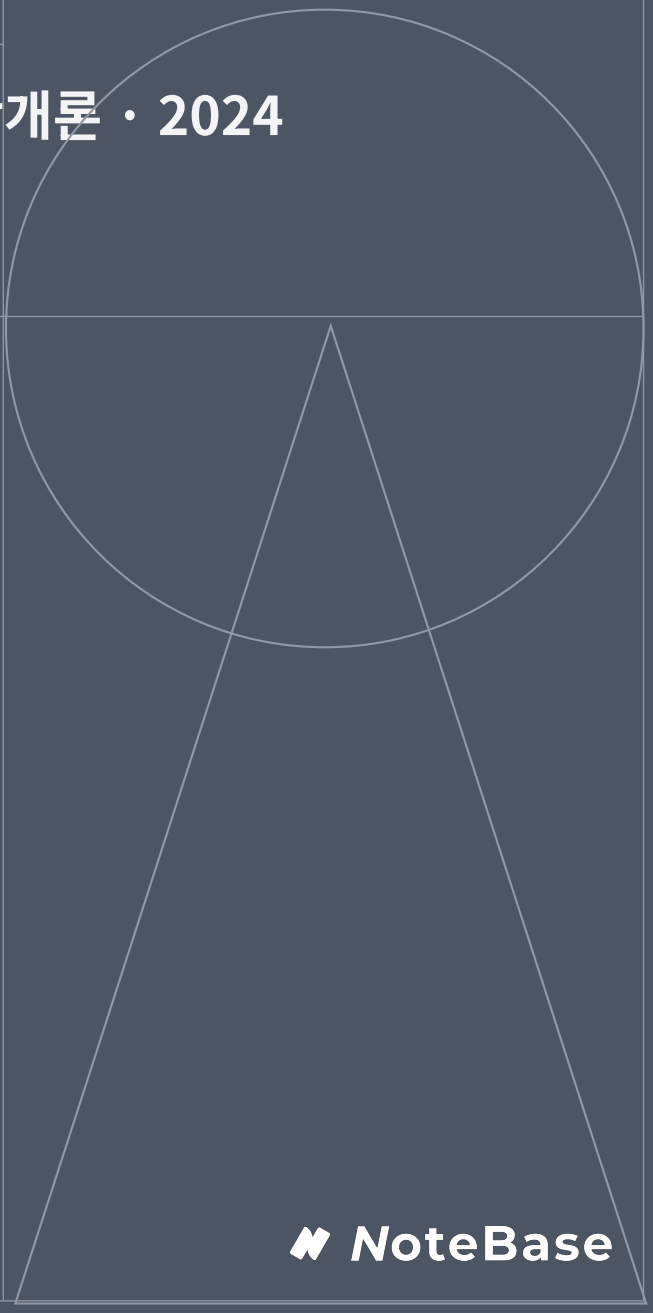
기출OX 해설집

공무원/국가직/9급 · 통계학개론 · 2024



NoteBase

기출OX 해설집

공무원/국가직/9급 · 통계학개론 · 2024

공무원/국가직/9급 · 통계학개론 · 2024

문 1. 다음은 대통령 선거 출구 조사 방법에 대한 설명이다. 이에 해당하는 표본 추출 기법은?

모든 투표자를 대상으로 조사하는 것이 아니라, 투표자들이 나오는 것을 조사원이 출구 근처에서 보다가 일정 간격으로 투표자를 추출해 조사합니다. 간격이 너무 넓은 경우에는 조사원의 집중력 저하로 조사해야 할 투표자를 놓칠 가능성이 큽니다. 반대로 간격이 좁으면 조사원이 너무 바빠 응답자를 정확하게 선정하지 못할 수 있습니다. 16대 대선의 경우는 간격이 6이었고, 19대 대선은 5로 정해 조사했습니다.

① 단순임의추출법(simple random sampling)
② 계통추출법(systematic sampling)
③ 층화임의추출법(stratified random sampling)
④ 집락추출법(cluster sampling)

정답 2번

간단한 해설

투표자가 나오는 순서에 따라 **일정한 간격마다** 조사 대상을 뽑는 방식은 계통추출법이다. 추출 간격이 6에서 5로 달라져도 같은 간격을 반복해 적용한다는 원리는 같다.

공통 해설

계통추출법은 일정한 순서로 늘어선 대상에서 첫 추출 대상을 정한 뒤, **정해진 간격마다 하나씩** 표본을 뽑는 방법이다. 확률표본추출로 시행할 때는 첫 추출 대상을 무작위로 정하고 이후에는 같은 간격을 적용한다.

지문에서는 투표자가 출구로 나오는 순서를 따라 일정 간격으로 조사 대상을 선정한다. 간격이 6이면 6명마다 한 명씩, 간격이 5이면 5명마다 한 명씩 선정한다는 뜻이다. 이처럼 대상이 나오는 순서에 고정된 추출 간격을 적용하는 것이 계통추출법을 식별하는 기준이다.

NoteBase

문 2. 세 사건 A, B, C가 서로 독립이고 $P(A)=\frac{1}{4}$, $P(B)=3P(C)$, $P(A\cap B^c\cap C)=\frac{1}{48}$일 때, $P(B\cap C^c)$는?

① $\frac{1}{12}$
② $\frac{5}{12}$
③ $\frac{7}{12}$
④ $\frac{11}{12}$

정답 2번

간단한 해설

세 사건이 서로 독립이므로 **여사건이 포함된 교집합도 각 확률의 곱**으로 계산한다. 주어진 조건에서 $P(B)=\frac{1}{2}$, $P(C)=\frac{1}{6}$이므로, $P(B\cap C^c)=\frac{5}{12}$이다.

계산 과정

확률 $P(B)$를 x라 두면, $P(B)=3P(C)$에서 $P(C)=\frac{x}{3}$이다. 또한 여사건의 확률은 $P(B^c)=1-x$이다.

세 사건이 서로 독립이면 **일부를 여사건으로 바꾸어도 독립**이므로, 주어진 교집합의 확률은 $P(A\cap B^c\cap C)=P(A)P(B^c)P(C)$이다. 따라서 $\frac{1}{4}\cdot(1-x)\cdot\frac{x}{3}=\frac{1}{48}$이다.

양변에 12를 곱하면 $x(1-x)=\frac{1}{4}$이다. 이를 정리하면 $x^2-x+\frac{1}{4}=0$, 즉 $(x-\frac{1}{2})^2=0$이므로 $x=\frac{1}{2}$이다. 이 값은 사건 B가 일어날 확률이며, $P(C)=\frac{x}{3}=\frac{1}{6}$이다.

구하려는 것은 B**는 일어나고** C**는 일어나지 않을 확률**이다. $P(C^c)=1-\frac{1}{6}=\frac{5}{6}$이고 B와 C^c도 독립이므로, $P(B\cap C^c)=\frac{1}{2}\cdot\frac{5}{6}=\frac{5}{12}$이다. 이는 ②의 값과 일치한다.

NoteBase

문 3. 확률변수 X, Y에 대하여 $E[X(Y+1)]=9$, $E[Y(X+1)]=10$, $E(X+Y)=7$일 때, 공분산 $Cov(X, Y)$의 값은?

① −6

② −5

③ 5

④ 6

정답 1번

간단한 해설

주어진 식을 기대값의 선형성으로 전개하면 $E[X]=3$, $E[Y]=4$, $E[XY]=6$을 얻는다. 공분산은 **곱의 기대값에서 각 기대값의 곱을 뺀 값**이므로 $6-3\times4=-6$이다.

계산 과정

공분산은 두 변수 곱의 기대값에서 각 기대값의 곱을 뺀 값으로, $\mathrm{Cov}(X, Y)=E[XY]-E[X]E[Y]$이다. 따라서 $E[X]$, $E[Y]$, $E[XY]$를 구한다.

기대값의 선형성을 적용하면 $E[X(Y+1)]=E[XY]+E[X]=9$, $E[Y(X+1)]=E[XY]+E[Y]=10$이다. 두 번째 식에서 첫 번째 식을 빼면 **공통인 곱의 기대값이 소거**되어 $E[Y]-E[X]=1$을 얻는다.

나머지 조건은 $E[X]+E[Y]=7$이다. 이 식과 $E[Y]-E[X]=1$을 더하면 $2E[Y]=8$이므로 $E[Y]=4$이고, 따라서 $E[X]=3$이다.

$E[XY]+E[X]=9$에 $E[X]=3$을 대입하면 $E[XY]=6$이다. 이는 **두 변수를 곱한 뒤 구한 기대값**이며, 각 기대값을 곱한 $E[X]E[Y]=12$와 구별한다.

공분산 공식에 대입하면 $\mathrm{Cov}(X, Y)=6-3\times4=-6$이다. 구한 값 −6은 ①과 일치한다.

NoteBase

문 4. '두 확률변수 X, Y의 상관계수가 0', 즉 '$Corr(X, Y) = 0$'과 동치가 아닌 것은?

① $Var(X + Y) = Var(X - Y)$

② $E(XY) = E(X)E(Y)$

③ $Cov(-2X + 3, 3Y - 2) = 0$

④ $Cov(X + Y, X - Y) = 0$

정답 4번

간단한 해설

④의 공분산은 $\mathrm{Var}(X) - \mathrm{Var}(Y)$이므로, 0이라는 조건은 **두 확률변수의 분산이 같다**는 뜻이다. 이는 원래 두 확률변수의 **상관계수가 0**이라는 조건과 동치가 아니다.

선지별 해설

상관계수가 정의되려면 두 확률변수의 분산이 유한하고 양수여야 한다. $c = \mathrm{Cov}(X, Y)$라 두면, 상관계수는 공분산을 두 표준편차의 곱으로 나눈 값이므로 $\mathrm{Corr}(X, Y) = 0$과 $c = 0$은 동치이다. 아래 O/X는 각 조건이 이 무상관 조건과 동치인지를 나타낸다.

① O: 합과 차의 분산을 전개하면 두 분산의 합은 공통으로 들어가고, 공분산 항은 각각 $+2c$와 $-2c$이다. 따라서 $\mathrm{Var}(X + Y) = \mathrm{Var}(X - Y)$는 $4c = 0$, 즉 **공분산이 0**이라는 조건과 동치이다.

② O: 공분산의 정의는 $c = E(XY) - E(X)E(Y)$이다. 따라서 $E(XY) = E(X)E(Y)$이면 공분산이 0이고, 공분산이 0이면 이 등식이 성립한다.

③ O: **상수를 더해도 공분산은 변하지 않고**, 각 확률변수에 곱한 상수는 공분산 밖으로 나온다. 따라서 $\mathrm{Cov}(-2X + 3, 3Y - 2) = -6c$이다. 계수 −6은 0이 아니므로 이 공분산이 0이라는 조건은 $c = 0$과 동치이다.

④ X: 공분산을 전개하면 $\mathrm{Cov}(X + Y, X - Y)$는 $\mathrm{Var}(X) - \mathrm{Var}(Y)$가 된다. 전개 과정에서 $-\mathrm{Cov}(X, Y)$와 $\mathrm{Cov}(Y, X)$가 서로 상쇄되기 때문이다. 따라서 이 값이 0이라는 조건은 **두 분산이 같다는 조건**으로 고쳐 이해해야 하며, 무상관 조건과는 다르다. 예를 들어 분산이 양수인 확률변수에 대해 $Y = X$이면 두 분산이 같아서 이 공분산은 0이지만, $\mathrm{Corr}(X, Y) = 1$이다.

NoteBase

문 5. 두 연속확률변수 X, Y의 상관계수에 대한 설명으로 옳지 않은 것은?

① 상관계수가 0이면 X, Y는 독립이다.

② 상관계수는 측정단위의 영향을 받지 않는다.

③ 상관계수가 −0.7인 경우가 +0.5인 경우보다 선형관계가 강하다.

④ X, Y의 표본상관계수가 0.5일 때, 동일 자료에서 $3X$와 $5Y$의 표본상관계수도 0.5이다.

정답 1번

간단한 해설

상관계수 0은 **선형관계가 없다는 뜻**이며, 비선형적인 관계까지 없다는 뜻은 아니다. 따라서 상관계수가 0이어도 두 확률변수가 독립이라고 단정할 수 없다.

선지별 해설

① X: 상관계수는 두 변수의 선형관계의 방향과 강도를 나타낸다. 상관계수가 정의되는 경우 독립이면 상관계수가 0이지만, **상관계수가 0이라고 독립인 것은 아니다**. 예를 들어 X가 −1부터 1까지의 균등분포를 따르고 $Y = X^2$이면, 대칭성에 의해 상관계수는 0이다. 그러나 X의 값이 정해지면 Y의 값도 정해지므로 두 변수는 독립이 아니다.

② O: 상관계수는 공분산을 두 변수의 표준편차의 곱으로 나눈 값이다. 분자와 분모의 측정단위가 서로 상쇄되므로 **단위가 없는 수치**이다. 따라서 길이를 m에서 cm로 바꾸는 것과 같은 단위 환산은 상관계수에 영향을 주지 않는다.

③ O: 선형관계의 강도는 상관계수의 **절댓값**으로 비교한다. −0.7의 절댓값 0.7이 +0.5의 절댓값 0.5보다 크므로, −0.7인 경우가 더 강한 선형관계를 나타낸다. 음수와 양수의 차이는 각각 한 변수가 증가할 때 다른 변수가 감소하는 방향과 증가하는 방향을 뜻한다.

④ O: 각 변수에 **양의 상수를 곱해도** 표본상관계수는 변하지 않는다. 원자료의 표본공분산을 s_{XY}, 표본표준편차를 s_X, s_Y라 하면, $3X$와 $5Y$의 표본공분산은 $15s_{XY}$, 표본표준편차는 각각 $3s_X$, $5s_Y$가 된다. 따라서 $r_{3X,\,5Y} = \frac{15s_{XY}}{(3s_X)(5s_Y)} = \frac{s_{XY}}{s_X s_Y} = 0.5$이다.

NoteBase

문 6. 정규분포 $N(\mu, \sigma^2)$에서 표준편차(σ)를 이용하여 사분위수범위를 나타낸 것으로 옳은 것은? (단, z_α는 표준정규분포의 제100×(1−α) 백분위수이다)

① $z_{0.025} \times \sigma$

② $z_{0.25} \times \sigma$

③ $2 \times z_{0.025} \times \sigma$

④ $2 \times z_{0.25} \times \sigma$

정답 4번

간단한 해설

사분위수범위는 **제75백분위수와 제25백분위수의 차이**이다. 정규분포는 평균을 중심으로 대칭이므로 두 사분위수 사이의 폭은 $2z_{0.25}\sigma$이다.

계산 과정

사분위수범위는 제3사분위수에서 제1사분위수를 뺀 $Q_3 - Q_1$로, **가운데 50%가 차지하는 구간의 폭**이다. X가 평균 μ, 표준편차 σ인 정규분포를 따르면, $Z = \frac{X-\mu}{\sigma}$는 표준정규분포를 따른다. 따라서 표준정규분포의 분위수에 σ를 곱하고 μ를 더하면 원래 분포의 분위수가 된다.

제3사분위수: Q_3는 제75백분위수이다. 문제에서 z_α는 제100×(1−α) 백분위수이므로, $z_{0.25}$가 표준정규분포의 제75백분위수이다. 따라서 $Q_3 = \mu + \sigma z_{0.25}$이며, 평균보다 $\sigma z_{0.25}$만큼 오른쪽에 있다.

제1사분위수: Q_1은 제25백분위수이다. 표준정규분포는 **0을 중심으로 대칭**이므로 제25백분위수는 $-z_{0.25}$이다. 따라서 $Q_1 = \mu - \sigma z_{0.25}$이며, 평균보다 $\sigma z_{0.25}$만큼 왼쪽에 있다.

사분위수범위 계산: 두 사분위수의 차를 구하면 평균 μ는 소거되고, 평균의 양쪽에 있는 거리가 더해져 $Q_3 - Q_1 = 2z_{0.25}\sigma$가 된다. 이는 ④의 식과 일치하며, 사분위수범위의 단위는 X와 σ의 단위와 같다.

NoteBase

문 7. 미지의 모평균이 μ이고 모분산이 σ^2인 모집단으로부터 확률표본 X_1, X_2, X_3, X_4를 추출하여 μ에 대한 점추정량을 다음과 같이 정의하였다. 이 점추정량 중에서 μ에 대한 불편추정량이면서 그 중 최소분산을 갖는 것은? (단, $\sigma^2 > 0$이다)

○ $\hat{\mu}_1 = \frac{1}{5}(X_1 + X_2 + 2X_3 + X_4)$
○ $\hat{\mu}_2 = \frac{1}{4}(X_1 + X_2 + 2X_3 + X_4)$
○ $\hat{\mu}_3 = \frac{1}{5}(2X_1 + X_2 + X_3 + 2X_4)$
○ $\hat{\mu}_4 = \frac{1}{4}(2X_1 + X_2 + X_3)$

① $\hat{\mu}_1$
② $\hat{\mu}_2$
③ $\hat{\mu}_3$
④ $\hat{\mu}_4$

정답 1번

간단한 해설

불편추정량은 **기댓값이 모평균과 같은** 추정량이다. 이 조건을 만족하는 추정량끼리 분산을 비교하면, $\hat{\mu}_1$의 분산 $\frac{7}{25}\sigma^2$이 $\hat{\mu}_4$의 분산 $\frac{3}{8}\sigma^2$보다 작다.

NoteBase

선지별 해설

확률표본이므로 각 X_i는 서로 독립이고, 기댓값은 μ, 분산은 σ^2이다. 따라서 선형결합의 기댓값은 계수의 합에 μ를 곱하고, 분산은 **계수 제곱의 합**에 σ^2을 곱해 구한다. 불편추정량인지 먼저 확인한 뒤, 불편추정량들 사이에서 분산을 비교한다.

① $\hat{\mu}_1$: 계수의 합이 $\frac{1+1+2+1}{5}=1$이므로 $E(\hat{\mu}_1)=\mu$인 불편추정량이다. 분산은 $\mathrm{Var}(\hat{\mu}_1)=\frac{7}{25}\sigma^2$이다. 여기서 분자 7은 분자 계수의 제곱합 $1^2+1^2+2^2+1^2$이고, 분모 25는 5^2이다.

② $\hat{\mu}_2$: 계수의 합이 $\frac{5}{4}$이므로 $E(\hat{\mu}_2)=\frac{5}{4}\mu$이다. 기댓값이 모든 μ에 대해 μ와 같지는 않으므로 **불편추정량이 아니다**. 분산은 $\mathrm{Var}(\hat{\mu}_2)=\frac{7}{16}\sigma^2$이다.

③ $\hat{\mu}_3$: 분자 계수의 합은 $2+1+1+2=6$이므로 $E(\hat{\mu}_3)=\frac{6}{5}\mu$이다. 따라서 **불편추정량이 아니다**. 분자 계수의 제곱합은 $2^2+1^2+1^2+2^2=10$이므로 분산은 $\mathrm{Var}(\hat{\mu}_3)=\frac{10}{25}\sigma^2$이다.

④ $\hat{\mu}_4$: 계수의 합이 $\frac{2+1+1}{4}=1$이므로 $E(\hat{\mu}_4)=\mu$인 불편추정량이다. X_4의 계수는 0이며, 분자 계수의 제곱합은 $2^2+1^2+1^2=6$이다. 따라서 분산은 $\mathrm{Var}(\hat{\mu}_4)=\frac{6}{16}\sigma^2=\frac{3}{8}\sigma^2$이다.

불편추정량은 $\hat{\mu}_1$과 $\hat{\mu}_4$이다. $\sigma^2>0$이고 $\frac{7}{25}=\frac{56}{200}<\frac{75}{200}=\frac{3}{8}$이므로, **주어진 불편추정량 중 분산이 가장 작은 것**은 ①의 $\hat{\mu}_1$이다.

NoteBase

문 8. 확률변수 X의 확률분포가 다음과 같다.
다음 항목 중 옳은 것만을 모두 고르면?

x	-1	0	1	계
$P(X=x)$	(가)	(나)	0.25	1

ㄱ. (가)+(나) = $\frac{3}{4}$
ㄴ. $E(X) = \frac{1}{8}$일 때, (나) = $\frac{3}{8}$
ㄷ. $E(X) = \frac{1}{12}$일 때, 위 모집단에서 크기가 2인 확률표본을 추출한다면 표본평균 $\overline{X}$의 분산은 $Var(\overline{X}) = \frac{59}{144}$이다.

① ㄱ
② ㄴ
③ ㄱ, ㄷ
④ ㄴ, ㄷ

정답 1번

간단한 해설

확률의 총합이 1이므로 (가)+(나)는 3/4이다. ㄴ의 기댓값 조건에서 (나)는 5/8이며, ㄷ의 59/144는 **표본평균의 분산이 아니라 모집단의 분산**이다.

NoteBase

보기별 해설

표의 두 미지 확률을 $a = P(X=-1)$, $b = P(X=0)$로 나타낸다. 기댓값은 가능한 값에 각각의 확률을 곱하여 더한 값이므로 $E(X) = -a + \frac{1}{4}$이다. X = 0에 해당하는 항은 확률이 얼마이든 기댓값 계산에서 0이 된다.

ㄱ. O: 확률분포에서는 **모든 확률을 더하면 1**이다. 따라서 $a + b + \frac{1}{4} = 1$이고, $a + b = \frac{3}{4}$이므로 (가)+(나)는 3/4이다.

ㄴ. X: $E(X) = \frac{1}{8}$을 기댓값 식에 대입하면 $-a + \frac{1}{4} = \frac{1}{8}$이므로 $a = \frac{1}{8}$이다. 따라서 $b = \frac{3}{4} - \frac{1}{8} = \frac{5}{8}$이며, **(나)는 3/8이 아니라 5/8**이다.

ㄷ. X: $E(X) = \frac{1}{12}$이면 $a = \frac{1}{4} - \frac{1}{12} = \frac{1}{6}$이다. X를 제곱하면 −1과 1은 모두 1이 되고 0은 그대로이므로 $E(X^2) = a + \frac{1}{4} = \frac{5}{12}$이다. 분산은 제곱의 기댓값에서 기댓값의 제곱을 뺀 값이므로 $Var(X) = \frac{5}{12} - \left(\frac{1}{12}\right)^2 = \frac{59}{144}$이다.

크기가 2인 확률표본의 두 관측값은 서로 독립이고 모집단과 같은 분포를 따른다. 이때 **표본평균의 분산은 모집단 분산을 표본크기로 나눈 값**이므로 $Var(\overline{X}) = \frac{Var(X)}{2} = \frac{59}{288}$이다. 보기의 59/144는 모집단 분산이며, 표본평균의 분산으로는 59/288이 맞다.

옳은 보기는 ㄱ뿐이므로, ㄱ만 제시한 ①에 해당한다.

NoteBase

문 9. 독립변수 X와 종속변수 Y를 갖는, 절편이 있는 단순선형회귀모형에서 최소제곱법을 적용했을 때, 결정계수에 대한 설명으로 옳은 것은?

① 결정계수가 0이면 모든 잔차가 0이다.

② 결정계수가 큰 모형일수록 설명력이 낮은 모형이다.

③ 결정계수는 X, Y의 표본상관계수를 이용하여 구할 수 있다.

④ 만약 새로운 독립변수를 모형에 추가한다면 결정계수는 감소한다.

정답 3번

간단한 해설

절편이 있는 단순선형회귀에서 최소제곱법으로 구한 결정계수는 두 변수의 **표본상관계수의 제곱**이다. 따라서 표본상관계수를 r이라 하면 $R^2 = r^2$로 구한다.

선지별 해설

결정계수는 종속변수의 전체 변동 중 회귀모형이 설명하는 비율이다. 잔차는 관측값과 예측값의 차이이며, 잔차제곱합을 SSE, 관측값이 표본평균에서 벗어난 정도의 제곱합인 총제곱합을 SST라 하면 $R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$이다. 여기서는 결정계수가 정의되도록 종속변수의 총제곱합이 양수인 경우를 전제한다.

① X: 결정계수가 0이면 잔차제곱합과 총제곱합이 같아 모형이 종속변수의 변동을 설명하지 못한다. **모든 잔차가 0이면** 잔차제곱합도 0이므로 결정계수는 0이 아니라 1이다.

② X: 결정계수가 클수록 전체 변동에서 잔차가 차지하는 비율이 작고, **모형이 설명하는 비율이 크다**. 따라서 같은 종속변수와 같은 표본을 대상으로 비교하면 결정계수가 큰 모형일수록 표본 내 설명력이 높다.

③ O: **절편이 있는 단순선형회귀에 최소제곱법을 적용하면** 결정계수는 독립변수와 종속변수의 표본상관계수 r의 제곱인 $R^2 = r^2$이다. 상관계수의 부호는 관계의 방향을 나타내지만, 제곱하여 구하는 결정계수에는 방향이 반영되지 않는다.

④ X: 같은 표본에서 기존 독립변수를 유지한 채 새 독립변수를 추가하면 결정계수는 **증가하거나 그대로이다**. 새 변수의 계수를 0으로 두면 기존 모형과 같은 적합이 가능하므로, 최소제곱법으로 얻는 잔차제곱합은 커질 수 없다. 총제곱합은 변하지 않으므로 결정계수도 감소하지 않는다.

NoteBase

문 10. 어느 도시의 고등학생 중 남학생 100명과 여학생 100명을 무작위 추출한 뒤 수학과 영어 과목에 대한 선호도를 조사하여 다음과 같은 분할표를 얻었다. “성별에 따라 과목 선호도에 차이가 없다.”라는 귀무가설을 검정하기 위한 카이제곱검정통계량의 값이 8일 때, (가)의 값과 귀무가설 하에서의 이 검정통계량의 근사적인 분포를 바르게 연결한 것은? (단, (가) ≥ 50이다)

성별 \ 과목	수학	영어	계
남학생	(가)		100
여학생			100
계	100	100	200

① (가): 55, 근사적인 분포: 자유도가 1인 카이제곱분포

② (가): 55, 근사적인 분포: 자유도가 2인 카이제곱분포

③ (가): 60, 근사적인 분포: 자유도가 1인 카이제곱분포

④ (가): 60, 근사적인 분포: 자유도가 2인 카이제곱분포

정답 3번

간단한 해설

네 칸의 기대도수가 모두 50명이므로, 검정통계량 8과 **(가) ≥ 50**을 적용하면 (가)는 60명이다. 성별과 과목이 각각 두 범주이므로 근사분포는 **자유도가 1인 카이제곱분포**이다.

NoteBase

계산 과정

관측도수: (가)를 x명이라 두면, 남학생의 영어 선호 인원은 $100-x$명이다. 수학 선호 인원의 합계도 100명이므로 여학생의 수학 선호 인원은 $100-x$명이고, 여학생의 영어 선호 인원은 x명이다. 따라서 네 칸의 관측도수는 차례로 $x,\ 100-x,\ 100-x,\ x$이다.

기대도수: 귀무가설은 남학생과 여학생의 과목 선호 비율이 같다는 뜻이다. 전체 200명 중 수학과 영어를 선호하는 인원이 각각 100명이므로, 공통 선호 비율은 각 과목에 대해 1/2로 추정한다. 각 성별 표본이 100명씩이므로 네 칸의 기대도수는 모두 $E=\frac{100\cdot 100}{200}=50$명이다. 이는 **귀무가설하에서 각 칸에 예상되는 인원**이다.

검정통계량: 관측도수 O와 기대도수 E의 차이를 비교하는 피어슨 카이제곱검정통계량은 $\chi^2=\sum\frac{(O-E)^2}{E}$이며, 합은 네 칸에 대해 구한다. 각 칸에서 관측도수와 기대도수의 차이는 $x-50$ 또는 $-(x-50)$이므로 제곱하면 모두 같다. 따라서 $\chi^2=\frac{4(x-50)^2}{50}=8$이다.

(가)의 값: 위 식을 정리하면 $(x-50)^2=100$이므로 $x=40$ 또는 $x=60$이다. 문제의 **(가) ≥ 50 조건**에 따라 x = 60을 택한다. 즉, 남학생 중 수학 선호 인원은 60명이며, 네 칸의 관측도수는 60명, 40명, 40명, 60명이 된다.

근사분포: 합계 행과 열을 제외하면 성별 2개 범주와 과목 2개 범주로 이루어진 분할표이다. 자유도는 **(행 범주 수 − 1) × (열 범주 수 − 1)**이므로 (2 − 1) × (2 − 1) = 1이다. 각 칸의 기대도수가 50명으로 충분히 크므로, 귀무가설하에서 이 검정통계량은 자유도가 1인 카이제곱분포를 근사적으로 따른다.

따라서 (가) = 60과 자유도가 1인 카이제곱분포를 연결한 ③에 해당한다.

NoteBase

문 11. 어느 도시에서 하루에 발생하는 교통사고 건수를 지난 200일 동안 조사하여 다음 표와 같이 정리하였다. 이 자료에서 구한 하루 발생 교통사고 건수의 통계량의 값으로 옳지 않은 것은?

교통사고 건수	0	1	2	3	4	계
해당일 수	90	50	30	20	10	200

① 평균은 1.05이다.

② 중앙값은 1이다.

③ 최빈값은 0이다.

④ 제3사분위수는 3이다.

정답 4번

간단한 해설

자료를 작은 값부터 나열하면 150번째와 151번째 관측값이 모두 2건이다. 따라서 제3사분위수는 **3이 아니라 2**이다.

선지별 해설

① O: 평균은 사고 총건수를 관측일 수로 나눈 값이다. 사고 총건수는 표의 각 사고 건수에 해당일 수를 곱해 더한 $0 + 50 + 60 + 60 + 40 = 210$건이다. 따라서 평균은 $\frac{210}{200} = 1.05$건이다. 사고 건수 0~4를 단순히 평균내는 것이 아니라 **각 값이 관측된 횟수를 반영**해야 한다.

② O: 관측값이 200개로 짝수이므로 중앙값은 작은 값부터 나열했을 때 **100번째와 101번째 값의 평균**이다. 0건인 날이 90일이고 1건인 날이 50일이므로, 91번째부터 140번째까지의 값은 모두 1건이다. 두 가운데 값이 모두 1건이므로 중앙값은 1건이다.

③ O: 최빈값은 **가장 자주 관측된 값**이다. 표에서 해당일 수가 가장 많은 것은 90일인 0건이므로 최빈값은 0건이다. 90은 최빈값 자체가 아니라 그 값의 도수이다.

④ X: 제3사분위수는 자료를 작은 값부터 나열했을 때 75% 지점에 해당하는 값이다. 200개 자료의 상위 절반인 101번째부터 200번째까지에서 중앙의 두 값은 전체의 150번째와 151번째 값이다. 누적일 수는 1건 이하가 140일, 2건 이하가 170일이므로 141번째부터 170번째까지의 값은 모두 2건이다. 따라서 제3사분위수는 **3건이 아니라 2건**이다.

NoteBase

문 12. 분산이 같은 두 정규모집단 A와 B로부터 크기가 각각 8인 확률표본을 독립적으로 추출하였더니 표본분산이 각각 19와 13이었다. 두 모평균의 차 $\mu_A - \mu_B$에 대한 95% 신뢰구간의 길이는? (단, $t_\alpha(k)$는 자유도가 k인 t분포의 제$100 \times (1-\alpha)$ 백분위수를 나타내고, $t_{0.025}(7) = 2.365,\ t_{0.025}(14) = 2.145$ 이다)

① 4.29

② 4.73

③ 8.58

④ 9.46

정답 3번

간단한 해설

두 모집단의 분산이 같으므로 합동표본분산을 사용하면 표본평균 차의 표준오차는 2이다. 자유도 14의 임계값 2.145를 적용한 오차한계는 4.29이며, **신뢰구간의 길이는 오차한계의 두 배**인 8.58이다.

계산 과정

두 표본은 서로 독립이고, 두 모집단은 정규분포를 따르며 **모분산이 같다**. 따라서 공통 모분산을 합동표본분산으로 추정하는 두 표본 t 신뢰구간을 사용한다. 표본크기를 $n_A = n_B = 8$, 표본분산을 $s_A^2 = 19$, $s_B^2 = 13$으로 놓으면 자유도는 $8 + 8 - 2 = 14$이다. 95% 양측 신뢰구간은 양쪽 꼬리에 각각 0.025를 남기므로 임계값은 $t_{0.025}(14) = 2.145$이다.

합동표본분산은 각 표본분산을 **각 표본의 자유도로 가중평균**한 값이다. 두 표본의 자유도가 각각 7이므로 $s_p^2 = \frac{7 \cdot 19 + 7 \cdot 13}{14} = 16$이다. 이는 두 모집단이 공유하는 모분산의 추정값이다.

두 표본이 독립이므로 표본평균 차의 분산은 두 표본평균의 분산을 더해 구한다. 공통 모분산에 합동표본분산 16을 대입하면 표본평균 차의 추정 표준오차는 $SE = \sqrt{s_p^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$이며, $SE = \sqrt{16\left(\frac{1}{8} + \frac{1}{8}\right)} = 2$이다.

신뢰구간은 표본평균 차를 중심으로 양쪽에 오차한계를 더하고 빼서 구한다. 오차한계는 임계값과 표준오차의 곱인 $2.145 \times 2 = 4.29$이다. 구간의 길이는 상한에서 하한을 뺀 값이므로 **오차한계의 두 배**인 $2 \times 4.29 = 8.58$이며, ③의 값과 일치한다. 표본평균 차는 구간의 중심만 바꾸므로 길이를 구할 때는 필요하지 않다.

NoteBase

문 13. 앞면이 나올 확률이 p인 동전에 대해 가설 $H_0 : p = \frac{1}{2},\ H_1 : p > \frac{1}{2}$을 검정하려고 한다. 이 동전을 다섯 번 던져 보니 앞면이 네 번 나왔다면, 유의확률(p-value)은?

① $\frac{1}{16}$
② $\frac{2}{16}$
③ $\frac{3}{16}$
④ $\frac{4}{16}$

정답 3번

간단한 해설

대립가설이 앞면 확률의 증가를 주장하므로, 유의확률은 귀무가설 아래에서 **앞면이 4번 이상 나올 확률**이다. 앞면이 4번인 경우와 5번인 경우를 더하면 $\frac{3}{16}$이다.

계산 과정

다섯 번 중 앞면이 나온 횟수를 X라 하자. 각 던지기가 서로 독립이고 앞면 확률이 일정하다는 동전 던지기 모형에서 X는 이항분포를 따른다. 귀무가설 $H_0 : p = \frac{1}{2}$ 아래에서는 $X \sim \mathrm{Bin}(5, \frac{1}{2})$이다.

유의확률은 귀무가설이 참일 때 관측 결과와 같거나 대립가설을 더 강하게 지지하는 결과가 나올 확률이다. 대립가설이 $H_1 : p > \frac{1}{2}$이므로 앞면 횟수가 클수록 이를 더 지지한다. 따라서 관측값 4를 포함한 **오른쪽 꼬리확률** $P(X \geq 4) = P(X = 4) + P(X = 5)$를 구한다.

앞면이 정확히 4번 나오는 배열은 뒷면이 나올 위치에 따라 5가지이다. 귀무가설 아래에서 각 배열의 확률은 $(\frac{1}{2})^5 = \frac{1}{32}$이므로 $P(X = 4) = 5 \cdot \frac{1}{32} = \frac{5}{32}$이다. 앞면이 5번 나오는 배열은 하나이므로 $P(X = 5) = \frac{1}{32}$이다.

따라서 유의확률은 $\frac{5}{32} + \frac{1}{32} = \frac{6}{32} = \frac{3}{16}$이며, 선택지 ③의 값과 일치한다.

NoteBase

문 14. 다음은 연령대별로 스마트폰 과의존 위험군 비율을 조사한 것이다.
이 자료를 바탕으로 다음의 가설을 검정하고자 한다.
○ H_0: 연령대에 따라 과의존 위험군 비율에 차이가 없다.
○ H_1: 연령대에 따라 과의존 위험군 비율에 차이가 있다.
귀무가설 H_0가 참일 때 구한, 위험군에 속하는 사람 수의 추정된 기대도수와 관측도수 간에 차이가 존재한다.
이 차이의 절댓값이 가장 큰 연령대는?

연령대	10대 이하	20 ~ 30대	40 ~ 50대	60대 이상
위험군 비율	28 %	37 %	22 %	16 %
표본 크기	50명	100명	100명	50명

① 10대 이하
② 20□30대
③ 40□50대
④ 60대 이상

정답 2번

간단한 해설

귀무가설 아래에서 추정한 공통 위험군 비율은 27%이다. 20□30대는 관측도수 37명과 기대도수 27명의 **차이의 절댓값이 10명**으로 가장 크다.

NoteBase

선지별 해설

귀무가설이 참이면 **모든 연령대의 위험군 비율이 같으므로**, 전체 표본을 합친 위험군 비율을 공통 비율로 추정한다. 표의 위험군 비율에 각 연령대의 표본 크기를 곱하면 관측도수는 차례로 14명, 37명, 22명, 8명이다. 전체 표본은 50 + 100 + 100 + 50 = 300명이고, 위험군은 14 + 37 + 22 + 8 = 81명이므로 공통 비율의 추정값은 $\hat{p} = \frac{81}{300} = 0.27$이다. 각 연령대의 기대도수는 해당 연령대의 표본 크기에 0.27을 곱하여 구한다.

① 10대 이하: 관측도수는 50 × 0.28 = 14명이고, 기대도수는 50 × 0.27 = 13.5명이다. 차이의 절댓값은 |14 − 13.5| = 0.5명이다.

② 20☒30대: 관측도수는 100 × 0.37 = 37명이고, 기대도수는 100 × 0.27 = 27명이다. 차이의 절댓값은 |37 − 27| = 10명이다.

③ 40☒50대: 관측도수는 100 × 0.22 = 22명이고, 기대도수는 100 × 0.27 = 27명이다. 차이의 절댓값은 |22 − 27| = 5명이다.

④ 60대 이상: 관측도수는 50 × 0.16 = 8명이고, 기대도수는 50 × 0.27 = 13.5명이다. 차이의 절댓값은 |8 − 13.5| = 5.5명이다.

비교하는 값은 위험군 비율 자체가 아니라 **관측도수와 기대도수 차이의 절댓값**이다. 네 값 중 10명이 가장 크므로 ②의 20☒30대가 해당한다.

NoteBase

문 15. 어느 마트에서 특정 상품에 대한 진열대의 높이와 폭이 매출액에 영향을 주는지 알아보기 위해, 진열대 높이(A)를 상, 중, 하 3수준으로, 폭(B)을 대, 소 2수준으로 하여 실험하였다. 다른 조건들이 동일한 12일 동안 6개의 실험조건을 무작위 순서로 2일씩 적용하여 매출액을 조사한 후 얻은 분산분석표가 다음과 같다. 이에 대한 설명으로 옳지 않은 것은?

요인	제곱합	자유도	평균제곱	F-값
인자 A	1,544			
인자 B	12			
교호작용 $A\times B$	24			
오차	62			
계				

① 총제곱합은 1,642이다.

② 인자 B의 평균제곱은 교호작용 A x B의 평균제곱과 동일하다.

③ 오차 자유도는 6이다.

④ 유의성검정에서의 유의확률은 인자 A가 교호작용 A x B보다 크다.

정답 4번

간단한 해설

인자 A와 교호작용 A×B의 검정은 자유도가 같은 F분포를 사용한다. 인자 A의 F값이 더 크므로 **유의확률은 인자 A가 더 작다**.

NoteBase

선지별 해설

① O: 반복이 있는 이원배치 분산분석에서 총제곱합은 인자 A, 인자 B, 교호작용 A×B 및 오차의 제곱합을 더한 값이다. 표의 제곱합을 합하면 1,544 + 12 + 24 + 62 = 1,642이다.

② O: 평균제곱은 **제곱합을 해당 자유도로 나눈 값**이다. 인자 B는 2수준이므로 자유도가 2 − 1 = 1이고, 교호작용 A×B의 자유도는 (3 − 1) × (2 − 1) = 2이다. 따라서 인자 B의 평균제곱은 $MS_B = \frac{12}{1} = 12$, 교호작용의 평균제곱은 $MS_{A\times B} = \frac{24}{2} = 12$로 같다.

③ O: 3 × 2 = 6개의 실험조건마다 2회씩 관측하였다. 각 조건 안에서 오차를 추정하는 자유도는 2 − 1 = 1이므로, 전체 오차 자유도는 3 × 2 × (2 − 1) = 6이다. 즉, **각 실험조건 안의 반복 관측 간 변동**으로 오차를 추정한다.

④ X: 각 효과의 F값은 해당 평균제곱을 오차평균제곱으로 나눈 값이다. 인자 A의 자유도는 3 − 1 = 2이므로 $MS_A = \frac{1544}{2} = 772$이고, 오차평균제곱은 $MS_E = \frac{62}{6} = \frac{31}{3}$이다. 따라서 $F_A = \frac{772}{MS_E} \approx 74.71$, $F_{A\times B} = \frac{12}{MS_E} \approx 1.16$이다. 효과가 없다는 귀무가설 아래 두 검정 모두 분자 자유도 2, 분모 자유도 6인 F분포를 사용한다. 유의확률은 관측된 F값 이상의 값이 나올 확률이므로, **같은 F분포에서는 F값이 클수록 유의확률이 작다**. 따라서 인자 A의 유의확률은 교호작용 A×B의 유의확률보다 작다.

NoteBase

문 16. 다음은 자료 (x_{1i}, x_{2i}, y_i) $(i = 1, 2, \cdots, 23)$에 다중선형회귀모형 $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$를 최소제곱법으로 적합하여 얻은 분산분석표이다. 이에 대한 설명으로 옳은 것만을 모두 고르면? (단, ε_i는 $N(0, \sigma^2)$을 따르고 서로 독립이다)

요인	제곱합	자유도	평균제곱	$F-$값	$p-$값
회귀	420		(가)	(다)	< 0.001
잔차			(나)		
계	480				

ㄱ. 결정계수는 0.875이다.

ㄴ. 유의수준 1%에서 이 모형이 유의하다.

ㄷ. (가) + (나) + (다) = 283이다.

① ㄱ, ㄴ

② ㄱ, ㄷ

③ ㄴ, ㄷ

④ ㄱ, ㄴ, ㄷ

정답 4번

간단한 해설

회귀·잔차 평균제곱은 **각 제곱합을 해당 자유도로 나눈 값**이므로 각각 210과 3이다. F값은 $\frac{210}{3} = 70$이어서 세 빈칸의 합은 283이다.

NoteBase

보기별 해설

관측값은 23개이고 설명변수는 2개이다. 회귀 자유도는 설명변수의 수인 2이고, 잔차 자유도는 절편을 포함한 모수 3개를 추정하므로 $23 - 3 = 20$이다.

ㄱ. O: 결정계수는 **전체 변동 중 회귀모형이 설명하는 변동의 비율**이다. 절편이 포함된 최소제곱 회귀모형이므로 표의 회귀제곱합을 전체제곱합으로 나누어 $R^2 = \frac{420}{480} = 0.875$를 얻는다. 즉, 전체 변동의 87.5%를 이 모형이 설명한다.

ㄴ. O: 모형 전체의 F검정은 두 설명변수의 회귀계수가 모두 0이라는 귀무가설 $H_0 : \beta_1 = \beta_2 = 0$을 검정한다. 표의 회귀 행에 제시된 p값은 0.001보다 작아 유의수준 1%인 0.01보다 작다. 따라서 귀무가설을 기각하며, **적어도 하나의 회귀계수가 0이 아니므로** 모형 전체가 유의하다.

ㄷ. O: 전체제곱합은 회귀제곱합과 잔차제곱합의 합이므로 잔차제곱합은 $480 - 420 = 60$이다. 평균제곱은 제곱합을 해당 자유도로 나눈 값이므로 회귀평균제곱 (가)는 $\frac{420}{2} = 210$, 잔차평균제곱 (나)는 $\frac{60}{20} = 3$이다. F값은 **회귀평균제곱을 잔차평균제곱으로 나눈 값**이므로 (다)는 $\frac{210}{3} = 70$이다. 따라서 (가) + (나) + (다) = 210 + 3 + 70 = 283이다.

ㄱ, ㄴ, ㄷ이 모두 성립하므로 세 보기를 모두 포함한 ④와 일치한다.

NoteBase

문 17. 다음 설명 중 옳지 않은 것은?

① 유의수준 $\alpha = 0.01$에서 귀무가설을 기각하는 검정 결과가 나왔다면 유의수준 $\alpha = 0.05$에서도 귀무가설을 기각하게 된다.

② 모든 자연수 n에 대해 $t_{0.05}(n) < t_{0.05}(n+1)$이다(단, $t_{0.05}(k)$는 자유도가 k인 t분포의 제95백분위수).

③ 모든 자연수 n에 대해 $\chi^2_{0.05}(n) < \chi^2_{0.05}(n+1)$이다(단, $\chi^2_{0.05}(k)$는 자유도가 k인 카이제곱분포의 제95백분위수).

④ 두 사건 A_1, A_2가 서로 배반이고 $P(B) > 0$일 때, $P(A_1 \cup A_2 \mid B) = P(A_1 \mid B) + P(A_2 \mid B)$이다.

정답 2번

간단한 해설

②는 부등호 방향이 반대이다. t분포는 자유도가 커질수록 꼬리가 얇아지므로 **제95백분위수는 감소**하여 $t_{0.05}(n) > t_{0.05}(n+1)$이 성립한다.

선지별 해설

① O: 같은 자료에 같은 검정방법을 적용할 때, 유의수준이 커지면 귀무가설을 **기각하는 범위가 넓어진다**. 유의확률을 p라 하면 유의수준 0.01에서 기각되었다는 것은 $p \le 0.01$이라는 뜻이다. 따라서 $p \le 0.05$도 성립하므로 유의수준 0.05에서도 기각한다.

② X: 제95백분위수는 그 값 이하에 전체 확률의 95%가 놓이는 지점이다. t분포는 자유도가 작을수록 꼬리가 두껍고, 자유도가 커질수록 표준정규분포에 가까워진다. 따라서 양의 영역에 있는 제95백분위수는 자유도가 커질수록 **작아진다**. 모든 자연수 n에 대해 $t_{0.05}(n) > t_{0.05}(n+1)$으로 고쳐야 한다.

③ O: 자유도가 n인 카이제곱분포는 서로 독립인 표준정규확률변수 n개의 제곱합이 따르는 분포이다. 여기에 독립인 표준정규확률변수 하나의 제곱을 더하면 자유도가 $n+1$인 카이제곱분포가 된다. **음이 아닌 제곱항이 추가**되므로 분포가 큰 값 쪽으로 이동하고, 제95백분위수도 커진다. 따라서 $\chi^2_{0.05}(n) < \chi^2_{0.05}(n+1)$이 성립한다.

④ O: 두 사건이 배반이면 동시에 일어날 수 없으므로 $A_1 \cap A_2 = \varnothing$이다. 이에 따라 $A_1 \cap B$와 $A_2 \cap B$도 서로 배반이어서 두 사건의 합집합 확률은 각각의 확률을 더한 값이다. 조건부확률은 교집합 확률을 **같은 양수 $P(B)$로 나눈 값**이므로, 양변을 $P(B)$로 나누면 선지의 조건부확률 덧셈식이 성립한다.

NoteBase

문 18. 다음은 자료 (x_i, y_i) $(i = 1, 2, \cdots, n)$에 단순선형회귀모형 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$를 적합하여 구할 수 있는 값들이다. 다음 식 중 옳은 것은? (단, $\hat{\beta}_0$, $\hat{\beta}_1$은 각각 β_0, β_1의 최소제곱 추정값이다)

○ $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
○ $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
○ $e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$
○ $SSE = \sum_{i=1}^{n} e_i^2$
○ $S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$
○ $S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$
○ $S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$

① $\hat{\beta}_1 = \frac{S_{xy}}{S_{yy}}$
② $SSE = S_{yy} - \hat{\beta}_1 S_{xy}$
③ $r^2 = \frac{S_{xy}}{S_{xx} S_{yy}}$ (단, r^2은 결정계수)
④ $\sum_{i=1}^{n} y_i e_i = 0$

정답 2번

간단한 해설

절편을 포함한 최소제곱회귀에서는 **전체 변동이 회귀로 설명되는 변동과 잔차 변동으로 분해**된다. 전체 제곱합이 S_{yy}, 회귀로 설명되는 제곱합이 $\hat{\beta}_1 S_{xy}$이므로, 잔차제곱합은 $SSE = S_{yy} - \hat{\beta}_1 S_{xy}$이다.

NoteBase

선지별 해설

최소제곱법은 잔차제곱합을 가장 작게 하는 절편과 기울기를 구한다. 절편을 포함한 회귀모형의 정규방정식은 $\sum_{i=1}^{n} e_i = 0$, $\sum_{i=1}^{n} x_i e_i = 0$이다. 기울기를 유일하게 구하려면 $S_{xx} > 0$이어야 하며, 결정계수를 정의하려면 $S_{yy} > 0$이어야 한다.

① X: 기울기 추정값은 $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$이다. 정규방정식에 $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$를 대입하면 $S_{xy} = \hat{\beta}_1 S_{xx}$를 얻는다. 따라서 분모는 반응변수의 변동인 S_{yy}가 아니라 **설명변수의 변동**인 S_{xx}이다.

② O: 잔차를 $e_i = (y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})$로 쓰고 제곱합을 전개하면 $SSE = S_{yy} - 2\hat{\beta}_1 S_{xy} + \hat{\beta}_1^2 S_{xx}$이다. 여기에 $\hat{\beta}_1 S_{xx} = S_{xy}$를 적용하면 $SSE = S_{yy} - \hat{\beta}_1 S_{xy}$가 된다. 이는 **전체 제곱합에서 회귀로 설명되는 제곱합을 뺀 값**이 잔차제곱합이라는 뜻이다.

③ X: 결정계수는 전체 변동 중 회귀로 설명되는 변동의 비율이므로 $r^2 = 1 - \frac{SSE}{S_{yy}}$이다. ②의 관계와 기울기 공식을 대입하면 $r^2 = \frac{\hat{\beta}_1 S_{xy}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx} S_{yy}}$이다. 선지의 분자는 S_{xy}가 아니라 **그 제곱**인 S_{xy}^2이어야 한다.

④ X: 적합값을 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$라 하면, 정규방정식에 의해 $\sum_i \hat{y}_i e_i = 0$이다. 관측값은 $y_i = \hat{y}_i + e_i$이므로 $\sum_i y_i e_i = \sum_i e_i^2 = SSE$가 된다. 잔차와 곱한 합이 항상 0인 것은 **관측값이 아니라 적합값**이다. 관측값과 잔차의 곱의 합은 모든 잔차가 0인 완전 적합일 때에만 0이다.

NoteBase

문 19. 두 이산확률변수 X, Y의 결합확률질량함수가 다음과 같을 때, $P(Y=2X)$의 값은? (단, c는 상수이다)

$f(x,y)=\frac{x+y}{c}, \quad x=1,2,3, \quad y=1,2,3,4$

① $\frac{1}{18}$
② $\frac{1}{9}$
③ $\frac{1}{6}$
④ $\frac{2}{9}$

정답 3번

간단한 해설

$Y=2X$를 만족하는 가능한 값의 쌍은 **(1, 2)와 (2, 4)뿐**이다. 전체 확률의 합이 1이 되도록 구한 $c=54$를 이용해 두 쌍의 확률을 더하면 $P(Y=2X)=\frac{1}{6}$이다.

계산 과정

1. 상수 c 구하기: 결합확률질량함수는 각 값의 쌍이 나타날 확률을 뜻하므로, **가능한 모든 쌍의 확률을 더하면 1**이다. 따라서 $\sum_{x=1}^{3}\sum_{y=1}^{4}\frac{x+y}{c}=1$이다. 분자의 합에서 각 x는 y의 값 4개에 대해 네 번씩, 각 y는 x의 값 3개에 대해 세 번씩 더해진다. x의 값의 합은 1 + 2 + 3 = 6이고, y의 값의 합은 1 + 2 + 3 + 4 = 10이므로 $c=4\cdot 6+3\cdot 10=54$이다. 이 54는 전체 12개 쌍의 분자 값을 합한 값이다.

2. 사건에 포함되는 쌍 찾기: 주어진 범위에서 $y=2x$인 쌍은 (1, 2)와 (2, 4)이다. x가 3이면 y가 6이어야 하지만, **y의 범위가 1부터 4까지**이므로 (3, 6)은 포함되지 않는다.

3. 해당 확률 더하기: 두 쌍은 동시에 일어날 수 없으므로 결합확률을 더한다. $P(Y=2X)=f(1,2)+f(2,4)$이고, $f(1,2)=\frac{1+2}{54}=\frac{3}{54}$, $f(2,4)=\frac{2+4}{54}=\frac{6}{54}$이다. 따라서 $P(Y=2X)=\frac{3}{54}+\frac{6}{54}=\frac{1}{6}$이며, 이는 ③의 값과 일치한다.

NoteBase

문 20. 수준이 3개이고 반복수가 각각 $n_1 = 5,\ n_2 = 10,\ n_3 = 15$인 일원배치법으로부터 얻은 자료를 $y_{ij}\ (i = 1, 2, 3,\ j = 1, 2, \cdots, n_i)$라 하고, i번째 수준에서의 평균을 $\bar{y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$라 하자. $\bar{y}_{1\cdot} = 6,\ \bar{y}_{2\cdot} = 3,\ \bar{y}_{3\cdot} = 2$일 때, 분산분석표에서 처리제곱합은?

① 60

② 90

③ 270

④ 330

정답 1번

간단한 해설

처리제곱합은 각 수준의 평균과 전체 평균의 차이를 제곱한 뒤 **각 수준의 반복수로 가중하여 합한 값**이다. 반복수를 반영한 전체 평균은 3이며, 이에 따른 처리제곱합은 60이다.

계산 과정

처리제곱합은 수준별 평균이 전체 평균에서 얼마나 떨어져 있는지를 나타낸다. 수준별 반복수를 n_i, 수준별 평균을 $\bar{y}_{i\cdot}$, 전체 평균을 $\bar{y}_{\cdot\cdot}$라 하면, $SS_A = \sum_{i=1}^{3} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$이다. 각 수준의 평균이 전체 평균에서 벗어난 정도를 그 수준의 관측값 수만큼 반영하므로, 평균 차이의 제곱에 반복수를 곱한다.

전체 평균: 총 관측값 수는 $5 + 10 + 15 = 30$이다. **반복수가 서로 다르므로** 수준별 평균을 단순히 더해 3으로 나누지 않고, 각 수준의 자료 합을 모두 더한 뒤 총 관측값 수로 나눈다. 따라서 $\bar{y}_{\cdot\cdot} = \frac{5\cdot 6 + 10\cdot 3 + 15\cdot 2}{30} = 3$이다.

처리제곱합: 첫 번째 수준의 기여분은 $5(6-3)^2 = 45$이다. 두 번째 수준은 평균이 전체 평균과 같으므로 $10(3-3)^2 = 0$이고, 세 번째 수준은 $15(2-3)^2 = 15$이다. 이를 합하면 처리제곱합은 $45 + 0 + 15 = 60$으로, ①의 값과 일치한다.

NoteBase