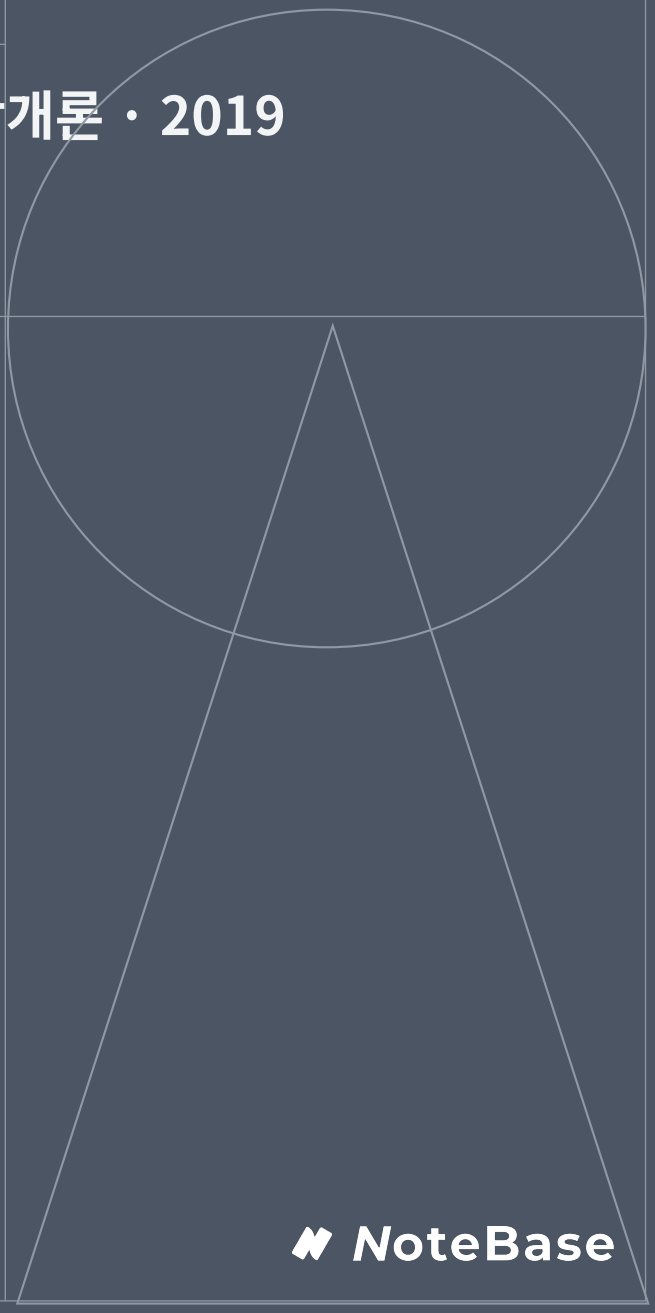
기출OX 해설집

공무원/국가직/9급 · 통계학개론 · 2019



NoteBase

기출OX 해설집

공무원/국가직/9급 · 통계학개론 · 2019

공무원/국가직/9급 · 통계학개론 · 2019

문 1. 다음은 자료 $x_1, x_2, \cdots, x_n$에 대한 표본평균($\bar{x}$)의 성질이다. 이에 대한 설명으로 옳은 것만을 모두 고르면?

ㄱ. 5개의 자료 1, 3, 5, 7, 9의 표본평균은 5이다.
ㄴ. 표본평균보다 값이 큰 자료 수와 작은 자료 수는 같다.
ㄷ. Σ_i=1^n(x_i-x)의 값은 항상 0이다.

① ㄱ, ㄴ
② ㄱ, ㄷ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 2번

간단한 해설

표본평균은 자료의 개수를 양쪽에 똑같이 나누는 값이 아니라, 각 자료에서 평균을 뺀 **편차의 합이 0**이 되게 하는 값이다. 따라서 평균보다 큰 자료와 작은 자료의 개수가 같을 필요는 없다.

NoteBase

보기별 해설

ㄱ. O: 표본평균은 **모든 자료의 합을 자료 수로 나눈 값**이다. 자료의 합은 1 + 3 + 5 + 7 + 9 = 25이고 자료 수는 5이므로, 표본평균은 $\bar{x} = \frac{25}{5} = 5$이다.

ㄴ. X: 표본평균보다 큰 자료 수와 작은 자료 수는 **항상 같지는 않다**. 예를 들어 자료가 1, 2, 9라면 평균은 $\frac{1+2+9}{3} = 4$이다. 평균 4보다 큰 자료는 9 하나이고, 작은 자료는 1과 2 두 개이다. 평균은 자료의 개수가 아니라 값의 합을 기준으로 정해진다.

ㄷ. O: 편차는 각 자료에서 표본평균을 뺀 값이다. 표본평균의 정의에 따라 $\sum_{i=1}^{n} x_i = n\bar{x}$이다. 편차를 모두 더하면 평균을 n번 빼므로 $\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x}$이다. 여기에 자료의 합을 대입하면 $n\bar{x} - n\bar{x} = 0$이 되어, **편차의 합은 항상 0**이다.

따라서 옳은 보기는 ㄱ과 ㄷ이며, 이를 묶은 선택지는 ②이다.

NoteBase

문 2. $P(A)=\frac{1}{3}$, $P(B)=\frac{1}{4}$, $P(B \mid A)+P(A \mid B)=\frac{7}{12}$**일 때,** $P(A \cap B)$**의 값은?**

① $\frac{1}{12}$
② $\frac{1}{6}$
③ $\frac{1}{4}$
④ $\frac{1}{3}$

정답 1번

간단한 해설

조건부확률의 정의를 적용하면 주어진 두 확률의 합은 **교집합 확률의 7배**이다. 따라서 $P(A \cap B)=\frac{7}{12} \times \frac{1}{7}=\frac{1}{12}$ 이다.

계산 과정

조건부확률은 조건으로 주어진 사건이 일어난 범위에서 다른 사건도 일어날 확률이다. 따라서 **교집합 확률을 조건 사건의 확률로 나누어** 구한다. 이 문제에서는 두 조건 사건의 확률이 모두 양수이므로 조건부확률을 정의할 수 있다.

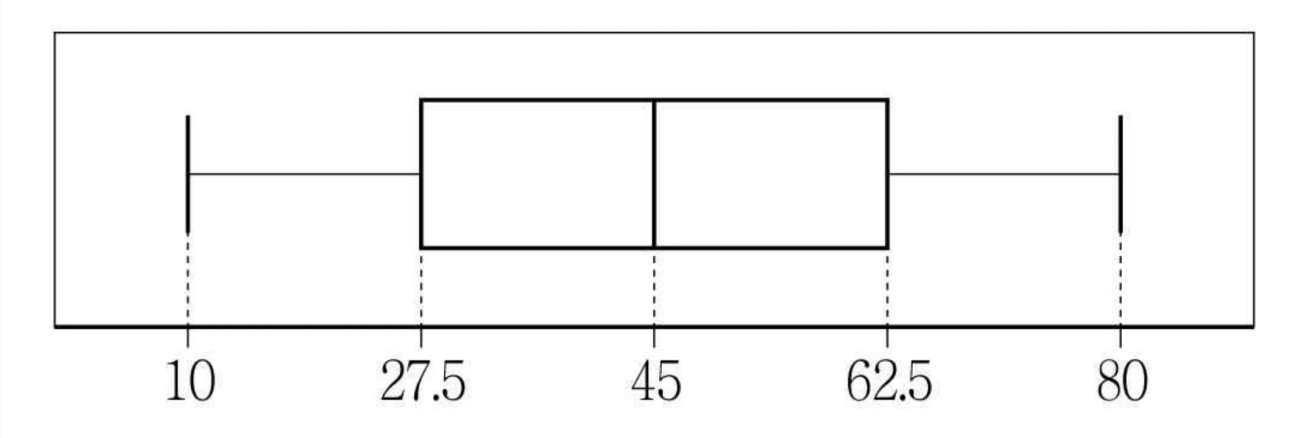
교집합 확률을 $x=P(A \cap B)$로 놓는다. A가 조건인 경우에는 $P(B \mid A)=\frac{x}{1/3}=3x$이고, B가 조건인 경우에는 $P(A \mid B)=\frac{x}{1/4}=4x$이다. 두 식의 분자는 같지만 **조건 사건이 달라 분모가 다르다**.

두 조건부확률의 합이 주어져 있으므로 $3x+4x=\frac{7}{12}$, 즉 $7x=\frac{7}{12}$이다. 양변을 7로 나누면 $x=\frac{7}{12} \times \frac{1}{7}=\frac{1}{12}$을 얻는다.

따라서 구하는 교집합 확률은 $P(A \cap B)=\frac{1}{12}$이며, 이에 해당하는 선택지는 ①이다.

NoteBase

문 3. 다음 상자그림(box plot)에 대한 설명으로 옳지 않은 것은?



① 최솟값은 10이다.

② 범위(range)는 70이다.

③ 45 이상의 값을 갖는 자료는 전체 자료의 35%이다.

④ 사분위수 범위는 35이다.

정답 3번

간단한 해설

상자 안의 세로선이 가리키는 45는 **중앙값**이다. 따라서 45 이상인 자료는 **전체의 적어도 50%**이므로 35%라는 설명은 맞지 않는다.

선지별 해설

① O: 그림의 왼쪽 수염 끝이 10을 가리키므로 최솟값은 10이다.

② O: 그림에서 최댓값은 오른쪽 수염 끝의 80, 최솟값은 왼쪽 수염 끝의 10이다. 범위는 **최댓값에서 최솟값을 뺀 값**이므로 80 − 10 = 70이다.

③ X: 상자 안의 세로선은 자료를 크기순으로 나열했을 때 가운데 위치를 나타내는 중앙값이며, 그림에서는 45이다. 따라서 45 이상인 자료는 **전체의 적어도 50%**이므로 '35%'로 볼 수 없다. 중앙값과 같은 자료가 여러 개이면 45 이상인 자료의 비율은 50%보다 클 수도 있다.

④ O: 상자의 왼쪽 경계 27.5는 제1사분위수이고, 오른쪽 경계 62.5는 제3사분위수이다. 사분위수 범위는 **제3사분위수에서 제1사분위수를 뺀 값**이므로 62.5 − 27.5 = 35이다.

NoteBase

문 4. $X_1, X_2, \ldots, X_n$**은 평균이** μ**, 분산이** σ^2**인 확률표본(random sample)이라고 하자. 표본평균** $\overline{X}$**에 대한 설명 중 옳은 것만을 모두 고르면?**

ㄱ. $\overline{X}$의 분산은 X_1의 분산보다 크다.
ㄴ. $\overline{X}$의 기댓값은 X_1의 기댓값과 같다.
ㄷ. $\overline{X}$의 분산은 n이 커질수록 작아진다.

① ㄱ, ㄴ
② ㄱ, ㄷ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 3번

간단한 해설

표본평균의 기댓값은 **모집단 평균과 같고**, 분산은 **모분산을 표본 크기로 나눈 값**인 σ^2/n이다. 여러 관측값의 평균을 내면 중심은 유지되면서 평균의 변동성은 줄어든다.

보기별 해설

확률표본의 각 관측값은 서로 독립이고 같은 분포를 따르므로, 모든 X_i의 기댓값은 μ, 분산은 σ^2이다.
표본평균은 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$로 정의된다.

ㄱ. X: 독립인 확률변수들의 합의 분산은 각 분산의 합이므로, $\sum_{i=1}^{n} X_i$의 분산은 $n\sigma^2$이다. 또 확률변수에 상수를 곱하면 분산에는 **그 상수의 제곱**이 곱해진다. 따라서 $\mathrm{Var}(\overline{X}) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$이다. 이는 X_1의 분산인 σ^2보다 크지 않으며, $n > 1$이고 $\sigma^2 > 0$이면 더 작다.

ㄴ. O: 기댓값의 선형성에 따라 $E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$이다. 각 기댓값이 μ이므로 $E(\overline{X}) = \frac{n\mu}{n} = \mu$이다. 따라서 표본평균과 개별 관측값 X_1은 **기댓값이 같다**.

ㄷ. O: 표본평균의 분산은 σ^2/n이다. **모분산이 양수로 일정할 때** 표본 크기 n이 커지면 분모가 커지므로 분산은 작아진다. 이는 표본을 반복해서 뽑을 때 표본평균이 모집단 평균 주위에 더 가깝게 모인다는 뜻이다.

옳은 보기는 ㄴ, ㄷ이므로 ③의 조합과 일치한다.

NoteBase

문 5. 세 확률변수 X_1, X_2, X_3은 서로 독립이고, 각각의 확률분포는 다음과 같다. 변환된 확률변수 $W = X_1 - 2X_2 + X_3$에 대해 $P(W > a) = 0.05$를 만족하는 실수 a의 값은? (단, z_α는 표준정규분포의 제 $100 \times (1-\alpha)$ 백분위수이다.)

$X_1 \sim N(2, 1^2), \quad X_2 \sim N(1, 1^2), \quad X_3 \sim N(2, 2^2)$

① $2+3\times z_{0.05}$

② $2+3\times z_{0.95}$

③ $2+4\times z_{0.05}$

④ $2+4\times z_{0.95}$

정답 1번

간단한 해설

W의 평균은 2이고 분산은 9이므로 **표준편차는 3**이다. 오른쪽 꼬리확률이 0.05인 경계에는 **상위 5%에 해당하는** $z_{0.05}$를 사용하므로 $a = 2 + 3z_{0.05}$이다.

계산 과정

평균: 기댓값은 각 확률변수의 계수를 그대로 곱해 더한다. 따라서 $E(W) = 2 - 2 \cdot 1 + 2 = 2$이다.

분산과 표준편차: 세 확률변수가 서로 독립이므로 공분산 항은 0이고, 각 분산에 **계수의 제곱**을 곱해 더한다. 주어진 분산은 각각 $1, 1, 4$이므로 $\mathrm{Var}(W) = 1 + (-2)^2 \cdot 1 + 4 = 9$이다. 음의 계수도 제곱하므로 분산을 빼지 않는다. 표준편차는 $\sqrt{9} = 3$이다.

분포와 표준화: 독립인 정규확률변수의 선형결합도 정규분포를 따르므로 $W \sim N(2, 3^2)$이다. 평균을 빼고 표준편차로 나누면 $Z = \frac{W-2}{3} \sim N(0, 1)$이 된다.

경곗값: $P(W > a) = 0.05$를 표준화하면 $P\left(Z > \frac{a-2}{3}\right) = 0.05$이다. 문제에서 z_α는 제$100(1-\alpha)$ 백분위수로 정의했으므로, **왼쪽 누적확률이 0.95**인 값은 $z_{0.05}$이다. 따라서 $\frac{a-2}{3} = z_{0.05}$이고, 이를 풀면 $a = 2 + 3z_{0.05}$로 ①의 식과 일치한다.

NoteBase

문 6. 어느 보험회사에서 도시 근로자의 평균 나이(μ)를 추정하기 위하여 64명을 임의로 추출하여 조사하였다. 64명 도시 근로자의 평균 나이가 36.38이고 표준편차가 11.07일 때, 모평균 μ에 대한 95% 신뢰구간은? (단, 표준정규분포를 따르는 확률변수 Z에 대하여 $P(Z \geq 1.96) = 0.025$, $P(Z \geq 1.645) = 0.05$이다)

① $\left(36.38 - 1.96 \times \frac{11.07}{8},\ 36.38 + 1.96 \times \frac{11.07}{8}\right)$

② $\left(36.38 - 1.96 \times \frac{11.07}{64},\ 36.38 + 1.96 \times \frac{11.07}{64}\right)$

③ $\left(36.38 - 1.645 \times \frac{11.07}{8},\ 36.38 + 1.645 \times \frac{11.07}{8}\right)$

④ $\left(36.38 - 1.645 \times \frac{11.07}{64},\ 36.38 + 1.645 \times \frac{11.07}{64}\right)$

정답 1번

간단한 해설

모평균의 신뢰구간에는 개인 나이의 표준편차가 아니라 **표본평균의 표준오차**인 $\frac{11.07}{\sqrt{64}} = \frac{11.07}{8}$을 사용한다. 95% 양측 신뢰구간의 **임계값은 1.96**이므로, 36.38세를 중심으로 약 2.71세씩 범위를 잡는다.

계산 과정

주어진 값은 표본크기 $n = 64$, 표본평균 $\bar{x} = 36.38$세, 표본표준편차 $s = 11.07$세이다. 모표준편차가 주어지지 않았으므로, 큰 표본에 대한 정규근사를 이용하여 모평균의 95% 신뢰구간을 $\bar{x} \pm 1.96\frac{s}{\sqrt{n}}$으로 구한다.

표준오차: 표본평균이 표본 추출에 따라 얼마나 달라지는지를 나타내는 값이다. 개인 나이의 표준편차를 **표본크기의 제곱근**으로 나누므로 $\frac{s}{\sqrt{n}} = \frac{11.07}{8} = 1.38375$세이다. 64로 나누는 ②와 ④는 표준오차의 분모를 잘못 적용한 식이다.

임계값: 95% 양측 신뢰구간에서는 구간 밖의 확률 0.05를 **양쪽 꼬리에 0.025씩** 배분한다. 주어진 $P(Z \geq 1.96) = 0.025$와 표준정규분포의 대칭성에 따라 임계값은 1.96이다. ③과 ④의 1.645를 쓰면 양쪽 꼬리 확률의 합이 0.10이므로 90% 양측 신뢰구간이 된다.

신뢰구간: 오차한계는 $1.96 \times \frac{11.07}{8} = 2.71215$세이다. 따라서 하한은 $36.38 - 2.71215 = 33.66785$세이고, 상한은 $36.38 + 2.71215 = 39.09215$세이다. 구하는 구간은 약 33.67세에서 39.09세까지이며, 표본평균 36.38에 $1.96 \times \frac{11.07}{8}$을 빼고 더한 ①의 식과 같다.

NoteBase

문 7. 다음은 정부의 미세먼지 관련 정책에 대한 남녀 간 지지도의 차이를 알아보기 위해서 남녀 각각 100명씩을 조사하여 얻은 결과표이다. 남성의 지지율(p_1)보다 여성의 지지율(p_2)이 더 큰지 유의수준 5%에서 검정할 때, Z 검정통계량의 값(Z_0)과 기각역을 옳게 짝 지은 것은? (단, z_α는 표준정규분포의 제$100\times(1-\alpha)$ 백분위수이다)

구분		지지 여부		합계
		지지함	지지하지 않음	
성별	남성	40	60	100
	여성	60	40	100
합계		100	100	200

① Z 검정통계량: $Z_0 = \frac{0.4-0.6}{\sqrt{\frac{0.5\times0.5}{100}+\frac{0.5\times0.5}{100}}}$, 기각역: $Z_0 \geq z_{0.05}$

② Z 검정통계량: $Z_0 = \frac{0.4-0.6}{\sqrt{\frac{0.4\times0.6}{100}+\frac{0.6\times0.4}{100}}}$, 기각역: $Z_0 \geq z_{0.05}$

③ Z 검정통계량: $Z_0 = \frac{0.4-0.6}{\sqrt{\frac{0.5\times0.5}{100}+\frac{0.5\times0.5}{100}}}$, 기각역: $Z_0 \leq -z_{0.05}$

④ Z 검정통계량: $Z_0 = \frac{0.4-0.6}{\sqrt{\frac{0.4\times0.6}{100}+\frac{0.6\times0.4}{100}}}$, 기각역: $Z_0 \leq -z_{0.05}$

정답 3번

간단한 해설

두 지지율이 같다는 귀무가설의 경계에서 **합동비율 0.5**로 표준오차를 구하면 $Z_0 \approx -2.828$이다. 여성의 지지율이 더 크다는 대립가설은 $p_1 - p_2 < 0$이므로, **좌측검정**의 기각역 $Z_0 \leq -z_{0.05}$를 적용한다.

NoteBase

계산 과정

1. 표본비율: 표의 '지지함' 열에서 남성은 100명 중 40명, 여성은 100명 중 60명이다. 표본크기는 $n_1 = n_2 = 100$이고, 표본지지율은 각각 $\hat{p}_1 = \frac{40}{100} = 0.4$, $\hat{p}_2 = \frac{60}{100} = 0.6$이다. 두 집단의 표본이 독립적으로 추출되었다는 전제로 두 모비율 차이의 Z 검정을 적용한다.

2. 가설과 검정 방향: 여성의 모지지율이 더 큰지를 확인하므로 대립가설은 $H_1 : p_1 < p_2$, 귀무가설은 $H_0 : p_1 \geq p_2$이다. 남성에서 여성의 지지율을 뺀 차이가 작을수록 대립가설을 지지하므로 **왼쪽 꼬리에서 기각**한다. 검정통계량의 표준오차는 귀무가설의 경계인 $p_1 = p_2$에서 계산한다.

3. 합동비율: 두 모비율이 같다는 경계조건에서는 공통 지지율을 **두 표본의 지지 인원을 합쳐 추정**한다. 따라서 $\hat{p} = \frac{40 + 60}{100 + 100} = 0.5$이다. 남녀의 표본비율 0.4와 0.6을 각각 분산에 대입하지 않고, 공통 추정치 0.5를 양쪽에 사용한다.

4. 표준오차: 독립인 두 표본비율 차이의 분산은 각 표본비율의 분산을 더한 값이다. 귀무가설의 경계에서 그 추정값은 $s_0^2 = \frac{0.5 \cdot 0.5}{100} + \frac{0.5 \cdot 0.5}{100} = 0.005$이다. 따라서 $s_0 = \sqrt{0.005} \approx 0.07071$이며, 이는 두 지지율이 같을 때 표본비율 차이의 변동 크기를 나타낸다. 각 집단에서 공통비율로 계산한 지지·비지지 기대 인원이 모두 50명이므로 정규근사를 적용하기에 충분하다.

5. 검정통계량: 관측된 표본비율 차이에서 귀무가설 경계의 차이 0을 빼고 표준오차로 나눈다. 따라서 $Z_0 = \frac{0.4 - 0.6 - 0}{\sqrt{0.005}} \approx -2.828$이다. 음수인 것은 남성의 표본지지율이 여성보다 낮기 때문이다.

6. 기각역과 선택지 연결: 문제의 정의에 따라 $z_{0.05}$는 표준정규분포의 95백분위수로 약 1.645이다. 유의수준 5%의 좌측 기각역은 $Z_0 \leq -z_{0.05}$, 즉 약 −1.645 이하이다. **합동비율을 사용한 검정통계량과 좌측 기각역**을 함께 제시한 것은 ③이다. 관측값 −2.828은 기각역에 속하므로 귀무가설을 기각하며, 여성의 모지지율이 더 크다고 판단할 통계적 근거가 있다.

NoteBase

문 8. 두 확률변수 X, Y의 상관계수에 대한 설명으로 옳은 것만을 모두 고르면?

ㄱ. X와 Y의 상관계수가 0이면 X와 Y가 서로 독립이다.
ㄴ. X와 Y가 서로 독립이면 상관계수가 0이다.
ㄷ. $P\left(Y=\frac{X}{2}+1\right)=1$이면 X와 Y의 상관계수는 $\frac{1}{2}$이다.

① ㄴ
② ㄱ, ㄴ
③ ㄱ, ㄷ
④ ㄴ, ㄷ

정답 1번

간단한 해설

두 변수가 **독립이면 공분산이 0**이므로 상관계수도 0이다. 다만 상관계수가 0이라는 사실만으로 독립을 보장할 수는 없다.

보기별 해설

상관계수는 두 변수의 선형관계를 나타내며, $\rho_{XY}=\frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}$로 정의한다. 여기서 σ_X, σ_Y는 각 변수의 표준편차이다. 이 문항은 상관계수가 정의되는 경우, 즉 두 변수의 분산이 유한하고 양수인 경우를 전제로 한다.

ㄱ. X: **상관계수가 0이어도 비선형적인 의존관계가 있을 수 있다.** 예를 들어 X가 −1, 0, 1을 각각 같은 확률로 취하고 $Y=X^2$이라고 하자. 대칭성에 의해 $E(X)=0$, $E(XY)=E(X^3)=0$이므로 공분산과 상관계수는 0이다. 그러나 X가 0이면 Y는 0이고, X가 −1 또는 1이면 Y는 1로 정해지므로 두 변수는 독립이 아니다. 따라서 '상관계수가 0이면 서로 독립이다'는 일반적으로 성립하지 않는다.

ㄴ. O: 독립이면 $E(XY)=E(X)E(Y)$가 성립한다. 공분산은 $\mathrm{Cov}(X,Y)=E(XY)-E(X)E(Y)$이므로 0이 된다. 이를 두 표준편차의 곱으로 나눈 상관계수도 0이다. 즉, **독립은 무상관을 보장**한다.

ㄷ. X: 주어진 확률이 1이라는 것은 $Y=\frac{X}{2}+1$이 확률 1로 성립한다는 뜻이다. Y는 X에 양의 상수를 곱하고 상수를 더한 변수이므로 두 변수는 **완전한 양의 선형관계**에 있다. 상수 1을 더하는 것은 공분산과 표준편차에 영향을 주지 않고, X에 1/2을 곱하면 공분산과 Y의 표준편차가 모두 같은 비율로 줄어 상관계수에서는 그 비율이 약분된다. 따라서 상관계수는 1이다. 보기의 1/2은 직선의 기울기이며 상관계수가 아니다.

옳은 보기는 ㄴ만이므로 ①의 구성과 일치한다.

NoteBase

문 9. 다음은 카이제곱통계량을 이용하여 두 변수가 서로 독립인지 알아보기 위한 관측도수의 2×2 분할표이다. 카이제곱(χ^2) 검정에 대한 설명으로 옳지 않은 것은? (단, 귀무가설이 참일 때 각 셀의 기대도수는 5 이상이고, 카이제곱통계량의 값은 k이다)

구분		변수2		합계
		범주1	범주2	
변수1	범주1	O_{11}	O_{12}	n_{1+}
	범주2	O_{21}	O_{22}	n_{2+}
합계		n_{+1}	n_{+2}	n

① 관측도수가 O_{11}인 셀의 기대도수는 $\frac{(n_{1+}) \times (n_{+1})}{n}$과 같다.

② 관측도수가 O_{11}인 셀의 기대도수와 O_{12}인 셀의 기대도수의 합은 n_{1+}와 같다.

③ X가 자유도 1인 카이제곱분포를 따를 때, 유의확률은 $P(X \leq k)$와 같다.

④ 전체 관측도수의 합과 전체 기대도수의 합은 같다.

정답 3번

간단한 해설

카이제곱 독립성 검정은 통계량이 클수록 두 변수의 독립성을 의심하므로, 유의확률은 **오른쪽 꼬리확률**인 $P(X \geq k)$이다. ③의 $P(X \leq k)$는 왼쪽 누적확률이다.

NoteBase

선지별 해설

귀무가설은 '두 변수가 서로 독립이다'이다. 표에서 n_{1+}, n_{2+}는 행 합계, n_{+1}, n_{+2}는 열 합계, n은 전체 관측도수이다. 기대도수는 독립이라는 가정 아래 각 셀에 예상되는 도수이며, E_{ij}로 나타낸다.

① O: 독립이면 두 범주가 동시에 나타날 확률은 각 범주의 확률을 곱한 값이다. 표의 주변합계로 각 확률을 추정하므로, 기대도수는 **행 합계와 열 합계의 곱을 전체 도수로 나눈 값**이다. 따라서 $E_{11} = \frac{n_{1+} n_{+1}}{n}$이다.

② O: 첫째 행의 두 셀은 변수2의 모든 범주를 포함한다. 두 열의 비율은 $\frac{n_{+1}}{n} + \frac{n_{+2}}{n} = 1$이므로, 첫째 행의 기대도수를 합하면 **그 행의 관측도수 합계가 그대로 유지**된다. 즉, $E_{11} + E_{12} = n_{1+}$이다.

③ X: 2×2 분할표의 자유도는 $(2-1) \times (2-1) = 1$이다. 각 셀의 기대도수가 5 이상이라는 조건에서 귀무가설 아래 검정통계량의 분포를 자유도 1인 카이제곱분포로 근사한다. 검정통계량은 각 셀에서 관측도수와 기대도수의 차이를 제곱하여 기대도수로 나눈 값을 더하므로, 값이 클수록 독립이라는 가정과의 불일치가 크다. 따라서 유의확률은 관측된 값 k 이상이 나올 **오른쪽 꼬리확률**인 $P(X \geq k)$로 구한다. 선지의 $P(X \leq k)$는 왼쪽 누적확률이므로 이를 $P(X \geq k)$로 고쳐야 한다.

④ O: 첫째 행의 기대도수 합은 n_{1+}, 둘째 행의 기대도수 합은 n_{2+}이다. 따라서 전체 기대도수의 합은 $n_{1+} + n_{2+} = n$으로, **전체 관측도수의 합과 같다**.

NoteBase

문 10. F-분포에 대한 설명으로 옳은 것만을 모두 고르면? (단, $F_\alpha(k_1, k_2)$는 분자의 자유도가 k_1이고 분모의 자유도가 k_2인 F-분포의 제$100 \times (1-\alpha)$ 백분위수이다)

ㄱ. 자유도 k_1, k_2에 대해 항상 $F_\alpha(k_1, k_2) \times F_{1-\alpha}(k_2, k_1) = 1$이다.
ㄴ. T가 자유도 k인 t-분포를 따를 때, 확률변수 $\frac{1}{T^2}$은 분자의 자유도가 k이고 분모의 자유도가 1인 F-분포를 따른다.
ㄷ. 서로 독립인 두 확률변수 Z_1과 Z_2가 표준정규분포를 따를 때, 확률변수 $\left(\frac{Z_1}{Z_2}\right)^2$은 분자의 자유도가 1이고 분모의 자유도가 1인 F-분포를 따른다.

① ㄱ, ㄴ
② ㄱ, ㄷ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 4번

간단한 해설

F-분포의 **역수는 분자와 분모의 자유도가 서로 바뀐 F-분포**를 따른다. 또한 서로 독립인 표준정규변수의 제곱은 각각 자유도 1인 카이제곱분포를 따르므로, 그 비는 두 자유도가 모두 1인 F-분포를 따른다.

NoteBase

보기별 해설

서로 독립인 $U \sim \chi^2_m$, $V \sim \chi^2_n$에 대해 $\frac{U/m}{V/n} \sim F(m, n)$이다. 각 카이제곱변수를 자신의 자유도로 나눈 뒤 비를 취하며, 분자와 분모의 순서가 자유도의 순서를 결정한다.

ㄱ. O: $X \sim F(k_1, k_2)$이면 $1/X \sim F(k_2, k_1)$이다. $0 < \alpha < 1$에서 $q = F_\alpha(k_1, k_2)$라 두면 $P(X \geq q) = \alpha$이다. 역수는 크기 순서를 뒤집으므로 $P(1/X \leq 1/q) = \alpha$이다. 따라서 $1/q$는 자유도가 뒤바뀐 F-분포의 **누적확률이 α인 백분위수**이고, 문제의 표기에서는 $F_{1-\alpha}(k_2, k_1)$에 해당한다. 즉 $F_{1-\alpha}(k_2, k_1) = 1/q$이므로 ㄱ의 두 백분위수의 곱은 1이다.

ㄴ. O: 자유도 k인 t-분포의 확률변수는 서로 독립인 $Z \sim N(0,1)$, $U \sim \chi^2_k$를 이용하여 $T = Z/\sqrt{U/k}$로 나타낼 수 있다. 따라서 $T^2 = \frac{Z^2/1}{U/k}$이고, $Z^2 \sim \chi^2_1$이므로 $T^2 \sim F(1, k)$이다. 역수를 취하면 $\frac{1}{T^2} = \frac{U/k}{Z^2/1}$이므로 **분자의 자유도는 k, 분모의 자유도는 1**이다.

ㄷ. O: 표준정규변수의 제곱은 자유도 1인 카이제곱분포를 따르므로 $Z_1^2 \sim \chi^2_1$, $Z_2^2 \sim \chi^2_1$이다. 원래 두 변수가 독립이므로 **각각 제곱한 변수도 서로 독립**이다. 또한 $\left(\frac{Z_1}{Z_2}\right)^2 = \frac{Z_1^2/1}{Z_2^2/1}$이므로 F-분포의 정의에 따라 분자와 분모의 자유도가 모두 1이다.

따라서 ㄱ, ㄴ, ㄷ이 모두 성립하며, 이 조합은 ④에 해당한다.

NoteBase

문 11. 중심극한정리에 대한 설명으로 ㉠, ㉡에 들어갈 말을 옳게 짝 지은 것은? (단, 모집단의 평균이 μ이고, 분산 σ^2은 존재한다)

표본크기가 충분히 클 때, 임의의 분포에서 추출한 확률표본의
(㉠)은 근사적으로 (㉡)를 따른다.

① ㉠: 표본평균, ㉡: 카이제곱분포

② ㉠: 표본평균, ㉡: 균등분포

③ ㉠: 표준화 표본평균, ㉡: 지수분포

④ ㉠: 표준화 표본평균, ㉡: 표준정규분포

정답 4번

간단한 해설

중심극한정리는 표본크기가 충분히 크면 **표준화 표본평균**이 근사적으로 **표준정규분포**를 따른다는 정리이다. 모집단 자체가 정규분포일 필요는 없지만 분산은 유한해야 한다.

빈칸별 해설

㉠ 표준화 표본평균: 같은 모집단에서 독립적으로 추출한 크기 n의 확률표본에서 표본평균을 $\bar{X}$라 하자. 모집단의 평균이 μ, 유한한 양의 분산이 σ^2이면 표본평균의 평균은 μ, 표준편차는 σ/ n이다. 표준화 표본평균은 Z_n=(X-μ)/(σ/ n)으로, **평균을 빼고 표본평균의 표준편차로 나눈 값**이다.

㉡ 표준정규분포: 중심극한정리에 따라 표본크기가 충분히 크면 Z_n의 분포는 평균 0, 분산 1인 **표준정규분포** $N(0, 1)$에 가까워진다. 표본평균 자체는 평균이 μ이고 분산이 σ^2/n이므로, 일반적으로 표준정규분포에 바로 대응하지 않는다. 모집단의 분포 모양이 아니라 표준화 표본평균의 근사분포를 묻는다는 점이 핵심이다.

㉠은 표준화 표본평균, ㉡은 표준정규분포이므로 이 짝에 해당하는 선택지는 ④이다.

NoteBase

문 12. 다음은 세 가지 속독법(A, B, C)에 따라 책 읽는 시간에 차이가 있는지 알아보기 위해 일원배치분산분석법을 적용하여 얻은 분산분석표이다. 각 속독법에 5명씩 15명을 임의로 배치하여 책을 읽게 한 후, 책 읽는 시간을 측정하였다. 이에 대한 설명으로 옳지 않은 것은?

구분	제곱합	자유도	평균제곱	$F-$값	$p-$값
처리	2156	(㉠)		16.84	0.0003
오차	768	(㉡)	(㉢)		
합계	2924	14			

① ㉠의 값은 3이다.

② ㉡의 값은 12이다.

③ ㉢의 값은 64이다.

④ 유의수준 1%에서 검정할 때, 세 가지 속독법에 따라 책 읽는 시간에 차이가 있다고 할 수 있다.

정답 1번

간단한 해설

㉠은 속독법의 개수가 아니라 **처리 자유도**이다. 속독법이 3가지이므로 처리 자유도는 3 − 1 = 2이다.

NoteBase

선지별 해설

① X: ㉠은 처리, 즉 속독법 간 변동의 자유도이다. 집단 수가 k이면 처리 자유도는 $k-1$이다. 세 집단의 평균과 전체 평균 사이의 편차에는 하나의 제약이 있으므로, 집단이 3개여도 독립적으로 변할 수 있는 편차는 2개이다. 따라서 ㉠**은 3이 아니라 2**이다.

② O: ㉡은 집단 내 변동에 해당하는 오차 자유도이다. 전체 관측값 수를 N, 집단 수를 k라 하면 오차 자유도는 $N-k$이다. 각 집단의 평균을 구하는 데 자유도를 하나씩 사용하므로 $15-3=12$이다. 표의 합계 자유도 14에서 처리 자유도 2를 빼도 12가 된다.

③ O: ㉢은 오차 평균제곱이다. 평균제곱은 **제곱합을 해당 자유도로 나눈 값**이므로, 표의 오차 제곱합 768과 오차 자유도 12를 이용하면 $\frac{768}{12}=64$이다.

④ O: 일원배치분산분석의 귀무가설은 세 속독법의 모집단 평균 책 읽는 시간이 모두 같다는 것이다. 표의 p값은 0.0003으로 유의수준 1%에 해당하는 0.01보다 작으므로 귀무가설을 기각한다. 따라서 **적어도 한 속독법의 평균 책 읽는 시간은 다르다**고 할 수 있다. 이 결과만으로 세 속독법의 평균이 각각 모두 다르다고 판단하는 것은 아니다.

NoteBase

문 13. 다음은 입학 시 수학 성적(X)과 1학년 때의 통계학 성적(Y)에 대하여 단순선형회귀모형 $Y_i = \alpha + \beta X_i + \varepsilon_i,\ i = 1, 2, \cdots, n$을 적용하여 얻은 결과이다. 이에 대한 설명으로 옳은 것은? (단, $F_\alpha(k_1, k_2)$는 분자의 자유도가 k_1이고 분모의 자유도가 k_2인 F-분포의 제$100 \times (1-\alpha)$ 백분위수를 나타내고, $F_{0.05}(1, 10) = 4.96,\ F_{0.05}(1, 11) = 4.84$이다. 그리고 $t_\alpha(k)$는 자유도가 k인 t-분포의 제$100 \times (1-\alpha)$ 백분위수를 나타내고, $t_{0.05}(10) = 1.812,\ t_{0.025}(10) = 2.228,\ t_{0.025}(11) = 2.201$이다.)

요인	제곱합	자유도	평균제곱	$F-$값
회귀	541.69	1	541.69	29.04
잔차	186.56	10	18.66	

	회귀계수	표준오차	$t-$값
상수항	30.04	10.14	2.96
X	0.90	0.17	5.34

① 자료의 개수 n는 11이다.

② 추정된 회귀직선은 $\hat{Y} = 10.14 + 0.17X$이다.

③ X와 Y 사이의 모상관계수(ρ)가 0인지 검정할 때, 귀무가설 $H_0 : \rho = 0$은 유의수준 5%에서 기각되지 않는다.

④ 추정된 회귀모형의 유의성을 검정할 때, 귀무가설(H_0: 회귀모형은 유의하지 않다)은 유의수준 5%에서 기각된다.

정답 4번

간단한 해설

회귀모형의 F값은 29.04로 **5% 임계값 4.96보다 크다**. 따라서 기울기가 0이라는 귀무가설을 기각하며, 회귀모형은 유의하다.

NoteBase

선지별 해설

① X: 상수항을 포함한 단순선형회귀에서는 절편과 기울기 두 모수를 추정하므로 잔차의 자유도는 $n-2$이다. 분산분석표의 잔차 자유도가 10이므로 $n-2=10$에서 **자료의 개수는 12**이다. 11은 전체 자유도인 $n-1$에 해당한다.

② X: 추정 회귀직선에는 표의 **회귀계수 열**을 사용한다. 상수항의 회귀계수는 30.04, X의 회귀계수는 0.90이므로 추정 회귀직선은 $\hat{Y}=30.04+0.90X$이다. 선지의 10.14와 0.17은 각각 절편과 기울기 추정량의 표준오차로, 회귀직선의 계수가 아니다.

③ X: 주어진 단순선형회귀모형에서 모상관계수가 0이라는 가설은 기울기가 0이라는 가설과 같은 선형 관계의 유무를 검정한다. 따라서 X 행의 t값 5.34를 사용한다. 대립가설은 $H_1:\rho\neq 0$이므로 **양측검정**을 하며, 잔차 자유도는 10이다. 유의수준 5%의 임계값은 $t_{0.025}(10)=2.228$이고, $|t|=5.34>2.228$이므로 귀무가설 $H_0:\rho=0$을 기각한다.

④ O: 상수항을 포함한 단순선형회귀모형의 유의성 검정은 $H_0:\beta=0$과 $H_1:\beta\neq 0$을 검정한다. 귀무가설은 X가 Y를 설명하는 선형 효과가 없다는 뜻이다. 검정통계량은 **회귀평균제곱을 잔차평균제곱으로 나눈 값**으로, $F=\frac{541.69}{18.66}\approx 29.04$이다. 분자 자유도는 1, 분모 자유도는 10이므로 5% 임계값은 $F_{0.05}(1,10)=4.96$이다. $29.04>4.96$이므로 귀무가설을 기각하고 회귀모형이 유의하다고 판단한다.

NoteBase

문 14. 확률변수 X와 Y의 분산과 공분산은 다음과 같다. 확률변수 W와 T를 각각 $W = 2X + 2$, $T = -Y + 1$ 이라고 할 때, W와 T의 상관계수는?

$V(X) = 25, \quad V(Y) = 16, \quad Cov(X, Y) = -10$

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

정답 2번

간단한 해설

상수를 더해도 상관계수는 변하지 않으며, 한 변수에만 음수를 곱하면 **상관계수의 부호가 반대로 바뀐다**. 원래 상관계수 $-\frac{1}{2}$는 변환 후 $\frac{1}{2}$이 된다.

계산 과정

변환 규칙: 상관계수는 공분산을 두 확률변수의 표준편차의 곱으로 나눈 값이다. 상수를 더하면 평균도 같은 만큼 이동하여 평균으로부터의 편차는 변하지 않는다. 따라서 분산과 공분산을 구할 때 더해진 상수 2와 1은 영향을 주지 않는다.

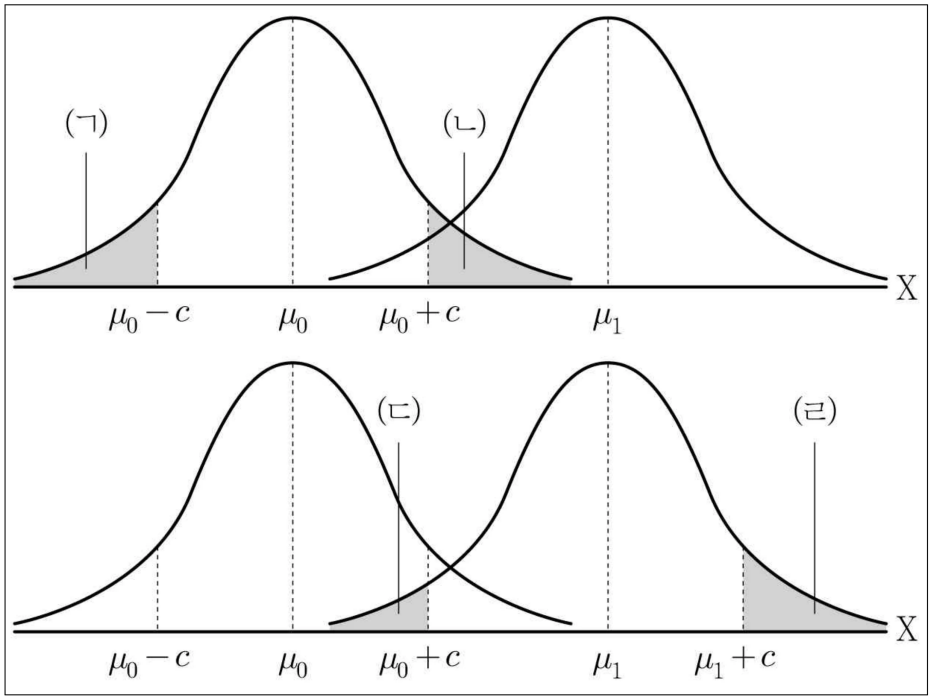
분산과 표준편차: 확률변수에 상수를 곱하면 분산에는 **그 상수의 제곱**이 곱해진다. 따라서 $V(W) = 2^2 \cdot 25 = 100$, $V(T) = (-1)^2 \cdot 16 = 16$이다. 상관계수의 분모에 쓰는 표준편차는 각각 $\sqrt{100} = 10$, $\sqrt{16} = 4$이다.

공분산: 두 확률변수에 각각 상수를 곱하면 공분산에는 **두 상수의 곱**이 곱해진다. 따라서 $\mathrm{Cov}(W, T) = 2 \cdot (-1) \cdot (-10) = 20$이다. 분산과 달리 계수를 제곱하지 않으므로, 음의 계수로 인해 공분산의 부호가 음수에서 양수로 바뀐다.

상관계수: 두 분산이 모두 양수이므로 $\rho_{WT} = \frac{\mathrm{Cov}(W, T)}{\sqrt{V(W)}\sqrt{V(T)}}$를 적용한다. 계산하면 $\rho_{WT} = \frac{20}{10 \cdot 4} = \frac{1}{2}$이다. 이는 단위가 없는 값이며, ②에 해당한다.

NoteBase

문 15. 확률변수 X는 $N(\mu, 1)$를 따를 때, 가설 $H_0 : \mu = \mu_0$ 대 $H_1 : \mu = \mu_1$에 대한 기각역이 $R = \{x : x \geq \mu_0 + c\}$로 주어진 경우, 다음 그림에서 제1종 오류를 범할 확률에 해당하는 영역 (A)과 제2종 오류를 범할 확률에 해당하는 영역(B)을 옳게 짝 지은 것은? (단, $\mu_1 > \mu_0$이고, $c > 0$이다)



① A: ㄱ, B: ㄷ

② A: ㄱ, B: ㄹ

③ A: ㄴ, B: ㄷ

④ A: ㄴ, B: ㄹ

정답 3번

간단한 해설

제1종 오류는 **귀무가설이 참인데 기각하는 경우**이므로, 평균이 μ_0인 곡선에서 기각 기준 오른쪽의 ㄴ에 해당한다. 제2종 오류는 **대립가설이 참인데 귀무가설을 기각하지 못하는 경우**이므로, 평균이 μ_1인 곡선에서 같은 기준 왼쪽의 ㄷ에 해당한다.

NoteBase

대응 관계 해설

기각역이 $X \geq \mu_0 + c$이므로 관측값이 이 기준 이상이면 귀무가설을 기각하고, 기준보다 작으면 기각하지 않는다. 그림에서 왼쪽 곡선은 평균이 μ_0인 귀무가설하의 분포이고, 오른쪽 곡선은 평균이 μ_1인 대립가설하의 분포이다. 오류확률은 실제로 참인 가설의 분포에서 해당 영역의 넓이로 구한다.

(A) 제1종 오류 → ㄴ: 귀무가설이 참일 때 이를 기각하는 오류이다. 따라서 **평균이 μ_0인 곡선** 아래에서 $\mu_0 + c$ 오른쪽 넓이를 찾으면 된다. 이 확률은 $\alpha = P_{\mu_0}(X \geq \mu_0 + c)$이며, 그림 위쪽의 ㄴ이다. ㄱ은 $\mu_0 - c$ 왼쪽 넓이로, 주어진 기각역에 포함되지 않는다.

(B) 제2종 오류 → ㄷ: 대립가설이 참일 때 귀무가설을 기각하지 못하는 오류이다. 따라서 **평균이 μ_1인 곡선** 아래에서 $\mu_0 + c$ 왼쪽 넓이를 찾으면 된다. 이 확률은 $\beta = P_{\mu_1}(X < \mu_0 + c)$이며, 그림 아래쪽의 ㄷ이다. ㄹ은 $\mu_1 + c$ 오른쪽 넓이로, 귀무가설을 기각하지 않는 영역과 다르다.

따라서 (A)는 ㄴ, (B)는 ㄷ으로 연결되며, ③의 짝과 일치한다.

NoteBase

문 16. 다음은 다이어트 종류에 따라 체중 감량 효과에 차이가 있는지 알아보기 위해 분산분석을 시행한 결과표이다. 이 결과에서 알 수 있는 내용으로 옳지 않은 것은? (단, $F_{\alpha}(k_1, k_2)$는 분자의 자유도가 k_1이고 분모의 자유도가 k_2인 F-분포의 제100×$(1-\alpha)$ 백분위수를 나타내고, $F_{0.05}(3, 26) = 2.98$, $F_{0.025}(3, 26) = 3.67$이다)

요인	제곱합	자유도	평균제곱	F-값
다이어트	6	3		20
오차	2.6			
합계	8.6	29		

① 다이어트 종류는 4가지이다.

② F-값은 오차의 평균제곱을 처리의 평균제곱으로 나눈 값이다.

③ F-값과 분자의 자유도 3, 분모의 자유도가 26인 F-분포를 이용하여 유의확률(p-값)을 구할 수 있다.

④ 유의수준 5%에서 다이어트 종류에 따라 체중 감량 효과에 차이가 있다고 할 수 있다.

정답 2번

간단한 해설

②는 분자와 분모를 바꾸어 설명했다. 분산분석의 F-값은 **처리의 평균제곱을 오차의 평균제곱으로 나눈 값**이다.

NoteBase

선지별 해설

① O: 일원분산분석에서 처리의 자유도는 **집단 수보다 1 작다**. 표에서 다이어트의 자유도가 3이므로 다이어트 종류는 3 + 1 = 4가지이다.

② X: 처리는 다이어트 종류를 뜻하며, 처리의 평균제곱은 집단 간 변동을, 오차의 평균제곱은 집단 내 변동을 나타낸다. F-값은 **집단 간 변동을 집단 내 변동으로 나누어** 평균 차이가 큰지 판단하는 통계량이다. 평균제곱은 제곱합을 해당 자유도로 나눈 값이므로, 처리의 평균제곱은 $\frac{6}{3} = 2$이다. 오차의 자유도는 $29 - 3 = 26$이고 오차의 평균제곱은 $\frac{2.6}{26} = 0.1$이다. 따라서 $F = \frac{2}{0.1} = 20$이며, 선지의 나눗셈 방향은 반대이다.

③ O: F-통계량의 분자 자유도는 처리의 자유도 3이고, 분모 자유도는 오차의 자유도 26이다. 모든 다이어트의 모집단 평균 감량이 같다는 귀무가설 아래에서 F-통계량은 이 두 자유도를 갖는 F-분포를 따른다. 유의확률은 이 분포에서 **관측된 F-값 20 이상이 나올 확률**이므로, $p = P(F_{3,\,26} \geq 20)$으로 구할 수 있다.

④ O: 표의 F-값 20은 유의수준 5%의 임계값 2.98보다 크므로, 모든 다이어트의 모집단 평균 감량이 같다는 귀무가설을 기각한다. 따라서 다이어트 종류에 따른 체중 감량 효과에 차이가 있다고 판단한다. 이는 **적어도 한 집단의 모집단 평균이 다르다**는 뜻이며, 모든 다이어트 사이에 각각 차이가 있다는 뜻은 아니다.

NoteBase

문 17. 자료의 수가 n인 표본 (x_i, y_i) $(i = 1, 2, \ldots, n)$에 대해 다음 두 회귀모형 M_1과 M_2를 적용하여 분석하고자 한다. 두 모형에 대한 설명으로 옳은 것만을 모두 고르면?

$M_1 : Y_i = \alpha + \varepsilon_i$,
$M_2 : Y_i = \alpha + \beta X_i + \varepsilon_i$ $(i = 1, 2, \ldots, n)$
ㄱ. 모형 M_1에서 $\hat{Y}_i = \overline{Y}$이다.
ㄴ. 모형 M_2의 결정계수는 0 이상이다.
ㄷ. 모형 M_2의 회귀제곱합은 모형 M_1의 회귀제곱합보다 크거나 같다.

① ㄱ, ㄴ
② ㄱ, ㄷ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 4번

간단한 해설

상수항만 있는 모형은 모든 관측값을 **표본평균**으로 예측한다. 여기에 설명변수를 추가해 최소제곱법으로 적합하면 **잔차제곱합이 늘지 않으므로**, 평균을 기준으로 모형이 설명하는 변동은 줄지 않는다.

NoteBase

보기별 해설

두 모형을 통상적인 최소제곱법으로 적합하여 판단한다. 결정계수는 관측값의 변동이 있어 총제곱합이 양수인 경우를 전제로 한다.

ㄱ. O: 모형 M_1은 설명변수 없이 상수항만 있으므로 모든 적합값이 $\hat{\alpha}$로 같다. 최소제곱법은 $\sum_{i=1}^{n}(Y_i-\alpha)^2$을 최소화하며, 그 조건은 $\sum_{i=1}^{n}(Y_i-\hat{\alpha})=0$이다. 따라서 $\hat{\alpha}=\overline{Y}$이므로 **모든 적합값이 표본평균**인 $\hat{Y}_i=\overline{Y}$가 된다.

ㄴ. O: 총제곱합 SST는 관측값과 표본평균 차이의 제곱합, 회귀제곱합 SSR은 적합값과 표본평균 차이의 제곱합, 잔차제곱합 SSE는 관측값과 적합값 차이의 제곱합이다. 모형 M_2처럼 **상수항을 포함한 최소제곱회귀**에서는 $SST=SSR+SSE$가 성립한다. 제곱합은 0 이상이므로, 결정계수 $R^2=\frac{SSR}{SST}$는 0≤ R^2≤1을 만족한다.

ㄷ. O: 모형 M_1에서는 적합값이 표본평균이므로 $SSR_1=\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2=0$이다. 모형 M_2의 회귀제곱합도 제곱합이므로 SSR_2≥0=SSR_1이다. 또한 M_2에서 $\beta=0$으로 두면 M_1이 되므로, M_2의 최소제곱 적합은 M_1보다 잔차제곱합을 크게 만들지 않는다. 같은 자료에서 SST는 일정하므로, **잔차제곱합이 감소하면 회귀제곱합은 증가**하며 설명변수의 추가로 적합이 개선되지 않으면 두 회귀제곱합이 같을 수도 있다.

ㄱ, ㄴ, ㄷ이 모두 성립하므로 이를 함께 제시한 ④에 해당한다.

NoteBase

문 18. 어떤 자판기에서 판매되는 음료수 용량은 모평균이 μ(mL)이고, 모표준편차가 5mL인 확률분포를 따른다고 한다. 이 자동판매기에서 임의로 추출한 100개 음료수의 표본평균이 150mL일 때, 가설 $H_0 : \mu = \mu_0$ 대 $H_1 : \mu \neq \mu_0$에 대한 유의수준 α에서 귀무가설을 기각하지 못하는 μ_0의 범위는? (단, z_α는 표준정규분포의 제$100 \times (1-\alpha)$ 백분위수이다)

① $\left(150 - \frac{1}{2}z_\alpha,\quad 150 + \frac{1}{2}z_\alpha\right)$
② $\left(150 - \frac{1}{2}z_{\alpha/2},\quad 150 + \frac{1}{2}z_{\alpha/2}\right)$
③ $\left(150 - \frac{1}{4}z_\alpha,\quad 150 + \frac{1}{4}z_\alpha\right)$
④ $\left(150 - \frac{1}{4}z_{\alpha/2},\quad 150 + \frac{1}{4}z_{\alpha/2}\right)$

정답 2번

간단한 해설

표본평균의 표준오차는 0.5mL이며, 양측검정에서는 **양쪽 꼬리에 유의수준을 절반씩** 배분한다. 기각하지 못하는 가설값은 150mL와의 차이가 $\frac{1}{2}z_{\alpha/2}$mL보다 작은 범위이다.

계산 과정

표본평균을 $\bar{X}$라 하면, 모표준편차 $\sigma = 5$mL와 표본크기 n = 100이 주어져 있다. 표본평균의 표준오차는 $\frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{100}} = \frac{1}{2}$mL이다. 모표준편차 자체가 아니라 **표본평균의 표준오차**로 표준화한다.

귀무가설 $H_0 : \mu = \mu_0$ 아래에서 검정통계량은 $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$이다. 모집단이 정규분포라는 조건은 없지만, 표본크기가 100으로 크므로 중심극한정리에 따른 표준정규근사를 이용한다. 대립가설이 $\mu \neq \mu_0$인 **양측검정**이므로 양쪽 꼬리확률을 각각 $\alpha/2$로 두며, 임계값은 $\pm z_{\alpha/2}$이다.

관측된 표본평균 150mL를 대입하면 기각하지 못하는 조건은 $-z_{\alpha/2} < \frac{150 - \mu_0}{1/2} < z_{\alpha/2}$이다. 이를 μ_0에 관해 정리하면 $150 - \frac{1}{2}z_{\alpha/2} < \mu_0 < 150 + \frac{1}{2}z_{\alpha/2}$가 된다. 즉, 가설값과 관측된 표본평균의 차이가 임계값에 표준오차를 곱한 오차한계보다 작아야 한다.

따라서 구하는 범위는 $\left(150 - \frac{1}{2}z_{\alpha/2},\, 150 + \frac{1}{2}z_{\alpha/2}\right)$mL이며, ②와 일치한다.

NoteBase

문 19. 다음은 금연 프로그램에 참석한 120명을 대상으로 직업군에 따라 금연 성공률에 차이가 있는지 조사한 분할표이다. 이에 대한 설명으로 옳은 것은? (단, $O_{ij}(i=1,2,3,4,\ j=1,2)$는 (i,j) 셀에서 얻어진 관측도수이고, $E_{ij}(i=1,2,3,4,\ j=1,2)$는 귀무가설이 참일 때 (i,j) 셀에서 얻어진 기대도수이다. $\chi^2_\alpha(k)$는 자유도가 k인 카이제곱분포의 제$100\times(1-\alpha)$ 백분위수이고, χ^2은 검정통계량이다.)

구분		금연		합계
		성공함(1)	성공하지 못함(2)	
직업군	사무직(1)	15	15	30
	자영업(2)	15	10	25
	교육관련(3)	12	18	30
	노동직(4)	18	17	35
합계		60	60	120

① 각 직업의 성공률을 $p_i(i=1,2,3,4)$라고 할 때, 귀무가설은 $H_0: p_1=p_2=p_3=p_4=\frac{1}{4}$이다.

② $E_{11}=15$이다.

③ 카이제곱 검정통계량은 $\chi^2=\sum_{i=1}^{4}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{O_{ij}}$이다.

④ 유의수준 5%에서 검정할 때, 기각역은 $\chi^2\geq\chi^2_{0.025}(2)$이다.

정답 2번

간단한 해설

직업군별 성공률이 같다는 귀무가설 아래에서는 각 직업군에 전체 성공 비율을 적용해 **기대도수**를 구한다.
사무직은 30명이고 전체 성공 비율은 $\frac{60}{120}$이므로, 사무직 성공자의 기대도수는 $E_{11}=30\times\frac{60}{120}=15$명이다.

NoteBase

선지별 해설

① X: 검정하려는 것은 네 직업군의 **금연 성공률이 같은지**이다. 따라서 귀무가설은 $H_0 : p_1 = p_2 = p_3 = p_4$이며, 대립가설은 적어도 한 직업군의 성공률이 다르다는 것이다. 각 성공률은 서로 다른 직업군 안에서의 비율이므로 그 합이 1일 필요가 없고, 공통 성공률을 1/4로 정할 근거도 없다. 귀무가설 아래 공통 성공률은 전체 성공자 60명을 전체 참석자 120명으로 나눈 $\hat{p} = \frac{60}{120} = \frac{1}{2}$로 추정한다.

② O: 귀무가설 아래에서는 직업군과 금연 성공 여부가 독립이므로, 기대도수는 **해당 행 합계와 열 합계의 곱을 전체 합계로 나눈 값**이다. 첫 번째 행은 사무직, 첫 번째 열은 성공함에 해당한다. 사무직의 행 합계는 30명, 성공함의 열 합계는 60명, 전체 합계는 120명이므로 $E_{11} = \frac{30 \times 60}{120} = 15$명이다. 이는 직업군별 성공률이 같을 때 사무직 30명 중 성공할 것으로 기대되는 인원이다.

③ X: 피어슨 카이제곱 검정통계량은 $\chi^2 = \sum_{i=1}^{4}\sum_{j=1}^{2}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$이다. 각 셀에서 관측도수와 기대도수의 차이를 제곱한 뒤 **관측도수가 아니라 기대도수로 나누어** 합한다. 이렇게 하면 기대도수의 크기를 고려해 차이를 평가할 수 있으며, 통계량이 클수록 귀무가설에서 예상한 도수와 실제 도수의 차이가 크다는 뜻이다.

④ X: 직업군 4개와 금연 성공 여부 2개로 이루어진 분할표이므로 자유도는 $(4-1)(2-1) = 3$이다. 이 검정에서는 관측도수가 기대도수에서 크게 벗어날 때 귀무가설을 기각하므로 **오른쪽 꼬리에 유의수준 5%를 배정**한다. 문제의 백분위수 표기에 따르면 기각역은 $\chi^2 \geq \chi^2_{0.05}(3)$이다. 선지는 자유도를 2로 쓰고 꼬리확률을 0.025로 나누어 적용한 점이 잘못되었다.

NoteBase

문 20. 다음 설명 중 옳은 것만을 모두 고르면?

ㄱ. 유의확률(p-값)이 유의수준보다 작을 때 귀무가설을 기각한다.
ㄴ. 모수 θ에 관한 불편추정량(unbiased estimator)의 기댓값은 θ이다.
ㄷ. 검정에서 제1종 오류의 확률을 줄이면 제2종 오류의 확률도 줄어든다.

① ㄱ
② ㄱ, ㄴ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 2번

간단한 해설

같은 표본크기와 검정 방식에서 기각 기준을 엄격하게 하면 **제1종 오류는 줄지만 제2종 오류는 일반적으로 늘어난다**. 귀무가설이 거짓이어도 기각하기 어려워지기 때문이다.

보기별 해설

ㄱ. O: 유의확률 p는 귀무가설이 참이라고 가정할 때 관측된 결과만큼 또는 그보다 더 극단적인 결과가 나올 확률이다. 유의수준 α는 귀무가설을 기각할 판단 기준이다. **유의확률이 유의수준보다 작으면**, 관측된 결과가 귀무가설 아래에서 충분히 드물다고 보아 귀무가설을 기각한다.

ㄴ. O: 모수 θ의 추정량을 $\hat{\theta}$라고 할 때, **불편추정량은 기댓값이 추정하려는 모수와 같은 추정량**으로, $E(\hat{\theta}) = \theta$를 만족한다. 같은 방식으로 표본을 반복 추출하여 추정값을 구하면 그 추정값들의 평균이 모수를 중심으로 한다는 뜻이며, 개별 표본의 추정값이 항상 모수와 같다는 뜻은 아니다.

ㄷ. X: 제1종 오류는 참인 귀무가설을 기각하는 오류이고, 제2종 오류는 거짓인 귀무가설을 기각하지 못하는 오류이다. 같은 표본크기와 검정 방식에서 기각역을 줄이면 참인 귀무가설을 잘못 기각할 확률은 낮아지지만, 거짓인 귀무가설도 기각하기 어려워진다. 따라서 이 조건에서는 **제1종 오류의 확률을 줄이면 제2종 오류의 확률은 일반적으로 증가한다**고 고쳐야 한다.

옳은 보기는 ㄱ과 ㄴ이며, 이 조합은 ②에 해당한다.

NoteBase