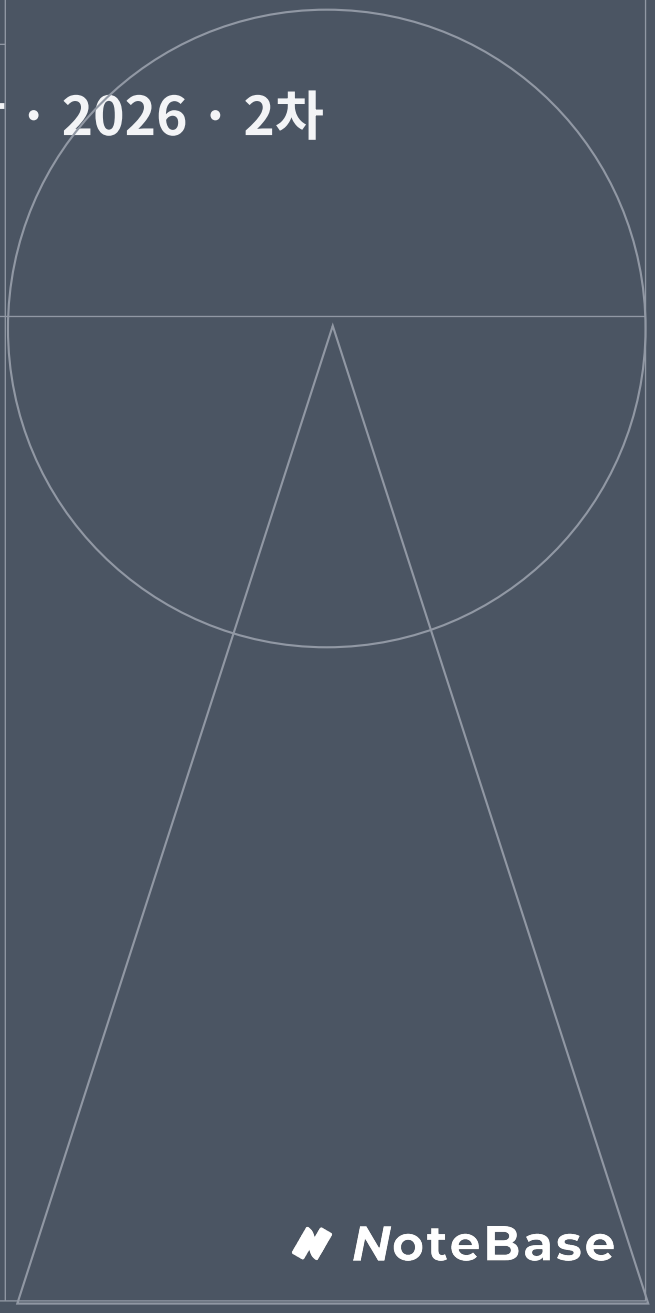
기출OX 해설집

공무원/국가직/7급 · 통계학 · 2026 · 2차



NoteBase

기출OX 해설집

공무원/국가직/7급 · 통계학 · 2026 · 2차

공무원/국가직/7급 · 통계학 · 2026 · 2차

문 1. 다음은 통계학을 수강하는 학생 30명 성적에 대한 줄기-잎 그림이다. 제1사분위수와 최빈값을 더한 값은?

4	1 4
5	1 1 4 8
6	1 1 1 2 2 3
7	1 2 3 3 5 6 7
8	1 1 2 5 6 7 7 8
9	1 4 8

① 122

② 124

③ 126

④ 128

정답 1번

간단한 해설

제1사분위수는 **하위 15개 성적의 중앙값**인 61점이고, 최빈값도 가장 많이 나타나는 61점이다. 두 값의 합은 61 + 61 = 122점이다.

NoteBase

계산 과정

자료 읽기: 줄기는 십의 자리, 잎은 일의 자리를 나타낸다. 예를 들어 줄기 4의 잎 1은 41점이다. 잎 하나가 학생 한 명의 성적에 해당하며, 그림에는 30개의 성적이 작은 값부터 정렬되어 있다.

제1사분위수: 자료를 크기순으로 반씩 나누었을 때 **아래쪽 절반의 중앙값**을 구한다. 30명의 아래쪽 절반은 작은 순서로 1번째부터 15번째까지이고, 이 15개 성적의 가운데는 전체에서 8번째 성적이다. 40점대가 2명, 50점대가 4명이므로 60점대 전까지 6명이다. 이어지는 7·8·9번째 성적은 모두 61점이므로 제1사분위수는 61점이다.

최빈값: 전체 자료에서 **출현 횟수가 가장 많은 값**이다. 줄기 6에서 잎 1이 세 번 나타나므로 61점의 도수는 3이다. 다른 성적은 모두 두 번 이하로 나타나므로 최빈값은 61점이다.

두 값의 합: 제1사분위수와 최빈값을 더하면 61 + 61 = 122점이며, 이는 ①의 122에 해당한다.

NoteBase

문 2. 세 사건 A, B, C에 대해 A와 B는 서로 독립, A와 C도 서로 독립, B와 C는 서로 배반이다. $P(A) = 0.2$, $P(B) = 0.3$, $P(C) = 0.5$일 때, $P(A \cup B \cup C)$의 값은?

① 0.78

② 0.84

③ 0.90

④ 0.96

정답 2번

간단한 해설

세 사건의 확률을 더한 뒤 **중복된 교집합 확률**을 빼야 한다. 배반 조건으로 겹침이 없는 부분을 제외하면, 독립 조건으로 구한 교집합 확률 0.06과 0.10을 빼서 합집합 확률 0.84를 얻는다.

계산 과정

배반 조건: B와 C는 **동시에 일어날 수 없으므로** $P(B \cap C) = 0$이다. 세 사건이 모두 일어나려면 B와 C도 동시에 일어나야 하므로 $P(A \cap B \cap C) = 0$이다.

독립 조건: 두 사건이 독립이면 **교집합 확률은 각 확률의 곱**이다. 따라서 $P(A \cap B) = 0.2 \cdot 0.3 = 0.06$, $P(A \cap C) = 0.2 \cdot 0.5 = 0.10$이다. 이 값들은 각각 두 사건이 동시에 일어날 확률이다.

합집합 계산: 구하는 값은 세 사건 중 적어도 하나가 일어날 확률이다. 각 사건의 확률을 더하면 0.2 + 0.3 + 0.5 = 1.0이다. 이 합에는 두 사건이 동시에 일어나는 부분이 두 번씩 포함되므로 교집합 확률을 한 번씩 뺀다. 세 사건의 교집합은 0이므로 다시 더할 부분은 없다. 빼야 할 확률은 0.06 + 0.10 = 0.16이므로 $P(A \cup B \cup C) = 1.0 - 0.16 = 0.84$이다.

따라서 구한 확률 0.84는 ②의 값과 일치한다.

NoteBase

문 3. 정규분포 $N(\mu, 1)$을 따르는 확률표본 X_1, X_2, X_3을 이용하여 모평균 μ를 추정하려고 한다. 다음에 제시된 μ에 대한 추정량 $\hat{\mu}_A, \hat{\mu}_B, \hat{\mu}_C$ 중 평균제곱오차(MSE)가 작은 것부터 순서대로 바르게 나열한 것은?

$\hat{\mu}_A = \frac{1}{2}X_1 + \frac{1}{2}X_2$
$\hat{\mu}_B = \frac{1}{2}X_1 + \frac{1}{4}X_2 + \frac{1}{4}X_3$
$\hat{\mu}_C = \frac{3}{2}X_1 - \frac{1}{4}X_2 - \frac{1}{4}X_3$

① $\hat{\mu}_A, \hat{\mu}_B, \hat{\mu}_C$
② $\hat{\mu}_A, \hat{\mu}_C, \hat{\mu}_B$
③ $\hat{\mu}_B, \hat{\mu}_A, \hat{\mu}_C$
④ $\hat{\mu}_C, \hat{\mu}_A, \hat{\mu}_B$

정답 3번

간단한 해설

세 추정량은 계수의 합이 모두 1인 불편추정량이므로 **MSE는 분산과 같다**. 확률표본은 서로 독립이고 각각의 분산이 1이므로, **계수 제곱의 합**이 작을수록 MSE가 작다.

올바른 순서

$\hat{\mu}_B \to \hat{\mu}_A \to \hat{\mu}_C$

평균제곱오차는 $\mathrm{MSE}(T) = E[(T-\mu)^2]$이며, 분산과 편향 제곱의 합이다. 확률표본이므로 각 관측값은 서로 독립이고, $E(X_i) = \mu$, $\mathrm{Var}(X_i) = 1$이다. 세 추정량은 계수의 합이 모두 1이어서 기댓값이 μ인 불편추정량이다. 따라서 **편향이 0이므로 분산만 비교**하면 된다. 독립인 확률변수의 선형결합에서는 각 계수를 제곱하여 해당 분산에 곱한 뒤 더한다.

$\hat{\mu}_B$: 분산은 $\left(\frac{1}{2}\right)^2 + \left(\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right)^2 = \frac{3}{8}$이다. 이것이 세 추정량 중 가장 작은 MSE이다.

$\hat{\mu}_A$: 분산은 $\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 = \frac{1}{2}$이다. $\frac{1}{2} = \frac{4}{8} > \frac{3}{8}$이므로 MSE가 앞의 추정량보다 크다.

$\hat{\mu}_C$: 분산은 $\left(\frac{3}{2}\right)^2 + \left(-\frac{1}{4}\right)^2 + \left(-\frac{1}{4}\right)^2 = \frac{19}{8}$이다. **음의 계수도 분산을 계산할 때는 제곱**하므로 양의 값으로 더한다. 이 값은 앞의 두 추정량의 MSE보다 크므로 마지막에 놓인다.

NoteBase

문 4. 자료 $(x_i, y_i)(i = 1, 2, \cdots, 10)$에 단순선형회귀모형 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$를 최소제곱법으로 적합한 회귀계수의 추정값은 $\hat{\beta}_0 = 5,\ \hat{\beta}_1 = 2$이다. $\sum_{i=1}^{10} x_i = 120$일 때 $\sum_{i=1}^{10} y_i$의 값은? (단, ε_i는 $N(0, \sigma^2)$을 따르고 서로 독립이다)

① 190

② 240

③ 290

④ 340

정답 3번

간단한 해설

상수항을 포함한 최소제곱회귀에서는 **잔차의 합이 0**이므로 관측값의 합과 적합값의 합이 같다. 따라서 구하는 합은 $10 \times 5 + 2 \times 120 = 290$이다.

계산 과정

관측값은 10개이며, 적합된 회귀식은 $\hat{y}_i = 5 + 2x_i$이다. 잔차는 관측값에서 적합값을 뺀 $e_i = y_i - \hat{y}_i$이다. **상수항을 포함한 최소제곱법**에서는 잔차제곱합을 최소화하는 상수항을 구할 때 잔차의 합이 0이 되므로, $\sum_{i=1}^{10} y_i = \sum_{i=1}^{10} \hat{y}_i$가 성립한다.

적합값을 모두 더하면 상수항 5가 10번 더해지고, 기울기 2에는 모든 x_i의 합 120이 곱해진다. 따라서 $\sum_{i=1}^{10} \hat{y}_i = 10 \cdot 5 + 2 \cdot 120 = 290$이다. 여기서 50은 상수항의 기여분이고, 240은 설명변수에 의한 기여분이다.

따라서 관측값의 합 $\sum_{i=1}^{10} y_i$는 290으로, ③의 값과 일치한다.

NoteBase

문 5. 세 가지 백신에 따른 항체 생성률의 차이가 있는지를 확인하기 위해 백신 $k(k=1,2,3)$를 접종한 후, 항체가 미생성된 개체 수(O_{k0})와 항체가 생성된 개체 수(O_{k1})를 조사한 결과는 다음과 같다.

항체 생성 여부 / 백신	미생성	생성	계
백신 1	O_{10}	O_{11}	n_1
백신 2	O_{20}	O_{21}	n_2
백신 3	O_{30}	O_{31}	n_3
계	O_0	O_1	n

백신 k를 접종한 후 항체가 생성된 개체의 비율을 p_k라 할 때, 가설 H_0: '$p_1=p_2=p_3$이다' 대 H_1: 'p_1, p_2, p_3 중 적어도 하나는 다르다'에 대해 카이제곱검정을 하고자 한다. 귀무가설이 참일 때 백신 k를 접종한 후 항체가 미생성된 개체의 추정 기대도수를 $\hat{E}_{k0}$라 하고 항체가 생성된 개체의 추정 기대도수를 $\hat{E}_{k1}$이라 하자. 이에 대한 설명으로 옳은 것만을 모두 고르면?

ㄱ. $\hat{E}_{10}+\hat{E}_{11}=n_1$이다.

ㄴ. $\sum_{k=1}^{3}\hat{E}_{k1}=O_1$이다.

ㄷ. 카이제곱 검정통계량의 자유도는 2이다.

① ㄱ

② ㄱ, ㄷ

③ ㄴ, ㄷ

④ ㄱ, ㄴ, ㄷ

정답 4번

간단한 해설

귀무가설 아래에서는 모든 백신에 **전체 표본의 항체 생성 비율**을 공통으로 적용해 기대도수를 구한다. 이렇게 구한 기대도수는 **관측도수의 행 합계와 열 합계를 보존**한다.

NoteBase

보기별 해설

표에서 n_k는 백신 k의 접종 개체 수, O_1은 전체 항체 생성 개체 수이며, 전체 개체 수는 $n = n_1 + n_2 + n_3$이다. 귀무가설은 세 백신의 항체 생성률이 같다는 것이므로, 공통 생성률을 $\hat{p} = \frac{O_1}{n}$으로 추정한다. 따라서 생성 기대도수는 $\hat{E}_{k1} = n_k\hat{p}$, 미생성 기대도수는 $\hat{E}_{k0} = n_k(1 - \hat{p})$이다.

ㄱ. O: 백신 1의 접종 개체는 항체 미생성과 생성 중 하나에 속한다. 따라서 $\hat{E}_{10} + \hat{E}_{11} = n_1(1 - \hat{p}) + n_1\hat{p} = n_1$이다. 두 기대도수의 합도 **해당 백신의 접종 개체 수**와 같다.

ㄴ. O: 모든 백신의 생성 기대도수를 더하면 $\sum_{k=1}^{3} \hat{E}_{k1} = \hat{p} \sum_{k=1}^{3} n_k = n\hat{p}$이다. 공통 생성률을 전체 표본에서 추정했으므로 $n\hat{p} = n \cdot \frac{O_1}{n} = O_1$이다. 즉, 생성 열의 기대도수 합계는 **실제 항체 생성 개체 수**와 일치한다.

ㄷ. O: 백신 종류가 3개 행, 항체 생성 여부가 2개 열인 분할표의 동질성 검정이다. 카이제곱 검정통계량이 귀무가설 아래에서 근사적으로 따르는 카이제곱분포의 자유도는 $(3 - 1) \cdot (2 - 1) = 2$이다. **행 합계와 열 합계를 고정**하면 첫 두 행의 생성 도수만 독립적으로 정할 수 있고, 나머지 도수는 합계에 맞춰 결정된다.

ㄱ, ㄴ, ㄷ이 모두 성립하므로 이들을 모두 포함한 조합은 ④이다.

NoteBase

문 6. 다음 자료의 중앙값과 산술평균이 각각 50, 55일 때, $x - y$의 값은? (단, $x \geq y$이다)

60, 30, 40, 50, 70, 40, x, y

① 20
② 30
③ 40
④ 50

정답 4번

간단한 해설

산술평균으로부터 $x + y = 150$이고, 중앙값이 50이 되려면 **크기순으로 네 번째와 다섯 번째 값이 모두 50**이어야 한다. 따라서 $y = 50$, $x = 100$이므로 두 값의 차이는 50이다.

계산 과정

산술평균은 자료의 합을 자료 수로 나눈 값이다. 자료가 8개이고 산술평균이 55이므로 전체 합은 $8 \times 55 = 440$이다. 주어진 여섯 수의 합은 $60 + 30 + 40 + 50 + 70 + 40 = 290$이므로 $x + y = 440 - 290 = 150$이다.

$x \geq y$이고 $x + y = 150$이므로 $2x \geq 150$, 즉 $x \geq 75$이다. 따라서 x는 주어진 여섯 수보다 모두 크다. 자료 수가 짝수인 8개이므로 중앙값은 **크기순으로 네 번째와 다섯 번째 값의 산술평균**이다.

y를 제외하면 50보다 작은 값은 30, 40, 40의 세 개이고, 50보다 큰 값은 60, 70, x의 세 개이다. $y < 50$이면 네 번째 값은 50보다 작고 다섯 번째 값은 50이어서 중앙값이 50보다 작다. 반대로 $y > 50$이면 네 번째 값은 50이고 다섯 번째 값은 50보다 커서 중앙값이 50보다 크다. 따라서 **중앙값이 50이 되려면** $y = 50$이어야 하며, 이때 가운데 두 값은 모두 50이다.

$x + y = 150$에 $y = 50$을 대입하면 $x = 100$이다. 따라서 구하는 값은 $x - y = 100 - 50 = 50$이며, ④의 50과 일치한다.

NoteBase

문 7. 어느 대학교는 메일을 스팸 메일과 정상 메일로 분류하는 프로그램을 사용하고 있다. 이 프로그램은 스팸 메일의 99%를 스팸 메일로 분류하고, 정상 메일의 1%는 스팸 메일로 분류한다. 이 대학교 전체 메일의 1%는 스팸 메일이다. 이 프로그램이 어떤 메일을 스팸 메일로 분류했을 때, 해당 메일이 정상 메일일 확률은?

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

정답 1번

간단한 해설

스팸으로 분류되는 메일은 실제 스팸 메일과 잘못 분류된 정상 메일로 나뉘며, 두 경우가 전체 메일에서 차지하는 비율은 각각 0.99%로 같다. 따라서 **스팸으로 분류된 메일 중 정상 메일의 비율**은 $\frac{1}{2}$, 즉 50%이다.

계산 과정

실제 스팸인 사건을 S, 정상인 사건을 N, 프로그램이 스팸으로 분류하는 사건을 D라 하자. 구하려는 값은 분류 결과 D가 주어졌을 때 정상일 조건부확률 $P(N \mid D)$이다. 전체 메일 중 스팸은 1%, 정상은 99%이다.

실제 스팸이면서 스팸으로 분류될 확률은 $P(S \cap D) = 0.01 \cdot 0.99 = 0.0099$이다. 전체 메일의 1%인 스팸 중 99%가 스팸으로 분류되므로, 전체 메일의 0.99%가 이 경우에 해당한다.

정상이면서 스팸으로 잘못 분류될 확률은 $P(N \cap D) = 0.99 \cdot 0.01 = 0.0099$이다. 정상 메일의 오분류율 1%는 **전체 메일의 99%인 정상 메일을 기준**으로 적용하므로, 전체 메일에서 이 경우가 차지하는 비율도 0.99%이다.

스팸으로 분류된 메일은 위의 두 경우로 빠짐없이 나뉘고 두 경우는 서로 겹치지 않는다. 따라서 $P(D) = 0.0099 + 0.0099 = 0.0198$이다. 이는 전체 메일의 1.98%가 스팸으로 분류된다는 뜻이다.

조건부확률은 **스팸으로 분류된 메일 전체를 분모**로 하여, 그중 정상 메일의 비율을 구한다. 따라서 $P(N \mid D) = \frac{P(N \cap D)}{P(D)}$이고, 그 값은 $\frac{0.0099}{0.0198} = \frac{1}{2}$이다. 정상 메일일 확률은 50%이며, ①의 $\frac{1}{2}$에 해당한다.

NoteBase

문 8. 볼펜을 제조하는 어느 회사에서는 자사 제품의 판매 비율 p를 추정하고자 한다. 판매되는 볼펜을 대상으로 임의추출한 n개 표본 중 자사 제품의 비율을 이용하여 p를 추정할 때, 95.44% 신뢰수준에서 p에 대한 추정량의 오차한계를 0.05 이하로 하는 최소 표본의 크기는? (단, 확률변수 Z가 표준정규분포를 따를 때 $P(Z \geq 1) = 0.1587$, $P(Z \geq 2) = 0.0228$이다)

① 100

② 400

③ 900

④ 1,600

정답 2번

간단한 해설

95.44% 신뢰수준의 임계값은 2이며, 판매 비율을 모를 때는 **표본비율의 분산이 가장 큰 경우**인 $p = 0.5$를 적용한다. 이때 오차한계가 0.05 이하가 되려면 최소 400개를 추출해야 한다.

계산 과정

표본비율과 오차한계: 표본 중 자사 제품의 비율을 $\hat{p}$라 하면, 독립적인 임의추출에서 $\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}$이다. 표본크기가 충분히 클 때 표본비율의 분포를 정규분포로 근사하므로, 오차한계는 $E = z\sqrt{\frac{p(1-p)}{n}}$으로 계산한다. 여기서 z는 양측 신뢰수준에 대응하는 표준정규분포의 임계값이다. 오차한계 0.05는 비율의 차이로 5%포인트를 뜻한다.

임계값 결정: 표준정규분포는 좌우 대칭이고 $P(Z \geq 2) = 0.0228$이므로, $P(-2 \leq Z \leq 2) = 1 - 2 \cdot 0.0228 = 0.9544$이다. 따라서 **양쪽 꼬리확률을 제외한** 95.44% 신뢰수준에는 $z = 2$를 사용한다.

미지의 판매 비율 처리: p에 대한 사전 정보가 없으므로, 특정 비율을 임의로 가정하지 않고 **오차한계가 가장 커지는 경우**를 기준으로 표본 크기를 정한다. p(1-p)=14-≤ft(p-12)^2이므로 $p = 0.5$일 때 최댓값 0.25를 갖는다. 이 경우를 기준으로 하면 다른 판매 비율에서도 같은 오차한계 조건을 충족한다.

최소 표본 크기 계산: 위의 최댓값을 대입하면 최대 오차한계는 E_=2(0.25)/(n)=(1)/(n)이다. 따라서 (1)/(n)≤ 0.05를 만족해야 하며, 이를 풀면 $n \geq \left(\frac{1}{0.05}\right)^2 = 400$이다. 표본 크기는 정수이고 경계값 400에서 최대 오차한계가 정확히 0.05이므로, 최소 표본 크기는 400개이며 ②의 값과 일치한다.

NoteBase

문 9. 모평균이 μ인 정규분포를 따르는 모집단에서 임의추출한 10개 자료의 표본평균이 11.5이다. 가설 $H_0 : \mu = 10$ 대 $H_1 : \mu \neq 10$에 대한 검정통계량의 p-값이 0.052일 때, 이에 대한 설명으로 옳은 것만을 모두 고르면?

ㄱ. 가설 $H_0 : \mu = 10$ 대 $H_1 : \mu \neq 10$에 대하여 유의수준 5%에서 귀무가설을 기각한다.
ㄴ. 가설 $H_0 : \mu = 10$ 대 $H_1 : \mu > 10$에 대한 p-값은 0.026이다.
ㄷ. 모평균 μ에 대한 95% 신뢰구간은 10을 포함한다.

① ㄴ
② ㄱ, ㄷ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 3번

간단한 해설

표본평균이 기준값 10보다 크므로, 오른쪽 단측검정의 p-값은 **양측 p-값의 절반**인 0.026이다. 양측 p-값 0.052는 0.05보다 커서 **대응하는 95% 신뢰구간에 10이 포함**된다.

보기별 해설

ㄱ. X: p-값은 귀무가설이 참일 때 관측된 검정통계량만큼 또는 그보다 극단적인 값이 나올 확률이다. p-값이 유의수준 이하이면 귀무가설을 기각한다. 여기서는 0.052 > 0.05이므로 유의수준 5%에서 **귀무가설을 기각하지 않는다**. 이는 모평균이 10이라는 사실을 확정한다는 뜻은 아니다.

ㄴ. O: 정규모집단의 평균에 대한 양측검정에서는 귀무가설 아래 검정통계량의 분포가 좌우 대칭이므로, 양측 p-값은 관측값에 대응하는 한쪽 꼬리확률의 두 배이다. 표본평균 11.5가 귀무가설의 평균 10보다 크므로 관측된 검정통계량은 양수이고, $H_1 : \mu > 10$의 방향과 일치한다. 따라서 **오른쪽 단측검정**의 p-값은 $\frac{0.052}{2} = 0.026$이다.

ㄷ. O: 같은 검정에 대응하는 양측 95% 신뢰구간은 유의수준 5%의 양측검정에서 기각되지 않는 모평균 값들의 집합이다. 모평균을 10으로 둔 양측검정의 p-값이 0.052로 0.05보다 크므로, **95% 신뢰구간은 10을 포함**한다.

옳은 보기는 ㄴ, ㄷ이며, 이 조합은 ③에 해당한다.

NoteBase

문 10. 자료 $(x_i, y_i)(i = 1, 2, \cdots, 11)$에서 표본평균 $\bar{x}$, $\bar{y}$는 각각 2, 3이다.
이 자료에 단순선형회귀모형 $y_i = \beta x_i + \varepsilon_i$를 최소제곱법으로 적합하여 얻은 β의 추정량을 $\hat{\beta}$라 할 때, 이에 대한 설명으로 옳은 것만을 모두 고르면? (단, ε_i는 $N(0, \sigma^2)$을 따르고 서로 독립이다)

ㄱ. 잔차제곱합의 자유도는 10이다.
ㄴ. 추정 회귀직선은 (2,3)을 지난다.
ㄷ. $\hat{\beta}$은 β의 비편향추정량(unbiased estimator)이다.

① ㄷ
② ㄱ, ㄴ
③ ㄱ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 3번

간단한 해설

주어진 모형은 **절편 없이 원점을 지나는 회귀모형**이다. 표본평균점 (2,3)을 지난다는 성질은 절편을 함께 추정할 때 성립하므로, 이 모형에 그대로 적용할 수 없다.

보기별 해설

주어진 모형은 절편 없이 기울기 β 하나만 추정한다. 잔차제곱합을 최소로 만드는 추정량은 $\hat{\beta} = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$이며, 합은 11개 관측값 전체에 대한 합이다. $\bar{x} = 2$이므로 모든 x_i가 0인 경우는 아니며, 분모는 양수이다.

ㄱ. O: 잔차제곱합의 자유도는 관측값 수에서 추정한 회귀계수의 수를 뺀 값이다. 이 모형은 **절편을 추정하지 않고 기울기 하나만 추정**하므로 자유도는 $11 - 1 = 10$이다. 절편과 기울기를 모두 추정하는 모형의 $11 - 2 = 9$와 구별해야 한다.

ㄴ. X: 절편을 포함해 최소제곱법으로 적합한 회귀직선은 표본평균점을 지나지만, 이 모형의 회귀직선은 **원점을 지나도록 제한**되어 있다. (2,3)을 지나려면 $\hat{\beta} = \frac{3}{2}$이어야 한다. 그러나 $\bar{x} = 2, \bar{y} = 3$이라는 정보만으로는 $\frac{\sum_i x_i y_i}{\sum_i x_i^2} = \frac{3}{2}$가 성립하지 않는다. 따라서 이 회귀직선이 평균점 (2,3)을 지난다고 보장할 수 없다.

ㄷ. O: 모형의 $y_i = \beta x_i + \varepsilon_i$를 추정량 식에 대입하면 $\hat{\beta} = \beta + \frac{\sum_i x_i \varepsilon_i}{\sum_i x_i^2}$이다. 설명변수 값들을 고정하여 보면 분모는 상수이고, 주어진 오차분포에서 $E(\varepsilon_i) = 0$이므로 분자의 기댓값도 0이다. 따라서 $E(\hat{\beta}) = \beta$로, 추정량의 기댓값이 실제 모수와 같은 **비편향추정량**이다.

옳은 보기는 ㄱ, ㄷ이며, 이 조합은 ③에 해당한다.

NoteBase

문 11. 어느 놀이공원은 검은 공과 흰 공이 들어 있는 상자에서 흰 공을 뽑으면 선물을 증정하고 있다. 상자 속의 흰 공의 비율 p는 $\frac{1}{2}$로 알려져 있다. 이 비율이 $\frac{1}{4}$로 낮아졌다는 주장을 확인하기 위하여 다음 가설에 대한 검정을 실시하고자 한다. 상자에서 임의로 복원추출한 4개의 공이 모두 검은색이면 귀무가설을 기각한다고 할 때, 이 검정법의 제1종 오류를 범할 확률과 검정력을 바르게 연결한 것은?

$H_0 : p = \frac{1}{2}$ 대 $H_1 : p = \frac{1}{4}$

① 제1종 오류를 범할 확률: $\frac{1}{32}$, 검정력: $\frac{1}{256}$
② 제1종 오류를 범할 확률: $\frac{1}{32}$, 검정력: $\frac{81}{256}$
③ 제1종 오류를 범할 확률: $\frac{1}{16}$, 검정력: $\frac{1}{256}$
④ 제1종 오류를 범할 확률: $\frac{1}{16}$, 검정력: $\frac{81}{256}$

정답 4번

간단한 해설

제1종 오류 확률은 **귀무가설이 참일 때 기각할 확률**이고, 검정력은 **대립가설이 참일 때 기각할 확률**이다. 두 경우 모두 네 공이 전부 검은색일 확률을 계산하므로, 각각 $\frac{1}{16}$과 $\frac{81}{256}$이다.

계산 과정

기각 조건: p는 흰 공의 비율이므로 한 번에 검은 공을 뽑을 확률은 $1 - p$이다. **복원추출이므로 각 추출은 독립**이고 검은 공을 뽑을 확률도 일정하다. 따라서 네 공이 모두 검은색일 확률은 $(1 - p)^4$이며, 이 사건이 귀무가설의 기각에 해당한다.

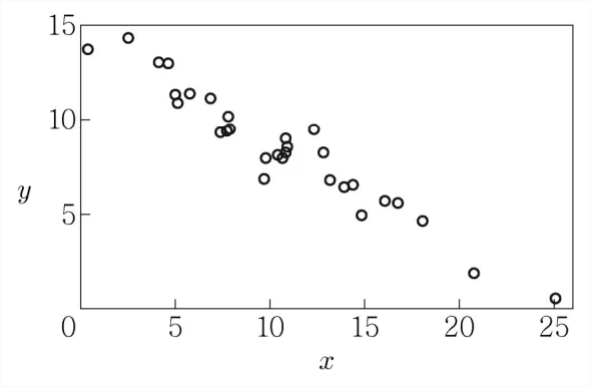
제1종 오류 확률: 제1종 오류는 **참인 귀무가설을 기각**하는 오류이다. 귀무가설의 $p = \frac{1}{2}$를 적용하면 한 번에 검은 공을 뽑을 확률은 $1 - \frac{1}{2} = \frac{1}{2}$이다. 따라서 제1종 오류 확률은 $\alpha = \left(\frac{1}{2}\right)^4 = \frac{1}{16}$이다.

검정력: 검정력은 **대립가설이 참일 때 귀무가설을 기각**할 확률이다. 대립가설의 $p = \frac{1}{4}$를 적용하면 한 번에 검은 공을 뽑을 확률은 $1 - \frac{1}{4} = \frac{3}{4}$이다. 따라서 검정력은 $\left(\frac{3}{4}\right)^4 = \frac{81}{256}$이다. 이는 흰 공의 비율이 실제로 1/4로 낮아졌을 때 이 검정법이 그 변화를 탐지할 확률이다.

제1종 오류 확률 $\frac{1}{16}$, 검정력 $\frac{81}{256}$의 연결은 ④와 일치한다.

NoteBase

문 12. 다음과 같은 산점도를 갖는 자료 (x_i, y_i)에 단순선형회귀모형 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$를 최소제곱법으로 적합하여 얻은 추정 회귀직선이 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$일 때, 다음 값 중 가장 큰 것은? (단, ε_i는 $N(0, \sigma^2)$을 따르고 서로 독립이다)



① 추정 회귀직선의 기울기($\hat{\beta}_1$)

② x와 y의 표본상관계수

③ x와 $\hat{y}$의 표본상관계수

④ y와 $\hat{y}$의 표본상관계수

정답 4번

간단한 해설

산점도는 음의 선형관계를 보이지만, 최소제곱법으로 구한 **관측값과 적합값의 상관계수는 양수**이다. 이 상관계수는 x와 y의 표본상관계수의 절댓값과 같으며, 나머지 세 값은 모두 음수이다.

NoteBase

선지별 해설

산점도에서 x가 커질수록 y가 작아지고, 점들이 하나의 직선에 완전히 놓여 있지는 않다. 따라서 x와 y의 표본상관계수를 r_{xy}라 하면 $-1 < r_{xy} < 0$이다. 아래 관계는 절편을 포함한 단순선형회귀모형을 최소제곱법으로 적합했을 때 성립한다.

① 추정 회귀직선의 기울기: x와 y의 표본표준편차를 각각 s_x, s_y라 하면 $\hat{\beta}_1 = r_{xy}\frac{s_y}{s_x}$이다. 표준편차의 비는 양수이므로 기울기는 **표본상관계수와 같은 부호**를 가지며, 이 자료에서는 음수이다.

② x와 y의 표본상관계수: 산점도의 **오른쪽 아래로 향하는 경향**에 따라 음수이다. 직선 주위에 흩어진 정도는 상관관계의 강도와 관련되며, 완전한 직선관계가 아니므로 −1보다는 크다.

③ x와 $\hat{y}$의 표본상관계수: 적합값은 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$로 계산된다. 기울기가 음수이므로 x와 적합값은 **완전한 음의 선형관계**에 있고, 표본상관계수는 −1이다. 원래 관측값의 흩어짐과 달리, 적합값은 모두 추정 회귀직선 위에 놓인다.

④ y와 $\hat{y}$의 표본상관계수: 최소제곱법에서는 잔차 $e_i = y_i - \hat{y}_i$와 적합값의 표본공분산이 0이다. 따라서 $y_i = \hat{y}_i + e_i$에서 관측값과 적합값의 공분산은 적합값의 분산과 같아 **양수**이다. 구체적으로 $r_{y\hat{y}} = \sqrt{R^2} = |r_{xy}|$이며, 여기서 R^2는 결정계수이다. 이 자료에서는 $0 < r_{y\hat{y}} < 1$이다.

①·②·③은 모두 음수이고 ④는 양수이므로, 가장 큰 값은 ④의 관측값과 적합값 사이의 표본상관계수이다.

NoteBase

문 13. 어떤 작물의 생산량이 비료(A)의 4가지 수준과 토양(B)의 3가지 수준에 따라 차이가 있는지를 알아보고자 한다. 모든 수준 조합마다 동일한 반복수로 측정하여 얻은 분산분석표의 일부는 다음과 같다. 이에 대한 설명으로 옳은 것만을 모두 고르면?

요인	제곱합	자유도	평균제곱	$F-$값	$p-$값
비료(A)	145		48.3	9.67	0.0002
토양(B)	(가)				< 0.0001
교호작용 $(A\times B)$	35	(나)			0.3564
오차	120	24	5.0		
계	600				

ㄱ. (가)의 값은 300이다.

ㄴ. (나)의 값은 5이다.

ㄷ. 모든 수준 조합에서 반복수는 4이다.

① ㄱ

② ㄱ, ㄷ

③ ㄴ, ㄷ

④ ㄱ, ㄴ, ㄷ

정답 1번

간단한 해설

총제곱합은 **비료·토양·교호작용·오차의 제곱합을 더한 값**이다. 따라서 토양의 제곱합 (가)는 600 − 145 − 35 − 120 = 300이다.

NoteBase

보기별 해설

ㄱ. O: 모든 수준 조합의 반복수가 같은 이원분산분석에서는 총제곱합이 비료(A), 토양(B), 교호작용, 오차의 제곱합으로 분해된다. 표에서 총제곱합은 600이고 나머지 세 제곱합은 각각 145, 35, 120이므로, 토양의 제곱합 (가)는 $600 - 145 - 35 - 120 = 300$이다.

ㄴ. X: 비료의 수준 수를 $a = 4$, 토양의 수준 수를 $b = 3$이라 하면 교호작용의 자유도는 $(a-1)(b-1)$이다. 각 요인의 독립적인 수준 차이 수를 곱하므로 $(4-1) \times (3-1) = 6$이다. 따라서 (나)는 **5가 아니라 6**이다.

ㄷ. X: 각 수준 조합의 반복수를 r이라 하면, 조합마다 평균을 추정한 뒤 남는 오차 자유도는 $r-1$이다. 수준 조합이 $4 \times 3 = 12$개이므로 전체 오차 자유도는 $12(r-1)$이다. 표의 오차 자유도 24를 대입하면 $12(r-1) = 24$에서 $r = 3$이다. 따라서 각 수준 조합의 반복수는 **4가 아니라 3**이다.

옳은 보기는 ㄱ뿐이며, 이는 ①의 구성과 일치한다.

NoteBase

문 14. 다음은 자료 $(x_{1i}, x_{2i}, y_i)(i = 1, 2, \ldots, 20)$에 다중선형회귀모형 $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$를 최소제곱법으로 적합한 결과이다. 이 회귀모형에 대한 총제곱합(SST)은 980, 회귀제곱합(SSR)은 538일 때, 이에 대한 설명으로 옳은 것만을 모두 고르면? (단, ε_i는 $N(0, \sigma^2)$을 따르고 서로 독립이다)

계수	추정 회귀계수	표준오차	$t-$값	$p-$값
β_0	-6.053	15.663	-0.386	0.704
β_1	0.955	0.354	2.699	0.015
β_2	0.477	0.294	1.624	0.123

ㄱ. 잔차평균제곱(MSE)은 26이다.

ㄴ. 모든 설명변수는 유의수준 5%에서 유의하다.

ㄷ. 동일한 신뢰수준에서 β_1에 대한 신뢰구간의 폭이 β_2에 대한 신뢰구간의 폭보다 넓다.

① ㄱ

② ㄱ, ㄷ

③ ㄴ, ㄷ

④ ㄱ, ㄴ, ㄷ

정답 2번

간단한 해설

ㄴ에서 x_2의 회귀계수에 대한 p값은 0.123으로 0.05보다 크므로, **모든 설명변수가 유의한 것은 아니다**.

NoteBase

보기별 해설

ㄱ. O: 절편이 포함된 회귀모형에서 잔차제곱합 SSE는 총제곱합 SST에서 회귀제곱합 SSR을 뺀 값이므로, SSE = 980 − 538 = 442이다. 관측값은 20개이고 절편과 두 기울기를 추정하므로 **잔차 자유도는 20 − 3 = 17**이다. 잔차평균제곱은 잔차제곱합을 잔차 자유도로 나눈 값이므로 MSE = $\frac{442}{17}$ = 26이다.

ㄴ. X: 각 설명변수의 유의성은 다른 설명변수를 모형에 둔 상태에서 해당 회귀계수가 0인지 검정하여 판단한다. 표의 p값을 유의수준 0.05와 비교하면, β_1은 0.015 < 0.05이므로 유의하지만 β_2는 0.123 > 0.05이므로 유의하지 않다. 따라서 **설명변수 중 유의한 것은 x_1뿐**이며, 모든 설명변수가 유의하다는 진술은 성립하지 않는다.

ㄷ. O: 주어진 독립 정규오차 모형에서 회귀계수 β_j의 양측 신뢰구간은 $\hat{\beta}_j \pm t_{1-\alpha/2,\,17}\mathrm{SE}(\hat{\beta}_j)$이다. 여기서 $\mathrm{SE}(\hat{\beta}_j)$는 추정 회귀계수의 표준오차이고, t 임계값은 잔차 자유도 17과 신뢰수준 $1-\alpha$로 정해진다. 신뢰구간의 폭은 $2t_{1-\alpha/2,\,17}\mathrm{SE}(\hat{\beta}_j)$이므로 **같은 신뢰수준과 자유도에서는 표준오차가 클수록 폭이 넓다**. 표에서 $\hat{\beta}_1$의 표준오차는 0.354, $\hat{\beta}_2$의 표준오차는 0.294이므로 β_1의 신뢰구간이 더 넓다.

따라서 옳은 보기는 ㄱ, ㄷ이며, ②의 조합과 일치한다.

NoteBase

문 15. 자격증 시험의 응시자 중 100명을 임의추출하여 온라인 강의 수강과 자격증 시험 합격 여부를 조사한 결과는 다음과 같다. 온라인 강의 수강 여부와 자격증 시험 합격 여부는 연관성이 없다는 귀무가설을 유의수준 10%에서 검정하고자 할 때, 카이제곱 검정통계량의 값과 검정 결과를 바르게 연결한 것은? (단, $\chi^2_\alpha(k)$는 자유도가 k인 카이제곱분포의 제$100 \times (1-\alpha)$ 백분위수를 나타내고, $\chi^2_{0.1}(1) = 2.71$, $\chi^2_{0.1}(3) = 6.25$ 이다)

온라인 강의 수강 여부 \ 합격 여부	합격	불합격	계
수강함	40	20	60
수강하지 않음	30	10	40
계	70	30	100

① 카이제곱 검정통계량의 값: $\frac{50}{63}$, 검정 결과: 귀무가설을 기각함

② 카이제곱 검정통계량의 값: $\frac{50}{63}$, 검정 결과: 귀무가설을 기각하지 않음

③ 카이제곱 검정통계량의 값: $\frac{200}{63}$, 검정 결과: 귀무가설을 기각함

④ 카이제곱 검정통계량의 값: $\frac{200}{63}$, 검정 결과: 귀무가설을 기각하지 않음

정답 2번

간단한 해설

독립을 가정한 기대도수와 관측도수의 차이를 이용하면 카이제곱 검정통계량은 $\frac{50}{63} \approx 0.794$이다. 자유도 1에서 유의수준 10%의 임계값 2.71보다 작으므로 귀무가설을 **기각하지 않는다**.

NoteBase

계산 과정

1. 검정의 대상과 관측도수: 귀무가설은 온라인 강의 수강 여부와 시험 합격 여부가 서로 독립이라는 것이고, 대립가설은 두 변수가 연관되어 있다는 것이다. 표에서 수강자의 합격·불합격 인원은 각각 40명·20명이고, 미수강자는 각각 30명·10명이다. 행 합계는 60명·40명, 열 합계는 70명·30명이며 전체 인원은 100명이다.

2. 기대도수: 두 변수가 독립이면 각 집단에서도 전체 합격·불합격 비율이 적용된다. 따라서 각 셀의 기대도수는 **행 합계와 열 합계의 곱을 전체 인원수로 나눈 값**이다. 수강자의 합격 기대도수는 $\frac{60 \cdot 70}{100} = 42$명, 불합격 기대도수는 $\frac{60 \cdot 30}{100} = 18$명이다. 미수강자의 합격 기대도수는 $\frac{40 \cdot 70}{100} = 28$명, 불합격 기대도수는 $\frac{40 \cdot 30}{100} = 12$명이다. 모두 5 이상이므로 카이제곱 근사를 적용하기에 충분하다.

3. 검정통계량: 각 셀의 관측도수를 O, 기대도수를 E라 하면 피어슨 카이제곱 검정통계량은 $\chi^2 = \sum \frac{(O-E)^2}{E}$이다. 합은 네 셀 전체에 대해 구하며, 관측도수와 기대도수의 차이가 클수록 독립 가정과의 불일치가 커진다. 위 순서대로 차이는 −2명, 2명, 2명, −2명이므로 차이의 제곱은 모두 4이다. 따라서 $\chi^2 = 4\left(\frac{1}{42} + \frac{1}{18} + \frac{1}{28} + \frac{1}{12}\right)$이고, 이를 합하면 $\chi^2 = \frac{50}{63} \approx 0.794$이다. 검정통계량에는 단위가 없다.

4. 자유도와 검정 결과: 행과 열이 각각 두 범주인 분할표이므로 자유도는 $(2-1)(2-1) = 1$이다. 독립 가정과의 불일치가 큰 경우에 기각하므로 **검정통계량이 상측 임계값보다 큰지** 비교한다. 유의수준 10%에서 임계값은 $\chi^2_{0.1}(1) = 2.71$이며, $\frac{50}{63} \approx 0.794 < 2.71$이므로 귀무가설을 기각하지 않는다. 이는 ②의 연결과 일치한다. 두 변수 사이에 연관성이 있다는 증거가 충분하지 않다는 뜻이며, 독립이 입증되었다는 뜻은 아니다.

NoteBase

문 16. 두 철강 공장 A, B에서 생산하는 소재의 인장강도는 모분산이 각각 σ_A^2, σ_B^2인 정규분포를 따른다고 한다. 두 철강 공장 A, B의 소재를 각각 13개, 10개 임의추출하여 구한 인장강도의 표본분산은 각각 $s_A^2 = 27, s_B^2 = 9$이다. 두 공장에서 생산하는 소재의 인장강도가 서로 독립일 때, 두 모분산의 비 $\frac{\sigma_A^2}{\sigma_B^2}$에 대한 95 % 신뢰구간은? (단, $F_\alpha(k_1, k_2)$는 분자와 분모의 자유도가 각각 k_1, k_2인 F분포의 제$100 \times (1-\alpha)$ 백분위수를 나타내고, $F_{0.025}(12, 9) = 3.87,\ F_{0.025}(9, 12) = 3.44$이다)

① $\left(\frac{3}{3.87},\ 3 \times 3.44\right)$

② $\left(\frac{3}{3.87},\ 3 \times 3.87\right)$

③ $\left(\frac{3}{3.44},\ 3 \times 3.44\right)$

④ $\left(\frac{3}{3.44},\ 3 \times 3.87\right)$

정답 1번

간단한 해설

표본분산의 비는 3이며, 모분산 비의 신뢰구간은 이를 **F분포의 양쪽 임계값으로 나누어** 구한다. 아래쪽 임계값은 자유도를 바꾼 위쪽 임계값의 역수이므로, 95% 신뢰구간은 $\left(\frac{3}{3.87},\ 3 \cdot 3.44\right)$이다.

계산 과정

1. 표본분산의 비와 자유도: 구하려는 모분산 비를 $r = \frac{\sigma_A^2}{\sigma_B^2}$라 두면, 표본분산의 비는 $\frac{s_A^2}{s_B^2} = \frac{27}{9} = 3$이다. 두 표본은 서로 독립인 정규모집단에서 추출되었으므로, 각 표본분산을 모분산으로 나눈 값의 비는 F분포를 따른다. 따라서 $T = \frac{s_A^2/s_B^2}{r}$는 분자 자유도 $13 - 1 = 12$, 분모 자유도 $10 - 1 = 9$인 F분포를 따른다.

2. 양쪽 임계값: 95% 신뢰구간은 양쪽 꼬리에 각각 2.5%를 남겨 구한다. 문제의 표기에서 위쪽 임계값은 $F_{0.025}(12, 9) = 3.87$이다. F분포는 **역수를 취하면 분자·분모의 자유도가 서로 바뀌므로**, 아래쪽 임계값은 $F_{0.975}(12, 9) = \frac{1}{F_{0.025}(9, 12)}$, 즉 $\frac{1}{3.44}$이다.

3. 모분산 비에 대해 정리: 중앙 95%에 해당하는 부등식에 관측된 표본분산의 비 3을 대입하면 $\frac{1}{3.44} \le \frac{3}{r} \le 3.87$이다. 모분산 비는 양수이므로 이를 r에 대해 풀면 $\frac{3}{3.87} \le r \le 3 \cdot 3.44$가 된다. **큰 임계값으로 나눈 값이 하한**이고, 작은 임계값으로 나눈 값이 상한이다.

따라서 95% 신뢰구간은 $\left(\frac{3}{3.87},\ 3 \cdot 3.44\right)$로, ①의 구간과 일치한다. 같은 인장강도에 대한 두 분산의 비이므로 단위는 없다.

NoteBase

문 17. 세 그룹 A, B, C에서 각각 4개, 5개, 6개의 관측값을 조사한 결과 그룹별 표본평균과 표본분산은 다음과 같다. 그룹에 따라 평균의 차이가 있는지를 알아보기 위한 분산분석에서 오차제곱합과 그 자유도를 바르게 연결한 것은?

그룹	자료의 수	표본평균	표본분산
A	4	19	7.0
B	5	10	9.5
C	6	14	8.0

① 오차제곱합: 99, 자유도: 12

② 오차제곱합: 99, 자유도: 14

③ 오차제곱합: 180, 자유도: 12

④ 오차제곱합: 180, 자유도: 14

정답 1번

간단한 해설

오차제곱합은 **각 그룹 내부의 편차제곱합을 더한 값**이므로 표본분산에 각 그룹의 자료 수에서 1을 뺀 값을 곱해 합산한다. 이 값은 99이고, 자유도는 전체 자료 수 15에서 그룹 수 3을 뺀 12이다.

계산 과정

오차제곱합은 각 관측값과 그 관측값이 속한 **그룹의 표본평균 사이 편차**를 제곱하여 모두 더한 값이다. 그룹 i의 자료 수를 n_i, 표본분산을 s_i^2라 하면, 표본분산은 그룹 내부의 편차제곱합을 $n_i - 1$로 나눈 값이다. 따라서 각 그룹의 편차제곱합은 $(n_i - 1)s_i^2$이고, 세 그룹의 오차제곱합은 $SSE = \sum_{i=1}^{3} (n_i - 1)s_i^2$이다.

표의 자료 수와 표본분산을 이용하면 A의 편차제곱합은 $(4 - 1) \times 7.0 = 21$, B는 $(5 - 1) \times 9.5 = 38$, C는 $(6 - 1) \times 8.0 = 40$이다. 이들은 각 그룹 내부의 변동량이며, 합산하면 오차제곱합은 $21 + 38 + 40 = 99$이다.

각 그룹에서는 표본평균을 기준으로 한 편차의 합이 0이므로, **그룹마다 자유도가 하나씩 줄어든다**. 따라서 오차 자유도는 $\sum_{i=1}^{3} (n_i - 1)$이며, $(4 - 1) + (5 - 1) + (6 - 1) = 12$이다. 전체 자료 수에서 그룹 수를 빼도 $15 - 3 = 12$로 같다. 14는 전체 자료를 하나의 평균 기준으로 다루는 총제곱합의 자유도이다.

오차제곱합 99와 오차 자유도 12의 연결은 ①에 해당한다.

NoteBase

문 18. 5명의 학생을 대상으로 특별 교육 프로그램 이수 전후의 학업성취도를 측정한 결과는 다음과 같다. 이수 전후 학업성취도의 평균이 같은지를 검정하기 위한 t-검정통계량의 절댓값은? (단, 이수 전후 학업성취도 차이는 서로 독립이고, 정규분포를 따른다)

학생	A	B	C	D	E
이수 전	43	35	32	18	22
이수 후	41	41	42	32	44

① 2.0

② 2.5

③ 3.0

④ 3.5

정답 2번

간단한 해설

같은 학생의 이수 전후 점수이므로 **학생별 차이에 대한 대응표본 t-검정**을 한다. 이수 후에서 이수 전을 뺀 차이의 평균은 10, 그 평균의 표준오차는 4이므로 검정통계량의 절댓값은 $\frac{10}{4} = 2.5$이다.

NoteBase

계산 과정

1. 검정 방법과 가설: 같은 학생을 두 번 측정했으므로 **학생별 이수 전후 차이**를 하나의 관측값으로 사용한다. 학생 i의 차이를 d_i=이수 후 점수-이수 전 점수로 두고, 차이의 모평균을 μ_d라 하면 검정할 가설은 $H_0 : \mu_d = 0$, $H_1 : \mu_d \neq 0$이다. 차이들이 서로 독립이고 정규분포를 따른다는 조건에서, 모분산을 표본분산으로 추정하는 대응표본 t-검정을 적용한다. 표본크기는 n = 5이며 귀무가설 아래 자유도는 5 − 1 = 4이다.

2. 차이의 표본평균: 표의 A부터 E까지 이수 후 점수에서 이수 전 점수를 빼면 차이는 −2, 6, 10, 14, 22이다. 합은 50이므로 차이의 표본평균은 $\bar{d} = \frac{50}{5} = 10$이다. 이는 학생들의 점수가 평균적으로 10만큼 높아졌다는 뜻이다.

3. 차이의 표본분산: 각 차이에서 평균 10을 빼면 −12, −4, 0, 4, 12이고, 편차제곱합은 144 + 16 + 0 + 16 + 144 = 320이다. 표본평균을 이용했으므로 편차제곱합을 **n − 1로 나누어** 표본분산을 구한다. 따라서 $s_d^2 = \frac{320}{5-1} = 80$이다.

4. 평균 차이의 표준오차: 검정통계량의 분모에는 개별 차이의 표준편차가 아니라 **차이의 표본평균에 대한 표준오차**를 사용한다. 그 값은 (s_d)/(n)=(80)/(5)=4이다.

5. 검정통계량: 귀무가설에서 평균 차이는 0이므로 t=(d-0)/(s_d/ n)=(10)/(4)=2.5이다. 따라서 검정통계량의 절댓값은 2.5로, ②의 값과 일치한다. 점수 차이를 그 표준오차로 나눈 값이므로 검정통계량에는 단위가 없다.

NoteBase

문 19. 1부터 8까지의 숫자가 각각 하나씩 적혀 있는 8장의 카드가 있다. 이 중에서 3장의 카드를 임의로 비복원추출할 때, 선택된 3장의 카드에 적혀 있는 수 중에서 가장 작은 값을 X라 하자. X의 기댓값 $E(X)$는?

① $\frac{3}{2}$
② $\frac{7}{4}$
③ 2
④ $\frac{9}{4}$

정답 4번

간단한 해설

최솟값이 k이려면 **카드 k를 포함하고 나머지 두 장은 k보다 큰 수에서 골라야 한다**. 이 조건으로 각 최솟값의 확률을 구해 가중평균하면 $E(X)=\frac{9}{4}$이다.

계산 과정

1. 추출 경우의 수와 범위: 비복원추출이므로 서로 다른 카드 3장을 고르며, 순서를 무시한 모든 조합은 같은 확률로 선택된다. 전체 경우의 수는 $\binom{8}{3}=56$이다. 최솟값이 6일 때는 $6\cdot7\cdot8$을 고를 수 있지만, 7 이상이면 그보다 큰 카드가 두 장 남지 않는다. 따라서 X가 가질 수 있는 값은 1부터 6까지이다.

2. 최솟값의 확률: $X=k$가 되려면 k**는 반드시 고르고 나머지 두 장은** $k+1$**부터 8까지에서 골라야 한다**. 따라서 $P(X=k)=\frac{\binom{8-k}{2}}{56}$이고, $k=1,2,\ldots,6$이다. 각 최솟값에 해당하는 경우의 수는 차례로 21, 15, 10, 6, 3, 1이다.

3. 기댓값의 가중합: 기댓값은 각 값에 그 값이 나올 확률을 곱해 더한 $E(X)=\sum_{k=1}^{6} kP(X=k)$이다. 각 경우의 수에 최솟값을 곱하면 차례로 21, 30, 30, 24, 15, 6이므로, 가중합의 분자는 $21+30+30+24+15+6=126$이다.

4. 결과: 전체 56개 조합에 대한 최솟값의 평균은 $E(X)=\frac{126}{56}=\frac{9}{4}$이다. 이는 선택지 ④의 값과 일치한다.

NoteBase

문 20. 자유도가 10인 t분포를 따르는 확률변수 T가 평균으로부터 사분위수 범위의 m배 이상 떨어져 있는 값을 극단값으로 정의하자. 확률변수 T가 극단값을 가질 확률이 0.02 이하가 되도록 하기 위한 m의 최솟값을 m_0이라 할 때, $14m_0$의 값은? (단, $t_\alpha(k)$는 자유도가 k인 t분포의 제$100 \times (1-\alpha)$ 백분위수를 나타내고, $t_{0.25}(10) = 0.70$, $t_{0.02}(10) = 2.36$, $t_{0.01}(10) = 2.76$이다)

① 23.6

② 27.6

③ 47.2

④ 55.2

정답 2번

간단한 해설

사분위수 범위는 $0.70 - (-0.70) = 1.40$이다. 극단값 확률 0.02는 **양쪽 꼬리확률의 합**이므로 한쪽 꼬리확률이 0.01인 경계값 2.76을 사용하며, $1.40m_0 = 2.76$에서 $14m_0 = 27.6$을 얻는다.

계산 과정

자유도가 10인 t분포는 0을 중심으로 대칭이며 평균도 0이다. 제1사분위수 Q_1과 제3사분위수 Q_3는 각각 누적확률이 0.25, 0.75인 값이다. 주어진 $t_{0.25}(10) = 0.70$과 대칭성에 따라 $Q_1 = -0.70$, $Q_3 = 0.70$이다. 따라서 사분위수 범위는 $Q_3 - Q_1 = 0.70 - (-0.70) = 1.40$이다.

평균으로부터 사분위수 범위의 m배 이상 떨어진다는 조건은 $|T| \geq 1.40m$이다. 대칭성에 의해 $P(|T| \geq 1.40m) = 2P(T \geq 1.40m)$이므로, 이 확률이 0.02 이하가 되려면 **한쪽 꼬리확률이 0.01 이하**여야 한다. 주어진 $t_{0.01}(10) = 2.76$은 오른쪽 꼬리확률이 0.01인 경계값이다.

경계값이 커질수록 꼬리확률이 감소하므로, 조건을 만족하는 최소 배수는 $1.40m_0 = 2.76$일 때 얻어진다. 따라서 $m_0 = \frac{2.76}{1.40}$이고, $14m_0 = 14 \times \frac{2.76}{1.40} = 27.6$이다. 이는 ②의 27.6에 해당한다.

NoteBase

문 21. 두 확률변수 X, Y의 결합확률밀도함수가 다음과 같을 때, $P\left(X > \frac{1}{2}\right)$의 값은?

f(x,y)=cases6(1-x-y), & x>0, y>0, x+y<1\\0, & 그 외cases

① $\frac{1}{8}$
② $\frac{1}{6}$
③ $\frac{1}{4}$
④ $\frac{1}{2}$

정답 1번

간단한 해설

결합확률밀도함수가 양수인 영역에서 $X > \frac{1}{2}$에 해당하는 부분을 적분한다. 이때 **주어진 x에 따라 y의 상한이 1 − x로 달라지며**, 계산한 확률은 $\frac{1}{8}$이다.

계산 과정

적분 범위: 밀도가 양수인 조건은 $x > 0,\ y > 0,\ x + y < 1$이다. 따라서 구하려는 사건에서는 $\frac{1}{2} < x < 1$이고, 각 x에서 **y의 범위는** $0 < y < 1 - x$이다. 이 영역에 걸쳐 결합확률밀도함수를 적분하면 해당 사건의 확률을 얻는다.

X의 주변확률밀도함수: Y에는 별도의 제한이 없으므로, 주어진 x에서 가능한 모든 y에 대해 밀도를 적분한다. $0 < x < 1$에서 $f_X(x) = \int_0^{1-x} 6(1 - x - y)dy$이다. y에 대한 원시함수는 $6(1 - x)y - 3y^2$이므로, 양 끝값을 대입하면 $f_X(x) = 3(1 - x)^2$이다. 이는 Y의 값을 모두 합쳐 얻은 X만의 확률밀도이다.

확률 계산: 구한 주변확률밀도함수를 x가 1/2보다 큰 구간에서 적분한다. $P(X > \frac{1}{2}) = \int_{1/2}^{1} 3(1 - x)^2 dx$이다. 원시함수는 $-(1 - x)^3$이므로, $P(X > \frac{1}{2}) = 0 - (-\frac{1}{8}) = \frac{1}{8}$이다. 이 값은 선택지 ①과 일치한다.

NoteBase

문 22. 네 가지 약품 A, B, C, D에 대해 각각 3회씩 반복하여 측정한 약효는 다음과 같다. 약품에 따라 약효의 차이가 있는지를 검정하기 위한 F-검정통계량의 값은?

약품 A	약품 B	약품 C	약품 D
3	2	0	-1
2	1	-1	-2
1	0	-2	-3

① 1

② 8

③ 10

④ 30

정답 3번

간단한 해설

F-검정통계량은 **집단 간 평균제곱을 집단 내 평균제곱으로 나눈 값**이다. 이 자료에서는 두 평균제곱이 각각 10과 1이므로 $F = \frac{10}{1} = 10$이다.

NoteBase

계산 과정

검정의 설정: 약품이라는 한 요인에 따른 네 집단의 평균을 비교하므로 일원분산분석을 적용한다. 귀무가설은 네 약품의 모평균이 모두 같다는 것이다. 통상적인 F-검정은 관측값의 독립성, 각 집단의 정규성, 집단 간 등분산성을 가정한다. 집단 수 k = 4, 집단별 관측 수는 3, 전체 관측 수 N = 12이다.

평균 계산: 표의 각 열에서 A는 3, 2, 1이므로 평균이 2이고, B는 2, 1, 0이므로 평균이 1이다. C는 0, −1, −2이므로 평균이 −1이고, D는 −1, −2, −3이므로 평균이 −2이다. 집단별 관측 수가 같으므로 전체 평균은 $\bar{x} = \frac{2 + 1 - 1 - 2}{4} = 0$이다.

집단 간 변동: 집단 간 제곱합은 각 집단 평균과 전체 평균의 차이를 제곱한 뒤 **해당 집단의 관측 수를 곱하여** 더한다. 평균 차이의 제곱은 A부터 차례로 4, 1, 1, 4이므로 $SS_B = 3(4 + 1 + 1 + 4) = 30$이다. 집단 간 자유도는 k − 1 = 3이므로 집단 간 평균제곱은 $MS_B = \frac{30}{3} = 10$이다.

집단 내 변동: 집단 내 제곱합은 각 관측값이 **자기 집단 평균에서 벗어난 정도**의 제곱을 모두 더한 값이다. 네 집단 모두 평균에 대한 편차가 1, 0, −1이므로 각 집단의 제곱합은 $1^2 + 0^2 + (-1)^2 = 2$이다. 따라서 오차제곱합은 $SS_E = 4 \cdot 2 = 8$이고, 집단 내 자유도는 N − k = 8이다. 오차평균제곱은 $MS_E = \frac{8}{8} = 1$로, 약품별 평균 차이를 제외한 집단 내부의 변동을 나타낸다.

F값 계산: 집단 간 변동이 집단 내부의 오차 변동에 비해 얼마나 큰지를 비교하기 위해 두 평균제곱의 비를 구한다. $F = \frac{MS_B}{MS_E} = \frac{10}{1} = 10$이며, 분자와 분모의 단위가 같아 F값에는 단위가 없다. 계산값 10은 ③의 값과 일치한다.

NoteBase

문 23. 자료 (x_i, y_i) $(i = 1, 2, \cdots, n)$에 단순선형회귀모형 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$를 최소제곱법으로 적합하여 얻은 β_0의 추정값, 회귀제곱합, 잔차평균제곱을 각각 $\hat{\beta}_0$, SSR, MSE라 하자. 설명변수를 $x_i^* = 2x_i$로 변환한 자료 (x_i^*, y_i) $(i = 1, 2, \cdots, n)$에 단순선형회귀모형 $y_i = \beta_0^* + \beta_1^* x_i^* + \varepsilon_i^*$를 최소제곱법으로 적합하여 얻은 β_0^*의 추정값, 회귀제곱합, 잔차평균제곱을 각각 $\hat{\beta}_0^*$, SSR^*, MSE^*라 하자. 이에 대한 설명으로 옳은 것만을 모두 고르면? (단, ε_i, ε_i^*는 각각 $N(0, \sigma^2)$, $N(0, (\sigma^*)^2)$을 따르고 서로 독립이다)

ㄱ. $\hat{\beta}_0 = \hat{\beta}_0^*$
ㄴ. $SSR = SSR^*$
ㄷ. $MSE = MSE^*$

① ㄱ
② ㄱ, ㄴ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 4번

간단한 해설

설명변수를 2배로 해도 기울기가 절반으로 조정되어 **적합값과 잔차가 그대로 유지된다**. 따라서 절편과 회귀제곱합은 바뀌지 않으며, 잔차제곱합과 자유도가 같아 잔차평균제곱도 변하지 않는다.

보기별 해설

설명변수를 $x_i^* = 2x_i$로 바꾸면 $a + bx_i = a + \frac{b}{2}x_i^*$이다. 절편은 그대로 두고 기울기를 절반으로 조정하면 변환 전후에 같은 적합값을 만들 수 있다. 따라서 두 모형에서 최소화하는 잔차제곱합과 최소제곱 적합값은 같다.

ㄱ. O: 기울기 추정값은 $\hat{\beta}_1^* = \frac{\hat{\beta}_1}{2}$이고 설명변수의 평균은 $\bar{x}^* = 2\bar{x}$이다. 절편 추정값은 $\hat{\beta}_0^* = \bar{y} - \hat{\beta}_1^* \bar{x}^*$로 구한다. 여기서 $\hat{\beta}_1^* \bar{x}^* = \hat{\beta}_1 \bar{x}$이므로 **절편 추정값은 변하지 않는다**. 즉, $\hat{\beta}_0^* = \hat{\beta}_0$이다.

ㄴ. O: 회귀제곱합은 적합값이 반응변수의 평균에서 얼마나 떨어져 있는지를 나타내며, $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$이다. 변환 전후에 **적합값과 반응변수의 평균이 모두 같으므로** $SSR^* = SSR$이다.

ㄷ. O: 잔차는 관측값과 적합값의 차이이다. 두 모형에서 잔차가 같으므로 잔차제곱합 $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$도 같다. 두 모형 모두 같은 n개 관측값으로 절편과 기울기 두 모수를 추정하므로 **잔차 자유도도 n − 2로 같다**. 따라서 $MSE = \frac{SSE}{n-2}$도 변하지 않아 $MSE^* = MSE$이다.

ㄱ, ㄴ, ㄷ이 모두 성립하므로 이들을 모두 포함하는 조합은 ④이다.

NoteBase

문 24. 어느 공장에서 생산하는 반도체 웨이퍼 1장당 결함 수는 평균이 0.5인 포아송 분포를 따르고, 평균이 $m(m>0)$인 포아송 분포의 확률질량함수는 다음과 같다. 웨이퍼당 결함 수가 서로 독립일 때, 갑과 을의 대화에서 (가), (나)에 들어갈 값을 바르게 연결한 것은? (단, 확률변수 Z가 표준정규분포를 따를 때 $P(Z \geq 1) = 0.1587$, $P(Z \geq 2) = 0.0228$이다)

f(x)=cases (e^-mm^x)/(x!), & x=0,1,2,⋯\ 0, & 그 외 cases
갑: 포아송 분포의 성질을 이용하면 웨이퍼 5장에서 총 결함이 2개 이상일 확률은 (가)임을 알 수 있어.
을: 나는 50장의 웨이퍼를 고려해 보겠어. 중심극한정리를 이용하면 50장의 웨이퍼에서 총 결함이 30개 이상일 확률은 (나)로 근사할 수 있어.

① (가): $1 - e^{-0.5}$, (나): 0.0228

② (가): $1 - e^{-0.5}$, (나): 0.1587

③ (가): $1 - 3.5e^{-2.5}$, (나): 0.0228

④ (가): $1 - 3.5e^{-2.5}$, (나): 0.1587

정답 4번

간단한 해설

독립인 포아송 확률변수의 합도 포아송 분포를 따르므로, 5장의 총 결함 수는 **평균 2.5**를 갖고 (가)는 $1 - 3.5e^{-2.5}$이다. 50장의 총 결함 수는 평균 25, 표준편차 5이므로, 30개 이상일 확률은 **평균보다 1표준편차 이상 클 확률**인 0.1587로 근사한다.

NoteBase

계산 과정

웨이퍼 한 장의 결함 수를 X_i, n장의 총 결함 수를 $S_n = \sum_{i=1}^{n} X_i$라 하자. 각 X_i는 평균과 분산이 모두 0.5인 포아송 확률변수이다. **서로 독립인 포아송 확률변수의 합**도 포아송 분포를 따르며, 그 평균과 분산은 각각 더해지므로 S_n의 평균과 분산은 모두 $0.5n$이다.

(가): 5장의 총 결함 수 S_5는 평균 2.5인 포아송 분포를 따른다. '2개 이상'의 여사건은 **0개 또는 1개**이므로 P(S_5≥2)=1-\P(S_5=0)+P(S_5=1)\이다. 주어진 확률질량함수에 평균 2.5를 대입하면 $P(S_5 = 0) = e^{-2.5}$, $P(S_5 = 1) = 2.5e^{-2.5}$이다. 따라서 P(S_5≥2)=1-3.5e^-2.5이다.

(나): 50장의 총 결함 수 S_{50}는 평균 25, 분산 25를 가지므로 표준편차는 $\sqrt{25} = 5$이다. 독립이고 동일한 분포를 따르는 50개 결함 수의 합에 중심극한정리를 적용하면, **평균을 빼고 표준편차로 나눈 값** $\frac{S_{50} - 25}{5}$는 표준정규분포로 근사된다. 연속성 보정 없이 30을 그대로 표준화하면 $\frac{30 - 25}{5} = 1$이다. 따라서 P(S_50≥30)≈P(Z≥1)=0.1587이다.

(가)는 $1 - 3.5e^{-2.5}$, (나)는 0.1587이므로 두 값을 연결한 선택지는 ④이다.

NoteBase

문 25. 금융상품 A, B의 월간 수익률을 각각 확률변수 R_A, R_B라 하자. 금융상품 A, B에 분산 투자하였을 때의 수익률 R_P는 다음과 같다. R_A, R_B의 분산과 공분산이 다음과 같을 때, R_P의 분산 $Var(R_P)$를 최소로 하는 상수 a의 값은?

R_P=aR_A+(1-a)R_B(단, 0≤ a≤ 1)
$Var(R_A) = 5, \quad Var(R_B) = 20, \quad Cov(R_A, R_B) = -5$

① $\frac{2}{7}$
② $\frac{3}{7}$
③ $\frac{4}{7}$
④ $\frac{5}{7}$

정답 4번

간단한 해설

투자 수익률의 분산은 **공분산 항까지 포함한 이차식**이다. 이를 완전제곱식으로 정리하면 $35(a-\frac{5}{7})^2+\frac{15}{7}$ 이므로, 제곱항이 0이 되는 $a=\frac{5}{7}$에서 최소가 된다.

계산 과정

a는 금융상품 A의 투자 비중이고, $1-a$는 B의 투자 비중이다. 주어진 조건 0≤ a≤1에서 월간 수익률 $R_P = aR_A + (1-a)R_B$의 분산을 최소화한다.

두 확률변수의 합의 분산은 각 변수의 분산에 두 변수의 공분산의 두 배를 더한 값이다. 상수를 곱하면 분산에는 그 상수의 제곱이, 공분산에는 두 상수의 곱이 반영된다. 따라서 각 가중 수익률의 분산은 $5a^2$, $20(1-a)^2$이고, **공분산에 따른 항**은 $2a(1-a)(-5) = -10a(1-a)$이다.

이 세 항을 합쳐 정리하면 다음과 같다.
$Var(R_P) = 35a^2 - 50a + 20$
$= 35(a-\frac{5}{7})^2 + \frac{15}{7}$

완전제곱식에서 제곱항은 항상 0 이상이고 그 계수 35가 양수이므로, **제곱항이 0이 될 때** 분산이 가장 작다. 따라서 $a=\frac{5}{7}$이며, 이 값은 0≤ a≤1을 만족한다. A에 전체 투자금의 5/7, B에 2/7를 투자하는 비중으로, 선택지 ④와 일치한다.

NoteBase