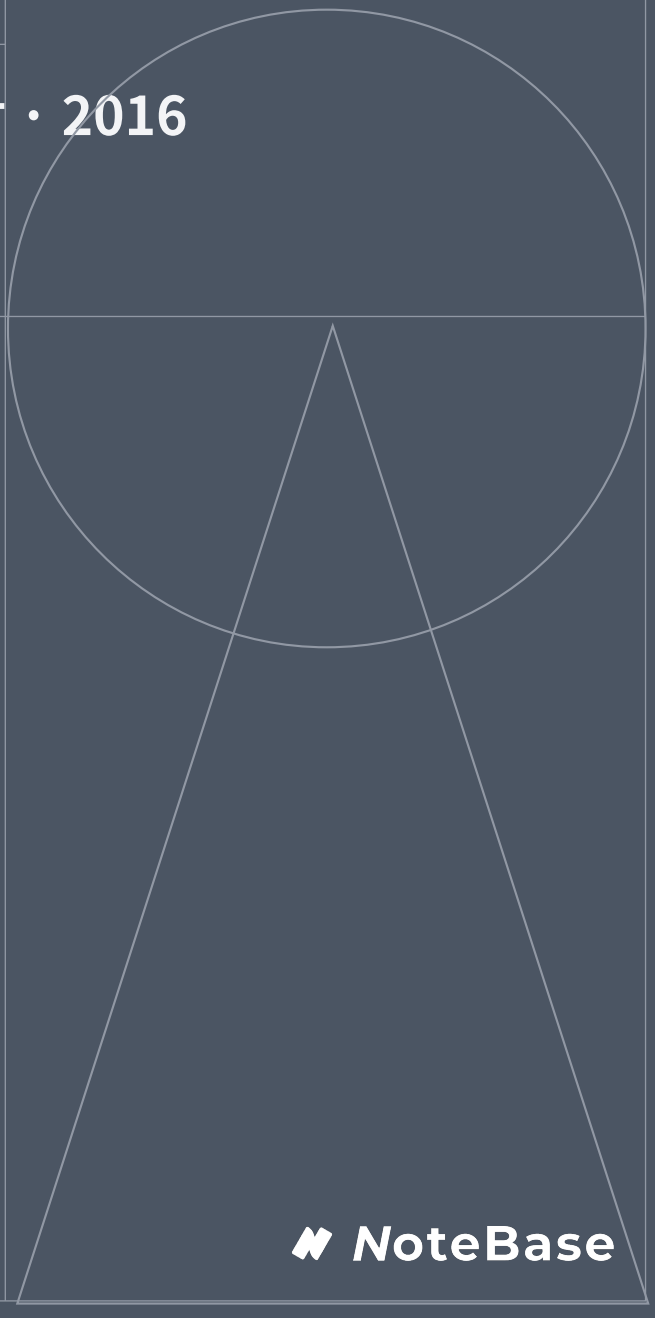
기출OX 해설집

공무원/국가직/7급 · 통계학 · 2016



NoteBase

기출OX 해설집

공무원/국가직/7급 · 통계학 · 2016

공무원/국가직/7급 · 통계학 · 2016

문 1. 구간 (-1, 3)에서 균일분포(uniform distribution)를 따르는 확률변수의 중앙값(median)과 사분위범위(interquartile range)는?

① 중앙값: 0, 사분위범위: 2

② 중앙값: 0, 사분위범위: 3

③ 중앙값: 1, 사분위범위: 2

④ 중앙값: 1, 사분위범위: 3

정답 3번

출제 포인트

이 문제는 균일분포 $U(-1, 3)$의 중앙값과 사분위수를 이용해 사분위범위를 계산하는 문제이다.

조건 정리

주어진 분포는 $U(a, b)$이고 $a = -1, b = 3$이다.
구해야 할 값은 중앙값과 사분위범위이다.
균일분포는 구간 전체에 동일한 밀도를 가지므로 분위수는 구간을 같은 비율로 나누어 구한다.

사용할 공식

균일분포 $U(a, b)$의 p분위수는 $Q_p = a + p(b - a)$이다.
중앙값은 $Q_{0.5} = (a + b)/2$이고, 사분위범위는 $Q_3 - Q_1 = Q_{0.75} - Q_{0.25}$이다.
이 공식을 쓰는 이유는 균일분포의 누적분포함수가 구간 $[a, b]$에서 선형이어서 분위수가 구간 길이의 비율만큼 이동한 값으로 정해지기 때문이다.

계산 과정

1. 구간 길이는 $b - a = 3 - (-1) = 4$이다.
2. 중앙값은 $Q_{0.5} = -1 + 0.5 \times 4 = 1$이다.
3. 제1사분위수는 $Q_{0.25} = -1 + 0.25 \times 4 = 0$이다.
4. 제3사분위수는 $Q_{0.75} = -1 + 0.75 \times 4 = 2$이다.
5. 사분위범위는 $Q_3 - Q_1 = 2 - 0 = 2$이다.

NoteBase

선지별 해설

1번 X: 사분위범위 2는 맞지만 중앙값 0이 틀리다. 균일분포 $U(-1, 3)$의 중앙값은 구간의 한가운데인 1이다. 틀린 표현은 '중앙값 0'이고, 올바르게는 '중앙값 1'이다.
2번 X: 중앙값 0도 틀리고 사분위범위 3도 틀리다. 중앙값은 1이고 사분위범위는 2이다. 틀린 표현은 '중앙값 0, 사분위범위 3'이고, 올바르게는 '중앙값 1, 사분위범위 2'이다.
3번 O: 중앙값 1, 사분위범위 2가 계산 결과와 정확히 일치한다.
4번 X: 중앙값 1은 맞지만 사분위범위 3이 틀리다. 3은 사분위범위가 아니라 제3사분위수 Q_3와 혼동한 값이다. 틀린 표현은 '사분위범위 3'이고, 올바르게는 '사분위범위 2'이다.

오답 포인트

사분위범위는 Q_3 자체가 아니라 $Q_3 - Q_1$이라는 점이 함정이다. 이 문항에서는 $Q_3 = 2$, $Q_1 = 0$인데 3이라는 값은 구간 끝점이나 분위수 위치를 잘못 섞어 계산할 때 나오기 쉽다.

다시 나오면 이렇게 판단

균일분포 문항은 먼저 구간 끝점 a, b를 확인하고, $Q_p = a + p(b - a)$로 중앙값과 사분위수를 구한 뒤, 사분위범위는 반드시 $Q_3 - Q_1$로 마무리하면 된다.

암기 포인트

균일분포 $U(a, b)$에서는 중앙값이 중점 $(a + b)/2$이고, 사분위범위는 구간 길이의 절반 $(b - a)/2$이다.

NoteBase

문 2. $X_1, \cdots, X_n$이 평균 μ, 분산 σ^2인 정규모집단에서의 임의표본(random sample)일 때, $\sum_{i=1}^{n}(X_i-\mu)^2$의 평균과 분산은? (단, $n \geq 2$)

① 평균: $(n-1)\sigma^2$, 분산: $2(n-1)\sigma^2$

② 평균: $(n-1)\sigma^2$, 분산: $2(n-1)\sigma^4$

③ 평균: $n\sigma^2$, 분산: $2n\sigma^2$

④ 평균: $n\sigma^2$, 분산: $2n\sigma^4$

정답 4번

출제 포인트

이 문제는 정규표본에서 $\sum_{i=1}^{n}(X_i-\mu)^2$의 카이제곱 분포를 이용해 평균과 분산을 구하는 문제이다.

조건 정리

주어진 값: $X_1, \cdots, X_n$은 평균 μ, 분산 σ^2인 정규모집단의 임의표본이다.
구해야 할 값: $Y = \sum_{i=1}^{n}(X_i-\mu)^2$의 평균과 분산이다.
주의할 조건: 표본평균 $\bar{X}$가 아니라 모평균 μ를 기준으로 제곱합을 만든다는 점이 핵심이다.

사용할 공식

$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \sim \chi^2_n$
$E(\chi^2_n) = n, \qquad \mathrm{Var}(\chi^2_n) = 2n$
따라서 $Y = \sigma^2\chi^2_n$이므로 $E(Y) = \sigma^2 n = n\sigma^2$, $\mathrm{Var}(Y) = \sigma^4 \cdot 2n = 2n\sigma^4$이다.
이 공식을 쓰는 이유는 정규표본에서 모평균 μ를 기준으로 표준화한 제곱합이 자유도 n인 카이제곱분포를 따르기 때문이다.

계산 과정

1. 각 표본을 표준화하면 $Z_i = \frac{X_i-\mu}{\sigma}$이고 $Z_i \sim N(0,1)$이다.
2. 따라서 $\sum_{i=1}^{n} Z_i^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \sim \chi^2_n$이다.
3. $Y = \sum_{i=1}^{n}(X_i-\mu)^2 = \sigma^2\chi^2_n$이다.
4. 평균은 $E(Y) = \sigma^2 E(\chi^2_n) = \sigma^2 \cdot n = n\sigma^2$이다.
5. 분산은 $\mathrm{Var}(Y) = \sigma^4 \mathrm{Var}(\chi^2_n) = \sigma^4 \cdot 2n = 2n\sigma^4$이다.

NoteBase

선지 판단

1번: X. 평균 $(n-1)\sigma^2$가 틀렸다. 모평균 μ를 알고 있을 때의 제곱합이므로 자유도 손실이 없어서 평균은 $n\sigma^2$이다. 올바르게는 평균 $n\sigma^2$, 분산 $2n\sigma^4$이다.
2번: X. 분산 $2(n-1)\sigma^4$가 틀렸다. 자유도 $n-1$은 $\bar{X}$를 사용한 $\sum(X_i-\bar{X})^2$에서 나타난다. 이 문항은 μ를 사용하므로 올바르게는 분산 $2n\sigma^4$이다.
3번: X. 평균 $n\sigma^2$는 맞지만 분산 $2n\sigma^2$가 틀렸다. $Y=\sigma^2\chi_n^2$이므로 분산은 σ^2가 아니라 σ^4가 곱해져야 한다. 올바르게는 분산 $2n\sigma^4$이다.
4번: O. $\frac{Y}{\sigma^2}\sim\chi_n^2$이므로 평균은 $n\sigma^2$, 분산은 $2n\sigma^4$가 된다. 제시된 평균과 분산이 모두 맞다.

오답 포인트

출제자가 노린 함정은 $\sum(X_i-\mu)^2$와 $\sum(X_i-\bar{X})^2$를 혼동하게 만드는 점이다. $\bar{X}$가 들어가면 자유도는 $n-1$이지만, μ가 들어가면 자유도는 n이다. 또 분산 계산에서 상수배의 분산이 σ^4로 커진다는 점을 놓치면 3번처럼 $2n\sigma^2$를 고르기 쉽다.

다시 나오면 이렇게 판단

정규표본의 제곱합 문제는 먼저 기준이 μ인지 $\bar{X}$인지 확인한다. 그다음 카이제곱의 자유도를 정하고, 마지막으로 평균은 상수배가 σ^2, 분산은 상수배 제곱이 σ^4로 붙는다는 순서로 계산하면 된다.

암기 포인트

$\sum(X_i-\mu)^2$는 자유도 n, $\sum(X_i-\bar{X})^2$는 자유도 $n-1$이다. 상수 σ^2를 곱하면 평균은 σ^2, 분산은 σ^4가 붙는다.

NoteBase

문 3. 평균이 μ, 분산이 σ^2인 무한모집단으로부터의 임의표본 X_1, X_2, X_3, X_4가 주어졌을 때, μ에 대한 다음 추정량 중 분산이 가장 작은 것은? (단, $\sigma^2 > 0$)

① $\frac{X_1 + X_2 + X_3 + X_4}{4}$
② $\frac{3X_1 + X_2 + X_3 + X_4}{6}$
③ $\frac{0.5X_1 + 1.5X_2 + X_3 + 2X_4}{5}$
④ $\frac{2X_1 + X_2 + 2X_3}{5}$

정답 1번

출제 포인트

이 문제는 독립인 표본들의 선형결합 추정량에서 분산이 $\sigma^2\sum w_i^2$로 결정된다는 점을 이용해 가장 작은 분산을 찾는 문제이다.

조건 정리

주어진 값은 X_1, X_2, X_3, X_4가 평균 μ, 분산 σ^2를 가지는 무한모집단의 임의표본이라는 점이다.
구해야 할 값은 각 추정량의 분산 크기 비교이다.
주의할 조건은 표본들이 서로 독립이므로 분산 계산에서 공분산항이 0이 된다는 점이다.

사용할 공식

$T = \sum w_i X_i, \quad E(T) = \mu\sum w_i, \quad Var(T) = \sigma^2\sum w_i^2$
이 공식을 쓰는 이유는 각 보기의 추정량이 X_i들의 가중평균 꼴이고, 독립 표본에서는 분산이 각 가중치의 제곱합에 비례하기 때문이다.

계산 과정

① $\frac{X_1 + X_2 + X_3 + X_4}{4}$의 가중치는 각각 14,14,14,14이므로
Var=σ^2≤ft(4×≤ft(14)^2)=(σ^2)/(4)
② $\frac{3X_1 + X_2 + X_3 + X_4}{6}$의 가중치는 12,16,16,16이므로
Var=σ^2≤ft(≤ft(12)^2+3×≤ft(16)^2)=σ^2≤ft(14+112)=(σ^2)/(3)
③ $\frac{0.5X_1 + 1.5X_2 + X_3 + 2X_4}{5}$의 가중치는 0.1, 0.3, 0.2, 0.4이므로
$Var = \sigma^2(0.1^2 + 0.3^2 + 0.2^2 + 0.4^2) = 0.30\sigma^2 = \frac{3\sigma^2}{10}$
④ $\frac{2X_1 + X_2 + 2X_3}{5}$의 가중치는 0.4, 0.2, 0.4, 0이므로
$Var = \sigma^2(0.4^2 + 0.2^2 + 0.4^2) = 0.36\sigma^2 = \frac{9\sigma^2}{25}$
비교하면 14σ^2 < (3)/(10)σ^2 < 13σ^2 < (9)/(25)σ^2이므로 ①의 분산이 가장 작다.

NoteBase

선지별 해설

① O
분산은 $\frac{\sigma^2}{4}$이다. 네 표본에 동일한 가중치 14를 준 경우라 가중치 제곱합이 가장 작아 최소 분산이 된다.

② X
분산은 $\frac{\sigma^2}{3}$이다. X_1에 가중치 12를 과하게 주어 제곱합이 커진다.
틀린 판단은 이 추정량의 분산이 최소라는 생각이다. 올바르게는 ①보다 분산이 크므로 최소가 아니다.

③ X
분산은 $\frac{3\sigma^2}{10}$이다. 가중치가 0.1, 0.3, 0.2, 0.4로 고르지 않아 제곱합이 0.30이 된다.
틀린 판단은 이 추정량이 더 효율적이라는 생각이다. 올바르게는 ①의 $\frac{\sigma^2}{4}$보다 크므로 최소가 아니다.

④ X
분산은 $\frac{9\sigma^2}{25}$이다. X_4를 아예 쓰지 않고 나머지 표본에 큰 가중치를 주어 제곱합이 가장 커진다.
틀린 판단은 일부 표본을 빼면 분산이 줄어든다는 생각이다. 올바르게는 표본을 고르게 활용한 ①보다 분산이 크므로 최소가 아니다.

오답 포인트

평균 μ의 추정량이라고 해서 아무 가중평균이나 같은 분산을 가지는 것은 아니다. 합이 1인 가중치라도 한 표본에 큰 가중치를 몰아주면 제곱합 $\sum w_i^2$가 커져 분산이 증가한다는 점이 함정이다.

다시 나오면 이렇게 판단

이런 문제는 먼저 각 추정량을 $\sum w_i X_i$ 꼴로 바꿔 가중치 합이 1인지 확인하고, 그다음 독립 표본이면 곧바로 $\sum w_i^2$만 비교하면 된다. 가중치가 가장 고르게 분산된 추정량의 분산이 가장 작다.

암기 포인트

독립 표본의 가중평균 분산은 $\sigma^2 \sum w_i^2$이고, 가중치가 균등할수록 분산이 작다.

NoteBase

문 4. 표본의 크기가 n인 단순선형회귀모형 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1, \cdots, n$에 대하여 최소제곱법으로 구한 추정식이 $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$이고 잔차가 $e_i = Y_i - \hat{Y}_i$일 때, 옳은 것만을 모두 고르면?

ㄱ. $\sum_{i=1}^{n} e_i = 0$
ㄴ. $\sum_{i=1}^{n} X_i e_i = 0$
ㄷ. $\sum_{i=1}^{n} Y_i e_i = 0$

① ㄱ

② ㄷ

③ ㄱ, ㄴ

④ ㄱ, ㄴ, ㄷ

정답 3번

출제 포인트

이 문제는 단순선형회귀모형에서 최소제곱법으로 얻는 잔차의 직교조건을 묻는 문제이다.

핵심 개념

절편이 포함된 최소제곱회귀에서는 정규방정식 때문에 잔차의 합이 0이 되고, 설명변수와 잔차의 곱의 합도 0이 된다.
또한 $Y_i = \hat{Y}_i + e_i$ 이므로 $\sum Y_i e_i = \sum \hat{Y}_i e_i + \sum e_i^2$ 이다. 여기서 $\sum \hat{Y}_i e_i = 0$ 이지만 $\sum e_i^2$ 는 일반적으로 0이 아니므로 ㄷ은 성립하지 않는다.

NoteBase

선지별 해설

① X
ㄱ만 고른 선택지이다. $\sum e_i = 0$ 는 맞지만, 최소제곱법에서는 $\sum X_i e_i = 0$ 도 함께 성립하므로 부족하다. 틀린 표현은 'ㄱ만 옳다'이고, 올바르게는 'ㄱ, ㄴ이 옳다'이다.

② X
ㄷ만 고른 선택지이다. $\sum Y_i e_i$ 는 일반적으로 0이 아니라 잔차제곱합 $\sum e_i^2$ 와 연결된다. 틀린 표현은 'ㄷ이 옳다'이고, 올바르게는 'ㄷ은 틀리고 ㄱ, ㄴ이 옳다'이다.

③ O
ㄱ, ㄴ을 고른 선택지이다. 절편이 있는 최소제곱회귀에서는 잔차합이 0이고, 설명변수와 잔차도 직교하므로 $\sum X_i e_i = 0$ 이다. 따라서 정답 조합이 맞다.

④ X
ㄱ, ㄴ, ㄷ을 모두 고른 선택지이다. ㄱ과 ㄴ은 맞지만 ㄷ이 틀리므로 전체 선택지는 틀리다. 틀린 표현은 'ㄷ까지 옳다'이고, 올바르게는 'ㄱ, ㄴ만 옳다'이다.

오답 포인트

잔차가 설명변수와 직교한다는 성질을 잔차가 원변수 Y_i 와도 직교한다고 잘못 확장하면 ㄷ을 옳다고 보기 쉽다. Y_i 는 적합값과 잔차의 합이므로 Y_i 와 잔차의 곱의 합에는 $\sum e_i^2$ 가 남는다.

다시 나오면 이렇게 판단

절편이 포함된 최소제곱회귀가 보이면 먼저 $\sum e_i = 0$, $\sum X_i e_i = 0$ 의 두 정규방정식을 떠올린다. 그다음 Y_i 가 나오면 바로 $Y_i = \hat{Y}_i + e_i$ 로 분해해서 잔차제곱합이 남는지 확인하면 된다.

암기 포인트

OLS 잔차는 상수항, 설명변수, 적합값과는 직교하지만 원래 반응변수와는 일반적으로 직교하지 않는다.

NoteBase

문 5. 비료의 양(X)이 곡물 수확량(Y)에 미치는 영향을 알아보기 위해 자료를 수집하여 정리한 결과 X의 표본분산이 16, Y의 표본분산이 25이고 X와 Y의 표본상관계수가 0.8일 때, X를 독립변수(설명변수)로 하고 Y를 종속변수(반응변수)로 하여 최소제곱법으로 구한 추정 회귀직선의 기울기는?

① 0.512

② 0.640

③ 1.000

④ 1.250

정답 3번

출제 포인트

단순회귀식에서 Y를 X에 회귀시킬 때 기울기 b를 상관계수와 표준편차로 구하는 문제이다.

조건 정리

주어진 값: $s_X^2 = 16$, $s_Y^2 = 25$, $r = 0.8$
구해야 할 값: Y를 X에 대한 최소제곱 회귀직선으로 적합했을 때의 기울기 b
주의할 조건: 회귀계수 공식에는 분산이 아니라 표준편차 s_X, s_Y를 넣는다.

사용할 공식

$b = r\frac{s_Y}{s_X}$
이 공식을 쓰는 이유: Y를 X에 대한 단순회귀로 적합하면 최소제곱 회귀직선의 기울기는 상관계수에 표준편차의 비를 곱한 값으로 정해진다.

계산 과정

1. 표본분산에서 표준편차를 구하면 $s_X = \sqrt{16} = 4$, $s_Y = \sqrt{25} = 5$이다.
2. 기울기 공식에 대입하면 $b = 0.8 \times \frac{5}{4}$이다.
3. 계산하면 $b = 0.8 \times 1.25 = 1.0$이므로 기울기는 1.000이다.

NoteBase

선지별 해설

① X
- 0.512는 $0.8 \times \frac{16}{25}$처럼 분산비를 직접 넣었을 때 나오는 값이다.
- 틀린 표현은 분산을 그대로 쓴 점이다. 올바르게는 표준편차비 $\frac{5}{4}$를 써서 1.000이 된다.

② X
- 0.640은 상관계수 0.8를 제곱한 결정계수 r^2 값이다.
- 틀린 표현은 결정계수와 회귀기울기를 혼동한 점이다. 올바르게는 r^2가 아니라 $r\frac{S_Y}{S_X}$를 계산한다.

③ O
- $b = 0.8 \times \frac{5}{4} = 1.000$이므로 맞는 선택지이다.

④ X
- 1.250은 표준편차비 $\frac{5}{4}$만 계산하고 상관계수 0.8을 곱하지 않았을 때의 값이다.
- 틀린 표현은 상관계수를 누락한 점이다. 올바르게는 $0.8 \times \frac{5}{4} = 1.000$이다.

오답 포인트

이 문항은 분산과 표준편차를 혼동하거나, 결정계수 r^2를 회귀기울기로 착각하거나, 상관계수 0.8을 끝까지 곱하지 않는 실수를 노린다. 분산이 주어져도 공식에 바로 넣지 않는다는 점이 핵심 함정이다.

다시 나오면 이렇게 판단

단순회귀에서 Y를 X에 회귀시키면 먼저 $b = r\frac{S_Y}{S_X}$를 떠올린다. 그다음 주어진 값이 분산이면 표준편차로 바꾼 뒤 Y의 표준편차를 분자, X의 표준편차를 분모에 두고 계산하면 된다.

암기 포인트

회귀기울기 $b_{Y|X} = r\frac{S_Y}{S_X}$이다. 분산이 보이면 먼저 제곱근을 취해 표준편차로 바꾼다.

NoteBase

문 6. 어느 지역에서 남녀 40명씩을 무작위로 뽑아 음식 선호도를 조사한 결과가 표와 같다. 이 표를 바탕으로 남녀의 음식 선호도 간에 차이가 없다는 것을 귀무가설로 하여 유의수준 5%에서 검정할 때, 카이제곱 검정통계량의 값과 검정 결과가 바르게 연결된 것은? (단, $\chi^2_\alpha(k)$는 자유도가 k인 카이제곱분포의 $100 \times (1-\alpha)$번째 백분위수를 나타내고, $\chi^2_{0.05}(4) = 9.49$, $\chi^2_{0.05}(9) = 16.92$, $\chi^2_{0.05}(10) = 18.31$이다)

성별 \ 음식	A	B	C	D	E	합계
남	6	4	6	12	12	40
여	6	12	12	4	6	40

① 검정통계량의 값: 12, 검정 결과: 귀무가설을 기각함

② 검정통계량의 값: 12, 검정 결과: 귀무가설을 기각하지 않음

③ 검정통계량의 값: 15, 검정 결과: 귀무가설을 기각함

④ 검정통계량의 값: 15, 검정 결과: 귀무가설을 기각하지 않음

정답 1번

출제 포인트

이 문제는 2×5 분할표에 대한 카이제곱 독립성 검정에서 기대도수, 자유도, 검정통계량을 계산하고 귀무가설의 기각 여부를 판단하는 문제이다.

조건 정리

주어진 값: 남자 40명, 여자 40명이며 관측도수는 남자 (6, 4, 6, 12, 12), 여자 (6, 12, 12, 4, 6)이다.
구해야 할 값: 카이제곱 검정통계량과 유의수준 5%에서의 검정 결과이다.
주의할 조건: 귀무가설은 남녀의 음식 선호도에 차이가 없다는 것이고, 자유도는 (행의 수-1)(열의 수-1)로 계산한다.

사용할 공식

기대도수는 E_ij=((행합)(열합))/(전체합)이다.
검정통계량은 $\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ 이다.
자유도는 $(r-1)(c-1)$이다.
이 공식을 쓰는 이유: 성별과 음식 선호도처럼 두 범주형 변수의 독립 여부를 분할표로 검정하는 상황이기 때문이다.

NoteBase

계산 과정

1. 열합은 A 12, B 16, C 18, D 16, E 18이고 전체합은 80이다.
2. 각 행의 합이 40이므로 기대도수는 남녀 모두 (6, 8, 9, 8, 9)이다.
3. 검정통계량은
$\chi^2 = \left(\frac{(6-6)^2}{6} + \frac{(4-8)^2}{8} + \frac{(6-9)^2}{9} + \frac{(12-8)^2}{8} + \frac{(12-9)^2}{9}\right) \times 2$
$= (0 + 2 + 1 + 2 + 1) \times 2 = 12$
이다.
4. 자유도는 $(2-1)(5-1) = 4$이고, 임계값은 $\chi^2_{0.05}(4) = 9.49$이다. 따라서 $12 > 9.49$이므로 귀무가설을 기각한다.

선지별 해설

① O: 검정통계량이 12이고 자유도 4에서 임계값 9.49보다 크므로 귀무가설을 기각한다. 검정통계량의 값과 결론이 모두 맞다.
② X: 검정통계량 12는 맞지만 결론이 틀리다. $12 > 9.49$이므로 '귀무가설을 기각하지 않음'이 아니라 '귀무가설을 기각함'이 맞다.
③ X: 결론은 기각이 맞지만 검정통계량이 틀리다. '검정통계량의 값 15'가 틀린 표현이고, 올바른 값은 12이다.
④ X: 검정통계량과 결론이 모두 틀리다. '검정통계량의 값 15'가 아니라 12이고, '귀무가설을 기각하지 않음'이 아니라 '귀무가설을 기각함'이 맞다.

오답 포인트

행합이 모두 40이므로 기대도수가 각 열합의 절반이라는 점을 바로 잡아내지 못하면 검정통계량을 잘못 계산하기 쉽다. 또한 검정통계량만 구하고 임계값 9.49와의 비교를 생략하면 ②처럼 결론을 반대로 고르기 쉽다.

다시 나오면 이렇게 판단

독립성 검정 문제는 열합으로 기대도수를 먼저 만들고, 각 셀에서 $\frac{(O-E)^2}{E}$를 계산해 모두 더한 뒤, 자유도 $(r-1)(c-1)$에 맞는 임계값과 비교하여 기각 여부를 판단한다.

암기 포인트

$2 \times c$ 독립성 검정의 자유도는 $c - 1$이고, 검정통계량이 임계값보다 크면 귀무가설을 기각한다.

NoteBase

문 7. 다음은 반복수가 다른 일원배치법 모형이다. 위 모형에 대하여 기술한 것으로 옳지 않은 것은? (단, $n_i \geq 2,\ i=1,\cdots,k,\ k \geq 2$이고, $n=\sum_{i=1}^{k} n_i$, $\bar{Y}=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$, $\bar{Y}_i=\frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$이며, $F_\alpha(\nu_1,\nu_2)$는 분자의 자유도가 ν_1, 분모의 자유도가 ν_2인 F 분포의 $100\times(1-\alpha)$번째 백분위수를 나타낸다)

○ $Y_{ij}=\mu+\alpha_i+\epsilon_{ij},\ \ i=1,\cdots,k,\ j=1,\cdots,n_i$

○ $\epsilon_{ij} \sim N(0,\sigma^2)$이고 서로 독립, $i=1,\cdots,k,\ j=1,\cdots,n_i$

○ $\sum_{i=1}^{k} n_i\alpha_i=0$

① 총제곱합은 $\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2-n\bar{Y}^2$이고 자유도는 $n-1$이다.

② 처리제곱합은 $\sum_{i=1}^{k} n_i\bar{Y}_i^2-n\bar{Y}^2$이고 자유도는 $k-1$이다.

③ 오차제곱합(잔차제곱합)은 $\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2-\sum_{i=1}^{k} n_i\bar{Y}_i^2$이고 자유도는 $n-k$이다.

④ 통계량 $\dfrac{\sum_{i=1}^{k} n_i\bar{Y}_i^2-n\bar{Y}^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2-\sum_{i=1}^{k} n_i\bar{Y}_i^2}$은 분자의 자유도가 $k-1$, 분모의 자유도가 $n-k$인 F 분포를 따른다.

정답 4번

출제 포인트

반복수가 다른 일원배치법에서 총제곱합, 처리제곱합, 오차제곱합의 분해와 검정통계량이 제곱합의 비가 아니라 평균제곱의 비라는 점을 묻는 문제이다.

조건 정리

모형은 $Y_{ij}=\mu+\alpha_i+\varepsilon_{ij}, \varepsilon_{ij}\sim N(0,\sigma^2)$ 서로독립, $\sum_{i=1}^{k} n_i\alpha_i=0$이다.
전체 표본수는 $n=\sum_{i=1}^{k} n_i$, 집단평균은 $\bar{Y}_i=\frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$, 전체평균은 $\bar{Y}=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$이다.
일원배치 분산분석에서는 $SST=SSA+SSE$, 자유도는 $(n-1)=(k-1)+(n-k)$로 분해된다.

NoteBase

사용할 공식

$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - n\bar{Y}^2$

$SSA = \sum_{i=1}^{k} n_i \bar{Y}_i^2 - n\bar{Y}^2$

$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^{k} n_i \bar{Y}_i^2$

$F = \frac{MSA}{MSE} = \frac{SSA/(k-1)}{SSE/(n-k)}$

이 문제는 각 선지가 제곱합과 자유도, 그리고 최종 검정통계량을 정확히 연결했는지를 확인하면 판단할 수 있다.

선지별 해설

① O. 총제곱합은 전체 자료의 변동을 뜻하므로 $SST = \sum\sum Y_{ij}^2 - n\bar{Y}^2$로 쓴다. 전체평균 1개를 추정하므로 자유도는 $n-1$이다.

② O. 처리제곱합은 집단평균들 사이의 변동이므로 $SSA = \sum n_i \bar{Y}_i^2 - n\bar{Y}^2$가 맞다. 집단이 k개이면 제약 $\sum n_i \alpha_i = 0$ 아래 독립적인 처리효과 자유도는 $k-1$이다.

③ O. 오차제곱합은 각 집단 내부 변동이므로 $SSE = SST - SSA = \sum\sum Y_{ij}^2 - \sum n_i \bar{Y}_i^2$이다. 각 집단에서 평균을 하나씩 추정하므로 자유도는 $n-k$이다.

④ X. 제시된 식은 $\frac{SSA}{SSE}$로서 제곱합의 비일 뿐이다. F분포를 따르는 것은 제곱합을 자유도로 나눈 평균제곱의 비인 $\frac{SSA/(k-1)}{SSE/(n-k)}$이다. 틀린 표현은 '이 통계량이 곧 F분포를 따른다'는 부분이고, 올바르게는 '자유도로 나눈 평균제곱의 비가 분자 자유도 $k-1$, 분모 자유도 $n-k$인 F분포를 따른다'이다.

오답 포인트

4번은 처리제곱합과 오차제곱합의 식 자체는 맞게 제시한 뒤, 마지막 단계에서 평균제곱으로 나누는 절차를 숨겨 놓은 함정이다. 제곱합 SS와 평균제곱 MS를 같은 것으로 보면 F통계량을 잘못 판단하게 된다.

다시 나오면 이렇게 판단

분산분석 선지는 먼저 SST, SSA, SSE 식과 자유도를 각각 확인한 뒤, 검정통계량이 나오면 반드시 '자유도로 나누었는가'를 본다. F검정은 항상 제곱합의 비가 아니라 평균제곱의 비로 판단한다.

암기 포인트

일원배치 분산분석은 $SST = SSA + SSE$, 검정통계량은 $F = MSA/MSE$이다. SS만 있으면 아직 F가 아니다.

NoteBase

문 8. 자유도가 ν인 t 분포의 $100 \times (1-\alpha)$번째 백분위수는 $t_\alpha(\nu)$, 분자의 자유도가 1이고 분모의 자유도가 ν인 F 분포의 $100 \times (1-\alpha)$번째 백분위수는 $F_\alpha(1, \nu)$로 나타낸다. $F_\alpha(1, \nu) = 3$일 때 옳은 것은?

① $t_\alpha(\nu) = 9$

② $t_{\alpha/2}(\nu) = 9$

③ $t_\alpha(\nu) = \sqrt{3}$

④ $t_{\alpha/2}(\nu) = \sqrt{3}$

정답 4번

출제 포인트

이 문제는 t분포와 F분포의 관계 $T_\nu^2 \sim F(1, \nu)$를 이용하여 백분위수를 연결하는 문제이다.

조건 정리

주어진 값: $F_\alpha(1, \nu) = 3$
구해야 할 값: $t_\alpha(\nu)$ 또는 $t_{\alpha/2}(\nu)$와의 관계
주의할 조건: T^2로 바꾸면 양쪽 꼬리가 합쳐지므로 α가 $\alpha/2$로 나뉜다.

사용할 공식

$T_\nu^2 \sim F(1, \nu)$ 이다. 따라서 $F_\alpha(1, \nu) = t_{\alpha/2}(\nu)^2$ 이다.
이 공식을 쓰는 이유: 분자의 자유도가 1인 F분포는 같은 자유도의 t분포를 제곱한 분포이므로 백분위수도 제곱 관계로 연결된다.

계산 과정

1. $F_\alpha(1, \nu) = 3$ 이면 $t_{\alpha/2}(\nu)^2 = 3$ 이다.
2. t분포의 상측 백분위수는 양수이므로 $t_{\alpha/2}(\nu) = \sqrt{3}$ 이다.
3. 따라서 보기 중 맞는 식은 ④이다.

선지별 해설

① X: $t_\alpha(\nu) = 9$는 제곱 관계와 꼬리확률 관계를 모두 잘못 쓴 식이다. 올바르게는 9가 아니라 3이 제곱값이고, $t_\alpha(\nu)$가 아니라 $t_{\alpha/2}(\nu)$가 연결된다.
② X: $t_{\alpha/2}(\nu) = 9$는 꼬리확률 첨자는 맞지만 제곱을 풀 때 값을 잘못 쓴 식이다. 올바르게는 $t_{\alpha/2}(\nu)^2 = 3$, 따라서 $t_{\alpha/2}(\nu) = \sqrt{3}$ 이다.
③ X: $t_\alpha(\nu) = \sqrt{3}$는 제곱을 푼 값 자체는 맞지만 첨자가 틀렸다. 올바르게는 양쪽 꼬리 분할 때문에 $t_{\alpha/2}(\nu) = \sqrt{3}$ 이다.
④ O: $T_\nu^2 \sim F(1, \nu)$와 대칭성을 쓰면 $F_\alpha(1, \nu) = t_{\alpha/2}(\nu)^2$ 이므로 $F_\alpha(1, \nu) = 3$에서 $t_{\alpha/2}(\nu) = \sqrt{3}$ 가 된다.

NoteBase

오답 포인트

이 문항의 함정은 두 가지이다. 첫째, F분포 백분위수와 t분포 백분위수를 연결할 때 제곱 관계를 빼먹기 쉽다. 둘째, $T^2 \le c$는 $-\sqrt{c} \le T \le \sqrt{c}$를 뜻하므로 한쪽 꼬리 α가 아니라 양쪽으로 나뉜 $\alpha/2$를 써야 한다.

다시 나오면 이렇게 판단

분자의 자유도가 1인 F분포가 나오면 먼저 $T_\nu^2 \sim F(1, \nu)$를 떠올린다. 그다음 제곱을 벗길 때는 값에 제곱근을 취하고, 꼬리확률은 대칭성 때문에 $\alpha \to \alpha/2$로 바뀌는지를 확인한다.

암기 포인트

분자 자유도 1인 F는 t^2이고, 백분위수 관계는 $F_\alpha(1, \nu) = t_{\alpha/2}(\nu)^2$ 이다.

NoteBase

문 9. $X_1, X_2, \cdots, X_{16}$**이 평균** μ**, 분산** σ^2**인 정규모집단에서의 임의표본일 때, 가설** $H_0 : \sigma^2 = 30$ **대** $H_1 : \sigma^2 \geq 30$**을 유의수준 5%에서 검정하기 위한 기각역(rejection region)은? (단,** μ**는 미지의 모수이고** $\overline{X} = \frac{1}{16}\sum_{i=1}^{16} X_i$**이며,** $\chi^2_\alpha(k)$**는 자유도가** k**인 카이제곱분포의** $100 \times (1-\alpha)$**번째 백분위수를 나타낸다.)**

① $\sum_{i=1}^{16}(X_i - \overline{X})^2 \geq 2\chi^2_{0.05}(15)$

② $\sum_{i=1}^{16}(X_i - \overline{X})^2 \geq 30\chi^2_{0.05}(15)$

③ $\sum_{i=1}^{16}(X_i - \overline{X})^2 \geq 2\chi^2_{0.05}(16)$

④ $\sum_{i=1}^{16}(X_i - \overline{X})^2 \geq 30\chi^2_{0.05}(16)$

정답 2번

출제 포인트

이 문제는 모평균을 모르는 정규모집단에서 모분산 검정을 할 때 검정통계량의 자유도와 기각역을 묻는 문제이다.

조건 정리

주어진 값: 표본크기는 16, 귀무가설은 $H_0 : \sigma^2 = 30$, 대립가설은 $H_1 : \sigma^2 > 30$이다.
구해야 할 값: 유의수준 5%에서의 기각역이다.
주의할 조건: μ를 모르므로 $\overline{X}$를 사용한 제곱합의 자유도는 $n-1 = 15$이다.

사용할 공식

정규모집단에서 모평균을 모를 때
$\frac{\sum_{i=1}^{16}(X_i - \overline{X})^2}{\sigma_0^2} \sim \chi^2(15)$
이다.
이 공식을 쓰는 이유: 귀무가설 아래에서 검정하려는 값은 모분산 σ^2이고, 모평균이 미지이므로 표본평균으로 중심화한 제곱합이 자유도 $n-1$인 카이제곱분포를 따른다.

계산 과정

1. 귀무가설에서 $\sigma_0^2 = 30$이다.
2. 검정통계량은 $T = \frac{\sum_{i=1}^{16}(X_i - \overline{X})^2}{30}$이고, $T \sim \chi^2(15)$이다.
3. 대립가설이 $\sigma^2 > 30$이므로 큰 값에서 기각하는 우측검정이다.
4. 문제에서 $\chi^2_{0.05}(15)$를 자유도 15인 카이제곱분포의 95번째 백분위수로 정의했으므로 기각역은 $T \geq \chi^2_{0.05}(15)$이다.
5. 이를 원래 제곱합으로 옮기면
$\sum_{i=1}^{16}(X_i - \overline{X})^2 \geq 30\chi^2_{0.05}(15)$
가 된다.

NoteBase

선지별 해설

① X

- 틀린 이유: 임계값 앞의 상수가 2로 되어 있는데, 검정통계량의 분모에는 귀무가설의 모분산 $\sigma_0^2 = 30$이 들어가야 한다.
- 틀린 표현 → 올바르게는 $\sum_{i=1}^{16}(X_i - \overline{X})^2 \geq 30\chi^2_{0.05}(15)$이다.

② O

- 맞는 이유: 모평균이 미지이므로 자유도는 $16 - 1 = 15$이고, $H_1 : \sigma^2 > 30$이므로 우측검정에서 $\sum_{i=1}^{16}(X_i - \overline{X})^2 \geq 30\chi^2_{0.05}(15)$가 기각역이 된다.

③ X

- 틀린 이유: 상수 2도 잘못되었고 자유도도 16으로 잘못 잡았다. 모평균을 모르면 자유도는 $n - 1$이다.
- 틀린 표현 → 올바르게는 상수는 30, 자유도는 15가 되어 $\sum_{i=1}^{16}(X_i - \overline{X})^2 \geq 30\chi^2_{0.05}(15)$이다.

④ X

- 틀린 이유: 상수 30은 맞지만 자유도를 16으로 둔 점이 틀렸다. $\overline{X}$를 추정에 사용했으므로 자유도 1이 줄어든다.
- 틀린 표현 → 올바르게는 $\chi^2_{0.05}(16)$이 아니라 $\chi^2_{0.05}(15)$를 써야 한다.

오답 포인트

이 문항의 함정은 두 가지이다. 하나는 분산 검정에서 제곱합 앞의 계수를 귀무가설의 모분산 30이 아니라 엉뚱한 수로 바꾸는 실수이고, 다른 하나는 모평균이 미지일 때도 자유도를 n으로 쓰는 실수이다. 특히 $\overline{X}$가 들어가면 자유도는 반드시 1 감소한다.

다시 나오면 이렇게 판단

정규모집단의 분산 검정은 먼저 모평균의 기지 여부를 확인한 뒤 자유도가 n인지 $n - 1$인지 정한다. 그다음 대립가설이 $>$이면 우측검정, $<$이면 좌측검정으로 방향을 정하고, 마지막에 검정통계량의 분모에 귀무가설의 분산 σ_0^2를 넣어 기각역을 원래 식으로 되돌린다.

암기 포인트

모평균 미지 분산검정은 $\frac{\sum(X_i - \overline{X})^2}{\sigma_0^2} \sim \chi^2(n-1)$이다.

NoteBase

문 10. $X_1, X_2, \cdots, X_n$**이 서로 독립이고 구간** $(0, 1)$**에서 균일분포(uniform distribution)를 따른다 하자. 이 중 가장 작은 확률변수를** Y**라 할 때,** $0 < y < 1$**에서** Y**의 누적분포함수** $F(y)$**는?**

① y^n

② $(1-y)^n$

③ $(1-e^{-y})^n$

④ $1-(1-y)^n$

정답 4번

출제 포인트

이 문제는 독립인 균일분포 표본의 최솟값 $Y = \min(X_1, \ldots, X_n)$의 누적분포함수를 여사건으로 구하는 문제이다.

조건 정리

주어진 값: $X_1, X_2, \ldots, X_n$은 서로 독립이고 각각 구간 $(0, 1)$에서 균일분포를 따른다.
구해야 할 값: $0 < y < 1$에서 최솟값 Y의 누적분포함수 $F_Y(y) = P(Y \le y)$이다.
주의할 조건: 최솟값은 직접 $P(Y \le y)$를 세기보다 여사건 $P(Y > y)$를 쓰는 편이 가장 간단하다.

사용할 공식

$F_Y(y) = P(Y \le y) = 1 - P(Y > y)$이다.
또한 $Y > y$는 모든 표본이 y보다 큰 경우와 같으므로 $P(Y > y) = P(X_1 > y, \ldots, X_n > y)$이다.
독립이므로 $P(Y > y) = \prod_{i=1}^{n} P(X_i > y)$이고, 균일분포 $(0, 1)$에서는 $P(X_i > y) = 1 - y$이다.
이 공식을 쓰는 이유는 최솟값 사건이 '모두 y보다 크다'라는 곱사건으로 바로 바뀌기 때문이다.

계산 과정

1. $Y = \min(X_1, \ldots, X_n)$이므로 Y>y X_1>y,$\cdots$,X_n>y이다.
2. 따라서 $P(Y > y) = P(X_1 > y, \ldots, X_n > y) = \prod_{i=1}^{n} P(X_i > y) = (1-y)^n$이다.
3. 그러므로 $F_Y(y) = P(Y \le y) = 1 - P(Y > y) = 1 - (1-y)^n$이다.

NoteBase

선지별 해설

① X
y^n은 $P(X_1 \le y, \dots, X_n \le y)$로서 최댓값 $\max(X_1, \dots, X_n)$의 누적분포함수이다. 틀린 표현은 최솟값의 분포를 최댓값의 분포와 혼동한 점이며, 올바르게는 $F_Y(y) = 1 - (1-y)^n$이다.

② X
$(1-y)^n$은 $P(Y > y)$에 해당하는 생존함수이다. 틀린 표현은 여사건을 처리하지 않고 그대로 누적분포함수라고 본 점이며, 올바르게는 $F_Y(y) = 1 - (1-y)^n$이다.

③ X
$(1-e^{-y})^n$은 이 문제의 균일분포와 맞지 않는 식이다. 이는 지수분포 계열이나 다른 분포의 최댓값 형태에서 나올 수 있는 식이지, $(0,1)$ 균일분포 표본의 최솟값 누적분포함수가 아니다. 올바르게는 각 표본에 대해 $P(X_i > y) = 1 - y$를 써서 $1 - (1-y)^n$을 얻는다.

④ O
최솟값 Y에 대해 $F_Y(y) = 1 - P(Y > y) = 1 - P(X_1 > y, \dots, X_n > y) = 1 - (1-y)^n$이므로 옳다.

오답 포인트

최솟값 문제에서 $P(Y \le y)$와 $P(Y > y)$를 뒤바꾸기 쉽다. 특히 ①처럼 최댓값의 공식 y^n을 끌어오거나, ②처럼 생존함수 $(1-y)^n$를 누적분포함수로 착각하는 함정이 자주 나온다.

다시 나오면 이렇게 판단

순서통계량에서 최솟값이 보이면 먼저 $P(Y > y)$를 쓴다. 그다음 '모든 표본이 y보다 크다'로 바꾸고, 독립이면 각 확률을 곱한 뒤 마지막에 여사건으로 $1 -$를 붙이면 된다.

암기 포인트

최솟값의 누적분포함수는 '전부 초과'의 여사건이므로 $F_{\min}(y) = 1 - [P(X > y)]^n$ 형태로 정리된다.

문 11. 두 사건 A와 B에 대하여 $P(A) = 0.3$이고 $P(A \cup B) = 0.7$이다. A와 B가 독립이 되기 위한 $P(B)$의 값은?

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

정답 1번

출제 포인트

이 문제는 독립사건의 합집합 확률식 $P(A \cup B) = P(A) + P(B) - P(A \cap B)$와 $P(A \cap B) = P(A)P(B)$를 이용해 $P(B)$를 구하는 문제이다.

조건 정리

주어진 값: $P(A) = 0.3$, $P(A \cup B) = 0.7$
구해야 할 값: $P(B)$
주의할 조건: A와 B가 독립이면 $P(A \cap B) = P(A)P(B)$가 된다.

사용할 공식

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$, $P(A \cap B) = P(A)P(B)$이다.
이 공식을 쓰는 이유: 발문이 독립 조건에서 합집합 확률을 주고 한 사건의 확률을 묻고 있으므로, 교집합을 곱으로 바꾸어 미지수 $P(B)$를 세울 수 있다.

계산 과정

1. $P(B) = x$라고 두면 독립이므로 $P(A \cap B) = 0.3x$이다.
2. 합집합 확률식에 대입하면 $0.7 = 0.3 + x - 0.3x$이다.
3. 정리하면 $0.4 = 0.7x$이므로 $x = \frac{4}{7}$이다. 따라서 $P(B) = \frac{4}{7}$이다.

선지별 해설

① O: $P(B) = \frac{4}{7}$를 대입하면 $0.3 + \frac{4}{7} - 0.3 \cdot \frac{4}{7} = 0.7$이 되어 조건을 만족한다.
② X: 숫자 $\frac{1}{2}$가 틀렸다. 이를 대입하면 $0.3 + 0.5 - 0.15 = 0.65$로 0.7이 되지 않는다. 올바르게는 $P(B) = \frac{4}{7}$이다.
③ X: 숫자 $\frac{4}{9}$가 틀렸다. 이를 대입하면 $0.3 + \frac{4}{9} - 0.3 \cdot \frac{4}{9} = \frac{11}{18}$로 0.7이 되지 않는다. 올바르게는 $P(B) = \frac{4}{7}$이다.
④ X: 숫자 $\frac{2}{5}$가 틀렸다. 이는 $0.7 - 0.3 = 0.4$만 보고 교집합을 빼는 항을 무시할 때 나오는 값인데, 독립이면 $P(A \cap B) = P(A)P(B)$를 반드시 반영해야 한다. 올바르게는 $P(B) = \frac{4}{7}$이다.

NoteBase

오답 포인트

독립이라는 말을 보고도 교집합을 곱으로 바꾸지 않거나, 합집합 확률식에서 $-P(A \cap B)$ 항을 빼먹으면 오답이 된다. 특히 ④는 $P(A \cup B) - P(A)$만 계산한 전형적인 함정이다.

다시 나오면 이렇게 판단

합집합 확률과 독립 조건이 함께 나오면 먼저 $P(A \cap B) = P(A)P(B)$로 바꾼 뒤
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$에 대입해 미지수를 푼다.

암기 포인트

독립이면 교집합은 곱이고, 합집합은 두 확률의 합에서 교집합을 한 번 뺀 값이다.

NoteBase

문 12. 두 확률변수 X와 Y의 결합확률밀도함수가 다음과 같을 때, 상수 c의 값은? (단, π는 원주율을 의미한다)

f(x,y)=cases c, & 1 < x^2 + y^2 < 9, x < y \ 0, & 그 밖의 영역 cases

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

정답 2번

출제 포인트

이 문제는 결합확률밀도함수의 전체 적분값이 1이라는 조건을 이용해 상수 c를 구하는 문제이다.

조건 정리

- 지지영역은 $1 < x^2 + y^2 < 9$ 이므로 반지름이 1과 3인 두 원 사이의 원환 영역이다.
- 추가 조건은 $x < y$ 이므로 직선 $y = x$를 기준으로 한쪽 반영역만 남는다.
- 구해야 할 값은 $\iint f(x, y)dxdy = 1$을 만족하는 상수 c이다.

사용할 공식

$\iint_D cdxdy = c \cdot \text{Area}(D) = 1$
이 공식을 쓰는 이유는 밀도함수가 영역 D에서 상수이므로 전체 확률은 상수 c에 지지영역의 넓이를 곱한 값이 되기 때문이다.

계산 과정

1. $1 < x^2 + y^2 < 9$의 전체 넓이는 반지름 3인 원의 넓이에서 반지름 1인 원의 넓이를 뺀 값이므로 $9\pi - \pi = 8\pi$이다.
2. 조건 $x < y$는 원환 영역을 직선 $y = x$로 정확히 절반으로 나누므로 실제 지지영역의 넓이는 $\frac{8\pi}{2} = 4\pi$이다.
3. 따라서 $c \cdot 4\pi = 1$이고, $c = \frac{1}{4\pi}$이다.

NoteBase

선지별 해설

① X: $\frac{1}{8\pi}$는 원환 전체 넓이 8π만 쓰고 조건 $x < y$를 빼먹었을 때 나오는 값이다. 틀린 표현은 지지영역을 전체 원환으로 본 점이고, 올바르게는 $x < y$ 때문에 넓이가 절반인 4π가 되어 $c = \frac{1}{4\pi}$이다.
② O: 지지영역의 넓이가 4π이므로 $c \cdot 4\pi = 1$, 따라서 $c = \frac{1}{4\pi}$가 된다.
③ X: 4π는 지지영역의 넓이일 뿐 밀도상수 c가 아니다. 틀린 표현은 넓이와 밀도상수를 뒤바꾼 점이고, 올바르게는 그 넓이의 역수인 $\frac{1}{4\pi}$가 상수 c이다.
④ X: 8π는 조건 $x < y$를 무시한 전체 원환의 넓이일 뿐이며 밀도상수가 아니다. 틀린 표현은 면적 자체를 c로 둔 점이고, 올바르게는 실제 지지영역 넓이 4π의 역수인 $\frac{1}{4\pi}$이다.

오답 포인트

가장 흔한 함정은 $1 < x^2 + y^2 < 9$만 보고 원환 전체 넓이 8π를 쓴 뒤, $x < y$가 영역을 절반으로 줄인다는 점을 놓치는 것이다. 또 넓이 4π를 구한 뒤 그 값을 그대로 c로 착각하는 역수 혼동도 자주 나온다.

다시 나오면 이렇게 판단

상수형 결합밀도는 먼저 지지영역의 모양과 넓이를 구한 뒤, 상수×넓이=1로 정규화한다. 원이나 원환에 직선 조건이 붙으면 대칭으로 절반인지부터 확인하면 된다.

암기 포인트

상수밀도는 c=1/지지영역 넓이이고, $x < y$는 $y = x$ 대각선을 기준으로 반영역을 뜻한다.

NoteBase

문 13. 관측값 $x_1, x_2, \cdots, x_{10}$을 다음과 같이 변환하였다. x_i 자료의 변동계수를 CV_x, y_i 자료의 변동계수를 CV_y, z_i 자료의 변동계수를 CV_z라 할 때, CV_x, CV_y, CV_z 간의 대소관계가 옳은 것은? (단, $x_i > 0,\ i = 1, 2, \cdots, 10$)

$y_i = 2x_i,\ z_i = x_i + 10,\ i = 1, 2, \cdots, 10.$

① $CV_x < CV_y,\ CV_x > CV_z$

② $CV_x < CV_y,\ CV_x < CV_z$

③ $CV_x = CV_y,\ CV_x > CV_z$

④ $CV_x = CV_y,\ CV_x < CV_z$

정답 3번

출제 포인트

변동계수 $CV = \frac{s}{\bar{x}}$가 배수변환에는 변하지 않고, 양의 상수 가산에는 감소한다는 점을 묻는 문제이다.

핵심 개념

변동계수는 표준편차를 평균으로 나눈 상대적 산포도이다. $y_i = 2x_i$이면 평균과 표준편차가 모두 2배가 되어 $CV_y = \frac{2s_x}{2\bar{x}} = CV_x$가 된다. $z_i = x_i + 10$이면 표준편차는 그대로이고 평균만 $\bar{x} + 10$으로 커지므로 $CV_z = \frac{s_x}{\bar{x}+10} < \frac{s_x}{\bar{x}} = CV_x$이다. 문제에서 $x_i > 0$이므로 $\bar{x} > 0$라는 점이 이 부등호의 근거이다.

NoteBase

선지별 해설

① X
- $y_i = 2x_i$이면 평균과 표준편차가 같은 비율로 커지므로 $CV_x < CV_y$는 성립하지 않는다.
- 틀린 표현은 'CV_x < CV_y'이다. 올바르게는 'CV_x = CV_y'이다.
- $z_i = x_i + 10$이면 평균만 커져 변동계수는 감소하므로 $CV_x > CV_z$는 맞는 판단이다.

② X
- 첫째 비교인 $CV_x < CV_y$가 틀리다. $y_i = 2x_i$에서는 변동계수가 변하지 않는다.
- 틀린 표현은 'CV_x < CV_y'이다. 올바르게는 'CV_x = CV_y'이다.
- 둘째 비교인 $CV_x < CV_z$도 틀리다. $z_i = x_i + 10$에서는 평균이 커져 CV_z가 더 작아진다.
- 틀린 표현은 'CV_x < CV_z'이다. 올바르게는 'CV_x > CV_z'이다.

③ O
- $y_i = 2x_i$이면 $CV_y = CV_x$이고, $z_i = x_i + 10$이면 $CV_z < CV_x$이므로 $CV_x = CV_y$, $CV_x > CV_z$가 모두 맞다.

④ X
- $CV_x = CV_y$는 맞다. 배수변환은 변동계수를 바꾸지 않는다.
- 그러나 $CV_x < CV_z$는 틀리다. 상수 10을 더하면 평균만 증가하므로 CV_z는 오히려 작아진다.
- 틀린 표현은 'CV_x < CV_z'이다. 올바르게는 'CV_x > CV_z'이다.

오답 포인트

분산이나 표준편차의 성질만 떠올리고 평균이 분모에 들어간다는 점을 놓치면 헷갈리기 쉽다. $x_i + 10$에서 표준편차가 그대로라는 사실만 기억하고 변동계수도 그대로라고 착각하는 것이 대표적 함정이다.

다시 나오면 이렇게 판단

변동계수 문제는 평균 변화와 표준편차 변화를 따로 적은 뒤 마지막에 CV=표준편차평균로 합쳐 판단한다. cX는 평균과 표준편차가 함께 c배가 되어 CV가 같고, $X + a$는 표준편차는 그대로이고 평균만 바뀌어 CV가 달라진다.

암기 포인트

cX는 변동계수 불변, $X + a$는 평균이 양수이면 변동계수 감소이다.

NoteBase

문 14. 지수분포의 확률밀도함수는 다음과 같다. 어느 동사무소에 민원인이 도착하여 민원서류를 발급받기까지 기다리는 시간이 평균 2분인 지수분포를 따른다고 할 때, 한 민원인이 3분 이상 기다릴 확률은?

f(x)=casesλ e^-λ x, & x>0 \ 0, & x≤ 0cases, λ>0

① $e^{-2/3}$

② $e^{-3/2}$

③ e^{-3}

④ e^{-6}

정답 2번

출제 포인트

이 문제는 지수분포의 평균과 모수의 관계 $E(X) = 1/\lambda$를 이용해 $P(X \geq 3)$를 구하는 문제이다.

조건 정리

- 주어진 값: 평균 대기시간은 2분이다.
- 구해야 할 값: 한 민원인이 3분 이상 기다릴 확률이다.
- 주의할 조건: 지수분포에서 평균은 $1/\lambda$이므로 λ 자체가 2가 아니라 1/2이다.

사용할 공식

지수분포 $X \sim Exp(\lambda)$에서 평균은 $E(X) = 1/\lambda$, 꼬리확률은 $P(X \geq t) = e^{-\lambda t}$이다.

이 공식을 쓰는 이유:
문제는 평균이 주어진 지수분포에서 일정 시간 이상 기다릴 확률을 묻고 있으므로 평균으로 λ를 구한 뒤 꼬리확률 공식을 적용하면 된다.

계산 과정

1. 평균이 2분이므로 $1/\lambda = 2$에서 λ는 $\lambda = 1/2$이다.
2. 구하려는 확률은 $P(X \geq 3) = e^{-\lambda \cdot 3}$이다.
3. λ를 대입하면 $P(X \geq 3) = e^{-(1/2) \cdot 3} = e^{-3/2}$이다.

NoteBase

선지별 해설

① X
- $e^{-2/3}$는 평균 2분과 대기시간 3분의 비만 거꾸로 섞은 값일 뿐, 지수분포의 꼬리확률 $e^{-\lambda t}$에 맞지 않는다.
- 틀린 표현은 확률값을 $e^{-2/3}$로 둔 부분이다. 올바르게는 확률이 $e^{-3/2}$이다.

② O
- 평균 2분이면 λ는 1/2이고, 따라서 $P(X \geq 3) = e^{-(1/2)\cdot 3} = e^{-3/2}$가 된다.
- 지수분포에서 3분 이상 기다릴 확률을 정확히 계산한 값이므로 옳다.

③ X
- e^{-3}은 λ를 1로 잘못 둔 값이다. 평균이 2분이라는 조건을 반영하면 λ는 1이 아니라 1/2이다.
- 틀린 표현은 지수를 -3으로 둔 부분이다. 올바르게는 지수가 $-3/2$이므로 확률은 $e^{-3/2}$이다.

④ X
- e^{-6}은 평균 2를 그대로 λ로 오해해 $e^{-2\cdot 3}$처럼 계산한 값이다. 평균과 모수는 역수 관계이므로 이렇게 계산하면 안 된다.
- 틀린 표현은 λ를 2로 둔 부분이다. 올바르게는 $\lambda = 1/2$이므로 확률은 $e^{-3/2}$이다.

오답 포인트

평균이 2분이라는 문장을 보고 λ=2로 바로 두는 실수가 가장 흔하다. 지수분포는 평균이 $1/\lambda$이므로 먼저 역수를 취해야 하며, 그다음에 $e^{-\lambda t}$의 지수에 시간을 곱해야 한다.

다시 나오면 이렇게 판단

지수분포 계산형 문제는 먼저 평균과 λ의 역수 관계를 확인하고, 다음으로 $P(X \geq t) = e^{-\lambda t}$에 시간을 대입하는 순서로 판단한다.

암기 포인트

지수분포는 평균이 $1/\lambda$, 3분 이상 같은 꼬리확률은 $e^{-\lambda t}$ 형태로 계산된다.

NoteBase

문 15. 표본의 크기가 12인 어느 자료에 단순선형회귀모형 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ i = 1, 2, \cdots, 12$를 적용하여 얻은 분산분석표의 일부가 다음과 같을 때, 결정계수 r^2의 값은?

요인	제곱합	자유도	평균제곱	F
회귀				15.0
오차(잔차)			4.0	
합계				

① 0.56

② 0.58

③ 0.60

④ 0.62

정답 3번

출제 포인트

이 문제는 단순선형회귀의 분산분석표에서 F값과 오차평균제곱을 이용해 결정계수 r^2를 구하는 문제이다.

조건 정리

주어진 값은 표본크기 12, 회귀의 $F = 15.0$, 오차(잔차)의 평균제곱 $MSE = 4.0$이다.
단순선형회귀이므로 회귀 자유도는 1, 오차 자유도는 $12 - 2 = 10$이다.
구해야 할 값은 결정계수 r^2이다.

사용할 공식

사용할 관계식은 $F = MSR/MSE$, $SSR = MSR \times df_R$, $SSE = MSE \times df_E$, $r^2 = SSR/(SSR + SSE)$이다.
이 공식을 쓰는 이유는 분산분석표가 회귀제곱합과 오차제곱합을 직접 주지 않고 F값, 오차평균제곱, 표본크기만 주고 있기 때문이다.

계산 과정

1. $F = MSR/MSE = 15$이고 $MSE = 4$이므로 $MSR = 15 \times 4 = 60$이다.
2. 단순선형회귀에서 회귀 자유도는 1이므로 $SSR = 60 \times 1 = 60$이다.
3. 오차 자유도는 $n - 2 = 10$이므로 $SSE = 4 \times 10 = 40$이다.
4. 따라서 $r^2 = SSR/(SSR + SSE) = 60/(60 + 40) = 0.60$이다.

NoteBase

선지별 해설

① X: 0.56은 계산 결과와 다르다. 틀린 표현은 '결정계수가 0.56이다'이고, 올바르게는 '결정계수는 0.60이다'이다.
② X: 0.58은 정답값이 아니다. 틀린 표현은 '결정계수가 0.58이다'이고, 올바르게는 '결정계수는 0.60이다'이다. 이 값은 자유도를 잘못 잡아 $n-2$ 대신 다른 수를 넣을 때 나올 수 있는 오답이다.
③ O: $r^2 = 0.60$이므로 맞는 선택지이다. 회귀가 전체 변동의 60%를 설명한다는 뜻이다.
④ X: 0.62는 SSE나 자유도를 잘못 계산했을 때 나오는 값이다. 틀린 표현은 '결정계수가 0.62이다'이고, 올바르게는 '결정계수는 0.60이다'이다.

오답 포인트

단순선형회귀의 오차 자유도를 $n-2$가 아니라 $n-1$이나 다른 값으로 착각하면 오답이 된다. 또한 F값에서 바로 결정계수를 읽는 것이 아니라 $MSR \to SSR$, $MSE \to SSE$로 바꾼 뒤 r^2를 계산해야 한다.

다시 나오면 이렇게 판단

회귀 분산분석표 문제는 먼저 자유도부터 정리한 뒤 $F = MSR/MSE$로 MSR을 구하고, SSR과 SSE로 환산해서 $r^2 = SSR/SST$를 계산하면 된다.

암기 포인트

단순선형회귀에서는 회귀 자유도 1, 오차 자유도 $n-2$, 그리고 $r^2 = F/(F+n-2)$가 바로 연결된다.

NoteBase

문 16. 과거 연구에 따르면 병원에서 사망하는 사람들에 대한 사망원인 A, B, C, D, 기타의 비율이 각각 15%, 20%, 20%, 30%, 15%라고 한다. 어떤 병원의 사망자 중 무작위로 뽑힌 200명의 사망원인별 분포가 다음과 같다. 이 병원 사망자의 사망원인별 비율이 과거 연구 결과와 같다는 것을 귀무가설로 하여 유의수준 5%에서 검정하고자 한다. 카이제곱 검정통계량의 자유도와 검정 결과가 바르게 연결된 것은? (단, $\chi^2_\alpha(k)$는 자유도가 k인 카이제곱분포의 $100 \times (1-\alpha)$번째 백분위수를 나타내고, $\chi^2_{0.05}(4) = 9.49$, $\chi^2_{0.05}(5) = 11.07$이다)

사망원인	A	B	C	D	기타
사망자수	35	40	50	60	15

① 검정통계량의 자유도: 4, 검정 결과: 귀무가설을 기각함

② 검정통계량의 자유도: 4, 검정 결과: 귀무가설을 기각하지 않음

③ 검정통계량의 자유도: 5, 검정 결과: 귀무가설을 기각함

④ 검정통계량의 자유도: 5, 검정 결과: 귀무가설을 기각하지 않음

정답 1번

출제 포인트

이 문제는 주어진 사망원인 비율을 기대도수로 바꾼 뒤 카이제곱 적합도 검정의 자유도와 기각 여부를 판단하는 문제이다.

조건 정리

주어진 비율은 A 15%, B 20%, C 20%, D 30%, 기타 15%이다.
표본크기는 200명이다.
관측도수는 A 35, B 40, C 50, D 60, 기타 15이다.
귀무가설은 이 병원의 사망원인별 비율이 과거 연구 결과와 같다는 것이다.

사용할 공식

적합도 검정통계량은 $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$ 이다.
이 공식을 쓰는 이유는 범주형 자료의 관측도수와 귀무가설이 주는 기대도수의 차이를 한꺼번에 비교하는 문제이기 때문이다.

NoteBase

계산 과정

기대도수는 비율에 200을 곱해 A 30, B 40, C 40, D 60, 기타 30이 된다.
검정통계량은 $\chi^2 = \frac{(35-30)^2}{30} + \frac{(40-40)^2}{40} + \frac{(50-40)^2}{40} + \frac{(60-60)^2}{60} + \frac{(15-30)^2}{30}$ 이다.
따라서 $\chi^2 = \frac{25}{30} + 0 + \frac{100}{40} + 0 + \frac{225}{30} = 0.833\cdots + 2.5 + 7.5 = 10.833\cdots$ 이다.
범주 수가 5개이고 모비율을 자료로 추정한 것이 없으므로 자유도는 $5-1=4$이다.
임계값은 $\chi^2_{0.05}(4) = 9.49$ 이므로 $10.833\cdots > 9.49$ 이다. 따라서 귀무가설을 기각한다.

선지별 해설

① O: 자유도는 5개 범주에서 1을 뺀 4이고, 검정통계량이 10.833\cdots로 9.49보다 크므로 귀무가설을 기각한다.
② X: 자유도 4는 맞지만 검정 결과가 틀렸다. 틀린 표현은 '귀무가설을 기각하지 않음'이고, 올바르게는 검정통계량이 임계값보다 커서 '귀무가설을 기각함'이다.
③ X: 검정 결과는 기각이 맞지만 자유도가 틀렸다. 틀린 표현은 '자유도 5'이고, 올바르게는 범주 수 5에서 1을 뺀 '자유도 4'이다.
④ X: 자유도와 검정 결과가 모두 틀렸다. 틀린 표현은 '자유도 5, 귀무가설을 기각하지 않음'이고, 올바르게는 '자유도 4, 귀무가설을 기각함'이다.

오답 포인트

적합도 검정에서 범주가 5개이면 자유도를 그대로 5로 두기 쉽다. 그러나 자유도는 항상 범주 수에서 1을 뺀다. 또 기대도수를 비율에 표본크기를 곱해 구하지 않거나, 검정통계량과 임계값의 대소 비교를 거꾸로 읽으면 ②처럼 잘못 판단하게 된다.

다시 나오면 이렇게 판단

적합도 검정은 먼저 기대도수를 만든 뒤 χ^2 값을 계산하고, 그다음 자유도 $k-1$를 구해 임계값과 비교하는 순서로 판단한다. 모비율을 따로 추정하지 않았다면 자유도는 범주 수에서 1만 뺀다.

암기 포인트

적합도 검정은 기대도수 계산 후 χ^2 비교, 자유도는 기본적으로 범주 수 -1이다.

NoteBase

문 17. X_1과 X_2는 서로 독립이며 표준정규분포를 따르는 확률변수일 때, 옳지 않은 것은?

① $\frac{X_1 - 2X_2}{\sqrt{5}}$의 분포는 표준정규분포이다.
② $\frac{X_1 + X_2}{\sqrt{(X_1 - X_2)^2}}$의 분포는 자유도가 1인 t 분포이다.
③ $\frac{X_1^2}{X_2^2}$의 분포는 자유도가 1인 카이제곱분포이다.
④ $\frac{(X_1 - X_2)^2}{(X_1 + X_2)^2}$의 분포는 분자의 자유도가 1, 분모의 자유도가 1인 F 분포이다.

정답 3번

출제 포인트

독립인 표준정규변수의 선형결합과 제곱비가 어떤 표준분포로 이어지는지를 묻는 문제이다.

핵심 개념

X_1, X_2가 서로 독립인 표준정규분포이면 선형결합도 정규분포가 된다. 또 $U = (X_1 + X_2)/\sqrt{2}$, $V = (X_1 - X_2)/\sqrt{2}$로 두면 U, V는 서로 독립인 표준정규분포이고, U^2, V^2는 각각 자유도 1인 카이제곱분포가 된다. 따라서 정규분포와 카이제곱분포의 결합에서는 t 분포와 F 분포를 바로 연결할 수 있다.

선지별 해설

① O: $X_1 - 2X_2$는 독립인 정규변수의 선형결합이므로 정규분포이다. 평균은 0이고 분산은 $1 + 4 = 5$이므로 $(X_1 - 2X_2)/\sqrt{5}$의 분산은 1이 되어 표준정규분포가 된다.
② O: $U = (X_1 + X_2)/\sqrt{2}$, $V = (X_1 - X_2)/\sqrt{2}$로 두면 U, V는 서로 독립인 표준정규분포이다. 따라서 $(X_1 + X_2)/\sqrt{(X_1 - X_2)^2} = U/\sqrt{V^2} = U/\sqrt{\chi_1^2/1}$ 이므로 자유도가 1인 t 분포이다.
③ X: X_1^2와 X_2^2는 각각 자유도 1인 카이제곱분포이지만, 그 비 $(X_1^2/1)/(X_2^2/1)$는 자유도 $(1, 1)$인 F 분포이다. 틀린 표현은 '자유도가 1인 카이제곱분포'이고, 올바르게는 '자유도 $(1, 1)$인 F 분포'이다.
④ O: $X_1 + X_2$와 $X_1 - X_2$는 공분산이 0인 결합정규변수이므로 서로 독립이고 각각 $N(0, 2)$를 따른다. 따라서 $((X_1 - X_2)^2/2)/((X_1 + X_2)^2/2)$는 자유도 $(1, 1)$인 F 분포이며, 원식 $(X_1 - X_2)^2/(X_1 + X_2)^2$도 같은 분포이다.

오답 포인트

③처럼 '각각 카이제곱분포이다'와 '그 비도 카이제곱분포이다'를 혼동하기 쉽다. 카이제곱분포 두 개의 비는 카이제곱분포가 아니라 F 분포로 넘어간다는 점이 함정이다.

다시 나오면 이렇게 판단

정규변수의 선형결합이면 평균과 분산을 먼저 계산해 정규분포 여부를 본다. 제곱 하나는 카이제곱분포, 독립인 카이제곱 두 개의 비는 F 분포, 정규변수를 독립 카이제곱의 제곱근으로 나누면 t 분포로 연결하면 된다.

NoteBase

암기 포인트

표준정규의 제곱은 χ_1^2, 독립인 χ^2의 비는 F, 정규분포를 χ^2의 제곱근으로 나누면 t 분포이다.

NoteBase

문 18. 평균이 μ, 분산이 4인 정규모집단으로부터 크기 10인 임의표본을 뽑아 가설 $H_0 : \mu = 28$ 대 $H_1 : \mu > 28$을 유의수준 α에서 검정할 때, 모평균 $\mu = 30$에서의 검정력(power)은? (단, Z는 표준정규분포를 따르는 확률변수이고 z_α는 표준정규분포의 $100 \times (1-\alpha)$번째 백분위수를 나타낸다)

① $P(Z > z_\alpha - \sqrt{10})$
② $P(Z > z_\alpha + \sqrt{10})$
③ $P\left(Z > z_\alpha - \frac{\sqrt{10}}{2}\right)$
④ $P\left(Z > z_\alpha + \frac{\sqrt{10}}{2}\right)$

정답 1번

출제 포인트

이 문제는 분산이 알려진 정규모집단에서 오른쪽 단측검정의 기각역을 세운 뒤, 실제 모평균이 $\mu = 30$일 때 그 기각역에 들어갈 확률, 즉 검정력을 구하는 문제이다.

조건 정리

주어진 값:
- 모집단 분산은 4이므로 표준편차는 $\sigma = 2$이다.
- 표본크기는 $n = 10$이다.
- 가설은 $H_0 : \mu = 28, H_1 : \mu > 28$이다.
- 유의수준은 α이다.

구해야 할 값:
- $\mu = 30$에서의 검정력이다.

주의할 조건:
- 검정력은 H_0가 아니라 실제 평균 $\mu = 30$일 때 기각확률로 계산한다.

사용할 공식

검정통계량은 $Z = \frac{\bar{X} - 28}{2/\sqrt{10}}$이다.
기각역은 $Z > z_\alpha$이므로 $\bar{X} > 28 + \frac{2z_\alpha}{\sqrt{10}}$이다.
검정력은
$P\left(\bar{X} > 28 + \frac{2z_\alpha}{\sqrt{10}} \mid \mu = 30\right)$
로 계산한다.
이 공식을 쓰는 이유는 모집단이 정규분포이고 분산 4가 알려져 있으므로 모평균 검정에 표준정규 검정통계량을 바로 사용할 수 있기 때문이다.

NoteBase

계산 과정

1. 유의수준 α에서의 기각역을 세우면 $\frac{\bar{X}-28}{2/\sqrt{10}} > z_\alpha$이다.
2. 이를 $\bar{X}$에 대해 정리하면 $\bar{X} > 28 + \frac{2z_\alpha}{\sqrt{10}}$이다.
3. 실제 평균이 $\mu = 30$일 때 $\bar{X} \sim N\left(30, \frac{4}{10}\right)$이므로
$P\left(\bar{X} > 28 + \frac{2z_\alpha}{\sqrt{10}}\right) = P\left(\frac{\bar{X}-30}{2/\sqrt{10}} > \frac{28+\frac{2z_\alpha}{\sqrt{10}}-30}{2/\sqrt{10}}\right)$
4. 오른쪽 항을 정리하면
$\frac{28+\frac{2z_\alpha}{\sqrt{10}}-30}{2/\sqrt{10}} = z_\alpha - \sqrt{10}$
이므로 검정력은 $P(Z > z_\alpha - \sqrt{10})$이다.

선지별 해설

1번: O
- 기각역을 $\bar{X} > 28 + \frac{2z_\alpha}{\sqrt{10}}$로 세운 뒤, $\mu = 30$에서 표준화하면 $P(Z > z_\alpha - \sqrt{10})$가 된다.

2번: X
- $P(Z > z_\alpha + \sqrt{10})$는 실제 평균 30을 대입할 때 기준값 28과의 차이 $30-28=2$를 더해야 할 것을 반대로 처리한 식이다.
- 틀린 표현은 $z_\alpha + \sqrt{10}$이다. 올바르게는 $z_\alpha - \sqrt{10}$이다.

3번: X
- $P\left(Z > z_\alpha - \frac{\sqrt{10}}{2}\right)$는 모집단 표준편차를 2가 아니라 4로 잘못 써서 표준오차를 $\frac{4}{\sqrt{10}}$로 계산할 때 나오는 값이다.
- 틀린 표현은 $\frac{\sqrt{10}}{2}$이다. 올바르게는 분산 4의 제곱근을 써서 $\sqrt{10}$가 되어야 한다.

4번: X
- 이 식은 2번의 부호 오류와 3번의 표준편차 오류가 함께 들어간 경우이다.
- 틀린 표현은 $z_\alpha + \frac{\sqrt{10}}{2}$이다. 올바르게는 $z_\alpha - \sqrt{10}$이다.

오답 포인트

이 문항의 함정은 검정력 계산에서 기준이 되는 분포가 H_0 아래 분포가 아니라 실제 평균 $\mu = 30$ 아래 분포라는 점이다. 또 분산 4를 표준편차 4로 착각하면 $\frac{\sqrt{10}}{2}$가 나오므로 3번과 4번으로 흔들리기 쉽다.

다시 나오면 이렇게 판단

단측 검정의 검정력 문제는 먼저 H_0 기준으로 기각역을 세우고, 다음에는 실제 평균 아래에서 그 기각역에 들어갈 확률로 바꾸어 계산하면 된다. 이때 분산이 주어지면 반드시 표준편차 σ=분산를 다시 확인해야 한다.

암기 포인트

검정력은 기각역은 H_0로 만들고 확률은 실제 μ로 계산하는 값이다.

NoteBase

문 19. 어느 지역에서 남자와 여자의 체질량지수 평균이 같은지 알아보기 위하여 남자 22명과 여자 10명을 무작위로 추출한 후 체질량지수를 조사한 결과가 다음과 같다. 일원배치 분산분석법을 적용하여 분산분석표를 작성할 때 오차평균제곱(잔차평균제곱)의 값은?

성별	n	표본평균	표본분산
남자	22	21.30	1.50
여자	10	20.90	0.80

① 1.15

② 1.25

③ 1.28

④ 1.29

정답 4번

출제 포인트

이 문제는 일원배치 분산분석에서 집단별 표본분산으로 오차평균제곱 MSE를 구하는 문제이다.

조건 정리

- 주어진 값: 남자 $n_1 = 22, \bar{x}_1 = 21.30, s_1^2 = 1.50$, 여자 $n_2 = 10, \bar{x}_2 = 20.90, s_2^2 = 0.80$이다.
- 구해야 할 값: 오차평균제곱 MSE이다.
- 주의할 조건: 일원배치 분산분석의 오차제곱합은 $\sum(n_i - 1)s_i^2$, 잔차자유도는 $N - k$이다.

사용할 공식

$SSE = \sum_{i=1}^{k} (n_i - 1)s_i^2, \qquad MSE = \frac{SSE}{N-k}$

이 공식을 쓰는 이유는 오차평균제곱이 집단 내 변동을 자유도로 나눈 pooled variance이기 때문이다.

계산 과정

1. 오차제곱합은

$SSE = (22 - 1) \times 1.50 + (10 - 1) \times 0.80 = 21 \times 1.50 + 9 \times 0.80 = 31.5 + 7.2 = 38.7$ 이다.

2. 전체 표본수는 $N = 22 + 10 = 32$, 집단 수는 $k = 2$이므로 잔차자유도는 $N - k = 32 - 2 = 30$이다.

3. 따라서 $MSE = \frac{38.7}{30} = 1.29$ 이다.

NoteBase

선지별 해설

① X
- 1.15는 $\frac{1.50+0.80}{2}=1.15$인 단순평균이다. 오차평균제곱은 두 표본분산의 산술평균이 아니라 집단 내 자유도를 반영한 pooled variance이다.
- 틀린 표현은 "두 표본분산을 그대로 평균낸 값"이다. 올바르게는 (n_i-1) 가중과 $N-k$ 자유도를 반영한 값을 써야 한다.

② X
- 1.25는 $\frac{38.7}{31}\approx 1.25$로 계산한 값이다. 이는 분모를 총자유도 $N-1$로 둔 경우이며, 일원배치 분산분석의 오차평균제곱 분모와 다르다.
- 틀린 표현은 "분모를 31로 둔 계산"이다. 올바르게는 분모를 잔차자유도 $N-k=30$으로 둔다.

③ X
- 1.28은 $\frac{22\times 1.50+10\times 0.80}{32}=\frac{41}{32}\approx 1.28$에서 나오는 값이다. 이는 표본분산 앞에 n_i를 곱해 집단 내 자유도 (n_i-1)를 무시한 계산이다.
- 틀린 표현은 "오차제곱합을 $\sum n_i s_i^2$로 계산한 값"이다. 올바르게는 $\sum(n_i-1)s_i^2$를 사용한다.

④ O
- $SSE=38.7$, $N-k=30$이므로 $MSE=38.7/30=1.29$이다. 계산 결과와 일치하므로 옳다.

오답 포인트

이 문항의 함정은 표본분산이 보이면 그냥 평균을 내거나 분모를 습관적으로 $N-1$로 두는 실수이다. ANOVA의 오차평균제곱은 집단 내 자유도 (n_i-1)로 만든 오차제곱합을 잔차자유도 $(N-k)$로 나누는 값이라는 점이 핵심이다.

다시 나오면 이렇게 판단

집단별 표본분산이 주어진 ANOVA 문항은 먼저 $SSE=\sum(n_i-1)s_i^2$를 만들고, 다음에 분모를 반드시 $N-k$로 두는 순서로 판단한다.

암기 포인트

오차평균제곱은 "집단 내 제곱합 ÷ 잔차자유도"이므로 $MSE=SSE/(N-k)$, $SSE=\sum(n_i-1)s_i^2$이다.

NoteBase

문 20. 다음 자료에 회귀모형 $Y_i = \beta X_i^2 + \varepsilon_i,\ i = 1, 2, 3, 4$를 적용할 때, 최소제곱법에 의한 β의 추정값은?

X	-1	0	1	2
Y	3	1	4	14

① $\frac{10}{3}$
② $\frac{7}{2}$
③ $\frac{18}{5}$
④ $\frac{29}{6}$

정답 2번

출제 포인트

절편이 없는 회귀모형 $Y_i = \beta X_i^2 + \varepsilon_i$에서 최소제곱추정량 $\hat{\beta}$를 구하는 문제이다.

조건 정리

주어진 값은 $X = (-1, 0, 1, 2), Y = (3, 1, 4, 14)$이다.
모형의 설명변수는 X_i가 아니라 $Z_i = X_i^2$이므로 $Z = (1, 0, 1, 4)$로 바꿔 본다.
구해야 할 값은 절편이 없는 회귀식의 $\hat{\beta}$이다.

사용할 공식

절편이 없는 단순회귀 $Y_i = \beta Z_i + \varepsilon_i$의 최소제곱추정량은 $\hat{\beta} = \frac{\sum Z_i Y_i}{\sum Z_i^2}$이다.
이 문제에서는 $Z_i = X_i^2$이므로 $\hat{\beta} = \frac{\sum X_i^2 Y_i}{\sum X_i^4}$를 쓴다.

계산 과정

$X_i^2 = (1, 0, 1, 4), X_i^4 = (1, 0, 1, 16)$이다.
따라서 $\sum X_i^2 Y_i = 1 \cdot 3 + 0 \cdot 1 + 1 \cdot 4 + 4 \cdot 14 = 63$이다.
또한 $\sum X_i^4 = 1 + 0 + 1 + 16 = 18$이므로 $\hat{\beta} = \frac{63}{18} = \frac{7}{2}$이다.

NoteBase

선지별 해설

① X: $\frac{10}{3}$은 계산된 $\hat{\beta}=\frac{7}{2}$와 다르다. 틀린 표현은 추정값이 $\frac{10}{3}$이라는 부분이다. 올바르게는 $\hat{\beta}=\frac{7}{2}$이다.
② O: 최소제곱법으로 구한 추정값이 $\frac{63}{18}=\frac{7}{2}$이므로 맞는 선지이다.
③ X: $\frac{18}{5}$는 계산 결과와 다르다. 틀린 표현은 추정값이 $\frac{18}{5}$이라는 부분이다. 올바르게는 $\hat{\beta}=\frac{7}{2}$이다.
④ X: $\frac{29}{6}$은 $\sum X_iY_i/\sum X_i^2$처럼 원래 모형을 $Y_i=\beta X_i+\varepsilon_i$로 잘못 본 경우의 오답이다. 이 문제의 설명변수는 X_i^2이므로 올바르게는 $\hat{\beta}=\sum X_i^2Y_i/\sum X_i^4=\frac{7}{2}$이다.

오답 포인트

이 문제의 함정은 설명변수를 X_i로 보면 안 되고 반드시 X_i^2로 바꿔야 한다는 점이다. 특히 ④처럼 $\sum X_iY_i/\sum X_i^2$ 공식을 적용하면 다른 회귀모형의 추정량을 구하게 된다.

다시 나오면 이렇게 판단

절편이 없는 회귀식이 보이면 먼저 설명변수를 모형의 형태에 맞는 새 변수로 치환한다. 그다음 $\hat{\beta}=\sum z_iY_i/\sum z_i^2$ 공식을 적용해 분자와 분모를 따로 계산하면 된다.

암기 포인트

$Y=\beta z+\varepsilon$ 꼴의 절편 없는 회귀에서는 $\hat{\beta}=\sum zY/\sum z^2$이고, 이 문제에서는 $z=X^2$이다.

NoteBase