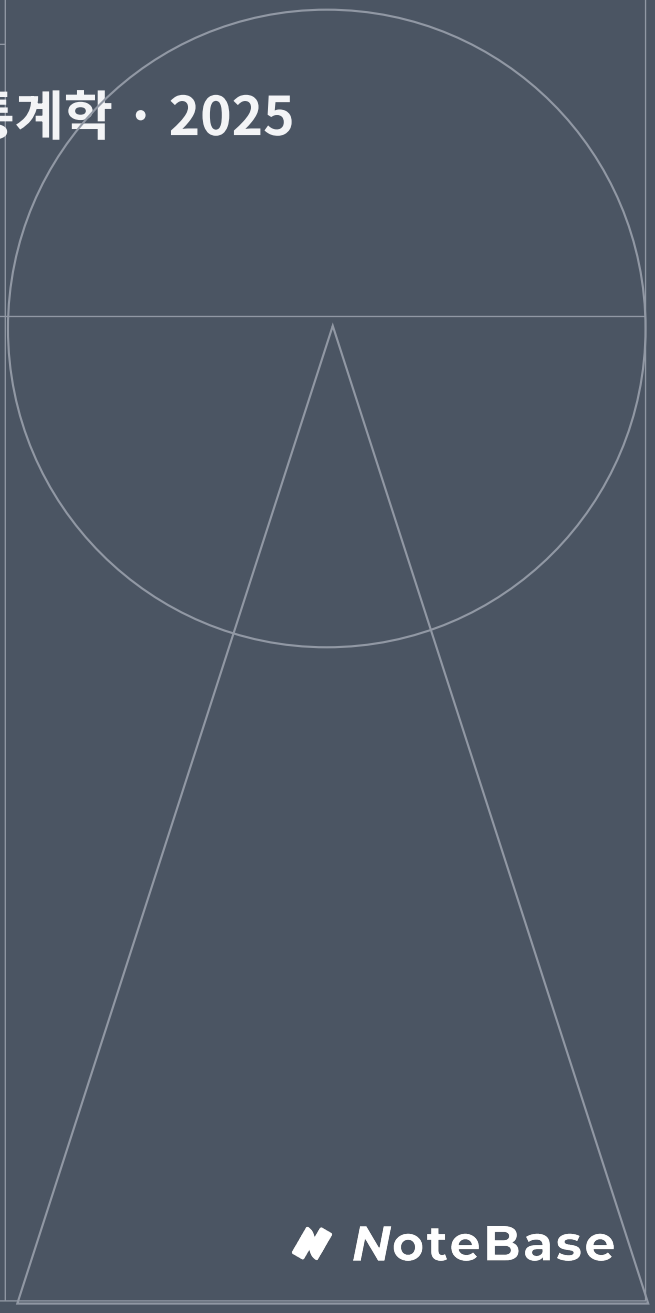
기출OX 해설집

공무원/경찰/공채(경위) · 통계학 · 2025



NoteBase

기출OX 해설집

공무원/경찰/공채(경위) · 통계학 · 2025

공무원/경찰/공채(경위) · 통계학 · 2025

문 1. 다음 중 범주형 자료(질적 자료)에 해당되는 것은?

① 키

② 몸무게

③ 허리둘레

④ 성별

정답 4번

간단한 해설

성별은 수치의 크기가 아니라 **어느 범주에 속하는지를 나타내는 질적 자료**이다. 키·몸무게·허리둘레는 통상 측정한 양의 크기를 나타내는 양적 자료이다.

선지별 해설

범주형 자료는 대상의 종류나 속성을 구분하는 값이고, 양적 자료는 수량이나 측정값의 크기를 나타낸다. 숫자로 기록했는지만이 아니라 그 숫자가 실제 양을 뜻하는지를 기준으로 구별한다.

① X: 키는 길이를 측정한 값이므로 양적 자료이다. 예를 들어 cm 단위의 키 차이는 실제 길이의 차이를 뜻한다. 통상 연속형 자료로 다루며, 키를 별도로 구간화해 범주로 바꾼 경우와는 구별한다.

② X: 몸무게는 무게의 크기를 측정한 양적 자료이다. kg 단위로 측정한 값은 차이나 비율을 비교할 수 있다. 저체중·정상체중 등으로 다시 분류한 범주가 아니라 측정값 자체를 묻고 있다.

③ X: 허리둘레는 신체 둘레의 길이를 측정한 양적 자료이다. cm 등의 단위로 표현되는 크기이며, 측정기기의 정밀도에 따라 반올림해 기록하더라도 자료의 기본 성격이 질적 자료로 바뀌지는 않는다.

④ O: 성별은 사람을 범주로 분류하는 질적 자료이며 통상 명목척도로 다룬다. 범주에 숫자 코드를 부여하더라도 그 숫자는 구분을 위한 표지일 뿐, 양의 크기나 범주 사이의 순서를 의미하지 않는다.

NoteBase

문 2. 다음 중 분포의 산포를 확인하기 위해 사용할 수 있는 측도는?

① 평균

② 표준편차

③ 중앙값

④ 최빈값

정답 2번

간단한 해설

표준편차는 자료가 평균 주위에 얼마나 흩어져 있는지를 나타내는 산포 측도이므로 ②가 옳다.
평균·중앙값·최빈값은 중심 위치를 나타낸다.

선지별 해설

① (X) 평균은 관측값을 모두 더하여 개수로 나눈 값으로, **분포의 중심 위치**를 나타낸다. 평균이 같은 자료라도 흩어진 정도는 다를 수 있으므로 평균만으로 산포를 알 수 없다.

② (O) 표준편차는 분산의 음이 아닌 제곱근으로, **관측값이 평균 주위에 흩어진 정도**를 나타내는 산포 측도이다. 원자료와 같은 단위를 가지며, 값이 클수록 평균에서 떨어진 정도가 크다.

③ (X) 중앙값은 자료를 크기순으로 정렬했을 때 가운데에 놓이는 값으로, **중심 위치의 측도**이다. 자료가 짝수 개이면 가운데 두 값의 평균을 사용하며, 흩어진 정도 자체를 나타내지는 않는다.

④ (X) 최빈값은 가장 자주 나타나는 값으로, **자료를 대표하는 중심 측도**이다. 빈도가 가장 높은 값이 무엇인지 보여 주지만 전체 자료의 산포를 측정하지는 않는다.

NoteBase

문 3. 표본공간 S에서 정의된 두 사건 A와 B가 서로 배반일 때 항상 성립하는 것은?

① $A \cap B = S$

② $A \cap B = \emptyset$

③ $A \cup B = S$

④ $A \cup B = \emptyset$

정답 2번

간단한 해설

서로 배반인 두 사건은 동시에 일어날 수 없으므로 **공통 원소가 없다**. 따라서 $A \cap B = \emptyset$이며, 합집합에 대한 별도의 조건은 요구되지 않는다.

선지별 해설

① X: $A \cap B = S$는 표본공간의 모든 결과가 A와 B에 공통으로 속한다는 뜻이다. 배반이면 두 사건에 **공통으로 속하는 결과가 없어야** 하므로 교집합은 S가 아니라 공집합이다.

② O: 두 사건이 배반이라는 것은 동시에 일어날 수 없다는 뜻이다. 따라서 공통 원소가 없고, 정의에 따라 항상 $A \cap B = \emptyset$가 성립한다.

③ X: 배반이라는 조건만으로 **합집합이 표본공간 전체를 덮을 필요는 없다**. 예를 들어 주사위에서 A가 1이 나오는 사건, B가 2가 나오는 사건이면 서로 배반이지만 3~6이 나오는 결과는 합집합에 속하지 않는다. 배반에 더하여 합집합이 S라는 조건까지 있으면 두 사건은 서로 여사건이다.

④ X: $A \cup B = \emptyset$이려면 **A와 B가 모두 공집합**이어야 한다. 배반은 교집합이 비어 있다는 조건일 뿐 각 사건 자체가 비어 있다는 뜻은 아니다. 주사위의 1이 나오는 사건과 2가 나오는 사건은 배반이지만 합집합은 비어 있지 않다.

따라서 배반일 때 항상 성립하는 것은 ②이다.

NoteBase

문 4. 확률 변수 X가 다음과 같은 확률밀도함수를 가질 때,
$f(x) = \frac{1}{2}e^{-|x-1|}, \quad -\infty < x < \infty$
X의 중앙값은?

① −1

② 0

③ 1

④ 2

정답 3번

간단한 해설

밀도함수는 $|x-1|$에 의해 결정되므로 **1을 중심으로 대칭**이다. 1의 양쪽에 각각 $\frac{1}{2}$의 확률이 놓이므로 중앙값은 1이다.

핵심 풀이

연속확률변수의 중앙값은 누적확률이 $\frac{1}{2}$가 되는 값이다. 주어진 밀도함수 $f(x) = \frac{1}{2}e^{-|x-1|}$는 x와 1 사이의 거리만으로 결정되므로 **1을 중심으로 좌우 대칭**이다.

실제로 임의의 t≥0에 대해 $f(1-t) = f(1+t) = \frac{1}{2}e^{-t}$이다. 전체 확률이 1이고 한 점의 확률은 0이므로, 1의 왼쪽과 오른쪽에 각각 절반의 확률이 놓인다.

직접 확인하면 x≤1에서 $|x-1| = 1-x$이므로 P(X≤1)= ∫ _-∞^1(1)/(2)e^x-1 dx=(1)/(2)이다. 또한 밀도가 모든 실수에서 양수이므로 누적분포함수는 엄격히 증가하여 이 중앙값은 유일하다.

따라서 중앙값은 1이며, ③에 해당한다.

NoteBase

문 5. 가설검정에서 귀무가설이 거짓임에도 불구하고 귀무가설을 채택하는 오류를 범할 확률은?

① 검정력(power)

② 유의수준

③ 유의확률

④ 1-검정력

정답 4번

간단한 해설

거짓인 귀무가설을 기각하지 못하는 것은 **제2종 오류**이다. 같은 대립가설 아래에서 올바르게 기각할 확률인 검정력과 서로 보완되므로, 그 오류 확률은 1-검정력이다.

개념 구분

제2종 오류는 귀무가설이 거짓인데도 이를 기각하지 않는 오류이다. 그 확률을 β라고 한다. 문제의 '귀무가설을 채택'은 여기서는 '기각하지 못함'을 뜻하며, 검정 결과 귀무가설이 참임을 증명했다는 의미는 아니다.

검정력은 특정한 대립가설이 참일 때 귀무가설을 올바르게 기각할 확률이다. 같은 대립가설 아래에서 기각과 비기각은 서로 배반적이며 전체 경우를 이루므로 **검정력과 제2종 오류 확률의 합은 1**이다. 따라서 검정력=1-β, β=1-검정력이다.

유의수준은 참인 귀무가설을 기각하는 제1종 오류 확률을 통제하는 기준이다.

유의확률은 귀무가설 아래에서 관측된 검정통계량만큼 또는 그보다 대립가설을 지지하는 방향으로 극단적인 결과를 얻을 확률이다. 둘 다 거짓인 귀무가설을 기각하지 못하는 확률 자체가 아니다.

따라서 구하는 확률은 1-검정력이며, ④와 일치한다.

NoteBase

문 6. 연속확률변수 X의 확률밀도함수 $f(x)$에 대한 설명으로 옳지 않은 것은?

① 모든 실수 x에 대해 $P(X = x) = 0$

② 모든 실수 x에 대해 $f(x) \geq 0$

③ ∫ _-∞^∞ f(x)dx=1

④ $a < b$인 실수 a, b에 대해
$P(a < X < b) < P(a \leq X \leq b)$

정답 4번

간단한 해설

연속확률변수에서는 **한 점의 확률이 0**이므로 구간의 끝점을 포함해도 확률이 늘어나지 않는다. 따라서 P(a X b)=P(a≤ X≤ b)이다.

선지별 해설

① O: 확률밀도함수를 갖는 연속확률변수에서 한 점에 해당하는 구간의 길이는 0이므로 그 점의 확률도 0이다. 따라서 모든 실수 x에 대해 $P(X = x) = 0$이다. **밀도값 f(x) 자체가 그 점의 확률인 것은 아니다**.

② O: 확률밀도함수는 구간에 적분하여 확률을 구하는 비음수 함수이다. 따라서 f(x)≥0을 만족한다. 다만 **밀도값은 1을 넘을 수 있으며**, 1 이하이어야 하는 것은 구간의 확률이다.

③ O: 확률변수는 전체 실수 범위 안의 어떤 값을 가지므로 전체 확률은 1이다. 이에 따라 밀도함수 아래의 전체 넓이는 $\int_{-\infty}^{\infty} f(x)dx = 1$이다.

④ X: 닫힌 구간은 열린 구간에 양 끝점 a와 b를 추가한 것이다. 연속확률변수는 $P(X = a) = P(X = b) = 0$ 이므로 **끝점 포함 여부가 확률에 영향을 주지 않는다**. 따라서 올바른 관계는 P(a X b)=P(a≤ X≤ b)= ∫ _a^b f(x) dx이며, 엄격한 부등식이 아니다.

따라서 옳지 않은 것은 ④이다.

NoteBase

문 7. 다음 중 정규분포에 대한 설명으로 옳지 않은 것은?

① 확률밀도함수가 대칭이다.

② 확률밀도함수의 봉우리가 하나이다.

③ 왜도가 0보다 크다.

④ 확률밀도함수는 모평균과 모분산의 값에 의해 결정된다.

정답 3번

간단한 해설

정규분포는 평균을 중심으로 좌우 대칭이므로 **왜도가 0**이다. 왜도가 0보다 크다는 설명은 오른쪽으로 치우친 분포의 특징을 잘못 연결한 것이다.

선지별 해설

평균이 μ, 분산이 $\sigma^2 > 0$인 정규분포의 밀도는 $f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$이다. 평균은 중심 위치를, 분산은 퍼짐의 정도를 결정한다.

① O: 밀도함수는 평균과의 차이의 제곱 $(x-\mu)^2$에 의해 결정된다. 평균에서 같은 거리만큼 왼쪽과 오른쪽으로 떨어진 두 점의 밀도가 같으므로 μ를 중심으로 대칭이다.

② O: 밀도는 $x=\mu$에서 가장 크고, 평균에서 멀어질수록 작아진다. 따라서 평균 위치에 하나의 봉우리가 있는 단봉분포이며, 평균·중앙값·최빈값이 모두 μ로 일치한다.

③ X: 정규분포의 왜도는 **0보다 큰 값이 아니라 0**이다. 왜도 $E[(X-\mu)^3]/\sigma^3$에서 대칭적인 좌우 편차의 세제곱이 서로 상쇄되어 분자가 0이 된다. 양의 왜도를 갖는 비대칭분포와 구별해야 한다.

④ O: 정규분포에서는 모평균 μ와 모분산 σ^2를 정하면 위 밀도함수가 완전히 결정된다. 이는 정규분포족의 성질이며, 모든 종류의 분포가 평균과 분산만으로 결정된다는 뜻은 아니다.

문 8. 어떤 복합비료 양 x와 농작물 수확량 y의 단순회귀모형을 적합한 결과 결정계수(R^2) 0.49를 얻었다. 이때 복합비료 양 x와 농작물 수확량 y의 표본상관계수는?

① 0.24

② 0.49

③ 0.70

④ 0.98

정답 3번

간단한 해설

절편이 있는 단순선형회귀에서는 $R^2 = r^2$이므로 **상관계수의 절댓값은 0.70**이다. 부호는 결정계수만으로 알 수 없지만, 제시된 선택지에서 이 관계를 만족하는 값은 0.70이다.

계산 과정

절편을 포함하는 통상적인 단순선형회귀에서 결정계수와 표본상관계수 r의 관계는 $R^2 = r^2$이다. 따라서 $r^2 = 0.49$이고, $|r| = \sqrt{0.49} = 0.70$이다.

결정계수는 제곱값이므로 **상관계수의 부호까지 알려 주지는 않는다**. 회귀기울기가 양수이면 r = 0.70이고 음수이면 r = −0.70이다. 문제의 본문에는 그 부호를 확정할 조건이 없으므로, 비료와 수확량이라는 소재만으로 양의 상관을 단정할 수는 없다.

다만 제시된 선택지에서는 **0.70만 제곱하여 0.49가 된다**. 0.24, 0.49, 0.98은 이 관계를 만족하지 않으므로, 선택지 중 가능한 값은 ③의 0.70이다.

NoteBase

문 9. t 분포에 대한 설명으로 옳지 않은 것은?

① 자유도가 커질수록 확률밀도함수가 정규분포의 확률밀도함수에 가까워진다.

② 자유도가 작아질수록 정규분포에 비해 확률밀도함수의 꼬리가 두꺼워진다.

③ 확률밀도함수가 대칭이다.

④ 중앙값은 0보다 크다.

정답 4번

간단한 해설

중심 t분포는 **0을 중심으로 대칭이며 중앙값이 0**이다. 따라서 중앙값이 0보다 크다는 ④가 옳지 않다.

선지별 해설

① (O) 자유도 k가 커질수록 t분포는 **표준정규분포에 가까워진다.** $T = Z/\sqrt{U/k}$에서 서로 독립인 $Z \sim N(0,1)$, $U \sim \chi^2_k$를 생각하면, 자유도가 커질수록 분모가 1에 가까워져 표준정규분포와의 차이가 줄어든다.

② (O) 자유도가 작을수록 **표준정규분포보다 꼬리가 두꺼워지는 정도가 커진다.** 중심에서 멀리 떨어진 값이 나타날 가능성이 상대적으로 크다는 뜻이다. 따라서 작은 표본에서 평균을 추론할 때 정규분포보다 큰 임계값이 필요할 수 있다.

③ (O) 중심 t분포의 확률밀도함수는 **0을 중심으로 좌우 대칭**이다. 같은 크기의 양수와 음수에서 밀도값이 같고, 0의 양쪽 확률은 각각 0.5이다.

④ (X) 중앙값은 0보다 큰 것이 아니라 **자유도와 관계없이 0**이다. 연속인 중심 t분포에서 $P(T \le 0) = 0.5$이기 때문이다. 자유도가 작아 평균이 존재하지 않는 경우에도 중앙값은 0으로 정의된다.

NoteBase

문 10. 이산확률변수 X가 다음과 같은 확률질량함수를 가질 때,
$f(x)=\frac{5-x}{10}, \quad x=1,2,3,4$
X의 평균은?

① 1

② 2

③ 3

④ 4

정답 2번

간단한 해설

평균은 각 값을 그 확률로 가중한 합이다. **값과 해당 확률을 곱해 더하면** $E[X]=\frac{4+6+6+4}{10}=2$가 된다.

계산 과정

이산확률변수의 평균은 가능한 각 값에 그 값이 나올 확률을 곱하여 더한 기댓값이다. 따라서 $E[X]=\sum_{x} x f(x)$를 사용한다.

주어진 $f(x)=\frac{5-x}{10}$에 $x=1,2,3,4$를 대입하면 확률은 각각 $\frac{4}{10}, \frac{3}{10}, \frac{2}{10}, \frac{1}{10}$이다. 이 확률들은 모두 음이 아니고 합이 1이다.

각 값을 확률로 가중하여 더하면
aligned&E[X]=1·(4)/(10)+2·(3)/(10)&+3·(2)/(10)+4·(1)/(10)=(4+6+6+4)/(10)=2aligned
이다. 확률이 서로 다르므로 네 값을 같은 비중으로 평균내면 안 된다.

따라서 X의 평균은 2이며, ②에 해당한다.

NoteBase

문 11. 다음은 60명에게는 온라인 수업을, 40명에게는 강의실 수업을 받게 하고 학기 말에 시험을 보게 하여 시험 합격 여부를 나타낸 표이다. 강의실 수업과 온라인 수업의 합격률이 같다는 귀무가설 하에서 공통 합격률 p의 추정값은?

	강의실 수업	온라인 수업
합격	30	30
불합격	10	30
계	40	60

① 0.40

② 0.50

③ 0.60

④ 0.70

정답 3번

간단한 해설

합격률이 같다는 귀무가설 아래에서는 **두 집단의 합격자와 응시자를 합쳐** 공통 비율을 추정한다.
p=(30+30)/(40+60)=0.60이다.

계산 과정

귀무가설은 강의실 수업과 온라인 수업의 모집단 합격률이 공통으로 p라는 것이다. 이 가정 아래에서는 두 집단을 따로 추정하지 않고 **전체 합격자 수를 전체 응시자 수로 나눈 통합 비율**로 p를 추정한다.

표에서 강의실 수업은 합격 30명, 불합격 10명으로 총 40명이고, 온라인 수업은 합격 30명, 불합격 30명으로 총 60명이다. 따라서 전체 합격자는 60명, 전체 응시자는 100명이다.

공통 합격률의 추정값은 p=(30+30)/(40+60)=(60)/(100)=0.60이다. 두 집단의 크기가 다르므로 개별 합격률 0.75와 0.50을 단순 평균하지 않고 **표본 크기로 가중**하여 통합한다.

따라서 공통 합격률의 추정값은 0.60이며, 선택지는 ③이다.

NoteBase

문 12. 평균이 μ, 분산이 σ^2인 모집단으로부터 크기 $2n$인 확률표본 $X_1, X_2, \ldots, X_{2n}$을 추출하였을 때 표본평균 $\overline{X}$의 분산 $Var(\overline{X})$는?

① $\frac{\sigma^2}{n}$
② $\frac{\sigma^2}{2n}$
③ $\frac{\sigma^2}{3n}$
④ $\frac{\sigma^2}{4n}$

정답 2번

간단한 해설

독립·동일분포 표본의 평균은 **모분산을 표본 크기로 나눈 분산**을 갖는다. 표본 크기가 2n이므로 표본평균의 분산은 $\frac{\sigma^2}{2n}$이다.

핵심 풀이

크기가 2n인 확률표본은 서로 독립이고 동일한 모집단 분포를 따르므로 각 관측값의 분산은 σ^2이며 서로 다른 관측값 사이의 공분산은 0이다. 표본평균은 X=(1)/(2n) Σ _i=1^2nX_i이다.

독립인 확률변수의 합의 분산은 각 분산의 합이고, 상수배의 분산에서는 상수를 제곱한다. 따라서 Var(X)=(1)/((2n)^2) Σ _i=1^2nVar(X_i)=(2nσ^2)/(4n^2)=(σ^2)/(2n)이다.

핵심은 **모분산을 실제 표본 크기 2n으로 나누는 것**이다. 모집단 평균 μ는 이 분산식에 들어가지 않으며, 이 결과에 모집단의 정규성은 필요하지 않다. 표본평균의 표준편차인 표준오차와 분산도 구별해야 한다.

따라서 표본평균의 분산은 $\frac{\sigma^2}{2n}$이며, ②와 일치한다.

NoteBase

문 13. 두 확률변수 X, Y와 상수 a, b, c, d에 대해 공분산 $Cov(aX+b, cY+d)$의 값은?

① $Cov(X, Y)$

② $Cov(X, Y) + bd$

③ $acCov(X, Y)$

④ $acCov(X, Y) + bd$

정답 3번

간단한 해설

공분산은 평균에서의 편차로 계산하므로 **상수를 더해도 변하지 않고, 각 변수의 배율은 곱으로 반영**된다. 따라서 $\mathrm{Cov}(aX+b, cY+d) = ac\mathrm{Cov}(X, Y)$이다.

핵심 풀이

공분산이 존재한다고 할 때, 공분산은 두 변수가 각자의 평균에서 벗어난 정도를 곱한 값의 평균이다. 즉 $\mathrm{Cov}(U, V) = E[(U - E[U])(V - E[V])]$이다.

$U = aX + b$이면 $U - E[U] = a(X - E[X])$이고, $V = cY + d$이면 $V - E[V] = c(Y - E[Y])$이다. 따라서 aligned&Cov(aX+b,cY+d)&=acE[(X-E[X])(Y-E[Y])]=acCov(X,Y)aligned
이다.

더하는 상수 b, d는 중심화 과정에서 사라지고, 곱하는 상수 a, c는 공분산에 곱해진다. 그러므로 배율 ac가 빠진 ①·②와 불필요한 bd가 더해진 ②·④는 일반적인 변환 공식이 아니다. 구한 식에 해당하는 선택지는 ③이다.

NoteBase

문 14. 2개의 설명변수로 반응변수를 예측하는 다중선형회귀모형을 적합하여 다음 분산분석표를 얻었다. 결정계수 (R^2)는?

요인	제곱합	자유도	평균제곱	F 값
회귀	1,400	2	700	11.67
잔차	600	10	60	
계	2,000	12		

① 0.30

② 0.70

③ 0.09

④ 0.64

정답 2번

간단한 해설

결정계수는 **회귀제곱합을 전체제곱합으로 나눈 값**이다. $R^2 = 1400/2000 = 0.70$이므로 정답은 ②이다.

계산 과정

결정계수는 전체 변동 중 회귀모형이 설명하는 비율이다. 회귀제곱합을 SSR, 잔차제곱합을 SSE, 전체제곱합을 SST라고 하면 $R^2 = SSR/SST = 1 - SSE/SST$이다.

표에서 $SSR = 1400$, $SSE = 600$, $SST = 2000$이므로 $1400 + 600 = 2000$이 성립한다. 따라서 $R^2 = 1400/2000 = 0.70$이다. 같은 값은 $1 - 600/2000 = 0.70$으로도 확인할 수 있다.

결정계수는 단위가 없는 비율이며, 이 모형이 표본에서 반응변수 전체 변동의 **70%**를 설명한다는 뜻이다. 구하는 값은 0.70이므로 정답은 ②이다.

주의할 점은 수정결정계수와의 구별이다. 자유도를 반영하면 $\overline{R}^2 = 1 - (600/10)/(2000/12) = 0.64$이지만, 문제는 수정결정계수가 아닌 R^2를 물었다. 또한 $600/2000 = 0.30$은 설명되지 않은 변동의 비율이다.

NoteBase

문 15. $X_1, \cdots, X_{30}$**이 균일분포** $U(0, 1)$**로부터 추출된 확률표본일 때,** $Y = X_1 + \cdots + X_{30}$**의 중심극한정리에 의한 근사 분포는?**

① $N\left(15, \frac{15}{2}\right)$

② $N\left(30, \frac{15}{2}\right)$

③ $N\left(15, \frac{5}{2}\right)$

④ $N(30, 30)$

정답 3번

간단한 해설

독립인 균일확률변수 30개의 합은 평균이 $30 \cdot \frac{1}{2} = 15$, 분산이 $30 \cdot \frac{1}{12} = \frac{5}{2}$이다. **중심극한정리**에 따라 $N\left(15, \frac{5}{2}\right)$로 근사한다.

계산 과정

확률표본이므로 $X_1, \ldots, X_{30}$은 서로 독립이며 같은 균일분포를 따른다. 구하려는 변수는 표본평균이 아니라 **30개 확률변수의 합**이므로, 각 변수의 평균과 분산을 각각 더한다.

균일분포 $U(a, b)$의 평균은 $\frac{a+b}{2}$, 분산은 $\frac{(b-a)^2}{12}$이다. 따라서 $U(0, 1)$에서는 $E[X_i] = \frac{1}{2}$, $\mathrm{Var}(X_i) = \frac{1}{12}$이다.

합의 평균은 $E[Y] = 30 \cdot \frac{1}{2} = 15$이다. 독립성에 의해 공분산 항이 0이므로 합의 분산은 $\mathrm{Var}(Y) = 30 \cdot \frac{1}{12} = \frac{5}{2}$이다.

중심극한정리에 따라 $\frac{Y-15}{\sqrt{5/2}}$의 분포를 표준정규분포로 근사한다. 따라서 Y의 근사 분포는 $N\left(15, \frac{5}{2}\right)$이며, 여기서 두 번째 모수는 표준편차가 아니라 분산이다. 이는 유한한 30개 합의 정확한 분포가 정규분포라는 뜻은 아니다.

따라서 근사 분포 $N\left(15, \frac{5}{2}\right)$는 ③과 일치한다.

NoteBase

문 16. 서로 독립인 두 확률변수 X와 Y에 대한 다음 설명 중 옳지 않은 것은? 단, a와 b는 상수이다.

① $E(aX+bY)=aE(X)+bE(Y)$

② $E(aX-bY)=aE(X)-bE(Y)$

③ $Var(aX+bY)=a^2Var(X)+b^2Var(Y)$

④ $Var(aX-bY)=a^2Var(X)-b^2Var(Y)$

정답 4번

간단한 해설

분산에서는 상수 계수가 **제곱**되고, 독립인 두 변수의 공분산은 0이다. 따라서
$\mathrm{Var}(aX-bY)=a^2\mathrm{Var}(X)+b^2\mathrm{Var}(Y)$로 분산 항을 더한다.

선지별 해설

아래에서는 문제의 기댓값과 분산이 유한하게 존재하는 경우의 성질을 사용한다.

① O: 기댓값은 선형이므로 $E(aX+bY)=aE(X)+bE(Y)$이다. 이 성질은 **독립 여부와 관계없이** 성립한다.

② O: 선형성에서 Y의 계수를 −b로 두면 $E(aX-bY)=aE(X)-bE(Y)$가 된다. 따라서 차의 기댓값에서는 빼기 부호가 그대로 유지된다.

③ O: 일반적으로
aligned&Var(aX+bY)=a^2Var(X)&+b^2Var(Y)+2abCov(X,Y)aligned
이다. X와 Y가 **독립이면 공분산이 0**이므로 제시된 두 분산 항의 합이 된다.

④ X: $aX-bY$에서 Y의 계수는 −b이지만 분산에는 그 계수의 제곱이 들어가므로 $(-b)^2=b^2$이다. 독립 조건 아래의 올바른 식은 $\mathrm{Var}(aX-bY)=a^2\mathrm{Var}(X)+b^2\mathrm{Var}(Y)$이다. 즉 **차의 분산도 두 분산 항을 더한다**.

따라서 옳지 않은 것은 ④이다.

NoteBase

문 17. 어느 옷가게에 1시간당 남성과 여성 방문 고객 수는 각각 평균 4명, 6명이고 서로 독립인 포아송분포를 따른다. 이때 1시간당 남성과 여성을 모두 합한 전체 손님의 수를 확률변수 X라 할 때, 확률변수 X의 분산은?

① 20

② 16

③ 14

④ 10

정답 4번

간단한 해설

포아송분포는 **평균과 분산이 같고**, 독립인 두 확률변수의 합은 분산도 더할 수 있다. 따라서 전체 고객 수의 분산은 4 + 6 = 10이다.

계산 과정

1시간 동안의 남성 방문 고객 수를 M, 여성 방문 고객 수를 F라 두면 $M \sim \text{Poisson}(4)$, $F \sim \text{Poisson}(6)$이다. 포아송분포는 **평균과 분산이 모수와 같다**. 따라서 $\text{Var}(M) = 4$, $\text{Var}(F) = 6$이다.

전체 고객 수는 $X = M + F$이다. 합의 분산은 $\text{Var}(X) = \text{Var}(M) + \text{Var}(F) + 2\text{Cov}(M, F)$이고, 두 고객 수가 **독립이므로 공분산은 0**이다. 따라서 분산은 4 + 6 = 10이다.

구한 분산 10에 해당하는 선택지는 ④이다. 20, 16, 14는 각 집단의 포아송 분산과 독립성으로 계산한 합의 분산에 맞지 않는다.

NoteBase

문 18. 확률변수 $Z_1, Z_2, \cdots, Z_k$가 각각 표준정규분포 $N(0,1)$을 따르고 서로 독립일 때, $\sum_{i=1}^{k} Z_i^2$의 분포는?

① 자유도 k인 t 분포

② 자유도 k인 카이제곱(χ^2) 분포

③ 자유도 $(1, k)$인 F 분포

④ 자유도 $(k, 1)$인 F 분포

정답 2번

간단한 해설

독립인 표준정규변수 k개의 제곱합은 자유도 k인 카이제곱분포를 따른다. 따라서 정답은 ②이다.

핵심 풀이

서로 독립인 표준정규 확률변수 k개의 제곱합은 자유도가 k인 카이제곱분포를 따른다. 따라서 $Q = \sum_{i=1}^{k} Z_i^2 \sim \chi_k^2$이며, 해당하는 선택지는 ②이다.

핵심 조건은 각 변수가 평균 0·분산 1인 표준정규분포를 따르고 서로 독립이라는 점이다. 독립적인 제곱 항 하나가 자유도 1에 해당하므로 k개를 더한 자유도는 k이다.

t분포는 독립인 표준정규변수 Z와 $U \sim \chi_k^2$를 이용한 $Z/\sqrt{U/k}$의 분포이다. F분포는 서로 독립인 두 카이제곱변수를 각 자유도로 나눈 값의 비로 정의한다. 문제의 식은 비율이 아니라 제곱합이므로 t분포나 F분포가 아니다.

NoteBase

문 19. 귀무가설이 사실이라는 가정하에서 검정통계량 값이 대립가설을 지지하는 방향으로 관측될 확률은?

① 유의수준(α)

② 제2종 오류를 범할 확률(β)

③ 검정력(power)

④ p-값(p-value)

정답 4번

간단한 해설

귀무가설이 참이라는 조건에서 관측값만큼 또는 더 극단적인 검정통계량이 나올 확률이 p-값이다. 귀무가설 자체가 참일 확률과는 다르다.

개념 구분

문항에서 가리키는 개념은 p-값이다. 정확히는 **귀무가설이 참일 때 관측값만큼 또는 그보다 더 대립가설을 지지하는 검정통계량이 나올 확률**을 뜻한다. 단순히 어느 방향에 속할 확률이나, 관측된 한 점의 확률을 뜻하지 않는다.

예를 들어 검정통계량 T가 클수록 대립가설을 지지하는 단측검정에서 관측값이 t_{obs}이면 $p = P_{H_0}(T \geq t_{\text{obs}})$이다. 작은 값이 대립가설을 지지하면 왼쪽 꼬리를, 양측검정이면 검정에서 정한 양쪽의 극단성 기준을 적용한다.

유의수준 α는 귀무가설이 참인데도 이를 기각하는 제1종 오류확률에 대해 미리 정하는 기준이다. 반면 p-값은 실제 관측한 자료에 따라 계산되며, 통상 $p \leq \alpha$이면 귀무가설을 기각한다.

제2종 오류확률 β는 대립가설이 참일 때 귀무가설을 기각하지 못할 확률이다. 검정력은 같은 대립가설하에서 귀무가설을 기각할 확률 $1 - \beta$이다. 둘 다 어떤 대립가설의 참값을 전제하느냐에 따라 달라지며, 귀무가설하의 관측값 극단성을 나타내는 p-값과 다르다.

따라서 문항에 해당하는 것은 ④ p-값이다. 이는 귀무가설이 참일 확률을 계산한 값이 아니라, 귀무가설을 가정하여 자료의 극단성을 평가한 값이다.

NoteBase

문 20. X의 확률밀도함수가 다음과 같다.
$f(x) = cx + 1,\ 0 \le x \le 2$
이때, f(x)가 확률밀도함수가 되기 위한 c는?

① −1

② −0.5

③ 0.5

④ 1

정답 2번

간단한 해설

확률밀도함수의 **전체 적분값은 1**이어야 하므로 $2c + 2 = 1$에서 c = −0.5를 얻는다. 이때 밀도는 구간 내에서 1부터 0까지 감소하여 음수가 되지 않는다.

계산 과정

확률밀도함수는 정의된 구간에서 음수가 아니어야 하고, 전체 적분값이 1이어야 한다. 주어진 지지구간은 0≤x≤2이므로 먼저 $\int_0^2 (cx + 1)dx = 1$을 적용한다.

적분하면 $\left[\frac{c}{2}x^2 + x\right]_0^2 = 2c + 2 = 1$이다. 따라서 $2c = -1$이고 c=-12=-0.5를 얻는다.

이 값을 대입한 밀도는 $f(x) = 1 - \frac{x}{2}$이다. 구간의 왼쪽 끝에서 1, 오른쪽 끝에서 0이며 그 사이에서 선형으로 감소하므로 **주어진 구간 전체에서 음수가 아니다**. 따라서 정규화 조건뿐 아니라 비음수 조건도 만족한다.

따라서 c = −0.5이며, ②와 일치한다.

NoteBase

문 21. 확률변수 Z는 평균이 0이고 표준편차가 1인 표준정규분포를 따른다고 한다. 이때 $P(Z = 0)$은?

① 0

② 0.5

③ 0.95

④ 0.975

정답 1번

간단한 해설

표준정규분포는 연속분포이므로 **한 점의 확률은 0**이다. 평균이 0이라는 사실은 $P(Z = 0)$을 0.5로 만들지 않는다.

핵심 풀이

표준정규분포는 연속분포이므로 확률은 밀도곡선 아래의 구간 넓이로 구한다. 한 점 0은 구간의 폭이 0이므로 **정확히 그 값이 될 확률은 0**이다. 따라서 $P(Z = 0) = \int_0^0 f(z)dz = 0$이다.

평균이 0이고 좌우 대칭이라는 사실에서 얻는 0.5는 $P(Z \leq 0)$ 또는 $P(Z \geq 0)$의 값이다. 이는 한 점의 확률이 아니라 **분포의 한쪽 절반에 해당하는 누적 확률**이다. 또한 0에서 밀도곡선의 높이가 양수라고 해도 그 높이를 한 점의 확률로 해석하지 않는다.

따라서 구하는 확률은 0이며, 선택지는 ①이다.

NoteBase

문 22. 성공확률이 0.2인 베르누이 시행을 독립적으로 100번 반복하였다. 성공횟수를 확률변수 X라 할 때, X의 평균과 표준편차는?

① 평균: 20, 표준편차: 4

② 평균: 20, 표준편차: 16

③ 평균: 16, 표준편차: 4

④ 평균: 16, 표준편차: 16

정답 1번

간단한 해설

독립적인 베르누이 시행의 성공횟수는 이항분포를 따르므로 평균은 $100 \times 0.2 = 20$, 분산은 $100 \times 0.2 \times 0.8 = 16$이다. **표준편차는 분산의 제곱근**이므로 4이다.

계산 과정

각 시행이 독립이고 성공확률 $p = 0.2$가 일정하므로, $n = 100$회의 성공횟수 X는 $X \sim \text{Binomial}(100, 0.2)$인 이항분포를 따른다.

이항분포의 평균은 $E[X] = np$이다. 따라서 평균은 $100 \times 0.2 = 20$이다.

분산은 $\text{Var}(X) = np(1-p)$이므로 $100 \times 0.2 \times 0.8 = 16$이다. 문제에서 요구한 것은 분산이 아니라 **분산의 양의 제곱근인 표준편차**이므로, $\sigma_X = \sqrt{16} = 4$이다.

따라서 평균 20, 표준편차 4인 ①에 해당한다. ②는 분산 16을 표준편차로 잘못 놓았고, ③·④는 분산 16을 평균으로 잘못 놓은 값이다.

NoteBase

문 23. 앞면이 나올 확률은 0.4, 뒷면이 나올 확률은 0.6인 동전을 2번 던질 때, 앞면과 뒷면이 각각 한 번씩 나올 확률은?

① 0.52

② 0.48

③ 0.24

④ 0.76

정답 2번

간단한 해설

앞면과 뒷면이 한 번씩 나오는 순서는 **앞면–뒷면과 뒷면–앞면 두 가지**이다. 확률은 $2 \times 0.4 \times 0.6 = 0.48$이므로 정답은 ②이다.

계산 과정

통상적인 동전 던지기 모형에서 두 시행은 서로 독립이다. 앞면을 H, 뒷면을 T라 하면 각각 한 번씩 나오는 경우는 HT**와** TH **두 가지**이다.

독립이므로 각 순서의 확률은 곱으로 구한다. $P(HT) = 0.4 \times 0.6 = 0.24$, $P(TH) = 0.6 \times 0.4 = 0.24$이다.

두 경우는 동시에 일어날 수 없는 서로 배반인 사건이므로 확률을 더한다. 따라서 P(HT 또는 TH)=0.24+0.24=0.48이다. 확률은 단위 없는 값이며, 백분율로는 48%이다. 정답은 ②이다.

0.24는 한 가지 순서만 센 값이다. 두 번 모두 같은 면이 나올 확률은 $0.4^2 + 0.6^2 = 0.52$이므로, 여사건으로 계산해도 $1 - 0.52 = 0.48$로 일치한다.

NoteBase

문 24. 어느 항공기의 도착 지연시간(단위: 분)은 0분과 60분 사이의 균일분포를 따른다. 이 항공기의 도착 지연시간을 X라 할 때 X의 평균과 분산은?

① 평균: 20, 분산: 300

② 평균: 20, 분산: 120

③ 평균: 30, 분산: 300

④ 평균: 30, 분산: 120

정답 3번

간단한 해설

균일분포 $U(0, 60)$의 평균은 **구간의 중점**인 30분이다. 분산은 $\frac{60^2}{12} = 300$분^{2}이다.

계산 과정

지연시간이 0분부터 60분까지의 연속균일분포를 따르므로 $X \sim U(0, 60)$이다. 균일분포는 구간 안의 밀도가 일정하며, 같은 길이의 부분구간은 같은 확률을 갖는다.

일반적으로 $X \sim U(a, b)$이면 평균은 구간의 중점 $E[X] = \frac{a+b}{2}$이다. 따라서 $E[X] = \frac{0+60}{2} = 30$분이다.

균일분포의 분산은 $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$이다. 구간의 길이가 60분이므로 $\mathrm{Var}(X) = \frac{(60-0)^2}{12} = \frac{3600}{12} = 300$분^{2}이다. **분산은 구간 길이의 제곱을 12로 나눈 값**이며, 표준편차나 평균과 혼동하지 않아야 한다.

따라서 평균 30, 분산 300인 ③과 일치한다. 평균의 단위는 분이고 분산의 단위는 분^{2}이다.

NoteBase

문 25. 서로 독립인 두 사건 A와 B에 대해 항상 성립하는 것은?

① $P(A \cap B) = P(A)P(B)$

② $P(A \cap B) = P(A) + P(B)$

③ $P(A \cup B) = P(A)P(B)$

④ $P(A \cup B) = P(A) + P(B)$

정답 1번

간단한 해설

독립인 두 사건은 **동시에 발생할 확률이 각 확률의 곱**과 같다. 따라서 $P(A \cap B) = P(A)P(B)$이며, 독립과 배반을 혼동해서는 안 된다.

선지별 해설

① O: 두 사건이 독립이라는 것은 **교집합의 확률이 각 사건의 확률의 곱**과 같다는 뜻이다. 따라서 정의에 의해 $P(A \cap B) = P(A)P(B)$가 항상 성립한다.

② X: 독립인 두 사건이 모두 발생할 확률은 확률의 합이 아니라 곱이다. 예를 들어 독립이고 각 확률이 1/2이면 교집합의 확률은 12×12=14이지만 두 확률의 합은 1이다.

③ X: 확률의 곱은 독립인 두 사건의 교집합에 해당한다. 합집합은 중복 부분을 빼야 하므로 $P(A \cup B) = P(A) + P(B) - P(A)P(B)$이다. 각각 확률이 1/2인 독립 사건에서는 합집합의 확률은 3/4이고 곱은 1/4이다.

④ X: 두 확률을 그대로 더하려면 **교집합의 확률이 0**이어야 한다. 독립이라는 조건은 배반을 의미하지 않으므로 일반적으로 $P(A)P(B)$를 빼야 한다. 두 사건의 확률이 모두 양수이면 독립에 의해 교집합의 확률도 양수이므로 제시된 덧셈식은 성립하지 않는다.

따라서 항상 성립하는 것은 ①이다.

NoteBase

문 26. 다음 중 선형회귀분석의 가정으로 옳지 않은 것은?

① 설명변수와 반응변수 간의 선형 관계가 있다.

② 오차항의 분산이 설명변수의 값과 관계없이 일정하다.

③ 모든 오차항이 서로 독립이다.

④ 설명변수들이 서로 독립이다.

정답 4번

간단한 해설

회귀분석에서 구별해야 할 것은 **오차항 사이의 독립성**과 설명변수 사이의 관계이다. 설명변수들은 서로 관련되어도 되며, 회귀계수의 유일한 추정을 위해 필요한 것은 **완전 다중공선성의 배제**이지 통계적 독립이 아니다.

선지별 해설

① O: 기본적인 선형회귀모형은 설명변수가 주어졌을 때 반응변수의 **조건부 평균이 선형식으로 표현**된다고 가정한다. 관측값 자체가 모두 직선 위에 있어야 한다는 뜻은 아니며, 평균에서 벗어난 부분은 오차항으로 나타낸다.

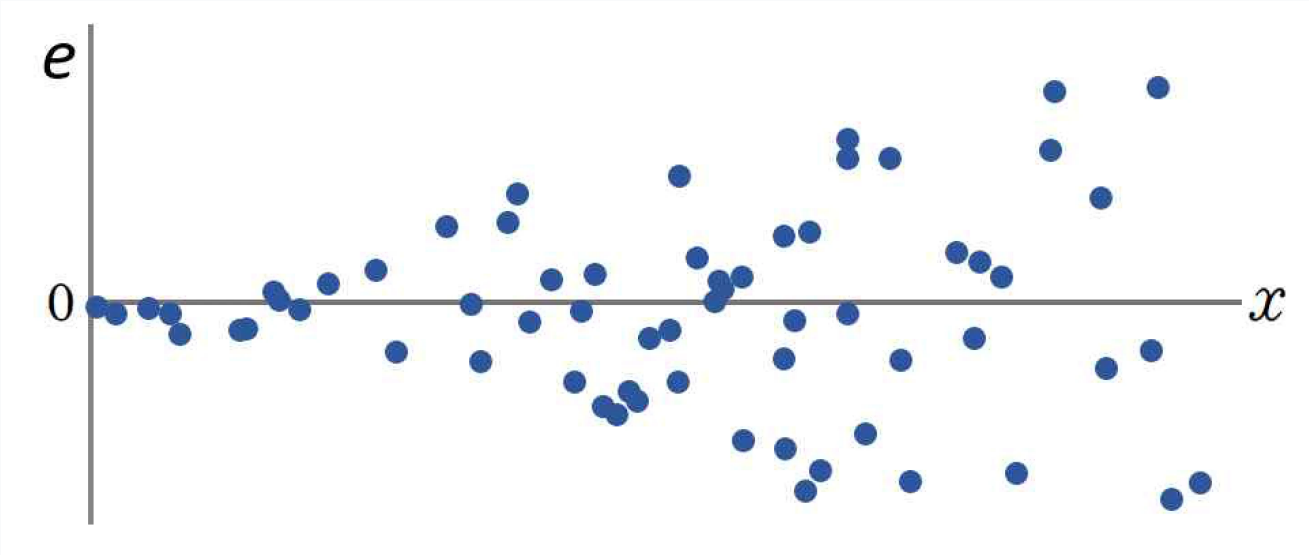
② O: 등분산성은 설명변수의 값에 따라 오차의 분산이 달라지지 않는다는 가정이다. 조건부 분산을 $\mathrm{Var}(\varepsilon_i \mid X) = \sigma^2$로 두므로, 설명변수 수준에 따라 잔차의 퍼짐이 체계적으로 커지는 현상은 이 가정에 어긋난다.

③ O: 통상적인 고전적 회귀모형에서는 서로 다른 관측의 오차항이 독립이라고 가정한다. 이는 한 관측의 오차가 다른 관측의 오차와 연계되어 발생하지 않는다는 뜻이다. 다만 최소제곱추정량의 일부 성질에는 독립보다 약한 오차 간 무상관 조건만으로 충분하다.

④ X: **설명변수끼리 통계적으로 독립일 필요는 없다**. 다중회귀에서 설명변수 사이에 상관관계가 있어도 분석할 수 있다. 회귀계수를 유일하게 식별하려면 한 설명변수가 나머지 설명변수의 정확한 선형결합으로 표현되는 완전 다중공선성을 배제해야 하며, 이는 설명변수 간 독립과 다른 조건이다.

NoteBase

문 27. 선형회귀분석 결과 나타난 설명변수 x와 잔차 e의 그래프가 아래 그림과 같을 때, 다음 설명 중 가장 적절한 것은?



① 현재의 선형회귀모형이 적절하다.

② 오차항 사이에 상관관계가 있으므로 다른 모형을 고려해야 한다.

③ 설명변수 값이 커짐에 따라 오차항의 분산이 커지므로 다른 모형을 고려해야 한다.

④ 설명변수와 반응변수가 선형함수로 표현될 수 없으므로 다른 모형을 고려해야 한다.

정답 3번

간단한 해설

그림에서는 설명변수 x가 커질수록 잔차가 0의 위아래로 더 넓게 퍼진다. 이는 **오차분산이 일정하지 않은 이분산성**을 시사하므로 분산 구조를 고려한 분석이 필요하다.

NoteBase

선지별 해설

가로축은 설명변수 x, 세로축은 잔차 e이다. 작은 x에서는 점들이 0 주변에 좁게 모이고, 큰 x에서는 양수와 음수 방향으로 넓게 퍼지는 부채꼴 모양이 보인다. 이 그림의 핵심은 잔차의 중심보다 퍼짐의 변화이다.

① X: 통상적인 선형회귀모형의 등분산 가정은 설명변수의 값이 달라도 오차분산이 일정하다는 것이다. 그림에서는 잔차의 수직 퍼짐이 커지므로, 등분산을 가정한 현재 모형을 그대로 적절하다고 보기 어렵다. 다만 이 사실만으로 평균 관계의 선형성까지 반드시 틀렸다고 할 수는 없다.

② X: 오차항 사이의 상관관계는 서로 다른 관측치의 오차가 연관되어 있다는 뜻이다. 이를 진단하려면 관측 순서나 시간에 따른 잔차 패턴, 잔차 간 상관 등 별도의 정보를 확인해야 한다. 설명변수에 따라 잔차의 폭이 넓어지는 그림만으로 오차들 사이의 상관을 결론 내릴 수 없다.

③ O: x가 증가할수록 잔차의 퍼짐이 커지므로 **설명변수 값에 따라 오차분산이 증가하는 이분산성**이 가장 직접적인 해석이다. 분산함수를 고려한 모형이나 적절한 가중최소제곱 등을 검토할 수 있다. 평균 모형이 적절하다면 이분산에 강건한 표준오차로 추론을 보완하는 방법도 있으므로, 반드시 비선형 평균 모형으로 바꾸어야 한다는 뜻은 아니다.

④ X: 평균 관계가 선형으로 적절하지 않으면 잔차의 중심이 설명변수에 따라 곡선 형태로 변하는 등의 체계적인 패턴을 살핀다. 제시된 그림에서 뚜렷한 특징은 오른쪽으로 갈수록 잔차의 폭이 넓어지는 것이므로, 설명변수와 반응변수의 평균 관계가 선형함수로 표현될 수 없다고 단정하는 것은 적절하지 않다.

NoteBase

문 28. 다음은 무엇에 대한 설명인가?

$X_1, \cdots, X_n$이 평균 μ와 분산 σ^2을 가지는 모집단에서 추출한 확률표본이고 표본 크기 n이 충분히 클 때, 표본평균 $\overline{X}$의 표본분포는 정규분포 $N\left(\mu, \frac{\sigma^2}{n}\right)$에 근사한다.

① 베이즈정리

② 중간값정리

③ 바수(Basu)정리

④ 중심극한정리

정답 4번

간단한 해설

중심극한정리는 독립·동일분포와 유한한 분산 등의 조건에서 **표본평균의 분포가 큰 표본에서 정규분포에 가까워짐**을 설명한다. 정규근사의 대상은 모집단 자체가 아니라 표본평균이다.

대응 관계 해설

큰 표본의 평균 분포 → 중심극한정리: 서로 독립이고 동일한 분포에서 추출한 표본이 유한한 평균과 양의 유한한 분산을 가지면, 표본 크기가 커질수록 표준화한 표본평균의 분포가 표준정규분포에 가까워진다. 이를 표본평균 자체로 표현하면 X≈ N≤ft(μ,(σ^2)/(n))이다.

근사의 대상과 조건: 정규분포에 가까워지는 것은 **개별 관측값의 분포가 아니라 표본평균의 표본분포**이다.
모집단 자체가 정규분포일 필요는 없으며, 필요한 표본 크기는 모집단의 비대칭성·꼬리 특성 등에 따라 달라진다.
모집단이 정규분포이면 표본평균도 표본 크기와 관계없이 정확히 정규분포를 따른다.

다른 정리와의 구별: 베이즈정리는 관측된 정보에 따라 조건부확률을 갱신하는 관계이다. 중간값정리는 연속함수가 구간 양 끝의 함수값 사이의 값을 적어도 한 번 갖는다는 정리이다. 바수정리는 완비충분통계량과 부수통계량의 독립성에 관한 정리이므로, 큰 표본평균의 정규근사라는 설명과 다르다.

따라서 제시된 설명은 중심극한정리이며, ④와 일치한다.

NoteBase

문 29. 회귀분석을 통해 다음과 같은 분산분석표를 얻었다. 이에 대한 다음 설명 중 옳지 않은 것은?

요인	제곱합	자유도	평균제곱	F 값	p-값
회귀	2700	1	2700	90	(ㄴ)
잔차	600	20	(ㄱ)		
계	3300	21			

① (ㄱ)은 30이다.

② (ㄴ)은 $F(1, 20)$분포 확률밀도함수에서 90보다 크거나 같은 확률이다.

③ 단순회귀모형에 대한 분산분석표이다.

④ 분석에 사용된 자료의 개수는 $n = 21$이다.

정답 4번

간단한 해설

분산분석표의 **전체 자유도 21은** $n - 1$이다. 따라서 자료의 개수는 $n = 22$이므로 ④가 옳지 않다.

선지별 해설

① (O) 잔차 평균제곱은 잔차제곱합을 잔차 자유도로 나눈 값이다. 따라서 (char"3131) = 600/20 = 30이다. 회귀 평균제곱과의 비도 2700/30 = 90으로 표의 F값과 일치한다.

② (O) 회귀모형의 유의성을 검정할 때 귀무가설 아래에서 검정통계량은 **분자 자유도 1, 분모 자유도 20인 F 분포**를 따른다. 관측값이 90이므로 (char"3134) = $P(F_{1,20} \geq 90)$이다. 이는 밀도함수의 높이가 아니라 90 오른쪽 꼬리의 면적을 뜻한다.

③ (O) 절편을 포함한 통상적인 회귀모형에서 회귀 자유도는 설명변수의 수이다. 표의 회귀 자유도가 1이므로 **설명변수가 하나인 단순회귀모형**에 해당한다.

④ (X) 전체 자유도는 표본 수 자체가 아니라 $n - 1$이다. $n - 1 = 21$이므로 $n = 22$이며, 21이라는 설명은 잘못이다. 단순회귀의 잔차 자유도도 $n - 2 = 22 - 2 = 20$으로 표와 일치한다.

NoteBase

문 30. 흡연과 폐암이 서로 관련이 있는가를 알아보기 위해 다음의 데이터를 얻었다. 흡연과 폐암이 관련이 없다는 가정하에 셀 (ㄱ)의 기대도수는?

폐암	흡연 여부		계
	흡연	비흡연	
있음	15(ㄱ)	5	20
없음	10	10	20
계	25	15	40

① 12.5

② 15

③ 17.5

④ 20

정답 1번

간단한 해설

독립 가정 아래의 기대도수는 **행 합계 × 열 합계 ÷ 전체 도수**이다. 셀 ㄱ은 폐암 있음·흡연에 해당하므로 $\frac{20 \times 25}{40} = 12.5$이다.

계산 과정

흡연과 폐암이 관련이 없다는 가정은 두 분류변수가 독립이라는 뜻이다. 독립일 때 각 셀의 확률은 해당 행과 열의 주변확률의 곱이므로, 기대도수는 **행 합계와 열 합계의 곱을 전체 도수로 나눈 값**이다.

셀 ㄱ은 '폐암 있음' 행과 '흡연' 열이 만나는 곳이다. 표에서 해당 행 합계는 20명, 열 합계는 25명이고 전체는 40명이다. 셀 안의 15명은 실제 관측도수이며, 독립 가정 아래에서 구할 기대도수와 구별한다.

따라서 $E_{\text{char"3131}} = \frac{20 \times 25}{40} = 12.5$명이다. 같은 뜻으로 전체 흡연 비율 25/40을 폐암이 있는 20명에 적용한 결과이다. **기대도수는 정수일 필요가 없다**.

따라서 셀 ㄱ의 기대도수는 12.5이며, 선택지는 ①이다.

NoteBase

문 31. 확률분포에 대한 다음 설명 중 옳지 않은 것은?

① 확률변수 X가 정규분포 $N(\mu, \sigma^2)$를 따를 때, $Y = \frac{X-\mu}{\sigma}$는 표준정규분포를 따른다.

② 평균이 θ인 지수분포는 감마분포 $Gamma(1, \theta)$이다.

③ 확률변수 T가 자유도 r인 $t(r)$분포를 따를 때, T^2은 $F(1, r)$분포를 따른다.

④ 확률변수 X가 $F(1, r)$분포를 따를 때, $Y = \frac{1}{X}$는 $F(1, \frac{1}{r})$분포를 따른다.

정답 4번

간단한 해설

F분포는 독립인 두 카이제곱확률변수를 각각의 자유도로 나눈 값의 비이다. 역수를 취하면 **두 자유도의 순서가 교환**되므로, $X \sim F(1, r)$일 때 $1/X \sim F(r, 1)$이다.

선지별 해설

① O: 표준편차가 양수인 정규확률변수에서 평균을 빼고 표준편차로 나누면 평균 0, 분산 1로 표준화된다. 정규성도 유지되므로 $Y = (X-\mu)/\sigma \sim N(0, 1)$이다.

② O: 감마분포를 모양 모수 α와 **척도 모수 θ**로 표시하면 평균은 $\alpha\theta$이다. α = 1일 때 밀도는 $f(x) = \frac{1}{\theta}e^{-x/\theta}$로, x > 0에서 정의되는 평균 θ의 지수분포이다. 두 번째 모수를 비율 모수로 쓰는 다른 표기에서는 평균이 그 역수이므로, 여기서는 척도 모수 표기를 사용한다.

③ O: 서로 독립인 $Z \sim N(0, 1)$과 $U \sim \chi^2(r)$에 대해 $T = Z/\sqrt{U/r}$로 나타낼 수 있다. 제곱하면 $T^2 = (Z^2/1)/(U/r)$이며, $Z^2 \sim \chi^2(1)$이므로 $T^2 \sim F(1, r)$이다.

④ X: F 확률변수의 역수를 취하면 **분자와 분모의 자유도 순서가 바뀐다**. $X = (A/1)/(B/r)$에서 A와 B가 각각 자유도 1, r인 독립 카이제곱확률변수이면, $1/X = (B/r)/(A/1)$이므로 $Y \sim F(r, 1)$이다. 자유도 r을 역수로 바꾸어 $F(1, 1/r)$로 만드는 것이 아니다.

NoteBase

문 32. 어느 실험실에서 3명의 연구원이 시료 분석을 하기에 연구원마다 시료 측정에 차이가 있는지 알아보고자 한다. 3명의 연구원이 같은 실험 장비로 표준시료를 4회씩 측정하는, 즉, 총 12회의 실험을 완전히 랜덤하게 수행하여 다음과 같은 분산분석표를 얻었다. 이에 대한 다음 설명 중 옳지 않은 것은?

요인	제곱합	자유도	평균제곱	F 값	p-값
연구원		(ㄱ)		0.486	0.631
잔차		(ㄴ)			
계					

① 일원배치법에 해당하는 분산분석표이다.

② (ㄱ)은 2이다.

③ 유의수준 α=0.05에서 검정 결과, 연구원마다 시료 측정에 차이가 있다.

④ (ㄴ)은 9이다.

정답 3번

간단한 해설

표의 p-값은 0.631로 유의수준 0.05보다 크므로 **연구원별 평균이 같다는 귀무가설을 기각하지 못한다**. 따라서 연구원 사이에 유의한 측정 차이가 있다고 결론 내릴 수 없다.

NoteBase

선지별 해설

연구원이라는 한 요인의 수준이 3개이고, 각 연구원이 4회씩 측정하여 전체 관측 수는 $N = 12$이다. 귀무가설은 세 연구원의 평균 측정값이 같다는 $H_0 : \mu_1 = \mu_2 = \mu_3$이고, 대립가설은 적어도 하나의 평균이 다르다는 것이다. 표에는 연구원 요인의 $F = 0.486$, $p = 0.631$이 제시되어 있다.

① O: 비교하는 요인은 연구원 한 가지이며 그 수준이 세 연구원이다. 동일 장비와 표준시료를 사용하고 총 12회의 실험을 완전히 랜덤하게 수행하므로, 제시된 표는 일원배치 분산분석에 해당한다. 연구원이 세 명이라는 이유로 요인이 세 개인 것은 아니다.

② O: 집단 수가 $k = 3$일 때 집단 간, 즉 연구원 요인의 자유도는 $k - 1$이다. 따라서 (ㄱ)은 $3 - 1 = 2$이다.

③ X: 표의 p-값을 유의수준과 비교하면 $0.631 > 0.05$이므로 귀무가설을 기각하지 못한다. 따라서 **연구원마다 평균 측정값에 차이가 있다는 통계적 근거가 충분하지 않다**. 이는 세 평균이 반드시 같다는 사실을 입증한 것도 아니다. 일반적으로 이 검정을 기각했더라도 모든 연구원 쌍에 차이가 있다고 바로 결론 내리는 것은 아니다.

④ O: 잔차, 즉 집단 내 자유도는 전체 관측 수에서 집단 수를 뺀 $N - k$이다. 따라서 (ㄴ)은 $12 - 3 = 9$이다. 전체 자유도는 $N - 1 = 11$이며, 요인 자유도와 잔차 자유도의 합 $2 + 9 = 11$과 일치한다.

NoteBase

문 33. 범주형 자료에 대한 동일성검정(동질성검정)을 실시하려 한다. 다음 설명 중 옳지 않은 것은?

① 귀무가설은 각 범주가 동일한 분포를 따른다고 가정한다.

② 귀무가설 하에서 각 셀의 기대도수와 실제 관측값의 차이가 충분히 작을 때 귀무가설을 기각할 수 있다.

③ 카이제곱통계량을 검정통계량으로 사용할 수 있다.

④ 동일성검정의 검정 절차는 독립성검정의 검정 절차와 동일하다.

정답 2번

간단한 해설

기대도수와 관측도수가 가까우면 귀무가설과 자료의 불일치가 작다. 동질성검정에서는 **불일치를 나타내는 카이제곱통계량이 충분히 클 때** 귀무가설을 기각한다.

선지별 해설

① O: 동질성검정의 귀무가설은 비교하는 여러 집단에서 **범주별 확률분포가 같다**는 것이다. 예를 들어 각 집단의 찬성·반대 비율 구성이 동일한지를 검정한다. 모든 범주의 확률이 서로 같아야 한다는 균등분포 가정은 아니다.

② X: 관측도수가 귀무가설에서 예상한 기대도수와 가까우면 귀무가설과 자료의 불일치가 작다. 귀무가설을 기각하는 방향은 차이가 작은 때가 아니라 **표준화하여 합산한 불일치가 충분히 큰 때**이다. 즉 카이제곱통계량이 유의수준에 따른 우측 임계값보다 클 때 기각한다.

③ O: 관측도수를 O, 기대도수를 E라 하면 피어슨 검정통계량 $\chi^2 = \sum \frac{(O-E)^2}{E}$를 사용할 수 있다. 각 셀의 차이를 제곱하여 기대도수로 나눈 값을 합산한다. 관측의 독립성과 충분한 기대도수 등 카이제곱 근사의 적용 조건을 확인해야 한다.

④ O: 두 검정은 분할표에서 E_ij=행 합계×열 합계전체 도수로 기대도수를 구하고, 카이제곱통계량과 자유도 $(r-1)(c-1)$를 이용하는 **계산 절차가 같다**. 다만 동질성검정은 여러 모집단의 범주 분포가 같은지를, 독립성검정은 한 모집단의 두 분류변수가 독립인지를 묻는다는 점에서 표본 설계와 가설 해석이 다르다.

따라서 옳지 않은 설명은 ②이다.

NoteBase

문 34. 두 모집단의 모분산이 같은지를 알아보기 위해 다음 가설을 고려한다.
$H_0 : \sigma_X^2 = \sigma_Y^2$
귀무가설 하에서 검정통계량 $T = \frac{S_X^2}{S_Y^2}$의 분포는? 단, S_X^2과 S_Y^2은 각 모집단에서 추출한 확률표본의 표본분산이고, n_X와 n_Y는 표본 크기다.

① $t(n_X + n_Y - 2)$

② $\chi^2(n_X + n_Y)$

③ $F(n_X - 1, n_Y - 1)$

④ 감마분포, 즉, $Gamma(n_X, n_Y)$

정답 3번

간단한 해설

두 정규모집단에서 서로 독립적으로 표본을 추출했다는 전제하에서, 모분산이 같으면 표본분산의 비는 **F분포**를 따른다. 분자 자유도는 $n_X - 1$, 분모 자유도는 $n_Y - 1$이다.

핵심 풀이

두 모분산의 동일성을 검정하는 표준적인 F검정은 **정규모집단과 두 표본의 독립성**을 전제로 한다. 제시문에는 이 조건들이 명시되어 있지 않으므로, 아래 분포는 이 전제하에서 성립하는 것으로 해석한다. 모분산이 같다는 조건만으로 표본분산의 비가 항상 F분포를 따르는 것은 아니다.

정규모집단의 표본분산을 모분산으로 나누고 표본 크기에서 1을 뺀 값을 곱하면 카이제곱분포를 따른다. 따라서 $U = \frac{(n_X - 1)S_X^2}{\sigma_X^2}$는 $\chi^2(n_X - 1)$을 따르고, $V = \frac{(n_Y - 1)S_Y^2}{\sigma_Y^2}$는 $\chi^2(n_Y - 1)$을 따른다. 표본평균을 추정한 뒤 편차를 계산하므로 각 자유도는 표본 크기보다 1 작다.

두 표본이 독립이면 U와 V도 독립이다. **독립인 두 카이제곱변수를 각각의 자유도로 나눈 뒤 비를 취하면** F분포를 따르므로, $\frac{U/(n_X - 1)}{V/(n_Y - 1)} \sim F(n_X - 1, n_Y - 1)$이다.

이 비는 $\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$와 같다. 귀무가설 $\sigma_X^2 = \sigma_Y^2$ 아래에서는 모분산이 약분되므로 $T = \frac{S_X^2}{S_Y^2} \sim F(n_X - 1, n_Y - 1)$이다. 즉, 분자의 X 표본과 분모의 Y 표본에 대응하는 **두 자유도를 따로 유지**하며, 이는 ③의 분포와 일치한다.

카이제곱분포는 위에서 모분산으로 표준화한 각 표본분산에 적용된다. 이 문항의 통계량은 그 하나가 아니라 두 표본분산의 비이므로, 두 카이제곱변수의 비에서 얻는 F분포로 연결해야 한다.

NoteBase

문 35. 어느 회사에 입사를 희망한 지원자의 영어점수는 평균이 700이고 표준편차가 100인 정규분포를 따른다고 한다. 상위 5%를 선택하기 위한 커트라인은? 단, 표준정규분포를 따르는 확률변수 Z의 확률 $P(Z \geq 1.65) = 0.05$이다.

① 825

② 845

③ 865

④ 885

정답 3번

간단한 해설

상위 5%의 경계는 평균에서 **표준편차의 1.65배만큼 높은 점수**이다. 따라서 700 + 1.65 × 100 = 865점이다.

계산 과정

영어점수를 X, 커트라인을 c라고 하면 상위 5%의 조건은 $P(X \geq c) = 0.05$이다. 평균은 700점, 표준편차는 100점이므로 표준화하면 $Z = \frac{X - 700}{100}$이다.

문제에서 P(Z≥1.65)=0.05를 주었으므로 $\frac{c - 700}{100} = 1.65$가 되어야 한다. **상위 꼬리의 기준이므로 평균보다 높은 쪽의 양수 1.65**를 사용한다.

이를 풀면 $c = 700 + 1.65 \times 100 = 865$이다. 표준편차 100을 곱해 원래 점수 단위로 되돌리고 평균 700을 더한 값이다.

따라서 주어진 표준정규분포 기준에 따른 커트라인은 865점이며, ③과 일치한다.

NoteBase

문 36. 두 확률변수 X와 Y의 피어슨 상관계수가 0.9라 할 때, $X^* = 0.5X + 0.1$과 $Y^* = 0.4Y + 0.2$의 피어슨 상관계수는?

① 0.18

② 0.48

③ 0.81

④ 0.90

정답 4번

간단한 해설

피어슨 상관계수는 **양의 배율을 곱하거나 상수를 더해도 변하지 않는다**. 두 변환의 배율 0.5와 0.4가 모두 양수이므로 변환 후 상관계수도 0.90이다.

핵심 풀이

피어슨 상관계수는 $\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$이다. 상수 0.1과 0.2를 더하는 평행이동은 평균에서의 편차를 바꾸지 않으므로 공분산과 표준편차에 영향을 주지 않는다.

배율에 따라 $\mathrm{Cov}(X^*, Y^*) = 0.5 \times 0.4\mathrm{Cov}(X, Y)$이고, $\sigma_{X^*} = 0.5\sigma_X$, $\sigma_{Y^*} = 0.4\sigma_Y$이다. 따라서 $\rho_{X^*Y^*} = \frac{0.5 \times 0.4\mathrm{Cov}(X, Y)}{(0.5\sigma_X)(0.4\sigma_Y)} = \rho_{XY} = 0.9$이다.

두 배율이 모두 양수이므로 상관계수는 변하지 않는다. 분자의 배율만 반영한 0.18은 분모의 표준편차 변화가 빠진 값이고, 이동 상수를 더하거나 상관계수를 제곱할 이유도 없다. 구한 값 0.90에 해당하는 선택지는 ④이다.

NoteBase

문 37. 평균이 μ이고 분산이 σ^2인 정규분포로부터 다음과 같은 자료를 얻었다. 모평균의 최대가능도추정치(maximum likelihood estimate)는?

1.0 2.5 2.9 3.2 3.4 5.0

① 2.9

② 3.0

③ 3.1

④ 3.2

정답 2번

간단한 해설

정규모집단의 모평균 최대가능도추정치는 **표본평균**이다. 여섯 관측값의 합이 18.0이므로 μ=(18.0)/(6)=3.0이다.

계산 과정

정규모집단에서 얻은 독립 표본의 모평균 μ에 대한 최대가능도추정치는 **표본평균**이다. 모분산을 알고 있는 경우와 모평균·모분산을 함께 추정하는 경우 모두 같은 결과를 얻는다.

그 이유는 정규표본의 로그가능도에서 μ가 관여하는 부분이 $-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$이기 때문이다. 양의 분산을 고정하면 가능도를 최대화하는 것은 편차제곱합을 최소화하는 것과 같고, 그 값은 $\mu=\bar{x}$에서 얻어진다.

주어진 여섯 값의 합은 $1.0+2.5+2.9+3.2+3.4+5.0=18.0$이다. 따라서 μ_MLE=x=(18.0)/(6)=3.0이다.

모평균의 최대가능도추정치는 3.0이므로 ②와 일치한다.

NoteBase

문 38. 새로운 법안에 대한 찬성 여부를 성별에 따라 나누어 조사한 결과가 아래 표와 같다. 성별과 찬성 여부가 서로 독립인지 검정하기 위한 카이제곱통계량의 값은?

	찬성	반대
남성	30	20
여성	20	30

① 0

② 1

③ 4

④ $\frac{25}{6}$

정답 3번

간단한 해설

독립일 때 네 칸의 기대빈도는 모두 25명이다. **각 칸의 (관측빈도-기대빈도)^2/기대빈도가 1**이므로 카이제곱통계량은 $1+1+1+1=4$, 정답은 ③이다.

계산 과정

관측빈도는 남성의 찬성·반대가 각각 30명·20명, 여성의 찬성·반대가 각각 20명·30명이다. 성별 합계는 각각 50명이고, 찬성·반대 합계도 각각 50명이므로 전체는 100명이다.

독립이라는 귀무가설 아래 각 칸의 기대빈도는 **해당 행 합계와 열 합계의 곱을 전체 합계로 나눈 값**이다. 따라서 네 칸 모두 $E_{ij} = 50 \times 50/100 = 25$명이다.

연속성 보정을 적용하지 않은 Pearson 카이제곱통계량은 $\chi^2 = \sum_{i,j}(O_{ij} - E_{ij})^2/E_{ij}$이다. 대입하면 aligned&^2=(30-25)^2/25+(20-25)^2/25&+(20-25)^2/25+(30-25)^2/25&=1+1+1+1=4aligned 이다.

따라서 구하는 통계량은 **4**이며 정답은 ③이다. 빈도로 계산한 검정통계량은 단위가 없다. 자유도 $(2-1)(2-1)=1$과 통계량 값 4를 혼동하지 않아야 한다.

NoteBase

문 39. 모평균의 차이 검정을 위해 대응비교를 하려 한다. 다음 설명 중 가장 적절한 것은?

① 서로 다른 두 모집단에서 두 표본을 서로 독립적으로 추출한 경우 두 모집단의 차이를 확인하기 위해 사용할 수 있다.

② 하나의 모집단에서 두 표본을 서로 독립적으로 추출하였을 때 사용할 수 있다.

③ 새로운 정책에 대한 지지율이 성별에 따라 차이가 있는지 확인하기 위해 사용할 수 있다.

④ 고혈압 환자들에게 혈압강하제를 투여하기 전과 후 혈압 차이가 있는지를 확인하기 위해 사용할 수 있다.

정답 4번

간단한 해설

대응비교는 **같은 대상의 전후 측정값**처럼 일대일로 짝지어진 자료에 사용한다. 같은 환자의 약물 투여 전후 혈압은 환자별 차이를 구해 평균 변화가 있는지 검정할 수 있다.

선지별 해설

① X: 두 표본을 서로 독립적으로 추출했다면 관측값 사이에 자연스러운 일대일 짝이 없다. 이런 평균 차이는 **독립표본 비교**로 다루며, 관측값을 임의로 짝지어 대응비교를 적용하지 않는다.

② X: 같은 모집단에서 추출했더라도 두 표본이 서로 독립이라면 대응 자료가 아니다. 대응비교 여부는 모집단이 하나인지보다 **동일 대상의 반복 측정이나 의미 있는 짝이 있는지**에 달려 있다.

③ X: 성별에 따른 정책 지지율의 차이는 통상 서로 다른 응답자 집단의 비율을 비교하는 문제이다. 제시된 내용에는 대응 관계가 없으며, 평균의 대응비교보다는 두 비율 비교나 분할표 분석이 적합하다.

④ O: 같은 환자에게서 투여 전과 후의 혈압을 측정하면 **환자별로 두 측정값이 짝**을 이룬다. 각 환자의 전후 차이를 하나의 관측값으로 만들어 평균 차이가 0인지를 검정할 수 있다. 대응표본 t 검정은 환자 간 차이값의 독립성과 차이값 분포에 대한 정규성 또는 충분한 표본 크기 등의 조건을 확인하여 적용한다.

따라서 대응비교에 가장 적절한 사례는 ④이다.

NoteBase

문 40. $P(A \cup B) = 0.85$**이고** $P(A \cup B^C) = 0.75$**일 때, 확률** $P(A)$**는?**

① 0.4

② 0.5

③ 0.6

④ 0.7

정답 3번

간단한 해설

두 사건 $A \cup B$와 $A \cup B^C$의 합집합은 전체 표본공간이고, **교집합은 A**이다. 포함배제 공식으로 $P(A) = 0.85 + 0.75 - 1 = 0.60$을 얻는다.

계산 과정

B^C는 B의 여집합이므로 $B \cup B^C = \Omega$, $B \cap B^C = \varnothing$이다. $E = A \cup B$, $F = A \cup B^C$라 두면 $E \cup F = \Omega$이고, 분배법칙에 의해 $E \cap F = A \cup (B \cap B^C) = A$이다.

포함배제 공식 $P(E \cup F) = P(E) + P(F) - P(E \cap F)$에 대입하면 $1 = 0.85 + 0.75 - P(A)$이다. 따라서 $P(A) = 0.85 + 0.75 - 1 = 0.60$이다.

두 합집합의 확률을 더하면 전체 표본공간을 한 번, **공통 부분 A를 한 번 더** 세므로 1을 빼면 A의 확률만 남는다. 이 계산에는 A와 B의 독립성이 필요하지 않다. 구한 값 0.6에 해당하는 선택지는 ③이며, 나머지 0.4, 0.5, 0.7은 주어진 관계를 만족하지 않는다.

NoteBase