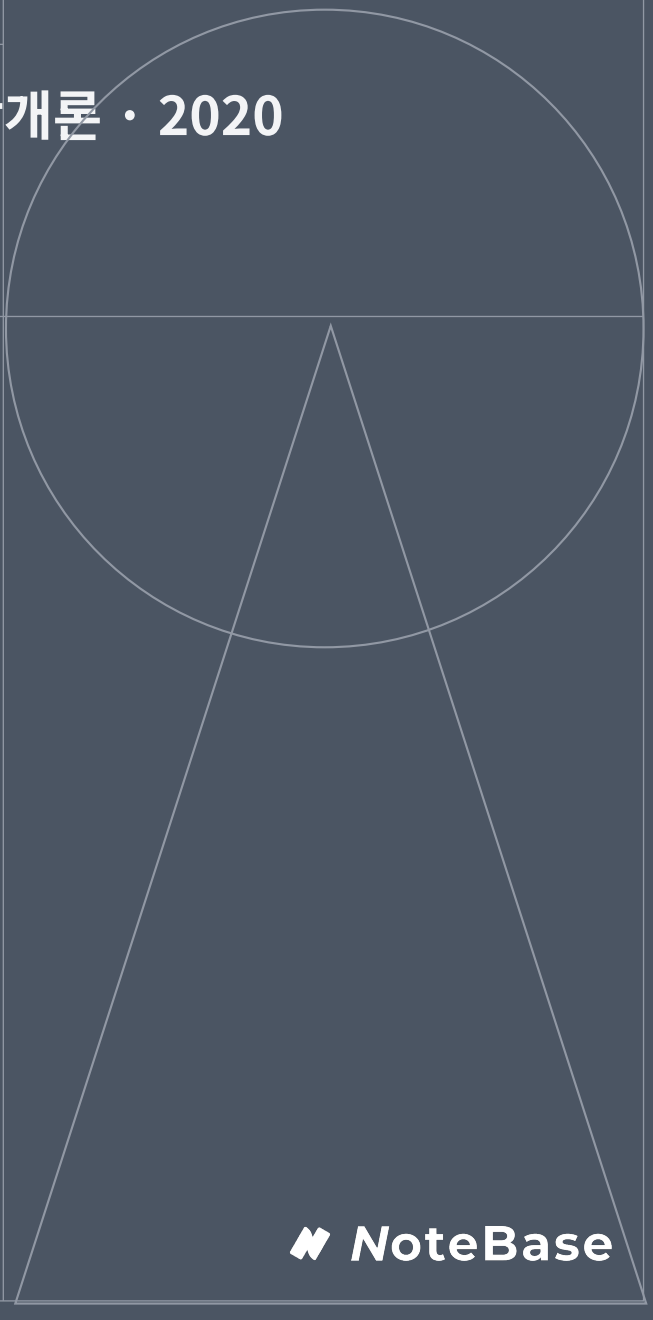
기출OX 해설집

공무원/국가직/9급 · 통계학개론 · 2020



NoteBase

기출OX 해설집

공무원/국가직/9급 · 통계학개론 · 2020

공무원/국가직/9급 · 통계학개론 · 2020

문 1. 다음 자료는 어느 고등학교 A, B반의 학생이 하루에 섭취하는 물의 양을 측정한 것이다. A반은 리터(L) 단위로 측정하고, B반은 밀리리터(mL) 단위로 측정하였다. A, B반 학생이 섭취하는 물의 양에 대한 산포를 비교하고자 할 때 가장 적합한 측도는?

A반(L)	1.0	1.2	0.5	0.8	2.0	1.5	1.4	1.2	1.0	0.9
B반(mL)	500	700	450	600	1,000	300	500	500	1,500	400

① 표준편차

② 사분위수 범위

③ 제50백분위수

④ 변동계수

정답 4번

간단한 해설

A반은 L, B반은 mL로 측정했으므로 단위의 영향을 받지 않는 산포 척도가 필요하다. **변동계수**는 표준편차를 평균으로 나눈 값으로, **평균에 비해 얼마나 흩어져 있는지**를 단위 없이 비교한다.

NoteBase

핵심 풀이

자료의 첫째 행은 A반의 하루 물 섭취량을 L로, 둘째 행은 B반의 섭취량을 mL로 나타낸다. 1 L = 1,000 mL이므로 두 행의 숫자 크기를 그대로 비교하면 측정 단위의 차이가 섞인다.

변동계수는 평균이 양수인 자료에서 $CV = \frac{s}{\bar{x}}$로 구한다. 여기서 s는 표준편차, $\bar{x}$는 평균이다. 물의 양은 0이 실제로 양이 없음을 뜻하므로 평균에 대한 상대적 산포를 해석할 수 있다. L를 mL로 바꾸면 평균과 표준편차가 모두 1,000배가 되어 **두 값의 비율은 변하지 않는다**. 따라서 변동계수가 클수록 평균에 비해 섭취량이 더 크게 흩어져 있다고 해석하며, 두 반의 상대적 산포를 비교하는 데 적합하다.

표준편차와 사분위수 범위도 산포 척도이지만 **원자료의 측정 단위를 갖는다**. 사분위수 범위는 제3사분위수에서 제1사분위수를 뺀 값으로, 이 두 척도를 직접 비교하려면 먼저 단위를 통일해야 한다. 제50백분위수는 중앙값으로, 산포가 아니라 자료의 중심 위치를 나타낸다.

NoteBase

문 2. 두 사건 A, B에 대하여 $P(A)=\frac{1}{3}$, $P(A \mid B)=\frac{1}{9}$, $P(A \cap B)=\frac{1}{12}$일 때, 옳은 것만을 모두 고르면?

ㄱ. 사건 A와 B는 서로 배반이다.
ㄴ. 사건 A와 B는 서로 독립이다.
ㄷ. $P(A \cup B)=1$이다.

① ㄴ
② ㄷ
③ ㄱ, ㄴ
④ ㄴ, ㄷ

정답 2번

간단한 해설

조건부확률 식에서 $P(B)=\frac{3}{4}$를 얻는다. 두 사건의 확률을 더한 뒤 **교집합의 확률을 한 번 빼면** $P(A \cup B)=1$이 된다.

보기별 해설

ㄱ. X: 두 사건이 배반이면 동시에 일어날 수 없으므로 $A \cap B=\varnothing$이고 교집합의 확률은 0이다. 그런데 $P(A \cap B)=\frac{1}{12}>0$이므로 두 사건은 **배반이 아니다**.

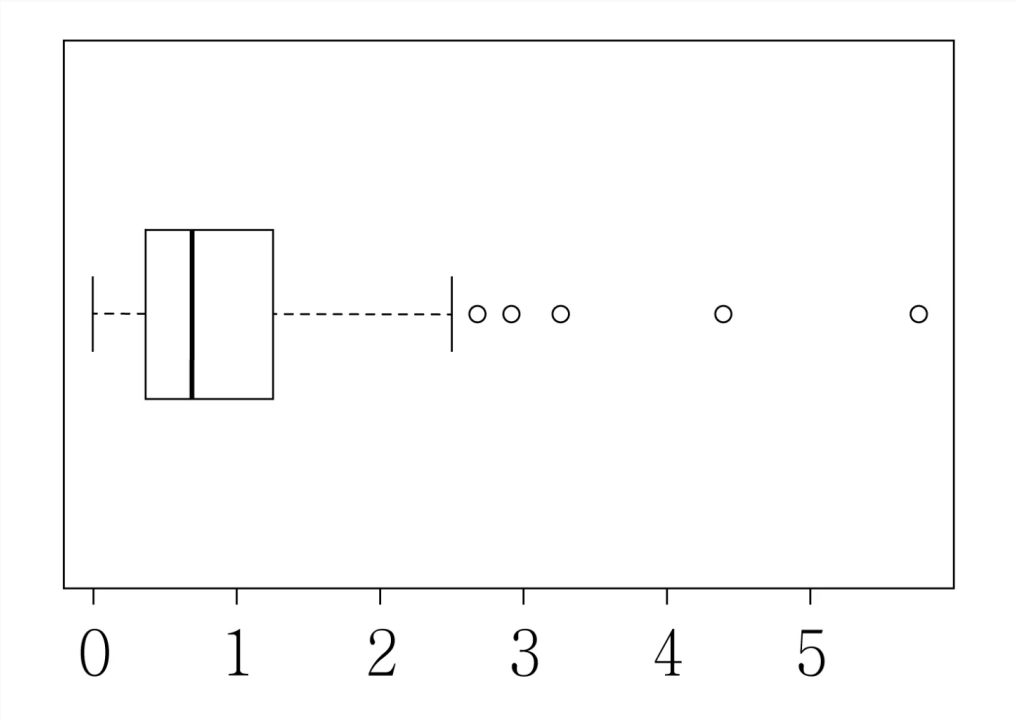
ㄴ. X: 두 사건이 독립이면 B의 발생 여부가 A의 확률을 바꾸지 않으므로, $P(B)>0$일 때 $P(A \mid B)=P(A)$가 성립해야 한다. 주어진 조건에서는 $P(A \mid B)=\frac{1}{9}$, $P(A)=\frac{1}{3}$로 서로 다르므로 두 사건은 **독립이 아니다**.

ㄷ. O: 조건부확률의 정의에 따라 $P(A \cap B)=P(A \mid B)P(B)$이다. 주어진 값을 넣으면 $\frac{1}{12}=\frac{1}{9}P(B)$이므로 $P(B)=\frac{3}{4}$이다. 합집합의 확률을 구할 때는 두 사건의 확률을 더하면서 중복으로 센 **교집합의 확률을 한 번 빼야 한다**. 따라서 $P(A \cup B)=\frac{1}{3}+\frac{3}{4}-\frac{1}{12}=1$이다.

옳은 보기는 ㄷ만이므로 ②에 해당한다.

NoteBase

문 3. 다음 상자그림에 대한 설명으로 옳지 않은 것은?



① 자료의 분포는 오른쪽으로 긴 꼬리를 가지는 비대칭 형태이다.

② 자료의 평균이 중앙값보다 작다.

③ 자료의 중앙값은 1보다 작다.

④ 자료의 왜도(skewness)는 정규분포의 왜도보다 크다.

정답 2번

간단한 해설

오른쪽 수염이 길고 큰 값 쪽에 이상값들이 있어 오른쪽 꼬리가 긴 분포로 해석한다. 이런 분포에서는 큰 값들이 평균을 끌어올리므로, 일반적으로 **평균이 중앙값보다 크다**.

NoteBase

선지별 해설

① O: 오른쪽 수염이 왼쪽 수염보다 길고, 오른쪽 수염 밖에도 여러 이상값이 있다. 큰 값 쪽으로 자료가 멀리 뻗어 있으므로 **오른쪽 꼬리가 긴 비대칭 분포**로 해석한다.

② X: 평균은 모든 관측값의 크기를 반영하므로 오른쪽의 큰 값들에 의해 커진다. 반면 중앙값은 자료를 크기순으로 나열했을 때 가운데 위치의 값이어서 극단적인 값의 영향을 상대적으로 덜 받는다. 따라서 이와 같은 오른쪽 꼬리가 긴 분포에서는 일반적으로 **평균이 중앙값보다 크다**고 해석하며, 선지의 '작다'는 관계가 반대이다.

③ O: 상자 안의 굵은 세로선이 **중앙값**을 나타낸다. 이 선은 가로축의 0과 1 사이에 있으므로 중앙값은 1보다 작다.

④ O: 왜도는 분포가 어느 방향으로 비대칭인지를 나타내는 지표이다. 그림처럼 오른쪽 꼬리가 긴 분포는 **양의 왜도**를 갖는 것으로 해석한다. 정규분포는 좌우 대칭이므로 왜도가 0이어서, 그림의 분포는 정규분포보다 왜도가 크다.

NoteBase

문 4. 두 확률변수의 상관계수에 대한 설명으로 옳지 않은 것은?

① 상관계수는 항상 −1과 1 사이의 값이다.

② 공분산이 0이면 상관계수도 0이다.

③ 상관계수의 절댓값이 1에 가까울수록 두 확률변수는 강한 양의 상관관계가 있다.

④ 상관계수는 두 확률변수의 선형 관계를 나타내는 측도이다.

정답 3번

간단한 해설

상관계수의 **절댓값은 선형 관계의 강도**를, **부호는 관계의 방향**을 나타낸다. 절댓값이 1에 가까워도 상관계수가 음수이면 강한 음의 상관관계이다.

선지별 해설

두 확률변수 X, Y의 분산이 유한하고 0보다 클 때, 상관계수는 공분산을 두 표준편차의 곱으로 나눈 값이다. 즉, $\rho = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$이며, 여기서 σ_X, σ_Y는 각각 X, Y의 표준편차이다.

① O: 공분산의 절댓값은 두 표준편차의 곱을 넘을 수 없으므로, 이를 나눈 상관계수는 **−1 이상 1 이하**이다. 양 끝값인 −1과 1도 포함한다.

② O: 상관계수가 정의되는 조건에서는 분모인 $\sigma_X \sigma_Y$가 양수이다. 따라서 **공분산이 0이면** 상관계수도 0이다.

③ X: 절댓값이 1에 가까울수록 강해지는 것은 **선형 관계의 강도**이다. 양의 관계인지 음의 관계인지는 **상관계수의 부호**로 구분한다. 예를 들어 상관계수가 −0.9이면 절댓값은 1에 가깝지만 강한 음의 상관관계이다. 따라서 '강한 양의 상관관계'는 '강한 양 또는 음의 상관관계'로 고쳐야 한다.

④ O: 상관계수는 두 변수가 **직선적인 관계를 이루는 방향과 강도**를 나타낸다. 따라서 상관계수가 0이라는 것은 선형 관계가 없다는 뜻이며, 비선형 관계까지 없다는 뜻은 아니다.

NoteBase

문 5. 어떤 전자부품의 불량률은 0.1이다. 전체 생산된 부품에서 임의로 추출된 100개의 부품 가운데 불량품의 개수를 확률변수 X라 할 때, 옳지 않은 것은?

① $P(X=10) > P(X=90)$

② $P(X \geq 1) = 1$

③ X의 기댓값은 10이다.

④ X의 표준편차는 3이다.

정답 2번

간단한 해설

불량품이 하나도 없을 확률이 $P(X=0) = 0.9^{100} > 0$이므로, **적어도 하나가 불량품일 확률은 1보다 작다**. 여사건을 이용하면 $P(X \geq 1) = 1 - 0.9^{100}$이다.

선지별 해설

각 부품의 불량 여부가 서로 독립이고 불량확률이 일정하다는 이항모형을 적용한다. 불량품 개수 X는 시행 횟수 $n=100$, 불량확률 $p=0.1$인 이항분포 $B(100, 0.1)$을 따르며, 한 부품이 양품일 확률은 0.9이다.

① O: 불량품이 정확히 10개일 때와 90개일 때 해당 부품을 고르는 경우의 수는 $\binom{100}{10} = \binom{100}{90}$으로 같다. 이항확률의 비를 구하면 $\frac{P(X=10)}{P(X=90)} = 9^{80} > 1$이다. 경우의 수는 같지만 **불량보다 양품일 확률이 더 크므로**, 불량품 10개와 양품 90개가 나올 확률이 그 반대보다 크다.

② X: 적어도 하나가 불량품인 사건의 **여사건은 100개가 모두 양품인 사건**이다. 모두 양품일 확률은 $P(X=0) = 0.9^{100} > 0$이므로, $P(X \geq 1) = 1 - 0.9^{100} < 1$이다. 따라서 선지의 확률 1은 $1 - 0.9^{100}$으로 고쳐야 한다.

③ O: 이항분포의 기댓값은 $E(X) = np$이다. 각 부품이 불량품 개수에 기여하는 평균이 0.1이므로, 100개의 기여를 합하면 $E(X) = 100 \cdot 0.1 = 10$이다. 즉 100개를 추출하는 일을 반복할 때 **불량품 개수의 장기적인 평균은 10개**이다.

④ O: 이항분포의 분산은 $V(X) = np(1-p)$이므로, $100 \times 0.1 \times 0.9 = 9$이다. **표준편차는 분산의 양의 제곱근**이므로 $\sqrt{9} = 3$이며, 불량품 개수의 표준편차는 3개이다.

NoteBase

문 6. 두 확률변수 X, Y에 대한 설명으로 옳지 않은 것은?

① $X + Y$의 기댓값 $E(X + Y) = E(X) + E(Y)$이다.

② $E(XY) = E(X)E(Y)$이면 확률변수 X와 Y는 서로 독립이다.

③ 확률변수 X와 Y가 서로 독립이면 두 확률변수의 차 $X - Y$의 분산 $Var(X - Y) = Var(X) + Var(Y)$이다.

④ 확률변수 X와 Y가 서로 독립이면 두 확률변수의 공분산 $Cov(X, Y) = 0$이다.

정답 2번

간단한 해설

$E(XY) = E(X)E(Y)$는 공분산이 0인 **무상관**을 뜻하며, 이것만으로 **독립이라고 할 수는 없다**. 독립이면 이 등식이 성립하지만 그 역은 일반적으로 성립하지 않는다.

선지별 해설

이하에서는 기댓값·분산·공분산이 존재하는 경우를 전제로 한다.

① O: 기댓값의 선형성에 따라 $E(X + Y) = E(X) + E(Y)$이다. 합의 기댓값은 각 기댓값의 합이 되며, 이 성질에는 **독립이라는 조건이 필요하지 않다**.

② X: $E(XY) = E(X)E(Y)$이면 공분산이 0이므로 무상관이지만, **무상관이 독립을 보장하지는 않는다**. 독립은 한 변수의 값에 관한 정보가 다른 변수의 확률분포를 바꾸지 않는 성질로, 무상관보다 강한 조건이다. 예를 들어 X가 $-1, 0, 1$을 각각 1/3의 확률로 취하고 $Y = X^2$이라고 하자. 대칭성으로 $E(X) = E(XY) = 0$이므로 $E(XY) = E(X)E(Y)$가 성립한다. 그러나 전체에서 $Y = 0$일 확률은 1/3인데, $X = 0$임을 알면 $Y = 0$이 확실해지므로 두 변수는 독립이 아니다. 선지는 '서로 독립이면 $E(XY) = E(X)E(Y)$이다'로 고쳐야 한다.

③ O: 일반적으로 두 확률변수 차의 분산은 두 분산의 합에서 공분산의 2배를 뺀 값이다. **독립이면 공분산이 0**이므로 $\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$이다. 뺄셈이어도 분산은 편차를 제곱하여 측정하므로 두 분산을 빼는 것이 아니다.

④ O: 공분산은 $\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y)$이다. **독립이면 곱의 기댓값이 기댓값의 곱**이 되므로 오른쪽의 두 항이 같아져 공분산은 0이 된다.

NoteBase

문 7. 모집단에서 크기가 16인 표본을 임의로 추출하여 구한 모평균의 95 % 신뢰구간이 (20, 25)이다. 다음 설명 중 옳은 것만을 모두 고르면?

ㄱ. 모평균이 구간 (20, 25) 안에 있을 확률은 0.95이다.
ㄴ. 표본의 95 %가 구간 (20, 25) 안에 있다.
ㄷ. 표본평균이 구간 (20, 25) 안에 있다.

① ㄷ
② ㄱ, ㄴ
③ ㄱ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 1번

간단한 해설

통상적인 모평균의 양측 신뢰구간은 **표본평균 ± 오차한계**로 구성되므로 표본평균은 그 구간 안에 있다. 95 %는 표본값의 포함 비율이 아니라 **구간 추정 방법의 신뢰수준**이다.

보기별 해설

ㄱ. X: 빈도주의에서 모평균은 **고정된 값**이다. 표본을 관측하여 구간 (20, 25)가 정해진 뒤에는 그 구간이 모평균을 포함하거나 포함하지 않으므로, 포함할 확률이 0.95라고 해석하지 않는다. 95 % 신뢰수준은 같은 모집단에서 크기 16인 표본을 반복 추출하여 같은 방법으로 구간을 만들 때, 장기적으로 그 구간들의 약 95 %가 모평균을 포함한다는 뜻이다.

ㄴ. X: 모평균의 신뢰구간은 **개별 관측값이 아니라 모평균을 추정하는 구간**이다. 따라서 16개 표본값 중 95 %가 (20, 25)에 들어 있다는 의미가 아니다. 이 신뢰구간만으로는 표본값 중 몇 개가 해당 구간에 들어 있는지 알 수 없다.

ㄷ. O: 통상적인 모평균의 양측 신뢰구간은 **표본평균을 중심으로 양쪽에 오차한계를 더하고 빼서** 구한다. 따라서 표본평균은 구간의 중점인 $\bar{x} = \frac{20 + 25}{2} = 22.5$이고, 22.5는 구간 (20, 25) 안에 있다.

옳은 보기는 ㄷ만이므로, 이에 해당하는 선택지는 ①이다.

NoteBase

문 8. 평균이 1이고, 분산이 4인 정규모집단에서 추출한 확률표본 $X_1, \cdots, X_9$의 표본평균이 $\overline{X}$이다. 실수 $\alpha(0 < \alpha < \frac{1}{2})$에 대해 $P(\overline{X} \leq 1 - 2c) = \alpha$를 만족시키는 c의 값은? (단, z_α는 표준정규분포의 제 $100 \times (1 - \alpha)$ 백분위수이다)

① z_α

② $z_{1-\alpha}$

③ $\frac{1}{3}z_\alpha$

④ $\frac{1}{3}z_{1-\alpha}$

정답 3번

간단한 해설

표본평균의 **표준오차**는 $\frac{2}{3}$이므로 주어진 경계를 표준화하면 $-3c$가 된다. z_α는 오른쪽 꼬리확률이 α인 값이므로, **왼쪽 누적확률**이 α인 값은 $-z_\alpha$이며 $c = \frac{1}{3}z_\alpha$이다.

계산 과정

1. 표본평균의 분포: 확률표본 $X_1, \ldots, X_9$은 서로 독립이고 평균 1, 분산 4인 같은 정규분포를 따른다. 따라서 표본평균 $\overline{X}$도 정규분포를 따르며, 평균은 1, 분산은 $\frac{4}{9}$이다. 표본평균의 표준편차인 **표준오차**는 $\sqrt{\frac{4}{9}} = \frac{2}{3}$이다.

2. 확률조건의 표준화: 표본평균에서 평균 1을 빼고 표준오차로 나누면 $Z = \frac{\overline{X}-1}{2/3}$은 표준정규분포를 따른다. 경계값 $1 - 2c$는 $\frac{(1-2c)-1}{2/3} = -3c$로 바뀌므로, 주어진 조건은 $P(Z \leq -3c) = \alpha$이다.

3. 분위수의 부호: 문제에서 z_α는 누적확률이 $1 - \alpha$인 값이므로 $P(Z \leq z_\alpha) = 1 - \alpha$이다. 표준정규분포는 **0을 중심으로 대칭**이어서 $P(Z \leq -z_\alpha) = \alpha$이다. 또한 $0 < \alpha < \frac{1}{2}$이므로 $z_\alpha > 0$이고, 왼쪽 누적확률이 α인 경계는 음수이다.

4. c의 계산: 표준정규분포의 누적분포함수는 엄격히 증가하므로, 누적확률이 같은 두 경계는 일치한다. 따라서 $-3c = -z_\alpha$이고, 양변을 −3으로 나누면 $c = \frac{1}{3}z_\alpha$이다. 이는 ③의 식과 일치한다.

NoteBase

문 9. 어느 지역의 2019년 쌀 생산량을 추정하기 위해 이 지역 9개의 농지를 임의로 추출하여 단위면적당 쌀 생산량을 조사하였더니 표본평균 1,150 kg, 표본표준편차 150 kg이다. 이 지역의 2019년 단위면적당 평균 쌀 생산량이 2018년의 단위면적당 평균 쌀 생산량 1,100 kg보다 크다는 주장을 검정하고자 한다. 귀무가설이 참일 때, t 검정통계량의 값과 자유도 k를 옳게 짝 지은 것은? (단, 단위면적당 쌀 생산량은 정규분포를 따른다)

① t 검정통계량의 값: 1, k: 8

② t 검정통계량의 값: 1, k: 9

③ t 검정통계량의 값: $\frac{1}{3}$, k: 8

④ t 검정통계량의 값: $\frac{1}{3}$, k: 9

정답 1번

간단한 해설

표본평균과 검정 기준인 1,100 kg의 차이 50 kg을 **표본평균의 추정 표준오차** 50 kg으로 나누면 t 검정통계량은 1이다. 모표준편차 대신 표본표준편차를 사용하므로 자유도는 $n-1=8$이다.

계산 과정

검정 설정: 2019년 단위면적당 평균 쌀 생산량을 μ라 하면, 검정할 주장은 $\mu > 1{,}100$이다. 귀무가설의 경계값 $\mu_0 = 1{,}100$ kg을 기준으로 계산한다. 주어진 값은 표본크기 n = 9, 표본평균 $\bar{x} = 1{,}150$ kg, 표본표준편차 s = 150 kg이다. 정규모집단에서 임의추출했고 **모표준편차를 알 수 없어 표본표준편차로 추정**하므로 일표본 t 검정을 사용한다.

표준오차: 표본평균의 변동 크기는 개별 관측값의 변동 크기를 표본크기의 제곱근으로 나누어 구한다. 따라서 추정 표준오차는 $\frac{s}{\sqrt{n}} = \frac{150}{\sqrt{9}} = 50$ kg이다. 이는 개별 농지 생산량의 표준편차 150 kg과 구별되는 **표본평균의 변동 크기**이다.

t 검정통계량: 표본평균이 귀무가설의 기준값에서 표준오차의 몇 배만큼 떨어져 있는지 계산한다. $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{1{,}150-1{,}100}{50} = 1$이다. 표본평균이 기준값보다 표준오차 1배만큼 크다는 뜻이며, 분자와 분모의 kg 단위가 소거되어 t 값에는 단위가 없다.

자유도: 표본분산은 표본평균으로부터의 편차를 이용한다. 편차의 합이 0이라는 제약 때문에 9개 편차 중 독립적으로 정할 수 있는 것은 8개이므로, 귀무가설의 경계에서 검정통계량은 **자유도 n − 1의 t 분포**를 따른다. 따라서 $k = n-1 = 9-1 = 8$이다.

t 검정통계량 1과 자유도 8의 짝은 ①과 일치한다.

NoteBase

문 10. 다음 중 서로 독립인 두 확률변수 X_1, X_2에 대해 옳은 것만을 모두 고르면? (단, ~는 확률변수가 해당 확률분포를 따르는 것을 나타낸다)

ㄱ. $X_1 \sim P(m_1), X_2 \sim P(m_2)$이면 두 확률변수의 합 $X_1 + X_2 \sim P(m_1 + m_2)$이다. 단, $P(m)$은 평균이 m인 포아송분포
ㄴ. $X_1 \sim \chi^2(m_1), X_2 \sim \chi^2(m_2)$이면 두 확률변수의 합 $X_1 + X_2 \sim \chi^2(m_1 + m_2)$이다. 단, $\chi^2(m)$은 자유도가 m인 카이제곱분포
ㄷ. $X_1 \sim B(m_1, p), X_2 \sim B(m_2, p)$이면 두 확률변수의 합 $X_1 + X_2 \sim B(m_1 + m_2, p)$이다. 단, $B(m, p)$는 베르누이 시행횟수가 m이고 성공률이 p인 이항분포
ㄹ. $X_1 \sim N(\mu_1, \sigma^2), X_2 \sim N(\mu_2, \sigma^2)$이면 두 확률변수의 차 $X_1 - X_2 \sim N(\mu_1 - \mu_2, 2\sigma^2)$이다. 단, $N(\mu, \sigma^2)$은 평균이 μ, 분산이 σ^2인 정규분포

① ㄱ, ㄷ
② ㄱ, ㄴ, ㄹ
③ ㄴ, ㄷ, ㄹ
④ ㄱ, ㄴ, ㄷ, ㄹ

정답 4번

간단한 해설

독립인 두 정규확률변수의 차도 정규분포를 따른다. 평균은 두 평균의 차이지만 **분산은 두 분산의 합**이므로, ㄹ의 차의 분산은 $2\sigma^2$이다.

NoteBase

보기별 해설

ㄱ. O: **독립인 포아송확률변수의 합**은 각 평균을 더한 값을 평균으로 하는 포아송분포를 따른다. 두 독립적인 사건 발생 횟수를 합치면 전체 발생 횟수도 포아송분포를 따르며, 평균 발생 횟수는 $m_1 + m_2$가 된다. 따라서 $X_1 + X_2 \sim P(m_1 + m_2)$이다.

ㄴ. O: 카이제곱분포는 독립인 표준정규확률변수들의 제곱합으로 나타낼 수 있으며, 제곱합에 포함된 변수의 수가 자유도이다. 독립인 두 카이제곱확률변수를 더하면 두 제곱합이 합쳐져 **자유도도 더해진다**. 따라서 $X_1 + X_2 \sim \chi^2(m_1 + m_2)$이다.

ㄷ. O: 이항확률변수는 독립적인 베르누이 시행에서 얻은 성공 횟수이다. 두 이항확률변수가 독립이고 **성공률이 같은** p이므로, 두 성공 횟수의 합은 총 $m_1 + m_2$번의 시행에서 얻은 성공 횟수로 볼 수 있다. 따라서 $X_1 + X_2 \sim B(m_1 + m_2, p)$이다.

ㄹ. O: 독립인 정규확률변수의 선형결합은 정규분포를 따른다. 차의 평균은 $E(X_1 - X_2) = \mu_1 - \mu_2$이다. 독립이므로 공분산이 0이고, 분산은 각 계수의 제곱을 곱해 더하여 $\sigma^2 + (-1)^2\sigma^2 = 2\sigma^2$가 된다. 즉, **뺄셈이어도 분산은 더한다**. 따라서 $X_1 - X_2 \sim N(\mu_1 - \mu_2, 2\sigma^2)$이다.

ㄱ, ㄴ, ㄷ, ㄹ이 모두 성립하므로, 이 네 보기를 모두 포함한 조합은 ④이다.

NoteBase

문 11. 다음 귀무가설(H_0)과 대립가설(H_1)에 대한 검정법의 제2종 오류에 해당하는 것은?

$H_0 : \mu = 10$ 대 $H_1 : \mu \neq 10$

① $\mu = 10$일 때 귀무가설을 기각하는 것
② $\mu = 15$일 때 귀무가설을 기각하는 것
③ $\mu = 10$일 때 귀무가설을 기각하지 않은 것
④ $\mu = 15$일 때 귀무가설을 기각하지 않은 것

정답 4번

간단한 해설

제2종 오류는 **거짓인 귀무가설을 기각하지 않는 것**이다. 실제로 $\mu = 15$이면 귀무가설 $\mu = 10$은 거짓이므로, 이를 기각하지 않는 경우가 해당한다.

개념 구분

제2종 오류: 귀무가설이 실제로 거짓인데도 검정에서 이를 기각하지 않는 오류이다. 이 문제의 귀무가설은 $H_0 : \mu = 10$이므로, 실제로 $\mu = 15$이면 대립가설 $H_1 : \mu \neq 10$이 참이다. 이때 **거짓인 귀무가설을 기각하지 않은 경우**가 제2종 오류이며, ④가 이에 해당한다.

제1종 오류와의 구분: 제1종 오류는 **참인 귀무가설을 기각하는 오류**이다. 따라서 $\mu = 10$일 때 귀무가설을 기각하는 ①은 제1종 오류이다. 반면 $\mu = 15$일 때 기각하는 ②와 $\mu = 10$일 때 기각하지 않는 ③은 실제 귀무가설의 참·거짓에 맞는 판단이므로 오류가 아니다.

NoteBase

문 12. 어떤 공무원이 종합민원실에 방문하는 민원인의 수가 요일별로 고르게 분포되어 있다고 생각한다. 이 가설을 확인하고자 4주간 요일별로 방문하는 민원인의 수를 정리하여 다음 표를 작성하였다. 이 가설의 검정방법에 대한 설명으로 옳지 않은 것은? (단, $\chi^2_\alpha(k)$는 자유도가 k인 카이제곱분포의 제100×(1−α) 백분위수를 나타낼 때 $\chi^2_{0.05}(4)=9.49, \chi^2_{0.05}(5)=11.07$이다.)

요일	월	화	수	목	금	계
민원인의 수	120	80	80	100	120	500

① 귀무가설이 참일 때 민원인의 각 요일별 방문 확률은 0.2이다.

② 귀무가설이 참일 때 월요일에 방문하는 민원인의 수의 기대도수는 100이다.

③ 이 가설을 검정하기 위한 카이제곱통계량의 값은 20이다.

④ 유의수준 5%에서 요일별 방문 확률은 같다고 할 수 없다.

정답 3번

간단한 해설

관측도수와 기대도수의 차이를 제곱한 뒤 **기대도수로 나누어 합산**해야 한다. 각 요일의 기대도수가 100이므로 이 표의 카이제곱통계량은 **20이 아니라 16**이다.

NoteBase

선지별 해설

요일별 방문 확률이 모두 같다는 귀무가설을 카이제곱 적합도 검정으로 확인한다.

① O: 귀무가설은 월요일부터 금요일까지 다섯 요일의 방문 확률이 모두 같다는 것이다. 다섯 확률의 합은 1이므로 각 요일의 방문 확률은 $\frac{1}{5} = 0.2$이다.

② O: 기대도수는 **전체 도수에 귀무가설 아래의 해당 범주 확률을 곱한 값**이다. 표에서 4주간 방문한 민원인은 총 500명이므로 월요일의 기대도수는 $500 \times 0.2 = 100$명이다. 다른 네 요일의 기대도수도 각각 100명이다.

③ X: 요일별 방문 확률이 균등한지 확인하는 카이제곱 적합도 검정이다. 각 요일의 관측도수를 O_i, 기대도수를 E_i라 하면 검정통계량은 $\chi^2 = \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i}$이다. 차이를 제곱하여 부호가 상쇄되지 않게 하고, 기대도수로 나누어 각 요일의 차이를 평가한다. 표의 관측도수는 월요일부터 차례로 120, 80, 80, 100, 120이고 기대도수는 모두 100이다. 월요일의 기여도는 $\frac{(120-100)^2}{100} = 4$이며, 나머지 요일도 같은 방식으로 계산하면 화요일 4, 수요일 4, 목요일 0, 금요일 4이다. 이를 합하면 $4 + 4 + 4 + 0 + 4 = 16$이므로, 통계량의 값은 **20이 아니라 16**이다.

④ O: 범주는 다섯 요일이고 자료에서 별도로 추정한 모수가 없으므로 자유도는 $5 - 1 = 4$이다. 관측도수와 기대도수의 차이가 클수록 통계량이 커지므로 **오른쪽 꼬리를 기각역**으로 삼는다. 유의수준 5%의 임계값은 문제에서 주어진 $\chi^2_{0.05}(4) = 9.49$이며, 계산한 통계량 16이 이를 초과하므로 귀무가설을 기각한다. 따라서 유의수준 5%에서 요일별 방문 확률이 모두 같다고 할 수 없다.

NoteBase

문 13. 다음 분산분석에 대한 설명으로 옳은 것만을 모두 고르면?

ㄱ. 3개 이상 모집단의 평균비교를 위한 통계 기법이다.
ㄴ. 각 비교 대상 모집단은 서로 다른 분산을 가져야 한다.
ㄷ. 일원배치법은 각 처리수준에서 반복수가 일정해야 한다.
ㄹ. 반복이 있는 이원배치법에서는 두 요인의 교호작용을 검정할 수 있다.

① ㄱ, ㄷ
② ㄱ, ㄹ
③ ㄴ, ㄷ
④ ㄴ, ㄹ

정답 2번

간단한 해설

일원배치 분산분석은 **처리수준별 반복수가 달라도** 적용할 수 있다. 반복이 있는 이원배치에서는 한 요인의 효과가 다른 요인의 수준에 따라 달라지는 **교호작용**을 오차와 구분하여 검정할 수 있다.

NoteBase

보기별 해설

ㄱ. O: 분산분석은 **여러 모집단의 평균이 모두 같은지**를 검정하는 기법으로, 세 개 이상 모집단의 평균 비교에 사용된다. 집단 간 변동을 집단 내 변동과 비교하여 평균 차이가 우연한 오차로 설명될 수 있는지 판단한다. 평균이 모두 같다는 가설을 기각하면 적어도 한 모집단의 평균이 다르다는 뜻이며, 모든 평균이 서로 다르다는 뜻은 아니다.

ㄴ. X: 고전적인 분산분석에서는 비교 대상 모집단의 분산이 서로 다르다는 조건이 아니라 **분산이 같다는 등분산성**을 가정한다. 이 가정 아래 집단 내 변동을 합쳐 공통 오차분산을 추정하고, 이를 집단 간 변동과 비교한다.

ㄷ. X: 일원배치법의 '일원'은 분석하는 **요인이 하나**라는 뜻이며, 처리수준별 반복수가 같아야 한다는 뜻이 아니다. 반복수는 각 처리수준에서 얻은 관측값의 수이다. 반복수가 같으면 균형설계, 서로 다르면 불균형설계이며, 일원배치 분산분석은 두 경우 모두 가능하다.

ㄹ. O: 교호작용은 **한 요인의 효과가 다른 요인의 수준에 따라 달라지는 현상**이다. 반복이 있는 이원배치법에서는 각 요인 수준의 조합에서 관측을 반복하므로, 동일한 조합 내의 변동으로 오차분산을 추정할 수 있다. 이를 기준으로 두 요인의 개별 효과뿐 아니라 교호작용의 유의성도 검정할 수 있다.

옳은 보기는 ㄱ, ㄹ이며, 이 조합은 ②에 해당한다.

NoteBase

문 14. 키가 큰 사람의 연봉이 높다는 주장이 있다. 이를 검정하기 위해 MBA 졸업생 250명의 키(X)와 연봉(Y)을 조사하여 단순선형회귀분석한 결과의 일부가 다음 표와 같다. 키에 대한 회귀계수의 최소제곱추정량의 값 ㉠은? (단, 키에 대한 회귀계수는 양수이다)

<분산분석표>

	자유도	제곱합	평균제곱	F값	$p-$값
회귀	1	960	960	16	0.00
잔차	248				
계	249				

	계수	표준 오차	t값	$p-$값
Y 절편	17.93	11.48	1.56	0.12
키	㉠	0.20		0.00

① 0.8

② 1.0

③ 2.0

④ 4.0

정답 1번

간단한 해설

설명변수가 하나인 단순선형회귀에서는 기울기 검정의 $F = t^2$ 관계가 성립한다. **기울기가 양수**이므로 t값은 4이고, 여기에 표준오차 0.20을 곱하면 회귀계수 추정값은 0.8이다.

NoteBase

계산 과정

검정통계량의 관계: 키 X가 설명변수이고 연봉 Y가 반응변수이며, 구하려는 ㉠은 키에 대한 기울기의 최소제곱추정값 $\hat{\beta}_1$이다. 절편을 포함하고 **설명변수가 하나인 단순선형회귀**에서는 기울기가 0이라는 귀무가설 $H_0 : \beta_1 = 0$에 대한 분산분석의 F통계량이 기울기의 t통계량을 제곱한 값과 같다. 따라서 $F = t^2$를 이용한다.

t값 계산: 분산분석표의 회귀 행에서 F값은 16이다. 따라서 $t^2 = 16$이므로 t값의 절댓값은 4이다. 기울기의 t통계량은 $t = \frac{\hat{\beta}_1}{\mathrm{SE}(\hat{\beta}_1)}$이고 표준오차는 양수이므로, **기울기와 t값의 부호는 같다**. 문제에서 기울기가 양수라고 했으므로 $t = 4$이다.

㉠ 계산: 계수표의 키 행에서 표준오차는 0.20이다. 기울기 추정값은 t값에 표준오차를 곱하여 구하므로 $\hat{\beta}_1 = 4 \cdot 0.20 = 0.8$이다. 이는 키가 1단위 증가할 때 예측 연봉이 0.8단위 증가한다는 뜻이며, 각 단위는 원자료에서 사용한 키와 연봉의 단위를 따른다.

따라서 ㉠은 0.8이며, ①의 값과 일치한다.

NoteBase

문 15. 두 인자 A와 B가 어떤 특성값에 영향을 주는가에 대해 인자의 수준 조합에서 동일하게 반복실험을 행하여 분산분석을 한 결과의 일부가 다음 표와 같다. 이에 대한 설명으로 옳지 않은 것은?

<분산분석표>

요인	제곱합	자유도	평균제곱	$p-$값
인자 A	9	3	3	
인자 B	2			
교호작용 $A\times B$	900	6	150	㉠
오차	20			
계	931	23		

① 인자 A의 수준 수는 4이고, 인자 B의 수준 수는 3이다.

② 인자 A, B의 수준 조합에서 반드시 두 번의 반복실험을 해야 한다.

③ 유의수준 5%에서 ㉠이 0.05보다 작으면 인자 A와 B의 각각에 대한 유의성 검정은 의미가 없다.

④ 유의수준 5%에서 ㉠이 0.05보다 작으면 교호작용을 오차항에 풀링(pooling)하여 분산분석을 다시 실시한다.

정답 4번

간단한 해설

㉠은 교호작용의 p-값이므로 0.05보다 작으면 **교호작용이 유의하다**. 이때는 교호작용을 모형에 유지해야 하며, **오차항에 풀링해서는 안 된다**.

NoteBase

선지별 해설

① O: A와 B의 수준 수를 각각 a, b라 하면, 표에서 A의 자유도가 3이므로 $a-1=3$, $a=4$이다. 교호작용의 자유도는 $(a-1)(b-1)=6$이므로 $3(b-1)=6$, $b=3$이다. 따라서 A는 4수준, B는 3수준이다.

② O: 표의 총 자유도가 23이므로 전체 관측값은 23 + 1 = 24개이다. 수준 조합은 $4\times3=12$개이고, 각 조합에서 같은 횟수로 반복했으므로 조합당 반복수 r은 $r=\frac{24}{12}=2$이다. 여기서 '반드시 두 번'은 **주어진 분산분석표에 부합하는 반복수**가 2회라는 뜻이며, 모든 이원배치 실험에서 반복수를 2회로 정해야 한다는 뜻은 아니다.

③ O: ㉠은 교호작용의 p-값이다. ㉠이 0.05보다 작으면 교호작용이 없다는 귀무가설을 기각한다. 교호작용이 있다는 것은 한 인자의 효과가 다른 인자의 수준에 따라 달라진다는 뜻이다. 따라서 다른 인자의 수준을 평균하여 얻는 주효과만으로 각 인자의 영향을 해석하면 수준별 차이를 놓칠 수 있다. 선지의 '의미가 없다'는 **교호작용을 무시한 주효과 검정만으로 효과를 해석하기 어렵다**는 취지이다. 주효과 검정 자체가 수학적으로 불가능하다는 뜻은 아니며, 다른 인자의 수준별 효과인 단순효과나 수준 조합별 평균을 살펴야 한다.

④ X: 풀링은 무시할 수 있다고 판단한 효과의 제곱합과 자유도를 오차항에 합쳐 오차분산을 다시 추정하는 것이다. ㉠이 0.05보다 작으면 교호작용이 유의하므로 이를 오차에 합치면 실제 효과를 오차로 처리하게 된다. 이때는 **교호작용을 모형에 유지**해야 한다. 교호작용의 오차항 풀링은 교호작용이 유의하지 않고 무시할 수 있다고 판단할 때 검토한다.

NoteBase

문 16. 흡연과 심근경색의 연관성을 평가하기 위하여 특정 병원을 방문하여 심근경색으로 진단된 환자 중에서 300명, 심근경색이 아닌 환자 중에서 600명을 각각 임의추출하여 흡연 여부를 조사한 분할표가 다음과 같다. 이 분할표에 대한 설명으로 옳은 것은?

흡연 여부 \ 심근경색 여부	예	아니오	계
예	150	150	300
아니오	150	450	600
계	300	600	900

① 동질성 검정의 카이제곱통계량의 자유도는 2이다.

② 동질성 검정의 카이제곱통계량의 값은 $\chi^2 = \frac{(150-100)^2}{150} + \frac{(150-200)^2}{150} + \frac{(150-200)^2}{150} + \frac{(450-400)^2}{450}$이다.

③ 분할표로부터 심근경색이 발병한 환자에서의 흡연율을 추정할 수 있다.

④ 분할표로부터 흡연한 사람에서의 심근경색 발생률을 추정할 수 있다.

정답 3번

간단한 해설

심근경색 환자 집단에서 임의추출했으므로 **심근경색 환자에서의 흡연율**을 추정할 수 있다. 표의 '심근경색 여부: 예' 열에서 전체 300명 중 흡연자가 150명이므로 추정치는 $\frac{150}{300} = 50\%$이다.

NoteBase

선지별 해설

① X: 합계 행과 열을 제외하면 흡연 여부 2개 범주와 심근경색 여부 2개 범주로 이루어진 분할표이다. 동질성 검정의 자유도는 행 범주 수와 열 범주 수에서 각각 1을 뺀 값의 곱이므로 $(2-1)(2-1)=1$이다. 따라서 자유도는 2가 아니라 **1**이다.

② X: 관측도수를 O, 두 집단의 흡연율이 같다는 가정 아래의 기대도수를 E라고 하면 카이제곱통계량은 $\chi^2=\sum\frac{(O-E)^2}{E}$이다. 각 셀의 기대도수는 E=행 합계·열 합계900로 구한다. 예를 들어 흡연하면서 심근경색인 셀의 기대도수는 $\frac{300\cdot 300}{900}=100$이다. 표의 첫째 행 왼쪽부터 차례로 기대도수는 100, 200, 200, 400이며, **각 항의 분모에는 기대도수**가 들어가야 한다. 선지는 분모에 관측도수인 150, 150, 150, 450을 넣었다. 올바른 각 항의 값은 25, 12.5, 12.5, 6.25이므로 카이제곱통계량은 25 + 12.5 + 12.5 + 6.25 = 56.25이다.

③ O: '심근경색 여부: 예' 열은 심근경색 환자 중 임의추출한 300명을 나타내며, 이 중 '흡연 여부: 예'에 해당하는 사람은 150명이다. 따라서 $\frac{150}{300}=50\%$로 **조사 대상 병원의 심근경색 환자에서의 흡연율**을 추정할 수 있다. 발병한 환자라는 조건 안에서 흡연자의 비율을 구하는 것이다.

④ X: 이 조사는 흡연자들을 먼저 뽑아 발병 여부를 조사한 것이 아니라, **발병 여부별 표본 수를 미리 정하여 추출**한 조사이다. 표본의 흡연자 300명 중 심근경색 환자가 150명이더라도, 그 비율은 환자 300명과 비환자 600명이라는 표본 배분에 영향을 받는다. 따라서 이 표만으로 흡연자의 심근경색 발생률을 추정할 수 없다. 추정할 수 있는 것은 심근경색 환자 집단이나 비환자 집단 각각에서의 흡연율이며, 흡연자의 발생률을 구하려면 별도의 추적 자료 등 추가 정보가 필요하다.

NoteBase

문 17. 어떤 질병의 새로운 치료약을 개발하였다. 다음 표는 복용량에 따라 이 질병 증상의 경감기간을 알아보기 위해 10명의 환자에 대하여 실험한 자료의 일부이다. 복용량과 경감기간의 관계를 단순선형회귀모형 $y_i = \alpha + \beta x_i + \varepsilon_i \ (i = 1, 2, \cdots, 10)$에 적용하였다. 이때 ε_i는 서로 독립이며 평균이 0, 분산이 σ^2인 정규분포를 따른다. $\sum_{i=1}^{10}(x_i - \bar{x})^2 = 90$, $\sum_{i=1}^{10}(y_i - \bar{y})^2 = 1{,}000$, $\sum_{i=1}^{10}(x_i - \bar{x})(y_i - \bar{y}) = 270$일 때, 옳지 않은 것은? (단, $\bar{x}, \bar{y}$는 각각 x, y의 표본평균이다)

번호	1	2	3	⋯	8	9	10
복용량(x)	3	3	4	⋯	8	8	9
경감기간(y)	9	5	12	⋯	18	24	22

① 회귀계수 β의 최소제곱추정량의 값은 3이다.

② 복용량과 경감기간의 상관계수는 0.9이다.

③ 결정계수는 0.81이다.

④ σ^2의 불편추정량의 값은 $\frac{190}{9}$이다.

정답 4번

간단한 해설

절편과 기울기 두 모수를 추정하므로 오차분산의 불편추정량은 잔차제곱합을 **10 − 2 = 8로 나눈 값**이다.
잔차제곱합은 190이므로, ④의 $\frac{190}{9}$가 아니라 $\frac{190}{8} = 23.75$이다.

NoteBase

선지별 해설

표에서 복용량은 x, 경감기간은 y이다. 복용량의 편차제곱합을 $S_{xx} = 90$, 경감기간의 편차제곱합을 $S_{yy} = 1000$, 두 변수의 편차곱합을 $S_{xy} = 270$이라 쓰자. 표본 수는 $n = 10$이다.

① O: 최소제곱법은 관측값과 회귀직선의 예측값 사이 차이의 제곱합을 최소로 한다. 절편을 포함한 단순선형회귀에서 기울기의 추정값은 **편차곱합을 설명변수의 편차제곱합으로 나눈 값**이므로, $\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{270}{90} = 3$이다.

② O: 표본상관계수는 편차곱합을 두 변수의 편차제곱합 곱의 제곱근으로 나누어 구한다. $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$에 주어진 값을 대입하면 $r = \frac{270}{\sqrt{90 \cdot 1000}} = 0.9$이다. 편차곱합이 양수이므로 상관계수의 부호도 양수이다.

③ O: **절편을 포함한 단순선형회귀**의 결정계수는 표본상관계수의 제곱과 같다. 따라서 $R^2 = r^2 = 0.9^2 = 0.81$이다. 이는 표본에서 경감기간의 전체 변동 중 81%가 복용량에 대한 회귀직선으로 설명된다는 뜻이다.

④ X: 잔차제곱합 SSE는 관측값과 회귀직선의 예측값 사이 차이의 제곱합이다. 최소제곱 회귀직선에 대해 $SSE = S_{yy} - \hat{\beta}S_{xy}$이므로, 그 값은 $1000 - 3 \times 270 = 190$이다. 오차분산 σ^2의 불편추정량은 잔차제곱합을 **잔차의 자유도**로 나눈 값이다. 절편과 기울기 두 모수를 추정하므로 자유도는 $10 - 2 = 8$이다. 따라서 $\hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{190}{8} = 23.75$이며, 분모를 9로 둔 $\frac{190}{9}$는 불편추정량의 값이 아니다.

NoteBase

문 18. 다음 표에서 x'와 y'는 각각 설명변수(x)와 반응변수(y)를 표준화한 변수 $x' = \frac{x-\bar{x}}{s_x}$, $y' = \frac{y-\bar{y}}{s_y}$의 자료이다. 위의 표준화된 변수 (x', y')에 대해 단순선형회귀분석을 최소제곱법으로 수행할 때, 회귀계수의 추정값에 대한 설명으로 옳은 것은? (단, $\bar{x}, \bar{y}$는 각각 x, y의 표본평균이고, s_x, s_y는 각각 x, y의 표본표준편차이고, $s_x \neq s_y$이다)

번호	1	2	3	4	5	6	7	8	9	10
x'	-1.7	0.0	0.4	0.0	0.0	1.8	-0.9	0.9	-0.9	0.4
y'	-1.4	0.1	0.5	-0.6	-0.3	1.9	-0.1	1.0	-1.3	0.2

① 원 변수 (x, y)에 대한 회귀계수의 추정값과 동일하다.

② 원 변수 (x, y)에 대한 회귀계수의 추정값의 역수와 동일하다.

③ 원 변수 (x, y)의 상관계수와 동일하다.

④ 원 변수 (x, y)의 결정계수와 동일하다.

정답 3번

간단한 해설

두 변수를 표준화하면 표본표준편차가 모두 1이 되므로, 단순선형회귀의 기울기는 **원 변수의 상관계수**와 같다. 결정계수는 이 상관계수를 제곱한 값이다.

핵심 풀이

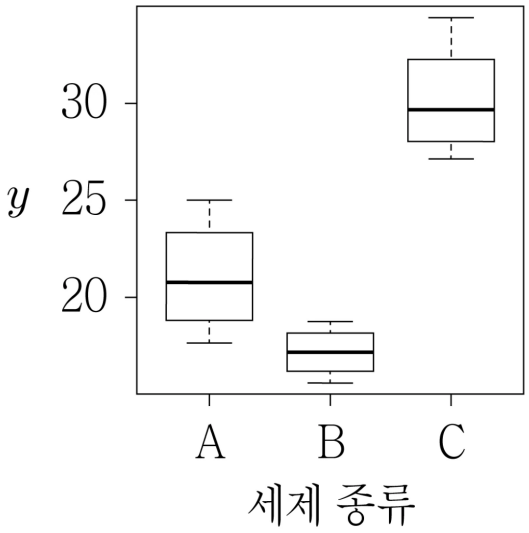
표준화의 효과: 표의 x', y'는 각 변수에서 표본평균을 빼고 표본표준편차로 나눈 자료이다. 따라서 두 변수는 **평균이 0, 표본표준편차가 1**이다. 또한 양의 표준편차로 나누는 표준화는 상관계수를 바꾸지 않으므로, 원 변수와 표준화된 변수의 표본상관계수는 같은 r이다.

회귀기울기와의 연결: 절편을 포함한 단순선형회귀에서 원자료의 최소제곱 기울기는 $\hat{b} = r\frac{s_y}{s_x}$이다. 즉, 상관계수에 두 변수의 표준편차 비율을 곱한다. 표준화된 자료에서는 그 비율이 1이므로 $\hat{b}' = r \cdot \frac{1}{1} = r$이다. 따라서 표준화된 변수의 회귀기울기는 **원 변수의 상관계수**와 동일하다. 표의 수치를 일일이 계산하지 않아도 이 관계로 판단할 수 있다.

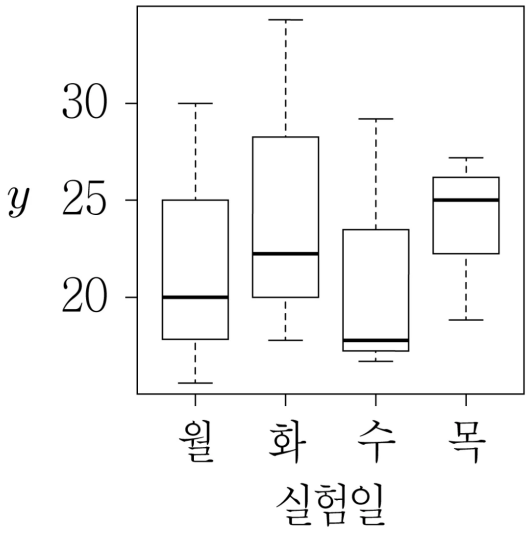
다른 양과의 구분: 원자료의 기울기와는 $\hat{b}' = \hat{b}\frac{s_x}{s_y}$의 관계이다. 문제에서 $s_x \neq s_y$이므로 표준화 전후의 기울기를 일반적으로 같은 값으로 둘 수 없으며, 원자료 기울기의 역수를 취하는 것도 아니다. 한편 절편을 포함한 단순선형회귀의 결정계수는 $R^2 = r^2$이다. **기울기는 상관계수 자체**이고, 결정계수는 그 제곱이라는 차이가 있다.

NoteBase

문 19. 3가지 종류의 유기농 세제(A, B, C)에 대해 박테리아의 성장을 지연시키는 효과를 비교하고자 한다. 실험실에서 하루에 3번의 실험만 가능하므로 실험일 월, 화, 수, 목을 블록으로 하여 확률화완전블록설계를 실시하여 다음과 같은 결과를 얻었다. 분산분석을 이용하여 세제 종류별 박테리아 성장 지연 효과의 차이를 검정하고자 한다. 분석결과에 대한 설명으로 옳은 것만을 모두 고르면?



<그림 1> 세제 종류별 상자그림



<그림 2> 실험일별 상자그림

<분산분석표>

요인	자유도	제곱합	평균제곱	F값	$p-$값
세제(처리)		360	180	36	0.00
실험일(블록)		30	10	2	0.22
오차		30	5		
계					

ㄱ. 유의수준 5%에서 세제 종류별 박테리아의 성장을 지연시키는 효과의 차이는 유의하다.

ㄴ. 유의수준 5%에서 실험일간 특성값(y)의 산포가 매우 작다고 할 수 있다.

ㄷ. 귀무가설이 참일 때 세제에 대한 검정통계량의 확률분포는 분자의 자유도는 3, 분모의 자유도는 6인 F 분포이다.

ㄹ. 세제 종류별 상자그림에서 범위가 가장 큰 세제는 B이다.

① ㄱ, ㄴ

NoteBase

② ㄱ, ㄷ

③ ㄱ, ㄴ, ㄷ

④ ㄴ, ㄷ, ㄹ

정답 1번

간단한 해설

세제 효과의 p값은 표에 0.00으로 표시되어 유의수준 5%보다 작으므로 **세제 종류에 따른 평균 효과의 차이가 유의하다**. 실험일 효과의 p값은 0.22이므로 실험일 간 유의한 평균 차이는 확인되지 않는다.

보기별 해설

확률화완전블록설계에서는 세제 3종을 처리로, 실험일 4일을 블록으로 두고 실험일의 영향을 분리하여 세제의 평균 효과를 비교한다. 각 날짜에 세제별로 한 번씩 실험하므로 관측값은 총 12개이다. 통상적인 분산분석은 처리와 블록의 효과가 더해지는 모형에서 오차의 독립성·정규성·등분산성을 전제로 한다.

ㄱ. O: 분산분석표의 '세제(처리)' 행에서 F값은 36이고, p값은 0.00으로 표시되어 0.05보다 작다. 따라서 세 세제의 평균 성장 지연 효과가 모두 같다는 귀무가설을 기각한다. 이는 **적어도 한 세제의 평균 효과가 다르다**는 뜻이며, 모든 세제 쌍의 차이가 각각 유의하다는 뜻은 아니다. 표의 0.00은 반올림된 표시이므로 실제 p값이 정확히 0이라는 의미도 아니다.

ㄴ. O: '실험일(블록)' 행의 p값은 0.22로 0.05보다 크므로, 실험일 효과가 없다는 귀무가설을 기각하지 못한다. 이 문항의 '실험일간 특성값의 산포가 매우 작다'는 표현은 **실험일 간 평균 차이가 통계적으로 유의하지 않다**는 취지로 해석한다. 다만 이 검정은 평균 차이를 검정하는 것이므로, 이 결과만으로 실제 산포가 매우 작거나 실험일 효과가 전혀 없다고 입증한 것은 아니다.

ㄷ. X: 세제 검정의 분자 자유도는 **세제 수에서 1을 뺀 값**이므로 $3-1=2$이다. 분모는 오차의 자유도로, 처리와 블록의 효과를 분리한 뒤 남는 자유도인 $(3-1)(4-1)=6$이다. 따라서 귀무가설 아래 세제 검정통계량은 $F_{2,6}$ 분포를 따른다. 분자 자유도 3은 세제가 아니라 실험일 4개에 대응하는 블록 자유도이다.

ㄹ. X: 범위는 **최댓값에서 최솟값을 뺀 값**이다. 그림 1에는 별도로 표시된 이상값이 없으므로 각 세제의 위아래 수염 끝 사이의 간격을 비교하면 된다. B는 이 간격이 A와 C보다 짧으므로 범위가 가장 큰 세제가 아니라 가장 작은 세제이다. 상자 자체의 높이는 가운데 50% 자료의 산포인 사분위범위이므로 전체 범위와 구별한다.

따라서 옳은 보기는 ㄱ, ㄴ이며, 이 조합에 해당하는 선택지는 ①이다.

NoteBase

문 20. 어떤 직육면체 주사위의 1의 눈이 있는 면이 작아서 1의 눈이 나올 확률 p가 $\frac{1}{6}$보다 작을 것으로 예상된다. 이에 대한 가설 $H_0 : p = \frac{1}{6}$ 대 $H_1 : p < \frac{1}{6}$을 검정하기 위해 이 주사위를 10번 던져 1의 눈이 나온 횟수를 확률변수 Y라 할 때, 옳은 것은?

① $\frac{Y}{10}$는 p의 불편추정량(unbiased estimator)이 아니다.

② Y의 관측값이 클수록 귀무가설을 기각한다.

③ Y의 확률분포는 이항분포 $B\left(10, \frac{1}{6}\right)$을 따른다.

④ Y의 관측값이 2일 때 유의확률(p-값)은 $P(X \leq 2)$이다. 단, 확률변수 X는 이항분포 $B\left(10, \frac{1}{6}\right)$을 따른다.

정답 4번

간단한 해설

대립가설이 $p < \frac{1}{6}$이므로 1의 눈이 나온 횟수가 **작을수록** 대립가설을 지지한다. 관측값이 2일 때 p-값은 **귀무가설 아래에서 2 이하가 나올 확률**인 $P(X \leq 2)$이다.

선지별 해설

각 투척을 독립 시행으로 보고 매번 1의 눈이 나올 확률이 p로 같다고 하면, 10번 중 1의 눈이 나온 횟수는 $Y \sim B(10, p)$를 따른다.

① X: 불편추정량은 기댓값이 추정하려는 모수와 같은 추정량이다. 이항분포에서 $E(Y) = 10p$이므로 $E\left(\frac{Y}{10}\right) = \frac{10p}{10} = p$이다. 따라서 표본비율 $\frac{Y}{10}$는 p의 **불편추정량이다**.

② X: 대립가설은 $p < \frac{1}{6}$이므로, 1의 눈이 적게 나올수록 대립가설에 부합한다. 따라서 귀무가설은 Y의 관측값이 **충분히 작을 때** 기각하며, 관측값이 클수록 기각하는 것이 아니다.

③ X: Y의 분포는 $B(10, p)$이다. $B\left(10, \frac{1}{6}\right)$은 **귀무가설이 참일 때의 분포**이며, 검정 대상인 미지의 확률 p를 항상 $\frac{1}{6}$로 확정할 수는 없다.

④ O: p-값은 귀무가설이 참이라고 가정할 때, 관측값과 같거나 대립가설을 더 강하게 지지하는 값이 나올 확률이다. 이 검정에서는 작은 횟수가 대립가설을 지지하므로, 관측값이 2이면 **0회·1회·2회가 나올 확률을 합한다**. 귀무가설 아래의 확률변수 $X \sim B\left(10, \frac{1}{6}\right)$를 사용하면 p-값은 $P(X \leq 2)$이다.

NoteBase