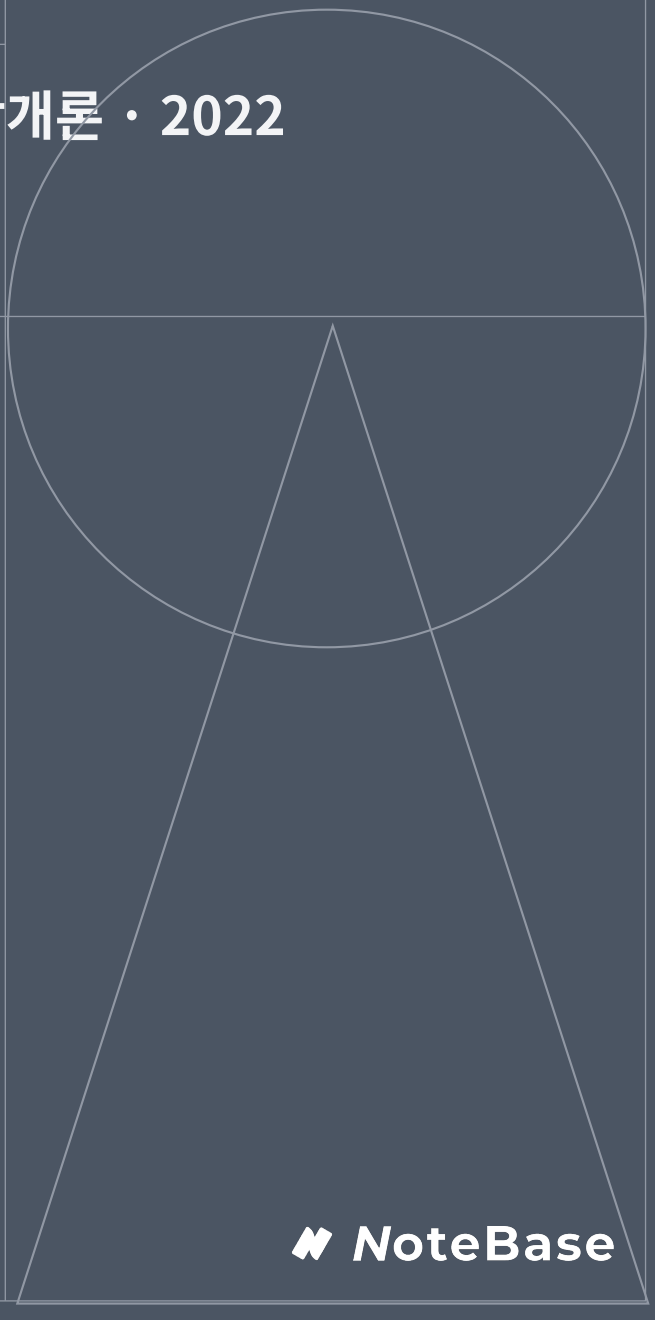
기출OX 해설집

공무원/국가직/9급 · 통계학개론 · 2022



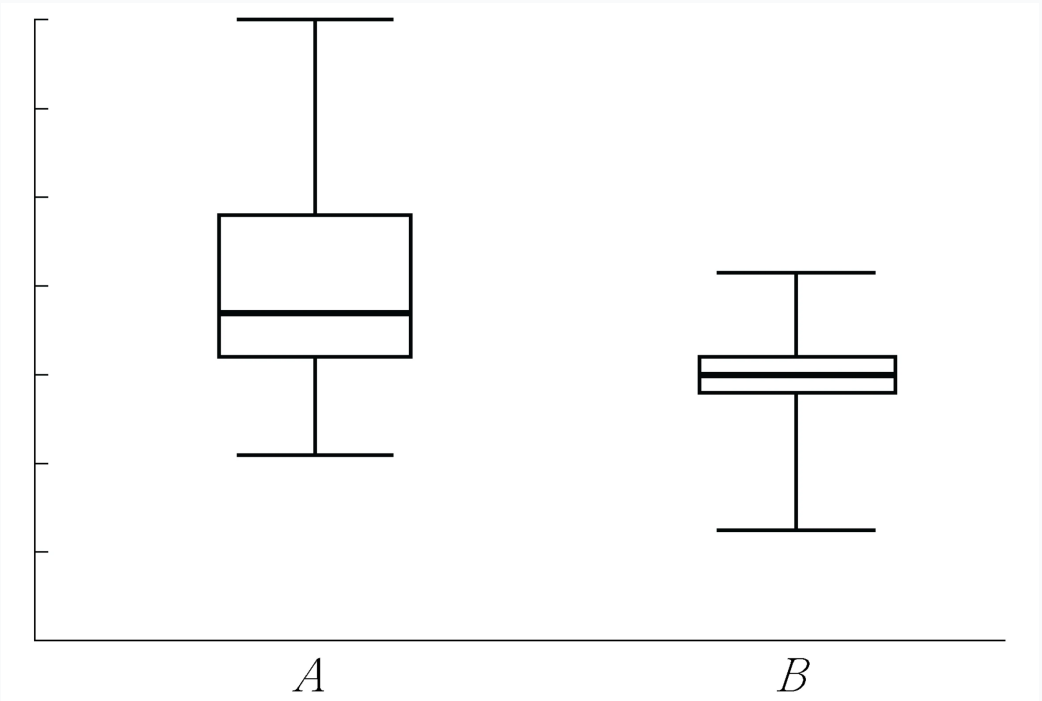
NoteBase

기출OX 해설집

공무원/국가직/9급 · 통계학개론 · 2022

공무원/국가직/9급 · 통계학개론 · 2022

문 1. 두 지역 A, B의 소득에 대한 상자 그림(box plot)은 다음과 같다. 이에 대한 설명으로 옳은 것만을 모두 고르면?



ㄱ. 최솟값은 A 지역이 B 지역보다 크다.

ㄴ. 중앙값은 A 지역이 B 지역보다 크다.

ㄷ. 범위는 A 지역이 B 지역보다 작다.

① ㄱ, ㄴ

② ㄱ, ㄷ

③ ㄴ, ㄷ

④ ㄱ, ㄴ, ㄷ

정답 1번

NoteBase

간단한 해설

범위는 **최댓값과 최솟값의 차이**로, 위·아래 수염 끝 사이의 간격을 비교한다. 이 간격이 A 지역에서 더 크므로 A 지역의 범위도 더 크다.

보기별 해설

ㄱ. O: 그림에서 아래쪽 수염 끝은 최솟값을 나타낸다. A 지역의 아래쪽 수염 끝이 B 지역보다 높은 위치에 있으므로 **최솟값은 A 지역이 더 크다**.

ㄴ. O: 중앙값은 자료를 크기순으로 나열했을 때 가운데에 해당하는 값으로, **상자 안의 가로선**으로 표시된다. A 지역의 가로선이 B 지역보다 높은 위치에 있으므로 중앙값은 A 지역이 더 크다.

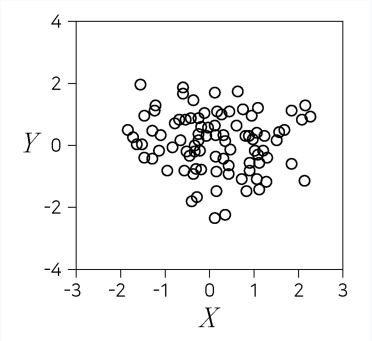
ㄷ. X: 범위는 최댓값에서 최솟값을 뺀 값이다. 이 그림에서는 위쪽 수염 끝부터 아래쪽 수염 끝까지의 간격에 해당하며, A 지역의 간격이 B 지역보다 더 크다. 따라서 범위는 A 지역이 B 지역보다 작은 것이 아니라 **더 크다**.

옳은 보기는 ㄱ, ㄴ이므로 이 조합에 해당하는 선택지는 ①이다.

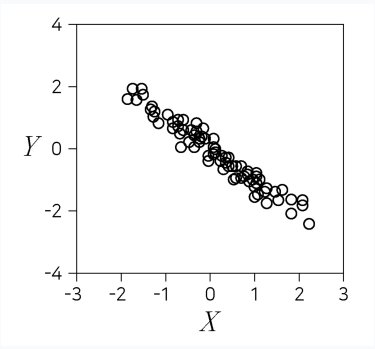
NoteBase

문 2. 다음 제시된 두 변수 X와 Y에 대한 산점도 중 X와 Y의 표본상관계수(피어슨의 표본상관계수)가 가장 큰 것은?

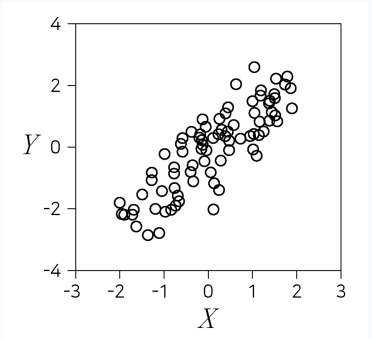
①



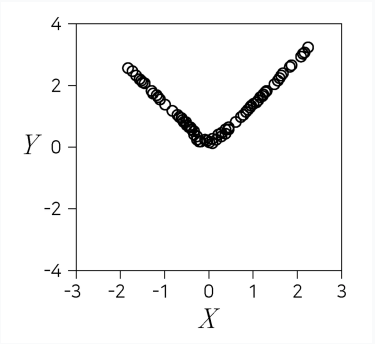
②



③



④



정답 3번

간단한 해설

③의 점들은 오른쪽으로 갈수록 위로 올라가는 직선 주위에 모여 있어 **뚜렷한 양의 선형관계**를 보인다. 피어슨 상관계수는 곡선에 얼마나 밀집했는지가 아니라 **직선 관계의 방향과 강도**를 나타낸다.

NoteBase

선지별 해설

피어슨의 표본상관계수 r은 두 변수의 선형관계의 방향과 강도를 나타내며, −1 이상 1 이하의 값을 갖는다. 우상향 직선에 가까울수록 1에, 우하향 직선에 가까울수록 −1에 가까워진다. 이 문제는 절댓값이 아니라 상관계수 자체가 가장 큰 산점도를 묻는다.

①: 점들이 넓게 흩어져 있고, X가 증가할 때 Y가 일정한 방향으로 변하는 경향이 뚜렷하지 않다. 따라서 선형관계가 약하여 상관계수는 0에 가깝다.

②: 점들이 왼쪽 위에서 오른쪽 아래로 향하는 직선 주위에 밀집해 있다. X가 증가할수록 Y가 감소하므로 **강한 음의 선형관계**이며, 상관계수는 −1에 가깝다. 절댓값이 크더라도 음수이므로 양의 상관계수보다 작다.

③: 점들이 왼쪽 아래에서 오른쪽 위로 향하는 직선 주위에 모여 있다. X가 증가할수록 Y도 증가하는 **뚜렷한 양의 선형관계**이므로, 상관계수는 양수이며 ①보다 1에 가깝다.

④: 점들이 **U자형의 비선형관계**를 보인다. 왼쪽에서는 X가 증가할수록 Y가 감소하고 오른쪽에서는 증가하므로, 두 구간의 경향이 상쇄되어 전체의 양의 선형관계는 ③보다 약하다. 점들이 곡선에 밀집해 있다는 사실만으로 피어슨 상관계수가 1에 가까워지는 것은 아니다.

네 산점도를 선형관계의 방향과 강도로 비교하면, 뚜렷한 양의 선형관계를 보이는 ③의 표본상관계수가 가장 크다.

NoteBase

문 3. 통계적 가설검정에서 제1종 오류확률에 대한 설명으로 옳은 것은?

① 귀무가설이 참일 때 귀무가설을 기각할 확률

② 귀무가설이 참일 때 귀무가설을 기각하지 않을 확률

③ 귀무가설이 거짓일 때 귀무가설을 기각할 확률

④ 귀무가설이 거짓일 때 귀무가설을 기각하지 않을 확률

정답 1번

간단한 해설

제1종 오류는 **참인 귀무가설을 기각**하는 잘못이다. 그 오류확률은 귀무가설이 참이라는 조건에서 귀무가설을 기각할 확률이다.

개념 구분

귀무가설이 참일 때: 실제로 참인 귀무가설을 기각하면 제1종 오류가 발생한다. 제1종 오류확률은 **귀무가설이 참이라는 조건에서 기각할 확률**이며, 보통 α로 나타낸다. 같은 조건에서 귀무가설을 기각하지 않는 것은 올바른 판단이므로, 그 확률은 $1-\alpha$이다.

귀무가설이 거짓일 때: 거짓인 귀무가설을 **기각하지 않으면 제2종 오류**가 발생하며, 그 확률을 보통 β로 나타낸다. 반대로 거짓인 귀무가설을 기각하는 것은 올바른 판단이고, 그 확률 $1-\beta$를 검정력이라고 한다. 따라서 제1종 오류와 제2종 오류는 귀무가설의 실제 참·거짓과 검정의 기각 여부를 함께 구분해야 한다.

NoteBase

문 4. 단순선형회귀모형 $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ (i = 1, 2, \ldots, 22)$에서 최소제곱법으로 회귀식을 추정한 후 잔차제곱합을 구하였더니 2,100이었다. 이때 오차항 분산(σ^2)의 불편추정값은? (단, ε_i는 $N(0, \sigma^2)$를 따르고 서로 독립이다)

① 100

② 105

③ 110

④ 120

정답 2번

간단한 해설

절편과 기울기 두 모수를 추정하므로 **잔차의 자유도는 22 − 2 = 20**이다. 잔차제곱합을 이 자유도로 나눈 $\frac{2100}{20} = 105$가 오차항 분산의 불편추정값이다.

계산 과정

관측값의 수는 n = 22이고, 최소제곱법으로 구한 잔차제곱합은 SSE = 2,100이다. 잔차는 관측값과 추정한 회귀식의 예측값 사이의 차이이며, SSE는 이 잔차들을 제곱하여 더한 값이다.

단순선형회귀모형에서는 **절편과 기울기 두 모수를 추정**하므로 잔차의 자유도는 $n - 2$이다. 따라서 이 문제의 잔차 자유도는 $22 - 2 = 20$이다. 주어진 서로 독립인 정규오차 조건에서 $E(\text{SSE}) = (n - 2)\sigma^2$이므로, SSE를 관측값 수 n이 아니라 잔차 자유도로 나누어야 불편추정값을 얻는다.

오차항 분산의 불편추정량은 $s^2 = \frac{\text{SSE}}{n-2}$이다. 주어진 값을 대입하면 $s^2 = \frac{2100}{22-2} = 105$이다. 따라서 구하는 분산의 불편추정값은 105이며, ②의 값과 일치한다.

NoteBase

문 5. 코로나바이러스 감염 상황과 진단키트의 진단 확률이 다음과 같을 때, 한 사람의 진단키트 결과가 음성이라면 이 사람이 실제 음성일 확률은?

◯ 전체 인구 중 코로나바이러스 양성은 20%이다.
◯ 진단키트에서 실제 음성인 사람을 음성으로 판단할 확률이 70%이다.
◯ 진단키트에서 실제 양성인 사람을 양성으로 판단할 확률이 80%이다.

① $\frac{13}{15}$
② $\frac{9}{10}$
③ $\frac{14}{15}$
④ $\frac{29}{30}$

정답 3번

간단한 해설

키트에서 음성이 나온 경우에는 **실제 음성과 실제 양성의 위음성**이 모두 포함된다. 전체 인구 중 키트 음성은 60%, 실제 음성이면서 키트 음성인 경우는 56%이므로, 구하는 확률은 $\frac{56}{60} = \frac{14}{15}$이다.

계산 과정

실제 음성인 사건을 A, 키트 결과가 음성인 사건을 B라고 하자. 구하는 값은 $P(A \mid B)$이다. 실제 양성이 20%이므로 $P(A) = 0.8$이고, 실제 음성을 음성으로 판단할 확률은 $P(B \mid A) = 0.7$이다. 실제 양성을 양성으로 판단할 확률이 0.8이므로, 실제 양성을 음성으로 판단할 확률은 $1 - 0.8 = 0.2$이다.

실제 음성이면서 키트 결과도 음성일 확률은 실제 음성의 비율에 그 집단에서 음성이 나올 확률을 곱한다. 따라서 $P(A \cap B) = 0.8 \cdot 0.7 = 0.56$이다. 이는 전체 인구의 56%가 두 조건을 모두 만족한다는 뜻이다.

키트 음성은 **실제 음성에서 나온 음성과 실제 양성에서 나온 음성**으로 나뉜다. 두 경우는 서로 겹치지 않고 키트 음성인 경우를 모두 포함하므로 더한다. 실제 양성이면서 키트 음성일 확률은 $0.2 \cdot 0.2 = 0.04$이므로, $P(B) = 0.56 + 0.04 = 0.60$이다.

이제 기준 집단을 전체 인구에서 **키트 결과가 음성인 사람들**로 좁힌다. 그중 실제 음성인 비율을 구하므로 $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$이다. 값을 대입하면 $\frac{0.56}{0.60} = \frac{14}{15}$이며, 이는 ③의 값과 일치한다.

NoteBase

문 6. 두 연속형 확률변수 X와 Y가 독립일 때, 이에 대한 설명으로 옳지 않은 것은?

① X와 Y의 공분산은 0이다.

② X와 Y의 합에 대한 분산은 X와 Y의 분산의 합이다.

③ X와 Y의 곱에 대한 기댓값은 X와 Y의 기댓값의 곱이다.

④ X가 주어졌을 때 Y의 조건부확률밀도함수는 X의 주변확률밀도함수와 같다.

정답 4번

간단한 해설

독립이면 X의 값을 알아도 Y의 분포는 달라지지 않는다. 따라서 X가 주어졌을 때 Y의 조건부확률밀도함수는 X가 아니라 **Y의 주변확률밀도함수**와 같다.

선지별 해설

아래 기댓값·공분산·분산 관계는 필요한 기댓값과 분산이 존재한다는 전제에서 성립한다.

① O: 독립인 확률변수는 $E[XY] = E[X]E[Y]$를 만족한다. 공분산은 $\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]$이므로, **독립이면 공분산은 0**이다.

② O: 일반적으로 두 확률변수의 합의 분산은 각 분산의 합에 공분산의 두 배를 더한 값이다. 독립이면 **공분산 항이 0**이 되므로, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$가 성립한다.

③ O: 독립이면 결합확률밀도함수는 $f_{X,Y}(x, y) = f_X(x)f_Y(y)$로 분해된다. 따라서 곱의 기댓값을 구하는 적분도 X에 관한 부분과 Y에 관한 부분으로 분리되어, **곱의 기댓값이 기댓값의 곱**인 $E[XY] = E[X]E[Y]$가 성립한다.

④ X: $f_X(x) > 0$인 x에서 Y의 조건부확률밀도함수는 $f_{Y \mid X}(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$이다. 독립 조건에 따라 분자에 $f_X(x)f_Y(y)$를 대입하면 $f_X(x)$가 약분되어 $f_Y(y)$가 남는다. 즉 X의 값을 조건으로 주어도 Y의 분포는 바뀌지 않으므로, 선지의 'X의 주변확률밀도함수'는 **'Y의 주변확률밀도함수'**로 고쳐야 한다.

NoteBase

문 7. 다음은 단순선형회귀모형을 적합하여 얻은 분산분석표이다.
이에 대한 설명으로 옳지 않은 것은?

요인	제곱합	자유도	평균제곱	$F-$값
회귀	540		(가)	(다)
잔차	200		(나)	
계	740	11		

① (가)의 값은 540이다.

② (나)의 값은 20이다.

③ (다)의 값은 27이다.

④ 모형의 결정계수는 $\frac{200}{740}$이다.

정답 4번

간단한 해설

결정계수는 전체 변동 중 **회귀모형이 설명하는 변동의 비율**이므로 $R^2 = \frac{540}{740}$이다. ④의 $\frac{200}{740}$은 모형이 설명하지 못한 잔차 변동의 비율이다.

NoteBase

선지별 해설

분산분석표에서 제곱합은 변동의 크기를 나타내며, 전체제곱합 740은 회귀제곱합 540과 잔차제곱합 200으로 나뉜다. 절편을 포함한 단순선형회귀의 회귀 자유도는 1이므로, 전체 자유도가 11이면 잔차 자유도는 $11 - 1 = 10$이다.

① O: 평균제곱은 **각 제곱합을 해당 자유도로 나눈 값**이다. 회귀제곱합은 540, 회귀 자유도는 1이므로 회귀평균제곱 (가)는 $\frac{540}{1} = 540$이다.

② O: 잔차 자유도는 전체 자유도 11에서 회귀 자유도 1을 뺀 10이다. 따라서 잔차평균제곱 (나)는 $\frac{200}{10} = 20$이다.

③ O: 회귀모형의 F값은 **회귀평균제곱을 잔차평균제곱으로 나눈 값**이다. 따라서 (다)는 $F = \frac{540}{20} = 27$이다.

④ X: 결정계수는 전체제곱합 중 **회귀제곱합이 차지하는 비율**이다. 따라서 $R^2 = \frac{540}{740} = \frac{27}{37}$로 고쳐야 한다. 선지의 $\frac{200}{740}$은 잔차제곱합을 전체제곱합으로 나눈 값으로, 모형이 설명하지 못한 변동의 비율인 $1 - R^2$이다.

NoteBase

문 8. 다음은 시장조사에서 세탁기 색에 대한 선호도를 알아보기 위해 최근 판매된 세탁기의 색을 조사한 결과이다.
선호하는 색이 4:4:1:1의 비율로 분포한다는 가설을 검정하려 할 때, 이에 대한 설명으로 옳은 것만을 모두 고르면?

색	흰색	검은색	상아색	회색	합계
빈도	40	24	10	6	80

ㄱ. 검정통계량은 5이다.
ㄴ. 검정통계량의 자유도는 4이다.
ㄷ. 검은색의 기대도수는 32이다.

① ㄱ, ㄴ
② ㄱ, ㄷ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 2번

간단한 해설

색이 네 범주이고 가설의 비율이 모두 주어져 있으므로, 카이제곱 적합도검정의 자유도는 **4 − 1 = 3**이다. 전체 도수가 80으로 고정되어 있어 세 범주의 도수를 정하면 나머지 하나도 결정되기 때문이다.

NoteBase

보기별 해설

관측도수가 가설에서 정한 비율과 부합하는지 확인하는 카이제곱 적합도검정이다. 표의 관측도수는 흰색 40, 검은색 24, 상아색 10, 회색 6이며 합계는 80이다. 기대도수는 **전체 도수에 가설의 비율을 곱한 값**이므로, 같은 색 순서로 32, 32, 8, 8이다.

ㄱ. O: 관측도수를 O_i, 기대도수를 E_i라 하면 검정통계량은 $\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$이다. 관측도수와 기대도수의 차이는 흰색 8, 검은색 −8, 상아색 2, 회색 −2이다. 흰색과 검은색은 각각 $\frac{8^2}{32} = 2$, 상아색과 회색은 각각 $\frac{2^2}{8} = 0.5$만큼 기여한다. 이를 더하면 검정통계량은 2 + 2 + 0.5 + 0.5 = 5이다.

ㄴ. X: 가설에서 네 범주의 비율을 모두 정해 주었으며, 표본으로 별도의 모수를 추정하지 않는다. 이 경우 자유도는 **범주 수에서 1을 뺀 값**이므로 4 − 1 = 3이다. 전체 도수가 80으로 고정되어 있어 네 범주의 도수를 독립적으로 정할 수 없기 때문이다. 따라서 자유도 4는 3으로 고친다.

ㄷ. O: 검은색은 가설의 비율 4:4:1:1 중 두 번째 4에 해당한다. 전체 비율의 합이 10이므로 검은색의 기대도수는 $80 \times \frac{4}{10} = 32$이다. 표에 나타난 24는 실제로 관측된 도수이고, **32는 가설이 성립할 때 기대하는 도수**이다.

옳은 보기는 ㄱ과 ㄷ이므로 ②의 조합과 일치한다.

NoteBase

문 9. 분포에 대한 설명으로 옳지 않은 것은?

① 확률변수 F가 자유도 (df_1, df_2)인 F 분포를 따를 때, $\frac{1}{F}$은 자유도 (df_2, df_1)인 F 분포를 따른다.

② 평균이 μ이고 분산이 σ^2인 정규모집단에서 임의로 추출한 n개의 확률변수에 대한 표본평균 $\overline{X}$는 $N\left(\mu, \frac{\sigma^2}{n}\right)$를 따른다.

③ 확률변수 $X_1, X_2, \cdots, X_n$이 서로 독립이며 표준정규분포를 따르면 $X_1^2 + X_2^2 + \cdots + X_n^2$은 자유도가 n인 카이제곱분포를 따른다.

④ 평균이 μ이고 분산이 σ^2인 모집단에서 임의로 추출한 5개의 확률변수에 대한 표본평균이 $\overline{X}$일 때, $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$는 근사적으로 표준정규분포를 따른다.

정답 4번

간단한 해설

모집단의 분포를 모르는 상태에서 **표본 크기가 5라는 이유만으로 정규근사를 보장할 수는 없다**. 중심극한정리는 표본 크기가 충분히 클 때 적용하며, 모집단이 정규분포라면 표본 크기가 작아도 표본평균은 정확히 정규분포를 따른다.

선지별 해설

① O: 서로 독립인 카이제곱 확률변수 U, V의 자유도가 각각 df_1, df_2이면, $F = \frac{U/df_1}{V/df_2}$는 해당 자유도의 F 분포를 따른다. 역수를 취하면 $\frac{1}{F} = \frac{V/df_2}{U/df_1}$가 되어 **분자와 분모의 자유도 순서가 바뀐다**. 따라서 역수의 F 분포 자유도는 (df_2, df_1)이다.

② O: 정규모집단에서 독립적으로 추출한 표본의 평균은 **표본 크기와 관계없이 정확히 정규분포**를 따른다. 표본평균의 기댓값은 μ이고, 독립인 표본들의 합을 표본 크기로 나누므로 분산은 $\mathrm{Var}(\overline{X}) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$이다. 따라서 $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$이다.

③ O: 카이제곱분포는 **서로 독립인 표준정규 확률변수들의 제곱합**으로 정의된다. 제곱하여 더하는 독립 표준정규 확률변수가 n개이므로 $\sum_{i=1}^{n} X_i^2 \sim \chi_n^2$이며, 자유도는 n이다.

④ X: 모집단이 정규분포라는 조건이 없고 표본 크기도 5이므로, 표본평균을 표준화한 값이 근사적으로 표준정규분포를 따른다고 일반화할 수 없다. 모집단의 평균과 유한한 양의 분산이 존재하고 표본을 독립적으로 추출할 때, 중심극한정리에 따라 **표본 크기가 충분히 커지면** $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$의 분포가 표준정규분포에 가까워진다. 필요한 표본 크기는 모집단 분포의 비대칭성이나 꼬리의 형태 등에 따라 달라진다. 모집단이 정규분포라면 n = 5에서도 이 Z는 근사가 아니라 정확히 표준정규분포를 따른다.

NoteBase

문 10. 세 가지 필터의 성능을 비교하기 위해 일원배치 분산분석법(one-way analysis of variance)을 실시하여 얻은 분산분석표의 일부가 다음과 같을 때 (가)의 값은?

요인	제곱합	자유도	평균제곱	$F-$값
처리	50			(가)
오차	20	6		
계	70			

① 2.5

② 5.0

③ 7.5

④ 15.0

정답 3번

간단한 해설

분산분석의 F값은 **처리 평균제곱을 오차 평균제곱으로 나눈 값**이다. 처리 평균제곱은 25, 오차 평균제곱은 $\frac{10}{3}$이므로 F값은 7.5이다.

계산 과정

처리 자유도: 필터 종류가 처리에 해당하며, 처리 수는 3이다. 일원배치 분산분석에서 **처리 자유도는 처리 수보다 1 작으므로** 3 − 1 = 2이다. 각 처리 평균과 전체 평균의 차이에는 하나의 제약이 있어 독립적인 차이는 2개이다.

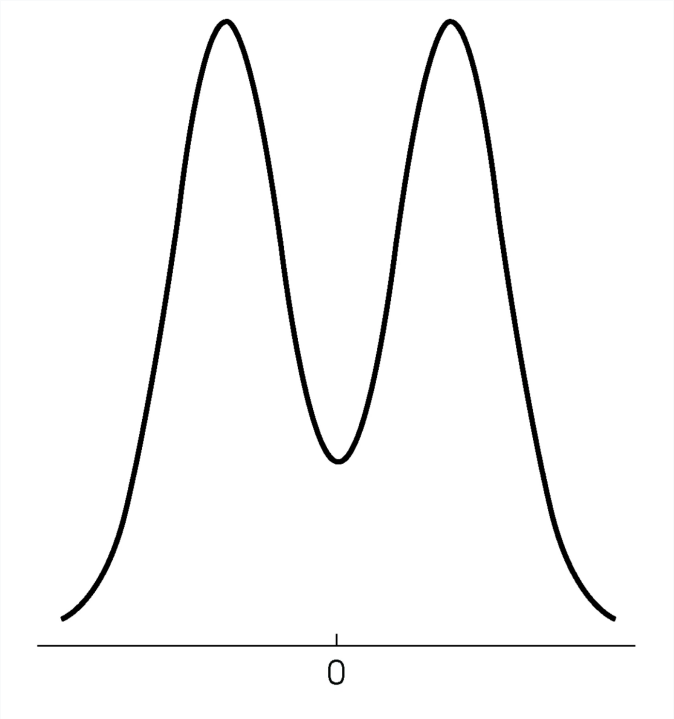
처리 평균제곱: 평균제곱은 **제곱합을 해당 자유도로 나눈 값**이다. 표의 처리 제곱합이 50이고 처리 자유도가 2이므로, 처리 평균제곱은 $\frac{50}{2} = 25$이다. 이는 필터 종류 간 변동을 자유도당 크기로 나타낸 값이다.

오차 평균제곱: 표의 오차 제곱합은 20, 오차 자유도는 6이므로 오차 평균제곱은 $\frac{20}{6} = \frac{10}{3}$이다. 이는 같은 필터 종류 안에서 나타나는 변동을 자유도당 크기로 나타낸 값이다.

F값: 처리 간 변동을 처리 내 변동과 비교하기 위해 **처리 평균제곱을 분자, 오차 평균제곱을 분모**로 둔다. 따라서 $F = \frac{25}{\frac{10}{3}} = 7.5$이며, (가)는 ③의 7.5에 해당한다. F값은 같은 단위의 평균제곱끼리 나눈 비이므로 단위가 없다.

NoteBase

문 11. 다음은 어느 확률변수의 확률밀도함수이다. 이에 대한 설명으로 옳은 것만을 모두 고르면? (단, 함수는 0을 기준으로 좌우 대칭이다)



ㄱ. 평균은 0이다.
ㄴ. 중앙값은 0이다.
ㄷ. 최빈값의 절댓값은 중앙값보다 크다.

① ㄱ
② ㄱ, ㄴ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 4번

간단한 해설

그래프의 대칭 중심인 0은 중앙값이지만, 최빈값은 **밀도가 가장 높은 두 봉우리의 가로축 값**이다. 두 최빈값 모두 0이 아니므로 그 절댓값은 중앙값 0보다 크다.

NoteBase

보기별 해설

ㄱ. O: 평균은 각 값에 확률을 가중치로 주어 구한 값이다. 평균이 존재하는 좌우 대칭 분포에서는 대칭 중심에서 같은 거리만큼 떨어진 양수와 음수의 기여가 서로 상쇄된다. 이 분포의 **대칭 중심은 0**이므로 평균은 0이다.

ㄴ. O: 중앙값은 전체 확률을 절반씩 나누는 위치이다. 확률밀도함수 아래의 전체 넓이는 1이고, 이 그래프는 0을 기준으로 좌우 대칭이므로 **0의 왼쪽과 오른쪽 넓이가 각각 0.5**이다. 따라서 중앙값은 0이다.

ㄷ. O: 연속확률변수의 최빈값은 확률밀도함수가 가장 높아지는 가로축 값이다. 그림에서는 0이 두 봉우리 사이의 골에 있고, 가장 높은 봉우리는 0의 양쪽에 하나씩 있다. 따라서 **최빈값은 0이 아닌 음수와 양수 두 개**이며, 두 최빈값의 절댓값은 모두 중앙값 0보다 크다.

ㄱ, ㄴ, ㄷ이 모두 성립하므로 ④의 조합에 해당한다.

NoteBase

문 12. 어느 지역에서 20대 남자의 몸무게는 평균이 μ이고 표준편차가 8인 정규분포를 따른다고 한다. 이 지역에서 임의추출한 20대 남자 64명의 평균 몸무게가 66일 때, 모평균 μ에 대한 95% 신뢰구간은? (단, 단위는 kg이며, 표준정규분포를 따르는 확률변수 Z에 대하여 $P(Z \geq 1.96) = 0.025$, $P(Z \geq 1.645) = 0.05$이다)

① (64.040, 67.960)

② (64.355, 67.645)

③ (65.755, 66.245)

④ (65.794, 66.206)

정답 1번

간단한 해설

표본평균의 표준오차는 1 kg이며, 95% 신뢰구간에서는 **양쪽 꼬리에 각각 2.5%를 남기는 임계값 1.96**을 사용한다. 따라서 표본평균 66 kg에서 1.96 kg씩 빼고 더한 (64.040, 67.960) kg이 된다.

계산 과정

모집단이 정규분포를 따르고 **모표준편차가 8 kg으로 알려져 있으므로**, 표준정규분포를 이용하여 모평균의 신뢰구간을 구한다. 표본크기는 $n = 64$, 관측된 표본평균은 $\bar{x} = 66$ kg, 모표준편차는 $\sigma = 8$ kg이다.

표본평균의 표준오차는 $\frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{64}} = 1$ kg이다. 이는 개인 몸무게의 표준편차가 아니라, **64명의 평균 몸무게가 표본마다 변동하는 정도**를 나타낸다.

95% 양측 신뢰구간은 중앙에 확률 0.95를 두므로 양쪽 꼬리확률은 각각 $\frac{1-0.95}{2} = 0.025$이다. 주어진 $P(Z \geq 1.96) = 0.025$와 표준정규분포의 대칭성에 따라 임계값은 1.96이다. 표준화한 표본평균 $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$가 표준정규분포를 따르므로, 신뢰구간은 $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$으로 구한다.

오차한계는 $1.96 \times 1 = 1.96$ kg이다. 하한은 $66 - 1.96 = 64.040$ kg, 상한은 $66 + 1.96 = 67.960$ kg이다. 따라서 모평균의 95% 신뢰구간은 (64.040, 67.960) kg으로, ①의 구간과 일치한다.

NoteBase

문 14. 확률변수 X에 대하여 $Var(3X+9)=36$, $E(X(X+1))=10$을 만족할 때, $E(X)$의 값은? (단, $E(X)>0$이다)

① 1

② 2

③ 3

④ 4

정답 2번

간단한 해설

분산이 4이므로 평균 $\mu = E(X)$는 $\mu^2+\mu-6=0$을 만족한다. 두 근 중 **양수 조건**을 만족하는 값은 2이다.

계산 과정

평균을 $\mu = E(X)$라고 하자. 분산은 **상수를 더해도 변하지 않고, 상수배하면 그 상수의 제곱배**가 된다. 따라서 $Var(3X+9)=9Var(X)=36$이므로 $Var(X)=\frac{36}{9}=4$이다.

기댓값의 선형성에 따라 $E(X(X+1))=E(X^2)+\mu=10$이다. 한편 분산 공식 $Var(X)=E(X^2)-\mu^2$에서 $E(X^2)=4+\mu^2$이다. 이는 제곱의 기댓값이 평균의 제곱에 분산을 더한 값이라는 뜻이다.

이를 앞의 기댓값 조건에 대입하면 $\mu^2+\mu+4=10$, 즉 $\mu^2+\mu-6=0$이다. 인수분해하면 $(\mu-2)(\mu+3)=0$이므로 $\mu=2$ 또는 $\mu=-3$이다. $E(X)>0$ **조건**에 따라 음수인 근을 제외하면 $E(X)=2$이며, 선택지 ②의 값과 일치한다.

NoteBase

문 15. 독립인 두 모집단 1과 2에 대한 확률분포는 각각 다음과 같다.
모집단 i에서 임의로 추출한 100개의 확률변수의 평균을 $\overline{X_i}$라 할 때, $\overline{X_1}+\overline{X_2}$의 근사분포는?

구분	x	0	1	2
모집단 1	$P(X_1=x)$	$\frac{2}{3}$	$\frac{1}{3}$	0
모집단 2	$P(X_2=x)$	0	$\frac{1}{3}$	$\frac{2}{3}$

① $N(2,\frac{2}{900})$
② $N(2,\frac{4}{900})$
③ $N(4,\frac{2}{900})$
④ $N(4,\frac{4}{900})$

정답 2번

간단한 해설

표본평균의 분산은 **모분산을 표본크기 100으로 나눈 값**이며, 독립인 두 표본평균을 더하면 분산도 더한다. 두 평균의 합은 평균이 2, 분산이 $\frac{4}{900}$이고, 중심극한정리에 따라 정규분포로 근사한다.

NoteBase

계산 과정

모집단 1: 표에서 0과 1의 확률은 각각 $\frac{2}{3}, \frac{1}{3}$이다. 기댓값은 각 값에 그 확률을 곱해 더하므로 $E(X_1) = 0 \cdot \frac{2}{3} + 1 \cdot \frac{1}{3} = \frac{1}{3}$이다. 또한 $E(X_1^2) = \frac{1}{3}$이므로, 분산 공식 $\mathrm{Var}(X) = E(X^2) - \{E(X)\}^2$을 적용하면 $\mathrm{Var}(X_1) = \frac{1}{3} - \left(\frac{1}{3}\right)^2 = \frac{2}{9}$이다.

모집단 2: 1과 2의 확률이 각각 $\frac{1}{3}, \frac{2}{3}$이므로 $E(X_2) = 1 \cdot \frac{1}{3} + 2 \cdot \frac{2}{3} = \frac{5}{3}$이다. 제곱의 기댓값은 $E(X_2^2) = 1^2 \cdot \frac{1}{3} + 2^2 \cdot \frac{2}{3} = 3$이며, $\mathrm{Var}(X_2) = 3 - \left(\frac{5}{3}\right)^2 = \frac{2}{9}$이다.

표본평균의 평균과 분산: 각 모집단에서 독립적으로 임의추출한 100개 관측값의 평균은 기댓값이 모평균과 같고, **분산은 모분산의 100분의 1**이다. 따라서 $E(\overline{X_1}) = \frac{1}{3}, E(\overline{X_2}) = \frac{5}{3}$이고, 두 표본평균의 분산은 각각 $\frac{2/9}{100} = \frac{2}{900}$이다. 이는 원래 관측값보다 표본평균의 흩어짐이 작다는 뜻이다.

두 표본평균의 합: 평균은 $E(\overline{X_1} + \overline{X_2}) = \frac{1}{3} + \frac{5}{3} = 2$이다. 두 표본이 독립이므로 공분산이 0이며, **합의 분산은 두 분산의 합**이다. 따라서 $\mathrm{Var}(\overline{X_1} + \overline{X_2}) = \frac{2}{900} + \frac{2}{900} = \frac{4}{900}$이다.

근사분포: 두 모집단은 유한한 분산을 가지며 표본크기가 각각 100이므로, 중심극한정리에 따라 각 표본평균을 정규분포로 근사한다. 독립인 두 정규확률변수의 합도 정규분포를 따르므로, 두 표본평균의 합의 근사분포는 $N\left(2, \frac{4}{900}\right)$이다. 여기서 두 번째 모수는 표준편차가 아니라 분산이며, 이 분포는 ②와 일치한다.

NoteBase

문 16. 모평균과 모분산이 알려지지 않고 서로 독립인 두 개의 정규 모집단에서 각각 25개씩 표본을 임의추출한 후, 두 모집단의 모분산을 비교할 때 사용하는 분포는?

① 표준정규분포

② 자유도가 24인 t 분포

③ 자유도가 24인 카이제곱분포

④ 자유도가 (24, 24)인 F 분포

정답 4번

간단한 해설

서로 독립인 두 정규모집단의 모분산을 비교할 때는 **두 표본분산의 비가 따르는 F 분포**를 이용한다. 각 표본의 크기가 25이므로 분자와 분모의 자유도는 각각 $25 - 1 = 24$이다.

핵심 풀이

두 표본분산: 각 모집단의 모분산을 σ_1^2, σ_2^2, 표본분산을 S_1^2, S_2^2이라 하자. 모평균을 모르므로 표본평균을 이용하여 표본분산을 구하며, 각 표본에서 자유도는 표본크기보다 1 작다. 정규모집단에서는 $\frac{(n_i - 1)S_i^2}{\sigma_i^2}$이 자유도 $n_i - 1$인 카이제곱분포를 따른다. 따라서 이 문제에서는 두 표본에 대해 각각 자유도 24인 카이제곱변량을 얻는다.

분산 비교: **서로 독립인 두 카이제곱변량을 각각의 자유도로 나눈 값의 비**는 F 분포를 따른다. 두 모집단이 독립이므로 두 표본분산도 독립이다. 모분산이 같다는 귀무가설 $H_0 : \sigma_1^2 = \sigma_2^2$ 아래에서 표본분산의 비 $\frac{S_1^2}{S_2^2}$는 분자 자유도 24, 분모 자유도 24인 F 분포를 따른다. 따라서 두 모분산을 비교할 때 사용하는 분포는 ④의 자유도 (24, 24)인 F 분포이다.

다른 분포와의 구분: 카이제곱분포는 **한 정규모집단의 모분산**에 관한 추론에 이용한다. 두 모분산을 비교하는 이 문제에서는 표본분산 두 개의 비를 사용하므로 F 분포가 필요하다. 표준정규분포나 t 분포를 이용하는 평균에 관한 추론과도 구별한다.

NoteBase

문 17. 성별과 와인 생산지의 선호도가 관련이 있는지를 알아보기 위해 100명을 임의추출한 후 성별과 와인 생산지 선호도를 조사하여 작성한 분할표가 다음과 같다. 성별과 와인 생산지 선호도가 관련이 있는지를 검정하는 카이제곱 검정통계량의 값은?

	프랑스	칠레	미국	행의 합
남	x_{11}	x_{12}	x_{13}	$R_{1\cdot}$
여	x_{21}	x_{22}	x_{23}	$R_{2\cdot}$
열의 합	$C_{\cdot 1}$	$C_{\cdot 2}$	$C_{\cdot 3}$	

① Σ_i=1^2Σ_j=1^3≤ft(x_ij-≤ft((R_i· × C_· j)/(100)))÷≤ft((R_i· × C_· j)/(100))

② Σ_i=1^2Σ_j=1^3≤ft(x_ij-≤ft((R_i· × C_· j)/(100)))÷≤ft((R_i· × C_· j)/(100))^2

③ Σ_i=1^2Σ_j=1^3≤ft(x_ij-≤ft((R_i· × C_· j)/(100)))^2÷≤ft((R_i· × C_· j)/(100))

④ Σ_i=1^2Σ_j=1^3≤ft(x_ij-≤ft((R_i· × C_· j)/(100)))^2÷≤ft((R_i· × C_· j)/(100))^2

정답 3번

간단한 해설

카이제곱 독립성 검정통계량은 각 칸의 **관측도수와 기대도수의 차이를 제곱한 뒤 기대도수로 나눈 값**을 모두 더한다. 두 변수가 독립일 때의 기대도수는 해당 행의 합과 열의 합을 곱하여 전체 인원 100으로 나눈 값이다.

NoteBase

공통 해설

성별과 와인 생산지 선호도의 관련성은 '두 변수가 독립이다'라는 귀무가설을 세워 검정한다. 표의 행은 남·여의 두 범주이고, 열은 프랑스·칠레·미국의 세 범주이다. 각 칸의 관측도수를 x_{ij}, 해당 행의 합을 $R_{i\cdot}$, 해당 열의 합을 $C_{\cdot j}$라 한다.

독립이라면 두 범주에 함께 속할 확률은 각 범주에 속할 확률의 곱이다. 따라서 행 비율과 열 비율을 곱하고 전체 인원 100을 곱하여 기대도수 $E_{ij} = 100 \cdot \frac{R_{i\cdot}}{100} \cdot \frac{C_{\cdot j}}{100} = \frac{R_{i\cdot} C_{\cdot j}}{100}$을 구한다. 이는 두 변수가 독립이라고 할 때 해당 칸에 예상되는 인원이다.

카이제곱 검정통계량은 $\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(x_{ij} - E_{ij})^2}{E_{ij}}$이다. 관측도수와 기대도수의 차이를 제곱하여 양·음의 차이가 서로 상쇄되지 않게 하고, 기대도수로 나누어 각 칸의 빈도 규모를 반영한다. 여섯 칸에서 이렇게 구한 값을 모두 더하므로, 독립일 때 예상한 분포와 실제 관측 분포의 차이가 클수록 통계량도 커진다.

위 식의 기대도수에 $\frac{R_{i\cdot} C_{\cdot j}}{100}$을 넣으면 문제에서 요구한 식이 된다. 핵심은 **분자의 차이는 제곱하고 분모의 기대도수는 제곱하지 않는 것**이다. 차이를 제곱하지 않거나 기대도수의 제곱으로 나누는 식은 이 검정통계량과 다르다.

NoteBase

문 18. 다음은 서로 독립인 두 지역 A와 B에서 실시하는 복지 프로그램 수혜자의 평균연령을 비교하기 위하여 두 지역의 복지 프로그램 수혜자 중에서 A 지역에서 100명, B 지역에서 200명을 임의 추출하여 연령을 조사한 결과이다.

A 지역에서 복지 혜택을 받는 사람의 평균 연령을 μ_A라고 하고, B 지역에서 복지 혜택을 받는 사람의 평균 연령을 μ_B라고 할 때, 가설 $H_0 : \mu_A = \mu_B$ 대 $H_1 : \mu_A \neq \mu_B$를 검정하기 위한 p-값(유의확률)은? (단, Z는 표준정규분포를 따르는 확률변수이다)

	A 지역	B 지역
표본평균	38	41
표본표준편차	8	10

① $P\left(|Z| \geq \left|\frac{41-38}{\sqrt{\frac{10^2}{200}+\frac{8^2}{100}}}\right|\right)$
② $P\left(Z \geq \left|\frac{41-38}{\sqrt{\frac{10^2}{200}+\frac{8^2}{100}}}\right|\right)$
③ $P\left(|Z| \leq \frac{41-38}{\sqrt{\frac{10^2}{200}+\frac{8^2}{100}}}\right)$
④ $P\left(Z \leq \frac{41-38}{\sqrt{\frac{10^2}{200}+\frac{8^2}{100}}}\right)$

정답 1번

간단한 해설

대립가설이 두 모평균이 다르다는 것이므로 **양측검정**이다. 따라서 유의확률은 표준화한 평균 차이의 절댓값이 관측값 이상이 되는 **양쪽 꼬리확률**이다.

NoteBase

계산 과정

표에서 A 지역의 표본크기·표본평균·표본표준편차는 각각 100, 38, 8이고, B 지역은 각각 200, 41, 10이다. 두 표본이 독립이므로 **표본평균 차이의 분산은 각 표본평균의 분산을 더하여** 구한다. 모집단분산을 표본분산으로 추정하면 평균 차이의 표준오차 추정값은 다음과 같다.

$SE = \sqrt{\frac{10^2}{200} + \frac{8^2}{100}} = \sqrt{1.14}$

귀무가설 아래에서는 두 모평균의 차이가 0이다. 두 표본이 모두 크므로 표본평균 차이를 표준오차로 나눈 통계량에 표준정규근사를 적용한다. B 지역에서 A 지역을 뺀 표본평균 차이는 41 − 38 = 3이므로 관측 검정통계량은 다음과 같다.

$z_{\text{obs}} = \frac{41 - 38}{\sqrt{1.14}} \approx 2.81$

대립가설이 $\mu_A \neq \mu_B$이므로 어느 지역의 모평균이 더 큰지 방향을 정하지 않는다. 따라서 귀무가설 아래에서 **관측값만큼 또는 그보다 극단적인 차이가 양쪽 방향으로 나타날 확률**을 합한다. 이 정규근사에서 유의확률은 다음과 같다.

$p = P(|Z| \geq |z_{\text{obs}}|)$

표준정규분포는 0을 중심으로 대칭이므로 $p = 2P(Z \geq |z_{\text{obs}}|)$로도 나타낼 수 있다.

②의 식은 오른쪽 꼬리확률만 포함한다. 따라서 이 문제에서 필요한 양측 유의확률의 절반이다.

③의 식은 관측 통계량이 양수이므로 두 경계 사이의 **중앙확률**을 나타낸다. 이는 양쪽 꼬리확률인 유의확률의 여사건에 해당한다.

④의 식은 양수인 관측 통계량까지의 왼쪽 누적확률이다. 오른쪽 꼬리를 제외한 영역이므로 양측 유의확률을 나타내지 않는다.

따라서 관측 통계량의 절댓값 이상인 양쪽 꼬리확률을 나타내는 ①이 해당한다.

NoteBase

문 19. 다음은 두 인자 A와 B에 대하여 이원배치법을 적용한 실험 결과의 분산분석표 일부이다. 실험에서 인자의 수준조합에 대한 반복실험수는 동일하다. 이에 대한 설명으로 옳지 않은 것은?

요인	제곱합	자유도	평균제곱	$F-$값	$p-$값
인자 A	700	1			0.011
인자 B	900	3			0.040
교호작용 $A\times B$	1,200	(가)		(나)	0.034
오차	800	8			

① (가)는 3이고 (나)는 4이다.

② 분석에 사용한 자료의 수는 모두 15개이다.

③ 인자 A와 인자 B의 수준의 수는 각각 2와 4이다.

④ 인자 A에 대한 주효과는 유의수준 5%에서 유의하다.

정답 2번

간단한 해설

표에서 전체 자유도는 1 + 3 + 3 + 8 = 15이다. 전체 자유도는 자료 수보다 1 작으므로, 분석에 사용한 자료는 15개가 아니라 **16개**이다.

NoteBase

선지별 해설

① O: 교호작용의 자유도는 **두 인자의 자유도를 곱한 값**이므로 (가)는 $1 \times 3 = 3$이다. 평균제곱은 제곱합을 자유도로 나눈 값이므로, 교호작용의 평균제곱은 $\frac{1200}{3} = 400$, 오차평균제곱은 $\frac{800}{8} = 100$이다. F값은 해당 요인의 평균제곱을 오차평균제곱으로 나누므로 $F = \frac{400}{100} = 4$이다. 따라서 (나)는 4이다.

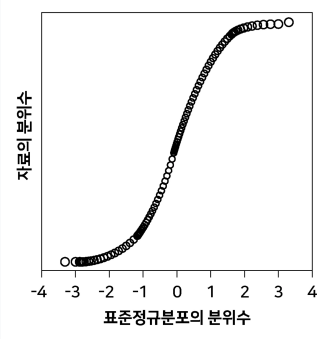
② X: 인자 A, 인자 B, 교호작용, 오차의 자유도를 합하면 전체 자유도는 $1 + 3 + 3 + 8 = 15$이다. **전체 자유도는 자료 수에서 1을 뺀 값**이므로, 자료 수는 $15 + 1 = 16$개이다. 선지의 15개는 전체 자유도와 자료 수를 혼동한 값이다.

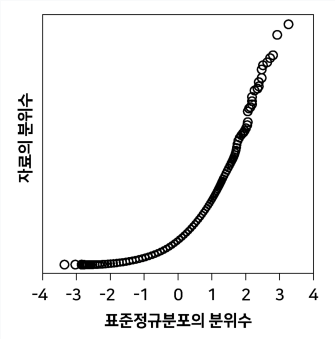
③ O: 각 인자의 자유도는 **수준 수에서 1을 뺀 값**이다. 표에서 인자 A의 자유도는 1이므로 수준 수는 $1 + 1 = 2$이고, 인자 B의 자유도는 3이므로 수준 수는 $3 + 1 = 4$이다.

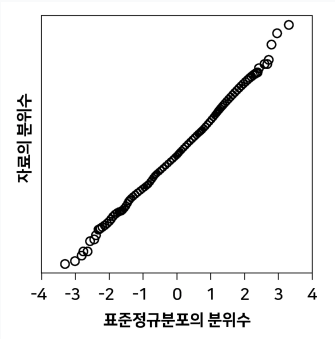
④ O: 표에서 인자 A의 p값은 0.011로, **유의수준 0.05보다 작다**. 따라서 인자 A의 주효과가 없다는 귀무가설을 기각하며, 인자 A의 주효과는 유의수준 5%에서 유의하다.

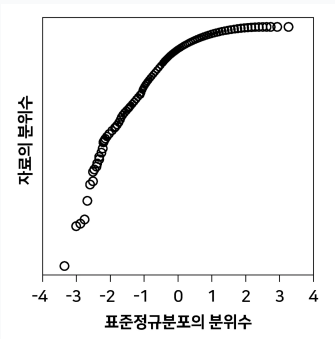
NoteBase

문 20. 정규분포로부터 임의추출한 자료의 정규확률그림(정규분포 분위수대분위수 그림, normal probability plot)은?

① 

② 

③ 

④ 

정답 3번

간단한 해설

정규분포에서 뽑은 자료의 정규확률그림에서는 점들이 **우상향하는 직선 부근**에 놓인다. 표본추출에 따른 작은 흔들림은 있을 수 있으나, 전체적으로 뚜렷하게 휘는 형태와는 구별된다.

NoteBase

도식 해설

가로축은 표준정규분포의 분위수이고, 세로축은 자료의 분위수이다. 정규확률그림은 같은 누적확률에 해당하는 두 분위수를 짝지어 나타낸다. 평균이 μ, 표준편차가 σ인 정규분포의 분위수는 표준정규분포의 분위수 z와 $y = \mu + \sigma z$의 관계를 이룬다. 따라서 정규분포 자료의 점들은 **기울기가 양수인 직선 부근**에 놓인다.

③은 왼쪽 아래에서 오른쪽 위로 점들이 거의 일정한 기울기를 유지하며 늘어서 있다. 양 끝의 일부 점이 조금 벗어나더라도 전체적인 직선 형태가 유지되므로, 정규분포에서 임의추출한 자료의 정규확률그림에 해당한다.

①은 양 끝이 완만하고 가운데가 가파른 S자 형태이다. ②는 오른쪽으로 갈수록 기울기가 커지고, ④는 오른쪽으로 갈수록 기울기가 작아진다. 이들은 모두 **전체적으로 휘는 경향**이 뚜렷하여 정규분포 자료의 전형적인 직선 형태와 다르다.

NoteBase