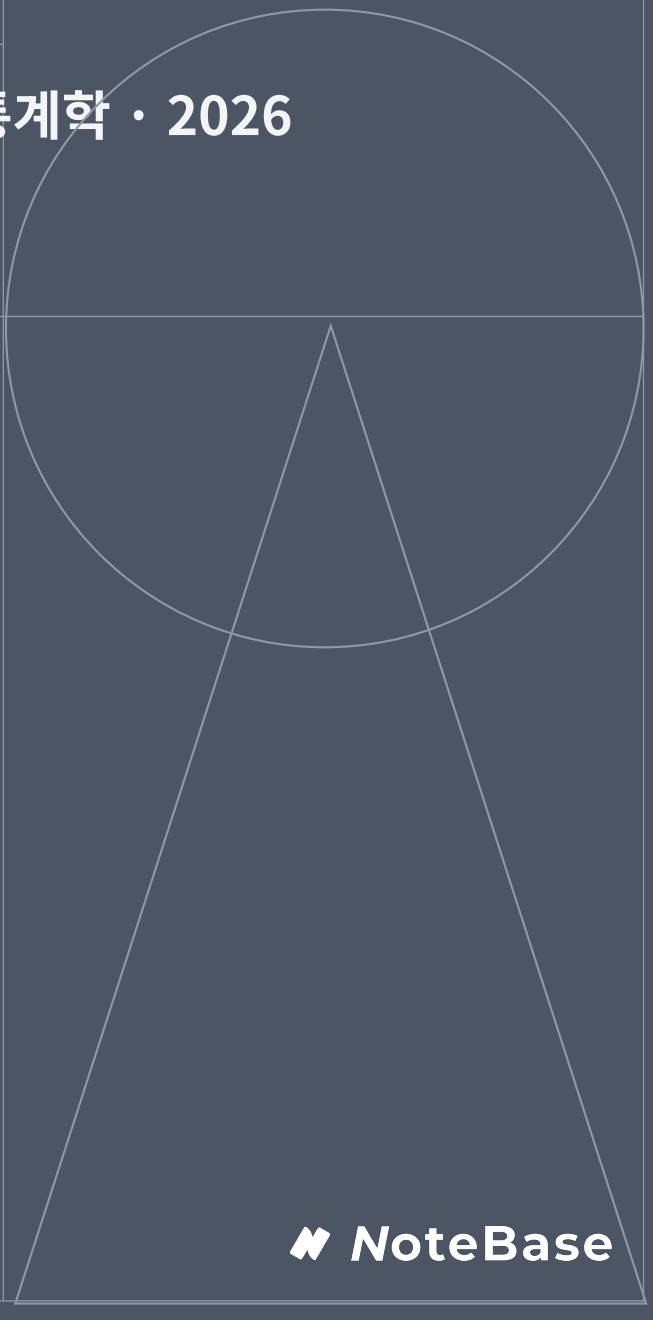
기출OX 해설집

공무원/경찰/공채(경위) · 통계학 · 2026



NoteBase

기출OX 해설집

공무원/경찰/공채(경위) · 통계학 · 2026

공무원/경찰/공채(경위) · 통계학 · 2026

문 1. 다음 중 연속형 자료에 해당되는 것은?

① 흡연 여부

② 섭씨 온도

③ 혈액형

④ 성별

정답 2번

간단한 해설

연속형 자료는 측정값이 일정 구간에서 연속적으로 변할 수 있는 수치형 자료이다. **섭씨 온도**는 소수 단위로도 측정할 수 있으므로 연속형에 해당한다.

선지별 해설

① X: 흡연 여부는 흡연자와 비흡연자처럼 **범주를 구분하는 질적 자료**이다. 0과 1로 표시하더라도 숫자는 범주의 기호일 뿐, 연속적인 측정량을 뜻하지 않는다.

② O: 섭씨 온도는 20도와 21도 사이에서도 20.5도처럼 다양한 값을 가질 수 있는 **연속형 수치 자료**이다. 온도계의 표시 자릿수가 제한되어도 측정 대상인 온도의 연속형 성격은 달라지지 않는다.

③ X: 혈액형은 A형·B형·AB형·O형 등 **종류를 구별하는 명목형 자료**이다. 혈액형 사이에 수치적 간격이나 크기 순서를 정할 수 없으므로 연속형 자료가 아니다.

④ X: 이 문항에서 성별은 범주를 구분하는 **명목형 자료**로 다룬다. 연속적인 크기를 측정한 값이 아니므로 연속형 자료에 해당하지 않는다.

NoteBase

문 2. 다음 중 산포도(흩어진 정도)의 측도로 사용할 수 없는 것은?

① 사분위수 범위

② 변이계수(변동계수)

③ 표준편차

④ 중앙값

정답 4번

간단한 해설

중앙값은 자료를 정렬했을 때의 **중심 위치**를 나타내는 값이지 흩어진 정도를 나타내는 값이 아니다. 따라서 산포도 측도로 사용할 수 없는 것은 ④이다.

선지별 해설

O/X는 각 항목이 산포도의 측도인지에 대한 판단이다.

① (O) 사분위수 범위는 **자료의 가운데 50%가 퍼져 있는 폭**이다. 제1사분위수를 Q_1, 제3사분위수를 Q_3라 하면 $IQR = Q_3 - Q_1$로 계산한다. 양 끝의 극단값보다 중앙부 자료의 흩어진 정도를 나타내는 산포도 측도이다.

② (O) 변이계수는 **평균에 비한 상대적인 산포도**이다. 표본평균과 표본표준편차를 각각 $\overline{x}$, s라 하면 $CV = \frac{s}{\overline{x}}$로 나타내며, 필요하면 100을 곱해 백분율로 표시한다. 평균이 양수이고 비율 비교가 의미 있는 자료에서 단위나 평균 수준이 다른 집단의 변동성을 비교하는 데 활용한다. 평균이 0이면 이 비율은 정의되지 않는다.

③ (O) 표준편차는 **평균을 중심으로 자료가 흩어진 정도**를 나타낸다. 평균과의 편차를 제곱하여 구한 분산의 0 이상인 제곱근이며, 원자료와 같은 단위를 사용한다. 값이 클수록 평균으로부터의 전반적인 흩어짐이 크다.

④ (X) 중앙값은 자료를 크기순으로 정렬했을 때 **가운데 위치를 나타내는 대표값**이다. 자료 개수가 홀수이면 가운데 값, 짝수이면 가운데 두 값의 평균을 사용한다. 분포의 중심 위치를 나타낼 뿐, 그 값만으로 자료가 얼마나 넓게 퍼져 있는지는 알 수 없으므로 산포도 측도가 아니다.

NoteBase

문 3. 다음 중 범주형 자료의 요약을 위한 방법으로 적절하지 않은 것은?

① 상자그림(box plot)

② 분할표(contingency table)

③ 파이차트(pie chart)

④ 막대그래프(bar plot)

정답 1번

간단한 해설

상자그림은 **수치형 자료의 중앙값·사분위수·산포**를 나타내는 도구이다. 범주별 빈도나 비율을 요약하는 방법이 아니므로 적절하지 않은 것은 ①이다.

선지별 해설

범주형 자료는 각 관측값이 어느 범주에 속하는지를 나타내므로, 요약할 때에는 **범주별 빈도와 비율**에 주목한다.

① X: 상자그림은 수치형 자료의 중앙값과 사분위수, 자료의 퍼짐 및 이상치 후보를 보여 준다. 범주 이름 자체에 수치적 크기나 간격이 있는 것은 아니므로 범주형 자료를 요약하는 일반적인 방법으로 적절하지 않다.

② O: 분할표는 둘 이상의 범주형 변수를 교차 분류하여 각 범주 조합의 빈도나 비율을 나타낸다. 두 변수 사이의 관련성을 살펴보는 데에도 활용한다.

③ O: 파이차트는 전체에서 각 범주가 차지하는 비율을 부채꼴의 크기로 나타낸다. 범주별 구성비를 요약하는 데 적절하다.

④ O: 막대그래프는 각 범주의 빈도나 상대도수를 막대의 높이로 나타낸다. 범주 사이의 빈도나 비율을 비교하는 데 적절하다.

NoteBase

문 4. 숫자 아닌 자료들을 분석할 때 알 수 있는 것은?

① 자료값들의 평균(mean)

② 자료값들의 중앙값(median)

③ 자료값들의 최빈값(mode)

④ 자료값들의 분산(variance)

정답 3번

간단한 해설

자료가 숫자가 아니어도 각 범주가 나온 횟수는 셀 수 있으므로, **가장 자주 나타난 범주인 최빈값**을 구할 수 있다. 범주 사이에 수치적 간격이나 순서가 없어도 빈도 비교는 가능하다.

선지별 해설

① X: 평균은 자료값을 더한 뒤 자료의 개수로 나눈 값이다. 색깔·혈액형 같은 비수치적 범주는 **값의 합과 나눗셈에 의미가 없으므로** 일반적으로 산술평균을 구할 수 없다. 범주에 임의의 번호를 붙여 그 번호의 평균을 계산해도 범주 자체의 평균이 되지는 않는다.

② X: 중앙값은 자료를 크기순으로 정렬했을 때 가운데 위치의 값이다. 따라서 **의미 있는 순서가 없는 명목형 자료**에서는 중앙값을 정할 수 없다. 만족도 등 순서가 있는 범주형 자료라면 중앙 범주를 논할 수 있지만, 단순히 숫자가 아닌 자료라는 조건만으로 순서가 주어지는 것은 아니다.

③ O: 최빈값은 가장 높은 빈도로 나타나는 값 또는 범주이다. 예를 들어 빨강·파랑·빨강이라는 자료에서는 빨강이 가장 많이 나타나므로 최빈값은 빨강이다. **출현 횟수만 비교하면 되므로** 숫자가 아닌 자료에도 적용할 수 있다. 최고 빈도가 같은 범주가 여럿이면 최빈 범주도 여러 개일 수 있다.

④ X: 분산은 각 자료값이 평균에서 떨어진 정도를 제곱하여 평균한 산포의 척도이다. 수치적 차이와 평균이 정의되어야 하므로 **일반적인 비수치 범주에는 분산을 적용할 수 없다**. 범주형 자료의 분포는 각 범주의 빈도나 비율 등으로 파악한다.

NoteBase

문 5. 표본공간 S에서 정의된 두 사건 A와 B에 대하여 옳지 않은 것은?

① $P(S) = 1$

② $A \subset B$이면 $P(A) \leq P(B)$

③ $P(A^c) = 1 - P(A)$, 단 A^c는 A의 여사건

④ $A \cap B = \phi$이면 A와 B는 서로 독립이다.

정답 4번

간단한 해설

배반은 두 사건이 **동시에 일어날 수 없음**을 뜻하고, 독립은 한 사건의 발생이 다른 사건의 확률에 영향을 주지 않는 관계이다. 확률이 모두 양수인 배반사건은 독립일 수 없으므로 두 개념을 같게 보면 안 된다.

선지별 해설

① O: 표본공간은 가능한 모든 결과를 포함하므로 **전체 확률은 1**이다. 이는 확률의 기본 공리로, $P(S) = 1$이 성립한다.

② O: $A \subset B$이면 B는 A와 나머지 부분 $B \setminus A$로 나뉜다. 따라서 $P(B) = P(A) + P(B \setminus A)$이고 확률은 음수가 아니므로 $P(A) \leq P(B)$이다.

③ O: 사건과 그 여사건은 서로 겹치지 않으면서 표본공간 전체를 이룬다. 따라서 $P(A) + P(A^c) = 1$이므로 $P(A^c) = 1 - P(A)$이다.

④ X: $A \cap B = \emptyset$은 두 사건이 **배반**이라는 뜻이지 독립을 보장하지 않는다. 독립의 조건은 $P(A \cap B) = P(A)P(B)$인데, 배반이면 왼쪽은 0이다. 두 사건의 확률이 모두 양수라면 오른쪽은 양수이므로 독립이 아니다. 배반과 독립이 함께 성립하려면 적어도 한 사건의 확률이 0이어야 한다.

NoteBase

문 6. 두 사건 A와 B가 서로 독립일 때 성립하지 않는 것은?

① $P(A \cap B) = P(A)P(B)$

② $P(A \mid B) = P(A)$

③ $P(B \mid A) = P(B)$

④ $P(A \cup B) = P(A) + P(B)$

정답 4번

간단한 해설

독립은 **배반을 뜻하지 않는다**. 독립인 두 사건의 합사건 확률은 $P(A \cup B) = P(A) + P(B) - P(A)P(B)$이므로, 독립이라는 조건만으로 성립하지 않는 것은 ④이다.

선지별 해설

독립은 한 사건의 발생이 다른 사건의 확률을 바꾸지 않는다는 뜻이다. 조건부확률을 사용하는 식은 조건이 되는 사건의 확률이 양수일 때 해석한다.

① O: 두 사건이 독립이라는 정의에 따라 $P(A \cap B) = P(A)P(B)$가 성립한다.

② O: $P(B) > 0$이면 $P(A \mid B) = P(A \cap B)/P(B) = P(A)$이다. 독립이므로 B의 발생을 알아도 A의 확률은 바뀌지 않는다.

③ O: $P(A) > 0$이면 $P(B \mid A) = P(A \cap B)/P(A) = P(B)$이다. 독립 관계는 대칭이므로 A의 발생을 알아도 B의 확률은 바뀌지 않는다.

④ X: 확률의 덧셈법칙에서는 두 사건에 공통으로 포함된 부분을 한 번 빼야 한다. 독립이면 그 부분의 확률이 $P(A)P(B)$이므로 올바른 식은 $P(A \cup B) = P(A) + P(B) - P(A)P(B)$이다. 교집합 확률이 0이라는 조건 없이 단순히 두 확률을 더할 수는 없다.

NoteBase

문 7. 두 사건 A와 B가 서로 독립이고, $P(A) = 0.3$, $P(B) = 0.2$이다. $P(A \cup B)$는 얼마인가?

① 0.41

② 0.42

③ 0.43

④ 0.44

정답 4번

간단한 해설

두 사건이 독립이므로 함께 발생할 확률은 $0.3 \times 0.2 = 0.06$이다. 두 확률의 합에서 이 **중복 부분을 한 번 빼면** 합사건의 확률은 0.44이다.

계산 과정

구할 값은 두 사건 중 적어도 하나가 발생할 확률 $P(A \cup B)$이다. 주어진 값은 $P(A) = 0.3$, $P(B) = 0.2$이며, 두 사건은 서로 독립이다.

먼저 **독립 조건으로 교집합 확률**을 구한다. 두 사건이 함께 발생할 확률은
$P(A \cap B) = P(A)P(B) = 0.3 \times 0.2 = 0.06$이다.

두 확률을 단순히 더하면 교집합을 두 번 세므로, 덧셈법칙에 따라 **중복된 교집합을 한 번 뺀다**. 따라서
aligned&P(A B)=P(A)+P(B)&-P(A B)=0.3+0.2-0.06=0.44aligned
이다.

구한 확률은 0.44, 즉 44%이며 ④와 일치한다.

NoteBase

문 8. 확률변수 X에 대해 $E(X)$는 기댓값(expectation), $Var(X)$는 분산(variance), $SD(X)$는 표준편차(standard deviation)을 나타낸다. 다음 중 맞지 않는 것은? (단, a와 b는 실수이다.)

① $E(aX+b)=aE(X)+b$

② $Var(aX+b)=a^2Var(X)$

③ $SD(aX+b)=aSD(X)$

④ $E(X^2)\geq\{E(X)\}^2$

정답 3번

간단한 해설

a는 음수일 수도 있지만 표준편차는 음수가 될 수 없다. 따라서 표준편차에는 **배율의 절댓값**을 곱하여 $SD(aX+b)=|a|SD(X)$로 써야 하므로 ③이 옳지 않다.

선지별 해설

문항에서 사용하는 기댓값과 분산이 유한하게 존재하는 경우를 전제로 한다.

① (O) 기댓값의 **선형성**에 따라 $E(aX+b)=aE(X)+E(b)=aE(X)+b$이다. 상수 b를 더하면 전체 위치와 기댓값이 b만큼 이동한다.

② (O) 상수 b는 값과 평균을 함께 이동시키므로 평균에서의 편차에는 영향을 주지 않는다. $(aX+b)-E(aX+b)=a\{X-E(X)\}$이므로 편차를 제곱한 평균인 분산은 $Var(aX+b)=a^2Var(X)$가 된다. 즉, 분산은 **배율의 제곱**만큼 변한다.

③ (X) 표준편차는 분산의 0 이상인 제곱근이므로 $SD(aX+b)=\sqrt{a^2Var(X)}=|a|SD(X)$이다. 따라서 a가 아니라 **a의 절댓값**을 곱해야 한다. a가 음수이고 원래 표준편차가 양수이면 선지의 우변이 음수가 되어 표준편차의 성질에 어긋난다.

④ (O) 분산은 $Var(X)=E(X^2)-\{E(X)\}^2$이고 항상 0 이상이다. 이를 정리하면 $E(X^2)\geq\{E(X)\}^2$를 얻는다. 두 값이 같은 경우는 **분산이 0인 경우**이다.

NoteBase

문 9. 다음은 대학생 2,000명의 음주와 흡연 습관을 조사한 결과표이다. 흡연자들 중 음주를 하는 학생들의 비율은?

	음주	비음주	합계
흡연	300	200	500
비흡연자	600	900	1,500
합계	900	1,100	2,000

① 0.40

② 0.45

③ 0.60

④ 0.75

정답 3번

간단한 해설

흡연자 전체 500명을 분모로, 그중 음주자 300명을 분자로 사용한다. **흡연자 중 음주자의 비율**은 300/500 = 0.60이다.

계산 과정

'흡연자들 중'이라는 조건이 있으므로 비율의 기준은 전체 학생 2,000명이 아니라 **흡연자 500명**이다. 표의 흡연 행에서 음주 300명과 비음주 200명을 더하면 합계 500명이다.

구하는 분자는 흡연과 음주를 모두 하는 학생 수이다. 표의 **흡연 행과 음주 열이 만나는 칸**에 있는 300명을 사용한다.

따라서 흡연자 중 음주자의 비율은 $\frac{300}{500} = 0.60$이다. 이는 60%이므로 ③과 일치한다.

NoteBase

문 10. 다음은 1개의 동전을 연속하여 3회 던져서 앞면의 수의 횟수 X의 확률분포함수를 나타낸 표이다. 빈칸 (ㄱ)에 들어갈 값은?

X	0	1	2	3
$P(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	(ㄱ)

① $\frac{1}{8}$
② $\frac{2}{8}$
③ $\frac{3}{8}$
④ $\frac{4}{8}$

정답 1번

간단한 해설

확률분포표에서 **모든 확률의 합은 1**이다. 이미 주어진 세 확률의 합이 $\frac{7}{8}$이므로 빈칸은 $1-\frac{7}{8}=\frac{1}{8}$이다.

계산 과정

표의 빈칸 (ㄱ)은 앞면이 3번 나올 확률인 $P(X=3)$이다. 3회 던졌을 때 앞면 횟수 X가 가질 수 있는 값은 0, 1, 2, 3뿐이며, 이 네 사건은 서로 겹치지 않고 모든 결과를 포함한다. 따라서 **표에 있는 확률의 합은 1**이다.

표에서 $P(X=0)=\frac{1}{8}$, $P(X=1)=\frac{3}{8}$, $P(X=2)=\frac{3}{8}$이므로 주어진 확률의 합은 $\frac{1}{8}+\frac{3}{8}+\frac{3}{8}=\frac{7}{8}$이다. 전체 확률에서 이를 빼면 $P(X=3)=1-\frac{7}{8}=\frac{1}{8}$이다.

따라서 (ㄱ)에 들어갈 값은 $\frac{1}{8}$이며, ①과 일치한다.

NoteBase

문 11. X와 Y의 결합확률질량함수가 다음 표와 같을 때, X의 기대값 $E(X)$의 값은?

	$X=-1$	$X=0$	$X=1$
$Y=-1$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
$Y=0$	$\frac{1}{8}$	0	$\frac{1}{8}$
$Y=1$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

① 0
② $\frac{1}{8}$
③ $\frac{2}{8}$
④ $\frac{3}{8}$

정답 1번

간단한 해설

기대값은 각 X값을 확률로 가중한 평균이다. **음수 -1과 양수 1의 확률이 각각 3/8로 같아 서로 상쇄**되고, X가 0인 경우의 기여도 0이다.

NoteBase

계산 과정

표에서 열은 $X=-1, 0, 1$, 행은 $Y=-1, 0, 1$을 나타낸다. 중앙의 $P(X=0, Y=0)$만 0이고 나머지 여덟 칸은 모두 1/8이므로, 전체 확률의 합은 1이다.

X의 기대값을 구하려면 먼저 **같은 X값의 열에서 Y의 모든 경우를 합산**하여 주변확률을 구한다. 따라서 $P(X=-1)=3/8$, $P(X=0)=2/8$, $P(X=1)=3/8$이다. 이때 Y의 특정 값이 조건으로 주어진 것이 아니므로 조건부확률로 나누지 않는다.

이산확률변수의 기대값은 각 값에 해당 확률을 곱해 더한 **확률가중평균**이다.
aligned&E(X)=Σ_x xP(X=x)=(-1)×(3/8)&+0×(2/8)+1×(3/8)=0aligned
이다. 음수와 양수의 기여가 같은 크기로 상쇄된다.

따라서 $E(X)=0$이며 ①과 일치한다. 문제에서 별도의 물리적 단위는 제시되지 않았다.

NoteBase

문 12. 확률변수 X와 Y의 결합확률질량함수가 아래와 같다. 틀린 것은?

	$X=-1$	$X=0$	$X=1$
$Y=0$	0	$\frac{1}{3}$	0
$Y=1$	$\frac{1}{3}$	0	$\frac{1}{3}$

① $E(X)=0$

② $E(XY)=0$

③ $Cov(X,Y)=0$

④ X와 Y가 서로 독립이다.

정답 4번

간단한 해설

$P(X=0,Y=0)=1/3$이지만 $P(X=0)P(Y=0)=1/9$이다. 결합확률이 주변확률의 곱과 다르므로 두 변수는 독립이 아니며, 틀린 것은 ④이다. 공분산이 0이어도 **독립이 보장되지는 않는다**.

NoteBase

선지별 해설

표에서 양의 확률을 갖는 점은 $(X, Y) = (-1, 1), (0, 0), (1, 1)$이며 각 확률은 1/3이다. 기댓값은 각 값에 해당 확률을 곱해 더하고, 독립 여부는 결합확률과 주변확률의 곱을 비교한다.

① O: 각 X 열의 확률을 합하면 $P(X=-1) = P(X=0) = P(X=1) = 1/3$이다. 따라서 E(X)=(-1)×13+0×13+1×13=0이다.

② O: 양의 확률을 갖는 세 점에서 XY의 값은 각각 $-1, 0, 1$이다. 따라서 E(XY)=(-1)×13+0×13+1×13=0이다.

③ O: $Y=1$ 행의 확률 합은 2/3이므로 $E(Y) = 2/3$이다. 공분산 공식에 대입하면 aligned&Cov(X,Y)=E(XY)&-E(X)E(Y)=0-0×23=0aligned 이다.

④ X: 독립이라면 **모든 값의 쌍에서 결합확률이 주변확률의 곱과 같아야 한다**. 그런데 표의 $X=0, Y=0$ 칸은 1/3이고, 해당 열과 행의 확률 합은 각각 1/3이므로 그 곱은 1/9이다. 두 값이 다르므로 X와 Y는 독립이 아니다.

NoteBase

문 13. 확률 변수 X와 Y의 상관계수가 0에 가까울 때의 정확한 해석은?

① X와 Y가 아무 관계가 없다.

② X와 Y 사이에 곡선 관계가 존재한다.

③ X와 Y 사이에 직선 관계가 약하다.

④ X와 Y 사이에 곡선 관계가 약하다.

정답 3번

간단한 해설

상관계수는 두 변수의 **직선적 관계**를 측정하므로 0에 가까울수록 직선 관계가 약하다. 비선형 관계가 남아 있을 수 있어 모든 관계가 없다고 단정할 수는 없다.

개념 구분

이 문항에서 상관계수는 두 변수 사이의 **선형, 즉 직선적 관계의 방향과 강도**를 나타내는 피어슨 상관계수이다. 값이 1에 가까우면 강한 양의 직선 관계, -1에 가까우면 강한 음의 직선 관계를 뜻한다. 0에 가까우면 일정한 직선 방향으로 함께 변하는 경향이 약하다는 의미이다.

직선 관계가 약하다는 사실을 **아무 관계도 없거나 서로 독립이라는 뜻으로 확대할 수는 없다**. 예를 들어 U자 모양의 강한 비선형 관계가 있어도 상관계수가 0에 가까울 수 있다. 그러나 그러한 관계가 가능하다는 것과 반드시 존재한다는 것은 다르므로, 곡선 관계의 존재나 강약 역시 상관계수가 0에 가깝다는 정보만으로 결정할 수 없다.

따라서 정확한 해석은 X와 Y 사이의 직선 관계가 약하다는 것이며, ③과 일치한다.

NoteBase

문 14. 확률변수 X와 Y의 분산을 각각 $Var(X)$, $Var(Y)$라 하고 X와 Y의 공분산을 $Cov(X, Y)$라 할 때, 다음 설명 중 옳지 않은 것은?

① X와 Y가 독립이면 $Cov(X, Y)$는 항상 0이다.

② $Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y)$

③ $Cov(X, Y)$는 음수가 될 수도 있다.

④ $Var(X - Y)$는 음수가 될 수도 있다.

정답 4번

간단한 해설

분산은 평균에서 벗어난 정도를 제곱하여 평균한 값이므로 **항상 0 이상**이다. 두 확률변수의 차이도 하나의 확률변수이므로 그 분산이 음수가 되지는 않는다.

선지별 해설

① O: 문항처럼 분산과 공분산이 유한하게 존재하는 상황에서, 독립이면 **곱의 기대값이 기대값의 곱으로 분리**되어 $E(XY) = E(X)E(Y)$이다. 따라서 $Cov(X, Y) = E(XY) - E(X)E(Y) = 0$이다. 다만 공분산이 0이라는 사실만으로 독립을 결론 낼 수는 없다.

② O: 선형결합의 분산은
aligned&Var(aX+bY)=a^2Var(X)&+b^2Var(Y)+2abCov(X,Y)aligned
이다. 여기에 $a = 1, b = -1$을 대입하면 **공분산 항의 부호가 음수**가 되어
$Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y)$가 된다. 이 등식에는 독립 조건이 필요하지 않다.

③ O: 공분산은 두 변수의 평균으로부터의 편차를 곱한 값의 기대값이므로 **음수도 가능**하다. 예를 들어 분산이 양수인 X에 대해 $Y = -X$이면 $Cov(X, Y) = -Var(X) < 0$이다.

④ X: $Z = X - Y$로 놓으면 $Var(X - Y) = Var(Z) = E[(Z - E(Z))^2]$이다. 제곱한 값은 항상 0 이상이므로 그 기대값인 **분산도 음수가 될 수 없다**. 분산 전개식에 빨셈이 나타나더라도 전체 결과는 0 이상이다.

NoteBase

문 15. 확률변수 X, Y와 상수 a, b, c, d에 대하여 상관계수 $\mathrm{Corr}(aX+b, cY+d)$와 같은 값은? (단, a와 c는 양수이다.)

① $\mathrm{Corr}(X, Y)$

② $ac\mathrm{Corr}(X, Y)$

③ $ac\mathrm{Corr}(X, Y) + b + d$

④ $ac\mathrm{Corr}(X, Y) + bd$

정답 1번

간단한 해설

상관계수에서는 양의 배율이 공분산과 표준편차의 곱에 함께 적용되어 상쇄되고, 상수의 덧셈도 영향을 주지 않는다. 따라서 **양의 선형 변환 전후의 상관계수는 같다**.

핵심 풀이

상관계수는 공분산을 두 표준편차의 곱으로 나눈 값이다. σ_X, σ_Y를 각 변수의 표준편차라 하면 $\mathrm{Corr}(X, Y) = \mathrm{Cov}(X, Y)/(\sigma_X\sigma_Y)$이다. 이 식은 두 분산이 유한하고 양수여서 상관계수가 정의되는 경우에 적용한다.

상수 b, d를 더하면 평균도 같은 만큼 이동하므로 평균으로부터의 편차는 바뀌지 않는다. 따라서 **상수의 덧셈은 공분산과 표준편차를 바꾸지 않는다**. 공분산은 $\mathrm{Cov}(aX+b, cY+d) = ac\mathrm{Cov}(X, Y)$이다.

$a, c > 0$이므로 변환된 표준편차는 각각 $a\sigma_X$, $c\sigma_Y$이다. 따라서 $\mathrm{Corr}(aX+b, cY+d) = \frac{ac\mathrm{Cov}(X, Y)}{(a\sigma_X)(c\sigma_Y)} = \mathrm{Corr}(X, Y)$이다. **분자와 분모에 같은 배율이 생겨 상쇄된다**.

배율 ac가 남는 것은 공분산의 변환식이며, 상관계수에는 표준편차로 나누는 과정이 추가된다. 또 위치 이동량 $b+d$나 bd가 상관계수에 더해지지 않으므로 ②~④는 일반적으로 성립하는 변환식이 아니다.

따라서 주어진 양의 배율 조건에서 같은 값은 $\mathrm{Corr}(X, Y)$이며 ①과 일치한다.

NoteBase

문 16. 확률변수 X와 Y의 분산이 각각 $Var(X)=3$, $Var(Y)=2$이고 공분산이 $Cov(X,Y)=-\sqrt{3}$이다. $Var(2X-\sqrt{3}Y)$는 얼마인가?

① 10

② 20

③ 30

④ 40

정답 3번

간단한 해설

선형결합의 분산에는 공분산항도 더해야 하며, 이 문제에서는 그 항이 **양수인 12**가 된다. 두 분산항 12와 6을 합치면 30이므로 정답은 ③이다.

계산 과정

두 확률변수의 선형결합에서는 **각 분산뿐 아니라 공분산항도 포함**한다. 적용할 식은
aligned&Var(aX+bY)=a^2Var(X)&+b^2Var(Y)+2abCov(X,Y)aligned
이다.

이 문제의 계수는 $a=2$, $b=-\sqrt{3}$이다. 따라서 두 분산항은 $2^2\times 3=12$, $(-\sqrt{3})^2\times 2=6$이다. 공분산항은 $2\times 2\times(-\sqrt{3})\times(-\sqrt{3})=12$로, **음수인 계수와 음수인 공분산이 곱해져 양수**가 된다.

세 항을 합하면 $Var(2X-\sqrt{3}Y)=12+6+12=30$이다. 따라서 정답은 ③이다.

NoteBase

문 17. A 산업단지의 공장의 크기가 다음과 같은 특성을 가지고 있다. 가장 작은 공장의 크기는 15에이커이고, 전체의 10%는 25에이커보다 작고, 전체의 25%는 40에이커보다 작으며 중앙값은 55에이커, 75%는 100에이커보다 작고, 90%는 125에이커보다 작다. 가장 큰 공장은 200에이커이다. 사분위수 범위(Interquartile Range, IQR)를 구하시오.

① 185

② 100

③ 60

④ 55

정답 3번

간단한 해설

사분위수 범위는 **제3사분위수에서 제1사분위수를 뺀 값**이다. 하위 25%와 75%의 경계가 각각 40에이커와 100에이커이므로 사분위수 범위는 60에이커이다.

계산 과정

사분위수 범위는 자료의 가운데 50%가 퍼져 있는 구간의 폭이며, $IQR = Q_3 - Q_1$로 계산한다. 여기서 Q_1은 하위 25%의 경계인 제1사분위수이고, Q_3은 하위 75%의 경계인 제3사분위수이다.

문제에서 전체의 25%가 40에이커보다 작으므로 $Q_1 = 40$에이커이고, 전체의 75%가 100에이커보다 작으므로 $Q_3 = 100$에이커이다. 최솟값·최댓값, 10%·90%의 경계와 중앙값은 사분위수 범위 계산에 사용하지 않는다.

따라서 $IQR = 100 - 40 = 60$에이커이다. 이는 선택지 ③의 60에 해당한다.

NoteBase

문 18. A 지역 휘발유의 리터 당 가격이 평균 1,800원이고 표준편차가 50원이다. A 지역 어떤 주유소에 리터 당 휘발유 가격이 1,900원일 때, 표준화 값인 z-값을 구하시오.

① 1.5

② 2

③ 2.5

④ 3

정답 2번

간단한 해설

1,900원은 평균보다 100원 높으며, 이는 표준편차 50원의 **2배**이다. 따라서 $z = \frac{1,900 - 1,800}{50} = 2$이다.

계산 과정

표준화 값 z는 관측값이 평균에서 표준편차의 몇 배만큼 떨어져 있는지를 나타낸다. 관측 가격을 x, 평균을 μ, 표준편차를 σ라고 하면 **평균과의 차이를 표준편차로 나누어** $z = \frac{x - \mu}{\sigma}$로 구한다. 분모에는 분산이 아니라 표준편차를 사용한다.

문제에서 x = 1,900원, μ = 1,800원, σ = 50원이다. 같은 리터 당 가격 기준으로 비교하므로 평균과의 차이는 1,900 - 1,800 = 100원이다. 이를 표준편차로 나누면 $z = \frac{1,900 - 1,800}{50} = \frac{100}{50} = 2$이다.

이 가격은 평균보다 **표준편차의 2배만큼 높다**. 가격 차이와 표준편차의 단위가 상쇄되어 z에는 단위가 없으며, 구한 값 2는 ②와 일치한다.

NoteBase

문 19. 창고에 100개의 제품이 있는데, 그 중 20개가 불량품이고, 80개가 양품이다. 정확한 불량률이 0.2라는 것을 모르기에, 불량률을 추정하기 위해 표본을 복원추출로 10개 뽑는다. 뽑힌 10개의 표본 중 불량품의 개수를 확률변수 X라 하면 X의 분산은?

① 1.4

② 1.6

③ 1.8

④ 2.0

정답 2번

간단한 해설

복원추출에서는 불량 확률 0.2가 유지되므로 불량품 개수는 이항분포를 따른다. **이항분포의 분산** $np(1-p)$에 대입하면 $10 \times 0.2 \times 0.8 = 1.6$이다.

계산 과정

한 번 뽑을 때 불량품일 확률은 $p = 20/100 = 0.2$이고, 양품일 확률은 $1 - p = 0.8$이다. 표본을 뽑는 사람이 불량률을 모른다는 사정과 실제 모집단의 불량 확률은 구별한다.

추출 횟수는 $n = 10$이다. **복원추출**이므로 매번 모집단 구성과 불량 확률이 유지되고 각 추출을 독립인 시행으로 본다. 따라서 불량품의 개수는 이항분포 $X \sim \mathrm{Bin}(10, 0.2)$를 따른다.

독립인 베르누이 시행의 합인 이항확률변수의 분산은 $np(1-p)$이다. 이에 따라
$\mathrm{Var}(X) = 10 \times 0.2 \times 0.8 = 1.6$이다. 복원추출이므로 비복원추출에서 쓰는 유한모집단 수정계수를 곱하지 않는다.

문제가 구하는 값은 **불량품 개수의 분산**이지 평균 $np = 2$가 아니다. 분산 1.6은 ②와 일치한다.

NoteBase

문 20. A 은행에서는 오전 10시부터 11시 사이에 평균 60명의 손님이 방문한다고 한다. 10분간 도착한 손님의 수를 확률변수 X라고 할 때, X의 분산은?

① 1

② 6

③ 10

④ 60

정답 3번

간단한 해설

손님 도착 수를 포아송분포로 모형화하면 **평균과 분산이 같다**. 10분간 평균 도착 수는 $60 \times \frac{10}{60} = 10$이므로 분산도 10이다.

출제 포인트

시간에 맞게 평균 도착 수를 환산한 뒤, 포아송분포의 평균과 분산이 같다는 성질을 적용하는 문제이다.

조건 정리

1시간인 60분 동안 평균 60명이 방문한다. X는 10분 동안 도착한 손님의 수이며, 구할 값은 X의 분산이다. 이 문항은 평균 도착률이 일정한 포아송 도착 모형을 전제로 풀이한다.

사용할 공식

포아송분포는 일정한 시간 동안 발생한 사건의 횟수를 나타내는 분포이다. 모수 λ는 해당 시간 동안의 평균 발생 횟수이며, $E(X) = Var(X) = \lambda$이다. 여기서는 손님의 도착 횟수에 이 모형을 적용하므로, 10분간 평균 도착 수가 곧 X의 분산이다.

계산 과정

1. 분당 평균 도착 수는 $\frac{60}{60} = 1$명이다.
2. 10분간 평균 도착 수는 $\lambda = 1 \times 10 = 10$이다.
3. 포아송분포의 분산은 평균과 같으므로 $Var(X) = 10$이다.

선지 판단

① X: 1은 분당 평균 도착 수이다. X는 10분간 도착 수이므로, 분산은 1이 아니라 10이다.
② X: 10분간 평균 도착 수는 $60 \times \frac{10}{60} = 10$이다. 포아송분포의 분산도 10이므로, 6을 10으로 고친다.
③ O: 10분간 평균 도착 수가 10이고, 포아송분포는 평균과 분산이 같으므로 X의 분산은 10이다.
④ X: 60은 1시간 동안의 평균 도착 수이다. X의 관측 시간은 10분이므로, 그 시간에 맞게 환산한 분산은 10이다.

NoteBase

오답 포인트

④의 60은 1시간의 평균이고, ①의 1은 1분의 평균이다. 포아송분포에서 평균과 분산이 같더라도, 먼저 X가 정의된 10분에 맞춰 평균을 환산해야 한다.

다시 나오면 이렇게 판단

포아송 도착 모형에서는 평균 도착 수를 관측 시간에 비례해 환산한다. 그 값을 λ로 두고, 평균과 분산을 구할 때 모두 λ를 사용한다.

암기 포인트

포아송분포: 평균=분산=해당 시간의 평균 발생 횟수.

NoteBase

문 21. 확률변수 X의 평균이 2, 분산이 9이다. 이때 $2X$의 평균과 분산은?

① 평균: 2, 분산 9

② 평균: 4, 분산 18

③ 평균: 4, 분산 36

④ 평균: 8, 분산 36

정답 3번

간단한 해설

확률변수에 2를 곱하면 **평균은 2배, 분산은 4배**가 된다. 평균은 $2 \times 2 = 4$이고, 분산은 $2^2 \times 9 = 36$이다.

계산 과정

주어진 값은 $E(X) = 2$, $\mathrm{Var}(X) = 9$이다. 확률변수의 모든 값에 상수 2를 곱하면 평균도 2배가 되므로, 기댓값의 선형성에 따라 $E(2X) = 2E(X) = 2 \times 2 = 4$이다.

분산은 평균으로부터의 편차를 제곱한 값의 평균이다. X를 2배 하면 평균도 2배가 되어 각 편차가 2배가 되고, **편차의 제곱은 4배**가 된다. 따라서 $\mathrm{Var}(2X) = 2^2\mathrm{Var}(X) = 4 \times 9 = 36$이다.

따라서 평균은 4, 분산은 36이며, ③과 일치한다. 분산을 18로 계산하는 것은 평균처럼 분산에도 계수 2만 곱한 오류이다.

NoteBase

문 22. 평균이 10, 분산이 25인 모집단으로부터 확률표본 100개를 추출하였을 때, 표본평균 $\bar{X}$의 평균과 분산은?

① 평균: 10, 분산 25

② 평균: 10, 분산 0.25

③ 평균: 0.1, 분산 25

④ 평균: 0.1, 분산 0.25

정답 2번

간단한 해설

표본평균의 평균은 모평균 10과 같고, 분산은 **모분산을 표본 크기로 나누어** $\frac{25}{100} = 0.25$가 된다. 따라서 정답은 ②이다.

계산 과정

모평균은 $\mu = 10$, 모분산은 $\sigma^2 = 25$, 표본의 크기는 $n = 100$이다. 확률표본의 관측값들을 $X_1, \ldots, X_{100}$이라 하면 표본평균은 $\overline{X} = \frac{1}{100}\sum_{i=1}^{100} X_i$이다.

기댓값의 선형성에 따라 $E(\overline{X}) = \frac{1}{100} \times 100 \times 10 = 10$이다. 여러 관측값을 평균 내더라도 **평균의 중심은 모평균과 같다**.

표본 관측값들이 독립이므로 합의 분산은 각 분산의 합이고, 100으로 나누는 배율은 분산에서 제곱된다. 따라서 $Var(\overline{X}) = \frac{1}{100^2} \times 100 \times 25 = \frac{25}{100} = 0.25$이다. 즉, **표본평균의 분산은 모분산을 표본 크기로 나눈 값**이다. 평균 10, 분산 0.25에 해당하는 정답은 ②이다.

NoteBase

문 23. 확률변수 Z가 평균이 0이고 표준편차가 1인 표준정규분포를 따른다고 한다. 이때 $P(Z > 0)$은?

① 0

② 0.5

③ 0.9

④ 1.0

정답 2번

간단한 해설

표준정규분포는 **0을 중심으로 좌우 대칭**이고 $P(Z = 0) = 0$이다. 따라서 0보다 큰 구간의 확률은 전체의 절반인 0.5이다.

핵심 풀이

표준정규분포의 밀도곡선은 평균 0을 중심으로 좌우 대칭이다. 따라서 **0의 왼쪽과 오른쪽 확률이 같다**는 성질을 이용한다.

정규분포는 연속분포이므로 정확히 한 점을 취할 확률은 $P(Z = 0) = 0$이다. 이는 Z0인 구간 전체의 확률이 0이라는 뜻은 아니다.

전체 확률이 1이고 가운데 한 점의 확률이 0이므로 P(Z0)+P(Z0)=1이다. 두 항은 대칭성으로 같아서 P(Z0)=1/2=0.5이며 ②와 일치한다. 나머지 선택지의 값은 좌우에 확률이 절반씩 나뉜다는 조건과 맞지 않는다.

NoteBase

문 24. 확률변수 X가 아래의 확률밀도함수를 갖는다. X의 중앙값은?
f(x)=cases-(3)/(4)(x+2)x, & (-2<x<0)\\0 & (그 외)cases

① −2

② −1

③ 1

④ 2

정답 2번

간단한 해설

확률밀도함수와 그 정의 구간이 **-1을 중심으로 대칭**이므로 -1의 왼쪽과 오른쪽 확률이 각각 1/2이다. 따라서 중앙값은 -1이다.

계산 과정

연속확률변수의 중앙값 m은 그 왼쪽과 오른쪽의 확률을 각각 1/2로 나누는 값이다. 따라서 주어진 밀도함수에 대해 P(X≤ m)= ∫ _-2^mf(x) dx=12을 만족해야 한다.

밀도함수를 완전제곱식으로 나타내면 f(x)=34\1-(x+1)^2\이다. 밀도가 양수인 구간 (-2, 0)은 -1을 중심으로 대칭이고, 밀도함수도 $x = -1$을 축으로 대칭이다. 그러므로 **-1의 양쪽에 놓인 확률이 같다**.

왼쪽 확률을 적분하여 확인하면
aligned& ∫ _-2^-1-34(x^2+2x) dx&=≤ft[-(x^3)/(4)-(3x^2)/(4)]_-2^-1=-12-(-1)=12aligned
이다. 구간 안에서는 밀도가 양수이므로 누적확률이 엄격히 증가하여 이 중앙값은 유일하다.

따라서 중앙값은 -1이며, 선택지 ②와 일치한다.

NoteBase

문 25. 성공확률이 p, 시행횟수가 n인 이항분포 $B(n, p)$를 따르는 확률변수 X에 대한 설명으로 옳지 않은 것은?

① 확률변수 $X_1, \ldots, X_n$이 서로 독립이고 동일한 성공확률 p를 가지는 베르누이 확률변수일 때 $X = \sum_{i=1}^{n} X_i$이다.

② X의 기대값은 np, 분산은 $np(1-p)$이다.

③ X가 가질 수 있는 값은 1부터 n까지의 자연수이다.

④ 확률변수 Y가 이항분포 $B(m, p)$를 따르고 X와 Y가 독립일 때, $X + Y$는 이항분포 $B(n+m, p)$를 따른다(m은 자연수).

정답 3번

간단한 해설

이항분포는 성공 횟수를 나타내므로 **한 번도 성공하지 않는 0도 가능한 값**이다. 가능한 성공 횟수의 범위는 0부터 n까지이며, 1부터라고 한 설명이 잘못되었다.

선지별 해설

① O: 각 시행의 성공을 1, 실패를 0으로 나타내면 성공확률이 p인 베르누이 확률변수가 된다. 서로 독립인 n번의 시행에서 성공한 횟수는 $X = \sum_{i=1}^{n} X_i$이므로 이 합이 이항분포 $B(n, p)$를 따른다. **시행의 독립성과 동일한 성공확률**이 이항분포의 조건이다.

② O: 각 베르누이 확률변수의 기대값은 p, 분산은 $p(1-p)$이다. 기대값의 선형성에 의해 $E(X) = \sum_{i=1}^{n} E(X_i) = np$이고, 시행들이 독립이므로 공분산이 0이어서 $\mathrm{Var}(X) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = np(1-p)$이다.

③ X: X는 n번 중 성공한 횟수이므로 **0을 포함한 정수**를 값으로 갖는다. 모두 실패하면 0, 모두 성공하면 n이 되며, 일반적인 $0 < p < 1$에서 가능한 값은 $0, 1, \ldots, n$이다. 따라서 '1부터 n까지'가 아니라 '0부터 n까지'로 고쳐야 한다. 경계인 p = 0이나 p = 1에서는 각각 0 또는 n에 확률이 집중된다.

④ O: X와 Y는 각각 성공확률이 같은 n번과 m번의 시행에서 얻은 성공 횟수이다. 두 확률변수가 독립이면 두 묶음의 성공 횟수를 합한 분포는 **성공확률이 같은 독립 이항분포의 가법성**에 따라 $B(n+m, p)$가 된다. 성공확률이 다르거나 X와 Y가 독립이 아니면 이 결론을 일반적으로 적용할 수 없다.

NoteBase

문 26. 분포에 대한 아래의 설명 중 항상 옳은 것은?

① 표준정규분포는 t 분포에 비해 꼬리가 두껍다.

② 두 확률변수 Z_1, Z_2가 표준정규분포를 따를 때, $Z_1^2 + Z_2^2$은 자유도가 2인 카이제곱 분포를 따른다.

③ t 분포를 따르는 확률변수를 제곱한 확률변수는 F 분포를 따른다.

④ t 분포는 모분산에 대한 추론 시 사용하는 분포이다.

정답 3번

간단한 해설

자유도가 n인 t 확률변수를 제곱하면 **분자 자유도 1, 분모 자유도 n인 F 분포**가 된다. 표준정규변수의 제곱합이 카이제곱분포가 되려면 변수 사이의 독립성도 필요하다.

선지별 해설

① X: 유한한 자유도의 t 분포는 표준정규분포보다 꼬리가 두껍다. 따라서 중심에서 멀리 떨어진 값이 나올 확률이 상대적으로 크며, 자유도가 커질수록 표준정규분포에 가까워진다. 선지는 두 분포의 꼬리 두께를 반대로 설명했다.

② X: 각각 표준정규분포를 따른다는 조건만으로는 충분하지 않고 **두 확률변수가 서로 독립**이어야 한다. 독립인 경우 $Z_1^2 + Z_2^2$가 자유도 2인 카이제곱분포를 따른다. 예를 들어 $Z_2 = Z_1$이면 둘 다 표준정규분포를 따르지만 제곱합은 $2Z_1^2$이므로 자유도 2인 카이제곱분포가 아니다.

③ O: 독립인 $Z \sim N(0,1)$과 $U \sim \chi_n^2$에 대해 $T = \frac{Z}{\sqrt{U/n}}$는 자유도 n인 t 분포를 따른다. 이를 제곱하면 $T^2 = \frac{Z^2/1}{U/n}$이고, 분자의 Z^2는 자유도 1인 카이제곱분포를 따른다. 따라서 **독립인 두 카이제곱변수를 각각 자유도로 나눈 비**의 정의에 따라 $T^2 \sim F_{1,n}$이 된다.

④ X: t 분포는 대표적으로 정규모집단에서 모분산을 모를 때 **모평균에 대한 추론**에 사용한다. 표본크기가 n일 때 $(\bar{X}-\mu)/(S/\sqrt{n})$는 자유도 n−1인 t 분포를 따른다. 정규모집단의 모분산 자체를 추론할 때는 $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$를 이용하므로 카이제곱분포가 해당한다.

NoteBase

문 27. 어느 회사에서 하나의 새로운 상품을 개발하였다. 기존 고객 명단에서 임의로 100명을 추출하여 신상품 구입 예약을 받은 결과 그 중 50명이 구입을 희망하였다. 이 표본자료에 기반하여 신상품 구입을 원하는 고객의 비율 p를 추정하고자 한다. 이 추정량의 표준오차의 추정치는?

① 0.01

② 0.05

③ 0.1

④ 0.2

정답 2번

간단한 해설

표본비율은 p=50/100=0.5이다. 표본비율의 표준오차에서 모비율을 **표본비율로 대체**하면 $\sqrt{0.5(1-0.5)/100} = 0.05$가 된다.

계산 과정

표본크기는 n = 100이고 구입 희망 고객 수는 50명이다. 모비율 p의 추정량으로 표본비율 p를 사용하므로 관측된 추정값은 p=(50)/(100)=0.5이다.

일반적인 독립 베르누이 표본모형에서 표본비율의 분산은 Var(p)=(p(1-p))/(n)이다. 표준오차는 이 분산의 제곱근이며, 알 수 없는 모비율 p를 표본비율로 바꾸어 추정한다. 따라서 SE(p)=p(1-p)n을 적용한다. 모집단 크기가 주어지지 않았으므로 유한모집단 보정은 적용하지 않는 통상적인 추정식이다.

값을 대입하면 SE(p)=(0.5×0.5)/(100)=√(0.0025)=0.05이다. 0.0025는 **분산의 추정치**이고, 이를 제곱근으로 바꾼 0.05가 **표준오차의 추정치**이다.

따라서 추정된 표준오차는 0.05로 ②와 일치한다. 비율의 단위로 0.05는 5%포인트에 해당한다.

NoteBase

문 28. 정규모집단의 평균 μ의 구간추정에 대한 설명 중 옳은 것은?

① 같은 신뢰수준을 가지는 조건 하에서 신뢰구간의 길이가 길수록 더 좋은 구간추정량이다.

② 대한민국 성인 평균 키 μ에 대한 95% 구간추정치가 (157, 172)이면 $P(157 < \mu < 172) = 0.95$이다.

③ 모든 조건이 동일할 때, 표본의 수가 커지면 신뢰구간의 길이가 짧아진다.

④ 모든 조건이 동일할 때, 신뢰수준(confidence level)이 95%에서 99%로 커지면 신뢰구간의 길이가 짧아진다.

정답 3번

간단한 해설

다른 조건이 같으면 표본크기가 커질수록 **표본평균의 표준오차가 작아져** 신뢰구간이 짧아진다. 모표준편차를 아는 경우 구간 길이는 $1/\sqrt{n}$에 비례한다.

선지별 해설

① X: 같은 신뢰수준에서 다른 조건이 같다면 신뢰구간이 **짧을수록 추정의 정밀도가 높다**. 긴 구간은 모평균의 위치를 더 넓은 범위로 제시하므로, 같은 포함 확률을 유지한다면 더 좋은 추정이라고 할 수 없다.

② X: 빈도주의 구간추정에서 모평균 μ는 고정된 값이고 표본에 따라 달라지는 것은 신뢰구간이다. 95% 신뢰수준은 같은 방법으로 표본추출과 구간 계산을 반복할 때 **그 구간들의 약 95%가 실제 모평균을 포함한다**는 뜻이다. 이미 계산된 고정 구간 (157, 172)은 μ를 포함하거나 포함하지 않으므로, 그 구간에 μ가 있을 확률을 0.95라고 해석하지 않는다.

③ O: 모표준편차 σ를 아는 정규모집단에서 모평균의 신뢰구간 길이는 $2z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$이다. 신뢰수준과 σ가 같으면 표본크기 n이 커질수록 표준오차가 작아져 구간이 짧아진다. 모표준편차를 몰라 t 구간을 쓰는 경우에도 표본표준편차 등 다른 조건이 같다면 n의 증가와 자유도 증가가 구간 폭을 줄이는 방향으로 작용한다.

④ X: 다른 조건이 같을 때 신뢰수준을 높이면 더 높은 포함 비율을 확보해야 하므로 신뢰구간은 **길어진다**. 표준정규분포를 이용한 양측 구간의 임계값은 95%에서 약 1.96, 99%에서 약 2.576이므로 같은 표준오차에 더 큰 값을 곱하게 된다.

NoteBase

문 29. 다음은 어떤 통계적 방법론에 대한 설명인가?

설명변수 X가 주어졌을 때, 반응변수 Y의 조건부 기댓값 $E(Y|X=x)$를 추론하고자 하는 통계적 방법론

① 분산분석
② 회귀분석
③ 군집분석
④ 표본조사 방법론

정답 2번

간단한 해설

설명변수가 주어졌을 때 반응변수의 **조건부 기댓값**을 추론하는 것은 회귀분석의 목적이다. 개별 관측값 자체가 아니라, 설명변수의 값에 따른 반응변수의 평균적인 관계를 다룬다.

공통 해설

회귀분석은 설명변수와 반응변수의 관계를 조건부 평균 함수로 나타내고, 자료를 이용해 그 관계를 추정하는 통계적 방법이다. 지문의 $E(Y \mid X=x)$는 설명변수의 값이 x로 주어졌을 때 반응변수 Y의 평균을 뜻한다. 이 평균은 **개별 반응변수 값과 구별**되며, 같은 설명변수 값을 갖더라도 개별 관측값은 서로 다를 수 있다.

예를 들어 설명변수가 공부 시간이고 반응변수가 시험 점수라면, 공부 시간이 같은 학생들의 평균 점수가 공부 시간에 따라 어떻게 달라지는지를 추정하는 것이다. 이처럼 지문은 설명변수가 주어진 조건에서 반응변수의 평균을 추론하는 회귀분석을 가리킨다.

NoteBase

문 30. 분산분석(Analysis of Variance, ANOVA)에 대한 설명 중 옳지 않은 것은?

① 여러 집단의 분산을 비교하기 위한 분석 방법이다.

② 자료의 총변동을 분해하여 집단 간 차이가 있는지 파악한다.

③ 검정통계량의 분포는 귀무가설 하에서 F 분포를 따른다.

④ 귀무가설이 기각되었을 시 사후검정으로 다중비교를 실시한다.

정답 1번

간단한 해설

분산분석은 집단별 분산 자체가 아니라 **여러 집단의 모평균이 같은지** 검정하는 방법이다. 총변동을 집단 간 변동과 집단 내 변동으로 나누어 평균 차이가 우연한 오차에 비해 큰지 판단한다.

선지별 해설

① X: 분산분석의 검정 대상은 **집단 간 모평균의 차이**이다. 집단이 k개이면 귀무가설은 $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ 이고, 대립가설은 모든 모평균이 같지는 않다는 것이다. 여러 집단의 분산 자체가 같은지 비교하는 검정과 구별한다.

② O: 일원분산분석에서는 각 관측값이 전체 평균에서 벗어나는 총변동을 집단 평균 간 차이에 따른 변동과 같은 집단 안의 오차변동으로 나눈다. 집단 간 변동이 집단 내 변동에 비해 충분히 크면 모평균이 모두 같다는 설명을 기각한다.

③ O: 고전적 일원분산분석에서 관측값의 독립성, 집단별 정규성, 등분산 가정하에 검정통계량은 $F = \frac{MS_B}{MS_W}$이다. k개 집단, 총 N개 관측값이면 귀무가설하에서 분자 자유도 k−1, 분모 자유도 N−k인 F 분포를 따른다. MS_B는 집단 간 제곱합을 k−1로, MS_W는 집단 내 제곱합을 N−k로 나눈 평균제곱이다.

④ O: 전체 F 검정에서 귀무가설을 기각해도 **어느 집단 사이에 차이가 있는지**는 바로 알 수 없다. 구체적인 차이를 확인하려면 사후 다중비교를 실시하며, 여러 비교에 따른 제1종 오류의 증가를 제어하는 방법을 사용한다. 예를 들어 Tukey 방법은 집단 평균의 쌍별 차이를 비교한다.

NoteBase

문 31. 통계적 가설검정에서 귀무가설이 사실임에도 불구하고 귀무가설을 기각할 오류를 범할 확률의 최대 허용 한계를 의미하는 통계 용어는 무엇인가?

① 검정력

② 제 I종 오류

③ 유의수준

④ 유의확률

정답 3번

간단한 해설

참인 귀무가설을 기각하는 것은 제1종 오류이고, 그 오류 확률에 대해 미리 정한 **최대 허용 한계가 유의수준**이다. 오류 자체와 오류 확률의 허용 기준을 구별한다.

개념 구분

유의수준 α: 귀무가설이 참인데도 기각할 확률을 어느 수준까지 허용할지 검정 전에 정한 기준이다. 이 문항은 오류의 이름이 아니라 **오류 확률의 최대 허용 한계**를 묻는다. 예를 들어 유의수준 0.05는 제1종 오류 확률이 5%를 넘지 않도록 검정을 설계한다는 뜻이다.

제1종 오류: 실제로 참인 귀무가설을 기각하는 잘못된 판단 그 자체이다. 유의수준은 이 오류를 통제하는 확률 기준이므로 두 용어는 동일하지 않다.

검정력: 특정 대립가설이 참일 때 귀무가설을 올바르게 기각할 확률이다. 참인 귀무가설을 잘못 기각할 확률의 허용 한계가 아니라, 거짓인 귀무가설을 탐지하는 능력을 나타낸다.

유의확률(p값): 귀무가설을 전제로 관측된 검정통계량만큼 또는 그보다 더 대립가설 방향으로 극단적인 결과가 나올 확률이다. 표본자료에서 계산하며, 미리 정한 유의수준과 비교하여 기각 여부를 판단한다. 유의수준 자체나 귀무가설이 참일 확률을 뜻하지 않는다.

따라서 요구한 통계 용어는 유의수준이며 선택지 ③과 일치한다.

NoteBase

문 32. 어느 한 선거구에서 지역에 따라 A, B, C 세 정당 후보자에 대한 지지도가 다른지 알아보기 위해 각 도시에서 100명씩 조사한 결과 다음의 자료를 얻었다. 지역에 따라 정당 후보자 지지도가 다른지 검정하고자 하는 카이제곱검정 통계량의 자유도는?

	A	B	C	계
지역 1	35	30	35	100
지역 2	50	26	24	100
지역 3	40	33	27	100
지역 4	30	50	20	100

① 6

② 9

③ 12

④ 20

정답 1번

간단한 해설

분할표의 범주는 **지역 4개와 정당 3개**이며, '계'는 범주에 포함하지 않는다. 카이제곱 동질성검정의 자유도는 $(4-1)(3-1)=6$이다.

NoteBase

계산 과정

각 지역에서 100명씩 조사하여 정당 지지도의 분포가 같은지 비교하므로 카이제곱 동질성검정을 적용한다. 귀무가설은 네 지역의 정당별 지지 비율이 모두 같다는 것이다.

표에는 지역 1~4의 네 행과 정당 A, B, C의 세 열이 있다. 따라서 $r = 4$, $c = 3$이다. 오른쪽의 '계'는 각 행의 합계 100명을 나타내므로 **별도의 범주로 세지 않는다**.

자유도 공식은 $df = (r - 1)(c - 1)$이다. 각 지역에서 세 정당 비율의 합이 1이므로 독립적인 비율은 두 개씩 있어 대립가설 아래에서는 $4 \times 2 = 8$개이고, 귀무가설 아래에서는 모든 지역에 공통인 두 개뿐이다. 그 차이인 $8 - 2 = 6$이 자유도가 된다.

공식에 대입하면 $df = (4 - 1)(3 - 1) = 3 \times 2 = 6$이다. 100명이라는 표본크기나 개별 도수는 검정통계량 계산에 사용하지만, 이 문제의 자유도는 범주의 수로 결정한다. 따라서 자유도는 6으로 ①과 일치한다.

NoteBase

문 33. 한 개의 주사위와 한 개의 동전을 동시에 던질 때, 주사위의 짝수의 눈과 동전의 앞면이 나올 확률은?

① $\frac{1}{8}$
② $\frac{1}{6}$
③ $\frac{1}{4}$
④ $\frac{1}{2}$

정답 3번

간단한 해설

주사위의 짝수 사건과 동전의 앞면 사건은 **서로 독립이므로 각 확률을 곱한다**. 따라서 구하는 확률은 $\frac{3}{6} \times \frac{1}{2} = \frac{1}{4}$이다.

계산 과정

주사위와 동전이 공정하고 두 시행이 서로 독립인 통상적인 조건에서 계산한다. 주사위의 짝수 눈이 나오는 사건을 A, 동전의 앞면이 나오는 사건을 B라 하자.

주사위의 가능한 눈은 1부터 6까지이고, 짝수 눈은 2, 4, 6의 세 가지이다. 따라서 $P(A) = \frac{3}{6} = \frac{1}{2}$이다. 동전의 앞면과 뒷면은 같은 확률로 나오므로 $P(B) = \frac{1}{2}$이다.

두 조건이 **동시에 성립하는 교집합 사건**의 확률을 구한다. 독립사건의 곱셈법칙을 적용하면 $P(A \cap B) = P(A)P(B) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$이다. 전체 12가지 결과 중 (2, 앞면), (4, 앞면), (6, 앞면)의 3가지가 해당한다고 세어도 같은 결과가 나온다.

구한 확률 $\frac{1}{4}$은 단위가 없는 값이며, 이에 해당하는 선택지는 ③이다.

NoteBase

문 34. 남학생과 여학생에 따라 영어, 수학, 국어, 세 과목에 대한 선호도가 다른지 조사한 결과가 아래 표와 같다. “남학생과 여학생의 선호도가 같다”는 귀무가설 하에서 남학생 영어 선호도의 추정된 기대돗수는?

	영어	수학	국어	표본 크기
남학생	100	80	70	250
여학생	110	60	80	250
합계	210	140	150	500

① 102

② 103

③ 104

④ 105

정답 4번

간단한 해설

두 집단의 선호도가 같다는 가정에서는 **전체의 영어 선호 비율을 남학생 수에 적용**한다. 남학생 영어 선호도의 추정 기대도수는 $250 \times \frac{210}{500} = 105$명이다.

계산 과정

기대도수는 귀무가설이 성립할 때 각 칸에서 기대되는 인원수이다. 남학생과 여학생의 선호도가 같다면 두 집단은 과목별 선호 확률을 공유하므로, 두 집단을 합친 표본에서 영어 선호 비율을 추정한다.

표에서 남학생 수는 250명이고, 영어를 선호하는 학생은 남학생 100명과 여학생 110명을 합한 210명이다. 전체 표본은 500명이므로 공통 영어 선호 비율의 추정값은 $\frac{210}{500}$이다. 수학·국어 선호자도 전체 표본 500명에 포함되며, 분모를 영어 선호자나 남학생 수로만 잡지 않는다.

남학생 영어 칸의 기대도수는 남학생 수에 이 공통 비율을 곱한다. 이는 해당 행 합계와 열 합계의 곱을 전체 합계로 나누는 계산이다.
$E = 250 \times \frac{210}{500} = 105$
표의 100명은 실제 관측도수이고, 105명은 **동일한 선호도를 가정해 추정한 기대도수**이므로 서로 구별한다.

따라서 기대도수는 105명이며 선택지 ④와 일치한다.

NoteBase

문 35. 로지스틱 회귀분석에 대한 설명 중 옳지 않은 것은?

① 설명변수와 반응변수가 모두 범주형변수일 때에만 적용 가능한 분석 방법이다.

② 로짓(logit) 연결함수를 사용해 예측값을 0과 1 사이로 제한한다.

③ 오즈비(odds ratio)를 이용해 모형 해석이 가능하다.

④ 금융회사에서 고객의 신용카드 발급 여부를 판단할 때 사용할 수 있는 분석 방법이다.

정답 1번

간단한 해설

로지스틱 회귀분석은 범주형 반응변수의 발생확률을 모형화한다. **설명변수는 범주형뿐 아니라 연속형도 사용할 수 있다**는 점이 핵심이다.

선지별 해설

① X: 이항 로지스틱 회귀에서는 반응변수가 성공·실패처럼 두 범주로 나뉘지만, **설명변수까지 범주형일 필요는 없다**. 나이·소득처럼 연속적인 변수도 넣을 수 있고, 범주형 설명변수는 보통 지시변수로 표현한다. 따라서 두 변수가 모두 범주형일 때에만 적용한다는 제한이 잘못되었다.

② O: 성공확률을 p라 하면 로짓 연결함수는 $\log\frac{p}{1-p}=\eta$이고, $\eta=\beta_0+\beta_1x_1+\cdots+\beta_kx_k$는 선형예측값이다. 이를 확률로 되돌리는 **역로짓 변환**을 적용하면 $p=\frac{e^{\eta}}{1+e^{\eta}}$가 되어 유한한 선형예측값에 대한 예측확률이 0보다 크고 1보다 작아진다. 0과 1 사이로 제한되는 것은 로짓 값 자체가 아니라 예측확률이다.

③ O: 오즈는 사건이 일어날 확률과 일어나지 않을 확률의 비 $p/(1-p)$이다. 다른 설명변수를 고정하고 상호작용항이 없는 모형에서 해당 설명변수가 1단위 증가하면 오즈는 e^{β_j}배가 된다. 이 오즈비가 1보다 크면 오즈가 증가하고, 1보다 작으면 감소한다. **오즈비는 확률 자체의 비와는 다르다**.

④ O: 발급·미발급처럼 두 범주로 구분되는 결과는 이항 로지스틱 회귀로 모형화할 수 있다. 고객 특성을 설명변수로 사용해 한 범주에 속할 확률을 추정한 뒤 기준값과 비교하여 분류하는 방식이 가능하다. 실제 발급 결정에는 별도의 심사 기준 등이 함께 적용될 수 있으나, 이 방법을 분류에 활용할 수 있다는 설명은 타당하다.

NoteBase

문 36. 강우량 정도에 따른 대기오염 제거의 정도를 알아보기 위해 데이터를 이용하여 추정된 회귀직선 $\hat{y} = 110 - 5x$를 얻었다. 강우량이 $x = 10$일 때, 대기오염 제거 정도의 추정값은?

① 10

② 30

③ 40

④ 60

정답 4번

간단한 해설

추정된 회귀직선에 **강우량 $x = 10$을 대입**하면 "0177=110-5×10=60이다.

계산 과정

주어진 회귀직선은 "0177=110-5x이다. 여기서 x는 강우량이고, "0177는 그 강우량에서의 대기오염 제거 정도의 추정값이다. 절편은 110, 기울기는 −5이다.

강우량이 $x = 10$이므로 "0177=110-5×10=110-50=60이다. 기울기에 10을 곱한 값은 −50이지만, 추정값을 구하려면 **절편 110을 함께 반영**해야 한다.

따라서 대기오염 제거 정도의 추정값은 60으로 ④와 일치한다.

NoteBase

문 37. 회귀분석을 통해 다음과 같은 분산분석표를 얻었다.
오차항의 모분산 σ^2의 추정치는?

요인	제곱합	자유도	평균제곱	F 값
회귀	200	1	200	4.0
잔차	300	6	50	
계	500	7		

① 50
② 200
③ 300
④ 500

정답 1번

간단한 해설

회귀모형의 오차 모분산은 **잔차평균제곱(MSE)**으로 추정한다. 표에서 잔차평균제곱이 50이므로 오차 모분산의 추정치는 50이다.

NoteBase

핵심 풀이

오차항의 모분산은 회귀식으로 설명되지 않는 변동의 크기이다. 통상적인 등분산 회귀모형에서 그 불편추정량은 잔차제곱합을 잔차 자유도로 나눈 **잔차평균제곱(MSE)**이다. 회귀계수를 자료로 추정하면서 사용한 자유도를 반영하기 위해 잔차 자유도로 나눈다.

표의 '잔차' 행을 보면 제곱합은 300, 자유도는 6, 평균제곱은 50이다.
σ^2=MSE=(300)/(6)=50
이미 제시된 평균제곱 50이 구하려는 모분산의 추정치이므로 다시 자유도로 나누지 않는다.

200은 회귀제곱합이면서 회귀 자유도가 1이어서 회귀평균제곱과도 같은 값이다. 300은 잔차제곱합으로, 아직 잔차 자유도로 나누기 전의 값이며, 500은 총제곱합이다. 따라서 이 값들을 오차 모분산의 추정치와 혼동하지 않는다. F값 4.0도 회귀평균제곱과 잔차평균제곱의 비율이지 분산 추정치 자체는 아니다.

따라서 오차 모분산의 추정치는 50이며 선택지 ①과 일치한다.

NoteBase

문 38. 어떤 제품을 만들 때 처리액의 농도가 제품의 강도에 영향을 미치는지 조사하기 위해 네 가지 농도에서 각 농도 당 5회를 반복하여 총 20회의 실험을 랜덤하게 처리한 후 강도를 측정했다. 이 자료에 일원배치법을 적용하여 작성한 분산분석표가 아래와 같다. (ㄱ)의 F값은?

요인	제곱합	자유도	평균제곱	F값	p-값
처리	1800	3		(ㄱ)	
잔차	1600	16			
합계					

① 6

② 7

③ 8

④ 9

정답 1번

간단한 해설

분산분석의 F값은 **처리평균제곱을 잔차평균제곱으로 나눈 값**이다. 따라서 $F = \frac{1800/3}{1600/16} = \frac{600}{100} = 6$이다.

계산 과정

농도 수준은 4개이고 각 수준에서 5회씩 반복했으므로 전체 관측값 수는 $N = 4 \times 5 = 20$이다. 일원배치 분산분석에서 처리 자유도는 $4 - 1 = 3$, 잔차 자유도는 $20 - 4 = 16$으로, 표의 자유도와 일치한다.

평균제곱은 **각 제곱합을 해당 자유도로 나눈 값**이다. 표에서 처리 제곱합은 1800이므로 처리평균제곱은 $MS_T = 1800/3 = 600$이다. 이는 농도별 평균의 차이에 따른 변동을 자유도 1당으로 환산한 값이다.

잔차 제곱합은 1600이므로 잔차평균제곱은 $MS_E = 1600/16 = 100$이다. 이는 같은 농도 안의 우연한 변동을 바탕으로 오차분산을 추정한 값이다.

농도별 모평균이 모두 같은지 검정할 때는 처리에 따른 변동을 오차변동과 비교하므로 $F = MS_T/MS_E = 600/100 = 6$을 사용한다. **제곱합끼리 바로 나누거나 분자·분모를 바꾸지 않는다**. 두 평균제곱은 같은 단위를 가지므로 F값에는 단위가 없다. 따라서 (ㄱ)은 6이며, 해당 선택지는 ①이다.

NoteBase

문 39. 독립변수 x와 종속변수 y 사이에 직선 관계를 가정하고 단순선형회귀 모형 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, 2, \cdots, n$ (ε_i들은 서로 독립이고, ε_i들이 각각 정규분포 $N(0, \sigma^2)$을 따름)을 적용시켰다. β_0, β_1의 최소제곱추정량을 각각 b_0, b_1이라 하고, x=(1)/(n) Σ _i=1^nx_i, y=(1)/(n) Σ _i=1^ny_i, $\hat{y}_i = b_0 + b_1 x_i$라 하자. 아래의 내용 중 틀린 것은?

① b_1= Σ _i=1^n(x_i-x)(y_i-y) Σ _i=1^n(x_i-x)^2

② $b_0 = \bar{y} - b_1 \bar{x}$

③ (1)/(n-1) Σ _i=1^n(y_i-y_i)^2은 σ^2의 불편(비편향)(unbiased) 추정량이다.

④ Σ _i=1^ny_i= Σ _i=1^ny_i

정답 3번

간단한 해설

절편과 기울기 두 계수를 추정하므로 잔차 자유도는 **n − 2**이다. 오차 모분산의 불편추정량은 잔차제곱합을 n − 1이 아니라 n − 2로 나눈 값이다.

NoteBase

선지별 해설

① O: 최소제곱법은 잔차제곱합을 최소로 만드는 절편과 기울기를 구한다. x값들이 모두 같지 않을 때 기울기는 다음과 같다.
$b_1 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$
분자는 x와 y의 함께 변하는 정도이고, 분모는 x의 변동을 나타내므로 x의 변화에 대한 y의 선형 변화율을 추정한다.

② O: 절편에 관한 정상방정식에서 잔차의 합이 0이므로 $\overline{y} = b_0 + b_1\overline{x}$이다. 따라서 $b_0 = \overline{y} - b_1\overline{x}$가 성립한다. 이는 **절편을 포함한 최소제곱회귀선이 표본평균점** $(\overline{x}, \overline{y})$을 지난다는 뜻이다.

③ X: 절편과 기울기 두 모수를 추정하므로 잔차 자유도는 n − 2이다. 잔차제곱합을 SSE라고 하면 $E(\text{SSE}) = (n-2)\sigma^2$이므로 불편추정량은 다음과 같다.
σ^2=SSEn-2=(1)/(n-2) Σ _i=1^n(y_i-"0177_i)^2
선지처럼 n − 1로 나누면 기댓값이 $\frac{n-2}{n-1}\sigma^2$이 되어 모분산을 작게 추정한다. 평균 하나만 추정하는 표본분산의 자유도 n − 1을 **그대로 적용할 수 없다**. 잔차분산을 추정하려면 양의 잔차 자유도가 필요하다.

④ O: 절편을 포함한 최소제곱법에서는 Σ _i=1^n(y_i-"0177_i)=0이라는 정상방정식이 성립한다. 이를 정리하면 Σ _i=1^ny_i=Σ _i=1^n"0177_i이다. 관측값의 합과 적합값의 합이 같은 것은 이 **잔차합이 0이라는 성질**에 따른다.

따라서 틀린 진술은 잔차분산 추정의 분모를 n − 1로 제시한 ③이다.

NoteBase

문 40. 평균이 μ, 분산이 σ^2인 정규모집단으로부터 크기 n인 확률표본 $X_1, \dots, X_n$을 추출하였을 때, 표본평균 $\overline{X}_n$, 표본분산 S_n^2, 표본표준편차 S_n이 다음과 같다.

$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2,$

$S_n = \sqrt{S_n^2}, \quad S_n > 0.$

다음 설명 중 옳지 않은 것은?

① S_n^2은 σ^2의 불편(비편향)(unbiased) 추정량이다.

② $\frac{(n-1)S_n^2}{\sigma^2}$은 자유도가 $n-1$인 카이제곱 분포를 따른다.

③ $\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}$은 표준정규분포를 따른다.

④ $\overline{X}_n$은 μ의 불편 추정량이다.

정답 3번

간단한 해설

정규모집단에서도 모표준편차 대신 표본표준편차를 쓰면, 표본평균을 표준화한 통계량은 **표준정규분포가 아니라 자유도 n−1인 t분포**를 따른다.

선지별 해설

① O: 불편 추정량은 기댓값이 추정하려는 모수와 같은 추정량이다. 표본평균으로부터의 편차는 합이 0이 되어 자유도 하나가 줄어든다. 이때 $E[\sum_{i=1}^{n}(X_i - \overline{X}_n)^2] = (n-1)\sigma^2$이므로, 분모를 **n이 아닌 n−1**로 둔 표본분산은 $E[S_n^2] = \sigma^2$를 만족한다.

② O: 정규모집단의 독립인 확률표본에서는 $\frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}{\sigma^2} = \frac{(n-1)S_n^2}{\sigma^2}$이 **자유도 n−1인 카이제곱분포**를 따른다. 모평균 대신 표본평균을 사용하여 편차를 구하므로 n개의 편차에 합이 0이라는 제약 하나가 생긴다.

③ X: 정규모집단에서는 $Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$이 표준정규분포를 따른다. 또한 $U = \frac{(n-1)S_n^2}{\sigma^2}$은 자유도 n−1인 카이제곱분포를 따르며 Z와 독립이다. 따라서 $\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} = \frac{Z}{\sqrt{U/(n-1)}}$은 **자유도 n−1인 t분포**를 따른다. 표본표준편차를 사용하는 분모에도 표본에 따른 변동이 있으므로 표준정규분포라고 할 수 없다.

④ O: 표본평균의 기댓값은 $E[\overline{X}_n] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{n\mu}{n} = \mu$이다. 따라서 표본평균은 **모평균의 불편 추정량**이다. 이 성질은 각 표본의 기댓값이 모평균이고 기댓값이 존재하면 성립하며, 정규분포라는 조건 자체가 꼭 필요한 것은 아니다.

NoteBase