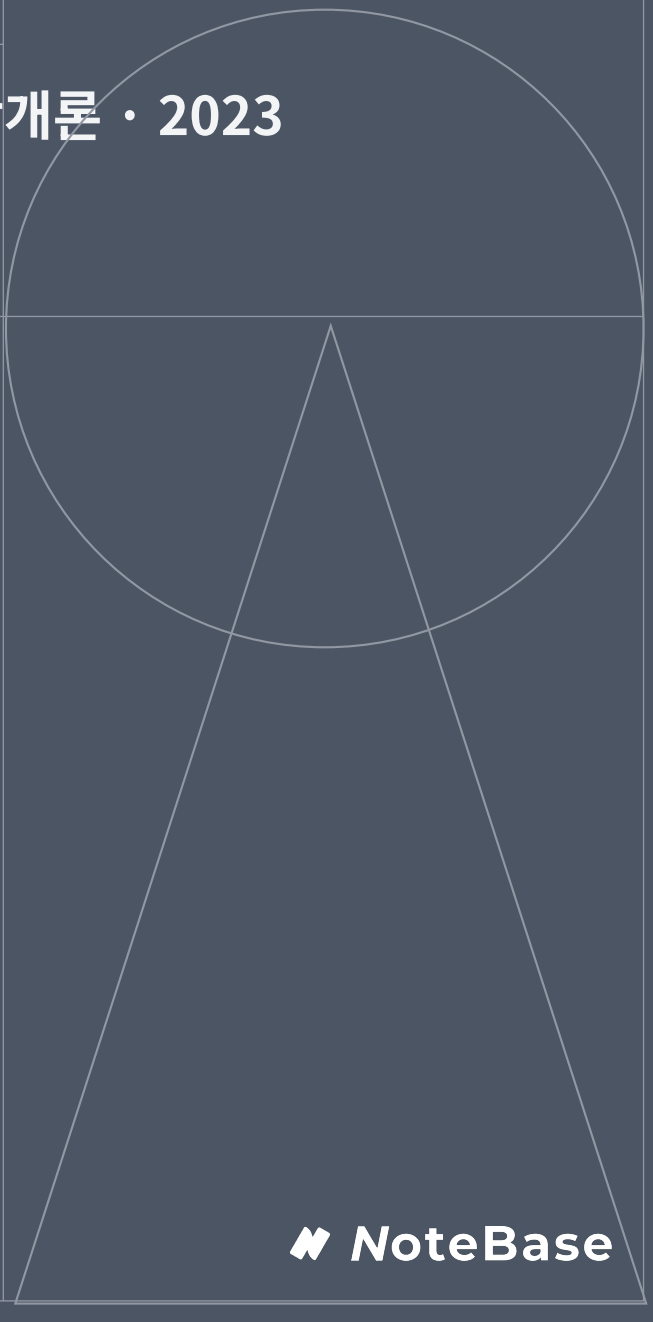
기출OX 해설집

공무원/국가직/9급 · 통계학개론 · 2023



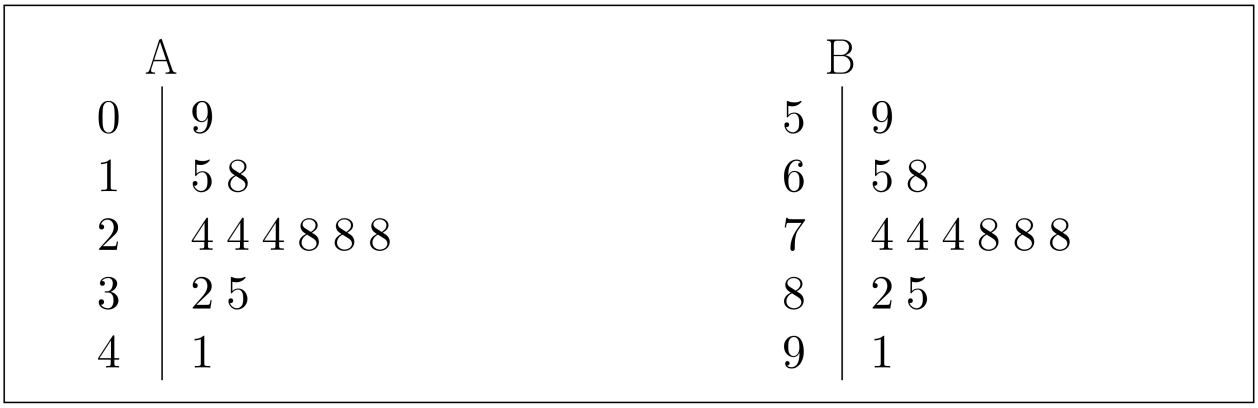
NoteBase

기출OX 해설집

공무원/국가직/9급 · 통계학개론 · 2023

공무원/국가직/9급 · 통계학개론 · 2023

문 1. 두 자료 A, B의 줄기-잎 그림에 대한 설명으로 옳지 않은 것은?



A		B	
0	9	5	9
1	5 8	6	5 8
2	4 4 4 8 8 8	7	4 4 4 8 8 8
3	2 5	8	2 5
4	1	9	1

① 변동계수는 서로 같다.

② 범위는 서로 같다.

③ 사분위수범위는 서로 같다.

④ 표준편차는 서로 같다.

정답 1번

간단한 해설

B는 A의 모든 값에 50을 더한 자료이므로 표준편차는 같지만 평균은 더 크다. 변동계수는 **표준편차를 평균으로 나눈 값**이므로, **B의 변동계수가 A보다 작다**.

NoteBase

선지별 해설

줄기-잎 그림에서 A의 9, 15, 18 등이 B에서는 각각 59, 65, 68 등으로 나타난다. 두 그림은 잎의 배열이 같고 B의 줄기가 A보다 각각 5 크므로, B는 A의 모든 값에 50을 더한 자료이다.

① X: 변동계수는 평균에 대한 표준편차의 비이다. 평균을 $\bar{x}$, 표준편차를 s라 하면 $CV = \frac{s}{\bar{x}}$이다. 두 자료의 표준편차는 같고 0보다 크지만, B의 평균은 A의 평균보다 50 크다. 두 평균이 모두 양수이므로 분모가 더 큰 **B의 변동계수가 A보다 작다**. 따라서 '서로 같다'는 설명을 이렇게 고쳐야 한다.

② O: 범위는 **최댓값에서 최솟값을 뺀 값**이다. A의 범위는 41 − 9 = 32이고, B의 범위도 91 − 59 = 32이다. 최댓값과 최솟값에 같은 수를 더하면 그 차이는 변하지 않는다.

③ O: 사분위수범위는 제3사분위수에서 제1사분위수를 뺀 값으로, 가운데 50% 자료가 퍼진 정도를 나타낸다. 모든 값에 50을 더해도 자료의 크기 순서는 유지되므로, **두 사분위수가 모두 50씩 증가**한다. 따라서 두 사분위수의 차이인 사분위수범위는 같다.

④ O: 표준편차는 각 자료가 평균에서 얼마나 떨어져 있는지를 나타낸다. A의 각 값에 50을 더하면 평균도 50 증가하므로 **각 값과 평균의 차이는 그대로**이다. 이 차이를 제곱하여 구하는 분산이 변하지 않으므로, 분산의 제곱근인 표준편차도 같다.

NoteBase

문 2. 두 확률변수 X와 Y의 결합확률분포가 다음과 같을 때, 이에 대한 설명으로 옳은 것은?

Y \ X	0	2
0	$\frac{1}{8}$	$\frac{1}{8}$
2	$\frac{3}{8}$	$\frac{3}{8}$

① $E(X)=\frac{3}{4}$

② $Var(X)=\frac{1}{2}$

③ $E(X+Y)=\frac{3}{2}$

④ $P(X+Y=2)=\frac{1}{2}$

정답 4번

간단한 해설

$X+Y=2$가 되는 경우는 $(X,Y)=(0,2)$**와** $(2,0)$이다. 두 경우의 결합확률을 더하면 $\frac{3}{8}+\frac{1}{8}=\frac{1}{2}$이다.

NoteBase

선지별 해설

표의 열은 X의 값, 행은 Y의 값을 나타내며, 각 칸은 두 값이 동시에 나타날 결합확률이다. 한 변수의 특정 값에 대한 확률은 해당 열 또는 행의 결합확률을 더해 구한다.

① X: 표의 각 열에서 확률을 더하면 $P(X=0)=\frac{1}{2}$, $P(X=2)=\frac{1}{2}$이다. 기댓값은 **가능한 값에 그 값의 확률을 곱해 더한 값**이므로 $E(X)=0\cdot\frac{1}{2}+2\cdot\frac{1}{2}=1$이다. 따라서 선지의 기댓값은 3/4이 아니라 1이다.

② X: $E(X^2)=0^2\cdot\frac{1}{2}+2^2\cdot\frac{1}{2}=2$이다. 분산은 **제곱의 기댓값에서 기댓값의 제곱을 뺀 값**이므로 $Var(X)=E(X^2)-[E(X)]^2$에 따라 $Var(X)=2-1^2=1$이다. 선지의 분산은 1/2이 아니라 1이다.

③ X: 표의 각 행에서 확률을 더하면 $P(Y=0)=\frac{1}{4}$, $P(Y=2)=\frac{3}{4}$이다. 따라서 $E(Y)=0\cdot\frac{1}{4}+2\cdot\frac{3}{4}=\frac{3}{2}$이다. 기댓값의 선형성은 **독립 여부와 관계없이** 성립하므로 $E(X+Y)=E(X)+E(Y)$이다. 이를 적용하면 $E(X+Y)=1+\frac{3}{2}=\frac{5}{2}$이므로 선지의 3/2은 Y만의 기댓값이다.

④ O: X와 Y가 각각 0 또는 2를 취하므로 합이 2가 되는 경우는 $(X,Y)=(0,2)$와 $(2,0)$이다. 표에서 Y가 2인 행과 X가 0인 열이 만나는 칸의 확률은 3/8이고, Y가 0인 행과 X가 2인 열이 만나는 칸의 확률은 1/8이다. 두 경우는 **서로 겹치지 않으므로 확률을 더한다**. 따라서 $P(X+Y=2)=\frac{3}{8}+\frac{1}{8}=\frac{1}{2}$이다.

NoteBase

문 3. 정규분포를 따르는 확률변수 X에 대하여 $P(X \le 2) = 0.2$이고 $P(X \le 12) = 0.8$일 때, $P(2 \le X \le 7)$의 값은?

① 0.35

② 0.3

③ 0.25

④ 0.2

정답 2번

간단한 해설

정규분포는 **평균을 중심으로 좌우 대칭**이므로, 누적확률이 0.2와 0.8인 두 값 2와 12의 중점 7이 평균이다. 따라서 $P(X \le 7) = 0.5$이고, 구하는 확률은 $0.5 - 0.2 = 0.3$이다.

계산 과정

평균 구하기: 정규분포의 평균을 μ라 하자. 주어진 조건에서 $P(X \le 2) = 0.2$이고 $P(X > 12) = 1 - 0.8 = 0.2$이므로, 2의 왼쪽과 12의 오른쪽 꼬리확률이 같다. 정규분포는 평균을 중심으로 대칭이므로 **같은 꼬리확률에 대응하는 두 값은 평균에서 같은 거리**에 있다. 따라서 $\mu = \frac{2+12}{2} = 7$이다.

평균까지의 누적확률: 정규분포에서 평균의 왼쪽과 오른쪽 확률은 각각 0.5이다. 평균이 7이므로 $P(X \le 7) = 0.5$이다.

구간확률 계산: 정규분포는 연속분포이므로 $P(X = 2) = 0$이다. 따라서 **7까지의 누적확률에서 2까지의 누적확률을 빼면** 2 이상 7 이하의 확률을 얻는다. $P(2 \le X \le 7) = P(X \le 7) - P(X \le 2)$이므로, 구하는 값은 $0.5 - 0.2 = 0.3$이다.

계산한 확률 0.3은 ②의 값과 일치한다.

NoteBase

문 4. 두 확률변수 X와 Y가 각각 성공의 확률이 0.3과 0.7인 베르누이 분포를 따르고 서로 독립일 때, $P(X \neq Y)$의 값은?

① 0.09

② 0.21

③ 0.49

④ 0.58

정답 4번

간단한 해설

두 값이 다르려면 **하나는 1이고 다른 하나는 0**이어야 한다. 독립인 두 변수에 대해 각 경우의 확률을 곱해서 구한 뒤 더하면 $P(X \neq Y) = 0.09 + 0.49 = 0.58$이다.

계산 과정

베르누이 확률변수는 성공하면 1, 실패하면 0의 값을 갖는다. 따라서 $P(X = 1) = 0.3$, $P(X = 0) = 0.7$이고, $P(Y = 1) = 0.7$, $P(Y = 0) = 0.3$이다. 두 값이 다른 경우는 $(X, Y) = (1, 0)$과 $(0, 1)$ 두 가지이다.

첫째, X만 성공하는 경우이다. 두 변수가 **서로 독립이므로 결합확률은 각 확률의 곱**이다. 따라서 $P(X = 1, Y = 0) = 0.3 \cdot 0.3 = 0.09$이다.

둘째, Y만 성공하는 경우이다. 같은 독립성에 따라 $P(X = 0, Y = 1) = 0.7 \cdot 0.7 = 0.49$이다.

두 경우는 **동시에 일어날 수 없는 사건**이므로 확률을 더한다. 따라서 $P(X \neq Y) = 0.09 + 0.49 = 0.58$이며, 이는 ④의 값과 일치한다.

NoteBase

문 5. 어느 회사에서 경력에 따라 연봉에 차이가 있는지 알아보기 위해 1년 미만, 1년 이상 3년 미만, 3년 이상의 경력을 가진 직원을 각각 100명씩 임의추출하여 연봉(단위: 만 원)을 조사한 결과 다음과 같은 분할표를 얻었다.
"경력에 따른 연봉에 차이는 없다."라는 가정하에서 구한 기대도수에 대한 설명으로 옳지 않은 것은?

경력 \ 연봉	3,000 미만	3,000 이상 4,000 미만	4,000 이상 5,000 미만	5,000 이상
1년 미만	30	50	20	0
1년 이상 3년 미만	20	30	40	10
3년 이상	10	10	30	50

① 기대도수와 관측도수 차의 절댓값의 최솟값은 0이다.

② 기대도수와 관측도수 차의 절댓값의 최댓값은 30이다.

③ 기대도수의 최솟값은 10이다.

④ 기대도수의 최댓값은 30이다.

정답 3번

간단한 해설

경력에 따른 연봉 차이가 없다고 가정하면 각 경력 집단에 전체 표본의 연봉 구간별 비율을 동일하게 적용한다. 각 집단의 기대도수는 연봉 구간 순서대로 20·30·30·20명이므로, 기대도수의 **최솟값은 20명**이다.

NoteBase

선지별 해설

기대도수는 경력 집단별 연봉 분포가 같다는 가정하에서 각 칸에 예상되는 인원수이다. 행 합계를 R_i, 열 합계를 C_j, 전체 인원수를 n이라 하면 $E_{ij} = \frac{R_i C_j}{n}$으로 구한다. 이는 각 집단의 인원수에 전체 표본의 해당 연봉 구간 비율을 곱한 것이다. 각 행의 합계는 100명, 전체는 300명이며, 연봉 구간 순서대로 열 합계는 60·90·90·60명이다. 따라서 세 경력 집단 모두 기대도수가 $\frac{100 \cdot 60}{300} = 20$, $\frac{100 \cdot 90}{300} = 30$, 30, 20명이다.

① O: 경력 1년 이상 3년 미만이고 연봉이 3,000만 원 이상 4,000만 원 미만인 칸은 관측도수와 기대도수가 모두 30명이다. 차의 절댓값은 0이며, **절댓값은 음수가 될 수 없으므로** 최솟값은 0이다.

② O: 연봉 구간 순서대로 기대도수와 관측도수 차의 절댓값은 경력 1년 미만에서 10·20·10·20명, 1년 이상 3년 미만에서 0·0·10·10명, 3년 이상에서 10·20·0·30명이다. 이 중 최댓값은 30명이다. 이는 경력 3년 이상이고 연봉이 5,000만 원 이상인 칸에서 **관측도수 50명과 기대도수 20명의 차이**이다.

③ X: 모든 칸의 기대도수는 20명 또는 30명이므로 **최솟값은 10명이 아니라 20명**이다. 연봉 3,000만 원 미만과 5,000만 원 이상 구간의 기대도수가 각 경력 집단에서 모두 20명이다.

④ O: 연봉 3,000만 원 이상 4,000만 원 미만과 4,000만 원 이상 5,000만 원 미만 구간의 열 합계는 각각 90명이다. 두 구간의 기대도수는 각 경력 집단에서 30명이며, 나머지 구간은 20명이므로 최댓값은 30명이다.

NoteBase

문 6. 정규분포를 따르는 두 모집단 $N(\mu_X, \sigma_X^2)$과 $N(\mu_Y, \sigma_Y^2)$에서 각각 독립표본 $X_1, X_2, \ldots, X_5$와 $Y_1, Y_2, \ldots, Y_6$을 임의추출하여 구한 표본분산을 순서대로 24와 9라고 하자. 확률변수 W는 분자와 분모의 자유도가 각각 4와 5인 F분포를 따를 때, 가설 $H_0 : \frac{\sigma_X^2}{\sigma_Y^2} = 1$ 대 $H_1 : \frac{\sigma_X^2}{\sigma_Y^2} > 1$에 대한 검정에서 유의확률과 같은 것은?

① $P\left(W \geq \frac{3}{8}\right)$
② $P\left(W \geq \frac{8}{3}\right)$
③ $P\left(W \geq \frac{15}{32}\right)$
④ $P\left(W \geq \frac{32}{15}\right)$

정답 2번

간단한 해설

귀무가설 아래에서 표본분산의 비는 자유도 4와 5인 F분포를 따르며, 관측값은 $\frac{24}{9} = \frac{8}{3}$이다. 대립가설이 모분산비가 1보다 크다는 것이므로, 유의확률은 이 관측값 이상의 **오른쪽 꼬리확률**이다.

계산 과정

표본크기는 $n_X = 5$, $n_Y = 6$이고, 표본분산은 $S_X^2 = 24$, $S_Y^2 = 9$이다. 각 표본분산의 자유도는 **표본크기에서 1을 뺀 값**이므로, 분자의 자유도는 $5 - 1 = 4$, 분모의 자유도는 $6 - 1 = 5$이다.

정규모집단에서는 $\frac{(n-1)S^2}{\sigma^2}$이 자유도 $n-1$인 카이제곱분포를 따른다. 서로 독립인 카이제곱변수를 각각의 자유도로 나눈 뒤 그 비를 구하면 F분포가 된다. 따라서 귀무가설 $\sigma_X^2 = \sigma_Y^2$ 아래에서 검정통계량 $T = \frac{S_X^2}{S_Y^2}$는 분자와 분모의 자유도가 각각 4와 5인 F분포를 따른다.

주어진 표본분산을 대입하면 검정통계량의 관측값은 $t_{\text{obs}} = \frac{24}{9} = \frac{8}{3}$이다. 이는 표본에서 구한 X의 분산이 Y의 분산의 $\frac{8}{3}$배라는 뜻이며, 분산끼리 나눈 비이므로 단위는 없다.

대립가설은 $\frac{\sigma_X^2}{\sigma_Y^2} > 1$이므로, 검정통계량이 **클수록 대립가설을 지지**한다. 유의확률은 귀무가설 아래에서 관측값 이상으로 큰 검정통계량이 나올 확률이다. 문제의 W가 귀무가설 아래의 검정통계량과 같은 분포를 따르므로, 유의확률은 $P\left(W \geq \frac{8}{3}\right)$이며 ②의 식과 일치한다.

NoteBase

문 7. 고혈압 환자 20명 중에서 임의추출한 10명으로 이루어진 집단 A에는 신약을 처방하고 나머지 10명으로 이루어진 집단 B에는 기존 약을 처방한 뒤 4주 후에 혈압을 측정하였다. 다음은 두 집단의 혈압이 각각 동일한 분산을 갖는 정규분포를 따른다고 가정하고 분석한 결과이다. 공통분산의 추정량을 S_p^2이라고 할 때, k와 S_p^2의 값을 바르게 연결한 것은? (단, $t_\alpha(k)$는 자유도가 k인 t분포의 제$100 \times (1-\alpha)$ 백분위수이다)

공통분산의 추정량의 자유도는 k이고 $t_{0.025}(k) = 2.1$이라고 할 때, 두 집단의 평균 혈압 차에 대한 95% 신뢰구간은 $(-12.2,\ -3.8)$이다.

① k: 9, S_p^2: 10
② k: 9, S_p^2: 20
③ k: 18, S_p^2: 10
④ k: 18, S_p^2: 20

정답 4번

간단한 해설

두 집단이 각각 10명이므로 공통분산 추정량의 **자유도는 18**이다. 신뢰구간의 오차한계 4.2를 임계값 2.1로 나누면 평균 차의 표준오차가 2이므로, **공통분산 추정값**은 $S_p^2 = 20$이다.

계산 과정

자유도: 두 집단은 서로 다른 환자 10명씩으로 구성되어 있고, 혈압이 같은 분산을 갖는 정규분포를 따른다고 가정한다. 따라서 독립 두 표본의 등분산 t 신뢰구간을 적용한다. 각 집단에서 표본평균을 추정하므로 분산 추정에 쓰이는 자유도는 각각 10 - 1 = 9이다. 공통분산 추정량은 **두 집단의 자유도를 합쳐** 사용하므로 $k = (10-1) + (10-1) = 18$이다.

오차한계: 평균 차의 신뢰구간은 표본평균 차를 중심으로 좌우 대칭이다. 따라서 주어진 구간의 **전체 길이의 절반**이 오차한계이며, $E = \frac{-3.8-(-12.2)}{2} = 4.2$이다.

평균 차의 표준오차: 표본평균 차를 $D = \bar{X}_A - \bar{X}_B$, 그 표준오차 추정값을 s_D라고 하면, 95% 신뢰구간은 $D \pm t_{0.025}(k)s_D$이다. 오차한계는 임계값과 표준오차의 곱이므로 $4.2 = 2.1s_D$이고, $s_D = \frac{4.2}{2.1} = 2$이다. 이 값은 **표본평균 차의 변동 크기**를 추정한 것으로, 개별 환자 혈압의 표준편차와 구별한다.

공통분산: 독립인 두 표본에서는 표본평균 차의 분산이 각 표본평균의 분산을 더한 값이다. 각 집단의 표본크기가 10이므로 $s_D^2 = S_p^2\left(\frac{1}{10} + \frac{1}{10}\right)$이고, 이를 정리하면 $2^2 = \frac{S_p^2}{5}$이다. 따라서 $S_p^2 = 5 \cdot 2^2 = 20$이다.

구한 값은 $k = 18$, $S_p^2 = 20$으로, ④의 연결과 일치한다.

NoteBase

문 8. 자료 (x_i, y_i) $(i = 1, 2, \ldots, 6)$에 단순선형회귀모형 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$를 최소제곱법으로 적합하여 다음 값을 얻었다. x와 y의 표본상관계수가 음수일 때, β_1의 추정값은? (단, $\hat{y}_i$는 y_i의 적합값이다)

$$\sum_{i=1}^{6}(y_i - \hat{y}_i)^2 = 150, \quad \sum_{i=1}^{6}(y_i - \bar{y})^2 = 270, \quad \sum_{i=1}^{6}(x_i - \bar{x})^2 = 30$$

① 2

② -2

③ $\sqrt{5}$

④ $-\sqrt{5}$

정답 2번

간단한 해설

회귀제곱합은 270 − 150 = 120이므로 기울기 추정값의 제곱은 $\frac{120}{30} = 4$이다. 표본상관계수가 **음수**이므로 기울기 추정값은 −2이다.

계산 과정

회귀제곱합 계산: 절편을 포함한 단순선형회귀모형을 최소제곱법으로 적합하면 **총제곱합은 회귀제곱합과 잔차제곱합의 합**으로 분해된다. 주어진 총제곱합은 270, 잔차제곱합은 150이므로 회귀제곱합은 270 − 150 = 120이다. 이는 관측값의 전체 변동 중 회귀직선으로 설명되는 부분이다.

기울기의 크기 계산: 기울기의 최소제곱추정값을 $\hat{\beta}_1$이라 하면, 적합한 회귀직선은 표본평균점 $(\bar{x}, \bar{y})$을 지나므로 $\hat{y}_i - \bar{y} = \hat{\beta}_1(x_i - \bar{x})$이다. 양변을 제곱하여 여섯 관측값에 대해 합하면 회귀제곱합은 $\hat{\beta}_1^2 \sum_{i=1}^{6}(x_i - \bar{x})^2$이 된다. 따라서 $120 = 30\hat{\beta}_1^2$이고, $\hat{\beta}_1^2 = \frac{120}{30} = 4$이므로 기울기의 절댓값은 2이다.

부호 결정: 표본상관계수를 r, 두 변수의 표본표준편차를 각각 s_x, s_y라 하면 $\hat{\beta}_1 = r\frac{s_y}{s_x}$이다. 두 표본표준편차가 양수이므로 **기울기와 표본상관계수의 부호는 같다**. 주어진 상관계수가 음수이므로 $\hat{\beta}_1 = -2$이며, 이는 ②에 해당한다. 이 기울기는 x가 1단위 증가할 때 y의 적합값이 2단위 감소한다는 뜻이다.

NoteBase

문 9. 다음 표는 세 학생 집단 A, B, C의 시험성적을 비교하기 위해 각 집단에서 4명씩 임의추출하여 작성한 분산분석표의 일부이다. 집단 A와 B의 시험성적의 모평균이 동일하다고 가정하고 하나의 집단으로 묶어서 집단 C와 비교하기 위한 분산분석표를 다시 작성한다고 할 때, (가) ~ (라) 중에서 변하지 않는 것만을 모두 고르면?

요인	제곱합	자유도	평균제곱	$F-$값
처리	(가)			
오차		(나)		
계	(다)	(라)		

① (가), (나)

② (가), (다)

③ (나), (라)

④ (다), (라)

정답 4번

간단한 해설

집단을 합쳐도 **12명의 관측값과 전체 평균은 그대로**이므로 총제곱합과 총자유도 11은 변하지 않는다. 모평균이 같다는 가정만으로 두 집단의 표본평균까지 같아지는 것은 아니므로 처리 제곱합의 불변은 보장되지 않는다.

NoteBase

계산 과정

전체 학생 수는 3 × 4 = 12명이다. A와 B를 합치면 집단별 인원은 8명과 4명이 되고, 집단 수는 3개에서 2개로 줄어든다. 표의 (가)는 처리 제곱합, (나)는 오차 자유도, (다)는 총제곱합, (라)는 총자유도이다.

(가) 처리 제곱합: 각 집단의 표본평균이 전체 평균에서 얼마나 떨어져 있는지를 집단별 인원으로 가중하여 합한 값이다. A와 B의 표본평균을 각각 a, b라 하면, 두 집단의 인원이 같으므로 합친 집단의 평균은 $h = \frac{a+b}{2}$이다. 합병하면 A와 B 사이의 차이에 해당하는 제곱합이 처리 제곱합에서 빠져 오차 제곱합으로 옮겨간다. 그 크기는 $\Delta = 4(a-h)^2 + 4(b-h)^2$이며, $\Delta = 2(a-b)^2$로 정리된다. 따라서 합병 후 처리 제곱합은 합병 전보다 $2(a-b)^2$만큼 작아진다. 두 표본평균이 같은 특수한 경우에는 유지되지만, **모평균이 같아도 표본평균은 다를 수 있다**는 점에서 항상 변하지 않는 항목은 아니다.

(나) 오차 자유도: 전체 관측값 수를 N, 집단 수를 k라 하면 오차 자유도는 $N-k$이다. 각 집단의 평균을 하나씩 추정하기 때문에 집단 수만큼 자유도가 줄어든다. 합병 전에는 $12-3=9$, 합병 후에는 $12-2=10$이다. 따라서 **오차 자유도는 9에서 10으로 증가**한다.

(다) 총제곱합: 각 관측값 x_i와 전체 평균 $\bar{x}$의 차이를 제곱하여 합한 값으로, $\mathrm{SST} = \sum_{i=1}^{12}(x_i-\bar{x})^2$이다. 집단을 합치는 것은 집단 구분만 바꾸는 것이므로 **관측값과 전체 평균이 모두 유지**된다. 따라서 총제곱합은 변하지 않는다.

(라) 총자유도: 전체 관측값에서 전체 평균 하나를 추정하므로 전체 관측값 수에서 1을 뺀 값이다. 합병 전후 모두 관측값이 12개이므로 $12-1=11$로 유지된다.

따라서 변하지 않는 항목은 (다), (라)이며, 이는 ④의 조합에 해당한다.

NoteBase

문 10. 어느 영업사원에게 걸려오는 전화통화 수는 한 시간에 평균 6인 포아송분포를 따른다. 두 확률변수 U와 V는 각각 평균이 2와 6인 포아송분포를 따른다고 할 때, 오전 9시 이후 이 영업사원에게 처음 걸려오는 전화가 같은 날 오전 9시 20분 이후일 확률과 같은 것은?

① $P(U=0)$

② $P(U=1)$

③ $P(V=0)$

④ $P(V=1)$

정답 1번

간단한 해설

첫 전화가 오전 9시 20분 이후에 오려면 앞선 20분 동안 **전화가 한 통도 오지 않아야 한다**. 이 구간의 평균 통화 수는 2회이므로 구하는 확률은 $P(U=0)$이다.

계산 과정

1. 사건 바꾸기: 오전 9시부터 9시 20분까지 걸려오는 전화 수를 N이라 하자. 처음 전화가 오전 9시 20분 이후에 온다는 것은 그 전 20분 동안 **도착한 전화 수가 0**이라는 뜻이다. 따라서 구할 확률은 $P(N=0)$이다.

2. 20분 동안의 평균 구하기: 시간당 평균 6회인 포아송 도착 모형에서는 구간의 평균 통화 수가 **시간 길이에 비례**한다. 20분은 $\frac{20}{60}=\frac{1}{3}$시간이므로 이 구간의 평균은 $6\cdot\frac{1}{3}=2$회이다. 따라서 N은 평균 2인 포아송분포를 따른다.

3. 확률 계산하기: 평균이 2인 포아송분포의 확률질량함수는 $P(N=k)=\frac{e^{-2}2^k}{k!}$이다. 전화가 한 통도 오지 않는 경우는 $k=0$이므로 $P(N=0)=\frac{e^{-2}2^0}{0!}=e^{-2}$이다.

주어진 U도 평균 2인 포아송분포를 따르므로 $P(N=0)=P(U=0)$이다. 이는 ①의 확률과 같다.

NoteBase

문 11. 자료의 퍼진 정도를 나타내는 측도 중에서 원자료와 같은 측정 단위를 갖는 것만을 모두 고르면?

ㄱ. 표준편차
ㄴ. 사분위수범위
ㄷ. 분산
ㄹ. 변동계수

① ㄱ, ㄴ
② ㄱ, ㄷ
③ ㄴ, ㄹ
④ ㄷ, ㄹ

정답 1번

간단한 해설

표준편차는 분산의 제곱근을 취하므로 **원자료와 같은 단위**를 갖는다. 사분위수범위도 **같은 단위를 가진 두 사분위수의 차이**이므로 원자료의 단위가 유지된다.

보기별 해설

ㄱ. 표준편차: 분산의 제곱근으로, 편차를 제곱하면서 생긴 단위의 제곱을 제곱근으로 되돌린다. 따라서 **원자료와 같은 단위**를 갖는다. 예를 들어 원자료가 cm 단위이면 분산은 cm^2, 표준편차는 cm 단위이다.

ㄴ. 사분위수범위: 제3사분위수에서 제1사분위수를 뺀 값인 $Q_3 - Q_1$이다. 두 사분위수는 모두 원자료와 같은 단위를 가지며, **두 값의 차이에도 그 단위가 유지**된다.

ㄷ. 분산: 각 관측값과 평균의 차이를 제곱한 값을 바탕으로 산포를 나타낸다. 편차의 단위는 원자료와 같지만 이를 제곱하므로, 분산의 단위는 **원자료 단위의 제곱**이다. 원자료가 cm 단위이면 분산은 cm^2 단위이므로 같은 측정 단위를 갖지 않는다.

ㄹ. 변동계수: 평균이 0이 아닐 때 표준편차를 평균으로 나눈 비율인 표준편차평균이다. 분자와 분모의 단위가 같아 서로 약분되므로 **측정 단위가 없는 상대적 산포도**이다. 비율에 100을 곱해 백분율로 표시하기도 하지만 원자료의 측정 단위를 갖지는 않는다.

원자료와 같은 측정 단위를 갖는 것은 ㄱ과 ㄴ이며, 이 조합은 ①에 해당한다.

NoteBase

문 12. 동전 한 개를 연속해서 10번 던질 때, 처음 6번의 시행에서는 뒷면이 2번 나오고 나머지 4번의 시행에서는 앞면이 2번 나올 확률은?

① $\frac{15}{2^9}$
② $\frac{15}{2^{10}}$
③ $\frac{45}{2^9}$
④ $\frac{45}{2^{10}}$

정답 3번

간단한 해설

처음 6번과 나머지 4번은 **서로 독립인 시행 구간**이므로, 각 구간에서 요구한 횟수가 나올 확률을 곱한다. 두 조건을 동시에 만족할 확률은 $\frac{45}{2^9}$이다.

계산 과정

동전 던지기의 통상적인 모형에 따라 앞면과 뒷면이 나올 확률을 각각 1/2로 보고, 각 시행은 서로 독립이라고 가정한다. 따라서 각 구간에서 특정 면이 나온 횟수에는 **이항분포**를 적용한다.

처음 6번에서 뒷면이 2번: 뒷면이 나올 두 자리를 고르는 방법은 $\binom{6}{2} = 15$가지이다. 각 배열에는 뒷면 2번과 앞면 4번이 있으며, 한 배열의 확률은 $\left(\frac{1}{2}\right)^6$이다. 따라서 이 구간의 조건을 만족할 확률은 $\binom{6}{2}\left(\frac{1}{2}\right)^6 = \frac{15}{2^6}$이다.

나머지 4번에서 앞면이 2번: 앞면이 나올 두 자리를 고르는 방법은 $\binom{4}{2} = 6$가지이다. 각 배열에는 앞면 2번과 뒷면 2번이 있으므로, 이 구간의 조건을 만족할 확률은 $\binom{4}{2}\left(\frac{1}{2}\right)^4 = \frac{6}{2^4}$이다.

두 조건의 동시 충족: 두 사건은 **서로 겹치지 않는 독립 시행 구간**에 관한 것이므로 확률을 곱한다. 구하는 확률은 $\frac{15}{2^6} \cdot \frac{6}{2^4} = \frac{90}{2^{10}} = \frac{45}{2^9}$이며, 이는 ③의 값과 일치한다.

NoteBase

문 13. 다음 표는 단순선형회귀모형 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ $(i = 1, 2, \ldots, 26)$에 대한 분산분석표의 일부이다. 이 모형의 결정계수의 값은? (단, ε_i는 $N(0, \sigma^2)$을 따르고 서로 독립이다)

요인	제곱합	자유도	평균제곱	$F-$값
회귀				5
잔차	120			
계		25		

① $\frac{4}{29}$
② $\frac{5}{29}$
③ $\frac{6}{29}$
④ $\frac{7}{29}$

정답 2번

간단한 해설

결정계수는 전체 변동 중 **회귀모형이 설명하는 변동의 비율**이다. 잔차제곱합 120과 F값 5로부터 회귀제곱합은 25이므로, 결정계수는 $\frac{25}{25+120} = \frac{5}{29}$이다.

NoteBase

계산 과정

결정계수의 의미: 회귀제곱합 SSR은 모형이 설명하는 변동, 잔차제곱합 SSE는 설명하지 못하는 변동이다. 절편을 포함한 최소제곱 회귀모형에서는 총제곱합이 SST = SSR + SSE로 분해되므로, 결정계수는 $R^2 = \frac{\text{SSR}}{\text{SST}}$이다.

잔차평균제곱 계산: 관측값은 26개이고 절편과 기울기 두 모수를 추정하므로, **잔차 자유도는 26 − 2 = 24**이다. 이는 표의 전체 자유도 25에서 회귀 자유도 1을 뺀 값과도 같다. 표의 잔차제곱합 120을 자유도로 나누면 잔차평균제곱은 $\text{MSE} = \frac{120}{24} = 5$이다.

회귀제곱합 계산: F값은 **회귀평균제곱을 잔차평균제곱으로 나눈 값**이므로 $F = \frac{\text{MSR}}{\text{MSE}}$이다. 표에서 F값이 5이므로 회귀평균제곱은 $\text{MSR} = 5 \cdot 5 = 25$이다. 설명변수가 하나라 회귀 자유도는 1이므로, 회귀제곱합도 $\text{SSR} = 25 \cdot 1 = 25$이다.

결정계수 계산: 총제곱합은 25 + 120 = 145이다. 따라서 $R^2 = \frac{25}{145} = \frac{5}{29}$이며, 이는 전체 변동 중 모형이 설명하는 비율로 단위가 없다. 이 값은 ②의 값과 일치한다.

NoteBase

문 14. 두 확률변수 X와 Y가 이변량 정규분포를 따르고 상관계수가 ρ일 때, 가설 $H_0 : \rho = 0$ 대 $H_1 : \rho > 0$을 검정하고자 한다. 표본의 크기가 n이고 표본상관계수가 r일 때, 위 가설을 검정하기 위한 검정통계량 $T = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$는 H_0하에서 자유도가 $(n-2)$인 t분포를 따른다고 한다. $n = 11$이고 $r = 0.6$일 때, 유의수준 5%에서 검정 방법에 대한 설명으로 옳은 것은? (단, $t_\alpha(k)$는 자유도가 k인 t분포의 제 $100 \times (1-\alpha)$ 백분위수를 나타낼 때, $t_{0.025}(9) = 2.26$, $t_{0.05}(9) = 1.83$이다.)

① T의 값이 $t_{0.025}(9)$보다 커서 귀무가설을 기각한다.

② T의 값이 $t_{0.025}(9)$보다 작아서 귀무가설을 기각하지 못한다.

③ T의 값이 $t_{0.05}(9)$보다 커서 귀무가설을 기각한다.

④ T의 값이 $t_{0.05}(9)$보다 작아서 귀무가설을 기각하지 못한다.

정답 3번

간단한 해설

대립가설이 $\rho > 0$이므로 **오른쪽 단측검정**을 하며, 유의수준 5%의 임계값은 $t_{0.05}(9) = 1.83$이다. 검정통계량 T = 2.25가 이를 초과하므로 귀무가설을 기각한다.

계산 과정

검정통계량과 자유도: 두 확률변수가 이변량 정규분포를 따르므로, 문제에서 주어진 $T = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$를 이용한다. 여기서 n은 표본의 크기, r은 표본상관계수이다. 귀무가설 $H_0 : \rho = 0$하에서 T는 자유도 $n - 2 = 11 - 2 = 9$인 t분포를 따른다.

관측값 계산: $\sqrt{n-2} = \sqrt{9} = 3$, $\sqrt{1-r^2} = \sqrt{1-0.6^2} = \sqrt{0.64} = 0.8$이다. 따라서 $T = \frac{3 \cdot 0.6}{0.8} = 2.25$이다. 이는 관측된 표본상관계수 0.6을 귀무가설하의 t분포와 비교할 수 있도록 변환한 값이며, 단위는 없다.

기각역 설정: 대립가설이 $H_1 : \rho > 0$이므로 표본상관계수 r이 클수록 대립가설을 지지하고 T도 커진다. 따라서 **오른쪽 꼬리에 유의수준 5% 전부를 배정**한다. $t_\alpha(k)$는 오른쪽 꼬리확률이 α인 값이므로, 기각역은 $T > t_{0.05}(9) = 1.83$이다. $t_{0.025}(9) = 2.26$은 유의수준 5%의 양측검정에서 양쪽 꼬리에 2.5%씩 배정할 때 사용하는 양의 임계값이다.

판정: 관측값 T = 2.25는 임계값 1.83보다 크므로 귀무가설을 기각한다. 즉, 유의수준 5%에서 상관계수가 양수라는 대립가설을 지지할 통계적 근거가 있다. 이는 T가 $t_{0.05}(9)$보다 커서 귀무가설을 기각한다는 ③의 설명과 일치한다.

NoteBase

문 15. 수준 수가 4인 인자 A와 수준 수가 3인 인자 B의 각 수준조합에서 3회씩 반복 실험을 하였다. 다음 표는 인자 A와 B의 교호작용(interaction)이 유의하지 않아서 교호작용을 오차항에 포함한 후 작성한 분산분석표의 일부이다.

유의수준 5%에서 검정할 때, 인자 A와 B의 유의성에 대한 설명으로 옳은 것은? (단, $F_{\alpha}(k_1, k_2)$는 분자와 분모의 자유도가 각각 k_1, k_2인 F분포의 제$100 \times (1-\alpha)$ 백분위수를 나타낼 때, $F_{0.05}(2, 30) = 3.32$, $F_{0.05}(3, 30) = 2.92$이다.)

요인	제곱합	자유도	평균제곱	$F-$값
인자 A	54			
인자 B	72			
오차				
계	396			

① 인자 A, B는 모두 유의하다.

② 인자 A는 유의하고, 인자 B는 유의하지 않다.

③ 인자 A는 유의하지 않고, 인자 B는 유의하다.

④ 인자 A, B는 모두 유의하지 않다.

정답 3번

간단한 해설

교호작용을 오차에 합친 뒤의 **오차평균제곱은 9**이다. 이를 분모로 계산한 F값은 A가 2로 임계값 2.92보다 작고, B가 4로 임계값 3.32보다 크므로 A는 유의하지 않고 B는 유의하다.

NoteBase

계산 과정

자유도: A의 수준 수는 4, B의 수준 수는 3이고 각 조합에서 3회 반복하므로 관측값은 $4 \times 3 \times 3 = 36$개이다. 총자유도는 $36 - 1 = 35$, A의 자유도는 $4 - 1 = 3$, B의 자유도는 $3 - 1 = 2$이다. **교호작용을 오차항에 포함**했으므로 합쳐진 오차의 자유도는 $35 - 3 - 2 = 30$이다.

오차평균제곱: 표의 제곱합 열에서 총제곱합은 396, A와 B의 제곱합은 각각 54, 72이다. 따라서 합쳐진 오차제곱합은 $396 - 54 - 72 = 270$이다. 평균제곱은 제곱합을 해당 자유도로 나눈 값이므로 오차평균제곱은 $MS_E = \frac{270}{30} = 9$이다. 이는 오차분산의 추정값으로, 두 인자의 F검정에서 공통 분모가 된다.

검정 기준: 각 인자에 대해 수준별 모평균이 모두 같다는 귀무가설을 검정한다. 인자의 평균제곱이 오차평균제곱에 비해 충분히 큰지 판단하므로 검정통계량은 F=인자의 평균제곱MS_E이다. 유의수준 5%에서 F값이 해당 자유도의 **상위 5% 임계값보다 크면** 귀무가설을 기각한다.

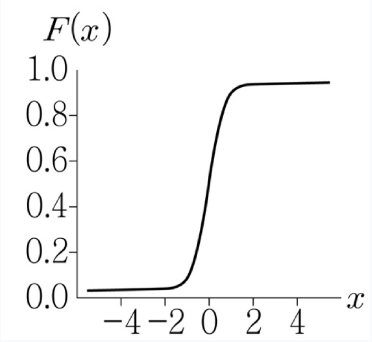
인자 A: 평균제곱은 $MS_A = \frac{54}{3} = 18$, 검정통계량은 $F_A = \frac{18}{9} = 2$이다. 분자 자유도 3, 분모 자유도 30의 임계값 2.92보다 작으므로 귀무가설을 기각하지 못한다. 따라서 A는 유의수준 5%에서 유의하지 않다.

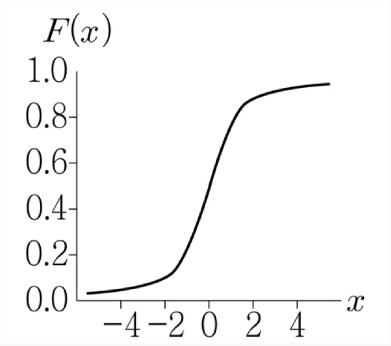
인자 B: 평균제곱은 $MS_B = \frac{72}{2} = 36$, 검정통계량은 $F_B = \frac{36}{9} = 4$이다. 분자 자유도 2, 분모 자유도 30의 임계값 3.32보다 크므로 귀무가설을 기각한다. 따라서 B는 유의수준 5%에서 유의하다.

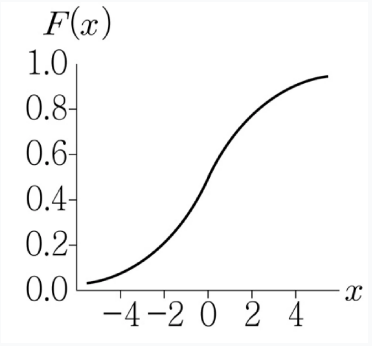
A는 유의하지 않고 B는 유의하므로 ③의 설명과 일치한다.

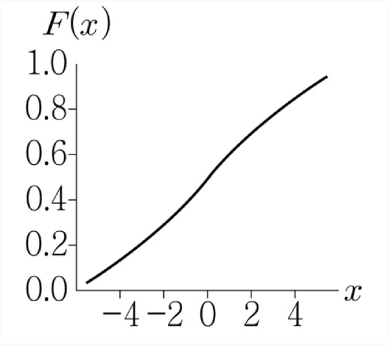
NoteBase

문 16. 다음은 평균은 같고 분산이 서로 다른 정규분포의 누적분포함수 $F(x)$ $(-\infty < x < \infty)$의 그림이다. 분산이 가장 큰 것은?

① 

② 

③ 

④ 

정답 4번

간단한 해설

평균이 같은 정규분포는 분산이 클수록 확률이 넓게 퍼져, 누적분포함수가 **평균 부근에서 더 완만하게 증가한다**.
④의 곡선은 네 그림 중 이 구간의 기울기가 가장 작다.

NoteBase

선지별 해설

누적분포함수 $F(x) = P(X \leq x)$는 x 이하의 확률을 나타내며, 그 기울기는 확률밀도이다. 정규분포의 평균을 μ, 표준편차를 σ라 하면 평균에서의 기울기는 $F'(\mu) = \frac{1}{\sigma\sqrt{2\pi}}$이다. 따라서 평균이 같을 때 **중앙의 기울기가 작을수록 표준편차와 분산이 크다**. 네 그림은 가로축과 세로축의 눈금이 같으므로 중앙의 기울기를 직접 비교할 수 있다.

①: 중앙에서 가장 가파르게 증가한다. 확률이 평균 주변의 좁은 범위에 집중되어 있어 네 분포 중 분산이 가장 작다.

②: 중앙의 증가가 ①보다 완만하지만 ③과 ④보다는 가파르다. 분산은 ①보다 크고 ③과 ④보다 작다.

③: 중앙의 증가가 ①과 ②보다 완만하지만 ④보다는 가파르다. 분산은 ①과 ②보다 크고 ④보다 작다.

④: 중앙에서 가장 완만하게 증가한다. 같은 정도의 누적확률 증가가 더 넓은 x 구간에 걸쳐 나타나므로 확률이 가장 넓게 퍼져 있다.

분산은 ① < ② < ③ < ④ 순으로 크므로, 분산이 가장 큰 그림은 ④이다.

문 17. 자료 (x_i, y_i) $(i = 1, 2, \ldots, 10)$로부터 다음 결과를 얻었을 때, 이 자료를 최소제곱법으로 적합한 단순선형회귀식은? (단, r는 x와 y의 표본상관계수이다)

$$\bar{x} = 2, \quad \bar{y} = 3, \quad r = \frac{1}{3}, \quad \sum_{i=1}^{10} (x_i - \bar{x})^2 = 1, \quad \sum_{i=1}^{10} (y_i - \bar{y})^2 = 36$$

① $\hat{y} = -1 + 2x$

② $\hat{y} = -2 + x$

③ $\hat{y} = -4 + 2x$

④ $\hat{y} = -4 + x$

정답 1번

간단한 해설

회귀직선의 기울기는 상관계수에 **표준편차의 비**를 곱한 값이므로 2이다. 이 직선은 평균점 (2, 3)을 지나므로 절편은 −1이고, 적합식은 $\hat{y} = -1 + 2x$이다.

계산 과정

절편을 포함한 단순선형회귀식은 $\hat{y} = a + bx$이다. 최소제곱법은 관측값과 예측값의 차이인 잔차의 제곱합을 최소로 하도록 절편 a와 기울기 b를 정한다. 이때 $b = r\frac{s_y}{s_x}$, $a = \bar{y} - b\bar{x}$이며, s_x와 s_y는 각각 x와 y의 표본표준편차이다.

기울기 계산: 두 표본분산의 분모가 모두 $n - 1 = 9$이므로 표준편차의 비에서는 **공통 분모가 약분된다**. 따라서 $\frac{s_y}{s_x} = \sqrt{\frac{36/9}{1/9}} = \sqrt{\frac{36}{1}} = 6$이다. 여기에 상관계수를 곱하면 $b = \frac{1}{3} \cdot 6 = 2$이다. 이는 x가 1 증가할 때 예측값 $\hat{y}$가 2 증가한다는 뜻이다.

절편 계산: 절편을 포함한 최소제곱 회귀직선은 **두 변수의 평균점** $(\bar{x}, \bar{y})$를 지난다. 따라서 $a = \bar{y} - b\bar{x} = 3 - 2 \cdot 2 = -1$이다.

기울기 2와 절편 −1을 대입한 회귀식은 $\hat{y} = -1 + 2x$이며, ①과 일치한다.

NoteBase

문 18. 어느 사과 농장에서 수확한 사과의 평균 무게 μ에 대한 가설 $H_0 : \mu = 350$ 대 $H_1 : \mu \neq 350$을 검정하려고 한다. 이 농장에서 수확한 사과 49개를 임의추출하여 구한 μ에 대한 95% 신뢰구간이 (349, 361)일 때, 이에 대한 설명으로 옳지 않은 것은?

① 유의수준 5%에서 검정할 때 H_0을 기각할 수 없다.

② 유의수준 2%에서 검정할 때 H_0을 기각할 수 없다.

③ 위 가설에 대한 유의확률은 0.01보다 크다.

④ 위 가설에 대한 유의확률은 0.03보다 작다.

정답 4번

간단한 해설

95% 신뢰구간 (349, 361)에 귀무가설의 모평균 350이 포함되므로, 유의확률은 **0.05보다 크다**. 따라서 ④의 '0.03보다 작다'는 설명과 맞지 않는다.

선지별 해설

같은 방법으로 구성한 신뢰구간과 양측검정을 대응시키면, 95% 신뢰구간은 유의수준 5%에서 기각되지 않는 모평균값들의 범위이다. 유의확률을 p라 할 때, p가 유의수준 이하이면 귀무가설을 기각하고 더 크면 기각하지 않는다.

① O: 귀무가설의 모평균 350은 95% 신뢰구간 (349, 361)의 **내부에 포함**된다. 따라서 유의수준 5%의 양측검정에서 $H_0 : \mu = 350$을 기각할 수 없으며, 유의확률 p는 0.05보다 크다.

② O: 유의확률 p가 0.05보다 크므로 0.02보다도 크다. 유의수준을 5%에서 2%로 낮추면 **기각 기준이 더 엄격**해지므로, 5%에서 기각하지 못한 귀무가설을 2%에서도 기각할 수 없다.

③ O: 350이 95% 신뢰구간 내부에 있으므로 유의확률 p는 0.05보다 크다. 0.05는 0.01보다 크므로 p도 0.01보다 크다.

④ X: 유의확률 p는 **0.03보다 작은 것이 아니라 0.05보다 크다**. 만약 p가 0.03보다 작다면 0.05보다도 작아 유의수준 5%에서 귀무가설을 기각해야 하지만, 이는 350이 95% 신뢰구간 내부에 있다는 사실과 모순된다.

NoteBase

문 19. 알코올 중독자와 비중독자 간에 질병 A의 유병률이 차이가 있는지 알아보기 위해 어느 지역의 알코올 중독자와 비중독자 중에서 각각 76명과 109명을 임의추출하여 질병 A의 유무에 대한 다음과 같은 분할표를 얻었다. 분석 목적과 표집 과정을 고려할 때, 이 지역에 대한 설명으로 옳은 것은?

질병 A의 유무 \ 중독 여부	중독자	비중독자	계
있음	53	28	81
없음	23	81	104
계	76	109	185

① 알코올 중독자의 비율은 $\frac{76}{185}$으로 추정된다.

② 질병 A의 유병률은 $\frac{81}{185}$로 추정된다.

③ 알코올 중독자 중에서 질병 A의 유병률은 $\frac{53}{76}$으로 추정된다.

④ 질병 A의 유병자 중에서 알코올 중독자의 비율은 $\frac{53}{81}$으로 추정된다.

정답 3번

간단한 해설

알코올 중독자 집단에서 임의추출한 76명 중 질병 A가 있는 사람은 53명이므로, **중독자 집단 내 유병률**은 $\frac{53}{76}$으로 추정한다. 두 집단의 표본 수를 따로 정했으므로, 표본의 중독자 구성비를 지역 전체의 구성비로 볼 수는 없다.

NoteBase

선지별 해설

유병률은 해당 집단에서 질병이 있는 사람이 차지하는 비율이다. 이 조사는 지역 전체에서 185명을 한꺼번에 임의추출한 것이 아니라, 중독자와 비중독자 집단에서 각각 76명과 109명을 따로 임의추출했다. 따라서 각 집단 내부의 유병률은 추정할 수 있지만, 두 집단을 합쳐 지역 전체에 관한 비율을 추정할 때는 모집단의 집단별 구성비를 고려해야 한다.

① X: $\frac{76}{185}$은 전체 표본 185명 중 중독자 76명이 차지하는 비율이다. 이 비율은 **집단별로 정한 표본 수**에 따른 것으로, 지역 전체에서 중독자가 차지하는 비율을 추정한 값이 아니다. 지역의 중독자 비율을 알려면 모집단의 집단별 인원이나 지역 전체를 대표하는 별도의 표본 정보가 필요하다.

② X: 질병 A가 있는 표본은 중독자 53명과 비중독자 28명을 합한 81명이므로, $\frac{81}{185}$은 표본 전체의 유병 비율이다. 지역 전체의 유병률을 추정하려면 중독자와 비중독자의 집단별 유병률을 **모집단의 집단별 구성비로 가중평균**해야 한다. 그 구성비가 주어지지 않았으므로 표본을 단순히 합친 비율을 지역 전체의 유병률 추정값으로 쓸 수 없다.

③ O: 표의 중독자 열에서 질병 A가 있는 사람은 53명, 없는 사람은 23명으로 합계가 76명이다. 중독자 집단 안에서 임의추출했으므로, **중독자 표본 76명을 분모**로 하고 그중 유병자 53명을 분자로 한 $\frac{53}{76}$으로 이 지역 중독자 집단의 유병률을 추정할 수 있다.

④ X: $\frac{53}{81}$은 표본의 유병자 81명 중 중독자 53명이 차지하는 비율이다. 그러나 유병자 81명은 지역의 유병자 집단에서 직접 임의추출한 표본이 아니라, 중독 여부에 따라 따로 뽑은 두 표본에서 나온 사람들이다. 따라서 이 비율은 집단별 표본 수의 영향을 받는다. 지역 유병자 중 중독자의 비율을 추정하려면 집단별 유병률과 함께 **모집단의 중독자·비중독자 구성비**를 반영해야 하며, 주어진 정보만으로는 그 비율을 정할 수 없다.

NoteBase

문 20. 어느 지역 남성의 흡연율 p를 추정하기 위해 이 지역 남성 n명을 임의추출하여 구한 흡연자의 표본비율을 $\hat{p}$라고 하자. 정규근사를 이용할 때, 모든 p에 대하여 부등식 $P(|\hat{p}-p| \le 0.098) \ge 0.95$를 만족하기 위한 n의 최솟값은? (단, 표준정규분포를 따르는 확률변수 Z에 대하여 $P(Z \ge 1.96) = 0.025$이다)

① 64

② 100

③ 400

④ 625

정답 2번

간단한 해설

모든 흡연율에서 조건을 만족하려면 **표본비율의 분산이 가장 큰** $p = 0.5$를 기준으로 삼는다. 이때 95% 확률에 해당하는 오차한계가 (0.98)/(n)이므로, 이를 0.098 이하로 만드는 최소 표본크기는 100명이다.

계산 과정

1. 표본비율의 분포: p는 이 지역 남성의 실제 흡연율, n은 표본크기, $\hat{p}$는 표본에서 흡연자가 차지하는 비율이다. 표본비율의 평균은 p이고 분산은 $\frac{p(1-p)}{n}$이다. 문제에서 지정한 정규근사를 적용하면 표준화한 값 $Z = \frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$은 근사적으로 표준정규분포를 따른다.

2. 확률 조건을 오차한계로 바꾸기: 표준정규분포의 대칭성과 주어진 꼬리확률에 따라 $P(|Z| \le 1.96) = 0.95$이다. 따라서 실제 흡연율과 표본비율의 차이가 $1.96\sqrt{\frac{p(1-p)}{n}}$ 이내일 확률이 약 0.95이다. 이 값이 **허용오차 0.098 이하**이면 요구한 확률 조건을 만족하므로, $1.96\sqrt{\frac{p(1-p)}{n}} \le 0.098$이 되도록 한다.

3. 모든 흡연율에 적용하기: p(1-p)≤14이고 등호는 $p = 0.5$일 때 성립한다. 같은 표본크기에서 이때 표준오차와 오차한계가 가장 크다. 그러므로 **가장 큰 오차한계**를 기준으로 $1.96\sqrt{\frac{1}{4n}} \le 0.098$을 만족시키면 다른 모든 흡연율에서도 조건을 만족한다.

4. 최소 표본크기 계산: 위 부등식은 (0.98)/(n)≤ 0.098이다. 이를 풀면 n≥ 10, 즉 $n \ge 100$을 얻는다. 표본크기는 정수이므로 최솟값은 100명이며, 이는 ②의 100에 해당한다.

NoteBase