포함수 $F(x)$를 구하고, 최솟값 문제는 여사건 $1-[1-F(c)]^n$, 최댓값 문제는 $[F(c)]^n$ 형태로 접근한다고 정리해 둔다. 독립인 표본의 개수 n에 맞춰 거듭제곱을 취하는 것이 핵심이다.

암기 포인트

독립인 표본 n개의 최솟값에 대해 $P(\min(X_1, \dots, X_n) < c) = 1-[1-F(c)]^n$이고, 최댓값에 대해 $P(\max(X_1, \dots, X_n) < c) = [F(c)]^n$임을 공식처럼 기억해 둔다.

NoteBase