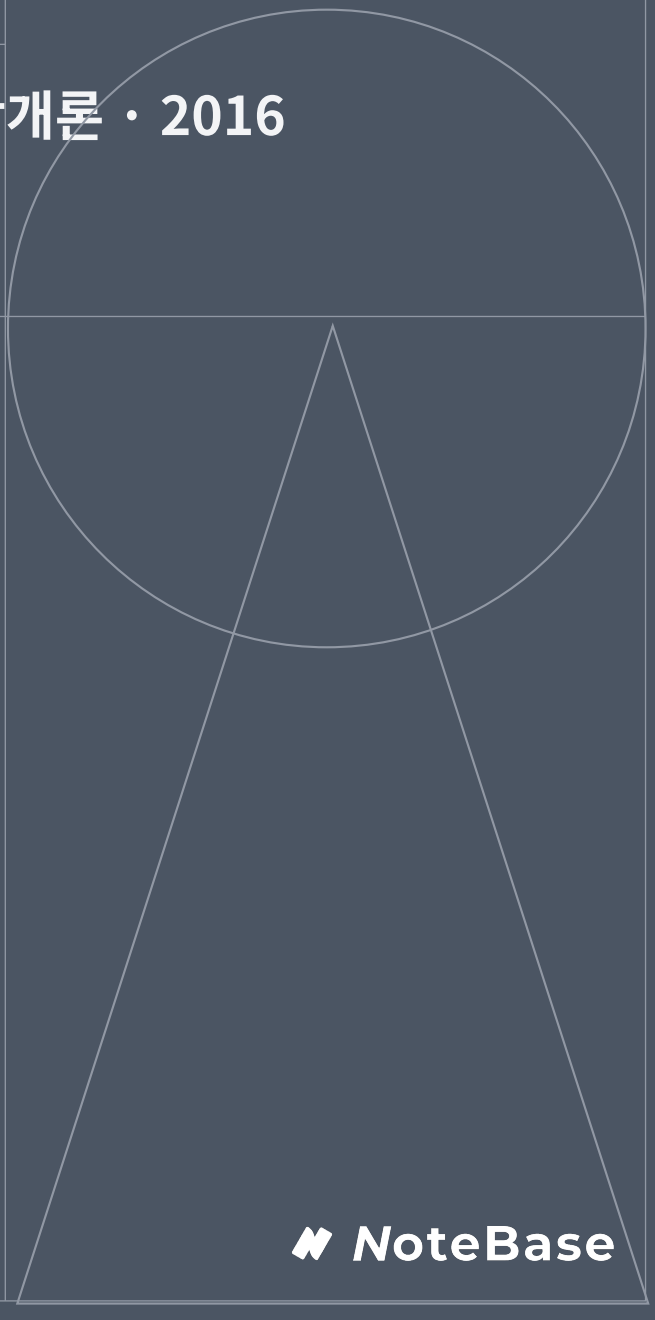
기출OX 해설집

공무원/국가직/9급 · 통계학개론 · 2016



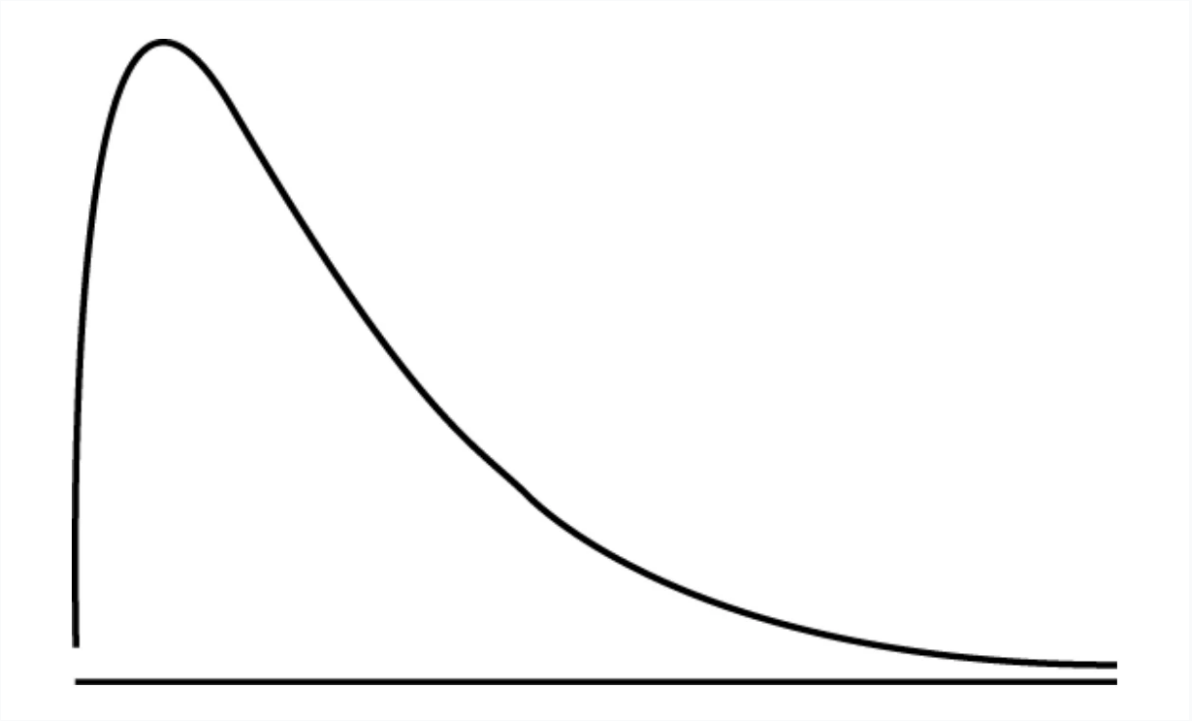
NoteBase

기출OX 해설집

공무원/국가직/9급 · 통계학개론 · 2016

공무원/국가직/9급 · 통계학개론 · 2016

문 1. 그림과 같이 분포가 오른쪽으로 긴 꼬리를 가지는 확률변수에 대해 최빈값(mode), 평균(mean), 중앙값(median)을 작은 것부터 순서대로 바르게 나열한 것은?



① 최빈값 < 평균 < 중앙값

② 중앙값 < 평균 < 최빈값

③ 최빈값 < 중앙값 < 평균

④ 평균 < 중앙값 < 최빈값

정답 3번

간단한 해설

오른쪽으로 긴 꼬리에 있는 **큰 값들이 평균을 오른쪽으로 끌어당긴다**. 그림과 같은 전형적인 분포에서 최빈값은 봉우리의 위치에 있고, 중앙값은 최빈값과 평균 사이에 놓인다.

NoteBase

올바른 순서

최빈값 → 중앙값 → 평균

최빈값: 확률밀도곡선의 **높이가 가장 높은 지점의 가로축 값**이다. 그림에서는 봉우리가 왼쪽에 있으므로 세 대표값 중 가장 작은 위치에 놓인다.

중앙값: 곡선 아래 전체 넓이를 **좌우 절반씩 나누는 가로축 값**이다. 곡선의 높이를 반으로 나누는 값과는 다르다. 그림에서는 봉우리 오른쪽에도 확률이 넓게 퍼져 있어, 누적확률이 절반에 도달하는 위치가 최빈값보다 오른쪽에 놓인다.

평균: 각 값에 그 값이 나타날 확률을 가중하여 구하는 대표값이다. 중앙값은 확률을 절반씩 나누는 위치로 정해지지만, 평균에는 **오른쪽 꼬리의 큰 값들의 크기까지 반영**된다. 따라서 그림과 같은 전형적인 오른쪽 꼬리 분포에서는 평균이 중앙값보다 오른쪽에 놓인다.

NoteBase

문 2. 작년 입사 지원자 중 60%가 남자이고 40%가 여자인 어느 회사에서 남자 지원자 중 8%, 여자 지원자 중 6%가 합격하였을 때 이 회사의 작년 입사 지원자 중 합격한 사람의 비율은?

① 7.2%

② 7.4%

③ 7.6%

④ 7.8%

정답 1번

간단한 해설

남녀 합격률을 **지원자 비중 60%, 40%로 가중평균**해야 하므로, 전체 지원자 중 합격한 사람의 비율은 7.2%이다.

계산 과정

전체 지원자는 남자 60%, 여자 40%로 나뉜다. 남자 합격률 8%와 여자 합격률 6%는 각각 해당 성별의 지원자만을 기준으로 한 비율이다. 전체 합격률을 구하려면 **각 성별의 지원자 비중에 그 성별의 합격률을 곱한 뒤 더한다**.

남자 합격자의 비율: $0.6 \times 0.08 = 0.048$이므로, 남자 합격자는 전체 지원자의 4.8%이다. 이는 남자 지원자 중의 비율 8%를 전체 지원자 기준으로 바꾼 값이다.

여자 합격자의 비율: $0.4 \times 0.06 = 0.024$이므로, 여자 합격자는 전체 지원자의 2.4%이다.

전체 합격률: 남자 합격자와 여자 합격자는 서로 겹치지 않으므로 두 비율을 더한다. 4.8% + 2.4% = 7.2%이며, 이는 ①의 7.2%에 해당한다.

NoteBase

문 3. 두 이산확률변수 X와 Y의 결합확률분포표가 다음과 같을 때, $2X - Y$의 기대값은?

X \ Y	-1	0	1
1	0.1	0.1	0.2
2	0.1	0.2	0.3

① 2.6

② 2.7

③ 2.8

④ 2.9

정답 4번

간단한 해설

기대값의 **선형성**에 따라 $E(2X - Y) = 2E(X) - E(Y)$이다. 결합확률표에서 구한 $E(X) = 1.6$, $E(Y) = 0.3$을 적용하면 기대값은 2.9이다.

계산 과정

기대값의 선형성: 구하려는 값은 $2X - Y$의 확률가중평균이다. 기대값은 각 변수의 기대값에 계수를 곱해 더할 수 있으므로 $E(2X - Y) = 2E(X) - E(Y)$를 이용한다. 이 성질에는 **두 변수의 독립이 필요하지 않다**.

X의 기대값: 표에서 X는 행에 표시되어 있다. 같은 행의 확률을 모두 더하면 Y의 값과 관계없이 해당 X값이 나올 확률인 **주변확률**을 얻는다. X가 1일 확률은 0.1 + 0.1 + 0.2 = 0.4이고, X가 2일 확률은 0.1 + 0.2 + 0.3 = 0.6이다. 따라서 $E(X) = 1 \cdot 0.4 + 2 \cdot 0.6 = 1.6$이다.

Y의 기대값: Y는 열에 표시되어 있으므로 **각 열의 확률을 더한다**. Y가 −1일 확률은 0.1 + 0.1 = 0.2, 0일 확률은 0.1 + 0.2 = 0.3, 1일 확률은 0.2 + 0.3 = 0.5이다. Y가 0인 경우는 기대값에 0을 더하므로,
$E(Y) = (-1) \cdot 0.2 + 1 \cdot 0.5 = 0.3$이다.

기대값 계산: $E(2X - Y) = 2 \cdot 1.6 - 0.3 = 2.9$이다. 이 값은 ④의 2.9와 일치한다.

NoteBase

문 4. 다음은 어느 정책에 대하여 세 도시 A, B, C의 만족도를 조사한 결과이다.
만약 세 도시 A, B, C에서 이 정책에 대한 만족도의 분포가 동일하다면 B 도시의 '만족' 셀의 기대도수는?

만족도 / 도시	만족	보통	불만족	합계
A	34	39	12	85
B	29	31	26	86
C	33	29	19	81
합계	96	99	57	252

① $\frac{96 \times 86}{252}$
② $\frac{29}{86} \times \frac{29}{96} \times 252$
③ $\frac{96}{3} \times \frac{86}{3} \times \frac{1}{29}$
④ 29

정답 1번

간단한 해설

세 도시의 만족도 분포가 같다면 B 도시에도 **전체의 만족 비율**을 적용한다. 따라서 기대도수는 B 도시의 조사 인원 86명에 전체 만족 비율 $\frac{96}{252}$를 곱한 $\frac{96 \times 86}{252}$이다.

NoteBase

계산 과정

자료 확인: 표에서 B 도시의 행 합계는 86명, '만족' 열의 합계는 96명, 전체 조사 인원은 252명이다. B 도시의 '만족' 셀에 적힌 29명은 실제 조사에서 얻은 관측도수이다.

계산 원리: 기대도수는 **세 도시의 만족도 분포가 동일하다는 가정** 아래 해당 셀에서 예상되는 인원이다. 이 가정에서는 각 도시에 공통된 만족 비율을 전체 자료로 추정하므로, 그 비율은 $\frac{96}{252}$이다. 각 도시의 조사 인원이 다르므로 만족 인원도 도시별 조사 인원에 비례하여 배분한다.

기대도수 계산: B 도시의 조사 인원에 공통 만족 비율을 곱하면 $86 \times \frac{96}{252} = \frac{96 \times 86}{252}$로, 약 32.76명이다. 이는 ①의 식과 일치한다. 기대도수는 가정에 따라 계산한 값이므로 관측도수 29명과 달라도 되고, 정수일 필요도 없다.

NoteBase

문 5. 인자 A의 처리수준 A_1, A_2, A_3에 대하여 A_1에서 5회, A_2에서 10회, A_3에서 15회를 랜덤하게 실험하였다. 이 실험을 통하여 얻은 자료의 일원배치 분산분석에서 처리제곱합의 자유도(㉠)와 잔차제곱합의 자유도(㉡)를 바르게 연결한 것은?

① ㉠: 2, ㉡: 26

② ㉠: 2, ㉡: 27

③ ㉠: 3, ㉡: 26

④ ㉠: 3, ㉡: 27

정답 2번

간단한 해설

처리수준은 3개이고 총 실험횟수는 30회이다. 일원배치 분산분석에서 **처리 자유도는 처리수준 수에서 1을 뺀 값**이므로 2이고, **잔차 자유도는 총 관측값 수에서 처리수준 수를 뺀 값**이므로 27이다.

계산 과정

처리수준 수를 k, 총 관측값 수를 N이라 하면, 이 실험에서는 $k = 3$이고 총 관측값 수는 $5 + 10 + 15 = 30$이다. 수준별 반복횟수가 달라도 각 수준의 관측값 수를 합하여 전체 크기를 구한다.

㉠ 처리제곱합의 자유도: 처리제곱합은 **처리수준별 평균 사이의 차이**에 따른 변동을 나타낸다. 전체평균을 기준으로 한 수준별 평균의 차이에는 하나의 제약이 있으므로 자유도는 $k - 1$이다. 따라서 처리 자유도는 $3 - 1 = 2$이다.

㉡ 잔차제곱합의 자유도: 잔차제곱합은 각 관측값이 해당 처리수준의 평균에서 벗어난 정도에 따른 변동을 나타낸다. 각 수준에서 평균을 하나씩 추정하면 그 수준의 잔차 합이 0이 되므로 **수준마다 자유도가 하나씩 줄어든다**. 따라서 잔차 자유도는 $N - k$이며, $(5 - 1) + (10 - 1) + (15 - 1) = 4 + 9 + 14 = 27$이다.

따라서 ㉠은 2, ㉡은 27이며, 이 두 값을 연결한 선택지는 ②이다.

NoteBase

문 6. 가설검정에서 검정력(power)에 대한 설명으로 옳은 것은?

① 참인 귀무가설을 기각할 확률

② 참인 귀무가설을 기각하지 않을 확률

③ 거짓인 귀무가설을 기각할 확률

④ 거짓인 귀무가설을 기각하지 않을 확률

정답 3번

간단한 해설

검정력은 귀무가설이 실제로 거짓일 때, 이를 **기각하여 잘못된 가설을 찾아낼 확률**이다.

공통 해설

검정력은 **귀무가설이 거짓이라는 조건**에서 그 귀무가설을 기각할 확률이다. 실제로 성립하지 않는 귀무가설을 검정이 얼마나 잘 가려내는지를 나타내므로, '거짓인 귀무가설을 기각할 확률'에 해당한다.

후보를 구별할 때는 귀무가설의 실제 상태와 검정의 결정을 함께 확인한다. '참인 귀무가설'을 전제로 한 두 확률은 검정력의 조건과 다르다. 귀무가설이 거짓이더라도 **기각하지 않으면 잘못된 가설을 찾아내지 못한 경우**이므로 검정력에 포함되지 않는다.

NoteBase

문 7. 두 확률변수 X와 Y의 상관계수 ρ에 대한 설명으로 항상 옳은 것만을 모두 고른 것은?

ㄱ. X와 Y가 서로 독립이면 $\rho = 0$이다.
ㄴ. $\rho = 1$이면 $Y = X$이다.
ㄷ. $X + 1$과 $\frac{Y+2}{2}$의 상관계수는 ρ이다.

① ㄱ
② ㄱ, ㄷ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 2번

간단한 해설

상관계수가 1이라는 것은 두 변수가 **양의 기울기를 가진 직선 관계**에 있다는 뜻이다. 기울기가 1이고 절편이 0이어야 하는 것은 아니므로, 반드시 $Y = X$가 성립하지는 않는다.

보기별 해설

상관계수는 $\rho = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$로, 공분산을 두 변수의 표준편차로 나눈 값이다. 문제에서 상관계수가 주어졌으므로 두 변수의 분산은 유한하고 양수여서 상관계수가 정의되는 경우를 다룬다.

ㄱ. O: 두 변수가 독립이면 $E(XY) = E(X)E(Y)$이므로 공분산은 0이다. 따라서 상관계수도 0이다. 즉, **독립이면 무상관**이다.

ㄴ. X: $\rho = 1$이면 양의 상수 a와 상수 b에 대하여 $Y = aX + b$가 확률 1로 성립한다. 이는 **양의 완전한 선형관계**를 뜻하며, 반드시 $a = 1, b = 0$인 것은 아니다. 예를 들어 분산이 양수인 X에 대해 $Y = 2X + 1$로 두면 상관계수는 1이지만 $Y = X$일 필요는 없다.

ㄷ. O: 상수를 더해도 각 값과 평균의 차이는 변하지 않으므로 공분산과 표준편차에 영향을 주지 않는다. 또한 한 변수에 양수를 곱하면 공분산과 그 변수의 표준편차에 같은 배수가 곱해져 상관계수에서 상쇄된다. 여기서 $X + 1$은 평행이동이고, $\frac{Y+2}{2} = \frac{Y}{2} + 1$은 양수 1/2배와 평행이동이므로 **상관계수의 크기와 부호가 모두 유지**된다.
따라서 변환한 두 변수의 상관계수는 ρ이다.

항상 옳은 보기는 ㄱ, ㄷ이며, 이 조합은 ②에 해당한다.

NoteBase

문 8. 어느 지역의 남녀 출생률이 같은지를 검정하기 위하여 이 지역에서 태어난 100명의 신생아를 임의로 추출하여 조사하였더니 이 중 남아가 57명이었다. 이 지역의 남녀 출생률에 대한 설명으로 옳은 것은? (단, Z가 표준정규분포를 따르는 확률변수일 때,
$P(Z \geq 1.96) = 0.025, P(Z \geq 1.645) = 0.05, P(Z \geq 1.282) = 0.1$**로 가정한다)**

① 유의수준 5%에서 남녀 출생률이 같다고 할 수 없다.

② 유의수준 5%에서 남아의 출생률이 여아의 출생률보다 더 크다고 할 수 있다.

③ 유의수준 10%에서 남아의 출생률이 여아의 출생률보다 더 크다고 할 수 있다.

④ 유의수준 10%에서 여아의 출생률이 남아의 출생률보다 더 크다고 할 수 있다.

정답 3번

간단한 해설

검정통계량은 $z = 1.4$로, 남아 출생률이 더 큰지를 보는 **우측 단측검정**의 10% 임계값 1.282보다 크다. 따라서 유의수준 10%에서는 남아 출생률이 여아 출생률보다 크다고 판단할 수 있다.

NoteBase

선지별 해설

남아의 출생률을 p라 하면 여아의 출생률은 $1-p$이므로, 남녀 출생률이 같다는 귀무가설은 $H_0 : p = 0.5$이다. 표본비율은 $\hat{p} = \frac{57}{100} = 0.57$이다. 귀무가설 아래 남아와 여아의 기대도수가 각각 50명으로 충분히 크므로 표본비율의 정규근사를 적용한다. 표준오차는 $\sqrt{\frac{0.5 \cdot 0.5}{100}} = 0.05$이고, 검정통계량은 $z = \frac{0.57 - 0.5}{0.05} = 1.4$이다. 이는 관측 비율이 귀무가설의 비율보다 표준오차의 1.4배만큼 크다는 뜻이다.

① X: 남녀 출생률이 다른지를 판단하려면 대립가설을 $p \neq 0.5$로 두는 **양측검정**을 한다. 유의수준 5%에서는 양쪽 꼬리에 각각 2.5%를 배분하므로 기각 기준은 $|z| \geq 1.96$이다. 관측값 1.4는 기각역에 속하지 않으므로, 남녀 출생률이 다르다고 할 근거가 충분하지 않다. 이는 두 출생률이 같다는 것을 입증했다는 뜻은 아니다.

② X: 남아 출생률이 더 크다는 주장은 $p > 0.5$이므로 우측 단측검정을 한다. 유의수준 5%의 기각 기준은 $z \geq 1.645$인데, 1.4는 이에 미치지 못한다. 따라서 이 유의수준에서는 **남아 출생률이 더 크다고 할 근거가 충분하지 않다**.

③ O: 대립가설 $p > 0.5$에 대한 우측 단측검정에서 유의수준 10%의 기각 기준은 $z \geq 1.282$이다. 관측값 1.4가 이 기준을 넘으므로 귀무가설을 기각하고, 남아 출생률이 여아 출생률보다 크다고 판단할 수 있다. **같은 방향의 검정이라도 유의수준이 커지면 기각역이 넓어진다**는 점에서 ②와 판단이 달라진다.

④ X: 여아 출생률이 더 크다는 것은 남아 출생률이 0.5보다 작다는 뜻이므로 대립가설을 $p < 0.5$로 두는 **좌측 단측검정**을 한다. 표준정규분포의 대칭성에 따라 유의수준 10%의 기각 기준은 $z \leq -1.282$이다. 관측값 1.4는 기각역에 속하지 않으므로, 여아 출생률이 더 크다고 할 근거가 충분하지 않다.

NoteBase

문 9. 다음은 어느 식이요법이 몸무게에 미치는 영향을 알아보기 위하여 임의로 추출한 10명의 식이요법 전후 몸무게를 측정하여 얻은 결과이다.

식이요법 전과 후의 몸무게 평균을 각각 μ_1, μ_2라 할 때, 귀무가설 $H_0 : \mu_1 = \mu_2$ 대 대립가설 $H_1 : \mu_1 \neq \mu_2$에 대한 검정의 유의확률 (p-값)은?

(단, $t_1 = \frac{73.9 - 72.4}{\sqrt{\frac{9 \times 4.01^2 + 9 \times 3.34^2}{18}} \times \sqrt{\frac{1}{10} + \frac{1}{10}}}$, $t_2 = \frac{1.5}{2.68/\sqrt{10}}$이며 T_1은 자유도가 18인 t 분포를 따르는 확률변수이며 T_2는 자유도가 9인 t 분포를 따르는 확률변수이다)

(단위 : kg)

실험대상	1	2	3	4	5	6	7	8	9	10	평균	표준편차
식이요법 전	68	76	74	71	71	72	75	83	75	74	73.9	4.01
식이요법 후	67	77	74	74	69	70	71	77	71	74	72.4	3.34
차이	1	-1	0	-3	2	2	4	6	4	0	1.5	2.68

① $P(T_1 > t_1)$

② $P(T_2 > t_2)$

③ $2P(T_1 > |t_1|)$

④ $2P(T_2 > |t_2|)$

정답 4번

간단한 해설

같은 10명의 전후 몸무게이므로 **개인별 차이를 이용하는 대응표본 t 검정**을 적용하며, 자유도는 9이다.
대립가설이 평균의 차이가 0이 아니라는 **양측검정**이므로 유의확률은 $2P(T_2 > |t_2|)$이다.

NoteBase

계산 과정

자료의 대응 관계: 식이요법 전후의 측정값은 **같은 사람끼리 짝지어진 자료**이다. 각 사람의 차이를 D_i=식이요법 전 몸무게-식이요법 후 몸무게로 두면, 검정할 모평균 차이는 $\mu_D = \mu_1 - \mu_2$이다. 표의 '차이' 행에서 차이의 표본평균은 1.5 kg, 표본표준편차는 2.68 kg이며, 표본 크기는 10이다. 전후를 독립된 두 표본으로 취급하는 t_1은 이 자료의 대응 관계를 반영하지 못한다.

검정통계량과 자유도: 사람 간 차이가 서로 독립이고 차이의 모집단이 정규분포를 따른다는 가정 아래 대응표본 t 검정을 적용한다. 귀무가설은 $H_0 : \mu_D = 0$이므로 검정통계량은 $t = \frac{\bar{D} - 0}{s_D/\sqrt{n}}$이다. 분모 $s_D/\sqrt{n}$은 **차이의 표본평균의 표준오차**를 뜻한다. 주어진 값을 대입하면 $t_2 = \frac{1.5}{2.68/\sqrt{10}} \approx 1.77$이다. 10개의 차이에서 평균 하나를 추정하므로 자유도는 $10 - 1 = 9$이며, 귀무가설 아래 검정통계량은 문제의 T_2와 같은 분포를 따른다.

양측 유의확률: 대립가설 $H_1 : \mu_D \neq 0$은 몸무게의 감소와 증가를 모두 포함한다. 따라서 p-값은 귀무가설 아래에서 검정통계량의 절댓값이 관측된 $|t_2|$ 이상일 확률이다. t 분포는 0을 중심으로 대칭이므로 **양쪽 꼬리의 확률**은 오른쪽 꼬리 확률의 두 배이다. 따라서 $p = 2P(T_2 > |t_2|)$이며, 이는 ④의 식에 해당한다. $P(T_2 > t_2)$만 사용하면 한쪽 꼬리만 반영한다.

NoteBase

문 10. Σ_i=1^n(X_i-X)^2=80, Σ_i=1^n(Y_i-Y)^2=360인 두 변수 X와 Y에 대하여 단순회귀모형 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ i = 1, 2, \cdots, n$에서 추정된 회귀직선이 $\hat{Y}_i = 0.8 + 2.0 \times X_i$일 때, 잔차제곱합 Σ_i=1^n(Y_i-Y_i)^2은? (단, X=(1)/(n)Σ_i=1^nX_i, Y=(1)/(n)Σ_i=1^nY_i이다)

① 40

② 160

③ 200

④ 280

정답 1번

간단한 해설

절편을 포함한 최소제곱 회귀에서는 **총제곱합 = 회귀제곱합 + 잔차제곱합**이 성립한다. 기울기가 2.0이므로 회귀제곱합은 $2.0^2 \cdot 80 = 320$이고, 잔차제곱합은 360 − 320 = 40이다.

계산 과정

총제곱합 확인: 총제곱합 T는 관측값이 표본평균에서 벗어난 정도의 제곱합으로, $T = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = 360$이다. **절편을 포함한 최소제곱 회귀**에서는 $T = R + E$가 성립한다. 여기서 R은 회귀직선이 설명하는 변동인 회귀제곱합이고, E는 회귀직선이 설명하지 못하는 변동인 잔차제곱합이다.

회귀제곱합 계산: 절편을 포함한 최소제곱 회귀직선은 표본평균점 $(\overline{X}, \overline{Y})$을 지나므로, 주어진 기울기 2.0을 이용하면 $\hat{Y}_i - \overline{Y} = 2.0(X_i - \overline{X})$이다. 따라서 회귀제곱합 $R = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$은 $R = 2.0^2 \cdot 80 = 320$이다. 편차를 제곱하여 더하므로 **기울기도 제곱하여 곱한다**. 절편 0.8은 평균을 뺄 때 소거된다.

잔차제곱합 계산: 총제곱합에서 회귀직선이 설명하는 변동을 빼면 잔차제곱합이 된다. $E = T - R$이므로 360 − 320 = 40이며, 이는 ①의 40과 일치한다.

NoteBase

문 11. 다음은 18명의 지능지수(IQ)를 측정한 자료를 줄기-잎 그림으로 정리한 것이다. 이 자료의 중앙값은?

줄기	잎
7	8
8	59
9	2899
10	14458
11	689
12	34
13	8

① 102
② 103
③ 104
④ 105

정답 3번

간단한 해설

자료가 18개이므로 중앙값은 크기순으로 나열했을 때 **9번째와 10번째 값의 평균**이다. 두 값이 모두 104이므로 중앙값은 104이다.

NoteBase

계산 과정

줄기-잎 판독: 줄기는 일의 자리를 제외한 앞부분이고 잎은 일의 자리이다. 예를 들어 줄기 7의 잎 8은 78, 줄기 10의 잎 4는 104를 뜻한다. 각 행의 잎 수를 더하면 자료는 1 + 2 + 4 + 5 + 3 + 2 + 1 = 18개이다.

가운데 순위 확인: 중앙값은 자료를 크기순으로 나열했을 때 가운데에 놓이는 값이다. 자료 수가 짝수인 18개이므로 **9번째와 10번째 값의 평균**을 구한다. 줄기 7·8·9에 속한 자료가 모두 7개이므로, 줄기 10의 잎 1·4·4·5·8은 각각 8번째부터 12번째 자료에 해당한다. 따라서 9번째와 10번째 값은 모두 104이다.

중앙값 계산: $\frac{104+104}{2} = 104$이다. 따라서 이 자료의 중앙값은 IQ 104이며, ③의 104에 해당한다.

NoteBase

문 12. 음료를 판매하는 회사에서 나온 어느 제품의 한 개당 용량은 평균이 100 ml, 표준편차가 10 ml인 정규분포를 따른다고 할 때, 이 제품에서 임의로 추출한 25개의 표본평균이 97 ml 이상 102 ml 이하일 확률은? (단, 아래의 표는 Z가 표준정규분포를 따르는 확률변수일 때 P(Z≤z)의 값에 상응하는 z의 값을 나타낸 것이다)

$P(Z \leq z)$	z
0.8413	1.0
0.9332	1.5
0.9772	2.0

① 0.6826

② 0.7745

③ 0.8185

④ 0.9104

정답 2번

간단한 해설

표본평균의 표준편차는 개별 제품의 표준편차를 표본 크기의 제곱근으로 나눈 **2 ml**이다. 따라서 97 ml 이상 102 ml 이하인 구간을 표준화하면 $-1.5 \leq Z \leq 1$이고, 표의 누적확률과 대칭성을 이용하면 그 확률은 0.7745이다.

NoteBase

계산 과정

표본평균의 분포: 25개 제품의 용량을 독립인 표본으로 보고 그 평균을 $\bar{X}$라 하자. 모집단이 정규분포이므로 **표본평균도 정규분포**를 따른다. 표본평균의 평균은 모집단 평균과 같은 100 ml이고, 표준편차인 표준오차는 $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2$ ml이다. 이는 개별 제품의 용량이 아니라 25개를 묶어 구한 평균의 변동 크기이다.

구간의 표준화: 표준정규분포표를 이용하려면 평균 100을 빼고 **표본평균의 표준편차 2**로 나눈다. $Z = \frac{\bar{X} - 100}{2}$이므로, 아래 경계 97 ml는 $\frac{97 - 100}{2} = -1.5$, 위 경계 102 ml는 $\frac{102 - 100}{2} = 1$에 대응한다. 따라서 구하려는 확률은 $P(-1.5 \le Z \le 1)$이다.

누적확률의 차: 표에서 $P(Z \le 1) = 0.8413$, $P(Z \le 1.5) = 0.9332$이다. 표준정규분포는 0을 중심으로 대칭이므로 $P(Z \le -1.5) = 1 - 0.9332 = 0.0668$이다. **위 경계까지의 누적확률에서 아래 경계까지의 누적확률을 빼면** 두 경계 사이의 확률을 얻는다. 연속분포에서는 한 점의 확률이 0이므로 경계의 포함 여부는 계산값에 영향을 주지 않는다.

$P(97 \le \bar{X} \le 102) = 0.8413 - 0.0668 = 0.7745$이므로, 구한 확률은 ②의 0.7745와 일치한다.

NoteBase

문 13. 평균이 μ, 분산이 σ^2인 모집단에서 추출한 임의표본(random sample)의 표본평균에 대한 설명으로 옳은 것만을 모두 고른 것은? (단, $0 < \sigma^2 < \infty$이다)

ㄱ. 모집단이 정규분포를 따를 때 표본평균의 분포는 정규분포이다.
ㄴ. 표본의 크기가 커질수록 표본평균의 분산은 커진다.
ㄷ. 표본평균은 모평균의 불편추정량이다.

① ㄱ
② ㄱ, ㄷ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 2번

간단한 해설

표본평균의 분산은 $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$이므로, 표본의 크기가 커질수록 **분산은 작아진다**. 표본을 더 많이 평균하면 모평균 주위의 변동이 줄어든다는 뜻이다.

보기별 해설

임의표본 $X_1, \ldots, X_n$은 서로 독립이고 같은 모집단 분포를 따르는 확률변수들이다. 표본의 크기를 n이라 하면 표본평균은 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$이다.

ㄱ. O: 서로 독립인 정규확률변수의 합과 평균은 정규분포를 따른다. 따라서 모집단이 정규분포이면 $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$이다. 이때는 **표본의 크기가 작아도 정확히 정규분포**를 따른다.

ㄴ. X: 각 표본의 분산은 σ^2이고 서로 독립이므로 표본합의 분산은 $n\sigma^2$이다. 확률변수를 n으로 나누면 분산은 n^2으로 나누어지므로, $\mathrm{Var}(\bar{X}) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$이다. 모분산이 일정한 양수일 때 표본의 크기가 커질수록 표본평균의 분산은 **커지는 것이 아니라 작아진다**. 이 관계에는 모집단의 정규성도 필요하지 않다.

ㄷ. O: 각 표본의 기댓값이 μ이므로 그 평균의 기댓값도 $E(\bar{X}) = \mu$이다. 추정량의 기댓값이 추정하려는 모수와 같으면 불편추정량이라고 한다. 따라서 표본평균은 **모평균의 불편추정량**이며, 이 성질도 모집단의 정규성과 관계없이 성립한다.

옳은 보기는 ㄱ과 ㄷ이므로 ②의 조합에 해당한다.

NoteBase

문 14. 다음은 어느 지역에서 80명을 임의로 추출하여 출생한 계절을 조사한 결과이다.
이 지역의 계절별 출생률이 같다는 귀무가설에 대한 검정에서 카이제곱 검정통계량의 값과 유의수준 5%에서의 검정결과에 대한 설명으로 옳은 것은? (단, $\chi^2_\alpha(k)$는 자유도가 k인 카이제곱분포의 $(1-\alpha)\times 100$번째 백분위수를 나타내고, $\chi^2_{0.05}(3)=7.8147$, $\chi^2_{0.05}(4)=9.4877$이다.)

출생계절	봄	여름	가을	겨울
관측도수	28	12	16	24

① 검정통계량의 값이 8로 계절별 출생률이 같다고 할 수 있다.

② 검정통계량의 값이 8로 계절별 출생률이 같다고 할 수 없다.

③ 검정통계량의 값이 9로 계절별 출생률이 같다고 할 수 있다.

④ 검정통계량의 값이 9로 계절별 출생률이 같다고 할 수 없다.

정답 2번

간단한 해설

계절별 출생률이 같다면 각 계절의 기대도수는 20명이며, 이를 기준으로 구한 **카이제곱 검정통계량은 8**이다.
자유도 3의 5% 임계값 7.8147보다 크므로 **귀무가설을 기각**하여 계절별 출생률이 같다고 할 수 없다.

NoteBase

계산 과정

기대도수: 표의 관측도수는 봄 28명, 여름 12명, 가을 16명, 겨울 24명으로 총 80명이다. 귀무가설은 네 계절의 출생확률이 모두 1/4이라는 것이고, 대립가설은 적어도 한 계절의 출생확률이 다르다는 것이다. 따라서 귀무가설 아래 각 계절의 기대도수는 $E_i = 80 \times \frac{1}{4} = 20$명이다.

검정통계량: 관측도수가 가정한 확률분포에 부합하는지 확인하므로 카이제곱 적합도검정을 이용한다. 각 계절의 관측도수를 O_i, 기대도수를 E_i라고 하면 검정통계량은 $\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$이다. 이는 관측도수와 기대도수의 차이를 제곱한 뒤 **기대도수로 나누어 합한 값**이므로, 차이가 클수록 커진다. 봄·여름·가을·겨울의 차이는 각각 8, −8, −4, 4이므로 $\chi^2 = \frac{64 + 64 + 16 + 16}{20} = 8$이다. 검정통계량에는 단위가 없다.

자유도와 기각 기준: 네 계절의 도수 합계가 80명으로 고정되어 있어 세 계절의 도수가 정해지면 나머지 하나도 정해진다. 출생확률은 귀무가설에서 이미 주어져 별도로 추정한 모수가 없으므로 **자유도는 4 − 1 = 3**이다. 모든 기대도수가 20으로 충분히 크므로 카이제곱분포를 이용한 근사가 적절하다. 유의수준 5%에서는 $\chi^2 > \chi^2_{0.05}(3) = 7.8147$일 때 귀무가설을 기각한다. 범주가 네 개라는 이유로 자유도 4의 임계값 9.4877을 쓰면 안 된다.

검정결과: $8 > 7.8147$이므로 귀무가설을 기각한다. 관측된 계절별 차이는 유의수준 5%에서 통계적으로 유의하므로, 계절별 출생률이 같다고 할 수 없다. 검정통계량 8과 이 검정결과를 함께 제시한 것은 ②이다.

NoteBase

문 15. X와 Y가 확률변수이고 a와 b가 상수일 때, 이에 대한 설명으로 옳지 않은 것은?

① $E(aX+b)=aE(X)+b$

② $Var(aX+b)=a^2Var(X)$

③ $Cov(aX,bY)=ab\times Cov(X,Y)$

④ X와 Y가 서로 독립이면 $Var(X-Y)=Var(X)-Var(Y)$이다.

정답 4번

간단한 해설

서로 독립인 확률변수는 차이의 분산을 구할 때도 **각 분산을 더한다**. 따라서 ④의 분산을 빼는 식은 $\mathrm{Var}(X-Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)$로 고쳐야 한다.

선지별 해설

기댓값·분산·공분산은 해당 값이 존재하는 경우를 전제로 한다.

① O: 기댓값은 확률변수의 값을 확률에 따라 가중평균한 것이다. 기댓값의 선형성에 따라 $\mathrm{E}(aX+b)=a\mathrm{E}(X)+b$이다. 모든 값에 상수를 곱하거나 더하면 **평균에도 같은 변환**이 적용된다.

② O: 분산은 평균으로부터의 편차를 제곱하여 평균한 값이다. 상수 b를 더하면 값과 평균이 함께 이동하므로 편차는 변하지 않고, a를 곱하면 편차도 a배가 된다. 따라서 분산은 **a의 제곱배**가 되어 $\mathrm{Var}(aX+b)=a^2\mathrm{Var}(X)$이다.

③ O: 공분산은 두 확률변수 각각의 평균으로부터의 편차를 곱한 뒤 기댓값을 취한 것이다. X에 a를 곱하면 X의 편차가 a배, Y에 b를 곱하면 Y의 편차가 b배가 되므로 편차의 곱은 ab배가 된다. 따라서 $\mathrm{Cov}(aX,bY)=ab\mathrm{Cov}(X,Y)$이며, **독립 여부와 관계없이** 성립한다.

④ X: 두 확률변수 차이의 분산은 두 분산의 합에서 공분산의 2배를 뺀 값이다. X와 Y가 서로 독립이면 $\mathrm{Cov}(X,Y)=0$이므로 $\mathrm{Var}(X-Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)$이다. Y의 계수가 −1이어도 분산에는 그 제곱이 적용되므로 **Y의 분산을 빼지 않고 더한다**.

NoteBase

문 16. 다음은 회사 A, B, C에서 생산하는 직물의 인장강도를 비교하기 위하여 각 회사별로 제품 4개씩을 임의로 추출하여 실험한 결과이다. 이 결과를 이용하여 일원배치 분산분석표를 작성할 때, ㉠과 ㉡의 값은?

(단위 : kg/cm^2)

회사	A	B	C
인장강도	55	52	49
	54	50	51
	52	51	52
	55	51	52
표본평균	54	51	51
표본분산	2	2/3	2

요인	제곱합	자유도	평균제곱	F 값
처리	㉠			
잔차	㉡			
계				

① ㉠: 6, ㉡: 10

② ㉠: 6, ㉡: 14

③ ㉠: 24, ㉡: 10

④ ㉠: 24, ㉡: 14

정답 4번

간단한 해설

처리제곱합은 **회사별 평균과 전체평균의 차이**로 계산하므로 ㉠은 24이다. 잔차제곱합은 **각 표본분산에 회사별 자유도 3을 곱해 합한 값**이므로 ㉡은 14이다.

NoteBase

계산 과정

전체평균: 회사별 표본 수가 모두 4개이므로 전체평균은 회사별 표본평균 54, 51, 51의 단순평균이다. 따라서 $\bar{x} = \frac{54 + 51 + 51}{3} = 52$이며, 단위는 kg/cm^2이다.

㉠ 처리제곱합: 회사 사이의 변동을 나타내므로, 각 회사의 평균과 전체평균의 차이를 제곱한 뒤 **해당 회사의 표본 수를 곱하여** 합한다. 회사별 평균에서 전체평균 52를 뺀 편차는 각각 2, −1, −1이고, 회사별 표본 수는 4개이다. 따라서 SS_처리=4\2^2+(-1)^2+(-1)^2\이다. 즉, $4 \times (4 + 1 + 1) = 24$이다.

㉡ 잔차제곱합: 각 회사 안에서 관측값이 그 회사의 평균으로부터 흩어진 정도를 합한 값이다. 표본분산은 편차제곱합을 **표본 수에서 1을 뺀 자유도**로 나눈 값이므로, 각 회사의 편차제곱합은 표본분산에 $4 - 1 = 3$을 곱해 구한다. 따라서 SS_잔차=3·2+3·(2)/(3)+3·2이고, 회사별 편차제곱합은 각각 6, 2, 6이므로 잔차제곱합은 6 + 2 + 6 = 14이다.

두 제곱합의 단위는 모두 인장강도 단위의 제곱인 $(kg/cm^2)^2$이다.

㉠은 24, ㉡은 14로, ④의 값과 일치한다.

NoteBase

문 17. 표본상관계수 r에 대한 설명으로 옳은 것만을 모두 고른 것은?

ㄱ. r가 −1이면 산점도에서 모든 관측값은 일직선 위에 있다.
ㄴ. r가 −0.5이면 단순회귀분석에서 결정계수는 0.25이다.
ㄷ. r가 양수이면 단순회귀분석에서 기울기의 추정값도 양수이다.

① ㄱ, ㄴ
② ㄱ, ㄷ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 4번

간단한 해설

단순회귀에서 **상관계수와 기울기의 부호는 같고**, 결정계수는 상관계수의 제곱인 $R^2 = r^2$이다. 따라서 음의 상관관계라도 결정계수는 양수가 될 수 있다.

보기별 해설

여기서 단순회귀는 절편을 포함하여 최소제곱법으로 적합하는 경우이다. 표본상관계수는 두 변수 사이의 선형관계의 방향과 강도를 나타낸다.

ㄱ. O: $r = -1$은 **완전한 음의 선형관계**를 뜻한다. 한 변수가 증가할 때 다른 변수가 일정한 비율로 감소하므로, 산점도의 모든 관측점은 기울기가 음수인 하나의 직선 위에 놓인다.

ㄴ. O: 절편을 포함한 최소제곱 단순회귀에서는 **결정계수가 상관계수의 제곱**과 같다. 따라서 $R^2 = r^2 = (-0.5)^2 = 0.25$이다. 이는 종속변수의 전체 변동 중 25%를 회귀모형이 설명한다는 뜻이다. 상관계수의 음의 부호는 관계의 방향을 나타내며, 제곱하면 사라진다.

ㄷ. O: 설명변수와 종속변수의 표본표준편차를 각각 s_x, s_y라 하면, 기울기의 최소제곱 추정값은 $\hat{\beta}_1 = r\frac{s_y}{s_x}$이다. 상관계수가 정의되려면 두 표준편차가 모두 양수이므로, 표준편차의 비율도 양수이다. 따라서 **기울기의 추정값은 상관계수와 같은 부호**를 가지며, $r > 0$이면 기울기의 추정값도 양수이다.

ㄱ, ㄴ, ㄷ이 모두 성립하므로 이들을 모두 포함한 조합은 ④이다.

NoteBase

문 18. 어느 공장에서 생산되는 제품의 불량률이 5%라고 한다. 이 공장에서 생산되는 제품을 임의로 추출하여 추출된 순서대로 조사할 때 10번째 제품이 첫 번째 불량품일 확률은?

① 0.05×0.95^{9}

② 0.05×0.95^{10}

③ $\binom{10}{1} \times 0.05^{9} \times 0.95$

④ $\binom{10}{1} \times 0.05 \times 0.95^{9}$

정답 1번

간단한 해설

10번째 제품이 첫 불량품이려면 **앞의 9개는 모두 양품이고 10번째만 불량품**이어야 한다. 독립 시행으로 보면 그 확률은 $0.95^{9} \cdot 0.05$이며, 불량품의 위치가 정해져 있어 조합계수를 곱하지 않는다.

계산 과정

각 제품의 불량 여부가 서로 독립이고 불량 확률이 0.05로 일정한 시행으로 모형화한다. 제품 하나가 양품일 확률은 $1 - 0.05 = 0.95$이다.

첫 불량품이 나오는 순서를 T라 하자. $T = 10$이 되려면 **1번째부터 9번째까지 모두 양품**이고 10번째 제품은 불량품이어야 한다. 앞의 9개가 모두 양품일 확률은 0.95^{9}이고, 10번째가 불량품일 확률은 0.05이다.

독립인 사건들이 함께 일어날 확률은 각 확률을 곱하므로 $P(T = 10) = 0.95^{9} \cdot 0.05$이다. 이는 ①의 식과 같다. 여기서는 **불량품의 위치가 10번째로 고정**되어 있으므로 조합계수를 곱하지 않는다. ④처럼 $\binom{10}{1}$을 곱하면 위치에 관계없이 10개 중 불량품이 정확히 1개일 확률이 된다.

NoteBase

문 19. 두 변수 X와 Y에 대한 10개의 관측쌍 $(x_1, y_1), (x_2, y_2), \cdots, (x_{10}, y_{10})$에 대하여 x_i와 y_i를 다음과 같이 표준화하였다.
$z_{1i} = \frac{x_i - \bar{x}}{s_x}, \quad z_{2i} = \frac{y_i - \bar{y}}{s_y}, \quad i = 1, 2, \cdots, 10.$ **Σ _i=1^10z_1iz_2i=0.9일 때, X와 Y의 표본상관계수는? (단, x=Σ _i=1^10x_i/10, y=Σ _i=1^10y_i/10, s_x=Σ _i=1^10(x_i-x)^2/9, s_y=Σ _i=1^10(y_i-y)^2/9이다.)**

① 0.1

② -0.1

③ 0.9

④ -0.9

정답 1번

간단한 해설

표본표준편차를 사용해 표준화했으므로, 표준화한 두 값의 곱을 모두 더한 뒤 **표본 크기에서 1을 뺀 9로 나누어야** 한다. 따라서 표본상관계수는 $r = \frac{0.9}{9} = 0.1$이다.

계산 과정

관측쌍이 10개이므로 표본 크기는 n = 10이다. 주어진 표본표준편차는 편차제곱합을 **n − 1 = 9로 나누어** 구한 것이다. 같은 기준으로 표본공분산을 구하면 $s_{xy} = \frac{1}{9}\sum_{i=1}^{10}(x_i - \bar{x})(y_i - \bar{y})$이다.

표본상관계수는 표본공분산을 두 표본표준편차의 곱으로 나눈 값이므로 $r = \frac{s_{xy}}{s_x s_y}$이다. 각 편차를 해당 표준편차로 나눈 값이 z_{1i}, z_{2i}이므로, 이를 대입하면 $r = \frac{1}{9}\sum_{i=1}^{10} z_{1i}z_{2i}$가 된다. 주어진 0.9는 **표준화한 값들의 곱의 합**이므로, 그 자체가 상관계수는 아니다.

주어진 합을 대입하면 $r = \frac{0.9}{9} = 0.1$이다. 표본상관계수는 단위가 없는 값이며, 계산 결과는 ①의 0.1에 해당한다.

NoteBase

문 20. 단순회귀모형 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ \ i = 1, 2, \cdots, n$에서 최소제곱법으로 구한 추정회귀식이 $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$이고, 잔차가 $e_i = Y_i - \hat{Y}_i$일 때, 이에 대한 설명으로 옳은 것만을 모두 고른 것은? (단, X=(1)/(n)Σ_i=1^nX_i, Y=(1)/(n)Σ_i=1^nY_i이다)

ㄱ. $\overline{Y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{X}$
ㄴ. Σ_i=1^ne_i=0
ㄷ. Σ_i=1^nY_i e_i=0

① ㄱ, ㄴ
② ㄱ, ㄷ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 4번

간단한 해설

절편을 포함한 최소제곱 회귀에서는 **정규방정식**에 따라 잔차의 합과 설명변수·잔차의 곱의 합이 모두 0이다. 이로부터 회귀직선은 표본평균점을 지나며, 적합값과 잔차의 곱의 합도 0이 된다.

보기별 해설

최소제곱법은 잔차제곱합 $\sum_{i=1}^{n} e_i^2$을 최소로 하는 계수를 구한다. 절편과 기울기에 대해 각각 미분하여 0으로 놓으면 정규방정식 $\sum_i e_i = 0$, $\sum_i X_i e_i = 0$을 얻는다. 이하 모든 합의 범위는 $i = 1, \ldots, n$이다.

ㄱ. O: 첫 번째 정규방정식에 잔차의 정의를 대입하면 $\sum_i Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_i X_i$이다. 양변을 n으로 나누면 $\overline{Y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{X}$를 얻는다. 즉, 추정회귀직선은 **표본평균점** $(\overline{X}, \overline{Y})$를 지난다.

ㄴ. O: **절편을 포함하여 추정**하므로 절편에 대한 최소화 조건이 $\sum_i e_i = 0$이다. 따라서 관측값과 적합값의 차이인 잔차를 모두 더하면 0이며, 잔차의 표본평균도 0이다.

ㄷ. O: 적합값은 $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$이므로, 잔차와 곱하여 합하면 $\sum_i \hat{Y}_i e_i = \hat{\beta}_0 \sum_i e_i + \hat{\beta}_1 \sum_i X_i e_i$이다. 두 정규방정식에 의해 오른쪽의 두 합이 모두 0이므로 $\sum_i \hat{Y}_i e_i = 0$이다. 이는 **적합값 벡터와 잔차 벡터가 직교**한다는 뜻이다.

ㄱ, ㄴ, ㄷ이 모두 성립하므로 이들을 모두 포함한 ④에 해당한다.

NoteBase