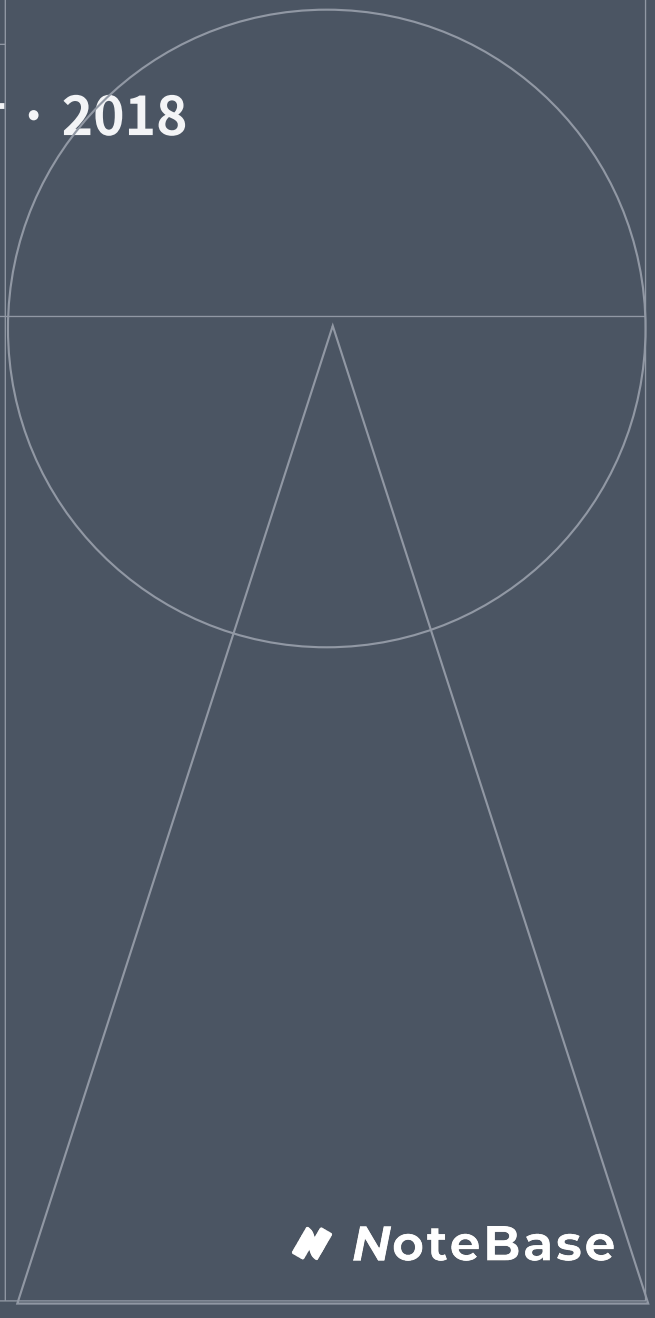
기출OX 해설집

공무원/국가직/7급 · 통계학 · 2018



NoteBase

기출OX 해설집

공무원/국가직/7급 · 통계학 · 2018

공무원/국가직/7급 · 통계학 · 2018

문 1. 다음 자료의 중앙값과 산술평균에 대한 설명으로 옳은 것은?

13, 17, 16, 19, 14, 11, 15

① 중앙값이 산술평균보다 크다.

② 중앙값이 산술평균보다 작다.

③ 중앙값과 산술평균이 같다.

④ 알 수 없다.

정답 3번

출제 포인트

이 문제는 주어진 자료의 중앙값과 산술평균을 각각 계산한 뒤 두 값을 비교하는 문제이다.

조건 정리

주어진 값은 13, 17, 16, 19, 14, 11, 15이다.
구해야 할 값은 중앙값과 산술평균의 크기 관계이다.
중앙값은 자료를 크기순으로 나열한 뒤 가운데 값을 찾고, 산술평균은 전체 합을 자료 수로 나누어 구한다.

사용할 공식

중앙값은 자료를 오름차순으로 정렬했을 때 한가운데 있는 값이다.
산술평균은 자료의 합자료의 수이다.
이 공식을 쓰는 이유는 문제에서 두 대표값의 정의에 따라 실제 값을 구해 비교해야 하기 때문이다.

계산 과정

1. 자료를 크기순으로 정리하면 11, 13, 14, 15, 16, 17, 19이다.
2. 자료가 7개이므로 가운데인 4번째 값은 15이고, 중앙값은 15이다.
3. 전체 합은 $13 + 17 + 16 + 19 + 14 + 11 + 15 = 105$이고, 산술평균은 $\frac{105}{7} = 15$이다.
4. 중앙값과 산술평균이 모두 15이므로 두 값은 같다.

NoteBase

선지별 해설

① X: 중앙값은 15, 산술평균도 15이므로 중앙값이 더 크지 않다. 틀린 표현은 "중앙값이 산술평균보다 크다"이고, 올바르게는 "중앙값과 산술평균이 같다"이다.
② X: 중앙값은 15, 산술평균도 15이므로 중앙값이 더 작지 않다. 틀린 표현은 "중앙값이 산술평균보다 작다"이고, 올바르게는 "중앙값과 산술평균이 같다"이다.
③ O: 중앙값은 15이고 산술평균도 15이므로 두 값이 같다는 설명이 옳다.
④ X: 주어진 자료만으로도 중앙값과 산술평균을 모두 계산할 수 있으므로 알 수 없는 것이 아니다. 틀린 표현은 "알 수 없다"이고, 올바르게는 "중앙값과 산술평균이 같다"이다.

오답 포인트

자료를 정렬하지 않고 중앙값을 찾거나, 합은 맞게 구하고도 자료 수 7로 나누지 않으면 오답이 나온다. 이 문항의 함정은 중앙값과 평균을 감으로 비교하게 만드는 점이며, 실제로 계산하면 두 값이 정확히 같아진다.

다시 나오면 이렇게 판단

대표값 비교 문제는 먼저 자료를 정렬해 중앙값을 구하고, 다음에 전체 합을 자료 수로 나누어 평균을 구한 뒤 마지막에 두 값을 비교하는 순서로 판단한다.

암기 포인트

중앙값은 정렬 후 가운데 값이고, 산술평균은 합을 개수로 나눈 값이다.

NoteBase

문 2. 확률변수 X가 평균이 12이고 분산이 25인 정규분포를 따른다. 확률변수 $Y = 20 - 2X$일 때, $P(Y < 16)$과 같은 값은? (단, Z는 표준정규분포를 따르는 확률변수이다)

① $P(X < 2)$

② $P(X > -2)$

③ $P(Z > 2)$

④ $P(Z > -2)$

정답 4번

출제 포인트

이 문제는 정규분포를 따르는 확률변수를 일차변환한 뒤 부등식을 표준화하여 같은 확률을 찾는 문제이다.

조건 정리

주어진 값은 $X \sim N(12, 25)$이므로 평균은 12, 표준편차는 5이다.
구해야 할 값은 $P(Y < 16)$이다.
주의할 조건은 $Y = 20 - 2X$에서 계수가 음수이므로 부등호 방향이 바뀐다는 점이다.

사용할 공식

표준화 공식은 $Z = \frac{X - \mu}{\sigma}$이다.
또한 $P(Y < 16) = P(20 - 2X < 16)$처럼 먼저 X에 대한 부등식으로 바꾼 뒤 표준화한다.
이 공식을 쓰는 이유는 문제에서 Z가 표준정규분포를 따른다고 주어졌고, 보기들도 X 또는 Z의 확률 형태로 제시되어 있기 때문이다.

계산 과정

1. $P(Y < 16) = P(20 - 2X < 16)$이다.
2. $-2X < -4$이므로 음수 -2로 나누면 부등호가 바뀌어 $X > 2$이다.
3. 이를 표준화하면 $P(X > 2) = P\left(Z > \frac{2 - 12}{5}\right) = P(Z > -2)$이다.
4. 따라서 $P(Y < 16)$와 같은 값은 $P(Z > -2)$이다.

NoteBase

선지 판단

1번 X: $P(X<2)$는 계산 결과인 $P(X>2)$와 부등호 방향이 반대이다. 틀린 표현은 $X<2$이고, 올바르게는 $X>2$이다.
2번 X: $P(X>-2)$는 기준값이 잘못되었다. $Y<16$을 정리하면 $X>2$가 되므로 틀린 표현은 -2이고, 올바르게는 2이다.
3번 X: $P(Z>2)$는 표준화 과정에서 부호를 잘못 처리한 경우이다. $\frac{2-12}{5}=-2$이므로 틀린 표현은 $Z>2$이고, 올바르게는 $Z>-2$이다.
4번 O: $P(Z>-2)$는 $P(Y<16)=P(X>2)$를 표준화한 결과와 정확히 같다.

오답 포인트

가장 흔한 함정은 $20-2X<16$에서 음수로 나눌 때 부등호가 바뀐다는 점을 놓치는 것이다. 또 표준화에서 $2-12=-10$의 부호를 놓치면 3번처럼 $P(Z>2)$를 고르게 된다.

다시 나오면 이렇게 판단

정규분포 계산형은 먼저 원하는 사건을 원래 변수 하나의 부등식으로 정리하고, 그다음 평균을 빼고 표준편차로 나누어 Z로 바꾼다. 이때 음수로 나누는 순간 부등호 방향이 바뀌는지부터 확인하면 된다.

암기 포인트

일차변환 확률은 먼저 원변수 부등식으로 되돌리고, 표준화는 $Z=\frac{X-\mu}{\sigma}$로 처리한다.

NoteBase

문 3. 다음은 분산이 같고 서로 독립인 두 정규모집단 A와 B에서 임의로 표본을 추출하여 정리한 결과이다. 두 모집단의 공통분산에 대한 추정량인 합동표본분산(pooled sample variance)과 합동표본분산의 자유도를 구하면? (단, 두 모집단의 공통분산은 알려져 있지 않다)

모집단	표본크기	표본분산
A	5	2
B	6	4

① 합동표본분산: $\frac{28}{9}$, 자유도: 9

② 합동표본분산: $\frac{28}{9}$, 자유도: 10

③ 합동표본분산: $\frac{34}{10}$, 자유도: 9

④ 합동표본분산: $\frac{34}{10}$, 자유도: 10

정답 1번

출제 포인트

이 문제는 분산이 같은 두 독립 정규모집단에서 합동표본분산과 그 자유도를 구하는 문제이다.

조건 정리

주어진 값: A는 표본크기 5, 표본분산 2이고 B는 표본크기 6, 표본분산 4이다.
구해야 할 값: 합동표본분산과 자유도이다.
주의할 조건: 두 모집단의 공통분산은 같지만 알려져 있지 않으므로 각 표본분산을 자유도 $n_i - 1$로 가중한 합동표본분산 공식을 사용한다.

사용할 공식

합동표본분산은 $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$이다.
자유도는 $n_1 + n_2 - 2$이다.
이 공식을 쓰는 이유: 두 표본이 서로 독립이고 두 모집단의 분산이 같다고 가정하므로, 각 표본분산을 각자의 자유도로 가중해 하나의 공통분산 추정량으로 합친다.

NoteBase

계산 과정

1. A의 가중합은 $(5-1)\times 2=8$이다.
2. B의 가중합은 $(6-1)\times 4=20$이다.
3. 분자는 $8+20=28$, 분모는 자유도 합인 $5+6-2=9$이므로 합동표본분산은 $s_p^2=\frac{28}{9}$이다.
4. 합동표본분산의 자유도도 같은 값인 9이다.

선지별 해설

① O. 합동표본분산은 $\frac{28}{9}$, 자유도는 9이므로 제시된 값이 모두 맞다.
② X. 합동표본분산 $\frac{28}{9}$는 맞지만 자유도 10이 틀리다. 자유도는 $n_1+n_2-2=5+6-2=9$이다. 틀린 표현 '자유도 10' → 올바르게는 '자유도 9'이다.
③ X. 자유도 9는 맞지만 합동표본분산 $\frac{34}{10}$이 틀리다. 34는 $5\times 2+6\times 4$처럼 표본크기로 직접 가중했을 때 나오는 값이고, 합동표본분산은 (n_i-1)로 가중해야 한다. 틀린 표현 '합동표본분산 $\frac{34}{10}$' → 올바르게는 '합동표본분산 $\frac{28}{9}$'이다.
④ X. 합동표본분산과 자유도를 모두 틀렸다. $\frac{34}{10}$은 표본크기로 가중하고 분모를 10으로 둔 오답값이고, 자유도도 10이 아니라 9이다. 틀린 표현 '합동표본분산 $\frac{34}{10}$, 자유도 10' → 올바르게는 '합동표본분산 $\frac{28}{9}$, 자유도 9'이다.

오답 포인트

표본분산을 합칠 때 표본크기 n로 곧바로 가중하거나 분모를 전체 표본크기에서 1만 빼는 실수가 자주 나온다. 이 문제의 함정은 각 집단의 자유도 $n-1$를 써야 한다는 점과 합동표본분산의 분모가 자유도 합이라는 점이다.

다시 나오면 이렇게 판단

분산이 같은 두 집단의 합동표본분산 문제는 먼저 각 집단의 자유도 $n-1$를 구한 뒤, $(n_1-1)s_1^2+(n_2-1)s_2^2$를 계산하고 마지막에 n_1+n_2-2로 나누는 순서로 판단한다.

암기 포인트

합동표본분산은 표본크기 자체가 아니라 자유도 $n-1$로 가중하고, 분모도 자유도 합 n_1+n_2-2이다.

NoteBase

문 4. 자료 $X_i = i$와 다음의 변환으로 생성된 Y_i에 대하여 자료 (X_i, Y_i) 사이의 표본상관계수가 가장 큰 것은? (단, $i = 1, 2, \cdots, 9$)

① $Y_i = |X_i - 5|$

② $Y_i = -(X_i - 5)^2$

③ $Y_i = \frac{X_i}{2} + 1$

④ $Y_i = -X_i + 1$

정답 3번

출제 포인트

이 문제는 표본상관계수의 성질을 이용해 변환된 Y_i와 X_i의 선형 관계 방향과 강도를 비교하는 문제이다.

조건 정리

주어진 값은 $X_i = i$ $(i = 1, 2, \cdots, 9)$이다.
X_i의 평균은 $\bar{X} = 5$이고, 중심화하면 $X_i - 5 = -4, -3, -2, -1, 0, 1, 2, 3, 4$가 된다.
표본상관계수는 공분산의 부호와 선형성의 정도로 판단할 수 있다.

사용할 공식

표본상관계수는 $r_{XY} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$ 이다.
이 공식을 쓰는 이유는 각 변환이 X와 Y 사이의 공분산 부호와 선형성을 어떻게 바꾸는지 직접 비교할 수 있기 때문이다.

계산 과정

1. X_i를 평균 5를 기준으로 보면 좌우가 완전히 대칭이다.
2. ① $Y_i = |X_i - 5|$와 ② $Y_i = -(X_i - 5)^2$는 $X_i = 5 - k$와 $X_i = 5 + k$에서 같은 Y_i값을 가지므로, $(X_i - \bar{X})(Y_i - \bar{Y})$가 좌우에서 서로 상쇄되어 공분산이 0이 된다. 따라서 상관계수도 0이다.
3. ③ $Y_i = \frac{X_i}{2} + 1$은 X_i의 양의 일차변환이므로 모든 점이 정확히 같은 증가 직선 위에 놓여 상관계수는 1이다.
4. ④ $Y_i = -X_i + 1$은 X_i의 음의 일차변환이므로 모든 점이 정확히 같은 감소 직선 위에 놓여 상관계수는 -1이다.
5. 네 값은 $0, 0, 1, -1$이므로 가장 큰 표본상관계수는 1이고 정답은 ③이다.

NoteBase

선지별 해설

① X
-4와 4, -3과 3처럼 $X_i - 5$가 부호만 반대인 쌍에서 $Y_i = |X_i - 5|$는 동일하다. 그래서 공분산 합이 서로 상쇄되어 상관계수는 0이다.
틀린 표현은 이 변환이 가장 큰 상관계수를 만든다는 판단이다. 올바르게는 선형 증가 관계가 아니라 V자형 대칭 관계이므로 상관계수는 0이다.

② X
$Y_i = -(X_i - 5)^2$도 $X_i = 5$를 기준으로 좌우대칭인 역포물선 형태이다. 좌우에서 같은 Y_i값이 나오므로 공분산이 0이고 상관계수도 0이다.
틀린 표현은 이 변환이 가장 큰 상관계수를 만든다는 판단이다. 올바르게는 완전한 선형관계가 아니라 대칭적 비선형관계이므로 상관계수는 0이다.

③ O
$Y_i = \frac{X_i}{2} + 1$은 X_i의 양의 일차변환이다. 양의 일차변환은 점들의 순서와 직선성을 그대로 보존하므로 표본상관계수는 정확히 1이다.
따라서 네 선택지 중 표본상관계수가 가장 크다.

④ X
$Y_i = -X_i + 1$은 완전한 음의 일차관계이므로 상관계수는 -1이다. 절댓값은 1이지만 값 자체는 가장 큰 것이 아니라 가장 작은 쪽이다.
틀린 표현은 이 변환이 가장 큰 상관계수를 만든다는 판단이다. 올바르게는 완전한 음의 선형관계이므로 상관계수는 -1이다.

오답 포인트

절댓값이 1이면 무조건 크다고 착각하기 쉽다. 상관계수는 절댓값이 아니라 실제 값의 대소를 비교해야 하므로 $1 > -1$을 먼저 확인해야 한다. 또 대칭적인 비선형관계를 선형관계로 오해하면 ①과 ②를 잘못 고르기 쉽다.

다시 나오면 이렇게 판단

상관계수 비교 문제는 먼저 양의 일차변환인지, 음의 일차변환인지, 대칭적 비선형변환인지를 구분하면 된다. 양의 일차변환이면 1, 음의 일차변환이면 -1, 평균을 기준으로 좌우대칭이면 공분산이 상쇄되어 0으로 판단할 수 있다.

암기 포인트

양의 일차변환 $aX + b$ $(a > 0)$은 상관계수 1, 음의 일차변환 $aX + b$ $(a < 0)$은 상관계수 -1, 좌우대칭 비선형변환은 공분산 0이 자주 나온다.

NoteBase

문 5. 공정한 주사위를 5 또는 6의 눈이 처음으로 나올 때까지 던지는 횟수를 확률변수 X라 할 때, 조건부확률 $P(X \le 2 \mid X \le 4)$의 값은?

① $\frac{65}{81}$
② $\frac{9}{13}$
③ $\frac{5}{9}$
④ $\frac{4}{9}$

정답 2번

출제 포인트

이 문제는 성공확률이 $\frac{1}{3}$인 기하분포에서 조건부확률 $P(X \le 2 \mid X \le 4)$를 구하는 문제이다.

조건 정리

- 주어진 값: 한 번 던져 5 또는 6이 나올 확률은 $\frac{2}{6} = \frac{1}{3}$이다.
- 구해야 할 값: $P(X \le 2 \mid X \le 4)$이다.
- 주의할 조건: $X \le 2$이면 항상 $X \le 4$이므로 $P(X \le 2 \mid X \le 4) = \frac{P(X \le 2)}{P(X \le 4)}$로 계산한다.

사용할 공식

$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$, $P(X \le n) = 1 - (1-p)^n$이다.

이 공식을 쓰는 이유:
처음으로 5 또는 6이 나올 때까지의 횟수 X는 성공확률 $p = \frac{1}{3}$인 기하분포이고, 사건 $A = \{X \le 2\}$, $B = \{X \le 4\}$에서는 $A \subset B$이므로 $P(A \cap B) = P(A)$가 된다.

계산 과정

1. $P(X \le 2) = 1 - \left(\frac{2}{3}\right)^2 = 1 - \frac{4}{9} = \frac{5}{9}$이다.
2. $P(X \le 4) = 1 - \left(\frac{2}{3}\right)^4 = 1 - \frac{16}{81} = \frac{65}{81}$이다.
3. 따라서
$P(X \le 2 \mid X \le 4) = \frac{P(X \le 2)}{P(X \le 4)} = \frac{\frac{5}{9}}{\frac{65}{81}} = \frac{5}{9} \cdot \frac{81}{65} = \frac{9}{13}$
이다.

NoteBase

선지별 해설

① X
- $\frac{65}{81}$는 $P(X \le 4)$ 값 자체이다. 조건부확률에서는 이 값을 분모로 한 번 더 나누어야 하므로 틀리다.
- 틀린 표현은 $\frac{65}{81}$이다. 올바르게는 $P(X \le 2 \mid X \le 4) = \frac{9}{13}$이다.

② O
- 계산 결과가 $\frac{9}{13}$이므로 맞다. $X \le 2$가 $X \le 4$ 안에 포함된다는 점을 이용하면 바로 조건부확률을 구할 수 있다.

③ X
- $\frac{5}{9}$는 $P(X \le 2)$ 값일 뿐이고 조건 $X \le 4$를 반영하지 않은 값이다. 조건부확률의 분모 $P(X \le 4)$를 빠뜨려서 나온 오답이다.
- 틀린 표현은 $\frac{5}{9}$이다. 올바르게는 $\frac{\frac{5}{9}}{\frac{65}{81}} = \frac{9}{13}$이다.

④ X
- $\frac{4}{9}$는 $P(X > 2) = \left(\frac{2}{3}\right)^2$로, 2회 안에 성공하지 못할 확률이다. 구하려는 조건부확률과 사건 자체가 다르므로 틀리다.
- 틀린 표현은 $\frac{4}{9}$이다. 올바르게는 $P(X \le 2 \mid X \le 4) = \frac{9}{13}$이다.

오답 포인트

조건부확률을 묻는데 $P(X \le 2)$나 $P(X \le 4)$만 따로 계산하고 끝내는 실수가 가장 흔하다. 특히 $X \le 2 \subset X \le 4$라는 포함관계를 보면 $P(A \mid B) = \frac{P(A)}{P(B)}$로 바로 가야 하는데, 분모를 빠뜨리거나 여사건 $P(X > 2)$와 섞으면 ③, ④ 같은 오답이 나온다.

다시 나오면 이렇게 판단

처음 성공할 때까지의 횟수 문제가 나오면 먼저 기하분포인지 확인하고, 그다음 조건부확률의 두 사건이 포함관계인지 본다. 포함관계가 잡히면 $P(A \mid B) = \frac{P(A)}{P(B)}$로 정리한 뒤 각 누적확률을 $1 - (1-p)^n$으로 계산하면 된다.

암기 포인트

기하분포에서 $P(X \le n) = 1 - (1-p)^n$이고, 포함관계 $A \subset B$이면 $P(A \mid B) = \frac{P(A)}{P(B)}$이다.

NoteBase

문 6. 평균이 μ, 분산이 σ^2인 정규모집단에서 8개의 임의표본을 추출하여 얻은 표본평균이 $\bar{x}$이고 표본표준편차가 s일 때, 가설 $H_0 : \mu = \mu_0$ 대 $H_1 : \mu > \mu_0$에 대한 검정통계량의 값은

$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{8}}$
이다. T_ν를 자유도 ν인 t 분포를 따르는 확률변수라 할 때, 유의확률(p-값)은? (단, 분산 σ^2은 알려져 있지 않으며, t_0은 0이 아니다)

① $P(T_7 > t_0)$

② $P(T_7 > -t_0)$

③ $P(T_8 > t_0)$

④ $P(T_8 > -t_0)$

정답 1번

출제 포인트

이 문제는 모분산을 모르는 정규모집단의 일표본 평균검정에서 자유도와 우측검정의 p-값 정의를 정확히 적용하는지를 묻는다.

조건 정리

주어진 값은 표본크기 8, 검정통계량 $t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{8}}$, 대립가설 $H_1 : \mu > \mu_0$이다.
구해야 할 값은 이 검정의 p-값이다.
주의할 조건은 모분산을 모르므로 정규분포가 아니라 t분포를 쓰고, 자유도는 $n - 1$이라는 점이다.

사용할 공식

우측 단측검정의 p-값은 $p = P(T_\nu > t_0)$이다.
여기서 T_ν는 귀무가설 아래의 검정통계량 분포이고, 모분산을 모를 때 일표본 t검정의 자유도는 $\nu = n - 1$이다.
이 공식을 쓰는 이유는 문제의 검정통계량이 표본표준편차 s를 사용한 일표본 평균검정 통계량이기 때문이다.

계산 과정

1. 표본크기가 8이므로 자유도는 $\nu = 8 - 1 = 7$이다.
2. 대립가설이 $H_1 : \mu > \mu_0$이므로 기각역과 p-값은 모두 우측꼬리 기준으로 잡는다.
3. 따라서 p-값은 관측된 t_0보다 크거나 같은 방향의 확률이고, 연속분포이므로 보기 표기대로 $P(T_7 > t_0)$로 쓸 수 있다.

NoteBase

선지 판단

1번 O: 표본크기 8에서 모분산 미지인 일표본 t검정이므로 자유도는 7이다. 또 우측 단측검정 $H_1 : \mu > \mu_0$의 p-값은 우측꼬리확률이므로 $P(T_7 > t_0)$가 맞다.
2번 X: $-t_0$를 넣으면 관측통계량의 부호를 뒤집은 확률이 되어 우측검정의 p-값이 아니다. 틀린 표현은 $-t_0$이고, 올바르게는 t_0이다.
3번 X: 자유도를 8로 둔 점이 틀리다. 표본표준편차 s를 사용한 일표본 t검정의 자유도는 $n-1$이므로 틀린 표현은 T_8이고, 올바르게는 T_7이다.
4번 X: 자유도를 8로 둔 점도 틀리고, 꼬리확률의 기준점을 $-t_0$로 둔 점도 틀리다. 올바른 표현은 $P(T_7 > t_0)$이다.

오답 포인트

함정은 두 가지이다. 첫째, 표본크기 8을 그대로 자유도로 착각하면 T_8을 고르게 된다. 둘째, t분포의 대칭성만 보고 $-t_0$를 넣으면 1번과 2번이 비슷해 보이는데, 문제에서 $t_0 \neq 0$라고 한 이유가 바로 이 혼동을 막기 위한 것이다.

다시 나오면 이렇게 판단

일표본 평균검정 문제는 먼저 모분산의 기지 여부를 보고 z검정인지 t검정인지 정한다. 그다음 자유도를 $n-1$로 놓고, 대립가설 방향이 $>$, $<$, $\neq$ 중 무엇인지에 따라 우측꼬리, 좌측꼬리, 양측꼬리 p-값을 정하면 된다.

암기 포인트

모분산 미지의 일표본 평균검정은 자유도 $n-1$, 우측검정의 p-값은 관측통계량 오른쪽 꼬리확률이다.

NoteBase

문 7. 표본의 크기가 10인 어느 자료에 단순선형회귀모형 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ i = 1, 2, \cdots, 10$을 적용하여 최소제곱법으로 얻은 오차제곱합(잔차제곱합)이 36일 때, Y_i의 분산에 대한 불편추정량의 값은? (단, ε_i는 서로 독립이며 평균이 0, 분산이 σ^2인 정규분포를 따른다)

① 2.0

② 3.6

③ 4.0

④ 4.5

정답 4번

출제 포인트

이 문제는 단순선형회귀에서 오차분산 σ^2의 불편추정량이 잔차제곱합을 자유도 $n-2$로 나눈 값임을 묻는 문제이다.

조건 정리

주어진 값은 표본크기 10, 잔차제곱합 36이다.
구해야 할 값은 Y_i의 분산, 즉 오차분산 σ^2의 불편추정량이다.
단순선형회귀에서는 β_0, β_1 두 개의 모수를 추정하므로 잔차의 자유도는 $10-2=8$이다.

사용할 공식

사용할 공식은 $\hat{\sigma}^2 = \frac{SSE}{n-2}$ 이다.
이 공식을 쓰는 이유는 단순선형회귀에서 절편과 기울기 두 개를 추정한 뒤 남는 잔차 자유도가 $n-2$이기 때문이다.

계산 과정

1. 잔차제곱합은 $SSE = 36$이다.
2. 표본크기 10에서 추정한 회귀계수는 두 개이므로 자유도는 $10-2=8$이다.
3. 따라서 불편추정량은 $\hat{\sigma}^2 = \frac{36}{8} = 4.5$ 이다.

NoteBase

선지 판단

1번 X: 2.0은 $\frac{36}{18}$에 해당하는 값으로, 단순선형회귀의 분산 불편추정량에서 쓰는 분모가 아니다. 2.0이 아니라 4.5가 되어야 한다.
2번 X: 3.6은 $\frac{36}{10}$으로 표본크기 n으로 나눈 값이다. 회귀모형에서는 모수 두 개를 추정했으므로 n이 아니라 $n-2$로 나누어야 하므로 3.6이 아니라 4.5가 맞다.
3번 X: 4.0은 $\frac{36}{9}$로 자유도를 $n-1$로 잘못 둔 값이다. 단순선형회귀의 잔차 자유도는 $n-2$이므로 4.0이 아니라 4.5가 맞다.
4번 O: $\frac{36}{10-2}=\frac{36}{8}=4.5$이므로 불편추정량의 값과 일치한다.

오답 포인트

표본분산 공식을 떠올려 분모를 $n-1$로 두거나, 잔차제곱합을 그냥 n으로 나누는 실수가 가장 흔하다.
회귀에서는 추정한 모수의 개수만큼 자유도가 줄어든다는 점이 핵심 함정이다.

다시 나오면 이렇게 판단

회귀모형의 분산 추정 문제는 먼저 잔차제곱합인지 확인하고, 다음에 추정한 모수 개수만큼 자유도를 차감한다.
단순선형회귀이면 항상 분모를 $n-2$로 둔다.

암기 포인트

단순선형회귀의 오차분산 불편추정량은 잔차제곱합을 $n-2$로 나눈 값이다.

NoteBase

문 8. 다음은 어느 제품에 포함된 두 인자 A, B의 수준에 따라 제품의 품질이 어떤 영향을 받는지 알아보기 위해, 반복이 있는 이원배치 분산분석법을 적용하여 얻은 분산분석표의 일부이다. 이에 대한 설명으로 옳지 않은 것은?

요인	제곱합	자유도	F-값	p-값
인자 A	14	1		0.0155
인자 B				0.0987
교호작용 A × B	2	1		0.3293
오차(잔차)	40	20		

① 인자 A의 F-값은 7이다.

② 인자 B는 두 가지 수준을 갖는다.

③ 유의수준 5%에서 제품의 품질은 인자 B의 수준에 따라 영향을 받는다고 할 수 있다.

④ 유의수준 5%에서 인자 A와 인자 B의 교호작용에 따른 효과가 있다고 할 수 없다.

정답 3번

출제 포인트

이 문제는 반복이 있는 이원배치 분산분석표에서 잔차평균제곱을 구해 F-값을 복원하고, 각 요인의 p-값으로 주효과와 교호작용의 유의성을 판단하는 문제이다.

조건 정리

주어진 값은 인자 A의 제곱합 14, 자유도 1, 교호작용 $A \times B$의 제곱합 2, 자유도 1, 오차의 제곱합 40, 자유도 20이다.
구해야 할 내용은 인자 A의 F-값, 인자 B의 수준 수, 유의수준 5%에서의 유의성 판단이다.
주의할 조건은 오차평균제곱 $MS_E = SS_E/df_E$, 요인의 자유도 = 수준 수 - 1, 유의성 판단 기준은 $p < 0.05$라는 점이다.

NoteBase

사용할 공식

$MS_E = \frac{SS_E}{df_E} = \frac{40}{20} = 2$

F=MS_요인MS_E

$df_B = b - 1$

이 공식을 쓰는 이유는 분산분석표에서 제곱합과 자유도로 평균제곱을 구하고, 평균제곱의 비로 F-값을 판단하며, 요인의 자유도로 수준 수를 바로 복원할 수 있기 때문이다.

계산 과정

1. 오차평균제곱은 $MS_E = 40/20 = 2$이다.
2. 인자 A의 평균제곱은 $MS_A = 14/1 = 14$이므로 $F_A = 14/2 = 7$이다.
3. 인자 B의 자유도는 표의 구조상 1이므로 수준 수는 $b = 1 + 1 = 2$이다.
4. 유의수준 5%에서는 인자 A의 p-값 0.0155는 유의하고, 인자 B의 p-값 0.0987과 교호작용의 p-값 0.3293은 유의하지 않다.

선지별 해설

① O: 인자 A의 F-값은 $F_A = (14/1)/(40/20) = 14/2 = 7$이므로 맞는 설명이다.
② O: 인자 B의 자유도는 1이고, 요인의 자유도는 수준 수 - 1이므로 인자 B의 수준 수는 2이다.
③ X: 인자 B의 p-값은 0.0987로 0.05보다 크므로 유의수준 5%에서는 제품의 품질이 인자 B의 수준에 따라 영향을 받는다고 할 수 없다. 틀린 표현은 '영향을 받는다고 할 수 있다'이고, 올바르게는 '유의수준 5%에서는 영향을 받는다고 할 수 없다'이다.
④ O: 교호작용 $A \times B$의 p-값은 0.3293으로 0.05보다 크므로 유의수준 5%에서는 교호작용 효과가 있다고 할 수 없다. 따라서 이 선지는 맞는 설명이다.

오답 포인트

인자 B의 p-값 0.0987이 0.05에 가깝다는 이유로 유의하다고 착각하기 쉽다. 이 문항은 값이 애매하게 작아 보이더라도 유의수준 5% 기준에서는 0.05와 직접 비교해야 한다는 점, 그리고 발문이 '옳지 않은 것'이라는 점을 함께 묻는 함정이 있다.

다시 나오면 이렇게 판단

이원배치 분산분석표 문제는 먼저 오차평균제곱을 구한 뒤 필요한 F-값을 복원하고, 요인 자유도로 수준 수를 확인한 다음, 각 p-값을 유의수준과 바로 비교해 선지를 판단하면 된다. 마지막에는 발문이 '옳은 것'인지 '옳지 않은 것'인지 다시 확인하면 실수를 줄일 수 있다.

암기 포인트

이원배치 분산분석에서는 요인 자유도 = 수준 수 - 1이고, 유의성은 $p < 0.05$인지로 판단한다.

NoteBase

문 9. 다음은 X가 Y에 미치는 영향을 알아보기 위해 자료를 수집하여 정리한 결과이다. X가 독립변수(설명변수)이고 Y가 종속변수(반응변수)인 단순선형회귀모형을 적용하여 최소제곱법으로 얻은 추정회귀식은?

	X	Y
표본평균	6	10
표본표준편차	2	6
표본상관계수	0.5	

① $\hat{Y} = \frac{1}{6}X + 1$
② $\hat{Y} = \frac{1}{6}X + 9$
③ $\hat{Y} = \frac{3}{2}X + 1$
④ $\hat{Y} = \frac{3}{2}X + 9$

정답 3번

출제 포인트

표본평균, 표본표준편차, 표본상관계수로 단순선형회귀식 $\hat{Y} = a + bX$의 기울기와 절편을 구하는 문제이다.

조건 정리

주어진 값: $\bar{X} = 6$, $\bar{Y} = 10$, $s_X = 2$, $s_Y = 6$, $r = 0.5$이다.
구해야 할 값: 추정회귀식 $\hat{Y} = a + bX$이다.
주의할 조건: 단순선형회귀의 기울기는 $r \times s_Y/s_X$이고, 절편은 $\bar{Y} - b\bar{X}$이다.

사용할 공식

기울기 $b = r\frac{s_Y}{s_X}$, 절편 $a = \bar{Y} - b\bar{X}$이다.
이 공식을 쓰는 이유: 단순선형회귀의 최소제곱추정 직선은 상관계수와 표준편차 비로 기울기가 정해지고, 회귀직선은 항상 평균점 $(\bar{X}, \bar{Y})$을 지난다.

계산 과정

1. 기울기는 $b = 0.5 \times \frac{6}{2} = 0.5 \times 3 = 1.5 = \frac{3}{2}$이다.
2. 절편은 $a = 10 - \frac{3}{2} \times 6 = 10 - 9 = 1$이다.
3. 따라서 추정회귀식은 $\hat{Y} = \frac{3}{2}X + 1$이다.

NoteBase

선지별 해설

① X. 기울기 $\frac{1}{6}$은 $r\frac{s_X}{s_Y} = 0.5 \times \frac{2}{6}$처럼 표준편차 비를 거꾸로 쓴 값이다. 틀린 표현은 기울기 $\frac{1}{6}$이며, 올바르게는 기울기 $\frac{3}{2}$이다.

② X. 기울기 $\frac{1}{6}$도 틀리고 절편 9도 틀리다. 절편은 $\bar{Y} - b\bar{X} = 10 - 9 = 1$이어야 하므로, 올바르게는 $\hat{Y} = \frac{3}{2}X + 1$이다.

③ O. 기울기 $\frac{3}{2}$와 절편 1이 계산값과 일치한다. 회귀직선이 평균점 $(6, 10)$을 지나므로 $10 = \frac{3}{2} \times 6 + 1$도 만족한다.

④ X. 기울기 $\frac{3}{2}$는 맞지만 절편 9가 틀리다. 틀린 표현은 절편 9이며, 올바르게는 절편 1이어서 식은 $\hat{Y} = \frac{3}{2}X + 1$이다.

오답 포인트

기울기 공식에서 s_Y/s_X를 s_X/s_Y로 뒤집거나, 절편을 $\bar{Y} - b\bar{X}$ 대신 다른 값으로 두면 오답이 된다. ④는 기울기는 맞고 절편만 틀린 전형적인 함정이다.

다시 나오면 이렇게 판단

회귀식 문제는 먼저 $b = r\frac{s_Y}{s_X}$로 기울기를 구하고, 다음에 $a = \bar{Y} - b\bar{X}$로 절편을 구한 뒤 평균점 $(\bar{X}, \bar{Y})$ 대입으로 검산하면 된다.

암기 포인트

단순회귀는 '기울기 $r \times \frac{s_Y}{s_X}$, 절편은 평균점 통과'로 정리된다.

NoteBase

문 10. π_1은 처리 A에서의 성공확률이고 π_2는 처리 B에서의 성공확률이다. 처리변수와 반응변수 간의 오즈비가 2일 때, 옳은 것은? (단, 오즈비 = 처리 A의 성공오즈처리 B의 성공오즈이며, π_1과 π_2는 0보다 크고 1보다 작다)

		반응	
		성공	실패
처리	A	π_1	$1-\pi_1$
	B	π_2	$1-\pi_2$

① 처리 A의 성공오즈가 0.5이면 처리 B의 성공확률은 0.2이다.

② 처리 A의 성공확률이 0.5이면 처리 B의 성공오즈는 0.2이다.

③ 처리 A의 성공확률은 처리 B의 성공확률의 2배이다.

④ 처리 A의 성공오즈는 처리 B의 성공확률의 2배이다.

정답 1번

출제 포인트

오즈와 확률의 변환, 그리고 오즈비 =처리 A의 성공오즈처리 B의 성공오즈 관계를 이용해 각 선지를 판별하는 문제이다.

조건 정리

주어진 값: 오즈비는 2이다.
구해야 할 값: 각 선지가 오즈비 2와 일치하는지 여부이다.
주의할 조건: 성공오즈는 확률이 아니라 $\frac{p}{1-p}$이고, 확률 p와 오즈는 서로 다른 값이다.

사용할 공식

성공오즈는 $\frac{p}{1-p}$이고, 오즈비는 $\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = 2$이다.
이 공식을 쓰는 이유: 문제에서 처리 A와 B의 성공오즈 비율을 직접 주었으므로, 먼저 오즈를 비교한 뒤 필요하면 확률로 다시 바꾸어야 하기 때문이다.

NoteBase

계산 과정

1. 오즈비 2는 처리 A의 성공오즈가 처리 B의 성공오즈의 2배라는 뜻이다.
2. 따라서 처리 A의 성공오즈를 알면 처리 B의 성공오즈는 그 절반이고, 처리 B의 성공오즈를 알면 처리 A의 성공오즈는 그 2배이다.
3. 확률이 필요하면 오즈 o를 $p = \frac{o}{1+o}$로 바꾸어 계산한다.

선지별 해설

① O: 처리 A의 성공오즈가 0.5이면 처리 B의 성공오즈는 $0.5 \div 2 = 0.25$이다. 따라서 처리 B의 성공확률은 $\frac{0.25}{1+0.25} = 0.2$이므로 옳다.
② X: 처리 A의 성공확률이 0.5이면 처리 A의 성공오즈는 $\frac{0.5}{1-0.5} = 1$이다. 그러면 처리 B의 성공오즈는 $1 \div 2 = 0.5$이므로 0.2가 아니다. 틀린 표현은 "처리 B의 성공오즈는 0.2이다"이고, 올바르게는 "처리 B의 성공오즈는 0.5이다"이다.
③ X: 오즈비가 2라는 것은 성공확률의 비가 2라는 뜻이 아니다. 예를 들어 처리 B의 성공확률이 0.2이면 처리 B의 성공오즈는 0.25이고 처리 A의 성공오즈는 0.5이므로 처리 A의 성공확률은 $\frac{0.5}{1.5} = \frac{1}{3}$이다. 이는 0.4가 아니므로 틀리다. 틀린 표현은 "처리 A의 성공확률은 처리 B의 성공확률의 2배이다"이고, 올바르게는 "처리 A의 성공오즈는 처리 B의 성공오즈의 2배이다"이다.
④ X: 오즈비 2는 처리 A의 성공오즈가 처리 B의 성공오즈의 2배라는 뜻이다. 처리 B의 성공확률과는 직접 2배 관계가 성립하지 않는다. 틀린 표현은 "처리 B의 성공확률의 2배"이고, 올바르게는 "처리 B의 성공오즈의 2배"이다.

오답 포인트

오즈와 확률을 같은 값처럼 취급하면 ②, ③, ④에 모두 속기 쉽다. 이 문항의 함정은 오즈비가 확률비가 아니라는 점과, 오즈 $\frac{p}{1-p}$를 확률 p로 다시 변환해야 한다는 점이다.

다시 나오면 이렇게 판단

오즈비 문제는 먼저 "2배"가 무엇의 2배인지 확인한 뒤, 오즈끼리 비교하고, 확률을 묻는 선지에서만 $p = \frac{o}{1+o}$로 변환해 판단한다.

암기 포인트

오즈비 2는 확률 2배가 아니라 오즈 2배이다.

NoteBase

문 11. 표본의 크기가 n인 자료 (X_i, Y_i), $i = 1, 2, \cdots, n$에 단순선형회귀모형을 적용하여 최소제곱법으로 얻은 추정회귀식이 $\hat{Y}_i = 0.8 - 1.6X_i$이고 결정계수가 0.64일 때, 자료 (X_i, Y_i) 사이의 표본상관계수는?

① -0.80

② -0.64

③ 0.64

④ 0.80

정답 1번

출제 포인트

단순선형회귀에서 $R^2 = r^2$ 관계와 회귀계수의 부호로 표본상관계수의 부호를 판단하는지를 묻는 문제이다.

조건 정리

- 주어진 값: 추정회귀식의 기울기는 -1.6이고 결정계수는 0.64이다.
- 구해야 할 값: 자료 (X_i, Y_i)의 표본상관계수 r이다.
- 주의할 조건: 단순선형회귀에서는 $R^2 = r^2$이고, 기울기와 상관계수의 부호는 같다.

사용할 공식

$R^2 = r^2, \quad r = \pm\sqrt{R^2}$

이 공식을 쓰는 이유는 단순선형회귀에서 결정계수가 표본상관계수의 제곱이기 때문이다. 또한 회귀식의 기울기 b와 r의 부호는 같으므로 기울기가 음수이면 r도 음수이다.

계산 과정

1. 결정계수 0.64에서 표본상관계수의 크기를 구하면 $|r| = \sqrt{0.64} = 0.80$이다.
2. 추정회귀식 $\hat{Y}_i = 0.8 - 1.6X_i$의 기울기 -1.6이 음수이므로 상관계수도 음수이다.
3. 따라서 $r = -0.80$이다.

NoteBase

선지별 해설

① O

- 단순선형회귀에서는 $R^2 = r^2$이다. $R^2 = 0.64$이므로 $|r| = 0.80$이고, 기울기 -1.6이 음수이므로 $r = -0.80$이다.

② X

- -0.64는 결정계수 0.64에 부호만 붙인 값이다. 상관계수는 $r^2 = 0.64$를 만족해야 하므로 크기가 0.80이어야 한다. 틀린 표현은 '상관계수가 -0.64이다'이고, 올바르게는 '상관계수는 -0.80이다'이다.

③ X

- 0.64는 결정계수 자체의 값일 뿐 상관계수는 아니다. 틀린 표현은 '상관계수가 0.64이다'이고, 올바르게는 '상관계수의 크기는 0.80이다'이다.

④ X

- 0.80은 $\sqrt{0.64}$까지만 반영한 값이다. 기울기가 음수이므로 상관계수도 음수여야 한다. 틀린 표현은 '상관계수가 0.80이다'이고, 올바르게는 '상관계수는 -0.80이다'이다.

오답 포인트

결정계수와 상관계수를 같은 값으로 보는 실수가 가장 흔하다. 이 문항은 먼저 $\sqrt{0.64} = 0.80$으로 크기를 구한 뒤, 회귀기울기 부호가 음수라는 점으로 최종 부호를 정해야 한다.

다시 나오면 이렇게 판단

단순선형회귀에서 상관계수를 묻으면 먼저 $R^2 = r^2$로 크기를 구하고, 그다음 회귀기울기의 부호로 r의 부호를 결정하면 된다.

암기 포인트

단순회귀에서는 '결정계수는 상관계수의 제곱, 상관계수의 부호는 기울기와 동일'로 정리된다.

NoteBase

문 12. 다음은 어느 대학교 A 과목의 중간고사 성적을 요약한 도수분포표의 일부이다. 중앙값이 속한 계급구간은?

점수	도수	상대도수
20점 미만		0.1
20점 이상 ~ 40점 미만	6	
40점 이상 ~ 60점 미만		0.2
60점 이상 ~ 80점 미만	30	
80점 이상		0.1
합계		1.0

① 20점 이상 ~ 40점 미만

② 40점 이상 ~ 60점 미만

③ 60점 이상 ~ 80점 미만

④ 80점 이상

정답 3번

출제 포인트

상대도수와 도수를 함께 이용해 총도수와 누적도수를 복원한 뒤 중앙값이 속한 계급구간을 찾는 문제이다.

자료 읽는 순서

1. 상대도수의 합이 1이라는 점을 이용해 총도수 N을 구한다.
2. 각 계급의 도수를 모두 복원하고 누적도수를 계산한다.
3. $N/2$번째와 $N/2 + 1$번째 자료가 들어 있는 계급구간을 찾는다.

핵심 근거

표에서 도수와 상대도수를 합치면 $0.1N + 6 + 0.2N + 30 + 0.1N = N$이므로 $N = 60$이다.
따라서 각 계급의 도수는 차례로 6, 6, 12, 30, 6이고 누적도수는 6, 12, 24, 54, 60이다.
중앙값은 30번째와 31번째 자료 사이에 있으므로 이 두 값이 포함되는 계급구간은 60점 이상~80점 미만이다.

NoteBase

선지별 해설

① X
- 근거 위치: 20점 이상~40점 미만 행의 도수 6과 그 앞 계급의 상대도수 0.1에서 복원한 도수 6이다.
- 40점 미만의 누적도수는 12이므로 중앙값 위치인 30번째, 31번째 자료에 도달하지 못한다.
- 틀린 표현은 '20점 이상~40점 미만'이고, 올바르게는 '60점 이상~80점 미만'이다.

② X
- 근거 위치: 40점 이상~60점 미만 행의 상대도수 0.2이다.
- 총도수 60을 곱하면 이 계급의 도수는 12이고, 60점 미만의 누적도수는 24이다.
- 중앙값 위치는 24를 넘어서므로 이 구간이 아니다.
- 틀린 표현은 '40점 이상~60점 미만'이고, 올바르게는 '60점 이상~80점 미만'이다.

③ O
- 근거 위치: 60점 이상~80점 미만 행의 도수 30이다.
- 60점 미만 누적도수는 24이고 80점 미만 누적도수는 54이므로 30번째와 31번째 자료가 모두 이 구간에 들어간다.
- 따라서 중앙값이 속한 계급구간은 60점 이상~80점 미만이다.

④ X
- 근거 위치: 80점 이상 행의 상대도수 0.1이다.
- 이 계급의 도수는 6이므로 누적도수 55번째부터 60번째 자료에 해당한다.
- 중앙값 위치보다 훨씬 뒤에 있으므로 이 구간이 아니다.
- 틀린 표현은 '80점 이상'이고, 올바르게는 '60점 이상~80점 미만'이다.

오답 포인트

상대도수가 섞여 있을 때 바로 중앙값 계급만 감으로 고르면 틀리기 쉽다. 특히 40점 이상~60점 미만의 상대도수 0.2를 보고 중간쯤이라고 오해하기 쉽지만, 중앙값은 비율의 크기가 아니라 누적도수가 30번째와 31번째를 포함하는지로 판단한다.

다시 나오면 이렇게 판단

도수분포표에서 중앙값 계급을 묻으면 먼저 총도수를 완성한 뒤 누적도수를 적는다. 그다음 전체 자료 수가 짝수이면 $N/2$번째와 $N/2 + 1$번째 자료가 함께 들어 있는 계급구간을 고르면 된다.

암기 포인트

중앙값 계급은 가운데값의 위치를 누적도수로 끼워 넣어 찾는 값이다.

NoteBase

문 13. 다음은 확률변수 X의 확률분포이다. 확률변수 $Y = X^2$일 때, X와 Y의 공분산 $Cov(X, Y)$는?

x	0	1
$P(X=x)$	$\frac{1}{2}$	$\frac{1}{2}$

① 0
② $\frac{1}{4}$
③ $\frac{1}{2}$
④ 1

정답 2번

출제 포인트

이 문제는 $Y = X^2$를 먼저 단순화한 뒤 공분산 $\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y)$를 계산하는 문제이다.

조건 정리

주어진 값은 $P(X = 0) = \frac{1}{2}, P(X = 1) = \frac{1}{2}$이다.
구해야 할 값은 $\mathrm{Cov}(X, Y)$이다.
주의할 조건은 X가 0과 1만 가지므로 $X^2 = X$라는 점이다.

사용할 공식

$\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y)$이다.
이 공식을 쓰는 이유는 공분산을 직접 묻고 있고, 이 문항에서는 $Y = X^2$를 이용해 $E(Y)$와 $E(XY)$를 바로 구할 수 있기 때문이다.

계산 과정

1. $X \in \{0, 1\}$이므로 $Y = X^2 = X$이다.
2. 따라서 $E(X) = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}, E(Y) = \frac{1}{2}$이다.
3. 또 $XY = X \cdot X = X^2 = X$이므로 $E(XY) = \frac{1}{2}$이다.
4. 따라서 $\mathrm{Cov}(X, Y) = \frac{1}{2} - \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$이다.

NoteBase

선지별 해설

① X: 공분산이 0이 되려면 $E(XY) = E(X)E(Y)$이어야 한다. 이 문항에서는 $Y = X$가 되므로 공분산은 0이 아니다. 틀린 표현은 '공분산이 0이다'이고, 올바르게는 '공분산은 $\frac{1}{4}$이다.'이다.
② O: 계산 결과 $\mathrm{Cov}(X, Y) = \frac{1}{4}$이므로 옳다. $Y = X^2$를 바로 $Y = X$로 바꾸면 가장 빠르게 도달할 수 있다.
③ X: $\frac{1}{2}$는 $E(X)$, $E(Y)$, $E(XY)$ 값으로는 맞지만 공분산 값은 아니다. 틀린 표현은 '공분산이 $\frac{1}{2}$이다'이고, 올바르게는 '공분산은 $\frac{1}{4}$이다.'이다.
④ X: 1은 X나 Y가 가질 수 있는 값 중 하나일 뿐 공분산이 아니다. 틀린 표현은 '공분산이 1이다'이고, 올바르게는 '공분산은 $\frac{1}{4}$이다.'이다.

오답 포인트

함정은 $Y = X^2$를 새로운 변수처럼 복잡하게 보는 데 있다. 이 문제에서는 X가 0과 1만 가지므로 $X^2 = X$가 즉시 성립하고, 이를 놓치면 $E(X)$나 $E(XY)$와 공분산을 혼동하기 쉽다.

다시 나오면 이렇게 판단

확률변수가 0과 1만 가지면 먼저 거듭제곱을 단순화한다. 그다음 Y를 X로 바꿀 수 있는지 확인하고, 공분산은 $E(XY) - E(X)E(Y)$ 또는 분산으로 연결해 계산한다.

암기 포인트

0과 1만 가지는 변수는 $X^2 = X$이므로 $Y = X^2$이면 곧 $Y = X$이고, 공분산은 $\mathrm{Var}(X) = \frac{1}{4}$로 이어진다.

NoteBase

문 14. 모분산이 각각 σ_A^2과 σ_B^2이고 서로 독립인 두 정규모집단 A와 B에 대하여 다음과 같은 가설을 설정하였다. 모집단 A에서 9명, 모집단 B에서 7명을 임의로 추출하여 얻은 모집단 A와 B의 표본분산이 각각 4와 1일 때, 다음 중 이 가설을 검정하기 위한 검정통계량의 값과 유의수준 5%에서 검정결과를 옳게 짝지은 것은? (단, $F_\alpha(m,n)$은 분자의 자유도가 m, 분모의 자유도가 n인 F 분포의 $(1-\alpha)\times 100$번째 백분위수를 나타내고, $F_{0.05}(9,7)=3.68$, $F_{0.05}(8,6)=4.15$이다)

$H_0 : \frac{\sigma_A^2}{\sigma_B^2} = 1$ 대 $H_1 : \frac{\sigma_A^2}{\sigma_B^2} > 1$

① 검정통계량의 값: 2, 검정결과: H_0을 기각함

② 검정통계량의 값: 2, 검정결과: H_0을 기각하지 않음

③ 검정통계량의 값: 4, 검정결과: H_0을 기각함

④ 검정통계량의 값: 4, 검정결과: H_0을 기각하지 않음

정답 4번

출제 포인트

이 문제는 두 정규모집단의 모분산비에 대한 단측 F검정에서 검정통계량과 기각역을 정확히 구하는 문제이다.

조건 정리

주어진 값:
표본크기는 A가 9, B가 7이다.
표본분산은 A가 4, B가 1이다.
가설은 $H_0 : \sigma_A^2/\sigma_B^2 = 1, H_1 : \sigma_A^2/\sigma_B^2 > 1$이다.
주의할 조건:
F검정의 자유도는 표본크기 자체가 아니라 각각 n_A-1, n_B-1을 쓴다.

사용할 공식

검정통계량은 $F = \frac{S_A^2}{S_B^2}$이다.
귀무가설이 맞으면 $F \sim F(n_A-1, n_B-1)$를 따른다.
이 공식을 쓰는 이유:
두 독립 정규모집단의 모분산비 σ_A^2/σ_B^2를 검정하는 문제이므로 분산비에 대한 F검정을 사용한다.

NoteBase

계산 과정

1. 검정통계량을 계산하면 $F = \frac{4}{1} = 4$이다.
2. 자유도는 분자가 $9 - 1 = 8$, 분모가 $7 - 1 = 6$이므로 $F \sim F(8, 6)$를 사용한다.
3. 유의수준 5% 단측검정의 기각역은 $F > F_{0.05}(8, 6) = 4.15$이다.
4. 계산된 값 4는 4.15보다 작으므로 H_0를 기각하지 않는다.

선지별 해설

① X

검정통계량을 2로 본 것이 틀리다. 실제 검정통계량은 4/1 = 4이다. 또한 4는 4.15보다 작으므로 기각도 하지 않는다. 틀린 표현은 "검정통계량의 값 2, H_0를 기각함"이고, 올바르게는 "검정통계량의 값 4, H_0를 기각하지 않음"이다.

② X

결론은 기각하지 않는 것이 맞지만 검정통계량을 2로 본 것이 틀리다. 검정통계량은 $\frac{S_A^2}{S_B^2} = \frac{4}{1} = 4$이다. 틀린 표현은 "검정통계량의 값 2"이고, 올바르게는 "검정통계량의 값 4"이다.

③ X

검정통계량 4는 맞지만 기각 여부가 틀리다. 이 문제는 자유도 (8, 6)의 임계값 4.15와 비교해야 하므로 4는 기각역에 들어가지 않는다. 틀린 표현은 "H_0를 기각함"이고, 올바르게는 "H_0를 기각하지 않음"이다.

④ O

검정통계량을 4로 계산하고, 이를 $F_{0.05}(8, 6) = 4.15$와 비교하여 $4 < 4.15$이므로 H_0를 기각하지 않는다고 판단한 것이 맞다.

오답 포인트

표본크기 9와 7을 그대로 자유도로 넣어 $F_{0.05}(9, 7)$과 비교하거나, 표본분산비 4/1를 잘못 계산해 2로 만드는 실수가 자주 나온다. 이 문제의 함정은 자유도와 검정통계량을 동시에 흔드는 데 있다.

다시 나오면 이렇게 판단

모분산비 검정이 나오면 먼저 표본분산비로 F를 계산하고, 다음에 자유도를 반드시 $(n_A - 1, n_B - 1)$로 잡은 뒤, 마지막으로 단측인지 양측인지 확인하여 임계값과 비교하면 된다.

암기 포인트

모분산비 F검정은 "분산비 먼저, 자유도는 각각 하나씩 뺀다"로 정리된다.

NoteBase

문 15. 확률변수 X가 시행횟수는 4이고 성공확률은 p인 이항분포를 따른다. 가설 $H_0 : p = \frac{1}{3}$ 대 $H_1 : p < \frac{1}{3}$에 대한 기각역이 '$X \leq 0$'일 때, $p = \frac{1}{4}$에서의 검정력(power)은?

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

정답 2번

출제 포인트

이 문제는 기각역이 $X \leq 0$로 주어진 이항검정에서 $p = \frac{1}{4}$일 때의 검정력 $P(X \leq 0 \mid p = \frac{1}{4})$을 구하는 문제이다.

조건 정리

주어진 값:
- $X \sim Bin(4, p)$
- 귀무가설: $H_0 : p = \frac{1}{3}$
- 대립가설: $H_1 : p < \frac{1}{3}$
- 기각역: $X \leq 0$
- 검정력을 구할 지점: $p = \frac{1}{4}$

구해야 할 값:
- $p = \frac{1}{4}$에서의 검정력

주의할 조건:
- X는 성공횟수이므로 $X = 0$은 4번 모두 실패한 경우이다.
- 검정력은 기각할 확률이지 비기각할 확률이 아니다.

NoteBase

사용할 공식

검정력은 대립가설 쪽의 특정 모수값에서 기각역에 들어갈 확률이다.
$\text{Power}(p) = P(\text{reject } H_0 \mid p)$

이 문제에서는 기각역이 $X \le 0$이므로
$\text{Power}(\frac{1}{4}) = P(X \le 0 \mid p = \frac{1}{4}) = P(X = 0 \mid p = \frac{1}{4})$이다.

또한 이항분포에서 $P(X = 0) = (1 - p)^4$이다.

이 공식을 쓰는 이유:
기각역이 성공횟수 0인 경우 하나뿐이므로 $X = 0$의 확률만 계산하면 검정력이 바로 나온다.

계산 과정

1. 기각역 $X \le 0$는 정수형 확률변수에서는 $X = 0$과 같다.
2. 따라서 검정력은 $P(X = 0 \mid p = \frac{1}{4})$이다.
3. 이항분포 공식에 대입하면 $P(X = 0) = (1 - \frac{1}{4})^4 = (\frac{3}{4})^4$이다.

NoteBase

선지별 해설

① X
- $(\frac{1}{4})^4$는 p^4이므로 4번 모두 성공한 $P(X=4)$에 해당한다.
- 이 문제의 기각역은 $X=0$이므로 필요한 값은 $P(X=4)$가 아니라 $P(X=0)$이다.
- 틀린 표현은 $(\frac{1}{4})^4$이고, 올바르게는 $(1-\frac{1}{4})^4=(\frac{3}{4})^4$이다.

② O
- 검정력은 $p=\frac{1}{4}$에서 기각역 $X\le 0$에 들어갈 확률이다.
- $X\le 0$는 곧 $X=0$이므로 $P(X=0\mid p=\frac{1}{4})=(\frac{3}{4})^4$이다.
- 따라서 ②가 옳다.

③ X
- $1-(\frac{3}{4})^4$는 $P(X\ge 1)$로서 비기각역에 들어갈 확률이다.
- 검정력은 기각확률이어야 하므로 여집합을 택하면 안 된다.
- 틀린 표현은 $1-(\frac{3}{4})^4$이고, 올바르게는 $(\frac{3}{4})^4$이다.

④ X
- $1-(\frac{1}{4})^4$는 $1-P(X=4)$일 뿐이며 기각역 확률과 직접 관련이 없다.
- 이는 모두 성공의 여집합을 계산한 값으로, $X=0$의 확률과는 다른 값이다.
- 틀린 표현은 $1-(\frac{1}{4})^4$이고, 올바르게는 $(\frac{3}{4})^4$이다.

오답 포인트

기각역 $X\le 0$를 보고도 $X=0$ 대신 $X=4$를 떠올리면 ①에 끌리기 쉽다.
또 검정력을 기각하지 않을 확률의 여집합과 섞어 생각하면 ③을 고르기 쉽다.
이 문제의 핵심 함정은 성공확률 p와 실패확률 $1-p$, 그리고 기각확률과 비기각확률을 뒤바꾸는 데 있다.

다시 나오면 이렇게 판단

단측검정의 검정력을 묻는 문제는 먼저 기각역을 사건으로 바꾼다.
그다음 그 사건을 대립가설 쪽의 지정된 모수값에 대입해 확률을 계산한다.
이 문제처럼 기각역이 $X\le 0$이면 바로 $X=0$으로 줄인 뒤 $P(X=0)=(1-p)^n$을 적용하면 된다.

암기 포인트

이항분포에서 $P(X=0)=(1-p)^n$, $P(X=n)=p^n$이다.
검정력은 대립가설 쪽 모수값에서 기각역에 들어갈 확률이다.

NoteBase

문 16. 안경 렌즈의 코팅 처리 방법에 따른 마모도 차이를 알아보기 위하여, 코팅 처리 방법 A, B, C, D에 대하여 반복수가 각각 5, 4, 7, 6인 일원배치 분산분석법을 적용하여 얻은 분산분석표의 일부는 다음과 같다. ㉠과 ㉡의 값은?

요인	제곱합	자유도	F-값
처리	24		㉡
오차(잔차)		㉠	
합계	60		

① ㉠: 17, ㉡: $\frac{17}{6}$

② ㉠: 17, ㉡: 4

③ ㉠: 18, ㉡: $\frac{17}{6}$

④ ㉠: 18, ㉡: 4

정답 4번

출제 포인트

이 문제는 일원배치 분산분석표에서 자유도와 평균제곱을 이용해 오차 자유도와 F값을 구하는 문제이다.

조건 정리

주어진 값: 처리방법은 A, B, C, D의 4수준이고 반복수는 각각 5, 4, 7, 6이다. 처리 제곱합은 24이고 합계 제곱합은 60이다.
구해야 할 값: 오차 자유도 ㄱ과 처리의 F값 ㄴ이다.
주의할 조건: 일원배치 분산분석에서 오차 자유도는 전체 관측수에서 처리 수를 뺀 값이고, F값은 처리 평균제곱을 오차 평균제곱으로 나눈 값이다.

NoteBase

사용할 공식

총 관측수 $N = 5 + 4 + 7 + 6 = 22$, 처리 수 $k = 4$이다.
오차 자유도는 $N - k$이다.
오차 제곱합은 $SSE = SST - SSA$이다.
처리 평균제곱은 $MSA = SSA/(k - 1)$, 오차 평균제곱은 $MSE = SSE/(N - k)$이다.
F값은 $F = MSA/MSE$이다.
이 공식을 쓰는 이유: 분산분석표의 빈칸은 자유도, 제곱합, 평균제곱의 관계로 채우고, 최종 F값은 처리 변동과 오차 변동의 비로 계산하기 때문이다.

계산 과정

1. 총 관측수는 22이므로 전체 자유도는 $N - 1 = 21$이다.
2. 처리 자유도는 $k - 1 = 3$이므로 오차 자유도는 $21 - 3 = 18$이다. 따라서 ㄱ=18이다.
3. 오차 제곱합은 $60 - 24 = 36$이다.
4. 처리 평균제곱은 $24/3 = 8$이고 오차 평균제곱은 $36/18 = 2$이다.
5. 따라서 $F = 8/2 = 4$이므로 ㄴ=4이다.

선지별 해설

① X: ㄱ=17, ㄴ=17/6은 둘 다 틀리다. 오차 자유도는 $N - k = 22 - 4 = 18$이어야 하므로 'ㄱ=17'이 틀린 표현이다. 올바르게는 ㄱ=18이다. 또한 F값은 $8/2 = 4$이므로 'ㄴ=17/6'도 틀리다. 올바르게는 ㄴ=4이다.
② X: ㄱ=17, ㄴ=4에서 'ㄱ=17'이 틀리다. 오차 자유도는 전체 자유도에서 처리 자유도를 빼거나 $N - k$로 직접 계산하므로 18이 맞다. 올바르게는 ㄱ=18이다.
③ X: ㄱ=18은 맞지만 ㄴ=17/6이 틀리다. 17/6은 분산분석표 계산 결과와 맞지 않는다. 처리 평균제곱은 $24/3 = 8$, 오차 평균제곱은 $36/18 = 2$이므로 $F = 4$이다. 올바르게는 ㄴ=4이다.
④ O: ㄱ=18, ㄴ=4는 모두 맞다. 반복수 합으로 총 관측수 22를 구한 뒤 오차 자유도를 $22 - 4 = 18$로 계산하고, 제곱합으로부터 $F = 4$를 구하면 된다.

오답 포인트

반복수의 합을 관측수로 보지 않고 자유도를 바로 17로 착각하기 쉽다. 또한 합계 제곱합과 처리 제곱합만 보고 바로 F값을 추정하면 평균제곱 계산 단계를 빼먹어 17/6 같은 오답에 흔들릴 수 있다.

다시 나오면 이렇게 판단

분산분석표 빈칸 문제는 먼저 N, k, 각 자유도를 정리한 뒤 제곱합으로 오차 제곱합을 구하고, 마지막에 평균제곱과 F값을 계산하는 순서로 판단한다.

암기 포인트

일원배치 분산분석은 오차 자유도 $= N - k$, F=(처리 제곱합/처리 자유도)/(오차 제곱합/오차 자유도)이다.

NoteBase

문 17. 확률변수 X가 시행횟수는 8이고 성공확률은 $\frac{1}{2}$인 이항분포를 따를 때, $(X-a)^2$의 기댓값이 최소가 되도록 하는 상수 a와 이때의 최솟값을 구하면?

① a: 2, 최솟값: 0

② a: 2, 최솟값: 1

③ a: 4, 최솟값: 0

④ a: 4, 최솟값: 2

정답 4번

출제 포인트

이 문제는 $E[(X-a)^2]$를 분산과 평균으로 분해하여 최솟값을 구하는 문제이다.

조건 정리

주어진 값: $X \sim Bin(8, \frac{1}{2})$이다.
구해야 할 값: $E[(X-a)^2]$가 최소가 되게 하는 상수 a와 그 최솟값이다.
주의할 조건: 이항분포의 평균과 분산을 먼저 구하면 계산이 바로 정리된다.

사용할 공식

$E[(X-a)^2] = Var(X) + (E[X]-a)^2$
이 공식은 평균으로부터의 제곱편차를 분산과 평균 차이의 제곱으로 분해한 식이므로, 어떤 a에서 최소가 되는지 바로 판단할 수 있다.

계산 과정

1. $X \sim Bin(8, \frac{1}{2})$이므로 평균은 $E[X] = np = 8 \cdot \frac{1}{2} = 4$이다.
2. 분산은 $Var(X) = np(1-p) = 8 \cdot \frac{1}{2} \cdot \frac{1}{2} = 2$이다.
3. 따라서 $E[(X-a)^2] = 2 + (4-a)^2$이다.
4. $(4-a)^2$가 최소일 때는 $a = 4$이고, 그때 최솟값은 2이다.

NoteBase

선지별 해설

① X
- $a = 2$이면 $E[(X-a)^2] = 2 + (4-2)^2 = 6$이므로 최솟값이 0일 수 없다.
- 틀린 표현은 "$a = 2$, 최솟값 0"이다. 올바르게는 "$a = 4$, 최솟값 2"이다.

② X
- $a = 2$는 평균 4와 다르므로 제곱항 $(4-a)^2$가 0이 되지 않아 최소가 아니다.
- 틀린 표현은 "$a = 2$"이다. 올바르게는 최소가 되게 하는 값이 $a = 4$이다.

③ X
- $a = 4$는 맞지만 그때 값은 $Var(X) = 2$이므로 최솟값이 0이 아니다.
- 틀린 표현은 "최솟값 0"이다. 올바르게는 "최솟값 2"이다.

④ O
- $a = 4$일 때 $(4-a)^2 = 0$이 되어 $E[(X-a)^2] = 2$가 된다.
- 평균에 맞춰야 제곱오차 기대값이 최소가 되고, 그 최소값은 분산이 된다.

오답 포인트

제곱오차 기대값의 최소가 0이라고 착각하기 쉽다. 그러나 확률변수가 상수가 아닌 이상 $E[(X-a)^2]$의 최소값은 0이 아니라 분산이다.

다시 나오면 이렇게 판단

$E[(X-a)^2]$ 꼴이 나오면 먼저 $Var(X) + (E[X]-a)^2$로 바꾼다. 그다음 평균에 해당하는 값으로 a를 맞추고, 최소값은 분산으로 읽으면 된다.

암기 포인트

$E[(X-a)^2]$의 최적 a는 평균, 최소값은 분산이다.

NoteBase

문 18. 두 확률변수 X와 Y 사이의 상관계수가 0일 때, 옳은 것만을 모두 고르면?

ㄱ. $E(XY) = E(X)E(Y)$
ㄴ. $Var(X - Y) = Var(X) + Var(Y)$
ㄷ. X와 Y의 결합확률분포가 이변량정규분포이면 X와 Y는 서로 독립이다.

① ㄱ, ㄴ
② ㄱ, ㄷ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 4번

출제 포인트

이 문항은 상관계수 0이 뜻하는 공분산 0의 의미와, 이변량정규분포에서 무상관이 독립으로 이어지는 특별한 경우를 묻는 문제이다.

핵심 개념

상관계수 ρ_{XY}가 0이면 $\mathrm{Cov}(X, Y) = 0$이다.
공분산이 0이면 $E(XY) = E(X)E(Y)$가 되고, 분산 공식 $\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(X, Y)$에서 교차항이 사라진다.
일반 분포에서는 무상관이 곧 독립은 아니지만, 결합분포가 이변량정규분포이면 무상관과 독립이 동치가 된다.

선지별 해설

① X: ①은 ㄱ, ㄴ만 고른 보기인데 ㄷ도 옳으므로 불완전한 보기이다. 올바르게는 ㄱ, ㄴ, ㄷ이 모두 옳다.
② X: ②는 ㄱ, ㄷ만 고른 보기인데 ㄴ도 옳다. $\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(X, Y)$에서 공분산이 0이므로 ㄴ은 참이다. 올바르게는 ㄱ, ㄴ, ㄷ이 모두 옳다.
③ X: ③은 ㄴ, ㄷ만 고른 보기인데 ㄱ도 옳다. 상관계수 0이면 공분산이 0이므로 $E(XY) - E(X)E(Y) = 0$, 즉 $E(XY) = E(X)E(Y)$이다. 올바르게는 ㄱ, ㄴ, ㄷ이 모두 옳다.
④ O: ㄱ은 공분산 0의 정의에서 바로 성립하고, ㄴ은 분산 공식에서 교차항이 0이 되어 성립하며, ㄷ은 이변량정규분포의 성질상 무상관이면 독립이므로 성립한다. 따라서 ㄱ, ㄴ, ㄷ을 모두 고른 ④가 정답이다.

오답 포인트

가장 흔한 함정은 무상관과 독립을 항상 같은 뜻으로 보는 착각이다. 일반적으로는 독립이면 무상관이지만, 무상관이라고 항상 독립은 아니다. 다만 ㄷ처럼 결합분포가 이변량정규분포라는 조건이 붙으면 예외적으로 독립이 된다.

NoteBase

다시 나오면 이렇게 판단

상관계수 0이라는 조건이 보이면 먼저 공분산 0으로 바꿔서 $E(XY) = E(X)E(Y)$와 분산 공식을 확인한다. 그다음 독립 여부는 분포 조건을 따로 본다. 정규성 조건이 없으면 무상관과 독립을 동일시할 수 없고,
이변량정규분포이면 독립까지 결론낼 수 있다.

암기 포인트

무상관은 공분산 0이고, 이변량정규에서는 무상관이면 독립이다.

NoteBase

문 19. 어느 대학교에서 재학생 1인당 소유하고 있는 신용카드 개수가 평균이 0.5인 포아송분포를 따른다는 귀무가설을 카이제곱검정하고자 한다. 이를 위하여 이 대학교 재학생 1000명을 임의추출하여 각 재학생이 소유하고 있는 신용카드 개수를 조사하여 다음과 같은 표를 작성하려고 한다. 이에 대한 설명으로 옳은 것은? (단, 확률변수 X가 평균이 0.5인 포아송분포를 따를 때, $P(X \le 0) = 0.607$, $P(X \le 1) = 0.910$, $P(X \le 2) = 0.986$, $P(X \le 3) = 0.998$, $P(X \le 4) = 1.000$이다)

신용카드 개수	0	1	2	3	4 이상	합계
도수	n_0	n_1	n_2	n_3	n_4	1000

① 카이제곱검정통계량의 값이 작을수록 유의확률(p-값)은 작아진다.

② 귀무가설이 참일 때 카이제곱검정통계량은 자유도가 5인 카이제곱분포를 따른다.

③ 귀무가설이 참일 때 소유하고 있는 신용카드 개수가 3인 재학생의 기대도수는 12이다.

④ 귀무가설이 참일 때 이 대학교에 재학 중인 임의의 학생이 소유하고 있는 신용카드 개수가 1일 확률은 0.293이다.

정답 3번

출제 포인트

포아송분포의 누적확률을 개별확률과 기대도수로 바꾸고, 카이제곱 적합도검정의 자유도와 p-값의 방향을 판단하는 문제이다.

조건 정리

귀무가설은 재학생 1인당 신용카드 개수 X가 평균 0.5인 포아송분포를 따른다는 내용이다.
주어진 누적확률에서 개별확률을 구하면 $P(X = 0) = 0.607$, $P(X = 1) = 0.910 - 0.607 = 0.303$,
$P(X = 2) = 0.986 - 0.910 = 0.076$, $P(X = 3) = 0.998 - 0.986 = 0.012$,
$P(X \ge 4) = 1 - 0.998 = 0.002$이다.
표본크기가 1000명이므로 기대도수는 각각 607, 303, 76, 12, 2이다.

사용할 공식

기대도수는 $n \times p_i$로 구한다.
카이제곱 적합도검정의 자유도는 모수를 추정하지 않으면 $k - 1$이다.
카이제곱검정의 p-값은 관측된 검정통계량보다 큰 쪽의 상측확률이다.
이 공식을 쓰는 이유는 이 문항이 포아송분포를 가정한 범주별 기대도수와 카이제곱 적합도검정의 기본 성질을 함께 묻고 있기 때문이다.

NoteBase

계산 과정

1. 주어진 누적확률을 차감하여 각 범주의 개별확률을 구한다.
2. 각 개별확률에 1000을 곱해 기대도수를 구한다.
3. 구한 기대도수와 카이제곱검정의 성질로 각 선지의 O/X를 판단한다.

선지별 해설

① X: 카이제곱검정에서 검정통계량이 작을수록 귀무가설과의 차이가 작으므로 상측확률인 p-값은 커진다. 틀린 표현은 "값이 작을수록 유의확률이 작아진다"이고, 올바르게는 "값이 클수록 유의확률이 작아진다"이다.
② X: 범주가 5개이고 평균 0.5는 표본에서 추정한 값이 아니라 주어졌으므로 자유도는 $5-1=4$이다. 틀린 표현은 "자유도가 5인 카이제곱분포"이고, 올바르게는 "자유도가 4인 카이제곱분포"이다.
③ O: $P(X=3)=P(X\le 3)-P(X\le 2)=0.998-0.986=0.012$이므로 기대도수는 $1000\times 0.012=12$이다. 선지의 수치가 기대도수 계산과 일치한다.
④ X: $P(X=1)=P(X\le 1)-P(X\le 0)=0.910-0.607=0.303$이다. 틀린 표현은 "확률은 0.293이다"이고, 올바르게는 "확률은 0.303이다"이다.

오답 포인트

누적확률 $P(X\le k)$를 바로 개별확률 $P(X=k)$로 착각하면 ③과 ④를 동시에 틀리기 쉽다. 또 카이제곱검정의 p-값이 상측확률이라는 점을 놓치면 ①의 방향을 반대로 판단하기 쉽다.

다시 나오면 이렇게 판단

포아송표나 누적확률표가 주어지면 먼저 차감해서 개별확률로 바꾸고, 그다음 $n\times p_i$로 기대도수를 구한다. 카이제곱 적합도검정 선지는 자유도 $k-1$과 "검정통계량이 클수록 p-값은 작아진다"는 방향을 함께 확인하면 된다.

암기 포인트

누적확률 차감으로 개별확률을 구하고, 기대도수는 $n\times p$, 적합도검정 자유도는 기본적으로 $k-1$이다.

NoteBase

문 20. 다음은 어느 자료에 다중선형회귀모형(multiple linear regression model)
$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i$,
$i = 1, 2, \cdots, 16$**을 적용하여 최소제곱법으로 얻은 분산분석표의 일부이다.**

요인	제곱합	자유도	평균제곱
회귀			50
오차(잔차)			
합계	300		

또한, 다중선형회귀모형 $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_5 X_{5i} + e_i$, $i = 1,2,\cdots,16$을 이 자료에 적용하여 최소제곱법으로 얻은 분산분석표의 일부가 다음과 같다.

요인	제곱합	자유도	평균제곱
회귀	210		
오차(잔차)			
합계			

이때, 유의수준 5%에서 다음의 가설을 검정하려고 한다.

$H_0 : \beta_3 = \beta_4 = 0$
대
H_1 : β_3과 β_4 중 적어도 하나는 0이 아니다

이를 검정하기 위한 F−값과 검정 결과를 옳게 짝지은 것은? (단, $F_\alpha(m,n)$은 분자의 자유도가 m, 분모의 자유도가 n인 F 분포의 $(1-\alpha)\times 100$번째 백분위수를 나타내고, $F_{0.05}(1,10) = 4.96$, $F_{0.05}(1,12) = 4.75$, $F_{0.05}(2,10) = 4.10$, $F_{0.05}(2,12) = 3.89$이다. ϵ_i는 서로 독립이며 평균이 0, 분산이 σ_ϵ^2인 정규분포를 따르고, e_i는 서로 독립이며 평균이 0, 분산이 σ_e^2인 정규분포를 따른다)

① F-값: $\frac{8}{3}$, 검정결과: H_0을 기각함
② F-값: $\frac{8}{3}$, 검정결과: H_0을 기각하지 않음
③ F-값: 4, 검정결과: H_0을 기각함
④ F-값: 4, 검정결과: H_0을 기각하지 않음

정답 4번

출제 포인트

이 문제는 축소모형과 전체모형의 분산분석표를 이용해 $H_0 : \beta_3 = \beta_4 = 0$에 대한 부분 F-검정통계량을 구하고 기각 여부를 판단하는 문제이다.

조건 정리

전체모형은 설명변수 5개와 절편을 포함하므로 회귀 자유도는 5, 오차 자유도는 $16 - 6 = 10$이다.
전체모형의 회귀평균제곱이 50이므로 회귀제곱합은 $50 \times 5 = 250$이다.
총제곱합이 300이므로 전체모형의 오차제곱합은 $300 - 250 = 50$이다.
축소모형의 회귀제곱합은 210이므로 오차제곱합은 $300 - 210 = 90$이다.
검정할 제약 수는 β_3, β_4 두 개이므로 2이다.

NoteBase

사용할 공식

부분 F-검정통계량은
$F = \frac{(SSE_R - SSE_F)/q}{SSE_F/(n - p_F - 1)}$
이다.
여기서 SSE_R는 축소모형의 오차제곱합, SSE_F는 전체모형의 오차제곱합, q는 제약 수이다.
이 공식을 쓰는 이유는 β_3, β_4를 제외한 축소모형이 전체모형에 포함되는 nested model이기 때문이다.

계산 과정

1. 전체모형의 오차평균제곱을 구하면
$MSE_F = \frac{SSE_F}{10} = \frac{50}{10} = 5$
이다.
2. 부분 F-값은
$F = \frac{(90 - 50)/2}{50/10} = \frac{40/2}{5} = \frac{20}{5} = 4$
이다.
3. 임계값은 $F_{0.05}(2, 10) = 4.10$이므로 $4 < 4.10$이다.
4. 따라서 유의수준 5%에서 H_0를 기각하지 않는다.

선지 판단

1번: X. 제시된 $F = \frac{8}{3}$은 분모에 전체모형의 $MSE_F = 5$ 대신 축소모형의 $MSE_R = 90/12 = 7.5$를 잘못 넣을 때 나오는 값이다. 틀린 표현은 "F값이 $\frac{8}{3}$"과 "H_0를 기각함"이다. 올바르게는 $F = 4$이고 H_0를 기각하지 않는다.
2번: X. 검정결과 "H_0를 기각하지 않음"은 맞지만 F값이 틀렸다. 틀린 표현은 "F값이 $\frac{8}{3}$"이다. 올바르게는 전체모형의 오차평균제곱을 분모에 써서 $F = 4$가 된다.
3번: X. F값 4는 맞지만 검정결과가 틀렸다. 틀린 표현은 "H_0를 기각함"이다. 올바르게는 임계값 4.10보다 작으므로 H_0를 기각하지 않는다.
4번: O. 부분 F-값은 4이고, $F_{0.05}(2, 10) = 4.10$보다 작으므로 유의수준 5%에서 H_0를 기각하지 않는다는 판단이 맞다.

오답 포인트

축소모형과 전체모형을 비교하는 부분 F-검정에서는 분모를 항상 전체모형의 오차평균제곱으로 둔다.
축소모형의 오차제곱합 90에서 바로 90/12를 분모로 쓰면 $\frac{8}{3}$이 나와 오답으로 흐른다. 또 $F = 4$를 구해 놓고도 임계값 4.10과의 대소를 반대로 보면 3번에 속는다.

다시 나오면 이렇게 판단

부분 F-검정 문제는 전체모형과 축소모형의 SSE를 먼저 구한 뒤, 제약 수 q를 분자 자유도에 놓고 전체모형의 오차 자유도를 분모 자유도에 둔다. 마지막에는 계산한 F값과 주어진 임계값의 대소만 비교하면 된다.

NoteBase

암기 포인트

부분 F-검정은 "분자: $SSE_R - SSE_F$, 분모: 전체모형 MSE_F"로 정리된다.

NoteBase