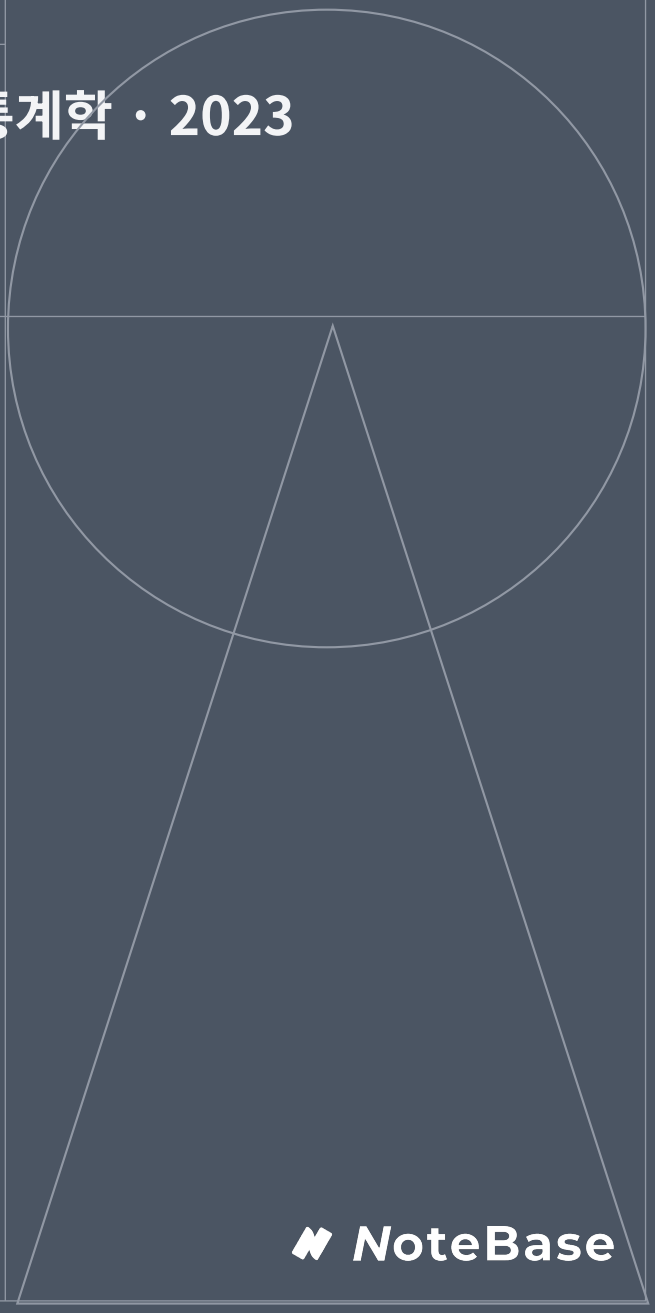
기출OX 해설집

공무원/경찰/공채(경위) · 통계학 · 2023



NoteBase

기출OX 해설집

공무원/경찰/공채(경위) · 통계학 · 2023

공무원/경찰/공채(경위) · 통계학 · 2023

문 1. 표준편차가 4인 정규분포로부터 나온 확률표본 16개의 표본평균이 10일 때 모평균 μ에 대한 95% 신뢰구간은?

① $10 \pm 1.65 \times 4$

② $10 \pm 1.96 \times 4$

③ 10 ± 1.65

④ 10 ± 1.96

정답 4번

간단한 해설

모표준편차를 알고 있는 정규모집단이므로 95% 양측 임계값 1.96을 사용한다. **표본평균의 표준오차**는 $\frac{4}{\sqrt{16}} = 1$ 이므로 신뢰구간은 10 ± 1.96이다.

계산 과정

적용 조건: 모집단이 정규분포이고 모표준편차 $\sigma = 4$가 주어져 있으므로, 모평균의 신뢰구간에는 **표준정규분포의 임계값**을 사용한다. 표본 크기는 $n = 16$, 표본평균은 $\overline{x} = 10$이다. 모표준편차를 알고 있으므로 표본이 작다는 이유만으로 t분포를 사용하지 않는다.

표준오차: 신뢰구간의 오차한계에는 개별 관측값의 표준편차 4가 아니라 **표본평균의 표준오차**가 들어간다.
$SE(\overline{X}) = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{16}} = 1$

95% 양측 임계값: 구간 밖의 확률 5%를 양쪽 꼬리에 2.5%씩 나누면 표준정규분포의 임계값은 약 1.96이다.
따라서 오차한계는 $1.96 \times 1 = 1.96$이다.

신뢰구간: 표본평균을 중심으로 오차한계를 더하고 뺀다.
$\overline{x} \pm 1.96\frac{\sigma}{\sqrt{n}} = 10 \pm 1.96 \times 1 = 10 \pm 1.96$
즉 하한은 8.04, 상한은 11.96이다.

따라서 모평균의 95% 신뢰구간은 10 ± 1.96으로, ④에 해당한다.

NoteBase

문 2. 확률변수 Y의 표준편차가 2일 때 확률변수 $Z = 3Y$의 분산은?

① 36

② 18

③ 6

④ 12

정답 1번

간단한 해설

확률변수에 상수를 곱하면 분산은 **그 상수의 제곱배**가 된다. Y의 분산은 $2^2 = 4$이므로 Z의 분산은 $3^2 \times 4 = 36$이다.

계산 과정

표준편차와 분산: 표준편차는 분산의 양의 제곱근이다. Y의 표준편차가 2이므로 $\mathrm{Var}(Y) = 2^2 = 4$이다.

상수배의 분산: $Z = 3Y$이면 평균도 3배가 되어 평균으로부터의 편차가 3배가 된다. 분산은 편차의 제곱을 평균한 값이므로 **분산은 9배**가 된다. 따라서 $\mathrm{Var}(Z) = 3^2\mathrm{Var}(Y) = 9 \times 4 = 36$이다.

결과: 구하는 분산은 36이므로 ①에 해당한다. Z의 표준편차는 $3 \times 2 = 6$이지만, 발문은 표준편차가 아니라 분산을 묻는다.

NoteBase

문 3. 성공확률이 p인 베르누이 시행을 독립적으로 반복 시행하여 k번 성공할 때까지의 총 시행횟수를 확률변수 X라 하자. 이때 확률변수 X는 어떤 분포를 따르는가?

① 이항분포

② 초기하분포

③ 기하분포

④ 음이항분포

정답 4번

간단한 해설

독립 베르누이 시행에서 **k번째 성공이 나올 때까지의 총 시행횟수**는 음이항분포를 따른다. 성공 횟수는 고정하고, 이를 달성하는 데 필요한 시행횟수가 변한다.

개념 구분

음이항분포는 성공확률이 일정한 독립 베르누이 시행에서 **정해진 횟수의 성공이 나올 때까지 기다리는 시행횟수**의 분포이다. 여기서는 성공 횟수 k가 정해져 있고, 마지막인 X번째 시행에서 k번째 성공이 나타난다. 따라서 총 시행횟수 X는 k 이상인 정숫값을 갖는다.

이항분포는 **총 시행횟수를 미리 고정**한 뒤 그 안에서 나온 성공 횟수를 센다. 이 문항처럼 성공 횟수에 도달할 때까지 시행을 계속하는 실험과 구별된다.

초기하분포는 성공·실패로 나뉜 유한 모집단에서 정해진 개수를 비복원추출할 때의 성공 개수를 센다. 추출에 따라 남은 모집단의 구성이 달라지므로, 성공확률이 일정한 독립 반복 시행이라는 이 문항의 조건과 다르다.

기하분포는 **첫 성공이 나올 때까지**의 시행횟수를 나타낸다. 음이항분포에서 k = 1인 특수한 경우이며, 일반적인 k번 성공을 다루는 분포 이름은 음이항분포이다. 음이항분포를 성공 전 실패 횟수로 정의하는 관례도 있지만, 이 문항은 실패 횟수에 k를 더한 총 시행횟수를 센다.

따라서 X의 분포는 ④ 음이항분포이다.

NoteBase

문 4. 다음 중 확률표본추출방법이 아닌 것은?

① 가용표본추출(Available Sampling)

② 단순무작위추출(Simple Random Sampling)

③ 집락(군집)추출(Cluster Sampling)

④ 층화추출(Stratified Sampling)

정답 1번

간단한 해설

가용표본추출은 조사하기 쉬운 대상을 편의에 따라 고르는 **비확률표본추출**이다. 무작위 절차에 따라 각 단위의 추출확률을 알 수 있도록 설계하는 확률표본추출과 구별된다.

선지별 해설

① X: 가용표본추출은 접근하거나 응답을 얻기 쉬운 사람 등 이용 가능한 대상을 선택하는 편의표본추출이다. **무작위 추출에 의해 알려진 추출확률을 부여하지 않으므로** 확률표본추출방법이 아니다. 예를 들어 조사자가 만난 사람 중 응답에 동의한 사람만 편의대로 고르면 모집단의 각 구성원이 뽑힐 확률을 설계상 알 수 없다.

② O: 단순무작위추출은 모집단에서 정해진 크기의 표본을 뽑을 때 **가능한 모든 표본이 동일한 확률**로 선택되게 하는 방법이다. 추첨이나 난수를 이용하므로 구성원별 추출확률을 알 수 있는 확률표본추출에 해당한다.

③ O: 집락추출은 모집단을 지역·학교 등 여러 집락으로 나누고 **집락을 무작위로 선택**하는 방법이다. 선택된 집락의 구성원을 모두 조사하거나, 그 안에서 다시 확률적으로 일부를 추출한다. 무작위 선택 절차와 추출확률이 정해져 있으므로 확률표본추출방법이다.

④ O: 층화추출은 모집단을 관련 특성에 따라 서로 겹치지 않는 층으로 나눈 뒤 **각 층에서 무작위로 표본을 추출**하는 방법이다. 모든 층에서 표본을 확보한다는 점에서 일부 집락을 선택하는 집락추출과 구별된다. 층별 추출비율이 같지 않더라도 정해진 확률적 절차로 뽑으면 확률표본추출에 해당한다.

NoteBase

문 5. 프로야구 경기에서 연장전을 할 확률이 0.2라고 할 때 10번의 경기에서 연장전을 치르는 경기의 횟수 X는 이항분포 $Bin(10, 0.2)$를 따른다. 이때 적어도 한 경기 이상이 연장전으로 이어질 확률은?

① $P(X=1) = \binom{10}{1}(0.2)^1(0.8)^9$

② $1 - P(X=1) = 1 - \binom{10}{1}(0.2)^1(0.8)^9$

③ $1 - P(X=0) = 1 - \binom{10}{0}(0.2)^0(0.8)^{10}$

④ $P(X=0) = \binom{10}{0}(0.2)^0(0.8)^{10}$

정답 3번

간단한 해설

'적어도 한 경기 이상'의 여사건은 **연장전이 한 번도 없는 경우**이다. 따라서 구하는 확률은
$1 - P(X=0) = 1 - (0.8)^{10}$이다.

핵심 풀이

X는 10경기 중 연장전을 치르는 경기 수이다. '적어도 한 경기 이상'은 정확히 한 경기가 아니라 $X \geq 1$이라는 뜻이다. 그 여사건은 **연장전이 한 번도 없는 경우**, 즉 X = 0이다.

주어진 이항분포에서 한 경기가 연장전으로 이어지지 않을 확률은 0.8이다. 10경기 모두 연장전이 없을 확률은 $P(X=0) = \binom{10}{0}(0.2)^0(0.8)^{10} = (0.8)^{10}$이다. 따라서 1회부터 10회까지의 확률을 따로 더할 필요 없이 여사건을 이용한다.

구하는 확률은 $P(X \geq 1) = 1 - P(X=0) = 1 - (0.8)^{10}$이며, 약 0.8926이다.

①은 정확히 한 경기만 연장전인 확률이고,

②는 그 여사건으로 연장전이 없거나 두 경기 이상인 확률이다.

④는 연장전이 한 번도 없는 확률이다.

따라서 **연장전이 없는 확률을 1에서 뺀** 식을 제시한 ③과 일치한다.

NoteBase

문 6. 상자그림(Box Plot)으로부터 알 수 없는 통계량은?

① 최솟값(Minimum), 최댓값(Maximum)

② 제1사분위수, 제3사분위수

③ 중앙값(Median)

④ 평균(Mean)

정답 4번

간단한 해설

기본 상자그림은 중앙값·사분위수·극단값으로 자료를 요약하므로 **평균을 직접 표시하지 않는다**. 이 요약값만으로는 관측값 전체의 합을 알 수 없어 평균도 일반적으로 결정할 수 없다.

선지별 해설

상자그림은 정렬된 자료의 위치와 산포를 보여 주는 그림이다. O/X는 평균을 별도로 추가 표시하지 않은 기본 상자그림에서 해당 통계량을 알 수 있는지를 뜻한다.

① O: 기본 상자그림의 양쪽 수염 끝은 자료의 최솟값과 최댓값을 나타낸다. 이상값을 따로 점으로 표시하는 방식에서는 수염 끝이 전체 극단값과 다를 수 있으므로 **이상값 점까지 포함해** 가장 작은 값과 가장 큰 값을 확인한다.

② O: 상자의 두 경계는 제1사분위수와 제3사분위수이다. 각각 아래에서 약 25%, 75%에 해당하는 위치를 나타내며, 두 값의 차이가 사분위범위이다.

③ O: 상자 안의 선은 **중앙값**이다. 자료를 크기순으로 정렬했을 때 가운데에 놓이는 값이며, 평균을 나타내는 선이 아니다.

④ X: 평균은 관측값 전체의 합을 관측값 수로 나눈 값이다. 기본 상자그림에는 개별 관측값 전체나 그 합이 제시되지 않으며, 같은 사분위수·중앙값·극단값을 갖는 자료라도 평균은 다를 수 있다. 따라서 기본 상자그림만으로 평균을 일반적으로 알 수 없다.

NoteBase

문 7. 서로 독립인 두 확률변수 X, Y가 각각 $N(0, 1), N(1, 3)$을 따를 때, 확률 $P(X + Y \geq 4.92)$는 얼마인가? (단, $P(|Z| \geq 1.96) = 0.05, P(|Z| \geq 1.645) = 0.1$)

① 0.025

② 0.05

③ 0.075

④ 0.1

정답 1번

간단한 해설

독립인 두 정규변수의 합은 평균 1, 분산 4인 정규분포를 따르므로 4.92를 표준화하면 1.96이다. 주어진 0.05는 **양쪽 꼬리의 합**이므로 오른쪽 꼬리 확률은 $\frac{0.05}{2} = 0.025$이다.

계산 과정

합의 분포: 정규분포 표기 $N(\mu, \sigma^2)$에서 두 번째 값은 분산이다. 따라서 X의 평균과 분산은 각각 0과 1, Y의 평균과 분산은 각각 1과 3이다. S = X + Y로 놓으면 평균은 0 + 1 = 1이다. 두 변수가 **독립이므로 공분산이 0**이어서 합의 분산은 1 + 3 = 4가 된다. 독립인 정규변수의 합도 정규분포이므로 $S \sim N(1, 4)$이며 표준편차는 2이다.

표준화: 임계값에서 평균을 빼고 **분산이 아닌 표준편차**로 나눈다.
$Z = \frac{S-1}{2} \sim N(0, 1)$

aligned&P(X+Y≥4.92)&=P≤ft(Z≥(4.92-1)/(2))=P(Z≥1.96)aligned

한쪽 꼬리 확률: 주어진 P(|Z|≥1.96)=0.05는 왼쪽과 오른쪽 꼬리를 합한 확률이다. 표준정규분포는 0을 중심으로 대칭이므로 두 꼬리의 확률은 같다.
P(Z≥1.96)=(P(|Z|≥1.96))/(2)=(0.05)/(2)=0.025

따라서 구하는 확률은 0.025이며 ①에 해당한다.

NoteBase

문 8. 확률변수 X가 평균 $\mu = 60$과 분산 $\sigma^2 = 16$을 가질 때 체비셰프 부등식에 의한 확률 $P(54 < X < 66)$의 하한은?

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

정답 3번

간단한 해설

체비셰프 부등식에서 평균으로부터 6 이상 떨어질 확률은 (16)/(6^2)=49 이하이다. **그 여사건인 구간 안의 확률**은 1-49=59 이상이다.

계산 과정

주어진 평균은 60, 분산은 16이다. 두 경계 54와 66은 평균에서 각각 6만큼 떨어져 있으므로, 구할 확률은 P(54 X66)=P(|X-60|6)이다. 여기서 6은 **평균에서 경계까지의 거리**이며 표준편차가 아니다. 표준편차는 $\sigma = \sqrt{16} = 4$이다.

분산이 유한한 확률변수에 대해 체비셰프 부등식은 $P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$이다. 정규분포 같은 특정 분포를 가정하지 않아도 적용할 수 있다. 평균에서 6 이상 떨어질 확률의 상한은 P(|X-60|≥6)≤(16)/(6^2)=49이다.

구하려는 열린 구간의 여사건은 **양쪽 경계를 포함한 바깥 영역**이다. 따라서
aligned&P(|X-60|6)=1-P(|X-60|≥6)≥1&-49=59aligned
이다.

체비셰프 부등식에 의한 하한은 59로, ③에 해당한다. 이는 보장되는 최소 수준이며, 실제 구간확률이 반드시 이 값과 같다는 뜻은 아니다.

NoteBase

문 9. 단순회귀분석에서 독립변수 X와 종속변수 Y의 관계를 파악하기 위해 수집한 자료 10개에 대해 다음과 같은 요약값을 얻었다. 최소제곱법에 의해 추정된 회귀선은?

$$\sum_{i=1}^{10} X_i = 40, \quad \sum_{i=1}^{10} X_i^2 = 175, \quad \sum_{i=1}^{10} Y_i = 50, \quad \sum_{i=1}^{10} Y_i^2 = 215, \quad \sum_{i=1}^{10} X_i Y_i = 179$$

① $\hat{Y} = 10.6 - 1.4X$

② $\hat{Y} = 10.6 - 2.33X$

③ $\hat{Y} = 8.6 - 1.4X$

④ $\hat{Y} = 8.6 - 2.33X$

정답 1번

간단한 해설

최소제곱 회귀선의 **기울기**는 X와 Y의 중심화 곱합을 X의 중심화 제곱합으로 나눈 값으로, -1.4이다. 두 변수의 평균을 지나는 조건에서 절편은 10.6이므로 회귀선은 "0176=10.6-1.4X이다.

계산 과정

1. 표본평균과 회귀식: 추정할 회귀선을 "0176=a+bX라 두면, a는 절편이고 b는 기울기이다. 자료가 10개이므로 두 변수의 평균은 $\overline{X} = \frac{40}{10} = 4$, $\overline{Y} = \frac{50}{10} = 5$이다.

2. 기울기 계산: 최소제곱법은 관측값과 예측값의 차이인 **잔차의 제곱합을 최소화**한다. 절편을 포함한 단순회귀에서 기울기는 $b = \frac{S_{XY}}{S_{XX}}$로 구한다. 여기서 S_{XX}는 X가 평균에서 벗어난 정도의 제곱합으로, $S_{XX} = 175 - \frac{40^2}{10} = 15$이다. S_{XY}는 X와 Y가 각 평균에서 벗어난 값들의 곱을 합한 것으로, $S_{XY} = 179 - \frac{40 \cdot 50}{10} = -21$이다. 따라서 $b = \frac{-21}{15} = -1.4$이다.

3. 절편과 회귀선: 절편을 포함한 최소제곱 회귀선은 **두 변수의 표본평균으로 이루어진 점** $(\overline{X}, \overline{Y})$를 지난다. 따라서 $a = \overline{Y} - b\overline{X}$에 평균과 기울기를 대입하면 $a = 5 - (-1.4) \cdot 4 = 10.6$이다. 추정 회귀선은 "0176=10.6-1.4X이며 ①과 일치한다. X가 1 증가할 때 Y의 예측값은 1.4 감소한다는 뜻이다.

NoteBase

문 10. 다음은 자동차 부품을 생산하는 어느 공장의 현황이다. 이 공장에서 생산되는 제품 가운데 임의로 뽑힌 1개의 제품이 불량품일 때, 기계 1에서 생산되었을 확률은?

	생산량	불량률
기계 1	30%	1%
기계 2	40%	3%
기계 3	30%	2%

① $\frac{1}{7}$
② $\frac{2}{7}$
③ $\frac{3}{7}$
④ $\frac{4}{7}$

정답 1번

간단한 해설

전체 제품 중 기계 1에서 나온 불량품의 비중은 0.003이고, 전체 불량률은 0.021이다. **불량품 전체를 기준으로 다시 비중을 구하면** $\frac{0.003}{0.021} = \frac{1}{7}$이다.

NoteBase

계산 과정

기계 1에서 생산되었다는 사건을 A, 불량품이라는 사건을 D라 한다. 구하는 것은 **불량품으로 뽑혔다는 조건 아래의 생산 기계 확률**인 $P(A \mid D)$이다. 기계 1 자체의 불량률 $P(D \mid A)$와는 조건의 방향이 다르다.

표에서 기계 1·2·3의 생산량 비중은 각각 0.3, 0.4, 0.3이고, 각 기계의 불량률은 0.01, 0.03, 0.02이다. 전체 제품 중 각 기계에서 나온 불량품의 비중은 생산량 비중과 해당 불량률을 곱해 구한다.

기계 1: $0.3 \times 0.01 = 0.003$

기계 2: $0.4 \times 0.03 = 0.012$

기계 3: $0.3 \times 0.02 = 0.006$

어느 제품이든 세 기계 중 하나에서 생산되므로, **전체 불량률은 기계별 생산 비중으로 가중**하여 더한다.
$P(D) = 0.003 + 0.012 + 0.006 = 0.021$

전체 불량품 중 기계 1의 불량품이 차지하는 비중을 구하는 베이즈 정리를 적용한다.
$P(A \mid D) = \frac{P(A)P(D \mid A)}{P(D)} = \frac{0.003}{0.021} = \frac{1}{7}$
분모는 전체 제품 수가 아니라 불량품 전체에 해당하는 확률이다. 따라서 구한 확률은 ①과 일치한다.

NoteBase

문 11. $X_1, X_2, \cdots, X_{12}$**가 균일분포** $U(0, 2)$**의 확률표본이라 할 때 확률변수 Y=Σ_i=1^12X_i의 분포는 중심극한정리에 의해 근사적으로 정규분포** $N(\mu, \sigma^2)$**을 따른다. 이때** Y**의 평균과 분산은 각각 얼마인가?**

① 평균: 12, 분산: 4

② 평균: 12, 분산: 6

③ 평균: 6, 분산: 4

④ 평균: 6, 분산: 6

정답 1번

간단한 해설

각 표본값의 평균은 1, 분산은 1/3이다. Y는 독립인 12개 표본값의 **합**이므로 평균과 분산을 각각 더하여 평균 12, 분산 4를 얻는다.

계산 과정

개별 표본값의 평균·분산: 연속균일분포 $U(a, b)$의 평균은 $\frac{a+b}{2}$, 분산은 $\frac{(b-a)^2}{12}$이다. 따라서 X_i에 대해 $E(X_i) = \frac{0+2}{2} = 1$, $\mathrm{Var}(X_i) = \frac{(2-0)^2}{12} = \frac{1}{3}$이다.

합의 평균: Y는 표본평균이 아니라 12개 표본값의 합이다. 기댓값의 선형성에 따라 $E(Y) = \sum_{i=1}^{12} E(X_i) = 12 \times 1 = 12$이다.

합의 분산: 확률표본이므로 표본값들은 서로 독립이다. 따라서 **공분산이 0**이 되어 분산을 더할 수 있으며, $\mathrm{Var}(Y) = \sum_{i=1}^{12} \mathrm{Var}(X_i) = 12 \times \frac{1}{3} = 4$이다.

정규근사와 선택지: 중심극한정리를 적용한 근사분포는 $N(12, 4)$이다. 평균 12와 분산 4는 위 조건에서 정확하게 계산되는 값이며, 근사라는 말은 합의 분포 모양을 정규분포로 나타내는 데 해당한다. 따라서 평균 12, 분산 4인 ①과 일치한다.

NoteBase

문 12. $X_1, X_2, \cdots, X_{10}$**은 정규분포** $N(\mu_1, \sigma^2)$**의 확률표본이고** $Y_1, Y_2, \cdots, Y_{15}$**은 정규분포** $N(\mu_2, \sigma^2)$**의 확률표본이다.** X**와** Y**는 서로 독립이며** σ^2**은 모르는 상수이다.** S_1^2**과** S_2^2**이 각각** X**와** Y**의 표본분산일 때 공통분산** σ^2**의 추정량인 통합표본분산(Pooled Sample Variance)** S_p^2**를 잘 나타낸 것은?**

① $\frac{1}{23}[9S_1^2 + 14S_2^2]$
② $\frac{1}{2}[S_1^2 + S_2^2]$
③ $\frac{1}{23}[14S_1^2 + 9S_2^2]$
④ $\frac{1}{25}[15S_1^2 + 10S_2^2]$

정답 1번

간단한 해설

통합표본분산은 두 표본분산을 **각 표본의 자유도**로 가중평균한 값이다. 자유도가 각각 9와 14이므로 $S_p^2 = \frac{9S_1^2 + 14S_2^2}{23}$이다.

계산 과정

적용 조건: 두 정규모집단의 분산은 공통으로 σ^2이므로, 두 표본의 분산 정보를 통합하여 추정할 수 있다.
통합표본분산은 표본분산을 **각 표본의 자유도로 가중평균**한 값이다. 두 모평균이 같을 필요는 없으며, 각 표본은 자기 표본평균으로부터의 산포를 계산한다.

편차 제곱합과 자유도: 첫 표본의 크기는 10이므로 표본분산의 자유도는 9이다. 따라서 첫 표본의 평균에 대한 편차 제곱합은 $9S_1^2$이다. 둘째 표본의 크기는 15이므로 자유도는 14이며, 편차 제곱합은 $14S_2^2$이다. **두 표본평균을 각각 추정**했으므로 전체 자유도는 10 + 15 − 2 = 23이다.

통합 계산: 두 집단 내 편차 제곱합을 더하고 자유도의 합으로 나눈다.
$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$
$S_p^2 = \frac{9S_1^2 + 14S_2^2}{23}$
표본 크기가 서로 다르므로 단순 평균이 아니라 자유도 9와 14를 각각의 표본분산에 대응시켜 가중한다.

따라서 통합표본분산을 나타낸 선택지는 ①이다.

NoteBase

문 13. 표본의 크기가 n인 어느 자료에 단순회귀모형 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2)$을 적용하여 모형의 유의성 검정을 하였다. 다음 설명 중 틀린 것은?

① 모형의 유의성 검정을 위한 귀무가설은 $H_0 : \beta_1 = 0$이다.

② Y_i와 ε_i의 분산은 같다.

③ 분산분석 검정을 위한 검정 통계량은 귀무가설 하에서 $F(1, n-1)$분포를 따른다.

④ 최소제곱법에 의해 얻은 $\hat{\beta}_1$은 β_1의 불편추정량이다.

정답 3번

간단한 해설

절편과 기울기를 추정하는 단순회귀모형의 **잔차 자유도는 n − 2**이다. 따라서 모형 유의성 검정의 F 통계량은 귀무가설 아래에서 $F(1, n-2)$ 분포를 따른다.

선지별 해설

절편을 포함하는 통상적인 단순선형회귀모형의 유의성 검정이다. 설명변수 값은 고정하거나 주어진 값에 조건부로 해석하며, 오차의 독립성·정규성·등분산성 및 설명변수의 변동이 있는 검정 가능한 표본을 전제로 한다.

① O: 모형 유의성 검정은 설명변수가 반응변수의 평균과 선형적으로 관련되는지를 검정한다. 귀무가설은 $H_0 : \beta_1 = 0$, 대립가설은 H_1:_1"22600이다. 절편이 0인지 검정하는 것이 아니다.

② O: 고정된 X_i에서 $\beta_0 + \beta_1 X_i$는 상수이므로 이를 오차에 더해도 분산은 바뀌지 않는다. 따라서 $\mathrm{Var}(Y_i) = \mathrm{Var}(\varepsilon_i) = \sigma^2$이다. 설명변수 자체를 확률변수로 다룰 때에는 이 설명을 **주어진 설명변수 값에서의 조건부 분산**으로 이해해야 하며, 주변분산까지 항상 같다고 일반화하지 않는다.

③ X: 회귀제곱합 SSR의 자유도는 1이고, 잔차제곱합 SSE의 자유도는 **n − 2**이다. 표본 n개로 절편과 기울기 두 모수를 추정하기 때문이다. 따라서 $F = \frac{SSR/1}{SSE/(n-2)}$는 귀무가설 아래에서 $F(1, n-2)$ 분포를 따른다. n − 1은 전체제곱합의 자유도이며 분모의 잔차 자유도가 아니다.

④ O: 설명변수 값이 주어졌을 때 오차의 평균이 0인 모형에서 최소제곱 기울기 추정량은 참 기울기에 평균 0인 오차들의 가중합을 더한 형태가 된다. 따라서 E(β_1)=_1이므로 불편추정량이다. 불편성은 반복 표집에서의 평균에 관한 성질이지, 한 번 얻은 추정값이 항상 참값과 같다는 뜻은 아니다.

NoteBase

문 14. 확률변수 X의 기댓값 $E(e^{tX})$를 X의 적률생성함수라 할 때, 성공확률이 p인 베르누이 분포의 적률생성함수는? (단, $p+q=1$)

① $p+qe^t$

② $q+pe^t$

③ $1+e^{pt}$

④ $e^{pt}+e^{qt}$

정답 2번

간단한 해설

베르누이 변수는 실패 시 0, 성공 시 1이므로 **각 함수값에 실패·성공 확률을 곱해 더하면** $E(e^{tX})=qe^0+pe^t=q+pe^t$이다.

핵심 풀이

성공확률이 p인 베르누이 확률변수는 성공하면 1, 실패하면 0의 값을 갖는다. 따라서 $P(X=1)=p$, $P(X=0)=q=1-p$이다. **p와 q는 X의 값이 아니라 각 값이 나올 확률**이다.

이산확률변수의 기댓값은 각 함수값에 그 확률을 곱해 더한다. 적률생성함수의 정의를 적용하면 $M_X(t)=E(e^{tX})=qe^{t\cdot 0}+pe^{t\cdot 1}=q+pe^t$이다. 실패할 때의 함수값은 1이고, 성공할 때의 함수값은 e^t이다.

①의 $p+qe^t$는 성공과 실패의 확률을 반대로 배치한 식이다. p와 q가 같은 특수한 경우에는 값이 같지만, 주어진 성공확률 p에 대한 일반식은 아니다.

적률생성함수는 **t = 0에서 값이 1**이어야 한다. $E(e^{0X})=E(1)=1$이기 때문이다. ③의 $1+e^{pt}$와 ④의 $e^{pt}+e^{qt}$는 모두 t = 0에서 2가 되므로 적률생성함수가 될 수 없다.

따라서 구하는 적률생성함수는 $q+pe^t$이며, ②에 해당한다.

NoteBase

문 15. 표준오차에 대한 설명 중 틀린 것은?

① 표본평균의 분포가 가지는 표준편차, 즉 표본평균의 확률분포가 갖는 흩어짐의 정도를 의미한다.

② 표본의 크기가 클수록 표준오차는 커진다.

③ 일반적으로 표준오차는 모집단의 표준편차보다 작다.

④ 표본평균의 표준편차를 표준오차라 한다.

정답 2번

간단한 해설

서로 독립인 동일한 모집단의 표본에서 표본평균의 표준오차는 $\sigma/\sqrt{n}$이다. 모집단 표준편차가 같으면 **표본 크기가 커질수록 표준오차는 작아진다**.

선지별 해설

① O: 표준오차는 통계량의 표본분포가 갖는 표준편차이며, 이 문항에서는 표본평균의 표준오차를 다룬다. 같은 크기의 표본을 반복해서 뽑아 각각 평균을 구할 때 **표본평균들이 얼마나 흩어지는지**를 나타낸다. 한 표본 안의 개별 관측값들이 흩어지는 정도와는 구별된다.

② X: 서로 독립이고 같은 모집단에서 추출한 표본의 크기를 n, 모집단 표준편차를 σ라 하면 SE(X)=σ/√(n)이다. 따라서 σ가 일정할 때 n이 커지면 분모가 커져 표준오차는 **커지는 것이 아니라 작아진다**. 표본 크기를 4배로 늘리면 표준오차는 절반이 된다.

③ O: 위와 같은 독립표본에서 SE(X)=σ/√(n)이므로, **n이 1보다 크고 σ가 양수이면** 표본평균의 표준오차는 모집단 표준편차보다 작다. n = 1이면 둘이 같으므로 모든 경우에 반드시 작다고 일반화하는 것은 아니다. 여러 관측값을 평균내면 개별 관측값보다 변동이 줄어드는 관계를 설명한 선지이다.

④ O: 표본평균의 표본분포에 대한 표준편차를 표본평균의 표준오차라고 한다. 모집단 표준편차 σ를 모르면 표본표준편차 s를 이용해 $s/\sqrt{n}$으로 추정한다. 이때 s는 **개별 관측값의 산포**를, 표준오차는 표본을 바꾸어 추출할 때 생기는 평균의 산포를 나타낸다.

NoteBase

문 16. 확률변수 X의 누적분포함수 $F(x)$가 다음과 같이 주어져 있다.
이때 X의 기댓값은?
F(x)=cases0, & x<0\\(1)/(2)x^2, & 0≤ x<1\\(1)/(2), & 1≤ x<2\\1, & 2≤ xcases

① $\frac{2}{5}$
② $\frac{2}{4}$
③ $\frac{4}{3}$
④ $\frac{5}{4}$

정답 3번

간단한 해설

$0 < x < 1$에서는 밀도가 x이고, **x = 2에 확률 1/2이 따로 모여 있다**. 두 부분을 모두 반영하면 E[X]=∫ _0^1 x^2 dx+2·12=43이다.

계산 과정

누적분포함수는 $F(x) = P(X \le x)$이다. 주어진 함수는 0과 1 사이에서 연속적으로 증가하고, 2에서 불연속적으로 증가한다. 따라서 **연속적인 부분과 한 점에 모인 확률질량을 함께** 반영해야 한다.

$0 < x < 1$에서 연속 부분의 밀도는 $f(x) = F'(x) = x$이다. 이 구간의 확률은 ∫ _0^1 x dx=12이다. $1 < x < 2$에서는 누적분포함수가 일정하므로 확률이 없고, 0과 1에서도 누적분포함수의 점프가 없어 점확률은 0이다.

점 2에서의 확률은 **누적분포함수의 점프 크기**로 구한다. 왼쪽 극한은 F(2-)=12이고 $F(2) = 1$이므로, P(X=2)=F(2)-F(2-)=12이다.

기댓값은 연속 부분에서 값과 밀도를 곱한 적분에, 점 2의 값과 확률을 곱한 항을 더한다.

aligned&E[X]=∫ _0^1 x· x dx+2·12&=≤ft[(x^3)/(3)]_0^1+1=13+1=43aligned

따라서 기댓값은 ③과 일치한다. 밀도만 적분하여 점 2의 확률질량을 빠뜨리면 전체 기댓값을 구한 것이 아니다.

NoteBase

문 17. 분포에 대한 설명 중 틀린 것은?

① 두 확률변수 X, Y가 서로 독립이며 각각 이항분포 $Bin(n, p_1), Bin(m, p_2)$를 따를 때 $Z = X + Y$는 $p_1 = p_2$가 성립하면 이항분포를 따른다.

② 두 확률변수 X, Y가 서로 독립이며 각각 지수분포 $\exp(\theta_1), \exp(\theta_2)$를 따를 때 $Z = X + Y$는 $\theta_1 = \theta_2$가 성립하면 지수분포를 따른다.

③ 두 확률변수 X, Y가 서로 독립이며 카이제곱분포를 따를 때 $Z = X + Y$는 카이제곱분포를 따른다.

④ 두 확률변수 X, Y가 서로 독립이며 표준정규분포를 따를 때 $U = X + Y, V = X - Y$로 정의하면 두 확률변수 U, V는 같은 분포를 갖는다.

정답 2번

간단한 해설

같은 모수를 갖는 독립인 지수분포 두 개의 합은 지수분포가 아니라 **형상모수 2의 감마분포**를 따른다. 지수분포에 해당하는 형상모수 1이 합산되어 2가 되기 때문이다.

선지별 해설

① O: 두 이항변수가 독립이고 **성공확률이 동일**하면, 각각의 성공 횟수를 더한 것은 전체 시행의 성공 횟수이다. 공통 성공확률을 p라 하면 $X + Y \sim Bin(n + m, p)$이다. 시행 횟수 n과 m은 서로 달라도 된다.

② X: 같은 모수의 독립인 지수분포 두 개를 더하면 **형상모수가 2인 감마분포**, 즉 얼랑분포를 따른다. 지수분포는 형상모수가 1인 감마분포에 해당하고, 동일한 척도 또는 동일한 율을 가진 독립 감마변수의 합에서는 형상모수가 더해진다. 따라서 두 지수변수의 합은 원래와 같은 지수분포가 아니다.

③ O: 독립인 카이제곱변수의 합은 **자유도를 더한 카이제곱분포**를 따른다. X와 Y의 자유도가 각각 r과 s라면 $X + Y \sim \chi^2(r + s)$이다. 카이제곱변수를 독립 표준정규변수들의 제곱합으로 나타낼 수 있으므로, 두 독립 제곱합을 합치면 자유도가 합쳐진다.

④ O: 독립 표준정규변수의 선형결합은 정규분포이다. U = X + Y와 V = X − Y의 평균은 모두 0이고, 독립성 때문에 공분산 항이 0이므로 분산도 모두 2이다. 특히 $Var(X - Y) = Var(X) + (-1)^2 Var(Y) = 2$이므로, 두 변수는 모두 $N(0, 2)$를 따른다. Y를 뺀다고 해서 분산까지 빼는 것은 아니다.

따라서 틀린 설명은 ②이다.

NoteBase

문 18. 공평한 동전 4개를 동시에 던졌을 때, 적어도 1개의 앞면이 나올 확률은?

① $\frac{1}{2}$
② $\frac{1}{4}$
③ $\frac{15}{16}$
④ $\frac{1}{16}$

정답 3번

간단한 해설

적어도 하나의 앞면이 나오는 사건의 여사건은 **네 동전이 모두 뒷면인 사건**이다. 따라서 구하는 확률은 1-≤ft(12)^4=(15)/(16)이다.

계산 과정

'적어도 1개'는 앞면이 1개, 2개, 3개 또는 4개 나오는 경우를 모두 포함한다. 이 사건의 여사건은 **앞면이 하나도 나오지 않는 사건**, 즉 네 동전이 모두 뒷면인 사건이다.

통상적인 동전 던지기 실험처럼 네 동전의 결과를 서로 독립으로 본다. 공평한 동전은 뒷면이 나올 확률이 12이므로, 모두 뒷면일 확률은 ≤ft(12)^4=116이다. 독립이므로 네 확률을 곱한다.

따라서 적어도 하나의 앞면이 나올 확률은 1-116=(15)/(16)이며, ③에 해당한다. ④의 116은 구하려는 사건이 아니라 모두 뒷면인 여사건의 확률이다.

NoteBase

문 19. 베이즈 추론의 세 가지 기본요소에 해당되지 않는 것은?

① 사전분포(Prior Distribution)

② 가능도 함수(Likelihood Function)

③ 표본분포(Sample Distribution)

④ 사후분포(Posterior Distribution)

정답 3번

간단한 해설

베이즈 추론은 **사전분포**에 자료의 **가능도**를 반영해 **사후분포**를 구하는 과정이다. 표본분포는 이 세 요소 중 하나로 별도 열거되는 항목이 아니다.

선지별 해설

① O: 사전분포는 현재 자료를 관측하기 전에 모수에 대해 갖는 불확실성을 확률분포로 나타낸 것이다. 모수를 θ라 하면 $p(\theta)$로 표시하며, 기존 정보나 가정을 추론에 반영하는 출발점이다.

② O: 가능도 함수는 관측된 자료 x를 고정하고, 그 자료가 나타날 확률 또는 확률밀도를 **모수 θ의 함수**로 보는 것이다. $L(\theta; x) = p(x \mid \theta)$로 쓰며, 서로 다른 모수값이 관측 자료를 얼마나 잘 설명하는지 반영한다. 가능도 자체가 θ에 대한 확률분포로 정규화되어 있어야 하는 것은 아니다.

③ X: 표본분포는 반복해서 표본을 추출할 때 표본평균 같은 통계량이 가지는 확률분포를 말한다. 베이즈 추론도 자료의 확률모형을 이용하지만, 이 문제에서 묻는 세 가지 기본요소는 **사전분포·가능도 함수·사후분포**이다. 자료의 모형이 가능도를 제공한다는 사실과 표본분포를 세 요소 중 하나로 별도 나열하는 것은 구별된다.

④ O: 사후분포는 자료를 관측한 뒤 모수에 대한 불확실성을 갱신한 분포이다. $p(\theta \mid x) \propto p(x \mid \theta)p(\theta)$이므로 사전분포와 가능도를 곱하고 전체 확률이 1이 되도록 정규화해 얻는다. 즉 사후분포는 **관측 자료를 반영한 모수의 분포**이다.

NoteBase

문 20. 어느 나라 국민의 2%가 특정 질병에 걸려있다고 한다. 이 병의 검진방법에 의하면 감염자의 98%가 양성 반응을 보이며, 비감염자의 경우도 3%가 거짓 양성 반응을 보인다고 한다. 한 사람이 검진 결과 양성을 보였다면, 이 사람이 실제로 병에 걸렸을 확률은?

① 0.060

② 0.196

③ 0.400

④ 0.490

정답 3번

간단한 해설

전체 중 실제 감염·양성인 비율은 0.0196이고, 거짓 양성을 포함한 양성 비율은 0.049이다. **양성 전체를 기준으로 실제 감염자의 비중**을 구하면 $\frac{0.0196}{0.049} = 0.400$이다.

계산 과정

감염된 사건을 D, 양성 반응이 나온 사건을 T라 한다. 주어진 확률은 $P(D) = 0.02$, $P(D^c) = 0.98$, $P(T \mid D) = 0.98$, $P(T \mid D^c) = 0.03$이다. 구하는 것은 **양성인 사람 중 실제 감염자의 비율**인 $P(D \mid T)$이며, 감염자 중 양성일 확률과 조건의 방향이 반대이다.

실제로 감염되어 양성인 경우와 비감염자인데 양성인 경우를 각각 구한다.
$P(D \cap T) = 0.02 \times 0.98 = 0.0196$
$P(D^c \cap T) = 0.98 \times 0.03 = 0.0294$
두 식의 0.98은 각각 감염자의 양성 반응률과 전체 중 비감염자의 비율이므로 의미가 다르다.

양성 반응은 위 두 경우를 모두 포함한다. 따라서 전체 양성 확률은 $P(T) = 0.0196 + 0.0294 = 0.049$이다. **거짓 양성도 분모에 포함**해야 하며, 감염자의 양성 반응률 0.98만으로 사후확률을 판단할 수 없다.

베이즈 정리에 따라 양성 전체에서 실제 감염·양성이 차지하는 비중을 구한다.
$P(D \mid T) = \frac{P(D \cap T)}{P(T)} = \frac{0.0196}{0.049} = 0.400$

따라서 구하는 확률은 40%이며 ③과 일치한다.

NoteBase

문 21. 확률변수 X의 적률생성함수가 다음과 같이 주어졌다.
$M(t) = E(e^{tX}) = [pe^t + 1 - p]^n, \quad 0 < p < 1$
주어진 적률생성함수를 한 번 미분한 다음 $t = 0$을 대입하면 X의 평균, 즉 $E(X) = M'(0)$임이 알려져 있다. 이때 X의 평균은?

① p

② np

③ $np(1-p)$

④ $p(1-p)$

정답 2번

간단한 해설

주어진 적률생성함수에 **연쇄법칙**을 적용하면 $M'(t) = n[pe^t + 1 - p]^{n-1}pe^t$이다. 여기에 t = 0을 대입하면 대괄호 안이 1이 되어 평균은 np이다.

계산 과정

적용할 관계: 문제에서 평균은 $E(X) = M'(0)$이라고 주어졌다. 따라서 $M(t) = [pe^t + 1 - p]^n$을 t에 관해 한 번 미분한 뒤 t = 0을 대입한다.

연쇄법칙으로 미분: 바깥의 n제곱을 미분한 값에 안쪽 함수의 미분값을 곱한다. n과 p는 t에 대한 상수이고, $pe^t + 1 - p$의 미분값은 pe^t이다. 따라서 $M'(t) = n[pe^t + 1 - p]^{n-1}pe^t$이다.

t = 0 대입: $e^0 = 1$이고 $p + 1 - p = 1$이므로 $M'(0) = n \times 1^{n-1} \times p = np$이다. 따라서 구하는 평균은 np이며 ②에 해당한다. **미분하기 전에 t = 0을 대입**해서 얻는 $M(0) = 1$은 평균을 구한 값이 아니다.

NoteBase

문 22. 성별에 따라 색상에 대한 선호도 차이가 있는지를 알아보기 위해 어느 학교 남녀학생 각각 500명씩을 임의로 추출하여 노랑, 빨강, 파랑 세 가지 색상에 대한 선호도를 조사한 결과가 다음과 같다. 만일, 성별에 따라 선호하는 색상에 차이가 없다면 빨강을 선호하는 여학생 수에 대한 기대도수(Expected Frequency)는?

	노랑	빨강	파랑	합계
남학생	170	150	180	500
여학생	170	200	130	500
합 계	340	350	310	1,000

① 200

② 175

③ 350

④ 150

정답 2번

간단한 해설

성별에 따른 선호 차이가 없다면 전체 자료의 빨강 선호 비율을 여학생에게도 적용한다. **여학생 행 합계 500명과 빨강 열 합계 350명**을 이용하면 기대도수는 $\frac{500 \times 350}{1,000} = 175$명이다.

NoteBase

계산 과정

가정의 의미: 성별에 따라 색상 선호에 차이가 없다는 것은 **남학생과 여학생의 색상별 선호 비율이 같다**는 뜻이다. 노랑·빨강·파랑의 선호 비율이 서로 모두 같다는 뜻은 아니므로, 학생 수를 단순히 세 색으로 똑같이 나누지 않는다.

표에서 필요한 값: 여학생 행의 합계는 500명이다. 빨강 열의 합계는 남학생 150명과 여학생 200명을 합한 350명이며, 전체 학생 수는 1,000명이다. 성별 간 선호 비율이 같다는 가정 아래 공통 빨강 선호 비율은 전체 자료로부터 $\frac{350}{1{,}000}$으로 추정한다.

기대도수 계산: 여학생 수에 공통 빨강 선호 비율을 곱한다. 이는 분할표에서 **해당 행 합계와 열 합계의 곱을 전체 합계로 나누는** 계산이다.
$E_{\text{여학생, 빨강}} = 500 \times \frac{350}{1{,}000} = 175$
표에 나타난 여학생의 빨강 선호 관측도수 200명과, 차이가 없다는 가정 아래 계산한 기대도수 175명은 구분된다.

따라서 기대도수는 175명이며 ②에 해당한다.

NoteBase

문 23. 자동차 부품을 생산하는 각 회사의 제품품질이 균일한지 알아보기 위해 네 회사로부터 부품 3개씩을 임의로 추출하여 부품의 강도를 측정한 결과 다음과 같은 분산분석표를 얻었다. a의 값은?

요인	제곱합	자유도	F 비
처리	a		4
잔차			
계	70		

① 25

② $\frac{23}{1120}$

③ $\frac{3}{140}$

④ 42

정답 4번

간단한 해설

처리 자유도는 3, 잔차 자유도는 8이므로 **평균제곱의 비**를 사용해야 한다. $\frac{a/3}{(70-a)/8} = 4$를 풀면 처리 제곱합은 a = 42이다.

NoteBase

계산 과정

회사가 4개이고 각 회사에서 부품 3개씩을 추출했으므로 집단 수 k = 4, 전체 관측 수 N = 12이다.
일원분산분석에서 처리 자유도는 $k-1=3$, 잔차 자유도는 $N-k=8$이다. 전체 자유도는 11이며, 두 자유도의 합과 일치한다.

표의 a는 **처리 제곱합**이다. 전체 제곱합 70은 처리 제곱합과 잔차 제곱합의 합이므로 잔차 제곱합은 $70-a$이다.

F 비는 **제곱합의 비가 아니라 각 제곱합을 자유도로 나눈 평균제곱의 비**이다. 따라서 처리 평균제곱은 $\frac{a}{3}$, 잔차 평균제곱은 $\frac{70-a}{8}$이며, 표의 F 비 4를 적용하면 다음과 같다.
$\frac{a/3}{(70-a)/8}=4$

양변을 정리하여 a를 구한다.
$8a=12(70-a)$
$20a=840$
$a=42$

검산하면 잔차 제곱합은 28, 처리 평균제곱은 $\frac{42}{3}=14$, 잔차 평균제곱은 $\frac{28}{8}=3.5$이다. 따라서 $F=\frac{14}{3.5}=4$로 표와 일치한다. 구한 처리 제곱합 42는 ④와 일치한다.

NoteBase

문 24. $X_1, X_2, \cdots, X_9$**는 감마분포** $Gam(4, 3)$**의 확률표본이고,** $\overline{X}$**는 표본평균이다. 중심극한정리를 이용하여 구한** $P(8.71 \leq \overline{X} \leq 15.92)$**의 근사확률은?**
(단, $P(|Z| \geq 0.65) = 0.5156$**,** $P(|Z| \geq 0.55) = 0.5824$**,** $P(|Z| \geq 1.96) = 0.05$**,** $P(|Z| \geq 1.645) = 0.1$**)**

① 0.451

② 0.549

③ 0.85

④ 0.925

정답 4번

간단한 해설

표본평균의 평균은 12, **표준오차는 2**이므로 구간을 표준화하면 −1.645부터 1.96까지이다. 주어진 양측 꼬리확률을 각각 절반으로 나누어 빼면 근사확률은 1 − 0.05 − 0.025 = 0.925이다.

계산 과정

이 문항의 $Gam(4, 3)$는 형상모수 4와 척도모수 3을 사용하는 표기이다. 형상모수가 α, 척도모수가 θ이면 평균은 $\alpha\theta$, 분산은 $\alpha\theta^2$이므로 $E(X_i) = 4 \times 3 = 12$, $\mathrm{Var}(X_i) = 4 \times 3^2 = 36$이다. 둘째 모수를 발생률로 사용하는 다른 표기와 구별한다.

크기 9의 독립·동일분포 확률표본이므로 E(X)=12, Var(X)=(36)/(9)=4이다. 따라서 **표본평균의 표준오차는 2**이다. 개별 관측값의 표준편차 6을 그대로 사용하지 않는다.

중심극한정리를 이용하면 X를 평균 12, 분산 4인 정규분포로 근사한다. 표준화한 양쪽 경계는 $\frac{8.71-12}{2} = -1.645$, $\frac{15.92-12}{2} = 1.96$이므로, 구할 확률은 P(-1.645≤ Z≤1.96)로 근사된다.

제시된 확률은 **양측 꼬리확률**이다. 표준정규분포의 대칭성에 따라 P(Z-1.645)=(0.1)/(2)=0.05, P(Z1.96)=(0.05)/(2)=0.025이다. 따라서 구간확률은 1 − 0.05 − 0.025 = 0.925이다.

중심극한정리로 구한 근사확률은 0.925이며, ④에 해당한다. 표본평균이 정확히 정규분포를 따른다는 뜻은 아니다.

NoteBase

문 25. $X_1, X_2, \cdots, X_n$**이 정규분포** $N(\mu, \sigma^2)$**의 확률표본일 때, 확률변수** $Y = (n-1)S^2/\sigma^2$**은 자유도** $(n-1)$**인 카이제곱** $\chi^2(n-1)$ **분포를 따른다. 다음 기댓값은 알려졌다고 하자.**
$E\left[\frac{1}{S}\right] = \sqrt{\frac{n-1}{2}} \cdot \frac{\Gamma\left(\frac{n}{2}-1\right)}{\Gamma\left(\frac{n-1}{2}\right)} \cdot \frac{1}{\sigma}$
$n = 5$**일 때** $\frac{C}{S}$**가** $\frac{1}{\sigma}$**의 불편추정량이면** C**의 값은? (단,** S^2**은 표본분산,** C**는 상수,** $\Gamma(1/2) = \sqrt{\pi}$**)**

① $\sqrt{\pi}$
② $\sqrt{\frac{\pi}{2}}$
③ $\frac{1}{\sqrt{\pi}}$
④ $\sqrt{\frac{2}{\pi}}$

정답 4번

간단한 해설

n = 5를 주어진 식에 넣으면 $E[1/S] = \sqrt{\pi/2}/\sigma$이다. 불편추정량이 되려면 이 계수의 **역수**를 곱해야 하므로 $C = \sqrt{2/\pi}$이다.

계산 과정

불편추정량의 조건: 추정량의 기댓값이 추정하려는 모수와 같아야 한다. 따라서 $E[C/S] = 1/\sigma$가 되어야 한다. C는 상수이므로 $E[C/S] = CE[1/S]$이며, 주어진 기댓값에서 $1/\sigma$ 앞에 곱해지는 **계수의 역수**를 C로 정하면 된다.

감마함수 값: n = 5를 대입하면 $\frac{n-1}{2} = 2$, (n)/(2)-1=32이다. 감마함수의 관계 $\Gamma(z+1) = z\Gamma(z)$를 이용하면 (3/2)=12(1/2)=√(π)2이고, $\Gamma(2) = 1$이다.

기댓값 계산: 위 값을 문제에 주어진 식에 대입하면 E[1/S]=2·(√(π)/2)/(1)·1σ=(π)/(2)·1σ이다. 따라서 $1/S$는 그대로는 $1/\sigma$의 불편추정량이 아니며, 계수 $\sqrt{\pi/2}$를 보정해야 한다.

상수 결정: C(π)/(2)·1σ=1σ이므로 C=√(2π)이다. 이 값을 곱하면 $E[C/S] = 1/\sigma$가 되어 불편성을 만족하며, 해당 값은 선택지 ④이다.

NoteBase

문 26. 어느 도시 성인들의 키(단위: cm)와 몸무게(단위: kg)의 관계를 알아보기 위해 이 도시의 성인 100명을 임의로 뽑아 키를 독립변수, 몸무게를 종속변수로 갖는 단순 회귀모형 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2)$를 적합한 결과, 회귀식 $\hat{Y} = 0.6X - 32$과 결정계수 $R^2 = 0.75$을 얻었다. 다음 설명 중 틀린 것은?

① 오차인 ε_i와 종속변수 Y_i는 동일한 분산을 갖는다.

② 키 X가 $\overline{X}$일 때 몸무게 Y는 $\overline{Y}$이다.

③ 키가 10cm 증가하면 평균 몸무게는 6kg 증가한다.

④ 적합된 회귀식에 의해 키가 몸무게를 설명하는 정도는 75%이므로 키와 몸무게의 피어슨 상관계수는 0.75이다.

정답 4번

간단한 해설

절편을 포함하는 단순회귀에서 결정계수는 **피어슨 상관계수의 제곱**이다. 기울기 0.6이 양수이므로 $r = +\sqrt{0.75} \approx 0.866$이며 0.75가 아니다.

선지별 해설

키 X를 설명변수, 몸무게 Y를 반응변수로 한 회귀모형과 표본 100명에서 얻은 적합 회귀식을 구분한다. 회귀식의 값은 주어진 키에서의 평균 몸무게를 예측한 값이며, 개개인의 실제 몸무게와 반드시 같지는 않다.

① O: 주어진 회귀모형에서 X_i를 고정하면 $\beta_0 + \beta_1 X_i$는 상수이고, Y_i의 변동은 오차 ε_i에서 생긴다. 따라서 이 조건에서 두 변수의 분산은 σ^2로 같다. 이는 **키가 주어진 조건에서의 분산**에 관한 설명이며, 서로 다른 키를 가진 사람들을 모두 섞은 몸무게의 주변분산까지 오차분산과 같다는 뜻은 아니다.

② O: 적합된 회귀식에 관한 설명으로 보면, 절편을 포함한 최소제곱 회귀선은 **표본평균점** $(\overline{X}, \overline{Y})$를 지난다. 절편이 β_0=Y-β_1X이므로 "0176(X)=0.6X-32=Y이다. 선지의 Y는 이 문맥에서 회귀선 위 예측값을 가리키며, 키가 평균인 개인의 실제 몸무게가 반드시 표본평균과 같다는 의미는 아니다.

③ O: 적합 회귀식의 기울기는 키 1cm 차이에 대한 예측 평균 몸무게의 차이 0.6kg을 뜻한다. 따라서 키가 10cm 더 큰 경우 회귀식이 예측하는 평균 몸무게는 $0.6 \times 10 = 6$kg 더 크다. 이는 회귀관계의 해석이며, 한 사람의 키를 늘리면 몸무게가 반드시 그만큼 늘어난다는 인과적 보장은 아니다.

④ X: 결정계수 0.75는 표본 몸무게의 평균 주위 변동 중 적합 회귀식이 설명하는 비율이 75%라는 뜻이다. 절편을 포함하는 단순회귀에서는 $R^2 = r^2$이고 r의 부호는 회귀기울기의 부호와 같다. 따라서 기울기가 0.6인 이 자료에서는 $r = +\sqrt{0.75} = \frac{\sqrt{3}}{2} \approx 0.866$이다. 결정계수 0.75를 상관계수에 그대로 대입한 부분이 잘못되었다.

NoteBase

문 27. X_1, X_2, X_3, X_4**는 균일분포** $U(0, 1)$**의 확률표본이고,** $X_{(1)} \le X_{(2)} \le X_{(3)} \le X_{(4)}$**는** X_1, X_2, X_3, X_4**의 순서통계량이라 하자.** $E[X_{(2)}] = a$**,** $P(X_{(4)} < 1/2) = b$**라 하면** $a + b$**는?**

① $\frac{9}{16}$
② $\frac{11}{35}$
③ $\frac{25}{62}$
④ $\frac{37}{80}$

정답 4번

간단한 해설

두 번째 순서통계량의 평균은 $\frac{2}{5}$이다. 최댓값이 1/2보다 작다는 것은 **네 표본값이 모두 1/2보다 작다**는 뜻이므로 그 확률은 $\frac{1}{16}$이고, 합은 $\frac{37}{80}$이다.

계산 과정

a의 계산: $X_{(2)}$는 네 관측값을 작은 순서로 정렬했을 때 **두 번째로 작은 값**이다. U(0,1)에서 뽑은 크기 n의 확률표본의 k번째 순서통계량은 $Beta(k, n+1-k)$ 분포를 따르며 평균은 $\frac{k}{n+1}$이다. 여기서는 n = 4, k = 2이므로
$a = E[X_{(2)}] = \frac{2}{4+1} = \frac{2}{5}$

b의 계산: $X_{(4)}$는 표본의 최댓값이다. 최댓값이 1/2보다 작으려면 **네 관측값이 모두 1/2보다 작아야** 한다. 각 관측값은 U(0,1)을 따르므로 그 확률은 1/2이고, 원래 표본들이 독립이므로 확률을 곱한다.
b=P≤ft(X_(4)(1)/(2))=_i=1^4P≤ft(X_i(1)/(2))=≤ft((1)/(2))^4=(1)/(16)

합의 계산: 두 결과를 같은 분모로 통분한다.
$a + b = \frac{2}{5} + \frac{1}{16} = \frac{32}{80} + \frac{5}{80} = \frac{37}{80}$

따라서 구하는 값은 $\frac{37}{80}$이며 ④에 해당한다.

NoteBase

문 28. 상관계수의 성질에 대한 설명 중 틀린 것은?

① 공분산과 상관계수의 부호는 서로 같다.

② 상관계수는 -1과 +1 사이의 값을 갖는다.

③ 두 변수 간 관계의 성질과 강도를 알려준다.

④ 적합된 회귀선 기울기가 양(+)이면 상관계수는 음(-)의 값을 갖는다.

정답 4번

간단한 해설

절편을 포함하는 단순최소제곱회귀에서 기울기는 β_1=r(s_y)/(s_x)이다. 표준편차의 비가 양수이므로 **기울기가 양수이면 상관계수도 양수**이다.

선지별 해설

여기서 상관계수는 두 변수의 선형관계를 나타내는 피어슨 상관계수 r이다. 두 변수의 표준편차가 모두 양수인 경우를 전제로 하며, 표본공분산을 s_{xy}, 각 표본표준편차를 s_x, s_y라 하면 $r = \frac{s_{xy}}{s_x s_y}$이다.

① O: 표준편차의 곱 $s_x s_y$는 양수이므로 공분산을 이 값으로 나누어도 부호는 바뀌지 않는다. 따라서 **공분산과 상관계수의 부호는 같다**. 공분산이 0이면 상관계수도 0이다.

② O: 코시·슈바르츠 부등식에 의해 $-1 \le r \le 1$이다. r = 1이면 완전한 양의 선형관계, r = -1이면 완전한 음의 선형관계이며, 이 양 끝값도 허용된다.

③ O: 상관계수의 부호는 선형관계의 방향을 나타내고, 절댓값은 선형관계의 강도를 나타낸다. 절댓값이 1에 가까울수록 점들이 직선에 더 가깝게 놓인다. 다만 **선형관계의 척도**이므로 r = 0이어도 비선형관계는 존재할 수 있으며, 상관관계만으로 인과관계를 증명할 수는 없다.

④ X: 절편을 포함하는 단순최소제곱회귀에서 Y를 X에 회귀한 기울기는 β_1=(s_xy)/(s_x^2)=r(s_y)/(s_x)이다. 표준편차의 비가 양수이므로 **회귀선 기울기와 상관계수의 부호는 같다**. 따라서 기울기가 양수이면 상관계수도 양수이며, 음수라는 설명은 반대이다.

NoteBase

문 29. 공평한 육면체 주사위를 던져 나오는 수를 X라 하고 조건부 분포 $Y|X$는 성공확률이 0.2인 이항분포 $Bin(X, 0.2)$를 따른다고 하자. 이때 확률변수 Y의 평균 $E(Y)$는?

① 0.3

② 0.5

③ 0.7

④ 0.9

정답 3번

간단한 해설

조건부 이항분포의 평균은 $E(Y \mid X) = 0.2X$이다. **이를 X에 대해 다시 평균**하면
$E(Y) = 0.2E(X) = 0.2 \times 3.5 = 0.7$이다.

계산 과정

주사위 결과가 x로 정해졌을 때 Y는 시행횟수 x, 성공확률 0.2인 이항분포를 따른다. 이항분포의 평균은 시행횟수와 성공확률의 곱이므로 $E(Y \mid X = x) = 0.2x$이고, $E(Y \mid X) = 0.2X$이다. **조건부 평균도 X에 따라 변한다**는 점이 중요하다.

공평한 육면체 주사위의 각 눈은 확률 16로 나온다. 따라서 $E(X) = \frac{1+2+3+4+5+6}{6} = 3.5$이다.

전체 평균은 각 조건부 평균을 해당 조건이 발생할 확률로 가중평균하여 구한다. **총기댓값 법칙**을 적용하면
aligned&E(Y)=E\E(Y X)\=E(0.2X)&=0.2E(X)=0.2×3.5=0.7aligned
이다. 이 법칙을 쓰기 위해 X와 Y가 독립일 필요는 없다.

따라서 Y의 평균은 0.7이며, ③에 해당한다.

NoteBase

문 30. 단순선형회귀모형 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ \ i = 1, 2, \ldots, n$에 대한 설명 중 틀린 것은?

① 오차항 ε_i는 관측될 수 없는 값이기 때문에 오차항의 추정량인 잔차를 이용하여 등분산성, 독립성, 정규성에 대한 타당성을 조사한다.

② 오차항 ε_i는 서로 상관을 가지고, 평균이 0, 분산이 σ^2인 동일한 분포를 따른다.

③ β_0는 회귀직선의 절편이다.

④ β_1은 회귀직선의 기울기이다.

정답 2번

간단한 해설

고전적 단순선형회귀모형은 서로 다른 관측의 오차가 **상관되지 않는 것**을 기본으로 한다. 오차들이 서로 상관을 가진다는 설명은 이 가정과 반대이다.

선지별 해설

이 문항은 고전적 단순선형회귀의 오차 가정과 회귀계수의 의미를 묻는다. 회귀직선은 $E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i$로 나타내며, 오차는 실제 반응값과 이 조건부 평균의 차이이다.

① O: 모수의 참값을 모르므로 오차 ε_i를 직접 계산할 수 없다. 대신 관측값에서 적합값을 뺀 **잔차** e_i=Y_i-"0176_i를 이용한다. 잔차와 적합값의 그림으로 산포 변화, 관측 순서에 따른 잔차 그림으로 연속적 패턴, 정규확률그림으로 정규성의 타당성을 살핀다. 잔차는 추정 과정의 영향을 받으므로 오차와 완전히 같은 값이나 동일한 독립 확률변수로 취급하는 것은 아니다.

② X: 기본 오차 가정은 $E(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$이며, 서로 다른 오차 사이에는 $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$을 둔다. 정규오차를 이용하는 통상적인 추론에서는 독립인 정규오차를 가정한다. 따라서 '서로 상관을 가진다'가 아니라 **서로 상관되지 않는다**로 고쳐야 한다. 평균과 분산이 같다는 사실만으로 독립성이 자동으로 성립하는 것은 아니다.

③ O: β_0는 회귀직선의 절편으로, X = 0에서의 반응변수 조건부 평균을 나타낸다. 다만 X = 0이 자료의 관측 범위에서 멀리 떨어져 있으면 절편에 현실적인 의미를 부여하는 데 주의가 필요하다.

④ O: β_1은 회귀직선의 기울기로, X가 1단위 증가할 때 반응변수의 조건부 평균이 얼마나 변하는지를 나타낸다. 양수이면 평균이 증가하고 음수이면 감소하며, 회귀기울기만으로 인과관계가 입증되는 것은 아니다.

NoteBase

문 31. 분산분석의 성질에 대한 설명 중 틀린 것은?

① 각 집단의 관측된 데이터 수는 반드시 같아야 한다.

② 3개 이상 집단의 평균을 비교하는 분석방법이다.

③ 각 집단의 분포가 정규분포이고, 개별 관측값은 독립이라는 가정이 필요하다.

④ 총변동을 집단 간의 변동과 집단 내의 변동으로 분해하여 집단 내의 변동보다 집단 간의 변동이 더 큰지를 분석한다.

정답 1번

간단한 해설

분산분석에서는 **집단별 표본 크기가 반드시 같을 필요가 없다**. 표본 수가 다른 불균형 자료도 분석할 수 있으며, 각 집단의 표본 수와 자유도를 계산에 반영한다.

선지별 해설

① X: 집단마다 관측값 수가 같은 자료를 균형자료라고 하지만, 분산분석이 균형자료에서만 가능한 것은 아니다. **집단별 표본 크기가 달라도** 해당 표본 수를 반영하여 집단 간 제곱합과 집단 내 제곱합을 구하고 검정할 수 있다. 다만 표본 수가 다를 때도 독립성·정규성·등분산성 등 적용하는 분석법의 가정을 검토해야 한다.

② O: 분산분석은 여러 집단의 모집단 평균이 같은지 동시에 비교하는 방법이며, 특히 세 집단 이상을 비교할 때 활용된다. 귀무가설은 **모든 집단의 평균이 같다**는 것이고, 대립가설은 적어도 한 집단의 평균이 다르다는 것이다. 두 집단에도 적용할 수 있으므로 '세 집단 이상에서만 가능하다'는 뜻으로 해석하지 않는다.

③ O: 고전적인 일원분산분석의 F 검정에서는 각 집단의 오차가 정규분포를 따르고 관측값들이 독립이라는 가정을 둔다. 또한 집단들의 오차분산이 같다는 **등분산성**을 가정한다. 이 선지는 그 가정 중 정규성과 독립성을 설명한 것이다.

④ O: 총제곱합은 집단 간 제곱합과 집단 내 제곱합으로 나뉜다. 실제 검정에서는 두 제곱합을 그대로 비교하지 않고 **각각의 자유도로 나눈 평균제곱**을 비교한다. 집단 수가 k, 전체 관측값 수가 N이면 $MS_B = SS_B/(k-1)$, $MS_W = SS_W/(N-k)$이며 $F = MS_B/MS_W$를 구한다. 이 값이 귀무가설 아래 기대되는 범위보다 충분히 커서 정해진 유의수준의 임계값을 넘으면 평균이 모두 같다는 가설을 기각한다.

NoteBase

문 32. 눈덩이표본추출방법(Snowball Sampling)이 필요한 경우로 가장 적합한 조사대상끼리 짝지어진 것은?

① 도서지역 주민, 교도소 수감자

② 마약중독자, 불법체류자

③ 불법체류자, 국민연금 수령자

④ 기초노령연금 수급자, 문맹자

정답 2번

간단한 해설

눈덩이표본추출은 **명단 확보와 직접 접촉이 어려운 집단을 참여자의 소개망으로 찾아가는 방법**이다.
마약중독자와 불법체류자는 상태 노출을 꺼리거나 통상적인 표본틀에서 누락될 수 있어 이 방법이 유용한 대표 사례이다.

NoteBase

선지별 해설

눈덩이표본추출은 처음 접촉한 조사 참여자가 다른 적격 참여자를 소개하고, 그 소개를 이어 가며 표본을 늘리는 방식이다. 비교 기준은 **조사대상의 명단을 확보하기 어렵고, 참여자 간 연결을 통해 접근할 필요가 큰가**이다. 일반적인 눈덩이표본은 비확률표본으로, 소개망에서 벗어난 사람이 빠지거나 특정 관계망에 치우칠 수 있다.

① 도서지역 주민·교도소 수감자: 도서지역은 이동이 어렵더라도 지역·가구를 기준으로 표본을 구성할 수 있고, 교도소 수감자는 기관이 관리하는 명부가 존재한다. 실제 조사에는 접근 허가 등이 필요하지만, 두 집단 모두 참여자의 소개 없이는 찾기 어려운 집단이라고 보기는 어렵다.

② 마약중독자·불법체류자: 두 집단은 조사대상이 자신의 상태를 드러내기 꺼리거나 통상적인 명단에서 누락될 수 있어 **완전한 표본틀 확보와 직접 접촉이 어렵다**. 이미 협조를 얻은 참여자의 소개망을 통해 추가 참여자를 찾는 방식이 유용하므로 두 대상 모두 눈덩이표본추출의 전형적인 적용 사례이다.

③ 불법체류자·국민연금 수령자: 앞의 대상은 소개망을 통한 접근이 유용할 수 있다. 반면 국민연금 수령자는 수급 관련 행정자료를 바탕으로 모집단을 파악할 수 있는 집단이므로, 두 대상이 모두 눈덩이표본추출에 특히 적합한 짝은 아니다.

④ 기초노령연금 수급자·문맹자: 연금 수급자는 행정상 관리되는 집단이다. 글을 읽고 쓰는 데 어려움이 있는 사람은 조사 방식의 조정이나 적격자 확인이 필요할 수 있지만, 문자 조사에 응하기 어렵다는 사실 자체가 소개망으로 표본을 뽑아야 한다는 뜻은 아니다. 두 집단 모두에 눈덩이표본추출의 필요성이 큰 조합은 아니다.

따라서 표본틀 확보와 접근의 어려움이라는 공통 기준에 가장 부합하는 짝은 ②이다.

NoteBase

문 33. 다음과 같은 확률함수를 갖는 분포에서 얻은 두 확률표본을 X_1, X_2라 하자. 이때 확률 $P(X_1 + X_2 \leq 1)$는?
f(x)=casesx, & 0≤ x≤ 1,\\2-x, & 1<x≤ 2.cases

① $\frac{1}{8}$
② $\frac{1}{11}$
③ $\frac{1}{17}$
④ $\frac{1}{24}$

정답 4번

간단한 해설

두 변수가 음이 아니므로 합이 1 이하인 영역에서는 둘 다 밀도의 첫 구간에 속한다. 독립성에 따른 결합밀도 xy를 **합이 1 이하인 삼각형 영역**에서 적분하면 $\frac{1}{24}$이다.

계산 과정

적분영역: 두 변수는 0부터 2 사이의 값을 갖는다. X_1+X_2≤1이려면 각각도 0과 1 사이여야 하며, $X_1 = x$일 때 $X_2 = y$의 범위는 0≤ y≤1-x이다. 따라서 필요한 영역은 **단위 정사각형 전체가 아니라 합이 1 이하인 삼각형**이다.

결합밀도: 두 변수는 확률표본이므로 독립이며 결합밀도는 각 밀도의 곱이다. 위 삼각형 영역에서는 두 변수 모두 밀도의 첫 구간에 속하므로 $f(x) = x$, $f(y) = y$를 사용한다. 따라서 이 영역의 결합밀도는 $f_{X_1, X_2}(x, y) = xy$이다. 두 번째 구간의 $2 - x$는 이번 적분영역에서 사용되지 않는다.

확률 계산: 삼각형 영역 위에서 결합밀도를 적분한다.
P(X_1+X_2≤1)= ∫ _0^1 ∫ _0^1-xxy dy dx
$= \int_0^1 \frac{x(1-x)^2}{2} dx$
$= \frac{1}{2}\int_0^1 (x - 2x^2 + x^3) dx$
$= \frac{1}{2}\left(\frac{1}{2} - \frac{2}{3} + \frac{1}{4}\right) = \frac{1}{24}$

따라서 구하는 확률은 $\frac{1}{24}$이며 ④에 해당한다.

NoteBase

문 34. 다음 중 변동계수에 대한 설명 중 틀린 것은?

① 변동계수는 평균을 표준편차로 나눈 값이다.

② 관측값의 산포 정도를 수치화하여 상대적으로 비교할 때 이용한다.

③ 평균의 차이가 큰 두 집단의 산포 정도를 비교할 때 이용한다.

④ 관측값의 단위가 서로 다른 두 집단의 산포 정도를 비교할 때 이용한다.

정답 1번

간단한 해설

변동계수는 **표준편차를 평균으로 나눈 값**으로, 평균에 비해 자료가 얼마나 흩어져 있는지를 나타낸다. 평균을 표준편차로 나눈다는 설명은 분자와 분모가 뒤바뀌었다.

선지별 해설

① X: 변동계수는 **표준편차를 평균으로 나눈 값**이다. 표본평균을 x, 표본표준편차를 s라고 하면 CV=(s)/(x)이며, 백분율로 나타내기도 한다. 선지는 분자와 분모를 뒤바꾸었다. 보통 평균이 양수이고 의미 있는 절대 영점이 있는 비율척도 자료에 사용하며, 평균이 0이면 정의되지 않는다.

② O: 변동계수는 표준편차가 평균에 비해 얼마나 큰지를 나타내는 **상대적 산포도**이다. 예를 들어 변동계수가 0.1이면 표준편차가 평균의 10%라는 의미이다. 표준편차 자체가 나타내는 절대적 흩어짐과 구별된다.

③ O: 평균 수준이 크게 다른 집단에서는 표준편차만 비교하면 평균에 대한 상대적 변동을 놓칠 수 있다. 각 집단의 표준편차를 해당 평균으로 나누면 **평균 규모를 기준으로 산포를 비교**할 수 있다. 다만 평균이 0에 가까우면 작은 평균 변화에도 변동계수가 크게 변하므로 해석에 주의한다.

④ O: 표준편차와 평균은 같은 단위를 가지므로 그 비인 변동계수에는 **측정 단위가 남지 않는다**. 따라서 서로 다른 단위로 측정한 집단의 상대적 산포를 비교할 때 사용할 수 있다. 다만 의미 있는 절대 영점이 있는 비율척도여야 하며, 섭씨·화씨처럼 기준점이 임의적인 척도에 무조건 적용하는 것은 아니다.

NoteBase

문 35. 회귀분석(Regression Analysis)과 상관분석(Correlation Analysis)에 대한 설명 중 틀린 것은?

① 회귀분석이란 독립변수의 특정한 값에 따른 종속변수의 값을 예측하는 기법이다.

② 상관분석이란 두 변수 사이의 밀접성(선형관계)의 강도(Strength)와 방향(Direction)을 나타내는 수치를 구하는 기법이다.

③ 상관분석만으로도 두 변수 사이에 관계가 있는지는 물론, 관계의 원인과 결과인 인과관계까지를 손쉽게 규명할 수 있다.

④ 회귀분석이나 상관분석을 할 때, 두 변수 사이의 관계를 개략적으로 알아보기 위해 산포도(Scatter Plot)를 그려서 활용한다.

정답 3번

간단한 해설

두 변수가 함께 변한다는 사실만으로 어느 변수가 다른 변수의 원인인지 정해지지 않는다. 제3의 변수나 반대 방향의 영향도 가능하므로 **상관분석만으로 인과관계를 규명할 수 없다**.

선지별 해설

회귀분석은 설명변수에 따른 반응변수의 관계를 모형화하여 예측에 이용하고, 상관분석은 두 변수의 연관성 정도를 수치로 요약한다. 이 문항의 상관분석은 선형관계를 나타내는 피어슨 상관분석의 맥락이다.

① O: 회귀분석에서는 독립변수와 종속변수의 역할을 정하고 적합한 회귀식에 독립변수의 값을 대입하여 종속변수의 값을 예측한다. 통상적인 평균 회귀식은 주어진 독립변수 값에서의 **조건부 평균**을 추정하므로, 개별 관측값에는 예측값과 다른 오차가 있을 수 있다.

② O: 피어슨 상관계수 r은 두 변수 사이 선형관계의 방향과 강도를 나타낸다. 부호는 같은 방향 또는 반대 방향의 관계를, 절댓값은 선형관계의 강도를 보여 준다. $-1 \le r \le 1$이며, $r = 0$이라고 해서 비선형관계까지 모두 없다는 뜻은 아니다.

③ X: 상관관계가 관찰되어도 X가 Y에 영향을 주는지, Y가 X에 영향을 주는지, 제3의 변수가 둘 모두에 영향을 주는지 상관계수만으로 구별할 수 없다. 인과관계를 판단하려면 연구 설계, 시간적 선후관계, 교란요인의 통제 및 식별 가정 등 추가 근거가 필요하다. **강한 상관도 인과관계의 충분조건은 아니다**.

④ O: 산포도는 두 변수의 관측값 쌍을 점으로 나타낸 그림이다. 관계의 방향과 선형·곡선 형태, 이상값, 군집 등을 눈으로 확인하여 회귀모형의 선택이나 상관계수의 해석에 활용한다. 하나의 상관계수가 가릴 수 있는 자료의 구조를 먼저 살펴보는 데 유용하다.

NoteBase

문 36. 김 대리가 받는 이메일의 20%에는 '행운'이란 단어가 포함되어 있다. '행운'이란 단어가 포함된 이메일의 60%는 광고이고, '행운'이란 단어가 포함되지 않은 이메일의 30%도 광고이다. 김 대리가 받은 한 이메일이 광고일 때, 이 이메일에 '행운'이란 단어가 포함될 확률은?

① $\frac{1}{5}$
② $\frac{3}{5}$
③ $\frac{24}{100}$
④ $\frac{1}{3}$

정답 4번

간단한 해설

'행운'이 포함된 광고는 전체의 12%, 광고 전체는 36%이다. **광고 전체를 기준**으로 비율을 구하면 (0.12)/(0.36)=13이다.

계산 과정

사건과 조건: '행운'이 포함된 사건을 W, 광고인 사건을 A라고 하자. 주어진 확률은 $P(W) = 0.2$, $P(A \mid W) = 0.6$, $P(A \mid W^c) = 0.3$이며, 단어가 포함되지 않을 확률은 $P(W^c) = 0.8$이다. 구하려는 것은 **광고라는 조건에서 단어가 포함될 확률**인 $P(W \mid A)$로, 주어진 $P(A \mid W)$와 조건의 방향이 반대이다.

각 광고의 비율: '행운'이 포함된 광고의 비율은 $P(W \cap A) = P(W)P(A \mid W) = 0.2 \times 0.6 = 0.12$이다. '행운'이 포함되지 않은 광고의 비율은 $P(W^c \cap A) = 0.8 \times 0.3 = 0.24$이다. 두 값 모두 전체 이메일을 기준으로 한 비율이다.

전체 광고의 비율: 광고는 '행운'이 포함된 경우와 포함되지 않은 경우로 나뉘며 두 경우는 겹치지 않는다. 따라서 전확률법칙으로 $P(A) = 0.12 + 0.24 = 0.36$을 얻는다. 조건이 광고이므로 이 **전체 광고의 비율이 분모**가 된다.

조건부확률 계산: P(W A)=(P(W A))/(P(A))=(0.12)/(0.36)=13이다. 즉 광고 이메일 중 '행운'이 포함된 비율은 1/3이며, 선택지 ④에 해당한다.

NoteBase

문 37. 어느 회사에서 현재의 직위가 회사가 마련한 정년연장안에 대한 의견에 영향을 미치는지 알아보기 위해 전체 직원 1,000명을 대상으로 설문조사를 실시하였다. 조사 결과의 일부가 표에 주어졌다. 현재의 직위와 정년연장안에 대한 의견이 서로 독립이라고 가정할 때, a에 들어갈 기대도수(Expected Frequency)는 얼마인가?

	정년연장 찬성	정년연장 반대	합계
대리 이하	a		700
과장 이상			300
합 계	350	650	1,000

① 455

② 245

③ 105

④ 195

정답 2번

간단한 해설

직위와 의견이 독립이면 대리 이하 집단의 찬성 비율도 전체 찬성 비율과 같아진다. **행 합계와 열 합계를 전체 인원으로 연결**하면 기대도수는 $\frac{700 \times 350}{1,000} = 245$명이다.

NoteBase

계산 과정

표에서 구하려는 셀은 **'대리 이하'이면서 '정년연장 찬성'**인 경우이다. 해당 행 합계는 700명, 해당 열 합계는 350명, 전체 인원은 1,000명이다.

직위와 의견이 독립이라면 직위별 찬성 비율은 전체 찬성 비율과 같아야 한다. 전체 찬성 비율은 $\frac{350}{1,000} = 0.35$이므로 대리 이하 700명에게도 이 비율을 적용한다.

독립을 가정한 교차표의 기대도수는 **행 합계와 열 합계의 곱을 전체 합계로 나눈 값**이다.
$E_{ij} = \frac{R_i C_j}{N}$
여기서 R_i는 해당 행 합계, C_j는 해당 열 합계, N은 전체 합계이다. 따라서 $a = \frac{700 \times 350}{1,000} = 245$이다.

이는 독립이라는 가정 아래 예상되는 도수이며, 실제로 조사된 해당 셀의 관측도수를 확정한 것은 아니다. 구한 기대도수는 245명으로 ②와 일치한다.

NoteBase

문 38. $X_1, X_2, \cdots, X_n$**이 정규분포** $N(\mu, \sigma^2)$**의 확률표본일 때 모평균** μ**에 대한 95% 신뢰구간을 구하고자 한다. 모집단의 표준편차** σ**를 알고 있는 경우** $\sqrt{n}(\overline{X}-\mu)/\sigma$**는 정규분포를 따르며, 모집단의 표준편차** σ**를 모르는 경우** $\sqrt{n}(\overline{X}-\mu)/S$**는 자유도** $(n-1)$**인** $t(n-1)$**분포를 따르고 표본크기** $n=5$**인 경우** $E(S)=0.94\sigma$**가 된다.** σ**를 모르는 경우와 비교하여** σ**를 아는 경우 정보가 많아 신뢰구간의 크기가 감소할 것으로 예상된다. 이때 신뢰구간의 크기가 평균적으로 얼마나 감소하는가? (단,** $z_{0.025}=1.96$**,** $t_{0.025}(4)=2.78$**)**

① 약 13%

② 약 20%

③ 약 25%

④ 약 33%

정답 3번

간단한 해설

σ를 모를 때는 구간 길이에 포함된 S를 **기대값 0.94σ**로 바꾸어 평균 길이를 비교한다. σ를 아는 경우의 상대 길이는 $\frac{1.96}{2.78\times 0.94}\approx 0.75$이므로 약 25% 감소한다.

계산 과정

σ를 아는 경우: 95% 신뢰구간은 $\overline{X}\pm 1.96\frac{\sigma}{\sqrt{5}}$이다. 구간 길이는 상한과 하한의 차이이므로 오차한계의 두 배이다.

$L_z = 2\times 1.96\frac{\sigma}{\sqrt{5}}$

표본평균에 따라 구간의 위치는 달라지지만, σ와 표본 크기가 정해져 있으므로 이 길이는 일정하다.

σ를 모르는 경우: 자유도가 5 − 1 = 4이므로 구간 길이는 $L_t = 2\times 2.78\frac{S}{\sqrt{5}}$이다. 이 길이는 표본표준편차 S에 따라 달라지므로 **주어진 S의 기대값**을 이용해 평균 길이를 구한다.

$E(L_t) = 2\times 2.78\frac{E(S)}{\sqrt{5}}$

$= 2\times 2.78\times 0.94\frac{\sigma}{\sqrt{5}}$

감소율: 비교 기준은 **σ를 모를 때의 평균 구간 길이**이다. 두 길이에 공통인 $\frac{2\sigma}{\sqrt{5}}$를 약분하면

$\frac{L_z}{E(L_t)} = \frac{1.96}{2.78\times 0.94}\approx 0.7500$

따라서 σ를 아는 경우의 길이는 기준 길이의 약 75%이고, 감소율은

$\frac{E(L_t)-L_z}{E(L_t)}\times 100\approx 25\%$

이다.

따라서 평균 구간 길이는 약 25% 감소하므로 ③이다.

NoteBase

문 39. 모평균 μ에 대한 신뢰구간의 크기를 1/4로 줄이기 위해서는 원래의 표본 크기(n)를 어떻게 변화시켜야 하는가?

① 4배 증가

② 4배 감소

③ 4^2배 증가

④ 4^2배 감소

정답 3번

간단한 해설

신뢰수준과 변동성이 같을 때 신뢰구간의 길이는 **표본크기의 제곱근에 반비례**한다. 길이를 1/4로 줄이려면 표본크기를 $4^2 = 16$배로 늘려야 한다.

계산 과정

이 문항은 **신뢰수준과 모집단의 변동성을 일정하게 유지**한 채 표본크기만 바꾸는 통상적인 신뢰구간 설계 관계를 묻는다. 모표준편차가 알려진 정규모형에서 모평균 신뢰구간의 전체 길이는 L(n)=2z_1-α/2(σ)/(n)이다. 여기서 z는 고정된 신뢰수준에 따른 임계값이고, σ는 모표준편차이다.

나머지 조건이 같으면 구간 길이는 **표본크기의 제곱근에 반비례**한다. 새 표본크기를 m이라고 하면 (L(m))/(L(n))=√‾(nm)이다. 전체 길이 대신 오차한계인 반길이를 사용해도 비율은 같다.

길이를 원래의 1/4로 줄이려면 √‾(nm)=14이어야 한다. 양변을 제곱하면 nm=116이므로 $m = 16n$이다.

따라서 표본크기를 원래의 $4^2 = 16$배로 늘려야 하므로 ③에 해당한다. 표본크기를 4배로 늘리면 길이는 1/2로만 줄어든다. 소표본 t구간에서는 표본크기에 따라 임계값도 달라지므로, 이 제곱근 관계를 정확한 고정 비율로 적용하는 조건과 구별한다.

NoteBase

문 40. 다음 A, B 두 회사는 조사를 실시하여 조사결과를 향후 정책수립에 반영하고자 한다. 다음과 같은 조사 목적에 적합한 귀무가설(H_0)과 대립가설(H_1)은?

	현황과 목표
A 회사	현재 월평균 냉장고 판매량은 500대인데, 새로운 판매보너스 지급정책이 월평균 판매량을 증가시킬 수 있는지 조사하고자 함.
B 회사	생산제품의 용기에 담긴 내용물의 중량이 750g에 미달하거나 초과하는 경우 내용물을 담는 기계를 재조정하려고 조사하고자 함.

① 회사: A 회사, $H_0: \mu > 500$, $H_1: \mu \leq 500$
회사: B 회사, $H_0: \mu \neq 750$, $H_1: \mu = 750$

② 회사: A 회사, $H_0: \mu = 500$, $H_1: \mu > 500$
회사: B 회사, $H_0: \mu = 750$, $H_1: \mu \neq 750$

③ 회사: A 회사, $H_0: \mu = 500$, $H_1: \mu \neq 500$
회사: B 회사, $H_0: \mu > 750$, $H_1: \mu \leq 750$

④ 회사: A 회사, $H_0: \mu \neq 500$, $H_1: \mu = 500$
회사: B 회사, $H_0: \mu \leq 750$, $H_1: \mu > 750$

정답 2번

간단한 해설

A 회사는 판매량의 **증가**를 확인하므로 우측 단측검정을 한다. B 회사는 중량의 **미달과 초과 모두**를 확인하므로 양측검정을 한다.

NoteBase

핵심 풀이

가설의 기준: 조사로 확인하려는 변화의 방향을 대립가설에 담고, 귀무가설에는 비교 기준이 되는 모평균을 둔다. 여기서 μ는 A 회사에서는 새 정책 아래 월별 판매량의 모평균, B 회사에서는 내용물 중량의 모평균을 뜻한다.

A 회사: 현재 월평균 판매량은 500대이고 새 판매보너스 정책이 이를 증가시키는지가 관심이다. 따라서 **500대보다 큰 방향**을 검정하여 $H_1 : \mu > 500$으로 둔다. 선택지의 귀무가설은 기준값을 나타내는 $H_0 : \mu = 500$이다. 우측 단측검정의 귀무가설을 μ≤500으로 표현할 때에도 검정의 기준 경계는 500이다. 감소까지 포함하는 양측 대립가설은 조사 목적과 다르다.

B 회사: 목표 중량은 750g이며, 이보다 미달해도 초과해도 기계를 재조정하려 한다. 선지의 모평균 검정에서는 $H_0 : \mu = 750$, H_1:μ"2260750으로 두어 **양쪽 방향의 이탈**을 함께 살핀다. 초과만을 검정하는 우측 단측검정으로는 미달이라는 관심 방향을 포함하지 못한다. 이 모평균 검정은 모든 개별 제품의 중량이 정확히 750g임을 보장하는 검정은 아니다.

선택지 연결: A 회사의 대립가설은 $\mu > 500$, B 회사의 대립가설은 μ"2260750이고 각각의 귀무가설은 기준값과의 등식이다. 이 두 설정을 모두 제시한 선택지는 ②이다.

NoteBase