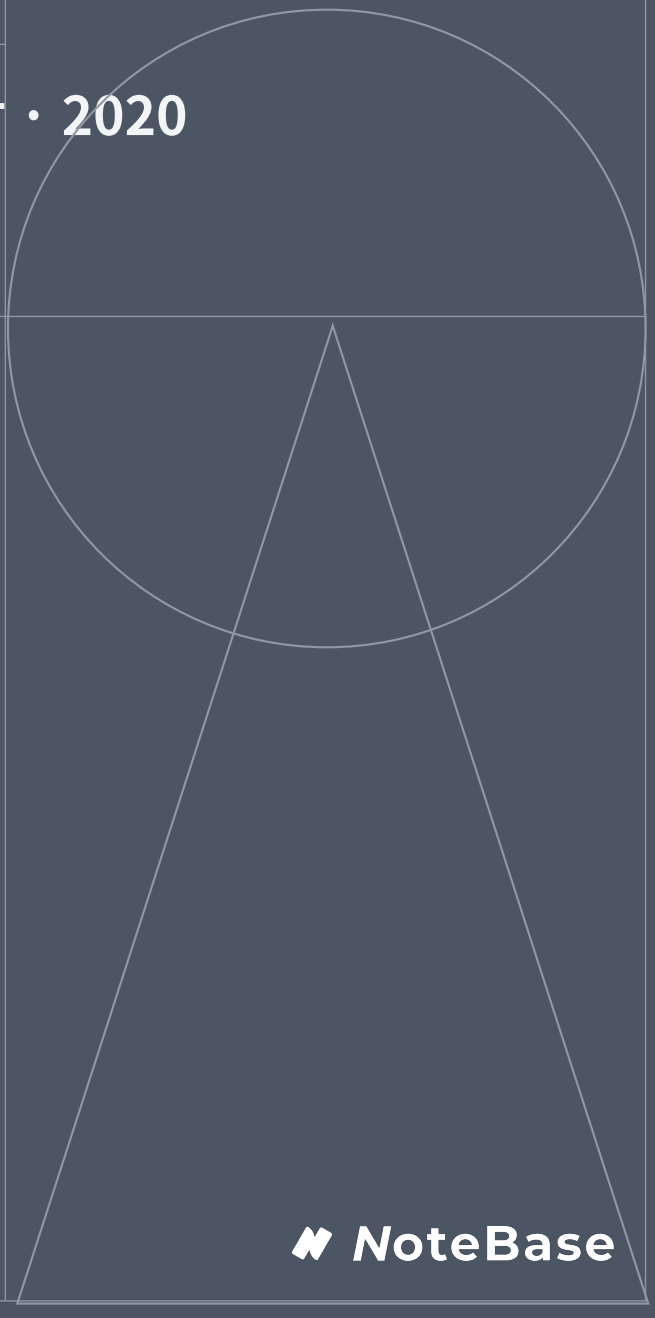
기출OX 해설집

공무원/국가직/7급 · 통계학 · 2020



NoteBase

기출OX 해설집

공무원/국가직/7급 · 통계학 · 2020

공무원/국가직/7급 · 통계학 · 2020

문 1. 다음은 A 지역에서 20일간 측정된 미세먼지의 양을 요약한 것이다. 이에 대한 설명으로 옳은 것은? (단, $100\alpha\%$ 절사평균은 관측값의 양쪽에서 각각 일정비율 α의 극단값을 버린 나머지 관측값들의 평균이다)

표본크기	평균	표준편차	중앙값	최솟값	최댓값
20	50.0	58.9	20	10	180

① 변동계수는 1보다 작다.

② 평균을 중심으로 대칭인 분포이다.

③ 범위(range)는 180이다.

④ 5% 절사평균은 45.0이다.

정답 4번

출제 포인트

이 문제는 표의 요약통계량으로 변동계수, 분포의 대칭성, 범위, 절사평균을 판단하는 문제이다.

조건 정리

주어진 값: 표본크기 20, 평균 50.0, 표준편차 58.9, 중앙값 20, 최솟값 10, 최댓값 180
구해야 할 값: 각 선지의 진위를 판단할 수 있는 변동계수, 범위, 5% 절사평균
주의할 조건: 5% 절사평균은 관측값의 양끝에서 각각 5%씩 제거한 뒤 남은 값들의 평균이다.

사용할 공식

변동계수는 $CV = \frac{s}{\bar{x}}$이다.
범위는 range=최댓값-최솟값이다.
5% 절사평균은 양끝 5%씩 제거한 뒤 남은 값들의 평균이다.
이 공식을 쓰는 이유: 선지들이 모두 표에 없는 파생값이나 분포의 형태를 묻고 있으므로 요약통계량의 정의로 직접 판단해야 하기 때문이다.

NoteBase

계산 과정

1. 변동계수는 $\frac{58.9}{50.0} = 1.178$이므로 1보다 크다.
2. 범위는 $180 - 10 = 170$이다.
3. 평균이 50.0이고 표본크기가 20이므로 전체 합은 $50.0 \times 20 = 1000$이다.
4. 5% 절사평균은 20개 자료의 양끝에서 각각 1개씩 제거하므로 10과 180을 뺀 합 $1000 - 10 - 180 = 810$을 남은 18개로 나눈 값이다.
5. 따라서 5% 절사평균은 $\frac{810}{18} = 45.0$이다.
6. 대칭 분포라면 평균과 중앙값이 비슷해야 하는데 여기서는 평균 50.0, 중앙값 20으로 차이가 크고 최댓값 쪽 꼬리가 더 길어 우측으로 치우친 분포로 볼 수 있다.

선지별 해설

① X: "변동계수는 1보다 작다"가 틀린 표현이다. 변동계수는 $\frac{58.9}{50.0} = 1.178$이므로 1보다 크다. 올바르게는 "변동계수는 1보다 크다"이다.
② X: "평균을 중심으로 대칭인 분포"가 틀린 표현이다. 대칭 분포라면 평균과 중앙값이 비슷해야 하는데 평균은 50.0, 중앙값은 20으로 차이가 크다. 또한 최솟값과 최댓값의 위치도 평균을 기준으로 대칭적이지 않다.
올바르게는 "오른쪽 꼬리가 긴 우측왜도 분포로 해석할 수 있다"이다.
③ X: "범위는 180"이 틀린 표현이다. 범위는 최댓값에서 최솟값을 뺀 값이므로 $180 - 10 = 170$이다.
올바르게는 "범위는 170이다"이다.
④ O: 5% 절사평균은 20개 자료에서 양끝 5%씩, 즉 각각 1개씩 제거한 평균이다. 전체 합 1000에서 10과 180을 빼면 810이고, 이를 18개로 나누면 45.0이므로 옳은 설명이다.

오답 포인트

변동계수는 표준편차가 평균보다 큰지만 계산하면 되는데 수치를 직접 나누지 않고 감으로 판단하기 쉽다. 또 범위를 최댓값 자체로 착각하거나, 5% 절사평균을 전체의 5%만 제거하는 평균으로 오해하기 쉽다. 대칭성도 평균만 보고 판단하지 말고 평균과 중앙값의 차이, 양끝 꼬리 길이를 함께 봐야 한다.

다시 나오면 이렇게 판단

요약통계 문항은 먼저 정의를 대입해 파생값을 계산하고, 그다음 분포 모양을 평균과 중앙값으로 점검하는 순서로 판단한다. 절사평균은 "양쪽에서 각각 몇 개를 버리는가"를 먼저 계산하고, 범위는 항상 최댓값에서 최솟값을 뺀다.

암기 포인트

변동계수는 표준편차 나누기 평균, 범위는 최댓값 빼기 최솟값, 절사평균은 양끝을 같은 비율로 버린 평균이다.

NoteBase

문 2. 측정 단위의 영향을 받지 않는 통계량만을 모두 고르면?

ㄱ. 표준편차(standard deviation)
ㄴ. 변동계수(coefficient of variation)
ㄷ. 산술평균(arithmetic mean)
ㄹ. 중앙값(median)
ㅁ. 피어슨 상관계수(Pearson correlation coefficient)
ㅂ. 공분산(covariance)

① ㄱ, ㄴ
② ㄴ, ㅁ
③ ㄷ, ㄹ
④ ㅁ, ㅂ

정답 2번

출제 포인트

측정 단위가 남는 통계량과 표준화되어 단위가 소거되는 통계량을 구분하는 문제이다.

핵심 개념

표준편차는 자료의 산포를 원자료와 같은 단위로 나타낸 값이다.
변동계수는 표준편차를 평균으로 나눈 상대적 산포도이므로 단위가 소거된다.
산술평균과 중앙값은 대표값이므로 원자료와 같은 단위를 가진다.
피어슨 상관계수는 공분산을 각 변수의 표준편차로 나누어 표준화한 값이므로 단위가 없다.
공분산은 두 변수 단위의 곱을 가지므로 측정 단위의 영향을 받는다.

선지별 해설

① X: ㄴ은 변동계수이므로 단위가 없지만, ㄱ 표준편차는 자료와 같은 단위를 가진다. 틀린 표현은 ㄱ을 포함한 부분이며, 올바르게는 ㄴ만 단위의 영향을 받지 않는다.
② O: ㄴ 변동계수와 ㅁ 피어슨 상관계수는 모두 표준화된 비율이라 단위가 소거된다. 따라서 단위의 영향을 받지 않는 통계량만 묶은 조합이 맞다.
③ X: ㄷ 산술평균과 ㄹ 중앙값은 모두 자료의 대표값이므로 원자료와 같은 단위를 가진다. 올바르게는 ㄷ과 ㄹ을 제외해야 하며, 단위의 영향을 받지 않는 것은 ㄴ과 ㅁ이다.
④ X: ㅁ은 단위가 없지만 ㅂ 공분산은 두 변수 단위의 곱을 가진다. 틀린 표현은 ㅂ을 포함한 부분이며, 올바르게는 ㅁ 대신 ㄴ과 함께 묶어야 한다.

NoteBase

오답 포인트

표준편차와 공분산을 모두 산포 관련 통계량으로만 기억하면 단위 여부를 놓치기 쉽다. 이 문항은 원단위 통계량과 표준화 통계량을 섞어 제시한 점이 함정이다.

다시 나오면 이렇게 판단

통계량이 평균, 중앙값, 표준편차처럼 원자료의 크기를 직접 나타내면 원단위가 유지된다. 같은 단위끼리 나누어 만든 비율이나 표준화 지표이면 단위가 소거된다고 판단할 수 있다.

암기 포인트

표준편차·평균·중앙값은 원단위, 변동계수·상관계수는 무단위, 공분산은 단위의 곱이다.

NoteBase

문 3. 공정한 동전을 던지는 놀이를 하고 있다. 확률변수 X를 앞면이 나올 때까지 동전을 던진 횟수라 하자. X가 1이면 1,000원, 2이면 800원, 3이면 400원, 4 이상이면 0원을 상금으로 받는다고 할 때, 참가비가 800원이라면 이 놀이에서 참가자가 얻는 이익의 기대 금액은?

① -100원

② -50원

③ 0원

④ 100원

정답 2번

출제 포인트

이 문제는 X의 분포를 이용해 상금의 기대값을 구한 뒤 참가비를 뺀 기대이익을 계산하는 문제이다.

조건 정리

- 주어진 값: $X=1$이면 상금 1,000원, $X=2$이면 800원, $X=3$이면 400원, $X\geq 4$이면 0원이다.
- 구해야 할 값: 참가비 800원을 고려한 참가자의 기대이익이다.
- 주의할 조건: 공정한 동전이므로 $P(X=1)=\frac{1}{2}$, $P(X=2)=\frac{1}{4}$, $P(X=3)=\frac{1}{8}$, $P(X\geq 4)=\frac{1}{8}$이다.

사용할 공식

기대이익 = 기대상금 - 참가비이다.
또는 직접 E(이익)=Σ (각 경우의 이익 × 그 확률)로 계산할 수 있다.
이 공식을 쓰는 이유는 각 경우의 상금과 확률이 모두 주어져 있으므로 확률가중평균으로 평균 이익을 구할 수 있기 때문이다.

계산 과정

1. 각 경우의 이익을 구하면 $X=1$일 때 200원, $X=2$일 때 0원, $X=3$일 때 -400원, $X\geq 4$일 때 -800원이다.
2. 따라서 기대이익은 $200\times\frac{1}{2}+0\times\frac{1}{4}+(-400)\times\frac{1}{8}+(-800)\times\frac{1}{8}$이다.
3. 계산하면 $100+0-50-100=-50$이므로 기대이익은 -50원이다.

NoteBase

선지별 해설

① X
- '-100원'이 틀린 표현이다. 계산 결과 기대이익은 -100원이 아니라 -50원이다.

② O
- 각 경우의 이익과 확률을 곱해 더하면 기대이익이 -50원이 되므로 옳다.

③ X
- '0원'이 틀린 표현이다. 기대상금이 참가비와 같다면 0원이지만, 실제 기대상금은 750원이므로 기대이익은 0원이 아니라 -50원이다.

④ X
- '100원'이 틀린 표현이다. 첫 번째에 앞면이 나올 때의 이익 200원만 보면 100원처럼 착각할 수 있으나, $X = 3$과 $X \geq 4$에서 손해가 발생하므로 올바르게는 -50원이다.

오답 포인트

상금의 기대값만 보고 끝내거나 참가비 800원을 빼지 않으면 오답이 된다. 또한 $X \geq 4$의 확률을 한 번에 $\frac{1}{8}$로 처리해야 하는데, 이를 빠뜨리면 기대이익이 실제보다 크게 계산된다.

다시 나오면 이렇게 판단

확률게임 문제는 먼저 경우별 순이익을 만든 뒤 그 순이익에 확률을 곱해 더하는 순서로 판단한다. 상금 기대값을 먼저 구했다면 마지막에 참가비나 비용을 반드시 반영한다.

암기 포인트

기대이익은 '상금 기대값 - 참가비'가 아니라 '경우별 순이익의 기대값'으로 바로 써도 같은 값이 나온다.

NoteBase

문 4. 확률변수 X가 이항분포 $B(16, p)$를 따른다. 분산 $Var(X)$가 최대일 때의 평균 $E(X)$는?

① 4

② 6

③ 8

④ 10

정답 3번

출제 포인트

이 문제는 이항분포의 분산 $Var(X) = np(1-p)$가 최대가 되는 조건을 이용해 평균 $E(X) = np$를 구하는 문제이다.

조건 정리

주어진 값: $X \sim B(16, p)$
구해야 할 값: 분산이 최대일 때의 평균 $E(X)$
주의할 조건: 분산의 최댓값 자체를 묻는 것이 아니라, 그때의 평균을 묻는다.

사용할 공식

이항분포에서 평균은 $E(X) = np$, 분산은 $Var(X) = np(1-p)$이다.
분산 $np(1-p)$는 $p(1-p)$가 최대일 때 최대가 되므로, $p = \frac{1}{2}$에서 최댓값을 가진다.

계산 과정

1. 분산은 $Var(X) = 16p(1-p)$이다.
2. $p(1-p)$의 최댓값은 $p = \frac{1}{2}$일 때이고, 그 값은 $\frac{1}{4}$이다.
3. 따라서 분산이 최대일 때 $p = \frac{1}{2}$이다.
4. 이때 평균은 $E(X) = np = 16 \times \frac{1}{2} = 8$이다.

NoteBase

선지별 해설

① X

4는 평균값이 아니다. 분산이 최대일 때 $p=\frac{1}{2}$이므로 평균은 $16\times\frac{1}{2}=8$이다. 4는 분산의 최댓값 $16\times\frac{1}{4}=4$를 평균으로 잘못 읽었을 때 나오는 값이다. 올바르게는 평균이 8이다.

② X

6은 $p=\frac{3}{8}$일 때의 평균값일 뿐, 분산이 최대일 때의 평균값이 아니다. 분산 최대 조건에서는 반드시 $p=\frac{1}{2}$가 되어야 하므로 올바르게는 8이다.

③ O

분산이 최대일 때 $p=\frac{1}{2}$이고, 평균은 $E(X)=16\times\frac{1}{2}=8$이다. 따라서 8이 옳다.

④ X

10은 $p=\frac{5}{8}$일 때의 평균값일 뿐이다. $p=\frac{5}{8}$이면 분산은 $16\times\frac{5}{8}\times\frac{3}{8}$로 최대가 아니다. 올바르게는 평균이 8이다.

오답 포인트

분산의 최댓값과 그때의 평균을 서로 바꾸어 읽기 쉽다. 이 문항에서는 $Var(X)$의 최대값이 4라는 사실보다, 그 최대가 되는 $p=\frac{1}{2}$를 먼저 찾고 다시 $E(X)=np$에 넣는 순서가 핵심이다.

다시 나오면 이렇게 판단

이항분포에서 분산 최대 조건이 나오면 먼저 $p(1-p)$의 최대가 $p=\frac{1}{2}$에서 나온다는 점을 확인한다. 그다음 문제에서 묻는 값이 평균인지 분산인지 구분해서 $E(X)=np$ 또는 $Var(X)=np(1-p)$에 대입해 판단한다.

암기 포인트

이항분포는 분산 최대일 때 $p=\frac{1}{2}$, 따라서 평균은 $\frac{n}{2}$이다.

NoteBase

문 5. A 대학교 재학생 중 40%가 남학생, 60%가 여학생이다. 남학생 중 기숙사 입주자의 비율을 p_1, 여학생 중 기숙사 입주자의 비율을 p_2라 할 때, $\frac{p_1}{p_2} = 3$이다. 기숙사 입주자 중 남학생의 비율을 a_1, 기숙사 입주자 중 여학생의 비율을 a_2라 할 때 $\frac{a_1}{a_2}$의 값은?

① $\frac{1}{3}$

② $\frac{1}{2}$

③ 2

④ 3

정답 3번

출제 포인트

이 문제는 전체 집단의 성별 구성비와 성별별 기숙사 입주비율을 결합하여 기숙사 입주자 내부의 성별 비율 a_1/a_2를 구하는 문제이다.

조건 정리

주어진 값은 남학생 비율 40%, 여학생 비율 60%, 그리고 $p_1/p_2 = 3$이다.
구해야 할 값은 기숙사 입주자 중 남학생 비율 a_1와 여학생 비율 a_2의 비 a_1/a_2이다.
주의할 조건은 p_1, p_2가 각각 성별 내부의 입주비율이라는 점이므로, 먼저 실제 기숙사 입주 남학생 수와 여학생 수의 비로 바꾸어야 한다는 점이다.

사용할 공식

(a_1)/(a_2)=기숙사 입주 남학생 수기숙사 입주 여학생 수=(0.4N p_1)/(0.6N p_2)=(2)/(3)·(p_1)/(p_2)
이 공식을 쓰는 이유는 a_1과 a_2가 기숙사 입주자 집단 내부의 구성비이므로, 남녀 재학생 수에 각 성별의 입주비율을 곱해 실제 입주자 수의 비를 만들어야 하기 때문이다.

계산 과정

1. 전체 재학생 수를 N이라 두면 남학생 수는 $0.4N$, 여학생 수는 $0.6N$이다.
2. 기숙사 입주 남학생 수는 $0.4Np_1$, 기숙사 입주 여학생 수는 $0.6Np_2$이다.
3. 따라서 $a_1/a_2 = (0.4Np_1)/(0.6Np_2) = (2/3) \cdot (p_1/p_2) = (2/3) \cdot 3 = 2$이다.

NoteBase

선지 판단

① X
1/3은 계산 결과와 다르다. a_1/a_2는 $(0.4/0.6) \times (p_1/p_2)$로 계산해야 하므로 올바른 값은 2이다. 틀린 표현은 1/3이며, 올바르게는 2이다.

② X
1/2는 기숙사 입주자 비율을 단순한 감으로 처리한 값일 뿐, 성별 구성비와 입주비율을 함께 반영한 값이 아니다. 올바르게는 $(0.4/0.6) \times 3 = 2$이므로 틀린 표현은 1/2이고, 올바르게는 2이다.

③ O
기숙사 입주자 중 남학생 대 여학생의 비는 $(0.4Np_1) : (0.6Np_2)$이므로 $(2/3) \times 3 = 2$가 된다. 따라서 $a_1/a_2 = 2$이므로 옳다.

④ X
3은 $p_1/p_2 = 3$만 그대로 답으로 본 경우의 오답이다. 기숙사 입주자 내부 비율을 구할 때는 남녀 재학생의 기본 비율 0.4 : 0.6도 함께 반영해야 하므로 틀린 표현은 3이고, 올바르게는 2이다.

오답 포인트

이 문제의 함정은 p_1/p_2가 이미 최종 답이라고 착각하는 데 있다. p_1, p_2는 각 성별 내부의 조건부 비율이고, a_1, a_2는 기숙사 입주자 집단 내부의 구성비이므로 서로 기준 집단이 다르다. 조건부 비율을 바로 비교하지 말고 실제 입주자 수의 비로 바꾸어야 한다.

다시 나오면 이렇게 판단

이런 문제는 전체 수를 N으로 두고 각 집단의 실제 인원수를 먼저 만든 뒤 비교하면 된다. 전체 구성비와 집단별 비율이 함께 나오면 `구성비 × 집단별 비율 = 실제 부분집단 수`의 순서로 판단하면 된다.

암기 포인트

기숙사 입주자 내부 비율은 `전체 성별 구성비 × 성별별 입주비율`을 비교해서 구한다. 조건부 비율과 최종 구성비를 그대로 같은 것으로 두지 않는다.

NoteBase

문 6. 연속형 확률변수 X의 확률밀도함수(probability density function)가 다음과 같을 때, 옳은 것만을 모두 고르면?

f(x)=casesx+(1)/(2), & 0 ≤ x ≤ a\\0, & 그 밖의 경우cases

ㄱ. a의 값은 2이다.
ㄴ. $f(2) = \frac{5}{2}$
ㄷ. $P(0 \le X \le \frac{1}{2}) = \frac{3}{8}$

① ㄷ
② ㄱ, ㄴ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 1번

출제 포인트

확률밀도함수의 정규화 조건 $\int f(x)dx = 1$로 a를 구하고, 그 값을 이용해 함수값과 구간확률을 판단하는 문제이다.

조건 정리

주어진 값: $f(x) = x + \frac{1}{2}$ for $0 \le x \le a$, 그 밖에는 0이다.
구해야 할 값: a의 값과 ㄱ, ㄴ, ㄷ의 진위이다.
주의할 조건: 확률밀도함수는 전체 적분값이 1이어야 하고, 정의구간 밖에서는 함수값이 0이다.

사용할 공식

$\int_{-\infty}^{\infty} f(x)dx = 1$
이 공식을 쓰는 이유: 주어진 식이 확률밀도함수이므로 전체 면적이 1이 되어야 하고, 여기서는 밀도가 $0 \le x \le a$에서만 정의되므로 $\int_0^a (x + \frac{1}{2})dx = 1$로 놓아 a를 결정할 수 있다.

NoteBase

계산 과정

1. 정규화 조건으로 $\int_0^a \left(x+\frac{1}{2}\right)dx = 1$ 이다.
2. 적분하면 $\left[\frac{x^2}{2}+\frac{x}{2}\right]_0^a = \frac{a^2}{2}+\frac{a}{2} = 1$ 이므로 $a^2+a-2=0$이다.
3. $(a+2)(a-1)=0$이므로 $a=-2, 1$이고, 구간 $0 \le x \le a$를 만족해야 하므로 $a=1$이다.
4. 따라서 ㄱ의 'a의 값은 2이다'는 거짓이다.
5. $2 > a = 1$이므로 $f(2)=0$이다. 따라서 ㄴ의 '$f(2)=\frac{5}{2}$'는 거짓이다.
6. $P\left(0 \le X \le \frac{1}{2}\right) = \int_0^{1/2}\left(x+\frac{1}{2}\right)dx = \left[\frac{x^2}{2}+\frac{x}{2}\right]_0^{1/2} = \frac{1}{8}+\frac{1}{4} = \frac{3}{8}$ 이므로 ㄷ은 참이다.

선지별 해설

1번 O: 1번은 ㄷ만 고른 선택지이다. 계산 결과 ㄱ은 거짓, ㄴ은 거짓, ㄷ은 참이므로 옳다.

2번 X: 2번은 ㄱ, ㄴ을 고른 선택지이다. 'ㄱ: $a=2$'가 틀렸고 'ㄴ: $f(2)=\frac{5}{2}$'도 틀렸다. 올바르게는 '$a=1$, $f(2)=0$'이다.

3번 X: 3번은 ㄴ, ㄷ을 고른 선택지이다. ㄷ은 참이지만 ㄴ은 거짓이므로 틀리다. 올바르게는 '$f(2)=0$이고 ㄷ만 참'이다.

4번 X: 4번은 ㄱ, ㄴ, ㄷ을 모두 고른 선택지이다. ㄱ과 ㄴ이 거짓이므로 틀리다. 올바르게는 '$a=1$, $f(2)=0$, $P(0 \le X \le \frac{1}{2}) = \frac{3}{8}$이므로 ㄷ만 참'이다.

오답 포인트

정규화 조건으로 a를 먼저 확정하지 않고 식 $x+\frac{1}{2}$에 바로 $x=2$를 대입하면 ㄴ을 잘못 판단하기 쉽다. 이 문제의 핵심 함정은 '정의구간 밖에서는 밀도가 0'이라는 점과 '전체 적분값이 1'이라는 조건을 동시에 놓치는 데 있다.

다시 나오면 이렇게 판단

밀도함수 선지형은 먼저 전체 적분이 1이 되도록 미정수를 구하고, 다음에 정의구간 안팎을 확인해 함수값을 판단하며, 마지막에 필요한 확률만 적분해 참인 진술 조합을 고르면 된다.

암기 포인트

확률밀도함수는 '전체 적분 1, 구간 밖 0'을 먼저 확인하면 선지 판단이 바로 정리된다.

NoteBase

문 7. 지난달 A 드라마 시청률이 50%였다. 이번 달의 A 드라마 시청률이 지난달과 다른지를 이항검정법으로 검정하기 위하여 10세대를 랜덤 추출하여 이번 달 A 드라마를 시청한 세대수(X)를 조사하고자 한다. 유의수준 10%에서 귀무가설(H_0)의 기각역은? (단, p는 이번 달의 A 드라마 시청률이다)

$H_0 : p = \frac{1}{2},\ H_1 : p \neq \frac{1}{2}$

① $X = 0$ 또는 $X = 10$

② $X \leq 1$ 또는 $X \geq 9$

③ $X \leq 2$ 또는 $X \geq 8$

④ $X \leq 3$ 또는 $X \geq 7$

정답 2번

출제 포인트

이 문제는 $X \sim Bin(10, 0.5)$인 이항분포에서 양측검정의 기각역을 유의수준 10%에 맞게 정하는 문제이다.

조건 정리

주어진 값: $n = 10, H_0 : p = \frac{1}{2}, H_1 : p \neq \frac{1}{2}$
구해야 할 값: 유의수준 10%에서의 기각역
주의할 조건: 대립가설이 $p \neq \frac{1}{2}$인 양측검정이므로 좌우 꼬리를 함께 본다.

사용할 공식

귀무가설 하에서 $X \sim Bin(10, 0.5)$이므로
$P(X = k) = \binom{10}{k}\left(\frac{1}{2}\right)^{10}$
이다.

이 공식을 쓰는 이유:
표본 10세대에서 이번 달 시청 세대 수 X는 성공확률 0.5인 이항분포를 따르며, 기각역은 귀무가설 하에서 그 구간의 확률이 0.1 이하가 되도록 정해야 하기 때문이다.

계산 과정

1. 양측검정이므로 $X \leq c$ 또는 $X \geq 10 - c$ 꼴의 대칭 기각역을 본다.
2. P(X ≤ 1 또는 X ≥ 9)=2×≤ft\100+1012^10\=2×(1+10)/(1024)=(22)/(1024)≈0.0215이다.
3. 한 단계 넓힌 P(X ≤ 2 또는 X ≥ 8)=2×≤ft\100+101+1022^10\=2×(1+10+45)/(1024)=(112)/(1024)≈0.1094로 0.1을 넘는다.
4. 따라서 유의수준 10%를 넘지 않는 가장 넓은 대칭 기각역은 $X \leq 1$ 또는 $X \geq 9$이다.

NoteBase

선지별 해설

① X
P(X=0 또는 10)=2×≤ft((1)/(2))^10=(2)/(1024)≈0.0020이므로 유의수준 10%에 비해 지나치게 좁은 기각역이다. "$X=0$ 또는 $X=10$이 기각역이다"가 틀렸고, 올바르게는 "$X \le 1$ 또는 $X \ge 9$가 기각역이다"이다.

② O
P(X ≤ 1 또는 X ≥ 9)=(22)/(1024)≈0.0215≤0.1이다. 반면 $X \le 2$ 또는 $X \ge 8$까지 넓히면 유의수준을 초과하므로, 이 선택지가 유의수준 10%를 만족하는 기각역이다.

③ X
P(X ≤ 2 또는 X ≥ 8)=(112)/(1024)≈0.1094이므로 유의수준 10%를 초과한다. "$X \le 2$ 또는 $X \ge 8$이 기각역이다"가 틀렸고, 올바르게는 "$X \le 1$ 또는 $X \ge 9$가 기각역이다"이다.

④ X
P(X ≤ 3 또는 X ≥ 7)=2×(1+10+45+120)/(1024)=(352)/(1024)≈0.3438이므로 유의수준 10%를 크게 초과한다. "$X \le 3$ 또는 $X \ge 7$이 기각역이다"가 틀렸고, 올바르게는 "$X \le 1$ 또는 $X \ge 9$가 기각역이다"이다.

오답 포인트

양측검정이라는 이유만 보고 좌우 대칭만 맞추면 된다고 생각하면 ③처럼 유의수준을 초과한 구간을 고르게 된다. 이 문항의 함정은 "가장 넓은 대칭 구간"이 아니라 "귀무가설 하 확률이 0.1 이하인 구간"이라는 점이다.

다시 나오면 이렇게 판단

이항 양측검정의 기각역 문제는 귀무가설 하 분포를 먼저 정하고, 좌우 꼬리의 누적확률을 계산하여 유의수준을 넘지 않는 가장 바깥 구간까지 채우는 방식으로 판단한다.

암기 포인트

양측 이항검정의 기각역은 대칭성만으로 정하지 않고, 귀무가설 하 꼬리확률의 합이 유의수준 이하인지로 정한다.

NoteBase

문 8. 표준편차가 10인 정규모집단에서 모평균(μ)에 대한 귀무가설 $H_0 : \mu = 10$ 대 대립가설 $H_1 : \mu = 12$를 검정하고자 한다. 크기 100의 표본을 랜덤 추출하여 표본평균 $\bar{X}$를 구한 후 기각역으로 $R : \bar{X} \geq 11.5$를 사용할 때, 제2종 오류의 확률은? (단, Z는 표준정규분포를 따르는 확률변수를 나타낸다)

① $P(Z < -0.05)$

② $P(Z < -0.5)$

③ $P(Z > 0.15)$

④ $P(Z > 1.5)$

정답 2번

출제 포인트

이 문제는 기각역이 주어졌을 때 대립가설이 참인 상황에서 제2종 오류 확률 β를 표준화하여 구하는 문제이다.

조건 정리

주어진 값: 모표준편차는 10, 표본크기는 100, 귀무가설은 $H_0 : \mu = 10$, 대립가설은 $H_1 : \mu = 12$이다.
구해야 할 값: 제2종 오류 확률 $\beta = P(\bar{X} < 11.5 \mid \mu = 12)$이다.
주의할 조건: 제2종 오류는 대립가설이 참인데도 귀무가설을 기각하지 못하는 확률이므로 $\mu = 12$를 기준으로 계산한다.

사용할 공식

$\beta = P(\bar{X} \notin R \mid \mu = 12) = P(\bar{X} < 11.5 \mid \mu = 12)$
$Z = \frac{\bar{X} - 12}{10/\sqrt{100}} \sim N(0, 1)$
이 공식을 쓰는 이유는 대립가설 $\mu = 12$가 참일 때 표본평균의 분포를 표준정규분포로 바꾸면 제2종 오류 확률을 바로 읽을 수 있기 때문이다.

계산 과정

1. 표본평균의 표준오차는 $10/\sqrt{100} = 1$이다.
2. 기각역이 $R : \bar{X} \geq 11.5$이므로 기각하지 않는 영역은 $\bar{X} < 11.5$이다.
3. 따라서 제2종 오류 확률은 $\beta = P(\bar{X} < 11.5 \mid \mu = 12)$이다.
4. 이를 표준화하면 $\beta = P\left(\frac{\bar{X} - 12}{1} < \frac{11.5 - 12}{1}\right) = P(Z < -0.5)$ 이다.

NoteBase

선지별 해설

1번: X. $P(Z < -0.05)$는 기준값 11.5와 평균 12의 차이인 0.5를 표준오차 1이 아니라 10으로 잘못 나눈 값이다. 틀린 표현은 -0.05이고, 올바르게는 -0.5이므로 $P(Z < -0.5)$가 된다.
2번: O. 제2종 오류는 $\mu = 12$일 때 $\bar{X} < 11.5$가 나올 확률이므로 표준화 결과는 정확히 $P(Z < -0.5)$이다.
3번: X. 부등호 방향과 부호를 뒤집은 선택지이다. 틀린 표현은 $P(Z > 0.15)$이고, 올바르게는 왼쪽 꼬리확률 $P(Z < -0.5)$이다.
4번: X. 1.5는 귀무가설 평균 10을 기준으로 $(11.5 - 10)/1$처럼 계산한 값이라서 제2종 오류 정의와 맞지 않는다. 제2종 오류는 대립가설 평균 12를 기준으로 보아야 하므로 올바르게는 $P(Z < -0.5)$이다.

오답 포인트

제2종 오류를 구하면서도 무의식적으로 귀무가설 평균 10을 대입하거나, 기각역의 반대영역을 잡아 놓고도 부등호 방향을 바꾸는 실수가 자주 나온다. 이 문항은 '대립가설이 참인데 기각하지 못함'이라는 정의를 끝까지 유지하는지가 함정이다.

다시 나오면 이렇게 판단

가설검정 오류 문항은 먼저 어떤 오류인지 정한 다음, 그 오류가 발생하는 상황의 참인 평균을 잡고, 마지막에 기각역의 반대영역인지 여부를 확인하여 표준화한다.

암기 포인트

제1종 오류는 H_0 기준, 제2종 오류는 H_1 기준이라는 구분이 핵심이다.

NoteBase

문 9. T는 자유도가 k인 t분포를 따르는 확률변수를 나타내고, $t_\alpha(k)$는 자유도가 k인 t분포의 제100$(1-\alpha)$백분위수를 나타낼 때, 옳은 것만을 모두 고르면? (단, $k>2$, $0<\alpha<1$이다)

ㄱ. T의 평균은 0이다.
ㄴ. α가 고정되어 있을 때, k가 증가할수록 $t_\alpha(k)$는 감소한다.
ㄷ. k가 고정되어 있을 때, α가 증가할수록 $t_\alpha(k)$는 감소한다.

① ㄴ
② ㄱ, ㄴ
③ ㄱ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 3번

출제 포인트

이 문제는 자유도 k인 t분포의 평균 존재 조건과 $t_\alpha(k)=F_k^{-1}(1-\alpha)$의 단조성을 함께 묻는 문제이다.

핵심 개념

t분포는 0을 중심으로 대칭이고, 자유도 $k>1$이면 평균이 0이다.
$t_\alpha(k)$는 자유도 k인 t분포의 제100$(1-\alpha)$백분위수이므로 $t_\alpha(k)=F_k^{-1}(1-\alpha)$이다.
k가 커질수록 t분포는 표준정규분포에 가까워져 양의 상측 분위수는 작아지지만, 음의 하측 분위수는 0 쪽으로 커질 수 있다.

선지별 해설

① X: 틀린 부분은 'ㄴ만 옳다'이다. ㄱ은 $k>2$이면 당연히 $k>1$도 만족하므로 평균이 0이어서 옳다. ㄴ은 항상 옳지 않다.
$\alpha<0.5$이면 $t_\alpha(k)$가 양수여서 k가 커질수록 감소하지만, $\alpha>0.5$이면 $t_\alpha(k)$가 음수여서 k가 커질수록 0에 가까워지며 증가한다. 실제로는 'ㄱ, ㄷ이 옳다'가 맞다.
② X: 틀린 부분은 'ㄴ이 옳다'이다. ㄱ은 옳지만 ㄴ은 위와 같은 이유로 $0<\alpha<1$ 전체에서 성립하지 않는다. 실제로는 'ㄱ, ㄷ이 옳다'가 맞다.
③ O: ㄱ은 t분포의 평균 존재 조건이 $k>1$이라는 점에서 옳다. ㄷ은 $t_\alpha(k)=F_k^{-1}(1-\alpha)$에서 α가 증가하면 $1-\alpha$가 감소하므로 분위수도 감소하여 옳다.
④ X: 틀린 부분은 'ㄴ이 옳다'이다. ㄱ과 ㄷ은 옳지만 ㄴ은 α의 범위를 $0<\alpha<1$ 전체로 두면 항상 감소한다고 할 수 없다. 실제로는 'ㄱ, ㄷ'만 고른 ③이 맞다.

NoteBase

오답 포인트

보통 검정표에서 쓰는 $\alpha = 0.05, 0.01$만 떠올리면 ㄴ이 맞아 보인다. 그러나 문제는 $0 < \alpha < 1$ 전체를 가정하므로, $\alpha > 0.5$에서 $t_\alpha(k)$가 음수가 된다는 점을 놓치면 ㄴ을 잘못 옳다고 판단하게 된다.

다시 나오면 이렇게 판단

분위수 기호가 나오면 먼저 $t_\alpha(k) = F_k^{-1}(1-\alpha)$로 바꿔 본다. 그다음 α 변화는 확률 $1-\alpha$의 증감으로, k 변화는 분포가 0 중심으로 얼마나 퍼져 있는지로 나누어 판단하면 된다. 특히 α 범위가 $0 < \alpha < 1$ 전체인지, 통상적인 작은 유의수준만 상정하는지부터 확인해야 한다.

암기 포인트

t분포는 $k > 1$이면 평균 0, $t_\alpha(k) = F_k^{-1}(1-\alpha)$이므로 $\alpha \uparrow \Rightarrow t_\alpha(k) \downarrow$이다. 다만 $k \uparrow$일 때는 분위수의 부호에 따라 증감 방향이 달라질 수 있다.

NoteBase

문 10. 단순선형회귀모형 $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, 2, \cdots, 10$을 최소제곱법으로 적합시킨 회귀식이 $\hat{y}_i = 1 + \frac{1}{3}x_i$이고 $\sum_{i=1}^{10} \hat{y}_i = 70$일 때, $E[Y \mid x]$의 95% 신뢰구간의 길이가 최소가 되는 x의 값은? (단, ε_i는 $N(0, \sigma^2)$를 따르며 서로 독립이다)

① 12

② 18

③ 21

④ 24

정답 2번

출제 포인트

이 문제는 단순선형회귀에서 평균반응 $E[Y|x]$의 신뢰구간 길이가 최소가 되는 점이 $x = \bar{x}$임을 이용해 $\bar{x}$를 구하는 문제이다.

조건 정리

주어진 값은 표본크기 10, 적합회귀식 $\hat{y} = 1 + \frac{1}{3}x$, 그리고 $\sum_{i=1}^{10} \hat{y}_i = 70$이다.
구해야 할 값은 $E[Y|x]$의 95% 신뢰구간 길이가 최소가 되는 x이다.
최소제곱법으로 구한 회귀직선은 항상 점 $(\bar{x}, \bar{y})$를 지난다는 성질을 사용한다.

사용할 공식

평균반응 $E[Y|x]$의 95% 신뢰구간은
$\hat{y}(x) \pm t_{0.975,\,8} s \sqrt{\frac{1}{10} + \frac{(x - \bar{x})^2}{S_{xx}}}$
의 꼴이다.
신뢰구간의 길이는 위 식의 제곱근 부분에 비례하므로 $(x - \bar{x})^2$가 최소인 $x = \bar{x}$에서 가장 짧아진다.
또한 최소제곱 회귀직선은 $\hat{y}(\bar{x}) = \bar{y}$를 만족한다.

계산 과정

1. 최소제곱법에서는 잔차의 합이 0이므로 $\sum \hat{y}_i = \sum y_i$이고, 따라서 $\bar{y} = \frac{70}{10} = 7$이다.
2. 회귀직선은 $(\bar{x}, \bar{y})$를 지나므로 $7 = 1 + \frac{1}{3}\bar{x}$이다.
3. 따라서 $\frac{1}{3}\bar{x} = 6$, $\bar{x} = 18$이다.
4. 평균반응의 신뢰구간 길이는 $x = \bar{x}$에서 최소이므로 구하는 값은 18이다.

NoteBase

선지 판단

1번: X. 신뢰구간 길이가 최소가 되는 점은 $x = \bar{x}$인데 계산된 $\bar{x}$는 18이다. 틀린 표현은 "최소가 되는 x는 12이다"이고, 올바르게는 "최소가 되는 x는 18이다"이다.
2번: O. $\bar{y} = 7$, $7 = 1 + \frac{1}{3}\bar{x}$에서 $\bar{x} = 18$이므로 평균반응의 신뢰구간 길이가 최소가 되는 x는 18이다.
3번: X. 21은 회귀직선이 지나는 중심점 $\bar{x}$ 계산 결과와 일치하지 않는다. 틀린 표현은 "최소가 되는 x는 21이다"이고, 올바르게는 "최소가 되는 x는 18이다"이다.
4번: X. 24 역시 $\hat{y}(\bar{x}) = \bar{y}$를 이용해 얻는 $\bar{x} = 18$과 다르다. 틀린 표현은 "최소가 되는 x는 24이다"이고, 올바르게는 "최소가 되는 x는 18이다"이다.

오답 포인트

신뢰구간 길이의 최소점을 직접 미분하려고 하다가 핵심 성질을 놓치기 쉽다. 이 문항의 함정은 회귀직선이 $(\bar{x}, \bar{y})$를 지난다는 사실과 평균반응 신뢰구간의 분산항이 $x = \bar{x}$에서 최소라는 사실을 연결하지 못하게 만드는 점이다.

다시 나오면 이렇게 판단

회귀에서 평균반응의 신뢰구간이나 예측값의 분산이 가장 작은 x를 묻는 문항은 먼저 $x = \bar{x}$인지 확인한다.
그다음 회귀직선이 $(\bar{x}, \bar{y})$를 지난다는 성질로 $\bar{x}$를 계산하면 된다.

암기 포인트

단순회귀에서 평균반응 신뢰구간의 중심 최소점은 항상 $x = \bar{x}$이고, 최소제곱 회귀직선은 항상 $(\bar{x}, \bar{y})$를 지난다.

NoteBase

문 11. 어느 병원에 내원한 호흡기 질환자 중 200명을 랜덤 추출하여 계절별로 정리하면 다음과 같다. '호흡기 질환자의 계절별 비율은 동일하다'는 귀무가설을 검정하기 위해 적합한 검정법과 유의수준 5%에서의 검정 결과를 바르게 연결한 것은? (단, 자유도가 k_1, k_2인 F분포의 제100$(1-\alpha)$백분위수를 $F_\alpha(k_1, k_2)$로 나타낼 때 $F_{0.05}(1,3) = 10.13$, $F_{0.05}(1,4) = 7.71$이며, 자유도가 k인 χ^2분포의 제100$(1-\alpha)$백분위수를 $\chi^2_\alpha(k)$로 나타낼 때 $\chi^2_{0.05}(3) = 7.81$, $\chi^2_{0.05}(4) = 9.49$이다)

계절	봄	여름	가을	겨울
환자 수	40	65	40	55

① 검정법: F, 검정 결과: 귀무가설을 기각하지 않음

② 검정법: F, 검정 결과: 귀무가설을 기각함

③ 검정법: χ^2, 검정 결과: 귀무가설을 기각하지 않음

④ 검정법: χ^2, 검정 결과: 귀무가설을 기각함

정답 4번

출제 포인트

이 문제는 계절별 비율이 동일한지에 대한 카이제곱 적합도 검정에서 검정법 선택, 자유도 계산, 기각 여부 판단을 묻는 문제이다.

조건 정리

주어진 값: 봄 40명, 여름 65명, 가을 40명, 겨울 55명, 전체 200명이다.
구해야 할 값: 계절별 비율이 동일하다는 귀무가설에 대한 적절한 검정법과 기각 여부이다.
주의할 조건: 범주가 4개이므로 귀무가설 아래 기대도수는 각 50명이고 자유도는 $4-1=3$이다.

사용할 공식

검정법은 일원 카이제곱 적합도 검정이다.
사용할 통계량은 $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$ 이다.
이 공식을 쓰는 이유는 한 표본을 여러 범주로 나눈 뒤 관측도수와 귀무가설이 주는 기대도수를 비교하는 상황이기 때문이다.

NoteBase

계산 과정

귀무가설은 호흡기 질환자의 계절별 비율이 모두 1/4로 같다는 것이다.
기대도수는 각 계절마다 $200 \times 1/4 = 50$명이다.
검정통계량은 $\chi^2 = \frac{(40-50)^2}{50} + \frac{(65-50)^2}{50} + \frac{(40-50)^2}{50} + \frac{(55-50)^2}{50} = 2 + 4.5 + 2 + 0.5 = 9$ 이다.
자유도는 3이고 임계값은 $\chi^2_{0.05}(3) = 7.81$이므로 $9 > 7.81$에서 유의수준 5%에서 귀무가설을 기각한다.

선지별 해설

① X. F검정은 두 분산의 비 또는 분산분석에서 평균 차이를 다루는 검정이므로 범주별 비율 동일성 검정에 맞지 않는다. 틀린 표현은 '검정법이 F'인 부분이며, 올바르게는 '검정법은 χ^2 적합도 검정'이다.
② X. 귀무가설 기각 여부만 보면 맞지만 검정법이 틀렸다. 틀린 표현은 '검정법이 F'인 부분이며, 올바르게는 '검정법은 χ^2 적합도 검정이고 귀무가설은 기각한다'이다.
③ X. 검정법은 χ^2로 맞지만 계산값이 임계값보다 크므로 결론이 틀렸다. 틀린 표현은 '귀무가설을 기각하지 않음'이며, 올바르게는 '귀무가설을 기각함'이다.
④ O. 계절별 비율 동일성은 χ^2 적합도 검정으로 본다. 계산된 검정통계량이 9이고 임계값이 7.81이므로 유의수준 5%에서 귀무가설을 기각하는 판단이 맞다.

오답 포인트

범주가 4개라고 해서 자유도를 4로 두면 $\chi^2_{0.05}(4) = 9.49$와 비교하게 되어 기각하지 않는다는 오답으로 흐르기 쉽다. 이 문제의 자유도는 '범주 수 - 1'이라서 3이다. 또 여러 계절을 비교한다고 해서 바로 F검정을 떠올리면 평균 비교와 비율 비교를 혼동하게 된다.

다시 나오면 이렇게 판단

한 표본을 여러 범주로 나누어 '비율이 같다'를 묻는 문항이면 먼저 χ^2 적합도 검정을 떠올린다. 그다음 기대도수 계산, 자유도 $k-1$ 확인, 검정통계량 계산, 임계값 비교 순서로 판단하면 된다.

암기 포인트

한 표본의 범주별 비율 동일성은 χ^2 적합도 검정, 자유도는 '범주 수 - 1'이다.

NoteBase

문 12. 단순선형회귀모형 $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, 2, \cdots, n$을 최소제곱법으로 적합시켰을 때 결정계수와 같은 값을 갖는 것만을 모두 고르면? (단, ε_i는 $N(0, \sigma^2)$을 따르고 서로 독립이며, $\hat{\beta}_1$는 β_1에 대한 추정량, SSE는 잔차제곱합, $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$, $S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$, $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$이다)

ㄱ. $1 - \frac{SSE}{S_{yy}}$

ㄴ. $\frac{\hat{\beta}_1^2 S_{xx}}{S_{yy}}$

ㄷ. $\frac{(S_{xy})^2}{S_{xx}S_{yy}}$

① ㄱ

② ㄱ, ㄴ

③ ㄴ, ㄷ

④ ㄱ, ㄴ, ㄷ

정답 4번

출제 포인트

단순선형회귀에서 결정계수 R^2를 $1 - \frac{SSE}{S_{yy}}, \frac{\hat{\beta}_1^2 S_{xx}}{S_{yy}}, \frac{S_{xy}^2}{S_{xx}S_{yy}}$로 서로 연결할 수 있는지를 묻는 문제이다.

조건 정리

주어진 것은 단순선형회귀모형과 $SSE, S_{xx}, S_{yy}, S_{xy}, \hat{\beta}_1$의 정의이다.
구해야 할 것은 각 진술 ㄱ, ㄴ, ㄷ이 결정계수와 같은 값인지 여부이다.
주의할 조건은 단순선형회귀에서 $SST = S_{yy}$, $R^2 = 1 - \frac{SSE}{S_{yy}} = \frac{SSR}{S_{yy}}$, $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$를 함께 써야 한다는 점이다.

사용할 공식

결정계수는 $R^2 = 1 - \frac{SSE}{S_{yy}} = \frac{SSR}{S_{yy}}$이다.
또한 단순선형회귀에서 $SSR = \hat{\beta}_1 S_{xy} = \hat{\beta}_1^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}}$, $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$이다.
이 공식을 쓰는 이유는 보기 ㄱ, ㄴ, ㄷ이 모두 R^2의 서로 다른 표현인지 확인하는 문제이기 때문이다.

계산 과정

1. ㄱ은 $R^2 = 1 - \frac{SSE}{S_{yy}}$와 정확히 같다.
2. $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$이므로 $\hat{\beta}_1^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}}$이다.
3. 따라서 ㄴ의 $\frac{\hat{\beta}_1^2 S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}}$이고, 이는 ㄷ과 같다.
4. 한편 $SSR = \frac{S_{xy}^2}{S_{xx}}$이므로 $\frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{SSR}{S_{yy}} = R^2$이다.
5. 결국 ㄱ, ㄴ, ㄷ이 모두 결정계수와 같은 값이다.

NoteBase

선지별 해설

① X: ①은 ㄱ만 골랐으므로 틀리다. ㄴ과 ㄷ도 각각 R^2와 같으므로 올바르게는 'ㄱ, ㄴ, ㄷ'을 모두 골라야 한다.
② X: ②는 ㄱ, ㄴ만 골랐으므로 틀리다. ㄷ의 $\frac{S_{xy}^2}{S_{xx}S_{yy}}$도 R^2이므로 올바르게는 'ㄱ, ㄴ, ㄷ'이다.
③ X: ③은 ㄴ, ㄷ만 골랐으므로 틀리다. ㄱ의 $1-\frac{SSE}{S_{yy}}$도 결정계수의 기본식이므로 올바르게는 'ㄱ, ㄴ, ㄷ'이다.
④ O: ④는 ㄱ, ㄴ, ㄷ을 모두 골랐고, 세 식이 모두 R^2의 동치표현이므로 옳다.

오답 포인트

ㄴ과 ㄷ만 보고 서로 같은 식이라는 점에서 멈추고, ㄱ의 $1-\frac{SSE}{S_{yy}}$를 결정계수의 기본식으로 연결하지 못하면 ③으로 흔들리기 쉽다. 반대로 ㄱ만 익숙하고 $\hat{\beta}_1=\frac{S_{xy}}{S_{xx}}$, $SSR=\frac{S_{xy}^2}{S_{xx}}$ 연결을 놓치면 ①이나 ②로 오답을 고르기 쉽다.

다시 나오면 이렇게 판단

회귀에서 결정계수를 묻는 문제는 먼저 $R^2=1-\frac{SSE}{S_{yy}}=\frac{SSR}{S_{yy}}$를 세운 뒤, $\hat{\beta}_1=\frac{S_{xy}}{S_{xx}}$로 보기를 SSR 형태로 바꿔 본다. 보기가 서로 다른 기호를 쓰더라도 결국 R^2로 환원되면 모두 옳은 진술이다.

암기 포인트

단순회귀에서는 $R^2=1-\frac{SSE}{S_{yy}}=\frac{SSR}{S_{yy}}=\frac{S_{xy}^2}{S_{xx}S_{yy}}$이고, $\hat{\beta}_1=\frac{S_{xy}}{S_{xx}}$를 끼워 넣으면 ㄴ도 바로 R^2가 된다.

NoteBase

문 13. 다음은 서로 다른 세 가지 비료에 따른 평균수확량의 차이가 있는지 알아보기 위한 분산분석표이다. 이에 대한 설명으로 옳은 것만을 모두 고르면? (단, 자유도가 k_1, k_2인 F분포의 제100$(1-\alpha)$백분위수를 $F_\alpha(k_1, k_2)$로 나타낼 때, $F_{0.05}(2, 27) = 3.35$, $F_{0.05}(3, 26) = 2.98$이다)

요인	제곱합	자유도	평균제곱	F-값
처리	17.956	( )	8.978	10.125
오차	23.941	( )	0.887	
전체	41.897	29		

ㄱ. 유의수준 5%에서 세 가지 비료에 따른 평균수확량은 모두 다르다.

ㄴ. $F_{m,n}$이 자유도가 m, n인 F분포를 따르는 확률변수일 때 유의확률은 $P(F_{2,27} > 10.125)$이다.

ㄷ. 각 처리의 반복수가 모두 같다면 각 처리에서 9개씩 조사한 것이다.

① ㄴ

② ㄷ

③ ㄱ, ㄴ

④ ㄱ, ㄴ, ㄷ

정답 1번

출제 포인트

이 문제는 일원분산분석표에서 자유도, 전체 표본수, 검정통계량의 의미를 복원해 ㄱ, ㄴ, ㄷ의 참거짓을 판별하는 문제이다.

조건 정리

주어진 값은 처리 제곱합 17.956, 처리 평균제곱 8.978, 오차 제곱합 23.941, 오차 평균제곱 0.887, 전체 자유도 29, 검정통계량 10.125이다.
구해야 할 값은 처리 자유도, 오차 자유도, 전체 표본수, 그리고 각 진술의 참거짓이다.
주의할 조건은 분산분석에서 귀무가설 기각이 곧 모든 평균의 쌍별 차이를 뜻하지는 않는다는 점이다.

NoteBase

사용할 공식

일원분산분석에서는 df_처리=k-1, df_오차=N-k, df_전체=N-1, F=MS_처리/MS_오차를 사용한다.
또한 자유도는 $df = SS/MS$로 복원할 수 있다.
이 공식을 쓰는 이유는 표에 제곱합, 평균제곱, 전체 자유도가 함께 주어져 있어 누락된 자유도와 표본수를 직접 계산할 수 있기 때문이다.

계산 과정

1. 처리 자유도는 $17.956/8.978 = 2$이므로 2이다. 따라서 처리 수는 $k = 2 + 1 = 3$이다.
2. 전체 자유도는 29이므로 전체 표본수는 $N = 29 + 1 = 30$이다.
3. 오차 자유도는 $N - k = 30 - 3 = 27$이다. 이는 $23.941/0.887 \approx 27$과도 일치한다.
4. 검정통계량은 $F = 8.978/0.887 = 10.125$이다. 이는 임계값 $F_{0.05}(2, 27) = 3.35$보다 크므로 귀무가설을 기각한다.
5. 귀무가설 기각의 의미는 세 평균이 모두 같지는 않다는 뜻이며, 적어도 하나의 평균이 다르다는 뜻이다.

선지별 해설

① O
- ㄴ만 옳다. 유의확률은 관측된 검정통계량보다 큰 오른쪽 꼬리확률이므로 $P(F_{2,27} > 10.125)$이다. 반면 ㄱ은 '모두 다르다'로 과도하게 결론을 확장한 표현이고, ㄷ은 반복수가 9가 아니라 10이어서 틀리다.

② X
- ②는 ㄷ만 옳다고 본 선택지인데 ㄷ이 틀리므로 오답이다. 전체 표본수는 30이고 처리 수는 3이므로 반복수가 모두 같다면 각 처리의 반복수는 $30/3 = 10$이다.
- 틀린 표현은 '각 처리에서 9개씩 조사'이다. 올바르게는 '각 처리에서 10개씩 조사'이다. 선택지는 'ㄴ'으로 고쳐져야 한다.

③ X
- ③은 ㄱ, ㄴ이 옳다고 본 선택지인데 ㄱ이 틀리므로 오답이다. $F = 10.125 > 3.35$이므로 귀무가설은 기각되지만, 이것만으로 세 평균이 모두 서로 다르다고 단정할 수는 없다.
- 틀린 표현은 ㄱ의 '모두 다르다'이다. 올바르게는 '적어도 하나의 평균은 다르다'이다. 선택지는 'ㄴ'으로 고쳐져야 한다.

④ X
- ④는 ㄱ, ㄴ, ㄷ이 모두 옳다고 본 선택지인데 ㄱ과 ㄷ이 모두 틀리므로 오답이다. ㄴ만 옳다.
- 틀린 표현은 ㄱ의 '모두 다르다'와 ㄷ의 '9개씩'이다. 올바르게는 '적어도 하나의 평균은 다르다', '각 처리에서 10개씩 조사하였다'이다. 선택지는 'ㄴ'으로 고쳐져야 한다.

NoteBase

오답 포인트

분산분석에서 귀무가설을 기각하면 곧바로 '세 집단 평균이 모두 다르다'고 받아들이는 실수가 잦다. 이 검정은 적어도 하나의 평균 차이만 보장한다. 또 전체 자유도 29를 표본수 29로 착각하면 반복수를 9로 잘못 계산하게 된다.

다시 나오면 이렇게 판단

분산분석표 문제는 먼저 df_전체=N-1로 총 표본수를 복원하고, df_처리=k-1로 처리 수를 확인한 뒤, *F*값과 임계값을 비교해 귀무가설 기각 여부를 판단하는 순서로 풀면 된다. 그다음 '기각의 의미가 적어도 하나의 평균 차이인지, 모든 평균 차이인지'를 반드시 구분한다.

암기 포인트

ANOVA에서 유의하다는 말은 적어도 하나의 평균이 다름을 뜻하고, 3처리에서 전체 자유도 29이면 총 표본수는 30이므로 균등반복수는 10이다.

NoteBase

문 14. 다음은 인자 A, B의 수준수가 각각 3이고, 각 처리에서 반복수가 3인 이원배치법 분산분석표의 일부이다. 이에 대한 설명으로 옳은 것은? (단, 자유도가 k_1, k_2인 F분포의 제100$(1-\alpha)$백분위수를 $F_\alpha(k_1, k_2)$로 나타낼 때 $F_{0.05}(2, 22) = 3.44$, $F_{0.05}(3, 20) = 3.10$이며, 인자 A, B의 교호작용은 고려하지 않는다)

요인	제곱합	자유도	평균제곱	$F-$값
인자 A	100			
인자 B	20			
오차				
계	230			

① 인자 A의 F-값이 인자 B의 F-값보다 작다.

② 인자 B의 유의확률은 0.05보다 작다.

③ 오차항의 자유도는 20이다.

④ 오차평균제곱은 5이다.

정답 4번

출제 포인트

이 문제는 이원배치법 분산분석표에서 자유도, 오차제곱합, 오차평균제곱, F값을 계산해 선지를 판정하는 문제이다.

조건 정리

주어진 값: 인자 A의 제곱합은 100, 인자 B의 제곱합은 20, 총제곱합은 230이다.
구해야 할 값: 오차항의 자유도와 오차평균제곱, 각 인자의 F값이다.
주의할 조건: 인자 A, B는 각각 수준 수가 3이고, 각 처리 반복 수가 3이며, 교호작용은 고려하지 않는다.

NoteBase

사용할 공식

자유도는 전체 $abn-1$, 인자 A는 $a-1$, 인자 B는 $b-1$, 오차는 전체 자유도에서 인자 자유도를 뺀 값이다.
오차제곱합은 $SSE = SST - SSA - SSB$이다.
평균제곱은 $MS = SS/df$, 검정통계량은 $F = MS_{\text{factor}}/MS_E$이다.
이 공식을 쓰는 이유: 주어진 표가 이원배치 분산분석표의 일부이므로 빈칸은 분산분석의 기본 관계식으로 채워야 하기 때문이다.

계산 과정

전체 관측수는 $3 \times 3 \times 3 = 27$이므로 전체 자유도는 $27 - 1 = 26$이다.
인자 A와 인자 B의 자유도는 각각 $3 - 1 = 2$이다.
교호작용을 고려하지 않으므로 오차 자유도는 $26 - 2 - 2 = 22$이다.
오차제곱합은 $230 - 100 - 20 = 110$이다.
오차평균제곱은 $110/22 = 5$이다.
인자 A의 평균제곱은 $100/2 = 50$, 인자 B의 평균제곱은 $20/2 = 10$이다.
따라서 $F_A = 50/5 = 10$, $F_B = 10/5 = 2$이다.

선지별 해설

① X: 인자 A의 F값은 10이고 인자 B의 F값은 2이므로 인자 A의 F값이 더 크다. 틀린 표현은 '작다'이고, 올바르게는 '인자 A의 F값이 인자 B의 F값보다 크다'이다.
② X: 인자 B의 검정통계량은 $F_B = 2$이고 임계값은 $F_{0.05}(2, 22) = 3.44$이다. $2 < 3.44$이므로 귀무가설을 기각하지 못해 유의확률은 0.05보다 크다. 틀린 표현은 '0.05보다 작다'이고, 올바르게는 '인자 B의 유의확률은 0.05보다 크다'이다.
③ X: 교호작용을 고려하지 않는 조건에서 오차 자유도는 전체 자유도 26에서 인자 A, B의 자유도 2와 2를 뺀 22이다. 틀린 표현은 '20'이고, 올바르게는 '오차항의 자유도는 22이다'이다.
④ O: 오차제곱합이 110이고 오차 자유도가 22이므로 오차평균제곱은 $110/22 = 5$이다. 계산 결과와 일치하므로 옳다.

오답 포인트

교호작용을 고려하지 않는다는 조건을 놓치면 오차 자유도를 $ab(n-1) = 18$로 잘못 잡기 쉽다. 이 문항은 그 실수를 유도한 뒤 오차평균제곱과 F값까지 연쇄적으로 틀리게 만드는 함정이다.

다시 나오면 이렇게 판단

이원배치 분산분석표 문제는 먼저 교호작용 포함 여부를 확인하고, 그다음 전체 자유도와 각 요인 자유도를 정리한 뒤 오차제곱합, 오차평균제곱, F값 순서로 계산하면 된다.

NoteBase

암기 포인트

분산분석표는 자유도부터 정리하면 SSE, MSE, F값이 한 줄로 이어진다.

NoteBase

문 15. 다중선형회귀모형 $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i,\ \ i = 1, 2, \cdots, n$에서 모수 $\beta_0, \beta_1, \beta_2$를 추정하기 위해 사용할 수 있는 방법은? (단, 오차항 ε_i는 서로 독립이며 평균 0, 분산 σ^2인 확률변수이다)

① 최소제곱법(least squares method)

② 최대우도법(maximum likelihood estimation)

③ 우도비 검정법(likelihood ratio test)

④ 분산분석법(ANOVA)

정답 1번

출제 포인트

이 문제는 다중선형회귀모형에서 주어진 가정만으로 사용할 수 있는 회귀계수 추정법과 추정법이 아닌 절차를 구분하는 문제이다.

핵심 개념

최소제곱법: 잔차제곱합을 최소로 하여 회귀계수 $\beta_0, \beta_1, \beta_2$를 추정하는 방법이다 / 결정 차이축: 회귀계수를 직접 추정하는가 / 키워드: 잔차제곱합 최소, 회귀계수 추정
최대우도법: 우도를 최대화하여 모수를 추정하는 방법이지만, 우도를 세우려면 오차분포를 구체적으로 가정해야 한다 / 결정 차이축: 추정법이지만 분포가정이 추가로 필요한가 / 키워드: 우도, 정규성 가정
우도비 검정법: 두 모형 또는 가설의 우도를 비교하는 가설검정 방법이다 / 결정 차이축: 추정이 아니라 검정인가 / 키워드: 모형 비교, 가설검정
분산분석법: 제곱합을 분해하여 모형이나 요인의 유의성을 판단하는 분석 절차이다 / 결정 차이축: 회귀계수 추정이 아니라 유의성 판단인가 / 키워드: ANOVA, 제곱합 분해

NoteBase

선지별 해설

① O
최소제곱법은 다중선형회귀에서 $\beta_0, \beta_1, \beta_2$를 추정하는 대표적인 방법이다.
문항에 제시된 회귀모형과 오차 조건 아래에서 바로 적용할 수 있는 추정법이므로 옳다.

② X
틀린 부분은 주어진 조건만으로 최대우도법을 사용할 수 있다는 취지이다.
최대우도법도 추정법이지만, 우도를 구성하려면 오차항의 분포를 구체적으로 정해야 하므로 보통 정규성 같은 추가 가정이 필요하다.
올바르게는 오차가 특정 분포, 대표적으로 정규분포를 따른다는 가정이 추가되면 최대우도법을 사용할 수 있다이다.

③ X
틀린 부분은 우도비 검정법을 모수 추정 방법으로 본 표현이다.
우도비 검정법은 제한모형과 비제한모형의 우도를 비교하여 가설을 검정하는 절차이지 회귀계수를 추정하는 방법이 아니다.
올바르게는 우도비 검정법은 추정법이 아니라 가설검정 방법이다이다.

④ X
틀린 부분은 분산분석법을 회귀계수 추정 방법으로 본 표현이다.
분산분석법은 회귀제곱합과 오차제곱합을 비교하여 모형 전체의 유의성을 판단하는 데 쓰이며, $\beta_0, \beta_1, \beta_2$ 자체를 추정하는 절차가 아니다.
올바르게는 분산분석법은 회귀계수 추정법이 아니라 모형의 유의성 분석 방법이다이다.

오답 포인트

회귀분석에서 자주 등장하는 용어라는 이유만으로 모두 회귀계수 추정법으로 묶으면 틀리기 쉽다. 특히 ②는 추정법이라는 점 때문에 그럴듯하지만, 이 문항은 주어진 조건만으로 가능한 방법을 묻기 때문에 추가 분포가정이 필요한 최대우도법은 정답이 아니다.

다시 나오면 이렇게 판단

먼저 선지가 추정법인지 검정법인지부터 가른다. 추정법으로 남은 선택지에 대해서는 주어진 가정만으로 바로 적용되는지, 아니면 정규성 같은 추가 분포가정이 필요한지를 확인하면 된다.

암기 포인트

회귀계수 추정은 최소제곱법, 우도비와 ANOVA는 검정, 최대우도법은 분포가정이 붙는 추정법이다.

NoteBase

문 16. X_1과 X_2는 서로 독립이고 평균이 10, 분산이 1인 정규분포를 따르는 확률변수이다. Y가 X_1과 X_2 중 최댓값일 때, Y가 10보다 클 확률은?

① 0.5

② 0.625

③ 0.75

④ 0.875

정답 3번

출제 포인트

이 문제는 독립인 두 정규확률변수의 최댓값에 대해 여사건을 이용하여 확률을 구하는 문제이다.

조건 정리

주어진 값: X_1과 X_2는 서로 독립이고 각각 평균 10, 분산 1인 정규분포를 따른다.
구해야 할 값: $Y = \max(X_1, X_2)$일 때 $P(Y > 10)$이다.
주의할 조건: 최댓값이 10보다 크지 않으려면 두 변수 모두 10 이하여야 한다.

사용할 공식

$P(Y > 10) = 1 - P(Y \le 10) = 1 - P(X_1 \le 10, X_2 \le 10)$
$= 1 - P(X_1 \le 10)P(X_2 \le 10)$
이 공식을 쓰는 이유는 Y가 두 값의 최댓값이므로 $Y \le 10$이 되는 경우가 두 값이 모두 10 이하인 경우와 정확히 같기 때문이다. 또 X_1과 X_2가 독립이므로 결합확률을 곱으로 계산할 수 있다.

계산 과정

1. 평균이 10인 정규분포는 10을 중심으로 대칭이므로 $P(X_1 \le 10) = 0.5$, $P(X_2 \le 10) = 0.5$이다.
2. 따라서 $P(Y \le 10) = P(X_1 \le 10, X_2 \le 10) = 0.5 \times 0.5 = 0.25$이다.
3. 그러므로 $P(Y > 10) = 1 - 0.25 = 0.75$이다.

NoteBase

선지별 해설

① X
0.5는 한 변수만 놓고 $P(X_i > 10)$ 또는 $P(X_i \le 10)$를 본 값이다. 이 문제는 최댓값 Y를 묻기 때문에 두 변수를 함께 봐야 한다. 틀린 표현은 최댓값 사건을 한 변수 사건처럼 본 점이며, 올바르게는
$P(Y > 10) = 1 - (0.5)^2 = 0.75$이다.

② X
0.625는 여사건을 잘못 계산한 값이다. $P(Y \le 10)$은 0.25이므로 1에서 빼면 0.75가 되어야 한다. 틀린 표현은 $P(X_1 \le 10, X_2 \le 10)$을 잘못 처리한 점이며, 올바르게는 독립성을 써서 $0.5 \times 0.5 = 0.25$로 계산한다.

③ O
$Y > 10$은 적어도 하나가 10보다 큰 경우이다. 그 여사건은 두 변수 모두 10 이하인 경우이고, 그 확률은
$0.5 \times 0.5 = 0.25$이다. 따라서 $P(Y > 10) = 1 - 0.25 = 0.75$이므로 옳다.

④ X
0.875는 최댓값 사건의 여사건을 과도하게 작게 본 값이다. 두 변수 중 하나만 10 이하인 경우까지 여사건으로 잘못 밀어 넣으면 이런 식의 과대계산이 나온다. 틀린 표현은 $Y \le 10$의 조건을 잘못 해석한 점이며, 올바르게는 두 변수 모두 10 이하여야 하므로 확률은 0.25, 정답 확률은 0.75이다.

오답 포인트

평균이 10인 정규분포에서 $P(X \le 10) = 0.5$라는 사실만 보고 바로 0.5를 찍기 쉽다. 이 문항의 핵심 함정은 '최댓값이 10보다 크다'를 '한 변수 하나가 10보다 크다'와 혼동하는 점이다.

다시 나오면 이렇게 판단

최댓값이나 최솟값 확률은 직접 세기보다 여사건으로 바꾸어 본다. 최댓값이 기준값보다 크다는 사건은 '둘 다 기준값 이하'의 여사건으로, 최솟값이 기준값보다 작다는 사건은 '둘 다 기준값 이상'의 여사건으로 바꾸면 계산이 빨라진다.

암기 포인트

최댓값은 여사건으로 푼다. P(>a)=1-P(모두 ≤ a)이다.

NoteBase

문 17. 과목 A의 점수는 $N(50, 5^2)$을 따르고, 과목 B의 점수는 $N(60, 10^2)$을 따른다. 갑돌이의 과목 A, B 점수는 각각 45점, 80점이고, 갑순이의 과목 A, B 점수는 각각 50점, 70점이다. 다음 설명 중 옳은 것만을 모두 고르면? (단, Z가 $N(0, 1)$을 따를 때 $P(Z \le 1) = 0.84$, $P(Z \le 2) = 0.98$이고, 어떤 사람의 과목 성적이 해당 과목의 점수 분포에서 하위 p% 지점에 위치하면 이 사람의 해당 과목 백분위점수는 p라고 정의한다)

ㄱ. 두 과목 원점수의 합은 갑돌이가 더 높다.
ㄴ. 두 과목 표준화점수(z-score)의 합은 갑돌이와 갑순이가 같다.
ㄷ. 두 과목 백분위점수의 합은 갑순이가 더 높다.

① ㄱ, ㄴ
② ㄱ, ㄷ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 4번

출제 포인트

평균과 표준편차가 다른 두 분포에서 원점수, 표준화점수, 백분위점수를 각각 비교할 수 있는지를 묻는 문제이다.

조건 정리

- 과목 A: 평균 50, 표준편차 5
- 과목 B: 평균 60, 표준편차 10
- 갑돌: A 45점, B 80점
- 갑순: A 50점, B 70점
- 백분위 계산에 필요한 값: $P(Z \le 1) = 0.84$, $P(Z \le 2) = 0.98$, 따라서 $P(Z \le 0) = 0.50$, $P(Z \le -1) = 0.16$이다.

사용할 공식

표준화점수는 $z = \frac{x - \mu}{\sigma}$이다. 백분위점수는 해당 z값의 누적확률이다.
이 공식을 쓰는 이유는 과목 A와 B의 평균과 표준편차가 달라 원점수를 바로 비교하면 상대적 위치를 알 수 없기 때문이다.

NoteBase

계산 과정

1. 원점수 합을 비교하면 갑돌은 $45+80=125$, 갑순은 $50+70=120$이므로 ㄱ은 참이다.
2. 표준화점수를 구하면 갑돌은 과목 A에서 $z=\frac{45-50}{5}=-1$, 과목 B에서 $z=\frac{80-60}{10}=2$이므로 합이 1이다. 갑순은 과목 A에서 $z=\frac{50-50}{5}=0$, 과목 B에서 $z=\frac{70-60}{10}=1$이므로 합이 1이다. 따라서 ㄴ은 참이다.
3. 백분위점수를 구하면 갑돌은 A에서 16, B에서 98이므로 합이 114이다. 갑순은 A에서 50, B에서 84이므로 합이 134이다. 따라서 갑순의 백분위점수 합이 더 커서 ㄷ은 참이다.

선지별 해설

① X

- ①은 ㄱ, ㄴ만 옳다고 본 선택지이다. 실제로는 ㄱ이 참이고, ㄴ도 두 사람의 z합이 모두 1이어서 참이며, ㄷ도 갑순의 백분위 합 134가 갑돌의 114보다 커서 참이다. 틀린 표현은 'ㄱ, ㄴ'만 고른 부분이며, 올바르게는 'ㄱ, ㄴ, ㄷ'이다.

② X

- ②는 ㄱ, ㄷ만 옳다고 본 선택지이다. 그러나 ㄴ도 참이다. 갑돌의 표준화점수 합은 $-1+2=1$, 갑순의 표준화점수 합은 $0+1=1$로 같기 때문이다. 틀린 표현은 ㄴ을 제외한 부분이며, 올바르게는 'ㄱ, ㄴ, ㄷ'이다.

③ X

- ③은 ㄴ, ㄷ만 옳다고 본 선택지이다. 그러나 ㄱ도 참이다. 원점수 합은 갑돌 125, 갑순 120이므로 갑돌이 더 높다. 틀린 표현은 ㄱ을 제외한 부분이며, 올바르게는 'ㄱ, ㄴ, ㄷ'이다.

④ O

- ④는 ㄱ, ㄴ, ㄷ이 모두 옳다고 본 선택지이다. 원점수 합은 갑돌이 더 높고, 표준화점수 합은 두 사람이 같고, 백분위점수 합은 갑순이 더 높으므로 세 진술이 모두 참이다.

오답 포인트

원점수 합과 백분위 합을 같은 기준으로 비교하면 틀리기 쉽다. 과목마다 평균과 표준편차가 다르면 같은 10점 차이도 상대적 의미가 달라지므로 z값과 누적확률을 따로 계산해야 한다. 또한
$P(Z \le -1)=1-0.84=0.16$을 바로 떠올리지 못하면 ㄷ 판단에서 실수하기 쉽다.

다시 나오면 이렇게 판단

서로 다른 분포의 점수를 비교하는 문제는 원점수 비교, 표준화점수 비교, 백분위 비교를 분리해서 판단한다. 먼저 각 점수를 z값으로 바꾸고, 필요하면 누적확률로 백분위를 구하면 각 진술의 참거짓이 정리된다.

암기 포인트

원점수 비교는 단순 합산, 분포가 다른 점수의 상대 비교는 z, 상대적 위치 비교는 백분위이다.

NoteBase

문 18. 두 정규모집단에서 각각 랜덤 추출한 서로 독립인 표본 1과 표본 2의 통계값이 다음과 같다. 이에 대한 설명으로 옳은 것만을 모두 고르면? (단, 자유도가 k_1, k_2인 F분포의 제100$(1-\alpha)$백분위수를 $F_\alpha(k_1, k_2)$로 나타낼 때, $F_{0.025}(9, 11) = 3.59$, $F_{0.025}(11, 9) = 3.91$, $F_{0.05}(9, 11) = 2.90$, $F_{0.05}(11, 9) = 3.10$이다)

구분	표본크기	평균	표준편차
표본 1	10	10	3
표본 2	12	15	2

ㄱ. 유의수준 5%에서 두 집단의 모분산은 다르다고 할 수 없다.

ㄴ. 두 집단의 모분산이 같다는 가정하에, 두 표본의 분산을 사용하여 공통분산의 불편추정값을 구하면 6.25이다.

ㄷ. 두 집단의 모분산이 같다는 가정하에, 두 집단의 모평균이 같다는 귀무가설과 같지 않다는 대립가설을 검정하기 위해 검정통계량의 값 $\frac{5}{2.5 \times \sqrt{\frac{1}{10} + \frac{1}{12}}}$를 사용할 수 있다.

① ㄱ, ㄴ

② ㄱ, ㄷ

③ ㄴ, ㄷ

④ ㄱ, ㄴ, ㄷ

정답 4번

출제 포인트

등분산성의 F 검정, 공통분산의 불편추정, 등분산 가정하 두 평균의 t 검정통계량을 함께 판단하는 문제이다.

조건 정리

- 표본 1은 표본크기 10, 표본평균 10, 표준편차 3이므로 분산은 $s_1^2 = 9$이다.
- 표본 2는 표본크기 12, 표본평균 15, 표준편차 2이므로 분산은 $s_2^2 = 4$이다.
- 모분산 비교에서는 큰 표본분산을 분자로 두어 $F = \frac{9}{4} = 2.25$를 본다.
- 등분산 가정이 가능하면 공통분산 S_p^2와 pooled t 통계량을 사용한다.

사용할 공식

F=s_큰^2s_작은^2, $S_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$, $t = \frac{\bar{X}_i - \bar{X}_j}{S_p\sqrt{1/n_1 + 1/n_2}}$이다.
이 공식을 쓰는 이유는 두 표본이 정규모집단에서 나온 서로 독립인 표본이므로 모분산 비교에는 F 검정, 등분산 가정하 평균 비교에는 pooled t 검정을 적용할 수 있기 때문이다.

NoteBase

계산 과정

1. 모분산 검정통계량은 $F = \frac{9}{4} = 2.25$이다.
2. 유의수준 5% 양측검정의 기각역은 $F > F_{0.025}(9, 11) = 3.59$ 또는 $F < \frac{1}{F_{0.025}(11, 9)} = \frac{1}{3.91} \approx 0.256$이다. 2.25는 기각역에 들어가지 않으므로 두 모분산이 다르다고 할 수 없다.
3. 공통분산의 불편추정값은 $S_p^2 = \frac{9 \cdot 9 + 11 \cdot 4}{10 + 12 - 2} = \frac{81 + 44}{20} = 6.25$이다.
4. 따라서 $S_p = \sqrt{6.25} = 2.5$이고, 평균 비교의 검정통계량은 $t = \frac{\bar{X}_2 - \bar{X}_1}{S_p\sqrt{1/10 + 1/12}} = \frac{15 - 10}{2.5\sqrt{1/10 + 1/12}} = \frac{5}{2.5\sqrt{1/10 + 1/12}}$로 둘 수 있다.

선지별 해설

① X: ①은 ㄱ, ㄴ만 옳다고 본 조합이다. 그러나 ㄷ도 $\bar{X}_2 - \bar{X}_1$ 순서로 검정통계량을 잡으면 제시된 식 그대로 사용할 수 있으므로 ①은 일부만 고른 조합이다. 틀린 표현은 'ㄱ, ㄴ만'이고, 올바르게는 'ㄱ, ㄴ, ㄷ 모두'이다.
② X: ②는 ㄱ, ㄷ만 옳다고 본 조합이다. ㄴ에서 공통분산의 불편추정값이 6.25로 정확하므로 ②는 ㄴ을 빠뜨려 틀리다. 틀린 표현은 'ㄱ, ㄷ만'이고, 올바르게는 'ㄱ, ㄴ, ㄷ 모두'이다.
③ X: ③은 ㄴ, ㄷ만 옳다고 본 조합이다. ㄱ도 $F = 2.25$가 5% 양측검정의 기각역에 들어가지 않아 옳으므로 ③은 ㄱ을 빠뜨려 틀리다. 틀린 표현은 'ㄴ, ㄷ만'이고, 올바르게는 'ㄱ, ㄴ, ㄷ 모두'이다.
④ O: ㄱ은 유의수준 5%에서 두 모분산이 다르다고 할 수 없다는 뜻이므로 맞다. ㄴ은 $S_p^2 = 6.25$로 계산되어 맞다. ㄷ은 pooled 표준편차가 2.5이고 평균 차를 $15 - 10 = 5$로 잡으면 제시된 검정통계량을 그대로 사용할 수 있으므로 맞다.

오답 포인트

양측 F 검정에서 임계값을 $F_{0.05}$로 바로 쓰면 ㄱ을 잘못 판단하기 쉽다. 또한 표준편차 3, 2를 그대로 넣지 않고 먼저 분산 9, 4로 바꿔야 하며, 공통분산은 단순평균이 아니라 자유도 가중평균이라는 점을 놓치기 쉽다.
ㄷ에서는 평균 차의 부호가 아니라 절댓값 크기와 분모 구조를 본다는 점이 함정이다.

다시 나오면 이렇게 판단

두 정규모집단 비교 문제는 먼저 표준편차를 분산으로 바꾼 뒤 F 검정으로 등분산 여부를 판단하고, 등분산이면 S_p^2를 자유도 가중평균으로 구한 다음 pooled t 통계량으로 평균 비교 선지를 정리하는 순서로 판단한다.

암기 포인트

두 표본 등분산 검정은 큰 분산을 위에 둔 F로 먼저 거르고, 등분산이면 S_p^2는 자유도 가중평균이며, t의 부호는 평균을 뺀 순서에 따라 달라진다.

NoteBase

문 19. 단순선형회귀모형 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ \ i = 1, 2, \cdots, 11$에 최소제곱법을 적용하여 얻은 분산분석표의 일부가 다음과 같을 때, 옳지 않은 것은? (단, ε_i는 $N(0, \sigma^2)$을 따르고 서로 독립이며, 자유도가 k_1, k_2인 F분포의 제100$(1-\alpha)$백분위수를 $F_\alpha(k_1, k_2)$로 나타낼 때 $F_{0.01}(1, 9) = 10.56$이다)

요인	제곱합	자유도	평균제곱	F−값
회귀	75			
오차(잔차)	45			
합계	120			

① X, Y의 표본상관계수는 $\sqrt{0.625}$이다.

② 오차항 분산의 불편추정값은 5이다.

③ F-값은 15이다.

④ 유의수준 5%에서 회귀모형이 유의하다.

정답 1번

출제 포인트

단순선형회귀의 분산분석표에서 SSR, SSE, SST와 자유도를 이용해 MSE, F, R^2를 구하고, $R^2 = r^2$에서 상관계수의 부호까지 판단할 수 있는지를 묻는 문제이다.

조건 정리

표본수는 11이므로 전체 자유도는 10, 회귀 자유도는 1, 오차 자유도는 9이다.
제곱합은 $SSR = 75, SSE = 45, SST = 120$이다.
발문은 옳지 않은 것을 고르는 문제이다.

사용할 공식

단순선형회귀에서는 $MSR = SSR/1, MSE = SSE/(n-2), F = MSR/MSE, R^2 = SSR/SST = r^2$이다.
이 공식을 쓰는 이유는 분산분석표의 제곱합과 자유도만으로 회귀모형의 유의성, 오차분산의 추정값, 상관계수의 크기를 모두 연결해서 판단할 수 있기 때문이다.

NoteBase

계산 과정

1. 오차 자유도는 $n-2=11-2=9$이다.
2. 오차항 분산의 불편추정값은 $MSE=SSE/9=45/9=5$이다.
3. 회귀 평균제곱은 $MSR=SSR/1=75$이므로 $F=MSR/MSE=75/5=15$이다.
4. 결정계수는 $R^2=SSR/SST=75/120=0.625$이므로 $r^2=0.625$이다. 따라서 표본상관계수는 $r=\pm\sqrt{0.625}$이고, 부호는 회귀계수 $\hat{\beta}_1$의 부호에 따라 정해진다.
5. $F=15$는 $F_{0.01}(1,9)=10.56$보다 크므로 1% 유의수준에서도 유의하고, 따라서 5% 유의수준에서도 유의하다.

선지 판단

1번: X. $R^2=0.625$에서 바로 알 수 있는 것은 $r^2=0.625$뿐이다. 선지의 틀린 표현은 '표본상관계수는 $\sqrt{0.625}$이다'라는 단정이다. 올바르게는 '표본상관계수는 $\pm\sqrt{0.625}$이며, 부호는 $\hat{\beta}_1$의 부호와 같다'이다.
2번: O. 오차항 분산의 불편추정값은 $MSE=SSE/(n-2)=45/9=5$이므로 맞는 설명이다.
3번: O. $F=MSR/MSE=(75/1)/(45/9)=75/5=15$이므로 맞는 설명이다.
4번: O. $F=15$는 제시된 $F_{0.01}(1,9)=10.56$보다 크다. 1%에서도 유의하므로 5%에서는 당연히 유의하다.

오답 포인트

가장 흔한 함정은 $R^2=r^2$를 보고 상관계수를 바로 양수 $\sqrt{R^2}$로 적는 실수이다. 이 문제에서는 기울기의 부호 정보가 없으므로 상관계수의 절댓값만 알 수 있다.

다시 나오면 이렇게 판단

단순회귀 분산분석표 문항은 먼저 자유도 $1, n-2, n-1$를 채우고, 다음에 MSE와 F를 계산한 뒤, 마지막에 $R^2=r^2$에서 상관계수의 부호가 확정되는지까지 확인하면 된다.

암기 포인트

단순선형회귀에서는 $R^2=r^2$, $MSE=SSE/(n-2)$, $F=MSR/MSE$이며, 상관계수의 부호는 $\hat{\beta}_1$의 부호와 같다.

NoteBase

문 20. 중위소득 이상인 가구와 미만인 가구를 각각 100개씩 랜덤 추출하여 특정 정책에 대한 가구주의 의견을 조사한 결과 다음과 같은 찬반 교차표를 얻었다. 이에 대한 설명으로 옳은 것만을 모두 고르면?

구분	찬성	반대	합
중위소득 미만	60	40	100
중위소득 이상	40	60	100
합	100	100	200

소득 수준에 따라 찬성률의 차이가 없다는 귀무가설과 차이가 있다는 대립가설을 검정하기 위해 다음 두 가지 통계값을 사용하고자 한다.

$$a = \frac{0.6 - 0.4}{\sqrt{0.5 \times 0.5 \times \left(\frac{1}{100} + \frac{1}{100}\right)}}$$

$$b = \frac{(60-50)^2}{50} + \frac{(40-50)^2}{50} + \frac{(60-50)^2}{50} + \frac{(40-50)^2}{50}$$

ㄱ. a를 사용하여 정규검정법으로 검정할 수 있다.
ㄴ. b를 사용하여 χ^2검정법으로 검정할 수 있다.
ㄷ. a를 사용한 검정의 유의확률이 b를 사용한 검정의 유의확률보다 작다.

① ㄴ
② ㄱ, ㄴ
③ ㄱ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 2번

출제 포인트

두 독립표본의 비율 차이 검정에서 정규검정 통계량과 2 × 2 분할표의 카이제곱 검정통계량이 어떻게 연결되는지를 묻는 문제이다.

NoteBase

조건 정리

중위소득 미만 가구의 찬성 비율은 $\hat{p}_1 = 60/100 = 0.6$이고, 중위소득 이상 가구의 찬성 비율은 $\hat{p}_2 = 40/100 = 0.4$이다.
귀무가설은 두 집단의 찬성률이 같다는 $H_0 : p_1 = p_2$이다.
귀무가설 아래 합동추정 비율은 $\hat{p} = (60 + 40)/200 = 0.5$이다.

사용할 공식

두 비율 차이의 정규검정 통계량은
$a = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$
이다.
같은 자료를 2 × 2 분할표로 보면 Pearson 카이제곱 통계량은
b=Σ ((관측도수-기대도수)^2)/(기대도수)
이다.
이 공식을 쓰는 이유는 이 문제가 두 집단의 비율 비교를 분할표 형태로 제시하고 있어 정규검정과 카이제곱검정을 모두 적용할 수 있기 때문이다.

계산 과정

1. 정규검정 통계량은
$a = \frac{0.6 - 0.4}{\sqrt{0.5 \times 0.5 \times (\frac{1}{100} + \frac{1}{100})}} = \frac{0.2}{\sqrt{0.005}} \approx 2.828$
이므로 a를 사용해 정규검정법으로 검정할 수 있다.
2. 각 셀의 기대도수는 모두 50이므로
$b = \frac{(60-50)^2}{50} + \frac{(40-50)^2}{50} + \frac{(40-50)^2}{50} + \frac{(60-50)^2}{50} = 8$
이다. 따라서 b를 사용해 자유도 1의 χ^2 검정법으로 검정할 수 있다.
3. 자유도 1인 두 검정은 $b = a^2$ 관계를 가진다. 실제로 $a^2 \approx 2.828^2 = 8 = b$이므로 두 검정의 유의확률은 같다.

NoteBase

선지별 해설

① X

`ㄴ`만 고른 선택지이다. `ㄱ`도 옳으므로 틀렸다. 틀린 표현은 `ㄴ만 옳다`이고, 올바르게는 `ㄱ, ㄴ이 옳다`이다.

② O

`ㄱ, ㄴ`을 고른 선택지이다. $a \approx 2.828$은 두 비율 차이의 정규검정 통계량이므로 `ㄱ`은 옳다. $b = 8$은 같은 자료의 Pearson 카이제곱 통계량이므로 `ㄴ`도 옳다.

③ X

`ㄷ`이 틀려서 오답이다. 자유도 1에서는 $b = a^2$이므로 정규검정과 카이제곱검정의 유의확률은 같고, `a를 사용한 검정의 유의확률이 더 작다`고 볼 수 없다. 틀린 표현은 `ㄱ, ㄷ이 옳다`이고, 올바르게는 `ㄱ, ㄴ이 옳다`이다.

④ X

`ㄷ`까지 포함해 고른 선택지이므로 틀렸다. `ㄱ`과 `ㄴ`은 옳지만 `ㄷ`은 틀리다. 틀린 표현은 `ㄱ, ㄴ, ㄷ이 옳다`이고, 올바르게는 `ㄱ, ㄴ만 옳다`이다.

오답 포인트

`ㄷ`에서 서로 다른 검정법이므로 유의확률도 다를 것이라고 착각하기 쉽다. 그러나 2×2 분할표에서 두 비율 비교를 할 때 자유도 1의 Pearson 카이제곱 통계량은 두 비율 차이의 Z 통계량 제곱과 같아 결론과 유의확률이 일치한다.

다시 나오면 이렇게 판단

두 집단의 비율 비교가 나오면 먼저 합동추정 비율로 Z 통계량을 만들고, 같은 자료를 2×2 분할표로 보았을 때 $\chi^2(1) = Z^2$ 관계가 성립하는지 확인한다. 이 관계가 보이면 검정 방법만 다를 뿐 결론과 유의확률은 같다고 판단할 수 있다.

암기 포인트

두 비율 비교의 2×2 분할표에서는 자유도 1의 χ^2 검정통계량이 Z^2와 같다.

NoteBase