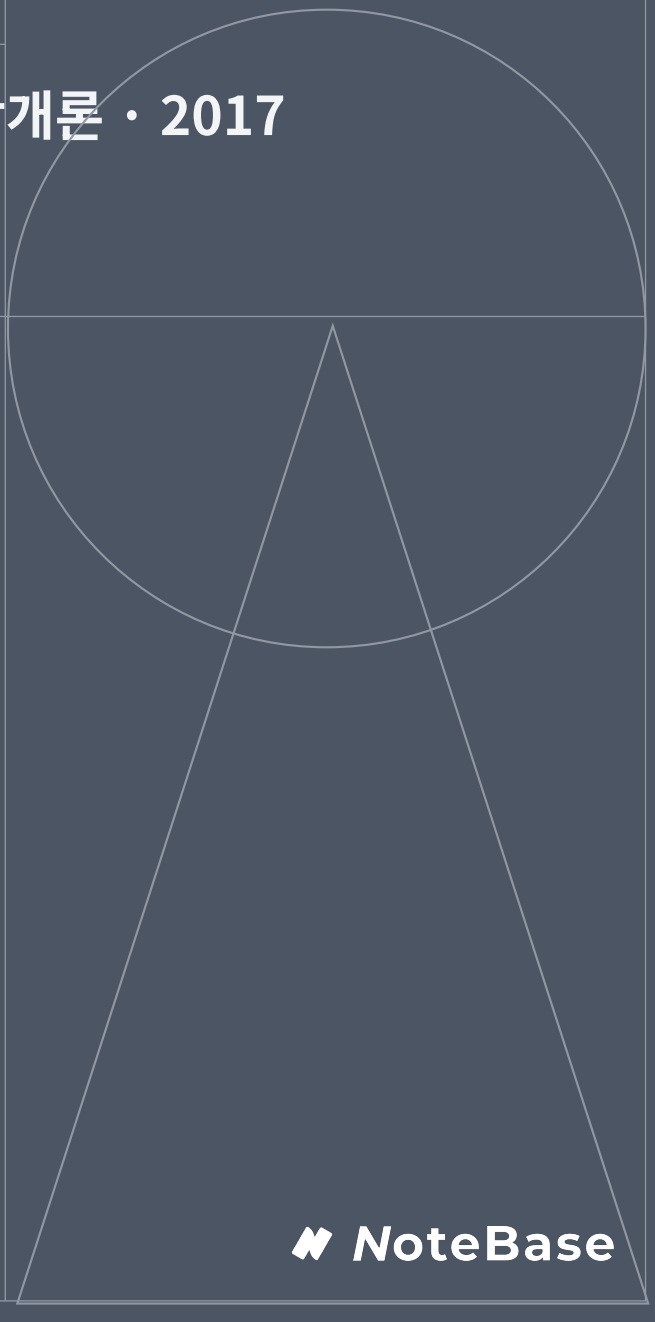
기출OX 해설집

공무원/국가직/9급 · 통계학개론 · 2017



NoteBase

기출OX 해설집

공무원/국가직/9급 · 통계학개론 · 2017

공무원/국가직/9급 · 통계학개론 · 2017

문 1. 다음 자료에 대한 설명으로 옳지 않은 것은?

0, 0, 0, 0, 0, 0, 1, 3, 5, 7

① 평균은 $\frac{8}{5}$이다.

② 중앙값(median)은 0이다.

③ 최빈값(mode)은 0이다.

④ 10% 절사평균(trimmed mean)은 1이다.

정답 4번

간단한 해설

10% 절사평균은 작은 쪽과 큰 쪽에서 **각각 자료의 10%씩** 제외하고 구한다. 이 자료에서는 0과 7을 하나씩 제외하므로 남은 8개 자료의 평균은 $\frac{9}{8}$이며, 1이 아니다.

선지별 해설

① O: 평균은 자료의 합을 자료의 개수로 나눈 값이다. 자료는 10개이고 합은 1 + 3 + 5 + 7 = 16이므로, 평균은 $\frac{16}{10} = \frac{8}{5}$이다.

② O: 중앙값은 크기순으로 정렬한 자료의 가운데 값이다. 자료가 10개로 짝수이므로 **5번째 값과 6번째 값의 평균**을 구한다. 두 값이 모두 0이므로 중앙값은 $\frac{0+0}{2} = 0$이다.

③ O: 최빈값은 **가장 자주 나타나는 값**이다. 0은 6번 나타나고 1, 3, 5, 7은 각각 한 번씩 나타나므로 최빈값은 0이다.

④ X: 10% 절사평균은 크기순으로 정렬한 자료의 **양끝에서 각각 10%씩 제외**한 뒤 남은 자료로 구한 평균이다. 자료가 10개이므로 가장 작은 값 0과 가장 큰 값 7을 하나씩 제외한다. 남은 자료는 0, 0, 0, 0, 0, 1, 3, 5로, 개수는 8개이고 합은 9이다. 따라서 10% 절사평균은 1이 아니라 $\frac{9}{8}$이다.

NoteBase

문 2. 다음 측도 중에서 자료의 단위에 의존하지 않는 것만을 모두 고른 것은?

ㄱ. 변동계수(coefficient of variation)
ㄴ. 상관계수(correlation coefficient)
ㄷ. 결정계수(coefficient of determination)

① ㄱ, ㄴ
② ㄱ, ㄷ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 4번

간단한 해설

세 측도는 모두 **분자와 분모의 단위가 약분되는 무차원 측도**이다. 따라서 cm를 m로 바꾸는 것처럼 측정 단위를 환산해도 값이 달라지지 않는다.

보기별 해설

ㄱ. 변동계수: 평균이 0이 아닐 때 표준편차를 평균으로 나눈 값으로, $CV = \frac{s}{\bar{x}}$이다. 여기서 s는 표준편차, $\bar{x}$는 평균이다. **표준편차와 평균은 원자료와 같은 단위**를 가지므로 나누면 단위가 사라진다. 단위를 환산하면 표준편차와 평균이 같은 비율로 바뀌어 변동계수는 유지된다.

ㄴ. 상관계수: 두 변수 사이의 선형 관계의 방향과 강도를 나타내며, $r = \frac{s_{xy}}{s_x s_y}$로 계산한다. 여기서 s_{xy}는 공분산, s_x와 s_y는 각 변수의 표준편차이며 두 표준편차가 모두 0이 아니어야 한다. 공분산의 단위는 두 변수 단위의 곱이므로, 이를 **두 표준편차의 곱으로 나누면 단위가 약분**된다. 따라서 각 변수의 측정 단위를 환산해도 상관계수는 변하지 않는다.

ㄷ. 결정계수: 절편을 포함한 최소제곱 회귀에서 종속변수의 전체 변동 중 회귀모형이 설명하는 변동의 비율이다. 전체 변동이 0이 아닐 때 정의되며, 설명된 변동과 전체 변동은 모두 종속변수 단위의 제곱으로 측정된다. 따라서 **같은 단위의 변동끼리 나눈 비율**에는 단위가 남지 않으며, 단위를 환산해도 결정계수는 유지된다.

따라서 단위에 의존하지 않는 측도는 ㄱ, ㄴ, ㄷ이며, 이는 ④의 조합과 일치한다.

NoteBase

문 3. 어느 지역의 남녀 간 소비생활 만족도의 분포가 같은지 알아보고자 이 지역에 거주하는 남녀 각각 100명을 임의로 추출하여 조사한 결과가 다음과 같다.

이를 검정하기 위한 카이제곱 검정통계량의 값이 2.1이고 유의확률 (p-값)이 0.35일 때, 이 검정에 대한 설명으로 옳지 않은 것은?

(단위 : 명)

성별 \ 만족도	불만족	보통	만족	합계
남	34	40	26	100
여	25	43	32	100

① 귀무가설은 '남녀 간 소비생활 만족도의 분포가 같다.'이다.

② 귀무가설이 참일 때, 카이제곱 검정통계량의 자유도는 3이다.

③ 유의수준 5%에서 귀무가설을 기각할 수 없다.

④ 이 검정을 동질성검정(homogeneity test)이라고 한다.

정답 2번

간단한 해설

표는 남·여 2개 집단과 불만족·보통·만족 3개 범주로 구성되며, 합계 열은 범주에 포함하지 않는다. 카이제곱 동질성검정의 자유도는 $(2-1)(3-1)=2$이므로 **3이 아니라 2**이다.

NoteBase

선지별 해설

① O: 귀무가설은 남녀의 소비생활 만족도 분포가 같다는 것이다. 즉 불만족·보통·만족 각각에 속하는 **모집단 비율이 남녀에서 모두 같다**고 가정한다. 대립가설은 적어도 한 범주의 비율이 남녀에서 다르다는 것이다.

② X: 카이제곱 동질성검정에서 집단 수가 r, 응답 범주 수가 c이면 자유도는 $(r-1)(c-1)$이다. 표의 행은 남·여 2개이고, 응답 범주는 불만족·보통·만족 3개이다. **합계 열은 별도의 응답 범주가 아니므로** 제외한다. 따라서 자유도는 $(2-1)(3-1)=2$이며, 선지의 3을 2로 고쳐야 한다.

③ O: 유의확률이 유의수준 이하이면 귀무가설을 기각한다. 주어진 유의확률 0.35는 유의수준 0.05보다 크므로 **귀무가설을 기각할 수 없다**. 남녀의 만족도 분포가 다르다고 판단할 통계적 근거가 부족하다는 뜻이며, 두 분포가 같다는 사실을 입증한 것은 아니다.

④ O: 동질성검정은 **서로 다른 모집단에서 한 범주형 변수의 분포가 같은지** 검정한다. 이 문제는 남녀 집단에서 각각 100명을 추출하여 소비생활 만족도라는 한 범주형 변수의 분포를 비교하므로 동질성검정에 해당한다. 한 모집단에서 뽑은 표본으로 두 범주형 변수 사이의 관련성을 조사하는 독립성검정과는 표본 추출 방식과 검정 목적이 다르다.

NoteBase

문 4. 다섯 가지 고기 포장 방법에 따라 박테리아 번식의 차이를 알아보고자 반복수가 같은 일원배치 분산분석법(one-way analysis of variance)을 적용하여 얻은 분산분석표의 일부는 다음과 같다. 이에 대한 설명으로 옳지 않은 것은?

요인	제곱합	자유도	평균제곱	F-값	p-값
처리			㉡	㉢	㉤
오차		㉠	㉣		
합계		14			

① ㉠의 값은 10이다.

② ㉡의 값은 ㉣의 값과 ㉢의 값의 곱이다.

③ 귀무가설은 '다섯 가지 고기 포장 방법에 따라 박테리아 번식에 차이가 없다.'이다.

④ ㉤의 값이 0.05보다 크면, 유의수준 5%에서 다섯 가지 고기 포장 방법에 따라 박테리아 번식에 차이가 있다고 할 수 있다.

정답 4번

간단한 해설

㉤은 p값이므로 0.05보다 크면 유의수준 5%에서 **귀무가설을 기각할 수 없다**. 따라서 포장 방법에 따라 박테리아 번식의 모평균에 차이가 있다고 판단할 근거가 부족하다.

NoteBase

선지별 해설

① O: 표의 합계 자유도는 전체 관측값 수에서 1을 뺀 값이다. 전체 관측값 수를 N이라 하면 $N-1=14$이므로 $N=15$이다. 포장 방법이 다섯 가지이므로 처리 자유도는 $5-1=4$이고, **오차 자유도** ㉠은 $14-4=10$이다.

② O: ㉡은 처리 평균제곱, ㉢은 오차 평균제곱, ㉣은 F값이다. 처리 평균제곱을 MS_T, 오차 평균제곱을 MS_E라 하면, F통계량은 **처리 평균제곱을 오차 평균제곱으로 나눈 값**으로 $F=\frac{MS_T}{MS_E}$이다. 이를 정리하면 $MS_T=F\cdot MS_E$이므로 ㉡은 ㉢과 ㉣의 곱이다.

③ O: 일원배치 분산분석은 여러 집단의 모평균이 같은지를 검정한다. 각 포장 방법의 박테리아 번식 모평균을 μ_1부터 μ_5까지로 나타내면, 귀무가설은 $\mu_1=\mu_2=\mu_3=\mu_4=\mu_5$이다. 따라서 포장 방법에 따른 차이가 없다는 것은 **다섯 모평균이 모두 같다**는 뜻이다. 대립가설은 이들 모평균이 모두 같지는 않다는 것이다.

④ X: 표의 ㉤은 p값이다. 이는 귀무가설이 참일 때 관측된 F값 이상으로 큰 값이 나올 확률이다. 유의수준 5%에서는 p값이 0.05보다 크면 **귀무가설을 기각하지 못한다**. 따라서 포장 방법에 따라 박테리아 번식의 모평균에 차이가 있다고 할 수 없다. 차이가 있다고 판단하는 것은 p값이 유의수준 0.05 이하일 때이다. 귀무가설을 기각하지 못했다고 해서 다섯 모평균이 같다는 사실이 입증된 것은 아니다.

NoteBase

문 5. 도시 A에 거주하는 사람 중에서 1%가 질병 D에 걸렸고 나머지 99%는 질병 D에 걸리지 않았다고 한다. 질병 D에 걸린 사람이 어떤 진단 시약에 대해 양성 반응을 보일 확률이 질병 D에 걸리지 않은 사람이 양성 반응을 보일 확률의 33배라고 한다. 도시 A에 거주하는 사람 중에서 임의로 뽑힌 사람이 이 진단 시약에 대해 양성 반응을 보였을 때, 이 사람이 질병 D에 걸렸을 확률은?

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

정답 1번

간단한 해설

질병 보유 비율과 각 집단의 양성 반응 확률을 함께 반영하면, 양성 반응자 중 질병에 걸린 사람과 걸리지 않은 사람의 비는 **1 : 3**이다. 따라서 양성 반응자가 질병에 걸렸을 확률은 $\frac{1}{4}$이다.

계산 과정

조건 정리: 질병 D에 걸린 사건을 D, 양성 반응을 보인 사건을 T라 하자. 질병에 걸리지 않은 사건은 D^c이며, $P(D) = 0.01$, $P(D^c) = 0.99$이다. 질병에 걸리지 않은 사람의 양성 반응 확률을 p라 두면, $P(T \mid D^c) = p$, $P(T \mid D) = 33p$이다.

질병 여부별 양성 반응 확률: 전체 주민 중 질병에 걸리면서 양성인 확률은 $P(D \cap T) = 0.01 \times 33p = 0.33p$이다. 질병에 걸리지 않으면서 양성인 확률은 $P(D^c \cap T) = 0.99p$이다. 각 집단 안에서 양성일 확률에 **그 집단이 전체 주민에서 차지하는 비율**을 곱한 것이다.

전체 양성 반응 확률: 양성 반응자는 질병에 걸린 사람과 걸리지 않은 사람으로 나뉘며, 두 경우는 겹치지 않는다. 따라서 $P(T) = 0.33p + 0.99p = 1.32p$이다.

양성일 때의 질병 확률: 이제 **양성 반응자 전체를 기준**으로 질병에 걸린 사람의 비율을 구한다. 조건부확률의 정의에 따라 $P(D \mid T) = \frac{P(D \cap T)}{P(T)}$이므로, $P(D \mid T) = \frac{0.33p}{1.32p} = \frac{1}{4}$이다. 양성 반응이 가능한 상황에서는 $p > 0$이므로 공통인 p가 약분되어, 그 값을 따로 알 필요가 없다.

구한 확률은 $\frac{1}{4}$, 즉 25%이며 ①의 값과 일치한다.

NoteBase

문 6. 모평균의 추정과 가설검정에 대한 설명으로 옳은 것만을 모두 고른 것은?

ㄱ. 신뢰도(confidence level)가 낮아질수록 신뢰구간의 길이가 짧아진다.
ㄴ. 검정력(power)은 귀무가설이 참일 때 귀무가설을 기각할 확률이다.
ㄷ. 유의수준 5%에서 기각된 귀무가설은 유의수준 10%에서도 기각된다.

① ㄱ, ㄴ
② ㄱ, ㄷ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 2번

간단한 해설

검정력은 **귀무가설이 거짓일 때 이를 기각할 확률**이다. 귀무가설이 참일 때 기각할 확률은 검정력이 아니라 제1종 오류의 확률이다.

보기별 해설

ㄱ. O: 표본크기와 표준오차 등 다른 조건이 같을 때, 신뢰도를 낮추면 신뢰구간 계산에 쓰는 임계값이 작아진다. 이에 따라 **오차한계가 줄어들어** 신뢰구간의 길이가 짧아진다.

ㄴ. X: 검정력은 **귀무가설이 거짓일 때 이를 기각할 확률**이다. 거짓인 귀무가설을 기각하지 못하는 제2종 오류의 확률을 β라고 하면 검정력은 1 − β이다. 보기의 '귀무가설이 참일 때 귀무가설을 기각할 확률'은 제1종 오류의 확률에 해당한다.

ㄷ. O: 같은 자료와 검정 방법에서 유의수준을 5%에서 10%로 높이면 **기각역이 넓어진다**. 따라서 5% 기각역에 속했던 검정통계량은 10% 기각역에도 속한다. p값으로 보아도 5% 이하인 p값은 10% 이하이므로 기각 결정이 유지된다.

옳은 보기는 ㄱ과 ㄷ이므로, 이 조합에 해당하는 선택지는 ②이다.

NoteBase

문 7. 자료 (X_i, Y_i), $(i = 1, 2, \ldots, 10)$에 단순선형회귀모형을 적용하여 얻은 분산분석표의 일부는 다음과 같다.
최소제곱법으로 추정된 회귀직선이 $\hat{Y}_i = 3 - 2.5X_i$일 때, X와 Y의 표본상관계수와 유의수준 5%에서 회귀직선이 유의하지 않다는 귀무가설에 대한 검정 결과가 옳게 짝지어진 것은?

요인	제곱합	자유도	평균제곱	$F-$값	$p-$값
회귀	15				0.002
잔차	6				
합계	21	9			

① 표본상관계수: $\sqrt{\frac{5}{7}}$, 검정 결과: 귀무가설 기각함
② 표본상관계수: $-\sqrt{\frac{5}{7}}$, 검정 결과: 귀무가설 기각함
③ 표본상관계수: $\sqrt{\frac{5}{7}}$, 검정 결과: 귀무가설 기각하지 않음
④ 표본상관계수: $-\sqrt{\frac{5}{7}}$, 검정 결과: 귀무가설 기각하지 않음

정답 2번

간단한 해설

회귀제곱합을 전체제곱합으로 나누면 $r^2 = \frac{5}{7}$이고, 회귀직선의 기울기가 **음수**이므로 $r = -\sqrt{\frac{5}{7}}$이다. 표의 p-값 0.002가 **유의수준 0.05보다 작으므로** 회귀직선이 유의하지 않다는 귀무가설을 기각한다.

NoteBase

계산 과정

결정계수: 표에서 회귀제곱합은 15, 잔차제곱합은 6, 전체제곱합은 21이다. 절편을 포함한 최소제곱 단순선형회귀에서는 **결정계수가 표본상관계수의 제곱**과 같다. 따라서 r^2=회귀제곱합전체제곱합이고, $r^2 = \frac{15}{21} = \frac{5}{7}$이다. 이는 Y의 전체 변동 중 회귀직선이 설명하는 비율이다.

표본상관계수의 부호: 기울기 추정값을 $\hat{\beta}_1$, X와 Y의 표본표준편차를 각각 s_X, s_Y라 하면 $\hat{\beta}_1 = r\frac{s_Y}{s_X}$이다. 표준편차의 비율은 양수이므로 **상관계수와 기울기의 부호는 같다**. 주어진 회귀직선의 기울기가 −2.5이므로 $r = -\sqrt{\frac{5}{7}}$이다.

회귀직선의 유의성 검정: 모집단 회귀계수를 β_1이라 할 때, 회귀직선이 유의하지 않다는 귀무가설은 $H_0 : \beta_1 = 0$이고 대립가설은 $H_1 : \beta_1 \neq 0$이다. 표의 회귀 행에 주어진 p-값은 0.002이다. **p-값이 유의수준보다 작으면 귀무가설을 기각**하므로, 0.002 < 0.05에 따라 귀무가설을 기각한다. 즉, 유의수준 5%에서 X와 Y 사이의 선형관계가 통계적으로 유의하다.

표본상관계수 $-\sqrt{\frac{5}{7}}$와 '귀무가설 기각함'을 짝지은 것은 ②이다.

NoteBase

문 8. 확률변수 X와 Y가 다음 조건을 만족할 때, 옳은 것은?

$E(X)=0, \quad Var(X)=4,$
$E(Y)=-2, \quad Var(Y)=8,$
$Cov(X,Y)=-5$

① $E(2X-3Y)=-6$

② $Var(X-Y)=17$

③ $Cov(X,X-Y)=-1$

④ $Corr(3X+3,2Y-4)=-\frac{5}{4\sqrt{2}}$

정답 4번

간단한 해설

두 확률변수에 각각 **양의 상수를 곱하거나 상수를 더해도 상관계수는 변하지 않는다**. 따라서 $Corr(3X+3,2Y-4)=Corr(X,Y)=-\frac{5}{4\sqrt{2}}$이다.

선지별 해설

① X: **기댓값은 선형성**을 가지므로 $E(2X-3Y)=2E(X)-3E(Y)$이다. 주어진 값을 대입하면 $2\cdot 0-3\cdot(-2)=6$이므로, 선지의 −6은 6으로 고쳐야 한다. 이 성질에는 두 확률변수의 독립성이 필요하지 않다.

② X: 두 확률변수의 차의 분산에는 **공분산 항도 반영**해야 한다. $Var(X-Y)=Var(X)+Var(Y)-2Cov(X,Y)$이므로 $4+8-2\cdot(-5)=22$이다. 공분산이 음수이므로 이를 두 배 하여 빼면 분산은 4 + 8보다 커진다. 선지의 17은 22로 고쳐야 한다.

③ X: 공분산은 각 변수에 대해 선형이며, **자기 자신과의 공분산은 분산**이다. 따라서 $Cov(X,X-Y)=Cov(X,X)-Cov(X,Y)$이고, $Var(X)-Cov(X,Y)=4-(-5)=9$이다. 선지의 −1은 9로 고쳐야 한다.

④ O: 상관계수는 공분산을 두 표준편차의 곱으로 나눈 값이다. 따라서 $Corr(X,Y)=\frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$이고, 주어진 값으로 계산하면 $\frac{-5}{\sqrt{4}\sqrt{8}}=-\frac{5}{4\sqrt{2}}$이다. 상수를 더하거나 빼는 것은 평균만 옮기므로 공분산과 분산을 바꾸지 않는다. 또한 X와 Y에 각각 3과 2를 곱하면 공분산은 6배, 두 표준편차의 곱도 6배가 되어 비율이 유지된다. 두 배율이 **모두 양수이므로 상관계수의 부호도 유지**되어 $Corr(3X+3,2Y-4)=-\frac{5}{4\sqrt{2}}$이다.

NoteBase

문 9. 표본의 크기가 n인 자료 (X_i, Y_i)에 단순선형회귀모형 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ (i = 1, 2, \ldots, n)$을 적용하고자 한다. 최소제곱법으로 추정된 회귀직선이 $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$일 때, 옳은 것만을 모두 고른 것은?

ㄱ. $X = x_0$에서 Y의 평균반응에 대한 신뢰구간의 길이는 x_0가 $\overline{X}$에 가까울수록 짧아진다.
ㄴ. 회귀제곱합(regression sum of squares)은 β_1^2 Σ _i=1^n(X_i-X)^2과 같다.
ㄷ. Y와 $\hat{Y}$의 표본상관계수의 제곱은 Y와 X의 표본상관계수의 제곱과 같다.

① ㄱ, ㄴ
② ㄱ, ㄷ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 4번

간단한 해설

절편을 포함한 단순선형회귀에서 예측값은 X에 기울기를 곱하고 상수를 더한 값이다. 상관계수가 정의되는 경우, 상수를 더해도 상관계수는 변하지 않고 음수를 곱하면 부호만 바뀌므로 **상관계수의 제곱은 같다**.

NoteBase

보기별 해설

절편을 포함한 최소제곱 회귀직선은 표본평균점 $(\overline{X}, \overline{Y})$을 지난다. 따라서 $\hat{Y}_i - \overline{Y} = \hat{\beta}_1(X_i - \overline{X})$이다. 설명변수의 편차제곱합을 $S_{XX} = \sum_{i=1}^{n}(X_i - \overline{X})^2$으로 두며, 회귀계수를 추정할 수 있도록 $S_{XX} > 0$인 경우를 다룬다.

ㄱ. O: 오차의 독립성·등분산성·정규성을 가정하는 통상적인 회귀 추론에서, 평균반응 추정값의 표준오차는 $s\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{X})^2}{S_{XX}}}$이다. 여기서 s는 잔차로 추정한 오차의 표준편차이다. 신뢰구간의 길이는 이 표준오차에 자유도 $n-2$인 t분포의 임계값과 2를 곱한 값이다. 같은 자료와 신뢰수준에서는 x_0가 $\overline{X}$에 가까울수록 거리의 제곱항이 작아지므로 길이가 짧아지며, **표본평균에서 가장 짧다**.

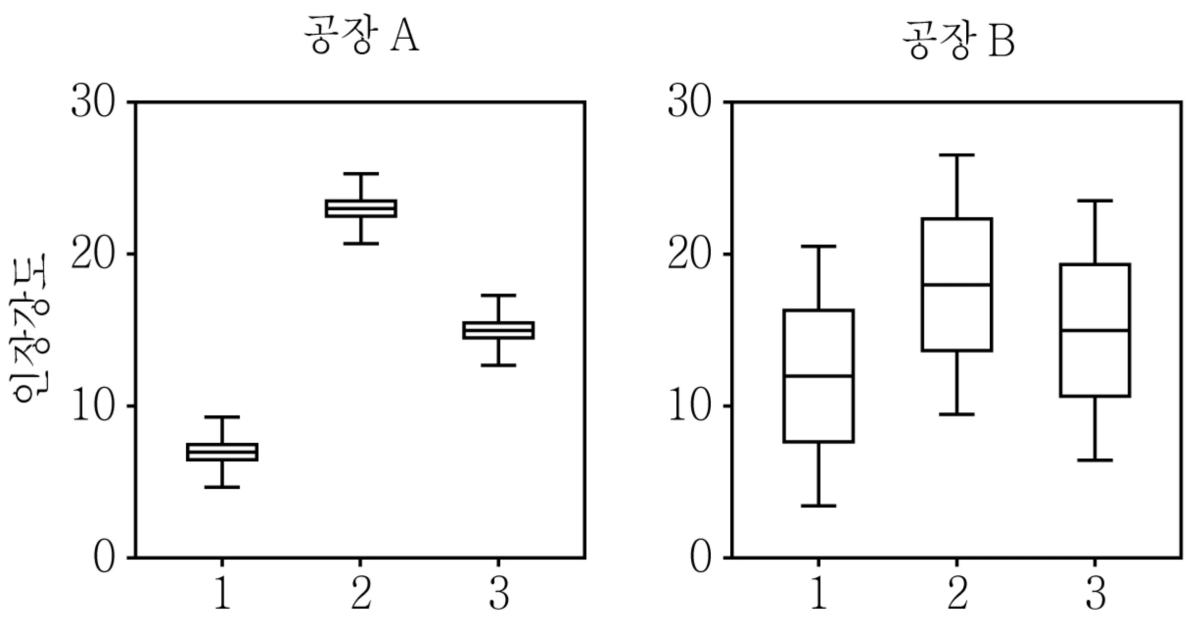
ㄴ. O: 회귀제곱합은 관측값 자체의 변동이 아니라 **예측값이 반응변수의 표본평균에서 벗어난 정도**를 합한 값으로, $\mathrm{SSR} = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$이다. 여기에 $\hat{Y}_i - \overline{Y} = \hat{\beta}_1(X_i - \overline{X})$을 대입하면 $\mathrm{SSR} = \hat{\beta}_1^2 S_{XX}$이 된다. 이는 보기의 식과 같다.

ㄷ. O: 예측값은 $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$이므로, X를 상수배하고 평행이동한 값이다. 표본상관계수는 상수를 더해도 변하지 않으며, 양수를 곱하면 그대로이고 음수를 곱하면 부호만 반대가 된다. 따라서 상관계수가 정의되는 경우 $r_{Y,\hat{Y}}$는 $r_{Y,X}$와 절댓값이 같아 $r^2_{Y,\hat{Y}} = r^2_{Y,X}$이다. **기울기가 음수여도 제곱의 동일성은 유지된다**. 다만 $\hat{\beta}_1 = 0$이면 예측값이 상수가 되어 Y와 예측값의 상관계수가 정의되지 않으므로, 이 진술은 상관계수가 정의되는 경우를 전제로 한다.

ㄱ, ㄴ, ㄷ이 모두 성립하므로 이를 모두 포함한 조합은 ④이다.

NoteBase

문 10. 공장 A와 B에서 각각 제조방법 1, 2, 3에 따라 생산되는 어떤 제품의 인장강도에 차이가 있는지 알아보고자 반복수가 같은 일원배치 분산분석법을 적용하려고 한다. 각 공장에서 측정한 제품의 인장강도에 대한 총제곱합은 같고, 제조방법별 인장강도의 상자그림(box plot)은 다음과 같다. 공장 A와 B의 평균처리제곱(mean square for treatment)을 각각 $MSTr_A$와 $MSTr_B$, 평균오차제곱(mean square for error)을 각각 MSE_A와 MSE_B라고 할 때, 이들 간의 대소 관계를 바르게 나타낸 것은?



① $MSTr_A < MSTr_B$, $MSE_A < MSE_B$

② $MSTr_A < MSTr_B$, $MSE_A > MSE_B$

③ $MSTr_A > MSTr_B$, $MSE_A < MSE_B$

④ $MSTr_A > MSTr_B$, $MSE_A > MSE_B$

정답 3번

간단한 해설

공장 A는 같은 제조방법 안에서 인장강도가 더 좁게 모여 있어 **오차변동이 작다**. 두 공장의 총제곱합이 같으므로 A의 처리변동은 더 크며, 반복수도 같아 $MSTr_A > MSTr_B$, $MSE_A < MSE_B$가 된다.

NoteBase

도식 해설

처리제곱합 $SSTr$은 제조방법별 평균 사이의 변동을, 오차제곱합 SSE는 같은 제조방법 안에서 관측값이 해당 방법의 평균 주위로 흩어진 변동을 나타낸다. 평균처리제곱과 평균오차제곱은 이 제곱합들을 각각의 자유도로 나눈 값이다.

상자그림의 가로축은 제조방법 1·2·3이고, 세로축은 인장강도이다. 두 그림의 세로축 눈금은 같으므로 산포를 직접 비교할 수 있다. A에서는 각 방법의 상자와 수염이 짧지만, B에서는 각 방법의 상자와 수염이 훨씬 길다. 따라서 **제조방법 내부의 산포**가 A에서 더 작으며, 동일 반복수에서 $SSE_A < SSE_B$로 판단한다. 상자 안의 가로선은 중앙값이므로 이를 평균값으로 놓고 처리제곱합을 계산하지 않는다.

일원배치 분산분석에서는 총제곱합 SST가 $SST = SSTr + SSE$로 분해된다. 문제에서 **두 공장의 총제곱합이 같다**고 하였으므로 오차제곱합이 작은 A는 처리제곱합이 더 크다. 즉, $SSTr_A > SSTr_B$이다. 전체 변동이 같더라도 그중 제조방법 사이의 차이가 차지하는 몫은 달라질 수 있다.

각 제조방법의 반복수를 n이라 하면 두 공장 모두 처리 수는 3, 전체 관측 수는 3n이다. 따라서 처리 자유도는 2, 오차 자유도는 3n − 3으로 같다. $MSTr = \frac{SSTr}{2}$, $MSE = \frac{SSE}{3n-3}$이므로 **같은 양의 자유도로 나누어도 대소 관계는 유지된다**. 따라서 $MSTr_A > MSTr_B$, $MSE_A < MSE_B$이며, 이 관계를 나타낸 선지는 ③이다.

NoteBase

문 11. 두 도시의 평균가구소득에 차이가 없다는 귀무가설을 검정하기 위하여 두 도시에서 각각 6 가구와 8 가구를 임의로 추출하여 조사하였다. 두 도시의 가구소득의 분산이 동일하다는 가정하에서 이 자료에 대한 t-검정통계량의 값이 −1.85, 유의확률(p-값)이 0.09일 때, t-검정통계량의 자유도와 유의수준 5%에서 검정 결과가 옳게 짝지어진 것은?

① 자유도: 12, 검정 결과: 귀무가설 기각함

② 자유도: 12, 검정 결과: 귀무가설 기각하지 않음

③ 자유도: 13, 검정 결과: 귀무가설 기각함

④ 자유도: 13, 검정 결과: 귀무가설 기각하지 않음

정답 2번

간단한 해설

두 모집단의 분산이 같다고 가정하는 두 표본 t-검정의 자유도는 $6 + 8 - 2 =$ **12**이다. 유의확률 0.09가 유의수준 0.05보다 크므로 **귀무가설을 기각하지 않는다**.

계산 과정

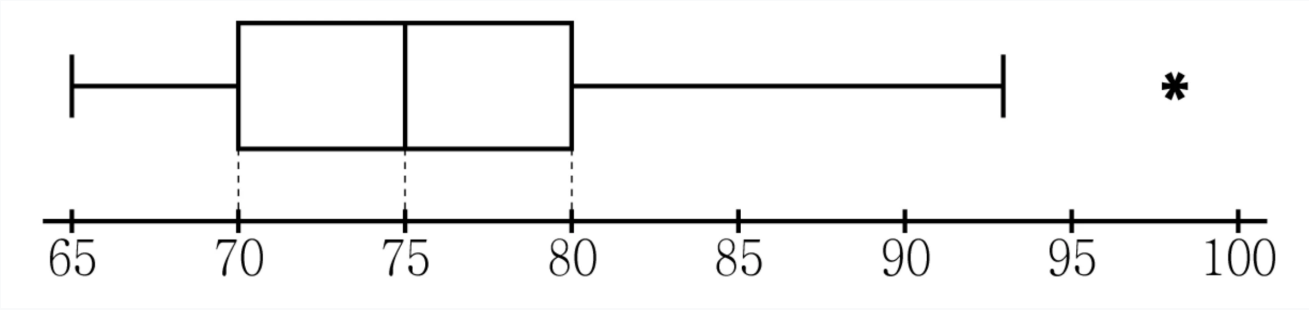
자유도: 두 도시의 표본크기를 각각 $n_1 = 6$, $n_2 = 8$이라 하자. **두 모집단의 분산이 같다는 가정**하에서는 두 표본의 분산을 합쳐 공통 분산을 추정하는 두 표본 t-검정을 사용한다. 각 표본에서 평균을 하나씩 추정하므로 분산 추정에 쓰이는 자유도는 각각 $n_1 - 1$, $n_2 - 1$이다. 이를 합하면 $n_1 + n_2 - 2$이므로, 이 검정통계량의 자유도는 $6 + 8 - 2 = 12$이다.

검정 결과: 귀무가설은 두 도시의 모집단 평균가구소득이 같다는 $H_0 : \mu_1 = \mu_2$이고, 대립가설은 $H_1 : \mu_1 \neq \mu_2$이다. 유의확률은 귀무가설 아래에서 관측된 결과만큼 또는 그보다 더 극단적인 결과가 나올 확률이다. **유의확률이 유의수준 이하일 때** 귀무가설을 기각한다. 여기서는 $0.09 > 0.05$이므로 기각하지 않는다. 검정통계량 −1.85로부터 구한 유의확률이 이미 주어졌으므로 임계값을 별도로 구할 필요는 없다. 이는 두 도시의 평균가구소득이 다르다고 판단할 증거가 유의수준 5%에서 충분하지 않다는 뜻이며, 두 평균이 같다고 입증한 것은 아니다.

자유도 12와 '귀무가설 기각하지 않음'을 짝지은 선택지는 ②이다.

NoteBase

문 12. 어느 대학교에서 통계학을 수강한 학생들의 중간고사 점수에 대한 상자그림이 다음과 같다. 이 자료에 대한 설명으로 옳지 않은 것은?



① 중앙값은 75이다.

② 제25백분위수(25th percentile)는 70이다.

③ 제3사분위수(Q_3)는 80이다.

④ 범위(range)는 30보다 작다.

정답 4번

간단한 해설

범위는 **이상값을 포함한 최댓값에서 최솟값을 뺀 값**이다. 그림에서 최솟값은 65점이고 별표로 표시된 최댓값은 95점보다 크므로, 범위는 30점보다 크다.

선지별 해설

① O: 상자 안의 세로선은 자료를 크기순으로 나열했을 때 가운데에 해당하는 **중앙값**을 나타낸다. 이 선이 가로축의 75에 있으므로 중앙값은 75점이다.

② O: 상자의 왼쪽 경계는 **제1사분위수**를 나타내며, 이는 자료의 하위 25%에 해당하는 경곗값인 제25백분위수이다. 왼쪽 경계가 가로축의 70에 있으므로 제25백분위수는 70점이다.

③ O: 상자의 오른쪽 경계는 자료의 하위 75%에 해당하는 경곗값인 **제3사분위수**를 나타낸다. 오른쪽 경계가 가로축의 80에 있으므로 제3사분위수는 80점이다.

④ X: 범위는 자료 전체의 최댓값에서 최솟값을 뺀 값으로, **별도로 표시된 이상값도 포함**한다. 그림에서 왼쪽 수염 끝의 최솟값은 65점이고, 오른쪽 별표로 표시된 최댓값은 95점보다 크다. 95 − 65 = 30이므로 실제 범위는 30점보다 크다. 따라서 '30보다 작다'는 '30보다 크다'로 고쳐야 한다. 오른쪽 수염 끝은 별표로 표시된 값보다 작으므로 자료 전체의 최댓값이 아니다.

NoteBase

문 13. 20대와 30대에서 100명을 임의로 추출하여 두 제품 A와 B에 대한 선호도를 조사한 결과가 다음과 같다. 연령대와 제품에 대한 선호도가 서로 독립이라는 귀무가설을 검정하기 위한 카이제곱 검정통계량의 값과 유의수준 5%에서 검정 결과가 옳게 짝지어진 것은? (단, $\chi^2_\alpha(k)$는 자유도가 k인 카이제곱분포의 제 $(1-\alpha)\times 100$ 백분위수를 나타내고, $\chi^2_{0.05}(1)=3.84,\ \chi^2_{0.05}(2)=5.99$이다)

(단위 : 명)

제품 / 연령대	A	B
20대	20	30
30대	30	20

① 검정통계량의 값: 4, 검정 결과: 귀무가설 기각함

② 검정통계량의 값: 4, 검정 결과: 귀무가설 기각하지 않음

③ 검정통계량의 값: 5, 검정 결과: 귀무가설 기각함

④ 검정통계량의 값: 5, 검정 결과: 귀무가설 기각하지 않음

정답 1번

간단한 해설

연령대와 제품 선호도가 독립이면 네 칸의 기대도수는 모두 25명이며, 관측도수와의 차이로 구한 **카이제곱 검정통계량은 4**이다. 자유도 1의 5% 임계값 3.84보다 크므로 **독립이라는 귀무가설을 기각**한다.

NoteBase

계산 과정

1. 자료와 가설: 표에서 20대는 A 20명·B 30명, 30대는 A 30명·B 20명이다. 각 연령대의 합계와 각 제품의 합계는 모두 50명이며, 전체는 100명이다. 귀무가설은 연령대와 제품 선호도가 서로 독립이라는 것이고, 대립가설은 서로 독립이 아니라는 것이다.

2. 기대도수: 독립이라면 어느 연령대에서도 제품별 선호 비율이 같아야 한다. 각 칸의 기대도수는 **해당 행 합계와 열 합계의 곱을 전체 도수로 나눈 값**이다. 따라서 네 칸 모두 $E = \frac{50 \cdot 50}{100} = 25$명이다. 이는 독립이라는 가설 아래 각 연령대에서 A와 B를 각각 25명씩 선호할 것으로 기대한다는 뜻이다. 모든 기대도수가 5 이상이므로 카이제곱 근사를 적용할 수 있다.

3. 검정통계량: 관측도수를 O, 기대도수를 E라 하면 피어슨 카이제곱 검정통계량은 $\chi^2 = \sum \frac{(O-E)^2}{E}$이다. 네 칸의 관측도수는 기대도수 25명보다 각각 5명 적거나 많으므로, 각 칸의 기여도는 $\frac{5^2}{25} = 1$이다. 따라서 $\chi^2 = 1 + 1 + 1 + 1 = 4$이다. 이 값은 독립일 때 기대되는 도수와 실제 도수의 차이를 합산한 것으로, 클수록 독립 가설에 불리한 증거가 된다.

4. 자유도와 검정 결과: 2행 2열 분할표의 자유도는 $(2-1) \times (2-1) = 1$이다. 카이제곱 독립성 검정에서는 **검정통계량이 오른쪽 꼬리의 임계값보다 크면** 귀무가설을 기각한다. 주어진 자유도 1의 5% 임계값은 3.84이며, $4 > 3.84$이므로 연령대와 제품 선호도가 독립이라는 귀무가설을 기각한다.

검정통계량 4와 귀무가설 기각을 짝지은 ①에 해당한다.

NoteBase

문 14. 확률분포에 대한 설명 중 옳지 않은 것은?

① Z가 표준정규분포를 따를 때, −Z도 표준정규분포를 따른다.

② X가 자유도가 10인 t분포를 따를 때, X^2은 분자의 자유도가 1, 분모의 자유도가 10인 F분포를 따른다.

③ X가 이항분포 B(10, p)를 따를 때, 10−X도 이항분포 B(10, p)를 따른다.

④ X_1과 X_2가 성공의 확률이 p인 베르누이분포를 따르고 서로 독립일 때, X_1+X_2는 이항분포 B(2, p)를 따른다.

정답 3번

간단한 해설

이항분포에서 X가 성공 횟수이면 10−X는 **실패 횟수**이다. 따라서 10−X는 성공확률이 p인 분포가 아니라 $B(10, 1-p)$를 따르며, $p = \frac{1}{2}$일 때만 원래 분포와 같아진다.

선지별 해설

① O: 표준정규분포는 평균이 0이고 분산이 1이며, **0을 중심으로 좌우 대칭**이다. Z의 부호를 바꾸면 분포를 좌우로 뒤집게 되지만 대칭성 때문에 분포가 변하지 않으므로 −Z도 표준정규분포를 따른다.

② O: 자유도가 10인 t분포는 $X = \frac{Z}{\sqrt{V/10}}$으로 나타낼 수 있다. 여기서 Z는 표준정규변수, V는 자유도가 10인 카이제곱변수이며 둘은 서로 독립이다. 제곱하면 $X^2 = \frac{Z^2}{V/10}$이고, Z^2은 자유도가 1인 카이제곱분포를 따른다. F분포는 **서로 독립인 두 카이제곱변수를 각각의 자유도로 나눈 값의 비**이므로, X^2은 분자의 자유도가 1, 분모의 자유도가 10인 F분포를 따른다.

③ X: $X \sim B(10, p)$는 성공확률이 p인 독립 시행 10회에서 성공 횟수가 X라는 뜻이다. 10−X는 같은 시행에서 실패한 횟수이고, 각 시행의 실패확률은 **1−p**이므로 $10 - X \sim B(10, 1-p)$이다. $p = \frac{1}{2}$인 특별한 경우에는 두 분포가 같지만, 일반적인 p에 대해 $B(10, p)$를 따른다고 할 수는 없다.

④ O: 베르누이변수는 성공하면 1, 실패하면 0의 값을 갖는다. X_1과 X_2가 **서로 독립이고 성공확률이 동일하게 p**이므로, 그 합은 성공확률이 p인 독립 시행 2회의 성공 횟수이다. 따라서 $X_1 + X_2 \sim B(2, p)$이다.

NoteBase

문 15. 두 확률변수 X_1과 X_2가 평균이 μ, 분산이 σ^2인 모집단에서 추출한 임의표본(random sample)일 때, 확률변수 $X_1(X_1+X_2)$의 기댓값은?

① $\mu^2+\sigma^2$

② $\mu^2+2\sigma^2$

③ $2\mu^2+\sigma^2$

④ $2\mu^2+2\sigma^2$

정답 3번

간단한 해설

임의표본이므로 X_1과 X_2는 **서로 독립**이다. 따라서 $E[X_1X_2]=\mu^2$이고, $E[X_1^2]=\sigma^2+\mu^2$이므로 구하는 기댓값은 $2\mu^2+\sigma^2$이다.

계산 과정

주어진 조건: 임의표본은 같은 모집단 분포를 따르며 서로 독립인 확률변수들이다. 따라서 $E[X_1]=E[X_2]=\mu$이고 $\mathrm{Var}(X_1)=\sigma^2$이다.

식의 전개: $X_1(X_1+X_2)=X_1^2+X_1X_2$이다. 기댓값의 선형성에 따라 두 항의 기댓값을 각각 구하여 더한다.

제곱항의 기댓값: 분산은 $\mathrm{Var}(X_1)=E[X_1^2]-(E[X_1])^2$이므로 $E[X_1^2]=\sigma^2+\mu^2$이다. **제곱의 기댓값**은 평균의 제곱에 분산을 더한 값이다.

곱항의 기댓값: X_1과 X_2가 **독립이므로** 곱의 기댓값을 각 기댓값의 곱으로 나눌 수 있다. 따라서 $E[X_1X_2]=E[X_1]E[X_2]=\mu^2$이다.

결과: $E[X_1(X_1+X_2)]=(\sigma^2+\mu^2)+\mu^2$이므로 구하는 값은 $2\mu^2+\sigma^2$이며, ③의 식과 일치한다.

NoteBase

문 16. 20개의 시점에서 관측한 시장포트폴리오의 수익률(X)과 A 회사 주식의 수익률(Y)에 대해 단순선형회귀모형을 적용하여 최소제곱법으로 추정된 회귀직선이 $\hat{Y} = 1 + 2X$이다. X와 Y의 표본분산이 각각 4와 32일 때, X와 Y의 표본상관계수는?

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

정답 1번

간단한 해설

절편을 포함한 최소제곱 단순선형회귀에서 **상관계수는 기울기에 두 표준편차의 비를 곱한 값**이다.
표본표준편차는 각각 2와 $4\sqrt{2}$이므로, 표본상관계수는 $\frac{1}{\sqrt{2}}$이다.

계산 과정

회귀기울기와 상관계수의 관계: 회귀직선 $\hat{Y} = 1 + 2X$의 기울기를 b라 하면 b = 2이다. 표본표준편차를 s_X, s_Y, 표본공분산을 s_{XY}라 하자. **절편을 포함한 최소제곱 단순선형회귀**에서는 $b = \frac{s_{XY}}{s_X^2}$이고, 표본상관계수의 정의는 $r = \frac{s_{XY}}{s_X s_Y}$이다. 따라서 $r = b\frac{s_X}{s_Y}$로 계산한다. 회귀기울기는 X와 Y의 단위에 영향을 받지만, 상관계수는 표준편차의 비로 이를 조정한 무단위 값이다.

표본표준편차: 주어진 4와 32는 **표준편차가 아니라 분산**이므로 제곱근을 취한다. 따라서 $s_X = \sqrt{4} = 2$, $s_Y = \sqrt{32} = 4\sqrt{2}$이다.

표본상관계수 계산: $r = 2 \cdot \frac{2}{4\sqrt{2}} = \frac{1}{\sqrt{2}}$이다. 기울기가 양수이므로 상관계수도 양수이며, 계산한 값은 ①의 $\frac{1}{\sqrt{2}}$과 일치한다.

NoteBase

문 17. 두 확률변수 X와 Y의 상관계수 $Corr(X, Y)$에 대한 설명으로 옳은 것은?

① $X = -0.5Y + 1$일 때, $Corr(X, Y)$는 -0.5이다.

② X와 Y가 서로 독립일 때, $Corr(X, Y)$가 0이 아닐 수 있다.

③ $Corr(X, Y)$와 $Corr(X, -Y)$의 합은 0이다.

④ $Corr(X, Y) = 0$일 때, X와 Y는 서로 독립이다.

정답 3번

간단한 해설

한 변수에 -1을 곱하면 **상관계수의 부호만 반대로 바뀐다**. 따라서 $\text{Corr}(X, -Y) = -\text{Corr}(X, Y)$이므로 두 상관계수의 합은 0이다.

선지별 해설

상관계수는 두 확률변수의 선형 관계의 방향과 강도를 나타내며, $\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$로 정의한다. 여기서 σ_X, σ_Y는 각 변수의 표준편차이다. 이하에서는 상관계수가 정의되도록 두 변수의 분산이 유한하고 양수인 경우를 전제한다.

① X: $X = -0.5Y + 1$은 X와 Y 사이에 **완전한 음의 선형 관계**가 있다는 뜻이므로 상관계수는 -1이다. Y가 증가할 때 X는 언제나 일정한 비율로 감소한다. 기울기 -0.5는 변화량의 비율이며, 표준편차로 표준화한 상관계수의 값과는 다르다. 상수 1을 더하는 것은 변수의 위치만 옮기므로 상관계수에 영향을 주지 않는다.

② X: 두 변수가 독립이면 $E[XY] = E[X]E[Y]$이므로 공분산은 0이다. 따라서 **독립이면 상관계수는 0**이며, 0이 아닐 수 있다는 설명은 성립하지 않는다.

③ O: Y에 -1을 곱하면 공분산은 $\text{Cov}(X, -Y) = -\text{Cov}(X, Y)$로 바뀌지만, -Y의 표준편차는 Y의 표준편차와 같다. 따라서 **크기는 같고 부호는 반대**인 $\text{Corr}(X, -Y) = -\text{Corr}(X, Y)$가 성립하여 두 상관계수의 합은 0이다.

④ X: 상관계수 0은 **선형 상관이 없다는 뜻이며 독립을 보장하지 않는다**. 예를 들어 X가 -1, 0, 1을 각각 1/3의 확률로 갖고 $Y = X^2$이라고 하자. 대칭성에 의해 $E[X] = E[X^3] = 0$이므로 공분산은 $E[X^3] - E[X]E[X^2] = 0$이고 상관계수도 0이다. 그러나 X가 0이면 Y는 반드시 0이고, X가 -1 또는 1이면 Y는 반드시 1이므로 두 변수는 독립이 아니다.

NoteBase

문 18. 확률변수 X가 정규분포 $N(6, 2^2)$을 따르고 확률변수 Y가 정규분포 $N(4, 2^2)$을 따르며 서로 독립일 때, $P(X < 6 < Y)$의 값은? (단, Z가 표준정규분포를 따르는 확률변수일 때, $P(Z \leq -1) = 0.16$이다)

① 0.08

② 0.17

③ 0.34

④ 0.42

정답 1번

간단한 해설

$X < 6$일 확률은 평균을 기준으로 한 대칭성에 의해 0.5이고, $Y > 6$일 확률은 표준화와 대칭성을 이용하면 0.16이다. 두 확률변수가 **서로 독립**이므로 두 사건이 동시에 일어날 확률은 $0.5 \times 0.16 = 0.08$이다.

계산 과정

사건의 분해: $X < 6 < Y$는 $X < 6$**과** $Y > 6$**이 동시에 성립**한다는 뜻이다. 두 확률변수가 독립이므로 이 두 사건도 독립이며, 각각의 확률을 곱하여 구한다.

$X < 6$의 확률: X의 평균은 6이고 표준편차는 2이다. 정규분포는 **평균을 중심으로 좌우 대칭**이므로 $P(X < 6) = 0.5$이다.

$Y > 6$의 확률: Y의 평균은 4이고 표준편차는 2이므로, 6은 평균보다 표준편차의 1배만큼 큰 값이다. 표준화하면 $P(Y > 6) = P\left(Z > \frac{6-4}{2}\right) = P(Z > 1)$이다. 표준정규분포의 대칭성과 연속성에 의해 $P(Z > 1) = P(Z \leq -1) = 0.16$이다. 이 0.16은 **표준정규분포의 오른쪽 꼬리 확률**로, Y가 6을 넘을 확률이다.

확률의 곱: 독립성을 적용하면 $P(X < 6 < Y) = P(X < 6) \cdot P(Y > 6)$이다. 따라서 $0.5 \times 0.16 = 0.08$이며, 이는 ①의 값과 일치한다.

NoteBase

문 19. 위치 A에서 0 또는 1의 메시지 x를 송신하면 위치 B에서는 x에 표준정규분포를 따르는 확률변수 ε을 더한 값 $R = x + \varepsilon$으로 수신한다. 위치 B에서는 수신된 메시지 R가 0.3 이하이면 0으로 판독하고, 0.3보다 크면 1로 판독한다. 위치 A에서 1을 송신할 때, 위치 B에서 수신한 메시지를 0으로 판독할 확률은? (단, Z가 표준정규분포를 따르는 확률변수일 때, $P(Z \leq 0.7) = 0.76$이다)

① 0.24

② 0.26

③ 0.48

④ 0.76

정답 1번

간단한 해설

1을 송신했는데 0으로 판독되려면 $1 + \varepsilon \leq 0.3$, 즉 $\varepsilon \leq -0.7$이어야 한다. **표준정규분포의 대칭성**에 따라 이 확률은 $1 - 0.76 = 0.24$이다.

계산 과정

송신값 반영: 1을 송신하므로 $x = 1$을 수신식에 대입하면 $R = 1 + \varepsilon$이다. 여기서 ε은 평균 0, 분산 1인 표준정규분포를 따르는 잡음이다. 따라서 수신값은 송신값 1을 중심으로 변동한다.

0으로 판독되는 조건: 수신값이 **0.3 이하**일 때 0으로 판독하므로 구할 사건은 $1 + \varepsilon \leq 0.3$이다. 양변에서 1을 빼면 $\varepsilon \leq -0.7$이다. 즉, 잡음이 수신값을 1에서 0.3 이하로 낮추는 경우의 확률을 구한다.

확률 계산: 표준정규분포는 0을 중심으로 대칭이므로 $P(\varepsilon \leq -0.7) = P(Z \geq 0.7)$이다. 또한 연속분포에서는 한 점의 확률이 0이므로 $P(Z \geq 0.7) = 1 - P(Z \leq 0.7)$이다. 주어진 누적확률 0.76을 적용하면 $1 - 0.76 = 0.24$이다. 따라서 1을 송신했을 때 0으로 판독할 확률은 0.24로, ①의 값과 일치한다.

NoteBase

문 20. 확률변수 X가 이항분포 $B(5, p)$를 따를 때, p에 대한 다음과 같은 가설을 검정하려고 한다.
$H_0 : p = \frac{1}{2}$ **대** $H_1 : p \neq \frac{1}{2}$
기각역이 '$|X - 2.5| > 2$**'일 때,** $p = \frac{2}{3}$**에서 제2종의 오류를 범할 확률은?**

① $\frac{11}{81}$
② $\frac{22}{81}$
③ $\frac{59}{81}$
④ $\frac{70}{81}$

정답 4번

간단한 해설

제2종의 오류는 **귀무가설이 거짓인데도 기각하지 않는** 오류이다. 실제 모수가 $p = \frac{2}{3}$일 때 기각역인 0과 5를 제외한 1≤ X≤4의 확률을 구하면 $\frac{70}{81}$이다.

계산 과정

기각역 확인: X는 0부터 5까지의 정수 값을 갖는다. $|X - 2.5| > 2$를 풀면 $X < 0.5$ 또는 $X > 4.5$이므로 기각역은 **0과 5**이다. 따라서 귀무가설을 기각하지 않는 경우는 $X = 1, 2, 3, 4$이다.

오류 확률의 기준: 제2종의 오류는 귀무가설이 거짓일 때 이를 기각하지 않는 오류이다. 여기서는 실제 모수가 $p = \frac{2}{3}$이므로 $H_0 : p = \frac{1}{2}$는 거짓이다. 따라서 **실제 모수가 2/3인 이항분포**에서 1≤ X≤4의 확률을 계산한다.

기각확률 계산: 이항분포에서 $X = 0$은 5번 모두 실패하는 경우이고, $X = 5$는 5번 모두 성공하는 경우이다. 각 시행이 독립이고 실패확률은 $1 - p = \frac{1}{3}$이므로 다음과 같다.
$P(X = 0) = \left(\frac{1}{3}\right)^5 = \frac{1}{243}$
$P(X = 5) = \left(\frac{2}{3}\right)^5 = \frac{32}{243}$
두 경우는 동시에 일어날 수 없으므로 기각확률은 $\frac{1 + 32}{243} = \frac{11}{81}$이다.

제2종의 오류 확률: 기각하지 않을 확률은 기각할 확률의 여사건 확률이므로 $\beta = 1 - \frac{11}{81} = \frac{70}{81}$이다. 이 값은 선택지 ④의 $\frac{70}{81}$과 일치한다.

NoteBase