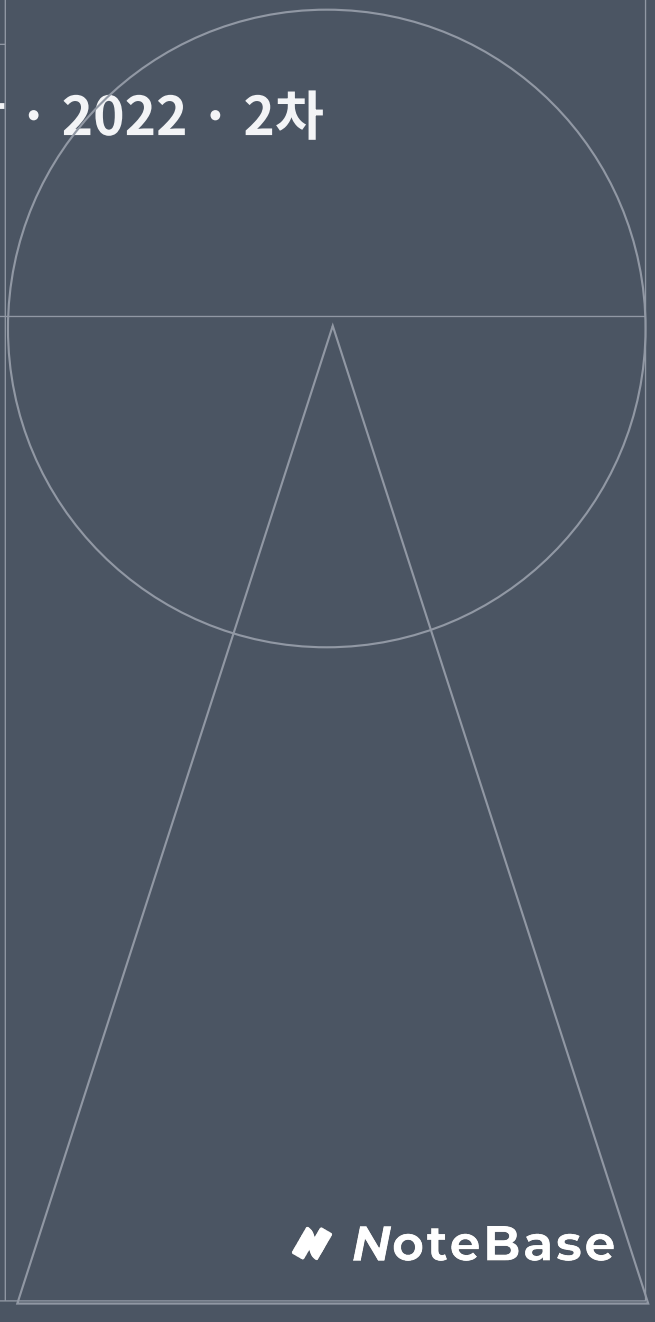
기출OX 해설집

공무원/국가직/7급 · 통계학 · 2022 · 2차



NoteBase

기출OX 해설집

공무원/국가직/7급 · 통계학 · 2022 · 2차

공무원/국가직/7급 · 통계학 · 2022 · 2차

문 1. 어떤 학교에서 35명의 학생을 무작위로 추출하여 수학 성적을 조사한 결과 평균과 중앙값은 각각 70점, 76점이었고, 동점자는 없었다. 5명의 학생을 추가로 뽑아 수학 성적을 조사하여 다음 값을 얻었을 때, 40명 학생의 수학 성적 평균과 중앙값을 바르게 연결한 것은?

77, 76, 61, 81, 75

① 평균: 70.5, 중앙값: 75

② 평균: 70.5, 중앙값: 76

③ 평균: 71.0, 중앙값: 75

④ 평균: 71.0, 중앙값: 76

정답 2번

출제 포인트

기존 자료의 평균과 중앙값이 주어졌을 때 새로 추가된 점수들을 반영하여 전체 평균과 중앙값을 다시 구하는 문제이다.

조건 정리

주어진 값: 기존 35명의 평균은 70점, 중앙값은 76점이다. 동점자가 없으므로 기존 35명을 오름차순으로 놓으면 18번째 값이 76점이다.
추가된 값: 77, 76, 61, 81, 75이다.
구해야 할 값: 40명의 평균과 중앙값이다.
주의할 조건: 40명은 짝수명이므로 중앙값은 20번째와 21번째 값의 평균으로 정한다.

NoteBase

사용할 공식

평균은 전체 합계를 인원수로 나눈 값이다. 즉, 평균은 총점의 증감을 먼저 계산해 구한다.
중앙값은 자료를 정렬했을 때 가운데 오는 값이다. 40명처럼 짝수 개 자료에서는 20번째 값과 21번째 값의 평균이 중앙값이 된다.
이 공식을 쓰는 이유: 이 문제는 일부 점수만 추가되었으므로 기존 평균으로 기존 총점을 먼저 복원한 뒤 새 점수를 더하면 전체 평균을 바로 구할 수 있다. 중앙값도 전체를 새로 정렬할 때 76보다 작은 값, 76과 같은 값, 76보다 큰 값의 개수를 세면 결정된다.

계산 과정

1. 기존 35명의 총점은 $35 \times 70 = 2450$이다.
2. 추가된 5명의 총점은 $77 + 76 + 61 + 81 + 75 = 370$이다.
3. 새 총점은 $2450 + 370 = 2820$이므로 새 평균은 $2820 \div 40 = 70.5$이다.
4. 기존 35명은 동점자가 없고 중앙값이 76이므로 76보다 작은 값이 정확히 17개, 76보다 큰 값이 정확히 17개이다.
5. 새 점수 중 76보다 작은 값은 61, 75의 2개이고, 76과 같은 값은 76의 1개이며, 76보다 큰 값은 77, 81의 2개이다.
6. 따라서 40명을 정렬하면 76보다 작은 값이 $17 + 2 = 19$개이고, 20번째와 21번째 값은 모두 76이다. 중앙값은 $(76 + 76) \div 2 = 76$이다.

선지별 해설

① X: 평균 70.5는 맞지만 중앙값 75가 틀리다. 76보다 작은 값이 19개이므로 20번째와 21번째 값은 모두 76이다. 틀린 표현은 "중앙값 75"이고, 올바르게는 "중앙값 76"이다.
② O: 평균은 새 총점 2820을 40명으로 나누어 70.5가 되고, 중앙값은 20번째와 21번째 값이 모두 76이므로 76이다.
③ X: 평균 71.0과 중앙값 75가 모두 틀리다. 평균은 $2820 \div 40 = 70.5$이므로 틀린 표현은 "평균 71.0"이고, 올바르게는 "평균 70.5"이다. 중앙값도 "75"가 아니라 "76"이다.
④ X: 중앙값 76은 맞지만 평균 71.0이 틀리다. 추가된 5명의 합계를 잘못 더하거나 35명 평균을 그대로 비율처럼 다루면 71.0 같은 오답이 나온다. 틀린 표현은 "평균 71.0"이고, 올바르게는 "평균 70.5"이다.

오답 포인트

중앙값을 새로 들어온 점수들만 보고 75나 76 중 하나로 감으로 고르는 경우가 많다. 이 문항의 함정은 기존 35명에서 76보다 작은 값이 정확히 17개라는 사실을 이용해야 한다는 점이다. 평균도 기존 평균 70에서 점수가 조금 올랐다고 보고 71.0으로 성급히 잡으면 틀린다.

NoteBase

다시 나오면 이렇게 판단

평균과 중앙값이 함께 나오면 평균은 총점으로, 중앙값은 정렬 후 가운데 위치로 따로 판단한다. 특히 기존 중앙값과 동점 없음이 주어지면 중앙값 기준으로 작은 값과 큰 값의 개수를 먼저 확정한 뒤 새 자료를 끼워 넣어 20번째와 21번째 값을 찾으면 된다.

암기 포인트

평균은 합계, 중앙값은 자리수로 판단한다. 짝수 개 자료의 중앙값은 가운데 두 값의 평균이다.

NoteBase

문 2. 다음은 헌혈한 사람 중 80명을 무작위로 뽑아 혈액형을 정리한 표이다. 헌혈한 사람의 혈액형 비(O형, A형, B형, AB형 순)가 3:4:2:1이라는 귀무가설을 유의수준 10%에서 검정하기 위한 검정통계량의 값과 귀무가설의 검정결과를 바르게 연결한 것은? (단, $\chi^2_\alpha(k)$는 자유도가 k인 카이제곱분포의 $(1-\alpha)$분위수를 나타내고, $\chi^2_{0.1}(3)=6.25$, $\chi^2_{0.1}(4)=7.78$이다)

혈액형	O	A	B	AB	합계
빈도	26	40	8	6	80

① 검정통계량의 값: $\frac{20}{3}$, 귀무가설의 검정결과: 기각함

② 검정통계량의 값: $\frac{20}{3}$, 귀무가설의 검정결과: 기각하지 못함

③ 검정통계량의 값: $\frac{43}{6}$, 귀무가설의 검정결과: 기각함

④ 검정통계량의 값: $\frac{43}{6}$, 귀무가설의 검정결과: 기각하지 못함

정답 1번

출제 포인트

이 문제는 카이제곱 적합도 검정에서 기대도수를 구하고 검정통계량을 계산한 뒤 임계값과 비교할 수 있는지를 묻는다.

조건 정리

주어진 값: 관측도수는 O형 26, A형 40, B형 8, AB형 6이다.
귀무가설: 혈액형의 비는 O:A:B:AB = 3:4:2:1이다.
전체 표본수: 80명이다.
기대도수: 비율의 합이 10이므로 각각 80×3/10, 80×4/10, 80×2/10, 80×1/10에서 24, 32, 16, 8이다.
주의할 조건: 범주 수가 4개이므로 자유도는 $4-1=3$이다.

사용할 공식

검정통계량은 $\chi^2=\sum\frac{(O_i-E_i)^2}{E_i}$이다.
이 공식을 쓰는 이유: 관측도수와 귀무가설이 주는 기대도수의 차이를 범주별로 비교하는 적합도 검정 문제이기 때문이다.

NoteBase

계산 과정

1. 기대도수를 구하면 $(24, 32, 16, 8)$이다.
2. 검정통계량은 $\chi^2 = \frac{(26-24)^2}{24} + \frac{(40-32)^2}{32} + \frac{(8-16)^2}{16} + \frac{(6-8)^2}{8}$ 이다.
3. 각 항은 $1/6, 2, 4, 1/2$이므로 $\chi^2 = 1/6 + 2 + 4 + 1/2 = 20/3$이다.
4. 자유도 3에서 유의수준 10%의 임계값은 6.25이고, $20/3 \approx 6.67 > 6.25$이므로 귀무가설을 기각한다.

선지 판단

1번: O. 검정통계량이 20/3이고 자유도 3의 임계값 6.25보다 크므로 귀무가설을 기각한다.
2번: X. 검정통계량 20/3까지는 맞지만 $20/3 > 6.25$이므로 '기각하지 못함'이 틀리다. 올바르게는 '검정통계량의 값은 20/3이고 귀무가설은 기각함'이다.
3번: X. 43/6은 검정통계량 계산이 틀린 값이다. AB형 항 $\frac{(6-8)^2}{8}$은 1/2인데 이를 1로 잘못 계산하면 43/6이 나온다. 올바르게는 '검정통계량의 값은 20/3이고 귀무가설은 기각함'이다.
4번: X. 43/6도 잘못된 값이고, 설령 그 값을 쓰더라도 6.25보다 크므로 '기각하지 못함'이 될 수 없다. 올바르게는 '검정통계량의 값은 20/3이고 귀무가설은 기각함'이다.

오답 포인트

기대도수는 비율을 전체 표본수 80에 곱해 먼저 구해야 하고, 각 범주의 $\frac{(O-E)^2}{E}$를 끝까지 정확히 더해야 한다. 특히 마지막 AB형 항 $\frac{4}{8} = \frac{1}{2}$를 1로 잘못 더하면 오답값 43/6이 나온다.

다시 나오면 이렇게 판단

적합도 검정 문제는 비율을 기대도수로 바꾸고, χ^2 각 항을 계산한 뒤, 자유도 범주 수-1에 맞는 임계값과 비교하는 순서로 판단한다.

암기 포인트

적합도 검정은 기대도수 계산 → $\chi^2 = \sum \frac{(O-E)^2}{E}$ 계산 → 자유도 $k-1$로 임계값 비교의 흐름이다.

NoteBase

문 3. 자료 (x_i, y_i) $(i = 1, 2, \cdots, 20)$로부터 다음 값을 얻었을 때, 이 자료를 최소제곱법으로 적합한 단순선형회귀식은?

$$\sum_{i=1}^{20} x_i = 100,\ \sum_{i=1}^{20} x_i^2 = 1{,}200,\ \sum_{i=1}^{20} y_i = 1{,}000,$$
$$\sum_{i=1}^{20} y_i^2 = 70{,}000,\ \sum_{i=1}^{20} x_i y_i = 6{,}400$$

① $\hat{Y} = 40 + 2x$

② $\hat{Y} = 20 + 2x$

③ $\hat{Y} = 40 + x$

④ $\hat{Y} = 20 + x$

정답 1번

출제 포인트

이 문제는 단순선형회귀식의 기울기와 절편을 $S_{xx}, S_{xy}, \bar{x}, \bar{y}$로 계산하는 문제이다.

조건 정리

주어진 값은 $n = 20, \sum x_i = 100, \sum x_i^2 = 1,200, \sum y_i = 1,000, \sum x_i y_i = 6,400$이다.
구해야 할 값은 최소제곱법으로 적합한 단순선형회귀식 $\hat{Y} = a + bx$의 절편 a와 기울기 b이다.
주의할 조건은 합계값을 바로 대입하지 말고 먼저 평균과 편차제곱합으로 바꾸어 계산해야 한다는 점이다.

사용할 공식

단순선형회귀식은 $\hat{Y} = a + bx$이고, 기울기와 절편은 $b = \frac{S_{xy}}{S_{xx}}, a = \bar{y} - b\bar{x}$로 구한다.
여기서 $S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}, S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$이다.
이 공식을 쓰는 이유는 문제에서 원자료가 아니라 합계형 요약통계를 주었으므로, 최소제곱 추정량을 요약통계에서 바로 복원해야 하기 때문이다.

계산 과정

1. 평균을 구하면 $\bar{x} = 100/20 = 5, \bar{y} = 1000/20 = 50$이다.
2. $S_{xx} = 1200 - 100^2/20 = 1200 - 500 = 700$이다.
3. $S_{xy} = 6400 - (100 \times 1000)/20 = 6400 - 5000 = 1400$이다.
4. 기울기는 $b = S_{xy}/S_{xx} = 1400/700 = 2$이다.
5. 절편은 $a = \bar{y} - b\bar{x} = 50 - 2 \times 5 = 40$이다.
6. 따라서 회귀식은 $\hat{Y} = 40 + 2x$이다.

NoteBase

선지별 해설

1번 O: 계산 결과 절편이 40, 기울기가 2이므로 $\hat{Y} = 40 + 2X$가 맞다.
2번 X: 기울기 2는 맞지만 절편 20이 틀렸다. 틀린 표현은 "20"이고, 올바르게는 $\hat{Y} = 40 + 2X$이다. 이 값은 절편 계산에서 $50 - 2 \times 5$를 잘못 처리했을 때 나오는 오답이다.
3번 X: 절편 40은 맞지만 기울기 1이 틀렸다. 틀린 표현은 "+x"이고, 올바르게는 "+2x"이다. 이 값은 S_{xy}와 S_{xx}를 같게 보거나 1400/700 계산을 잘못했을 때 나오는 오답이다.
4번 X: 절편과 기울기가 모두 틀렸다. 틀린 표현은 "20"과 "+x"이고, 올바르게는 $\hat{Y} = 40 + 2X$이다. 이 식은 절편과 기울기를 모두 잘못 구한 경우의 오답이다.

오답 포인트

이 문항의 함정은 $\sum x_i$, $\sum y_i$, $\sum x_i y_i$를 바로 회귀계수 공식에 넣지 않고 반드시 $\bar{x}$, $\bar{y}$, S_{xx}, S_{xy}로 바꾸어야 한다는 점이다. 절편은 마지막에 $a = \bar{y} - b\bar{x}$로 구하므로, 기울기를 구한 뒤 평균을 다시 대입하는 순서를 놓치면 20 같은 오답이 나온다.

다시 나오면 이렇게 판단

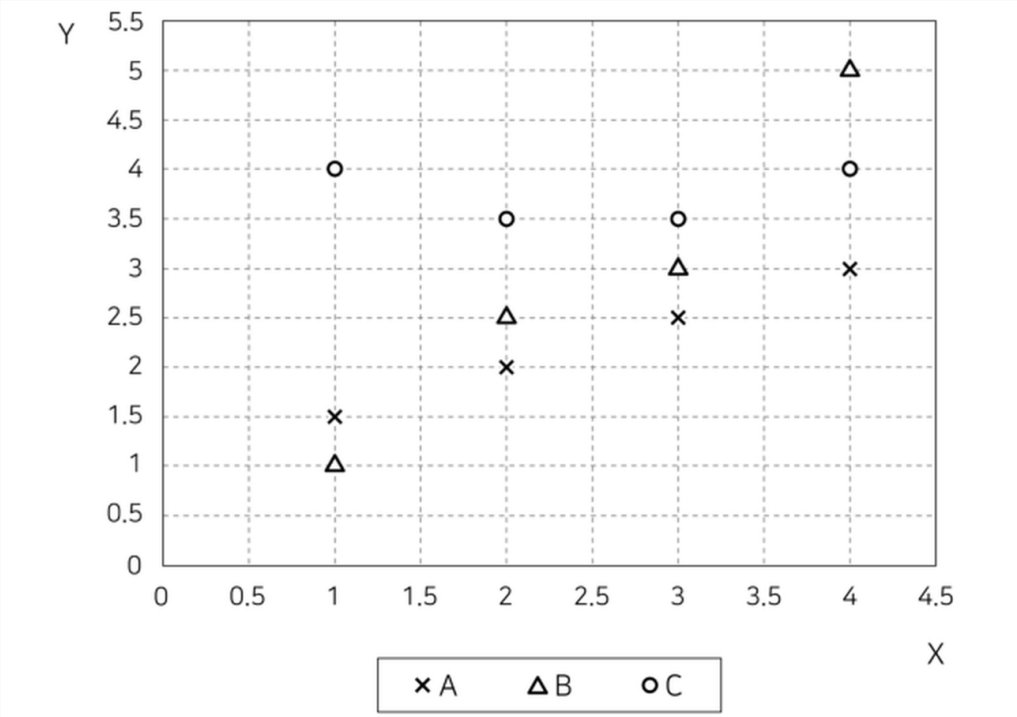
요약통계로 회귀식을 묻는 문제는 먼저 $\bar{x}$, $\bar{y}$를 구하고, 다음에 S_{xx}, S_{xy}를 계산한 뒤, 마지막에 b와 a를 순서대로 구하면 된다. 기울기를 먼저, 절편을 나중에 구하는 흐름을 유지하면 실수를 줄일 수 있다.

암기 포인트

단순선형회귀는 $b = \frac{S_{xy}}{S_{xx}}$, $a = \bar{y} - b\bar{x}$ 순서로 정리하면 된다.

NoteBase

문 4. 다음은 A, B, C 표본의 두 변수 X와 Y에 대한 산점도이다. X와 Y의 피어슨(Pearson) 표본상관계수가 큰 것부터 순서대로 나열한 것은?



① A, B, C

② A, C, B

③ B, A, C

④ C, B, A

정답 1번

출제 포인트

이 문제는 산점도에서 점들이 직선에 얼마나 가깝게 늘어서는지를 보고 피어슨 표본상관계수의 크기를 비교할 수 있는지를 묻는다.

NoteBase

선지별 해설

1번: O. A는 네 점이 정확히 한 직선 위에 있어 완전한 양의 선형관계이므로 상관계수가 가장 크다. B도 양의 상관이지만 점들이 직선에서 조금 벗어나므로 A보다 작다. C는 좌우가 평균을 기준으로 대칭이라 선형 공분산이 0이 되어 가장 작다.
2번: X. "A, C, B"가 틀린 표현이다. C가 B보다 크지 않다. C는 x가 커져도 y가 일정하게 증가하거나 감소하지 않아 선형상관이 0이고, B는 뚜렷한 양의 선형관계가 있어 C보다 크다. 올바른 순서는 A, B, C이다.
3번: X. "B, A, C"가 틀린 표현이다. B는 양의 상관이지만 완전한 직선관계가 아니므로 A보다 작다. A가 가장 크고 그다음이 B이다. 올바른 순서는 A, B, C이다.
4번: X. "C, B, A"가 틀린 표현이다. C는 가장 큰 것이 아니라 가장 작다. C는 평균을 기준으로 위아래 편차가 상쇄되어 선형상관이 0이고, A는 완전한 양의 선형관계이므로 가장 크다. 올바른 순서는 A, B, C이다.

오답 포인트

점들이 오른쪽으로 갈수록 대체로 올라간다는 이유만으로 B와 C를 함께 양의 상관으로 묶으면 틀린다. 피어슨 상관계수는 증가 경향 자체보다 직선성의 강도를 본다. C처럼 가운데를 기준으로 위아래가 상쇄되는 모양은 산점도가 퍼져 보여도 선형상관은 0이 된다.

다시 나오면 이렇게 판단

먼저 점들이 한 직선 위에 정확히 놓이는 표본이 있는지 본다. 그다음 증가 방향이면 양의 상관, 감소 방향이면 음의 상관으로 판단한다. 마지막으로 직선에서 벗어나는 정도가 클수록 상관계수의 절댓값이 작아진다고 정리하면 된다.

암기 포인트

피어슨 상관계수는 직선성의 방향과 강도를 재는 값이다. 완전한 직선이면 가장 크거나 가장 작고, 대칭적으로 상쇄되면 0이 된다.

NoteBase

문 5. 두 정규모집단 $N(\mu_1, \sigma_1^2)$과 $N(\mu_2, \sigma_2^2)$으로부터 각각 크기가 n_1, n_2이며 독립인 확률표본을 뽑아 가설 $H_0 : \sigma_1^2 = \sigma_2^2$ 대 $H_1 : \sigma_1^2 > \sigma_2^2$을 검정통계량 $F_0 = \frac{S_1^2}{S_2^2}$을 사용하여 유의수준 α로 검정할 때, 귀무가설의 기각역은? (단, S_1^2과 S_2^2는 두 표본의 분산이고, $F_\alpha(k_1, k_2)$는 자유도가 k_1, k_2인 F분포의 $(1-\alpha)$분위수를 나타낸다)

① $F_0 > F_{1-\alpha}(n_1 - 1, n_2 - 1)$

② $F_0 > F_\alpha(n_2 - 1, n_1 - 1)$

③ $F_0 > \frac{1}{F_\alpha(n_1 - 1, n_2 - 1)}$

④ $F_0 > \frac{1}{F_{1-\alpha}(n_2 - 1, n_1 - 1)}$

정답 4번

출제 포인트

이 문제는 두 정규모집단 분산의 단측 검정에서 F 검정통계량의 기각역을 올바른 자유도와 분위수 표기로 쓰는지를 묻는다.

조건 정리

두 모집단은 정규분포이고 표본은 서로 독립이다.
귀무가설은 $H_0 : \sigma_1^2 = \sigma_2^2$, 대립가설은 $H_1 : \sigma_1^2 > \sigma_2^2$이다.
검정통계량은 $F_0 = \frac{S_1^2}{S_2^2}$이고, 귀무가설 아래에서 $F_0 \sim F(n_1 - 1, n_2 - 1)$이다.

사용할 공식

상측 단측검정이므로 기각역은 $F_0 > F_\alpha(n_1 - 1, n_2 - 1)$이다.
또한 F분포의 역수 관계로 $F_\alpha(v_1, v_2) = \frac{1}{F_{1-\alpha}(v_2, v_1)}$가 성립한다.
이 공식을 쓰는 이유는 대립가설이 $\sigma_1^2 > \sigma_2^2$이므로 S_1^2/S_2^2가 충분히 큰 쪽에서만 귀무가설을 기각해야 하기 때문이다.

계산 과정

1. 대립가설이 $\sigma_1^2 > \sigma_2^2$이므로 검정통계량 $F_0 = \frac{S_1^2}{S_2^2}$가 커질수록 대립가설에 유리하다.
2. 따라서 기각역은 우측꼬리 확률이 α가 되도록 잡은 임계값보다 큰 영역이다.
3. 문제의 표기에서 $F_\alpha(n_1 - 1, n_2 - 1)$는 $(1-\alpha)$분위수이므로 정답 형태는 $F_0 > F_\alpha(n_1 - 1, n_2 - 1)$이다.
4. 보기 ④의 $\frac{1}{F_{1-\alpha}(n_2 - 1, n_1 - 1)}$는 위 역수 관계로 바로 $F_\alpha(n_1 - 1, n_2 - 1)$와 같으므로 ④가 맞다.

NoteBase

선지 판단

1번: X. 상측검정이라는 방향은 맞지만 임계값 첨자가 틀렸다. 문제에서 F_{α}가 $(1-\alpha)$분위수인데, $F_{1-\alpha}(n_1-1, n_2-1)$는 α분위수라 너무 작은 값이다. 틀린 표현은 $F_{1-\alpha}(n_1-1, n_2-1)$이고, 올바르게는 $F_{\alpha}(n_1-1, n_2-1)$이다.
2번: X. 자유도를 뒤바꾼 점이 틀렸다. $F_0 = \frac{S_1^2}{S_2^2}$의 분자 자유도는 n_1-1, 분모 자유도는 n_2-1이어야 한다. 틀린 표현은 $F_{\alpha}(n_2-1, n_1-1)$이고, 올바르게는 $F_{\alpha}(n_1-1, n_2-1)$이다.
3번: X. 역수 관계를 쓸 때 첨자와 자유도를 함께 바꿔야 하는데 그중 첨자만 맞고 자유도 순서가 틀렸다. $\frac{1}{F_{\alpha}(n_1-1, n_2-1)}$는 정답 임계값과 같은 값이 아니다. 올바른 형태는 $\frac{1}{F_{1-\alpha}(n_2-1, n_1-1)}$이다.
4번: O. $\frac{1}{F_{1-\alpha}(n_2-1, n_1-1)} = F_{\alpha}(n_1-1, n_2-1)$이므로 상측 단측검정의 정확한 기각역이다.

오답 포인트

이 문제의 함정은 F_{α}의 뜻을 평소 표기와 반대로 읽는 실수와, 역수 관계를 적용하면서 자유도 순서를 같이 바꿔야 한다는 점을 놓치는 데 있다. ①은 첨자 함정, ②는 자유도 위치 함정, ③은 역수 공식의 불완전 적용 함정이다.

다시 나오면 이렇게 판단

분산 검정은 먼저 대립가설 방향을 보고 우측검정인지 좌측검정인지 정한다. 그다음 검정통계량의 분자 자유도와 분모 자유도를 그대로 적고, 역수 관계를 쓰는 경우에는 첨자와 자유도 순서를 동시에 바꾸는지 확인하면 된다.

암기 포인트

F검정에서 S_1^2/S_2^2를 쓰면 자유도는 항상 (n_1-1, n_2-1) 순서이고, $F_{\alpha}(v_1, v_2) = 1/F_{1-\alpha}(v_2, v_1)$이다.

NoteBase

문 6. 주머니 안에 흰 공과 검은 공이 섞여 있다. 전체 공 10개 중 흰 공의 개수에 대한 다음 가설을 고려한다. 주머니에서 공 세 개를 무작위로 비복원추출한 후, 세 개의 공이 모두 흰색이면 귀무가설을 기각한다. 이때 제2종 오류를 범할 확률의 최댓값은?

◯ 귀무가설 (H_0): 흰 공의 개수는 5개 이하이다.
◯ 대립가설 (H_1): 흰 공의 개수는 6개 이상이다.

① $\frac{1}{12}$
② $\frac{1}{6}$
③ $\frac{5}{6}$
④ $\frac{11}{12}$

정답 3번

출제 포인트

제2종 오류의 의미를 정확히 잡고, 대립가설 범위에서 그 확률이 가장 커지는 경계값을 찾아 조합확률로 계산하는 문제이다.

조건 정리

전체 공은 10개이고, 흰 공의 개수를 W라 둔다.
귀무가설은 $H_0 : W \leq 5$이고 대립가설은 $H_1 : W \geq 6$이다.
주머니에서 공 3개를 비복원추출한 뒤 세 개가 모두 흰 공이면 H_0를 기각한다.
제2종 오류는 실제로는 H_1가 참인데도 H_0를 기각하지 못하는 경우이다.

사용할 공식

제2종 오류 확률은
β(W)=P(기각하지 않음 W≥ 6)=1-P(세 개 모두 흰 공 W)
이다.
비복원추출에서 세 개 모두 흰 공일 확률은
P(세 개 모두 흰 공 W)=W3103
이다.
흰 공의 수 W가 커질수록 세 개 모두 흰 공일 확률은 커지므로, 제2종 오류 확률은 대립가설에서 가장 불리한 경계값 $W = 6$에서 최댓값이 된다.

NoteBase

계산 과정

1. 대립가설 $H_1 : W \geq 6$에서 제2종 오류 확률 $\beta(W)$의 최댓값을 구해야 한다.
2. W가 클수록 세 개 모두 흰 공일 확률 $\frac{\binom{W}{3}}{\binom{10}{3}}$는 커지므로, 기각하지 못할 확률 $\beta(W)$는 작아진다.
3. 따라서 최댓값은 $W = 6$일 때 나온다.
4. 이때 기각확률은
P(세 개 모두 흰 공 W=6)=63103=(20)/(120)=(1)/(6)
이다.
5. 따라서 제2종 오류 확률은
$\beta(6) = 1 - \frac{1}{6} = \frac{5}{6}$
이다.

선지별 해설

① X: $\frac{1}{12}$는 제2종 오류 확률의 최댓값이 아니다. 올바르게는 $W = 6$일 때 $\beta(6) = \frac{5}{6}$이다. 이 값은 조합 계산이나 보완확률 처리를 잘못한 경우에 나오는 오답이다.
② X: $\frac{1}{6}$은 제2종 오류 확률이 아니라 $W = 6$일 때의 기각확률, 즉 세 개 모두 흰 공이 뽑힐 확률이다. 올바르게는 제2종 오류 확률의 최댓값은 그 보완확률인 $\frac{5}{6}$이다.
③ O: 대립가설에서 제2종 오류 확률이 가장 큰 경우는 흰 공 수가 가장 적은 $W = 6$일 때이다. 이때 $\beta(6) = 1 - \frac{\binom{6}{3}}{\binom{10}{3}} = 1 - \frac{20}{120} = \frac{5}{6}$이므로 옳다.
④ X: $\frac{11}{12}$는 제2종 오류 확률의 최댓값이 아니다. 올바르게는 $\frac{5}{6}$이다. 이 값은 세 개 모두 흰 공일 확률을 $\frac{1}{12}$로 잘못 계산한 뒤 다시 보완확률을 취한 경우에 해당한다.

오답 포인트

이 문항의 함정은 제2종 오류를 바로 계산하지 않고, 먼저 기각확률과 혼동하는 데 있다. 또 대립가설 전체에서 최댓값을 묻는데 $W = 6$이라는 경계값을 잡지 못하면 계산 방향이 흔들린다.

다시 나오면 이렇게 판단

가설검정 확률 문항은 먼저 오류의 뜻을 사건으로 번역한다. 그다음 대립가설 범위에서 그 확률이 가장 커지거나 작아지는 경계값을 찾고, 마지막에 조합확률과 보완확률을 구분해 계산하면 된다.

암기 포인트

제2종 오류는 대립가설이 참인데 기각하지 못하는 경우이고, 단조적으로 움직이는 검정에서는 최댓값이 경계값에서 자주 나온다.

NoteBase

문 7. Y는 이항분포 $B(n,p)$를 따르는 확률변수이다. 성공 확률 p의 추정량 $\hat{p}$를 $\frac{Y+2}{n+3}$라고 할 때, 편의(bias)는?

① $\frac{-3p}{n+3}$
② 0
③ $\frac{1-2p}{n+3}$
④ $\frac{2-3p}{n+3}$

정답 4번

출제 포인트

이 문제는 이항분포 $B(n,p)$에서 추정량의 기댓값을 구해 편의 $E[\hat{p}]-p$를 계산할 수 있는지를 묻는다.

조건 정리

주어진 값은 $Y \sim B(n,p)$, 추정량은 $\hat{p}=\frac{Y+2}{n+3}$이다.
구해야 할 값은 추정량의 편의이다.
편의는 추정량의 기댓값에서 모수 p를 뺀 값이다.

사용할 공식

이항분포의 기댓값은 $E[Y]=np$이고, 편의는 $\text{Bias}(\hat{p})=E[\hat{p}]-p$이다.
이 공식을 쓰는 이유는 발문이 추정량의 편의를 직접 묻고 있으므로 먼저 추정량의 기댓값을 계산해야 하기 때문이다.

계산 과정

1. 추정량의 기댓값은 $E[\hat{p}]=E\left[\frac{Y+2}{n+3}\right]=\frac{E[Y]+2}{n+3}$이다.
2. $Y \sim B(n,p)$이므로 $E[Y]=np$이고, 따라서 $E[\hat{p}]=\frac{np+2}{n+3}$이다.
3. 편의는 $E[\hat{p}]-p=\frac{np+2}{n+3}-p=\frac{np+2-pn-3p}{n+3}=\frac{2-3p}{n+3}$이다.

선지 판단

1번: X. $-\frac{3p}{n+3}$라고 보면 상수항 2를 빠뜨린 값이다. 편의는 $\frac{2-3p}{n+3}$이므로 틀렸다. 틀린 표현은 분자의 2가 누락된 점이고, 올바르게는 $\frac{2-3p}{n+3}$이다.
2번: X. 편의가 0이 되려면 $2-3p=0$, 즉 $p=\frac{2}{3}$인 특수한 경우여야 한다. 일반적인 p에 대해 항상 0은 아니다. 올바르게는 편의를 식으로 계산해 $\frac{2-3p}{n+3}$로 써야 한다.
3번: X. $\frac{1-2p}{n+3}$는 분자 정리에서 상수항과 p의 계수를 잘못 계산한 값이다. $\frac{np+2}{n+3}-p$를 통분하면 분자는 $2-3p$가 된다. 올바르게는 $\frac{2-3p}{n+3}$이다.
4번: O. $E[\hat{p}]=\frac{np+2}{n+3}$이므로 편의는 $\frac{np+2}{n+3}-p=\frac{2-3p}{n+3}$이다. 계산 결과와 일치하므로 옳다.

NoteBase

오답 포인트

편의를 묻는 문제에서 $E[\hat{p}]$까지만 구하고 끝내거나, p를 빼는 과정에서 통분을 대충 처리하면 1번이나 3번 같은 오답이 나온다. 특히 분자 정리에서 $np - pn$은 소거되지만 상수 2와 $-3p$는 그대로 남는다는 점이 함정이다.

다시 나오면 이렇게 판단

추정량의 편의를 묻는 계산형은 추정량 식 확인, 기댓값 계산, 마지막에 모수 차감의 순서로 판단한다. 선형성으로 $E[\hat{p}]$를 먼저 구한 뒤 반드시 p를 빼서 통분까지 마쳐야 한다.

암기 포인트

편의는 $E[\hat{\theta}] - \theta$이고, 이항분포에서는 $E[Y] = np$이다.

NoteBase

문 8. $X_1, X_2, \cdots, X_{10}$**은 구간** $(2, 4)$**에서의 균일분포(uniform distribution)로부터 뽑힌 확률표본이다. 이에 대한 설명으로 옳은 것만을 모두 고르면? (단,** $X_{(10)} = \max\{X_1, X_2, \cdots, X_{10}\}$**이다)**

ㄱ. $E(X_1^2) = \frac{28}{3}$
ㄴ. $2 < x < 4$일 때, $P(X_{10} > x) = 2 - \frac{1}{2}x$이다.
ㄷ. $E(X_{(10)}) = \frac{19}{5}$

① ㄱ, ㄴ
② ㄱ, ㄷ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 1번

출제 포인트

이 문제는 균일분포 $U(2, 4)$의 $E(X^2)$와 최댓값 순서통계량 $X_{(10)}$의 분포, 기대값을 정확히 구할 수 있는지를 묻는다.

조건 정리

표본은 X_1, X_2, , X_10iidU(2,4)이다.
$X_{(10)} = \max(X_1, X_2, \ \dots, X_{10})$이다.
ㄱ, ㄴ, ㄷ의 참거짓을 먼저 판정한 뒤 조합 선지를 고르면 된다.

사용할 공식

균일분포 $U(a, b)$에서 $E(X^2) = \frac{a^2 + ab + b^2}{3}$이다.
또한 최댓값 순서통계량은 $P(X_{(n)} \le x) = F(x)^n$, $E(X_{(n)}) = a + \frac{n}{n+1}(b - a)$를 쓴다.
이 문제의 ㄴ, ㄷ은 $X_{(10)}$에 대한 진술이므로 원래 분포의 누적분포함수 $F(x) = \frac{x-2}{2}$를 먼저 세워야 한다.

계산 과정

1. ㄱ: $E(X_1^2) = \frac{2^2 + 2 \cdot 4 + 4^2}{3} = \frac{4 + 8 + 16}{3} = \frac{28}{3}$이므로 참이다.
2. ㄴ: $2 < x < 4$에서 $F(x) = \frac{x-2}{2}$이므로 $P(X_{(10)} > x) = 1 - P(X_{(10)} \le x) = 1 - \left(\frac{x-2}{2}\right)^{10}$이다. 제시식 $2 - \frac{1}{2}x$는 최댓값의 확률이 아니라 단일 표본의 꼬리확률 $P(X_1 > x)$에 해당하므로 거짓이다.
3. ㄷ: $E(X_{(10)}) = 2 + \frac{10}{11}(4 - 2) = 2 + \frac{20}{11} = \frac{42}{11}$이다. 따라서 $\frac{19}{5}$라는 진술은 거짓이다.
4. 참인 것은 ㄱ만이므로 정답 조합은 1번이다.

NoteBase

선지별 해설

1번: O. ㄱ만 옳다. ㄱ은 $E(X_1^2)=\frac{28}{3}$으로 맞고, ㄴ은 올바르게는 $P(X_{(10)}>x)=1-\left(\frac{x-2}{2}\right)^{10}$이어야 하며, ㄷ은 올바르게는 $E(X_{(10)})=\frac{42}{11}$이다.
2번: X. 'ㄱ, ㄷ'이라고 했지만 ㄷ이 틀리므로 옳지 않다. 틀린 표현은 $E(X_{(10)})=\frac{19}{5}$이고, 올바르게는 $E(X_{(10)})=\frac{42}{11}$이다. 올바른 조합은 'ㄱ'만인 1번이다.
3번: X. 'ㄴ, ㄷ'이라고 했지만 ㄴ과 ㄷ이 모두 틀리므로 옳지 않다. ㄴ은 최댓값의 분포를 한 개 표본의 꼬리확률처럼 선형식으로 쓴 것이 오류이고, ㄷ은 최댓값의 기대값을 잘못 계산한 것이다. 올바른 조합은 1번이다.
4번: X. 'ㄱ, ㄴ, ㄷ'이라고 했지만 참인 진술은 ㄱ 하나뿐이다. 올바르게는 ㄴ을 $P(X_{(10)}>x)=1-\left(\frac{x-2}{2}\right)^{10}$, ㄷ을 $E(X_{(10)})=\frac{42}{11}$로 고쳐야 한다. 따라서 전체 조합도 1번으로 정리된다.

오답 포인트

ㄴ에서 최댓값 $X_{(10)}$의 확률을 구하면서 단일 표본 X_1의 꼬리확률 $P(X_1>x)=2-\frac{x}{2}$를 그대로 쓰기 쉽다. ㄷ에서도 최댓값의 기대값이 상한 4에 가깝다는 직관만 보고 근삿값을 넣으면 틀린다.

다시 나오면 이렇게 판단

균일분포와 순서통계량이 함께 나오면 먼저 원래 분포의 $F(x)$를 쓰고, 그다음 $P(X_{(n)}\le x)=F(x)^n$ 또는 $E(X_{(n)})=a+\frac{n}{n+1}(b-a)$를 적용한다. 원표본의 확률과 최댓값의 확률을 섞지 않으면 판단이 안정된다.

암기 포인트

균일분포 $U(a,b)$의 최댓값 평균은 $a+\frac{n}{n+1}(b-a)$, 최댓값 분포는 $F(x)^n$으로 정리된다.

NoteBase

문 9. 400명의 사람을 무작위로 뽑아 200명씩 두 그룹으로 나누고 각 그룹에 서로 다른 치료제 A, B를 복용시킨 후 치유된 비율을 조사한 결과, 치료제 A의 치유율은 40%, 치료제 B의 치유율은 60%로 나타났다. p_A, p_B를 각각 치료제 A, B의 모집단 치유율이라고 하고 가설 $H_0 : p_A = p_B$ 대 $H_1 : p_A < p_B$를 검정할 때, p-값(유의확률)은? (단, $\Phi(x)$는 표준정규분포의 누적분포함수이다)

① $\Phi(-1)$

② $\Phi(-2)$

③ $\Phi(-3)$

④ $\Phi(-4)$

정답 4번

출제 포인트

이 문제는 두 집단 비율 차이의 가설검정에서 귀무가설 $H_0 : p_A = p_B$ 아래 풀드 비율을 사용해 검정통계량과 p-값을 구하는 문제이다.

조건 정리

주어진 값: A집단은 200명 중 치유율 40%이므로 성공 수가 80명이고, B집단은 200명 중 치유율 60%이므로 성공 수가 120명이다.
구해야 할 값: 대립가설 $H_1 : p_A < p_B$에 대한 왼쪽꼬리 p-값이다.
주의할 조건: 귀무가설이 $p_A = p_B$이므로 표준오차 계산에는 각 집단의 표본비율 0.4, 0.6이 아니라 풀드 비율을 넣는다.

사용할 공식

검정통계량은 $Z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}(1-\hat{p})(1/n_A + 1/n_B)}}$이다.
여기서 풀드 비율은 $\hat{p} = \frac{x_A + x_B}{n_A + n_B}$이다.
이 공식을 쓰는 이유는 귀무가설이 두 모비율의 동일성을 가정하므로 공통 모비율의 추정치로 분산을 계산해야 하기 때문이다.

계산 과정

1. 성공 수는 $x_A = 200 \times 0.4 = 80$, $x_B = 200 \times 0.6 = 120$이다.
2. 풀드 비율은 $\hat{p} = \frac{80 + 120}{200 + 200} = \frac{200}{400} = 0.5$이다.
3. 표준오차는 $\sqrt{0.5(1 - 0.5)(1/200 + 1/200)} = \sqrt{0.25 \times 0.01} = \sqrt{0.0025} = 0.05$이다.
4. 검정통계량은 $Z = \frac{0.4 - 0.6}{0.05} = \frac{-0.2}{0.05} = -4$이다.
5. 대립가설이 $p_A < p_B$이므로 p-값은 $P(Z \leq -4) = \Phi(-4)$이다.

NoteBase

선지별 해설

① X
- $\Phi(-1)$은 검정통계량이 -1일 때의 왼쪽꼬리 확률인데, 이 문항의 검정통계량은 -4이다.
- 틀린 표현은 "p-값이 $\Phi(-1)$이다"이고, 올바르게는 "p-값은 $\Phi(-4)$이다"이다.

② X
- $\Phi(-2)$는 표준오차를 0.1처럼 잘못 크게 잡아 $\frac{-0.2}{0.1} = -2$로 계산할 때 나오는 값이다.
- 틀린 표현은 "표준오차를 $\sqrt{1/200 + 1/200}$처럼 계산한다"이고, 올바르게는 "표준오차는 $\sqrt{\hat{p}(1-\hat{p})(1/200 + 1/200)} = 0.05$이다"이다.

③ X
- $\Phi(-3)$는 정확한 풀드 검정을 적용한 결과가 아니며, 표준오차를 약 0.0667로 잘못 계산할 때 생기는 오답이다.
- 틀린 표현은 "검정통계량이 -3이다"이고, 올바르게는 "검정통계량은 -4이므로 p-값은 $\Phi(-4)$이다"이다.

④ O
- 귀무가설 아래 풀드 비율이 0.5이고 이에 따른 표준오차가 0.05이므로 검정통계량은 -4이다.
- 왼쪽꼬리 검정이므로 p-값은 $\Phi(-4)$가 맞다.

오답 포인트

함정은 두 가지이다. 첫째, 귀무가설이 $p_A = p_B$인데도 풀드 비율을 쓰지 않고 표준오차를 계산하면 검정통계량이 틀어진다. 둘째, 대립가설이 $p_A < p_B$인 왼쪽꼬리 검정인데 방향을 놓치면 $\Phi(-4)$ 대신 다른 꼬리확률을 고르게 된다.

다시 나오면 이렇게 판단

두 비율 검정 문제는 먼저 각 집단의 성공 수를 구하고, 귀무가설이 두 비율의 동일성인지 확인한 뒤 풀드 비율로 표준오차를 계산한다. 그다음 $\hat{p}_A - \hat{p}_B$의 부호와 대립가설의 방향을 맞춰 왼쪽꼬리인지 오른쪽꼬리인지 정하면 p-값이 바로 정해진다.

암기 포인트

두 모비율 동일성 검정은 귀무가설 아래 풀드 비율을 쓰고, 왼쪽꼬리 p-값은 $\Phi(z)$ 형태로 읽는다.

NoteBase

문 10. 100명 중 1명꼴로 감염되는 질병이 있다. 이 질병에 대한 감염 여부 검사에서 질병에 감염된 사람은 90%의 확률로, 질병에 감염되지 않은 사람은 5%의 확률로 양성으로 판정된다. 검사의 정확성을 높이기 위해 같은 검사를 두 번 실시하기로 한다. 무작위로 뽑힌 어떤 사람이 두 번의 검사에서 모두 양성으로 판정되었을 때, 실제로 그 질병에 감염되었을 확률은? (단, 두 번의 검사는 서로 독립이고, 양성 판정 확률은 동일하다고 가정한다)

① $\frac{2}{13}$
② $\frac{9}{20}$
③ $\frac{361}{757}$
④ $\frac{36}{47}$

정답 4번

출제 포인트

베이즈 정리로 두 번 모두 양성이라는 사후정보를 반영해 실제 감염 확률을 구하는 문제이다.

조건 정리

주어진 값
- 유병률: $P(D) = 1/100 = 0.01$
- 민감도: $P(+|D) = 0.9$
- 거짓양성률: $P(+|D^c) = 0.05$
- 두 검사는 서로 독립이다.

구해야 할 값
- $P(D|++)$

주의할 조건
- 두 번 모두 양성이므로 양성 확률을 한 번이 아니라 제곱해 반영해야 한다.

사용할 공식

$$P(D|++) = \frac{P(D)P(++|D)}{P(D)P(++|D) + P(D^c)P(++|D^c)}$$

이 공식을 쓰는 이유는 감염 여부를 알 수 없는 상태에서 두 번 모두 양성이라는 관측 결과가 주어졌기 때문이다. 따라서 사전확률인 유병률과 조건부확률인 두 번 양성될 확률을 함께 반영하는 베이즈 정리를 써야 한다.

NoteBase

계산 과정

1. 감염자라면 두 번 모두 양성일 확률은 $P(++|D) = 0.9^2 = 0.81$이다.
2. 비감염자라면 두 번 모두 양성일 확률은 $P(++|D^c) = 0.05^2 = 0.0025$이다.
3. 따라서

$P(D|++) = \frac{0.01 \times 0.81}{0.01 \times 0.81 + 0.99 \times 0.0025} = \frac{0.0081}{0.010575} = \frac{324}{423} = \frac{36}{47}$

계산 결과 실제 감염 확률은 36/47이다.

선지별 해설

① X

- 2/13은 한 번 양성일 때의 사후확률 $\frac{0.01 \times 0.9}{0.01 \times 0.9 + 0.99 \times 0.05}$에서 나오는 값이다. 이 문항은 두 번 모두 양성이므로 한 번 검사만 반영한 계산은 부족하다. 틀린 값 2/13 → 올바르게는 36/47이다.

② X

- 9/20은 비감염자가 두 번 모두 양성일 확률 0.99×0.05^2를 분모에 제대로 넣지 않은 계산에서 나오는 오답이다. 두 번 검사에서는 거짓양성률도 제곱되어야 한다. 틀린 값 9/20 → 올바르게는 36/47이다.

③ X

- 361/757은 두 번 양성이라는 조건 아래의 사후확률을 정확히 반영하지 못한 값이다. 분모는 $0.01 \times 0.9^2 + 0.99 \times 0.05^2$로 계산해야 하므로 이 값은 맞지 않는다. 틀린 값 361/757 → 올바르게는 36/47이다.

④ O

- 계산 결과 $P(D|++) = 36/47$이므로 맞는 선지이다. 두 번 모두 양성이라는 정보가 추가되면서 한 번 양성일 때보다 감염 확률이 크게 올라간다.

오답 포인트

이 문항의 함정은 두 번 검사를 했는데도 한 번 양성일 때의 베이즈 계산을 그대로 쓰거나, 비감염자의 두 번 양성 확률을 0.05가 아니라 0.05^2로 처리해야 한다는 점을 놓치는 데 있다. 특히 ①은 한 번 양성의 정답값이라서 가장 쉽게 끌리는 오답이다.

다시 나오면 이렇게 판단

검사 문제는 유병률, 민감도, 거짓양성률을 먼저 적고, 관측 결과가 한 번 양성인지 두 번 양성인지에 따라 조건부확률을 정확히 만든 뒤 베이즈 정리를 적용하면 된다. 반복 검사는 독립이면 해당 확률을 제곱해 반영한다.

암기 포인트

반복 검사 사후확률은 베이즈 정리에 반복 조건부확률을 곱해 넣는다. 두 번 양성이면 0.9^2, 0.05^2처럼 제곱이 들어간다.

NoteBase

문 11. 다음은 직장인 40명을 무작위로 뽑아 네 가지의 점심 메뉴 중 선호하는 메뉴를 조사한 것이다. 이 자료를 바탕으로 "네 가지 메뉴에 대한 선호도에 차이가 없다."라는 귀무가설을 검정하려고 한다. 관측도수와 귀무가설하에서의 기대도수 간 차이를 제곱한 값이 가장 큰 메뉴는?

메뉴	인원 수
김치찌개	18
돈가스	10
햄버거	7
자장면	5

① 김치찌개

② 돈가스

③ 햄버거

④ 자장면

정답 1번

출제 포인트

이 문제는 카이제곱 적합도 검정에서 각 범주의 기대도수를 먼저 구하고, 관측도수와 기대도수의 차이 제곱값 $(O_i - E_i)^2$을 비교할 수 있는지를 묻는다.

조건 정리

주어진 값: 전체 40명, 메뉴 4종, 관측도수는 김치찌개 18, 돈가스 10, 햄버거 7, 자장면 5이다.
구해야 할 값: 각 메뉴의 $(O_i - E_i)^2$ 중 가장 큰 메뉴이다.
주의할 조건: 귀무가설은 네 메뉴 선호도에 차이가 없다는 뜻이므로 각 메뉴의 기대도수는 모두 10이다.

사용할 공식

기대도수는 $E_i = n/k$, 비교값은 $(O_i - E_i)^2$이다.
이 공식을 쓰는 이유: 귀무가설이 네 메뉴의 선호 확률이 모두 같다는 뜻이므로 각 범주의 기대도수는 전체 40명을 4개 범주에 균등 배분한 10명이다. 문제는 카이제곱 통계량 전체가 아니라 차이의 제곱값만 비교하라고 했으므로 각 메뉴의 $(O_i - E_i)^2$만 계산하면 된다.

NoteBase

계산 과정

1. 기대도수는 메뉴별로 40/4 = 10이다.
2. 김치찌개는 $(18-10)^2 = 64$, 돈가스는 $(10-10)^2 = 0$이다.
3. 햄버거는 $(7-10)^2 = 9$, 자장면은 $(5-10)^2 = 25$이다.
4. 가장 큰 값은 64이므로 김치찌개가 정답이다.

선지 판단

1번 O: 김치찌개의 차이 제곱값은 $(18-10)^2 = 64$로 네 메뉴 중 가장 크다.
2번 X: 돈가스의 차이 제곱값은 $(10-10)^2 = 0$이다. 틀린 표현은 '가장 크다'이고, 올바르게는 '귀무가설과 정확히 일치하므로 차이 제곱값이 0이다'이다.
3번 X: 햄버거의 차이 제곱값은 $(7-10)^2 = 9$이다. 틀린 표현은 '가장 크다'이고, 올바르게는 '김치찌개보다 작고 자장면보다도 작다'이다.
4번 X: 자장면의 차이 제곱값은 $(5-10)^2 = 25$이다. 틀린 표현은 '가장 크다'이고, 올바르게는 '돈가스와 햄버거보다 크지만 김치찌개의 64보다는 작다'이다.

오답 포인트

기대도수를 40으로 두거나 차이의 부호만 보고 판단하면 틀리기 쉽다. 이 문항은 차이 자체가 아니라 차이의 제곱값을 비교하는 문제이므로 먼저 기대도수 10을 만든 뒤 네 범주를 같은 기준으로 계산해야 한다.

다시 나오면 이렇게 판단

균등분포 귀무가설이 보이면 먼저 기대도수를 전체 수 ÷ 범주 수로 맞춘다. 그다음 각 범주에 대해 (관측도수-기대도수)^2 또는 필요하면 (관측도수-기대도수)^2/기대도수를 계산해 가장 큰 항을 고르면 된다.

암기 포인트

적합도 검정의 출발점은 기대도수부터 정하고, 그다음 차이 제곱을 비교하는 것이다.

NoteBase

문 12. 다음 A, B, C, D 4개 지역에서 5명씩을 무작위로 뽑아 조사한 출근 소요시간에 대한 설명으로 옳지 않은 것은?

지역	출근 소요시간(분)
A	10, 20, 30, 40, 50
B	15, 25, 35, 45, 55
C	20, 40, 60, 80, 100
D	15, 35, 55, 75, 95

① A와 B 지역의 표준편차는 같다.

② A와 C 지역의 변동계수는 같다.

③ B와 C 지역의 변동계수는 같다.

④ C와 D 지역의 표준편차는 같다.

정답 3번

출제 포인트

표준편차는 모든 자료에 같은 수를 더하거나 빼도 변하지 않고, 변동계수는 $CV = \frac{s}{\bar{x}}$ 이므로 배수 변화에는 유지되지만 평행이동에는 달라진다는 점을 묻는 문제이다.

조건 정리

A 지역의 평균은 30이고 B 지역은 A 지역의 모든 값에 5를 더한 자료이다.
C 지역은 A 지역의 모든 값을 2배한 자료이고 D 지역은 C 지역의 모든 값에 5를 뺀 자료이다.
따라서 각 선지는 평행이동과 배수 변화가 표준편차와 변동계수에 미치는 영향을 비교하면 된다.

사용할 공식

변동계수는 $CV = \frac{s}{\bar{x}}$ 이다.
자료가 $Y = aX + b$ 로 바뀌면 평균은 $\bar{Y} = a\bar{X} + b$, 표준편차는 $s_Y = |a|s_X$ 이다.
이 공식을 쓰는 이유는 각 지역 자료가 A 지역 자료의 평행이동 또는 배수 변환으로 바로 표현되기 때문이다.

NoteBase

계산 과정

1. A 지역은 평균이 30이고 편차가 -20, -10, 0, 10, 20이므로 표준편차는 $s_A = 10\sqrt{2}$ 이다.
2. B 지역은 A 지역에 5를 더한 자료이므로 $s_B = s_A$, 평균은 35이다.
3. C 지역은 A 지역의 2배 자료이므로 $s_C = 2s_A = 20\sqrt{2}$, 평균은 60이다. 따라서 $CV_C = \frac{20\sqrt{2}}{60} = \frac{10\sqrt{2}}{30} = CV_A$ 이다.
4. D 지역은 C 지역에서 5를 뺀 자료이므로 $s_D = s_C = 20\sqrt{2}$, 평균은 55이다.
5. 따라서 $CV_B = \frac{10\sqrt{2}}{35}$, $CV_C = \frac{10\sqrt{2}}{30}$ 이므로 두 값은 같지 않다.

선지별 해설

① O: B 지역은 A 지역의 모든 값에 5분을 더한 자료이다. 같은 수를 더하는 평행이동은 산포를 바꾸지 않으므로 표준편차는 같다.
② O: C 지역은 A 지역의 모든 값을 2배한 자료이다. 평균과 표준편차가 모두 2배가 되므로 $CV = \frac{s}{\bar{x}}$ 값은 그대로여서 변동계수는 같다.
③ X: B 지역은 평균만 35로 커지고 표준편차는 A와 같은 반면, C 지역은 평균이 60이고 표준편차도 2배가 된다. 따라서 $CV_B = \frac{10\sqrt{2}}{35}$, $CV_C = \frac{10\sqrt{2}}{30}$ 이라서 같지 않다. 틀린 표현은 "같다"이고, 올바르게는 "B와 C 지역의 변동계수는 같지 않다. C 지역의 변동계수가 더 크다."이다.
④ O: D 지역은 C 지역의 모든 값에서 5분을 뺀 자료이다. 같은 수를 빼는 평행이동도 표준편차를 바꾸지 않으므로 C와 D의 표준편차는 같다.

오답 포인트

표준편차와 변동계수를 같은 방식으로 생각하면 틀리기 쉽다. 표준편차는 평행이동에 불변이지만, 변동계수는 분모인 평균이 함께 들어가므로 평행이동에는 변한다. 반대로 모든 값을 같은 배수로 곱하면 평균과 표준편차가 함께 같은 배수로 커져 변동계수는 유지된다.

다시 나오면 이렇게 판단

자료열이 다른 자료의 $+b$ 형태인지 $\times a$ 형태인지 먼저 본다. $+b$ 이면 표준편차만 비교하고, $\times a$ 이면 평균과 표준편차가 함께 변한다는 점을 이용해 변동계수까지 바로 판단한다.

암기 포인트

같은 수를 더하고 빼면 표준편차는 그대로이고, 같은 배수를 곱하면 평균과 표준편차가 함께 변한다. 변동계수는 배수 변화에는 그대로이고 평행이동에는 달라진다.

NoteBase

문 13. 다음은 자료 $(x_{1i}, x_{2i}, x_{3i}, y_i)$ $(i = 1, 2, \cdots, 38)$에 다중선형회귀모형(multiple linear regression model) $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$를 최소제곱법으로 적합하여 얻은 회귀분석 결과이다. 이에 대한 설명으로 옳지 않은 것은? (단, ε_i는 $N(0, \sigma^2)$을 따르고, 서로 독립이다)

요인	제곱합	자유도	평균제곱	$F-$값	$p-$값
회귀	4540.6	3	1513.5	22.8	0.000
잔차	2262.0	34	66.5		
전체	6802.6	37			

모형	비표준화 계수	$t-$값	$p-$값
상수	77.89	3.79	0.001
x_1	-19.33	-4.64	0.000
x_2	0.48	7.93	0.000
x_3	-0.11	-0.66	0.514

① 종속변수(Y)의 전체 변동 중에서 이 모형에 의하여 설명되는 변동이 차지하는 비율은 $\frac{4540.6}{6802.6}$이다.

② 유의수준 1%에서 이 모형은 유의하다.

③ 유의수준 5%에서 가설 $H_0 : \beta_3 = 0$ 대 $H_1 : \beta_3 \neq 0$을 검정할 때, H_0를 기각하지 못한다.

④ 추정된 회귀식은 $\hat{Y} = 77.89 - 19.33x_1 + 0.48x_2$이다.

정답 4번

출제 포인트

이 문제는 회귀분석 결과표에서 결정계수, 모형의 유의성, 개별 회귀계수의 유의성, 추정 회귀식의 형태를 정확히 읽는 문제이다.

자료 읽는 순서

1. 분산분석표에서 회귀제곱합과 전체제곱합을 확인해 설명력과 모형 전체 유의성을 본다.
2. 계수표에서 각 변수의 추정계수, t값, p값을 확인해 개별 계수의 유의성을 판단한다.
3. 유의하지 않은 변수가 있더라도 추정 회귀식에는 해당 추정계수가 포함되는지를 확인한다.

NoteBase

선지별 해설

1번: O. 종속변수 Y의 전체 변동 중 모형이 설명하는 비율은 결정계수 R^2이며, 분산분석표에서 $SSR/SST = 4540.6/6802.6$로 읽는다.
2번: O. 모형 전체에 대한 F검정의 p값이 0.000이므로 유의수준 1%에서 귀무가설을 기각한다. 따라서 이 회귀모형은 유의하다.
3번: O. x_3의 p값이 0.514이므로 유의수준 5%에서 $H_0 : \beta_3 = 0$를 기각하지 못한다. 따라서 3번 선지의 설명은 맞다.
4번: X. 계수표의 추정치는 상수 77.89, x_1의 계수 -19.33, x_2의 계수 0.48, x_3의 계수 -0.11이다. 따라서 추정된 회귀식은 $\hat{Y} = 77.89 - 19.33x_1 + 0.48x_2 - 0.11x_3$이다. 틀린 표현은 x_3 항을 빠뜨린 부분이며, 올바르게는 x_3 항까지 포함해야 한다.

오답 포인트

유의하지 않은 계수는 회귀식에서 자동으로 제거된다고 오해하면 4번을 맞는 선지로 착각한다. 유의성 검정 결과와 추정 회귀식의 표기 여부는 다른 문제이다.

다시 나오면 이렇게 판단

회귀표 문제는 먼저 분산분석표로 모형 전체의 설명력과 유의성을 확인하고, 다음으로 계수표에서 각 변수의 p값을 본다. 마지막에는 추정 회귀식을 묻는지, 유의한 변수만 남긴 축약모형을 묻는지 문장을 구분해 판단한다.

암기 포인트

유의하지 않은 계수도 추정 회귀식에는 남고, 제거 여부는 별도의 모형선택 문제이다.

NoteBase

문 14. 세 가지 다이어트 방법에 따라 체중 감량에 차이가 있는지를 알아보기 위해 15명의 사람을 무작위로 뽑아 5명씩 세 그룹으로 나눈 후 각 그룹에 서로 다른 다이어트 방법을 적용하여 실험을 진행하였다. 자료를 분석하여 다음과 같은 분산분석표를 얻었을 때, 이에 대한 해석으로 옳은 것은? (단, $F_\alpha(k_1, k_2)$는 자유도가 k_1, k_2인 F분포의 $(1-\alpha)$분위수를 나타내고, $F_{0.1}(2, 12) = 2.81$, $F_{0.05}(2, 12) = 3.89$, $F_{0.01}(2, 12) = 6.93$이다)

요인	제곱합	자유도	평균제곱	F-값
처리			150	
오차				
전체	540			

① 유의수준 10%에서 차이가 없다.

② 유의수준 10%에서는 차이가 있으나 유의수준 5%에서는 차이가 없다.

③ 유의수준 5%에서는 차이가 있으나 유의수준 1%에서는 차이가 없다.

④ 유의수준 1%에서 차이가 있다.

정답 4번

출제 포인트

이 문제는 일원분산분석표의 빈칸을 채워 F값을 계산한 뒤 임계값과 비교하여 유의수준별 집단 평균 차이 여부를 판단하는 문제이다.

조건 정리

세 집단이므로 처리 자유도는 2이다.
전체 표본 수가 15명이므로 전체 자유도는 14이고 오차 자유도는 12이다.
처리 평균제곱이 150이고 전체 제곱합이 540으로 주어져 있다.

사용할 공식

분산분석에서는 처리 제곱합 $SSTR = MSTR \times (k-1)$, 오차 제곱합 $SSE = SST - SSTR$, 오차 평균제곱 $MSE = SSE/(N-k)$, 검정통계량 $F = MSTR/MSE$를 사용한다.
이 문제는 처리 평균제곱과 전체 제곱합이 주어져 있으므로 먼저 $SSTR$와 SSE를 구한 뒤 F값을 계산해야 한다.

NoteBase

계산 과정

처리 자유도는 $k-1=3-1=2$이다.
처리 제곱합은 $SSTR=150\times 2=300$이다.
오차 제곱합은 $SSE=540-300=240$이다.
오차 자유도는 $15-3=12$이므로 오차 평균제곱은 $MSE=240/12=20$이다.
따라서 검정통계량은 $F=150/20=7.5$이다.
$7.5>F_{0.01}(2,12)=6.93$이므로 유의수준 1%에서도 귀무가설을 기각한다.

선지 판단

1번 X: 계산된 F값 7.5는 $F_{0.1}(2,12)=2.81$보다 크므로 유의수준 10%에서 차이가 없다가 아니라 차이가 있다가 맞다.
2번 X: 유의수준 10%에서만 차이가 있고 5%에서는 차이가 없다는 진술은 틀리다. $7.5>F_{0.05}(2,12)=3.89$이므로 유의수준 5%에서도 차이가 있다.
3번 X: 유의수준 5%에서는 차이가 있으나 1%에서는 차이가 없다는 진술도 틀리다. $7.5>F_{0.01}(2,12)=6.93$이므로 1%에서도 차이가 있다.
4번 O: F값이 1% 임계값 6.93보다 크므로 유의수준 1%에서 세 다이어트 방법의 평균 체중 감량에는 차이가 있다고 본다.

오답 포인트

유의수준이 작을수록 기준이 더 엄격하다는 점을 놓치면 3번에 흔들리기 쉽다. 1%에서 유의하면 5%와 10%에서도 모두 유의하므로, 가장 엄격한 1% 기준을 통과했는지만 확인하면 나머지 선지는 함께 정리된다.

다시 나오면 이렇게 판단

분산분석 선지는 먼저 자유도, 제곱합, 평균제곱으로 F값을 계산하고, 그다음 임계값과 큰지 작은지를 비교한다. 계산된 F값이 더 작은 유의수준의 임계값까지 넘으면 더 큰 유의수준에서도 모두 유의하다고 판단한다.

암기 포인트

일원분산분석은 $F=MSTR/MSE$이고, 1%에서 유의하면 5%와 10%에서도 유의이다.

NoteBase

문 15. 콘크리트 건조방법(A)과 혼합방법(B)이 압축강도에 영향을 미치는지를 알아보고자 한다. 건조방법은 3수준, 혼합방법은 4수준을 고려하고, 각 수준의 조합에서 3회 반복 실험하여 다음과 같은 분산분석표를 얻었을 때, 유의수준 5%에서 유의한 요인만을 모두 나열한 것은? (단, $A \times B$는 인자 A, 인자 B의 교호작용(interaction)을 나타내고, $F_\alpha(k_1, k_2)$는 자유도가 k_1, k_2인 F분포의 $(1-\alpha)$분위수를 나타내며, $F_{0.05}(2, 24) = 3.40, F_{0.05}(3, 24) = 3.01, F_{0.05}(6, 24) = 2.51$이다)

요인	제곱합	자유도	평균제곱	$F-$값
인자 A	8.0			
인자 B	8.7			
A×B	15.6			
오차	24.0			

① 인자 A, 인자 B

② 인자 A, A × B

③ 인자 B, A × B

④ 인자 A, 인자 B, A × B

정답 2번

출제 포인트

이 문제는 이원분산분석표에서 자유도, 평균제곱, F값을 계산해 유의한 주효과와 교호작용을 판정하는 문제이다.

조건 정리

주어진 값: 건조방법 A는 3수준, 혼합방법 B는 4수준, 각 조합에서 3회 반복이다.
구해야 할 값: 인자 A, 인자 B, A×B 중 유의수준 5%에서 유의한 요인이다.
주의할 조건: 자유도는 $df_A = 2, df_B = 3, df_{A\times B} = 6, df_E = 24$이고, 오차평균제곱을 먼저 구해야 한다.

사용할 공식

$MS = SS/df$, $F_A = MS_A/MS_E, F_B = MS_B/MS_E, F_{A\times B} = MS_{A\times B}/MS_E$이다.
이 공식을 쓰는 이유: 분산분석은 각 요인의 평균제곱을 오차평균제곱과 비교해 유의성을 판단하기 때문이다.

NoteBase

계산 과정

총 관측수는 $3 \times 4 \times 3 = 36$이므로 총 자유도는 35이고, 오차 자유도는 $3 \times 4 \times (3-1) = 24$이다.
오차평균제곱은 $MS_E = 24.0/24 = 1.0$이다.
인자 A는 $MS_A = 8.0/2 = 4.0$, $F_A = 4.0/1.0 = 4.0$이므로 $F_{0.05}(2, 24) = 3.40$보다 크다.
인자 B는 $MS_B = 8.7/3 = 2.9$, $F_B = 2.9/1.0 = 2.9$이므로 $F_{0.05}(3, 24) = 3.01$보다 작다.
A×B는 $MS_{A \times B} = 15.6/6 = 2.6$, $F_{A \times B} = 2.6/1.0 = 2.6$이므로 $F_{0.05}(6, 24) = 2.51$보다 크다.

선지 판단

1번 X: 인자 A는 유의하지만 A×B도 $2.6 > 2.51$로 유의하다. A×B를 누락한 점이 틀렸고, 올바르게는 인자 A와 A×B가 유의하다.
2번 O: 인자 A는 $4.0 > 3.40$, A×B는 $2.6 > 2.51$로 유의하고 인자 B는 $2.9 < 3.01$로 유의하지 않다. 유의한 요인만 고르면 인자 A와 A×B이다.
3번 X: A×B는 유의하지만 인자 B는 $2.9 < 3.01$로 유의하지 않다. 인자 B를 포함한 점이 틀렸고, 올바르게는 인자 A와 A×B가 유의하다.
4번 X: 인자 A와 A×B는 유의하지만 인자 B는 유의하지 않다. 인자 B까지 모두 포함한 점이 틀렸고, 올바르게는 인자 A와 A×B만 유의하다.

오답 포인트

교호작용 자유도 6과 오차평균제곱 1.0을 놓치면 A×B의 유의성을 빼먹기 쉽다. 또 인자 B의 F값 2.9가 3.01에 매우 가까워 부등호를 거꾸로 읽으면 3번이나 4번을 고르기 쉽다.

다시 나오면 이렇게 판단

이원분산분석 선지는 각 요인의 자유도 계산, 오차평균제곱 계산, 각 F값과 해당 임계값 비교의 순서로 판단하면 된다. 주효과와 교호작용은 각각 따로 비교해야 한다.

암기 포인트

이원분산분석은 MS_E를 먼저 구한 뒤 각 요인의 $F = MS/MS_E$를 해당 자유도의 임계값과 비교해 유의성을 정한다.

NoteBase

문 16. 다음은 어느 나라의 2019~2021년의 1인당 국민소득이다. 이 기간 동안 1인당 국민소득의 연평균 증가율은?

연도	2019	2020	2021
1인당 국민소득(달러)	10,000	15,000	30,000

① $100(\sqrt{3}-1)\%$

② $100\sqrt{3}\%$

③ 75%

④ 175%

정답 1번

출제 포인트

이 문제는 시작값과 최종값으로부터 복리 기준의 연평균 증가율을 구하는 문제이다.

조건 정리

주어진 값: 2019년 1인당 국민소득은 10,000달러, 2021년 1인당 국민소득은 30,000달러이다.
구해야 할 값: 2019년부터 2021년까지 2년간의 연평균 증가율이다.
주의할 조건: 연평균 증가율은 해마다 일정한 비율로 증가한다고 보는 복리 기준으로 계산한다.

사용할 공식

연평균 증가율은 ≤ft(최종값시작값)^1/n-1로 구한다.
이 공식을 쓰는 이유: 연도별 증가율을 단순 평균하는 값이 아니라, 시작값이 최종값이 되도록 만드는 일정한 연간 증가율을 묻기 때문이다.

계산 과정

1. 전체 증가배수는 $\frac{30000}{10000}=3$이다.
2. 2019년에서 2021년까지는 증가 구간이 2개이므로 $n=2$이다.
3. 연평균 증가율은 $3^{1/2}-1=\sqrt{3}-1$이다.
4. 퍼센트로 바꾸면 $100(\sqrt{3}-1)\%$이다.

NoteBase

선지 판단

1번 O: 계산 결과가 $100(\sqrt{3}-1)\%$이므로 맞다.
2번 X: $100\sqrt{3}\%$는 연평균 증가배수 $\sqrt{3}$에 100을 곱한 값일 뿐, 증가율 계산에서 필요한 -1을 빠뜨렸다. 틀린 표현은 $100\sqrt{3}\%$이고, 올바르게는 $100(\sqrt{3}-1)\%$이다.
3번 X: 75%는 2019→2020의 50%와 2020→2021의 100%를 단순 평균한 값이다. 연평균 증가율은 단순평균이 아니라 복리평균이므로 틀리다. 틀린 값은 75%이고, 올바르게는 $100(\sqrt{3}-1)\%$이다.
4번 X: 175%는 연평균 증가율 값이 아니다. 시작값이 3배가 되었다는 사실을 잘못 환산한 값으로 볼 수 있으며, 연평균 증가율은 반드시 복리식으로 구해야 한다. 틀린 값은 175%이고, 올바르게는 $100(\sqrt{3}-1)\%$이다.

오답 포인트

연평균 증가율을 물으면 해마다의 증가율을 단순 평균하면 안 된다. 특히 3번처럼 50%와 100%를 평균해 75%로 처리하는 실수가 대표적이고, 2번처럼 증가배수에서 -1을 빼지 않는 실수도 자주 나온다.

다시 나오면 이렇게 판단

연평균 증가율 문제는 먼저 시작값과 최종값으로 전체 배수를 만든 뒤, 기간 수만큼 거듭제곱근을 취하고 마지막에 1을 빼는 순서로 판단한다.

암기 포인트

연평균 증가율은 단순평균이 아니라 ≤ft(최종값시작값)^1/n-1의 복리 개념이다.

NoteBase

문 17. A, B 지역에서 각각 “갑, 을, 병, 정 네 후보의 지지도는 동일하다.”라는 귀무가설을 검정하기 위해 A 지역에서 110명, B 지역에서 1,100명을 무작위로 뽑아 조사한 지지자 수가 다음과 같을 때, 이에 대한 설명으로 옳은 것만을 모두 고르면? (단, χ_A^2, χ_B^2는 각각 A, B 지역에서의 카이제곱 통계량 값이고, p_A, p_B는 각각 A, B 지역에서의 p-값(유의확률)이다)

지역 \ 후보	갑	을	병	정
A	30	40	25	15
B	300	400	200	200

ㄱ. 위의 내용은 동질성검정에 해당한다.
ㄴ. $\frac{\chi_B^2}{\chi_A^2} < 10$
ㄷ. $\frac{p_B}{p_A} < 1$

① ㄱ, ㄴ
② ㄱ, ㄷ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 3번

출제 포인트

이 문제는 카이제곱 적합도검정의 유형을 구분하고, 각 지역의 카이제곱 통계량과 p값의 크기를 비교할 수 있는지를 묻는다.

조건 정리

A지역 표본수는 110명, B지역 표본수는 1,100명이다.
귀무가설은 각 지역에서 갑, 을, 병, 정 네 후보의 지지도가 동일하다는 것이다.
따라서 기대도수는 A지역에서 각 27.5명, B지역에서 각 275명이다.

사용할 공식

$X^2 = \sum \frac{(O-E)^2}{E}$
이 공식을 쓰는 이유는 각 지역에서 관측도수와 '네 후보 지지도가 동일하다'는 귀무가설 아래의 기대도수를 비교하는 적합도검정 상황이기 때문이다.

NoteBase

계산 과정

A지역은 $E = 27.5$이므로 $X_A^2 = \frac{(30-27.5)^2}{27.5} + \frac{(40-27.5)^2}{27.5} + \frac{(25-27.5)^2}{27.5} + \frac{(15-27.5)^2}{27.5} = 11.818\ldots$이다.
B지역은 $E = 275$이므로 $X_B^2 = \frac{(300-275)^2}{275} + \frac{(400-275)^2}{275} + \frac{(200-275)^2}{275} + \frac{(200-275)^2}{275} = 100$이다.
따라서 $\frac{X_B^2}{X_A^2} = \frac{100}{11.818\ldots} \approx 8.46 < 10$이므로 ㄴ은 참이다.
두 검정의 자유도는 모두 $4-1=3$이므로 카이제곱 통계량이 더 큰 B지역의 p값이 더 작아 $p_B < p_A$가 되어 ㄷ도 참이다.

선지 판단

① X: 이 조합은 거짓인 ㄱ을 포함하고 참인 ㄷ을 빠뜨려 틀리다. 틀린 부분은 ㄱ이고, 올바르게는 위 내용이 동질성검정이 아니라 각 지역별 적합도검정이라는 점이다.
② X: 이 조합도 거짓인 ㄱ을 포함하므로 틀리다. 올바른 조합은 ㄴ, ㄷ이다.
③ O: ㄴ은 $\frac{X_B^2}{X_A^2} \approx 8.46 < 10$이므로 참이고, ㄷ은 자유도가 같을 때 $X_B^2 > X_A^2$이면 $p_B < p_A$이므로 참이다. 따라서 옳은 조합은 ㄴ, ㄷ이다.
④ X: 이 조합은 ㄱ을 포함한 것이 틀린 부분이다. 동질성검정이라면 A와 B의 지지 분포가 서로 같은지를 직접 비교해야 하지만, 여기서는 각 지역 안에서 네 후보 지지도가 동일한지를 따로 검정한다.

오답 포인트

표본이 두 개라는 이유만으로 동질성검정이라고 판단하면 틀린다. 이 문항은 A와 B의 분포를 서로 비교하는 문제가 아니라, 각 지역에서 균등분포인지 아닌지를 각각 검정하는 적합도검정이라는 점이 함정이다.

다시 나오면 이렇게 판단

카이제곱 문제는 먼저 한 표본을 이론분포와 비교하는지, 여러 모집단의 분포를 서로 비교하는지부터 구분한다. 그다음 기대도수를 구해 X^2를 계산하고, 자유도가 같으면 통계량이 클수록 p값이 더 작다고 판단한다.

암기 포인트

이론분포 대 관측도수는 적합도검정이고, 모집단 간 분포 비교는 동질성검정이다. 자유도가 같으면 X^2가 클수록 p값은 작다.

NoteBase

문 18. 실내온도가 생산량에 영향을 미치는지를 알아보기 위하여 5가지의 실내온도에서 6회씩 실험을 실시하여 총 30개의 측정값을 얻었다. 일원배치 분산분석을 수행한 결과 처리평균제곱의 값이 250이고, 오차제곱합이 처리제곱합의 1.5배일 때, "실내온도는 생산량에 영향을 미치지 않는다."라는 귀무가설을 검정하기 위한 F-검정통계량의 값은?

① $\frac{23}{6}$
② $\frac{25}{6}$
③ $\frac{23}{5}$
④ $\frac{26}{5}$

정답 2번

출제 포인트

이 문제는 일원배치 분산분석에서 처리평균제곱과 오차제곱합의 관계를 이용해 F 검정통계량을 구하는 문제이다.

조건 정리

주어진 값은 처리수준 수 5개, 각 수준 반복수 6회, 총 관측치 수 30개, 처리평균제곱 $MST = 250$, 오차제곱합 SSE는 처리제곱합 $SSTR$의 1.5배이다.
구해야 할 값은 귀무가설 검정을 위한 F 통계량이다.
주의할 조건은 처리평균제곱이 이미 평균제곱이므로 먼저 자유도 4를 곱해 처리제곱합으로 되돌려야 한다는 점이다.

사용할 공식

일원배치 분산분석의 검정통계량은 $F = \frac{MST}{MSE}$이다.
또한 $MST = \frac{SSTR}{a-1}$, $MSE = \frac{SSE}{N-a}$이다.
이 공식을 쓰는 이유는 문제에서 처리효과 유무를 검정하는 일원배치 분산분석 상황을 직접 제시했고, 처리평균제곱과 오차제곱합이 함께 주어졌기 때문이다.

계산 과정

1. 처리수준 수가 5개이므로 처리 자유도는 $a - 1 = 4$이다.
2. $MST = 250$이므로 $SSTR = 250 \times 4 = 1000$이다.
3. 오차제곱합은 처리제곱합의 1.5배이므로 $SSE = 1.5 \times 1000 = 1500$이다.
4. 총 관측치 수가 30개이므로 오차 자유도는 $N - a = 30 - 5 = 25$이다.
5. 따라서 $MSE = \frac{1500}{25} = 60$이다.
6. 최종적으로 $F = \frac{250}{60} = \frac{25}{6}$이다.

NoteBase

선지별 해설

① X. F 값은 $\frac{23}{6}$이 아니라 $\frac{25}{6}$이다. 처리제곱합을 1000, 오차평균제곱을 60으로 계산하면 반드시 $\frac{25}{6}$이 나오므로 주어진 조건과 맞지 않는다.
② O. $MST = 250$, $MSE = 60$이므로 $F = \frac{250}{60} = \frac{25}{6}$이다. 계산 결과와 정확히 일치한다.
③ X. $\frac{23}{5}$는 주어진 자유도와 제곱합 관계로는 나오지 않는 값이다. 오차 자유도를 25로 두고 MSE를 계산하면 60이므로 F 값은 $\frac{23}{5}$가 아니라 $\frac{25}{6}$이어야 한다.
④ X. $\frac{26}{5}$ 역시 주어진 조건에서 계산되지 않는 값이다. 처리평균제곱을 그대로 분자에 두고 오차평균제곱을 정확히 구하면 F 통계량은 $\frac{26}{5}$가 아니라 $\frac{25}{6}$이다.

오답 포인트

처리평균제곱 MST와 처리제곱합 $SSTR$을 혼동하면 처음부터 값이 어긋난다. 또 오차제곱합 SSE를 구한 뒤 다시 오차 자유도 25로 나누어 MSE를 만들어야 하는데, 이 단계를 빼먹으면 보기의 오답처럼 잘못된 분수가 나오기 쉽다.

다시 나오면 이렇게 판단

분산분석 계산형 문제는 평균제곱인지 제곱합인지부터 구분한 뒤, 자유도로 제곱합과 평균제곱을 서로 변환하고 마지막에 $F = \frac{MS\ \text{between}}{MS\ \text{within}}$ 형태로 정리하면 된다.

암기 포인트

일원배치 분산분석은 $MST = \frac{SSTR}{a-1}$, $MSE = \frac{SSE}{N-a}$, $F = \frac{MST}{MSE}$의 순서로 연결된다.

NoteBase

문 19. 확률변수 X와 Y의 표준화된 확률변수를 각각 W와 Z라 하자. 이에 대한 설명으로 옳은 것만을 모두 고르면?

ㄱ. $\mathrm{Cov}(Z, W) = \mathrm{Corr}(X, Y)$
ㄴ. $\mathrm{Corr}(W, Z) = \mathrm{Cov}(W, Z)$
ㄷ. $\mathrm{Cov}(W, Z) = -1$이면 $P(W + Z = 0) = 1$이다.

① ㄱ, ㄴ
② ㄱ, ㄷ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 4번

출제 포인트

이 문제는 표준화한 확률변수의 분산, 공분산, 상관계수의 관계와 상관계수 -1의 의미를 묻는다.

핵심 개념

표준화 변수 $W = \frac{X - E(X)}{\sigma_X}, Z = \frac{Y - E(Y)}{\sigma_Y}$ 는 평균이 0이고 분산이 1이다.
따라서 $\mathrm{Cov}(W, Z) = \mathrm{Corr}(W, Z)$ 이고, 이는 $\mathrm{Corr}(X, Y)$ 와도 같다.
또한 상관계수가 -1이면 두 변수는 완전한 음의 선형관계에 있으므로 표준화 변수에서는 $Z = -W$ 이다.

선지별 해설

1번: X. ㄱ은 $\mathrm{Cov}(Z, W) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} = \mathrm{Corr}(X, Y)$ 이므로 옳고, ㄴ도 표준화 변수의 표준편차가 1이어서 옳다. 다만 ㄷ도 옳으므로 1번의 조합은 불완전하다. 틀린 표현은 'ㄱ, ㄴ'만 고른 부분이고, 올바르게는 'ㄱ, ㄴ, ㄷ'이다.
2번: X. ㄱ은 옳고, ㄷ도 $\mathrm{Cov}(W, Z) = -1 \Rightarrow \mathrm{Corr}(W, Z) = -1$ 이므로 옳다. 그러나 ㄴ도 함께 옳다. 틀린 표현은 'ㄱ, ㄷ'만 고른 부분이고, 올바르게는 'ㄱ, ㄴ, ㄷ'이다.
3번: X. ㄴ과 ㄷ은 모두 옳다. 그러나 ㄱ도 표준화 공분산의 정의에서 옳다. 틀린 표현은 'ㄴ, ㄷ'만 고른 부분이고, 올바르게는 'ㄱ, ㄴ, ㄷ'이다.
4번: O. ㄱ은 $\mathrm{Cov}(Z, W) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} = \mathrm{Corr}(X, Y)$ 이라서 옳다. ㄴ은 표준화 변수의 분산이 1이므로 $\mathrm{Corr}(W, Z) = \frac{\mathrm{Cov}(W, Z)}{1 \cdot 1} = \mathrm{Cov}(W, Z)$ 이라서 옳다. ㄷ은 공분산이 -1이면 상관계수도 -1이고, 표준화 변수에서는 $Z = -W$ 가 성립하므로 $P(W + Z = 0) = 1$ 이라서 옳다.

NoteBase

오답 포인트

표준화하면 공분산과 상관계수가 같아진다는 사실은 쉽게 떠올리지만, 상관계수 -1이 단순히 '매우 작은 값'이 아니라 정확한 완전 음의 선형관계라는 점을 놓치기 쉽다. 또한 조합형 보기에서는 ㄱ, ㄴ만 확인하고 ㄷ을 끝까지 검토하지 않으면 1번으로 오판하기 쉽다.

다시 나오면 이렇게 판단

표준화 문제가 나오면 먼저 평균 0, 분산 1을 적고 시작한다. 그다음 공분산을 상관계수 형태로 바꿔 보고, 상관계수가 ± 1 이면 두 변수가 완전한 선형관계라는 결론까지 연결한다. 조합형 보기에서는 각 진술의 O/X를 먼저 확정한 뒤 마지막에 보기 번호를 고른다.

암기 포인트

표준화하면 분산은 1이고 공분산은 상관계수와 같으며, 상관계수가 -1이면 표준화 변수는 $Z = -W$ 이다.

NoteBase

문 20. 주머니 안에 흰 공 1개, 검은 공 2개가 들어 있다. 공을 한 개씩 무작위로 꺼낼 때 꺼낸 공이 검은 공이면 그 공을 흰 공으로 교체하여 주머니 속에 넣는다. 처음으로 흰 공이 나올 때까지 꺼낸 횟수를 확률변수 X라고 할 때, $E(9X)$의 값은?

① 13

② 16

③ 17

④ 20

정답 3번

출제 포인트

검은 공을 뽑을 때마다 그 공이 흰 공으로 바뀐다는 조건에서 확률변수 X의 분포를 구하고, 선형성 $E(9X) = 9E(X)$를 적용하는 문제이다.

조건 정리

주머니의 시작 상태는 흰 공 1개, 검은 공 2개이다.
검은 공을 뽑으면 그 검은 공을 흰 공으로 바꾸어 다시 넣는다.
처음으로 흰 공이 나올 때까지의 횟수가 X이므로 가능한 값은 1,2,3뿐이다.

사용할 공식

기댓값 공식은 $E(X) = \sum xP(X = x)$이다.
또한 상수배의 기댓값은 $E(9X) = 9E(X)$이다.
이 공식을 쓰는 이유는 이 문제의 정답이 확률변수 X의 평균적인 시행 횟수를 수치로 묻고 있기 때문이다.

계산 과정

처음에 흰 공을 바로 뽑을 확률은 $\frac{1}{3}$이므로 $P(X = 1) = \frac{1}{3}$이다.
처음에 검은 공을 뽑으면 상태가 흰 공 2개, 검은 공 1개로 바뀌므로 $P(X = 2) = \frac{2}{3} \cdot \frac{2}{3} = \frac{4}{9}$이다.
두 번 연속 검은 공을 뽑아야 $X = 3$이므로 $P(X = 3) = \frac{2}{3} \cdot \frac{1}{3} = \frac{2}{9}$이다.
$E(X) = 1 \cdot \frac{1}{3} + 2 \cdot \frac{4}{9} + 3 \cdot \frac{2}{9} = \frac{17}{9}$
따라서
$E(9X) = 9E(X) = 9 \cdot \frac{17}{9} = 17$

NoteBase

선지별 해설

① X

17이 아니므로 틀리다. 실제 계산값은 $E(9X) = 17$이다. 틀린 부분은 기댓값 수치이다. 올바르게는 13이 아니라 17이다.

② X

17이 아니므로 틀리다. 실제 계산값은 $E(9X) = 17$이다. 틀린 부분은 기댓값 수치이다. 올바르게는 16이 아니라 17이다.

③ O

$P(X=1)=\frac{1}{3}$, $P(X=2)=\frac{4}{9}$, $P(X=3)=\frac{2}{9}$를 반영하면 $E(X)=\frac{17}{9}$이고, 따라서 $E(9X)=17$이다. 계산 결과와 일치하므로 옳다.

④ X

17이 아니므로 틀리다. 실제 계산값은 $E(9X) = 17$이다. 틀린 부분은 기댓값 수치이다. 올바르게는 20이 아니라 17이다.

오답 포인트

검은 공을 뽑은 뒤 주머니 구성이 바뀐다는 점을 놓치면 매 시행의 확률을 고정된 값으로 처리하게 된다. 특히 $X=3$까지 가능하다는 점을 빼먹거나, $P(X=2)$를 $\frac{2}{3}\cdot\frac{1}{3}$으로 잘못 두는 실수가 자주 나온다.

다시 나오면 이렇게 판단

상태가 바뀌는 확률과정 문제는 가능한 X의 값부터 모두 적고, 각 값이 나오기까지의 경로확률을 순서대로 곱해 분포를 완성한 뒤 $E(X)=\sum xP(X=x)$로 정리하면 된다.

암기 포인트

상태가 변하는 기댓값 문제는 가능한 값 나열, 경로확률 계산, $\sum xP(X=x)$ 적용의 순서로 정리된다.

NoteBase

문 21. 자료 (x_i, y_i) $(i = 1, 2, \cdots, 7)$에 단순선형회귀모형 $Y_i = a + bx_i + \varepsilon_i$를 최소제곱법으로 적합한 결과, 회귀제곱합은 5, 오차(잔차)제곱합은 5이었다. $x_i^* = 5 + \frac{x_i}{10}$, $y_i^* = 5 + 10y_i$로 변환하여 단순선형회귀모형 $Y_i^* = c + dx_i^* + \varepsilon_i^*$를 최소제곱법으로 적합할 때, 이에 대한 설명으로 옳은 것만을 모두 고르면? (단, $\varepsilon_i, \varepsilon_i^*$는 평균 0, 분산이 각각 σ^2, $(\sigma^*)^2$인 정규분포를 따르는 서로 독립인 확률변수이고, $t_\alpha(k)$는 자유도가 k인 t 분포의 $(1-\alpha)$분위수를 나타내며, $t_{0.05}(5) = 2.02$, $t_{0.025}(5) = 2.57$이다)

ㄱ. 총제곱합은 100이다.
ㄴ. 결정계수는 0.5이다.
ㄷ. 가설 $H_0 : d = 0$ 대 $H_1 : d \neq 0$을 검정할 때, 유의수준 5%에서 귀무가설(H_0)을 기각하지 못한다.

① ㄴ
② ㄱ, ㄷ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 3번

출제 포인트

이 문제는 반응변수와 설명변수의 선형변환이 제곱합, 결정계수, 기울기 검정통계량에 어떤 영향을 주는지를 묻는 문제이다.

조건 정리

주어진 값: 원래 모형의 회귀제곱합은 5, 오차제곱합은 5, 표본크기는 7이다.
구해야 할 값: 변환된 모형의 총제곱합, 결정계수, $H_0 : d = 0$의 기각 여부이다.
주의할 조건: $y_i^* = 5 + 10y_i$는 평행이동과 10배 확대이므로 제곱합을 100배로 바꾸고, $x_i^* = 5 + \frac{x_i}{10}$는 기울기의 크기만 바꾸므로 결정계수와 기울기의 t검정값은 그대로 유지된다.

사용할 공식

총제곱합은 $SST = SSR + SSE$, 결정계수는 $R^2 = \frac{SSR}{SST}$이다.
단순회귀에서 기울기 검정은 $t^2 = \frac{SSR}{SSE/(n-2)}$를 이용할 수 있다.
이 공식을 쓰는 이유: 이 문항은 변환 전의 SSR, SSE, n이 주어져 있으므로 변환 후 제곱합의 배수와 원래 모형의 R^2, t값을 바로 연결할 수 있기 때문이다.

NoteBase

계산 과정

1. 원래 모형의 총제곱합은 $SST = SSR + SSE = 5 + 5 = 10$이고, 결정계수는 $R^2 = \frac{5}{10} = 0.5$이다.
2. $y_i^* = 5 + 10y_i$이므로 평균에서의 편차는 10배가 된다. 따라서 $SSR^* = 100 \times 5 = 500$, $SSE^* = 100 \times 5 = 500$, $SST^* = 500 + 500 = 1000$이다.
3. 기울기 검정은 변환 전후에 동일하다. 원래 모형에서 $t^2 = \frac{5}{5/(7-2)} = 5$이므로 $t = \sqrt{5} \approx 2.236$이다. 이는 $t_{0.025}(5) = 2.57$보다 작으므로 유의수준 5%에서 $H_0 : d = 0$을 기각하지 못한다.

선지별 해설

① X
이 선택지는 ㄴ만 옳다고 본다. 그러나 $R^2 = 0.5$이므로 ㄴ은 옳고, $t = \sqrt{5} \approx 2.236 < 2.57$이므로 ㄷ도 옳다. 틀린 표현은 "ㄴ"만 고른 부분이며, 올바르게는 "ㄴ, ㄷ"이다.

② X
이 선택지는 ㄱ, ㄷ을 옳다고 본다. ㄷ은 옳지만 ㄱ의 "총제곱합은 100이다"는 틀리다. y를 10배 하면 제곱합은 100배가 되므로 총제곱합은 $10 \times 100 = 1000$이다. 또한 ㄴ도 옳다. 올바르게는 "ㄴ, ㄷ"이다.

③ O
이 선택지는 ㄴ, ㄷ을 고른다. $R^2 = \frac{SSR}{SST} = \frac{5}{10} = 0.5$이므로 ㄴ은 옳다. 또 $t = \sqrt{5} \approx 2.236 < 2.57$이므로 유의수준 5%에서 $H_0 : d = 0$을 기각하지 못하므로 ㄷ도 옳다.

④ X
이 선택지는 ㄱ, ㄴ, ㄷ이 모두 옳다고 본다. 그러나 ㄱ의 "총제곱합은 100이다"가 틀리다. 올바른 값은 1000이다. ㄴ과 ㄷ만 옳으므로 올바르게는 "ㄴ, ㄷ"이다.

오답 포인트

평행이동과 배수변환의 효과를 섞어 보는 함정이 있다. y에 상수를 더하는 것은 제곱합에 영향을 주지 않지만, y를 10배 하는 것은 제곱합을 100배로 만든다. 반면 결정계수와 기울기 검정의 t값은 선형변환 뒤에도 바뀌지 않는다는 점을 놓치기 쉽다.

다시 나오면 이렇게 판단

회귀식 변환 문제가 나오면 먼저 y변환이 제곱합에 주는 배수를 계산하고, 그다음 R^2와 기울기 t검정이 불변인지 확인한 뒤 각 진술의 참거짓을 판정하면 된다.

암기 포인트

y를 k배 하면 제곱합은 k^2배가 되고, R^2와 단순회귀 기울기 검정의 t값은 선형변환으로 변하지 않는다.

NoteBase

문 22. 확률변수 X의 누적분포함수 $F(x)$가 다음과 같을 때, 이에 대한 설명으로 옳은 것만을 모두 고르면?

F(x)=cases0, & x ≤ 0\^2, & 0 < x < (1)/(2)\\1, & x ≥ (1)/(2)cases
ㄱ. $F(x)$는 $x = 0,\ \frac{1}{2}$에서 불연속이다.
ㄴ. $P(X = \frac{1}{2}) = 0$이다.
ㄷ. 구간 $0 < x < \frac{1}{2}$에서 X의 확률밀도함수를 $f(x)$라 할 때,
$f(x) = 2x$이다.

① ㄴ

② ㄷ

③ ㄴ, ㄷ

④ ㄱ, ㄴ, ㄷ

정답 2번

출제 포인트

이 문제는 누적분포함수의 연속성, 점질량 $P(X = a) = F(a) - F(a -)$, 그리고 열린구간에서의 밀도함수 $f(x) = F'(x)$를 함께 판정하는 문제이다.

조건 정리

주어진 누적분포함수는 다음과 같다.
F(x)=cases 0, & x≤ 0 ; x^2, & 0<x<(1)/(2) ; 1, & x≥ (1)/(2) cases
따라서 $x = 0$과 $x = \frac{1}{2}$에서는 함수값과 극한을 비교해야 하고, 열린구간 $0 < x < \frac{1}{2}$에서는 미분으로 밀도함수를 판단한다.

사용할 공식

연속성은 함수값과 좌우극한이 같은지로 판단한다.
한 점의 확률은 $P(X = a) = F(a) - F(a -)$로 구한다.
밀도함수는 절대연속인 구간에서 $f(x) = F'(x)$이다.
이 공식을 쓰는 이유는 선지가 각각 불연속점, 한 점의 확률, 열린구간의 밀도를 직접 묻고 있기 때문이다.

계산 과정

1. $x = 0$에서는 $F(0) = 0$, $\lim_{x\to 0-} F(x)=0$, $\lim_{x\to 0+} F(x)= \lim_{x\to 0+} x^2 = 0$이므로 연속이다.
2. $x = \frac{1}{2}$에서는 $F(\frac{1}{2}) = 1$, $\lim_{x\to (1/2)-} F(x)=(\frac{1}{2})^2 = \frac{1}{4}$이므로 점프 크기는 $1 - \frac{1}{4} = \frac{3}{4}$이고 불연속이다.
3. 따라서 $P(X = \frac{1}{2}) = F(\frac{1}{2}) - F(\frac{1}{2} -) = 1 - \frac{1}{4} = \frac{3}{4}$이다.
4. 열린구간 $0 < x < \frac{1}{2}$에서는 $F(x) = x^2$이므로 $f(x) = F'(x) = 2x$이다.

NoteBase

선지 판단

1번 X: 1번은 ㄴ만 옳다고 본 선지이다. 그런데 ㄴ은 $P\left(X=\frac{1}{2}\right)=0$이라고 했지만 실제로는 점프확률이 $\frac{3}{4}$이므로 틀리다. 바른 서술은 'ㄴ. $P\left(X=\frac{1}{2}\right)=\frac{3}{4}$이다.'이다.
2번 O: 2번은 ㄷ만 옳다고 본 선지이다. $0<x<\frac{1}{2}$에서 $F'(x)=2x$이므로 ㄷ은 옳다. 또한 ㄱ, ㄴ은 틀리므로 ㄷ만 고른 2번이 정답이다.
3번 X: 3번은 ㄴ, ㄷ이 옳다고 본 선지이다. ㄷ은 옳지만 ㄴ이 틀려서 전체 선지는 틀리다. 바른 조합은 'ㄷ만 옳다'이다.
4번 X: 4번은 ㄱ, ㄴ, ㄷ이 모두 옳다고 본 선지이다. ㄱ은 $x=0$에서 연속이라는 점을 놓쳐 틀리고, ㄴ도 점질량 $\frac{3}{4}$를 0으로 본 점에서 틀리다. 바른 조합은 'ㄷ만 옳다'이다.

오답 포인트

ㄱ의 '$x=0,\frac{1}{2}$에서 불연속'은 경계점마다 모두 점프가 있다고 착각하게 만들고, ㄴ의 '$P(X=\frac{1}{2})=0$'은 한 점의 확률을 밀도값처럼 0으로 처리하게 만든다. 누적분포함수 문제에서는 함수식만 보지 말고 경계점의 좌극한과 함수값 차이를 먼저 봐야 한다.

다시 나오면 이렇게 판단

누적분포함수 문제는 경계점마다 $F(a)$, $F(a-)$, $F(a+)$를 먼저 비교하고, 점프가 있으면 그 차이를 $P(X=a)$로 읽는다. 그다음 열린구간에서만 미분하여 밀도함수를 판단하면 된다.

암기 포인트

누적분포함수는 '경계점은 극한 비교, 점프는 한 점의 확률, 열린구간은 미분'으로 정리된다.

NoteBase

문 23. $X_1, \cdots, X_8$과 $Y_1, \cdots, Y_8$은 각각 두 정규모집단 $N(2, \sigma^2)$과 $N(1, \sigma^2)$에서 추출한 서로 독립인 확률표본이고, $\overline{X}$, $\overline{Y}$, S_p가 다음과 같을 때, $P(\overline{X} - \overline{Y} \geq 1 + c \times S_p) = 0.1$을 만족하는 상수 c의 값은? (단, $t_\alpha(k)$는 자유도가 k인 t분포의 $(1-\alpha)$분위수이다)

$$\overline{X} = \frac{1}{8}\sum_{i=1}^{8} X_i$$

$$\overline{Y} = \frac{1}{8}\sum_{i=1}^{8} Y_i$$

$$S_p = \sqrt{\frac{1}{14}\left(\sum_{i=1}^{8}(X_i - \overline{X})^2 + \sum_{i=1}^{8}(Y_i - \overline{Y})^2\right)}$$

① $\frac{1}{4}t_{0.1}(14)$

② $\frac{1}{3}t_{0.1}(14)$

③ $\frac{1}{2}t_{0.1}(14)$

④ $t_{0.1}(14)$

정답 3번

출제 포인트

이 문제는 두 독립정규표본의 공통분산이 미지인 경우의 pooled two-sample t 통계량을 세워 확률조건을 임계값 식으로 바꾸는 문제이다.

조건 정리

표본크기는 두 집단 모두 8이다.

E(X- Y)=2-1=1이다.

공통 분산 σ^2는 미지이므로 문제에서 정의한 풀드표준편차 S_p를 사용한다.

사용할 공식

두 집단의 분산이 같고 미지일 때의 통계량은

T=((X- Y)-(μ_X-μ_Y))/(S_p√(1/8+1/8))

이고 자유도는 $8 + 8 - 2 = 14$이다.

이 문제는 $\mu_X - \mu_Y = 2 - 1 = 1$이고 공통분산을 S_p로 추정하므로 이 공식을 사용한다.

NoteBase

계산 과정

$\sqrt{1/8 + 1/8} = \sqrt{1/4} = 1/2$
이므로
T=((X- Y)-1)/(S_p/2)=2\(X- Y)-1\S_p t(14)
이다.
주어진 조건
P(X- Y≥ 1+cS_p)=0.1
은
P≤ft(2\(X- Y)-1\S_p≥ 2c)=0.1
즉
$P(T \geq 2c) = 0.1$
과 같다.
따라서 $t_{0.1}(14)$의 정의에 의해 $2c = t_{0.1}(14)$이고,
$c = \frac{1}{2}t_{0.1}(14)$
이다.

선지 판단

① X: $c = \frac{1}{4}t_{0.1}(14)$이면 $2c = \frac{1}{2}t_{0.1}(14)$가 되어 상위 10% 임계값보다 절반으로 줄어든다. 틀린 표현은 계수 $\frac{1}{4}$이고, 올바르게는 $c = \frac{1}{2}t_{0.1}(14)$이다.
② X: $\sqrt{1/8 + 1/8}$를 잘못 계산한 값이다. 틀린 표현은 계수 $\frac{1}{3}$이고, 올바르게는 $\sqrt{1/4} = 1/2$이므로 $c = \frac{1}{2}t_{0.1}(14)$이다.
③ O: T=2\(X- Y)-1\/S_p t(14)이므로 $P(T \geq 2c) = 0.1$에서 $2c = t_{0.1}(14)$가 된다. 따라서 $c = \frac{1}{2}t_{0.1}(14)$가 맞다.
④ X: 사건 (X- Y)-1≥ cS_p를 그대로 $T \geq c$로 두어 표준화 계수 2를 빠뜨린 값이다. 틀린 표현은 계수 1이고, 올바르게는 $2c = t_{0.1}(14)$이므로 $c = \frac{1}{2}t_{0.1}(14)$이다.

오답 포인트

이 문제의 함정은 $S_p\sqrt{1/n_1 + 1/n_2}$에서 $\sqrt{1/8 + 1/8} = 1/2$를 놓치거나, 부등식을 t 통계량으로 바꿀 때 앞의 2를 빠뜨리는 데 있다.

다시 나오면 이렇게 판단

두 독립정규표본에서 공통분산 미지 조건이 보이면 먼저 pooled two-sample t 통계량을 세우고, 그다음 확률사건을 그 통계량의 부등식으로 바꾸어 임계값 $t_\alpha(\nu)$와 직접 연결하면 된다.

암기 포인트

X- Y의 pooled t 통계량 분모는 $S_p\sqrt{1/n_1 + 1/n_2}$이고, $n_1 = n_2 = 8$이면 분모가 $S_p/2$가 된다.

NoteBase

문 24. 어떤 식품이 혈중 콜레스테롤의 양에 영향을 주는지 알아보기 위해 10명의 피실험자를 무작위로 추출하여 각 피실험자에 대해 식품을 섭취한 경우와 섭취하지 않은 경우의 혈중 콜레스테롤의 양을 측정하였다. "식품을 섭취한 경우와 섭취하지 않은 경우의 혈중 콜레스테롤의 양에 차이가 있다."라는 대립가설을 검정하기 위한 t-검정통계량의 값이 -1.9일 때, 이에 대한 설명으로 옳은 것은? (단, $t_\alpha(k)$는 자유도가 k인 t분포의 $(1-\alpha)$분위수를 나타내고, $t_{0.05}(9)=1.83$, $t_{0.025}(9)=2.26$, $t_{0.05}(18)=1.73$, $t_{0.025}(18)=2.10$ 이다)

① 대응(쌍체)비교를 실시하며 유의수준 5%에서 차이가 없다.

② 대응(쌍체)비교를 실시하며 유의수준 5%에서 차이가 있다.

③ 독립인 표본에 의한 두 집단 비교를 실시하며 유의수준 5%에서 차이가 없다.

④ 독립인 표본에 의한 두 집단 비교를 실시하며 유의수준 5%에서 차이가 있다.

정답 1번

출제 포인트

이 문제는 같은 10명에게 섭취 여부 두 조건을 모두 측정한 상황에서 대응표본 t-검정을 선택하고, 양측검정의 기각 여부를 판단하는 문제이다.

조건 정리

- 피실험자 10명 각각에 대해 섭취한 경우와 섭취하지 않은 경우를 모두 측정하였다.
- 같은 사람을 두 번 측정했으므로 두 표본은 독립이 아니라 대응되어 있다.
- 대립가설은 "차이가 있다"이므로 양측검정이다.
- 검정통계량은 $t=-1.9$이고, 대응표본 t-검정의 자유도는 $n-1=9$이다.

사용할 공식

대응표본 t-검정에서는 각 사람의 차이값 d_i를 만들고, $t=\frac{\bar{d}}{s_d/\sqrt{n}}$를 사용한다.
이 공식을 쓰는 이유는 두 관측값이 같은 사람에게서 나온 짝지어진 자료이기 때문이다. 따라서 독립 두 표본 검정보다 차이값 1개 표본의 문제로 보는 것이 맞다.

계산 과정

1. 같은 10명에게 두 조건을 모두 측정했으므로 검정 방법은 대응표본 t-검정이다.
2. 대응표본 t-검정의 자유도는 $n-1=10-1=9$이다.
3. 대립가설이 "차이가 있다"이므로 유의수준 5% 양측검정의 임계값은 $t_{0.025}(9)=2.26$이다.
4. $|t|=1.9$이고 $1.9<2.26$이므로 귀무가설을 기각하지 못한다.
5. 따라서 유의수준 5%에서 혈중 콜레스테롤 양에 차이가 있다고 볼 수 없다.

NoteBase

선지별 해설

① O
같은 10명에게 두 경우를 모두 측정했으므로 대응표본 비교가 맞다. 또한 양측검정에서
$|t| = 1.9 < 2.26 = t_{0.025}(9)$이므로 유의수준 5%에서 차이가 없다고 본다.

② X
대응표본 비교를 택한 점은 맞지만, "유의수준 5%에서 차이가 있다"가 틀렸다. 틀린 표현은 "차이가 있다"이고, 올바르게는 "유의수준 5%에서 차이가 있다고 볼 수 없다"이다.

③ X
틀린 표현은 "독립인 표본에 의한 두 집단 비교"이다. 같은 10명을 반복 측정했으므로 독립 두 표본이 아니라 대응표본 비교가 맞다. 결과 판단 자체는 차이가 없다는 점이 맞지만, 검정 방법 설명이 잘못되었다.

④ X
틀린 표현은 "독립인 표본에 의한 두 집단 비교"와 "차이가 있다"이다. 올바르게는 "대응표본 비교를 실시하며, 유의수준 5%에서 차이가 있다고 볼 수 없다"이다.

오답 포인트

출제자는 "10명의 피실험자를 무작위로 추출"했다는 문장 때문에 독립 두 표본으로 오해하게 만들었다. 핵심은 표본 추출 방식이 아니라 같은 사람에게 두 조건을 모두 측정했는지 여부이다. 또 대립가설이 "차이가 있다"이면 양측검정이므로 $t_{0.05}(9) = 1.83$이 아니라 $t_{0.025}(9) = 2.26$를 써야 한다.

다시 나오면 이렇게 판단

같은 대상에게 전후·유무·처치별 값을 모두 재면 먼저 대응표본인지부터 판단한다. 그다음 대립가설이 양측인지 단측인지 확인하고, 대응표본이면 자유도 $n-1$, 독립 두 표본이면 문제 조건에 맞는 다른 자유도를 적용해 임계값과 비교한다.

암기 포인트

같은 사람 두 번 측정은 대응표본, "차이가 있다"는 양측검정, 대응표본 자유도는 $n-1$이다.

NoteBase

문 25. 사건 A, B, C는 서로 독립이고, $P(A \cap B \cap C) = \frac{4}{27}$이다. $P(A) = P(B) = 2P(C)$일 때 $P(A \cup B) - P(B \cap C)$의 값은?

① $\frac{2}{9}$
② $\frac{2}{3}$
③ $\frac{7}{9}$
④ $\frac{5}{6}$

정답 2번

출제 포인트

이 문제는 독립사건의 곱셈법칙으로 각 사건의 확률을 구한 뒤 합사건과 교사건 확률을 계산할 수 있는지를 묻는다.

조건 정리

주어진 값은 $P(A \cap B \cap C) = \frac{4}{27}$, $P(A) = P(B) = 2P(C)$이다.
구해야 할 값은 $P(A \cup B) - P(B \cap C)$이다.
사건 A, B, C는 서로 독립이므로 $P(A \cap B \cap C) = P(A)P(B)P(C)$를 사용한다.

사용할 공식

$P(A \cap B \cap C) = P(A)P(B)P(C)$, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$이다.
독립 조건이 있으므로 $P(A \cap B) = P(A)P(B)$, $P(B \cap C) = P(B)P(C)$로 바꿀 수 있다.

계산 과정

$P(C) = x$라고 두면 $P(A) = P(B) = 2x$이다.
그러면 $(2x)(2x)(x) = \frac{4}{27}$이므로 $4x^3 = \frac{4}{27}$, $x^3 = \frac{1}{27}$, $x = \frac{1}{3}$이다.
따라서 $P(C) = \frac{1}{3}$, $P(A) = P(B) = \frac{2}{3}$이다.
$P(A \cup B) = \frac{2}{3} + \frac{2}{3} - \frac{2}{3} \cdot \frac{2}{3} = \frac{8}{9}$이다.
또한 $P(B \cap C) = \frac{2}{3} \cdot \frac{1}{3} = \frac{2}{9}$이므로 $P(A \cup B) - P(B \cap C) = \frac{8}{9} - \frac{2}{9} = \frac{2}{3}$이다.

NoteBase

선지 판단

1번 X: $\frac{2}{9}$은 $P(B \cap C)$ 값일 뿐, 문제에서 구하는 $P(A \cup B) - P(B \cap C)$ 값이 아니다. 틀린 수치 → 올바르게는 구하는 값이 $\frac{2}{3}$이다.
2번 O: 계산 결과 $P(A \cup B) - P(B \cap C) = \frac{2}{3}$이므로 옳다.
3번 X: $\frac{7}{9}$은 계산 결과와 일치하지 않는다. 합사건 확률과 교사건 확률을 정확히 구하면 $\frac{8}{9} - \frac{2}{9} = \frac{2}{3}$이다. 틀린 수치 → 올바르게는 $\frac{2}{3}$이다.
4번 X: $\frac{5}{6}$도 계산 결과와 다르다. 독립을 이용해 $P(A \cap B)$와 $P(B \cap C)$를 각각 곱으로 계산하면 최종값은 $\frac{2}{3}$이다. 틀린 수치 → 올바르게는 $\frac{2}{3}$이다.

오답 포인트

세 사건이 독립이라는 사실은 세 사건의 교집합뿐 아니라 두 사건의 교집합에도 그대로 적용된다. $P(B \cap C) = \frac{2}{9}$만 구하고 끝내거나, $P(A \cup B)$에서 $P(A \cap B)$를 빼지 않으면 오답이 된다.

다시 나오면 이렇게 판단

확률비가 주어지면 하나의 미지수로 두고 독립사건의 곱셈식에 대입해 각 사건의 확률부터 확정한다. 그 다음 합사건 공식과 교사건 공식을 차례로 적용하면 계산이 흔들리지 않는다.

암기 포인트

독립사건은 교집합을 곱으로 바꾸고, 합사건은 더한 뒤 겹치는 부분을 한 번 뺀다.

NoteBase