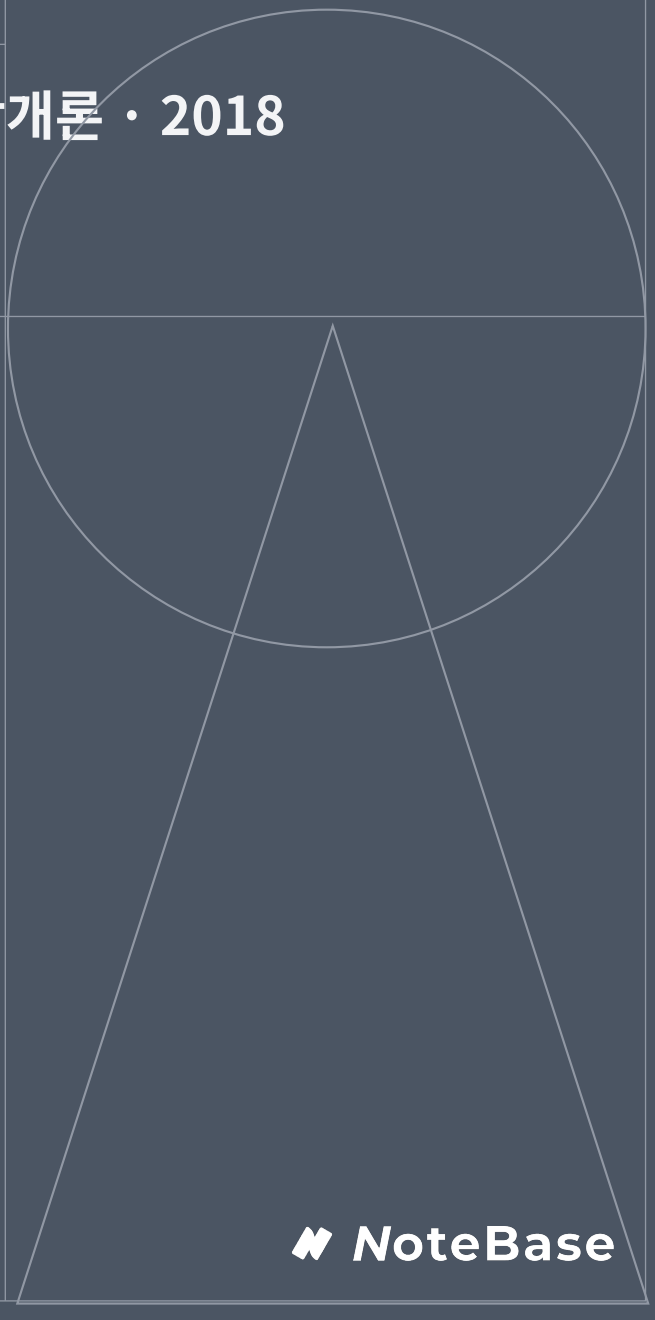
기출OX 해설집

공무원/국가직/9급 · 통계학개론 · 2018



NoteBase

기출OX 해설집

공무원/국가직/9급 · 통계학개론 · 2018

공무원/국가직/9급 · 통계학개론 · 2018

문 1. 상자그림(box plot)으로부터 알 수 없는 통계량은?

① 중앙값(median)

② 25 % 절사평균(trimmed mean)

③ 사분위수 범위(interquartile range)

④ 범위(range)

정답 2번

간단한 해설

25% 절사평균은 자료의 양쪽 끝에서 각각 25%를 제외한 뒤 **남은 자료의 평균**을 구한 값이다. 상자그림에는 가운데 자료 각각의 값이나 합이 나타나지 않으므로 이를 알 수 없다.

공통 해설

상자그림에서 알 수 있는지는 **그림에 표시된 요약값만으로 구할 수 있는가**를 기준으로 판단한다. 상자 안의 선은 중앙값을, 상자의 양 끝은 제1사분위수와 제3사분위수를 나타낸다. 따라서 중앙값을 읽을 수 있고, 제3사분위수에서 제1사분위수를 빼면 사분위수 범위를 구할 수 있다.

범위는 최댓값에서 최솟값을 뺀 값이다. 수염이 최솟값과 최댓값까지 이어지는 상자그림에서는 수염의 양 끝으로 구한다. 이상치를 수염 밖에 별도 점으로 표시하는 방식에서는 **이상치까지 포함한 전체 자료의 양 끝**을 읽어야 한다.

반면 25% 절사평균은 자료를 크기순으로 정렬하여 아래쪽 25%와 위쪽 25%를 제외하고, 가운데 남은 자료들의 합을 그 개수로 나눈 값이다. 상자그림은 가운데 자료가 놓이는 구간을 요약하지만 **그 안의 개별 값과 합은 보여 주지 않는다**. 따라서 중앙값·사분위수 범위·범위와 달리 25% 절사평균은 상자그림만으로 결정할 수 없다.

NoteBase

문 2. 휴대전화를 제조하는 세 회사 A, B, C의 시장 점유율은 각각 50 %, 30 %, 20 %로 알려져 있다. 신제품의 출시가 시장 점유율에 영향을 미치는지 알아보기 위하여 C회사가 신제품을 출시하고 6개월 후, 200명의 휴대전화 사용자를 임의로 추출하여 다음과 같이 조사하였다. 시장 점유율의 변화가 있는지 알아보기 위하여 사용하는 검정법으로 옳은 것은?

회사명	A	B	C
관측도수	98	48	54

① 자유도가 2인 카이제곱분포를 이용한 적합도 검정

② 자유도가 2인 카이제곱분포를 이용한 독립성 검정

③ 자유도가 3인 카이제곱분포를 이용한 적합도 검정

④ 자유도가 3인 카이제곱분포를 이용한 독립성 검정

정답 1번

간단한 해설

사용 회사라는 하나의 범주형 변수의 분포가 기존 비율 50%·30%·20%에 부합하는지 확인하므로 **카이제곱 적합도 검정**을 사용한다. 세 범주의 비율이 미리 주어져 있으므로 자유도는 **3 − 1 = 2**이다.

핵심 풀이

검정법: 적합도 검정은 한 범주형 변수의 관측도수가 주어진 확률분포에 부합하는지 확인하는 검정이다.
여기서는 신제품 출시 6개월 후의 회사별 시장 점유율이 기존의 50%·30%·20%와 같은지를 검정한다. 표의 관측도수는 A사 98명, B사 48명, C사 54명이고, 기존 점유율이 유지된다는 가정 아래 기대도수는 200 × 0.5 = 100명, 200 × 0.3 = 60명, 200 × 0.2 = 40명이다. 관측도수와 기대도수의 차이를 카이제곱 통계량으로 평가하며, 기대도수도 모두 충분히 크다. 독립성 검정은 **두 범주형 변수 사이의 관련성**을 검정할 때 사용한다. 이 자료는 사용 회사라는 한 변수만 조사했으므로 독립성 검정에 해당하지 않는다.

자유도: 범주는 A·B·C의 세 개이지만, 도수의 합이 200명으로 고정되어 있어 두 범주의 도수가 정해지면 나머지 하나는 자동으로 정해진다. 또한 기준 비율은 표본에서 추정한 값이 아니라 문제에서 미리 주어진 값이므로, 모수 추정에 따른 자유도 차감은 없다. 따라서 자유도는 **범주 수에서 1을 뺀** 3 − 1 = 2이며, 자유도가 2인 카이제곱분포를 이용한 적합도 검정을 적용한다.

NoteBase

문 3. 선형회귀모형에 대한 가정으로 옳지 않은 것은?

① 오차의 등분산성

② 오차 간의 독립성

③ 독립변수와 종속변수의 선형성

④ 독립변수 간 상관관계 존재성

정답 4번

간단한 해설

독립변수 사이에 **상관관계가 존재해야 한다는 가정은 없다**. 서로 상관될 수는 있지만, 한 독립변수가 다른 독립변수들의 정확한 일차결합으로 표현되는 **완전 다중공선성은 없어야 한다**.

선지별 해설

① O: 오차의 등분산성이란 **독립변수의 값에 관계없이 오차의 분산이 일정하다**는 뜻이다. 독립변수 값이 커질수록 오차의 흩어짐이 커지거나 작아지는 이분산성과 구별된다.

② O: 오차 간의 독립성은 **서로 다른 관측치의 오차가 통계적으로 독립이다**라는 가정이다. 예를 들어 시계열 자료에서 앞 시점의 오차와 뒤 시점의 오차가 연관되면 이 가정에 문제가 생긴다.

③ O: 독립변수의 값이 주어졌을 때 **종속변수의 평균이 독립변수의 선형함수로 표현된다**는 가정이다. 개별 관측값에는 오차가 더해지므로 모든 관측점이 정확히 직선이나 평면 위에 있어야 한다는 뜻은 아니다.

④ X: 독립변수 간 상관관계의 존재는 선형회귀모형의 필수 가정이 아니다. 여러 독립변수는 서로 상관될 수 있지만, 한 변수가 다른 변수들의 정확한 일차결합으로 표현되는 **완전 다중공선성은 없어야 한다**. 완전 다중공선성이 있으면 각 독립변수의 효과를 분리하여 회귀계수를 유일하게 추정할 수 없다.

NoteBase

문 4. 두 변수의 상관계수(correlation coefficient)에 대한 설명으로 옳은 것은?

① 항상 0과 1 사이의 값이다.

② 0에 가까울수록 두 변수는 서로 독립이다.

③ 공분산과 부호가 같다.

④ 두 변수의 인과관계를 설명하는 값이다.

정답 3번

간단한 해설

상관계수는 공분산을 두 변수의 표준편차의 곱으로 나눈 값이다. 표준편차가 모두 양수이면 분모가 양수이므로 **공분산과 부호가 같다**.

선지별 해설

상관계수는 두 변수의 선형적 관계의 방향과 강도를 나타낸다. 분산이 유한하고 0보다 큰 두 변수 X, Y에 대해 $\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$로 정의하며, 여기서 σ_X, σ_Y는 각 변수의 표준편차이다.

① X: 상관계수의 범위는 0부터 1까지가 아니라 **−1부터 1까지**이다. 음수이면 한 변수가 증가할 때 다른 변수는 감소하는 방향의 선형적 관계를, 양수이면 함께 증가하는 방향의 선형적 관계를 나타낸다.

② X: 상관계수가 0에 가깝다는 것은 **선형적 관계가 약하다**는 뜻이며, 독립이라는 뜻은 아니다. 상관계수가 0이어도 비선형 관계가 있을 수 있다. 두 변수가 독립이고 분산이 유한하며 양수이면 상관계수는 0이지만, 그 역은 일반적으로 성립하지 않는다.

③ O: 상관계수의 분모는 양수인 두 표준편차의 곱이므로, 나눗셈을 해도 **공분산의 부호가 유지된다**. 공분산이 양수이면 상관계수도 양수이고, 음수이면 음수이며, 공분산이 0이면 상관계수도 0이다.

④ X: 상관계수는 두 변수의 **선형적 연관성**을 나타내며, 그 자체로 인과관계를 설명하지는 못한다. 두 변수에 공통으로 영향을 주는 다른 요인 때문에 상관이 나타날 수도 있으므로, 상관계수가 크다는 사실만으로 한 변수가 다른 변수의 원인이라고 결론 내릴 수 없다.

NoteBase

문 5. 서로 독립인 두 확률변수 X와 Y의 결합확률분포표가 다음과 같을 때, $P(X=2, Y=1)$의 값은?

X \ Y	1	3
2	a	b
4	0.2	0.4

① $\frac{1}{15}$
② $\frac{2}{15}$
③ $\frac{1}{5}$
④ $\frac{4}{15}$

정답 2번

간단한 해설

두 확률변수가 **독립**이므로 결합확률은 각 주변확률의 곱이다. 표에서 구한 $P(X=2)=\frac{2}{5}$와 $P(Y=1)=\frac{1}{3}$을 곱하면 구하는 확률은 $\frac{2}{15}$이다.

계산 과정

1. 행의 합으로 X의 주변확률을 구한다. 표의 행은 X의 값, 열은 Y의 값을 나타낸다. X=4 행의 두 확률을 더하면 $P(X=4)=0.2+0.4=0.6$이다. 전체 확률의 합은 1이므로 $P(X=2)=1-0.6=0.4$이다.

2. 독립 조건으로 Y의 주변확률을 구한다. **독립이면 결합확률은 주변확률의 곱**이므로 $P(X=4, Y=1)=P(X=4)P(Y=1)$이다. 표에서 이 결합확률이 0.2이므로 $P(Y=1)=\frac{0.2}{0.6}=\frac{1}{3}$이다. 이는 X=4 행의 합 중 Y=1인 칸이 차지하는 비율이며, 독립이므로 X의 값에 따라 달라지지 않는다.

3. 구하는 칸의 확률 a를 계산한다. X=2 행과 Y=1 열이 만나는 칸이므로 $a=P(X=2)P(Y=1)$이다. 따라서 $a=\frac{2}{5}\cdot\frac{1}{3}=\frac{2}{15}$이며, 이는 ②의 값과 일치한다.

NoteBase

문 6. 가설 $H_0 : \mu = 4$ 대 $H_1 : \mu = 6$ 에 대한 검정통계량 T의 분포는 $N(\mu, 0.5^2)$을 따르고 검정의 기각역이 $\{T > 5\}$이다. 이 검정법의 제1종의 오류와 제2종의 오류를 범할 확률을 옳게 짝지은 것은? (단, 표준정규분포를 따르는 확률변수 Z에 대하여, $P(Z > 1) = 0.159, P(Z > 2) = 0.023$이다)

① 제1종의 오류를 범할 확률: 0.023, 제2종의 오류를 범할 확률: 0.023

② 제1종의 오류를 범할 확률: 0.159, 제2종의 오류를 범할 확률: 0.023

③ 제1종의 오류를 범할 확률: 0.023, 제2종의 오류를 범할 확률: 0.159

④ 제1종의 오류를 범할 확률: 0.159, 제2종의 오류를 범할 확률: 0.159

정답 1번

간단한 해설

제1종 오류는 **평균이 4일 때 기각할 확률**, 제2종 오류는 **평균이 6일 때 기각하지 않을 확률**이다. 경계값 5는 각 평균에서 표준편차 0.5의 두 배만큼 떨어져 있으므로, 두 오류확률은 각각 0.023이다.

계산 과정

분포와 기각 기준: 검정통계량 T의 평균은 μ, 표준편차는 0.5이므로 $Z = \frac{T-\mu}{0.5}$로 표준화한다. $T > 5$이면 귀무가설을 기각하고, $T \le 5$이면 기각하지 않는다.

제1종 오류확률: **귀무가설이 참인데 이를 기각**하는 확률이다. 따라서 평균이 4인 분포에서 $T > 5$인 오른쪽 꼬리확률을 구한다. 경계값을 표준화하면 $\frac{5-4}{0.5} = 2$이므로, $\alpha = P(T > 5 \mid \mu = 4) = P(Z > 2) = 0.023$이다.

제2종 오류확률: **대립가설이 참인데 귀무가설을 기각하지 못하는** 확률이다. 따라서 평균이 6인 분포에서 $T \le 5$인 왼쪽 꼬리확률을 구한다. 경계값을 표준화하면 $\frac{5-6}{0.5} = -2$이므로, $\beta = P(T \le 5 \mid \mu = 6) = P(Z \le -2)$이다. 표준정규분포는 0을 중심으로 대칭이고 연속분포이므로 $P(Z \le -2) = P(Z > 2) = 0.023$이다.

제1종 오류확률과 제2종 오류확률은 모두 0.023이며, 이 짝은 ①에 해당한다.

NoteBase

문 7. 다음은 여행 프로그램 A, B, C에 대한 선호도가 연령대와 상관성이 있는지 알아보기 위하여 설문조사를 한 결과이다. 프로그램에 대한 선호도가 연령대와 상관성이 없다는 귀무가설을 검정하기 위한 카이제곱 검정통계량이 14.15일 때, 유의수준 5%에서 검정 결과로 옳은 것은? (단, $\chi^2_\alpha(n)$은 자유도가 n인 카이제곱분포의 제 $100 \times (1-\alpha)$ 백분위수이고 $\chi^2_{0.05}(6) = 12.59$, $\chi^2_{0.05}(8) = 15.51$이다)

	A	B	C
20대	6	27	19
30대	8	36	17
40대	21	45	33
50대	14	18	6

① 검정통계량의 값이 12.59보다 크므로 프로그램의 선호도가 연령대와 상관성이 없다고 할 수 있다.

② 검정통계량의 값이 12.59보다 크므로 프로그램의 선호도가 연령대와 상관성이 없다고 할 수 없다.

③ 검정통계량의 값이 15.51보다 작으므로 프로그램의 선호도가 연령대와 상관성이 없다고 할 수 있다.

④ 검정통계량의 값이 15.51보다 작으므로 프로그램의 선호도가 연령대와 상관성이 없다고 할 수 없다.

정답 2번

간단한 해설

표에는 연령대가 4개, 프로그램이 3개이므로 독립성 검정의 **자유도는 6**이다. 검정통계량 14.15가 임계값 12.59보다 커서 **상관성이 없다는 귀무가설을 기각**한다.

NoteBase

계산 과정

검정할 가설: 귀무가설은 '프로그램 선호도와 연령대가 독립이다', 대립가설은 '두 변수가 독립이 아니다'이다. 독립이라는 것은 연령대에 따라 프로그램별 선호 비율이 달라지지 않는다는 뜻이다.

자유도: 표의 행은 20대·30대·40대·50대의 4개 연령대이고, 열은 A·B·C의 3개 프로그램이다. 행이 r개, 열이 c개인 분할표의 독립성 검정에서 자유도는 $(r-1)(c-1)$이다. 행합과 열합이 정해지면 일부 칸의 빈도가 나머지 칸의 빈도로 결정되기 때문이다. 따라서 자유도는 $(4-1)\times(3-1)=6$이다.

임계값과 검정통계량 비교: 카이제곱 독립성 검정은 관측빈도와 독립을 가정한 기대빈도의 차이가 클수록 검정통계량이 커지므로 **오른쪽 꼬리에서 귀무가설을 기각**한다. 자유도 6, 유의수준 5%의 임계값은 문제에서 주어진 12.59이다. 주어진 검정통계량은 14.15이므로 $14.15 > 12.59$로 기각역에 들어간다. 15.51은 **자유도 8일 때의 임계값**이므로 이 표의 판단 기준으로 사용할 수 없다.

검정 결과: 귀무가설을 기각하므로 유의수준 5%에서 프로그램 선호도와 연령대가 독립이라고 할 수 없다. 즉, 두 변수 사이에 관련성이 있다는 통계적 근거가 있으며, '12.59보다 크므로 상관성이 없다고 할 수 없다'는 ②의 설명에 해당한다.

NoteBase

문 8. 어느 지역에 사는 주민의 월 소득은 정규분포 $N(200, 20^2)$을 따른다. 이 지역에서 두 명의 주민을 임의로 추출하여 월 소득을 조사할 때 옳지 않은 것은? (단, 표준정규분포를 따르는 확률변수 Z에 대하여 $P(Z > 1.645) = 0.05$이다)

① 두 명 모두 월 소득이 200보다 작을 확률은 0.25이다.

② 두 명 중 최소한 한 명의 월 소득이 구간 (167.1, 232.9)에 속할 확률은 0.99이다.

③ 한 명의 월 소득이 167.1보다 작고 다른 한 명의 월 소득이 232.9보다 클 확률은 0.05이다.

④ 두 명의 월 소득 평균이 구간 (167.1, 232.9)에 속할 확률은 0.9보다 크다.

정답 3번

간단한 해설

한 사람이 하위 5% 구간에, 다른 사람이 상위 5% 구간에 속해야 하므로 두 확률을 곱한다. 어느 사람이 낮은 소득을 갖는지에 따라 **서로 겹치지 않는 두 경우**가 있어, 그 확률은 $2 \times 0.05 \times 0.05 = 0.005$이다.

선지별 해설

임의추출한 두 사람의 월 소득을 서로 독립인 X_1, X_2로 둔다. 각 소득은 평균 200, 표준편차 20인 정규분포를 따른다. 한 사람의 소득을 X라 하면 표준화한 값은 $Z = \frac{X-200}{20}$이다. 경계값 167.1과 232.9는 각각 표준화하면 -1.645와 1.645이다. 정규분포의 대칭성과 주어진 조건에 따라 양쪽 꼬리확률은 각각 0.05이고, 가운데 구간에 속할 확률은 $1 - 0.05 - 0.05 = 0.9$이다.

① O: 정규분포는 평균을 중심으로 대칭이므로 한 사람의 소득이 200보다 작을 확률은 0.5이다. 두 사람의 소득은 **서로 독립**이므로 두 명 모두 200보다 작을 확률은 $0.5 \times 0.5 = 0.25$이다.

② O: 한 사람의 소득이 구간 (167.1, 232.9) 밖에 있을 확률은 0.1이다. **최소한 한 명이 구간 안에 있는 사건의 여사건**은 두 명 모두 구간 밖에 있는 사건이다. 두 명 모두 구간 밖에 있을 확률은 $0.1 \times 0.1 = 0.01$이므로, 구하는 확률은 $1 - 0.01 = 0.99$이다.

③ X: 첫 번째 사람이 167.1보다 작고 두 번째 사람이 232.9보다 클 확률은 $0.05 \times 0.05 = 0.0025$이다. 두 사람의 역할이 바뀐 경우도 같은 확률을 가지며, **두 경우는 동시에 일어날 수 없다**. 따라서 한 명은 167.1보다 작고 다른 한 명은 232.9보다 클 확률은 $2 \times 0.0025 = 0.005$이다. 선지의 0.05는 0.005로 고쳐야 한다.

④ O: 두 사람의 소득 평균을 $\bar{X} = \frac{X_1 + X_2}{2}$라 하면, 독립인 정규확률변수의 평균이므로 $\bar{X} \sim N\left(200, \frac{20^2}{2}\right)$이다. 평균은 그대로 200이지만 **표준편차는 원래 소득의 표준편차보다 작다**. 구체적으로 표준편차는 $\frac{20}{\sqrt{2}}$이다. 표준화한 변수 $W = \frac{\bar{X} - 200}{\frac{20}{\sqrt{2}}}$는 표준정규분포를 따르므로 $P(167.1 < \bar{X} < 232.9) = P(|W| < 1.645\sqrt{2})$이다. 이 구간은 $-1.645 < W < 1.645$보다 넓으므로 그 확률은 0.9보다 크다.

NoteBase

문 9. 다음은 볍씨의 네 가지 종류에 따라 벼의 평균 수확량에 차이가 있는지 알아보기 위한 분산분석표의 일부이다. 유의수준 5%에서 이 가설을 검정할 때, ㉠, ㉡의 값과 검정 결과를 옳게 짝지은 것은?

요인	제곱합	자유도	$F-$값	$p-$값
볍씨의 종류	6	㉠	㉡	0.0519
오차	4	8		
계	10	11		

① ㉠: 3, ㉡: 4, 검정 결과: 유의하지 않다

② ㉠: 3, ㉡: 4, 검정 결과: 유의하다

③ ㉠: 4, ㉡: 5, 검정 결과: 유의하지 않다

④ ㉠: 4, ㉡: 5, 검정 결과: 유의하다

정답 1번

간단한 해설

볍씨가 네 종류이므로 요인의 자유도는 **4 − 1 = 3**이고, 요인평균제곱을 오차평균제곱으로 나눈 F값은 4이다.
표의 p값 0.0519가 유의수준 0.05보다 크므로 평균 수확량의 차이는 **통계적으로 유의하지 않다**.

NoteBase

계산 과정

검정 가설: 네 종류 볍씨의 모평균 수확량을 각각 $\mu_1, \mu_2, \mu_3, \mu_4$라 하면, 귀무가설은 $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$이다. 대립가설은 네 모평균이 모두 같지는 않다는 것으로, 적어도 한 종류의 모평균이 다르다는 뜻이다.

㉠ 자유도: 일원분산분석에서 요인의 자유도는 **집단 수에서 1을 뺀 값**이다. 볍씨가 네 종류이므로 ㉠은 4 − 1 = 3이다. 표에서도 전체 자유도 11에서 오차 자유도 8을 빼면 3이 된다.

㉡ F값: 평균제곱은 제곱합을 해당 자유도로 나눈 값이다. 표의 요인 제곱합은 6이므로 요인평균제곱은 $\frac{6}{3} = 2$이고, 오차 제곱합은 4이므로 오차평균제곱은 $\frac{4}{8} = 0.5$이다. F통계량은 집단 간 변동이 집단 내 오차변동에 비해 얼마나 큰지를 보는 값으로, **요인평균제곱을 오차평균제곱으로 나누어** 구한다. 따라서 $F = \frac{2}{0.5} = 4$이므로 ㉡은 4이다. F값은 단위가 없는 비율이다.

검정 결과: 유의수준 5%는 0.05이다. 표의 p값은 0.0519로 0.05보다 크므로 **귀무가설을 기각하지 못한다**. 즉, 볍씨 종류에 따른 평균 수확량 차이가 통계적으로 유의하다고 볼 근거가 부족하다. 이는 네 모평균이 같다는 사실이 입증되었다는 뜻은 아니다.

따라서 ㉠은 3, ㉡은 4이고 검정 결과는 유의하지 않으므로 ①의 조합과 일치한다.

NoteBase

문 10. 공기업 기관장의 재임기간은 적어도 10년이라고 주장하는 정부는 이 사실을 입증하기 위하여 25개의 공기업을 조사하였다. 25개 공기업 기관장의 평균 재임기간이 10.5년이라고 할 때, 옳은 것만을 모두 고른 것은? (단, 재임기간의 모표준편차는 3년으로 알려져 있으며, 표준정규분포를 따르는 확률변수 Z에 대하여, $P(Z > 1.645) = 0.05, P(Z > 2.576) = 0.005$이다)

ㄱ. 대립가설은 '공기업 기관장의 재임기간은 10년 이하이다.'이다.

ㄴ. Z 검정통계량의 값은 $\frac{5}{6}$이다.

ㄷ. 유의수준이 5%일 때 Z 검정통계량과 1.645를 비교한다.

ㄹ. 기각역의 형태는 $\{\overline{X} < c\}$이다.

① ㄱ, ㄴ

② ㄱ, ㄹ

③ ㄴ, ㄷ

④ ㄷ, ㄹ

정답 3번

간단한 해설

정부의 주장을 뒷받침하려는 방향은 평균 재임기간이 10년보다 길다는 쪽이므로 **우측 단측검정**을 한다.
검정통계량은 $\frac{5}{6}$이며, 유의수준 5%에서는 상위 꼬리확률이 0.05인 임계값 1.645와 비교한다.

NoteBase

보기별 해설

모평균 재임기간을 μ, 표본평균을 $\overline{X}$라 하자. 표본크기는 25, 관측된 표본평균은 10.5년이며, 모표준편차는 3년으로 알려져 있다. 이 조건에서 문제의 Z 검정을 적용한다.

ㄱ. X: 정부의 주장을 입증하려면 평균 재임기간이 10년보다 길다는 방향의 증거가 필요하다. 따라서 가설은 $H_0 : \mu \le 10, H_1 : \mu > 10$으로 설정하고, 귀무가설의 경계값인 10년을 기준으로 검정한다. 대립가설은 '10년 이하'가 아니라 **'10년 초과'**이다.

ㄴ. O: 표본평균의 표준오차는 $\frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{25}} = 0.6$년이다. 이는 개별 기관장의 재임기간이 아니라 **표본평균의 변동크기**를 나타낸다. 검정통계량은 $Z = \frac{\overline{X} - 10}{\sigma/\sqrt{n}}$이므로, 관측값은 $\frac{10.5 - 10}{0.6} = \frac{5}{6}$이다. 검정통계량에는 단위가 없다.

ㄷ. O: 대립가설이 $\mu > 10$이므로 **오른쪽 꼬리에 유의수준 5%를 배정**한다. 주어진 $P(Z > 1.645) = 0.05$에 따라 임계값은 1.645이며, 검정통계량이 이를 초과하면 귀무가설을 기각한다. 관측값 $\frac{5}{6}$은 1.645보다 작으므로 귀무가설을 기각하지 못한다.

ㄹ. X: 우측 단측검정에서는 **표본평균이 클수록** 평균 재임기간이 10년보다 길다는 대립가설을 지지한다. 따라서 기각역은 $\{\overline{X} > c\}$의 형태이다. 여기서 c는 표본평균에 대한 임계값이며, 보기의 $\{\overline{X} < c\}$는 좌측 단측검정의 기각역 형태이다.

옳은 보기는 ㄴ, ㄷ이므로 ③의 조합과 일치한다.

NoteBase

문 11. $P(A) = 0.4$**,** $P(B) = 0.3$**,** $P(B \mid A) = 0.5$**일 때,** $P(A \cup B)$**의 값은?**

① 0.45

② 0.5

③ 0.55

④ 0.7

정답 2번

간단한 해설

조건부확률로 구한 두 사건의 **교집합 확률은 0.2**이다. 합집합 확률은 두 사건의 확률을 더한 뒤 중복된 교집합 확률을 한 번 빼므로 0.4 + 0.3 − 0.2 = 0.5이다.

계산 과정

1. 교집합 확률 구하기: $P(B \mid A) = 0.5$는 **A가 일어났다는 조건에서 B가 일어날 확률**이 0.5라는 뜻이다. $P(A) = 0.4 > 0$이므로 조건부확률의 정의 $P(B \mid A) = \frac{P(A \cap B)}{P(A)}$를 적용한다. 따라서 $P(A \cap B) = P(A) \cdot P(B \mid A)$이며, $0.4 \times 0.5 = 0.2$이다. 이 값은 A와 B가 모두 일어날 확률이다.

2. 합집합 확률 구하기: $A \cup B$는 A와 B 중 적어도 하나가 일어나는 사건이다. 두 사건의 확률을 단순히 더하면 둘 다 일어나는 경우가 두 번 포함되므로 **교집합 확률을 한 번 뺀다**. 덧셈법칙은 $P(A \cup B) = P(A) + P(B) - P(A \cap B)$이다. 주어진 값과 앞에서 구한 0.2를 대입하면 0.4 + 0.3 − 0.2 = 0.5로, ②의 값과 일치한다.

NoteBase

문 12. 확률변수 X가 정규분포 $N(100, 10^2)$를 따른다고 할 때, 옳지 않은 것은?

① $X/10$의 분산은 1이다.

② $X/100$의 표준편차는 1이다.

③ $(X - 100)$의 분산은 100이다.

④ $(X - 100)/10$은 표준정규분포를 따른다.

정답 2번

간단한 해설

확률변수를 100으로 나누면 표준편차도 **100으로 나누어진다**. 따라서 $X/100$의 표준편차는 $\frac{10}{100} = 0.1$이며, 1이 아니다.

선지별 해설

$X \sim N(100, 10^2)$이므로 평균은 100, 분산은 100, 표준편차는 10이다.

① O: 확률변수에 상수를 곱하면 평균과의 편차에도 그 상수가 곱해진다. 분산은 편차를 제곱하여 평균한 값이므로 **배율의 제곱**만큼 변한다. 따라서 $\mathrm{Var}(X/10) = \frac{100}{10^2} = 1$이다.

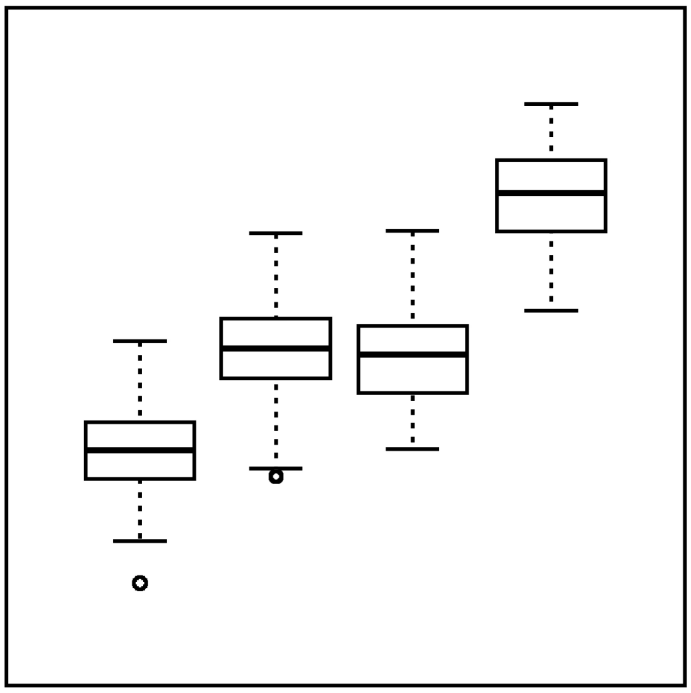
② X: 표준편차는 분산의 양의 제곱근이므로 확률변수에 곱한 **상수의 절댓값**만큼 변한다. 따라서 $X/100$의 표준편차는 $\frac{10}{100} = 0.1$이다. 선지의 1을 0.1로 고쳐야 한다.

③ O: 확률변수에서 상수를 빼면 평균도 같은 만큼 줄어들어 평균과의 편차는 그대로이다. 따라서 **분산은 변하지 않으며**, $\mathrm{Var}(X - 100) = 100$이다.

④ O: $Z = \frac{X-100}{10}$은 평균 100을 빼고 표준편차 10으로 나누는 **표준화**이다. 평균은 0이 되고 분산은 $\frac{100}{10^2} = 1$이 된다. 정규확률변수를 이처럼 선형변환한 결과도 정규분포를 따르므로 $Z \sim N(0, 1)$, 즉 표준정규분포를 따른다.

NoteBase

문 13. 다음은 4개 시의 단위면적당 주택가격의 자료를 이용하여 그린 상자그림이다. 단위면적당 평균 주택가격이 4개 시에 따라 차이가 있는지 검정하기 위한 분석법은?



① 분산분석

② 상관분석

③ 교차분석

④ 시계열분석

정답 1번

간단한 해설

시를 기준으로 나눈 네 집단의 주택가격 평균을 비교하는 문제이므로 **일원분산분석**을 사용한다. 집단 간 평균 차이를 집단 내 자료의 변동과 비교하여 판단한다.

NoteBase

공통 해설

분산분석은 여러 집단의 모집단 평균이 같은지를 검정하는 방법이다. 이 문제에서 집단을 나누는 요인은 시이고, 비교하는 수치형 변수는 단위면적당 주택가격이다. 하나의 요인으로 구분된 네 집단의 평균을 비교하므로 **일원분산분석**에 해당한다.

귀무가설은 네 시의 모집단 평균 주택가격이 모두 같다는 것이고, 대립가설은 적어도 한 시의 평균이 다르다는 것이다. 각 시 안에서도 주택가격은 서로 다를 수 있으므로, 시별 표본평균의 차이만 보지 않고 **집단 간 변동을 집단 내 변동과 비교**한다. 집단 내 변동에 비해 집단 간 변동이 충분히 크면 평균이 모두 같다는 귀무가설을 기각한다. 이 결과만으로 모든 시의 평균이 서로 다르다고 결론짓는 것은 아니다.

통상적인 일원분산분석은 관측값의 독립성, 집단별 정규성, 집단 간 등분산성을 전제로 한다. 제시된 네 상자그림은 시별 분포를 비교하는 자료이며, 상자 안의 굵은 선은 평균이 아니라 중앙값이다. 따라서 상자그림의 위치 차이만으로 평균 차이의 통계적 유의성을 확정할 수는 없다.

NoteBase

문 14. 서로 다른 값을 가지는 세 개 이상의 연속형 자료를 요약할 때 사용하는 기술통계량에 대한 설명으로 옳은 것은?

① 자료의 개수가 적을수록 평균과 분산에서 극단값의 영향은 작아진다.

② 최댓값이 현재 값보다 큰 값으로 바뀌면 중앙값은 커진다.

③ 최댓값이 α만큼 증가하고, 최솟값이 α만큼 감소하여도 변동계수(coefficient of variation) 값은 바뀌지 않는다.

④ 평균보다 작은 자료의 개수가 평균보다 큰 자료의 개수보다 많으면 중앙값은 평균보다 작다.

정답 4번

간단한 해설

평균보다 작은 자료가 더 많으면 **크기순으로 놓은 자료의 가운데도 평균보다 낮아진다**. 자료 수가 홀수이면 가운데 값이 평균보다 작고, 짝수이면 가운데 두 값의 평균이 전체 평균보다 작다.

선지별 해설

① X: 자료 수가 적을수록 극단값 하나가 평균과 분산에 미치는 영향은 일반적으로 **더 커진다**. 자료 수가 n일 때 한 값만 Δ만큼 바꾸면 평균은 Δ/n만큼 변하므로, 같은 크기의 변화라도 자료 수가 적을수록 평균의 변화가 크다. 분산도 평균에서 벗어난 정도를 제곱하여 합한 뒤 자료 수 또는 자료 수에서 1을 뺀 값으로 나누므로, 적은 자료에서는 극단값 하나의 큰 편차가 차지하는 비중이 커진다.

② X: 중앙값은 자료를 크기순으로 정렬했을 때 가운데에 있는 값이다. 자료가 세 개 이상이므로 **최댓값만 더 크게 바꾸어도 가운데 값들은 바뀌지 않는다**. 홀수 개일 때는 가운데 값이, 짝수 개일 때는 가운데 두 값의 평균이 그대로이므로 중앙값은 변하지 않는다.

③ X: 변동계수는 평균이 0이 아닐 때 표준편차를 평균으로 나눈 값으로, 표본 자료에서는 $CV = s/\bar{x}$로 나타낸다. 최댓값을 α만큼 늘리고 최솟값을 같은 양만큼 줄이면 합계가 같아 **평균은 그대로이지만 표준편차는 커진다**. 두 극단값이 각각 변하지 않은 평균에서 더 멀어져 편차제곱합이 증가하기 때문이다. 따라서 변동계수도 바뀐다.

④ O: 자료를 크기순으로 정렬하면 평균보다 작은 값들이 앞쪽에 놓인다. 서로 다른 값들이므로 평균과 같은 자료는 많아야 하나이고, 평균보다 작은 자료의 개수가 평균보다 큰 자료의 개수보다 많으면 **가운데 위치가 평균보다 작은 쪽에 포함된다**. 자료 수가 홀수이면 가운데 값 자체가 평균보다 작다. 짝수이면 가운데 두 값 중 적어도 하나는 평균보다 작고 다른 하나도 평균 이하이므로, 두 값의 평균인 중앙값은 전체 평균보다 작다.

NoteBase

문 15. 다음 기댓값 또는 분산 중에서 값이 가장 큰 것은?

① 구간 [1,3]에서 정의된 균일(uniform)분포를 따르는 확률변수 X에 대하여 $E(X)$의 값

② 평균 $\frac{1}{2}$인 확률변수 X에 대하여 $E(6X)$의 값

③ 서로 독립이고 분산이 각각 3과 1인 확률변수 X와 Y에 대하여 $Var(X-Y)$의 값

④ 이항분포 $B(20, 0.2)$를 따르는 확률변수 X에 대하여 $Var(X)$의 값

정답 3번

간단한 해설

서로 독립인 확률변수는 **차의 분산도 두 분산의 합**으로 구하므로 $Var(X-Y)=3+1=4$이다. 이 값이 나머지 선지의 기댓값 또는 분산보다 크다.

선지별 해설

① 구간 [1,3]의 균일분포는 구간의 중점을 기준으로 대칭이므로 기댓값은 **양 끝값의 평균**이다. 따라서 $E(X)=\frac{1+3}{2}=2$이다.

② 기댓값의 선형성에 따라 확률변수에 상수를 곱하면 기댓값에도 **그 상수를 그대로 곱한다**. 따라서 $E(6X)=6E(X)=6\cdot\frac{1}{2}=3$이다.

③ 두 확률변수가 **독립이면 공분산이 0**이다. 차의 분산에서 빼는 쪽의 계수 −1은 제곱되어 $(-1)^2=1$이 되므로, 분산을 빼지 않고 더한다. 따라서 $Var(X-Y)=Var(X)+Var(Y)$이고, 그 값은 $3+1=4$이다.

④ 이항분포 $B(n,p)$에서 n은 시행 횟수, p는 각 시행의 성공확률이며, **분산은** $np(1-p)$이다. 여기서는 $n=20, p=0.2$이므로 $Var(X)=20\cdot0.2\cdot0.8=3.2$이다.

네 값을 비교하면 $4>3.2>3>2$이므로, ③의 분산이 가장 크다.

NoteBase

문 16. 다음 자료에 단순선형회귀모형 $Y_i = \alpha + \beta X_i + \varepsilon_i$, $(i = 1, 2, \cdots, 7)$을 적용하려고 한다. 회귀분석 결과가 다음과 같을 때, 옳은 것만을 모두 고른 것은?

i	1	2	3	4	5	6	7
X	1	2	3	5	7	9	8
Y	6	9	10	12	15	13	17

	회귀계수	표준오차	$t-$값	$p-$값
상수항	6.3695	1.4031	4.539	0.006
x	1.0690	0.2432	4.395	0.007
$R^2 = 0.794,\ R^2_{adj} = 0.753,\ F = 19.32\ (p-\text{값} = 0.007)$				

ㄱ. 유의수준 5%에서 귀무가설 $H_0 : \beta = 0$을 기각한다.

ㄴ. 결정계수는 75% 이상이다.

ㄷ. X와 Y의 표본상관계수는 0.9보다 크다.

ㄹ. 유의수준 1%에서 단순선형회귀직선은 통계적으로 유의하다.

① ㄱ, ㄴ

② ㄴ, ㄷ

③ ㄱ, ㄴ, ㄹ

④ ㄱ, ㄷ, ㄹ

정답 3번

간단한 해설

절편을 포함한 단순선형회귀에서 결정계수는 **표본상관계수의 제곱**이다. 이 자료는 기울기가 양수이므로, 결정계수 0.794에 대응하는 표본상관계수는 약 0.891로 **0.9보다 작다**.

NoteBase

보기별 해설

ㄱ. O: 기울기에 관한 귀무가설 $H_0 : \beta = 0$은 X와 Y 사이에 선형관계가 없다는 가설이다. 회귀분석 표의 x행에서 t값은 4.395, p값은 0.007이다. 이 p값이 **유의수준 0.05보다 작으므로** 귀무가설을 기각한다. 상수항의 p값 0.006은 절편에 대한 검정값이므로 여기서 사용할 값이 아니다.

ㄴ. O: 결정계수 R^2는 Y의 전체 변동 중 회귀직선으로 설명되는 변동의 비율이다. 표에 제시된 $R^2 = 0.794$는 **전체 변동의 79.4%를 설명한다**는 뜻이므로 75% 이상이다. 수정결정계수 $R^2_{adj} = 0.753$과 구별한다.

ㄷ. X: **절편을 포함한 단순선형회귀**에서는 결정계수와 표본상관계수 r 사이에 $R^2 = r^2$이 성립하고, r의 부호는 추정 기울기의 부호와 같다. 추정 기울기가 1.0690으로 양수이므로 $r = \sqrt{0.794} \approx 0.891$이다. 따라서 표본상관계수는 0.9보다 큰 것이 아니라 약 0.891로, 0.9보다 작다.

ㄹ. O: 회귀직선의 유의성은 회귀모형의 F검정으로 판단한다. 표에서 F값은 19.32, p값은 0.007이므로 **유의수준 0.01에서도 유의하다**. 설명변수가 하나인 이 모형에서는 회귀직선의 F검정과 기울기에 대한 양측 t검정이 같은 귀무가설을 검정하므로, ㄱ과 동일한 p값으로 판단할 수 있다.

옳은 보기는 ㄱ, ㄴ, ㄹ이므로 이 조합에 해당하는 것은 ③이다.

NoteBase

문 17. $X_1, X_2, \cdots, X_n$**이 정규분포** $N(\mu, \sigma^2)$**을 따르는 모집단에서의 임의표본(random sample)이고** $\overline{X}$**를 표본평균,** S^2**을 표본분산이라고 할 때, 가설** $H_0 : \mu = \mu_0$ **대** $H_1 : \mu \neq \mu_0$**의 검정에 대한 설명으로 옳지 않은 것은? (단, 모집단** σ^2**은 알려져 있지 않으며,** $t_\alpha(k)$**는 자유도가** k**인** t **분포의 제 100×(1−α) 백분위수이다)**

① 검정통계량으로 $t = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$을 사용한다.

② 검정통계량 t에 대하여 $|t| < t_{\alpha/2}(n-1)$일 때 귀무가설을 기각한다.

③ 유의수준 5%에서 귀무가설을 기각하지 못하면 μ의 95% 신뢰구간에 μ_0가 포함된다.

④ 유의확률(p-값)이 유의수준보다 작으면 귀무가설을 기각한다.

정답 2번

간단한 해설

②는 기각역의 부등호 방향을 거꾸로 제시했다. 양측 t 검정에서는 검정통계량의 절댓값이 **임계값보다 클 때** 귀무가설을 기각한다.

선지별 해설

① O: 정규모집단의 분산을 모르므로 모표준편차 대신 표본표준편차 S를 사용한다. 표본평균의 표준오차를 $S/\sqrt{n}$으로 추정하면 검정통계량은 $t = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$이다. 귀무가설이 참일 때 이 통계량은 **자유도가** $n-1$**인 t 분포**를 따른다.

② X: 대립가설이 $\mu \neq \mu_0$이므로 표본평균이 가정한 모평균보다 크게 벗어나거나 작게 벗어나는 경우를 모두 살피는 양측검정이다. 유의수준 α를 양쪽 꼬리에 $\alpha/2$씩 배분하므로 기각 조건은 $|t| > t_{\alpha/2}(n-1)$이다. 선지의 **작다는 부등호를 크다는 부등호로** 고쳐야 한다. $|t| < t_{\alpha/2}(n-1)$이면 통계량이 중앙 영역에 있으므로 귀무가설을 기각하지 않는다.

③ O: 이 검정에 대응하는 모평균의 95% 신뢰구간은 $\overline{X} \pm t_{0.025}(n-1)\frac{S}{\sqrt{n}}$이다. 유의수준 5%에서 귀무가설을 기각하지 못한다는 것은 $|t| \leq t_{0.025}(n-1)$이라는 뜻이다. 즉, $\overline{X}$와 μ_0 사이의 거리가 신뢰구간의 오차한계를 넘지 않으므로 **신뢰구간에** μ_0**가 포함된다**.

④ O: p-값은 귀무가설이 참이라는 가정 아래 관측된 검정통계량만큼 또는 그보다 극단적인 값이 나올 확률이다. 이 양측검정에서는 관측된 절댓값 이상의 양쪽 꼬리확률을 합한다. p-값이 **유의수준보다 작으면** 귀무가설 아래에서 드문 결과가 관측되었다고 판단하여 귀무가설을 기각한다.

NoteBase

문 18. 다음은 일원분산분석(one-way ANOVA)을 수행하여 얻은 분산분석표의 일부이다. 이에 대한 설명으로 옳지 않은 것은?

요인	자유도	$F-$값	$p-$값
처리	2	2.78	0.08
오차	27		

① 처리의 수는 3이다.

② 자료의 수는 30이다.

③ 유의수준 5%에서 각 처리의 평균이 모두 동일하다는 귀무가설을 채택한다.

④ 유의수준 1%에서 기각역이 $\{F > c\}$일 때 $c < 2.78$이다.

정답 4번

간단한 해설

관측된 F값 2.78의 p값은 0.08로, 유의수준 1%인 0.01보다 크다. 오른쪽 꼬리확률을 0.01로 줄이려면 임계값을 더 크게 잡아야 하므로 **임계값 c는 2.78보다 크다**.

NoteBase

선지별 해설

① O: 처리의 수를 k라고 하면 일원분산분석에서 **처리 자유도는** $k-1$이다. 표의 처리 자유도가 2이므로 $k-1=2$에서 $k=3$이다.

② O: 전체 자료의 수를 N, 처리의 수를 k라고 하면 **오차 자유도는** $N-k$이다. 각 처리에서 평균 하나씩을 추정하므로 전체 자료 수에서 처리 수만큼 자유도가 줄어든다. 표의 오차 자유도가 27이고 처리 수가 3이므로 $N-3=27$에서 $N=30$이다.

③ O: 귀무가설은 세 처리의 모평균이 모두 같다는 $H_0 : \mu_1 = \mu_2 = \mu_3$이다. 표의 p값 0.08이 유의수준 0.05보다 크므로 **귀무가설을 기각하지 않는다**. 선지의 '채택한다'는 이 의미로 이해한다. 평균이 같다는 사실이 입증되었다는 뜻은 아니며, 평균이 다르다고 판단할 증거가 충분하지 않다는 뜻이다.

④ X: 분산분석은 F값이 클수록 귀무가설에 반하는 증거가 커지는 오른쪽 꼬리 검정이다. 귀무가설 아래에서 표의 p값은 $P(F>2.78)=0.08$을 뜻한다. 유의수준 1%의 임계값 c는 $P(F>c)=0.01$을 만족해야 한다. 오른쪽 꼬리확률을 0.08에서 0.01로 줄이려면 경계값을 오른쪽으로 옮겨야 하므로, 선지의 부등호는 $c>2.78$로 고쳐야 한다.

NoteBase

문 19. 정규분포를 따르는 모집단에서 n개의 임의표본을 추출하여 모평균 μ에 대한 추론을 하려고 한다. 옳은 것만을 모두 고른 것은? (단, 모표준편차는 알려져 있는 값이다)

ㄱ. n이 일정할 때 μ에 대한 신뢰구간의 길이는 신뢰수준이 증가할수록 길어진다.
ㄴ. 오차의 한계가 d로 주어질 때 μ를 추정하기 위한 표본의 크기는 신뢰수준이 증가할수록 커진다.
ㄷ. 가설 $H_0 : \mu = 5$ 대 $H_1 : \mu > 5$에서 Z 검정의 유의확률 (p-값)은 표본평균의 관측값이 증가할수록 작아진다.
ㄹ. 가설 $H_0 : \mu = 5$ 대 $H_1 : \mu > 5$에서 Z 검정의 검정력은 μ가 5보다 클수록 증가한다.

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

정답 4번

간단한 해설

우측검정에서는 실제 모평균이 5에서 오른쪽으로 멀어질수록 표본평균이 **기각 기준을 넘을 확률**이 커진다.
따라서 다른 조건이 같으면 실제 모평균이 5보다 커질수록 검정력이 증가한다.

NoteBase

보기별 해설

모표준편차를 σ, 표본평균을 $\bar{X}$라 하면, 정규모집단의 임의표본에서 표본평균의 표준오차는 σ/ n이다. 신뢰수준은 $1-\alpha$로 나타내고, z_q는 표준정규분포에서 그 값 이하일 확률이 q인 분위수로 둔다.

ㄱ. O: 모표준편차를 알 때 모평균의 양측 신뢰구간은 X± z_1-α/2(σ)/(n)이다. 그 길이는 2z_1-α/2(σ)/(n)이다. 표본 크기와 모표준편차가 일정하면, 신뢰수준이 높아질수록 임계값 $z_{1-\alpha/2}$가 커져 **신뢰구간의 길이가 길어진다**. 더 높은 신뢰수준을 얻으려면 모평균을 포함하도록 구간을 더 넓게 잡아야 한다.

ㄴ. O: 오차의 한계는 신뢰구간 길이의 절반으로, z_1-α/2(σ)/(n)이다. 이를 주어진 d 이하로 유지하려면 $n \geq \left(\frac{z_{1-\alpha/2}\sigma}{d}\right)^2$을 만족해야 한다. 신뢰수준이 높아지면 임계값이 커지므로, **같은 오차의 한계를 유지하려면 표본 크기를 늘려야 한다**. 실제 표본 크기는 이 하한 이상인 정수로 정한다.

ㄷ. O: 귀무가설 $H_0 : \mu = 5$ 아래에서 검정통계량 Z=(X-5)/(σ/ n)는 표준정규분포를 따른다. 대립가설이 $H_1 : \mu > 5$인 우측검정이므로, 유의확률은 **귀무가설 아래에서 관측된 검정통계량 이상이 나올 확률**이다. 표본 크기와 모표준편차가 일정할 때 표본평균의 관측값이 증가하면 Z의 관측값도 커진다. 이에 따라 오른쪽 꼬리확률인 유의확률은 작아진다.

ㄹ. O: 검정력은 대립가설에 해당하는 실제 모평균 아래에서 귀무가설을 기각할 확률이다. 유의수준 α, 표본 크기와 모표준편차를 고정하면, 기각 기준은 c=5+z_1-α(σ)/(n)이고 $\bar{X} > c$일 때 기각한다. 실제 모평균이 5보다 큰 범위에서 더 커지면 표본평균의 분포가 오른쪽으로 이동하므로, **고정된 기각 기준 c를 넘을 확률이 증가한다**. 따라서 검정력도 증가한다.

ㄱ, ㄴ, ㄷ, ㄹ이 모두 성립하므로 이들을 모두 포함한 조합은 ④이다.

NoteBase

문 20. 다음은 지난 1년간 두 손해보험 상품 A, B에 대하여 계약자 및 해지자의 수를 조사한 결과이다. 두 손해보험 상품의 계약 해지 비율이 서로 같다고 할 수 있는지 유의수준 5%에서 검정할 때 옳지 않은 것은?

	계약자 수	계약 해지자 수
A	300	60
B	200	40

① 두 손해보험 상품에 대한 계약 해지 비율의 추정값은 같다.

② 귀무가설이 참일 때 조사 대상 전체 인원에 대해 손해보험 상품의 계약을 해지한 사람의 수는 이항분포를 따른다.

③ 두 손해보험 상품의 계약 해지 비율의 추정값에 대한 표준오차는 같다.

④ 검정통계량의 값은 0이다.

정답 3번

간단한 해설

두 상품의 해지 비율 추정값은 모두 0.2이지만, **표준오차는 계약자 수에도 좌우된다**. 계약자가 300명인 A보다 200명인 B의 해지 비율 추정값이 더 크게 변동하므로, B의 표준오차가 더 크다.

NoteBase

선지별 해설

상품 A, B의 모집단 계약 해지 비율을 각각 p_A, p_B라 하면, 검정 가설은 $H_0 : p_A = p_B$, $H_1 : p_A \neq p_B$이다. 각 계약자의 해지 여부를 서로 독립인 베르누이 시행으로 보는 두 표본 비율 검정 모형을 적용한다.

① O: 표의 계약자 수를 분모, 계약 해지자 수를 분자로 삼는다. A의 해지 비율 추정값은 $\hat{p}_A = \frac{60}{300} = 0.2$, B는 $\hat{p}_B = \frac{40}{200} = 0.2$이므로 둘 다 20%이다.

② O: 귀무가설이 참이면 두 상품의 해지확률은 공통값 p이다. 독립인 각 계약자의 해지 여부는 해지확률이 p인 베르누이 시행이므로, 전체 500명 중 해지자 수 X는 **공통 해지확률을 갖는 독립 시행의 성공 횟수**이다. 따라서 $X \sim \text{Bin}(500, p)$이다.

③ X: 표본비율의 표준오차는 $\sqrt{\frac{p(1-p)}{n}}$이며, 이를 추정할 때는 p에 표본비율을 대입한다. 두 상품 모두 표본비율이 0.2이지만 **표본크기가 서로 다르므로 표준오차도 다르다**. A의 추정 표준오차는 $\sqrt{\frac{0.2 \cdot 0.8}{300}} \approx 0.0231$, B는 $\sqrt{\frac{0.2 \cdot 0.8}{200}} \approx 0.0283$이다. 따라서 표준오차는 B가 더 크다.

④ O: 귀무가설 아래에서 공통 해지확률은 두 집단을 합쳐 $\hat{p} = \frac{60+40}{300+200} = 0.2$로 추정한다. 독립인 두 표본비율 차이의 추정 표준오차를 s라 하면 $s = \sqrt{0.2 \cdot 0.8\left(\frac{1}{300} + \frac{1}{200}\right)}$이다. 두 비율의 차이를 이 표준오차로 나누는 검정통계량은 $Z = \frac{\hat{p}_A - \hat{p}_B}{s} = 0$이다. **관측된 두 비율의 차이가 0**이므로 검정통계량도 0이며, 유의수준 5%의 양측검정에서 귀무가설을 기각하지 않는다.

NoteBase